Big Data

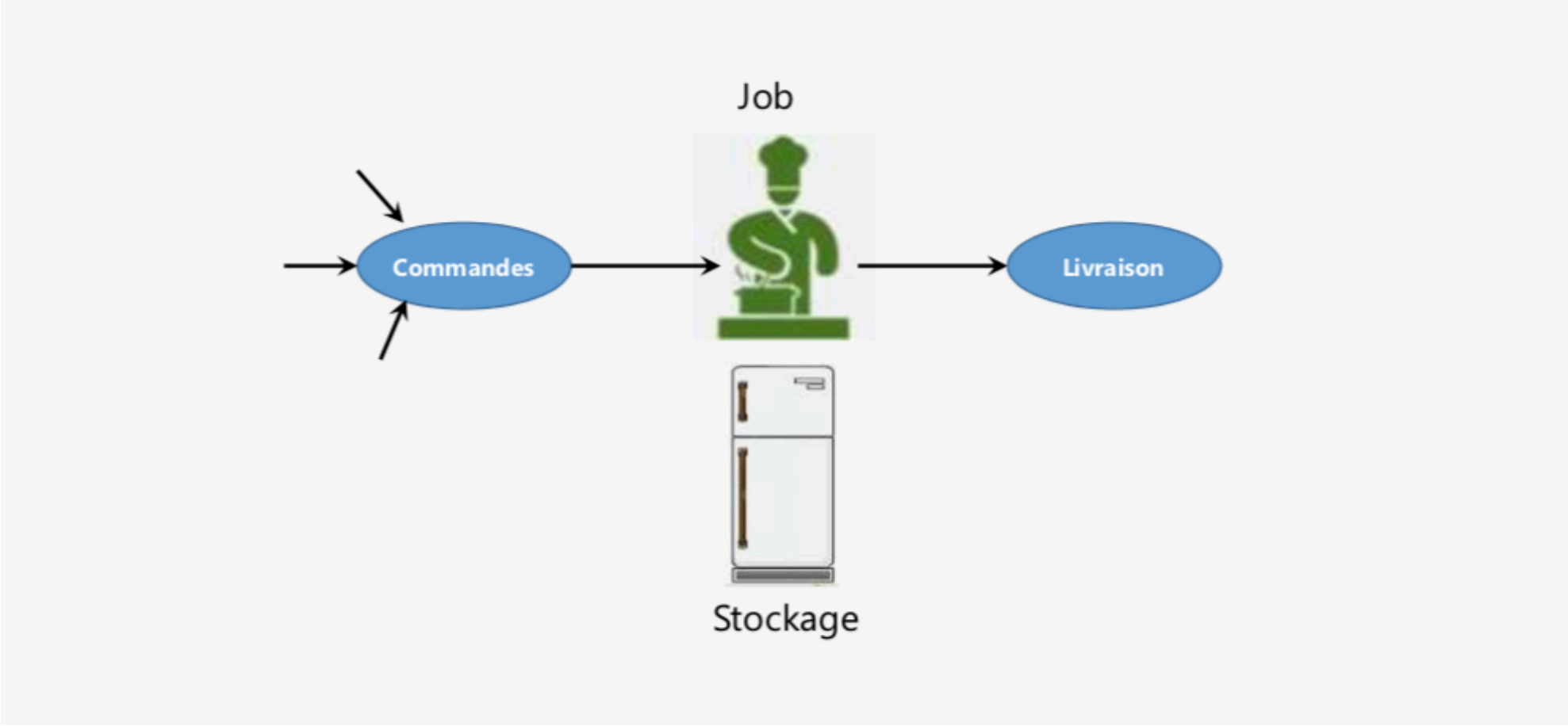
Chapitre 2: Hadoop

- Présentation du framework
- HDFS
- Map Reduce

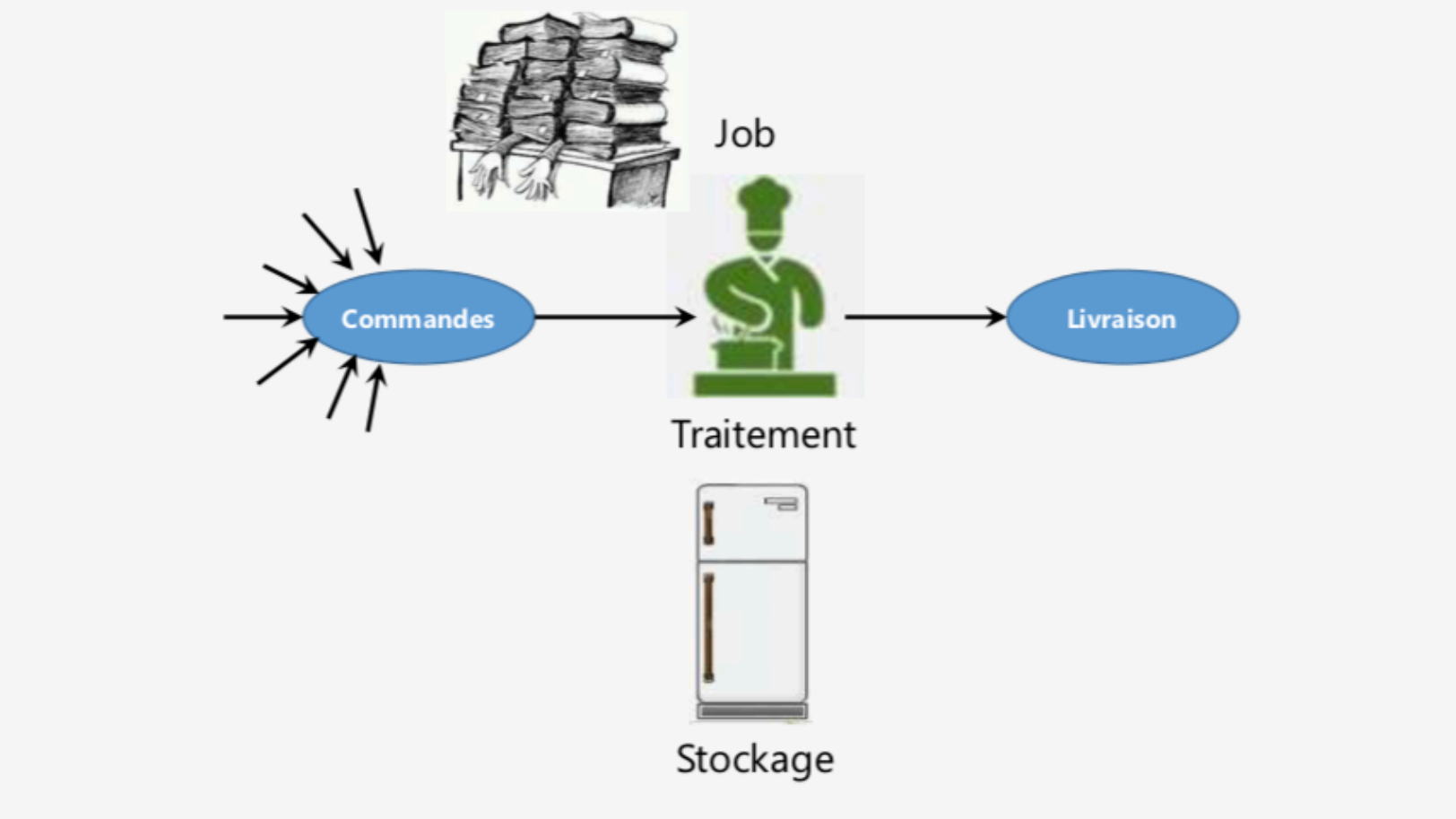
Présentation du framework

- Le projet Hadoop consiste en deux grandes parties:
  - Stockage des données : HDFS (Hadoop Distributed File System)
  - Traitement des données : MapReduce / Yarn
- 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 à votre cluster, au fur et à mesure que les données augmentent

