Big Data Engineering Project on Azure

Maram Alshehri
Afnan Alshehri
Munira Alhumaidan

TABLE OF CONTENTS

01 **Project Overview**

02 **About Data**

03 **Data Ingestion**

04 **Data Transformation**

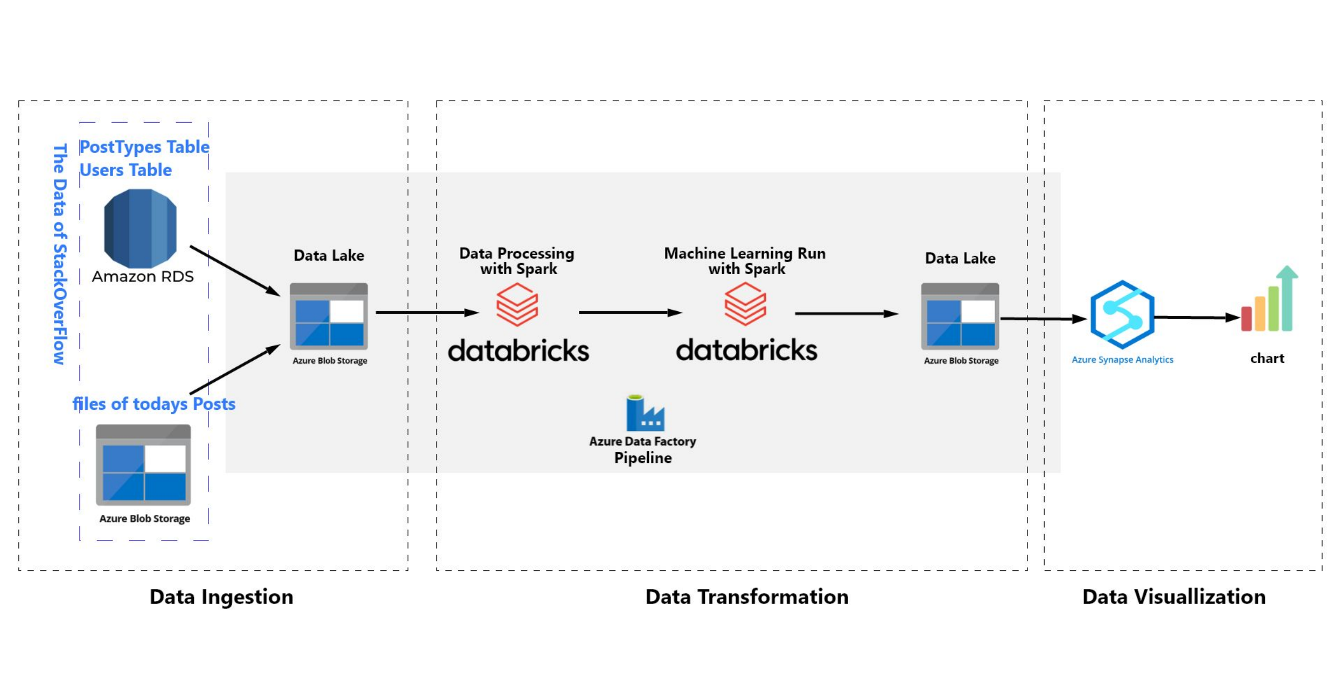
05 **Data Visualization**

06 **Future Work**

01 Project Overview

Project Overview

Our Azure Big Data Project involves ingesting data from two sources into Azure Data Lake, where it undergoes transformation and machine learning. The ML results are reintegrated, and Azure Synapse connects to generate insightful reports, ensuring efficient data handling and informed decision-making.



