SIDDHARTH INSTITUTE OF ENGINEERING & TECHNOLOGY::PUTTUR
(AUTONOMOUS)
Siddharth Nagar, Narayanavanam Road, Puttur – 517583

QUESTION BANK (DESCRIPTIVE)

Subject with Code: BIG DATA TECHNOLOGIES (25MC9125) **Course & Branch**: MCA
Year & Sem: II-MCA & I-Sem **Regulation:** R25

UNIT –I

Introduction and Fundamentals of Big Data

1	Define Big Data and explain its key characteristics of Volume, Velocity, Variety, and Veracity.	**[L1][C01]**	**[12M]**
2	a) Explain why Big Data has become important for modern organizations, highlighting its key drivers.	**[L2][C01]**	**[6M]**
	b) Discuss mobile business intelligence and crowd-sourcing analytics with suitable examples.	**[L6][C01]**	**[6M]**
3	a) Explain the concept of unstructured data and describe its role in Big Data analytics.	**[L2][C03]**	**[6M]**
	b) Classify the application of Big Data in risk assessment and management.	**[L4][C01]**	**[6M]**
4	a) Describe the role of Big Data in web analytics with a suitable example.	**[L1][C05]**	**[6M]**
	b) Explain how Big Data techniques help in fraud detection and prevention.	**[L2][C05]**	**[6M]**
5	a) Visualize the role of Big Data in credit risk management with an example.	**[L1][C05]**	**[6M]**
	b) Discuss how Big Data supports algorithmic trading in financial markets.	**[L6][C05]**	**[6M]**
6	Explain the applications of Big Data in the healthcare industry.	**[L2][C05]**	**[12M]**
7	Evaluate the role of Big Data in advertising, with focus on targeted advertisement.	**[L5][C06]**	**[12M]**
8	Discuss various Big Data technologies used for storing, processing, and analyzing large datasets.	**[L6][C05]**	**[12M]**
9	a) Express open source technologies that support the Big Data ecosystem.	**[L6][C03]**	**[6M]**
	b) Explain the relationship between cloud computing and Big Data.	**[L2][C04]**	**[6M]**
10	Determine the concepts of inter-firewall and trans-firewall analytics in the context of Big Data security.	**[L3][C03]**	**[12M]**

UNIT –II

Introduction to NoSQL Databases and Big Data Models

1	Derive the concept of NoSQL databases and discuss why they are preferred over relational databases for Big Data.	**[L3][C02]**	**[12M]**
2	Define the document data model and compare it with the key-value data model.	**[L1][C03]**	**[12M]**
3	a) Explain the graph database model with a suitable example.	**[L2][C02]**	**[6M]**
	b) Discuss schema-less databases and their advantages over schema-based systems.	**[L6][C03]**	**[6M]**
4	a) Evaluate the concept of materialized views and their use in NoSQL databases.	**[L4][C02]**	**[6M]**
	b) Discuss the various distribution models used in NoSQL databases.	**[L6][C02]**	**[6M]**
5	Compute the concept of sharding and how it helps in scaling databases.	**[L3][C02]**	**[12M]**
6	a) Explain peer-to-peer replication and compare it with master-slave replication.	**[L2][C03]**	**[6M]**
	b) Discuss the combined use of sharding and replication in distributed databases.	**[L6][C02]**	**[6M]**
7	a) Discuss the idea of relaxing consistency and its trade-offs (CAP theorem).	**[L6][C03]**	**[6M]**
	b) Compare the concept of version stamps and their role in maintaining consistency.	**[L5][C04]**	**[6M]**
8	a) Analyze the map-reduce concept for processing large datasets.	**[L4][C02]**	**[6M]**
	b) Explain MapReduce calculations are composed for complex data processing tasks.	**[L2][C02]**	**[6M]**
9	Discuss partitioning and combining operations in a MapReduce computation.	**[L6][C02]**	**[12M]**
10	Compare the aggregate data model with the relational data model, highlighting the advantages of NoSQL for Big Data applications.	**[L2][C02]**	**[12M]**