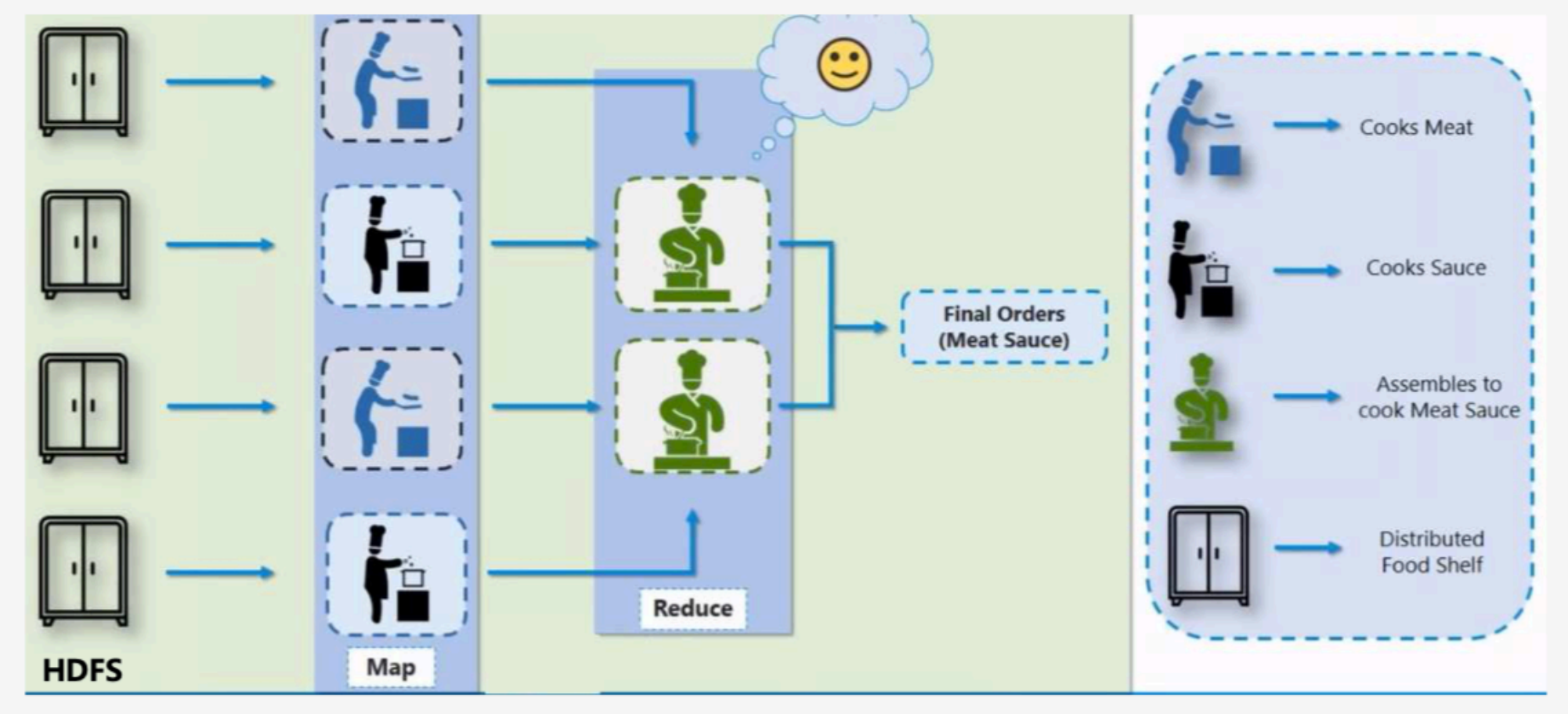
Présentation du framework



Présentation du framework

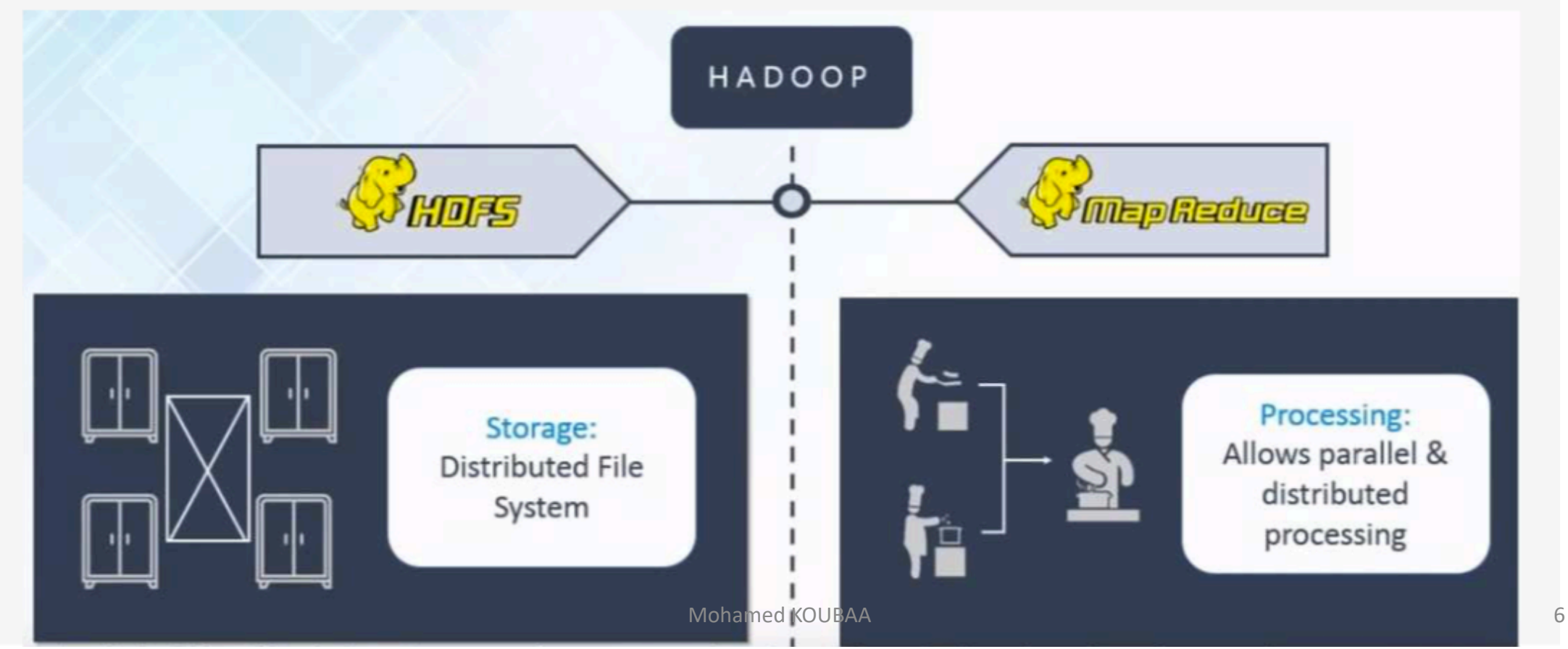


Présentation du framework



