IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 8, Issue 2, February 2019

Data Analysis with Hadoop

Bhawna Makhija[1], Chhaya Amrute[2], Durga Nandan Mishra[3], Rashmi Tembhurne[4], Vijay V. Chakole[5]

Electronics Department, K.D.K.C.E, Nagpur[1,2,3,4,5]

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 inn 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 are widely used for storing and managing Big Data. Analyzing Big Data is a challenging task as it involved large distributed file system.

Keywords: Big Data, HDFS, Map Reduced, Cluster

I. INTRODUCTION

The amount of data in the world has been increasing exponentially. This data in petabytes of amount is called “big data”. Big data is an evolving term that describes any voluminous amount of structured, semi-structured and unstructured data that has the potential to be mined for information. Although big data doesn't refer to any specific quantity, the term is often used when speaking about petabytes and exabytes of data. Analysis of such a large amount of data is a challenge for IT companies. So the solution is to provide more manageable softwares. Big data also brings new opportunities and challenges in IT companies, Ecommerce and academia. There are many alternative recommendation services but effectively recommending services are need of time. These are the valuable tools to help users deal with services overload. Examples of such practical applications are existing customer records to predict trends, social media logs, CDs, EBooks, webpages, gadgets, video and music streaming or even food. For example, large retailer might have huge amounts of data, tens of millions of customers and millions of distinct catalog items. Many applications require the results set to be returned in realtime, in no more than half a second, while still producing high-quality recommendations. New customers typically have extremely limited information, based on only a few purchasesor product ratings. Older customers can have a glut of information, based on thousands of purchases and ratings. Customer data is volatile: Each interaction provides valuable customer data, and the algorithm must respond immediately to new information. So, an efficient service recommendation system is needed. 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. Most existing service recommender systems are only based on a single numerical rating to represent a service's utility as a whole. In fact, evaluating a service through multiple criteria and taking into account of user feedback can help to make more effective recommendations for the users. Existing Approaches solve the scalability problem by dividing dataset. But their method doesn't have favorable scalability and efficiency if the amount of data grows.

Motivated by these observations, In paper we have addressed these challenges through the following contributions:

(1) A keyword aware service recommendation method named KASR, is proposed in this paper which is based on user-based Collaborative Filtering (CF) algorithm.

(2) Keywords extracted from reviews of previous users are used to indicate their preferences For efficiency and scalability we implement it on distributed computing platform hadoop.

Hadoop uses MapReduce programming as a computing framework. Most recommendation algorithms start by finding a set of customers whose purchased and rated items overlap the user’s purchased and rated items. The algorithm aggregates items from these similar customers, eliminates items the user has already purchased or rated, and recommends the remaining items to the user. Two popular versions of these algorithms are collaborative filteringand cluster models. Other algorithms including search-based methods and item-to-item collaborative filtering focus on finding similar items, not similar customers. For each of the user’s purchased and rated items, the algorithm attempts to find similar items. It then aggregates the similar items and recommends them.

DOI 10.17148/IJARCCE.2019.8242

II. AIM AND OBJECTIVE

The main aim is 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. 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 HadoopMapReduce 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.

DOI 10.17148/IJARCCE.2019.8242

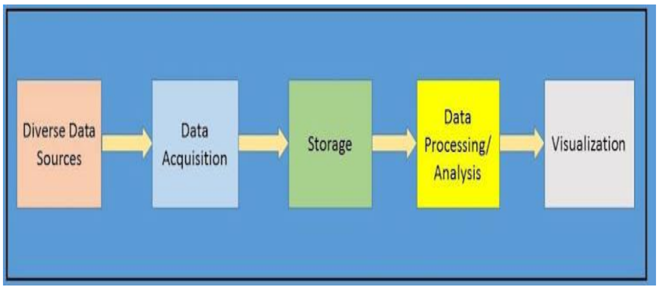


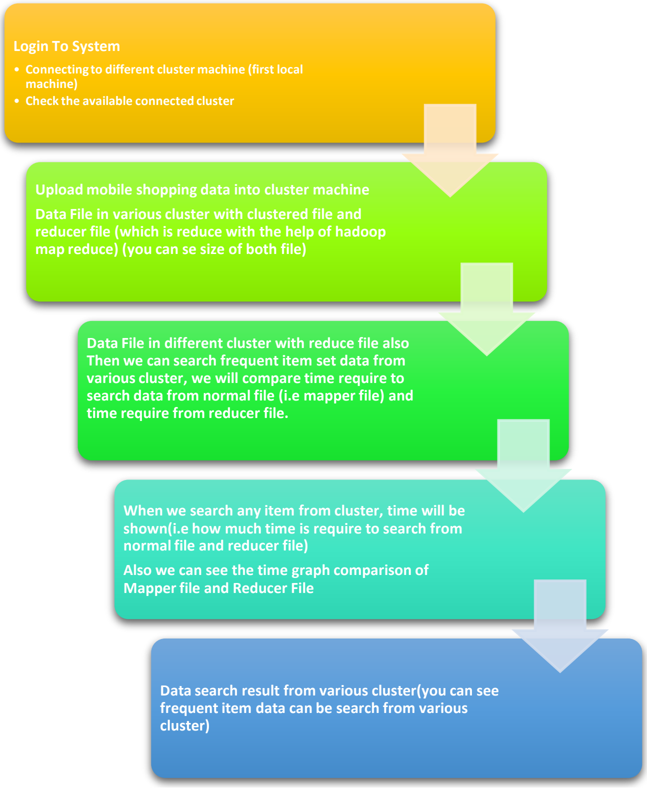
Fig.(1):-Methodology

IV. MAPREDUCE AND HADOOP PROJECT FLOW

1. The user logins into the 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

V. DATA DOWNLOADING ALGORITHM



Login To System
- Connecting to different cluster machine (first local machine)
- Check the available connected cluster

