Étude d’approfondissement

Social Data Engineering

Walid Bibi

2013/2014

Plan de la présentation

I- Le contexte

II- Pourquoi le social data engineering

III- Hadoop

Les réseaux sociaux

Facebook : 800 millions d'utilisateurs
Twitter : 200 millions d'abonnés
Google + : 10 millions d'abonnés
Youtube : 490 millions de visiteurs uniques chaque mois

Par jour : - 50 millions de tweet
- 60 millions d'actualités facebook
- 1 milliard de mises à jour de profils facebook

→ Big data = péta octets de données non structurées (email + images + données sociales…)

Utilité ? Pour qui?

Objectif : générer du chiffre d'affaires.
attirer / fidéliser les clients.

Des questions existentielles : Qui achète le plus ?
Quel est le client le plus précieux ?
Que recherche le client ?

L'utilité des réseaux sociaux : établir une stratégie marketing.

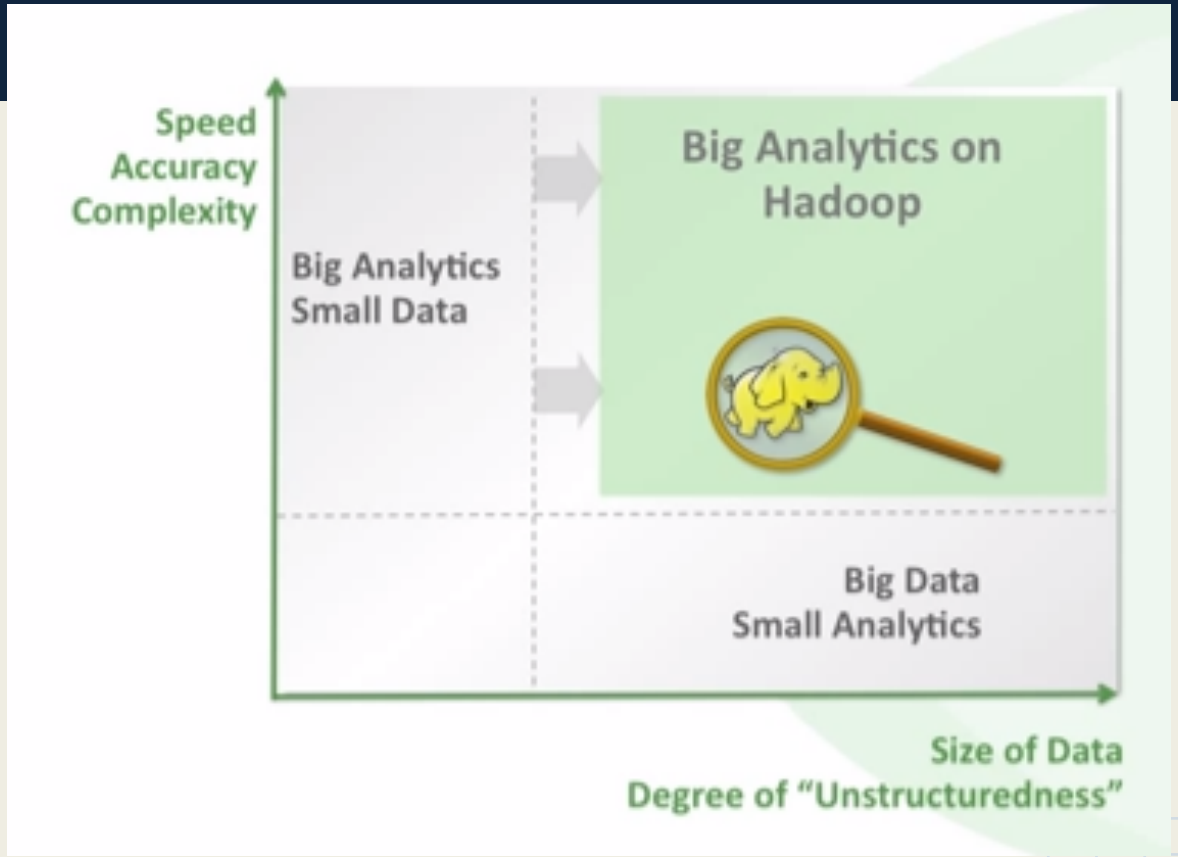
Problèmes pour les entreprises

Données incomplètes sur les clients (ERP,CRM).

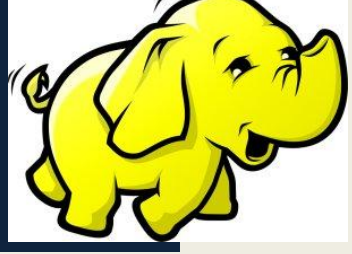
Gros volumes d'informations présentes dans les réseaux sociaux.

