Team Members:
1. Niraj Sai Prasad - NUID 001006514
2. Sindhu Swaroop - NUID 001006558
3. Vatsal Doshi - NUID 002776613
4. Avani Kala - NUID 002772623

Project Report for Topic:

Twitter: Summing up the Pulse of the Internet

P1: Project Proposal

Topic: Twitter: Summing up the Pulse of the Internet
Data Model: Document (NoSQL)
Target Platform: Azure SQL Multi-Model
Objective/Scope:
● Scrap real-time tweets, Twitter user details, URLs and images off of Twitter
● Perform sentiment analysis on tweets to find the general pulse of Twitterati
● Compare statistics across tweets and find the most trending type of tweets for a particular period of time
● Find the top 10 users who have the most tweets and look for any correlation between the number of tweets and their followers and following
● Perform location-based analysis on trending hashtags across regions (group by location)
● Perform analysis on the top 10 most followed Twitter users to understand what gender, age group and nationalities they most appeal to
Visualizations Tool: Tableau/ Power BI

P3: Implementation

1) A brief description of the implementation process

The aim of our project is to use ETL tools to *extract* data from twitter, *transform* the data and *load* it onto our Azure database.

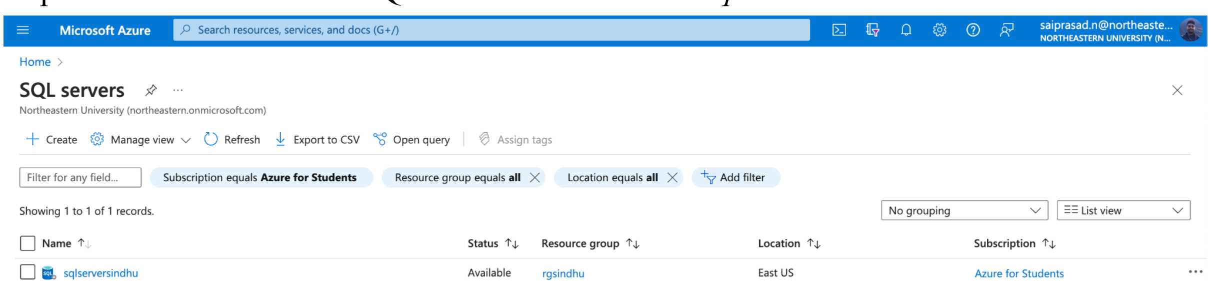
1. First, we wrote a Python script that includes API calls to Twitter to extract tweet data and some tweet details like twitter user, location, hashtags, tweet text and follower count. The script also performs some transformation on the data like data cleaning.
2. Next we created our SQL server, source blob, destination database, and batch pools. For the pools, we used a standard A2 windows VM instance with two nodes and two CPUs.
3. We then incorporated the python script in a batch job ("*runpythonscript*") which is essentially the start of our data pipeline. This batch job runs the python script, extracts data from Twitter into data frames, puts the data into CSV files which are then automatically uploaded into Azure blobs (our data source).
4. Upon successful loading of data into the source, the second step in our pipeline is multiple jobs to copy the data from source (blob storage) to destination (database) using column mapping in Azure.
5. For the document data model, we linked a Cosmos DB resource instance to the data factory as a destination for the hashtag data.

6. For the graph data model, we created Nodes and Edges files to be pushed to our SQL destination database.
7. As a third step in our pipeline, we also added a job to delete all CSV files from the blob once the data is copied into the destination.

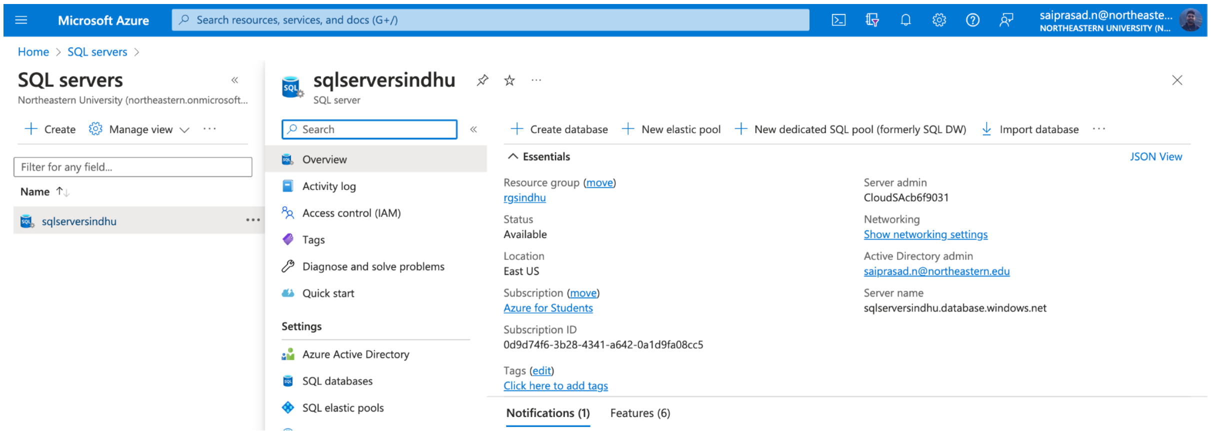
We then plan to perform analysis and visualizations on this data using PowerBI.

2) Screenshots which capture key steps in implementing the database and other architecture components

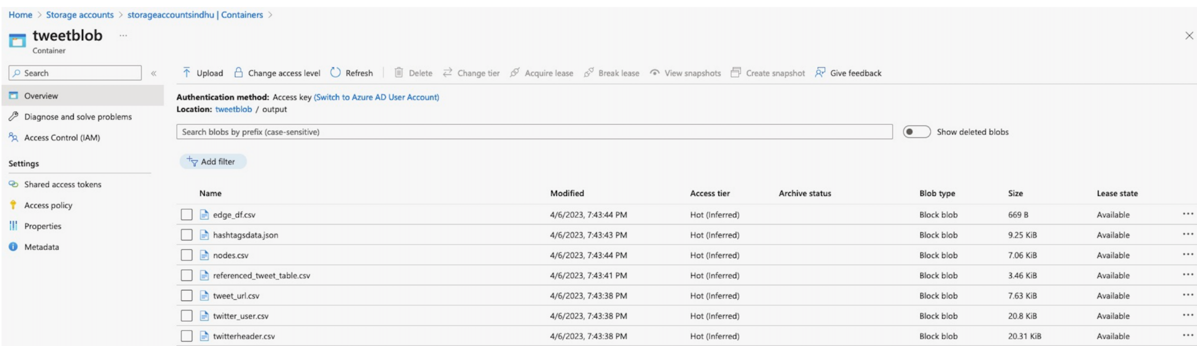
Step 1: We created our own SQL server with the name *sqlserversindhu.database.windows.net*



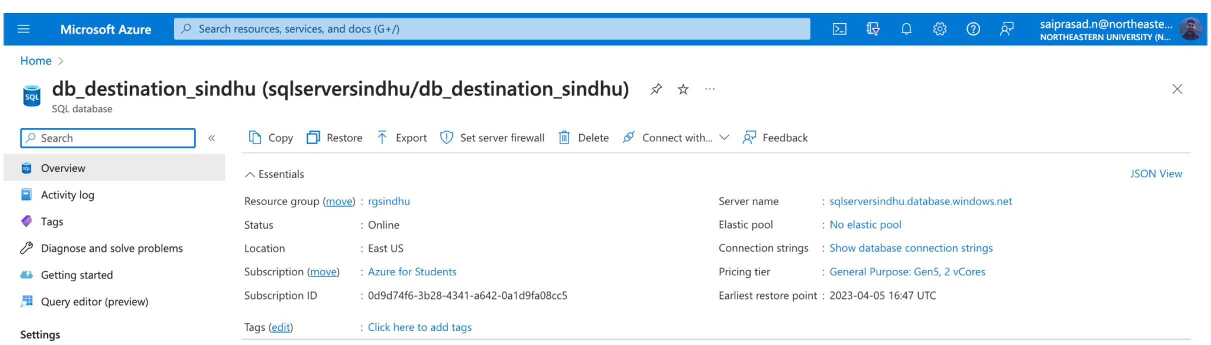
Server Properties:



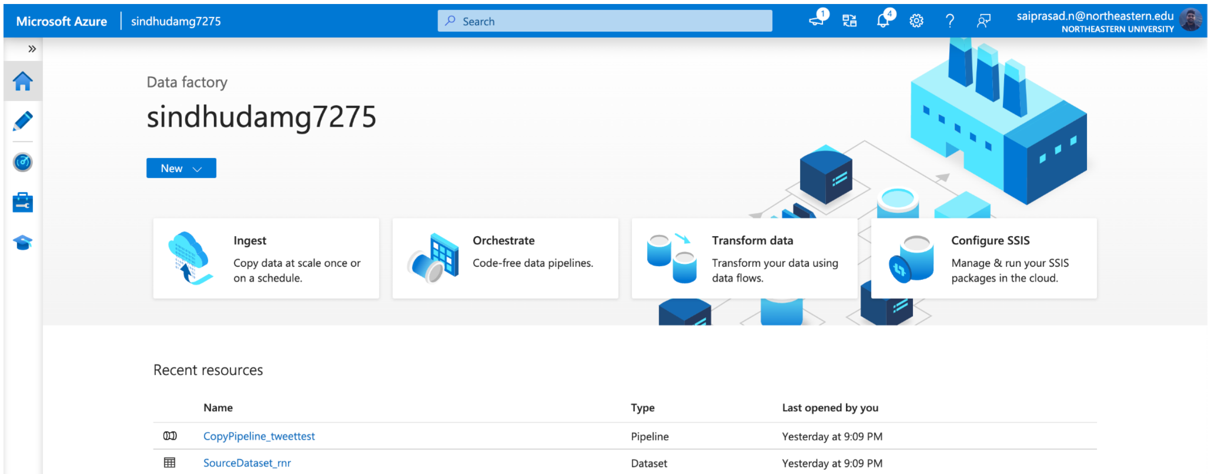
Step 2: We created the source azure storage blob container as *tweetblob* in the SQL server.



