Roll No.: _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _

Amrita Vishwa Vidyapeetham

Amrita School of Engineering, Coimbatore

B.Tech. Degree Examinations – April/May 2018

Eighth Semester

Computer Science and Engineering

CSE459 Big Data Analytics

[Time: Three hours Maximum: 100 Marks]

Answer all questions

PART A **(10 x 3 =30 Marks)**

1. Identify the data type for the given examples:

   (a) JSON documents

   (b) Rador data

   (c) Social Security numbers

2. Define the following big data characteristics

   (a) Veracity and validity

   (b) Volatility

   (c) Variability

3. Compare Hadoop 1 and Hadoop 2.

4. What are the values of default block size in Hadoop 1 and Hadoop 2? Is it possible to change the block size?

5. What is dynamic partitioning in Hive? When it is used?

6. Briefly elucidate the usage of “GROUP”, “DISTINCT”,” ORDER BY” keywords in Pig scripts.

7. List the types of NoSQL databases with suitable example.

8. Write the equivalent mongo DB queries for the given SQL commands:
   (a) Select name, salary from employees where designation =” Manager”
   (b) Select *from employees order by salary
   (c) Select * from employees where salary >30000

9. Give short notes about tunable consistency in Cassandra

10. Find the Euclidean distance between the two points (-1,2,3) and (4,0, -3)

PART B **(7 x 10 =70 Marks)**

11. (a) How does HDFS Daemons (Name node, Data node and Secondary name nodes) work to maintain the HDFS file system? Explain with suitable diagrams. (5)
(b) Write a map-reduce program that display the count of all words that begin with letter 't' in the document (5)

12. (a) A Client having some E –commerce data which belongs to India operations in which each state (38 states) operations mentioned in as a whole. Perform the following task using Hive. (5)
(i) Creation of table all states with 3 column names such as state, district, and enrolment.
(ii)Loading data into table all states.
(iii) Creation of partition table with state as partition key.
(iv) Using buckets cluster, the states

(b) Write hive queries for the following: (5)
(i)Create a database named as RETAIL store.
(ii)Create a table retail with the fields txnno, custno, amount, category, product, city, state.
(iii)Load the data into table
(iv)Find the total number of records in the table.
(v) Find the total no of records based on city.

13. (a) Consider the following toy dataset, write a PIG script for the following operations: (6)
(i)Select products whose quantity is greater than or equal to 1000.
(ii)Select products whose quantity is greater than 1000 and year is 2001
(iii)Select products with year not in 2000

Year	**Product**	**Quantity**
2000	iphone	1000
2001	Iphone	1500
2002	Iphone	2000
2000	Nokia	1200
2001	Nokia	1500
2002	Nokia	900

(b) Perform a map reduce program to count the occurrence of the words using Pig. (4)

14. (a)Give MongoDB queries for the following: (6)

(i)Create a collection “movies”

(ii)Insert the following data in to the movie collection:

title : The Hobbit: An Unexpected Journey
writer : J.R.R. Tolkein
year : 2012
franchise : The Hobbit

title : The Hobbit: The Desolation of Smaug
writer : J.R.R. Tolkein
year : 2013
franchise : The Hobbit

title : The Hobbit: The Battle of the Five Armies
writer : J.R.R. Tolkein
year : 2012
franchise: The Hobbit
synopsis : Bilbo and Company are forced to engage in a war against an array of combatants and keep the Lonely Mountain from falling into the hands of a rising darkness.

(iii)Get all the documents with franchise set to "The Hobbit"

(iv) Add a synopsis to "The Hobbit: An Unexpected Journey”: "A reluctant hobbit, Bilbo Baggins, sets out to the Lonely Mountain with a spirited group of dwarves to reclaim their mountain home - and the gold within it - from the dragon Smaug."

(v) Get all titles released after the year 2000.

(b) Consider the following document which contains the name of product and price. (4)

{name: xxx, price:9}

{name: xxx, price:12}

{name: bbb, price:8}

{name: yz, price:3}

{name: yz, price:5}

Write a MapReduce program to count the price for all the items with same name.

15. Assume the example of a social music service songs data set having an id, song order, album, artist, Song id, title. The table uses a id and song order as a primary key. Write a Cassandra queries for the following: (10)

(i) Create a key space named as Music
(ii) Change the Music key space as current working directory

(iii) Create a table called songs with the above-mentioned fields.

(iv) Insert the following values in to the table

Id	Song order	Album	Artist	Song id	title
626092	1	Tres Hombres	ZZ top	a3e	La Grange
626094	2	We must obey	Fu Manchu	a1f	Moving in stereo
626096	3	Roll away	Back Door slam	b0i	Outside Woman Blues

(v) Display the details of artist “Fu Manchu.

(vi) Sort the data in descending order based on song order and display only 10 records.

(vii) Create an index for the artists filed.

16. Determine the regression equation by using the regression slope coefficient and intercept value as shown in the regression table given below. (10)

X Values	Y Values
55	52
60	54
65	56
70	58
80	62

17. Apply Apriori algorithm with minimum support 50% for the given dataset and find all frequent item sets. (10)

TID	Onion	Potato	Burger	milk	juice
1	1	1	1	0	0
2	0	1	1	1	0
3	0	0	0	1	1
4	1	1	0	1	0
5	1	1	1	0	1
6	1	1	1	1	1

****