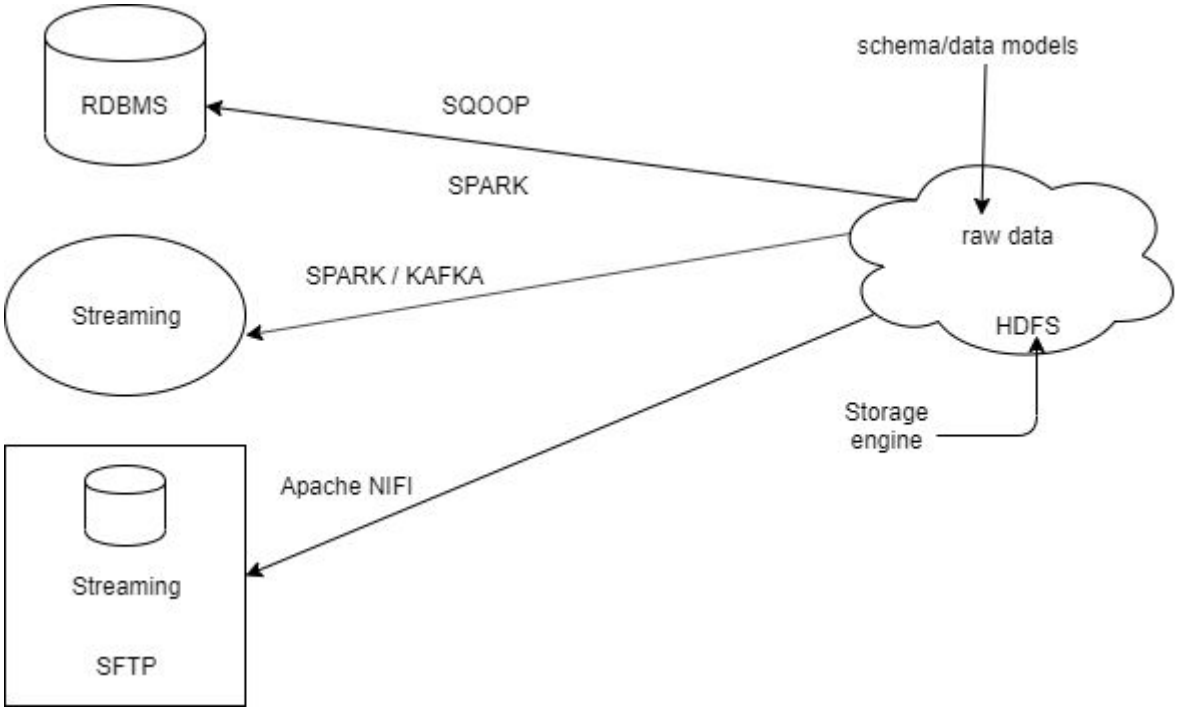
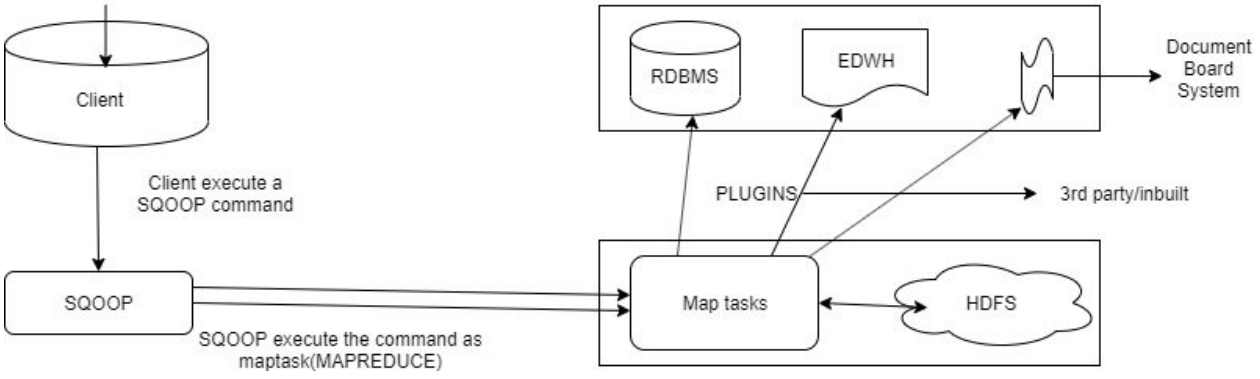
➔ DATA INGESTION :



- Typically the first task in using the hadoop cluster is getting your big data into HDFS.You have several options to choose from and typically you may need to use more than one tool depending on the sources of your big data.
- NOTE : When putting data into hadoop do not forget one of the essentials of hadoop, no schema is applied when the data goes in.In other words keep your big data in its raw format(same as source) and worry about applying schema to it later when you transform and analyse the data.

