IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 18, Issue 3, Ver. IV (May-Jun. 2016), PP 04-12
www.iosrjournals.org

A Hadoop-based Retail E-Commerce Weblog Analysis System

Pooja D. Savant[1], Debnath Bhattacharyya[2]
[1](*Department of Information Technology, Bharati Vidyapeeth Deemed University College of Engineering Pune 411043, India)*
[2](*Department of Information Technology, Bharati Vidyapeeth Deemed University College of Engineering Pune 411043, India)*

Abstract: *In the modern age, the act of purchasing underwent a drastic revolution due to the rapid development of e-commerce. For challenging their competitors over such kind of modern setting, which offers more than one way of completing a sales transaction, sellers sense the need to make use of novel plans of action for alluring more customers as well as keeping the existing customers. To set the plans of action, sellers have to analyze the massive web data logs generated by the retail sites. To handle such kind of massive data processing, the traditional text software and Relational DataBase Management System technology has been facing a bottleneck and the results displayed are not very satisfactory. To solve this problem, we are proposing this system which will process these massive web logs with the help of Hadoop as well as allow sellers for linking with purchasers over more than one means at a completely different height through the process of controlling huge capacities of fresh data accessible presently. Purchaser's online search and browsing behavior data and purchase history can give sellers quick insights into their preferences, wants, and desires. By analyzing these massive web logs, sellers are able to visually explore the data quickly and they can use that information to create targeted "micro" segments as well as personalized promotions.*

Keywords: *Hadoop Framework, Distributed File System, MapReduce Software Framework, Pig Latin Language, Apache Sqoop, Tableau Software*

I. Introduction

In recent years, due to the rapid development of retail market over Internet Technologies (ecommerce), the experience of shopping has changed dramatically. As weblog data increased, the retailers started facing problem with the analysis of web data logs generated by the retail sites. We execute the analysis of massive web data logs with the help of Hadoop and Big data. A similarity coefficient is a function which computes the degree of similarity between a pair of text objects. There are a large number of similarity coefficients such as Jaccard, Dice and Cosine coefficients. Out of these, we use Dice coefficient algorithm to perform a comparative analysis and generate comparison results for the product details given by user.

First we get the product details data of our website as well as for other retailer web sites for products comparison. Dice coefficient algorithm for product match strategy gives the average between the words or description for product matching between different retailers. This algorithm will get the description of each product and compare the description of other product in same category and check how many words are matching and give that much % as matching dice coefficient. It is calculated as shown in (1):

$$\left(\frac{\text{Matching word count}}{\text{Total no. of words in description}}\right) X\ 100\ =\ \textit{Dice coefficient}\ \% \qquad \ldots \quad (1)$$

After matching is done we give the system the nearest matching product details and retailer name for comparison. Once done with product matching, we directly compare the matched product price from other retailers with our website, display visual representations of results generated by our system and allow business user to change selling price of sales articles belonging to our website based on sales recommendation as well as sales article comparison of our sales article with sales articles of other retail sites belonging to alike category for increasing sale and business of our website.

II. Literature Study

The numerous kinds of long-established software dealing with the subject matter of documents and collections of individual pieces of information structured to recognize the dependency between stored items of information have had to endure congestion. A growing and accepted reality which poses a difficulty is that the current methodology is unsuited for the management of massive records of individual pieces of information and

as a result it is unsuccessful in laying its hands on the style that is concealed inside these large records of individual pieces of information over the Internet [1].

Across the current environment of interconnected networks, the study of sources utilized for saving basic information structure is evolving into a mandatory job for studying a purchaser's habits for increasing the promotion and deals in order to study record containing individual pieces of information to obtain the desired insight arising out of these records with the usage of various data mining techniques to automatically discover and extract information from Web documents and services. For being able to study these massive collections of individual pieces of information, a methodology for the handling of these massive collections of individual pieces of information and splitting them across more than one processors with the objective of performing operations on these individual fragments of facts over a smaller span and a trustworthy depository for separate fragments of information are required. Hadoop structure supplies a trustworthy depository containing separate fragments of information through usage of HDFS in combination with MapReduce estimating convention that executes a methodology for the handling of huge collections of individual pieces of information by splitting the separate fragments of information across more than one computers across Hadoop cluster with the objective of handling them to process the outcome more quickly. Hadoop gives us a superior way to save initially and perform queries afterwards, which enables us for heaping the individual pieces of information to the HDFS as well as for performing queries formulated with the help of Pig Latin language. This enables us to lessen the reply period as well as heap across user node [2].

