Le génie pour l'industrie

RAPPORT TECHNIQUE
PRÉSENTÉ À L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
DANS LE CADRE DU COURS GTI792.

MODÉLISATION ET CONCEPTION D'UNE BASE DE DONNÉES BIGDATA (IMPALA). RAPPORT D'ANALYSES DE LA PUISSANCE DES SIGNAUX DES OPÉRATEURS MOBILES.

RONALD LESSAGE
LESR16128802

DÉPARTEMENT DE GÉNIE LOGICIEL ET DES TI

Professeur-superviseur

Alain April

MONTRÉAL, 09 DÉCEMBRE 2014
RONALD AUTOMNE 2014

REMERCIEMENTS.

Mes remerciements s'adressent en premier lieu à mon superviseur, le professeur Alain April, pour sa confiance et ses conseils qui m'ont permis de progresser sans cesse durant la réalisation de ce projet. Je remercie toute l'équipe pédagogique de l'école technologie supérieure pour m'avoir donné des outils qui m'ont permis d'assurer la partie théorie de ce projet. Je tiens à remercier tout particulièrement et à témoigner toute ma reconnaissance à Monsieur Thierry Marechal créateur de Snoobe pour l'expérience enrichissante et pleine d'intérêt qu'il m'a fait vivre durant la réalisation de ce projet. Je remercie également, Monsieur Olivier Mirandette développeur et enseignant en BigData–temps réel. Et enfin, je tiens à remercier Sébastien Bonami et David Lauzon pour leur soutien et conseil.

ANALYSE DE LA PUISSANCE DES SIGNAUX DES OPÉRATEURS MOBILES GRÂCE À LA TECHNOLOGIE IMPALA.

RONALD LESSAGE
LESR16128802

RÉSUMÉ

Snoobe est un logiciel (c.-à-d. une application mobile) gratuit de comparaison de forfaits téléphoniques mobiles pour la technologie Androïde. Ce logiciel a été créé par la force et la détermination de Thierry Maréchal, le fondateur de Snoobe, qui un jour a découvert qu'il paye un forfait qui n'existe plus sur le marché depuis 3 ans sans en être informé. Après avoir corrigé la situation, il pense qu'il n'est pas le seul à faire face à ce genre de problème et qu'il existe probablement une multitude d'utilisateurs, comme lui, qui paie trop pour des services dont ils n'ont pas besoin ou qu'ils n'utilisent que rarement. Des millions de dollars sont perdus par cette clientèle, et ce au profit des fournisseurs de services.

Ce projet, d'initiation au domaine du BigData a été ardu. Plusieurs obstacles se sont présentés, mais grâce à l'encouragement, la détermination et surtout l'envie d'apprendre une nouvelle technologie, j'ai réussi à aller jusqu'au bout. J'ai réalisé, à travers cette expérience, combien il est parfois difficile de cerner avec précision les besoins d'un client. La meilleure façon de comprendre et répondre aux besoins d'un client est de travailler en collaboration avec lui et de le faire participer à la réalisation du projet**.** Afin de réaliser ce projet, il y a eu un grand nombre de rencontres, avec Thierry, afin de préciser un cas d'utilisation BigData qui sera décrit dans les chapitres suivants. Je tiens à réitérer mes remerciements à Thierry qui était toujours disponibles pour répondre à mes questions.

L'objectif de ce projet

La figure 1, qui suit, présente une vue d'ensemble du cas d'utilisation proposé pour ce projet. On peut y voir que pour une localisation précise il est possible d'obtenir la puissance du signal de tous les opérateurs mobiles disponibles.

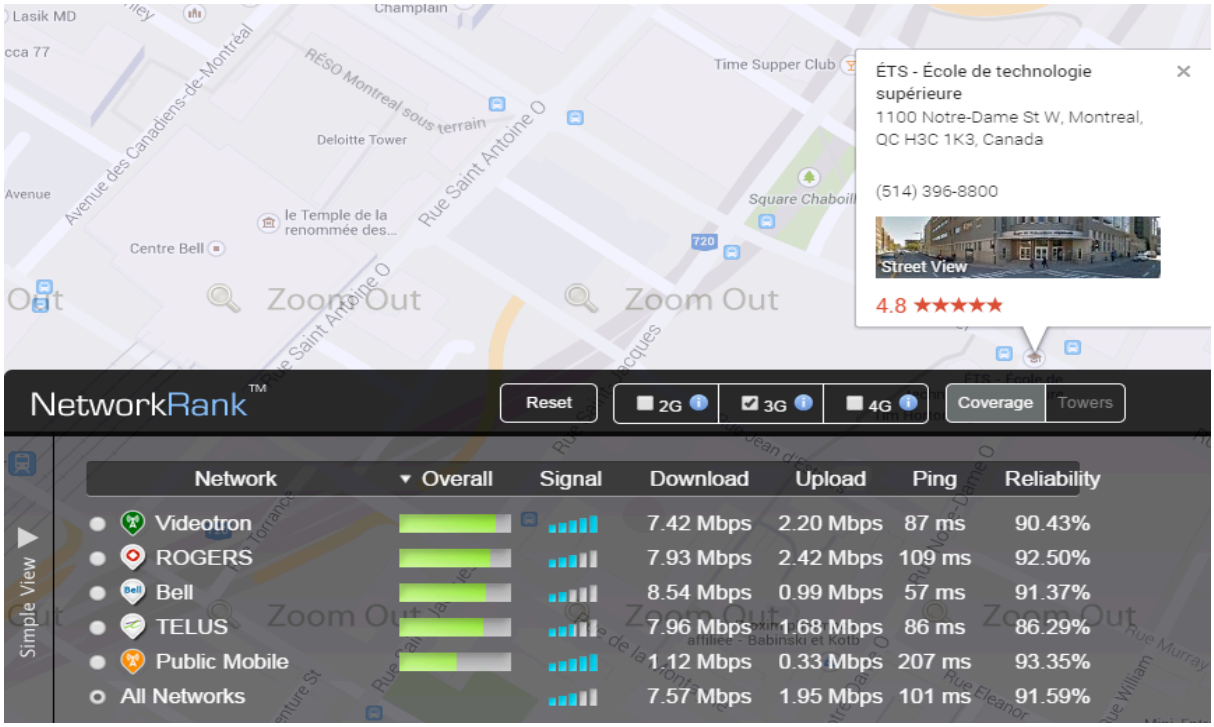


Figure 1: Prototype du projet à réaliser. [1]

Mon projet de fin d'études consiste à :

- Préciser un cas d'utilisation BigData,
- Identifier et installer la technologie BigData en vue de la réalisation d'un prototype;
- Concevoir le schéma NoSQL, créer des données simulées et charger la base de données Impala.

TABLE DES MATIÈRES

Page

LISTE DES TABLEAUX

LISTE DES FIGURES

Page

LISTE DES ABRÉVIATIONS, SIGLES ET ACRONYMES

Alterysx	Est une société de logiciels Américain fournit des produits permettant l'analyse avancées sur des données [2].
API	Application Programming Interface.
BigData	D'écrire un volume massif de données structurées et non structurées, qui est si grand, qu'il est difficile de traiter en utilisant des techniques de base de données et de logiciels traditionnels [3].
CDH	Distribution de Hadoop par Cloudera «Cloudera's Distribution Including Apache Hadoop».
CLI	Interface de ligne de commande.
Deflate	Algorithme LZ77 et codage de Huffman
Download	Télécharger ou recevoir des données d'un système local à partir d'un système distant,
HBase	Base de données de Hadoop « Hadoop Data Base».
HDFS	Hadoop Distributed File system
HUE	Hadoop User Expérience Interface utilisateur utilisé pour développer des Applications Hadoop.
IMPALA	Moteur de recherche et d'analyse des données stockées dans HDFS et Hbase.
JVM	Machine virtuelle de Java « Java Virtual Machine».
JSON	JavaScript Object Notation.
Microstrategy	Chef de file du marché international de la Business Intelligence, fournit des logiciels intégrés de reporting, d'analyse et de pilotage qui permettent aux entreprises d'analyser les données disséminées dans leur structure afin de prendre des décisions métier plus efficace [4].
Pentaho	Société qui fournit des logiciels libres permettant l'intégration et l'analyse des données.
Prepaid	Plan de service mobile prépayé sans contrat.
Postpaid	Plan de service mobile nécessitant l'Abonnement mobile avec contrat.
Pig	C'est une plate-forme de haut niveau pour la création de programmes MapReduce utilisés avec Hadoop.
Qlikview	Qlik est une société de logiciels basée à Radnor, en Pennsylvanie. Qlik est le fournisseur de QlikView qui est un logiciel permettant l''intégration et l'analyse des données [5].
Snoobe	Application gratuite de comparaison de forfait de téléphonie mobile pour téléphone intelligent Android.
SGBD	Système de gestion de base de données
Spark	Suite de logiciels libres, développés par l'université Berkeley pour le traitement analytique de données à grande échelle/ou Bigdata.
SSH	Sécure Shell. Est un protocole de réseau pour sécuriser les communications de données, à distance de ligne de commande de connexion.
Snappy	Bibliothèque de compression et décompression il vise pour une vitesse très

	élevée et une compression raisonnable
Upload	Transfert de données à partir d'un ordinateur local à un ordinateur distant.

LISTE DES SYMBOLES ET UNITÉS DE MESURE

BI Business Intelligence.

BZIP2 Programme de compression de fichier gratuit et open source qui utilisent l'algorithme de Burrows-Wheeler**.**

ETL Extraction, Transformation et chargement de données.

GZIP est un format de fichier et un logiciel utilisé pour la compression et décompression de fichiers

SAP Systèmes d'applications et produits.

INTRODUCTION

De nos jours, la plupart des gens possèdent un téléphone mobile afin d'être joignables et accessibles en tout le temps. Si anciennement le téléphone était exclusivement utilisé pour des appels téléphoniques, aujourd'hui il est intelligent et il offre des centaines, voir, des milliers d'applications qui visent à répondre à toutes sortes de besoins des utilisateurs.

Selon plusieurs études, 75% [6] des utilisateurs de téléphone mobiles utilisent un forfait qui ne leur convient pas. De manière générale, les utilisateurs ont tendance à surestimer leurs besoins et paient trop pour des services dont ils n'ont pas besoin ou qu'ils n'utilisent que rarement. Des centaines de millions d'utilisateurs à travers le monde ont ce problème et actuellement il n'y a pas de solution à cette situation. Il est estimé que si les utilisateurs de téléphones mobiles utilisaient un forfait qui correspond à leur profil, ils pourraient économiser, en moyenne, 30% sur leur facture mensuelle. Collectivement, cela représente des milliards de dollars qui sont gaspillés par la clientèle, et ce au profit des fournisseurs de services.

CHAPITRE 1

PRÉSENTATION

1.1 L'application de Snoobe.

Snoobe [7] : est un outil de comparaison de forfaits sur technologie Androïde cette application vise à aider les utilisateurs à reconnaitre leurs besoins réels en consommation de données (c.-à-d. 3G, LTE). Elle permet aussi de comparer plus de 100 forfaits mobiles aux États-Unis et au Canada, et recommande aux utilisateurs les meilleurs forfaits selon leurs besoins réels.

.

1.2 Problématique spécifique du projet

Comme dans la réalisation de tous les projets, il y a des obstacles à surmonter, l'ingénieur doit prendre les bonnes décisions afin d'assurer le succès du projet et gérer les attentes des clients. Lors de la réalisation de ce projet, les principaux défis ont été :

- Bien cerner le besoin du client : pour commencer ce projet, un document de vision a été rédigé avec l'objectif de mieux comprendre et de formaliser les besoins du client;
- Préciser un cas d'utilisation à partir de tous les besoins exprimés par le client : il a été, difficile de trouver un cas d'utilisation BigData qui vise à la partie «back-en» de l'application Snoobe. À la suite de plusieurs rencontres avec Thierry, nous avons réussi à trouver un cas d'utilisation portant sur « l'analyse de la puissance du signal d'un opérateur à un endroit donné»;
- Définir la source de données : étant donné que le projet nécessite l'utilisation de technologie BigData, il a fallu générer une grande quantité de données pour notre prototype;
- Maitriser la nouvelle technologie BigData : il a été nécessaire d'installer les technologies localement et sur Amazon ce qui a été un défi.

