# Data Ingestion and Analysis with Azure Databricks



Cleber Zumba
Data Engineer

# Sumário

# 1 Version

This document was created by Cleber Zumba de Souza and can be freely distributed, as long as the source is mentioned.

| Version | Action | Data |
|---|---|---|
| **1.0** | Document creation | 2024/10/05 |

# 2 Introduction

In this project, I developed a data engineering solution integrating AWS, Azure, and Databricks technologies. The main goal was to create an efficient and scalable data pipeline that spans from data ingestion to transformation and analysis, ensuring secure storage and robust processing. The data processed in this pipeline is related to emergency calls from the San Francisco Fire Department.

*Context and Objectives*

In modern data environments, it is essential to integrate different cloud platforms and tools to optimize the data workflow. This project was designed to:

- **Data Ingestion:** Read data from an S3 bucket on AWS.
- **Centralized Storage:** Store the data in a Data Lake in Azure Data Lake Storage (ADLS).
- **Processing and Analysis:** Use Databricks to transform, analyze, and store the transformed data.

*Data Overview*

The data used in this pipeline are records of emergency calls answered by the San Francisco Fire Department. This data contains critical information such as the nature of the call, the neighborhood where it occurred, and response times. Analyzing this data is vital to improving emergency services and public safety.

*Overview of Technologies Used*

**1. AWS S3:** Amazon's cloud storage service where raw data is initially stored.

**2. Azure Resource Group:** Groups and manages all project-related resources in Azure.

**3. Azure Storage Account:** Provides secure and scalable storage for data in Azure.

**4. Azure Data Factory:** Data integration service used to orchestrate and automate the movement of data between AWS S3 and ADLS services.

**5. Azure Data Lake Storage (ADLS):** Centralized storage solution in Azure that allows efficient and secure storage of data.

**6. Databricks:** Apache Spark-based data processing and analytics platform used to transform, analyze, and store data.

*Project Workflow*

1. **Data Ingestion with Azure Data Factory:**
    - A data pipeline in Azure Data Factory reads the data file stored in the AWS S3 bucket.
    - The data is then transferred and stored in the Data Lake in Azure Data Lake Storage (ADLS).
2. **Processing and Analysis with Databricks:**
    - Databricks integrates with ADLS to access the stored data.
    - Using the processing power of Apache Spark, the data is transformed and analyzed as per the project needs.
    - The transformed data is then stored back in ADLS or other destinations as needed.

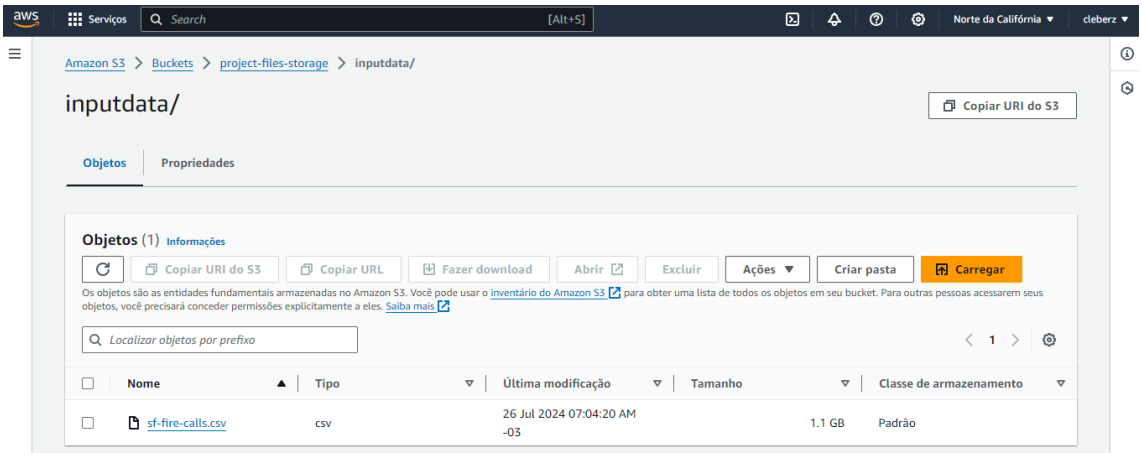
# 3 Data Ingestion

## 3.1 Introduction

In this chapter, I detailed the data ingestion process, from reading the file stored in the Amazon S3 bucket to transferring the data to Azure Data Lake Storage (ADLS). This is the first step in the data pipeline and is crucial to ensuring that the raw data is available for subsequent processing and analysis.

## 3.2 Initial Storage on Amazon S3

The San Francisco Fire Department's call data is initially stored in an Amazon S3 bucket. The image below shows the storage structure of the sf-fire-calls.csv file in the S3 bucket:

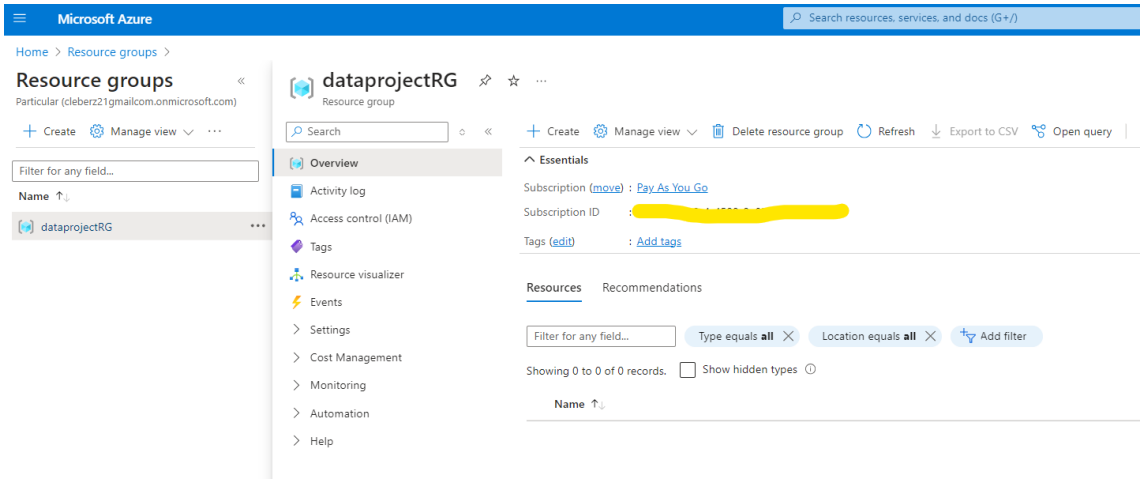



- **Bucket**: project-files-storage
- **Directory**: inputdata
- **File**: sf-fire-calls.csv
- **Size**: 1.1 GB
- **Last Modified**: July 26, 2024, 07:04 AM