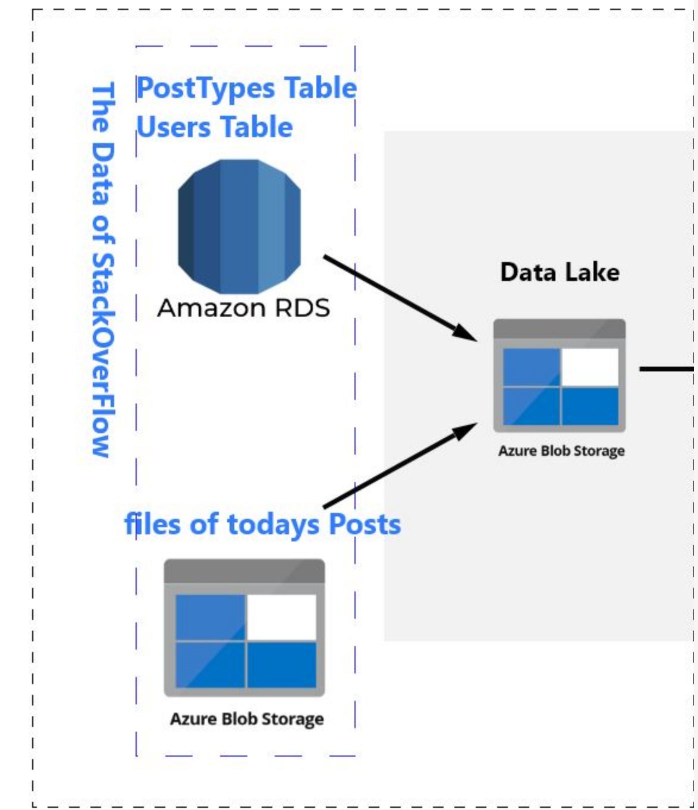
02 About Data

About Data

Data Source: The dataset is sourced from Stack Overflow, containing records of daily online posts, along with information on post types and users.

RDS (Relational Database Service): Users and PostTypes tables are hosted on RDS PostgreSQL database. These tables undergo weekly updates following Slowly Changing Dimension (SCD) type 1 methodology, where only new records are retained, and old records are overwritten.

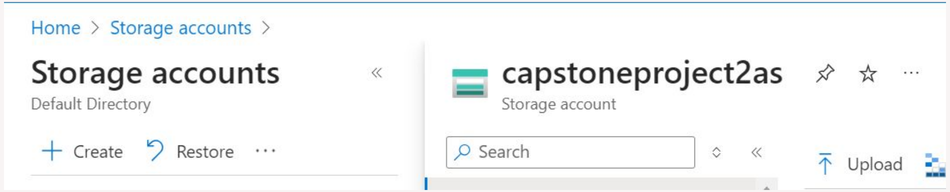
Azure Storage Blob: The daily posts data is stored in Azure Storage Blob in parquet format. Multiple files exist, necessitating the copying of all files into the storage blob for comprehensive data access and analysis.



