ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 5, Issue 5, May 2016

An Efficient Online Shopping System Using Map Reduce Framework in Big data

Vinodhini. M[1], Manju. A[2]
[1] PG Scholar,
Department of Computer Science and Engineering,
Saveetha Engineering College,
Tamil Nadu, India

[2] Assistant Professor,
Department of Computer Science and Engineering,
Saveetha Engineering College,
Tamil Nadu, India

***Abstract*- Big data concerns the massive volumes of dataset usually huge, sparse, incomplete, uncertain, complex or dynamic, which are mainly coming from multiple and autonomous sources. The most fundamental challenge for big data application is to explore the large volume of data and extract useful information or knowledge for future actions. Online shopping application provide users with the products in their stock and will render the comparison within their products only. Thereby limiting the users to analyze before buying a product. Shoppers may be unsatisfied with the limited number of choices in available product as per their requirement. There by the shoppers has to switch between multiple sites in order to match the requirements. This may sound as a hectic process, the shoppers may get frustrated due to lot of time utility. This system crabs the data from various web application and load its dataset collaboratively using crawling technique and process the batch jobs in a distributed and parallel processing way using HDFS (Hadoop Distributed File System). It allow the shoppers to analyze, get recommendations, to pick products and add to cart irrespective of the service provider. This system stands unique as it does not rely on the single service provider. The cart can be reviewed at any time and can be processed. All the information will be securely and precisely stored in user session. This results in an effective data analysis, to achieve fast response, scalable and an efficient precise service comparison.**

Keywords*- *Big data, E-commerce, Hadoop, HDFS, Relevance clustering, Collaborative filtering, Map reduce, Service Comparison, Recommendation.

I. INTRODUCTION

Big data is a term that describes large volume of data, both structured and unstructured that inundates a business on a day-to-day basis. Big data can be analyzed for insights that lead to better decisions and strategic business moves. Big data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision-making.

Big data concerns the massive volumes of dataset usually huge, sparse, incomplete, uncertain, complex or dynamic, which are mainly coming from multiple and autonomous sources. The most fundamental challenge for big data application is to explore large volumes of data and extract useful information or knowledge for future actions. In day to day life, we will need to buy lots of goods or products from a shop. It may be food items, electronic items, house hold items etc. Now a days, it is really hard to get some time to go out and get them by ourselves due to busy life style or lot of works.

In order to solve this, e-commerce websites have been started. Using these websites, we can buy goods or products online just by visiting the website and ordering the item online by making payments online. The current online shopping system provides the user to purchase the goods in online, user can choose different products based on categories, online payments, delivery services are done. It provides only limited number of products available in the online shopping website.

Thereby limiting the users to analyze before buying a product and the shoppers may be unsatisfied with the limited number of choices in available product as per their requirement. Thus the shoppers has to switch between multiple sites in order to match the requirement and this results a hectic process, shoppers may get frustrated due to the lot of time utility. So it is necessary to propose new system which helps in building a website where massive amount of products are available in a single service provider.

This massive amount of products are crawled from multiple service provider and loads its dataset collaboratively using crawling technique and process the batch jobs in a distributed and parallel processing way using HDFS (Hadoop Distributed File System) in an efficient way.

It allows the shoppers to analyze, get recommendations, can pick products and add to cart irrespective of the service provider and hence covering the disadvantages of the system and making the buying easier and helping the vendor to reach wider market. This system stands unique as it does not rely on the single service provider. The cart can be reviewed at any time and can be processed. All the information will be securely and precisely stored in user's session. This results in an effective data analysis, to achieve fast response, scalable and an efficient precise service comparison.

II. Related work

In recent years online shopping system becomes a trend in which data mining and data warehousing plays a vital role in processing and storing the data. As the days passed the shoppers who buy the products online increases such that product in the website also increases. The techniques and algorithms used in existing online shopping system are clustering algorithm, classification model, prediction analysis in recommendation system. Clustering algorithm [2] groups the shoppers who have similar preferences in buying the product from e-commerce website. The users who buy the products are given an unique identifier and then clustered as a group.

