Data Analysis of Mobile Shopping Using Hadoop Inspired Mapreduce

Ms. Kalyani Jiwtode[1], Mr. Hirendra Hazare[2]
[1]Dept of CSE
[2]Assit. Professor, Dept of CSE
[1, 2] Ballarpur Institute of Technology (BIT), Ballarpur.

Abstract- *We live in on-demand, on command digital universe with data prolife ring by institution, individuals and machines at a very high rate. This data is categories as "Big Data" due to its sheer volume, variety and velocity. Most of this data is unstructured, quasi-structured or semi structured and it is heterogeneous in nature. The volume and the heterogeneity of data with the speed it is generated, makes it difficult for the present computing infrastructure to manage Big Data. Traditional data management, warehousing and analysis system fall short of tools to analyze this data. Due to its specific nature of Big Data, it is stored in distributed file system architectures. Hadoop and HDFS by Apache is widely used for storing and managing Big Data. Analyzing Big Data is a challenging task as it evolved large distributed file system.*

Keywords- Big Data, HDFS, Map Reduced, Cluster.

I. INTRODUCTION

Data mining is one of the most prominent areas in modern technologies for retrieving meaningful information from huge amount of unstructured and distributed data using parallel processing of data. There is huge advantage to Educational sector of following Data Mining Techniques to analyse data input from students, feedbacks, latest academic trends etc which helps in providing quality education and decision-making approach for students to increase their career prospects and right selection of courses for industrial trainings to fulfil the skill gap pertains between primary education and industry hiring students. Data Mining has great impact in academic systems where education is weighed as primary input for societal progress.

Big data is the emerging field of data mining. It is a term for datasets that are so large or complex that traditional data processing application software is incompetent to deal with them. Big data includes gathering of data for storage and analysis purpose which gain control over operations like searching, sharing, visualization of data, query processing, updating and maintain privacy of information. In Big data, here is extremely large dataset that is analysed computationally to reveal patterns, trends and associations. It deals with unstructured data which may include MS Office files, PDF, Text etc whereas structured data may be the relational data. Hadoop is one technique of big data and answer to problems related to handling of unstructured and massive data. Hadoop is an open-source programming paradigm which performs parallel processing of applications on clusters. Big Data approach can help colleges, institutions, universities to get a comprehensive aspect about the students. It helps in answering questions related to the learning behaviours, better understanding and curriculum trends, and future course selection for students which helps to create captivating learning experiences for students. The problem of enormously large size of dataset can be solved using Map Reduce Techniques. Map Reduce jobs run over Hadoop Clusters by splitting the big data into small chunks and process the data by running it parallel on distributed clusters.

II. OBJECTIVE

The main objective to build Semantic Similarity Based Rank Boosting Approach on Hadoop using Map Reduce for Big data applications.

Motivation behind this project is that with the success of the Web 2.0, more and more companies capture large-scale information about their customers, providers, and operations. The rapid growth of the number of customers, services and other online information yields service recommender systems in "Big Data" environment, which poses critical challenges for service recommender systems. Moreover, in most existing service recommender systems, such as hotel reservation systems and restaurant guides, the ratings of services and the service recommendation lists presented to users are the same. They have not considered users' different preferences, without meeting users' personalized requirements.
Objective:

- To Present a personalized Service recommendation list and recommending the most appropriate services to the users effectively

- Semantic similarity-based approach is used for finding keywords which are having similar meaning for more accuracy
- Distinguish the positive and negative preferences of the users from their reviews to make predictions more accurate.

III. LITERATURE SURVEY

Implementation of Hadoop Operations for Big Data Processing in Educational Institutions: Education plays an important role in maintaining the economic growth of a country. The objective of this paper is to focus on the impact of cloud computing on educational institutions by using latest big data technology to provide quality education. Our educational systems have a large amount of data. Big Data is defined as massive sets of data that is so large or so complex that it is very difficult to process by using conventional applications and software technologies. This has resulted in the penetration of Big Data technologies and tools into education, to process the large amount of data involved. In this paper we discuss what Cloud and Hadoop is, and its types, operations and services offered. Hence it has an advantage which will surely help the students when used in an appropriate way.

Predicting Student Performance Using Map Reduce: Data mining and machine learning depend on classification which is the most essential and important task. Many experiments are performed on Student datasets using multiple classifiers and feature selection techniques. Many of them show good classification accuracy. The existing work proposes to apply data mining techniques to predict Students dropout and failure. But this work doesn’t support the huge amount of data. It also takes more time to complete the classification process. So, the time complexity is high. To improve the accuracy and reduce the time complexity, the Map Reduce concept is introduced. In this work, the deadline constraint is also introduced. Based on this, an extensional Map Reduce Task Scheduling algorithm for Deadline constraints (MTSD) is proposed. It allows user to specify a job’s (classification process in data mining) deadline and tries to make the job to be finished before the deadline. Finally, the proposed system has higher classification accuracy even in the big data and it also reduced the time complexity.

