Index

□□□

MODULE I

CHAPTER 1

Introduction to Big Data and Hadoop

University Prescribed Syllabus w.e.f Academic Year 2022-2023

1.1 Introduction to Big Data- Big Data characteristics and Types of Big Data

1.2 Traditional vs. Big Data business approach

1.3 Case Study of Big Data Solutions

1.4 Concept of Hadoop, Core Hadoop Components; Hadoop Ecosystem

1.1 INTRODUCTION TO BIG DATA AND HADOOP

GQ. Firstly, We need to know "what is data?"

(1) Now a day the amount of data created by various advanced technologies like Social networking sites, E-commerce etc. is very large. It is really difficult to store such huge data by using the traditional data storage facilities.

(2) Until 2003, the size of data produced was 5 billion gigabytes. If this data is stored in the form of disks it may fill an entire football field. In 2011, the same amount of data was created in every two days and in 2013 it was created in every ten minutes. This is really tremendous rate.

(3) In this topic, we will discuss about big data on a fundamental level and define common concepts related to big data. We will also see in deep about some of the processes and technologies currently being used in this field.

1.1.1 What is Big Data ?

GQ. What is Big Data?

1. Big Data is a massive collection of data that continues to grow dramatically over time.
2. It is a data set that is so huge and complicated that no typical data management technologies can effectively store or process it.
3. Big Data is like regular data, but it is much larger. A data which are very large in size.
4. Normally we work on data of size MB(WordDoc ,Excel) or maximum GB(Movies, Codes) but data in Peta bytes i.e. 1015 byte size is called Big Data.
5. It is stated that almost 90% of today's data has been generated in the past 3 years.

1.1.2 Sources of Big Data

There are various sources of big data. Now a days in number of fields such huge data get created. Following are the some of fields.

1. **Stock Exchange :** The data in the share market regarding information about prices and status details of shares of thousands of companies is very huge.

2. **Social Media Data :** The data of social networking sites contains information about all the account holders, their posts, chat history, advertisements etc. On topmost sites like facebook and whatsapp, there are literally billions of users.
3. **Video sharing portals :** Video sharing portals like youtube, Vimeo etc. contains millions of videos each of which requires lots of memory to store.
4. **Search Engine Data :** The search engines like Google and Yahoo holds lot much of metadata regarding various sites.

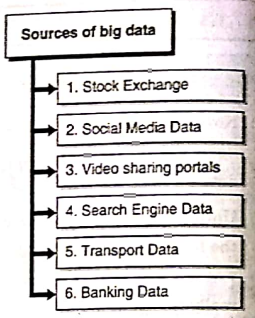


Fig. 1.1.1 : Sources of big data

5. **Transport Data :** Transport data contains information about model, capacity, distance and availability of various vehicles.
6. **Banking Data :** The big giants in banking domain like SBI or ICICI hold large amount of data regarding huge transactions of account holders.

1.2 BIG DATA CHARACTERISTICS

GQ. What are Characteristics of Big Data?
UQ. Describe any five characteristics of Big Data. **MU - Dec. 17, 5 Marks**
UQ. Explain what characteristic of Social Networks make it Big Data. **MU - May 18, 5 Marks**
GQ. Explain Big data along with 5V's

(1) Volume represents the volume i.e. amount of data that is growing at a high rate i.e. data volume in Petabytes.

(2) Value refers to turning data into value. By turning accessed big data into values, businesses may generate revenue.

(3) Veracity refers to the uncertainty of available data. Veracity arises due to the high volume of data that brings incompleteness and inconsistency.

(4) Visualization is the process of displaying data in charts, graphs, maps, and other visual forms.

(5) Variety refers to the different data types i.e. various data formats like text, audios, videos, etc.

(6) Velocity is the rate at which data grows. Social media contributes a major role in the velocity of growing data.

(7) Virality describes how quickly information gets spread across people to people (P2P) networks.

1.2.1 Volume

- As it follows from the name, big data is used to refer to enormous amounts of information.
- We are talking about not gigabytes but terabytes and petabytes of data.
- The IoT (Internet of Things) is creating exponential growth in data.
- The volume of data is projected to change significantly in the coming years.
- Hence, 'Volume' is one characteristic which needs to be considered while dealing with Big Data.

Volume

[Data at Rest]

- Terabytes, Petabytes • Records/Arch • Table/Files • Distributed

1.2.2 Variety

- Variety refers to heterogeneous sources and the nature of data, both structured and unstructured.
- Data comes in different formats – from structured, numeric data in traditional databases to unstructured text documents, emails, videos, audios, stock ticker data and financial transactions.
- This variety of unstructured data poses certain issues for storage, mining and analysing data.
- Organizing the data in a meaningful way is no simple task, especially when the data itself changes rapidly.
- Another challenge of Big Data processing goes beyond the massive volumes and increasing velocities of data but also in manipulating the enormous variety of these data.

