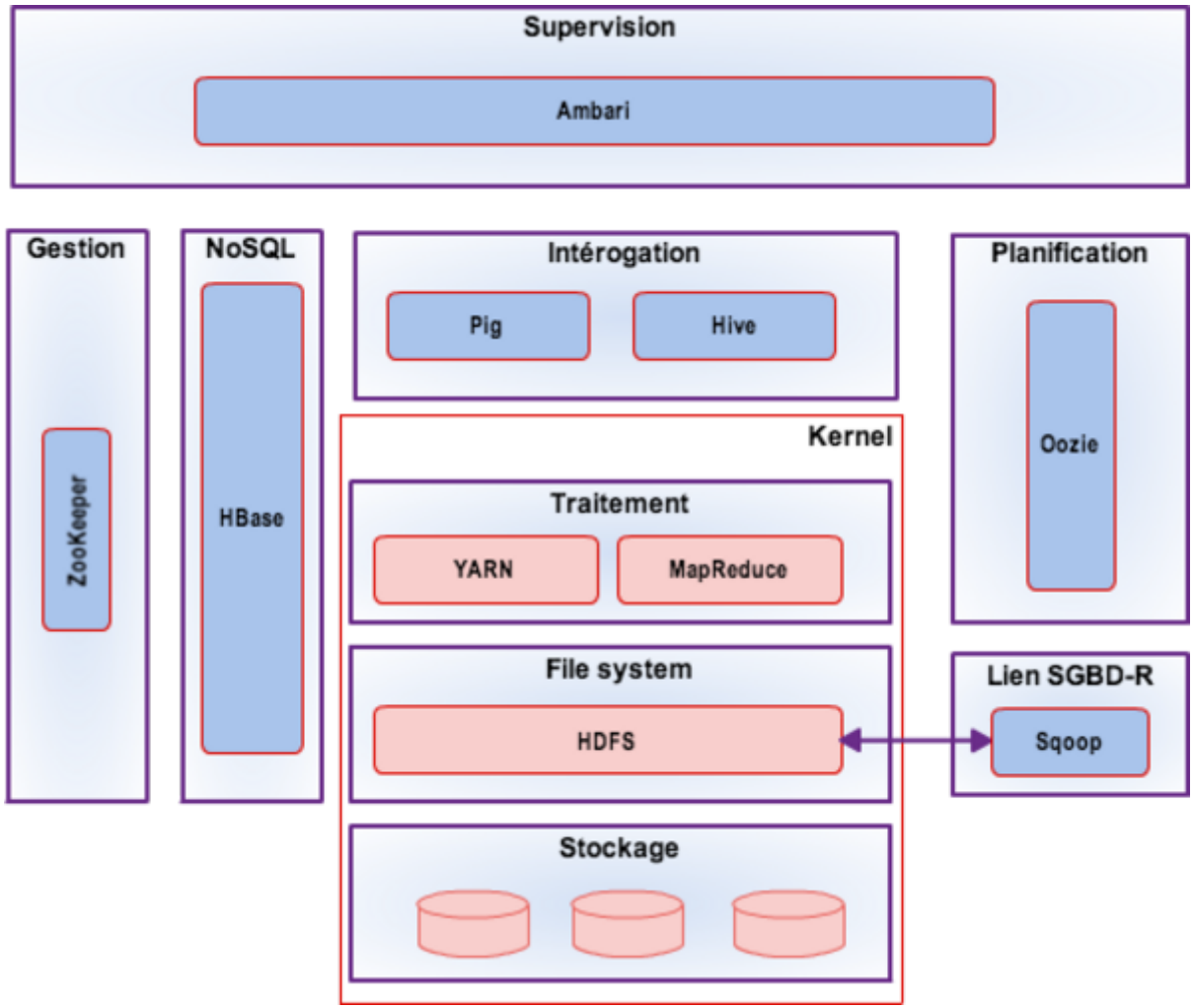
**Table des matières**

***Définitions*** ..... 3
***Généralités*** ..... 4
***Composants*** ..... 4
Hive & HBase ..... 4
Flume ..... 5
Hadoop ..... 5
Pig ..... 5

### Définitions

- Hadoop HDFS : système de fichiers scalable et distribué ;
- Hadoop Mapreduce : framework logiciel de traitement des données ;
- YARN : permet la gestion de l'état du cluster et des ressources et la gestion de l'exécution des jobs ;
- HBase : base de données d'Hadoop NoSQL, scalable et distribuée. HBase est une base de données distribuée disposant d'un stockage structuré pour les grandes tables. Comme BigTable, HBase est une base de données orientée colonnes ;
- Hive : logiciel d'analyse de données permettant d'utiliser Hadoop avec une syntaxe proche du SQL. Hive a été initialement développé par Facebook ;
- Flume : framework permettant d'intégrer des données à Hadoop ;
- Pig : logiciel d'analyse de données comparable à Hive, mais qui utilise le langage Pig Latin. Pig a été initialement développé par Yahoo ;
- Zookeeper : logiciel de gestion de configuration pour systèmes distribués, basé sur le logiciel Chubby développé par Google. ZooKeeper est utilisé entre autres pour l'implémentation de HBase.
- Mahout : implémentations d'algorithmes d'apprentissage automatique distribués sur Hadoop (machine learning) ;
- Sqoop : interface permettant de transférer des données entre les bases de données relationnelles et Hadoop ;
- Oozie : utilisée pour gérer et coordonner les tâches de traitement de données à destination de Hadoop ;
- Ambari : supervision et administration de clusters Hadoop

**Schéma de synthèse** [1)]




### Généralités

- Ecosystème Hadoop
- BigData : quelques mythes

### Composants

#### Hive & HBase

- HBase or Hive ? Hive vs. HBase
- Présentation Hadoop & HBase

**Flume**

- Introduction à Flume

**Hadoop**

Quelques remarques :

- Hadoop is not good to process transactions due to its lack random access ;
- It is not good when the work cannot be parallelized or when there are dependencies within the data, that is, record one must be processed before record two ;
- It is not good for low latency data access ;
- Not good for processing lots of small files although there is work being done in this area, for example, IBM's Adaptive MapReduce ;
- And it is not good for intensive calculations with little data.

- **Monter un cluster Hadoop**
- Dimensionner un cluster Hadoop
- Cluster Hadoop single-node
- Cluster Hadoop multi-node
- Démarrer un secondary node sur un node quelconque
- Quotas HDFS, fsck, etc.
- Simuler un crash d'un datanode

**Pig**

- Pig Tutorial part I

1)
http://blog.ippon.fr/2013/05/14/big-data-la-jungle-des-differentes-distributions-open-source-hadoop/