03 Data Ingestion

1. Data Lake

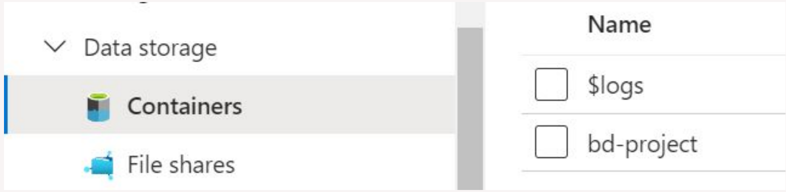
STEP 1: create a storage account



capstoneproject2as

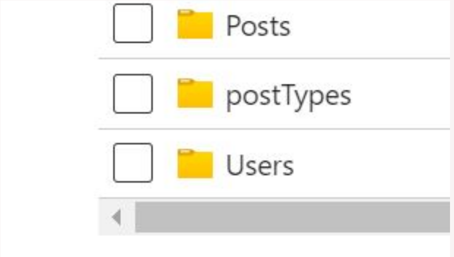
Storage account

STEP 2: Create a container in the storage account that was created



	Name
□	$logs
□	bd-project

STEP 3: create three folders: posts,postType,users



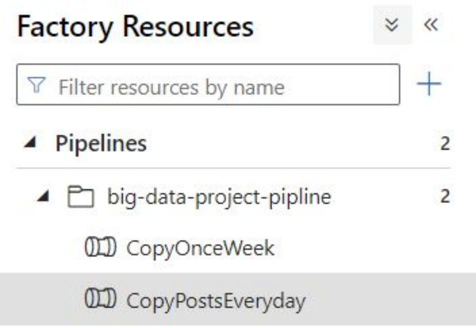
□ Posts

□ postTypes

□ Users

2. Data Factory

STEP 1: Create 2 pipelines copyOnceWeek and copyPostsEveryday

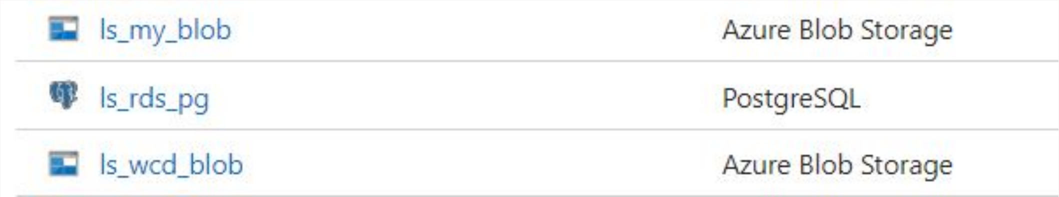


Factory Resources

Filter resources by name

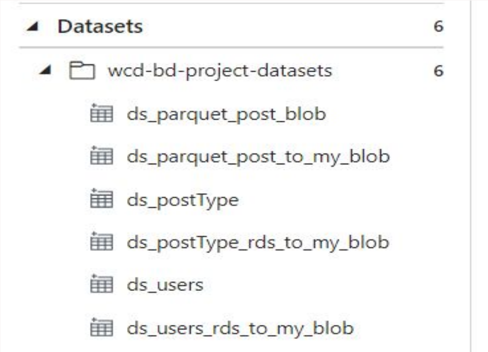
- Pipelines 2
  - big-data-project-pipline 2
    - CopyOnceWeek
    - CopyPostsEveryday

STEP 2: Create 3 linked services



ls_my_blob	Azure Blob Storage
ls_rds_pg	PostgreSQL
ls_wcd_blob	Azure Blob Storage

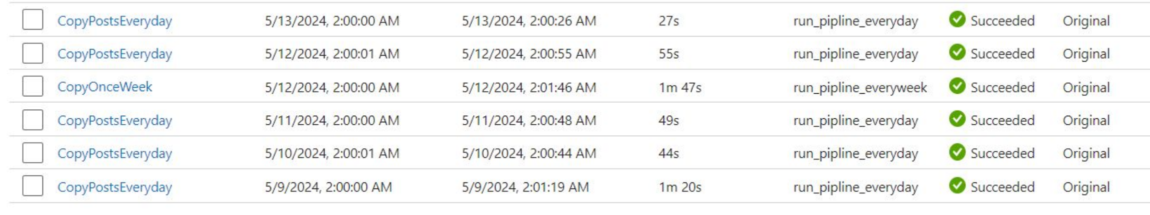
STEP 3: Create 6 datasets



- Datasets 6
  - wcd-bd-project-datasets 6
    - ds_parquet_post_blob
    - ds_parquet_post_to_my_blob
    - ds_postType
    - ds_postType_rds_to_my_blob
    - ds_users
    - ds_users_rds_to_my_blob

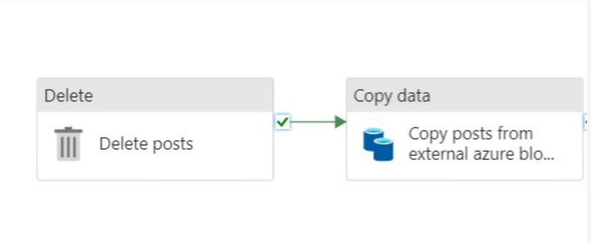
2. Data Factory

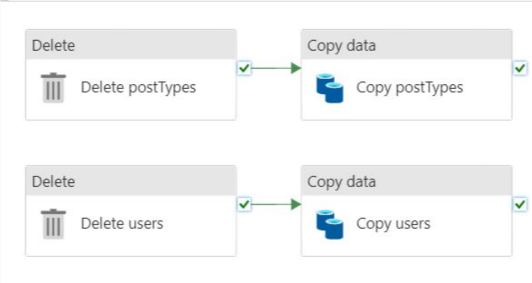
STEP 2: Create trigger for each pipeline



