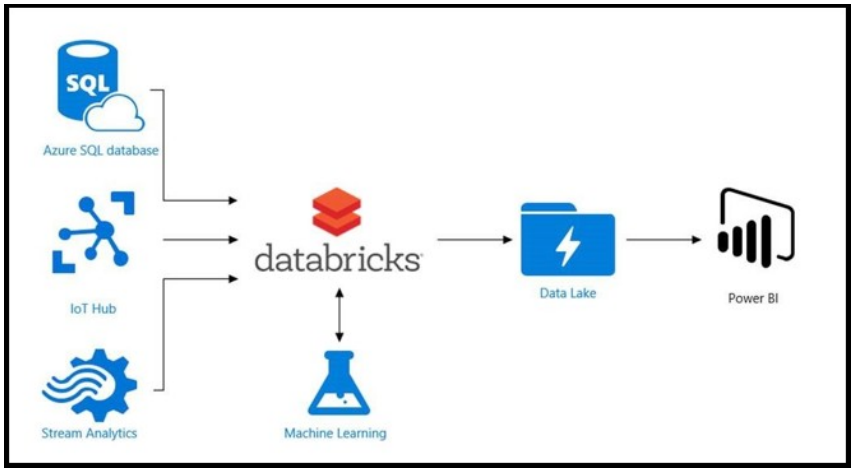
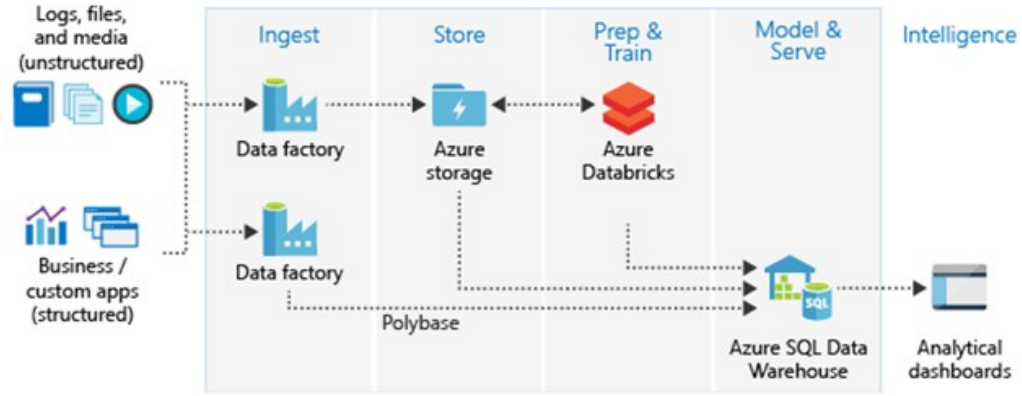
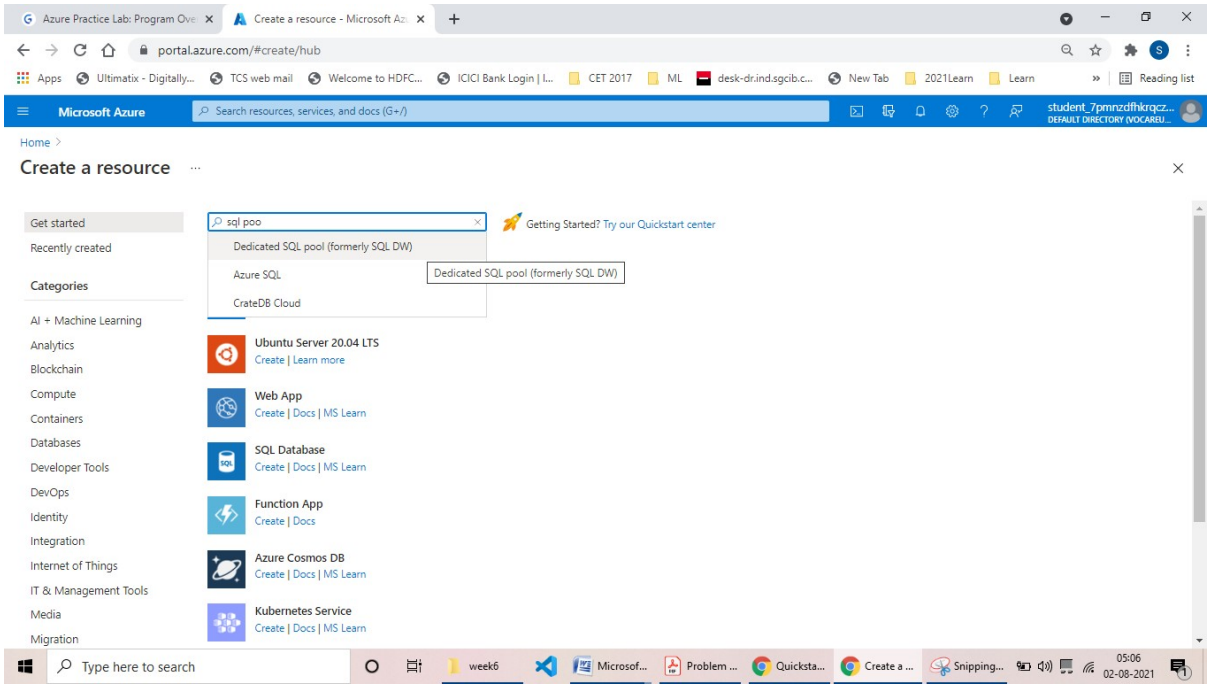
Contents

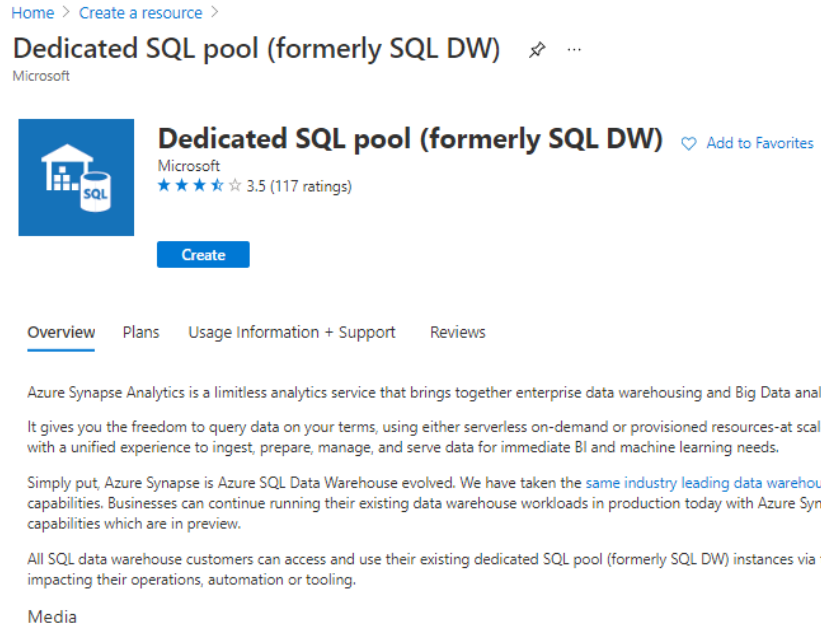
AZURE DATABRICKS



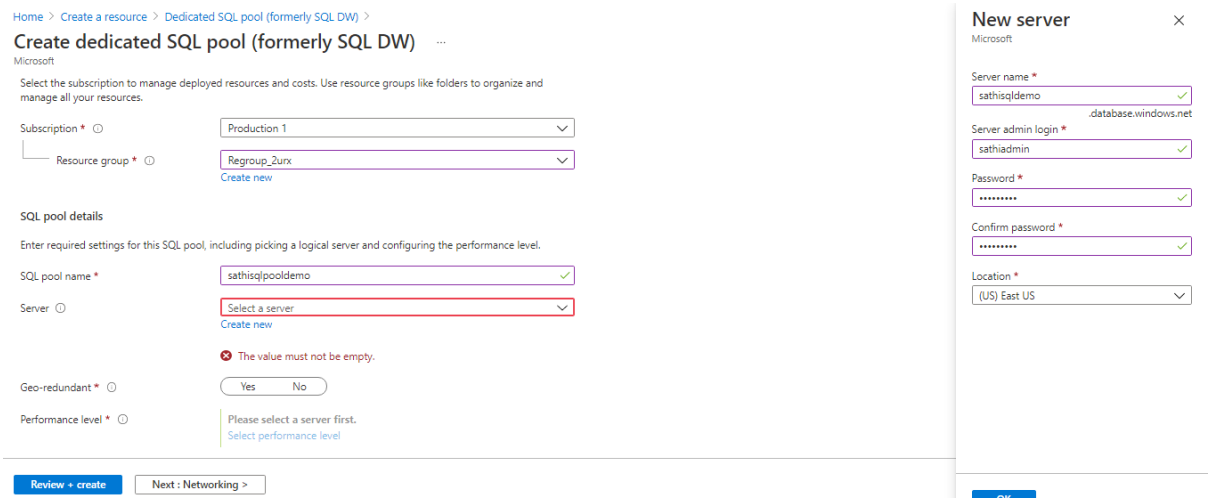


Create an Azure Synapse SQL Pool –Dedicated Sql pool





Create new server



Create dedicated SQL pool (formerly SQL DW) ...

