**Sqoop**

Se utiliza para hacer consumo de datos para realizar una ingesta de los datos.

* SQL to Hadoop
* Transfiere datos a hadoop u otros.
* Sqoop consume un conjunto de datos (sql u otros) para hacer una ingesta a hadoop.
* Uno puede especificar la distribución del map.
* Si no se especifica la llave primaria este la genera automáticamente.

**Pig**

* Es un flujo de datos.
* Cualquier tipo de información la puede transformar.
* Todo es una tabla de dos dimensiones.
* Cada instrucción genera un map reduce.

**HIVE**

* Es una infraestructura tipo warehouse.
* Aprovecha todos los beneficios de hadoop.a
* Permite hacer consultas ad-hoc, se pueden hacer con queries tipo sql.
* Es similar a las consultas sql.
* Se basa en un catalogo de metadatos para describir todas las consultas que se hacen.
* **Metastore(HCatalog):**
  + La estructura de datos: bd, tablas y columnas.
  + Mantiene las bases de datos relacionales
  + La estructura es independiente de la información.
  + Schema on read, los datos se evalúan de forma tardía hasta que se utilizar para consultar.
  + Funcionamiento:
    - Toma la consulta, la parsea y optimiza la consulta generando el plan de consulta y envía los Jobs para retornar lo solicitado.

**Impala**

* Impala ejecuta el map reduce de manera natural.
* Tiene una baja latencia.
* Es mejor que hive.
* Soporta un dialecto sql, es muy similar a hive.

-Impala tiene un mejor performace.

- Impala no tiene tolerancia a fallos que hive si lo tiene.

- Impala es para consultas más rápido.

- Impala usa un uso intensivo de la memoria, hive es mejor para tolerar los fallos.

- Impala es mas para usar en tiempo real por su velocidad.

- Hive es mas para procesamiento bash y Impala es mas para el procesamiento rápido.

**WorkBench** es una herramienta de Cloudera que es para poder usar lenguajes como R o Python para administrar proyectos sobre ciencia de datos.