Apache Pig is planned for being suitable at an optimum point which lies within descriptive manner of SQL as well as the manner of MapReduce of using a computer's native language for specification of a systematic order of statements, functions and commands to complete a computational task. Apache Pig is completely executed as well as it also accumulates the Pig Latin language in the form of a series of MapReduce jobs that describe the physical operators used by Pig for implementing an ordered set of commands using other set of commands which implement a particular job instead of machine hardware, giving no evidence about the manner in which their execution takes place inside MapReduce that undergo processing over Hadoop that is nothing but a publicly accessible, map-reduce execution. Apache Pig drastically lessens period required for progress as well as implementation of its jobs for performing the study of individual pieces of information in contrast to the adoption of Hadoop precisely as it is. The new troubleshooting setting which is made available in combination with Apache Pig gives rise to profits incurred by increased pace of work [3].

A measure described as couple-based differentiating capability evaluates the differentiating capability about likeness standard of measurement for inspecting the differences between items refined for sale established over particular features. Across the preliminary stage of handling individual pieces of information, every file which is to undergo further development, is distributed to a separate server program in a computer, over a scattered system of connections which implement the trade policy for a request. Numerous records over the Internet are scattered over separate server programs in computers, over a scattered system of connections which implement the trade policy for a request, are subjected to initial assembling of authored data in the form of an approved sequence with the help of Linux FTP particular collection of directives which are used by end points across a telecommunication connection, when they have to perform communication. After this process of initial assembling, these records over the Internet are unified in the form of a solitary resource for storing all these records. After these storage resources complete the process of splitting, the predefined block capacity undergoes heaping across HDFS named folders with the help of numerous commands. A significant data unification issue about existing objects belonging to a pair of intentionally created examination collections containing existing objects or articles refined for sale that exhibit a resemblance towards a pair of extremely distinct shopping compartments like furniture and clothes is solved by this preliminary stage of data-handling [4].

The utilization of likeness measure for calculating the likeness amidst information entities which possess the ability of having single or multiple features dependent on each other. A likeness indication amidst purchasers related to sheet modifications undergoes computation using a method like Dice's coefficient which computes the similarity coefficients. With the usage of MapReduce estimating convention, the Dice coefficient method undergoes execution, for analyzing how the similarity coefficients act in relation to distinct information amounts as well as cluster capacities [5].

III. Proposed Work

Our Retail Shopping Weblog Analysis System architecture is shown in Fig. 1. Our three retail shopping sites aapna.com, yebhi.com, wobhi.com will be deployed using Eclipse Luna as Integrated Development Environment with three Apache Tomcat Application Servers and three MySQL depositories respectively. Each of these three MySQL depositories of retail shopping sites like aapna.com, yebhi.com, wobhi.com will have three distinct databases web_log_db, yebhi_web_log_db, wobhi_web_log_db, each of which will consist of distinct tables which will save the information related to user details, website details, product details about our three distinct retail shopping websites like aapna.com, yebhi.com as well as wobhi.com.

We consider aapna.com as the retail shopping site which will enable the business user to perform the comparative study of sales articles and find out the resultant article match by using Dice coefficient method, for this we have included one more table for saving the information related to Dice coefficient sales article match in the database web_log_db for aapna.com retail shopping site.

The administrators of our three retail shopping sites like aapna.com, yebhi.com, wobhi.com will possess the rights to add various sales article categories, subordinate categories, sales article name, sales article depiction, sales article images, sales article selling price, the offered price for a particular sales article by the retail shopping business organization as well as promo type through End User Web Interface (EUWI). Each of the three massive web transaction logs of our three distinct shopping sites like aapna.com, yebhi.com and wobhi.com will consist of the information like sales article prices as well as sales price as per the promotions available on different seasons. We will join together these massive logs by our three distinct retail shopping websites aapna.com, yebhi.com as well as wobhi.com into one voluminous retail shopping web log as input for performing a comparative study on it through our system.

