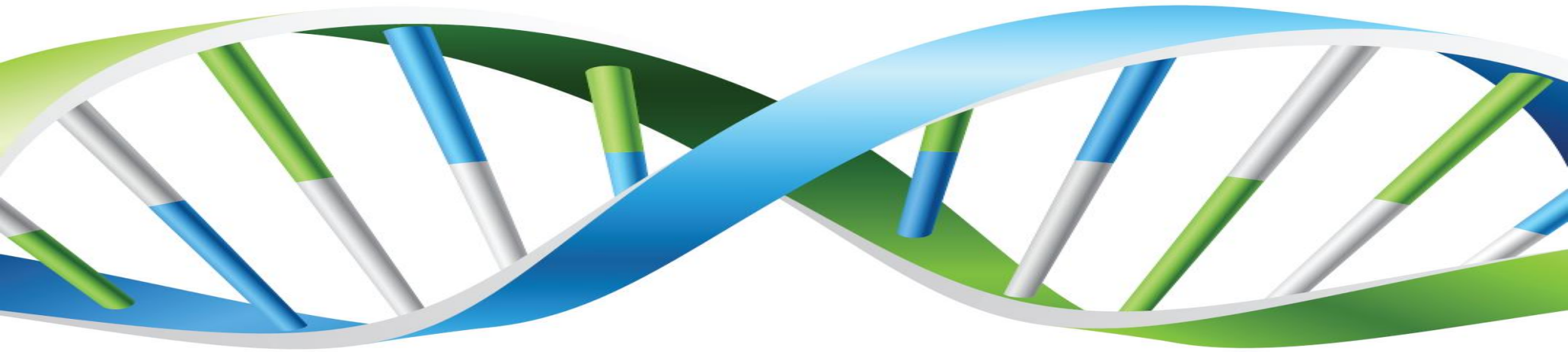
Big Data – Connectivity through SQOOP

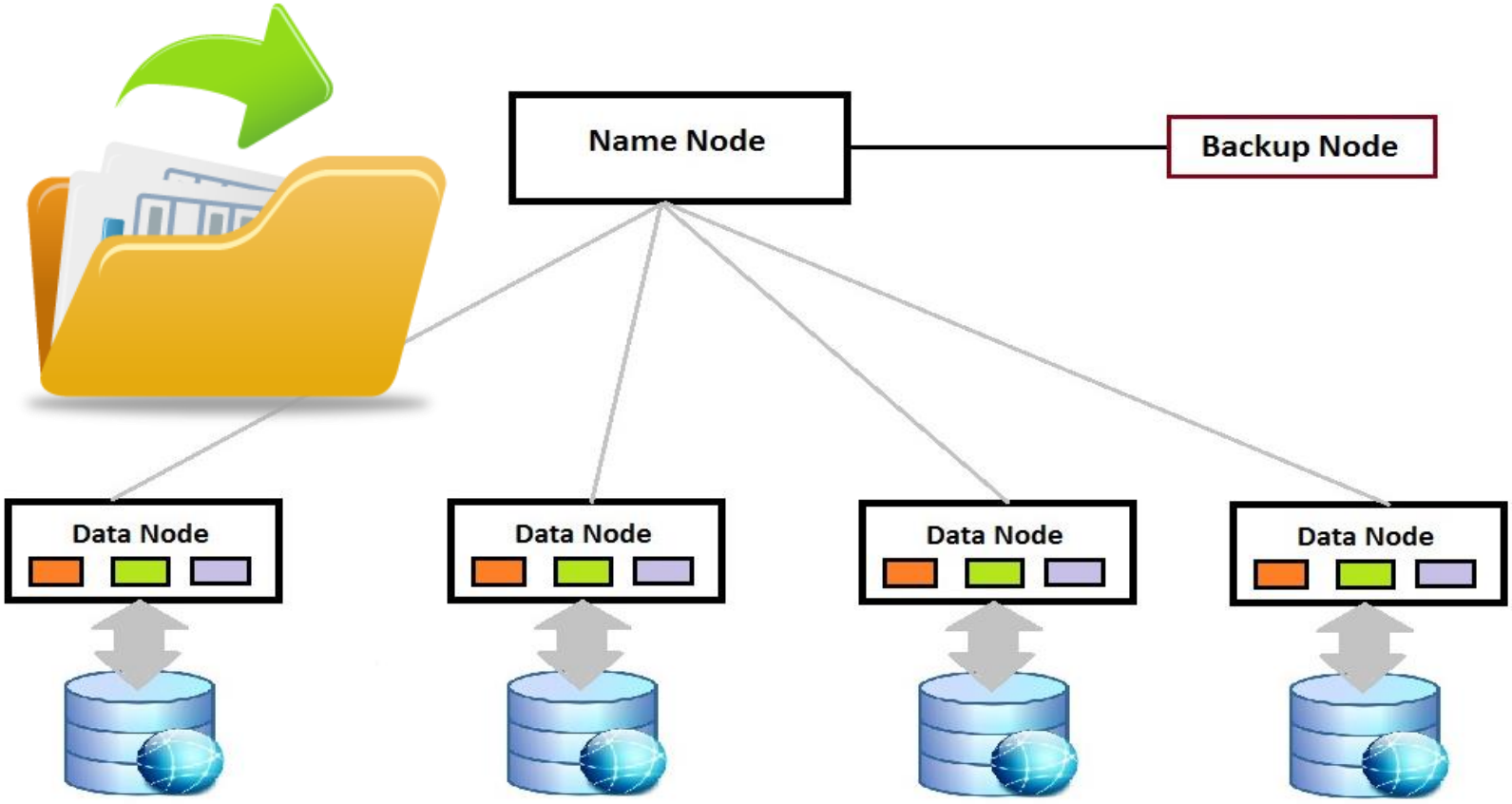


Agenda

- Hadoop
- Relational Database
- SQOOP
- Solution Architecture
- Features of SQOOP
- How SQOOP Works
- SQOOP Commands
- SQOOP Import and Export
- How SQOOP Import works
- How SQOOP Export works

