Nancy Vien (1895337)
Nancy Badran (1895359)
Alejandra Jimenez Nieto (1895997)

Architecture matérielle et logicielle Big Data
420-BD3-BB Gr. 425

Projet – Étude de cas – Printemps 2018

Travail présenté
à Hafed Benteftita

Collège de Bois-de-Boulogne
26 mai 2018

1. Introduction

Comme dans des plusieurs domaines, les applications de données massives ont changé la façon dont nous analysons et gérons les données dans le secteur des soins de santé. Ces applications ont le potentiel de réduire le coût des traitements, de prédire des épidémies et de les prévenir. Les données massives que les professionnels de la santé peuvent collecter durant leurs parcours professionnels, constituent une source très riche d'informations qui sera dédiée à être extraite et exploitée pour en obtenir des résultats.

Notre application Big Data repose sur un ensemble de données massives et aidera à prévenir les cas d'allergies respiratoires causés en premier lieu par des facteurs climatiques.

2. Données et besoins d'affaires

2.1 Données

Nous disposons d'un ensemble de données provenant de plusieurs sources. Ces données sont semi-structurées en format CSV. Elles portent sur plusieurs aspects, en voici quelques uns :

2.1.1 Indicateurs de la qualité de l'air[1].

Les deux principaux polluants qui composent le smog sont l'ozone troposphérique et les particules (PT : Particules Totales d'un diamètre maximal d'environ 100 micromètres, P10 : Particules d'un diamètre inférieur à 10 micromètres, P2.5 : Particules d'un diamètre inférieures à 2.5 micromètres). Les particules sont des fragments en suspension dans l'air sous forme solide ou liquide[2]. La taille des particules détermine la portée des dommages que celles-ci causent à la sante (il y a un lien entre les particules et de troubles respiratoires tels que l'asthme, la bronchite et l'emphysème).

La forte densité de smog est associée à la saison estivale pour la présence de rayons du soleil et des températures élevées. Le smog hivernal (à cause de la contribution des particules au lieu de l'ozone) étant une préoccupation sérieuse, quand l'air en stagnation accumule les polluants provenant du chauffage au bois et de l'utilisation de véhicules[3].

Les indicateurs de la qualité de l'air suivent les mesures de l'exposition des Canadiens à l'ozone troposphérique et aux particules fines (PM2,5). Les indicateurs d'exposition à l'ozone et aux P2,5 sont des concentrations moyennes pondérées en fonction de la population observées dans les stations de surveillance au Canada pendant la saison chaude (d'avril à septembre)[4].

---

[1] https://www.canada.ca/fr/environnement-changement-climatique/services/indicateurs-environnementaux/qualite-air.html
[2] https://www.canada.ca/fr/environnement-changement-climatique/services/pollution-atmospherique/polluants/principaux-contaminants/matieres-particulaires.html
[3] https://www.canada.ca/fr/environnement-changement-climatique/services/pollution-atmospherique/enjeux/smog-causes-effets.html
[4] http://www.statcan.gc.ca/pub/16-251-x/16-251-x2007000-eng.htm

2.1.2 Inventaire d'émissions polluantes atmosphériques[5] :

Monoxyde de carbone (CO), Composés organiques volatils (COV).

2.1.3 Asthme en Canada

Décès selon la cause : Maladies de l'appareil respiratoire[6] , Personnes avec maladie pulmonaire obstructive chronique[7].

2.1.4 Données climatiques

En tenant compte du fait qu'il existe une relation directe entre les maladies respiratoires et le climat, on va utiliser des données historiques avec des information de climatique. Il y a différents sources, par exemple sources qui viennent de différentes stations de mesure[8] :

2.2 Caractéristiques des données

Dimensions :

Volume	Grande charge de données
Variété	Plusieurs types de données et des données stratégiques : démographique, géolocalisation
Temporel	Non
Véracité	Oui
Visibilité	Oui
Valeur	La valeur des données augmentent en ajoutant des dossiers qui comportent un historique sur les patients.

2.2 Analyse des besoins applicatifs

2.2.1 Besoin d'affaires

Développement d'un système de prévision de l'asthme et des allergies respiratoires au Canada. Cette application permet faire une prévision des allergies respiratoires ou encore l'asthme à une localisation précise et une période spécifique. L'application peut aussi prévenir ou éviter les symptômes d'une personne déjà allergique en un endroit spécifique et à un moment donné et ainsi d'empêcher que cette allergie ne se complique en asthme ou encore en rhinite chronique.

