International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 1 | ISSN : 2456-3307

Big Data : The Futuristic Promising Savoir

Tawseef Ayoub Shaikh, Umar Badr Shafeeque, Maksud Ahamad

Department of Computer Engineering, Aligarh Muslim University, Uttar Pradesh, India

ABSTRACT

Big Data, a new jackpot in the world of vocabulary is the recent hot term which has made itself omnipresent in debate and occupied its place on almost every lip. Data as usual is somehow known to everyone and now that data is not only data, it's Big Data. Big but how much? "Big Data" is typically considered to be a data collection that has grown so large it can't be effectively or affordably managed (or exploited) using conventional data management tools: e.g., classic relational database management systems (RDBMS) or conventional search engines, depending on the task at hand. Big Data is more a concept than a precise term. Some apply the "Big Data" label only to petabyte-scale data collections (> one million GB). For others, a Big Data collection may house 'only' a few dozen terabytes of data. More often, however, Big Data is defined situation ally rather than by size. Specifically, a data collection is considered "Big Data" when it is so large an organization cannot effectively or affordably manage or exploit it using conventional data management tools. Why is Big Data different from any other data that we have dealt with in the past? IBM defined Big Data having 4 V's as its key characteristics such as: Volume, Velocity, Variety, and Veracity.

Keywords : Big Data, RDBMS, IDC, IoT, CSP, Healthcare, Security

I. INTRODUCTION

1.1: Volume

Poor fellow, he suffers from files—Aneurin Bevan

Data is everywhere ranging from Online shopping sites, banks, healthcare, business, credit card, web logs, Social Networks, Streaming data, Smart phones, Sensors as in Internet of Things (IoT). The St. Anthony Falls Bridge (which replaced the 2007 collapse of the I-35W Mississippi River Bridge) in Minneapolis has more than 200 embedded sensors positioned at strategic points to provide a fully comprehensive monitoring system where all sorts of detailed data is collected and even a shift in temperature and the bridge's concrete reaction to that change is available for analysis. IDC estimates that in 2010 alone they generated enough digital information worldwide to fill a stack of DVDs reaching from the earth to the moon and back. Volume is the scale and size of the data available today. Most organizations were already struggling with the increasing size of their databases as the Big Data tsunami hit the data stores. Fortune magazine claimed to have created 5 exabytes of digital data in recorded time until 2003. In 2011, the same amount of data was created in two days. By 2013, that time period is expected to shrink to just 10 minutes. A decade ago, organizations typically counted their data storage for analytics infrastructure in terabytes. They have now graduated to applications requiring storage in petabytes. This data is straining the analytics infrastructure in a number of industries. For a communications service provider (CSP) with 100 million customers, the daily location data could

amount to about 50 terabytes, which, if stored for 100 days, would occupy about 5 petabytes.

A clear look on this Ocean of data can be had by having a glance on below facts:

- ✓ In the year 2000, 800,000 petabytes (PB) of data were stored in the world [1].
- ✓ In 2008, number of devices connected to Internet exceeded world population.
- ✓ In 2020, 40 zettabyte of data will be there that is 57 times the number of grains of sand on all beaches in the world.
- ✓ Face book has 40 petabyte of data and captures 100 TB/day and makes 800 million updates per day.
- ✓ Yahoo has 60PB of data and has 250 million tweets per day.
- ✓ Twitter captures 8TB/day.
- ✓ EBay has 40PB of data and captures 50TB/day [1].
- ✓ New York stock exchange 1TB data every day.
- ✓ YouTube users upload more than 48 hours of videos every minute and has 4 million views per day.
- ✓ Google gets 1 Billion queries per day.
- ✓ 90% of all data produced so far is only in last two years and it will be 44 times in 2020 than in 2009.
- ✓ 2.5 Quintillion Bytes/ day.
- ✓ In 2012 Health Care data reached 500 petabyte and is expected to reach 25000 petabyte in 2020 and Medical data doubles every 5 years [2].
- ✓ US health care has already reached to a mark of 150 exabytes [2].

1.2: Velocity

It's the speed at which data is produced, analyzed and stored. There are two aspects to velocity, one representing the throughput of data and the other representing latency. Throughput represents the data in the pipes. The amount of global mobile data is growing at a 78 percent compounded growth rate and is expected to reach 10.8 exabytes per month in 2016 as consumers share more pictures and videos. To analyze this data, the corporate analytics infrastructure is seeking bigger pipes and massively parallel processing. Latency is the other measure of velocity [3].

