# Hadoop

### Définition d’Hadoop

• **Hadoop :** framework libre et open source écrit en Java qui facilite la création d'applications distribuées et échelonnables pour le stockage et le traitement de gros volumes de données

• Hadoop est apparu pour résoudre les problèmes des 3V (Volume, Vélocité, Variété) des données. Il permet aux applications de travailler avec des milliers de nœuds et des pétaoctets de données.

### Avantages d’Hadoop

• **Gestion de gros volumes de données :** Hadoop peut traiter des quantités massives de données en les fractionnant en blocs et en les distribuant sur des nœuds.

• **Efficacité de stockage :** Hadoop utilise le système de stockage distribué HDFS pour stocker les données de manière redondante et fiable.

• **Bonne capacité de récupération de données :** En cas de défaillance d'un nœud, les données sont automatiquement répliquées sur d'autres nœuds.

• **Évolutivité horizontale :** Hadoop permet d'ajouter facilement de nouveaux nœuds au cluster pour augmenter la capacité de traitement.

• **Moindre cout :** Hadoop s'exécute sur du matériel standard et utilise des logiciels open source, ce qui réduit les couts par rapport aux solutions propriétaires.

### Architecture d’Hadoop

Hadoop est composé de plusieurs modules :

- Hadoop YARN : gestionnaire de ressources qui surveille les ressources disponibles dans le cluster.

- Hadoop MapReduce : moteur de traitement de données intégré à Hadoop, responsable de la gestion des fichiers et du traitement distribué.

- Hadoop Common : ensemble de fonctionnalités pour l'administration et la planification du système.

- Hadoop Distributed File System (HDFS) : système de stockage distribué pour les données d'Hadoop.

### Outils basés sur Hadoop

- MapReduce : outil de mise en œuvre du paradigme de programmation parallèle du même nom.

- HBase : base de données distribuée avec stockage structuré pour les grandes tables.

- Hive : logiciel d'analyse de données avec une syntaxe proche du SQL, initialement développé par Facebook.

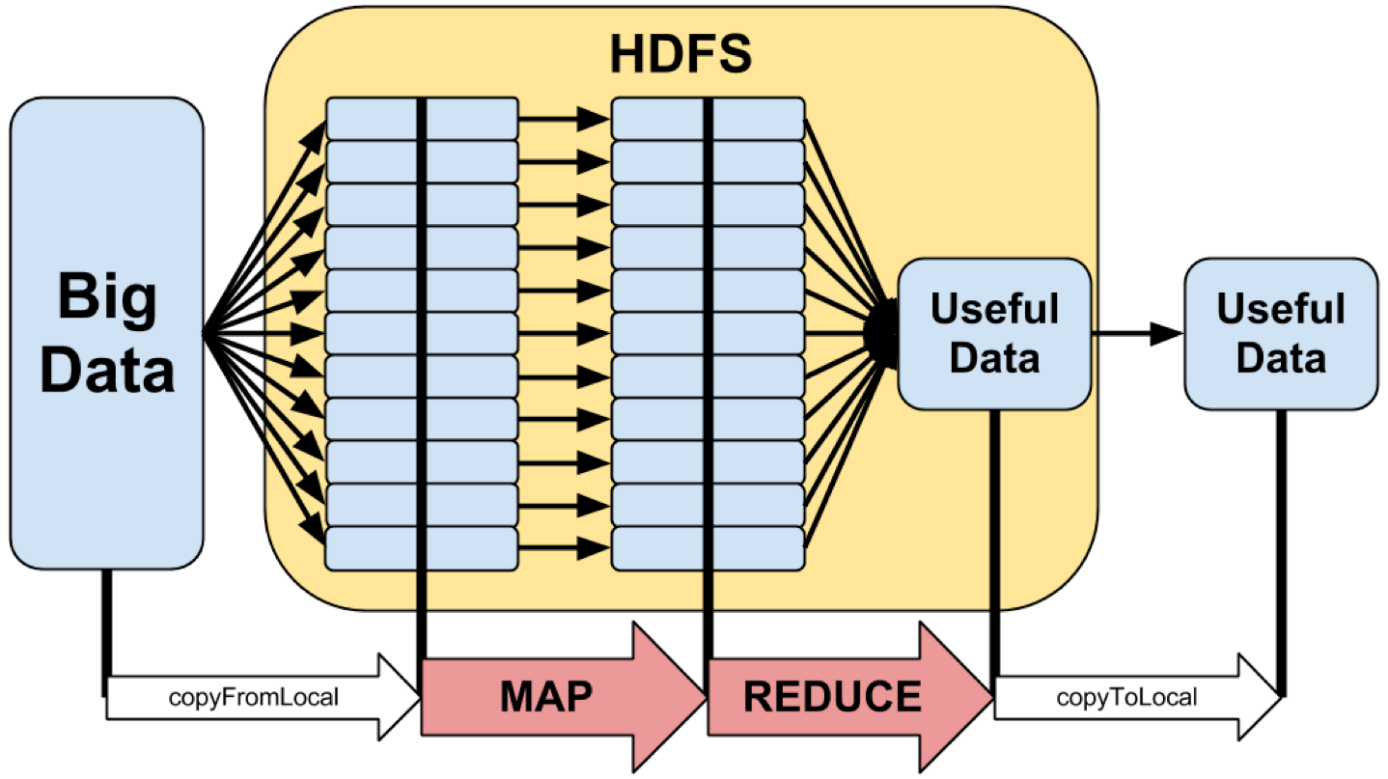
- Pig : logiciel d'analyse de données utilisant le langage Pig Latin, initialement développé par Yahoo!