Variety

[Data in many Forms]

- Structured
- Unstructured
- Text
- Multimedia

1.2.3 Veracity

- Veracity describes whether the data can be trusted. Veracity refers to the uncertainty of available data.
- Veracity arises due to the high volume of data that brings incompleteness and inconsistency.
- Hygiene of data in analytics is important because otherwise, you cannot guarantee the accuracy of your results.
- Because data comes from so many different sources, it's difficult to link, match, cleanse and transform data across systems.
- However, it is useless if the data being analysed are inaccurate or incomplete.
- Veracity is all about making sure the data is accurate, which requires processes to keep the bad data from accumulating in your systems.

Veracity

[Data in Doubt]

- Trustworthiness
- Authenticity
- Accurate
- Availability

1.2.4 Velocity

- Velocity is the speed in which data is grows, process and becomes accessible.
- A data flows in from sources like business processes, application logs, networks, and social media sites, sensors, Mobile devices, etc.
- The flow of data is massive and continuous.
- Most data are warehoused before analysis, there is an increasing need for real-time processing of these enormous volumes.
- Real-time processing reduces storage requirements while providing more responsive, accurate and profitable responses.
- It should be processed fast by batch, in a stream-like manner because it just keeps growing every years.

Velocity

[Data in Motion]

- Streaming
- Batch
- Real / Near Time
- Processes

1.2.5 Value

- It refers to turning data into value. By turning accessed big data into values, businesses may generate revenue.
- Value is the end game. After addressing volume, velocity, variety, variability, veracity, and visualization – which takes a lot of time, effort and resources – you want to be sure your organization is getting value from the data.
- For example, data that can be used to analyze consumer behavior is valuable for your company because you can use the research results to make individualized offers.

Value

[Data into Money]

- Statistical
- Events
- Correlations

1.2.6 Visualization

- Big data visualization is the process of displaying data in charts, graphs, maps, and other visual forms.
- It is used to help people easily understand and interpret their data at a glance, and to clearly show trends and patterns that arise from this data.
- Raw data comes in a different formats, so creating data visualizations is process of gathering, managing, and transforming data into a format that's most usable and meaningful.
- Big Data Visualization makes your data as accessible as possible to everyone within your organization, whether they have technical data skills or not.

Visualization

[Data Readable]

- Readable
- Accessible
- Presentation
- Visual Forms

1.2.7 Virality

- Virality describes how quickly information gets spread across people to people (P2P) networks.
- It is measures how quickly data is spread and shared to each unique node.
- Time is a determinant factor along with rate of spread.

Virality

[Data Spread]

- P2P
- Shared
- Rate of Spread

1.3 EXAMPLES OF BIG DATA APPLICATIONS

GQ. List the examples of big data. **(2 Marks)**

GQ. Explain the examples of big data. **(6 Marks)**

There are various big data applications as shown in Fig 1.3.1

1. **Fraud detection**
   - Fraud detection is a Big Data application example for businesses which has operations like any type of claims or transaction processing.
   - Number of times the detection of fraud is concluded long after the fact. At this point the damage has been already done all that's left is to decrease the harm and revise policies to prevent it in future.

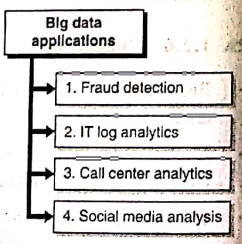


Fig. 1.3.1 : Big data applications

   - The Big Data platforms can analyze claims and transactions of businesses. They identify large-scale patterns across many transactions or detect anomalous behaviour of a some user. This helps to avoid the fraud.

2. **IT log analytics**
   - An enormous quantity of logs and trace data is generated in IT solutions and IT departments. Many times such data go unexamined: organizations simply don't have the manpower or resource to go through all such information.

   - Big data has the ability to quickly identify large-scale patterns to help in diagnosing and preventing problems. It helps the organization with a large IT department.

3. **Call center analytics**
   - Now we turn to the customer-facing Big Data application examples, of which call center analytics are particularly powerful. Without a Big Data solution, much of the insight that a call center can provide will be ignored or exposed later.
   - By making sense of time/quality resolution metrics, the Big Data solutions are able to identify recurring problems or customer and staff behaviour patterns. Big data can also capture and process call content itself.

4. **Social media analysis**
   - With the help of Social media we can observe the real-time insights into how the market is responding to products and campaigns.
   - With the help of these insights, it is possible for companies to adjust their pricing, promotion, and campaign placement to get optimal results.

1.4 TYPES OF BIG DATA

GQ. What are different Types of Big Data ?