Microsoft

Configure network access and connectivity for your server. The configuration selected below will apply to the selected server 'sathisqldemo' and all databases it manages. Learn more

Network connectivity

Choose an option for configuring connectivity to your server via public endpoint or private endpoint. Choosing no access creates with defaults and you can configure connection method after server creation. Learn more

Connectivity method * ⓘ

- ○ No access
- ◉ Public endpoint
- ○ Private endpoint

Firewall rules

Setting 'Allow Azure services and resources to access this server' to Yes allows communications from all resources inside the Azure boundary, that may or may not be part of your subscription. Learn more
Setting 'Add current client IP address' to Yes will add an entry for your client IP address to the server firewall.

Allow Azure services and resources to access this server * — No | Yes

Add current client IP address * — No | Yes

Create dedicated SQL pool (formerly SQL DW) ...

Microsoft

*Basics *Networking *Additional settings Tags Review + create

Customize additional configuration parameters including collation & data source.

Data source

Start with a blank SQL pool, restore from a backup or select sample data to populate your new database.

Use existing data * — None | Backup | Sample

AdventureWorksDW will be created as the sample database.

SQL pool collation

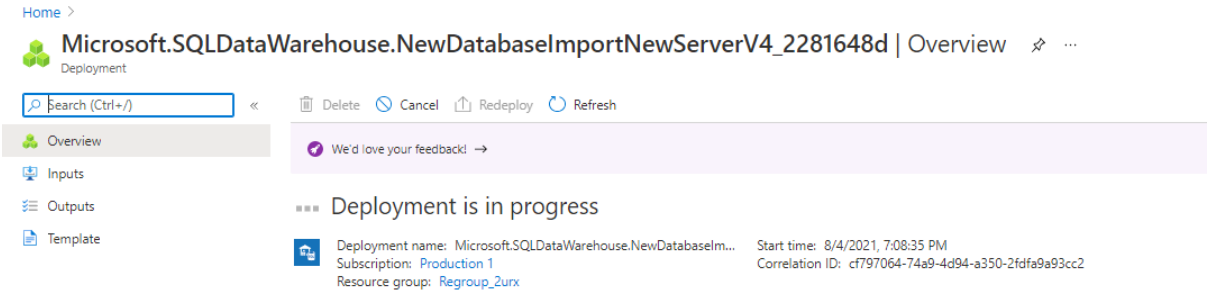
Collation defines the rules that sort and compare data, and cannot be changed after SQL pool creation. The default collation is SQL_Latin1_General_CP1_CI_AS. Learn more

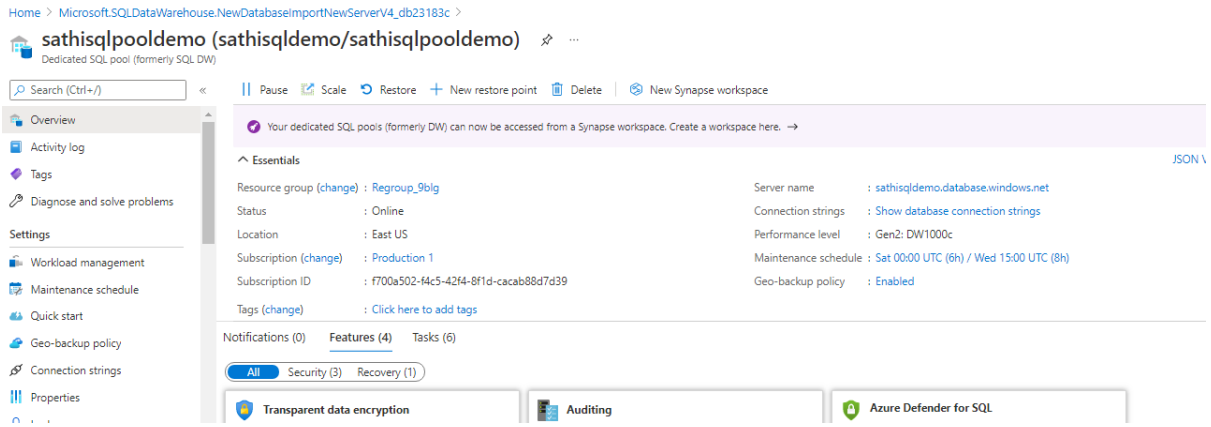
Collation ⓘ — SQL_Latin1_General_CP1_CI_AS

Azure Defender for SQL

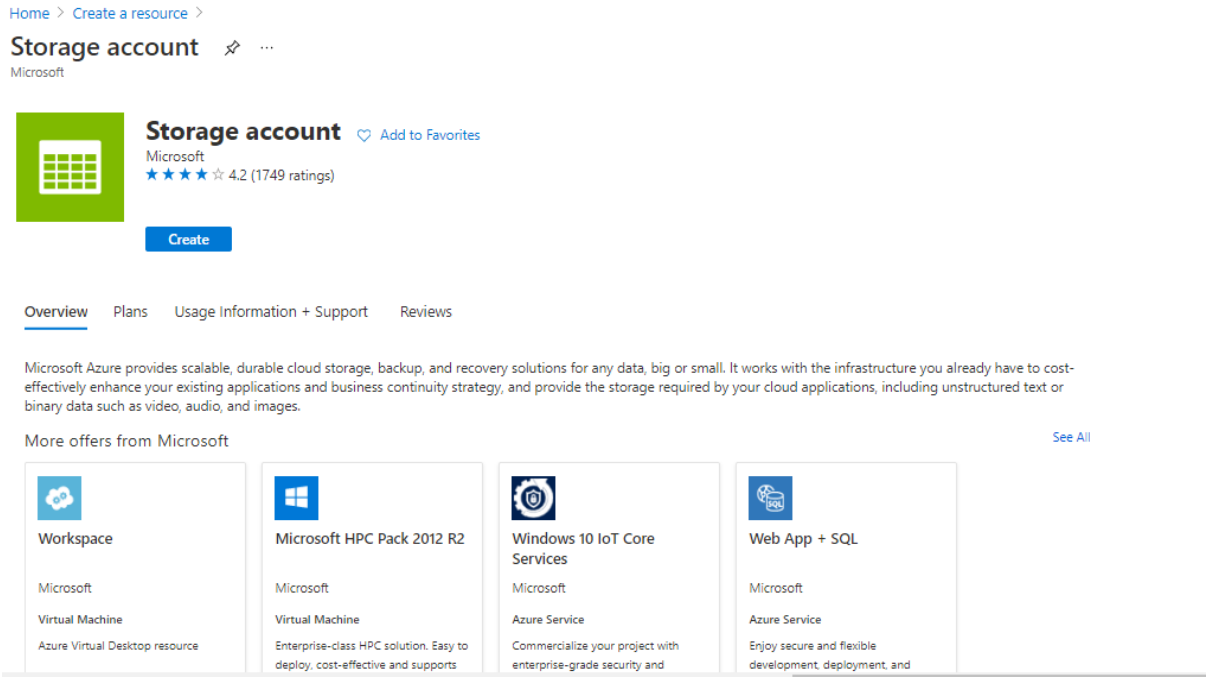
Protect your data using Azure Defender for SQL, a unified security package including vulnerability assessment and

