Uyen Dang

16227610

CS5590 Big Data Programming – Summer 2020

PART 1 - Use Sqoop to import and export mySQL Tables to HDFS.

Step 1: show database:

```
mysql> show databases;
+--------------------+
| Database           |
+--------------------+
| information_schema |
| cm                 |
| db1                |
| firehose           |
| hue                |
| metastore          |
| mysql              |
| nav                |
| navms              |
| oozie              |
| retail_db          |
| rman               |
| sentry             |
+--------------------+
13 rows in set (0.00 sec)
```

Step 2: Create available database and select database

```
mysql> create database db1;
Query OK, 1 row affected (0.00 sec)

mysql> use db1;
Database changed
```

Step 3: Create table and insert value

```
mysql> create table acad(emp_id INT NOT NULL AUTO_INCREMENT, emp_name VARCHAR(10
0), emp_sal INT, PRIMARY KEY (emp_id));
Query OK, 0 rows affected (0.04 sec)
```

```
mysql> insert into acad values(5,"sanam",50000),(6,"opra",6000),(7,"yella",70000
);
Query OK, 3 rows affected (0.02 sec)
Records: 3  Duplicates: 0  Warnings: 0

mysql> select * from acad;
+--------+----------+---------+
| emp_id | emp_name | emp_sal |
+--------+----------+---------+
|      5 | sanam    |   50000 |
|      6 | opra     |    6000 |
|      7 | yella    |   70000 |
+--------+----------+---------+
3 rows in set (0.00 sec)
```

