

Hadoop Ecosystem

Composants de base de Hadoop



Core components of Hadoop:
Hadoop common, HDFS,
MapReduce, and YARN

- Hadoop Core (le noyau): utilitaires et bibliothèques de support.
 - HDFS : stockage distribué.
 - MapReduce : traitement parallélisé du Big Data.
 - YARN : gestionnaire de ressources.

Composants étendus de l'écosystème Hadoop

- Packages supplémentaires installés au-dessus du noyau Hadoop.
 - Ingestion, stockage, analyse et accès aux données.
 - Les composants travaillent ensemble pour traiter le Big Data.

Ingest Data
Examples: Flume & Sqoop

Store Data
Examples: HDFS & HBase

Process and Analyze Data
Examples: Pig & Hive

Access Data
Examples: Impala & Hue

- ingestion de données
- chargement des données
- importation des données

Ingestion des données

- Flume
 - collecte, agrégation et transfert de données.
 - Architecture flexible basée sur des flux continus.
- Sqoop
 - transfert en masse entre SGBDR et Hadoop.
 - Sqoop génère automatiquement du code MapReduce.

Stockage dans Hadoop

- HBase
 - Base NoSQL orientée colonnes.
 - Accès aléatoire rapide basé sur des tables de hachage.
- Cassandra
 - base NoSQL distribuée sans point unique de défaillance.

Analyse et traitement des données

- Pig
 - langage procédural pour analyser de grands volumes.
 - Suit une suite de commandes pour transformer les données.
- Hive
 - langage déclaratif pour créer des rapports.
 - Fonctionne côté serveur dans un cluster.

Accès aux données affinées

- Impala
 - accès rapide aux données pour utilisateurs non techniques.
 - Permet des requêtes interactives.
- Hue
 - interface utilisateur Hadoop.
 - Supporte Pig, SQL, flux de travail, navigation des fichiers.



Hadoop Ecosystem

OOZIE

(workflow)

oogop

(Data Collection)



(Data Collection)

HCatalog
(Table & schema management)

(Managing & monitoring)



Data access layer



Pig
(Scripting)



HIVE
(sql query)

Data intelligence



mahout
(Machine learning)



DRILL
(Interactive analysis)

Data serialization



Thrift
(cross-language services)
(JSON)

APACHE
H
BASE

HBase
(Columnar store)

MapReduce
(Data processing)



YARN
(cluster resource management)

HDFS
(Hadoop Distributed File System)

