DELINEATION OF THE DIMENSIONS OF BIG DATA, AND A REVIEW ON THE CHALLENGES, OPPORTUNITIES, AND TOOLS OF BIG DATA

Aseema Sultana[1], Farnaz Khatoon[2]
[1-2]*Dept. Of Information Science and Engineering, HKBK CE, Bangalore, India*
[1]aseemas.is@hkbk.edu.in, [2]farnazk.is@hkbk.edu.in

***Abstract*— "Big Data", that brings us a world full of benefits from smart cars that could help us avoid accidents, calling an ambulance if we happen to collide with a vehicle, an advertisement that focuses on our interested products from a previous search, to implantable devices that can monitor our health every second and immediately notify our doctor if something goes wrong. Big data analytics provide a universe of such possibilities and much more. There is more to the term Big Data. It comes in so many dimensions, be it its incoming velocity, its volume, variety, veracity or variability. Although we reap so many benefits from it, it could also lead to many issues including but not restricted to- privacy concerns. This paper discusses the dimensions of big data, challenges that come with big data, its opportunities and a few big data tools.**

***Keywords*— Big Data, Volume, Velocity, Variety, Veracity, Variability of Big Data, Challenges and Opportunities of big data, Big Data Tools**

I. INTRODUCTION

In 1992, WalMart was said to be the first enterprise to store a terabyte of data, as in [1]; and back then it was considered to be big data. However, now anyone can buy a USB drive with a storage capacity of a terabyte. So what is Big Data? And what distinguishes Big Data from very large data? The answer to this question is very simple; big data is infinitely flowing data that is so huge that just acquiring this data submerges one's ability to process it with traditional methods. For example, the weather indicators are capable of forecasting today's and tomorrow's weather, may be even a week's weather. But the forecast for next month or entire year might not be impressive enough since analysing such huge amount of data might not be effectively possible in real time. As pointed out by Eric Schmidt (Google CEO), every two days we are creating as much information as we did since the dawn of civilization till 2003, as mentioned in [2]. The ever increasing information size has changed the way we store and process data. The information just keeps exploding day by day. To be able to manage this data and extract useful information out of it is the primary task for data scientists' and analysts.

II. THE 5 DIMENSIONS OF BIG DATA

When the 5 V's combine and are expanding at an increasing rate, this is when we know that we have big data. The five main dimensions or characteristics of big data are: Volume, Velocity, Variety, Veracity and Variability.

A. VOLUME

Volume is probably the best known characteristic of big data; it refers to how big the data can be. When we talk about volume of big data, we refer to quantities of data that reach almost incomprehensible proportions. For Example, Facebook stores user photographs. Each user stores lot of photographs. This fact might not seem unusual, but when we really think over, Facebook has more users than the number of people in a country with maximum population in the world. Hence Facebook is storing hundreds of billions of images, which is a lot to handle.

B. VELOCITY

Velocity here refers to how fast the tsunami of data keeps flowing in. The data may be coming in so fast that by the time we count it, the count would have increased. In these situations, we could just throw up our hands and say that a mean, a standard deviation, and a random sample are impossible to calculate. Take the same example of Facebook; its users upload hundreds of millions of photos a day. Facebook has to accept a tsunami of images each day, process and store it somehow and should be able to retrieve it later as well.

Another example is tweets on twitter. Let's say we are watching our favourite reality show on TV and every weekend we vote for our favourite contestant on the show. Say we want to know how the rest of the world is feeling about our favourite contestant and all the other contestants right now. How is it possible to know? We know that everybody tweet about the popular show. One way is to subject the tweets to sentiment analysis. The feed of Twitter data is obviously big and keeps increasing each second. We can use stream algorithms, Reservoir sampling, and other algorithms to compute stats on the fly based on the data received to date, and then project the trends.

C. VARIETY

Variety is one of the biggest challenges of big data. When we talk about traditional databases, we deal with data in rows and columns. All data in databases is in these rows and columns. In previous section, we discussed Facebook storing images, twitter handling tweets which are two different varieties of data. If we look into the world there and n varieties of such data; more examples include- sensor data, encrypted packets, video recordings, audio songs, web pages and blogs, email messages and many more to count. All these are coming from different sources and in different forms. Each one of these is very different from the other. And most of the data is unstructured. Which means it does not fit into our traditional rows and columns application. This data diversity is one of the key dimensions of big data.

