Sqoop

A tool designed to transfer the data between Relational databases and Hadoop [Hbase, Hive]

Sqoop Flow & Features

Key Features

- **Full Load**
- **Incremental load**
- **Parallel import & export**
- **Load data directly to Hive/Hbase**

Key Notes on Import & Export Data transfer - Sqoop

Import : Each row in a table will be treated as Record in HDFS as part of Sqoop import and records will be stored as Avro ,sequence file, Parquet file

Export : Set of files from HDFS to RDBMS and ensure the target table exists in database. Files are read & Parsed into set of records according to user specified delimiters

- "--query" or "e"
- "-m 1" or "–num-mappers 1"
- --password or –P

Mappers divide their work during Sqoop import execution :

1. It takes one records to get metadata and build java file
2. 2. use java file and build the jar file

3..boundary val to get min and max value on primary key column

4. max-min/4 to get the split size

Sqoop Hands on commands – Cloudera

Connect to Mysql from cloudera Terminal:

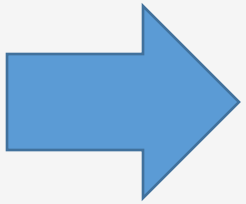
mysql -u root -p
password :cloudera

How to get the databases & tables in mysql terminal:

show databases;
show tables;
use retail_db; ----->Database in the name of "retail_db"

How to get the list of databases from terminal :

sqoop-list-databases \
--connect "jdbc:mysql://quickstart.cloudera:3306" \
--username retail_dba \
--password cloudera



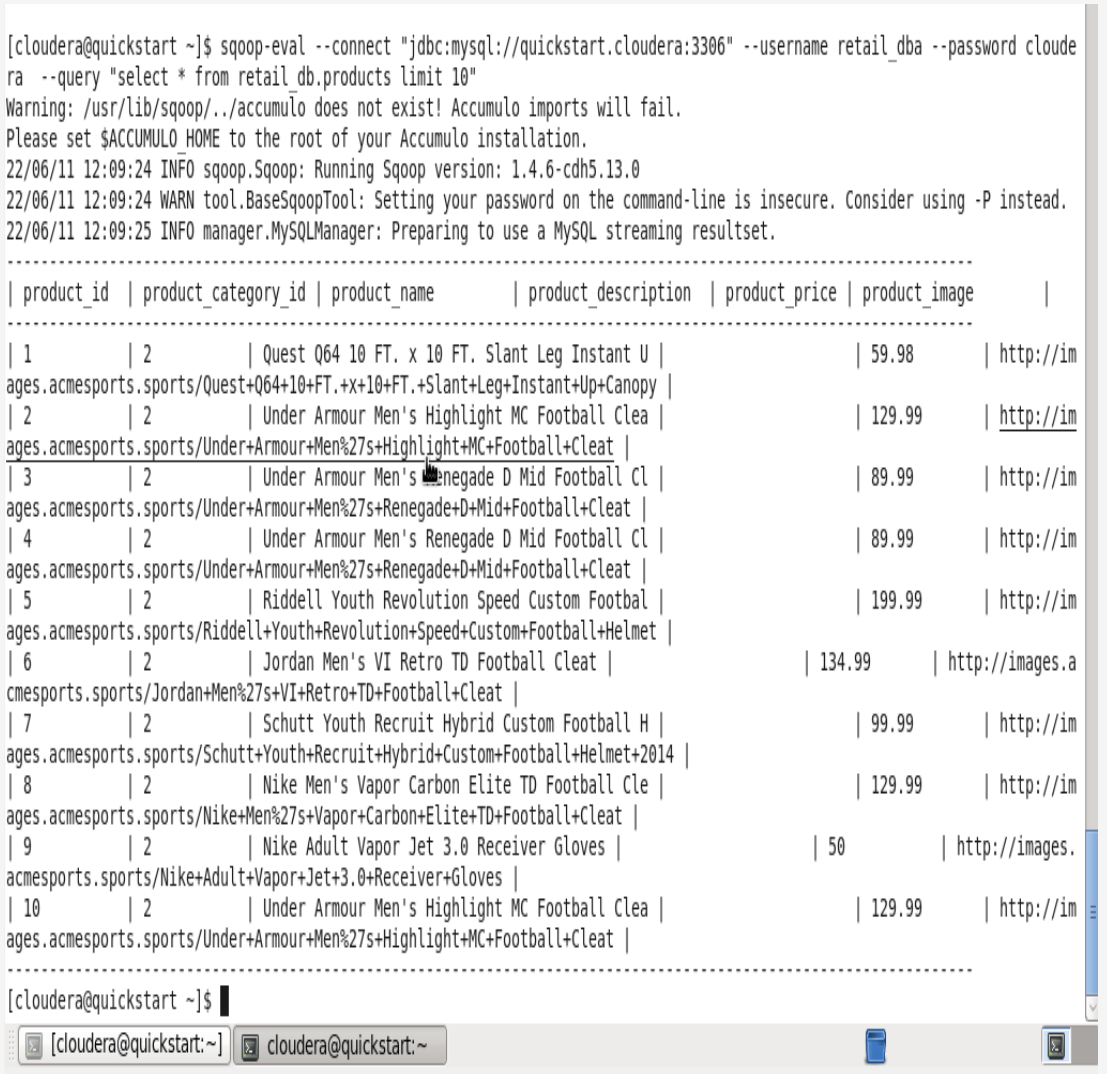
```
cloudera@quickstart:~
File  Edit  View  Search  Terminal  Help
[cloudera@quickstart ~]$ mysql -u root -p
Enter password:
Welcome to the MySQL monitor.  Commands end with ; or \g.
Your MySQL connection id is 15
Server version: 5.1.73 Source distribution

Copyright (c) 2000, 2013, Oracle and/or its affiliates. All rights reserved.

Oracle is a registered trademark of Oracle Corporation and/or its
affiliates. Other names may be trademarks of their respective
owners.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

mysql> show databases;
+--------------------+
| Database           |
+--------------------+
| information_schema |
| cm                 |
| firehose           |
| hue                |
| metastore          |
| mysql              |
| nav                |
| navms              |
| oozie              |
| retail_db          |
| rman               |
| sentry             |
+--------------------+
12 rows in set (0.22 sec)

mysql> use retail_db;
Reading table information for completion of table and column names
You can turn off this feature to get a quicker startup with -A
```

