Lab: Big Data on the Cloud – Analyzing NYC Taxi Data from S3 using EC2

Objective

This lab demonstrates the power of cloud computing by processing a large public dataset directly from AWS S3 using an EC2 instance. Students will:

- Launch an EC2 instance with proper IAM permissions
- Access the NYC Taxi public dataset hosted on Amazon S3
- Read and analyze the data using Python (pandas & Boto3)

Section 1: Launching the EC2 Instance

1. Go to **Amazon EC2 > Instances > Launch Instance**
2. Name your instance: `nyc-taxi-lab`
3. Choose an AMI: **Amazon Linux 2**
4. Choose an instance type:
    - For demo: `t3.micro` (free tier)
    - For performance comparison: `t3.large` (if available in sandbox)
5. Under **Key pair**, choose `vockey`
6. Under **Network Settings**:
    - Allow **SSH** from your IP
7. Under **Advanced Settings**:
    - IAM Role: choose `LabRole`
8. Click **Launch Instance**

Section 2: Connect to EC2 and Set Up Environment

Connect using **Session Manager** or SSH.

Update and install dependencies:

```
bash
CopyEdit
sudo yum update -y
sudo yum install -y python3 pip
pip3 install boto3 pandas
```

Section 3: Python Script to Load Data from Public S3 Bucket

Use this script to load and analyze the NYC Yellow Taxi data:

```
import boto3
import pandas as pd
from io import BytesIO

# S3 client
s3 = boto3.client('s3')

# Download CSV file from public dataset
bucket = 'nyc-tlc'
key = 'trip\_data\_yellow\_csv/yellow_tripdata_2019-01.csv'  # Choose a single
month to avoid memory issues

response = s3.get_object(Bucket=bucket, Key=key)
df = pd.read_csv(BytesIO(response['Body'].read()))

# Display sample and analysis
print("Sample rows:")
print(df.head())

print("\nTrip Distance Statistics:")
print(df['trip_distance'].describe())

print("\nAverage trip distance by hour:")
df['tpep_pickup_datetime'] = pd.to_datetime(df['tpep_pickup_datetime'])
df['hour'] = df['tpep_pickup_datetime'].dt.hour
print(df.groupby('hour')['trip_distance'].mean())
```

Section 4: Teaching Points

- **Dataset size**: Show how large datasets can be accessed directly from S3 without downloading.
- **Power of EC2**: Compare performance on a micro vs. larger instance (if allowed).
- **IAM Roles**: Explain how the `LabRole` allows EC2 to access S3 securely without managing credentials.
- **Cost optimization**: Show how to spin up resources only when needed and shut them down after.

Cleanup

After the lab:

1. Stop or terminate the EC2 instance.
2. Revoke any additional permissions if created.
3. Remind students that sandbox resources will auto-delete when time expires.