ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 10, October 2017

Big Data Concepts, Challenges and Solution in Hadoop Ecosystem

Dr. Ujjwal Agarwal
Lecturer (I.T.), Salalah College of Technology, Salalah, Oman

Abstract: **Data becomes big data when its volume, variety, and velocity exceed the abilities of our systems architecture and algorithm. This paper discusses about three major sources of big data: machine generated data, people generated data and organization generated data, 6V's of Big Data: volume, velocity, variety, valence, veracity and value along with we discussed the different variety of data: structured, semi-structured and un-structured data like sensor, images, PDF, CSV, JSON, RDMS, database, table data etc. out of which approximately 5% of available data is in structured form rest other data is in either un-structured or semi structured. Big data is facing lots of challenges due to volume, variety and other complexity in the data. Hadoop is the platform where we can find all our solution related to big data to store process and analysis purpose. The main objective of this paper to describe how Hadoop can solve different challenges of Big data by using HDFS (Hadoop distributed file System), Map Reduce and Hadoop Ecosystem components like Hive, Sqoop, HBase, Pig, spark, Flume, Kafka etc.**

Index Terms—Big Data, 6 V's, Structured data, un-structured data, Hadoop, HDFS, Hadoop Ecosystem

1. INTRODUCTION

It is hard to avoid mention of Big Data anywhere we turn today. There is broad recognition of the value of data, and products obtained through analyzing it [1]. Annual data production is increasing day by day. By 2020 rate of data generation will reach ten times greater than the data generated nowadays [4]. For example the size of a single sequenced human genome is approximately 200 gigabytes [6]. Big data is a growing term that describes any capacious amount of structured , semi structured and un- structured data that has the potential to be mined for information [9].

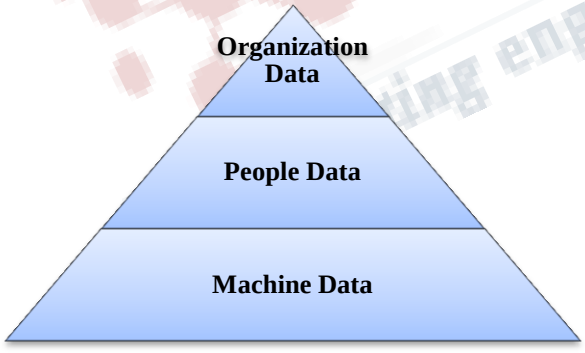


Figure 1.1 –Big Data coming from

1.1 Source of data: These data is coming from machine, people and organization.

1.1.1 Machine generated data: It is the major source of Big Data. With machine generated data, we refer to data generated from real time sensors in industry machinery or vehicles like car, aircraft. Data comes from various sensors, cameras, satellites, log files, bio informatics, activity tracker, personal health care track and many other sense data resources. [10]. Biggest source of big data, about 90% of data is generated by the machine, and there are many devices like Body temperature, Heart beat monitor, sleep monitor devices etc., they are generation huge of data every hour, every day. The widespread availability of the smart devices and their interconnectivity led to a new term being coined, The Internet of Things (IoT). Think of a world of smart devices at home, in your car, in the office, city, remote rural areas, the sky, even the ocean, all connected and all generating data.

1.1.2 People generated data: People are generating terabyte/petabyte amounts of data every day. As social media is very popular among every one. We are doing various activities on social media websites like Facebook, Twitter, LinkedIn, Instagram or online photo sharing sites like Flickr, or Picasa. In addition a huge amount of information gets generated via blogging and commenting, internet searches, more via text messages. Email, and through personal documents. Most of this data is unstructured, as there is no proper format or well defined structure is available. This data is very huge, this data can be in the .txt, .pdf, .csv, .json or it can be any format.

1.1.3 Organization generated data: This data is highly structured form of data. Organizations storing their data on some type of RDBMS like SQL, Oracle, and MS Access etc. This data is located in a fixed format inside the field or file/table. Traditionally, IT has managed and processed organization generated data in both operational and business intelligence system. Organization stores the data for current and future use as well as analysis of past.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 10, October 2017

1.2. 6V's of big data: Volume, velocity, variety, valence, veracity and Value.

