# Probability of Developing Favorable Attitude Towards Online Shopping

**Aarthi. M[1] , Karthika shree. S[2] ,Mrs.R.Reena[3], Mrs. R.K.Kapilavani[4]**
Department of Computer Science And Engineering
[1,2] Student, Prince Shri venkateshwara padmavathy college of engineering, Chennai, India
[3,4] Assistant Professor, Prince Shri venkateshwara padmavathy college of engineering, Chennai, India

***Abstract***- *Online shopping websites provide users with the products in their stock and will provide the comparison within their products alone. Thereby limiting the users to analyze before buying a product. Shoppers may get frustrated as they have to switch between multiple sites in order to match their requirements. This may sound as a hectic process. This system crabs the data from various web application and load its dataset collaboratively using crawling technique and process the batch jobs in a distributed and parallel processing way using HDFS (Hadoop Distributed File System) allowing the shoppers to analyze the features, get recommendations, to pick products and add to cart. The major part is that they can request the same for their own affordable rate. This is a greater advantage for the users who prefer economic online shopping. The cart can be reviewed at any time and can be processed. All the information will be securely and precisely stored in user session. This results in an effective data analysis, to achieve fast response, scalable and economical online shopping.*

***Keywords***- Big data, HDFS, k-means clustering, economic online shopping, Map reduce.

## I. INTRODUCTION

The era of Big Data has arrived. In the early 19th century, our capability for data generation has never been so powerful ever since the invention of the information technology. Online discussions provide a new means to identify the public interests and generate feedback in the real-time, and are mostly appealing compared to media, such as radio or TV broadcasting. An example is Flicker, a public picture sharing site, which received 1.8 million photos per day, on average, from February to March 2012. Assuming the size of each photo is 2 megabytes (MB), this requires 3.6 terabytes (TB) storage every single day. Indeed, as an old saying states: “a picture is worth a thousand words,” the billions of pictures on Flicker are a treasure tank for us to explore the human society, social events, public affairs, disasters, and so on, only if we have the power to harness the enormous amount of data. The above examples demonstrate the growth of Big Data applications where data collection has grown tremendously and it is beyond the ability of commonly used software tools to capture, manage, and process within a “tolerable elapsed time.” The most fundamental challenge for Big Data applications is to explore the large volumes of data and extract useful information or knowledge for future actions. Although researchers have confirmed that interesting patterns can be discovered from the SKA data, existing methods can only work in an offline fashion and are incapable of handling this Big Data scenario in real time. As a result, the unprecedented data volumes require an effective data analysis and prediction platform to achieve fast response and real-time classification for such Big Data. Big data is a term that describes large volume of data, both structured and unstructured that inundates a business on a day-to-day basis. Big data can be analyzed for insights that lead to better decisions and strategic business moves. Big data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision-making.

Big data concerns the massive volumes of dataset usually huge, sparse, incomplete, uncertain, complex or dynamic, which are mainly coming from multiple and autonomous sources. The most fundamental challenge for big data application is to explore large volumes of data and extract useful information or knowledge for future actions. In day to day life, we will need to buy lots of goods or products from a shop. It may be food items, electronic items, house hold items etc. Now-a-days, it is really hard to get some time to go out and get them by ourselves due to busy life style or lot of works. In order to solve this, e-commerce websites have been started. Using these websites, we can buy goods or products online just by visiting the website and ordering the item online by making payments online. The current online shopping system provides the user to purchase the goods in online, user can choose different products based on categories, online payments, delivery services are done. It provides only limited number of products available in the online shopping website. Thereby limiting the users to analyze before buying a product and the shoppers may be unsatisfied with the limited number of choices in available product as per their rate requirement. Thus the shoppers has to switch between multiple sites in order to match the requirement and this results a hectic process, shoppers may get frustrated due to the lot of time utility. So it is necessary to propose new system which helps in building a website where massive amount of products are available in a single service provider.

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

It allows the shoppers to analyze the configurations and features of the products, get recommendations, can pick products and add to cart irrespective of the service provider and they can request their affordable reasonable rate to the seller. Hence this covers the disadvantages of the system and makes buying easier and helps the vendor to reach wider market. The cart can be reviewed by the user at any time and can be processed. All the information will be securely and precisely stored in user's session. This results in an effective data analysis, to achieve fast response, scalable and an efficient precise service comparison.

