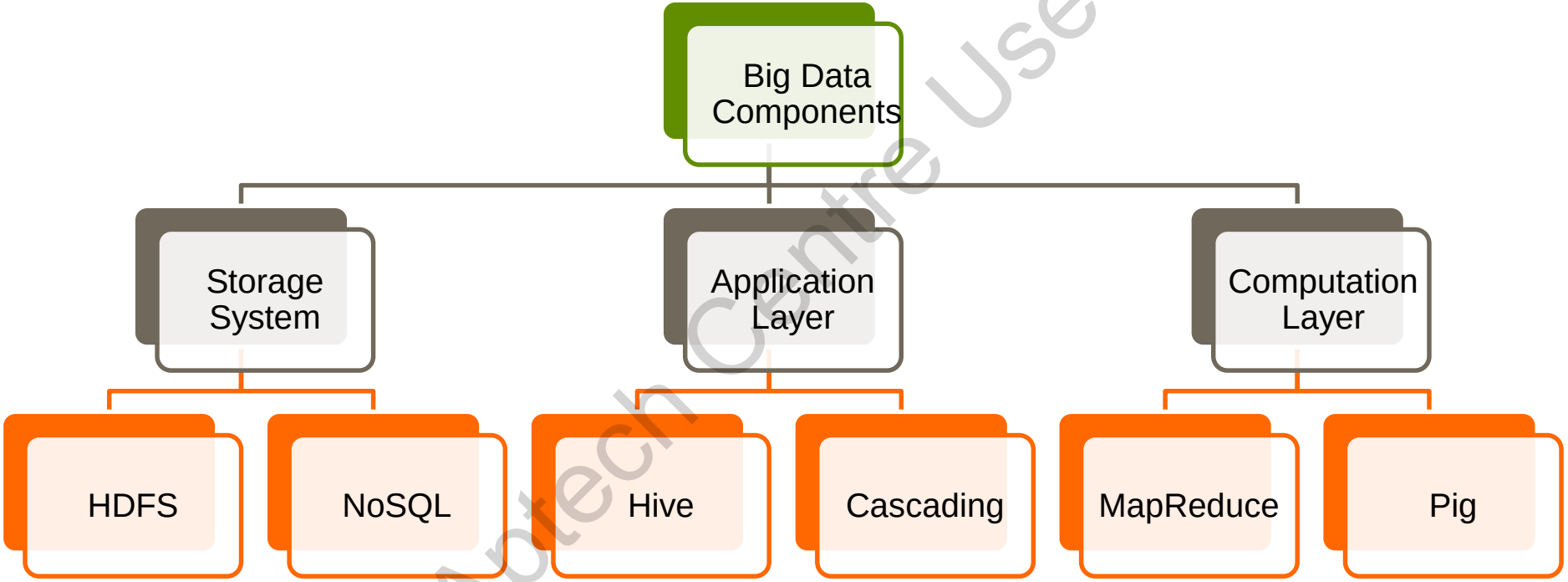
SESSION 4

Components of Big Data and Adoption of Technologies

Objectives

- Describe the main components of Big Data platform
- Explain the adoption of new technologies in Big Data

Components of Big Data



Big Data Technologies and Tools

- Different tools and technologies used for working with Big Data are classified based on the following:

- High-Availability Distributed Object-Oriented Platform (Hadoop)
- Data Storage and Management
- Data Cleaning
- Data Mining
- Data Analysis
- Data Visualization
- Data Integration
- Data Languages
- Data Extraction
- Big Data Collection Methods

Hadoop (1-2)

- An open-source software framework used for distributed storage of very large database on computer clusters
- Allows the user to scale data up and down
- Features of Hadoop are:
  - It secures and maintains the data by preserving and storing the replica of the data
  - It concentrates on scaling depending on the usage of data
  - It can delete and detect an unsuccessful task and unsuccessful data transaction
  - It recovers the data and automatically restores it

Hadoop (2-2)

- Hadoop Platform Stack consists of:

HDFS

Hive

HBase

Pig

Data Storage and Management

- Tools that offer storage facilities are:

Hadoop

- Provides huge amount of data storage, massive processing power, and concurrent tasks

Cloudera

- An enterprise solution to support businesses to manage their Hadoop ecosystem
- Helps an organization to create an enterprise data hub, thus providing better access

MongoDB

- Handles unstructured data, semi-structured data, and frequently modified data

Talend

- Works on Master Data Management (MDM)
- An open-source software, it is free of cost and organizations can employ them at any stage
- Merges real-time data, applications, and processes with embedded data quality and stewardship

Data Cleaning

- Data needs to be cleaned thoroughly before mining
- Companies that will refine and reshape data set are:

OpenRefine

- Also, called as GoogleRefine
- An open-source tool used to clean the messy data
- Cleans the unstructured data quickly and easily and it is very user-friendly

DataCleaner

- A tool used to convert messy semi-structured data sets into structured data
- Reads only structured and clean data
- Offers data warehousing and data management services

Data Mining

- The process of determining insights within a database as opposed to extracting data from Web pages into databases
- Tools for data mining are:

RapidMiner	IBM SPSS Modeler	Oracle Data Mining	Teradata	FramedData	Kaggle
• A data tool used for predictive analysis • PayPal, Deloitte, eBay, and Cisco	• A suite used for data mining • Includes text analysis, entity analytics, decision management, and optimization	• Allows users to identify insights. • Identifies customer behavior and targets best customers	• Provides end to end solutions and services • It is a host of services including implementation, business consulting, training, and tech support	• The apt tool if user churn is of concern • Does not require any user intervention	• The world's largest data science community • Solves a tough problem or some data mining issue