1.3 **Objectifs à atteindre**:

Il y a trois objectifs à atteindre pour réaliser ce projet:

- Installer l'environnement BigData cdh5 sur Amazon puis d'identifier la structure et le format des données nécessaires pour la base de données.
- Générer une grande quantité de données.
- Analyser les résultats obtenus en opérant localement sur un LapTop et sur une grappe Amazon.

CHAPITRE 2 PRÉSENTATION DES TECHNOLOGIES BIGDATA.

Bâtir une infrastructure où vous pouvez stocker et analyser des données de toute taille et de tout type est devenu un défi. Les entreprises génèrent une quantité massive de données ce qui nécessite une nouvelle approche pour stocker et analyser ces données. Les systèmes traditionnels (c.-à-d. relationnel) de gestion de base de données (c.-à-d. relationnel) ne suffisent pas toujours à cette tâche. Ce projet permet de s'initier aux différentes technologies du BigData et à apprécier leurs avantages et inconvénients en réalisant un projet réel. Les sections qui suivent présentent les technologies BigData considérées pour ce projet.

2.1 Hadoop

Apache Hadoop[8] est un cadriciel disponible en logiciel libre permettant à ses utilisateurs le traitement distribué de grande quantité de données à travers de grappes d'ordinateurs, et ce, à l'aide de modèles de programmation simples. Une grappe d'ordinateurs Hadoop n'a pas besoin d'être configurée avec un système RAID. Hadoop est composée de deux sous-systèmes le Hadoop Distributed File (HDFS), et le MapReduce. Les sociétés telles que Cloudera, Hortonworks et MapR, sont positionnées comme des acteurs clés dans l'écosystème Hadoop[9].

2.2 HDFS

HDFS [10] est un système de stockage distribué qui offre la possibilité d'enregistrer des données en les dupliquant. L'avantage de HDFS c'est qu'il est conçu et accessible en logiciel libre, au lieu d'exiger des licences d'utilisation couteuses auprès de fournisseurs traditionnels de logiciels d'entrepôt de données. Il est tolérant aux pannes, car les blocs de fichiers sont répliqués sur plusieurs nœuds de données (c.-a-d. de la grappe d'ordinateurs) et son système de fichiers peut détecter et récupérer des défaillances matérielles. Le cas d'utilisation typique pour lequel HDFS a été conçu est d'écrire une fois, mais lire plusieurs fois les données. Il accepte des fichiers de grande taille allant de 10MB à 100 TB (c.-à-d. fichiers texte, fichiers de log, etc.). Une fois que les fichiers sont écrits dans les répertoires HDFS, ils ne sont pas

censés changer et les données peuvent être immédiatement analysées avec le traitement des données de l'algorithme de Hadoop (par exemple à l'aide de MapReduce), ou avec un moteur de recherche distribué comme Impala [11] qui sera décrit dans les chapitres suivants.

2.3 MapReduce

L'algorithme de traitement des données de Hadoop est basé sur la technologie MapReduce [12]. Les données sont aussi immuables, c'est-à-dire qu'elles ne sont pas modifiées au cours du processus. Cette immuabilité se prête bien à la programmation parallèle, car il n'y a pas le besoin de synchroniser les opérations entre les moteurs d'exécution.

2.3.1 Inconvénients de MapReduce.

L'inconvénient de MapReduce[16]. Il y a deux problèmes principaux observés avec l'utilisation de la technologie MapReduce.

- Le premier est que les concepts de programmation fonctionnelle ne sont pas faciles à implémenter.
- Le second c'est que, MapReduce est conçu pour les processus de traitement par lots de longue durée. Il nécessite un certain temps pour traiter les données et cracher le résultat. Il peut prendre des heures, voir plusieurs jours, en fonction du volume de données.

2.4 Hbase

Hbase[13] : est une infrastructure superposée à Hadoop, qui n'utilise pas la technologie de MapReduce. Hbase enregistre les tables par colonnes alors que traditionnellement elles sont stockées par rangées, ce qui lui permet de réduire grandement le temps de recherche parce qu'on ne lit que les informations des colonnes dont on a besoin. Hbase ne supporte pas la mise à jour et les suppressions de données. L'utilisation de Hbase ne serait donc pas une bonne solution pour la réalisation de ce projet, car nous avons besoin du mode lecture et écriture de données.

2.4.1 Inconvénient de Hbase

II n'y a aucun moyen simple et efficace de faire des jointures avec Hbase.

2.5 Hive

Hive [14] : a été développé, à l'origine, par Facebook. Hive est une infrastructure d'entrepôt de données superposées à d'hadoop. Hive permet de faire de l'analyse, des résumés et des requêtes dans Hadoop. C'est un langage ressemblant beaucoup à SQL.

2.5.1 Inconvénient de Hive.

L'inconvénient de Hive [15] c'est qu'il utilise des fonctions de MapReduce qui sont très lentes. Hive ne supporte pas facilement les opérations de mise à jour ou d'effacement des données.

2.6 Impala

Impala [17] est un entrepôt de données qui permet l'utilisation efficace de plusieurs unités de traitement indépendantes, qui fonctionnent en parallèle. Impala permet de coordonner et exécuter les requêtes d'un utilisateur à travers une grande quantité de données. L'avantage de cette nouvelle technologie est que ce qui aurait normalement pris quelques minutes peut être exécuté en quelques secondes. Impala n'a pas de concept de serveur «maître». À la place de ce concept chaque «daemon Impala» est capable de réaliser toutes les responsabilités d'un moteur de recherche. Le moteur d'exécution d'Impala a été conçu pour être extrêmement rapide et efficace.

2.6.1 Architecture Impala.

Les différents composants d'Impala sont

- **Impala daemon** : sert à recevoir des demandes à partir d'une variété de sources.
- **Impala statestore** : vérifie le bon fonctionnement du «daemon», sur tous les nœuds d'une grappe d'ordinateurs. Si un nœud Impala est déconnecté en raison d'une panne de matériel ou de réseau, le «daemon statestore» informe tous les autres nœuds, afin que les futures requêtes puissent éviter de faire des demandes aux nœuds inaccessibles.
- **Impala metada et metastore :** est une base de données permettant l'enregistrement d'une grande quantité de données.
- **Serveur de catalogue** : sert à synchroniser toutes les métadonnées c'est-à-dire les tables avec le «daemon» Impala.

La figure ci-dessous présente la répartition de ces différents composants **:**

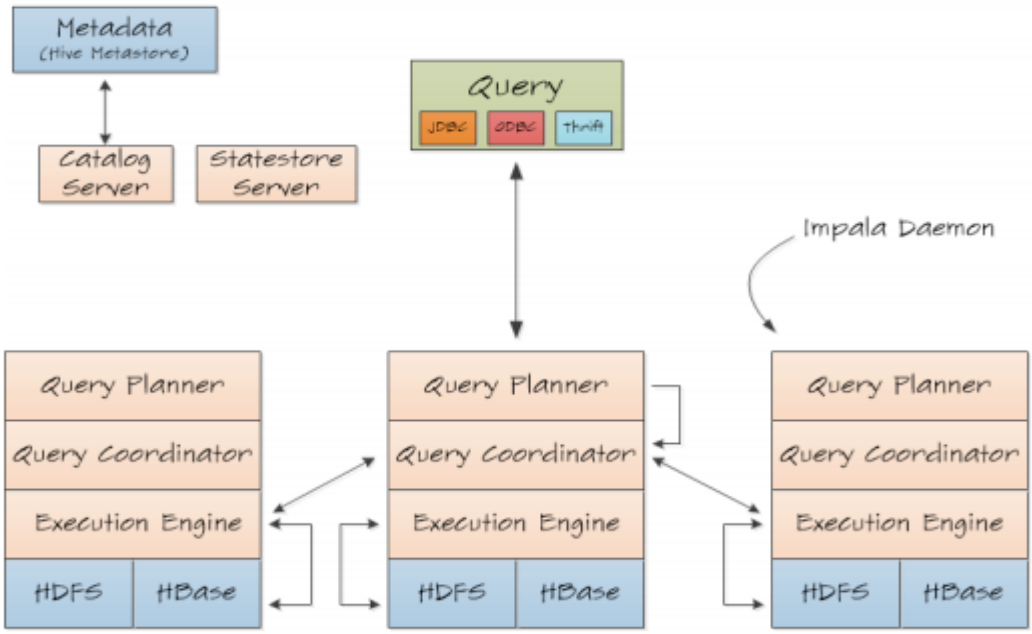


Figure 2 Présentation de l'architecture Impala [18].

2.6.2 Format supporté par impala

Le tableau ci-dessous présente les différents formats supportés par Impala

Format Prise en charge par Impala	Type de fichier
Fichier texte et séquence qui peuvent être compressés avec ➢ Snappy ➢ GZIP ➢ BZIP	➢ Avro ➢ RCFile ➢ LZO texte file ➢ paquet

Tableau 1 Formats supportés par Impala. [19]

Type de fichier	Format	Compression codec	CREATE?	INSERT?
Paquet	Structurée	Snappy, GZIP. Snappy est utilisé actuellement par default.	Oui	Oui. On peut utiliser des requêtes pour créer une table, insérer et chargé des données.
Avro	Structurée	Snappy, GZIP, deflate, BZIP2	Oui, avec la version 1.4.0 d'Impla. Avant pour créer des tables il fallait utiliser Hive.	Non. Utilisé la commande LOAD DATA pour charger des fichiers qui sont déjà dans un format correct ou utilisé Hive,
RCFile	Structurée	Snappy, GZIP, deflate, BZIP2	Oui.	Non. Utilisé la commande LOAD DATA pour charger des fichiers qui sont déjà dans un format correct ou utilisé Hive
Fichier Texte	Non structurée	LZO, gzip, bzip, Snappy	Oui, format de fichier texte non compressé, avec des valeurs	Oui, on peut utiliser des requêtes créer insérer et

			séparées par des caractères ASCII.	charger des données si la compression LZO est utilisée on doit passer par Hive. Si on utilise d'autres types de compression on doit utiliser la commande LOAD DATA.
Fichier de séquence	Structurée	Snappy, GZIP, deflate, BZIP2	Non, chargées les données à travers la commande LOAD DATA ou utilise Hive	Oui. Dans la version 2.0 d'Impala

Tableau 2 Règles sur la façon d'utiliser les types de fichier sur Impala [20].

Il est intéressant de connaitre comment sont utilisés ces différents formats, la figure 3 ci-dessous présente l'avantage de ces différents formats. Il est avantageux d'utiliser Impala paquet comme le montre et en plus, afin d'avoir un meilleur résultat de combiner le format Paquet avec le format Snappy (voir figure 4). Si le format de fichier de parquet est utilisé comme source de données d'entrée, il peut accélérer le traitement de la requête à une vitesse multiple.

La figure ci-dessous présente le nombre de requêtes que les systèmes pourraient traiter en une heure.

