Market Basket Analysis with Mining Association Rule(ShoppingBasketSystem)

Sayali Borse[1], Prachi Chorghe[2], Poonam Fate[3], Shivraj Pisal[4]

[1] *Student, Computer Engg, Dr.D.Y.Patil college of engg Akurdi, Maharashtra, India*
[2] *Student, Computer Engg, Dr.D.Y.Patil college of engg Akurdi, Maharashtra, India*
[3] *Student, Computer Engg, Dr.D.Y.Patil college of engg Akurdi, Maharashtra, India*
[4] *Student, Computer Engg, Dr.D.Y.Patil college of engg Akurdi, Maharashtra, India*

ABSTRACT

The proposed system uses market basket analysis method by mining association rules on the items. Market basket analysis helps to take decision for business strategies by mining association rules among items purchased together. Association Rules is data mining technique which is used for identifying the relation between item sets along with proper algorithm. The proposed system uses apriori and FIC (Frequent Itemset Counting) algorithm which is used to determine association rules which highlight general trends in the database. Apriori is easy to parallelized. Association rules are used to generate new knowledge which is require to determine frequent item sets. Proposed system achieves parallelism by using hadoop libraries and Map Reduce which is one of the popular data mining algorithms. So FIC Algorithm and Ec-Apriori Algorithm will solve our existing problems and give proper desired output.

Keyword : - *Association rules mining, Market Basket Analysis, data mining, frequent itemset, Apriori,support, confidence.*

I. INTRODUCTION

Now a days data from any source and any field is becoming very important for business and enterprises. Because that data can help them to grow their business and to gain more profit by analyzing it. Huge Data is extracted and mined to get useful information. Mined data can contain information useful for business purpose. This can also be termed as business intelligence. To handle business there is need of business intelligence and risk management skills. The analyzed records can be used for intelligent marketing. Analysis gives predictions and facts based on the given input. Analysis reduces human efforts. Earlier the analysis used to be carried out manually. Manual analysis is very tedious job and requires a lot of time. Despite being time consuming it is also error prone. Later on some tools are used for analysis but with high cost and low speed of computation. The reference number should be shown in square bracket [1]. However the authors name can be used along with the reference number in the running text. The order of reference in the running text should match with the list of references at the end of the paper.

This paper introduces a system for big shopping malls or markets which allows them to analyze their sales data by mining association rules and accordingly devise the sales purchase strategy and earn profit. The input data is a huge and complex which includes items purchased by the customers with other details about the item. The proposed system has used Hadoop to perform computation parallely. Apriori algorithm is used to determine association rules and highlight the general trends in data. Also this algorithm is easy to parallelized. Product selling strategies may include cross selling, giving discounts on a particular product, sell products in associate pair. This strategies can be predicted from the results generated by this system. Our system will provide the accurate results and supporting the result will be graphs and textual feedback. The graphical representation will help in understanding the result even more easily for the common man. There is no requirement that the person who is performing the analysis should be a technical person. Even a common man will be able to understand the results of analysis and can make decisions based on the feedback given by the system.

II. RELATED WORK

To parallelize the task hadoop libraries are use along with the mapreduce algorithm. Mapreduce is the technique used for parallel computation for large dataset. Thus mapreduce implementation is responsible for distributing and running the different task simultaneously. Performance and power of using Hadoop wordcount and mapreduce is analyzed. Wordcount is not memory bounded i.e. increasing memory speed will not impact on the performance[2].
Frequent itemset mining this concept is widely in the data mining techniques which is very popular this day's. The problem arise with infrequent itemset mining from traditional dataset and to solve this problem consider weighted itemset to discover infrequent itemset[3].
There are various phases in the life cycle of big data such as 1. Data generation: huge amount of data is generated from various sources i.e. retail market therefore it is hard system to handle them. 2. Data storage: This phase refers to storing and managing large-scale dataset. Data management means set of software deployed to manage and query large data set. And provide interfaces to interact with and analyze stored data. 3. Data processing: it refers to the process of data collection, data transmission, pre-processing and extracting useful information. Another important issue is privacy data. Access restricted for data provided to system to retain the protection[4].
Association rule mining is used to find out the relationship between the items. The paper[5] summarizes, analyses and compares the various association algorithms such as Classical association rule mining algorithm, Data set partitioning algorithm, Depth first search , Breadth first search. Association rule mining is discussed using several aspects such as width, partition, depth, sampling and incremental updating [5]. The strategy used in Apriori algorithm is to separate tasks of association rule mining into two steps: 1) The generation of frequent itemsets: frequent itemsets are generated by iteration that satisfy minimum support threshold.
2) Generating association rules: extract high confidence rules from the frequent itemsets found, which are the strong association rules. In the first step for mining frequent itemsets, the apriori algorithm will produce a large set of candidate items, and in order to generate frequent itemsets, scanning the databases require to loop through pattern matching to inspect candidates, computation time of this step will be much larger than the second step, first step is the core of the algorithm[5].
In the retail sector it is hard to find frequent itemset due to large data so solution to this is given in the paper[6]. The apriori algorithm is used to find frequent itemset and the patterns from the large databases to make better business decisions. It resolved the problem of decision making. The algorithm firstly find out the support of all itemset and generate association rules by combining large itemset or frequent set[6].

