15 JUN 2023

Reg. No.												

Question Paper Code	11893

B.E. / B.Tech. - DEGREE EXAMINATIONS, APRIL / MAY 2023

Fifth Semester

Information Technology

20ITPC502 – BIG DATA ESSENTIALS

(Regulations 2020)

Duration: 3 Hours Max. Marks: 100

PART - A (10 × 2 = 20 Marks)

Answer ALL Questions

Marks, K-Level, CO

1. Define "big data" and under what conditions it is given that name. *2,K1,CO1*
2. Identify the characteristics of Big Data *2,K2,CO1*
3. Show the key advantages in Hadoop. *2,K1,CO2*
4. Discuss the importance of DFS. *2,K1,CO2*
5. How can a key value pair is formed? *2,K1,CO3*
6. Compare MapReduce and YARN *2,K2,CO3*
7. Write about the key design principles of Pig Latin. *2,K1,CO4*
8. List out the data types in Hive. *2,K1,CO4*
9. Identify the components and features of Spark. *2,K2,CO5*
10. What is mean by Matrix Multiplication *2,K1,CO5*

PART - B (5 × 13 = 65 Marks)

Answer ALL Questions

11. a) What are the benefits of Big Data? Discuss challenges under Big Data. How Big Data Analytics can be useful in the development of smart cities. *13,K2,CO1*

OR

b) List various applications of big data. Explain how it can be used to improve business for a superstore. *13,K2,CO1*

12. a) Explain core architecture of Hadoop with suitable block diagram. Discuss role of each component in detail. *13,K2,CO2*

OR

b) Explain about Hadoop distributed file system (HDFS) architecture with neat diagram. *13,K2,CO2*

13. a) How will you perform Job scheduling, shuffle and sorting using YARN. *13,K2,CO3*

OR

b) Consider a collection of literature survey made by a researcher in the form of a text document with respect to cloud and big data analytics. Analyze the above data using Map Reduce, write a program to count the occurrence of pre dominant key words *13,K2,CO3*

14. a) (i) How can you create and manage the data bases in Hive, explain the steps with example. *7,K2,CO4*

(ii) What are the steps to followed Squirrel running on Apache Hive with a neat diagram. *6,K2,CO4*

OR

b) (i) How can you run the Pig script in Local and Distributed mode explain with your own example. *7,K2,CO4*

(ii) Write a syntax for Pig program with suitable example. *6,K2,CO4*

15. a) Explain in detail about CUDA Programming Model and Memory model. *13,K2,CO5*

OR

b) What is Apache Spark? What are the advantages of using Apache Spark over Hadoop? Explain in brief four major libraries of Apache Spark. *13,K2,CO5*

PART - C (1 × 15 = 15 Marks)

16. a) Suppose we have created a Hive partition table which is partitioned by a column named city. We are getting data which are having Empty/Null value for the partition column (city) and have to load these data into the hive table with dynamic partition as it is having multiple city records in the data set. In which partition the records, with an empty value for city column, will be available? *15,K3,CO6*

OR

b) (i) There is a json file with following content :- *7,K3,CO6*

{“dept_id”:101,”e_id”:[10101,10102,10103]}
{“dept_id”:102,”e_id”:[10201,10202]}

and data is loaded into spark dataframe say mydf, having below dtypes

dept_id: bigint, e_id: array<bigint>

What will be the best way to get the e_id individually with dept_id ?

(ii) Suppose you have two dataframe df1 and df2, both have below columns :- *8,K3,CO6*

df1 => id, name, mobno
df2 => id, pincode, address, city

After joining both the dataframe on the basis of key i.e id, while selecting id, name, mobno, pincode, address, city, you are getting an error ambiguous column id. How would you resolve it ?