# TITLE: FACEBOOK MUTUAL FRIENDS ANALYSIS USING MAP REDUCE

# TABLE OF CONTENTS

|  |  |  |
| --- | --- | --- |
|  | **CHAPTERS** | **PG.NO** |
| **1** | **Abstract** | **3** |
| **2** | **Discovery Phase:**  **2.a Problem statement/ Initial Hypothesis**  **2.b Literature Study**  **2.c S/w and H/w Requirements** | **4**  **4**  **5**  **10** |
| **3** | **Data Preparation Phase:**  **3.a Data set Description**  **3.b Data Pre-Processing** | **10**  **10**  **12** |
| **4** | **Model planning:**  **4.a Modules** | **13**  **13** |
| **5** | **Model Building Phase**  **5a. Implementation of Module 1 & 2**  **5b. Implementation of Module 3 & 4** | **15**  **16**  **18** |
| **6** | **Communicating Results Phase**  **6a. Output of Implementation of Module 1 & 2**  **6b. Output of Implementation of Module 3 & 4** | **19**  **20**  **20** |
| **7** | **Conclusion** | **21** |
| **8** | **References** | **22** |

# 1. Abstract

The best friendship recommendations often come from friends. The key idea is that if two people have a lot of mutual friends, but they are not friends, then the system should recommend them to be connected to each other. Let’s assume that the friendships are undirected: if A is a friend of B then B is also a friend of A. This is the most common friendship system used in Facebook, Google+, Linkedin, and several social networks. It’s not difficult to extend it to directed friendship system used in twitter;

In the current era of big data, high volumes of valuable data can be easily collected and generated. Social networks are examples of generating sources of this big data. Users in these social networks are often linked by some interdependency such as friendship. As these big social networks keep growing, there are situations in which an individual user wants to find popular groups of friends so that he can recommend the same groups to other users. At the initial stage, we are taking a list of person's and their friends list. The friend's list is sent to the record reader and the map stage sorts out the friends of two persons at an instance using the principle of one to many in alphabetical order. In the map output, the two friends are taken as a key and their mutual friends are retrieved as a value pair. Once the map phase is successfully completed, then in the sorting phase the friends are sorted in the alphabetical order. The output is displayed as an adjacency matrix or the list of friends and their mutual friends in the alphabetical order.

# DISCOVERY PHASE

# 2.a Problem Statement

Smartphones without social media usage in the daily lifestyle of people is unthinkable. That much effect has been created in the lifestyle of people by smartphone and social media. There are many social media such as Facebook, Twitter, etc., As per 2017 statistics, nearly 1.37 billion daily active users for Facebook. Every user contributes some type of data in structured or semi-structured or unstructured data format. Business owners utilize this data to understand customer needs and their behavior to make profit in their business. Facebook data analysis is the process of collecting, analyzing Facebook data and visualizing extracted results to the end user.

The user data is collected from Facebook based on their activities. User behavior, number of likes, number of posts, type of posts, their comments, etc. are stored by the database server. Comments by the user in unstructured formats, while other data in structured and semi-structured format. Petabytes of data is generated by Facebook users. So, Hadoop, MapReduce and related big data concepts are used in this project to analyze the data. Mutual friend analysis is one such field we are going to explore in this project by using MapReduce framework. As these big social networks keep growing, there are situations in which an individual user wants to find popular groups of friends so that he can recommend the same groups to other users. In this project, a big data analytic solution that uses the MapReduce model in mining these big social networks for discovering groups of frequently connected users for friend recommendation.

# 2.b Literature Survey

## Paper 1:

**Title:** Collecting Facebook data for big data research

**Authors:** Ioan Dragon ; Razvan Zota

## Methodology:

* This paper aims to provide insights on how to collect Facebook user data based on Facebook Graph API V2.0 or newer. A proof of concept application is developed that invites users to participate in the research and collect certain data. User's possible circle of friends is built, based on their interaction with others.
* The larger context of this research includes collecting relevant big data for a telecom business support system integration. The receiver of this research can be a telecom operator or a telecom services reseller.

**Published in:** [2017 16th RoEduNet Conference: Networking in Education and Research (RoEduNet)](https://ieeexplore.ieee.org/xpl/conhome/8116724/proceeding)

## Paper 2:

**Title:** Big data mining of social networks for friend recommendation

**Authors:** [Fan Jiang](https://ieeexplore.ieee.org/author/37395108700); [Carson K. Leung](https://ieeexplore.ieee.org/author/37279073000); [Adam G. M. Pazdor](https://ieeexplore.ieee.org/author/37085875927)

## Methodology:

* In the current era of big data, high volumes of valuable data can be easily collected and generated. Social networks are examples of generating sources of this big data. Users in these social networks are often linked by some interdependency such as friendship. As these big social networks keep growing, there are situations in which an individual user wants to find popular groups of friends so that he can recommend the same groups to other users. In this paper, a big data analytic solution that uses the MapReduce model in mining these big social networks for discovering groups of frequently connected users for friend recommendation. Evaluation results show the efficiency and practicality of our data analytic solution in mining big social networks and recommending friends.

**Published in:** [2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)](https://ieeexplore.ieee.org/xpl/conhome/7736513/proceeding)

## 

## Paper 3:

**Title:** MapReduce Functions to Analyse sentiment information from social Big data

**Authors:** llkyu ha; Bonghyun Back; Byoungchul Ahn

**Methodology:**

* Opinion mining, which extracts meaningful opinion information from large amounts of social multimedia data, has recently arisen as a research area. In particular, opinion mining has been used to understand the true meaning and intent of social networking site users. It requires efficient techniques to collect a large amount of social multimedia data and extract meaningful information from them. Therefore, in this paper, a method to extract sentiment information from various types of unstructured social media text data from social networks by using a parallel Hadoop Distributed File System (HDFS) to save social multimedia data and using MapReduce functions for sentiment analysis. The proposed method has stably performed data gathering and data loading and maintained stable load balancing of memory and CPU resources during data processing by the HDFS system. The proposed MapReduce functions have effectively performed sentiment analysis in the experiments. Finally, the sentiment analysis results of the proposed system are very close to those of manual processes.

## Paper 4:

**Title:** Social Media Big Data Analytics for Demand Forecasting: Development and Case Implementation of an Innovative Framework

**Authors:** Rehan Iftikhar ; Mohammad Saud Khan

## 

## Methodology:

* Social media big data offers insights that can be used to make predictions of products' future demand and add value to the supply chain performance. The paper presents a framework for improvement of demand forecasting in a supply chain using social media data from Twitter and Facebook.
* The proposed framework uses sentiment, trend, and word analysis results from social media big data in an extended Bass emotion model along with predictive modelling on historical sales data to predict product demand. The forecasting framework is validated through a case study in a retail supply chain. It is concluded that the proposed framework for forecasting has a positive effect on improving accuracy of demand forecasting in a supply chain.

**Published in:** [Journal of Global Information Management (JGIM)](https://www.igi-global.com/journal/journal-global-information-management/1070) 28(1)

## Paper 5:

**Title:** An Improved Algorithm for Mutual Friends Recommendation Application of SNS in Hadoop

**Authors:** Y.Lei

**Methodology:**

* In the Social network system, existing “people you might know” or “Mutual friends” recommending applications are commonly utilized to list two-hop mediate relationships that one could have with another in order to tighten the bonds among groups. However, as the number of SNS users increase dramatically, the relationship data gets so huge that the performance of the Mutual Friends recommendation system becomes an urgent problem considering the developers’ requirements. This paper proposes a sorting algorithm in Hadoop-a parallel computing framework, to enhance the efficiency of “Mutual friends” recommendation process by taking advantage of the novel map reduce model. In the revised application, the original sorting algorithm of intermediate data, merge sort, is replaced by a more time saving sorting approach which introduces a B-Tree like data structure, 2-3 Tree to store the user friendship data and conducts the sorting process. As the number of users increases, the revised user defined map reduce functions perform better than the conventional design in a time consuming aspect.

