Recommendation System by Using Hadoop Framework and Mahout Interfaces

[1] Yatisha Bhoir, [2] Deepali Patil
[1]Lecturer, [2]Lecturer
[1]Computer Engineering
[1]Shree L.R.Tiwari College of Engineering,Miraroad,India

Abstract : Big data is a term for massive data sets having large, more varied and complex structure with the difficulties of storing, analyzing and visualizing for further processes or results. Hadoop is an open source distributed processing framework that manages data processing and storage for big data applications running in clustered systems. Recommendation system provides the facility to understand a person's taste and find new, desirable content for them automatically based on the pattern between their likes and rating of different items [5]. Although people's tastes vary, they do follow patterns. People tend to like things that are similar to other things they like as well as another similar behavioral person likes. Sometimes these types of patterns can be related with the relevancy of items. On the other hand, we could figure out what items are similar to what we already liked, again by looking to others apparent preferences. In this, we proposed a movie recommendation system for large amount of data available on the web in the form of ratings, reviews, opinions, complains, remarks, feedbacks, and comments about any item (product, event, individual and services) using Hadoop Framework. Depending on the taste of the person, a list of movies would be recommended to them. Mahout is a data mining framework that normally runs coupled with the Hadoop infrastructure at its background to manage huge volumes of data. We have implemented Mahout Interfaces for analyzing the data provided by review and rating site for movies.

Index Terms: Big-data, Pattern, Recommendation, Review, Hadoop, Mahout.

I. INTRODUCTION

The large amount of data available on the web in the form of ratings, ranks, reviews, opinions, complain, remarks, feedback, and comments about any item (product, event, individual and services) can be used for making correct decision [5]. Now a days many other communication portals like Facebook, WhatsApp, twitter, etc. are available where user can ask their friends and followers about any product, service, event, and issues before making any decision. To handle these huge amount of data (also called Big Data), we require a robust system. Now the question remains: How do we analyze that amount of text data? The most popular answer to this is: Hadoop framework. Hadoop is an open-source framework for developing and executing distributed applications that process very large amount of data. [1][2].

Recommendation system provides the facility to understand a person's taste and find new, desirable content for them automatically based on the pattern between their likes and rating of different items [1]. Recommendation is a popular technique that provides close recommendations based on user information such as previous purchases, clicks, and ratings. Amazon uses this technique to display a list of recommended items that you might be interested in, drawing information from your past actions [3]. There are recommender engines that work behind Amazon to capture user behavior and recommend selected items based on your earlier actions. Facebook uses the recommender technique to identify and recommend the "people you may know list". These recommendations are based on two filtering techniques namely collaborative and content –based filtering techniques [2][5]. Collaborative filtering produces recommendations based on, and only based on, knowledge of user's relationship with items, product and services. These techniques require no knowledge of properties of items and characteristics [2]. Content-based recommendations are based on attributes of items. Here suggestions are based on the content related to items and their aspects [2]. One more approach has become more popular where both the filtering techniques can be applied on different levels of recommendation system, called hybrid filtering technique [2][7].

In this project, we have used hybrid filtering technique. We proposed a movie recommendation system for large amount of data available on the web in the form of ratings, reviews, opinions, complains, remarks, feedbacks, and comments about any item (product, event, individual and services) using Hadoop Framework. We have implemented Mahout Interfaces for analyzing the data provided by review and rating site for movies. Depending on the taste of the person, a list of movies would be recommended to them. We have implemented statistical analysis on the reviews which are generated by various users.

II. RELATED WORK

Many papers are available in literature on Big Data Analysis for Recommendation System using different platforms.

A. Big data analytics-recommendation system with Hadoop Framework[1]

This paper proposed movie recommendation system for large amount data available on the web in the form of ratings, reviews, opinions, complain, remarks, feedback, and comments about any item (product, event, individual and services) using Hadoop Framework. Depending on the taste of the person a list of movies would be recommended to him.

B. Design of large-scale content-based recommender system using Hadoop MapReduce framework [2]

In this paper, the improved MapReduce based data preprocessing and Content based recommendation algorithms are proposed and implemented using Hadoop framework. Also, graphical user interfaces are developed to interact with the recommender system. Experimental results on Amazon product co-purchasing network metadata show that Hadoop distributed cluster environment is an efficient and scalable platform for implementing large scale recommender system.

C. Design of large-scale mobile advertising recommendation system [3]

This paper proposed a mobile advertising recommendation system using item-based Top-N recommendation algorithm based on Hadoop framework and then propose a solution based on the Hadoop framework to build a distributed recommendation system. They also introduced advertising server into the system. Seller will pay for their advertising when it is clicked by users, so that this platform can make profit. They used data analysis algorithm based on MapReduce programming model to achieve good performance and high scalability. The experiments show that system can improve efficient clearly.

D. Building hybrid recommendation system based on Hadoop framework [4]

This paper proposed MapReduce framework and compared the differences time consumption of common seral hybrid recommendation algorithm with parallel hybrid recommendation algorithm by using different clusters. Here they have applied the experiments by using Each Movie data sets to exploit the advantages of parallel algorithm. Also, from the experiments they can get how improved parallel K-Mean algorithm by using two methods such as distance measure method and initial centroids method which are based on MapReduce can achieve higher accuracy as compare to tradition K-Mean algorithm and performance can improve.

E. Big Data Analysis: Recommendation System with Hadoop Framework[5]

This paper proposed a recommendation system for the large amount of data available on the web in the form of ratings, reviews, opinions, complaints, remarks, feedback, and comments about any item (product, event, individual and services) using Hadoop Framework. They have implemented Mahout Interfaces for analyzing the data provided by review and rating site for movies.

F. Cloud Based Predictive Analytics: Text Classification, Recommender Systems and Decision Support [6]

This paper presents a detailed study of technologies based on Hadoop and MapReduce available over the cloud for large-scale data mining and predictive analytics. Although some studies may have shown that cloud technologies relying on the MapReduce framework do not perform as well as parallel database management systems, e.g., with ad hoc queries and interactive applications, MapReduce has still been widely used by many organizations for big data storage and analytics. A number of MapReduce based tools are broadly available over the cloud. In this work they explore the Apache Hive data warehousing solution and particularly its Mahout data mining libraries for predictive analytics. They present results in the context of text classification, recommender systems and decision support. They develop prototype tools in these areas and discuss outcomes from the study useful to researchers and other professionals in cloud computing and application domains.

