Nouvelle approche d'adaptation de modèle et d'intégration des données pour les recherches de médecine de précision

par

Fodil BELGHAIT

THÈSE PRÉSENTÉE À L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
COMME EXIGENCE PARTIELLE À L'OBTENTION
DU DOCTORAT EN GÉNIE
Ph. D.

MONTRÉAL, LE 05 OCTOBRE 2020

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

PRÉSENTATION DU JURY

CETTE THÈSE A ÉTÉ ÉVALUÉE

PAR UN JURY COMPOSÉ DE :

M. Alain April, directeur de thèse
Département de génie logiciel et TI à l'École de technologie supérieure

M. Christian Desrosiers, co-directeur
Département de génie logiciel et TI à l'École de technologie supérieure

M. Marc Paquet, président du jury
Département de génie des systèmes à l'École de technologie supérieure

M. Abdelaoued Gherbi, membre du jury
Département de génie logiciel et des TI à l'École de technologie supérieure

M. Robert Dupuis, examinateur externe
Professeur département informatique UQAM

ELLE A FAIT L'OBJET D'UNE SOUTENANCE DEVANT JURY ET PUBLIC

LE 22 SEPTEMBRE 2020

À L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

REMERCIEMENTS

Mes principaux remerciements s'adressent à mon directeur de recherche, le professeur Alain April pour la confiance qu'il m'a accordée, pour ses commentaires pertinents et surtout pour ses précieuses remarques et son soutien continu qui m'a grandement aidé à compléter cette thèse.

Je voudrais aussi exprimer mes vifs remerciements aux Dr Pavel Hamet, Dr Johanne Tremblay et les membres de leur équipe de recherche pour leur aide dans la réalisation de l'expérimentation du prototype. Je remercie aussi tous les membres de jury pour leur temps et effort à réviser cette thèse et leurs commentaires précieux.

Je remercie tous ceux et celles qui durant toutes ces cinq années, m'ont soutenu, motivé, écouté, orienté et qui, de près ou de loin ont contribué à la bonne réalisation de cette thèse. Je dédie cette thèse à ma famille, principalement à mes parents, mes frères, ma chère épouse Khadija et mes trois chers enfants Leila, Chakib et Sarah.

Nouvelle approche d'adaptation de modèle et d'intégration des données pour les recherches de médecine de précision

Fodil BELGHAIT

RÉSUMÉ

L'adoption massive des technologies de la génétique à haut débit, des puces de séquençage des données génétiques à haute vitesse a permis une accélération des recherches complexes en matière de médecine de précision grâce à l'utilisation d'un volume considérable de données médicales hétérogènes. Avec de telles données à leur disposition, les chercheurs et cliniciens peuvent élaborer des stratégies adaptées à chaque individu pour proposer des traitements préventifs et thérapeutiques plus précis en ciblant des sous-groupes de patients sur des maladies spécifiques à partir d'un grand nombre de données génétiques, cliniques, démographiques, et autres types d'information reliée aux modes de vie des patients. Les techniques de séquençage de nouvelle génération jouent un rôle clé dans la recherche en médecine de précision; cependant, la complexité, la diversité, l'hétérogénéité, le volume des données et l'inaptitude des logiciels d'analyse de données disponibles à effectuer leur fonction facilement sur cet ensemble de données représentent un défi de taille pour une utilisation directe par le chercheur. Cette thèse, du domaine du génie logiciel, propose une piste de solution pour les chercheurs en médecine de précision. Elle propose une nouvelle approche pour adapter et intégrer toutes les données nécessaires à ces recherches dans un seul endroit, accompagnées par un cycle de recherche de médecine de précision visant à simplifier l'étape de préparation des analyses. Cette recherche tente d'adresser ces problématiques auxquelles sont confrontés les chercheurs de ce domaine. Pour valider cette proposition, les résultats d'une expérimentation d'analyse de données d'une recherche du domaine de médecine de précision effectuée à l'aide d'un prototype expérimental sur un réel cas d'une recherche qui consiste à développer un modèle prédictif qui permettrait d'identifier les patients à risque de développer la maladie d'insuffisances rénales chronique (de l'anglais Chronic Kidney Disease (CKD)) chez les patients atteints par le diabète de type 2 (DT2).

Mots-clés : médecine de précision, génotypage, base de données, infonuagique, données massives, bio-informatique.

New modeling and data processing approach for precision medicine research

Fodil BELGHAIT

ABSTRACT

Massive adoption of high-throughput genetics technologies, deoxyribonucleic acid (DNA) chips, and high-speed sequencing has accelerated complex research in precision medicine through the use of high-volume considerable amount of heterogeneous medical data. With such data at their disposal, researchers and clinicians can develop strategies tailored to each individual to provide more precise preventive and therapeutic treatments by targeting subgroups of patients on specific diseases from a large amount of data. genomics, clinical, demographic, and other types of information related to patients' lifestyles. Next Generation Sequencing (NGS) techniques play a key role in research in precision medicine; however, the complexity, diversity, heterogeneity, volume of data and the inability of the data analysis software available to perform their function easily on this dataset represents a challenge for direct use by the researcher. This thesis, from the field of software engineering, proposes a solution track for researchers in precision medicine. It proposes a new approach to adapt the data model to the research new data requirements, unify the storage location and format of all research data accompanied by a precision medicine research cycle aimed at simplifying the phase of preparing medical precision medicine analyzes. This research attempts to address the main issues facing researchers in this field. To validate this proposition, the results of a precise medicine research data analysis experiment were carried out using an experimental prototype on a real case of a research which consists in developing a predictive model that would make it possible to identify the patients at risk of developing Chronic Kidney Disease (CKD) disease in patients with type 2 diabetes (T2D).

Keywords: precision medicine, genotyping, database, cloud computing, big data, bioinformatics.

TABLE DES MATIÈRES

Page

LISTE DES TABLEAUX

Page

LISTE DES FIGURES

Page

LISTE DES ALGORITHMES

LISTE DES ABRÉVIATIONS, SIGLES ET ACRONYMES

Apache Avro	Cadriciel de sérialisation des données. Il utilise le format JSON pour la définition des types de données
ADN	Acide désoxyribonucléique
ADVANCE	(Atherosclerotic Disease, Vascular FuNction & GenetiCEpidemiology). Une étude clinique approuvée par le conseil d'approbation institutionnel de l'université Stanford
ADAM	Projet proposé par l'université Berkeley, en Californie, permettant d'effectuer des transformations, des analyses et des requêtes sur de très grands volumes de données génétiques
API	Interface de Programme d'Application (Application Program Interface)
AWS	Amazon Web Services
BDD	Base De Données
CKD	Chronic Kidney Disease
CIM	Classification internationale des Maladies
CP	Current Procedural Terminology
CRCHUM	Centre de Recherche du Centre Hospitalier de l'Université de Montréal
DPI	Dossier patient informatisé
DT2	Diabète de type 2
eGFR	estimated Glomerular Filtration Rate
FN	Faux négatifs/False Negatives
FP	Faux Positifs/False Positives
GATK	Genome Analysis Kit
GWAS	Genome Wide Association
GEO	Gene Expression Omnibus
GE	General Electric
GO et GOS	Gig Octet(s)
HGI	Haemoglobin Glycation Index

IAM	Identity and Access Management
IaaS	Infrastructure as a Service
IP	Transmission Control Protocol/Internet Protocol
IROmiCS	Information Retrieval for Omic and Clinical Sciences
JSON	JavaScript Object Notation
PaaS	Plateforme as a Service
LR	Linear Regression
MP	Médecine de Précision
MPP	Massively Parallel Processing/Traitement Massivement Parallèle
MRC	Maladie Rénale Chronique
NCI	National Cancer Institute
NGS	Next Generation Sequencing
NIH	National Institutes of Health
NIST	National Institute of Standards and Technology
NN	Neural Network
RAG	Risk Allele Genotype
RAVEL	Recherche d'Information et Visualisation dans le Dossier patient Informatisé
RF	Random Forest
RS ID	Reference SNP Cluster ID
SAM	Sequence Alignment/Map format
SNP	Single Nucleotide Polymorphism
SaaS	Software as a Service
S/O	Sans Objet
SSH	Secure Shell
TP	True Positives
TN	True Negatives
UACR	Urine Albumin to Creatinine Ratio
VCF	Variant Call Format

VP	Vrai Positifs
VN	Vrai Négatifs
VCPU	Virtual Central Processing Unit, aussi appelé processeur virtuel

INTRODUCTION

0.1 Problématique et motivation de la recherche

L'adoption massive des technologies de la génétique à haut débit, des puces d'Acide DésoxyriboNucléique (ADN) et du séquençage du génome a permis une accélération de recherches complexes en matière de médecine de précision grâce à l'utilisation d'un volume considérable de données médicales hétérogènes (Davis-Turak et coll. 2017). Les termes « médecine de précision », « médecine translationnelle » et « recherche translationnelle » s'utilisent de manière interchangeable dans les ouvrages spécialisés (Duffy, 2018)(Feldman, 2015). Dans cette thèse, le terme « médecine de précision » sera utilisé pour désigner ce nouveau champ de recherche médicale qui cible la prévention et le traitement personnalisé d'une maladie en prenant en compte les différences individuelles au niveau des gènes, l'historique des traitements, l'environnement et le mode de vie du patient.

Cette approche de recherche personnalisée pour le traitement futur des patients permettra aux médecins et aux chercheurs de prévoir, avec une plus grande précision, quel type de traitement et quelles stratégies de prévention peuvent être appliqués à des groupes de patients qui souffrent d'une maladie particulière. Elle permettra également d'élaborer des stratégies thérapeutiques et préventives pour cibler les besoins spécifiques des patients en se basant sur les caractéristiques génétiques, phénotypiques, psychosociales et les biomarqueurs, qui différencient les individus les uns des autres (Vassy, Korf, et Green 2015). Pour être efficace, une recherche en médecine de précision requiert l'intégration des données génétiques et cliniques des patients afin de mieux comprendre leur maladie (Wolkenhauer et coll. 2014).

Le domaine de recherche en médecine de précision effectue des analyses sur un domaine très vaste de maladies. Les besoins informationnels nécessaires pour effectuer ces analyses varient d'une recherche à une autre, de plus, les volumes des données à traiter sont importants et nécessitent de grandes ressources de calcul. Selon les articles récents publiés qui traitent de la

médecine de précision, ce domaine émergent de recherche médicale fait face aux deux défis majeurs :

1- Le premier défi est lié à la quantité croissante des données à traiter dans les recherches de médecine de précision. Les chercheurs présentement font face à des défis majeurs concernant le stockage et le traitement des données génétiques et cliniques même quand il s'agit d'un petit nombre d'individus et de leurs gènes à analyser(Karen He, Dongliang Ge, 2017). De plus, afin d'obtenir de meilleurs résultats de prédiction, les chercheurs aimeraient explorer à l'aide d'un plus grand nombre d'individus et de gènes;

2- Le deuxième défi est lié à la nature des données et des technologies actuellement disponibles. Les chercheurs sont contraints à travailler avec plusieurs logiciels en même temps (Chloé Cabot, Lina F. Soualmia, 2015). Ceci est nécessaire, car ils doivent intégrer différentes données du patient qui proviennent de différentes sources de données. En effet, plusieurs logiciels de recherche ont été développés et mis à la disposition des chercheurs, mais chacun de ces logiciels est spécialisé pour traiter des formats des données spécifiques (Chloé Cabot, Lina F. Soualmia, 2015). En plus, les besoins informationnels nécessaires pour faire des analyses en médecine de précision varient d'une recherche à une autre. Les chercheurs doivent donc pouvoir adapter facilement le modèle des données pour chaque analyse. Les logiciels de recherche disponibles actuellement ne permettent pas facilement l'adaptation de modèle des données par le chercheur aux nouveaux besoins informationnels requis par les recherches. Dans ce contexte, le chercheur de ce domaine fait face à un nombre important de logiciels libres et commerciaux qui doivent être maîtrisés par lui ou par son équipe de bio-informaticiens. Cette situation limite la capacité des chercheurs à faire de la recherche rapidement et à faibles coûts dans ce domaine (Belghait, Kanzki, et April 2018). En effet, ces dernières années ont vu l'émergence de nombreux logiciels de recherche en médecine de précision qui visent à offrir des solutions innovantes dans le but de recueillir et d'analyser un plus grand volume de données génétiques et cliniques pouvant être utilisées dans le cadre de la recherche en médecine de précision. Cependant, ces logiciels commerciaux ont tendance à être chers, restreints à des domaines de recherche précis et spécialisés pour effectuer des analyses spécifiques. Par

exemple, des logiciels commerciaux de recherche pour les maladies de Cancer (Jang et coll. 2018, Bhuvaneshwar et coll. 2016), des logiciels commerciaux de recherche pour le diabète, d'autres sont spécialisés dans les maladies reliées à la pression artérielle (Turner, Schwartz, & Boerwinkle, 2007), et bien d'autres. Souvent les chercheurs veulent faire des analyses croisées, mais ils se retrouvent coincés par les limites des modèles des données contenues dans ces logiciels qui ne peuvent pas facilement s'adapter à leurs besoins.

La littérature du domaine précise aussi qu'il y a actuellement une absence de logiciels de recherche en médecine de précision qui sont capables de traiter les très grands volumes de données génétiques et cliniques qui augmentent quotidiennement (Prosperi et coll. 2018). De plus, il est rapporté que ces logiciels ne s'adaptent pas facilement au caractère dynamique et à la diversité des besoins informationnels des recherches en médecine de précision. Ils ne permettent pas non plus facilement d'intégrer les données nécessaires aux recherches provenant de multiples sources et qui ont différents formats de données (Prosperi et coll. 2018).

En résumé, le domaine de recherche en médecine de précision fait face aux deux défis principaux suivants :

1. Les données requises pour réaliser les recherches dans ce domaine proviennent de différentes sources, sous formes très diversifiées, et comportent une grande volumétrie qui ne cesse de croître (Karen He, Dongliang Ge, 2017). De plus, les logiciels de recherche en médecine de précision actuels ont de la difficulté à traiter ces très grands volumes de données. (Prosperi et coll. 2018);
2. Les logiciels de recherche en médecine de précision sont incapables de s'adapter facilement au caractère dynamique et diversifié des besoins informationnels des recherches en médecine de précision (Prosperi et coll. 2018). Les besoins informationnels varient plusieurs fois au cours du même cycle de recherche ce qui nécessite l'adaptation de modèle et l'intégration des données plusieurs fois (Belghait, Kanzki, et April 2018).

Ces deux défis limitent la réactivité, la créativité et la capacité des chercheurs en médecine de précision à effectuer des analyses croisées, comportant plusieurs facteurs, qui influencent le développement de pathologies. Conséquemment, cette situation ralentit considérablement la vitesse des découvertes du domaine. C'est précisément sur cette problématique que cette thèse vise à contribuer à des solutions.

0.2 But, objectif et question de recherche

0.2.1 But de la recherche

Le but de cette recherche est d'améliorer le processus itératif de préparation et d'analyse des données des recherches en médecine de précision afin d'atténuer les problématiques (présentées à la page précédente) auxquelles font face présentement les chercheurs du domaine.

0.2.2 Objectif de la recherche

Pour atteindre ce but, cette recherche vise à atteindre l'objectif suivant :

Améliorer l'approche d'adaptation de modèle des données utilisé dans le cycle de recherche en médecine de précision afin de permettre aux chercheurs de pouvoir continuellement adapter le modèle des données, d'une recherche, aux besoins informationnels changeants et permettre d'intégrer toutes ces données dans une seule base de données permettant ainsi d'effectuer des analyses sur des quantités massives de données.

Les objectifs spécifiques de ce projet de recherche sont :

1) Proposer une nouvelle façon d'adapter un modèle de données dans le contexte de recherche en médecine de précision afin de permettre aux chercheurs de répondre aux besoins informationnels hétérogènes, spécifiques et évolutifs de chacune de leurs recherches;
2) Permettre aux chercheurs du domaine de médecine de précision d'intégrer toutes les données dont ils ont besoin pour faire leur analyse de données dans une même base de données et dans le même format de stockage;

3) Optimiser le cycle de recherche actuel pour dégager les chercheurs des activités sans valeurs ajoutées et le rendre plus adaptés aux cycles de recherche itérative.

0.2.3 Questions de recherche

Afin d'atteindre cet objectif, cette recherche vise à répondre aux trois questions suivantes :

1. Comment permettre aux chercheurs en médecine de précision d'adapter eux-mêmes plus dynamiquement et efficacement, le modèle de ces données afin qu'il puisse répondre aux besoins informationnels hétérogènes, spécifiques et évolutifs de chacune de leurs recherches en médecine de précision ?
2. Comment intégrer toutes ces données hétérogènes nécessaires à une recherche en médecine de précision, c.-à-d. les regrouper dans un même format de stockage, dans un seul modèle et dans la même base de données ?
3. Comment aider ces chercheurs à réaliser plus efficacement (c.-à-d. avec moins de personnel, moins d'efforts et plus rapidement) le cycle de recherche de manière itérative ? C'est-à-dire récupérer les résultats des analyses précédentes, ajuster le tir en fonctions des résultats obtenus, ajouter/modifier des données et refaire le cycle de recherche autant de fois que nécessaire jusqu'à l'obtention des réponses aux questions de recherche ?

0.3 Organisation de la thèse

Pour répondre aux questions de recherches énoncées à la section précédente, cette thèse a été structurée en quatre chapitres et d'une conclusion de la thèse :

- INTRODUCTION : ce chapitre présente la problématique, la motivation de la présente recherche, le but et les objectifs ainsi que les questions de recherche;
- Chapitre 1 : ce chapitre présente la méthodologie de recherche adoptée pour répondre aux questions de recherche posées, ensuite il décrit la problématique à résoudre, les étapes de recherche planifiée et les résultats attendus de la solution proposée afin de pouvoir évaluer à quel point cette proposition pourrait résoudre les problématiques soulevées;

- Chapitre 2 : ce chapitre présente l'état de l'art des problèmes rencontrés par les chercheurs du domaine de la médecine de précision ainsi qu'une synthèse des travaux publiés qui tentent de résoudre ces problématiques. Plus spécifiquement l'emphase est placée sur les sujets suivants :
    - Les logiciels d'analyse de données utilisées par les chercheurs du domaine de la médecine de précision;
    - Les défis technologiques dans le traitement des données rencontrés par ces chercheurs;
    - Les orientations futures des logiciels de recherches de médecine de précision;
- Chapitre 3: ce chapitre présente les décisions de conception du nouveau modèle de cycle de recherche;
- Chapitre 4 : ce chapitre décrit les décisions de conception d'un prototype expérimental permettant de démontrer les avantages de l'approche d'adaptation de modèle et d'intégration des données proposées. À la suite de la mise en œuvre et des essais de ce prototype expérimental, une étude de cas est effectuée afin d'appuyer un chercheur lors d'une analyse de données du domaine de médecine de précision;
- Chapitre 5 : finalement, le dernier chapitre conclu cette recherche et fait une synthèse des résultats observés lors de la mise en œuvre du modèle proposé dans l'étude de cas, fait un rappel des contributions principales, de son originalité, fait un rappel des questions de recherche et précise les limites des contributions concernant la problématique soulevée dans cette thèse. En fin de chapitre, des recommandations d'avenues de recherche et de solutions pour le futur de cette recherche sont abordées.

CHAPITRE 1

MÉTHODOLOGIE DE RECHERCHE

Un cadre méthodologique de recherche populaire, du domaine du génie logiciel, est le cadre de Basili qui est adapté pour les recherches exploratoires du domaine (Basili, Selby, et Hutchens 1986). Il est utilisé pour illustrer la planification des activités de cette recherche qui ont été structurées en quatre phases: la définition de la recherche, la planification de la recherche et de l'expérimentation, le développement d'une solution originale et la préparation de son expérimentation/sa validation; et finalement et l'interprétation des résultats obtenus (voir tableau 1.1).

1.1 Phase 1: Définition

Cette première phase de la recherche consiste à bien préciser, cerner et documenter la problématique principale de la recherche. Ainsi, le lecteur peut bien comprendre ce qui est inclus et exclu de la portée de cette recherche spécifique. Premièrement, la motivation de la recherche est rappelée. Elle est suivie par la présentation des objectifs généraux. Ensuite la/les questions de recherches précises sont présentées et finalement l'identification des utilisateurs potentiels est listée. Le tableau 1.1, ci-dessous, décrit chaque élément de cette première phase de la définition précise de la recherche.

Tableau 1.1 Cadre de Basili : Phase 1- Définition de la recherche

Étape	Description
Motivation	La médecine de précision est un domaine émergent de recherche qui cible la prévention et le traitement personnalisé des maladies. Il prend en compte les différences individuelles au niveau des gènes, l'historique des traitements, l'environnement et le mode de vie des patients. Ces études, uniques à chaque patient, nécessitent l'intégration et le traitement d'une quantité de plus en plus importante de données de toute provenance incluant génétiques et cliniques. Il a été amplement publié que les chercheurs du domaine éprouvent actuellement des difficultés à effectuer ces études, principalement à cause des obstacles reliés aux quantités importantes de données et des nombreuses technologies et logiciels impliqués.
Objectifs	Cette recherche vise à atteindre l'objectif suivant: Améliorer l'approche d'adaptation de modèle des données utilisé dans le cycle de recherche en médecine de précision afin de permettre aux chercheurs de pouvoir continuellement adapter le modèle des données, d'une recherche, aux besoins informationnels changeants et permettre d'intégrer toutes ces données dans une seule base de données permettant ainsi d'effectuer des analyses sur des quantités massives de données.
Proposition /But	Concevoir une nouvelle approche d'adaptation de modèle et d'intégration des données pour les recherches de médecine de précision qui va permettre aux chercheurs de : 1. Adapter dynamiquement et continuellement par eux même le modèle des données afin qu'ils puissent répondre aux besoins informationnels hétérogènes et spécifiques de chacune de leurs recherches ? 2. Intégrer toutes les données hétérogènes nécessaires à une recherche en médecine de précision, c.-à-d. les regrouper dans le même format de stockage, dans un seul modèle et dans la même base de données? 3. Réaliser plus efficacement (c.-à-d. avec moins de personnes impliquées, moins d'effort et plus rapidement) le cycle de recherche de manière itérative
Utilisateurs	Les résultats de cette recherche seront utilisés par les trois catégories d'utilisateurs suivants : 1. Chercheurs du domaine de médecine de précision; 2. Éditeur de logiciel du domaine de recherche en médecine de précision; 3. Étudiants et chercheurs du domaine du génie logiciel qui travaillent à développer des solutions technologiques pour ce domaine de recherche.

1.2 Phase 2 : Planification

Lors de la phase de planification, les activités de la recherche et les livrables à produire afin d'atteindre les objectifs et répondre aux questions de recherche sont établis. Cette phase commence par une revue littéraire spécialisée pour bien comprendre les défis, les obstacles actuels et les solutions actuellement disponibles aux chercheurs du domaine de la médecine de précision lors de l'obtention et la préparation des données pour leur recherche et les processus/technologies dont ils disposent actuellement. Le Tableau 1.2, de la page suivante, présente les différentes étapes et activités de cette phase de la recherche ainsi que les déclencheurs et livrables de chacune d'elle.

Tableau 1.2 Cadre de Basili : Phase 2- Planification de la recherche

Étapes du projet	Déclencheurs/Entrées	Livrables
1.Définition et validation de la recherche	Examen écrit et oral	Clarification de la définition et validation de la direction de la recherche
2.Revue de littérature	La revue de la littérature couvrira les six sujets principaux suivants : 1. Les publications de recherche du domaine de la médecine de précision et ses défis en technologies de l'information; 2. Les logiciels de recherche de médecine de précision : les défis technologiques liés au stockage et aux traitements des données massives complexes et hétérogènes; 3. L'intégration des données requises provenant de différentes sources; 4. La décision des chercheurs concernant les différents logiciels à utiliser : choisir d'utiliser un logiciel existant ou de concevoir leur propre logiciel de médecine de précision; 5. Les directions futures des logiciels des recherches du domaine de médecine de précision et les pistes de solutions envisagées.	1. Listes des défis auxquels font face les chercheurs du domaine de la médecine de précision; 2. Liste des logiciels de recherche du domaine de médecine de précision ; 3. Solutions technologiques potentielles;

Étapes du projet	Déclencheurs/Entrées	Livrables
2- Revue de littérature (suite)	6. La méthode de mesure qui doit être préalablement établie avant l'expérimentation de cette recherche.	4. Publication d'articles de conférence.
1- Réalisation des travaux de recherches	1. Validation du projet de recherche	1. Définition de la nouvelle approche d'adaptation de modèle et d'intégration des données; 2. Création d'un prototype expérimental; 3. Validation de la nouvelle approche avec un réel cas de recherche; 4. Dépôt du document de la thèse.

1.3 Phase 3 : Développement et exécution des activités de la recherche

Au cours de cette phase, la majeure partie des efforts de la réalisation de la présente recherche sera effectuée à cette étape. Chacune des étapes de la nouvelle approche d'adaptation de modèle et d'intégration des données proposée pour les chercheurs en médecine de précision sera définie, conçue, développée et mise en œuvre dans un prototype expérimental qui sera validé. À la suite de la collecte de mesures, concernant le processus actuel dans le laboratoire du Dr Hamet au CRCHUM, la nouvelle approche d'adaptation de modèle et d'intégration des données sera validée à l'aide d'une étude de cas d'une recherche en médecine de précision. Deux publications (Belghait, Kanzki, et April 2018)(Belghait et April, 2018) ont été produites pour présenter les progrès des travaux lors de cette étape de la recherche dans le but de partager et valider la proposition de solution avec la communauté scientifique. Le Tableau 1.3, de la page suivante, présente les différentes activités de cette étape de la recherche.

Tableau 1.3 Cadre de Basili: Phase 3- Développement et exécution des activités de la recherche

Conception	Validation	Analyse/ Résultats
1. Préciser la séquence des activités d'une recherche de médecine de précision comportant des données massives personnalisées; 2. Concevoir l'approche d'adaptation de modèle fondée sur le modèle de données ADAM; 3. Établir une liste d'objectifs mesurables	1. Effectuer des essais de personnalisation du modèle de données et des interfaces de programme d'application (API); 2. Analyse préliminaire des données de la cohorte de l'étude clinique d'ADVANCE (Atherosclerotic Disease, Vascular FuNction, & GenetiCEpidemiology), une étude clinique approuvée par le conseil d'approbation; institutionnel de l'université de Stanford; 3. Définition d'une étude de cas, avec l'équipe de chercheurs du Dr Hamet, pour valider l'approche d'adaptation proposée pour une étude de cas d'une analyse réelle typique; 4. Définition de la méthode de mesure et des mesures qui seront collectées avant et après l'étude de cas; 5. Sondage pour obtenir les mesures de chaque objectif avant l'étude de cas; 6. Valider les hypothèses de la recherche avec la réalisation de l'étude de cas à l'aide du prototype expérimental.	1. Mesures définies; 2. Définition et conception détaillée de l'approche d'adaptation de modèle d'intégration des données et de l'architecture d'un prototype expérimental; 3. Expérimentation d'API validées sur AWS pour chaque composant du prototype expérimental; 4. Données prêtes pour l'étude de cas; 5. Mesures de référence obtenues; 6. Publication d'un article de journal présentant la proposition de solution.

Conception	Validation	Analyse/ Résultats
4. Conception de la première version des différents API sur AWS d'un prototype expérimental; 5. Développement et validation de chaque composant du prototype expérimental.	1. Essais en laboratoire de l'approche d'adaptation de modèle et d'intégration à l'aide de données publiques à l'aide d'un prototype expérimental; 2. Préparation pour l'étude de cas au laboratoire du Dr Hamet; 3. Validation des objectifs, des données et des étapes.	1. Composants logiciels du prototype validés; 2. Guide de déploiement du prototype sur AWS; 3. Prototype expérimental déployé et prêt pour l'étude de cas.
1. Validation de la proposition à l'aide d'une étude de cas réelle avec l'équipe de recherche en médecine de précision du Dr Hamet.	1. Réalisation de l'étude de cas et collecte des mesures de chaque objectif; 2. Revue des résultats de l'étude de cas avec l'équipe de recherche du laboratoire du Dr Hamet.	1. Modèle des données nécessaires pour la réalisation de l'étude de cas; 2. Description détaillée de l'étude de cas et collecte des données.

1.4 Phase 4 : Interprétation des résultats

Au cours de cette dernière phase, les activités suivantes seront réalisées : réviser le contexte d'interprétation de l'étude de cas, analyser et extrapoler les résultats obtenus, réflexion concernant le modèle proposé pour son utilisation par les utilisateurs visés par la recherche et finalement, identifier les travaux futurs pour faire évoluer cette proposition de solution. Le Tableau 1.4 ci-dessous présente les étapes de cette phase d'interprétation des résultats de la recherche.

Tableau 1.4 Cadre de Basili adapté aux projets de recherche génie logiciel : Phase 4- Interprétation

Contexte d'interprétation	Extrapolation des résultats	Perspectives futures
L'étude de cas utilisant un prototype expérimental mettant en œuvre la nouvelle approche d'adaptation de modèle et d'intégration de données a démontré avec succès qu'il est possible de : *1.* Adapter un modèle des données existant pour y ajouter les nouveaux besoins informationnels d'une recherche; *2.* Charger ces données dans le modèle des données adapté; *3.* Intégrer les nouvelles données avec les données déjà existantes dans le modèle, et ce sans grands efforts du chercheur; *4.* Ajuster facilement la capacité de traitement du logiciel proposé au volume de données de la recherche, ce qui permet au chercheur de faire des analyses sur de grands volumes de données.	Discuter des résultats de l'essai de l'approche proposée d'adaptation de modèle et d'intégration des données : 1. Comment cette approche pourrait-elle être étendue à d'autres laboratoires; 2. Quelles sont les limites de l'étude de cas effectuée; 3. Quelles sont les limites du prototype logiciel actuel.	1. Conduire des validations plus détaillées avec d'autres chercheurs et d'autres études de médecine de précision; 2. Création d'une interface utilisateur plus conviviale que les scripts actuels pour faciliter l'utilisation des différents composants du prototype expérimental par un chercheur néophyte en technologies de l'information; *3.* Améliorer l'appui à l'utilisation de modèles de prédiction.

À la suite de cette planification détaillée, la figure 1, ci-dessous, présente un résumé de la démarche méthodologique de cette recherche représentée graphiquement.

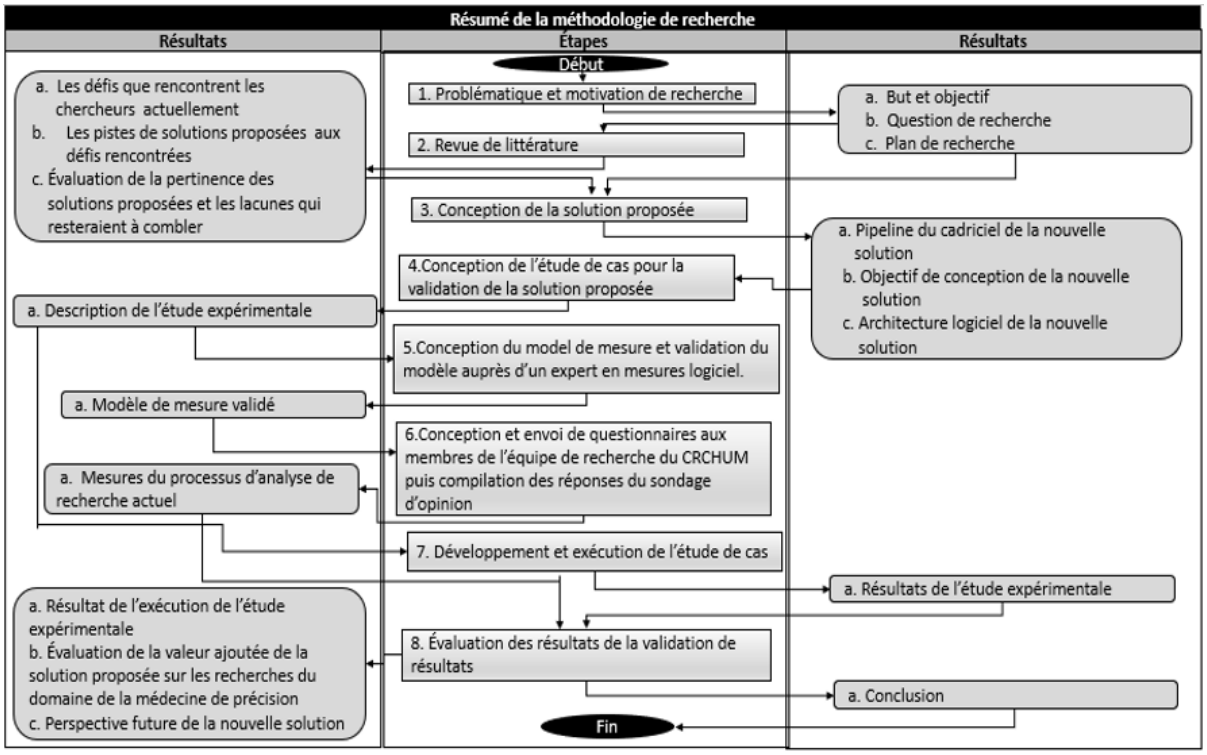


Figure 1.1 Représentation sommaire des activités du cadre de Basili de cette recherche

La méthodologie adoptée pour la réalisation de cette recherche est composée de huit étapes :

1. Problématique et motivation de recherche : la première étape consiste à documenter une problématique qui n'a pas été encore résolue;
2. Revue de littérature : une fois le problème à résoudre bien défini et documenté, à l'aide de questions de recherches et d'objectifs spécifiques à réaliser, une revue de littérature sera effectuée portant sur tous les aspects qui couvrent cette problématique dans le but d'identifier des défis et les pistes des solutions proposées dans d'autres recherches;
3. Conception de la solution proposée : à la suite de la cueillette des informations, la conception d'une solution proposée sera effectuée;
4. Conception de l'étude de cas pour la validation de la solution proposée : cette étape consiste à concevoir une étude de cas d'une réelle analyse de recherche afin d'expérimenter la solution proposée;

5. Conception du modèle de mesure et validation du modèle auprès d'un expert en mesures logiciel : afin de mesurer la valeur ajoutée de la solution proposée, s'assurer que cette solution va résoudre la problématique soulevée et va permettre d'atteindre les objectifs, un modèle de mesure de la solution sera développé et validé auprès d'un expert de la mesure logiciel;
6. Conception et envoi de questionnaires aux membres de l'équipe de recherche du CRCHUM puis compilation des réponses du sondage d'opinion : cette étape va permettre de réunir les informations nécessaires pour mesurer le processus actuel et puis les comparer avec les résultats à obtenir à partir de l'expérimentation de la nouvelle solution avec l'étude de cas défini dans la quatrième étape;
7. Développement et exécution de l'étude de cas : au cours de cette étape, l'analyse de recherche définie dans l'étude de cas sera réalisée en utilisant la solution proposée et les mesures d'évaluation de la nouvelle solution identifiée au cours de la cinquième étape seront capturées;
8. Évaluation des résultats de la validation de résultats : dans cette étape, les mesures du processus actuel obtenu à partir des réponses des membres de l'équipe de recherche du CRCHUM, seront comparées à ceux recueillis au cours de la réalisation de l'analyse de l'étude de cas afin d'évaluer à quel point la nouvelle solution a permis de résoudre la problématique soulevée.

Le chapitre suivant présente la revue de littérature. Elle couvre les sujets principaux suivants :

- Les publications de recherche du domaine de la médecine de précision, ses défis en technologies de l'information, du point de vue des chercheurs;
- Les logiciels de recherche de médecine de précision : les défis technologiques liés au stockage et aux traitements des données massives complexes et hétérogènes;
- L'intégration des données requises provenant de différentes sources;
- Le dilemme auquel les chercheurs font face entre le choix de l'utilisation des logiciels de recherche, autant commerciaux que libres versus la création de leur propre logiciel (plateforme de médecine de précision);

- Les directions futures des logiciels des recherches du domaine de médecine de précision et les pistes de solutions envisagées pour résoudre les problèmes rencontrés par les chercheurs au cours de la phase de préparation des données spécifiques de leurs recherches;
- La méthode de mesure qui doit être préalablement établie avant l'expérimentation de cette recherche.

CHAPITRE 2

REVUE DE LITTÉRATURE

2.1 Introduction

Lors de la revue de littérature, il a été constaté que les termes « médecine de précision », « médecine translationnelle » et « recherche translationnelle » sont utilisés de manière interchangeable dans les publications spécialisées(Duffy, 2018)(Feldman, 2015). Plusieurs articles qui traitent de ce sujet et qui abordent les problèmes rencontrés par les chercheurs de ces domaines et les solutions proposées ont été analysés.

Plusieurs définitions du terme médecine de précision sont énoncées dans la littérature, celle de l'institut national de santé des États-Unis (de l'anglais : National Institutes of Health (NIH)) est parmi celles qui sont les plus acceptées par la communauté scientifique. Le NIH définit la médecine de précision comme étant « an emerging approach for disease treatment and prevention that takes into account individual variability in genes, environment, and lifestyle for each person.»(Garrido et coll. 2017). Le terme médecine de précision est devenu la bannière de plusieurs projets de recherche biomédicale dans plusieurs grands pays comme les États-Unis, la Chine, ceux de l'Union européenne et bien d'autres encore.

Ce domaine de recherche est composé de deux phases importantes. Premièrement, faire des diagnostics précis, et deuxièmement prescrire des traitements personnalisés. Afin de pouvoir diagnostiquer de manière précise, il est impératif de construire des profils individuels précis des patients à partir de leurs données génétiques, cliniques, démographiques, ainsi qu'à l'aide de données sur leur historique familial et médical, des données sur le risque de leur exposition à des facteurs environnementaux, des résultats de tests de laboratoire récents et toute autre donnée disponible. La création de ces profils requiert l'importation et l'intégration d'une très grande quantité de données à partir de plusieurs systèmes existants. Cette étape est

incontournable pour la recherche de médecine de précision et à cette étape du processus, le chercheur fait face à plusieurs défis majeurs :

- Le caractère diversifié et hétérogène des types des données ajouté à la quantité grandissante de ces données. En effet, l'intégration de toutes ces données hétérogènes et l'extraction des informations utiles à partir de ces données à des usages cliniques sont l'un des défis majeurs des chercheurs de ce domaine (Xue et coll. 2016);
- Les données proviennent de plusieurs systèmes. Ceci soulève deux problèmes principaux : premièrement, souvent les standards de représentations de données utilisées sont différents et deuxièmement les technologies utilisées par ces systèmes sont souvent incompatibles entre elles (Xue et coll. 2016);
- Les volumes de données sont importants et grandissants. L'exploitation de ces données nécessite des infrastructures informatiques très puissantes et coûteuses pour emmagasiner, gérer et analyser ces données (Xue et coll. 2016, Belghait et April 2018).

Ces dernières années ont vu l'émergence de nombreux logiciels de recherche commerciaux et libres qui offrent des solutions innovantes dans le but de recueillir et d'analyser un grand volume de données génétiques et cliniques pouvant être utilisées dans le cadre des recherches de médecine de précision. Elles nécessitent que les chercheurs puissent extraire les données des systèmes de Dossier Patient Informatisé (DPI), par exemple, qui contient les données cliniques des patients. Ces logiciels sont souvent limités au format des données de chaque DPI (Chloé Cabot, Lina F. Soualmia, 2015).

Canual et ses collègues (Canuel et coll. 2015) ont analysé sept caractéristiques des logiciels de médecine de précision : le contenu de l'information (c.-à-d. les données cliniques et génétiques), l'environnement pour l'accès sécurisé et la gestion de la confidentialité des informations, les supports d'analyse, les interfaces utilisateurs de visualisation, l'interopérabilité entre les systèmes, la configuration du système et les langages de programmation et les systèmes d'exploitation utilisés par ces logiciels. Ces caractéristiques vont être utiles pour évaluer les logiciels de préparation et d'analyse des données de médecine de précision actuellement disponibles pour les chercheurs du domaine.

2.2 Les logiciels de médecine de précision

Ces dernières années, de nombreux logiciels de recherche, autant commerciaux que libres, et qui visent à appuyer les chercheurs dans leurs activités de collecte, gestion et d'analyse des données cliniques et génétiques ont vu le jour. Il a été rapporté que ces logiciels offrent des interfaces utilisateurs et des cadres de programmations restreints aux formats de données supportés par des systèmes DPI spécifiques (Chloé Cabot, Lina F. Soualmia, 2015). La revue de littéraire concernant cette problématique a démontré que les auteurs ont une volonté de diminuer les dépendances technologiques entre ces logiciels et les systèmes DPI. La liste ci-dessous présente six exemples de logiciels qui adoptent cette tendance:

1. Le logiciel BRISK (Biology-Related Information Storage Kit) (Canuel et coll. 2015) développé par l'université de British Columbia au Canada et distribué en logiciel libre, est un assemblage d'applications Web qui donnent accès aux informations du phénotype et du génotype d'un patient. Il permet aux chercheurs d'analyser des études d'association GWAS (Genome-Wide Association Study) (Tan, Tripp, & Daley, 2011).
2. Le logiciel commercial iCOD (Integrated Clinical Omics Database) (Canuel et coll. 2015), développée par l'université de Tokyo de médecine dentaire au Japon, donne aux chercheurs la possibilité de recueillir et combiner des données sur les cas de carcinome hépatocellulaire (CHC) et de les visualiser sur une carte représentant l'interrelation entre les données. Les chercheurs peuvent ainsi analyser les corrélations possibles entre les données cliniques, telles que les phénotypes, et les informations moléculaires du patient (Tan, Tripp, et Daley 2011).
3. tranSMART est un logiciel libre développé par l'entreprise pharmaceutique américaine Johnson & Johnson aux États-Unis (Canuel et coll. 2015). Son objectif est d'analyser les données intégrées du patient dans le but de générer des hypothèses de recherche, les valider et faire des découvertes dans les cohortes, ce qui est incontournable en médecine de précision. Ce logiciel possède une architecture logicielle de « n tiers » et est programmé à l'aide de la technologie Grails (« Groovy on Rails ») un cadriciel de développement rapide (c.-à-d. Agile) pour développer des applications Web à l'aide du langage Java. Ce

logiciel de recherche gère les données à l'aide de bases de données relationnelles qui possèdent une structure de données inspirée du modèle des données libre i2b2 (Scheufele et coll. 2014).

4. OncDRS (Oncology Data Retrieval System) (Orechia et coll. 2015, Londin and Barash 2015) est un logiciel libre qui permet d'interroger et intégrer des données cliniques et génétiques provenant de sources hétérogènes. Il s'appuie sur le logiciel SPARKS (Système Synergique de Connaissance provenant du patient et de la recherche)(Canuel et coll. 2015).
5. deCODE Genetics (Hakonarson, Gulcher, & Stefansson, 2003), est un logiciel commercial qui permet l'accès aux Interfaces de Programmes d'Application (API) de manière sécurisée lors de la gestion des données des patients et de l'interaction avec d'autres systèmes sur internet. L'entreprise deCODE a mis au point une technologie qui va donner aux patients le contrôle de leurs données en leur permettant de décider comment elles sont partagées dans le domaine de la médecine de précision et s'appuie sur une méthodologie de cartographie du génome en utilisant une banque de données généalogique de la population Icelandaise créée par deCODE (Hakonarson, Gulcher, et Stefansson 2003);
6. IROmiCS (Information Retrieval for Omic and Clinical Sciences) (Chloé Cabot, Lina F. Soualmia, 2015), logiciel libre, qui repose sur le modèle des données cliniques RAVEL (Recherche d'Information et Visualisation dans le Dossier Patient Informatisé) (Lelong et coll. 2014). Il a pour objectif de représenter et rechercher des données génétiques dans une base de données relationnelle.