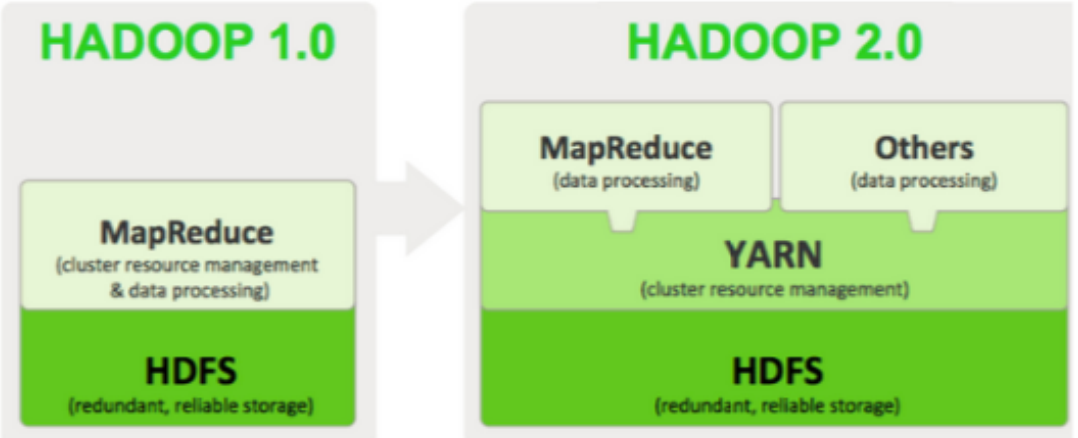
Eco système Hadoop

Hadoop est Framework qui permet de **Stocker** et **Traiter** une grande quantité de données de manière **parallèle** et **distribuée**