Sqoop Eval concept

Using Sqoop eval you can connect to database and run the select to see the data from cloudera terminal

Command :

```
sqoop -eval \
--connect
"jdbc:mysql://quickstart.cloudera:3
306" \
--username retail_dba \
--password cloudera  \
--query "select * from products
limit 10"
```



How to create a database and table in Mysql

Login to cloudera and connect to Mysql terminal :

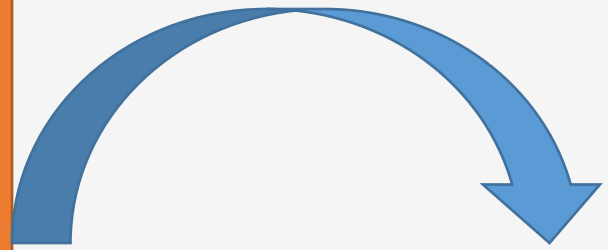
```
create database testing;

use testing;
```

Create table statement :

```
create table customers
(id int,
lastname varchar(200),
firstname varchar(150),
address varchar(200),
mobile int,
city varchar(200));
```

Insert statement :

```
insert into customers (id,lastname,firstname,address,mobile,city)
values
(100,'kumar','ram','nehru road near post office',8220212387,'chennai'),
(101,'hasan','kamal','patalia road near baskin robins',8120212387,'pune')
(103,'jose','diana','main junction road near jj
complex',8220215387,'chennai')
(104,'karthick','ram','kk nagar road',8220612387,'mumbai')
(105,'sundaram','soma','JJ road near marriat hotel',8520212387,'pune');
```



```
mysql> select * from testing.customers;
+------+----------+-----------+----------------------------------+------------+---------+
| id   | lastname | firstname | address                          | mobile     | city    |
+------+----------+-----------+----------------------------------+------------+---------+
|  100 | kumar    | ram       | nehru road near post office      | 2147483647 | chennai |
|  101 | hasan    | kamal     | patalia road near baskin robins  | 2147483647 | pune    |
|  103 | jose     | diana     | main junction road near jj complex | 2147483647 | chennai |
+------+----------+-----------+----------------------------------+------------+---------+
3 rows in set (0.11 sec)
```

