Advances in Computational Sciences and Technology
ISSN 0973-6107 Volume 10, Number 6 (2017) pp. 1691-1705
© Research India Publications
http://www.ripublication.com

A Hybride Product Recommendation Model Using Hadoop Server for Amazon Dataset

Jitendra Soni*[1]
Assistant Professor, Department of Computer Engineering, Institute of Engineering & Technology, Devi Ahilya Vishwavidyalaya, Indore M.P., India.

Imran Uddin[2]
Lecturer, Department of Computer Science and Engineering
Prestige Institute of Engineering Management and Research, Indore M. P., India.
E-mail: iuddin39@gmail.com

Abstract

Recommendation in daily life simplifies the searching process to everyone. In day-to-day life, we face the situation of getting the advertisement on web browsers as per our searching history. Internet user in today's world having the value of billion. Every required information is easily accessible using web browsers and internet. The new enhancing field of internet is online shopping sites. From a piece of rack to cars and high cost gold, everything is available online. Whenever customer attempt to perform search operation, need to browse huge amount of data. Many recommender systems are designed as per the need, but as scalability increases processing and analysis becomes tougher. Our work proposed a solution in which collaborative filtering algorithm will be used. The work analyzed customer behavior and presents the efficient technique. Apache Hadoop Server is used to employ it. As data regarding customer query become large hence Big Data processing need to be performed. Each product has certain popularity index and similarity. Based on the customer searching, recommendations are given. Amazon dataset is used in evaluation and recommendation. Performance evaluation is always necessary hence computation time for individual and cluster of customer is main factor in analysis.

Keywords: BigData, Hadoop, Amazon Dataset, Product Recommendation, Collaborative Filtering, Customized customer behavior analysis

1. INTRODUCTION

In today world out of ten people, eight are using internet. Hence data on sites and browsers are increasing at fast pace. Several devices are responsible for huge amount of data generation. Mobile, laptops, televisions, online shopping sites, social networking sites can be big source for data generation. Rate of data generation is increasing with the proportion of time. A study concludes that 90% of total data has been generated into last 3 years. Online shopping and social networking sites are responsible for this big generation. Hence, at data centers of Google, Amazon, Facebook terabytes of data is reducing every day. Here, the study observes that this large amount of data carries variety of characteristics and can be good source for knowledge generation. Data mining algorithms can be use for characteristics extraction. Popularity of web portals reflects from originality or relevance of data available on portals. Accurate and relevant results can help to increase the popularity as well user satisfaction limits [1].

Every online well-known environment, is trying to enhance the business as much as possible. Any government organization, self-businesses is trying to gain more economy from internet world in these days. Any possible thing, which one can think of, is available in shopping sites today. Hence, there is strong need of the environment, which dealt with all such new requirements. Novel efficient systems, which can resolve current challenges, are required to design. While designing this system the privacy, analysis of the system should be high enough to meet the customer needs. The architecture designed for this should be distributed instead of centralized to recover from any kind of disaster activity. Load balancing and distribution can also be good way for simplification of work [2].

As we known cloud, computing has achieved a different level of use in the current era. Hadoop ecosystem can also be used for distribution and load handling purpose. It is capable to develop any kind of scalable and distributed applications. Hadoop works on the concept of distribution system so MapReduce component has been developed to achieve these phenomena. Hadoop Server has been developed and manage by

Apache Software Foundation comes with open source license [1][3].

To store such vast amount of data database like MySQL are not enough, hence new database are also framed for such purposes. Examples of such database can be MongoDB, NoSQL and pig. Every database follow some file structure Hadoop uses HDFS which is Hadoop distributed file system. Data warehouses are designed, in Hadoop the database is named as Hive [4].

Map Reduce perform mapping of data and based on the mapping only desired data is reduced. Hence only desired data is filtered instead of complete data. The language used in MySQL is replaced by language like Pig which took a bit more time to query such large dataset. Instead such querying language is capable of handling large dataset. Once database is designed and queried the connectivity with programming model like Hive is performed using Sqoop. For simplest need and small requirement the Hadoop has designed the specific tools which can be used.

Most of the time Hadoop is used in Linux environment, which work more effectively in cost-minimization manner. Usually the ma-reduce is highly used by lots of business organization and websites. The reduction of data from terabytes of data solves a huge problem of answer resolving.