D. VERACITY

One of the most important aspects to be considered in the incoming big data is the uncertainty or impreciseness of data. By definition, unstructured data contains a significant amount of uncertain and imprecise data. The level of uncertainty and imprecision varies on a case by case basis yet must be factored.

Veracity is all about making sure the data is accurate, which requires processes to keep the bad data from accumulating in our systems. The simplest example is contacts that enter your marketing automation system with false names and inaccurate contact information.

E. VARIABILITY

Not to be confused with the term variety, variability refers to data whose meaning constantly keeps changing. In other words it is data that is loaded into a system at an inconsistent speed. Consider as an example; a coffee shop may offer 6 different blends of coffee, but if you get the same blend every day and it tastes different every day; that is variability. The same is true of data; if the meaning is constantly changing it can have a huge impact on your data homogenization.

Variability is often confused with variety. Say you have bakery that sells 10 different breads. That is variety. Now imagine you go to that bakery three days in a row and every day you buy the same type of bread but each day it tastes and smells different. That is variability.

III. CHALLENGES ASSOCIATED WITH BIG DATA

There are a lot of hurdles in managing big data that probe as major challenges that need immediate attention. Few of them include but not limited to- the size of data, privacy and security, heterogeneity of data, incompleteness, data access and sharing, timeliness, analytical challenges, manpower, and technical challenges such as fault tolerance, scalability and quality of data as in [3], [4] and [5]. We will discuss a few of them in this paper.

A. SIZE

The size of big data is the biggest challenge and the first thing that comes into mind when we think about big data. The word 'Big' in big data itself describes it all. Big data has large size of data sets. It was easier to solve this challenge earlier with processors but as the data is becoming huge and huge day after day, the processors are no more able to handle this overwhelming data fast enough. Data storage is another aspect to be considered for this huge data and there is always need for a better storage system to be able to store and retrieve such large data.

B. INCOMPLETENESS AND HETEROGENEITY OF DATA

When we talk about incomplete data, what do we mean to have incomplete data? Incomplete data sets are created when a subject under scrutiny does not have information regarding one or more of the relevant variables. And if you're analysing incomplete data, your insights, if any, may only be half-correct. So, what can be done about the incomplete data? Two steps can be taken to ensure complete data is collected: one is to assess your source; meaning- Do not assume anything. Make sure the places from where you aggregate information are accredited and thorough. The second step is to use a reliable application. Purchasing a data collection and analysis program from a developer that focuses solely on constructing analytics solutions will ensure the solution has received the appropriate amount of attention.

C. PRIVACY AND SECURITY

This is the most sensitive issue in Big Data. There may be enormous benefits from big data analytics, but there is also a massive potential for exposure of data that could cause serious damage to individuals or organisations. People generate enormous amount of data every day- where they travel, where and what they eat, whom they communicate with, what they buy, where they exercise, which is their favourite coffee shop, where and when they go for swimming, how many hours they sleep and what not. By this they are more vulnerable to exposure in ways unimaginable. Such sensitive and detailed information in the hands of marketers, government, organizations, or even common people, can affect anything from having a job, boarding a plane to even getting loan. Proper action needs to be taken to improve security against the above said scenarios in this forever connected world. Few measures that individuals can take to lower privacy risks are- To quit sharing more information on social media, post images or any information only with selected people instead of doing it with entire public. Not to provide personal information to any businesses that is not necessary. And restricting others from sharing our information in public without our knowledge.

D. TIMELINESS

Timing is everything. Let’s imagine a scenario in which we have a system that has several inputs and several outputs and we are measuring them. In this rapidly evolving system, a second or even a few microseconds between one reading and another can mean one dataset is mismatched against another. In this example it’s essential to ensure that all our data is timely and being interpreted in real-time to ensure the best optimisation of the system.

Due to the rapid changes in big data, the timeliness of some data is very short. If companies can’t collect the required data in real time or deal with the data needs over a very long time, then they may obtain out-dated and invalid information. Processing and analysis based on these data will produce useless or misleading conclusions, eventually leading to decision-making mistakes by enterprises.

