# Documentation technique

## Diagramme d’architecture





Extraction des données en format csv



Stockage



Structuration des données + Jointures



Nettoyage des données



Partitionnement   
+   
Requêtage des données



Extraction et affichage des données

## Description des flux

Les données une fois extraites des différentes bases de données sont stockées sur HDFS. Les données seront ensuite traitées par Yarn et Map Reduce mais elles resteront sur HDFS. Ensuite, les données seront organisées avec Hive (mais resteront stockées sur HDFS grâce à des tables externes). Nous créerons une table Hive où nous stockerons les différents indicateurs du SI pilotage. Depuis cette table, nous réaliserons des visualisations sur Tableau avec l’aide de Tableau Server.

## Organisation des données

Les données dans les tables externes Hive seront organisées comme ceci :



