Explain the types of digital data with example

11.a)i)~~Explain in detail about various classification of big data analyties.~~ CO1,K2 8

(ii) A manufacturing company is looking to optimize its production processes using big data analytics. Identify the data sources and analytics techniques used to improve and increase operational efficiency CO1,K3 8

b)i)Explain about some of the terminologies used in the big data environments. CO1,K2

(ii) A retail chain is experiencing a sudden drop in sales for a specific product category. How would you approach analyzing this situation using big data analytics? What steps would you take to identify the root causes? CO1,K3

12a) Explain the semantic representation of MapReduce algorithm and its workflow with suitable example.CO2,K2

b) Illustrate the concept about Hadoop distributed file system with neat diagram. CO2,K2

13) a)Construct the MongoDB query to create a student collection and insert student details such as rollno, name, year of study, department, subject and CGPA for 8 students. Finally show the student Collection using pretty function. CO3,K3

b)Develop the Cassandra query for the following

i) To insert 4 data into the column family "student_info"
ii) To view the data from the table "student_info"
iii) To view only those records where the rollno column either has a value 1 or 2 or 3 CO3,K3

14a)i) Explain the architecture of **Apache Hive** and how it interacts with **Hadoop**. CO4,K2

ii) Develop a query in HQL to **join two tables** and perform a **group by** operation? CO4,K3

b) i)Demonstrate the role of **Pig** in the **Hadoop ecosystem** and how it simplifies **data processing**. CO4,K2

ii) Develop a Pig script that performs a **JOIN operation** between two datasets? CO4,K3

15)a) i)Outline the **architecture of Apache Spark** and its key components.CO5,K2

ii) Explain how tools like **Spark SQL**, **MLlib**, and **GraphX** are used in the Spark ecosystem.CO5,K2

b)i)Summarize the KAFKA architecture with neat sketch CO5,K2

ii)Illustrate the role of Kafka in combination with **Spark** to handle real-time streaming data in a **financial trading** scenario? CO5,K2

2 Marks

1. Identify the Structured, Unstructured and Semi structured data for the following examples.

   Email

   Images

   Chat Conversations

   Relations / Tables C01, K3

2. Interpret CAP Theorem. CO1,K2
3. Compare RDBMS and Hadoop. C02,K2
4. Identify how the hardware failures in distributed computing and how it is overcome. CO2, K3
5. Construct the MongoDB query to create a collection by the name "Student" and insert two documents. CO3,K3
6. Interpret Hinted Handsoffs in Cassandra CO3, K2
7. Show the Embedded Metastore in Hive CO4,K2
8. Outline the two types of running pig and execution modes of pig. CO4,K2
9. Apply **Spark Streaming** to process real-time sensor data from IoT devices? CO5,K3
10. Interpret Apache KAFKA and list five major API's in KAFKA CO5,K2

Set 2

2Marks

1. Identify the Structured, Unstructured and Semi structured data for the following examples.

   CCTV Footage

   XML Document

   Facebook
   Spreadsheet CO1,K3

2. Infer – Brewer's Theorem CO1, K2

3. Outline the key aspects of Hadoop. CO2, K2
4. Show the high-level architecture of Hadoop with its components. CO2,K2
5. Construct the MongoDB query to create a collection by the name "Employee" and insert two documents. Co3,k3
6. Construct the Cassandra query to delete a row (where Roll no =2) from the table "student_info". Co3,k3.
7. Show the Remote meta store in Hive. Co4,k2
8. Identify the key differences between the Apache Hive and Apache Pig. Co4,k3
9. Interpret Apache spark.co5,k2
10. Outline the various applications of Apache Kafka. Co5,k2

10 Marks

11. a).i) Explain the key differences between structured, semi structured and unstructured data with an example sources. CO1, K2

    a).ii). A retail company wants to optimize its marketing strategy by analyzing customer behavior and preferences. The company has collected a large dataset containing customer demographics, purchase history, browsing behavior, and social media interactions. How would you, as a data scientist, identify patterns and relationships in this dataset to inform business strategy, and what technology and mathematical expertise would you employ to achieve this? Co1,k3

    b). i) Outline seven top analytic tools in Big Data and its usages. CO1, K2

    b) ii) Identify the difference between Parallel and Distributed systems in Big Data Environment. C01,K3

12. A). Explain the MapReduce programming architecture and its workflow. Co2,k2

    B). Construct the MapReduce program to sort the data by student name (Value).

    Input Data:

    1001,Jams,45

    1002,Bharathi,39

    1003,Anbu,44

1004,Nandhu,38

1005,Subhasri,33 CO2,K3

13). A). Construct the MongoDB query to create a employee collection and insert employee details such as Employee ID, name, department and salary for 5 employees. Co3,k3

B). Construct the Cassandra query to count the no. of times a particular book is issued from the library by the student. CO3,K3

14). A). i) Explain the architecture of Apache Hive and its meta store types. CO4,K2

A).ii). Summarize the Hive Query Language. CO4,K2

B).i). Explain the various relational operators in Pig. CO4,K2

ii).Explain the EVAL Function in Pig CO4,K2

15). A).i). Explain the architecture of Apache spark and explain two important innovations of Spark CO5,K2

B) i). Explain the Architecture of Apache KAFKA with its components. CO5,K2