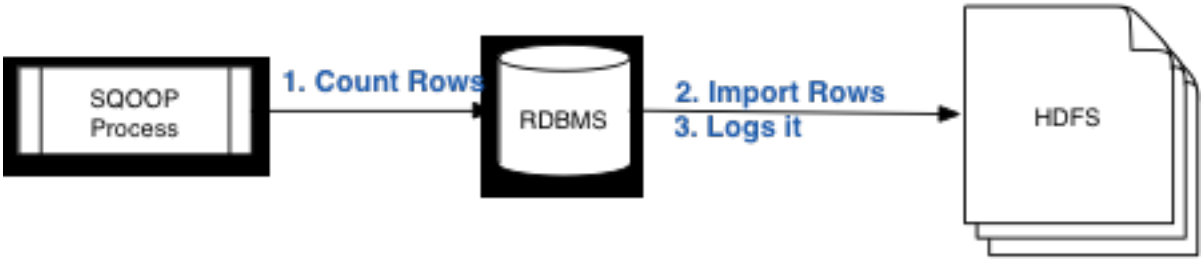
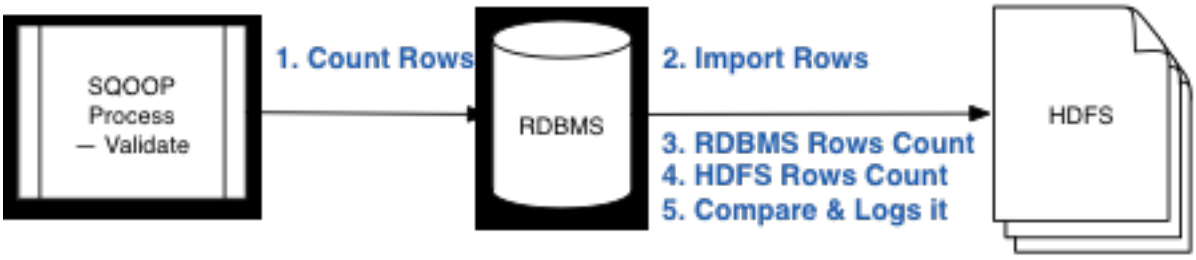
SQOOP Validate:

The Sqoop validate option is used to compare the row counts between source and target after data imported into HDFS. When the rows are deleted or added during the imports, Sqoop tracks this change and updates the log file.

SQOOP process:



SQOOP process with Validate option:



SQOOP Test scenario:

1.Insert more than 4 million Rows into MYSQL table.

-- In my test table have **14,134500 rows**

2.Import data using SQOOP into hdfs with validate option.

Example: sqoop import --connect jdbc:mysql://localhost/hive --username hive --password hive --table test **--validate** --verbose --delete-target-dir -m 1

3.Delete some rows on mysql table.

3.Delete some rows on mysql table.

Example: delete from test where column = ‘column_value’. It deletes 68850

4.Output:

17/08/14 15:25:44 INFO mapreduce.Job: Counters: 30

File System Counters

FILE: Number of bytes read=0

FILE: Number of bytes written=160073

FILE: Number of read operations=0

FILE: Number of large read operations=0

FILE: Number of write operations=0

HDFS: Number of bytes read=87

HDFS: Number of bytes written=1352442000

HDFS: Number of read operations=4

HDFS: Number of large read operations=0

HDFS: Number of write operations=2

Job Counters

Launched map tasks=1

Other local map tasks=1

Total time spent by all maps in occupied slots (ms)=153958

Total time spent by all reduces in occupied slots (ms)=0

Total time spent by all map tasks (ms)=76979

Total vcore-milliseconds taken by all map tasks=76979

Total megabyte-milliseconds taken by all map tasks=118239744

Map-Reduce Framework

Map input records=14134500

Map output records=14134500

Input split bytes=87

Spilled Records=0

Failed Shuffles=0

Merged Map outputs=0

GC time elapsed (ms)=2409

CPU time spent (ms)=59620

Physical memory (bytes) snapshot=268062720

Physical memory (bytes) snapshot=268062720

Virtual memory (bytes) snapshot=3242549248

Total committed heap usage (bytes)=170917888

File Input Format Counters

Bytes Read=0

File Output Format Counters

Bytes Written=1352442000

17/08/14 15:25:44 INFO mapreduce.ImportJobBase: Transferred 1.2596 GB in 97.074 seconds (13.2867 MB/sec)

17/08/14 15:25:44 INFO mapreduce.ImportJobBase: Retrieved 14134500 records.

17/08/14 15:25:44 DEBUG mapreduce.ImportJobBase: Validating imported data.

17/08/14 15:25:44 DEBUG manager.SqlManager: No connection paramenters specified. Using regular API for making connection.

17/08/14 15:25:56 INFO mapred.ClientServiceDelegate: Application state is completed**. FinalApplicationStatus=SUCCEEDED**. Redirecting to job history server

17/08/14 15:25:57 INFO mapreduce.JobBase: Validating the integrity of the import using the following configuration

Validator : org.apache.sqoop.validation.RowCountValidator

Threshold Specifier : org.apache.sqoop.validation.AbsoluteValidationThreshold

Failure Handler : org.apache.sqoop.validation.AbortOnFailureHandler

17/08/14 15:25:57 DEBUG validation.RowCountValidator: Validating data using row counts: Source [**14065650**] with Target[**14134500**]

17/08/14 15:25:57 DEBUG validation.AbsoluteValidationThreshold: Absolute Validation threshold comparing 14065650 with 14134500

17/08/14 15:25:57 DEBUG util.ClassLoaderStack: Restoring classloader: sun.misc.Launcher$AppClassLoader@215be6bb

17/08/14 15:25:57 ERROR tool.ImportTool: Error during import: Error validating row counts