Spend Analysis in Super Market Data

[#1] A. Ishwaria , [#2] T. Nandhini, [#3]. S. Suseela
IV IT Students, Department of Computer Science and Engineering , Periyar Maniammai University, Vallam , Thanjavur, India.
Assistant Professor, Department of Computer Science and Engineering, Periyar Maniammai University, Vallam , Thanjavur, India.

***Abstract* - Super Market analysis is an important component of analytical system in retail organizations to determine the placement of goods, designing sales promotions for different segments of customers to improve customer satisfaction and hence the profit of the Supermarket. These issues for a leading supermarket are addressed here using frequent itemset mining. The frequent itemsets are mined from the Super Market database using sqoop and hadoop. A retailer must know the needs of customer and adapt to them. Super Market analysis is one possible way to find out which items can be put together . Super Market analysis gives retailer good information about related sales on group of goods basics Customers who buys bread often also buy several products related to bread like milk, butter or jam. It makes sense that these groups of goods also must be located side-by-side in order to remind customers of related items and to lead them through the center in a logical manner.Super Market analysis is one of the data mining methods focusing on discovering purchasing patterns by extracting associations or co-occurrences from a store's transactional data. Super Market analysis determines the products which are bought together and to reorganize the supermarket layout, and also to design promotional campaigns such that product's purchase can be improved. Hence, the Market consumer behaviors need to analyzed, which can be done through different data mining techniques.**

Keywords: Big data, NoSql, Super Market, Map reduce.

I. INTRODUCTION

One of the challenges for companies that have invested heavily in customer data collection is how to extract important information from their vast customer databases and product feature databases, in order to gain competitive advantage. Several aspects of market basket analysis have been studied in academic literature, such as using customer interest profile and interests on particular products for oneto- one marketing , purchasing patterns in a multi-store environment been intensively used in many companies as a means to discover product associations and base a retailer's promotion strategy on them. Informed decision can be made easily about product placement, pricing, promotion, profitability and also finds out, if there are any successful products that have no significant related elements. Similar products can be found so those can be placed near each other or it can be cross-sold.

A retailer must know the needs of customers and adapt to them. Market basket analysis is one possible way to find out which items can be put together. Market basket analyses gives retailer good information about related sales on group of goods basis Customers who buy s bread often also buy several products related to bread like milk, butter or jam. It makes sense that these groups are placed side by side in a retail center so that customers can access them quickly. Such related groups of goods also must be located side-by-side in order to remind customers of related items and to lead them through the center in a logical manner. Market basket analysis is one of the data mining methods focusing on discovering purchasing patterns by extracting associations or co-occurrences from a store's transactional data. Market basket analysis determines the products which are bought together and to reorganize the supermarket layout, and also to design promotional campaigns such that products' purchase can be improved. Hence, the Market consumer behaviors need to be analyzed, which can be done through different data mining techniques.

II. BIG DATA

Big data can be characterized by the volume, velocity, and variety of data that is generated [6].

A. Volume Data volume is the amount of data available to an organization or a firm, as long as it can access it, it is not necessarily owned by the organization. The value of data records will decrease in proportion to age, type, richness, as data volume increases. Since 2012, about 2.5 Exabyte of data are created each day, and that number is doubling every 40 months or so [11].

B. Velocity Velocity is the speed of creation, streaming, and aggregation of data. Data velocity can also refer to the rate at which data may enter the organization. The velocity dimension shifts the data into a continuous flow of information rather than discrete packages of data [12].

C. Variety Data variety is a measure of the richness of the data representation, it takes the form of messages, updates, and images posted to social networks; readings from sensors; GPS signals from cell phones, etc. [11]. From an analytic perspective, it is the biggest challenge to effectively using large volumes and incompatible data formats.

III LITERATURE SUREVEY

To enhance the processing of Supermarket, we have a proposed a series of Big Data in SuperMarket by using Hadoop. There are many techniques proposed in order to efficiently process large volume of medical record which has explained below:

1]. Fan W, Bifet A (2013). Mining big data: current status, and forecast to the future. ACM SIGKDD Explorations Newsletter, 14(2), pp. 1–5.This paper proposed a frame work which focus on improving the performance of MapReduce workloads and maintain the system. DHSA will focuses on the maximum utilization of slots by allocating map (or reduce) slots to map and reduce tasks dynamically.