E. COLLABORATION

Collaboration is the key to making most of the big data. It is a skill we need to harness in order to take advantage of big data properly and leverage it to make researchers more productive: There needs to be interaction between skill sets – between data scientists with knowledge of processing and analysis. The ability to collaborate and communicate with others to solve problems is essential. Big data is community based, we must build groups to solve problems and use platforms to gather data and use that to innovate. There needs to be continual feedback from the users of Big Data to improve its potential.

IV. OPPORTUNITIES OF BIG DATA

Big data throws an ocean of opportunities in various fields. It is an umbrella which encompasses all sorts of data which exists today. It is not limited to only technology domains but is also extended to fields like healthcare, banking, government and many more as mentioned in [4].

A. HEALTHCARE

Big data takes information from multiple sources such as electronic patient record; clinical decision support system including medical imaging, physician's written notes and prescription, pharmacy and laboratories; clinical data; and machine generated sensor data to gain insights. The integration of clinical, individual health and behavioural data helps to develop a robust treatment system, which can reduce the cost and at the same time, improve the quality of treatment as in [6].

B. TELECOMMUNICATION

For any mobile service provider, the cost of getting a new customer is higher than retaining an existing customer. To retain an existing customer, the service providers try to improve customer experience. This is done by analysing a number of factors such as the customer's personal information (their gender, age, language preference etc.), customer preferences, and internet usage. And then offer them a personalized service. This is nothing but targeted marketing. The call patterns are identified and different call plans are offered to different customers. With the diffusion of Smartphones, based on analysis of real-time location and behavioural data, location-based services/context-based services can be offered to the customers when requested. This would increase the adoption of mobile services, as mentioned in [6].

C. MARKETING AND SALES

Big Data can serve as a key ingredient towards better marketing. It is not the data that leads to successful marketing, but the insights that we gain from the data that help in taking better decisions. Insights like who the customers are, their location, their needs, what factors influence customer loyalty, what are their likes and dislikes. All this can help marketers attract better customers.

D. BANKING

Using big data in banking industries is very advantageous. From fraud detection and prevention where big data allows banks to make sure that no unauthorised transactions are made, to enhanced compliance reporting where banks access customer needs and use big data to fulfil them. Big data gives banks deep insights into customer habits and patterns by tracing each customer's transactions, their commonly accessed services, their credit card usage and much more. This allows banks to target customers with relatable marketing and personalized product offerings as discussed in [7].

E. EDUCATION

Big data can be used to make the learning process of every student a better task. Individual students can be analysed for their learning progression. The information from devices that students use can be collected to check their learning skills and analyse their strong and weak points. We can provide them with customised learning that can provide clearer picture on the subjects the students learn.

F. MEDIA AND ENTERTAINMENT

Big Data is too big to ignore. It can be used as a massive force for boosting one's business. Big Data is helping media companies create new revenue sources. With smartphones and associated digital media becoming the major source of entertainment, now it is possible to reach out to a larger chunk of digitally connected audience. Big Data facilitates zeroing in on the right content that such audience will prefer. It helps advertisers and businesses pinpoint the exact preferences of customers. It also gives a better understanding about what type of content viewers watch at what time and duration by predicting the interest of the target audience. In a live streaming scenario, Big Data also helps advertisers to tweak their broadcasts real-time to deliver a far enriched and personalized media experience.

V. BIG DATA TOOLS

A. HADOOP

Hadoop is an open source programming framework that allows parallel or distributed processing of huge data sets. It is written in Java originally developed by Doug Cutting who named it after his son's toy elephant as in [8]. The term Hadoop refers to not only the base package but also th entire Hadoop ecosystem, as in [9]The Current Apache Hadoop ecosystem consists of the Hadoop Kernel, Mapreduce, HDFS and a numbers of various components like Apache Hive, HBase and Zookeeper, pig, Oozie and spark, as in [10]. Reference [11] explains that Hadoop runs code across a cluster of computers. Data is initially divided into directories and files. Files are divided into uniform sized blocks and then are distributed across various cluster nodes for further processing.

