**Course Project:Big Data Concepts**

**Author: Preetham Reddy Vinnamala**

**Email:** [**savinn@iu.edu**](mailto:savinn@iu.edu)

**1. Introduction**

In the realm of data analytics, the ability to decipher trends and patterns in the global music industry presents both a challenging and exhilarating opportunity. My project embarked on this journey, focusing on a dataset that encapsulates the top trending songs across over 73 countries. This dataset, comprising 127306 entries, reveals not just the popularity of songs but a myriad of dimensions that define what makes music resonate with audiences worldwide. This dataset is a treasure trove for any data enthusiast, from daily and weekly movements in song rankings to musical attributes like danceability, energy, and tempo.

My objective was to harness advanced data analytics tools and techniques to unravel the intricacies hidden within these musical trends. This exploration was more than a mere academic pursuit; it was a quest to understand the pulse of global music preferences and how various factors shape them. Through this project, I aimed to enhance my data-handling skills and contribute meaningfully to the broader discourse on music trends and their implications in the digital age.

**2.Background**

The motivation behind delving into the world of trending songs stemmed from my fascination with how music preferences vary and evolve across different societies. Music, in its universal language, touches every corner of the globe, and analyzing its trends offers insights into cultural dynamics and shifts in global tastes. This dataset’s profound impact on understanding global music trends drew me in, urging me to explore the intricate relationship between musical attributes and their popularity.

The project was not just about numbers and charts; it was a journey to appreciate what makes a song appealing to diverse audiences. It provided an avenue to contribute to understanding global music trends, a topic that often intertwines with cultural, social, and economic factors. By utilizing tools like Google Cloud Platform, Google Collab, Big Query, and Looker Studio, I was equipped to dissect this extensive dataset, uncovering patterns and trends that narrate the story of contemporary music preferences.

### 3.Methodology

#### **3.1.Plan**

My journey into analyzing the world of music began with a clear vision – to uncover patterns and insights within the top trending songs across various countries. The plan was to navigate a comprehensive dataset, apply advanced data analytics techniques, and visually represent the findings. This project wasn't just about data; it was about the stories behind the numbers, the rhythms that resonate globally, and the melodies that define cultures.

#### **3.2. Obtain**

The initial step in my analytical journey was the careful selection of a dataset that not only aligned with my project's objectives but offered a rich tapestry of data to explore. After a thorough search on Kaggle, I discovered the "Top Spotify Songs" dataset, an expansive collection that captures the essence of the global music landscape.

This extensive dataset consists of 127,306 entries, each representing a unique song trending across over 73 countries. It is not just the volume of data that makes this dataset remarkable, but its depth and diversity. The dataset encompasses various musical attributes, providing a comprehensive view of what constitutes a top song in today's global music scene. Key attributes of the dataset include:

* **Basic Information:**  Each entry contains the Spotify ID, name, and artists, offering a clear identification of the song and its creators.
* **Ranking and Movement:** Crucial metrics like daily\_rank, daily\_movement, and weekly\_movement shed light on the song's popularity dynamics, indicating how each track fares daily and weekly.
* **Geographic and Temporal Data:** The country and snapshot\_date fields provide geographical and temporal contexts, allowing for an analysis of trends over time and across different regions.
* **Popularity Metrics:** The popularity score, along with indicators like is\_explicit, provides insights into the audience's reception and the content nature of the tracks.
* **Musical Characteristics:** A suite of attributes, including danceability, energy, key, loudness, mode, speechiness, acousticness, instrumentalness, liveness, valence, and tempo, offers a detailed analysis of the songs' musical qualities.
* **Duration and Structure:** The duration\_ms and time\_signature give an understanding of the songs' length and rhythmic structure.
* **Album Information:** The album\_name and album\_release\_date provide additional context about the album from which each song originates.

By choosing this dataset, I aimed to delve into the multifaceted world of music, exploring how various elements converge to define what resonates with listeners globally. The dataset's comprehensive nature promised a wealth of insights, setting the stage for a deep dive into the analytics of music popularity.

#### **3.3. Assure - Infrastructure Setup**

With the dataset in hand, I moved to the Google Cloud Platform (GCP), creating a bucket named bucket\_spotify to ensure a centralized and secure storage location. This step was crucial for maintaining the integrity and accessibility of the data throughout the analytical process.

A screenshot of a computer

Description automatically generated

#### **3.4. Transform – Infrastructure Step**

After making sure that the data is loaded to the bucket. I accessed the Google Colab with the same user credentials and accessed the dataset directly from the GCP bucket, leveraging the cloud's computational power for data processing. The dataset underwent a series of preprocessing steps, cleaning and structuring the data for more effective analysis.

