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 – Big Data Analytics
III Year / V Semester

Unit 3 – DATA ANALYTICAL FRAMEWORKS

Topic - RDBMS vs Hadoop

Hadoop

Hadoop Conceptual Layer

- It is conceptually divided into Data Storage Layer which stores huge volumes of data Data Processing Layer which processes data in parallel to extract richer and meaningful insights from data.

High-Level Architecture of Hadoop

- Hadoop is distributed Master-Slave Architecture. Master Node is known as Name Node and slave nodes are known as DataNodes.

- Master HDFS: Its main responsibility is partitioning the data storage across the slave nodes. It also keeps track of locations of data on DataNodes.

- Master MapReduce: It decides and schedules computation task on slave nodes.

RDBMS vs Hadoop

Why not RDBMS?

- RDBMS is not suitable for storing and processing large files, images and videos.

- RDBMS is not a good choice when it comes to advanced analytics involving machine learning.

PARAMETERS	RDBMS	HADOOP
System	Relational Database Management System.	Node based flat structure
Data	Suitable for structures data	Suitable for structured, unstructured data. Supports variety of data formats in real time such as XML, JSON, text based flat file formats, etc.
Processing	OLTP	Analytical, Big Data Processing
Performance	Data processing in GB's	Data Processing in PB's
Choice	When data needs consistent relationship	Big data processing, which does not require any consistent relationship between data

RDBMS vs Hadoop

Parameter	RDBMS	HADOOP
Software license	Proprietary	Open source
Project Environment	One project with multiple components	Eco System suite of java based projects
Architecture	Designed for client server architecture	Designed for distributed architecture
Hardware	High usage require high end server	Designed to run on commodity hardware.
File System	Relies on OS file system	Based on distributed file system
Updates	Stable product	Still evolving
Transactions	Support ACID transactions	Support BASE
Schema	Schema required on write	Schema required on read
Processor	Needs expensive hardware or high-end processor to store huge volumes of data.	In a hadoop cluster, a node requires only a processor, a network card, and few hard drives.
Cost	Cost around $10,000 to $14,000 per terabytes of storage	Cost around $4,000 per terabytes of storage.

THANK YOU