Big Data : Challenge & Opportunities

Gheysen Jérémy

16 mars 2016

1 Introduction

Hermes = société de services IT 3 mois de formation dès l'entrée en entreprise. Hermes fait du service : propose des solutions en fonction des problèmes du client, gestion de projet.. Actifs en finances/assurances/télécom/pharmaceutique...

2 Business Intelligence

Panel d'applications & technologies permettant de stocker/analyser des données permettant de prendre des décisions. **Predictive analytics** : Solution statistique data-mining permettant de... **BI and predictive analytics** : Types de questions ? Cb de clients vont pouvoir visiter mon magasin la semaine prochaine, tendance pour les recettes futures ? => Tendance dans le futur (pour la predictive analysis). Pour le BI : collecte des informations pour permettre de créer de la connaissance qui a du sens. Permet de prendre des décision et de créer des actions qui vont modifier notre business.

3 Data Warehouse

Ferme de données, avoir une vérité unique au sein de la société. Grouper tout en un seul endroit et créer une seule vision, une seule vérité. Comment fait-on ? On part de différentes sources -> processus ETL (extraction-transformation-load) qui perlet de charger toutes les données. Puis différents moyens d'exploiter les données. Autre déf : Ensemble de donnée orientées sur un sujet, intégrées, non-volatiles et pouet...

- Orientées sujet : différents sujet
- Intégrées : encodage unique au sein du datawarehouse.
- non-volatile : rentre dans le data warehouse et n'a pas de vocation d'en sortir excepté pour l'extraire.
- Historisation : Certain horizon de temps 60-90 jours , peut contenir une notion de temps (pour l'opérationnel) ; 5-10 ans (data warehouse) contient d'office une notion de temps

2 approches du data-wh Approche top down - créer un modèle complet en ayant une vue complète de la société. Modélise puis crée la représentation physique. Pb -> Bcp de temps pour la phase analyse. Av -> Vue globale de la société , on aura déjà tout pensé Approche Bottom - up : se base sur des sujets métiers, l'un après l'autre, (travaille avec data marks, mart ? et les fait évoluer). Pb -> bcp de retravail à faire.

4 Digitalization

Avec un ensemble de donnée, on peut modéliser la vie de quelqu'un et ainsi la capitaliser. Habitudes de vie par exemple. **Qu'est ce que le big data ?** Au début : récolte de données, stockage puis organisation des données. Google lui a voulu indexé ttes les pages sur le net, mais énormément de données. Donc il a fallu trouver un new way pour effectuer ce stockage. Google avait un pb de big data : 3 axes - volume de données - vélocité (tenir compte de ttes les mises à jour faites sur les sites) - variété (texte, images, vidéos...) Solution :

Infrastructure en cluster, plusieurs machines distribuées en data file center. Au dessus google file system : plutot que de stocker un fichier volumineux sur une machine. Division d'un fichier en blocs, + réplications des blocs sur différentes machines, car énormément de machines et donc crashs fréquents. Ensuite Google a créé big table : base de donnée distribuée, données stockées sous formes de clés valeurs. (3 clés pour avoir une valeur). 3 ème grosse invention de Google : Map-reduce = façon de programmé adaptée à la programmation sur fichiers distribués. Distribution du calcul des morceaux fichiers sur les machines les stockant et envoi du résulat sur la machine "mère".

5 Use case

Que faire de tous les outils, comment les optimiser ? Donnes -> ETL -> extractions. Méthodes traditionnelles d'ETL sont cassées car trop de volume et trop de variété. On va venir interroger les bdd avec des outils spécifiques et on les stocke dans HDFS, on interroge avec d'autres outils puis les stocke dans le datawh

6 Data Science

L'art tourner les données en actions. But est de fournir des datas products, produit infos permettant d'agir. Pourquoi est-ce différent de l'analyse classique ? On tire des conclusions de faits , formuler hypothèse sur relations et modèles (approche déductive), + questionnement sur ce qui arrivera en faisant telle ou telle chose, approche explorative (approche inductive). Business Intelligence plus orientée vers le passé, sur l'analyse des faits passé. En data science on interagit sur les données, on essaie de formuler des hypothèses sur le futur. Plusieurs étapes :

- Acquérir les données, attention aux données erronées, abberantes, manquantes. Techniques existantes pour les supprimer ou compléter en fct du pb.
- Preparation
- Analyse, mise en place de modèles (prédiction pour la suggestion de films par ex), si après 2 sem on remarque que ça va pas, on remet en question le modèle, modifie ou crée un nouveau.
- Act, réalisation, au niveau du client. **Google Flu Trends** Prédire en real-time l'évolution de la grippe. En fonction des areas de recherche de la grippe, prédire la propagation de l'épidémie. Très bien fonctionné au début. En 2011, overestimation de l'influence de la grippe. Les gens ont trop recherché sur le net sous l'influence des médias donc le flu trend n'a pas pris ça en compte et le modèle n'était plus très adapté du coup.