

PPD : Hadoop Spark

Présenté par : Samuel Lebas et Ryad Tarafi

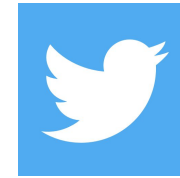
Qu'est-ce que Hadoop?

- Conçu en 2004 par Doug Cutting
- Framework Open Source
- Écrit en Java
- Aide à créer des applications distribuées
- Résiste aux pannes (répliquations des données)



Utilisation de Hadoop

- Stockage moins coûteux de très gros volumes données (Po)
- Réseaux sociaux
- Site de e-commerce
- Analyse de fichiers non structurés (log)



Système distribué par fichiers Hadoop (HDFS)

Schéma de principe du HDFS

- Système de stockage
- Possibilités de stocker des pétaoctets de données
- Traitement parallèle et distribué
- Tolérance aux erreurs avec la réplication des données

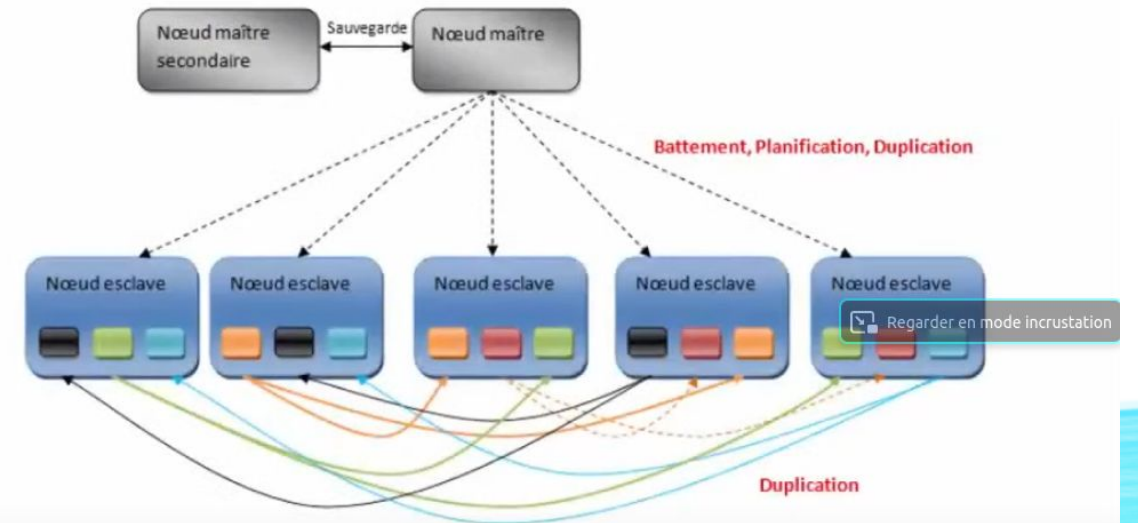
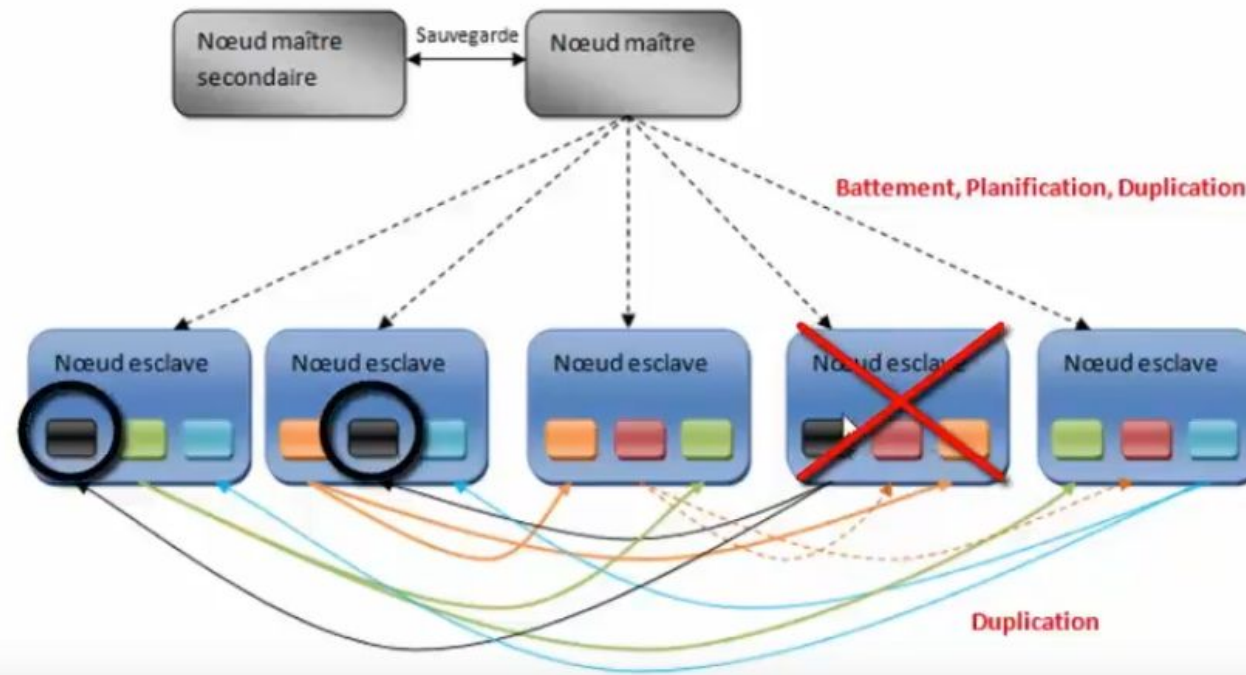


Schéma de principe du HDFS



Qu'est-ce que Apache Spark ?

