Introduction to SQOOP

Agenda

- What is Sqoop
- Why Sqoop?
- How Sqoop Works
- Sqoop Architecture
- Sqoop Import
- Sqoop Export

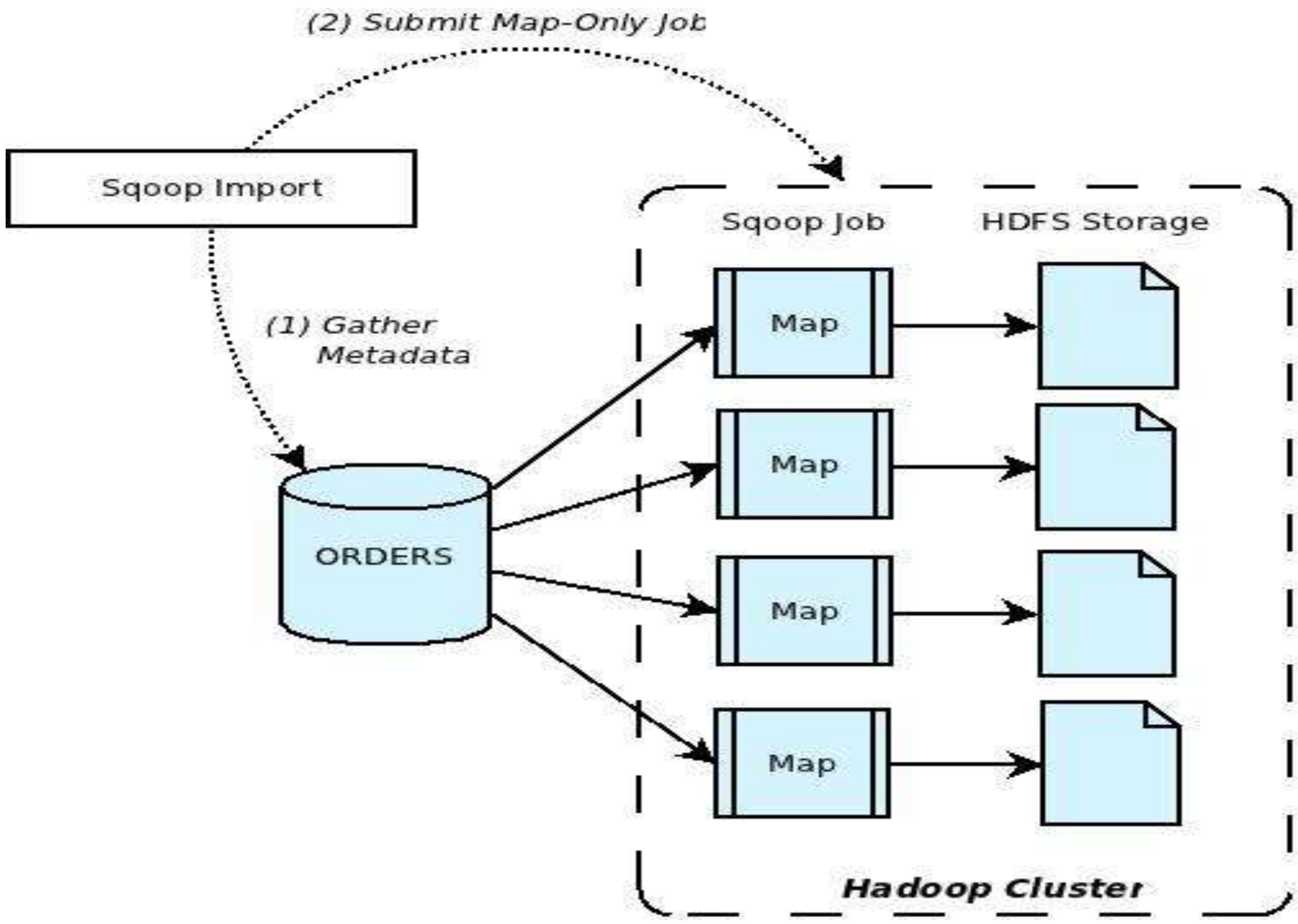
What is Sqoop

- Apache Sqoop is a tool designed for efficiently transferring bulk data between Apache Hadoop and structured datastores such as relational databases.
- Sqoop imports data from external structured datastores into HDFS or related systems like Hive and HBase.
- Sqoop can also be used to export data from Hadoop and export it to external structured datastores such as relational databases and enterprise data warehouses.

