الأكاديمية السعودية الرقمية
SAUDI DIGITAL ACADEMY

Capstone Project 2

Big Data

Noof Alsubhi
Rahaf Alwaladi
Maryam Alsubhi

Big Data

In today's digital landscape, Big Data is not just a buzzword; it's a pervasive force that influences every aspect of our lives. From the way we interact on social media to the personalized recommendations we receive, Big Data has become an integral part of our daily experiences.

Table of contents

01 Project Overview

02 Data Overview

03 Project Steps

05 Future Work

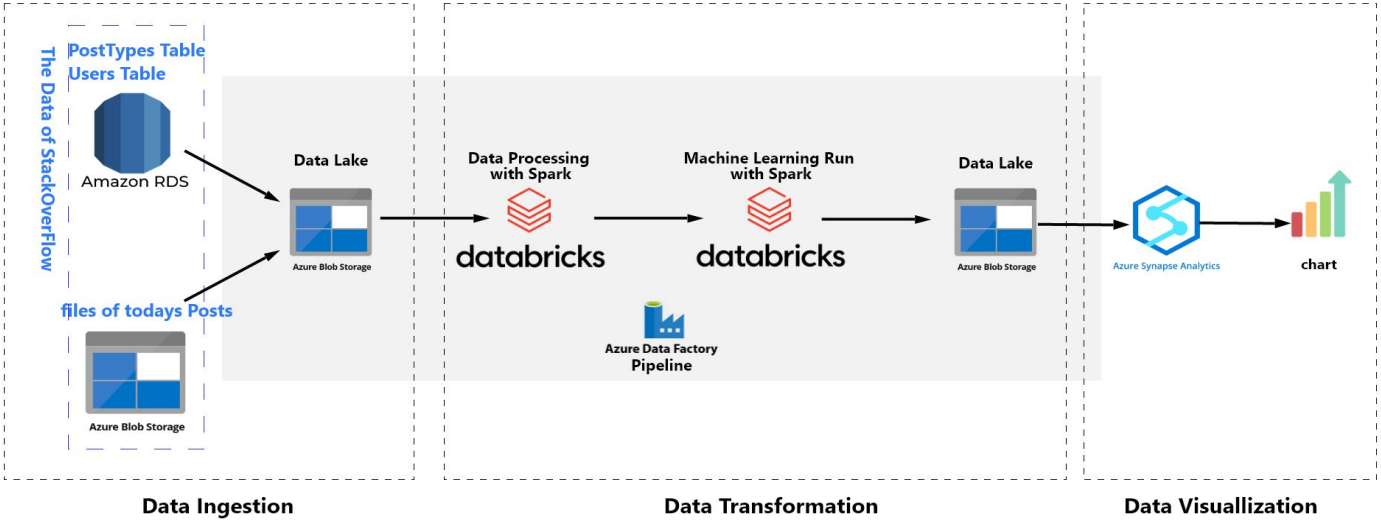
06 Conclusion

Project Overview

Project Goal

The purpose of this project is to build a data processing pipeline on Azure that ingests, transforms, and analyzes data from multiple sources to generate insights about the StackOverflow community.