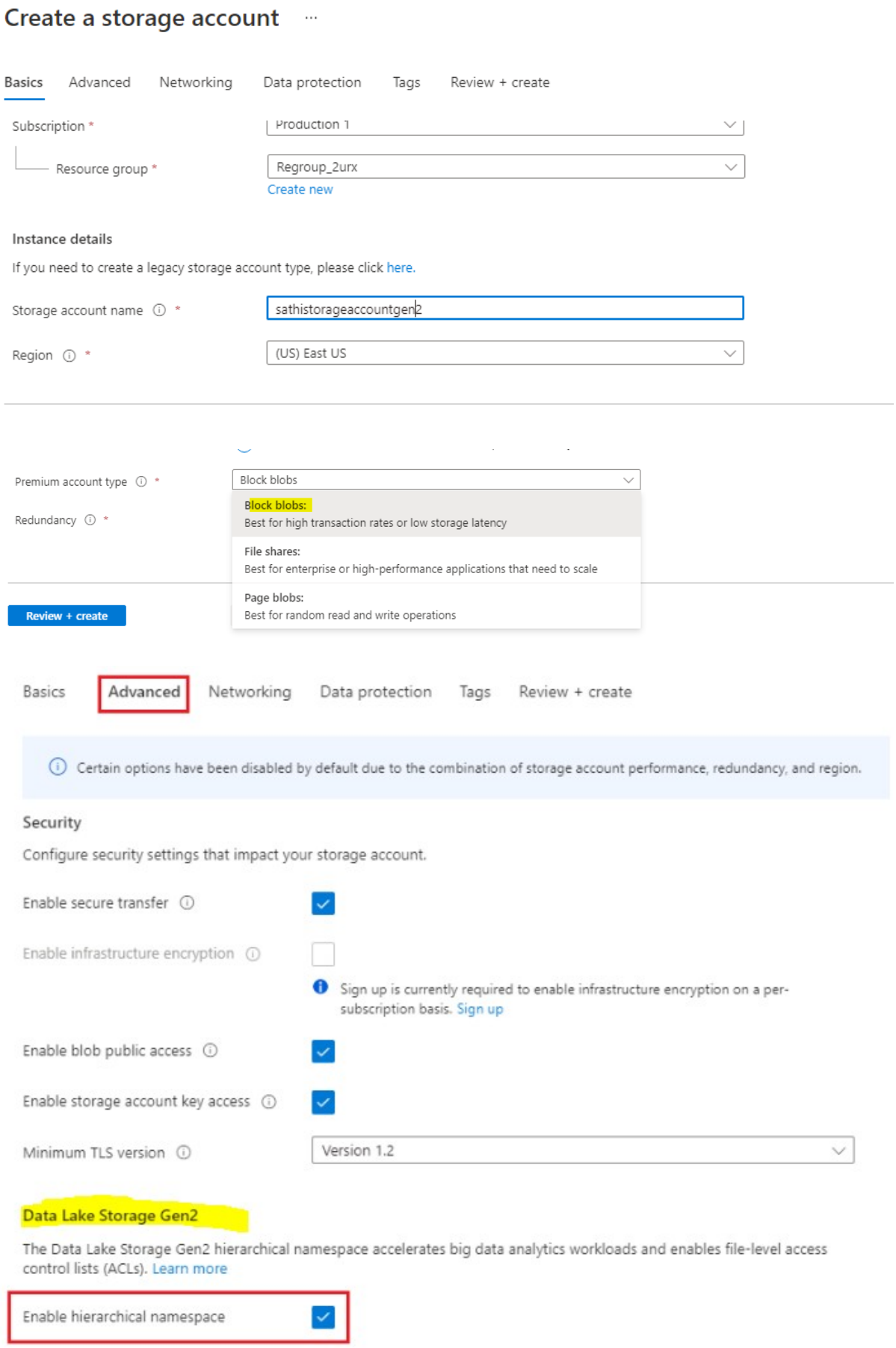
Review + create | < Previous | Next : Tags >

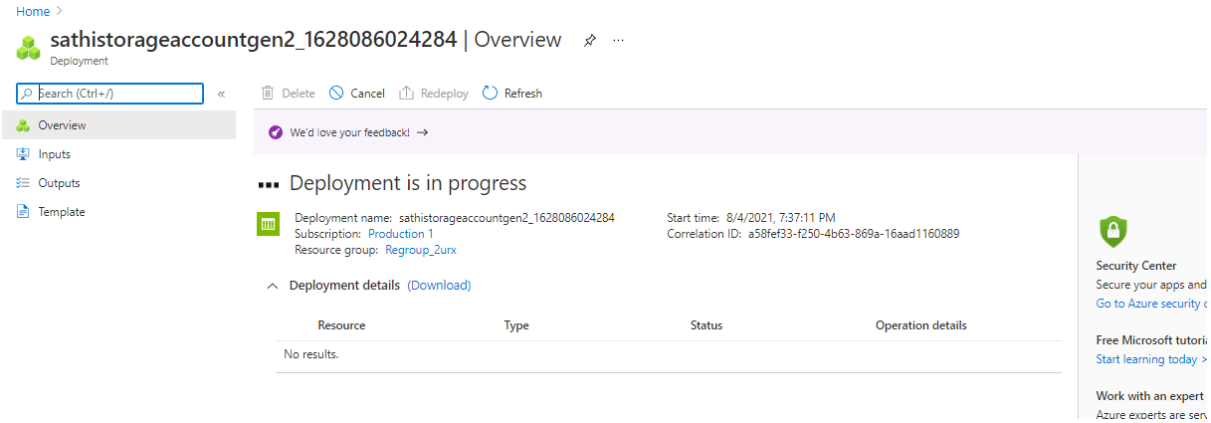




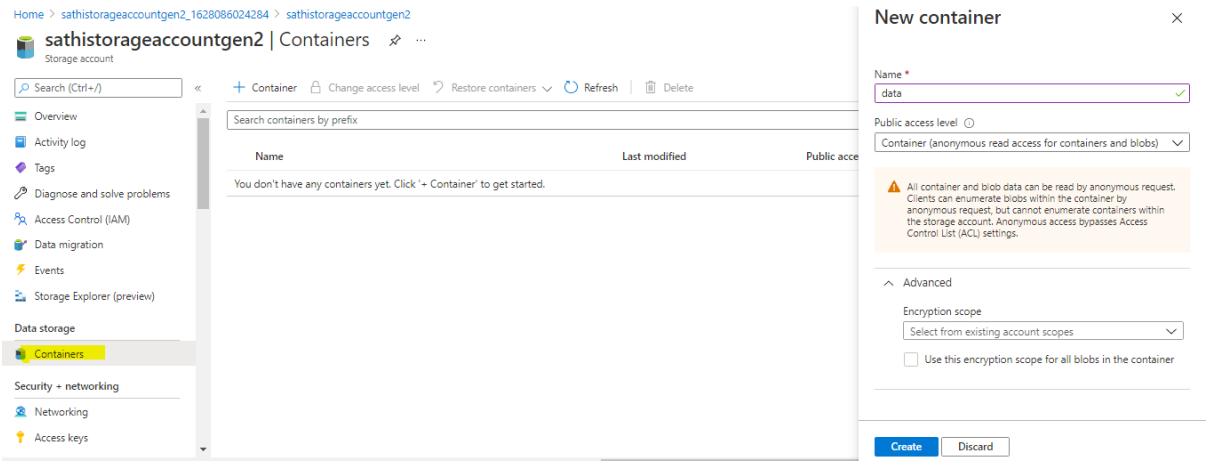
Create An Azure Blob storage account-Gen2 Data lake




Create a storage account
Basics
Advanced
Networking
Data protection
Tags
Review + create
Subscription *
Production 1
Resource group *
Regroup_2urx
Create new
Instance details
If you need to create a legacy storage account type, please click here.
Storage account name *
sathistorageaccountgen2
Region *
(US) East US
Premium account type *
Block blobs
Redundancy *
Block blobs:
Best for high transaction rates or low storage latency
File shares:
Best for enterprise or high-performance applications that need to scale
Page blobs:
Best for random read and write operations
Review + create
Basics
Advanced
Networking
Data protection
Tags
Review + create
Certain options have been disabled by default due to the combination of storage account performance, redundancy, and region.
Security
Configure security settings that impact your storage account.
Enable secure transfer
Enable infrastructure encryption
Sign up is currently required to enable infrastructure encryption on a per-subscription basis. Sign up
Enable blob public access
Enable storage account key access
Minimum TLS version
Version 1.2
Data Lake Storage Gen2
The Data Lake Storage Gen2 hierarchical namespace accelerates big data analytics workloads and enables file-level access control lists (ACLs). Learn more
Enable hierarchical namespace