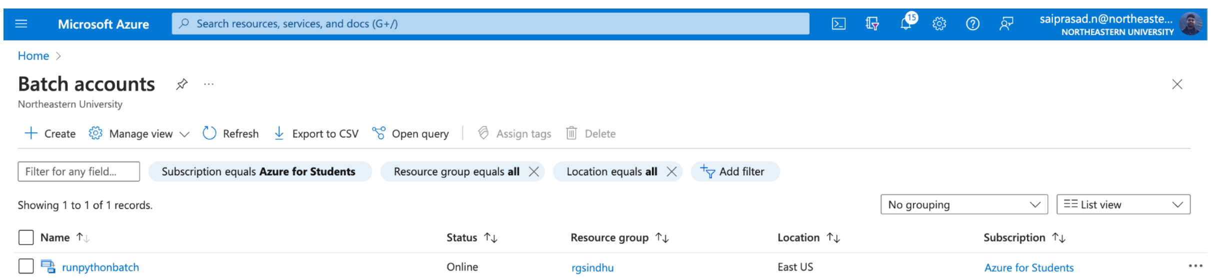
Step 3: We created the destination azure SQL database as *db_destination_sindhu* in the SQL server.

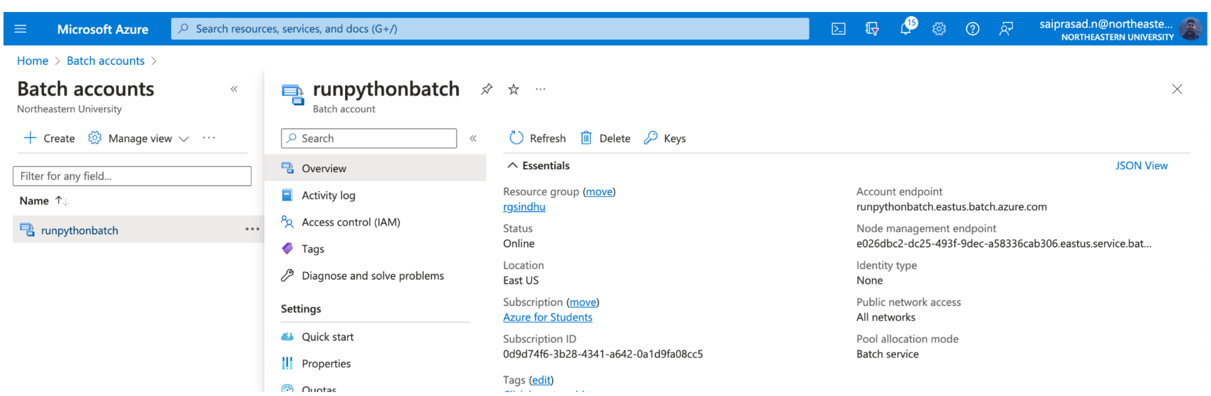


Step 4: Next we created a Data Factory to build a pipeline to move the data from source to destination using ETL operations. We set the source of the data factory as *tweetblob* and destination as *db_destination_sindhu* and *sindhucosmosdb*.

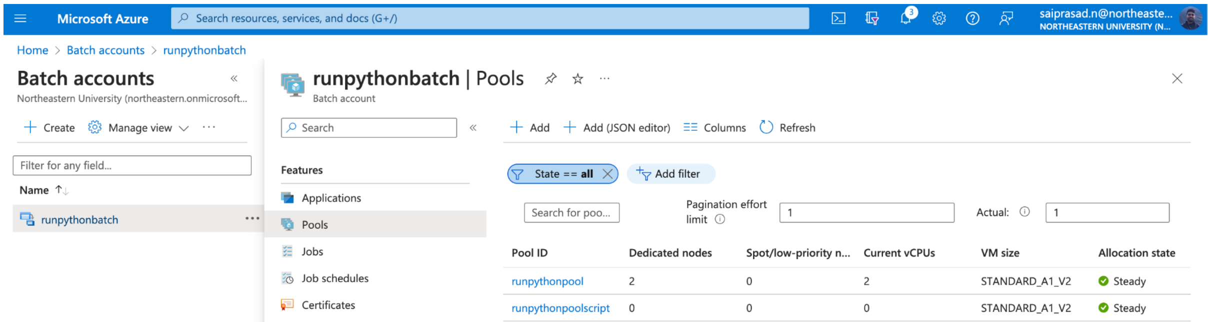


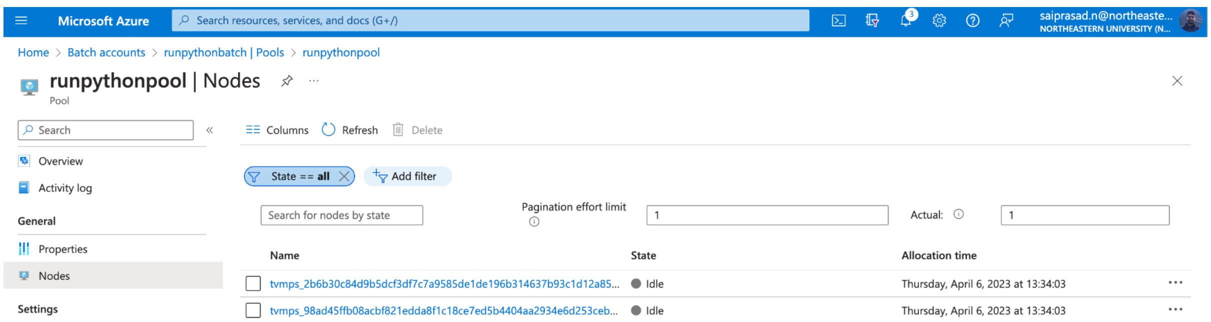
Step 5: We then created a batch account and a batch job *runpythonbatch* to run the python script that extracts twitter data via API.



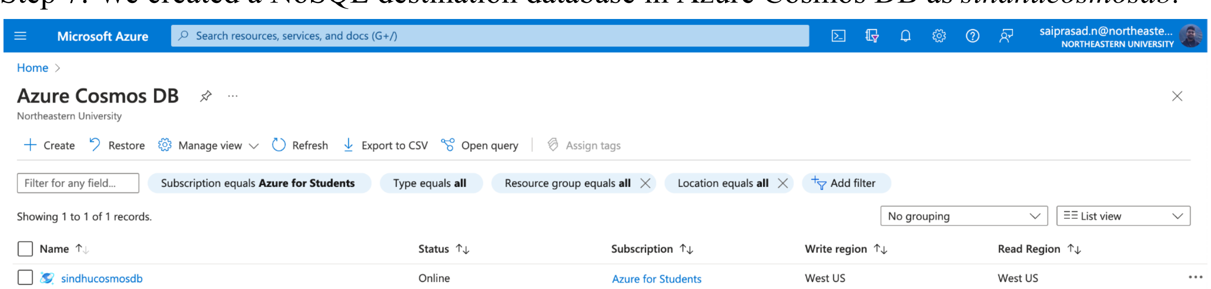


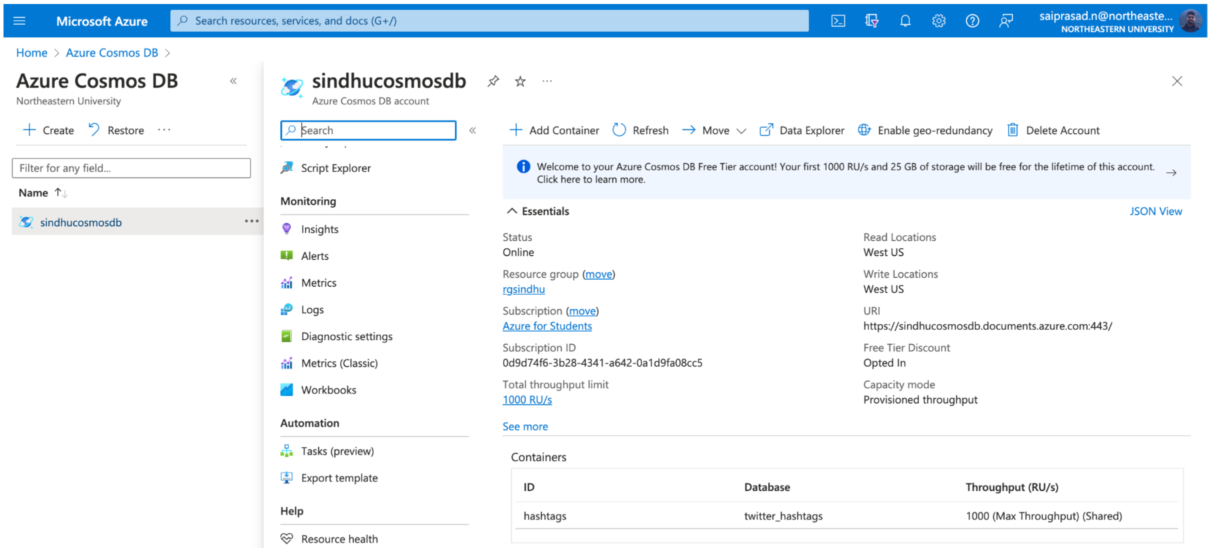
Step 6: We configured a pool *runpythonpool* with two nodes and two CPUs each of which are small virtual machines that execute the jobs depending on which machine has the bandwidth.