G. A Framework for Recommender Systems in E-Commerce Based on Distributed Storage and Data-Mining

A recommender system in E-Commerce is an intermediary program (or an agent) with a user interface that automatically and intelligently generates a list of information which suits an individual's needs. In this paper, a new recommender system framework based on Distributed Storage and Data mining is proposed. The recommendation systems work by collecting data from users, commodities, and transactions, examples of implicit data collection include the following: observing the items that a user views in an online store, keeping a record of the items that a user purchases online, obtaining a list of items that a user has listened to or watched on his/her computer, et al. This data is stored by Hadoop distributed file system (HDFS) and mined in Map Reduce that is a programming model and an associated implementation for processing and generating large data sets

H. High dimensional datasets using Hadoop mahout machine learning algorithms[8]

As the Data increasing very drastically day-to-day, it is a major issue to manage and organize the data very efficiently. This emerged the necessity of machine learning techniques. With the Fast development of Networking, data storage and the data collection capacity, Machine learning cluster algorithms are now rapidly expanding in all science and engineering domains such as Pattern recognition, data mining, bioinformatics, and recommendation systems. So as to support the scalable machine learning framework with MapReduce and Hadoop support, they are using Apache Mahout to manage the High Voluminous data. Various Cluster problems such as Cluster Tendency, Partitioning, Cluster Validity, and Cluster Performance can be easily overcome by Mahout clustering algorithms. Mahout manages data in four steps i.e., fetching data, text mining, clustering, classification and collaborative filtering. In the proposed approach, various datatypes such as Numeric, Characters and Image datasets are classified in the several categories i.e., Collaborative Filtering, Clustering, Classification or Frequent Item set Mining. Some of the Pre-clustering techniques are also implemented such as EDBE, ECCE, and Extended Co-VAT. A non-Hadoop Cluster named Taste recommendation Framework is also implemented.

I. Review based service recommendation for big data [9]

Success of web 2.0 brings online information overload. An exponential growth of customers, services and online information has been observed in last decade. It yields big data investigation problem for service recommendation system. Traditional recommender systems often put up with scalability, lack of security and efficiency problems. Users preferences are almost ignored. So, the requirement of robust recommendation system is enhanced now a days. This paper proposed review-based service recommendation to dynamically recommend services to the users. Keywords are extracted from passive users reviews and a rating value is given to every new keyword observed in the dataset. Sentiment analysis is performed on these rating values and top-k services recommendation list is provided to users. To make the system more effective and robust Hadoop framework is used.

J. Design of friend recommender system using apache Hadoop [10]

Success of web 2.0 brings online information overload. An exponential growth of customers, services and online information has been observed in last decade. It yields big data investigation problem for service recommendation system. Traditional recommender systems often put up with scalability, lack of security and efficiency problems. Users preferences are almost ignored. So, the requirement of robust recommendation system is enhanced now a days. This paper proposed review-based service recommendation to dynamically recommend services to the users. Keywords are extracted from passive users reviews and a rating value is given to every new keyword observed in the dataset. Sentiment analysis is performed on these rating values and top-k services recommendation list is provided to users. To make the system more effective and robust Hadoop framework is used.

K. Addressing Cold Start Problem in Recommendation System Using Custom Built Hadoop Ecosystem[11]

In the modern digital era, the most common problem is being surrounded by the e-Commerce and social networking sites which provides accurate and timely recommendations based on user interests. When we have user-related data, it will be easier to give recommendations to those users based on their previous activity in that particular website. But cold start problem arises when there is no information related to a user or an item, because they are new to that particular website. Hadoop is one of the distributed huge data processing frameworks that provides an effective solution to overcome these cold start problems in real time scenarios. Here, they have used the tools like Sqoop, Hive and build a model that address and resolve certain cold start issues in recommendation systems.

L. CSRS: Customized service recommendation system for big data analysis using map reduce[12]

Recommender systems have been important helpful software system techniques and tools giving desirable recommendation to items that are valuable to end-users. In past era, the measure of services, users, and online data have expanded very quickly, so it recognizes the Big data that is the reason there is examination issue for recommender systems. As a result, efficiency and scalability are issues in existing service recommender systems when the analyzing or processing Big Data. Also, the same rankings and ratings are provided by existing service recommender systems to various users without counting his/her distinctive preferences and in this manner, doesn't meet user's customized essentials. In this paper, it's proposed to utilize service recommendation techniques, which goes for showing a customized service recommendation set and recommender systems are recommending the foremost valuable services to end-users. To improve efficiency and scalability issue in massive data surroundings, it is implemented in Hadoop framework which uses map reduce parallel processing model.

M. A Study on Key Techniques of Wisdom Campus Information Recommendation Platform Based on Big Data[13]

This thesis employs the mainstream Hadoop framework to store and analyze the large-scale data generated in the campus environment. The system will first collect, process and save data by using the HDFS distributed file system. And then data mining and intelligent analysis of the data will be carried out. Last, the recommendation system framework which deals with large dataset will be put forward based on the above two steps. Therefore, a modified collaborative filtering recommendation algorithm based on the students' similarity will be achieved and the accuracy as well as the efficiency of recommendation will be improved.

N. A comparative study of recommendation algorithms in e-commerce[14]

The recommender systems are used extensively around the world to increase sales and profits. A proper recommender system improves overall user shopping experience. In the e-commerce industry, this approach helps acquire potential market opportunities. This paper focuses on the comparison of different recommendation algorithms and provides a guideline to choose the best algorithm. Better and appropriate recommendations can be achieved by analyzing a large amount of data by using Hadoop MapReduce framework, along with appropriate GUI

O. Big data analytics framework to identify crop disease and recommendation a solution [15]

Due to technology the term data is replaced by transforming big data in many fields. Rapidly advancements in the technology causes agricultural data enter into the era of big data. Traditional tools and techniques are unable to store and analyze this massive amount of data. To store and analyze this type of data parallel computing and analyze paradigm is required. Big data analytic is used as a solution to this. In this paper big data analytic agriculture framework is developed that identify disease based on symptoms similarity and recommend a solution based on high similarity. To achieve this objective Hadoop and Hive tools has been used. The data is collected, cleansed and normalized. Data is collected from laboratory reports, web sites etc. then cleansing of data is done that is important information is extracted from unstructured redundant data. In the next step normalization is done that is features are extracted from cleaned data. Normalized data is uploaded on HDFS and save in a file supported by hive. HiveQL is a SQL like query language and used to analyze the agricultural data. It finds out disease name based on crop disease symptoms and purpose a solution based on evidence from historical data. Result is represented in form of graphs that will useful for recommending a solution that is highly used or high symptoms similarity.