□	CopyPostsEveryday	5/13/2024, 2:00:00 AM	5/13/2024, 2:00:26 AM	27s	run_pipline_everyday	✓ Succeeded	Original
□	CopyPostsEveryday	5/12/2024, 2:00:01 AM	5/12/2024, 2:00:55 AM	55s	run_pipline_everyday	✓ Succeeded	Original
□	CopyOnceWeek	5/12/2024, 2:00:00 AM	5/12/2024, 2:01:46 AM	1m 47s	run_pipline_everyweek	✓ Succeeded	Original
□	CopyPostsEveryday	5/11/2024, 2:00:00 AM	5/11/2024, 2:00:48 AM	49s	run_pipline_everyday	✓ Succeeded	Original
□	CopyPostsEveryday	5/10/2024, 2:00:01 AM	5/10/2024, 2:00:44 AM	44s	run_pipline_everyday	✓ Succeeded	Original
□	CopyPostsEveryday	5/9/2024, 2:00:00 AM	5/9/2024, 2:01:19 AM	1m 20s	run_pipline_everyday	✓ Succeeded	Original

STEP 4: Create copy activity and delete activity for copyPostsEveryday pipeline and 2 copy activity and 2 delete activity for copyOnceWeek pipeline





04

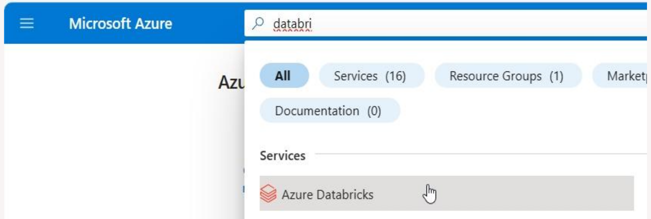
Data Transformation

1. Databricks Mount

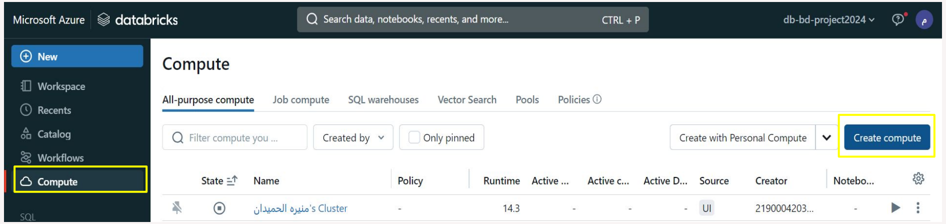
Main Goal: synchronizing the Databricks directory with Azure storage container

Step 1: Set up an Azure Databricks workspace, computing cluster, and notebook.

1.

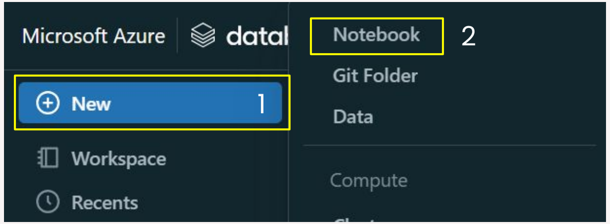


2.



State	Name	Policy	Runtime	Active ...	Active c...	Active D...	Source	Creator	Notebo...
○	منيره الحميدان's Cluster	-	14.3	-	-	-	UI	2190004203...	-

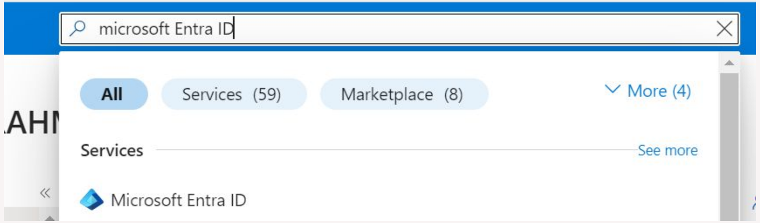
3.