III. PROPOSED SYSTEM

The proposed system is being implemented using Hadoop libraries. The data that the system is taking as a input is in the form of excel sheets. It is taken from big shopping malls or markets using the google drives. Input data is very much complex as it contains different fields like different transactions, product names, their prices, location of purchase, latitude, longitude etc. Then input will be divided into different chunks according to map reduce functions. The proposed system uses apriori algorithm which is used to determine association rules which highlight general trends in the database. Then it will analyze and make result parallely. The following figure shows the system architecture of proposed system.

Components of system

• Data creation: Data is taken in the form of .csv file so that data can be created and edited. Data can be imported from Google drive in text (.csv or .txt) file.
• Data initialization: It checks the file format of input records to be inserted. It requires the proper row and column format which includes different fields.
• Temporary data processing: The data is being processed is stored temporarily and creates temporary files.
• Middleware: It consists of Map Reduce functions that performs using Hadoop libraries. Whole large data is divided into multiple chunks and data it will be processed parallely. Apriori algorithm is applied on it at the same time.
• Result: Result is generated in the texual and graphical format that contains overall analysis of items, items to be sell in associate pair, periodic analysis of products, recommendation that will increase profit in the business. Ex.Cross Selling of the Items, Proper Placement of Items, giving some percent off on the price of product.

