Big Data Capstone Project

-2-

Presented by:Amjad Al Salem,Amal Al Sarrar,Ghadah Almutairi
Sadeem AlFajih

Objectives

1. Ingest data from AWS RDS and Azure Storage Blob, transforming it using Databricks.
2. Apply Machine Learning to classify post topics and generate top 10 topic reports.
3. Demonstrate Azure's scalable Big Data engineering capabilities.

Data Overflow

stack**overflow**

AWS RDS

azure blob storage

Preview data

Linked service: ls_my_blob

Object:

	id	AcceptedAnswerId	AnswerCount	Body
1	96470	0	0	Thanks to Hallgrim, here is the code I ended up with: <pre> <code>ScreenCapture = System.Windows.Interop.Imaging.CreateBitmapSourceFromHBitmap(bmp.GetHbitmap(), IntPtr.Zero, System.Windows.Int32Rect.Empty, BitmapSizeOptions.FromWidthAndHeight(width, height)); </ code> </pre> I also ended up binding to a BitmapSource instead of a BitmapImage as in my original question
2	99188	0	0	You can use the test command: <pre> <code>test -d $(OBJDIR) || mkdir $(OBJDIR) </code> </pre>

Posts table

Linked service: ls_rds_pg

Object: raw_st.users

	id	age	creationdate	displayname	downvotes	emailhash	location	reputation
1	173274	NULL	2017-09-14T00:00:00	waywardspooky	0	NULL		41
2	173275	NULL	2017-09-14T00:00:00	Phill	0	NULL	San Francisco, CA	30

