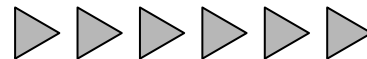


# Bans and Bubbles: Analysis of Moderation in VTuber Streams

Presentation by Jedi's Team



# TABLE OF CONTENTS

1. Abstract

2. Goals and Objectives

3. Tools and Techniques

4. ER Diagram

5. System Architecture

6. ELT Process

7. Exploratory Data Analysis

8. Website

9. Dashboard

10. Project Results and Deliverables

11. Future Work

12. Conclusion

# Abstract

- In the dynamic realm of online entertainment, Virtual YouTubers, or VTubers, have risen to prominence, captivating audiences with their mesmerizing blend of 2D, 3D, and anime-style avatars, coupled with their charismatic personalities.
- To gain a deeper understanding of this phenomenon, we conducted a comprehensive analysis of millions of Live Chats, Super Chats, and Moderation Events from VTubers' live stream data collected from Kaggle.
- Our quest is to assess the usefulness and impact of VTubers' content through concise analytics. We aimed to help viewers make informed decisions about their online entertainment choices by identifying VTubers with a consistent track record of positive content creation. Additionally, we sought to provide guidance on affiliations to help new YouTubers make informed decisions about which affiliations to prefer for a more positive and sustainable content creation journey.

# Goals And Objectives

- Assess the Usefulness and Impact of VTubers' Content: By analyzing the metrics such as chat engagement, Super Chat donations, and moderation events, we will evaluate the effectiveness and resonance of VTubers' content with their viewers.
- Provide Guidance on Affiliations for Aspiring VTubers: Identifying VTubers with a consistent track record of positive content creation aims to facilitate affiliation agencies in providing crucial support. This, in turn, opens avenues for lucrative brand collaborations and sponsorships, contributing significantly to the growth and success of aspiring VTubers under their guidance.
- We have achieved this goal by utilizing the data orchestration tools of the Azure ecosystem, involving the loading of data into Azure Data Lake, processing in Azure Databricks, Exploratory Data Analysis in Synapse, and Visualization using PowerBi. Our ultimate objective is to develop a user-friendly website showcasing the insights and findings derived from our comprehensive investigation.

# Data Description

- The data is in the form of Parquet files, where about 13GB of Vtuber Data is collected from Kaggle.
- Vtuber Data consists of the information about the Vtuber Channel, Vtuber Chat data, Vtuber Superchat Data, Vtuber Chat Statistics, Vtuber Super Chat Statistics and information about Ban and Deletion Events.
- The data is ingested into the Azure Data Lake, where all the collected Parquet files are stored and fetched for further Data Processing.
- The Collected files are further merged into 4 separate files based on the context and relevance of the information available in them:
  - Merged Data File for Channel, Chats, Ban and Deletion files
  - Merged Data File for Channel and Chat Statistics File
  - Merged Data File for Channel and Super Chat Statistics File
  - Merged Data File for Channel and Super Chat File

# Initial Raw Data Files

Microsoft Azure

Search resources, services, and docs (G+)

poojan.gagrani@jsu.edu  
JSU.EDU (JSU6.ONMICROSOFT...)

Home > vtuberdata | Containers >

vtubercontainer

Container

Search

Upload Change access level Refresh Delete Change tier Acquire lease Break lease View snapshots Create snapshot Give feedback

Overview

Diagnose and solve problems

Access Control (IAM)

Settings

Shared access tokens

Access policy

Properties

Metadata

Authentication method: Access key (Switch to Microsoft Entra user account)

Location: vtubercontainer

Search blobs by prefix (case-sensitive)

Show deleted blobs







Add filter

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state	
<input type="checkbox"/> ban_events.parquet	11/16/2023, 1:59:45 AM	Hot (Inferred)		Block blob	25.64 MiB	Available	***
<input type="checkbox"/> channels.csv	11/22/2023, 12:09:05 PM	Hot (Inferred)		Block blob	276.93 KiB	Available	***
<input type="checkbox"/> chat_stats.csv	11/22/2023, 12:09:05 PM	Hot (Inferred)		Block blob	686.55 KiB	Available	***
<input type="checkbox"/> chats_2022-01.parquet	11/16/2023, 12:33:37 PM	Hot (Inferred)		Block blob	4.25 GiB	Available	***
<input type="checkbox"/> chats_2022-02.parquet	11/16/2023, 3:32:38 PM	Hot (Inferred)		Block blob	3.78 GiB	Available	***
<input type="checkbox"/> chats_2022-03.parquet	11/16/2023, 3:51:25 PM	Hot (Inferred)		Block blob	4.28 GiB	Available	***
<input type="checkbox"/> chats_2022-04.parquet	11/16/2023, 4:34:40 PM	Hot (Inferred)		Block blob	4.44 GiB	Available	***
<input type="checkbox"/> chats_2022-05.parquet	11/16/2023, 4:34:25 PM	Hot (Inferred)		Block blob	4.37 GiB	Available	***
<input type="checkbox"/> chats_2022-06.parquet	11/16/2023, 4:33:12 PM	Hot (Inferred)		Block blob	4.02 GiB	Available	***
<input type="checkbox"/> deletion_events.parquet	11/16/2023, 10:44:09 AM	Hot (Inferred)		Block blob	227.91 MiB	Available	***
<input type="checkbox"/> superchat_stats.csv	11/22/2023, 12:09:05 PM	Hot (Inferred)		Block blob	751.16 KiB	Available	***
<input type="checkbox"/> superchats_2022-01.parquet	11/16/2023, 10:34:38 AM	Hot (Inferred)		Block blob	18.03 MiB	Available	***
<input type="checkbox"/> superchats_2022-02.parquet	11/16/2023, 1:24:55 PM	Hot (Inferred)		Block blob	13.58 MiB	Available	***
<input type="checkbox"/> superchats_2022-03.parquet	11/16/2023, 1:24:07 PM	Hot (Inferred)		Block blob	15.38 MiB	Available	***
<input type="checkbox"/> superchats_2022-04.parquet	11/16/2023, 1:24:12 PM	Hot (Inferred)		Block blob	15.41 MiB	Available	***
<input type="checkbox"/> superchats_2022-05.parquet	11/16/2023, 1:25:56 PM	Hot (Inferred)		Block blob	15.42 MiB	Available	***
<input type="checkbox"/> superchats_2022-06.parquet	11/16/2023, 1:24:24 PM	Hot (Inferred)		Block blob	13.74 MiB	Available	***

# Merged Data Files

Microsoft Azure

Search resources, services, and docs (G+/)



poojan.gagrani@sjsu.edu  
SJSU.EDU (SJSU.ONMICROSOFT...)

