

## A Study on Big Data and its Applications in Retail Sector

G.S. Sra<sup>1\*</sup>, R. Kaur<sup>2</sup>

<sup>1\*</sup>Dept. CE, Yadavindra College of Engineering, Punajbi University Patiala, Kalanwali, India

<sup>2</sup>Dept. CE, Yadavindra College of Engineering, Punajbi University Patiala, Talwandi Sabo, India

\*Corresponding Author: [gagandeepsra.ind@gmail.com](mailto:gagandeepsra.ind@gmail.com)

Available online at: [www.isroset.org](http://www.isroset.org)

Received 14<sup>th</sup> May 2017, Revised 26<sup>th</sup> May 2017, Accepted 19<sup>th</sup> Jun 2017, Online 30<sup>th</sup> Jun 2017

**Abstract** – Big Data is no longer an unpopular word anymore. Any data which poses a challenge for currently existing database technologies is termed as Big Data. Today only big giants in the different fields like retail, medical, stock exchange etc. can think of handling it because of expensive and huge infrastructure involved in handling it. This research paper is based on highlighting how retail and cellular industry makes its use in targeting customers. In today's busy life no one loves to receive unnecessary junk mails or messages. It is because of this that companies want to be crystal clear about each of its customer interest and his or her pattern of shopping and expenditure. For handling such a huge database of millions of customers, companies make use of powerful framework like Apache Hadoop. Apache Hadoop is one such framework which is capable of handling huge databases via its several components. It makes use of Map Reduce technology.

**Keywords** - Apache Hadoop; Big Data; Map Reduce; retail industry.

### I. INTRODUCTION

“Big data” – which undeniably means numerous things to numerous people – is not limited to the area of technology. As big data is the junction of huge data from numerous sources than people have ever gotten, it signifies a cultural shift too in the means retailers interact with clients in a right way. This core influence of big data is that which makes it a business critical and why retailers in all over the world are benefiting it to alter their processes, their organizations and, soon, the entire industry. Enormous data can be created as consumers and businesses go into the area of universal connectivity that describes the Internet of Everything (IoE) world. IoE associate data, processes, people & things to allow the transmission of information and produce new possibilities for business novelty. Devices & Sensors provide data from earlier separate processes and their components, increasing the part of data in making of decisions crosswise the whole retail enterprise.

Analytics are driving the move from merchant-driven business models—where the product is the differentiator—to digital models, where every decision is informed by data. Brand associations are becoming extra closely aligned with distinct shopper favorites, making a brand association that is everchanging from “nice to have” time related proposal-dependent bond to “must have” digital camaraderie depending on broad insights and empathetic of the consumer. To attain this serious diversity, retailers are dependent less on progressively smaller product cycles and more on the suffering distinguisher of association and client experience

formed through planned use of data and analytics. Companies progressively know that their skill to contest is bond to their capability to produce and harness worth from data, and are looking for new methods to look at big data and past [1, 2].

### II. CHALLENGES RELATED WITH BIG DATA

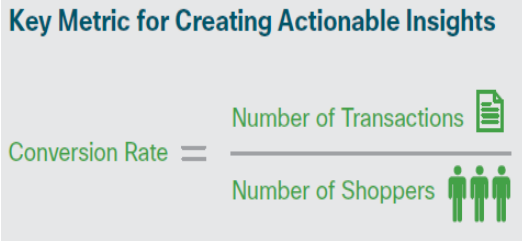
**Data Volume** – It refers to the enormous amount of data that is been created each second, each minute and each hour of the day. 571 websites are created in every minute. Entire 625000 GB of data is moved from one end to another in every single minute of internet, may be terms of mails, posts, pictures etc. If we burn the amount of data present on planet earth today on DVDs and pile them in the form of a stack one upon another, the pile will be such huge that one can climb it and touch the moon, come back to earth and again repeat this process once.

**Data Velocity** – Data is being created at such high velocity that companies are finding it difficult to cope up with such high speed. They have to establish their infrastructure in such a manner that it is capable of handling such generated data Social media, E-Commerce has rapidly increased the speed and richness of data used for different business transactions.

**Data Variety** - All the data being generated is totally diverse consisting of raw, structured, semi structured and even unstructured data which is hard to be managed by the available traditional systems for analytic. Mismatched data formats and data structures represent significant challenges that can lead to analytic collapse [1, 6].

