

SRM VALLIAMMAI ENGINEERING COLLEGE

(An Autonomous Institution)

SRM Nagar, Kattankulathur – 603 203

DEPARTMENT OF CYBER SECURITY

QUESTION BANK



VII SEMESTER - IV YEAR

PCY305 – BIG DATA SECURITY

Regulation – 2023

Academic Year: 2026 – 2027 (ODD)

Prepared by

Mr. S.Giridharan, Assistant Professor / CYS



SRM VALLIAMMAI ENGINEERING COLLEGE



SRM Nagar, Kattankulathur-603203
DEPARTMENT OF CYBER SECURITY

QUESTION BANK

SUBJECT : PCY305– BIG DATA SECURITY
SEM / YEAR : VII SEMESTER/ IV YEAR

UNIT –I INTRODUCTION TO BIG DATA

Introduction – Distributed File System – Big Data and its importance – Four Vs – Drivers for Big Data – Big Data Analytics – Big Data Applications – Algorithms using MapReduce – Matrix–Vector Multiplication by MapReduce.

UNIT –I [PART-A]

Q.No	Questions	Competence	Level
1	Define Big Data.	Remember	BTL-1
2	What is a Distributed File System?	Remember	BTL-1
3	State the importance of Big Data in modern computing.	Understand	BTL-2
4	List the Four Vs of Big Data.	Remember	BTL-1
5	Define Volume in Big Data.	Remember	BTL-1
6	What is meant by Velocity in Big Data?	Remember	BTL-1
7	Explain the term Variety in Big Data.	Understand	BTL-2
8	What is Veracity in the context of Big Data?	Remember	BTL-1
9	List any four drivers for Big Data.	Remember	BTL-1
10	Define Big Data Analytics.	Remember	BTL-1
11	State two objectives of Big Data Analytics.	Understand	BTL-2
12	List any four applications of Big Data.	Remember	BTL-1
13	What is MapReduce?	Remember	BTL-1
14	Define the Map phase in the MapReduce programming model.	Remember	BTL-1
15	Define the Reduce phase in MapReduce.	Remember	BTL-1
16	What is meant by a key–value pair in MapReduce?	Understand	BTL-2
17	What is Matrix–Vector Multiplication?	Remember	BTL-1
18	Why is Matrix–Vector Multiplication suitable for MapReduce?	Understand	BTL-2
19	Distinguish between traditional data processing and Big Data processing.	Understand	BTL-2
20	Summarize the role of MapReduce in Big Data processing.	Understand	BTL-2

UNIT –I [PART-B]

Q.No	Questions	Marks	Competence	Level
1	Explain the concept of Big Data and discuss its importance in handling modern data-intensive applications.	16	Apply	BTL-3
2	Analyze the characteristics of Big Data using the Four Vs with suitable examples.	16	Analyze	BTL-4
3	Discuss the major drivers responsible for the rapid growth of Big Data and evaluate their impact on organizations.	16	Evaluate	BTL-5
4	Explain the architecture and working principles of a Distributed File System and analyze its advantages in Big Data environments.	16	Analyze	BTL-4
5	Apply Big Data Analytics concepts to explain how organizations derive meaningful insights from large-scale datasets.	16	Apply	BTL-3

6	Analyze various real-world applications of Big Data across different domains and justify their significance.	16	Analyze	BTL-4
7	Explain the MapReduce programming model with a detailed illustration of its execution process.	16	Apply	BTL-3
8	Analyze the roles of the Map and Reduce phases in processing large datasets efficiently.	16	Analyze	BTL-4
9	Apply the MapReduce paradigm to solve a large-scale data processing problem and explain each stage involved.	16	Apply	BTL-3
10	Explain the algorithm for Matrix–Vector Multiplication using MapReduce and illustrate the execution flow.	16	Apply	BTL-3
11	Analyze the advantages and limitations of using MapReduce for Matrix–Vector Multiplication.	16	Analyze	BTL-4
12	Compare conventional data processing techniques with Big Data processing approaches and evaluate their suitability for large-scale applications.	16	Evaluate	BTL-5
13	Evaluate the effectiveness of Distributed File Systems in supporting scalable Big Data processing.	16	Evaluate	BTL-5
14	Analyze the relationship between Big Data characteristics and the design of MapReduce-based processing systems.	16	Analyze	BTL-4
15	Apply Big Data concepts to propose an appropriate solution for processing massive datasets in a business environment.	16	Apply	BTL-3
16	Evaluate the contribution of MapReduce algorithms in improving scalability, reliability, and performance of Big Data applications.	16	Evaluate	BTL-5
17	Analyze the challenges encountered in Big Data processing and recommend suitable approaches based on the MapReduce framework.	16	Evaluate	BTL-5

UNIT –II INTRODUCTION TO HADOOP

Big Data – Apache Hadoop and Hadoop Ecosystem – Moving Data in and out of Hadoop – Understanding Inputs and Outputs of MapReduce – Data Serialization.