Home > vtuberdata | Containers >

vtuberfinal

Container

Search

«

Upload

Change access level

Refresh

Delete

Change tier

Acquire lease

Break lease

View snapshots

Create snapshot

Give feedback

Overview

Diagnose and solve problems

Access Control (IAM)

Settings

Shared access tokens

Access policy

Properties

Metadata





**Authentication method:** Access key (Switch to Microsoft Entra user account)

**Location:** vtuberfinal

Search blobs by prefix (case-sensitive)

Show deleted blobs

Add filter

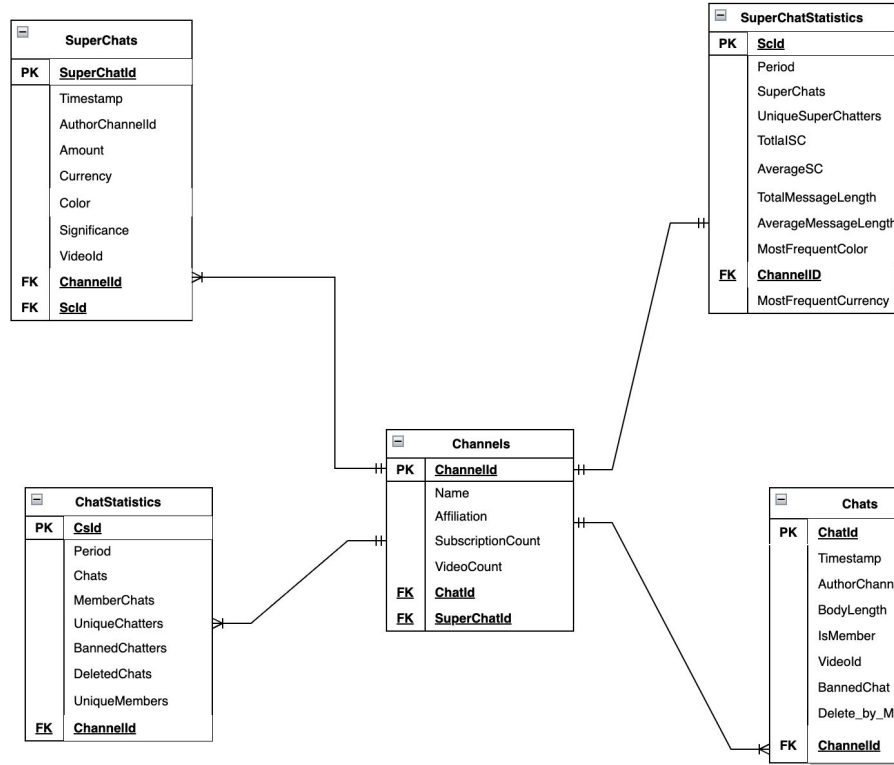
	Name	Modified	Access tier	Archive status	Blob type	Size	Lease state	
<input type="checkbox"/>	 superchat_stats_channel.txt	11/26/2023, 2:13:52 PM	Hot (Inferred)		Block blob	276.23 KiB	Available	***
<input type="checkbox"/>	 chat_channel_merged.txt	11/26/2023, 2:07:04 PM	Hot (Inferred)		Block blob	1.51 GiB	Available	***
<input type="checkbox"/>	 chat_stats_channel.txt	11/26/2023, 2:07:08 PM	Hot (Inferred)		Block blob	274.1 KiB	Available	***
<input type="checkbox"/>	 superchat_channel.csv	11/26/2023, 2:10:34 PM	Hot (Inferred)		Block blob	153.78 MiB	Available	***

