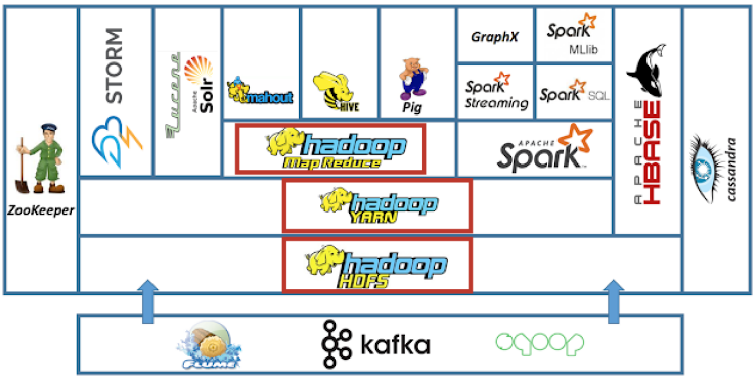
Écosystème Hadoop

Introduction à Hadoop

Qu'est-ce que Hadoop ?

- Un framework open-source pour stocker et traiter de grandes quantités de données sur des clusters d'ordinateurs.
- Fournit les outils suivants :
  - **HDFS** : système de fichiers distribué.
  - **YARN** : gestion des ressources du cluster.
  - **MapReduce** : modèle de programmation pour le traitement parallèle.



HDFS : Système de fichiers distribué Hadoop

Caractéristiques principales

- Un système répliqué pour tolérance aux pannes.
- Division des fichiers en blocs de 64 ou 128 Mo distribués sur les nœuds.
- Gestion optimisée pour minimiser les transferts de données.

Architecture

- **NameNode** : gère les métadonnées (structure des fichiers et localisation des blocs).
- **DataNode** : stocke les blocs de données et gère leur réplication.
- **Secondary NameNode** : sauvegarde des métadonnées pour tolérance aux pannes.

Mécanismes

- **Lecture** : Le client interroge le NameNode pour localiser les blocs, puis accède directement aux DataNodes.
- **Écriture** : Les blocs sont à la fois écrits et répliqués sur différents DataNodes.

YARN : Gestionnaire de ressources

Définition

- Acronyme de Yet Another Resource Negotiator .
- Gestion des ressources pour les applications distribuées.
- Permet l'exécution d'applications variées au-delà de MapReduce.

Composants principaux

- **Resource Manager** : planifie et attribue les ressources au cluster.
- **Node Manager** : gère les conteneurs individuels des nœuds.
- **Application Master** : supervise l'exécution des applications.

Hugo Tondenier

Avantages de YARN

- **Évolutivité** : Permet de traiter des charges de travail plus importantes.
- **Haute disponibilité** : Gestion des redondances.
- **Polyvalence** : Compatibilité avec différents types d'applications.

MapReduce : Modèle de programmation

Principes

- **Map** : Transforme les données en paires (clef, valeur).
- **Shuffle et sort** : Trie les paires par clef et les regroupe.
- **Reduce** : Effectue des calculs sur les paires regroupées.

Flux de traitement

1. Diviser les données en blocs et assigner les tâches aux nœuds Map.
2. Les sorties des nœuds Map sont triées et transférées aux nœuds Reduce.
3. Les nœuds Reduce traitent les données et produisent le résultat final.