* **Code for establishing connection from Google Colab to the GCP bucket:**  
    
  

A screen shot of a computer program

Description automatically generated  
  
I began by installing the necessary Google Cloud libraries in the Colab environment. This setup was imperative to facilitate the interaction between Colab and the GCP's extensive functionalities.

Upon completing the installations, the next step was authentication. This critical phase ensured my Colab session had secured access and the appropriate permissions to interact with GCP resources. The project was then configured to connect with the specific GCP project, in this case, targeting the 'spotify\_songs' project. I defined the storage client and specified the bucket, aptly named 'bucket\_spotify'. This was followed by pinpointing the precise dataset file, 'universal\_top\_spotify\_songs.csv', to be utilized. The final leg of this setup involved creating bucket and blob objects in Google Cloud Storage, facilitating the direct retrieval of the dataset. The file was then downloaded and parsed into a pandas DataFrame, ready for the subsequent data processing stages.

* **Preprocessing Steps**  
    
  A screen shot of a computer program

  Description automatically generated  
    
    
  Given the dataset's regular updates and its structured composition, the preprocessing phase required targeted interventions:  
  .
  + **Addressing Missing Values:** The dataset predominantly had missing values in the 'country' column, totaling 1,708 entries. I replaced these missing values with 'Unknown' to maintain data integrity and continuity. For the remaining minimal null values, roughly around 25 rows, I opted for removal, considering their insignificant proportion in the dataset.
  + **Normalization of the 'Mode' Attribute:** I observed notable discrepancies in the 'mode' attribute. To rectify this, I employed z-score calculations to identify and filter out anomalies, retaining only those values with z-scores below 3. This approach effectively normalized the data, mitigating outlier impacts.
  + **Correlation Analysis:** A crucial step in the preprocessing stage was the construction of a correlation heatmap. This analysis provided insights into the interrelationships among various musical attributes, aiding in identifying key factors for subsequent visualizations.
  + **Popularity Metric Analysis:** A focal point of my analysis was the 'popularity' metric. I computed this attribute's Interquartile Range (IQR) to discern and exclude extreme values. This step was essential to concentrate the analysis on tracks that genuinely captured listener interest, reflecting authentic music trends.
  + **Visual Exploration of Data Distributions:** I generated box plots and histograms for all numerical attributes to conclude the preprocessing. These visual representations were instrumental in unveiling the distribution patterns and variances within the dataset's principal numerical features, thereby setting a solid foundation for the following in-depth analytical exploration.

