International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-2, July 2019

Optimized Recommendation System for E Commerce on Product Features and User Behavior

Chemmalar Selvi G, Iqjot Singh

***Abstract:** Big data is a late of huge information stored in it and all we need is to dig into get the important information out of it and create a useful system which can be very helpful in improving the current scenario. There are various applications where big data is being used and even there are few fields that are learning techniques to go with big data and evaluate their work and get an improve decision. This paper particularly concentrates on the e commerce system which is highly trending on the market field. [20] E commerce also known as electronic commerce is a market place which gives you a platform to enjoy various services from both buyers as well as sellers. It is a place with various varieties are provided that can help the consumer to choose from and the buyer can get a platform where he can show case his product and get millions of the customer at the same time and he does not have to look for site all the time, it's the system that take care of it. Now big data is playing a vital role in e commerce as it reads about user behavior and provides him a suitable product that he may need according to his behavior and query. There are various machine learning algorithms that are working on this and improving the services. [11] Basically in this paper we will read the user information and combine it with the product attributes and get a suitable suggestion for the user that will be most likely to be purchased by him. In the existing system we just look at one part of the case and give suggestion but in this paper we looked at both the sides, that is we looked after the product entities (the attributes and features that it poses) and the user behavior (the information given by the user and its previous history) that will better prediction and improve the system. Moreover for the optimized working of the system we included an enhanced version of HPCA scheduling algorithm for the Hadoop distributed file system also known as HDFS, which is very suitable for the heterogeneous system, the existing algorithm looks after the overall capacity of the node and then the tasks were assigned but here we will consider the health and the left over capacity of the nodes and arrange the queue for the same which will be refreshed all the time after the task is completed by any node.[18] The aim of the paper is to provide fast and most suitable suggestions to the users which can play a vital role in improving the sales of the company and getting the target done soon and faster.*

Index Terms: *About four key words or phrases in alphabetical order, separated by commas.*

Revised Manuscript Received on 30 July 2019.
***** Correspondence Author

(Chemmalar Selvi G.), School of Information Technology and Engineering, VIT Vellore, India.

Iqjot Singh, Department of Computer Applications, VIT Vellore,India.

I. INTRODUCTION

E Commerce or electronic commerce is a term which is suitable for any type of commercial, or business transactions which includes the exchange of information over the Internet medium.[9] It ranges from small organizations to a large one. The biggest advantage that the E commerce gave to the human kind is that they allow millions of customers to exchange services and goods electronically and also with no hurdles like distance or time or any other factors. An E commerce provides us a platform to get exposed to many people, know their choices and also get updated about the latest trends going in the market or in their domain and hence make their product up to the mark combined with all latest technologies and with necessary changes. With the help of an E commerce platform we can view and judge various products any time and at any location, just we need to take our smart device open it up and get connected to the Internet and then surf with millions of products and get aware about the new stuffs [12].

Fig. 1.a. E Commerce

You can also compare with other goods and company without getting irritated or exhausted and even after comparing you can make a good decision of buying the stuff. The only disadvantage about E commerce is that it does not give us a touch and feel facility, instead we need to judge the product on the basis of the pictures or on the description provided by the company and we have to trust those words and buy on the risk, although there are several schemes like cash back policy, 30days return and all which had solved this problem to a high extend and hence developed the trust of the user or customers.

Types of E commerce [13]

E commerce is not only describing about the relationship between entrepreneurs and customers but it also depicts the relations between various businesses and administrations, so we can say that there are four types of E commerce models, namely:

1. Business-to-Consumer (B2C)
2. Consumer-to-Consumer (C2C)
3. Business-to-Business (B2B)
4. Consumer-to-Business (C2B)

Business-to-Consumer (B2C): It is a kind of E Commerce model in which businesses sell their services to the customers over internet like ordering food in restaurant using apps like swiggy or foodpanda, shopping in the websites and so on. It is normal like a traditional market

Consumer-to-Consumer (C2C): It is a kind of E Commerce model in which consumers sell their product to other consumers, like in olx we sell out our products those are not in use.

Business-to-Business (C2C): It is a kind of E Commerce model in which different buyers come up and sell product or services to each other, its like giving the product to retailer from manufacturer.

Consumer-to-Business (C2B): It is a kind of E Commerce model in which the group of consumers comes up and makes a product which a company buys, it is a total reversal of traditional marketing. The common example can be giving graphic services or digital services and so on.