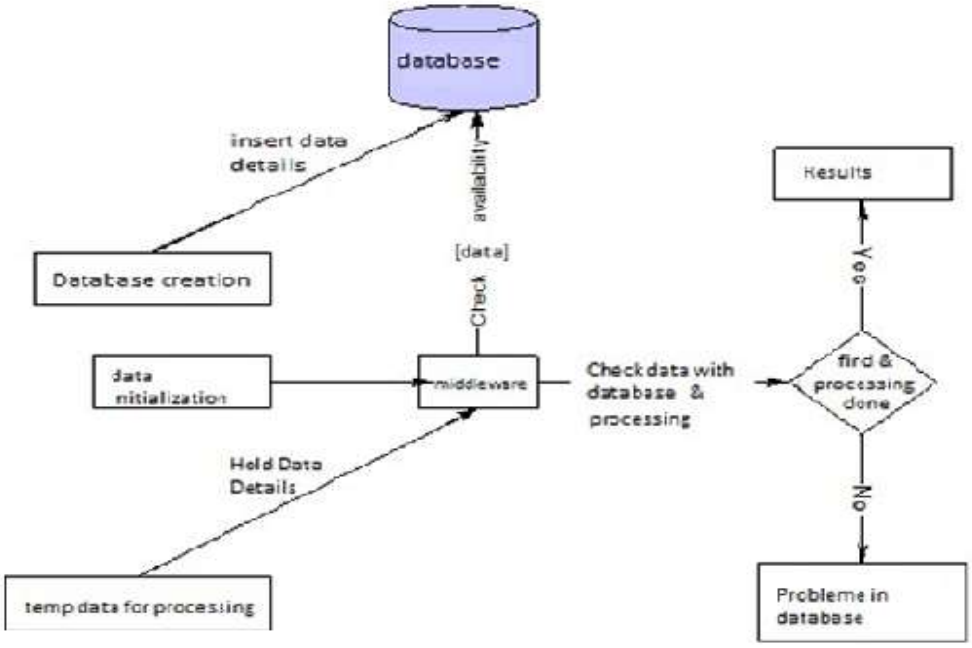

Fig -1 : System Architecture

IV. ALGORITHM

A. Map Reduce algorithm

It is a programming model used for processing huge structured or unstructured data sets. It runs in the background of Hadoop to provide scalability and easy data processing solutions. MapReduce provides analytical capabilities for analyzing huge volumes of complex data. It splits the huge data into multiple independent chunks which are processed by map tasks in a completely parallel manner. It uses mainly three functions. Map function-It takes input key and generate a set of intermediate key-value pair. Reduce function- The reduce function is used to merge all the intermediate values associated with the same intermediate key. Before the reduce function, sorting is performed on the key value pair. After sorting the merging process begins where the sorting is merged into a list. After this merging process the actual reduce function starts. Map tasks may vanish at different times, so the reduce task starts copying their outputs as soon as each completes. This is known as the copy/shuffle phase of the reduce task. Driver function: It communicates with the Hadoop framework and species the configuration elements required to run a MapReduce job. This involves aspects such as telling Hadoop which Mapper and Reducer classes to use, where to send the input data and in what format, and where to place the output data and how to format it.

The MapReduce algorithm consist of two important tasks, namely Map and Reduce. The Map task takes a set of data and converts it into another set of data, where each elements are broken down into tuples like key-value pairs. The Reduce task takes the output from the Map as an input and combines those data tuples key-value pairs into a small set of tuples. The reduce task is always performed after the map job. As shown in the illustration, the MapReduce algorithm performs the following actions

- Tokenize Tokenizes the tweets into maps of tokens and writes them as key-value pairs.
- Filter Filters unwanted words from the maps of tokens and writes the filtered maps as key-value pairs.
- Count Generates a token counter per word.
- Aggregate Counters Prepares an aggregate of similar counter values into small manageable units.

B. Apriori Algorithm

It is a Data mining technique and one of the influential algorithm. Massive amounts of data continuously being collected and stored as transactions. Those collected data can be useful from the business prospective and used to find frequent itemsets for boolean association rules. Frequent item set mining leads to the discovery of associations and correlations among items. Apriori uses a bottom up approach, in which there is a step known as candidate generation. In candidate generation, frequent subsets are extended one item at a time. Database containing trasactions are operated by apriori algorithm. For example, collections of items bought by customers. Apriori gives the affinity analysis and association rule learning, which encompasses a large set of analytic techniques focused at uncovering the associations and connections between specific objects. For example, maybe people who buy bread and casting butter, also tend to buy milk. A retailer can use this information to inform marketing, Store layout (arranging products which are frequently brought together to enhance business or improve profit), Drive recommendation engines (like Amazons, flipkart's recommendation of another product.) In the proposed system, the shopping mall data will be given to apriori algorithm along with map reduce technique. Apriori helps to find the co-relation and association between the different data items. The algorithm will take all the data items and find relation between them. The analysis result will be displayed as list of more likely brought items together, which will help to decide market strategies, in store layout. The list will be in the descending order, the most likely brought item to least brought item.

Advantages of using Apriori: The algorithm makes use of large itemset property. The method can be easily parallelized. The algorithm is easy from implementation point of view.

Support

Count of that itemset in the total number of trasactions. i.e. The support supp(X) of an attribute X is defined as the proportion of transactions in the data set which contain the attribute.

supp(X)= no. of attributes which contain the comments X / total no. of transactions.

Confidence

conf(X->Y)= supp(X U Y)/supp(X).

Frequent itemsets

The sets of items that have minimum support. All the subsets of a frequent itemset must be frequent for e.g. PQ is a frequent itemset P and Q must also be frequent.

Steps of Algorithm

Step 1: Let k=1

Step 2: Generate frequent itemsets of length 1.

Step 3:Repeat until no new frequent itemsets are identified.

- Generate length (k+1) candidate itemset of length k that are infrequent
- Prune candidate itemsets containing subsets of length k that are infrequent
- Count the support of each candidate by scanning the database
- Eliminate candidates that are infrequent, leaving only those that are frequent.