Now discussing the file system of Hadoop, which is HDFS commonly called as Hadoop distributed file system. The technology used at the back end is java. The advantage of such large file system is that, the scalability and efficiency can be gained properly. In vast quantity of data there are high chances that faults and code breaks will occur. Hadoop efficiently managed it, as the replication is performed at distributed servers so any loss of information from one end can be recovered from other. The data stored in databases of HDFS can be of wide variety hence structured and structure-less commonly called as schema-less data can be access in HDFS. The entire communication will not be abort still continuously manages in best possible way.

The distribution file system for better accessing suffers some drawbacks one is security threat to network. The more wide the network more is the chances of its failure. Basic security need in Hadoop is quite required, but even the basic operations of security like encryption on such vast data is very much complex. To identify each individual big data even contains the personal identifiable information of each. If anyone can access that information then it is just the betrayal to information security. Hence the one which can access information as well as proper classification of data based on which information is how much important criticality is assigned. Hadoop architecture and its key parts are shown in below figure.

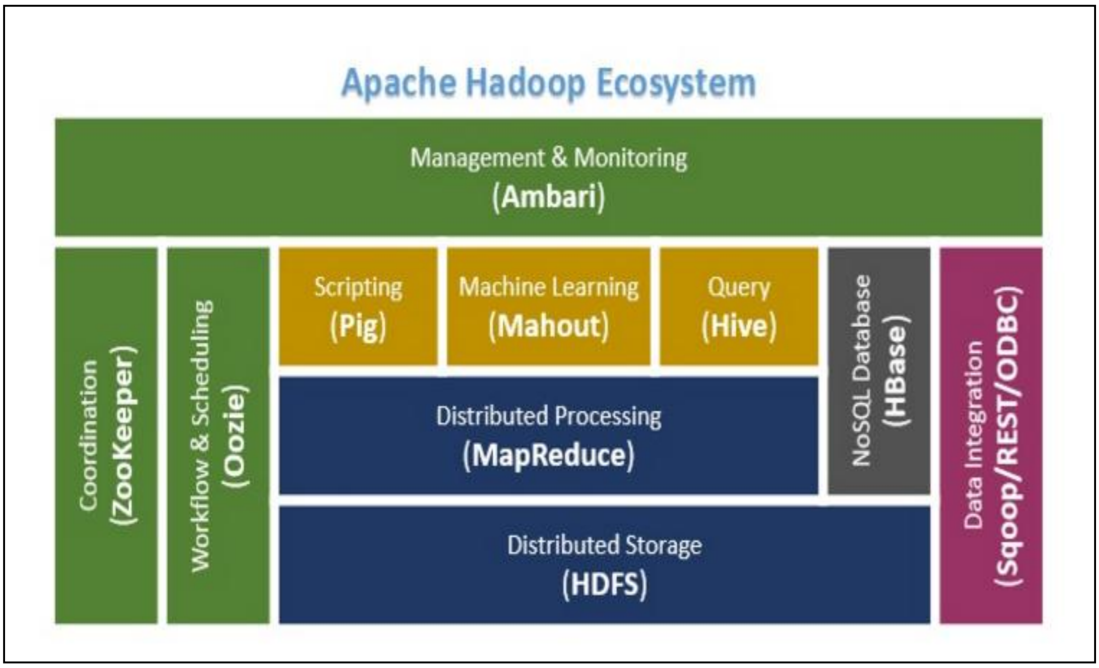


Figure 1: Block Representation of Hadoop Components[1]

