Université de Paris
Dauphine

NATIXIS

Stage de fin d'études

Analyse et Recherche d'informations A travers les technologies Big Data

Auteur :
Mamadou Noumou
BARRY

Maitre de stage :
Madame Génivieve Jomié
Manager :
M.Amine Doukkali

Table des matières

Remerciements

Remerciements

Je tiens à exprimer mes remerciements et toute ma gratitude a mon responsable de master monsieur Jamal Atif, ma tutrice de stage madame Génievieve Jomier, mes managers monsieur Amine Doukkali et monsieur Edem pour avoir accepté de m'encadrer pendant ce stage et pour la confiance qu'ils m'ont accordée.

Je remercie les membres du Jury qui ont accepté d'examiner mon travail, le personnel enseignant et administratif de l'Université Paris Dauphine. Monsieur Raphaël Fourrier Professeur au cnam, je te suis reconnaissant pour le temps et l'effort que tu as consacrer à nos discussions.

Je tiens à remercier mon épouse BARRY BAH Aissata qui a été toujours là pour moi et qui m'a soutenu, a tous les membres de ma famille de leur soutien continu et plus particulièrement mon défunt père qui m'a toujours aider à aller plus loin et qui me motivais ; ma mère Hadja Binta Diallo qui m'a accordé sa bénédiction, son amour et ses encouragements continus.

Je remercie tous mes amis et plus particulièrement Jerome Tonnellier.

Introduction

Le présent rapport est une présentation du travail effectué dans le cadre de mon stage de fin d'études à la direction des systèmes d'information GM-Global Markets Distribution EBusiness de Natixis.

Contexte :

Le Big Data permet de traiter les données non-structurées encore très peu exploitées pour catégoriser les clients qui traitent avec Natixis sur le périmètre des produits de change (Forex). Dans ce contexte, Natixis souhaite travailler dans le cadre d'un stage de fin d'études sur l'analyse et la recherche d'informations des données récoltées au travers deux streams de données :

- Le premier Stream de données est axé sur les informations des clients de Natixis et sur les opérations d'achat/vente qu'ils effectuent avec Natixis.
- Le deuxième Stream de données concerne les informations marché (le carnet de liquidité et les informations financières).

Pour réaliser ce travail, j'ai commencé par l'élaboration d'une étude approfondie des technologies Big Data et son ecosystem, du spark, des bases de données NOsql, du data mining et de machine learning.

Ensuite a travers cette étude j'ai mis en place d'une architecture Big data. Pendant ce travail j'ai utiliser une machine virtuelle pour faire les differents test.

1 Présentation de l'entreprise

Natixis est la banque de financement, de gestion, d'assurance et de services financiers du Groupe BPCE, deuxième acteur bancaire en France avec 36millions de clients à travers ses deux réseaux, Banque Populaire et Caisse d'Epargne.

Avec plus de 16000 collaborateurs, Natixis intervient dans trois domaines d'activités au sein desquels elle dispose d'expertises métiers fortes : la Banque de Grande Clientèle, l'Épargne l'Assurance et les Services Financiers Spécialisés.

Elle accompagne de manière durable, dans le monde entier, sa propre clientèle d'entreprises, d'institutions financières et d'investisseurs institutionnels et la clientèle de particuliers, professionnels et PME des deux réseaux du Groupe BPCE.

1.1 Présentation de la DSI-GM-Global Markets-Distribution

2 Description de la mission de stage

3 Le stage

4 Partie 1 : Choix des technologies

4.1 La technologie Big Data

4.2 Définition

4.3 Conception d'une architecture lambda

4.4 Comparaison des bases de données NOSQL

Comparation des bases de données NOSQL : Cassandra, MongoDB et Hbase

Les bases données NOSQL (Not Only SQL) sont comparer sous le fameux théorème CAP (Consistency, Availibility, Partition tolerance).

Consistency : toutes données sont présenteé sous la meme forme sur tous les nœuds du réseau. Une mise a jour rapide.

Availibility : disponible et accessible à tout instant et chaque requête reçoit une réponse qui confirme si elle a été traitée avec succès ou non.

Partition tolerance : Le système doit pouvoir fonctionner lorsque différents nœuds sont isolés consécutivement à une rupture du réseau. Le théorème stipule qu'au aucun système distribué ne peut satisfaire les trois conditions en même temps.

Cassandra : C'est une base donnée NOSQL avec un Stockage de données par colonnes, il est facile d'ajouter une colonne avec un schéma dynamique. Il supporte les semi-structurées avec une indexation de chaque colonne, un passage à l'échelle horizontal et un langage de requête CQL (Cassandra Query Language) proche du SQL.

Il permet une répartition robuste sur plusieurs serveurs. C'est pour le traitement à froid à fin d'analyse conçu pour les applications en ligne qui ont besoin d'une vitesse et d'une disponibilité élevées. Des données organisées dans un keystore (équivalent d'une base donnée relationnelle). correspond au AP (disponibilité et tolerance au partitionnement) du théorème CAP.