Users table

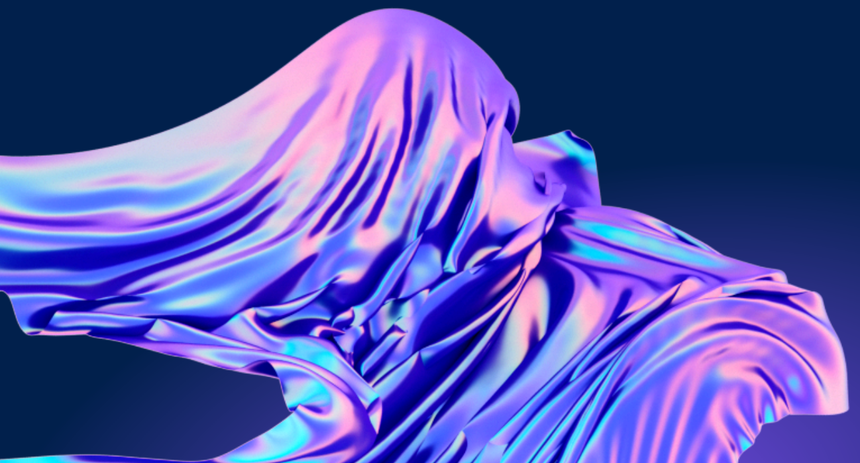
Linked service: ls_rds_pg

Object: raw_st.posttypes

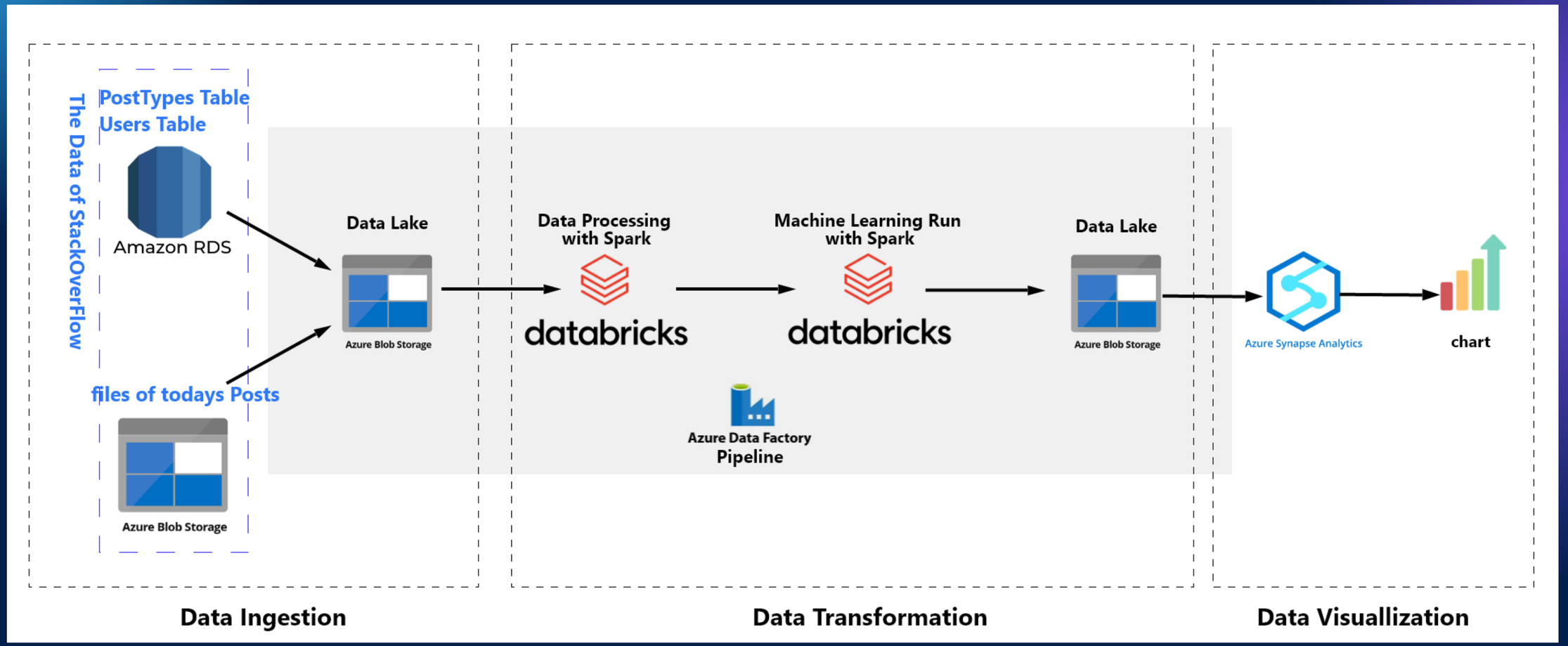
	id	type
1	1	Question
2	2	Answer
3	3	Wiki

PostTypes table

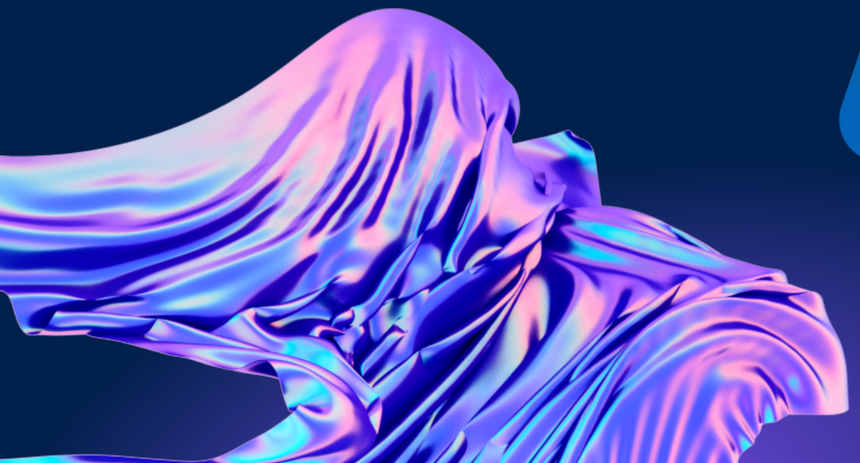
Why big data was used?



