International Journal of Applied Engineering and Technology ISSN: 2277-212X (Online)
An Open Access, Online International Journal Available at http://www.cibtech.org/jet.htm
2016 Vol. 6 (1) January-March, pp.23-30/Siddiqua et al.
Research Article

DATA MINING CHALLENGES OF SMART RECOMMENDER IN ECOMMERCE

***Syeda Saba Siddiqua[1], Sohail Mohammed[2] and Ashfaq Ahmed Khan[3]**
[1]*Computer Science and Engineering Department, Asifia College of Engineering, Hyderabad*
[2]*Hadoop Developer, Saama Technology, Pune*
[3]*Computer Science Department, Visionary Degree College, Hyderabad*
*****Author for Correspondence*

ABSTRACT
Some 15 to 20 years ago business leaders wanted more data to inform their decisions. But, now a day they are finding themselves soaked into the data that decision making has become more difficult. Organization need mined data for new answers, collection methods, and delivery systems; for that Organizations have to look at the type of data involved structured, unstructured or semi structured as well as latency and complexity. Data sets have their own unique challenges to be faced. They're most difficult to search, store, share and analyze, so any business creating large data sets must have to embrace big data management with the right tools and efficient architectures. This study focus on retail industries as it generates bulk of data each day. Most of the organizations in retail industry use the collaborative and content base filtering for recommenders, but still almost 56% of the retailers do not get the insight enough of the consumer to utilize the big data for competitive advantage. Therefore, we will see how hadoop framework can be used for processing of big data. So, we will research on new technology Hadoop, which is most advance and high in demand analytical tool and also known for fast processing of data at the lowest cost.

Keywords*: *Big Data, Data Mining, Hadoop, Map / Reduce and Recommender

INTRODUCTION
Data are an organization's lifeblood and without that an organization cannot function. Tons of data getting generated in the companies in the recent years, and many organizations-health care, IT, communication, manufacturing, etc. - are dealing with extremely large data sets as they are expanding faster than ever before. One of the most common challenges of big data in any sector is the quality of the data (Computer Science Corporation, 2012). If the data of an organization are inconsistent, variable or incomplete, the insight these could yield becomes doubtful, and the organization might not want to risk acting on it. Hence, analyzing large so-called 'big data' will become a key basis of competition, supporting new impression of productivity growth, consumer surplus and innovation.

The potential value of Business Analytics to organizations has been a frequent topic of discussion across industries over the couple of years. Big Data have compelled organizations to install and stabilize Enterprise data warehouses, Resource Planning (ERP) systems, and Business Intelligence (BI) tools in an effort to more precisely capture operational data and use it to predict future business performance and consumer purchasing trend (Scottmadden, 2014). This availability of data has put forth the apparent question in ecommerce—how do we use it for our competitive advantage by setting personalized recommenders to the individual consumer with right content in the right format at that time so as to convert him from browser to purchaser?

Tools like collaborative filtering and content based filtering are used by organizations in retail industry for recommenders, but still almost 56% of the retailers do not get the insight enough of the consumer to utilize that big data for competitive advantage (Wegener and Sinha, 2013). Organizations have started to realize the importance of big data and are slowly moving towards big data processing tool, hadoop, which is the most advanced, sophisticated, and high in demand data analytical tool. Hadoop uses distributed file system, which runs the job on a cluster of commodity systems, for data processing of large amounts of data in low cost and at high speed. So, we will research on the new technology which is known for fast

processing of the data at the lowest cost and see how it overcomes the challenges of the ecommerce industry.

At first we go with the introduction where it has been discussed, how the big data management tools can leverage the business growth. In section two data processing of big scale data by hadoop technology is discussed, here we can observed how traditional approaches handles data processing and later on hadoop data processing is discussed where it can be seen how hadoop overcomes the challenges of ecommerce industry. In section three collaborative filtering using hadoop file system and map reduce architecture is discussed. In last section quantitative methodology is discussed to understand the trend of the shoppers and their attitude towards the product recommenders. A questionnaire was prepared to understand the trend of the shoppers and their attitude towards the product recommenders.

Data Processing of Big Scale Data by Hadoop Technology

The emergence of ETL tools and Data Integration platforms in the late 90s and early 2000s made mining of data important in every industry by extracting data from key source systems, cleansing and relating that data through transformations, then loading it into the warehouse (Rogers, 2013). Worked well initially, as businesses got to customize their views of information combined from multiple systems, they realized new insights but they also came up with new challenges. They increased the requests for new customized views with new data sources.

