SNS COLLEGE OF TECHNOLOGY

(An Autonomous Institution)
Re-accredited by NAAC with A+ grade, Accredited by NBA(CSE, IT, ECE, EEE & Mechanical)
Approvedy by AICTE, New Delhi, Recognized by UGC, Affiliated to Anna University, Chennai

Topic: Modern Data Analytic Tools

Course

16CAT702
Big Data Analytics

Unit I

Introduction to Big data

Elective

III Semester / II MCA

Modern Big Data Analytic Tools

Data Storage and Management

- Apache Cassandra database is widely used to provide an **effective management** of large amounts of data.
- Supports **replication** of data across multiple data centers for scalability.
- Offers very good **fault tolerance** and low latency

- MongoDB is an open source NoSQL database which is **cross-platform compatible** with many built in features
- It is ideal for the business that needs **fast** and real-time data for instant decisions
- It is ideal for the users who want **data-driven** experiences

Modern Big Data Analytic Tools

Data Storage and Management

- The Apache Hadoop software library is a big data **framework. HDFS** is used for storing data.
- It allows **distributed** processing of large data sets across clusters of computers.
- It is designed to **scale up** from single servers to thousands of machines.

- MySQL is an open source **RDBMS.**
- It is used for storing **structured data** and is one of the widely used database.
- It uses **basic SQL** instead of a specialized SQL variant for itself which makes it easier to learn.

Modern Big Data Analytic Tools

Data Storage and Management

Modern Big Data Analytic Tools

Data Mining

- Data mining is the process of discovery insights within a database.
- Teradata is one of the popular data mining tool.

- One of the popular data mining tool to extract useful insights from database.
- Open source tool

Modern Big Data Analytic Tools

Data Ingestion and Data Acquisition

- Data ingestion is the process of getting data into the Hadoop Ford which can be done sqoop, flume or storm

Modern Big Data Analytic Tools

Data Cleansing

•Powerful tool for working messy data. cleaning it; transforming it from one format into another.

•OpenRefine helps to **explore large datasets** with ease.

•Upload your **cleaned** data to a central database.

•Helps the user can **discover, structure, clean,** enrich and publish data of all shapes and sizes.

•Suppose large data volumes, cloud and other deployment options.

•Connected desktop application to transform data for downstream analytics and visualization.

Modern Big Data Analytic Tools

Data Analysis

- Open source big data tool which files the **gaps** of Apache Hadoop concerning data processing.
- Spark can handle **both** batch data and real time data
- As Spark does **in memory data processing** it processes data much faster than traditional disk processing.

- Hive is an **Open source software** big data tool.
- It allows programmers **analyze** large data sets on Hadoop
- It helps with **querying** and managing large datasets real fast

Modern Big Data Analytic Tools

Data Analysis

- Programming model or pattern that is used to access big data stored in the Hadoop File System **(HDFS).**

- Facilitates processing by splitting petabytes of **data smaller chunks**.

- The logic is executed on the server where the **data already resides** which makes the process quicker

Modern Big Data Analytic Tools

Data Visualization

tableau

- Tableau is the most powerful, secure and flexible end to end analytics platform for your data.

- It is the visualization tool used in the business intelligence Industry.

- Tableau can extract data from any database , be it Excel, PDF, Oracle or even Amazon web services.

Power BI

- Power BI is an **business analytics** service by Microsoft.

- It provides interactive **visualizations** and **BI** capabilities with an interface simple enough for end users.

- It can **connect** with just an Excel spreadsheet or bring together cloud based and on premises data warehouses.

Modern Big Data Analytic Tools

Modern Big Data Analytic Tools

An open-source Big Data analytics tools, Hadoop offers massive storage for all kinds of data.

Pros:

Hadoop’s core strength is its HDFS (Hadoop Distributed File System), which holds all types of data, video, images, JSON, XML and plain texts across the same file system.

Very useful for research and development purposes.

Offers easy data access.

Extremely scalable

Cons:

Data redundancy can often cause disk space problems.

For improved efficiency, I/O operations should have been optimized.

Pricing: With the Apache License, this Big Data Analytics tool is free to use.

Modern Big Data Analytic Tools

Xplenty

This cloud-based Big Data Analytics tool for integrating, analyzing and preparing data brings all data sources together. Its intuitive graphical interface allows you with ETL, ELT or replication.

Pros:

It is a cloud network that is elastic and scalable.

You can immediately access a range of data stores and a diverse collection of data transformation components.

By using the rich expression language of Xplenty, you can incorporate complex data preparation functions.

It offers a customized and flexible API component.

Cons:

There is no option for monthly subscription.

Pricing: It has a price model focused on subscriptions and can be tried for free for 7 days.

Modern Big Data Analytic Tools

CLOUDERA

CDH is a complete open-source Big Data Analytics tool and includes Apache Hadoop, Apache Spark, Apache Impala, and many more on its free distribution site. It enables you to acquire, store, manage, discover, model and distribute limitless data.

