Total No. of Questions: 6 *Total No. of Printed Pages:2*

Enrollment No.....

Faculty of Engineering
End Sem Examination May-2024
CS3ED07 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. What are the main components of big data? **1**
(a) HDFS (b) MapReduce
(c) YARN (d) All of these

ii. On which of the following platforms does Hadoop run? **1**
(a) Debian (b) Cross Platform
(c) Bare Metal (d) All of these

iii. A _____ node acts as the Slave and is responsible for executing a Task assigned to it by the JobTracker. **1**
(a) MapReduce (b) Mapper
(c) TaskTracker (d) JobTracker

iv. Pig operates in mainly how many nodes? **1**
(a) Two (b) Three (c) Four (d) Five

v. The process of _____ data involves converting it from one form to another. **1**
(a) Extracting (b) Transforming
(c) Loading (d) None of these

vi. Which of the following is NOT a supported data format for import/export in Sqoop? **1**
(a) CSV (b) Avro (c) Parquet (d) JSON

vii. What is a task in Apache Storm? **1**
(a) An instance of a spout or a bolt
(b) A message sent between spouts and bolts
(c) A log file generated by Apache Storm
(d) A database table used by Apache Storm

viii. Among the following options which component deals with ingesting streaming data into Hadoop? **1**
(a) Oozie (b) Hive (c) Kafka (d) Flume

ix. A database table used by Apache Storm- **1**
(a) Factor analysis
(b) Coefficient of partial correlation
(c) Coefficient of partial regression
(d) Coefficient of determination

x. The primary Machine Learning API for Spark is now the _____ based API. **1**
(a) DataFrame (b) Dataset (c) RDD (d) All of these

Q.2 i. Define big data engineering. **2**

ii. What are the three types of data? Explain with the examples. **3**

iii. What is the HashMap? Explain with an example. Also write the Java code for the HashMap. **5**

OR iv. Construct a KD tree and its graph for the following pairs: **5**
(6,2),(7,1),(2,9),(3,6),(4,8),(8,4),(5,3),(1,5), (9,5)

Q.3 i. Write any three applications of NOSQL. **3**

ii. What is NOSQL databases? Explain the types of NOSQL in detail. **7**

OR iii. Explain the Pig in detail with its architecture and features. **7**

Q.4 i. What is ETL? Explain. **3**

ii. Explain the Apache Oozie with its three types of jobs in detail. **7**

OR iii. Discuss the Apache Sqoop import and export method with a suitable diagram. **7**

Q.5 i. What is real time data processing? Explain with examples. **4**

ii. What is the Apache storm? Explain its core components with a diagram. **6**

OR iii. What is Apache flume? Write the steps to configure a flume agent. **6**

Q.6 Attempt any two:

i. Explain the K-Means clustering algorithm in detail. **5**

ii. Explain any one classification algorithm provided in Spark MLlib. **5**

iii. Discuss the regression analysis using linear and non-linear regression models. **5**

******

Marking Scheme

CS3ED07 (T) Big Data Engineering

Q.1	i)	D		**1**
	ii)	B		**1**
	iii)	C		**1**
	iv)	A		**1**
	v)	B		**1**
	vi)	D		**1**
	vii)	D		**1**
	viii)	A		**1**
	ix)	D		**1**
	x)	A		**1**