2014 : Spark a gagné le Daytona GraySort Contest

- Conçu en 2009 par Matei Zaharia
- Framework Open Source sous licence BSD
- 2013 : Licence Apache Software Foundation
- Fonctionnement en mémoire (calcul dans la mémoire)
- 100 fois plus rapide dans la mémoire et même 10 fois plus rapide sur une disque dur classique que Hadoop MapReduce



Applications utilisant Apache Spark

- Analyse prédictive et machine learning
- Opérations d'accès aux données (SQL)
- Traitement et extraction de texte
- Traitement temp réel

Data science and Machine learning



SQL analytics and BI

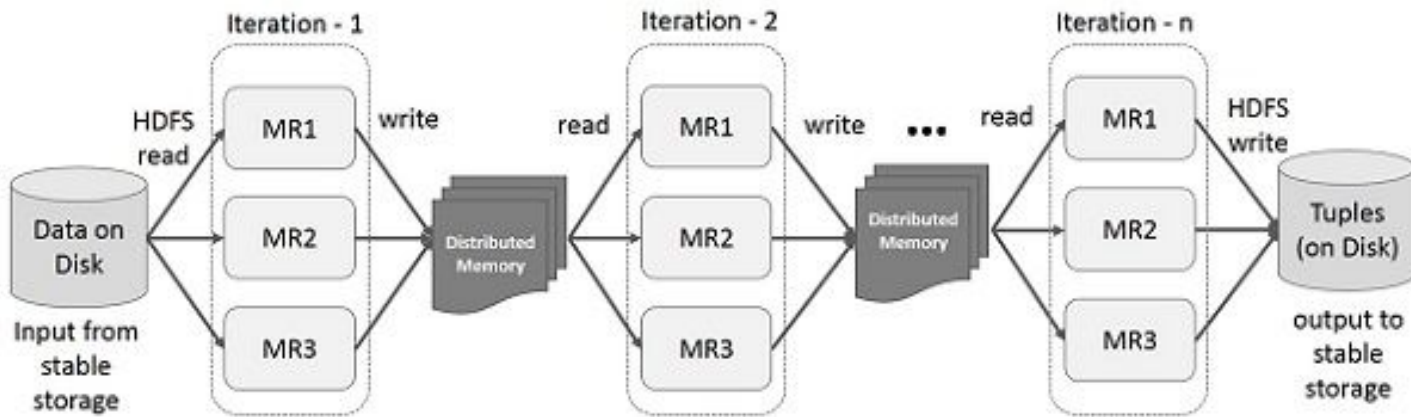


Storage and Infrastructure



Apache Spark RDD

Opération itérative

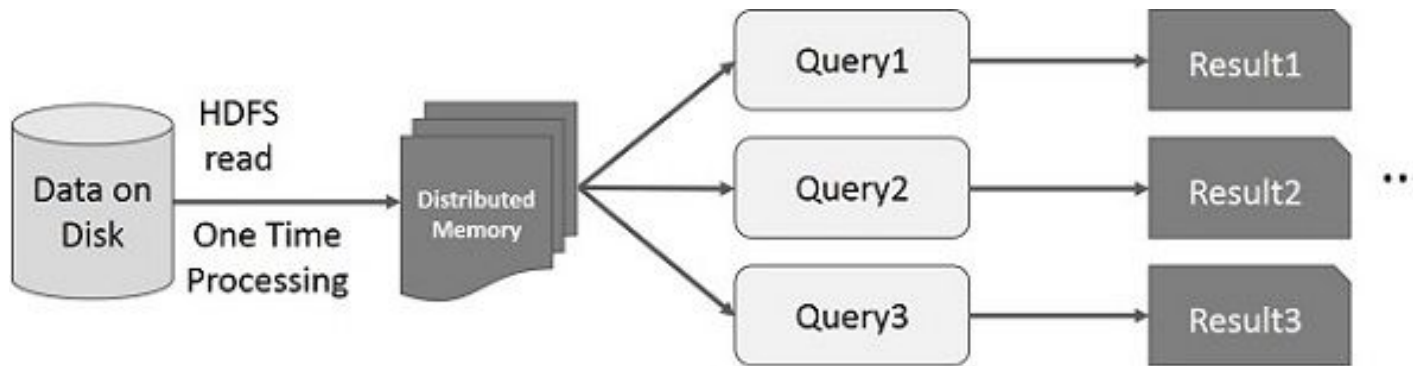


Stockage des résultats intermédiaire sur une mémoire distribuée au lieu d'un disque dur classique ce qui permet au système d'être plus rapide

Attention, si la RAM distribuée ne dispose plus de place suffisante, les résultats intermédiaires (états du job) seront stockés sur le disque.

Apache Spark RDD

Opération interactive



Si différentes requêtes sont exécutées sur le même jeu de données de manière répétée alors ce jeu de données sera gardé en mémoire pour un meilleur temps d'exécution

Hadoop vs Spark

Hadoop

- Coûts inférieurs
- Traitement des données
- Évolutif
- Meilleure sécurité

Spark

- Performance
- Traitement des données
- Évolutif
- Efficace pour le machine learning



DE VOTRE ATTENTION <3