III. PROBLEM STATEMENT

"Enhancing Performance of Recommender Systems" deal with improving performance of recommender systems applicable to various domains. Goal of this work is to make recommendation methods more accurate and applicable to broader range of real-life needs [2][6]. Different applications generate different categories of data. Depending on nature of data, data processing varies. Suitable representation of data would also optimize performance of a recommender system. Hence, need is to identify suitable data processing scheme along with the best suited data representation to optimize performance of recommender systems pertaining to different domains [8].

The large amount of data available on the web in the form of ratings, ranks, reviews, opinions, complain, remarks, feedback, and comments about any item (product, event, individual and services) can be used for making correct decision. Now a days many other communication portals like Facebook, WhatsApp, twitter, etc. are available where user can ask their friends and followers about any product, service, event, and issues before making any decision. To handle these huge amount of data (also called Big Data), we require a robust system [6]. Now the question remains: How do we analyze that amount of text data? The most popular answer to this is: Hadoop framework. Hadoop is an open-source framework for developing and executing distributed applications that process very large amount of data. Now to recommend data to various users, we use recommendation system. Recommendation system provides the facility to understand a person's taste and find new, desirable content for them automatically based on the pattern between their likes and rating of different items [11]. Amazon uses this technique to display a list of recommended items that you might be interested in, drawing information from your past actions. There are recommender engines that work behind Amazon to capture user behavior and recommend selected items based on your earlier actions [13]. Facebook uses the recommender technique to identify and recommend the "people you may know list". There are various sites available on the internet who recommends movies to users. Now for correct recommendation, we need efficient recommendation system [14].

IV. PROPOSED SOLUTION

So, in this project "Big Data Analysis: Recommendation System with Hadoop Framework" we are developing movie recommendation system by using Hadoop Framework and Mahout Interfaces. We proposed a movie recommendation system to recommend movies to the users based on their past choices. We used Hadoop framework and Mahout interfaces to analyze big data. Here we used a hybrid filtering technique to filter different types of reviews, opinions, remarks, comments, complains etc. Because recommendations are based on ratings, ranks, content, reviewer's behavior, and timing of review generated by different reviewers. We have implemented statistical analysis on the reviews generated by different uses. Recommendation by applying the weightage of summarized reviews and opinions on the rating of item.

A. Recommendation System Platform with Hadoop

Hadoop provide a framework for storage, management, and retrieval of the big amount of data known as Big Data. Many high-level languages, predictive analysis algorithms, and other tools for different task can be integrated with Hadoop framework [14]. For the above-mentioned processes for recommendation system we will configure Mahout and Hive with Hadoop framework. HDFS is a distributed file system designed to run on commodity hardware. It is highly fault-tolerant, designed to be deployed on low-cost hardware. It requires a few POSIX for working with different environment systems. MapReduce is a programming model designed for processing large volumes of data in parallel by dividing the work into a set of independent tasks [12]. Hive is a data warehousing solution built on top of Hadoop. It provides SQL like query language named HiveQL. The Apache Mahout free machine learning library's goal is to build scalable machine learning tools and data mining framework for use on analyzing big data on a distributed manner [5].

B. Mahout Architecture and Algorithms:

Mahout is a project of the Apache Software Foundation to produce free implementations of distributed, scalable machine learning algorithms focused primarily in the areas of collaborative filtering, clustering and classification [14]. As per mahout architecture (Figure-1), machine learning algorithms with different performance measurement methods are implemented in distributed computing environment. In this work we will implement mahout interfaces with Hadoop framework. Evaluation for the different implementations is actually very time-consuming. The strength of Mahout lies in that it is possible to save time in the evaluation of the different combinations of the parameters. It provides standard interface for the evaluation of a Recommender System. Mahout provides classes for the evaluation of a recommender system [5].

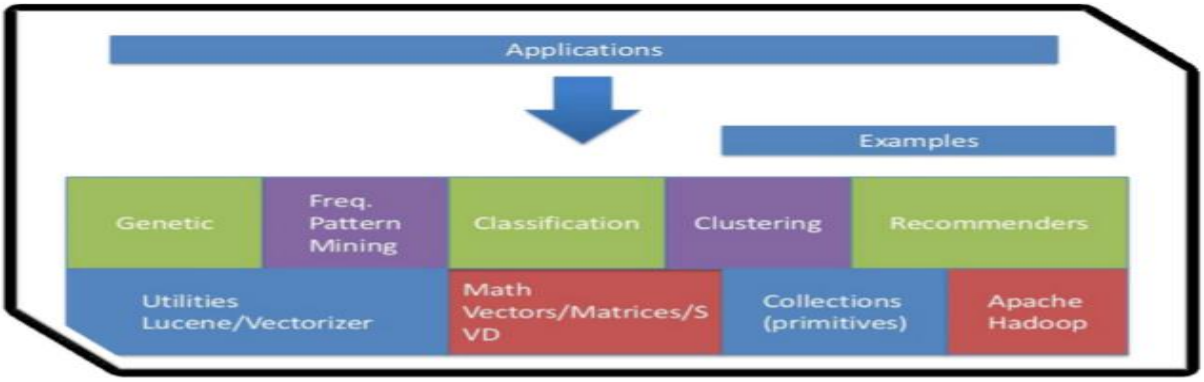

Fig.1 Mahout Architecture

C. Mahout Recommender Engine

Mahout has a non-distributed, non-Hadoop-based recommender engine. You should pass a text document having user preferences for items. And the output of this engine would be the estimated preferences of a particular user for other items. The components provided by Mahout to build a recommender engine are as follows:

1. Data Model
2. User Similarity
3. Item Similarity
4. User Neighborhood
5. Recommender

D. Building a Recommender using Mahout

Here are the steps to develop a simple recommender:

Step1: Create Data Model Object

The constructor of **Pearson Correlation Similarity** class requires a data model object, which holds a file that contains the Users, Items, and Preferences details of a product. The **Data Model** object requires the file object, which contains the path of the input file

Step2: Create User Similarity Object

Create **User Similarity** object using **Pearson Correlation Similarity** class

Step3: Create User Neighborhood object

This object computes a "neighborhood" of users like a given user. There are two types of neighborhoods:

Nearest N user Neighborhood - This class computes a neighborhood consisting of the nearest *n* users to a given user. "Nearest" is defined by the given User Similarity.

