International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 07, July - 2017

Challenges and Opportunities with Big Data

Prarthana T.V
Computer Science and Engineering
B.N.M Institute of Technology,
Bangalore, India

Jebah Jayakumar
Computer Science and Engineering
B.N.M Institute of Technology,
Bangalore, India

***Abstract* - We are living in an era of big data – an age of huge information. Big Data incorporate endless information. Big Data is getting larger in industries and providing better business. It has changes the world in the terms of predicting customer's behavior. Another buzz word these days is social networks and relation between two of these is very obvious yet complicated. Both Big Data and Social network are interdependent to each other as most of the data is generating from social networking sites but not all the Big Data is useful. The actual challenge of Big Data is not collecting it but managing it and making sense out of it. Such huge data is becoming a new technology focus both in science and in industry and motivate technology shift to data centric architecture and operational models. There is a vital need to define the basic information/semantic models, architecture components and operational models that together comprise a so-called Big Data Ecosystem. This paper mainly focuses on introducing the concept of big data, the terminology used and the architectural details of the big data ecosystem. Researchers and practitioners are trying to look into the future of big data so that more and more benefit can be taken out of the large amount of data. Hadoop - MapReduce is getting prominent in the field of Big Data. Big Data is being used in several research areas related to health-care, location based services, satellite information usage as well as online advertising and retail marketing. This also tries to establish a connect between big data, online social networking, Hadoop, MapReduce.**

Keywords - Big Data, Hadoop, Mapreduce

I. INTRODUCTION

With the vast development of internet, it has become really easy to stay connected and communicate with your friends and family through various social and professional networking channels like Facebook, Twitter, Linkedin, and Skype. With such strong social networking in place, the volume of data being collected, stored, and analyzed has exploded especially with the activity on the Web and mobile devices. The total amount of data in the world was 4.4 zettabytes in 2013. That is set to rise steeply to 44 zettabytes by 2020. Almost 2.5 Exabyte of data is being produced per day on internet. Facebook unveiled that 500+ terabytes of new data is created by its users every day. When faced with this quantity of data human-powered systems quickly become infeasible. This has led to a rise in the so-called big data. Arrival of Internet of Things (IoT) will increase the level of data getting created every day. Apart from social networks, traffic update, weather update and mobile phones are also creating huge amount of data. For instance, a half of terabytes of data is being generated by a Boeing 787 in one flight. According to an estimation done by IBM, 2.5 quintillion bytes of data are getting created every day and 90% of data in the world has been created in last few years.

Such huge data is being created in variety of fields like - biotech, energy, IoT, healthcare, automotive, space and deep sea explorations, cybersecurity, social media, telecom, consumer electronics, manufacturing, gaming and entertainment – the list goes on. Recent research has found that less than 0.5 percent of that data is actually being analyzed for operational decision making. Analysis of this data in the right direction would enable to users to take decisions either strategically to make important long-term decisions, or in real-time to make operational decisions. Such analysis can be achieved only if meaningful information can be extracted from unstructured data. This high volume, high variety data is mostly unstructured and and traditional platforms are not sufficient for analyzing this data. It requires adequate infrastructure and programming paradigms capable of processing large amount of data. Hadoop, the most known open-source implementation of the Map Reduce paradigm, is widely employed in such frameworks. It helps in performing all data analytical tasks in two functions Map() and Reduce() using basic concept of key-value pairs. It provides high scalability and powerful fault tolerance. MapReduce is playing significant role in Big Data analysis.

II. DEFINITION OF BIG DATA

Despite the "Big Data" became a new buzz-word, there is no consistent definition of Big Data, nor detailed analysis of this new emerging technology. Most discussions until now have been going in blogosphere where active contributors have generally converged on the most important features and incentives of the Big Data.

Big Data can be defined to have the following 5V properties:

1) Volume,
2) Velocity,
3) Variety that constitute native/original Big Data properties,
4) Value
5) Veracity as acquired as a result of data initial classification and processing in the context of a specific process or model.

Volume: This indicates the huge size of data being produced. Amount of data being generated at web using social networks like facebook, twitter, flicker etc is increasing exponentially. The data produced is estimated using zettabytes rather than terabytes. Not just social media but other sources such as sensors and weblogs are also

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 07, July - 2017

contributing in this phenomenon of data generation at big amount.

