Zhihao Xue
CM 3357
Lab6
CSSE490 Dr. Mohan
Thur,Oct 22, 2015

- What does MSCK repair command do?
  Hive stores a list of partitions for each table in its metastore. If, however, new partitions are directly added to HDFS (say by using hadoop fs -put command), the metastore (and hence Hive) will not be aware of these partitions unless the user runs ALTER TABLE table]_name ADD PARTITION commands on each of the newly added partitions.

  ```
  MSCK REPAIR TABLE table_name;
  ```

- Explain the difference between order by, sort by commands with an example.
  Hive supports SORT BY which sorts the data per reducer. The difference between "order by" and "sort by" is that the former guarantees total order in the output while the latter only guarantees ordering of the rows within a reducer. If there are more than one reducer, "sort by" may give partially ordered final results.

  order by:guarantees global ordering, but does this by pushing all data through just one reducer. This is basically unacceptable for large datasets. You end up one sorted file as output.

  sort by: orders data at each of N reducers, but each reducer can receive overlapping ranges of data. You end up with N or more sorted files with overlapping ranges.

  ```
  SELECT time, user_id, action FROM user_action_table ORDER BY user_id,
      time
  SELECT time, user_id, action FROM user_action_table
  DISTRIBUTE BY user_id SORT BY time
  ```

- Explain the purpose of the distribute by command.
  Hive uses the columns in Distribute By to distribute the rows among reducers. All rows with the same Distribute By columns will go to the same reducer. However, Distribute By does not guarantee clustering or sorting properties on the distributed keys.
  distribute by:ensures each of N reducers gets non-overlapping ranges of x, but doesn't sort the output of each reducer. You end up with N or unsorted files with non-overlapping ranges.
  CLUSTER BY is the replace.

- Explain command
  Bucketing: the bucketing specified at table creation is not enforced when the table is written to, and so it is possible for the table's metadata to advertise properties which are not upheld by the table's actual layout. This should obviously be avoided.

```
CREATE TABLE user_info_bucketed(user_id BIGINT, firstname STRING, lastname
    STRING)
COMMENT 'A bucketed copy of user_info'
PARTITIONED BY(ds STRING)
CLUSTERED BY(user_id) INTO 256 BUCKETS;
set hive.enforce.bucketing = true;
FROM user_id
INSERT OVERWRITE TABLE user_info_bucketed
PARTITION (ds='2009-02-25')
SELECT userid, firstname, lastname WHERE ds='2009-02-25';
```

UNION ALL:Hive versions prior to 1.2.0 only support UNION ALL (bag union), in which duplicate rows are not eliminated.In Hive 1.2.0 and later, the default behavior for UNION is that duplicate rows are removed from the result.With the optional ALL keyword, duplicate-row removal does not occur and the result includes all matching rows from all the SELECT statements.

```
SELECT *
FROM (
  select_statement
  UNION ALL
  select_statement
) unionResult
```