→ 20% présentent une utilité marketing.

Pas de technologie pour traiter les données des réseaux sociaux.


Speed
Accuracy
Complexity
Big Analytics
Small Data
Big Analytics on
Hadoop
Big Data
Small Analytics
Size of Data
Degree of "Unstructuredness"

Hadoop



Framework open source écrit en java.

Réaliser des traitements de volumes de données en masse.

Traitement massivement parallèle = répartir les données de calcul sur plusieurs noeuds : HDFS

Gestion de calcul : MapReduce

Hadoop Distributed File System

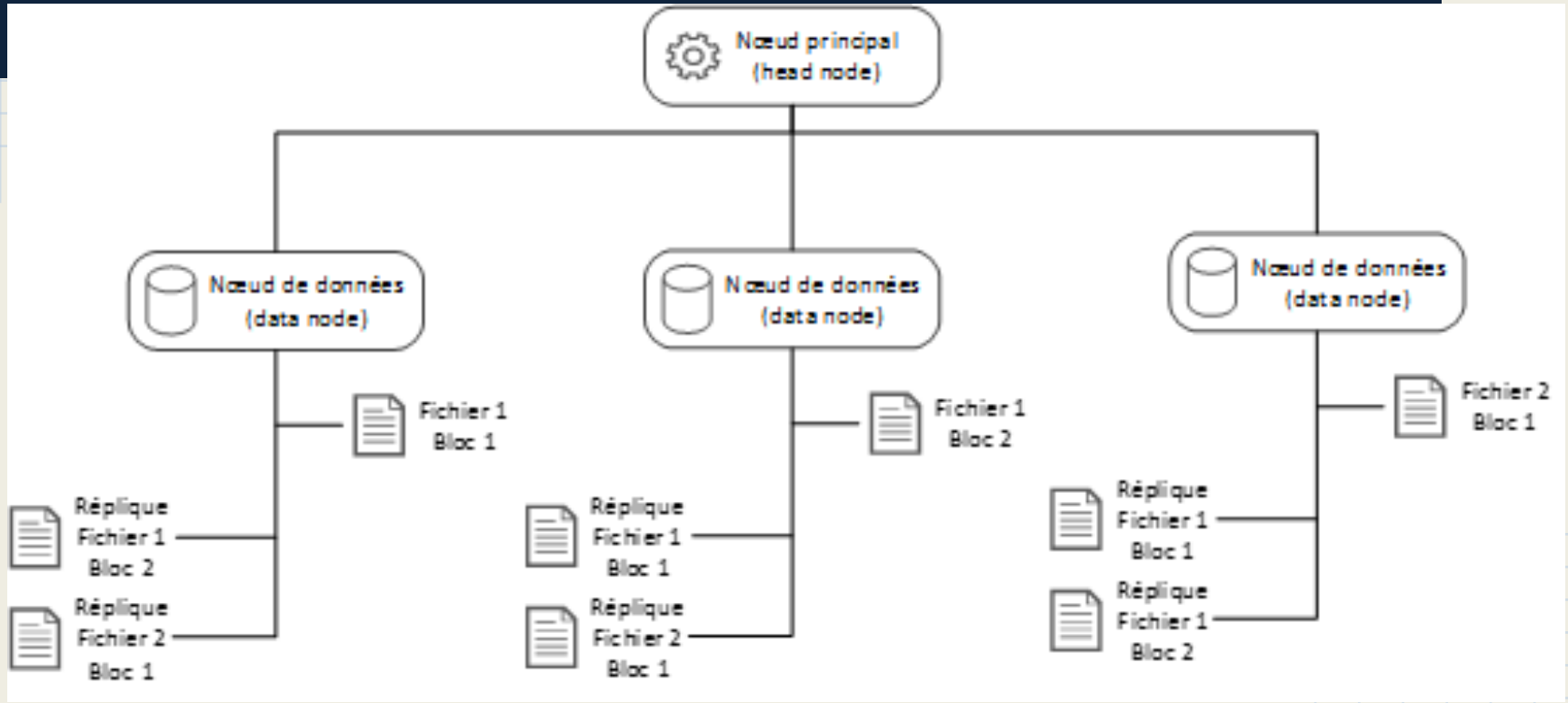
- Système distribué.

- Création d’applications échelonnables.

Architecture :

  - NameNode : répertorier où sont stockées les données.

  - DataNode : stockage brute des données.