There are three types of Big Data Analytics:

1. Unstructured 2. Structured 3. Semi-structured

1.4.1 Type #1 : Unstructured

- Any data with unknown form or the structure is classified as unstructured data. In addition to the size being huge, un-structured data poses multiple challenges in terms of its processing for deriving value out of it.
- Typical example of unstructured data is, a heterogeneous data source containing a combination of simple text files, images, videos like search in Google Engine.
- Now a day organizations have wealth of data available with them but unfortunately they don't know how to derive value out of it since this data is in its raw form or unstructured format.
- Human Generated Data Machine Generated Data.
- Unstructured – Example: The output returned by 'Google Search'

1.4.1(A) Characteristics of Unstructured Data

(1) Data neither conforms to a data model nor has any structure.

(2) Data can not be stored in the form of rows and columns as in Databases.

(3) Data does not follows any semantic or rules.

(4) Data lacks any particular format or sequence.

(5) Data has no easily identifiable structure.

(6) Due to lack of identifiable structure, it can not used by computer programs easily.

1.4.1(B) Sources of Unstructured Data

(1) Web pages (2) Images (JPEG, GIF, PNG, etc.)

(3) Videos (4) Memos

(5) Reports (6) Word documents and PowerPoint presentations

(7) Surveys

1.4.1(C) Advantages and Disadvantages of Unstructured Data

Advantages

1. Its supports the data which lacks a proper format or sequence.
2. The data is not constrained by a fixed schema.
3. Very Flexible due to absence of schema.
4. Data is portable.
5. It is very scalable.
6. It can deal easily with the heterogeneity of sources.
7. These type of data have a variety of business intelligence and analytics applications.

Disadvantages

1. It is difficult to store and manage unstructured data due to lack of schema and structure.
2. Indexing the data is difficult and error prone due to unclear structure and not having pre-defined attributes. Due to which search results are not very accurate.
3. Ensuring security to data is difficult task.

1.4.2 Type #2 : Structured

- Any data that can be stored, accessed and processed in the form of fixed format is termed as a "Structured" data.
- Over the period of time, talent in computer science have achieved greater success in developing techniques for working with such kind of data (where the format is well known in advance) and also determining value out of it.
- When size of such data grows to a huge extent, typical sizes are being in the range of multiple zettabyte. Data stored in a relational database management system in one example of a structured data.
- **Structured data** is the data which conforms to a data model, has a well define structure, follows a consistent order and can be easily accessed and used by a person or a computer program.
- Structured data is usually stored in well-defined schemas such as Databases. It is generally tabular with column and rows that clearly define its attributes.
- SQL (Structured Query language) is often used to manage structured data stored in databases.

1.4.2(A) Characteristics of Structured Data

- Data conforms to a data model and has easily identifiable structure.
- Data is stored in the form of rows and columns.

Example : Database

- Data is well organised so, Definition, Format and Meaning of data is explicitly known.
- Data resides in fixed fields within a record or file.
- Similar entities are grouped together to form relations or classes.
- Entities in the same group have same attributes.
- Easy to access and query, So data can be easily used by other programs.
- Data elements are addressable, so efficient to analyse and process.

1.4.2(B) Sources of Structured Data

(1) SQL Databases (2) Spreadsheets such as Excel

(3) OLTP Systems (4) Online forms

(5) Sensors such as GPS or RFID tags (6) Network and Web server logs

(7) Medical devices

1.4.2(C) Advantages of Structured Data

1. Structured data have a well defined structure that helps in easy storage and access of data.
2. Data can be indexed based on text string as well as attributes. This makes search operation hassle-free.
3. Data mining is easy i.e knowledge can be easily extracted from data.
4. Operations such as Updating and deleting is easy due to well structured form of data.
5. Business Intelligence operations such as Data warehousing can be easily undertaken.
6. Easily scalable in case there is an increment of data.
7. Ensuring security to data is easy.

Structured - Example

Employee_Table

Employee_ID	Employee_Name	Gender	Department	Salary_In_lacs
1	XYX	MALE	FINANCE	850000
2	ABC	MALE	ADMIN	250000
3	PQR	FEMALE	SALES	350000
4	MNR	FEMALE	FINANCE	600000

1.4.3 Type #3 : Semi Structured

- Semi structured is the third type of big data. Semi-structured data can contain both the forms of data.
- Semi-structured data pertains to the data containing both the formats mentioned above, that is, structured and unstructured data.
- To be precise, it refers to the data that although has not been classified under a particular repository (database), yet contains vital information or tags that segregate individual elements within the data.

- Web application data, which is unstructured, consists of log files, transaction history files etc.
- Online transaction processing systems are built to work with structured data wherein data is stored in relations (tables).
- Semi-structured data is data that does not conform to a data model but has some structure. It lacks a fixed or rigid schema. It is the data that does not reside in a rational database but that have some organizational properties that make it easier to analyze. With some processes, we can store them in the relational database.

1.4.3(A) Characteristics of Semi-structured Data

1. Data does not conform to a data model but has some structure. Data can not be stored in the form of rows and columns as in Databases
2. Semi-structured data contains tags and elements (Metadata) which is used to group data and describe how the data is stored.
3. Similar entities are grouped together and organized in a hierarchy. Entities in the same group may or may not have the same attributes or properties.
4. Does not contain sufficient metadata which makes automation and management of data difficult.
5. Size and type of the same attributes in a group may differ.
6. Due to lack of a well-defined structure, it can not used by computer programs easily.