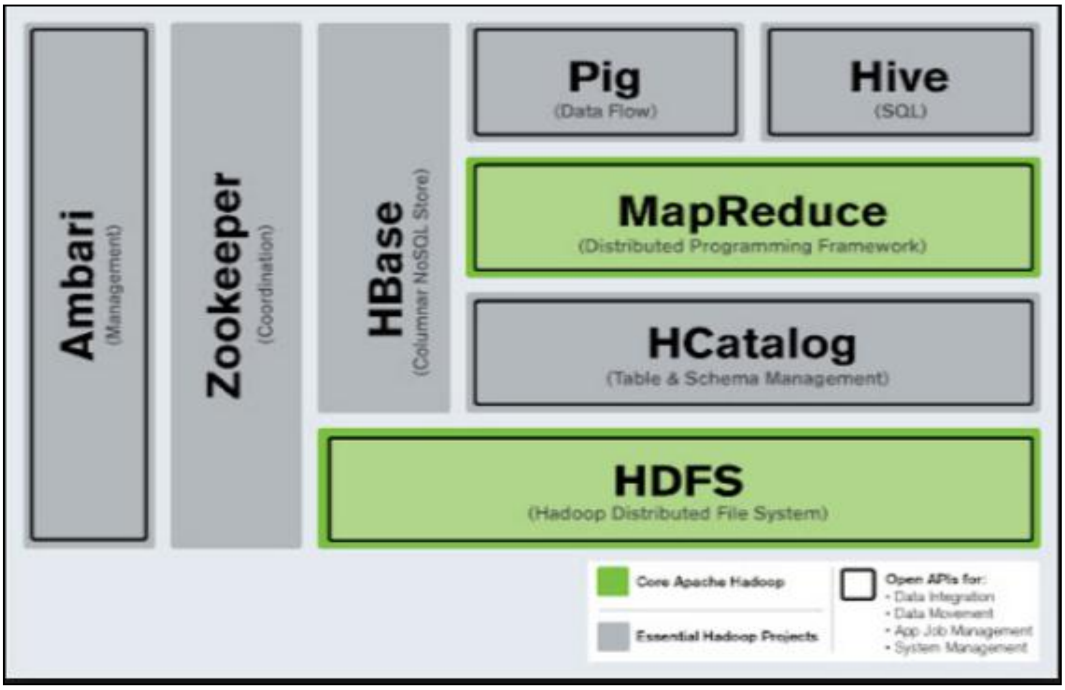


Figure 2: Block Representation of Core Apache Hadoop Server [5]

2. LITERATURE REVIEW

Anindita A Khade[3] suggest that if one is capable of properly analyzing information then the buyer behavior can be useful to give suggestion from the next time. Customer analysis in such big data helps in understanding the behavior in more efficient way. These results in enhancing business economy in terms of optimization, money value, fraud detection and other such needs. Author suggests that customer can use wide variety of approaches with decision trees algorithm. Decision trees are wide known and highly used classification approach. The analysis of customer using it gives best results. Wide variety of algorithm are designed which handles any kind of data and retrieves the best results. These approaches result in giving the better performance of the recommendations. But for this work, the visualization is necessary. The visualizations of this help in understanding the customer patterns to retrieve the results in enhancing way. A decision tree with visualization makes the perfect combination.

The diagram of the system with the flow can be given as:

- The input is carried out from HDFS of customer dataset and loaded.
- Decision tree algorithm work by designing the required instances , all those instances are invoked on requirement basis
- Map Reduce is then used in Hadoop, current node belonging to the instance or not is analyzed with the help of map function. If in case the output index is not covered, it returns value and class level of instance.
- Value and class label with combination of index is counted with the help of reduce function and printing of it is performed.
- Parameters like entropy, gain, ratio of attributes are measured.
- The HDFS decision tree which is designed in Map Reduce framework will process the input dataset
- All the decision tree designed is then store using HDFS
- If any new data set arrives from the source which is any browser it accept and tested
- The categorization of data is performed and rules regard to it is framed.
- Decision tress will help in generating visualization for the given dataset, this

visualization can be in form of pie-chart, graph etc.