# 4 Layers of Service Provisioning and Configuration in Azure

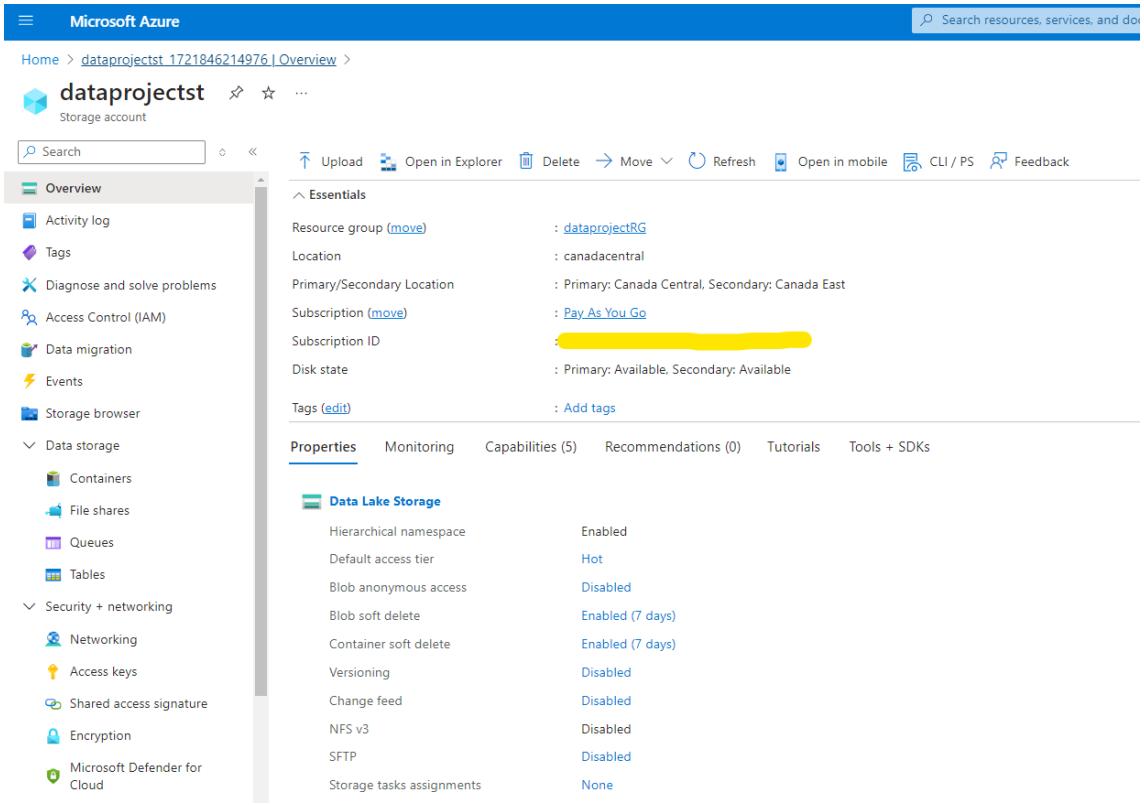
Before performing the data transfer, it was necessary to provision and configure the services in Azure. The steps include creating a Resource Group, a Storage Account, a Container, and a Key Vault.

## 4.1 Resource Group



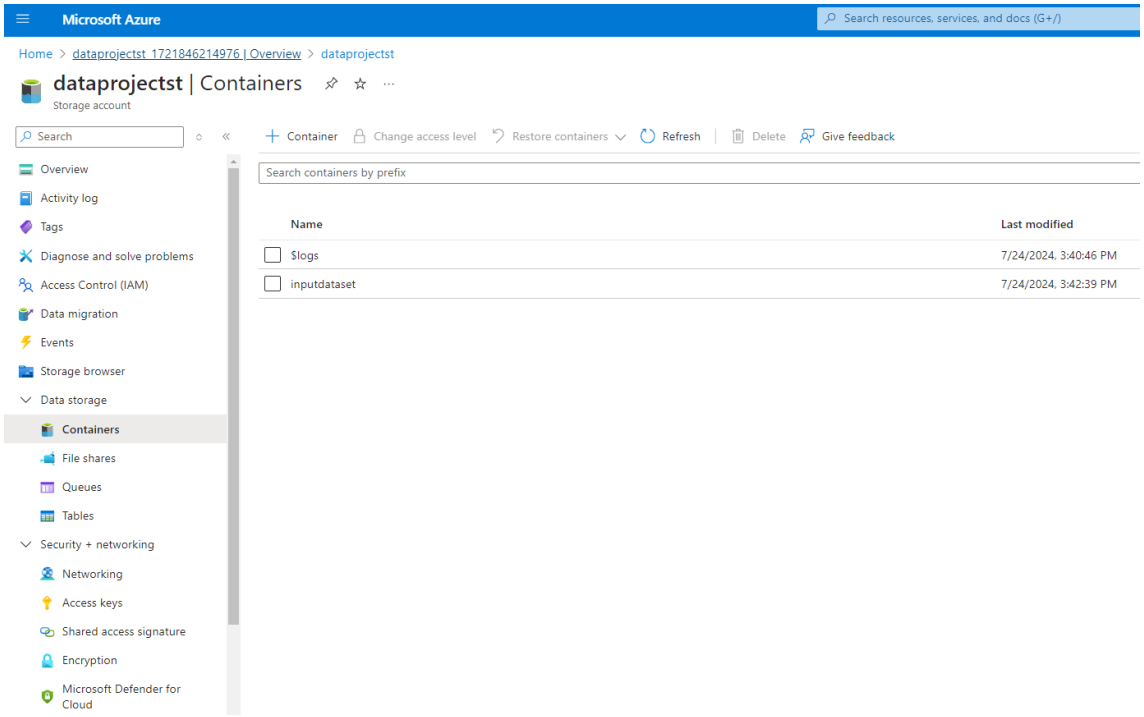

A Resource Group is created to group and manage all project-related resources in Azure. This makes it easier to organize and manage resources.

## 4.2 Storage Account



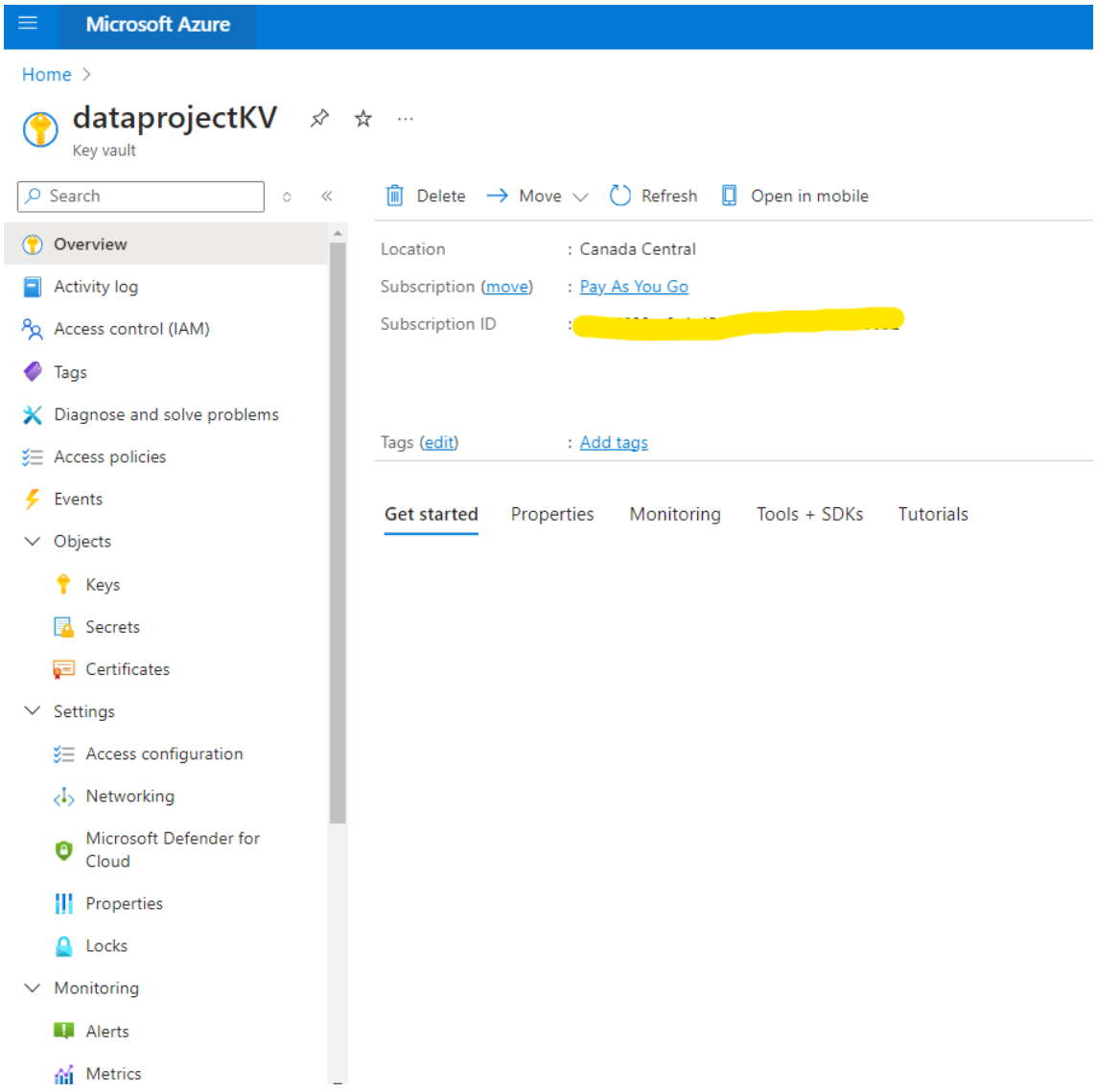

A Storage Account is created to provide secure and scalable storage for your data. This storage account is essential for storing data transferred from S3.