1.4.3(B) Sources of semi-structured Data:

(1) E-mails (2) XML and other markup languages

(3) Binary executables (4) TCP/IP packets

(5) Zipped files (6) Integration of data from different sources

(7) Web pages

1.4.3(C) Advantages and Disadvantages of Semi-structured Data

Advantages

1. The data is not constrained by a fixed schema.
2. Flexible i.e Schema can be easily changed.
3. Data is portable.

4. It is possible to view structured data as semi-structured data.
5. Its supports users who can not express their need in SQL.
6. It can deal easily with the heterogeneity of sources.

Disadvantages

1. Lack of fixed, rigid schema make it difficult in storage of the data.
2. Interpreting the relationship between data is difficult as there is no separation of the schema and the data.
3. Queries are less efficient as compared to structured data.

Semi-structured - Example

- User can see semi-structured data as a structured in form but it is actually not defined with e.g. a table definition in relational DBMS.
- Personal data stored in a XML file:

<rec><name>Prashant Rao</name><sex>Male</sex><age>35</age></rec><rec><name>Seema R.</name><sex>Female</sex><age>41</age></rec><rec><name>Satish Mane</name><sex>Male</sex><age>29</age></rec>

1.5 DIFFERENCE BETWEEN STRUCTURED, SEMI-STRUCTURED AND UN-STRUCTURED DATA

GQ. What is difference between structured, semi-structured and Un-Structured Data ?

Properties	Structured data	Semi-structured data	Unstructured data
Technology	It is based on Relational database table	It is based on XML/RDF(Resource Description Framework).	It is based on character and binary data
Transaction management	Matured transaction and various concurrency techniques	Transaction is adapted from DBMS not matured	No transaction management and no concurrency
Version management	Versioning over tuples, row, tables	Versioning over tuples or graph is possible	Versioned as a whole

Properties	Structured data	Semi-structured data	Unstructured data
Flexibility	It is schema dependent and less flexible	It is more flexible than structured data but less flexible than unstructured data	It is more flexible and there is absence of schema
Scalability	It is very difficult to scale DB schema	It's scaling is simpler than structured data	It is more scalable.
Robustness	Very robust	New technology, not very spread	–
Query performance	Structured query allow complex joining	Queries over anonymous nodes are possible	Only textual queries are possible

1.6 TRADITIONAL VS. BIG DATA BUSINESS APPROACH

GQ. What is Traditional Data & BigData ?

GQ. Explain in detail Traditional vs. Big Data Business Approach. (5 Marks)

1. Traditional Data

- Traditional data is the structured data which is being majorly maintained by all types of businesses starting from very small to big organizations.
- In traditional database system a centralized database architecture used to store and maintain the data in a fixed format or fields in a file. For managing and accessing the data Structured Query Language (SQL) is used.

2. Bigdata

- We can consider big data an upper version of traditional data. Big data deal with too large or complex data sets which is difficult to manage in traditional data-processing application software.
- It deals with large volume of both structured, semi structured and unstructured data. Volume, Velocity and Variety, Veracity and Value refer to the 5'V characteristics of big data.
- Big data not only refers to large amount of data it refers to extracting meaningful data by analyzing the huge amount of complex data sets.

GQ. Compare big data analytics with traditional data mining. MU - Dec. 18, 5 Marks

Sr. No.	Traditional Data	Big Data
1.	Traditional data is generated in enterprise level.	Big data is generated in outside and enterprise level.
2.	Its volume ranges from Gigabytes to Terabytes.	Its volume ranges from Petabytes to Zettabytes or Exabytes.
3.	Traditional database system deals with structured data.	Big data system deals with structured, semi structured and unstructured data.
4.	Traditional data is generated per hour or per day or more.	But big data is generated more frequently mainly per seconds.
5.	Traditional data source is centralized and it is managed in centralized form.	Big data source is distributed and it is managed in distributed form.
6.	Data integration is very easy.	Data integration is very difficult.
7.	Normal system configuration is capable to process traditional data.	High system configuration is required to process big data.
8.	The size of the data is very small.	The size is more than the traditional data size.
9.	Traditional data base tools are required to perform any data base operation.	Special kind of data base tools are required to perform any data base operation.
10.	Normal functions can manipulate data.	Special kind of functions can manipulate data.
11.	Its data model is strict schema based and it is static.	Its data model is flat schema based and it is dynamic.
12.	Traditional data is stable and inter relationship.	Big data is not stable and unknown relationship.