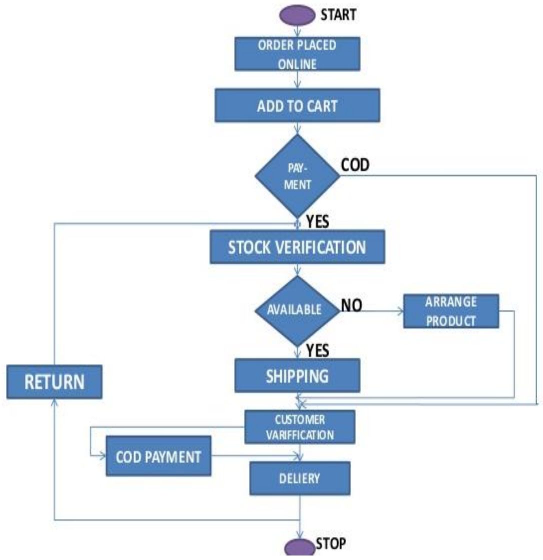

Fig.1. b. Process of shopping in E Commerce

II. LITERATURE SURVEY

Michael R. Anderson and Michael Cafarella [1] told that applying machine learning to a huge amount of datasets had now become a trend and also it is very important for the system to give a proper recommendation and create a sophisticated system. The authors of the paper looked after the problem of writing long codes and performing small changes within the huge code as per the requirements. Machine learning is a process of training a machine or system to learn new features and serve to the user as the best and to make this thing happen we need a proper set of features or attributes which can be analyzed properly and make the decision making easy and sorted. In short, we can say that the success of the trained system is fully dependent on the features those are chosen. Sometimes feature engineering becomes a tedious and a time-consuming experience because we have to deal with big and real-time data so the code becomes iterative and difficult to handle. The authors introduced Zombie which is a data-centric system that is used to speed up the feature engineering by intelligently choosing the appropriate feature for the system and serving an optimized version of an inner loop system in feature engineering. Their approach was to decrease the feature extraction time by providing an effective rule-based input selection technique. The system is divided into two stages; firstly the data of similar elements are grouped together to an index and then in the second step online querying is done in which the system output a high-quality subset data which can provide the useful feature vectors and then this subset will play a role for evaluating the feature code. Suhang Wang, Jiliang Tang, Yilin Wang, Huan Liu [2] described about the current hierarchal structure present within the recommender system. Basically various items shows a certain and a unique hierarchy in the real world scenario and in the same way we can structure the user preferences and that can be easily studied by anyone. The new studies and research showed that combining the hierarchy of the user and the item can improve the suggestions and recommendations given by the recommender system. But the problem that was faced and addressed by the authors was that the implicit preferences were not available from the user and this was the next difficult task to do. Authors proposed a framework through which they can get those implicit details from the user and not only that It will help us to get both the details of user implicit and explicit when combined with the system. This will help the system to work on a uniform framework and will help in using the resources less and smartly. They perform experiments on real datasets and analyzed the result which was useful and fast as well. Nitin Sodera, Akshi Kumar [3] had addressed on various problems on the available recommender system and presented a study on them. With the high availability of the information and the resource it is very important to keep track of all the resource and an accurate filtering tool is required as well to sort out or filter the vital information. There is a proper system named recommender system which is very useful in providing the information and recommendations. Recommender system is a particular tool which study our likings and disliking on various attributes and conditions and give a recommended product to us. Their paper has complete literature survey on the various existing system in the IT sector and how fruitful they are to the scenarios and the conditions and moreover they have addressed various problems that come up on these systems and they gave a brief overview of the system to the users for their better understandings and clear view. Chen and et point out about the fact of how the scheduling algorithms in MapReduce cannot be used when it comes to solving real-time problems. In this paper, they come up with a technique for real-time problems with Hadoop framework [4].

The proposed algorithm takes into consideration a lot of factors such as load balancing, properties of every machine that is within the cluster. After performing tests on it, it was found that the proposed algorithm can prevent starvation for short jobs, the data-locality increased by 15% [4]. The main purpose of all this is to provide a solution for handling issues related to big data analytics and also to increase the performance of MR and HDFS. These algorithms have a huge role for the satisfactory result of the recommendations, as they schedule the files in a proper and in an efficient way so thata accessing of these file randomly could be fast and the programmer as well as the user does not have to wait for long. In this paper, Liu and et al [5] explained the importance of big data and the need of E commerce. This importance and need has led to the development of techniques that can handle and process large amount of data [5]. The authors focus on the future growth for technology related to large data and also about the influence that data has made on the Chinese E commerce. All the encounters faced are taken into consideration and solutions are put forward. Pingping Dong [6] discussed about various E commerce websites which are available under the big data domain. The author told how the firms used big data insights to get their sale high and increased. The author told that big data not only played a role in increasing the sales but it also increase the security and user comfortability within the site, which helps in maintaining the user and vendor relationship. The author gave many examples of various kinds of e commerce including B2b, C2B, C2C and so on. With this study he concluded that big data had added a new opportunity in the constructing the electronic website and evaluation for the various subparts is easy and highly accessible. From the paper "Handling Big Data using a Data-Aware HDFS and Evolutionary Clustering Technique" which was published in "IEEE Transactions on Big Data" we get to know the data-aware module for the Hadoop eco-system. This paper proposes a framework which allows Hadoop to manage the distribution of data and its placement based on the cluster analysis on the data itself. This allowed parallel processing of queries for data on HDFS that required less resource usage. In E commerce the parallel processing is very important because every time the data come in is needed to be computed and resulted then only a recommendation can be powerful in nature.[7] Dr.S.Suguna, M.Vithya, J.I.Christy Eunaicy [8] proposed that how we can use Hadoop MapReduce to analyze the big data. Authors addressed the issue that how it is difficult to apply data mining techniques within the present amount of data. Basically web mining is technology that helps in extracting the useful and required information from the available data coming from various web resources and these log files are well maintained and preserved in the web servers for the various computational tasks. Basically the e commerce website studies about the customer and it behavior towards its site and various products available there and on that basis it gives suggestions and recommendations and there are many web mining algorithms which are very useful for the study of user and its behavior so the author proposed that we can process the algorithms using the MapReduce and design a framework that can compute result effectively and efficiently.

From the paper "A Personalized Recommendation Engine for Prediction of Disorder using Big Data Analytics" which was published in "IEEE International Conference on Innovations in Green Energy and Healthcare Technologies", the Big Data is told as a promising solution in healthcare sector. The data related to the patient and its well being constitutes to be a Big-Data problem. This paper focuses on prediction and diagnosis of disorders based on the size of the data as large sized data helps to predict the results of a particular occurrence very easily. SO by the help of this paper we came know how the predictive algorithm work on the user basis and how we can easily relate to other use cases. In E commerce there are bunch of people which shares the same interest and a person with same interests tends to have same outcome and result, so predicting a good recommendation can be easy with these algorithms.[9] In the paper "Big Data Analytics : Hadoop and Tools" which was published in "2016 IEEE Bombay Section Symposium (IBSS)", he typically used data analytics tools are compared. Here the Big Data characteristics describe the exponential increase in challenges that arise in traditional data analysis and processing. The authors talked about how the huge data can be processed well and fast with Hadoop. The problem that is mostly faced with big data is how we can capture it and then computing the huge data, which can be petabytes in size, but the mapreduce algorithms and HDFS system it can be done and in parallel basis which helps the computation of teh data to be done fast and easily. In E commerce, every second more than 100 events are recorded and each record plays a very vital role for the product analysis and it scoring, so here the Hadoop and its component come into the scenario where we can access and compute the large data fast and easily[10].

Reference	Technology	Advantages	Disadvantages
[1] Anderson, M. R., & Cafarella, M. (2016, May). Input selection for fast feature engineering. In Data Engineering (ICDE), 2016 IEEE 32nd International Conference on (pp. 577-588). IEEE.	ZOMBIE - a data-centric system that is used to speed up the feature engineering by intelligently choosing the appropriate feature for the system and serving an optimized version of an inner loop system in feature engineering.	1.)Fast feature selection 2.)optimized loop handling algorithm	1.) Explicit features were ignored. 2.) System Workload Problem
[2] Wang, S., Tang, J., Wang, Y., & Liu, H. (2018). Exploring Hierarchical Structures for Recommender Systems. IEEE Transactions on Knowledge and Data Engineering.	Explicit and implicit hierarchal framework	1.) Better visualization 2.) Good decision making 3.) Implicit and explicit both are considered	1.) Time Consuming
[3] Sodera, N., & Kumar, A. (2017, May). Open problems in recommender systems diversity. In Computing, Communication and Automation (ICCCA), 2017 International Conference on (pp. 82-87). IEEE.	Survey paper for various recommender systems available.		
[4] Chen, F., Liu, J., & Zhu, Y. (2017, June). A Real-Time Scheduling Strategy Based on Processing Framework of Hadoop. In Big Data (BigData Congress), 2017 IEEE International Congress on (pp. 321-328). IEEE.	Scheduling Algorithms in HDFS 1.)HDMA 2.)LERDA	1.)Data locality was increased 2.) Mapping was improved 3.) No starvation effect	1.) Not a secure system. 2.) System overload problem for huge datasets
[5] Liu, J., Sun, L., Higgs, R., Zhang, Y., & Huang, Y. (2017, July). The Electronic Commerce in the Era of Internet of Things and Big Data. In Computational Science and Engineering (CSE) and Embedded and Ubiquitous Computing (EUC), 2017 IEEE International Conference on (Vol. 2, pp. 360-363). IEEE.	Analyzing the technologies for big data in e commerce		