## II. RELATED WORK

In recent years online shopping system becomes a trend in which data mining and data warehousing plays a vital role in processing and storing the data. As the days passed the shoppers who buy the products online increases such that product in the website also increases. The techniques and algorithms used in existing online shopping system are clustering algorithm, classification model, prediction analysis in recommendation system. Clustering algorithm[6] groups the shoppers who have similar preferences in buying the product from e-commerce website. The users who buy the products are given an unique identifier and then clustered as a group.

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

Predictive analysis[12] is applied in product recommendation, predictive search. This prediction forms patterns to predict the product to be chosen by the user from the given input data, past and historical data. It also uses machine learning algorithm to analyze the data and make predictions. The relationship between the customer and the e-commerce websites is managed using association rules. This association rules[13] determines the buying habits of customers and set offers to them in e-commerce websites. It also identifies the frequent item sets that are occurring simultaneously in the database. With the fast development of networking, data storage, and the data collection capacity, Big data[14] are now rapidly expanding in all the fields for data processing. Big data processing mainly depends on parallel programming models like map reduce, as well as providing a cloud computing platform of big data services for the public. Map reduce is a batch-oriented parallel computing model. There is still a certain gap in performance with relational databases. Improving the performance of map reduce and enhancing the real-time nature of large-scale data processing have received a significant amount of attention, with map reduce parallel programming being applied to many machine learning and data mining algorithms. Data mining algorithms[15] usually need to scan through the training data for obtaining the statistics to solve or optimize model parameters. It calls for intensive computing to access the large-scale data frequently. To improve the efficiency of algorithms, visual et al proposed a general-purpose parallel programming method, which is applicable to a large number of machine learning algorithms based on the simple map reduce programming model on multi core processors.

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

Data node[16] is a type of slave node which is used to save data and task tracker in data node which is used to track on the ongoing jobs which are coming from name node. A small-size hadoop cluster includes one master node and multiple slave nodes. A slave node acts as a data node in HDFS architecture and acts as task tracker in map reduce framework. These types of clusters are used in only non-standard applications. Again large-size cluster [17] includes a dedicated name node which manages all file system index or metadata, a secondary name node which periodically generates snapshots of name node to prevent or recover from name node failure and reducing loss of data. Similarly in map reduce framework a standalone job tracker server which manages job scheduling and several task trackers are running on data nodes or slaves. Skew reduce improves the execution speed and runtime is improved [18] even if the cost functions provided by user were not perfectly accurate but that are good enough to identify expensive results in the data. The existing system allows the user to search the products by specifying the price range, size, color etc. One more technique is emerging where the user can view the products by discounted rates. Here, the user will have to pay the amount displayed for that particular product as such.

## III. FAVORABLE ONLINE SHOPPING SYSTEM USING MAP REDUCE FRAMEWORK

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

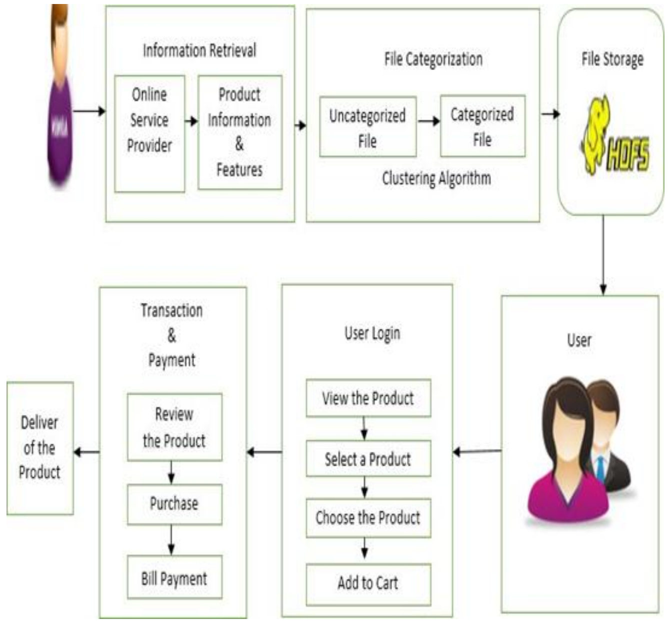

Fig1. Architecture of existing online shopping system using map reduce

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

### A. Big data environment

Sample Web Applications built so that the users can compare their products with different Service Providers. The Applications uses sample datasets that has been crawled in Amazon previously. Similar The datasets were prepared for other Applications too using the Meta model that has been crawled earlier. Each Data Set was loaded independently in Various Web Applications. Features and other specifications have been loaded differently for each Application based on the Service Providers Requirement. These Applications have been deployed in Web Servers so that the Application is Up and Running. Web Services have been written on each Web Application so that any third party can communicate with Secure Authentication.

