Name : Tejas Redkar
Roll No : PC - 44, Panel - C
PRN : 10322 10937
BDT - 2, Batch - I [BDT Roll No. 22]

Big Data Technologies - I Lab
Assignment no : 01

* Problem Statement :

Installation of Big Data tools

* Objectives :

1) To learn concepts of Bigdata.
2) To learn how to install & use different big data tools.

* Theory :

Bigdata : It is the combination of the structured, semi-structured & unstructured data collected by organization that can be mined for organization. & used in machine learning projects, predictive modelling & other advanced analytics applications. Systems that process & store big data have become a common component of data management architectures in organizations, combined with tools that support big data analytics uses.

It is often characterized by the three V's:

- the large volume of data in many environments
- the wide variety of data types frequently stored in big data systems.
- the ~~variety~~ velocity at which much of the data is generated, collected & processed.

Examples of big data are transaction processing systems, customer databases, documents, emails, medical records, mobile apps & social networks.

* Big Data Tools:

A big data tool is a software that extracts information from various complex data types & sets, and then processes these to provide meaningful insights. Here are a few:

a) Hadoop:

This open source software frame work processes data sets of big data with the help of the MapReduce programming model written in Java it provides cross-platform support. This is one of the popular big data tools used by most Fortune 50 companies including Amazon Web Services, IBM, Intel & Facebook.

- It is highly scalable, provides fast access to data.
- Offers a robust system

- Offers flexibility.

b) Hive & Pig

Hive: It is built on top of Hadoop & is used to process structured data in Hadoop. It was developed by Facebook. It provides various types of querying language which is frequently known as Hive Query language.

Pig: It is used for analysis of a large amount of data. It is abstracted over MapReduce. It is used to perform all kinds of data manipulation operations in Hadoop. It provides the Pig-Latin language to write the code that contains many inbuilt functions like join, filter etc.

c) Mongo DB

It came into limelight in 2010 & is a free, open-source platform & a document-oriented database that is used to store high volume of data. It uses collections & documents for storage & its document consists of key-value pairs which are considered a basic unit of MongoDB

Features of MongoDB:

- Written in C++: It's a schema-less DB & can hold varieties of document inside.

- Simplifies stack : With the help of Mongo, a user can easily store files without any disturbances.

d) Spark.
It is another frame work that is used to process data & perform numerous tasks on a large scale. It is also used to process data via multiple computers with the help of distributing tools. It allows users to run in their preferred language. Real-time processing : Spark can handle real-time streaming

e) AWS
It provides the broadest selection of analytic services that fit all your data analytics needs & enables organization of all sizes & industries to reinvent their ~~busi~~ business with data.

f) Snowflake
It's easy to use cloud data platform with data warehouse as a service & cloud data which provides a cloud based single solution to Big Data management needs.

* Platform : 64-bit open source Linux / Windows.

* Conclusion: Hence, I learned different tools of Bigdata technologies

CLASSMATE
Date :
Page :

* FAQ's

Q1) Explain V's in Big Data

Ans 1) Volume : As the term implies, big data analytic entail handling & analyzing vast amount of data.

2) Velocity : It denotes the speed at which data is generated

3) Variety : It refers to the diversity of data types & sources.

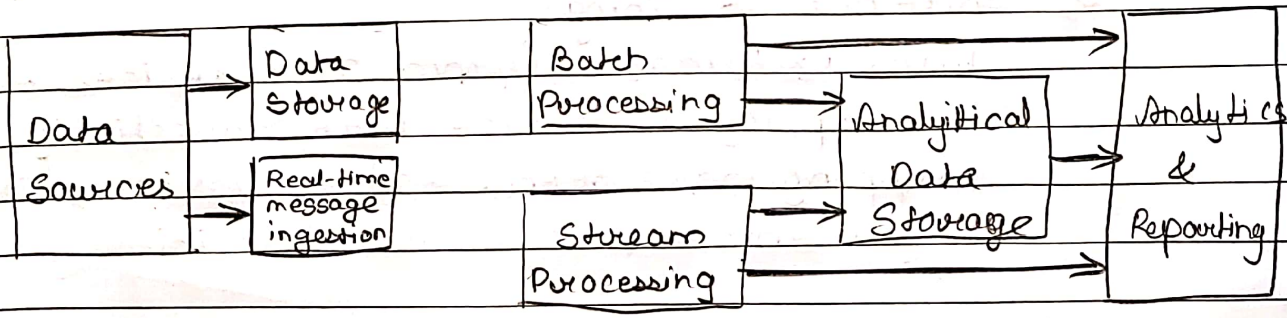
4) Variability : Big data often contain noisy & incomplete data points, which can obscure valuable insights.

5) Veracity : It performs to the accuracy & the authenticity of the data

6) Value : A successful big data analytics strategy must generate value.

7) Visualization : It plays a vital role in data analytics, as it involves presenting the analyzed data in a visually comprehensible manner.

Q2) Explain Architecture of Big Data Systems



Big Data Architecture is the foundation for big data analytics. It is the overarching system used to manage large amounts of data so that can be analyzed for business purposes, steer data analysis analytics & provide an environment in which big data analytic tools can extract vital business information from otherwise ambiguous data. The big data architecture framework serves as a reference blueprint for big data infrastructure & solutions logically defining how big data solutions will work, the component that will be used, how information will flow & security details.

Q.3) Explain Bigdata applications in any three domination domains.

Ans i) Media & Entertainment:

Big Data provides actionable points of information about millions of individuals. Now, publishing environments are tailoring advertisements & contents to appeal consumers. These insights are gathered through various data mining activities.

2) Internet of Things

Data extracted from IOT devices provides a mapping of device interconnectivity. Such mapping have been used by various companies & governments to increase efficiently.

3) Government:

The use & adoption of Big Data within governmental processes allow efficiencies in terms of cost productivity & innovation.

05/09/23