A COMPREHENSIVE STUDY ON BIG DATA APPLICATIONS AND CHALLENGES

DDD Suribabu[1] and K Venkanna Naidu[2]

[1]Head&Assoc Professor, Dept. of Cse., DNR College of Engg& Tech.Bhimavaram
[2]Head&Assoc Professor, Dept. of Ece., DNR College of Engg& Tech.Bhimavaram

ABSTRACT

Big Data has gained much interest from the academia and the IT industry. In the digital and computing world, information is generated and collected at a rate that quickly exceeds the boundary range. As information is transferred and shared at light speed on optic fiber and wireless networks, the volume of data and the speed of market growth increase. Conversely, the fast growth rate of such large data generates copious challenges, such as the rapid growth of data, transfer speed, diverse data, and security. Even so, Big Data is still in its early stage, and the domain has not been reviewed in general. Hence, this study expansively surveys and classifies an assortment of attributes of Big Data, including its nature, definitions, rapid growth rate, volume, management, analysis, and security. This study also proposes a data life cycle that uses the technologies and terminologies of Big Data. Map/Reduce is a programming model for efficient distributed computing. It works well with semi-structured and unstructured data. A simple model but good for a lot of applications like Log processing and Web index building.

KEYWORDS

Big Data, HBase, Hadoop, MapReduce, Heterogeneity

1. INTRODUCTION

Big Data is proficient for business application and is rapidly escalating as a segment of IT industry. It has generated significant interest in various fields including the manufacture of health care machines, Banking transactions, Social media, Satellite imaging. Traditionally data is stored in a highly structured format to maximize its informational contents. Conversely, current data volumes are driven by both unstructured and semi structured data. Billions of individuals are various mobile devices and as a result of this technological revolution.

These people are generating tremendous amounts of data through the increased use of such devices. Particularly remote sensors continuously produce much heterogeneous data that are either structured or unstructured. This data also is termed as Big Data. Big Data is characterized by 3 aspects namely; (a) Numerous data (b) Categorization of data cannot be done in regular relational data bases (c) Processing of data can be generated and captured quickly. Consequently end to end processing can be obstructed by the translation between structured data in relational systems of DBMS and unstructured data for analytics. Big Data is a compilation of very huge

DOI: 10.5121/ijci.2016.5409

data sets with a great diversity of types so that it becomes complicated to process by using state of the art data processing approaches.

In general a data set can be called a big data if it is formidable to perform capture, visualization, analysis at current technologies. The Essential characteristics of Big Data are followed with five V's namely, Variety, Velocity, Volume, Virality, viscosity. Volume is described as the relative size of the data to the processing capability. Viscosity measures the resistance to flow in the volume of data. Here the resistance may come from different data sources. Virality is described as faster distribution of information across B2B networks (Business to Business).Variety depicts the spread of data types from machine to machine and adds new data types to traditional transactional data. Velocity is described as a frequency at which the data is generated, Shared and captured.

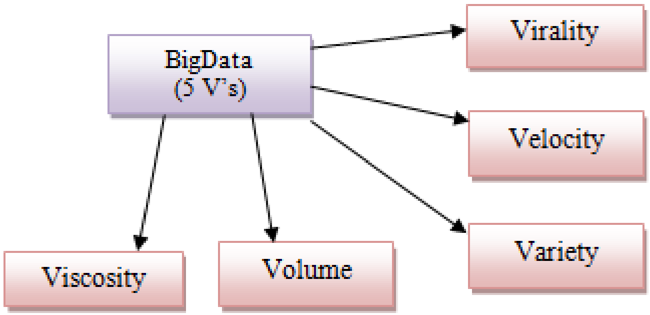

Fig 1: Essential characteristics of Big Data

2. BIG DATA ARCHITECTURE

The architecture of Big Data must be coordinated with the support infrastructure of the organization. Today, all of the data used by organizations are stagnant. Data is increasingly sourced from various fields that are disorganized and messy, such as information from machines or sensors and large sources of public and private data. Big Data technology improves performance, assists innovation in the products and services of business models, and provide decision making support. Big Data technology aim to lessen hardware and processing costs and to substantiate the value of Big Data before committing significant company resources. Properly managed Big Data are accessible, reliable, secure, and manageable. Hence, Big Data applications can be applied in various complex scientific disciplines (either single or interdisciplinary), including atmospheric science, astronomy, medicine, biology, genomics, and biogeochemistry.