Project Architecture



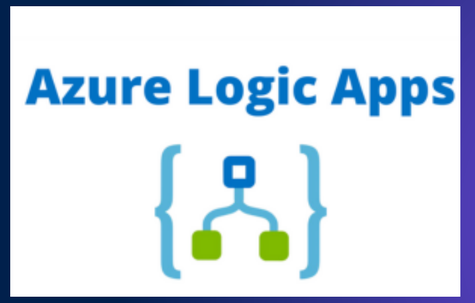
Azure Demo



Azure Services

Cybersecurity best practices

Big Data Handling



Automation

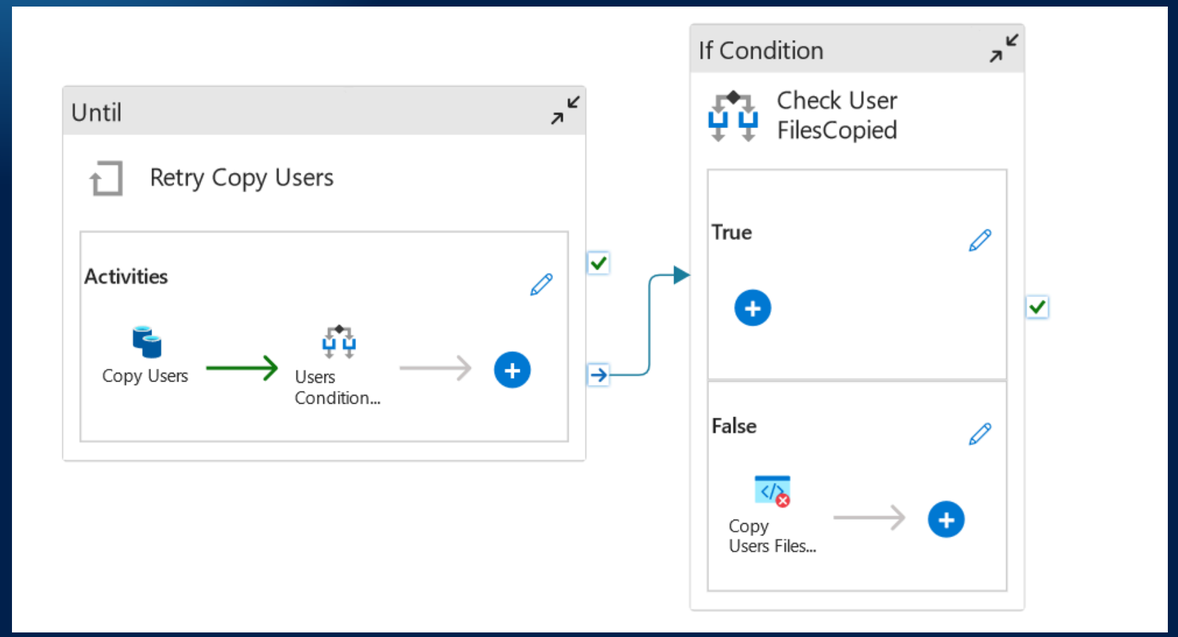
DevOps and CI/CD best Practices

Data Visualization

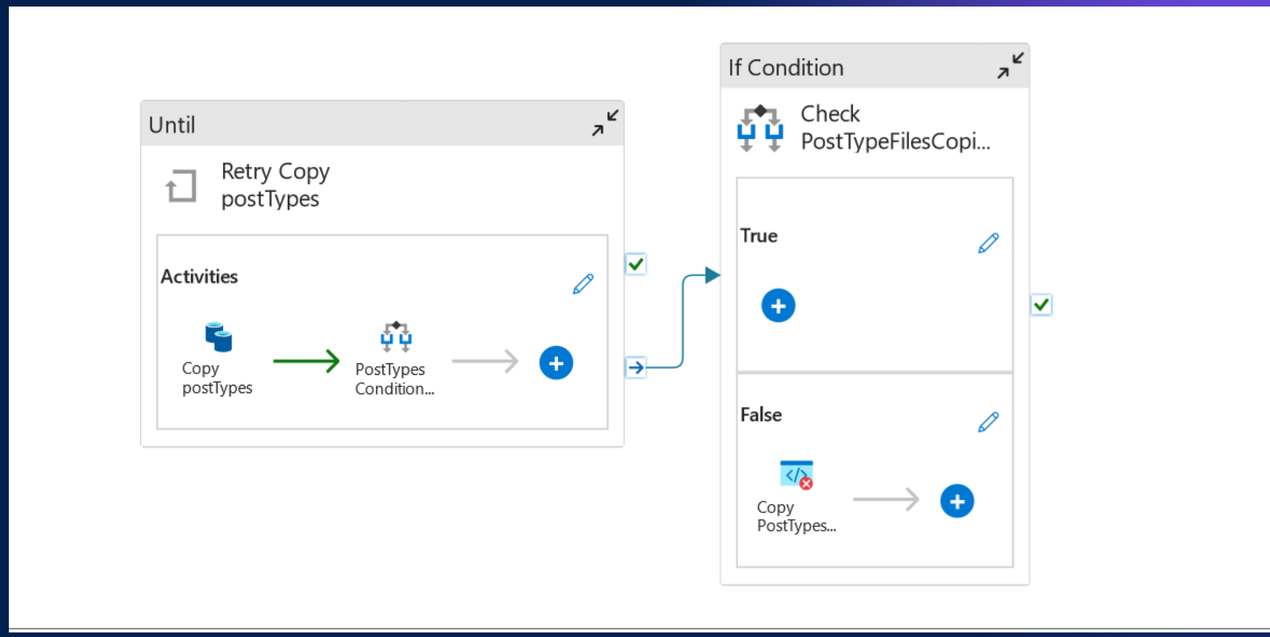
Daily Pipeline

Posts

Weekly Pipeline

