VIT

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act 1956)

Continuous Assessment Test II – Jun2023

Programme	:	M.TECH CSE(Business Analytics)-INTEGRATED	Semester	:	Fall-Inter 2023-24
Course Title	:	Big Data Framework	Code	:	CSE3120
			Class Nbr(s)	:	CH2022232500911 CH2022232500912
Faculty (s)	:	Dr. G. Suganeshwari Dr. R. Priyadarshini	Slot	:	F1
Time	:	1½ Hours	Max. Marks	:	50

Answer all the Questions

1.		Assume that you are recruited as a big data analyst intern for an IT consultancy. There was a recent article published by the consultancy firm "Ernst& Shine" which said that Hadoop continues to generate a positive Return On Investment (ROI) for enterprises. The same firm also predicted that Spark is poised to expand its market share in the enterprise computing space to handle huge amount of data in the finance domain, better in terms of in-memory computation and fault tolerence. Prepare a technical report for analyzing various characteristics of the data that has been collected during your internship. Compare and contrast both the frameworks with suitable diagrams.	10
2.		Consider the following code snippet in Spark, which performs a series of operations on a dataset: Line 1: data = sc.parallelize([1, 2, 3, 4, 5, 6, 7]) Line 2: mapped_data = data.map(lambda x: (x, x**2)) Line 3: filtered_data = mapped_data.filter(lambda x: x[1] > 10) Line 4: reduced_data = filtered_data.reduceByKey(lambda a, b: a + b) Line 5: sorted_data = reduced_data.sortByKey(ascending=False) Line 6: final_result = sorted_data.collect() In the context of this code, explain the following concepts: (2 marks each) 1. Lazy evaluation and its benefits in Spark. (3 marks) 2. The reduceByKey operation aggregates values based on the key in the given code snippet. Describe how the groupByKey can be used instead to achieve a similar outcome. (5 marks) 3. What is the final output for the above given code snippet? (2 marks) Provide a detailed explanation.	10

3. Perform transformation and action operations, display the results and write the code using python / scala. 10

i. Given the dataset [3, 5, 7, 9, 11, 13, 17] return the values in key /value pair where the key is the index of the corresponding vlaue in the array and value is three times the original value in the array, by applying a suitable transformation function.

(3 marks)

ii. Perform sort operation on the following data [(2,2), (2,3),(1,3),(1,4)], and reduce the resultant dataset using the samekey.(3 marks)

iii. Perform suitable transformation operations for the following words on the RDD such as ("Tomcruise", "Lenovo", "Joedanith", "Anvisha", "Jacky", "Jimmy", "Acer", "Zenith"), group the words and filter the same by more than 3 vowels and more than 3 consonants. (4 marks)

4. Consider the employee.txt file with the following attributes (Empid, Emp name, Salary, Dept. ID, Dept. name) separated by a tab. 10

i) Find the employees with the highest salary (2 Marks)

ii) Display the details of the employees with the lowest salary (2Marks)

iii) Count the number of employees in each department (2 Marks)

iv) Filter the employees details who belong to same department (2 Marks)

v) Find the average of salaries dispersed in all departments (2 Marks)

Emp id	Emp name	Salary	Dept. ID	Dept name
101	Joe	80000	21	Finance
102	Jhon	70000	32	Human Resource
103	Jill	90000	21	Automobile
104	Jack	65000	33	Sales
105	Sam	100000	32	Human Resource

Perform the above operations using Pyspark.

5. Write a Spark program to read the first n prime numbers from a text file named as "prime_nums.txt" and print the sum of those numbers at console. Whereas each row of the input file contains m numbers separated by spaces.(6 Marks) 10

Draw the RDD lineage for the above program.(4 Marks)

⇔⇔⇔

Reg. No.:

Name :

VIT

Vellore Institute of Technology

Continuous Assessment Test I– May 2023

Programme	:	**M.TECH CSE(Business Analytics)-INTEGRATED**	Semester	:	**Fall-Inter 2022-23**
Course Title	:	**Big Data Framework**	Code	:	**CSE3120**
			Class Nbr(s)	:	**CH2022232500911 CH2022232500912**
Faculty (s)	:	**Dr. G. Suganeshwari Dr. R. Priyadarshini**	Slot	:	**F1**
Time	:	**1½ Hours**	Max. Marks	:	**50**

Answer all the Questions

1. *A large-scale organization relies on a Hadoop cluster for data storage and data processing. The organization's data architecture includes a primary NameNode responsible for metadata management. Consider a situation where the primary NameNode encounters a failure. Illustrate the approaches to retain the metadata information during the NameNode failure.* 10

2. A healthcare dataset includes medical records and patient history with attributes like patient name, age, disease, CT scans images, X-rays images, and patient video screening. 10

   i) Identify and explain the key characteristics of the healthcare data. (5 marks)

   ii) List and illustrate the challenges while storing healthcare data in traditional database systems. (5 marks)

3. Write a mapper and reducer jobs to find the highest salary of an employee in each city for a given employee dataset. The employee dataset contains attributes such as employee id, employee's name, city, and salary. The sample of the given dataset is as follows. 10

   001, Harit, Delhi, 20000

   002, Hardy, Agra, 20000

   003, Amit, Delhi, 12000

   004, Anil, Delhi, 15000

   005, Deepak, Delhi, 34000

006, Fahed, Agra, 45000

007,Ravi, Patna, 98777

008, Avinash, Punjab, 120000

009, Saajan, Punjab, 54000

And also, draw the process flow of your map reduce program for the given sample employee dataset.

4. A popular streaming platform wants to enhance its movie recommendation system to provide personalized recommendations to its users. Analyze the stages involved in the data analytics lifecycle for the personalized movie recommendation system. 10

5. You are working as a data engineer in a large organization that recently adopted Hadoop for its data processing needs. The organization is interested in understanding the evolution of Hadoop (i.e., Hadoop1, Hadoop 2, Hadoop 3). And also discuss the commands which are used to access the HDFS in different versions of Hadoop. 10

⇔⇔⇔