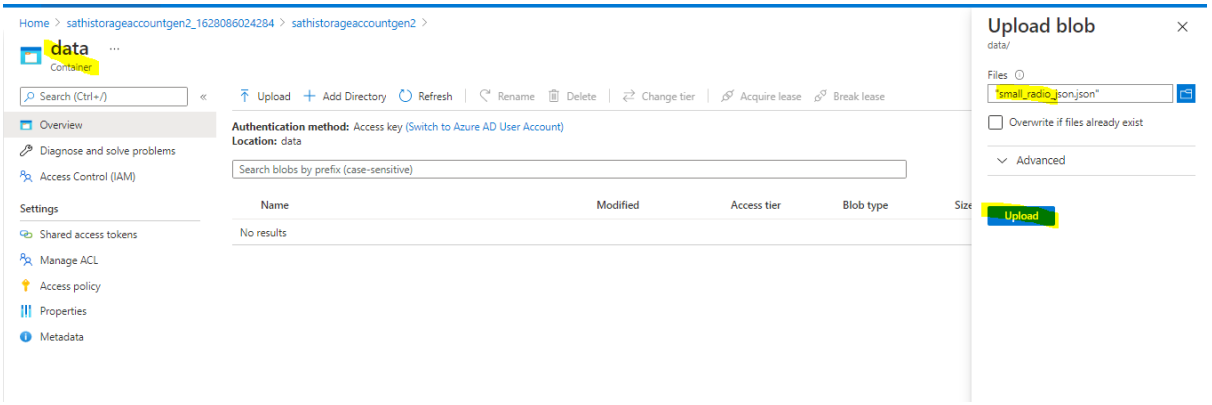


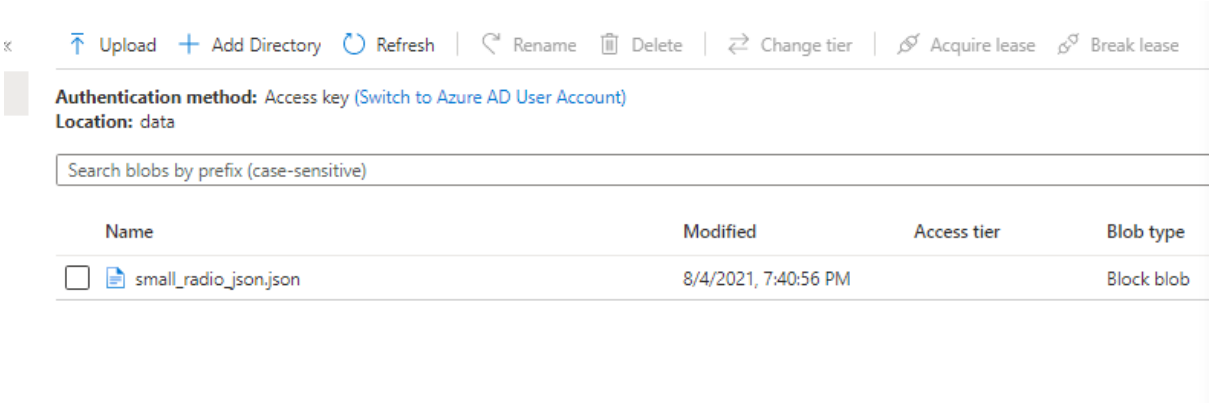
Create a container and upload json file



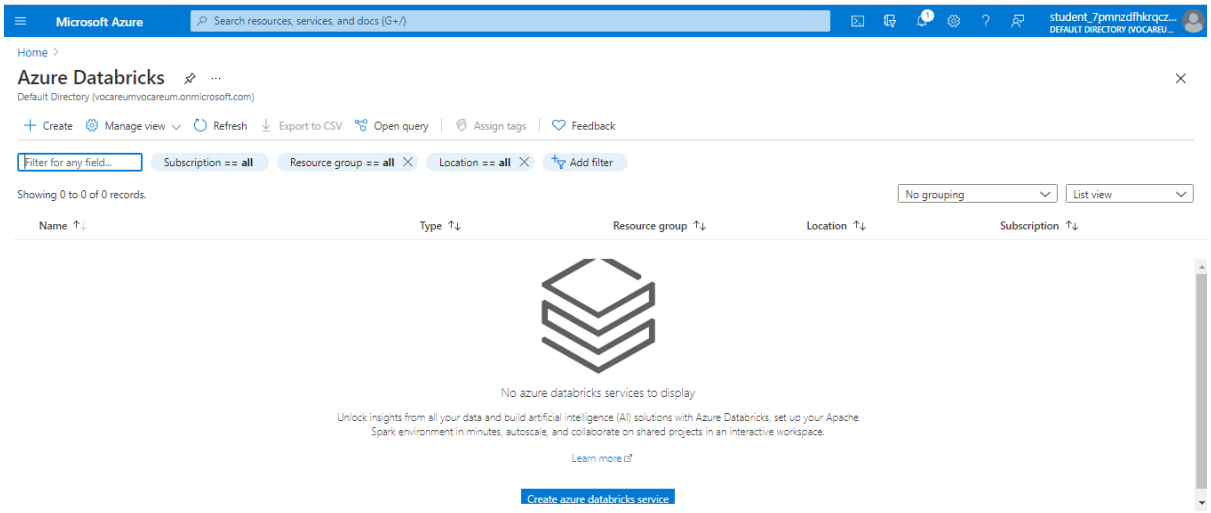
Load the following dataset into the storage account

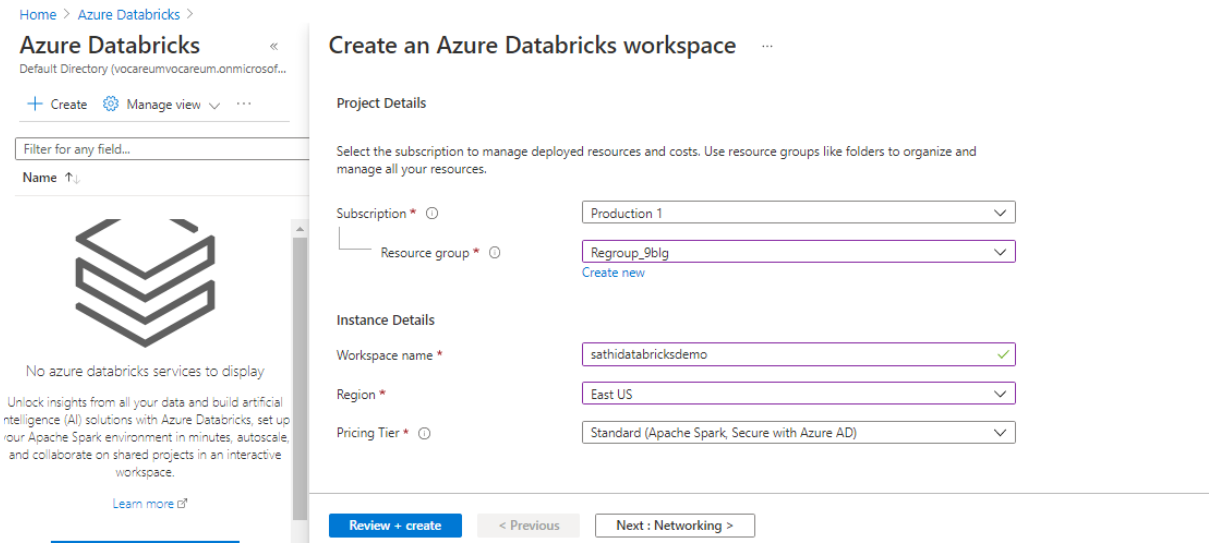
https://raw.githubusercontent.com/Azure/usql/master/Examples/Samples/Data/json/radiowebsite/small_radio_json.json

