CA687I - Cloud Systems

Midway Report Submission

23rd February 2021

Lecturer: Alessandra Mileo

Student ID	Name	Email
20214350	Colin Hanily	colin.hanily2@mail.dcu.ie
21210014	Neal Knauf	neal.knauf2@mail.dcu.ie
20215470	Kay-Ti Tan	kayti.tan3@mail.dcu.ie
20216275	Rana Zeeshan Ali Shahid	rana.shahid2@mail.dcu.ie
20216268	Darren Rooney	darren.rooney37@mail.dcu.ie

Assignment 1

1. **What dataset are you going to work on?**
   Spotify Music Dataset, with 175k+ songs dating from 1921 to 2021. The well-structured dataset contains 19 columns with metadata such as year of release, tempo, genre, key and popularity offering a wide variety of analytical approaches.

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

   Apache Hadoop Framework - Used to analyse 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. Scikit-learn.

   Hadoop Streaming - Used to connect our machine learning model 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 correlations 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 we would like to display our findings is needed before choosing a suitable data visualisation technology.

   Google Drive – Team Reports and task breakdown sharing.

   Movavi – Demo video recording.

   Slack/Whatsapp/Zoom – Team Communication.

   Github – Storing project code and documentation.

3. **What analytics (e.g. what you want to gain from the analysis on the data)?**
   a) To see what makes a song popular, and how the characteristics of popular/unpopular songs have changed over the last century.

   b) Analyse the dataset on a yearly level, giving a clear overview of the most popular songs by year/decade, and the correlating characteristics of these songs with respect to popularity.

   c) Analyse the dataset on a macro level to see what the characteristics of a song that has current popularity are. The datasets "popularity" field gives the current popularity of a song, which 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.

d) 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 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.

4. 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	Team	To do	
Create HDFS	Neal	To do	
Upload HDFS to GCP	Neal	To do	
Cleanse/prune data	KT	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/Rana	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	