[6] Dong, P. (2015, September). Research on the evaluation of E commerce website under the environment of big data. In Instrumentation and Measurement, Computer, Communication and Control (IMCCC), 2015 Fifth International Conference on(pp. 288-292). IEEE.	Us of big data in e commerce		
[7] Hajeer, M. H., & Dasgupta, D. (2017). Handling Big Data Using a Data-Aware HDFS and Evolutionary Clustering Technique. IEEE Transactions on Big Data.	Encoding technique for generic algorithm in a distributed system	1.) Parallel processing for the data. 2.) Less resource usage 3.) Efficient for OLAP worloads	1.) Computation workload. 2.) No dynamic update.
[8] Suguna, S., Vithya, M., & Eunaicy, J. C. (2016, August). Big data analysis in E commerce system using HadoopMapReduce. In Inventive Computation Technologies (ICICT), International Conference on (Vol. 2, pp. 1-6). IEEE.	Web mining algorithm using Hadoop	1.)Highly robust 2.) Fast output 3.) Less overload 4.) Well managed data. 5.) Best for distributed system and parallel processing	
[9] Shobana, V., & Kumar, N. (2017, March). A personalized recommendation engine for prediction of disorders using big data analytics. In Innovations in Green Energy and Healthcare Technologies (IGEHT), 2017 International Conference on (pp. 1-4). IEEE.	Recommendation system in healthcare sector	1.) Prediction of disease. 2.) Recommendation of proper treatment. 3.) Precautions available.	
[10] Sogodekar, M., Pandey, S., Tupkari, I., & Manekar, A. (2016, December). Big data analytics: hadoop and tools. In Bombay Section Symposium (IBSS), 2016 IEEE (pp. 1-6). IEEE.	Comparison of tools used for big data analytics		

III. EXISTING SYSTEM

In the existing e-commerce system, the prediction are made by analyzing the behavior of the user, which let them to choose from the recommended products and with the help of big data we are able to track the user favorites or the like products and help them to suggest the product which can be most likely brought by him.[15] This is done by various data mining algorithms like Aprori algorithm, KNN, Decision tree and so on. Aprori algorithm: gets the combination of the products and recommend on the basis of combos. KNN: K – Nearest Neighbor checks for the attributes and then give the possibility of user buying the product or not. Decision Tree: Splits the data into splitting factor (based on information gain and gini index) and then the decision is made.[14] There are more algorithms followed up but we see how the recommendations are made and suggested to the user. Now let’s study about the kinds of recommender system existing in the scenario.[19]

There are three kinds of recommender system namely:

1. Collaborative filtering
2. Content-based filtering

3. Hybrid recommender system

Collaborative Filtering: It explores the idea that relationship existing between products and people’s interest. The organization use collaborative filtering to analyze this relation and give a recommendation to the user which he may like or enjoy. The relationship is made on the basis of customers past purchase records or views or in all we can say his behavior towards any product and that creates a connection with different users that posses the same kind of interest and hence the recommendation is made which gets a high chance of being liked by the user. The popular example for collaborative filtering is Netflix; everything which is displayed on their site is carried out by the selection of their customer, which if made regularly or frequently enough, then it gets turned in to a recommendation. [14]Netflix ranks these recommendations in such a way that highest ranking items are more visible to the user in order to be get clicked or selected by them. The advantage of the collaborative filtering is that is gives a broader exposure to the consumer for many different products and this helps in the continuity of the relationship between site and the user and hence creates a better experience. The problem with this kind of filtering is that if a new product comes in, then we does not have sufficient amount of the event history that can help in making a recommendation and it’s the point where this system fails.

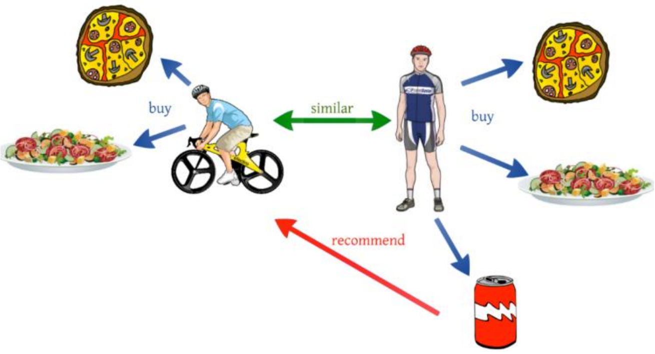

Fig.2. a. Collaborative Filtering

Content-based filtering: Content based filtering focuses on the description or we can say attributes of product and then compares it with the profile or preferences provided by the user and gives a recommendation.[7] Keywords or labels play a vital role in content based filtering as these words help the system to indicate the preferences of the user and provide recommendation for its likings. The data about the user is taken explicitly (by the likes or rating provided with it) or implicitly (by the click events or view events) so it means more input provided by the user will give a better recommendation. [2]Example: suppose we are running an online video streaming website where different movies are displayed and are of different genres so what our recommender system will do, it will look for the users interest and behavior and then sees the product attributes and if matches are there then it provides a suggestion and creates a recommendation which may be most liked by the user. And also this kind of recommendation gives more value to the product attributes and becomes a personalized one. It removes the problem of cold start that is a new products also gets the engagements and are pushed up and also it is free from user that is it does not need user interferences only the profile is required but this system lacks when a new user comes in as the user profile is not having enough of the information that a recommendation can be made.[14]

