Big Data Analytics Issues and Challenges: A Survey

Dr D. Lakshmi Sreenivasa Reddy
Associate Professor,
Chaitanya Bharathi Institute of Technology
Hyderabad, India
drreddycsejntuh@cbit.ac.in

Dr Mudumbi Krishna Murthy
Professor, CSE
Adisankara Engineering College
Gudur, Andhra Pradesh
krishnamudimbi @gmail.com

Abstract: In the digital world data accumulation is increasing more and more. Billions of devices are connected worldwide already and expected to connect 50 billons of devices by the year 2020. All devices produce Peta bytes of information by transforming and sharing with light speed on optical and wireless networks. The fast growth of such large data facing numerous issues and challenges in big data analytics, as the rapid growth of variety of data, hardware platforms, software, speed and security. Utilizing the data in decision making is one of the big challenges. Scaling the platforms and frameworks are other important challenges. This paper surveys the characteristics of big data including its characteristics, issues and challenges, right choices of platforms and models depending on their computational requirements and challenges.

Key Words: *Big data Analytics, Hadoop, Big data platforms, Map Reduce, Machine Learning, Data Mining.*

*****

I. INTRODUCTION

Under the massive increment of data globally in various forms traditional data bases and datasets are not useful to dig business value. Whereas Big data includes structured, unstructured and semi structured data. In addition big data analytics brings about many opportunities to find business values, hidden patterns and in depth understanding of the business. High level industries are interested in high level of big data potentiality. Many government organizations are planning to establish research in big data analytics to gain benefits from big data. Google processes internet data of hundred petabytes, Face book processes 10 petabytes of data per month. Alibaba is another data processing company to process share market data. Seventy two hours length of videos is uploaded every minute to Youtube. The rapid growth of internet of things (IOT) accumulates the data by sensors every second. Now a day's cloud computing provides safe guarding of the data storage and it provides the access and controlling of the data from anywhere. The increasingly growth of data needs how to store and maintain such heterogeneous datasets with minimum requirements of hardware and software infrastructure. Considering the heterogeneous, scalable, real time, complex type of big data, we can "mine" the datasets effectively at different levels of granularity during the analysis and modeling so as to forecast its intrinsic properties and improve the decision making. Different sources of big data have given in Fig 1.

Figure 1: Sources of Big data.

In general big data defined as "datasets which could not be gathered, managed, and processed by general purpose computers within a required scope.

II. CHARACTERISTICS OF BIG DATA

According to Gartner the challenges and issues of big data describes 3V's model. i.e the increase of Volume ,Velocity and Variety. Volume means the generation and collection of masses of data. Velocity means the speed of data flowing; it demands the timeliness of collection and analyzes big data. Variety indicates the different types of data, which includes structured, semi-structured and unstructured data such as table, audio, webpage, video, text and combination of all these. According to IDC the most influential research organization describes the four characteristics of big data as 4Vs'. [1] With this the characteristics of big data summarized as Volume, Variety, Velocity, and the extra thing added to the previous definition of big data is Value (huge value but very low density). The 4Vs definition was widely familiar since it highlights the necessity of big data in exploring the enormous hidden values. The above definition says that business values can be discovered from datasets with massive scale, different types, and speedy generation. Jay Parikh, from Facebook, said that, “You could only own a bunch of data other than big data if you do not utilize the collected data.” [2] The four Vs’ concept shown in Fig.2

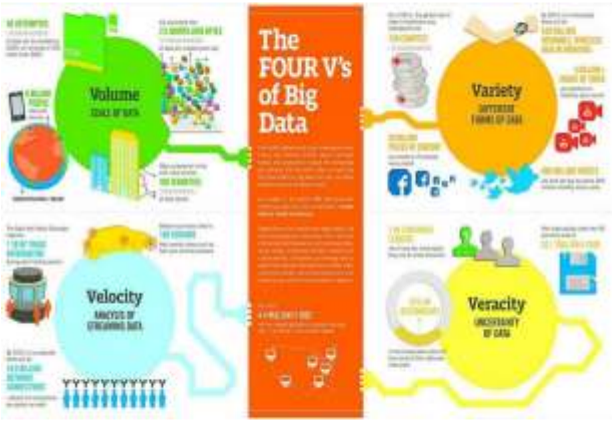

Figure 2: 4V’s of Big data.

