Faculty of Engineering
End Sem (Even) Examination May-2022
CS3ED04 Big Data Engineering

Programme: B.Tech.   Branch/Specialisation: CSE

Duration: 3 Hrs.   **Maximum Marks: 60**

Note: All questions are compulsory. Internal choices, if any, are indicated. Answers of Q.1 (MCQs) should be written in full instead of only a, b, c or d.

Q.1   i.   The average time required to search for an element in a hash table is _____.   **1**
            (a) O(1)   (b) O(n)   (c) log n   (d) None of these

      ii.   The base data structure of a Bloom filter is-   **1**
            (a) Array   (b) Bit Vector (c) Queue   (d) Hashmap

      iii. Pig Architecture does not consist of-   **1**
            (a) Pig Latin Program   (b) Map-Reduce plan
            (c) Hadoop Execution   (d) Batch Processing

      iv.   Which of the following can be used to store data on cloud?   **1**
            (a) DynamoDB   (b) Amazon S3
            (c) Pig   (d) Both (a) and (b)

      v.   What is T in ETL?   **1**
            (a) Translate   (b) Transform   (c) Transpose   (d) Table

      vi.   Which of the following component is not a part of flume agent?   **1**
            (a) Flume data sources   (b) Flume data channel
            (c) Flume data storage   (d) Flume data Sinks

      vii. Which one of the following is true for Apache Storm?   **1**
            (a) Is a non-distributed real-time big data-processing system
            (b) It is a distributed real-time big data-processing system.
            (c) It is a distributed batch-processing system.
            (d) It is a data warehouse system.

      viii. Apache Spark can be used with which of the following data source?   **1**
            (a) MySQL   (b) HBase   (c) Kafka   (d) All of these

      ix.   Which one of the following is not a big data visualization technique?   **1**
            (a) Symbol maps   (b) Line charts
            (c) Pie charts   (d) Area maps

P.T.O.

      x.   Which of the following is not a Supervised Machine learning technique?   **1**
            (a) Regression   (b) Clustering
            (c) Classification   (d) None of these.

Q.2   i.   What is Big Data?   **2**
      ii.   Explain Reduce phase in MapReduce.   **3**
      iii. Consider following points in a 2-D plane: (3, 6), (17, 15), (13, 15), (6, 12), (9, 1), (2, 7), (10, 19). Draw a KD tree by inserting the points one by one.   **5**

OR   iv.   What is bloom filter? Explain in detail how bloom filter works with the help of an example.   **5**

Q.3   i.   Differentiate among HBase and HDFS.   **4**
      ii.   Explain In memory distributed processing using Apache spark in detail.   **6**

OR   iii. How Amazon S3 works? What are various storage classes provided by Amazon S3?   **6**

Q.4   i.   What are various features provided by Sqoop?   **3**
      ii.   Explain how workflow management for Hadoop is done using Apache Oozie.   **7**

OR   iii. Describe steps involved in ETL process along with various validations/verifications preformed on each stage.   **7**

Q.5   i.   What are various applications of streaming data in industry?   **3**
      ii.   How real-time data pipelines are build using Apache Storm? Explain in detail.   **7**

OR   iii. Explain in detail streaming data using Apache Flume?   **7**

Q.6   Attempt any two:
      i.   Show steps of implementing regression in Spark MLLib.   **5**
      ii.   Explain anyone clustering algorithm with example.   **5**
      iii. Explain any one classification algorithm with example.   **5**

******

Marking Scheme
CS3ED04 Big Data Engineering

Q.1 i.   The average time required to search for an element in a hash table is   **1**
_____.
(a) O(1)

ii.   The base data structure of a Bloom filter is-   **1**
(b) Bit Vector

iii.   Pig Architecture does not consist of-   **1**
(d) Batch Processing

iv.   Which of the following can be used to store data on cloud?   **1**
(c) Pig

v.   What is T in ETL?   **1**
(b) Transform

vi.   Which of the following component is not a part of flume agent?   **1**
(c) Flume data storage

vii.   Which one of the following is true for Apache Storm?   **1**
(b) It is a distributed real-time big data-processing system.

viii.   Apache Spark can be used with which of the following data source?   **1**
(d) All of these

ix.   Which one of the following is not a big data visualization technique?   **1**
(d) Area maps

x.   Which of the following is not a Supervised Machine learning technique?   **1**
(b) Clustering

Q.2	i.	As per the explanation	2 Marks	**2**
	ii.	As per the explanation	3 Marks	**3**
	iii.	As per the solution	5 Marks	**5**
OR	iv.	What is bloom filter	1.5 Marks	**5**
		Explain in detail how bloom filter works with the help of an example	3.5 Marks	

Q.3	i.	As per the explanation	4 Marks	**4**
	ii.	As per the explanation	6 Marks	**6**
OR	iii.	How Amazon S3 works?	3 Marks	**6**
		What are various storage classes provided by Amazon S3?	3 Marks	

Q.4	i.	As per the explanation	3 Marks	**3**
	ii.	As per the explanation	7 Marks	**7**
OR	iii.	Describe steps involved in ETL process along with various validations/verifications preformed on each stage.	4 Marks / 3 Marks	**7**

Q.5	i.	As per the explanation	3 Marks	**3**
	ii.	As per the explanation	7 Marks	**7**
OR	iii.	As per the explanation	7 Marks	**7**

Q.6		Attempt any two:		
	i.	As per the explanation	5 Marks	**5**
	ii.	Explain anyone clustering algorithm with example.	3 Marks / 2 Marks	**5**
	iii.	Explain any one classification algorithm with example.	3 Marks / 2 Marks	**5**

******