Le Tableau 2.1 présente une évaluation sommaire de ces logiciels commerciaux et libres disponibles par rapport à sept caractéristiques les plus référencées dans la littérature pour leur intérêt par les chercheurs du domaine.

À l'aide des articles qui ont été publiés concernant les logiciels commerciaux et les logiciels libres en médecine de précision, disponibles au moment de rédiger cette thèse, sept caractéristiques les plus référencées dans la littérature pour leur intérêt ont été identifiées. Ces caractéristiques ont été identifiées et analysées pour chacun. L'objectif de cette analyse est de

représenter s'ils possèdent ou non ces caractéristiques d'intérêt (O=Oui, N=Non) à partir des données publiées disponibles:

1. La première caractéristique d'intérêt vise les données génétiques : est-ce que le logiciel commercial ou le logiciel libre permet de traiter les données génétiques ?
2. La deuxième caractéristique d'intérêt s'interroge sur la capacité du logiciel commercial ou du logiciel libre à accepter toutes les données autres que les données génétiques et qui sont nécessaires à la recherche en médecine de précision ?
3. La troisième caractéristique d'intérêt concerne l'adaptabilité du modèle des données : est-ce que le logiciel commercial ou le logiciel libre est capable de satisfaire les besoins en données supplémentaires, qui n'existent pas déjà dans le logiciel commercial ou le logiciel libre et qui est nécessaire à toute nouvelle recherche, et cela sans un effort considérable afin que les chercheurs puissent élargir le champ de leurs recherches sans aucune limitation ?
4. La quatrième caractéristique d'intérêt porte sur l'élasticité automatique des bases de données: est-ce que le logiciel commercial ou le logiciel libre permet au chercheur de télécharger de très grands volumes de données, sans limitation, quel que soit le modèle des données ou l'infrastructure matérielle requise ?
5. La cinquième caractéristique d'intérêt s'interroge sur la capacité des logiciels à traiter de grands volumes de données : le logiciel permet-il aux chercheurs de faire des études sur de grandes quantités de données ?
6. La sixième caractéristique d'intérêt a trait au rééchelonnement de la capacité de traitement de l'infrastructure du logiciel commerciale ou du logiciel libre: est-il possible d'adapter le logiciel pour fournir aux chercheurs des moyens pour changer et ajuster la capacité de traitement de l'infrastructure du logiciel de recherche selon les besoins spécifiques des recherches ?

7. Cette dernière caractéristique d'intérêt vise la capacité du logiciel commercial ou du logiciel libre de permettre la reproductibilité d'une recherche : est-ce que le logiciel donne aux chercheurs la possibilité de reproduire une recherche antérieure (c.-à-d. du début à la fin) à tout moment, et ce sans trop d'efforts ?

Tableau 2.1 Évaluation des logiciels de recherche de médecine de précision (février 2019)

#	7 caractéristiques d'intérêt des logiciels de médecine de précision	Logiciels de recherche de médecine de précision					
		1	2	3	4	5	6
		BRISK (Tan et coll., 2011)	iCOD (Shimokawa et coll., 2010) (Alessandro, Meyer, & Klingmüller, 2013)	TranSMART (Scheufele et coll. 2014) (Murphy et coll. 2006)	OncDRS (Orechia et coll., 2015) (Londin & Barash, 2015)	décote (Hakonarson et coll., 2003)	IROmiCS (Chloé Cabot, Lina F. Soualmia, 2015; Soualmia, Darmoni, & Soualmia, 2015)
1	Permets l'exploration des données génétiques	O	O	O	O	O	O
2	Permets l'exploration des données autres que génétiques (clinique, historique médical, etc.)	O	O	O	O	O	O
3	Permets à son modèle des données d'être ajusté/bonifié avec de nouvelles données	N	N	N	N	N	N
4	Fournis une infrastructure flexible	N	N	N	N	N	N
5	Permets de faire une étude sur une grande quantité de données	O	O	O	O	O	O
6	Permets le rééchelonnement de la capacité de traitement de l'infrastructure informatique	N	N	N	N	N	N
7	Permets de reproduire une étude sans effort considérable	O	O	O	O	O	O

Lors de cette revue des caractéristiques de six des logiciels commerciaux et des logiciels libres les plus publiés concernant la médecine de précision (présentés au Tableau 2.1 ci-dessus), il a été constaté qu'aucun d'eux ne permet d'adapter le modèle des données afin de répondre aux besoins informationnels variables de leurs recherches spécifiques ni de leur permettre d'ajuster la capacité de traitement du logiciel (c.-à-d. la puissance de traitement) afin de s'adapter à la volumétrie variable et grandissante de leurs données. Tel qu'il a déjà été mentionné, les données utilisées dans une recherche de médecine de précision sont très variées et d'une grande volumétrie. Les défis technologiques que posent leurs captures, leurs structurations et leurs traitements créent des goulots d'étranglement et empêchent de nombreux chercheurs de réaliser eux-mêmes des recherches en médecine de précision. Ceci a pour effet de freiner l'utilisation des résultats potentiels de ces découvertes dans les soins cliniques (Wolkenhauer et coll. 2014).

Notez que, dans cette thèse, les problèmes liés à la sécurité des données ne seront pas abordés. Les questions de la sécurité d'accès et de la gestion de la confidentialité des données sont des problèmes majeurs qui touchent non seulement la médecine de précision, mais tout le domaine des analyses sur les données du domaine de la santé et cette question mérite qu'on lui consacre, à elle seule, un autre projet de recherche.

2.3 Défis technologiques concernant les données

La grande diversité et la volumétrie des données requises dans les recherches de médecine de précision nécessitent l'utilisation et la maîtrise d'un grand nombre de technologies du domaine de l'informatique. Les défis reliés à l'obtention, la préparation et le traitement de ces données représentent un réel problème pour les chercheurs du domaine et particulièrement pour les petits centres de recherche situés dans les hôpitaux universitaires. Trois défis technologiques majeurs, auxquels ces chercheurs doivent faire face, en ce qui concerne la préparation et le traitement des données ont été identifiés : le premier est de pouvoir traiter facilement un très grand volume de données, le deuxième est d'être capable de gérer leur

complexité/hétérogénéité et le troisième qui de permettre de facilement intégrer ces données dans une seule base de données afin d'en faciliter le traitement, la gestion et l'analyse subséquente.

2.3.1 Facilité de traitement d'un grand volume de données

Les avancées récentes, dans le domaine de la technologie de séquençage de nouvelle génération (NGS), ont entraîné la production d'une très grande quantité de données génétiques, qui sont maintenant au cœur des recherches en médecine de précision (Gullapalli et coll. 2012). Il y a d'abord eu le projet de séquençage du génome humain (F S Collins et Mansoura 2001), qui a permis la création de nombreuses bases de données accessibles publiquement. D'ailleurs, celui-ci continue d'augmenter la quantité d'information disponible année après année (Galperin, 2004). Au cours des huit dernières années, la taille du Sequence Read Archive (SRA), la base de données du NIH (National Institutes of Health ou Institut américain de la Santé), a augmenté de façon exponentielle. Nous savons maintenant qu'il y a plus de trois milliards de paires de bases sur un génome humain. Le séquençage d'un génome complet génère donc plus de 100 gigaoctets de données. Ces données sont actuellement stockées dans des formats de fichier génétique d'un format spécialisé, par exemple le format Binary Aligmnent Map (BAM) qui est une représentation binaire du format d'alignement de séquence (SAM), ou dans le format de fichier Variant Call Format (VCF).

En plus de ces grandes quantités de données génétiques, la recherche en médecine de précision requiert l'utilisation des données du patient (c.-à-d. les données cliniques). La quantité des données cliniques a aussi augmenté de façon considérable au cours des six dernières années, notamment avec l'adoption massive des systèmes informatiques de gestion des dossiers patients informatisés (DPI) (Jamoon, Eric; Ninee, Yang; Hing, 2016). Les données cliniques extraites des DPI, peuvent s'avérer extrêmement volumineuses. Par exemple, dans un article de recherche récent, on a estimé que les données de 20 000 patients inscrits dans ce projet de

recherche en médecine personnalisée représentaient à peu près 3,3 gigaoctets de données (Karen He, Dongliang Ge, 2017).

2.3.2 Complexité / hétérogénéité des données

Il a été aussi discuté précédemment que les logiciels de médecine de précision doivent composer avec de nombreuses sources de données hétérogènes (Ideker et coll. 2001). Hormis les données génétiques, les types des données cliniques extraites des DPI varient grandement; ainsi, pour chaque patient, ces données peuvent inclure, selon la Classification Internationale des Maladies (CIM) : des codes; des médicaments; des traitements; des codes de procédure (de l'anglais Current Procedural Terminology (CPT)); des résultats d'analyses de laboratoire; des notes de cliniciens; ainsi que: des régimes alimentaires; et des activités physiques. Tel qu'il sera discuté plus tard dans cette thèse, des données considérées comme des données démographiques de patients jouent de plus en plus un rôle important lors de recherches en médecine de précision.

2.3.3 Intégration des données

La médecine de précision pose aussi un autre défi important au niveau des données : la nécessité de les intégrer dans une seule et unique base de données (Jure et coll. 2014, Wolkenhauer et coll. 2014, Dubitzky, Krebs, et Eils 2001). Cette activité est nécessaire, car elle permet et facilite la conception de modèles d'analyse prédictifs, adaptés à chaque patient, dans lesquels divers types de données telles que les données : cliniques, génétiques, moléculaires, pathologiques, physiologiques, d'imagerie médicale et démographique qui décrivent l'environnement et le mode de vie du patient pourront faire l'objet d'analyse simultanément (Duffy, 2018). D'autres problématiques de cohabitation de données démographiques, cliniques et génétiques ont été rapportées par des chercheurs. Elles incluent le manque de normalisation des formats de données, causée par les nombreux fournisseurs

d'équipements de NGS qui compétitionnent entre eux et utilisent des formats propriétaires. Ceci crée une grande diversité de formats qui doivent être utilisables par les logiciels du domaine, ce qui se répercute aussi dans l'utilisation d'une grande variété de techniques de stockage de données (Baro et coll. 2015).

2.3.4 Gestion de l'adaptation des modèles des données aux nouveaux besoins informationnels

Nous avons déjà parlé de la nécessité d'intégrer toutes les données requises par ces recherches dans une seule base de données. Nous avons aussi dit que l'évolution constante des besoins informationnels de ces recherches pose un défi d'adaptation des modèles des données de ces bases de données. Cette problématique de gestion de l'adaptation des modèles des données n'est pas récente. Il faut comprendre qu'une base de données possède un modèle des données (aussi nommé un schéma de données) qui décrit les relations entre chaque donnée. Il a été observé que chaque recherche du domaine de la médecine de précision nécessite des données de toutes sortes et qui changent assez souvent. À chacun de ces changements, c'est-à-dire : l'ajout, la modification ou le retrait d'une donnée, le chercheur doit retravailler le modèle de sa base de données et cela crée des complications assez rapidement. Conséquemment, plusieurs auteurs ont publié sur cette question.

Castelli (Castelli, 1998) propose une stratégie basée sur la réutilisation des modèles des données existants pour gérer l'adaptation d'un modèle des données à de nouveaux besoins informationnels. Sa stratégie consiste à toujours maintenir une documentation à jour des dernières adaptations effectuées sur le modèle des données. Selon l'auteur, cette stratégie simplifie le processus d'adaptation des modèles des données, mais ne résout pas la problématique de la gestion des adaptations répétées de ces modèles, surtout dans un contexte où les adaptations des modèles des données dans le domaine des recherches en médecine de précision sont nombreuses et nécessitent des délais très courts. Kupfer et ses collègues (Kupfer, Eckstein, Neumann, & Mathiak, 2006) soulèvent aussi la problématique de la diversité des bases de données des recherches, dans le domaine de la santé, et les difficultés rencontrées par les chercheurs de ce domaine concernant l'adaptation des modèles des données aux nouveaux

besoins informationnels requis pour les recherches. Ils proposent une approche basée sur la génération automatique des ontologies de bases des données et le mappage de ces ontologies avec les modèles des données. Selon ces auteurs, cette approche garantit la synchronisation des ontologies et des modèles des données, mais ne résout pas le problème de la gestion des adaptations successives effectuées sur les modèles de données. Une troisième approche publiée est celle de la gestion des versions des modèles (Andany, Lconard, et Palisser 1991, Sven-Eric Lautemann 1997). L'objectif de cette approche est que plusieurs applications utilisent des versions différentes d'un même modèle des données. Les auteurs précisent cependant que la mise en œuvre de cette approche pose encore problème lors de la gestion des données à travers les versions de différents modèles.

Ces propositions ne permettent pas de solutionner l'objectif de cette recherche, qui vise à ce que la même application (c.-à-d. l'application d'intégration et d'analyse des données) puisse fonctionner avec des versions différentes d'un même modèle des données : c'est-à-dire des modèles des données obtenues à partir des mises à jour continues des versions antérieures d'un même modèle des données. Il faudra donc innover et tenter de trouver une solution à cette problématique importante du domaine de la recherche en médecine de précision. La prochaine section présente un processus typique de recherche de médecine de précision.

2.3.5 Approche itérative dans le processus de recherche

Les recherches du domaine de la médecine de précision sont basées essentiellement sur l'évaluation d'hypothèses en effectuant l'exploration de données médicales. Les chercheurs réévaluent la pertinence de leurs hypothèses en fonction des résultats obtenus à partir de l'exploration de ces données. On pourrait penser que ce type de recherche utilise un processus qui est un hybride entre la recherche qualitative et la recherche quantitative.

Les approches d'analyses itératives de données cadrent bien avec la nature des recherches en médecine de précision. C'est-à-dire qu'une itération d'analyse de données dans ce type de recherche n'est pas une répétition mécanique de la tâche d'analyse, mais plutôt un processus

de réflexion continue sur les connaissances extraites à partir des données analysées au cycle précédent, permettant l'affinement progressif des hypothèses et des questions de recherche (Srivastava et Hopwood 2009). Ces auteurs proposent un processus itératif pour les analyses de données qualitatives basées sur trois questions. La première est que me disent les données? la deuxième est qu'est-ce que je veux savoir? et enfin quelle est la relation dialectique entre ce que les données me disent et ce que je veux savoir? (Drongelen, 2001) adopte aussi un processus itératif et propose que chaque itération du processus de recherche puisse s'adapter en fonction des principes suivants : les questions de recherches peuvent changer d'une itération à une autre en fonction : des données utilisées; de la capacité du matériel informatique disponible/utilisé à traiter ces données; des méthodes d'analyse utilisées; et des résultats intermédiaires obtenus lors d'une itération de recherche.

2.3.6 Choisir entre : utiliser un logiciel commercial ou un logiciel libre existant; ou créer son propre logiciel de recherche en médecine de précision

Quelques choix se présentent, aux chercheurs de ce domaine, à propos de solution logiciel qu'ils peuvent utiliser dans leurs recherches: utiliser un logiciel existant (c.-à-d. commercial ou logiciel libre) ou construire leur propre logiciel de recherche en médecine de précision.

Dans cette section, des solutions envisageables sont abordées. Si un chercheur en médecine de précision n'utilise pas l'un des logiciels commerciaux ou l'un des logiciels libres présentés au Tableau 2.1, quelles seraient les solutions alternatives ?

Advenant qu'un chercheur décide de concevoir sa propre solution, une des premières pistes de solution possible est d'utiliser les infrastructures infonuagiques, Google Cloud, AWS ou Microsoft Azure, pour ne nommer que celles-ci, pour créer son propre logiciel de recherche en médecine de précision. Cette option est une possibilité, étant donné que les fournisseurs de services infonuagiques mettent à la disposition de tous, des services/applications sous forme de services Web qu'ils peuvent utiliser pour développer des logiciels adaptés à leurs besoins. Cette approche d'intégration d'application sous forme de service Web exige un haut niveau de compétences en informatique et bio-informatique dans l'équipe du chercheur en médecine de précision. Les risques associés à cette approche sont de ne pas avoir le bon savoir-faire à

l'interne; de ne pas utiliser les bonnes technologies et logiciels au départ; et de se retrouver potentiellement avec un logiciel mal conçu, ce qui génère des coûts d'adaptation élevés lors de la préparation de chaque étude. Pour mitiger certains de ces risques, il est souvent conseillé de contracter un cabinet de conseil spécialisé, tel que Databricks (Databricks, 2019) ou Informatica qui pourra guider ces choix technologiques et potentiellement concevoir une partie de ce logiciel.

Une autre option consiste à identifier et sélectionner un service commercial (c.-à-d. une solution existante) de médecine de précision déjà disponible dans le nuage. Ces services en ligne, rendent disponibles la plupart des fonctionnalités nécessaires, déjà éprouvées, et sont accompagnés d'un service de support et d'aide à l'utilisation. Ces services permettent de charger des données en toute sécurité, de construire un pipeline de séquences d'activités pour l'étude à effectuer, et ce à partir des applications existantes. Finalement, il est possible d'ajuster la capacité de traitement de l'infrastructure des serveurs virtuels et des disques si nécessaire. Certains de ces services peuvent aussi permettre de reproduire facilement des études déjà faites. Des fournisseurs de services tel que DNANexus (DNANexus, 2019) et DRAGEN (Illumina, 2019) font de la publicité concernant ces services, mais peu d'analyses indépendantes et d'informations publiques sont disponibles concernant la qualité, la satisfaction et la tarification. De nouveaux services similaires ont aussi été annoncés, par exemple, celui proposé par Deloitte et Vineti (Deloitte, 2019), qui ont prévu de retravailler le logiciel de thérapie génique développé en collaboration avec l'entreprise Général Electric (GE) et la Mayo Clinic. Finalement, Général Électric et l'entreprise Roche se sont associés et ont annoncé qu'ils allaient offrir une solution à l'avenir.

En somme, l'utilisation d'un service infonuagique nécessite que le chercheur en médecine de précision puisse effectuer un audit (ou essai) préalable du service afin d'en sélectionner un parmi ceux qui sont disponibles qui convient à ses besoins. Il doit ensuite avoir la capacité financière de payer pour ce service. Étant donné que ces données ne sont pas disponibles publiquement au moment d'écrire cette thèse, il faut supposer qu'il faudra allouer un assez important budget pour ce genre de service.

Dans le cas où le chercheur n'a pas un grand budget et qu'il désire exploiter l'expertise interne de son groupe de recherche (c.-à-d. l'expertise de bio-informaticiens) pour créer son propre logiciel de recherche en médecine de précision, les solutions existantes présentées au Tableau 2.1 peuvent être un bon point de départ. De plus, il peut examiner l'utilisation de technologies émergentes du domaine du logiciel libre, car leur utilisation ne nécessite pas de coûts de licence additionnelle. Comme présenté au Tableau 2.1, tous les logiciels commerciaux et les logiciels libres présentés supportent les types de données requises par ce chercheur (c.-à-d. autant les données génétiques que les données cliniques). Cependant, chaque logiciel du Tableau 2.1 a ses particularités et ses limites propres au niveau du type et du format de données qu'il supporte. De plus, il a été identifié qu'aucun ne supporte la fonction de l'adaptabilité de modèle des données aux nouveaux besoins informationnels. En ce qui concerne l'échelonnabilité de l'infrastructure, encore une fois, aucun des logiciels étudiés ne discute de cette caractéristique dans ses publications. Lorsque le logiciel n'est pas déployé sur une infrastructure infonuagique, acquérir, installer et gérer l'infrastructure matérielle risque d'être très coûteux pour le chercheur. D'autre part, les solutions présentées au Tableau 2.1 sont étroitement liées à des technologies spécifiques qu'elles utilisent, par exemple I2B2/TranSMART, qui est offert par la faculté de médecine de Harvard, utilise le modèle des données i2b2 et la technologie de présentation des résultats Redshift (Jure et coll. 2014, Dubitzky, Krebs, et Eils 2001). Dans les articles consultés, il n'y avait aucun avis sur l'expérience utilisateur concernant cette combinaison de technologies.

Cette section a présenté les défis reliés à la conception d'un logiciel de recherche en médecine de précision à l'aide de logiciels libres. La section suivante décrit les directions/tendances futures des logiciels de recherche du domaine de médecine de précision.

2.3.7 Direction future des logiciels de recherche de médecine de précision

Dans la section précédente, aucun logiciel de médecine de précision (c.-à-d. présentés au Tableau 2.1) n'a obtenu un score complet (c.-à-d. une évaluation positive (un « oui ») pour tous les critères). Il faut donc se demander comment ces logiciels vont évoluer dans le futur. Selon la tendance actuelle, afin de mieux répondre aux défis des chercheurs du domaine, les logiciels ou les plateformes logicielles d'analyse en médecine de précision du futur, devront utiliser les technologies du domaine des données massives (c.-à-d. du Big Data) et de l'infonuagique.

2.3.7.1 Solutions basées sur les technologies des données massives (Big Data)

Concernant les besoins des bio-informaticiens et des chercheurs en médecine de précision, les technologies émergentes du domaine des données massives offrent de nouvelles possibilités en rendant disponible gratuitement un écosystème de technologies libres qui ont le potentiel de transformer significativement la manière d'effectuer leurs recherches.

Le terme Big Data (en français données massives) est couramment associé avec l'utilisation de technologies de bases de données NoSQL, appelées aussi « Not only SQL » (c.-à-d. pas seulement SQL) ainsi qu'à l'utilisation de moteurs d'analyse pour le traitement distribué sur des grappes de serveurs virtuels. Ces technologies récentes sont capables de traiter des quantités massives de données à petits frais, contrairement aux technologies conventionnelles telles que les bases de données relationnelles, qui s'avèrent moins efficaces dans le contexte des données massives menant à des analyses plus restreintes (c.-à-d. du cas par cas). Le domaine des données massives a été défini de différentes manières selon les auteurs (Baro et coll. 2015). Une des définitions les plus répandues des données massives est fournie par les technologies, qui prennent en compte les 5V : Volume, Vitesse, Variété, Véracité et Valeur (Dave et Kamal 2017). Parmi les technologies des données massives émergentes, il y a

Hadoop, Spark et les bases de données NoSQL telles que HBase et Impala. L'analyse à l'aide de technologies de données massives permet aussi de recueillir, manipuler et analyser une quantité massive de données de type très variées, incluant des données génétiques et des données provenant des systèmes DPI ce qui permettrait de révéler des tendances cachées et des corrélations, mais aussi de permettre des analyses menant à des conclusions nouvelles grâce à ces technologies nouvelles. Finalement, de par sa popularité grandissante, l'analyse à l'aide de technologies des données massives est de plus en plus utilisée dans différents champs de recherche similaires (Chute et coll. 2014).

2.3.7.2 Solutions basées sur les technologies infonuagiques

Selon l'Institut National des Normes et de la Technologie (de l'anglais : National Institute of Standards and Technology (NIST)) l'infonuagique est un « modèle de fonctionnement qui permet un accès réseau pratique et sur demande à un ensemble partagé de ressources informatiques configurables (c.-à-d. des réseaux, des serveurs virtuels, du stockage, des applications et des services), qui peut être rapidement approvisionné et disponible sans trop d'efforts de gestion et avec une interaction limitée avec les opérateurs.

L'utilisation de l'infonuagique favorise la disponibilité et est composée de cinq caractéristiques essentielles : « On-demand self-service, Broad network access, resource pooling, Rapid elasticity, Measured service », de trois modèles de services : « Software as a Service (SaaS), Platform as a Service (PaaS), Infrastructure as a Service (IaaS) » et de quatre modèles de déploiement : « Private cloud, Community cloud. Public infonuagique, Hybrid infonuagique » (Peter Mell et Tim Grance 2011).

Afin de permettre l'échelonnabilité automatique de la capacité de traitement, ce qui est incontournable pour les logiciels de médecine de précision, les plateformes commerciales infonuagiques publiques telles que AWS-EC2, Google Cloud et Microsoft Azure offrent des services permettant le traitement massivement parallèle (MPP - Massively Parallel Processing). L'utilisation de ces services est une alternative intéressante, car seulement une minorité de laboratoires de recherche peuvent se permettre l'achat, l'installation et la gestion

de leurs propres grappes de serveurs virtuels, car cela est très coûteux et requiert beaucoup d'expertise en informatique. Les fournisseurs des services infonuagiques proposent aussi à leurs clients des infrastructures matérielles prêtes à utiliser, des briques logicielles et des services de programmation prêts pour modification qui peuvent réduire les coûts d'acquisition, de développement et de maintenance logicielle (O'Driscoll, Daugelaite, et Sleator 2013). Étant donné que les clients de l'infonuagique ne paient que pour les services qu'ils utilisent, cette approche de solution devient une avenue économique pour la gestion et le traitement de quantité massive de données génétiques et cliniques. En comparaison la création et la maintenance des bases de données locales, par les chercheurs œuvrant dans des petits centres de recherche hospitaliers, nécessitent de grands investissements en matériel pour traiter et stocker ces quantités massives de données. Ces budgets seraient mieux investis pour effectuer de la recherche plutôt que : d'acheter du matériel informatique, qui doit être typiquement remplacé tous les cinq à sept ans; et défrayer les salaires de nombreux experts en informatique.

Hadoop et Sparks (The Apache Software Foundation, 2018)(Foundation Apache Software, 2019) sont des moteurs d'analyse pour le traitement distribué de données massives. Ils sont disponibles chez tous les fournisseurs publics de services infonuagiques. Ces technologies sont très bien adaptées et déjà utilisées pour développer des logiciels destinés à traiter de quantités massives de données génétiques et cliniques. De nombreux logiciels de traitement de données génétiques ont été développés à l'aide de Hadoop, par exemple : Crossbow est un pipeline de logiciel évolutif pour toute l'analyse du reséquençage du génome (O'Driscoll, Daugelaite, et Sleator 2013); GATK (Genome Analysis Kit) est un kit d'analyse du génome qui se concentre sur le génotypage et la détection de variants (McKenna et coll. 2010); Hadoop-BAM (Niemenmaa et coll. 2012) est une librairie JAVA pour la manipulation des données de NGS sur Hadoop. Elle agit comme une couche d'intégration (c.-à-d. un intermédiaire) entre les applications d'analyse et la version binaire des fichiers avec le format SAM (de l'anglais Sequence Alignement/Map (SAM)) qui sont traitées à l'aide de la technologie Hadoop (Niemenmaa et coll. 2012).

L'une des nombreuses initiatives permettant d'améliorer la performance du traitement des données génétiques est le projet ADAM de l'Université Berkeley. ADAM regroupe un ensemble de formats, d'API et sert à exécuter le traitement des données génétiques par étape (Massie et coll. 2013). ADAM propose un pipeline échelonnable pour traiter les données génétiques dans un cadre informatique distribué très performant. Il utilise la technologie libre des données massives Spark (Zaharia et coll. 2016) pour son moteur d'analyse très efficace de traitement distribué de données massives et l'utilisation de la technologie Parquet pour un stockage très efficace et rapide d'accès. Parquet est un format de compression et de stockage de données orienté colonne permettant d'accéder à des quantités massives de données très rapidement. (Massie et coll. 2013).

Les recherches du domaine de la médecine de précision, des données massives et du traitement distribué de données massives ont évolué de manière relativement isolée ces dernières années (Haas et coll. 2001, Chung et Wong 1999, Etzold, Ulyanov, et Argos 1996). Une revue de littérature des articles publiés sur ce sujet a permis d'identifier les opportunités d'améliorations suivantes pour les logiciels de médecine de précision du futur :

- Une plus grande synergie entre les différents domaines de recherche pourrait contribuer à résoudre les défis dans le traitement des données de médecine de précision;
- Ces différents domaines de recherche font face à des obstacles similaires en ce qui concerne l'utilisation de la technologie des données massives. Parmi les obstacles les plus significatifs, il y a l'intégration des bases de données variées, le manque de normes pour traiter les données de NGS, les processus bio-informatiques (c.-à-d. les pipelines) et les systèmes de soutien à grande échelle des décisions pour les données cliniques, empêchant la création d'un modèle des données global qui peut satisfaire les besoins des chercheurs dans le domaine de la médecine de précision;
- Les logiciels d'analyse proposés sont surtout limités à des domaines de recherches spécifiques de la médecine de précision et sont statiques, c.-à-d. ils ne permettent pas facilement l'inclusion de nouvelles dimensions de données; ils sont difficilement

adaptables aux besoins spécifiques des chercheurs dans la recherche en médecine de précision;

- Aucune structure complète ne permettant aux chercheurs d'installer, de configurer leur propre environnement et d'adapter le modèle des données pour soutenir leurs recherches spécifiques n'est disponible.

Par conséquent, compte tenu des défis mentionnés ci-dessus, il est possible de constater qu'une avenue prometteuse de recherche serait d'essayer d'identifier une nouvelle approche d'adaptation de modèles et d'intégration des données pour profiter de toutes ces avancées technologiques. Il serait intéressant aussi de permettre au chercheur d'opérer lui-même ses logiciels à l'aide d'interfaces utilisateurs conviviales. Il a été observé qu'actuellement, la majorité des logiciels répertoriés nécessitent l'utilisation d'instructions (c.-à-d. de commandes de programmation) en ligne de commande qui sont assez complexes et qui ont pour but de configurer les paramètres d'analyse et lancer leur exécution. De nombreux chercheurs ne disposent pas du niveau d'expertise technique en informatique nécessaire pour utiliser ce genre d'interface très technique et ainsi, cette approche est un frein pour faire progresser leurs travaux de recherche et obtenir des découvertes rapidement. En plus de ne pas faciliter l'expérience utilisateur, les logiciels actuels offrent une fonctionnalité complexe qui rend difficile la possibilité de reproduire une analyse, ce qui est une fonctionnalité incontournable dans ce domaine (Sandve et coll. 2013).

En conclusion, les chercheurs de médecine de précision s'appuient sur deux éléments principaux :

1. Des logiciels de médecine de précision qui doivent composer avec des quantités de données qui augmentent de manière exponentielle (Galperin, 2004), qui proviennent de nombreuses sources des données hétérogènes (Ideker et coll. 2001) et qui doivent être intégré dans la même base de données (Jure et coll. 2014, Wolkenhauer et coll. 2014, Dubitzky, Krebs, et Eils 2001). Une évaluation de six logiciels les plus référencés dans la littérature pour leur intérêt par les chercheurs du domaine a été effectuée (Tableau 2.1) par rapport à sept caractéristiques et aucun de ces logiciels n'a été évalué positif

pour tous les critères utilisés dans l'évaluation. De plus, les chercheurs doivent composer avec le dilemme entre utiliser un logiciel commercial ou un logiciel libre existant; ou créer son propre logiciel de recherche en médecine de précision;

2. Un processus de recherche itératif qui nécessite des adaptations récurrentes au modèle des données. Cette problématique n'est pas récente.

Dans la section suivante, la méthode de mesure de logiciel à utiliser dans cette thèse sera présentée.

2.4 La méthode de mesure en génie logiciel

Avant de passer à la proposition/conception de la nouvelle approche, il est important d'établir les bases théoriques de la mesure du succès ou de l'échec de l'expérimentation du modèle proposé qui aura lieu lors de l'étude de cas. D'après (Jacquet and Abran 1997), la mesure, dans le domaine du génie logiciel, doit suivre une méthode de mesure constituée de trois étapes (voir la Figure 2.1) (Jacquet and Abran 1997):

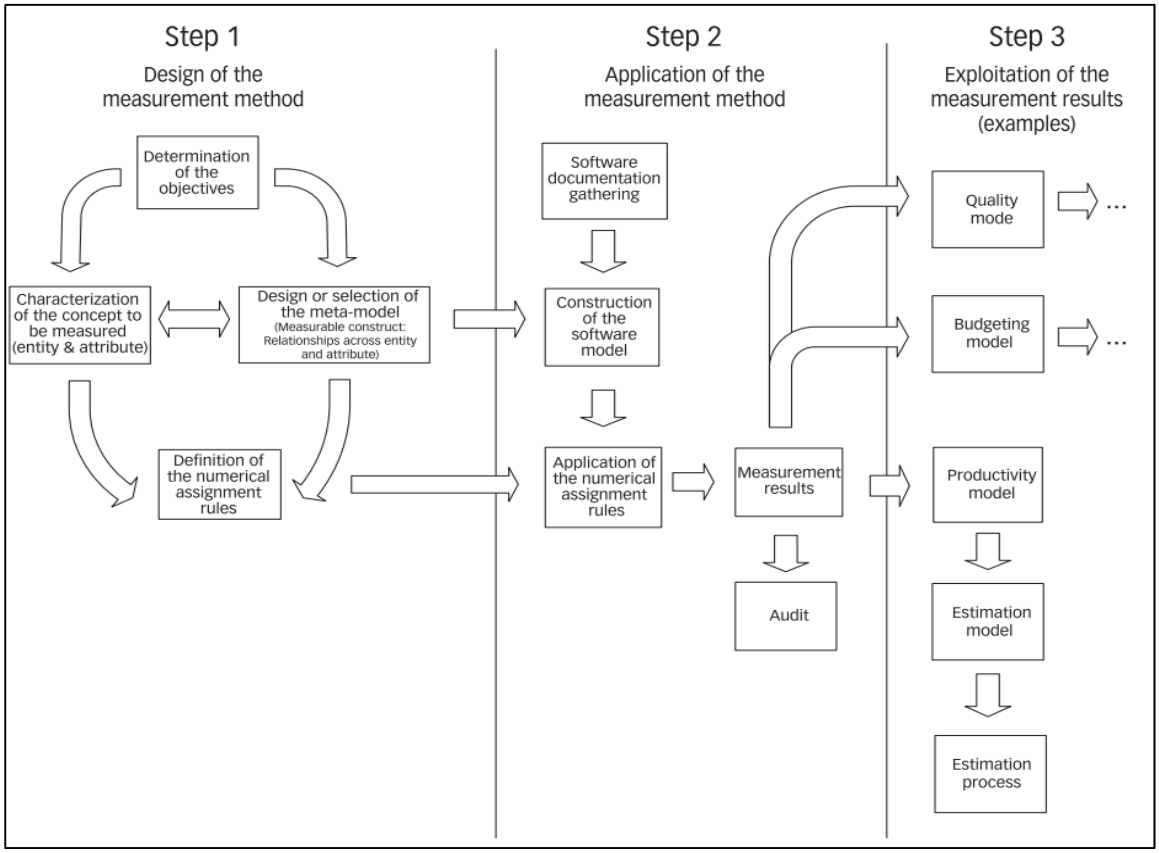


Figure 2.1 Modèle contextuel de mesure adapté de (Jacquet & Abran, 1997)

Le modèle conceptuel proposé comporte trois étapes afin de s'assurer de bien choisir des mesures qui correspondent aux objectifs de la recherche. Voici les étapes qui seront suivies pour établir les mesures qui seront collectées lors de l'étude de cas de cette recherche :

- **Étape 1 :** concevoir la méthode de mesure. Les intrants de cette étape sont :
  - La description de l'objectif de la mesure pour chaque objectif;
  - La description de l'ensemble des concepts et techniques requis pour concevoir la mesure.

  Les extrants de cette étape sont les concepts à mesurer pour chaque objectif de la méthode de mesure;
- **Étape 2 :** appliquer la méthode de mesure. Les intrants de cette étape sont :
  - Les entités à mesurer;

- La méthode de mesure conçue à l'étape précédente.

Les extrants de cette étape sont:

- Les résultats de la mesure pour chaque objectif;
- Le degré de précision de ces résultats de mesure.

- **Étape 3 :** l'utilisation des résultats de la mesure.

Pour cette recherche, les étapes 1 et 2 seront précisées, pour chaque objectif/mesure lors de la préparation de l'étude de cas. Cette description est présentée dans le prochain chapitre à la section 4.5. Ce niveau de détail est nécessaire, car le prototype expérimental, qui doit être conçu, doit être en mesure de collecter ces mesures. De plus, des questionnaires seront développés et envoyés, aux chercheurs du laboratoire du Dr Hamet, afin de collecter ces mesures, dites mesures de références, concernant le processus actuel.

2.5 Résumé

Ce chapitre conclut la deuxième étape de la recherche, celle de la revue de littérature. Il a été présenté la description des recherches du domaine de la médecine de précision, les logiciels disponibles et utilisés par certains chercheurs de ce domaine, les défis technologiques concernant le traitement de très grands volumes de données, la gestion de la complexité et de l'hétérogénéité de ces données, la nécessité de l'intégration de ces données qui proviennent de plusieurs sources et sous différents formats, les différentes options dont disposent présentement les chercheurs pour réaliser ces recherches et les directions futures des logiciels de recherche de médecine de précision.

Ce chapitre a permis d'identifier les défis auxquels font face les chercheurs en médecine de précision à chaque fois qu'ils entreprennent une nouvelle analyse et les différentes pistes de solutions dont ils disposent actuellement. Les tableaux suivants présentent un résumé des défis et de ces pistes de solutions futures :

a. Résumé des défis

Le Tableau 2.2 suivant présente un sommaire de tous les défis identifiés dans la revue de littérature auxquels font face les chercheurs actuellement et les conséquences de chaque défi sur les chercheurs du domaine de médecine de précision.

Tableau 2.2 Liste des défis des recherches de médecine de précision

Défis	Conséquences/Impact	Section de la thèse
1. Données de recherche diverses et hétérogènes (Xue et coll. 2016)	• Problèmes d'intégration des données.	2.1
2. Volume des données croissant à un rythme exponentiel (Xue et coll. 2016, Belghait et April 2018)	• Capacité de traitements et de stockage des données insuffisantes. • L'intégration et le traitement des données prennent beaucoup de temps parfois impossible avec les moyens informatiques disponibles.	
3. Plusieurs sources des données (Xue et coll. 2016, Ideker et coll. 2001).	• Standards de représentation des données différents; • Technologies différentes.	
4. Modèles des données des logiciels d'analyse de données actuels sont statiques	• Incapacité de supporter la variabilité des besoins informationnels des recherches.	2.2
5. Infrastructure des logiciels de recherche sont ou bien statique ou bien ne permettent pas une auto-extension	• Inaptitude d'adapter la capacité de traitement des logiciels de recherche aux volumes des données de recherches croissantes.	
6. Logiciels de recherche très liés aux technologies de l'infonuagique utilisées	• Dépendances envers les fournisseurs de l'infonuagique.	

b. Résumé des pistes de solutions existantes

Le Tableau 2.3 suivant présente un sommaire des pistes de solution proposée par les recherches antérieure et identifiée dans la revue de littérature. Pour chaque piste, les avantages et les inconvénients de la solution proposée sont présentés dans le tableau.

Tableau 2.3 Liste des pistes des solutions existantes

Piste de solution	Avantage/Inconvénient	Section de la thèse
1. Utiliser un service infonuagique existant de médecine de précision comme DNANexus et DRAGEN	**a. Avantages :** • Solutions éprouvées avec le type de recherches dans lesquels elles sont spécialisées; • Expertise technique fournie par le fournisseur de service. **b. Inconvénients** • Dépendances envers les fournisseurs de service de logiciels de recherche; • Coût en général très élevé; • Logiciels spécialisés, pas adaptés pour faire des recherches complexes et qui ont des besoins informationnels variés.	2.3.6

Piste de solution	Avantage/Inconvénient	Section de la thèse
2. Utiliser les services des logiciels de recherches existants comme BRISK, iCOD, I2B2, etc.	**a. Avantages** • Solutions éprouvées avec le type de recherches dans lesquels elles sont spécialisées; • Expertise technique fournie par le fournisseur de service. **b. Inconvénients** • Logiciels spécialisés ne supportent pas la fonction de l'adaptabilité de leur modèle des données aux nouveaux besoins informationnels; • Infrastructure statique, ne permet pas l'échelonnabilité de la capacité de traitement en fonction du volume de données croissant.	2.2

De plus, le modèle théorique de mesure, du domaine du génie logiciel, qui sera précisé/conçu plus en détail avant l'expérimentation, a été présenté. Ce modèle théorique de mesure vise à clarifier l'objectif des mesures, les concepts mesurés et la méthode de mesure qui serviront pour collecter chaque mesure associée aux objectifs de la recherche qui ont été définis à la section 0.2. Ainsi ces mesures devront être collectées avant et après l'étude de cas de manière à statuer sur le succès/échec du modèle original proposé dans cette thèse.

Le chapitre suivant présente la proposition de la nouvelle approche d'adaptation de modèle et d'intégration de données de cette thèse. Dans un premier temps, la définition de la nouvelle approche est présentée, suivie des étapes proposées et la conception d'un prototype expérimental permettant d'expérimenter sa mise en œuvre dans une étude de cas.

CHAPITRE 3

PROPOSITION D'UNE NOUVELLE APPROCHE D'ADAPTATION DE MODÈLE ET D'INTÉGRATION DES DONNÉES REQUISES POUR LES RECHERCHES DU DOMAINE DE MÉDECINE DE PRÉCISION

Ce chapitre vise principalement à présenter les décisions de conception de l'approche d'adaptation de modèle et d'intégration de données. Cette proposition vise à répondre aux différents défis identifiés au début de ce projet de recherche. Il est structuré en deux sections principales :

- La section 3.1 introduit la définition de la nouvelle approche d'adaptation de modèle et d'intégration des données. Elle qui est composée de quatre éléments principaux. La description et les assises sur lesquelles se base chacun de ces éléments seront décrites dans les sous-sections suivantes :
  - 3.1.1 Approche dynamique d'adaptation de modèle des données : cette section décrit en détail la proposition d'une nouvelle approche d'adaptation de modèle des données;
  - 3.1.2 Intégration continue des données : cette section décrit comment cette nouvelle approche assure l'intégration continue des données dans une seule base de données en utilisant progressivement différentes versions du modèle des données;
  - 3.1.3 Approche d'analyse itérative : cette section présente comment cette approche peut être utilisée dans le contexte de la proposition d'une nouvelle approche d'adaptation de modèle et d'intégration de données en médecine de précision;
  - 3.1.4 Cycle de recherche avec la nouvelle solution proposée : cette section définit comment les éléments de la nouvelle approche s'intègrent ensemble dans un cycle de recherche de médecine de précision;
- La section 3.2 présente les décisions de conception d'un prototype expérimental qui permet de mettre en œuvre la nouvelle approche d'analyse dans une étude de cas. Elle décrit la vue d'ensemble de l'architecture logicielle du prototype expérimental ainsi

que la description est les assises de chaque composante logicielle pour la nouvelle approche d'adaptation de modèle et d'intégration des données proposée.

3.1 Proposition de la nouvelle approche d'analyse de données

La nouvelle approche d'adaptation de modèle et d'intégration des données proposée est basée sur quatre éléments principaux :

1) une méthodologie pour la gestion des adaptations successives du modèle des données utilisé pour la réalisation des recherches de médecine de précision;
2) une approche, une mécanique et des outils pour l'intégration des données dans un modèle des données qui doivent être adaptés continuellement aux besoins informationnels des chercheurs;
3) une approche itérative d'exécution du cycle de recherche rendu possible grâce aux deux composants précédents.
4) un nouveau cycle de recherche qui intègre les trois composants précédents. La Figure 3.1 suivante illustre ces quatre composants et les relations de chevauchement entre eux.

