International Journal of Research in Advent Technology, Vol.2, No.9, September 2014
E-ISSN: 2321-9637

Big Data: Mining the Web Data and Its Causes on Retail Marketing

S.Ramanjaneyulu[1],K.Vijay Bhaskar[2],K.Naresh Babu[3]
CSE,Asst.professor,Geethanjali College of engineering & Technology[1]
CSE,Asst.professor,Geethanjali College of engineering & Technology[2]
IT,Asst.professor,Geethanjali College of engineering & Technology[3]
ramanji.csit@gmail.com[1], vijaybhaskarchamp@gmail.com[2], naresh.kosuri@gmail.com[3]

Abstract: The main theme of e-marketing: attracting the consumer in the highly competitive internet market place. Most of the consumers are not satisfied with e-marketing due to lack of virtual interaction. Identifying the web experience components and understanding their tests and preferences as inputs in the online decision making process are the first step in developing and delivering an attractive online presence likely to have the maximum impact on the internet users. Here we introduce one of the most popular data mining approaches using HDFS for item sets storage is to find frequent item sets from a transaction dataset we proposed MapReduce. Finding frequent item sets the apriori algorithm is used in MapReduce programming. Once frequent item sets are obtained, it is straightforward to generate association rules with confidence larger than or equal to a user specified minimum confidence.

Keywords: virtual interaction, HDFS, MapReduce

1. INTRODUCTION

Retail industry collects large amount of data on sales and customer shopping history. The quantity of data collected continues to expand rapidly, especially due to the increasing ease, availability and popularity of the business conducted on web, or e-commerce. Retail industry provides a rich source for data mining. Retail data mining can help identify customer behavior, discover customer shopping patterns and trends, improve the quality of customer service, achieve better customer retention and satisfaction, enhance goods consumption ratios design more effective goods transportation and distribution policies and reduce the cost of business.
Some of the retail applications of data mining are in following areas:

Customer Segmentation: Customer segmentation is a vital ingredient in a retail organization's marketing recipe. It can offer insights into how different segments respond to shifts in demographics, fashions and trends. For example it can help classify customers in the following segments:

- Customers who respond to new promotions
- Customers who respond to new product launches
- Customers who respond to discounts
- Customers who show propensity to purchase specific products

2. WHAT IS BIG DATA?

Big Data are high-volume, high-velocity, and/or high-variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization.

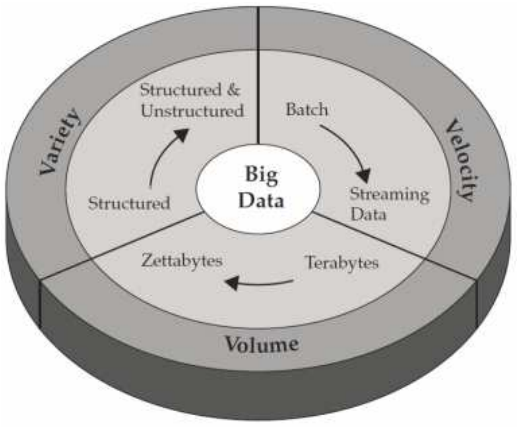

Figure 2.1: Characteristics of Big Data

3. MINING FREQUENT ITEMS ARE SELECTED BY THE CONSUMERS FROM BIG DATA

a. Data storage using HDFS framework:

HDFS is a fault tolerant and self-healing distributed file system designed to turn a cluster of industry standard servers into a massively scalable pool of storage. Developed specifically for large-scale data processing workloads where scalability, flexibility and throughput are critical, HDFS accepts data in any format regardless of schema, optimizes for high bandwidth streaming, and scales to proven deployments of 100PB and beyond.

Key HDFS Features:

- **Scale-Out Architecture** - Add servers to increase capacity
- **High Availability** - Serve mission-critical workflows and applications
- **Fault Tolerance** - Automatically and seamlessly recover from failures
- **Flexible Access** – Multiple and open frameworks for serialization and file system mounts
- **Load Balancing** - Place data intelligently for maximum efficiency and utilization
- **Tunable Replication** - Multiple copies of each file provide data protection and computational performance
- **Security** - POSIX-based file permissions for users and groups with optional LDAP integration