## Paper 6:

**Title:** Social Influence Analysis in Social Networking Big Data: Opportunities and Challenges

**Authors:** [Sancheng Peng](https://ieeexplore.ieee.org/author/37535929200); [Guojun Wang](https://ieeexplore.ieee.org/author/37086108662); [Dongqing Xie](https://ieeexplore.ieee.org/author/37086108535)

## Methodology:

* Social influence analysis has become one of the most important technologies in modern information and service industries. It will definitely become an essential mechanism to perform complex analysis in social networking big data. It is attracting an increasing amount of research ranging from popular topics extraction to social influence analysis, including analysis and processing of big data, social influence evaluation, influential users identification, and information diffusion modeling. A comprehensive investigation of social influence analysis, and the characteristics, architecture of social influence based on social networking big data is provided. The relationship between big data and social influence analysis is also discussed. In addition, research challenges relevant to real-world issues based on social networking big data in social influence analysis are discussed, focusing on research issues such as scalability, data collection, dynamic evolution, causal relationships, network heterogeneity, evaluation metrics, and effective mechanisms. The goal is to provide a broad research guideline of existing and ongoing efforts via social influence analysis in large-scale social networks, and to help researchers better understand the existing work, and design new algorithms and methods for social influence analysis.

**Published in:** [IEEE Network](https://ieeexplore.ieee.org/xpl/RecentIssue.jsp?punumber=65) ( Volume: 31, [Issue: 1](https://ieeexplore.ieee.org/xpl/tocresult.jsp?isnumber=7828253), January/February 2017)

## Paper 7:

**Title:** Big Data Techniques of Google, Amazon, Facebook and Twitter

**Authors:** [Thulara N. Hewage](https://www.researchgate.net/publication/scientific-contributions/Thulara-N-Hewage-2139626894) ; [Malka N. Halgamuge](https://www.researchgate.net/publication/profile/Malka-Halgamuge); [Ali Syed](https://www.researchgate.net/publication/scientific-contributions/Ali-Syed-2139635773); [Gullu Ekici](https://www.researchgate.net/publication/profile/Gullu-Ekici)

## Methodology:

* Amazon, Facebook and Twitter gained enormous advantages from big data methodologies and techniques. This paper will show a comparative analysis based on big data techniques about social media companies such as Google, Amazon, Facebook and Twitter to undertake a comparative analysis. Google has invented many techniques by using big data methods to strategize against competitors. As an illustration, Google required the data “warehousing” approach to store trillion of data related to Facebook, since Facebook owns more than one billion users and Twitter owns 300 million active users correspondingly equally to Amazon. Since all these organizations required a data warehouse approach, facebook and Twitter are both the only social media companies that have different requirements. The requirement of big data is high and these entire requirements partially depend on each other as it is completely isolated. Facebook uses Scuba as data management tool, consumes millions of rows of data per second and expires millions of data per second also analyses live data.

## Paper 8:

**Title:** A Big Data approach to the future of death online

**Authors:** Carl J Ohman; David Watson

**Methodology:**

Internet users leave vast volumes of online data behind when passing away, commonly referred to as digital remains. The phenomenon is gaining increasing traction within the academic community. Sociologists and anthropologists are increasingly turning their gaze towards the new types of ‘para-social’ relationships and the ‘continuing bonds’ that we shape with the online dead. Online death has rapidly become a booming and diverse research area the future accumulation of profiles belonging to deceased Facebook users. The analysis suggests that a minimum of 1.4 billion users will pass away before 2100 if Facebook ceases to attract new users as of 2018. If the network continues expanding at current rates, however, this number will exceed 4.9 billion. In both cases, a majority of the profiles will belong to non-Western users. This paper shows the emerging scholarship on digital preservation and stresses the challenges arising from curating the profiles of the deceased. An exclusively commercial approach to data preservation poses important ethical and political risks that demand urgent consideration. We call for a scalable, sustainable, and dignified curation model that incorporates the interests of multiple stakeholders.

**2.c Software and Hardware Requirements**

## Software Requirements

* Linux OS
* Hadoop & MapReduce
* Facebook API
* HIVE
* Cloudera VM
* **Hardware Requirements**
* Hard Disk – 1 TB or Above
* RAM required – 4 GB or Above
* Processor – Core i3 or Above

**DATA PREPARATION PHASE**

**3.a Dataset Description**

The dataset is obtained from the study material for a course available on Stanford.edu. CS246: Mining Massive Data Sets (https://web.stanford.edu/class/cs246/info.html). The dataset contains 50000 UserIDs number from 1. The CSV file beginning with a number indicates the UID is friends with many other UIDs and each person's friend list ends with no being separated by comma ",".

DATASET URL: <https://www.dbtsai.com/assets/blog/2013/01/soc-LiveJournal1Adj.txt>

**Dataset:**

0 1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32,33,34,35,36,37,38,39,40,41,42,43,44,45,46,47,48,49,50,51,52,53,54,55,56,57,58,59,60,61,62,63,64,65,66,67,68,69,70,71,72,73,74,75,76,77,78,79,80,81,82,83,84,85,86,87,88,89,90,91,92,93,94

1 0,5,20,135,2409,8715,8932,10623,12347,12846,13840,13845,14005,20075,21556,22939,23520,28193,29724,29791,29826,30691,31232,31435,32317,32489,34394,35589,35605,35606,35613,35633,35648,35678,38737,43447,44846,44887,49226,49985,623,629,4999,6156,13912,14248,15190,17636,19217,20074,27536,29481,29726,29767,30257,33060,34250,34280,34392,34406,34418,34420,34439,34450,34651,45054,49592

2 0,117,135,1220,2755,12453,24539,24714,41456,45046,49927,6893,13795,16659,32828,41878

3 0,12,41,55,1532,12636,13185,27552,38737

4 0,8,14,15,18,27,72,80,15326,19068,19079,24596,42697,46126,74,77,33269,38792,38822

5 0,1,20,2022,22939,23527,30257,32503,35633,41457,43262,44846,49574,31140,32828

6 0,21,98,2203,3238,5040,8795,9843,9847,15294,17874,18286,18311,18320,20553,35699,35776,38736,38750,38800,543,575,11879,12682,14943,15283,18332,18560,18625,25247,33080,34412,35785,35822,42231

7 0,31993,40218,40433,1357,21843

8 0,4,38,46,72,85,24777,83,33380

9 0,6085,18972,19269

10 0,12,16,30,6027,13793,23557,29581,35477,35617,44310

11 0,1754,6027,7789,11142,12633,17898,19049,22486,26970,27554,27585,27591,27679,29576,32631,34906,41444

12 0,3,10,16,29,38,41,55,1085,1532,7714,27679,29379,35195,38737,43121,30,83,85,89,13285,27655

13 0,12584,32064,27,37,111,129,274,1383,1600,2141,7284,9172,13207,16519,18122,19051,23525,25177,30071,32045,33439,35589,39022,44412,44575,47887

14 0,4,19,19079,42697,444,42748

15 0,4,27,80

16 0,10,12,18,30,38,89,12570,19044,29319,35477,53,83,9745,15520,19010,30062,31337

17 0,19,26,28,95,128,134,150,6157,7284,12570,20016,20533,20599,42704,49678,53,29872,31337,31347,44505

18 0,4,16,30,89,2406,2411,12562

19 0,14,17,439,1100,1694,1705,2413,2644,2646,2659,2678,3734,3926,7463,9892,10240,13076,18163,19388,20290,23202,23512,25195,25239,25256,26887,27736,27808,29260,35585,44824,47445,49678,50,543,623,627,1001,1234,1343,2019,2062,2142,2324,2648,3131,3298,3573,3574,3748,3895,3931,4990,4993,5170,5487,5490,5588,5685,6995,7718,8703,8706,8768,9289,9735,10175,11399,11416,12313,12680,13661,13829,13849,13951,14070,14160,14182,15186,15356,16027,17235,17438,18075,18147,18148,18338,18977,19358,19427,19430,19983,20070,20283,20557,21822,21873,22357,22520

20 0,1,5,12846,22939,28193,29724,29791,30691,31232,34394,35589,44887,49574

**3.b Data Pre-processing – Data Cleaning**

The dataset contains 50000 UserIDs number from 1. The CSV file beginneing with a number indiactes the UID is friend with many other UIDs and each person friend list ends with no being separated by comma ",".

In the dataset the data is separated by tab spaces and spaces making it difficult to analyse a particular UID's friends. Therefore, by data cleaning process we replace all the irregular tab spaces and spaces with just by the space.