As the importance of decision making and complexity of big data analysis, now these days many people extended the characteristics of big data from four to ten. According to William Vorhies, of Thousand Oaks, USA, The below are the ten possible characteristics.

a) Volume: Volume is a key contributor to the problem of why traditional relational database management systems (RDBMS, data warehouses as we know them today) fail to handle Big Data.

b) Variety: Variety describes different formats of data that do not lend themselves to storage in structured relational database systems. This includes a long list of data such as documents, emails, social media text messages, video, still images, audio, graphs, and the output from all types of machine-generated data from sensors, devices, RFID tags, machine logs, cell phone GPS signals, DNA analysis devices, and more. This type of data is characterized as unstructured or semi-structured and has existed all along.

c) Velocity: Meaning of Velocity is to describe data-in-motion, for example, the stream of readings taken from a sensor or the web log history of page visits and clicks by each visitor to a web site. This can be thought of as a fire hose of incoming data that needs to be captured, stored, and analyzed. Consistency and completeness of fast streams of data are one concern. Matching them to specific outcome events, a challenge raised under Variety is another. Velocity also incorporates the characteristics of timeliness or latency. i.e the data being captured at a rate or with a lag time that makes it useful. A second dimension of Velocity is how long the data will be valuable. Is it permanently valuable or does it rapidly age and lose its meaning and importance. The third dimension of Velocity is the speed with which it must be stored and retrieved. This is one of the major determinants of NoSQL storage, retrieval, analysis, and deployment architecture that companies must work through today. Yahoo, for example the ads that pop up have been selected specifically for you based on the capture, storage, and analysis of customers current web visit, prior web site visits, and a mash up of external data stored in a NoSQL DB like Hadoop and added to the analytics. Amazon or Netflix recommends the purchases or views based on the visits. The architecture of capture, analysis, and deployment must support real-time turnaround (in these case fractions and must do this consistently over thousands of new visitors each minute. Real Time Big Data Analytics (RTBDA) is one of the main frontiers of development in Big Data today.

d) Veracity: Necessary and sufficient data to test many different hypotheses, vast training samples for rich micro-scale model-building and model validation, micro-grained “truth” about every object in data collection is important.

e) Validity: Consistency in terms of availability or interval of reporting, accuracy, when data contains many extreme values is important. It presents a statistical problem to determine what to do with these ‘outlier’ values and whether they contain a new and important signal or are just noisy data. Data quality, governance, master data management (MDM) on massive, diverse, distributed, heterogeneous, “unclean” data collections needed to consider.

f) Value: The important V, characterizing the business value and potential of big data to transform your organization from top to bottom (including the bottom line) depends on this V

g) Variability: Dynamic, evolving, spatiotemporal data, time series, seasonal, and any other type of non-static behavior in your data sources, customers, objects of study, etc are under this category.

h) Venue: Central data bases are not sufficient for big data, distributed heterogeneous data from multiple platforms, from different owners’ systems, with different access and formatting requirements, private vs. public cloud are important considerations now.

i) Vocabulary: Schema, data models, semantics, ontologies, taxonomies, and other content- and context-based metadata that describe the data’s structure, syntax, content, and provenance is also very important characteristic.

j) Vagueness: There should not be any confusion over the meaning of big data, in selecting tools models and platforms.

III. CHALLENGES OF BIG DATA

Traditional database management systems are based on relational database management systems (RDBMS). However such RDBMS are useful only for structure data, not useful for any other type of data like semi structured non-structured data. The traditional RDBMS could not handle huge volume and heterogeneity of big data. Cloud computing meets some requirements in handling the huge data and infrastructure. For solutions of permanent storage and management of large-scale disordered datasets, distributed file systems [3] and No-SQL [4] databases are good choices. The key challenges [5-7] are discussed in the literature.

a) Data Representation:

Different datasets have different levels of data in structure, type, semantics, granularity, organization, and accessibility. The Aim of data representation is to convert data into more meaningful for further analysis. Otherwise the data representation will reduce the value of originality and may give obstruct to the data analysis. Well-organized data version can reflect the good data structure, type and class with integrated technologies to facilitate required operations from different databases.