Velocity: The data is being generated at very high speed. Velocity indicates the attribute of high speed at which data streams are arriving continuously. Handling the data streams coming at high velocity is a real challenge in the mining of Big Data.

Variety: Most of the data arriving is heterogeneous and unstructured. There are multiple sources and format of data like images, text, videos, sensor's output data etc. Here the challenge is to extract and analyze meaningful information from unstructured data.

Big Data, now a day, not only classified by its volume, velocity and variety but also by two newly introduced attributes which are described as follows:
Veracity: Data coming from several sources is noisy. It needs cleaning and elimination of noise so that data quality can be maintained and data analysis can be accurate and useful.
Value: Today's data has a cost. This data can be sold and bring money to company. It is important for a company to understand and estimate the cost of data storage and understand the value it has in market to understand if it's beneficial to invest money in that data. For instance, is it worth to spend 100 dollar on the resources needed for storing and processing data whose market value is 10 dollar or which will not bring good amount to company.

According to Gartner glossary, the definition of Big Data is "high volume, velocity and variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making".

III. APPLICATIONS OF BIG DATA

A. Banking and Securities

The Securities Exchange Commission (SEC) is using big data to monitor financial market activity. They are currently using network analytics and natural language processors to catch illegal trading activity in the financial markets.

Retail traders, Big banks, hedge funds and others in the financial markets use big data for trade analytics used in high frequency trading, pre-trade decision-support analytics, sentiment measurement, Predictive Analytics etc.

B. Communications, Media and Entertainment

Organizations in this industry simultaneously analyze customer data along with behavioral data to create detailed customer profiles that can be used to:

- Create content for different target audiences
- Recommend content on demand
- Measure content performance
- A case in point is the Wimbledon Championships (YouTube Video) that leverages big data to deliver detailed sentiment analysis on the tennis matches to TV, mobile, and web users in real-time.
- Spotify, an on-demand music service, uses Hadoop big data analytics, to collect data from its millions of users worldwide and then uses the analyzed data to give informed music recommendations to individual users.
- Amazon Prime, which is driven to provide a great customer experience by offering, video, music and Kindle books in a one-stop shop also heavily utilizes big data.

C. Healthcare Providers

Some hospitals, like Beth Israel, are using data collected from a cell phone app, from millions of patients, to allow doctors to use evidence-based medicine as opposed to administering several medical/lab tests to all patients who go to the hospital. A battery of tests can be efficient but they can also be expensive and usually ineffective.

Free public health data and Google Maps have been used by the University of Florida to create visual data that allows for faster identification and efficient analysis of healthcare information, used in tracking the spread of chronic disease.

D. Education

Big data is used quite significantly in higher education. For example, The University of Tasmania. An Australian university with over 26000 students has deployed a Learning and Management System that tracks among other things, when a student logs onto the system, how much time is spent on different pages in the system, as well as the overall progress of a student over time.

E. Manufacturing and Natural Resources

In the natural resources industry, big data allows for predictive modeling to support decision making that has been utilized to ingest and integrate large amounts of data from geospatial data, graphical data, text and temporal data. Areas of interest where this has been used include; seismic interpretation and reservoir characterization.

F. Government

In public services, big data has a very wide range of applications including: energy exploration, financial market analysis, fraud detection, health related research and environmental protection.

Some more specific examples are as follows:

Big data is being used in the analysis of large amounts of social disability claims, made to the Social Security Administration (SSA), that arrive in the form of unstructured data. The analytics are used to process medical information rapidly and efficiently for faster decision making and to detect suspicious or fraudulent claims.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 07, July - 2017

The Food and Drug Administration (FDA) is using big data to detect and study patterns of food-related illnesses and diseases. This allows for faster response which has led to faster treatment and less death.

The Department of Homeland Security uses big data for several different use cases. Big data is analyzed from different government agencies and is used to protect the country.

G. Insurance

Big data has been used in the industry to provide customer insights for transparent and simpler products, by analyzing and predicting customer behavior through data derived from social media, GPS-enabled devices and CCTV footage. The big data also allows for better customer retention from insurance companies.

When it comes to claims management, predictive analytics from big data has been used to offer faster service since massive amounts of data can be analyzed especially in the underwriting stage. Fraud detection has also been enhanced.