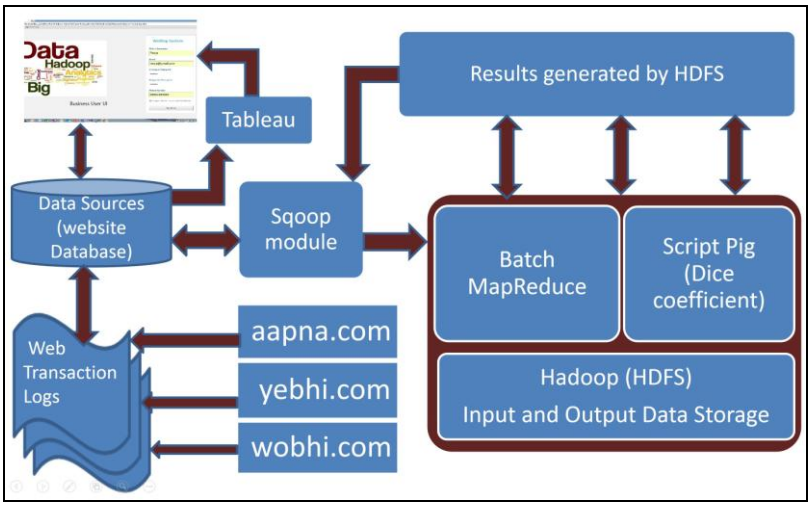

Figure 1. architecture of retail shopping weblog analysis system

This massive retail shopping weblog information will be shifted within Hadoop Distributed File System using Apache Sqoop for further batch processing using MapReduce and for executing Dice coefficient method using a tailor-made Apache Pig UDF to perform a comparative study of the sales articles belonging to alike categories and calculate the percentage of corresponding sales article match. The sales articles of alike category will undergo a comparison firstly amidst aapna.com and yehbi.com, secondly amidst yebhi.com and wobhi.com and lastly amidst wobhi.com and aapna.com respectively and the nearest sales article match will be calculated as a percentage using Dice coefficient method.

The resultant sales article match percentage generated by Dice coefficient algorithm will be 100% if both sales article depictions form an exact match. If the resultant sales article match percentage is less than 100%, then we will offer a list of distinct threshold percentage values as options to the business user. The business user can select one threshold percentage value from the list of options offered according to the business requirement and perform a comparative study to calculate the resultant sales article matching percentage with the help of Dice coefficient method using Pig User Defined Function. These results generated will be saved within Hadoop Distributed File System. The results generated by the massive sales article web logs will be moved into MySQL relation-based depository for visually exploring massive retail shopping results in the form of graphs generated with the utilization of Tableau software by connecting Tableau to MySQL relation-based depository.

These graphs and resultant sales articles of distinct retail shopping sites belonging to alike categories which form an exact match or the nearest sales article resultant match which underwent comparison using Dice coefficient method, along with the sales article match percentage calculated using Dice coefficient will be displayed to the business user of our retail shopping website aapna.com when he performs registration, logs in successfully to aapna.com and selects the desired sales article name, its category and its subordinate category from the drop-down lists. After selecting the desired sales article, the business user will have the option to view the graphs generated using Tableau and the business user will be allowed to change selling price of sales articles belonging to our website based on sales recommendation as well as sales article comparison of our sales article with sales articles of other retail sites

belonging to alike category for improving the sales for retail shopping website aapna.com for alluring more purchasers and preserving the already existing purchasers.
Our Hadoop based retail e-commerce weblog analysis system will utilize the technologies as shown:

3.1. Shell Script
We are using Shell Script to write control logic like script calling, performing the cleanup of output directory, etc.
3.2. MySQL
We are using MySQL to store the data of site like product details, user details. In our case, log means the product details at any time on site which is stored in MySQL table.
3.3. Apache Sqoop
We are using Apache Sqoop to first get stored product data from MySQL to HDFS and secondly to put data from HDFS to MySQL.
3.4. Apache Pig
We are using Apache Pig to get data from all sources and apply the Dice-coefficient logic on it and calculate the product-match.
3.5. Steps of Dice Coefficient Algorithm with Example
Our website is aapna.com and the other retailer's websites are yebhi.com and wobhi.com respectively and the various products offered by various retailers are given in different tables as follows:

Step 1: In the first step, we get the product details data of our site aapna.com as well as for other retailer's web sites yebhi.com and wobhi.com for products comparison.

We can obtain such data from different retailer's database. Sample data for database of different retailer's aapna.com, yebhi.com and wobhi.com is as given in Table 1, Table 2 and Table 3 respectively:

Table 1. Sample Data from database of retailer website aapna.com

Retailer	Manufacturer	Product UID	Product Description	Regular Price (Rs.)	Sale Price (Rs.)
aapna.com	Sony	DSCW830	Sony Cyber Shot DSC W830 with a warranty of one year.	190000	119999
aapna.com	Samsung	RT27JARYESA	Samsung Double Door 253 Liters Frost Free Refrigerator shiny steel	67980	67862
aapna.com	Whirlpool	FR258CLS3S	Whirlpool NEO FR258 CLS 3S Double Door 245 Liters Frost Free Refrigerator Price	67890	66789