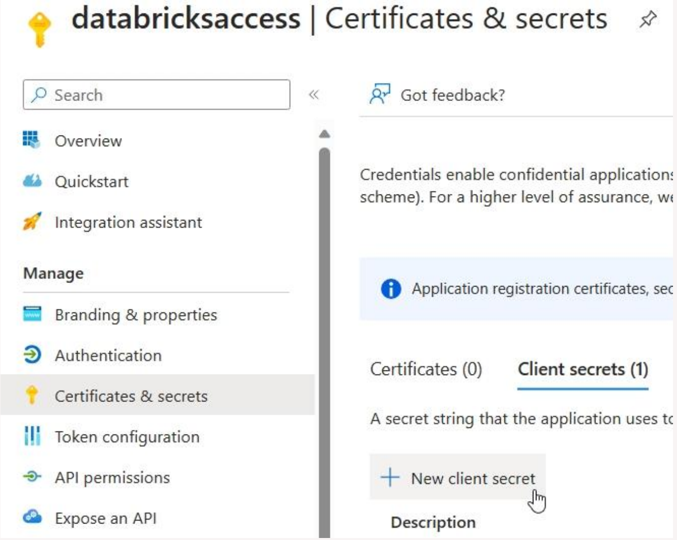
1. Databricks Mount

Step 2: In Azure, authorize Azure Databricks to access your Storage container.

1.



2.

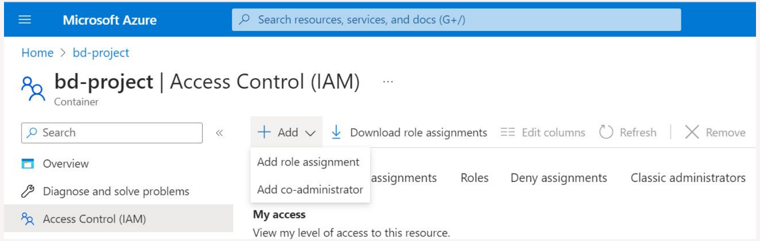


Credentials enable confidential applications scheme). For a higher level of assurance, we

Application registration certificates, sec

A secret string that the application uses to

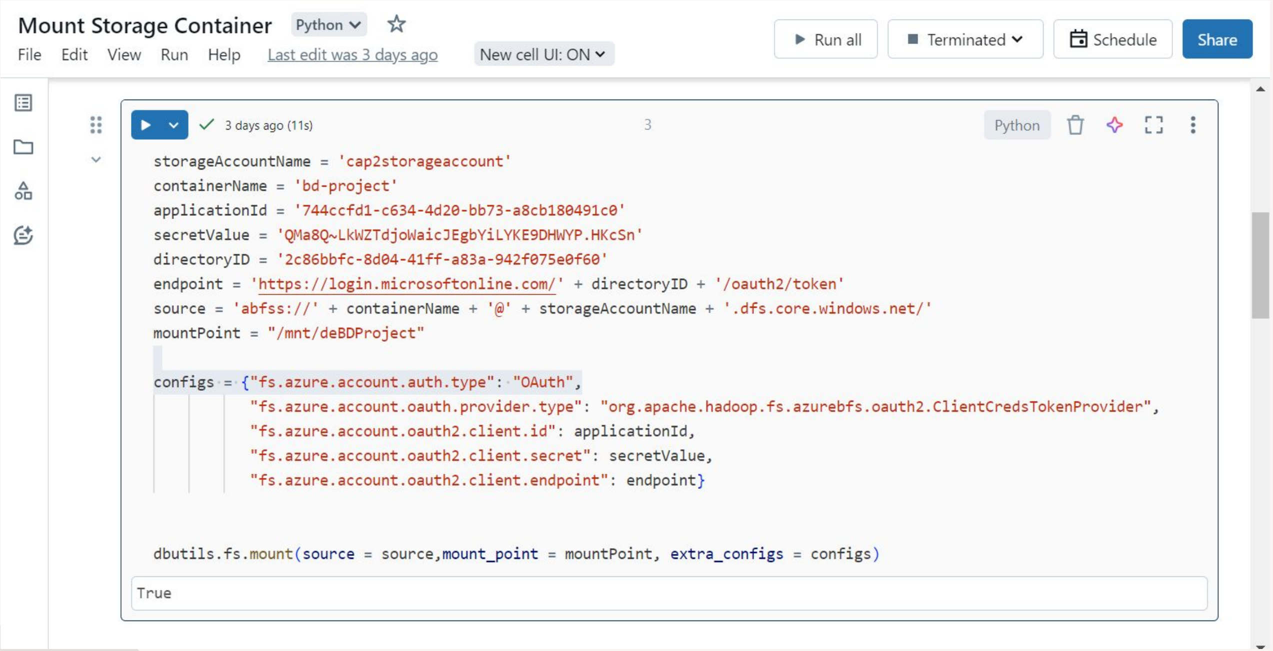
3.