UNIT –II [PART-A]

Q.No	Questions	Competence	Level
1	Define Apache Hadoop.	Remember	BTL-1
2	What is the Hadoop Ecosystem?	Remember	BTL-1
3	State the primary objectives of Apache Hadoop.	Understand	BTL-2
4	List the major components of the Hadoop Ecosystem.	Remember	BTL-1
5	What is meant by moving data into Hadoop?	Remember	BTL-1
6	What is meant by moving data out of Hadoop?	Remember	BTL-1
7	Why is data movement important in Hadoop?	Understand	BTL-2
8	Define MapReduce input.	Remember	BTL-1
9	What is a MapReduce InputSplit?	Remember	BTL-1
10	Define RecordReader in MapReduce.	Remember	BTL-1
11	What is the output of the Map phase?	Remember	BTL-1
12	What is the output of the Reduce phase?	Remember	BTL-1
13	Define Data Serialization.	Remember	BTL-1
14	State the purpose of serialization in Hadoop.	Understand	BTL-2
15	List the characteristics of an efficient serialization framework.	Understand	BTL-2
16	Distinguish between input and output in the MapReduce framework.	Understand	BTL-2

17	Differentiate Hadoop from traditional data processing systems.	Understand	BTL-2
18	Mention any four advantages of Apache Hadoop.	Remember	BTL-1
19	Explain the role of Hadoop in Big Data processing.	Understand	BTL-2
20	Summarize the importance of data serialization in distributed computing.	Understand	BTL-2

UNIT –II [PART-B]

Q.No	Questions	Marks	Competence	Level
1	Explain the architecture and working principles of Apache Hadoop and discuss its significance in Big Data processing.	16	Apply	BTL-3
2	Analyze the major components of the Hadoop Ecosystem and explain how they support distributed data processing.	16	Analyze	BTL-4
3	Apply Hadoop concepts to demonstrate how large-scale data can be stored and processed efficiently in a distributed environment.	16	Apply	BTL-3
4	Explain the process of moving data into and out of Hadoop with suitable illustrations.	16	Apply	BTL-3
5	Analyze the challenges associated with data movement in Hadoop and recommend suitable approaches to overcome them.	16	Analyze	BTL-4
6	Explain the complete flow of input processing in the MapReduce framework, highlighting the functions of InputFormat, InputSplit, and RecordReader.	16	Apply	BTL-3
7	Analyze the output generation process in MapReduce and explain the significance of intermediate and final outputs.	16	Analyze	BTL-4
8	Apply the concepts of MapReduce inputs and outputs to design an efficient data processing workflow for a distributed application.	16	Apply	BTL-3
9	Explain the concept of data serialization in Hadoop and discuss its role in improving distributed communication.	16	Apply	BTL-3
10	Analyze the characteristics of an effective serialization mechanism and evaluate its impact on Hadoop performance.	16	Evaluate	BTL-5
11	Compare Hadoop with conventional data processing frameworks and evaluate their suitability for large-scale applications.	16	Evaluate	BTL-5
12	Analyze how Hadoop Ecosystem components collectively support scalable and fault-tolerant Big Data processing.	16	Analyze	BTL-4
13	Evaluate the effectiveness of Hadoop as a platform for processing structured, semi-structured, and unstructured data.	16	Evaluate	BTL-5
14	Apply Hadoop data movement concepts to develop a strategy for transferring enterprise data efficiently between external systems and Hadoop.	16	Apply	BTL-3
15	Analyze the importance of serialization in distributed applications and justify its necessity in Hadoop environments.	16	Analyze	BTL-4
16	Evaluate the role of Apache Hadoop in addressing the limitations of traditional database systems for Big Data applications.	16	Evaluate	BTL-5
17	Analyze the complete data processing workflow in Hadoop, beginning with data ingestion and concluding with the generation of MapReduce outputs.	16	Analyze	BTL-4

UNIT –III HADOOP ARCHITECTURE

Hadoop Architecture – Hadoop Distributed File System (HDFS) – Common Hadoop Shell Commands – Anatomy of File Write and Read – NameNode – Secondary NameNode – DataNode – Hadoop MapReduce Paradigm – Map and Reduce Tasks – Job and Task Trackers – Cluster Setup – SSH and Hadoop Configuration – HDFS Administration – Monitoring and Maintenance.

