Project 3 : MapReduce

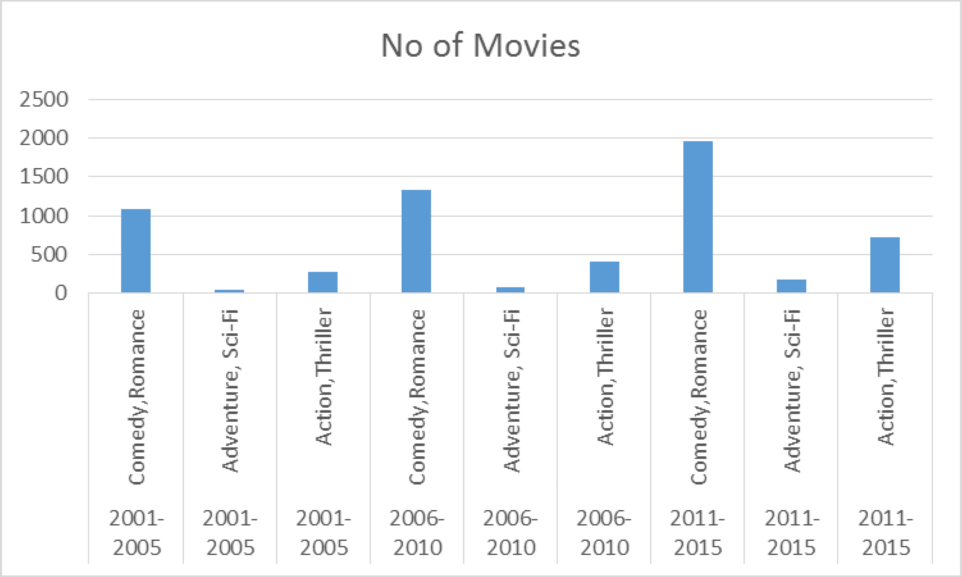
Overall Status

We have gone through the pdf file shared in the canvas. It gave us an insight of how map reduce and Hadoop process the data. We have downloaded the oracle virtual box, Ubuntu and Hadoop softwares which are required as part of this project. We have written the code by using the map reduce paradigm and we will get the output as the number of movies that are released of a particular genre combination for a particular pre defined period.

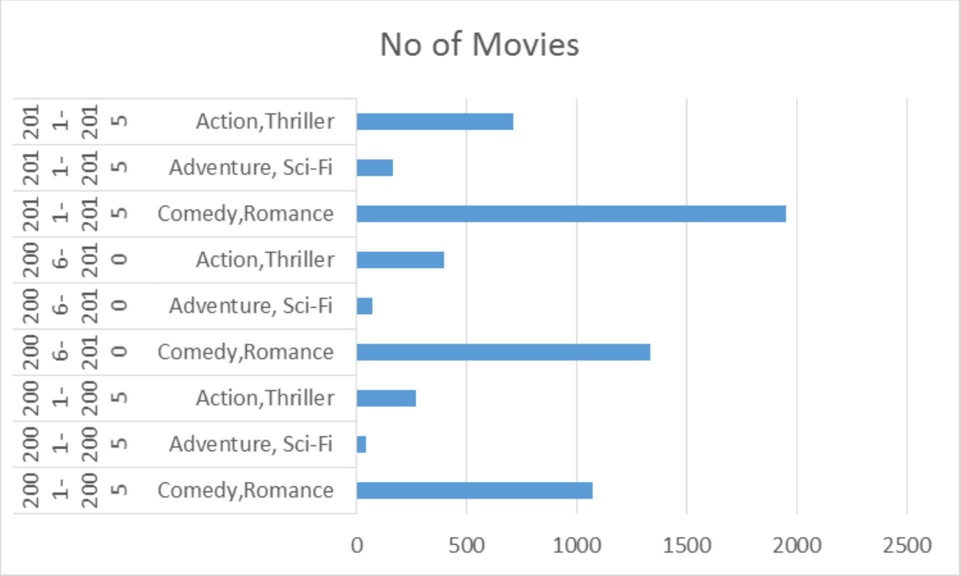
Analysis Results

Task1:MapReduce
From the results, the below pictures inferences that can explain the diagram. We can notice that Comedy and Romance has the highest amount of releases where as adventure and sci-fi has the lowest amount of releases and similarly other details can also be easily fetched just by looking at the picture below.