Project Architecture



Tools

Data factories

databricks

Azure Synapse Analytics

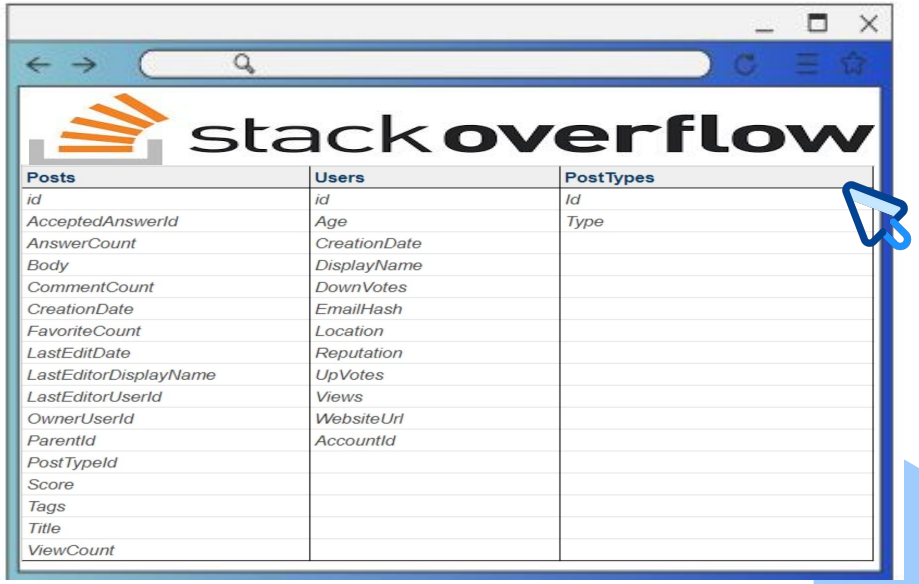
amazon
RDS

Azure Blob Storage

Storage
accounts

Data Overview

This dataset is from StackOverFlow, a popular online IT developer community. It recorded the daily online posts, it also include the posts' type and users' information.



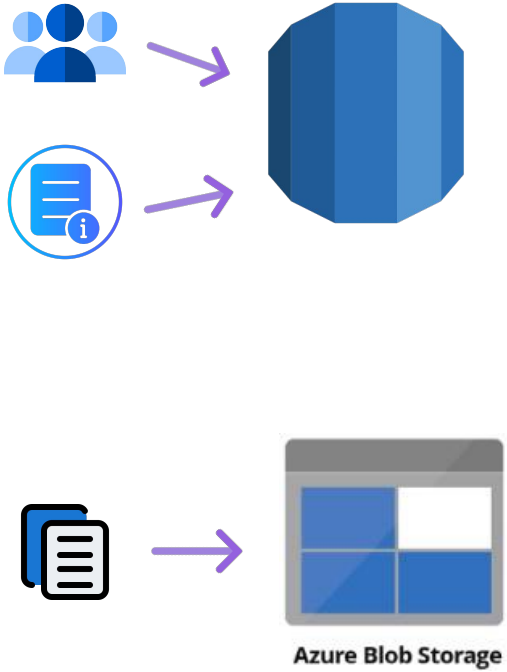
Posts	Users	PostTypes
id	id	Id
AcceptedAnswerId	Age	Type
AnswerCount	CreationDate	
Body	DisplayName	
CommentCount	DownVotes	
CreationDate	EmailHash	
FavoriteCount	Location	
LastEditDate	Reputation	
LastEditorDisplayName	UpVotes	
LastEditorUserId	Views	
OwnerUserId	WebsiteUrl	
ParentId	AccountId	
PostTypeId		
Score		
Tags		
Title		
ViewCount		

Data Overview

The tables are located in two data sources:

- **RDS:** Users and PostTypes tables are stored on RDS postgres. This database are going to be updated once a week

- **Azure Storage Blob:** The Posts data are in parquet format and it is going to be updated daily



Project Steps

Data Ingestion

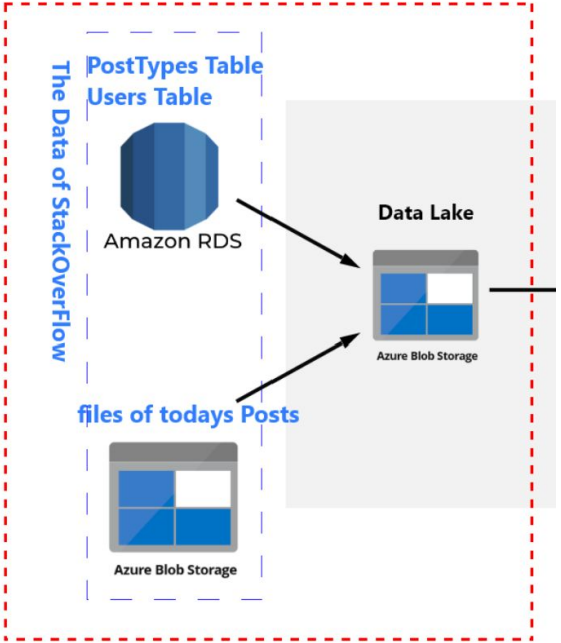
Data Ingestion

Sources:

