Import MySQL Data Using Sqoop

Exercise Path: ~/workspace/sqoop

MySQL database: lanier

MySQL tables: accounts

webpage

HDFS paths: lanier/accounts

lanier/webpage

**In this exercise, you will import tables from MySQL into HDFS using Sqoop.**

**Importing a Table From MySQL to HDFS**

You can use Sqoop to look at the table layout in MySQL. With Sqoop, you can also

import the table from MySQL to HDFS.

1. Open a new terminal window.
2. Run the sqoop help command to familiarize yourself with the options in Sqoop:

$ sqoop help

1. List the tables in the lanier database:

$ sqoop list-tables

--connect jdbc:mysql://localhost/lanier

--username [username] --password [password]

1. Run the sqoop import command to see its options:

$ sqoop import --help

1. Use Sqoop to import the accounts table in the lanier database and save it in HDFS under lanier:

$ sqoop import

--connect jdbc:mysql://localhost/lanier

--username [username] --password [password]

--table accounts

--target-dir lanier/accounts

--null-non-string '\\N'

The --null-non-string option tells Sqoop to represent null values as \N,

which makes the imported data compatible with Hive.

1. Optional: While the Sqoop job is running, try viewing it in the Hue Job Browser or YARN Web UI, as you did in the previous exercise.

**Viewing the Imported Data**

Sqoop imports the contents of the specified tables to HDFS. You can use the command line or the Hue File Browser to view the files and their contents.

7. List the contents of the accounts directory:

* + hdfs dfs -ls lanier/accounts

NOTE: Output of Hadoop processing jobs is saved as one or more numbered“partition” files. Partitions are covered later in the course.

1. Use either the Hue File Browser or the -tail option to the hdfs command to view the last part of the file for each of the MapReduce partition files, for example:

$ hdfs dfs -tail lanier/accounts/part-m-00000

$ hdfs dfs -tail lanier/accounts/part-m-00001

$ hdfs dfs -tail lanier/accounts/part-m-00002

$ hdfs dfs -tail lanier/accounts/part-m-00003

1. The first six digits in the output are the account ID. From the last line of the last file, take note of highest account ID (Account ID *129764*) because you will use it in the next step.

**Importing Incremental Updates**

As Lanier adds new accounts, the account data in HDFS must be updated as accounts are created. You can use Sqoop to append these new records.

Run the add\_new\_accounts.py script to add the latest accounts to MySQL.

$ python workspace/sqoop/ add\_new\_accounts.py

1. Incrementally import and append the newly added accounts to the accounts directory. Use Sqoop to import on the last value on the acct\_num column largest account ID:

$ sqoop import

--connect jdbc:mysql://localhost/lanier --username [username] --password [password]

--incremental append

--null-non-string '\\N'

--table accounts

--target-dir lanier/accounts --check-column acct\_num

--last-value <largest\_acct\_num>

NOTE: Refer to question #9 for the largest\_acct\_num ID number.

1. List the contents of the accounts directory to verify the Sqoop import: $ hdfs dfs -ls lanier/accounts
2. You should see three new files. Use Hadoop’s cat command to view the entire contents of these files.

$ hdfs dfs -cat lanier/accounts/part-m-0000[456]

**Importing Data Using an Alternate Field Delimiter**

1. We also want to import the webpage table to HDFS, but first, use the sqoop eval command to look at a few records in that table:

$ sqoop eval

--query "SELECT \* FROM webpage LIMIT 10"

--connect jdbc:mysql://localhost/lanier

--username [username] --password [password]

Notice that the values in the last column contain commas. By default, sqoop uses commas as field separators, but because the data itself uses commas, we cannot do that this time.

1. Use Sqoop to import the webpage table, but use the tab character (\t) instead of the default (comma) as the field terminator.

$ sqoop import

--connect jdbc:mysql://localhost/lanier

--username [username] --password [password]

--table webpage

--target-dir lanier/webpage

--fields-terminated-by "\t"

1. Using Hue or the hdfs command line utility, review the data files imported to the lanier/webpage directory. Take note of the structure of the data; you will use this data in the next exercises.

**END**