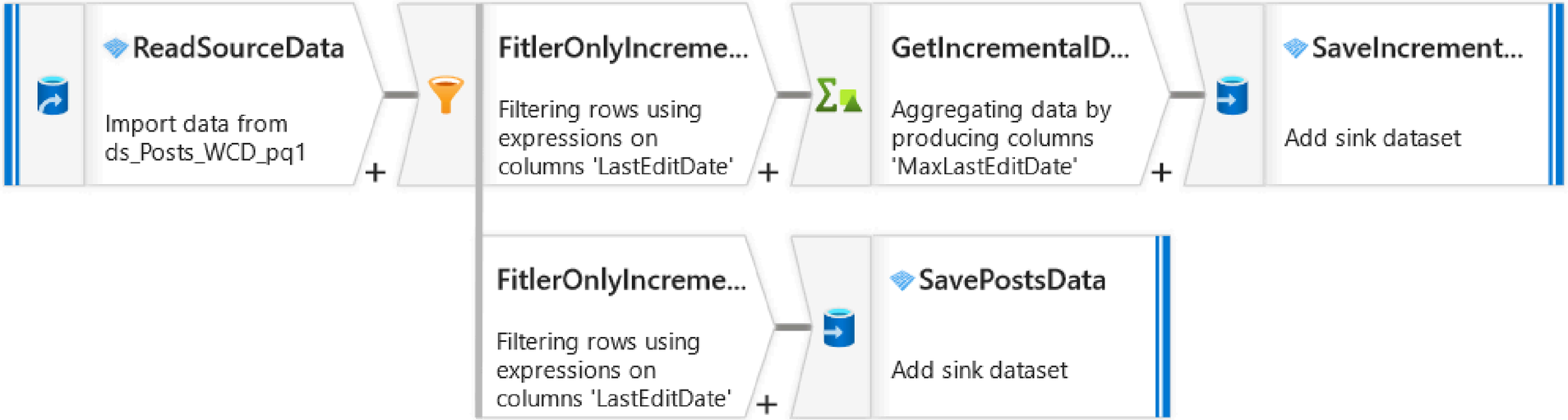


Users

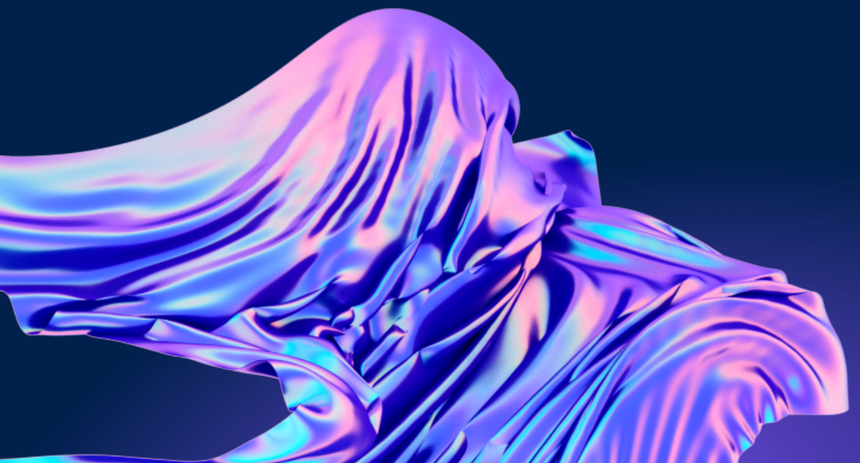


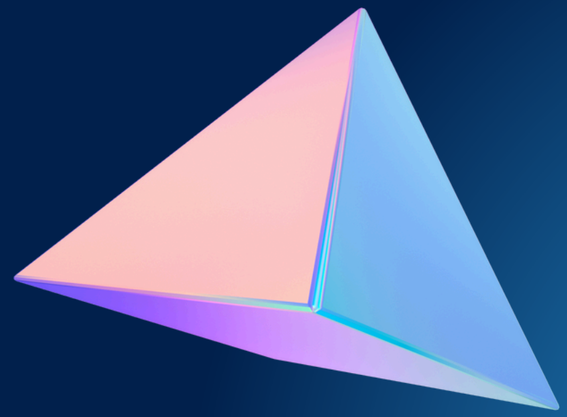
PostTypes

Daily Pipeline Data Flow



How data was processed with Databricks?






Databricks Demo

Databricks

Step1

- We mounted our storage container to the Azure Databricks directory to access our data.