UNIT –III

Hadoop Ecosystem and Data Processing

1	a) Explain various data formats used in Hadoop for storing and processing large datasets.	**[L2][C03]**	**[6M]**
	b) Discuss how data is analyzed using Hadoop, describe the overall processing approach.	**[L6][C01]**	**[6M]**
2	Describe Hadoop streaming and explain how it enables writing MapReduce programs in non-Java languages.	**[L2][C04]**	**[12M]**
3	a) Define Hadoop pipes and its role in interfacing with C++ programs.	**[L1][C03]**	**[6M]**
	b) Describe the basic concepts of HDFS, including roles of NameNode and DataNode.	**[L2][C03]**	**[6M]**
4	a) Explain the design principles of the Hadoop Distributed File System in HDFS.	**[L2][C03]**	**[6M]**
	b) Describe the data flow process during file read and write operations in HDFS.	**[L2][C03]**	**[6M]**
5	Illustrate the concept of serialization and its importance in Hadoop data processing.	**[L3][C05]**	**[12M]**
6	a) Analyze Avro and explain its role as a serialization framework in Hadoop.	**[L4][C05]**	**[6M]**
	b) Distinguish the structure and use of Avro schemas with a suitable example.	**[L5][C05]**	**[6M]**
7	a) Discuss various file-based data structures used in Hadoop, such as SequenceFile.	**[L6][C05]**	**[6M]**
	b) Analyze the concept of data integrity in Hadoop and the mechanisms used to ensure it.	**[L4][C03]**	**[6M]**
8	Compare different file-based data structures in Hadoop in terms of performance and use cases.	**[L2][C05]**	**[12M]**
9	a) Explain the MapFile data structure and its advantages over SequenceFile.	**[L2][C04]**	**[6M]**
	b) Discuss the codecs supported by Hadoop for compression and decompression.	**[L2][C04]**	**[6M]**
10	Discuss the overall Hadoop I/O pipeline covering data integrity, compression, and serialization with a real-time example.	**[L6][C03]**	**[12M]**

UNIT –IV

MapReduce Programming and Workflow Execution

1	a) Explain the overall MapReduce workflow with a neat diagram.	**[L2][C04]**	**[6M]**
	b) Describe the various stages involved in executing a MapReduce program.	**[L1][C04]**	**[6M]**
2	Discuss the role of test data and local tests in validating MapReduce programs.	**[L6][C04]**	**[12M]**
3	a) Describe the classic MapReduce architecture with its components.	**[L2][C04]**	**[6M]**
	b) Explain the role of JobTracker and TaskTracker in classic MapReduce.	**[L2][C04]**	**[6M]**
4	Compare classic MapReduce and YARN, highlighting the improvements introduced by YARN.	**[L2][C04]**	**[12M]**
5	a) Derive the types of failures that can occur in classic MapReduce and how they are handled.	**[L3][C04]**	**[6M]**
	b) Discuss failure handling mechanisms in YARN.	**[L6][C04]**	**[6M]**
6	Explain the concept of job scheduling in Hadoop and describe different scheduling algorithms (FIFO, Fair, Capacity).	**[L2][C04]**	**[12M]**
7	a) Discuss the shuffle-and-sort phase of MapReduce in detail.	**[L6][C04]**	**[6M]**
	b) Explain task execution in MapReduce, covering the concept of speculative execution.	**[L2][C04]**	**[6M]**
8	a) Evaluate different input formats supported by MapReduce with examples.	**[L5][C04]**	**[6M]**
	b) Differentiate output formats supported by MapReduce with examples.	**[L4][C04]**	**[6M]**
9	a) Discuss how a custom Input Format can be created and used in a MapReduce program.	**[L6][C02]**	**[6M]**
	b) Explain the role of Combiner and Partitioner in optimizing a MapReduce job.	**[L2][C02]**	**[6M]**
10	Discuss a real-world use case and design a complete MapReduce workflow (mapper, reducer, combiner) and to solve it.	**[L6][C06]**	**[12M]**

UNIT –V

Big Data Storage Solutions

1	Discuss the implementation architecture of HBase, covering RegionServers and HMaster.	**[L6][C05]**	**[12M]**
2	a) Derive various HBase client APIs used for accessing data.	**[L3][C05]**	**[6M]**
	b) Illustrate CRUD operations in HBase with suitable examples.	**[L2][C05]**	**[6M]**
3	a) Discuss HBase praxis, covering schema design considerations and best practices.	**[L6][C05]**	**[6M]**
	b) Evalute the Cassandra data model with its key components.	**[L5][C05]**	**[6M]**
4	a) Illustrate Cassandra architecture and operations with suitable examples using CQL commands.	**[L6][C05]**	**[6M]**
	b) Explain various Cassandra client APIs used for accessing data.	**[L2][C05]**	**[6M]**
5	How Cassandra integrates with Hadoop for Big Data processing, and compare HBase and Cassandra data models.	**[L2][C05]**	**[12M]**
6	a) Distinguish how tables and databases are created and managed using HiveQL DDL.	**[L4][C06]**	**[6M]**
	b) Explain HiveQL data manipulation language (DML) commands with examples.	**[L2][C06]**	**[6M]**
7	Define how data is loaded and inserted into Hive tables using DML statements.	**[L1][C06]**	**[12M]**
8	a) Explain various HiveQL query operations, including joins and subqueries, with examples.	**[L2][C06]**	**[6M]**
	b) Explain HiveQL data definition language (DDL) commands with examples.	**[L2][C06]**	**[6M]**
9	Discuss partitioning and bucketing in Hive and their role in query optimization.	**[L6][C06]**	**[12M]**
10	Compare Hive with a traditional RDBMS and explain scenarios in which Hive is the preferred solution for Big Data analytics.	**[L2][C06]**	**[12M]**

Prepared by**: **Mr. P Balaji**, **Associate Professor**, **Dept. of MCA, SIETK.