Sqoop Import command [Default mapper :4]

```
sqoop import \
--connect "jdbc:mysql://quickstart.cloudera:3306"/retail_db \
--username root \
--password cloudera  \
--table orders        \
--target-dir /sqoop_practice
```

Output with 4Mappers in sqoop_practice directory :

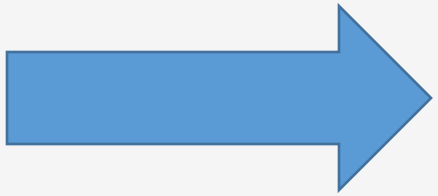
```
[cloudera@quickstart ~]$ hadoop fs -ls /sqoop_practice/
Found 5 items
-rw-r--r--   1 cloudera supergroup          0 2022-06-11 12:49 /sqoop_practice/_SUCCESS
-rw-r--r--   1 cloudera supergroup     741614 2022-06-11 12:49 /sqoop_practice/part-m-00000
-rw-r--r--   1 cloudera supergroup     753022 2022-06-11 12:49 /sqoop_practice/part-m-00001
-rw-r--r--   1 cloudera supergroup     752368 2022-06-11 12:49 /sqoop_practice/part-m-00002
-rw-r--r--   1 cloudera supergroup     752940 2022-06-11 12:49 /sqoop_practice/part-m-00003
[cloudera@quickstart ~]$
```

```
[cloudera@quickstart ~]$ hadoop fs -tail /sqoop_practice/part-m-00000
17198,2013-11-09 00:00:00.0,642,CLOSED
17199,2013-11-09 00:00:00.0,7246,PENDING_PAYMENT
17200,2013-11-09 00:00:00.0,4846,PENDING_PAYMENT
17201,2013-11-09 00:00:00.0,10506,PENDING_PAYMENT
17202,2013-11-09 00:00:00.0,4145,PROCESSING
17203,2013-11-09 00:00:00.0,6725,COMPLETE
17204,2013-11-09 00:00:00.0,3960,CLOSED
17205,2013-11-09 00:00:00.0,2715,CLOSED
17206,2013-11-09 00:00:00.0,2848,PROCESSING
17207,2013-11-09 00:00:00.0,8986,COMPLETE
17208,2013-11-09 00:00:00.0,1364,CLOSED
17209,2013-11-09 00:00:00.0,336,CLOSED
17210,2013-11-09 00:00:00.0,12143,PENDING_PAYMENT
17211,2013-11-09 00:00:00.0,1595,COMPLETE
17212,2013-11-09 00:00:00.0,11387,COMPLETE
17213,2013-11-09 00:00:00.0,6166,COMPLETE
17214,2013-11-09 00:00:00.0,585,CLOSED
17215,2013-11-09 00:00:00.0,8326,COMPLETE
17216,2013-11-09 00:00:00.0,5729,COMPLETE
```

Sqoop Import command [Default mapper :1]

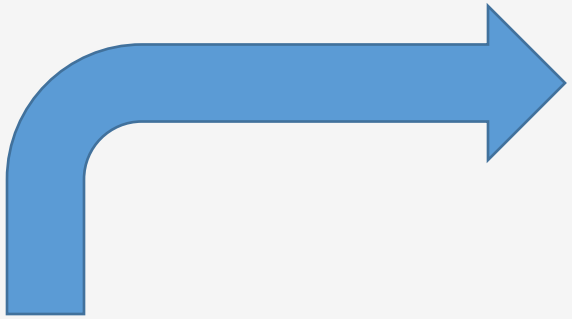
```
sqoop import --connect "jdbc:mysql://quickstart.cloudera:3306"/retail_db  --username root --password cloudera  --table orders -m 1 --target-dir /sqoop_practice_m1
```