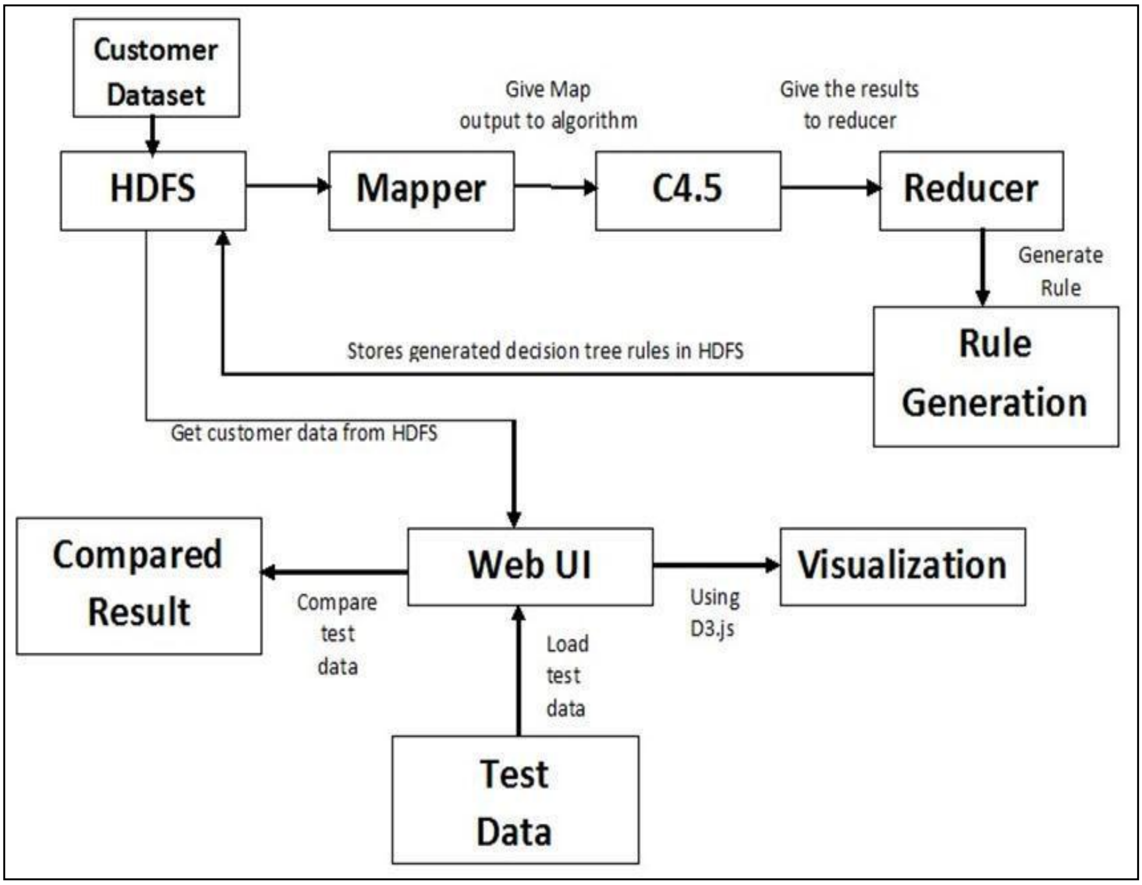


Figure 3: Existing System Architecture

As increase in data in web world the traditional system doesn't cooperate as it were not capable of handling such heavy loads. Our work designed the system using the concept of MapReduce framework along with decision tress. In it distributed file architecture is used. Cloud based technique by the help of Hadoop is used. The outcomes obtained were visualized using the tools of Hadoop.

Praveena Mathew [4] uses the recommendation system in which collaboration of three approaches namely content based filtering, book recommendation system and collaborative filtering is performed. It also uses association mining to have better recommendation. Just like grocery the book recommendation can help any student in better searching, hence the all the above given algorithm are combined to efficiently designed the recommendation of buyer's interest.

For better book searching, certain steps are designed in algorithm which can be given as:

1. The collections of books are retrieving from data warehouse and data cleaning is then performed. The data cleaning will result in bifurcating the unimportant books from the important ones. Thus the data present which is book in given case in retrieve for mining.

2. Further classification of books on the given subject is then done by extracting the even more relevant ones and data is further more refined, the process is known as data processing.

3. From the set of hundred only ten will remain in this stage by using the concept of filtering transactions. It is even minute level sub categorization. The process is called as filtering of transaction.

4. Now it may be possible that the extract ten books may be of different content. This content based filtering is performed on the basis of preferences given by user. Say for example among the given ten books, if book1 is searched and book6, book7 and book9 are related. Then the recommendation will be given of book6, book7 and book9. This scenario is called as content based filtering.

5. The recommendation given above is having the possibility that the book3 is also related but based on the opinion and popularity of books the others are suggested and this is not. This is key feature of collaborative filtering

6. Final recommendation from all the given books are performed which have highest rating among all. The comparisons from all the books are performed and rating is given.