The average similarities of the individual members in the cluster are calculated for predicting the new customer cluster. Grouping of products as cluster depends upon the same category. Clustering algorithms [4] such as k-means algorithm is used in large dataset which is partitioned into clusters; this is done based on property set and similar products are recognized from massive dataset. It deals with the large scale data which employs distributed solution based on the categories and its properties. Classification model [7] uses top down greedy search to form a decision tree structure on cluster in order to test the attribute on the cluster. ID3 algorithm is one of the classification algorithm that is widely used in online shopping system. The complexity is high in the arrangement of the products**.**

Predictive analysis [6] is applied in product recommendation, predictive search. This prediction forms patterns to predict the product to be chosen by the user from the given input data, past and historical data. It also uses machine learning algorithm to analyze the data and make predictions. The relationship between the customer and the e-commerce websites is managed using association rules. This association rules [13] determines the buying habits of customers and set offers to them in e-commerce websites. It also identifies the frequent item sets that are occurring simultaneously in the database. With the fast development of networking, data storage, and the data collection capacity, Big data are now rapidly expanding in all the fields for data processing. Big data [17] processing mainly depends on parallel programming models like map reduce, as well as providing a cloud computing platform of big data services for the public. Map reduce is a batch-oriented parallel computing model. There is still a certain gap in performance with relational databases.

Improving the performance of map reduce and enhancing the real-time nature of large-scale data processing have received a significant amount of attention, with map reduce parallel programming being applied to many machine learning and data mining algorithms. Data mining algorithms [23] usually need to scan through the training data for obtaining the statistics to solve or optimize model parameters. It calls for intensive computing to access the large-scale data frequently. To improve the efficiency of algorithms, visual et al [1] proposed a general-purpose parallel programming method, which is applicable to a large number of machine learning algorithms based on the simple map reduce programming model on multi core processors.

Ten classical data mining algorithms [5] are realized in the framework including locally weighted linear regression, k-means, logistic regression, naive bayes, linear support vector machines, the independent variable analysis, Gaussian discriminate analysis, expectation maximization, and back propagation neural networks. Big data [14] literally concerns about data volume, HACE theorem has the key characteristics of the big data are 1) huge with heterogeneous and diverse data sources, 2) autonomous with distributed and decentralized control, and 3) complex and evolving in data and knowledge associations.

To support big data mining, high-performance computing platforms are required, which impose systematic designs to unleash the full power of the Big data. Hadoop tool [19] uses Hadoop Distributed File System (HDFS) as the primary distributed storage. HDFS is well known for distributed, scalable and portable file-system using commodity hardware. Name node/Master node is having metadata information about whole system such as data stored on data nodes, free space, active nodes, passive nodes, job tracker, task tracker and many other configuration files such as replication of data.

Data node [15] is a type of slave node which is used to save data and task tracker in data node which is used to track on the ongoing jobs which are coming from name node. A small-size hadoop cluster includes one master node and multiple slave nodes. A slave node acts as a data node in HDFS architecture and acts as task tracker in map reduce framework. These types of clusters are used in only non-standard applications. Again large-size cluster [21] includes a dedicated name node which manages all file system index or metadata, a secondary name node which periodically generates snapshots of name node to prevent or recover from name node failure and reducing loss of data. Similarly in map reduce framework a standalone job tracker server which manages job scheduling and several task trackers are running on data nodes or slaves. Skew reduce [25] improves the execution speed and runtime is improved even if the cost functions provided by user were not perfectly accurate but that are good enough to identify expensive results in the data.

III. Online shopping system using map reduce framework

In proposed online shopping system, sample dataset are stored and distributed in different web servers using web service access call that uses SOAP protocol. It connects the web servers one another which acts as web applications. The information and features of the products are classified using relevance clustering algorithm thus it processes dataset as batch jobs in order to change the uncategorized files to categorized files in a distributed and Parallel processing way. The products in the dataset are converted to serializable object by representing their product id in the web servers and to reduce the dataset size in order to store in memory. The dataset are stored in HDFS (Hadoop Distributed File System).

The Gateway application is developed that crabs the data from the various distributed web servers using web crawling technique and then crawled resources are converted into a reduced object using map reduce algorithm, reduced object contains all necessary information providing service comparison and recommendation. Service comparison is used to have a clear comparison in choosing the products from the online shopping site. The recommendations were given based on the QoS, availability, delivery, offers, price and specifications of the particular product. It allows the shoppers to analyze, get recommendations, can pick products and add to cart irrespective of the service provider. It stands unique as it does not rely on the single service provider.