### III. NEED TO UNDERSTAND CONVERSION RATE

In specialty retail, it is generally accepted that conversion is the standard for comparative performance measurement. Conversion is a foundational metric—the first of the predictive analytics building blocks and a requirement for creating high-performance retailing in the IoE (Internet of Everything) age. It regularizes sales performance against the mutable chance related with each store, aisle, and type. There are different definitions of conversion, but the true basic calculation is quite simple: Conversion is the measure of transactions generated by a population of shoppers (Fig. 3). For example, if 1,000 shoppers generate 800 transactions, the result is an 80% conversion rate.



$$\text{Conversion Rate} = \frac{\text{Number of Transactions}}{\text{Number of Shoppers}}$$

Fig. 3. Formula for conversion rate

Some retailers, particularly in fast-moving consumer products segments, use the number of transactions as a proxy for conversion, assuming that all shoppers generate at least one transaction. This is fundamentally incorrect and can cause businesses to overlook key problem areas. Of course, in grocery and other areas where conversion is usually very high, some of the most valuable conversion comparisons actually come from aisle-, category- and display-level analysis.

Let's examine a typical store-to-store comparison using sales revenue and transaction count. If Store A reports POS revenue of \$10,000 and 60 transactions, and Store B reports \$20,000 and 100 transactions, the conclusion might be that Store B is the better performer. However, once shopper traffic is taken into consideration, we may find that Store A had 100 shoppers that week and Store B had 200. We now must consider the performance results in the context of one very important variable: opportunity. What opportunity did each store have that week to generate sales? Let's look at the population of shoppers. Using this definition of the opportunity variable, we see that Store A actually converted 60% of its opportunity, while Store B only converted 50%. The sales numbers and ATV are of course better in Store B, but over time they can create a false sense that the business is healthy, when in fact more shoppers may be leaving without purchasing. If Store B could actually convert at the rate of Store A (60%) and keep its average transaction value (\$200/transaction) the same, it would contribute \$4,000 more

each week, even without any improvement in traffic. That would be an additional \$208,000 per year from a single store! [8, 9].

### IV. BIG DATA IN RETAIL EXAMPLES

#### A. Hotel Chain Uses Big Data to Increase Bookings

The smart hoteliers recognized that cancelled flights often leave travelers in need of a place to sleep overnight.

The company provided freely accessible weather and flight annulment information, prearranged by groupings of hotel and airport sites, and made an algorithm which factored weather severity, travel conditions, time of the day and cancellation rates by airport and airline among other variables. With its big data insights, and identification that travelers will be utilizing mobile modules for this use case, the company transported beleaguered mobile ads to stranded explorers and form it easy for them to book a nearby hotel. With this it was found that concerned hoteliers made profit which increased by 10%.

#### B. Pizza Chain Earns More Dough in Bad Weather

Slightly like to the above instance, a pizza chain utilizes a mobile app and mobile marketing methods to transport coupons depending on worst weather or where power cuts leave customers incapable to cook. This mobile and location-dependent advertising movement attains a 20% response rate.

#### C. Music distributor Applies Big Data for Demand Planning

Record label EMI uses big data to measure and forecast product demand. After distributing or leaking music, the company measures consumption on its own social networks and additionally acquires third party listening pattern data from popular music streaming services, song identification apps or 'second screen' social media collators. The data is aggregated by demographics, locations and subcultures and helps the music distributor transport selective publicity and predict product demand with a great sureness level. This concept is applicable to other retailers who can also aggregate feeds from social networks to build an understanding of how new products will be received by new or existing markets, or even how their products and company reputation are perceived among the public.

### V. TECHNOLOGIES HANDLING BIG DATA

#### A. MPP – Massively Parallel Processing

Massive Parallel Processing (MPP) is the “shared nothing” method of parallel computing. It is a kind of computing where the process is being completed by several CPUs employed in parallel to perform a single task. One of the most significant differences between a Symmetric Multi-Processing or SMP and Massive Parallel Processing is that with MPP, each of the several CPUs has its individual

memory to support it in avoiding a probable hold up that the user may get by means of SMP when entire of the CPUs attempt to access the memory at simultaneously.