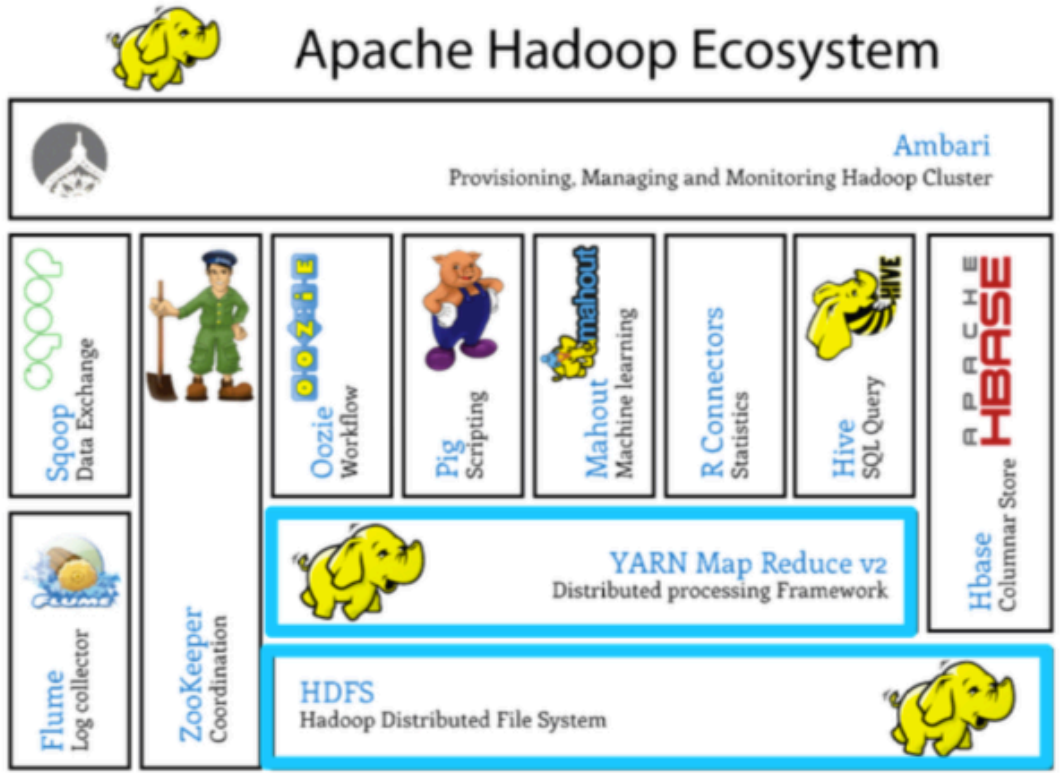
Eco système Hadoop

- Hadoop est un Framework libre et open source écrit en Java, destiné à faciliter la création d'applications massivement distribuées avec des milliers de nœuds.
  - Au niveau du stockage distribué des données (des pétaoctets de données) : HDFS
  - Au niveau Traitement de données : Map Reduce
  - Avec la prise en charge de caractéristiques fondamentales :
    - Haute disponibilité, Scalabilité, Tolérance aux panes, Reprise après échec, Sécurité (HPC : High Performance Computing)
- Le Framework Hadoop de base se compose des modules suivants :
  - **Hadoop Distributed File System (HDFS)** : le système de fichiers distribué
  - **Hadoop YARN** (Yet Another Resource Negotiator): Système de Gestion des Ressources du Cluster
  - **Hadoop MapReduce** : Traitement distribué de données