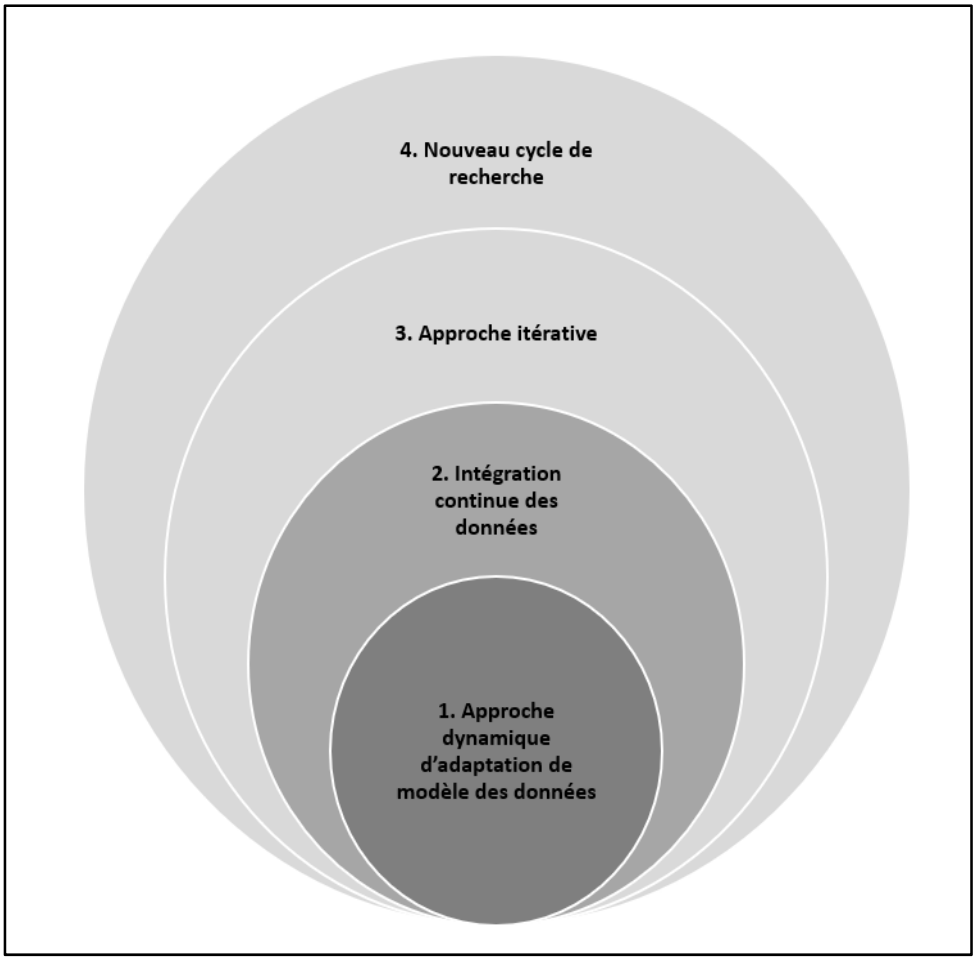


Figure 3.1 Composants de la nouvelle approche d'adaptation de modèle et d'intégration des données

3.1.1 Approche dynamique d'adaptation de modèle des données

Une approche dynamique d'adaptation de modèle des données permettra de répondre à la première question de recherche (voir section 0.2.3). Trois approches principales ont été proposées dans la littérature pour gérer les adaptations successives des schémas de base de données (section 2.3.4 Gestion des adaptations des modèles des données). Il a été constaté qu'au moment de rédiger cette thèse, aucune de ces approches ne permet la gestion de l'adaptation dynamique du modèle des données aux nouveaux besoins informationnels des recherches.

Le processus d'adaptation automatique du modèle des données original qui est proposé consiste à récupérer toujours la dernière version du modèle des données et de le bonifier à

l'aide d'un sous schéma de données destiné à emmagasiner les nouvelles données requises pour effectuer l'analyse lors de l'itération suivante. Afin de faciliter cette étape importante, un processus automatisé prendrait en charge la génération automatique du nouveau modèle des données. La figure 4 illustre le fonctionnement proposé du processus d'adaptation de modèle des données proposé.

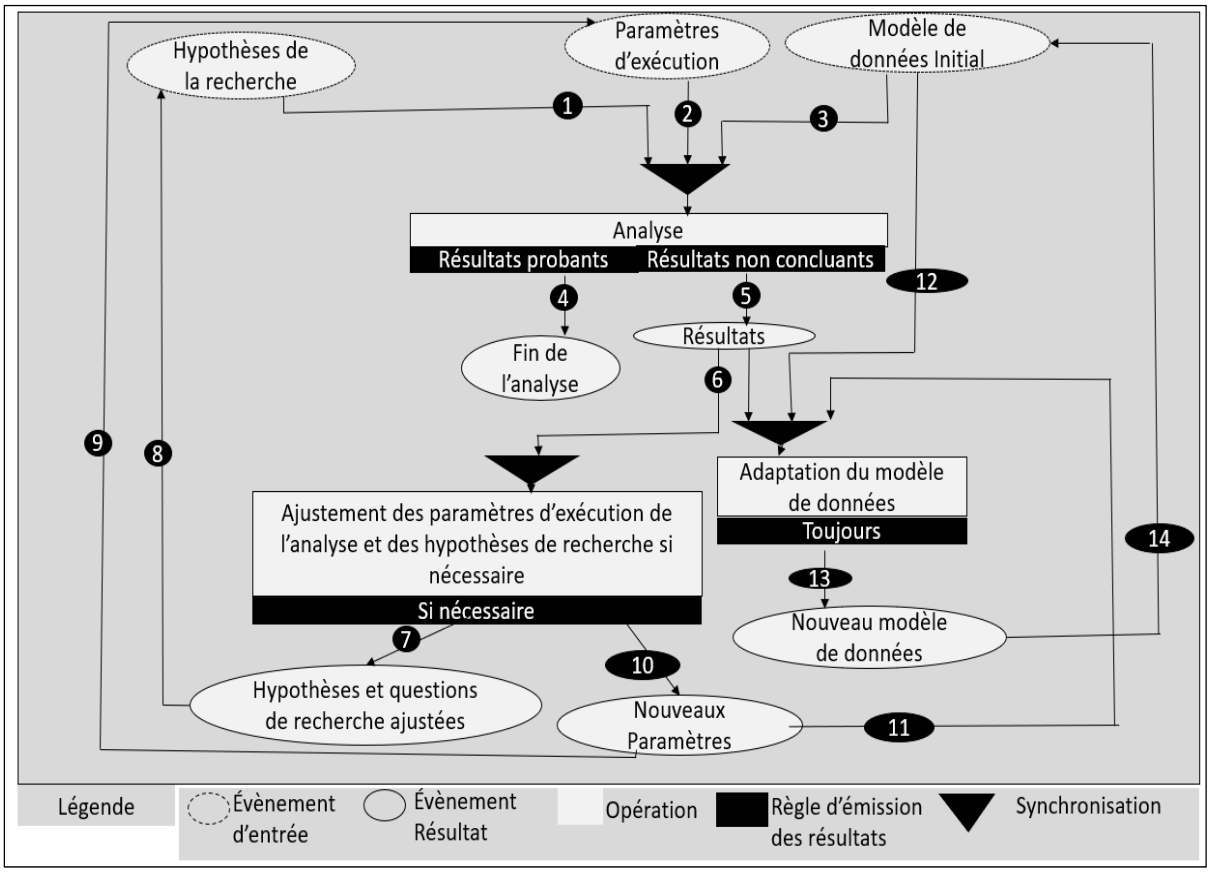


Figure 3.2 Processus de gestion de l'adaptation de modèle des données

Expliquons ici son fonctionnement. Lors de la première itération, le chercheur effectue son analyse en utilisant les hypothèses de recherche de départ, les paramètres d'exécution et le modèle des données initial. Si les résultats sont probants, le chercheur compile les résultats et termine l'analyse (voir l'étape 4 de la figure 3.2). Dans le cas contraire, il évalue les résultats obtenus, si nécessaires réévalue et réajuste les hypothèses de recherches et les paramètres

d'exécution (voir les étapes 7 et 10 de la figure 3.2). En utilisant ces nouveaux paramètres d'exécution et le modèle des données initial, il adapte le modèle des données pour obtenir un nouveau modèle des données capable de recevoir les données additionnelles nécessaires pour l'analyse de l'itération suivante (voir l'étape 13 de la figure 3.2). Ensuite il refait un autre cycle de recherche (voir les étapes 11, 14, 8 et 9 de la figure 3.2). Le modèle de données peut être adapté plusieurs fois dans une même recherche. Un modèle des données est un ensemble d'enregistrements de données (de l'anglais « record ») qui est l'équivalent d'un objet « entité » dans l'approche de modélisation de données entités-relation (Lescourret, Genest, Barnouin, Chassagne, & Faye, 1993).

3.1.2 Intégration continue des données

La composante d'intégration continue des données permettra de répondre à la deuxième question de recherche (voir section 0.2.3). La revue de littérature a permis de faire ressortir l'importance du besoin de l'intégration continue des données pour ce domaine de recherche (Jure et coll. 2014, Wolkenhauer et coll. 2014, Dubitzky, Krebs, et Eils 2001) et décris les défis qui empêchent sa réalisation (Baro et coll. 2015) (Duffy, 2018).

La figure 3.3 suivante illustre les étapes de la stratégie d'intégration continue des données proposées dans cette thèse. Nous avons vu à la section précédente qu'au cours d'une même recherche, le modèle de données peut subir plusieurs adaptations successives. Les activités d'adaptation de modèle de données (qui sont itératives) sont décrites par les étapes suivantes :

- Au début de la recherche, les chercheurs explorent la liste des informations disponibles dans les sources des données qui leur sont disponibles et identifient les besoins informationnels nécessaires pour réaliser leur recherche. C'est la version initiale des besoins informationnels "BI v0";
- Une fois la liste des besoins définie, un processus automatisé va créer la version initiale du modèle des données "MD v0";
- Une fois ce modèle des données initial disponible, un processus automatisé d'intégration des données va permettre d'intégrer les données correspondant au "BI v0" dans la base de données ;

- Une fois l'échantillon v0 des données est disponible dans la base de données, les chercheurs effectuent leurs analyses;
- En fonction des résultats obtenus, les chercheurs ajustent leurs hypothèses de recherche, identifient une nouvelle liste de besoin informationnel (BI v1), et répètent le cycle de l'étape 1 jusqu'à l'étape 4 de la figure 3.3, en créant de nouvelles versions du modèle des données "MD v1".

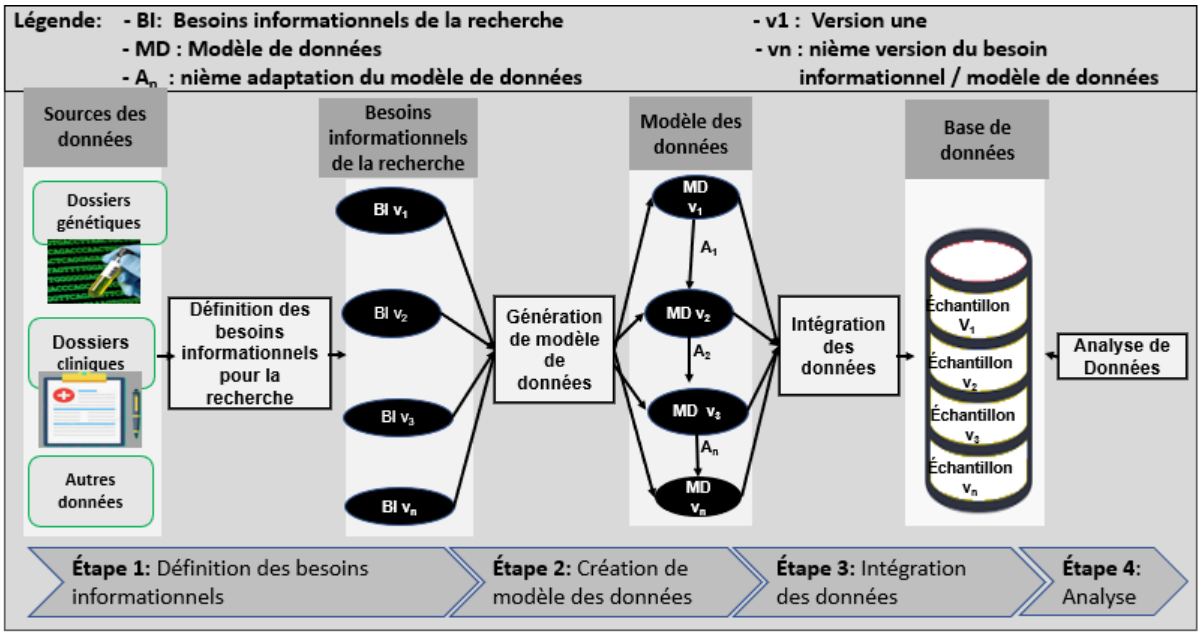


Figure 3.3 Stratégie de l'application de transformations successives aux modèles des données

La figure 3.4 suivante illustre un exemple de l'évolution du contenu de la base de données au cours des différentes itérations d'une recherche : dans la première itération, seulement les données génétiques et les données démographiques ont été identifiées et intégrées dans la base de données. Dans la deuxième itération, les données cliniques s'avéraient nécessaires pour compléter la recherche, le modèle de données a été adapté pour inclure les nouveaux besoins informationnels (les données cliniques) et ils ont été ajoutés à la base de données. De même, dans la troisième itération, les données de l'ethnie, le pays et la région ont été identifiés comme nécessaires à la recherche et ils ont été ajoutés à la base de données. Dans la nième itération, les données de laboratoire ont été ajoutées au modèle de données avec la fonction de génération

de modèle de données et qui a été utilisé par la fonction d'intégration des données pour ajouter ces données au contenu de la base de données.

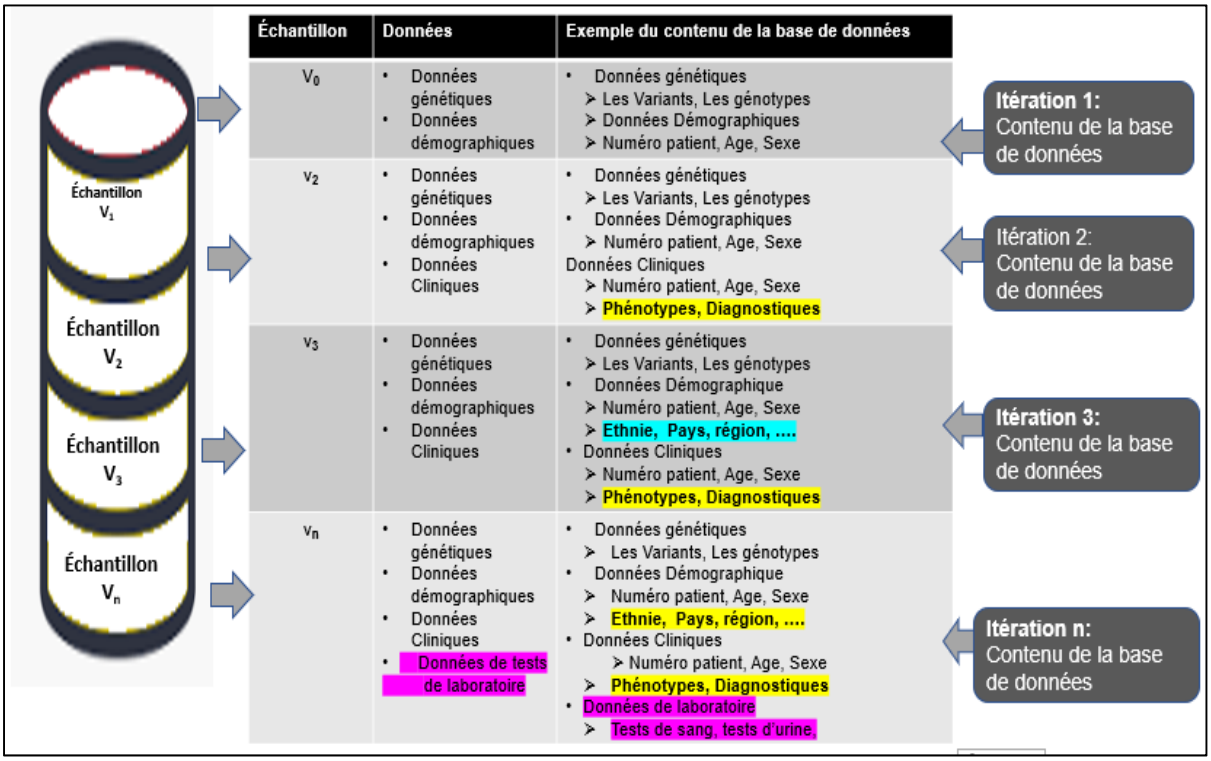


Figure 3.4 Exemple d'évolution du contenu de la base de données

Plus précisément chaque échantillon de données « v_i » dans la base de données est obtenu à l'aide de la formule mathématique suivante :

Soit:

- *MD-v_i: Le modèle des données de la ième itération*
- *Gen : Fonction de génération du modèle des données*
- *Integ : Fonction d'intégration des données*
- *Échantillon-v_i : l'échantillon de données de la ième itération d'analyse*
- A_i *: la ième adaptation du modèle des données.*

$$\underline{Échantillon\ vi = Integ}(\mathrm{Gen}(\mathrm{Ai}(\mathrm{MD} - \mathrm{vi}))$$

Pour s'assurer de l'efficacité des analyses sur des quantités massives de données, les deux choix de mise en œuvre ont été faits :

- Le format Avro a été choisi pour décrire le modèle de données. Avro est un cadriciel de sérialisation binaire compact qui utilise le format JSON.(The Apache Software Foundation, 2018).
- Le format de fichier Apache Parquet a été choisi pour le stockage des données. Ce format présente les avantages suivants :
    - Les données sont présentées en colonnes, c.-à-d. toutes les données du premier attribut sont sauvegardées, puis ceux de l'attribut suivant et ainsi de suite;
    - Les données sont divisées en page. Chaque page est constituée de blocs. Chaque bloc contient un certain nombre de colonnes de données.
    - Les pages sont stockées dans un format binaire;
    - Les fichiers binaires peuvent être compressés par un logiciel de compression.

La figure 3.5 suivante montre un exemple de représentation des données dans un format parquet.

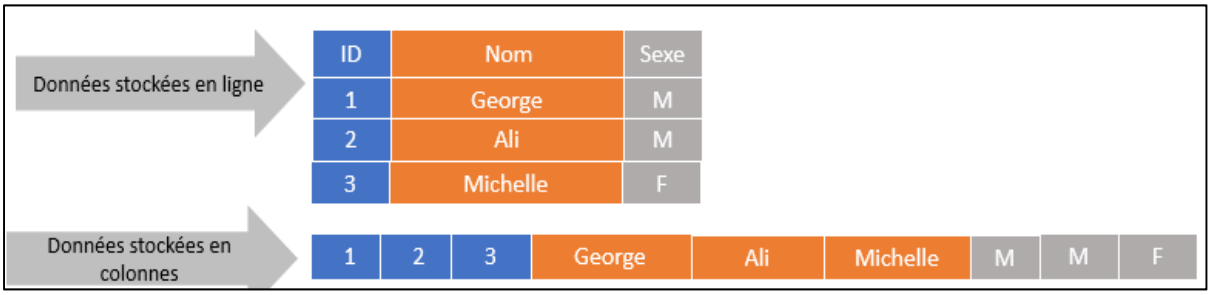


Figure 3.5 Exemple de stockage des données présentées en colonnes

3.1.3 Approche itérative

Nous avons vu que les recherches du domaine de la médecine de précision sont complexes de par leur nature même. Elles requièrent l'exploration de quantités massives de données provenant de plusieurs sources et ces données sont de types très différents. Afin d'obtenir des résultats probants, les chercheurs doivent répéter plusieurs fois leurs analyses en affinant

continuellement les paramètres d'exécution, les données et en ajustant la question de recherche jusqu'à ce qu'ils obtiennent des résultats concluants. Comme il a été mentionné lors de la revue de littérature (section 2.3.5 Approche itérative dans le processus de recherche), certains cycles de recherche en médecine de précision proposent l'utilisation d'un processus itératif (Drongelen, 2001) (Srivastava et Hopwood 2009).

L'utilisation d'un processus expérimental itératif est donc privilégiée par la solution proposée par cette thèse, car c'est l'approche expérimentale qui semble offrir la flexibilité requise pour les adaptations de modèle de données lors des itérations chez les chercheurs du domaine. L'utilisation de cette approche, appuyée par des technologies du domaine des données massives et de l'infonuagique, tentera de répondre à la question de recherche qui vise à permettre de réaliser plus facilement un cycle de recherche de manière itérative, c.-à-d. récupérer les résultats des analyses précédentes, ajuster le tir en fonctions des résultats obtenus et ajouter/modifier les données pour entraîner (ou ré entraîner) un modèle prédictif pour affiner l'analyse. La figure 3.6 ci-dessous illustre un exemple de la séquence des étapes d'analyse typiques dans le contexte d'une approche itérative.

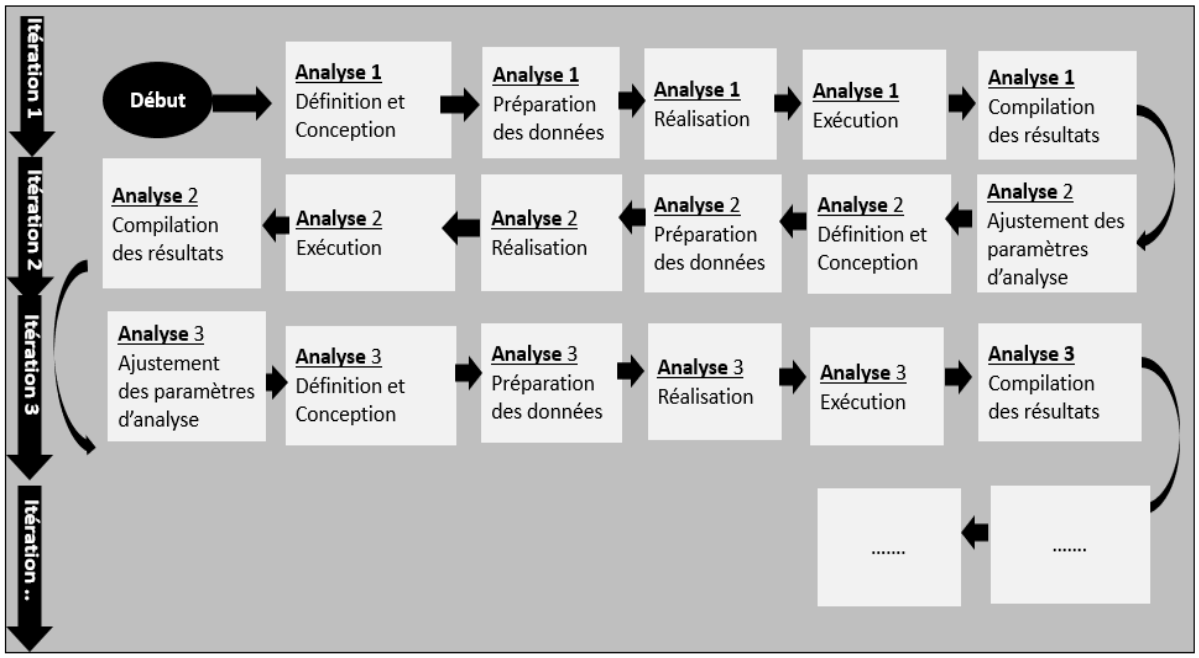


Figure 3.6 Processus expérimental itératif d'analyse des données

Dans cette approche, chaque itération est donc une nouvelle analyse raffinée et, conséquemment, les mêmes étapes sont répétées à chaque itération. Ces étapes sont:

- La définition et la conception de l'analyse : au mieux de leurs connaissances, les chercheurs définissent les hypothèses de leurs analyses, spécifient le type d'analyse de données à faire, les données à explorer et les seuils des valeurs des données à utiliser dans l'analyse;
- La préparation des données : les chercheurs réunissent (c.-à-d. collectent ou trouvent) les données nécessaires. Cette étape est très fastidieuse étant donné la complexité, la diversité des types et des sources des données, les contraintes technologiques reliées à la nature et la volumétrie des données obtenues. Pour adresser cette problématique, une approche novatrice dynamique d'adaptation de modèle des données est proposée dans cette thèse (voir figure 3.2);
- La réalisation de l'analyse : pour chaque itération d'analyse, et pour chaque échantillon de données, les chercheurs créent une nouvelle version du modèle d'analyse : nouveau modèle des données, nouveaux paramètres d'exécution et des objectifs d'analyse révisés. La révision des objectifs de la nouvelle analyse est faite en fonction des résultats obtenus dans les itérations précédentes;
- L'exécution de l'analyse : dans cette étape, les chercheurs exécutent les modèles d'analyses et produisent des résultats;
- L'interprétation des résultats : dans cette étape, les chercheurs compilent et analysent tous les résultats obtenus afin d'en tirer des conclusions;
- Ajustement des paramètres d'analyse : dans cette étape, les chercheurs ajustent les hypothèses de la recherche en fonction des conclusions tirées de l'étape précédente (et des itérations précédentes) et déterminent de nouveaux paramètres d'analyse (c.-à-d. autant au niveau des données que du modèle d'analyse) à utiliser pour l'itération suivante.

Ce processus est répété plusieurs fois jusqu'à ce que le chercheur obtienne des résultats qui lui permettent de valider ses hypothèses de recherches.

3.1.4 Nouveau cycle de recherche

À la section “3.1.3 Approche itérative” précédente, l’approche itérative proposée a été introduite. Dans cette section, les étapes de chaque itération vont être clarifiées et normalisées ainsi que l’orchestration de ces étapes pour compléter une itération du cycle de recherche.

Le cycle de recherche typique du domaine contient sept étapes consécutives :

1) Spécifier une hypothèse;
2) Identifier les données;
3) Adapter le modèle des données;
4) Intégrer ces données dans une base de données;
5) Faire une analyse;
6) Obtenir des résultats;
7) Comparer les résultats à l’hypothèse de départ.

Comme vu au Tableau 2.1 de la revue de littérature, aucun des logiciels de recherche consulté ne permettait l’adaptation automatique de modèle des données afin de s’adapter aux nouveaux besoins informationnels des recherches. L’introduction d’une nouvelle approche d’adaptation de modèle et d’intégration des données permettrait d’améliorer ce cycle avec l’ajout de nouvelles étapes et l’optimisation de certaines étapes du cycle actuel.

Le nouveau cycle de recherche proposée par cette thèse, illustrée à la figure 3.7 ici-bas, est composé de trois phases principales : la définition de l’analyse de la recherche; la préparation des environnements nécessaires pour effectuer une analyse subséquente; et l’exécution des étapes de l’analyse des données.

L’objectif de cette proposition de cycle de recherche vise à prendre des décisions concernant: l’adaptation dynamique de modèle des données, et l’approche d’analyse itérative. Le processus du nouveau cycle de recherche proposé contribue à répondre à la première et à la troisième question de recherche (voir section 0.2.3).

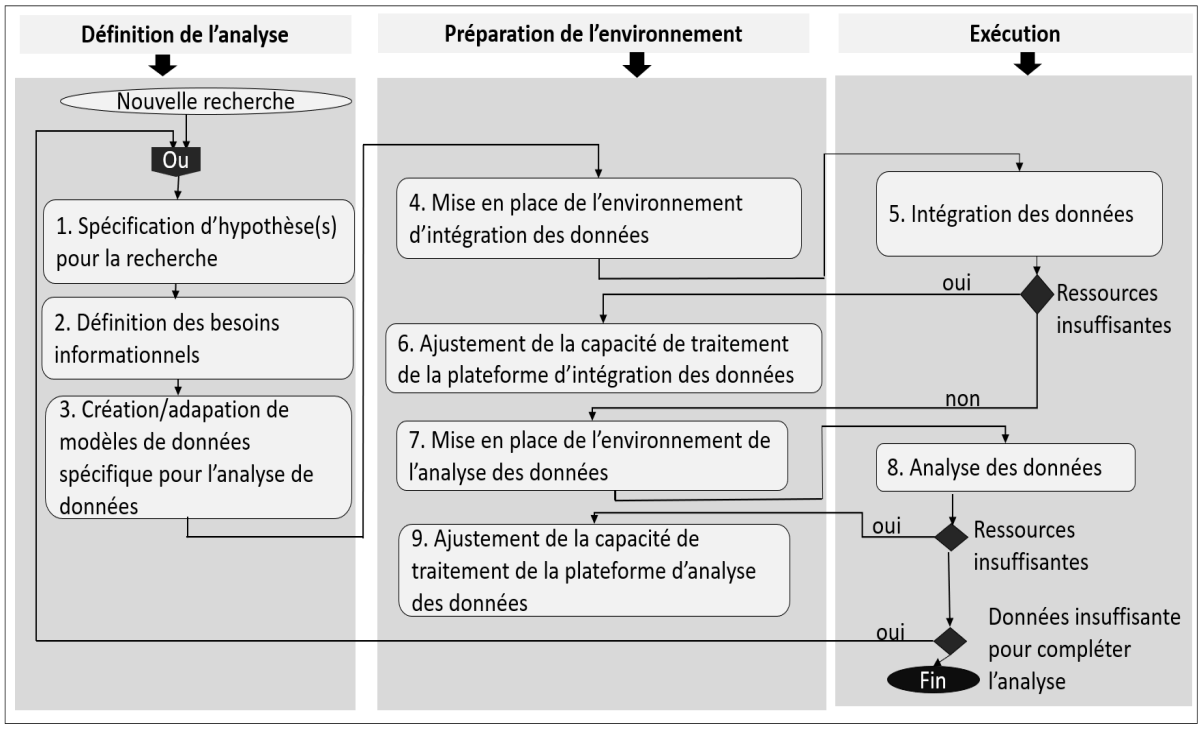


Figure 3.7 Nouveau cycle de recherche proposé

Les détails de chacune des étapes du cycle recherche illustrée à la figure 3.7, sont expliquées dans les sections suivantes.

3.1.4.1 Étape 1 : Définition de la recherche

Cette étape consiste essentiellement à définir tous les besoins informationnels requis pour réaliser les analyses de données prévues et, par la suite, préparer ce modèle des données qui sera utilisé ultérieurement lors de l'analyse.

1. Spécification d'hypothèse(s) pour la recherche

Cette étape consiste à définir la liste des hypothèses de la nouvelle recherche en médecine de précision. Dans le cas où les résultats obtenus d'une itération précédente de recherche s'avèrent

insuffisants pour valider ces hypothèses, les chercheurs, dans cette étape du cycle de recherche, peuvent réviser et ajuster ces hypothèses en fonction des résultats obtenus et refaire un autre cycle de recherche.

2. Définition des besoins informationnels

Cette étape d'identification de la liste détaillée des informations requises est typiquement manuelle et relève de la compétence des chercheurs dans leurs domaines d'expertises médicales. Cette étape est exécutée au début de chaque nouveau cycle de recherche.

3. Création/adaptation de modèles des données spécifique pour la recherche

Une fois la liste des informations (c.-à-d. les éléments de données) identifiée, la prochaine étape vise à collecter ces données à partir des différentes sources de données et de les enregistrer dans une base de données. Dans le cycle de recherche, une étape automatisée pour créer un modèle des données personnalisé, qui pourra être adapté aux besoins informationnels spécifiques et évolutifs de leur recherche a été prévue.

Cette étape de création et d'adaptation de modèle des données présente le processus de génération dynamique de modèle des données et les actions nécessaires pour mettre en œuvre ce processus :

- **Processus dynamique de génération de modèle des données**

  La Figure 3.8 suivante illustre le modèle de traitement du processus de génération du modèle des données de la recherche.

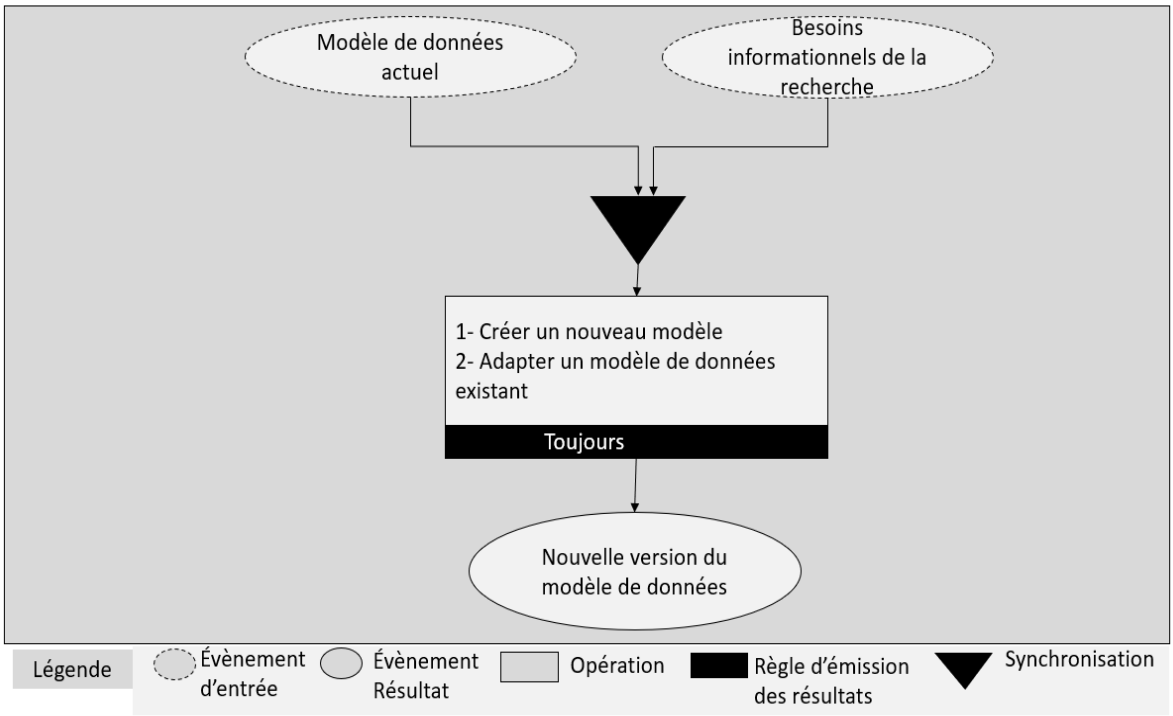


Figure 3.8 Modèle conceptuel de traitement du processus de génération de modèle des données

La mission du processus d'aide à la génération de modèle des données est de s'assurer de créer une nouvelle version du modèle des données à partir de deux intrants : Les besoins informationnels de la recherche; et la version actuelle du modèle des données. Décrivons plus en détail ces intrants :

- Les besoins informationnels de la recherche peuvent être : les besoins initiaux que les chercheurs ont pu déterminer au démarrage de la recherche; ou de nouveaux besoins qui ont été spécifiés par les chercheurs à la suite d'analyses provenant d'itérations précédentes de la recherche;
- La version actuelle du modèle des données : les chercheurs peuvent utiliser n'importe quel modèle des données provenant de recherches antérieures et le bonifier avec leurs nouveaux besoins ou bien simplement concevoir un nouveau modèle des données;
    - C'est donc un processus de génération dynamique de modèle des données qui est proposé dans cette thèse (Belghait, Kanzki, et April 2018) et qui permet aux chercheurs de :

- Mettre en place un processus itératif de réalisation de recherche dans leurs organisations. En effet, il est très courant qu'au début de la recherche, les chercheurs aient juste un aperçu global sur l'étendue de leur recherche et les besoins informationnels nécessaires, et au fur et aux mesures qu'ils progressent dans leurs analyses, de nouvelles perspectives apparaissent et de nouveaux besoins informationnels deviennent indispensables pour le succès de la recherche;
- Obtenir un modèle des données mature avec le temps qui leur permet de faire plusieurs recherches et ce dans des délais plus courts.

- **Mise en œuvre du processus dynamique de génération de modèle des données**

L'objectif d'un modèle des données adapté à ce domaine est qui doit utiliser des techniques de pointe de compression des données, car il va héberger une grande quantité de données qui : premièrement, peuvent provenir de plusieurs sources de données; deuxièmement, peuvent être de types très diversifiés; et, finalement, va être utilisée par un processus d'analyse de données à grande échelle. Ce modèle de données doit donc répondre à ces trois contraintes efficacement. Étant donné la popularité grandissante de l'utilisation de technologies de traitement des données massives sur le nuage (Belghait & April, 2018) qui permet un traitement efficace d'un grand volume de données, il a été décidé que le générateur de modèle des données va créer un modèle des données qui utilisera une des technologies de traitement des données massives efficace sur l'infonuagique. Dans le chapitre suivant, les choix technologiques et les raisons derrière ces choix seront présentés et expliqués en détail.

3.1.4.2 Étape 2 : Préparation des environnements

Une fois la liste des besoins informationnels définie (figure 3.3, Étape 1), les chercheurs vont devoir exécuter les étapes de la génération du modèle des données (figure 3.3, étape 2), intégrer ces données dans ce modèle des données (Figure 3.3, Étape 3) et puis explorer ces données (Figure 3.3, Étape 4). Cette phase est composée de quatre activités principales :

1. **Mise en place de la plateforme d'intégration des données :** en utilisant l'API de configuration du logiciel d'intégration (voir l'API 3 Configurateur d'infrastructure matérielle à la Figure 3.10); le chercheur configurera une grappe de serveurs virtuels qu'il pourra utiliser afin de migrer les données vers le modèle de données adapté. Le chercheur contrôlera lui-même la taille de la grappe et le type de serveurs virtuels à utiliser. Le budget dont il dispose pour faire la recherche, le délai visé et le volume des données sont les principaux éléments qui influencent ses choix de configuration de l'environnement d'intégration des données;
2. **Ajustement de la capacité de traitement de la plateforme d'intégration des données :** si, au cours du processus d'intégration des données, la capacité de traitement de la grappe de serveurs virtuels configurée s'avère insuffisante pour intégrer tout le volume de données assez rapidement, le chercheur pourra reconfigurer l'environnement d'intégration soit en ajoutant plus de serveurs virtuels ou bien remplacer la taille des serveurs virtuels configurés par d'autres, plus puissants, en utilisant cet API de configuration (voir l'API 3 Configurateur d'infrastructure matérielle à la Figure 3.10);
3. **Mise en place de la plateforme d'analyse des données :** une fois toutes les données intégrées et stockées dans cette base de données, le chercheur pourra détruire/éliminer l'environnement d'intégration (c.-à-d. fermer définitivement les serveurs virtuels utilisés dans la grappe). Il sera facile de recréer cette infrastructure plus tard si nécessaire et ainsi ne pas avoir à payer pour des serveurs virtuels inutilisés. Pour effectuer l'étape d'analyse des données, le chercheur créera un nouvel environnement d'analyse en utilisant la procédure et l'API de configuration d'environnement d'analyse. Encore une fois, le chercheur contrôlera la configuration de cet

environnement sur demande, et selon le type d'analyse et le volume de données, il pourra choisir le nombre et le type de serveurs virtuels à utiliser;

4. **Ajustement de la capacité de traitement de la plateforme d'analyse :** pendant le déroulement du processus d'analyse des données, s'il s'avère que l'environnement d'analyse manque de capacité de traitement pour certains types d'analyse, le chercheur pourra facilement reconfigurer l'environnement en y ajoutant des serveurs virtuels additionnels sur demande, et les fermer automatiquement aussitôt qu'il aurait terminé son analyse en utilisant la procédure d'ajustement du logiciel de recherche de médecine de précision (voir l'API 3 Configurateur d'infrastructure matérielle à la Figure 3.10).

3.1.4.3 Phase 3 : Exécution

La phase d'exécution inclut deux activités principales :

- **Intégration des données :** l'objectif de cette activité est d'intégrer toutes les données identifiées nécessaires pour la réalisation de la recherche et qui peuvent provenir de plusieurs sources de données. Chaque source de données utilise typiquement un format particulier de données. Dans le cadre de cette recherche, il n'a pas été considéré d'offrir une fonctionnalité qui prévoit le traitement de tous les types de formats populaires du domaine. Conséquemment, dans une première version du prototype expérimental conçu, deux formats de données populaires ont été pris en compte : le format de données provenant de bases de données relationnelles et le format des données génétiques stockées dans des fichiers textes au format de fichier de génotype d'Oxford (c.-à-d. les fichiers avec une extension « .gen »). L'intégration automatisée de ces deux types de données populaires est effectuée en utilisant l'API d'intégration des données (c.-à-d. l'API 2 présenté à la Figure 3.10) qui permet d'intégrer les données dans un même format (c.-à-d. dans un format de fichiers binaires au format Parquet) dans une base de données. La durée de l'étape d'intégration des données dépend de la capacité de traitement choisie lors de la phase de préparation de l'environnement (section : 3.1.4.2 Étape 2 : Préparation des environnements);

- **Analyse des données :** une fois les données chargées et intégrées dans la base de données, le chercheur pourra utiliser un outil/logiciel d'analyse de son choix pour nettoyer, structurer et analyser ces données.

3.2 Conception du prototype expérimental pour valider la solution proposée

3.2.1 Aperçue générale des décisions technologiques et d'architecture du prototype

Un prototype logiciel expérimental doit être conçu afin d'expérimenter et de valider les propositions théoriques de conception décrites dans les sections précédentes. Ce prototype logiciel devra donc permettre de rencontrer les exigences suivantes : débutons par l'exigence qui vise à « … permettant ainsi d'effectuer des analyses sur des quantités massives de données » de l'objectif de recherche (présentée à la section 0.2.2). Pour rencontrer cette exigence, une première décision de conception est que le prototype devra être en mesure de présenter les fonctionnalités comme des services infonuagiques sur demande (SaaS) (Belghait & April, 2018) et ainsi supporter le modèle de livraison d'infrastructure informatique comme un service (IaaS). Ces deux décisions de conception permettent d'offrir une fonctionnalité simple d'ajustement de la capacité élastique de calcul pour le chercheur. Le prototype logiciel devra ainsi offrir l'accès à des serveurs virtuels en mode IaaS, sur demande, et les rendre disponibles en quelques minutes. De plus si le chercheur a besoin de capacité additionnelle (ou de capacité moindre), il pourra faire varier le nombre de serveurs virtuels selon le besoin.

Une deuxième décision de conception est qu'il sera avantageux d'effectuer le développement du prototype expérimental à l'aide des logiciels libres afin de minimiser les coûts de licences, d'avoir le contrôle sur l'évolution du code source de la plateforme de médecine de précision et finalement profiter au maximum de briques logicielles libres de droits qui sont déjà disponibles et éprouvées pour réutilisation étant donné que toute la communauté des utilisateurs de ces logiciels participe à l'évolution de ces logiciels.