The idea behind MPP is truly just that of the all-purpose parallel computing in which the concurrent implementation of some grouping of manifold cases of programmed instructions and data on numerous processors in so that the outcome can be gotten a lot extra efficient and fast.

Massively parallel processing (MPP) is a method of cooperative processing of the similar program by two or extra processors. Each processor handles different threads of the program, and each processor itself has its own operating system and dedicated memory. A messaging interface is required to allow the different processors involved in the MPP to arrange thread handling. Sometimes, an application may be handled by thousands of processors working collaboratively on the application. MPP is a complex process demanding some jobs to be shared between all involved processors. Messages are swapped between processors through an interconnection of data tracks during MPP. MPP is classically got in applications like decision support systems and data warehouses. Supercomputers are also instances of MPP architecture.

The Massively Parallel Processing relational database architecture spreads data over a number of independent servers, or nodes, in a manner transparent to those using the database. Big Data environments often use analytic MPP systems usually called “shared-nothing” databases. In this the nodes that make up the cluster operate independently and communicate via a network but do not share disk or memory resources. With modern multi-core CPUs, MPP databases can be configured to treat each core as a node and run tasks in parallel on a single server.

Big Data can be a big headache for organizations that have outgrown the practicality and usefulness of single-server analytical tools, especially where self-service reporting is a high priority. That’s why successful Big Data users are financing in Massively Parallel Processing (MPP) hardware, a scalable computer architecture that leverages numerous commodity CPUs possibly hundreds or thousands to tackle huge scale analysis. Because companies collect and store extra and extra granular data from a widening number of business operations and other sources, databases have grown into huge information silos containing millions, even billions, of data records. The compounding consequence of database magnitude and quantity, together with growing user need for interactive transactional and supervisory decision support, has needed analytical systems capable to assimilate information from numerous processes, and gage to very big data sizes without forgoing ease of use, query performance or uptime.

### *B. NoSQL (Not Only SQL)*

NoSQL [7, 5] technology was founded by leading internet corporations including Google, Facebook, Amazon and LinkedIn to overawed the boundaries of 40-year-old relational database expertise for usage with modern web applications. Today, enterprises are accepting NoSQL for a rising amount of cases, a possibility that is determined by four interrelated megatrends: Internet of Things, Big Users, Big Data & Cloud Computing.

Relational and NoSQL data models are very different. The relational model receipts data and splits it into numerous interrelated tables that encompass rows and columns. Tables reference each other through foreign keys that are kept in columns as well. When seeing data, the wanted material has to be fetched from numerous tables (often hundreds in today’s enterprise applications) and united before it can be given to the application. Likewise, when scripting data, the script wants to be coordinated and performed on numerous tables [9].

### *C. Hadoop and Map Reduce*

Hadoop [7, 8] is a java based framework that is efficient for processing large data sets in a distributed computing environment. Hadoop is sponsored by Apache Software Foundation. The creator of Hadoop was Doug Cutting and he baptized the framework after his child’s satiated toy elephant. Applications are made run on systems with thousands of nodes making use of thousands of terabytes via Hadoop. Distributed file system in Hadoop eases fast data transmission amongst nodes and permits incessant processes of the system even if node failure happens. This idea decreases the risk of calamitous system failure even if numerous nodes become inoperative. The motivation behind working of Hadoop is Google’s Map reduce that is a software framework in which application under contemplation is fragmented into amount of minor parts [5, 6].

MapReduce [8] is a framework originally developed at Google that allows for easy large scale distributed computing across a number of domains. The Apache Hadoop software library is a framework that permits for the distributed processing of huge data sets crosswise clusters of computers by means of modest programming models. It is designed to gage up from solitary servers to thousands of machines, each offering local calculation and storing. Hadoop MapReduce contains several phases, each with an significant set of processes supporting to go to your goal of getting the responses you essential from big data. The process begins with a user appeal to track a MapReduce program and remains until the outcomes are written back to the HDFS.

