SQOOP

Training Session

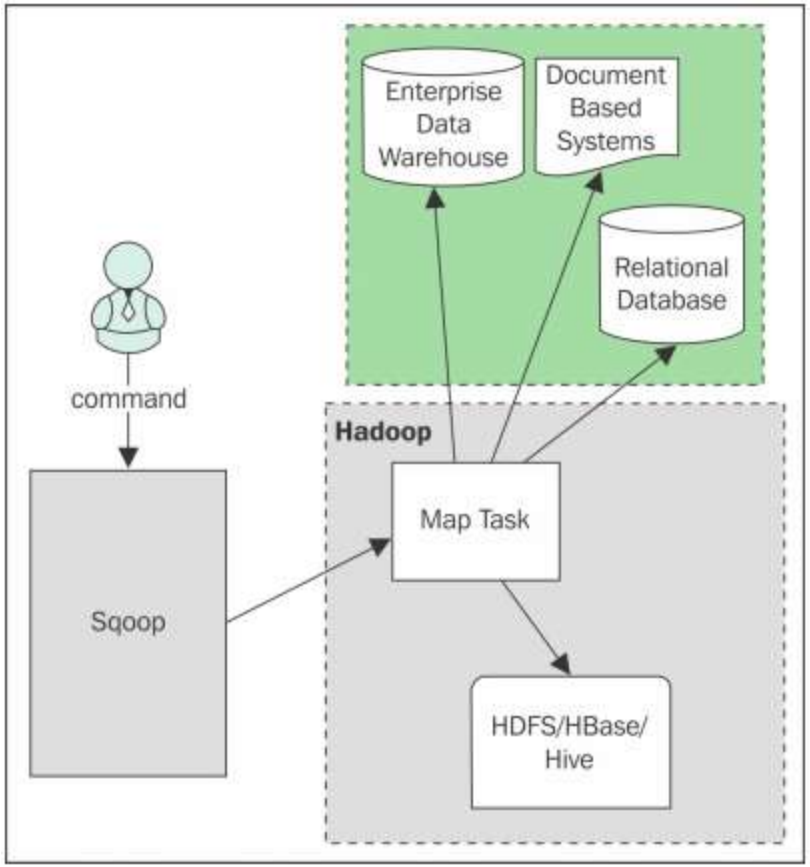
Overview

- Apache Sqoop is a tool designed for efficiently transferring bulk data between Apache Hadoop and structured datastores such as relational databases
- Sqoop can be used to import data from a RDBMS such as MySQL or Oracle into HDFS
- Sqoop automates most of this process, relying on the database to describe the schema for the data to be imported
- Manual process would involves tasks like
  - Export data from RDBMS in comma or tab seperated format
  - Load the files into HDFS
- Sqoop uses MapReduce to import and export the data, which provides parallel operation as well as fault tolerance

Overview

- SQOOP is data ingestion tool.
- SQOOP is a tool designed for transfer data between HDFS and RDBMS such as MySQL, Oracle etc.
- Export data back to RDBMS.
- Simple as user specifies the “what” and leave the “how” to underlying processing engine.
- No development, No Java is required.
- Developed by cloudera.

Sqoop Design



- Sqoop command initiated by the client fetches the metadata of the tables, columns, and data types, according to the connectors and drivers interfaces.
- The import or export is translated to a Map-only Job program to load the data in parallel between the databases and Hadoop.
- Clients should have the appropriate connector and driver for the execution of the process.

