THE CATHOLIC UNIVERSITY OF EASTERN AFRICA

A. M. E. C. E. A

P.O. Box 62157
00200 Nairobi - KENYA
Telephone: 891601-6
Ext 1022/23/25

MAIN EXAMINATION

JANUARY – APRIL 2022

FACULTY OF SCIENCE

DEPARTMENT OF COMPUTER SCIENCE

REGULAR PROGRAMME

CMT 428: HADOOP FRAMEWORK AND DFS

Date: APRIL 2022	**Duration: 2 Hours**
INSTRUCTIONS: Answer Question ONE and any TWO Questions	

SECTION A

QUESTION ONE

a) Use the following statement to answer the questions that follow:

> ***"****A MapReduce is a data processing tool which is used to process the data parallelly in a distributed form. It was developed in 2004, on the basis of paper titled as "MapReduce: Simplified Data Processing on Large Clusters," published by Google****"***

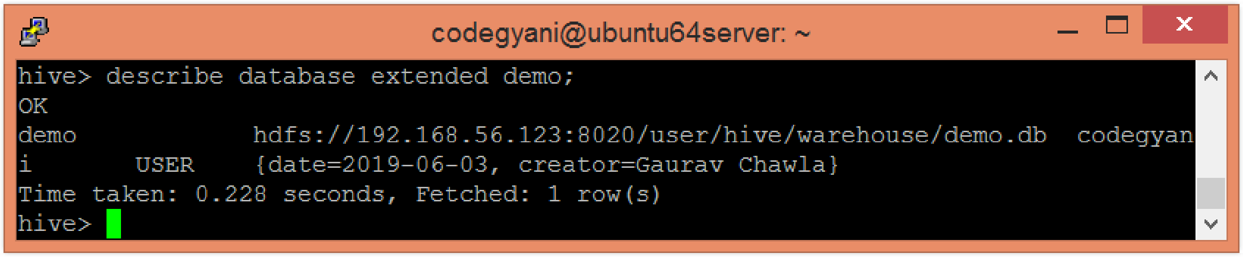
i) Outline **THREE** to be following when implementing above statement in MapReduce **[3 Marks]**

ii) With clear diagram and steps, map and reduce the above statements using keys and values **[7 Marks]**

iii) Citing THREE examples of each, discuss **THREE** usage of MapReduce **[6 Marks]**

b) Describe **FOUR** characteristics of HBase **[4 Marks]**

c) Write a snippet code that would give the following output in Hive. **[5 Marks]**



```
hive> describe database extended demo;
OK
demo            hdfs://192.168.56.123:8020/user/hive/warehouse/demo.db  codegyan
i       USER    {date=2019-06-03, creator=Gaurav Chawla}
Time taken: 0.228 seconds, Fetched: 1 row(s)
hive>
```

d) what happens, when the server hosting a MemStore that has not yet been flushed crashes? Explain **[5 Marks]**

Q2.

a) Briefly explain **THREE** components that Hadoop offers **[6 Marks]**

b) With examples, explain the role of Hadoop in the following cases: **[8 Marks]**

i) Link analysis

ii) Graph data processing

iii) Data stream mining

iv) Large-scale machine learning

c) Outline **FOUR** ssymptoms of over fitting **[4 Marks]**

d) In Spark, what is the difference between Action operations and Transformation operations? Give **TWO** examples of each. **[2 Marks]**

Q3. a) In Spark, draw a figure that shows the difference between a Narrow Dependency and a Wide Dependency? And Then, Give one example operation in each case.

[7 Marks]

b) Highlight **THREE** modes in which Hadoop can be run. **[3 Marks]**

c) Discuss why we use HDFS for applications having large data sets and not when there are lot of small files. **[6 Marks]**

d) Differentiate between the following: **[4 Marks]**

i) Reducer and Combiner

ii) Correlation and causality

Q4.

a) Using clear diagram, illustrate the how data can be moved into HDFS/hive/hbase from MySQL/ PostgreSQL/Oracle/SQL Server/DB2 and vise versa in Sqoop. **[6 Marks]**

b) Discuss **THREE** components that makes Hadoop an apache framework **[6 Marks]**

c) Compare and contrast RDBMS vs Hadoop **[8 Marks]**

Q5.

a) Discuss **THREE** important objectives for a Hadoop Distributed File System (HDFS) **[6 Marks]**

b) Differentiate between static partitioning and dynamic partitioning. **[4 Marks]**

c) Given the following operations/algorithms:
K-Means Clustering, Logistic Regression, Aggregation operation, Selection operation, Building the model of Naïve Bayes Classifier, Page Rank
State which one(s) will have similar performance (and also mention why) if executed on either Hadoop or Spark infrastructures. Assume the input data in all cases is initially stored on HDFS **[10 Marks]**

END