03/11/2023

El MapReduce se utiliza para la paralelización de datos. SE utiliza 10 maps por nodo. Lo mas eficiente es hacer 128Mb por bloque de información. Importante, no se puede utilizer para tiempo real.

Formatos de fichero

* Textfile (predeterminado)
* Avro
* Parquet (csv)

Hive

Hacer grandes consultas en HDFS al estilo SQL. Utiliza el lenguaje HiveQL. Su objetivo es que Hadoop se comporte lo mas parecido a un Data Warehouse. Los datos son estructurados

* Dificultad de uso: API de Java complicada de usar
* Orientado a operaciones Batch