ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

A Survey on Eat-Out Recommender in Hadoop

Seema Redekar[1], Pinkesh Nayak[2], Poorna Nilkund[3], Ruchi Shetty[4]

Assistant Professor, Department of Information Technology, SIES Graduate School of Technology

Nerul, Navi Mumbai, India[1]

B.E Student, Department of Information Technology, SIES Graduate School of Technology

Nerul, Navi Mumbai, India[2]

B.E Student, Department of Information Technology, SIES Graduate School of Technology

Nerul, Navi Mumbai, India[3]

B.E Student, Department of Information Technology, SIES Graduate School of Technology

Nerul, Navi Mumbai, India[4]

ABSTRACT: Hadoop is an open source framework to process a large amount of data. Even though there are lots of existing recommendation systems it's still an upcoming and unexplored field that has greater potential for better advancements. In this system we have introduced item based collaborative filtering for taking user based ratings and we have proposed Hive to process petabytes of data and Sqoop to transfer data between Hadoop and relational database. The proposed system is reliable and fault tolerant when compared to the existing recommendation systems as it collects the ratings from the user to predict the interest and analyses the item to find the features.

KEYWORDS: Big Data, Hadoop, Recommendation, Hive, Sqoop.

I. INTRODUCTION

Big data refers to a large amount of data which can be analysed to reveal new insights and used for optimised decision making. All the relational databases can handle only structured data but while handling large amount of unstructured data, big data comes into picture. Big data enables the storage of data in an efficient manner. The three main characteristics of big data are volume, velocity and variety. Volume refers to the large amount of data generated from various sources such as government agencies, pharmaceutical industries and social networking. The data generated are in terabytes, petabytes or even more than that. Velocity refers to massive amount of data generated per second. Variety refers to different types of data such as structured, unstructured and semi structured data. The exponential growth of data has led to challenges in storing and processing it. Traditional databases have become inadequate in processing of such data. Hadoop was introduced to solve the above problem. Hadoop is an open source framework used for analyzing huge amount of data in parallel.

In this paper, the proposed system recommends restaurants using item based collaborative filtering algorithm. Data processing will be done using Hive and Sqoop will be used for transferring data between Hadoop and relational database. Section 1 describes the introduction. Section 2, the literature survey. Section 3 describes the need of the proposed system and section 4, the proposed system.

DOI: 10.15680/IJIRCCE.2017. 0503038

II. RELATED WORK

Authors in [1] have used map reduce. It has three mapper class and reducer class because of which the performance and scalability of the system has increased. In this they have compared the user based collaborative filtering algorithm and item based collaborating filtering algorithm with the single node and multiple node cluster. Pearson correlation is used to calculate the similarity. Thus Hadoop works well for large amount of dataset.

As mentioned in literature Survey [2] the issues in big data handling and how to solve this problem using Hadoop cluster, Hadoop distributed file system and map reduce framework has been discussed. Hadoop does parallel processing to process large amount of data. Data used today in an organization is growing rapidly in a range of terabytes to petabytes of data. The issues in the big data is to handle large amount of data using relational databases. As it faced problems like scalability and fault tolerance to handle unstructured data, Apache Hadoop is an open source software which is reliable, scalable and can be used in distributed computing. In this the master node divides the problem into sub problem and then sends results to the master node. Key value pair is maintained. In HDFS the data is stored in file system. A file is split into one or more blocks. DataNode stores these blocks. NameNode determines the mapping of blocks to Datanode [2]. For preventing file system corruption and reducing loss of data, Secondary NameNode generates snapshots of the name node's memory structure. Authors in this have performed experiment on the text processing analysis and earthquake data analysis.

Authors in [3] have shown the difference between the traditional data analytics and big data analytics. Hadoop is an analytical software to handle large amount of unstructured data. Application of the big data analytics are business, social and medical application. With the advancement of big data, large amount of the data can be accessed very quickly.

As stated in literature survey [4] analysis of parameters such as amount of time taken by the map and the reduce on the big data using Hadoop distributed file system and the map reduce is done. Memory utilization of the mapper and the reducer are calculated for storage and processing. Data are of three types : structured, semi structured and unstructured data. Big data processing requires fast retrieval so that it can speed up the process. Hadoop is a tool that provides the storage to big data. Apache Hadoop uses both distributed file system and map reduce. Authors in this have analyzed different memory parameters using snapshot. Different memory parameters are virtual memory, heap usage and physical memory and the information is stored in the form of snapshot.