Sr. No.	Traditional Data	Big Data
13.	Traditional data is in manageable volume.	Big data is in huge volume which becomes unmanageable.
14.	It is easy to manage and manipulate the data.	It is difficult to manage and manipulate the data.
15.	Its data sources includes ERP transaction data, CRM transaction data, financial data, organizational data, web transaction data etc.	Its data sources includes social media, device data, sensor data, video, images, audio etc.

1.7 CASE STUDY OF BIG DATA SOLUTIONS

GQ. Explain Case Study of BIG Data Solutions

- Undoubtedly Big Data has become a major game change in most part of the cutting edge industries over the last few years.
- As Big Data keeps on going day by day, the number of various organizations that are adopting Big Data keeps on expanding.

Let's discuss example

- An e-commerce site XYZ (having 100 million users) wants to offer a gift voucher of 100$ to its top 10 customers who have spent the most in the previous year.
- Moreover, they want to find the buying trend of these customers so that company can suggest more items related to them.
- Issues: Huge amount of unstructured data which needs to be stored, processed and analyzed.
- Storage: This huge amount of data, Hadoop uses HDFS (Hadoop Distributed File System) which uses commodity hardware to form clusters and store data in a distributed fashion. It works on Write once, read many times principle.
- Processing: Map Reduce paradigm is applied to data distributed over network to find the required output.
- Analyze: Pig, Hive can be used to analyze the data.

- Cost: Hadoop is open source so the cost is no more an issue.
- Where are businesses finding uses for Big Data ?

Walmart

- Biggest retailer in the world and world's biggest organization by revenue. Approx. 2 million workers and 20000 stores in 28+ nations.
- It started to use Big Data concept in earlier stage.
- It used data mining to find designs pattern that can be used to give product suggestions to client, depending on which products were brought together.
- Based on data mining result, it has expanding its conversion rate of customers. Main target of Walmart is to holding customers and enhance their experience.
- Hadoop and NoSQL technologies are used to furnished these customers real time data to gathered from various sources and their effective valuable use.

Uber

- It is the best option for individuals around the globe when moving people and making conveyances.
- It utilizes individuals information of the user to intently monitor which features of services are used.
- To analyze usage pattern and to figure out where the services should be more engaged.
- It focuses around the oraganic market of the services because of which the costs of services gave changes.
- The use of data is surge pricing and its influences the rate of demand.

Netflix

- It is very popular entertainment company work in online on-request web based video streaming for its customers.
- It has been determined to be able to predict what precisely its customers will appreciate viewing with Big Data.
- Recently, Netflix begun positioning itself as a content creator, not simply a distribution medium which is solidly said based on data analytics.

- Data likes are recommendation engines take care of customers watch, regularly playback halted, ratings and so on.
- It has incorporates with Hadoop, Hive and Pig and other traditional business intelligence.

1.8 CONCEPT OF HADOOP

1.8.1 What is Hadoop ?

Hadoop is an open-source software Platform for storing massive volumes of data and running applications on clusters (groups) of commodity software. It gives us the massive data storage capability, massive computational power and the ability to handle different virtually limitless jobs that can be a running job, waiting jobs or tasks. Its main essential component is to support growing big data technologies, thereby support forward-thinking analytics like Predictive analytics, Machine learning and data mining. Hadoop has the capability to handle different modes of data such as structured, unstructured and semi-structured data. It gives us the elasticity to collect, process, and investigate data that the old data warehouses concept failed to do.

1.8.2 History of Hadoop

- The Hadoop was introduced by Doug Cutting and Mike Cafarella in 2002. Its beginning was the Google File System paper, printed by Google.
- In the year 2002, Doug Cutting and Mike Cafarella started to work on a project of Apache Nutch. It is an open source i.e. free web crawler software project.
- While working on Apache Nutch, they were facing some issue with big data. To store that data, they have invested lot of money which becomes the challenging of that project for completion.
- Due to this problem appearance of Hadoop came into existence.
- In 2003, Google presented a file system known as GFS (Google file system). It is a registered distributed file system developed to provide effective access to data.
- In year 2004, Google released the concept of a white paper on Map Reduce.
- This technique makes simpler the data processing on large clusters(groups).
- In 2005, Doug Cutting and Mike Cafarella presented a new file system known as NDFS (Nutch Distributed File System), this file system also contains Map reduce. In 2006, Doug Cutting resign Google and joined Yahoo. Based on the Nutch project,

Dough Cutting announce a new project Hadoop with a file system known as HDFS (Hadoop Distributed File System).

- Hadoop first version 0.1.0 was released and Doug Cutting gave named his project Hadoop after his son's toy elephant. In 2007, Yahoo successfully run two clusters of 1000 machines.
- In 2008, Hadoop became the quickest system to sort 1 terabyte of data on a 900-node cluster in 209 seconds. In 2013, Hadoop 2.2 was released. And Currently In 2017, Hadoop 3.0 was released.