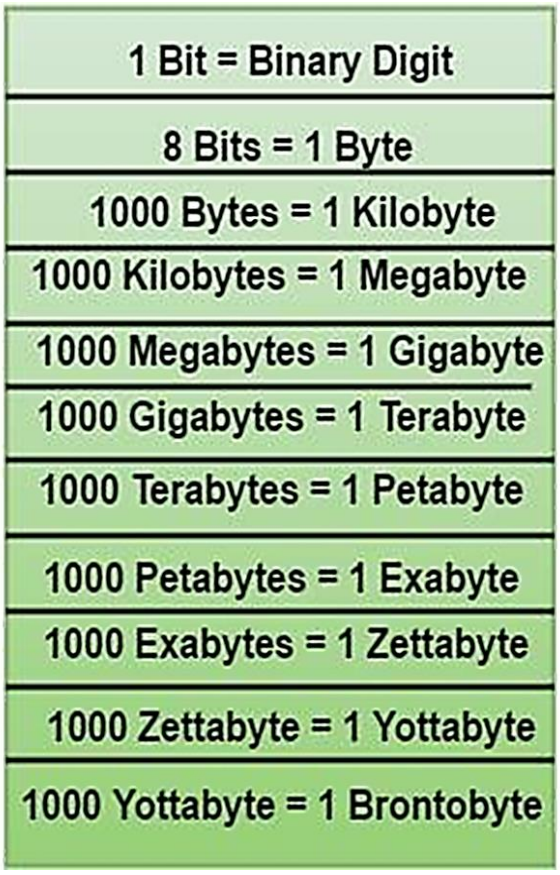

Figure 1: Different measuring units in Big Data

1.3: Variety

It refers to the Complexity of the data. Initially Data was stored in the tables like Relational tables which were predefined structure. But with the data available from diverse sources and possessing diverse formats e.g. in case of Healthcare data comes in the form of Clinical Notes, Lab Tests, Medical Images, Streams from Smart Sensors, it is the utmost need to integrate these diverse data formats so as derive the productive knowledge, which is not possible from a single source of data.

1.4: Veracity

Parameter used to measure the Quality, validity and volatility so as to be sure about the accuracy of data, reliability of the data source, context within analysis. Unlike carefully governed internal data, most Big Data comes from sources outside our control and therefore suffers from significant correctness or accuracy problems. Veracity represents both the credibility of the data source as well as the suitability of the data for the target audience.

II. BIG DATA APPLICATIONS

Big Data has laid its marks on almost every sphere of life. Below are some of the very few areas where it can be harnessed for productive benefits:

i) In Banking: The use of customer data invariably raises [3] privacy issues. By uncovering hidden connections between seemingly unrelated pieces of data, big data analytics could potentially reveal sensitive personal information. Research indicates that 62% of bankers are cautious in their use of big data due to privacy issues [4].

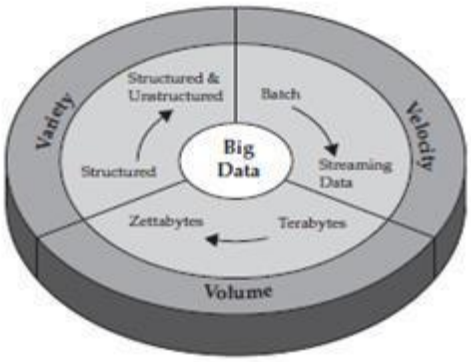

Figure 2 : IBM characteristics of Big Data by its V's

ii) In Stock: A private stock exchange in Asia uses in database analytics to establish a comprehensive system to detect abusive trading patterns to detect fraud [5].

iii) In Credit Cards: Credit card companies rely on the Speed and accuracy of in database analytics to identify possible fraudulent transactions [6]. By storing years' worth of usage data, they can flag atypical amounts, locations, and retailers, and follow up with cardholders before authorizing suspicious activity.

iv) In Enterprise: For enterprises around the world, in many industries, in-database analytics are providing a competitive advantage. When data doesn't have to commute to work and back, it can deliver faster insights that help businesspeople make informed decisions in real time for less expense than traditional data analysis tools [7].

v) In Consumer Goods: A maker of consumer products collects consumer preference and purchasing data extracted from surveys, purchases, web logs, product reviews from online retailers, phone conversations with customer call centers, even raw text picked up from around the Web [8, 9]. Their ambitious goal: to collect everything being said and communicated publicly about their products and extract meaning from it. By doing this, the company develops a nuanced understanding of why certain products succeed and why others fail. They can spot trends that can help them feature the right products in the right marketing media. Amazon gets 30% of Sales because of Recommendation.