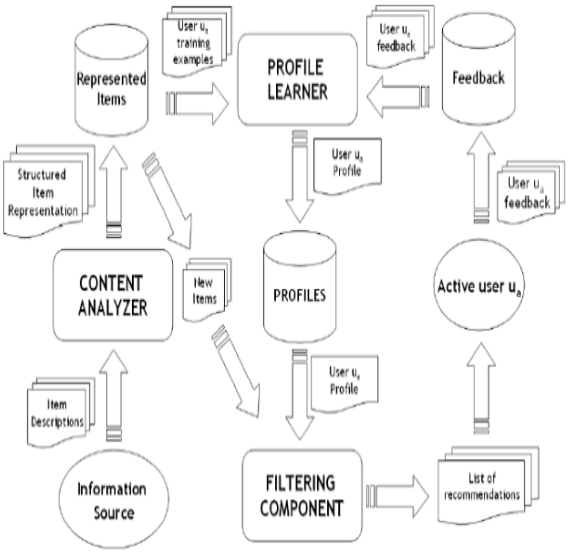

Fig. 2.b Content Based Filtering

Hadoop Distributed File System (HDFS)

We all know HDFS plays a very important role for handling big data, basically it is the most preferred because of the various advantages provided by it like elasticity and the better user interface with better understandings. [19]HDFS is the component of hadoop which is used to store the files in a fancy format i.e. it stores the files to different nodes and then accessibility of the data becomes fast as the data are accessed from the nodes and there are various components that look after different tasks assigned to them. [20] The components of HDFS were name-node, data-node, task-tracker and job-tracker but with the entry of YARN the things are slightly changed and now there is a new scenario. Previously the hadoop consists of two major components namely HDFS and mapreduce. HDFS was a normal file system which was used to store all sort of data like binary file text file and so on but now it had resource manager which handles a lot of work of hadoop at user end framework known as application framework, yarn has now become a system on which not only we can run MapReduce but also other applications whether it is any query or any other application. Previously hadoop was a drone application that is only one process can be worked out but now it supports parallel processing of the data at various levels. [11] By the entry of YARN we get resource manager which looks after how the data is stored and on which node it will be stored.

There are many algorithms made for scheduling the nodes and getting with the data but we are particularly interested in HCPA algorithm which looks after the health of the node, priority of the task, capacity of the node and availability of the node. We create a queue which is refreshed after the tasks are finished and then the node with highest score gets the data and compute on it. This is very efficient and a successful algorithm for both heterogeneous and homogeneous nodes and system and made system a satisfied to a high extend.[17]

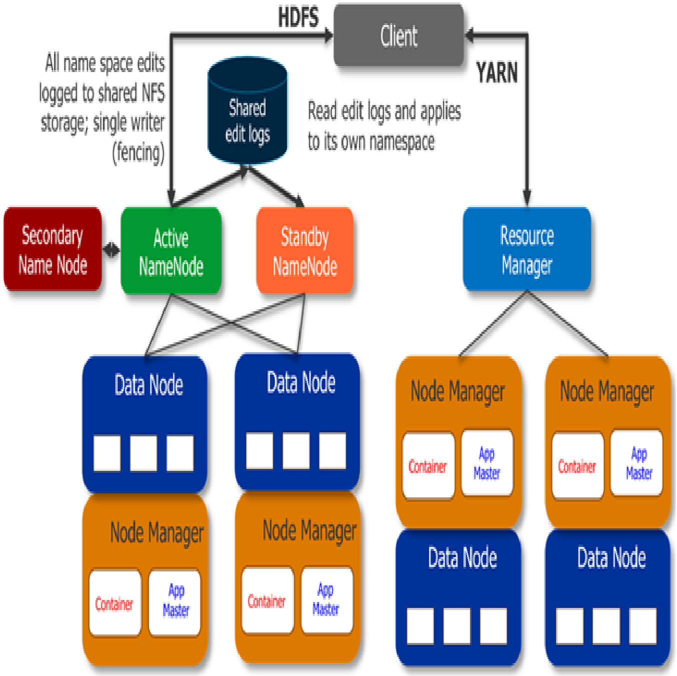

Fig. 2.c. HDFS with YARN

IV. CHALLENGES IN EXISTING SYSTEM

1. Both the recommender systems were not able to control the new entries of the product or user and hence the products were never boost up.
2. The data scarcity was a problem noticed in e commerce and the use was not able to get personalized recommendation.
3. Feature selection and dimensionality reduction is a big task or continuous data.
4. HPCA scheduling looks after overall capacity of the nodes and still passes the tasks.
5. HPCA scheduling don’t let other nodes to handle the task so not every nodes are utilized.
6. Some attributes become dominant in the recommendation system and hence those are needed to be looked after.

V. PROPOSED METHOD

A. Block Diagram

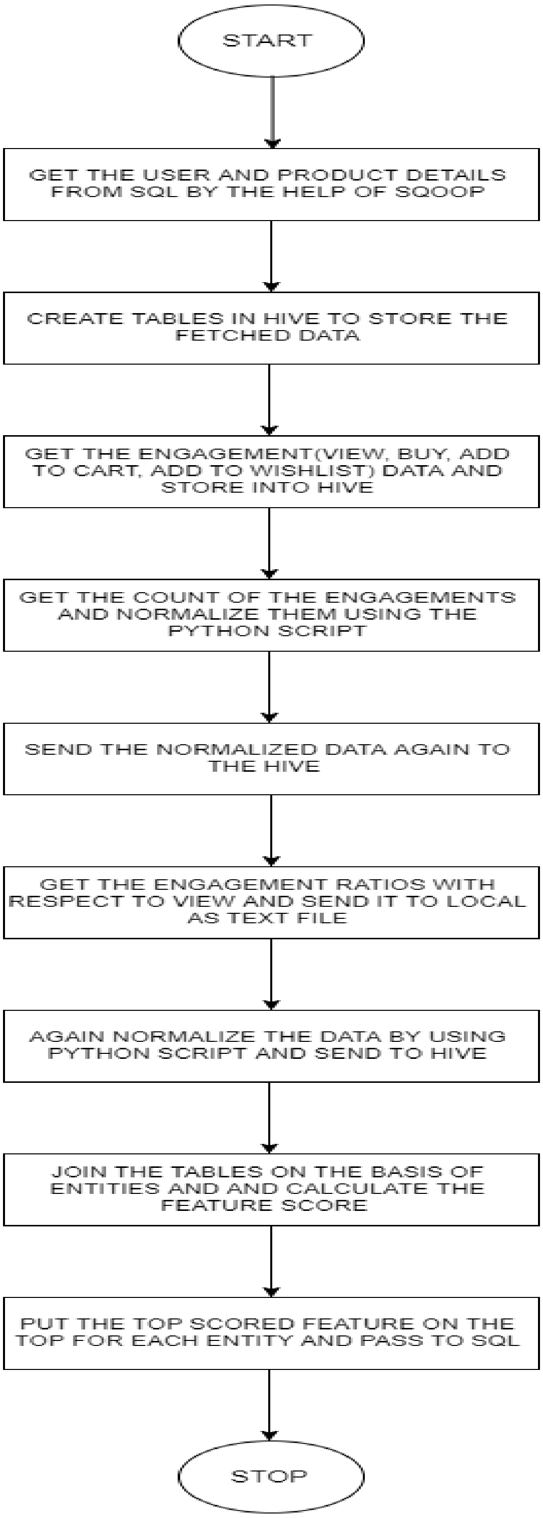
START
GET THE USER AND PRODUCT DETAILS FROM SQL BY THE HELP OF SQOOP
CREATE TABLES IN HIVE TO STORE THE FETCHED DATA
GET THE ENGAGEMENT(VIEW, BUY, ADD TO CART, ADD TO WISHLIST) DATA AND STORE INTO HIVE
GET THE COUNT OF THE ENGAGEMENTS AND NORMALIZE THEM USING THE PYTHON SCRIPT
SEND THE NORMALIZED DATA AGAIN TO THE HIVE
GET THE ENGAGEMENT RATIOS WITH RESPECT TO VIEW AND SEND IT TO LOCAL AS TEXT FILE
AGAIN NORMALIZE THE DATA BY USING PYTHON SCRIPT AND SEND TO HIVE
JOIN THE TABLES ON THE BASIS OF ENTITIES AND AND CALCULATE THE FEATURE SCORE
PUT THE TOP SCORED FEATURE ON THE TOP FOR EACH ENTITY AND PASS TO SQL
STOP