1.8.3 Features of Hadoop

1. **Suitable for Big Data Analysis :** As Big Data manages to be distributed and unstructured in nature, Hadoop clusters are well-matched for analysis of Big Data. Meanwhile it is processing logic (not the actual data) that flows to the computing nodes and less network bandwidth is spent. This concept is called as data locality which helps to increase the productivity of Hadoop based applications.
2. **Scalability :** Hadoop clusters can easily be scaled to any amount by adding extra cluster nodes and thus allows for the growth of Big Data. Also, scaling does not require adjustments to application logic.
3. **Fault Tolerance :** Hadoop network has a facility to duplicate the input data on to other cluster nodes. So, in the event of a cluster node failure the data processing can still process data by using data stored on another cluster node.

1.8.4 Advantages of Hadoop

1. **Fast :** In HDFS the data distributed over the cluster and mapped such a way which helps in faster recovery. Even the tools to process the data are often on the same servers, thus reducing the processing time can be efficient way to manage the data. It also processes terabytes of data in minutes and Peta bytes in hours.
2. **Scalable :** Hadoop cluster can be extended by just adding nodes in the cluster so failure chance can be less.
3. **Cost Effective :** Hadoop is open source and uses commodity hardware to store data. it is cheaper as compared to traditional RDMS.
4. **Tough to failure :** HDFS has the property with which it can duplicate data over the network, so if one node is down or some other network failure happens, then Hadoop takes the backup data and use it. Normally, data are replicated thrice but the replication factor is configurable.

1.8.5 Challenges of Hadoop

1. Hadoop is a complex distributed system with low-level Application programming interface.
2. Specialized skills are required for using Hadoop and prevent most developers from efficiently bringing solutions.
3. Business logic and infrastructure APIs have no clear separation therefore burdening come on app developers.
4. Automated testing of end-to-end solutions is unfeasible or terrible.
5. Common data patterns often require but does not support data steadiness and accuracy.
6. Hadoop is more than just disconnected storage.
7. Hadoop is a various collection of many open source projects.
8. Understanding multiple technologies and hand-coding combination between them is difficult.
9. Significant effort is wasted on simple tasks like data absorptions and ETL (Extract, Transform, Load).
10. Real-time and batch ingestion requires extremely integration numerous components.
11. Different processing models require data to be stored in specific ways so that data can be handle easily.

1.8.6 Architecture of Hadoop

UQ. Explain the physical architecture of Hadoop. **(MU - Dec. 18, 4 Marks)**

Hadoop basically has Master-Slave Architecture for storing data and distributed processing of data by using MapReduce and HDFS methods. In Hadoop, master or slave system can be set up on the cloud or on premise.

1. **NameNode :** NameNode represents all files and directory which is used in the namespace.
2. **DataNode :** DataNode helps you to manage the states of an HDFS node and allows you to cooperate with its blocks.

3. **Master Node :** The master node allows you to conduct parallel processing of data by use Hadoop MapReduce.

4. **Slave node :** The slave nodes are the supplementary machines in the Hadoop cluster which permits you to store data to conduct complex calculations. And all these slave node derives with Task Tracker and DataNode which allows to synchronize the processes with the NameNode and Job Tracker correspondingly.

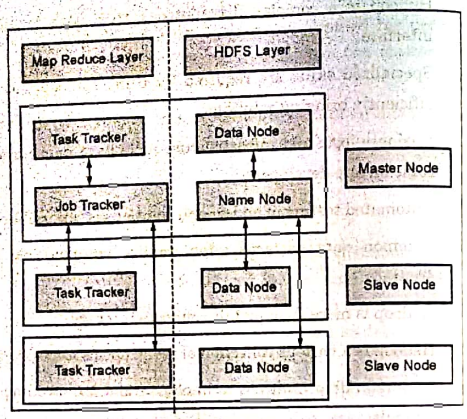


Fig. 1.8.1

1.9 CORE HADOOP COMPONENTS, HADOOP ECOSYSTEM

1.9.1 Core Hadoop Components

UQ. Explain Hadoop ecosystem with core components. (MU - Dec. 18, 4 Marks)

Hadoop concept is an open-source big data technology platform which allows computer networks to perform complex processing and gives results that are forever available even when a condition arises when a few nodes are in unavailable state for functional processing. There are a few important Hadoop core components that it can perform through several cloud-based platforms.

The core components are described as follows :

1. **The Distributed File System :** The furthermost important of the Hadoop core components is the idea of the Distributed File System. It permits the platform to access widely storage devices and use the basic tools to read the obtainable data as well perform the essential analysis. This is a unique file system because it lies overhead the individual file system of the network node computers and allows matchless functionality. The DFS of Hadoop can perform the required data

achievements without worrying about the operating system of the individual computers. This allows the network to service superior power and never face the problem of having to observe with the different computer systems available for use. It also allows the connection to other central components, such as MapReduce.