They wanted this information more frequently and distributed to larger audiences, requiring more data to be sourced, extracted, cleansed, related, and loaded. This traditional ETL approach, particularly the 'T', began to endure under the weight of these growing volumes of data day by day. In due course, a new approach to data integration was developed. The thought was to Extract the tables from source systems and then load those tables directly into the warehouse into a separate staging area. Use a common and widespread tool, SQL, to relate and combine those raw tables into the tables that are required for analytics and reporting. This approach was known as extract, load, and transfer (ELT) and is the most common approach used for data integration today. Scalability required to perform the 'T' was sorted out by this approach by integrating the growing sources and volumes of data, but there are trade-offs. The logic used for integration is written in SQL that is hard to maintain, requiring constant tuning as data modifies. It was also difficult to test and debug. There was a continuous management of harmonizing the transformation of staging tables with end user query load. It was very difficult to maintain lineage and manage metadata, besides that it fails to perform well most often than not.

By the replacement of ETL in Hadoop, you can provide clear answers to many of these questions. There are massive cost savings in just affordable data warehouse. Customers say that the cost of managing big data in a traditional data warehouse is 100 times the cost of managing the same in Hadoop cluster. One considers Hadoop framework based on return on investments. There is hype in the market for hadoop framework for processing of big data, but the results of data analytics for product recommenders and predictive analytics have proved that it is worth adapting to the latest technology for the organization to rake the profits

Product Recommenders

Web Analytics is a completely a Big Data application which is very closely related to the core of E-commerce. The biggest challenge faced by the Ecommerce Retail Industry today is "A recent study found that 59 percent of the retailers identified a lack of consumer insights as their top data-related pain point". One of the smartest applications derived from Web Analytics which are already successfully applied to ecommerce is Recommendation Systems. The Recommendation Systems reduces the search effort imposed on the customers while finding a product that suits their needs (Ghoshal, 2009). The smart recommenders do not just increase the conversion (browser to purchaser) rate of the customers but also increase the business (new customers). But the associated challenges with the implementation of recommender are: Will the recommender generate profit for an organization by overshadowing the time, effort, and money invested in implementing it? Will it understand the customer insight and provide the recommendation in a timely manner? Or will it provide the products that interest customer to become customers? All these questions can be answered by the best data processing tool.

Collaborative Filtering (CF)

The most used approach for the recommendations in many industries are collaborative and content based filtering approach. For our case study we will see the earlier one. The approach for the designing of recommender systems that has widely used is collaborative filtering (Zhao and Shang, 2010). Filtering by Collaborative methods is based on analyzing and collecting a large amount of Data/Information on behaviors, preferences or activities and prediction of the user. What users will like based on their similarity to that of others'. A key advantage of this filtering approach is that it does not depend on machine analyzable content and hence, it is capable of perfectly recommending intrinsic items such as movies without needing an "understanding" of the item itself. Various algorithms have been used for measuring item similarity or user similarity in recommender systems. Example The k-nearest neighbor (k-NN) approach and the Pearson Correlation.

For designing a model from a user's profile, we ought to have a distinction which is often made between implicit and explicit forms of data collection. Explicit data collection includes the following: Ask users to rate an item on a scale, Ask users to create a list of items that they like. Implicit data collection includes the following: Observe the items that users view in an online store, Maintain an individual record of the items that a user has watched on her/his computer or purchases online, Analyze the user's social network and discover similar likes as well as dislikes

The recommender systems compare the collected data to the similar and dissimilar data collected from others and calculate a list of recommended items for that user. Item-to-Item collaborative filtering (people who buy x also buy y) is one of the most famous examples of collaborative filtering algorithm.

Collaborative Filtering Using Hadoop File System

Scaling up the Item Based CF Recommendation Algorithm using Hadoop file system:

Since the inception of the Collaborative filtering techniques we have seen its wide spread usage in almost every ecommerce recommender system (Marcel, 2012). The remarkable growth of the number of products and customers in recent years causes some key challenges for recommender systems in which high quality and most efficient recommendations are required and more recommendations per second for hundreds of thousands of customers and products need to be performed. So, the improvement of efficiency and scalability of collaborative filtering algorithms become increasingly difficult and important.

