Name:

Enrolment No:

UPES

End Semester Examination, December 2024

Course: B.Tech Bio-Medical Engineering | **Semester : Fifth**

Program: Big Data Analytics | **Duration : 3 Hours**

Course Code: CSBA3012 | **Max. Marks: 100**

Instructions:

S. No.	Section A Short answer questions/ MCQ/T&F (20Qx1.5M= 30 Marks)	Marks	COs
Q	**For the question below write your answer as True or false**		
1)	The Hadoop Distributed File System (HDFS) is responsible for storing metadata about the files in Hadoop.	**1.5**	**1**
2)	MapReduce divides data into smaller tasks to improve processing efficiency.	**1.5**	**1**
3)	Flume is mainly used to process data in real time within a Hadoop cluster.	**1.5**	**1**
4)	In Hadoop, the DataNode stores the metadata, while the NameNode stores the actual data.	**1.5**	**1**
5)	HDFS Administration includes configuring and monitoring NameNode and DataNode.	**1.5**	**1**
6)	Jaql is designed for querying semi-structured and unstructured data.	**1.5**	**1**
7)	Pig is primarily used for structured data management.	**1.5**	**1**
8)	Hive enables SQL-based queries for analyzing Big Data within the Hadoop ecosystem.	**1.5**	**1**
9)	Only direct batch processing can be performed on Hadoop.	**1.5**	**1**
10)	Jaql does not support handling data output formats.	**1.5**	**1**
11)	The SPL (Streams Processing Language) is specifically designed for real-time data streaming.	**1.5**	**1**
12)	MapReduce tasks are always executed in a specific sequential order.	**1.5**	**1**
13)	Pig allows data loading, transformation, and storing functions.	**1.5**	**1**
14)	Hive supports schema-on-read, allowing schema definition only at data analysis.	**1.5**	**1**
15)	In Hadoop, HDFS is optimized for reading data sequentially rather than randomly.	**1.5**	**1**

16)	Pig Latin is the primary language used in Hive for data processing.	**1.5**	**1**
17)	Adapter Operators in SPL assist in linking and transforming different data formats.	**1.5**	**1**
18)	SPL’s windowing function enables the analysis of data in streams over time intervals.	**1.5**	**1**
19)	Business intelligence in Hadoop often uses indirect batch processing for data insights.	**1.5**	**1**
20)	Timing and coordination in SPL are important for managing the flow of streaming data.	**1.5**	**1**
Section B **(4Qx5M=20 Marks)**			
Q 1	Explain the difference between relational operators and utility operators in SPL.	**5**	**1**
Q 2	Compare and contrast structured and unstructured data, providing specific examples for each	**5**	**2**

Q 3	Solve the following Hive query, Consider a scenario where we have a large dataset containing customer transactions.
	SELECT customer_id, SUM(amount) AS total_spent
	FROM transactions
	GROUP BY customer_id
	ORDER BY total_spent DESC
	LIMIT 10;
		5	**1**

Q 4	Give an example query in Jaql that utilizes multiple data types.	**5**	**3**
Section C **(2Qx15M=30 Marks)**			
Q 1	What is Map Reduce? Explain working of various phases of Map Reduce with appropriate example and diagram.	**15**	**3**
Q 2	List the difference between partitioning and bucketing in Hive? When would you choose one over the other?	**15**	**3**
Section D **(2Qx10M=20 Marks)**			
Q 1	Explain the key features of Hive that differentiate it from other big data processing tools?	**10**	**2**
Q 2	Summarize the common data types in Jaql, and how are they used to handle Big Data queries? Provide an example.	**10**	**2**