# SQOOP

Sqoop is a tool designed to transfer data between Hadoop and relational databases or mainframes. You can use Sqoop to import data from a relational database management system (RDBMS) such as MySQL or Oracle or a mainframe into the Hadoop Distributed File System (HDFS), transform the data in Hadoop MapReduce, and then export the data back into an RDBMS.

Sqoop automates most of this process, relying on the database to describe the schema for the data to be imported. Sqoop uses MapReduce to import and export the data, which provides parallel operation as well as fault tolerance.

### Why SQOOP?

#### Technical Challenges:

When we pull data from a table call transaction which might contain millions of data, using a SQLload,

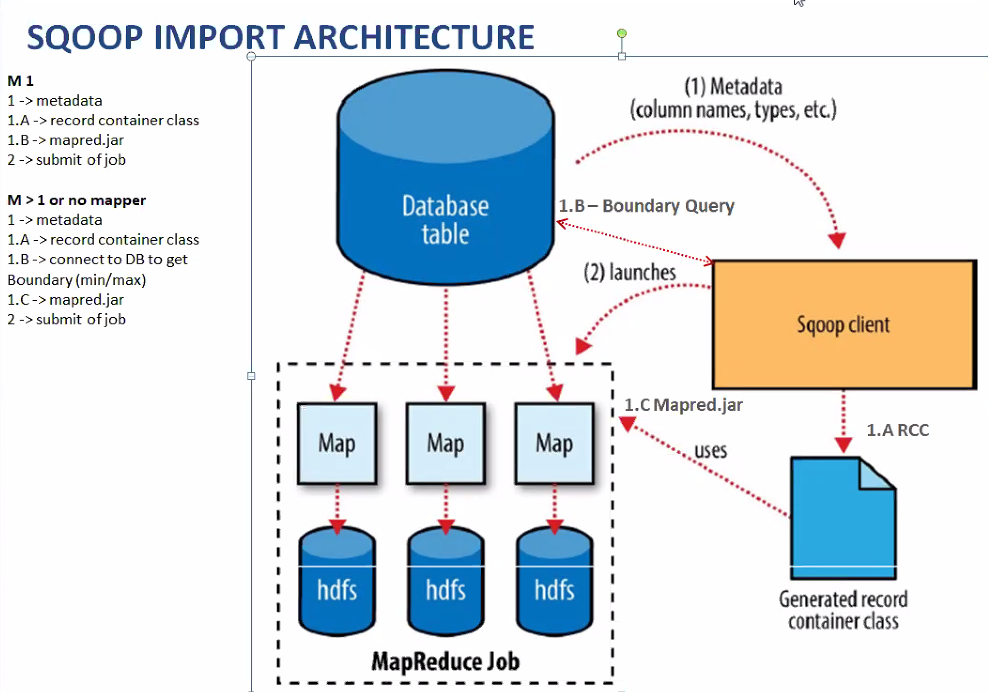
First Export 🡪 to file

Create a .sh file to load into HDFS into edge node using FTP/SFTP protocol

Ensure, transfer is complete or not

Data transfer happens using a single thread

# Sqoop import arch



* 1. Get meta data of database
  2. Record container class - create a java program having datatype convention data
  3. Mapred.jar – map reduce program – collection of data

1. Submitting of mapred.jar job to YARN
2. YARN architecture

## Sqoop Command Structure:

sqoop + toolname(import/export) + generic\_argument + tool\_specific\_args

sqoop list-tables --connect jdbc:mysql://localhost/custdb --username root --password root;

--direct 🡪 fetch in one shot using the source tool

--fetch size 🡪 how much data should be presented to Hadoop mappers at a time, default is 1000 rows at a time

When and Why to use single – and double -- ?

If we want to give more than one command then – is giv