## 4.3 Container

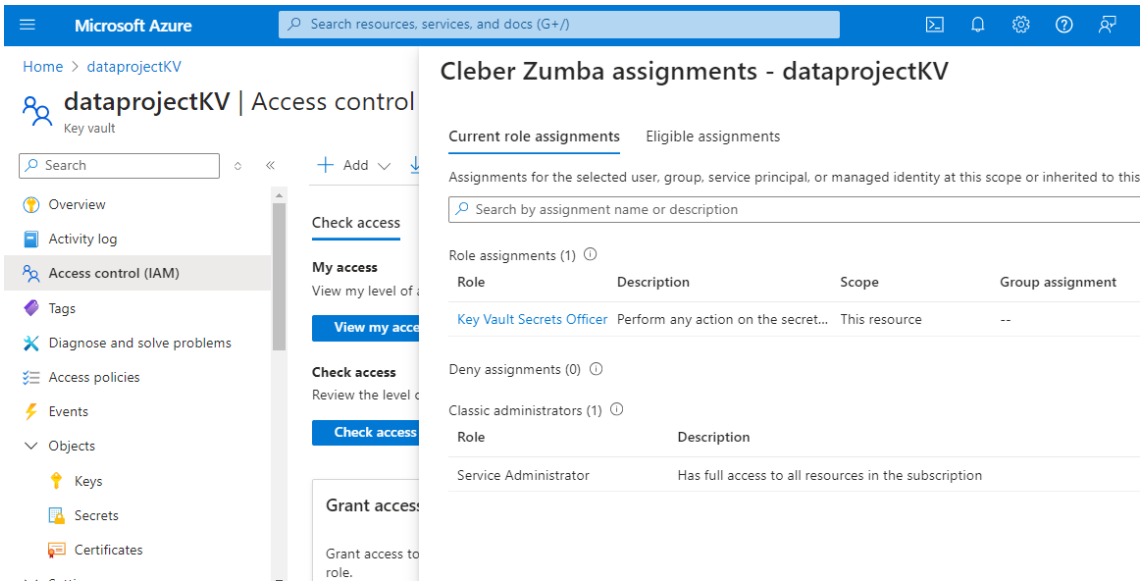



Within the Storage Account, a container was created to store the data in the desired format.