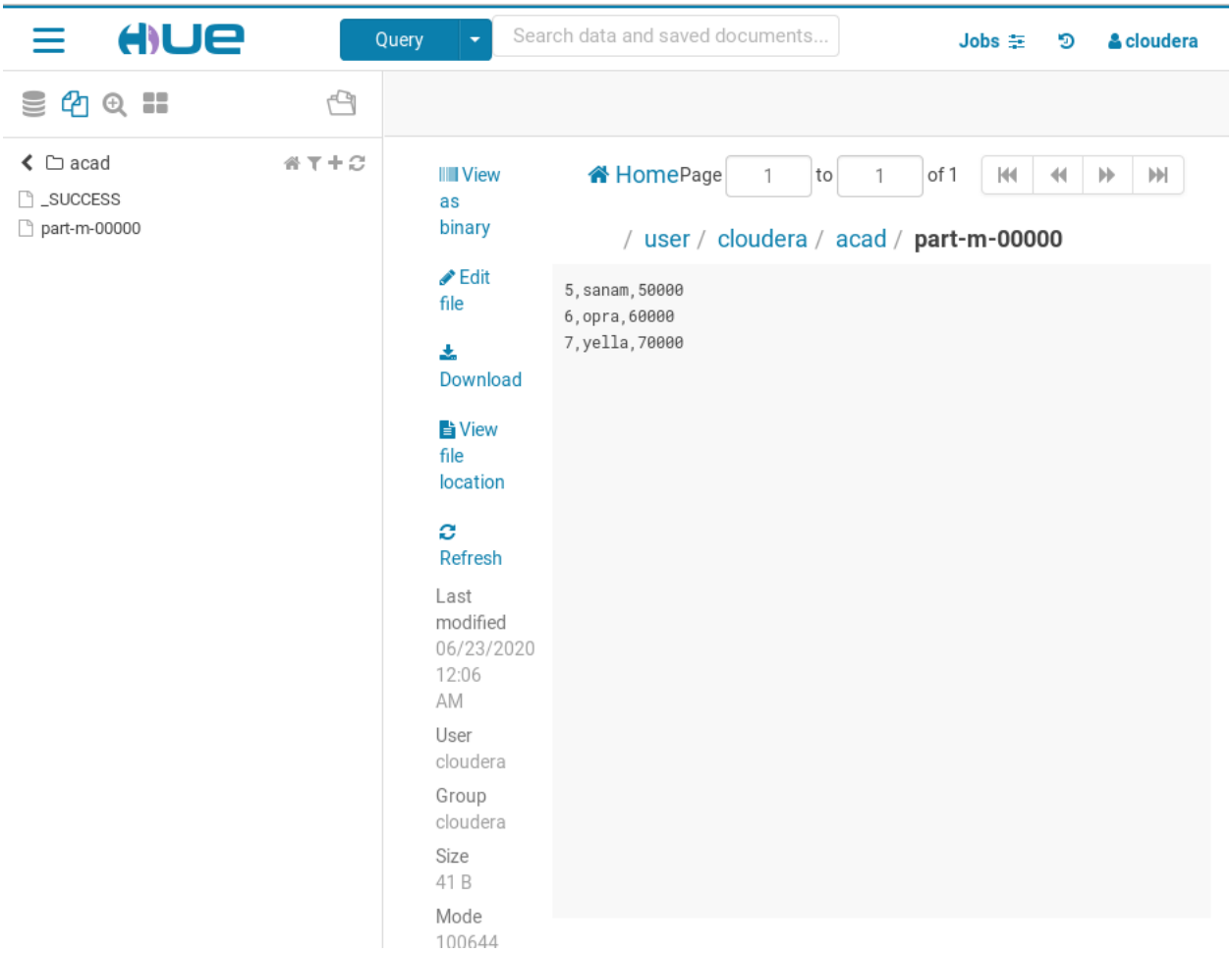
Step 4: Import Sqoop

```
[cloudera@quickstart Downloads]$ sqoop import --connect jdbc:mysql://localhost/d
b2 --username root --password cloudera --table acad --m 1
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
20/06/22 17:04:36 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5.13.0
20/06/22 17:04:36 WARN tool.BaseSqoopTool: Setting your password on the command-
line is insecure. Consider using -P instead.
20/06/22 17:04:37 INFO manager.MySQLManager: Preparing to use a MySQL streaming
resultset.
20/06/22 17:04:37 INFO tool.CodeGenTool: Beginning code generation
20/06/22 17:04:38 INFO manager.SqlManager: Executing SQL statement: SELECT t.* F
ROM `acad` AS t LIMIT 1
20/06/22 17:04:39 INFO manager.SqlManager: Executing SQL statement: SELECT t.* F
ROM `acad` AS t LIMIT 1
20/06/22 17:04:39 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/ha
doop-mapreduce
Note: /tmp/sqoop-cloudera/compile/8108ff668046406f38b7a3218b33d3e1/acad.java use
s or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
20/06/22 17:04:46 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-clou
dera/compile/8108ff668046406f38b7a3218b33d3e1/acad.jar
20/06/22 17:04:46 WARN manager.MySQLManager: It looks like you are importing fro
m mysql.
20/06/22 17:04:46 WARN manager.MySQLManager: This transfer can be faster! Use th
```

```
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=26692
                Total vcore-milliseconds taken by all map tasks=26692
                Total megabyte-milliseconds taken by all map tasks=27332608
        Map-Reduce Framework
                Map input records=3
                Map output records=3
                Input split bytes=87
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=367
                CPU time spent (ms)=2310
                Physical memory (bytes) snapshot=111063040
                Virtual memory (bytes) snapshot=1512013824
                Total committed heap usage (bytes)=60751872
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=41
20/06/22 17:06:20 INFO mapreduce.ImportJobBase: Transferred 41 bytes in 87.7575
seconds (0.4672 bytes/sec)
20/06/22 17:06:20 INFO mapreduce.ImportJobBase: Retrieved 3 records.


[cloudera@quickstart Downloads]$ hadoop fs -ls
Found 9 items
drwxr-xr-x   - cloudera cloudera          0 2020-06-16 05:33 ICP3
drwxr-xr-x   - cloudera cloudera          0 2020-06-16 20:13 ICP_3
drwxr-xr-x   - cloudera cloudera          0 2020-06-08 16:28 UyenDang
drwxr-xr-x   - cloudera cloudera          0 2020-06-22 17:06 acad
drwxr-xr-x   - cloudera cloudera          0 2020-06-08 17:50 icp1
drwxr-xr-x   - cloudera cloudera          0 2020-06-10 17:10 icp2
-rw-r--r--   1 cloudera cloudera         52 2020-06-15 17:46 icp3
drwxr-xr-x   - cloudera cloudera          0 2020-06-08 16:36 isp1
-rw-r--r--   1 cloudera cloudera    5590188 2020-06-08 16:25 shakespeare.txt
[cloudera@quickstart Downloads]$ hadoop fs -ls acad/
Found 2 items
-rw-r--r--   1 cloudera cloudera          0 2020-06-22 17:06 acad/_SUCCESS
-rw-r--r--   1 cloudera cloudera         41 2020-06-22 17:06 acad/part-m-00000
[cloudera@quickstart Downloads]$ hadoop fs -cat acad/*
5,sanam,50000
6,opra,60000
7,yella,70000
```