Once the book is recommended and buyer has ordered it then that purchased book is stored as entry. All the other books which is also ordered in past are also stored. Now this history will be used in content and collaborative filtering and ratings of books has changed based on the number of times it is purchased or viewed. Once we overcome this issue the best recommendation are provided. Now association based mining is performed which will associate the authors which users mostly relate with, the area of interest and stream in which one is performing shopping again and again. Thus from large set of data, lot of data can be filtered and efficiently managed. Such recommendation system is highly useful for student and faculties and is really good recommending scheme.

3. PROBLEM STATEMENT

Web browsers are accepting terabytes of data in blink of eye. Everyone is internet friendly these days and tries to utilize it in best possible ways. At one point the businessmen are trying to elaborate themselves as much as possible whereas the access point having user ants data frequently and flawlessly. Big data is way of storing data in structure and schema-less way, the data present may be in relational way or in non-relational way. Massive amount of data is present have variety of datasets. Analysis performance and conclusions on it quite tough to achieve and extract.

Big data is kind of data storage technique which is capable of storing such vast amount of data. It just changes the traditional scenario of storing and retrieving data. Large numbers of online sites are created in every single day with lots of new idea. Investors rated in top list are also from web world and trying to make it better. No matter what the product is the online sites are trying to make it available.

Of all the popular sites the one which the regular net user will answer is online shopping sites. It is present with wide variety of product and in wide ranges as well. Depends on what one is accessing it start to suggest the similar product to satisfy client as much as possible. This scenario saves lots of time. But it comes with certain issue if one starts to find better, its expectation start growing day to day. The customer wants to have data in time saving manner. In the worst traffic everyone prefers to shop from home. Everyone wants things of their personal interest with mindset choice. The behaviors of customer change all the recommendations say for example the kid and younger will mostly search the different things of their interest. Therefore based on customer behavior the facilities should be provided, which is quite tough. One way can be use of history of user shopping to make recommendations.

Analysis on user behavior help in drawing variety of algorithm, the examples of these can be discovering hidden relationship among customers, cause/effect analysis, and prediction and visualization behavior. Along with all this for even better system the recommendation plug-in are designed. These plug-in will help in even better shopping sites. For accuracy in results mining techniques like collaborative and content mining is merged with other mining solutions.

The study suggests that the work done have used the product recommendation by using the concept of popularity index and ranking. The estimation is figure out by customer view and its ranking. On which criteria the product should be categorized is big question till date. Different sites use different approaches. In usual form the sale of product is considered not the nature of it. Different algorithms are designed, based on different categories like product nature, product cost. Still the entire field is having wide chances of improvements. One such challenge is already existing stored data. The already stored data is not using big data. All the traditional work is carried out in small size data. The mining of relevant data is tough task, and data size is most considering factor. None of the solution technique uses a blend of behavior analysis and product nature.

The problem statement can be summarized as given below:

- Product recommendation algorithm should be enhanced according to large dataset

- Customer behavior should also be considered as the leading nature along with product characteristics.

The approach, which uses the similarity, popularity index and customer nature index in recommendation process, is yet to have certain improvements. Overcome such scenario is big challenge.

4. SOLUTION STATEMENT

In internet the volume of data is so large, that in one search the millions of result are returned. To get the one which we desire is such a major challenge. The situation worsens as the search return unwanted outcomes. The reason of these outcomes is poor performance of tools which are used in getting the desired data.

Recommendation system can be defined as the technique of presenting the data of user interest. The behavior of user is observed and suggestions are given in according to it. The set of required information is returned. Web personalization is one of the most used techniques in today scenario. In too much information loaded scenario, the web

The complete study concludes that web personalization in the field of product recommendation may help to improve the quality of content mining and recommend more useful and relevant piece of information.

The diagram below depicts the scenario suggested:

Data Collection & Preparation: Data loading is performed.