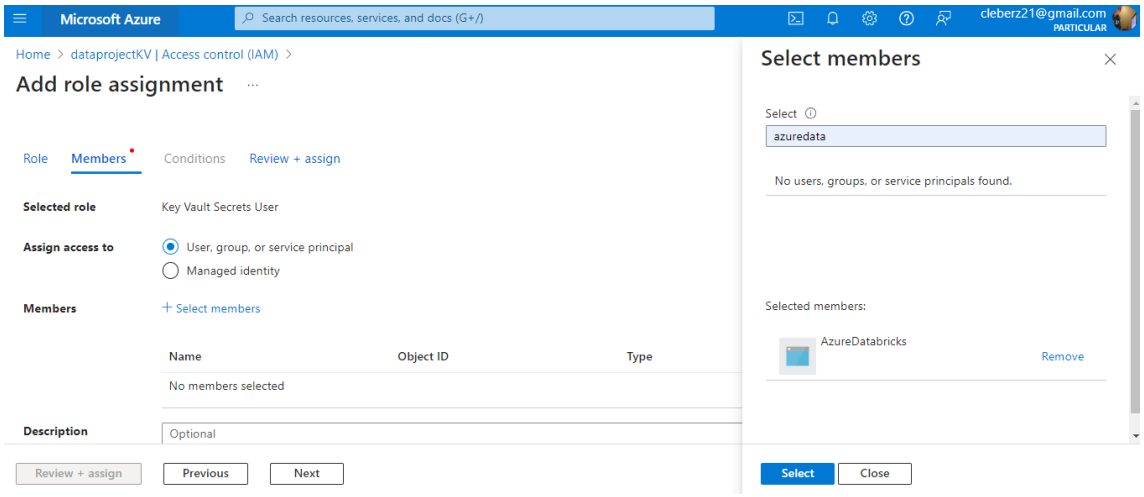
## 4.4 Key Vault



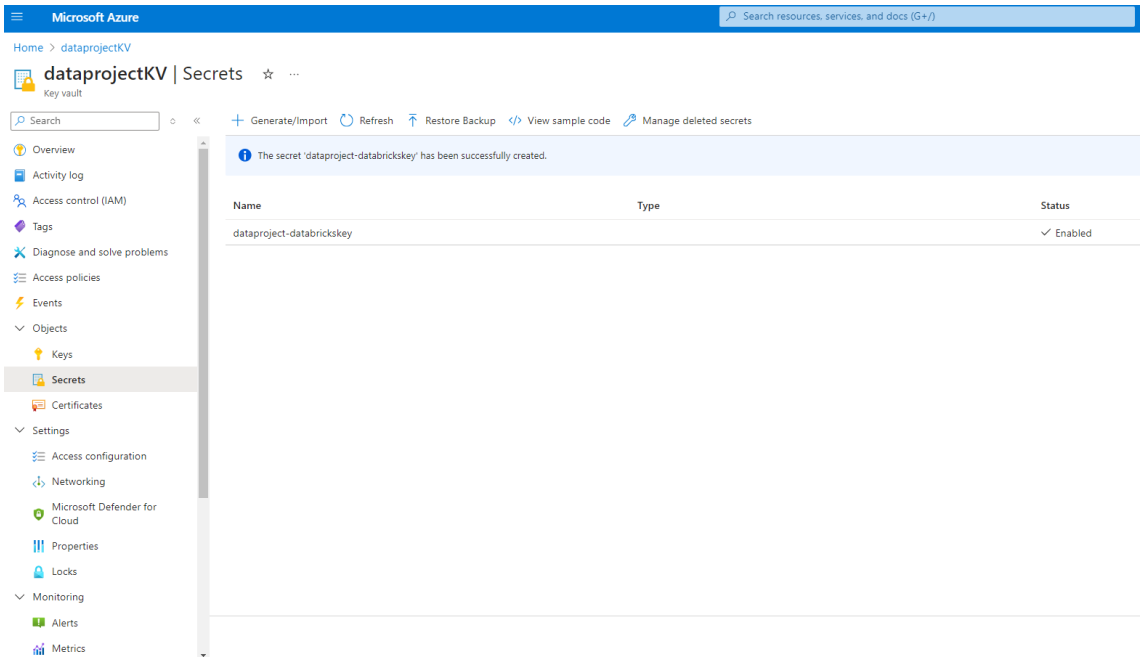

Key Vault Creation




Adding the Key Vault Secrets Officer role



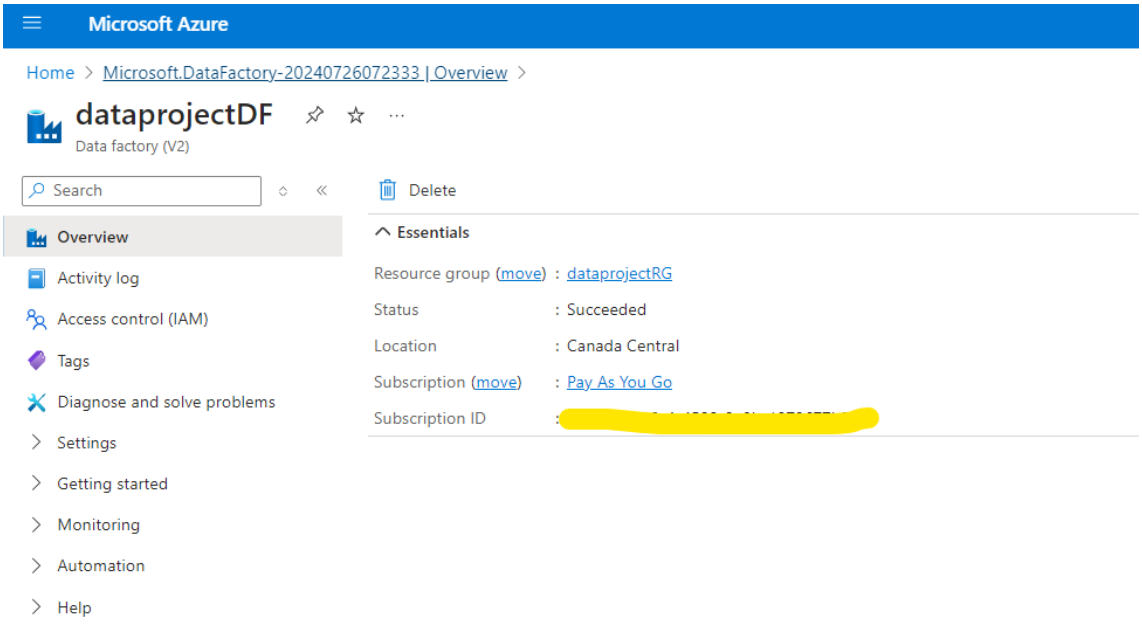

Adding Azure Databricks member to Key Vault Secrets Officer role



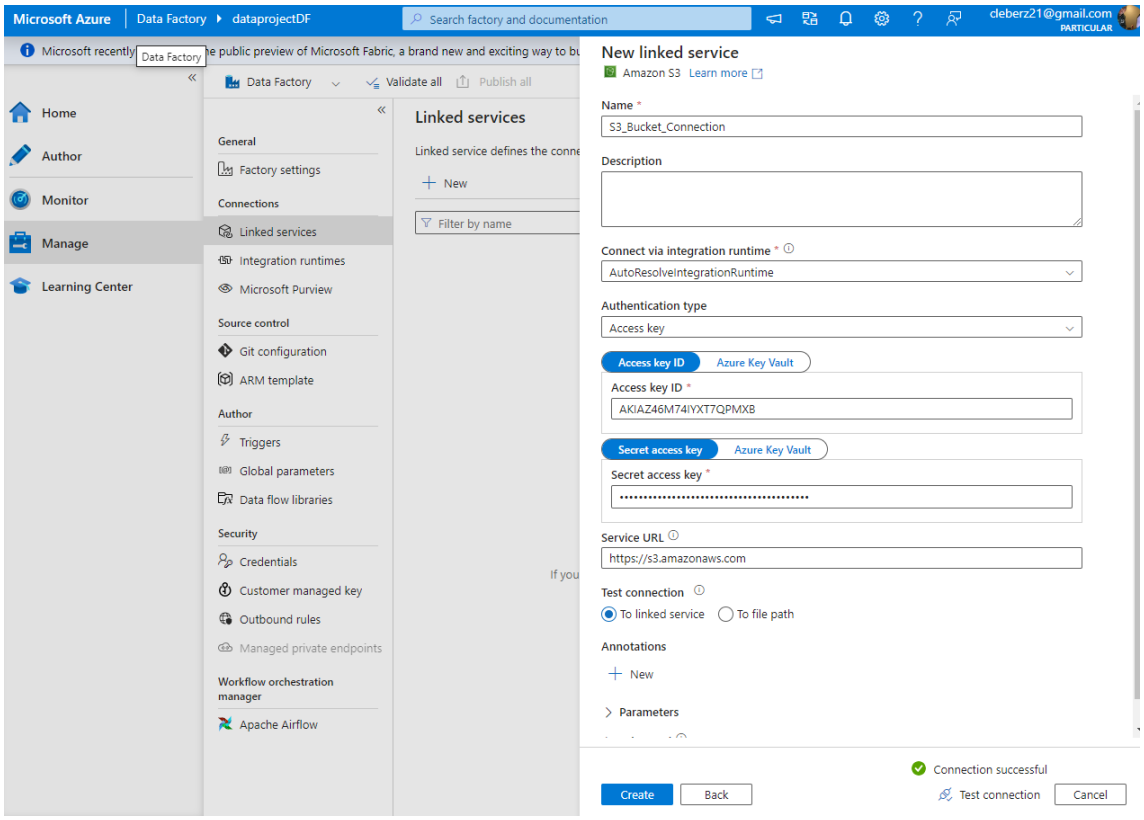

Creating a Secret Key Vault dataprojectKV

# 5 Data Transfer to Azure Data Lake Storage (ADLS)

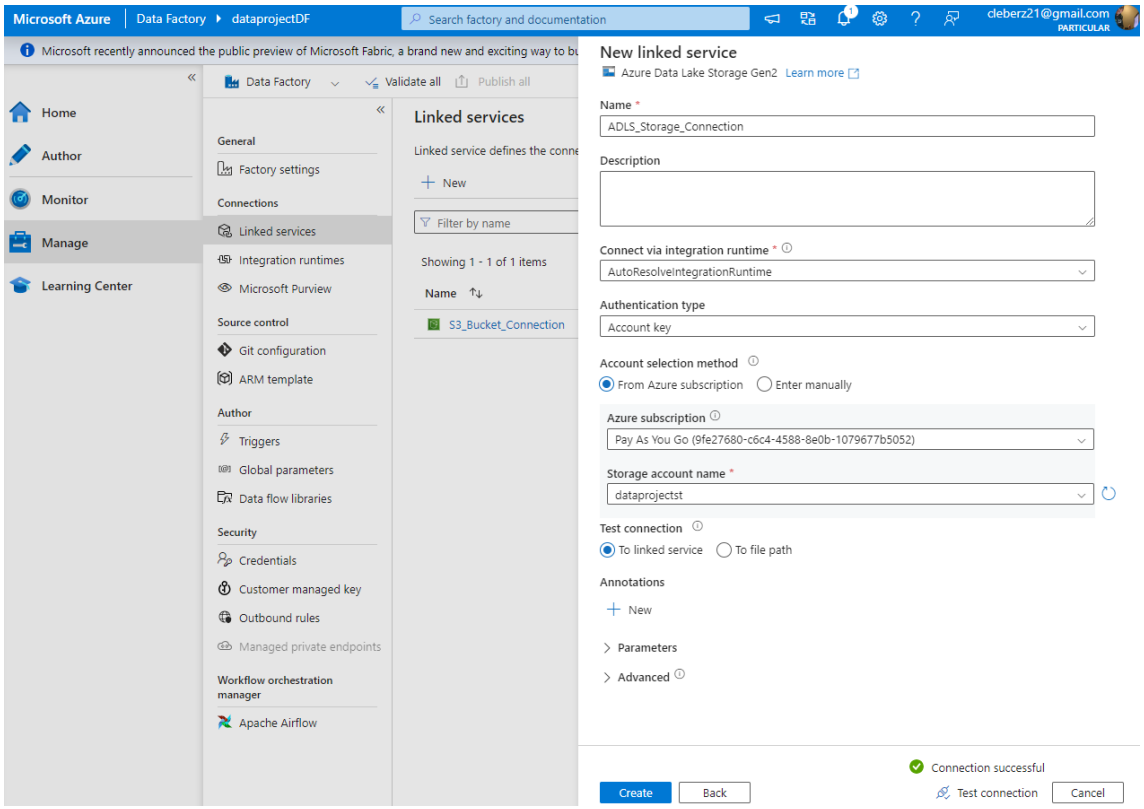
Using Azure Data Factory, I set up a data pipeline to read the file stored in the Amazon S3 bucket and transfer it to Azure Data Lake Storage (ADLS). This process involves:



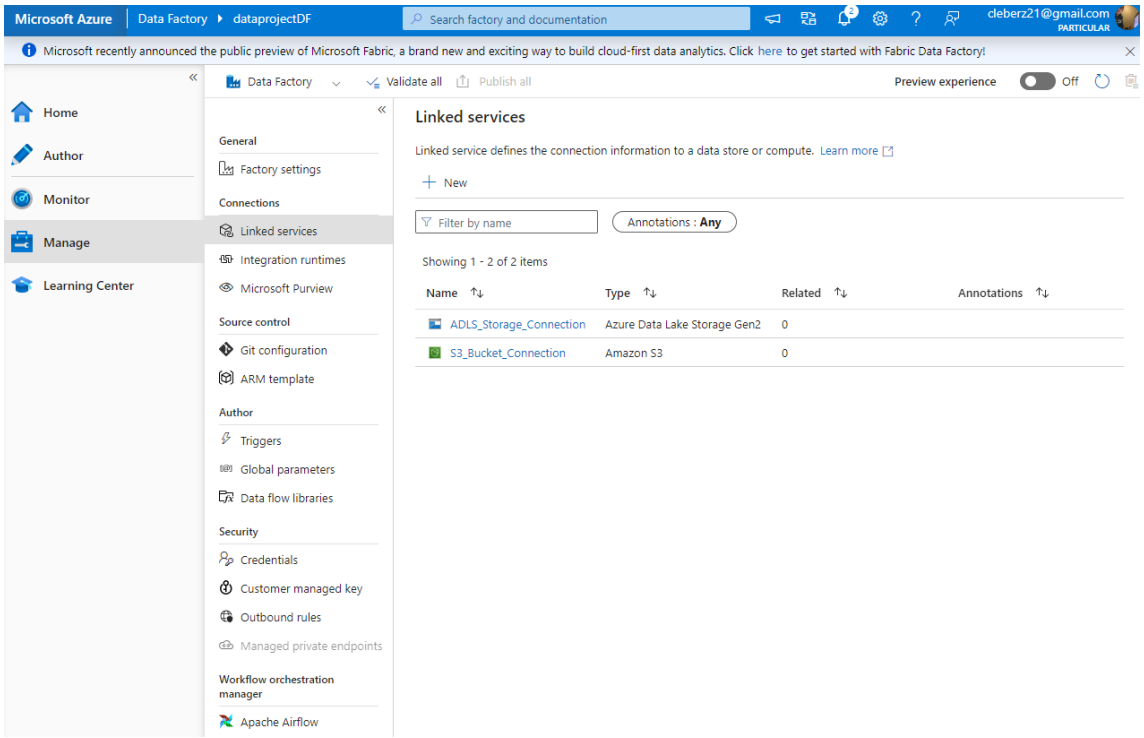

**Creating a Data Factory**



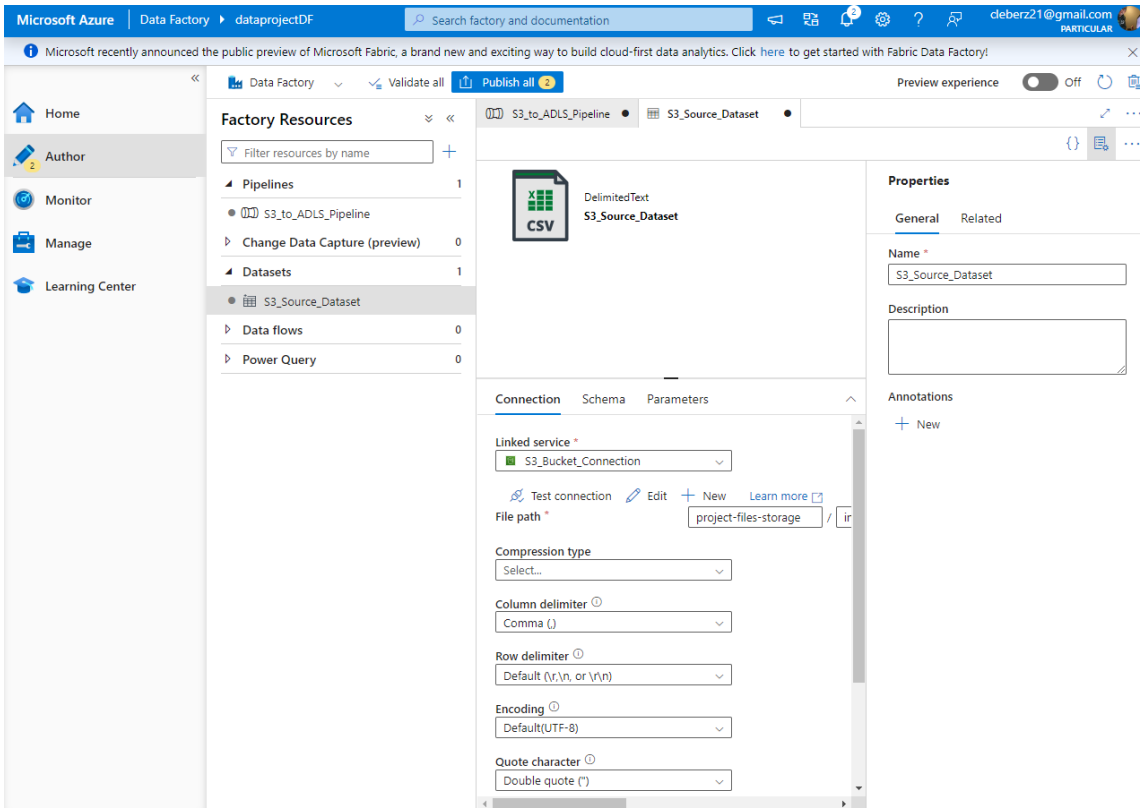

**Setting up a Linked Service in Azure Data Factory to connect to S3 bucket.**



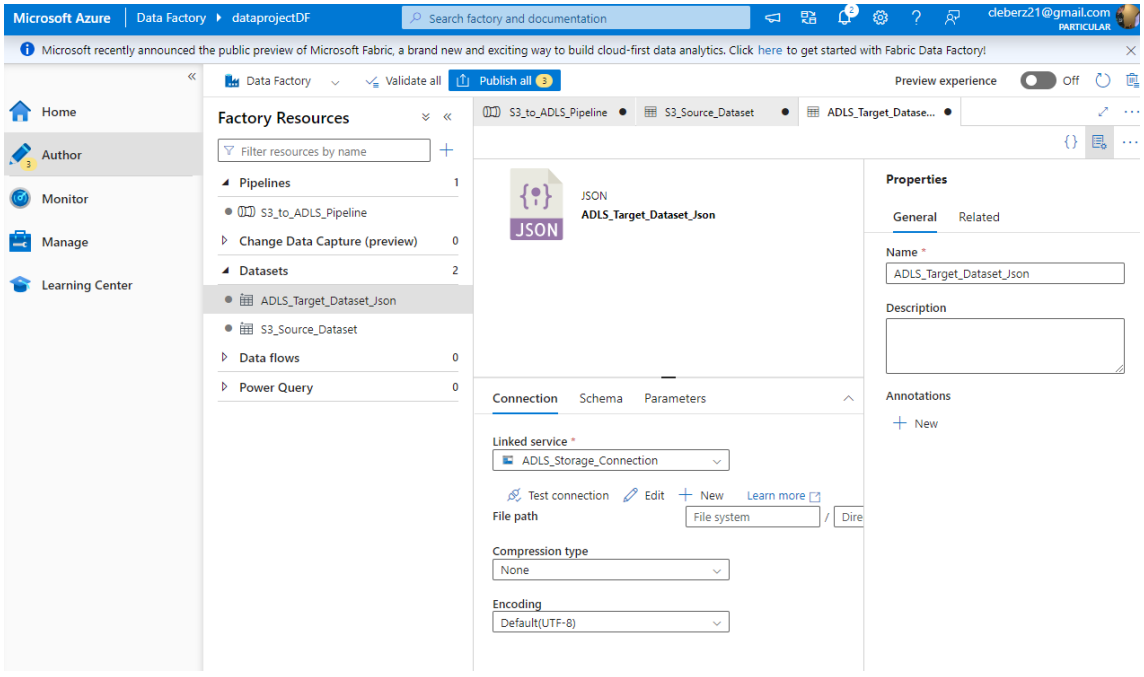

**Configuring a Linked Service in Azure Data Factory to connect to ADLS.**



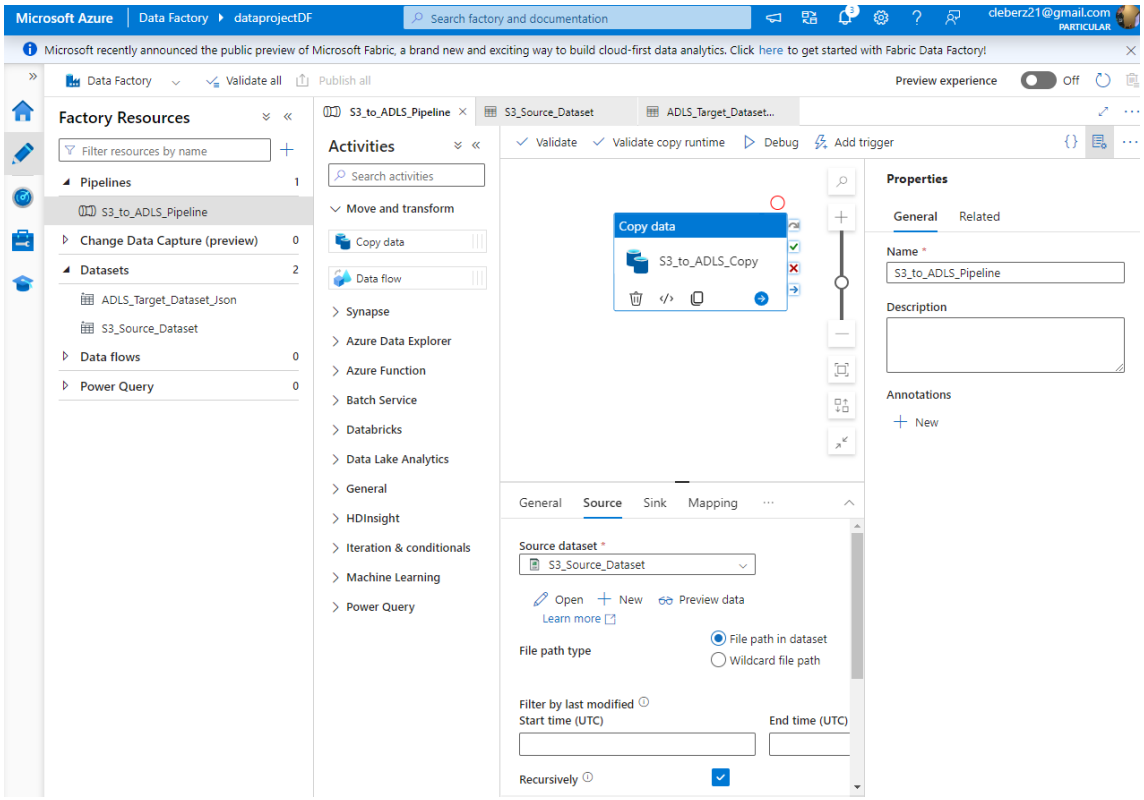

Linked Service created




**Creation of a Dataset that represents the source of the data**



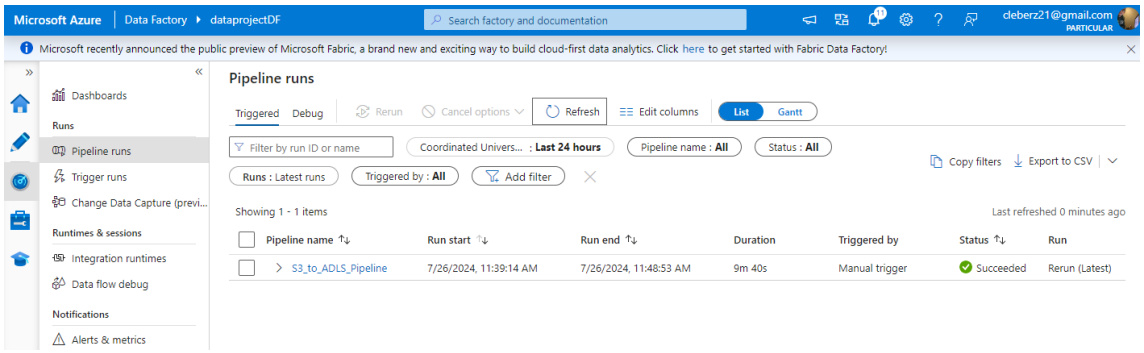