---

[5] https://www.canada.ca/fr/environnement-changement-climatique/services/pollution-atmospherique/publications/inventaire-emissions-polluants-atmospheriques-2016/chapitre-2-5.html
[6] http://www5.statcan.gc.ca/cansim/a26?lang=fra&id=1020530&p2=33#F2
[7] http://www.statcan.gc.ca/tables-tableaux/sum-som/l02/cst01/health104a-fra.htm

2.2.2 Besoins fonctionnels

Utilisateur non-inscrit

- Donne des informations pour prévenir les allergies respiratoires et asthme pour la population en général.
- Donne des recommandations pour éviter que des allergies se développent.

Utilisateur non allergique inscrit

- L'utilisateur répond à une liste de question pour étoffer son dossier.
- Système de recommandations personnalisées géolocalisées pour les abonnées allergiques.
- Le résultat obtenu aide à formuler différentes solutions pour prévenir quelques situations reliées à l'allergie.

Utilisateur ayant déjà une allergie connue ou l'asthme

- L'utilisateur entre des données personnelles journalières pour l'aider à maitriser et contrôler ses symptômes
- L'utilisateur reçoit des alertes personnalisées pour l'aider au contrôle et la planification de son allergie
- L'utilisateur entre des informations précises (questionnaire étoffé) et sur une base historique et contextuel l'application pourra lui prédire s'il développera une allergie ou encore une maladie respiratoire ou encore son pourcentage de risque.
- Les résultats obtenus donneront des recommandations pour éviter que l'allergie connue se complique en maladie respiratoires comme l'asthme ou la rhinite chronique
- Permet de prévenir l'asthme ou la rhinite chronique
- Permet d'éviter une crise d'asthme

2.2.3 Besoins non-fonctionnels

- Sécurité des données personnelles
- Disponibilité de l'application web

2.2.4 Besoins techniques

- Hébergement de l'application
- Trouver la plateforme adéquate pour le développement du projet.

3. Architecture préliminaire

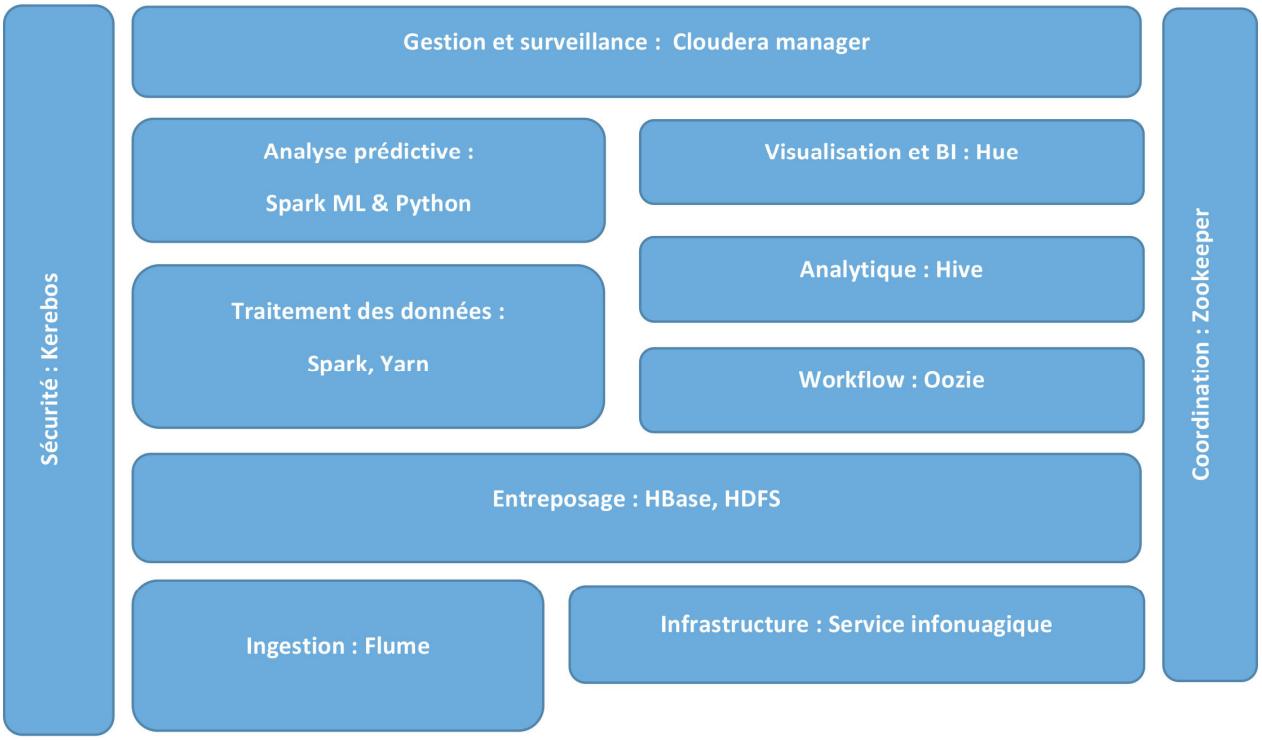
3.1 Contraintes technologiques

Le choix de des composantes de notre architecture est justifié par les contraintes suivantes :

- Un système d'ingestion de données structurées et semi-structurés
- Un système de traitement des données
- Un système d'analyse prédictive

- Un système pour le stockage de données

3.2 Élaboration de l'architecture



4. Feuille de route

Dans cette partie, nous allons expliquer la partie pratique de notre projet.