Nœud principal
(head node)
Nœud de données
(data node)
Nœud de données
(data node)
Nœud de données
(data node)
Fichier 1
Bloc 1
Fichier 1
Bloc 2
Fichier 2
Bloc 1
Réplique
Fichier 1
Bloc 2
Réplique
Fichier 2
Bloc 1
Réplique
Fichier 1
Bloc 1
Réplique
Fichier 2
Bloc 1
Réplique
Fichier 1
Bloc 1
Réplique
Fichier 1
Bloc 2

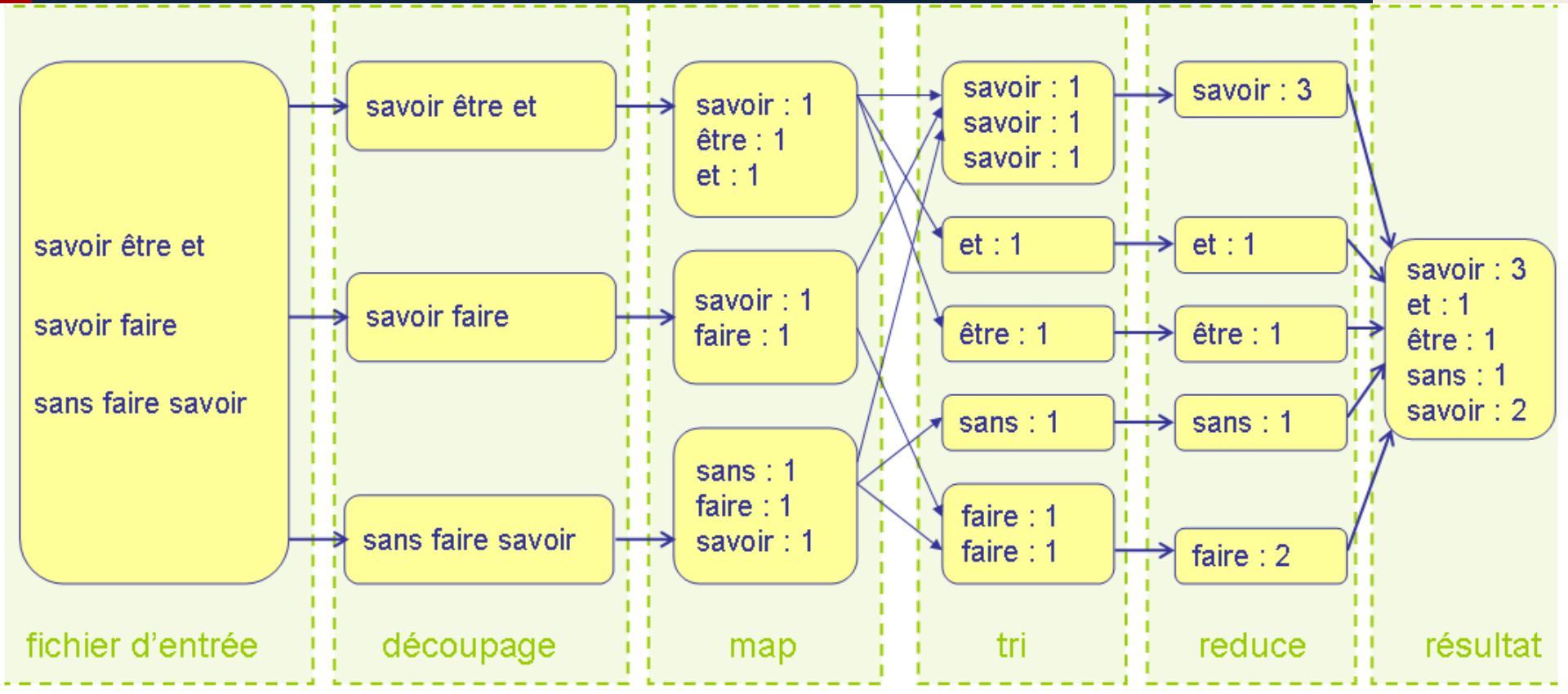
MapReduce

Technologie introduite par Google.