Upload mobile shopping data into cluster machine

Data File in various cluster with clustered file and reducer file (which is reduce with the help of hadoop map reduce) (you can se size of both file)

Data File in different cluster with reduce file also Then we can search frequent item set data from various cluster, we will compare time require to search data from normal file (i.e mapper file) and time require from reducer file.

When we search any item from cluster, time will be shown(i.e how much time is require to search from normal file and reducer file)

Also we can see the time graph comparison of Mapper file and Reducer File

Data search result from various cluster(you can see frequent item data can be search from various cluster)

Fig.(2):-Flow Chart

VI. CONCLUSION

Big Data is a data whose scale, ,and complexity require new architecture ,algorithms, and analytics to manage it and extract value and hidden knowledge from it. Today, Data is generated from various different sources and can arrive in the system at various rates. To process these large amounts of data is a big issue today. In this paper we discussed Hadoop tool for Big Data in details. Hadoop is the core platform for structuring Big Data, and solves the problem of

DOI 10.17148/IJARCCE.2019.8242

making it useful for analytic purposes. We also discussed some hadoop components which are used to support the processing of large data sets in distributed computing environment. In future we can use some clustering techniques and check the performance by implementing it in hadoop.

REFERENCES

[1]. Prathyusha Rani Merla; Yiheng Liang, "Data analysis using HadoopMapReduce environment" ,IEEE International Conference on Big data,2017.
[2]. Jinhua chen; Jing Tang, "Reaserch on architecture of education big data analysis system", IEEE 2nd International Conference on Big data analysis, 2017.
[3]. Chen Chen, "Construction of big data analysis and decision support platform based on cloud computing", Library theory and practice, vol. 05, pp. 101-104, 2016.
[4]. Hongtao Sun, QinhuaZheng, "The core technology application status and development trend of educational data", Journal of distance education, vol. 05, pp. 41-49, 2016.
[5]. Xueqi Cheng, Xiaolong Jin, Jing Yang, Jun Xu, "The progress and development trend of big data technology", Science and technology review, vol. 14, pp. 49-59, 2016.
[6]. Xuelong Li, "Overview of big data systems", Chinese Science: Information Science, vol. 01, pp. 1-44, 2015.
[7]. Ankita Saldhi; Dipesh Yadav, "Big data analysis using Hadoop cluster",IEEE International Conference on computational Intelligence and Computing Research,2014 .
[8]. Shankar ganesh manikandan ; Siddarth Ravi, "Big data analysis using Apache Hadoop" International Conference On it Convergence and Security (ICITCS),2014.
[9]. Xueqi Cheng, Xiaolong Jin, Yuanzhuo Wang, JiafengGuo, tie win Zhang, Guojie Li, "Big data system and analysis technology", Journal of software, vol. 09, pp. 1889-1908, 2014.
[10]. DunrenChe, MejdlSafran, and ZhiyongPeng, From Big Data to Big Data Mining: Challenges, Issues, and Opportunities, DASFAA Workshops 2013, LNCS 7827, pp. 1-15, 2013.CrossRef Google Scholar
[11]. XiaofengMeng, Xiang Ci, "Big data management: concepts technologies and challenges", Computer research and development, vol. 01, pp. 146-169, 2013.
[12]. Brad Brown, Michael Chui, and James Manyika, Are you ready for the era of big data-, McKinseyQuaterly, Mckinsey Global Institute, October 2011.
[13]. Jefry Dean and Sanjay Ghemwat, MapReduce:A Flexible Data Processing Tool, Communications of the ACM, Volume 53, Issuse.1, January 2010, pp 72-77.
[14]. Qi Zhang, Lu Cheng, RaoufBoutaba, "Cloud computing: state-of-the-art and research challenges", *Journal of Internet Services and Applications*, vol. 1, pp. 7-18, 2010.
[15]. Jefry Dean and Sanjay Ghemwat, .MapReduce: Simplified data processing on large clusters, Communications of the ACM, Volume 51 pp. 107-113, 2008
[16]. I. Foster, C. Kesselman, J. M. Nick, S. Tuecke, "Grid services for distributed system integration", *IEEE journals in Computer Science*, vol. 35, no. 6, pp. 37-46, 2002.
[17]. Byung-Hoon Park, HillolKargupta, "Distributed Data Mining: Algorithms Systems and Applications", *CiteSeerX*, pp. 341-358, 2002.
[18]. Statistics and Social Network of you Tube Videos, [online] Available:http://netsg.cs.sfu.ca/youtubedata/.
[19]. Apache Hadoop, [online] Available:https://en.wikipedia.org/wiki/Apache_Hadoop.
[20]. Multi Node Cluster Setup on AWS, [online] Available:https://blog.insightdatascience.com/spinning-up-a-free-hadoop-cluster-step-bv-step-c406d56bae42

DOI 10.17148/IJARCCE.2019.8242