Exam Date & Time: 08-May-2024 (02:30 PM - 05:30 PM)

MANIPAL ACADEMY OF HIGHER EDUCATION

VI SEMESTER B.TECH END SEMESTER EXAMINATIONS, MAY 2024

BIG DATA INTEGRATION AND PROCESSING [CRA 4056]

Marks: 50 **Duration: 180 mins.**

A

Answer all the questions.

Instructions to Candidates: Answer ALL questions Missing data may be suitably assumed

1) Differentiate between DBMS and BDMS by mentioning at least 3 differences. Justify the reason to prefer BDMS to handle streaming Data. (4)

A)

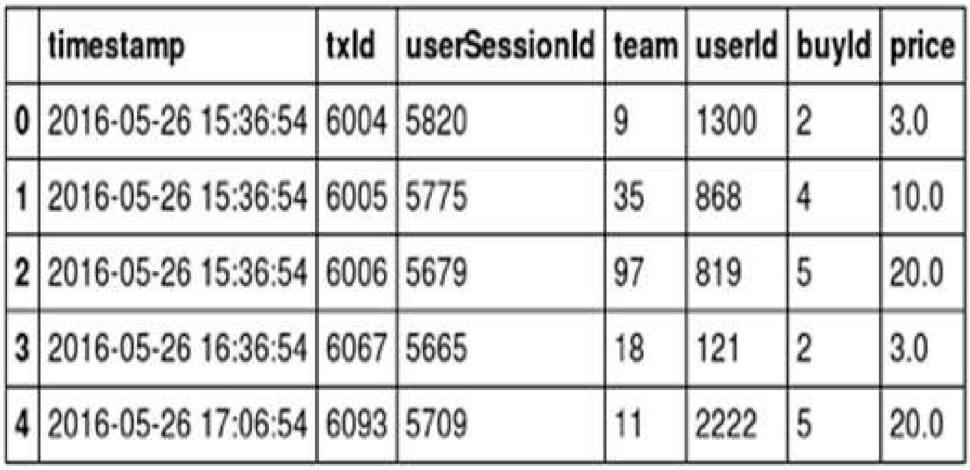
B) Consider the two CSV files, ‘buy-clicks.csv’ and ‘ad-clicks.csv’ shown in Figure 1B1 and 1B2. Apply Python Pandas concept and write the suitable code for implementing the following operations: (4)

a)Read both CSV files into 2 Pandas DataFrames.

b)View the contents of 2 DataFrames.

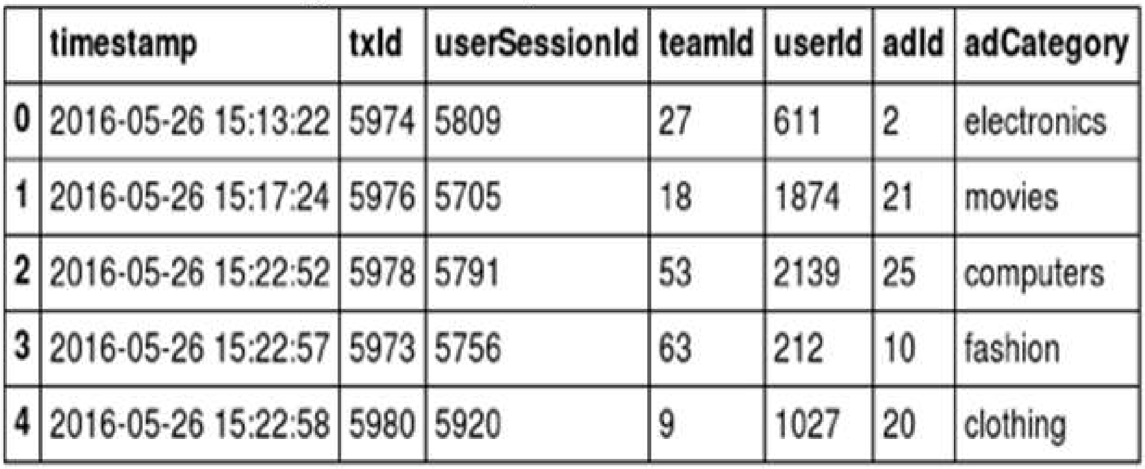
c)Calculate the average and sum of ‘prize’ column in the first DataFrame.

d)Combine two DataFrames by joining on a single column –‘userid’ using merge operation.



	timestamp	txId	userSessionId	team	userId	buyId	price
0	2016-05-26 15:36:54	6004	5820	9	1300	2	3.0
1	2016-05-26 15:36:54	6005	5775	35	868	4	10.0
2	2016-05-26 15:36:54	6006	5679	97	819	5	20.0
3	2016-05-26 16:36:54	6067	5665	18	121	2	3.0
4	2016-05-26 17:06:54	6093	5709	11	2222	5	20.0

Figure 1B1. Buy-clicks.csv



	timestamp	txId	userSessionId	teamId	userId	adId	adCategory
0	2016-05-26 15:13:22	5974	5809	27	611	2	electronics
1	2016-05-26 15:17:24	5976	5705	18	1874	21	movies
2	2016-05-26 15:22:52	5978	5791	53	2139	25	computers
3	2016-05-26 15:22:57	5973	5756	63	212	10	fashion
4	2016-05-26 15:22:58	5980	5920	9	1027	20	clothing

Figure 1B2. ad-clicks.csv

C) Describe Data Exchange problem with an example. (2)

2) Explain the concept of semi-structured data in the context of JSON. Illustrate MongoDB querying of JSON data using find. Provide examples to illustrate. (4)

A)

B) With a neat diagram, Explain the key components of the Aerospike data model and evaluate its use cases. (4)

C) Assuming the file data.csv is in the current directory. Write appropriate commands:

a.To load the file data.csv into a Pandas DataFrame df.

b.To view the first 10 rows in the DataFrame df

c.To know the number of rows and columns in the DataFrame df.

d.To calculate average of the column cost in the DataFrame df.

(2)

3) Depict the schematic representation of querying integrated data. (3)

A)

B) Outline the characteristics of the “big data” problem. Support your answer with valid examples. (4)

C) Illustrate the designs and issues related to mediated schema. (3)

4) List the aggregation functions. Explain the significance of aggregations in bigdata with suitable examples. (4)

A)

B) Explain in detail about Spark architecture with suitable diagram. (3)

C) Demonstrate the analytical operation and its significance in the big data pipelines with suitable examples. (3)

5) Identify the command used to export the result of MongoDB queries in the terminal shell. Explain the command in detail along with its arguments. (4)

A)

B) Explain in detail about spark steaming, its sources and creation and processing of DSteams. (3)

C) Describe MLib Algorithms and techniques with an example. (3)

-----End-----