Une troisième décision de conception est d'utiliser le modèle des données du projet ADAM de l'Université Berkeley qui est une brique logicielle libre qui suggère un modèle des données

générique pour traiter de très grandes quantités de données génétiques et utilise les technologies des données massives libres de droits telles que Avro et Spark. Le projet ADAM a été choisi pour les deux raisons suivantes : premièrement, il n'utilise que des logiciels libres du domaine des données massives; et deuxièmement il offre un modèle des données génétiques assez complet qui supporte des formats de fichier génétique populaires, par exemple : BAM, VCF, GEN et bien d'autres. Ainsi, cette décision de conception permet l'utilisation de logiciels libres adaptés à la problématique de cette recherche (c.-à-d. l'utilisation de technologies de données massives du domaine de la génétique), conjointement avec l'utilisation de services infonuagique. Cette décision de conception s'assure que le prototype logiciel puisse répondre aux trois questions de recherche présentées à la section 0.2.3.

La figure 3.9 suivante présente un aperçu général de l'architecture logicielle proposée pour le prototype expérimental qui vise à s'assurer de la mise en œuvre de l'approche itérative de recherche ainsi que l'adaptation dynamique du modèle des données. L'architecture logicielle proposée est découpée en quatre modules de fonctionnalités principales:

1) **Acquisition des Données :** ce premier module (voir en bas à gauche de la figure 3.9) permet de combiner les différentes sources de données et leur stockage. Pour la première version du prototype logiciel, l'acquisition sera limitée aux données stockées dans les bases de données relationnelles et celles stockées dans des fichiers textes et respectant le format de fichier de génotype d'Oxford (c.-à-d. les fichiers avec une extension « .gen »);
2) **Intégration des données :** nous avons vu qu'un des objectifs est de regrouper toutes les données nécessaires à la recherche dans une seule base de données et au sein d'un même modèle des données dynamiquement adaptable aux besoins informationnels changeants des recherches. De plus, l'intégration d'une quantité massive de données doit être exécutée rapidement à l'aide d'une infrastructure infonuagique échelonnable facilement. Ce deuxième module du prototype logiciel expérimental (voir en bas de la figure 3.9) va mettre en œuvre un processus dynamique d'adaptation de modèle des données pour accueillir toute donnée nécessaire provenant du modèle des données du

chercheur. Ceci nécessite de concevoir des APIs spécialisées qui permettent au chercheur d'adapter (c.-à-d. modifier lui-même) le modèle des données génétiques existant d'ADAM selon ses besoins.

3) **Analyse des données :** à l'aide de ce module (voir au bas de la figure 3.9), les chercheurs pourront utiliser un logiciel d'analyse de données de leur choix pour effectuer les analyses des données. Le développement d'une composante logicielle d'analyse ne fait pas partie de la portée de cette recherche. L'objectif du prototype logiciel se limite à fournir aux chercheurs une base de données efficace qui permet d'effectuer des analyses à l'aide d'autres outils/logiciels divers de leur choix;
4) **Applications cliniques :** cette étape de la recherche en médecine de précision ne fait pas partie de la portée de cette recherche. L'application clinique des résultats d'une recherche de précision est une question assez complexe et nécessite la répétabilité des résultats et la certification auprès d'organismes réglementaires.

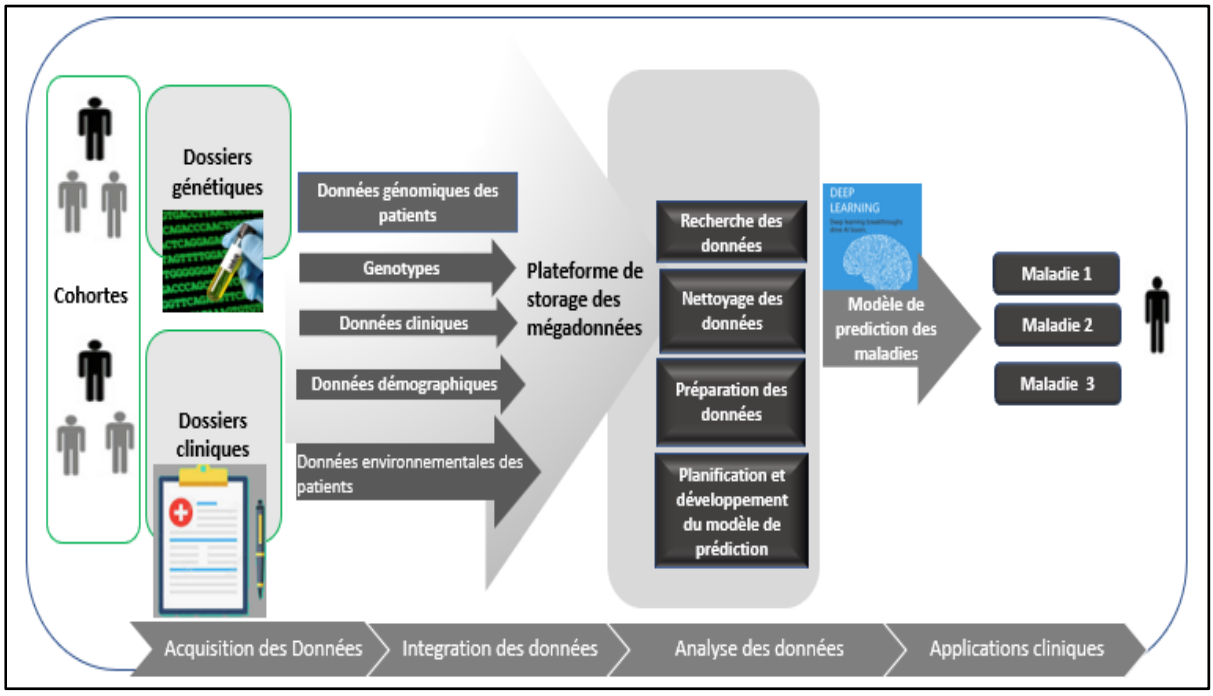


Figure 3.9 Aperçue générale de l'approche de solution proposée

3.2.2 Composantes logicielles du prototype expérimental

3.2.2.1 Conception des composantes logicielles du prototype expérimental

Cette section présente les concepts généraux des décisions de conception du prototype expérimental. Ces décisions visent à atteindre les objectifs décrits dans les chapitres précédents. Elle sera suivie d'une section qui décrira plus en détail les choix technologiques.

Afin d'aider à adresser les défis auxquels font face les chercheurs du domaine de la médecine de précision actuellement et atteindre les objectifs visés et énoncés à la section (0.2 But, Objectifs et Questions de recherche), le prototype expérimental doit aussi s'assurer d'atteindre les objectifs spécifiques de conception suivant :

- **Flexibilité du modèle des données :** le modèle des données doit pouvoir s'adapter facilement et rapidement afin de s'aligner avec n'importe quel objectif de recherche du domaine de la médecine de précision;
- **Évolutivité :** cette caractéristique qualité du prototype expérimental est essentielle afin de rencontrer trois exigences, plus particulièrement :
  - **L'adaptabilité de l'importation d'une grande quantité de données de patients:** les données des patients (c.-à-d. les données cliniques, génétiques, etc.) doivent pouvoir être transférées facilement sur un modèle des données personnalisé dans un court délai;
  - **L'élasticité de l'infrastructure informatique :** les chercheurs doivent être capables de :
    - Ajuster la capacité de traitement de l'infrastructure matérielle facilement pour pouvoir traiter un très grand volume de données dans des délais raisonnables. Ajouter et enlever des serveurs virtuels (c.-à-d. de la puissance de calcul) devrait être facile;
    - Ajouter/enlever des serveurs virtuels de l'infrastructure informatique devrait aussi permettre de faire des économies et réduire les coûts d'une recherche;

- **Évolutivité de l'analyse des données de la recherche :** les chercheurs doivent pouvoir réutiliser l'analyse de leurs données existantes et les aligner avec les nouvelles demandes d'analyse de données.

- **Reproductibilité (ou Réplication) :** cette caractéristique du prototype expérimental doit permettre aux chercheurs de répliquer facilement les résultats de toute analyse de données antérieure à partir des mêmes échantillons de données et dans les mêmes conditions d'analyse.

3.2.2.2 Conception détaillée du prototype expérimental

La Figure 3.10 suivante décrit la vue d'ensemble de la conception des composantes logicielles du prototype expérimental :

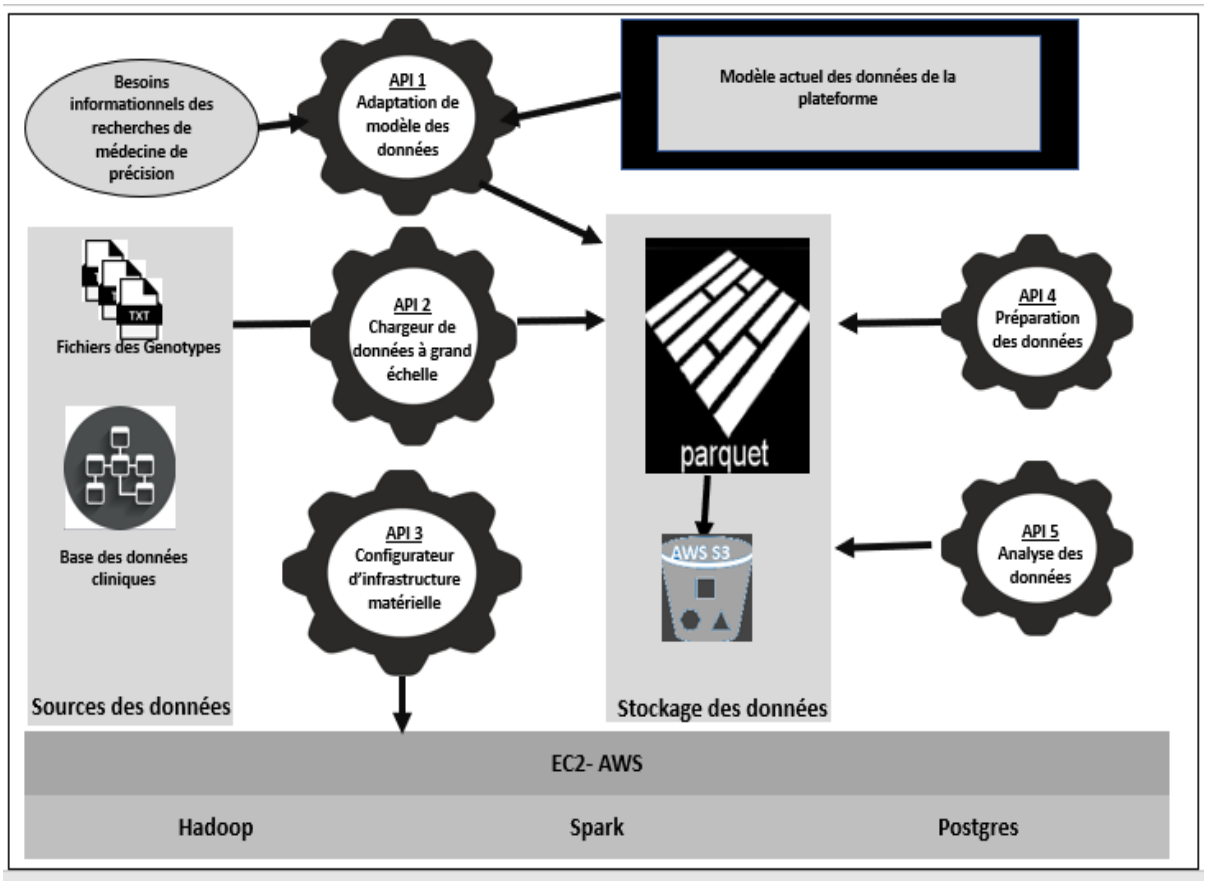


Figure 3.10 Conception détaillée et choix technologiques pour le prototype expérimental

Cette conception détaillée décrit les composants et les technologies utilisées pour réaliser les objectifs architecturaux et la conception de haut niveau. La section suivante décrit chaque composant de la conception illustrée dans la Figure 3.10 ci-haut plus les requis non fonctionnels de la solution.

a. **Traitement des données :** des interfaces applicatives (API) pour les traitements de données
   - **API 1. Adaptation de modèle des données :**
     - **Objectif :** cette API permet de créer un nouveau modèle des données ou bien adapter un modèle des données existant avec de nouveaux besoins informationnels;
     - **Mise en œuvre :** deux logiciels libres sont choisis pour développer cette API :
       - **Apache Maven :** un logiciel source libre pour la gestion et pour l'automatisation de la compilation et de la génération des binaires des programmes;
       - **Apache Avro :** un logiciel libre qui permet de définir un format de données capable de supporter le traitement distribué de grands volumes de données sur l'infonuagique. Ce format de données est supporté par une grande variété de données (The Apache Software Foundation, 2012);
       - Le langage Python pour Spark (PySpark) et les scripts Shell;
   - **API 2. Chargeur des données à grande échelle**
     - **Objectif :** cette API permet d'intégrer les données génétiques et cliniques dans le nouveau modèle des données implanté dans un format Avro. La première version de l'API supporte les données sauvegardées dans les bases de données relationnelles, et les données génétiques sauvegardées dans des fichiers texte et respectant le format de représentation des données génétiques d'Oxford (c.-à-d. les fichiers avec une extension « .gen »);

- **Mise en œuvre :** pour développer cette API, les choix technologiques sont :
  - Le langage Python pour Spark (PySpark) et les scripts Shell;
  - Apache Spark, un cadriciel des données massives qui permet le traitement parallèle en mémoire vive de très grands volumes de données. Afin de garantir un traitement rapide, ce cadriciel effectue des traitements distribués sur des grappes des serveurs virtuels. Il a été éprouvé dans le cadre du projet ADAM de l'Université Berkeley qui a démontré son efficacité pour résoudre des problématiques d'analyses génétiques.

- **API 3. Configurateur d'infrastructure matérielle**
  - **Objectif :** cette API permet aux utilisateurs d'installer et configurer de manière automatique leur infrastructure (c.-à-d. une grappe de serveurs virtuels) nécessaire pour l'intégration des données et celle pour l'analyse et l'exploration des données;
  - **Mise en œuvre :** pour le développement de cette API, le langage Python et les scripts Shell ont été sélectionnés.
- **API 4. Préparation des données**
  - **Objectif :** cette API permettra aux utilisateurs d'analyser les données chargées à partir des sources de données dans le nouveau modèle des données afin de les nettoyer et les préparer pour la phase d'analyse et d'exploration. En fonction de la qualité des données et du type de l'analyse que l'on voudrait faire, cette API peut nécessiter des adaptations afin de répondre aux besoins. Dans le prototype expérimental du logiciel, une API accompagnée d'exemples de code à utiliser sera développée. Elle servira pour nettoyer et préparer les données (c.-à-d. créer des échantillons des données, regroupement des données, etc.);
  - **Mise en œuvre :** le même langage de programmation et le même logiciel pour le développement de cette API ont été choisis: PySpark et Spark.

- **API 5. Analyse des données**
    - **Objectif :** cette API vise à mettre à la disposition des chercheurs une sorte de boîte à outils avec les logiciels d'analyse de données leur permettre d'explorer leur échantillon de données;
    - **Mise en œuvre :** dans la première version, les outils d'analyse d'Anaconda et PySpark sont mis à la disposition des chercheurs. Dans le futur proche il est planifié d'ajouter l'outil source libre d'apprentissages machine et d'intelligence artificielle H2O.

b. **Persistances des données :** les données seront placées dans le service d'archivage S3 de AWS. Ce service infonuagique permet de stocker de très grandes quantités de données à des coûts très raisonnables. Ce choix est motivé principalement par le coût de stockage très bas et la simplicité/fiabilité du service de stockage;

c. **Fiabilité:** la fiabilité du logiciel dépend de celle des services du fournisseur infonuagique utilisé. Cette première version du prototype expérimental sera basée sur les services infonuagiques de AWS.

3.3 Résumé

Ce chapitre a présenté l'approche proposée d'adaptation de modèle et d'intégration des données pour les chercheurs du domaine de médecine de précision afin de pallier aux problèmes auxquels ils font face actuellement à chaque fois qu'ils entreprennent une nouvelle recherche. Cette nouvelle approche a été présentée selon cinq perspectives :

a. Approche d'adaptation dynamique de modèle des données : cette section a présenté l'approche, le processus et les étapes qui vont permettre aux chercheurs de gérer leurs besoins informationnels particuliers à chaque itération d'analyse à travers l'adaptation dynamique du modèle des données;

b. **Intégration continue des données :** Cette section apporte des éléments de réponse à la deuxième question de recherche (voir section 0.2.3). Elle a présenté la stratégie d'intégration continue des données proposées dans cette thèse.

c. **Approche itérative d'analyse de données :** cette section a présenté les deux points suivants :
    i. La pertinence d'utiliser un processus expérimental itératif lors de la réalisation de recherche;
    ii. La description du processus expérimental itératif d'analyse des données à utiliser dans la nouvelle solution proposée.

d. **Cycle de recherche de médecine de précision :** dans cette section, une proposition d'un nouveau cycle de recherche de médecine de précision a été présentée. Il est basé sur les différents composants de la nouvelle approche d'adaptation de modèle et d'intégration des données présentés dans la Figure 3.1.

e. **Vue d'ensemble de la solution proposée :** une vue d'ensemble de la proposition de la nouvelle approche d'adaptation de modèles et d'intégration des données pour les recherches en médecine de précision décrit les fonctions logicielles principales nécessaires (Figure 3.9), à savoir : l'acquisition des données, l'intégration des données et la fonctionnalité de l'analyse de données;

f. **Architecture d'un prototype expérimental :** cette section a finalement présenté les différents composants logiciels et matériels nécessaires d'un prototype expérimental utilisant les concepts proposés, incluant : une liste d'interfaces applicatives développées pour réaliser les étapes de la recherche et les choix technologiques faits pour le développement et le déploiement d'un prototype logiciel pour effectuer une preuve de concept.

Pour démontrer le potentiel de cette nouvelle approche et appuyer les chercheurs dans la réalisation de leurs recherches de médecine de précision, une étude de cas portant sur une

recherche en médecine de précision conduite par le laboratoire de recherche du Dr Pavel Hamet à Montréal au Canada est planifiée. Cette recherche suivra chaque étape du cycle de recherche proposée à la figure 3.7 et en utilisant le prototype expérimental développé et préalablement testé. Le chapitre 4 suivant présente en détail les essais préparatoires du prototype expérimental, la planification de l'expérimentation (c.-à-d. l'étude de cas avec le Dr Hamet) et l'explication du déroulement de l'expérimentation; et une discussion des résultats obtenus.

CHAPITRE 4

VALIDATION DU PROTOTYPE ET EXPÉRIMENTATION

4.1 Introduction

Ce chapitre présente les activités de validation du prototype expérimental ainsi que son expérimentation lors d'une étude de cas, permettant d'évaluer et de valider la proposition d'approche d'adaptation de modèle et d'intégration des données, ainsi que le cycle de recherche proposé dans cette thèse.

Ce chapitre présente les six sujets suivants :

1. Les étapes de validation de chaque composant de la nouvelle approche d'adaptation de modèle et d'intégration des données;
2. L'étude de cas à réaliser à l'aide de la nouvelle approche proposée;
3. Le processus actuel de réalisation des recherches en médecine de précision du laboratoire du Dr Hamet au CRCHUM;
4. Les concepts mesurés et la méthode de mesure de chaque objectif de la recherche;
5. La réalisation d'une étude de cas avec la nouvelle approche;
6. Les résultats de l'expérimentation de la nouvelle approche.

Une étude de cas qui vise à valider l'approche proposée dans cette thèse a été définie avec la participation d'une équipe de chercheurs en médecine de précision au centre de recherche du centre hospitalier de l'Université de Montréal (CRCHUM). Les étapes suivantes ont été effectuées afin de planifier l'expérimentation et par la suite effectuer l'étude de cas:

1. Concevoir la méthode de mesure en précisant les concepts et les techniques pour chaque objectif à mesurer (Jacquet and Abran 1997);
2. Concevoir, valider et envoyer un questionnaire aux chercheurs pour obtenir les mesures de références du processus actuel afin de pouvoir les comparer avec celles obtenues à la suite de l'expérimentation;
3. Obtenir et documenter les réponses du questionnaire et valider les mesures obtenues;

4. Valider le prototype logiciel, en laboratoire, pour s'assurer qu'il effectue toutes les fonctions nécessaires et qu'il implémente fidèlement l'approche proposée dans cette thèse;
5. Préparer l'étude de cas avec l'équipe de recherche du Dr Hamet au CRCHUM;
6. Réaliser l'étude de cas et collecter les mesures.

4.2 Validation de l'approche proposée

Un prototype logiciel a été développé comportant quatre composants logiciels tels que décrits au chapitre précédent.

Lors du développement des composants, deux types de tests ont été effectués : premièrement, des tests unitaires sur chaque composant afin de vérifier son bon fonctionnement et deuxièmement, des tests de validation d'atteignabilité des objectifs de la recherche ont été effectués.

Premièrement, cette recherche vise à atteindre un objectif principal. Afin de pouvoir valider et mesurer à quel point la nouvelle approche d'adaptation de modèle et d'intégration des données permet d'atteindre cet objectif, des sous-objectifs ont été définis et des tests unitaires ont été effectués afin de s'assurer que la nouvelle approche répond à chacun d'eux. Les résultats détaillés de ces tests d'évaluation du nouveau cycle de recherche sont présentés à l'ANNEXE III.

Le Tableau 4.1 suivant présente la hiérarchie de l'objectif et des sous-objectifs mesurables de cette recherche. Les mesures de chaque sous-objectif sont présentées dans le modèle de mesure de la section 4.5 Objectifs, concepts mesurés et mesures.

Tableau 4.1 Hiérarchie des objectifs de la recherche

Objectif principal de recherche	Sous- objectifs
• **Objectif 1 :** Améliorer l'approche d'adaptation de modèle des données utilisé dans le cycle de recherche en médecine de précision afin de permettre aux chercheurs de pouvoir continuellement adapter le modèle des données, d'une recherche, aux besoins informationnels changeants et permettre d'intégrer toutes ces données dans une seule base de données permettant ainsi d'effectuer des analyses sur des quantités massives de données.	• **Objectif 1.1 :** permettre aux chercheurs de pouvoir continuellement adapter le modèle des données de la recherche aux nouveaux besoins informationnels et faire des analyses de données sur de grandes quantités de données; • **Objectif 1.2 :** améliorer et simplifier les tâches de préparation des données qui empêchent les chercheurs à concentrer leurs efforts (activités de gestion et d'administration de la plateforme d'analyse utilisée, intégration des données, etc.) et réduire les efforts (délais et ressources) requis pour la réalisation de ces tâches; • **Objectif 1.3** : améliorer et simplifier le processus de reproductibilité des analyses antérieures et réduire les efforts (délais et ressources) requis; • **Objectif 1.4 :** réduire les efforts (délais et ressources) requis pour l'adaptation de modèle des données dans le cycle de recherche.

Des tests unitaires ont été effectués sur chaque composant du prototype expérimental du logiciel qui met en œuvre la nouvelle approche. Les tableaux 4.2, 4.3 et 4.4 suivants décrivent les tests d'atteignabilité de chaque sous-objectif et les résultats obtenus.

Tableau 4.2 Liste des tests unitaires et résultats d'atteignabilité des objectifs 1.1 et 1.4

<table>
<tr><td>Objectifs</td><td>• Objectif 1.1: permettre aux chercheurs de pouvoir continuellement adapter le modèle des données de la recherche aux nouveaux besoins informationnels et faire des analyses de données sur de grandes quantités de données;
• Objectif 1.4: réduire les efforts (délais et ressources) requis pour l'adaptation de modèle des données dans le cycle de recherche.</td></tr>
<tr><td>Éléments de la nouvelle approche</td><td>• Approche dynamique d'adaptation de modèle des données (section 3.1.1 du chapitre 3)</td></tr>
<tr><td>Composants logiciels</td><td>• API 1 : adaptation de modèle des données (Figure 3.10)</td></tr>
<tr><td>Test de validation</td><td>• Objectifs du test : tester que l'API permet de réutiliser un modèle existant, ajouter de nouvelles classes et générer un nouveau modèle des données, puis vérifier le délai de création de nouveaux modèles des données
• Étapes de test
<table>
<tr><th>#</th><th>Test</th><th>Résultat</th></tr>
<tr><td>1</td><td>• Exécuter l'API en utilisant le modèle des données de ADAM</td><td>• Création d'un modèle des données en format Avro</td></tr>
<tr><td>2</td><td>• Dans un éditeur de texte, créer un nouveau fichier « patient.avdl », importer le fichier « bdg.avdl », et ajouter la définition d'une classe « Patients »;
• Exécuter l'API d'adaptation de modèle pour ajouter l'entité de données Patients</td><td>• Le modèle des données de ADAM est créé et adapté pour ajouter la classe Patients : « patients.avsc » prête à être utilisée.</td></tr>
<tr><td>3</td><td>• Éditer le fichier « patient.avdl », ajouter des informations à la définition de l'enregistrement « Patients », ajouter la définition d'une nouvelle classe « medicament »</td><td>• Les classes du test précédent plus une nouvelle classe : « medicament.avsc » a été créé</td></tr>
</table></td></tr>
</table>

Tableau 4.3 Liste des tests unitaires et résultats d'atteignabilité de l'objectif 1.2

<table>
<tr><td>Objectifs</td><td>• Objectif 1.2 : améliorer et simplifier les tâches de préparation des données qui empêchent les chercheurs à concentrer leurs efforts (activité de gestions et d'administration de la plateforme d'analyse utilisée, intégration des données, etc.) et réduire les efforts (délais et ressources) requis pour la réalisation de ces tâches;</td></tr>
<tr><td>Composants logiciels du prototype</td><td>• API 1 : adaptation de modèle des données (Figure 3.10)
• API 2 : chargeur de données à grande échelle (Figure 3.10)
• API 3 : configurateur de l'infrastructure matérielle (Figure 3.10)</td></tr>
<tr><td>Test</td><td>1. Objectifs du test : évaluer les efforts et les délais d'une itération d'analyse de données avec le prototype

2. Étapes de test :
<table>
<tr><th>#</th><th>Test</th><th>Résultat</th></tr>
<tr><td>1</td><td>• Exécuter l'API 1 d'adaptation de modèle des données
• Créer et configurer une grappe de dix serveurs virtuels de type « t2.medium »
• Exécuter l'API 2 pour intégrer les données de dix personnes
• Exécuter l'API 3 pour créer et configurer une grappe de serveurs virtuels (grappe de dix serveurs virtuels de type « t2.medium »</td><td>• Le processus au complet a duré environ deux heures depuis la préparation du fichier de définition de données « patients.avdl » jusqu'à la disponibilité des données dans la grappe d'analyse.</td></tr>
</table>
• </td></tr>
</table>

Tableau 4.4 Liste des tests unitaires et résultats d'atteignabilité de l'objectif 1.3

<table>
<tr><td>Objectifs</td><td>• Objectif 1.3 : améliorer et simplifier le processus de reproductibilité des analyses antérieures et réduire les efforts (délais et ressources) requis;</td></tr>
<tr><td>Composants logiciels du prototype</td><td>• API 3 : configurateur de l'infrastructure matérielle (Figure 3.10)</td></tr>
<tr><td>Test</td><td>• Objectifs du test : Vérifier que la nouvelle solution permet de recréer facilement l'environnement d'analyse et récupérer les données d'une analyse antérieure

• Étapes de test
<table>
<tr><th>#</th><th>Test</th><th>Résultat</th></tr>
<tr><td>1</td><td>• Créer un nouveau serveur virtuel et le configurer</td><td>• L'exécution de l'API 3 a permis, dans un délai moins de 15min, de configurer le nouveau serveur virtuel qui va être utilisé pour exécuter les APIs de gestion de la plateforme</td></tr>
<tr><td>2</td><td>• Exécuter le script automatisé de création de nouvelles grappes d'analyse (2 serveurs virtuels)</td><td>• L'API a permis dans un délai inférieur à cinq minutes, de créer et configurer une nouvelle grappe de serveurs virtuels prête à exécuter les scripts d'analyse.</td></tr>
<tr><td>3</td><td>• Récupérer les données de tests d'une analyse antérieure</td><td>• À l'aide de l'interface utilisateur de AWS, déplacer les données de l'ordinateur de bureau vers le service S3 de AWS.</td></tr>
</table></td></tr>
</table>

Cette section a permis de préciser que la mise en œuvre du prototype logiciel permet d'atteindre chacun des sous-objectifs de l'objectif principal de la recherche '0.2.2 Objectif de la recherche' en effectuant les tests unitaires. La prochaine étape a été de tester le comportement du prototype logiciel à l'aide d'un réel cas de recherche en médecine de précision. La section suivante présente en détail l'étude de cas qui va être utilisée avec le prototype logiciel.

4.3 Définition de l'étude de cas

L'étude de cas a été proposée par l'équipe de recherche en médecine de précision du laboratoire de recherche en médecine de précision du Dr Hamet au CRCHUM. Elle consiste à tenter de découvrir un modèle prédictif qui permettrait d'identifier les patients à risque de développer une insuffisance rénale chronique (CKD) chez les patients atteints de diabète de type 2 (DT2).

Pour effectuer cette analyse de médecine de précision, les données génétiques, cliniques et démographiques de patients seront utilisées afin d'alimenter les trois algorithmes de prédiction suivants : l'algorithme d'analyse prédictive de réseau de neurones (de l'anglais : Neural Network (NN)), l'algorithme d'analyse prédictive de forêt aléatoire (de l'anglais : Random Forest (RF)) et l'algorithme d'analyse prédictive de régression linéaire (de l'anglais : Linear Regression (LR)). Ces trois algorithmes ont été sélectionnés par les chercheurs à cause de leur popularité dans les publications récentes au sujet de la sélection génétique (Kourou et coll. 2015) (Endo, Shibata, & Tanaka, 2008).

L'objectif de l'étude de cas est de démontrer les capacités du modèle théorique d'analyse proposé par cette thèse qui est maintenant implanté dans le prototype logiciel. Il s'agit ici d'une nouvelle recherche en analyse de médecine de précision qui démarre par une première itération qui va effectuer une analyse avec des modèles prédictifs populaires. L'approche proposée par les chercheurs est d'identifier la liste des variants génétiques ayant des facteurs de risques potentiels pour le développement d'une insuffisance rénale. Ils ont été présélectionnés par les chercheurs du CRCHUM à partir de données GWAS disponibles en libre accès.

La portée de la validation du modèle théorique proposé par cette thèse est :

- Définir les besoins informationnels de la recherche de l'étude de cas, identifier les données nécessaires et les mettre telles quelles dans un endroit centralisé sur disque;
- Créer un modèle de données adapté à la recherche
- Intégration des données et gestion de la plateforme matérielle requise pour l'intégration des données;

- Colocalisés toutes les données dans une seule base de données et les rendre disponibles pour réaliser les analyses;
- Réaliser l'analyse de l'étude de cas;
- Démontrer que le modèle théorique de l'analyse proposé permet la reproduction des analyses antérieures.

Les données utilisées pour cette étude de cas ont été fournies par le laboratoire du Dr Hamet au CRCHUM. Toutes les données sont anonymisées et aucune information ne permet d'identifier les patients.

Une demande de certificat d'éthique a été soumise à l'ÉTS autorisant ainsi l'utilisation de l'échantillon de données anonymisées dans cette preuve de concept. Le responsable de l'éthique, à l'ÉTS, a conclu qu'aucun certificat d'éthique n'était requis vu que toutes les données sont anonymisées et aucune information dans l'échantillon utilisé ne permet d'identifier les patients auxquels appartiennent ces données.

Cette section a présenté les étapes et les résultats des validations unitaires effectuées sur les composantes de la nouvelle solution et a défini le réel cas d'étude à utiliser pour valider la solution proposée dans son ensemble. La section suivante va présenter le processus actuel de recherche de l'équipe du Dr Hamet au CRCHUM.

4.4 Description du processus actuel de réalisation des recherches

4.4.1 Composition de l'équipe de recherche

La Figure 4.1 suivante illustre la composition de l'équipe de recherche du laboratoire du Dr Hamet au moment de la préparation et la réalisation de l'étude de cas. L'équipe est formée de deux chercheurs séniors qui s'occupent de la définition des recherches et de l'interprétation des résultats et une équipe de bio-informaticiens qui s'occupe de la réalisation des différentes itérations d'analyse des données.

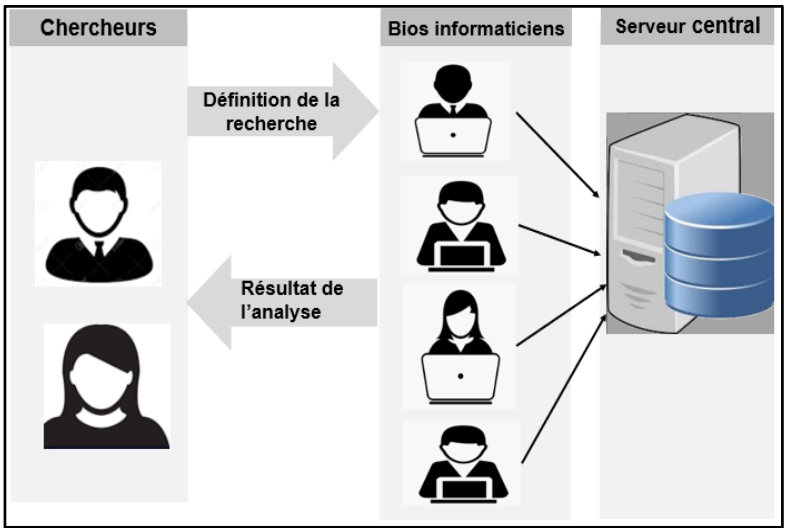


Figure 4.1 Composition de l'équipe de recherche actuelle

4.4.2 Description de l'infrastructure matérielle actuelle

L'équipe de recherche possède un serveur central physiquement installé dans le CRCHUM. Sur ce serveur sont installés les bases de données et les logiciels requis pour les analyses. Les bio-informaticiens travaillent sur leurs ordinateurs de bureau pour le développement des algorithmes d'analyses et faire les tests unitaires. Quand ces algorithmes sont validés, ils sont exécutés sur le serveur central en mode lot pendant la journée ou la nuit. Les bio-informaticiens doivent synchroniser entre eux pour ne pas envoyer les travaux d'analyse en même temps à cause de la capacité traitements du serveur central.

La section suivante présente le modèle de mesure (c.-à-d. concepts à mesurer, mesures et méthode de collecte de mesures) qui a été développé préalablement pour évaluer les résultats de l'utilisation du prototype logiciel.

4.5 Objectifs, concepts mesurés et mesures

Cette section précise les concepts mesurés et la méthode de mesure de chaque objectif de la recherche en préparation pour l'étude de cas. Nous avons vu, à la section 2.4 que la première étape de la conception d'un modèle de mesure selon (Jacquet and Abran 1997), décrit qu'il est nécessaire de préciser les concepts et la méthode de mesure pour chaque mesure planifiée pour

une expérimentation de génie logiciel. Afin d'évaluer d'une manière quantitative chaque objectif visé par la nouvelle approche d'adaptation de modèle des données et d'intégration des données dans le contexte d'une recherche en médecine de précision, une liste de mesures a été proposée initialement. Par la suite, cette liste a été envoyée, sous forme d'un questionnaire, aux membres de l'équipe de recherche du Dr Hamet qui ont travaillé sur plusieurs projets de recherche en médecine de précision avec les données de la cohorte d'ADVANCE qui vont être utilisées dans l'étude de cas. L'objectif de ce sondage est de vérifier, avec eux, préalablement, que des données existent actuellement pour ces concepts/mesures avant l'étude de cas. Ce questionnaire a été envoyé aux membres de l'équipe de recherche sous forme de sondage d'opinion via le site de sondage en ligne MonkeySurvey. Ce questionnaire comportait dix-huit questions (voir ANNEXE I et II).

4.5.1 Concepts mesurés et mesures

Chaque question du sondage vise à fournir des données nécessaires pour mesurer l'objectif de ce projet de recherche (préalablement énoncés à la section 0.2.2). L'ensemble des critères établis permettent de mesurer: des durées, des efforts, des simplifications qui révèlent les avantages de la solution proposée par cette thèse.

Cette section présente la conception du modèle de mesure nécessaire à l'évaluation du degré d'atteignabilité de chaque sous-objectif (Tableau 4.1-Hiérarchie des objectifs de la recherche). La conception du modèle de mesure nécessite pour chaque sous-objectif d'identifier une liste de concepts à mesurer, puis pour chaque concept identifié, établir la liste de mesures à évaluer et l'approche de collecte de ces mesures lors de l'étude de cas :

- **Objectif 1.1 :** permettre aux chercheurs de pouvoir continuellement adapter le modèle des données de la recherche aux nouveaux besoins informationnels et faire des analyses de données sur de grandes quantités de données.
    - **Concept à mesurer :**
        - **C1 :** La capacité des chercheurs de pouvoir adapter le modèle des données par eux même.

- **Mesures :**
  - **Mesure 1 :** le pourcentage des chercheurs capable d'adapter le modèle des données par eux même. C'est le rapport entre le nombre de chercheurs du CRCHUM capable d'adapter le modèle de données par eux même sur le nombre total des chercheurs.

$$\% = \frac{chercheurs\ capable\ d'adapter\ le\ modèle\ par\ eux\ même}{\text{nombre total des chercheurs}} \quad (1)$$

- **Approche de capture des mesures**
  - **Données de références :** la mesure 1 sera évaluée par l'analyse des réponses aux questions Q4 et Q5 de l'ANNEXE I;
  - **Données de validation du modèle proposé :** au cours de la réalisation de l'étude de cas, évaluer le nombre de personnes requises pour adapter le modèle des données aux nouveaux besoins informationnels.

- **Objectif 1.2 :** améliorer et simplifier les tâches de préparation des données (activités de gestion et d'administration de la plateforme d'analyse utilisée, intégration des données, etc.) et réduire les efforts (c.-à-d. délais et ressources) requis pour la réalisation de ces tâches;
  - **Concepts à mesurer :**
    - **C1 :** l'effort nécessaire pour gérer la plateforme matérielle. Cet effort est évalué à l'aide de deux paramètres : le temps consacré à l'exécution des tâches de gestion de la plateforme matérielle et le nombre de personnes impliquées dans la réalisation de ces tâches;
    - **C2 :** les efforts nécessaires pour exécuter les tâches d'adaptation de modèle des données et celles d'intégration des données. Cet effort est une durée de temps (heures et minutes) écoulé pendant l'exécution de ces tâches et prend aussi en compte le nombre de personnes requis pour la réalisation de ces tâches;
    - **C3 :** le pourcentage d'automatisation des tâches de préparation des données.

- **Mesures :**
  - **Mesure 2 :** le nombre de personnes impliquées dans le processus de mise à niveau de la capacité de traitement de l'infrastructure matérielle;
  - **Mesure 3 :** le temps (en nombre d'heures et de minutes) encouru pour la mise à niveau de la capacité de traitement de l'infrastructure matérielle. Ce critère calcule la durée de temps depuis l'envoi de la demande de mise à niveau jusqu'à l'exécution des tâches de mise à niveau par le personnel technique où la nouvelle infrastructure devient prête à être utilisée;
  - **Mesure 4 :** le pourcentage d'automatisation des tâches de gestion de l'infrastructure matérielle;
  - **Mesure 5 :** le temps (en heures et minutes) nécessaire pour intégrer les données dans la même base de données et les rendre prêts à être utilisées par les modèles d'analyses des chercheurs. Après l'adaptation de modèle des données aux besoins informationnels de l'itération courante, les chercheurs vont devoir importer les données manquantes à partir de leurs sources de données et les intégrer avec les données des itérations précédentes. Ce critère évalue le temps (en heures et minutes) nécessaire pour intégrer ces données dans la même base de données et les rendre prêts à être utilisées par les modèles d'analyses des chercheurs. Il permet de mesurer le degré d'amélioration du processus d'intégration des données;
  - **Mesure 6 :** le nombre de fois (c.-à-d. le nombre d'itérations d'analyse) que les chercheurs doivent refaire le cycle de recherche avant d'arriver à des résultats concluants. Les chercheurs doivent souvent refaire l'analyse plusieurs fois en modifiant les paramètres de la recherche : les hypothèses, les données de tests, etc. Ce critère contribue à l'évaluation approximative de la durée de réalisation d'une recherche en médecine de précision;

- **Mesure 7 :** le nombre de fois que les chercheurs doivent adapter le modèle des données. Dans chaque itération, les chercheurs en modifiant les paramètres de la recherche risquent d'avoir besoin de nouvelles données qui n'étaient pas prévues dans les échantillons initiaux. Afin d'ajouter les données manquantes, souvent, ils doivent changer le modèle des données pour l'adapter aux nouveaux besoins informationnels de l'itération en cour. Ce critère permet de mesurer le gain en temps dans la tâche d'adaptation de modèle des données;
- **Mesure 8** : le nombre de fois que les chercheurs doivent mettre à niveau la capacité de traitement de l'infrastructure matérielle utilisée dans l'analyse des données afin de pouvoir traiter les grands volumes de données. Durant le cycle de recherche, les chercheurs utilisent des données de tests différents, en ajoutant des données à chaque itération, la capacité de traitement de l'infrastructure matérielle risque de ne plus être capable de traiter les nouveaux volumes. Ce critère évalue la fréquence de mise à niveau de l'infrastructure durant le cycle de recherche. Il permet d'évaluer le gain en temps dans la gestion de l'infrastructure matérielle;

- **Approche de collecte des mesures**
  - **Données de références :** les mesures 1, 2, 3, 5, 6, 7 et 8 seront évaluées à l'aide des réponses aux questions de sondages envoyés aux chercheurs du CRCHUM (les questions Q10, Q9, Q7, Q2, Q3, et Q8 de l'ANNEXE I); les mesures 4 et 15 seront évaluées à l'aide de la réponse aux questions Q6 et Q5 de l'ANNEXE II;
  - **Données de validation du modèle proposé :** lors de la réalisation de l'étude de cas, capturer le temps (en heures et minutes) nécessaire pour accomplir chaque tâche reliée à la gestion de la plateforme.

- **Objectif 1.3 :** améliorer le processus de reproductibilité des analyses antérieures.
  - **Concept à mesurer :**
    - **C1 :** la capacité de reproduire des analyses antérieures;
    - **C2 :** l'effort nécessaire pour reproduire une analyse. Cet effort est évalué par deux paramètres : le temps nécessaire pour reproduire les analyses antérieures et le nombre de personnes requis pour la réalisation des tâches de reproduction de ces analyses.
  - **Mesures :**
    - **Mesure 9 :** le pourcentage des chercheurs capable de reproduire une analyse antérieure de médecine de précision. C'est le rapport entre le nombre de chercheurs du CRCHUM capable de reproduire des analyses antérieures sur le nombre total des chercheurs.

$$\% = \frac{chercheurs\ capable\ de\ reproduire\ une\ analys}{\text{nombre total des chercheurs}}$$

    - **Mesure 10 :** la taille des échantillons des données traitées par une analyse typique en médecine de précision dans le CRCHUM;
    - **Mesure 11 :** le nombre de personnes impliquées dans le processus de reproduction d'analyses antérieures;
    - **Mesure 12 :** le temps pour reproduire une analyse antérieure. Le nombre d'heures écoulées pendant l'exécution des tâches de reproduction d'une analyse antérieure.
  - **Approche de collecte des mesures**
    - **Données de références :** les questions Q1, Q2 et Q3 de l'ANNEXE II et la question Q1 de l'ANNEXE I permettront de collecter les quatre mesures;
    - **Données de validation du modèle proposé :** refaire l'analyse de l'étude de cas et collecter les mesures 9,10,11 et 12.

