POSSESION OF MOBILES IN EXAM IS UFM PRACTICE.

Name _____ **Enrollment No.** 200103065

Jaypee Institute of Information Technology, Noida

End Term Examination, Odd 2023
B.Tech VII Semester

Course Title : Big Data with Hadoop and Spark **Maximum Time : 2 Hours**
Course Code : 21B12CS411 **Maximum Marks : 35**

CO1	Understand big data challenges and need of Big data storage and computation tools
CO2	Apply Hadoop MapReduce and Spark to solve big data problems
CO3	Analyze big data using Pig, Hive, HBase and Spark tools for solving real word problems
CO4	Assess Hadoop & Spark for big data analytics
CO5	Implement big data application using Hadoop & Spark

Q1. [Marks 3*3, CO4(Assess Level)]

You have been hired as an analyst for planning flights to and from states. Use Spark to analyse flight data from a csv file, *India_Transport_statistics.csv*. Each row in the file represents a row in the format { **Dest_state_name, Origin_state_name, count_of_flights}**. Give commands to

a) Load data from the csv file in dataframe. Find the states with maximum number of flights originating from them.
b) Modify the created dataframe such that data can be accessed using HiveQL. Find the destination states with >10 flights.
c) Load data from the csv file in a RDD. Find the number of flights originating from Gujarat. Convert the RDD into a dataset using case class.

Q2. [Marks 6, CO2(Apply Level)]

Toogle has parsed large amount of data from the web in the format, as shown below. Create a UDF for a HTML parser to parse the data stored in *crawl.json* file, so that it can be applied on multiple datasets. Show all the steps to execute the UDF and display the output.

URL	CONTENT
www.sports.com	<html><body> This website provides sports materials </body></html>
www.textile.com	<html><body> This website provides dress materials </body></html>
www.books .com	<html><body><font-size=30> This website provides books materials </body></html>

Q3. [Marks 2+4, CO4 (Assess Level)]

a) Spark performs Lazy Evaluation. Discuss why and how.

b) 'In Spark, Pipelines are estimators whereas Pipeline Models (fitted pipelines) are transformers.' Justify this statement with the help of a complete example to predict the price of computers (Assume features of a dataset as required). Specify the statements as transformers and estimators.

Q4. [Marks 6, CO3 (Analyze Level)]

Data for customer-wise monthly sales of a store for the year 2023 is available in format **{month, customer_id, no_of_products_bought, total_orders_price}**. Give appropriate commands to load data into Hive tables using either partitioning or bucketing. Justify the choice made. Also, write a query to print the total sales of the store for each month.

Q5. [Marks 4, CO3 (Analyze Level)]

Use Pig to load 2 text files, *Record1.txt* and *Record2.txt*. Here, Record1.txt contains data in format **(player_id: Int, player_name: String, team_name: String)** and Record2.txt contains **(team_name: String, num_of_tournaments_won: Int)**. Give Pig commands to

a) Join the Pig relations

b) Find teams that have won > 5 tournaments

c) Group the relations such that a group contains records from both relations belonging to the same team.

Q6. [Marks 4, CO1 (Understand Level)]

Risk modelling is a big data driven operation for financial sector entities like banks. Previously, each branch of a bank maintained a legacy data warehouse isolated from others. Data such as checking & saving transactions, home mortgage details, credit card transactions and other financial details of every customer were restricted to a local database. Due to this, banks failed to determine a comprehensive risk portfolio of their customers. How can Hadoop support the risk modelling task of banks? Be specific in your answer.

Enrollment No.

Jaypee Institute of Information Technology, Noida

Test 1 Examination ODD SEM 2023

B. Tech. 7th Semester

Course Name: Big Data with Hadoop and Spark **Maximum Time: 1 Hr**

Course Code: 21B1CS411 **Maximum Marks: 20**

CO1	Understand Big Data Challenges and need of Big data storage and computation tools.
CO2	Apply Hadoop, Hbase, Mapreduce, spark to solve big data problems.
CO3	Analyze big data using PiG, Hive, Spark tools for solving real world problems.
CO4	Evaluate hadoop and Spark tools for big data analytics

Note: Attempt all the questions in order.

Q1.(a) Due to increase in the data volume, you are asked to re-engineer an online shopping website which is currently using MYSQL database server as its backend. The website fetches the data from the database and also insert/delete records in the database. Is Hadoop appropriate?

(b) When traditional file systems can handle file size in exabytes, what is the need for HDFS?

[CO1 (Understanding, 2+2 Marks]

Q2. METFLIX, a social media company, is using Hadoop to store and analyze the movie data with details as given in Table 1. You are assigned the task to write the code for map reduce functions to find the number of ratings and average ratings for each movie in the dataset. **[CO3 (Applying, 5 Marks]**

Table 1: Movie data

User_ID	Movie_ID	Ratings	Timestamp
101	242	3	888908021
121	302	4	882308876
145	377	1	878537890
96	51	2	898760560

Q3.(a) Being a database developer, suppose you create a table that contains details of all the transactions done by the customers of year 2016:

CREATE TABLE transaction_details(cust_id INT, amount FLOAT, month STRING, country STRING) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',';

After inserting 50,000 tuples in this table, your client wants to know the total revenue generated for each month. But, Hive is taking too much time in processing this query. Client cannot wait for so long. How will you solve this problem and list the steps that you will be taking in order to do so?

(b) How can you add a new partition in an existing partitioned table? **[CO3 (Analyzing, 4+2 Marks]**

Q4.(a) What will be the output of following queries and why?

(i) hive> SELECT * from numbers TABLESAMPLE(BUCKET 3 OUT OF 10 ON rand()) s;

(ii) hive> SELECT * from numbers TABLESAMPLE(BUCKET 3 OUT OF 10 ON num) s;

Assume, the numbers table has one num column with values 1-10.

(b) Justify answers with examples

_____ is a more straightforward way to generate hierarchical aggregations with subtotals and grand totals, while _____ provide greater flexibility and control to specify different combinations of groupings columns and levels of aggregation in your query.

[CO3 (Analyzing, 2.5+2.5 Marks]