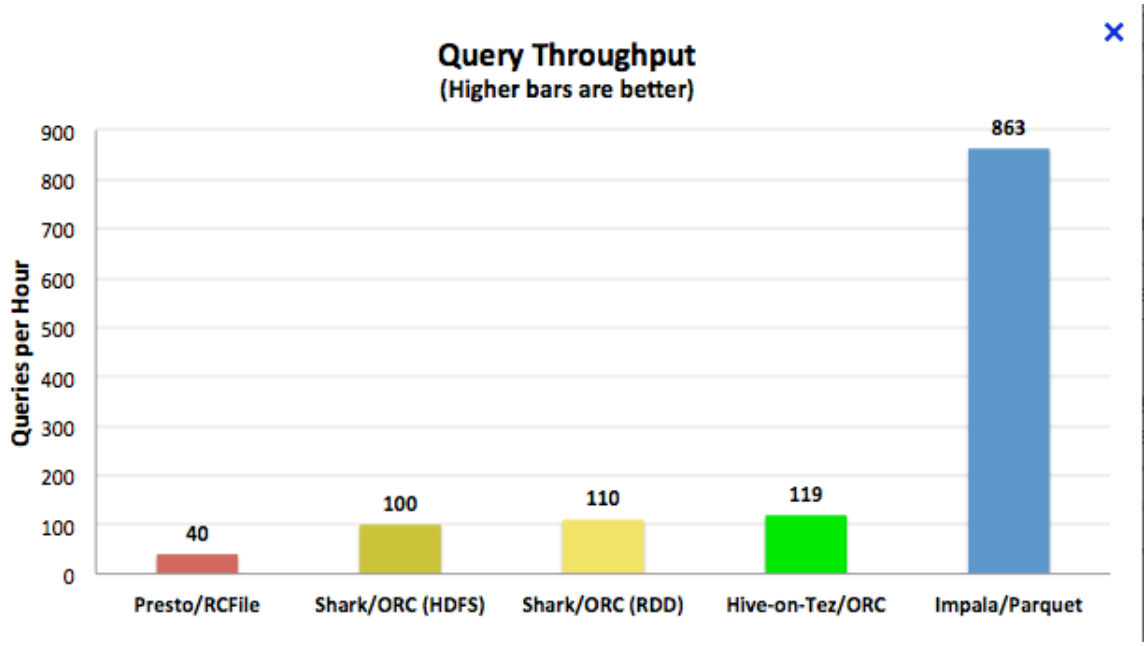


Figure3 Comparaison du nombre d'exécutions de requêtes, en heures, entre Impala et les autres technologies BigDta. [21]

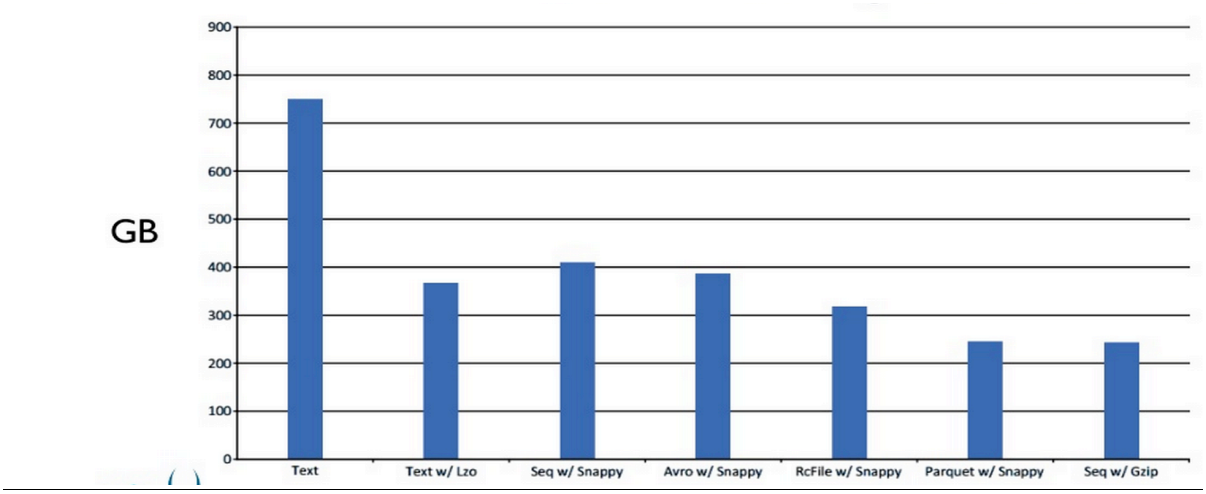


Figure 4 Optimisation de la base de données en utilisant différents formats. [22]

2.6.3 Avantages d'Impala.

- **La rapidité :**Impala permet d'exécuter des requêtes SQL sur Hadoop en quelques secondes.
- **Flexibilité :** il permet l'exécution de requêtes de données brutes et l'optimisation des formats de fichiers qui sont compatibles avec toutes les technologies BigData(c.-à-d. Hadoop, MapReduce, Hive, Pig, etc.). Les données peuvent être stockées sur HDFS ou HBase, et peuvent être traitées par des programmes MapReduce traditionnels, ainsi que par des requêtes ad hoc Impala. Il n'est pas nécessaire de déplacer des volumes importants de données pour effectuer un traitement analytique. Impala n'a pas été conçu pour remplacer MapReduce, mais plutôt pour le compléter et permettre de traiter, et d'analyser des données à sa sortie en utilisant un moteur à faible latence.
- **L'intégration facile du logiciel libre :** Impala est totalement intégré dans le cadriciel Hadoop (c.-à-d. ressources système, sécurité, etc.) et ne nécessite pas la migration des données dans des systèmes spécialisés ou des formats propriétaires. Son code source est ouvert et sous licence apache[23].
- **L'extensibilité :** Impala fonctionne avec des applications de Business Intelligence (BI) commerciale dont Microstrategy, Tableau,Qlikview, Pentaho, SAP, Alterysx et bien d'autres. Il n'est pas nécessaire d'utiliser des outils d'extraction, de

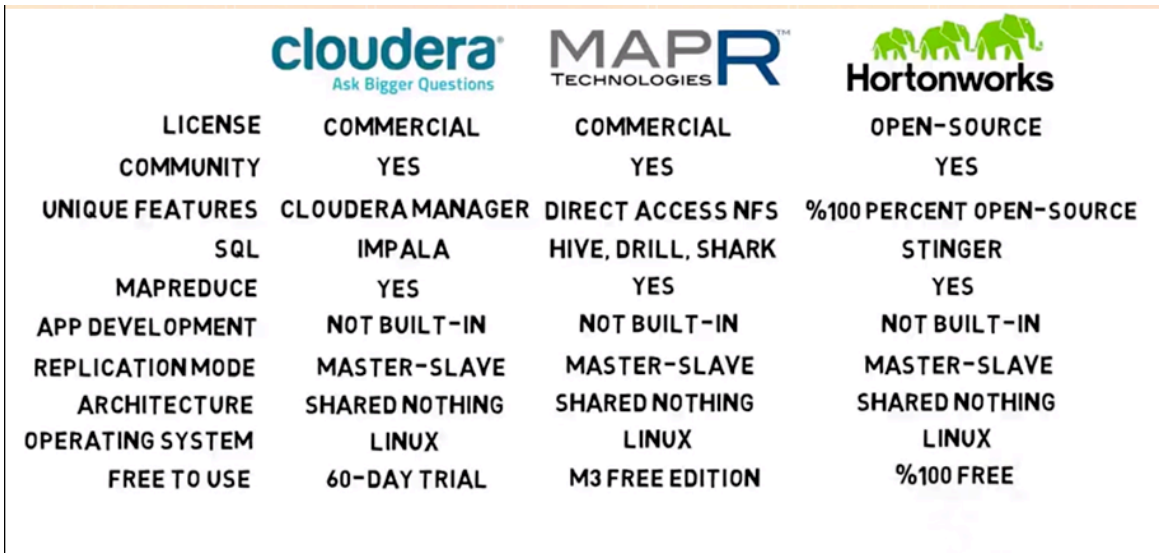
transformation et de chargement de données (ETL) pour charger des données dans Impala de Hadoop. Les analystes peuvent désormais interroger et analyser toutes ces données stockées dans un seul endroit. La fonctionnalité d'importation des données d'Impala est très simple, et est idéale pour le chargement d'un ensemble de données en vrac. Le moteur de stockage favori d'Impala est HDFS, qui est essentiellement un système de fichiers évolutif qui traverse de nombreux nœuds. Contrairement aux autres technologies de bases de données, Impala est capable d'analyser des données dans sa forme brute, ce qui signifie que vous pouvez simplement copier un fichier vers HDFS et immédiatement commencer à l'analyser avec Impala.

2.6.4 Inconvénients d'Impala.

Impala est une nouvelle technologie peu connue des spécialistes en génie logiciel. Sa syntaxe SQL légèrement différente des normes connues, ce qui nécessite une formation initiale. Si l'intégration d'Impala avec HDFS fonctionne très bien en revanche son » intégration n'est pas très efficace avec Hbase, Parquet et Amazon S3.

2.7 Comparaison des acteurs clés dans l'écosystème Hadoop.

Plusieurs acteurs sur le marché sont intéressés aux développements de l'écosystème Hadoop. La figure 5, ci-dessous, montre qu'Impala est une technologie largement utilisée par la société Cloudera. Étant donné que ce projet de fin d'études a choisi d'utiliser la technologie Impala, il sera nécessaire d'apprendre à installer à la configurer. Le chapitre suivant présente le paquet d'installation offert par Cloudera nommé cdh5.



	cloudera Ask Bigger Questions	MAPR TECHNOLOGIES	Hortonworks
LICENSE	COMMERCIAL	COMMERCIAL	OPEN-SOURCE
COMMUNITY	YES	YES	YES
UNIQUE FEATURES	CLOUDERA MANAGER	DIRECT ACCESS NFS	%100 PERCENT OPEN-SOURCE
SQL	IMPALA	HIVE, DRILL, SHARK	STINGER
MAPREDUCE	YES	YES	YES
APP DEVELOPMENT	NOT BUILT-IN	NOT BUILT-IN	NOT BUILT-IN
REPLICATION MODE	MASTER-SLAVE	MASTER-SLAVE	MASTER-SLAVE
ARCHITECTURE	SHARED NOTHING	SHARED NOTHING	SHARED NOTHING
OPERATING SYSTEM	LINUX	LINUX	LINUX
FREE TO USE	60-DAY TRIAL	M3 FREE EDITION	%100 FREE

Figure 5 : Comparaison des acteurs clés dans l'écosystème Hadoop[24].

2.7.1 Présentation de Cloudera

Cloudera [25] est une société de logiciels, des États-Unis, qui investit 50% de sa production d'ingénierie logicielle au projet Apache sous la licence gratuit (c.-à-d. Apache Hive, Apache Avro, Impala, etc.). D'un point de vue opérationnel, Impala est plus facile à gérer avec le gestionnaire de distribution Cloudera cdh5. Cloudera a préparé une application de gestion de distribution sophistiquée pour Hadoop. Elle est conçue pour se déployer, se configurer, surveiller et diagnostiquer les grappes d'ordinateurs facilement à partir d'une console de gestion Web centralisée. Elle est sous licence libre ce qui nous permet de la télécharger et de l'utiliser. Cloudera offre également une version entreprise, qui met à niveau ses capacités, l'ajout de fonctionnalités que les entreprises ont besoin (c.-à-dire la découverte, des alertes et des rapports).

2.7.2 Avantage de la distribution cdh5 de cloudera.

- Facile de modifier un paramètre de configuration ou d'installer une librairie supplémentaire sur plusieurs machines en un seul instant;

- Permets d'afficher les configurations qui sont reliées à l'option qui est en train d'être modifiée.

2.7.3 Désavantage de Cloudera Manager.

- Certains services doivent absolument être redémarrés à l'aide de l'interface graphique;
- Impossible de changer la configuration sans l'utilisation de l'interface graphique;
- Impossible de changer le nom d'hôte d'un serveur (c.-à-d. de changer le localhost vers le bon nom d'hôte) [26].

CHAPITRE 3

ENVIRONNEMENT DE DÉVELOPPEMENT CLOUDERA

3.1 Installation de Cloudera quickstart cdh5 sur Amazon

On peut utiliser Cloudera de plusieurs manières. Soit on l'installe et le configure directement sur une instance d'Amazon, soit on utilise une instance déjà configurée comme quickStart Cloudera, qui contient tout ce qu'on a besoin pour faire du BigData. Dans notre cas nous allons installer l'instance quickstart Cloudera cdh5 virtuel machine sur Amazon. Voici comment nous avons procédé (voir Annexe IV).

3.2 Installer les outils API EC2 Amazon

Premièrement, il est nécessaire d'installer les outils d'interface permettant de lancer des commandes depuis un Shell Windows. Téléchargez et décompressez le paquet « EC2 API Tools » et le place sur le bureau. On peut trouver ce paquet à l'adresse suivante [27]:

3.3 Vérifier la version et l'environnement de Java

Installez puis vérifiez la version et l'environnement de Java : il faut installer java version1.7 ou plus.

