Big Data con Hadoop y Spark

Módulo 01

Objetivos

En este módulo verás:

- Introducción a Sqoop
- Introducción a Hive
- Data Governance en Hadoop

Al final de la clase serás capaz de:

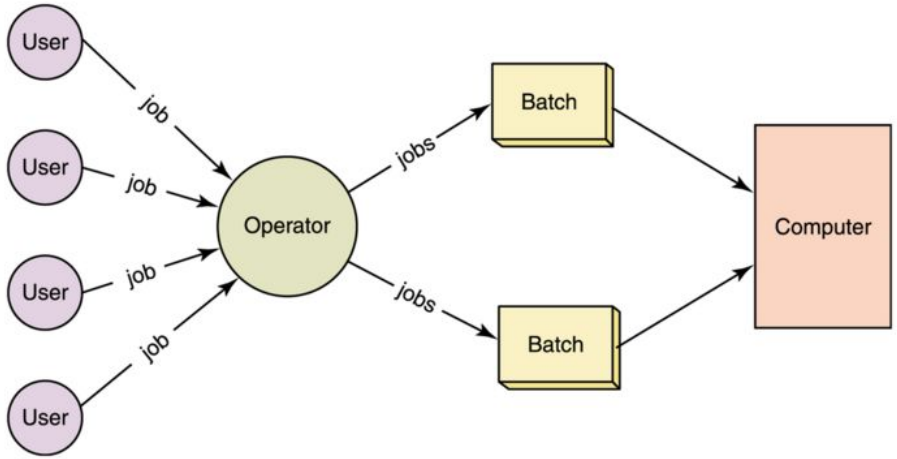
- Utilizar Hue como entorno de trabajo
- Identificar los casos de uso de Ranger y Atlas
- Procesar datos estructurados con Hive

Batch Processing

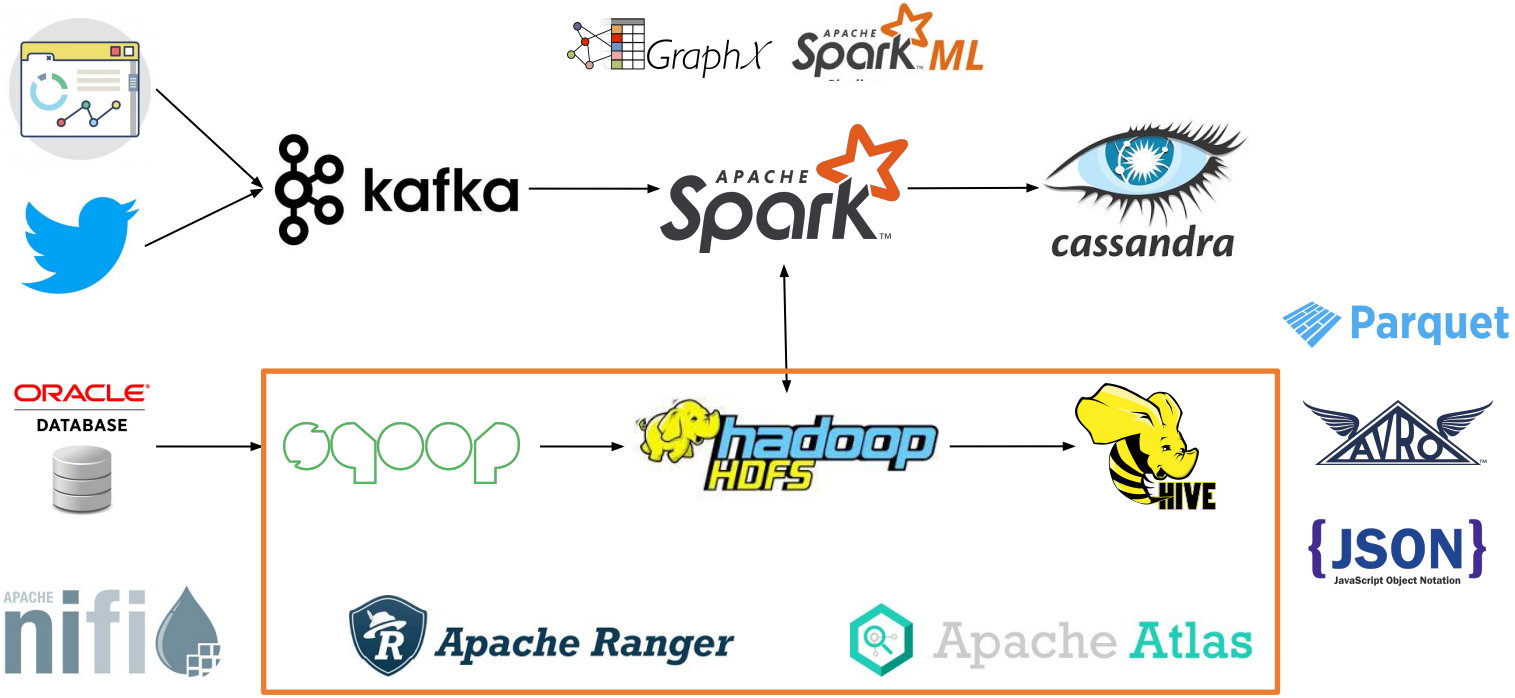
Es el procesamiento de transacciones por lote.