2.1 Hadoop Ecosystem:

Hadoop is tranquil of HBase, HCatalog, Pig, Hive, Oozie, Zookeeper, and Kafka; however, the most common components and well-known paradigms are Hadoop Distributed File System (HDFS) and Map Reduce for Big Data. HDFS. This paradigm is applied when the amount of data is excessively much for a single machine. HBase is a management system that is open-source, versioned, and distributed based on the Big Table of Google. Zookeeper maintains, configures, and names large amounts of data. HCatalog manages HDFS. It stores metadata and generates tables for large amounts of data. Hive structures warehouses in HDFS and other input sources, such as Amazon S3. The Pig framework generates a high-level scripting language (Pig Latin) and operates a run-time platform that enables users to execute Map Reduce on Hadoop. Mahout is a library for machine-learning and datamining. Oozie. In the Hadoop system, Oozie coordinates, executes and manages job flow. Avro serializes data, conducts remote procedure calls, and passes data from one program or language to another. Chukwa is a framework for data collection and analysis that is related to Map Reduce and HDFS. Flume is particularly used to aggregate and transfer large amounts of data (i.e., log data) in and out of Hadoop.

2.2 Hadoop Usage:

Explicitly used by

(1) Searching Yahoo, Amazon, Zvents
(2) Log processing Facebook, Yahoo,ContexWeb.Joost, Last.fm
(3) Analysis of videos and images New York Times, Eyelike
(4) Data warehouse Facebook, AOL
(5) Recommendation systems Facebook

