Big Data Analytics Part 1(Unit 3)

MBI304- Data Mining & Data Analytics

Mamta Sagar

Department of Bioinformatics

University Institute of Engineering & Technology, CSJM University, Kanpur

Introduction

- Irrespective of the size of the enterprise whether it is big or small, data continues to be a precious and irreplaceable asset. Data is present in homogeneous sources as well as in heterogeneous sources. The need of the hour is to understand, manage, process, and take the data for analysis to draw valuable insights. Digital data can be structured, semi-structured or unstructured data.

1. Unstructured data:

This is the data which does not conform to a data model or is not in a form which can be used easily by a computer program. About 80% data of an organization is in this format; for example, memos, chat rooms, PowerPoint presentations, images, videos, letters. researches, white papers, body of an email, etc.

2. Semi-structured data:

- Semi-structured data is also referred to as self-describing structure.This is the data which does not conform to a data model but has some structure. However, it is not in a form which can be used easily by a computer program. About 10% data of an organization is in this format; for example, HTML, XML, JSON, email data etc.

Structured data:

- When data follows a pre-defined schema/structure we say it is structured data. This is the data which is in an organized form (e.g., in rows and columns) and be easily used by a computer program. Relationships exist between entities of data, such as classes and their objects. About 10% data of an organization is in this format. Data stored in databases is an example of structured data.

Big data

- The term "Big Data" refers to the heterogeneous mass of digital data produced by companies and individuals whose characteristics (large volume, different forms, speed of processing) require specific and increasingly sophisticated computer storage and analysis tools. This article

intends to define the concept of Big Data, its concepts, challenges and applications, as well as the importance of Big Data Analytics

- Being a complex polymorphic object, its definition varies according to the communities that are interested in it as a user or provider of services. Invented by the giants of the web, the Big Data presents itself as a solution designed to provide everyone a real-time access to giant databases. Big Data is a very difficult concept to define precisely, since the very notion of big in terms of volume of data varies from one area to another. It is not defined by a set of technologies, on the contrary, it defines a category of techniques and technologies. This is an emerging field, and as we seek to learn how to implement this new paradigm and harness the value, the definition is changing. [2]

Characteristics of Big Data

- 1) Characteristics of Big Data The term Big Data refers to gigantic larger datasets (volume); more diversified, including structured, semi- structured, and unstructured (variety) data, and arriving faster (velocity) than before. These are the 3V.

- Volume: represents the amount of data generated, stored and operated within the system. The increase in volume is explained by the increase in the amount of data generated and stored, but also by the need to exploit it.

- Variety: represents the multiplication of the types of data managed by an information system. This multiplication leads to a complexity of links and link types between these data. The variety also relates to the possible uses associated with a raw data.

- -Velocity: represents the frequency at which data is generated, captured, and shared. The data arrive by stream and must be analyzed in real time.

b) 5V:

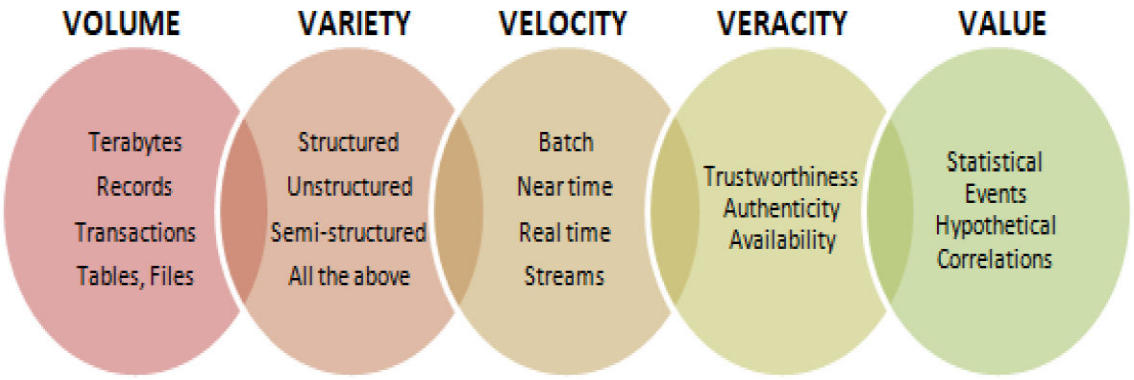


Fig. 2. *5V Concept*

- To this classical characterization, two other "V"s are important: - Veracity: level of quality, accuracy and uncertainty of data and data sources. -Value: the value and potential derived from data.

