**Big Data Project Proposal**

Supervised by: **Dr. Lydia Wahid**

Presented by:

| **BN** | **Sec** | **Name** |
| --- | --- | --- |
| 5 | 1 | Ahmed Khaled |
| 6 | 1 | Ahmed Sayed |
| 12 | 1 | Ahmed Mahmoud Mohamed |
| 25 | 1 | Hazem Abdo |

March **2023**

**Project Idea:**

The goal of this project is to use clustering algorithms to segment customers of an online retailer based on their purchasing behavior. The project will involve preprocessing and cleaning the Online Retail dataset and selecting relevant features, such as product ID, quantity, and price. The dataset will be scaled and outliers will be removed to ensure that the clustering algorithm provides accurate results.

**Work Plan:**

1. Data Preprocessing: Download the [dataset](https://archive.ics.uci.edu/ml/datasets/Online+Retail#) and preprocess it by removing missing values and duplicates, and transforming the data into a suitable format for clustering algorithms.
2. Feature Selection: Select the relevant features that are most important for clustering customers based on their purchasing behavior. In this dataset, the relevant features could be the customer ID, product ID, quantity, and price.
3. Feature Engineering: Create new features such as total amount spent, average purchase value, frequency of purchases, and recency of purchases.  
   Choose the clustering algorithm: Choose the clustering algorithm that best suits your data and business problem. For example, K-means clustering is a popular algorithm for clustering customer data based on their purchasing behavior.
4. Determine the optimal number of clusters: Experiment with different numbers of clusters to find the optimal solution in this data set. You can use techniques such as elbow method or silhouette score to determine the optimal number of clusters.
5. Apply the clustering algorithm: Apply the chosen clustering algorithm with the optimal number of clusters determined in the previous step. For example, you can use the Scikit-learn library in Python to apply K-means clustering.
6. Analyze the results of the clustering algorithm to identify the characteristics of each customer segment. This could involve visualizing the clusters using dimensionality reduction techniques, and interpreting the distribution of features across the clusters.
7. Evaluate the results: Evaluate the results by comparing them with the business problem and determining if the clustering provides valuable insights for the business.

In this project, MapReduce can be used for tasks such as filtering out missing values and duplicates, transforming the data into a suitable format, and scaling the data and in the feature engineering stage.