(Please start writing assignment/ test from here)

Q1 a Explain various components in Hadoop

A1 a Components are:

i) HDFS (Hadoop Distributed File System)

- HDFS is responsible for storing large data sets of structured or unstructured data across various nodes and metadata is kept as log files
- Two core components are:
  i) Name node
  ii) Data node
- HDFS works at heart of the system

ii) YARN (Yet Another Resource Navigator)

- YARN helps to manage resources across the clusters.
- It performs scheduling and resource allocation for Hadoop
- Three major components are:
  i) Resource Manager
  ii) Nodes Manager
  iii) Application Manager

iii) Map Reduce
- Map Reduce uses distributed and parallel algorithms to carry over the processing's logic and helps to write applications which transform big data sets into Manageable one.
- Two functions
  i) Map(): sort & filter data and organize in group
  ii) Reduce(): Aggregate the mapped data

iv) PIG
v) HIVE
vi) Mahout
vii) Apache Spark
viii) Apache H Base

Q1 b. Explain Apache Spark

A1ba. Apache Spark is an open source, distributed processing system used for big data workloads.

b. It handles all process consumptive tasks like batch processing, interactive or iterative real-time processing, visualizationetc.

c. It consumes in memory resources hence, faster than prior in optimization

d. Spark is best suited for real time data whereas Hadoop is suited for batch processing.

Q2.a. Explain any four challenges in data visualization

A2.a. Challenges are:

i) Diversity & heterogenity in big data creates a big problem while visualizing data

ii) Analysis speed is most challenging in BDA.

iii) Handling Big Data scalability, cloud computing and advanced GUI is a real challenge

iv) Data is unstructured, to visualize the data, we have to use metadata

v) Providing huge parallelization is a challenge in Big Data Visualization

vi) High complexity & dimensionality due to huge amount of data.

Q2.b. Analyze Big Data Visualization methods

A2.b. Big Data Visualization methods

i) Symbol Maps: Symbols on such maps differ in size, which makes them easy to compare

ii) Line Charts: a. Line charts allow looking at the behaviour of one or several variables over time and identifying the trends.
b. Gives shape to the change taking place, and can be used to show the volatility

iii) Pie charts: a. It shows the components of the whole.
b. The difference lies in the sources from which these companies take raw data.

iv> Bar charts : a. Bar charts allow comparing the values of different variables.
b. Companies can analyze their sales by category, costs etc.

v> Heat Maps : a. A heat map is a data viz. method showing relationship between 2 measure and rating information
b. Varying colours are used to show ratings.

vi> Word Cloud : a. A word cloud demonstrates how frequently a word appears in a block of text by connecting size with frequency
b. Bigger the word in the cloud, the more larger times it appears in the text.