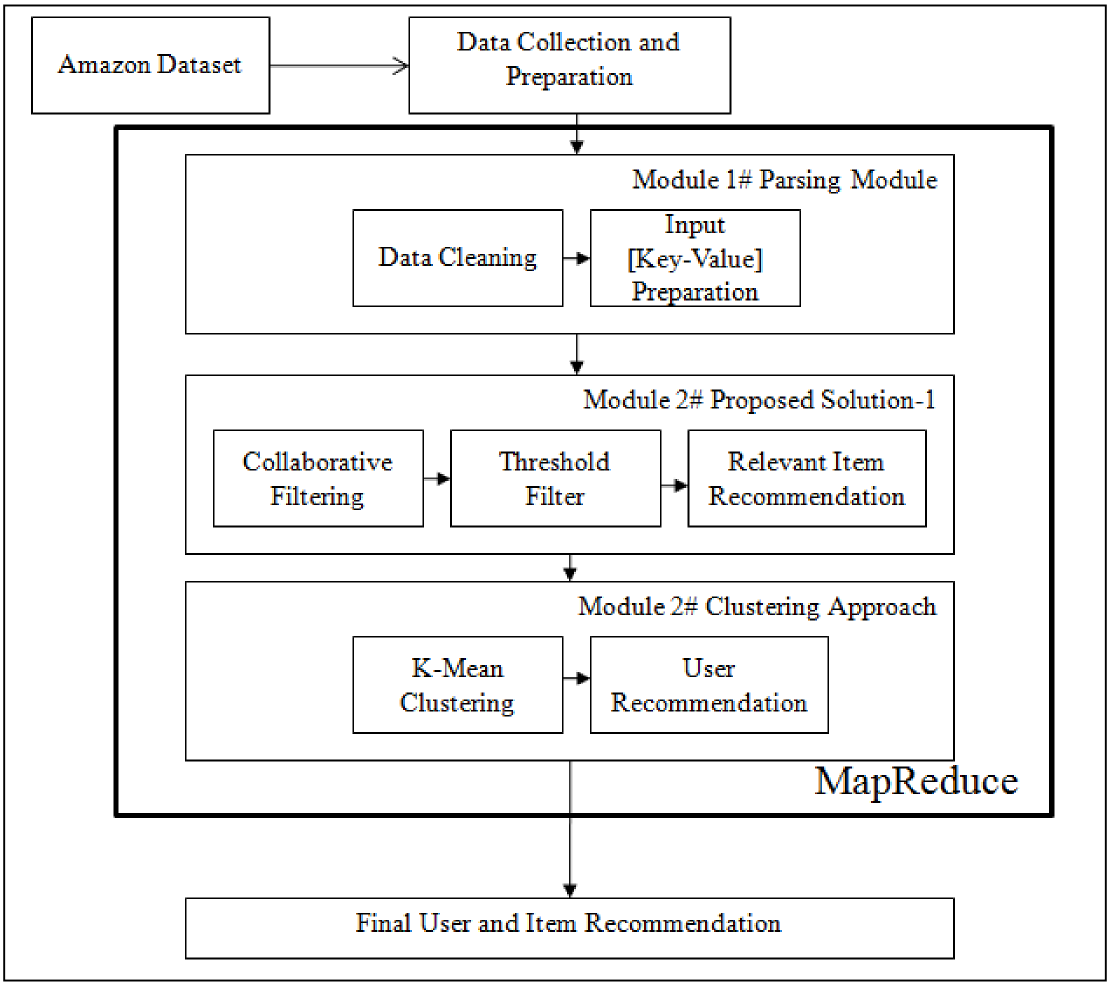

Amazon Dataset
Data Collection and Preparation
Module 1# Parsing Module
Data Cleaning
Input [Key-Value] Preparation
Module 2# Proposed Solution-1
Collaborative Filtering
Threshold Filter
Relevant Item Recommendation
Module 2# Clustering Approach
K-Mean Clustering
User Recommendation
MapReduce
Final User and Item Recommendation

Figure 4: Proposed Architecture

Step 1: Amazon Dataset has been considered with variable memory size from electronics and video games field.

Step 2: Next, complete dataset has been moved from normal location to HDFS.

Step 3: Data Collection and Preparation module load desire dataset from HDFS to Mysql for cleaning and processing purpose.

Afterwards, complete solution has been parsing through three different modules to generate results that are more accurate.