Output with 1 Mappers in sqoop_practice directory :



```
[cloudera@quickstart ~]$ hadoop fs -ls /sqoop_practice_m1
Found 2 items
-rw-r--r--   1 cloudera supergroup          0 2022-06-11 12:58 /sqoop_practice_m1/_SUCCESS
-rw-r--r--   1 cloudera supergroup    2999944 2022-06-11 12:58 /sqoop_practice_m1/part-m-00000
[cloudera@quickstart ~]$ hadoop fs -tail /sqoop_practice_m1/part-m-00000
014-06-12 00:00:00.0,4229,PENDING
68861,2014-06-13 00:00:00.0,3031,PENDING_PAYMENT
68862,2014-06-15 00:00:00.0,7326,PROCESSING
68863,2014-06-16 00:00:00.0,3361,CLOSED
68864,2014-06-18 00:00:00.0,9634,ON_HOLD
68865,2014-06-19 00:00:00.0,4567,SUSPECTED_FRAUD
68866,2014-06-20 00:00:00.0,3890,PENDING_PAYMENT
68867,2014-06-23 00:00:00.0,869,CANCELED
68868,2014-06-24 00:00:00.0,10184,PENDING
68869,2014-06-25 00:00:00.0,7456,PROCESSING
68870,2014-06-26 00:00:00.0,3343,COMPLETE
68871,2014-06-28 00:00:00.0,4960,PENDING
68872,2014-06-29 00:00:00.0,3354,COMPLETE
68873,2014-06-30 00:00:00.0,4545,PENDING
68874,2014-07-03 00:00:00.0,1601,COMPLETE
68875,2014-07-04 00:00:00.0,10637,ON_HOLD
68876,2014-07-06 00:00:00.0,4124,COMPLETE
68877,2014-07-07 00:00:00.0,9692,ON_HOLD
68878,2014-07-08 00:00:00.0,6753,COMPLETE
68879,2014-07-09 00:00:00.0,778,COMPLETE
68880,2014-07-13 00:00:00.0,1117,COMPLETE
68881,2014-07-19 00:00:00.0,2518,PENDING_PAYMENT
68882,2014-07-22 00:00:00.0,10000,ON_HOLD
68883,2014-07-23 00:00:00.0,5533,COMPLETE
```

Import All tables with Sequence file format

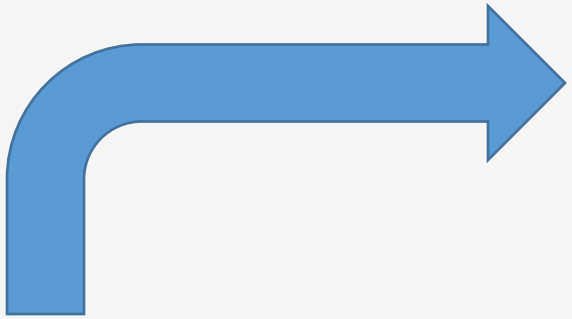


```
cloudera@quickstart ~]$ hadoop fs -ls  /user/cloudera/sqoopdir/
ound 6 items
rwxr-xr-x   - cloudera cloudera          0 2022-06-11 13:10 /user/cloudera/sqoopdir/categories
rwxr-xr-x   - cloudera cloudera          0 2022-06-11 13:12 /user/cloudera/sqoopdir/customers
rwxr-xr-x   - cloudera cloudera          0 2022-06-11 13:12 /user/cloudera/sqoopdir/departments
rwxr-xr-x   - cloudera cloudera          0 2022-06-11 13:13 /user/cloudera/sqoopdir/order_items
rwxr-xr-x   - cloudera cloudera          0 2022-06-11 13:14 /user/cloudera/sqoopdir/orders
rwxr-xr-x   - cloudera cloudera          0 2022-06-11 13:15 /user/cloudera/sqoopdir/products
cloudera@quickstart ~]$ hadoop fs -ls  /user/cloudera/sqoopdir/categories/
ound 5 items
rw-r--r--   1 cloudera cloudera          0 2022-06-11 13:10 /user/cloudera/sqoopdir/categories/_SUCCESS
rw-r--r--   1 cloudera cloudera        681 2022-06-11 13:10 /user/cloudera/sqoopdir/categories/part-m-00000
rw-r--r--   1 cloudera cloudera        642 2022-06-11 13:10 /user/cloudera/sqoopdir/categories/part-m-00001
rw-r--r--   1 cloudera cloudera        645 2022-06-11 13:10 /user/cloudera/sqoopdir/categories/part-m-00002
rw-r--r--   1 cloudera cloudera        630 2022-06-11 13:10 /user/cloudera/sqoopdir/categories/part-m-00003
cloudera@quickstart ~]$ sqoop import-all-tables --connect "jdbc:mysql://quickstart.cloudera:3306"/retail_db --username root
-password cloudera  --as-sequencefile -m 4 --warehouse-dir /user/cloudera/sqoopdir
```