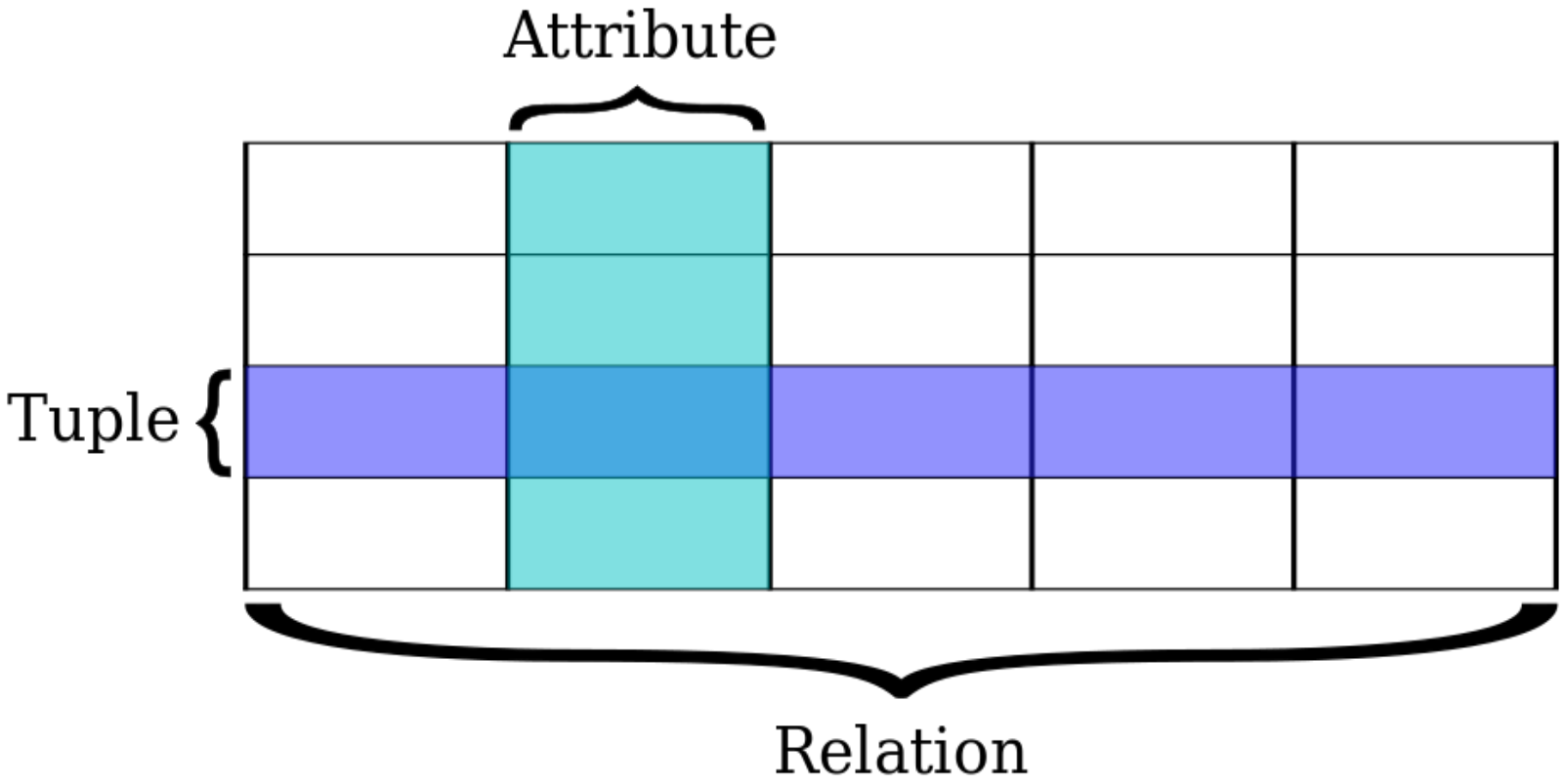
Hadoop

The Hadoop distributed File System(HDFS) is where data(unstructured) is stored in files and distributed over several nodes.



Relational Databases

The related and meaningful data (structured) is stored in two dimensional tables for analytics and reporting.



How to connect two worlds?

Hadoop:

This is File system based distributed data storage.

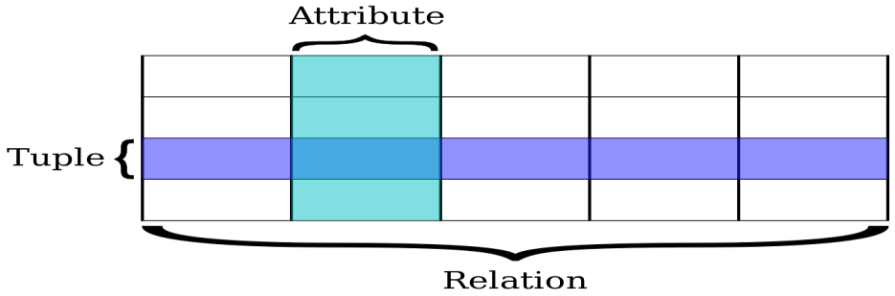
The data is processed using Map and Reduce mechanism.

The Hadoop is capable of processing Structured and Unstructured data

Relational Database:

The related and meaningful data (structured) is stored in two dimensional tables for analytics and reporting.