Blocks are replicated for handling hardware failure. Checking that the code was executed successfully, performing the sort that takes place between the map and reduce stages, sending the sorted data to a certain computer, writing the debugging logs for each job. Hadoop has two major layers namely: Processing/Computation layer (Map Reduce), and Storage layer (Hadoop Distributed File System).

1) *HDFS Architecture*: HDFS (Hadoop Distributed File system) is a highly fault tolerant parallel file system with a master/slave architecture whose objective includes storing and managing huge data sets, managing failures of hardware, data access and simplifying simple data coherency issues as in [12]. The HDFS architecture is as shown in Fig 1. The figure shows the HDFS master/slave architecture.

- *NameNode:* Each cluster consists of exactly one NameNode which is a master server that manages accesses to files w.r.t clients and also manages file system namespace. The responsibility of NameNode is to execute operations like Opening, Closing or renaming files.
- *DataNodes*: There can be n number of DataNodes depending upon the number of nodes in the cluster. The DataNodes manage storage of files. HDFS allows user data to be stored into files usually having the files itself split into more than one blocks and in turn these blocks stored in one or more DataNodes. The responsibility of DataNodes is to serve the read and write requests from the clients. Also, the DataNodes to perform block creation, cloning and deletion as per the instructions of the NameNode.

2) MapReduce: MapReduce is a software framework for distributed processing of vast amounts of data in a reliable, fault-tolerant manner. The two distinct phases of MapReduce are:

- Map Phase: In Map phase, the workload is divided into smaller sub-workloads. The tasks are assigned to Mapper, which processes each unit block of data to produce a sorted list of (key, value) pairs. This list, which is the output of mapper, is passed to the next phase. This process is known as shuffling.
- Reduce: In Reduce phase, the input is analyzed and merged to produce the final output which is written to the HDFS in the cluster.

B. SPARK

Apache spark is an Open-Source data analytics cluster computing framework that belongs to Hadoop Open Source community. It is built on top of Hadoop Distributed File System. Keeping the limitations of Hadoop into picture, Spark performs better than Hadoop in case certain specific applications. Spark is not restricted by two stage Map-Reduce paradigm. It provides the capabilities for in-memory cluster computing which allows user programs to load data into cluster's memory. The data loaded into main memory can be used repeatedly by subsequent database accesses and it speeds up the entire response time as in [9]. Spark is more useful in applications where jobs are iterative and analytics are interactive.

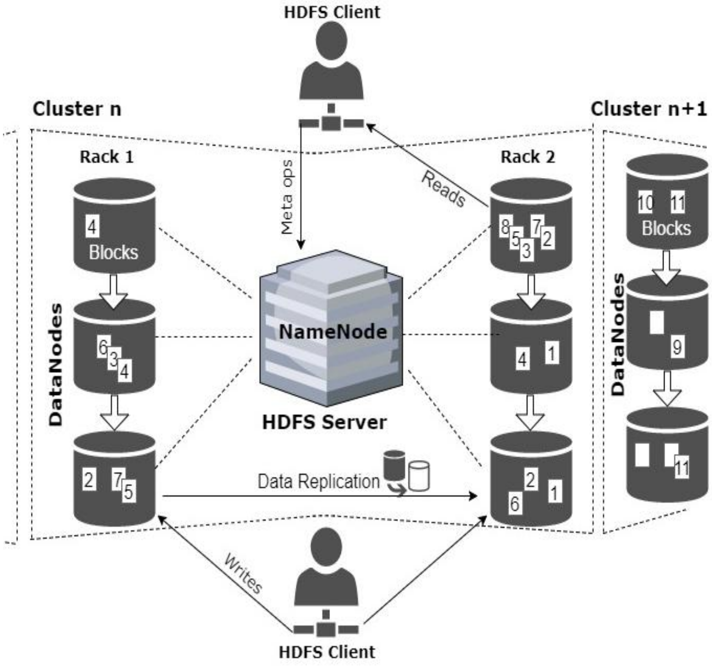

Fig 1: The HDFS Architecture

1) *The Resilient Distributed Datasets (RDD):* RDD is a read-only collection of objects partitioned across a set of machines that can be rebuilt if a partition is lost. The elements of an RDD need not exist in physical storage; instead, a handle to an RDD contains enough information to compute the RDD starting from data in reliable storage. This means that RDDs can always be reconstructed if nodes fail. As elements of the RDD need not exist in physical storage, and only a handle is enough for reconstructing it, the iterative and or interactive jobs will execute faster than that in Hadoop.

