Data Engineering Batch 1

PROJECT 2

Data Processing Pipeline

Akilesh K

k.akilesh123@gmail.com

Project statement

Implement a serverless data processing pipeline where Azure Data Factory orchestrates data workflows, and Azure Databricks is used as a serverless processing engine for on-demand analytics and transformations.

Project Overview

This project establishes a serverless code analysis pipeline in Azure. By leveraging Azure Data Factory (ADF), code seamlessly moves from a GitHub repository to secure Azure Blob Storage. Azure Databricks, empowered by PySpark, then analyses the data, offering valuable insights into functionality, quality, and potential issues. This automated process streamlines data analysis, reduces manual intervention, and fosters data-driven decision-making for developers.

About the Project

Dataset:

This dataset, named "organisations", contains information about 10,000 organisations. It was downloaded from the website Datablist: https://www.datablist.com/ and is formatted as a comma-separated values (CSV) file with the following schema:

Attributes:

Index: An integer representing the unique identifier for each organisation within the dataset.

Organization Id: A unique identifier for each organisation, potentially in a non-human-readable format.

Name: The official name of the organisation.

Website: The organisation's website URL (if available).

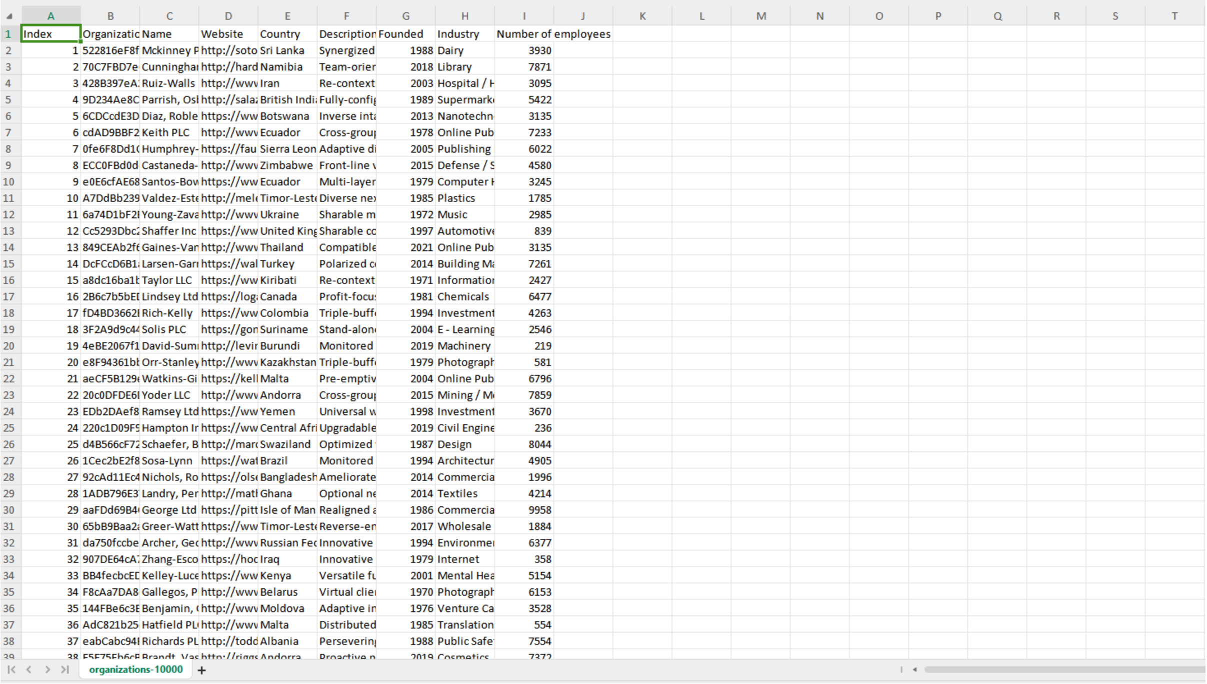
Country: The country in which the organisation is headquartered.

Description: A brief description of the organisation's activities or mission.

Founded: The year in which the organisation was founded.

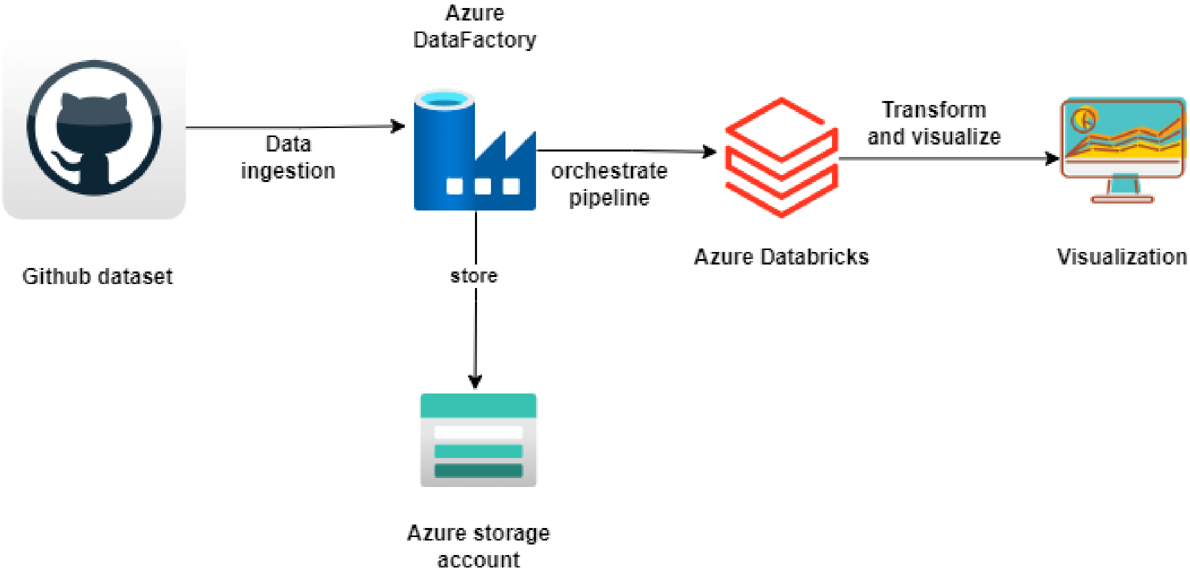
Industry: The primary industry sector in which the organisation operates.

Number of employees: The approximate number of employees working for the organisation (may not be entirely accurate)



	A	B	C	D	E	F	G	H	I
1	Index	Organizatic	Name	Website	Country	Description	Founded	Industry	Number of employees
2	1	522816eF8f	Mckinney P	http://soto	Sri Lanka	Synergized	1988	Dairy	3930
3	2	70C7FBD7e	Cunninghar	http://hard	Namibia	Team-orier	2018	Library	7871
4	3	428B397eA:	Ruiz-Walls	http://wwv	Iran	Re-context	2003	Hospital / H	3095
5	4	9D234Ae8C	Parrish, Osl	http://sala:	British Indi:	Fully-config	1989	Supermark	5422
6	5	6CDCcdE3D	Diaz, Roble	https://ww	Botswana	Inverse int:	2013	Nanotechn	3135
7	6	cdAD9BBF2	Keith PLC	http://wwv	Ecuador	Cross-grou	1978	Online Pub	7233
8	7	0fe6F8Dd1(	Humphrey-	https://fau	Sierra Leon	Adaptive di	2005	Publishing	6022
9	8	ECC0FBd0d	Castaneda-	http://wwv	Zimbabwe	Front-line v	2015	Defense / S	4580
10	9	e0E6cfAE68	Santos-Bow	https://ww	Ecuador	Multi-layer	1979	Computer H	3245
11	10	A7DdBb239	Valdez-Este	http://mel	Timor-Leste	Diverse nex	1985	Plastics	1785
12	11	6a74D1bF2I	Young-Zava	http://wwv	Ukraine	Sharable m	1972	Music	2985
13	12	Cc5293Dbc2	Shaffer Inc	https://ww	United King	Sharable cc	1997	Automotive	839
14	13	849CEAb2f6	Gaines-Var	http://wwv	Thailand	Compatible	2021	Online Pub	3135
15	14	DcFCcD6B1:	Larsen-Garr	https://wal	Turkey	Polarized c	2014	Building Ma	7261
16	15	a8dc16ba1t	Taylor LLC	https://ww	Kiribati	Re-context	1971	Informatior	2427
17	16	2B6c7b5bEI	Lindsey Ltd	https://log:	Canada	Profit-focu:	1981	Chemicals	6477
18	17	fD4BD3662I	Rich-Kelly	https://ww	Colombia	Triple-buff	1994	Investment	4263
19	18	3F2A9d9c44	Solis PLC	https://gon	Suriname	Stand-alon	2004	E - Learning	2546
20	19	4eBE2067f1	David-Sum	http://levir	Burundi	Monitored	2019	Machinery	219
21	20	e8F94361bt	Orr-Stanley	http://wwv	Kazakhstan	Triple-buff	1979	Photograph	581
22	21	aeCF5B129	Watkins-Gi	https://kell	Malta	Pre-emptiv	2004	Online Pub	6796
23	22	20c0DFDE6I	Yoder LLC	http://wwv	Andorra	Cross-grou	2015	Mining / M	7859
24	23	EDb2DAef8	Ramsey Ltd	https://ww	Yemen	Universal w	1998	Investment	3670
25	24	220c1D09F9	Hampton Ir	https://ww	Central Afri	Upgradable	2019	Civil Engine	236
26	25	d4B566cF72	Schaefer, B	http://marc	Swaziland	Optimized	1987	Design	8044
27	26	1Cec2bE2f8	Sosa-Lynn	https://wat	Brazil	Monitored	1994	Architectur	4905
28	27	92cAd11Ec4	Nichols, Ro	https://olse	Bangladesh	Ameliorate	2014	Commercia	1996
29	28	1ADB796E3	Landry, Per	http://matl	Ghana	Optional ne	2014	Textiles	4214
30	29	aaFDd69B4	George Ltd	https://pitt	Isle of Man	Realigned a	1986	Commercia	9958
31	30	65bB9Baa2a	Greer-Watt	https://ww	Timor-Leste	Reverse-en	2017	Wholesale	1884
32	31	da750fccbe	Archer, Gec	http://wwv	Russian Fec	Innovative	1994	Environmer	6377
33	32	907DE64cA7	Zhang-Esco	https://hoc	Iraq	Innovative	1979	Internet	358
34	33	BB4fecbcED	Kelley-Luce	https://ww	Kenya	Versatile fu	2001	Mental Hea	5154
35	34	F8cAa7DA8	Gallegos, P	http://wwv	Belarus	Virtual clier	1970	Photograph	6153
36	35	144FBe6c3E	Benjamin, (	http://wwv	Moldova	Adaptive in	1976	Venture Ca	3528
37	36	AdC821b25	Hatfield PL(	http://wwv	Malta	Distributed	1985	Translation	554
38	37	eabCabc94I	Richards PL	http://todd	Albania	Perseverin	1988	Public Safe	7554
39	38	E5E75Eb6cE	Brandt, Vas	http://riggs	Andorra	Proactive n	2019	Cosmetics	7372

Architecture



The architectural diagram depicts a systematic approach to data processing and analysis within the project. It begins with the acquisition of code from a GitHub repository, facilitated by Azure Data Factory (ADF). ADF orchestrates the transfer of code to Azure Blob Storage, where it is securely stored for further processing. Azure Databricks workspace is then provisioned to handle data processing tasks efficiently. Within the workspace, a Notebook activity configured within ADF executes PySpark code responsible for data analysis. This PySpark code loads the code data from Blob Storage into Spark DataFrame, conducts necessary transformations, and performs analysis using PySpark libraries. The results of this analysis are visualized using tools available within Azure Databricks, with options for generating visualizations such as pie charts to represent the code analysis findings. Additionally, if required, the processed data or analysis results can be stored back to Blob Storage for future reference or sharing.

Azure Resources Used for This Project

- **Azure Data Factory (ADF):**

An ADF instance will be created to orchestrate and automate the data transfer process between GitHub and Azure Blob Storage.

Linked Services: Configured to connect to GitHub using an appropriate authentication method (e.g., personal access token, OAuth) and Azure Blob Storage.

Datasets: Defined to represent the schema of the code files (e.g., text format) in the GitHub repository.

Pipelines: Created to define the sequence of activities for copying code from the GitHub repository to a specific blob storage container.

- **Azure Blob Storage:**

A blob storage container will be used as the destination for storing code transferred from the GitHub repository.

This stored code will be accessed by Azure Databricks for data processing tasks.