**Creation of a Dataset that represents the destination of the data.**

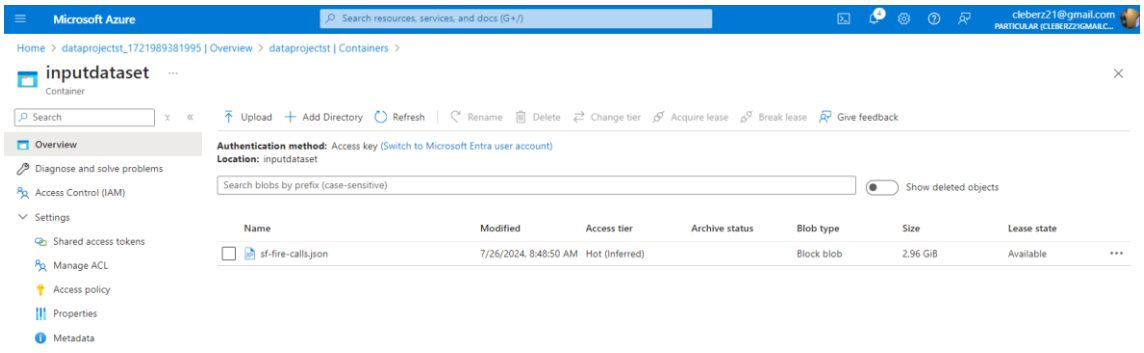



**Development of a Pipeline that copies data from S3 to ADLS.**

# 6 Data Pipeline Execution

The data pipeline has been configured to perform the data transfer efficiently and securely, ensuring that the sf-fire-calls.csv file is made available in the ADLS Data Lake, in Json format, for subsequent processing in Databricks.