cloudera@quickstart:~ | [cloudera@quickstart:~]

```
sqoop import-all-tables --connect
"jdbc:mysql://quickstart.cloudera:3306"/retail_db
--username root --password cloudera  --as-
sequencefile -m 4 --warehouse-dir
/user/cloudera/sqoopdir
```

Compress - Sqoop



```
[cloudera@quickstart ~]$ hadoop fs -ls  /sqoop_practice_compress/
Found 2 items
-rw-r--r--   1 cloudera supergroup          0 2022-06-11 13:31 /sqoop_practice_compress/_SUCCESS
-rw-r--r--   1 cloudera supergroup     471106 2022-06-11 13:31 /sqoop_practice_compress/part-m-00000.gz
[cloudera@quickstart ~]$
```

cloudera@quickstart:~ | [cloudera@quickstart:~]

```
sqoop import --connect
"jdbc:mysql://quickstart.cloudera:3306"/retail_db
--username root --password cloudera  --table
orders --compress -m 1 --target-dir
/sqoop_practice_compress
```

Mappers : 1
Compress : Gz algorithm

Subset of columns & where clause - Sqoop

```
sqoop import --connect "jdbc:mysql://quickstart.cloudera:3306"/retail_db --username root --password cloudera  --table orders --columns order_id,order_status --where "order_status in ('complete','closed')" -m 1 --target-dir /sqoop_practice_subset
```

```
68883,COMPLETE
[cloudera@quickstart ~]$ hadoop fs -tail  /sqoop_practice_subset/part-m-00000
OMPLETE
68742,COMPLETE
68745,CLOSED
68747,CLOSED
68753,COMPLETE
68755,COMPLETE
68756,CLOSED
68758,CLOSED
68759,COMPLETE
68760,COMPLETE
68761,COMPLETE
68764,COMPLETE
68768,COMPLETE
68772,COMPLETE
68773,CLOSED
68774,COMPLETE
68776,CLOSED
68778,COMPLETE
68781,CLOSED
68784,COMPLETE
68786,COMPLETE
68788,COMPLETE
68789,COMPLETE
```

Split By – Sqoop import

--split-by comes into picture when there is no primary key column and it is needed to indicate the column on which mappers should divide the work .In addition ensure that your primary key should not have lot of outliers to degrade your performance

```
sqoop import --connect
"jdbc:mysql://quickstart.cloudera:3306"/retail_db  --
username root --password cloudera  --table orders --–
split-by  --target-dir /sqoop_practice_splitby
```

Verbose & Append command

Append the data on the HDFS location :

```
sqoop import --connect "jdbc:mysql://quickstart.cloudera:3306"/retail_db --username root --password cloudera  --table orders --target-dir /sqoop_practice --append
```

Verbose flag to see more Logs & debugging info:

```
sqoop import --connect "jdbc:mysql://quickstart.cloudera:3306"/retail_db --username root --password cloudera  --table orders --target-dir /sqoop_practice --append
```

Key Commands

Sqoop job —list

Sqoop job —exec orders_Table_ingestion_job

Sqoop job —delete orders_Table_ingestion_job

Sqoop Job creation :

```
sqoop job --create orders_Table_ingestion_job
import --connect
"jdbc:mysql://quickstart.cloudera:3306"/retail_db  --
username root --password cloudera  --table orders -
-target-dir /sqoop_practice  --append
```

Happy learning !!!!!!!!!

Thank you