Step 5: go to Hue and see the data is successfully import to HDFS


Query
Search data and saved documents...
Jobs
cloudera
acad
_SUCCESS
part-m-00000
View as binary
Edit file
Download
View file location
Refresh
Last modified
06/23/2020 12:06 AM
User
cloudera
Group
cloudera
Size
41 B
Mode
100644
HomePage
1
to
1
of 1
/ user / cloudera / acad / part-m-00000
5,sanam,50000
6,opra,60000
7,yella,70000

PART2 - Create Hive Tables through HQL Script, Use Sqoop to import and export tables to Relational Databases

Step 1: Create a hive table

```
hive> CREATE TABLE employees(name STRING, salary FLOAT, subordinates array<strin
g>, deductions map<string,float>, address struct<street:string,city:string,state
:string,zip:int>) row format delimited fields terminated by ',' stored as textfi
le;
OK
Time taken: 4.406 seconds
```

Step 2: verify table and load data into path

```
hive> show tables;
OK
employees
movies
olympic
petrol
ratings
users
x
Time taken: 0.021 seconds, Fetched: 7 row(s)
hive> describe employees;
OK
name                    string
salary                  float
subordinates            array<string>
deductions              map<string,float>
address                 struct<street:string,city:string,state:string,zip:int>

Time taken: 0.198 seconds, Fetched: 5 row(s)
hive> LOAD DATA INPATH 'acad/' INTO TABLE employees;
Loading data to table default.employees
Table default.employees stats: [numFiles=1, totalSize=41]
OK
Time taken: 1.152 seconds
```

Step 3: Verify table in warehouse

```
[cloudera@quickstart Downloads]$ hadoop fs -ls /user/hive/warehouse/
Found 7 items
drwxrwxrwx   - cloudera supergroup          0 2020-06-22 17:30 /user/hive/wareho
use/employees
drwxrwxrwx   - cloudera supergroup          0 2020-06-17 17:14 /user/hive/wareho
use/movies
drwxrwxrwx   - cloudera supergroup          0 2020-06-17 16:28 /user/hive/wareho
use/olympic
drwxrwxrwx   - root     supergroup          0 2020-06-17 16:13 /user/hive/wareho
use/petrol
drwxrwxrwx   - cloudera supergroup          0 2020-06-17 17:11 /user/hive/wareho
use/ratings
drwxrwxrwx   - cloudera supergroup          0 2020-06-17 17:10 /user/hive/wareho
use/users
drwxrwxrwx   - root     supergroup          0 2020-06-17 16:02 /user/hive/wareho
use/x
```

Step 4: Create an empty table in mySQL:

```
mysql> create table empNew(empid INT, emp_name VARCHAR(100));
Query OK, 0 rows affected (0.02 sec)
```

