U.S.N.										

B.M.S. College of Engineering, Bengaluru-560019

Autonomous Institute Affiliated to VTU

September / October 2023 Supplementary Examinations

Programme: B.E **Semester: VI**
Branch: Information Science and Engineering **Duration: 3 hrs.**
Course Code: 20IS6PEBDA **Max Marks: 100**
Course: Big Data Analytics **Date: 22.09.2023**

Instructions: 1. Answer any FIVE full questions, choosing one full question from each unit.
2. Missing data, if any, may be suitably assumed.

Important Note: Completing your answers, compulsorily draw diagonal cross lines on the remaining blank pages. Revealing of identification, appeal to evaluator will be treated as malpractice.

UNIT - I

1 a) Is it important to analyse the unstructured data? Justify your answer with an example. **07**

b) Web Pages are said to be unstructured data even though they are defined by the HTML mark-up language, which has rich structure. Justify. **06**

c) Describe the Hadoop Components with a neat diagram. **07**

UNIT - II

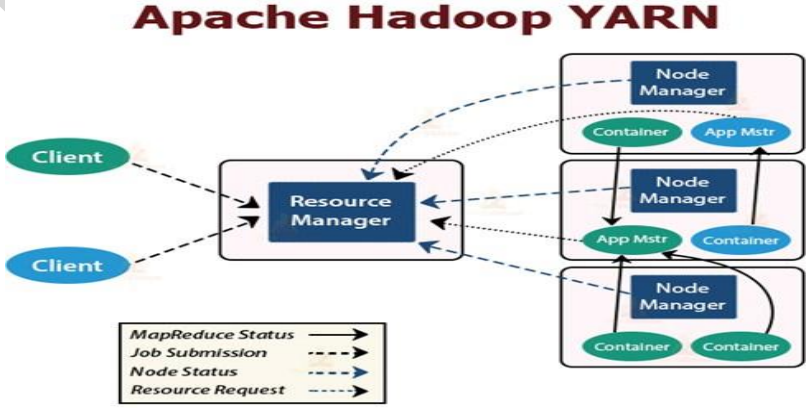
2 a) Assume a file by name example.txt whose contents are as follows: Dear,Bear,River,car,Car,River ,Deer,Car and Bear. Perform a word count on the sample.txt using MapReduce. **10**

b) Write the steps to describe the working Model of Mapreduce programming to perform its task. **10**

OR

3 a) Develop a mapreduce program in finding no of visitors who visited your linked in profile. **10**

b) For the given diagram provide the execution sequence ofoperations **10**



UNIT - III

4 a) Write CQL commands for creating a keyspace and table with different replication strategies for a database of storing information about the trains arrival and departure time **10**

b) Create tables in hive for voter with appropriate attributes and to store it in textfile, sequence file, rcfile , avro, orc, parquet file formats **10**

OR

5 a) i)Create a managed table and external table in Hive with name doctor with id, name,fees and area of specialization with comments "doctor details"fields ternminated by "\t" lines terminated by "\n" and stored in TEXT FILE. **10**

ii) Write hive commands to perform the following

I. change the name of the table
II. change the name of field from id to doc_id
III. change the data type of fees from float to double
IV. drop the column areaofspecialization
V. add column experience

b) How does Cassandra handle the cluster consistency process when replication owing nodes fails? **05**

c) Explain partitions and buckets in Hive with examples **05**

UNIT - IV

6 a) Depict the steps involved in anatomy of a spark job run **05**

b) Create RDD from a parallize collection **10**

i. from a csv file
ii. from a text file
iii. from JSON file
iv. From existing RDD

for the students with attributes id, name, marks, phone no, branch and semester.

c) i)Write commands in scoop to import data from MYSQL to HDFS specified directory. **05**

ii)
```
sqoop import \
--connect jdbc:mysql://localhost/userdb \
--username root \
--table student_add \
--m 1 \
--where "city ='sec-bad'" \
--target-dir /wherequery.
```

Predict the output of the following command applied for student_add table with fileds id, hno, street, city.

UNIT - V

7 a) Explain the limitations of flume **05**

b) Consider there are N number of nodes in a cluster. Analyze in detail how a leader node can be elected in a Zookeeper ensemble. **08**

c) Depict the Zookeeper workflow with a diagram. **07**

****