UNIT –III [PART-A]				
Q.No	Questions	Competence		Level
1	Define Hadoop Architecture.	Remember		BTL-1
2	What is Hadoop Distributed File System (HDFS)?	Remember		BTL-1
3	State the primary objective of HDFS.	Understand		BTL-2
4	Define NameNode.	Remember		BTL-1
5	What is the function of a DataNode?	Remember		BTL-1
6	Define Secondary NameNode.	Remember		BTL-1
7	Differentiate between NameNode and DataNode.	Understand		BTL-2
8	What is meant by Hadoop Shell Commands?	Remember		BTL-1
9	List any four commonly used HDFS shell commands.	Remember		BTL-1
10	What is meant by the anatomy of file write in HDFS?	Understand		BTL-2
11	Explain the anatomy of file read in HDFS.	Understand		BTL-2
12	Define the Hadoop MapReduce paradigm.	Remember		BTL-1
13	What is a Map Task in Hadoop?	Remember		BTL-1
14	What is a Reduce Task?	Remember		BTL-1
15	Define Job Tracker.	Remember		BTL-1
16	Define Task Tracker.	Remember		BTL-1
17	What is the purpose of SSH in Hadoop Cluster Setup?	Understand		BTL-2
18	State the need for Hadoop Configuration.	Understand		BTL-2
19	What is HDFS Administration?	Remember		BTL-1
20	Why are monitoring and maintenance important in Hadoop clusters?	Understand		BTL-2
UNIT –III [PART-B]				
Q.No	Questions	Marks	Competence	Level
1	Explain the architecture of Hadoop with a neat diagram and discuss the functions of its major components.	16	Apply	BTL-3
2	Analyze the architecture of HDFS and explain how it ensures reliable and scalable storage of Big Data.	16	Analyze	BTL-4
3	Explain the roles and responsibilities of NameNode, Secondary NameNode, and DataNode in the Hadoop Distributed File System.	16	Apply	BTL-3
4	Analyze the interaction between NameNode and DataNode during file storage and retrieval operations in HDFS.	16	Analyze	BTL-4
5	Explain the complete sequence involved in the anatomy of file write operation in HDFS with suitable illustrations.	16	Apply	BTL-3
6	Analyze the anatomy of file read operation in HDFS and discuss the mechanisms used to improve data access efficiency.	16	Analyze	BTL-4
7	Apply Hadoop shell commands to perform common HDFS file and directory management operations in a distributed environment.	16	Apply	BTL-3
8	Explain the Hadoop MapReduce paradigm and illustrate the execution flow from job submission to completion.	16	Apply	BTL-3
9	Analyze the functions of Map Tasks and Reduce Tasks in processing large datasets using the MapReduce framework.	16	Analyze	BTL-4
10	Explain the responsibilities of Job Tracker and Task Tracker and analyze their coordination in executing MapReduce jobs.	16	Analyze	BTL-4
11	Apply the concepts of Hadoop Cluster Setup to configure a basic Hadoop environment using SSH and Hadoop configuration files.	16	Apply	BTL-3

12	Analyze the significance of SSH in Hadoop cluster communication and evaluate its contribution to secure cluster management.	16	Evaluate	BTL-5
13	Explain the procedures involved in HDFS administration and discuss the responsibilities of an HDFS administrator.	16	Apply	BTL-3
14	Analyze various monitoring and maintenance activities performed in Hadoop clusters to ensure reliable system performance.	16	Analyze	BTL-4
15	Evaluate the effectiveness of HDFS architecture in providing fault tolerance, scalability, and high availability for Big Data applications.	16	Evaluate	BTL-5
16	Compare the roles of NameNode, Secondary NameNode, and DataNode and evaluate their collective contribution to efficient HDFS operation.	16	Evaluate	BTL-5
17	Analyze the complete workflow of storing, processing, administering, and maintaining data in a Hadoop cluster, highlighting the role of each architectural component.	16	Evaluate	BTL-5

UNIT –IV HADOOP ECOSYSTEM AND YARN

Hadoop Ecosystem Components – Schedulers (Fair Scheduler and Capacity Scheduler) – Hadoop 2.0 New Features – NameNode High Availability – HDFS Federation – MRv2 – YARN – Running MRv1 in YARN.