- **Objectif 1.4 :** réduire l'effort (les délais et le nombre de personnes) requis pour l'adaptation de modèle des données dans le cycle de recherche.
    - **Concept à mesurer :**
        - **C1 :** l'effort (délais et nombre de personnes) nécessaire pour configurer et adapter le modèle des données aux besoins informationnels de l'analyse. Cet effort sera évalué à l'aide de deux paramètres : premièrement, le délai (nombre d'heure et de minutes) écoulé depuis le début jusqu'à la fin de l'exécution de la tâche de configuration et d'adaptation de modèle et deuxièmement, le nombre de personnes impliquées dans la réalisation de cette tâche.
    - **Mesures :**
        - **Mesure 13 :** le pourcentage d'automatisation des tâches d'adaptation de modèle des données. L'activité d'adaptation de modèle de données étant composée de trois tâches : définition du modèle, création de la nouvelle version du modèle et préparation de l'infrastructure matérielle pour créer le modèle : le pourcentage d'automatisation et le rapport entre le nombre de tâches automatisées par le nombre total des tâches.

$$\% = \frac{nombre\ de\ tâches\ automatisées}{\text{nombre total des tâches}} \quad (3)$$

        - **Mesure 14 :** le nombre de personnes impliquées dans les processus de mise à jour et la création du modèle des données. Ce critère compte seulement les ressources techniques impliquées dans le processus d'adaptation de modèle des données. Il n'inclut pas les personnes impliquées dans le processus d'identifications des informations requises pour la nouvelle itération de recherche;
        - **Mesure 15 :** le temps (en nombre d'heures et de minutes) écoulé pendant l'adaptation de modèle des données utilisé dans l'analyse de données. Ce critère compte seulement le temps de création et d'adaptation de modèle des données. Le temps de l'identification,

des informations et leurs sources des données ne sont pas incluses. Il permet d'évaluer les efforts d'adaptation de modèle des données et par conséquent contribue à la mesure du degré d'amélioration du processus de gestion du modèle des données.

- **Approche de capture des mesures**
  - **Données de références :** la mesure 13 sera évaluée à l'aide de la réponse à la question Q5 de l'ANNEXE II, et les mesures 14 et 15 seront évaluées à l'aide des réponses aux questions Q5 et Q4 de l'ANNEXE I;
  - **Données de validation du modèle proposé :** lors de la réalisation de l'étude de cas, utiliser l'API 1 d'adaptation de modèle des données (voir Figure 3.10) pour adapter le modèle aux nouveaux besoins informationnels de la recherche et mesurer le temps d'adaptation de modèle des données.

4.5.2 Processus de collecte des mesures

Afin de mesurer les critères d'évaluation du processus d'analyse, deux questionnaires ont été préparés et envoyés aux trois bio-informaticiens du laboratoire de médecine de précision du Dr Hamet au CRCHUM : le premier sondage comportant dix questions (voir ANNEXE I). Il vise à évaluer le processus des nouvelles recherches. Le deuxième sondage comportant huit questions (voir ANNEXE II) et vise à évaluer l'aspect de reproductibilité des recherches antérieures. Chaque question de ces deux questionnaires vise à mesurer un ou plusieurs aspects du processus de recherche actuel de l'équipe de chercheurs en médecine de précision. Ce sont les mesures des données de références.

4.5.2.1 Présentation des résultats du sondage d'opinion

Suite à l'envoi des questions des deux sondages aux membres de l'équipe du laboratoire, 100% des réponses ont été reçues (voir ANNEXES I et II). Le sondage a été envoyé aux trois bio-informaticiens responsables de la réalisation des recherches. Le Tableau 4.5 suivant contient

la compilation des réponses par rapport aux critères de mesure du processus de recherche établie dans la section 4.5.1 concepts et mesure ci-haut.

Tableau 4.5 Réponses aux sondages concernant le processus d'analyse actuel

#	Mesure	Processus actuel
Mesure 1	Le pourcentage des chercheurs capable d'adapter le modèle des données par eux même.	0%[1]
Mesure 2	Le nombre de personnes impliquées dans le processus de mise à niveau de la capacité de traitement de l'infrastructure matérielle. (ANNEXE I, Q10).	3 personnes
Mesure 3	Le temps (nombre d'heure et de minutes) encouru pour la mise à niveau de la capacité de traitement de l'infrastructure matérielle. (ANNEXE I, Q9).	7 heures
Mesure 4	Le pourcentage d'automatisation des tâches de gestion de l'infrastructure matérielle. (ANNEXE II, Q6).	17%
Mesure 5	Le temps (heures et minutes) nécessaire pour intégrer les données dans la même base de données et les rendre prêts à être utilisées par les modèles d'analyses des chercheurs. (Q7, ANNEXE I)	4 heures
Mesure 6	Le nombre de fois (c.-à-d. le nombre d'itérations d'analyse) que les chercheurs doivent refaire le cycle de recherche avant d'arriver à des résultats concluants. (ANNEXE I, Q2).	Plus que 10 itérations
Mesure 7	Le nombre de fois que les chercheurs doivent adapter le modèle des données**.** (ANNEXE I, Q3).	Plus que 10 itérations
Mesure 8	Le nombre de fois que les chercheurs doivent mettre à niveau la capacité de traitement de l'infrastructure matérielle utilisée dans l'analyse des données afin de pouvoir traiter les grands volumes de données. (ANNEXE I, Q8).	1.66 fois
Mesure 9	Le pourcentage des chercheurs capable de reproduire une analyse de médecine de précision. (ANNEXE II, Q1).	Oui=66.67% Non=33.33%

[1] 3 personnes en moyenne sont impliquées dans le processus d'adaptation de modèle des données ce qui implique que les chercheurs ont toujours besoin d'aide pour faire personnaliser le modèle des données de leur recherches à leurs besoins.

#	Mesure	Processus actuel
Mesure 10	La taille des données traitées par une analyse typique en médecine de précision. (Q1, ANNEXE I) et (Q1 ANNEXE II).	6 GOS
Mesure 11	Le nombre de personnes impliquées dans le processus de reproduction d'analyses antérieures. (ANNEXE II, Q3)	2.33 personnes
Mesure 12	Le temps pour reproduire une analyse antérieure. C'est le nombre d'heures écoulées pendant l'exécution des tâches de reproduction d'une analyse antérieure. (ANNEXE II, Q2).	15.33 heures
Mesure 13	Le pourcentage d'automatisation des tâches d'adaptation de modèle des données. (ANNEXE II, Q5).	33%
Mesure 14	Le nombre de personnes impliquées dans les processus de mise à jour et la création du modèle des données (ANNEXE II, Q5).	3 personnes
Mesure 15	Le temps (nombre d'heure et de minutes) écoulé pendant l'adaptation du modèle des données utilisé dans l'analyse de données. (ANNEXE I, Q4).	18 heures

4.5.2.2 Analyse des réponses du sondage d'opinion

L'analyse des réponses des chercheurs a fait ressortir trois observations principales :

1. La taille des échantillons des données utilisées dans les recherches est relativement petite par rapport à la taille des échantillons de données disponibles. La cohorte d'ADVANCE utilisée par les chercheurs comporte un échantillon de 205 GOS, alors que les chercheurs rapportent qu'ils ont travaillé avec des échantillons de 1 GO pour le 1er chercheur, 6 GOS pour le 2e et 48 GOS pour le 3e. Le répondant 3, qui a travaillé avec le plus grand échantillon de données (c.-à-d. un échantillon de 48 GOS) a répondu qu'il n'était pas capable de reproduire son analyse. Cette information semble indiquer que le serveur central actuel localisé au CRCHUM ne soit plus en mesure de traiter des quantités massives de données;
2. Les réponses des chercheurs sur les efforts de certaines étapes sont très différents surtouts entre ceux qui travaillent sur des échantillons relativement petits (1 GO et 6

GOS) et ceux qui travaillent avec des échantillons de taille plus grande (48 GOS). Dans le 1er groupe, le délai de création du modèle des données est entre 2 à 5 heures alors que dans le 2e groupe le délai est de 10 jours (voir ANNEXE I et II pour la différence entre les différentes estimations). Ce constat mène aux trois conclusions suivantes :

a. Le processus actuel semble ne pas être normalisé entre les chercheurs, donc chacun a sa façon propre de travailler et possède des outils (c.-à-d. logiciels et infrastructures) personnalisés;
b. Dans le processus actuel, la taille des échantillons a un impact majeur sur les délais d'exécutions des tâches d'analyse;
c. Deux répondants parmi trois ont dit qu'il n'y a pas d'automatisation des tâches de gestion de la plateforme actuelle d'analyse (tâche de mise à niveau de l'infrastructure). L'opinion du troisième répondant est que le processus de gestion de la plateforme est automatisé à 50%, alors que selon lui le délai requis afin de mettre à niveau la plateforme est de plus de 10 heures. Ces réponses nous permettent de constater que le processus actuel de gestion de la plateforme de recherche est peu automatisé. Son pourcentage d'automatisation est établi à 17% d'automatisation (ANNEXE II, Q6).

3. Les écarts entre les réponses des trois membres de l'équipe de recherche sont notables. En effet, pour la question du degré d'automatisation de l'infrastructure d'analyse, l'écart-type est de 29% (ANNEXE II, Q17). Deux questions ont été posées pour l'estimation du délai de la phase de préparation des données, l'écart-type pour la première réponse était de 36.59% et 22.23% pour la deuxième réponse. Ces types d'écart ont été observés dans la majorité des réponses (voir ANNEXE I et ANNEXE II).

4.6 Exécution des étapes de l'analyse de la recherche pour la réalisation de l'étude de cas

La Figure 4.2 suivante présente le nouveau cycle de recherche dans le contexte de la réalisation de l'étude de cas. Avant de débuter une nouvelle recherche en médecine de précision, les

données brutes de la cohorte d'ADVANCE, qui inclut les données génétiques et les données cliniques, ont été copiées et stockées dans un compartiment[2] S3 de stockage de AWS.

Il est important ici de rappeler que toutes ces données brutes utilisées dans l'étude de cas sont complètement anonymisées, c'est-à-dire, les données de l'échantillon seules ne permettent pas de faire le lien entre les données et les patients.

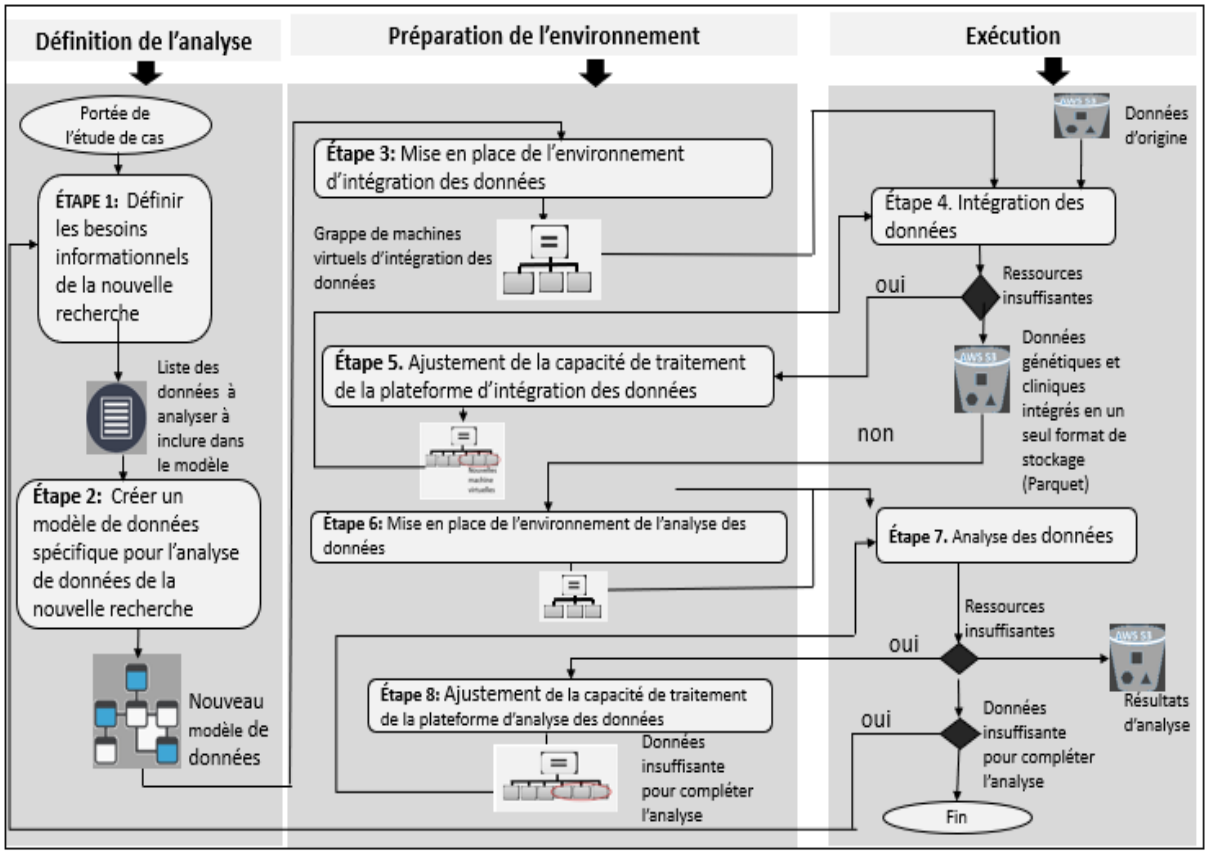


Figure 4.2 Flux d'exécution des étapes de l'étude de cas

Dans la section suivante, chaque étape de la Figure 4.2 est décrite.

4.6.1 Étape 1 : Définir les besoins informationnels de la nouvelle recherche

Les données cliniques utilisées pour cette étude de cas proviennent des données de la cohorte ADVANCE : une expérimentation clinique contrôlée incluant des patients atteints du diabète

[2] Un compartiment de stockage S3 d'AWS est un répertoire dans l'infonuagique d'Amazon où on peut stocker des fichiers

de type 2 (DT2) à haut risque de maladies vasculaires. Dans le cadre de cette étude effectuée en Angleterre, 11140 patients ont été recrutés parmi 215 centres dans 20 pays de l'Europe, de l'Australie, de l'Asie et de l'Amérique de nord. Les patients sélectionnés étaient âgés de 55 ans ou plus et ont été diagnostiqués pour le diabète de type 2 après l'âge de 30 ans. Un sous-groupe génotypé de 4098 patients d'origine caucasiens et atteints du DT2, provenant de la cohorte ADVANCE, a été analysé. Les phénotypes de base utilisés dans cette étude incluent l'âge, le sexe, le eGFR (estimated Glomerular Filtration Rate) et le génotype allèle à risque. Toutes les données cliniques et démographiques du patient étaient disponibles dans une base de données PostgresSQL, qui contenait 1,75 GO de données. Les données génétiques pour cette étude ont été générées en utilisant les variantes génétiques humaines sur Affymetrix Genome-Wide Human (à savoir les marqueurs SNP, Single Nucleotid Polymorphism), les « arrays » v5 et v6 et Affymetrix Axiom UK biobank (Affymetrix, Santa Clara, California, É.-U.) conformément aux protocoles standards recommandés par le fabricant. Une étape de filtrage dans le processus de contrôle de qualité a été appliquée aux génotypes obtenus (Hamet et coll. 2017). Trois imputations ont été effectuées séparément, la première sur les données génétiques générées avec le Genome-*Wide Human SNP Array 5.0,* la deuxième par celui du « array » 6.0 et la troisième par ceux générés par le *Affymetrix Axiom UK biobank.* Seuls les SNP ayant un ratio de qualité supérieure ou égale à 80% ont été retenus pour l'analyse (Patel, Chalmers, & Poulter, 2005). En raison de l'imputation, il n'y avait aucun génotype manquant dans la série de données. En conséquence, trois groupes de fichiers ont été créés pour représenter ces 4098 patients. Le Tableau 4.6 suivant montre la taille de chaque groupe de fichiers de la cohorte d'ADVANCE.

Tableau 4.6 Données utilisées dans l'analyse issue de la Cohorte d'ADVANCE

Affymetrix's GeneChip arrays	**Patients génotypés**	**Taille du fichier de génotype**
Genome-Wide Human SNP arrays 6.0	2394	101.578 GOS
Genome-Wide Human SNP arrays 5.0	1015	45 GOS
UK BioBank Axiom arrays	1294	59 GOS
Total	4703	205.578 GOS

4.6.2 Étape 2 : Créer un modèle des données spécifique pour l'analyse de données de la nouvelle recherche

Lors de cette étape, la dernière version du modèle des données de ADAM (adam-distribution-spark2_2.11-0.22.0) a été modifiée pour y inclure les définitions de toutes les données nécessaires à cette étude. Le principal objectif du prototype logiciel est de pouvoir répondre à n'importe quel besoin informationnel requis pour les analyses de données de médecine de précision. La composante d'intégration des données à grande échelle du prototype logiciel a été utilisée pour adapter le modèle des données ADAM existant et le bonifier avec tous les besoins informationnels requis par l'analyse de l'étude de cas. Nous avons vu que le but est de réunir toutes les données nécessaires à la recherche en un seul endroit et sous un même format efficace de stockage pour garantir l'efficacité de l'analyse. Le modèle des données de l'étude de cas est composé de trois sous-schémas de données: le premier sera utilisé pour le stockage des données génotypiques, le deuxième pour le stockage des données cliniques et le troisième pour le stockage des données d'analyse de la recherche.

Le modèle des données utilisé dans l'étude de cas est présenté dans la Figure 4.3 ci-dessous. Il est composé de trois sous-schémas : le premier est le modèle des données ADAM d'origine, il est présenté tel quel dans la partie haute du modèle. Les deux autres sous schémas présentés en bas du modèle, sont les deux sous-schémas ajoutés au modèle de données pour l'adapter aux requis de l'analyse de l'étude de cas. Elles comprennent deux sous-schémas : un sous-schéma des données cliniques et un autre pour les données d'analyse.

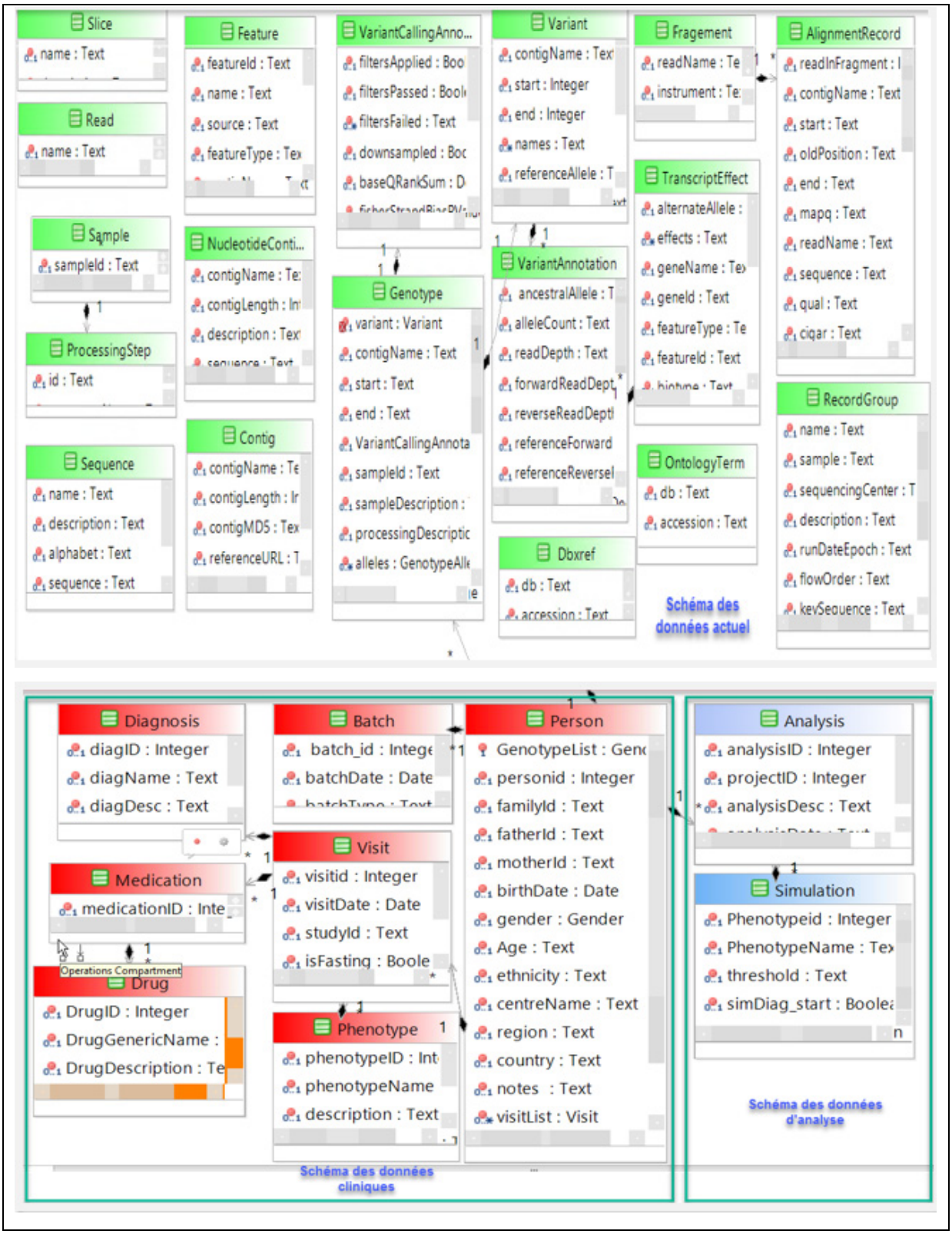


Figure 4.3 Adaptation de modèle des données ADAM

Le modèle des données à utiliser dans l'étude de cas sera donc composé par les trois sous-schémas de base de données suivants :

1. **Sous-schéma des données génétiques :** le schéma génétique du projet ADAM de l'université Berkeley (Massie et coll. 2013) offre un modèle spécialisé pour le stockage et l'accès efficace aux données de génotypage. Il est accompagné d'un pipeline de logiciels libres pour le traitement des données génétiques qui utilisent une architecture distribuée très performante sur les services infonuagiques. Il utilise Spark (Zaharia et coll. 2016), un moteur de traitement de données parallèle pour accéder rapidement aux données massives (Massie et coll. 2013). Afin de permettre la compatibilité avec les APIs du cadriciel ADAM et la solution proposée par cette thèse, la composante d'adaptation de modèle des données utilisera toujours la dernière version du modèle des données ADAM disponibles librement, ce qui permettra aux chercheurs de l'adapter à leurs besoins spécifiques.
2. **Sous-schéma des données cliniques :** pour cette étude de cas, le modèle des données présenté à la Figure 4.3 fournit aux chercheurs en médecine de précision des entités de données contenant les données démographiques et cliniques du patient. Pour le chercheur en médecine de précision, ceci est particulièrement utile pour les analyses croisées et l'exploration de ces données en simultané avec les données génétiques des patients. Typiquement, le chercheur désire pouvoir analyser à l'aide des cinq catégories de données suivantes : historique personnel, historique des visites et des diagnostics, phénotypes et traitement médical. Pour l'étude de cas, le modèle utilise des données issues de l'essai clinique ADVANCE, une cohorte de cas cliniques (Heller, 2009), (Patel et coll., 2005). L'essai a impliqué 215 centres collaborateurs répartis sur 20 pays et a été développé pour répartir en quatre groupes de traitement 11140 patients de façon aléatoire, hommes et femmes, tout âgées de 55 ans et plus, et diagnostiquées avec le diabète de type 2:

    1) groupe avec une diminution intensive de la glycémie et une baisse de la tension artérielle (avec une combinaison de perindopril /indapamide);

2) groupes ayant eu un traitement standard avec glucose et baisse de la tension artérielle;
3) groupes avec une diminution intensive de la glycémie et placébo;
4) un groupe ayant eu un traitement standard avec glucose et placébo.

Dans l'ensemble, le modèle des données cliniques, qui a été adapté ici, a pour but de fournir aux chercheurs en médecine de précision les données nécessaires pour conduire leur analyse à grande échelle en utilisant une base de données NoSQL normalisée pouvant être exploitée facilement à partir des technologies des données massives.

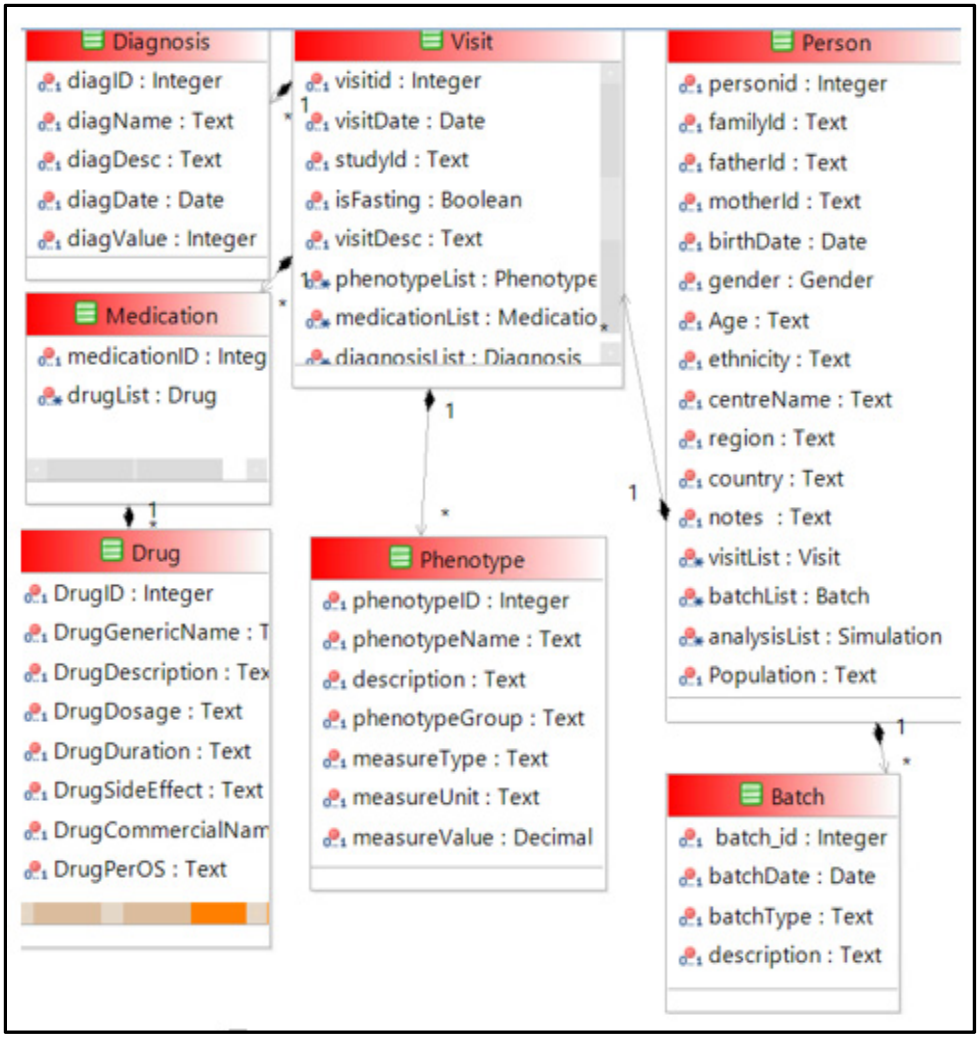


Figure 4.4 Modèle des données cliniques

Le sous-schéma contenant les données cliniques et les données démographiques est constitué des entités suivantes :

- **Historique personnel:** dans cette étude de cas, les données de l'historique clinique individuel étaient nécessaires, car elles servaient d'index principal pour toutes les informations du patient. L'historique médical est directement sauvegardé dans cette classe. Une référence à cette dernière sera contenue dans la classe « personne ». Des exemples de types de données sauvegardées dans cette catégorie seront présentés dans cette classe :

  - Données démographiques comme l'âge, l'ethnicité, la catégorie de population et des informations concernant la localité, le pays et la région d'origine du patient;
  - Une référence aux données génétiques et cliniques du patient ainsi qu'une référence à toutes les analyses des données disponibles.
- **Historique des visites :** la classe « historique des visites » contient toutes les informations permettant de garder une trace de chaque visite médicale durant l'expérimentation clinique ainsi que toutes les informations cliniques importantes recueillies pendant ces visites comme, par exemple, les phénotypes identifiés, la liste des diagnostics, les traitements prescrits et une liste de médicaments;
- **Historique des diagnostics :** la classe « historique des diagnostics » contient les informations les plus pertinentes, issues de chaque diagnostic du patient telles que le nom, la description du diagnostic et une valeur binaire pour indiquer si le patient a été diagnostiqué positivement ou négativement pour un problème de santé spécifique;
- **Phénotype :** cette classe contient les résultats de tests tels que des analyses de sang, d'urine et autres tests réalisés pour diagnostiquer des problèmes de santé. Les médecins vont souvent se servir de seuils établis pour définir des problèmes spécifiques. Avoir accès à cette information permettra de rechercher différents diagnostics en se basant sur les valeurs variables des seuils et de faire une analyse plus précise de ces données en les associant avec les données génétiques du patient;
- **Traitement médical** : cette classe a été utilisée afin de stocker les informations sur les médicaments, qui ont été prescrits aux patients. Elle inclut le nom commercial et générique des médicaments, la durée des traitements, le dosage et les effets secondaires vécus par le patient. Elle peut être combinée aux données génétiques et démographiques des patients pour permettre de proposer une médecine personnalisée.

3. **Sous-schéma pour l'analyse des données de la recherche :** La Figure 4.5 suivante illustre le sous-schéma d'analyse des données de chaque cycle de recherche.

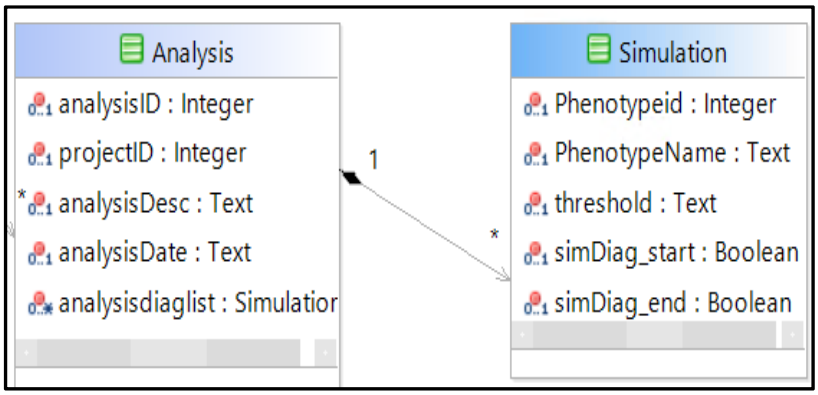


Figure 4.5 sous-schémas d'analyse des données de recherche.

Le sous-schéma contenant les données d'analyse est constitué des entités suivantes :

- **Analyse :** cette classe contient des informations générales qui identifient le projet sur lequel porte l'analyse ainsi qu'une référence à toutes les simulations (exécutions répétées de l'algorithme avec différents paramètres et différents seuils pour les paramètres) qui ont été réalisées pour chaque analyse;
- **Simulation :** pour chaque analyse, les chercheurs font plusieurs simulations sur chaque phénotype avec différents seuils et comparent les résultats. Cette catégorie sauvegarde toutes les informations qui identifient chaque simulation, le phénotype analysé avec le seuil de mesure utilisé. Cette information sera utilisée pour ajuster les valeurs des seuils critiques des données de tests et exécuter à nouveau les mêmes simulations avec différentes séries de données.

4.6.3 Étape 3 : Mise en place de l'environnement d'intégration des données

Une fois que le modèle des données a été adapté, l'infrastructure a été configurée pour pouvoir transférer les données à partir des sources de données vers l'infrastructure du nouveau logiciel. Au lieu d'utiliser un serveur virtuel configuré venant du fournisseur des services d'infonuagique (dans ce cas, AWS), des serveurs virtuels sans aucun logiciel installé à l'exception du système d'exploitation ont été utilisés et configurés pour créer la grappe de serveurs virtuels afin de réduire les coûts. Cette composante (API 3, Figure 3.10) va

automatiquement configurer la grappe avec les logiciels requis pour lancer le chargeur de données. Ce script va automatiquement configurer le nœud maître et tous les nœuds esclaves avec les logiciels requis pour l'adaptation de modèle et le chargement des données : Hadoop, Spark, Postgres; Apache Avro et Python. Une description des quatre étapes pour la mise en place du processus d'intégration des données est détaillée ci-dessous :

1. Spécifiez l'emplacement d'origine des données et la destination : dans un éditeur de texte, ouvrez le script settings.py et entrez l'emplacement d'origine et la base de destination des données cliniques et génétiques;
2. Préparez (démarrez et configurez) un serveur virtuel ou bien un ordinateur de bureau pour l'utiliser dans la gestion de la grappe de serveurs virtuels;
3. Créez et configurez les serveurs virtuels de la grappe : cette étape permet de créer et configurer les serveurs virtuels de la grappe à utiliser pour la préparation des données (adaptation de modèle des données, importation des données à partir de leurs emplacements d'origines, intégrations de toutes les données importées dans une seule base de données).

Les étapes 3 et 4 sont automatisées et sont exécutées via une interface visuelle. La Figure 4.6 suivante montre les options du menu de cette interface.

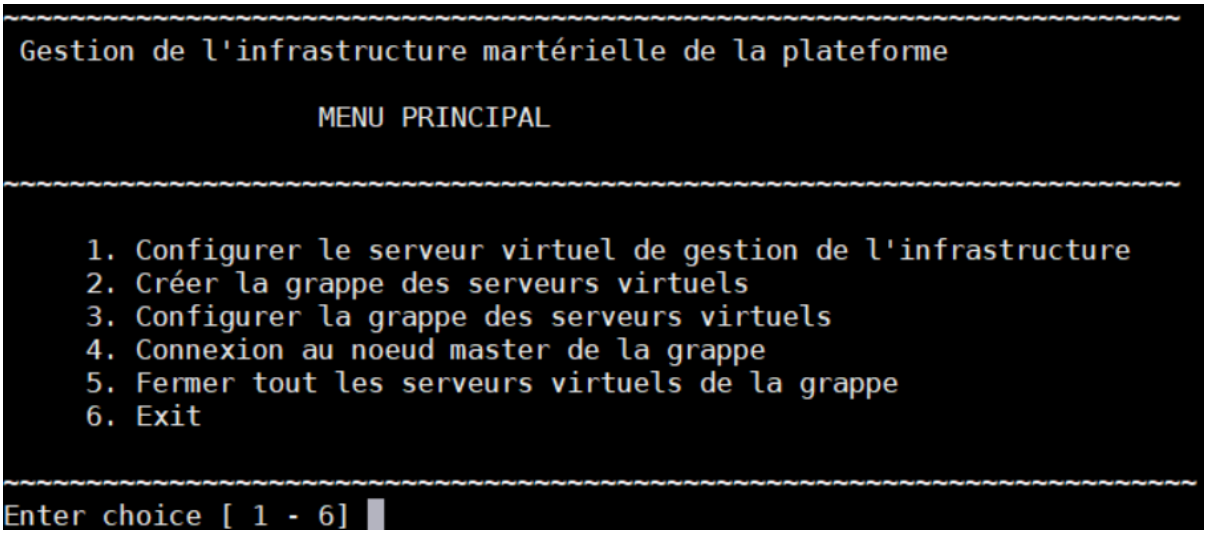


Figure 4.6 Gestion de l'infrastructure matérielle de la plateforme: Menu principal

5.6.4 Étape 4 : Intégration des données

Cette composante permet le transfert des données initiales vers le nouveau modèle des données adapté spécifique aux besoins de la recherche. Deux API pour le transfert des données ont été conçues et développées: une API de transfert des données génétiques qui va convertir et transférer les données brutes du format de fichier génotypique d'Oxford vers le format de fichiers binaires Parquet et une autre API pour transférer les données cliniques qui utilise à l'entrée le modèle de la base de données relationnelle (c.-à-d. dans cette étude de cas : la base de données Postgres) et convertit les données dans le format de fichier binaire Parquet.

Dans l'étude de cas, l'API d'intégration de données à grande échelle a été exécutée pour intégrer les données cliniques et génétiques dans la base de données créer à l'étape 2 précédente (c.-à-d. les importer à partir de leurs emplacements d'origines et les intégrer dans la base de données). Toutes les données nécessaires à la recherche (c.-à-d. données génétiques, cliniques, environnementales et démographiques) sont disponibles en un seul endroit (c.-à-d. colocalisées), sur un dépôt de données S3 de AWS. Elles sont enregistrées dans le format Parquet très efficace pour le traitement des données massives.

5.6.5 Étape 5 : Ajustement de la capacité de traitement de la plateforme d'intégration des données

Au cas où la configuration de la grappe d' intégration des données ne permettrait pas de traiter le très grand nombre de données à transférer, le chercheur peut stopper le processus d'intégration des données (s'il observe qu'il prend beaucoup trop de temps), supprimer toute donnée transférée, libérer les serveurs virtuels de la grappe à partir de la console EC2 AWS et configurer une nouvelle grappe comportant plus de serveurs virtuels en suivant les étapes décrites à l'Étape 3 : mise en place d'une grappe de serveurs virtuels pour la préparation des données.

Les APIs d'adaptation de modèle des données et d'intégration des données sont complètement automatisées et sont accessibles via l'interface utilisateur illustrée dans la Figure 4.7 suivante :

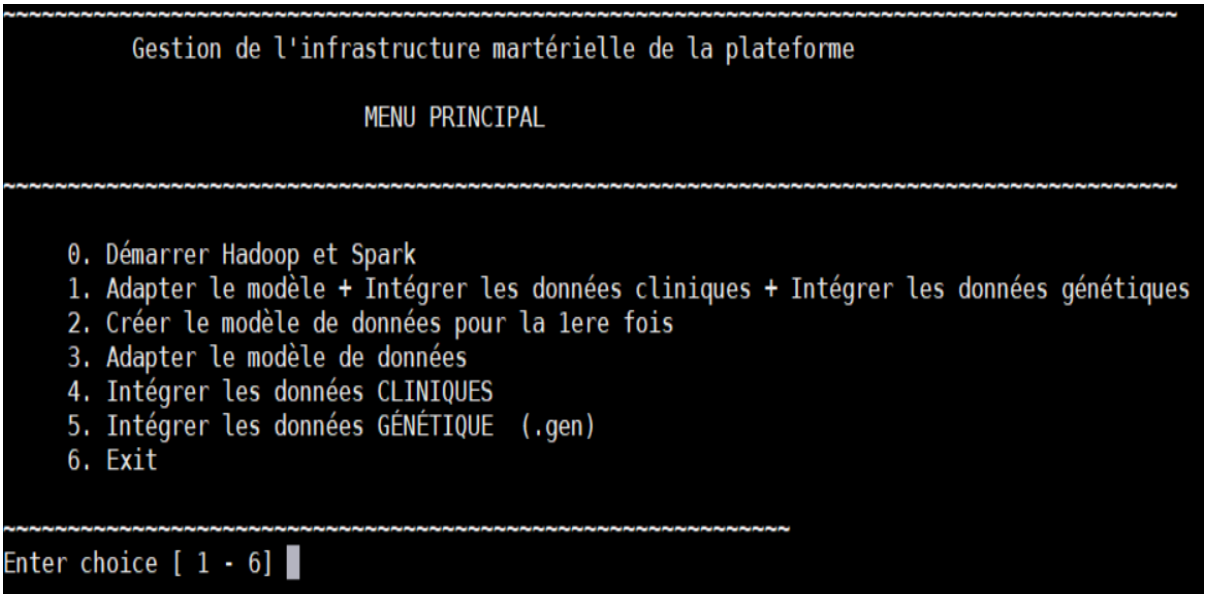


Figure 4.7 Configurateur de modèle des données et intégration des données: Menu principal

4.6.6 Étape 6 : Mise en place de l'environnement de l'analyse des données

Afin de préparer l'analyse des données, une grappe de serveurs virtuels doit être configurée et préparée. Cette étape est complètement automatisée. Les étapes suivantes ont été exécutées dans l'ordre pour créer et configurer la grappe de serveurs virtuels à utiliser pour l'exploration et l'analyse des données:

- À Partir de l'interface utilisateur d'AWS, un serveur virtuel Unix (Centos) non configuré a été démarré;
- Deux fichiers ont été copiés manuellement sur ce serveur virtuel. Un fichier de type ".pem" qui contient la clé d'accès aux serveurs virtuels de la grappe, et un script de configuration « sparkClustMgmt.sh »;

- L'exécution du script « sparkClustMgmt.sh » a permis de configurer le serveur virtuel de travail[3] et télécharger tous les scripts nécessaires à la configuration et à la gestion de grappe de serveurs virtuels d'analyse de données;
- Le script « sparkcl.sh », est un script qui permet de créer et de gérer une grappe de serveurs virtuels pour l'analyse des données. Les figures 20 et 21 suivantes respectivement illustrent les menus de création et de gestion de la grappe. Via ces deux interfaces utilisateur, le chercheur a le plein contrôle sur la gestion de sa plateforme d'analyse.

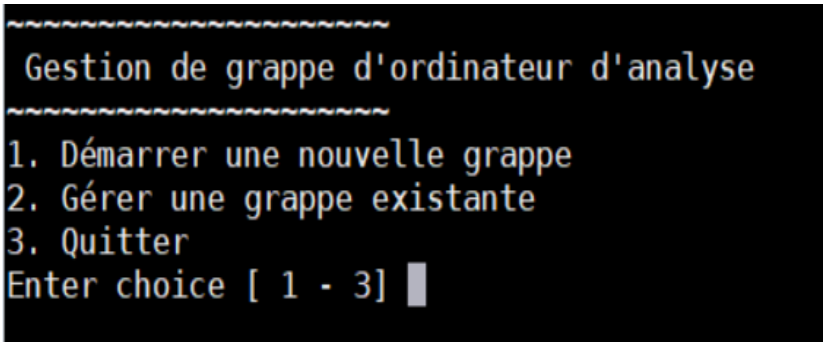


Figure 4.8 API de création de grappes de serveurs virtuels d'analyse

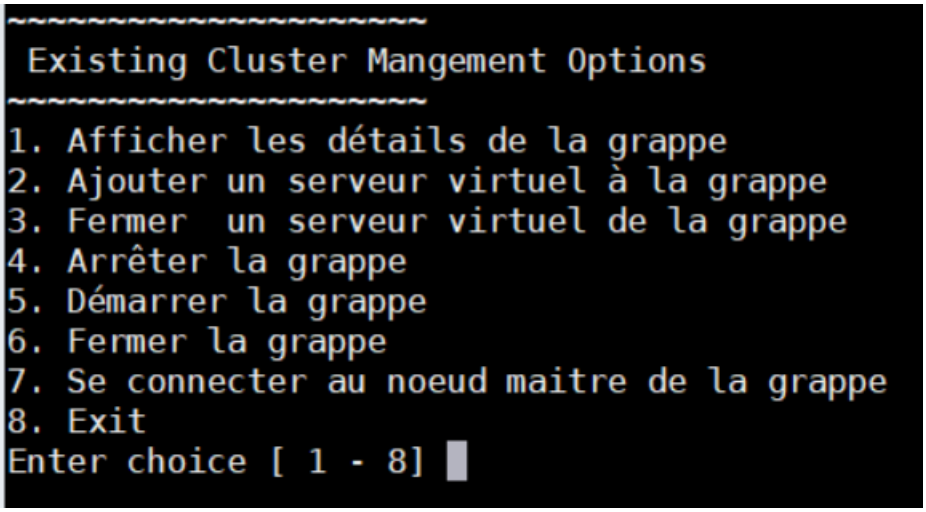


Figure 4.9 API de gestion de grappe de serveurs virtuels d'analyse de données

[3] Un serveur virtuel de travail est un serveur virtuel crée dans l'infonuagique de AWS pour être utilisé comme un ordinateur de travail pour la gestion de la grappe.

