# Big Data Analytics Project

Weightage: 12 %  Due Date:10th May 2023

Project is to be done in the groups assigned. No late Project will be accepted.

## HONOR POLICY

This Project is a learning opportunity that will be evaluated based on your ability to think, work through a problem in a logical manner. You may however discuss verbally or via email the assignment with your classmates or the course instructor, and use the Internet to do your research, but the written work should be your own. Plagiarized reports or code will get a zero. If indoubt, ask the course instructor. You should submit your assignment in a zip file following the naming convention.

**Naming Convention**: i21xxxx_i21xxxx_i21xxxx.zip

## Objectives:

In this project, you will be building a live product recommender system based on customer preferences using the Amazon product dataset. The data set can be downloaded from here. (32GB compressed). The project is divided into two phases.

## Phase 1: Data Processing and Analysis (Marks: 100)

- Load the Amazon product dataset into MongoDB using Apache Spark: In this phase, you will use Apache Spark to efficiently load the Amazon product dataset into MongoDB. You will need to consider the schema of the dataset and the structure of the MongoDB collections to optimize the data loading process. You will also need to handle any missing or malformed data as part of the data loading process. (50 Marks)
- Perform exploratory data analysis (EDA) on the dataset to gain insights into the data: Once the data is loaded into MongoDB, you will perform EDA to gain insights into the data. You will use tools such as Pandas and Matplotlib to visualize the data and identify patterns and trends in the data. You will also use statistical analysis to identify correlations and relationships in the data. Every one can have different analysis based on their perception of Data. (40 Marks)
- Preprocess and clean the data as necessary for training the model: After performing EDA, you will preprocess and clean the data as necessary for training the recommendation model. This may involve handling missing values, transforming categorical variables, normalizing the data, and feature engineering. (10 Marks)

## Phase 2: Model Training and Live Streaming (Marks: 150)

- Train a recommendation model using machine learning algorithms on the preprocessed data: In this phase, you will train a recommendation model using machine learning algorithms on the preprocessed data. You will use techniques such as collaborative filtering, content-based filtering, and matrix factorization to build the model. You will evaluate the model using metrics such as precision, recall, and F1 score. (50 Marks)
- Set up a Flask-based web application: you will set up a Flask-based web application. You will use Flask to build the web application and MongoDB to store the user preferences and recommendations. You will use REST APIs to communicate between the web application and the recommendation system. (50 Marks)
- Use Apache Kafka to stream real-time recommendations based on user preferences: Once the model is trained, you will use Apache Kafka to stream real-time recommendations based on user preferences to the webpage you created using Flask. (50 Marks)

Overall, this project will give you hands-on experience with big data processing, NoSQL databases, machine learning, web development, and stream processing. You will also develop skills in problem-solving, critical thinking, and communication.

# Submission

One of the group members is required to submit a report with supporting screenshots of EDA and everything with time stamps. Each member should also explain their findings in the report. Your Flask Website should be aesthetically pleasing. Make sure you create a virtual environment for this project and having an executable bash script is a plus point. Your Zip should have all the code files except for the dataset. Please note that you can get called for Demos.

**NOTE: Your Report must be PDF and you must define which group member did which work. **Your Zip file should not contain Dataset file as it will lead to zero marks in the project**.**