Table 2. Sample Data from database of retailer website yebhi.com

Retailer	Manufacturer	Product UID	Product Description	Regular Price (Rs.)	Sale Price (Rs.)
yebhi.com	Sony	DSCW830	Sony Cyber Shot DSC W830 with a warranty of one year and 10% discount.	100000	98999
yebhi.com	Samsung	RT27JARYESA	Samsung 253 Liters Double Door Frost Free Refrigerator.	67980	67862
yebhi.com	Whirlpool	FR258CLS3S	Whirlpool NEO FR258 CLS 3S Double Door.	67890	66789

Table 3. Sample Data from database of retailer website wobhi.com

Retailer	Manufacturer	Product UID	Product Description	Regular Price (Rs.)	Sale Price (Rs.)
wobhi.com	Sony	DSCW830	Sony Cyber Shot DSC W830 with 10% discount.	199000	109999
wobhi.com	Samsung		Samsung Refrigerator Double Door 253 Liters Frost Free 1year warranty	67980	67862
wobhi.com	Whirlpool		Whirlpool NEO FR258 CLS 3S Double Door 245 Liters Frost Free	67890	66789

Step 2: In the second step, once we will import data from databases to HDFS using Sqoop or MySQL import, we will validate information input position as well as HDFS import location by shell commands from input location validation script.

If the input location contains the data in it then it will internally call Script.pig which contains different logics like data filtration/cleanup, aggregation and also implementation of Dice Coefficient Algorithm.

3.5.1. Data Cleanup/Filtration: After loading all retailers' data in script for processing i.e. aapna.com, yebhi.com and wobhi.com, performs the cleanup matching logic by removing the unwanted columns like headers, the row with null product id's (UPC) i.e. Universal Product Code or with no price.
3.5.2. Matching Strategy: In this step we will perform/apply different types of Product Matching Strategy. The Product Matches Strategy consists of UPC_Matching.pig and Non_UPC_matching.pig.
3.5.3. UPC_Matching.pig: The entire products from same manufacturer have same UPC, so we filter out such items which have same product id (UPC) and directly compare such product price and assign such product dice coefficient as 100. In this step we will perform/apply different types of Product Matching Strategy. The Product Matches Strategy consists of UPC_Matching.pig and Non_UPC_matching.pig.
3.5.4. Non_UPC_Matching.pig: We will filter out those products for which UPC details are not there in data and then we apply dice coefficient algorithm on such product. While applying dice coefficient on such items we will consider different fields such as product description and compare the description of each product with our product in same category. With the help of Dice Coefficient, we check how many words are matching and give that much % as matching dice coefficient by using the formula as shown in (1).

After product matching is done we give the system the nearest matching product details and retailer name for comparison. Script.pig will in turn call sqoop_script.sh through which we can directly compare the matched product price from other retailers yebhi.com and wobhi.com with our website aapna.com and offer visual exploration of the results generated by our system and offer sales recommendation to the business user to change selling price of sales articles after comparison of sales articles of our retail shopping site aapna.com with sales articles of our competitor retail shopping sites yebhi.com as well as wobhi.com belonging to alike category for increasing the sale and business of our retail shopping site aapna.com.
Step 3: It is also possible to apply dice coefficient on other different fields.

IV. Result

The flow of our retail shopping weblog analysis program is given using Fig. 2:

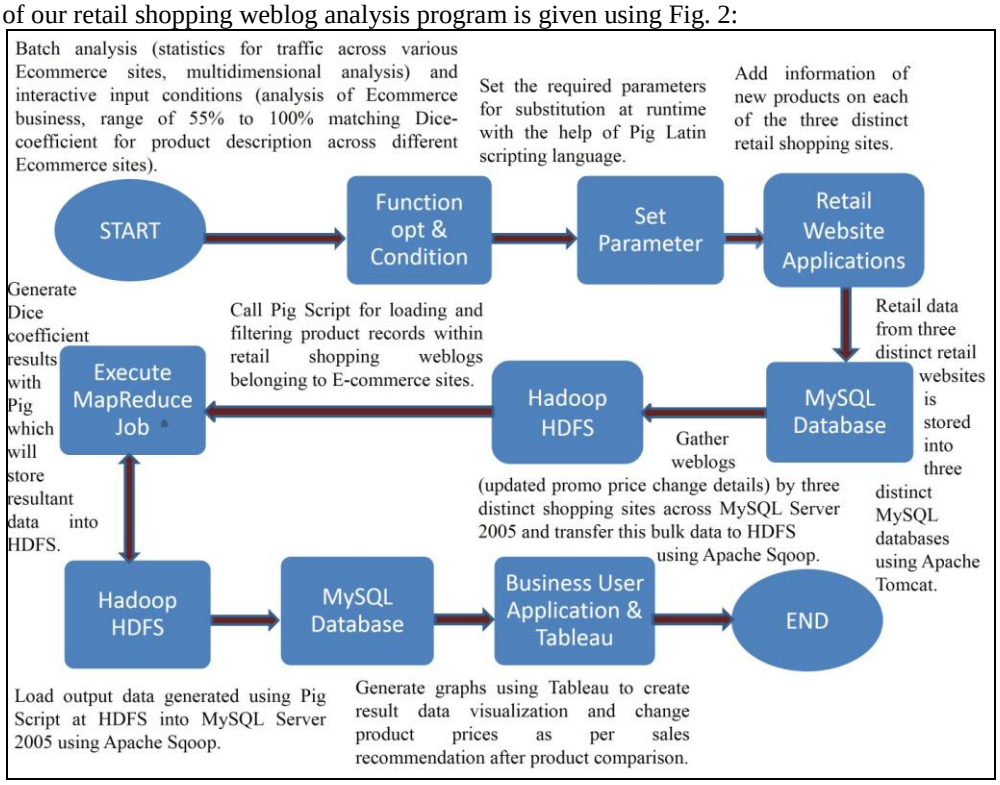

Figure 2. flow of retail shopping weblog analysis program

We login to CentOS 6.2 Machine with username as 'root' and password as 'tomtom'. We access Terminal and execute following three commands to get IP Address for accessing our Hadoop cluster using PuTTY which is made available without any cost for SSH as well as Telnet client execution for Windows as well as Unix Operating Systems:
service sshd start

service iptables stop ifconfig

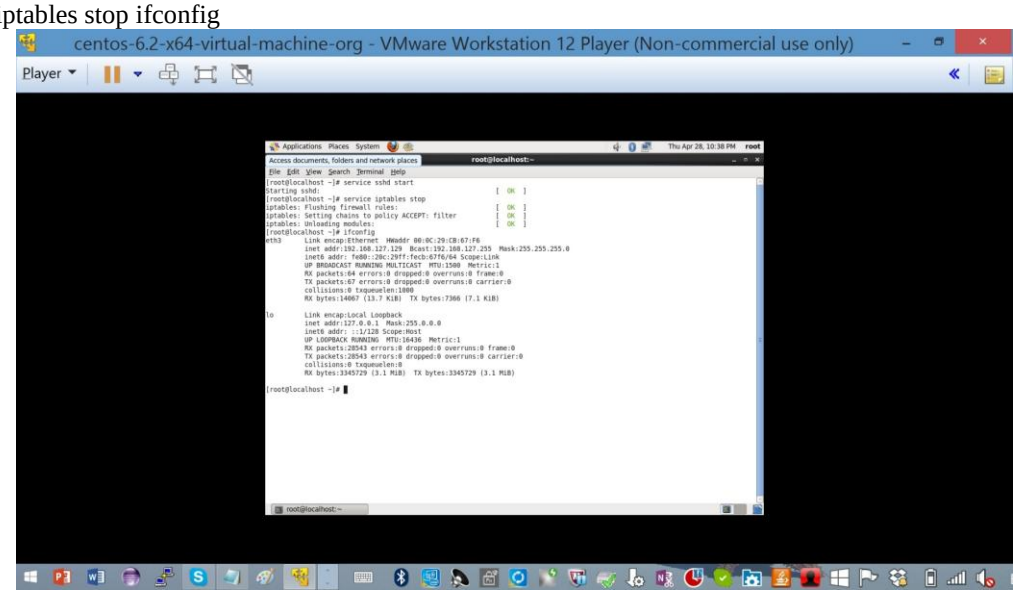

Figure 3. access ip address of hadoop cluster

We see InetAddress displayed in Fig. 3 which we use to access our Hadoop cluster. We use this InetAddress 192.168.127.129 to connect via PuTTY. Before starting MySQLWorkbench it is necessary to start MySQL services by utilizing command given below:

service mysqld start

Until this has been started, our MySQLWorkbench will not show the tables contained within each of the three distinct retail shopping website depositories like web_log_db for aapna.com, yebhi_web_log_db for yebhi.com as well as wobhi_web_log_db for wobhi.com. To view all the tables belonging to distinct retail shopping sites like aapna.com, yehbi.com and wobhi.com, we open MySQLWorkbench using desktop icon and in that we open default connection to view the tables.

