International Journal for Modern Trends in Science and Technology
Volume: 03, Issue No: 04, April 2017
ISSN: 2455-3778
http://www.ijmtst.com

Comparative Retail Data Analysis in Marketplaces using Hadoop

Prof A S Devare[1] | Prof N B Chopade[2] | Madhuri Mali[3] | Samridhi Jaiswal[4]

[1-4]Department of Computer Engineering, JSPM's JSCOE, Hadapsar, Pune, India.

To Cite this Article
Prof A S Devare, Prof N B Chopade, Madhuri Mali and Samridhi Jaiswal, "Comparative Retail Data Analysis in Marketplaces using Hadoop", *International Journal for Modern Trends in Science and Technology*, Vol. 03, Issue 04, 2017, pp. 124-130.

ABSTRACT

E-commerce portals are now trending in India. It is spreading in every place and customers are showing interest in using this portal effectively. While in this time, business of marketplaces are decreasing as they can't reach up to users. Therefore, we will be developing a portal search engine where all updates and advertisement of marketplaces will be available to promote their products. In other words, there will be "one stop" for all marketplaces. In this we can probe, screen, select, preserve product easily. It will also give the information related to availability and minimal assessment by comparing that product in different market places. And this record would help us to refined the product available based on the user input. It also is screening the product reviews and ratings. All this will contain big data. Apache Hadoop is open framework for distributed processing system can process large volume of data and then it will be processed using MapReduce technique where our database will be in HBase. Our goal is originally aims to provide consumers more information and to make them interactive with market places.

KEYWORDS: *MapReduce, HBase, Hadoop, E-commerce.*

I. INTRODUCTION

Hadoop is an open source framework for distributed storage and distributed processing of very large data sets on computer clusters built from commodity hardware. It was developed by "Apache Software Foundation". It is written in java and uses cross platform operating system. The core part of Apache Hadoop consists of a storage part known as Hadoop Distributed File System (HDFS) and a processing part called Mapreduce. Electronic Business which is also known as e-business, is the online business. It can also be defined as the business which is done with the help of internet or electronic data interchange. It is not confined to buying and selling of goods only, but it includes either activity like providing services to the customers, communicating with employees-commerce is the major part of e-business. Ecommerce is business transaction through electronic means, including internet, telephones, television and computer.

Nowadays it is growing in every place and customers are showing interest in using these portals effectively. In 2010 United Kingdom had biggest market in the world when measured. Now, it has become an important tool for small and large business worldwide, not only to sell to customers but also to engage them. It includes applications like online shopping, online tracking, online banking, electronic tickets, social networking etc.

Types of E-business models-

1] B2C: The business to consumer, or B2c, model of e-business sells products directly to retail consumers online.Amazon.com is an example of B2C model the e-business has only an online identity through which it offers a range of products to consumers.

2] B2B: The business to business, or B2B, model involves companies using the internet to conduct transaction with one other.B2B e-business accounts for more than 90 percent of all electronic commerce, according the U.S. These transactions are multifaceted and often involve multiple transactions each step of the supply chain.

3] C2B: Consumer to business, or C2B, is a unique e-business model in which consumers create value and demand for products. Reverse auctions are common characteristics of C2B models, in which consumers drive transactions and offer their own prices for products. The airline ticket website Priceline.com is an example of a C2B e-business model.

4] C2C: Consumer to Consumer, or C2C, e-business models enables consumers to behave as buyers and sellers in third-part-facilitated online marketplaces. The company brings together disparate buyers and sellers to conduct business.

As we can see due to this, trend of online shopping is increasing day by day in today's time. Websites have been implemented for the purchase of anything "online" using different sources by sitting at their own place. Like, firstly if we take an example of Amazon, it is nothing but the largest internet based retailer in the world. It was founded on July 5, 1994 at Seattle, Washington, U.S. It serves worldwide. It has separate retail websites in India. Secondly flipchart is India's the largest online book retailer. It took the initiative from India to work on this online shopping. It was found on 2007, at Karnataka India. Today it is the most superior E-business portal which is aggressively expanding and its roots deep into the Indian market and at the same time shifting the mindset of people i.e. from going and shopping from physical store to online be magnificent. Overall Flipchart is good, but it is facing some tough competition from its competitors like EBay and Amazon.

