Hive Lab Questions

Christopher Keinsley

Question 1 (4 points): Hive supports the MSCK repair command. What does it do? Explain with an example. (2 points for the explanation, 2 points for the example)

The MSCK repair command can repair the metadata in the metastore for a table. This is useful when we place files directly into hive's folders instead of making hive do the loading. This command can be used by msck repair <tableName>;

Question 2 (4 points): Explain the difference between order by, sort by commands with an example. (2 Points for the explanation, 2 points for the example)

Order by guarantees a total sort by shoving every row through one reducer. Sort by guarantees order only within each reducer. If there are more than 1 reducers, sort by may not return a total sort.

- Select * from Table order by colName
    - May return something like
        - 0
        - 1
        - 2
        - 3
- Select * from Table sort by colName
    - May return something like depending on the reducers
        - 0
        - 3
        - 1
        - 2

Question 3 (4 Points): Explain the purpose of the distribute by command (2 Points).

- Cluster by partitions the data by whatever column you give it when running map reduce jobs. This can be used to force the data to go to the same reducers. This is useful when running sorts, so that data can be sorted within a particular partition instead of sending all of the data through one reducer.

What happens when you add a sort by to the output of a distribute by? (1 Point)

- You receive data that is partitioned on the distribute by column and sorted within each column.

Is there an equivalent command that replaces the distribute by and sort by? What is it? (1 Point).

- Yes; cluster by.

Question 4 (8Points). Explain the following commands/concepts with an example. (2 points for explanation, 2 points for example)

a. Bucketing
    i. Bucketing is similar to partitioning on one field. Bucketing results in a fixed number of files because you can specify the number of buckets. Rows are then hashed on the bucketed field and placed in their respective buckets.
    ii. Create table bleh (field string) clustered by (field) into 200 buckets.
b. UNION ALL
    i. Union all combines all rows from different tables into one big table. This can be used to merge two tables with the same field types into one table.
    ii. Create Table combined as Select * from (select * from table1 UNION ALL select * from table2);