SHAKE THE FUTURE

CENTRALE NANTES

Bases de Données

Notions de Big Data

JY Martin

Plan

Vers le Big DATA

- Chaque jour : 2,5 trillons d'octets de données
- 90 % des données dans le monde ont été générées ces dernières années
  - Capteurs
  - Messages sur les médias sociaux
  - Images et vidéos
  - Achats en ligne
  - GPS
  - ...

SDSS

Sloan Digital Sky Survey

- Carte tridimensionnelle (1/3 de la voute céleste)
  - 470 millions d'astres
  - 2000 Galaxies
- 10 ans de projet pour comprendre la Voie Lactée et découvrir des exo-planètes
- Images d'1 peta pixels (10**15 pixels)
  = 500000 écrans HD pour la visualiser
  = 71 peta octets de données

Autres exemples

- Déforestation : PlanetarySKIN : 7 tera octets de données
- Astronomie : LSST : 30 tera de données / nuit
- Micro-organismes marins : Projet GOS : 2 tera octets de données
- Bio Chimie : Projet BSrC – 100 millions de molécules
- Cancer du foie : Projet ICGC : 200 tera octets d'analyses
- Détection d'épidémies en temps réel
- ...

Questions

Comment stocker de tels volumes de données ?

Comment les rendre accessibles ?

Comment effectuer des traitements sur de telles masses de données ?

Plan

Le Big Data

Big DATA = A l'intersection des 3 V

- Volume
  Les volumes de données à traiter sont de plus en plus gros
- Variety (Variété)
  Données de plus en plus complexes, de moins en moins structurées
- Velocity (Vitesse)
  Données de plus en plus saisies et traitées à la volée

Le Volume de Données

Le prix du stockage de données baisse régulièrement

Nombreuses solution de stockage fiables

PB : comment déterminer quelles données méritent d'être stockées ?

La Variété des données

- Les données sont de plus en plus non structurées ou semi-structurées
- Faut-il stocker la donnée sous plusieurs formats, plusieurs types ?
- Des données “obsolètes” peuvent être utiles pour des prises de décisions.
  - Comment les conserver ?
  - Sous quelle forme ?

La vitesse d'évolution

- Les données doivent être traitées de plus en plus rapidement (temps de calcul)
- Les données doivent être traitées de plus en plus rapidement (dans la chaine de traitement)

Une autre façon de voir les données

- Réseaux Sociaux
- Services de recommandations
- Analyse et prédiction du marché
- Relation plus personnelle avec le client
- Ciblage des publicités
- Réactivité
- ...

Vers de nouveaux métiers

Exemple : Data Scientist

- Spécialiste IT
- Développement, déploiement d'outils
- Gestion du parallélisme
- Statistiques
- Gestionnaire

Qu'est-ce que ça change ?

Les données sont là. Qu'est-ce qu'on en fait ?
=>Les sources guident la découverte

- Identification des données disponibles
- Plateforme d'exploration des données
- Quelles analyses doit-on conduire ?
- Introduction de nouvelles technologies pour capter d'autres informations

Plan

Les outils du Big Data

- Outils informatiques
- Outils Mathématiques

Les outils Mathématiques

Mise en oeuvre

- de modèle d'analyse
- de modèles de prédiction
- ...

Font que moyenne et écart type ne sont pas deo outils suffisants.
Nécessité de modèles plus évolués (voir Option Maths)

Les outils Informatiques

- Mémoriser l'information
- Traiter l'information

Mémoriser l'information

Big Data

- Beaucoup de données
- Des données pas forcément bien structurées
- Qu'on n'a pas le temps d'exploiter en temps réel

Mémoriser l'information

=>abandon du relationnel au profit du NoSQL

Si possible distribué

Traiter l'information

Modèles de calcul performants

- Donnée distribuée
- Volumineuse

=>Utiliser des outils de calcul parallèle

Hadoop, Spark

Plan

Conclusion

SBGDR / noSQL – Big Data

- Approche très différentes de la donnée
  - Acquisition (grosse masse de données)
  - Manipulation (Utilisation de fermes de serveurs)
  - Traitement (Map-Reduce)
- Ca fonctionne de manière très différente
- Ce n'est pas fait pour la même chose


Success
NEXT EXIT