Total No. of Questions: 6 *Total No. of Printed Pages:3*

Enrollment No......................................

Faculty of Engineering
End Sem Examination May-2023
CS3ED04 Big Data Engineering

Programme: B.Tech. Branch/Specialisation: CSE All

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. Assume suitable data if necessary. Notations and symbols have their usual meaning.

Q.1 i. Data in ______ bytes size is called Big Data. **1**

(a) Tera (b) Giga (c) Peta (d) Meta

ii. In Big Data environments, Velocity refers- **1**

(a) Data can arrive at fast speed

(b) Enormous datasets can accumulate within very short periods of time

(c) Velocity of data translates into the amount of time it takes for the data to be processed

(d) All of these

iii. Which is not a type of NoSql? **1**

(a) HBase (b) QBase (c) CouchDB (d) MongoDB

iv. Point out the correct statement- **1**

(a) MapReduce tries to place the data and the compute as close as possible

(b) Map Task in MapReduce is performed using the Mapper() function

(c) Reduce Task in MapReduce is performed using the Map() function

(d) All of these

v. ETL stands for- **1**

(a) Extract, Transfer and Load

(b) Extract, Transform and Load

(c) Extract, Time and Load

(d) Extract, Transform and Loss

vi. _______ are data processing units comprised of fact tables and dimensions from the data warehouse. **1**

(a) OLAP (b) Cubes (c) OLTP (d) None of these

P.T.O.

vii. Which of the following statements about data streaming is true? **1**

(a) Stream data is always unstructured data

(b) Stream data often has a high velocity

(c) Stream elements cannot be stored on disk

(d) Stream data is always structured data

viii. _________ is a distributed machine learning framework on top of Spark. **1**

(a) MLlib (b) Spark Streaming

(c) GraphX (d) RDDs

ix. Which of the following is finally produced by Hierarchical Clustering? **1**

(a) Final estimate of cluster centroids

(b) Tree showing how close things are to each other

(c) Assignment of each point to clusters

(d) All of these

x. What do we use to define a block of code in Python language? **1**

(a) Key (b) Brackets

(c) Indentation (d) None of these

Q.2 i. What is Big Data? **2**

ii. Explain Binary search tree Algorithm with suitable example. **3**

iii. Explain Hash Table and Hash Maps in details. **5**

OR iv. What is Map Reducing Techniques? Explain advantages of it **5**

Q.3 i. What is NoSQL? **2**

ii. Explain in memory distributed processing using Apache spark in detail. **8**

OR iii. How Amazon S3 works? What are various storage classes provided by Amazon S3? **8**

Q.4 i. What is Data Warehouse? **3**

ii. Explain different types of ETL operations in data warehousing. **7**

OR iii. Explain in detail Flume Workflow management for Hadoop using OOZIE Batch Processing on Cloud. **7**

Q.5 i. What is Processing of real time data? **4**

ii. How real-time data pipelines are build using Apache Storm? Explain in detail. **6**

OR iii. Explain in detail streaming data using Apache Flume? **6**

Q.6 Attempt any two:

i. Explain in detail 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 [T]- Big Data Engineering

Q.1	i)	C) Peta	**1**
	ii)	(D)All of the mentioned above.	**1**
	iii)	(B)QBase.	**1**
	iv)	(A) MapReduce tries to place the data and the compute as close as possible.	**1**
	v)	(B) Extract, Transform and Load.	**1**
	vi)	(B) Cubes.	**1**
	vii)	(B)Stream data often has a high velocity.	**1**
	viii)	(A) MLlib	**1**
	ix)	(B) tree showing how close things are to each other.	**1**
	x)	(C)Indentation.	**1**
Q.2	i.	Definition – 2 mark	**2**
	ii.	Definition – 2 mark Example - 1 mark	**3**
	iii.	Explanation Hash Table -2.5 marks Explanation Hash Maps -2.5 marks	**5**
OR	iv.	Map Reducing Techniques-2 mark Advantages each(Minimum 3) -1 mark	**5**
Q.3	i.	Definition – 2 mark	**2**
	ii.	Introduction -2 mark Application (1 each)-3 mark Explanation spark- 3 marks	**8**
OR	iii.	S3 work explanation -3 mark Various storage classs – 5mark(1 each)	**8**
Q.4	i.	Definition - 3 mark	**3**
	ii.	Each Etl Operation (Minimum 7)- 1 mark	**7**
OR	iii.	Explanation of Flume workflow – 4 mark Explanation of OOZIE Batch Processing- 3 mark	**7**
Q.5	i.	Explanation -3 mark	**4**
	ii.	Introduction- 2mark Application – 2 mark Diagram – 2 mark	**6**
OR	iii.	Introduction- 2mark Application – 2 mark Diagram – 2 mark	**6**
Q.6		Attempt Any two?	
	i.	Definition- 2 mark Explanation – 3 mark	**5**
	ii.	Definition- 2 mark Explanation – 3 mark	**5**
	iii.	Definition- 2 mark Explanation – 3 mark	**5**

******