2.3 System Architectures of Map Reduce and HDFS:

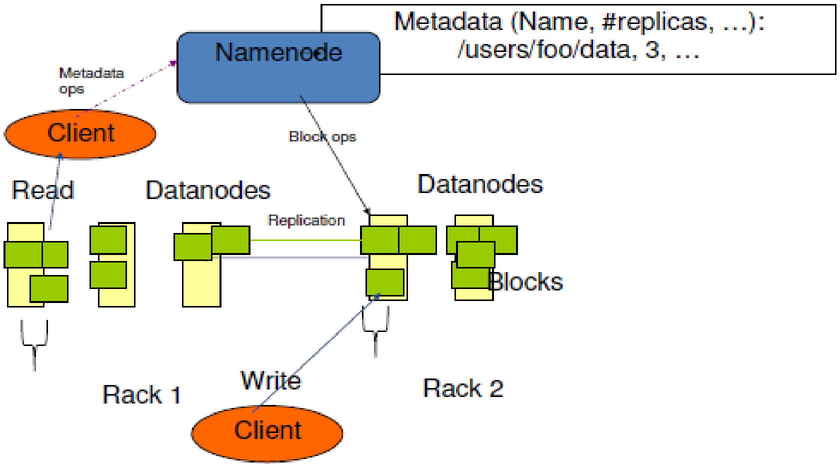

Fig 2: HDFS Architecture

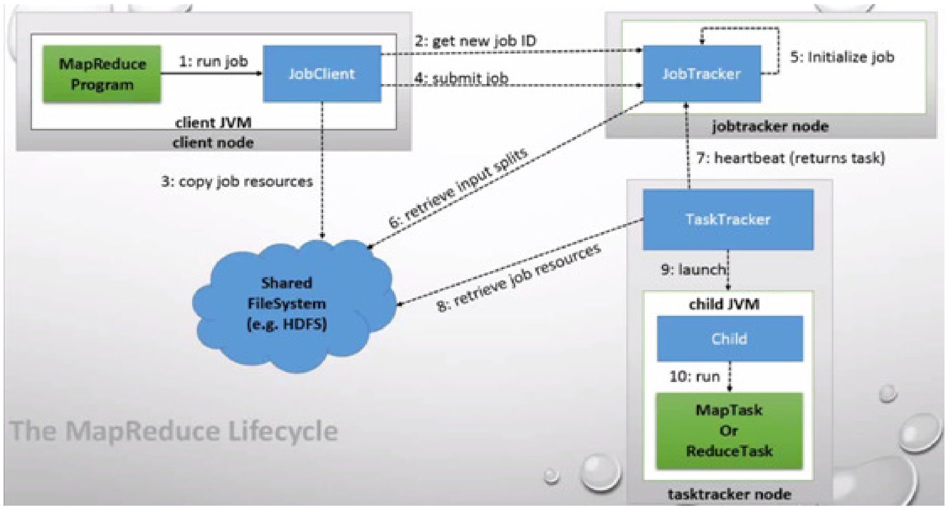

Fig 3: Map Reduce Life Cycle

Map Reduce is the hub of Hadoop and is a programming paradigm that enables mass scalability across numerous servers in a Hadoop cluster. In this cluster, each server encompasses a set of internal disk drives that are inexpensive. To improve performance, Map Reduce allocate workloads to the servers in which the processed data are stored. Data processing is scheduled based on the cluster nodes. A node may be consigned to a task that necessitates data foreign to that node.

2.3.1 MapReduce tasks:

(1) Input

(i) Data are loaded into HDFS in blocks and dispersed to data nodes
(ii) Blocks are replicated in case of failures
(iii) The name node tracks the blocks and data nodes

(2) Job Submits the job and its details to the JobTracker

(3) Job initialization

(i)The Job Tracker interacts with the TaskTracker on each data node
(ii) All tasks are scheduled

(4) Mapping

(i)The Mapper processes the data blocks
(ii) Key value pairs are listed

(5) Sorting

The Mapper sorts the list of key value pairs

(6) Shuffling

(i)The mapped output is conveyed to the Reducers
(ii) Values are rearranged in a sorted format

(7) Reduction Reducers merge the list of key value pairs to generate the final result

(8) Result
(i) Values are stored in HDFS
(ii) Results are replicated permitting to the configuration
(iii) Clients read the results from the HDFS

Map Reduce essentially communicate to two distinct jobs performed by Hadoop programs. The first is the map job, which involves procurement of a dataset and transforming it into another dataset. In these datasets, individual components are reviewed into tuples (key/value pairs). The reduction task receives inputs from map outputs and further divides the data tuples into small sets of tuples. Redundant data are stored in multiple areas across the cluster. The programming model tenacities failures inevitably by running portions of the program on a choice of servers in the cluster. Data can be distributed across a very large cluster of commodity components along with associated programming given the redundancy of data. This redundancy also endures faults and facilitates the Hadoop cluster to repair itself if the component of commodity hardware fails, especially given large amount of data. With this process, Hadoop can delegate workloads related to Big Data problems across large clusters of reasonable machines. The Map Reduce framework is convoluted, predominantly when complex transformational logic must be leveraged. Attempts have been generated by open source modules to simplify this framework, but these modules also use registered languages.

