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

Enrollment No.....

Faculty of Engineering

End Sem (Odd) Examination Dec-2019

CS3ED04 Big Data Engineering

Programme: B.Tech. Branch/Specialisation: CS

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. Input to the _____ is the sorted output of the mappers. **1**
(a) Reducer (b) Mapper (c) Shuffle (d) All of these

ii. All of the following accurately describe Hadoop, EXCEPT: **1**
(a) Open source (b) Real-time
(c) Java-based (d) Distributed computing approach

iii. Which of the following is not a NoSQL database? **1**
(a) SQL Server (b) MongoDB
(c) Cassandra (d) None of these

iv. In any MapReduce Job Hbase can be used as a **1**
(a) Metadata store (b) Data source
(c) Data node (d) Metadata node

v. _____ tool can list all the available database schemas. **1**
(a) Sqoop-list-tables (b) Sqoop-list-databases
(c) Sqoop-list-schema (d) Sqoop-list-columns

vi. Oozie Workflow jobs are Directed _____ graphs of actions. **1**
(a) Acyclical (b) Cyclical (c) Elliptical (d) All of these

vii. What are the different channel types in Flume? **1**
(a) Memory Channel (b) JDBC Channel
(c) File Channel (d) All of these

viii. Why Apache Storm is the first choice for real-time processing? **1**
(a) Easy to operate (b) Real fast
(c) Both (a) and (b) (d) None of these

ix. In cluster sampling, elements of selected clusters are classified as **1**
(a) Elementary units (b) Primary units
(c) Secondary units (d) Proportional units

P.T.O.

x. A multiple regression model has **1**
(a) Only one independent variable
(b) More than one dependent variable
(c) More than one independent variable
(d) None of these

Q.2 i. What do you mean by Big Data? **2**

ii. Differentiate between Hash Table and Hash Map. **3**

iii. How the MapReduce is used for distributed algorithm design? Explain with the help of example. **5**

OR iv. Explain the following terms: **5**
(a) Bloom Filters (b) KD Tree

Q.3 i. Define Apache spark. **2**

ii. How NoSQL is used for big data storage. Also explain different types of NoSQL databases and its characteristics. **8**

OR iii. Define the following terms: **8**
(a) HBase (b) Pig

Q.4 i. Define OOZIE Hadoop work flow with diagram. **3**

ii. Explain the concept of ETL with its operations. Also explain how its operation can be used for data warehousing with diagram. **7**

OR iii. How Sqoop and flume work flow management is used for ingesting data into big data platform? **7**

Q.5 i. Differentiate between streaming data and real time streaming data with its applications. **4**

ii. How Apache spark is used for streaming the data? Explain with suitable architecture. **6**

OR iii. Explain real time data pipeline with the help of Apache storm. **6**

Q.6 Attempt any two:

i. What do you mean by regression? Explain its types in details. **5**

ii. Explain types of clustering with examples. **5**

iii. How classification can be done using spark MLLib? **5**

******

Marking Scheme
CS3ED04 Big Data Engineering

Q.1	i.	Input to the _____ is the sorted output of the mappers. (a) Reducer		**1**
	ii.	All of the following accurately describe Hadoop, EXCEPT: (b) Real-time		**1**
	iii.	Which of the following is not a NoSQL database? (a) SQL Server		**1**
	iv.	In any MapReduce Job Hbase can be used as a (b) Data source		**1**
	v.	_____ tool can list all the available database schemas. (b) Sqoop-list-databases		**1**
	vi.	Oozie Workflow jobs are Directed _____ graphs of actions. (a) Acyclical		**1**
	vii.	What are the different channel types in Flume? (d) All of these		**1**
	viii.	Why Apache Storm is the first choice for real-time processing? (c) Both (a) and (b)		**1**
	ix.	In cluster sampling, elements of selected clusters are classified as (a) Elementary units		**1**
	x.	A multiple regression model has (b) More than one dependent variable		**1**
Q.2	i.	Big Data Definition contains 3 Vs		**2**
	ii.	Three differences between Hash Table and Hash Map. 1 mark for each	(1 mark *3)	**3**
	iii.	MapReduce Distributed algorithm design	3 marks 2 marks	**5**
OR	iv.	Explain the following terms: (a) Bloom Filters (b) KD Tree	 2.5 marks 2.5 marks	**5**
Q.3	i.	Definition of Apache spark.		**2**
	ii.	NoSQL Four types of NoSQL databases Its characteristics	2 marks 4 marks 2 marks	**8**
OR	iii.	(a) HBase (b) Pig	4 marks 4 marks	**8**
Q.4	i.	OOZIE Hadoop work flow with diagram.		**3**
	ii.	Concept of ETL with its operations Data warehousing with ETL	3 marks 4 marks	**7**
OR	iii.	Sqoop work flow management Flume work flow management	3.5 marks 3.5 marks	**7**
Q.5	i.	Difference b/w streaming data and real time streaming data 4 differences 1 mark for each	 (1 mark * 4)	**4**
	ii.	Apache spark for streaming Architecture	3 marks 3 marks	**6**
OR	iii.	Proper explanation using storm Apache storm.		**6**
Q.6		Attempt any two:		
	i.	Regression Its types	1 mark 4 marks	**5**
	ii.	Clustering Types of clustering with examples	1 mark 4 marks	**5**
	iii.	Classification and spark MLLib 2.5 marks for each	 (2.5 marks *2)	**5**

******