b) Redundancy reduction:

Data redundancy is quite common in big data. Without reduction of data redundancy and compression of the data leads us to wrong analysis results and increase the cost of the entire system on the premise of the values of the data are not affected. For example, sensor networks data is highly redundant data, which can be cleaned and compressed at the commands of magnitude.

c) Data refreshment: With advances of storage systems, generating data at unprecedented rates and scales are possible now. There are lot of challenges in current storage system It could not support such massive data. Hidden values in big data depend on data freshness. Therefore, an important principle data analytics must be developed to decide which part of the data should store and which is not.

d) Data Analytics: The analytical system should process the huge amount of heterogeneous data within a fraction of time. But the traditional database management systems (RDBMS) are designed with less scalability and expandability, which could not meet the required performance. Databases related to non relational structures have shown their unique benefits in processing of the unstructured data and have become important part of big data analytics. Some of the people are using both non relational and RDBMS. More research is needed to overcome this challenge of in-memory and sample data based on approximate analysis. We will address the some of the remedies to overcome these issues in subsequent chapters.

e) Data confidentiality: most of the data owners could not effectively maintain the data due to their limited storage capacity and software's. They depend on experts or tools to process such data, which increases the risks of safety. If transactional dataset considered, it contains a set of complete operating data to drive important business values. This type of data sets contains details of the lowest low abstract level data like credit card numbers and personal information. Therefore, big data analysis may be delivered to cloud service providers for processing only due to cost and security when they provide proper preventive measures.

IV. ISSUES RELATED SCALABILITY

The ability of the system in terms of data analysis is scaling. Incorporating different platforms and frameworks support big data processing, which improves scaling. There are two types of categories in to improve scaling.

IV.I. Horizontal scaling: Multiple independent machines are added to each other and distribute the work load among them to improve the processing capability. Typically different operating systems are running on separate machines in parallel. It increases the performance with small modifications. Financial investment to upgrade is relatively less and improves the scaling as much as possible. The main problem in this horizontal scaling is software. Software in parallel processing has not evolved much. Limited number of software is available now to increase performance.

a) *Horizontal scaling Platforms:* Some of the topmost horizontal platforms are peer to peer networks and Apache Hadoop. There is another important platform Spark, people are using more now a days. It overcomes the limitations of other platforms.

i) *Peer-to-peer networks*: It is one of the oldest distributed nature platforms [8-9]. It contains millions of machines connected, Massage Passing Interface (MPI) is the communication schema used in these networks to exchange the data between peers. Each machine stores the number of instances.

ii) *Apache Hadoop:* Apache Hadoop [10] is one of the open source framework for storing and analyzing large data using clusters of hardware. Hadoop is a fault tolerant platform designed to improve data processing using thousands of Machines. The mechanism of a Hadoop Stack is given in Figure 3. The Hadoop platform consists below two important components:

Hadoop Distributed File System (HDFS) [11] is used to store datasets across the clusters of commodity nodes providing high availability and fault tolerance.

Hadoop YARN [12] is a resource management layer and schedules the jobs across the cluster.

iii) MapReduce: The programming model used in Hadoop is MapReduce [13] which was invented by Dean and Ghemawat, Google Corporation. It is the basic data processing method used in Hadoop which breaks the entire task into mappers and reducers. Mappers read the data from

HDFS, process it and generates some intermediate results to the reducers. Reducers are used to aggregate the intermediate results to generate the final output which is sends to HDFS.

Hadoop job involves running of several mappers and reducers among different machines in the cluster.

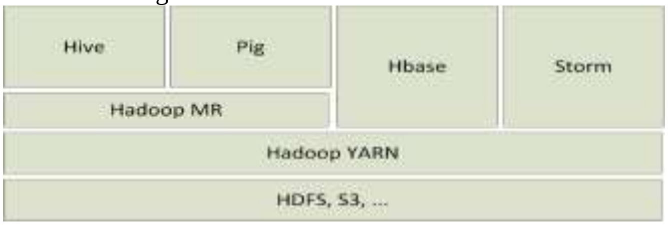

Figure 3. Hadoop Frame work.

iv) Spark: Spark is a next generation platform for big data analysis developed at University of California. It is an extension to Hadoop and designed to overcome the limitations of disk I/O and it improves the processing performance when compared to earlier systems. The major feature of Spark that makes it better is its capacity of performance of in-memory computations. It allows the datasets to be cached in memory, so that eliminating the Hadoop's disk overhead limitation for iterative tasks.