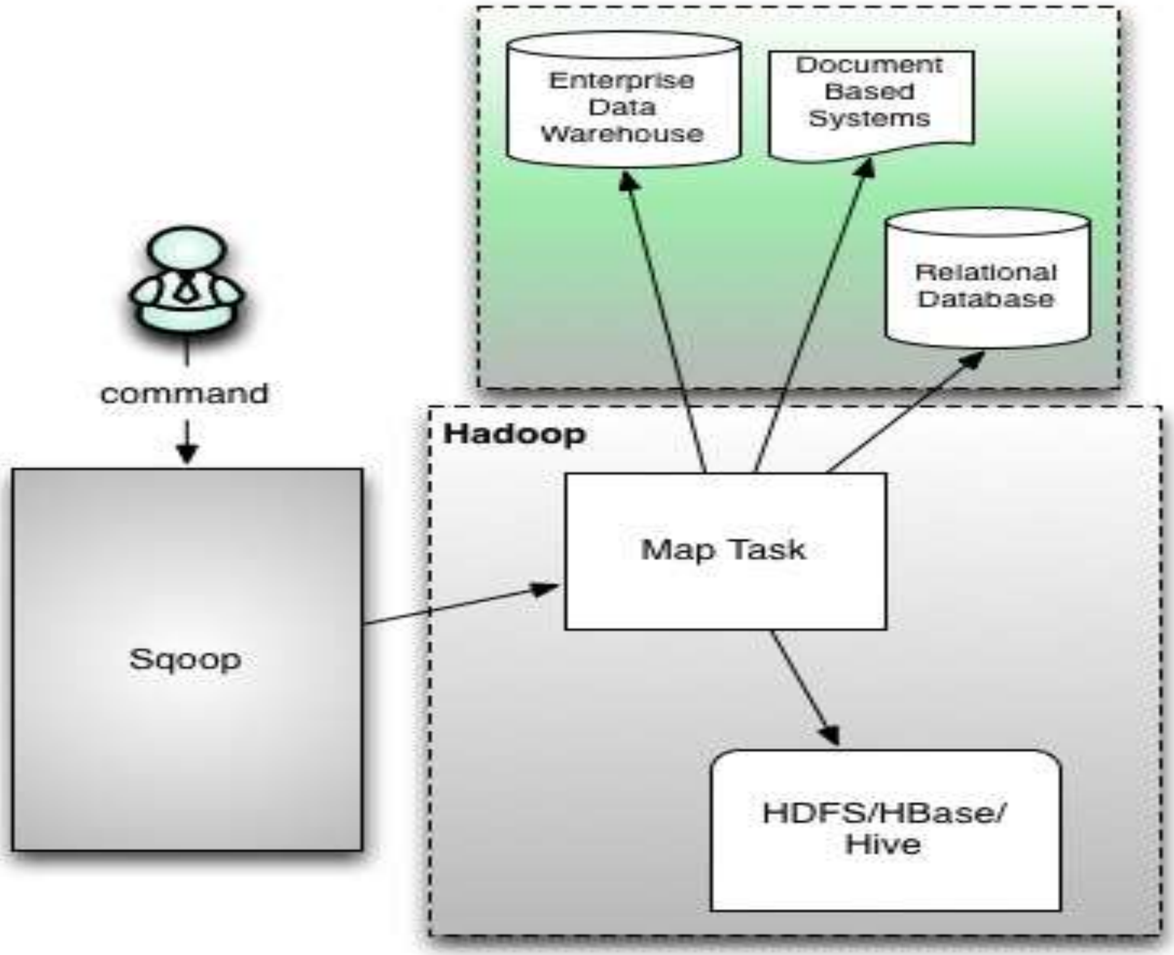
Why Sqoop?

- As more organizations deploy Hadoop to analyse vast streams of information, they may find they need to transfer large amount of data between Hadoop and their existing databases, data warehouses and other data sources
- Loading bulk data into Hadoop from production systems or accessing it from map-reduce applications running on a large cluster is a challenging task since transferring data using scripts is a inefficient and time-consuming task
- Allows data imports from external datastores and enterprise data warehouses into Hadoop
- Parallelizes data transfer for fast performance and optimal system utilization
- Copies data quickly from external systems to Hadoop
- Makes data analysis more efficient

How Sqoop Works



Sqoop Architecture



Sqoop Import

- sqoop import --connect jdbc:postgresql://hdp-master/sqoop_db --username sqoop_user --password postgres --table cities

Sqoop Export

```
sqoop export --connect jdbc:postgresql://hdp-master/sqoop_db --username sqoop_user --password postgres --table cities --export-dir cities
```