0 1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32,33,34,35,36,37,38,39,40,41,42,43,44,45,46,47,48,49,50,51,52,53,54,55,56,57,58,59,60,61,62,63,64,65,66,67,68,69,70,71,72,73,74,75,76,77,78,79,80,81,82,83,84,85,86,87,88,89,90,91,92,93,94

1 0,5,20,135,2409,8715,8932,10623,12347,12846,13840,13845,14005,20075,21556,22939,23520,28193,29724,29791,29826,30691,31232,31435,32317,32489,34394,35589,35605,35606,35613,35633,35648,35678,38737,43447,44846,44887,49226,49985,623,629,4999,6156,13912,14248,15190,17636,19217,20074,27536,29481,29726,29767,30257,33060,34250,34280,34392,34406,34418,34420,34439,34450,34651,45054,49592

2 0,117,135,1220,2755,12453,24539,24714,41456,45046,49927,6893,13795,16659,32828,41878

3 0,12,41,55,1532,12636,13185,27552,38737

4 0,8,14,15,18,27,72,80,15326,19068,19079,24596,42697,46126,74,77,33269,38792,38822

5 0,1,20,2022,22939,23527,30257,32503,35633,41457,43262,44846,49574,31140,32828

6 0,21,98,2203,3238,5040,8795,9843,9847,15294,17874,18286,18311,18320,20553,35699,35776,38736,38750,38800,543,575,11879,12682,14943,15283,18332,18560,18625,25247,33080,34412,35785,35822,42231

7 0,31993,40218,40433,1357,21843

8 0,4,38,46,72,85,24777,83,33380

9 0,6085,18972,19269

10 0,12,16,30,6027,13793,23557,29581,35477,35617,44310

**MODEL PLANNING**

**4.a Modules:**

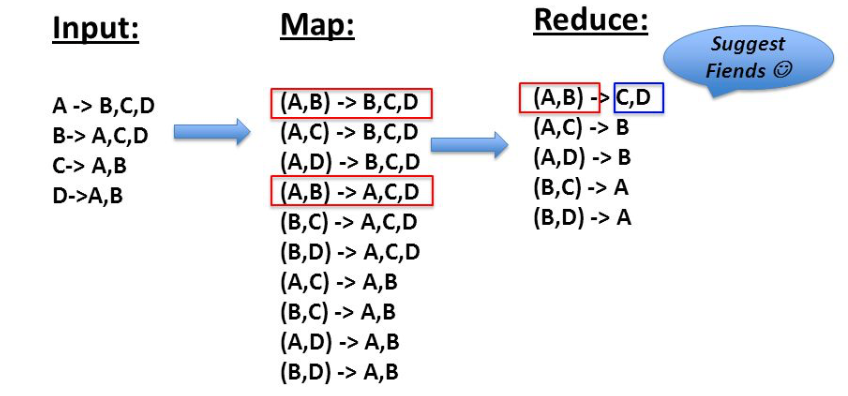
**Module-1: Mapper-1**

The map script splits data into words. Its output contains the word and 1 as its value. These tuples become the input of a reducer that sums the word count. ​​The first mapper takes each row and creates a small string output:

* + Split each row into a user number and a string of friends
  + Create pairs of the user with each friend with the smaller number first, then append a 0 to the pair.
  + Create combinations of the friends of the user, with the second friend always being larger than the first friend in the form “friend1, friend2, 1”.

**Module-2: Reducer-1**

The Mapper creates pairs of the first letter of each word with a 1. The Reducer counts the number of words having the same first letter. The first reducer orders and groups keys by value, and lists all the keys in order by id. The reducer sums values of 1s for each key. If there were no such lines that had the same key with a 0, then output “friend1 friend2 count of mutuals” for each pair that was not friends with each other.



**Module-3: Mapper-2**

* The second mapper takes this output and slightly changes it:

“f1 tab f2 numMutualFriends” and

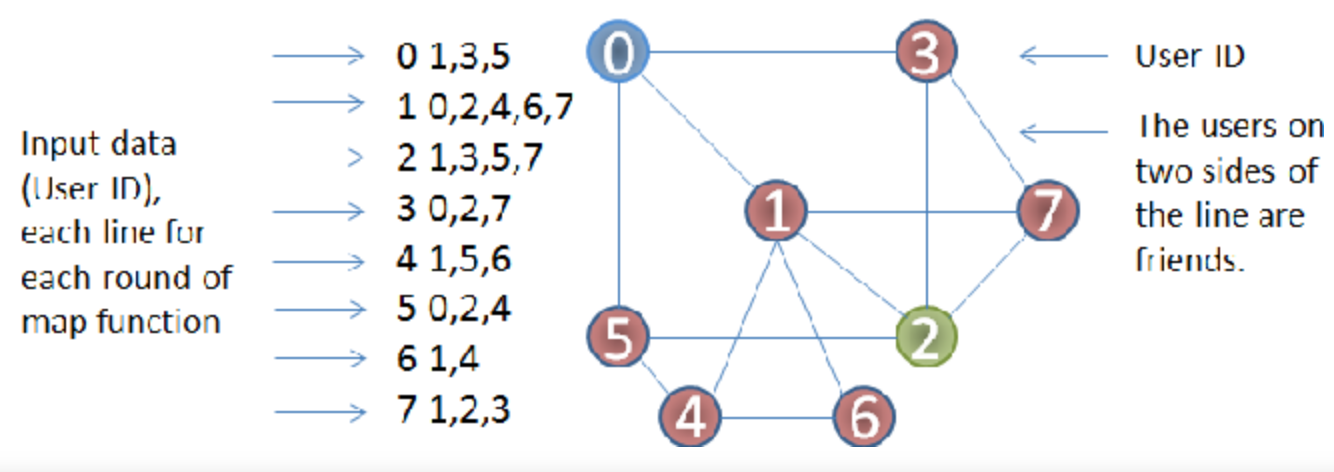
“f2 tab f1 numMutualFriends”

* Hadoop streamer orders these outputs by keys and sends them as inputs for the second reducer
* Input is ordered by user:

(user1 tab friend1, numMutualFriends)

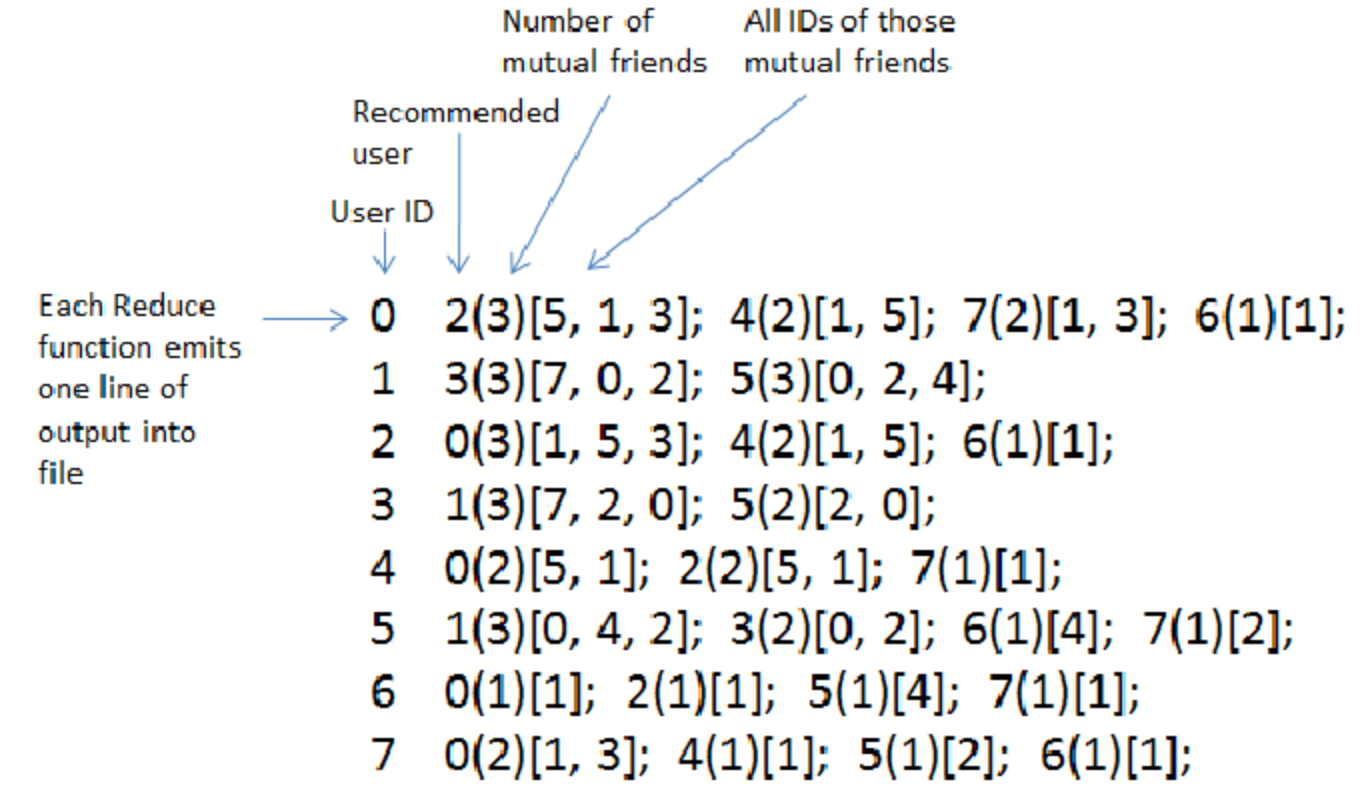
(user1 tab friend2, numMutualFriends) …..