WHAT IS BIG DATA ANALYTICS ?

- Big Data generally refers to data that exceeds the typical storage, processing, and computing capacity of conventional databases and data analysis techniques. As a resource, Big Data requires tools and methods that can be applied to analyze and extract patterns from large-scale data. [analytics-Najafabadi et al.]

Big Data Anlytics refers to the process of collecting, organizing, analyzing large data sets to discover different patterns and other useful information. Big data analytics is a set of technologies and techniques that require new forms of integration to disclose large hidden values from large datasets that are different from the usual ones, more complex, and of a large enormous scale. It mainly focuses on solving new problems or old problems in better and effective ways. [4]

A. Types of Big Data Analytics

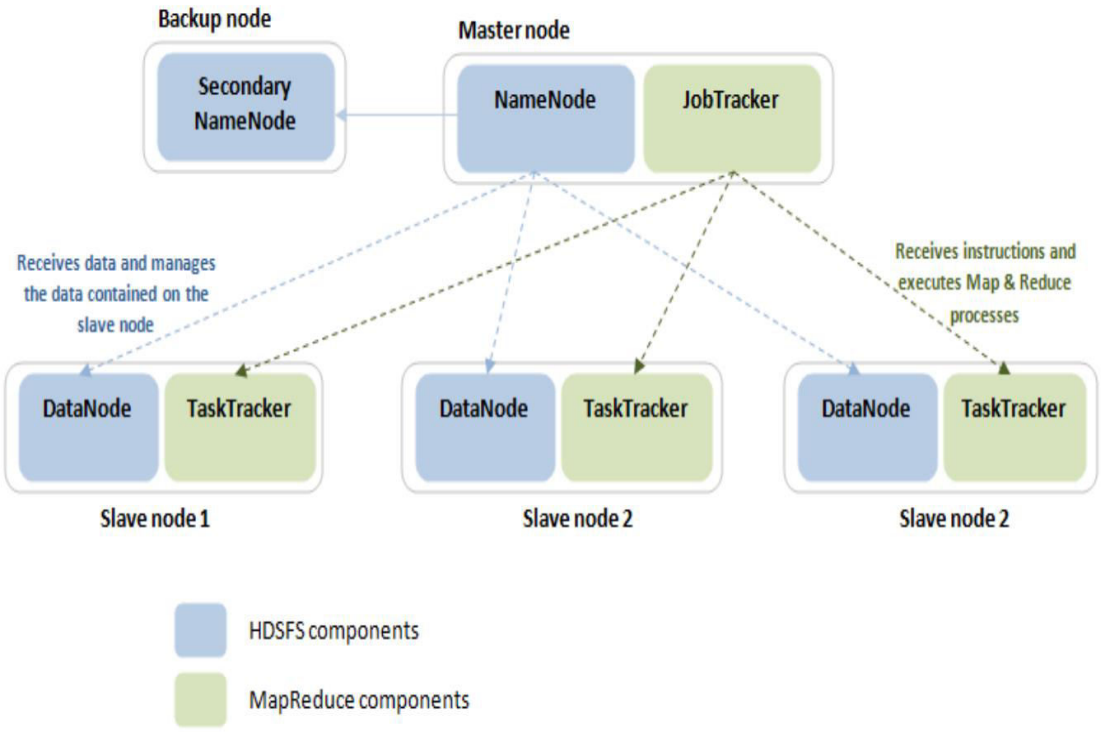
- a) Descriptive Analytics

- It consists of asking the question: What is happening? It is a preliminary stage of data processing that creates a set of historical data. Data mining methods organize data and help uncover patterns that offer insight. Descriptive analytics provides future probabilities and trends and gives an idea about what might happen in the future.

- b) Diagnostic Analytics It consists of asking the question: Why did it happen? Diagnostic analytics looks for the root cause of a problem. It is used to determine why something happened. This type attempts to find and understand the causes of events and behaviors.

- Predictive Analytics It consists of asking the question: What is likely to happen? It uses past data in order to predict the future. It is all about forecasting. Predictive analytics uses many techniques like data mining and artificial intelligence to analyze current data and make scenarios of what might happen.