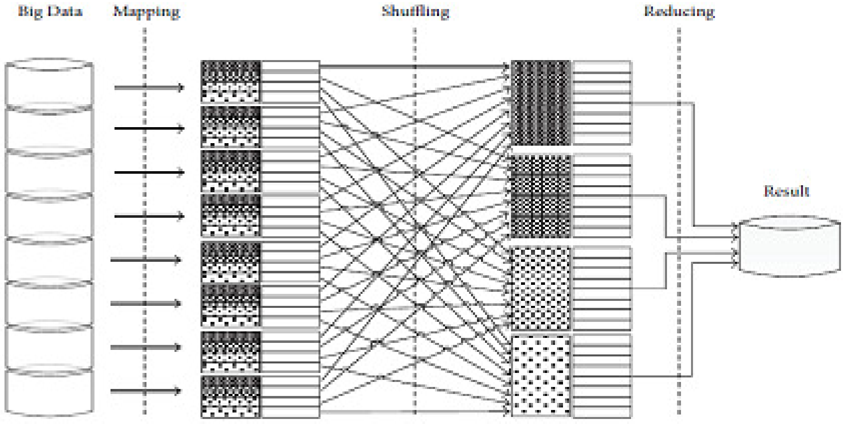

Fig 4: Map Reduce Architecture

2.3.2 Process flow

Record reader→Map→Combiner→Partitioner→Shuffle and sort→Reduce→Output format

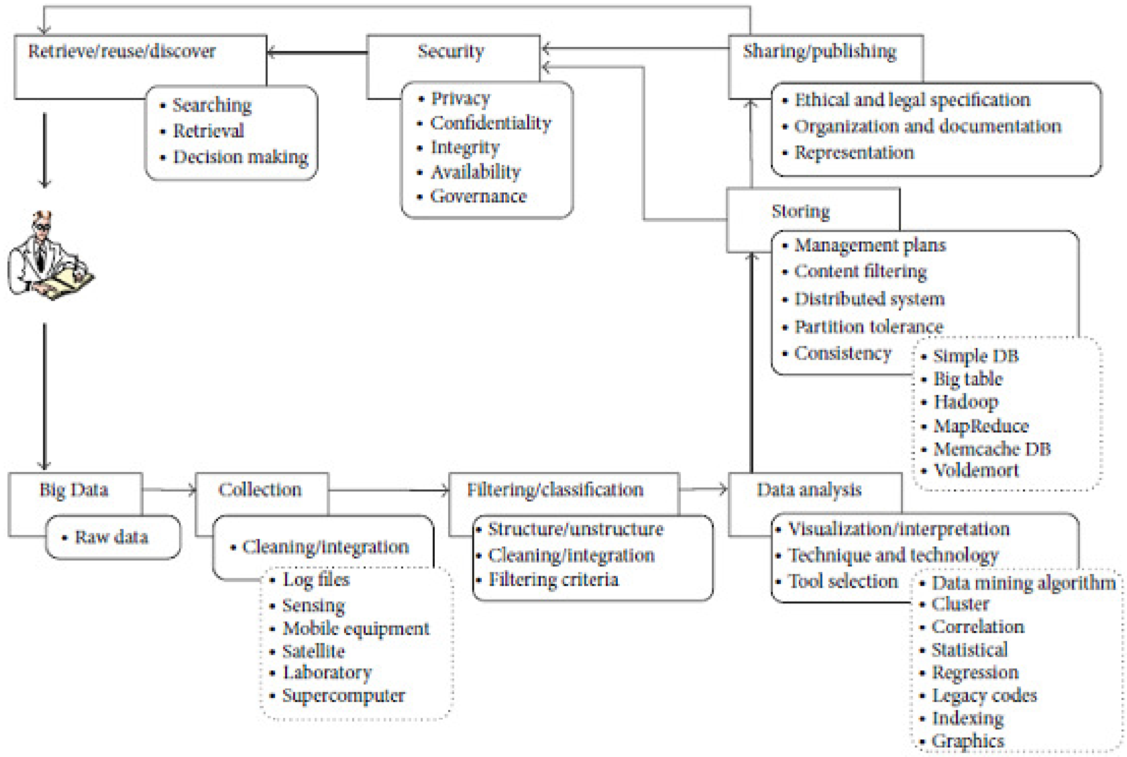

Figure 5: Proposed data life cycle using the technologies and terminologies of Big Data.

