Session 11

Assignment 2

|  |  |
| --- | --- |
| **Prepared For:** | AcadGild |
|  |  |
| **Document Approval:** | **AcadGild** |
|  |  |
|  |  |
|  |  |
|  |  |
| **Project Title:** | Session 11 – Assignment 2 |
|  |  |
| **Prepared By:** | Duncan Burgess |
|  |  |
|  | dburgess@duncb.com |
|  |  |
| **Primary Engineer:** | Duncan Burgess |
|  |  |
| **Document Reference:** | **Session 10 – Assignment 2** |
|  |  |
| **Start Date:** | 06/10/2017 |
|  |  |
|  |  |



# 

# Contents

[Contents 2](#_Toc495041775)

[Change History 3](#_Toc495041776)

[1. Problem Statement 4](#_Toc495041777)

[2. Dataset 5](#_Toc495041778)

[3. Solution 5](#_Toc495041779)

[3.1. Import into Hive table 5](#_Toc495041780)

[3.2. Performing Incremental Load 6](#_Toc495041781)

# Change History

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Document Revision** | **Date** | **Authored By** | **Authorised By** | **Sections Affected** | **Reason for Change** |
| Rev 01 | 06/10/2017 | Duncan Burgess |  | All | Initial release. |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

# Problem Statement

Perform incremental load in Hive

* Read from MySQL Table and load it in Hive table.
* Create hive table if it does not exist.
* If it exists, perform the incremental load.

# Dataset

The **optumstaff** table will be used

*mysql> select \* from optumstaff;*

*+--------+---------+-----------+-----------+------+----------------+------+------------+*

*| emp\_ID | company | firstname | lastname | age | city | jobs | skill |*

*+--------+---------+-----------+-----------+------+----------------+------+------------+*

*| 1 | Optumuk | Duncan | Burgess | 52 | Mawsley | 89 | windows |*

*| 2 | Optumuk | Mike | Tartaglia | 32 | Burton Latimer | 78 | windows |*

*| 3 | Optumuk | Ben | Ludford | 22 | Swindon | 90 | windows |*

*| 4 | Optumuk | Tracey | Baker | 21 | Birmingham | 93 | management |*

*| 5 | Optumuk | Chris | Ginn | 23 | Cardif | 14 | SQL |*

*| 6 | Optumuk | Chris | Reid | 23 | Leek | 87 | linux |*

*+--------+---------+-----------+-----------+------+----------------+------+------------+*

# Solution

Solutions below.

## Import into Hive table

Here we are using the Sqoop **import** command with the following parameters:

Connecting to the database **sqoop\_db** in the MySQL server using the **JDBC Driver**

Entering the system username (**root**) & password–P

Using 1 mapper

4 columns are selected from the MySQL table: ID, firstname, lastname and jobs to be imported into the Hive Table Specifying the MySQL table name.

Using hive import command, we create the hive table to load the data in in the query itself, named **optum\_incremental** and storing it in the **default** database.

**Command run**

*[cloudera@quickstart ~]$ sqoop import --connect jdbc:mysql://localhost/sqoop\_db --username 'root' -P --split-by emp\_ID --columns emp\_ID,firstname,lastname,jobs -input-fields-terminated-by ',' --table 'optumstaff' --target-dir '/sqoop\_hive' --hive-import --fields-terminated-by ',' --create-hive-table --hive-table default.optum\_incremental -m 1;*

**Results**

*17/10/04 20:50:42 INFO mapreduce.Job: Job job\_1507104061255\_0024 completed successfully*

*17/10/04 20:50:42 INFO mapreduce.Job: Counters: 30*

*File System Counters*

*FILE: Number of bytes read=0*

*FILE: Number of bytes written=152013*

*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=107*

*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)=13465*

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

*Total time spent by all map tasks (ms)=13465*

*Total vcore-milliseconds taken by all map tasks=13465*

*Total megabyte-milliseconds taken by all map tasks=13788160*

*Map-Reduce Framework*

*Map input records=6*

*Map output records=6*

*Input split bytes=87*

*Spilled Records=0*

*Failed Shuffles=0*

*Merged Map outputs=0*

*GC time elapsed (ms)=121*

*CPU time spent (ms)=1380*

*Physical memory (bytes) snapshot=123838464*

*Virtual memory (bytes) snapshot=1510166528*

*Total committed heap usage (bytes)=60882944*

*File Input Format Counters*

*Bytes Read=0*

*File Output Format Counters*

*Bytes Written=107*

*17/10/04 20:50:42 INFO mapreduce.ImportJobBase: Transferred 107 bytes in 38.3011 seconds (2.7937 bytes/sec)*

*17/10/04 20:50:42 INFO mapreduce.ImportJobBase: Retrieved 6 records.*

*17/10/04 20:50:42 INFO manager.SqlManager: Executing SQL statement: SELECT t.\* FROM `optumstaff` AS t LIMIT 1*

