International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 03 Issue: 04 | April-2016 www.irjet.net p-ISSN: 2395-0072

Commercial Product Analysis Using Hadoop MapReduce

Kshitij Jaju[1], Vishal Nehe[2],Abhishek Konduri[3]

[123] *BE IT, Department of Information Technology, Sinhgad College Of Engineering, Pune, Maharashtra, India*

Prof. Sonali Potdar [4]

Dept. of Information Technology, Sinhgad College Of Engineering, Pune, Maharashtra, India

---

Abstract - *In this electronic age, increasing range of organizations face the matter of explosion of knowledge. The databases employed by the organizations these days has been growing exponentially. There are unit sizable amount of sources generating this monumental information like totally different varieties of social networking sites, transactions, social networking sites, business processes, internet servers, etc. the info generated is in structured additionally as unstructured kind. Today's business applications area unit having enterprise options like giant scale, data-intensive, web-oriented and accessed from numerous devices as well as mobile devices. Process or analyzing such giant volume data} and break up out meaningful information kind it's a difficult task. The term "Big information" is employed for big data sets whose size is on the far side the capability of our ancient code tools to amass, administer, and method the info among a tolerable time period. Massive information sizes area unit a steady dynamic target starting from some dozen terabytes to several peta bytes {of information of knowledge of information} in a very single data set. Difficulties introduced owing to this embody acquire, storage, search, analysis and visual image.*

The commercial companies are currently making numerous unsuccessful attempts to identify the customer and their needs. Which also makes the customer face difficulties to fulfill their desired demands. Our aim would be to provide great assistance for the companies to improve their marketing strategies and increase sales. Thus developing a commercial product data analysis project and providing statistical data analysis using MapReduce technique to study and improve their sales by referring the immense data stored is the target we aim to achieve.

***Key Words*: Big Data, Hadoop, MapReduce, Data Analysis, Commercial Products**

1.INTRODUCTION

Incorporating offline and online data can help retailers to reshape targeting, resulting in incremental increases in e-commerce revenues.

Online shopping has developed as one of the most popular Internet activities, providing a variety of products for consumers and a ton of sales challenges for e-commerce players. Research proposes that online sales has increased 16.1% year-over-year from March 2010 through March 2011.

Today all online retailers are competing for the attention of the online consumer. A company can find real opportunities in applying advanced digital analytic techniques that will combine offline and online data sets to reform on-site store inventories. Let's consider the example of Best Buy as in how it took advantage of data in order to gain the attention of its customers. When Best Buy determined that 7% of its customers were responsible for 43% of its sales, the company divided its customers into several archetypes and redesigned stores to acknowledge the buying habits of these particular customer divisions, thereby improving the in-store experience and increasing same-store sales by 8.4%.

Even though optimization and other analytical techniques like A/B/multivariate testing, visitor

engagement, behavioural targeting and audience segmentation indicates towards a high likelihood of a customer's willingness to buy, transactional data such as offline sales are key indicators of actual sales. This paper provides acumen on how merging offline and online data can support modified targeting, resulting in incremental increases in e-commerce revenues.

1.1 BACKGROUND

Online and offline sales in most companies are unfortunately processed in technical silos.Many customers still want to touch and see products before they buy. Conventional study suggests that about 65% of all consumers will make an offline purchase as a result of online marketing. Thus mining offline data along with clickstream data provides a valuable resource for enhancing customer satisfaction, product development, sales forecasting, merchandising and visibility into the profit margin for products and services.

The capability to identify and reach customers at the most crucial point in their buying decision-making process, as assisted by Google's "Zero Moment of Truth" study, exhibited that in-store or online transactions are heavily influenced by intuition gained in the moment before a purchase decision is made. Thus, steering consumers toward a particular product or service and then bagging them at the point of sale is a very powerful solution.

This is why Google's search marketing is so appealing to marketers; consumers' desires are instantly telegraphed by their actions.