Through massive data from digital channels and social media, real-time monitoring of claims throughout the claims cycle has been used to provide insights.

Big Data Providers in this industry include: Sprint, Qualcomm, Octo Telematics, The Climate Corp.

H. Retail and Whole sale trade

Big data from customer loyalty data, POS, store inventory, local demographics data continues to be gathered by retail and wholesale stores.

In New York's Big Show retail trade conference in 2014, companies like Microsoft, Cisco and IBM pitched the need for the retail industry to utilize big data for analytics and for other uses including:

- Optimized staffing through data from shopping patterns, local events, and so on
- Reduced fraud
- Timely analysis of inventory

Social media use also has a lot of potential use and continues to be slowly but surely adopted especially by brick and mortar stores. Social media is used for customer prospecting, customer retention, promotion of products, and more.

I. Transportation

Some applications of big data by governments, private organizations and individuals include:

- Governments use of big data: traffic control, route planning, intelligent transport systems, congestion management (by predicting traffic conditions)
- Private sector use of big data in transport: revenue management, technological enhancements, logistics and for competitive advantage (by consolidating shipments and optimizing freight movement)
- Individual use of big data includes: route planning to save on fuel and time, for travel arrangements in tourism etc.

J. Energy and Utilities

- Smart meter readers allow data to be collected almost every 15 minutes as opposed to once a day with the old meter readers. This granular data is being used to analyze consumption of utilities better which allows for improved customer feedback and better control of utilities use.
- In utility companies the use of big data also allows for better asset and workforce management which is useful for recognizing errors and correcting them as soon as possible before complete failure is experienced.

IV. PLATFORMS TO WORK WITH BIG DATA

There are several big data platforms available with different characteristics and choosing the right platform requires an in-depth knowledge about the capabilities of all these platforms. Especially, the ability of the platform to adapt to increased data processing demands plays a critical role in deciding if it is appropriate to build the analytics based solutions on a particular platform.

Typically, when the user has to decide the right platforms to choose from, he/she will have to investigate what their application/algorithm needs are. One will come across a few fundamental issues in their mind before making the right decisions.

- How quickly do we need to get the results?
- How big is the data to be processed?
- Does the model building require several iterations or a single iteration?

Clearly, these concerns are application/algorithm dependent that one needs to address before analyzing the systems/platform-level requirements. At the systems level, one has to meticulously look into the following concerns:

- Will there be a need for more data processing capability in the future?
- Is the rate of data transfer critical for this application?
- Is there a need for handling hardware failures within the application?

This section primarily aims to provide simple analysis of these concerns and provide a score for each of the big data platforms with respect to these issues.

Scaling is the ability of the system to adapt to increased demands in terms of data processing. To support big data

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 07, July - 2017

processing, different platforms incorporate scaling in different forms. From a broader perspective, the big data platforms can be categorized into the following two types of scaling:

- Horizontal Scaling: Horizontal scaling involves distributing the workload across many servers which may be even commodity machines. It is also known as "scale out", where multiple independent machines are added together in order to improve the processing capability. Typically, multiple instances of the operating system are running on separate machines.

- Vertical Scaling: Vertical Scaling involves installing more processors, more memory and faster hardware, typically, within a single server. It is also known as "scale up" and it usually involves a single instance of an operating system.

A. *Horizontal scaling platforms*

Some of the prominent horizontal scales out platforms include peer-to-peer networks and Apache Hadoop. Recently, researchers have also been working on developing the next generation of horizontal scale out tools such as Spark [2] to overcome the limitations of other platforms.

Peer-to-Peer networks involve millions of machines connected in a network. It is a decentralized and distributed network architecture where the nodes in the networks (known as peers) serve as well as consume resources. It is one of the oldest distributed computing platforms in existence. Typically, Message Passing Interface (MPI) is the communication scheme used in such a setup to communicate and exchange the data between peers. Each node can store the data instances and the scale out is practically unlimited (can be millions of nodes). The major bottleneck in such a setup arises in the communication between different nodes. Broadcasting messages in a peer-to-peer network is cheaper but the aggregation of data/results is much more expensive.

