

Big Data Platforms

Résumé du Cours

Framework Hadoop

Pr. Btihal El Ghali

1 Introduction au Big Data

1.1 Évolution des Données

- De la création jusqu'en 2003 : 5 milliards de gigaoctets produits
- 2011 : Même volume créé tous les 2 jours
- 2013 : Même volume créé toutes les 10 minutes
- 90% des données mondiales générées ces dernières années

1.2 Échelle des Données

Unité	Taille	Exemple
Kiloctet (Ko)	10^3	20 Ko = Mots/jour d'une personne
Mégoctet (Mo)	10^6	3 Mo = Édition du NY Times
Gigaoctet (Go)	10^9	1 Go = Mots d'une vie
Téraoctet (To)	10^{12}	593 To = Mots mondiaux/jour
Pétaoctet (Po)	10^{15}	4 Po = Données Facebook/jour
Exaoctet (Eo)	10^{18}	605 Eo = Télescope SKA/jour
Zétaoctet (Zo)	10^{21}	40 Zo = Web 2020/an

1.3 Les 7V du Big Data

1. **Volume** : Quantité massive de données
2. **Vélocité** : Vitesse de génération et traitement
3. **Variété** : Diversité des sources et formats
4. **Véracité** : Fiabilité et qualité des données
5. **Valeur** : Transformation en informations utiles
6. **Visualisation** : Représentation graphique interactive
7. **Viralité** : Propagation et impact des données

2 Motivations

2.1 Limites des Systèmes Traditionnels

- **Bases de données traditionnelles** : Gérable jusqu'à 100 Go
- **Scale-up** : Limites de puissance processeur et mémoire
- **Systèmes distribués classiques** :
 - Bande passante limitée
 - Dépendances complexes entre nœuds
 - Faible tolérance aux pannes

2.2 Problèmes Identifiés

- **Data Bottleneck** : Taux de transfert 75 MB/sec → 22 min pour 100 GB
- **Stockage centralisé** : Copie des données vers processeurs
- **Défaillance partielle** : Perte de données et dégradation progressive

3 Big Data Use Cases

3.1 Exemples d'Applications

- **360° View of Customer** : Tableau de bord intégrant CRM, réseaux sociaux, historique
- **Prévention de Fraude** : Analyse prédictive des transactions suspectes
- **Sécurité Intelligente** : Analyse de logs pour détecter cyberattaques
- **Moteurs de Recommandation** : Suggestions personnalisées (Amazon)
- **Analyse Sentiment Réseaux Sociaux** : Monitoring et réponses appropriées
- **Internet of Things (IoT)** : Maisons/villes intelligentes, voitures connectées
- **Médecine** : Corrélation alimentation-maladies via réfrigérateurs intelligents

3.2 Scandale Cambridge Analytica

- 87 millions de profils Facebook collectés via application tierce
- Utilisation pour campagne électorale Donald Trump
- Conséquences : Perte de 50 milliards \$ pour Facebook

4 Framework Hadoop

4.1 Définition

- Framework open source basé sur Java
- Traitement Big Data en environnements distribués
- Projet Apache Software Foundation

4.2 Historique

- 2004 : Création NDFS par Doug Cutting et Mike Cafarella (Apache Nutch)
- 2006 : Hadoop devient sous-projet Apache Lucene
- 2008 : Yahoo propose Hadoop en open source
- 2011 : Version 1.0.0
- 2012 : Hadoop 2.0 avec introduction de YARN

4.3 Caractéristiques Principales

- **Distributed Computation** : Bring the computation to the data
- **Scalable** : Scale-out (ajout de machines) vs Scale-up
- **Fault-tolerant** : Duplication des données sur nœuds
- **Reliable** : Pas de perte de données malgré pannes
- **Communication minimale** : Réduction du trafic réseau
- **Open Source** : Économique
- **High Availability** : Disponibilité continue

4.4 Noyau de Hadoop

Hadoop 1.0 :

- HDFS (Hadoop Distributed File System) : Stockage distribué
- MapReduce : Traitement parallèle distribué

Hadoop 2.0 :

- YARN : Planification tâches et gestion ressources cluster

4.5 Architecture Cluster Hadoop

Types de Noeuds :

- **Master Node** : NameNode (HDFS) + Resource Manager (YARN)
- **Slave Nodes** : DataNode (HDFS) + Node Manager (YARN)

5 Écosystème Hadoop

5.1 Ingestion de Données

- **Sqoop** : Migration données BD relationnelle ↔ Hadoop
- **Flume** : Collecte et agrégation logs/messages vers HDFS

5.2 Stockage

- **HDFS** : Système de fichiers distribué
- **HBase** : Base NoSQL orientée colonnes

5.3 Traitement et Manipulation

- **Hive** : Requêtage SQL-like sur Hadoop
- **Pig** : Scripting ETL intuitif
- **Mahout** : Framework Machine Learning/AI

5.4 Coordination et Ordonnancement

- **Zookeeper** : Coordination et synchronisation services
- **Oozie** : Gestion et planification jobs Hadoop

5.5 Administration

- **Ambari** : Provisionnement, gestion et monitoring cluster

6 Installation Cloudera CDH

6.1 Étapes d'Installation

1. Télécharger VirtualBox (version adaptée à l'OS)
2. Télécharger Cloudera QuickStart VM

3. Extraire les fichiers du zip
4. Importer le fichier .ovf dans VirtualBox
5. Démarrer la VM Cloudera
6. Accéder au terminal : Applications → System Tools → Terminal

6.2 Cloudera Manager

- Lancement : `sudo /home/cloudera/cloudera-manager -force -express`
- URL d'accès : `http://quickstart.cloudera:7180`
- Fonctionnalités : Ajout clusters, racks, nœuds

Fin du résumé