Homework: Import Data from MySQL Using Sqoop

Files and Data Used in this Homework

Exercise directory: `$DEV1/exercises/sqoop`

MySQL Database: `loudacre`

MySQL Tables: `accounts`, `webpage`

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

Import the accounts table from MySQL

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 if necessary.
2. Run the `sqoop help` command to familiarize yourself with the options in Sqoop:

```
$ sqoop help
```

3. List the tables in the `loudacre` database:

```
$ sqoop list-tables \
--connect jdbc:mysql://localhost/loudacre \
--username training --password training
```

4. Run the `sqoop import` command to see its options:

```
$ sqoop import --help
```

5. Use Sqoop to import the `accounts` table in the `loudacre` database and save it in HDFS under `/loudacre`:

```
$ sqoop import \
  --connect jdbc:mysql://localhost/loudacre \
  --username training --password training \
  --table accounts \
  --target-dir /loudacre/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 and Impala.

6. *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.

View the imported data

Sqoop imports the contents of the specified tables to HDFS. You can use the hdfs 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 /loudacre/accounts
```

- **Note**: Output of Hadoop processing jobs is saved as one or more numbered "partition" files.

8. Use either the Hue File Browser or the HDFS `tail` command to view the last part of the file for each of the MapReduce partition files, e.g.:

```
$ hdfs dfs -tail /loudacre/accounts/part-m-00000
$ hdfs dfs -tail /loudacre/accounts/part-m-00001
$ hdfs dfs -tail /loudacre/accounts/part-m-00002
$ hdfs dfs -tail /loudacre/accounts/part-m-00003
```

9. The first six digits in the output are the account ID. Take note of highest account ID because you will use it in the next step.

Import incremental updates to accounts

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

10. Run the `add_new_accounts.py` script to add the latest accounts to MySQL.

```
$ $DEV1/exercises/sqoop/add_new_accounts.py
```

11. 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:

    Note: Use --last-value *<largest_acct_num>* - replace <largest_act_num> with the largest account number you found earlier.

12. List the contents of the `accounts` directory to verify the Sqoop import:

```
$ hdfs dfs -ls /loudacre/accounts
```

13. You should see three new files. Use Hadoop's `cat` command to view the entire contents of these files.

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

Import webpage data using an alternate field delimiter

14. We also want to import the webpage table to HDFS. But first look at a few records in that table using the `sqoop eval` command.

```
$ sqoop eval \
  --query "SELECT * FROM webpage LIMIT 10" \
  --connect jdbc:mysql://localhost/loudacre \
  --username training --password training
```

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 can't do that this time.

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

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

Optional Homework

Use Sqoop to import only accounts where the person lives in California (state = "CA") and has an active account (acct_close_dt IS NULL).

Note: This optional homework must output to a different HDFS directory. Otherwise, the original table will be overwritten and you will need to delete the directory and import the table again.