Syed Babar Ali
School of Science and Engineering

CS5304 Big Data Services & MLOps – Spring 2025 (Quiz # 2) – February 17th, 2025

Name: _____ Roll No: _____

Instructions

1- Read the Questions carefully to give concise Answers.
2- Answers outside the box will not be considered. Answers should be readable
3- Clearly mention any assumptions you have taken in your Answers.
4- Use of unfair means will result in an award of ZERO marks.

Q1. What are the 3Vs that constitute big data? Mention them & provide brief one-to-two-line explanation of each. {6 Marks}

(Answer in up to 2 lines for each category)

Volume: Terabytes to Petabytes and more
Variety: Structured and Unstructured data: eg, cvs, json, image, IoT …
Velocity: Accelerating rate of data ingestion and analysis

Q2. Mention 2 problems with Spark that Databricks solves. {4 Marks}

(Mention 2 Bullet Points)

- Databricks is a managed platform for running Apache Spark
- No cluster management
- No tedious maintenance tasks
- Point-and-click platform for developers that prefer a user interface
- Capabilities to automate aspects of data workloads with automated jobs
- Optimized autoscaling to resize a cluster intelligently

Q3. In Databricks, data is organized into 3 layers: bronze, silver, and gold. Briefly describe usage of each layer. {9 Marks}

(Answer in up to 3 lines per layer)

- Azure Databricks works well with a medallion architecture that organizes data into layers:
  - Bronze: Holds raw data.
  - Silver: Contains cleaned, filtered data.
  - Gold: Stores aggregated data that's useful for business analytics.

Q4.
a. Briefly describe the functionality of Spark Driver, Worker, and Executor. {9 Marks}
b. Which cloud resource is used to implement Spark Driver and Worker Nodes? {1 Marks}

(Answer should be concise)

- Spark applications run as independent sets of processes on a cluster, coordinated by the **SparkContext** object in your main program (called the driver program). 'Driver' runs the user's 'main' function and executes the various parallel operations on the worker nodes.
  - SparkContext sends tasks to the executors to run
  - Because the driver schedules tasks on the cluster, it should be run close to the worker nodes, preferably on the same local area network.
- Worker is the Node/Virtual Machine where your code runs.
- A process launched for an application on a worker node, that runs tasks and keeps data in memory or disk storage across them.
  - Each application gets its own executor processes, which stay up for the duration of the whole application and run tasks in multiple threads [on the nodes]