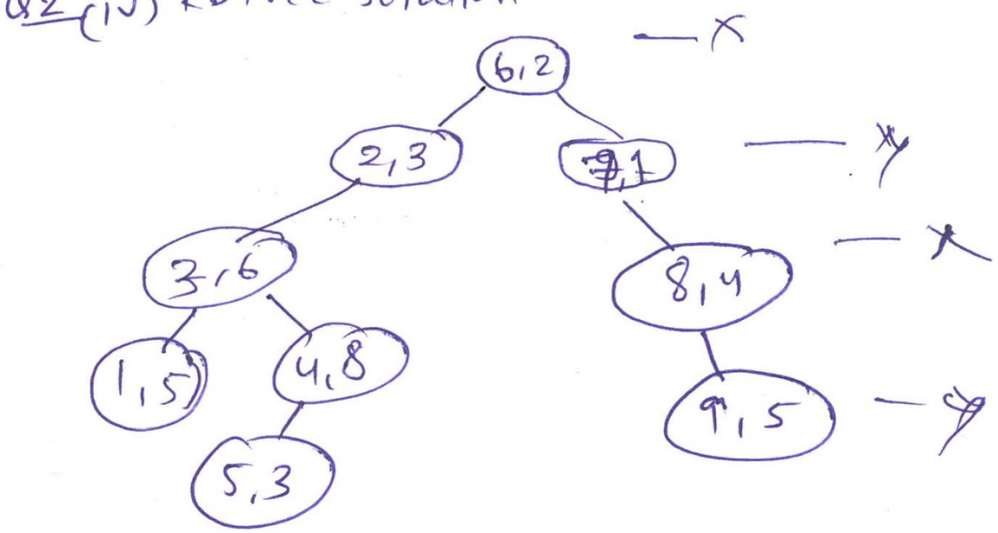
Q.2	i.	Define big data engineering. **Definition -2 Mark**	**2**
	ii.	What are the three types of data? Explain with the examples. **With Explaination-01 Mark for each (1x3=3)**	**3**
	iii.	What is the HashMap? Explain with an example.Also write the Java code for the HashMap. **Definition -02 Mark** **Example-01 Mark** **Java Code- 02 Mark**	**5**
OR	iv.	Construct a KD tree and its graph for the following pairs: (6,2),(7,1),(2,9),(3,6),(4,8),(8,4),(5,3),(1,5), (9,5) **Construction of KD-Tree step by step -03 Mark** **Graph Construction-02 Mark**	**5**
Q.3	i.	Write the applications of NOSQL.(Any Three) **01 Mark for Each (1x3=3)**	**3**
	ii.	What is NOSQL databases? Explain the types of NOSQL in detail. **Definition -01 Mark** **Explain any three types in detail-02 Marks of each(02x03=06)**	**7**
OR	iii.	Explain the Pig in detail with its architecture and features. **Definition- 02 Mark** **Architecture- 03 Mark** **Features- 02 Mark**	**7**
Q.4	i.	What is ETL?Explain. **Explain Extraction-01 Mark** **Explain Transformation -01 Mark** **Explain Load- 01 Mark**	**3**
	ii.	Explain the Apache Oozie with its three types of jobs in detail. **Definition- 1 Mark** **ExplainanThree Types - 2 Mark for each (2x3=6)**	**7**
OR	iii.	Discuss the Apache Sqoop Import and Export method with a suitable diagram. **Explanation of Apache Sqoop Import with its diagram- 3.5 Mark** **Explanation of Apache Sqoop Export with its diagram- 3.5 Mark**	**7**
Q.5	i.	What is real time data processing? Explain with examples. **Definition -02 Mark** **Examples any two -02 Mark**	**4**
	ii.	What is the Apache Storm? Explain its core components with a diagram. **Definition- 1 Mark** **Explain core components -03** **Diagram-02 Mark**	**6**
OR	iii.	What is Apache Flume? Write the steps to configure a Flume agent. **Definition - 1 Mark** **6 Steps- 1 Mark for each step**	**6**
Q.6		Attempt any two:	
	i.	Explain the K-Means Clustering algorithm in detail. **Explanation -02 Mark** **Algorithm-03 Mark**	**5**
	ii.	Explain any one classification algorithm provided in Spark MLlib. **Explanation of any one classification algorithm-05 Mark**	**5**
	iii.	Discuss the regression analysis using linear and non-linear regression models. **Linear Regression- 2.5 Mark** **Non-Linear Regression 2.5 Mark**	**5**

[2]

[3]

******

Q2 (iv) KD tree solution



******

P.T.O.