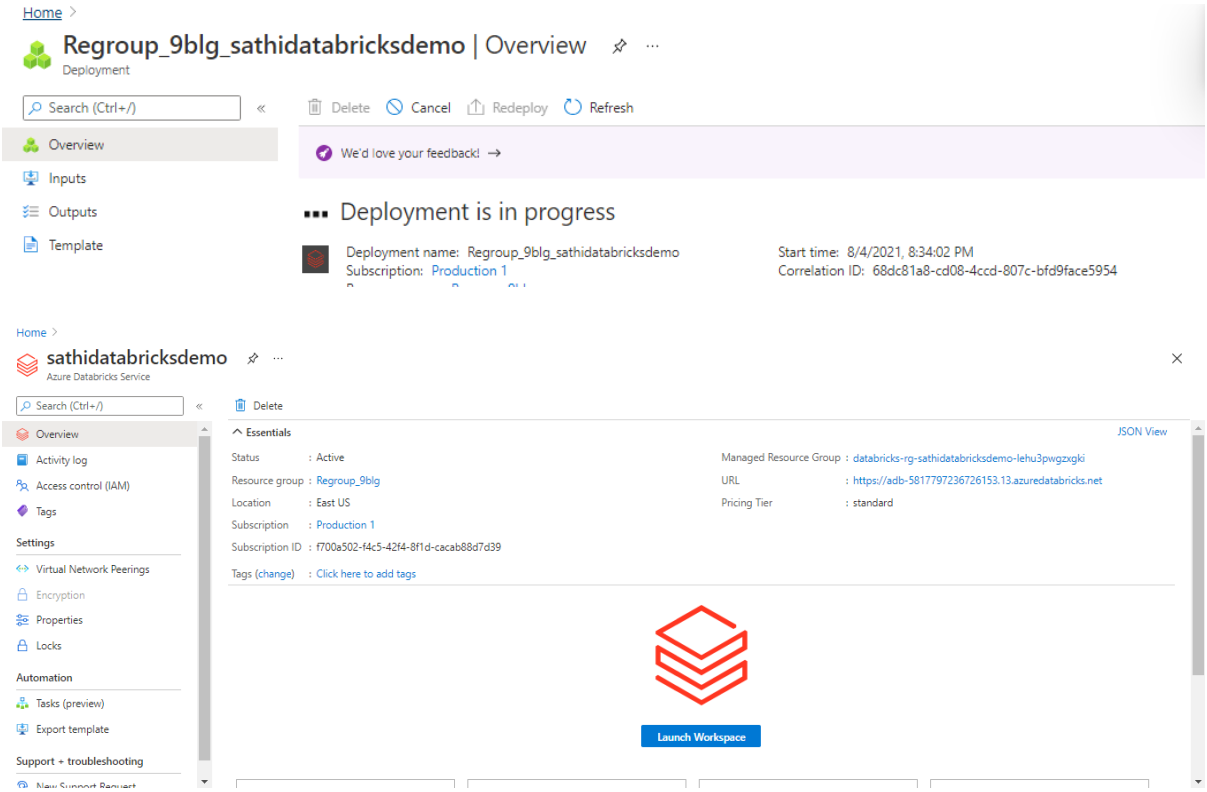




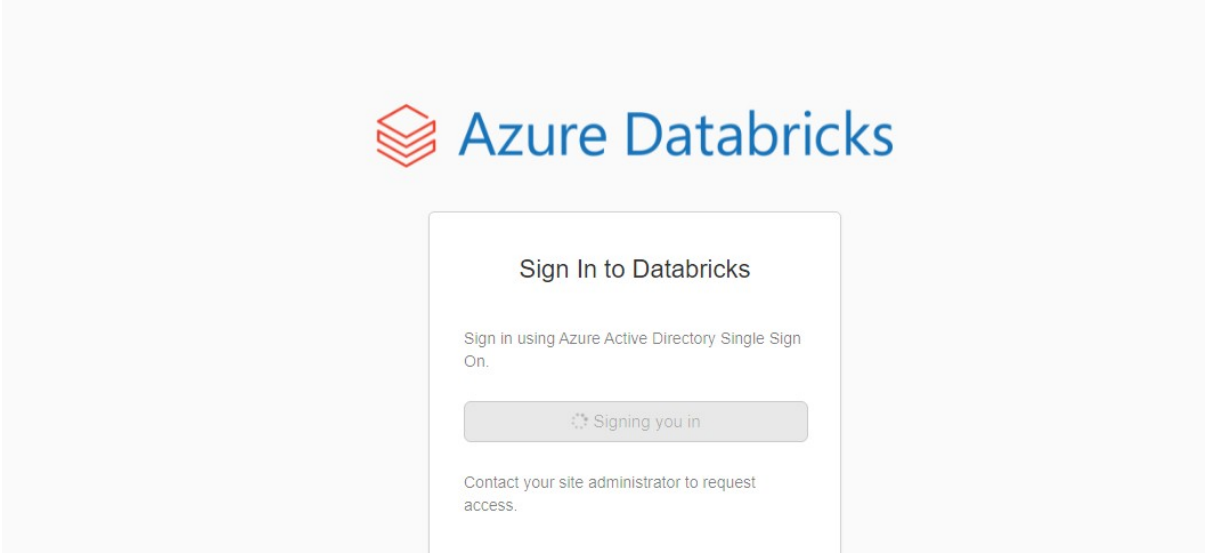
Create Azure DataBricks

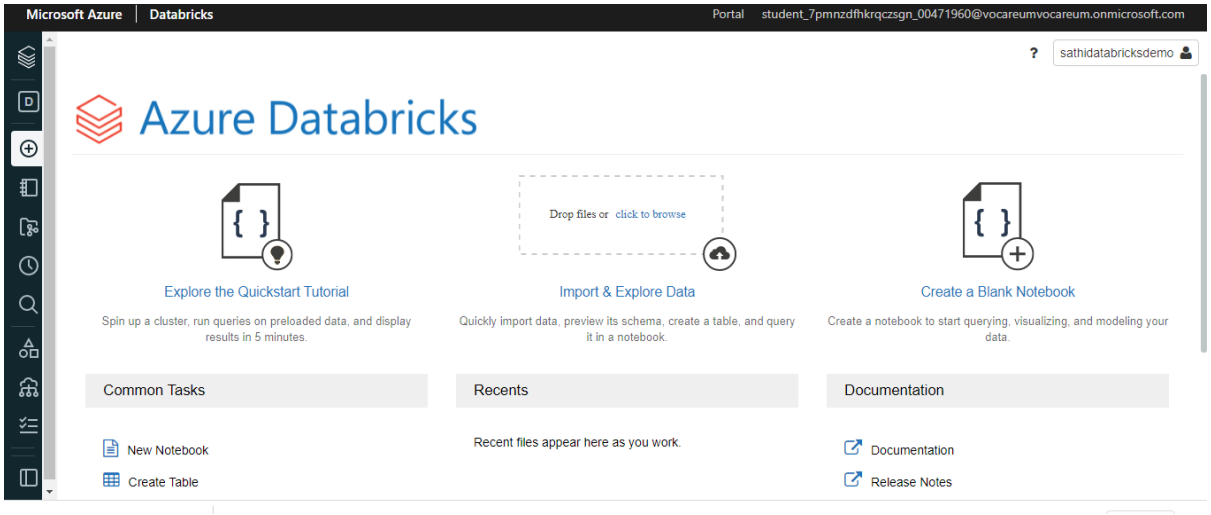




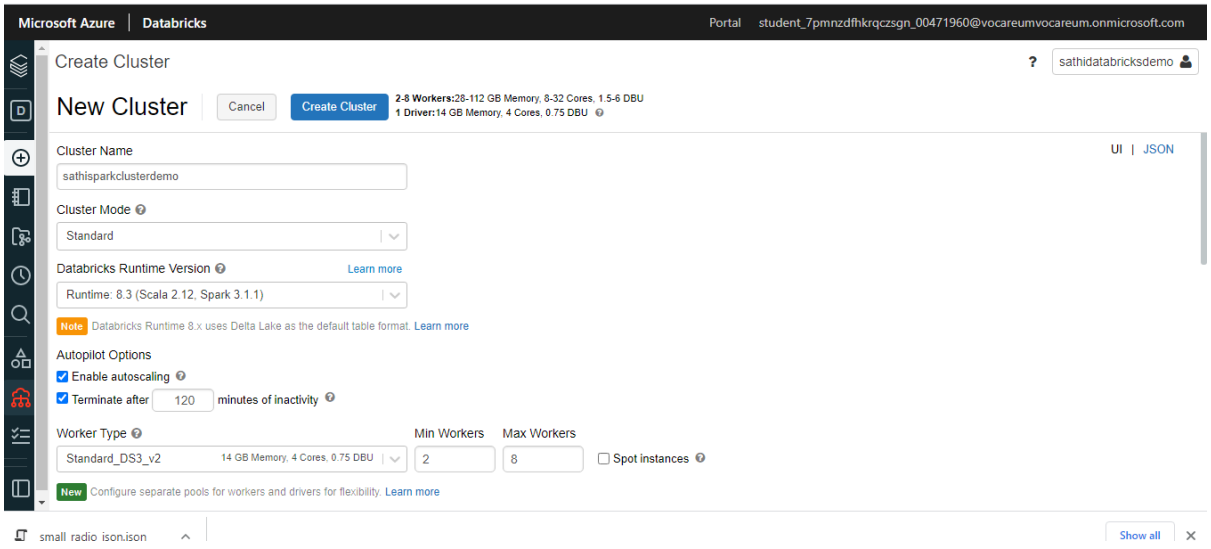


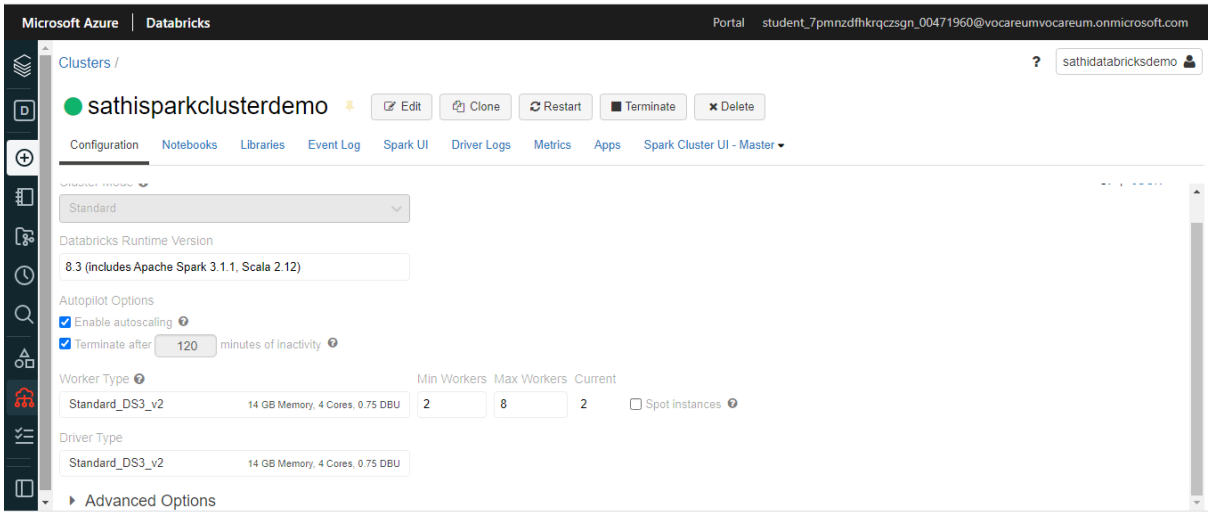
1. In the Azure portal, go to the Databricks workspace that you created, and then click **Launch Workspace**.



