A Review on Bigdata Technology

Ms.Raveena Muhatte[1], Mr.Nadeem Inamdar[2], Prof Hari Prasad Mal [3]
[1, 2, 3] SRTTC-VIT Kamshet Campus

Abstract-*The recent evolvement of technologies has led to an increase in the production of data in different forms. This data can be a source from user generated data, healthcare industries, Internet and financial companies, etc.[1]. Big data commonly refers to huge or massive bulk of data which is mostly in unstructured format. Big data is a term for data sets that are so large or complex that traditional data processing application software is inadequate to deal with them. The term big data is mostly used to simplify the use of predictive analytics, user behavior, or certain other advanced data analytics methods that pull out valuable data from huge data sets and which can be useful for the same[2]. Data sets grow rapidly in part because they are increasingly gathered by a huge number of sources which includes information-sensing devices, software logs, cameras, microphones, etc.[2]. In this paper we present a literature survey on big data and then discuss the basic terminologies that are used related to big data.*

Keywords-Big data analytics, data sets, Hadoop

I. INTRODUCTION

It is understood that we are living a technological era, which is proved by the massive volume of data from a variety of sources and its growing rate. The term ‘‘big-data’’ was evolved to capture the intense meaning of this data-explosion trend and indeed the data has been defined as the new way, which is expected to transform our society[1].

For example Next Generation Sequencing (NGS) is a classic big data application that deals with the dual challenge of vast amounts of raw heterogeneous. It requires workflow tools that are strong enough to process massive amounts of raw NGS data yet flexible enough to keep up with quickly changing research techniques. It also requires a way to meaningfully integrate data from Novartis with data from these large external organizations — such as 1000 Genomes, NIH’s GTEx (Genotype-Tissue Expression), and TCGA (The Cancer Genome Atlas) — paying particular attention to clinical, phenotypical, experimental, and other associated data[3].

The Novartis team chose Hadoop and Apache Spark to build a workflow system that allows them to integrate, process, and analyze diverse data for Next Generation Sequencing (NGS) research, while being responsive to advances in the scientific literature[3].

Big data is high volume, high velocity and high variety information need which require new forms of processing take decisions and also make predictions.

II. CHARACTERISTICS

Big data can be described by the following characteristics -

1. Volume: The quantity of the produced data is defined as its volume. The size of the data determines the value of deep knowledge and whether it can actually be considered as big data or not.

2. Variety: The type and nature of the data. This defines the variability in nature in the data produced through a particular source.

3. Velocity: In this characteristic, the speed at which the data is generated and processed so that it meets the demands and challenges for the growth and development of that particular system.

4. Variability: Inconsistency of the data set can obstruct the reliability and the capability to manage and handle it.

5. Veracity:The quality of gained data from different sources can differ in nature greatly and thus will affect the analysis process.

Data must be processed with various meaningful and accurate algorithms and methods for best outcomes. For example, to manage a factory one must consider both visible and invisible issues with various components. These methods or algorithms must be able to detect and address different issues such as machine degradation, component wear, etc. on the factory floor.[2]

III. BIG DATA ARTITECHTURE

Key architecture layers are the following:
File Systems- The big data uses distributed file systems which provide storage, fault tolerance, scalability, reliability, and availability[4].

Data Stores– Evolution of application databases with application specific databases instead of one size fits all[4].

Resource Managers– it provides the management of resource capabilities & support schedulers for more utilization and throughput[4].

Coordination– systems manage state, distributed coordination, consensus and lock management[4].

Computational Frameworks– a lot of work is happening at this layer with highly specialized compute frameworks for Streaming, Graph processing[4].

Data Analytics –Analytical tools & libraries, that support descriptive, predictive and machine learning[4].

Data Integration– these include the planning of tools for managing pipelines but also management of metadata[4].

Operational Frameworks – it provide scalable frameworks for monitoring[4].