3.4 Configurer l'interface EC2

Il est important de s'assurer qu'on obtient les résultats présentés aux figures ci-dessous. Il est nécessaire de configurer l'interface EC2 (CLI) installée de la manière suivante.

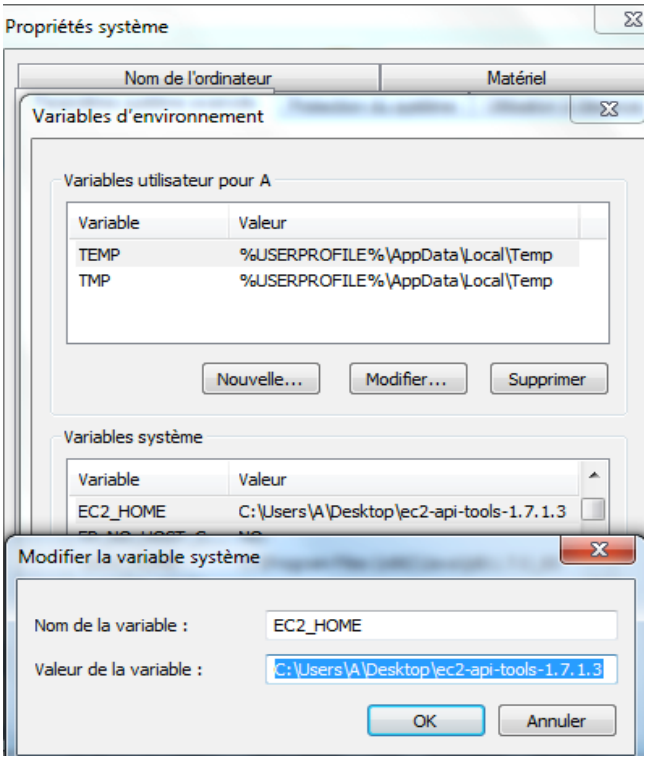


Figure 6 Configuration de l'environnement EC2 d'API Amazon.

Une fois que les configurations initiales sont terminées, il faut vérifier qu'il est possible d'afficher la liste des régions ec2 api Amazon : en utilisant la commande ec2-describe-regions.

3.5 Création d'une clé d'authentification

Pour accéder à notre instance il est incontournable de définir une clé d'accès permettant d'installer notre image virtuelle voir la figure ci-dessous.

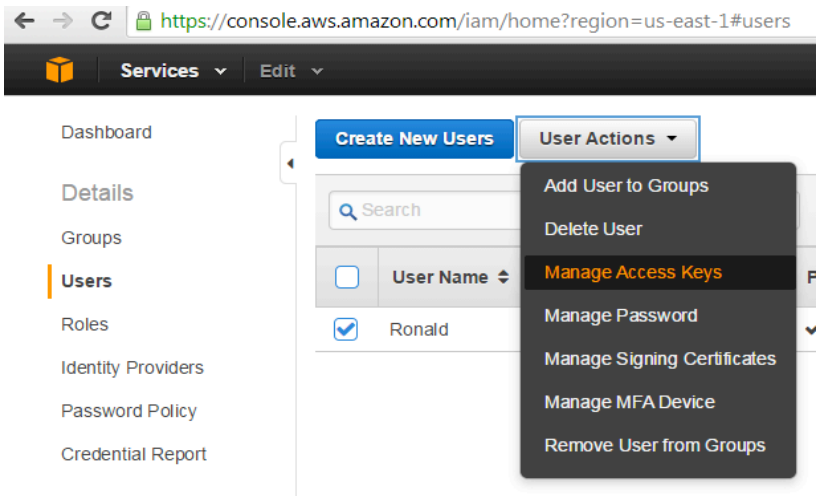


Figure 7 Création d'une clé privée.

3.6 Création du Bucket S3

Par la suite la création du bucket S3 sur Amazon, va contenir les métadonnées et les fichiers logs.

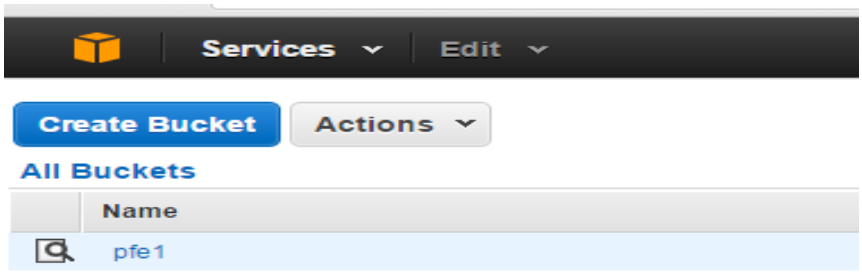


Figure 8: Création du bucket S3 sur Amazon.

3.7 Lancer l'installation de cloudera

Afin de lancer l'installation de Cloudera cdh5, depuis Windows, il suffit de taper la commande ci-dessous :

```
ec2-import-instance « D:\cloudera\cloudera-quickstart-vm-5.1.0-1-virtualbox-disk1.vmdk » -f VMDK -t m3.xlarge -a x86_64 -o %AWS_ACCESS_KEY% -w %AWS_SECRET_KEY% -p Linux -b pfe1
```

3.8 Conclusion

Dans ce chapitre nous avons appris comment installer et configurer Cloudera quickstart cdh5 sur Amazon. Cette plateforme contient toutes les technologies du BigData nécessaires pour réaliser ce projet de fin d'études. Une fois l'environnement installé, nous allons maintenant procéder, dans le chapitre suivant, à l'étude d'une étude de cas d'utilisation qui va nous permettre d'avoir une vision globale du comportement fonctionnel de cette nouvelle technologie, mais aussi de préciser les différents acteurs qui vont interagir avec ce cas d'utilisation et les fonctionnalités que doivent supporter l'application afin de répondre aux besoins de ses utilisateurs [28].

CHAPITRE 4

CONCEPTION ET OPTIMISATION DE LA BASE DE DONNÉES

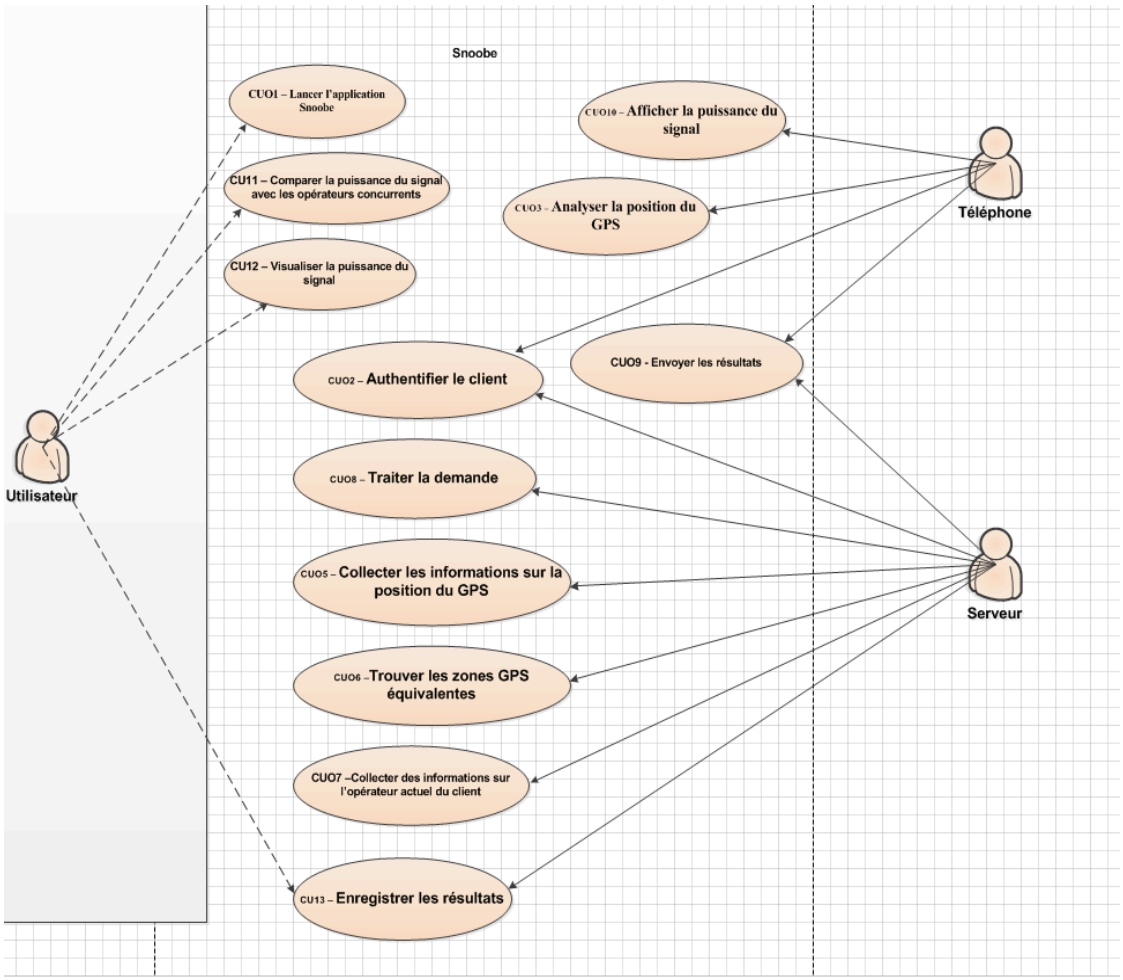
4.1 Introduction

Dans ce chapitre, nous abordons l'étude du modèle de donnée du cas d'utilisation étudié. Cette étape est très importante, car ce modèle de données va nous aider à préciser les exigences fonctionnelles du système. Il est aussi important de mentionner que les exigences non fonctionnelles, qui concernent la convivialité, la disponibilité, la performance, la fiabilité et la sécurité du système, décrites à la fin de ce chapitre seront aussi abordées. Dans le texte qui suit, les exigences fonctionnelles c'est-à-dire les cas d'utilisations sont présentées. Un cas d'utilisation permet d'identifier clairement les principaux acteurs et leurs interactions impliqués dans une transaction du système visé :

- AC01-Utilisateur
- AC01-Téléphone
- AC01-Serveurs

Les cas d'utilisation permettent non seulement d'identifier les acteurs, mais aussi les fonctionnalités qui doivent être supportées par l'application afin de répondre aux besoins de ses utilisateurs. La figure 9 ci-dessous présente l'interaction entre les acteurs avec le cas d'utilisation.

4.2 Étude d'un cas d'utilisation.



La figure 8 Diagramme des cas d'utilisation

Diagramme des Cas d'utilisation	**Brève description**
CUO1 : Lancer l'application Snoobe	L'utilisateur déclenche l'application Snoobe
CUO2 : Authentifier le client	L' utilisateur peut être Authentifier par le Téléphone et le serveur.
CUO3 : Analyser la position GPS	L' utilisateur effectue une analyse comparative de la puissance du signal.
CUO4 : Envoyer les informations	Synchronisation entre le Téléphon et le serveur.
CUO5 : Collecter les informations sur la position du GPS.	Le serveur recueille des informations sur la position du GPS.
CUO6 : Trouver les zones GPS équivalentes	Le serveur cherche et localise les zones du GPS équivalentes.
CUO7 : Collecter des informations sur l'opérateur actuel du client	Le serveur recueille des informations sur l'opérateur actuel du client.
CUO8; traiter la demande	Le serveur calcule la puissance du signal
CUO9 : Envoyer les résultats	L'utilisateur reçoit les résultats qui ont été transmis par le serveur.
CUO10 : Afficher la puissance du signal	L'utilisateur visualise la puissance du signal sur le téléphone.
CUO11 : Comparer la puissance du signal avec les opérateurs concurrents	L'utilisateur effectue une analyse comparative de la puissance du signal.
CUO12 : Visualiser la puissance du signal	L'utilisateur visionne la puissance du signal
CUO13 : Enregistrer les résultats	L'utilisateur sauvegarde les résultats

Tableau 3 résumé des cas d'utilisation .