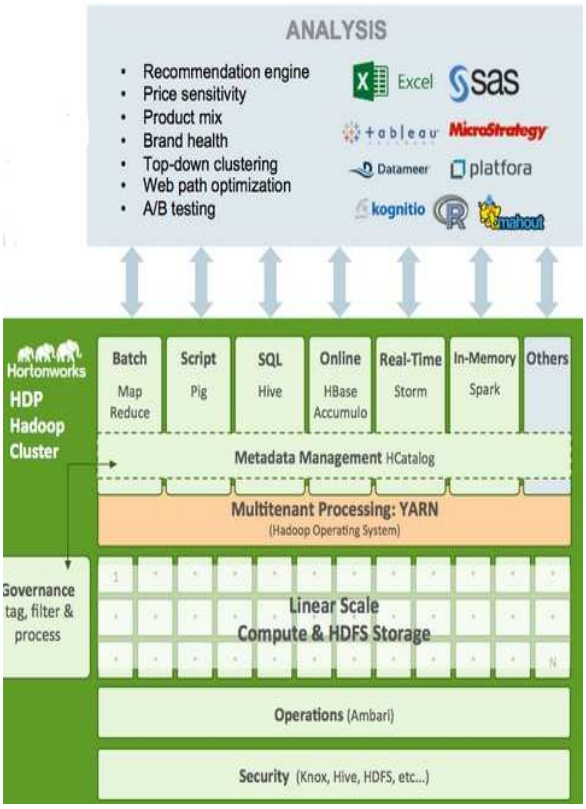

Figure 3.1: Analysis of HDFS

The building blocks of Hadoop

- NameNode
- DataNode
- Secondary NameNode
- JobTracker
- TaskTracker

Whatever the data items are collected by the different resources are stored in HDFS in the form of blocks. The storage portion of the Hadoop framework is provided by a distributed file system solution such as HDFS, while the analysis functionality is presented by MapReduce.

Consider the following table as sample items and their transaction ids.

TID	ITEM 1	ITEM2	ITEM3	ITEM4	ITEM5
T1	1	1	1	0	0
T2	1	1	1	1	1
T3	1	0	1	1	0
T4	1	0	1	1	1
T5	1	1	1	1	0

b. Data processing using MapReduce

The term MapReduce actually refers to two separate and distinct tasks that Hadoop programs perform. The first is the map job, which takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs). The reduce job takes the output from a map as input and combines those data tuples into a smaller set of tuples. As the sequence of the name MapReduce implies, the reduce job is always performed after the map job.

Key MapReduce Features:

- **Scale-out Architecture** - Add servers to increase processing power
- **Security & Authentication** - Works with HDFS and HBase security to make sure that only approved users can operate against the data in the system

- **Resource Manager** - Employs data locality and server resources to determine optimal computing operations
- **Optimized Scheduling** - Completes jobs according to prioritization
- **Flexibility** – Procedures can be written in virtually any programming language
- **Resiliency & High Availability** - Multiple job and task trackers ensure that jobs fail independently and restart automatically

c. Inputs and Outputs

The MapReduce framework operates exclusively on <key, value> pairs, that is, the framework views the input to the job as a set of <key, value> pairs and produces a set of <key, value> pairs as the output of the job, conceivably of different types.

The key and value classes have to be serializable by the framework and hence need to implement the Writable interface. Additionally, the key classes have to implement the WritableComparable interface to facilitate sorting by the framework.

Input and Output types of a MapReduce job:

(i/p) <k1,v1> -> **map** -> <k2, v2> -> **combine** -> <k2, v2> -> **reduce** -> <k3, v3> (o/p)

Map Function

```
Map(input_record){
…
Emit(k1,v1)
…
Emit(k2,v2)
…}
```

Reduce Function

```
reduce(key,values){
while(values.has_next){
aggregate=merge(values.next}
collect(key,aggregate)
}
```

A number of growing opportunities for Hadoop-MapReduce are emerging from a changing environment where Big Data affects IT budgets in two ways:

- Necessity to accommodate exponentially increasing amounts of data (processing, storage, analysis);
- Progressively cost-prohibitive pricing models imposed by established IT vendors.