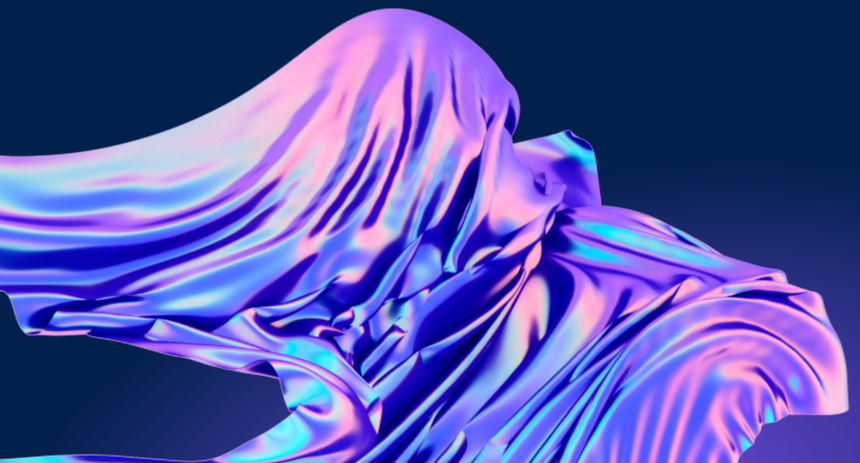
Step2

- We Ran Machine Learning code to train and deploy our model.

Step3

- We Conducted data transformation to prepare our data for analysis.