### B. Our Gateway Application and Batch Processing over the TSV Data:

Now Our Gateway Application is built which gives users with Recommendations and Comparisons between the Products in the Market .Generally the Resources provided by Various Web Servers are in TSV (Tab Separated Values) Format and should be Batch Processed before Proceeding. For that we use our own API for TSV Manipulation. The TSV files were parsed for data. Theses data’s are used for further processing (i.e. For Recommendation and comparison).

### C. Web Crawling for Resources and Map Reduce

The Users can register and can login to view Various Products Available in Market. This is done by writing a Web Service Client Process for each Service provider. It can connect to the Various Web Applications Web Service and can pull all the needed data’s to our backend. A huge Amount of data got accumulated now .Web crawling looks for web services provided by various web applications. The Crawled Resources are then reduced by MapReduce Framework and converted into a single object .This Reduced Object Contains all the necessary information for providing comparison and

Recommendations. Map Reduce It is a combination of a map task and multiple Reduce Tasks which is used for clubbing all the slave process outcomes.

### D. Picking Products from Recommendations and Purchase:

The Recommendations were given based on the QOS, Availability, Delivery, Offers, Price and Specifications of the particular product. The Users can pick any product so that our application provides with a most Genuine Recommendation and a set of Comparisons. The Users are provided with neat and clean indexes so that he can pick a best provider for a particular product. The picked products were added in Cart and can be purchased later. The User Cart is equipped with Case Based Recommender Systems.

It uses case-based reasoning (CBR) to identify and recommend the items that seem more suitable for completing a user's buying experience provided that he or she has already selected some items. The system models complete transactions as cases and recommended items come from the evaluation of those transactions. Because the cases aren't restricted to the user who purchased them, the developed system can generate accurate item recommendations for joint item selections, both for new and existing users. Having analyzed the previous transactions and identified the concepts within which concrete items appear, the given part of a new transaction is matched over the existing ones to find the more adequate solution. i.e. the best way to fill this basket. When the User initiates Transaction our Gateway will connect to the Banking Web Services directly on behalf of the Service Provider and Completes the transaction securely with help of OTP sent to their mail id given on User Registration .A Bank Account is needed for Complete the Transaction which can be created earlier through our Banking Application .The Process will be back to our Application as soon as the Transaction is over and the Purchased products will be reflected on the Bag List. i.e. Purchased Items List.