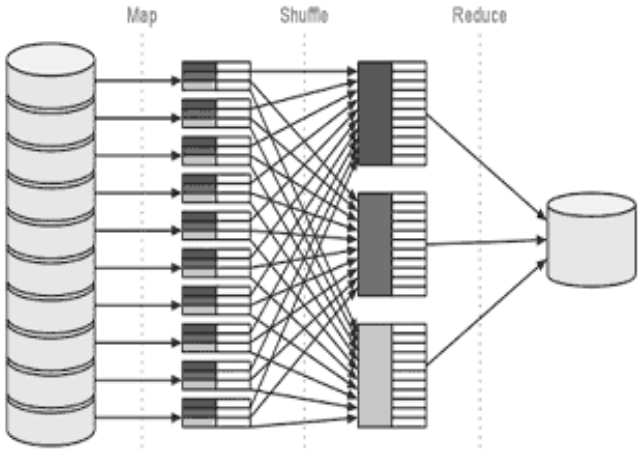

Figure 1: Map-Reduce Architecture

Figure 1 (Marcel, 2012), shows the architecture of the Hadoop framework having map phase, reduce phase, and shuffle & sort phase of hadoop distributed file system. As we have been discussing about the trillions of records of all the customers across the world to be processed at lowest cost and in a timely manner, we will see how hadoop processing proves to be the potential tool.

Research Article

Why go for Map-Reduce?: Map Reduce is an Apache framework originally developed at Google that allows easy large scale distributed computing of data across a number of domains; Apache Hadoop is an open source implementation of it. It can scale well up to many thousands of nodes in a cluster and can handle up to petabytes of data. For product recommendations on ecommerce where we have to find the similar products to a product customer is interested in, we have to calculate how similar pairs of items are (Marcel, 2012). For example, if someone searches for the movie DVD of Gladiator, the recommender would suggest the film Troy. So, we got to compute the similarity between two items. For this one has to find correlation between pairs of products, but if a shopping site has 50,000 products, potentially we will have to compute over 250 billion computations. Besides that the data of the correlation will be sparse because it's not likely that every pair of products will have some user interested in them. So, we have a very large and sparse dataset. And we also got to deal with the products' value based on 'time' aspect since the user interest in products changes with time, for that we need the correlation calculation done from time to time so that the results are up to date. The best way to handle with this scenario and problem is going after a divide and conquer pattern, for which Map-Reduce is a powerful framework and can be used to implement data mining algorithms.

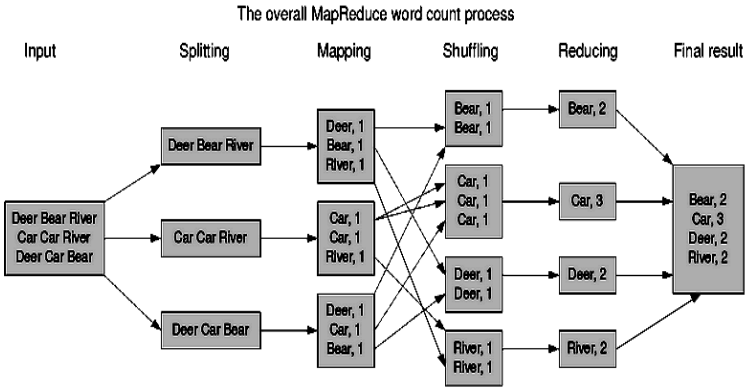

Figure 2: Shows the Map Reduce Job on Hadoop Cluster

The figure 2 (Marcel, 2012), shows the basic Map-reduce processing, where input node, the client which sends the job of count of all the words in the file which then will be split and send to individual nodes. There are 3 nodes below mapping column who work is to take the split data in the form of key value pair and assign the value '1' to all the words and save the intermediate key value pair its node. Once the map function is over it then be shuffled across the network to reducer where data be sorted and aggregated, finally it will be reduced for the final output file. In the above cluster there are 7 nodes (mappers and reducers) that define a cluster. A single file which is small here but if in petabytes will be handled by 7 nodes hence saves time and cost as they are commodity systems.

MATERIALS AND METHODS

Methodology

We have reviewed various papers and articles, written by big data analysts and have been reading research papers on big data from some time now and have read its potential to leverage business if comprehended in the most efficient manner. Then we started reading research paper to know how it can leverage e-commerce since this industry receives the most data each day, especially the big organizations like Amazon, ebay etc. Some of the research paper that talked about recommenders for e-commerce they lay stress on collaborative filtering of data but not much on how can this filtering be done fast by using Hadoop. For improvisation of recommenders, we have been studying Hadoop file distribution technology and its implications to leverage the speed of the data processing.