- d) Prescriptive Analytics It consists of asking the question: What should be done? It is dedicated to finding the right action to be taken. Descriptive analytics provides a historical data, and predictive analytics helps forecast what might happen. Prescriptive analytics uses these parameters to find the best solution.



General Architecture

HADOOP FOR BIG DATA APPLICATIONS

• Big Data are collections of information that would have been considered gigantic, impossible to store and process, a decade ago. The processing of such large quantities of data imposes particular methods. A classic database management system is unable to process as much information. Hadoop is an open source software product (or, more accurately, „software library framework‟) that is collaboratively produced and freely distributed by the Apache Foundation – effectively, it is a developer‟s toolkit designed to simplify the building of Big Data solutions. [5]

Hadoop

• Hadoop is a distributed data processing and management system. It contains many components, including: HDFS, YARN,Map Reduce. HDFS is a distributed file system that provides high-performance access to data across Hadoop clusters. [6]

• MapReduce is a core component of the Apache Hadoop software framework. Hadoop enables resilient, distributed processing of massive unstructured data sets across commodity computer clusters, in which each node of the cluster includes its own storage. MapReduce serves two essential functions: It parcels out work to various nodes within the cluster or map, and it organizes and reduces the results from each node into a cohesive answer to a query.[7]

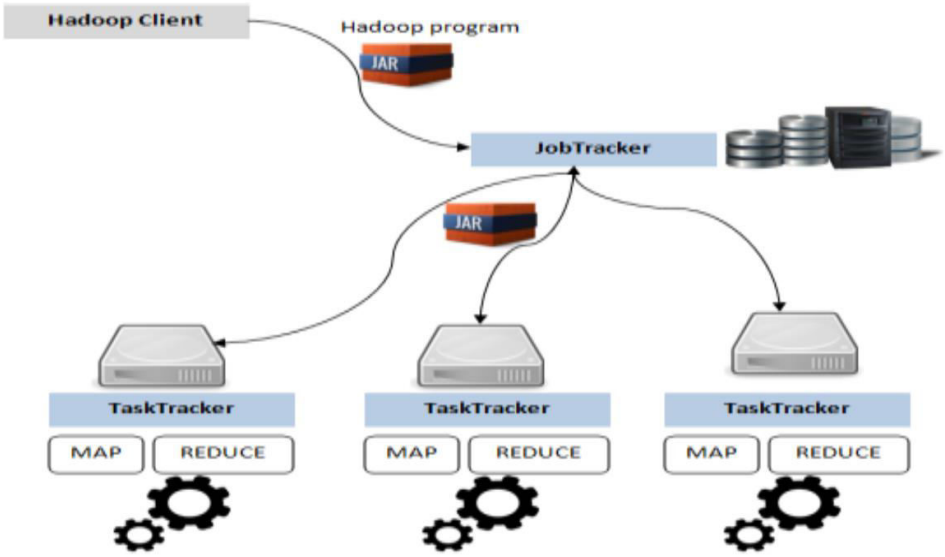


Fig. 3. Hadoop Architecture

- Hadoop relies on two servers: JobTracker: there is only one JobTracker per Hadoop cluster. It receives Map/Reduce tasks to run and organizes their execution on the cluster.When you submit your code to be executed on the Hadoop cluster, it is the JobTracker‟s responsibility to build an execution plan. This execution plan includes determining the nodes that contain data to operate on, arranging nodes to correspond with data, monitoring running tasks, and relaunching tasks if they fail.[8]

- •TaskTracker: several per cluster. Executes the Map/Reduce work itself (as a Map and Reduce task with the associated input data). The JobTracker server is in communication with HDFS; it knows where the Map/Reduce program input data is and where the output data must be stored. It can thus optimize the distribution of tasks according to the associated data.

- To run a Map/Reduce program, we must: •Write input data in HDFS

•Submit the program to the cluster's JobTracker. •Retrieve output

data from HDFS.

MAP REDUCE CONCEPT

- VMapReduce is a Java environment for writing programs intended for YARN. Java is not the simplest language for this, there are packages to import and class paths to provide. The data exchanged between Map and Reduce, in the entire job are pairs (key, value): -Key: it is any type of data: integer, text. . . -Value: it is any type of data. The two functions Map and Reduce receive and send such pairs.