We open the Eclipse Luna Integrated Development Environment and create a tailor-made Pig User Defined Function within that named as PigDiceCoefficient.java. We first compile and then export it in the form of a jar file with the help of following commands: javac–classpath hadoop-common-2.0.0-cdh4.7.0.jar:hadoop-mapreduce-client-core-2.0.0-cdh4.7.0.jar PigDiceCoefficient.java

jar –cvf mapreduce.jar DiceCoefficient*.class

We will then run DiceCoefficient.jar file by utilizing command given below:

hadoop jar mapreduce.jar DiceCoefficient input/ output/

We run Apache Sqoop which brings retail shopping weblog information from MySQL in HDFS. We create a new sqoop_get_mysql_to_hadoop.sh file which contains the following commands to shift information saved from MySQL in HDFS:

First we remove our Hadoop data location before importing data from MySQL table into HDFS.

hadoop fs -rm -r /user/web_log_db/tbl_site_product_dtl

hadoop fs -rm -r /user/wobhi_web_log_db/tbl_site_product_dtl

hadoop fs -rm -r /user/yebhi_web_log_db/tbl_site_product_dtl

sqoop import --connect jdbc:mysql://localhost/web_log_db --username root --table tbl_site_product_dtl --target-dir /user/web_log_db/tbl_site_product_dtl --fields-terminated-by '|'

sqoop import --connect jdbc:mysql://localhost/wobhi_web_log_db --username root --table tbl_site_product_dtl --target-dir /user/wobhi_web_log_db/tbl_site_product_dtl --fields-terminated-by '|'

sqoop import --connect jdbc:mysql://localhost/yebhi_web_log_db --username root --table tbl_site_product_dtl --target-dir /user/yebhi_web_log_db/tbl_site_product_dtl --fields-terminated-by '|'

Save and close sqoop_get_mysql_to_hadoop.sh file. Then we execute the following commands:

cd web_log

sh sqoop_get_mysql_to_hadoop.sh

To run Pig Script for Dice Coefficient Method, we utilize the following commands:

pig pig_product_dice_coefficient_calc_1.pig

To shutdown CentOS 6.2 Machine, we enter "init 0" inside terminal.

The administrator of aapna.com retail shopping website generates visual representation as given in Fig. 4 with information as given in Table 4 by utilizing Tableau:

Table 4. Visual representational information for no. of users on a monthly basis

Year	Month	No. of Users
2016	Jan-16	100
2016	Feb-16	200
2016	Mar-16	300

Such kind of visual exploration will enable administrator of aapna.com to estimate number of purchaser's allured by the website on a monthly basis to check if they are increasing per month or not Accordingly, a facility to change selling price of sales articles belonging to our website aapna.com based on sales recommendation as well as sales article comparison of our sales article with sales articles of other retail sites yebhi.com and wobhi.com belonging to alike category will be provided by our system to business user for increasing the number of purchaser's allured on a monthly basis by retail shopping site aapna.com.

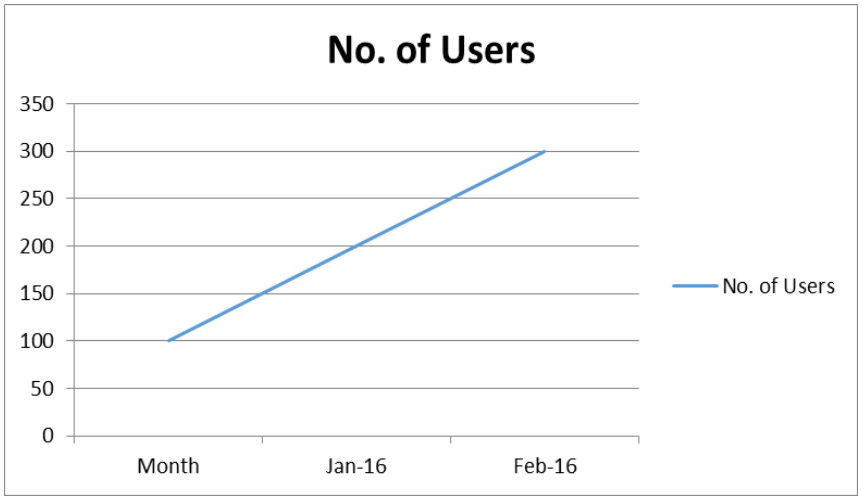

Figure 4. visual exploration of no. of users i.e. purchasers of sales articles on a month-wise basis

The business user of aapna.com retail shopping website generates visual representation as given in Fig. 5 with information as given in Table 5 by utilizing Tableau:

Table 5. Visual representational information for monthly sales articles by category

Year	Category	Articles by Category
2016	Electronics	2000
2016	Clothes	3000
2016	Shoes	2111

Such kind of visual exploration will enable the business user of retail shopping site aapna.com for estimating the number of sales articles purchased by customers on a monthly basis, so that the inventory of retail shopping site aapna.com can be managed in more efficient manner by the business user and the business user can employ various plans of action like change selling price of sales articles which are facing a decline in sales by purchasers on a monthly basis.