- Amazon RDS
- Azure Blob Storage

Destination:

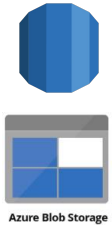
- Data Lake



Data Ingestion

Data Flow In ADF

Data Sources



Linked Services

ls_rds_pg

ls_wcd_blob

Datasets

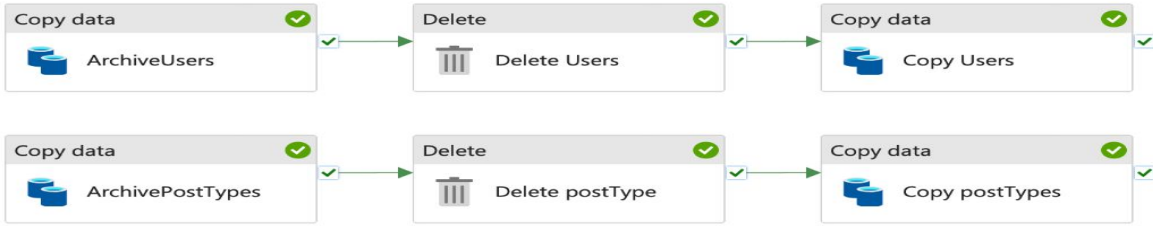
Datasets

Source-dataset 3

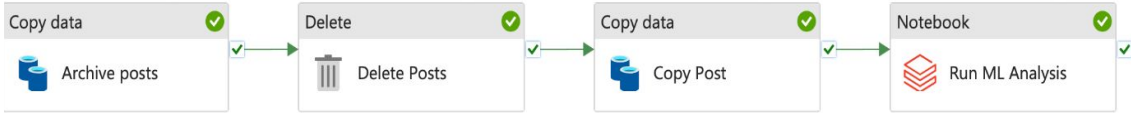
- ds_postType_table
- ds_users_table
- ds_posts_parquetfile

Data Pipelines

1. copyOnceWeek



2. copyPostsEveryday



Data Destination

ls_my_blob

Datasets

Datasets 9

Destinaton-dataset 6

- ds_Archive_posts
- ds_Archive_postTypes
- ds_Archive_Users
- ds_users_inblob
- ds_postType_inblob
- ds_posts_inblob

Containers

Authentication method: Access key (Switch to Microsoft Entra user account)
Location: bd-project / landing

Search blobs by prefix (case-sensitive)

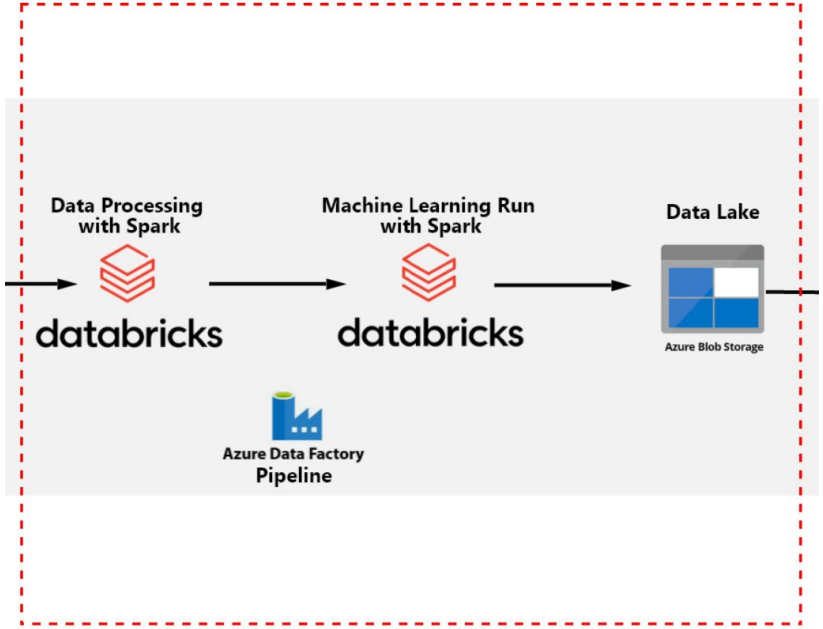
Name	Modified	Access tier
[..]		
Archives		
logs		
Posts		
postTypes		
Users		