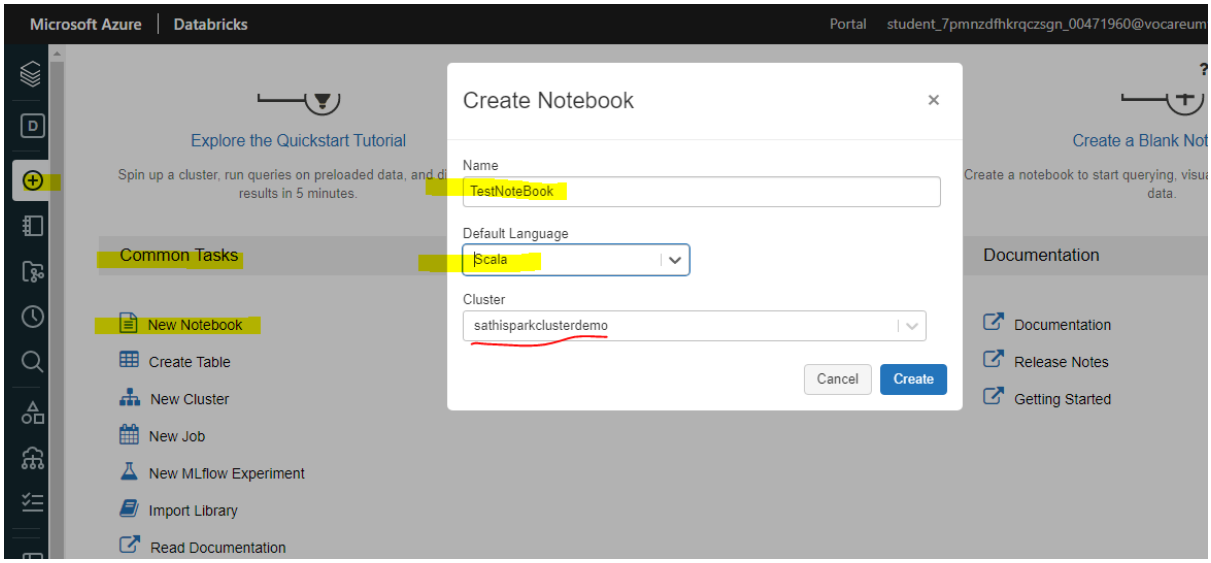


From redirected to the Azure Databricks portal. Click **New Cluster**.
Select **Create cluster**. Once the cluster is running, you can attach notebooks to the cluster and run Spark jobs



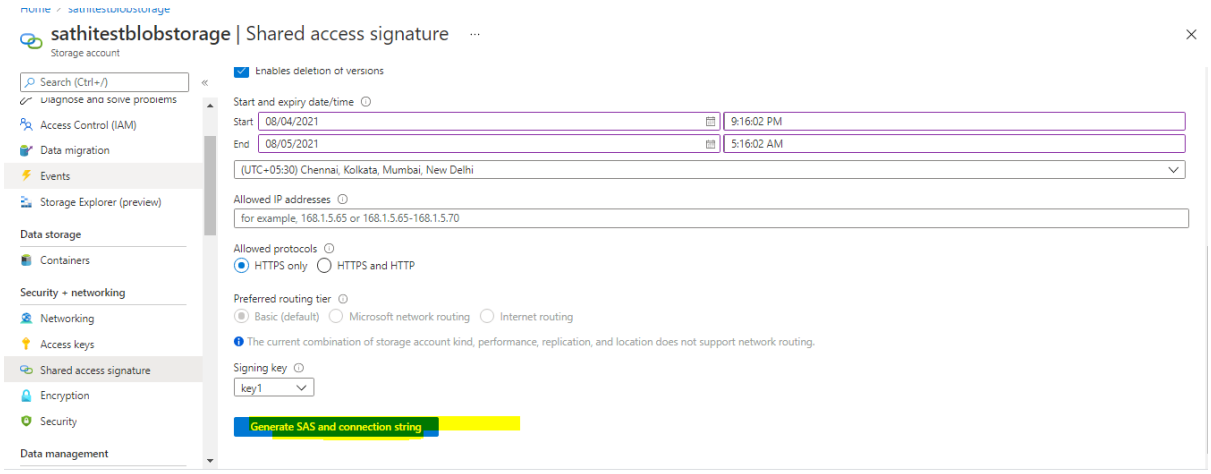


In the left pane, select **Azure Databricks**. From the **Common Tasks**, select **New Notebook**. Select the langugage that you wish to work on, later point of time we have option to change same

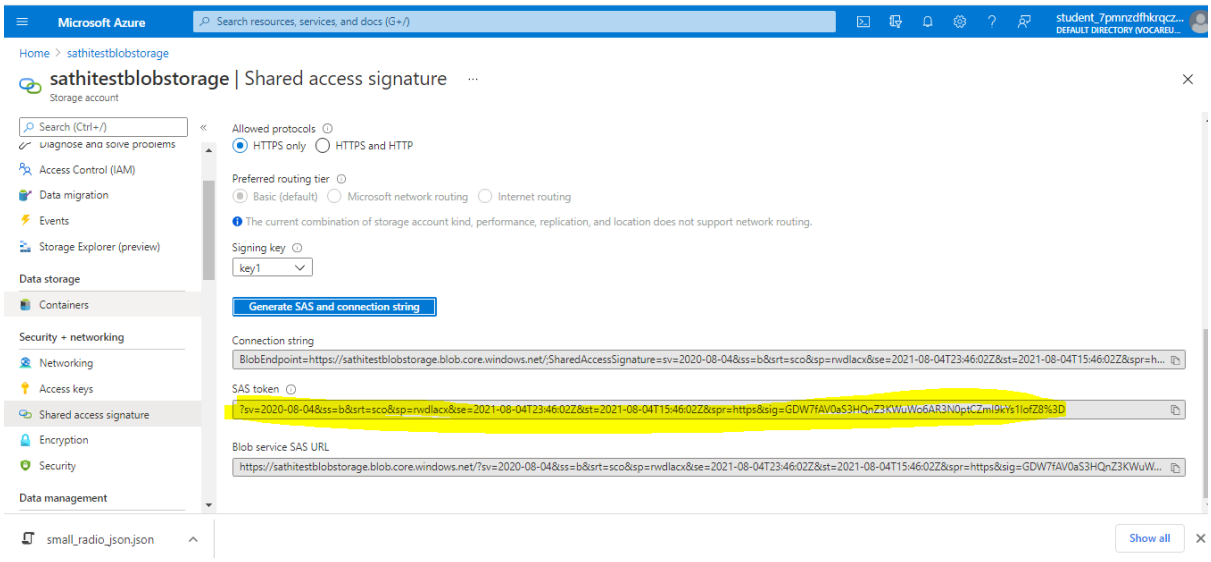


Generate sas token in storage account already created above , this is required as part of storage connection from spark cluster

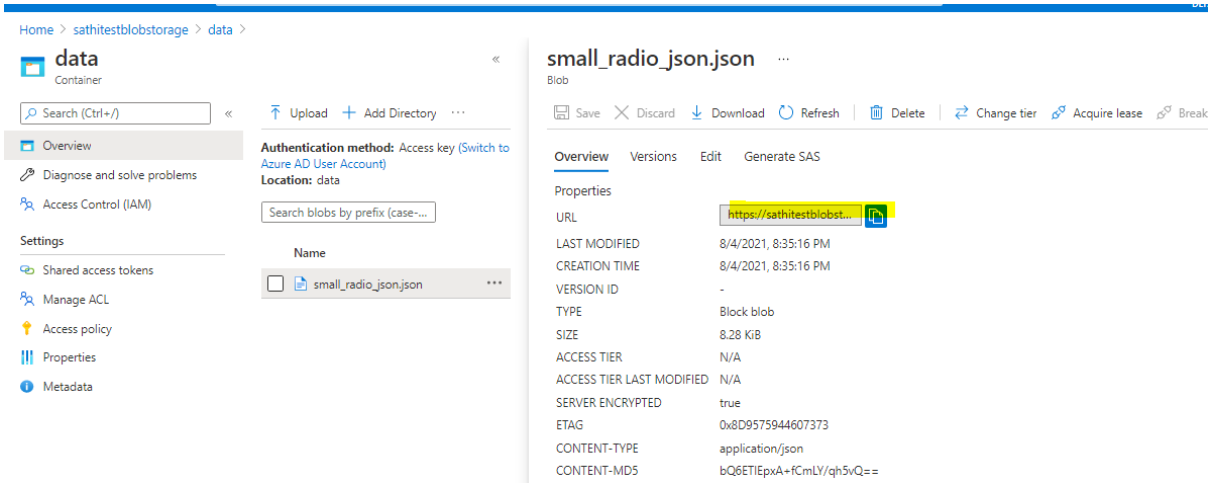
https://adamtheautomator.com/azure-sas-token/