Spark lets programmers construct RDDs in four ways: 1- From the shared file system, for example HDFS by parallelizing collection like an array. 2-

By dividing the array into number of slices. These slices can be sent to multiple nodes. 3- By transforming an existing RDD: A dataset with elements of type A can be transformed into a dataset with elements of type B using an operation called flatMap. 4- By Changing the Persistence of an existing RDD as mentioned in [9].

Though Spark is compatible with Hadoop and HDFS in many ways of its usage, it addresses certain applications in a specific manner and outperforms Hadoop. Spark provides three simple data abstractions for programming clusters: resilient distributed datasets (RDDs), and two restricted types of shared variables: broadcast variables and accumulators. The RDDs are Scala objects whose references can be stored in memory and the object can be recomputed from these references. This property of Spark makes it able to address the latency involved in iterative or machine learning algorithms, as in [9].

VI. CONCLUSIONS

Big Data is a data whose diversity and complexity require new architecture, techniques, algorithms, and analytics to manage it. Today, Data is generated from various sources in different forms and can arrive in the system at various rates. To process these large amounts of data is the biggest challenge today. In this paper we discussed what big data is, we highlighted its dimensions and saw how veracity and variability are equally important, we saw challenges associated with big data, and the opportunities that come with big data. This paper also gives a review on the tools for Big data. Hadoop is the core platform for structuring Big Data, and solves the problem of making it useful for analytics purposes. We also discussed HDFS architecture and MapRedure which are used to support the processing of large data sets in distributed computing environments. Then we discussed few basics on spark. In future we can use some clustering techniques and check the performance of the tools.

ACKNOWLEDGMENT

We thank Dr. Syed Mustafa, Professor and Head, Dept of Information Science and Engineering, HKBK CE for motivating us and providing their expertise that greatly assisted in writing the paper.

We are also grateful to the authors of papers and documents we have referenced. We would like to express our appreciation to the authors for sharing their pearls of wisdom.

REFERENCES

[1] Big Data or Infinite Data? By Dave Snell, PREDICTIVE ANALYTICS AND FUTURISM, Dec 2015.

[2] Spark: the Next-generation Processing Engine for Big Data, By Dihui Lai and Richard Xu, PREDICTIVE ANALYTICS AND FUTURISM, Dec 2015.

[3] Research Paper on Big Data and Hadoop, Iqbaldeep Kaur, Navneet Kaur, Amandeep Ummat, Jaspreet Kaur, Navjot Kaur; IJCST Vol. 7, Issue 4, Oct- Dec 2016.

[4] Big Data And Hadoop: A Review Paper, Rahul Beakta; RIEECE Volume 2, Spl. Issue 2 (2015).

[5] A Review Paper on Big Data and Hadoop, Harshawardhan S. Bhosale, Prof. Devendra P. Gadekar; International Journal of Scientific and Research Publications, Volume 4, Issue 10, October 2014.

[6] Big Data: Challenges, Opportunities, and Realities , Abhay Kumar Bhadani and Dhanya Jothimani; Indian Institute of Technology Delhi, India.

[7] http://bigdata-madesimple.com/role-big-data-banking-industry/

[8] A survey on Big Data concepts and Tools; D.Rajasekar, C Dhanamani, S.K. Sandhya; IJETAE.

[9] A Case Study Comparing Different Big-Data Handling Approaches Using Hadoop-Hive VS Spark-Shark, Aparna Shikhare, Swapna Kulkarni.

[10] A REVIEW: MAPREDUCE AND SPARK FOR BIG DATA ANALYTICS, Meenakshi Sharma, Vaishali Chauhan, Keshav Kishore; 5th International Conference on Recent Innovations in Science, Engineering and Management Venkateshwara Group of Institutions, Meerut (U.P.), India (ICRISEM-16).

[11] Big Data Analysis using Hadoop: A Survey; Rotsnarani Sethy and Mrutyunjaya Panda; International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 7, July 2015.

[12] The HDFS Architecture guide, https://hadoop.apache.org/.