(user 2 tab friend1…)



**Module-4: Reducer-2**

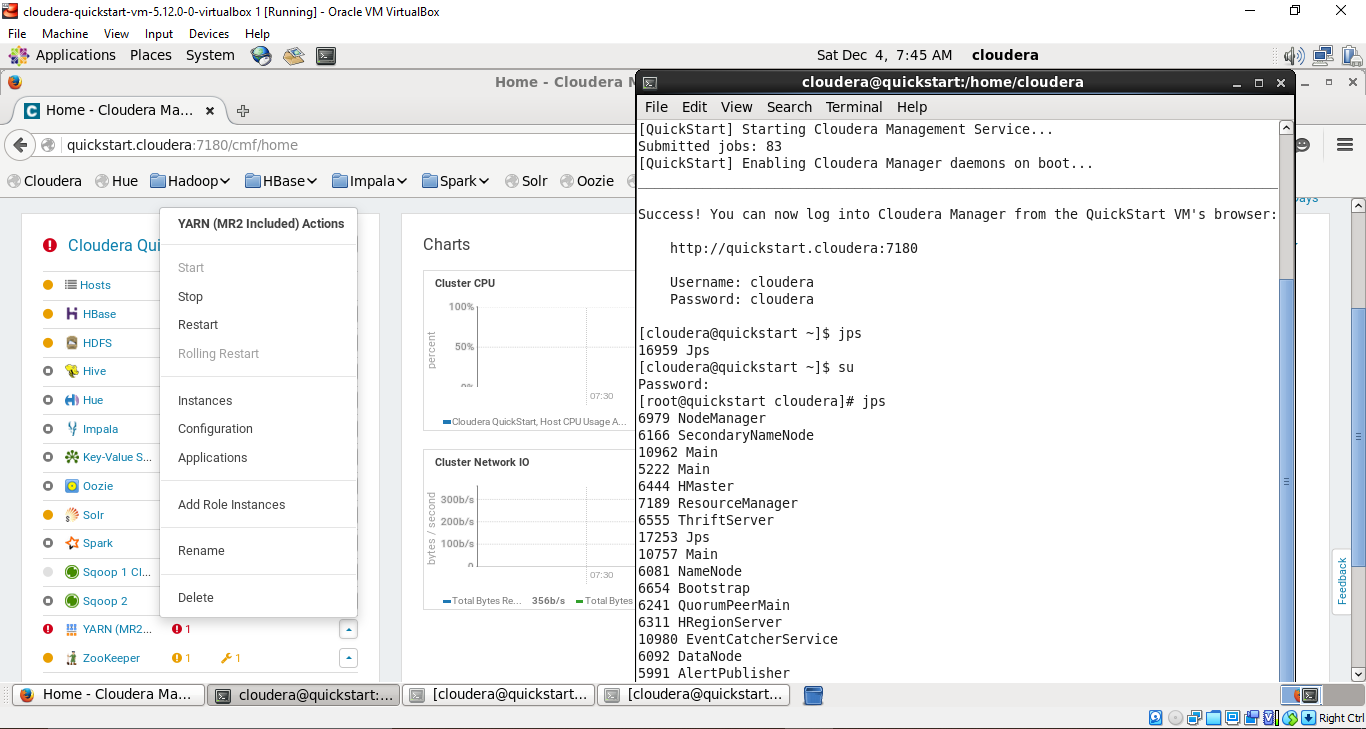
* The second reducer picks the top ten potential friends for each user.



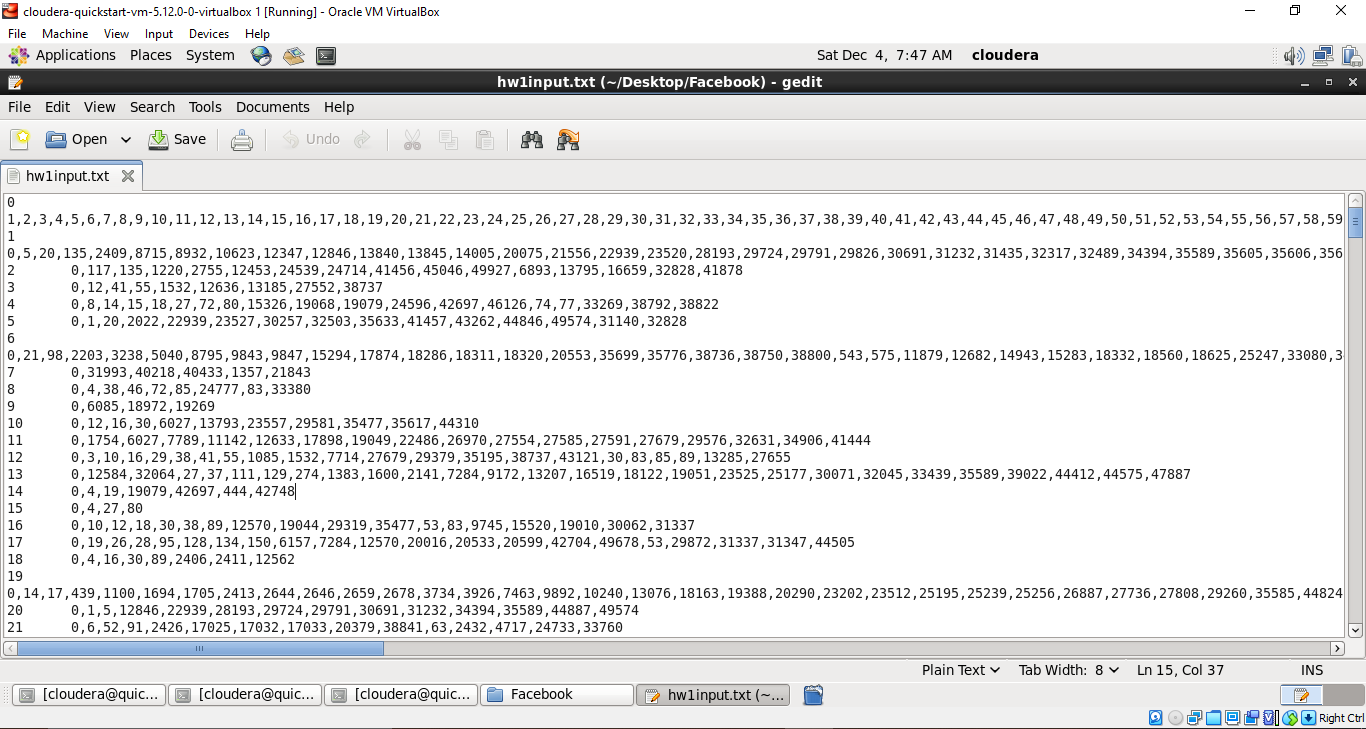
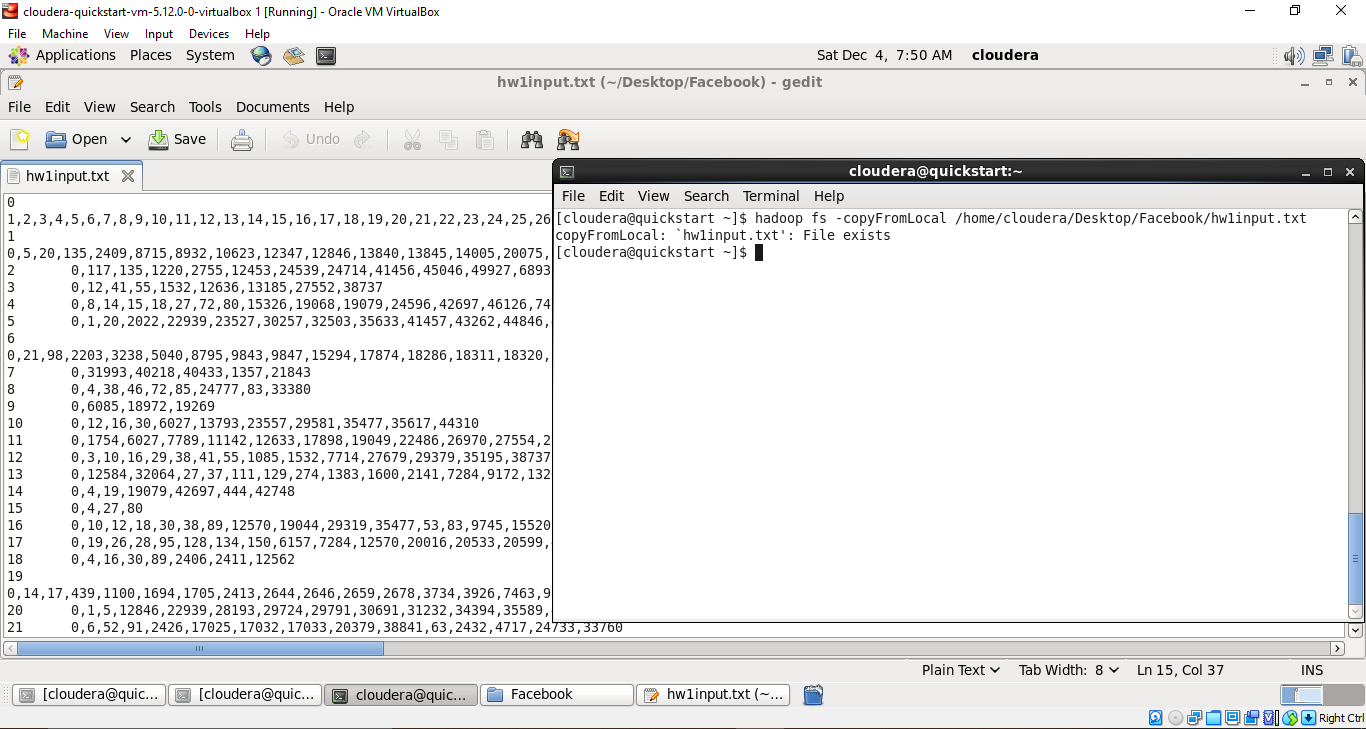
**MODULE BUILDING PHASE**

