Immersion BD et infostructure

Focus sur map-reduce

4 V fondamentaux : volume velocité variablité et variété
1ere approche nouvelle : pas de schéma, ou variable (V : variability)
Notion de variété : les sources de données sont hétérogène (web, fb, open data, gouvernement, flux vidéo, sms) (V : variety)
Les données sont en temps réel : chaque seconde sur internet il y a 300 000 tweets, sms, Les requêtes google dépasse le million, plus d'une heure de vidéo youtube publié.

Question fondamentale du Big Data : comment les exploiter ?

- ➔ Définir des patern, des corrélations
  ex : temps passé sur fb et temps en forme (humeur)
  identification des quartiers à risque d'incendie de new york
  - là où y'a des ravalement de façade, peu de risque (pas de conflit)
  - là où il y a des sous-locataires
  - Corrélation : immeuble à risque

La corrélation dans big data précède la causalité.

Peta = 1000 tera
Exa = 1 million de tera : 10 ** 18
Zetta = 1000 Exa : 10 * 24
IDC 2012 a analysé 1.8 Zetta octets de données produites en 2011 avec 50% de croissance par an

Dans les 3000 première année de l'humanité : 5 Exa octets de produit.
2012 : 5 Exa tout les 2 jours
2013 : toutes les 10 min
2015 : toutes les 10 secs

- ➔ Data Tsunami : ce grand volume arrive

Pour illustrer : http://onesecond.designly.com

Terme yotta : espionnage : surveiller réseau sociaux, ect
Ambition nsa : gérer big data du yota : 10 ** 32
Google en anglais signifie : 10 ** 99

Tsunami data : déluge data :

- Bottom up
- Temps reel

Big data : serveur (cloud), mobiquité (portable, tag nfc), réseaux sociaux, analyse

Application des big data : cartographie, océanographie, astronomie, génétique, médecine, épidémie, linguistique, Macro-economie, transport (VAMP), cartographie temps réel (FIRST), Commerce, tourisme (Matrium, reve, imajean + projet city wallet)
Résumé :

Avant big data : recherche scientifique type
Problème : état de l'art
Posé une hypothèse, trouver des impasses, des solutions
Avoir une intuition
Valider par des expériences, simulations, calculs

Avec big data :
Analyse informatique :
Identification de corrélations nouvelles →génératrice d'hypothèse
Emergence de découvertes

Pour pouvoir modéliser et interroger le web il faut un schéma : rdf et un langage de manipulation : sparql
Rdf permet de lier les données ensembles : principe du web sémantique

Rdf est un ensemble de triplet : <sujet, prédicat, objet>
Rdf / XML est une recommandation du W3C

Sparql : langage d'interrogation, ajout, modification et suppression des données RDF

---

3 approches pour faire du big data :

SQL 2 / SQL 3 (quelques petas octets)
NO SQL
NEW SQL

Echo système bd top dow sql

G / P / D -> gestion production distribution

SQL : données structuré (table, objet)
Approche top down

No SQL : donnée non structuré (web par url)
Approche bottom up

BASE :
Basically
Available
Scalabilty
Eventually consistent

CAP Theorem (en choisir 2)
Consistency } sql
Availability
Partitionning } nosql

Meilleures approches pour big data : no sql et new sql

4 types de base no sql :

- clé-valeur (suivi par hadoop et cassandra) : table de hachage
- orientées colonne : stockage par colonnes (pas par ligne)
- orientées document (mongodb) : pas de schéma, ensemble de clé valeur
- graphe (pour réseau sociaux : neo4j) : nœuds/lien/propriété, pas d'operateur ensembliste mais parcourt de graph

Map Reduce (à la base dans les calculs distribué):
Map : étape de transformation de données sous la forme d'un couple clé valeur
Reduce : on fusionne les enregistrements par clé pour obtenir le résultat final

4 phrases de map reduce :
splitting : sépare les mot
mapping : on compte l'occurrence des mot
suffling : on regroupe les valeurs
reduce : fusion des valeurs avec additions

Inconvénient : c'est celui qui a lancer la recherche qui traine le résultat, trop compliqué d'utiliser une sortie sans en connaître l'entrée

Prendre le meilleurs des 2 mondes (approche Stonebraker): faire un système amphibiens : passerelle entre sqgb(top down) et décisionnel (bottom up)