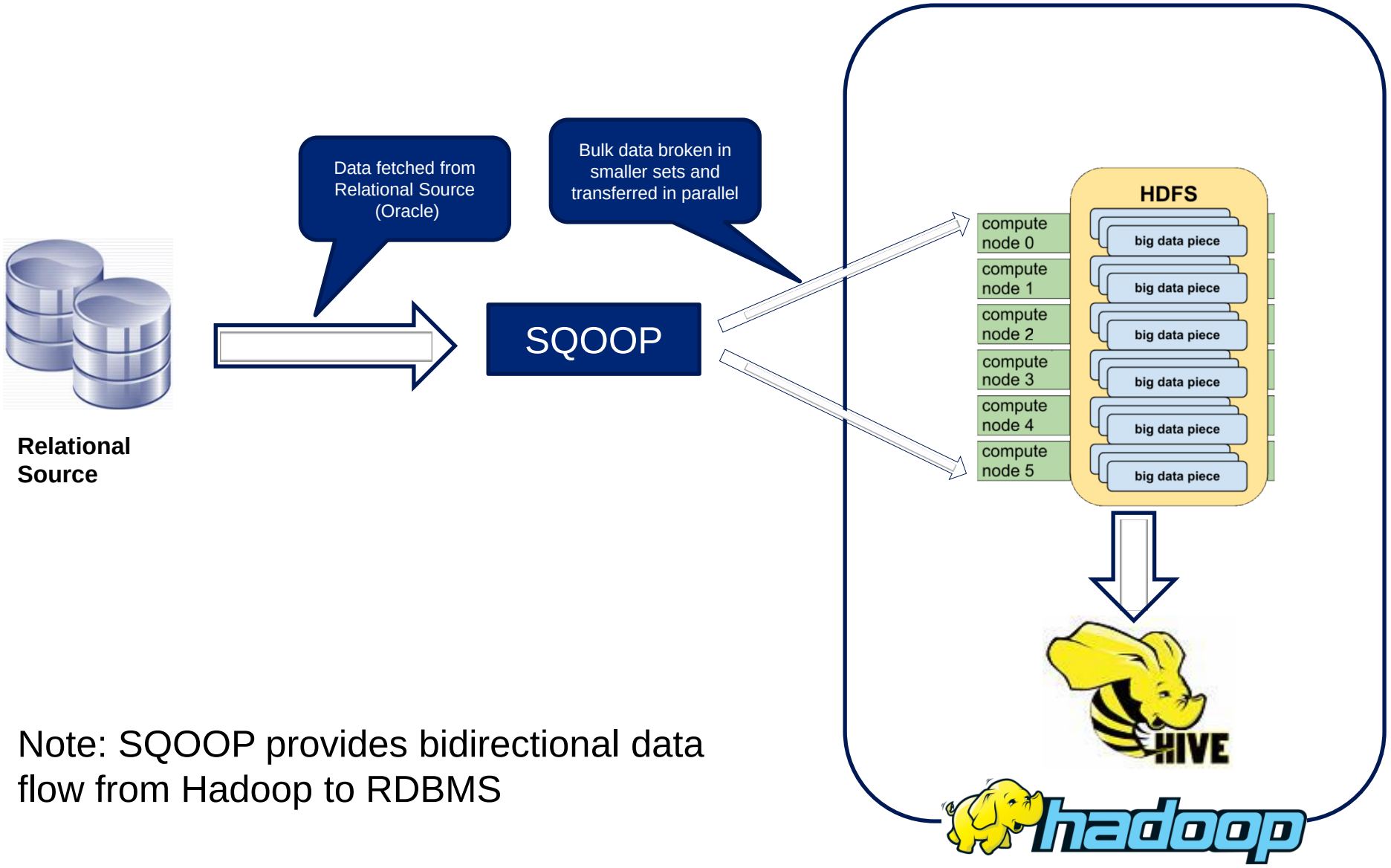
The data is processed using SQL engine.



SQOOP

- SQOOP, which stands for “SQL-to-Hadoop”, is a tool designed to transfer data between relational database(s) and Hadoop.
- It facilitates bidirectional exchange of data between relational databases (RDBMS) such as MySQL or Oracle and the Hadoop Distributed File System (HDFS).
- SQOOP can also import data into Hive.
- It uses MapReduce to read data from source tables.
- Supports incremental loads.
- Accepts vendor specific plug-ins for high performance import and export.

