esprit
Ecole Supérieure Privée
d'Ingénierie et de Technologies

Chp2 – Ecosystème Hadoop

2016-2017

Mme Ines Slimene
ines.slimen@esprit.tn

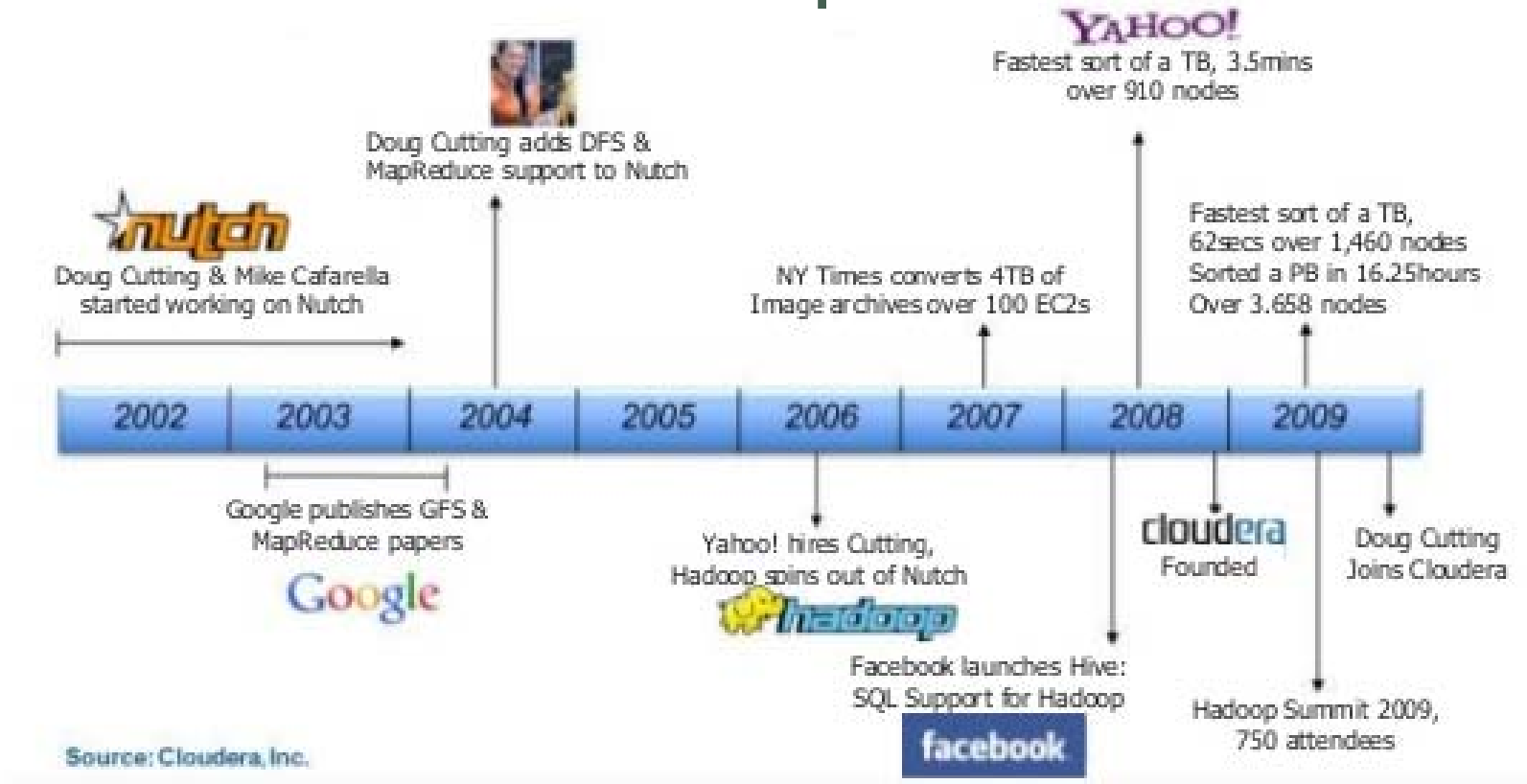
Plan module

- Introduction
- **Écosystème Hadoop**
- HDFS
- MapReduce
- Langages de requête Hadoop :Pig, Hive
- SGBDNR
  - Différences entre une BDNR et une BD relationnelle
  - Typologies des BD non relationnelles
- Etude d'un SGBDNR : HBase

Plan

- Historique
- Ecosystème Hadoop
- Architecture
- Distributions Hadoop
- Atout Hadoop
- Inconvénients de Hadoop

Histoire de Hadoop



Hadoop

- Stockage et traitement de données volumineuses
- Bien qu'il peut aussi bien fonctionner sur une seule machine, sa vraie puissance n'est visible qu'à partir d'un environnement composé de plusieurs ordinateurs.
- La multiplication de l'espace disque ne va pas avec l'accélération de la lecture des données.

Diviser les données en plusieurs parties pour les stocker sur plusieurs machines.