Figure 3.a: workflow of the recommender system

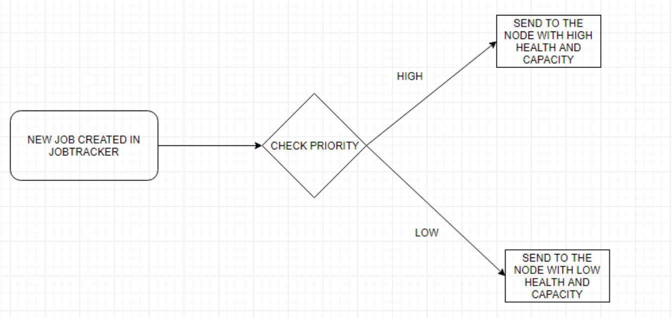
SEND TO THE NODE WITH HIGH HEALTH AND CAPACITY
HIGH
NEW JOB CREATED IN JOBTRACKER
CHECK PRIORITY
LOW
SEND TO THE NODE WITH LOW HEALTH AND CAPACITY

Figure 3.b. HDFS Scheduling

VI. MODULE DESCRIPTION

A. *SQOOP*

Sqoop is a tool which was designed by apache and is used to transfer data from a normal traditional database to HDFS. It can be defined as a path through which the structured data is entered into the Hadoop system and gets stored into our computer as per the requirement. With the help of sqoop, We can import data from a various platform like Oracle, MySQL and so on.[1] Generally, the communication between the applications with the relational database creates a huge amount data or we can say big data and this kind of big data is preserved in relational database servers. But now when Hadoop is there to handle all the big data problems, it needs a component or we can say a tool to interact with relational data for the transfer of big data that is stored in it; here the sqoop makes an entry and solves this big problem and gets a position in Hadoop ecosystem. [13]In short, we can define sqoop by saying as Hadoop to SQL and SQL to Hadoop. Sqoop uses two commands mainly sqoop export and sqoop import. Sqoop import will help in transferring the tables from our traditional database to HDFS and each row will be assumed as record and these records will be stored as a text data in HDFS for further task or computations as required by the system. The second command that comes in is sqoop export, by the name we can say that it is used to transfer the file from HDFS to back to a traditional database. The file passed as an input is filled with records which are termed as rows in SQL and these records are read first and then parsed into the records and can be delimited according to the user.[8]

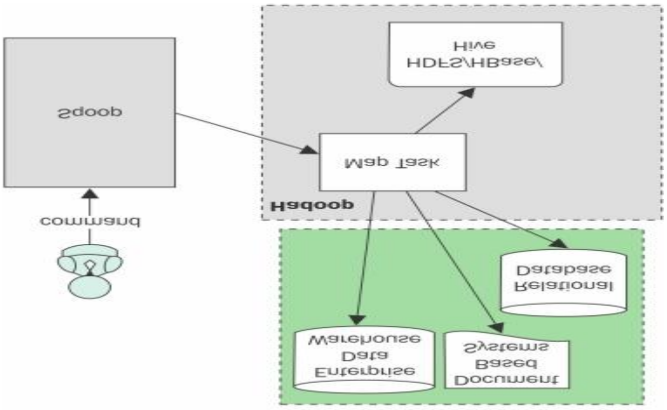

Figure 3.c. SQOOP Structure

B. *HIVE*

Hive can be termed as a tool for warehouse infrastructure which is used perform a computational task on the structured data or we can say that to process the structured data residing in Hadoop. By the entry of Hive the querying on the table and analyzing it had become easy and sort to a high extent.[9] Previously Hive was brought into existence by Facebook and afterwards, Apache Software Foundation took this tool and enhanced it by adding more features to hive and named it as Apache Hive. [10]We should consider that Hive is not designed for OLTP and not a language that is used for row-level updates or real-time queries and one more thing should be kept into mind that hive is not a relational database like other traditional ones.[7] Hive is particularly designed for OLAP and is used to process the structured data on HDFS and also stored the schema for it in the database. The big advantage of HiveQL is that it provides users a SQL kind of language to interact with the data and process it which is very fast to learn and easy to go with so it becomes familiar for the developer and moreover it provides us scalability and extensibility to our system.[14]

The process of executing a HiveQL query in the Hive tool:

1) First, the hive query is sent to the database driver (such as ODBC, JDBC, etc) by the help of command line or web UI for its execution.

2) Now the driver accepts the query and by the help of compiler the query is checked and all the requirements done by the query is noted down by the name of the plan.

3) The metadata request is being passed by the compiler to any metastore of the database.

4) Now when we get the metadata as a response from the metastore.

5) The compiler will observe and check the asked requirement and resends the plan to the driver. Till this step, we have completed the parsing and the compiling of the query.

6)Execution engine will accept the plan and execute it.

7) Now the job will be executed in a MapReduce style, basically, the jobTracker receives a job from execution engine and then JobTracker seeks for suitable TaskTracker to give the job and in this way, the MapReduce job is executed.