The Complete Solution …

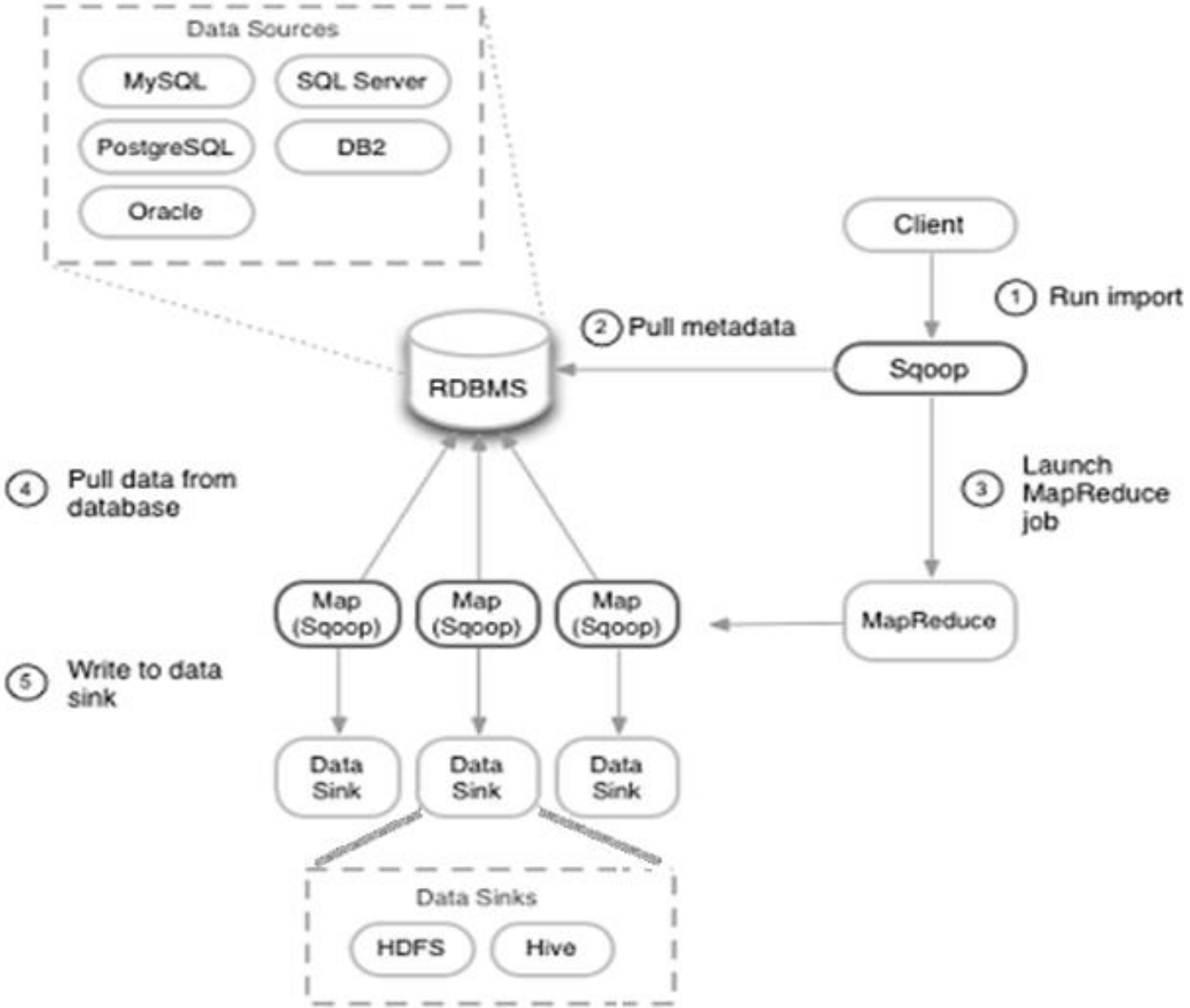


Note: SQOOP provides bidirectional data flow from Hadoop to RDBMS

Features of SQOOP

- Compatible with almost any JDBC enabled database.
- Generates Hive definition and auto-loads into Hive.
- Supports the use of WHERE clause.
- Supports incremental loads.
- Open source, comes bundled in Cloudera distributions.
- Apart from the built-in SQOOP connectors for a range of popular databases, various third party connectors are available for data stores ranging from EDW (Netezza, Teradata and Oracle) to NoSQL (Couchbase).

How SQOOP Works



SQOOP Commands

- **codegen** Generate code to interact with database records.
- **create-hive-table** import a table definition into Hive.
- **eval** Evaluate a SQL statement and display the results.
- **export** Export a HDFS directory to a database table.
- **help** List the available commands.
- **import** Import a table from a database to HDFS.
- **import-all-tables** Import tables from a database to HDFS.
- **list-databases** List available databases on a server.
- **list-table** List available tables in a database.
- **version** Display version information.

