BRAINWARE UNIVERSITY

Term End Examination 2023-2024
Programme – B.Tech.(CSE)-DS-2021
Course Name – Big Data and Analytics
Course Code - PEC-CSD601A
(Semester VI)

Full Marks : 60 **Time : 2:30 Hours**

[The figure in the margin indicates full marks. Candidates are required to give their answers in their own words as far as practicable.]

Group-A

(Multiple Choice Type Question) 1 x 15=15

1. *Choose the correct alternative from the following :*

(i) Select stream in the context of stream computing

a) A flowing river
b) A sequence of data elements over time
c) A static dataset
d) A graphical representation

(ii) Select the purpose of the Stream Data Model

a) To represent static data
b) To model streaming algorithms
c) To create visualizations
d) To store data in a database

(iii) Recall an example of a moment in stream computing

a) Mean
b) Streamflow
c) Visual representation
d) Sorting order

(iv) Select the primary advantage of using a decaying window in stream processing

a) Improved processing speed
b) Efficient memory usage
c) Real-time visualization
d) Increased window size

(v) Select Big Data Platform

a) A system for processing large datasets
b) A software for small-scale data analysis
c) A platform for social media management
d) A tool for graphic design

(vi) Select a challenge of big data platforms

a) Data security
b) Limited data sources
c) Slow processing speed
d) Single-user access

(vii) Explain the design of HDFS.

a) Distributed storage
b) Relational database
c) Client-server model
d) Peer-to-peer network

(viii) Collect the characteristics of a good MapReduce algorithm.

a) Scalable, Fault-tolerant, Efficient resource utilization
b) Memory-intensive, Serial processing
c) Sequential, I/O bound
d) Highly coupled, Monolithic

(ix) Collect the methods of data replication management in Hadoop's distributed file system.

a) Replication factor configuration, Rack awareness
b) Compression algorithms used
c) Encryption keys management
d) Data serialization formats

(x) Select the right option from following statements that is true about Apache Hive.

a) Hive queries are directly executed on HDFS.
b) Hive uses Pig Latin for data processing.
c) Hive provides a relational database-like interface.
d) Hive is optimized for real-time data ingestion.

(xi) Choose the correct option from following that is NOT a component of Apache Hive.

a) HiveQL
b) Hive Server
c) Hive Driver
d) Hive Master

(xii) Identify the proper option where IBM InfoSphere Streams primarily focus on.

a) Real-time analytics on streaming data
b) Batch processing of large datasets
c) Interactive querying and analysis
d) Data warehousing

(xiii) Choose the proper option for a key characteristic of IBM InfoSphere Streams.

a) It is optimized for batch processing.
b) It uses MapReduce for data processing.
c) It supports high-throughput, low-latency data processing.
d) It is primarily used for graph analytics.

(xiv) Select a key characteristic of the stream data model.

a) Storage of static, discrete entities
b) Continuous flows of data
c) Batch processing of data
d) Limited processing capabilities

(xv) Select technique is used for handling out-of-order events or late arrivals in stream processing systems.

a) Checkpointing
b) Replication
c) Stream partitioning
d) Time windows

Group-B

(Short Answer Type Questions) 3 x 5=15

2. State how do organizations derive value from unstructured data within big data environments. (3)
3. Develop an understanding of the design of HDFS-Java interfaces. (3)
4. Describe the purpose of filtering streams in stream computing and provide an example scenario. (3)
5. Explain how does Stream Computing differ from Batch Processing? (3)
6. Evaluate the history of Hadoop, and how did it evolve? (3)

OR

Analyze the key components of Hadoop, and how do they contribute to data processing? (3)

Group-C

(Long Answer Type Questions) 5 x 6=30

7. Observe how do organizations derive value from unstructured data within big data environments. (5)
8. Identify key challenges and ethical considerations in Real-Time Sentiment Analysis. (5)
9. Describe the Key Components of Stream Data Model. (5)
10. Analyze the role of shuffle and sort phase in MapReduce. (5)
11. Summarize the process how ZooKeeper ensures data consistency and reliability in distributed systems. (5)
12. Explain the concept of Hive in big data, and analyze the process how it fit into the Hadoop ecosystem. (5)

OR

Explain the key components of Hive. (5)

****