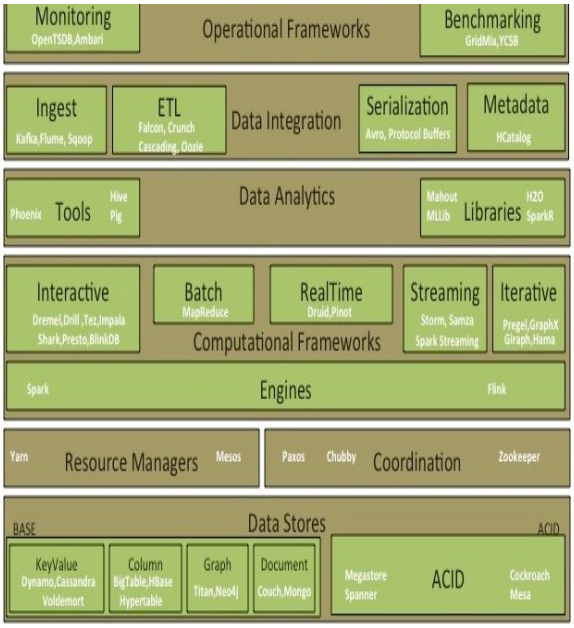

Fig. Big data architecture

IV. BIG DATA SYSTEM CHALLENGES

1. Privacy and security :

Information posted by users on their online profiles is likely to be used in creating a "users profile" so that can be further used by companies to develop their marketing strategies and to extend their services. Individual's privacy is still a delicate problem that can only be solved with drastic solutions. Allowing persons to choose whether they post or not information about them is a more secure way to achieve privacy, but will also cause software to "malfunction"[5].

2. Data access and sharing information :

Everything is available only if everyone shares it. Regarding persons, there is a difference between what is personal and what can be made public. The issue of what is personal and what is public mostly resides in the point of view of the services that they use. Regarding companies, this is a challenge that most refuse to overcome. The reason that companies don't want to share their own "Big Data" warehouse is more related to competitiveness and sensitive data that they have[5].

3. Human resource and manpower :

The skill of a data analyst must not be limited to the technical field. It should be expanded to research, analytical, interpretive and creative skills. Along with the organizations that train for data scientist the universities too must include education about big data and data analysis to produce skilled and expert employees[6].

4. Infrastructure faults :

A hardware system can only be reliable over a certain period of time. Intensive use and, rarely, production faults will most certainly result in a system malfunction. The challenge is to maintain the level of services that they provide[5].

V. TECHNOLOGIES

1. Hadoop:

Hadoop is an open source project hosted by Apache Software Foundation. Hadoop cluster uses a Master/Slave structure. Using Hadoop, large data sets can be processed across a cluster of servers and applications can be run on systems with thousands of nodes involving thousands of terabytes. It consists of many small sub projects. Hadoop consists of

- File System (The Hadoop File System)
- Programming paradigm (Map Reduce)

2. Hive :

Hive is a "SQL-like" bridge that allows conventional BI applications to run queries against a Hadoop cluster. It was developed originally by Facebook, but has been made open

source for some time now, and it's a higher-level abstraction of the Hadoop framework that allows anyone to make queries against data stored in a Hadoop cluster just as if they were manipulating a conventional data store. It amplifies the reach of Hadoop, making it more familiar for BI users[7].

3. PLATFORA

Perhaps the greatest limitation of Hadoop is that it is a very low-level implementation of Map Reduce, requiring extensive developer knowledge to operate. Between preparing, testing and running jobs, a full cycle can take hours, eliminating the interactivity that users enjoyed with conventional databases. PLATFORA is a platform that turns user's queries into Hadoop jobs automatically, thus creating an abstraction layer that anyone can exploit to simplify and organize datasets stored in Hadoop[7].

VI. FUTURE SCOPE AND DEVELOPMENT

As far as the future of big data is concerned it is sure that data volumes will continue to grow and the main reason for this is the drastic increase in the use of handheld devices, internet and other , which is expected that it will grow exponentially. SQL will remain as the standard for data. Tools for analysis without the presence of an analyst are set to take over. As per IDC half of all business analytics software will include intelligence where it is needed by 2020.

In other words it can be said that prescriptive analytics will be built into business software. Machine learning will have a far bigger role to play for data preparation and predictive analysis in businesses in the coming days. Privacy and security challenges related to big data will grow and by 2018, 50% of business ethics violations will be related to data. Autonomous agents and things like robots, autonomous vehicles, virtual personal assistant and smart devices will be a huge trend in the future. The International Institute for Analytics predicts that companies will use recruiting and internal training to budding data scientists to get their own problems done. More companies will try to derive their revenue from their data. The gap between insight and action in big data is going to reduce and more energy will be given to obtaining insights and execution rather than collecting big data.[6].

REFERENCES

[1] Q. Yan, J. Han, Y. Li, J. Zhou, and R.H. Deng, “Designing Leakage Resilient Password Entry onTouchscreen Mobile Devices”.
[2] https://en.wikipedia.org/wiki/Big_data
[3] https://mapr.com/blog/5-big-data-production-examples-healthcare/
[4] https://www.linkedin.com/pulse/100-open-source-big-data-architecture-papers-anil-madan
[5] http://www.dbjournal.ro/archive/13/13_4.pdf
[6] http://www.iaeme.com/MasterAdmin/uploadfolder/IJCET_07_04_002-2/IJCET_07_04_002-2.pdf
[7] http://www.techrepublic.com/blog/big-data-analytics/10-emerging-technologies-for-big-data/