We perform Hadoop streaming twice on the Mapper Reducer pairs. Hadoop streaming is a utility that comes with the Hadoop distribution. This utility allows you to create and run Map/Reduce jobs with any executable or script as the mapper and/or the reducer.

First, we start all the clusters in Cloudera Manager

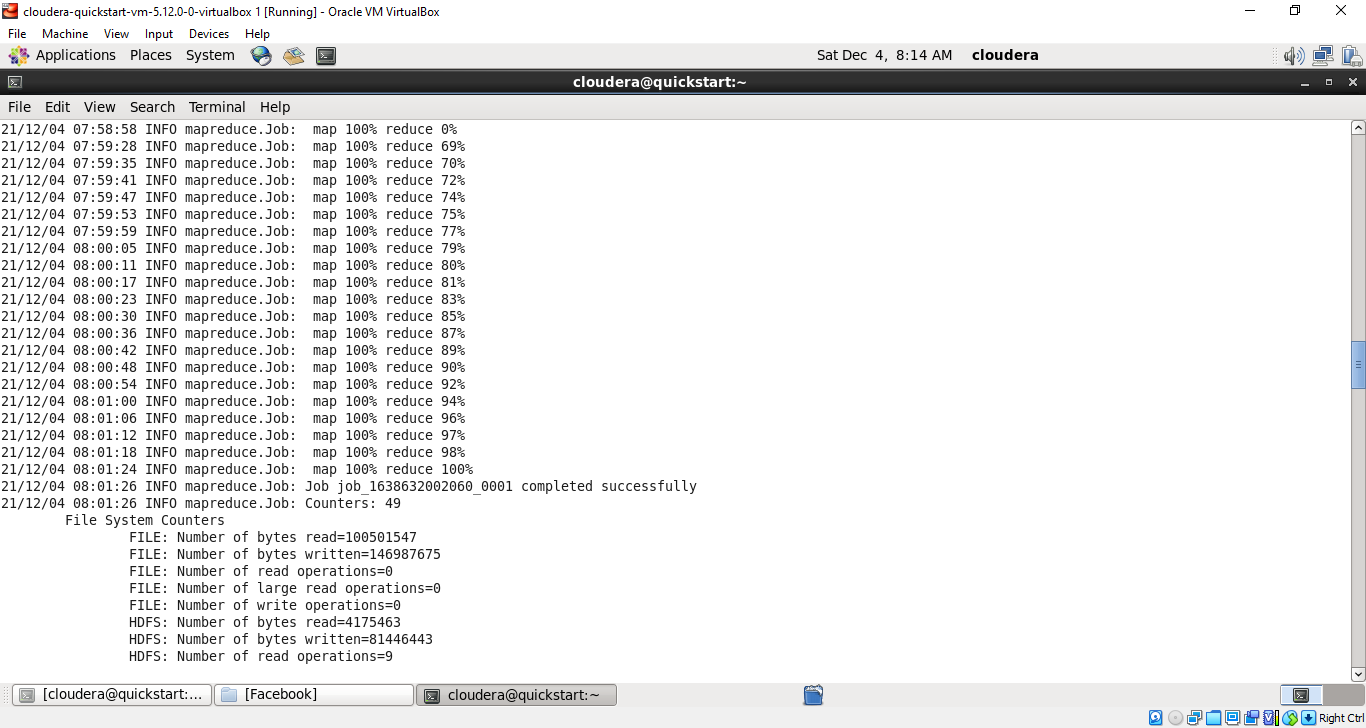
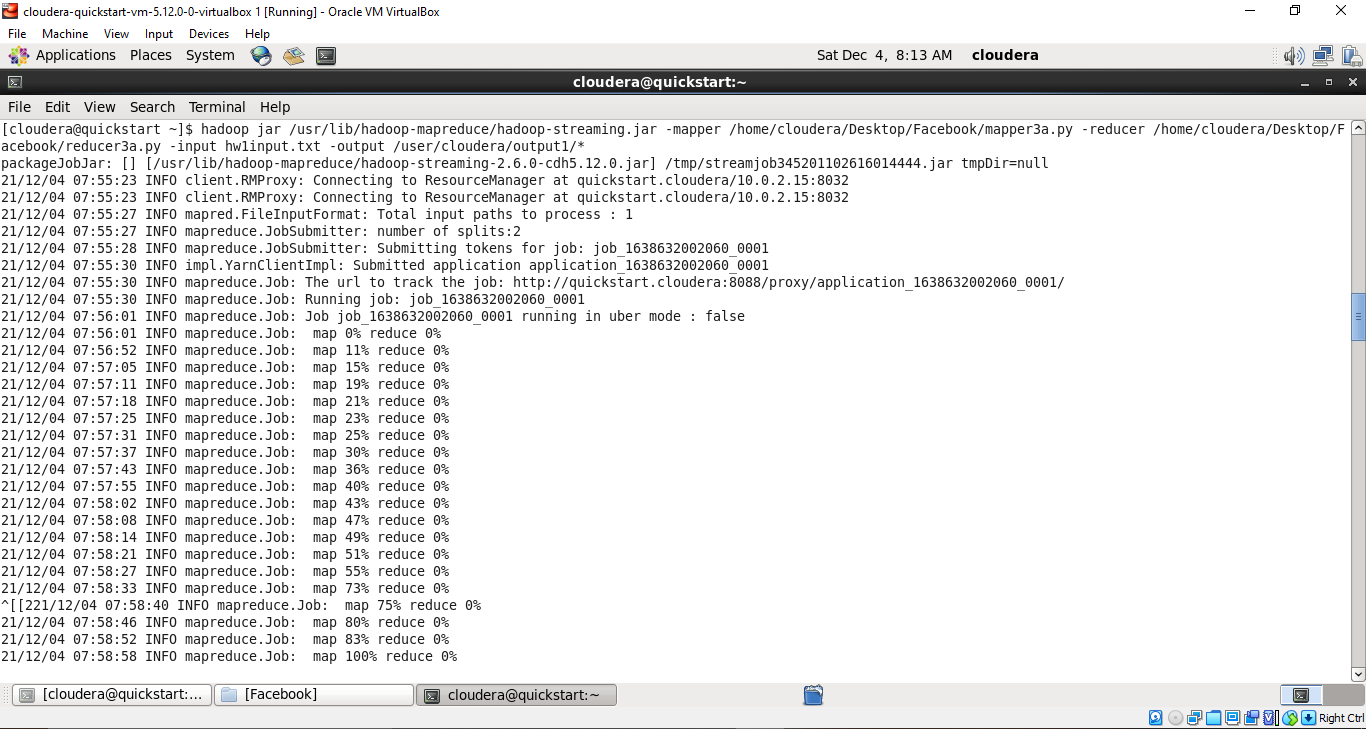