4.3 Les exigences fonctionnelles

En se basant sur le diagramme des cas d'utilisation de la figure 9, on constate que le système doit traiter et gérer assez rapidement une grande quantité de requêtes. L'identification des exigences fonctionnelles dans le modèle des cas d'utilisation va nous permettre de minimiser les risques et de valider rapidement la faisabilité du projet.

4.4 Les exigences non fonctionnelles

Performance au niveau de Back-end

- Le système doit enregistrer une grande quantité de données
- L'utilisateur ne doit pas attendre plus de 15 secondes pour avoir le résultat d'une recherche en temps réel.

Disponibilité : Les données doivent être disponibles et mises à jour.

Le chapitre suivant défini la structure et identifie le type de données nécessaires afin de permettre de répondre à ce cas d'utilisation, à savoir, l'analyse du signal de tous les opérateurs disponibles**.**

CHAPITRE 5

IMPLÉMENTATION

5.1 Structure et identification des données

Afin de définir la structure des données, je me suis inspiré des informations disponibles sur le site « d'Open Signal » [29] qui permet de générer des données pour notre projet. Chez Open Signal, il faut créer un compte pour obtenir une clé d'accès. Leur serveur retourne un fichier en format JSON (voir Annexe1). Une fois ces données obtenues, il faut faire la mise en forme. Cette étape a été difficile, car cela a nécessité d'appliquer plusieurs transformations. J'ai dû concevoir un programme Java permettant de stocker ces données dans un fichier de format CSV (c.-à-d. d'une taille d'environ 6 giga octets) que j'ai analysé à l'aide de la technologie Impala.

Étant donné qu'Impala permet l'analyse de fichiers de format CSV, j'ai importé ce fichier dans Impala et j'analyse la performance d'Impala suite à l'exécution de différentes requêtes. Par la suite je comparer les résultats d'exécution avec d'autres technologies BigData comme Spark, Hive et Hbase.

Pour récupérer les données, j'ai dû me connecter au le site d'Open Signal et obtenir un fichier, de format JSON, que je traite dans une base de données Impala (voir la structure de ce fichier en Annexe).

5.2 Création et importation des données

L'un des plus grands avantages d'impala c'est qu'on peut insérer et analyser directement des fichiers de tous types de formats. Voici la commande qui permet de créer notre table.

Table :analysesignal	
Colonne	**Type**
apiversion	string
distance	Int
latitude	double
longitude	double
averagerssiasu	double
averageRssidb	double
downloadSpeed	double
networkId	string
networkName	string
pingTime	double
reliability	double
uploadSpeed	double
networkType	string

Tableau 4 Structure de la base de données.

Je crée la table, de la base de données, à l'aide de la commande suivante :

```
create table analysesignal (apiVersion STRING, distance INT, latitude DOUBLE, longitude DOUBLE, averageRssiAsu DOUBLE, averageRssiDb DOUBLE, downloadSpeed DOUBLE, networkId STRING, networkName STRING, pingTime DOUBLE, reliability DOUBLE, uploadSpeed DOUBLE, networkType STRING) row format delimited fields terminated by ','
```

5.3 Les différentes manières pour se connecter à Impala.

Les utilisateurs peuvent se connecter à Impala à travers une variété de sources. Par exemple à l'aide d'ODBC, de JDBC, d'un Shell et même d'interfaces WEB. La figure 3 image ces options. Pour ce projet j'ai utilisé le Shell et l'interface Web HUE.

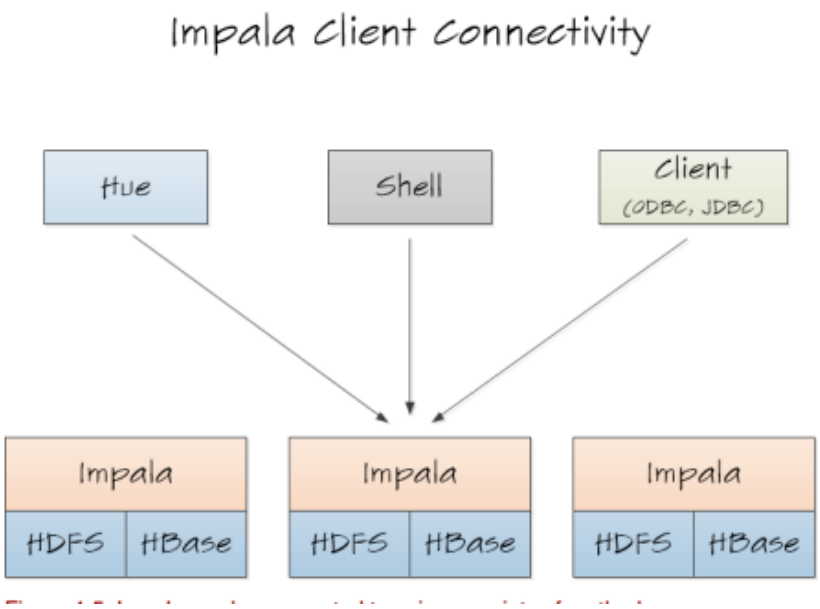


Figure 9: Les différentes manières pour se connecter à Impala [31]:

CHAPITRE 6

ANALYSE DES RÉSULTATS

Après avoir défini et identifié la structure et les types de donnés dans le chapitre précédent, ce chapitre décrit comment se connecter sur la base de données Impala, créer une table et insérer des données dans cette table.

6.1 Connexion à la base de données Impala

La figure ci-dessous présente l'interface permettant de se connecter à la base de données Impala, mais aussi aux autres technologies BigData (c.-à-d. Hbase ,Spark, Hive, Pig etc..)

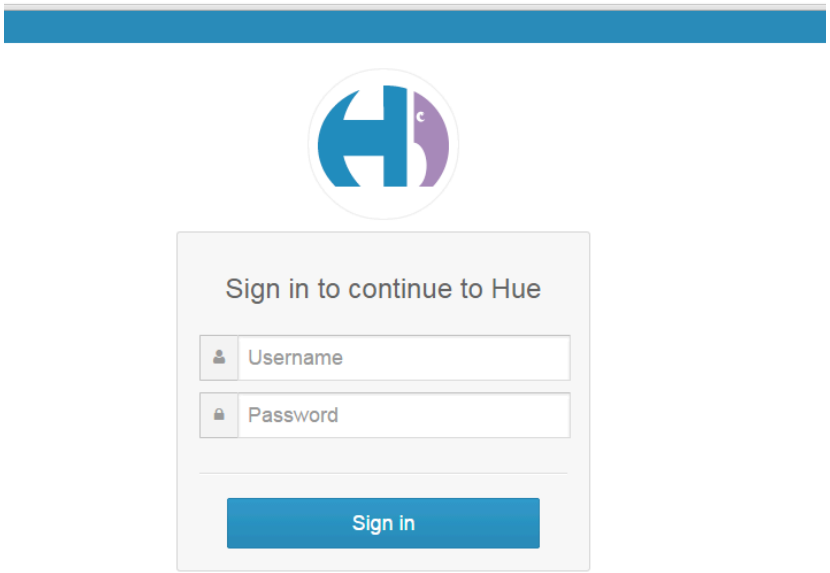
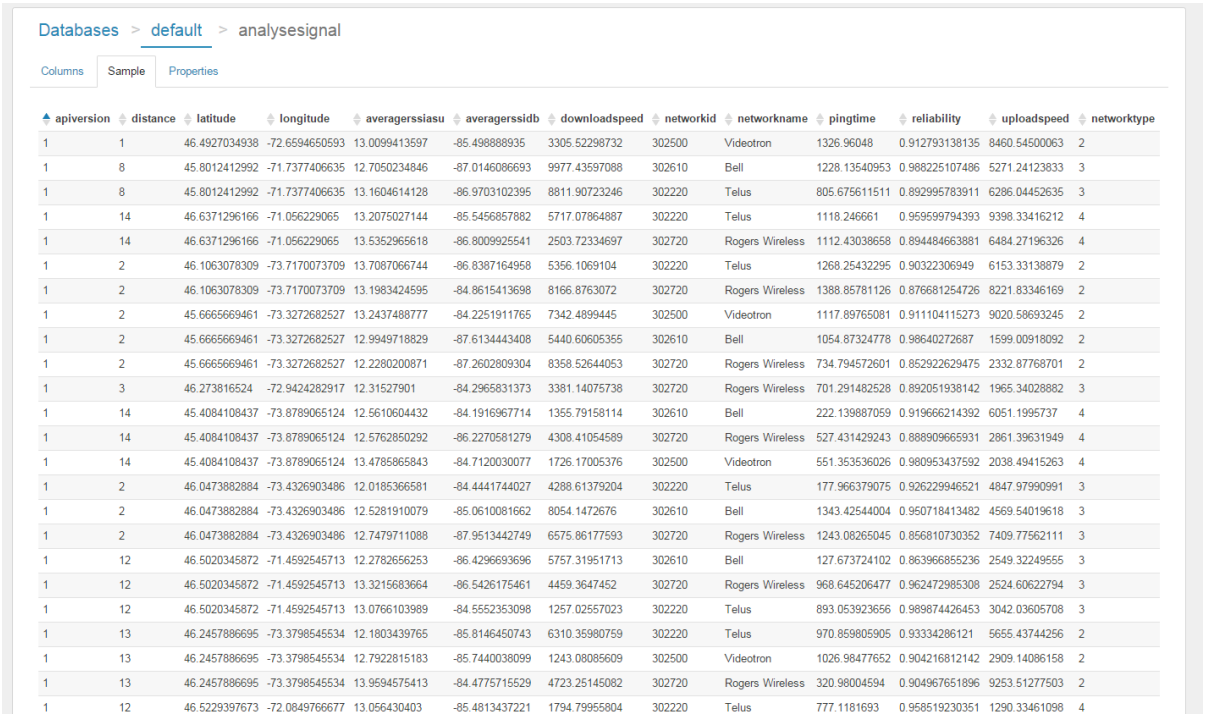


Figure 11 l'interface de connexion sur impala.

6.2 Importation de données à la base de données Impala.

Une fois que les données ont été traitées et misent au format, il suffit d'enregistrer les données dans le répertoire de HDFS et charger ces données dans la table décrite au chapitre précédent. Il est bon de signaler que cette table doit avoir la structure qui permet d'afficher les résultats décrits à la Figure 12 ci-dessous. Le chargement de données nécessite plusieurs minutes, mais une fois que ces données sont importées dans la base de données Impala la visualisation ne prend que quelques secondes .



