CSES21

AHMEDABAD UNIVERSITY
School of Engineering and Applied Science
2025-2026 Monsoon

CSC521 Big Data Analytics
End-Term Examination
Wednesday 03 December 2025, 14:00 – 16:00

Name:	Roll no.	A	U	2							Marks /50

1. Answer the following questions in short (Any 10). ***(20)***

(a) What are structured and unstructured data? Explain with example.

(b) Differentiate between big data and conventional data.

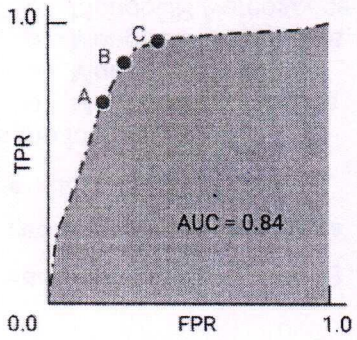
(c) Write different functions of Namenode and Datanode.

(d) Explain about data warehouse and data lake?

(e) What are the different components of Hadoop?

(f) If a cluster administrator increases the replication factor from 3 to 5 for higher data reliability, how will it affect the required Hadoop storage capacity? Explain the impact using the formula.

(g) You are developing a classifier to detect cancer from biopsy images. The positive class is cancer present and the negative class is cancer absent. The ROC curve of your classifier is shown below. Which of the following points on the ROC curve for your classifier is preferable? Give reasons.



(h) Why Mean Square Error (MSE) can't be used as the loss function in logistic regression? Write the cost function of logistic regression.

(i) Explain with example, How RDDs are resilient.

(j) I have trained my model and the training loss is 0.025 and the testing loss is 0.25. Is it a good model? What do you think is the issue in this model? Provide solution to tackle it.

(k) What is NoSql? How it is different from RDBMS?

2. Answer the following questions in detail (Any 6). ***(30)***

(a) What is Big Data? Explain velocity, veracity and value in Big data.

(b) Describe HDFS write architecture in detail.

(c) List different type of NoSQL databases. Explain any two of them in detail.

(d) Explain the difference between narrow and wide transformations with example in RDD.

(e) Consider below RDD. Explain how DAG scheduler will run it in single/multiple stage(s)? Clearly mention which transformations run in each stage and also their outputs.

```
rdd = sc.parallelize([("x", 2), ("y", 5), ("x", 4), ("z", 3), ("y", 1)], 2)

rdd1 = rdd.map(lambda x: (x[0], x[1] * 2))
rdd2 = rdd1.reduceByKey(lambda a, b: a + b)
result = rdd2.collect()
```

(f) Explain the followings:
- MongoDB Collections
- MongoDB Database
- MongoDB Documents
- MongoDB Namespace.

(g) Describe the working of Map-Reduce.

********** **BEST OF LUCK** **********