Atouts de Hadoop

- La gestion des défaillances : que ce soit au niveau du stockage ou traitement, les nœuds responsables de ces opérations durant le processus de Hadoop sont automatiquement gérés en cas de défaillance.
- La sécurité et persistance des données : il n'y a plus de soucis de perte de données.
- La montée en charge : garantie d'une montée en charge maximale.
- La complexité réduite : capacité d'analyse et de traitement des données à grande échelle.
- Le coût réduit : Hadoop est open source, et malgré leur massivité et complexité, les données sont traitées efficacement et à très faible coût

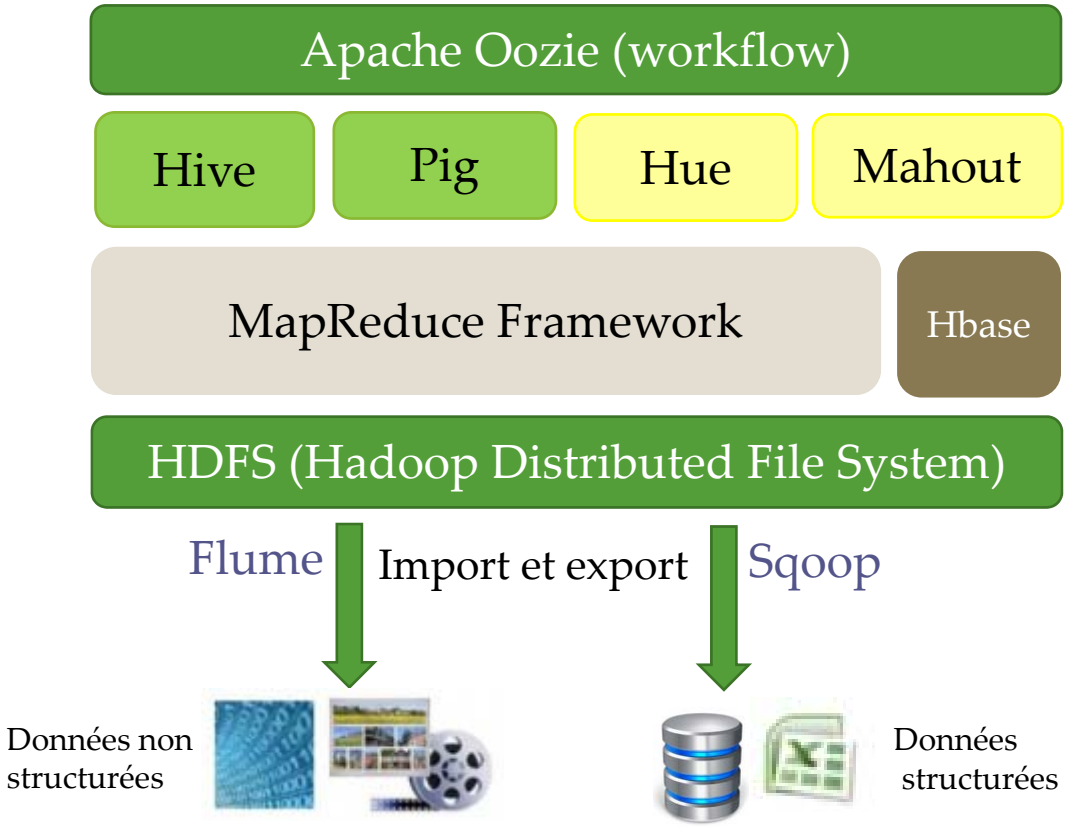
Hadoop

- Hadoop n'est pas conçu pour des requêtes temps réel ou de faible latence.
- Hadoop est performant dans le traitement des batch hors ligne d'un grand volume de données.
- Les SGBDNR sont meilleures pour les transactions en ligne et de faible latence.

Hadoop

- Le projet Hadoop consiste essentiellement en deux grandes parties:
  - Stockage des données : HDFS (Hadoop Distributed File System)
  - Traitement des données : MapReduce
- Principe :
  - Diviser les données
  - Les sauvegarder sur une collection de machines, appelées cluster
  - Traiter les données directement là où elles sont stockées, plutôt que de les copier à partir d'un serveur distribué
- Il est possible d'ajouter des machines au cluster, au fur et à mesure que les données augmentent

Ecosystème Hadoop



Ecosystème Hadoop

- Au début, Hadoop a été connu par ses deux principaux composants
  - **HDFS** : Hadoop Distributed FileSystem
  - **MapReduce** : framework de traitement des données distribuées

- Aujourd'hui, en plus de HDFS et MapReduce, on trouve plusieurs autres composants :
  - **HBase**: base de données Hadoop orientée colonne; supporte batch et lecture aléatoire.
  - **Oozie**: Planificateur et manager du workflow Hadoop
  - **Pig**: Langage de traitement de données
  - **Hive**: Data warehouse avec interface SQL.

Ecosystème Hadoop

- **Hue** :
  - Front-end graphique pour le cluster
  - Fournit
    - Un navigateur pour **HDFS** et HBase
    - Des éditeurs pour Hive, Pig, Impala et Sqoop
- **Mahout :**
  - Bibliothèque d'apprentissage automatique
  - Permet de :
    - Déterminer des éléments qu'un utilisateur pourra apprécier selon son comportement
    - Grouper des documents
    - Affecter automatiquement des catégories aux documents

Ecosystème Hadoop

- Connexion du **HDFS** à partir d'outils externes
  - **Sqoop** :
    - Prend les données à partir d'une base de données traditionnelle, et les met dans **HDFS**, comme étant des fichiers délimités, pour être traitées avec d'autres données dans le cluster
  - **Flume**:
    - Système distribué permettant de collecter, regrouper et déplacer efficacement un ensemble de données (des logs) à partir de plusieurs sources vers le **HDFS**

Distribution Hadoop

- Les Distributions Hadoop vise a résoudre les problèmes d'incompatibilité des versions.
- Les vendeurs des distributions vont :
  - Intégrer un jeu de test du produit Hadoop
  - Packager les produits Hadoop dans divers format d'installation
  - Les distributions peuvent fournir des scripts additionnels pour exécuter Hadoop
  - Quelques vendeurs peuvent choisir de reporter les fonctionnalités et les bugs corrigés faites par Apache

Vendeurs de distribution

- Cloudera Distribution for Hadoop (CDH)

- MapR Distribution

- Hortonworks Data Platform (HDP)

- IBM InfoSphere BigInsights