ThresholdUserNeighborhood - This class computes a neighborhood consisting of all the users whose similarity to the given user meets or exceeds a certain threshold. Similarity is defined by the given User Similarity.

Step4: Create Recommender Object

Create **User based Recommender** object. Pass all the above created objects to its constructor.

Step5: Recommend Items to a User

Recommend products to a user using the recommend () method of **Recommender** interface. This method requires two parameters. The first represents the user id of the user to whom we need to send the recommendations, and the second represents the number of recommendations to be sent.

E. Architecture of Recommender Engine

Figure-2 shows Architecture of Recommender Engine.

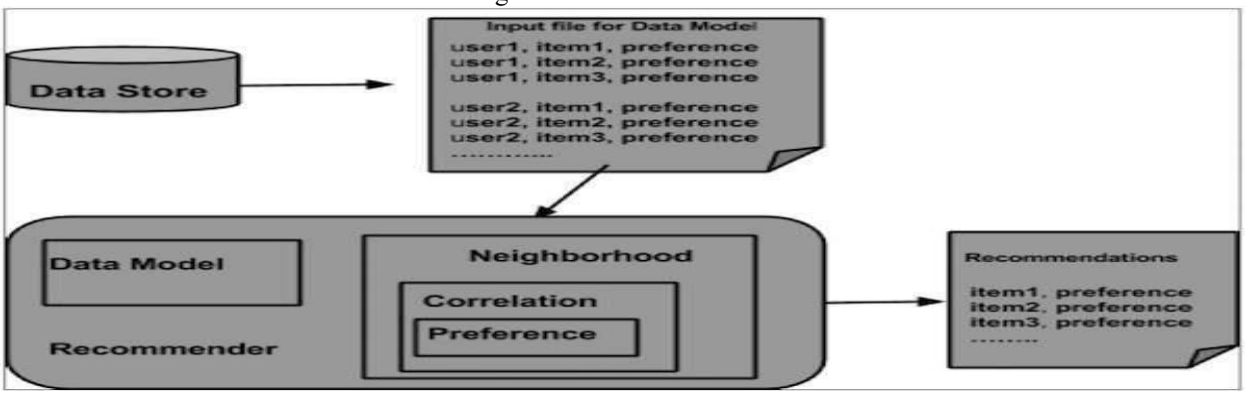

Fig.2 Architecture of Recommender Engine

F. Flowchart of project (Movies Recommender System)

Fig.3 shows the flow diagram of movies recommendation system.

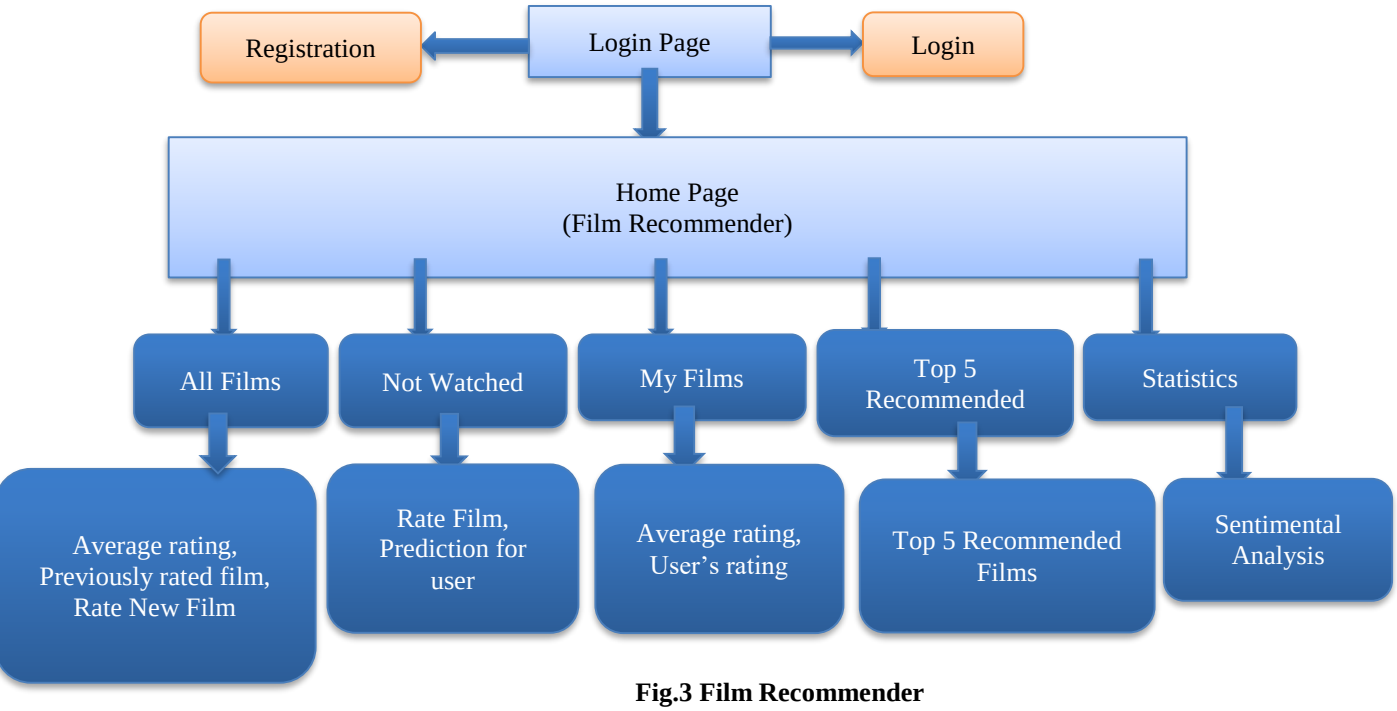

Fig.3 Film Recommender

V. RESULTS AND DISCUSSION

Fig.4 shows login page where user have to enter username and password to enter into recommendation page. For new user, Sign Up option is available where he needs to enter basic information for creating username and password.