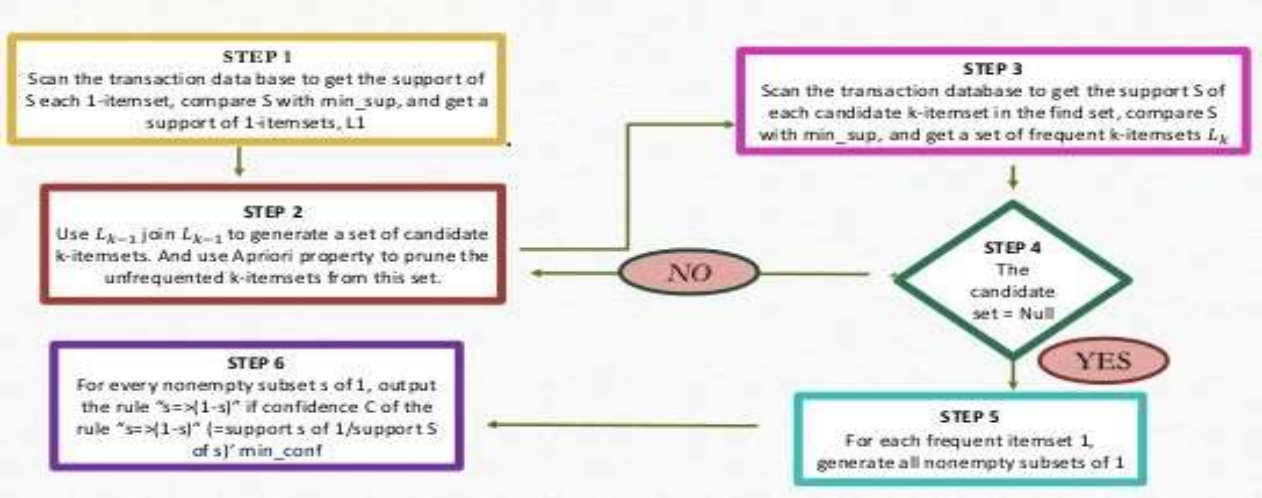

Fig -2 : Steps of Apriori Algorithm[7]

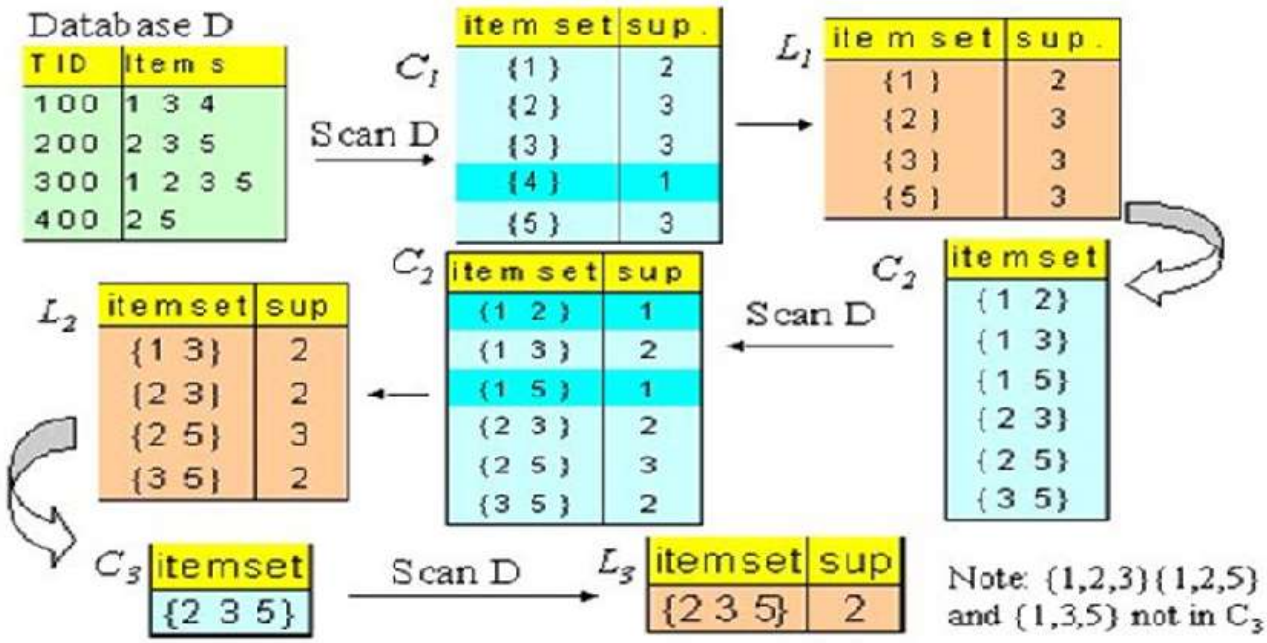

Fig -3 : Apriori Algorithm Example[8]

Fig.3 shows example of Apriori consist of one database D which contains number of transactions denoted by TID. Each transaction has some items purchased denoted in numbers. Whole database is scanned. Itemset and support is generated iteratively with eliminating item which has minimum support. At last most frequent associative itemset is generated.

