ISSN:2320-0790

An International Journal of Advanced Computer Technology

The Big Data

A GROWING TORRENT OF TECHNOLOGY

Prachi More[1], Latika Chaudhary[2], Sangita Panmand[3], Prof. Nilesh Shah[4]

Department of Computers, Padmabhushan Vasantdada Patil Pratishthan's College of Engineering, Mumbai

[1] prachi.more91@gmail.com

[2] latikac92@gmail.com

[3]sangitapanmand@gmail.com

Abstract: Demand and spurt in collections and accumulation of data has coined new term "Big Data" has begun. Accidently, incidentally and by interaction of people, information so called data is massively generated. This BIG DATA is to be smartly and effectively used Computer scientists, physicists, economists, mathematicians, political scientists, bio-informaticists, sociologists and many Variety of Intellegesia debate over the potential benefits and costs of analysing information from Twitter, Google, Facebook, Wikipedia and every space where large groups of people leave digital traces and deposit data. Given the rise of Big Data as both a phenomenon and a methodological persuasion, it is time to start critically interrogating this phenomenon, its assumptions and its biases. Big Data, which refers to the data sets that are too big to be handled using the existing database management tools, are emerging in many important applications, such as Internet search, business informatics, social networks, social media, genomics, and meteorology. Big Data presents a grand challenge for database and data analytics research. This paper is a blend of non-technical and introductory-level technical detail, ideal for the novice. We conclude with some technical challenges as well as the solutions that can be used to these challenges. Big Data differs from other data with five characteristics like volume, variety, value, velocity and complexity. The article will focus on some current and future cases and causes for BIG DATA.

Keywords: Large Datasets, Infinite Data Information, Advanced Data Warehouse

I. INTRODUCTION

Many resources have stated that every day, we create 2.5 quintillion bytes of data, so much that 90% of the data in the world today has been created in the last two years alone. Big Data is a large volume of data from various data sources such as social media, health care industry, web, genomics, cameras, medical records, aerial sensory technologies, satellite communication, everyday video recording, transactions in commercial world, cell phones communication in terms of text or pictures, and information sensing mobile device and even traditional enterprise data like customer, employer and so on. This data is **Big Data.** Big Data is a popular term used to describe the exponential growth, availability and use of information, both structured and unstructured. Much has been written on the big data trend and how it can serve as the basis for innovation, differentiation and growth. According to IDC, it is imperative that organizations and IT leaders focus on the ever-increasing volume, variety and velocity of information that forms big Data. "Big Data" refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze. This definition is intentionally subjective and incorporates a moving definition of how big a dataset needs to be in order to be considered big data; we assume that, as technology advances over time, the size of datasets that qualify as big data will also increase. [9]

II. REQUIREMENT

Metadata or Knowledge from data should be target of creator and user of the data right from organization till scientist. With the progress in science ,creators of data (i.e. different smart sensors, progress in satellite communication)—both explicit sensors like point-of-sales scanners and RFID tags, and implicit sensors like cell phones with GPSs and search activity. Harnessing both explicit and implicit human contribution leads to far more profound and powerful insights than traditional data analysis alone, e.g.: Google can detect regional flu outbreaks seven to ten days faster than the Centers for Disease Control and Prevention by monitoring increased

search term activity for phrases associated with flu systems. Satellite sensors can predict the possibility of tsunami or cyclones. MIT researchers were able to predict location and social interactions by analyzing patterns in geo/spatial/proximity data collected from students using GPS-enabled cell phones for a semester. IMMI captures media rating data by giving participants special cell phones that monitor ambient noise and identify where and what media (e.g., TV, radio, music, video games) a person is watching, listening to, or playing. Competitive advantage comes from capturing data more quickly, and building systems to respond automatically to that data. The practice of sensing, processing, and responding is arguably the hallmark of living things. We're now starting to build computers that work the same way. And we're building enterprises around this new kind of sense-and-respond computing infrastructure. As our aggregate behavior is measured and monitored, it becomes feedback that improves the overall intelligence of the system, a phenomenon Tim O'Reilly refers to as harnessing collective intelligence. With more data becoming publicly available, from the Web, from public data sharing sites like Info chimps, Swivel, and IBM's Many Eyes, from increasingly transparent government sources, from science organizations, from data analysis contests (e.g., Netflix), and so on, there are more opportunities for mashing data together and open sourcing analysis. Bringing disparate data sources together can provide context and deeper insights than what's available from the data in any one organization. Experimentation and models drive the analysis culture. At Google, the search quality team has the authority and mandate to fine-tune search rankings and results. To boost search quality and relevancy, they focus on tweaking the algorithms, not analyzing the data. Models improve as more data becomes available, e.g., Google's automatic language translation tools keep getting better over time as they absorb more data. Even Map Reduce programming was published by Google for big data which is again sensible.