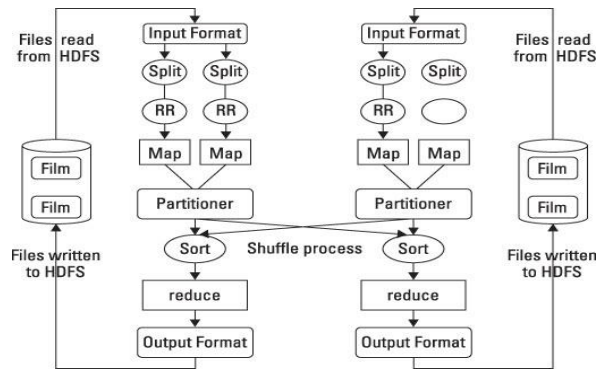


Fig. 6 Working of Map Reduce Technology

MapReduce is an architectural model for parallel processing of tasks on a distributed computing system. This algorithm was first described in a paper “MapReduce Simplified Data Processing on Large Clusters,” by Jeffery Dean and Sanjay Ghemawat from Google. This algorithm allows splitting of a single computation task to multiple nodes or computers for distributed processing.

**VI. IMPLEMENTATION**

The below in Fig 7 the snap shot of constructed database is shown. On running appropriate queries on this database in Hadoop framework desired results are obtained as shown in Fig. 8.

PRODUCT	TYPE	CATEGORY	PRICE_START_RANGE	PRICE_END_RANGE	MRP	SEASON	MONTH_OF_MAXSALE	MFG_
Film Escal Bluetooth Multimedia Speaker Sytem	NonConsumable	Electronic	2000	5000	5200	Both	January	20
Lenovo Vibe K5 Plus	NonConsumable	Electronic	7300	9999	10000	Both	March	20
Inspect and Mosquito Killer with Night Lamp	NonConsumable	Electronic	149	449	503	SUMMER	June	20
Budgetstone 8390 205/65/15 Tubeless Tyer	NonConsumable	Car_Tyer	5300	8400	9000	Both	May	20
FAB M 70W 120mm Ceiling Fan	NonConsumable	Electronic	999	1700	2000	Summer	March	20
EP03 3 Blade Table Fan	NonConsumable	Electronic	2500	3500	3900	Summer	March	20
Trans Air 4 Blade Exhaust Fan	NonConsumable	Electronic	1400	1700	1999	Both	May	20
Aura Metallic 47" 1200 cm Ceiling Fan	NonConsumable	Electronic	2000	2300	2500	Summer	June	20
Formula 1 Nutritional Shake Mix	Consumable	Health_Drink	1200	1900	2300	Both	May	20
Laser U11300W All Purpose Home Blower Tower Fan	NonConsumable	Electronic	2800	3800	4100	Both	All	20
Aqua supreme RO_UV_TDS 10 Litr Water Purifier	NonConsumable	Electronic	4000	8000	10000	Both	June	20
GC1010 Steam Iron	NonConsumable	Electronic	1000	1400	1700	Both	All	20
16 litre Unbreakable Non-Electric Water Purifier	NonConsumable	Non_Electric	700	1200	1544	Both	June	20
H114 Dry Iron	NonConsumable	Electronic	800	995	1156	Both	All	20
Cantor Wheels Air Cooler	NonConsumable	Electronic	9000	13555	19000	Summer	May	20