Authors in [5] have developed shuffling strategy to improve scalability and efficiency of word processor [5]. Shuffling is done between mapping and reducing phases and is used to globally exchange the intermediate data generated by mapping. It has been developed in distributed platform. Structured data is represented in spreadsheet and relational databases. Unstructured data is not represented in relational databases eg. Images, audio, video, email and word processing document. Semi structured data are XML and mark up languages [6]. Big data has features such as volume, velocity, variety, value and veracity. [7-8]. Volume consists of large amount of data. Velocity is the data collection and analysis to be performed fast and in time. Variety is the different types of data such as semi structured and unstructured data like audio, video, webpage and text. Value is the cost of collected or generated data. Veracity of data is to eliminate noise through sanitization to check the data accuracy. Big data processing framework is Hadoop open source framework whereas programming framework is map reduce. The text file was read and the occurrences of each word were counted by the authors. Key value pair of each word was prepared by mapper function by taking each line as input and divides it into words. Shuffling has been done using sorting by value as per the occurrences of each word. This is the input to the reducer. It sums the count of each word and produces a single value for a word. Reducer is used to combine the data. Hadoop distributed file system stores both the input and output file system. If the file is not in HDFS then it is fetched from the local file system.

As mentioned in literature Survey [9] , big data is a solution to the organization and extracts the accurate information from the data in less time span. Hadoop allows distributed processing for large amount of dataset. Authors in this have considered log file for extraction of information. It extracts meaningful information from access log file which is a text file.

Authors in [10] made a study on the Hadoop distributed File System. The study stated that by distributing the storage and computation across the machines of a cluster, the computational time can be reduced for analyzing big data when compared to single node processing.

III. NEED OF THE SYSTEM

- As the numbers of restaurants have grown, the need for recommendation system has also increased.
- For any online shopping portal, the main aim is user satisfaction, which goes for a toss without the recommendation system.
- Also without the recommendation system, the page would be very haphazard which would lead to great difficulty in locating the restaurant as per user's interest. This might lead to great disinterest among the users.
- Recommendation system tries to predict the interest of a user and recommends restaurants that match their interest correctly. Also, e-commerce business will be profited by the increase of sales.

IV. PROPOSED SYSTEM

User preference will be taken from web site and stored in .csv file or text file. This text file will be processed using hive script containing item based collaborative filtering algorithm. Hive is built on top of hadoop to analyze large amount of data and it uses a language called Hive querying language which is similar to SQL[11]. It can be used for analyzing both structured as well as unstructured data. The processed result will be stored in hadoop distributed file system (HDFS). To transfer these results from HDFS to relational database, Sqoop is used [12]. Sqoop is used to transfer data from HDFS and relational databases. From the results loaded in the database, the restaurant will be recommended to the user. The system architecture is shown in figure 4.1.

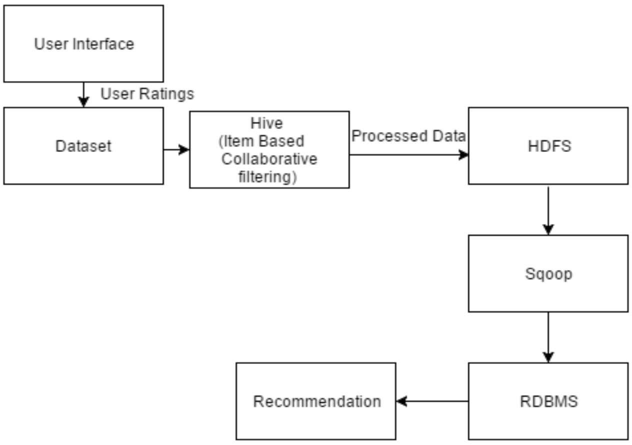

Figure 4.1: System Architecture

The flow of the system is shown in figure 4.2. When a user registers or logins, he can select any restaurant from the website. Once the restaurant is selected, the user can give his preference in the form of ratings. All these records of corresponding user & restaurant are recorded. A dataset is created in csv format with the attributes user id, restaurant id, ratings. Dataset is processed by running Hive script containing item based collaborative filtering algorithm.