Project Steps

Data Transformation

Data Transformation

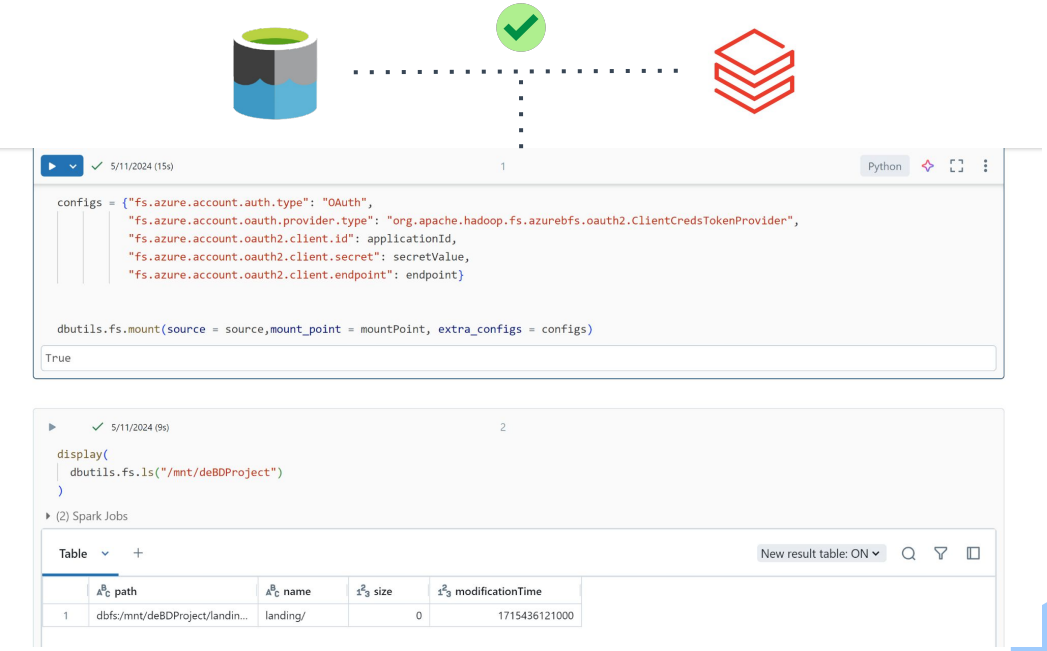
1. Mount Azure Storage container to the Azure Databricks
2. Using training data to train our ML model
   a. Load training data to notebook
   b. Join, filter, and clean the data
   c. Train the model
   d. Save the Model to Azure storage
3. Load model to apply it on our data
4. Save the result in BI folder
5. Add Databricks notebook activity in copyPostsEveryday pipeline



Data Transformation

Data Transformation

1. Mount Azure Storage container to the Azure Databricks



Data Transformation

2. Using training data to train our ML model
   a. Load training data to notebook
   b. Join, filter, and clean the data
   c. Train the model
   d. Save the Model to Azure storage