- **Azure Databricks:**

A Databricks workspace will be provisioned to perform data processing tasks using PySpark.

Databricks Notebooks: Utilized to write and execute PySpark code for code analysis, including tasks like:

Extracting data from the code files (e.g., lines of code, function names, comments).

Performing data transformations and cleaning (e.g., removing irrelevant content, formatting code).

Project Requirements

- **Data Source:**

Type: Identify the specific source of your code, such as a GitHub repository, Azure DevOps repository, or a local file system.

Connectivity: Ensure you have the necessary credentials and access methods to securely connect to this source from the Azure environment.

- **Azure Data Factory (ADF) Configuration:**

Instance: Create an Azure Data Factory (ADF) instance within your Azure subscription.

Linked Services:

Configure a linked service to connect to the GitHub repository using an appropriate authentication method (e.g., personal access token, OAuth).

Configure a linked service to connect to Azure Blob Storage for storing the transferred code.

Datasets: Define a dataset in ADF to represent the schema of the code files (e.g., text format) from the source location.

Pipelines: Create a pipeline in ADF to define the data transfer process:

Use a Copy Activity to copy code from the source location (e.g., GitHub repository) to a specific blob storage container.

Configure the copy activity with source and destination details, scheduling options (e.g., daily), and error handling mechanisms.

- **Data Storage:**

Azure Blob Storage: Utilize a blob storage container as the destination for storing the transferred code retrieved through ADF. This stored code will be accessed by Azure Databricks for further processing.

- **Data Processing with Azure Databricks:**

Provision an Azure Databricks workspace in your Azure subscription.

Define and implement data processing tasks using PySpark within the Databricks environment. PySpark provides a powerful framework for distributed data processing, enabling tasks such as data cleaning, transformation, aggregation, and analysis.

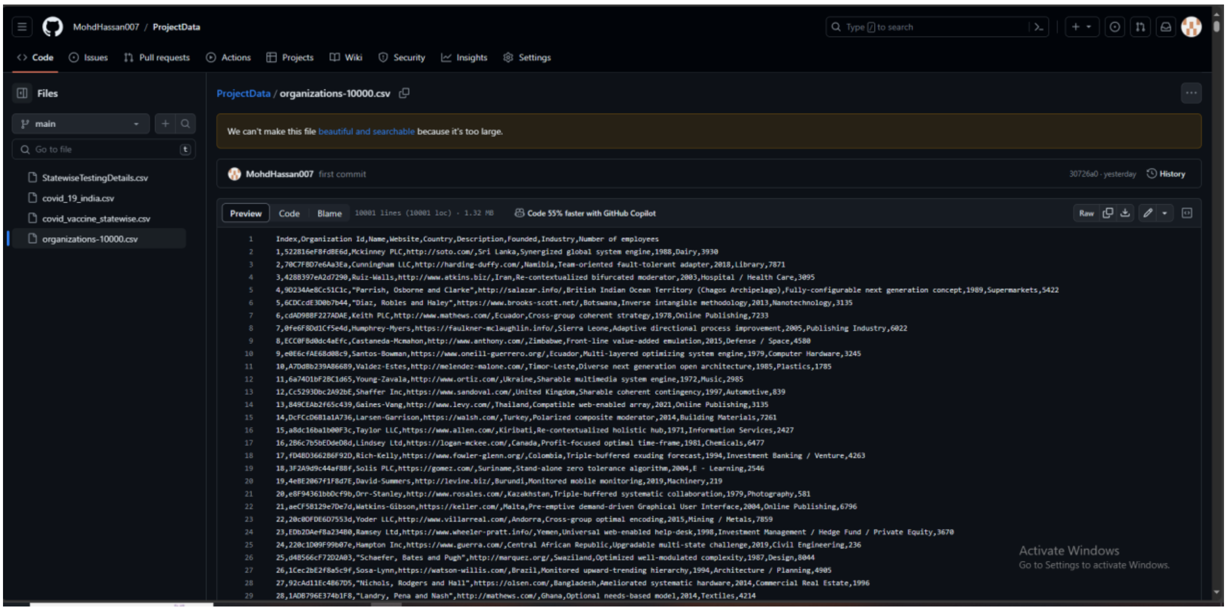
Use Databricks notebooks to write and execute PySpark code interactively, leveraging the scalability and performance of the Databricks runtime.

Tasks performed:

- Acquire data: Retrieve code from a GitHub repository.
- Transfer data: Utilize Azure Data Factory (ADF) to orchestrate the transfer of code from GitHub to Azure Blob Storage.
- Store data: Utilize Azure Blob Storage to securely store the transferred data.
- Process data:
- Provision an Azure Databricks workspace.
- Within ADF, add a Notebook activity: Configure the activity to use your Databricks workspace and specify the PySpark notebook responsible for code analysis.
- Within the PySpark notebook:
- Load data from Blob Storage into a Spark DataFrame.
- Implement data transformations and analysis using PySpark libraries.
- Visualize data: Generate visualizations (e.g., pie charts) depicting the results of code analysis using libraries within Databricks.

Implementation

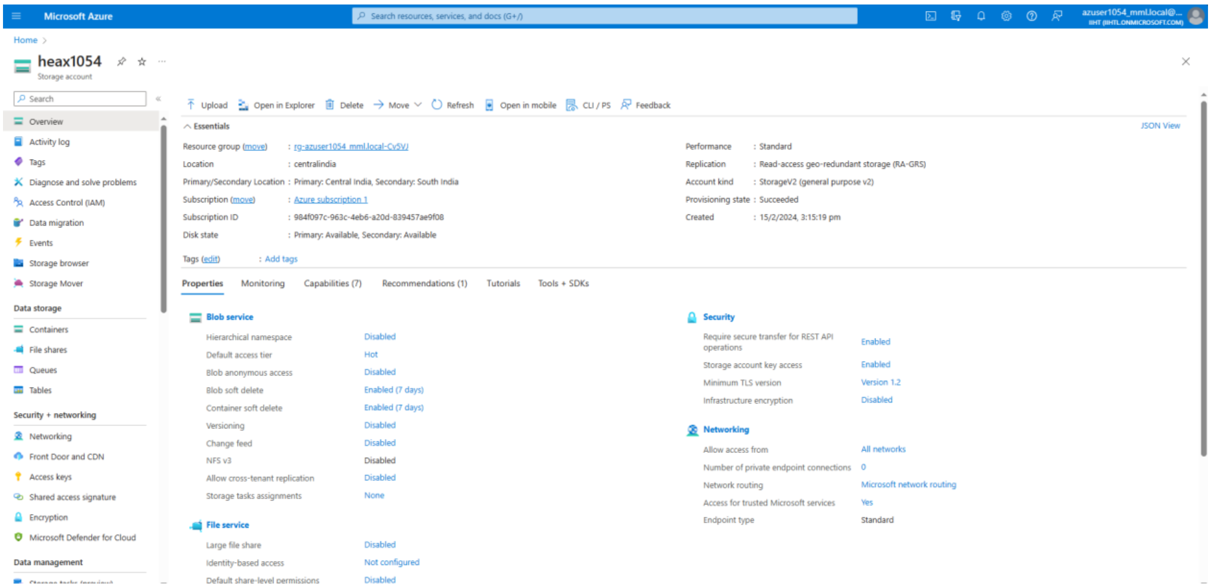
- **CSV data on github**


MohdHassan007 / ProjectData
Type / to search
Code
Issues
Pull requests
Actions
Projects
Wiki
Security
Insights
Settings
Files
main
Go to file
StatewiseTestingDetails.csv
covid_19_india.csv
covid_vaccine_statewise.csv
organizations-10000.csv
ProjectData / organizations-10000.csv
MohdHassan007 first commit
30726a0 · yesterday
History
Preview
Code
Blame
10001 lines (10001 loc) · 1.32 MB
Code 55% faster with GitHub Copilot
Raw
Activate Windows
Go to Settings to activate Windows.

We can't make this file beautiful and searchable because it's too large.

```
Index,Organization Id,Name,Website,Country,Description,Founded,Industry,Number of employees
1,522816eF8fdBE6d,Mckinney PLC,http://soto.com/,Sri Lanka,Synergized global system engine,1988,Dairy,3930
2,70C7FBD7e6Aa3Ea,Cunningham LLC,http://harding-duffy.com/,Namibia,Team-oriented fault-tolerant adapter,2018,Library,7871
3,4288397eA2d7290,Ruiz-Walls,http://www.atkins.biz/,Iran,Re-contextualized bifurcated moderator,2003,Hospital / Health Care,3095
4,9D234Ae8Cc51C1c,"Parrish, Osborne and Clarke",http://salazar.info/,British Indian Ocean Territory (Chagos Archipelago),Fully-configurable next generation concept,1989,Supermarkets,5422
5,6CDCcdE3D0b7b44,"Diaz, Robles and Haley",https://www.brooks-scott.net/,Botswana,Inverse intangible methodology,2013,Nanotechnology,3135
6,cdAD9BBF227ADAE,Keith PLC,http://www.mathews.com/,Ecuador,Cross-group coherent strategy,1978,Online Publishing,7233
7,0fe6F8Dd1Cf5e4d,Humphrey-Myers,https://faulkner-mclaughlin.info/,Sierra Leone,Adaptive directional process improvement,2005,Publishing Industry,6022
8,ECC0FBd0dc4aEfc,Castaneda-Mcmahon,http://www.anthony.com/,Zimbabwe,Front-line value-added emulation,2015,Defense / Space,4580
9,e0E6cfAE68d08c9,Santos-Bowman,https://www.oneill-guerrero.org/,Ecuador,Multi-layered optimizing system engine,1979,Computer Hardware,3245
10,A7DdBb239A86689,Valdez-Estes,http://melendez-malone.com/,Timor-Leste,Diverse next generation open architecture,1985,Plastics,1785
11,6a74D1bF2BC1d65,Young-Zavala,http://www.ortiz.com/,Ukraine,Sharable multimedia system engine,1972,Music,2985
12,Cc5293Dbc2A92bE,Shaffer Inc,https://www.sandoval.com/,United Kingdom,Sharable coherent contingency,1997,Automotive,839
13,849CEAb2f65c439,Gaines-Vang,http://www.levy.com/,Thailand,Compatible web-enabled array,2021,Online Publishing,3135
14,DcFCcD681a1A736,Larsen-Garrison,https://walsh.com/,Turkey,Polarized composite moderator,2014,Building Materials,7261
15,a8dc16ba1b00F3c,Taylor LLC,https://www.allen.com/,Kiribati,Re-contextualized holistic hub,1971,Information Services,2427
16,2B6c7b5bEDdeD8d,Lindsey Ltd,https://logan-mckee.com/,Canada,Profit-focused optimal time-frame,1981,Chemicals,6477
17,fD48D3662B6F92D,Rich-Kelly,https://www.fowler-glenn.org/,Colombia,Triple-buffered exuding forecast,1994,Investment Banking / Venture,4263
18,3F2A9d9c44af88f,Solis PLC,https://gomez.com/,Suriname,Stand-alone zero tolerance algorithm,2004,E - Learning,2546
19,4eBE2067f1F8d7E,David-Summers,http://levine.biz/,Burundi,Monitored mobile monitoring,2019,Machinery,219
20,e8F94361bbDcf9b,Orr-Stanley,http://www.rosales.com/,Kazakhstan,Triple-buffered systematic collaboration,1979,Photography,581
21,aeCF58129e7De7d,Watkins-Gibson,https://keller.com/,Malta,Pre-emptive demand-driven Graphical User Interface,2004,Online Publishing,6796
22,20c0DFDE6D7553d,Yoder LLC,http://www.villarreal.com/,Andorra,Cross-group optimal encoding,2015,Mining / Metals,7859
23,EDb2DAef8a234B0,Ramsey Ltd,https://www.wheeler-pratt.info/,Yemen,Universal web-enabled help-desk,1998,Investment Management / Hedge Fund / Private Equity,3670
24,220c1D09F99b07e,Hampton Inc,https://www.guerra.com/,Central African Republic,Upgradable multi-state challenge,2019,Civil Engineering,236
25,d4B566cF72D2A03,"Schaefer, Bates and Pugh",http://marquez.org/,Swaziland,Optimized well-modulated complexity,1987,Design,8044
26,1Cec2bE2f8a5c9f,Sosa-Lynn,https://watson-willis.com/,Brazil,Monitored upward-trending hierarchy,1994,Architecture / Planning,4905
27,92cAd11Ec4B67D5,"Nichols, Rodgers and Hall",https://olsen.com/,Bangladesh,Ameliorated systematic hardware,2014,Commercial Real Estate,1996
28,1ADB796E374b1F8,"Landry, Pena and Nash",http://mathews.com/,Ghana,Optional needs-based model,2014,Textiles,4214
```

