Ask Bigger Questions

CAP - Developing with Spark and Hadoop:

Homework Assignment Guide for Students

Homework: Select a Format for a Data File

Files and Data Used in this Homework

Exercise directory: `$DEV1/exercises/data-format/`

MySQL Database: `loudacre`

MySQL Table: `accounts`

In this exercise, you will use import data in Avro format and create an Impala/Hive table to access it.

1. Change directories to the exercise directory:

```
cd $DEV1/exercises/data-format/
```

2. Import the `accounts` table to an Avro data format.

```
$ sqoop import \
  --connect jdbc:mysql://localhost/loudacre \
  --username training --password training \
  --table accounts \
  --target-dir /loudacre/accounts_avro \
  --null-non-string '\\N' \
  --as-avrodatafile
```

3. View the files imported by Sqoop into HDFS.   What do you see when you try to view the content of the data files?

4. Sqoop generated a schema named `sqoop_import_accounts.avsc` in the current directory. Review this file and then copy it to the `/loudacre` directory in HDFS.

5. In Impala or Hive, create a table using this schema:

```
CREATE EXTERNAL TABLE accounts_avro
STORED AS AVRO
LOCATION '/loudacre/accounts_avro'
TBLPROPERTIES ('avro.schema.url'=
'hdfs:/loudacre/sqoop_import_accounts.avsc');
```

6. Confirm correct creation of the table by issuing a query such as

```
SELECT * FROM accounts_avro LIMIT 10
```

7. Optional: Use the `DESCRIBE` or `DESCRIBE FORMATTED` command to list the columns and data types of the `accounts_avro` table created from the Avro schema.

Optional Homework

8. Create a new table based on the existing `accounts_avro` table data, using Parquet for the storage format.

This is the end of the Homework