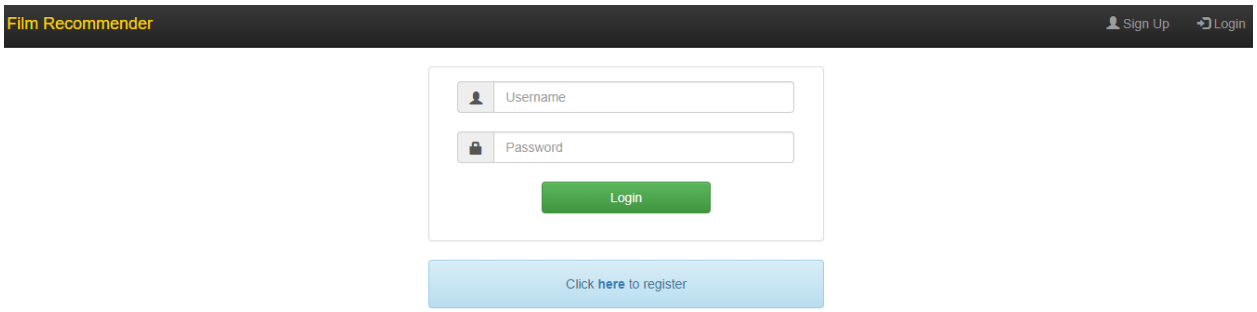

Fig.4 Login Page

After successful login, user will enter into the home page of Film Recommender. Home page includes All Films, Not Watched Films, My Films, Top 5 recommended films and Statistics. Fig.5 shows page of All Films where user can see all films included into the system and can see average rating given by all users to that particular film. User can give ratings to the films here. We can add any number of films as we are using Hadoop framework to deal with big data.

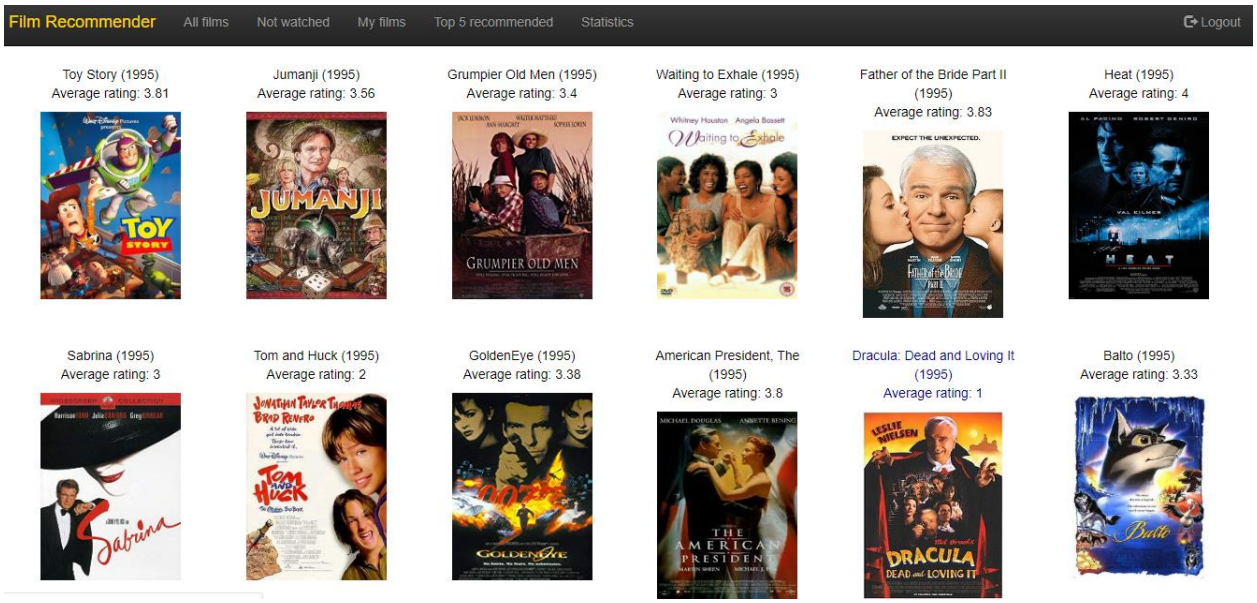

Fig.5 All Films

Fig. 6 shows the page where user can see how much rating he had gave previously to the particular film by clicking on that film poster. User can also see average rating given by all users to that film.

Fig.6 Previously Rated Film

Fig.7 shows, list of not watched films. This films are recommended for user based on previous choices. Here user can see description of film, Average rating given by all users and Prediction for user.