# Tools and Techniques

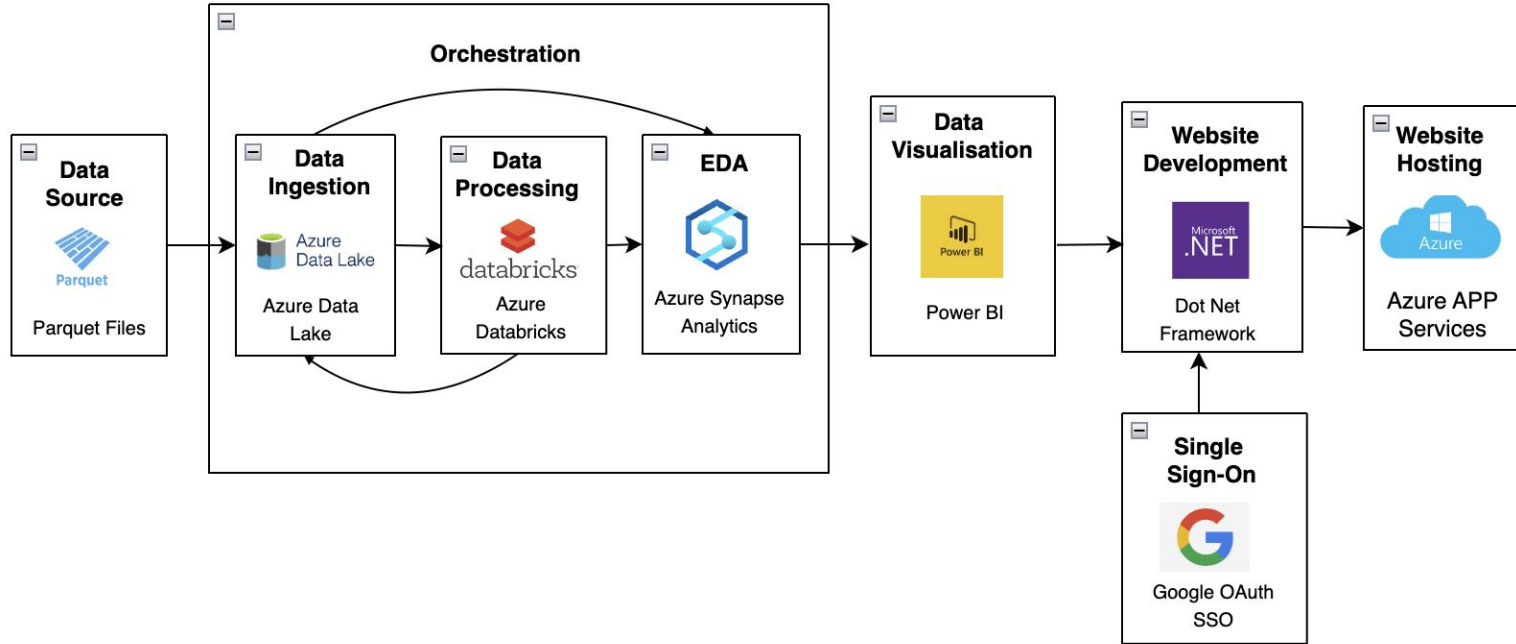
1. **Data Source** - Parquet Files
2. **Data Ingestion** - Azure Data Lake
3. **Data Processing** - Azure Databricks
4. **EDA** - Azure Synapse Analytics
5. **Analysis and Visualizations** - Power BI
6. **Website Development** - Dot Net Framework
7. **Single Sign On** - Google OAuth SSO
8. **Website Hosting** - Azure Web Services



# ER Diagram



# System Architecture



# ELT Process

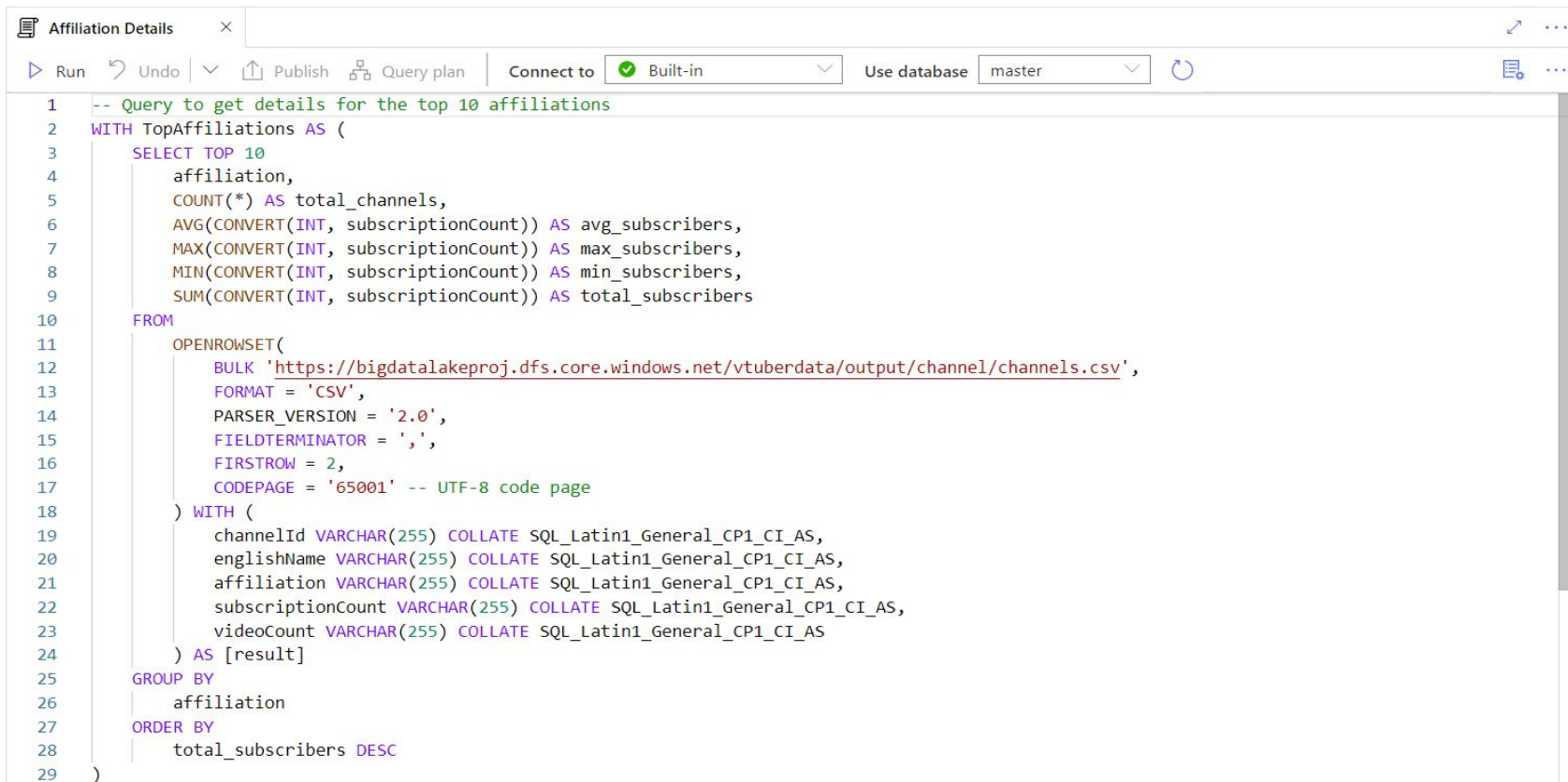
- The process begins with data sources in Parquet file format, known for efficient storage and performance on large datasets.
- Data is then ingested into the Azure Data Lake, a scalable storage repository that provides a massive scale data lake for big data analytics
- For further processing, Azure Databricks, an Apache Spark-based analytics platform is used. It provides collaborative notebooks, integrated workflows, and an interactive workspace that enable easy orchestration of complex data transformations.
- Azure Synapse Analytics, an integrated analytics service, allows for exploratory data analysis. It gives insights into the data through various analytical methods to inform the subsequent steps in the ELT process.