2" Chen H, Chiang R, Storey V (2012). Business Intelligence and analytics: From big data to Big Impact.

MIS Quarterly, 36(4), pp. 1165–1188.In this paper author proposed the potential and promise of big data analytics in healthcare. The paper provides a broad overview of big data analytics for healthcare researchers and practitioners. Big data analytics in healthcare is evolving into a promising field for providing insight from very large data sets and improving outcomes while reducing costs. Its potential is great; however there remain challenges to overcome.

IV. PROPOSED SYSTEM

Proposed concept deals with providing database by using Hadoop tool we can analyze no limitation of data and simple add number of machines to the cluster and we get results with less time, high throughput and maintenance cost is very less and we are using joins, partiations and bucketing techniques in Hadoop

Advantages:

- ✓ No data loss problem
- ✓ Efficient data processing

V. TECHNOLOGIES FOR BIG DATA ANALYTICS

1) Extraction, transformation and loading (ETL) ETL tools are a technology that allows to extract raw data from a variety of sources. ETL transform data according to a structure and load it into a data warehouse, by transforming data into a structured format, which can help to obtain meaningful informations [17].

2) Data warehouse Data warehouse is a database used for storing data which are specifically structured for analysis [18]. Data warehouses are used for managing the storage, retrieval and analysis of structured big data [18]. Data warehouses use ETL to transform big data to a structured format for storage and retrieval [16].

3) Distributed systems A distributed system is a set of multiple computers connected together are used to solve a computational problem [16]. Distributed systems are commonly used to analyze big datasets [19].

4) Hadoop Hadoop is an Apache open source software framework for writing applications which process large datasets in parallel on a distributed system [19]. Hadoop allows loading, storing and querying big datasets on multiple platforms using parallel analytics [3].

VI. METHODOLOGIES

MODULES:

- ➢ Data Prepossessing Module
- ➢ Data Ingestion Module With Sqoop
- ➢ Data Analytic Module With Hive
- ➢ Data Analytic Module With Pig
- ➢ Data Analytic Module With MapReduce
- ➢ Data Analytic Module With R

Data Preprocessing Module

In this module we have to create Data set for health it contain set of table such that Patient details, disease details, doctor details, billing details and payment details for last four years .and this data first provide in MySQL database with help of this dataset we analysis this project.

Data Migration Module with Sqoop

In this module we have to transfer the dataset into Hadoop(HDFS), that will be happen in this module. Sqoop is a command-line interface application for transferring data between relational databases and Hadoop.

In this module we fetch the dataset into Hadoop (HDFS) using Sqoop Tool. Using Sqoop we have to perform lot of the function, such that if we want to fetch the particular column or if we want to fetch the dataset with specific condition that will be support by Sqoop Tool and data will be stored in Hadoop (HDFS).

Data Analytic Module with Hive

Hive is a data ware house system for Hadoop. It runs SQL like queries called HQL (Hive query language) which gets internally converted to map reduce jobs. Hive was developed by Facebook. Hive supports Data definition Language (DDL), Data Manipulation Language (DML) and user defined functions.

In this module we have to analysis the dataset using HIVE tool which will be stored in Hadoop (HDFS).For analysis dataset HIVE using HQL Language. Using hive we perform Tables creations, joins, Partition, Bucketing concept. Hive analysis the only Structure Language.

Data Analytic Module with Pig

Apache Pig is a high level data flow platform for execution Map Reduce programs of Hadoop. The language for Pig is pig Latin. Pig handles both structure and unstructured language. It is also top of the map reduce process running background.

In this module also used for analyzing the Data set through Pig using Latin Script data flow language.in this also we are doing all operators, functions and joins applying on the data see the result.

Data Analytic Module with MapReduce

MapReduce is a processing technique and a program model for distributed computing based on java. The MapReduce algorithm contains two important tasks, namely Map and Reduce. In this module also used for analyzing the data set using MAP REDUCE. Map Reduce Run by Java Program.

VII. SYSTEM TECHNIQUES

MapReduce is a processing technique and a program model for distributed computing based on java. The MapReduce algorithm contains two important tasks, namely Map and Reduce. Map takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs). Secondly, reduce task, which takes the output from a map as an input and combines those data tuples into a smaller set of tuples. As the sequence of the name MapReduce implies, the reduce task is always performed after the map job.