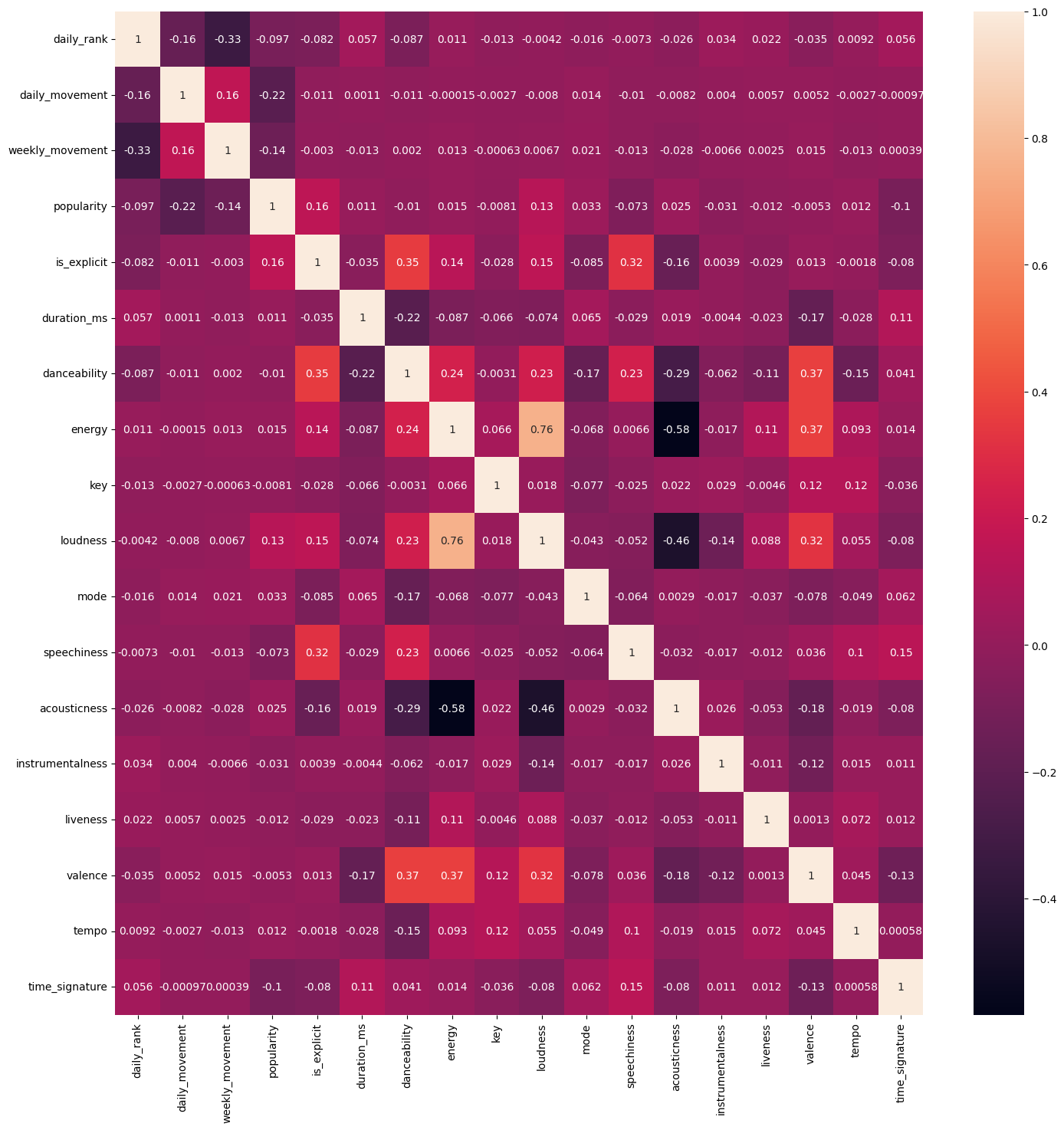


Fig 1

A graph with a blue rectangle and black lines

Description automatically generated  
Fig 2  
  
A graph with a blue rectangular object

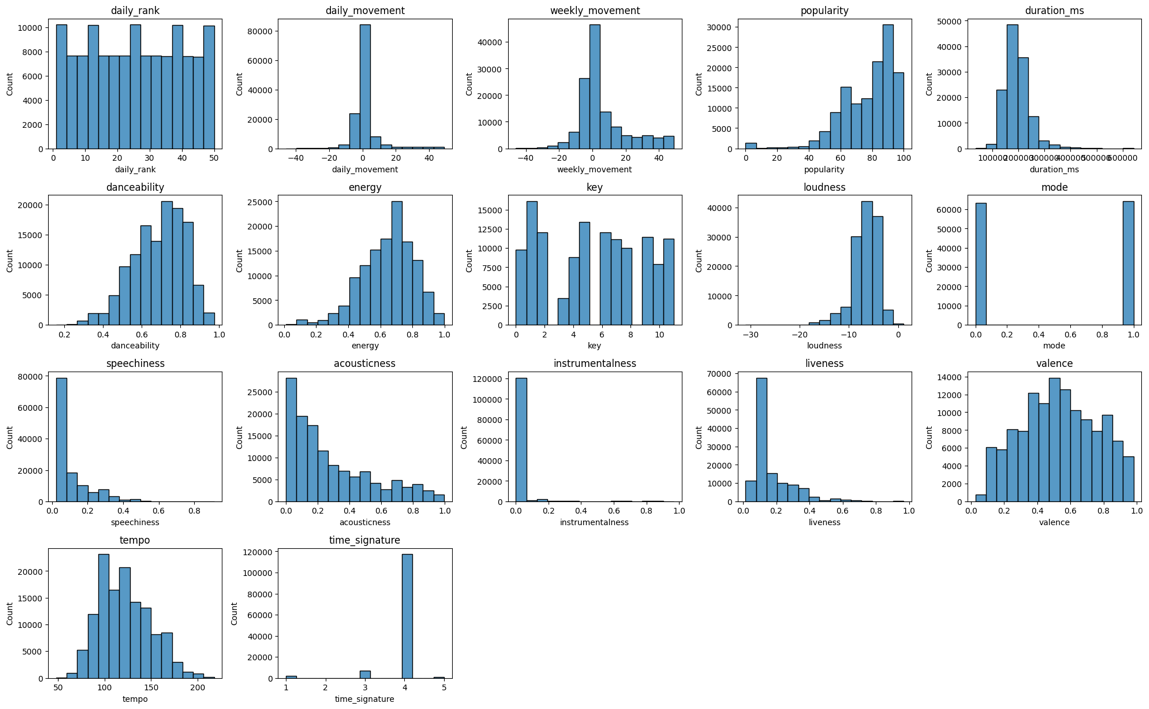
Description automatically generated with medium confidence  
Fig 3  
  


Fig 4

The transformation process revealed new patterns and insights, refining the raw data into a format ripe for deeper examination. The resulting file, named preprocessed\_spotify.csv, was a cleaner, more structured dataset ready for the next stages of analysis.