# ELT Process

- Data is visualized using Power BI, a business analytics service by Microsoft. It enables interactive visualizations and business intelligence capabilities with an interface simple enough for end users to create their own reports and dashboards.
- The data processed and visualized is then used in Website Development, leveraging the .NET Framework for robust and scalable web applications
- Integration with Single Sign-On (SSO) via Google OAuth ensures that users can seamlessly authenticate with their Google credentials.
- Throughout the process, Azure services manage the orchestration and optimization of the ELT pipeline, ensuring efficient data flow and transformation.

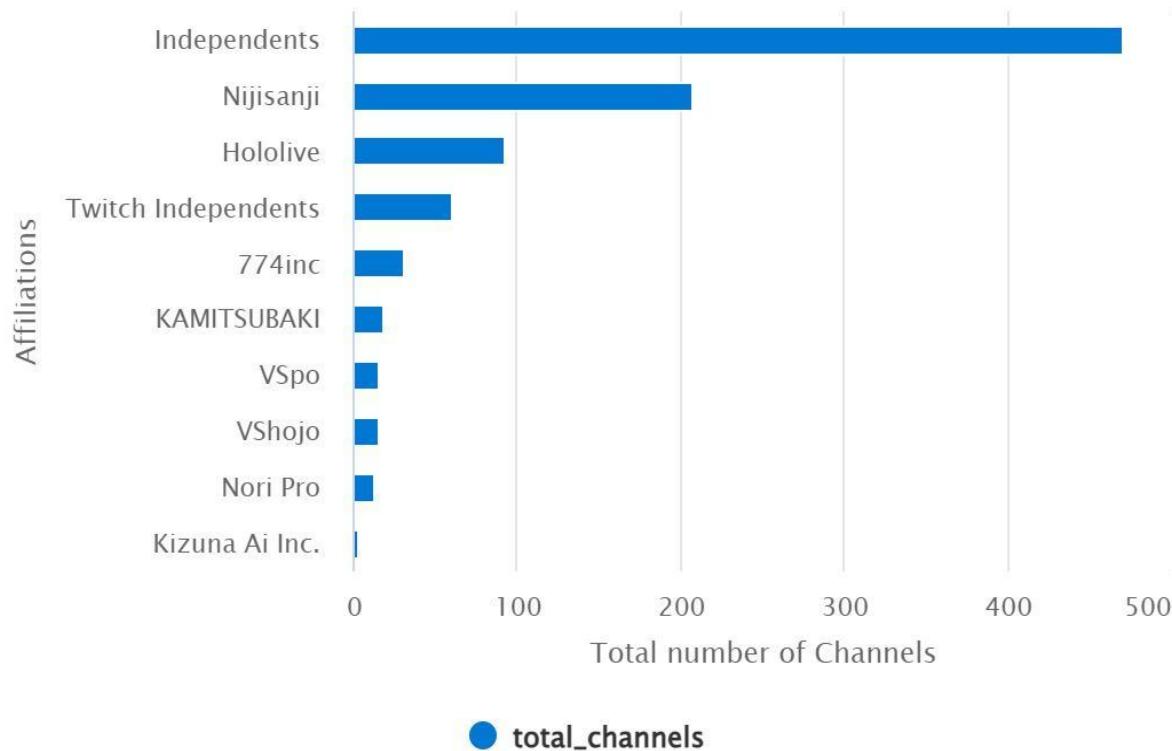
# Code Walkthrough.

# Exploratory Data Analysis (Code)



```
1  -- Query to get details for the top 10 affiliations
2  WITH TopAffiliations AS (
3      SELECT TOP 10
4          affiliation,
5          COUNT(*) AS total_channels,
6          AVG(CONVERT(INT, subscriptionCount)) AS avg_subscribers,
7          MAX(CONVERT(INT, subscriptionCount)) AS max_subscribers,
8          MIN(CONVERT(INT, subscriptionCount)) AS min_subscribers,
9          SUM(CONVERT(INT, subscriptionCount)) AS total_subscribers
10     FROM
11         OPENROWSET(
12             BULK 'https://bigdatalakeproj.dfs.core.windows.net/vtuberdata/output/channel/channels.csv',
13             FORMAT = 'CSV',
14             PARSER_VERSION = '2.0',
15             FIELDTERMINATOR = ',',
16             FIRSTROW = 2,
17             CODEPAGE = '65001' -- UTF-8 code page
18         ) WITH (
19             channelId VARCHAR(255) COLLATE SQL_Latin1_General_CP1_CI_AS,
20             englishName VARCHAR(255) COLLATE SQL_Latin1_General_CP1_CI_AS,
21             affiliation VARCHAR(255) COLLATE SQL_Latin1_General_CP1_CI_AS,
22             subscriptionCount VARCHAR(255) COLLATE SQL_Latin1_General_CP1_CI_AS,
23             videoCount VARCHAR(255) COLLATE SQL_Latin1_General_CP1_CI_AS
24         ) AS [result]
25     GROUP BY
26         affiliation
27     ORDER BY
28         total_subscribers DESC
29 )
```

