SNS COLLEGE OF ENGINEERING

Kurumbapalayam(Po), Coimbatore – 641 107
Accredited by NAAC-UGC with 'A' Grade
Approved by AICTE, Recognized by UGC & Affiliated to Anna University, Chennai

Department of Artificial Intelligence and Data Science

Course Name – Foundations of Data Science

III Year / V Semester

Unit 1 – Introduction

Topic 9: Tools BDA

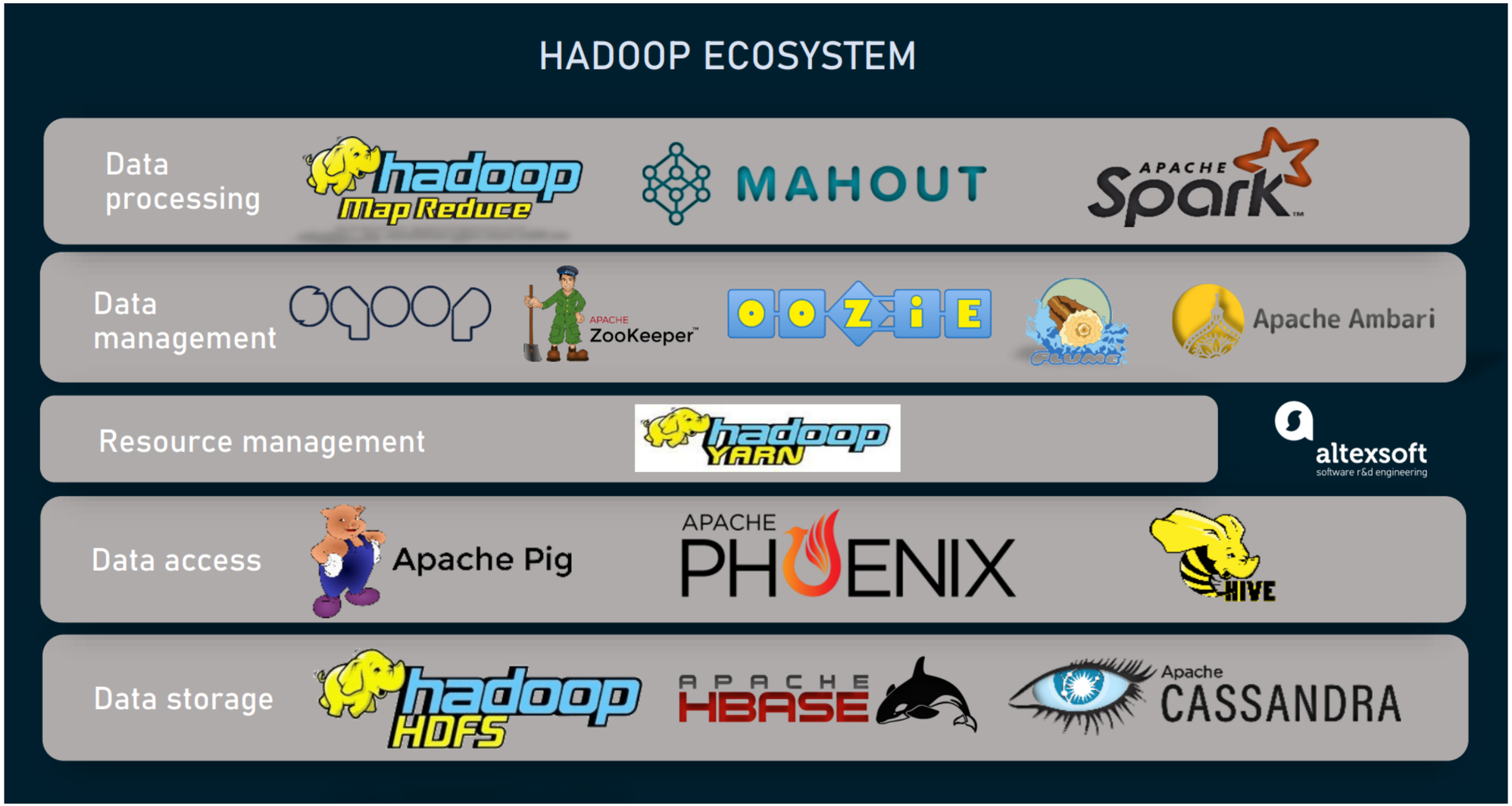
BDA Tools

Big data analytics cannot be narrowed down to a single tool or technology. Instead, several types of tools work together to help you collect, process, cleanse, and analyze big data. Some of the major players in big data ecosystems are listed below

Hadoop

It is an open-source framework that efficiently stores and processes big datasets on clusters of commodity hardware. This framework is free and can handle large amounts of structured and unstructured data, making it a valuable mainstay for any big data operation.