- **In azure ,create a storage account mentioning the location**



Essentials

Resource group (move) : rg-azuser1054_mml.local-Cv5VJ
Location : centralindia
Primary/Secondary Location : Primary: Central India, Secondary: South India
Subscription (move) : Azure subscription 1
Subscription ID : 984f097c-963c-4eb6-a20d-839457ae9f08
Disk state : Primary: Available, Secondary: Available

Performance : Standard
Replication : Read-access geo-redundant storage (RA-GRS)
Account kind : StorageV2 (general purpose v2)
Provisioning state : Succeeded
Created : 15/2/2024, 3:15:19 pm

Tags (edit) : Add tags

Blob service

Hierarchical namespace	Disabled
Default access tier	Hot
Blob anonymous access	Disabled
Blob soft delete	Enabled (7 days)
Container soft delete	Enabled (7 days)
Versioning	Disabled
Change feed	Disabled
NFS v3	Disabled
Allow cross-tenant replication	Disabled
Storage tasks assignments	None

File service

Large file share	Disabled
Identity-based access	Not configured
Default share-level permissions	Disabled

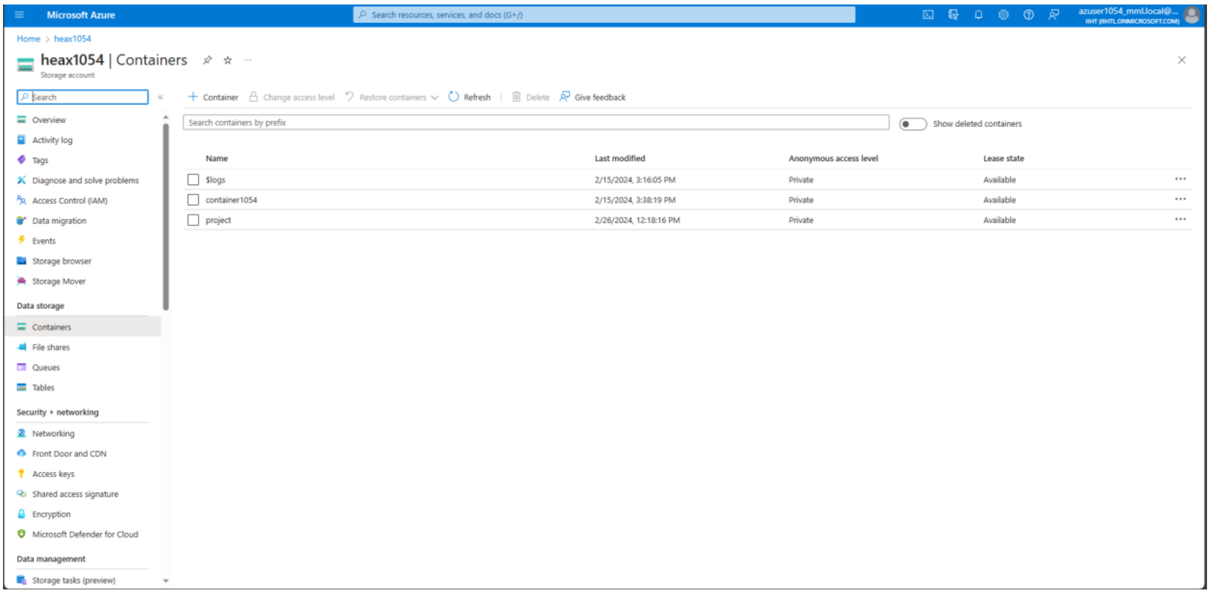
Security

Require secure transfer for REST API operations	Enabled
Storage account key access	Enabled
Minimum TLS version	Version 1.2
Infrastructure encryption	Disabled

Networking

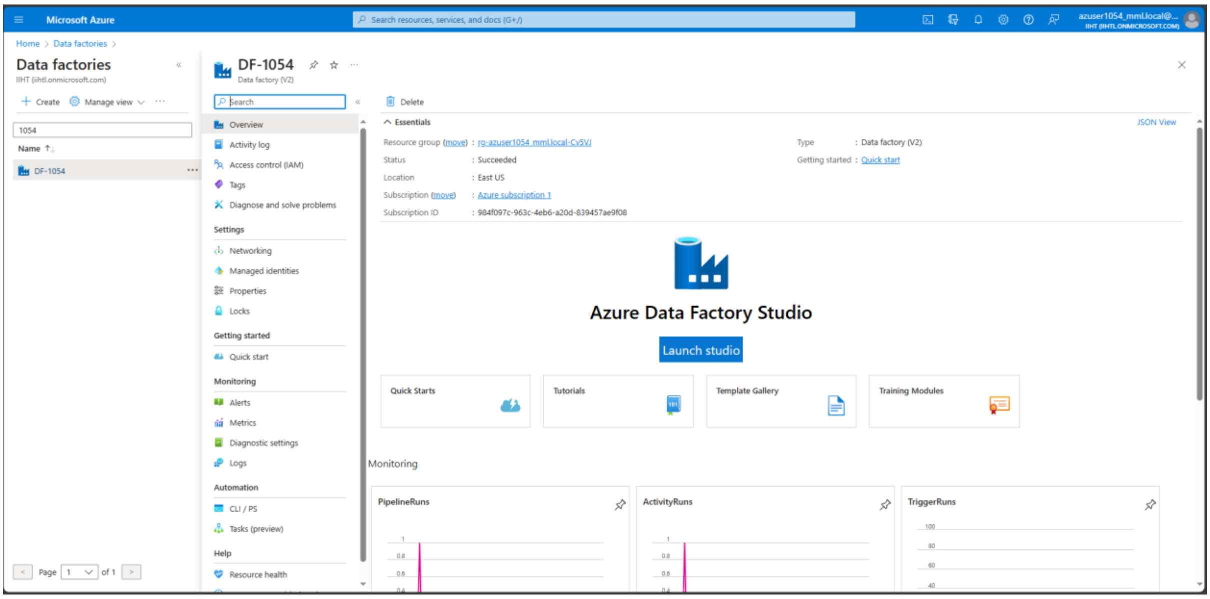
Allow access from	All networks
Number of private endpoint connections	0
Network routing	Microsoft network routing
Access for trusted Microsoft services	Yes
Endpoint type	Standard

- **create a new blob container**



	Name	Last modified	Anonymous access level	Lease state
□	$logs	2/15/2024, 3:16:05 PM	Private	Available
□	container1054	2/15/2024, 3:38:19 PM	Private	Available
□	project	2/26/2024, 12:18:16 PM	Private	Available

- **create a azure data factory account**



Essentials

Resource group (move) : rg-azuser1054_mml.local-Cv5VJ

Status : Succeeded

Location : East US

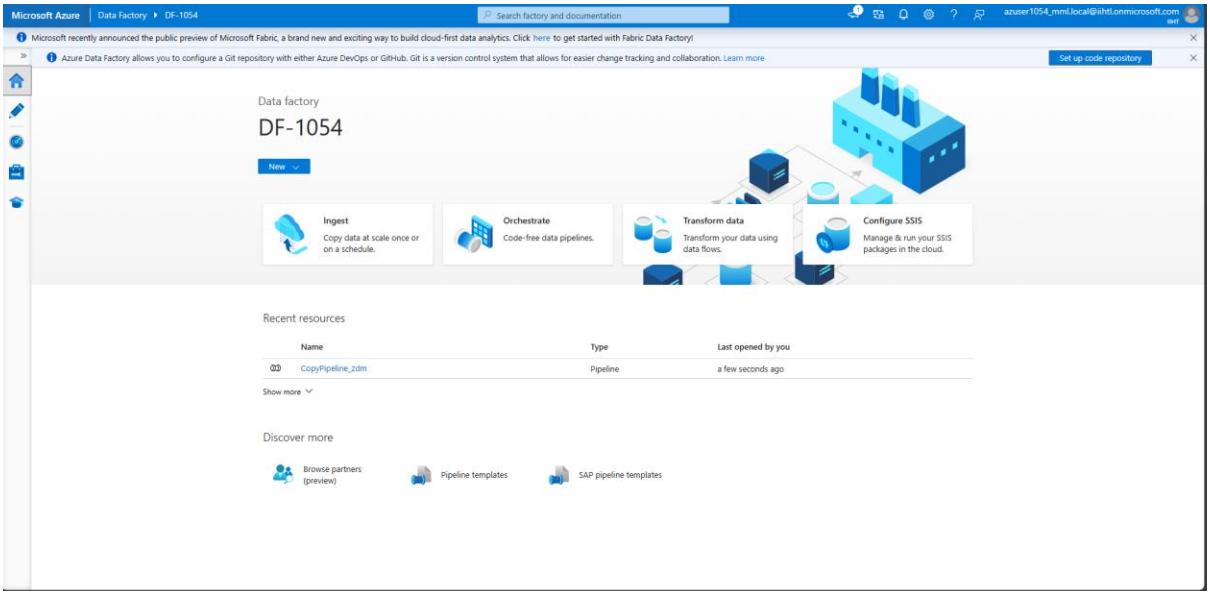
Subscription (move) : Azure subscription 1

Subscription ID : 984f097c-963c-4eb6-a20d-839457ae9f08

Type : Data factory (V2)

Getting started : Quick start

- **select ingest data to create a new data pipeline**


Microsoft Azure
Data Factory ▸ DF-1054
Search factory and documentation
azuser1054_mml.local@iihtl.onmicrosoft.com

Microsoft recently announced the public preview of Microsoft Fabric, a brand new and exciting way to build cloud-first data analytics. Click here to get started with Fabric Data Factory!

Azure Data Factory allows you to configure a Git repository with either Azure DevOps or GitHub. Git is a version control system that allows for easier change tracking and collaboration. Learn more

Data factory

DF-1054

Ingest
Copy data at scale once or on a schedule.

Orchestrate
Code-free data pipelines.

Transform data
Transform your data using data flows.

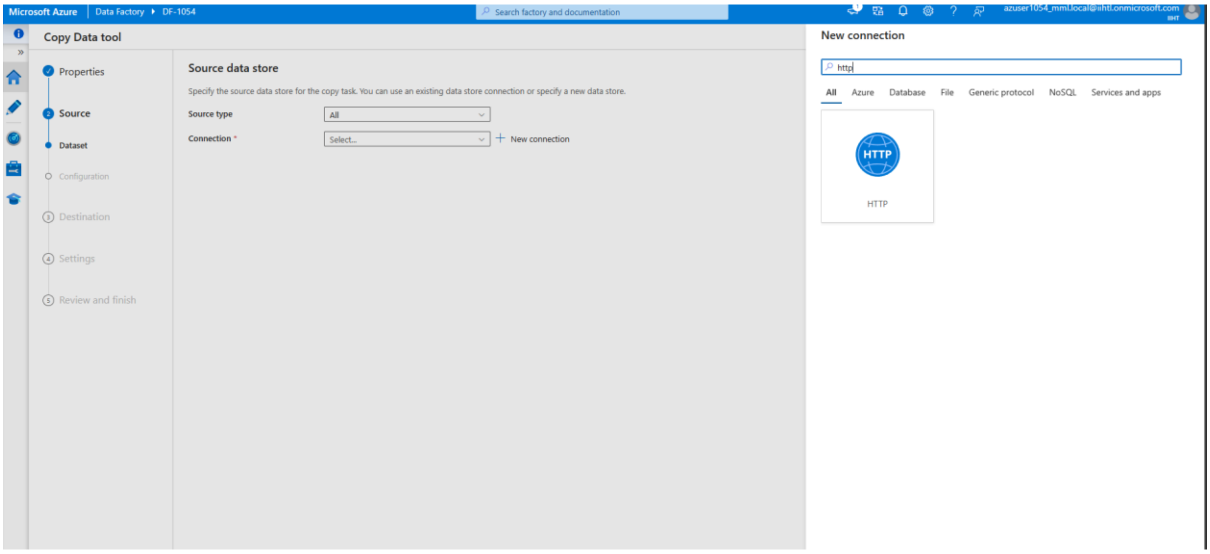
Configure SSIS
Manage & run your SSIS packages in the cloud.

Recent resources

Name	Type	Last opened by you
CopyPipeline_zdm	Pipeline	a few seconds ago

Discover more

- **For Source click on new connection and type http**


Microsoft Azure
Data Factory ▸ DF-1054
Search factory and documentation
azuser1054_mml.local@iihtl.onmicrosoft.com

Copy Data tool

Source data store

Specify the source data store for the copy task. You can use an existing data store connection or specify a new data store.