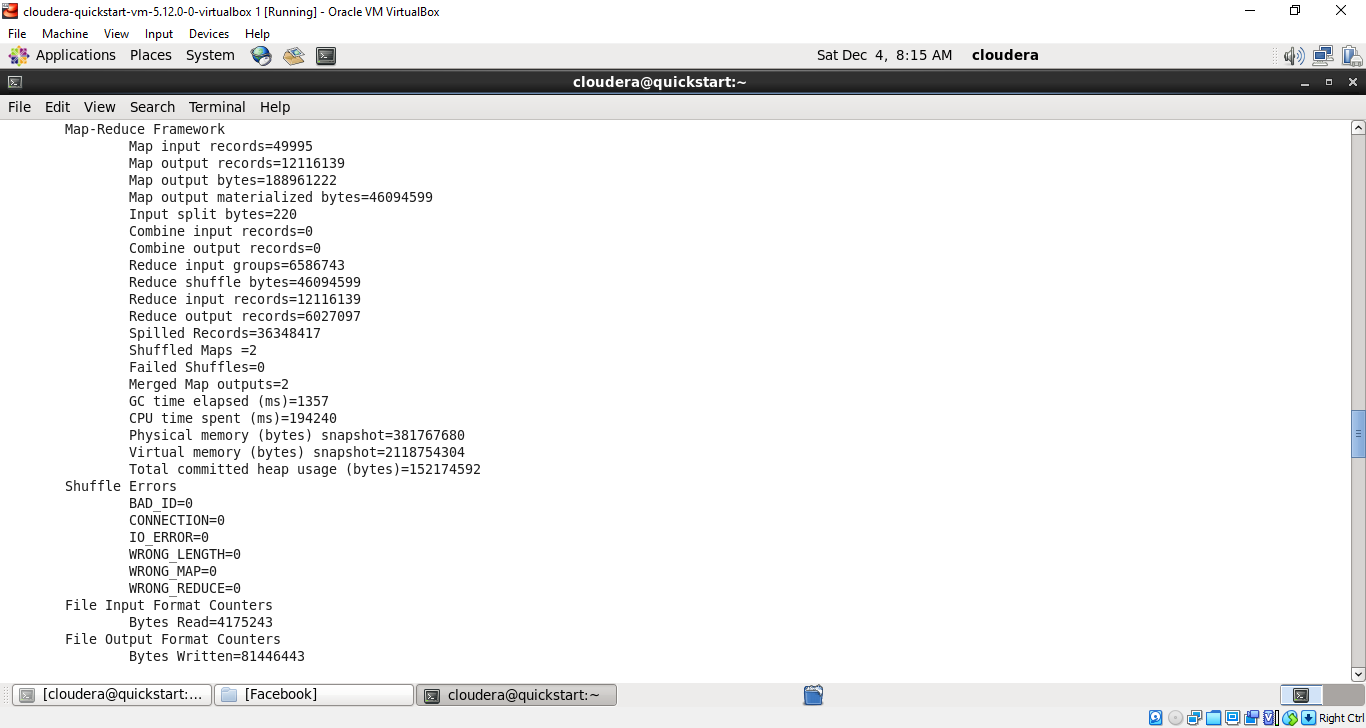
Next we input the Dataset into Hadoop File System by performing Hadoop CopyFromLocal.



**6.a Implementation of Module 1 & 2**

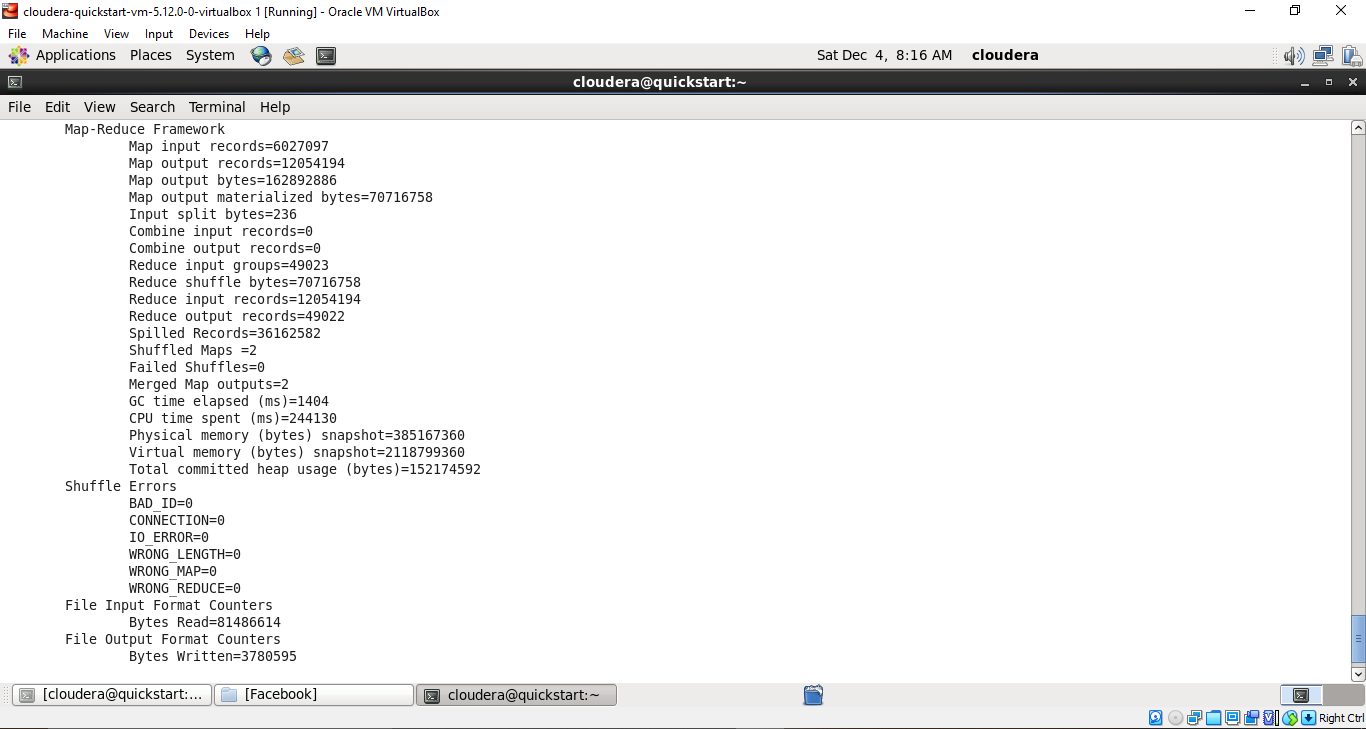
