BIG DATA AND ANALYTICS ASSIGNMENT # 4

Due Data: 8 April, 2023 **MARKS 100**

Assignments are to be done in same group as previous. No late assignments will be accepted.

HONOR POLICY
This assignment is a learning opportunity that will be evaluated based on your ability to think, work through a problem in a logical manner. You may however discuss verbally or via email the assignment with your classmates or the course instructor, and use the Internet to do your research, but the written work should be your own. Plagiarized reports or code will get a zero. If in doubt, ask the course instructor.

Kafka Assignment: Building a Complex Data Pipeline

In this assignment, you will build a complex data pipeline using Kafka. You will set up a Kafka cluster with multiple topics, write multiple producers and consumers to generate and consume data, and integrate the pipeline with external systems.

- **Install Kafka** on your system and set up a Kafka cluster with at least three brokers. Configure the cluster to use a replication factor of 3 and a retention period of 30 days.
- Create at least three Kafka topics with different replication factors, retention periods, and partition counts. Choose topics that are relevant to your use case (such as "user-events", "click-streams", or “product-inventory").
- **Write producers:** Write two producers that generate data in a format (such as JSON) and produce data to different topics. The producers should generate data based on a realistic financial transactions use case you can use any Live Stocks Data. The producers should use the Kafka client library.
- **Write consumers:** Write three consumers that consume data from different topics and process the data in different ways. For example, you could write a consumer that aggregates company events by company ID, a consumer that calculates real-time statistics on stock inventory, and a consumer that filters out fraudulent financial transactions. The consumers should use the Kafka client library.
- **Integrate with external systems:** Integrate the pipeline with one external system, such as a database i.e MongoDB or SQL.
- **Test the pipeline:** Start the producers, consumers, and external systems and verify that the data is being generated, consumed, and processed correctly. You should be able to see the data being printed out by the consumers or stored in the external system.

Submission

- A **README** file that includes:
  Instructions for setting up and running the pipeline.
  A description of the data being generated, consumed, and processed.
- The source code for the producers and consumers.
- Any additional files needed to run the pipeline (such as configuration files or scripts).
- Submit a video **Onedirve, Google Drive, Youtube** Link Explaining your approach How you configured replication factor and retention period. Each and every step of implementation.

Notes

- You are free to use any programming language and Kafka client library you prefer.
- You may use any data generation tools or frameworks that you like, as long as they are compatible with Kafka and the chosen programming language.