III. Architecture

The defining processing capabilities for big data architecture are to meet the volume, velocity, variety, and value requirements. Unique distributed (multi-node) parallel processing architectures have been created to parse these large data sets. There are differing technology strategies for real-time and batch processing requirements. For real-time, key-value data stores, such as NoSQL, allow for high performance, index-based retrieval. For batch processing, a technique known as "Map Reduce," filters data according to a specific data discovery strategy. (Map reduce is programming model for handling complex combination of several tasks. It runs ad-hoc query and works in two steps 1 Map: Queries are divided into sub queries and allocated to several nodes in the distributed system and processed in parallel. 2. Reduce: Results are assembled and delivered.)After the filtered data is discovered, it can be analyzed directly, loaded into other unstructured databases, sent to mobile devices, or merged into traditional data warehousing environment and correlated to structured data.

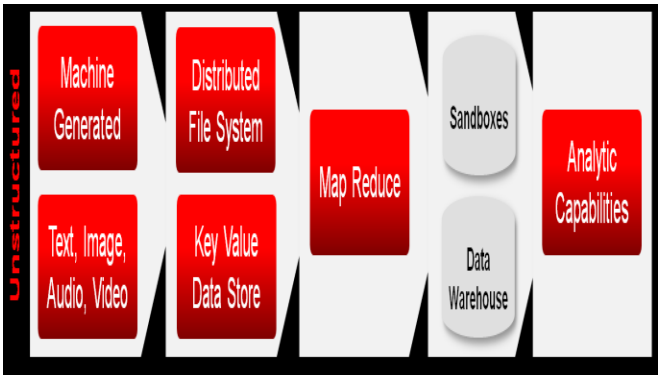

Figure 1: Big Data Information Architecture Capabilities

It is ideal to have analytic capabilities that combine a conventional BI platform along with big data visualization and query capabilities. And second, to facilitate analysis in the Hadoop environment, sandbox environments can be created. The Big Data architecture challenge is to meet the rapid use and rapid data interpretation requirements while at the same time correlating it with other data. [8]

IV. Dimensions

Most of theories confines as Big data are categorised as:

A. Volume.

IBM research finds that every day we add 2.5 quintillion bytes (2.5x10^18). Facebook alone adds 500 TB daily and Google processes about 1 Peta byte of data every hour**.** It is also sourced that 200 countries have been found twitting 500 millions tweets a day. Overcoming the volume issue requires technologies that store vast amounts of data in a scalable fashion and provide distributed approaches to querying or finding that data. Two options exist today: Apache Hadoop based solutions and massively parallel processing databases such as CalPont, EMC Green Plum, EXASOL, HP Vertica, IBM Netezza, Kognitio, ParAccel, and Teradata Kick fire. [6]

B. Velocity:

The rate of data growth is also phenomenal By Moore's law data is doubling every 2 years**.** Velocity describes the frequency at which data is generated, captured, and shared. The growth in sensor data from devices, and web based click stream analysis now creates requirements for greater real-time use cases. The velocity of large data streams power the ability to parse text, detect sentiment, and identify new patterns. Real-time offers in a world of

engagement require fast matching and immediate feedback loops so promotions align with geo location data, customer purchase history, and current sentiment. Key technologies that address velocity include streaming processing and complex event processing. NoSQL databases are used when relational approaches no longer make sense. In addition, the use of in-memory data bases (IMDB), columnar databases, and key value stores help improve retrieval of pre-calculated data. [6]

C. Variety:

The data that is getting added is also of numerous types like unstructured feeds, social media data, sensor data, and multi-media data. Data no longer fits into neat, easy to consume structures. New types include content, geo-spatial, hardware data points, location based, log data, machine data, metrics, mobile, physical data points, process, RFID's, search, sentiment, streaming data, social, text, and web. The addition of unstructured data such as speech, text, and language increasingly complicate the ability to categorize data. Some technologies that deal with unstructured data include data mining, text analytics, and noisy text analytics. [6] Different data with different characteristics highlight the importance and complexity required to solve context in big data. [6] Bringing such variety on high bandwidth of internet is another tactical game.