Module 1# Parsing Module

Step 4: Data Cleaning and Input Preparation method has been developed to remove unwanted anomaly and prepare input for mapper in form of Key and Value.

Step 5: Item recommendation algorithm has been developed. Here, Collaborative Filtering algorithm has been used to estimate the similarity between user characteristics and product type. Threshold value has been maintained to retrieve upper level recommendation.

Module 2# Item Recommendation Module

Step 6: Item recommendation algorithm has been developed. Here, Collaborative Filtering algorithm has been used to estimate the similarity between user characteristics and product type. Threshold value has been maintained to retrieve upper level recommendation.

Step 7: Next, A hybrid algorithm has been developed which create the cluster of relevant values from top recommendation and recommend product through user-product mapping.

Step 8: Next, A hybrid algorithm has been developed which create the cluster of relevant users and recommend product based on it.

Step 9: Final Recommendation based on both algorithm is generated as Outcome.

Step 10: The complete solution is implemented using Hadoop 2.7.1 server and Java technology.

Step 11: A Swing based user interface has been developed to seek user input.

Major significance of this project work is to extend the capability of product recommendation from small dataset to large dataset. Another one, it also involves two data mining approach K-Mean and Collaborative filtering, for final recommendation purpose. Involvement of these two algorithms would help us to sort the large data into small part with best result and once again filter them with double purity approach. The complete phenomena would provide the way of recommendation with double purity.

5. EXPERIMENTAL ANALYSIS

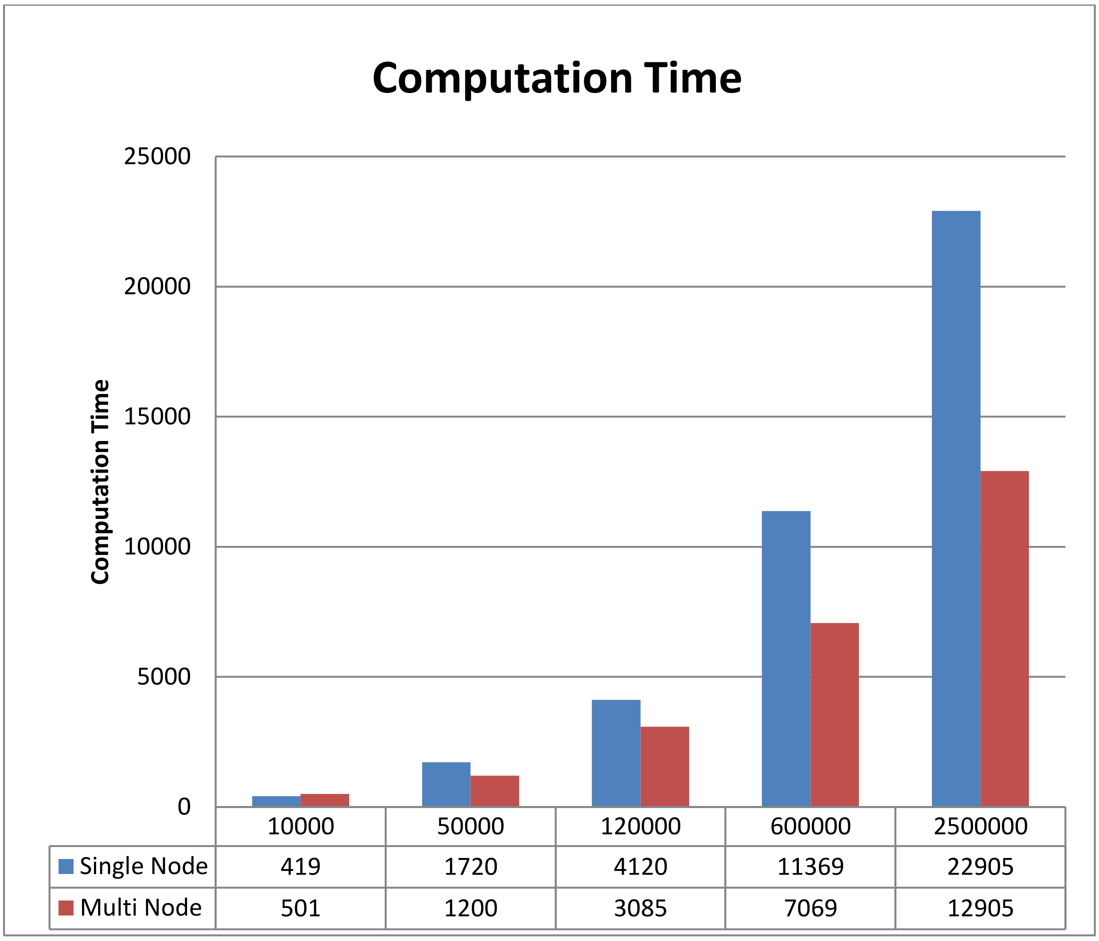
The real implementation of software system it to validate the feasibility of proposed solution. Another one, it would also help to evaluate the performance of proposed software architecture.