Data Analysis

- Data analysis is breaking down data and evaluating the impact of those patterns overtime
- Tools of data analysis are:

Qubole

- Simplify, speed up, and scale the data analytics workloads
- An enterprise level solution, which offers a free 30 days trial
- It is flexible and platform accessible

BigML

- Simplifies machine learning
- Provides powerful machine learning service with a user-friendly interface

Statwing

- Provides high level data analysis ranging from attractive visuals to complex analysis
- It has user-friendly blog on NFL data

Data Visualization

- Its main aim is to make data come alive
- Tools of Data Visualization are:

Tableau	Silk	CartoDB	Chartio	Plot.ly	DataWrapper
• A tool used to create maps, charts, scatter plots, and other visuals • It has released a Web Connector	• A simpler form of data visualization and analytical tool • It aids in creating interactive charts and maps	• It is meant for creating maps • It helps users to visualize location data without any programming knowledge	• A visual query language combining various data sources and executes queries • It allows users to schedule PDF reports	• A data tool that is used to create graph and 2D and 3D charts • It provides free version which creates private and unlimited public charts	• An open-source tool that is used to create embeddable charts • It has free version and paid version of the tool

Data Integration

- Tools of Data Integration are:

BlockSpring

- A unique program which includes the power of services such as IFTTT and Zapier
- It allows to connect to a host of third-party programs

Pentaho

- It uses drag and drop user interface to integrate the data
- It offers embedded analytics and business analytics services
- It is an enterprise package and users can request for a free trial of the product

Data Languages

- They can handle huge and complex datasets.
- Popular programming languages are:

R

- It is used for statistical computing and data science
- It was designed to execute matrix calculations – standard arithmetic functions
- It helps to produce easy visualizations based on these calculations

Python

- A general-purpose tool and is default choice for a developer
- Its user base has dedicated itself to producing libraries and extensions

Julia

- It is built for speed and scalability of operation when managing huge data sets
- It would group the assets of other analytics-oriented programming languages

Data Extraction

- It is the procedure of accumulating unstructured data from numerous sources and transforming the data to a structured table.
- It includes the huge diagnostic information logged by user devices
- Import.io is a popular tool for data extraction
- It allows users to convert Websites into machine readable and structured data with no coding required.

Big Data Collection Methods

- Different stages involved in Big Data collection are:

Collecting Data

- It comprises collecting of information from numerous means, such as information stockrooms, information sources, and information bazaars

Store

- It includes grouping the information into appropriate database servers and frameworks

Information Organization

- It involves masterminding and sorting the information on the basis of organized, semi-unstructured, and unstructured information

New Technologies Adoption

- There are multiple innovations available across the world for Big Data
- Some of the new technologies are:

Apache Pig	Apache Hive	Apache Spark
Apache Kafka	Presto	HBase

Apache Pig

- Is an innovation requiring 1/20th lines of programming code and 1/16th of development time as compared to Hadoop MapReduce
- During ETL phase, where raw data is cleaned to create datasets that users can consume for analysis, Apache Pig with a workflow system, such as Oozie, is a great choice
- Challenges with MapReduce are:

  - Hadoop developers have to write customized Java-based MapReduce code for operations: filter, projections, and join
  - It is challenging to handle n-stage jobs with Hadoop MapReduce

Apache Hive

- Apache Hive is used for data processing at data presentation phase

- It helps to perform data analysis more productively and also has improved query abilities

Apache Spark

- Recommended for following users:

Users who want to process and access all the data from prevailing Hadoop environment

Users who want to write applications rapidly in Python, Scala, Java, or R

Users who want to combine streaming, SQL, machine learning, and graph processing

Apache Kafka and Presto

- Apache Kafka was developed to resolve data movement problems amongst the Hadoop clusters
- LinkedIn uses Kafka to transmit more than 800 billion messages each day
- Presto helps with the number of queries and achieves the required results quicker, thus improving productivity
- It is an open-source SQL query solution in the Hadoop ecosystem for running interactive analytic queries on petabytes of data

HBase

- Mike Cafarella released the open-source code for the big table implementation known as HBase (Hadoop Database)
- It is a NoSQL database on top of Hadoop for a very large table having billions of rows and millions of columns
- It is used to deliver real-time read or write access to Big Data

Summary (1-2)

- The main components of Big Data are Storage system, Computation or logic layer, and application logic or interaction.
- In Big Data, a storage system can be either HDFS or NoSQL. The computation or the logic layer comprises MapReduce and Pig and the application logic or interaction can be Hive or Cascading.
- High-Availability Distributed Object-Oriented Platform (Hadoop) is a software framework which analyzes unstructured and structured data and distributes applications on different servers.
- A typical Hadoop Platform Stack comprises Hive, HDFS, HBase, and Pig.
- Apache Pig, Apache Hive, Apache Spark, Apache Kafka, Presto, and HBase are some of the new technologies that have been innovated in the world of Big Data.

Summary (2-2)

- Data storage and management tools used include Hadoop, Cloudera, MongoDB, and Talend.
- Some of the popular data cleaning tools used are OpenRefine and DataCleaner.
- Commonly used data mining tools are RapidMiner, IBM SPSS Modeler, Oracle, Teradata, FramedData, and Kaggle.
- Data analysis tools are Qubole and BigML.
- Some of the data visualization tools used are Tableau, Silk, CartoDB, and Chartio.
- Some of the data integration tools used are BlockSpring and Pentaho.