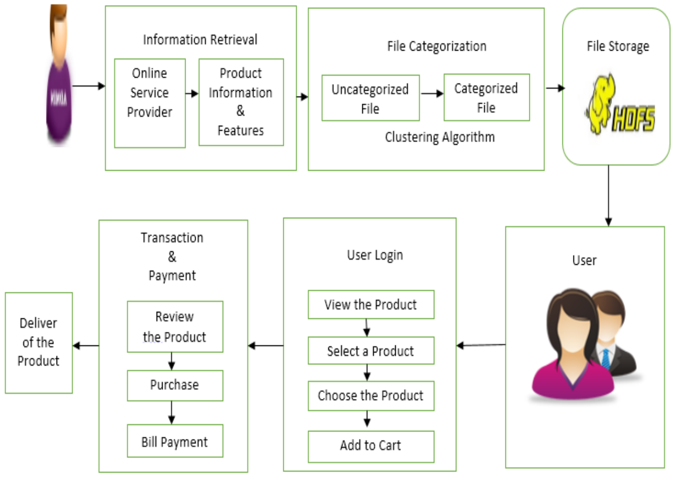

Fig1. Architecture of online shopping system using map reduce

A. Information Retrieval

Sample web applications were built using apache tomcat servers which acts as the web servers in order to store the massive amount of products, so that the users can compare their products with different service providers. The application uses sample dataset that has been crawled in amazon previously. Dataset was loaded and distributed independently in each web applications by web service access call that uses SOAP protocol. It is a protocol that uses xml to transmit the information between web servers. WSDL (Web Service Description Language) describes web services so that applications can connect to them easily. The web servers are connected through URL connectivity. The dataset which are in comma separated value are converted to tab separated value.

Features and other specifications have been loaded differently for each application based on the service providers requirement. Features and its information are classified and clustered using relevance clustering algorithm. It is used in order to separate features and its information of the each product. The first step is to eliminate the irrelevant features from the dataset; the relevant features are selected by the features having the value greater than the predefined threshold.

In the second step selected relevant features are used to generate the graph, divide the features using graph theoretic method, and then clusters are formed by using minimum spanning tree. In the third step the subsets features that are more related to the target class was selected. The offers and prices for each product are allocated dynamically.

B. File Categorization

The Online shopping gateway application is built in which the users will be provided with recommendations and comparisons between the products in the website. The resources provided by various web servers are in TSV (Tab Separated Values) format and it is preprocessed. The dataset which are distributed to the web servers are done by web services access call. Dataset are loaded in the web servers and processed as batch jobs for the conversion of object files. The tab separated value files were parsed for data. The Dataset are converted from text to serializable objects. The struts framework is used in which model view controller accepts the incoming request from web servers are accessed by web services to our gateway application.

The incoming request are accepted by the admin. The product list are listed in the online shopping gateway application. Admin can login by using the username and password, the product is selected from the online shopping gateway application. The selected product is preprocessed for the conversion of text to object files. The data which are uncategorized are categorized to an object and used for further processing. The conversion of text to object for product information and its features are done. This is the conversion of an object to a series of bytes, so that the object can be easily saved to the persistent storage and saves memory space.

C. Hadoop Storage

The Dataset are stored in HDFS where parallel processing is performed in different web applications. Hadoop Distributed File System is used in order to store a vast amount of data and hence fast accessing of data is done. The Hadoop Distributed File System (HDFS) is a distributed, scalable, and portable file-system written in java for the hadoop framework. HDFS stores large files typically in the range of gigabytes to terabytes across multiple machines. This uses map reduce algorithm which maps the data and reduces it by comparing the similarities among the products and reducing the products into a single object. Hadoop storage stores the large amount of data and can be retrieved in a very short period of time. This results in processing of the data in a parallel processing manner by the web servers.

D. User Login

The Dataset which are distributed to various web servers which acts as the web applications are crawled in the gateway application that is built. Crawling can be done by using URL connectivity and web service access call. The crawled resources are then reduced by map reduce framework and converted into a single object. The map reduce framework is just a computing framework which uses master slave concept. This conversion of reduced object is done separately in each web servers and it contains all the necessary information and features of the product for providing comparison of the products and cart based recommendations. A huge amount of data will be accumulated in a single service provider from various providers.