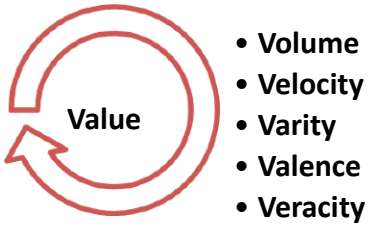
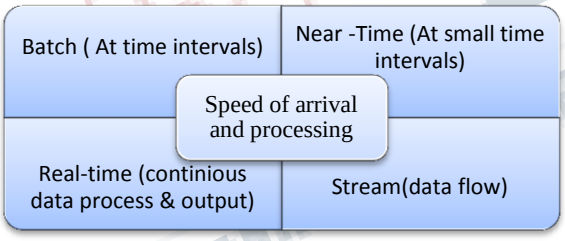


Figure: 1.2 6V's of Big Data

1.2.1 Volume: Volume is the big data aspect that relates to the total size of big data. This volume can come from large datasets being shared or many small data pieces and events being collected over time. Every minute 204 million emails are sent, 200,000 photos are uploaded, and 1.8 million likes are generated on Facebook. On YouTube, 1.3 million videos are viewed and 72 hours of video are uploaded. Large Synoptic Survey Telescope (LSST): ―Over 30 thousands gigabytes (30TB) of images will be generated every night during the decade –long LSST survey sky. [2]. this all generate a huge amount of data. Our traditional databases cannot manage this huge amount of data.

1.2.2 Velocity: Velocity refers to the increasing speed at which big data is created and the increasing speed at which the data needs to be stored and analyzed. Data may in the following way:-



2.3 Variety: There is variety of data is generated by different sources like Image data, text data, network data, geographic maps, computer generated simulations are only a few of the types of data we encounter every day, this variety of data may be structured or unstructured.

2.4 Veracity: Veracity of Big Data refers to the quality of the data, because big data can be noisy and uncertain. It can be full of biases, abnormalities and it can be imprecise. Data is of no value if it's not accurate, the results of big data analysis are only as good as the data being analyzed.

2.5 Valance: Simply put Valence refers to Connectedness, The more connected data is, the higher its valence. This can lead to complex molecules due to elements being interconnected through sharing electrons.

2.6 Value: Value is the total outcome after processing of big data. Does the data have value; if not is it worth being stored or collected? The analysis needs to be performed to meet the required purpose. The final output of all task is Value.

1.3.Types of data: Big data contains unstructured, semi structured or structured large amount of data [3]. This data is generated from different sources.

1.3.1 Structured Data: The data which can be co-related with the relationship keys, in a geeky word, RDBMS data! Maximum processing is happening on this type of data even today but then it constitutes around 5% of the total digital data.

1.3.2 Semi Structured Data: Semi structure data is not in a form of table or a specific structure but it can be converted to a defined structure. Examples of semi-structured: CSV, XML and JSON documents are semi structured documents, NoSQL databases are considered as semi structured. The most common form of semi structured data is .csv file, which can be easily converted to structured format.

1.3.3 Un-structured Data: All the left behind data having no structure at all, falls into this category and according to IDC estimate, it represents whopping 90% in share. Satellite images: This includes weather data or the data that the government captures in its satellite surveillance imagery. Just think about Google Earth, and you get the picture.
Scientific data: This includes seismic imagery, atmospheric data, and high energy physics. Photographs and video: This includes security, surveillance, and traffic video. Radar or sonar data: This includes vehicular, meteorological, and oceanographic seismic profiles. Text internal to your company: Think of all the text within documents, logs, survey results, and e-mails. Enterprise information actually represents a large percent of the text information in the world today. Social media data: This data is generated from the social media platforms such as YouTube, Facebook, Twitter, LinkedIn, and Flickr.
Mobile data: This includes data such as text messages and location information, chat information.
Website content: This is huge information which is available in the form of image, or video in YouTube, or text documents etc.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 10, October 2017

II CHALLENGES OF BIG DATA

Today, every minute, sees production of huge amounts of data. Every large company is stressed to find ways to make this data useful. However, this is not an easy task. The amount of data produced makes it very difficult to store, manage, analyze and utilize it. The handling of big data is very complex. Some challenges faced during its integration include uncertainty of data Management, big data talent gap, getting data into a big data structure, syncing across data sources, getting useful information out of the big data, volume, skill availability, solution cost, data storage and quality, security and privacy of data etc.

