****

**CA687I - Cloud Systems**

**Midway Report Submission**

**24th February 2021**

**Lecturer: Alessandra Mileo**

|  |  |  |
| --- | --- | --- |
| **Student ID** | **Name** | **Email** |
| 20214350 | Colin Hanily | colin.hanily2@mail.dcu.ie |
|  | Neal Knauf |  |
|  | Kay-Ti Tan |  |
|  | Rana Zeeshan Ali Shahid |  |
|  | Darren Rooney |  |

**Assignment 1**

1. **What dataset are you going to work on?**

[Spotify Music Dataset](https://www.kaggle.com/yamaerenay/spotify-dataset-19212020-160k-tracks), with 175k+ songs dating 1921 and 2021. The dataset contains 19 columns with metadata such as year of release, tempo, genre, key and popularity offering a wide a variety of analytical approaches

1. **What technology are you using for the analysis (and why)**

**Apache Hadoop** **Framework -** Used to store our data on a google cloud platform instance. Hadoop is a highly compatible big data framework that supports multiple languages which will be useful for our data cleansing, pruning queries and our machine learning model.

**Python –** Used to create the song popularity predictor machine learning model as it has straightforward syntax and well documented machine learning libraries e.g.sklearn.

**Hadoop Streaming** - Used to run our machine learning script and will be connected to our data visualisation tool to take user input.

**Apache Pig -** Runs on Hadoop making it simple to integrate, will be used to cleanse and prune our dataset for our data queries

**Apache Hive -** Runs on Hadoop and will be used to run our dataset queries to find data correlation and prepare our findings for visualisation

**Google Cloud Platform -** Used so that our data can be stored and queried on the cloud. Integration with Hadoop is well documented making this a suitable choice

**QlikView** or **Tableau -** Data visualisation. Further analysis of how would like to display our findings is needed before choosing a suitable.

**Google Drive** – Team Report and task breakdown sharing.

**Movavi** – Demo video recording.

**Slack/Whatsapp/Zoom** – Team Communication

**Github** – Storing project code and documentation

1. **What analytics (e.g. what you want to gain from the analysis on the data)**
2. To see what makes a song popular, and how the characteristics of popular/unpopular songs have changed over the last century.
3. Analyse the dataset on a yearly level, giving a clear overview of the most popular songs year by year, and the correlating characteristics of these songs with respect to popularity.
4. Analyse the dataset on a macro level to see what the characteristics of a song that has lasting popularity are. The datasets “popularity” field gives the current popularity of a song, we can use to analyse tracks from as far back as 1921 to see what, if any, correlation between a songs characteristics and its current popularity exists.
5. Build a song popularity predictor where a user can input the characteristics of a track and predict how popular the song would be today. This idea was inspired by an [article](https://www.theverge.com/2019/5/28/18642978/music-streaming-spotify-song-length-distribution-production-switched-on-pop-vergecast-interview#:~:text=Songs%20are%20getting%20dramatically%20shorter%20over%20time.&text=Songs%20have%20decreased%20in%20length,over%20the%20past%2018%20years.) that states the average length of a song has gradually decreased over the last decade as artists on Spotify are paid per song stream. Thus shorter songs are more likely to be played more often, generating more revenue and perceived popularity.
6. **The plan for the team roles and tasks**

|  |  |  |  |
| --- | --- | --- | --- |
| **Task** | **Main Contributor** | **Status** | **Notes** |
| Choose team lead | Team | Completed | Colin was chosen |
| Choose Dataset | Team | Completed | Spotify Music |
| Choose analysis to undertake | Team | Completed |  |
| Choose technologies to use | Team | Completed |  |
| Data Trends Exploration | Darren/KT | To do |  |
| Cleanse/Prune data | KT | To do |  |
| Create HDFS | Neal | To do |  |
| Upload HDFS to GCP | Neal | To do |  |
| Create relevant queries for data analysis | Darren | To do |  |
| Create machine learning model for popularity of user inputted song | Colin | To do |  |
| Connect HDFS to data visualisation tool | Neal | To do |  |
| Connect Machine Learning model to data visualisation tool | Colin | To do |  |
| Data visualisation | Rana | To do |  |
| Demo video | Rana | To do |  |
| Final Report | Team | To do |  |