Roll No.: _

Amrita Vishwa Vidyapeetham

Amrita School of Engineering, Coimbatore

B.Tech. Degree Examinations – March/April 2019

Eighth Semester

Computer Science and Engineering

15CSE334 Big Data Analytics

[Time : Three hours Maximum : 100 Marks]

CO	Course Outcomes
CO01	Understand fundamental concepts of Big Data and its technologies
CO02	Apply concepts of MapReduce framework for optimization
CO03	Analyze appropriate NoSQL database techniques for storing and processing large volumes of structured and unstructured data.
CO04	Apply data analytics solutions using Hadoop ecosystems
CO05	Explore modern reporting tools for Machine learning.

Answer all questions

PART A **(10 x 2 =20 Marks)**

1. Define Big Data and explain the Vs of Big Data. [CO01]
2. List the difference between SQL and NOSQL databases. [CO03]
3. Name the different commands for starting up and shutting down Hadoop Daemons. [CO02]
4. Replication causes redundancy then why it is pursued in HDFS? [CO02]
5. Name the default metastore in Hive. Suppose we have installed Apache Hive on top of Hadoop cluster using default metastore configuration. Then, what will happen if we have multiple clients trying to access Hive at the same? [CO04]
6. What is the advantage of storing data in RC File format. Illustrate with an example [CO04]
7. What is the function of UNION and SPLIT operators? Give examples. [CO03]
8. Write equivalent MongoDB queries [CO03]

MySQL	MongoDB
INSERT INTO users (user_id, age, status) VALUES ('bcd001', 45, 'A')	
SELECT * FROM users	
UPDATE users SET status = 'C' WHERE age > 25	

9. What is Gossip Protocol? [CO03]
10. Write two algorithm that support a) Supervised learning b) Unsupervised learning [CO05]

PART B **(8 x 10 =80 Marks)**

11. a) Explain the properties of CAP theorem and list databases that follow CAP theorem. (4)[CO01]

    b) Give 3 examples of NOSQL databases with their type (3)[CO01]

    c) Explain the following terms: (3)[CO01]

    i) Symmetric Multiprocessor system
    ii) Parallel processing system
    iii) Distributed system

12. a) Explain HDFS architecture and various HDFS daemons that run in HDFS system. (5)[CO02]

b) A Telecom company keeps records for its subscribers in specific format. Consider followingformat

FromPhoneNumber|ToPhoneNumber|CallStartTime|CallEndTime|STDFlag

Now we have to write a map reduce code to find out all phone numbers who are making more than 60 mins of STD calls. Here if STD Flag is 1 that means it was as STD Call.

What are the different phases of a mapper and reducer task? How are these phases invoked for the above program? Explain with a block diagram (5)[CO02]

Sample Input:

FromPhoneNumber|ToPhoneNumber|CallStartTime|CallEndTime|STDFlag

9665128505|8983006310|2015-03-01 09:08:10|2015-03-01 10:12:15|1

9665128505|8983006310|2015-03-01 07:08:10|2015-03-01 08:12:15|0

13. a) Write Queries in HIVE for HOSTEL table given below.Create the table and load the data (5)[CO04]

RoomNo	NoS	Floor	Name
1	4	1	ABC
2	4	1	ABC
3	2	2	ABC
4	2	1	XYZ
5	2	2	XYZ
6	2	2	XYZ

i) Display the total number of students in each hostel

ii) Display total number of rooms in each hostel that has less than 3 students

iii) Display all hostels that have their name ending with 'C'

iv) Display names of hostels that have more than 3 floors

v) Create hostel table using dynamic partition, partition the data using floor variable and insert the data into the table

b) Write a user defined function in hive to display all hostel names in hostel table. (5)[CO04]

14. a) Write Queries in PIG for train table given below.Create the table and load the data (6)[CO04]

Tid	Name	Place	Dist	Status	cost
1	Cheexp	che	450	travel	340
2	Blrexp	blr	380	travel	450
3	Delexp	del	560	travel	800
4	blrexp1	blr	670	reach	900
5	bmexpr	che	600	travel	1200
6	Sabex	che	1200	reach	5000
7	Vizexp	del	3400	reach	2300

i) Display trains that have a distance >600

ii) Display train that has the least cost to a city

iii) Display the total number of trains to same city

iv) Display per-km cost for each of the trains

v) update cost by 10% for each of the records

vi) Create a table train that has fields tid,name, place, where place is a tuple with source and destination information. Write the load command and display the records.

b) Illustrate with an example map reduce programming in PIG (4)[CO04]

15. a) What are the different data layers of Cassandra, How does read and write happens using these data layers. Explain with a diagram. (5)[CO03]

b) Write CQL to perfom the following activities. (5)[CO03]

i) Create a "OFFICE" keyspace with replica factor of 3 and simplestrategy class.

ii) Create a table "Employee" in college keyspace with <Eid, name, dept, email, salary> with Eid as the primary key. The email field will be represented as set datatype. Insert a record into the table.

iii) Display employee of sales dept.

iv) Insert a record into employee table such that the record is alive for 30 sec

v) Display total number of employees in the table

16. a) Write Queries in MONGODB for **subject table** given below. Create the table and insert records into the table (6)[CO03]

Cid	Name	Dept	Nos	Credit	Lab
1	C++	CSE	60	3	YES
2	OS	CSE	40	3	NO
3	PHY	SCI	35	4	YES
4	CHE	SCI	40	4	YES
5	BIGDATA	CSE	60	4	YES
6	SPM	CSE	60	3	NO
7	MAT	SCI	63	4	NO

i) Find the total number of students in each dept.

ii) Find courses in CSE that have strength less than 60.

iii) Count the total number of 3 credit courses in the table.

iv) Update record with name CHE, change credit value to 3.

v) Use cursors and display all courses in subject table.

vi) List two courses that have minimum no of students.

b) Write a map reduce program to count the number of subjects in each dept. (4)[CO03]

17. a) The values of x and their corresponding values of y are shown in the table below. Find the least square regression line y = a x + b. Estimate the value of y when x = 10. (5)[CO05]

x	0	1	2	3	4
y	2	3	5	4	6

b) Generate frequent items sets using Apriori algorithm for the given data: (5)[CO05]

Support Count = 2

TID	Items
1	Milk, Tea, Cake
2	Egg, tea, Cold Drink
3	Milk, Egg, Tea, Cold Drink
4	Egg, Cold Drink
5	Juice

18. a) Implement K means algorithm for the given data using K =2.Assume the 2 centroid are: m1 = (1, 1) and m2=(6, 6). The data points are (5)[CO05]

(1 , 1)
(1, 2)
(3,4)
(6, 6)
(6, 5)

b) Cluster theabove data given above using Complete Linkage Strategyand draw the dendogram. (5) [CO05]

****