National University of Computer and Emerging Sciences
Lahore Campus

Fundamental Of Big Data Analytics (DS2004)

Date: April 04, 2024

Course Instructor(s)

Ms. Mamoona Akbar

Ms. Asbah Khalid

Sessional-II Exam

Total Time (Hrs):	1
Total Marks:	30
Total Questions:	3

22L-7556 Roll No

BDS-4A Section

Mannan. Student Signature

IMPORTANT INSTRUCTIONS: Do not use pencil or red ink to answer the questions. In case of confusion or ambiguity make a reasonable assumption. **Only one solution will be marked against one question, carefully attempt questions on answer sheet.**

CLO # 3: Design and analysis the concepts of RDD in pySpark

Q1: [Marks 15] Consider the following database of the "BLOGs" websites. The website keeps track of the different users and blogs written by them on different topics. Each users id identified by a unique username . The websites also keep track of various comments given by users on the Blogs.

The field Bwriter in Blog table is a foreign key from user table and it gives the unique username of the Blog-writer and similarity of the field Cwritter in comment table is a foreign key and gives the username of the user who has given a comment on the Blog.

0026

USER

Uname	Age	Gender
Sara	25	F
Zara	42	F
Ali	15	M
Ahmad	19	M
Aliya	27	F
Tania	29	F
Hamza	34	M

TOPIC

TId	Name	Subject
1	Deep Learning	Computer Science
2	Big Data	Computer Science
3	Databases	Computer Science
4	Algorithms	Computer Science
5	Human Interactions	Philosophy

BLOG

BId	Bname	Bwriter	TopicId
10	BigData Frameworks	Ahmad	2
20	Generation Gap	Sara	5
100	Map Reduce	Hamza	2
30	The world of CNN	Ali	1
50	Cassandra	Ali	3
70	Neural Nets	Tania	1
60	MongoDB	Tania	3
120	Emerging trends	Sara	2
80	Hbase	Ali	3

COMMENT

CId	BlogId	Cwriter
1	20	Hamza
2	100	Hamza
3	20	Zara
20	80	Hamza
7	30	Zara
9	50	Zara
5	80	Ali
12	50	Ahmad
15	50	Tania

Write sql queries and rdd

a) Find the name and age of the users who have never written any blog and have naver given any comment.
b) Find the name of the users who have given comments on all the blogs written in the Computer science area.
c) Find the Cwriter whose comments on the topic name "Database"

CLO # 1: Basic Concept of RDD

Q2: [10 marks] Suppose you have a dataset containing information about students and their grades. Each record in the dataset consists of a tuple (student_id, [(subject, grade)]), where student_id is the unique identifier of the student and [(subject, grade)] represents a list of tuples containing subject and the corresponding grade obtained by the student in that subject.
Tasks:

- Calculate the average grade for each student across all subjects.
- Find the highest grade obtained by each student.

Note: Don't used DataFrame

1, ("Math", 80), ("Science", 75), ("English", 90)
2, ("Math", 85), ("Science", 92), ("English", 88)
3, ("Math", 78), ("Science", 80), ("English", 85)
4, ("Math", 90), ("Science", 87), ("English", 92)

CLO # 1: Basic concept of Hive

Question 3: Marks 5

a) What are the different types of tables supported by hive
b) Explain the role of metastore in Hive. Why is it important