C. FIC algorithm (Frequent Itemset Counting)

This algorithm is basically an ECLAT method. ECLAT is Equivalence Class Clustering and bottom up Lattice Traversal. Its a method for frequent itemset generation. It searches in DFS manner. FIC Algorithm is a distributed version of Eclat. It distributes the search space more evenly among different processing units. Earlier methods such as Partition algorithm, which oftenly worked on the methods of distributing the database or workload into n numbers of equally sized sub databases. These sub database is also called as Shards. Then these shards are mined separately to find the frequent item and then these results which are locally frequent are combined together and processed again to prune the globally infrequent items. This method is an expensive one. Apart from this it has large communication cost too, as the number of sets that is to be mined can be very large, hence the set that will be recounted will be also very large. Such sort of algorithms are generally avoided in Hadoop. But the proposed method will be free from such an anomalies as it will divide the search space rather than the data space. Hence this method will reduce the communication cost as no extra communication is required between the mappers and no checking for overlapping mining results has to be done. Concluding that for mining large dataset, memory-wise it is the best fit for FIC Algorithm i.e. Eclat using different sets. It divides the operation of this method in 3 parts and in order to maximally benefit the cluster environment. It distributes all the three parts among multiple mappers. And for this it utilizes the vertical database format.

Steps are as followed:

Finding the local Frequent Items: In this step, the vertical database is partitioned into sub databases and then divided to the separate mappers available. The mapper then removes each of the limited items from the sub databases. Then in the reduced phase, the frequent items are collected. k-FIs Generate: In this step, the set of limited itemsets of k-size are obtained. Firstly, the frequent items are divided amongst the m mappers. Each mapper separately combines the items. Then the reducer assigns these items to another batch of m mappers by using round robin method. Mining the Subtree: In this step the prefix tree is mined. Here the prefix tree is used as the big data. The prefix tree is a structure which describes an itemset, the exact path from the root to the node. The tree is divided into separate groups. Each group is then mined individually on different mappers.

D. Comparison between FIC and Apriori

Here in this, project is working on for mining Big Data as to extract information about frequently occurring itemsets. Although a number of methods are already existing but with some flaws. So the proposed system hereby present two methods namely FIC Algorithm and Ec-Apriori Algorithm which will solve our existing problems and give proper desired output. If the processing speed is important and the amount of data is less then system uses first method called FIC Algorithm and if our data is too large in size and processing is more important than the time taken then for such cases system works on second method called as Ec-Apriori Algorithm. So, system proposes these two algorithms to eradicate the current flaws in the existing method.

V. EXPERIMENTAL RESULTS

The proposed system consist of two algorithms for analysing most frequent associated itemsets. They are FIC (Frequent Itemset Counting) and Apriori algorithm. Following graphs shown in fig.4 and fig.5 shows the difference of efficiencies between FIC and Apriori with number of mappers and minimum support. These graphs shows that FIC algorithm is more efficient than Apriori algorithm as it requires less time for execution.
The Apriori algorithm generates number of iterations and each iteration generates a graph. First and second iteration graphs are shown in fig.6 and fig.7. First iteration consist of each items denoted in alphabets with its support. Second iteration consist of associated pair of items with their support. In such a way system generates each iteration values with their graph which shows most frequent and associated itemsets.