# Exploratory Data Analysis (Output)



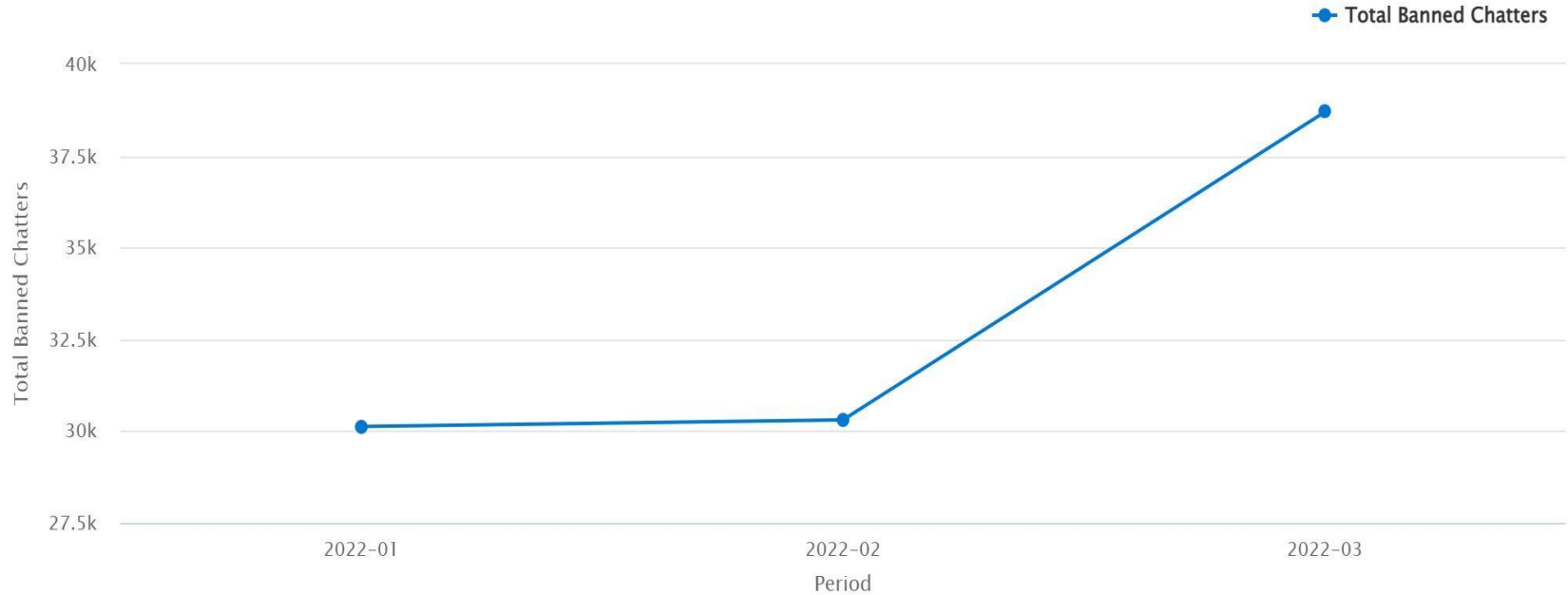
# Exploratory Data Analysis (Code)

```
vtuberdata | Max unique chatters ... | Total chats by channel | Total Banned Chatter... ×
Run | Undo | Publish | Query plan | Connect to Built-in | Use database master | Refresh

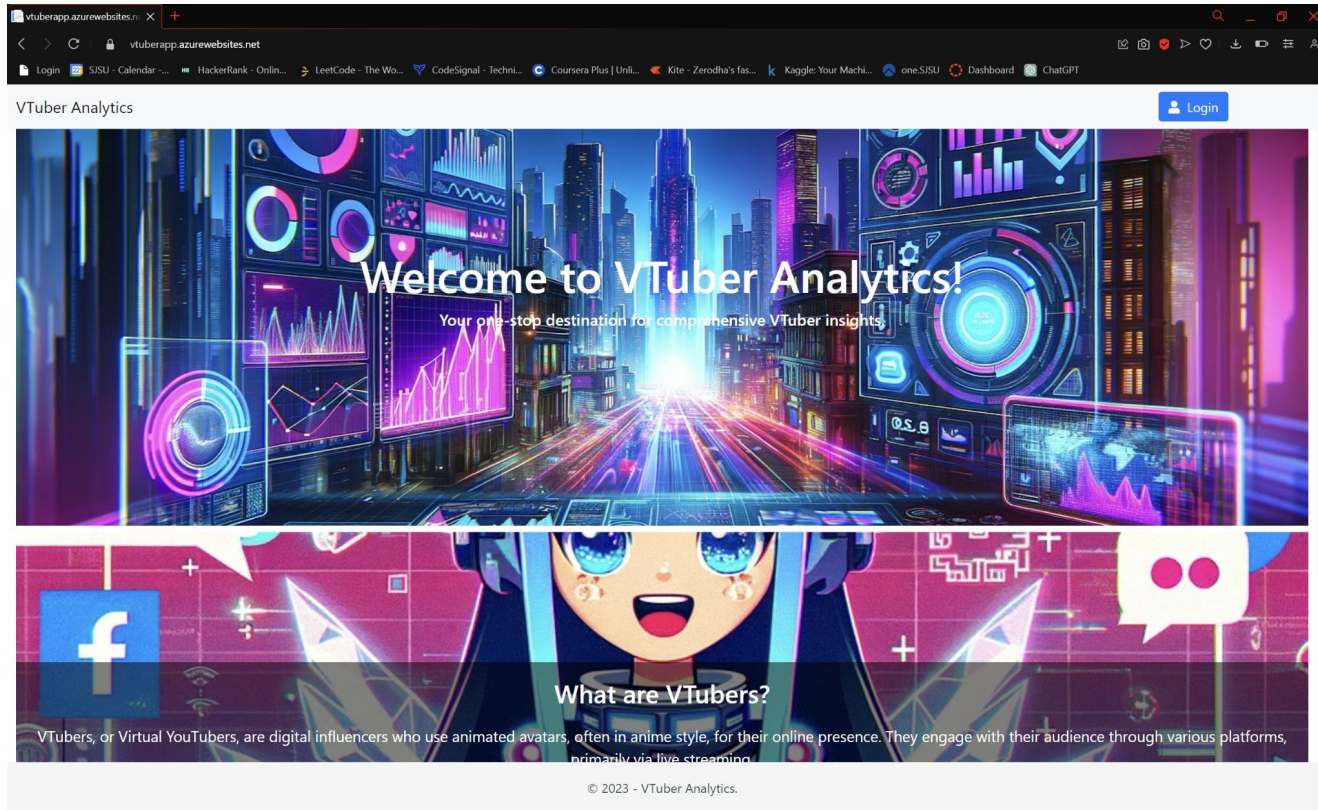
1  -- Exploratory Data Analysis (EDA) for Total Banned Chatters by Period
2
3  SELECT
4      period AS [Period],
5      SUM(bannedChatters) AS [Total Banned Chatters]
6  FROM
7      OPENROWSET(
8          BULK 'https://bigdatalakeproj.dfs.core.windows.net/vtuberdata/output/chat_stats/chat_stats.csv',
9          FORMAT = 'CSV',
10         PARSE_VERSION = '2.0',
11         FIELDTERMINATOR = ',',
12         FIRSTROW = 2,
13         CODEPAGE = '65001' -- UTF-8 code page
14     ) WITH (
15         channelId VARCHAR(255) COLLATE SQL_Latin1_General_CP1_CI_AS,
16         period VARCHAR(255) COLLATE SQL_Latin1_General_CP1_CI_AS,
17         chats INT,
18         memberChats INT,
19         uniqueChatters INT,
20         uniqueMembers INT,
21         bannedChatters INT,
22         deletedChats INT,
23         chatStatsId INT
24     ) AS [result]
25  GROUP BY
26      period
27  ORDER BY
28      period;
29
```