View my level of access to this resource.

1. Databricks Mount

Step 3: Mount your Storage container to the Azure Databricks directory so the Databricks can access it in the same way that it would a local file system.



```
storageAccountName = 'cap2storageaccount'
containerName = 'bd-project'
applicationId = '744ccfd1-c634-4d20-bb73-a8cb180491c0'
secretValue = 'QMa8Q~LkWZTdjoWaicJEgbYiLYKE9DHWYP.HKcSn'
directoryID = '2c86bbfc-8d04-41ff-a83a-942f075e0f60'
endpoint = 'https://login.microsoftonline.com/' + directoryID + '/oauth2/token'
source = 'abfss://' + containerName + '@' + storageAccountName + '.dfs.core.windows.net/'
mountPoint = "/mnt/deBDProject"

configs = {"fs.azure.account.auth.type": "OAuth",
           "fs.azure.account.oauth.provider.type": "org.apache.hadoop.fs.azurebfs.oauth2.ClientCredsTokenProvider",
           "fs.azure.account.oauth2.client.id": applicationId,
           "fs.azure.account.oauth2.client.secret": secretValue,
           "fs.azure.account.oauth2.client.endpoint": endpoint}


dbutils.fs.mount(source = source,mount_point = mountPoint, extra_configs = configs)
```

```
True
```

2. ML Model Training

Step 1: Join the Posts and the Posttypes tables, as we will utilize the Posttypes column in the Posttypes data to filter out the desired data from the Posts table. And then filter the required records.

Step 2: Prepare the data for machine learning training.

Step 3: Train the machine learning model.

Step 4: Save the model to an Azure storage folder so it can be used for future forecasts.

3. Achievement

- Accuracy for each Model

Model	Logistic Regression	Naive Bayes
for classes has count greater than 1	46%	47%
for top 20 tags	81%	74%
for top 10 tags	80%	72%

4. NLP Prediction

4.1 NLP Prediction Notebook

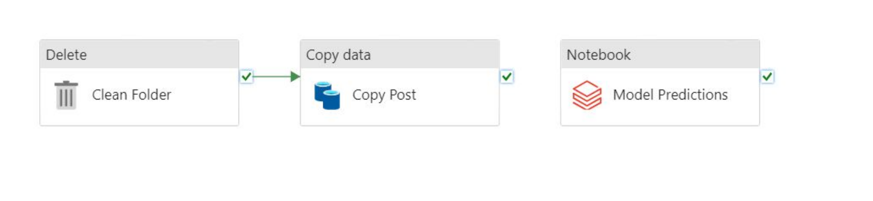
A Databricks notebook will execute the following steps:

- Load the posts data (Posts file) and the trained ML model.
- Define a User Defined Function (UDF) to perform data cleaning and transformation on the post content before feeding it to the model.
- Utilize the UDF to generate topic predictions for each post.
- Summarize the predicted topics and calculate the quantity of each topic.
- Save the resulting topic summary report (CSV file) to a designated Azure storage folder for BI access.