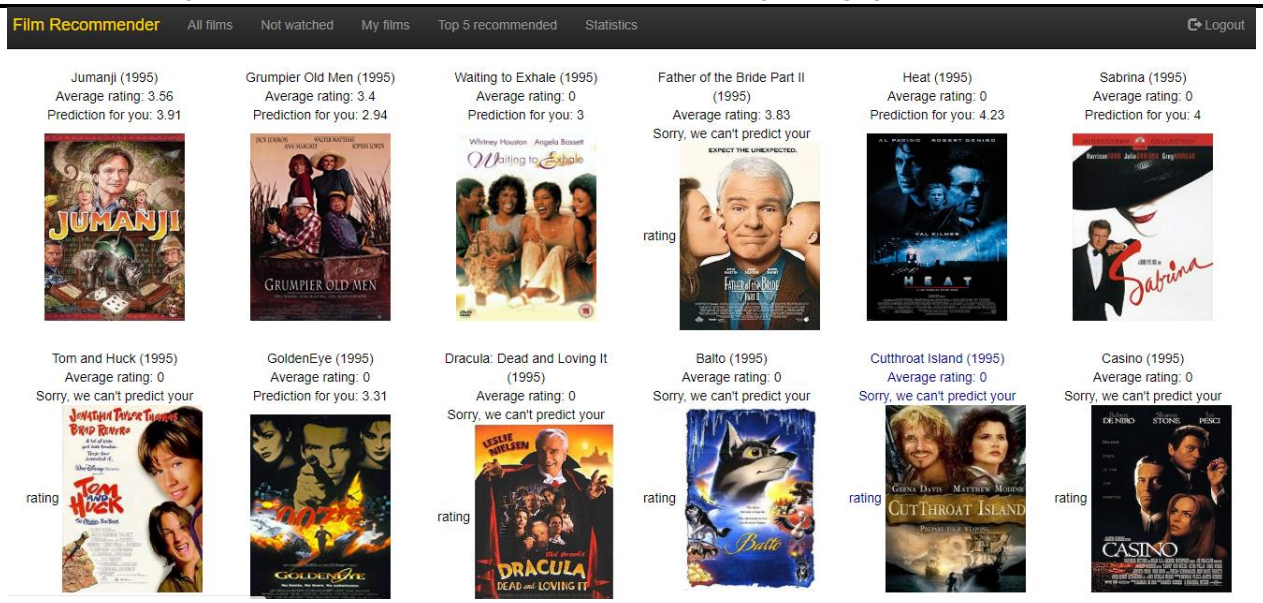

Fig.7 Not Watched Films

When user will click on the film included in Not Watched Film, he can see this page (Fig.8) to rate that particular film. User can see average rating given by all users.

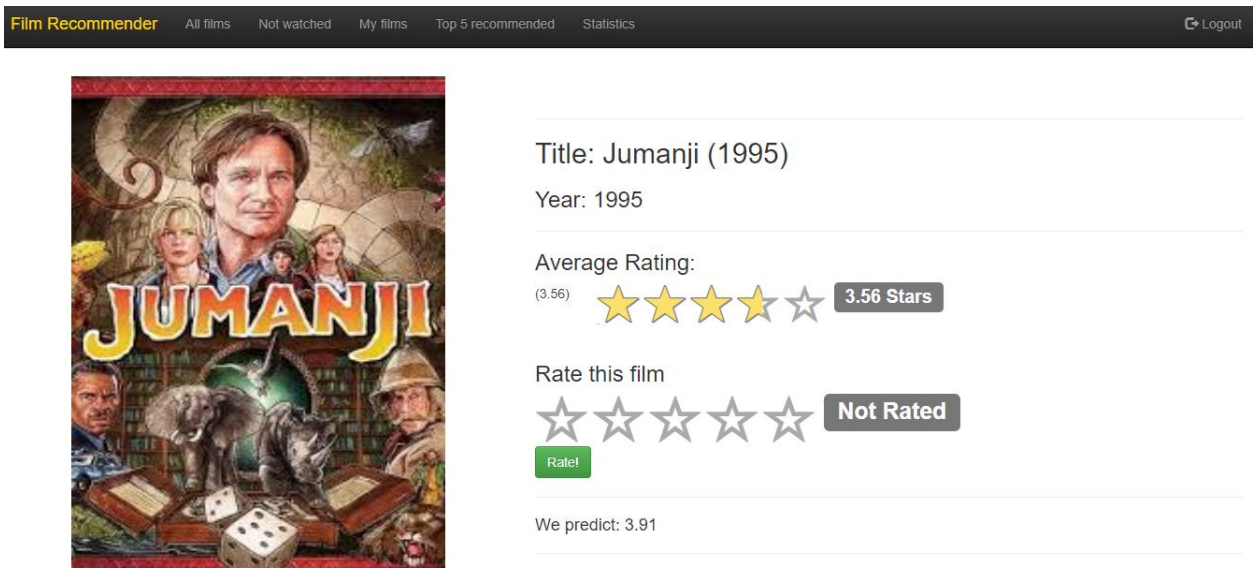

Fig.8 Rate Film

Fig.9 shows all the films rated by user. It is named as My Films. Here user can see average rating given by all users and rating giver by user to that film.

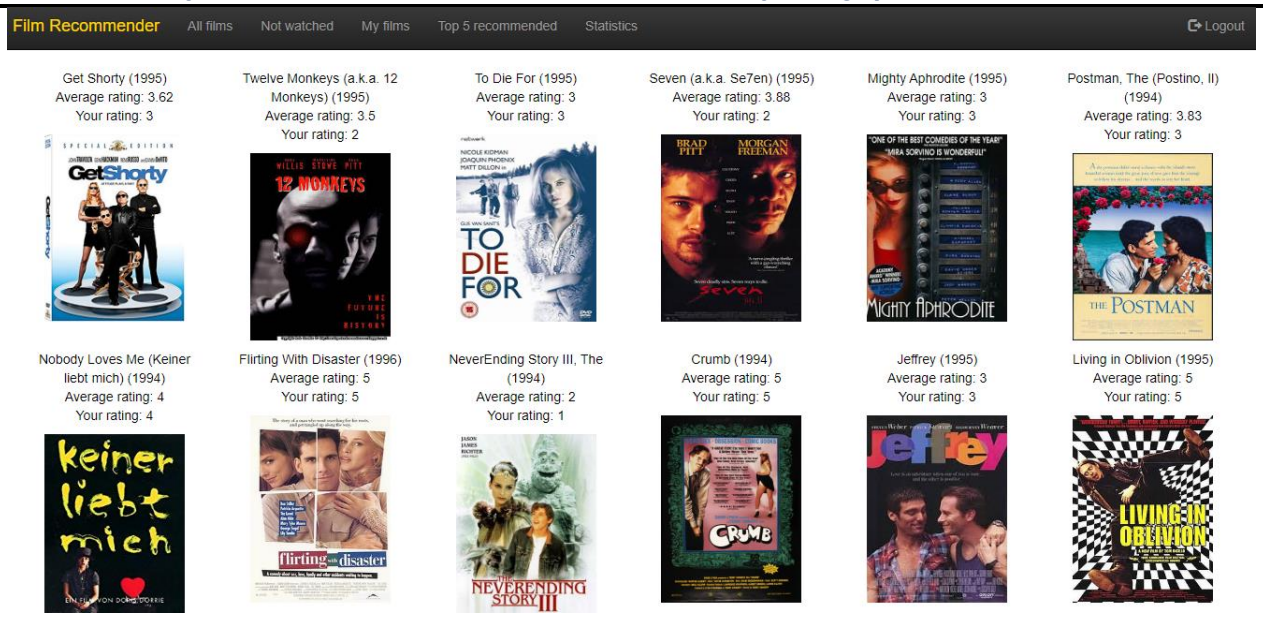

Fig.9 My Films

Fig.10 shows, list of top 5 recommended films for user based on previous preferences and list of films user have not seen yet. User can see average rating and prediction for user for that particular film.