# Exploratory Data Analysis (Output)

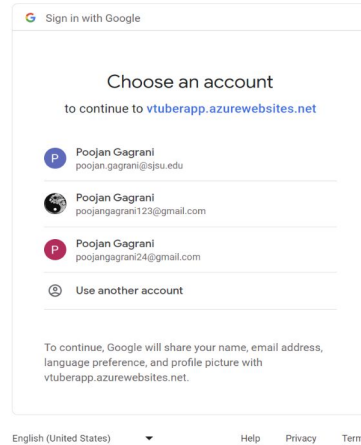
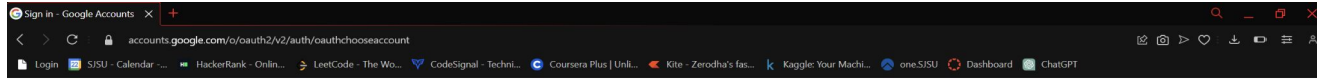


# Website



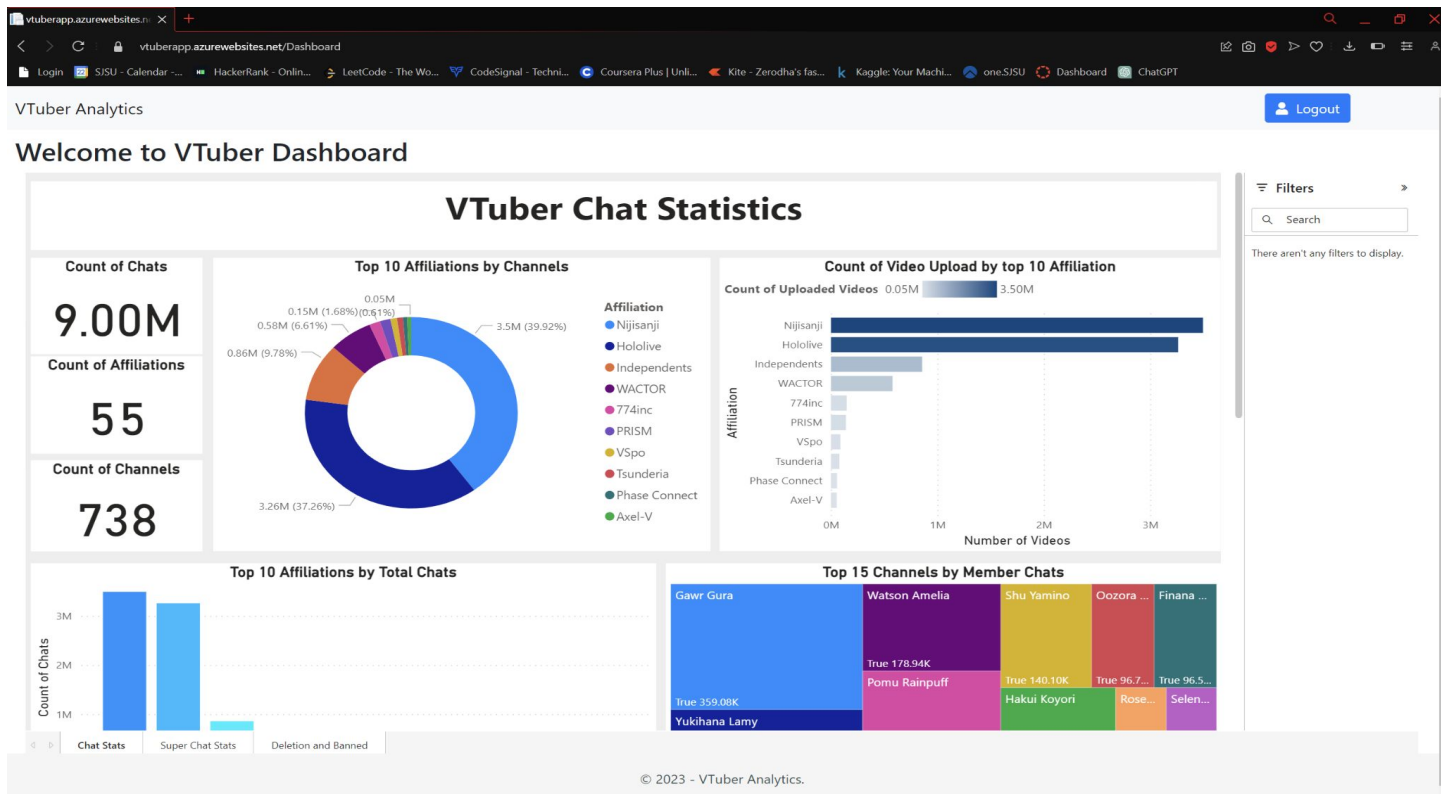
Landing page of website hosted at: <https://vtuberapp.azurewebsites.net>