Apache Hadoop is an open source framework for storing and processing large datasets using clusters of commodity hardware. Hadoop is designed to scale up to hundreds and even thousands of nodes and is also highly fault tolerant. The Hadoop platform contains the following two important components:

- Distributed File System (HDFS) is a distributed file system that is used to store data across cluster of commodity machines while providing high availability and fault tolerance.

- Hadoop YARN is a resource management layer and schedules the jobs across the cluster.

- The programming model used in Hadoop is MapReduce which was proposed by Dean and Ghemawat at Google. MapReduce is the basic data processing scheme used in Hadoop which includes breaking the entire task into two parts, known as mappers and reducers. At a high-level, mappers read the data from HDFS, process it and generate some intermediate results to the reducers. Reducers are used to aggregate the intermediate results to generate the final output which is again written to HDFS. A typical Hadoop job involves running several mappers and reducers across different nodes in the cluster. One of the major drawbacks of MapReduce is its inefficiency in running iterative algorithms. MapReduce is not designed for iterative processes.

Spark is a next generation paradigm for big data processing developed by researchers at the University of California at Berkeley. It is an alternative to Hadoop which is designed to overcome the disk I/O limitations and improve the performance of earlier systems. The major feature of Spark that makes it unique is its ability to perform in-memory computations. It allows the data to be cached in memory, thus eliminating the Hadoop's disk overhead limitation for iterative tasks. Spark is a general engine for large-scale data processing that supports Java, Scala and Python and for certain tasks it is tested to be up to 100× faster than Hadoop MapReduce when the data can fit in the memory and up to 10X faster when data resides on the disk. It can run on Hadoop Yarn manager and can read data from HDFS. This makes it extremely versatile to run on different systems.

B. *Vertical scaling platforms*

The most popular vertical scale up paradigms are High Performance Computing Clusters (HPC), Multicore processors, Graphics Processing Unit (GPU) and Field Programmable Gate Arrays (FPGA).

HPC clusters [27], also called as blades or supercomputers, are machines with thousands of cores. They can have a different variety of disk organization, cache, communication mechanism etc. depending upon the user requirement. These systems use well-built powerful hardware which is optimized for speed and throughput. Because of the top quality high-end hardware, fault tolerance in such systems is not problematic since hardware failures are extremely rare.

Multicore refers to one machine having dozens of processing cores [28]. They usually have shared memory but only one disk. Over the past few years, CPUs have gained internal parallelism. More recently, the number of cores per chip and the number of operations that a core can perform has increased significantly. Newer breeds of motherboards allow multiple CPUs within a single machine thereby increasing the parallelism.

Graphics Processing Unit (GPUs) is a specialized hardware designed to accelerate the creation of images in a frame buffer intended for display output [30]. Until the past few years, GPUs were primarily used for graphical operations such as video and image editing, accelerating graphics-related processing etc. However, due to their massively parallel architecture, recent developments in GPU hardware and related programming frameworks have given rise to GPGPU (general-purpose computing on graphics processing units) [31]. GPU has large number of processing

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 07, July - 2017

cores (typically around 2500+ to date) as compared to a multicore CPU.

FPGAs are highly specialized hardware units which are custom-built for specific applications [34]. FPGAs can be highly optimized for speed and can be orders of magnitude faster compared to other platforms for certain applications. They are programmed using Hardware descriptive language (HDL) [35]. Due to customized hardware, the development cost is typically much higher compared to other platforms. On the software side, coding has to be done in HDL with a low-level knowledge of the hardware which increases the algorithm development cost.

V. CHALLENGES

A. The practical issues of storing all the data

Most organizations' data is growing at a rate of 40 to 60 percent per year. Simply storing the data is becoming a real challenge. Companies are looking at options like data lakes, which will allow them to collect and store massive quantities of unstructured data in its native format. The problem is, data lakes have to be constructed wisely or they quickly become a useless wasteland where data goes to never be retrieved again.

B. Getting & Keeping the Talent Necessary to Make Use of Big Data Analytics

It takes a full set of hard and soft skills in order to be a successful data scientist. While the debate over the shortage of IT workers rages on, in the realm of data science, the shortage is a proven reality. One option for businesses at this stage is to develop their own data professionals in house. But this can be expensive and the results will often fall short. The other option is to work with an organization that specializes in big data. That way, the people are allowed to specialize in whatever the company does, while letting the data people handle the data stuff.