➔ SQOOP ARCHITECTURE :



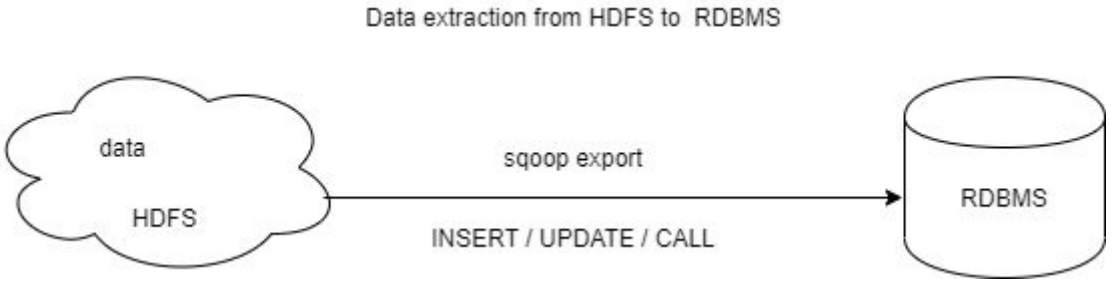
- Map task (default = 4) (used for parallel processing)

- Using map reduce to perform SQOOP commands provides parallel operations as well as fault tolerance.
- Apache SQOOP efficiently transfers bulk data between apache hadoop and structured data stores such as RDBMS.
- SQOOP can also be used to extract data from hadoop and export it to external structured data stores.
- Apache SQOOP does the following to integrate the bulk data moment between hadoop and structured data stores :
    - Import sequential datasets from mainframe.
    - Import direct to ORC file.(optimised row column data)
    - Parallel data transfer.
    - Load balancing.

- SQOOP import command :
    - **sqoop import <generic args> <sqoop args>**
    - Importing a table using sqoop :
        - **sqoop import --connect jdbc:mysql://host/nyse**
        - **--table StockPrices**
        - **--target-dir /data/stockprice/**
        - **--as-textfile**
        - The above sqoop command imports a database table named Stock Price into a folder in HDFS named /data/stockprice
    - Other useful import arguments include :
        - **--columns** (a comma separated list of the columns in the table to import)
        - **--fields-terminated-by** (specify the delimiter) (sqoop uses a comma(,) by default as a delimiter)
        - **--append** (the data is appended to an existing dataset in HDFS)
        - **--split-by** (the column used to determine how the data is split between mappers) (if you do not specify split by column then the primary key column is used)
        - **--m** (the number of map tasks to use)
        - **--query** (used instead of --table,the imported data are the resulting records from the given sql query)
        - **-- compress** (enable compression)

- Import specific column using SQOOP :
    - **sqoop import**
    - **--connect jdbc:mysql://host/nyse**
    - **--query "SELECT * FROM StockPrices s**
    - **WHERE s.Volume >= 1000000**
    - **AND \$CONDITIONS"**
    - **--target-dir /data/highvolume/**
    - **--as-textfile**
    - **--split-by StockSymbol**
    - Based on the above command only rows whose volume column is greater than 1000000 will be imported.
    - NOTE 1 : The $CONDITIONS token must appear somewhere in the where clause of your sequel query so that the data can be split between mappers.
    - NOTE 2 : If you use --query then you must also specify a --split-by column or the sqoop command will fail.
    - WARNING : You can use either --query or --table but attempting to define both results in an error.

➔ SQOOP export commands :



- ◆ SQOOP extract process will read a set of delimited text file from HDFS in parallel,parse them into records and insert them as new rows in a target database table.
- ◆ The sqoop export tools run in 3 modes :
    - Insert mode : The records being exported are inserted into the table using a sequel insert statement.

- Update mode : An update sql statement is executed for existing rows and an insert can be used for new rows.
- Call mode : A stored procedure is invoked for each record.

◆ SQOOP export arguments :

- **--export-dir**
- **--input-fields-terminated-by** (the input field delimiter)
- **--update-mode** (specify how updates are performed when new rows are found with non-matching keys in the database) (default values are update only and allows insert)

◆ SQOOP export commands :

- **sqoop export**
- **--connect jdbc:sql://host/mylogs**
- **--table LogData**
- **--export-dir /data/logfiles/**
- **--input-fields-terminated-by "\t"**
- Based on the command above the table log data needs to already exist in the mylogs database.

Links :
https://sqoop.apache.org/docs/1.4.6/SqoopUserGuide.html

https://sqoop.apache.org/
https://www.cloudera.com/products/open-source/apache-hadoop/apache-sqoop.html
https://www.dezyre.com/hadoop-tutorial/hadoop-sqoop-tutorial