A Hadoop 2.7.1 ecosystem with MySql 5.0 has been used for distribution and background processing purpose. This complete solution has been developed using Net beans 8.1 IDE with Pentium Quad-Core I3 processor and 8 GB RAM. Single Node and Multinode configuration can also be used to explore the performance of proposed solution. Different file size from different records has been used as the input from (10MB to 519 MB) and evaluated for Threshold value of 100 users.

The complete experimental analysis is observed in terms of computation time for single node and multi node with different data input.

Table 1: Result Analysis

File Size	8.6 MB	41 MB	99 MB	213 MB	519 MB
Number of Records	10,000	50000	1,20,000	6,00,000	250000
Single Node	419 sec	1720 sec	4120 sec	11369sec	22905sec
Multi Node	500 sec	1200 sec	3085sec	7069 sec	12905sec



	10000	50000	120000	600000	2500000
🟦 Single Node	419	1720	4120	11369	22905
🟥 Multi Node	501	1200	3085	7069	12905

Figure 5: Experimental Result Analysis

6. CONCLUSION

A work cannot be accepted until we derive the conclusion. It is the closing statement to justify the value of proposed solution. The complete work concludes that duel algorithm model helps to maintain the level of purity with respect to enhancement into dataset.

Following observations are derived from the proposed solution.

1. Minimum 419 sec and Maximum 22905 sec computation time has been recorded for single node.
2. Minimum 501 sec and Maximum 12905 sec computation time has been recorded for multi node.
3. Multi Node Solution performs better than single node for large dataset

4. The complete experiment has been performed on variable size dataset, which is divided according to number of records.
5. Increasing computation time has been recorded with enhancement into data size but a stable value has been observed in multi-node for low size data.

The complete work concludes that proposed solution can be used for product recommendation system and can perform excellent for large dataset.

REFERENCES

[1] Fatima Rodrigues, Bruno Ferreira, “Product recommendation based on shared customer’s behaviour” published in Conference on ENTERprise Information System / Internation Conference on Project MANagement / Conference on Health and Social Care Information System and Technologies, CENTERIS / projMAN / HCist 2016, October 5-7, 2016.

[2] Riyaz P A, Surekha Mariam Varghese, “A Scalable Product Recommendations using Collaborative Filtering in Hadoop for Bigdata” published in International Conference on Emerging Trends in Engineering, Science and Technology (ICETEST - 2015).

[3] Anindita A Khade, “Performing Customer Behavior Analysis using Big Data Analytics” published in 7th International Conference on Communication, Computing and Virtualization 2016.

[4] Ms. Praveena Mathew1,” Book Recommendation System through Content Based and Collaborative Filtering Method” .

[5] Meena Jha, Sanjay Jha, Liam O’Brien, “Combining Big Data Analytics with Business Process using Reengineering”.

[6] JunBo Xia, “E-commerce Product Recommendation Method based on Collaborative Filtering Technology” published in International Conference on Smart Grid and Electrical Automation 2016.

[7] **Amazon Dataset:** http://jmcauley.ucsd.edu/data/amazon/

[8] Mohammed Alodib, Zaki Malik “*A Big Data approach to enhance the integration of Access Control Policies for Web Services*” IEEE ICIS 2015, June 28-July 1 2015, Las Vegas, USA

[9] Xianfeng Yang and Liming Lian,et. al., 2014, “A New Data Mining Algorithm based on MapReduce and Hadoop,” Xinxiang University, Xinxiang Henan, P.R.CHINA