Fig. 7 Constructed database

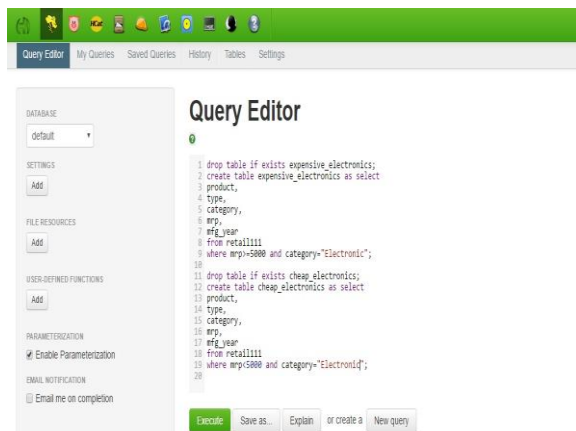


Fig. 8(a) Results obtained

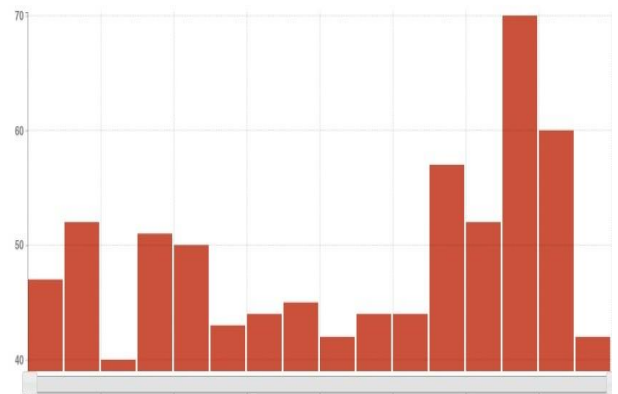


Fig. 8(b) Results obtained

**VII. PROMOTING BIG DATA ADOPTION**

*Commit initial efforts to customer-centric outcomes*

It is authoritative that governments attention big data initiatives on zones that can deliver the most worth to the business. For most retailers, this will mean beginning with customer analytics that enable them to offer better, more finely tailored products based on a better understanding of customer needs and predicted behavior patterns. Retail organizations can use customer insights to generate enhanced products, improve brand performance, drive customer loyalty, adjust pricing and improve customer satisfaction.

To effectively cultivate meaningful relationships with their customers across all retail channels (e.g., store, online, e-mail, mobile, etc.), retailers must connect with them in ways their customers perceive as valuable. The value may come through preferred features and pricing and more timely, informed or relevant interactions; it may also come as organizations improve their underlying operations in ways that enhance the overall consumer experience. Retailers should identify the processes that most directly affect customers, pick one and start; even small improvements matter as they often provide the proof points that demonstrate the value of big data and the incentive to do more. Analytics solutions fuel the insights from big data that are becoming essential in creating the level of depth in relationships that customers expect.

**VIII. CONCLUSION**

In this paper, we have introduced a vision of analytics as a new guiding principle for operating in today’s tumultuous retail environment. As we know that it is easy to lie with statistics, but it is difficult to tell truth without statistics, we’ve discussed the power of becoming a data driven decision-making culture, and shown how access to accurate, scalable, and actionable data can help retailers set a roadmap to success through a better understanding of their customers and of their store operations. We’ve also covered how data can reveal exposures as well as opportunities for the retailer.

As it is important to know who is not purchasing and why can be as important as understanding those who do purchase. The right insights enable a closer, stronger relationship with consumers.

### REFERENCES

- [1] A. Katal, M. Wazid, R. Goudar “*Big Data: Issues, Challenges, Tools and Good Practices*”, 2013 Sixth International Conference on Contemporary Computing (IC3), *Indian*, pp. 404-409, 2013.
- [2] S. Kaisler, F. Armour, J. Espinosa “*Big Data: Issues and Challenges Moving Forward*” International Conference on System Sciences, Hawaii, pp. 995-1004, 2013.
- [3] R. Kitchin, “*Big data and human geography: Opportunities, challenges and risks*”, *Dialogues in human geography*, Vol.3, Issue.3, pp.262-267.
- [4] J.Stuart Ward, Adam Barker “*Undefined by Data: A Survey of Big Data Definitions*”, School of Computer Science-University of St. Andrews, UK, pp.1-120, 2013.
- [5] M.S. Al-Hakeem, “*A Proposed Big Data as a Service (BDaaS) Model*”, *International Journal of Computer Sciences and Engineering*, Vol.4, Issue.11, pp.1-6, 2016.
- [6] Atsushi Sato, Runhe Huang, “*Retail Business Intelligence*”, IEEE International Conference on Data Science and Data Intensive Systems, USA, pp.211-219, 2015.
- [7] Sunny Kumar, “*Big Data Platform-A Review*”, *International Journal of Computer Sciences and Engineering*, Vol.3, Issue.10, pp.84-87, 2015.
- [8] Gagandeep Jagdev, “*Affirmative aspects of Big Data assures new revolution in retail sector*”, RTEST-2016 at Rayat-Bahra University, India, pp.23-28, 2016.
- [9] Gagandeep Jagdev, “*Constructive aspects of Big Data in retail sector for financial growth*”, CIOGEBBS-2016 at Maharaja Ranjit Singh State Technical University, India, pp. 77-79 April 2016.