- Spark : framework de traitement de données distribué avec mémoire partagée, compatible avec Hadoop.

## La solution Hadoop

### A

Le noyau d’Hadoop est constitué d’une partie de stockage : HDFS (Hadoop Distributed File System), et d’une partie de traitement appelée MapReduce



*Hadoop fractionne les fichiers en gros blocs et les distribue à travers les nœuds du cluster. Pour traiter les données, il transfère le code à chaque nœud et chaque nœud traite les données dont il dispose. Cela permet de traiter l'ensemble des données plus rapidement et plus efficacement que dans une architecture supercalculateur plus classique.*

Le framework Hadoop de base se compose des modules suivants :

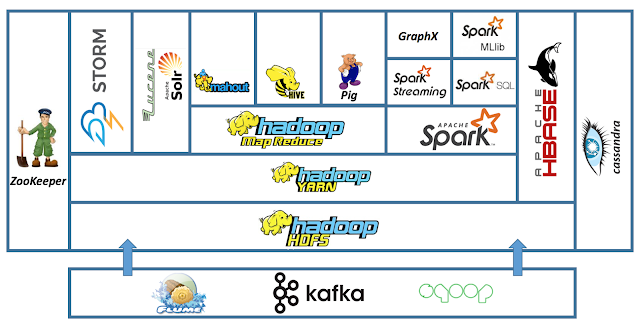
* Hadoop YARN : resources manager pour surveiller les ressources disponibles, machine séparée, on peut avoir plusieurs resources manager
* Hadoop MapReduce : programme MapReduce intégré à Hadoop qui gère les fichiers, shuffle and sort
* Hadoop Common : fonctionnalités d’Hadoop qui permettent le pilotage de l’ensemble (administration du système, scheduler)
* Hadoop Distributed File System (HDFS) : système de stockage distribué d’Hadoop

Ces 4 modules ensemble font l’architecture de base d’un système Hadoop. Chaque élément est remplaçable. Il peut vouloir un autre resources manager que YARN, utiliser NoSQL et pas HDFS, utiliser SPARK à la place de MapReduce

Spark : outil compatible avec Hadoop, envoyer du code dans Hadoop pour qu’il soit interpréter directement en MapReduce

### De nombreux outils basés sur Hadoop

* MapReduce : outil de mise en œuvre du paradigme de programmation parallèle du même nom
* HBase : base de données distribuée disposant d’un stockage structuré pour les grandes tables
* Hive : Logiciel d’analyse de données (initialement développé par Facebook) permettant d’utiliser Hadoop avec une syntaxe proche du SQL
* Pig : logiciel d’analyse de données (initialement développé par Yahoo!) comparable à Hive mais utilisant le langage Pig Latin
* Spark : framework de traitement de données distribué avec mémoire partagée



<https://blogs.perficient.com/2022/08/10/hadoop-ecosystem-components/>

## Lignes de commande : commandes HDFS

•

* hadoop fs -mkdir -p nom\_dossier
* hadoop fs -put purchases.txt input

commandes HDFS

* **lister le contenu à partir d’où on est  :** hadoop fs -ls
* **lister le contenu d’un répertoire :** hadoop fs -ls répertoire
* **prendre un fichier et le mettre dans Hadoop pour qu’il le transforme en shards :** hadoop fs -put /chemin/jusqu’à/ma/cible
* **récupérer un fichier d’un système Hadoop HDFS :** hadoop fs -get / chemin/jusqu’à/ma/cible
* **regarder le début d’un fichier :** hadoop fs -cat mon\_fichier | head
* **regarder la fin d’un fichier :** hadoop fs -tail mon\_fichier
* hadoop fs -mv old\_name.txt new\_name.txt
* hadoop fs -rm mon fichier

docker cp .\WordCount.jar hadoop-master:/root/WordCount.jar (copier un fichier d’un endroit à un autre : docker cp source destination)

hadoop jar WordCount.jar WordCount input output : demander à Hadoop de lire le fichier jar avec en entrée les info du dossier input à mettre dans le dossier output

*Hadoop jar fichier.jar programme à lancer source destination*

Spark

spark-shell

docker stats = pour voir la puissance utilisée par chacun des conteneurs

## A

A

•

—

## A

A

•

—