SQOOP Import and Export

SQOOP Import –

- Divide table into ranges using primary key max/min.
- Create mappers for each range.
- Mappers write to multiple HDFS nodes.
- Creates text or sequence files.
- Generates Java class for resulting HDFS file
- Generates Hive definition and auto loads into Hive.

SQOOP Export –

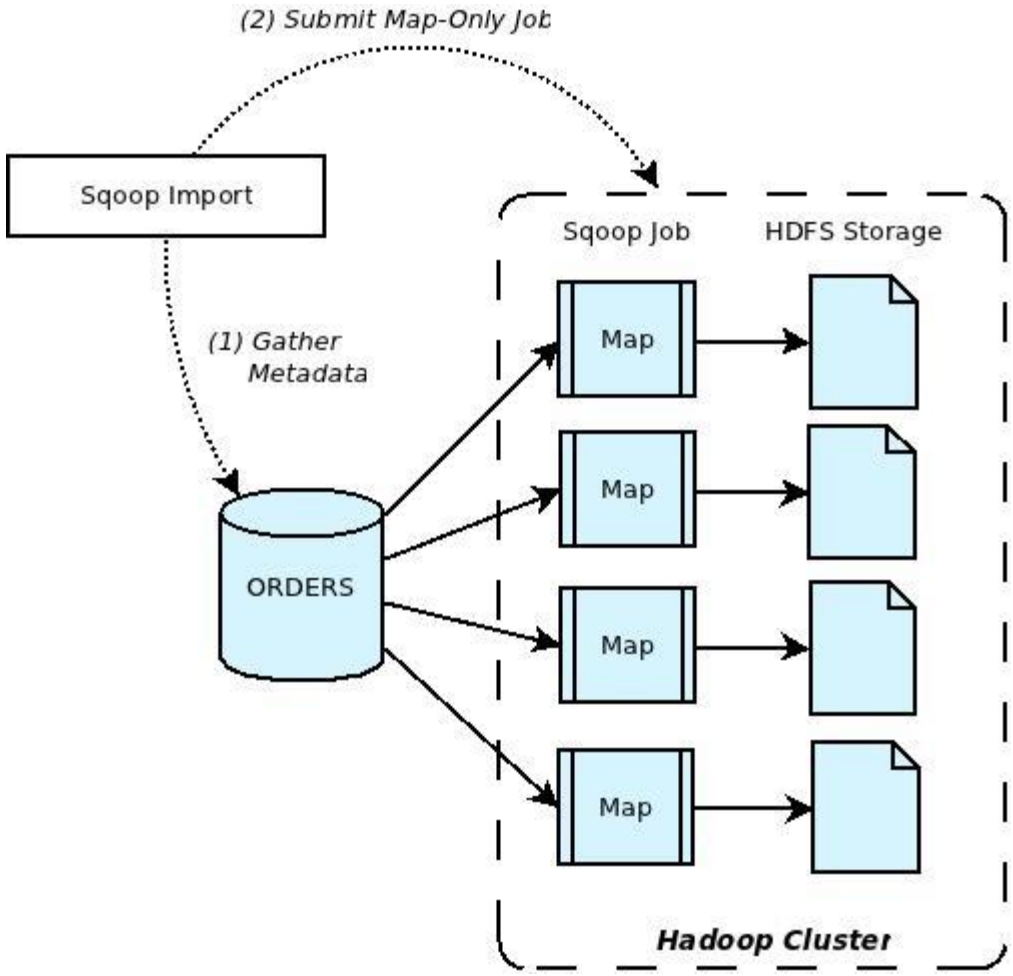
- Read files in HDFS directory via MapReduce.
- Bulk parallel insert into database table.

SQOOP Import

SQOOP import
--connect jdbc:mysql://localhost/testdb [Connection String]
--username USER_NAME [DB Username]
--password PASSWORD [DB Password]
--table TABLE_NAME [Table to be imported]

- SQOOP introspects the database to gather the necessary metadata for the data being imported.
- A Map-only Hadoop job is submitted to cluster by SQOOP.
- The Map-only job performs data transfer using the metadata captured in the previous step.
- The imported data is saved in a directory on HDFS based on the table being imported.
- By default, the files are comma delimited with new lines for different records.