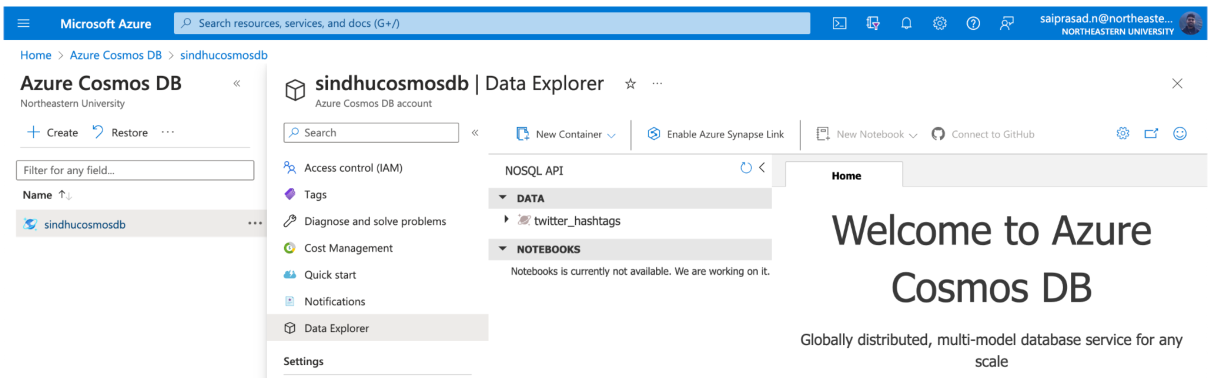




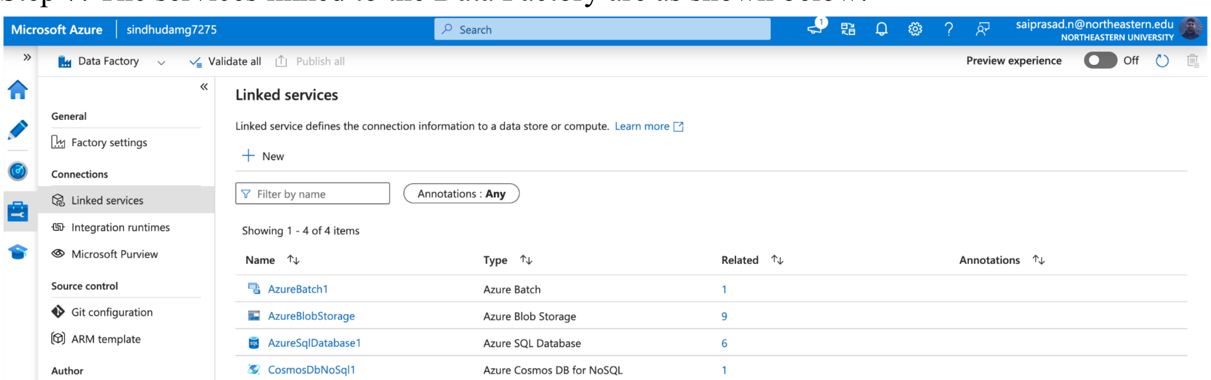
Step 7: We created a NoSQL destination database in Azure Cosmos DB as *sindhucosmosdb*.





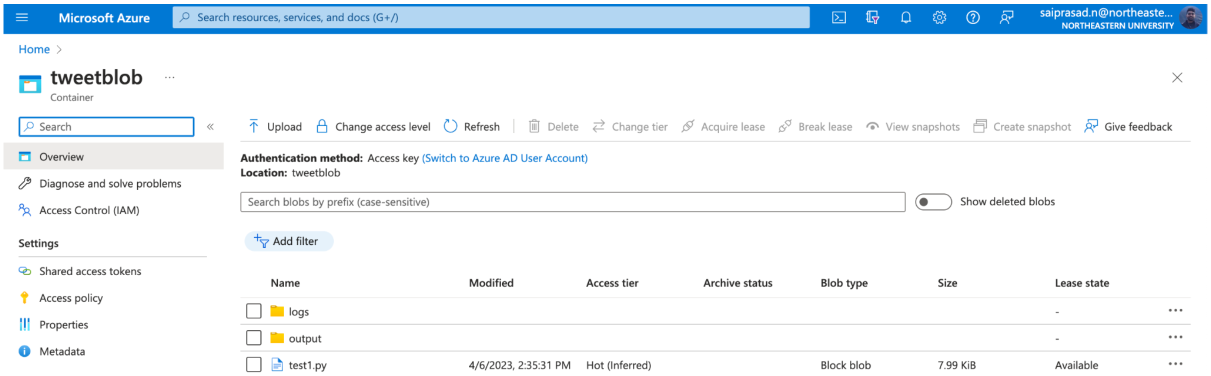


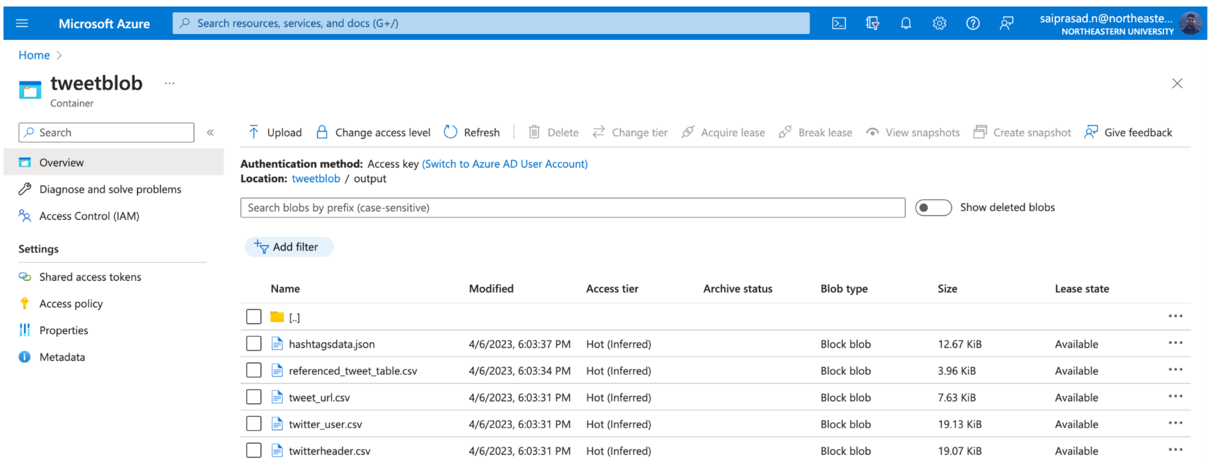
Step 7: The services linked to the Data Factory are as shown below:



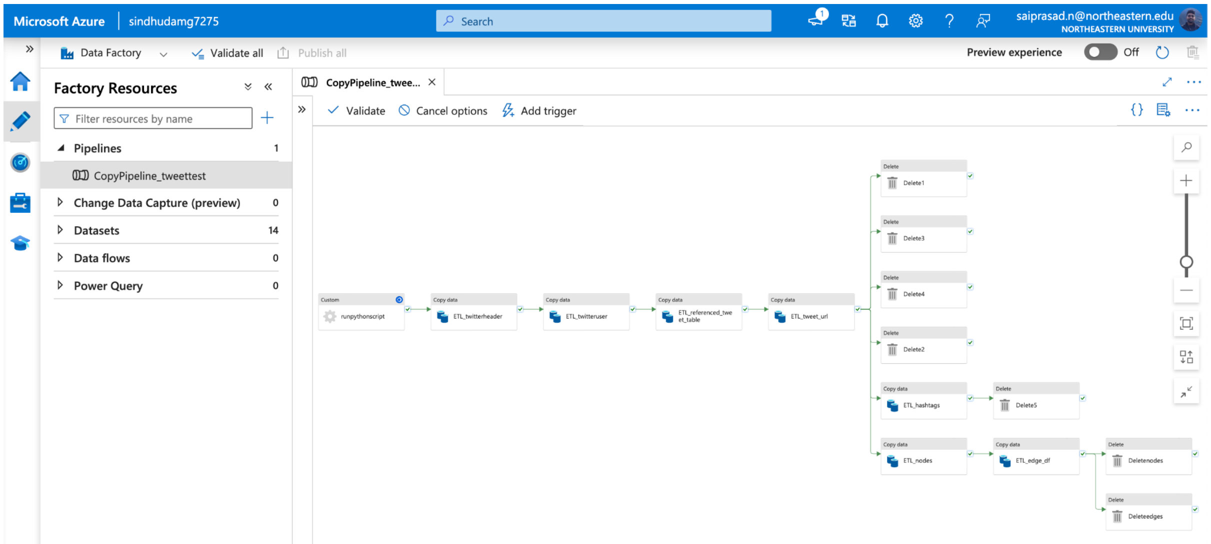
- Azurebatch1 refers to the batch job *runpythonbatch*
- AzureBlobStorage refers to the source storage blob *tweetblob*
- AzureSqlDatabase1 refers to the destination database *db_destination_sindhu*
- CosmosdbNosql1 refers to the NoSQL database *sindhucosmosdb*

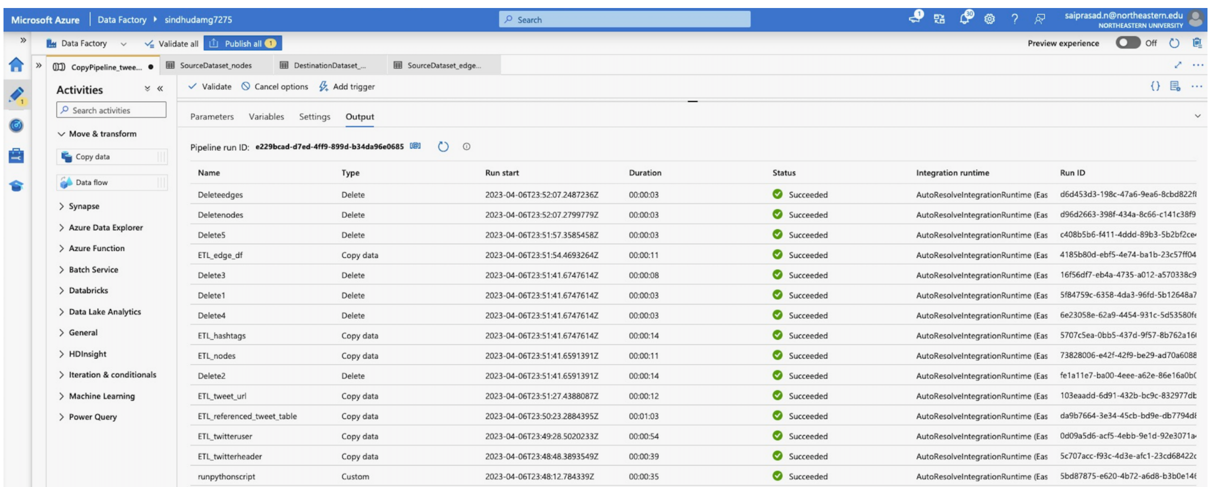
Step 8: After running the batch job in the data pipeline, the *tweetblob* looks like this:





Step 9: We created a data pipeline *CopyPipeline_tweettest* with 3 major steps to extract, copy data to destination and delete CSV files from source all in sequence.