Pros:

Complete and accurate distribution.

The Hadoop cluster is very well managed by the Cloudera Manager.

Simple to deploy.

The administration is less complicated.

High security and administration

Cons:

Few complicated user interfaces like CM service charts.

Several suggested installation methods are confusing.

Pricing: Cloudera edition of CDH is a free Big Data Analytics tool. However, if you are interested in learning about the cost of the Hadoop cluster then the rate per node is between $1000 and $2000.

Modern Big Data Analytic Tools

R is one of the most comprehensive Big Data analytics tool for statistical analysis. The software ecosystem is open-source, free, multi-paradigm, and diverse. The programming languages are C, Fortran, and R. Most extensively used by statisticians and data miners; its use cases include data processing, data manipulation, analysis, and visualization.

Pros:

The greatest value of R is the immensity of the ecosystem package.

Unparalleled Graphics and charting features.

Cons: Its shortcomings include memory management, speed, and security.

Pricing: The shiny server and R studio IDE are free.

Modern Big Data Analytic Tools

cassandra

Apache Cassandra is free of cost Big Data analytics tools designed to handle large quantities of data across many commodity servers, offering high-availability. The open-source NoSQL DBMS uses CQL (Cassandra Structure Language) to interact with the database.

Pros:

There is no single failure point.

It manages huge data really quick.

It has log-structured storage and linear scalability.

Cons:

Extra troubleshooting and maintenance work is required.

It could have boosted the clustering.

There is no row-level locking feature.

Pricing: Its subscription starts from $49 Per node per month.

Modern Big Data Analytic Tools

KNIME is an abbreviation for Konstanz Information Miner, which is an open-source Big Data Analytics tool. It is used for enterprise reporting, integration, data mining, data analytics, and business intelligence. It supports operating systems such as Linux, and Windows X.

Pros:

Quick to use ETL

It is very well integrated with other technologies and languages.

Rich set of algorithms.

Workflows are highly functional and structured.

A lot of manual tasks are automated.

There are no problems with stability.

Simple to configure.

Cons:

It covers nearly the whole of RAM.

Might have enabled graph database integration.

Pricing: It’s a free Big Data Analytics tool.

Modern Big Data Analytic Tools

Datawrapper

Datawrapper is an open-source Big Data Analytics tool for data visualization. It enables its users to produce clear, accurate, and embedded charts easily. It is broadly used in newsrooms across the world.

Pros:

Operates exceptionally well on any type of device – smartphone, laptop, or tablet.

Rapid and interactive responses.

Excellent export and customization options.

Cons:

Has limited options for color palettes.

Pricing: It offers free service.

Modern Big Data Analytic Tools

mongoDB.

MongoDB is a contemporary alternative to databases. It’s one of the best Big Data Analytics tools for working on data sets that vary or change frequently or the ones that are semi or unstructured. Some of the best uses of MongoDB include storage of data from mobile apps, content management systems, product catalogs, and more. Like Hadoop, you can’t get started with MongoDB instantly. You need to learn the tool from scratch and be aware of working on queries.

Pros:

Supports various platforms and technologies.

No install and maintenance hiccups.

Robust and cost-effective.

Cons:

It has a limited analytics resource.

Pricing: The SMB and corporate versions of MongoDB are paid, and their rates are available upon request.

Modern Big Data Analytic Tools

Lumify is one of the open-source Big Data Analytics tools to analyze and visualize large data. This Big Data Analytics tool's key features include full-text search, 2-dimensional and 3-dimensional graphical viewings, automated templates, multimedia analysis, real-time project-or workplace collaboration, to name but a few.

Pros:

Scalable and secure

A dedicated full-time development team backs it.

Supports the cloud-based environment and works excellently with Amazon's AWS.

Pricing: It is a free Big Data Analytics tool.

Modern Big Data Analytic Tools

HPCC SYSTEMS®

HPCC is an abbreviation for **High-Performance Computing Cluster**. This open-source Big Data Analytics tool is a complete Big Data solution over a highly scalable supercomputing platform. HPCC is also known as DAS (Data Analytics Supercomputer) and was developed by LexisNexis Risk Solutions. Written in C++ and ECL(Enterprise Control Language), it is based on a The architecture that enables data parallelism, pipeline parallelism, and system parallelism.

Pros:

High performance due to the commodity computing clusters based architecture.

Enables parallel data processing.

Agile, robust and highly scalable.

Cost-effective and comprehensive

Pricing: It’s a free Big Data Analytics tool.

Modern Big Data Analytic Tools

APACHE STORM™

Storm is a cross-platform and open-source Big Data Analytics tool from Apache. Written in Java and Clojure, Backtype and Twitter are the developers of the storm. Several big brands like Yahoo, Alibaba, and The Weather Channel, to name a few are organizations that use Storm.