There are more sites have been enabled, which offers online shopping of anything from anywhere. They provide various products from different brands and also show the price distribution of each. There are different payment facilities like credit card, cash on delivery, net banking etc. They also contain column in which one can give their review or any opinion related to any product or site which help others to take decisions properly. But still there are many things to improve in this field. People are not getting full satisfactory results

II. LIETRATURE REVIEW

E-commerce allows interacting consumers and sells directly through advertisements. But some advertisements are irrelevant to consumers. Therefore, they have proposed a consumer behaviour model using which relevant advertisement can be posted to consumer. Whenever they visit and only when the consumer clicks on particular advertisement then only system posts that advertisement again [1].Comparing product assessment is an important task for online websites and e-shoppers. Online merchant compares to the other competitor for search best price for specific product. Functional and non-functional requirement has been used in which they provide function like product page detection, product record detection, product attribute extraction, product attribute assignment [2].As we all know large amount of data is being generated from sensors, satellite and social media etc. And this data can be processed and used to take decisions properly. And for that, Hadoop framework uses mad and reduce phase. This phase uses technique which matches the stream data and similarity is calculated using thirteen semantic measures such as token-based-similarity, edit-based-similarity, similarity hybrid phonetic similarity as well as domain dependent natural's language processing measures. [3].Data is increase rapidly. The main purpose of this data collection to make thing simple for the user. The unstructured data is structured and processed by using the Mapreduce technique and automatic prediction of user's taste is done through collaborative filtering. Mapreduce is a shuffling strategy to perform filtering and aggregation of data analysis task [4].As growth of data is being increased and due to this large amount of data. It has become very difficult to make decision. Consumers are buying the product from where they want and capable. But before buying that particular product consumers prefer to see the reviews of that, which were given by others for that same. And for this rating as well as reviews i.e. any type of feedback, we use collaborative filtering technique. This technique will help an individual to make decision properly [5].Nowadays, web search engine has become a necessary platform for to get type of information or to shop anything. There is an agent is called as comparative shopping agent (CSAs). And these CSAs give us the average prices or any price dispersion for the search. They provide two type of search results organic and paid. In which paid results contain more CSAs and Venders than

organic. Their result suggests there is significant different for average price between venders inside CSAs between paid and organic results [6].

SR	Hadoop	Big Data	Algorithm/ Technology	Purpose
[1]	Yes	No	Consumer Behavior Model	Relevant advertisement posted when consumers visit the website.
[2]	No	Yes	Data Mining Records in Web pages(MDR)	Ecommerce portal contain huge number of products and it need to be discovers, extract the product properties and compare.
[3]	Yes	No	MapReduce	Matches the same stream data and calculates results.
[4]	Yes	No	MapReduce	Rapidly on Social networking sites large amount of data generated. MapReduce process the structured data and users taste is done through CF.
[5]	Yes	No	Collaborative Filtering	Before purchasing product on the basis of rating and reviews of existing consumers are useful to make decision for other consumers.
[6]			Comparative Shopping Agents	To the web shoppers, Comparative shopping agents (CSAs) give a list of vendors and price details of a specified product.
[7]	Yes	No	Collaborative Filtering	The recommendation processes for user is encapsulated in the Map function. The result describes that the design algorithms enable CF algorithm in Hadoop platform to take the good performance.
[8]	Yes	No	Data Packet Inspection	DPI is used to process big data and calculate various traffic functions.
[9]	Yes	No	MapReduce	On social media number of data is generated and there is data placement is a critical factor. RC file overcomes this factor. A fast and space-efficient data placement structure is very important to big data analytics in large-scale distributed systems.
[10]	Yes	No	Self-Adaptive MapReduce Algorithm	It was method to improve the efficiency of the MapReduce scheduling algorithms. It uses less amount of computation and gives high accuracy.
[11]	Yes	No	MapReduce, MR-SPS	MR-SPS, increases the scalability of the cluster and its performance by managing workload and data locality. CloudSim simulator use for simulation time and scalability.
[12]	Yes	No	Novel Prediction Model	OMO optimizes the overlap between map and reduce phases. The purpose is to reduce overall makespan by slot allocation. It shows the improvement in heavy shuffling jobs.
[13]	Yes	No	Collaborative Filtering	A CF algorithm based on the variance analysis of Attributes-value Preference(AP) is proposed in this paper.
[14]	Yes	No		The advantages of using parallelism, realized with Amazon Elastic MapReduce, over a sequential implementation for large datasets. After the job was executed sequentially, it was also executed on four clusters of one, three, five and eight Amazon EC2 virtual nodes.
[15]	Yes	No	MapReduce	HDFS are increasingly being used for processing large and unstructured data sets. Hadoop enables interacting with the MapReduce programming model while hiding the complexity of deploying, configuring and running the software components in the public or private cloud.
[16]	No	Yes	Kaal Algorithm	Address two very important issues for the implementation of high scale ESP applications: the lack of responsiveness in batch processing systems and the lack of scalability in ESP systems.
[17]	Yes	No	Collaborative Filtering	Implement user based CF algorithm on a cloud computing platform, namely Hadoop, to solve the scalability problem of CF. It partitions users into groups according to two basic principles, i.e., tidy arrangement of mapper number to overcome the initiation of mapper and partition task equally such that all processors finish task at the same time, can achieve linear speedup.
[18]	Yes	No	Collaborative Filtering	CF is one of the recommendation system considered to be effectively adapted in tourist's domain.
[19]	Yes	No	Collaborative Filtering	Location Based System Network (LSBN) has also emerged during this period. The LSBN's allows user to work on Point of Interest (POI's) for better service by sharing their experience and opinions about the place they have visited such as companies, restaurants, clubs etc.
[20]	Yes	No	Hadoop Benchmark	Benchmark is used to process all big data and networking components. This benchmark is used to mix data and process all storage and networking components of data of hadoop cluster. It also includes different data sizes for each mix job and to reflect customer usage