4.6.7 Étape 7 : Analyse des données

L'analyse des données est de loin l'activité la plus importante d'une étude de médecine de précision. Pour faciliter l'étape d'analyse, une composante (c.-à-d. fonctionnalité) du prototype logiciel permet l'exécution d'un script, facilement adaptable, qui fait le nettoyage des données qui vont servir à créer des scénarios d'analyse et à exécuter l'analyse de données. Le sous-schéma d'analyse et le chargeur de données à grande échelle vont, ensemble, assurer la réplicabilité de l'analyse. Pour simplifier cette tâche, le chercheur peut stocker les résultats intermédiaires dans le même emplacement S3 d'AWS avec les données transférées initialement. À l'aide du script de l'analyse des données sauvegardées et les données de résultats intermédiaires immédiatement disponibles par le logiciel de médecine de précision, une analyse pourra être répétée plus facilement.

Dans cette étude de cas, après l'intégration des données effectuées avec succès, l'API de configuration de la grappe de serveurs virtuels d'analyse des données a été utilisé pour configurer la grappe de serveurs virtuels Spark (qui est une grappe échelonnable) prête pour la phase d'analyse des données. Nous avons vu que les données ont été transférées telles quelles à partir des sources de données dans la base de données du prototype expérimental du logiciel. Lors de cette phase, ces données doivent maintenant être maîtrisées, préparées (c.-à-d. nettoyées et formatées) et ensuite analysées, par exemple dans notre étude de cas à l'aide de trois algorithmes de prédiction (réseau de neurones, régression logistique et celui de forêts aléatoires) pour tenter de découvrir un modèle prédictif pour une maladie. Mais avant de pouvoir effectuer ces analyses, regardons les étapes de collecte et de compréhension des données effectuées plus en détail.

L'étape d'analyse est composée de sept activités, la Figure 4.10 suivante montre les sept activités qui composent cette étape.

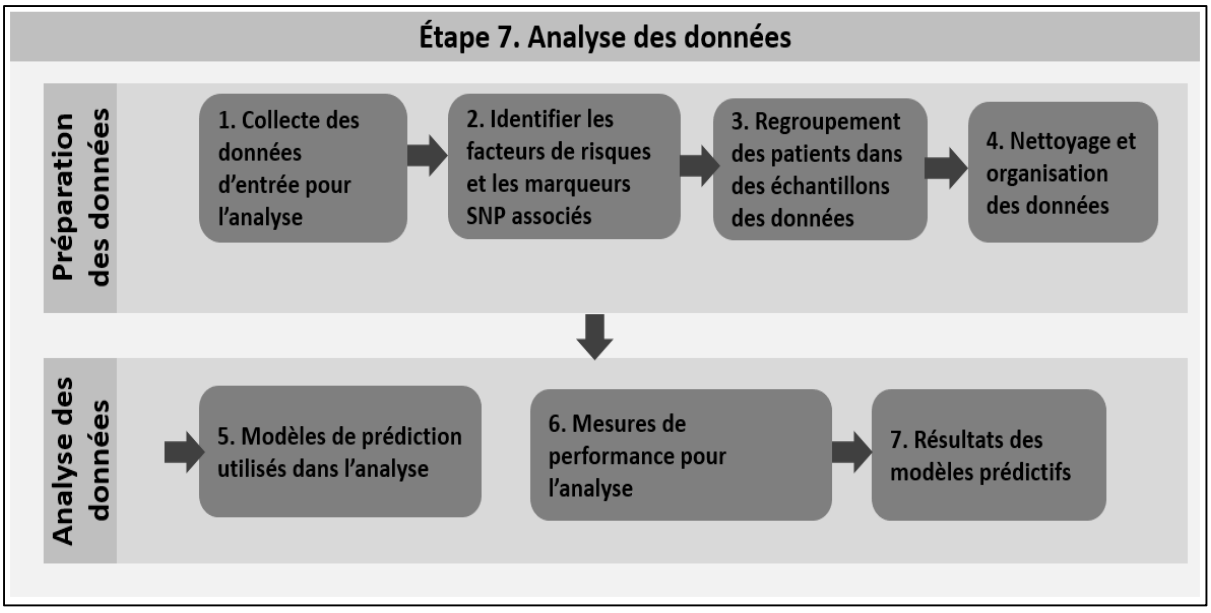


Figure 4.10 Étape 7 du cycle de recherche : Analyse des données

La description détaillée de chaque activité est présentée dans les sections suivantes.

4.6.7.1 Collecte et description des données d'entrée pour une recherche typique

Les étapes de collecte et de compréhension des données requièrent une grande partie des efforts du chercheur en médecine de précision. Dans notre étude de cas, les données de la cohorte ADVANCE ont été fournies sous deux formats :

- Les données génétiques étaient dans des fichiers plats. Ce sont des formats typiques produits par les puces de séquençage des données génétiques. Un premier fichier plat dans un format texte, à largeur fixe, contenant le numéro d'identification et le genre du patient (c.-à-d. féminin, masculin ou autre) a été utilisé. Ensuite 70 fichiers plats, utilisant un format texte à largeur variable ont été utilisés. Le format de ces 70 fichiers, présenté à la Figure 4.11 ci-dessous, est le suivant : chaque ligne contient les données génotypiques imputées pour un seul SNP. Les cinq premières entrées de chaque ligne contiennent: le numéro du chromosome, l'identifiant de la référence du SNP (c.-à-d. le RS ID), la position de la paire de bases du SNP, l'allèle codé A et l'allèle codé B. Les trois nombres suivants sur la ligne représentent les probabilités des trois génotypes AA,

AB, et BB, au niveau du SNP pour le premier individu de cette cohorte. Les trois nombres suivants sont les probabilités des données génotypiques pour le deuxième individu dans la cohorte. Les trois prochains nombres représentent le troisième individu et ainsi de suite. L'ordre des individus dans le fichier des données génotypique correspond à l'ordre des individus dans le fichier échantillon appelé « Sample »;

- En plus des données génétiques, des données cliniques et démographiques étaient disponibles. Ces données étaient disponibles dans une base de données relationnelle « Postgres » contenant des données cliniques et sociodémographiques des 4098 patients génotypés.

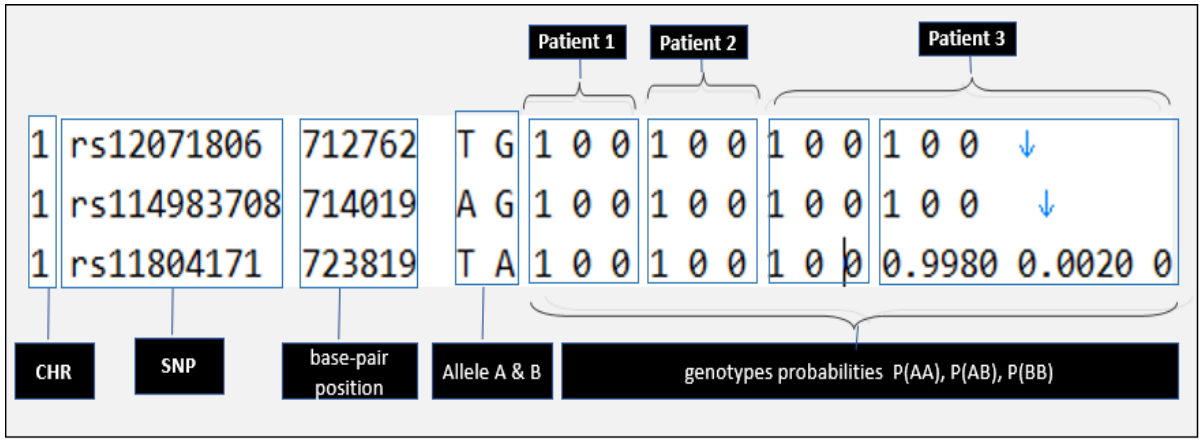


Figure 4.11 Formats de fichiers génotypiques

À cette étape, l'échantillon des données à utiliser dans l'analyse a été identifié, rassemblé et analysé pour bien s'assurer de comprendre la structure des données à traiter et à explorer, la prochaine étape est de définir et comprendre l'analyse à effectuer. La portée de l'analyse à réaliser est de développer un modèle prédictif qui permettrait d'identifier les patients à risque de développer la maladie d'insuffisances rénales chronique. Afin de développer le modèle prédictif, il est nécessaire d'identifier les paramètres de prédiction. Pour notre étude de cas, ces paramètres vont être identifiés et expliqués dans les deux sections suivantes (c.-à-d. 4.6.7.2 et 4.6.7.3).

4.6.7.2 Identifier les facteurs de risques et les marqueurs SNP associés

L'analyse de médecine de précision de notre étude de cas porte sur l'étude de l'albuminurie et la baisse du débit de filtration glomérulaire qui sont des manifestations de la néphropathie diabétique. Ces symptômes permettent de prédire une insuffisance rénale au stade terminal et une insuffisance rénale aigüe. Les chercheurs du laboratoire du Dr Hamet ont formé l'hypothèse que le SNP identifié en association avec le eGFR (estimation de la filtration glomérulaire) serait aussi associé à l'albuminurie (Ellis et coll. 2012). En se basant sur une analyse d'ouvrages spécialisés, ces chercheurs ont identifié 76 SNP associés au déclin des ratios eGFR ou celui de UACR (de l'anglais : Urine Albumin to Creatinine Ratio (UACR))(Ibrahim-Verbaas et coll. 2014) (Ellis et coll. 2012). Ces deux marqueurs de détérioration de la fonction rénale vont être utilisés dans la construction du modèle prédictif pour cette analyse de médecine de précision.

4.6.7.3 Regroupement des patients dans des échantillons des données

Les maladies du rein sont habituellement détectées chez les personnes diabétiques; 50 % des patients diabétiques montrent des signes de lésions ou d'atteintes rénales au cours de leur vie (Warram et coll. 1996), (Reenders et coll. 1993). Selon les chercheurs du domaine, c'est la première cause de maladie du rein au Canada (Warram et coll. 1996). La maladie rénale chronique (MRC) peut être observée dans une variété de conditions, y compris le diabète et l'hypertension. Mesurer le débit de filtration glomérulaire directement est la façon la plus précise de détecter des changements dans les conditions du rein. Le eGFR est un calcul basé sur un test de créatinine sérique. La créatinine est un déchet des muscles qui est transporté par le sang et filtré par les reins et est relâché dans les urines à un taux relativement constant. Quand la fonction rénale baisse, moins de créatinine est éliminée et sa concentration dans le sang augmente. Avec le test de créatinine, on obtient une estimation raisonnable du eGFR réel. Les patients vont être regroupés en fonction de la valeur de leur taux d'eGFR, une valeur de 60 ml/min/1.73m2 ou plus représente la concentration normale; tout patient avec un eGFR=60 ou en dessous pendant au moins trois mois (ces mesures se trouvent dans la base de données cliniques) sera considéré comme un cas de MRC.

Une fois que les données des patients et les paramètres de prédiction à utiliser pour cette analyse de médecine de précision ont été définis et compris, l'étape suivante est de préparer ces données pour l'exécution de l'analyse.

5.6.7.3 Nettoyage et organisation des données

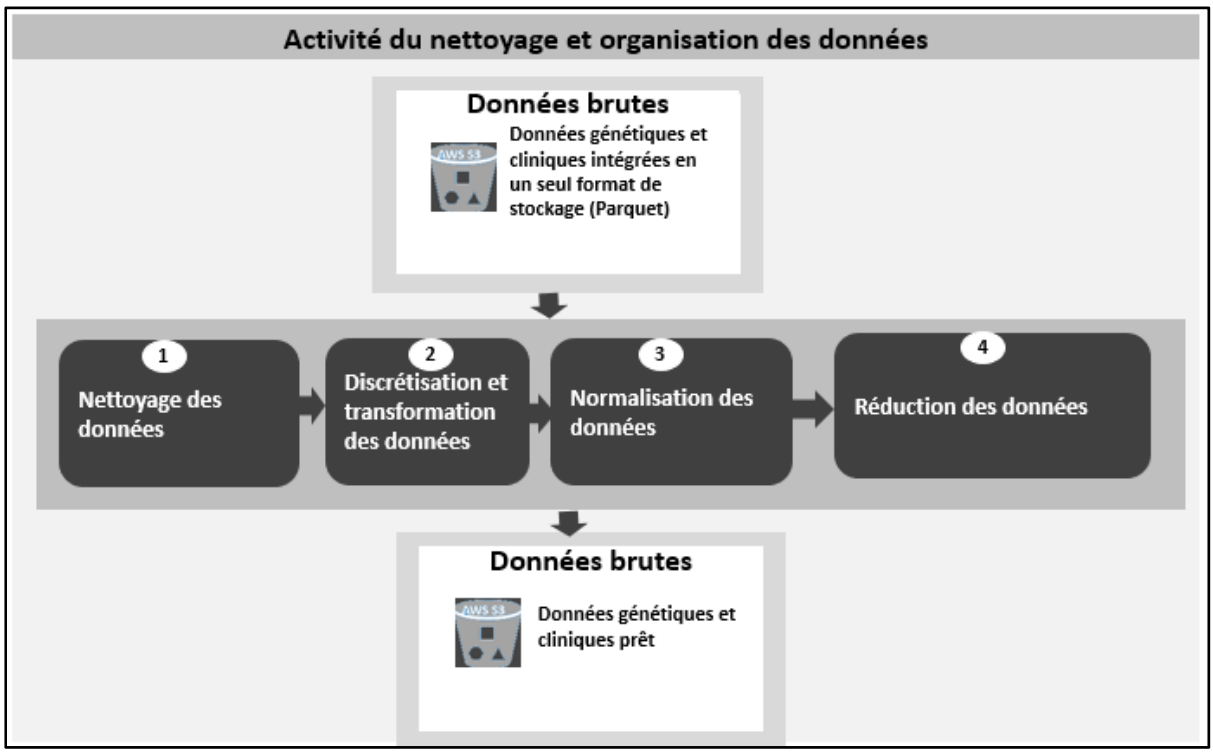


Figure 4.12 Activités de nettoyage et organisation des données

La Figure 4.12 ci-dessus montre les étapes qui composent l'activité de nettoyage et d'organisation des données. Cette activité a pour but de s'assurer que la qualité des données est suffisante pour exécuter des algorithmes d'intelligence artificielle (c.-à-d. l'apprentissage profond). Tout d'abord, les données collectées ont dû être extraites de la base de données clinique relationnelle (c.-à-d. une base de données Postgres) et les nombreux fichiers plats des données génotypiques ont dû être intégrés dans le modèle des données ADAM adapté. Elle est composée de quatre taches: la première est la tâche de nettoyage des données, elle consiste à enlever toutes les données erronées, la deuxième est celle de la discrétisation et transformation des données, elle consiste à convertir les données au format de type booléen en format binaire,

la troisième est la normalisation des données, elle consiste à préparer les données de sorte qu'elles puissent être utilisées par les trois algorithmes prévus (c.-à-d. normaliser les colonnes: âge, sexe et région) et la quatrième tâche de réductions des données, elle consiste à réduire les colonnes d'analyse de données vides et inutiles. Au terme de cette étape, toutes ces données sont maintenant réorganisées dans un seul tableau composé de 112 colonnes dont 76 colonnes étaient utilisées pour les données génétiques et 36 colonnes représentaient les données cliniques (c.-à-d. âge, sexe, région) donnant une taille de 1118 lignes (c.-à-d. une ligne par patient).

Ce tableau a été créé à partir de trois ensembles de données: un groupe de données cliniques (2394 patients), un groupe de données génétiques (15213486960 lignes), et la liste des SNP associés aux groupes à risque eGFR et UACR, et était composé de 5 colonnes et de 76 lignes. Le tableau contenait deux champs de calcul. Le génotype allèle à risque (RAG) se réfère aux valeurs calculées pour chaque SNP pour évaluer l'impact du SNP sur la maladie. Ces valeurs ont été obtenues (voir Algorithme 4.1 ci-dessus) en utilisant les probabilités de l'allèle de référence et l'allèle alterne avec la colonne allèle à risque pour chaque SNP.

Algorithme 4.1 Calcul du risque associé au Génotype (RAG)

Algorithme 1:

Si *risk allele = reference allele (RA)*

Alors *RiskAlleleGenotype = 2 * P(RA, RA) + P(RA, AA)*

Sinon *RiskAlleleGenotype = 2 * P(AA, AA) + P(RA, AA)*

Où P(RA,RA), P(RA,AA), et P(AA,AA) représentent respectivement les probabilités de l'allèle de référence (ou sauvage) et des allèles mutants

Le diagnostic produit une valeur booléenne qui va être utilisée pour l'étiquette de classification. Elle a été calculée (voir Algorithme 4.2 ci-dessous) à partir du eGFR provenant de la base de données clinique.

Algorithme 4.2 Logique d'identification des patients atteints du MRC

Algorithme 2:

Si *eGFR <= 60ml/min/1.73m2*

Alors *CKD = 1*

Sinon *CKD = 0*

Après avoir préparé les données, la section suivante va présenter les différents algorithmes de prédiction à évaluer pour développer le modèle de prédiction.

4.6.7.4 Modèles de prédiction utilisés dans l'analyse

Dans la phase d'analyse de l'étude de cas, trois modèles de classification ont été testés. Ces modèles ont été sélectionnés à cause de leur popularité dans les publications récentes au sujet de la sélection génétique (Kourou et coll. 2015) (Endo, Shibata, & Tanaka, 2008). Pour tester ces trois algorithmes, un script en utilisant les deux librairies « pyspark.ml.classification » et « pyspark.ml.evaluation » a été développé et exécuté. Ce script est sauvegardé avec les données intégrées dans le compartiment de stockage S3 pour permettre la reproductibilité de l'analyse.

1. **Algorithme Forêt aléatoire (Random Forest) :** cet algorithme pour la régression et la classification a été introduit pour la première fois en 2001 et n'a cessé de gagner en popularité depuis. C'est devenu l'une des méthodes de classification les plus utilisées, faisant compétition avec la régression logistique dans ces domaines. Il est communément utilisé en bio-informatique en raison de la précision de ses prédictions et la capacité d'interprétation de son modèle (Chen & Ishwaran, 2013). C'est un algorithme d'arbre de décision basé sur l'apprentissage automatique, qui est très adaptable aux problèmes du forage des données (ou « data mining ») « grand p » et « petit n » (Díaz-Uriarte & Alvarez de Andrés, 2006).
2. **Réseau de neurones (Neural Network) :** le réseau de neurones implique un type de réseau multicouche linéaire (ou à direction unique) (Stastny & Skorpil, 2007). Il s'inspire du réseau de neurones biologiques dans le cerveau humain en connectant les variables

prédites avec les variables explicatives, à partir de couches de nœuds[4] (cachés) intermédiaires. Dans cette analyse, deux couches d'entrées ont été utilisées. La couche d'entrée représente le nombre de caractéristiques c'est-à-dire le nombre de SNP + 2 (sexe et âge) pour diagnostiquer le patient (CKD = 1, Non-CKD = 0).

3. **Régression logistique (Logistic Regression) :** il s'agit d'une généralisation, de la régression linéaire (Liao & Chin, 2007). Il est surtout utilisé pour prédire les variables dépendantes multi classes ou binaires. Il exige que les réponses variables soient discrètes. Il est très utilisé dans les applications biostatistiques, où les réponses sont fréquemment binaires (deux classes), par exemple : les patients survivent ou meurent, ils sont atteints d'une maladie cardiovasculaire ou non, une condition particulière est présente ou absente.

Afin de choisir l'algorithme le plus adapté à l'analyse, les mesures à utiliser pour comparer la performance de chaque algorithme seront présentées dans la section suivante.

4.6.7.5 Mesures de performance des algorithmes de prédiction

Dans cette étude, quatre mesures de performance ont été utilisées : accuracy (ou exactitude) (Eq 1), « précision » correspondant à la proportion de vrais positifs (classés correctement) sur un ensemble de données classées comme positives (Eq 4), recall (ou rappel), correspondant à la proportion de vrais positifs de l'algorithme de classification par rapport au nombre total de positifs (Eq 3) et en dernier le f1-score (F-mesure) correspondant à la moyenne harmonique entre les métriques « précision » et « rappel » pour définir un bon équilibre (Eq 4) où tp (VP), tn (VN), fp (FP) et fn (FN) désignent respectivement, « true positive » (vrai positifs), « true négatives » (vrai négatifs), « false positive » (faux positifs) et « false negatives » (faux négatifs).

[4] Un nœud est un neurone dans le réseau de neurone

Algorithme 4.3 Calcul des paramètres de mesure

$$(Eq\ 1):\ \boldsymbol{accuracy} = \frac{\boldsymbol{tp}\ +\ \boldsymbol{tn}}{\boldsymbol{tp}\ +\ \boldsymbol{tn}\ +\ \boldsymbol{fp}+\ \boldsymbol{fn}}$$

$$(Eq\ 2):\boldsymbol{precision}:\ if\ \text{tp}\ +\ \text{fp}\ ==\ 0:\text{precision}\ =\ 0$$

$$\text{then precision} = 0 \quad else\ \ precision$$

$$= \frac{\text{tp}}{\text{tp}\ +\ \text{fp}}$$

$$(Eq\ 3):\ \boldsymbol{recall}\ =\ \frac{\text{tp}}{\text{tp}\ +\text{fn}}$$

$$(Eq\ 4)\ :\boldsymbol{f1-score}:\ if\ \ \text{precision}\ +\ \text{recall}\ ==\ 0\ \text{then}\ \text{f1}_{\text{score}}$$

4.6.7.6 Résultats obtenus à l'aide des modèles prédictifs

Les modèles ont été exécutés et évalués à partir de la mesure « d'exactitude ». Les résultats du Tableau 4.7 ci-dessous ont été obtenus en utilisant les résultats de 10 exécutions répétées du modèle. L'échantillon de 1118 patients a été divisé en deux échantillons dont 75 % des données sont utilisées pour l'entraînement des modèles et 25 % des données pour les tester.

Même avec un échantillon de taille relativement petite, les modèles prédictifs ont donné de meilleurs résultats avec l'algorithme Forêt aléatoire (RF) qu'avec les algorithmes de Régression linéaire (LR) et Réseaux de neurones (NN). Il a été constaté que le modèle NN a atteint une valeur d'exactitude de classification de 56,15% avec une valeur de précision de 65,70%, une valeur de rappel de 59,28% et une mesure F1 62,28%. LR a produit de meilleurs résultats que NN, atteignant une valeur d'exactitude de 62,54% avec une valeur de précision de 69,71%, une valeur de rappel de 68,54% et une mesure F1 de 68,12%. Cependant, le modèle RF a atteint une valeur d'exactitude de 64,60% avec une valeur de précision de 68,66%, une valeur de rappel de 77,53% et une mesure F1 de 72,82%. Le Tableau 4.7 suivant présente les résultats complets sous forme tabulaire.

Tableau 4.7 Résultats tabulaires de la formation des modèles prédictifs

Simulation	Neural Network (NN)				Linear Regression (LR)				Random Forest (RF)			
	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
1	0,5430	0,6380	0,5843	0,6100	0,6254	0,6971	0,6854	0,6912	0,646	0,6866	0,7753	0,7282
2	0,5739	0,6688	0,6011	0,6331	0,6254	0,6971	0,6854	0,6912	0,646	0,6866	0,7753	0,7282
3	0,5704	0,6626	0,6067	0,6334	0,6254	0,6971	0,6854	0,6912	0,646	0,6866	0,7753	0,7282
4	0,5326	0,6364	0,5506	0,5904	0,6254	0,6971	0,6854	0,6912	0,646	0,6866	0,7753	0,7282
5	0,5704	0,6646	0,6011	0,6313	0,6254	0,6971	0,6854	0,6912	0,646	0,6866	0,7753	0,7282
6	0,5533	0,6500	0,5843	0,6154	0,6254	0,6971	0,6854	0,6912	0,646	0,6866	0,7753	0,7282
7	0,5670	0,6605	0,6011	0,6294	0,6254	0,6971	0,6854	0,6912	0,646	0,6866	0,7753	0,7282
8	0,5704	0,6626	0,6067	0,6334	0,6254	0,6971	0,6854	0,6912	0,646	0,6866	0,7753	0,7282
9	0,5567	0,6561	0,5787	0,6149	0,6254	0,6971	0,6854	0,6912	0,646	0,6866	0,7753	0,7282
10	0,5773	0,6708	0,6067	0,6372	0,6254	0,6971	0,6854	0,6912	0,646	0,6866	0,7753	0,7282
Moyenne	0,5615	0,6570	0,5921	0,6228	0,6254	0,6971	0,6854	0,6912	0,646	0,6866	0,7753	0,7282
accuracy = (tp + tn) / (tp + tn + fp + fn); précision = tp / (tp + fp); recall = tp / (tp + fn); f1_score = 2 * precision * recall / (precision + recall)												

Au cours de l'exécution de l'analyse, c.-à-d. durant l'étape de préparation des données et l'étape d'exécution des algorithmes, des problèmes de capacité de traitement ont été rencontrés. La section suivante explique comment ces problèmes ont été résolus.

4.6.8 Étape 8 : Ajustement de la capacité de traitement de la plateforme lors de l'analyse des données

À plusieurs reprises, pendant la phase d'analyse, la taille de la grappe de serveurs virtuels a dû être rééchelonnée pour la rendre apte à traiter un très grand volume de données plus rapidement. La vitesse avec laquelle la taille d'une grappe de serveurs virtuels pouvait être ajustée au cours de l'analyse a été testée à l'aide de l'API d'ajustement de grappe de serveurs virtuels d'analyse de données (voir Figure 4.9). Cet API offre un menu contextuel qui comporte sept fonctions pour gérer les grappes de serveurs virtuels AWS: afficher les détails de la grappe, ajouter un nouveau serveur à la grappe pour augmenter sa capacité de traitement, retirer des serveurs virtuels de la grappe, arrêter une grappe de serveurs virtuels et démarrer une

grappe de serveurs virtuels, fermer définitivement une grappe de serveurs virtuels et se connecter au nœud maître d'une grappe de serveurs virtuels.

4.7 Reproductibilité de l'analyse

À la suite de la complétion d'une recherche de médecine de précision (c.-à-d. un certain nombre d'itérations), les données utilisées (données brutes et données intégrées) et les scripts d'analyse seront archivés. Puis ensuite, la grappe de serveurs virtuels utilisée pour l'intégration et l'exploration des données sera fermée[5]. La Figure 4.13 suivante présente les actions principales effectuées quand l'analyse de l'étude de cas a été complètement terminée et celles effectuées pour la reproduction de l'analyse complétée.

La figure 25 suivante illustre le scénario de reproduction d'analyse antérieure à partir des données déjà intégrées. Dans l'étude de cas, la reproductibilité de l'analyse a été testée à partir des données initiales afin de revalider le processus au complet.

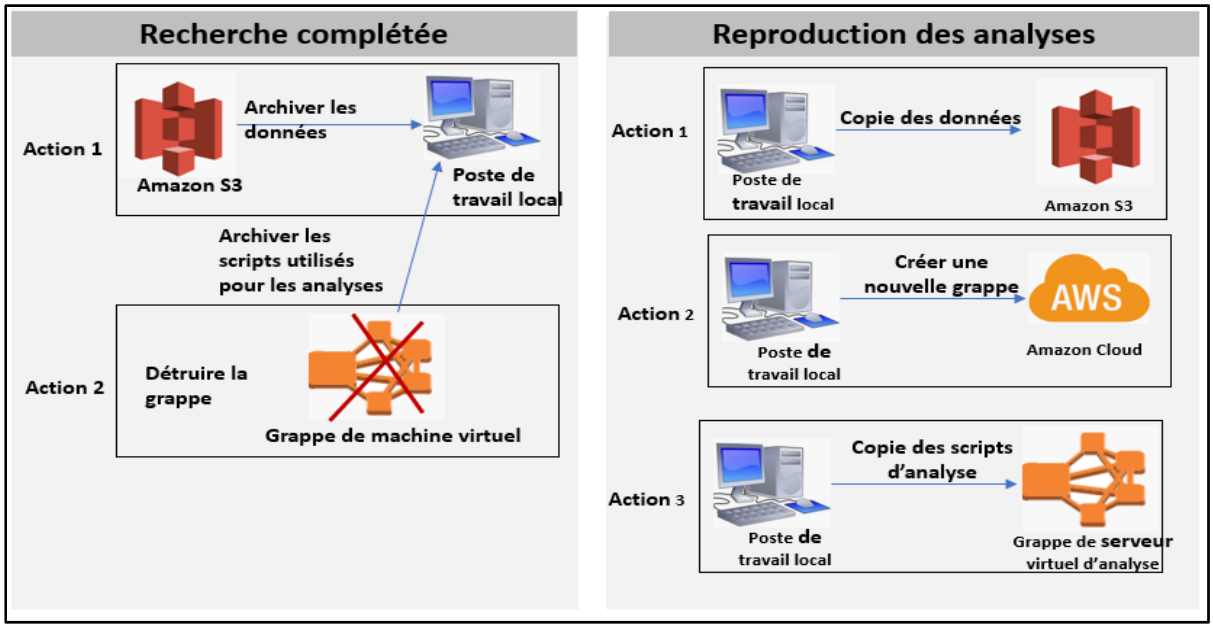


Figure 4.13 Étapes de reproductibilité d'analyses précédentes

[5] Dans le contexte d'infonuagique, une fois le travail est terminé, les serveurs virtuels sont fermés définitivement, puis recréer plus tard au besoin

Pour tester les scénarios de reproduction d'analyse dans l'étude de cas, après la complétion du processus d'analyse les étapes suivantes ont été exécutées:

- **Phase 1 : Actions de complétion de l'analyse de l'étude de cas**

  Après la complétion de l'analyse, les actions suivantes ont été faites :

  - Déplacer les données initiales (c.-à-d. les données brutes), celles intégrées et les scripts d'analyse utilisés (c.-à-d. le script de nettoyage, d'agrégation des données et le script du modèle prédictif) du service d'archivage S3 de AWS vers une base de données sur un poste de travail local;
  - Fermer la grappe de serveurs virtuels utilisés lors de l'intégration et l'analyse des données;
  - Fermer le serveur virtuel utilisé pour la gestion de la plateforme.

- **Phase 2 : Reproduction de l'analyse complétée**

  Les actions suivantes ont été faites afin de tester le processus de reproductibilité des analyses antérieures :

  - Une nouvelle base de données a été créée à l'aide du service d'archivage S3. Cette étape consistait à se connecter à l'interface utilisateur d'AWS et créer une base de données à l'aide du service S3;
  - Les données ont été ensuite déplacées du poste de travail local vers cette nouvelle base de données (c.-à-d. S3). Les données ont été copiées en utilisant un logiciel libre (c.-à-d. S3 Browser)(S3 Browser, 2019);
  - Créer et configurer un serveur virtuel de travail qui permet de démarrer l'API de gestion de la grappe de serveurs virtuels à utiliser pour la reproduction de l'analyse. Un serveur de type « t2.micro », gratuit, a été utilisé. La configuration du serveur de travail a été effectuée à l'aide de l'exécution d'un script automatisé;
  - Exécuter l'API de création et de configuration de la grappe de serveurs virtuels (l'*API 3 : Configurateur d'infrastructure matérielle*). L'API a été configuré pour créer 10 serveurs virtuels de type « m4.4xlarge»;
  - Exécuter l'API d'intégration des données;
  - Fermer la grappe des serveurs virtuels utilisée pour l'intégration des données;

- À l'aide d'un script automatisé, créer une grappe de serveurs virtuels d'analyse des données;
- Copier les scripts d'analyse des données (c.-à-d. les scripts de nettoyage, d'agrégation des données et celui du modèle prédictif) de la station de travail local vers le nœud maître de la grappe d'analyse;
- Exécuter les scripts d'analyse du modèle prédictif.

Tous les scripts utilisés sont des scripts développés lors de la réalisation de l'étude de cas. Aucun service d'AWS n'a été utilisé ni pour la réalisation de l'analyse ni pour la reproduction de l'analyse.

4.8 Présentation des résultats de l'étude de cas

Le but de cette étude de cas était de prouver la viabilité de l'approche proposée et de valider les avantages de la nouvelle approche d'adaptation de modèle et de traitement de données en s'appuyant sur une vraie étude de cas de recherche dans le domaine de la médecine de précision.

Cette étude a aussi permis de démontrer comment chaque caractéristique de la conception visée qui a été mise en œuvre et testée se comportait dans un cas d'étude réelle de médecine de précision.

Les résultats de cette étude de cas ont été analysés sous trois aspects:

1) Évaluer chaque étape du cycle de recherche de la nouvelle approche en utilisant le prototype expérimental du logiciel, le type des activités effectuées au cours de l'étape : sont-elles automatiques ou bien manuelles, à quel point sont-elles automatiques, le temps d'exécution de l'étape et les améliorations potentielles qu'on pourrait faire dans le futur sur le prototype logiciel?
2) Évaluer l'atteignabilité de l'objectif visé par la nouvelle approche. Pour chaque objectif, la solution proposée a été évaluée pour savoir à quel point cet objectif a été atteint;

3) Évaluer la faisabilité de l'étude de cas avec les autres logiciels d'analyse identifiés lors de la phase de la revue de littérature.

4.8.1 Résultats de la nouvelle approche du cycle de recherche

4.8.1.1 Présentation des résultats : nouvelle analyse

Cette section présente les résultats pour chaque étape du nouveau cycle de recherche décrit dans la Figure 3.5 ci-haut, un tableau avec les résultats obtenus, et les améliorations potentielles pour optimiser encore le processus.

Tableau 4.8 Évaluation de l'étape 1 du nouveau cycle de recherche

Étape 1	Spécification d'hypothèse(s) pour la recherche
Type	Manuelle, cette activité et hors de portée de la présente recherche.
Activités	Une activité exécutée par les chercheurs afin de définir les hypothèses de la recherche
Délai	Non Évalué,
Ressources humaines	Non évalué
Ressources matérielles	Non évalué
Améliorations futures	S/O

Tableau 4.9 Évaluation de l'étape 2 du nouveau cycle de recherche

Étape 2	Définition des besoins informationnels
Type	Manuelle, cette activité et hors de portée de la présente recherche.
Activités	Une activité exécuter par les chercheurs afin d'identifier : liste tous les besoins informationnels, sources des données, etc.
Délai	Non Évalué,
Ressources humaines	Non évalué
Ressources matérielles	Non évalué
Améliorations futures	S/O

Tableau 4.10 Évaluation de l'étape 3 du nouveau cycle de recherche

Étape 3	Création/adaptation de modèles de données spécifique pour l'analyse de données
Type	Semi-automatique : • L'adaptation de modèle des données. Elle consiste à manuellement ouvrir un fichier texte et ajouter les nouveaux besoins informationnels aux modèles existants; • La génération du modèle des données. Cette tâche est complètement automatisée.
Activités	Modifier le fichier de configuration AVRO dans un éditeur de texte pour l'adapter aux nouveaux requis de la recherche, puis l'API d'adaptation de modèle des données va automatiquement ajouter la définition des nouveaux besoins informationnels à la dernière version du modèle des données ADAM et créer un nouveau modèle de données.
Délai	Quinze minutes, incluant la modification manuelle du fichier de configuration AVRO et l'exécution de l'API d'adaptation de modèle des données.
Ressources humaines	1 personne
Ressources matérielles	1 serveur virtuel de type « t2.micro »
Améliorations futures	Créer une interface « utilisateur » pour configurer le processus d'adaptation de modèle des données.

Tableau 4.11 Évaluation de l'étape 4 du nouveau cycle de recherche

Étape 4	Mise en place de l'environnement d'intégration des données
Type	Automatique
Activités	Créer la grappe de serveurs virtuels en utilisant un script automatisé, puis lancer la commande de création et de configuration de la grappe à partir du menu contextuel.
Délai	15 minutes
Ressources humaines	1 personne
Ressources matérielles	1 serveur de type « t2.micro » offert gratuitement par AWS.
Améliorations futures	Créer une interface « utilisateur » facile à utiliser pour la création et la configuration de la grappe de serveurs virtuels d'intégration des données.

Tableau 4.12 Évaluation de l'étape 5 du nouveau cycle de recherche

Étape 5	Intégration des données
Type	Automatique
Activités	Exécuter les scripts d'intégration des données cliniques et ceux des données génétiques. (Échantillons de données: données d'environ 5000 patients (205 GOS).
Délai	• Intégration des données génétiques : 3 heures et 18 minutes; • Intégration des données cliniques : 1.4 minute.
Ressources humaines	1 personne
Ressources matérielles	10 serveurs virtuels de type m4.4xLarge
Améliorations futures	Ajouter des APIs d'intégration des données pour d'autres formats de fichiers des données génétiques. La version actuelle de l'API proposée ne supporte que le format de fichier Oxford « format .gen ».

Tableau 4.13 Évaluation de l'étape 6 du nouveau cycle de recherche

Étape 6	Ajustement de la capacité de traitement de la plateforme d'intégration des données
Type	Automatique
Activités	• Fermer la grappe d'intégration des données; • Créer une nouvelle grappe d'intégration en augmentant le nombre de serveurs virtuels ayant une capacité de traitement plus grande que celle utilisée précédemment.
Délai	• 5 minutes pour la fermeture de tous les serveurs virtuels de la grappe actuelle; • 15 minutes pour créer la nouvelle grappe.
Ressources humaines	1 personne
Ressources matérielles	10 serveurs virtuels de type m4.4xlarge
Améliorations futures	• Créer une API de gestion de grappe pour l'intégration des données et pour les analyses des données (fusionner l'API de gestion de la plateforme matériel pour l'intégration des données et celle pour l'analyse des données); • Crée une interface « utilisateur » conviviale pour la gestion de la plateforme matérielle d'intégration et d'analyse des données.

Tableau 4.14 Évaluation de l'étape 7 du nouveau cycle de recherche

Étape 7	Mise en place de l'environnement de l'analyse des données.
Type	Automatique
Activités	Exécuter un script automatique pour créer et configurer une grappe de serveurs virtuels pour l'analyse des données.
Délai	15 minutes
Ressources humaines	1 personne
Ressources matérielles	10 serveurs virtuels de type « m4.4xlarge »
Améliorations futures	• Créer un API pour gérer automatiquement les grappes de serveurs virtuels d'analyse des données indépendamment des technologies de l'infonuagique; • Crée une interface « utilisateur » conviviale pour la gestion de la grappe de serveurs virtuels de l'analyse des données.

Tableau 4.15 Évaluation de l'étape 8 du nouveau cycle de recherche

Étape 8	Analyse des données
Type	Automatique
Activités	Exécuter le script qui simule l'algorithme du modèle de prédiction.
Délai	• Environ 1 heure de configuration et d'exécution de l'algorithme de prédiction; • Note : ce délai n'inclut ni les efforts de programmation de l'algorithme ni les efforts d'analyse et interprétation des résultats de l'exécution.
Ressources humaines	1 personne
Ressources matérielles	10 serveurs virtuels de type « m4.4xlarge »
Améliorations futures	Ajouter des outils d'analyse de données sources libre dans l'API de configuration de la grappe de serveurs virtuels d'analyse.

Tableau 4.16 Évaluation de l'étape 9 du nouveau cycle de recherche

Étape 9	Ajustement de la capacité de traitement de la plateforme d'analyse des données.
Type	Automatique
Activités	Exécuter un script automatique qui utilise un outil source-libre pour gérer les serveurs virtuels Spark. Cet outil est spécifique pour les services infonuagiques de AWS.
Délai	• 50 secondes pour ajouter 1 serveur virtuel; • 3 minutes pour ajouter 5 serveurs virtuels à la grappe de serveurs virtuels.
Ressources humaines	1 personne
Ressources matérielles	10 serveurs virtuels de type « m4.4xlarge ».
Améliorations futures	• Créer un API pour gérer automatiquement la grappe d'analyse des données indépendamment des technologies de l'infonuagique; • Crée une interface « utilisateur » pour la gestion de la grappe de serveurs virtuels de l'analyse des données.

Le Tableau 4.17 suivant présente un sommaire des efforts dépensés pour réaliser une nouvelle analyse avec la nouvelle approche.

Tableau 4.17 Sommaire des efforts dépensés dans la réalisation d'une itération d'analyse

#	Étape	Délai	Nombre de personnes
1	Spécification d'hypothèse(s) pour la recherche	S/O	S/O
2	Définition des besoins informationnels	S/O	S/O
3	Création/adaptation de modèles de données spécifique pour l'analyse de données	15 min	1 personne
4	Mise en place de l'environnement d'intégration des données	15 min	1 personne
5	Intégration des données	3h18 min	1 personne à temps partiel. La personne est impliquée seulement dans le démarrage de l'API d'intégration des données.
6	Ajustement de la capacité de traitement de la plateforme d'intégration des données	20 min	1 personne
7	Mise en place de l'environnement de l'analyse des données	15 min	1 personne
8	Analyse des données	60 min	1 personne
9	Ajustement de la capacité de traitement de la plateforme d'analyse des données	4 min[6]	1 personne
Total		5H28min	

[6] Pour la grappe d'analyse, une grappe de six serveurs virtuels a été configurée initialement, puis, un serveur virtuel a été ajouté, puis ensuite, trois autres serveurs virtuels supplémentaires ont été ajoutés. Une grappe de dix serveurs virtuels a finalement été requise.

4.8.1.2 Présentation des résultats : reproduction d'analyses antérieures

Le Tableau 4.18 suivant illustre les étapes de reproduction d'une analyse antérieure. Le nombre de personnes requises pour la réalisation de tâches de l'analyse, et le délai encouru pour la complétion de chaque étape.