Research Article

We are using quantitative methodology to understand the trend of the shoppers and their attitude towards the product recommenders. It gives you the flexibility to ask various questions to large number of audience as once mentality totally differs from others. In addition to that different age groups have shown different trend in shopping online, so asking same questionnaires to different groups helps you to understand the way two individual thinks while buying a product, hence gives you an understanding of e shoppers thinking of the recommender.

We have gathered data based on the age group between 15 and 35, as almost 85% of this age group buys product online. Our population sampling contains right mix of Teen guys, teen girls, house wives, working women, and men.

Questionnaires and Scales

We have designed questions based on two criteria, Ease of use and Usefulness of recommenders. We used social network site, face book, and email as a platform to get the responses, besides that we have distributed pamphlets at one of the societal functions and asked the people to fill and return us back at the end of the event.

Questions were:

1. Your name, gender, and age? Name:_____ Gender:__ Age:__
2. What Percentage of time do you buy online?
3. Where do you invest the most? Can choose more than one?
   Electronic Gadgets, cloths, cosmetics, foot wares, or other
4. What makes you to buy online? Can choose more than one
   Ease of use useful
5. Do you use the suggested products to search the product that interests you? Choose one
   Strongly Agree Agree Neutral Disagree Strongly Disagree
6. Were personalized mails and recommendations useful? Choose one
   Strongly Agree Agree Neutral Disagree Strongly Disagree

RESULTS AND DISCUSSION

Table 1: Shows the total number of people surveyed

Age	Gender		Total
	Male	Female	
18	1	2	3
19	3	3	6
20	3	5	8
21	6	9	15
22	7	3	10
23	6	3	9
24	2	5	7
25	4	2	6
26	8	9	17
27	2	1	3
28	1	2	3
29	1	1	2
30	2	2	4
31	0	1	1
32	3	0	3
33	2	1	3
34	1	1	2
35	1	1	2
Total	53	51	104

Above table shows the data collected from age between 15 and 35, the age group known for being on Internet. Totally 104 responded to our survey request.

Table 2: Ease of Access/Usefulness

Males					
		Recommenders / Suggestions Useful	**Suggested Products Useful to Search Other Products**	**Ease of Use**	**Useful**
Strongly Agree		8	11	15	7
Agree		19	16	25	18
Neutral		20	16	10	23
Disagree		5	10	3	3
Strongly Disagree		1	0	0	0
	Total	53	53	53	51

Females					
		Recommenders / Suggestions Useful	**Suggested Products Useful to Search Other Products**	**Ease of Use**	**Useful**
Strongly Agree		7	13	17	9
Agree		20	14	23	22
Neutral		18	13	11	16
Disagree		4	8	0	4
Strongly Disagree		2	3	0	0
	Total	51	51	51	51

From Table 2 it can observe that, the data set is used to know the intent to use it in future. If you see the frequency of the people usage as per their satisfaction level you can see the major chuck of the active people who can be converted to buyers are the 'Agreeable' and the 'Neutral' people. Ones who agree with site and said in the survey that they will use the feature of recommended products for future purchase are the 47% and also 19% neutral.

But there will be various factors involved while purchasing, so consumers must be motivated and get involved into recommenders for their future purchases. It can be analyzed that most of the people feel that recommenders are easier to use, but does this leverage the business? Yes, but partially, because even if 90% of the people using a particular website are happy about the ease of the recommenders, but not buying products online then it is of no use.

So, it should be easier to use along with it giving the insight of the trend of the consumers by using the data mining tool and provide the best recommender which again should be easier to use along with appealing to people to increase the conversion (browser to purchaser) rate.

Table 3: Ease of Use / Like Cross Tabulation

		I Like					
		Strongly Agree	**Agree**	**Neutral**	**Disagree**	**Strongly Disagree**	**Total**
Ease of Use	Strongly Agree	5	12	2	3	0	22
	Agree	9	36	10	4	2	61
	Neutral	1	4	6	2	5	18
	Disagree	0	0	1	0	1	2
	Strongly Disagree	1	0	0	0	0	0
Total		16	52	19	9	8	104

The Table 3 shows the cross tabulation of ease to like which is helpful in getting the insight of the influence to use the recommender. As the ease of use will be the driving force in the customer to purchase a product, combined with liking of the recommendation will add to increase in conversion. Almost 80% of the people like the recommender while they think that it is easy to use. So, further enhancement to add to an ease of the consumer is not going to add to the business value that much than to have a smart recommender to leverage business.