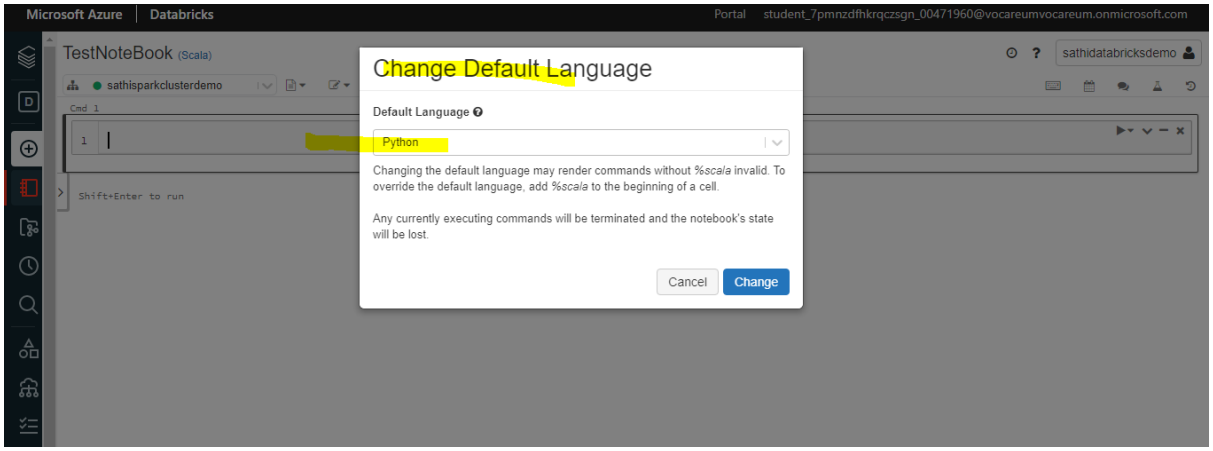
Copy for future refrence

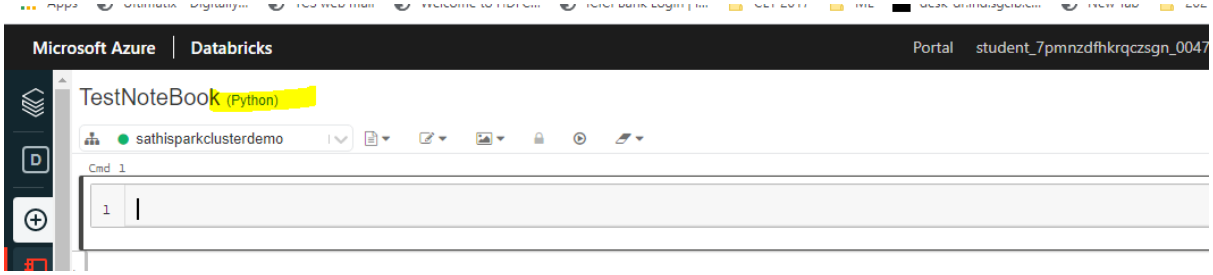


Copy uploaded json path for future reference



Change language to pythonin notebook





Extract the above json data into Databricks

Provide storage connection details

```
Cmd 1
1  blob_account_name = "sathitestblobstorage"
2  blob_container_name = "data"
3  blob_relative_path = "small_radio_json.json"
4  blob_sas_token = r"?sv=2020-08-04&ss=b&srt=sco&sp=rwdlacx&se=2021-08-04T23:46:02Z&st=2021-08-
   04T15:46:02Z&spr=https&sig=GDW7fAV0aS3HQnZ3KWuWo6AR3N0ptCZmI9kYs1lofZ8%3D"

Command took 0.04 seconds -- by student_7pmnzdfhkrqczsgn_00471960@vocareumvocareum.onmicrosoft.com at 8/4/2021, 9:52:07 PM on sathisparkclusterdemo
Cmd 2
```

Create datasource path

```
Cmd 2
1  wasbs_path = 'wasbs://%s@%s.blob.core.windows.net/%s' % (blob_container_name, blob_account_name, blob_relative_path)
2  spark.conf.set('fs.azure.sas.%s.%s.blob.core.windows.net' % (blob_container_name, blob_account_name), blob_sas_token)
3  print('Remote blob path: ' + wasbs_path)

Remote blob path: wasbs://data@sathitestblobstorage.blob.core.windows.net/userId
Command took 0.04 seconds -- by student_7pmnzdfhkrqczsgn_00471960@vocareumvocareum.onmicrosoft.com at 8/4/2021, 9:34:08 PM on sathisparkclusterdemo
```

`Load json file to dataframe`

```
Cmd 3
1  df = spark.read.json(wasbs_path)
2  print('Register the DataFrame as a SQL temporary view: source')
3  df.createOrReplaceTempView('source')

▸ (1) Spark Jobs
▸ df: pyspark.sql.dataframe.DataFrame = [artist: string, auth: string ... 15 more fields]
Register the DataFrame as a SQL temporary view: source
Command took 0.43 seconds -- by student_7pmnzdfhkrqczsgn_00471960@vocareumvocareum.onmicrosoft.com at 8/4/2021, 9:58:20 PM on sathisparkclusterdemo
```

Query

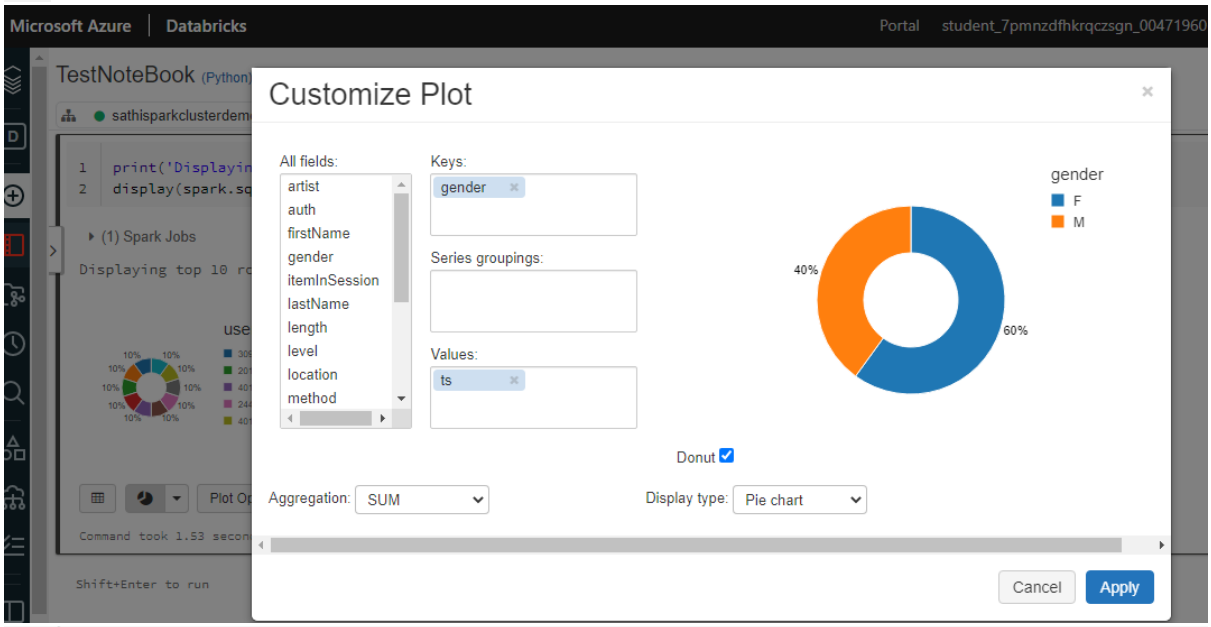
```
1  print('Displaying top 10 rows: ')
2  display(spark.sql('SELECT * FROM source LIMIT 10'))

▸ (1) Spark Jobs
Displaying top 10 rows:
```

	artist	auth	firstName	gender	itemInSession	lastName	length	level	location
1	El Arrebato	Logged In	Annalyse	F	2	Montgomery	234.57914	free	Killeen-Temple, TX
2	Creedence Clearwater Revival	Logged In	Dylann	M	9	Thomas	340.87138	paid	Anchorage, AK
3	Gorillaz	Logged In	Liam	M	11	Watts	246.17751	paid	New York-Newark-Jersey City, NY-
4	null	Logged In	Tess	F	0	Townsend	null	free	Nashville-Davidson--Murfreesboro-
5	Otis Redding	Logged In	Margaux	F	2	Smith	135.57506	free	Atlanta-Sandy Springs-Roswell, GA
6	Slightly Stoopid	Logged In	Alan	M	39	Morse	198.53016	paid	Chicago-Naperville-Elgin, IL-IN-WI
7	NOFX	Logged In	Gabriella	F	1	Shelton	130.2722	free	San Jose-Sunnyvale-Santa Clara,

```
Showing all 10 rows.
Command took 1.53 seconds -- by student_7pmnzdfhkrqczsgn_00471960@vocareumvocareum.onmicrosoft.com at 8/4/2021, 9:58:55 PM on sathisparkclusterdemo
```

PLOT




```
display(spark.sql('SELECT * FROM source LIMIT 10'))
```

Rename the column “location” to address

```
df.withColumnRenamed("location","address").printSchema()
```

```
Command took 1.53 seconds -- by student_7pmnzdfhkrqczsgn_00471960@vocareumvocareum.onmicrosoft.com at 8/4/2021, 9:58:55 PM on sathisparkclusterdemo
```