Tableau 4.18 Évaluation de l'étape de reproduction d'analyses existantes

#	Étape	Délai	Nombre de personnes
1	Créer une nouvelle base de données dans le service S3 en utilisant l'interface d'AWS	5 mins	1 personne
2	Déplacer les données de la base de données locale vers la base de données crée dans le service d'archivage d'AWS S3	3h18 mins [7]	1 personne à temps partiel. Elle démarre la copie des données.
3	En utilisant un script automatisé, créer et configurer un serveur virtuel de type t.2micro dans AWS pour démarrer (l'API 3 : configurateur d'infrastructure matérielle.)	5 mins	1 personne
4	Exécuter l'API de création et de configuration d'une grappe de 10 serveurs virtuels du type m4.4xlarge	15 mins	1 personne
5	Exécuter l'API d'intégration des données	3h 18 mins	1 personne à temps partiel (impliquée seulement pour exécuter l'API).
6	Fermer la grappe des serveurs virtuels utilisée pour l'intégration des données	5mns	1 personne
7	Créer une grappe de serveur virtuel d'analyse des données en utilisant un script automatisé	30 mins	1 personne
8	Copie les scripts d'analyse des données (scripts de nettoyage, d'agrégation des données et celui du modèle prédictif) de la station de travail local vers le nœud maître de la grappe d'analyse	5 mins	1 personne
9	Exécuter le script d'analyse du modèle prédictif	1 h	1 personne
Total		8h20 min	

[7] Le délai de 3h18mins dépend de la vitesse de connexion internet utilisée pour copier les données de l'ordinateur de bureau utilisé pour l'archivage des données de l'étude de cas vers la base de données infonuagique

4.8.2 Comparaison des résultats : étude de cas par rapport à la situation actuelle

Le Tableau 4.19 suivant présente l'amélioration observée lors d'une recherche qui utilise l'approche proposée (c.-à-d. l'étude de cas) versus l'utilisation du processus actuel de recherche au laboratoire. Il faut préciser ici qu'il y a eu dix itérations de la simulation (c'est-à-dire 10 itérations d'analyse) lors de la recherche effectuée dans l'étude de cas. Ce même nombre d'itérations représente une moyenne typique observée par les répondants du questionnaire lors d'une recherche typique. Aussi, il faut noter que la mesure des délais rapportée est directement reliée à la puissance des serveurs virtuels utilisée lors de l'étude de cas. Par exemple, un délai de 3 heures et 18 minutes a été observé pour l'intégration des données. Ce délai aurait pu être réduit à beaucoup moins si on avait utilisé : soit des serveurs virtuels plus puissants; soit un nombre plus grand dans les grappes de serveurs. Un des buts est de mesurer les délais de mise à niveau de l'infrastructure matérielle au besoin, la possibilité d'adapter le modèle des données à chaque fois que c'est requis par le chercheur et puis ensuite identifier les délais d'adaptation de modèle des données aux nouveaux besoins informationnels de l'analyse ainsi que le nombre de personnes impliquées dans ce processus.

Tableau 4.19 Performance du nouveau cycle de recherche par rapport au cycle actuel

#	Mesure	Processus actuel	Nouvelle approche
Mesure 1	Le pourcentage des chercheurs capable d'adapter le modèle des données par eux même.	0%	100%
Mesure 2	Le nombre de personnes impliquées dans le processus de mise à niveau de la capacité de traitement de l'infrastructure matérielle. (ANNEXE I, Q10).	3 personnes	1 personne
Mesure 3	Le temps (nombre d'heure et de minutes) encouru pour la mise à niveau de la capacité de traitement de l'infrastructure matérielle. (ANNEXE I, Q9)	7 heures	50 secondes pour ajouter un nouveau serveur virtuel à la grappe

#	Mesure	Processus actuel	Nouvelle approche
Mesure 4	Le pourcentage d'automatisation des tâches de gestion de l'infrastructure matérielle. (ANNEXE II, Q6).	17%	100%
Mesure 5	Le temps (heures et minutes) nécessaire pour intégrer les données dans la même base de données et les rendre prêts à être utilisées par les modèles d'analyses des chercheurs. (Q7, ANNEXE I)	4 heures	3heures 19 minutes
Mesure 6	Le nombre de fois (c.-à-d. le nombre d'itérations d'analyse) que les chercheurs doivent refaire le cycle de recherche avant d'arriver à des résultats concluants. (ANNEXE I, Q2).	Plus que 10 itérations	10 fois
Mesure 7	Le nombre de fois que les chercheurs doivent adapter le modèle des données**.** (ANNEXE I, Q3).	Plus que 10 itérations	10 fois
Mesure 8	Le nombre de fois que les chercheurs doivent mettre à niveau la capacité de traitement de l'infrastructure matérielle utilisée dans l'analyse des données afin de pouvoir traiter les grands volumes de données. (ANNEXE I, Q8).	1.66 fois	1.66 fois
Mesure 9	Le pourcentage des chercheurs capable de reproduire une analyse de médecine de précision. (ANNEXE II, Q1).	Oui=66.67 % Non=33.33 %	100%
Mesure 10	La taille des données traitées par une analyse typique en médecine de précision. (Q1, ANNEXE 1) et (Q1 ANNEXE II).	18 GOS	205,578 GOS
Mesure 11	Le nombre de personnes impliquées dans le processus de reproduction d'analyses antérieures. (ANNEXE II, Q3)	2.33 personnes	1 personne
Mesure 12	Le temps pour reproduire une analyse antérieure. C'est le nombre d'heures écoulées pendant l'exécution des tâches de reproduction d'une analyse antérieure. (ANNEXE II, Q2).	15.33 heures	4heures 25minutes
Mesure 13	Le pourcentage d'automatisation des tâches d'adaptation de modèle des données. (ANNEXE II, Q5).	33%	75%[8]

[8] La tâche de modifier le fichier texte contenant la définition du modèle est manuelle. La création de nouvelles bases de données et l'adaptation de base de données existante à partir des modifications effectuées dans le fichier de définition sont totalement automatisées.

#	Mesure	Processus actuel	Nouvelle approche
Mesure 14	Le nombre de personnes impliquées dans les processus de mise à jour et la création du modèle des données (ANNEXE I, Q5).	3 personnes	1 personne
Mesure 15	Le temps (nombre d'heure et de minutes) écoulé pendant l'adaptation de modèle des données utilisé dans l'analyse de données. (ANNEXE I, Q4).	18 heures	15[9] minutes

Malgré que l'échantillon des données utilisées dans l'étude de cas et onze fois plus grandes que la moyenne des échantillons utilisés par les trois chercheurs du CRCHUM (18 GOS vs 205GOS analysés dans l'étude cas), les délais des étapes d'analyses communes (adaptation de modèle des données, mise à niveau de l'infrastructure d'analyse, etc.) sont beaucoup plus courts avec le nouveau cycle de recherche qu'avec le cycle actuel et le nombre de personnes requis pour la réalisation de ces tâches et beaucoup moins que celui requis par le cycle de recherche actuel.

Ces résultats montrent que la solution proposée basée sur la nouvelle approche de modélisation et d'intégration des données permet de simplifier le processus d'analyse avec l'introduction de l'utilisation des APIs dans le processus d'analyse, de réduire les délais de réalisation de l'analyse, de diminuer le nombre de ressources impliquées, de permettre la reproductibilité des recherches antérieures, et finalement d'écourter la durée des recherches ce qui permet aux chercheurs de faire plus de recherche dans de courts délais. Ceci est un grand avantage pour le domaine de médecine de précision. En effet, la réalisation d'un réel cas d'analyse de recherche en utilisant la nouvelle solution proposée a permis les améliorations suivantes du processus actuel :

[9] 15 minutes = 14 minutes pour la modification du fichier qui contient la définition textuelle du modèle des données et 27 secondes pour la génération des fichiers binaires prêts à être utilisés du modèle de données (voir ANNEXE III)

- **Reproduction des analyses antérieures**
    - Le pourcentage de la capabilité de reproduire des analyses antérieures est de 100% (Tableau 4.19, mesure 9);
    - Le temps pour reproduire une analyse antérieure a été réduit d'environ 73% (Tableau 4.19, mesure 12);
    - Le nombre de personnes impliquées dans le processus de reproductibilité de l'analyse a été réduit de 57% (Tableau 4.19, mesure 11).
- **Réalisation de nouvelles recherches**
    - **Amélioration de la tâche de l'adaptation de modèle des données aux nouveaux besoins informationnels des recherches:**
        - Capabilité des chercheurs d'adapter leurs modèles des données par eux même a été amélioré par 100% (Tableau 4.19, mesure 1);
        - Le temps d'adaptation de modèle de donnée a été réduit de 98% (Tableau 4.19, mesure 15). Le délai est passé de 18 heures à 15 minutes;
        - Le nombre de personnes impliquées dans la tâche d'adaptation de modèle des données a été réduit de 67% (Tableau 4.19, mesure 14);
        - La tâche d'adaptation de modèle des données a été automatisée à 75% c'est ce qui a permis la réduction du délai de cette étape de l'analyse (Tableau 4.19, mesure 13).
    - **Amélioration de la tâche de la gestion de l'infrastructure matérielle utilisée pour les analyses des recherches**
        - Le nombre de personnes requis pour augmenter la capacité de traitement de la plateforme matérielle a été réduit de 67% (Tableau 4.19, mesure 2);
        - Le délai de mise à niveau de la capacité de traitement a été réduit considérablement. Sept heures avec le processus actuel, et 50 secondes environ pour ajouter un nouveau serveur virtuel à la grappe existante (Tableau 4.19, mesure 3);
        - La tâche de la gestion de la capacité de traitement de la plateforme a été automatisée à 100% (Tableau 4.19, mesure 4).

- **Amélioration de la tâche d'intégration des données**
  - **Cette tâche a été améliorée de deux perspectives :**
    - Le délai d'intégration de toutes les données cliniques et génétiques dans la base de données d'analyse a été réduit de 17% sachant que l'échantillon testé était 11 fois plus grand (18 GOS vs 205 GOS). Ce délai aurait pu être amélioré encore plus, si des serveurs virtuels plus puissants ont été utilisés. Dans la réalisation de l'étude de cas, le budget alloué à l'utilisation des serveurs virtuels était très limité (Tableau 4.19, mesure 10 et 5);
    - Toutes les données requises pour l'analyse (données génétiques provenant des fichiers textes et les données cliniques provenant d'une base de données relationnelle) ont été intégrées dans le même format de stockage (fichiers parquet adaptés pour les analyses des données massives) et intégrées dans la même base de données (AWS S3). Le processus actuel ne permet pas l'unification des types de données.

4.9 Résumé

Ce chapitre a décrit les activités de validation du prototype expérimental de la nouvelle approche d'adaptation de modèle et d'intégration des données et les résultats de l'expérimentation du prototype sur une étude de cas de recherche en de médecine de précision :

- **Activités de validation du prototype d'expérimental**
  - Décomposition de l'objectif de recherche principal en sous-objectifs mesurables, les tests effectués sur les différents composants de la solution et les résultats de ces tests;
  - Définition de l'étude de cas à utiliser pour expérimenter le prototype;
  - Description du processus actuel utilisé par l'équipe de recherche du Dr Hamet au CRCHUM;

- Définition du modèle de mesure à utiliser dans l'étude de cas afin d'évaluer la performance du prototype expérimental. Le modèle de mesure a été présenté de la manière suivante :
  - Pour chaque sous-objectif, les concepts à mesurer, les mesures et l'approche de capture de ces mesures ont été définis et présentés;
  - Les résultats de ces mesures à partir du processus de recherche actuel afin de les utiliser comme des données de références.
- Présentation des étapes de réalisation de l'étude via la description des activités effectuées lors de chaque étape du nouveau cycle de recherche;

- **Résultat de l'expérimentation du prototype expérimental**

Les résultats de l'expérimentation du prototype ont été présentés dans ce chapitre comme suis : en premier temps, les résultats de l'expérimentation du prototype sur la réalisation de nouvelle recherche et de celles de la reproductibilité de recherche antérieures ont été présentés, puis une comparaison entre ces résultats et la performance du processus actuel.

Cette comparaison a permis de montrer les gains apportés par l'utilisation de la nouvelle approche.

CHAPITRE 5

CONCLUSION, INTERPRÉTATION ET DISCUSSION

5.1 Sommaire

Ce chapitre présente la conclusion de cette thèse. Un sommaire de la recherche y est présenté et ses résultats y sont interprétés et discutés. La contribution de la recherche dans le contexte pratique de recherche du laboratoire du Dr Hamet, au CRCHUM, pour une recherche dans le domaine de la médecine de précision fait aussi objet de discussion. Enfin les limites de cette recherche, les recommandations et les propositions pour des recherches futures sont présentées.

Cette thèse proposait une nouvelle approche d'adaptation de modèle et d'intégration des données pour les recherches de médecine de précision. Elle s'appuie sur les avancées en technologies infonuagiques et aussi sur les avancées du traitement des données massives dans le domaine des données génétiques. Son but était de tenter de résoudre les deux défis auxquels les chercheurs sont confrontés à chaque fois qu'ils entreprennent une nouvelle recherche en médecine de précision :

1. Le premier défi est causé par le fait que les données requises pour réaliser les recherches dans ce domaine proviennent de différentes sources, sous formes très diversifiées, et comportent une grande volumétrie qui ne cesse de croître (Karen He, Dongliang Ge, 2017). De plus, les logiciels de recherche en médecine de précision actuels ont de la difficulté à traiter ces très grands volumes de données. (Prosperi et coll. 2018);
2. Le deuxième défi est que les logiciels de recherche en médecine de précision sont incapables de s'adapter facilement au caractère dynamique et diversifié des besoins informationnels des recherches en médecine de précision (Prosperi et coll. 2018). Les besoins informationnels varient plusieurs fois au cours du même cycle de recherche. Ceci nécessite l'adaptation de modèle et l'intégration des données plusieurs fois (Belghait, Kanzki, et April 2018).

Ces deux défis limitent la réactivité, la créativité et la capacité des chercheurs en médecine de précision à effectuer des analyses croisées, comportant plusieurs facteurs, qui influencent le développement de pathologies. Conséquemment, cette situation ralentit considérablement la vitesse des découvertes du domaine. C'est précisément sur cette problématique que cette thèse vise à contribuer à des solutions.

Deux pistes de solutions à ces deux défis sont discutées dans la littérature :

- Pour le premier défi, les solutions proposées se résument dans l'utilisation ou bien des services infonuagiques existants de médecine de précision comme DNANexus (DNANexus, 2019) et DRAGEN (Illumina, 2019) ou bien des logiciels de recherches existants comme BRISK, iCOD, I2B2, etc. Ces solutions, malgré leurs avantages, ils présentent plusieurs limitations : dépendances envers les fournisseurs de service de logiciels de recherche, coût en général très élevé, logiciels spécialisés pas adaptés pour faire des recherches complexes et qui ont des besoins informationnels variés et ne supportent pas la fonction de l'adaptabilité de leur modèle des données aux nouveaux besoins informationnels (voir Tableau 2.1) et infrastructure statique qui ne permet pas l'échelonnabilité de la capacité de traitement en fonction du volume de données croissant.

2. Pour le deuxième défi qui concerne l'évolution dynamique et la diversité des besoins informationnels des recherches, les solutions proposées se résument aux trois stratégies d'adaptation de modèle de données analysées et discutées dans la section (2.3.4 Gestion de l'adaptation des modèles des données aux nouveaux besoins informationnels). Ces solutions, malgré leurs avantages, aucune d'elle ne permet de résoudre la problématique des adaptations des modèles des données dans le domaine des recherches en médecine de précision qui sont nombreuses et qui sont requises dans des délais très courts.

Cette thèse se propose de résoudre ces deux défis par la proposition d'une solution basée sur quatre éléments principaux :

1) une méthodologie pour la gestion des adaptations successives du modèle des données utilisé pour la réalisation des recherches de médecine de précision;

2) une approche, une mécanique et des outils pour l'intégration des données dans un modèle des données qui doivent être adaptés continuellement aux besoins informationnels des chercheurs;

3) une approche itérative d'exécution du cycle de recherche rendu possible grâce aux deux composants précédents;

4) un nouveau cycle de recherche qui intègre les trois composants précédents.

Chacun de ces quatre éléments vise à réduire aux maximums l'impact de ces défis sur la réactivité, la créativité et la capacité des chercheurs en médecine de précision à effectuer des analyses croisées comportant plusieurs facteurs sur de grands volumes de données afin de permettre d'accélérer la cadence des découvertes dans le domaine de médecine de précision.

5.2 Interprétation et discussion des résultats

5.2.1 Analyse et discussion des résultats de la recherche

5.2.1.1 Réflexion sur les questions de recherche

La revue de littérature décrit plusieurs logiciels et cycles de recherche utilisés par les chercheurs en médecine de précision. Cependant, il est rapporté qu'aucun d'eux ne semble être suffisamment flexible pour appuyer efficacement la variabilité et l'évolutivité constante des besoins informationnels dans un contexte où le volume et l'hétérogénéité des données se complexifient continuellement. La tendance observée des solutions offerte à ces chercheurs est de leur proposer des logiciels de plus en plus spécialisés. Même en utilisant ces logiciels, les chercheurs sont souvent limités aux modèles de données contraignants disponibles dans ces logiciels (Belghait, Kanzki, et April 2018). Il a été rapporté aussi lors de la revue littéraire qu'il

y a progressivement de plus en plus de solutions qui proposent l'utilisation de technologies infonuagique pour tenter de résoudre le problème de l'échelonnabilité de l'infrastructure et pouvoir traiter de très grands volumes de données.

La motivation principale de cette thèse était d'améliorer le processus itératif de préparation et d'analyse des données des recherches en médecine de précision afin d'aider les chercheurs du domaine de la médecine de précision de solutionner les problèmes identifiés dans l'introduction de cette thèse. Pour atteindre ce but, trois questions de recherche ont été formulées dans la section introduction de cette thèse :

1. Comment permettre aux chercheurs en médecine de précision d'adapter eux-mêmes plus dynamiquement et efficacement, le modèle de ces données afin qu'il puisse répondre aux besoins informationnels hétérogènes, spécifiques et évolutifs de chacune de leurs recherches en médecine de précision?

Dans une même recherche, les chercheurs doivent souvent modifier le modèle des données pour l'adapter aux nouveaux besoins informationnels de leur recherche. Dans l'équipe de recherche du Dr Hamet, au CRCHUM, il a été rapporté que la fréquence d'adaptation du modèle de données est supérieure à dix fois par recherche (ANNEXE I, Q2). Aussi, il a été rapporté dans la revue de littérature que les logiciels spécialisés en médecine de précision disponible n'ont pas été conçus pour adresser le caractère dynamique et à la diversité des besoins informationnels des recherches en médecine de précision (Prosperi et coll. 2018). La solution proposée par cette thèse est la nouvelle approche d'adaptation dynamique de modèle de données (figure 3.2) qui permet aux chercheurs de pouvoir adapter par eux même de manière répétitive leur modèle de données aux tout nouveaux besoins informationnels de recherche. L'étude de cas réalisé dans cette thèse a permis de vérifier la véracité de l'hypothèse que cette nouvelle approche permet réellement aux chercheurs d'adapter par eux même leur modèle de données sans grands efforts.

Lors de cette étude de cas, l'adaptation du modèle des données a été facilitée via l'écriture de la définition des nouveaux besoins informationnels dans un fichier texte (respectant le format Avro), puis à l'aide d'un simple script automatisé, des fichiers binaires contenant la définition du nouveau modèle de données sont générés automatiquement. En utilisant ce prototype expérimental dans lequel la nouvelle approche a été mise en œuvre, une adaptation du modèle de données a été effectuée rapidement et simplement, en cours de recherche, par une seule personne. Le délai observé pour la création des fichiers binaires du nouveau modèle des données n'a nécessité que deux minutes. Les chercheurs du laboratoire de recherche du Dr Hamet ont rapporté qu'actuellement, sans cette innovation, dix-huit heures et trois personnes sont nécessaires pour adapter le modèle de données.

2. Comment intégrer toutes ces données hétérogènes nécessaires à une recherche en médecine de précision, c.-à-d. les regrouper dans un même format de stockage, dans un seul modèle et dans la même base de données?

   Les données requises pour la recherche en médecine de précision proviennent de plusieurs systèmes et comportent des types de données très diversifiées (Xue et coll. 2016). Afin de répondre à cette question, une nouvelle approche d'intégration continue des données a été proposée dans cette thèse (figure 3.3). Cette approche a été mise en œuvre dans le prototype expérimental et testée dans l'étude de cas réalisé dans cette thèse. Dans cette étude de cas, les données provenaient de la cohorte d'ADVANCE. Les données génétiques étaient toutes de type caractères et de type texte et étaient sauvegardées dans des fichiers textes et respectent le format de fichier de génotype d'Oxford (c.-à-d. les fichiers avec une extension « .gen »). Les données cliniques étaient des types numériques, caractères et textes et étaient sauvegardées dans une base de données relationnelle. Le prototype expérimental a fait convertir toutes les données génétiques et cliniques dans le format parquet (Parquet est un format de compression et de stockage de données orienté colonne permettant d'accéder à des quantités massives de données plus rapidement (Massie et coll. 2013)) et les a tous intégrées et colocalisées dans une seule base de données sous format de fichiers binaires stockés dans le service S3 de AWS;

3. Comment aider ces chercheurs à réaliser plus efficacement (c.-à-d. avec moins de personnel, moins d'efforts et plus rapidement) le cycle de recherche de manière itérative ? C'est-à-dire récupérer les résultats des analyses précédentes, ajuster le tir en fonctions des résultats obtenus, ajouter/modifier des données et refaire le cycle de recherche autant de fois que nécessaire jusqu'à l'obtention des réponses aux questions de recherche ?

Pour compléter une recherche dans ce domaine, les chercheurs ont besoin de faire dix itérations d'analyse en moyenne (ANNEXE I, Q2). Afin de répondre à cette question, cette thèse propose d'intégrer l'approche d'analyse itérative avec la nouvelle approche d'adaptation de modèle et d'intégration des données dans un nouveau cycle de recherche (voir figure 3.7). Ce nouveau cycle de recherche a été utilisé pour réaliser l'étude de cas réalisée dans le cadre de cette recherche.

Dans la section suivante, les résultats de l'étude de cas ont été analysés afin de faire une réflexion sur l'atteignabilité de chaque sous-objectif.

5.2.1.2 Réflexion sur les objectifs de la recherche

Cette recherche a un objectif principal d'amélioration de la situation pour les chercheurs. L'étude de cas a permis de vérifier que la nouvelle approche d'adaptation de modèle et d'intégration des données proposée dans cette thèse a permis une amélioration de la situation. Afin de discuter de l'atteignabilité de cet objectif, il a été décomposé en quatre sous-objectifs spécifiques et mesurable (voir Tableau 4.1 - Hiérarchie des objectifs de la recherche), puis vérifié à quel degré cet objectif a été atteint. Cette approche permet une réflexion sur un chiffrement possible de la notion d'amélioration dans ce contexte, car il peut être difficile d'établir le succès ou l'échec de la proposition théorique sans discuter d'une cible à atteindre. Aussi sans cette discussion il pourrait être difficile d'avoir un consensus sur le degré d'amélioration requis pour prononcer que l'innovation proposée contribuera significativement au domaine de recherche.

- **Objectif 1 : Améliorer l'approche d'adaptation de modèle des données utilisé dans le cycle de recherche en médecine de précision afin de permettre aux chercheurs de pouvoir continuellement adapter le modèle des données, d'une recherche, aux besoins informationnels changeants et permettre d'intégrer toutes ces données dans une seule base de données permettant ainsi d'effectuer des analyses sur des quantités massives de données:**
    - **Objectif 1.1 : permettre aux chercheurs de pouvoir continuellement adapter le modèle des données de la recherche aux nouveaux besoins informationnels et faire des analyses de données sur de grandes quantités de données :** cet objectif a été atteint. En effet l'étude de cas démontre qu'il est possible de modifier le modèle des données, en cours de recherche, autant de fois que les chercheurs en a besoin. Ainsi il est possible d'affirmer que le pourcentage des chercheurs capable d'adapter le modèle des données, par eux-mêmes, est maintenant de 100% (mesure 1 du Tableau 4.19).
    - **Objectif 1.2 : améliorer et simplifier les tâches de préparation des données :** cet objectif a été atteint, l'étude de cas démontre que l'utilisation de la nouvelle approche d'adaptation de modèle et d'intégration des données a permis de réduire le nombre de ressources impliqué dans le processus de préparation des données ainsi que le délai de réalisation de cette étape de manière notable (voir mesures 2,3,4,13,5,6,7 et 8 du Tableau 4.19);
    - **Objectif 1.3: améliorer et simplifier le processus de reproductibilité des analyses antérieures et réduire les efforts (délais et ressources) requis :** cet objectif a été atteint. En effet, l'étude de cas démontre que l'utilisation de la nouvelle approche d'adaptation de modèle et d'intégration des données avec l'approche itérative permet la reproductibilité d'analyse antérieure avec moins d'effort. Les résultats de l'étude de cas montrent que 33.33% des répondants du sondage d'opinion ont dit qu'ils n'étaient pas capables de refaire une analyse existante et 66.66% ont dit qu'ils étaient capables de refaire une analyse existante (ANNEXE II, Q1) sachant que les deux répondants travaillent avec des échantillons de données très petits pour des recherches du domaine de

médecine de précision, 1 GO pour le premier répondant et 6 GOS pour le deuxième (ANNEXE I, Q1). Avec le nouveau cycle de recherche, l'échantillon était de 205 GOS, et tout le processus de reproduction d'analyse a duré environ 4h25 et a nécessité l'effort d'une personne (voir mesures 9,10,11,12 du Tableau 4.19);

- **Objectif 2.2 : Réduire les efforts (délai et ressources) requis pour l'adaptation de modèle des données :** cet objectif a été atteint. En effet, le processus actuel d'adaptation de modèle des données requiert trois personnes et environ 18h de délai pour pouvoir adapter le modèle des données, alors qu'avec la nouvelle approche s'adaptation de modèle et d'intégration des données, une seule personne est requise pour cette tâche. Le délai d'adaptation de modèle a passé de 18 heures à 15 minutes, une amélioration de 98% (mesure 15, Tableau 4.19).

5.2.1.3 Réflexion concernant l'atteignabilité des objectifs du prototype expérimental

L'étude de cas nécessitait de créer un prototype logiciel expérimental pour mettre en œuvre la nouvelle approche d'adaptation de modèle et d'intégration des données. Afin d'aider à adresser les défis auxquels font face les chercheurs du domaine de la médecine de précision actuellement et atteindre les objectifs visés et énoncés à la section (0.2 But, Objectifs et Questions de recherche), le prototype expérimental doit aussi s'assurer d'atteindre les objectifs spécifiques :

- **Flexibilité du modèle des données**: cet objectif a été atteint. En effet, une fois les besoins informationnels de l'étude de cas identifiés, une adaptation au modèle des données d'analyse génétiques ADAM a été effectuée automatiquement en utilisant l'API de l'adaptation de modèle des données du prototype. La procédure d'adaptation de modèle des données a pris moins de 15 minutes. Une grande partie de ce temps a été utilisé pour l'ajout de nouveaux champs de données dans le fichier de définition du modèle des données. La flexibilité démontrée va au-delà de seulement l'adaptation de modèle des données ADAM facilement, il a été aussi observé qu'il serait aussi facile

d'adapter n'importe quel schéma Avro existant avec des nouveaux besoins en analyse de données. Cette nouvelle fonctionnalité va permettre aux chercheurs d'effectuer un processus progressif et itératif pour effectuer leur analyse de données de médecine de précision avec facilité;

1. **Évolutivité**: cet objectif a aussi été atteint. Cet objectif de conception a été évalué à l'aide des observations suivantes :

    - **L'adaptabilité de l'importation d'une grande quantité de données de patients** : Les éléments de la nouvelle approche d'intégration continue des données ont été mis en œuvre dans deux API distinctes dans le prototype expérimental. Ils permettent d'intégrer les données cliniques et génétiques dans le nouveau modèle des données. Ces APIs ont été conçues de façon à toujours inclure les nouvelles données dans le modèle, ce qui permet aux chercheurs d'enrichir leurs échantillons de données avec toute nouvelle donnée qu'ils sont susceptibles d'acquérir dans le futur et d'exécuter leur analyse sur de nouveaux échantillons de données;

    - **L'élasticité de l'infrastructure informatique :** cet objectif a été atteint puisque la configuration de l'infrastructure d'analyse et d'intégration des données, du prototype expérimental, pour l'adapter à la volumétrie traitée dans l'étude de cas a été effectuée en très peu de temps. Il a fallu moins de dix minutes pour configurer et démarrer la grappe de serveurs virtuels (contenant cinq serveurs virtuels); alors que la gestion de la grappe de serveurs virtuels pour l'analyse des données (c.-à-d. ajouter ou enlever des serveurs virtuels à la grappe) a duré 50 secondes pour l'ajout d'un seul serveur virtuel et trois minutes pour l'ajout de cinq serveurs virtuels. L'approche proposée offre ainsi une grande flexibilité aux chercheurs pour adapter la taille de la grappe, en temps réel, en fonction du temps désiré et des ressources financières disponibles pour la réalisation d'une l'analyse;

- **Évolutivité de l'analyse des données de la recherche :** l'atteinte de cet objectif n'a pas été vérifiée. Par contre la conception de l'architecture logicielle du prototype expérimental permettrait aux chercheurs d'utiliser la fonctionnalité de reproductibilité des analyses antérieures afin de récupérer les données et les résultats et les utiliser dans des analyses de nouvelles recherches.

2. **Reproductibilité (ou Réplication) de l'analyse des données de la recherche** : cet objectif a aussi été atteint. En effet, la reproductibilité d'analyse antérieure a été vérifiée lors de la réalisation de l'étude de cas via l'archivage de toutes les données de l'étude de cas après la complétion de l'étude, et puis le processus de reproductibilité d'analyse antérieure défini dans le chapitre 4 (section 3.7 Reproductibilité de l'analyse) a été utilisé pour reproduire la recherche de l'étude de cas. Les résultats obtenus (Tableau 4.18 - Évaluation de l'étape de reproduction d'analyses existantes) démontrent que la nouvelle approche d'adaptation de modèle et d'intégration des données permet la reproductibilité des recherches antérieures avec moins de ressources et dans des délais courts (3h43 minute).

   La plupart des logiciels existants en médecine de précision utilisent leur propre infrastructure pour le traitement des données massives, ce qui limite l'échelonnabilité de leur logiciel et les empêche de traiter un grand volume de données. Le nouveau logiciel proposé est à l'opposé de cela, offrant aux chercheurs un logiciel optimal sans limites en matière d'échelonnabilité.

5.2.1.4 Réflexion sur la faisabilité de la recherche de l'étude de cas à l'aide de logiciels

Cette nouvelle proposition de solution permet de réfléchir sur les caractéristiques souhaitées des logiciels futurs de recherche en médecine de précision. Elle est extraite des observations et conclusions de l'étude de cas ainsi que des observations des auteurs présentées dans la revue de littérature.

L'analyse des données de l'étude de cas a démontré que le prototype logiciel expérimental, qui met en œuvre la nouvelle approche de modélisation et d'intégration des données, possède

toutes les caractéristiques souhaitées d'un logiciel de médecine de précision future. Les autres logiciels existants, étudiés et présentés au Tableau 5.1 ci-dessous, échouent sur au moins une des caractéristiques souhaitées. Rappelons que dans la liste ci-dessous des logiciels étudiés, aucun ne permet que le modèle de données s'adapte, de façon dynamique, aux nouveaux besoins informationnels pendant une recherche; d'autre part, il y en a deux seulement qui utilisent une plateforme de données massives infonuagique. Le Tableau 5.1 ci-dessous présente le résultat de cette évaluation.

Tableau 5.1 Aperçu de la faisabilité de l'analyse des données de l'étude de cas avec d'autres logiciels de recherche de médecine de précision

#	Caractéristiques du logiciel de recherche de médecine de précision	Logiciel de médecine de précision						
		Prototype expérimental	1 BRISK(Tan et coll., 2011)	2 iCOD(Shimokawa et coll., 2010) (Alessandro et coll., 2013)	3 I2B2/TranSMART(Scheufele et coll. 2014) (Murphy et coll. 2006)	4 OncDRS(Orechia et coll, 2015) (Londin & Barash, 2015)	5 deCODE(Hakonarson et coll., 2003)	6 IROmiCS(Chloé Cabot, Lina F. Soualmia, 2015, Soualmia et coll., 2015)
1	Permets l'exploration des données génétiques	O	O	O	O	O	O	O
2	Exploration des données autres que génétiques (clinique, historique médical, etc.)	O	O	O	O	O	O	O
3	Permets à son modèle des données d'être ajusté/bonifié avec de nouvelles données	O	N	N	N	N	N	N
4	Fournis une infrastructure flexible	O	N	N	N	N	N	N
5	Permets de faire une étude sur une grande quantité de données	O	O	O	O	O	O	O
6	Permets le rééchelonnement de la capacité de traitement de l'infrastructure informatique	O	N	N	N	N	N	N
7	Permets de reproduire une étude sans effort considérable	O	N	N	N	N	N	N
8	Faisabilité de l'analyse de l'étude de cas	O	N	N	N	N	N	N

5.2.2 Réflexion concernant l'interprétation des résultats de la recherche

Les résultats de la recherche évalués et analysés dans la section 5.3.1 ci-haut ont été obtenus via l'expérimentation de la nouvelle approche d'adaptation de modèle et d'intégration des données sur un cas réel d'une analyse de données d'une recherche de médecine de précision.

La pertinence de ces résultats s'explique par les faits suivants :

- Les deux défis principaux présentés dans la section introduction et identifiés dans la revue de littérature et qui sont liés aux recherches de médecine de précision ont aussi été constatés dans le processus d'analyse du laboratoire du Dr Hamet de CRCHUM (ANNEXE I et ANNEXE II);
- Le prototype expérimental du logiciel qui met en œuvre la nouvelle approche d'adaptation de modèle et d'intégration des données a été testé sur une analyse d'un réel cas de recherche proposée par le laboratoire de recherche du Dr Hamet au CRCHUM. Elle consiste à développer un modèle prédictif pour identifier les patients à risque de développer la maladie d'insuffisances rénales chronique (de l'anglais Chronic Kidney Disease (CKD)) chez les patients atteints par le diabète de type 2 (DT2);
- Les données utilisées pour réaliser cette étude proviennent de la cohorte d'ADVANCE. C'est les mêmes données utilisées par l'équipe de recherche du CRCHUM, aussi utilisé dans le contexte de plusieurs autres recherches comme l'utilisation du HGI comme facteur de prédiction de complication causée par le diabète (van Steen et coll., 2018). Les données de cette cohorte sont stockées dans des types de données différents (texte, caractères, binaires, etc.), sous deux formes de stockages différents localisés dans deux bases de données différentes: une base de données relationnelle et une base de données fichiers textes respectant le format de fichiers de génotype d'Oxford (c.-à-d. les fichiers avec une extension « .gen »). Les données de cette cohorte sont un bon exemple des données traitées dans les recherches de médecine de précisions;
- L'expérimentation a été faite en utilisant le fournisseur infonuagique d'AWS, le seul service propriétaire utilisé est le service d'archivage S3 d'AWS, en déménageant les

données dans n'importe quel autre espace de stockage, on aurait pu faire la même expérimentation en utilisant la plateforme d'autres fournisseurs infonuagiques avec les deux modifications mineures suivantes :

- Remplacer les pilotes de connexion de AWS dans les APIs par ceux du fournisseur infonuagique à utiliser;
- Choisir un autre espace de stockage que le service S3 d'AWS;

La validation de la solution proposée a été faite en vérifiant chaque composant de la solution de manière unitaire (voir section 4.2 Validation de l'approche proposée) et en exécutant chaque étape du nouveau cycle de recherche en utilisant le prototype expérimental pour réaliser la recherche de l'étude de cas (voir 5.5 Exécutions des étapes de l'analyse de la recherche pour la réalisation de l'étude de cas).

La solution proposée par cette thèse est composée de quatre éléments principaux: une approche dynamique d'adaptation de modèle des données, une intégration continue des données, une approche d'analyse itérative et un nouveau cycle de recherche pour les recherches de médecine de précision basée sur l'orchestration des différentes étapes de l'analyse en utilisant la nouvelle approche de modélisation et d'intégration des données (Figure 3.5). La composante principale de la solution est celle de l'approche dynamique d'adaptation de modèle des données qui garantit aux chercheurs de toujours pouvoir adapter et personnaliser le modèle de données avec le minimum d'effort (mesure 14 et 15, Tableau 4.19). Cette composante permet aux chercheurs de toujours avoir un modèle des données mis à jour avec la définition de toutes les données requises pour leur analyse. En utilisant la composante d'intégration continue des données, les chercheurs vont pouvoir avoir une base de données continuellement enrichie par de nouvelles données, ce qui leur permet de pouvoir réaliser leurs recherches de manière itérative. L'optimisation du processus des réalisations de recherche est garantie par l'automatisation des étapes du nouveau cycle de recherche.

Dans la revue de littérature, trois approches de gestion d'adaptation de modèles des données ont été identifiées: la première est celle de Castelli (Castelli, 1998) qui propose une stratégie

basée sur une tenue rigoureuse de documentation de toutes les adaptations effectuées sur les modèles, la deuxième est celle de Kupfer et ses collègues (Kupfer, Eckstein, Neumann, & Mathiak, 2006) qui proposent une approche basée sur la génération automatique des ontologies des bases des données et le mappage de ces ontologies avec les modèles des données et une troisième de Andany et ses collègues (Andany, Léonard, et Palisser 1991, Sven-Eric Lautemann 1997) qui proposent une approche basée sur la gestion des versions des modèles.

L'approche dynamique d'adaptation de modèle des données proposée dans cette thèse se distingue de toutes ces approches par les deux éléments suivants :

L'approche ne nécessite aucune gestion ni de documentation, ni de version, ni d'anthologie de modèle des données;

Les trois approches ne peuvent pas être automatisées, de plus, avec la prolifération des changements dans le modèle, la tenue de la documentation, la gestion des versions et des ontologies deviennent très compliquées à gérer à notre avis. Dans l'approche dynamique proposée, la création des fichiers binaires contenant la définition des besoins informationnels est complètement automatisée, ce qui n'est pas le cas dans les autres approches. De plus, une bonne partie de la définition du modèle des données est automatisable. En effet, il est possible de générer automatiquement un fichier texte avec le format Avro requis par la nouvelle solution à partir des sources des données.

5.3 Pertinence de la recherche

5.3.1 Contributions originales de la recherche

Cette thèse a fait quatre contributions principales :

- Une approche dynamique d'adaptation et de personnalisation de modèle des données (voir Figure 3.2) permettant aux chercheurs d'adapter le modèle des données utilisé à tout nouveau besoin informationnel;
- Une nouvelle approche d'intégration continue des données basées sur l'unification du format et de stockage des données et la colocalisation de toutes les données

nécessaires aux recherches dans une seule base de données. Elle utilise la combinaison de deux technologies, celle du traitement des données massives et celle de l'infonuagique (voir figure 3.3);

- Une approche itérative pour la réalisation de recherche (voir Figure 3.6);
- Un nouveau cycle de recherche (voir Figure 3.7) formée de neuf étapes :

  1) Spécification d'hypothèse(s) pour la recherche;
  2) Définition des besoins informationnels;
  3) Création/adaptation de modèles de données spécifique pour l'analyse de données;
  4) Mise en place de l'environnement d'intégration des données;
  5) Intégration des données;
  6) Ajustement de la capacité de traitement de la plateforme d'intégration des données;
  7) mise en place de l'environnement de l'analyse des données;
  8) Analyse des données;
  9) Ajustements de la capacité de traitement de la plateforme d'analyse des données.

Tout au long du déroulement de cette recherche, les nouvelles contributions ont été publiées afin de valider auprès de la communauté scientifique la véracité et la pertinence des contributions cette thèse :

- Publication dans une conférence de santé numérique pour la vérification et la validation de la contribution du concept de modèle des données adaptable et personnalisable : Belghait, Fodil, Beatriz Kanzki, and Alain April. 2018. "ADAM Genomics Schema - Extension for Precision Medicine Research *." In *DH '18 Proceedings of the 2018 International Conference on Digital Health*, 4. https://doi.org/10.1145/3194658.3194669;
- Publication dans le journal international des tendances dans la recherche et développement pour la validation du nouveau cycle de recherche et du nouveau logiciel : Belghait, Fodil, and Alain April. 2018. "The Future of Large-Scale Precision Medicine Research Platforms: Preparing the Data for Analysis," publié dans le International Journal of Trend in Research and Development (IJTRD), ISSN: 2394-9333, Special Issue | ICTIMESH-18, décembre 2018, pp: 91–94. http://www.ijtrd.com/papers/IJTRD19226.pdf;

- Publication dans la conférence internationale de la santé publique numérique pour valider les résultats expérimentaux de notre recherche : Belghait, F., April, A., Hamet, P., Tremblay, J., & Desrosiers, C. (2019). A Large-scale and Extensible Platform for Precision Medicine Research. Proceedings of the 9th International Conference on Digital Public Health, 47–54. https://doi.org/10.1145/3357729.3357742.

5.3.2 Impacts attendus de la recherche sur l'industrie

La principale contribution de cette thèse est la proposition d'une nouvelle approche d'adaptation de modèle et d'intégration des données qui couvre l'ensemble du cycle de préparation des données d'une recherche du domaine de médecine de précision. Cette contribution principale aura les impacts principaux suivants :

Le premier impact sera sur les logiciels de recherche du domaine de médecine de précision. Cette thèse fournit aux concepteurs de logiciels de ce domaine les informations sur les éléments à prendre en considération dans la conception des logiciels de recherche afin de mieux servir les chercheurs du domaine et ainsi les découvertes vont se faire plus rapidement ce qui aura un impact positif sur la santé des patients;

Le deuxième impact est l'élimination de la contrainte des logiciels spécialisés dans des domaines précis de recherche de médecine de précision (logiciels pour les recherches de cancer du sein, logiciel pour les recherches de diabètes, logiciel pour les recherches de l'hypertension, etc.) et des modèles des données statiques, les chercheurs vont pouvoir faire des recherches croisées où ils pourront combiner les données de plusieurs catégories et provenant de plusieurs sources ce qui leur permettra d'explorer de nouvelles possibilités qu'ils ne puissent pas faire présentement à cause des contraintes techniques des logiciels disponibles. Désormais, leur créativité est la seule contrainte qui leur reste, et deuxièmement, avec l'élimination des coûts reliés à l'acquisition et à la gestion des infrastructures nécessaires pour la réalisation des analyses, la réduction des coûts reliés aux ressources techniques spécialisées pour l'exploitation de ces infrastructures et pour

l'exécution des étapes de préparation des données, les chercheurs vont disposer de plus de budget pour faire plus de recherche;

Le troisième impact est sur les logiciels de recherche de tous les domaines. Même si l'application visée par cette recherche était au départ pour le domaine de médecine de précision, les résultats et les constats de cette thèse peuvent être utilisés dans la conception des logiciels de recherche de n'importe quel domaine qui requiert une gestion dynamique de l'adaptation de modèle des données.

5.4 Limites de la solution proposée

La nouvelle approche d'adaptation et d'intégration de modèle a été développée pour améliorer le processus itératif de préparation et d'analyse des données des recherches en médecine de précision. Cependant, les limitations suivantes ont été rencontrées lors de la réalisation de cette recherche :

Le nombre de répondants aux questionnaires d'évaluation du processus actuel était limité aux chercheurs du CRCHUM seulement. Ceci limite l'étendu de la validation de la nouvelle approche à seulement un seul processus de recherche (processus du laboratoire du Dr Hamet du CRCHUM);

L'échantillon des données de l'étude de cas utilisé pour construire le modèle prédictif était relativement petit (1118 patients). Ceci n'avait pas d'impact sur la validation du processus de recherche, mais diminue de la véracité des résultats de l'analyse du modèle prédictif;

La revue de littérature a démontré que les données requises pour les recherches du domaine de médecine de précision peuvent provenir de nombreuses sources des données hétérogènes. L'approche proposée dans cette thèse a été validée pour seulement deux types de sources des données : données provenant de fichiers texte et les données provenant de base de données relationnelle. L'adaptation de modèle a été testée dans l'étude de cas seulement pour les types de données les plus standards (texte, caractère, numérique, binaire, etc.). Dans le domaine de médecine de précision, les données peuvent provenir

sous d'autres formes de données : données de type image, données de type audio ou autres. L'adaptation de modèle de données contenant ce type de données n'a pas été testée.