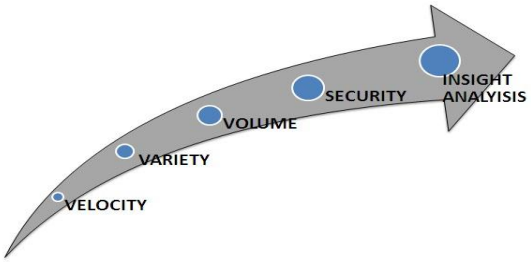


Figure 2: Challenges of Big Data

2.1 Velocity: If your organizations are generating new data at a rapid pace and needs to respond in real time, you have the velocity associated with big data. Most organizations that are involved in ecommerce, social media or IoT satisfy this criterion for big data.

2.2 Variety: If your data resides in many different formats, it has the variety associated with big data. For example, big data stores typically include email messages, word processing documents, images, video and presentations, as well as data that reside in structured relational database management systems (RDBMS).

2.3 Volume: Big data is any set of data that is so large that the organization that owns it faces challenges related to storing or processing it. In reality, trends like ecommerce, mobility, social media and the Internet of Things are generating so much information, that nearly every organization probably meets this criterion.

2.4 Security and privacy of the data: Once, companies and organizations figure out how to use big data, it gives them a varied range of opportunities. However, it also involves big risks when it comes to the security and the privacy of the data. The tools used for analysis, stores, manages, analyses, and utilizes the data from a different variety of sources. This ultimately leads to a risk of exposure of the data, making it highly vulnerable. Therefore, the production of more and more data increases security and privacy concerns. Thus making it essential for analysts and data scientists to consider these issues and deal with the data in a manner that will not lead to the disruption of privacy.

2.5 Insight Analysis: Challenges in Big Data analysis include data inconsistency and incompleteness, scalability, timeliness, and security [7, 8]. Prior to data analysis, data must be well constructed. However, considering the variety of datasets in Big Data, the efficient representation, access, and analysis of unstructured or semi-structured data are still challenging. Understanding the method by which data can be preprocessed is important to improve data quality and the analysis results. Datasets are often very large at several GB or more, and they originate from heterogeneous sources. Hence, current real-world databases are highly susceptible to inconsistent, incomplete, and noisy data. Therefore, numerous data preprocessing techniques, including data cleaning, integration, transformation, and reduction, should be applied to remove noise and correct inconsistencies. Each sub-process faces a different challenge with respect to data-driven applications. Thus, future research must address the remaining issues related to confidentiality. These issues include encrypting large amounts of data, reducing the computation power of encryption algorithms, and applying different encryption algorithms to heterogeneous data.

III HADOOP: SOLUTION FOR BIG DATA

Hadoop is a most major platform for storage, processing and analysis of big data. Hadoop is an apache open source program. Google had taken steps towards developing Hadoop through MapReduce concept. With the launching of MapReduce algorithm, Google has solved many problems regarding big data. MapReduce algorithm divides the task into small parts and assigns it to different nodes, and collects the result after processing. Using this solution by Google, Dough cutting and his team developed Hadoop platform. The first release of Hadoop was launched on 10th December 2011[5] and the first stable version (2.7.3) came into existence on 25th August 2016. It is used by many companies like Google, Facebook, Yahoo, YouTube, Twitter, LinkedIn etc. It simple uses the filesystem provided by Linux to store data. Hadoop has five daemons they are: NameNode, Secondary NameNode, DataNode, Job Tracker, and Task Tracker. Each daemons runs separately in its own JVM. Hadoop follows master-slave architecture means there is one master machine and multiple slave machine. The data you give to Hadoop is stored across these machines in the cluster. Two important components of Hadoop are: HDFS

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 10, October 2017

(Data Storage) and Map-Reduce (Analyzing and Processing).

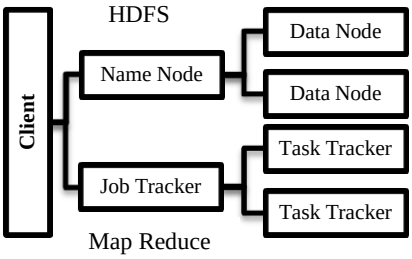


Figure: 3.1 Hadoop Architecture

3.1 HDFS: The first most important challenge is volume, because due to high volume our tradition data warehouses cannot manage all different set of huge data. Hadoop HDFS (Hadoop Database File System). HDFS is used to store very huge amount of data. HDFS follow master-slave architecture means there is one master machine called Name Node and multiple slave machine called Data Node. The data that we store in Hadoop is store in different clusters across the nodes.