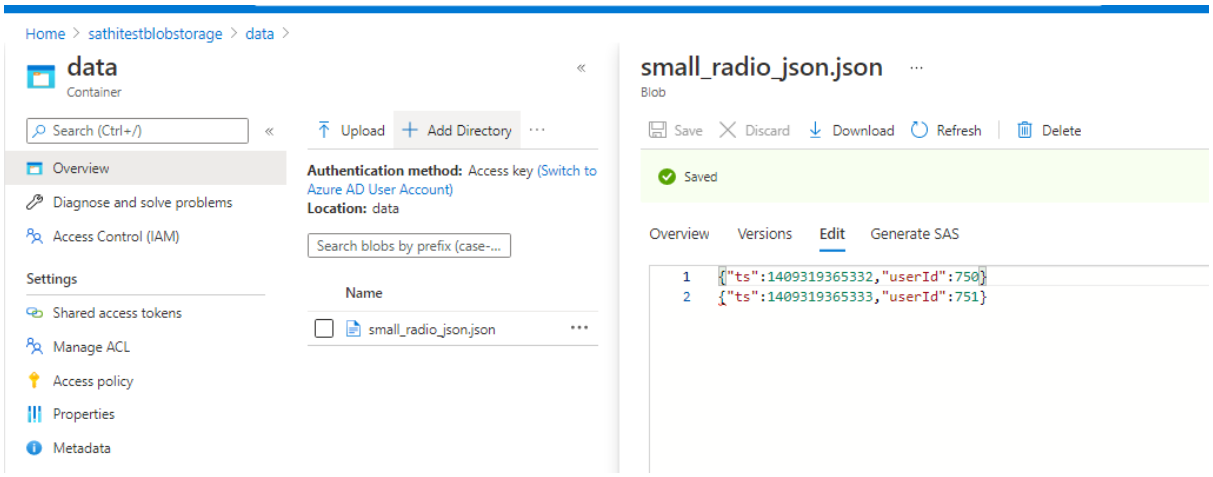
Cmd 5

```
1  df.withColumnRenamed("location","address").printSchema()
2
```

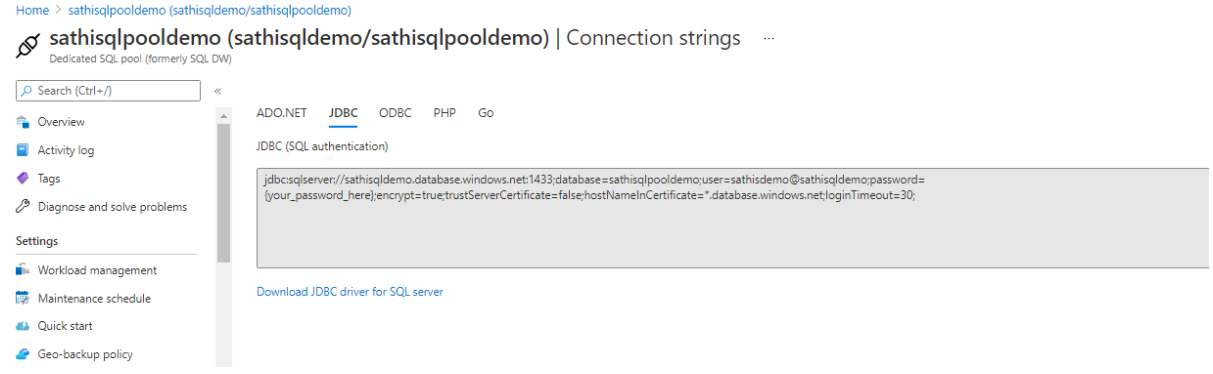
```
root
 |-- artist: string (nullable = true)
 |-- auth: string (nullable = true)
 |-- firstName: string (nullable = true)
 |-- gender: string (nullable = true)
 |-- itemInSession: long (nullable = true)
 |-- lastName: string (nullable = true)
 |-- length: double (nullable = true)
 |-- level: string (nullable = true)
 |-- address: string (nullable = true)
 |-- method: string (nullable = true)
 |-- page: string (nullable = true)
 |-- registration: long (nullable = true)
 |-- sessionId: long (nullable = true)
```

Load the transformed data into the created Synapse service

Edit json file in storage to make it simpler as below



Copy sql pool connection string details created in first step



connect

Cmd 6

```
jdbcHostname = "sathisqldemo.database.windows.net"
jdbcPort = "1433"
jdbcDatabase = "sathisqlpooldemo"
properties = {
 "user" : "sathisdemo",
 "password" : "Aomnmrs&1" }
```

Command took 0.04 seconds -- by student_7pmnzdfhkrqczsgn_00471960@vocareumvocareum.onmicrosoft.com at 8/4/2021, 10:47:15 PM on sathisparkclusterdemo

Cmd 7

```
url = "jdbc:sqlserver://{0}:{1};database={2}".format(jdbcHostname,jdbcPort,jdbcDatabase)
df = spark.read.json(wasbs_path)
```

▸ (1) Spark Jobs

▸ df: pyspark.sql.dataframe.DataFrame = [ts: long, userId: string]

Command took 10.77 seconds -- by student_7pmnzdfhkrqczsgn_00471960@vocareumvocareum.onmicrosoft.com at 8/4/2021, 10:47:20 PM on sathisparkclusterdemo

Verfiy loaded json data , dataframe and push data

Cmd 8

```
df.createOrReplaceTempView('source')
display(spark.sql('SELECT * FROM source LIMIT 10'))
```

▸ (1) Spark Jobs

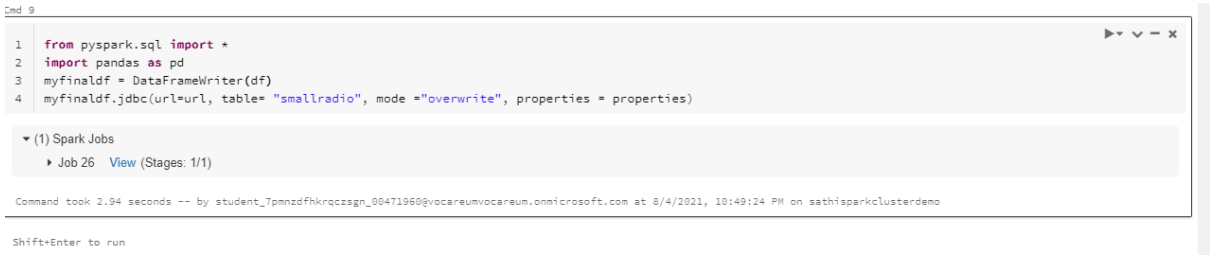
	ts	userId
1	1409319365332	750
2	1409319365333	751

Showing all 2 rows.

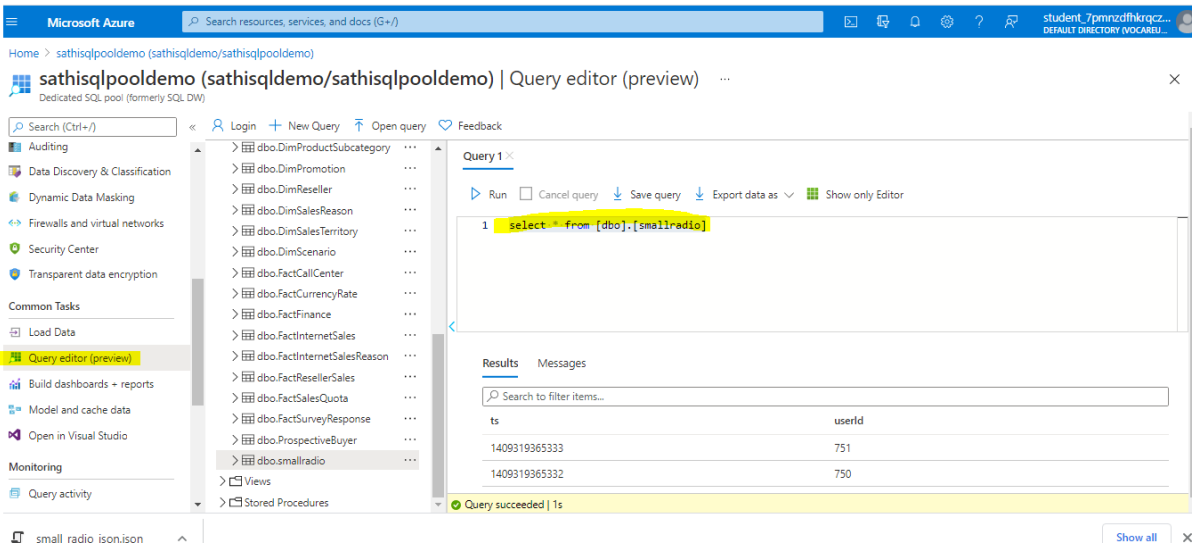
Command took 0.21 seconds -- by student_7pmnzdfhkrqczsgn_00471960@vocareumvocareum.onmicrosoft.com at 8/4/2021, 10:49:20 PM on sathisparkclusterdemo

Cmd 9

```
from pyspark.sql import *
import pandas as pd
myfinaldf = DataFrameWriter(df)
myfinaldf.jdbc(url=url, table= "smallradio", mode ="overwrite", properties = properties)
```



Verify published data in the pool by connecting to server to wich we have published



Refrences

Reference links:

1) https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/create-data-warehouse-portal
2) https://docs.microsoft.com/en-us/azure/storage/blobs/storage-quickstart-blobs-portal
3) https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-quickstart-create-account
4) https://docs.microsoft.com/en-us/azure/databricks/scenarios/quickstart-create-databricks-workspace-portal?tabs=azure-portal

5) https://docs.microsoft.com/en-us/azure/open-datasets/dataset-seattle-safety?tabs=pyspark