5.5 Recommandation

La nouvelle solution proposée et développée dans le contexte de cette thèse a été développée avec la perspective d'une preuve de concept afin de valider la pertinence et la viabilité des nouveaux concepts introduits par la présente recherche. Suite à la réalisation de l'étude de cas, les deux recommandations suivantes sont à considérer pour améliorer le processus de recherche dans le laboratoire de Dr Hamet au CRCHUM :

1. Plusieurs améliorations doivent être faites pour élever le niveau de maturité de la nouvelle solution afin qu'elle soit facilement exploitable par les chercheurs. Présentement, toutes les étapes automatisées du prototype expérimental sont faites via l'exécution des APIs dans la ligne de commande par les chercheurs, ceci nécessite un certain niveau d'expertise technique. Pour ce volet, on recommande les améliorations suivantes sur le prototype expérimental:
    - **Amélioration de la composante de l'entrée des données :** créer une interface visuelle pour la gestion du modèle des données. Une interface qui permette au chercheur d'analyser le contenu du modèle existant et d'ajouter les besoins informationnels requis pour sa recherche, générer une nouvelle version des scripts d'intégration des données pour inclure les nouveaux besoins informationnels ajoutés dans le modèle;
    - **Amélioration de la composante d'intégration des données** :
        - En utilisant les APIs du logiciel, créer une interface visuelle pour gérer l'infrastructure d'intégration de données, démarrer l'intégration des données et valider les données chargées dans le nouveau modèle des données de la recherche;
        - Présentement, le nouveau logiciel supporte deux types de données : les données stockées dans des bases de données relationnelles et les données génétiques stockées dans des fichiers texte respectant la norme du format de

fichier de génotype Oxford (.gen). Comme travaux futurs, on peut envisager d'ajouter des APIs pour supporter d'autres types de formats de données;

- **Amélioration de la composante d'analyse des données :**
  - Ajouter une interface visuelle pour la gestion de l'infrastructure d'analyse des données;
  - Présentement l'interface d'analyse supporte les langages de programmation comme Scala, PySpark et SparkR. On suggère de donner la possibilité au chercheur d'utiliser des logiciels d'analyse spécialisés comme H2O, DataRobot, RapidMinder, Weka, TensorFlow, etc.;
- Le prototype expérimental a été développé en utilisant les services infonuagiques de AWS. Les pilotes de connexion à ces services ont été programmés en dur dans les APIs du prototype. Une des améliorations futures est de développer des interfaces de connexion et mettre les pilotes de connexion dans ces interfaces comme paramètres d'entrée ce qui facilitera la portabilité de la solution à différents fournisseurs infonuagiques;
- Ajouter une nouvelle composante pour gérer les accès et la confidentialité des données des patients utilisées dans les analyses.

2. Introduire la nouvelle approche d'adaptation de modèle et d'intégration des données dans le processus de recherche actuel du laboratoire du Dr Hamet au CRCHUM.

5.6 Travaux futurs

La présente thèse présente plusieurs contributions à la recherche, cependant elle a besoin de plusieurs améliorations :

L'approche d'adaptation de modèle et d'intégration des données doit être validée par plusieurs études de cas;

Les questions de la sécurité d'accès, la gestion de la confidentialité et des implications éthiques de l'utilisation des données personnelles des patients (données cliniques, génétiques, etc.) surtout dans des plateformes infonuagiques sont des questions majeures

qui touchent non seulement la médecine de précision, mais tout le domaine des analyses sur les données du domaine de la santé. Ces questions ont été mises hors de la portée de cette recherche. Il est fortement recommandé de faire une recherche sur ces questions et intégrer les résultats de cette recherche à la nouvelle approche d'adaptation de modèle et d'intégration des données.

ANNEXE I

QUESTIONNAIRE 1 D'ÉVALUATION DU PROCESSUS DE NOUVELLES ANALYSES DE DONNÉES

Q1: In your research, what is the average size of the data samples you use in your data analysis?

Répondant 1	Répondant 2	Répondant 3	Moyenne	Écart-Type
1 GO	6 GO	48 GO	18.33 GO	25.81 GO

Q2: Usually we need many data analysis iteration before obtaining valid results. What is the average data analysis simulation number you used to do to complete a data analysis project?

Répondant 1	Répondant 2	Répondant 3	Moyenne	Écart-Type
>10 Itérations	>10 Itérations	>10 Itérations	>10 Itérations	0 itération

Q3: Usually, before we run a new data analysis iteration, we need to adjust the data sample to add new information in the data model. How frequently you used to adjust the data model of the data sample in a one research?

Répondant 1	Répondant 2	Répondant 3	Moyenne	Écart-Type
>10 fois	>10 fois	>10 fois	>10 fois	0 itération

Q4: Assuming you have already identified the information missing from the data model of the data sample and that are required in the new analysis. What is the time you used to spend to modify the data model to add/adjust the missing information? (Count only the time of modifying the logical data model and generating the physical data model: delay from requesting the data model modification until the delivery of the new data model and ready to be used for the new data analysis iteration).

Répondant 1	Répondant 2	Répondant 3	Moyenne	Écart-Type
<1 heures	2 à 5 heures	>10 jours	18 heures[10]	27.21 heures

Q5: How many people are involved in the process of adjusting the data model (Count only the people involved in modifying the data model: database administrators, system admin, database modeler, developer.).

Répondant 1	Répondant 2	Répondant 3	Moyenne	Écart-Type
4 personnes	1 personne	>5 personnes	3 personnes	2.08 personnes

Q6: Do you create new data repository in each data analysis to add the missing data?

Répondant 1	Répondant 2	Répondant 3	Moyenne	Écart-Type
Non	Oui	Oui	Oui=66.67% Non=33.33%	

Q7: How long it takes to load the missing data into the new data model and obtain a ready to use data sample?

Répondant 1	Répondant 2	Répondant 3	Moyenne	Écart-Type
<1 heure	<1 heure	>10 heures	4 heures	1.73 heures

Q8: In the case of big volume of data, sometimes, we need to upgrade the data analysis infrastructure (add more computers, upgrade the capacity of existing computers, etc.) to import the data from the data sources into the data analysis data model and/or to analyse the big volume of the data samples. During one data analysis cycle, how frequent did you need to upgrade the capacity of your data analysis infrastructure?

Répondant 1	Répondant 2	Répondant 3	Moyenne	Écart-Type
1 fois	1 fois	3 fois	1.66 fois	1.15 fois

[10] 18heures=1heure+3heures+10jours*5heures par jour)

Q9: What is the average time required to upgrade the data analysis infrastructure to be able to analyse the data volume? Count the delay time from requesting the infrastructure upgrade until the delivery time (time the infrastructure is ready to use)

Répondant 1	Répondant 2	Répondant 3	Moyenne	Écart-Type
<1 heure	> 10 heures	> 10 heures	7 heures	5.2 heures

Q10: How many people are involved in upgrading the infrastructure? (all IT resources involved in the infrastructure upgrade)

Répondant 1	Répondant 2	Répondant 3	Moyenne	Écart-Type
3 personnes	1 personne	>5 personnes	3 personnes	2 personnes

ANNEXE II

QUESTIONNAIRE 2 D'ÉVALUATION DU PROCESSUS DE REPRODUCTION DES ANALYSES PRÉCÉDENTES

Q1: Are you able to reproduce previous data analysis at any time?

Répondant 1	Répondant 2	Répondant 3	Moyenne	Écart-Type
Oui	Oui	Non	Oui=66.67% Non=33.33%	

Q2: What efforts are required to reproduce previous analysis? (Count the average time you usually spend in the configuration of the platform and preparing the data back to be able to reproduce the analysis)

Répondant 1	Répondant 2	Répondant 3	Moyenne	Écart-Type
4 heures	2 heures	40 heures	15.33 heures	22.23

Q3: How many people are generally involved in preparing the work environment (infrastructure and data) to reproduce previous analysis?

Répondant 1	Répondant 2	Répondant 3	Moyenne	Écart-Type
1 personne	1 personne	4 personnes	2.33 personnes	1.73 personnes

Q4: In the research data analysis cycle, what is total delay of data preparation phase? (Count time for data model configuration, data conversion, data loading until all the data are stored in the working repository and ready for the data analysis)

Répondant 1	Répondant 2	Répondant 3	Moyenne	Écart-Type
4 heures	77 heures	36 heures	39 heures	36.59 heures

Q5: What is the delay for the data preparation? Count delay for Data model preparation, data conversion, data loading, until the data is ready to be used for data analysis)

Répondant 1	Répondant 2	Répondant 3	Moyenne	Écart-Type
4 heures	5 heures	43 heures	17.33 heures	

Q5: At What percentage the data model configuration step is automated?

Répondant 1	Répondant 2	Répondant 3	Moyenne	Écart-Type
50 %	40%	11%	33%	20%

Q6: At what % the data analysis infrastructure management is automated? (adding and removing resources to the infrastructure, controlling the time of using the infrastructure)

Répondant 1	Répondant 2	Répondant 3	Moyenne	Écart-Type
0%	50%	0%	17%	29%

Q7: Do you use inhouse data analysis infrastructure or pay per usage infrastructure for your data analysis platform?

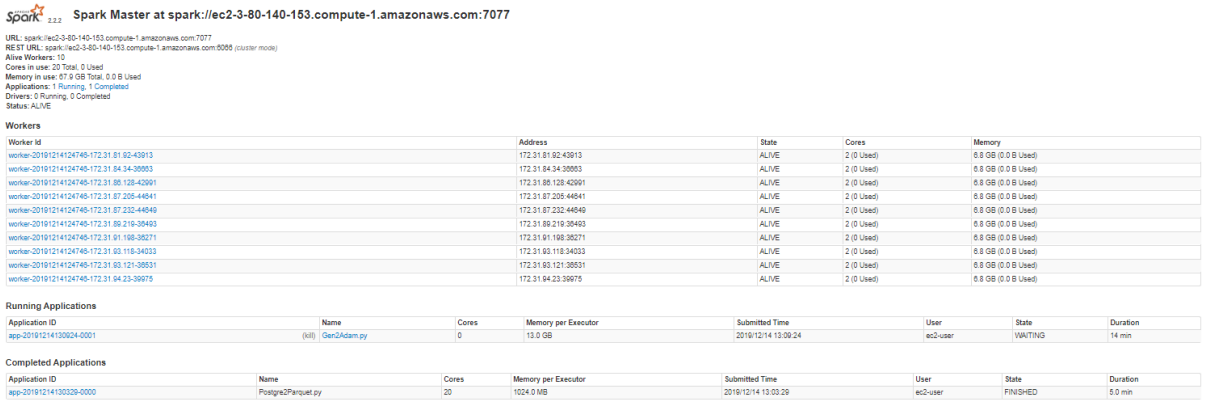
Répondant 1	Répondant 2	Répondant 3	Moyenne	Écart-Type
In house	In house	In house	In house=100%	0 %

ANNEXE III

RÉSULTATS DES TESTS D'ÉVALUATION DU NOUVEAU CYCLE DE RECHERCHE

#	Nb Patient	Nb Fichiers .gen	Nb serveurs virtuels	Type de serveurs	Création et configuration de grappe	Définition et création du modèle des données	Intégration des données cliniques	intégration des données génétiques
1	10	1 Fichiers 1.45 GB	10	t2.large	4min 47sec	27 sec	12 min	14 Min, aucun fichier[11] transféré
2	10	1 Fichiers 1.45 GB	2	m4.4xlarge	3min 56 sec	28 Sec	1.8 min	54 sec

Test 1



Spark 2.2.2 **Spark Master at spark://ec2-3-80-140-153.compute-1.amazonaws.com:7077**

URL: spark://ec2-3-80-140-153.compute-1.amazonaws.com:7077
REST URL: spark://ec2-3-80-140-153.compute-1.amazonaws.com:6066 *(cluster mode)*
Alive Workers: 10
Cores in use: 20 Total, 0 Used
Memory in use: 67.9 GB Total, 0.0 B Used
Applications: 1 Running, 1 Completed
Drivers: 0 Running, 0 Completed
Status: ALIVE

Workers

Worker Id	Address	State	Cores	Memory
worker-20191214124746-172.31.81.92-43913	172.31.81.92:43913	ALIVE	2 (0 Used)	6.8 GB (0.0 B Used)
worker-20191214124746-172.31.84.34-38663	172.31.84.34:38663	ALIVE	2 (0 Used)	6.8 GB (0.0 B Used)
worker-20191214124746-172.31.86.128-42991	172.31.86.128:42991	ALIVE	2 (0 Used)	6.8 GB (0.0 B Used)
worker-20191214124746-172.31.87.205-44641	172.31.87.205:44641	ALIVE	2 (0 Used)	6.8 GB (0.0 B Used)
worker-20191214124746-172.31.87.232-44649	172.31.87.232:44649	ALIVE	2 (0 Used)	6.8 GB (0.0 B Used)
worker-20191214124746-172.31.89.219-39493	172.31.89.219:39493	ALIVE	2 (0 Used)	6.8 GB (0.0 B Used)
worker-20191214124746-172.31.91.198-38271	172.31.91.198:38271	ALIVE	2 (0 Used)	6.8 GB (0.0 B Used)
worker-20191214124746-172.31.93.118-34033	172.31.93.118:34033	ALIVE	2 (0 Used)	6.8 GB (0.0 B Used)
worker-20191214124746-172.31.93.121-38531	172.31.93.121:38531	ALIVE	2 (0 Used)	6.8 GB (0.0 B Used)
worker-20191214124746-172.31.94.23-39975	172.31.94.23:39975	ALIVE	2 (0 Used)	6.8 GB (0.0 B Used)

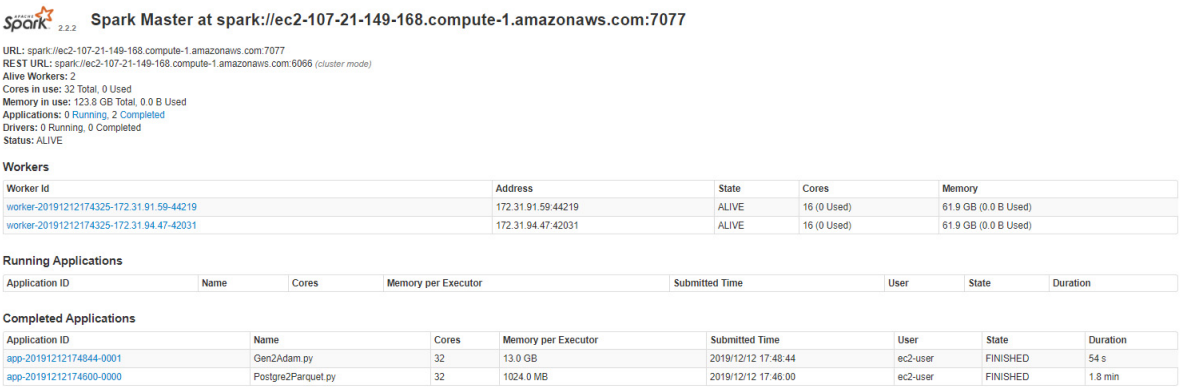
Running Applications

Application ID	Name	Cores	Memory per Executor	Submitted Time	User	State	Duration
app-20191214130924-0001 (kill)	Gen2Adam.py	0	13.0 GB	2019/12/14 13:09:24	ec2-user	WAITING	14 min

Completed Applications

Application ID	Name	Cores	Memory per Executor	Submitted Time	User	State	Duration
app-20191214130329-0000	PostgreSQL2Parquet.py	20	1024.0 MB	2019/12/14 13:03:29	ec2-user	FINISHED	5.0 min

[11] Il semble que le serveur virtuel choisi n'avait pas suffisamment de mémoire pour traiter le fichier génétique



Spark Master at spark://ec2-107-21-149-168.compute-1.amazonaws.com:7077

URL: spark://ec2-107-21-149-168.compute-1.amazonaws.com:7077
REST URL: spark://ec2-107-21-149-168.compute-1.amazonaws.com:6066 *(cluster mode)*
Alive Workers: 2
Cores in use: 32 Total, 0 Used
Memory in use: 123.8 GB Total, 0.0 B Used
Applications: 0 Running, 2 Completed
Drivers: 0 Running, 0 Completed
Status: ALIVE

Workers

Worker Id	Address	State	Cores	Memory
worker-20191212174325-172.31.91.59-44219	172.31.91.59:44219	ALIVE	16 (0 Used)	61.9 GB (0.0 B Used)
worker-20191212174325-172.31.94.47-42031	172.31.94.47:42031	ALIVE	16 (0 Used)	61.9 GB (0.0 B Used)

Running Applications

Application ID	Name	Cores	Memory per Executor	Submitted Time	User	State	Duration

Completed Applications

Application ID	Name	Cores	Memory per Executor	Submitted Time	User	State	Duration
app-20191212174844-0001	Gen2Adam.py	32	13.0 GB	2019/12/12 17:48:44	ec2-user	FINISHED	54 s
app-20191212174600-0000	Postgre2Parquet.py	32	1024.0 MB	2019/12/12 17:46:00	ec2-user	FINISHED	1.8 min

ANNEXE IV

DÉFINITION DES TERMES

Adaptation de modèle de données	Personnalisation et ajustement du modèle des données d'une base de données afin d'inclure tout nouveau besoin informationnel requis pour toute nouvelle analyse.
Cycle de recherche	Ce terme représente l'aspect itératif des activités de recherche en médecine de précision. Un cycle initial de recherche inclut les activités suivantes: spécifier une hypothèse, identifier des données, charger ces données dans une base de données, faire une analyse, obtenir des résultats, comparer les résultats à l'hypothèse de départ. Quand le chercheur n'atteint pas son objectif, il fait un ou plusieurs autres cycles additionnels de recherche : ajuster l'hypothèse, identifier d'autres données, adapter le modèle des données de la base de données aux nouveaux besoins informationnels, charger ces nouvelles données, effectuer une analyse, obtenir des résultats, comparer à l'hypothèse de départ.
Capacité de traitement	C'est la puissance de traitement d'un serveur virtuel. Elle est évaluée en fonction des composantes matérielles du serveur virtuel : le nombre de VCPU, la taille de mémoire vive, du type du disque dure, etc.
Délais	Le temps nécessaire pour réaliser une activité. Par exemple, le délai de réalisation de recherche, c'est le temps encouru global depuis la définition de la recherche jusqu'à l'obtention des réponses aux questions de recherche, le délai

	d'adaptation de modèle de données, est le temps encouru depuis la demande d'adaptations jusqu'à l'obtention d'un nouveau modèle adapté.
Données initiales	Ce sont les données sources dans leurs formats brutes
Données génotypiques	C'est les données de la composition génétiques d'un individu (composition allélique de tous les gènes de cet individu).
Échantillon des données	C'est l'ensemble des données identifiées pour être utilisées dans la réalisation d'une itération de recherche.
Échelonnable	Ce terme (ou le terme échelonnabilité) fait référence à la capabilité d'ajuster la capacité de traitement de la plateforme matérielle du processus d'analyse à la taille des bases de données utilisées.
Étape d'analyse des données	Cette étape fait référence aux tâches de nettoyage des données et à l'exploration de ces données.
Grappe	(Computer cluster en anglais). Est une technique qui consiste à regrouper plusieurs serveurs virtuels indépendants appelés nœuds afin de permettre une gestion globale de la capacité de traitements.
Intégration des données	Importation des données provenant de différentes sources dans un seul endroit et unification du format du stockage de ces données.
Ordinateur	Ce terme est utilisé dans cette thèse pour référencer un ordinateur de bureau ou bien un ordinateur portable utilisé par les chercheurs et bio-informaticiens pour réaliser leurs recherches.

Préparation des données	Ce terme fait référence aux étapes d'adaptation de modèle et d'intégration des données.
Processus d'analyse des données	Ce terme fait référence à la somme des étapes de préparation des données et celle d'analyse des données.
Modèle des données	Une illustration de la structure logique de la base de données plus les relations et les contraintes qui déterminent comment les données vont être stockées et accédées. Dans cette thèse ce terme réfère à la représentation logique plus la représentation physique du modèle (les fichiers binaires qui contiennent la définition textuelle des données).
Nœud esclave	Un serveur virtuel faisant partie d'une grappe Spark de serveurs virtuels et qui est utilisé pour exécuter les tâches assignées par le nœud maître de la grappe.
Nœud maître/Nœud master	Un serveur virtuel faisant partie d'une grappe Spark de serveur virtuel. Il contient une application appelée gestionnaire de ressources que le logiciel Spark utilise pour distribuer la charge de travail entre les autres serveurs virtuels qui composent la grappe.
Reproductibilité	La reproductibilité (synonyme de reproduction) d'une analyse de médecine de précision est une des conditions qui permettent d'inclure les observations réalisées durant une expérience dans le processus d'amélioration perpétuelle des connaissances scientifiques. Cette condition part du principe qu'on ne peut tirer de conclusions que d'une recherche bien décrite, qui est apparue plusieurs fois et reproduisible par des personnes différentes.
Serveur virtuel	Machine virtuelle dans un service infonuagique, par exemple une instance EC2 chez Amazon Web Services.

Serveur central	Est un ordinateur avec une grande capacité de traitement qui centralise plusieurs services utilisés en mode partagé par les bio-informaticiens.
Schéma de données	Appelée aussi “schéma de base de données” est la représentation visuelle d'une base de données entière. La représentation visuelle illustre comment les données qui composent la base de données se rapportent les uns aux autres.
Sous-schéma de données	Un sous-schéma de données est une partie d’un schéma de données. Il illustre la représentation visuelle d’un groupe de données qui hébergent les données d’un domaine précis. Exemple : schéma d’une base de données médicale peut-être composé de : sous-schémas des données cliniques, sous-schémas des données de laboratoires, sous-schémas des données d’imagerie médicale, etc.
Source(s) des données	Les bases de données d’où proviennent toutes les données à utiliser dans les recherches : bases de données des systèmes des Dossiers des Patients informatisés, des cohortes des données de santé, base des données des tests de laboratoire, base des données génétiques, etc.
Transformation des données	Ce terme fait référence à toute modification apportée sur le modèle des données : ajout, suppression ou modification de structure de données dans le modèle. C’est la fonction d’adaptation de modèle des données.

BIBLIOGRAPHIE

Acimovic, J., Auffray, C., Ballestrero, A. et coll. (2014). THE CASyM ROADMAP Implementation of Systems Medicine across Europe, version 2.0, *CASyMconsortium*, Dublin, Ireland, 31 p. En ligne : https://www.casym.eu/lw_resource/datapool/_items/item_394/casym_roadmap_2.pdf.

Andany, J., Léonard, M. et Palisser, C. (1991). Management Of Schema Evolution In Databases. *Proceedings of the 17th International Conference on Very Large Data Bases*, Barcelona, September, pp. 161–170. En ligne : http://www.inf.ufpr.br/eduardo/ensino/ci763/papers/vldb91-andany.pdf.

Baro, E., Degoul, S., Beuscart, R. et Chazard, E. (2015). Review Article Toward a Literature-Driven Definition of Big Data in Healthcare. *BioMed Research International*, vol. 2015:639021:1–9. En ligne: https://doi.org/10.1155/2015/639021.

Belghait, F. et April, A. (2018). The Future of Large-Scale Precision Medicine Research Platforms: Preparing the Data for Analysis. *Special Issue Published in International Journal of Trend in Research and Development (IJTRD)*, ISSN: 2394-9333, Special Issue | ICTIMESH-18, December 2018:91–94. En ligne: http://www.ijtrd.com/papers/IJTRD19226.pdf.

Belghait, F., Kanzki, B. et April, A. (2018). ADAM Genomics Schema - extension for precision medicine research. *DH '18 Proceedings of the 2018 International Conference on Digital Health*, Lyon, France, April 23-26, pp. 1–4. En ligne: https://doi.org/10.1145/3194658.3194669.

Belghait, F., April, A., Hamet, P., Tremblay, J., & Desrosiers, C. (2019). A Large-scale and Extensible Platform for Precision Medicine Research. Proceedings of the 9th International Conference on Digital Public Health, November 2019, pp. 47–54. https://doi.org/10.1145/3357729.3357742.

Bhuvaneshwar, K., Belouali, A., Singh, V. et coll. (2016). G-DOC Plus - An integrative bioinformatics platform for precision medicine. *BMC Bioinformatics*, 17(1):1–13. En ligne: https://doi.org/10.1186/s12859-016-1010-0.

Canuel, V., Rance, B., Avillach, P. et coll. (2015). Translational research platforms integrating clinical and omics data: A review of publicly available solutions. *Briefings in Bioinformatics*, 16(2):280–290. En ligne: https://doi.org/10.1093/bib/bbu006.

Castelli, D. (1998). A strategy for reducing the effort for database schema maintenance. Proceedings of the Second Euromicro Conference on Software Maintenance and Reengineering, Florence, Italy, pp. 29–35. En ligne: https://doi.org/10.1109/CSMR.1998.665729.

Cabot, C., Soualmia, L.F. et Darmoni S.J. (2015). Intégration de données cliniques et omiques pour la recherche d'information dans le Dossier Patient Informatisé. In C. AFIA (Ed.), Collection AFIA. Journées Francophones d'Ingénierie des Connaissances - IC 2015, Rennes, France, juillet 2015, hal-01179292. En ligne : https://www.researchgate.net/publication/280066011_Integration_de_donnees_cliniques_et_omiques_pour_la_recherche_d'information_dans_le_Dossier_Patient_Informatise.

Chen, X. et Ishwaran, H. (2013). Random Forests for Genomic Data Analysis. *Genomics*, 99(6):323–329. En ligne: https://doi.org/10.1016/j.ygeno.2012.04.003.

Chung, S.Y. et Wong, L. (1999). Kleisli: a new tool for data integration in biology. *Trends in Biotechnology*, 17(9):351–355. En ligne: https://doi.org/10.1016/S0167-7799(99)01342-6.

Chute, C.G., Ullman-Cullere, M., Wood, G.M. et coll. (2014). Some experiences and opportunities for big data in translational research. *Genetics in Medecine*, 15(10):802–809. En ligne: https://doi.org/10.1038/gim.2013.121.

Collins, F.S. et Mansoura, M.K. (2001). The Human Genome Project. Revealing the shared inheritance of all humankind. *7th Biennial Symposium on Minorities, the Medically Underserved and Cancer*, 91(February 2001):221–225. En ligne: https://www.researchgate.net/publication/12180013_The_Human_Genome_Project_Revealing_the_shared_inheritance_of_all_humankind.

D'Alessandro, L. A., Meyer, R. et Klingmüller, U. (2013). Hepatocellular carcinoma: a systems biology perspective. *Frontiers in Physiology*, February, vol. 4:1–6. En ligne: https://doi.org/10.3389/fphys.2013.00028.

Databricks. (2019). Unified Analytics Platform for Genomics. En ligne: https://databricks.com/product/genomics.

Davis-Turak, J., Courtney, S.M., Hazard, E.S. et coll. (2017). Genomics pipelines and data integration: challenges and opportunities in the research setting. *Expert Review of Molecular Diagnostics*, 17(3):225–237. En ligne: https://doi.org/10.1080/14737159.2017.1282822. (Consulté le 27 mars 2020).

Deloitte. (2019). Deloitte and Vineti Will Team on End-to-End Solution to Support Personalized Medicine. Contact PR Newswire, En ligne:

https://www.prnewswire.com/news-releases/deloitte-and-vineti-will-team-on-end-to-end-solution-to-support-personalized-medicine-300706127.html.

Díaz-Uriarte, R. et Alvarez de Andrés, S. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, 7(3):1–13. En ligne: https://doi.org/10.1186/1471-2105-7-3.

DNANexus. (2019). Integrate genomic data with other clinical data, including electronic medical records. DNANEXUC inc. En ligne: https://www.dnanexus.com/.

Drongelen, I.K. (2001). The iterative theory-building process: rationale, principles and evaluation. *Management Decision*, 39(7):503–512. En ligne: https://www.emerald.com/insight/content/doi/10.1108/EUM0000000005799/full/html. (Consulté le 27 mars 2020).

Dubitzky, W., Krebs, O. et Eils, R. (2001). Minding, OLAPing, and Mining Biological Data: Towards a Data Warehousing Concept in Biology. Proceedings of Network Tools and Applications in Biology (*NETTAB)*, May 17-18, Advance Biology Center, Genova, pp. 78–82. En ligne: https://www.semanticscholar.org/paper/Minding%2C-OLAPing%2C-and-Mining-Biological-Data%3A-a-in-Dubitzky-Krebs/32b132f2036f571d61309492cedd29a335727a9d.

Duffy, D. (2016). Problems, challenges and promises: perspectives on precision medicine. Briefings in Bioinformatics, 17(3):494–504. En ligne: https://doi.org/10.1093/bib/bbv060.

Ellis, J.W., Chen, M-H., Foster, M.C. et coll. (2012). Validated SNPs for eGFR and their associations with albuminuria. *Human Molecular Genetics*, 21(14):3293–3298. En ligne: https://doi.org/10.1093/hmg/dds138.

Endo, A., Shibata, T. et Tanaka, H. (2008). *Comparison of Seven Algorithms to Predict Breast Cancer Survival*. International Journal of Biomedical Soft Computing and Human Sciences, 13(2):11–16. En ligne: https://doi.org/10.24466/ijbschs.13.2_11.

Etzold, T., Ulyanov, A. et Argos, P. (1996). SRS: Information retrieval system for molecular biology data banks. In *Methods in Enzymology*, vol. 266:114–128. En ligne: https://doi.org/10.1016/S0076-6879(96)66010-8.

Feldman, A.M. (2015). Bench-to-Bedside; Clinical and Translational Research; Personalized Medicine; Precision Medicine-What's in a Name? *Clinical and Translational Science*, 8(3):171–173. En ligne: https://doi.org/10.1111/cts.12302.

Foundation Apache Software. (2019). Apache Hadoop. Retrieved October 17, 2019. En

ligne: https://hadoop.apache.org/.

Galperin, M.Y. (2004). The Molecular Biology Database Collection: 2004 update. *Nucleic Acids Research*, 32(90001):3–22. En ligne: https://doi.org/10.1093/nar/gkh143.

Garrido, P., Aldaz, A., Calleja, M. Á. et coll. (2017). Proposal for the Creation of a National Strategy for Precision Medicine in Cancer: A position statement of SEOM, SEAP and SEFH. *Farmacia Hospitalaria*, 41(6):688-691. En ligne: https://doi.org/10.7399/fh.10877.

Gullapalli, R., Lyons-Weiler, M., Petrosko, P. et coll. (2012). Clinical Integration of Next Generation Sequencing Technology. *NIH Public Access*, 32(4):585–599. En ligne: https://doi.org/10.1016/j.cll.2012.07.005.

Haas, L. M., Schwarz, P. M., Kodali, P. et coll. (2001). DiscoveryLink: A system for integrated access to life sciences data sources. *IBM Systems Journal DOI - 10.1147/Sj.402.0489*, 40(2):489–511. En ligne: https://doi.org/10.1147/sj.402.0489.

Hakonarson, H., Gulcher, J. J. R. et Stefansson, K. (2003). deCODE genetics, Inc. *Pharmacogenomics*, 4(2):209–215. En ligne: https://doi.org/10.1517/phgs.4.2.209.22627.

Hamet, P., Haloui, M., Harvey, F. et coll. (2017). PROX1gene CC genotype as a major determinant of early onset of type 2 diabetes in slavic study participants from Action in Diabetes andVascular Disease: Preterax and Diamicron MR Controlled Evaluation study. *Journal of Hypertension*, vol. 35:24–32. En ligne: https://doi.org/10.1097/HJH.0000000000001241.

Heller, S. R. (2009). A summary of the ADVANCE Trial. *Diabetes Care*, vol. 32:1–5. https://doi.org/10.2337/dc09-S339.

Ibrahim-Verbaas, C. A., Fornage, M., Bis, J. C. et coll. (2014). Predicting stroke through genetic risk functions: The CHARGE risk score project. *Stroke*, 45(2):403–412. https://doi.org/10.1161/STROKEAHA.113.003044.

Ideker, T., Thorsson, V., Ranish, J. A. et coll. (2001). Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*, vol.292:929–934. En ligne: https://doi.org/10.1126/science.292.5518.929.

Illumina. (2019). Illumina DRAGEN Bio-IT Platform. Retrieved March 17, 2019. En ligne: https://www.illumina.com/products/by-type/informatics-products/dragen-bio-it-platform.html?scid=2018-269ECL299.

Jacquet, J. P. et Abran, A. (1997). From software metrics to software measurement methods:

A process model. *Proceedings of the IEEE International Software Engineering Standards Symposium*, (July 1997), pp.128-135. En ligne: https://doi.org/10.1109/sess.1997.595954.

Jameson, L.J. et Longo, D. (2015). Precision Medicine—Personalized, Problematic, and Promising. *Obstetrical & Gynecological Survey*, 70(10):612–614. En ligne: https://doi.org/10.1097/01.ogx.0000472121.21647.38.

Jamoon, E., Ninee, Y., et Hing, E. (2016). Adoption of Certified Electronic Health Record Systems and Electronic Information Sharing in Physician Offices: United States, 2013 and 2014. *NCHS Data Brief*, 236(February):1–8. En ligne: https://www.ncbi.nlm.nih.gov/pubmed/26828707.

Jang, Y., Choi, T., Kim, J. et coll. (2018). An integrated clinical and genomic information system for cancer precision medicine. *BMC Medical Genomics*, 11(Suppl 2):96–116. En ligne: https://www.ncbi.nlm.nih.gov/pubmed/29697362.

Karen He, Dongliang Ge (2017). Big data analytics for genomic medicine. *International Journal of Molecular Sciences,* 18(2):1–18. https://doi.org/10.3390/ijms18020412.

Kourou, K., Exarchos, T.P., Exarchos, K.P. et coll. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, vol. 13:8–17. En ligne: https://doi.org/10.1016/j.csbj.2014.11.005.

Kupfer, A., Eckstein, S., Neumann, K. et Mathiak, B. (2006). Keeping Track of Changes in Database Schemas and Related Ontologies. *7th International Baltic Conference on Databases and Information Systems*, Vilnius, Lithuania, July 3-6, pp. 63–68. En ligne : https://www.researchgate.net/publication/224644290_Keeping_track_of_changes_in_database_schemas_and_related_ontologies.

Lelong, R., Merabti, T., Grosjean, J., Joulakian, M.B. et coll. (2014). Moteur de recherche sémantique au sein du dossier du patient informatisé : Langage de requêtes spécifique, Journées francophones d'informatique médicale, Fès, Maroc, 12-13 juin, pp. 139–151. En ligne: http://ceur-ws.org/Vol-1379/paper-12.pdf.

Lescourret, F., Genest, M., Barnouin, J. et coll. (1993). Data Modeling for Database Design in Production and Health Monitoring Systems for Dairy Herds. *Journal of Dairy Science*, 76(4):1053–1062. https://doi.org/10.3168/jds.S0022-0302(93)77434-2.

Liao, J.G., et Chin, K-V. (2007). Logistic regression for disease classification using microarray data: model selection in a large p and small n case. *Bioinformatics*, 23(15):1945–1951. En ligne: https://doi.org/10.1093/bioinformatics/btm287.

Londin, E.R., et Barash, C.I. (2015). Applied & Translational Genomics What is translational bioinformatics? *ATG*, vol. 6:1–2. En ligne: https://doi.org/10.1016/j.atg.2015.08.003.

Massie, M., Nothaft, F., Hartl, C. et coll. (2013). ADAM: Genomics Formats and Processing Patterns for Cloud Scale Computing, Electrical Engineering and Computer Sciences University of California at Berkeley, Technical Report No. UCB/EECS-2013207. En ligne: http://www.eecs.berkeley.edu/Pubs/TechRpts/2013/EECS-2013-207.html.

McKenna, A., Hanna, M., Banks, E. et coll. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, vol. 20:1297–1303.En ligne : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2928508/.

Meenu, D., et Jahangir, K. (2017). Identifying big data dimensions and structure. *2017 4th International Conference on Signal Processing, Computing and Control (ISPCC)*, Sept 21-23, Himachal Pradesh, India, pp. 163–168. En ligne: https://doi.org/10.1109/ISPCC.2017.8269669.

Mell, P. et Grance,T. (2011). The NIST Definition of Cloud Computing: Recommendations of the National Institute of Standards and Technology. In *National Institute of Standards and Technology*, Special Publication 800-145. 3 p. En ligne: https://doi.org/10.1136/emj.2010.096966.

Murphy, S. N., Mendis, M. E., Berkowitz, D. A. et coll. (2006). Integration of clinical and genetic data in the i2b2 architecture. *AMIA Annu Symp Proc*, (2), 1040. En ligne : https://www.researchgate.net/publication/6563736_Integration_of_Clinical_and_Genetic_Data_in_the_i2b2_Architecture.

Niemenmaa, M., Kallio, A., Schumacher, A. et coll. (2012). Hadoop-BAM: Directly manipulating next generation sequencing data in the cloud. *Bioinformatics*, 28(6):876–877. En ligne: https://doi.org/10.1093/bioinformatics/bts054.

O'Driscoll, A., Daugelaite, J. et Sleator, R.D. (2013). "Big data", Hadoop and cloud computing in genomics. *Journal of Biomedical Informatics*, 46(5):774–781. En ligne: https://doi.org/10.1016/j.jbi.2013.07.001.

Orechia, J., Pathak, A., Shi, Y. et coll. (2015). Applied & Translational Genomics OncDRS: An integrative clinical and genomic data platform for enabling translational research and precision medicine. *ATG*, vol. 6:18–25. https://doi.org/10.1016/j.atg.2015.08.005.

Patel, A., Chalmers, J., et Poulter, N. (2005). ADVANCE: Action in diabetes and vascular

disease. *Journal of Human Hypertension*, vol. 19:27–32. En ligne: https://doi.org/10.1038/sj.jhh.1001890.

Plase, D., Niedrite, L. et Taranovs, R. (2017). A Comparison of HDFS Compact Data Formats: Avro Versus Parquet. Mokslas - Lietuvos Ateitis, 9(3):267–276. En ligne: https://doi.org/10.3846/mla.2017.1033.

Prosperi, M., Min, J. S., Bian, J. et Modave, F. (2018). *Big data hurdles in precision medicine and precision public health*. vol.2:1–15. En ligne : https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-018-0719-2.

Reenders, K., de Nobel, E., van den Hoogen, H. J. et coll. (1993). Diabetes and its Long-term Complications in General Practice: a Survey in a Well-defined Population. *Family Practice*, 10(2):169–172. En ligne: http://dx.doi.org/10.1093/fampra/10.2.169.

S3 Browser. (2019). En ligne: https://s3browser.com/. (Consulté le 27 mars 2020).

Sandve, G.K., Nekrutenko, A., Taylor, J. et Hovig, E. (2013). Ten Simple Rules for Reproducible Computational Research. *PLOS Computational Biology*, 9(10):1–4. En ligne: https://doi.org/10.1371/journal.pcbi.1003285.

Scheufele, E., Aronzon, D., Coopersmith, R. et coll. (2014). tranSMART: An Open Source Knowledge Management and High Content Data Analytics Platform. *AMIA Joint Summits on Translational Science Proceedings*, April 7th, 2014:96–101. En ligne : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4333702/.

Shimokawa, K., Mogushi, K., Shoji, S., Hiraishi, A., Ido, K., Mizushima, H., et Tanaka, H. (2010). ICOD: An integrated clinical omics database based on the systems-pathology view of disease. *BMC Genomics*, 11(SUPPL.4):1–7. En ligne: https://doi.org/10.1186/1471-2164-11-S4-S19.

Soualmia, L. F., Darmoni, J., et Soualmia, L. F. (2015). Intégration de données cliniques et omiques pour la recherche d'information dans le Dossier Patient Informatisé. pp. 1–12. En ligne : https://hal.archives-ouvertes.fr/hal-01179292.

Srivastava, P., et Hopwood, N. (2009). A Practical Iterative Framework for Qualitative Data Analysis. *International Journal of Qualitative Methods*, 8(1):76–84. En ligne: https://doi.org/10.1177/160940690900800107.

Stastny, J. et Skorpil, V. (2007). Genetic Algorithm and Neural Network. *Proceedings of the 7th WSEAS International Conference on Applied Informatics and Communications*, vol. 7:345–349. En ligne: https://link.springer.com/chapter/10.1007/978-1-4615-2353-6_8.

Sven-Eric Lautemann. (1997). Schema Versions in Object-Oriented Database Systems. *Proceedings of the Fifth International Conference on Database Systems for Advanced Applications*, Melbourne, Australia, April 1-4, pp. 323–332. En ligne : https://www.worldscientific.com/doi/abs/10.1142/9789812819536_0034.

Tan, A., Tripp, B. et Daley, D. (2011). BRISK-research-oriented storage kit for biology-related data. *Bioinformatics*, 27(17):2422–2425. En ligne: https://doi.org/10.1093/bioinformatics/btr389.

The Apache Software Foundation. (2012). Apache Avro™ 1.8.1 Documentation. En ligne: https://avro.apache.org/docs/1.8.1/.

The Apache Software Foundation. (2018). Apache Spark™ is a unified analytics engine for large-scale data processing. En ligne: https://spark.apache.org/.

The Apache Software Foundation. (2018). Apache Avro™ 1.8.2 IDL. En ligne: https://avro.apache.org/docs/1.8.2/idl.html#arrays.

Turner, S. T., Schwartz, G. L. et Boerwinkle, E. (2007). *Personalized Medicine for High Blood Pressure Rationale for Personalized Medicine*. 50(1):1–5. En ligne: https://doi.org/10.1161/HYPERTENSIONAHA.107.087049.

Vassy, J. L., Korf, B. R. et Robert C. Green. (2015). How to know when physicians are ready for genomic medicine. *Sci Transl Med*, 344(6188):1173–1178. En ligne: https://www.ncbi.nlm.nih.gov/pubmed/25971999.

Warram, J. H., Gearin, G., Laffel, L., et Krolewski, A. S. (1996). Effect of duration of type I diabetes on the prevalence of stages of diabetic nephropathy defined by urinary albumin/creatinine ratio. *Journal of the American Society of Nephrology*, 7(6):930–937. En ligne: http://jasn.asnjournals.org/content/7/6/930.short.

Wolkenhauer, O., Auffray, C., Brass, O. et coll. (2014). Enabling multiscale modeling in systems medicine. *Genome Medicine*, 6(3):1–3. En ligne: https://doi.org/10.1186/gm538.

Xue, Y., Lameijer, E. W., Ye, K. et coll. (2016). Precision Medicine: What Challenges Are We Facing? *Genomics, Proteomics and Bioinformatics*, 14(2016):253–261. En ligne: ps://doi.org/10.1016/j.gpb.2016.10.001.

Zaharia, M., S. Xin, R., Wendell, P. et coll. (2016). Apache Spark: A unified engine for big data processing. In *Communications of the ACM*. Novembre, 59(11):56–65. En ligne: https://doi.org/10.1145/2934664.