UNIT –IV [PART-A]

Q.No	Questions	Competence	Level
1	Define the Hadoop Ecosystem.	Remember	BTL-1
2	List the major components of the Hadoop Ecosystem.	Remember	BTL-1
3	What is a Hadoop Scheduler?	Remember	BTL-1
4	Define the Fair Scheduler.	Remember	BTL-1
5	Define the Capacity Scheduler.	Remember	BTL-1
6	Differentiate between the Fair Scheduler and the Capacity Scheduler.	Understand	BTL-2
7	What is meant by Hadoop 2.0?	Remember	BTL-1
8	State any four new features introduced in Hadoop 2.0.	Remember	BTL-1
9	Define NameNode High Availability (HA).	Remember	BTL-1
10	Why is High Availability required in Hadoop?	Understand	BTL-2
11	What is HDFS Federation?	Remember	BTL-1
12	State the advantages of HDFS Federation.	Understand	BTL-2
13	Define MRv2 (MapReduce Version 2).	Remember	BTL-1
14	What is YARN?	Remember	BTL-1
15	Expand the acronym YARN.	Remember	BTL-1
16	List the major components of YARN.	Remember	BTL-1
17	State the functions of the ResourceManager in YARN.	Understand	BTL-2
18	What is the role of the NodeManager in YARN?	Understand	BTL-2
19	Explain the purpose of running MRv1 in YARN.	Understand	BTL-2
20	Summarize the significance of YARN in Hadoop 2.0.	Understand	BTL-2

UNIT –IV [PART-B]

Q.No	Questions	Marks	Competence	Level
1	Explain the Hadoop Ecosystem and discuss the functions of its major components in supporting Big Data processing.	16	Apply	BTL-3
2	Analyze the architecture of the Hadoop Ecosystem and explain how its components interact to process distributed data efficiently.	16	Analyze	BTL-4

3	Compare the Fair Scheduler and the Capacity Scheduler with suitable examples and evaluate their suitability for different application environments.	16	Evaluate	BTL-5
4	Explain the working principles of the Fair Scheduler and analyze its advantages and limitations in cluster resource management.	16	Analyze	BTL-4
5	Explain the Capacity Scheduler and evaluate its effectiveness in managing resources among multiple users in a Hadoop cluster.	16	Evaluate	BTL-5
6	Discuss the major enhancements introduced in Hadoop 2.0 and analyze how these improvements overcome the limitations of earlier Hadoop versions.	16	Analyze	BTL-4
7	Explain the concept of NameNode High Availability and discuss its implementation for improving Hadoop cluster reliability.	16	Apply	BTL-3
8	Analyze the architecture of NameNode High Availability and evaluate its contribution to fault tolerance in Hadoop.	16	Evaluate	BTL-5
9	Explain the concept of HDFS Federation and analyze its role in improving scalability and namespace management.	16	Analyze	BTL-4
10	Compare HDFS Federation with the traditional HDFS architecture and evaluate their performance in large-scale distributed environments.	16	Evaluate	BTL-5
11	Explain the architecture of MRv2 and discuss the improvements it provides over the earlier MapReduce model.	16	Apply	BTL-3
12	Explain the architecture and working of YARN, highlighting the functions of ResourceManager, NodeManager, and ApplicationMaster.	16	Apply	BTL-3
13	Analyze how YARN performs resource management and job scheduling efficiently in a Hadoop cluster.	16	Analyze	BTL-4
14	Evaluate the role of YARN in improving cluster utilization, scalability, and application management in Hadoop 2.0.	16	Evaluate	BTL-5
15	Explain the procedure for running MRv1 applications in a YARN environment and discuss its significance in maintaining backward compatibility.	16	Apply	BTL-3
16	Analyze the combined contributions of Hadoop Ecosystem components, HDFS Federation, and YARN toward building scalable and fault-tolerant Big Data platforms.	16	Evaluate	BTL-5
17	Evaluate the overall impact of Hadoop 2.0 enhancements, including NameNode High Availability, HDFS Federation, MRv2, and YARN, on the performance and reliability of Big Data processing systems.	16	Evaluate	BTL-5

UNIT – V SECURITY AND STORING SYSTEM

Hive Architecture and Installation – Comparison with Traditional Database – HiveQL – Querying Data – Sorting and Aggregating – MapReduce Scripts – Joins and Subqueries – HBase Concepts – Advanced Usage – Schema Design – Advanced Indexing – Pig – ZooKeeper – Monitoring a Cluster – HBase with ZooKeeper – Building Applications with ZooKeeper.