Deux types de fonctions :

- Map : étape de transformation de données
- Reduce : étape de fusion des enregistrements

Étape intermédiaire de tri

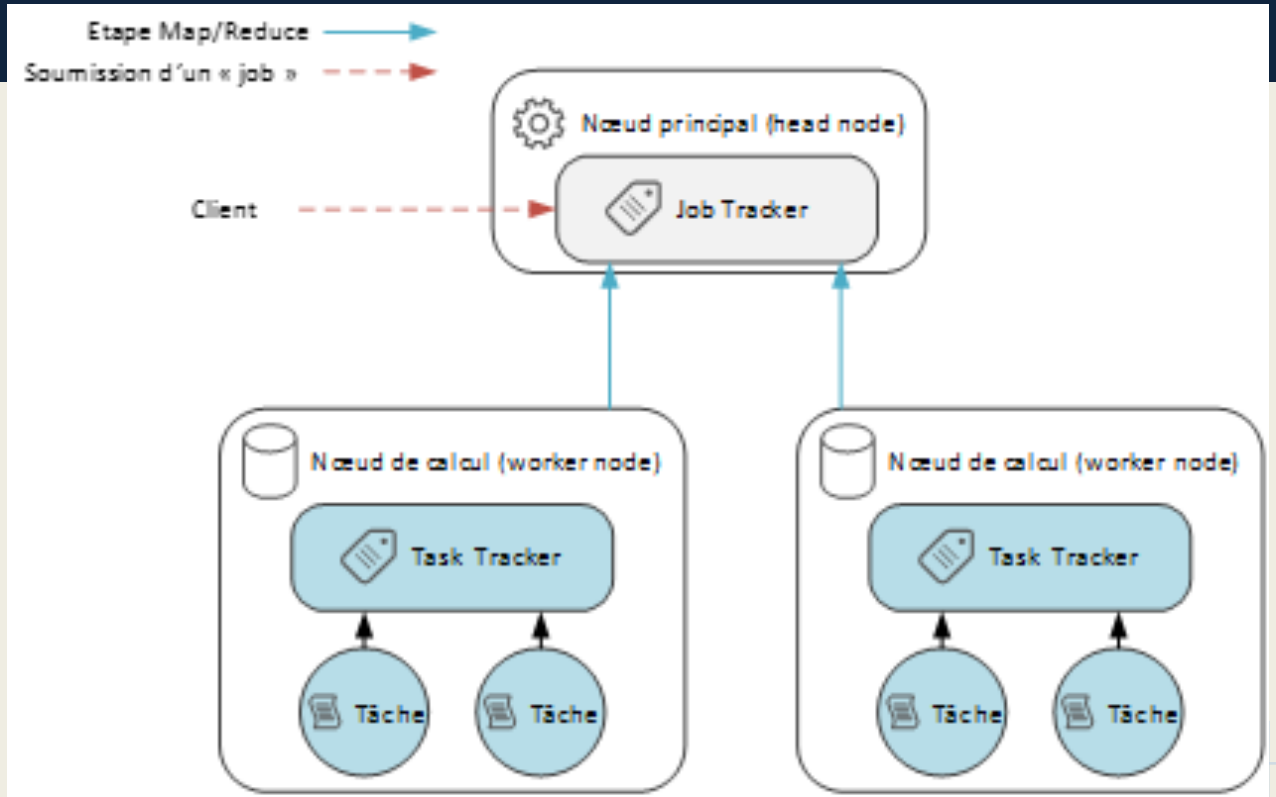

savoir être et
savoir faire
sans faire savoir
savoir être et
savoir faire
sans faire savoir
savoir : 1
être : 1
et : 1
savoir : 1
faire : 1
sans : 1
faire : 1
savoir : 1
savoir : 1
savoir : 1
et : 1
être : 1
sans : 1
faire : 1
faire : 1
savoir : 3
et : 1
être : 1
sans : 1
faire : 2
savoir : 3
et : 1
être : 1
sans : 1
savoir : 2
fichier d'entrée
découpage
map
tri
reduce
résultat

Exemple : nombre de visites d’un blog

class Main : on appelle un “job” = type de fichier en entrée.
le répertoire en entrée.
type de fichier en sortie.
le fichier de sortie qui contient le résultat.

Une classe MonMapperImpl : méthode map().

Une classe MonReducerImpl : méthode reduce().


Etape Map/Reduce
Soumission d'un « job »
Client
Nœud principal (head node)
Job Tracker
Nœud de calcul (worker node)
Task Tracker
Tâche
Tâche
Nœud de calcul (worker node)
Task Tracker
Tâche
Tâche

Conclusion

- Big Data : des pétaoctets de données non structurées.

- Objectif : trier ces données afin de les exploiter.

- Hadoop : outil le plus puissant pour cela grâce à HDFS et MapReduce.

- Utilisé par : twitter, facebook, linkedIn, amazon etc etc...

Sources

http://blogs.msdn.com/b/big_data_france/archive/2013/03/25/vous-avez-dit-hadoop-1-232-re-partie.aspx

http://www.it-expertise.com/exploiter-toute-la-puissance-des-donnees-des-medias-et-reseaux-sociaux/

http://www.liglab.fr/spip.php?article934