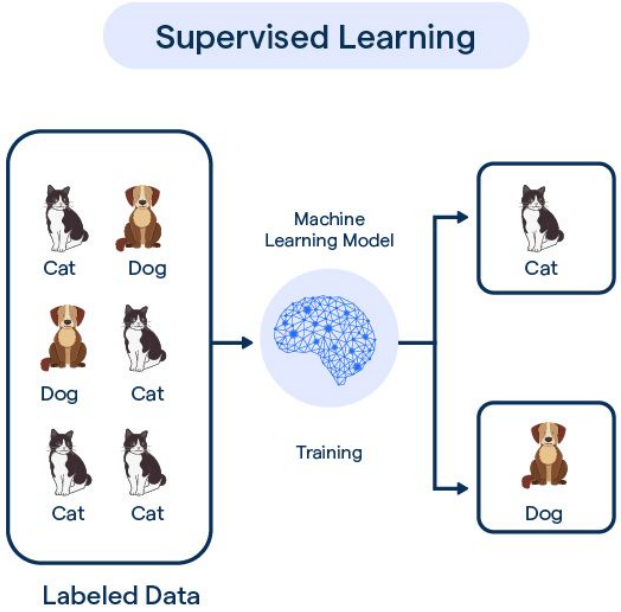
Authentication method: Access key (Switch to Microsoft
Location: bd-project / ml_training

Search blobs by prefix (case-sensitive)

Name
[..]
Posts
PostTypes.txt
users.csv

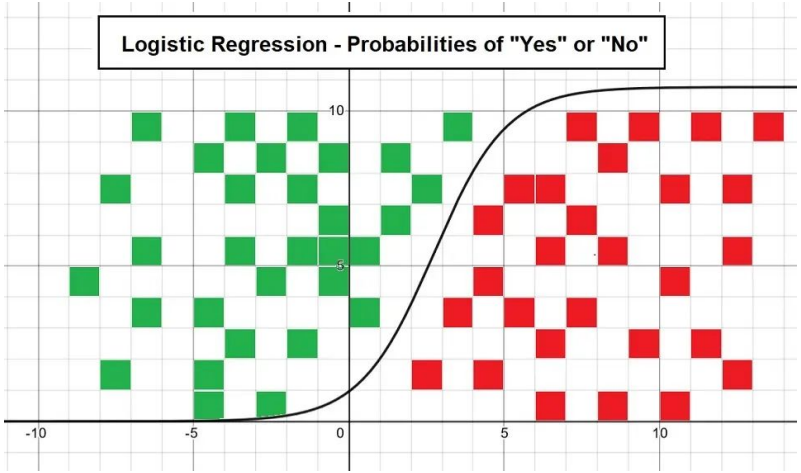
Data Transformation

features	1.2 label
{"vectorType":"sparse","length":8975,"indices":[5,33,91,144,161,317,373,532,612,791,1996,3421,4264,6044],"values":[6.04248601595263...	0
{"vectorType":"sparse","length":8975,"indices":[5,33,91,144,161,317,373,532,612,791,1996,3421,4264,6044],"values":[6.04248601595263...	8
{"vectorType":"sparse","length":8975,"indices":[5,33,91,144,161,317,373,532,612,791,1996,3421,4264,6044],"values":[6.04248601595263...	333
{"vectorType":"sparse","length":8975,"indices":[5,33,91,144,161,317,373,532,612,791,1996,3421,4264,6044],"values":[6.04248601595263...	829
{"vectorType":"sparse","length":8975,"indices":[6,8,28,47,51,67,76,117,123,180,190,206,213,237,266,346,415,568,579,972,1578,1594,350...	134
{"vectorType":"sparse","length":8975,"indices":[6,8,28,47,51,67,76,117,123,180,190,206,213,237,266,346,415,568,579,972,1578,1594,350...	826
{"vectorType":"sparse","length":8975,"indices":[0,1,5,13,36,50,90,98,132,161,176,204,210,318,354,372,414,554,640,1289,1574,1902,2879,...	3
{"vectorType":"sparse","length":8975,"indices":[0,1,5,13,36,50,90,98,132,161,176,204,210,318,354,372,414,554,640,1289,1574,1902,2879,...	659
{"vectorType":"sparse","length":8975,"indices":[0,1,3,4,5,7,12,16,17,23,33,35,51,57,58,60,63,65,73,80,82,88,95,110,113,114,120,121,122,1...	58
{"vectorType":"sparse","length":8975,"indices":[0,1,3,4,5,7,12,16,17,23,33,35,51,57,58,60,63,65,73,80,82,88,95,110,113,114,120,121,122,1...	75
{"vectorType":"sparse","length":8975,"indices":[0,1,3,4,5,7,12,16,17,23,33,35,51,57,58,60,63,65,73,80,82,88,95,110,113,114,120,121,122,1...	50
{"vectorType":"sparse","length":8975,"indices":[0,1,3,4,5,7,12,16,17,23,33,35,51,57,58,60,63,65,73,80,82,88,95,110,113,114,120,121,122,1...	391
{"vectorType":"sparse","length":8975,"indices":[3,8,46,56,70,126,225,243,516,946,1359,2213,4201,5389,5455],"values":[1.1668641449303...	12
{"vectorType":"sparse","length":8975,"indices":[3,8,46,56,70,126,225,243,516,946,1359,2213,4201,5389,5455],"values":[1.1668641449303...	13