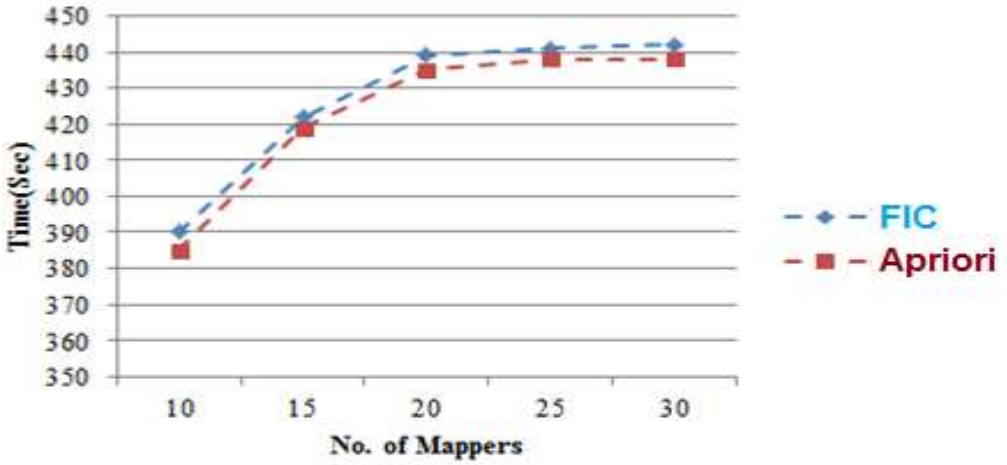

Fig-4: Timing comparison between FIC and Apriori

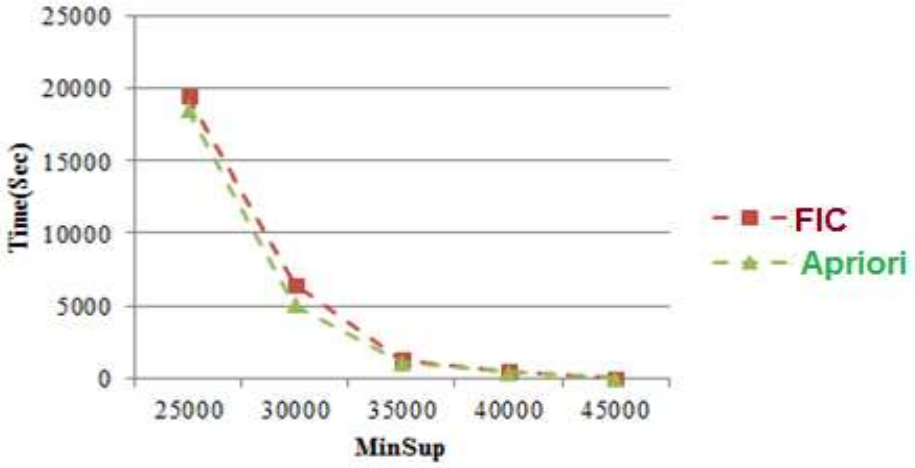

Fig-5: Timing comparision between FIC and Apriori.