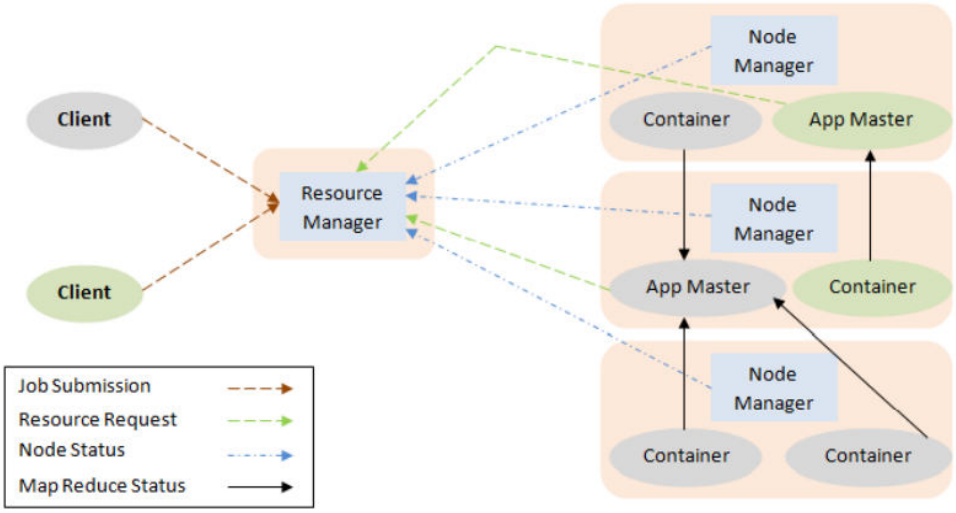


Fig. 6. General Architecture

- A. Map The Map function receives an input pair and can produce any number of pairs in output: none, one or more. The types of inputs and outputs are as desired. This very little constraint specification allows so many things. In general, the pairs Map are constituted as follows: • The text value is one of the rows or one of the n-tuples of the file to be processed • The key of type integer is the position of this line in the file.

- YARN launches a Map instance for each row of each file in the data to be processed. Each instance processes the row it has been assigned and produces output pairs.

- When designing a MapReduce treatment, we must think about the keys and values necessary so it can works. Reduce tasks receive a list of pairs with the same key and produce a pair that contains the expected result. This output pair can have the same key as the input.

C. Steps for a MapReduce job

- 1.Preprocessing of input data, eg: decompression of files

- 2.Split: Separate data into separately processable blocks and

formatted (key, value), eg in rows or tuples

- 3.Map: application of the map function on all the pairs (key, value) formed from the input data, this produces other pairs (key, value) output

- 4.Shuffle & Sort: redistribution of data so that the pairs produced by Map having the same keys are on the same machines

- 5. Reduce: Aggregation of pairs with the same key to get the final

result.

RM (Resource Manager):

- RM (Resource Manager): The central daemon of YARN. It manages resources assignments (CPU/Memory) when it comes to applications.
- It has two components: a scheduler which is in charge of resources allocation to the running application but it doesn‟t ensure restarting in case of task failure.
- The second component is the Application Manager which is in charge of App Masters management in the cluster. It ensures restarting of application masters on different nodes in case of failure.
- NM (Node Manager): The slave daemon of YARN. NM is responsible for containers monitoring their resource usage and reporting the same to the RM [10]. NM tracks the status of the node on which it is running. •AM (Application Master): There is only one application master per application. It negotiates resources from the RM and works with the NM. It manages the application life cycle. The AM acquires containers from the RM‟s scheduler before contacting the corresponding NMs to start the application‟s individual tasks. [11]

Healthcare

Healthcare organizations are using big data for everything from improving profitability to helping save lives. Healthcare companies, hospitals, and researchers collect massive amounts of data. But all of this data isn't useful in isolation. It becomes important when the data is analyzed to highlight trends and threats in patterns and create predictive models.