#### **3.5. Store**

Post-transformation, the preprocessed\_spotify.csv file was securely stored back in the GCP bucket. This storage safeguarded the data and made it readily available for further processing and analysis.

#### **3.6. BigQuery Integration and Visualization in Looker Studio**

The next phase involved BigQuery, where I uploaded the preprocessed data from the GCP bucket into a newly created dataset and table. This integration was a critical step, allowing me to dissect the data using various analytical lenses.

Following are SQL Queries and Outputs of respective Queries:

**Query 1: Top 50 artists with more popularity**

**A screenshot of a computer

Description automatically generated**

Fig 5

**Query 2: Danceability Score and Total no of Songs for that score**A screenshot of a computer

Description automatically generated

Fig 6

**Query 3: Energy of a Song and its Respective Popularity**

A screenshot of a computer

Description automatically generated

Fig 7

**Query 4: Artist their songs Loudness**A screenshot of a computer

Description automatically generated

Fig 8

**Query 5: Key and Count of Songs using that Key**A screenshot of a computer

Description automatically generated

Fig 9

**Query 6: Total count of songs in each of the countries by each artist**A screenshot of a computer

Description automatically generated

Fig 10

**Query 7: Popularity of a Songs based on three combined parameters.**

**A screenshot of a computer

Description automatically generated**

Fig 11

#### **3.7 Publish**

In the spirit of collaboration and contribution to the data science community, I published the cleaned data and associated Python scripts in a Git repository. This step not only marked the culmination of my project but also opened the door for others to explore, build upon, and derive their insights from this work. Please access the repository [here](https://github.iu.edu/savinn/INFO-I535-Top-Spotify-Songs).

**Data Pipeline**

A diagram of a diagram

Description automatically generated

Fig 12

**4. Results**

The following Visualizations are respectively arranged in the same order as that of query.

A graph of blue lines

Description automatically generated with medium confidence

Fig 13

A screenshot of a data

Description automatically generated

Fig 14

A graph with blue bars

Description automatically generated

Fig 15

A graph with different colored squares

Description automatically generated

Fig 16

A colorful circle with numbers and text

Description automatically generated

Fig 17

A graph of different colored bars

Description automatically generated

Fig 18

A graph with blue and pink lines

Description automatically generated

Fig 19

**5.Discussion**

* In Figure 13, the bar chart presents an overview of the average popularity among a selection of artists in the dataset. This visualization underscores the varying degrees of popularity, with some artists demonstrating a consistent appeal across the data sample. The average popularity metric is a quantitative proxy for the artists' current standing in the global music landscape, suggesting a correlation between their musical output and listener preferences**.**
* Figure 14 showcases a bar chart that correlates danceability scores with the count of songs possessing those scores. The prominence of higher danceability scores indicates a trend toward favoring songs with a solid rhythmic and danceable quality. This could reflect a cultural inclination towards upbeat and rhythmic music, which has implications for the production and marketing strategies in the music industry.
* Figure 15's bar chart scrutinizes the connection between a song's energy level and popularity. The visualization posits that while songs across the spectrum of energy levels can achieve popularity, there is a notable concentration around certain energy level. This pattern prompts a discussion on whether an optimal energy level makes a song more likely to be well-received by the audience.
* In Figure 16, the chart explores the relationship between artists and the loudness of their tracks. Loudness, an objective measure of sound intensity, varies significantly among artists, potentially alluding to the diversity in production techniques and the artists' musical styles. This plot invites further analysis of how loudness impacts listener perception and song popularity.
* The pie chart in Figure 17 delves into the musical keys in composing songs. The distribution across different keys might indicate the prevalence of specific keys in popular music or artists' preferences. The 'others' category suggests a wide variety of less standard keys, highlighting the musical diversity within the dataset.
* Figure 18's stacked bar chart compares the number of songs by various artists across different countries. This visual depiction offers insights into regional preferences, indicating which artists' works resonate more within specific countries. Such insights can inform targeted marketing and distribution strategies for artists and record labels.
* In Figure 19, the line chart investigates the interplay between three key attributes—danceability, acousticness, and energy—against the backdrop of song popularity. The intertwined lines suggest a complex relationship where no single attribute solely determines a song's popularity. Instead, the combination of these factors likely influences a track's success.