2. RELATED WORK DONE

Product analysis and product recommendations are something that each and every online retailers or any other retailers for that fact are trying to excel in. Nowadays the prime objective of any company is to carry out the product analysis in the most effective way and in turn avail the customers with their requirements and make them realize their actual needs. This would prove to be of great assistance for the companies to improve their marketing strategies and increase sales. There are various algorithms that provide heed to such product recommendations to work fine.

Recommendation algorithms are one such kind of algorithm. Recommendation algorithms are best recognized for their use on e-commerce Web sites. Here they use customer's interests as an input to produce a catalog of recommended items. Many applications make use of the products that buyers buy and explicitly rate to depict their personal interests, but they can also use other attributes, including products viewed, favorite artists, subject interests, and demographic data. Amazon.com uses recommendation algorithms to personalize the cart present on online store for every individual customer. The cart thoroughly changes depending upon consumer's interests, displaying gym accessories to people going to gym and baby toys to a new mother.

There are three common perspectives which helps to solve the problem of recommendation: traditional collaborative filtering, search-based methods, and cluster models. Here, amazon compares these perspectives with their algorithm, which is referred as item-to-item collaborative filtering by them. Amazon's algorithm's online computation unlike traditional collaborative filtering, scales independently of the number of products and number of customers in the product catalog. Their algorithm creates and gives recommendations in realtime, and produces high quality recommendations.

These are the systems that help select out similar things whenever you select something online. Netflix for example suggest other movies that you might want to watch, amazon suggests what kind of other products you might want to buy, Facebook will even suggest some other friends that you might want to befriend. Each of these systems operate using the same basic kind of algorithm. There are 2 basic types of algorithms that are at play when we talk about generating recommendations.

The first ones are called content based filtering. Content-based filtering can also be called as cognitive filtering, which recommends products on the basis of a comparison between the content of the products and a user profile. The content of each product is shown as a set of descriptors or terms, which are basically the words which occur in a document. Whereas the user profile is represented with the same terms and built up by analyzing the content of items which have been seen by the use.

And the second one is collaborative filtering. It relies upon not the qualities of the objects itself but how people i.e other users respond to the same objects.

Collaborative filtering can also be called as social filtering, which filters information by making use of the recommendations of other people. Collaborative filtering drives its existence on the basic idea that people who agreed in their assessment of certain products in the past are likely to agree again in the future.

3. PROPOSED WORK

In this paper the proposed system will provide companies or the online retailers a statistical data analysis using MapReduce technique to study and improve their sales by referring the immense data stored in our system.

For providing statistical data analysis we are going to create or produce various types of graphs which would showcase certain product analysis which would help the retailer with information to understand the purchase behavior of a buyer. This information will help the retailer to understand the buyer's needs and reorganize the store's layout accordingly, or even attract new buyers.

For supporting the told concept or methodology what we are going to do is understand the purchase behavior of the buyers. For this purpose apriori algorithm will be made use of. In data mining, Apriori is a typical algorithm for studying association rules. Apriori is designed to work on databases containing transactions (for example, collections of items bought by customers).

Association rules are usually used for products which are bought in lots or in combinations. Apriori algorithm or association rules might tell the retailer that customers who bought whey protein also bought oats.so this information can prove to be important to the retailers. For this purpose support and confidence are calculated. We are going to create a graph which shows all the products which has a confidence greater than a certain threshold.

The overall point of the algorithm (and data mining) is to extract useful information from immense data. For example, the information that a customer who purchases whey protein also tends to buy oats at the same time is acquired from the association rule:-

Support: The percentage of task-relevant data transactions for which the pattern is true.

$$Support(whey\ protein \rightarrow Oats) = \frac{No\ of\ transactions\ containing\ both\ whey\ protein\ and\ oats}{No\ of\ total\ transactions}$$

Confidence: The measure of certitude or reliability associated with each discovered pattern.

$$Confidence(whey\ protein \rightarrow Oats) = \frac{No\ of\ transactions\ containing\ both\ whey\ protein\ and\ oats}{No\ of\ total\ transactions\ containing\ keyboard}$$

3.1 SYSTEM WORKING

The challenge for e-commerce companies is that customers are often relatively far into the purchase funnel when they reach at a particular site. An e-commerce company that can capture the consumer's attention and activity further up the funnel and merge it with advanced analytics techniques can begin to assume the mantle of an analytics leader.