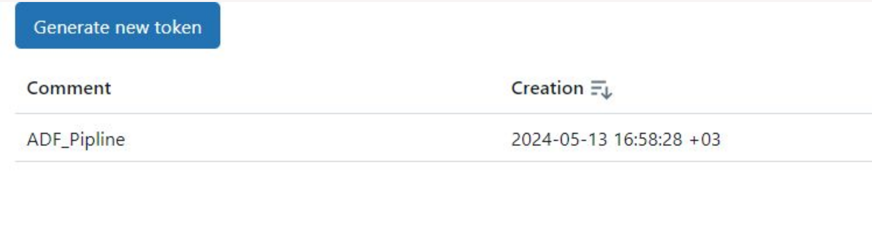
4. NLP Prediction

4.2 Data Factory

STEP 1: Create a Databricks notebook activity.

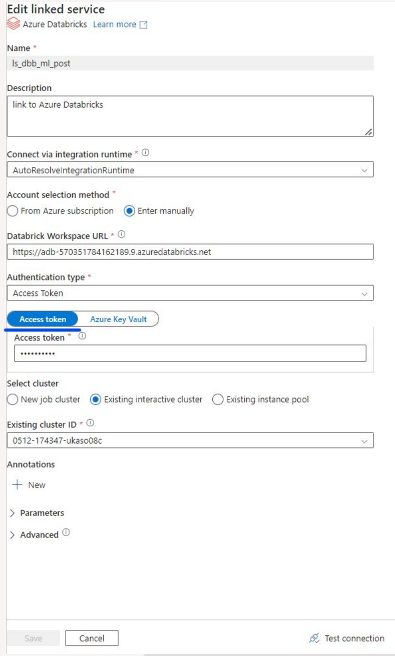


STEP 2: Generate Access token in Databricks workspace.



Comment	Creation
ADF_Pipline	2024-05-13 16:58:28 +03

STEP 3: Create a new Linked Service for connecting to Databricks workspace. Authentication Type: "Access token".



Edit linked service

Azure Databricks

Name *
ls_dbb_ml_post

Description
link to Azure Databricks

Connect via integration runtime *
AutoResolveIntegrationRuntime

Account selection method *
○ From Azure subscription ● Enter manually

Databrick Workspace URL *
https://adb-570351784162189.9.azuredatabricks.net

Authentication type *
Access Token

Access token | Azure Key Vault

Access token *
••••••••••

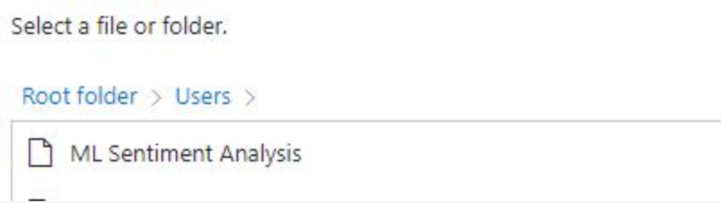
Select cluster
○ New job cluster ● Existing interactive cluster ○ Existing instance pool

Existing cluster ID *
0512-174347-ukaso08c

Annotations

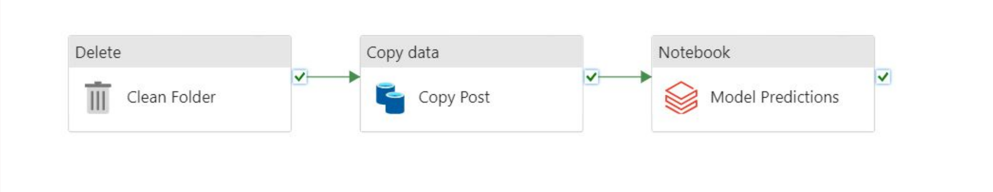
4.2 Data Factory

STEP 4: Complete the Databricks Notebook Activity Configuration



Select a file or folder.

