Shri Vile Parle Kelavani Mandal's

DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING

(Autonomous College Affiliated to the University of Mumbai)

NAAC Accredited with "A" Grade (CGPA : 3.18)

Academic Year (2021-22)

Year:3 Semester:VI

Program: B. Tech. (Computer Engg.)
Subject: Big Data Infrastructure
Date: 05/07/2022

Max. Marks: 75
Time: 10:30 am to 1:30 pm
Duration: 3 Hours

REGULAR EXAMINATION

Instructions: Candidates should read carefully the instructions printed on the question paper and on the cover page of the Answer Book, which is provided for their use.

(1) This question paper contains 2 pages.
(2) **All Questions are Compulsory.**
(3) All questions carry equal marks.
(4) **Answer to each new question is to be started on a fresh page.**
(5) **Figures in the brackets on the right indicate full marks.**
(6) **Assume suitable data wherever required, but justify it.**
(7) Draw the neat labeled diagrams, wherever necessary.

Question No.		**Max. Marks**
Q1 (a)	Mr. John, an associate research scientist at the National Aeronautics and Space Administration's (NASA's) Goddard Space Flight Center, explained the role of 'satellite data' in vector-borne disease research. Vector-borne diseases are illnesses that are transmitted by vectors, which include mosquitoes, ticks, and fleas. These vectors can carry infective pathogens such as viruses, bacteria, and protozoa, which can be transferred from one host (carrier) to another. These diseases account for a significant number of human illnesses and deaths each year. Data generation is one of NASA's primary functions, and has several big climate- and weather-related datasets that can be applied to research on vector-borne diseases. These datasets include 35 years of sea surface temperatures and vegetation patterns, 37 years of precipitation amounts, and 16 years of land surface temperatures. These datasets are useful for disease research because, for example, sea surface temperatures affect precipitation, which in turn affects land surface temperatures and vegetation, creating conditions under which different disease vectors emerge and are able to propagate and spread disease. In particular, long-term datasets such as these are valuable because they enable the detection of anomalies, which by themselves are not important but their persistence over time is. Long periods of abnormally wet or dry conditions affect vegetation and are important for creating conditions in which vectors flourish. Mr. John thinks of this as a big data case study. Which all characteristics of big data are appropriate for this case study? Justify these Big data characteristics with respect to the above case study.	[05]
	OR	
	Justify beyond a certain limit traditional databases fail to handle Big Data in social networking domain.	[05]

Q1 (b)	Determine avg ratings of movies using map and reduce methods. Input file has a series of lines containing movie number, user number, rating out of 10 and a time stamp. Create a short documentation which briefly describes mapper and reducer task.	[10]
Q2 (a)	There are two separate datasets of a sports complex: cust_details: It contains the details of the customer. [customer_id, fname,lname,age,profession] transaction_details: It contains the transaction record of the customer. [transaction_id, date, cust_id, amount, game_type, equipments, city, state,mode_of_payment] Write Hadoop mapreduce code on above mentioned schema of the datasets to find out how many times a particular customer has visited sports complex **OR** Write pseudo code to perform union, intersection and difference using map reduce with an example.	[05] [05]
Q2 (b)	What are different components of Kafka? How replication is implemented in Kafka?	[10]
Q3 (a)	Draw and explain HDFS architecture **OR** Draw and explain HBASE architecture	[10] [10]
Q3 (b)	Write HIVE queries to perform any 5 different operations.	[05]
Q4 (a)	i. Describe any 5 functions/algorithms provided by Apache Spark MLlib. ii. Write spark code to count occurrence of each word in a file.	[05] [05]
Q4 (b)	Compare and contrast - Apache Spark vs MapReduce **OR** What is the significance of Resilient Distributed Datasets in Spark?	[05] [05]
Q5 (a)	Do you think MongoDB can replace traditional RDBMS completely? justify your answer with suitable case study. **OR** Distinguish between traditional Database and NoSQL database. Write any 5 different operations of MongoDB.	[10] [10]
Q5 (b)	Write PIG operators to perform any 5 different operations.	[05]

All the Best!