In time, this situation will resolve itself, as colleges, universities, and technical institutions ramp up more educational programs to produce the data scientists, analysts, engineers, and other professionals that are needed. But that doesn't keep you competitive in the meantime.

C. Dealing with the Security Issues of Big Data

It is a reality that if there is an intention to collect, store, and use big data, then investment in adequate security needs to be done.

D. Handling the Various Sources of Data Available

Handling the volume of storage necessary and the velocity at which data is increasing is one thing. Managing enormous streams of data from various disparate sources, both inside and outside of the organization, is another matter entirely. When the enterprises' own data sources (like finance, operational, marketing, and other data) are combined with external sources (such as social media and industry data), it becomes truly diverse as well as exceptionally massive. There simply are not a lot of people out there who have learned how to build algorithms to successfully query these highly varied data sets and deliver useful meaning out of them.

E. The Time It Takes to Glean Value Out of the Data

Given the challenges listed above, it's easy to see how terribly time consuming big data analytics can be. It's difficult for many organizations to justify this amount of time and effort relative to the value it delivers. To put this into business speak, big data doesn't always deliver a powerful ROI. As data infrastructures, talent, and analytical tools and capabilities develop, this will eventually rectify.

VI. CONCLUSION

Here, we conclude that area of Big Data is arising as a pioneer because of it applications and uses in almost every major sector. To handle the challenges coming in the way of Big Data, data scientists need more attention and research as this area is going to be more deep, diverse and vast. In this paper we discussed some issue that revolve around Big Data and make it what it is. This is just a start of a new era of vast knowledge and applications based on Big Data.

REFERENCES

[1] S. Pandey, Dr. V. Tokekar, "Prominence of mapreduce in big data processing", Fourth International Conference on Communication Systems and Network Technologies, IEEE, 2014.

[2] Wei Fan, Albert Bifet, "Mining Big Data: Current Status, and Forecast to the Future", SIGKDD Explorations 14(2), ACM, 2012. 5

[3] D. Agrawal, P. Bernstein, E. Bertino, S. Davidson, U. Dayal, M. Franklin, et al, "Challenges and opportunities with big data–A community white paper developed by leading researchers across the United States," 2012.

[4] R. AKERKAR, C. BADICA, AND C. B. BURDESCU, "Desiderata for research in web intelligence, mining and semantics". In Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics (WIMS '12). ACM, New York, NY, USA,Article 0 , 5 pages. DOI=10.1145/2254129.2254131 http://doi.acm.org/10.1145/2254129.2254131.

[5] Gartner, http://www.gartner.com/it-glossary/bigdata.

[6] Global Research Data Infrastructures: Towards a 10-year vision for global research data infrastructures. Final Roadmap, March 2012. [Online].

Available: http://www.grdi2020.eu/Repository/FileScaricati/6bdc07fb-b21d-4b90-81d4-d909fdb96b87.pdf

[7] Riding the wave: How Europe can gain from the rising tide of scientific data. Final report of the High Level Expert Group on Scientific Data. October 2010. [Online]. Available at http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf

[8] Y.Demchenko, P.Membrey, P.Grosso, C. de Laat, "Addressing Big Data Issues in Scientific Data Infrastructure," in First International Symposium on Big Data and Data Analytics in Collaboration (BDDAC 2013). Part of The 2013 Int. Conf. on Collaboration Technologies and Systems (CTS 2013), May 20-24, 2013, San Diego, California, USA.

[9] NIST Big Data Working Group (NBD-WG). [Online]. Available: http://bigdatawg.nist.gov/home.php

[10] Definting Big Data Architetcure Framework: Outcome of the Brainstorming Session at the University of Amsterdam, 17 July 2013. Presented at NBD-WG, 24 July 2013 [Online]. Available: http://bigdatawg.nist.gov/_uploadfiles/M0055_v1_7606723276.pdf

[11] Yuri Demchenko, Cees de Laat, Peter Membrey , "Defining Architecture Components of the Big Data Ecosystem"

[12] Rise Of Big Data And Related Issues, Swati Sharma, IEEE INDICON 2015 1570175759

[13] Apache Mahout, http://en.wikipedia.org/wiki/Apache_Mahout

[14] Storm, https://storm.apache.org/

[15] Cassandra, http://en.wikipedia.org/wiki/Apache_Cassandra