apiversion	distance	latitude	longitude	averagerssiasu	averagerssidb	downloadspeed	networkid	networkname	pingtime	reliability	uploadspeed	networktype
1	1	46.4927034938	-72.6594650593	13.0099413597	-85.498888935	3305.52298732	302500	Videotron	1326.96048	0.912793138135	8460.54500063	2
1	8	45.8012412992	-71.7377406635	12.7050234846	-87.0146086693	9977.43597088	302610	Bell	1228.13540953	0.988225107486	5271.24123833	3
1	8	45.8012412992	-71.7377406635	13.1604614128	-86.9703102395	8811.90723246	302220	Telus	805.675611511	0.892995783911	6286.04452635	3
1	14	46.6371296166	-71.056229065	13.2075027144	-85.5456857882	5717.07864887	302220	Telus	1118.246661	0.959599794393	9398.33416212	4
1	14	46.6371296166	-71.056229065	13.5352965618	-86.8009925541	2503.72334697	302720	Rogers Wireless	1112.43038658	0.894484663881	6484.27196326	4
1	2	46.1063078309	-73.7170073709	13.7087066744	-86.8387164958	5356.1069104	302220	Telus	1268.25432295	0.90322306949	6153.33138879	2
1	2	46.1063078309	-73.7170073709	13.1983424595	-84.8615413698	8166.8763072	302720	Rogers Wireless	1388.85781126	0.876681254726	8221.83346169	2
1	2	45.6665669461	-73.3272682527	13.2437488777	-84.2251911765	7342.4899445	302500	Videotron	1117.89765081	0.911104115273	9020.58693245	2
1	2	45.6665669461	-73.3272682527	12.9949718829	-87.6134443408	5440.60605355	302610	Bell	1054.87324778	0.98640272687	1599.00918092	2
1	2	45.6665669461	-73.3272682527	12.2280200871	-87.2602809304	8358.52644053	302720	Rogers Wireless	734.794572601	0.852922629475	2332.87768701	2
1	3	46.273816524	-72.9424282917	12.31527901	-84.2965831373	3381.14075738	302720	Rogers Wireless	701.291482528	0.892051938142	1965.34028882	3
1	14	45.4084108437	-73.8789065124	12.5610604432	-84.1916967714	1355.79158114	302610	Bell	222.139887059	0.919666214392	6051.1995737	4
1	14	45.4084108437	-73.8789065124	12.5762850292	-86.2270581279	4308.41054589	302720	Rogers Wireless	527.431429243	0.888909665931	2861.39631949	4
1	14	45.4084108437	-73.8789065124	13.4785865843	-84.7120030077	1726.17005376	302500	Videotron	551.353536026	0.980953437592	2038.49415263	4
1	2	46.0473882884	-73.4326903486	12.0185366581	-84.4441744027	4288.61379204	302220	Telus	177.966379075	0.926229946521	4847.97990991	3
1	2	46.0473882884	-73.4326903486	12.5281910079	-85.0610081662	8054.1472676	302610	Bell	1343.42544004	0.950718413482	4569.54019618	3
1	2	46.0473882884	-73.4326903486	12.7479711088	-87.9513442749	6575.86177593	302720	Rogers Wireless	1243.08265045	0.856810730352	7409.77562111	3
1	12	46.5020345872	-71.4592545713	12.2782656253	-86.4296693696	5757.31951713	302610	Bell	127.673724102	0.863966855236	2549.32249555	3
1	12	46.5020345872	-71.4592545713	13.3215683664	-86.5426175461	4459.3647452	302720	Rogers Wireless	968.645206477	0.962472985308	2524.60622794	3
1	12	46.5020345872	-71.4592545713	13.0766103989	-84.5552353098	1257.02557023	302220	Telus	893.053923656	0.989874426453	3042.03605708	3
1	13	46.2457886695	-73.3798545534	12.1803439765	-85.8146450743	6310.35980759	302220	Telus	970.859805905	0.93334286121	5655.43744256	2
1	13	46.2457886695	-73.3798545534	12.7922815183	-85.7440038099	1243.08085609	302500	Videotron	1026.98477652	0.904216812142	2909.14086158	2
1	13	46.2457886695	-73.3798545534	13.9594575413	-84.4775715529	4723.25145082	302720	Rogers Wireless	320.98004594	0.904967651896	9253.51277503	2
1	12	46.5229397673	-72.0849766677	13.056430403	-85.4813437221	1794.79955804	302220	Telus	777.1181693	0.958519230351	1290.33461098	4

Figure 12 a : Affichage des résultats désirés.

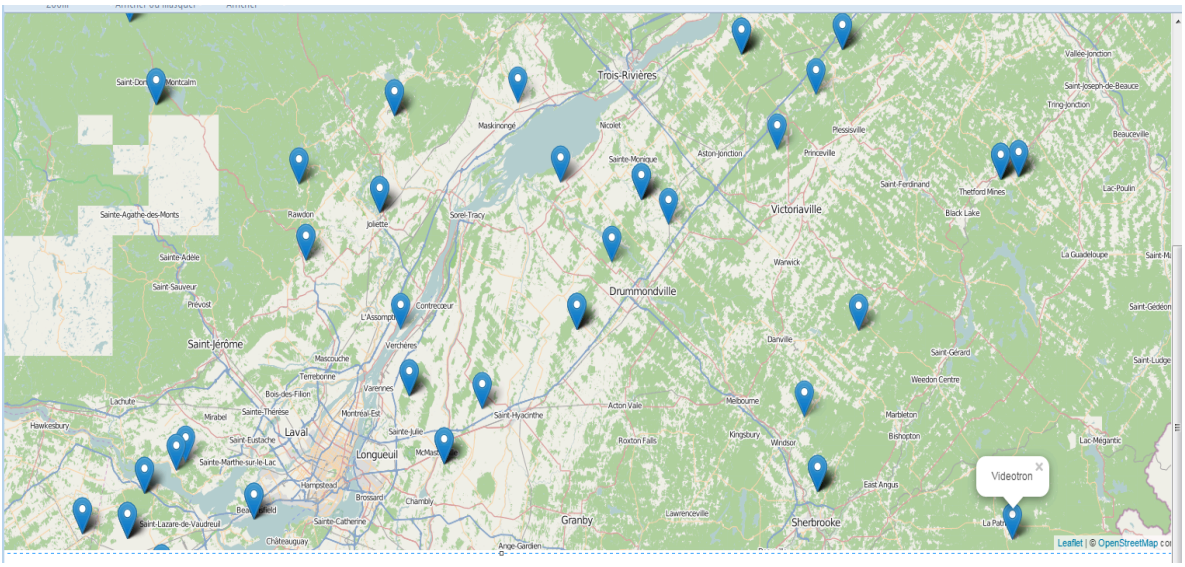


Figure 12 b Affichage des résultats désirés.

Il serait mieux de choisir beaucoup plus d'opérateurs, car

6.3 Comparaison de la technologie Impala avec les autres technologies BigData

Dans le but de comparer la technologie Impala avec les autres technologies BigData, j'ai utilisé un fichier, en format CSV, d'une taille d'environ 6 giga-octets qui contient 36,854,703 colonnes. C'est à partir de ces données que les commandes sont exécutées et que j'ai pu comparer le temps de réponse des différentes technologies (voir les commandes exécutées à l'annexe II).

Le tableau ci-dessous nous montre Clair que le temps en réponse est très rapide lorsqu'on exécute des requêtes sur la technologie Impala.

Requête: Temps(s)	1	2	3	4;	5	6	7	8	9
Impala	0.16	0.02	344.42	59.11	491.90	703.37	668.37	485.30	437.24
Hive	0.621	0.327	1154.518	4.144	1351.701	3093.228	1057.882	1057.822	744.797

Tableau 5 Résultats des requêtes en secondes effectué sur les technologies Impala et Hive

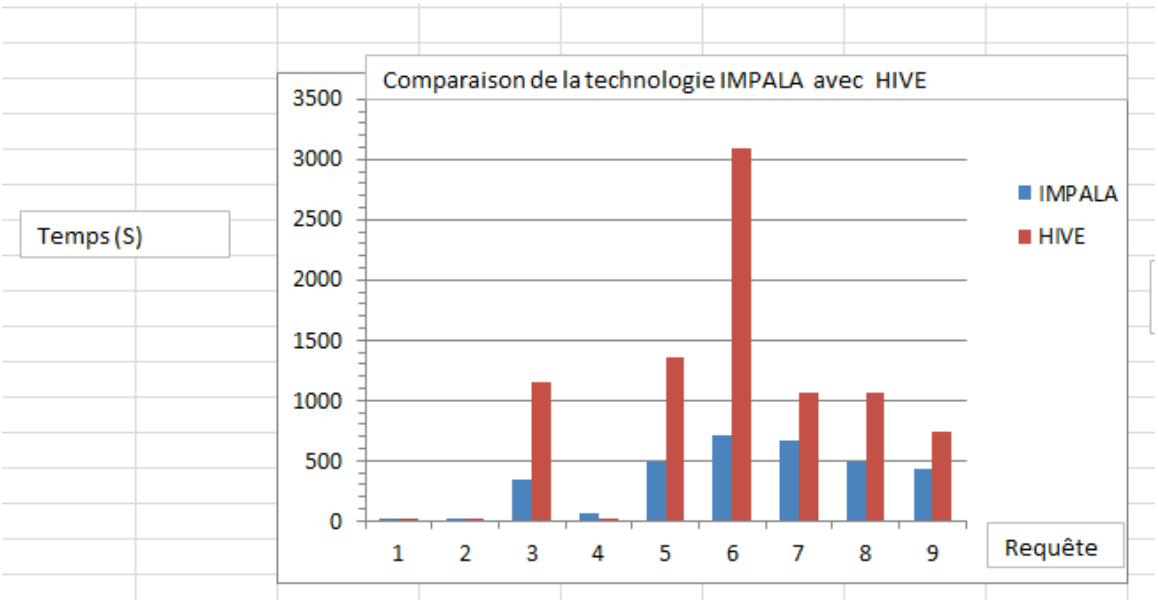


Figure-13 Comparaison de la technologie IMPALA avec HIVE.

6.4 Différence entre la technologie Impala avec la technologie Hive

Les principales différences entre Impala et Hive c'est que :

- Impala effectue un traitement en mémoire alors que Hive n'effectue aucune traite en mémoire.
- Hive Utilise MapReduce pour traiter les requêtes tandis que Impala utilise son propre moteur de traitement [31].

On peut conclure, que la technologie Impala est très performante, le temps réponse est très rapide ce qui est rassurant pour son utilisation pour l'application Snoobe qui souhait desservir un grand nombre d'utilisateurs dans le futur. Il serait possible d'utiliser d'autres formats de compression afin d'avoir de meilleur résultat.

CONCLUSION

Impala est une toute nouvelle technologie d'entrepôt de donnée BigData. Elle offre des améliorations sur la façon d'analyser, d'intégrer et de charger de grandes quantités de données. En comparant cette nouvelle technologie avec d'autres technologies similaires, du domaine du BigData, j'ai constaté que la technologie Impala est très avantageuse et qu'elle pourra être très pratique pour l'aspect « back-end » des analyses de grandes quantités de données pour les applications de Snoobe. Il sera ainsi possible d'afficher les informations concernant la puissance du signal de tous les opérateurs disponibles en quelques secondes, et ce à des milliers d'utilisateurs Snoobe. Ce projet ne m'a pas seulement permis de comprendre le fonctionnement de la technologie Impala, mais aussi l'ensemble des technologies du BigData qui sont actuellement des technologies nouvelles, puissantes et très en demande.

RECOMMANDATIONS

L'application Snoobe est une application qui sera utilisée par des milliers de clients, et ceci en temps-réel et, ou la rapidité et le temps de réponse joueront un rôle important. La technologie Impala est actuellement une des technologies BigData qui permet l'analyse d'une très grande quantité de données en quelques secondes. On pourrait utiliser aussi différents formats de compression afin d'augmenter encore cette vitesse. Pour afficher les informations depuis Impala vers un téléphone mobile Androïde voici les différentes étapes à suivre.

- Créer les vues nécessaires;
- créer un user « read only » dans Impala;
- configurer Impala pour fonctionner avec JDBC
- créé l'application Androïde;
- ajouter les permissions nécessaires dans la configuration de l'application pour qu'elle ait le droit de se connecter à internet;
- dans l'application Androïde, ouvrir une connexion vers Impala via le JDBC en utilisant le user « read only »;
- créer des champs pour que l'utilisateur entre les paramètres (distance, longitude, etc.);
- faire les requêtes nécessaires sur la Base de données et afficher le résultat.