7.1) While the jobs are being executed, the execution engine also start executing metadata operation with metastore and hence the work go in parallel.

8) The result is being fetched from the DataNodes and the resultant value is sent to the respective driver.

9) The driver accepts the result and sends it to the interface.

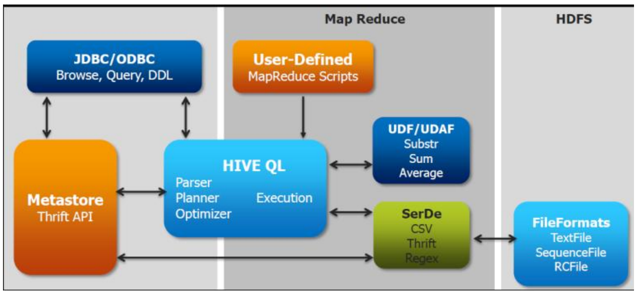

Figure 4.a. HIVE architecture

C. *Python*

Python is a high-level programming language which is now commonly used and accepted by many firms and organization because of the vast and a good number of libraries present which helps in coding and understanding of the code easy and fast. All we need to do is download a package and call its function and rest all is computed by the python.[21] Python is a language which is highly used in big data because of scalability feature which helps computing the large codes an efficient way. Java is another competitor for python but due its large scripting style it is not usually used. Python is dynamically written language that is we really don't need to specify the data type of any variable, it converts the variable into the required type by looking at the data which makes it easy for the programmer to type the program without being aware of the type casting stuff, moreover python have a vast range of data types available which are easy to understand and use. [12] The syntax of python is all about being in one line and keeping the code short and simple. There is nothing like braces and all, you just need to keep yourself simple and code simpler and this the beauty of the python, you perform long and huge operation like K means clustering and all but you only use a single function and here your task is done.[6] Python also helps you manage the memory efficiently and also supports alot of programming paradigms namely imperative, procedural, object-oriented and functional with a large number of available libraries.

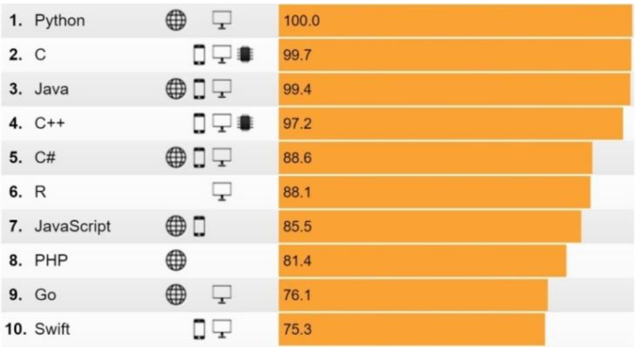

Figure 4.b.: Program Ranking

VII. PROPOSED DESIGN

STEP 1:

We will import the files from MySQL to HDFS by the means of sqoop. These are the tables which describes the user details and product details. These are the joined tables of various records available, thanks to purplle.com for their data for this project. So basically we will fetch the data which consists of the details of users as well as the product so that we can match their entities and get the result.

STEP2:

Now we will create tables in hive so that we can store the fetched MySQL into our hive for our computation. We are creating table just to save the details of user and the products so that we can join the further tables and get the score in the upcoming steps.

STEP: 3

Now by the help of clickstream data we are now capturing the events data, here the events we are considering is view, add to cart, add to wish list and buy and will rank these events accordingly. So we calling these events into hive and creating table of various days so that the product should get equal importance.

STEP 4:

Now we will use python scripts in order to normalize the data. Normalizing the data is very important step because this will help to give importance to all the products irrespective of the trends and promotions going on. In this step we will normalize the event count of the products and for that we need to send the data back to local so that python can fetch the file and work on it.

STEP 5:

After normalizing the data we need to resend it to the hive for the computation of the score. So basically in this step we are taking the normalized data and sending it to the hive for the further steps computation.

STEP 6:

Now we will calculate the ratio with respect of the view to other events because view is the first step that occurs and hence it is needed to be compared with the other events so that there won't be high weight-age event in the system.

STEP 7:

After getting the ratio now they are needed to be normalized in order to make the score so for that we need to send the data to the local so that python can access it and normalize the score easily

STEP 8:

After this call the normalize data to the hive and calculate the score by giving some weightage to various events and make the highest scored data for each entity on the top and push it to the MySQL.

VIII. RESULT & ANALYSIS

For testing our system, we took some events and observe the changes in them. We took few major events and check their result for a week and here are the results that we got.

The events those were taken into consideration are:

a) Feature Impression
b) Widget Impression
c) View

d) Add to Cart
e) Add to Wishlist
f) Notify Me
g) Widget Click
h) Feature Click
i) Buy

