Chapitre III : L'écosystème Hadoop

- Hadoop VS Ecosystème Hadoop
- Ecosystème Hadoop
- Hadoop version 2
- Conclusion

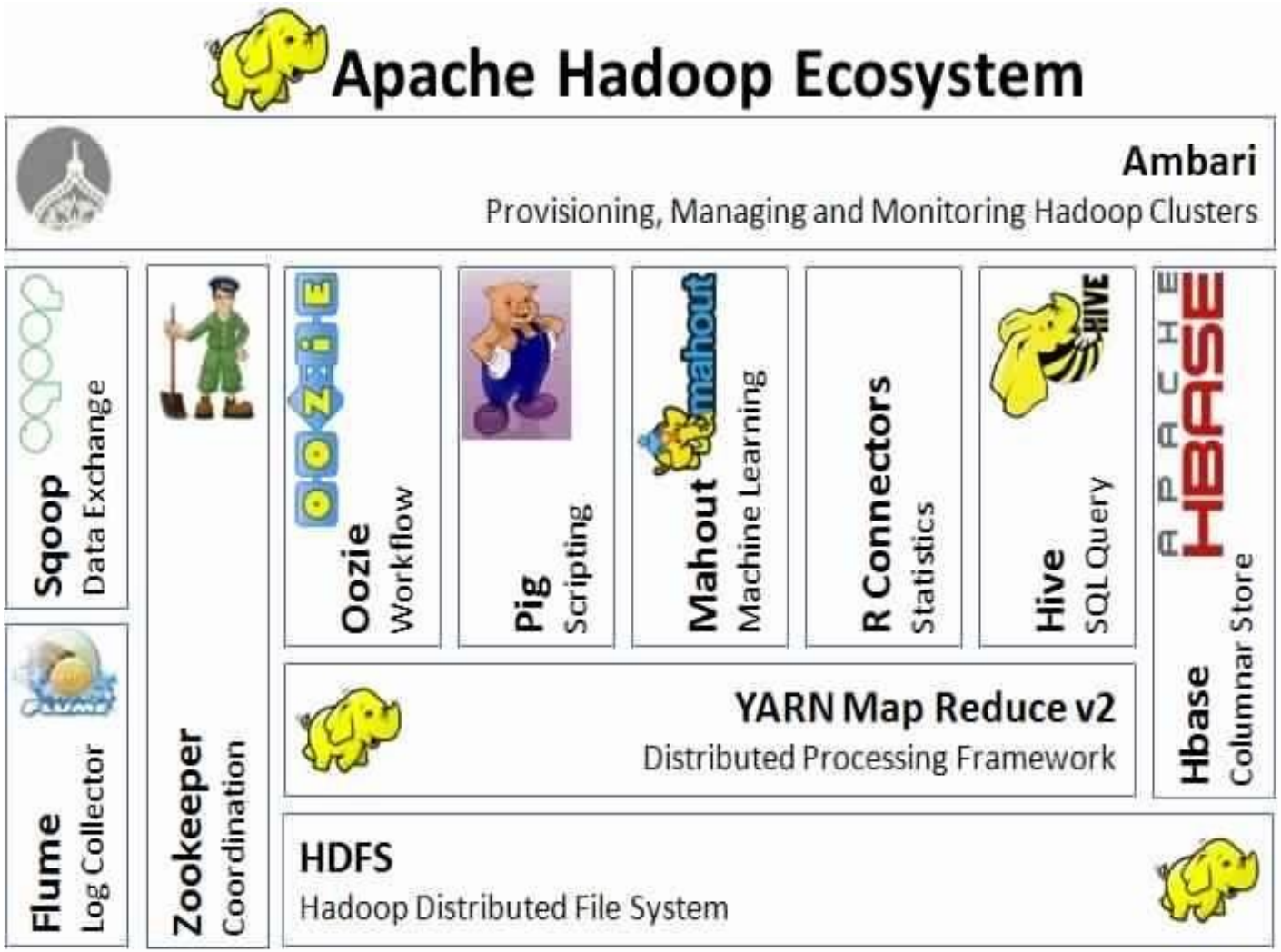
Hadoop VS Ecosystème Hadoop

- HADOOP est une plateforme open source permettant le stockage et le traitement des volumes de données massives d'une façon distribuée.
- Un écosystème HADOOP est l'ensemble des composants et outils qui forme un cluster HADOOP (HDFS, MapReduce, Apache Pig, Apache Hive, Hbase, etc.).

