**DATA/MSML 650**

**Final Project: First Progress**

**Title:** E-Commerce Predictive Analysis using AWS

**Team Members:** Tausif Khan, Varun Jain, Hsi-Chun Wang, Sidhant Chanana, Allen Mathews

**Problem Statement**

Using past sales and consumer data, we plan to predict and forecast the future consumer trends by leveraging the use of predictive machine learning regression models. To deploy these machine learning models effectively we plan on using AWS functionalities like AWS Kinesis and AWS Lambda to set up the data pipeline for our model.

**Members Planned Contribution**

1. **Tausif Khan:** Setup AWS environment by configuring IAM roles, policies, and VPC settings to make sure access to resources is secure to ensure a secure infrastructure for this project. This includes defining permissions for users, security policies, and configuring the network to control how resources interact within AWS.
2. **Varun Jain:** Manage the ETL pipeline to ensure data is collected, cleaned, and organized efficiently, making stored data easily accessible and ready for analysis. This involves setting up processes to handle data from various sources, transforming it into a usable format, and loading it into a central data storage for easy access.
3. **Hsi-Chun Wang:** Develop backend logic for serverless computing using Lambda functions integrated with Kinesis and DynamoDB to allow efficient data flow and storage without the need for servers.
4. **Sidhant Chanana:** Implement real time analytics to monitor and predict customer behavior by managing Kinesis data analytics applications, which enable the team to process data in real time.
5. **Allen Mathews:**  Create visual representations of data to show business decisions by building interactive dashboards with Amazon QuickSight, which can be used to interpret trends and insights, to make better business decisions.

**High-Level Approach**

1. Setting up and configuring the architecture including IAM roles/policies, VPC configurations and manage access control
2. Setting up the ETL pipeline into the data lake for analysis
3. Implementing backend logic for serverless computation and data interaction between services
4. Creating a Kinesis data analysis application
5. Creating Lambda functions
6. Integration of Lambda Functions with Kinesis
7. Analyzing and visualizing customer trends
8. Creating QuickSight Dashboards

**Implementation tools**

1. Data storage: DynamoDB, AWS Aurora
2. Computing: AWS Lambda, AWS Kinesis
3. Data Visualization: Amazon QuickSight

**Additional progress notes:**

1. Explored and found a few consumer datasets including fields such as destination, product, pricing etc. which makes the data rich in features and perfect to use for regression models.
2. Explored an ensemble approach to fit the best regression model and using comparative analysis to fit the optimal model.
3. For exploratory data analysis, to understand the correlation between various consumer features and target variables, we are using visualization for heatmaps.
4. The list of tools for implementation is not exhaustive and we might be integrating more tools into our project as per the requirement.
5. Researched IAM setup, multi-factor authentication, security groups, network access control lists, and VPC configurations and tested establishing S3 buckets while modifying access policies