Table 1: Literature Reviews

Collaborative filtering technique has been implemented on cloud computing platform called Hadoop. They also map reduce framework in that calculation process. But a problem is detected in this collaborative filtering is scalability i.e. if the volume of data is large then cost of collaborative would also be high. To overcome this problem cloud computing has been used. They mainly focused on map reduce model and map function [7].If we see, and then we will come to know that internet traffic is also experiencing with explosive growth. However, Deep Packet Inspection (DPI)is used to process this big data DPI is hadoop based and it is used for various like, live web traffic visiting mall, the leading e-shopping giant in China was investigated [8].One important and escaping application of big data happens in social networks on the Internet, where billions of people all over the world connect and the number of users along with their various activities is growing rapidly. In such a system, the data placement structure is critical factors that can aspect the warehouse performance in a fundamental way. A fast and space-efficient data placement structure is very important to big data analytics in large-scale distributed systems. RCFile is designed to meet all the four goals, and has been implemented on top of Hadoop [9]. The Scheduling algorithm of FIFO (FIRST IN FIRST OUT) is used in Hadoop as default in which the jobs are executed in the order of their arrival. LATE the dynamic scheduling technique is being introduced to schedule the jobs in the heterogeneous environment. The three principles of LATE algorithms are proposed in this paper are prioritizing tasks to speculate, selecting fast nodes to run on, capping speculative tasks to prevent thrashing [10]. We proposed a parallel scheduler for the YARN/MapReduce platform. Our scheduler is based on the MPI model, which allows users to execute parallel applications. This approach optimizes the capacities use of the resource manager which increases considerably the number of tasks that it can handle which improve the scalability of the system. The proposed algorithm is called MR-SPS for MapReduce Scalable Parallel Scheduler. To evaluate our algorithm, we have used the CloudSim simulator in java language. The experiments show that our algorithm is superior to others in term of simulation time and scalability [11].Main aim to develop an efficient scheduling scheme in MapReduce cluster to improve the resource utilization and reduce the total completion length of a given set of jobs. There are two techniques includes lazy start of reduce tasks, batch finish of map task. In lazy start of reduce tasks, to find the best timing to start reduce tasks so that there is sufficient time for reduce task to shuffle the intermediate data while the resource is allocated to server map task as such as possible [12]. Collaborative filtering is the state-of-the-art and widely applied method in personalized recommendation systems. A collaborative filtering algorithm based on the variance analysis of attributes-value preference (AP) is proposed. It makes full use of the difference of variance from the AP. The various algorithms are introducing Traditional Collaborative Filtering Algorithm, Neighbour-based recommendation through attributes value preference, Improved Attributes-value Preference that Incorporates with Variance Analysis, Collaborative Filtering Algorithm based on Variance Analysis of Attributes Value Preference [13].For the purpose of comparing a sequential solution to the join problem with a distributed MapReduce based approach, a large set of data was generated. The data is made of two sets - users and comments linked by the user's ID. The data was fed to both a sequential implementation of a join algorithm running on a single machine and a parallel one, using MapReduce, running on a cluster of varying size. This paper shows the advantages of using parallelism, realized with Amazon Elastic MapReduce, over a sequential implementation for large datasets [14].Nowadays, data is being generated with very high rate from different fields like business, scientific data, social networking sites etc. To analysis and process this large amount of data, there is a need of an application and clusters. These applications a Hadoop have also found model that works a distributed environment. HDFS and map reduce work efficiently and scalable to process huge amount of data. They have worked not only processing but also extraction of meaning and required information [15]. Data processing approach combines map reduce model and recent development in Event Stream Processing (ESP) ideas. Recent technologies like IOT, Mobile Computing, social networking creates more data processing problems. ESP having set of techniques to address such types of problems. Programming model that enables the map reduce for programmer who needs his application to support quick reaction of incoming data to quick more to the ESP system. Programming model also enables scalable functionality [16]. Offers on e-commerce websites are based on their decision made for advertisement of product. Decision is mostly made by websites for