Raw Data. Researchers, agencies, and organizations integrate the collected raw data and increase their value through input from specific program offices and scientific research projects. The data are distorted from their initial state and are stored in a value-added state, including web services.

Collection/Filtering/Classification. Data collection or generation is generally the first stage of any data life cycle. Large amounts of data are created in the forms of log file data and data from sensors, mobile equipment, satellites, laboratories, supercomputers, searching entries, chat records, posts on Internet forums, and micro blog messages. Log files are utilized in nearly all digital equipment; that is, web servers note the number of visits, clicks, click rates, and other property registers the web users in log files.

Sensing. Sensors are often used to extent the physical quantities, which are then altered into comprehensible digital signals for processing and storage.

Mobile Equipment. The functions of mobile devices have under wired progressively as their usage rapidly increases. As the features of such devices are complicated and as means of data acquisition are enriched, various data types are produced.

Data Analysis. Data analysis enables an organization to handle abundant information that can affect the business. Data analysis has two main objectives: to apprehend the relationships among features and to develop effective methods of data mining that can perfectly foresee future observations. Big Data analysis can be applied to special types of data.

Data Mining Algorithms. In data mining, hidden but potentially valuable information is extracted from large, incomplete, fuzzy, and noisy data.

Cluster Analysis. Cluster analysis groups objects statistically according to certain rules and features. For example, objects in the same group are extremely heterogeneous, whereas those in another group are exceedingly homogeneous.

Correlation Analysis. Correlation analysis regulates the law of relations amongst practical phenomena, including mutual restriction, correlation, and correlative dependence.

Statistical Analysis. Statistical analysis is established on statistical theory, which is a division of applied mathematics.

Regression Analysis. Regression analysis is a mathematical technique that can reveal correlations between one variable and others.

Heterogeneity. Data mining algorithms trace unknown patterns and homogeneous formats for analysis in structured formats.

Scalability. Demanding issues in data analysis embrace the management and analysis of huge amounts of data and the rapid increase in the size of datasets.

Accuracy. Data analysis is typically buoyed by relatively accurate data obtained from structured databases with limited sources.

Storing/Sharing/Publishing. Data and its resources are composed and analyzed for storing, sharing, and publishing to benefit audiences, the public, tribal governments, academicians, researchers, scientific partners, federal agencies, and other stakeholders (e.g., industries, communities, and the media). Large and extensive Big Data datasets must be stored and managed with reliability, availability, and easy accessibility; storage infrastructures must provide reliable space and a strong access interface that can not only analyze large amounts of data, but also store, manage, and establish data with relational DBMS structures. Storage capacity must be reasonable given the sharp increase in data volume; for this reason, research on data storage is necessary.

Security. This stage of the data life cycle describes the security of data, governance bodies, organizations, and agendas. It also clarifies the roles in data stewardship. Therefore, appropriateness in terms of data type and use must be considered in developing data, systems, tools, policies, and procedures to defend legitimate privacy, confidentiality, and intellectual property.

Retrieve/Reuse/Discover. Data retrieval warrants data quality, value addition, and data preservation by reusing existing data to discover new and valuable information. This area is specifically involved in various subfields, including retrieval, management, authentication, archiving, preservation, and representation. The classical approach to structured data management is divided into two parts: one is a schema to store the dataset and the other is a relational database for data retrieval.

3. CHALLENGES

With Big Data, users not only face abundant attractive prospects but also come across challenges. Such complications lie in data capture, storage, searching, sharing, analysis, and visualization. Challenges in Big Data analysis include data inconsistency and incompleteness, scalability, timeliness, and security. Prior to data analysis, data must be well constructed. Understanding the method by which data can be preprocessed is important to improve data quality and the analysis results. Privacy is major concern in outsourced data. Recently, some arguments have discovered how some security agencies are using data generated by individuals for their own benefits without permission.