Ecosystème 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 traitements sur ces données
  - La gestion et la coordination de la plateforme
  - Le monitoring du cluster

Ecosystème Hadoop



Ecosystème Hadoop

D'autres outils se situent directement au dessus de HDFS, tels que :

- **Hbase** : Base de données NoSQL orientée colonnes
- **Impala** : permet le requêtage de données directement à partir de HDFS (ou de Hbase) en utilisant des requêtes Hive SQL

D'autres outils permettent la gestion et administration de Hadoop, tels que :

- **Ambari** : outil pour le provisionnement, la gestion et le monitoring des clusters.
- **Zookeeper** : fournit un service centralisé pour maintenir les informations de configuration, de nommage et de synchronisation distribuée.

Ecosystème Hadoop

Certains outils se trouvent au dessus de la couche Yarn/MR, tel que:

- **Pig** : un langage de haut niveau dédié à l'analyse de gros volumes de données. Il s'adresse aux développeurs habitués à faire des scripts via Bash ou Python par exemple.
- **Hive** : un système d'entrepôt de données (Data Warehouse) pour Hadoop qui offre un langage de requête proche de SQL pour faciliter les agrégations, le requêtage ad-hoc et l'analyse de gros volumes de données stockés dans des systèmes de fichiers compatibles Hadoop.

Ecosystème Hadoop

- **Oozie :** est un système de flux de travail (workflow) dont l'objectif est de simplifier la coordination et la séquence de différents traitements (programme Map/Reduce, scripts Pig, etc.)
- **Mahout** : un système d'apprentissage automatique et d'analyse de données. Il implémente des algorithmes de classification et de regroupement automatique (Machine Learning, DataMining)
- **R Connectors**: permet l'accès à HDFS et l'exécution de requêtes Map/Reduce à partir du langage R

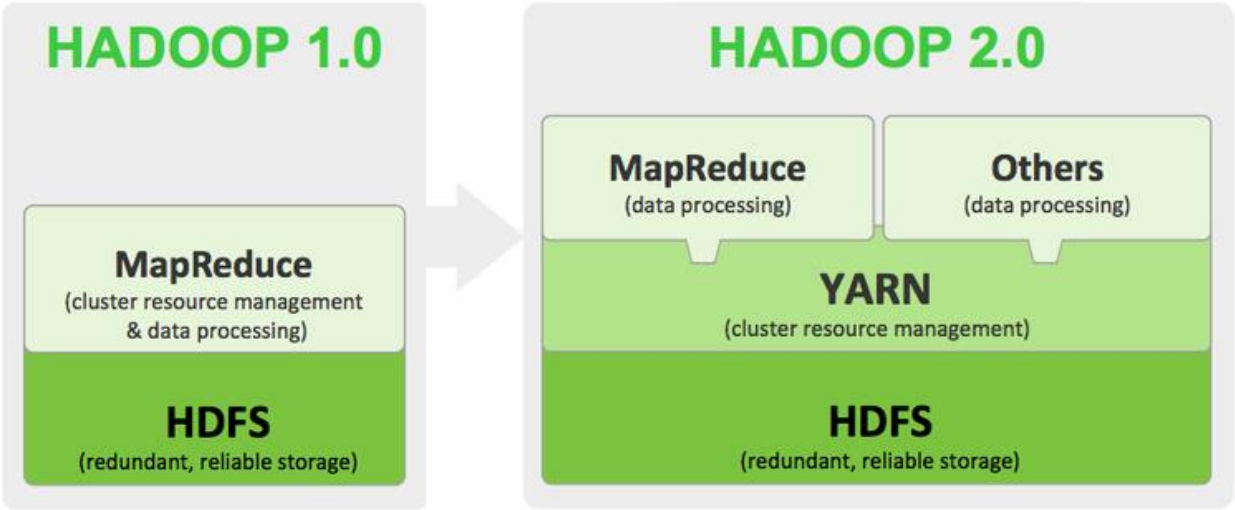
Ecosystème Hadoop

Certains outils permettent de connecter HDFS aux sources externes, tel que:

- **Sqoop** : un outil conçu pour **transférer** efficacement une masse de données entre Apache Hadoop et un stockage de données **structuré** tel que les bases de données **relationnelles.**
- **Flume :** un service distribué, fiable et disponible pour collecter efficacement, agréger et déplacer une grande quantité de logs.

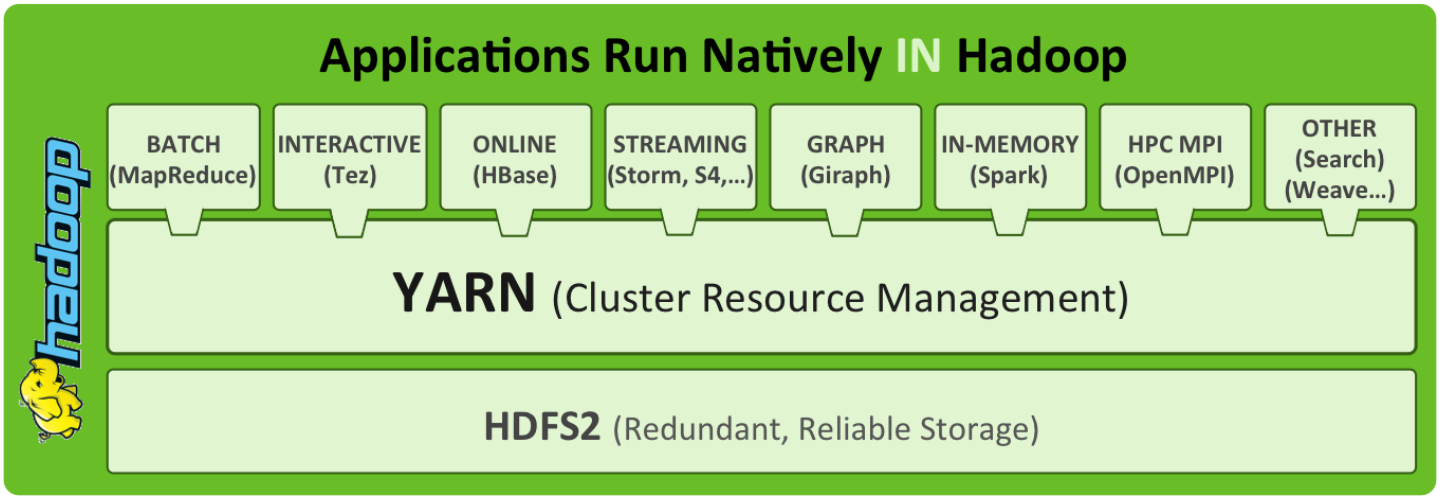
Hadoop version 2

- Découpler Hadoop de MapReduce
- Permettre à des frameworks alternatifs d'être portés directement sur Hadoop et HDFS.
- Une meilleure scalabilité, s'enrichir de nouveaux frameworks couvrant des besoins peu ou pas couverts avec MapReduce.



Hadoop version 2

- Client et cluster peuvent utiliser des versions différentes.
- Des protocoles de communication standardisés et documentés.
- Évolution du framework progressive avec rétro-compatibilité sans destruction des services.



Conclusion

- Hadoop n'utilise aucun principe foncièrement nouveau; il offre en revanche une très forte simplicité et souplesse de déploiement inconnues jusqu'à présent pour l'exécution facile de tâches parallèles variées.
- Grâce à Hadoop, même des structures limitées en taille/ressources peuvent facilement avoir accès à de fortes capacités de calcul: déploiement à bas coût de clusters en interne ou location de temps d'exécution via les services de cloud computing.