### E. Hadoop Storage

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

### F. Transaction Processing and Delivery of product

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

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

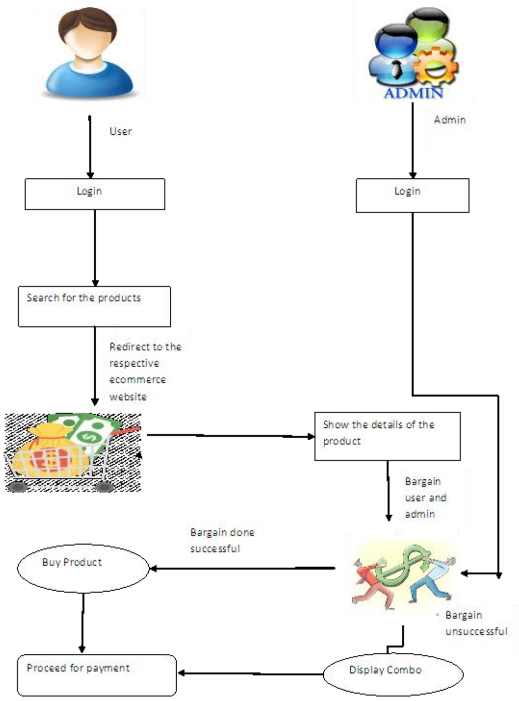

Fig.2 System architecture of proposed system

## IV. EXPERIMENTAL SETUP

The proposed system overcomes the problem of scalability in choosing the product from shopping website. Terabytes of products can be stored using HDFS and it can be accessed from anywhere at any time. It makes the product to be available up-to-date and to have a clear service comparison of products among the integrated website. It contains all dimensions of Big data. It provides the ease of gathering of the products to the shoppers. Since the dataset of product are stored in hadoop, the processing is fast and it is said to be robust. Service comparison system makes the customer, choose the products by having a clear comparison of the products in different web applications. The recommendation helps the customers to get the recommended products according to the products that are added to the cart. The customer details of this website are stored in bank server and it provides the confidentiality to each and every customer, the products that are purchased by each customer are stored as the record in customers account confidentially. This system stands unique as it does not rely on the single service provider. The cart can be reviewed at any time and can be processed. All the information will be securely and precisely stored in user session. This results in an effective data analysis, achieves fast response, scalable and an efficient precise service comparison.

User has an initial level Registration Process at the web end. The user can view the latest products at the bottom of the page. He should provide his own personal information for this process. The user can now login using his username and password. He will be displayed with products. They can buy the desired product or can bargain for the affordable rate of the product. Name, password, email id, city, contact number are the details to be provided for registration by the user. Then the user can submit and these details are stored in the database of the seller. Next time, when the user login, he can provide his username and password and then can logon. After user login process, he can enter into products page.

The user searches for the desired product among the multiple products displayed. He can choose the right product based on the rating provided by customers. All the products are displayed under respective categories. Each product is provided with Product ID, Price, Rating, Configuration, Features. If the rate is affordable to the user, he can add it to the cart by clicking on the “ADD TO CART” button.

1) Follow the payment instructions.
2) Pay for the product.
3) Buy the product

If the rate is not affordable to the user,

1) He can bargain for his affordable reasonable rate to the seller.
2) He enters the rate in the box provided by the seller side “ENTER THE BARGAIN RATE”.
3) After entering the rate, the user submits his request.

At the seller side, he receives the notification regarding the user’s request. He can either

a) Accept the rate proposal or
b) Quote the reasonable rate.

The user is responded with the seller’s response. After checking the features and configurations, the user can proceed onto the payment by clicking on the button “ADD TO CART”. For payment, the user has to A new account user can register by providing these following details

a) Account holder name
b) Email address
c) Contact number
d) City
e) Pin code
f) State

He can submit and can also make changes in the address, email address if necessary. These details are submitted to the database. When the user rate request is not accepted and when he is not ready to buy the desired product with the rate quoted by the seller. Then he can go for combo products suggested by the seller. He might like the offers and may proceed for payment.

Based on the “K MEANS CLUSTERING” algorithm, the combo products are displayed. The combo products are related to each other.

## K-MEANS CLUSTERING ALGORITHM

K-means clustering algorithm: It uses an iterative refinement for producing final result. This algorithm has following inputs: number of clusters K and the data set. The data set is a collection of features for each data point. The algorithm starts with initial estimates for the K centroids that can either be randomly generated or from the data set. The algorithm then iterates between two steps:

1. Data assignment:

Each centroid defines one of the clusters. In this step, each data point is assigned to its nearest centroid, based on the squared Euclidean distance. More formally, if ci is the collection of centroids in set C, then each data point x is assigned to a cluster based on where dist( · ) is the standard

(L2) Euclidean distance. Let the set of data point assignments for each ith cluster centroid be Si.

2. Centroid update:

In this step, the centroids are recomputed. This is done by taking the mean of all data points assigned to that centroid's cluster.

The algorithm iterates between steps one and two until a stopping criteria is met (i.e., no data points change clusters, the sum of the distances is minimized, or some maximum number of iterations is reached). It is guaranteed to converge to a result. The result may be a local optimum (i.e. not necessarily the best possible outcome), meaning that assessing more than one run of the algorithm with randomized starting centroids may give a better outcome.

**HOW TO CHOOSE K:**

This algorithm finds the clusters and data set labels for a particular pre-chosen K. In order to find the number of clusters in the data, the user needs to run the K-means clustering algorithm for a range of K values and compare the results. Generally, there is no method for determining exact value of K, but an accurate estimate can be obtained using the following techniques.