Los trabajos que pueden ejecutarse sin la interacción del usuario final, o pueden programarse para ejecutarse según lo permitan los recursos.

Ej. Reporte anual de ventas.



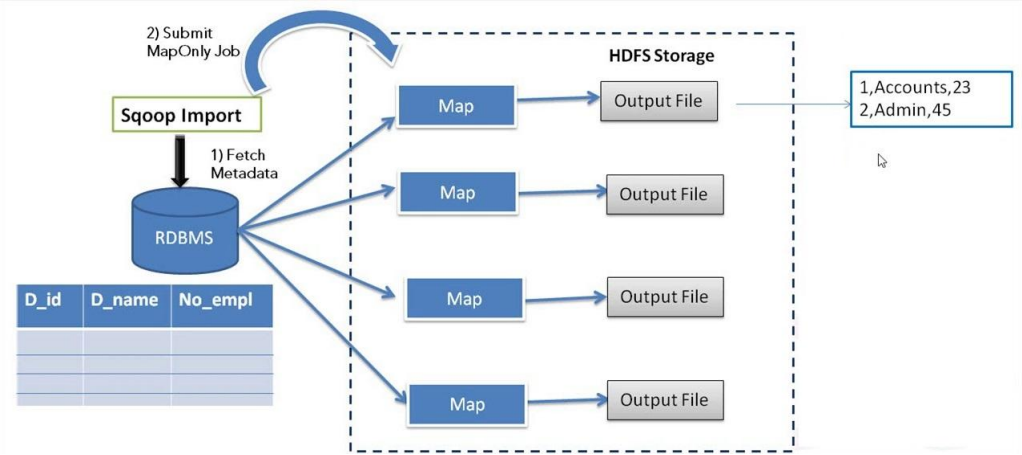
Frameworks


GraphX
APACHE Spark ML
kafka
APACHE Spark
cassandra
Parquet
ORACLE DATABASE
sqoop
hadoop HDFS
HIVE
AVRO
{JSON}
JavaScript Object Notation
APACHE nifi
Apache Ranger
Apache Atlas

Sqoop

Realiza transferencias de datos de tipo batch entre bases de datos relacionales y Hadoop

Utiliza JBDC para realizar la conexión con las bases de datos y MapReduce para ejecutar los jobs de manera distribuida



Ejemplo Sqoop

```
> sqoop import-all-tables \
    -m {{cluster_data.worker_node_hostname.length}} \
    --connect jdbc:mysql://{{cluster_data.manager_node_hostname}}:3306/retail_db \
    --username=retail_dba \
    --password=cloudera \
    --compression-codec=snappy \
    --as-parquetfile \
    --warehouse-dir=/user/hive/warehouse \
    --hive-import
```

Links de referencia

- **Sqoop** https://sqoop.apache.org/docs/1.4.7/SqoopUserGuide.html

Gracias