For this project I was tasked with provisioning a spark cluster using amazon web services. Using AWS, I used EMR to load and run some analysis on a IMDB’s dataset. Which you can find on [**Kaggle**](https://www.kaggle.com/datasets/ashirwadsangwan/imdb-dataset). For my analysis, I used Juptyer notebook to run my analysis . I also used a public S3 bucket to get this data for our analysis.

Techolonogies :

**AWS**:

* S3
* EC2
* Apache Spark

**Python Libraries :**

* Pandas
* Matplotlib