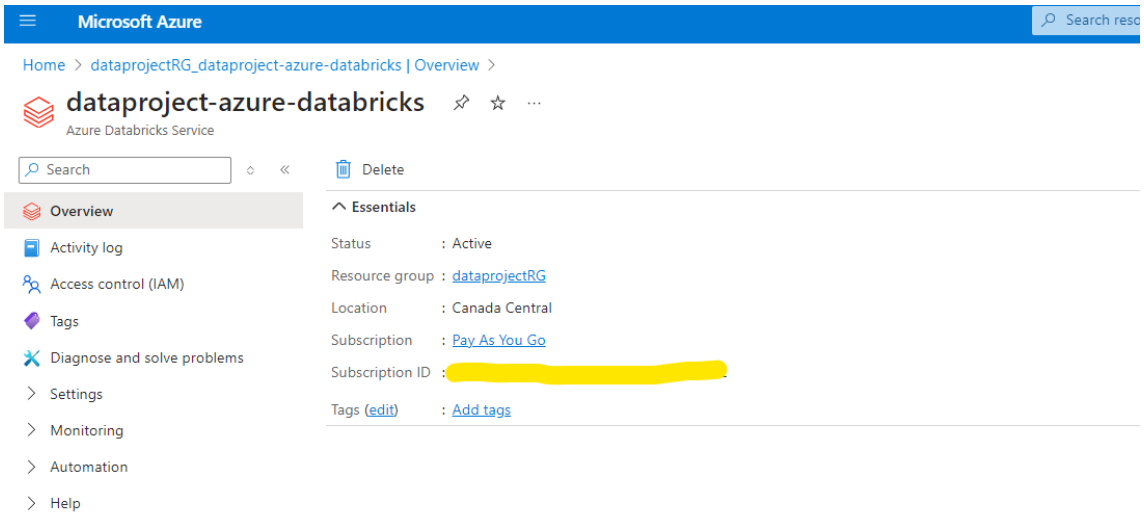


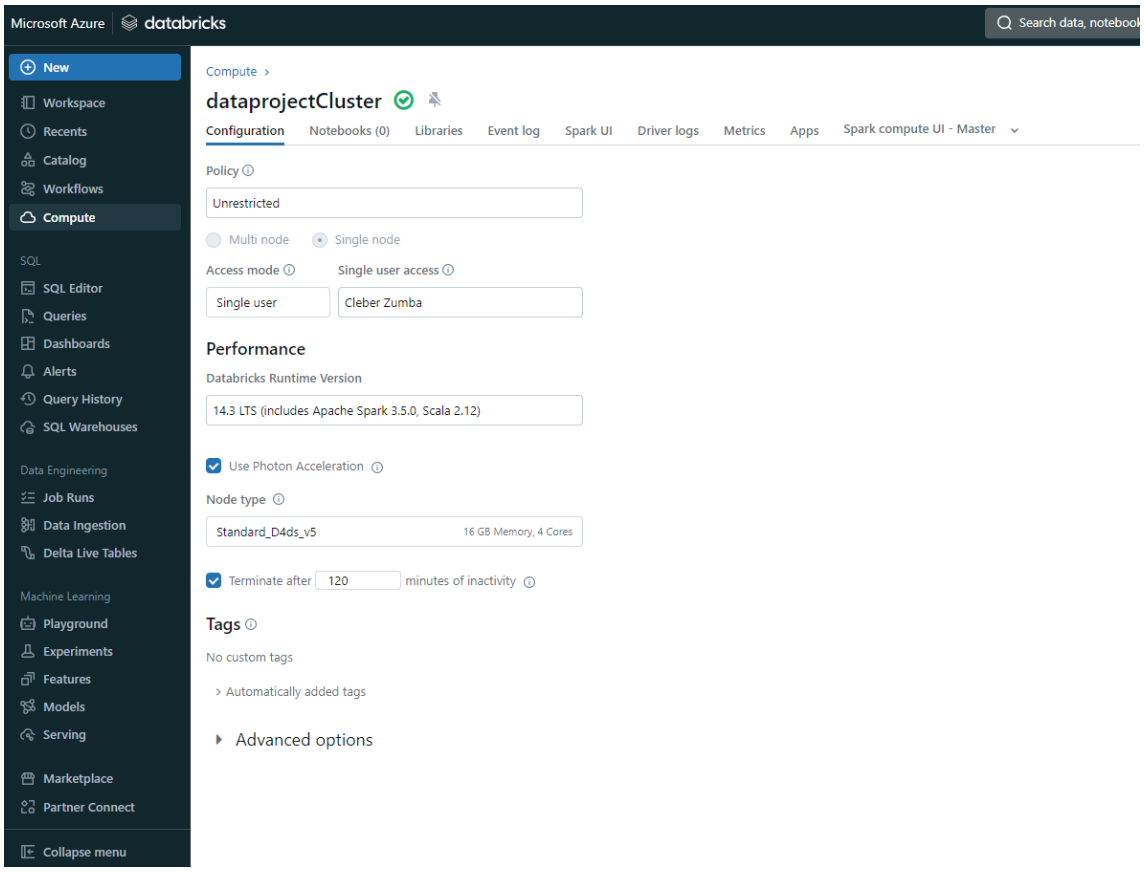




Data transfer completed

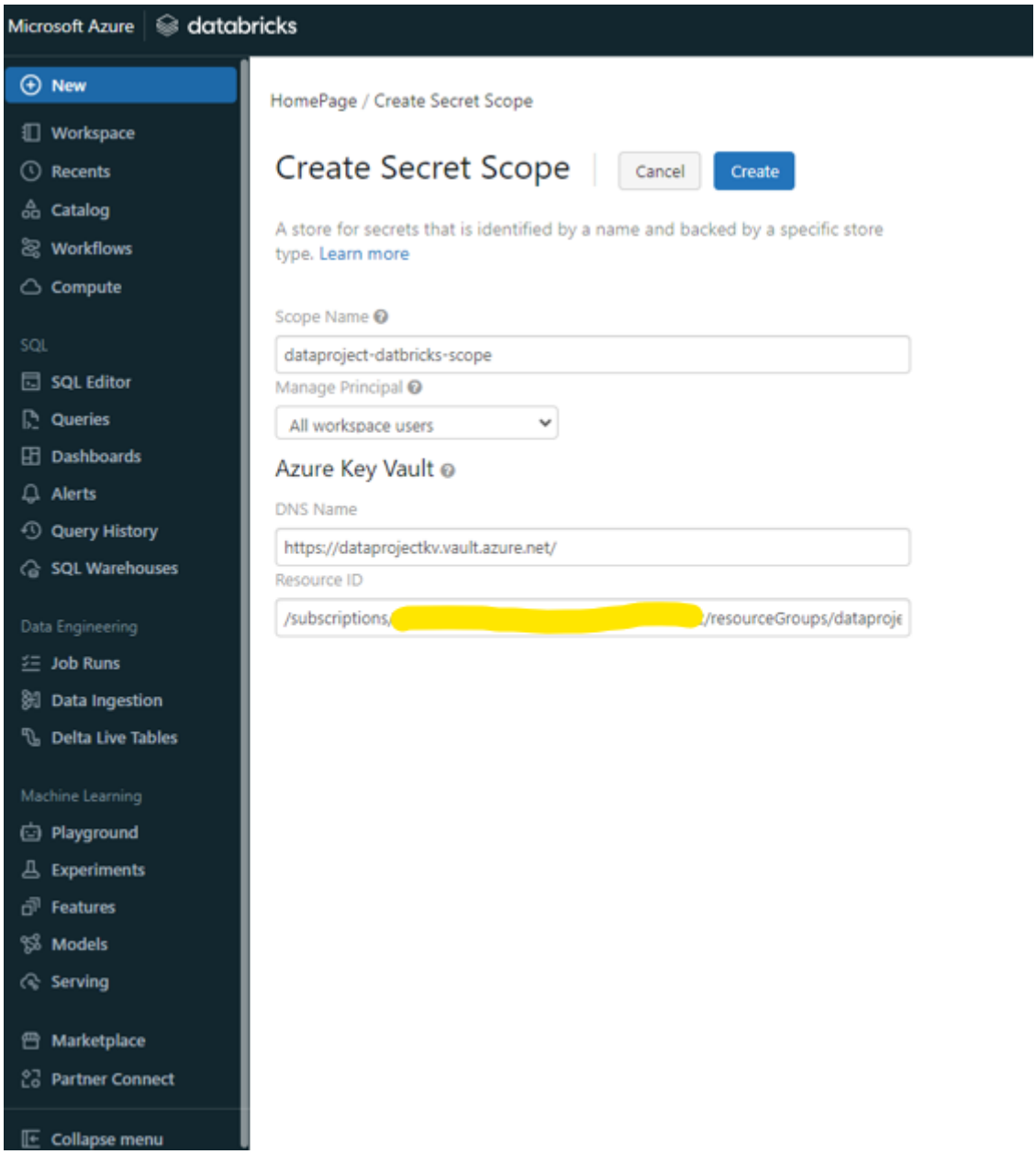
# 7 Creating Azure Databricks Workspace and Cluster

Azure Databricks has been provisioned to perform data transformation and analysis. This includes creating a workspace, cluster, and secret scope.

