Question Paper

Exam Date & Time: 30-May-2023 (02:30 PM - 05:30 PM)

MANIPAL ACADEMY OF HIGHER EDUCATION

VI SEMESTER B.TECH END SEMESTER EXAMINATIONS, MAY 2023

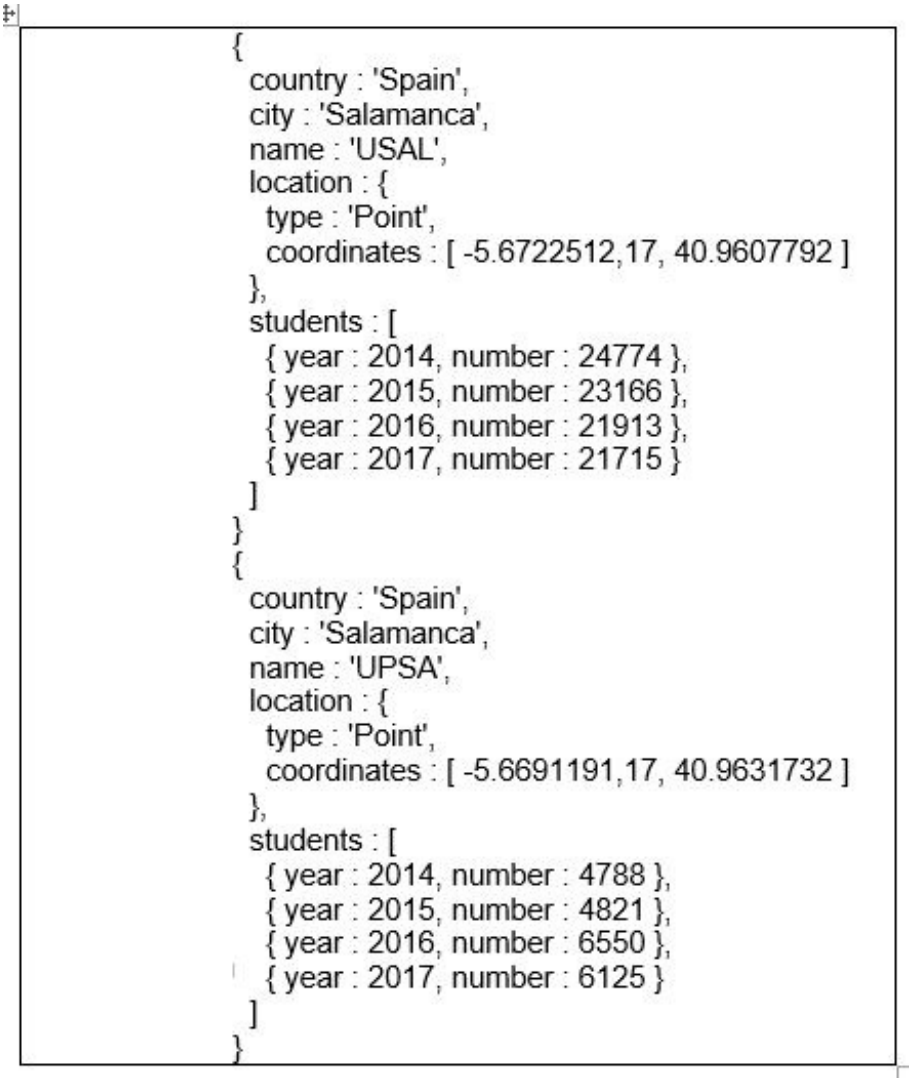
BIG DATA ANALYTICS [ICT 4034]

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

Answer all the questions.

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

1)		Big Data has the potential to significantly transform any business. What kind of value addition does this provide? Justify?	(5)
	A)		
	B)	Describe the structure of HDFS in hadoop ecosystem using a diagram.	(3)
	C)	Data that has high veracity and can be analyzed quickly have more value to a business. Justify with an example	(2)
2)		What is the significance of Resilient Distributed Dataset (RDD) in spark? Describe different ways to create RDDs in spark with example.	(5)
	A)		
	B)	Discuss the functionalities and communication protocols of datanode and namenode.	(3)
	C)	Identify and describe the three important components responsible for the query execution in Hive?	(2)
3)		Consider the sales.csv file with schema (product_name, price, payment_mode, city , country_of_client). Write hadoop map and reduce functions to compute the number of products sold in each country.	(5)
	A)		
	B)	What are the key differences between DataFrames and Resilient Distributed Datasets (RDDs)? Explain	(3)
	C)	Illustrate the rack-aware replica placement policy with an example.	(2)
4)		Determine how transformers and estimators fit in together using pipeline taking logistic regression as an example.	(5)
	A)		
	B)	Analyse the significance of different types of built-in streaming sources provided by Spark Streaming? Consider any two sources for each type. Write a pyspark code to create a new session and to get an existing session.	(3)
	C)	What is the purpose and importance of spark session in Apache Spark?	(2)
5)		The structure of 'universities' collection is given below in Figure Q5A.	(5)
	A)		



```
{
  country : 'Spain',
  city : 'Salamanca',
  name : 'USAL',
  location : {
    type : 'Point',
    coordinates : [ -5.6722512,17, 40.9607792 ]
  },
  students : [
    { year : 2014, number : 24774 },
    { year : 2015, number : 23166 },
    { year : 2016, number : 21913 },
    { year : 2017, number : 21715 }
  ]
}
{
  country : 'Spain',
  city : 'Salamanca',
  name : 'UPSA',
  location : {
    type : 'Point',
    coordinates : [ -5.6691191,17, 40.9631732 ]
  },
  students : [
    { year : 2014, number : 4788 },
    { year : 2015, number : 4821 },
    { year : 2016, number : 6550 },
    { year : 2017, number : 6125 }
  ]
}
```

Figure Q5A.

i. Create a collection named 'universities'

ii. Write a Mongodb query to display all the documents in the collection.

iii. Write a Mongodb query to find the country, university name and number of students in the year 2016.

iv. Write a Mongodb query to print the number of documents per university in the collection.

B) Compare and contrast the features of NoSQL and relational databases? (3)

C) Discuss briefly the working of any four common transformations supported by Spark Streaming on DStreams? (2)

-----End-----