Enrolment No. 33

NATIONAL FORENSIC SCIENCES UNIVERSITY, DELHI CAMPUS
B. Tech - M. Tech CSE (CS) – Semester V, October-November 2023
Mid Semester Examination

Subject Code: CTBTCSE SV P6 EL1
Subject Name: Big Data

Date: 06/11/2023
Time: 1 Hr 30 Min
Total Marks: 50

All the questions are compulsory.

Section A

1. Answer all questions. **10*2=20**

a) List the characteristics of Big Data.
b) What is the role of Sort and Shuffle in Map-Reduce?
c) What are the three key design principles pig latin?
d) What are the real time industry applications of Hadoop?
e) Explain Metastore in Hive.
f) How can you implement a custom variable? define
g) Write about combiner and partitioner.
h) Write about auto boxing and unboxing?
i) Why HIVE is relevant in Hadoop Eco system?
j) Specify the role of job tracker and task tracker in HDFS.

Section B **5*3=15**

2. a) Define structured, semi structured and un structured data with examples.
   b) Differentiate between Map-Reduce, PIG and HIVE.
   c) Explain about the implementation of Map reduce concept with an example.
   d) Differentiate between Google File System and Hadoop File System.
   e) What is a Data Node? How many instances of Data Node run on a Hadoop Cluster?

Section C **3*5=15**

3. a) Define HDFS. Discuss the HDFS Architecture and HDFS Commands in brief.

   b) Explain the architecture of HIVE with a neat sketch.

   c) Discuss in brief about the basic building blocks of Hadoop.

~~~~~~~~END~~~~~~~~
~~~~~~~~

Seat No.: _____ Enrolment No. _____

NATIONAL FORENSIC SCIENCES UNIVERSITY

B.Tech-M.Tech CSE(Cyber Security) - Semester - V – January - 2024

Subject Code: CTBTCSE SV P6 EL1 **Date: 08/01/2024**

Subject Name: Big Data

Time: 11:00 AM to 02:00 PM **Total Marks: 100**

Instructions:

1. **Write down each question on a separate page.**
2. **Attempt all questions.**
3. **Make suitable assumptions wherever necessary.**
4. **Figures to the right indicate full marks.**

Q.1 (a) Explain the main five characteristics of Hadoop. 05

(b) What are the benefits of Big Data? Discuss challenges under Big Data. 05

(c) In the Hadoop cluster, the name node is a single source of failure in HDFS, how is this failure managed in the HDFS cluster? 07

or

(c) Assume you have three files, and each file contains two columns (a key and a value in Hadoop terms) that represent a city and the corresponding temperature recorded in that city for the various measurement days. The city is the key, and the temperature is the value. For example: (Ahmedabad, 20). Out of all the data we have collected, you want to find the maximum temperature for each city across the data files (note that each file might have the same city represented multiple times). 07

Data: -

file1 – (Ahmedabad,38),(Mumbai,20),(Ahmedabad,33),(Jaipur,35)

file2 – (Mumbai,15),(Ahmedabad,39),(Banglore,30),(Banglore,33)

file3 – (Ahmedabad,40),(Jaipur,40),(Mumbai,36),(Delhi,37)

Using the MapReduce framework find out the maximum temperature for each city. What will be the output after each phase of MapReduce?

Q.2 (a) Explain the data model of Apache Pig. 05

(b) Explain zookeeper architecture. 05

(c) Explain Job Scheduling in YARN. How it is done in the case of (i) The Fair Scheduler (ii) The Capacity Scheduler. 07

or

(c) Explain the concept of a data frame in R with an example. Highlight its key characteristics and how it differs from other data structures. 07

Q.3 (a) Define HDFS. Discuss the HDFS Architecture. 08

(b) Discuss applications of big data in healthcare and transportation. 08

Q.4 (a) Explain different types of znodes in Zookeeper. 05

(b) What is HBase? Discuss the difference between HBASE and HDFS. 05

(c) Consider a file movies.txt with fields movie_id, movie_name, release_year, genre, movie_rating and with the following records. 07

movies.txt :

(1, 'A Space Odyssey ', 1968, 'Scifi', 5)

(2, 'The Godfather', 1972, 'Crime', 5)

(3, 'The Godfather', 1990, 'Crime', 3)

(4, 'City Lights', 1931, 'Comedy', 4)

Write below pig Latin commands/queries.

(i) Open pig shell.

(ii) Load movies records from HDFS with movie_id, movie_name, release_year, genre, movie_rating column names in relation 'movies'.

(iii) Group movies by their genre.

(iv) List all movies with the genre 'Crime'.

(v) List movie names that start with 'T'.

or

(c) Explain how Pig Latin scripts are compiled. (3 Marks) 07

Explain following commands in pig latin scripts. (i)Load (ii)Foreach (iii)Filter (iv)Dump. (4 Marks)

Q.5 (a) What is big data governance? Explain tools used for big data governance in Hadoop. 05

(b) Explain the following functions with examples in R programming. 05

(i) c() (ii) rep() (iii) seq() (iv) matrix() (v) diag()

(c) What is the difference between vector and list in R programming? (3 Marks) 07

Write code for the following. (1 Marks each)

a) Create a vector named ages containing the ages of five individuals: 25, 30, 22, 35, and 28.

b) Calculate and print the mean of the ages.

c) Create a new vector named ages_squared that contains the squared values of each age in the ages vector.

d) Display the content of the ages_squared vector.

or

(c) Explain the commissioning and decommissioning of nodes in Hadoop. 07

Q.6 (a) Explain Hadoop security through Kerberos and delegation tokens. 08

(b) Explain different types of data analytics with examples. 08

END OF PAPER