2. **MapReduce :** MapReduce is another of Hadoop core components that trusts two separate functions, which are required for execution of smart big data operations. The first operation is to read the data from a database and doing inserting operation of data it in a suitable format for performing the required analysis. This is the function which is known as a mapping activity. It basically allows a platform to make the data for the analytical requirements in a common format to allow any computer to do further procedures as per requirements. The next method carried out is mathematical operation. This operation is named as reduction (decrease), because it usually reduces the available map to a set of proper values. Together, these functions are existing in a single module and perform the whole operation which delivers information from available different data sources.

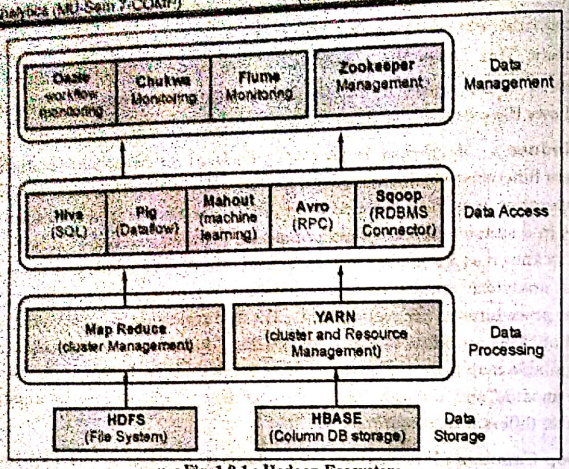
3. **Hadoop Common :** It is also one of the Hadoop core components and the tools which allow any computer to become part of the Hadoop network unrelatedly of the operating system or the present hardware. This module uses Java tools and parts that create a virtual machine and allows the Hadoop platform to store data under its path specific file system. This component named as common as it offers the required common functionality, which removes the difference between the different hardware nodes which may be connected to the network at any time and worldwide.

4. **YARN :** It is the component which accomplishes all the information sources that store the data and then route the required analysis. It is a system which accomplishes the available resources in a network group, as well as schedule the processing tasks to come up with a clever solution for every big data want on the system.

1.9.2 Hadoop Ecosystem Overview

UQ. How big data problems are handled by Hadoop system? (MU - May 19, 5 Marks)

Hadoop ecosystem is a platform or framework which assistances in solving the big data problems. It includes different components and services (ingesting, storing, analysing, and maintaining) inside of it. Most of the services available in the Hadoop ecosystem which contains the main four core components of Hadoop which include HDFS, YARN, MapReduce and Common. Hadoop ecosystem contains both Apache Open Source projects and other wide variation of marketable tools and solutions.



(1A9)Fig. 1.9.1 : Hadoop Ecosystem

Some of the open source examples are Spark, Hive, Pig, Sqoop and Oozie. As we have got some idea about what Hadoop ecosystem is, what it does, and what are its components, let's discuss each concept in detail.

Components of Hadoop Ecosystem

As we have seen an overview of Hadoop Ecosystem and well-known open-source examples, now we are going to discuss deeply the list of Hadoop Components individually and their specific roles in the big data processing. The components of Hadoop ecosystems are :

1. **HDFS :** Hadoop Distributed File System is the spine of Hadoop which runs on java language and allows to stores data in Hadoop applications. They act as a command interface to relate with Hadoop. There are two components of HDFS that are Data node and Name Node. Name node is the main node which manages file systems and activates all data nodes and maintains records of metadata updating. In some case of deletion of data, they automatically record in its Edit Log file. Data Node requires massive storage space due to the operation of reading and writing. They work according to the orders of the Name Node. The data nodes are hardware in the distributed system.

2. **HBASE :** It is an open-source framework storing all types of data and doesn't support the SQL database. They run on top of HDFS and they are written in java language. Most of the companies use them for capturing the features like supporting all types of data, high security, use of HBase tables etc. They play a dynamic role in analytical processing. The two major components of HBase are HBase master and Regional Server. The HBase master is answerable for load balancing in a Hadoop cluster and controls the failure occurred. They are answerable for performing management role. The role of the regional server would be a worker node and responsible for reading, writing data in the cache.
3. **YARN :** It is an important component in the ecosystem and named as operating system in Hadoop which delivers resource management and job scheduling task. The components are Resource and Node manager, Application manager and container. They also act as protectors across Hadoop clusters. They help in the dynamic allocation of cluster resources, increase in data centre process and allows multiple access engines.
4. **Sqoop :** It is a tool that helps in data transfer between HDFS and MySQL and gives hand-on to import and export of data and as well they have a connector for fetching and linking a data.
5. **Apache Spark :** It is an open-source cluster computing framework for data analytics and an important data processing engine. It is written in Scala and comes with packaged standard libraries. They are also used by many companies for their high processing speed and stream processing.
6. **Apache Flume :** It is a distributed service collecting a huge quantity of data from the source (web server) and moves back to its source and relocated to HDFS. It has its own three components are Source, sink, and channel.
7. **Apache Pig :** Data Handling of Hadoop is performed by Apache Pig and by using of Pig Latin Language. It helps in the reuse of code and easy to read and write code.
8. **Hive :** It is an open-source Platform software for execution of data warehousing ideas, it accomplishes to query for large data sets stored in HDFS. It is built on upper of the Hadoop Ecosystem, the language used is Hive Query language. The user submits the hive queries with metadata which converts SQL into Map-reduce jobs and directs to the Hadoop cluster which consists of one master and many slaves.
9. **Apache Drill :** Apache Drill is an open-source SQL engine which procedures on non-relational databases and File system. They are intended to support Semi-structured databases originate in Cloud storage. They have good Memory management abilities to maintain junk collection