Name	Type	Run start	Duration	Status	Integration runtime	Run ID
Deleteedges	Delete	2023-04-06T23:52:07.2487236Z	00:00:03	Succeeded	AutoResolveIntegrationRuntime (Eas	d6d453d3-198c-47a6-9ea6-8cbd822f
Deletenodes	Delete	2023-04-06T23:52:07.2799779Z	00:00:03	Succeeded	AutoResolveIntegrationRuntime (Eas	d96d2663-398f-434a-8c66-c141c38f9
Delete5	Delete	2023-04-06T23:51:57.3585458Z	00:00:03	Succeeded	AutoResolveIntegrationRuntime (Eas	c408b5b6-f411-4ddd-89b3-5b2bf2ce
ETL_edge_df	Copy data	2023-04-06T23:51:54.4693264Z	00:00:11	Succeeded	AutoResolveIntegrationRuntime (Eas	4185b80d-ebf5-4e74-ba1b-23c57ff04
Delete3	Delete	2023-04-06T23:51:41.6747614Z	00:00:08	Succeeded	AutoResolveIntegrationRuntime (Eas	16f56df7-eb4a-4735-a012-a570338c9
Delete1	Delete	2023-04-06T23:51:41.6747614Z	00:00:03	Succeeded	AutoResolveIntegrationRuntime (Eas	5f84759c-6358-4da3-96fd-5b12648a7
Delete4	Delete	2023-04-06T23:51:41.6747614Z	00:00:03	Succeeded	AutoResolveIntegrationRuntime (Eas	6e23058e-62a9-4454-931c-5d53580f
ETL_hashtags	Copy data	2023-04-06T23:51:41.6747614Z	00:00:14	Succeeded	AutoResolveIntegrationRuntime (Eas	5707c5ea-0bb5-437d-9f57-8b762a16
ETL_nodes	Copy data	2023-04-06T23:51:41.6591391Z	00:00:11	Succeeded	AutoResolveIntegrationRuntime (Eas	73828006-e42f-42f9-be29-ad70a6088
Delete2	Delete	2023-04-06T23:51:41.6591391Z	00:00:14	Succeeded	AutoResolveIntegrationRuntime (Eas	fe1a11e7-ba00-4eee-a62e-86e16a0b
ETL_tweet_url	Copy data	2023-04-06T23:51:27.4388087Z	00:00:12	Succeeded	AutoResolveIntegrationRuntime (Eas	103eaadd-6d91-432b-bc9c-832977d
ETL_referenced_tweet_table	Copy data	2023-04-06T23:50:23.2884395Z	00:01:03	Succeeded	AutoResolveIntegrationRuntime (Eas	da9b7664-3e34-45cb-bd9e-db7794d
ETL_twitteruser	Copy data	2023-04-06T23:49:28.5020233Z	00:00:54	Succeeded	AutoResolveIntegrationRuntime (Eas	0d09a5d6-acf5-4ebb-9e1d-92e3071a
ETL_twitterheader	Copy data	2023-04-06T23:48:48.3893549Z	00:00:39	Succeeded	AutoResolveIntegrationRuntime (Eas	5c707acc-f93c-4d3e-afc1-23cd68422
runpythonscript	Custom	2023-04-06T23:48:12.784339Z	00:00:35	Succeeded	AutoResolveIntegrationRuntime (Eas	5bd87875-e620-4b72-a6d8-b3b0e14

Step 10: We performed destination data validation to ensure data flow was smooth and we can now perform analysis.

RDBMS and Graph Tables:

```
select top 2 * from twitterheader;
select top 2 * from twitter_user;
select top 2 * from edge_df;
```

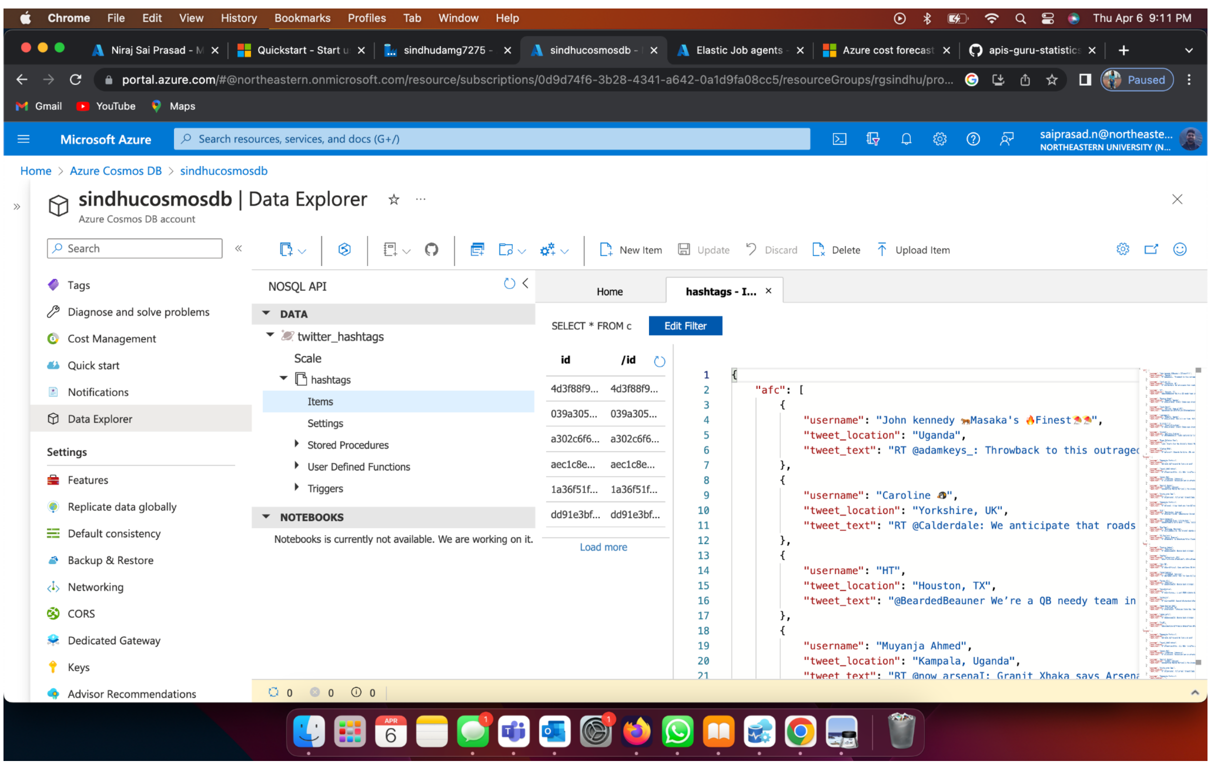
Results Messages

	tweet_id	twitter_handle	parent_tweet_id	tweet	tweet_time	location
1	1644124149864972289	226994736	None	RT @UtdDistrict: Out of #…	2023-04-06 23:45:18+00:00	London, UK↵
2	1644124143787319300	1582299311920058374	None	RT @fortun3e: This Liverp…	2023-04-06 23:45:16+00:00	London, England↵

	twitter_handle	user_name	profile_img_url	description	follower_count	following_count	joined_on
1	944228945515433990	TEE.| ™	http://pbs.twimg.com/prof…	H A T E I S H E A V Y L E…	701.0	1976.0	2017-12-22
2	1467119041567461377	Gunnertalk	http://pbs.twimg.com/prof…	Football is life, life is…	22.0	52.0	2021-12-04

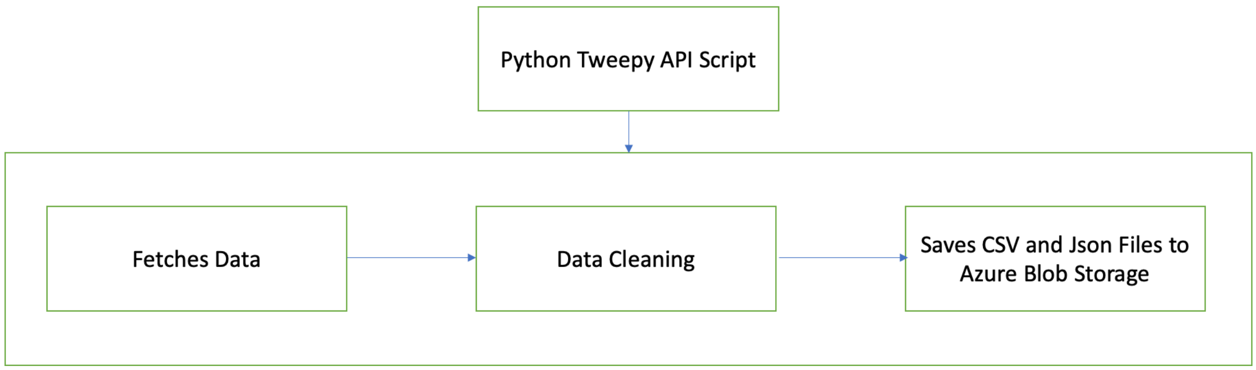
	_from	_to	type
1	nodes/FAYouthCup	nodes/Accra, Ghana	used_in↵
2	nodes/FAYouthCup	nodes/Accra, Ghana	used_in↵