D. Viscosity:

Viscosity measures the resistance to flow in the volume of data. This resistance can come from different data sources, friction from integration flow rates, and processing required turning the data into insight. Technologies to deal with viscosity include improved streaming, agile integration buses, and complex event processing. [6]

E. Virality:

Virality describes how quickly information gets dispersed across people to people (P2P) networks. Virality measures how quickly data is spread and shared to each unique node. Time is a determinant factor along with rate of spread. [6] All such V's implies that the humans analysing, detecting patterns and making sense of data need to have rich toolset at hand.

V. TECHNOLOGIES

The rising importance of Big-Data computing stems from advances in many different technologies:

A. Sensors:

Digital data are being generated by many different sources, including digital imagers (telescopes, video cameras, MRI machines), chemical and biological sensors (microarrays, environmental monitors), and even the millions of individuals and organizations generating web pages. [7]

B. Computer networks:

Data from the many different sources can be collected into massive data sets via localized sensor networks, as well as the Internet. [7]

C. Data storage:

Advances in magnetic disk technology have dramatically decreased the cost of storing data. For example, a one-terabyte disk drive, holding one trillion bytes of data, Costs around $100. As a reference, it is estimated that if all of the text in all of the books in the Library of Congress could be converted to digital form, it would add up to only around 20 terabytes..[7]

D. Cluster computer systems:

A new form of computer systems, consisting of thousands of "nodes," each having several processors and disks, connected by high-speed local-area networks, has become the chosen hardware configuration for data-intensive computing systems. These clusters provide both the storage capacity for large data sets, and the computing power to organize the data, to analyze it, and to respond to queries about the data from remote users. Compared with traditional high-performance computing, where the focus is on maximizing the raw computing power of a system, cluster computers are designed to maximize the reliability and efficiency with which they can manage and analyze very large data sets. The "trick" is in the software algorithms – cluster computer systems are composed of huge numbers of cheap commodity hardware parts, with scalability, reliability, and programmability achieved by new software paradigms. [7]

E. Cloud computing facilities:

The rise of large data centers and cluster computers has created a new business model, where businesses and individuals can *rent* storage and computing capacity, rather than making the large capital investments needed to construct and provision large-scale computer installations. For example, Amazon Web Services (AWS) provides both network-accessible storage priced by the gigabyte-month and computing cycles priced by the CPU-hour. Just as few organizations operate their own power plants, we can foresee an era where data storage and computing become utilities that are ubiquitously available. [7]

F. Data analysis algorithms:

The gigantic volumes of data require automated or semi automated analysis – techniques to detect patterns, identify anomalies, and extract knowledge. Again, the "trick" is in the software algorithms - new forms of computation, combining statistical analysis, optimization, and artificial intelligence, are able to construct statistical models from large collections of data and to infer how the system should

respond to new data. For example, Netflix uses machine learning in its recommendation system, predicting the interests of a customer by comparing her movie viewing history to a statistical model generated from the collective viewing habits of millions of other customers. [7]

VI. USES

Question is how do you en-cash the BIG DATA. Big Data has many application areas, intents and area of usability. Different intents can be combating crime, real time road way communication, quality of research, to prevent epidemics, to mark business areas. It can be used for complex physics, meteorology ,medicines and many other application areas across industry domains like financial industry, insurance, retail industry, Mobility, Health care,, ERP and CRM. The hopeful vision for big data is that organizations will be able to harness relevant data and use it to make the best decisions. Technologies today not only support the collection and storage of large amounts of data, they provide the ability to understand and take advantage of its full value, which helps organizations run more efficiently and profitably. For instance, with big data and big data analytics, it is possible to: Analyze millions of SKUs to determine optimal prices that maximize profit and clear inventory. Recalculate entire risk portfolios in minutes and understand future possibilities to mitigate risk. Mining the customer data for insights that drive new strategies for customer acquisition, retention, campaign optimization and next best offers. [3] It will indicate the pattern of customer mind set and project the possible incoming business ensuring a higher redemption rate .Send tailored recommendations to mobile devices at just the right time, while customers are in the right location to take advantage of offers. Analyze data from social media to detect new market trends and changes in demand.

