**Data Analysis Using Map/Reduce and Query Processing**

**Table of Contents**

Abstract………………………………………………………………………………2

Overall Status………………………………………………………………………3

Analysis…..…………………………………………………………………….3,4,5

File Descriptions………………………………………………………………….5

Division of Labour……………………………………………………………….5

Logical Errors………………………………………………………………………5

**Abstract**

In the first part, this project implements data analysis using Hadoop cluster. Here we have to generate the <key, Value> pair of year and genre from the given imdb dataset. Then it is reduced by adding the count of genre of specified year. In the second part, query is implemented to find top rated movies for one of the three periods and one of the genre.

**Overall Status**

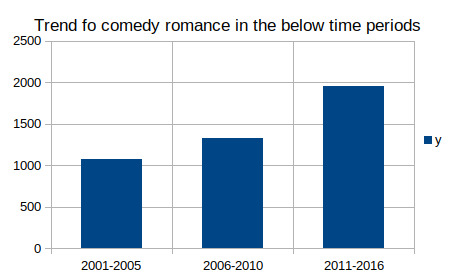
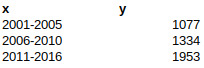
**Task-1**

* We have started with the installation of Hadoop and setup phase with the help of pdf provided.
* After the installation, we have successfully completed the execution of Word Count provided.
* Then we figured out to use three mappers for different years [2001-2005], [2006-2010], [2011-2015] and one reducer to count number of movies.

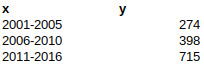
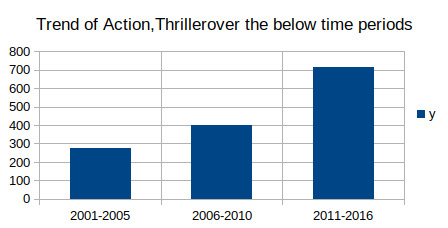
**Task-2**

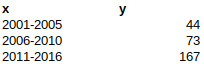
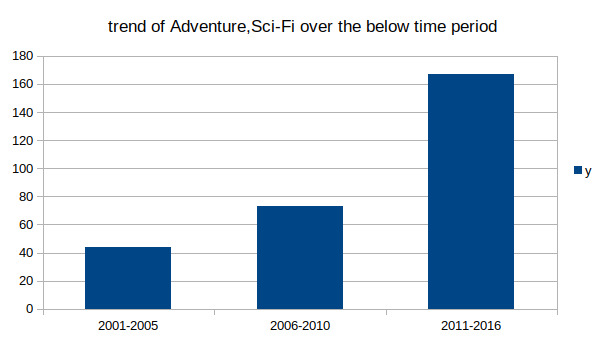
* We have started with getting the access to oracle account.
* Once we got access to oracle, started to explore individual tables provided in project description.
* Then implemented the query to find top 10 movies of one of the years and genres mentioned in task1.
* Then using EXPLAIN PLAN retrieved how the query is executed.

**Analysis Results**

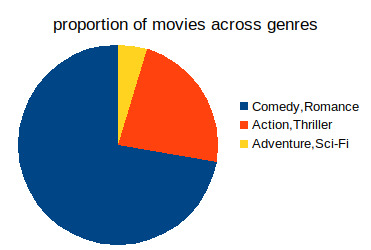
**Task-1**

There is growth in number of movies for genres comedy, Romance.

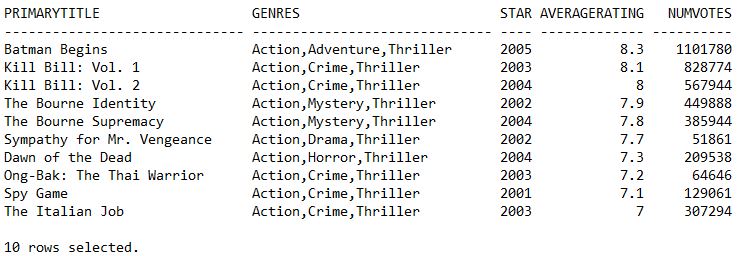
There is growth in number of movies for genres Action, Thriller.

****

There is growth in number of movies for genres Adventure, Sci-Fi.

****

**Task-2**



Query output for top 10 movies of genre (Action, Thriller) for year (2001-2005) which has number of votes > 50000.

**select PRIMARYTITLE, tr.AVERAGERATING, startyear, tr.numvotes, genres from sharmac.TITLE\_BASICS tb inner join sharmac.TITLE\_RATINGS tr on tb.TCONST=tr.TCONST where genres like '%Comedy%' and genres like '%Romance%' and ( STARTYEAR ='2001'OR STARTYEAR ='2002' OR STARTYEAR ='2003'OR STARTYEAR='2004' OR STARTYEAR ='2005') and titletype LIKE 'movie' and startyear NOT LIKE '\N' and numvotes >= 50000 order by averagerating desc fetch first 10 rows only;**

Explain plan for the above sql query -

(the output for the explain command is in the spoolOutput file)

• **ID 05**, TABLE ACCESS FULL scan of title\_ratings is done. Here, it also appplies the selection predicate numvotes>50000.

• **ID 06**, next ‘INDEX UNIQUE SCAN’ is performed.

• Both the outputs of 05 and 06 are passed to inner nested loop with operation **ID 04** which is then passed to **ID 03**

• **ID 07** is evaluated. TITLE\_BASICS table is accessed by TABLE ACCESS BY INDEX ROWID and selection predicate for start years and genre is evaluated here. The output of this is passed as well to ID 03 and joined there with the output of **ID 04** which was already passed

• Next ,a sort is performed in **ID 02** over averageRating.

• Next only the first 10 rows is picked.

**File Descriptions**

No new files are created as part of this project.

**Division of Labour**

Bhargavi (1001955763) and Santhosh (1001964729) implemented the code for Task1 Map/Reduce. Moreover, Avinash (1002023681) implemented the query for Task2.

**Logical Errors**

* Installing Hadoop and setting it up was bit difficult on windows. We were getting errors as Java home not found, later we reinstalled java and error was resolved.
* We were facing the below error

INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032

INFO ipc.Client: Retrying connect to server: 0.0.0.0/0.0.0.0:8032. Already tried 0 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)

Later we found that did not start the resource manager properly.

* Wrong arguments were passed so the output was overwritten and we got output for only one of the year.