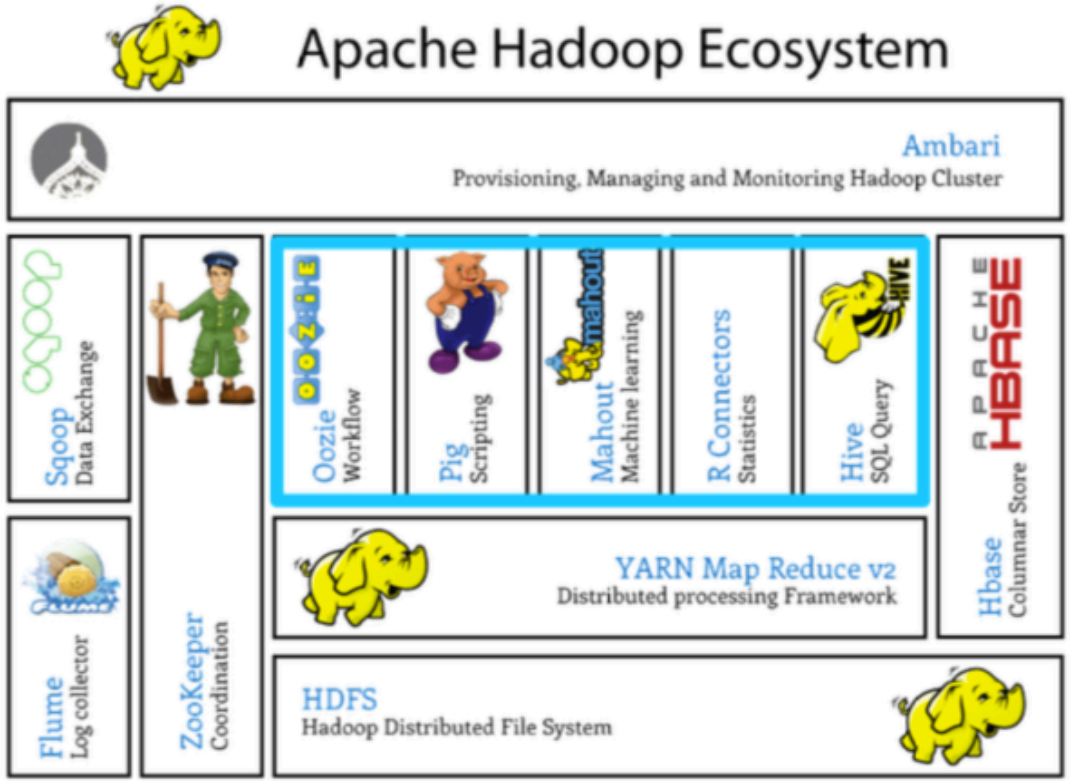
Eco système Hadoop

- En plus des briques de base Yarn Map Reduce/HDFS, plusieurs outils existent pour permettre:
  - L'extraction et le stockage des données de/sur HDFS
  - La simplification des opérations de traitement sur ces données
  - La gestion et coordination de la plateforme
  - Le monitoring du cluster