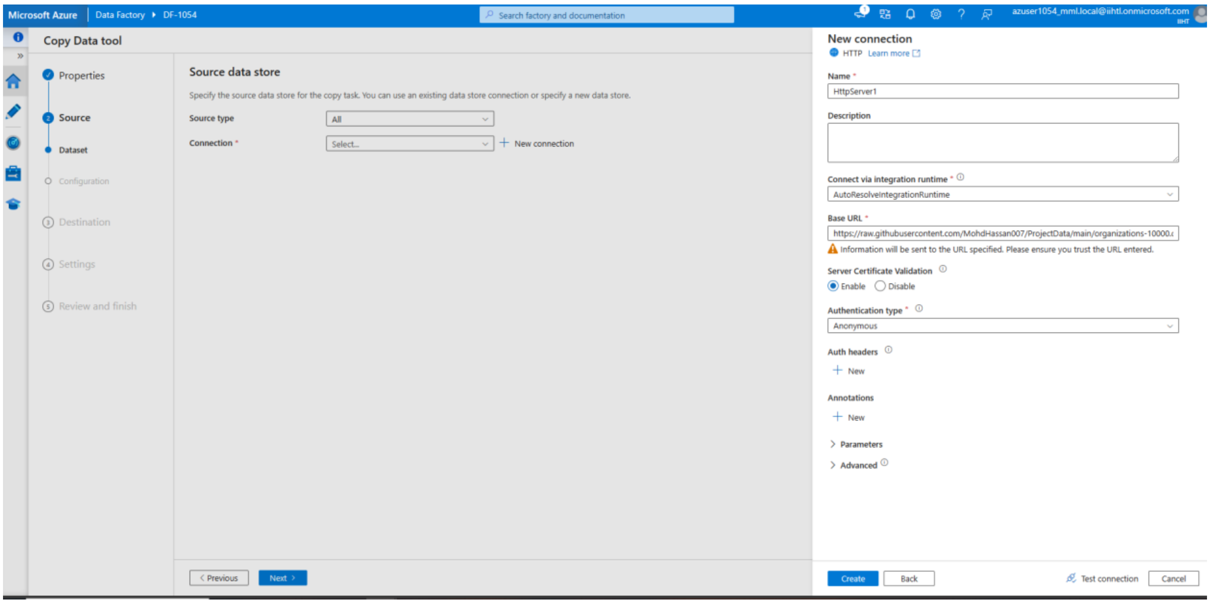
Source type: All

Connection *: Select...

New connection

http

- **Copy and paste the raw url of the csv file**



Source data store

Specify the source data store for the copy task. You can use an existing data store connection or specify a new data store.

Source type: All

Connection *: Select... + New connection

New connection

HTTP Learn more

Name *: HttpServer1

Description

Connect via integration runtime *: AutoResolveIntegrationRuntime

Base URL *: https://raw.githubusercontent.com/MohdHassan007/ProjectData/main/organizations-10000.c

⚠ Information will be sent to the URL specified. Please ensure you trust the URL entered.

Server Certificate Validation: ● Enable ○ Disable

Authentication type *: Anonymous

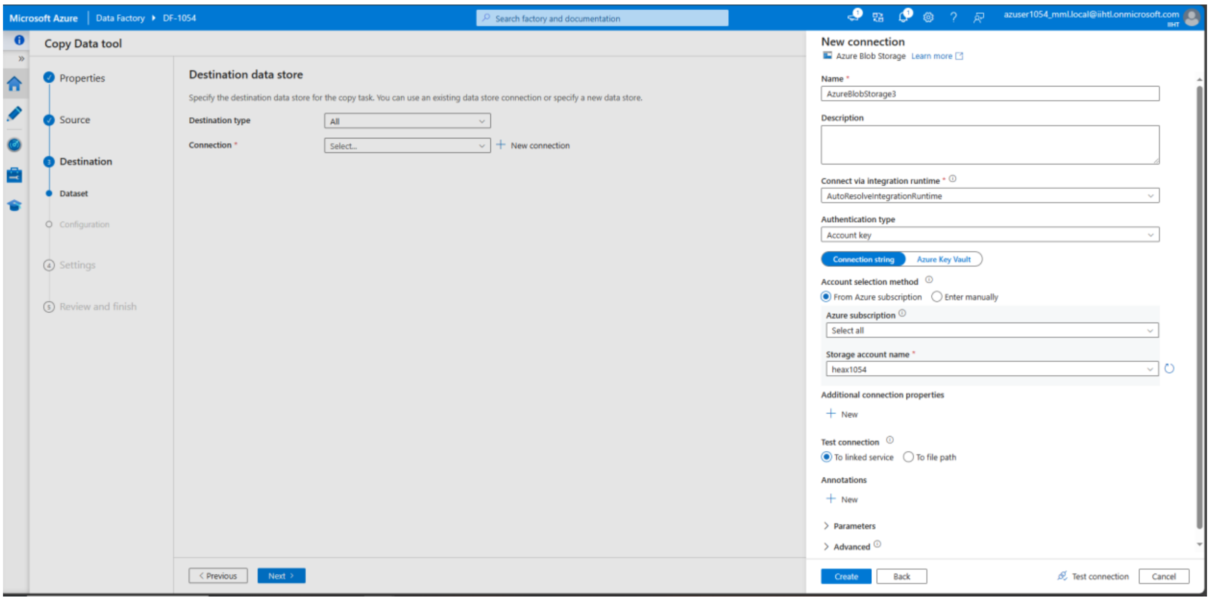
Auth headers: + New

Annotations: + New

> Parameters

> Advanced

- **Create a connection for destination blob storage**



Destination data store

Specify the destination data store for the copy task. You can use an existing data store connection or specify a new data store.

Destination type: All

Connection *: Select... + New connection

New connection

Azure Blob Storage Learn more

Name *: AzureBlobStorage3

Description

Connect via integration runtime *: AutoResolveIntegrationRuntime

Authentication type: Account key

Connection string | Azure Key Vault

Account selection method: ● From Azure subscription ○ Enter manually

Azure subscription: Select all

Storage account name *: heax1054

Additional connection properties: + New

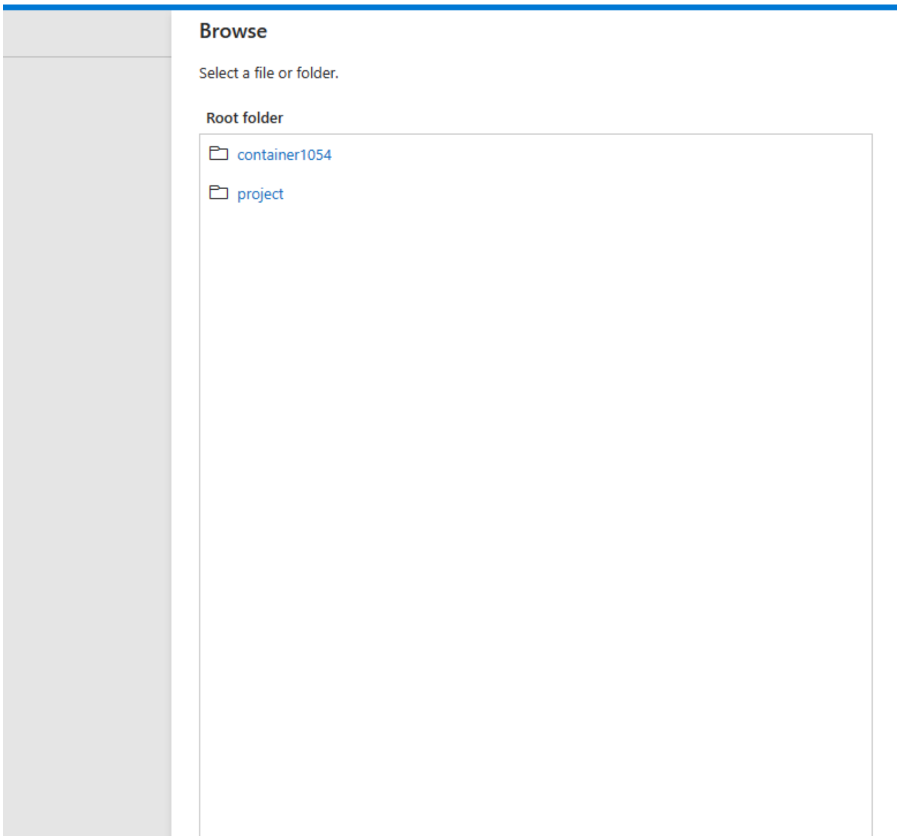
Test connection: ● To linked service ○ To file path

Annotations: + New

> Parameters

> Advanced

- **Select the container**



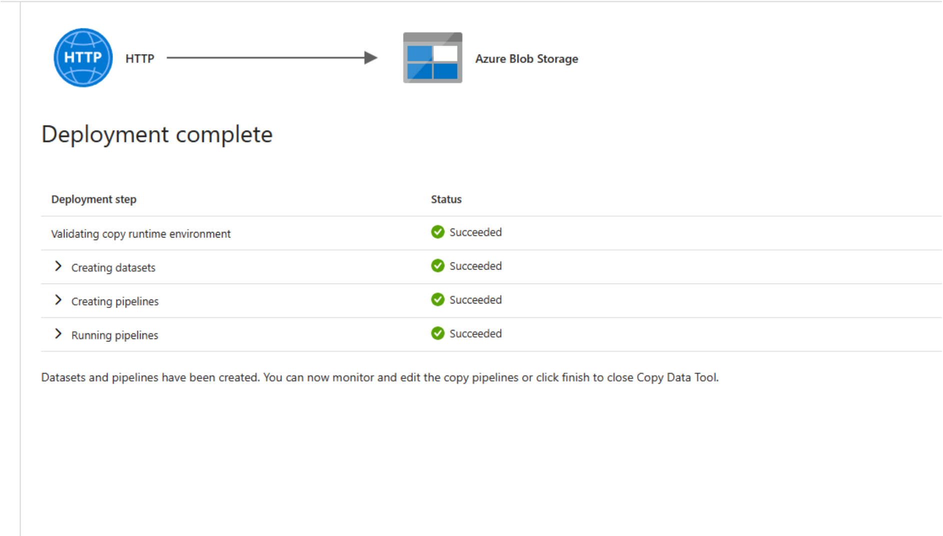
Browse

Select a file or folder.

Root folder

- container1054
- project

- **summary of the deployment of from SQL server to azure blob storage**

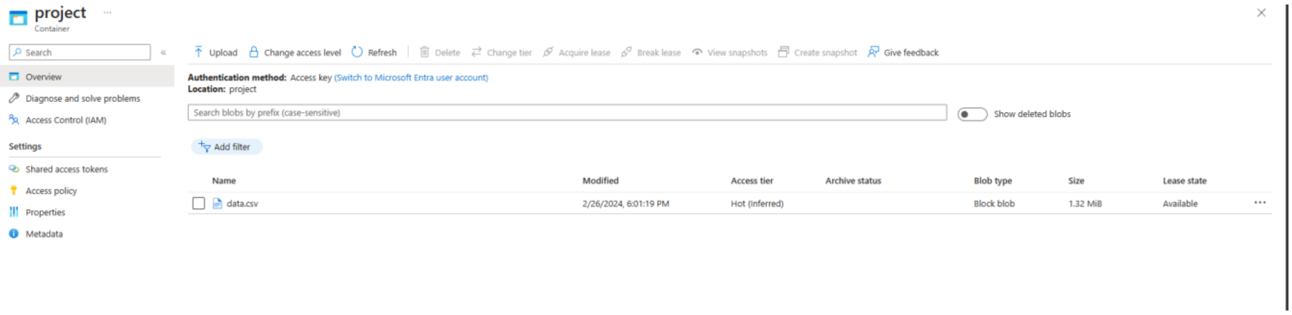


Deployment complete

Deployment step	Status
Validating copy runtime environment	✓ Succeeded
› Creating datasets	✓ Succeeded
› Creating pipelines	✓ Succeeded
› Running pipelines	✓ Succeeded

Datasets and pipelines have been created. You can now monitor and edit the copy pipelines or click finish to close Copy Data Tool.