Mode de distribution :

un mode de distribution décentralisé, chaque nœud est indépendant et il n'y pas besoin d'un serveur maître. Ne nécessitant pas un système de fichiers distribué avec une performance en écriture. Cassandra est un produit mature largement utilisé et très populaire. C'est une excellente solution pour bâtir un système de gestion de données volumineux et décentralisé.

Hbase :

C'est une base de données NOSQL de hadoop par défaut, avec un stockage de données par colonnes, une scalabilité linéaire, une absence de tout mécanisme d'indexation pour les colonnes. Respect le CP du théorème CAP (Consistance et tolérance au partionnement).

Mode de distribution : Nécessite un système de fichier distribué. Une architecture maitre –esclave (NameNode et les dataNodes).

MongoDB :

C'est une base donnée NOSQL avec un stockage de données par clé/valeur orienté document. Il supporte aussi des données structurées, mais complexe pour le CRUD (Create, Read,Update et Delete).

Mode de distribution : un mode de distribution centralisé. Il respect le CP du théorème CAP (Consistance et tolérance au partionnement).

4.5 Choix d'une base de données NOSQL

Après avoir manipulé quelques fonctionnalités, remarqué la performance et la solidité de MongoDB dans les travaux pratiques à l'université, j'ai aussi

installé et manipulé Apache Cassandra, Hbase pour faire des tests. Après une Vu les besoins dans ce service où on doit faire des calculs robustes pour avoir une disponibilité et une réponse à nos requêtes,mon choix de la base donnée NOSQL a été porté à Apache Cassandra qui est plus adapté à cette problématique.

A fin de valider ce choix, l'avis d'un expert Big Data etait importante. pour cela mon manager a organiser un entretien téléphonique avec un expert big data (hbase) de natixis.

4.6 Choix d'une distribution de hadoop et son ecosystème

4.7 Choix d'un ETL

4.8 Choix des langages de programmation

4.9 Spark

4.10 Spark SQL

4.11 Spark Streaming

4.12 Spark MLib

On peut mettre des mots en *italique*, en PETITES MAJUSCULES ou en `largeur fixe (machine à écrire)`.

Voici un deuxième paragraphe avec une formule mathématique simple : $e = mc^2$.

Un troisième avec des « guillemet français ».

5 Partie 2 : Présentation du travail

5.1 Conception et réalisation du projet

5.2 Modelisation

5.3 Stockage des données dans la base de données Hbase

5.4 Conception d'une fonction de ....

5.5 Analyse descriptive des données

6 Partie 3

Do you speak French? Does anybody here speak french?

6.1 Traitement en mode streaming

6.2 Analyse prédictive des données

- Liste classique ;
- un élément ;
- et un autre élément.

1. Une liste numéroté
2. deux
3. trois

Description C'est bien pour des définitions.

Deux Ou pour faire un liste spéciale.

Voici une référence à l'image de la figure 1 page 8 et une autre vers la partie 8 page 9.

On peut citer un livre [?] et on précise les détails à la fin du rapport dans la partie références.

Voici une note [1] de bas de page. Une deuxième [2] déclarée différemment. La même note [2].

1. Texte de bas de page
2. Il a deux références vers cette note

FIGURE 1 – BlogHiko | taille original

FIGURE 2 – BlogHiko | 50% de la largeur de la page

7 Apport des enseignements dans le développement de la solution

8 compétences acquisent

LaTeX est un langage et un système de composition de documents créé par Leslie Lamport en 198312. Plus exactement, il s'agit d'une collection de macro-commandes destinées à faciliter l'utilisation du « processeur de texte » TeX de Donald Knuth. Depuis 1993, il est maintenu par le LaTeX3 Project team. La première version utilisée largement, appelée LaTeX2.09, est sortie en 1984. Une révision majeure, appelée LaTeX2 epsilon est sortie en 1991.

Le nom est l'abréviation de Lamport TeX. On écrit souvent LaTeX, le logiciel permettant les mises en forme correspondant au logo.

Du fait de sa relative simplicité, il est devenu la méthode privilégiée d'écriture de documents scientifiques employant TeX. Il est particulièrement utilisé dans les domaines techniques et scientifiques pour la production de documents de taille moyenne ou importante (thèse ou livre, par exemple). Néanmoins, il peut aussi être employé pour générer des documents de types variés (par exemple, des lettres, ou des transparents).

Conclusion

Pour conclure, avec LaTeX on obtient un rendu impeccable mais il faut s'investir pour le prendre en main.

Références

[REF] Pirmin Lemberger, Marc Batty, Médéric Morel, Jean-Luc Raffaiëlli. *Big Data et Machine Learning*. edition, 2016.

[REF] Mohammed Guller. *Big Data Analytics with Spark*. O'Reilly, 2015.

[REF] youtub.com