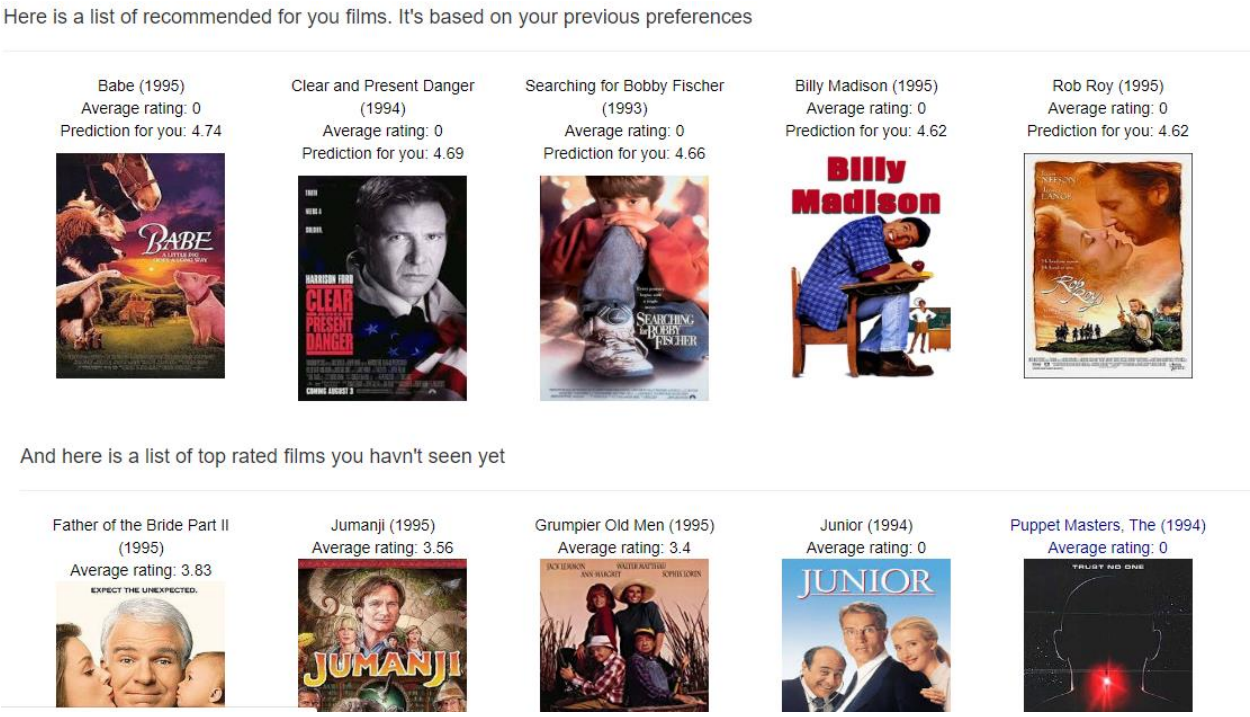

Fig.10 Top 5 Recommended Films

Fig. 11 shows the sentimental analysis for Film Recommender. It contains graph of positive, negative and neutral reviews given by all users. It also shows likes of the user from genre according to categories of movies like action, adventure, crime and so on.

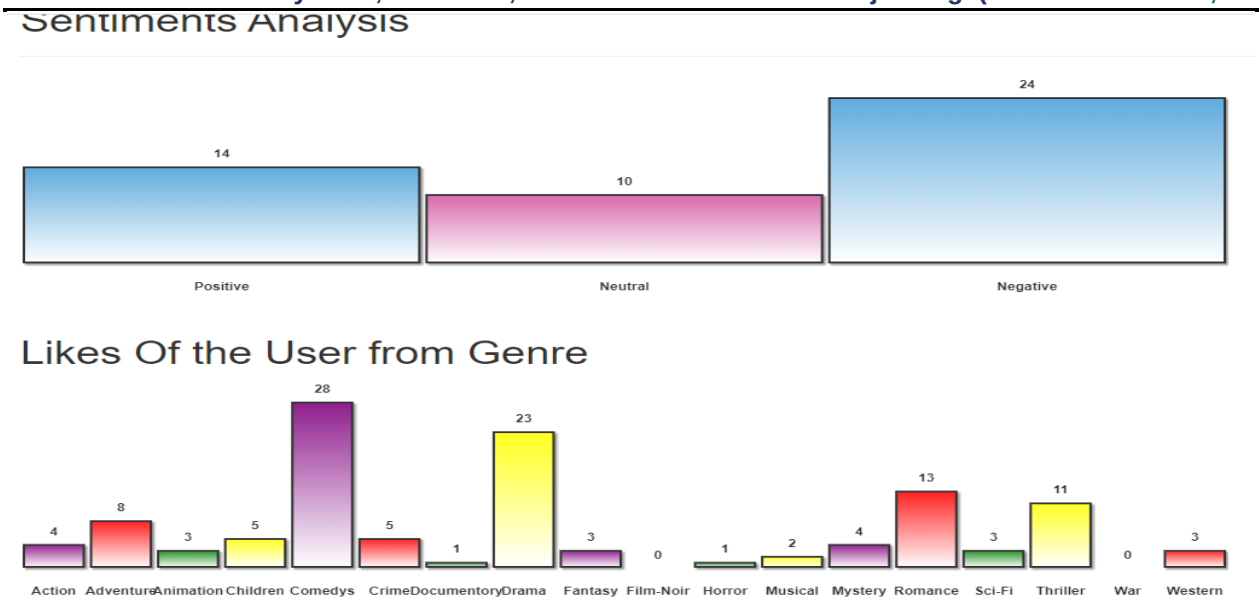

Fig.11 Sentimental Analysis

We have used phpMyAdmin database to store data and XAMPP server for database connectivity.