STEP 5: Link Activities



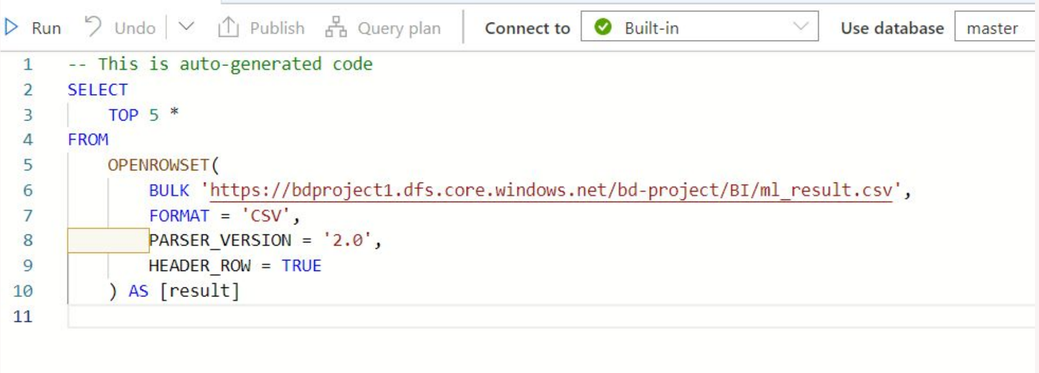
05

Data Visualization

Azure Synapse Analytics

A powerful tool for organizations seeking to unlock the full potential of their data. It offers a unified platform for data warehousing, big data analytics, and machine learning, enabling faster time to insights, improved decision making, and enhanced efficiency.

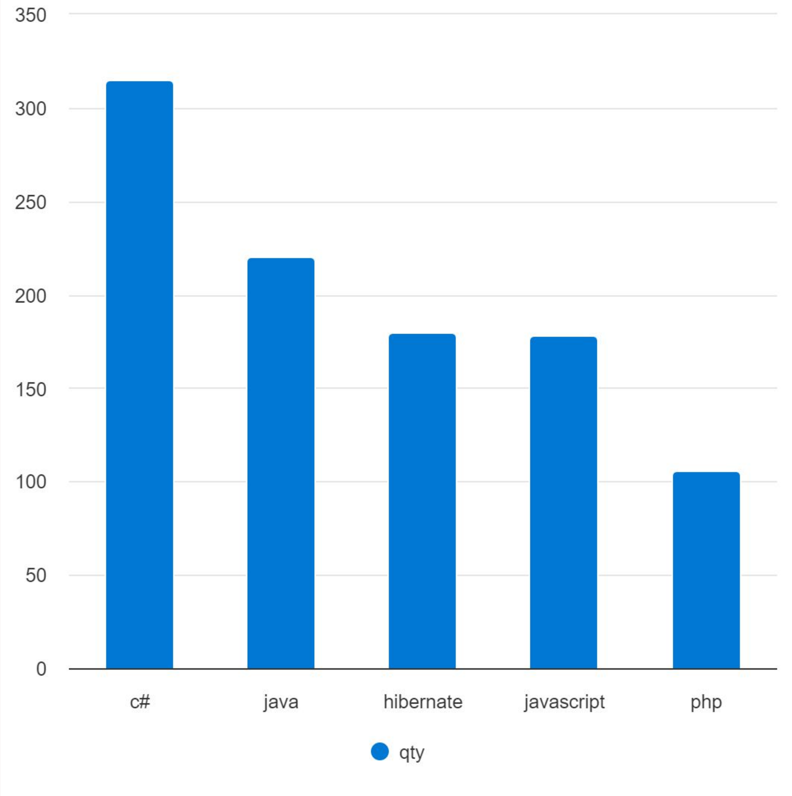
Azure Synapse Analytics



Connect to Built-in

Use database master

```
-- This is auto-generated code
SELECT
    TOP 5 *
FROM
    OPENROWSET(
        BULK 'https://bdproject1.dfs.core.windows.net/bd-project/BI/ml_result.csv',
        FORMAT = 'CSV',
        PARSER_VERSION = '2.0',
        HEADER_ROW = TRUE
    ) AS [result]
```



06 **Future Work**

Future Work

- Join users table with post-type and posts to provide additional features to the machine learning model and can help understand how different groups of users interact with different post types
- Develop and integrate more complex machine learning models in Databricks, possibly exploring deep learning technique for more accurate predictions and insights.
- Extend visualization capabilities by integrating with other BI tools like Power BI or Tableau for more interactive and user-friendly dashboards.

ThankYou
Any Questions!