- **The csv file is stored in the container**

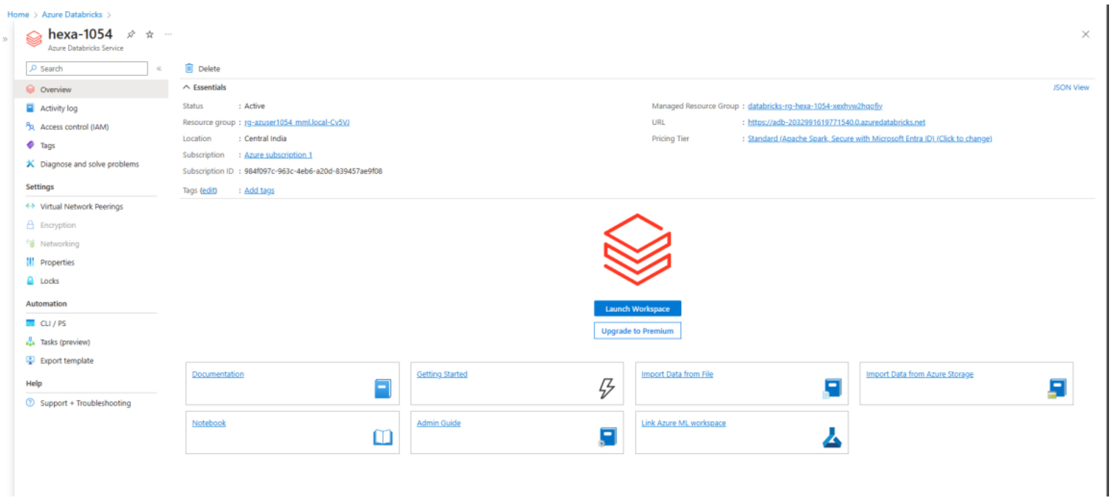


project
Container

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

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
data.csv	2/26/2024, 6:01:19 PM	Hot (Inferred)		Block blob	1.32 MiB	Available

- **Create a Azure databricks and launch it**



Home > Azure Databricks >

hexa-1054
Azure Databricks Service

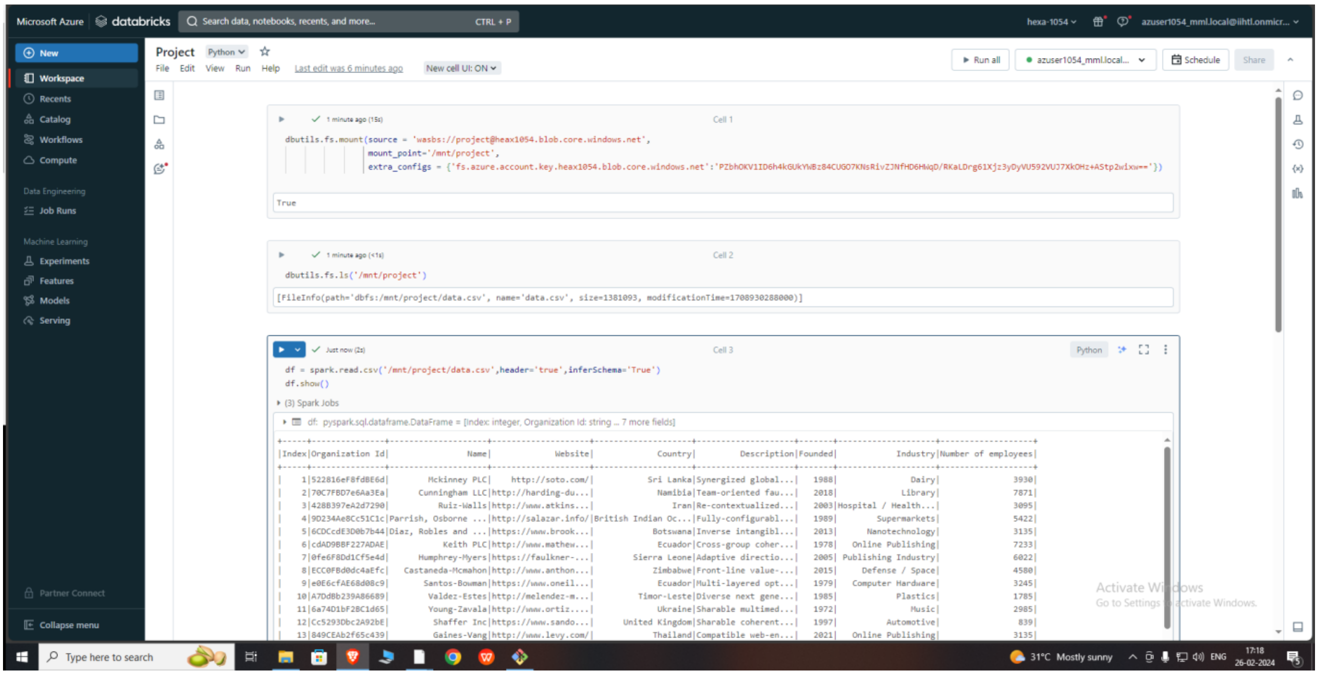
Essentials

Status : Active
Resource group : rg-azuser1054_mml.local-Cv5VJ
Location : Central India
Subscription : Azure subscription 1
Subscription ID : 984f097c-963c-4eb6-a20d-839457ae9f08
Tags (edit) : Add tags

Managed Resource Group : databricks-rg-hexa-1054-xexhyw2hqofiy
URL : https://adb-2032991619771540.0.azuredatabricks.net
Pricing Tier : Standard (Apache Spark, Secure with Microsoft Entra ID) (Click to change)

Mount blob storage to azure databricks

- **Add source line of the blob storage**
- **enter the mount points to store in databricks**
- **Configure extra_configs by mentioning the access key and pasting the key of the blob storage.**
- **list of the files mounted on Databricks**
- **display the data frame**



```
dbutils.fs.mount(source = 'wasbs://project@heax1054.blob.core.windows.net',
                 mount_point='/mnt/project',
                 extra_configs = {'fs.azure.account.key.heax1054.blob.core.windows.net':'PZbhOKV1ID6h4kGUkYWBz84CUGO7KNsRivZJNfHD6HWqD/RKaLDrg61Xjz3yDyVU592VUJ7XkOHz+AStp2wixw=='})
```

```
True
```

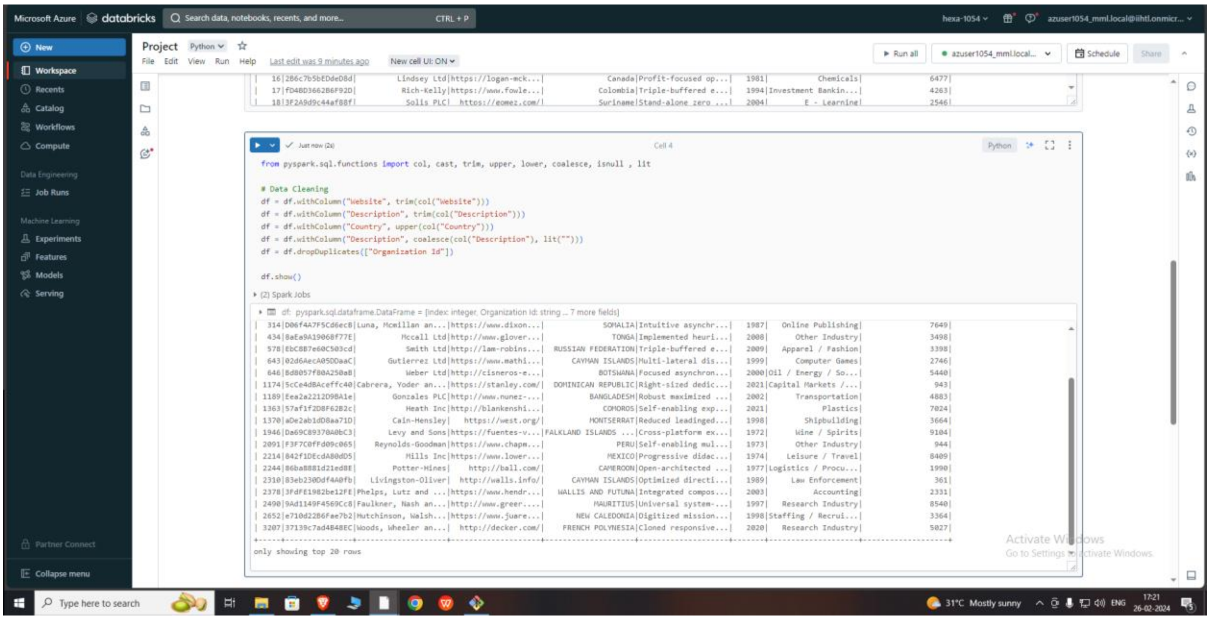
```
dbutils.fs.ls('/mnt/project')
```

```
[FileInfo(path='dbfs:/mnt/project/data.csv', name='data.csv', size=1381093, modificationTime=1708930288000)]
```

```
df = spark.read.csv('/mnt/project/data.csv',header='true',inferSchema='True')
df.show()
```

```
df:  pyspark.sql.dataframe.DataFrame = [Index: integer, Organization Id: string ... 7 more fields]
+-----+---------------+-------------------+-------------------+-------------------+-------------------+-------+-------------------+-------------------+
|Index|Organization Id|               Name|            Website|            Country|        Description|Founded|           Industry|Number of employees|
+-----+---------------+-------------------+-------------------+-------------------+-------------------+-------+-------------------+-------------------+
|    1|522816eF8fdBE6d|       Mckinney PLC|    http://soto.com/|          Sri Lanka|Synergized global...|   1988|              Dairy|               3930|
|    2|70C7FBD7e6Aa3Ea|     Cunningham LLC|http://harding-du...|            Namibia|Team-oriented fau...|   2018|            Library|               7871|
|    3|428B397eA2d7290|        Ruiz-Walls|http://www.atkins...|               Iran|Re-contextualized...|   2003|Hospital / Health...|               3095|
|    4|9D234Ae8Cc51C1c|Parrish, Osborne ...|http://salazar.info/|British Indian Oc...|Fully-configurabl...|   1989|       Supermarkets|               5422|
|    5|6CDCcdE3D0b7b44|Diaz, Robles and ...|https://www.brook...|           Botswana|Inverse intangibl...|   2013|     Nanotechnology|               3135|
|    6|cdAD9BBF227ADAE|          Keith PLC|http://www.mathew...|            Ecuador|Cross-group coher...|   1978|   Online Publishing|               7233|
|    7|0fe6F8Dd1Cf5e4d|      Humphrey-Myers|https://faulkner-...|       Sierra Leone|Adaptive directio...|   2005| Publishing Industry|               6022|
|    8|ECC0FBd0dc4aEfc|    Castaneda-Mcmahon|http://www.anthon...|           Zimbabwe|Front-line value-...|   2015|     Defense / Space|               4580|
|    9|e0E6cfAE68d08c9|       Santos-Bowman|https://www.oneil...|            Ecuador|Multi-layered opt...|   1979|  Computer Hardware|               3245|
|   10|A7DdBb239A86689|       Valdez-Estes|http://melendez-m...|        Timor-Leste|Diverse next gene...|   1985|           Plastics|               1785|
|   11|6a74D1bF2BC1d65|       Young-Zavala|http://www.ortiz....|            Ukraine|Sharable multimed...|   1972|              Music|               2985|
|   12|Cc5293Dbc2A92bE|        Shaffer Inc|https://www.sando...|     United Kingdom|Sharable coherent...|   1997|         Automotive|                839|
|   13|849CEAb2f65c439|        Gaines-Vang|http://www.levy.com/|           Thailand|Compatible web-en...|   2021|   Online Publishing|               3135|
```

- **Performing data preprocessing activity such trimming blank spaces from columns**

