**MID-TERM PROJECT**

**ETL WITH KAFKA AND HIVE**

**Name: MOHAMMED ADEEN SHAIK**

**PROG8451: Big Data integration and storage**

**Course Instructor: Saber Amini**

**STEP 1: GETTING API**

Getting API key / API token from the NewsApi to fetch values from their archives.

My API key: c3dcbb8a046e4261a9f70269aef1945f

**STEP 2: SETTING UP KAFKA BROKER**

Running Zookeeper and confluent to start Kafka.

Run this command in the zookeeper directory in your terminal:

bin/zkServer.sh start (make sure your broker ID and everything else is correct in config)

Run this command in the confluent directory:

bin/kafka-server-start etc/kafka/server.properties

**STEP 3: SETTING UP JUPYTER NOTEBOOK TO RUN KAFKA PRODUCER AND CONSUMER**

I am going with Jupyter Notebook to run my Python script for Kafka producer and consumer files.

I am fetching values of keyword= “apple “, to get news articles related to the apple company. Also, I am skipping all the articles that were removed or deleted.

**PRODUCER:**

A screenshot of a computer

Description automatically generated**CONSUMER:** **A screenshot of a computer

Description automatically generated**

**STEP 4: Setup a Kafka Consumer to read the data:**

Since I added this Jupyter notebook component to my cluster: etl-mid-term-project

I will process and send this data as a CSV file with name as ‘news.csv’ to my local directory:

Preprocessed data:

A white rectangular object with a black border

Description automatically generated with medium confidence

Processed data:

A screenshot of a computer

Description automatically generated

Next, Check-in your local directory where you saved the data as CSV file.:

A computer screen shot of a black screen

Description automatically generated

Now that a hive table has been created we will run a few queries to study the data and gain information.

A computer screen shot of a black screen

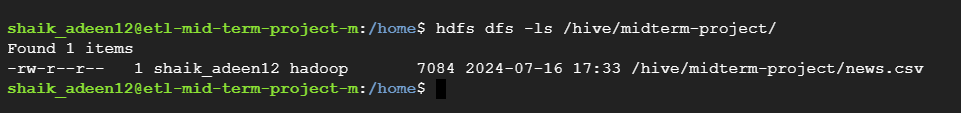
Description automatically generated

Now we upload this news.csv file to the HDFS directory: /hive/midterm-project/

A computer screen shot of a computer screen

Description automatically generated

Check if it is in the Hadoop directory:



**Step 5: Load the CSV to HIVE and run queries to gain insight on the data.**

Hive table ‘apple\_news’ using news.csv:

A screen shot of a computer screen

Description automatically generated

With this a hive table is made and now we can run our queries:

**QUERIES:**

1. Query to count the number of records / articles :

A screenshot of a computer

Description automatically generated

There are 16 articles or records/ rows

2. Query to List top 10 authors by the number of articles:

A screenshot of a computer

Description automatically generated

These are the authors with their articles count.

3. Query to Find the most frequent sources:

A black screen with white text

Description automatically generated

4. Query to List articles with the longest content:

A screenshot of a computer

Description automatically generated

This query shows the top 10 articles with the length of content.

**CONCLUSION:**

This ETL project utilized the NewsAPI to fetch news articles about Apple, streaming them through a Kafka broker. I efficiently transferred the data to a Hive table using a Kafka producer in a Jupyter Notebook within my Hadoop cluster on GCP. By querying the apple\_news table in Hive, I gained valuable insights into article trends, authors, sources, and content characteristics. This project demonstrates the effectiveness of Kafka and Hive for real-time data ingestion, storage, and analysis in cloud environments.

**THANK YOU**