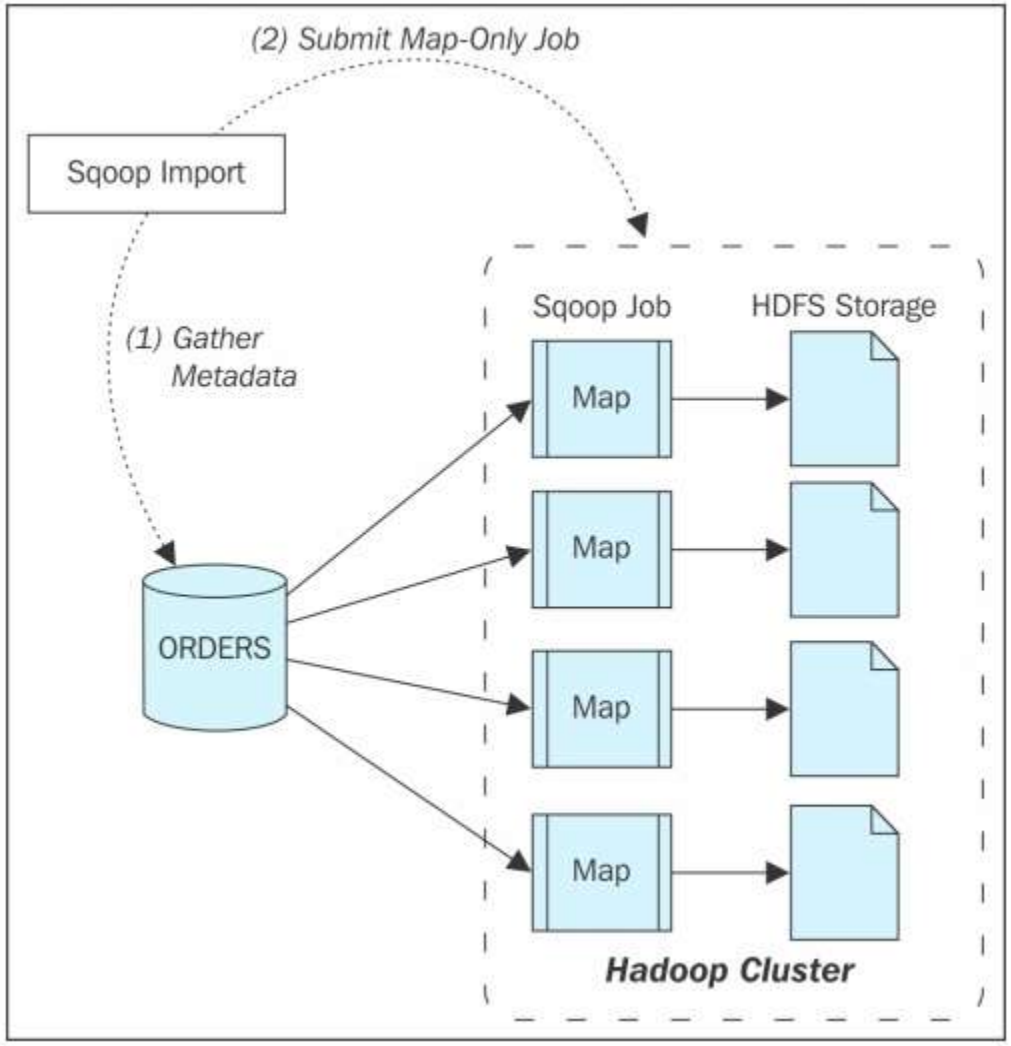
Sqoop Design

- Sqoop leverages map tasks for parallel import & export of relational databases.
- Fault tolerance as well as parallel processing is achieved this way
- Sqoop import → to HDFS, Hive & Hbase
- Data stored in any relational database with JDBC support can be directly imported into the Hive or HBase systems with Sqoop
- Sqoop export → from HDFS only

Sqoop Internals

Sqoop import is executed in two steps:

1. Gather metadata
2. Submit map only job



Sqoop Internals

- The dataset being transferred is sliced up into different partitions.
- A Map only Job is launched with individual mappers responsible for transferring a slice of dataset.
- Each record of the data is maintained in a type safe manner since SQOOP uses the database metadata to understand the data types.

Sqoop Command

sqoop list-databases --connect jdbc:mysql://localhost/ --username root –password root

sqoop: Hadoop executable

list-databases: Operation to perform

connect: JDBC URL of MySQL Server

--username: User name to use to connect to MySQL Server

--**password**: Password to use to connect to MySQL Server