Project Proposal
Group 2 : Cloud Comp and Big Data

ETL and Analysis of NYC Yellow Taxi Trip Data using Spark and AWS

Aditya Shah - as5069
Kris Patel - ksp177

Project Overview:

For our final project, we plan to work with the NYC Yellow Taxi Trip Dataset, which contains millions of taxi trip records collected across New York City. The goal is to develop an end-to-end ETL (Extract, Transform, Load) and data analysis pipeline utilizing Apache Spark and AWS Cloud.

The goal is to take a large real-world dataset that can't easily fit on one machine, clean and process it in a distributed environment, and then analyze it to find useful insights such as patterns in trip duration, fare amounts, and demand at different times of the day or in different parts of the city.

This project focuses on the data engineering side of big data handling, large data, cleaning it, running transformations efficiently, and generating analytics on the cloud.

Motivation:

We chose this dataset because it's both large and practical. The NYC Taxi data is well-known for being huge (hundreds of millions of rows), making it a great example for practicing real-world big data tools.

In most data science classes, we usually work with smaller datasets that run fine on a laptop. In this project, we aim to take a step further and utilize distributed computing tools, such as Spark, and cloud storage services like AWS S3 to explore how large-scale data processing operates in the real world.

Approach:

We plan to structure the project around the main ETL steps:

Extract:

Scrape the data from official NYC Open Data site.
Store it on AWS S3 so it can be accessed by Spark in the cloud.

Transform:

Use PySpark on an AWS EMR cluster (or Databricks) to clean and process the data.
Fix invalid or missing values (like negative fares or coordinates).
Create new columns like trip duration, speed, and fare per mile.
Aggregate data by hour, day, and location for analysis.

Load and Analyze:

Save the cleaned dataset back into S3 in Parquet format for efficiency.

Use Spark SQL or the Pandas API on Spark to run queries such as:

Average fare per mile by hour of day.
Total number of trips by pickup zone and day of week.
Busiest pickup and drop-off locations.

Finally, visualize results using Matplotlib, Seaborn, or Tableau/AWS QuickSight.

Tools and Technologies:

Apache Spark
AWS for cloud storage(S3) and computation
PySpark
Matplotlib/Seaborn for data visualization
GitHub for version control and collaboration

Tentative Timeline:

Week 1:	Explore dataset, set up AWS environment and Spark cluster
Week 2:	Data cleaning and preprocessing using PySpark
Week 3:	Perform transformations and create derived features
Week 4:	Run Spark SQL queries and generate analytics
Week 5:	Create visualizations and summarize key insights
Week 6:	Final documentation, testing, and presentation preparation

We believe this project is a good balance between challenge and practicality. It enables us to apply big data and cloud computing concepts to a real dataset without any overly complex machine learning or streaming systems. We'll get experience with data engineering and cloud technologies, which are essential skills for data professionals today.