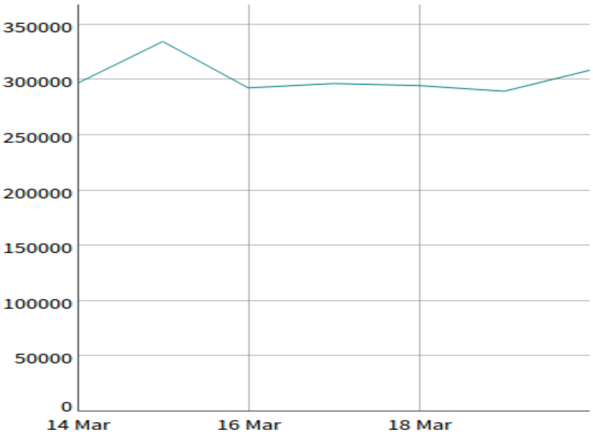

Fig 5.a. feature impression vs time

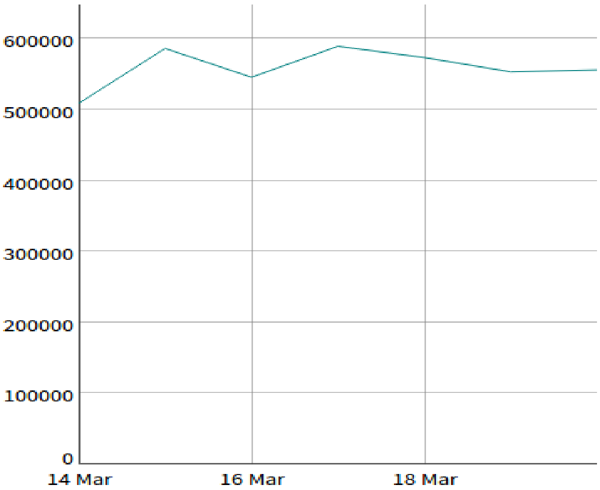

Fig 5.b. widget impression vs time

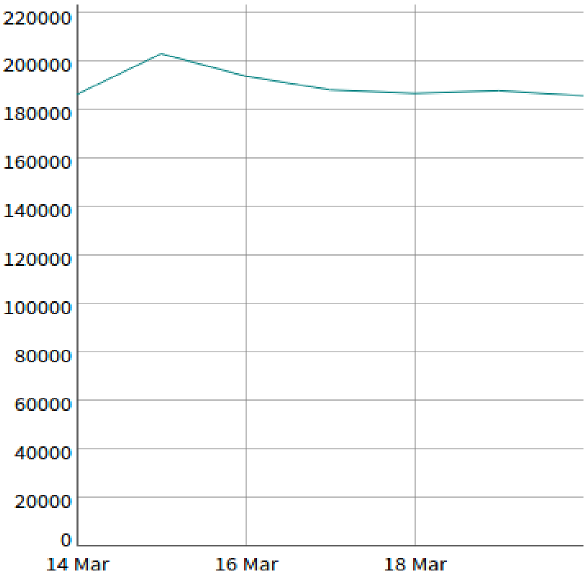

Fig 5.c. view vs time

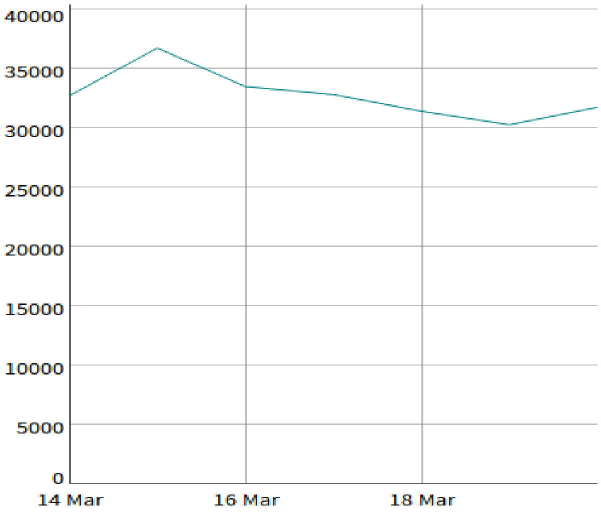

Fig 5.d. add_to_cart vs time

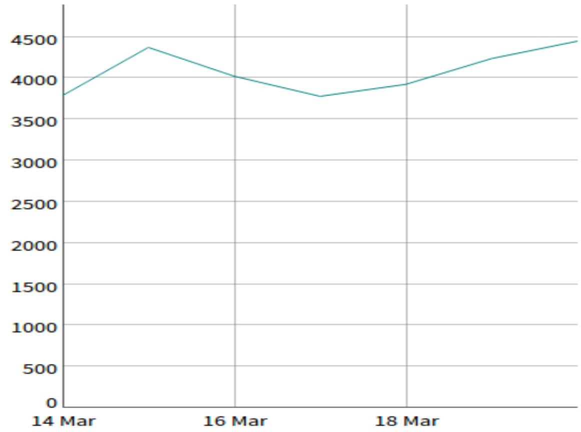

Fig 5.e. add_to_wishlist vs time

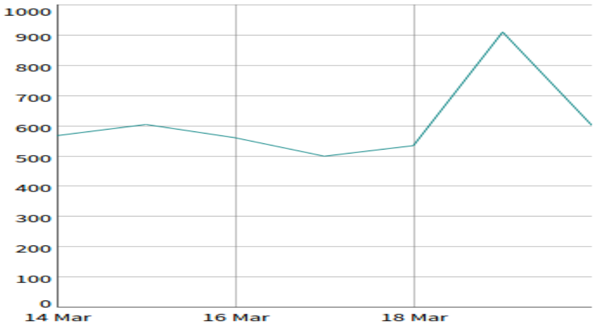

Fig 5.f. notify_me vs time

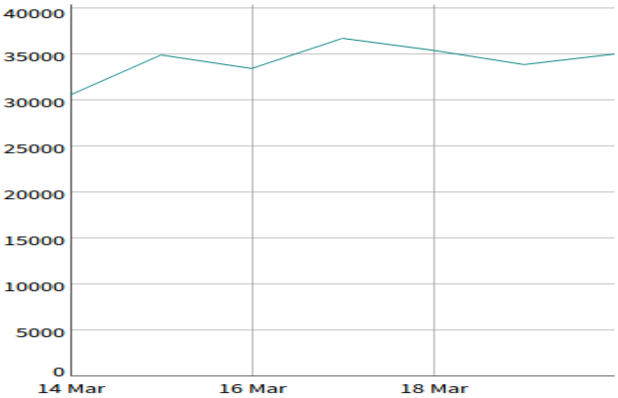

Fig 5.g. widget_click vs time

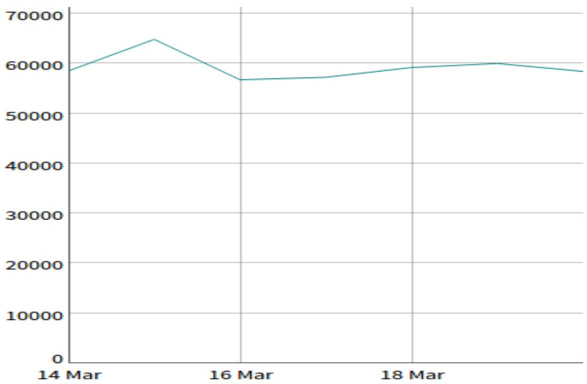

Fig 5.h. feature_click vs time

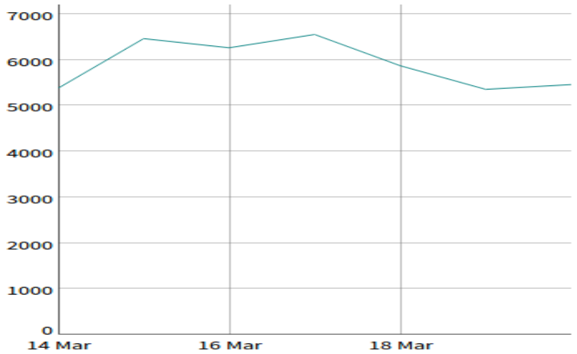

Fig 5.i. Buy vs time

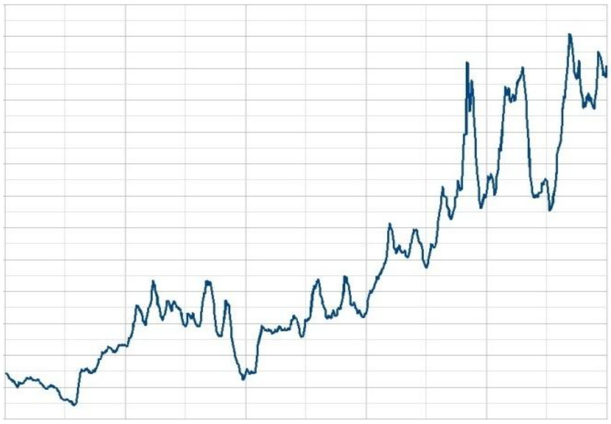

Fig. 6.a. increase the CTR rate (time vs CTR)