HDFS is block structured file system in which individual file is split into several blocks of equal size and stored across one or more machine in a cluster. HDFS blocks are 64 MB by default in Apache Hadoop and 128 MB by default in Cloudera Hadoop but it can be increased as per need. If file size is 10 MB and HDFS block size is 128 MB then it takes only 10 MB of space.

Name Node: Name Node is the controller/master of the system. Name node spreads data to data node. It stores the metadata of all the files in the HDFS. This metadata includes name, location of each block, block size and file permission.

3.2 Hadoop Ecosystem to handle variety of data: The next most important challenge is variety of data. The Hadoop Ecosystem consists of tools for data analysis and moving large amounts of unstructured, semi structur0ed and structured data into HDFS. These tools are specialized to handle variety of data, like sqoop is used to inject database/tables in HDFS. Hive, Pig, HBase is used to manage un-structured and structured data.

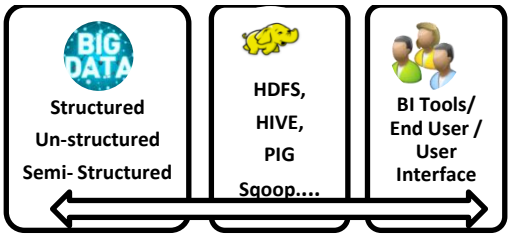


FIGURE 3.2: HADOOP PROCESS DATA MODEL

3.3 Security: The Hadoop ecosystem consists of various components. We need to secure all the other Hadoop ecosystem components. Hadoop community realized that more robust security controls were needed, and as a result, a team at Yahoo! decided to focus on authentication, and chose Kerberos as the authentication mechanism for Hadoop.

3.3.1 Mutual Authentication with Kerberos RPC (SASL/GSSAPI) on RPC Connections: SASL/ GSSAPI were used to implement Kerberos and mutually authenticate users, their processes, and Hadoop services on RPC connections.

3.3.2 Enforcement of HDFS file permissions: Access control to files in HDFS could be enforced by the NameNode based on file permissions - Access Control Lists (ACLs) of users and groups.

3.3.3 Network Encryption - Connections utilizing SASL can be configured to use a Quality of Protection (QoP) of confidential, enforcing encryption at the network level – this includes connections using Kerberos RPC and subsequent authentication using delegation tokens. Web consoles and MapReduce shuffle operations can be encrypted by configuring them to use SSL. Finally, HDFS File Transfer can also be configured for encryption.

3.3.4 Job Tokens to Enforce Task Authorization: Another security constrains can achieve by using Job tokens by the Job Tracker. Job tracker creates job tokens and transfers into TaskTrackers and ensures that the assigned job should be completed. The Task tracker responsible to complete the assign job. Tasks could also be configured to run as the user submitting the job, making access control checks simpler.

3.4 Insight Analysis: The main objective of Hadoop to analyze the data and find the hidden useful information. The big data is very huge in terms of volume, variety and velocity. HDFS provides us to store this huge volume of data. In Hadoop insight analysis is achieved by using MapReduce.

MapReduce breakdown big data in parallel for processing, it has two main steps first is Map and other is reduce. MapReduce can work on petabytes of data. The basic language of MapReduce is in Java but we van also work

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 10, October 2017

with Ruby, Python, R or Scala. Higher-level abstraction of Map Reduce is available. For example, a tool named Pig which is data flow language and translates them into MapReduce. Another tool, Hive, takes SQL queries and runs them using MapReduce for analysis of Big data.

