Technical Design Document: Analyzing Song Trends on Spotify to Improve the Popularity of Singers

Introduction:

The goal of this technical design document is to provide an overview of the technical infrastructure and tools used to analyze the "Spotify Charts" dataset and identify song trends to improve the popularity of singers in the music industry. This project aims to use exploratory data analysis, clustering, and regression techniques to analyze the dataset and provide insights into user preferences and trends.

Coding Language and Tools:

The project will use **Python** as the primary programming language for data cleaning, preprocessing, analysis, and visualization. We will use **PySpark** to manage and preprocess the large dataset and Pandas to manipulate and analyze the data. We will use **Matplotlib** and **Seaborn** libraries to create visualizations.

Infrastructure and System:

The project will use cloud computing resources provided by **Google Compute Engine and Google Colab**, which allow for collaborative coding and analysis in a cloud-based environment. We will use **Amazon S3** to store the cleaned and preprocessed dataset and directly store the visualizations generated from the analyses to the **S3 bucket** to facilitate easy access and sharing with stakeholders who may want to create websites or dashboards in the future.

Cloud Services:

We will be using **Google Cloud Platform (GCP)** to perform the analyses. Specifically, we will utilize **Google Compute Engine**, which provides cloud-based computing resources for virtual machines, in the form of **Google Colab**, which provides a cloud-based Jupyter Notebook environment. We will store the cleaned and preprocessed dataset on **Amazon S3**, which provides scalable and cost-effective storage.

Data Flow:

The "Spotify Charts" dataset will be collected using the Spotify API and stored in a CSV file format. The data will be preprocessed and cleaned using PySpark and Pandas, respectively. We will then perform exploratory data analysis and generate visualizations to identify patterns and trends in the data. We will also perform clustering and regression analyses to identify the factors that contribute to a song's success. Finally, the results will be stored in the S3 bucket for easy access and sharing.

Data Size: 3.48 GB

The "Spotify Charts" dataset contains 26,173,514 observations and 9 columns, covering countries around the world. The size of the dataset is relatively large, which requires the use of cloud-based computing resources and tools such as PySpark to manage and preprocess the data.

Cost Estimate:

For the Amazon S3 storage, the cost will depend on the amount of storage and data transfer used. For 10-15GB of storage, the cost would be approximately $1.50 - $2.25 per month. Data transfer costs depend on the amount of data accessed and transferred from the S3 bucket. Assuming moderate usage, the data transfer cost would be around $0.05 - $0.10 per month.

If one wants to dive in deeper into Machine Learning and AI, the cost of a standalone compute engine will depend on the type and number of virtual machines (VMs) used, as well as the duration of use. Assuming you need 100GB of compute engine with 16/32 RAM, you could use a custom VM with 16 or 32 vCPUs, 100GB of memory, and 1TB of SSD storage. The estimated cost for this VM would be approximately $800 - $1600 per month, depending on the selected region and the duration of use.

Additionally, there may be additional costs associated with using PySpark and Google Colab, such as increased memory and storage requirements. For instance, Google Colab Pro: $9.99/month for 16 GB RAM or $49.99/month for 32 GB RAM. However, these costs are likely to be minimal compared to the compute engine and S3 storage costs.

Overall, the estimated monthly cost of the infrastructure for this project would be around **$801.50 - $1602.25**, depending on the VM configuration and data transfer requirements. It's worth noting that this is just an estimate, and the actual cost may vary depending on factors such as usage patterns, data size, and resource requirements. It's important to regularly monitor the costs and adjust the infrastructure as necessary to ensure that it remains cost-effective.

Conclusion:

In summary, this technical design document provides an overview of the infrastructure and tools used to analyze the "Spotify Charts" dataset and identify song trends to improve the popularity of singers. The project will use Python as the primary programming language and Google Cloud Platform to perform the analyses in a cloud-based environment. The data flow will be managed using PySpark and Pandas, and the results will be stored in Amazon S3. The size of the dataset requires the use of cloud-based computing resources, and the cost estimate will depend on the usage of these resources.