Document Data on Cosmos:

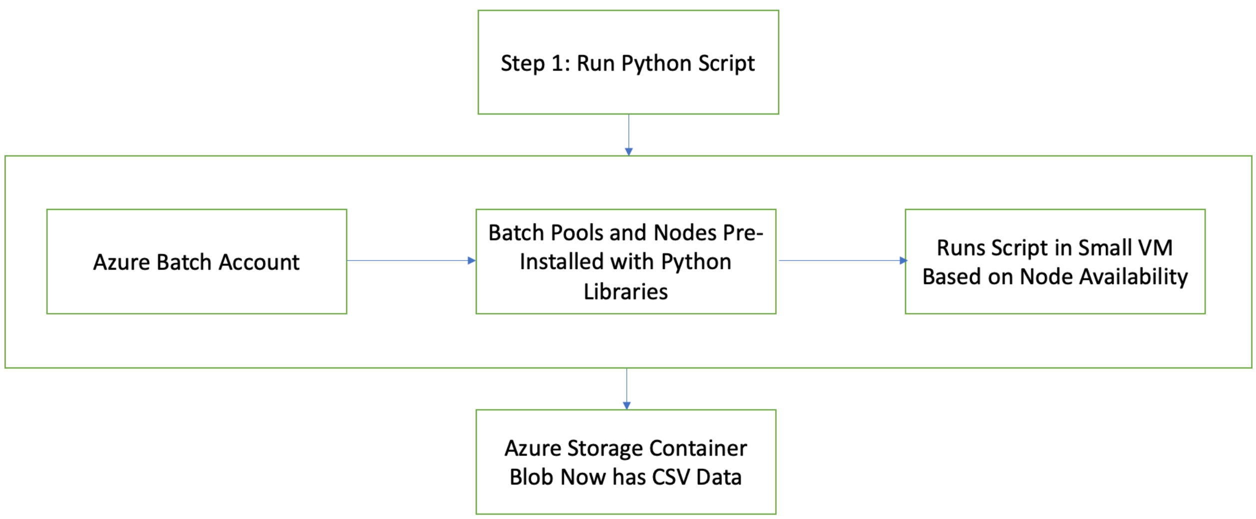


Architecture in Detail:

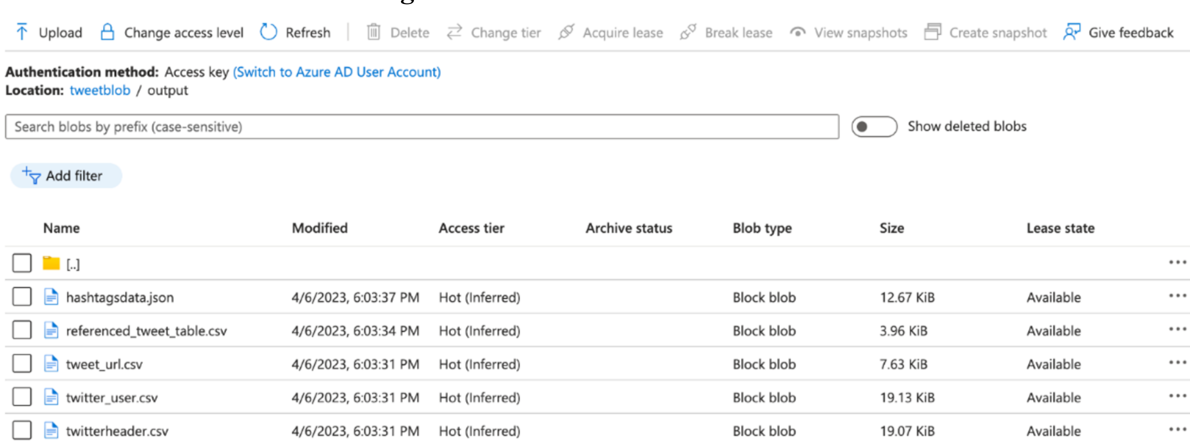
Data Source -



ADF -



CSV and Json files in Blob Storage:



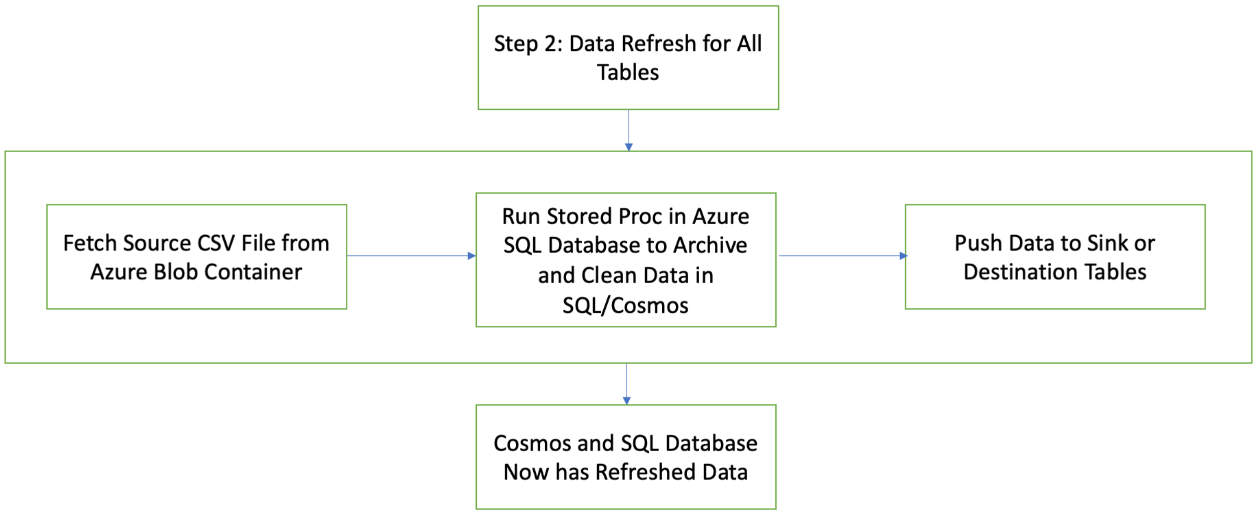
Upload | Change access level | Refresh | Delete | Change tier | Acquire lease | Break lease | View snapshots | Create snapshot | Give feedback

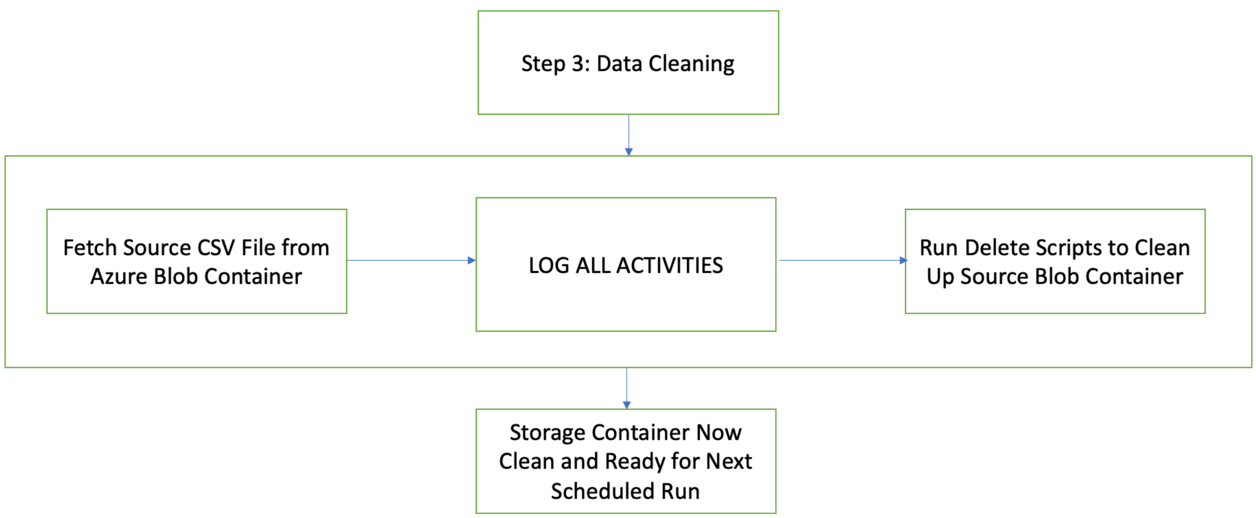
Authentication method: Access key (Switch to Azure AD User Account)
Location: tweetblob / output

Search blobs by prefix (case-sensitive) | Show deleted blobs

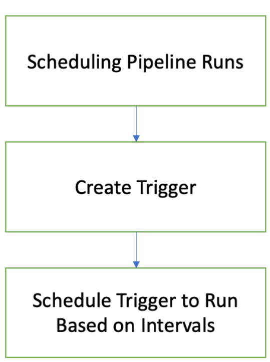
Add filter

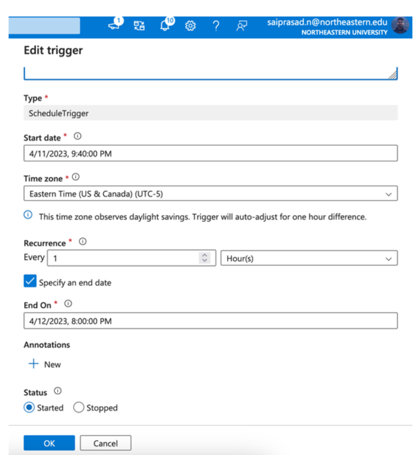
Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
[..]						
hashtagsdata.json	4/6/2023, 6:03:37 PM	Hot (Inferred)		Block blob	12.67 KiB	Available
referenced_tweet_table.csv	4/6/2023, 6:03:34 PM	Hot (Inferred)		Block blob	3.96 KiB	Available
tweet_url.csv	4/6/2023, 6:03:31 PM	Hot (Inferred)		Block blob	7.63 KiB	Available
twitter_user.csv	4/6/2023, 6:03:31 PM	Hot (Inferred)		Block blob	19.13 KiB	Available
twitterheader.csv	4/6/2023, 6:03:31 PM	Hot (Inferred)		Block blob	19.07 KiB	Available