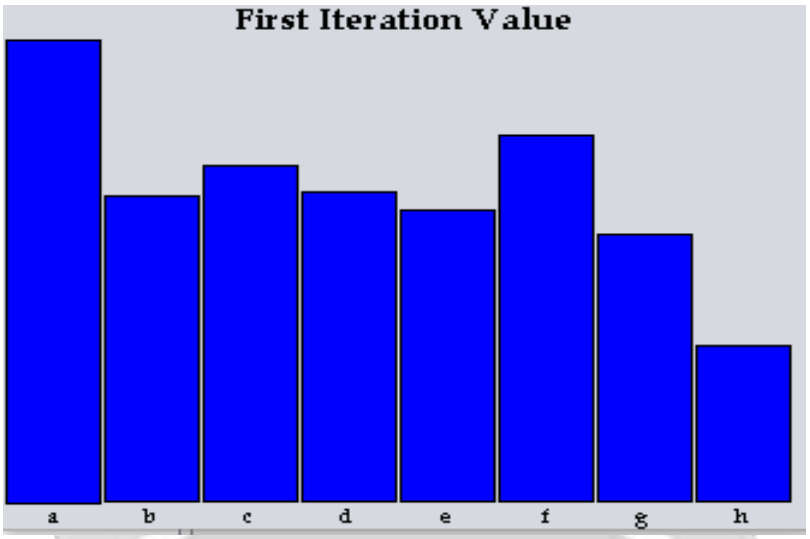

Fig. 6: First Iteration Graph

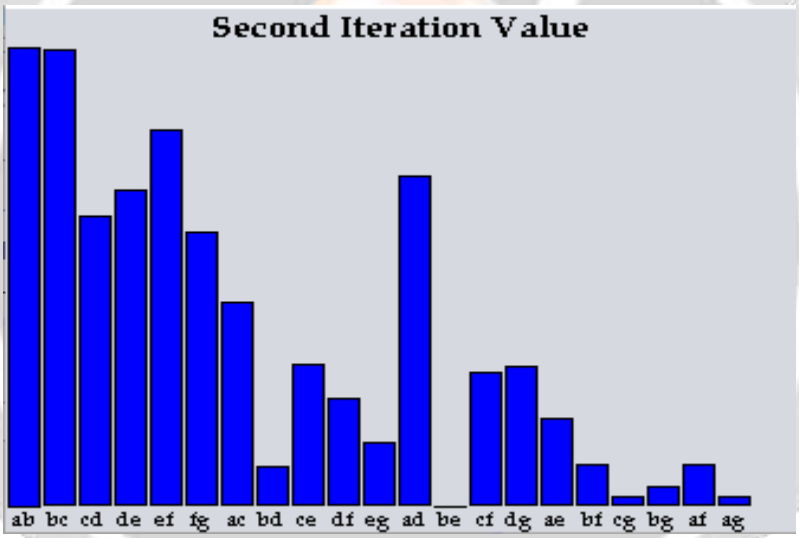

Fig. 7 : Second Iteration Graph

VI. RESULT AND FUTURE SCOPE

The successful implementation of the proposed system gives the analysis of whole shopping mall products. It will analyze the most frequent and associative items that are sold in a particular duration or in a particular region. It will recommend a better solution to gain more profit like particular item pair selling with giving an offer. The result will be generated in the form of graphs and text file. Graphs are stored in the file format by taking screenshots. It is very important from shopkeeper and retailers point of view. It will also display difference in time efficiency of FIC and Apriori algorithm. These results may use in deciding business strategies to gain more profit in their business in future. They can also use to avoid business risks. A retailer can use this information to inform marketing, Store layout (arranging products which are frequently brought together to enhance business or improve profit), Drive recommendation engines (like Amazons, flipkart's recommendation of another product.), cross selling of product(selling less sold product with higher sold product in a pair), associate pair selling like bread and butter, giving some discounts on a product etc.

VII. CONCLUSION

Data Analysis is very much important to avoid business risks and to gain more profit. But now a days it is becoming more time consuming, less accurate, need man power. Thus proposed system provide the effective tool to analyze the data using association rules provided by apriori along with Map Reduce function i.e. Hadoop. Result will generate frequent associative itemset, graphical representation, textual result stored in the file format which will used to decide different business strategies like selling associative product pair, cross selling, giving discounts etc. Generated result will be more accurate, easy to understand and requires less computation time.

REFERENCES

[1] Tsan-Ming,Hing kai chan and Xiaohang Yue,Recent develpoment in Big data analytics for Business operation and Risk management,IEEE transaction paper,2016
[2] JOSEPHA.ISSA,"PerformanceEvaluationandEstimationModelUsing Regression Method for Hadoop WordCount",IEEE Journal,2015
[3] Luca Cagliero and Paolo Garza"Infrequent Weighted Itemset Mining Using Frequent Pattern Growth",IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING,VOL. 26, NO. 4, APRIL 2014
[4] ABID MEHMOOD, IYNKARAN NATGUNANATHAN, YONG XIANG , (Senior Member, IEEE),GUANG HUA, (Member, IEEE), AND SONG GUO, (Senior Member, IEEE)"Protection of Big Data Privacy",IEEE Journal, date of publication April 27, 2016.
[5] Ruowu Zhong, Huiping Wang"Research of Commonly Used Association Rules Mining Algorithm in Data Mining.",International Conference on Internet Computing and Information Services,2011.
[6] lugendra Dongre,Gend Lal Prajapati,s. V. Tokekar "The Role of Apriori Algorithm for Finding the Association Rules in Data Mining.",IEEE 2014.
[7] https://www.slideshare.net/INSOFE/apriori-algorithm-36054672
[8] https://webdocs.cs.ualberta.ca/zaiane/courses/cmput499/slides/Lect10/sld054.html
[9] Ke Deng, Xin Li, Jiaheng Lu, and Xiaofang Zhou, Best Keyword Cover Search,2013
[10] VagelisHristidis,YaoWu,andLouiqaRaschid,EcientRankingonEntity Graphs with Personalized Relationships .