# Website



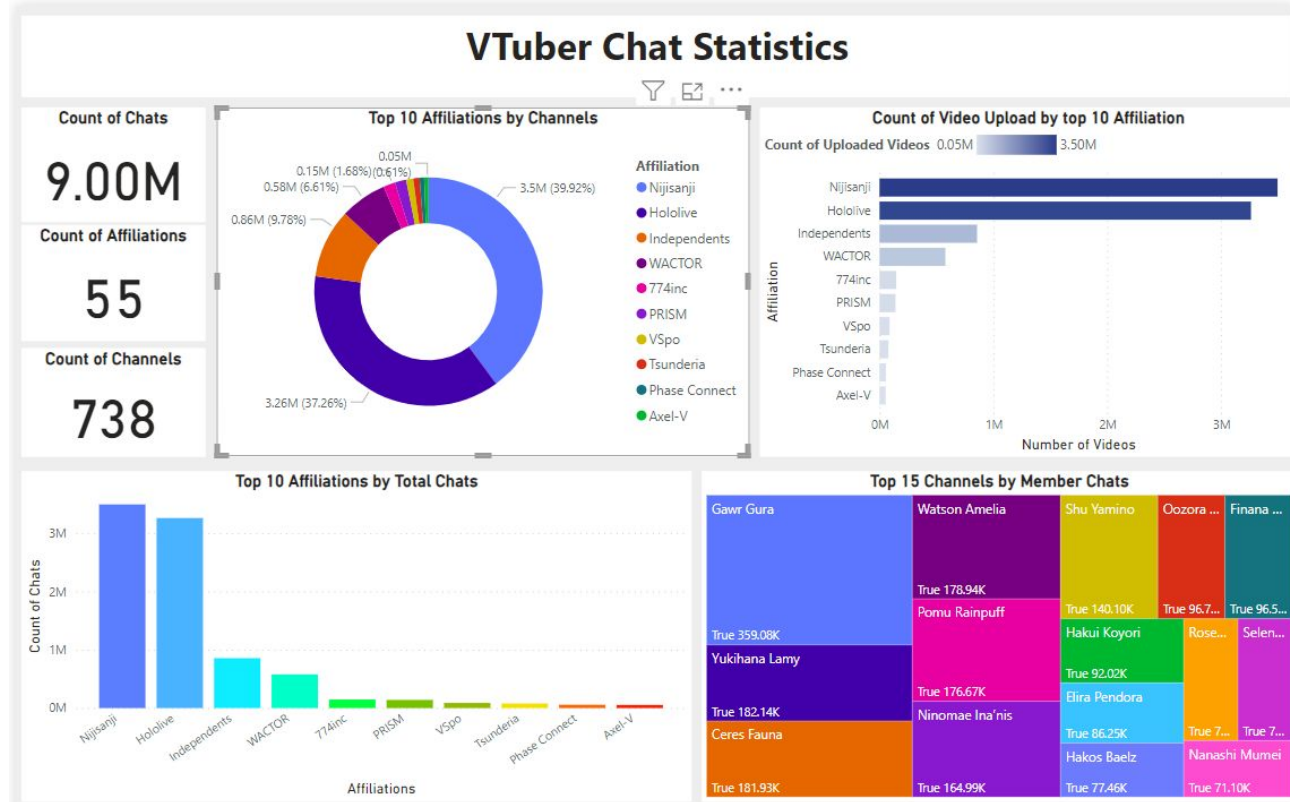
Login using Google OAuth (SSO)

# Website



Dashboard page after successful login

# Vtubers - Dashboard (Chat Stats)



Filters

Search

Filters on this visual

affiliation  
top 10 by Count of en...

Count of englishName  
is (All)

