Ask Bigger Questions

CAP - Developing with Spark and Hadoop:

Homework Assignment Guide for Students

© Copyright 2010-2015 Cloudera. All rights reserved. Not to be reproduced or shared without prior written consent from Cloudera.

Homework: Partition Data in Impala or Hive

Files and Data Used in this Homework

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

Data files (HDFS): `/loudacre/accounts_avro`

In this exercise you will create and load an Impala/Hive table with account data, partitioned by area code.

In the previous exercise you imported data from the accounts table using Sqoop, into a table called `accounts_avro`. In this exercise, you will create a new table with some of the account data, partitioned by area code (the first three digits of the phone number).

1. Create a new, empty table in Impala or Hive:

```
CREATE EXTERNAL TABLE accounts_by_areacode (
    acct_num INT,
    first_name STRING,
    last_name STRING,
    phone_number STRING)
  PARTITIONED BY (areacode STRING)
  ROW FORMAT DELIMITED
  FIELDS TERMINATED BY ','
  LOCATION '/loudacre/accounts_by_areacode';
```

© Copyright 2010-2015 Cloudera. All rights reserved. Not to be reproduced or shared without prior written consent from Cloudera.

2. In order to populate the new table, you will need to extract the area code from the phone number. Try executing the following query to demonstrate:

```
SELECT acct_num, first_name, last_name,
   phone_number, SUBSTR(phone_number,1,3) AS areacode
  FROM accounts_avro
```

3. Use the SELECT statement above in an INSERT INTO TABLE command to copy the specified columns to the new table, dynamically partitioning by area code.

4. Execute a simple query to confirm that the table was populated correctly, such as

```
SELECT * FROM accounts_by_areacode LIMIT 10
```

5. Using Hue or the hdfs command line interface, confirm that the directory structure of the accounts_by_areacode table includes partition directories. Review the data in the directories to verify that the partitioning is correct.

This is the end of the Homework

© Copyright 2010-2015 Cloudera. All rights reserved. Not to be reproduced or shared without prior written consent from Cloudera.