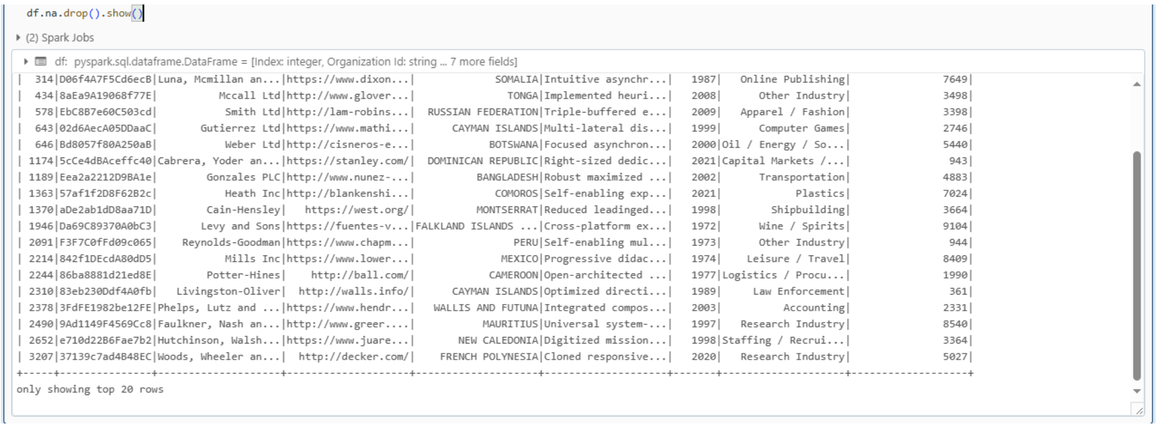


```
|   16|2B6c7b5bEDdeD8d|        Lindsey Ltd|https://logan-mck...|             Canada|Profit-focused op...|   1981|          Chemicals|               6477|
|   17|fD4BD3662B6F92D|         Rich-Kelly|https://www.fowle...|           Colombia|Triple-buffered e...|   1994|Investment Bankin...|               4263|
|   18|3F2A9d9c44af88f|          Solis PLC|  https://eomez.com/|           Suriname|Stand-alone zero ...|   2004|        E - Learning|               2546|
```

```
from pyspark.sql.functions import col, cast, trim, upper, lower, coalesce, isnull , lit

# Data Cleaning
df = df.withColumn("Website", trim(col("Website")))
df = df.withColumn("Description", trim(col("Description")))
df = df.withColumn("Country", upper(col("Country")))
df = df.withColumn("Description", coalesce(col("Description"), lit("")))
df = df.dropDuplicates(["Organization Id"])

df.show()
```

```
df:  pyspark.sql.dataframe.DataFrame = [Index: integer, Organization Id: string ... 7 more fields]
|  314|D06f4A7F5Cd6ecB|Luna, Mcmillan an...|https://www.dixon...|            SOMALIA|Intuitive asynchr...|   1987|   Online Publishing|               7649|
|  434|8aEa9A19068f77E|         Mccall Ltd|http://www.glover...|              TONGA|Implemented heuri...|   2008|      Other Industry|               3498|
|  578|EbC8B7e60C503cd|          Smith Ltd|http://lam-robins...| RUSSIAN FEDERATION|Triple-buffered e...|   2009|    Apparel / Fashion|               3398|
|  643|02d6AecA05DDaaC|      Gutierrez Ltd|https://www.mathi...|     CAYMAN ISLANDS|Multi-lateral dis...|   1999|      Computer Games|               2746|
|  646|Bd8057f80A250aB|          Weber Ltd|http://cisneros-e...|           BOTSWANA|Focused asynchron...|   2000|Oil / Energy / So...|               5440|
| 1174|5cCe4dBAceffc40|Cabrera, Yoder an...|https://stanley.com/| DOMINICAN REPUBLIC|Right-sized dedic...|   2021|Capital Markets /...|                943|
| 1189|Eea2a2212D9BA1e|       Gonzales PLC|http://www.nunez-...|         BANGLADESH|Robust maximized ...|   2002|      Transportation|               4883|
| 1363|57af1f2D8F62B2c|          Heath Inc|http://blankenshi...|            COMOROS|Self-enabling exp...|   2021|            Plastics|               7024|
| 1370|aDe2ab1dD8aa71D|       Cain-Hensley|   https://west.org/|         MONTSERRAT|Reduced leadinged...|   1998|        Shipbuilding|               3664|
| 1946|Da69C89370A0bC3|      Levy and Sons|https://fuentes-v...|FALKLAND ISLANDS ...|Cross-platform ex...|   1972|       Wine / Spirits|               9104|
| 2091|F3F7C0fFd09c065|    Reynolds-Goodman|https://www.chapm...|               PERU|Self-enabling mul...|   1973|      Other Industry|                944|
| 2214|842f1DEcdA80dD5|          Mills Inc|https://www.lower...|             MEXICO|Progressive didac...|   1974|     Leisure / Travel|               8409|
| 2244|86ba8881d21ed8E|       Potter-Hines|    http://ball.com/|           CAMEROON|Open-architected ...|   1977|Logistics / Procu...|               1990|
| 2310|83eb230Ddf4A0fb|    Livingston-Oliver|  http://walls.info/|     CAYMAN ISLANDS|Optimized directi...|   1989|     Law Enforcement|                361|
| 2378|3FdFE1982be12FE|Phelps, Lutz and ...|https://www.hendr...|  WALLIS AND FUTUNA|Integrated compos...|   2003|          Accounting|               2331|
| 2490|9Ad1149F4569Cc8|Faulkner, Nash an...|http://www.greer....|          MAURITIUS|Universal system-...|   1997|   Research Industry|               8540|
| 2652|e710d22B6Fae7b2|Hutchinson, Walsh...|https://www.juare...|      NEW CALEDONIA|Digitized mission...|   1998|Staffing / Recrui...|               3364|
| 3207|37139c7ad4B48EC|Woods, Wheeler an...|  http://decker.com/|   FRENCH POLYNESIA|Cloned responsive...|   2020|   Research Industry|               5027|
+-----+---------------+-------------------+-------------------+-------------------+-------------------+-------+-------------------+-------------------+
only showing top 20 rows
```

- **Drop rows with Null values**



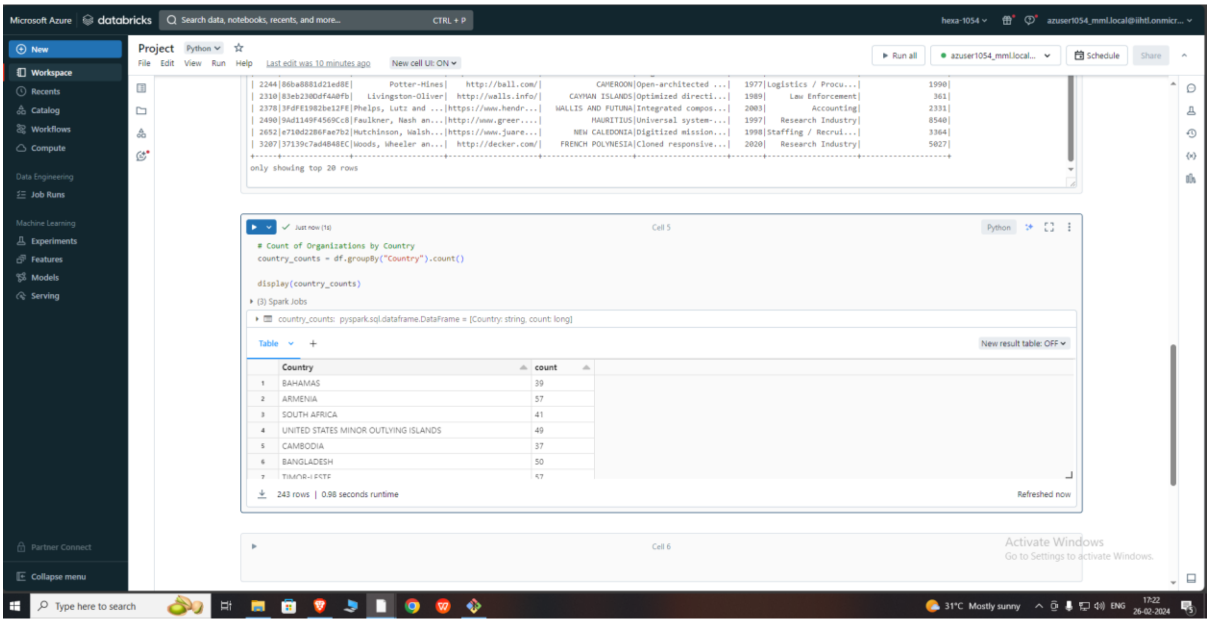
```
df.na.drop().show()
```

▸ (2) Spark Jobs

▸ df: pyspark.sql.dataframe.DataFrame = [Index: integer, Organization Id: string ... 7 more fields]

```
|  314|D06f4A7F5Cd6ecB|Luna, Mcmillan an...|https://www.dixon...|            SOMALIA|Intuitive asynchr...|  1987|   Online Publishing|              7649|
|  434|8aEa9A19068f77E|         Mccall Ltd|http://www.glover...|              TONGA|Implemented heuri...|  2008|      Other Industry|              3498|
|  578|EbC8B7e60C503cd|          Smith Ltd|http://lam-robins...| RUSSIAN FEDERATION|Triple-buffered e...|  2009|   Apparel / Fashion|              3398|
|  643|02d6AecA05DDaaC|      Gutierrez Ltd|https://www.mathi...|     CAYMAN ISLANDS|Multi-lateral dis...|  1999|      Computer Games|              2746|
|  646|Bd8057f80A250aB|          Weber Ltd|http://cisneros-e...|           BOTSWANA|Focused asynchron...|  2000|Oil / Energy / So...|              5440|
| 1174|5cCe4dBAceffc40|Cabrera, Yoder an...|https://stanley.com/| DOMINICAN REPUBLIC|Right-sized dedic...|  2021|Capital Markets /...|               943|
| 1189|Eea2a2212D9BA1e|       Gonzales PLC|http://www.nunez-...|         BANGLADESH|Robust maximized ...|  2002|      Transportation|              4883|
| 1363|57af1f2D8F62B2c|          Heath Inc|http://blankenshi...|            COMOROS|Self-enabling exp...|  2021|            Plastics|              7024|
| 1370|aDe2ab1dD8aa71D|       Cain-Hensley|   https://west.org/|         MONTSERRAT|Reduced leadinged...|  1998|        Shipbuilding|              3664|
| 1946|Da69C89370A0bC3|      Levy and Sons|https://fuentes-v...|FALKLAND ISLANDS ...|Cross-platform ex...|  1972|      Wine / Spirits|              9104|
| 2091|F3F7C0fFd09c065|   Reynolds-Goodman|https://www.chapm...|               PERU|Self-enabling mul...|  1973|      Other Industry|               944|
| 2214|842f1DEcdA80dD5|          Mills Inc|https://www.lower...|             MEXICO|Progressive didac...|  1974|    Leisure / Travel|              8409|
| 2244|86ba8881d21ed8E|       Potter-Hines|    http://ball.com/|           CAMEROON|Open-architected ...|  1977|Logistics / Procu...|              1990|
| 2310|83eb230Ddf4A0fb|   Livingston-Oliver|  http://walls.info/|     CAYMAN ISLANDS|Optimized directi...|  1989|     Law Enforcement|               361|
| 2378|3FdFE1982be12FE|Phelps, Lutz and ...|https://www.hendr...|  WALLIS AND FUTUNA|Integrated compos...|  2003|          Accounting|              2331|
| 2490|9Ad1149F4569Cc8|Faulkner, Nash an...|http://www.greer....|          MAURITIUS|Universal system-...|  1997|   Research Industry|              8540|
| 2652|e710d22B6Fae7b2|Hutchinson, Walsh...|https://www.juare...|      NEW CALEDONIA|Digitized mission...|  1998|Staffing / Recrui...|              3364|
| 3207|37139c7ad4B48EC|Woods, Wheeler an...|  http://decker.com/|   FRENCH POLYNESIA|Cloned responsive...|  2020|   Research Industry|              5027|
+-----+---------------+--------------------+--------------------+-------------------+--------------------+------+--------------------+------------------+
only showing top 20 rows
```

- **Count of Organisations by Country**



```
| 2244|86ba8881d21ed8E|       Potter-Hines|    http://ball.com/|           CAMEROON|Open-architected ...|  1977|Logistics / Procu...|              1990|
| 2310|83eb230Ddf4A0fb|   Livingston-Oliver|  http://walls.info/|     CAYMAN ISLANDS|Optimized directi...|  1989|     Law Enforcement|               361|
| 2378|3FdFE1982be12FE|Phelps, Lutz and ...|https://www.hendr...|  WALLIS AND FUTUNA|Integrated compos...|  2003|          Accounting|              2331|
| 2490|9Ad1149F4569Cc8|Faulkner, Nash an...|http://www.greer....|          MAURITIUS|Universal system-...|  1997|   Research Industry|              8540|
| 2652|e710d22B6Fae7b2|Hutchinson, Walsh...|https://www.juare...|      NEW CALEDONIA|Digitized mission...|  1998|Staffing / Recrui...|              3364|
| 3207|37139c7ad4B48EC|Woods, Wheeler an...|  http://decker.com/|   FRENCH POLYNESIA|Cloned responsive...|  2020|   Research Industry|              5027|
+-----+---------------+--------------------+--------------------+-------------------+--------------------+------+--------------------+------------------+
only showing top 20 rows
```

```
# Count of Organizations by Country
country_counts = df.groupBy("Country").count()

display(country_counts)
```