How SQOOP Import works

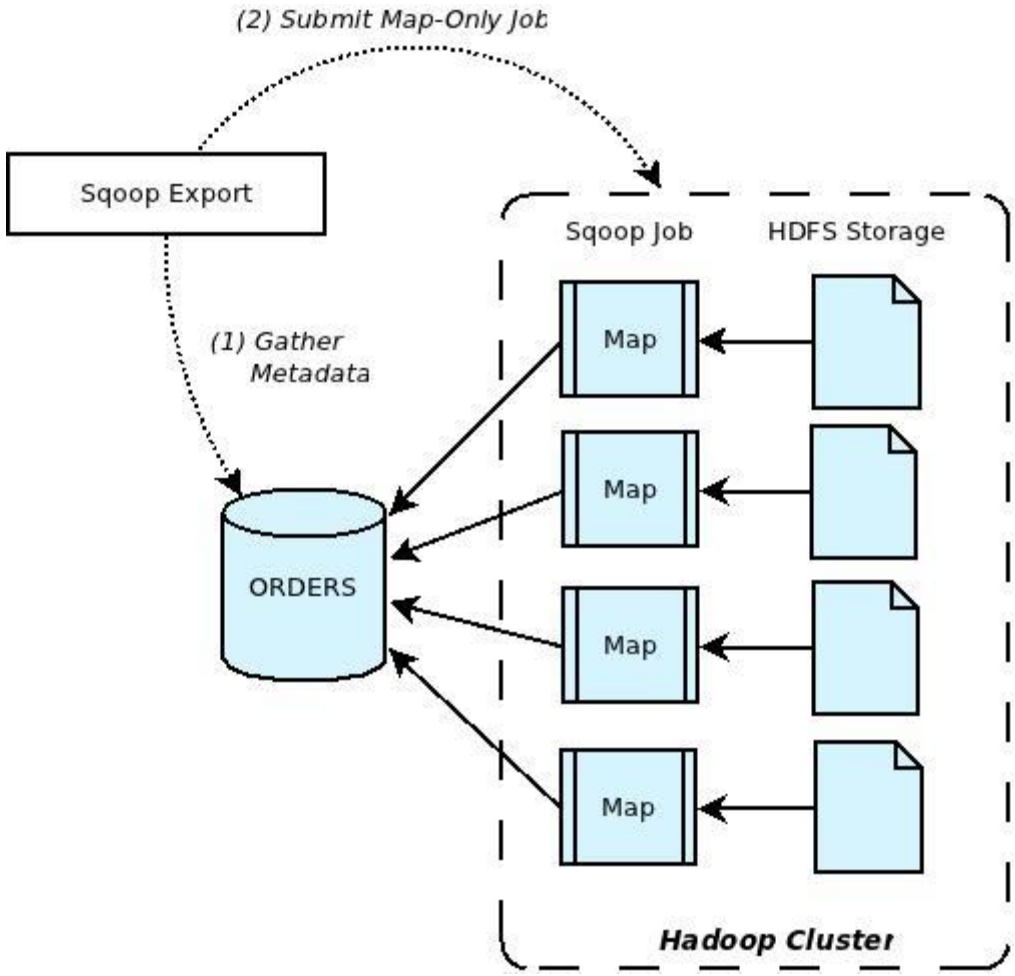


SQOOP Export

SQOOP export
--connect jdbc:mysql://localhost/testdb [Connection String]
--username USER_NAME [DB Username]
--password PASSWORD [DB Password]
--table TABLE_NAME [Table to be imported]
--export-dir DIR_NAME [Output directory]

- SQOOP introspects the database to gather the necessary metadata for the data being exported.
- The data to be exported is divided into splits.
- Individual map only jobs are used to push the splits to the database.
- Each map task performs this transfer over many transactions in order to ensure optimal throughput and minimal resource utilization.

How SQOOP Export works



References

- https://cwiki.apache.org/confluence/display/SQOOP/SQOOP+2
- https://blogs.apache.org/SQOOP/entry/apache_SQOOP_highlights_of_SQOOP
- http://wiki.apache.org/hadoop/Distributions%20and%20Commercial%20Support