vi) In Agriculture: A biotechnology firm uses sensor Data to optimize crop efficiency [10, 11]. It plants test crops and runs simulations to measure how plants react to various changes in condition. Its data environment constantly adjusts to changes in the attributes of various data it collects, including temperature, water levels, soil composition, growth, output, and gene sequencing of each plant in the test bed. These simulations allow it to discover the optimal environmental conditions for specific gene types.

vii) In Economy: Designed from the ground up to deal intelligently with commodity hardware,

Hadoop can help organizations transition to low cost servers. Information on human behavior is not only being collected on new multinational scales, but they are becoming more accessible than ever before thanks to an Open Data movement, in which organizations disclose their data to the public in order to uncover interesting patterns. Governments and social welfare organizations are able to collect information on larger dimensions, reaching new populations as collection technology moves from paper to tablet. Finally, classical measures of real economic activity, such as inflation and unemployment, are transformed from slow-moving [12].

viii) In Finance: A major financial institution grew up of using third-party credit scoring when evaluating new credit. Employee monitoring and surveillance. Predictive models, such as those that may be used by insurance underwriters to set premiums and loan officers to make lending decisions. Developing algorithms to forecast the direction of financial markets. Pricing illiquid assets such as real estate [12].

viii) In Conservation*:* Keeping data in a merged, isolated system provides business intelligence benefits and is both financially and ecologically sound.

ix) In Marketing*:* Marketers have begun to use facial Recognition software to learn how well their advertising succeeds or fails at stimulating interest in their products. A recent study published in the Harvard Business Review looked at what kinds of advertisements compelled viewers to continue watching and what turned viewers off. Among their tools was "a system that analyses facial expressions to reveal what viewers are feeling." The research was designed to discover what kinds of promotions induced watchers to share the ads with their social network, helping marketers create ads most likely to "go viral" and improve sales [12].

x) In Smart Phones: Perhaps more impressive, people now carry facial recognition technology in their pockets. Users of I Phone and Android smart phones have applications at their fingertips that use facial recognition technology for various tasks. For example, Android users with the remember app, can snap a photo of someone, then bring up stored information about that person based on their image when their own memory lets them down a potential boon for salespeople. I Phone users can unlock their device with recognize me, an app that uses facial recognition in lieu of a password. If deployed across a large enterprise, this app could save an average of $2.5 million a year in help-desk costs for handling forgotten passwords.

xi) In Telecom: Now a day's big data is used in Different fields. In telecom also it plays a very good role. Service providers are trying to compete in the cutthroat world of telecom services. Where mare and more subscribers rely on over-the-top (OTT) players as providers of value-added services are focused on increasing revenue, reducing open, chum and enhancing the customer experience as key business objectives. Operators believe that big data and advanced analytics will play a critical role in helping them meet their business objectives. In the same survey, respondents indicate critical use case scenarios in the context of big data and advanced Analytics where they are investing now and where they plan to invest in the next three years. Operators face an uphill challenge when they need to deliver new, compelling, and revenue generating services without overloading their networks and keeping their Running costs under control. The market demands new set of data management and analysis capabilities that can help service providers make accurate decisions by taking into account customer, network context and other critical aspects of their businesses. Most of these decisions must be made in real time, placing additional pressure on the operators. Real-time predictive analytics can help leverage the data that resides in their multitude systems, make it immediately accessible and help

correlate that data to generate insight that can help them drive their business forward.

xiii) In Health care: Traditionally, the health care industry has lagged behind other industries in the use of big data, part of the problem stems from resistance to change providers are accustomed to making treatment decisions independently, using their own clinical judgment, rather than relying on protocols based on big data [13, 14]. Other obstacles are more structural in nature. Many health care stakeholders have under invested in information technology because of uncertain returns. Although their older systems are functional, they have limited ability to standardize and consolidate data. The nature of health care industry itself also creates challenges: while there are many players, there is no way to easily share data among different providers or facilities, partly because of privacy concerns. Even within a single hospital, payer, or pharmaceutical company, important information often remains siloed within one group or department because organizations lack procedures for integrating data and communicating findings. Health care stakeholders now have access to promising new threads of knowledge. This information is a form of "big data," so called not only for its sheer volume but for its complexity, diversity, and timelines. Pharmaceutical industry exports, payers, and providers are now beginning to analyze big data to obtain insights. Although these efforts are still in their early stages, they could collectively help the industry address problems related to variability in Health care quality and escalating health care spend. Researchers can mine the data to see what treatment are more effective for particular conditions, identify patterns related to drug side effects or hospital read missions, and gains other important information that can help patients and reduce costs. Recent technologic advances in the industry have improved their ability to work with such data, even though the files are enormous and often have different database structures and technical characteristics.