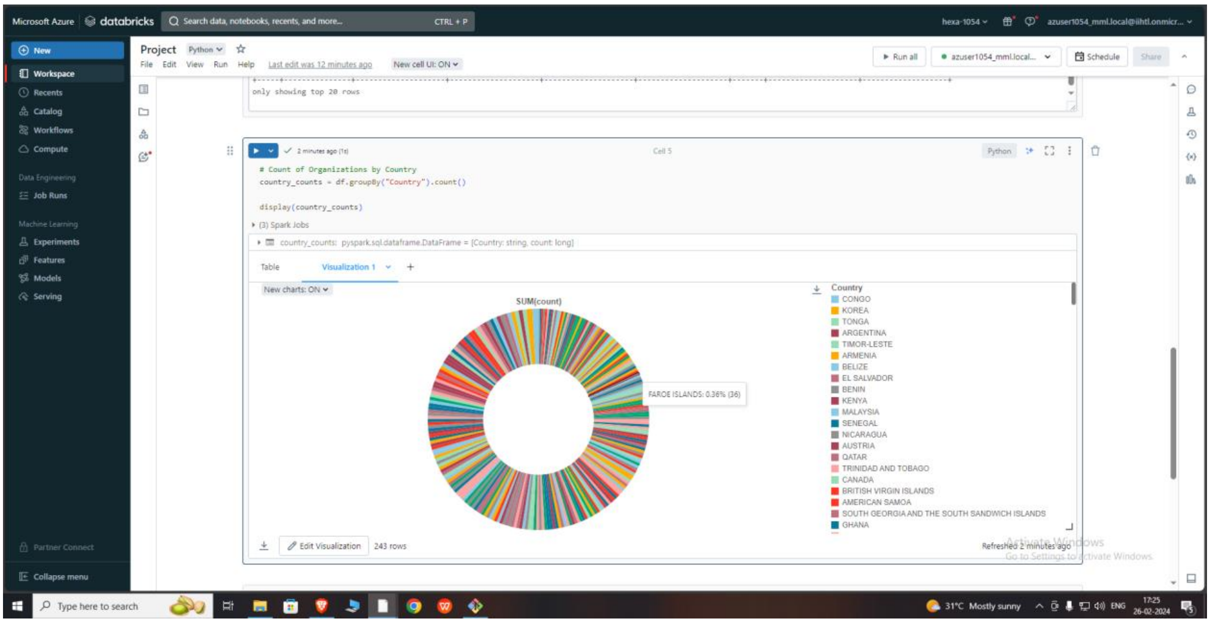
▸ (3) Spark Jobs

▸ country_counts: pyspark.sql.dataframe.DataFrame = [Country: string, count: long]

	Country	count
1	BAHAMAS	39
2	ARMENIA	57
3	SOUTH AFRICA	41
4	UNITED STATES MINOR OUTLYING ISLANDS	49
5	CAMBODIA	37
6	BANGLADESH	50
7	TIMOR-LESTE	57

243 rows | 0.98 seconds runtime

- **Visualising count of Organisation by Country using a pie chart**



```
only showing top 20 rows
```

Cell 5

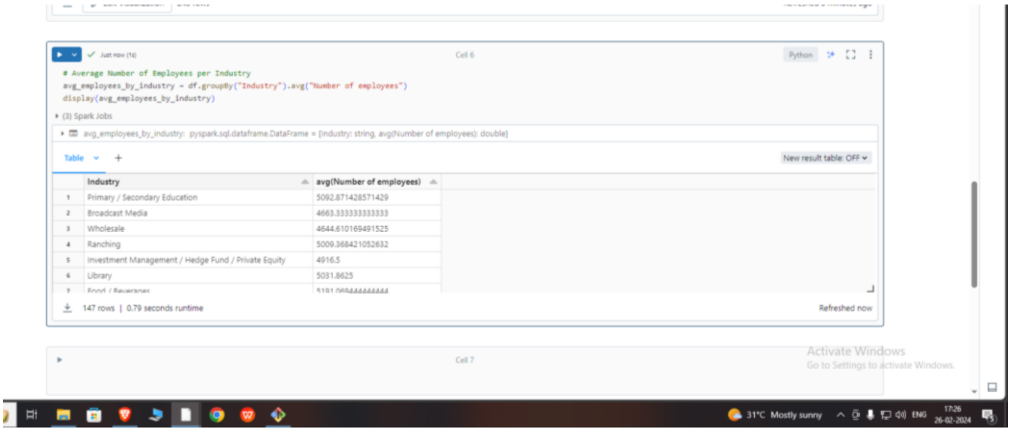
```
# Count of Organizations by Country
country_counts = df.groupBy("Country").count()

display(country_counts)
```

(3) Spark Jobs

country_counts: pyspark.sql.dataframe.DataFrame = [Country: string, count: long]

243 rows

- **Average number of Employees per industry**



Cell 6

```
# Average Number of Employees per Industry
avg_employees_by_industry = df.groupBy("Industry").avg("Number of employees")
display(avg_employees_by_industry)
```

(3) Spark Jobs

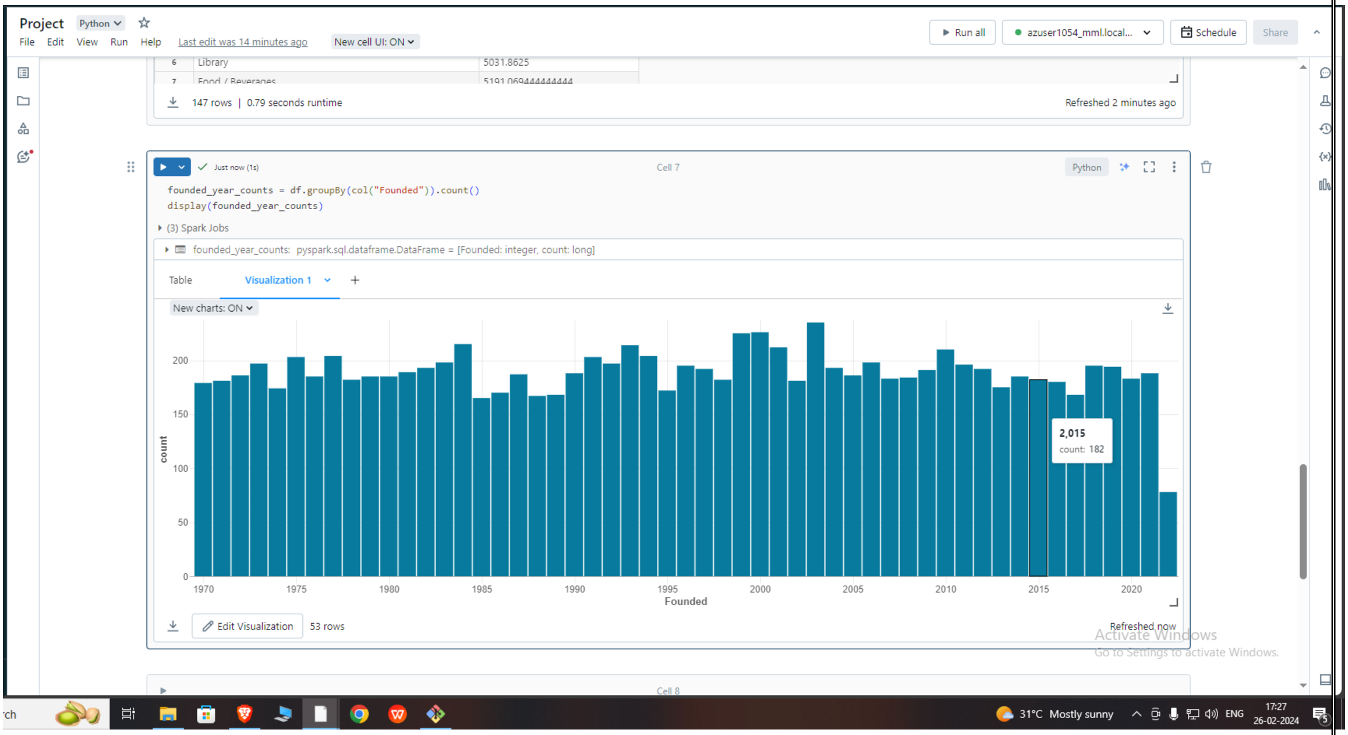
avg_employees_by_industry: pyspark.sql.dataframe.DataFrame = [Industry: string, avg(Number of employees): double]

	Industry	avg(Number of employees)
1	Primary / Secondary Education	5092.871428571429
2	Broadcast Media	4663.333333333333
3	Wholesale	4644.610169491525
4	Ranching	5009.368421052632
5	Investment Management / Hedge Fund / Private Equity	4916.5
6	Library	5031.8625
7	Food / Beverages	5191.069444444444

147 rows | 0.79 seconds runtime

Refreshed now

- **Visualising count of country grouped by the year in which they were founded using a bar graph**



6	Library	5031.8625
7	Food / Beverages	5191.069444444444

147 rows | 0.79 seconds runtime

Refreshed 2 minutes ago

```
founded_year_counts = df.groupBy(col("Founded")).count()
display(founded_year_counts)
```

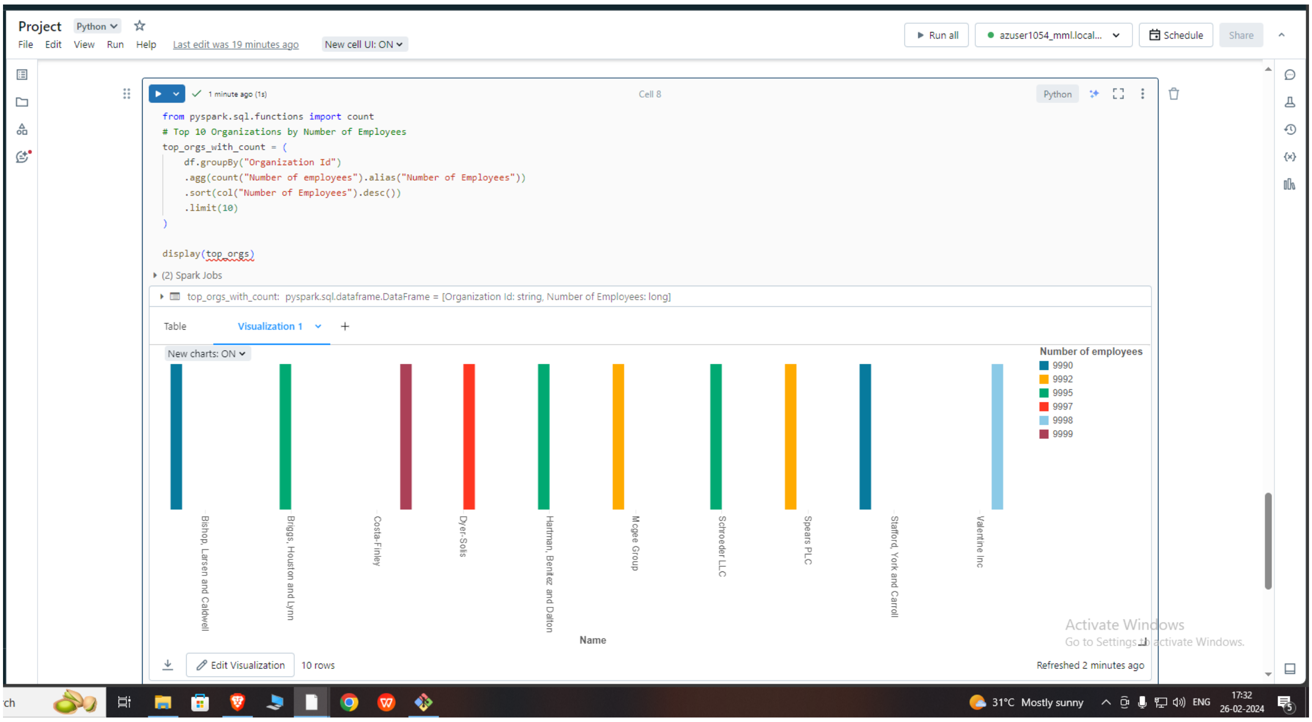
(3) Spark Jobs

founded_year_counts: pyspark.sql.dataframe.DataFrame = [Founded: integer, count: long]

53 rows

Refreshed now

- **Top 10 Organisations sorted by the number of employees**



```
from pyspark.sql.functions import count
# Top 10 Organizations by Number of Employees
top_orgs_with_count = (
    df.groupBy("Organization Id")
    .agg(count("Number of employees").alias("Number of Employees"))
    .sort(col("Number of Employees").desc())
    .limit(10)
)

display(top_orgs)
```