VIII. CONCLUSION AND FUTURE WORK

We have analyzed the result of our recommendations and we can see an increase in the graph in our recommendation system. Our system focuses on the user satisfaction on the basis of the features of the products so the user gets its favorable product in the top recommendations. We deal with the data which is real time in nature and compute the result which shows a hike. Not only this, our focus is towards getting the recommendation with parallel processing which is also important and the things are done very well and optimized. So we get the better hits on our recommendation and increased the count of the product. And we have implemented HPCA algorithm that helped in the fast computation of the data and the scheduling of the tasks was done properly as well. So we can say that our system is smart enough to work effectively as well as efficiently in many use cases. The graph can give a clear study of it.

REFERENCES

1. Anderson, M. R., & Cafarella, M. (2016, May). Input selection for fast feature engineering. In Data Engineering (ICDE), 2016 IEEE 32nd International Conference on (pp. 577-588). IEEE.
2. Wang, S., Tang, J., Wang, Y., & Liu, H. (2018). Exploring Hierarchical Structures for Recommender Systems. *IEEE Transactions on Knowledge and Data Engineering*.
3. Sodera, N., & Kumar, A. (2017, May). Open problems in recommender systems diversity. In Computing, Communication and Automation (ICCCA), 2017 International Conference on (pp. 82-87). IEEE..
4. Chen, F., Liu, J., & Zhu, Y. (2017, June). A Real-Time Scheduling Strategy Based on Processing Framework of Hadoop. In Big Data (BigData Congress), 2017 IEEE International Congress on (pp. 321-328). IEEE.
5. Liu, J., Sun, L., Higgs, R., Zhang, Y., & Huang, Y. (2017, July). The Electronic Commerce in the Era of Internet of Things and Big Data. In *Computational Science and Engineering (CSE) and Embedded and Ubiquitous Computing (EUC), 2017 IEEE International Conference on* (Vol. 2, pp. 360-363). IEEE.
6. Dong, P. (2015, September). Research on the evaluation of e-commerce website under the environment of big data. In *Instrumentation and Measurement, Computer, Communication and Control (IMCCC), 2015 Fifth International Conference on*(pp. 288-292). IEEE.
7. Hajeer, M. H., & Dasgupta, D. (2017). Handling Big Data Using a Data-Aware HDFS and Evolutionary Clustering Technique. *IEEE Transactions on Big Data*.
8. Suguna, S., Vithya, M., & Eunaicy, J. C. (2016, August). Big data analysis in e-commerce system using HadoopMapReduce. In *Inventive Computation Technologies (ICICT), International Conference on* (Vol. 2, pp. 1-6). IEEE.
9. Shobana, V., & Kumar, N. (2017, March). A personalized recommendation engine for prediction of disorders using big data analytics. In *Innovations in Green Energy and Healthcare Technologies (IGEHT), 2017 International Conference on* (pp. 1-4). IEEE.
10. Sogodekar, M., Pandey, S., Tupkari, I., & Manekar, A. (2016, December). Big data analytics: hadoop and tools. In Bombay Section Symposium (IBSS), 2016 IEEE (pp. 1-6). IEEE.
11. Kurniawati, D., & Triawan, D. (2017, November). Increased information retrieval capabilities on e-commerce websites using scraping techniques. In Sustainable Information Engineering and Technology (SIET), 2017 International Conference on (pp. 226-229). IEEE.
12. Wen, H., & Zhao, J. (2017, December). Aspect term extraction of E-commerce comments based on model ensemble. In Wavelet Active Media Technology and Information Processing (ICCWAMTIP), 2017 14th International Computer Conference on (pp. 24-27). IEEE.
13. Chu, P. M., & Lee, S. J. (2017, October). A novel recommender system for E-commerce. In Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), 2017 10th International Congress on (pp. 1-5). IEEE.
14. Sterlin, S., Sandhya, A., Merlin, S. S., & Sam, B. B. (2017, March). A review on e-commerce recommender applications. In Computation of Power, Energy Information and Commuincation (ICCPEIC), 2017 International Conference on (pp. 387-391). IEEE.
15. Putra, A. A., Mahendra, R., Budi, I., & Munajat, Q. (2017, November). Two-steps graph-based collaborative filtering using user and item similarities: Case study of E-commerce recommender systems. In Data and Software Engineering (ICoDSE), 2017 International Conference on (pp. 1-6). IEEE.
16. Yanfang, Q., & Chen, L. (2017, December). Research on E-commerce user churn prediction based on logistic regression. In Technology, Networking, Electronic and Automation Control Conference (ITNEC), 2017 IEEE 2nd Information (pp. 87-91). IEEE.

17. Archana, G. K., & Chakravarthy, V. D. (2015, February). HPCA: A node selection and scheduling method for Hadoop MapReduce. In Computing and Communications Technologies (ICCCT), 2015 International Conference on (pp. 368-372). IEEE.
18. Jing, W., & Tong, D. (2016, October). An Optimized Approach for Storing Small Files on HDFS-based on Dynamic Queue. In Identification, Information and Knowledge in the Internet of Things (IIKI), 2016 International Conference on (pp. 173-178). IEEE.
19. Atmaja, I. P. M., Saptawijaya, A., & Aminah, S. (2017, September). Implementation of change data capture in ETL process for data warehouse using HDFS and apache spark. In Big Data and Information Security (IWBIS), 2017 International Workshop on (pp. 49-55). IEEE.
20. Kulkarni, P. G., & Khonde, S. R. (2017, October). HDFS framework for efficient frequent itemset mining using MapReduce. In Intelligent Systems and Information Management (ICISIM), 2017 1st International Conference on (pp. 171-178). IEEE.
21. Zhang, M., Liu, F., Lu, Y., & Chen, Z. (2017, July). Workload Driven Comparison and Optimization of Hive and Spark SQL. In Information Science and Control Engineering (ICISCE), 2017 4th International Conference on (pp. 777-782). IEEE.