clearing stocks. Kaal algorithm used to generate frequent itemsets over stream of data. This itemsets will give an idea about offers to be made on purchase of offers [17].The rate of World Wide Web increases in recent years. Internet enables tourists to search and purchase service at any place. Recommendation system is applied for helping tourists to make personalize vacation plans. This recommendation system is capable of generating list of preference attraction for tourists. Collaborative filtering is one of the recommendation system considered to be effectively adapted in tourist's domain [18].When GPS enabled in smart phones, the gap between physical and virtual has been reduced. Location Based System Network (LSBN) has also emerged during this period. The LSBN's allows user to work on Point of Interest (POI's) for better service by sharing their experience and opinions about the place they have visited such as companies, restaurants, clubs etc. They have aimed to model user rating and their choices [19] .They has represented Hadoop benchmarks. This benchmark is used to mix data and process all storage and networking components of data of hadoopcluster.it also includes different data sizes for each mix job and to reflect customer usage. It is also easy to run [20].

III. PROBLEM DEFINATION

Due to unsatisfactory result of online shopping, we will be designing a server which will contain all the information and the updates of different showrooms and stores. This will help to get all the information of their need and get it done by going directly to that place which suited them best. Instead of roaming here and there, wasting their time or investing money on online products which don't give them full satisfaction and happiness, which they get by trying, checking and then buying.

IV. PROPOSE SYSTEM

Online Shopping System helps in purchasing of specific products and services online by selecting the listed products from portals (E-Commerce site). The proposed system helps in building a portal to purchase, sell products or goods online using online network. Purchasing of goods online, user can select different products based on department, online transaction, delivery services and hence covering the disadvantages of the existing system and making the purchasing easier and helping the dealer to reach wider market.

In day to day life, we will need to purchase lots of goods or products from a shop. It may be food items, electronic pieces, house hold pieces etc. Now days, it is really hard to get some time to go out and get them by ourselves due to busy in lots of works. In order to solve this, B2C E-Commerce portals have been started. Using these portals, we can purchase specific or products online just by visiting the portals and ordering the pieces online by making online transaction.

This existing system of purchasing best has several disadvantages. It requires lots of time to travel to the particular shop to purchase the specific. Since everyone is leading busy life now days, time means a lot to everyone. Also there is cost for travelling from house to shop. More over the shop from where we would like to purchase something may not be open 24*7*365. Hence we have to adjust our time with the salesperson's time or dealer's time.

In order to overcome these, we have e-commerce solution, i.e. one place where we can get all required goods/products online. The proposed system helps in building a portal to purchase, sell products or goods online using online network. Purchasing of goods online, user can choose different products based on department, online transaction, delivery services and hence covering the disadvantages of the existing system and making the purchasing easier and helping the dealer to reach vast market.