- Genomic research Big data can play in a significant role in genomic research. Using big data, researchers can identify disease genes and biomarkers to help patients pinpoint health issues they may face in the future. The results can even allow healthcare organizations to design personalized treatments. Challenges The volume of genome data is enormous, and running complex algorithms on the data is complicated and can require long processing times.
- Patient experience and outcomes Healthcare organizations seek to provide better treatment and improved quality of care—without increasing costs. Big data helps them improve the patient experience in the most cost-efficient manner.
- With big data, healthcare organizations can create a 360-degree view of patient care as the patient moves through various treatments and departments. Challenges Improving the patient experience requires a large volume of patient data, some of which could be multi- structured data, such as doctor notes or images. Additionally, to analyze patient journeys, path and graph analyses are often needed.
- Claims fraud For every healthcare claim, there can be hundreds of associated reports in a variety of different formats. This makes it extremely difficult to verify the accuracy of insurance

incentive programs and find the patterns that indicate fraudulent activity. Big data helps healthcare organizations detect potential fraud by flagging certain behaviors for further examination. Challenges Claims fraud analytics is a complex process that involves integrating different data sets, analyzing the claims data, and identifying complex fraud patterns.

- Healthcare billing analytics Big data can improve the bottom line. By analyzing billing and claims data, organizations can discover lost revenue opportunities and places where payment cash flows can be improved.

- This use case requires integrating billing data from various payers, analyzing a large volume of that data, and then identifying activity patterns in the billing data. Challenges Sifting through large volumes of data can be complicated, especially when it comes to integrating different data source

Oil and gas For the past few years, the oil and gas industry has been leveraging big data to find new ways to innovate. The industry has long made use of data sensors to track and monitor the performance of oil wells, machinery, and operations. Oil and gas companies have been able to harness this data to monitor well activity, create models of the Earth to find new oil sources, and perform many other value-added tasks.

- Predictive equipment maintenance Oil and gas companies often lack visibility into the condition of their equipment, especially in remote offshore and deep-water locations. Big data can help by providing insight so companies can predict the remaining optimal life of their systems and components, ensuring that their assets operate at optimum production efficiency

- Challenges Machine, log, and sensor data from different types of equipment comes in varying formats. Integrating all of this data can be difficult. Moreover, the data needs to be analyzed quickly and put into operation to effectively prevent downtime.

- Oil exploration and discovery Exploring for oil and gas can be expensive. But companies can make use of the vast amount of data generated in the drilling and production process to make informed decisions about new drilling sites. Data generated from seismic monitors can be used to find new oil and gas sources by identifying traces that were previously overlooked. Challenges To discover potential new oil deposits, companies will need to integrate and analyze an enormous volume of unstructured data.

Oil production optimization

- Oil production optimization Unstructured sensor and historical data can be used to optimize oil well production. By creating predictive models, companies can measure well production to understand usage rates. With deeper data analysis, engineers can determine why actual well outputs aren’t tallying with their predictions. Challenges This use case involves analyzing a large volume of data. Complex algorithms are also needed to identify the curve shape associated with that data to identify trends.

References

- Top big data analytics use cases, Oracle

- Big Data and Big Data Analytics: Concepts, Types and Technologies Author(s) : 1Youssra Riahi, 2 Sara Riahi , International Journal of Research and Engineering ISSN: 2348-7860 (O) | 2348-7852 (P) | Vol. 5 No. 9 | September-October 2018 | PP. 524-528

- The Big Data Revolution, Issues and Applications, Azzeddine Riahi, Sara Riahi- IJARCSSE, Volume 5, Issue 8
- Deep learning applications and challenges in big data analytics- Najafabadi et al. Journal of Big Data (2015) 2:1 DOI 10.1186/s40537- 014-0007-7
- BIG DATA ANALYTICS: CHALLENGES AND APPLICATIONS FOR TEXT, AUDIO, VIDEO, AND SOCIAL MEDIA DATA-International Journal on Soft Computing, Artificial Intelligence and Applications (IJSCAI), Vol.5, No.1, February 2016
- Big Data- The definitive guide to the revolution in business analytics- Fujitsu
- [7] http://searchcloudcomputing.techtarget.com/definiti on/MapReduce [8] http://www.informit.com/articles/article.aspx?p=20 08905 [9] http://www.informit.com/articles/article.aspx?p=20 08905

Big Data: Challenges, Opportunities, and Realities Abhay Kumar Bhadani Indian Institute of Technology Delhi, India Dhanya Jothimani Indian Institute of Technology Delhi, India

: Bhadani, A., Jothimani, D. (2016), Big data: Challenges, opportunities and realities, In Singh, M.K., & Kumar, D.G. (Eds.), Effective Big Data Management and Opportunities for Implementation (pp. 1-24), Pennsylvania, USA, IGI Globa