Big Data Analytics

Module-1 Introduction to Big Data and Hadoop

1.1 Introduction to Big Data -Big Data characteristics and Types of Big Data
1.2 Traditional vs. Big Data business approach
1.3 Case Study of Big Data Solutions
1.4 Concept of Hadoop, Core Hadoop Components; Hadoop Ecosystem
1.5 Hadoop Limitations

Importance Question:

1. What is big data? List the different characteristics of Big data.
2. Explain 5 V's of Big Data.
3. What are Big Data types. (Structured, Unstructured and Semi structured)
4. How Unstructured data are different from Structured data.
5. What are the different sources of big data?
6. What is Big Data Analytics? Explain the different types of analytics.
7. How Big Data approach is different than a Traditional business approach.
8. Compare the traditional data v/s big data.
9. List the different application of big data in different sector.
10. How big data analytics help in the health care sector.
11. What is the use case of big data in Banking sector.
12. How Big data analytics use in E-commerce.

13. What are the different challenges of big data, and their solutions?
14. What is Hadoop? Explain the core component of Hadoop.
15. What are the limitations of Hadoop?
16. Explain Hadoop Ecosystems with Diagrams.
17. How Pig and Hive are different from MapReduce in processing.
18. Which tools is used for import & export relation data between Relational DBMS & HDFS.
19. What is the role of Apache Spark in Big data analytics.

Module- 2 Hadoop HDFS and MapReduce

2.1Distributed File Systems: Physical Organization of Compute Nodes, Large-Scale File-System Organization.
2.2MapReduce: The Map Tasks, Grouping by Key, The Reduce Tasks, Combiners, Details of MapReduce Execution, Coping with Node Failures.
2.3Algorithms Using MapReduce: Matrix-Vector Multiplication by MapReduce, Relational-Algebra Operations, Computing Selections by MapReduce, Computing Projections by MapReduce, Union, Intersection, and Difference by MapReduce

Importance Question:

1. What is the distributed file system. List the key features of a distributed file system.
2. How distributed file system does differ from a traditional file system?

3. What is HDFS. Explain the HDFS main core components and their roles.
4. What are advantages of using replication factor > 1 in HDFS.
5. Explain the concept of data replication in Hadoop HDFS and its significance in a distributed environment.
6. How does Hadoop HDFS ensure fault tolerance and data reliability in the face of node failures?
7. What is the role of the Name Node and Data Nodes in Hadoop HDFS, and how do they interact to manage data storage?

8. Describe the concept of MapReduce and its role in big data processing.
9. What is the purpose of a Combiner in MapReduce, and how does it help optimize the data processing flow?

10. How can you perform relational-algebra operations like Join and Group By using MapReduce? Provide examples.
11. How does MapReduce handle Union, Intersection, and Difference operations on datasets?
12. Perform the matrix multiplication using MapReduce where $M = \begin{bmatrix} 5 & 4 \\ 2 & 3 \end{bmatrix}$ and $N = \begin{bmatrix} 3 & 2 & 5 \\ 2 & 5 & 3 \end{bmatrix}$
13. Find the average salary of all employees of every department using MapReduce.

Emp Id	Dept	Salary
01	CMPN	70 k
02	INFT	60 k
03	INFT	80 k
04	BIOM	80 k
05	CMPN	95 k
06	INFT	90 k
07	CMPN	90 k
08	CMPN	80 k
09	INFT	60 k
10	BIOM	50 k

Paper Pattern for MSE (30 Marks)

Q1. Solve any five questions (each question 2 marks) - 10 Marks.
1.a, 1.b, …….1.h (total 8 questions)

Q2. Solve anyone/two Questions (each question 10/5 marks) - 10 Marks.
2.a, 2.b, 2.c

Q3. Solve anyone/two Questions (each question 10/5 marks) - 10 Marks.
3.a, 3.b, 3.c