Examination Question Paper
PG-DBDA SEP 2021

Module Name: Big Data Technologies

Exam Type: Main

DATE: 24-01-2022 **Duration: 2.5 hrs** **Max. Marks: 40**

Exam Instructions:

1. You will be given 7 questions and out of that you need to answer 5 questions, each may carry a different weightage
2. Total marks assigned for this exam is 40
3. Total time provided to complete the assignment will be 2.5 hours
4. While submitting the exam, candidates need to submit the following:
   a. Source code (Java / SQL / Python)
   b. Input (as provided)
   c. Output
   d. Detailed steps in case of installation assignments
5. Candidates must mention the environment being used for the assignment. For e.g VirtualBox VM / Docker / Standalone.
6. Unless specified, candidates can use their implementation language (Java/Scala/Python) of choice.

Pre-requisites:

1. Each has to rely on his/her hardware.
2. Candidates have an option to either use standalone installation / VirtualBox / Docker images.
3. Candidates are assumed to have good hands on over the technology stack they use for installation. For e.g: If a candidate decides to use Docker, he/she should be well-versed with docker and no assistance will be provided for performing operations on docker.

Instructions for Uploading:

- Students should upload only zip file
- Kindly ensure before creating zip file, all the required files should be closed to avoid any corruption of the zip file
- Please make sure all file/folders are copied to the main folder before making zip file
- Zip File should be renamed full PRN followed by exam name. (For example: 210940125001-BDT.zip)
- Uploading is allowed only once

Q1. Below is the data for the fci_stock_position_commodity_Rice from the market. **10 Marks**

Sample data: (Please use the attached file: fci_Stock_Position_commodity_02_Rice-Raw_Tamil_Nadu-2021.csv)

Date	**Code**	**CommodityId**	**CommodityName**	**DistrictName**	**DistrictCode**	**Stock**	**CommodityStock**
2021-01-01	Tamil Nadu	2	Rice-Raw	COIMBATORE	SE12	717747.92	1174193.92
2021-01-01	Tamil Nadu	2	Rice-Raw	CHENNAI	SE14	96759.01	1174193.92
2021-01-01	Tamil Nadu	2	Rice-Raw	THOOTHUKKUDI	SE15	174754.11	1174193.92

Data - Record date
Code - District Code
CommodityId - Commodity ID
CommodityName - Name of commodity
DistrictName - Name of District
DistrictCode - Code for district
Stock - Stock as of Date
CommodityStock - Commodity stock as of Date

Software Stack needed: Spark, HDFS (using data from HDFS is optional)

Analyze it using Spark and answer the following questions (2 marks each):

1. Find all the districts participating in the data collection
2. Find the district with max Stock and CommodityStock
3. Find the district with max Total Stock, where Total Stock = Stock + CommodityStock
4. Find the average of all the Stock for “Chennai”
5. Find the district with min of avg of commodity stock

Note: Answer either Q2 or Q3

Q2. Create a hive table for analyzing market data. (fci_Stock_Position_commodity_02_Rice-Raw_Tamil_Nadu-2021.csv) Write Hive queries for the following: - **10 Marks**

Software Stack needed: Hive, HDFS

1. Create an external table using hive query language
2. Describe the external table.
3. Ingest data into the hive table using the load of dfs command.
4. Create another table partition on the hive table on the column. Select any appropriate column for partition.
5. Copy the contents from table created in step#1 into table created in step#4.

OR

Q3. Perform the following operations on HBase: **-10 Marks**

Software Stack needed: HDFS, HBase

1. Create a table “mytable” in Hbase with 2 column families, “mycolfamily1”, “mycolfamily2”
2. Insert data into two columns (named col1, col2) in first column family (“mycolfamily1”)
3. Retrieve the entire inserted data and display it on top of the screen.
4. Retrieve only the data from col1 of “mycolfamily1”
5. Drop the hbase table

Q4. Write the HDFS commands for the following operations. **– 10 Marks**

Stack needed: HDFS

1. Create a directory in HDFS
2. Ingest a text file into HDFS
3. List the contents of the directory created in step #1
4. Print the contents of the file ingested in step #2
5. Change the replication factor of file ingested in step #2
6. Display the changed replication factor
7. Change the permissions of file ingested in step #2 to 644
8. Change the ownership of a file ingested in step #2
9. Set the file created ingested in step 2, to size 0 (zero)
10. Delete the directory created in step 1

Q5. Create a pyspark streaming program to perform the following. **– 5 Marks**

1. Ingest the data from file dir (say from /tmp/inputDir) into HDFS.
2. The data should be copied into HDFS as soon as the file is copied into the input directory.
3. The data should be copied into parquet format.
4. Use the data provided in Q1

Note: Answer either Q6 or Q7

Q6. Create a Map-Reduce program for wordcount in README.txt using **any one** of the following options **– 5 Marks**

Software Stack needed: HDFS, Mapreduce, Java / Python + hadoop-streaming

1. Using Java for wordcount
2. Using Python and hadoop-streaming for wordcount

OR

Q7. Create an Apache Airflow to implement a pipeline for triggering the following actions. **– 5 Marks**

1. Any BashOperation (say, echo "PGDBDA")
2. Any pySpark Operation using Spark Operator Software

Stack needed: Airflow, Spark

All the best