Still, there are times when people don't prefer online shopping as they don't get full satisfactory results. They have to go for different places, for different varieties, which takes their lot of time. To make their work easy, there will be an online portal/platform/website which will be providing all the information and updates regarding all stores and showrooms. This will help customers to get all the information related to any product/item according their needs at their own place, instead of roaming here and there.

In that they can search anything according to their need, get all the updates related to it like price, availability etc. And then they can also book that item for limited period of time. At the end they just have to Take decision and can go directly to that place and buy without any confusion and doubt, which they get while shopping online.

Android smart phone: Android is an open source platform founded in October 2003 currently developed by Google. A smart phone is an mobile

phone with advanced mobile operating system which feature combine features of a personal computer with other features useful for mobile. Most smart phones can access the internet have a touch screen user interface can run app, music player.

Database: The shops database is designed using MYSQL. It provides interface with any database can be easily designed. The shops database consists

Inventory table - It provides information about the availability of the items, their unique id, product id etc.

Item table- It provides detailed information of each item from its manufacturing date, price, weight, etc.

Shop details- The customers information will be stored in this table including his address and phone number that will be used at the time of online payment.

Store details- This table will have detail information about the shops name, its branch and unique id That will be retrieved at the time of scanning of the shops barcode.

Final order table- This table maintains customer information about his purchases, total cost, session id and all those information that is required to generate a final bill.

Web server: A web server is server which can connect one device to another that is active in the internet and establish communication between them. Web server uses common protocol for communication such as HTTP. Web service is required to establish communication between Android device and Shops database to exchange information.

Steps to perform this operation:

1 The client registers his account and create login id with password.

XXXVII

2 Then send request to the web services.

3 The web services send this request to shop database.

4The shop database search the particular item from table and responds to web server with available information.

5 Next web services packed the item with related offers and send back to client.

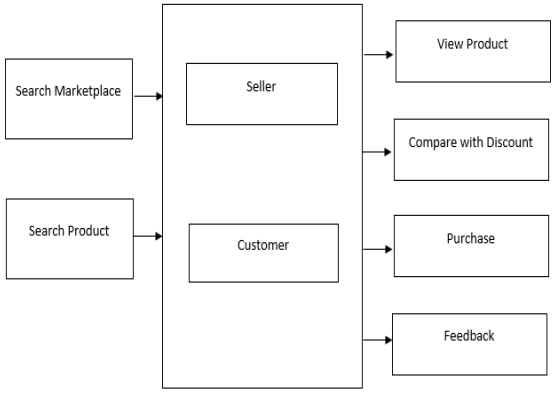

Fig1 Flow of System

V. CONCLUSION

Now as we conclude, instead of this growing trend of online shopping, still some customers are not satisfied. They buy products online, but they don't get what actually they thought of. There are always some issues regarding their quality, size etc. They invest still they don't get their exact thing. That's why they go outside themselves, to search for their need and get it after a long search. Therefore, by considering all this, there will be a portal, or we can say one platform where they will get all that which they were searching by going at different places. They will get all the information and updates regarding their search. As this portal will be containing the data of all the stores and showrooms. A big amount of data will be generated i.e. there will be Big Data. To handle all that, we will be using Hadoop and maps reduce concept and many more algorithms to sort them and present them in a pleasant manner.

ACKNOWLEDGEMENT

We would like give our special thanks to our guide, Prof.AvinashDevare and Prof.Nitin Chopade for guiding us in the examination in work. Our genuine thanks to Prof.H.A.Hingoliwala, HeadOf Department of Computer Science and Engineering, for his advantageous opinions and instructions. We would also like to give special thanks to Principal, Dr.M.D. Jadhav, for his support and stimulation in our work.

REFERENCES

[1] Thirumalaisamy Ragunathan, Sudheer Kumar Battula, Vedika Jorika, Ch Mounika, A U Sruthi, and Mucherla DivyaVani. "Advertisement Posting based on Consumer Behavior." Procedia Computer Science, 2015.