Step 5: Export sqoop from hive into mysql

```
[cloudera@quickstart Downloads]$ sqoop export --connect jdbc:mysql://localhost/d
b2 --username root --password cloudera --table empNew --export-dir /user/hive/wa
rehouse/employees -m 1
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
20/06/22 17:57:27 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5.13.0
20/06/22 17:57:27 WARN tool.BaseSqoopTool: Setting your password on the command-
line is insecure. Consider using -P instead.
20/06/22 17:57:27 INFO manager.MySQLManager: Preparing to use a MySQL streaming
resultset.
20/06/22 17:57:27 INFO tool.CodeGenTool: Beginning code generation
20/06/22 17:57:29 INFO manager.SqlManager: Executing SQL statement: SELECT t.* F
ROM `empNew` AS t LIMIT 1
20/06/22 17:57:29 INFO manager.SqlManager: Executing SQL statement: SELECT t.* F
ROM `empNew` AS t LIMIT 1
20/06/22 17:57:29 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/ha
doop-mapreduce
Note: /tmp/sqoop-cloudera/compile/7b873611c21fd271a3d85be86ab14d47/empNew.java u
ses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
20/06/22 17:57:36 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-clou
dera/compile/7b873611c21fd271a3d85be86ab14d47/empNew.jar
20/06/22 17:57:37 INFO mapreduce.ExportJobBase: Beginning export of empNew
20/06/22 17:57:37 INFO Configuration.deprecation: mapred.job.tracker is deprecat
```

```
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=23902
                Total vcore-milliseconds taken by all map tasks=23902
                Total megabyte-milliseconds taken by all map tasks=24475648
        Map-Reduce Framework
                Map input records=3
                Map output records=3
                Input split bytes=155
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=533
                CPU time spent (ms)=1440
                Physical memory (bytes) snapshot=111104000
                Virtual memory (bytes) snapshot=1508032512
                Total committed heap usage (bytes)=60751872
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=0
20/06/22 17:59:00 INFO mapreduce.ExportJobBase: Transferred 199 bytes in 77.444
seconds (2.5696 bytes/sec)
20/06/22 17:59:00 INFO mapreduce.ExportJobBase: Exported 3 records.
```

Step 6: The data is successfully export from Hive to mysql

```
mysql> select * from empNew;
+-------+----------+
| empid | emp_name |
+-------+----------+
|     5 | sanam    |
|     6 | opra     |
|     7 | yella    |
+-------+----------+
3 rows in set (0.01 sec)
```

PART 3: Perform three queries from databases

Step 1: Statistics

```
hive> analyze table employees compute statistics;
Query ID = cloudera_20200623201919_f08c1feb-6114-4580-ac3f-9e23904f72c8
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1592863501531_0003, Tracking URL = http://quickstart.cloudera
:8088/proxy/application_1592863501531_0003/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1592863501531_0003
Hadoop job information for Stage-0: number of mappers: 1; number of reducers: 0
2020-06-23 20:19:46,026 Stage-0 map = 0%,  reduce = 0%
2020-06-23 20:20:17,163 Stage-0 map = 100%,  reduce = 0%, Cumulative CPU 2.86 se
c
MapReduce Total cumulative CPU time: 2 seconds 860 msec
Ended Job = job_1592863501531_0003
Table default.employees stats: [numFiles=1, numRows=3, totalSize=41, rawDataSize
=38]
MapReduce Jobs Launched:
Stage-Stage-0: Map: 1   Cumulative CPU: 2.86 sec   HDFS Read: 3381 HDFS Write: 7
3 SUCCESS
Total MapReduce CPU Time Spent: 2 seconds 860 msec
OK
Time taken: 61.872 seconds
```

Step 2: Word Count

```
hive> select word,count(1) as count from(select explode(split(name,'//s')) as wo
rd from employees) temptable group by word;
Query ID = cloudera_20200623202222_8eb506e9-c4e4-4551-8167-5a3a56738b2f
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1592863501531_0004, Tracking URL = http://quickstart.cloudera
:8088/proxy/application_1592863501531_0004/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1592863501531_0004
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2020-06-23 20:22:32,563 Stage-1 map = 0%,  reduce = 0%
2020-06-23 20:22:59,063 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.0 sec
2020-06-23 20:23:23,975 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 5.68
sec
MapReduce Total cumulative CPU time: 5 seconds 680 msec
Ended Job = job_1592863501531_0004
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 5.68 sec   HDFS Read: 9633 HD
FS Write: 12 SUCCESS
Total MapReduce CPU Time Spent: 5 seconds 680 msec
OK
5       1
6       1
7       1
Time taken: 86.494 seconds, Fetched: 3 row(s)
```

Step 3: Identifying pattern

```
hive> select * from employees where salary = 60000 OR name LIKE 'C%';
OK
Time taken: 0.23 seconds
```