Synapse Demo



Synapse

01

- Visualized the most discussed topics

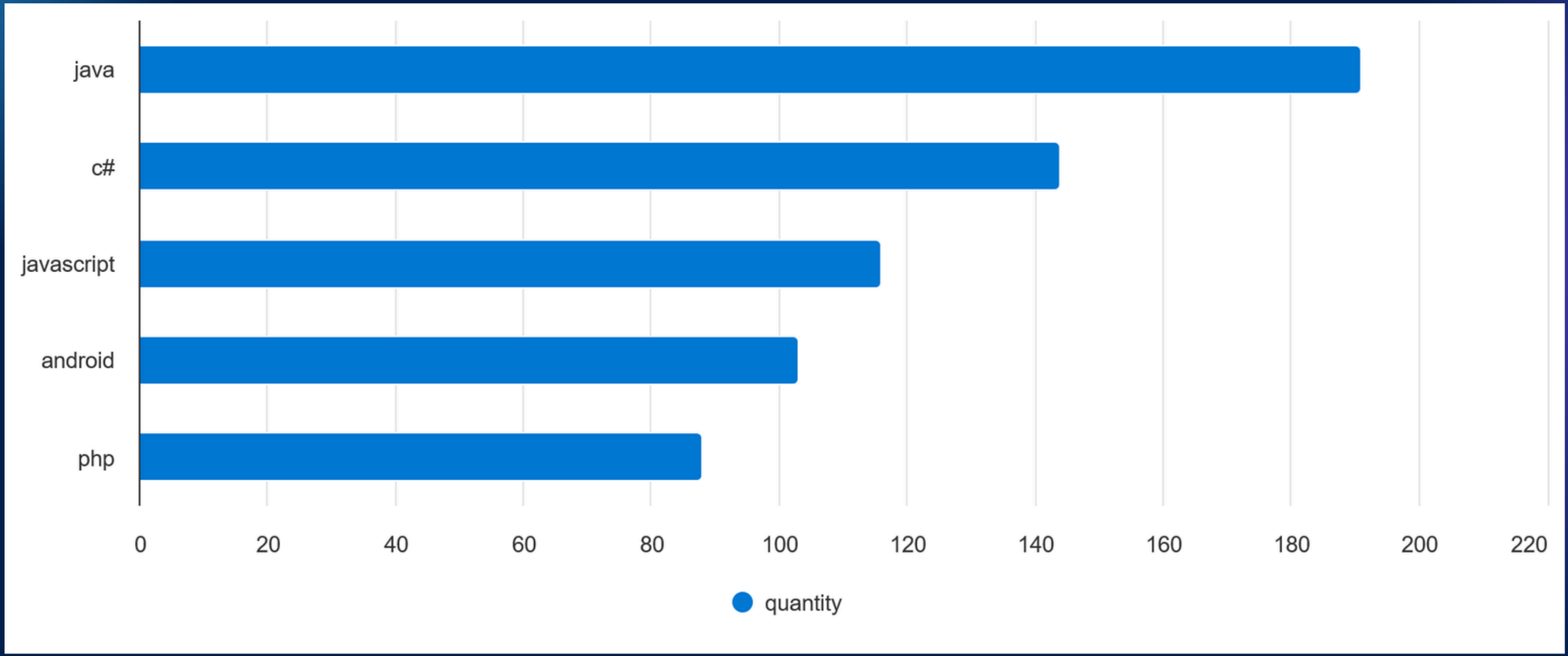
02

- Most & least frequently used tags

