Chapter 6 – Partitioning and Bucketing

Solution

1. Examine the files action.txt, comedy.txt and thriller.txt

```
$ less action.txt
$ less comedy.txt
$ less thriller.txt
```

2. Create a table called movies_part with 4 columns (movieid, movie_name, release_date, imdb_url) that is partitioned on genre. Use the appropriate row delimiter.

```
hive> CREATE TABLE movies_part (movieid int, movie_name string,
release_date string, imdb_url string)
PARTITIONED BY (genre string)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ',';
```

3. Load each text file from step 1 into a partition. The partitions should be called "action", "comedy" and "thriller".

```
hive> LOAD DATA LOCAL INPATH '/tmp/action.txt'
INTO TABLE movies_part
PARTITION(genre='action');

hive> LOAD DATA LOCAL INPATH '/tmp/comedy.txt'
INTO TABLE movies_part
PARTITION(genre='comedy');

hive> LOAD DATA LOCAL INPATH '/tmp/thriller.txt'
INTO TABLE movies_part
PARTITION(genre='thriller');
```

4. Describe the structure of the table. Notice the genre is a derived column in the table.

```
hive> DESCRIBE movies_part;
```

5. List the partitions for the movies_part table.

```
hive> SHOW PARTITIONS movies_part;
```

6. Look at the Hive warehouse to see the 3 subdirectories

```
$ hadoop fs -ls /user/hive/warehouse/movies_part
```

7. Create a table called rating_buckets with the same column definitions as UserRatings, but with 8 buckets, clustered on movieid.

```
hive> CREATE TABLE rating_buckets (userid int, movieid int,
rating int, unixtime int)
CLUSTERED BY (movieid) INTO 8 BUCKETS;
```

8. Use INSERT OVERWRITE TABLE to load the rows in UserRatings into rating_buckets. Don't forget to set mapred.reduce.tasks to 8.

```
hive> SET mapred.reduce.tasks = 8;
hive> INSERT OVERWRITE TABLE rating_buckets SELECT * FROM
UserRatings CLUSTER BY movieid;
```

9. View the 8 files that were created. They should be roughly even in size.

```
$ hadoop fs -ls /user/hive/warehouse/rating_buckets
```

10. Count the rows in bucket 3 using `TABLESAMPLE`.

```
hive> SELECT count(1) FROM rating_buckets
TABLESAMPLE (BUCKET 3 OUT OF 8);
```