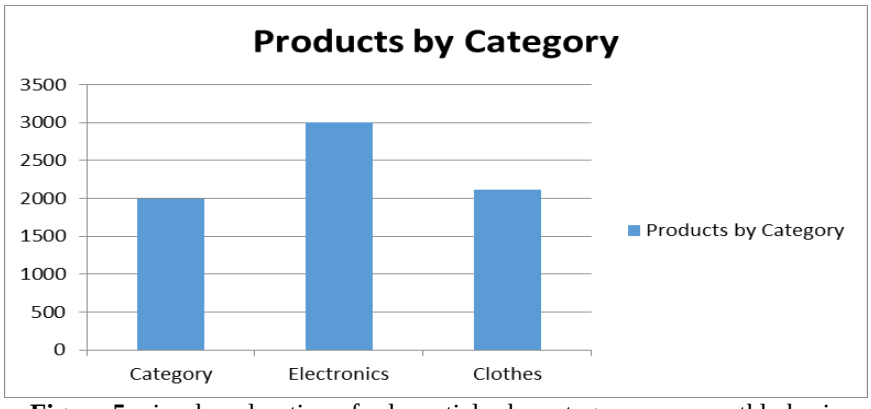

Figure 5. visual exploration of sales articles by category on a monthly basis

The business users as well as administrators of aapna.com retail shopping website generate visual representation as given in Fig. 6 with information as given in Table 6 by utilizing Tableau:

Table 6. Visual representational information for no. of Log Lines against time required in minutes for processing them

No. of Log Lines	Time (Minutes)
3000	10
4000	11
5000	11
6000	12
7000	12

Such kind of visual exploration enables the business users as well as administrators of retail shopping site aapna.com to estimate time taken for processing a particular count of lines of retail e-commerce information.

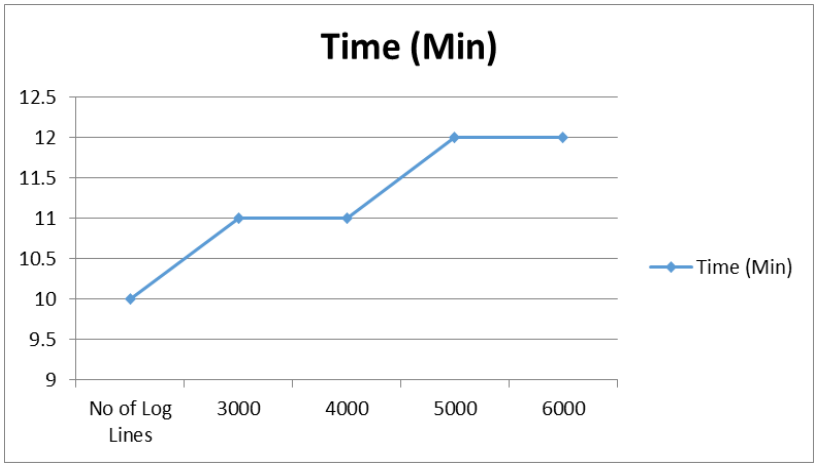

Figure 6. visual representation of hadoop retail shopping log processing time

This effectively demonstrates the power of Hadoop platform in processing voluminous amounts of data within less time in a fault-tolerant and dependable manner and enables the business users as well as administrators of retail shopping site aapna.com to acquire an accurate as well as thorough understanding which is invaluable over retail shopping log information by three distinct sites aapna.com, yebhi.com as well as wobhi.com and gain access to sales recommendation offered by our system like changing sales article price of sales articles belonging to our retail shopping site aapna.com after comparison with the prices charged for sales articles belonging to alike category by our competitors yebhi.com and wobhi.com within a matter of a few minutes which can be employed further to set plans of action to increase the sales as well as business of retail shopping site aapna.com and gain a valuable edge against its competitor retail shopping sites yebhi.com as well as wobhi.com and remain a step in front of them. The business user of a retail shopping website utilizing our retail shopping weblog analysis system may create targeted 'micro-segments' as well as personalized promotions for individual purchasers based on every purchaser's online search, browsing behavior data and purchase history which can give a quick understanding about every purchaser's preferences, wants, and desires and increase overall sales profit earned by retail shopping site and increase its business.

