Data Application Lab

Sunday Hive Quiz

1. Explain what is Hive?

Hive is an ETL and Data warehousing tool developed on top of Hadoop Distributed File System (HDFS). It is a data warehouse framework for querying and analysis of data that is stored in HDFS. Hive is an open-source-software that lets programmers analyze large data sets on Hadoop.

2. Mention key components of Hive Architecture?

Key components of Hive Architecture include:

- User Interface
- Compiler
- Metastore
- Driver
- Execute Engine

3. When do you choose "Internal Table" and "External Table" in Hive?

In Hive you can choose internal table,

- If the processing data available in local file system
- If we want Hive to manage the complete lifecycle of data including the deletion

You can choose External table,

- If processing data available in HDFS
- Useful when the files are being used outside of Hive

4. Why do we need buckets?

There are two main reasons for performing bucketing to a partition:

- A map side join requires the data belonging to a unique join key to be present in the same partition. But what about those cases where your partition key differs from that of join key? Therefore, in these cases you can perform a map side join by bucketing the table using the join key.
- Bucketing makes the sampling process more efficient and therefore, allows us to decrease the query time.

5. What is Map and Reduce?

Map takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs).

Reduce takes the output from a map as an input and combines those data tuples into a smaller set of tuples. After processing, it produces a new set of output, which will be stored in the HDFS.

6. Hive demo Practice

Step 1. create two text files in local system

```sh
mkdir -p /home/chenguang/demo/employee/
vi /home/chenguang/demo/employee/employee.txt
vi /home/chenguang/demo/employee/salaries.txt
```

Step 2. write below rows into the files separately

```sh
1,Jacky,Chan,40,China
2,Justin,Bibber,24,Canada
3,Kendrick,Lamar,30,United States
4,Lionel,Messi,30,Argentina
```

```sh
1,Jacky,Chan,60117
2,Justin,Bibber,62102
3,Kendrick,Lamar,66245
4,Lionel,Messi,37456
5 Jasper, Liao, 34142
```

Step 3. Create an empty table

```sql
CREATE TABLE employee (id INT, first_name STRING, last_name STRING, age INT, country STRING) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',';
```

```sql
CREATE TABLE salaries (id INT, first_name STRING, last_name STRING, salary INT) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',';
```

Step 4. Load the text file into this table

```sql
LOAD DATA LOCAL INPATH '/home/chenguang/demo/employee/salaries.txt' OVER-WRITE INTO TABLE salaries;

LOAD DATA LOCAL INPATH '/home/chenguang/demo/employee/employee.txt' OVERWRITE INTO TABLE employee;
```

Step 5. Query this table to validate it

SELECT * FROM employee;

Step 6. Inner join on id, and Left outer join table salaries (use salaries as left table) on id. And see the differences.

SELECT * FROM employee INNER JOIN salaries ON employee.id = salaries.id;
SELECT * FROM salaries LEFT OUTER JOIN employee ON employee.id = salaries.id;

****NOTE: student to replace 'chenguang' in paths with your own username**