Example: **Mapper for finding frequent items**

```
import java.io.IOException;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.MapReduceBase;
import org.apache.hadoop.Mapper;
import org.apache.hadoop..OutputCollector;
import org.apache.hadoop.mapred.Reporter;

public class MaxTemperatureMapper extends MapReduceBase implements Mapper<LongWritable, Text, Text, IntWritable>
{
void foundFrequentItemSet(int[] itemset, int support)
 {
              if (usedAsLibrary)
               {
         this.setChanged();
                    notifyObservers(itemset);
              }
              Else
              {
System.out.println(Arrays.toString(itemset) + " ("+ ((support / (double) numTransactions))+" "+support+")");
              }
              }
                          for (int i = 0; i < itemsets.size(); i++)
 {
if ((count[i] / (double) (numTransactions)) >= minSup)
 {
foundFrequentItemSet(itemsets.get(i),count[i]);
frequentCandidates.add(itemsets.get(i));
                                        }
                                        //else log("-- Remove candidate: "+ Arrays.toString(candidates.get(i)) + " is: "+ ((count[i] / (double) numTransactions)));
                          }


}
```

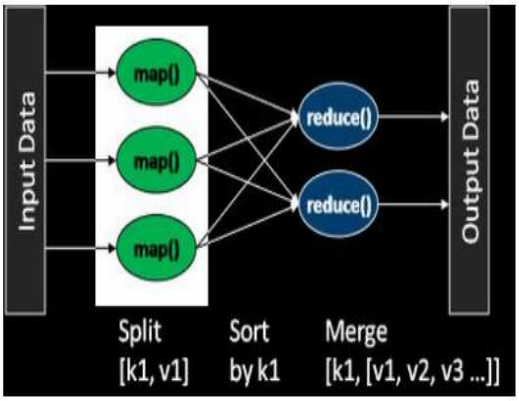

Figure courtesy of Apache/Hadoop

Consumer Buyer Behavior:

- The customer identifies a need.
- Looking for information.
- Checking out alternative products and suppliers Purchase decision.
- Using the product.

Information

Provide comprehensive product information for visitors, using photographs, videos and text to demonstrate product features and benefits. Include independent product reviews by other consumers to build trust and credibility for the products including delivery charges, as well as details of stock availability.

Mobile

Modify your website content so that it is easy to use on mobile devices such as smart phones or tablet computers. Website design must take account of the smaller screen size and limited broadband capacity of mobile devices. Designers of mobile websites recommend using text with minimal graphics for product information and ordering procedures that require few clicks.

Ordering

Make the ordering process simple, with the minimum number of clicks between order placement and checkout. Focus on making the process clear, secure and simple to reduce the risk of shoppers abandoning their carts – one of the biggest problems for online retailers, according to E-consultancy.

Checkout

Design the checkout page so that it does not frustrate customers at the final buying stage, according to Usability Sciences .Show the full cost, including price, taxes and delivery charges before requesting payment details. Ask customers for minimal address and payment information, and avoid asking regular customers to supply information they previously registered.

4. CONCLUSION

Marketing does not stop at understanding the buying processes of your customer however you need to understand their buying patterns and the market in which they operate. In this article we proposed the basic implementations of the storage using hadoop framework and processing will be performed by MapReduce. By using HiveQL we make this paper in an efficient way to collect the consumer inputs and provide better delivery mechanisms.

REFERENCES

[1] 1. Kumar , V . and Reinartz , W . J (2006) Customer relationship Management: A Databased Approach , Hoboken, NJ: John Wiley & Sons .

[2] Hughes , A . M . (2012) Strategic Database Marketing 4e: The Masterplan for Starting and Managing a Profi table, Customer-based Marketing Program McGraw-Hill Professional, USA.

[3] Davenport , T . H . (2009) Realizing the Potential of Retail Analytics: Plenty of Food for Those with the Appetite . Working Knowledge Report, Babson Executive Education

[4] Pattern-Based Strategy: Getting Value from Big Data. Gartner Group press release. July 2011.Available at http://www.gartner.com/it/page.jsp?id=1731916

[5] [Gon2008] Understanding individual human mobility patterns. Marta C. González, César A. Hidalgo, and Albert-László Barabási. Nature 453, 779-782 (5 June 2008)

[6] [LP+2009] Computational Social Science. David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer,Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, and Marshall Van Alstyne. Science 6 February 2009: 323 (5915), 721-723.

[7] [McK2011] Big data: The next frontier for innovation, competition, and productivity. James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela Hung Byers. McKinsey Global Institute. May 2011.

[8] [MGI2011] Materials Genome Initiative for Global Competitiveness. National Science and Technology Council. June 2011.