**Week 1 contents:**

————

**1. Big data introduction:**

- is a term for collection of large and complex datasets, that makes it difficult processing using relational database management tools or traditional data processing applications. (Mainly from TB size)

**2. Five V(s) in Bigdata**

- Volume: the size and amount of big data

- Variety: different types of data (structured, unstructured, semi-structured data)

- Velocity: the speed of process in receiving, storing and managing data.

- Veracity: the quality of data received.

- Value: the perspective of the business.

**3. Monolithic vs Distributed System** (vertical scale vs horizontal scale)

3.1. Storage

- Monolithic system is where the data stored, transformed, manipulated, consumed, managed from a single centralised data store.

- Distributed data mesh is highly decentralised or the platform version of micro services.

3.2. Processing / Computation

3.3. Scalability

**4. Hadoop - Introduction:** is a open source framework that is used to store and process large datasets ranging in size from gigabytes to petabytes of data.

**5. 3 Core Components of Hadoop (HDFS | MapReduce | YARN)**

- HDFS (Hadoop Distributed File System) (Storage): is the storage unit of Hadoop and responsible for storing different kinds of data as blocks in a distributed environment.

- MapReduce: is a framework for processing large datasets in parallel across Hadoop cluster.

- YARN (Yet Another Resource Negotiator): is the processing framework in Hadoop, which manages resources and provides an execution environment to the process.

**6. Hadoop Core & Ecosystem technologies:**

**HDFS (Hadoop Distributed File System**) is the backbone of Hadoop which runs on java language and stores data in Hadoop applications. They act as a command interface to interact with Hadoop.

The two components of HDFS - data node, name node.

Name node manages file systems and operates all data nodes and maintains records of metadata updating.

Data Node requires vast storage space due to reading and writing operations.

- Sqoop: (Data ingestion) is the tool for transferring data between Hadoop and RDBMS and gives hand-on to import and export data.

- Pig: (Data encrypting) was developed by Yahoo which works on pig Latin language, which is Query based language similar to SQL.

- Hive: (data warehouse) with the help of SQL methodology and interface, HIVE performs reading and writing of large data sets. However, its query language is called as HQL (Hive Query Language)

- Oozie: (scheduling) is a java web application that maintains many workflows in a Hadoop cluster. Having Web service APIs controls over a job is done anywhere. It is popular for handling Multiple jobs effectively.

- HBase: (NoSQL db) It’s a NoSQL database which supports all kinds of data and thus capable of handling anything of Hadoop Database. It provides capabilities of Google’s Bigtable, thus able to work on Big Data sets effectively.

**7. Challenges with Hadoop:**

1. MapReduce is very slow

2. Need to learn different components

**8. Cloud and Its advantages:**

1. **Cloud Computing** is a high Lebel abstraction procedure that focuses on business logic. This is a service delivered via the internet that aids you with the computing services without laying much importance on the infrastructural needs.

2. **Advantage of Cloud:**

- Scalable

- CapEx and OpEx (Capacity expenses vs Operation expenses)

- Agility

- GeoDistribution (Distributed expenses)

- Disaster Recovery

- Cost effective (Pay as you go)

3. **Cloud Types**:

i) Public ii) Private iii) Hybrid (public + private)

4. **On-premises vs private cloud:** (the main difference is the location of the infrastructure)

- On-premises data centres are built on the organisation’s premises. Private clouds are hosted on remote infrastructure.

- On-premises provide complete control and customisation, while private clouds offer scalability and cost-effectiveness.

- On-premises data require a high capital investment and are suitable for organisations with large IT budgets. Private clouds are suitable for small and medium-size organisation that requires cost-effective solutions.

5. **Apache Spark:** (replacement of compute engine MapReduce)

- Apache Spark is a framework for real-time data analytics in a distributed computing environment.

- It executes in-memory computations to increase the speed of data processing.

- It is 100x faster than MapReduce for large-scale data processing by exploiting in-memory computations and other optimisations.

5.1. Storage - Ex: HDFS | Amazon S3 | Azure ADLS Gen2 | Google Cloud | Local Storage

5.2 Resource Manager - Ex: YARN | Masos | Kubernetes

5.3 Spark in 10x to 100x faster than traditional MapReduce as it stores and processed the data in Memory.

**9. Database vs Data Warehouse vs Data Lake**

- Database is a storage location that houses structured data (Schema on write) - Ex: Oracle | PostgreSQL | MongoDB | Redis | Elasticsearch | Apache Cassandra.

- Data warehouse is a large storage location for data that we accumulate from a wide range of sources. Data Warehouses are popular with mid and large-size businesses as a way of sharing data and content across the team or department-siloed database. (Schema on write) - Ex: Snowflake | YellowBrick | Teradata

- Data lake is a large storage repository that holds a huge amount of raw data in its original format until you need it. Data lakes exploit the biggest limitation of data warehouses: their ability to be more flexible. (Schema on read) - Ex: Hadoop | Azure | Amazon S3

**10. Big Data Flow:**

**Data from multiple sources -> Data Ingestion -> Storage (Data Lake) ->**

**Processing -> Serving layers -> Visualisation Tools**

(Database (sourceDB) -> Sqoop (ingestion -> HDFS (data storage) -> MapReduce/SPARK (processing) -> Hive/MySQL -> Tableau / HIVE (Serving)

(MySQL RDBMS) -> Sqoop -> HDFS -> MapReduce / SPARK -> Hive/ mySQL -> Tableau

**11. Categories of computation:**

* Serverless Computation is a cloud-based service where a cloud provider manages the server. The cloud provider dynamically allots compute storage and resources as needed to execute each line of code.
* Serverful Computation involves the use of long-running physical servers to perform various computing tasks, such as running software programs and storing data. Serverful system can be quite expensive and complex to manage. The servers have to be purchased and maintained.

**12. HDFS Architecture:**

* Master vs Slave
* Name Node and Data Node
  + NameNode is the **master** node in the distributed environment, and it maintains the metadata information for the blocks of data stored in HDFS live block location, replication factors.
  + DataNode: is the **slave** nodes, which are responsible for storing data in HDFS. NameNode manages all the DataNodes.

**13. Role of Data Engineer:**

- Data engineer role is a bridge between data Owners and data consumers where they implement methods to improve data reliability and quality.