# Midterm Project

**PROG8451: BIG DATA INTEGRTION & STORAGE**

Aakash Baskaran | 8937042

Big Data Solutions Architecture, Conestoga College

PROG8451 – Big Data Integration & Storage

Professor Ms. Shanti

Technical contribution in ETL report of news API data analysis

In our final ETL project, each of the three team members played a vital role in the technical execution. Below is a summary of the individual technical contributions made by each member:

Team Members Name (Student Id) & Contributed Topic:

1. Aakash (100935802) – Setting up Kafka broker and Creating Producer Script.

2. Krishnapriya (100910070) – Creating Consumer Script and hive table.

3. Annadurai (100934897) – Running sample hive queries.

Setting up Kafka broker and Creating Producer Script:

I began the initial phase of our ELT (Extract, Load, Transform) project by establishing Apache Zookeeper and Apache Kafka. Apache Zookeeper functions as a distributed coordination service essential for managing Kafka's infrastructure. Concurrently, Apache Kafka operates as a distributed streaming platform, facilitating the handling of real-time data feeds. Subsequently, I crafted a script for the producer function, utilizing Python. This producer script is crucial as it is responsible for generating a news API token. Utilizing this script, I successfully generated the token, which specifically pertains to an article about the rising popularity of football.

Creating Consumer Script and hive table:

Following the successful generation of the token, I proceeded to develop a script for the consumer function, again using Python. This consumer script was instrumental in generating a CSV file, which I initially stored locally on my system. Next, I transferred this CSV file from my local storage to Hadoop, setting the stage for running Hive queries. Once the file was securely in the Hadoop environment, I took the next step of creating a database within Hive. Within this database, I established a table named 'articles\_table'.

Running sample hive queries:

After setting up a new table in Hive, I ran three different queries. The first query identified the authors who wrote the most articles about football. The second query found the top five sources with the most football articles. For the third query, to discover on which dates the most articles were published, I created a new table called `formatted\_articles\_table`. This was necessary because the date in the original table wasn't easy to read or use for analysis. I used the `UNIX\_TIMESTAMP` function to change the date format from a string to a more user-friendly format. After incorporating this new date format into the new table, I ran the third query and successfully identified the dates with the most articles published.