Chapter 5 – Query Execution

Solutions

1. Use EXPLAIN to view the execution plan for this query:

```
hive> EXPLAIN SELECT count(1) FROM Movies
WHERE movie_name LIKE '%Star%';
```

2. Find the "Filter Operator". What is it doing?
   This is scanning the rows and looking for ones that contain the string "Star".

3. Find the "Group By Operator". What is it doing?
   It is calculating the count(1).

4. Use EXPLAIN to view the execution plan for the query in step 6 of the previous lab.

```
hive> EXPLAIN SELECT movie_name, avg(rating) AS avgrating
FROM movies JOIN userratings
ON (movies.movieid = userratings.movieid)
WHERE sci_fi= 1
GROUP BY movie_name
ORDER BY avgrating DESC LIMIT 1;
```

5. How many MapReduce phases are required?
   3; Stage-1, Stage-2 and Stage-3

6. For the first MapReduce (Stage-1), what is the “Filter Operation” doing on the movies table?
   The Filter Operation is scanning the movies table and looking for records that have “sci_fi=1”. This accomplishes the “WHERE” clause of the query.

7. What is the reduce doing in Stage-1?
   The join of the movies table to the userratings table on movieid.

8. What is the general purpose of the second MapReduce (Stage-2)?
   Group the rows by movie_name and calculate avg(rating)

9. How does the ORDER BY get done?
   The third MapReduce (stage-3) outputs the avg(rating) for each group as the key. The MapReduce framework in Hadoop automatically sorts the records by key before giving them to the reducer.