Ref. No.: Ex/IT/PE/H/T/325B/2024

JADAVPUR UNIVERSITY
DEPARTMENT OF INFORMATION TECHNOLOGY
B.E. INFORMATION TECHNOLOGY, THIRD YEAR
SECOND SEMESTER EXAM 2024
SUBJECT: BIG DATA (HONS.)

Full Marks: 100 **Time: 3Hrs**

CO1 : (20)	**Answer any five** **5 X 4 =20** Q.1 i) What is the function of Job tracker and Task Tracker in HDFS? ii) What do you mean by a Hadoop Cluster and explain the various services hosted by the master node of a Hadoop Cluster. iii) What are the different modes in which Hadoop can be installed and what is the use of each mode from application and developer point of view? iv) Explain the uses of Name node, Data node and Secondary Name node in Hadoop Distributed File system. v) How does YARN allocate containers to the jobs? vi) Name the services running in slave part of Hadoop architecture.
CO2 : (20)	**Answer any four** **4 X 5 =20** Q.2 i) Consider a scenario where a company wants to analyze sales data alongside customer demographic data to understand buying patterns. Discuss how MapReduce can perform the relational join operation on these datasets. ii) Write a Map Reduce algorithm to get the Dot Product of two Large Vectors, assuming only non-zero elements of those vectors are given in input files and output file should show only non-zero entries (assuming two vectors are same size) v1=[5 4 0 1 2] and v2=[4 2 1 0 6]. iii) Explain the role of driver code, mapper code and reducer code within a map reduce program model by a suitable example. iv) With a neat diagram describe the phases of Map Reduce Job processing. v) Explain how data is partitioned before it is sent to the reducer if no custom partitioner is defined in Hadoop.
CO3 : (30)	**Answer any two** **15 X 2 =30** Q.3 i) a) State the procedural differences between sharding and replication. b) Explain the concept of quorum based write consistency in Cassandra. c) How the BASE properties of NoSQL databases do impact the design and scalability of distributed systems compared to the ACID properties of traditional relational databases? d) How can downtime be reduced in NoSQL databases? **4+5+4+2=15**

	ii) a) State how does MongoDB handles the issue of No schema, No DDL, No joins, and No transactions. b) Explain briefly the following features in MongoDB: Hierarchical Objects, MongoD, MongoS, Collections, BSON documents. c) Can you explain the concept of aggregation in MongoDB? **5+7+3 = 15** iii) a) How can organizations effectively scale and manage HBase region Servers to accommodate the growth of their data stores, while maintaining fast data lookup and fault tolerance across the cluster? b) What is the Hierarchy of Tables in Apache HBase? c) What are the operational commands of HBase? Mention few. d) Define compaction in HBase? e) How does sorted maps work in HBase **5 X 3 =15**
CO4 : (30)	**Answer All** **15 X 2 =30** Q.4 i) a) Consider a table named "website_traffic" in Hive, which contains records of website traffic with the following schema: *session_id:* Unique identifier for each session *browser_type:* Type of browser used by the visitor *session_duration:* Duration of the session in seconds Formulate a Hive query to calculate the average session duration for each browser type used by visitors, filtering out sessions shorter than 30 seconds, and displaying the results along with the browser types. b) Can you explain the role and purpose of Thrift Server, Metastore and Managed tables in context of Apache Hive? c) Where does the data of a Hive table get stored? **6+6+3 =15** ii) a) How does Spark leverage lineage information to recover from faults during computation? b) Mention few Spark Control operations with their functionalities. c) How does the DAG scheduler optimize task scheduling for performance improvement? **5+5+5 =15**