LISTE DE RÉFÉRENCES

[1]	Opensignal. (2014). Analyse de la puissance du signal. Consulté le 02 octobre 2014, sur Opensignal :http://opensignal.com/
[2]	Wikipédia. (2014, novembre 4). Entreprise Aletryx. Consulté le 14 novembre 2014, sur : http://en.wikipedia.org/wiki/Alteryx
[3]	Webopedia. (2014). BigData. consulté le 13 novembre 2014 sur : http://www.webopedia.com/TERM/B/big_data.html
[4]	Microstrategy. (2014). Entreprise microstrategy. Consulté le 14 novembre 2014 sur http://www.microstrategy.com/fr/a-propos/societe
[5]	Wikipédia. (2014, novembre 18). Logiciel qlikview. Consulté le 20 novembre 2014, sur : http://en.wikipedia.org/wiki/Qlik
[6]	BBC. (2011, avril 11). Les utilisateurs paient pour des services mobiles qu'ils n'utilisent pas. Consulté le 07 novembre 2014 sur : http://www.bbc.co.uk/news/technology-12996175
[7]	Snoobe. (2013). Application snoobe. Consulté le 03 novembre 2014 sur: http://www.snoobe.com/#!francais/c1dxv
[8]	Développez.(2014). Hadoop . Consulté le 10 novembre 2014 sur : http://mbaron.developpez.com/tutoriels/bigdata/hadoop/introduction-hdfs-map-reduce/
[9]	YouTube. (2014, février 24).Comparaison : Cloudera, MapR et Hortonworks . Consulté le 07 novembre 2014 sur : https://www.youtube.com/watch?v=WRfMrwyniqQ
[10]	Développez.(2014). HDFS. Consulté le 10 novembre 2014 sur : http://mbaron.developpez.com/tutoriels/bigdata/hadoop/introduction-hdfs-map-reduce/
[11]	Richard L.Saltzer et Istvan Szegedi. (2014). Impala: A distributed query engine. Impala in action.
[12]	Richard L.Saltzer et Istvan Szegedi. (2014). MapReduce : HDFS and MapReduce. Impala in action.
[13]	Anton.Z et David.L. (2012, octobre 29). Hbase. Consulté le 10 novembre 2014 sur : http://publicationslist.org/data/a.april/ref-382/03_RapportEtape.pdf
[14]	Wikipédia. (2014, novembre 16). Hive. Consulté le 18 novembre 2014 sur : http://en.wikipedia.org/wiki/Apache_Hive
[15]	Anton.Z et David.L. (2012, octobre 29). Hive. Consulté le 10 novembre 2014 sur : http://publicationslist.org/data/a.april/ref-382/03_RapportEtape.pdf
[16]	Anton.Z et David.L. (2012, octobre 29). Hive. Consulté le 10 novembre 2014 sur

	: http://publicationslist.org/data/a.april/ref-382/03_RapportEtape.pdf
[17]	Richard L.Saltzer et Istvan Szegedi. (2014). Impala: A distributed query engine. Impala in action.
[18]	Richard L.Saltzer et Istvan Szegedi. (2014). Architecture Impala. Impala in action
[19]	Richard L.Saltzer et Istvan Szegedi. (2014). Files format and storage engines. Impala in action.
[20]	Cloudera.(2014).Format de fichier. Consulté le 16 novembre 2014 sur : http://www.cloudera.com/content/cloudera/en/documentation/cloudera-impala/latest/topics/impala_avro.html
[21]	Cloudera. (2014). New choices in the Apache Hadoop Ecosystem: why impala Continues to Lead. Consulté le 12 novembre 2014 sur : http://blog.cloudera.com/blog/2014/05/new-sql-choices-in-the-apache-hadoop-ecosystem-why-impala-continues-to-lead/
[22]	Paquet. (2013, octobre 28). Strata conférence. Consulté le 05 octobre 2014 sur : http://parquet.incubator.apache.org/presentations/
[23]	GitHub. (2014).Real time for Hadoop. Consulté le 05 octobre 2014 sur: https://github.com/cloudera/impala
[24]	YouTube. (2014, février 24).Comparaison : Cloudera, MapR et Hortonworks . Consulté le 07 novembre 2014 sur : https://www.youtube.com/watch?v=WRfMrwyniqQ
[25]	Wikipédia. (2014, octobre 22). Cloudera. Consulté le 3 novembre sur : http://en.wikipedia.org/wiki/Cloudera
[26]	Anton.Z et David.L. (2012, octobre 29). Désavantage de Cloudera Manager. Consulté le 10 novembre 2014 sur : http://publicationslist.org/data/a.april/ref-382/03_RapportEtape.pdf
[27]	Amazon. (2013).EC2 API Tools. Consulté le 05 septembre 2014 sur http://aws.amazon.com/developertools/351
[28]	Wikipédia. (2014, octobre 8). Diagramme des Cas d'utilisation. Consulté le 11 novembre 2014 sur : http://fr.wikipedia.org/wiki/Diagramme_des_cas_d%27utilisation

[29]	Opensignal. (2014). Analyse de la puissance du signal. Consulté le 02 octobre 2014, sur Opensignal :http://opensignal.com/
[30]	Richard L.Saltzer et Istvan Szegedi. (2014). Impala: A distributed query engine. Impala in action.
[31]	Safari.(2014).Comparaison Impala avec Hive. Consulté le 25 novembre 2014, sur : https://www.safaribooksonline.com/library/view/learning-cloudera-impala/9781783281275/ch07s02.html

ANNEXE I LISTE DE REQUETES EFFECTUER SUR LES TECHNOLOGIES IMPALA ET HIVE

- Requêtes-1 **Créer la table analysesignal**

```
create table analysesignal (apiVersion STRING, distance INT, latitude
DOUBLE, longitude DOUBLE, averageRssiAsu DOUBLE,
averageRssiDb DOUBLE, downloadSpeed DOUBLE,
networkId STRING, networkName STRING, pingTime DOUBLE,
reliability DOUBLE, uploadSpeed DOUBLE, networkType STRING)
row format delimited fields terminated by ','
```

Requêtes-2 **Afficher de la structure de la table analysesignal sur**

```
DESCRIBE analysesignal;
```

- Requêtes-3 **Compter le nombre de lignes insérées dans la table analysesignal**

```
select count(*) from analysesignal;
```

- Requêtes-4 Afficher les données de la table analysesignal en limitant la recherche a 20,000

```
select *from analysesignal limit 20000
```

- Requêtes-5 Afficher la puissance maximale du signal.

```
select max(averagerssidb) from analysesignal;
```

- Requêtes-6 Afficher la puissance du signal avec le nom de l'opérateur correspondant

```
select networkname ,averagerssidb,from analysesignal where averagerssidb=-
84.00000000065167;
```

- Requêtes-7 Afficher la puissance moyenne du signal.

```
Select avg(averagerssidb) from analysesignal
```

- Requêtes-8 Afficher l'endroit où la puissance du signal est plus élevée avec le nom de l'opérateur correspondant.

```
select  latitude, longitude, networkname , averagerssidb from analysesignal; where averagerssidb=-84.00000000065167;
```

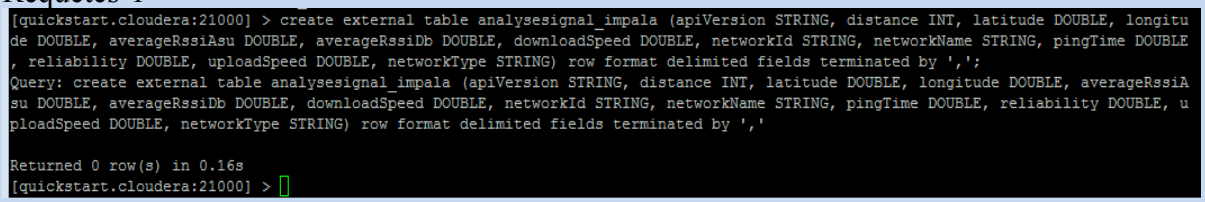
- Requêtes-9

```
select min(averagerssidb), = from analysesignal;
```

Requêtes exécutées sur les technologies Impala et Hive

- **Technologies Impala**

Requêtes-1



Requêtes -2

```
[quickstart.cloudera:21000] > DESCRIBE analysesignal;
Query: describe analysesignal
+----------------+--------+---------+
| name           | type   | comment |
+----------------+--------+---------+
| apiversion     | string |         |
| distance       | int    |         |
| latitude       | double |         |
| longitude      | double |         |
| averagerssiasu | double |         |
| averagerssidb  | double |         |
| downloadspeed  | double |         |
| networkid      | string |         |
| networkname    | string |         |
| pingtime       | double |         |
| reliability    | double |         |
| uploadspeed    | double |         |
| networktype    | string |         |
+----------------+--------+---------+
Returned 13 row(s) in 0.02s
```

Requêtes-3

```
[quickstart.cloudera:21000] > select count(*) from analysesignal;
Query: select count(*) from analysesignal
+----------+
| count(*) |
+----------+
| 36854703 |
+----------+
Returned 1 row(s) in 344.42s
```

Requêtes **-4**

```
+-----------+---------+-------------------+-------------------+-------------
Returned 20000 row(s) in 59.11s
[quickstart.cloudera:21000] >
```

Requêtes -5

```
                          > averagerssidb) from analysesignal;
Query: select max( averagerssidb) from analysesignal
+--------------------+
| max(averagerssidb) |
+--------------------+
| -84.00000000065167 |
+--------------------+
Returned 1 row(s) in 491.90s
```

Requêtes -6

```
[quickstart.cloudera:21000] > select networkname ,averagerssidb from analysesign
Query: select networkname ,averagerssidb from analysesignal where averagerssidb=
+-------------+--------------------+
| networkname | averagerssidb      |
+-------------+--------------------+
| Bell        | -84.00000000065167 |
+-------------+--------------------+
Returned 1 row(s) in 708.37s
```

Requêtes -7

```
[quickstart.cloudera:21000] > select avg(averagerssidb) from analysesignal;
Query: select avg(averagerssidb) from analysesignal
+--------------------+
| avg(averagerssidb) |
+--------------------+
| -85.99996234280546 |
+--------------------+
Returned 1 row(s) in 668.37s
```

Requêtes -8

```
[quickstart.cloudera:21000] > select  latitude, longitude, networkname , averagerssidb
Query: select latitude, longitude, networkname , averagerssidb from analysesignal   wh
+-----------------+-------------------+-------------+--------------------+
| latitude        | longitude         | networkname | averagerssidb      |
+-----------------+-------------------+-------------+--------------------+
| 45.8384935077136 | -71.14769160650097 | Bell        | -84.00000000065167 |
+-----------------+-------------------+-------------+--------------------+
Returned 1 row(s) in 485.30s
[quickstart.cloudera:21000] >
```

Requêtes-9

```
Query: select min(averagerssidb) from analysesignal
+--------------------+
| min(averagerssidb) |
+--------------------+
| -87.99999983450668 |
+--------------------+
Returned 1 row(s) in 437.24s
```

Technologie Hive :

Requête-1

```
hive> create external table analysesignal_hive (apiVersion STRING, distance INT,
    > networkId STRING, networkName STRING, pingTime DOUBLE, reliability DOUBLE,
OK
Time taken: 0.621 seconds
hive>
```

Requête -2

```
hive> DESCRIBE analysesignal;
OK
apiversion              string                  None
distance                int                     None
latitude                double                  None
longitude               double                  None
averagerssiasu          double                  None
averagerssidb           double                  None
downloadspeed           double                  None
networkid               string                  None
networkname             string                  None
pingtime                double                  None
reliability             double                  None
uploadspeed             double                  None
networktype             string                  None
Time taken: 0.327 seconds, Fetched: 13 row(s)
hive>
```

Requête-3

```
DFS Write: 9 SUCCESS
Total MapReduce CPU Time Spent: 1 minutes 50 seconds 930 msec
OK
36854703
Time taken: 1154.518 seconds, Fetched: 1 row(s)
```

Remarque : On remarque bien que Hive se base sur Mapreduce pour effectuer ces calculs.