10. **Apache Zookeeper :** It is an API that supports in distributed Organisation. Here a node called Z node which is created by an application in the Hadoop cluster. They do services like Synchronization, Configuration etc. Its categories out the time-consuming coordination in the Hadoop Ecosystem.

11. **Oozie :** Oozie is a java web application which maintains several workflows in a Hadoop cluster. Having Web service APIs controls over a job is done anywhere. It is widespread for handling Multiple jobs successfully.

1.9.3 Examples of Hadoop Ecosystem

Regarding map-reduce, we can see an example and use case. One such case is Skybox which uses Hadoop to analyse a huge volume of data. Hive can find simplicity on Facebook. Frequency of word count in a sentence using map-reduce; MAP performs by taking the count as input and perform functions such as Filtering and sorting and the reduce () consolidates the result. Hive example on taking students from different states from student databases using various DML commands.

1.9.4 Limitations

UQ. State limitations of Hadoop ecosystem. (MU - Dec. 18, 2 Marks)

1. Issue with Small Files	2. Slow Processing Speed
3. Support for Batch Processing only	4. No Real-time Data Processing
5. No Delta Iteration	6. Latency
7. Security	8. No Abstraction
9. No Caching	10. Lengthy Line of Code

1. **Issue with Small Files**

Hadoop is not suitable for the small data. **(HDFS) Hadoop distributed file system** wants the capability to professionally support the arbitrary reading of the small files since of it is high volume design. Small files are the main problematic in HDFS. A small file is expressively minor than the HDFS block size (default 128MB). If we are storing these vast numbers of small files, then HDFS cannot handle these files while HDFS is working accurately with a small unit of large files by storing large data sets rather than storing several small files. If there are several small files, then the **NameNode** will get burden meanwhile it stores the namespace of HDFS.

2. **Slow Processing Speed**

In Hadoop, with a support of the parallel and distributed algorithm the MapReduce procedure the large data sets. There are some tasks that we need to perform like Map and Reduce and thus the MapReduce needs a lot of time to complete these tasks thus by increasing the time interval. The Data is spread and handled over the cluster in MapReduce which increases the period and decreases the handling and execution speed.

3. **Support for Batch Processing only**

Hadoop supports the batch processing only and it does not procedure the streamed data and later complete performance is slower. The MapReduce framework of the Hadoop does not influence the memory of the **Hadoop cluster** to the extreme level.

4. **No Real-time Data Processing**

Apache Hadoop is for the operation of batch processing, which allow it to take a vast amount of data in input and execute it and generate the outcome. Even though batch processing is very well-organized for processing a data of high volume dependent on the size of the data that is being processes and the computational power of the system but basically an output can be delay so the Hadoop is not appropriate for Real-time data processing.

5. **No Delta Iteration**

Hadoop isn't well-organized for the constant processing basically the Hadoop doesn't support the repeated data flow i.e. sequence of phases during which the respective output of the earlier phase is the input to the succeeding phase.

6. **Latency**

The Hadoop MapReduce framework is that the comparatively slower so meanwhile its supporting the various format, structure and vast capacity of information or data. In **MapReduce**, Map takes the collection of the data and decodes it into the alternative set of data where the separate elements are fragmented down into **key-value pairs** and reduce the output from the map as input and process extra and MapReduce requires plenty of the time to accomplish these tasks thus by increasing the latency.

7. **Security**

Hadoop is challenges in handling the compound application. If the user doesn't know the way to enable a platform who is managing the platform that the data can be in the danger.

2. **Velocity :** In big data refers to the data transfer from various digital platform such as online systems, various sensors, social media, live web capture, etc. As we all known that social media messages get viral in seconds of time. Such scenario of big data represents to its velocity. In some cases, it needs to be analysed the data without storing it.
3. **Variety :** Refers to different types of data from many sources, it may be in structured, unstructured or semi-structured. Big data analytics provides the facility to integrate this data.
4. **Veracity :** Means Complexity which indicate that the data must be able to transfer via different multiple data centers such as the cloud and geographical zones. Managing or analyzing this huge data is very complex task.
5. **Value :** Is measure of visible or invisible benefits gained by organisation by the use of Big Data. It is more useful when the data from Big Data is converted into value then it gets used.

Chapter Ends...

□□□