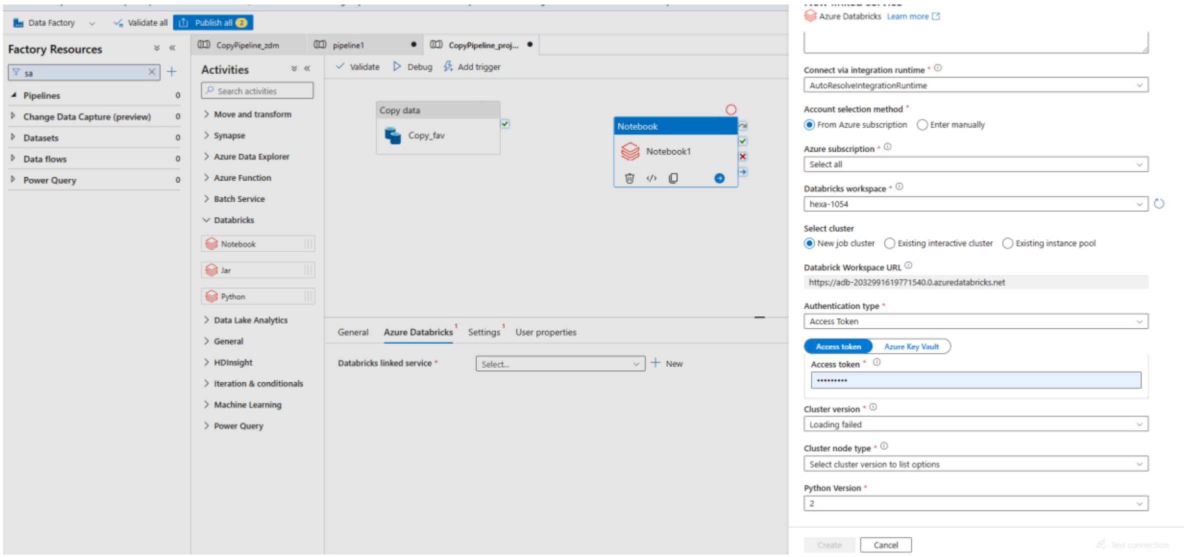
(2) Spark Jobs

top_orgs_with_count: pyspark.sql.dataframe.DataFrame = [Organization Id: string, Number of Employees: long]

10 rows

Refreshed 2 minutes ago

- **In Azure Datafactory drag and drop the Notebook activity and link it to the databricks Notebook**



Connect via integration runtime *: AutoResolveIntegrationRuntime

Account selection method *: ● From Azure subscription ○ Enter manually

Azure subscription *: Select all

Databricks workspace *: hexa-1054

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

Databrick Workspace URL: https://adb-2032991619771540.0.azuredatabricks.net

Authentication type *: Access Token

Access token | Azure Key Vault

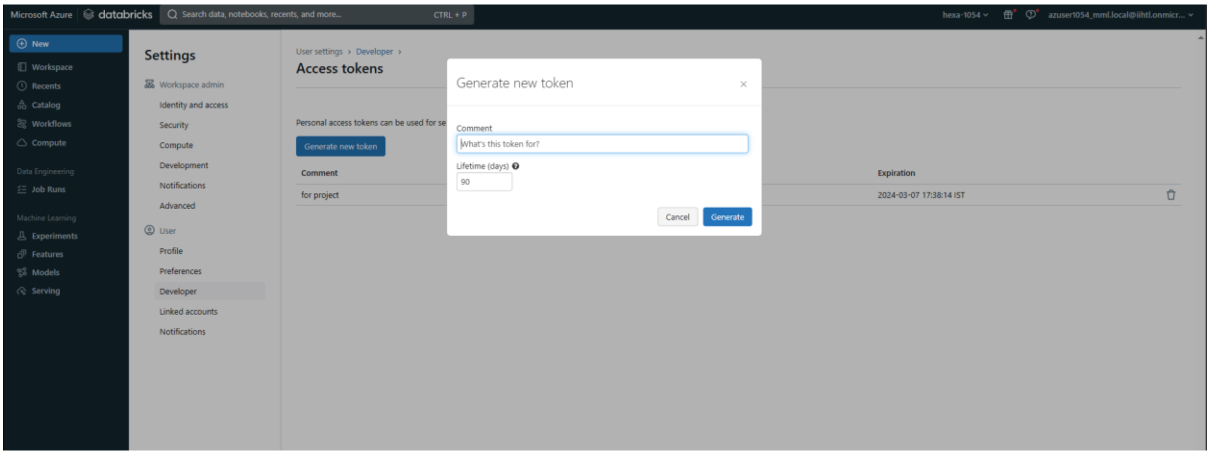
Access token *: ••••••••••

Cluster version *: Loading failed

Cluster node type *: Select cluster version to list options

Python Version *: 2

- **Copy the access token from Databricks Notebook by clicking on User settings -> Developer -> Generate new Token**



Access tokens

Personal access tokens can be used for se

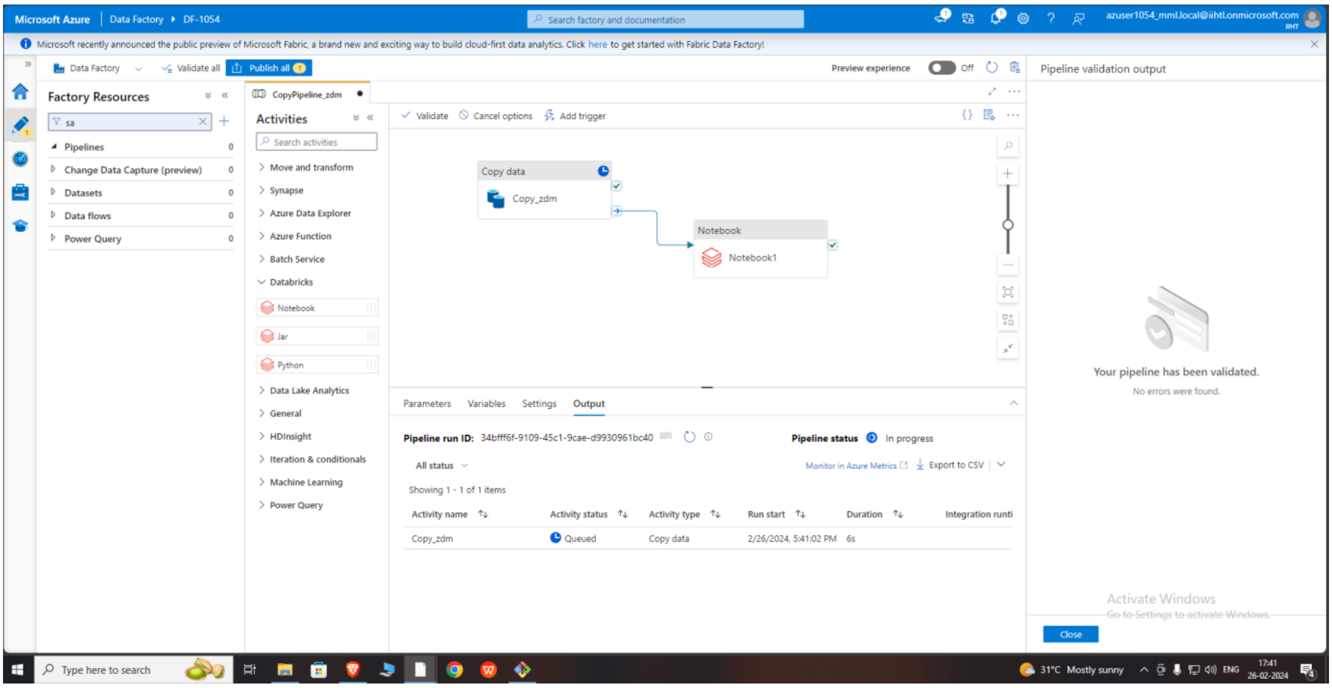
Generate new token

Comment: What's this token for?

Lifetime (days): 90

Comment	Expiration
for project	2024-03-07 17:38:14 IST

- **In Datafactory Validate the pipeline and then click on debug**



Microsoft recently announced the public preview of Microsoft Fabric, a brand new and exciting way to build cloud-first data analytics. Click here to get started with Fabric Data Factory!

Pipeline run ID: 34bfff6f-9109-45c1-9cae-d9930961bc40

Pipeline status: In progress

All status

Showing 1 - 1 of 1 items

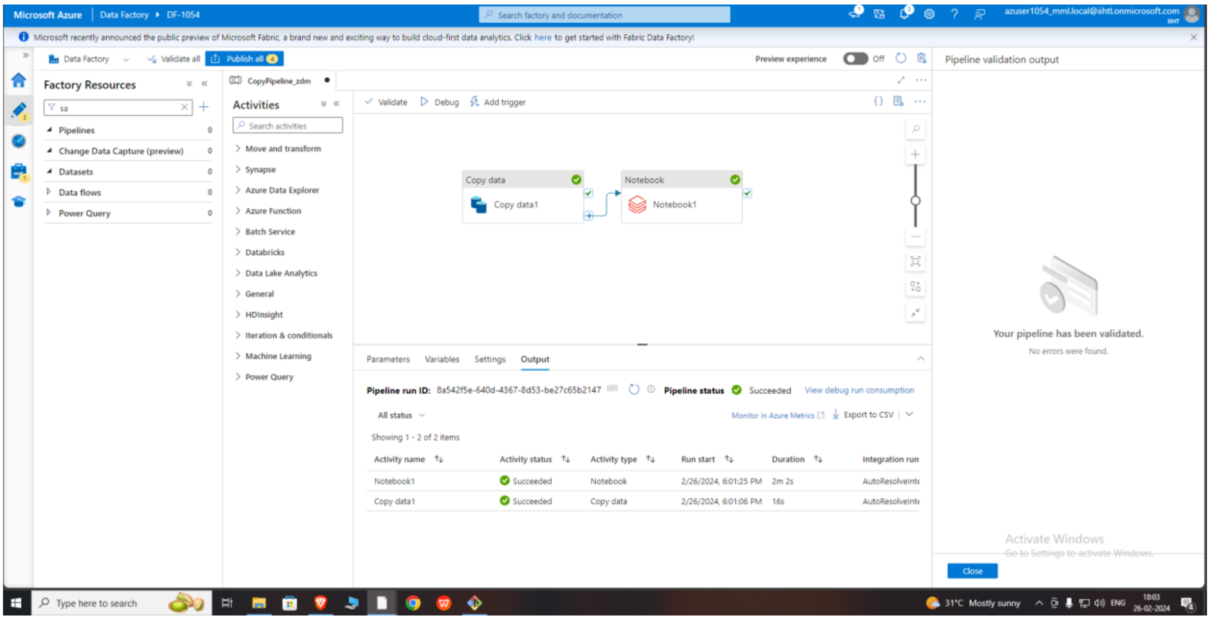
Activity name	Activity status	Activity type	Run start	Duration	Integration runti
Copy_zdm	Queued	Copy data	2/26/2024, 5:41:02 PM	6s	

Pipeline validation output

Your pipeline has been validated.

No errors were found.

- **The pipeline is running successfully**



Microsoft recently announced the public preview of Microsoft Fabric, a brand new and exciting way to build cloud-first data analytics. Click here to get started with Fabric Data Factory!

Pipeline run ID: 8a542f5e-640d-4367-8d53-be27c65b2147

Pipeline status: Succeeded

All status

Showing 1 - 2 of 2 items

Activity name	Activity status	Activity type	Run start	Duration	Integration run
Notebook1	Succeeded	Notebook	2/26/2024, 6:01:25 PM	2m 2s	AutoResolveInte
Copy data1	Succeeded	Copy data	2/26/2024, 6:01:06 PM	16s	AutoResolveInte

Pipeline validation output

Your pipeline has been validated.

No errors were found.

Conclusion

In conclusion, this project successfully demonstrated the integration of various Azure services to create an end-to-end data processing pipeline. Beginning with data acquisition from GitHub, we utilized Azure Data Factory to orchestrate data movement to Azure Blob Storage. Subsequently, Azure Databricks was employed for data processing and analysis, including preprocessing tasks and generating insightful visualizations. The seamless integration of these Azure services facilitated efficient data handling and analysis, showcasing the power of cloud-based data solutions in modern data engineering workflows. This project highlights the effectiveness of Azure's ecosystem in enabling scalable and efficient data processing pipelines for diverse use cases.