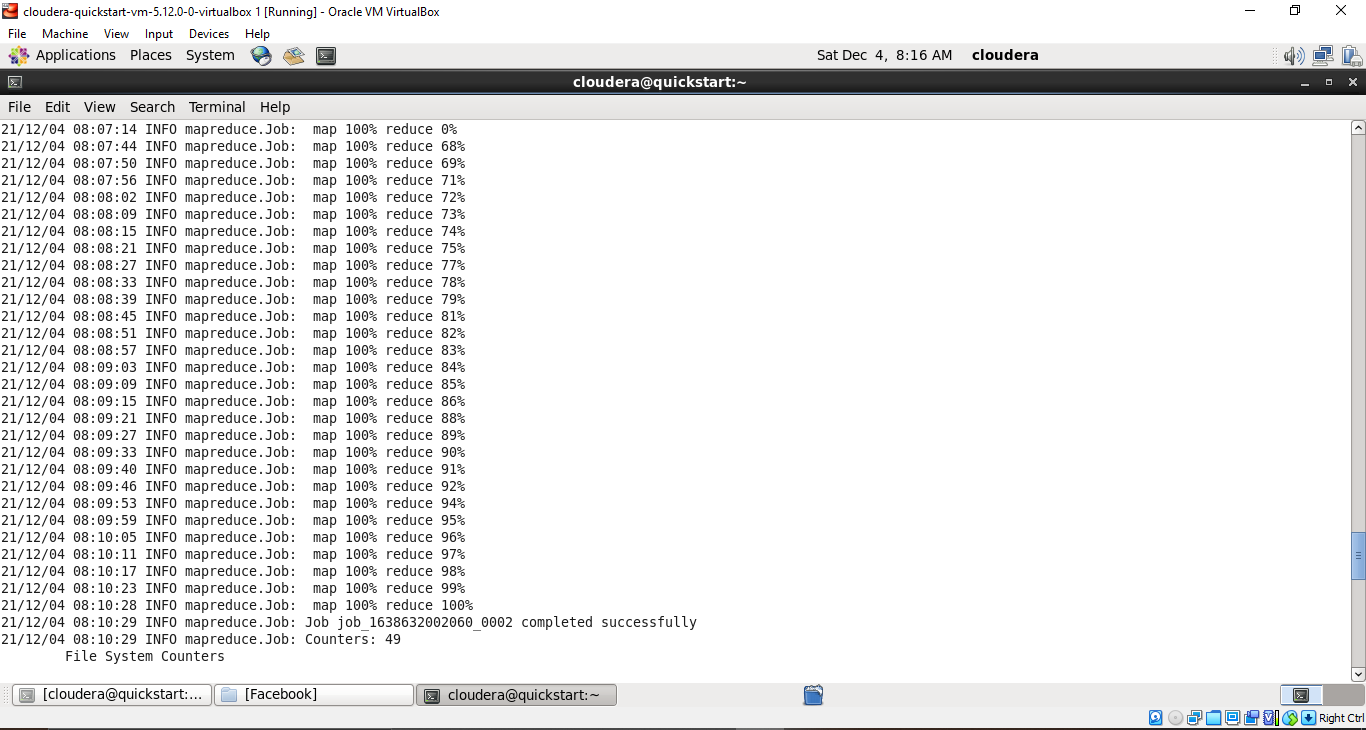
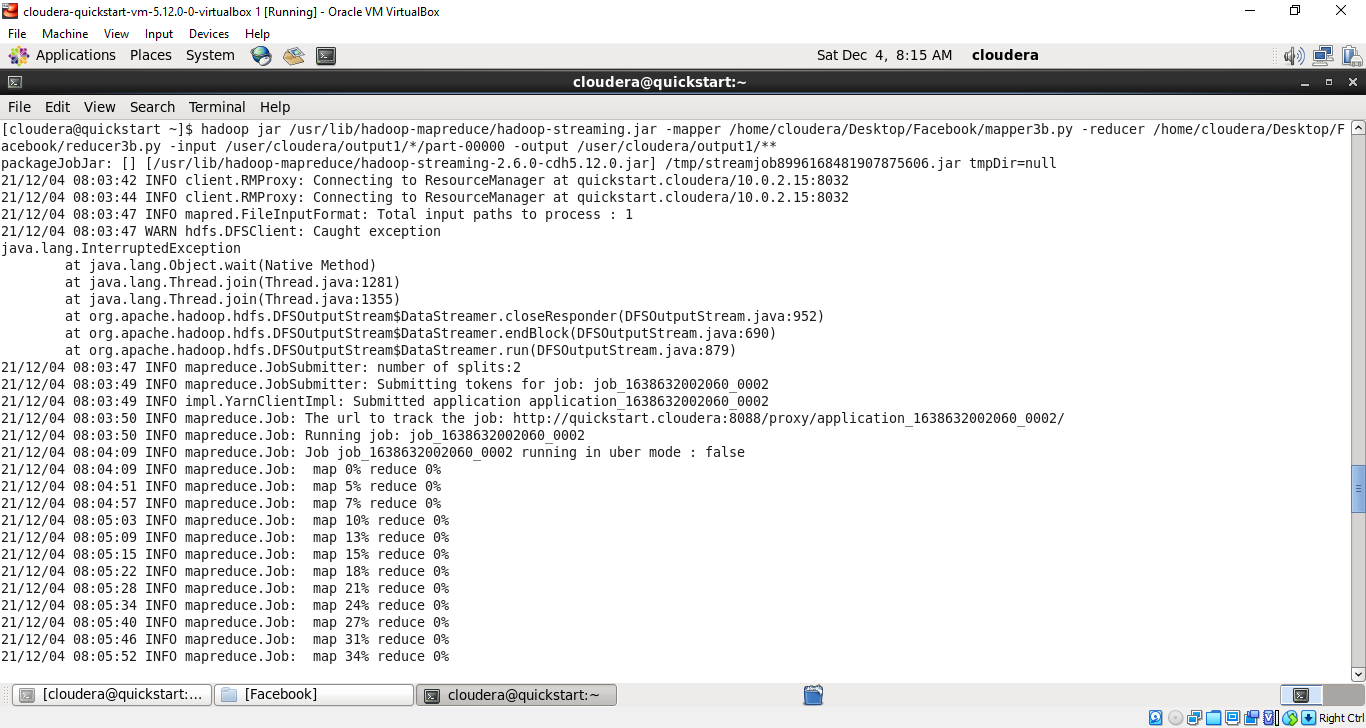
We then perform Hadoop Streaming on the first pair of mapper reducer. The output we get from streaming is stored in “ \* “ folder.



