Chapter 6 – Partitioning and Bucketing

Exercise

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

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

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

5. List the partitions for the movies_part table.

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

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

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

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

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