Pros:

There are many applications: real-time analysis, logging, ETL (Extract Transform Load), continuous computation, distributed RPC, machine learning.

Agile, reliable, and highly scalable.

Cons:

Difficult to understand and to use.

Have debugging complexity.

Pricing: It's a free Big Data Analytics tool.

Modern Analytic Tools

Modern Analytic Tools:
Current Analytic tools concentrate on three classes:

a) Batch processing tools
b) Stream Processing tools and
c) Interactive Analysis tools.

Modern Analytic Tools

Big Data Tools Based on Batch Processing:

Batch processing system :-

• Batch Processing System involves

– collecting a series of processing jobs and carrying them out periodically as a group (or batch) of jobs.

• It allows a large volume of jobs to be processed at the same time.

• An organization can schedule batch processing for a time when there is little activity on their computer systems, for example overnight or at weekends.

• One of the **most famous and powerful batch process-based Big Data tools is Apache Hadoop.**

It provides infrastructures and platforms for other specific Big Data applications.

Modern Analytic Tools

Stream Processing tools

- Stream processing – Envisioning (predicting) the life in data as and when it transpires.
- The key strength of stream processing is that **it can provide insights faster, often within milliseconds to seconds.**

– It **helps understanding the hidden patterns in millions of data records in real time.**

– It **translates into processing of data from single or multiple sources**

– in real or near-real time applying the desired business logic and emitting the processed information to the sink.

- Stream processing serves

– multiple

– resolves in today's business arena.

Real time data streaming tools are:

Modern Analytic Tools

a) **Storm**

- Storm is a ***stream processing engine without batch support,***
- a true ***real-time processing framework,***
- taking in a stream as an entire ‘event’ instead of series of small batches.
- *Apache Storm is a distributed real-time computation system.*
- It’s *applications are designed as directed acyclic graphs.*

Modern Analytic Tools

b) Apache flink

- Apache flink is

– an open source platform

– which **is a streaming data flow engine that provides communication fault tolerance and**

– **data distribution computation over data stream .**

– flink is a top level project of Apache flink is scalable data analytics framework that is fully compatible to hadoop .

– flink can **execute both stream processing and batch processing easily.**

– flink was **designed as an alternative to map-reduce.**

Modern Analytic Tools

c) Kinesis

– Kinesis as an out of the box **streaming data tool.**

– Kinesis **comprises of shards which Kafka calls partitions.**

For organizations that take **advantage of real-time or near real-time access to large stores of data,**

– Amazon Kinesis is great.

– Kinesis Streams solves a variety of streaming data problems.

– One common use is **the real-time aggregation of data which is followed by loading the aggregate data into a data warehouse.**

– Data is put into Kinesis streams.

– This ensures durability and elasticity.

Modern Analytic Tools

c) Interactive Analysis -Big Data Tools

• The **interactive analysis presents**

– the **data in an interactive environment,**

– allowing users to undertake their own analysis of information.

• **Users are directly connected to**

– the computer and hence can interact with it in real time.

• **The data can be :**

– reviewed,

– compared and

– analyzed

• ***in tabular or graphic format or both at the same time.***

Modern Analytic Tools

IA -Big Data Tools -

a) Google's Dremel is the google proposed an interactive analysis system in 2010. And named named Dremel.

– which is **scalable for processing nested data.**

– Dremel provides

• **a very fast SQL like interface to the data by using a different technique than MapReduce.**

• Dremel has a **very different architecture:**

– **compared with well-known Apache Hadoop, and**

– **acts as a successful complement of Map/Reduce-based computations.**

• **Dremel has capability to:**

– ***run aggregation queries over trillion-row tables in seconds***

– by means of:

•combining multi-level execution trees and

• columnar data layout.

Modern Analytic Tools

IA -Big Data Tools -

a) Google's Dremel is the google proposed an interactive analysis system in 2010. And named named Dremel.

– which is **scalable for processing nested data.**

– Dremel provides

• **a very fast SQL like interface to the data by using a different technique than MapReduce.**

• Dremel has a **very different architecture:**

– **compared with well-known Apache Hadoop, and**

– **acts as a successful complement of Map/Reduce-based computations.**

• **Dremel has capability to:**

– ***run aggregation queries over trillion-row tables in seconds***

– by means of:

•combining multi-level execution trees and

• columnar data layout.

Modern Analytic Tools

b) Apache drill

• **Apache drill is:**

– Drill is **an Apache open-source SQL query engine for Big Data exploration.**

– It is similar to Google's Dremel.

• For Drill, there is:

– **more flexibility to support**

• **a various different query languages,**

• **data formats and**

• **data sources.**

• Drill is designed from the **ground up to:**

– **support high-performance analysis on the semi-structured and**

– **rapidly evolving data coming from modern Big Data applications.**

• Drill **provides plug-and-play integration with existing Apache Hive and Apache HBase deployments.**

thank
you!