*17/10/04 20:50:42 INFO hive.HiveImport: Loading uploaded data into Hive*

*Logging initialized using configuration in jar:file:/usr/lib/hive/lib/hive-common-1.1.0-cdh5.12.0.jar!/hive-log4j.properties*

*OK*

*Time taken: 4.179 seconds*

*Loading data to table default.optum\_incremental*

*Table default.optum\_incremental stats: [numFiles=1, totalSize=107]*

*OK*

*Time taken: 0.841 seconds*

* Run hive to get hive shell:
* Use the SHOW TABLES; command to see if the table has indeed been created
* Then use the SELECT command to retrieve all the rows from the table
* The hive table has been successfully created too.

hive> select \* from optum\_incremental;

OK

1 Duncan Burgess 89

2 Mike Tartaglia 78

3 Ben Ludford 90

4 Tracey Baker 93

5 Chris Ginn 14

6 Chris Reid 87

Time taken: 0.891 seconds, Fetched: 6 row(s)

hive>

## Performing Incremental Load

Now, to show the feature of Incremental load in Sqoop, We have to insert more records into the MySQL table optumstaff as shown below:

*mysql> use sqoop\_db;*

*Database changed*

*mysql> insert into optumstaff values (7,'Optumuk','Stuart','Ford',24,'Edingborough',83,'linux');*

*Query OK, 1 row affected (0.01 sec)*

*mysql> insert into optumstaff values (8,'Optumuk','Shaun','Turner',38,'Leicester',67,'SQL');*

*Query OK, 1 row affected (0.01 sec)*

mysql> insert into optumstaff values (9,'Optumuk','Rosario','Marino',24,'Oxford',72,'windows');

Query OK, 1 row affected (0.03 sec)

After we’ve successfully added the additional rows of data to the MySQL table. We again transfer these new rows to the Hive table **optum\_incremental**, by using SQOOP

Here we are using the Sqoop **import** command with the following parameters:

Connecting to the database **sqoop\_db** in the MySQL server using the **JDBC Driver**

Entering the system username (**root**) & password (**cloudera**) –P

Using 1 mapper

I have only selected all the columns rather than the original 4 to test the increment fully be imported into the Hive Table to be created.

Using the parameter **--incremental append** to show that this is an append call

Using the parameter **--check emp\_ID** to specify the column to look at for the append (emp\_ID)

**Command run**

[cloudera@quickstart ~]$ sqoop import --connect jdbc:mysql://localhost/sqoop\_db --username 'root' -P --table 'optumstaff' --target-dir '/sqoop\_hive' --incremental append --check-column emp\_ID -m 1;

**Results**

*17/10/04 21:23:32 INFO mapreduce.ImportJobBase: Transferred 408 bytes in 39.1086 seconds (10.4325 bytes/sec)*

*17/10/04 21:23:32 INFO mapreduce.ImportJobBase: Retrieved 9 records.*

*17/10/04 21:23:32 INFO util.AppendUtils: Creating missing output directory - sqoop\_hive*

*17/10/04 21:23:32 INFO tool.ImportTool: Incremental import complete! To run another incremental import of all data following this import, supply the following arguments:*

*17/10/04 21:23:32 INFO tool.ImportTool: --incremental append*

*17/10/04 21:23:32 INFO tool.ImportTool: --check-column emp\_ID*

*17/10/04 21:23:32 INFO tool.ImportTool: --last-value 9*

*17/10/04 21:23:32 INFO tool.ImportTool: (Consider saving this with 'sqoop job –create')*

**Check HDFS and map file**

*[cloudera@quickstart ~]$ hadoop fs -ls /sqoop\_hive*

*Found 1 items*

*-rw-r--r-- 1 cloudera cloudera 408 2017-10-04 21:23 /sqoop\_hive/part-m-00000*

*[cloudera@quickstart ~]$ hadoop fs -cat /sqoop\_hive/p\**

[cloudera@quickstart ~]$ hadoop fs -cat /sqoop\_hive/p\*

1,Optumuk,Duncan,Burgess,52,Mawsley,89,windows

2,Optumuk,Mike,Tartaglia,32,Burton Latimer,78,windows

3,Optumuk,Ben,Ludford,22,Swindon,90,windows

4,Optumuk,Tracey,Baker,21,Birmingham,93,management

5,Optumuk,Chris,Ginn,23,Cardif,14,SQL

6,Optumuk,Chris,Reid,23,Leek,87,linux

7,Optumuk,Stuart,Ford,24,Edingborough,83,linux

8,Optumuk,Shaun,Turner,38,Leicester,67,SQL

9,Optumuk,Rosario,Marino,24,Oxford,72,windows

[cloudera@quickstart ~]$

**Note** missing rows and column information updated