The massive amount of products are stored in the gateway application where the user can register and then login to view various products available in online shopping application. This is done by writing a web service client process for each service provider. It can connect to the various web applications, web services can pull all the needed data to the backend. The user can register and login by using the user name and password to view and to shop the products. After selecting the product from the gateway application it gives the comparison of the products from the integrated website, so that the shopper can analyze and select the product from which website according to their choices. This makes the shopper to get satisfied by buying the product according to their choices. Recommendation system used here is the cart based recommendation system, the recommendations are done based on the products that are added to the cart. It helps the shopper to get the recommended products according to the products that are added to the cart.

E. Transaction Processing and Delivery of product

This system models complete transactions and recommended items come from the evaluation of those transactions. Having analyzed the previous transactions and identifies the concepts within which concrete items appear, the given part of a new transaction is matched over the existing ones to find the more adequate solution. When the user initiates transaction, the online shopping application will connect to the banking web services directly on behalf of the service provider and completes the transaction securely with help of OTP sent to their mail id given on user registration. Each and every customer must create a bank account by registering on this bank server.

A Bank account is needed to complete the transaction which can be created earlier through the banking application. The process will be back to our application as soon as the transaction is over and the purchased products will be added to the cart. The product which is purchased is delivered to the user. Delivery of product is done depending on the shipping address given by the user.

V. Experimental setup

The proposed system overcomes the problem of scalability in choosing the product from shopping website. Terabytes of products can be stored using HDFS and it can be accessed from anywhere at any time. It makes the product to be available up-to-date and to have a clear service comparison of products among the integrated website. It contains all dimensions of Big data. It provides the ease of gathering/processing of the products to the shoppers. Since the dataset of product are stored in hadoop the processing is fast and it is said to be fault tolerant. Service comparison system which makes the customer to choose the products by having a clear comparison of the products in three different web applications.

The recommendation system is based on cart based recommendation which helps the customers to get the recommended products according to the products that are added to the cart. The customer detail of this website are stored in bank server and it provides the confidentiality to each and every customer, the products that are purchased by each customer are stored as the record in customers account confidentially. This system stands unique as it does not rely on the single service provider. The cart can be reviewed at any time and can be processed. All the information will be securely and precisely stored in user session. This results in an effective data analysis, achieves fast response, scalable and an efficient precise service comparison.

VI. Conclusion

Online shopping system produces an integrated dataset from various websites that supports the scalability problem and reduces the time consumption which processes the system in an efficient manner. It stands unique as it does not rely on the single service provider. The information of the customer will be securely and precisely stored in user session. This results in an effective data analysis, to achieve fast response, scalable and an efficient precise service comparison.. In future, this system can be deployed in cloud, so that users from different places can make use of this system by accessing it through cloud storage. Dataset which covers terabytes of products can be stored and processed for shopping. Recommendation system can be enhanced for recommending the products based on the customer's preferences.

REFERENCES

[1] A. Arasu and H. Garcia-Molina, "Extracting structured data from web pages," in Proc. 2003 ACM SIGMOD, San Diego, CA, USA, pp. 337– 348.

[2] Bini Tofflin. R1, Kamala Malar. M2, Sivasakthi. S ,(2014) "A Relevant Clustering Algorithm for High-Dimensional Data".International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March.

[3] Colby Ranger, Ramanan Raghuraman, Arun Penmetsa, Gary Bradski, Christos sKozyrakis" Evaluating Map Reduce for Multi-core and Multi processor Systems".

[4] C.H. Chang, M. Kayed, M. R. Girgis, and K. F. Shaalan,(2006) "A survey of web information extraction systems," IEEE Trans. Knowl. Data Eng., vol. 18.

[5] Deepak S. Tamhane, Sultana N. Sayyad, (2015)" Big data analysis using hace theorem". International Journal of Advanced Research in Computer Engineering &Technology (IJARCET) Volume 4 Issue 1, January.

[6] F. Ashraf, T. Özyer, and R. Alhajj, "Employing clustering techniques for automatic information extraction from HTML documents,"IEEE Trans. Syst. Man Cybern. C, vol. 38, no. 5, pp. 660–673, Sept. 2008.

[7]Hassan A. Sleiman and Rafael Corchuelo" Trinity: On Using Trinary Trees for Unsupervised Web Data Extraction" IEEE transactions on knowledge and data engineering, vol. 26, no. 6, june 2014.

[8]H. A. Sleiman and R. Corchuelo, "A survey on region extractors from web documents," IEEE Trans. Knowl. *Data Eng.*, vol. 25, no.9, pp. 1960–1981, Sept. 2012.