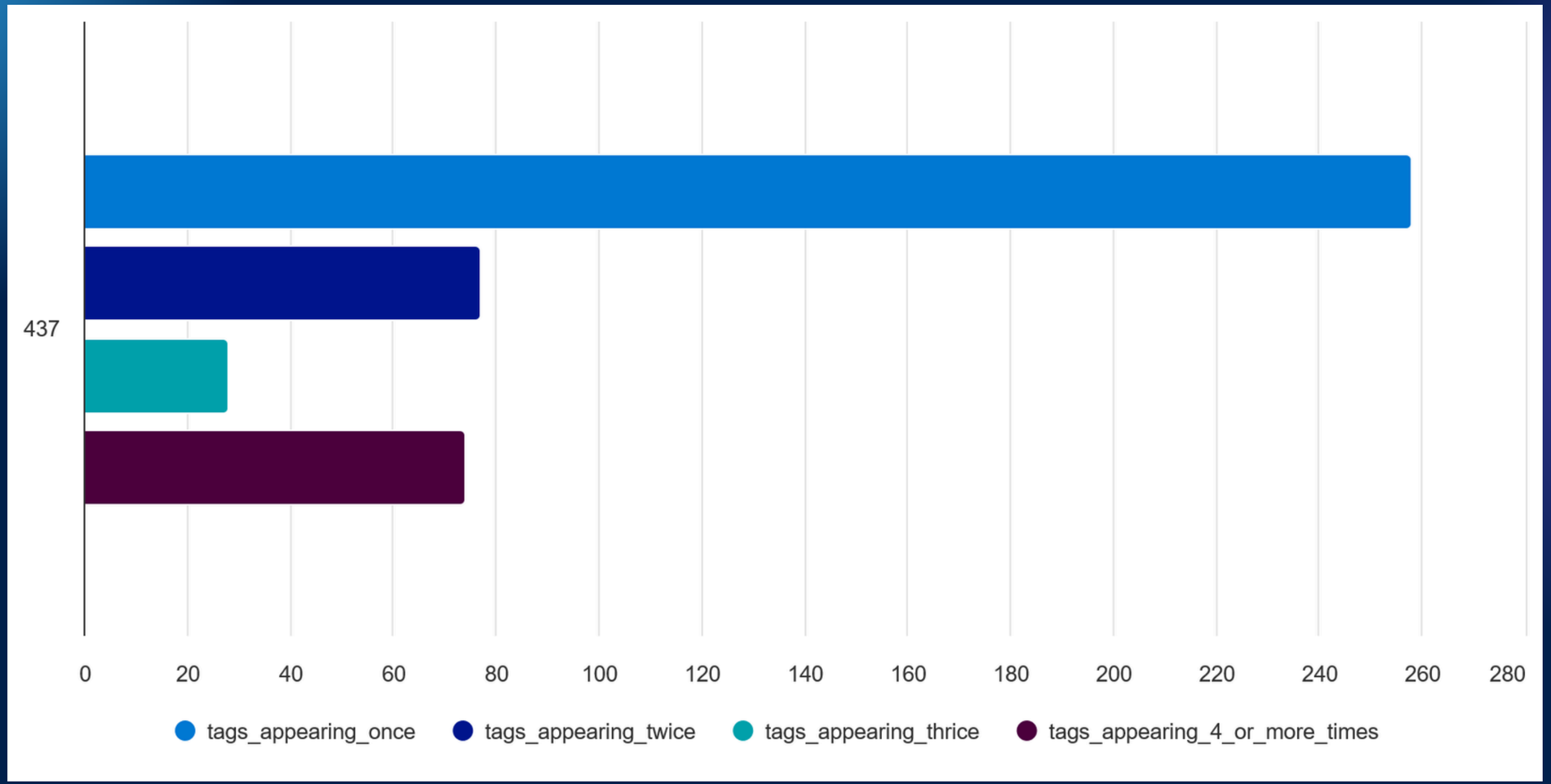
03

- Tags occurrence break down

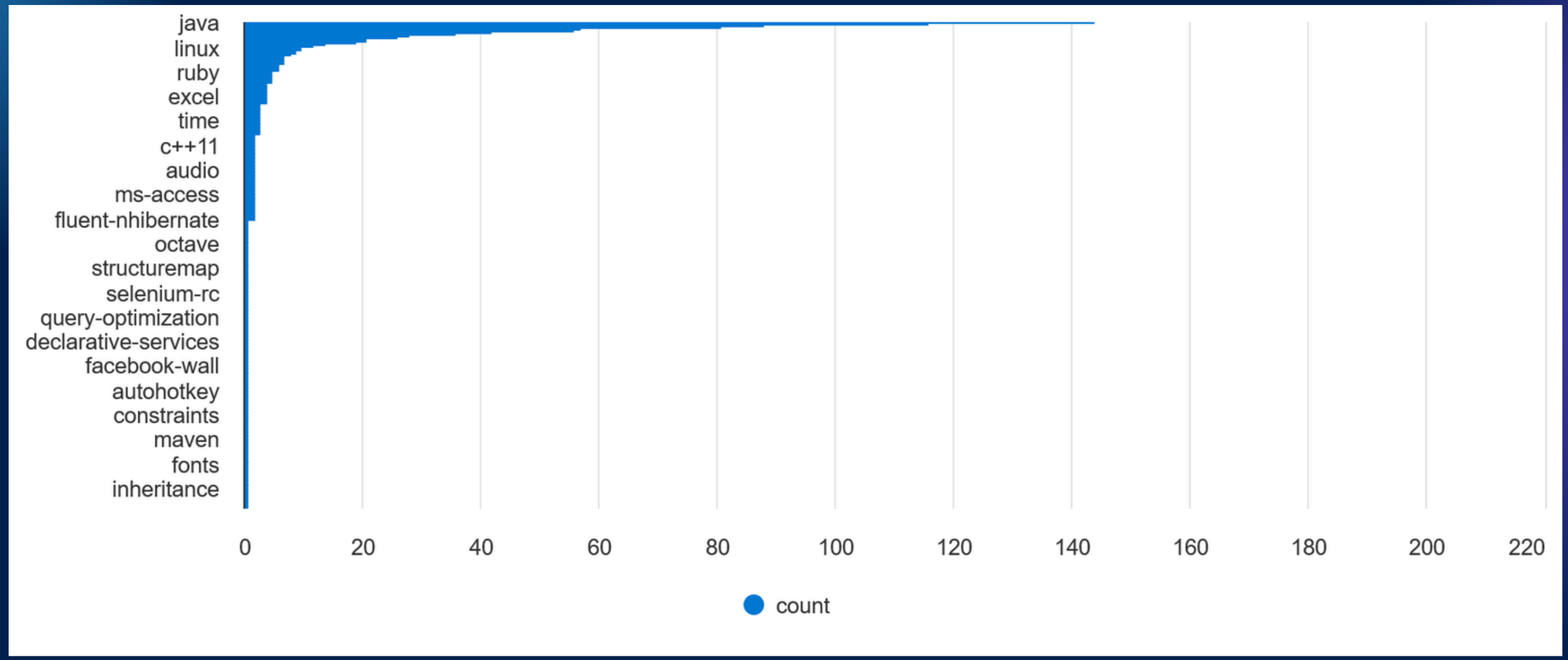
Result for the Top 5 predictions:



