**Realtime Data Processing Project**

# Tools used:

* **Storage**: Amazon S3
* **Realtime Streaming Tools**: Amazon EC2, Apache Nifi, Zookeeper, Snowpipe
* **Data Warehouse**: Snowflake

# Workflow:

## Data Generation and Collection:

Amazon EC2 instance is created to run Jupyter Notebook, Apache Nifi and Zookeeper. The code for fake data generation resides in the EC2 instance which runs for the interval of 5 minutes. A pipeline is created in Apache Nifi which picks the csv files from the EC2 local folder and puts the file in the S3 bucket.

## Data Loading:

Snowpipe is created and is connected to the S3 bucket via SQS notification. As new files arrive in S3 bucket, snowpipe runs and loads the data in the CUSTOMER\_RAW table.

## Data Transformation:

A stored procedure is created within snowflake which runs the MERGE command which checks the presence of any existing record. If record exists it updates the value to the new record else it inserts new record.