[9]H. A. Sleiman and R. Corchuelo, "An unsupervised technique to extract information from semi-structured web pages," in Proc.13th Int. Conf. WISE, Paphos, Cyprus, 2012, pp. 631–637.

[10] K. Simon and G. Lausen, "ViPER: Augmenting automatic information extraction with visual perceptions," in *Proc. 14th ACM Int. CIKM*, Bremen, Germany, 2005, pp. 381–388.

[11]Muslea, S. Minton, and C. A. Knoblock, "Hierarchical wrapper induction for semi structured information sources," Auton. Agents Multi-Agent Syst., vol. 4, no. 1–2, pp. 93–114, Mar./Jun. 2001.

[12] M. Álvarez, A. Pan, J. Raposo, F. Bellas, and F. Cacheda, "Extracting lists of data records from semi-structured web pages," *Data Knowl. Eng.*, vol. 64, no 2, pp. 491–509, Feb. 2008.s

[13]Ms. Saranya.K.S1, Ms.Anjana Prabhakaran ,Mr.Thomas George K,"Decision support system for crm in online shopping system".

[14]Sherin a, Dr s uma, saranya k, saranya vani M, "Survey on big data mining platforms, algorithms and challenges".

[15] S. Soderland, "Learning information extraction rules for semi structured and free text," Mach. Learn., vol. 34, no. 1–3, pp. 233–272, Feb. 1999

[16] Uyoyo Edosio,"Big data Analytics and its Application in E-commerce", case studies of adidas, walmart, amazon.com.

[17] V. Crescenzi and G. Mecca, "Automatic information extraction from large Websites," *J. ACM*, vol. 51, no. 5, pp. 731–779.

[18]V. Priyadharshini, K. Tharamaraiselvi, S. Kavitha, B. Sivaranjani" Extraction Of Unsupervised Web Data Using Trinity" International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 4, April 2015.

[19] Vishal A. Nawale1, Prof. Priya Deshpande2, "Survey on load balancing and data skew mitigation in map reduce applications", international journal of computer engineering & technology (ijcet).

[20] W. Meng, "ViDE: A vision-based approach for deep web data extraction," IEEE Trans. Knowl. *Data Eng.*, vol. 22, no. 3, pp. 447–460, Mar. 2010.

[21] Wu, F. Li, S. Mehrotra, and B. C. Ooi. "Query Optimization for Massively Parallel Data Processing". In SOCC, 2011.

[22]Xindong Wu, Fellow, IEEE, Xingquan Zhu, Senior Member, IEEE,Gong-Qing Wu, and Wei Ding, Senior Member, IEEE Data Mining with "Big Data transactions on knowledge and data engineering", vol. 26, no. 1, january 2014

[23]Xindong Wu, Fellow, IEEE, Xingquan Zhu, Gong-Qing Wu,Wei Ding" Data Mining with Big Data" Ieee transactions on knowledge and data engineering, vol. 26, no. 1, january 2014.

[24] Yingyi Bu, Bill Howe, Magdalena Balazinska, and Michael D. Ernst. Hadoop: Efficient iterative data processing on large clusters. Pvldb, 3(1):285 296, 2010

[25] Y. Kwon, M. Balazinska, and B. Howe, "A study of skew in map reduce applications," in Proc. of the Open Cirrus Summit, 2011.

[26] Y. Kwon, M. Balazinska, B. Howe, and J. Rolia. "Skew-resistant parallel processing of feature-extracting scientific user-defined functions". In Proc. of the First SOCC Conf., June 2010.

[27] Y. Kwon, M. Balazinska, B. Howe, and J. Rolia, "Skew tune: Mitigating skew in mapreduce applications," in Proc. of the ACM SIGMOD International Conference on Management of Data, 2012.

[28] Zhengkui Wang, Yan Chu, Kian-Lee Tan, Divyakant Agrawal, Amr EI Abbadi, Xiaolong Xu, "Scalable Data Cube Analysis over Big Data" appliarXiv:1311.5663v1 [cs.DB] 22 Nov 2013.

[29] Zhixian Zhang , Kenny Q. Zhu , Haixun Wang , Hongsong Li," Automatic Extraction of Top-k Lists from the Web.

[30] Zaharia et al. Improving Map Reduce Performance in Heterogeneous Environments. In OSDI, pages 29–42, 2008.