IV.II. Vertical Scaling: Vertical Scaling includes installing number of processors, more memory and faster hardware within a single server. It is also called as "scale up" and it usually uses a single operating system on single machine. Most of the software can easily take advantage of vertical scaling due to its single machine. It is Easy to manage and install hardware within a single machine. But it requires substantial amount to establish._System has to be more powerful to handle future workloads and initially the additional performance in not fully utilized. It is not possible to scale up vertically after a certain limit.

a)Vertical Scaling Platforms:

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

i) High performance computing (HPC) clusters: These are called 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.

ii) Graphics processing unit (GPU): 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 cores (typically around 2500+ to date) as compared to a multicore CPU.

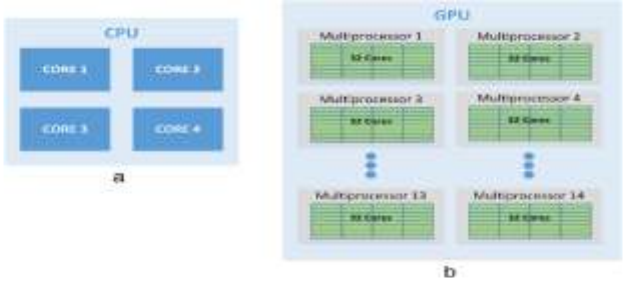

Fig.4: CPU and GPU Platforms

v) Conclusion and future directions

This paper surveys various data processing platforms that are currently available and discusses the advantages and drawbacks for each of them. Several details on each of these hardware platforms along with some of the popular software frameworks such as Hadoop and Spark are also provided.

References:

[1] Gantz J, Reinsel D (2011) Extracting value from chaos. IDC iView, pp 1–12

[2] Mayer-Sch¨onberger V, Cukier K (2013) Big data: a revolution that will transform how we live, work, and think. Eamon Dolan/Houghton Mifflin Harcourt

[3] Howard JH, Kazar ML, Menees SG, Nichols DA, Satyanarayanan M, Sidebotham RN, West MJ (1988) Scale and performance in a distributed file system. ACM Trans ComputSyst (TOCS) 6(1):51–81

[4] Cattell R (2011) Scalable SQL and NOSQL data stores. ACM SIGMOD Record 39(4):12–27

[5] Labrinidis A, Jagadish HV (2012) Challenges and opportunities with big data. Proc VLDB Endowment 5(12):2032–2033

[6] Chaudhuri S, Dayal U, Narasayya V (2011) An overview of business intelligence technology. Commun ACM 54(8): 88–98

[7] Agrawal D, Bernstein P, Bertino E, Davidson S, Dayal U, FranklinM, Gehrke J, Haas L, Halevy A, Han J et al (2012) Challenges and opportunities with big data. A community white paper developed by leading researches across the United States.

[8] Milojicic DS, Kalogeraki V, Lukose R, Nagaraja K, Pruyne J, Richard B, Rollins S, Xu Z (2002) Peer-to-peer computing, Technical Report HPL-2002-57, HP Labs.

[9] Steinmetz R, Wehrle K (2005) Peer-to-Peer Systems and Applications. Springer Berlin, Heidelberg.

[10] Hadoop. http://hadoop.apache.org

[11] Borthakur D (2008) HDFS architecture guide. HADOOP APACHE PROJECT. http://hadoop.apache.org/docs/r1.2.1/hdfs_design.pdf.

[12] Vavilapalli VK, Murthy AC, Douglas C, Agarwal S, Konar M, Evans R, Graves T, Lowe J, Shah H, Seth S (2013) Apache hadoop yarn: Yet another resource negotiator. In:

Proceedings of the 4th annual Symposium on Cloud Computing., p 5

[13] Dean J, Ghemawat S (2008) MapReduce: simplified data processing on large clusters. Commun ACM 51(1):107–113