xiv) Customer relationship management: The cost of retaining customers is significantly lower than the cost of replacing them, making the ability to identify customers at risk of churning vital [15]. Key Performance Indicators are used to describe customers, including demographic information and recent call patterns for each individual customer. Predictive models based on these fields use changes in customer call patterns that are consistent with call patterns of customers who have churned in the past to identify people having an increased churn risk. Customers identified as being at risk receive additional customer service or service options in an effort to retain them.

xv) Social network analysis: The increasing use of social networks, such as Facebook, Twitter, and Weibo (http://www.weibo.com/), has produced and is producing huge volume of data. Twitter posts more than 500 million tweets every day. Weibo is reported to have over 766 million active users per day in 2014. Business firms andother organizations are interested in discovering new business insight to increase business performance. By using advanced analytics, enterprises can analyze big data to learn about relationships underlying social networks that characterize the social behavior of individuals and groups. Using data describing the relationships, we are able to identify social leaders who influence the behavior of others in the network, and on the other hand, to determine which people are most affected by other network participants. We can also use diffusion analysis to identify the individuals most affected by the group leaders and target the marketing to them [16].

xvi) Transports and smart cities: A large number of data is being gathered every hour in today's cities, but there is surprisingly little global analysis that is being done on it. While combining data from multiple sources needs to be done in a careful way to preserve privacy, the benefits of being able to detect abnormal situations or discover surprising relations between events definitely make it worthwhile. This

area is a prime example of the need for combining very diverse types of information, and for presenting results in a flexible way.

xvii) Urban and physical planning: Data for urban and physical planning is collected and produced by local, regional and national authorities, but is not generally shared and used in an efficient manner. To this data from all available sources can be added and used. An important part of this is to create work processes from the early data.

III. CONCLUSION

Big Data has really changed every path of the present day life. There is no filed fully escaped from its effects. Data is the biggest asset nowadays and Data Scientist job is been expected as the sexiest job of the recent times. Big Displays a pivotal role in Personalization of things whether in Marketing, Healthcare, Purchase, Social Networks which help in better understanding of the customer behaviors, their likes, choices and accordingly there future prediction is made by analyzing upon their present data. Big Data is future to the IT sector. No field can fully escape form it. Big Data promises to be an everlasting career of the society and when analyzed properly it will defiantly change every bit of our life by changing our traditional way of living to the modern Smart Life. At the end I will finish it with a Proverb:

"If you are having data, mine it and take decisions. If you don't have data then take my Opinions".

IV. REFERENCES

[1]. Big Data Analytics by Dr Arvind Sethi

[2]. Big Data Analytics by Kim H. Pries and Robert Dunnigan

[3]. Analytics: The real-world use of big data Tom Inman, Vice President, IBM Software Group.

[4]. https://www.evry.com/globalassets/insight/bank2020/bank-2020---big-data---whitepaper.pdf Last visited 09-01-2018].

[5]. TP. Oberst, "Applications in Finance for BIG DATA", Advanced Strategic Technology, pp: 3-18, March 18, 2015.

[6]. Peter Groves, Basel Kayyali, David Knott, Steve Van Kuiken, "The big data revolution in health care," enter for US Health System,Reform Business Technology Office, published in January 2013.

[7]. S. Kavitha, RP. Vadhana and AN. Nivi," BIG DATA ANALYTICS IN FINANCIAL MARKET", IJRET: International Journal of Research in Engineering and Technology, Vol: 04 Issue: 02, pp: 422-427, Feb-2015.

[8]. "Data-driven healthcare organizations use big data analytics for big gains" by IBM software.

[9]. D. Dua , L. Aihua and L. Zhangb "Survey on the Applications of Big Data in Chinese Real Estate Enterprise", 1st International Conference on Data Science, ICDS 2014, Procedia Computer Science, Elsevier, Vol: 30, pp: 24-33, 2014.

[10]. "Deep learning applications and challenges in big data analytics" by Maryam M Najafabadi, Flavio Villanustre, Taghi M Khoshgoftaar, Naeem Seliya, Randall Wald and Edin Muharemagic.

[11]. MR. Bendre, RC. Thool and VR. Thool, "Big data in precision agriculture: Weather forecasting for future farming", 1st International Conference on Next Generation Computing Technologies (NGCT), pp: 4-5, Sept. 2015.

[12]. L. Einav "The Data Revolution and Economic Analysis", Stanford University and NBER Jonathan Levin, National Bureau of Economic Research, pp: 1-24, 2014.