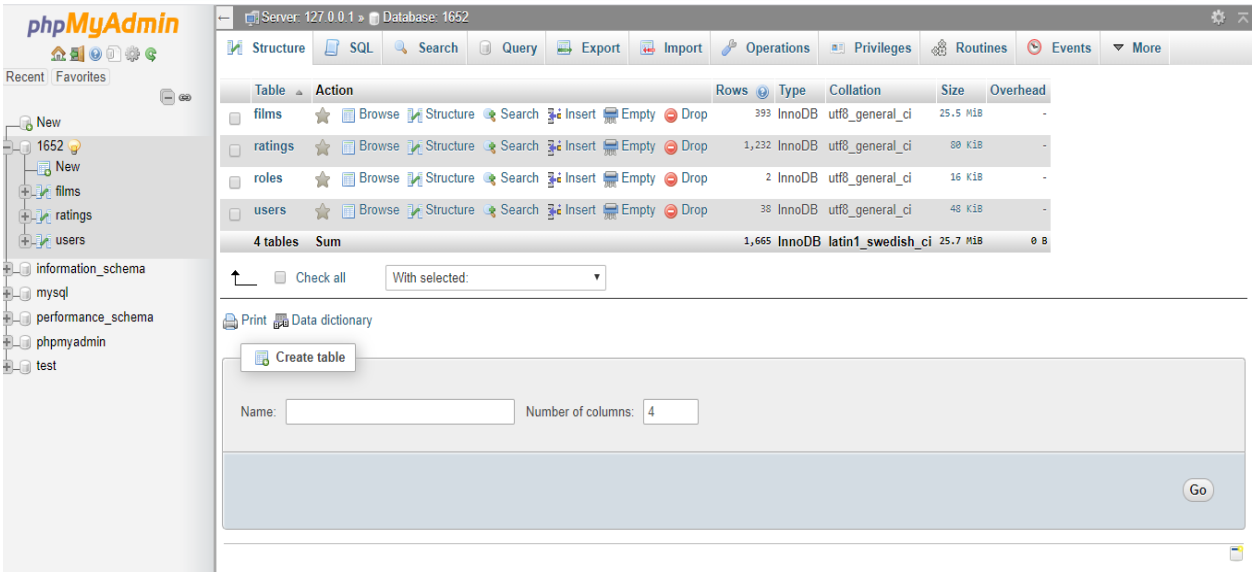

Fig.12 Database

VI. CONCLUSION

So, in this project "Big Data Analysis: Recommendation System with Hadoop Framework" we have developed a movie recommendation system by using Hadoop Framework and Mahout Interfaces. We proposed a movie recommendation system to recommend movies to the users based on their past choices. We used Hadoop framework and Mahout interfaces to analyze big data. Here we used a hybrid filtering technique to filter different types of reviews, opinions, remarks, comments, complains etc. Because recommendations are based on ratings, ranks, content, reviewer's behavior, and timing of review generated by different reviewers. We have implemented statistical analysis on the reviews generated by different users. Recommendation by applying the weightage of summarized reviews and opinions on the rating of item.

REFERENCES

[1]Sayali D. Kadam ; Siddhesh A. Vaidya, "Big data analytics-recommendation system with Hadoop Framework", in International Conference on Inventive Computation Technologies (ICICT), Vol. 3,2016.

[2]S. Saravanan, "Design of large-scale Content-based recommender system using Hadoop MapReduce framework", in Eighth International Conference on Contemporary Computing (IC3), Page s: 302 – 307,2015.

[3]Ye Tao,″ Design of large scale mobile advertising recommendation system ″in 4th International Conference on Computer Science and Network Technology (ICCSNT) , Vol. 01,Page s: 763 - 76, 2015.

[4]Arupananda Girish Prasad ; Lalit Kumar Vashishtha, "Building hybrid recommendation system based on Hadoop framework", in International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), Page s: 3493 - 3499, 2016.

[5]Jai Prakash Verma ;Bankim Patel ;Atul Patel, "Big Data Analysis: Recommendation System with Hadoop Framework", in IEEE International Conference on Computational Intelligence & Communication Technology, IEEE, Page s: 92 – 97, 2015.

[6]Klavdiya Hammond ; Aparna S. Varde, "Cloud Based Predictive Analytics: Text Classification, Recommender Systems and Decision Support" in IEEE 13th International Conference on Data Mining Workshops,IEEE, Page s: 607 – 612,2013.

[7Changjian Fu ; Zhihua Leng, "A Framework for Recommender Systems in E-Commerce Based on Distributed Storage and Data-Mining", in International Conference on E-Business and E-Government, Page s: 3502 – 3505, 2010.

[8]Asadi Srinivasulu ; Ch.D.V. Subbarao ; Jeevan Kumar Y, "High dimensional datasets using Hadoop mahout machine learning algorithms", in International Conference on Computing and Communication Technologies, Page s: 1 – 1, 2014.

[9]Khushboo R. Shrote ; A.V. Deorankar, "Review based service recommendation for big data" in 2nd International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB) , Page s: 470 – 474, 2016.

[10] Lakshay Nagpal ; Khurana, "Design of friend recommender system using apache Hadoop", in IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS) , Page s: 1 – 5, 2017.

[11] P.V. Sai Charan ; P. Ravi Kumar ; P. Mohan Anand, "Addressing Cold Start Problem in Recommendation System Using Custom Built Hadoop Ecosystem", in Second International Conference on Inventive Communication and Computational Technologies (ICICCT), Page s: 355 – 358,2018.

[12] Vijay M Bande ; Ganesh K Pakle," CSRS: Customized service recommendation system for big data analysis using map reduce", in International Conference on Inventive Computation Technologies (ICICT), Page s: 1 – 3, Vol. 3,2016.

[13] Haiyan Hu ; Duo Long, "A Study on Key Techniques of Wisdom Campus Information Recommendation Platform Based on Big Data", in International Conference on Intelligent Transportation, Big Data and Smart City, Page s: 56 – 58,2015.

[14]Reshma Chavan ; Debajyoti Mukhopadhyay," A comparative study of recommendation algorithms in e-commerce", in International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), Page s: 163 – 168,2017.

[15] Rupinder Kaur ; Raghu Garg ; Himanshu Aggarwal," Big data analytics framework to identify crop disease and recommendation a solution", in International Conference on Inventive Computation Technologies (ICICT), Page s: 1 – 5, Vol.2,2016