3.5 Hadoop Ecosystem Project: The Hadoop is not single software rather it's a group of different application, which works together to form a Hadoop Ecosystem. It includes Apache open source projects and a wide range of commercial tools and solutions. Most of the solution or projects are to provide services for Hadoop's four core elements (HDFS, MapReduce, YARN, and Common). However, many other commercial projects or application provides much other functionality. The table below provides detail of some Hadoop ecosystem projects and their purpose:-

HADOOP PROJECT NAME AND PURPOSE

PROJECT NAME	PURPOSE
Hive	A data warehouse infrastructure that provides data summarization and ad hoc querying." It's a system that gives users the tools to make powerful queries and get results often in real-time
Pig	Pig is a platform with a high-level query language built to handle large data sets
HBase	Base is a non-relational database management system that runs on top of HDFS. It is built to handle sparse data sets common to big data projects
Mahout	Mahout is a Apache project used for clustering, classification and filtering using MapReduce.
HCatalog	It is table storage management tool. Used to write data into the grid, by using Pig and Mapreduce.
Impala	It's an open source, native analytic database for Apache Hadoop. Impala is shipped by Cloudera, MapR, Oracle, and Amazon.
Sqoop	Sqoop is used to import and export RDBMS Database or table into the Hadoop. It's a command line interface, we can import data into HDFS and similarly export data from HDFS to relation tables.
ZooKeeper	Zookeeper is a service for coordination among Hadoop ecosystem.
Spark	Spark engine is used for data analysis on large datasets. Sparks uses its own data processing techniques. Spark uses the language called SCALA. Spark is using for graph analysis, real time data processing, complex operation and machine learning.
Strom	Apache Storm is a free and open source distributed real-time computation system. Storm makes it easy to reliably process unbounded streams of data, doing for real-time processing what Hadoop did for batch processing.
Solr	Apache Solr is the open source platform for searches of data stored in HDFS in Hadoop. Solr powers the search and navigation features of many of the world's largest Internet sites, enabling powerful full-text search and near real-time indexing. Whether users search for tabular, text, geo-location or sensor data in Hadoop, they find it quickly with Apache Solr.
Flume	A distributed, reliable, and available service for efficiently collecting,

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 10, October 2017

	aggregating, and moving large amounts of streaming event data.
Kafka	A messaging broker that is often used in place of traditional brokers in the Hadoop environment because it is designed for higher throughput and provides replication and greater fault tolerance.
Cassandra	Cassandra is distributed database for managing large amount of data, this data is stores on may commodity servers.

IV CONCLUSION

This paper presents the fundamental concepts of Big Data, its challenges and Hadoop as a solution to solve all challenges. These concepts include explaining how data is generated and make data as big data. The increase in data due to three major sources: machine, people and organization, only organization data is in form of structured data, rest machine and people data is in form of semi-structured or un-structured form. Big data have many challenges because of volume, variety and velocity of data. Hadoop is giving solution to manage, store, process and analysis of this huge data. Hadoop ecosystem has variety of application which gives higher level of abstraction to manage and analysis of data. Like sqoop is used to ingest structured data or Database/table from RDBMS to Hadoop HDFS (Hadoop Database File System). Similarly hive and Pig used to store and analysis of semi structured or un-structured data.

V REFERENCES

[1] Big Data. Nature (http://www.nature.com/ news/specials/bigdata/index.html), Sep 2008.

[2] http://blogs.worldbank.org/voices/meet-winners-and-finalists-firstwbg-big-data-innovation-challenge

[3] Vivekananth.P, Leo John Baptist.A. An Analysis of Big Data Analytics Techniques. International Journal of Engineering and Management Research. October-2015,Volume-5, Issue-5

[4] Hirak Kashyap, Hasin Afzal Ahmed, Nazrul Hoque, Swarup Roy, and Dhruba Kumar Bhattacharyya. Big Data Analytics in Bioinformatics: A Machine Learning Perspective. JOURNAL OF LATEX CLASS FILES, Sepetember 2014, Vol. 13, NO. 9.

[5] Hadoop release. Apache.org Apache software foundation. Retrieved 2016-11-27.
[6] R. J. Robison, How big is the human genome? Precision Medicine, January 2014.

[7] A. Labrinidis and H. Jagadish, "Challenges and opportunities with big data," Proceedings of the VLDB Endowment, vol. 5, no. 12, pp. 2032–2033, 2012. View at Google Scholar.

[8] R. T. Kouzes, G. A. Anderson, S. T. Elbert, I. Gorton, and D. K. Gracio, "The changing paradigm of data-intensive computing," IEEE Computer, vol. 42, no. 1, pp. 26–34, 2009. View at Publisher • View at Google Scholar • View at Scopus.

[9]searchcloudcomputing.techtarget.com/definition/ big-data-Big-Data

[10] IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661,p-ISSN: 2278-8727 PP 01-05