The major advantage of MapReduce is that it is easy to scale data processing over multiple computing nodes. Under the MapReduce model, the data processing primitives are called mappers and reducers. Decomposing a data processing application into mappers and reducers is

sometimes nontrivial. But, once we write an application in the MapReduce form, scaling the application to run over hundreds, thousands, or even tens of thousands of machines in a cluster is merely a configuration change. This simple scalability is what has attracted many programmers to use the MapReduce model.

The MapReduce Algorithm

- Generally MapReduce paradigm is based on sending the computer to where the data resides.
- MapReduce program executes in three stages, namely map stage, shuffle stage, and reduce stage.

Map stage : The map or mapper's job is to process the input data. Generally the input data is in the form of file or directory and is stored in the Hadoop file system (HDFS). The input file is passed to the mapper function line by line. The mapper processes the data and creates several small chunks of data.

Reduce stage: This stage is the combination of the Shuffle stage and the Reduce stage. The Reducer's job is to process the data that comes from the mapper. After processing, it produces a new set of output, which will be stored in the HDFS.

VIII.SYSTEM ARCHITECTURE:

The proposed method performs well in the general population as well as in sub-populations. Results indicate that the proposed model significantly improves predictions over established baseline methods analyzing electricity consumption. The goal of this study was to analyze how much of units consumed in last four years and how much amount they paid previous four year as the forecast for the following year.

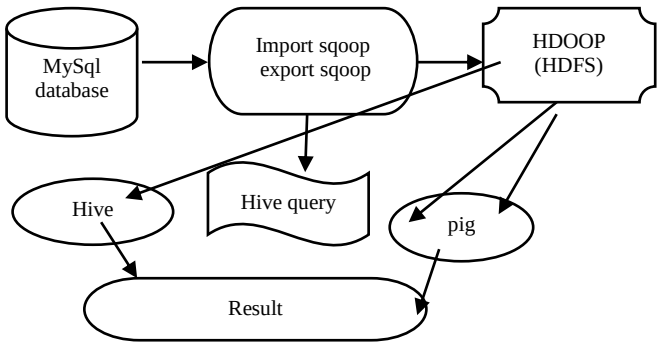

Fig:1.1

IX. RESULT AND DISCUSSION

From the household section of the Anantha store, sample market basket dataset is taken using the invoice copies or copy bills of the supermarket. 9620X302 sample Binary dataset is manipulated with Hadoop and Sqoop tool and the results are shown above. Hadoop and Sqoop tool are compared based on the frequent itemsets and association rules generated. Hadoop and Sqoop tool provides output only for very low support values. Very low support values are meaningless because it shows nothing about the customers' behavior.

X. CONCLUSION

This paper show how can big data analytics help to improve the retail business and can be applied in the sector and help improve marge. However, there are some barriers to using big data analytics such as the privacy of information and scalability of analytic algorithms. In order to help analyze big data, retailers can use analytic techniques and technologies to help analyze big data in order to help with supporting decision making.

XI. ACKNOWLEDGMENT

It is our great pleasure to thank our assistant professor Ms S. Suseela to encouraging us to do this paper presentation Effectively. We would also like to thank our dear parents for their support and the encouragement.

REFERENCES:

[1] Fan W, Bifet A (2013). Mining big data: current status, and forecast to the future. ACM SIGKDD Explorations Newsletter, 14(2), pp. 1–5.
[2] Davenport T (2013). Analytics 3.0. Harvard Busin.Rev.91(12), pp. 64– 72.
[3] Davenport T, Barth P, Bean R (2012). How "big data"is different. MIT Sloan Management Review, 54(1), pp. 22–24.I.
[4] McAfee A, Brynjolfsson, E. (2012). Big data: the management revolution. Harvard Business Review, 90(10), pp. 1–9.
[5] Chen H, Chiang R, Storey V (2012). Business Intelligence and analytics: From big data to Big Impact. MIS Quarterly, 36(4), pp. 1165–1188.
[6] Russom P (2011). Big data analytics. TDWI Best Practices Report, Fourth Quarter, pp. 1–35.
[7] Davenport T, Dyche J (2013). Big data in Big Companies (White paper). May 2013, pp. 1–31.
[8] Singh S, Singh N (2012). Big data analytics. In 2012 International Conference on Communication, Information & Computing Technology (ICCICT), pp. 1–4