# Amazon Reviews Classifier

October 3, 2022

## Group Members


| | |
|---|---|
| Ezra John Guia | 30031697 |
| Ivan Suyat | 30089089 |
| Long Ta | 30069130 |
| Suhaib Tariq | 30075751 |


## Key Project Objective

Develop a sentiment analyzer using Amazon review to predict whether a review is positive, neutral or negative. The source can be found here.

## Tools of Platforms Needed

- Coding:
  - GitHub
  - Git
  - Python
  - PySpark (Spark interface using Python API)
- Cloud
  - DataBricks
- Communication
  - Jira
  - Discord

## Details of the Project Management Plan

- GitHub and Git
  - Branches for various features
  - Repository for codebase
  - Version control is maintained using Git

- Use Jira to assign, manage, and track tasks
  - Create a Kanban board to track progress

Hadoops, Spark, …?

Possible data set:

- House price prediction using the following dataset House-Price-Prediction Cleaned Dataset | Kaggle
- Best Selling Albums By Duration (1990-2021) | Kaggle

Using Pyspark to analyze data, Pyspark provides API to run job in Spark PySpark Documentation — PySpark 3.3.0 documentation (apache.org)

AWS:

- Using AWS EMR
- There are 3 options with on how to use AWS EMR
- EC2 (Basically a VM)
- Serverless (pay for what you use only, ideally we should use this) (Charged per minute, cost 0.05264 per vCPU per hour and 0.0057785 per GB per hour) ( If you use ephemeral storage, the cost is 0.000111 per GB per hour)
- Can use Terraform, AWS CDK, CloudFormation to provision resources (Iac)

GCP:

- Using Dataproc
- Tutorial Running Spark Jobs in a Serverless Environment | Spark Day on Google Open Source Live - YouTube
- Don't think it's possible to use Iac
- Pricing: Pricing | Dataproc Serverless | Google Cloud

**Go with GCP because:**

- 300$ credit
- Cloud agnostic, we can switch over to AWS anytime
- We manage our own account + billing. If there is any shared data, it can be shared across all account