4.1 Mise en place de la plateforme

Vu que notre application utilisera des données massives, il serait raisonnable d'utiliser le plateforme Hadot 2.x. La platforme Hadoop sera installé sur une infrastructure nuageuse et prise en charge par Cloudera Manager.

4.2 Stockage de données

Pour une première phase, on s'est limité aux allergies respiratoires, c'est-à-dire les allergies liées aux changements climatiques, saisonniers et à la qualité de l'air. Notre Dataset préliminaire est de l'ordre de 1 Go plus le facteur de réplication HDFS. La taille des données mentionnées ci-dessus est estimée saisonnièrement, alors par année, il faudra multiplier par 4. À ne pas oublier, la capacité requise pour Map Reduce, pour la partie non-structurée, en plus de la taille de disque dur.

Nous estimons que le nombre minimal de nœuds requis est 2 Datanodes. Ce nombre augmentera au fur et à mesure que notre Dataset grossira.

4.3 Traitement

Pour le traitement des données, on a choisi de lancer un job Map Reduce en utilisant l'interface Hue. Pour l'entreposage des données intermédiaires (Staging Area), Hive a été notre choix. De plus, il y a une couche supérieure de programmation de Map Reduce Java assuré par Hue sur Cloudera Manager.

4.4 Collecte, transformation et chargement l'information dans le système

On a pris l'information climatique de la période de mai 2018 (Température (°C) / heure) à Montréal. Après, on a utilisé Map Reduce (Wordcount Job) pour obtenir les résultats suivants :

10.01
10.12
10.23
10.44
10.52
10.61
10.75
10.93
11.01
11.11
11.24
11.36
11.45
11.55
11.64
11.72
11.83
11.98
12.04
12.13
12.25
12.36
12.46
12.55
12.65
12.79
12.83
12.92
13.05
13.12
13.23
13.36
13.45
13.52
13.65
13.74
13.84
13.92
14.02
14.12
14.25
14.35
14.41
14.55
14.68

14.75
14.84
14.95
15.03
15.12
15.21
15.37
15.46
15.54
15.67
15.74
15.82
15.97
16.01
16.16
16.26
16.31
16.48
16.59
16.61
16.73
16.84
16.95
17.05
17.16
17.27
17.34
17.44
17.56
17.63
17.71
17.83
17.94
18.01
18.11
18.21
18.43
18.56
18.64
18.72
18.82
18.94
19.04
19.14
19.24
19.36
19.41
19.52
19.61
19.76
19.81
19.93
2.41
20.02
20.12
20.24
20.31
20.51
20.63
20.75
20.84
20.94
21.01
21.12
21.22
21.32

21.44
21.52
21.61
21.71
21.82
21.92
22.02
22.12
22.22
22.32
22.44
22.61
22.71
22.91
23.01
23.11
23.21
23.41
23.62
23.82
24.01
24.13
24.21
24.51
24.61
24.71
24.81
24.92
25.41
25.52
26.21
26.31
26.51
3.31
3.41
3.61
3.71
3.81
4.41
4.81
5.01
5.31
5.41
5.51
5.71
5.81
5.91
6.01
6.11
6.22
6.32
6.41
6.51
6.71
6.82
7.02
7.11
7.26
7.31
7.44
7.55
7.64
7.74
7.82
7.91
8.04

8.21
8.31
8.44
8.51
8.73
8.81
8.91
9.02
9.11
9.24
9.42
9.53
9.61
9.72
9.83

4.5 Analyses

Les températures plus fréquentes qu'on a trouvé dans le mois de mai de 2018 sont : 16.5 °C, 16.4 °C, 12.7 °C, 11.9 °C. On peut faire la comparaison de ces températures avec données de pollution mais aussi avec des données de maladies respiratoires dans le même période de temps et localisation.

5. Conclusion

En conclusion, cette première phase nous a permis de répondre aux besoins préliminaires du client. On a opté pour un cluster sur Cloudera Manager, avec un écosystème d'Hadoop consistant en une base HDFS, Hive et MapReduce Java.