Map Reduce as a Programming Model for Association Rules Algorithm on Hadoop as association rules widely used, it needs to study many problems, one of which is the generally larger and multi-dimensional datasets, and the rapid growth of the amount of data. Single processor’s memory and CPU resources are very limited, which makes the algorithm performance inefficient. Recently the development of network and distributed technology makes cloud computing a reality in the implementation of association rules algorithm. In this paper we describe the improved Apriori algorithm based on Map Reduce mode, which can handle massive datasets with a large number of nodes on Hadoop platform.

IV. PROPOSED WORK

Service recommendation method, for user’s personalized requirements, is proposed in this paper, which is based on a user-based Collaborative Filtering algorithm. In KASR, keywords extracted from reviews of previous users are used to indicate their preferences. Moreover, we implement it on Hadoop MapReduce as its computing framework. In KASR, keywords are used to indicate both of users' preferences and the quality of candidate services. A user-based CF algorithm is adopted to generate appropriate recommendations. KASR aims at calculating a personalized rating of each candidate service for a user, and then presenting a personalized service recommendation list and recommending the most appropriate services to him/her. Moreover, to improve the scalability and efficiency of our recommendation method in, we implement it by splitting the proposed algorithm into multiple Map Reduce phases.

1. Big Data and Environment:

Huge Collection of data is retrieved from open source datasets that are publicly available from major Travel Recommendation Applications. Big Data Schemas were analyzed and a Working Rule of the Schema is determined. The CSV(Comma separated values) files were read and manipulated using Java API that itself developed by us which is developer friendly ,light weighted and easily modifiable.

2. Batching and Preprocess:

The Traditional View of Service Recommender Systems that shows Top-K Results are displayed with Paginations with which a user can navigate Back and Forth of the Result sets. All Services Ratings and Reviews of Each Hotels are listed. Parts of Speech Tagger and Chucker Process are done on each and every review of all hotels for all countries in a Parallel and Distributed Manner as Batch jobs. The Master Job is Split up into ‘n’ no of small Batch jobs based on the slave machines Connected with the Master. POS Tagger tags each words of a review with its tags and the Clunker Process will take POS tagged output as input for Groping the Words based on meaning of the Review.

3. Digging in Big Data & Service Recommender Application:

The CSV Files in distributed Systems are invoked through Web Service Running in the Server Machine of the Host Process through a Web Service Client Process in the Recommendation System. The data that Retrieved to the Recommendation Systems are provided with a clean GUI and can be queried on Demand. Each and Every process on the Recommendation Application invokes Web Service which uses light weighted traversal of data using XML. The Users can Review each hotel and can post comments also. The Reviews gets updated to the CSV Files as it get retrieved. A User can scan or schedule a Travel highlighting his requirements in a detailed way that shows the Preference Keywords Set of the Active User. A Domain Thesaurus is built depending on the Keyword Candidate List and Candidate Services List. The Domain Thesaurus can be Updated Regularly to get accurate Results of the Recommendation System.

V. METHODOLOGY

MapReduce and Hadoop Project Flow

1. The logins into system.

2. Admin Panel

User sets the number of clusters, so for simulation on to the computer, if users set the 4 number of clusters, so data will be divided into 4 part and will be transfer two 4 client machines.

3. User uploads the dataset.

Then by applying algorithm, the file gets spitted to four clusters, i.e. folders. The mapper function makes the key value pairs and gives to the Reducer. The Reducer will take those key value pairs, processes it, aggregate the data to get the combine results The mapping will be present on separate cluster, that, on which cluster what type of data is available on which cluster The user search for the particular data, analyses the mappings and asks the particular to get the data.

VI. CONCLUSION

The Map Reduce approach is used for running jobs over HDFS. Using Map Reduce, the application can be scaled to run over multiple machines in a cluster and for that Hadoop cluster is used. The Map Reduce Framework consists of Map and Reduce Functions with single Resource Manager which acts as a master and one Node manager which acts as slave per cluster node.

REFERENCES

[1] Pratiyush Guleria ; Manu Sood,” Big data analytics: Predicting academic course preference using hadoop inspired mapreduce”, 2017 Fourth International Conference on Image Information Processing (ICIIP), Date Added to IEEE Xplore: 12 March 2018

[2] Sonali Agarwal, G. N. Pandey, M. D. Tiwari, “Data Mining in Education: Data Classification and Decision Tree Approach”, International Journal of e-Education, e-Business, e-Management and e-Learning, Vol. 2, No. 2, April 2012.

[3] Jongwook Woo, "Apriori-Map/Reduce Algorithm." Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2012.

[4] Xin YueYang,Zhen Liu,Yan Fu, “MapReduce as a Programming Model for Association Rules Algorithm on Hadoop”, Information Sciences and Interaction Sciences (ICIS), 2010 3rd International Conference on, pp. 99-102. IEEE, 2010.

[5] Katrina Sin,Loganathan Muthu,"Application of Big Data in Education Data Mining and Learning Analytics – A Literature Review" ,ICTACT Journal on Soft Computing,ISSN: 2229-6956 (online), Vol 5, Issue 4,July 2015.

[6] B. Manjulatha, Ambica Venna, K. Soumya, "Implementation of Hadoop Operations for Big Data Processing in Educational Institutions",International Journal of Innovative Research in Computer and Communication Engineering,ISSN(Online) : 2320-9801,Vol. 4, Issue 4, April 2016.