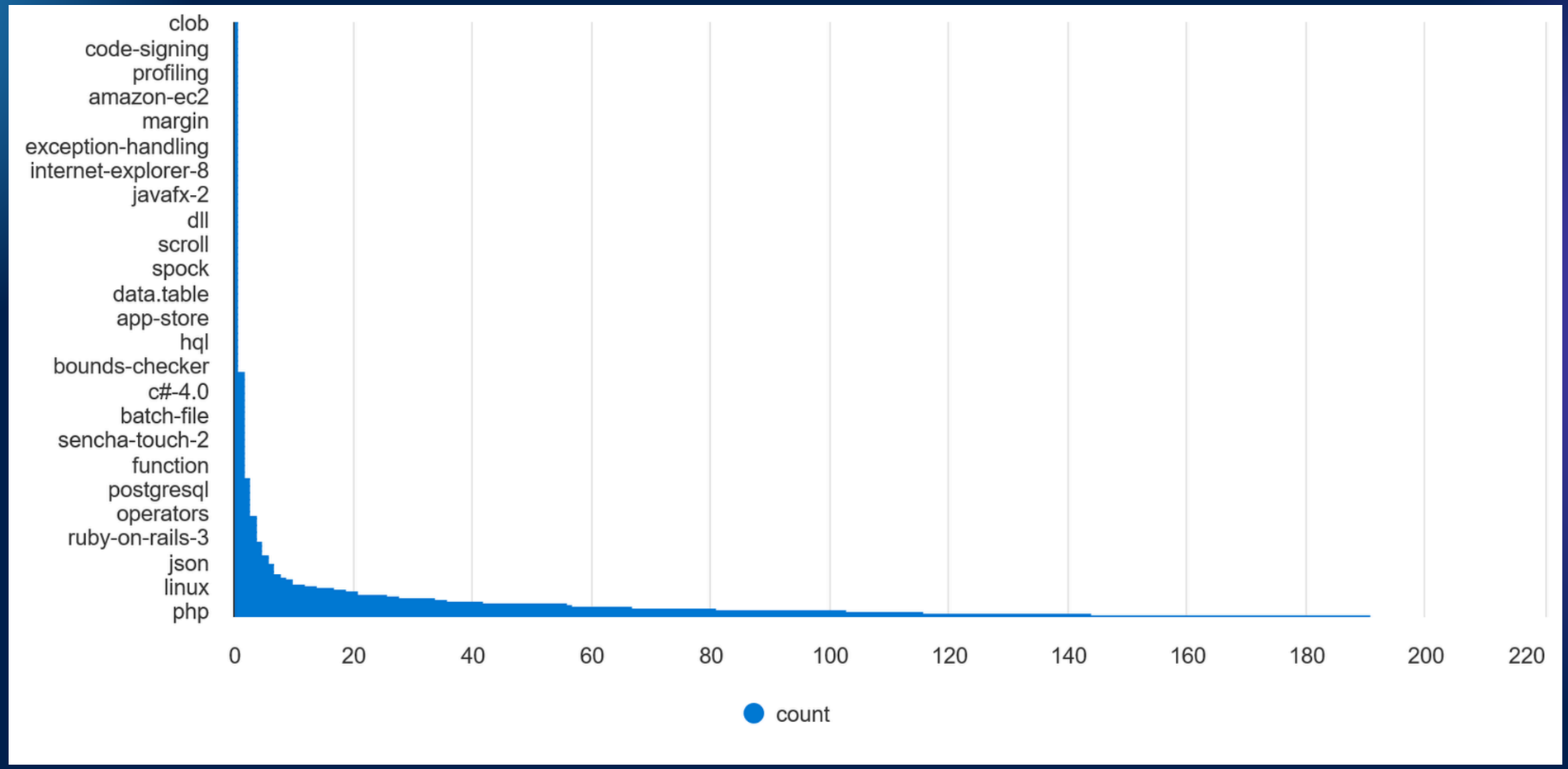
Result for the Tag Occurrence Breakdown:



Result for the Tags Count:

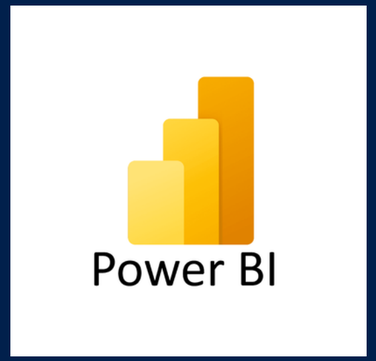


Result for the Least Frequently Used Tags:

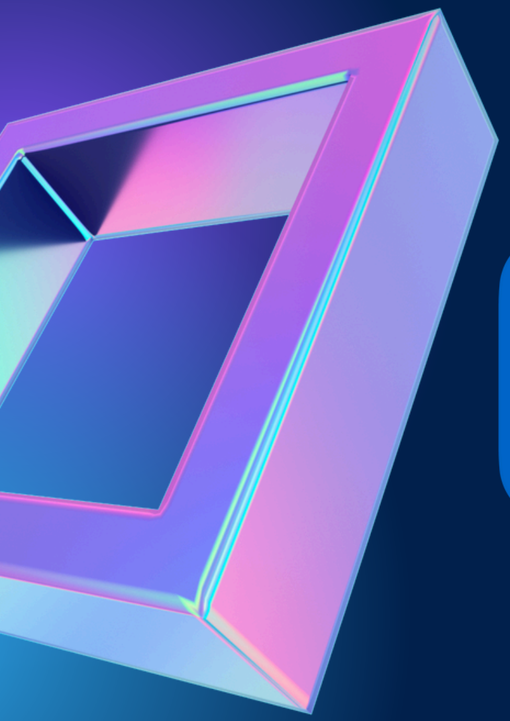


Improvements in Future

- Other visualization tools such as Power BI and Improving our presentation in case of none technical clients



Thank You



Amal Al Sarrar
Amalalsarrar@gmail.com

Amjad Al Salem
Amjadmalsalem@gmail.com

Ghadah Almutairi
Ghadaalmutiri2@gmail.com

Sadeem AlFajih
Eng.Sadeemah@gmail.com