One of the metrics that is commonly used to compare results across different values of K is the mean distance between data points and their cluster centroid. Since increasing the number of clusters will always reduce the distance to data points, increasing K will always decrease this metric, to the extreme of reaching zero when K is the same as the number of data points.

Thus, this metric cannot be used as the sole target. Instead, mean distance to the centroid as a function of K is plotted and the "elbow point," where the rate of decrease sharply shifts, can be used to roughly determine K. Some other techniques exist for validating K, including cross-validation, information criteria, the information theoretic jump method, the silhouette method, and the G-means algorithm. In addition, monitoring the distribution of data points across groups provides insight into how the algorithm is splitting the data for each K.

K-means clustering is a simple unsupervised learning algorithm that is used to solve clustering problems. It follows a simple procedure of classifying a given data set into a number of clusters, defined by the letter "k," which is fixed beforehand. The clusters are then positioned as points and all observations or data points are associated with the nearest cluster, computed, adjusted and then the process starts over using the new adjustments until a desired result is reached. K-means clustering has uses in search engines, market segmentation, statistics and even astronomy. K-means clustering is a method used for clustering analysis, especially in data mining and statistics. It aims to partition a set of observations into a number of clusters (k), resulting in the partitioning of the data into Voronoi cells. It can be considered a method of finding out which group a certain object really belongs to.

It is used mainly in statistics and can be applied to almost any branch of study. For example, in marketing, it can be used to group different demographics of people into simple groups that make it easier for marketers to target. Astronomers use it to sift through huge amounts of astronomical data; since they cannot analyze each object one by one, they need a way to statistically find points of interest for observation and investigation.

We are given a data set of items, with certain features, and values for these features (like a vector). The task is to categorize those items into groups. To achieve this, we will use the K-Means algorithm; an unsupervised learning algorithm.

## V. CONCLUSION

Online shopping system produces an integrated dataset from various websites that supports the scalability problem and reduces the time consumption which processes the system in an efficient manner. It stands unique as it enables the customers to bargain their affordable reasonable rate for their desired product. This is two way beneficial to both the customers and seller. This enables greater user satisfaction in buying their desired product with the affordable rate. The sellers can able to move the stocks quickly, reach wider markets and can reactivate the inactive customers through this. The information of the customer will be securely and precisely stored in user session. In future, this system can be deployed in cloud, so that users from different places can make use of this system by accessing it through cloud storage. Dataset which covers terabytes of products can be stored and processed for shopping.

Recommendation system can be enhanced for recommending the products based on the customer's preferences.

## REFERENCES

[1] Bandar Mohammed, Malek Mouhoub, “Evaluation of an Online Shopping System under Preferences and Constraints”, IEEE conf, 2014.

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

[3] Chia-Hui Chang, Mohammed Kayed, Moheb Ramzy Girgis, Khaled F. Shaalan,“A Survey of Web Information Extraction Systems”, IEEE Trans, VOL. 18, NO. 10, OCTOBER 2000

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

[5] N. Zhang, C. Li, N. Hassan, S. Rajasekaran, and G. Das, “On skyline groups,” IEEE Trans. on Knowl. Data Eng, vol. 26, no. 4,pp.942–956,2014.

[6] Prof. Sonali P. Kadam, “Clustering Based – A FAST Algorithm on High Dimensional Data”, (IC3I), 2015.

[7] Vinodhini.M and Manju.A, “An Efficient Online Shopping System Using Map Reduce Framework in Big data “ IEEE Trans. Volume 5, Issue 5, May 2016

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

[9] Xu Zhou, Kenli Li, Zhibang Yang, and Keqin Li, “Finding the optimal Skyline product combination under price promotion campaigns”, IEEE Trans. Volume: 31, Issue: 1 , Jan. 1 2019

[10] Y.-C. Chung, I.-F. Su, and C. Lee, “Efficient computation of combinatorial skyline queries,” Information Systems, vol. 38, no. 3,pp. 369–387, 2013.

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

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

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

[14] Crescenzi and G. Mecca, “Automatic information extraction from large Websites,” J. ACM, vol. 51, no. 5, pp. 731–779.

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

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

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

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

[19]