**5.Skills Implemented that I Learned during the Course**

Throughout my project, the teachings of the course were instrumental in guiding my methodology and execution. The course's modules on 'Cloud Computing' and 'Data Types and Sources' laid the groundwork for efficiently selecting and retrieving the "Top Spotify Songs" dataset from Kaggle. My subsequent data management strategy—centralizing storage on the Google Cloud Platform—directly applied the 'Virtualization' and 'Cloud Computing' concepts.

As I progressed through the project, the 'Lifecycles and Pipelines' and 'Ingest and Storage' concepts influenced my approach to data preprocessing, maintaining a streamlined workflow from raw data to refined analysis. Skills acquired in Modeling, processing, and Analytics were vital as I navigated the complexities of statistical analysis and data visualization, leading to insightful findings about global music trends. Moreover, the principles from 'Computing Principles and System Design' became evident in the systematic setup within Google Colab and the structured approach to coding. In the final stages, the 'Impact of Big Data' and 'Data Governance' modules resonated as I published my work, emphasizing the importance of sharing knowledge and fostering an open data community. This project was a harmonious blend of academic knowledge and practical application, demonstrating the real-world impact and relevance of the skills nurtured in this course.

**7.Failures Encountered**

I encountered a significant challenge during the phase of my project involving Google Cloud Dataproc. Linking Google Cloud Platform (GCP) to Google Colab was seamless; however, difficulties arose when attempting to create a cluster on Dataproc. My objective was to utilize this cluster to run a linear regression model as a PySpark job, but the cluster failed to materialize due to configuration complexities. Each attempt to initiate the cluster, even with varied configurations, resulted in a recurrent issue: my connections to the virtual machines were consistently refused.

To resolve this, I implemented a custom firewall policy within the VPC network settings, hoping to facilitate cluster creation. Despite these efforts, the persistent issue remained unresolved. The screenshots of my PySpark job code, alongside the VPC and cluster configuration settings, encapsulate the technical hurdles I faced. Dedicating approximately half a day to troubleshooting, I endeavored to debug the problem, but the cluster creation issue ultimately stood firm as a learning curve in my project's journey.

A screen shot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Fig 20

### 8.Conclusion

As I reflect on the culmination of this project, it's evident that the journey through the world of global music trends has been both enlightening and challenging. Extracting, analyzing, and visualizing data from the "Top Spotify Songs" dataset has reinforced the profound relationship between music and data analytics. I have pushed the boundaries of my analytical skills by successfully linking the Google Cloud Platform to Google Colab and leveraging BigQuery alongside Looker Studio for insightful visualizations. Despite encountering obstacles with Google Cloud Dataproc, which hindered my plans to run a PySpark job, I adapted and redirected my focus to ensure the project's progression. The lessons learned from these challenges have only deepened my resolve and expertise in data science.

This endeavor underscores the essence of resilience and adaptability in technical adversities. Although some goals remained unmet, the achievements in data processing, extracting meaningful insights, and sharing findings with the community are testaments to the successful application of academic principles to real-world scenarios. The visualizations crafted from this dataset serve not only as a representation of numerical data but also as a narrative that tells the story of what the world listens to. This project, while an end in itself, is also a gateway to further exploration and understanding of the intricate tapestry of global music preferences.

**9.References**

1. Top Spotify songs in 73 countries : <https://www.kaggle.com/datasets/asaniczka/top-spotify-songs-in-73-countries-daily-updated/>
2. Google Cloud VPC networks : <https://cloud.google.com/vpc/docs/vpc>
3. Dataproc: Qwik Start – Console : <https://www.cloudskillsboost.google/focuses/586?catalog_rank=%7B%22rank%22%3A7%2C%22num_filters%22%3A1%2C%22has_search%22%3Atrue%7D&parent=catalog&search_id=9028616>
4. Introduction to Cloud Dataproc: Hadoop and Spark on Google Cloud: <https://www.cloudskillsboost.google/focuses/672?catalog_rank=%7B%22rank%22%3A1%2C%22num_filters%22%3A0%2C%22has_search%22%3Atrue%7D&parent=catalog&search_id=9026299>
5. Use VPC firewall rules: <https://cloud.google.com/firewall/docs/using-firewalls>
6. Connect GCP Bucket to Google Colab [: https://pub.towardsai.net/connect-colab-to-gcs-bucket-using-gcsfuse-29f4f844d074](https://pub.towardsai.net/connect-colab-to-gcs-bucket-using-gcsfuse-29f4f844d074)