Step 2: Data Refresh for All Tables
Fetch Source CSV File from Azure Blob Container
Run Stored Proc in Azure SQL Database to Archive and Clean Data in SQL/Cosmos
Push Data to Sink or Destination Tables
Cosmos and SQL Database Now has Refreshed Data


Step 3: Data Cleaning
Fetch Source CSV File from Azure Blob Container
LOG ALL ACTIVITIES
Run Delete Scripts to Clean Up Source Blob Container
Storage Container Now Clean and Ready for Next Scheduled Run

Trigger:


Scheduling Pipeline Runs
Create Trigger
Schedule Trigger to Run Based on Intervals


saiprasad.n@northeastern.edu
NORTHEASTERN UNIVERSITY
Edit trigger
Type *
ScheduleTrigger
Start date *
4/11/2023, 9:40:00 PM
Time zone *
Eastern Time (US & Canada) (UTC-5)
This time zone observes daylight savings. Trigger will auto-adjust for one hour difference.
Recurrence *
Every
1
Hour(s)
Specify an end date
End On *
4/12/2023, 8:00:00 PM
Annotations
New
Status
Started
Stopped
OK
Cancel

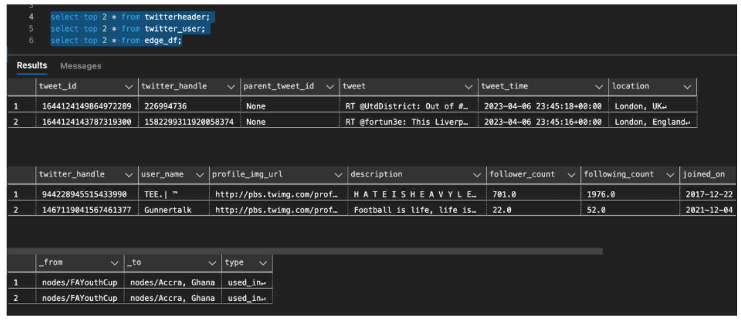
Data Destination -

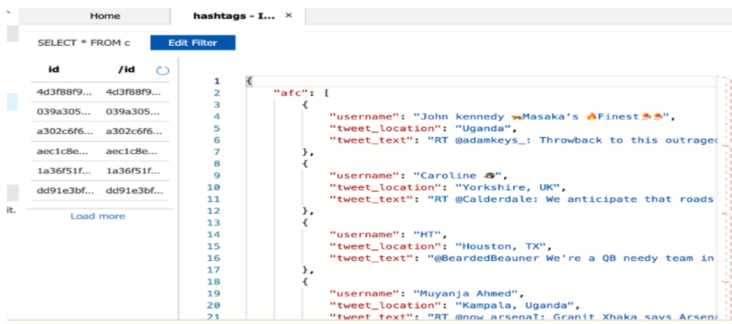
SQL Database:

- Twitter Header
- Twitter User
- Twitter URL
- Reference Tweet Table
- Hashtag Nodes
- Hashtag Edges

Cosmos Database:

- Hashtags Data Json Items



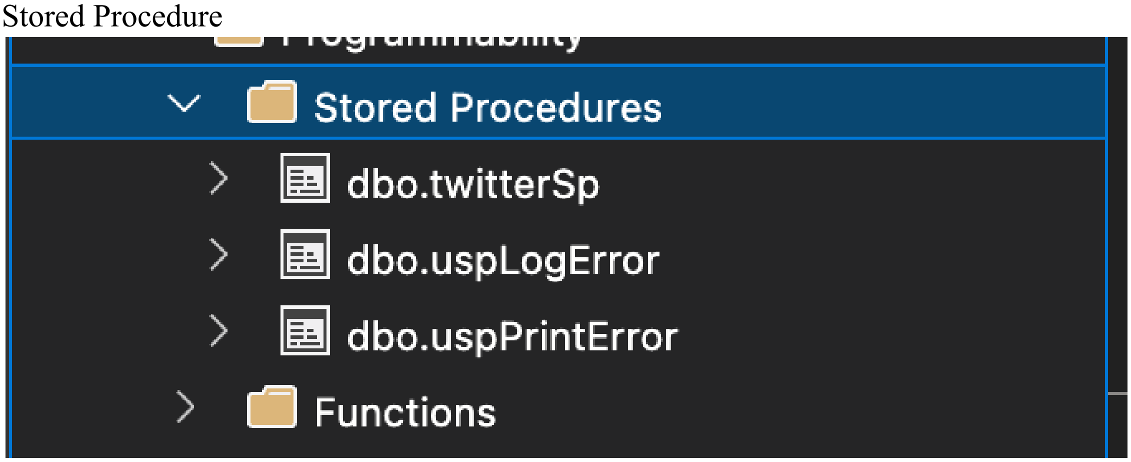


P4: Work with the Database

Functions Performed:

- We created a stored procedure to archive data that is already in the database tables.

Stored Procedure



Archive Tables

```sql
select * from sys.tables where name like '%archive%';
```

Results | Messages

	name	object_id	principal_id	schema_id	parent_object_id	type	type_desc	create_date
1	edgeArchive	1458104235	NULL	1	0	U	USER_TABLE	2023-04-11 19:
2	nodeArchive	1474104292	NULL	1	0	U	USER_TABLE	2023-04-11 19:
3	referencedTweetArchive	1426104121	NULL	1	0	U	USER_TABLE	2023-04-11 19:
4	tweetUrlArchive	1394104007	NULL	1	0	U	USER_TABLE	2023-04-11 19:
5	twitterHeaderArchive	1378103950	NULL	1	0	U	USER_TABLE	2023-04-11 18:
6	twitterUserArchive	1410104064	NULL	1	0	U	USER_TABLE	2023-04-11 19:

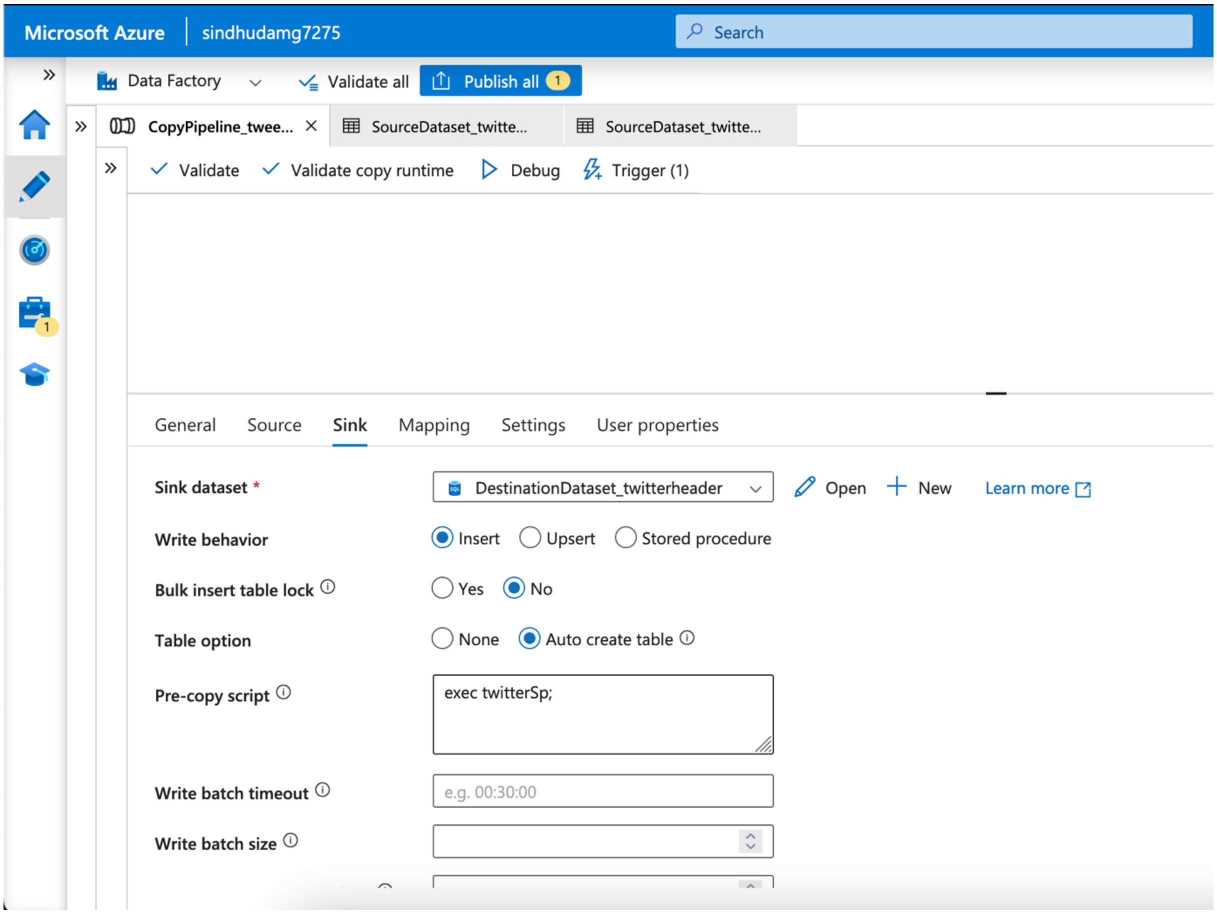
Count of Data in Archive Tables

```sql
select count(tweet_id) as cnt_records from tweetUrlArchive;
```