4. CONCLUSION

This paper presents the elementary concepts of Big Data. These concepts comprise the role of Big Data in the current environment of enterprise and technology. To augment the efficiency of data management, we have devised a data-lifecycle that uses the technologies and terminologies of BigData. The stages in this life cycle include collection, filtering, analysis, storage, publication, retrieval, and discovery. Data are also generated in different formats(unstructured and/or semi structured), which unfavorably affect data analysis, management, and storage. This deviation in data is accompanied by complexity and the development of further means of data acquisition. Big Data has developed such that it cannot be harnessed separately. Big Data is characterized by large systems, profits, and challenges. As a result, additional research is obligatory to address these issues and advance the efficient display, analysis, and storage of Big Data. To improve such research, capital investments, human resources, and pioneering ideas are the basic requirements.

REFERENCES

[1] C.L. Philip Chen, Chun-Yang Zhang, “Data-intensive applications, challenges, techniques and technologies: A survey on Big Data”, Information Sciences, www.elsevier.com/locate/ins, January 2014.

[2] Xindong Wu, Xingquan Zhu, Gong-Qing Wu, Wei Ding, “Data Mining with Big Data”, IEEE Transactions On Knowledge and Data Engineering, vol. 26, no. 1, January 2014,pp.97-107.

[3] “IBM What Is Big Data: Bring Big Data to the Enterprise,” http://www-01.ibm.com/software/data/bigdata/, IBM, 2012.

[4] Xindong Wu, Xingquan Zhu, Gong-Qing Wu, Wei Ding, “Data Mining with Big Data”, IEEE Transactions On Knowledge and Data Engineering, vol. 26, no. 1, January 2014,pp.97-107.

[5] unping Zhang, Fei-Yue Wang, Kunfeng Wang, Wei-Hua Lin, Xin Xu, Cheng Chen, Data-driven intelligent transportation systems: a survey, IEEE Trans. Intell. Trans. Syst. 12 (4) (2011) 1624–1639.

[6] D. Che, M. Safran, and Z. Peng, “From big data to big data mining: challenges, issues, and opportunities,” in Database Systems for Advanced Applications, B. Hong, X.Meng, L. Chen,W. Winiwarter, and W. Song, Eds., vol. 7827 of Lecture Notes in Computer Science, pp. 1–15, Springer, Berlin, Germany, 2013.

[7] Z. Sebepou and K. Magoutis, “Scalable storage support for data stream processing,” in Proceedings of the IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST ’10), pp. 1–6, Incline Village, Nev, USA, May 2010.

[8] A. Katal, M. Wazid, and R. H. Goudar, “Big data: issues, challenges, tools and good practices,” in Proceedings of the 6th International Conference on Contemporary Computing (IC3 '13), pp. 404–409, IEEE, 2013.

[9] A. Azzini and P. Ceravolo, “Consistent process mining over big data triple stores,” in Proceeding of the International Congress on Big Data (BigData '13), pp. 54–61, 2013.

[10] A. O’Driscoll, J. Daugelaite, and R. D. Sleator, “’Big data’, Hadoop and cloud computing in genomics,” Journal of Biomedical Informatics, vol. 46, no. 5, pp. 774–781, 2013.

[11] Y. Demchenko, P. Grosso, C. de Laat, and P. Membrey, “Addressing big data issues in scientific data infrastructure,” in Proceedings of the IEEE International Conference on Collaboration Technologies and Systems (CTS ’13), pp. 48–55, May 2013.

[12] Y. Demchenko, C. Ngo, and P. Membrey, “Architecture Framework and Components for the Big Data Ecosystem,” Journal of System and Network Engineering, pp. 1–31, 2013.

[13] M. Loukides, “What is data science? The future belongs to the companies and people that turn data into products,” AnOReilly Radar Report, 2010.

[14] A.Wahab,M. Helmy, H.Mohd, M. Norzali, H. F. Hanafi, and M. F. M. Mohsin, “Data pre-processing on web server logs for generalized association rules mining algorithm,” Proceedings of World Academy of Science: Engineering & Technology, pp. 48–53.