records are admissible where temperature is not equal to 9999 and air quality index is equal to 0 or 1 or 4 or 5 or 9. (10)

(b) What are the different data types in Pig? Explain with the help of examples. (5)

6. (a) The class XIIth Exam result, conducted by CBSE, is published in a text file containing following fields. It contains more than 35 lakh entries. This file is stored in a HDFS file system.

(i) Name of the student, (ii) Roll No., (iii) Percentage obtained, (iv) State and (v) City.

Write Hive scripts

– To load the data in a tabular structure.

– To create dynamic partitioning so as to do the analysis at "state" and, if required, at "city" level. (10)

(b) What is the basic architecture of Hive? (5)

7. (a) Describe Cassandra object model with the help of an example. (5)

(b) Explain the hierarchal relationship between Zookeeper, HMaster, Region Server, Region, Column Family and Columns in HBase object management model. (5)

(c) Explain how YARN manages tasks and resources in a Hadoop eco-system in order to run an application. (5)

Roll No. .... Total Pages : 4

504203

May 2024

M.Tech. (CE/CSE) II SEMESTER
Big Data Analytics (MCS-18-206)

Time : 3 Hours] [Max. Marks : 75

Instructions :

1. *It is compulsory to answer all the questions (1.5 marks each) of Part-A in short.*
2. *Answer any four questions from Part-B in detail.*
3. *Different sub-parts of a question are to be attempted adjacent to each other.*

PART-A

1. (a) What are the major factors that led to the genesis of Big Data Analytics? Briefly explain. (1.5)

(b) What challenges in RDBMS lead to the rise of NoSQL databases? (1.5)

(c) What benefits do data partitioning and data replication provide in HDFS design? (1.5)

(d) What are the four basic modules of the Hadoop framework? (1.5)

(e) What are the main components of YARN? (1.5)

(f) What does Map (k, v) → <k', v'>* and Reduce (k', <v'>*) → <k', v">* signifies? (1.5)

(g) What is the concept of a "super column" in Cassandra's data model? Give an example. (1.5)

(h) How do Read inconsistencies happen in Master-Slave configuration? (1.5)

(i) With the help of examples, explain the data model of Pig. (1.5)

(j) Schematically explain the relationship between HMaster, Region Server, and Region in HBase. (1.5)

PART–B

2. (a) What are the different types of NoSQL databases? Explain in brief. (5)

(b) Explain, with the help of examples, Write and Read Inconsistencies in Master-Slave and Peer-to-Peer replication techniques. (5)

(c) Design a Key-Value database aggregate, for a customer and Insurance policy (belonging to the customer). The insurance policy contains various policy items for a customer. The aggregate should cater to the query where policies belonging to a customer could be queried and all policies for the insurance company could be queried. Justify your design. (5)

3. (a) Explain the map-reduce workflow schematically highlighting. (5)

(i) Map stage.

(ii) Group by Key stage.

(iii) Reduce stage.

(b) Design a map-reduce workflow with input and output key and value tuple (K,V) for each stage for multiplying two matrices of NxN dimensions. Let us assume N is very large and files containing these matrices are in HDFS. Let us also assume that the mapper knows the indexing in each file. (5)

(c) For Twitter data containing Date, Message, and Location [other metadata...], design a map-reduce workflow with input and output key and value tuple (K,V) for each stage. The task is to find out the word count by the day. (5)

4. (a) For the statement written below, specify what kind of partitioning will take place and illustrate the same with an example. Write the create statement for student table.

```
Hive > Insert into studt_part
>Partitioned (program, semester, course)
>Select name, rollno, program,
>semester
>course
>From students;
```

(b) Design column families for a Banking system. You may assume it contains objects like customers, accounts, managers etc. (5)

(c) How do we resolve write inconsistencies in peer to peer replication model? Illustrate with the help of an example. (5)

5. (a) Write a Pig script to find out the maximum temperature by the year from a file containing following fields, a) Year, b) Temperature and c) Air Quality. Only those