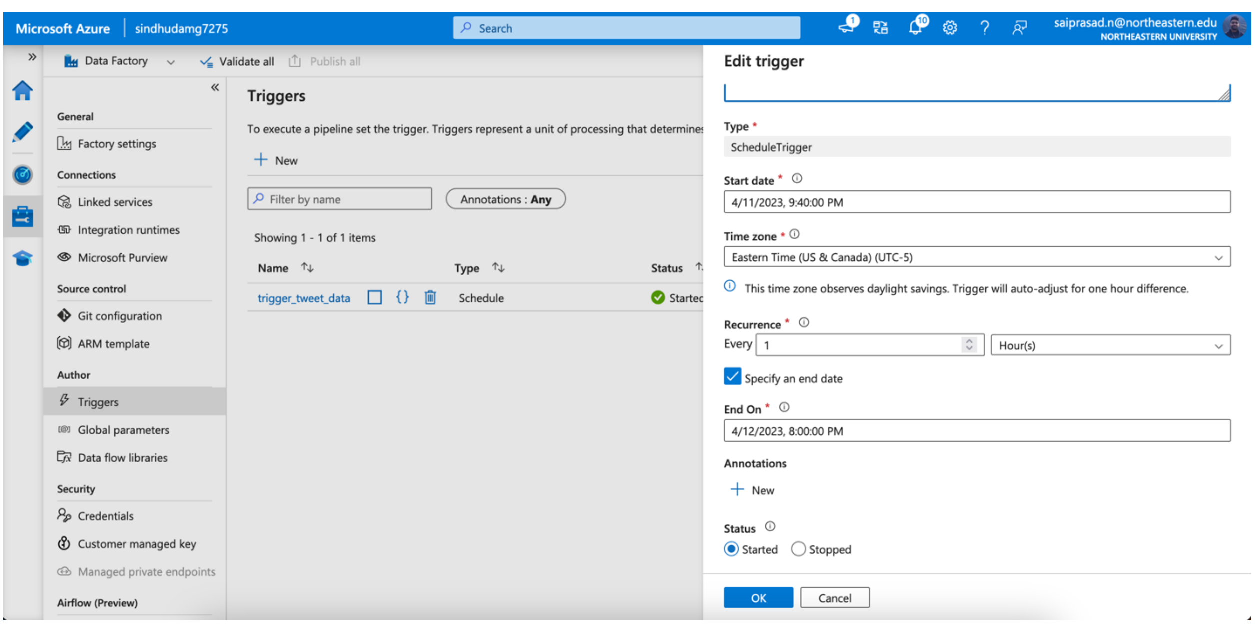
Results | Messages

	cnt_records
1	276

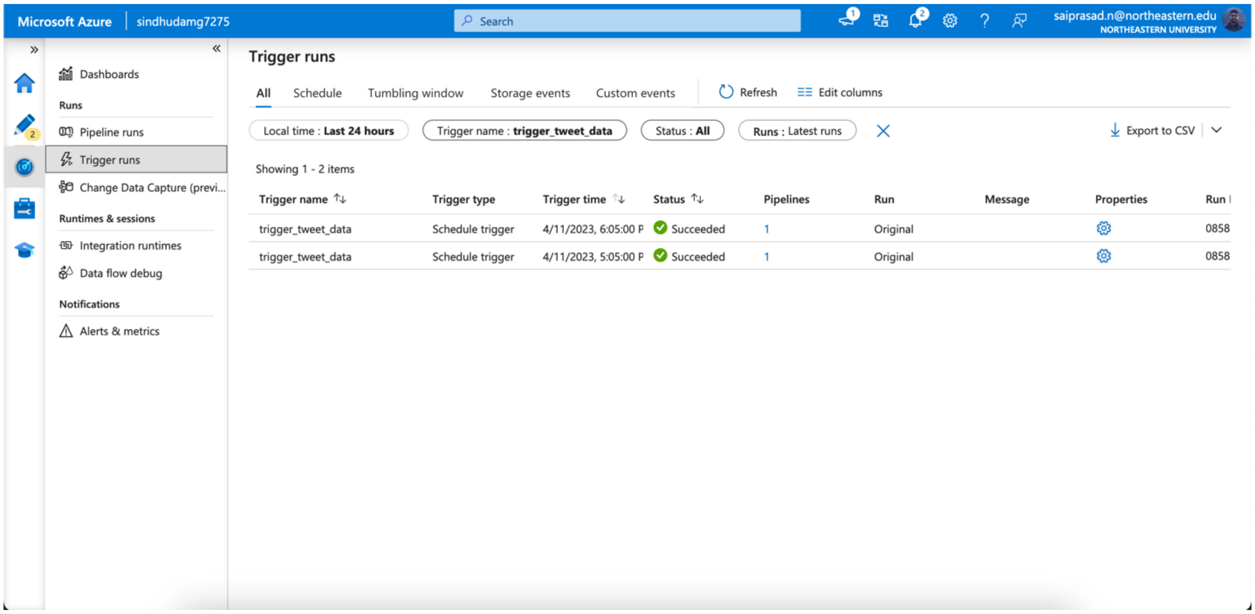
The first step of the pipeline run ensures that the stored procedure above runs:



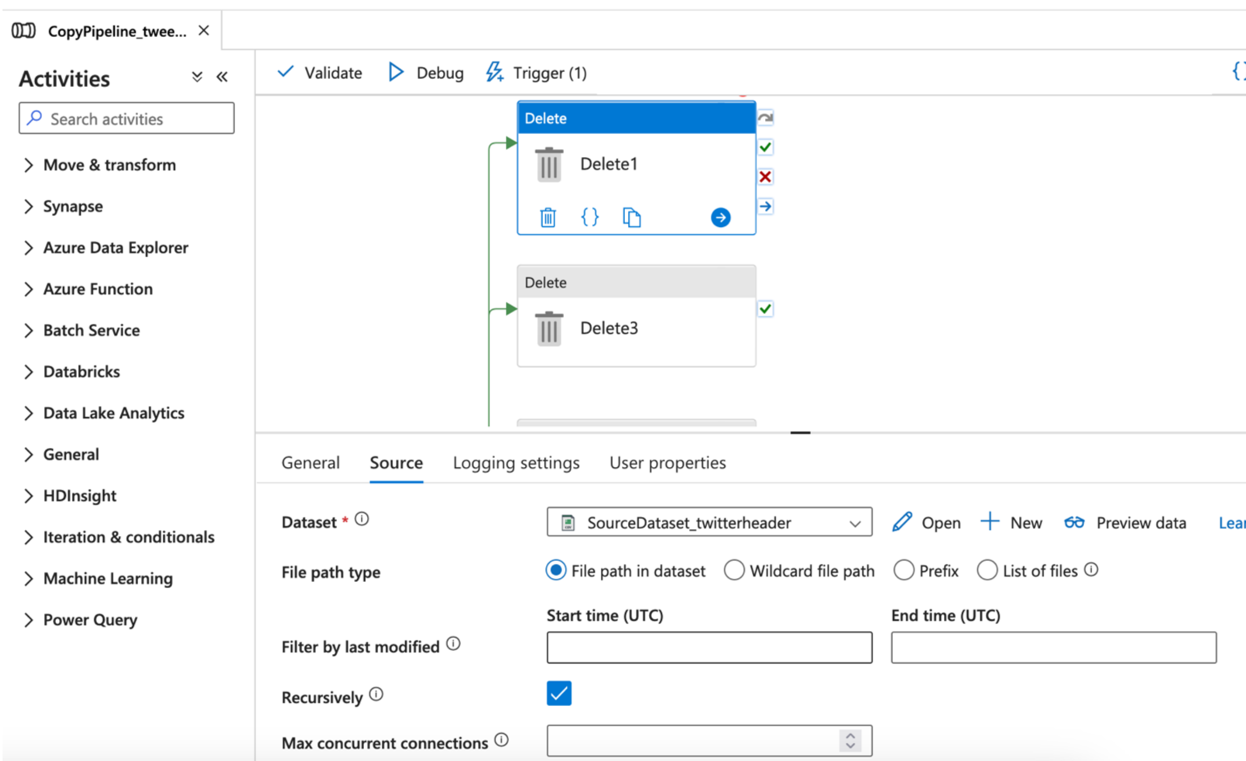
- We scheduled the ADF pipeline to run every hour so that we can run visualization tools over this.



Pipeline Runs:



- We also clean up the Azure blob storage after each run so that it does not clutter the container.

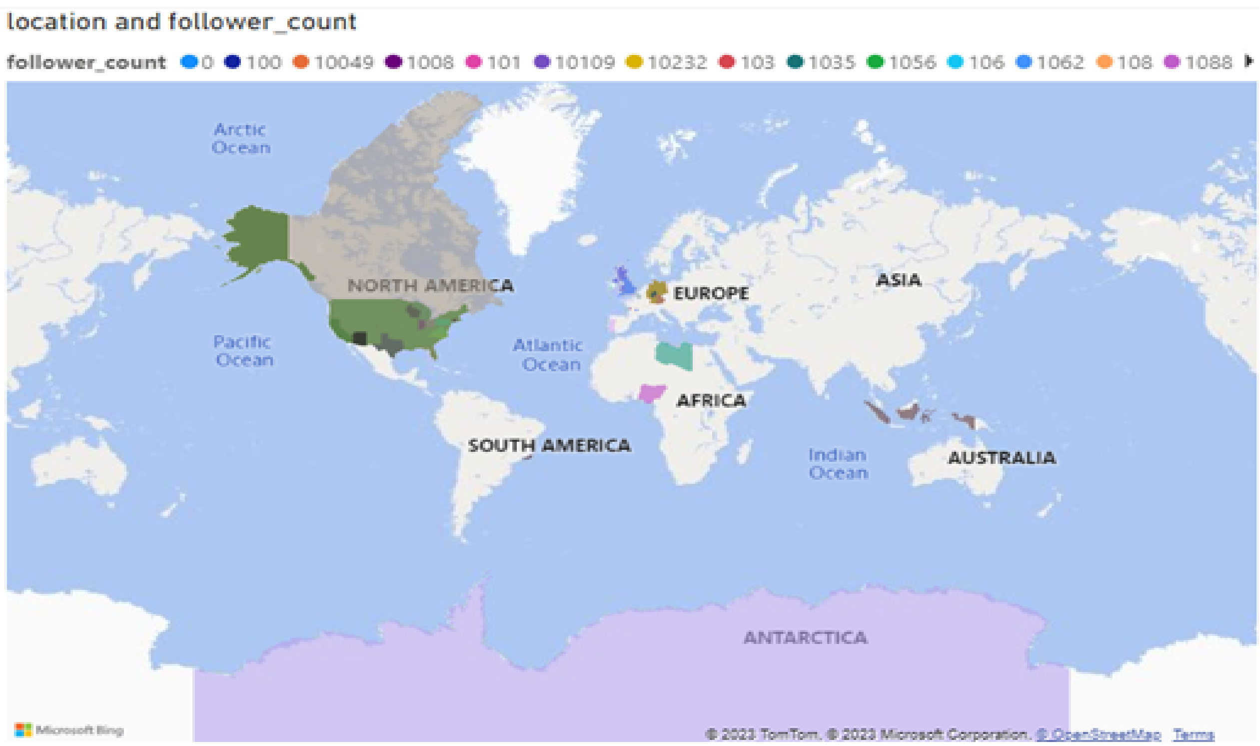


P5:

Visualizations

1. Location-based analysis

We have performed a location-based analysis on Power BI, and have added the location to the map visual, and the number of users as the legend.It provides a visual representation of where your users are located and how many users are in each location.The outcome we got from this analysis is a better understanding of where users are located and how many users are in each location. We are also able to identify patterns or trends in user behavior across different locations, which can help inform marketing and sales strategies. However, the depth of insights we can gain will depend on the complexity of the analysis and the additional layers of data we incorporate into the visualizations.

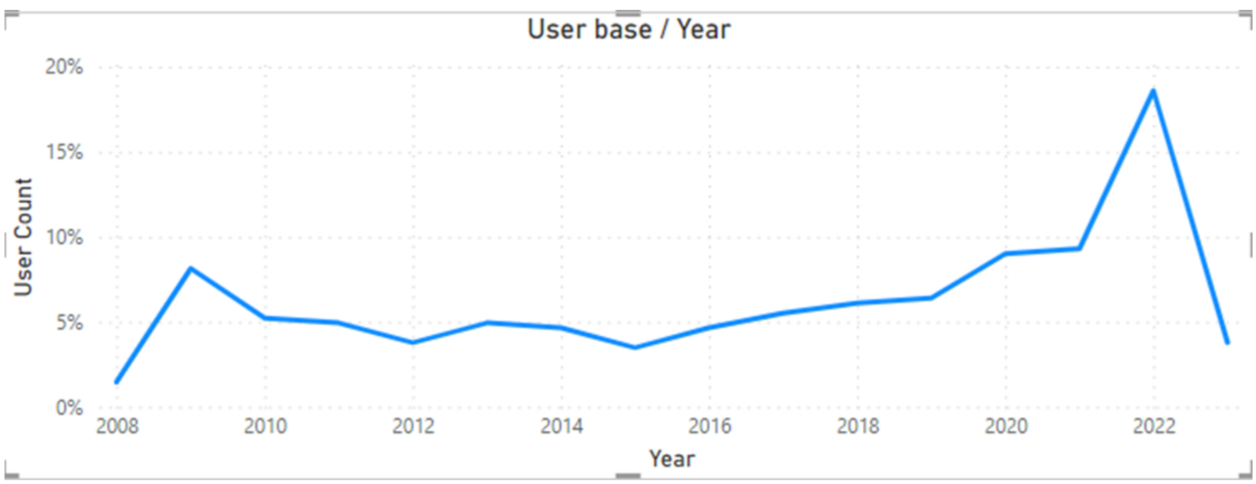


2. User base on Twitter for each year

We have performed a time-series analysis on the user base of Twitter using Power BI. We have plotted the test date on the x-axis and the percentage growth of users on the y-axis. This type of analysis is useful for understanding how the user base of Twitter has grown or changed over time.

The outcome we can expect from this analysis is a better understanding of the trends and patterns in Twitter's user growth. By analyzing the percentage growth of users over time, we can identify periods of rapid growth or decline and the factors that contributed to these changes. This information can be used to inform business decisions, such as marketing strategies, product development, or investment opportunities.

Furthermore, by visualizing the data in Power BI, we can create more compelling and informative presentations for stakeholders. This type of analysis can be a valuable tool for understanding the trajectory of a company or product over time, and making data-driven decisions based on that insight.



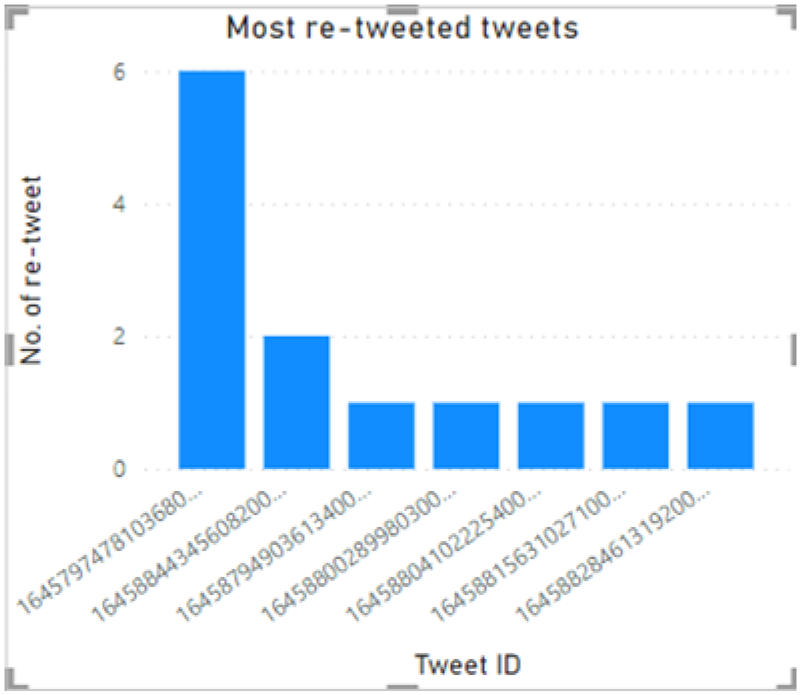
3. Analyze the number of retweets

We have performed an analysis of the most retweeted tweets using Power BI. We have used the parent tweet ID as the x-axis and the count of tweet IDs as the y-axis. This type of analysis can be useful for identifying the most popular or viral tweets on a specific topic or theme.

The outcome you can expect from this analysis is a better understanding of the tweets that have generated the most engagement and reach on Twitter. By analyzing the parent tweet ID and the count of tweet IDs, we can identify the tweets that have been retweeted the most and understand the content or messaging that resonated with users.

This information can be used to inform social media marketing strategies, as it provides insight into the types of content that are most likely to generate engagement and reach on Twitter.

Additionally, by visualizing the data in Power BI,we can create informative and visually appealing presentations to share with stakeholders.

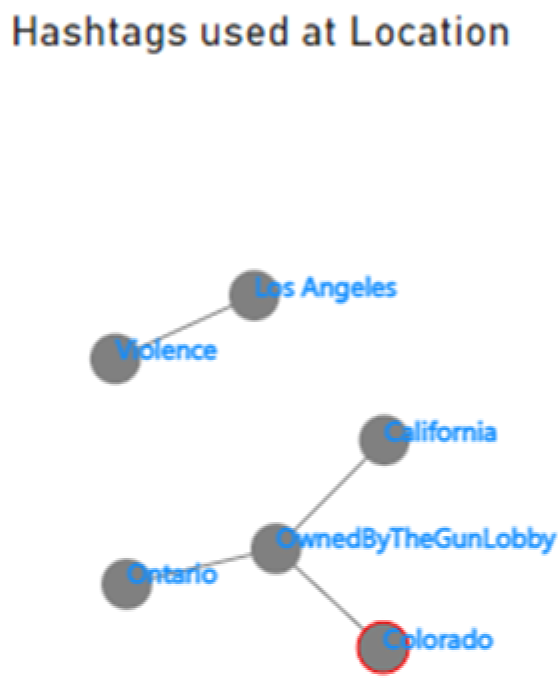


4. Correlation matrix for most popular hashtag

We have performed an analysis on the correlation matrix for the most popular hashtag using Power BI. We have used the network navigator visual and have added "_to" as the Source Node and "_from" as the Target Node. This type of analysis can be useful for identifying patterns and relationships between the most popular hashtag in a region.

The outcome you can expect from this analysis is a better understanding of the network of relationships between the most popular hashtag on the platform. By visualizing the correlation matrix in Power BI, we can identify which hashtags are most closely connected to each other, which users are most influential.

This information can be used to inform social media marketing strategies, as it provides insight into the influencers and key players in the platform's ecosystem. By gaining insight into the hashtags that have the most impact on the platform, we can refine our marketing strategies and better engage with our target audience.



5. Top 10 users with highest following

We have analyzed the top 10 most followed users on Twitter using Power BI. The outcome we can expect from this analysis is a better understanding of the distribution of followers among the top 10 most followed users on Twitter. By visualizing the data in a Treemap, we can easily identify which users have the most followers and compare their popularity to other users.

This information can be used to inform social media marketing strategies, as it provides insight into the most popular users on the platform and the types of content that are resonating with Twitter users. By gaining insight into which users have the most followers, companies can refine their social media marketing strategies and better engage with the targeted audience.


Sum of testfollower_Count by user_name
Spiro's Ghost
Gavin Mario Wax u...
Marko Silberhand
Angie
Baba
Glenn Roth
Spriter
Milu Ramalho
Eddie Smith
Mike Friedman