Joining geographic, time of day (e.g., morning, afternoon, evening, etc.), creative data and mapping these attributes against location-based and time-based sales data can highlight hidden interactions between online and offline sales activity. This means e-commerce companies can develop more effective multichannel strategies. The end result is that real additional sales can be increased. Also, e-commerce companies have an option to use advanced Web analytics to create a specific segment for customers who arrived online via offline campaigns. Though web analytic tools provide similar solutions, they are not by any means a full set of offerings nor do they create a single view of your customer. Organizations need to get all the data attributes, offline and online, into a single database, which would be further refined by advanced analytics techniques, and use the combined data for precision targeting.

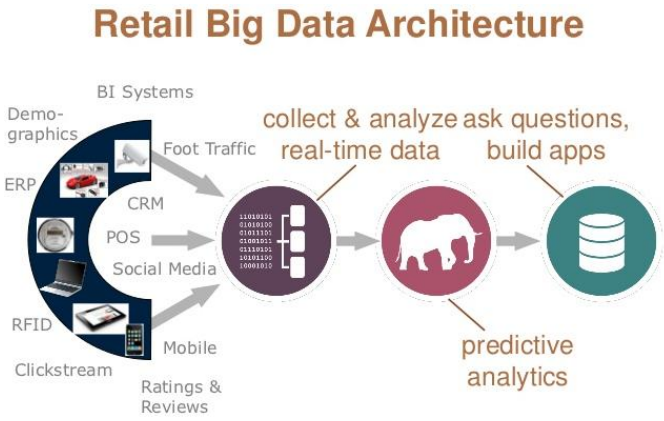

Figure illustrates that by combining Website data (i.e., clickstream information, etc.) with loyalty card insights, third-party information and sales data, e-commerce players can gain critical understanding into customer behavior. To reach the vanguard of the data-driven digital revolution, e-commerce players need to acquire key tools, analytic prowess and necessary skills. E-commerce players need to invest in the right advanced analytics tools and also gain access to mathematicians and statisticians to create models and interpret findings.

Organizations need to administer more complex data calculations to generate a more accurate picture of customer behavior and transactional activity. As many people may think, measuring offline marketing campaigns statistics is not an easy task. There are numerous different tools that can help companies improve data accuracy; however, many are third-party solutions that are not integrated into a Web analytics solution.

4. CONCLUSION

Thus, identifying the most important customers using online and offline data allows companies to more effectively identify those who are transaction oriented and can help boost their revenues. A data-driven optimized strategy is the key, and that is possible only with offline and online data integration. The end result helps companies to reach the right customer with the right product offerings at the right time and place.

ACKNOWLEDGEMENT

We take this opportunity to thank our project guide Prof. Sonali Potdar and Head of the Department Prof. N.J. Uke for their valuable guidance and for providing all the necessary facilities, which were indispensable in the completion of this project report. We are also thankful to all the staff members of the Department of Information Technology of Sinhgad College of Engineering, Vadgaon (B.K.) for their valuable time, support, comments, suggestions and persuasion. We would also like to thank the institute for providing the required facilities, Internet access and important books.

REFERENCES

[1] J. Dean and S. Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters", Google Research Publication, OSDI"04: 6th Symposium on Operating Systems Design and Implementation, pp. 137-149, 2004.

[2] S. Sathya, M. Victor Jose, "Application of Hadoop MapReduce Technique to Virtual Database System Design", International Conference on Emerging Trends in Electrical and Computer Technology (ICETECT), pp. 892-896, 2011.

[3]] K. Shvachko, H. Kuang, S. Radia, R. Chansler, "The Hadoop Distributed File System", IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST), Sunnyvale, California USA, vol. 10, pp. 1-10, 2010

[4] Apache-Hadoop, http://Hadoop.apache.org.