Table 4: Useful / Like Cross Tabulation

		I Like					
		Strongly Agree	**Agree**	**Neutral**	**Disagree**	**Strongly Disagree**	**Total**
Useful	Strongly Agree	7	5	1	0	0	13
	Agree	9	36	8	0	0	53
	Neutral	1	4	13	4	0	22
	Disagree	0	0	3	5	2	10
	Strongly Disagree	0	0	0	1	5	0
Total		17	45	25	10	7	104

Table 4 shows the cross tabulation of usefulness and the liking of the product. This data set is helpful in getting the insight of the attitude and the future interest to use by the consumer. Only around 60 percent of the people think recommenders are useful and hence this gives an insight that the recommender should be smart enough to get to the understanding of the consumer's purchase requirements.

For this to happen we need to take various factor into consideration and must dig further into billions of records for pattern matching using the distributed file system to get the smartest recommendations so that customer should increase its purchasing online. From the data it is clear that at least around 10-15% of the consumers who think that recommenders are easy to use but do not think that recommenders are smart enough to get them the best products.

The reason is very clear that still the retails' recommenders are not getting the insight into the behavior of the consumers to get the best product. This can be leveraged by running collaborative filtering methodology that includes various factors of the consumer behavior on to a bigger Hadoop cluster to fetch the smartest product in milliseconds at the least cost.

Conclusion

Now that we have known the impact of hadoop on the big data for the big organizations that want to tap in the potential market, there is also a way for small organizations to take advantage of this system. For small organization that cannot invest in big amounts and also cannot predict how many servers they need to install as they do not want to take the risk by investing hugely in the beginning, they can make use of Amazon web services (AWS). AWS is the cloud computing, it gives the services present in cloud. The advantage of using such services are you do not have to set up anything just use it by paying for it hourly. When you find huge demand increase the services and pay for it and when demand goes down cut of the services and pay less hence you save more. Once the organization understood that it is adjusted to it and started racking in money they can have their own cluster installed and setup hadoop technology over it. The implications as per now are very less as the system is fault tolerant, but there is big investment in the beginning in terms of cluster set up, huge salaries of hadoop engineers to work on it, and an administrators to monitor the cluster.

REFERENCES

Computer Science Corporation (2012). *Empower your Business with Data Overview Brochure* [Online]. Available: http://www.slideshare.net/GaldeMerkline/empower-your-businesswithdataoverviewbrochure18042012

Scottmadden (2014). *Business Analytics – How to Bridge the Gap Between Knowledge and Results* [Online]. Available: www.scottmadden.com/?a=strm&aid=590 [Accessed 20 July 2014].

Wegener R and Sinha V (2013). The value of Big Data: How analytics differentiates winners [Online]. Available: http://www.bain.com/publications/articles/the-value-of-big-data.aspx.

Rogers J (2013). *Is ETL Dead in the Age of Hadoop*? Available: http://www.wwpi.com/~wwpi/index.php?option=com_content&view=article&id=16125:is-etl-dead-in-the-age-of-hadoop&catid=331:ctr-exclusives&Itemid=2701750.

Chiky R, Ghisloti R and Aoul Z (2010). Development of a distributed recommender system using the Hadoop Framework [Accessed June 19, 2014].

Ghoshal A (2009). Do recommender systems always benefit firms by reducing consumer search effort? [Online]. Available: http://www.academia.edu/257394/DoRecommenderSystemsAlwaysBenefitFirmsbyReducingConsumerSearchEffort.

Zhao ZD and Shang MS (2010). User-Based Collaborative-Filtering Recommendation Algorithms on Hadoop. IEEE Xplore. *Knowledge Discovery and Data Mining* [Accessed 30 July 2014].

Marcel C (2012). *Artificial Intelligence in Motion*. Introduction to recommendations with map-reduce and mrjob. [Online] Available: http://aimotion.blogspot.com/2012/08/introduction-to-recommendations-with.html.

Hu J (2013). *Product Recommendation by Amazon* [Online]. Available: http://www.aboutdm.com/2013/01/product-recommendation-by-amazon.html.

Resnick P and Varian H (1997). Recommender systems. *Communications of the ACM* **30**(3) 56-58 [Accessed July 19, 2014].

Schafer JB, Konstan J and Riedl J (1999). Recommender Systems in E-Commerce. In: EC '99: *Proceedings of the First ACM Conference on Electronic Commerce, Denver, CO,* 158-166 [Accessed July 19, 2014].

Brin S and Page L (1998). The anatomy of a large-scale hypertextual {Web} search engine. *Computer Networks and ISDN Systems* **30**(1-7) 107-117 [Accessed 19 July 2014].