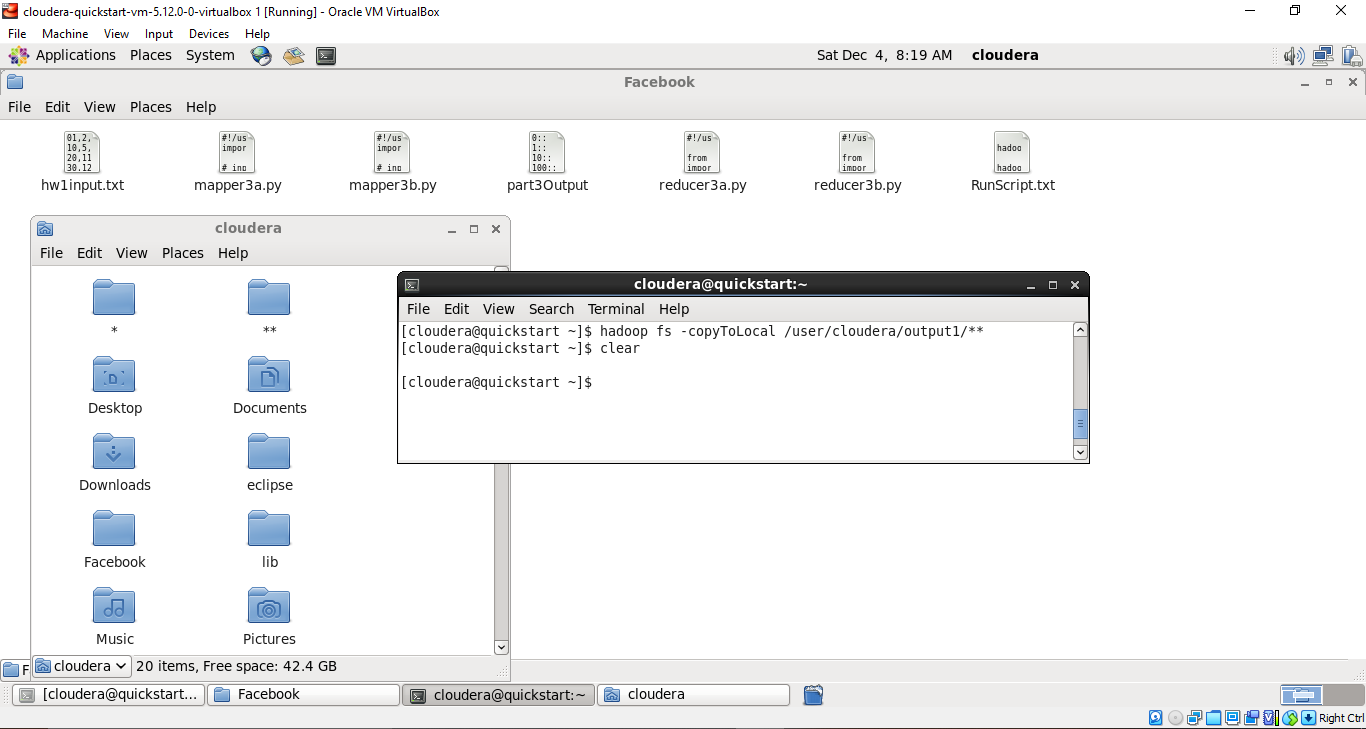


**6.b Implementation of Module 3 & 4**

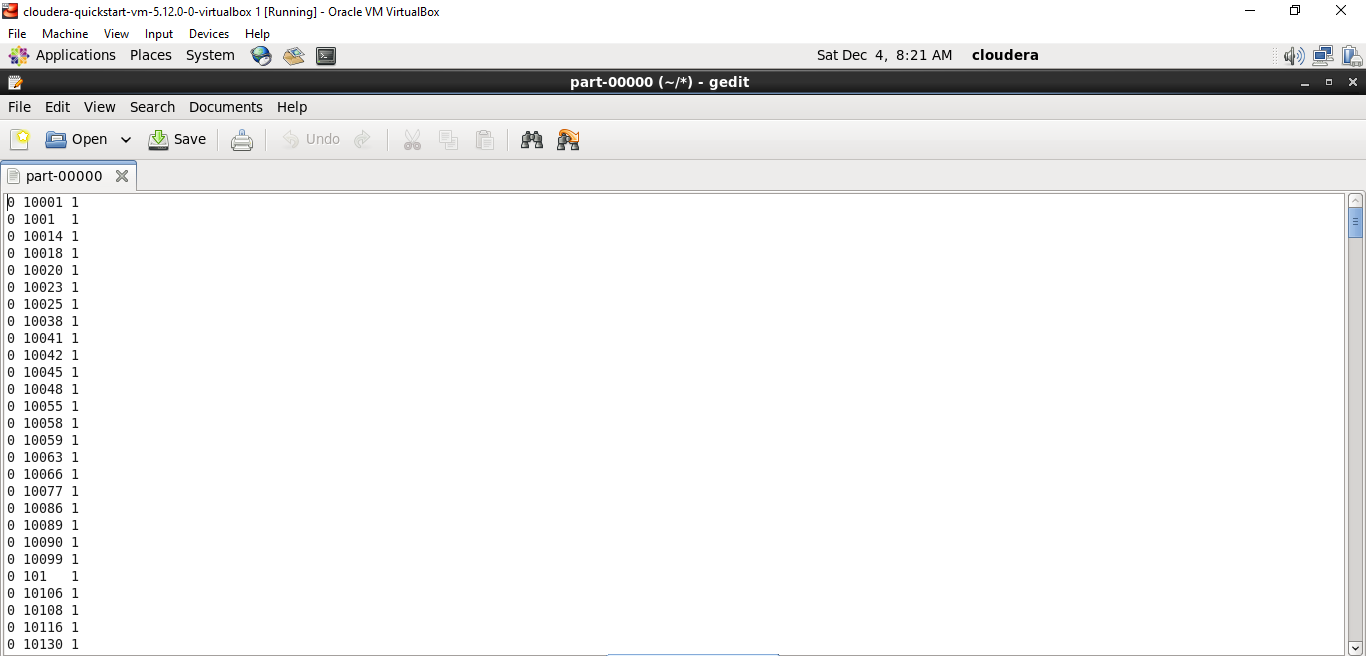
We then perform Hadoop Streaming on the second pair of mapper reducer. The output we get from streaming is stored in “ \*\* “ folder.

****

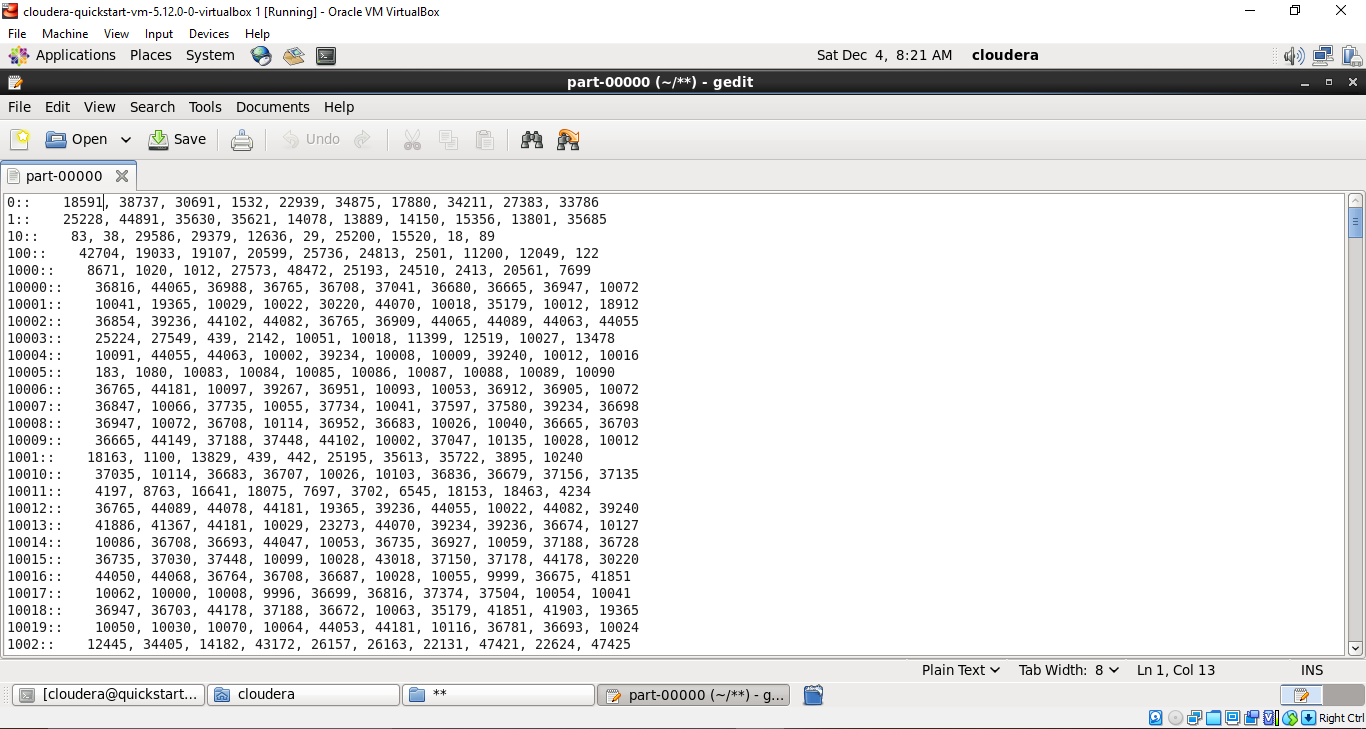
**COMMUNICATION OF RESULTS PHASE**

****

**6a. Output of Implementation of Module 1 & 2**

****

**6b. Output of Implementation of Module 3 & 4**

****

**CONCLUSION**

The user data is collected from Facebook based on their activities. User behavior, number of likes, number of posts, type of posts, their comments, etc. are stored by the database server. Comments by the user in unstructured formats, while other data in structured and semi-structured format. Petabytes of data is generated by Facebook users. So, Hadoop, MapReduce and related big data concepts are used in this project to analyze the data. Mutual friend analysis is one such field we are going to explore in this project by using MapReduce framework. As these big social networks keep growing, there are situations in which an individual user wants to find popular groups of friends so that he can recommend the same groups to other users. In this project, a big data analytic solution that uses the MapReduce model in mining these big social networks for discovering groups of frequently connected users for friend recommendation.

Hadoop map-reduce groups, maps and reduces into potential mutual friends list, very efficient when a person has large list of friends, MapReduce algorithm is used because of its accuracy and speed. In this project we applied a concept named Hadoop streaming which is a utility that enables us to create or run MapReduce scripts in any language either java or non-java as mapper or reducer. The output of this project suggests top 10 potential friends for a given user.

**REFERENCES**

1. Dragan, I., & Zota, R. (2017, September). Collecting Facebook data for big data research. In *2017 16th RoEduNet Conference: Networking in Education and Research (RoEduNet)* (pp. 1-3). IEEE.
2. Jiang, F., Leung, C. K., & Pazdor, A. G. (2016, August). Big data mining of social networks for friend recommendation. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (pp. 921-922). IEEE.
3. Ha, I., Back, B., & Ahn, B. (2015). MapReduce functions to analyze sentiment information from social big data. *International Journal of Distributed Sensor Networks*, *11*(6), 417502.
4. Iftikhar, R., & Khan, M. S. (2022). Social media big data analytics for demand forecasting: development and case implementation of an innovative framework. In *Research Anthology on Big Data Analytics, Architectures, and Applications* (pp. 902-920). IGI Global.
5. LEI, Y. (2015). An Improved Algorithm for Mutual Friends Recommendation Application of SNS in Hadoop. *法政大学大学院紀要. 情報科学研究科編= 法政大学大学院紀要. 情報科学研究科編*, (10), 1-6.
6. Peng, S., Wang, G., & Xie, D. (2016). Social influence analysis in social networking big data: Opportunities and challenges. *IEEE network*, *31*(1), 11-17.
7. Hewage, T. N., Halgamuge, M. N., Syed, A., & Ekici, G. (2018). Big Data Techniques of Google, Amazon, Facebook and Twitter. *J. Commun.*, *13*(2), 94-100.
8. Öhman, C. J., & Watson, D. (2019). Are the dead taking over Facebook? A Big Data approach to the future of death online. *Big Data & Society*, *6*(1), 2053951719842540.
9. https://hadoop.apache.org/docs/r1.2.1/streaming.html