VII. REAL TIME EXAMPLES

Big Data played vital role in elections last couple of year s in USA as well in our country at north India. 2012 USA election Obama's team tasked with data analysis using social media posts, voter list, decision making of voters mind set, designing effective policies to persuade them. The analysis provided crucial insights about the voters who are most like to switch sides and the required shooting points for the switch. It also helped to find out crucial geographies to amend changes for the voters. [12] RFID (radio frequency ID) systems generate up to 1,000 times the data of conventional bar code systems. 10,000 payment card transactions are made every second around the world. Wal-Mart handles more than 1 million customer transactions an hour. 340 million tweets are sent per day. That's nearly 4,000 tweets per second. Facebook has more than 901 million active users generating social interaction data. More than 5 billion people are calling, texting, tweeting and browsing websites on mobile phones. [3]

VIII. CHALLENGES

The "Bigness" of Big Data is posting new sets of challenges and threats. Much of the technology required for big-data computing is developing at a satisfactory rate due to market forces and technological evolution. For example, disk drive capacity is increasing and prices are dropping due to the ongoing progress of magnetic storage technology and the large economies of scale provided by both personal computers and large data centers. Other aspects require more focused attention, including:

A. High-speed networking:

Although one terabyte can be stored on disk for just $100, transferring that much data requires an hour or more within a cluster and roughly a day over a typical "high-speed" Internet connection. (Curiously, the most practical method for transferring bulk data from one site to another is to ship a disk drive via Federal Express.) These bandwidth limitations increase the challenge of making efficient use of the computing and storage resources in a cluster. They also limit the ability to link geographically dispersed clusters and to transfer data between a cluster and an end user. This disparity between the amount of data that is practical to store, vs. the amount that is practical to communicate will continue to increase. We need a "Moore's Law" technology for networking, where declining costs for networking infrastructure combine with increasing bandwidth.

B. Income or Return on Investment:

How visible it is to see the financial impact on the use of Big Data is really challenge. The new exciting arena like twitter, Facebook, LinkedIn used by industry from retail to large scale manufacturing, the veracity of inputs received from them remains a matter of concern. There is no straight measuring scale for measuring ROI from such social and business exercises. [12]

C. Cluster computer programming:

Programming large-scale, distributed computer systems is a longstanding challenge that becomes essential to process very large data sets in reasonable amounts of time. The software must distribute the data and computation across the nodes in a cluster, and detect and remediate the inevitable hardware and software errors that occur in systems of this scale. Major innovations have been made in methods to organize and program such systems, including the Map Reduce programming framework introduced by Google. Much more powerful and general techniques must be developed to fully realize the power of big-data computing across multiple domains. [7]

D. Extending the reach of cloud computing:

Although Amazon is making good money with AWS, technological limitations, especially communication bandwidth, make AWS unsuitable for tasks that require extensive computation over large amounts of data. In addition, the bandwidth limitations of getting data in and out of a cloud facility incur considerable time and expense. In an ideal world, the cloud systems should be geographically dispersed to reduce their vulnerability due to earthquakes and other catastrophes. But, this requires much greater levels of interoperability and data mobility. The Open Cirrus project is pointed in this direction, setting up an international tasted to allow experiments on interlinked cluster systems. On the administrative side, organizations must adjust to a new costing model. For example, government contracts to universities do not charge overhead for capital costs (e.g., buying a large machine) but they do for operating costs (e.g., renting from AWS). Over time, we can envision an entire ecology of cloud facilities, some providing generic computing capabilities and others targeted toward specific services or holding specialized data sets. [7]

E. Machine learning and other data analysis techniques:

As a scientific discipline, machine learning is still in its early stages of development. Many algorithms do not scale beyond data sets of a few million elements or cannot tolerate the statistical noise and gaps found in real-world data. Further research is required to develop algorithms that apply in real-world situations and on data sets of trillions of elements. The automated or semi automated analysis of enormous volumes of data lies at the heart of big-data computing for all application domains. [7]

F. Widespread deployment:

Until recently, the main innovators in this domain have been companies with Internet-enabled businesses, such as search engines, online retailers, and social networking sites. Only now are technologists in other organizations (including universities) becoming familiar with the capabilities and tools. Although many organizations are collecting large amounts of data, only a handful are making full use of the insights that this data can provide. We expect "big-data science" – often referred to as science – to be pervasive, with far broader reach and impact even than previous-generation computational science. [7]

G. Security and privacy:

Data sets consisting of so much, possibly sensitive data, and the tools to extract and make use of this information give rise to many possibilities for unauthorized access and use. Much of our preservation of privacy in society relies on current inefficiencies. For example, people are monitored by video cameras in many locations – ATMs, convenience stores, airport security lines, and urban intersections. Once these sources are networked together, and sophisticated computing technology makes it possible to correlate and analyze these data streams, the prospect for abuse becomes significant. In addition, cloud facilities become a cost-effective platform for malicious agents, e.g., to launch a bonnet or to apply massive parallelism to break a cryptosystem. Along with developing this technology to enable useful capabilities, we must create safeguards to prevent abuse. [7] There are other challenges need to be faced by the enterprises or media when handling Big Data is like capture, duration , storage , search , sharing, analysis, visualization.