Processed data will be stored in Hadoop Distributed File System (HDFS). Sqoop command is executed for transferring the processed data from HDFS to Relational database (RDBMS). The restaurants corresponding to restaurant selected by user are selected based on highest restaurant pair value & then recommended to the user. Advantages of Proposed System are scalable, cost effective and flexible.

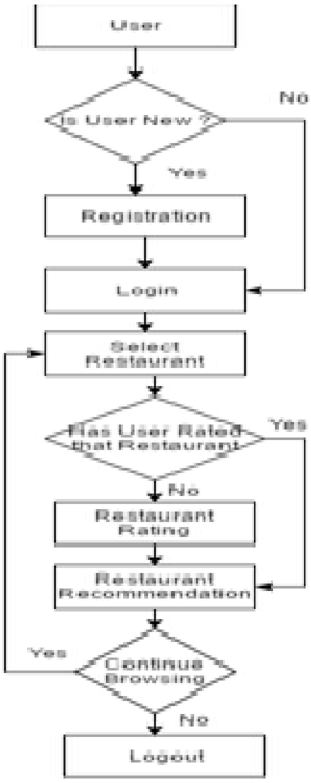

Figure 4.2 Flowchart of the proposed system

V. CONCLUSION

Over two decades of research and commercial development, recommendation systems have proved to be a successful technology to overcome the information overload that burdens users in modern online media. This recommendation system is to be built for recommending the restaurants to the users according to ratings of other users. This work can be extended for movies recommendation, music recommendation, website recommendation etc. This system uses Hadoop Mapreduce framework to get better and valuable insights from Big Data using Item based Collaborative filtering algorithm to analyze customer interests through ratings, and then analysis will be done to recommend the restaurants efficiently. The proposed system is more efficient and scalable as compared to the traditional system.

REFERENCES

1. Poonam Ghuli, Atanu Ghosh and Rajashree Shettar,” A Collaborative Filtering Recommendation Engine in a Distributed Environment”, IEEE International Conference on Contemporary Computing and Informatics (IC3I), 2014.
2. Aditya B. Patel, Manashvi Birla and Ushma Nair,”Addressing Big Data problem using Hadoop and Map Reduce”, IEEE International Conference on Engineering (NUiCONE) Nirma University, 2012.
3. Manjula Sanjay and B.H. Alamma, “An Insight into Big Data Analytics – Methods and Application”, IEEE International Conference on Inventive Computation Technologies (ICICT), 2016.
4. Amrit Pal, Sanjay Agrawal,” An Experimental Approach Towards Big Data forAnalyzing Memory Utilization on a Hadoop cluster using HDFS and MapReduce”, IEEE 1st International Conference on Networks & Soft Computing (ICNSC), 2014.

5. B. Mandal,S.Sethi and R.Kumar Sahoo," Architecture of efficient word processing using Hadoop MapReduce for big data applications", IEEE International Conference on Man and Machine Interfacing (MAMI), 2015.
6. Han Hu, Yong Gang Wen, Tat - Seng Chua and Xuelongl "Towards Scalable System for Big Data Analytics", IEEE, 652-687, June(2014)
7. M. Ferguson, "Architecting a big data platform for analytics", A whitepaper prepared for IBM, Armonk, New York (2012).
8. Shilpa and Manjit Kumar. "Big Data And Methodology-A Review", IJARCSSE, 991-995 Vol-3, October,(2013).
9. B. Kotival, A.Kumar, B.Pant and R.H.Goudar," Big Data: Mining of Log File through Hadoop", IEEE International Conference on Human Computer Interactions (ICHCI), 2013.
10. Konstantin Shvachko, Hairong Kuang, Sanjay Radia and Robert Chansler, "The Hadoop Distributed File System", IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST), 2010.
11. Chunk Lam,"Hadoop In Action" Available:[Online] https://books.google.co.in/books?isbn=1617291226.
12. TomWhite," HadoopThe.Definitive.Guide."3rd.Edition. Available: [Online]
13. http://www.isical.ac.in/~acmsc/WBDA2015/slides/hg/Oreilly.Hadoop.The.Definitive.Guide.3rd.Edition.Jan.2012.pdf.