BDA Tools


HADOOP ECOSYSTEM
Data processing
hadoop Map Reduce
MAHOUT
APACHE Spark
Data management
APACHE ZooKeeper
OOZIE
FLUME
Apache Ambari
Resource management
hadoop YARN
altexsoft
software r&d engineering
Data access
Apache Pig
APACHE PHOENIX
HIVE
Data storage
hadoop HDFS
APACHE HBASE
Apache CASSANDRA

BDA Tools

MapReduce

MapReduce is an essential component to the Hadoop framework serving two functions. The first is mapping, which filters data to various nodes within the cluster. The second is reducing, which organizes and reduces the results from each node to answer a query.

YARN stands for “Yet Another Resource Negotiator.” It is another component of second-generation Hadoop. The cluster management technology helps with job scheduling and resource management in the cluster.

Spark is an open source cluster computing framework that uses implicit data parallelism and fault tolerance to provide an interface for programming entire clusters. Spark can handle both batch and stream processing for fast computation.

Sqoop:

Enterprises that use Hadoop often find it necessary to transfer some of their data from traditional relational database management systems (RDBMSs) to the Hadoop ecosystem. Sqoop, an integral part of Hadoop, can perform this transfer in an automated fashion.

BDA Tools

NoSQL

NoSQL databases are non-relational data management systems that do not require a fixed scheme, making them a great option for big, raw, unstructured data.

NoSQL stands for “not only SQL,” and these databases can handle a variety of data models.

Popular NoSQL Databases

- MongoDB
- Couch DB
- Hbase
- Cassandra
- Redis
- Riak
- Neo4j
- Infinite graph
- Hypertable
- Amazon DynamoDB

BDA Tools

Visualization Tools

Tableau is an end-to-end data analytics platform that allows you to prep, analyze, collaborate, and share your big data insights. Tableau excels in self-service visual analysis, allowing people to ask new questions of governed big data and easily share those insights across the organization.

Microsoft Power BI

The Microsoft Power BI is the data visualization tool that is used for business intelligence type of data. It is and can be used for reporting, self-service analytics, and predictive analytics.

JupyteR A web-based application, JupyteR, is one of the top-rated data visualization tools that enable users to create and share documents containing visualizations, equations, narrative text, and live code. JupyteR is ideal for data cleansing and transformation, statistical modeling, numerical simulation, interactive computing, and machine learning.

RAW RAW, better-known as RawGraphs, works with delimited data such as TSV file or CSV file. It serves as a link between data visualization and spreadsheets. Featuring a range of non-conventional and conventional layouts, RawGraphs provides robust data security even though it is a web-based application.

BDA Tools

Visualization Tools

Zoho Analytics

Zoho Analytics is a Business Intelligence and Data Analytics software that can help you create wonderful looking data visualizations based on your data in a few minutes.

Sisense

Sisense is a business intelligence-based data visualization system and it provides various tools that allow data analysts to simplify complex data and obtain insights for their organization and outsiders.

IBM Cognos Analytics

IBM Cognos Analytics is an Artificial Intelligence-based business intelligence platform that supports data analytics among other things. You can visualize as well as analyze your data and share actionable insights with anyone in your organization.

BDA Tools

Python:
This is one of the most versatile programming languages that is rapidly being deployed for various applications including Machine Learning.

SAS:
SAS is an advanced analytical tool that is being used for working with huge volumes of data and deriving valuable insights from it.

R Studio
R Programming: R is the Number 1 programming language that is being used by Data Scientists for the purpose of statistical computing and graphical applications alike.

RapidMiner is a software package that allows data mining, text mining and predictive analytics. Rapidminer is a comprehensive data science platform with visual workflow design and full automation

THANK YOU