IX. SOLUTIONS

Smartest ways day by day to extract the core intelligence from big data and making it lightening fast on social network will yield tremendously.Re-invent new business possibilities, modernize operations, and make the best possible decisions in real time – with Big Data. [10] You need to manage the volume, variety and velocity of data, apply analytics for better insight and use this insight to help make better business decisions faster than the competition. And you need to accomplish these goals cost effectively. [11] With Big Data, you can re-imagine Decision making: Empower every employee to make informed and intelligent decisions. Business processes: Accelerate and improve processes using real-time Big Data insights Possibilities: Delight your customers with real-time engagement and differentiated experiences. Information technology: Simplify and modernize your IT strategies and landscape environment. Insights: View insights through a consumer's lens, and get instant answers to complex business questions.[10] Specific actions that the federal government could take include: Give the NSF a large enough budget increase that they can foster efforts in big-data computing without having to cut back on other programs. Possible economies of scale could be realized by consolidating these into a small number of "super data centers" provisioned as cloud computing facilities. This approach would provide opportunities for technologists to interact with and support domain scientists more effectively. These efforts should be coupled with large-scale networking research projects. An adversary with very modest financial resources could have access to supercomputer-class computer facilities. Big-data computing is perhaps the biggest innovation in computing in the last decade. We have only begun to see its potential to collect, organize, and process data in all walks of life. A modest investment by the federal government could greatly accelerate its development and deployment.[7] Convincing to the user how effective and cost saving information or tool is this from big data should be ball game of dynamics of big data.

X. CONCLUSIONS

Flood of data coming from many sources must be handled using some non traditional database tools which will provide different science and market value for the upcoming generation. One way of looking at big data is that it represents the large and rapidly growing volume of information that is mostly untapped by existing analytical applications and data warehousing systems. The actual technologies used will depend on the volume of data, the variety of data, the complexity of the analytical processing workloads involved, and the responsiveness required by the business. It will also depend on the capabilities provided by vendors for managing, administering, and governing the enhanced environment. These capabilities are important selection criteria for product evaluation. Big data, however, involves more than simply implementing new technologies. It requires senior management to understand the benefits of smarter and timelier decision making. It also requires business users to make realistic decisions about agility requirements for analyzing data and producing analytics, given tight IT budgets. Plethora of data coming from many sources must be handled using some non-traditional database tools. It provides more market value and systematic for the upcoming generation. Intelligence from big data will play vital role in any process and commercial activities.

XI. ACKNOWLEGEMENT

We would like to acknowledge ibm.com and various sites for the valuable information for helping us understand the technology. Secondly we would like to acknowledge photo courtesy of oracle.com for sharing their knowledge about our research topic.

XII. REFERENCES

[1] http://www-01.ibm.com/software/data/bigdata/

[2] www.sas.com/resources/asset/SAS_BigData_final.pdf

[3] http://www.sas.com/big-data

[4] http://my.safaribooksonline.com/book/technology-management/9780596803582/big-data-technologies-and-techniques-for-large-scale-data/why_big_data_matters

[5] theprofessionalspoint.blogspot.in/2012/12/real-time-big-data-defination-need-uses.html

[6] http://blog.softwareinsider.org/2012/02/27/mondays-musings-beyond-the-three-vs-of-big-data-viscosity-and-virality/

[7] cra.org/ccc/docs/init/Big_Data.doc

[8] www.oracle.com/technetwork/topics/entarch/articles/oea-big-data-guide-1522052.pdf

[9] www.journalistsresource.org/wp-content/uploads/2013/03/MGI_big_data_full_detail.pdf

[10] www54.sap.com/solution/big-data.html

[11] www.ibm.com/systems/x/solutions/analytics/bigdata.html

[12] www.forbes.com/sites/netapp/2012/11/06/big-data-election-surprising-starts