Requête-**4**

```
Time taken: 4.144 seconds, Fetched: 20000 row(s)
hive>
```

Requêtes-5

```
Job 0: Map: 24  Reduce: 1   Cumulative CPU: 189.51 sec   HDFS Read: 6300053841 HDFS Write:
 19 SUCCESS
Total MapReduce CPU Time Spent: 3 minutes 9 seconds 510 msec
OK
-84.00000000065167
Time taken: 1351.701 seconds, Fetched: 1 row(s)
hive>
```

Requêtes-6

```
Job 0: Map: 24   Cumulative CPU: 213.71 sec   HDFS Read: 6300053841 HDFS Write:
Total MapReduce CPU Time Spent: 3 minutes 33 seconds 710 msec
OK
Bell    -84.00000000065167
Time taken: 3093.228 seconds, Fetched: 1 row(s)
hive>
```

Requêtes-7

```
Time taken: 659.812 seconds, Fetched: 36854703 row(s)
hive> select  networktype from analysesignal;
```

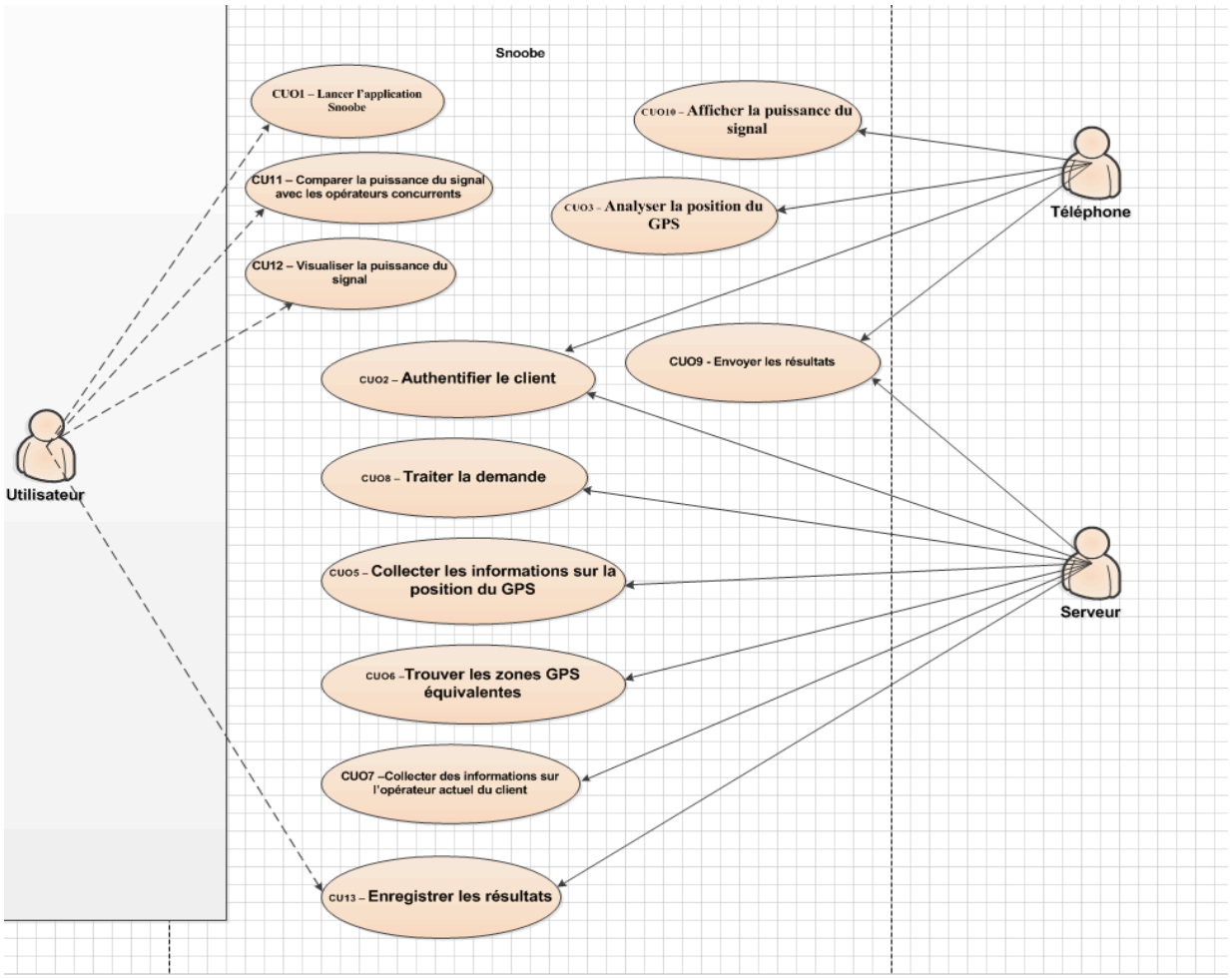
Requetes-8

```
DFS Write: 19 SUCCESS
Total MapReduce CPU Time Spent: 3 minutes 1 seconds 990 msec
OK
-85.99996234284177
Time taken: 1057.822 seconds, Fetched: 1 row(s)
hive>
```

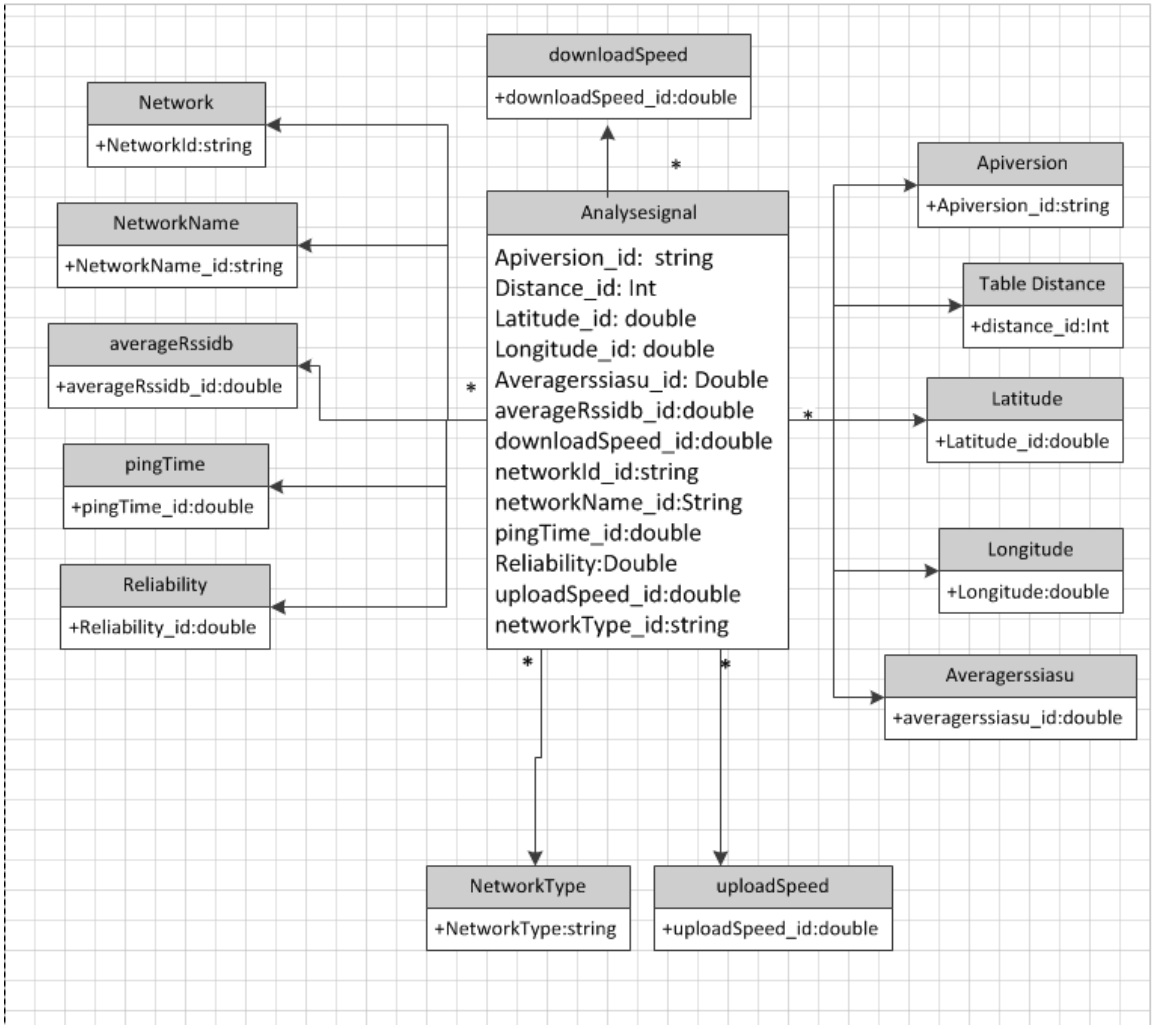
Requetes-9

```
Total MapReduce CPU Time Spent: 3 minutes 33 seconds 280 msec
OK
-87.99999983450668
Time taken: 744.797 seconds, Fetched: 1 row(s)
hive>
```

ANNEXE II DIAGRAMME DES CAS D'UTILISATION



ANNEXE III SCHÉMAS DE LA BASE DE DONNÉES :



ANNEXE IV
DOCUMENT DE VISION.

ANNEXE V PROCÉDURE D'INSTALLATION DE CLOUDERA CDH5 SUR AMAZON

ANNEXE VI DOUCUMENT D'ÉTAPE

ANNEXE VII IMPORTATION DES DONNÉES DE FORMAT JSON DANS IMPALA


```
{
 « apiVersion »: « 2 »,
 « latitude »: « 37.790 »,
 « longitude »: « -122.4058 »,
 « distance »: « 15 »,
 « network_type »: « 3g »,
 « perMinuteCurrent »: 0,
 « perMinuteLimit »: 10,
 « perMonthCurrent »: 6,
 "perMonthLimit": 2000,
 "networkRank": {
  « network310120 »: {
   « type3G »: {
    "networkName": "Sprint",
    "networkId": "310120",
    "networkType": "3",
    "averageRssiAsu": "16.443599",
    "averageRssiDb": "-80.112801",
    "sampleSizeRSSI": "909869",
    "downloadSpeed": "908.6300",
    "uploadSpeed": « 761.1343 »,
    « pingTime »: « 172.8134 »,
    « reliability »: « 85.3313518696118 »
   }
  },
  « network310260 »: {
   « type3G »: {
    « networkName »: « T-Mobile »,
    « networkId »: « 310260 »,
    « networkType »: « 3 »,
    « averageRssiAsu »: « 13.440021 »,
    « averageRssiDb »: « -86.119958 »,
    « sampleSizeRSSI »: « 571741 »,
```

« downloadSpeed »: « 5564.2263 »,
« uploadSpeed »: « 1433.0242 »,
« pingTime »: « 166.9662 »,
« reliability »: « 89.3065545565596 »
}
},
« network310150 »: {
« type3G »: {
« networkName »: « AT&T »,
« networkId »: « 310150 »,
« networkType »: « 3 »,
« averageRssiAsu »: « 14.935515 »,
« averageRssiDb »: « -83.128969 »,
« sampleSizeRSSI »: « 244537 »,
« downloadSpeed »: « 3130.4336 »,
« uploadSpeed »: « 1109.8630 »,
« pingTime »: « 179.1498 »,
« reliability »: « 88.947704386418 »
}
},
« network310004 »: {
« type3G »: {
"networkName": "Verizon",
"networkId": "310004",
"networkType": "3",
"averageRssiAsu": "13.531994",
"averageRssiDb": "-85.936012",
"sampleSizeRSSI": "129769",
"downloadSpeed": "605.9672",
"uploadSpeed": « 554.0638 »,
« pingTime »: « 119.3926 »,
« reliability »: « 85.8720930232623 »
}
},
« network31016 »: {

```
  « type3G »: {
    "networkName": "MetroPCS",
    "networkId": "31016",
    "networkType": "3",
    "averageRssiAsu": "14.435914",
    "averageRssiDb": "-84.128172",
    "sampleSizeRSSI": "91348",
    "downloadSpeed": "3393.6456",
    "uploadSpeed": "1136.6131",
    "pingTime": "211.6000",
    "reliability": « 87.3622586872615 »
  }
},
« network310016 »: {
  « type3G »: {
    "networkName": "cricKet",
    "networkId": "310016",
    "networkType": "3",
    "averageRssiAsu": "15.893620",
    "averageRssiDb": "-81.212760",
    "sampleSizeRSSI": « 42921 »
  }
},
« network310053 »: {
  « type3G »: {
    "networkName": "Virgin Mobile",
    "networkId": "310053",
    "networkType": "3",
    "averageRssiAsu": "17.228289",
    "averageRssiDb": "-78.543421",
    "sampleSizeRSSI": "22660",
    "downloadSpeed": "618.8333",
    "uploadSpeed": "463.9388",
    "pingTime": "180.6154",
    "reliability": « 90.6114819759706 »
```

```
      }
    }
  }
}
```

ANNEXE VIII LISTE DES PORTS À AUTORISER SUR AMAZON

Pour que notre système fonctionner normalement voici la liste des port à ouvrir sur Amazon :