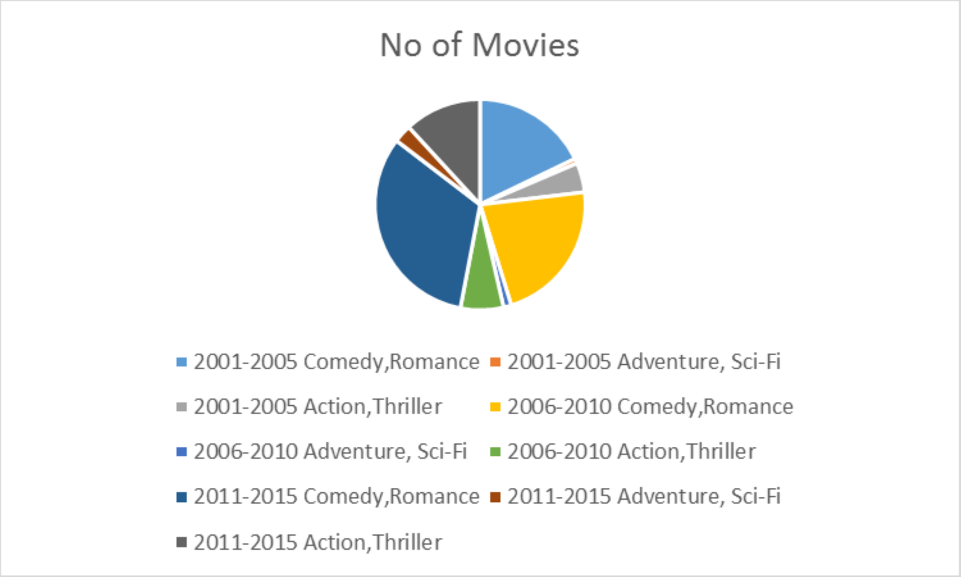
a)



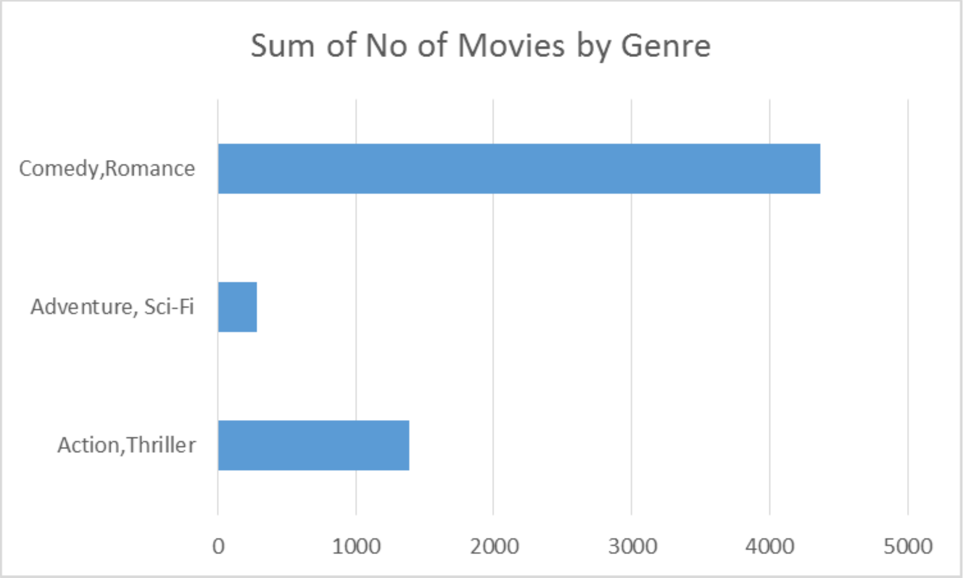
b)The bar graph below also depicts the inferences from the result set where we can actually get to know the number of movies that are released. We can notice from the figure that adventure and Sci-Fi has the least number of releases.



c)pie chart provides an overview of the results which is similar to that of bar graph but in a different pictorial representation. The diagram displays the information about the number of movies for a particular year and genre. The one which occupies more area in the below circle has the more value which means it has the highest number of releases.



d) The below picture depicts the total number of movies that are released for a given genre across a time period. Comedy,Romance genre has the highest releases.(same result as the above pictures has depicted)



Task 2 – SQL

For the sql part, the queries are executed in the oracle database and the results are captured in the spool file which have been uploaded as part of task 2. Please refer to it. With respect to the query plan generation, a tree structure of plan nodes representing the different actions is built. At first, the select statements are executed, then STOPKEY occurs as we have the condition rownum < 5,then the intermediate result set is ordered by the number of rows it has fetched, then a hash join operation is performed and then the table access is done by taking the attributes title_basics and title_ratings.

File Descriptions:

We just implemented two functions(mapper and reducer).Mapper was used to obtain the required data from the data set and we have created a key, value pair (genre combination) for processing the data and to send the intermediate output to the reducer.

Division of Labor:

We sat together in library and understood the concepts of Map Reduce. We have gone through the slides and the documents that are provided to us and cleared each other doubts through mutual help. Once when we made sure that we had the basic knowledge, We started working on the coding part.

Mahesh – Mapper and Task2

Harsha – Reducer and Task2

Error Handling:

a)Installation of Hadoop was a bit difficult as we have to go through the installation of the virtual machine, linux environment and a lot of commands for setting up the Hadoop environment. In doing this process, we have encountered few errors and we looked into the internet and resolved it.

b)There were few exceptions that popped up while executing the code and after carefully examining the coding part in mapper, we were able to resolve it

c) Identifying the key,value pairs was one of the vital and tough thing which defines the result and also the output of the reducer.