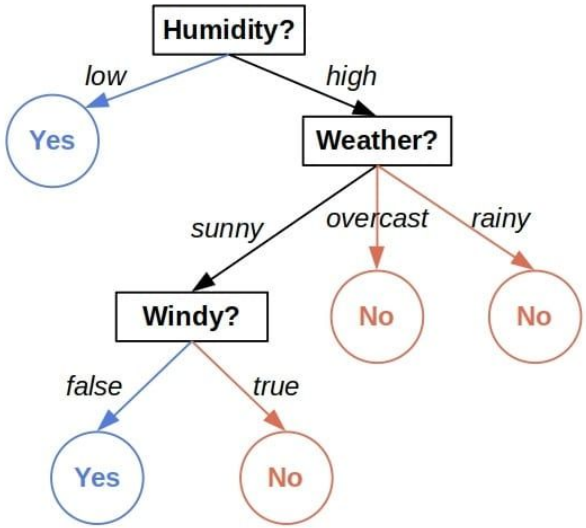
Data Transformation

1. **Logistic Regression:** is a type of regression analysis used for predicting binary outcomes (1/0, True/False, Yes/No) based on one or more predictor variables. It estimates the probability that a given input belongs to a certain class.



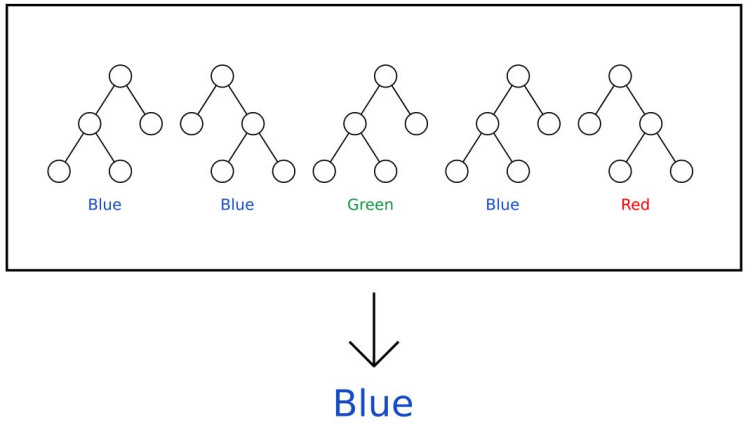
Data Transformation

2. **Decision Trees:** is a flowchart-like tree structure where each internal node represents a decision based on the value of a feature, each branch represents the outcome of that decision, and each leaf node represents a class label.



Data Transformation

3. **Random Forests:** it build many decision trees using different subsets of the data and features. Each tree makes a prediction, and the forest aggregates these predictions to produce the final output.