[2] Andrea Horch, Holger Kett and Anette Weisbecker. "Mining E-Commerce Data from E-Shop Websites." 2015 IEEE Trustcom/ BigDataSE/ISPA.

[3] S.Prabhakar Benny ,Dr s.vasavip,Anupriva, "Hadoop Framework For Entity Resolution with in High Velocity Strams",International Conference on Computional Modeling and Security(CMS 2014).

[4] Subramaniyaswamy v, Vijayakumar v, Logesh R and Indragandhi v. "Unstructured Data Analysis on Big Data using Map Reduce" 2015, Precedia Computer Science.

[5] Riyaz P A, Surekha Mariam Varghese. "A Scalable Product Recommendation using Collaborative Filtering in Hadoop for Bigdata." 2015, International Conference on Emerging Trends in Engineering, Science and Technology.

[6] Zhongming Ma, Kun Liao, Johnny Jiung-Yee Lee. "Examining Comparative Shopping Agents from Two Types of Search Results." International Conference on Computing, Engineering and Information.

[7] Zhi-Dan Zhao and Ming-Sheng Shang. "User-based Collaborative-Filtering Recommendation Algorithms on Hadoop." 2010, Third International Conference on Knowledge Discovery and Data Mining.

[8] Jiangtao Luo, Yan Liang, Wei Gao, Junchao Yang,"Hadoop based Deep Packet Inspection System for Traffic Analysis of E-Business websites",2014 IEE.

[9] Yongqiang He , Rubao Lee , Yin Huai , Zheng Shao , Namit Jain , Xiaodong Zhang , Zhiwei Xu . " RCFile: A Fast and Space-efficient Data Placement Structure in MapReduce-based Warehouse Systems." IEEE Transaction.

[10] R.Thangaselvi, S.Ananthbabu, S.Jagadeesh, R.Aruna. Improving the efficiency of MapReduce scheduling algorithm in Hadoop. IEEE Transaction.

[11] Rostom Mennour,Mohamed Batouche,Oussama Hannache. "MR-SPS: Scalable Parallel Scheduler for YARN/MapReduce Platform." 2015, IEEE International Conference on Service Operations And Logistics, And Informatics (SOLI).

[12] Jiayin Wang,Yi Yaot,Ying Mao,Bo Sbeng, Ningfang Mit, "OMO: Optimize MapReduce Overlap with a Good Start(Reduce) and a Good Finish (Map)."IEEE Trancation.

[13] Xiaoyun Wang, and Jintao Du. "A Collaborative Filtering Algorithm Based on Variance Analysis of Attributes-value Preference." 2009, International Conference on Management of e-Commerce and e-Government.

[14] Marko Lalic, Emina Memic, Faruk Kesan, Edita Gondzic, Nermin Smajic, Novica Nosovic. "Comparison of a Sequential and a MapReduce Approach to Joining Large Datasets." May 2013IEEE Transaction.

[15] Mohd Rehan Ghazi, Durgaprasad Gangodkar. "Hadoop, MapReduce and HDFS: A Developers Perspective." International Conference on Computer, Communication and Convergence (ICCC 2015).

[16] Andrey Brito,Andre Martin, Thomas Knauth, Stephan Creutz, Diogo Becker, Stefan Weigert, Christof Fetzer, "Scalable and Low-Latency Data Processing with StreamMapReduce." 2011, Third IEEE International.

[17] Hemant Chaudhary, Deepak Kumar Yadav, Rajat Bhatnagar Uddagiri Chandrashekhar. "MapReduce Based Frequent Itemset Itemset Mining Algorithm on stream Data." 2015, Global Conference on communication technology.

[18] Ziyang Jia, Wei Gao, Yuting Yang, and Xu Chen. "User Based Collaborative Filtering for Tourists Attraction Recommendations." 2015, IEEE International Conference on Computational Intelligence and Communication Technology.

[19] Xin Li ,Guandong Xu ,Enhong Chen, and Yu Zong. "Learning recency based comparative choice towards point of interest recommendation."2015,Expert System with Application.

[20] Vikram A. Saletore, Karthik Krishnan, Vish Viswanathan, Matthew E. Tolentino, "HcBench: Methodology, Development, and Characterization of a Customer Usage Representative Big Data/Hadoop Benchmark",2013 IEEE