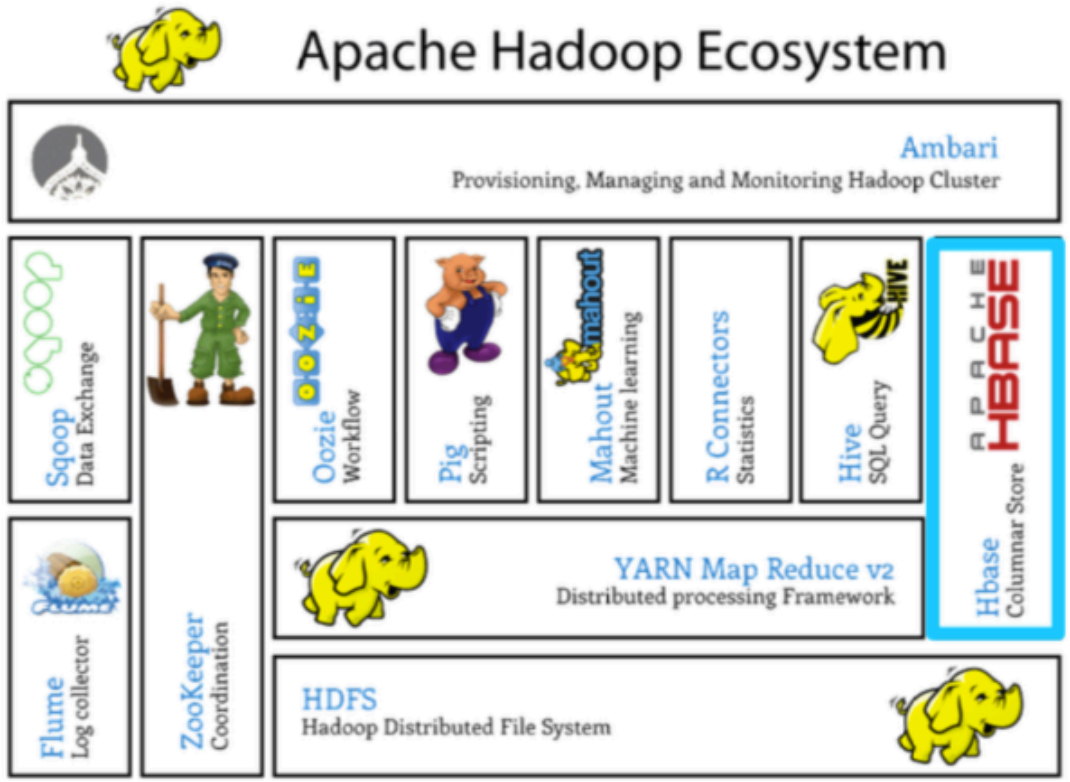
Eco système Hadoop

- Parmi ces outils, certains se trouvent au dessus de la couche Yarn/MR, tel que:
  - **Pig**: Langage de script
  - **Hive**: Langage proche de SQL (Hive QL)
  - **R Connectors**: permet l'accès à HDFS et l'exécution de requêtes Map/Reduce à partir du langage R
  - **Mahout**: bibliothèque de machine learning et mathématiques
  - **Oozie**: permet d'ordonnancer les jobs Map Reduce, en définissant des workflows



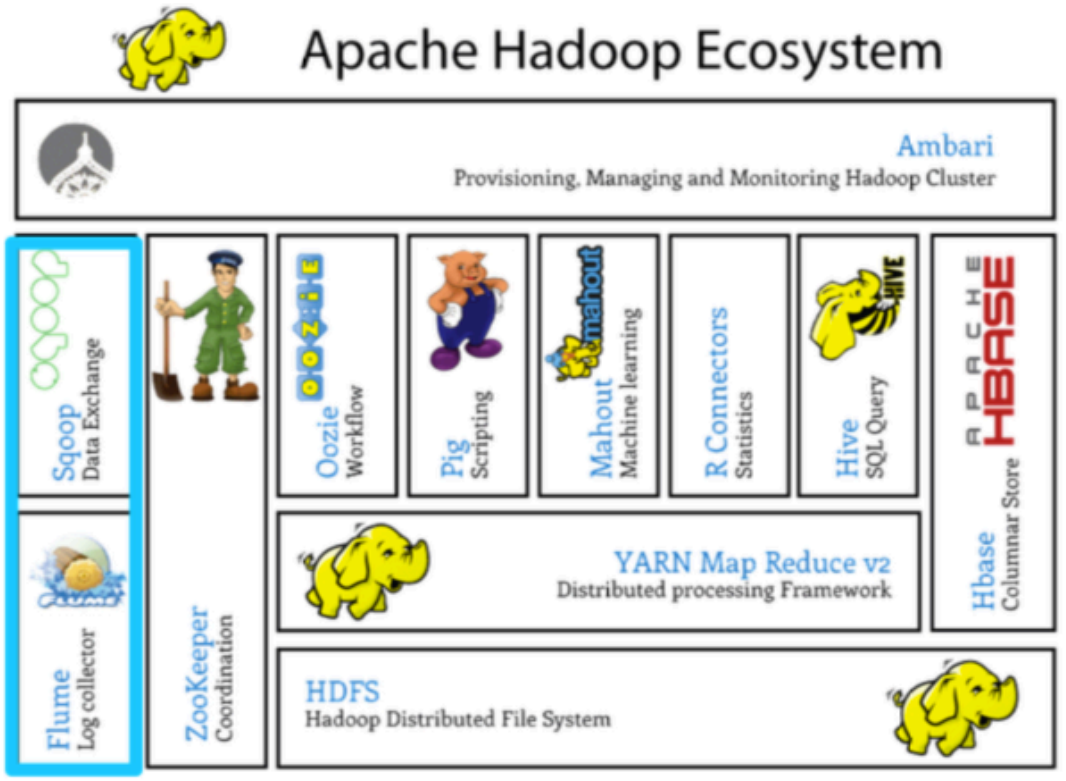
Eco système Hadoop

- D’autres outils sont directement au dessus de HDFS, tel que :
  - **Hbase** : Base de données NoSQL orientée colonnes
  - **Impala** (pas représenté dans la figure): Permet le requêtage de données directement à partir de HDFS (ou de Hbase) en utilisant des requêtes Hive SQL



Eco système Hadoop

- Certains outils permettent de connecter HDFS aux sources externes, tel que:
  - **Sqoop**: Lecture et écriture des données à partir de bases de données externes
  - **Flume**: Collecte de logs et stockage dans HDFS



Eco système Hadoop

- Enfin, d'autres outils permettent la gestion et administration de Hadoop, tel que:
  - **Ambari**: outil pour le provisionnement, gestion et monitoring des clusters
  - **Zookeeper**: fournit un service centralisé pour maintenir les information de configuration, de nommage et de synchronisation distribuée