Big Data Analytics

- SCSA1603

Name: Mohnish Devaraj
Reg No: 39110636
Section: C1

Assignment-1

PART-A

(1) Big Data

(2) 10^{15} byte

(3) d. All of the Above

(4) d. All of these

(5) d. All of the Above

PART-B

(1) **Big Data**

Big Data is a collection of data that is huge in volume, yet growing exponentially with time. It is data with so large size and complexity that none of traditional data management tools can store it or process it efficiently. Big data is also a data but with huge size.

(2) The 7Vs of Big Data: Volume, Velocity, Variety, Variability, Velocity, Veracity, Value and Visibility.

(3) Intelligent Data Analysis

Intelligent Data Analysis (IDA) is an interdisciplinary study that is concerned with the extraction of usefull Knowledge from data, drawing techniques from a variety of fields, such as artifical intelligence, high-performance computing, patterns recognition and statistics.

(4) The following are the Big data Analytics tools:

-> R and Python
-> Microsoft Excel
-> Tableau
-> Rapid Miner
-> KNIME
-> Power BI
-> Apache Spark
-> Qilk View
-> Talend
-> Splunk

(5) Sampling distribution is a statistics that determines the probability of an event based on data from a small group within a large population. Its primary purpose is to establish representative results of small samples of a comparatively large population. Since the population is too large to analyze, the smaller group is selected

and repeatedly sampled, or analyzed. the gathered data, or statistic, is used to calculate the likely occurence, or probability, of an event.

PART-C

① Big Data Analytics is a process that enables data scientists to make something out of the stack of big data generated. This analysis of big data is done using some tools that we reckon as big data analytics tools.

a) Altamira Lumify

Lumify is a big data fusion, analysis and visualization platform. Like all big data analytics tools, it too enables you to understand connections and explore the relationship between your data. Lumify is consider-ed as a good big data analytics tool because it facilitates its users to get a set of analytics options that include graph visualizations, full-text faceted search, dynamic histograms, interactive, geospatial views and collaborative work spaces that can be shared in real-time. Lumify offers both 2D and 3D graph visualization with automatic layouts. It also provides a plethora of options to analyze the links between different entities in a graph.

b) R - Programming

R - Programming is a domain-specific programming language specifically designed for statistical analysis, scientific computing, and data visualization using R Programming. R-Programming software helps data scientists to create statistics engines that can provide better and precise insights due to relevant and accurate data collection. The tools exhibit some feature that are: Effective data handling and storage facility, It provides tenacious and integrated tools for data analysis. R produces plots and graphics that are ready for publication.

c) Apache Hadoop

Apache Hadoop is an open-source software frame work for storing data and running application on clusters of commodity Hardware. Apache Hadoop is a framework that consists of a software ecosystem. Hadoop Distributed File System or HDFS and Map Reduce are the two primary components of Hadoop.

d) Mango DB

MangoDB is a document-oriented NOSQL database used to store high volumes of data. MangoDB is well known for its roburstness and this makes MangoDB different from Hadoop. Each database in Mango DB contains collections that in turn contain documents. However, the size, content and number of fields vary from document to document.

e) Rapid miner

Rapid miner is a software platform that is built for analysts who like to integrate data prep, machine learning and predictive model deployment. the icing on the cake here is that it is an open-source software tool, free of charge, for data and text mining. Rapid Miner offers the most powerful and intuitive graphical user interface for the design of the analysis process.

f) Apache Spark

Apache Spark is one of the most powerful open source big data analytics tools. It is a data processing frame work that can quickly possess very large data sets. It can also distribute data processing tasks

across multiple computers, either on its own or in conjunction with other distributed computing tools. Apache spark features in-built for streaming, SQL, machine learning, and graph processing support and earns the site as the speediest and common generator for big data transformation.

g) Microsoft Azure

Microsoft Azure formerly known as Windows Azure, is a public cloud computing perform handled by Microsoft. It provides a range of services that include computing, analytics, storage and networking. Windows Azure provides big data cloud offerings in two categories, standard and premium. It provides an enterprise-scale cluster for the organization so that they can run their big data workloads. The platform aims to offer information in real time in a way that is easy to manage even when used on the most advanced application.

h) Zoho Analytics

Zoho Analytics is a BI and Data Analytics software platform that helps its users to visually analyze data,

create visualization, and get a better and in-depth understanding of raw data. It allows its users to integrate multiple data sources that may include business applications, databases, cloud drives and more. It helps users generate dynamic, highly customizable, and actionable reports.

i) Xplenty

Xplenty is a cloud-based ETL solution that provides simple visualized simple visualized data pipelines. These pipelines allow data to flow automatically across sources and destinations. Xplenty Data has powerful on platform transformation tools that allow you to clean, normalize and transform data. whilst adhering to compliance best practices. The platform exhibits some features that make it a user-friendly platform: Easy Data transformation, REST API for connecting to any dat Source.

ii) Splice Machine

Splice Machine is a scale-out SQL Rotational Database Management System (RDBMS). It has ACID transaction, in memory analytics, and in-database machine learning.

combined. The big data analytics tools can scale from a few to thousands of nodes enabling applications at every scale. The Splice Machine optimizer automatically evaluates every query to the distributed HBase regions. It offers low latency row-based storage. Splice Machine's dual model leverages columnar external tables on cost-effectives storage on cloud storage, HDFS, or local files as Parquet, ORC, or Avro files with append-only functionalities. Splice Machine analytical computation maintains ACID properties with a special integration to our underlying row-based storage.