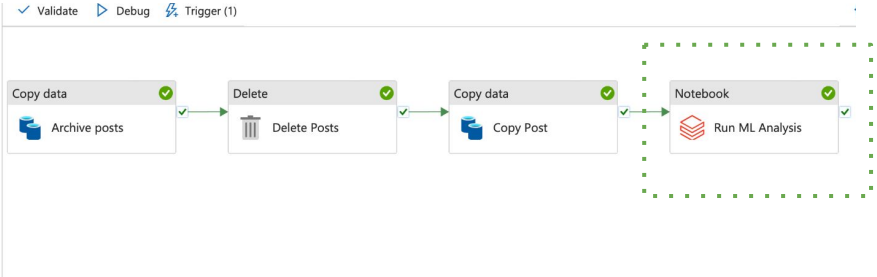


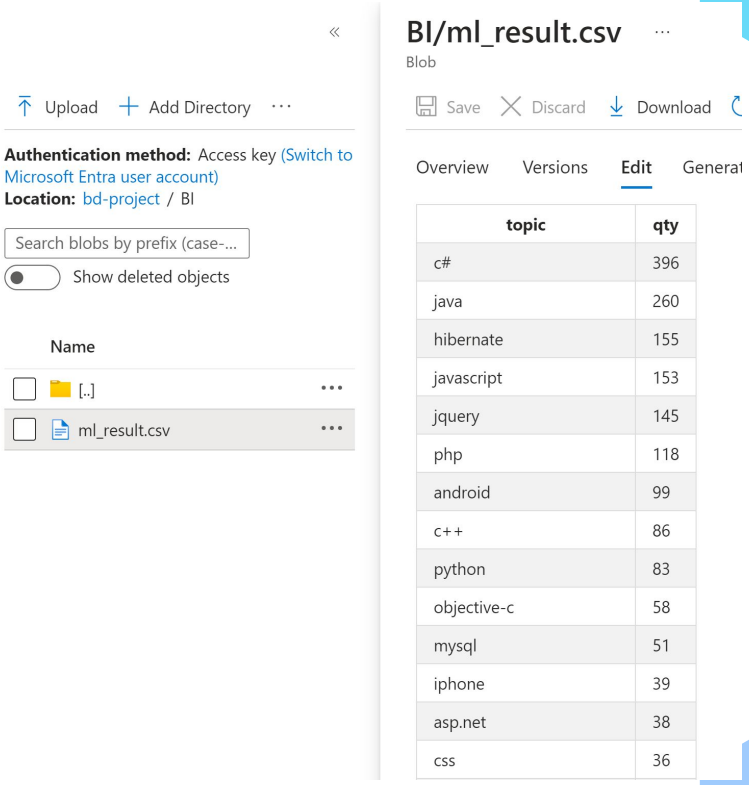
Data Transformation

Logistic Regression	Decision Trees	Random Forests
6 min	50 sec	23 sec
0.3482	0.1027	0.0923

Data Transformation

3. Load model to apply it on our data
4. Save the result in BI folder
5. Add Databricks notebook activity in copyPostsEveryday pipeline.





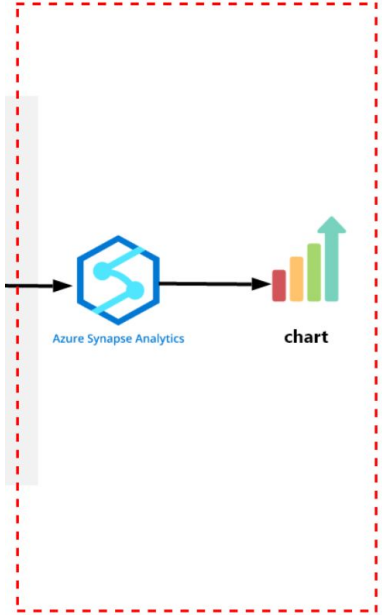
topic	qty
c#	396
java	260
hibernate	155
javascript	153
jquery	145
php	118
android	99
c++	86
python	83
objective-c	58
mysql	51
iphone	39
asp.net	38
css	36

Project Steps

Data Visualization

Data Visualization

- Azure Synaps
- Databricks Dashboard



Data Visuallization

Data Visualization

Azure Synaps: Machine Learning Insights:

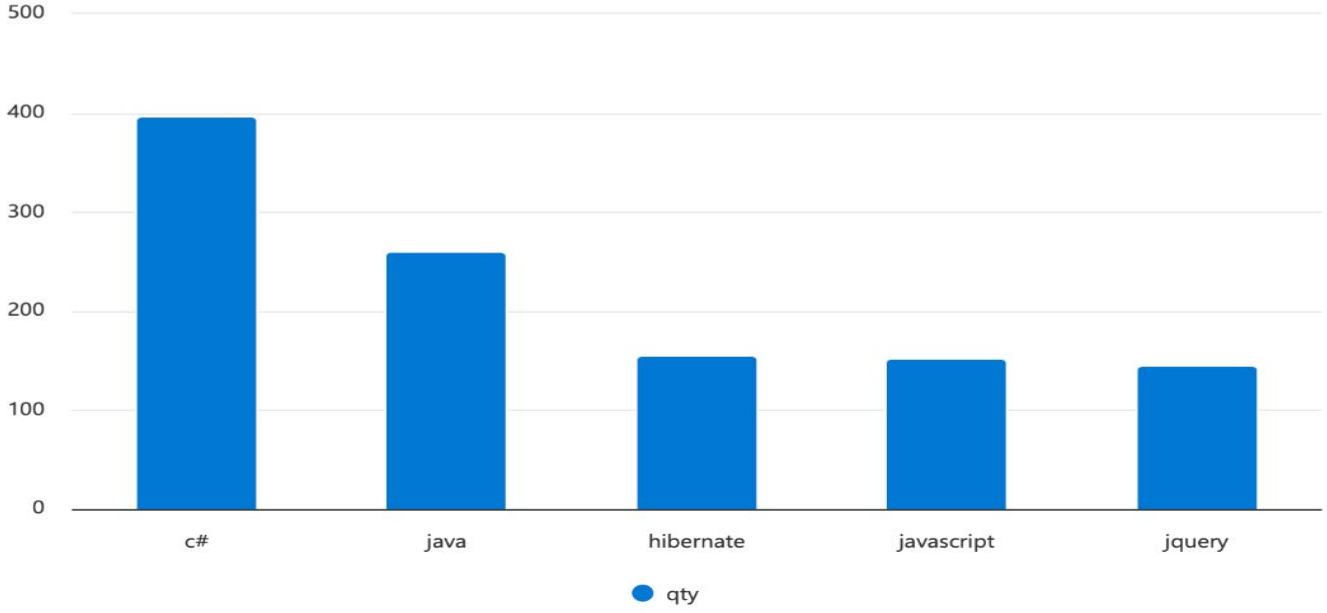
- Extracting Insights through ML
- Visualizing Top 5 Topics

Cancel | Undo | Publish | Query plan | Connect to: Built-in | Use database: master

```sql
-- This is auto-generated code
SELECT
    TOP 5 *
FROM
    OPENROWSET(
        BULK 'https://capstone2storagebigdata.dfs.core.windows.net/bd-project/BI/ml_result.csv',
        FORMAT = 'CSV',
        PARSER_VERSION = '2.0',
        HEADER_ROW = TRUE
    ) AS [result]
```

Data Visualization

Azure Synaps:



Data Visualization

Databricks Dashboard: Exploring Insights:

- EDA and Insightful Visualizations

Future Work

- Collect more data to improve the model's ability to generate accurate results.
- Expand our knowledge in Machine Learning to be able to choose a better model for our data.
- Enhance our data visualization by adding more visualizations and making the dashboard more interactive to dynamically explore and analyze specific subsets of the data.

Conclusion

- Ingested the data from two sources and stored them in Data Lake.
- Created two pipelines in ADF one for updating Posts daily and another for updating Users and Post's Types tables weekly.
- Trained our model and saved it in our Data lake Storage.
- Run the model on our data to give us result that will be used in Synapse.
- Generated a column chart in Synapse to show us the top 5 topics.

Thanks!

Do you have any questions?