Add data fields here

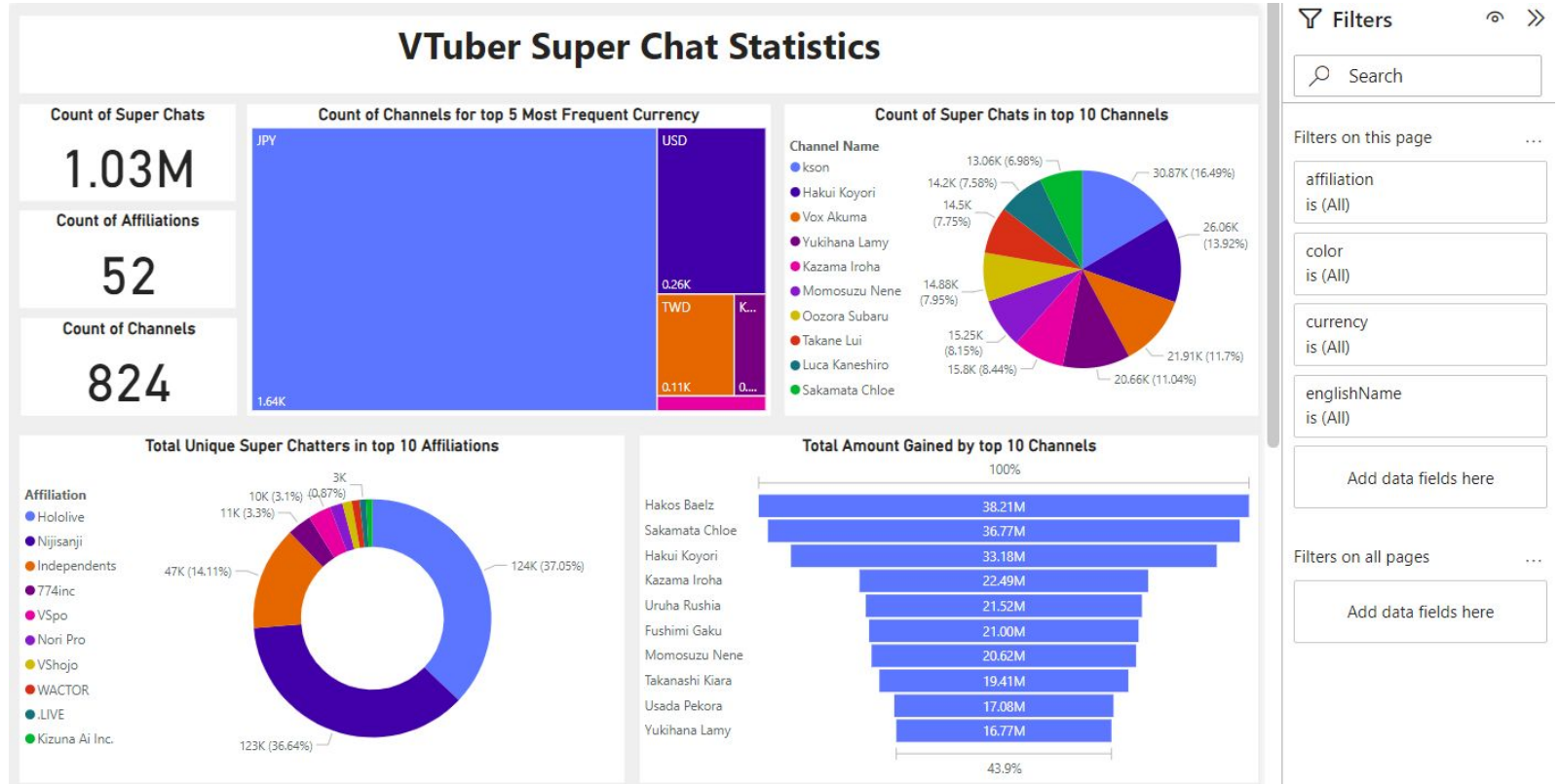
Filters on this page

Add data fields here

Filters on all pages

Add data fields here

# Vtubers - Dashboard (Superchat Stats)



# Vtubers - Dashboard (Deletion and Banned Stats)

## VTuber Banned and Deleted Chats Statistics

Count of Chats

9.00M

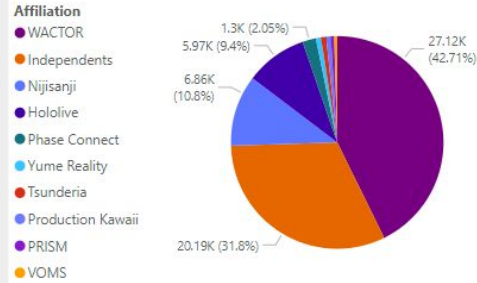
Count of Banned Chats

64.39K

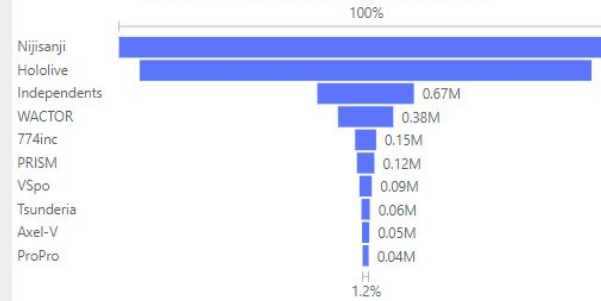
Count of Deleted Chats

8.35M

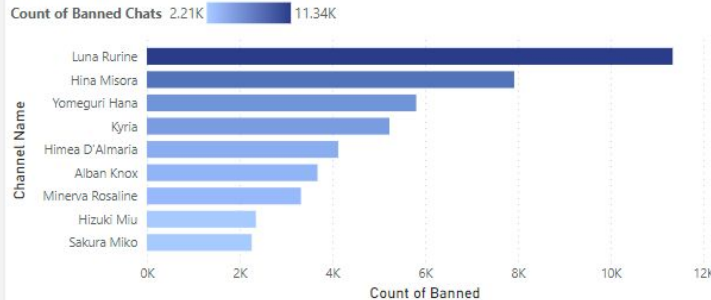
Count of Banned Chats for top 10 Affiliations



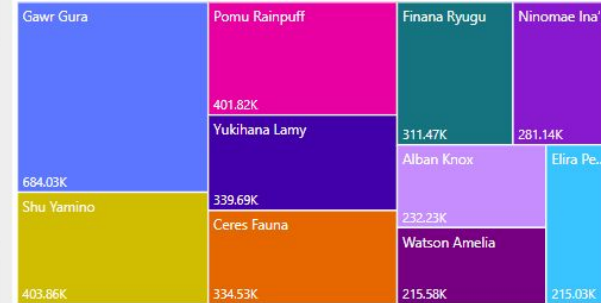
Count of Deleted Chats for top 10 Affiliations



Count of Banned Chats for top 10 Channels



Count of Deleted Chats for top 10 Channels



# Project Results & Deliverables

- Developed Interactive Analytics Dashboard for VTuber event data
- Performed In-Depth VTuber Engagement Analysis
- Gained insights on Content Moderation in VTuber streams
- Utilized Azure Synapse Analytics for EDA
- Created dynamic Data Visualizations with Power BI
- Implemented Google OAuth Single Sign-On



# Future Work

- Building an advanced pipeline which can handle even larger data (TB's) of data at scale.
- Better data preprocessing can be achieved using more capable spark clusters which handle & preprocess larger datasets.
- Incorporating more statistics when larger set of data is available.
- Enhancing website by integrating multiple login options such as AzureAD, Facebook and Twitter logins.
- Enabling user specific dashboards website e.g, chat stats dashboard for streamer which focuses on stats that are particular to that streamer which enables them to access insights at ease.

# Conclusion

- In conclusion, the project merges data orchestration with advanced analytics to enhance the virtual YouTuber ecosystem.
- It utilizes a suite of Azure services, Power BI, and .NET to provide actionable insights on live streams, promoting positive content creation and informed viewer engagement.
- This integrated approach offers a new dimension in supporting sustainable virtual communities.

Thank you!