V. Comparison Between Results Of Existing System And Proposed System

The comparison between the results generated by the proposed system and existing system is as given in Table 7:

Table 7. Comparison between results of proposed and existing system

Proposed System	Existing System
The users of our proposed system consist of top-management professionals such as business users, retailers as well as administrators of e-commerce websites.	The users of existing system consist of administrators of websites belonging to all domains.
In our proposed system, we have developed an application which utilizes log obtained by joining weblogs of three distinct retail shopping sites which consist of sales article cost as per promotions offered during diverse seasons.	Our existing system employs no such application which utilizes third-party logs.
In our proposed system, we have used Tableau software for better visual representation of retailer data.	In our existing system, no software is available to show the graph trends, so admin users have to do this manually using

	the result data.
In our proposed system, we have used Dice coefficient method to carry out the sales article statement differentiation amidst the sales articles owned by many shopping sites from identical subordinate category and calculate the product match percentage.	In system which is already present, no article tallying method has been executed.
In this system which we present, we have utilized tailored method for executing sales article matching algorithm.	In existing system, no matching system was there, so tailored functions were not utilized.
In this system which we present, we are providing a recommendation on sales.	In existing system, no such kind of recommendation is provided.
Our proposed system enables business users to create targeted "micro-segments" and personalized promotions for employing new plans of action to allure as well as preserve customers based on comparison of sales articles made available by different shopping sites from identical subordinate category.	System which is already present enables a system administrator to take website-related business decisions for a website.
Our proposed system utilizes Apache Sqoop which provides a facility to transfer the stored retailer trade object information saved by relation-based depository like MySQL into HDFS and also to place facts stored by HDFS into relation-based depository like MySQL.	System which is already present fails to supply any type of provision to move any kind of facts stored by relation-based depository into HDFS and also to place facts stored by HDFS into relation-based depository.
This system which we present, conducts a study of retail e-commerce weblogs.	The existing system performs analysis of generic web data logs.
Our proposed system promotes Business Intelligence and management of shopping site.	System which is already present promotes better generic site management but does not promote Business Intelligence.

VI. Conclusion

Our research presents a Retail Shopping Weblog analysis system using the Apache Hadoop framework along with recommendation system. Our project execution highlights that Hadoop MapReduce software framework can efficiently deal with the massive Retail Shopping Weblogs generated by the shopping sites and generate results very quickly in fault tolerant manner. Our system is very cost-effective as Hadoop uses economical systems or desktops rather than highly configured servers. The Dice coefficient method for product match strategy calculates the average amidst words or depiction for sales article match amidst many sellers as well as carries out a comparative analysis successfully and generate comparison results for the product details given by user. Apache Pig enables to finish the MapReduce job execution very quickly which in turn leads to rise in the profits earned by business organizations utilizing this technology. Tableau software connected to business user application enables the sellers to generate visual representation of results of analysis of massive retail shopping weblog data quickly to set in motion novel plans of action like change sales article price for the seller's website after comparison with the prices charged by competitors for sales articles belonging to alike category for alluring more and more purchasers as well as for preserving existing purchasers which further leads to increased gains for the business organizations utilizing this software.

VII. Future Scope

Through the analysis of massive retail shopping web logs generated by the retail shopping sites that contain valuable facts about every purchaser's online search, browsing behavior data and purchase history, the sellers can get quick insights into their preferences, wants, and desires. The sellers can visually explore the data quickly and they can use that information in the future, to create targeted "micro" segments as well as personalized promotions to increase their profit-rate as well as perform a better management of their inventory and reduce the company's sales article manufacturing costs and also predict the future trends of business most effectively. Currently, our system provides recommendation system only for sales and it can work efficiently only with huge web data logs. Our system can be used in E-commerce product comparison applications as well as in Cross-Selling recommendation applications or Up-Selling recommendation applications.

References

[1]. Chen Hau Wang, Ching Tsomg Tsai, Chia Chen Fan and Shyan Ming Yuan, "A Hadoop-based Weblog Analysis System", International Conference on Ubi-Media Computing and Workshops (UMEDIA), Ulaanbaatar, Mongolia, July 12-14, 2014, pp. 72 – 77.

[2]. Sayalee Narkhede, Tripti Baraskar, "HMR Log Analyzer: Analyze Web Application Logs Over Hadoop Mapreduce", International Journal of UbiComp (IJU), Vol. 4, No. 3, July 2013, pp. 41-51.

[3]. C. Olston, B. Reed, U. Srivastava, R. Kumar, and A. Tumkins, "Pig Latin: A Not-So-Foreign Language for Data-Processing", SIGMOD -2008, June 9-12, 2008, Vancouver, Canada, pp. 1099-1110.

[4]. Krisztian Balog, Norway, "On the investigation of similarity measures for product resolution", International Conference on Discovering Meaning On the Go in Large Heterogeneous Data, San Francisco, CA, USA, 2011, Pages 49-54.

[5]. Nayakam, GhanaShyam Nath, "Study of Similarity Coefficients Using MapReduce Programming Model", North Dakota State University Institutional Repository, Fargo, 2013.