UNIT – V [PART-A]

Q.No	Questions	Competence	Level
1	Define Apache Hive.	Remember	BTL-1
2	What is the architecture of Hive?	Remember	BTL-1
3	State the purpose of Hive installation.	Understand	BTL-2
4	Differentiate between Hive and a traditional relational database.	Understand	BTL-2
5	Define HiveQL.	Remember	BTL-1
6	What is meant by querying data in Hive?	Remember	BTL-1

7	Define sorting in HiveQL.	Remember	BTL-1
8	What is aggregation in HiveQL?	Remember	BTL-1
9	Define a MapReduce script in Hive.	Remember	BTL-1
10	What is a join in HiveQL?	Remember	BTL-1
11	Define a subquery in HiveQL.	Remember	BTL-1
12	What is Apache HBase?	Remember	BTL-1
13	State the purpose of schema design in HBase.	Understand	BTL-2
14	What is advanced indexing in HBase?	Understand	BTL-2
15	Define Apache Pig.	Remember	BTL-1
16	What is Apache ZooKeeper?	Remember	BTL-1
17	Explain the role of ZooKeeper in monitoring a Hadoop cluster.	Understand	BTL-2
18	Why does HBase use ZooKeeper?	Understand	BTL-2
19	State the importance of building applications with ZooKeeper.	Understand	BTL-2
20	Summarize the role of Hive and HBase in the Hadoop ecosystem.	Understand	BTL-2

UNIT – V [PART-B]

Q.No	Questions	Marks	Competence	Level
1	Explain the architecture of Apache Hive and discuss the steps involved in its installation and configuration.	16	Apply	BTL-3
2	Compare Hive with traditional relational database systems and evaluate their suitability for Big Data applications.	16	Evaluate	BTL-5
3	Explain the features and architecture of HiveQL and demonstrate how it is used for querying large datasets.	16	Apply	BTL-3
4	Apply HiveQL statements to perform data retrieval, filtering, sorting, and aggregation on Big Data.	16	Apply	BTL-3
5	Analyze the role of sorting and aggregation operations in improving the efficiency of Hive-based data analysis.	16	Analyze	BTL-4
6	Explain the execution of MapReduce scripts in Hive and discuss their significance in distributed data processing.	16	Apply	BTL-3
7	Analyze different types of joins and subqueries supported by HiveQL and evaluate their effectiveness in complex query processing.	16	Analyze	BTL-4
8	Explain the architecture and fundamental concepts of HBase and discuss its applications in Big Data environments.	16	Apply	BTL-3
9	Analyze advanced usage features of HBase and explain how they contribute to efficient storage and retrieval of large datasets.	16	Analyze	BTL-4
10	Apply HBase schema design principles to develop an efficient schema for storing large-scale distributed data.	16	Apply	BTL-3
11	Evaluate the importance of advanced indexing techniques in improving HBase query performance.	16	Evaluate	BTL-5
12	Explain the architecture and features of Apache Pig and analyze its role in Big Data processing.	16	Analyze	BTL-4
13	Explain the architecture and working principles of Apache ZooKeeper and discuss its role in distributed coordination.	16	Apply	BTL-3
14	Analyze how ZooKeeper assists in monitoring Hadoop clusters and maintaining coordination among distributed nodes.	16	Analyze	BTL-4
15	Explain the interaction between HBase and ZooKeeper and evaluate the benefits achieved through their integration.	16	Evaluate	BTL-5

16	Apply ZooKeeper concepts to design a reliable distributed application that requires coordination and synchronization among multiple nodes.	16	Apply	BTL-3
17	Evaluate the collective contributions of Hive, HiveQL, HBase, Pig, and ZooKeeper in developing scalable, reliable, and efficient Big Data applications.	16	Evaluate	BTL-5