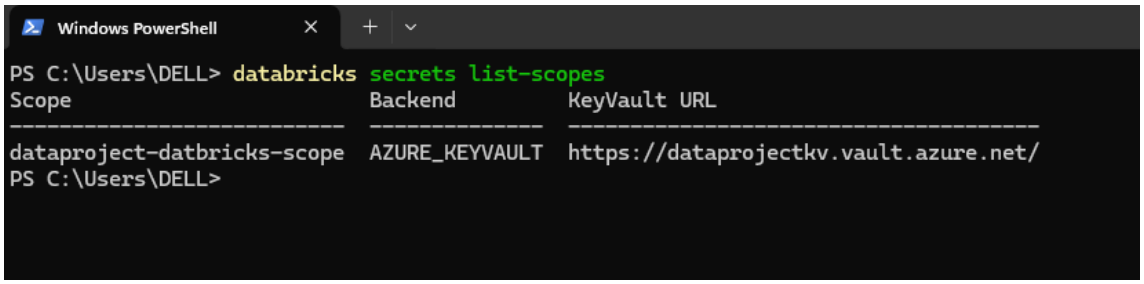





Creating a Cluster in Databricks

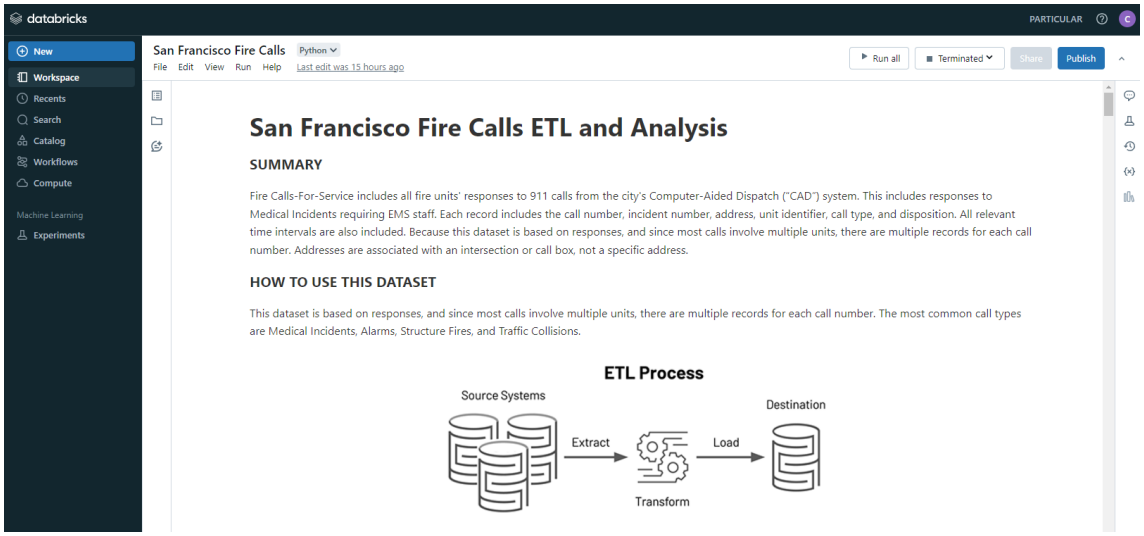


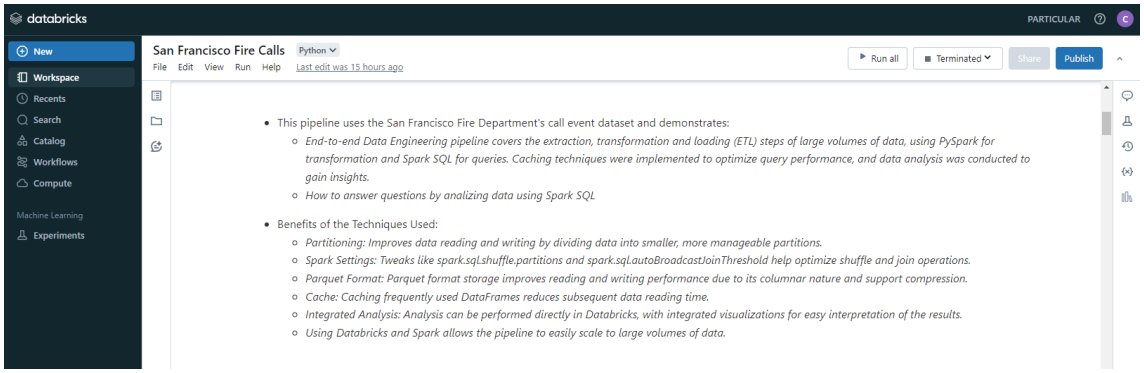

Creating a Secret Scope in Databricks

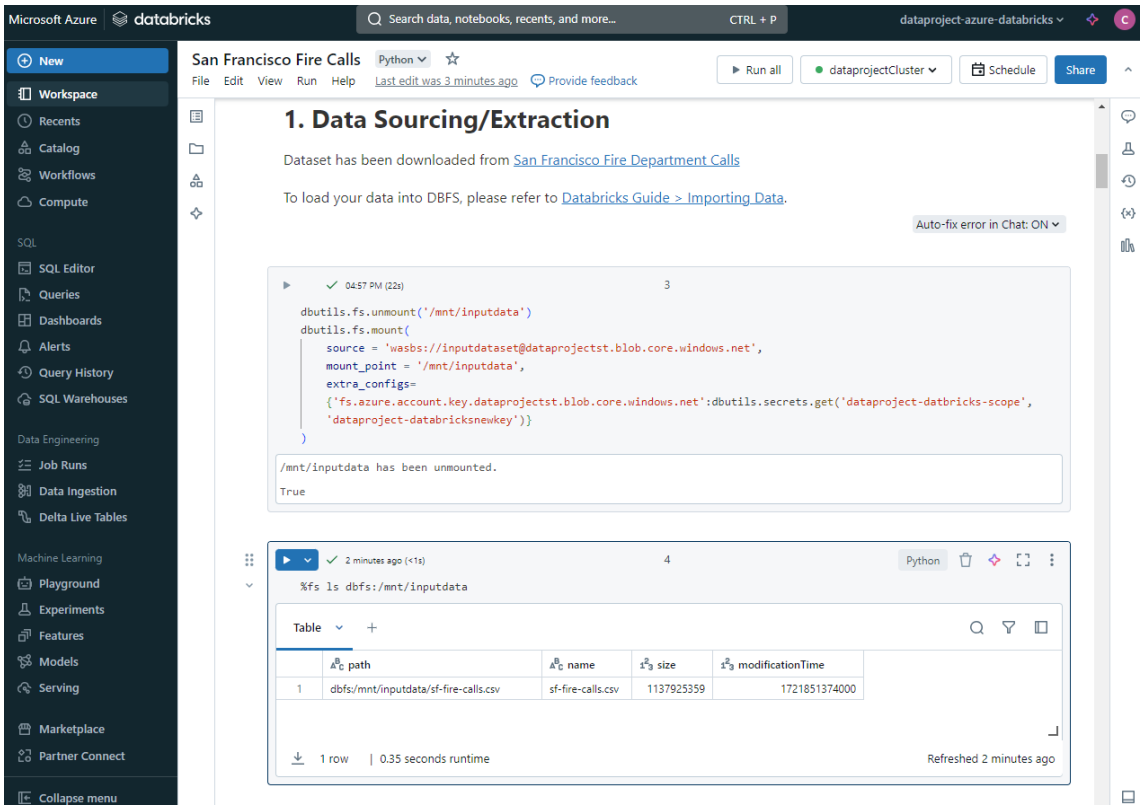


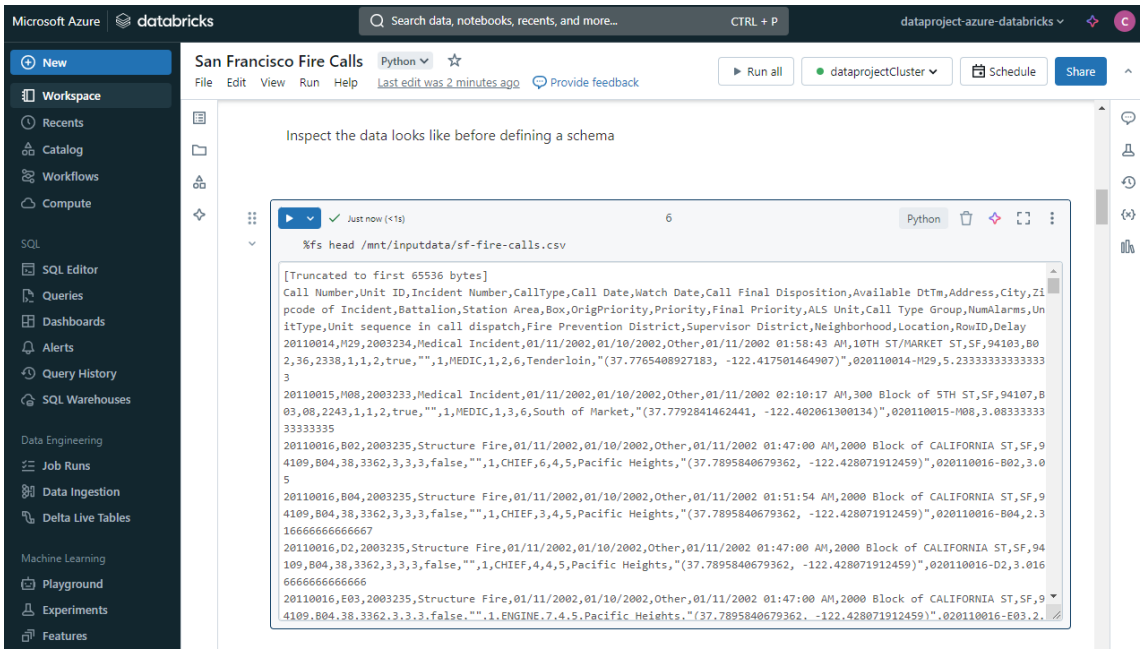

List Scope

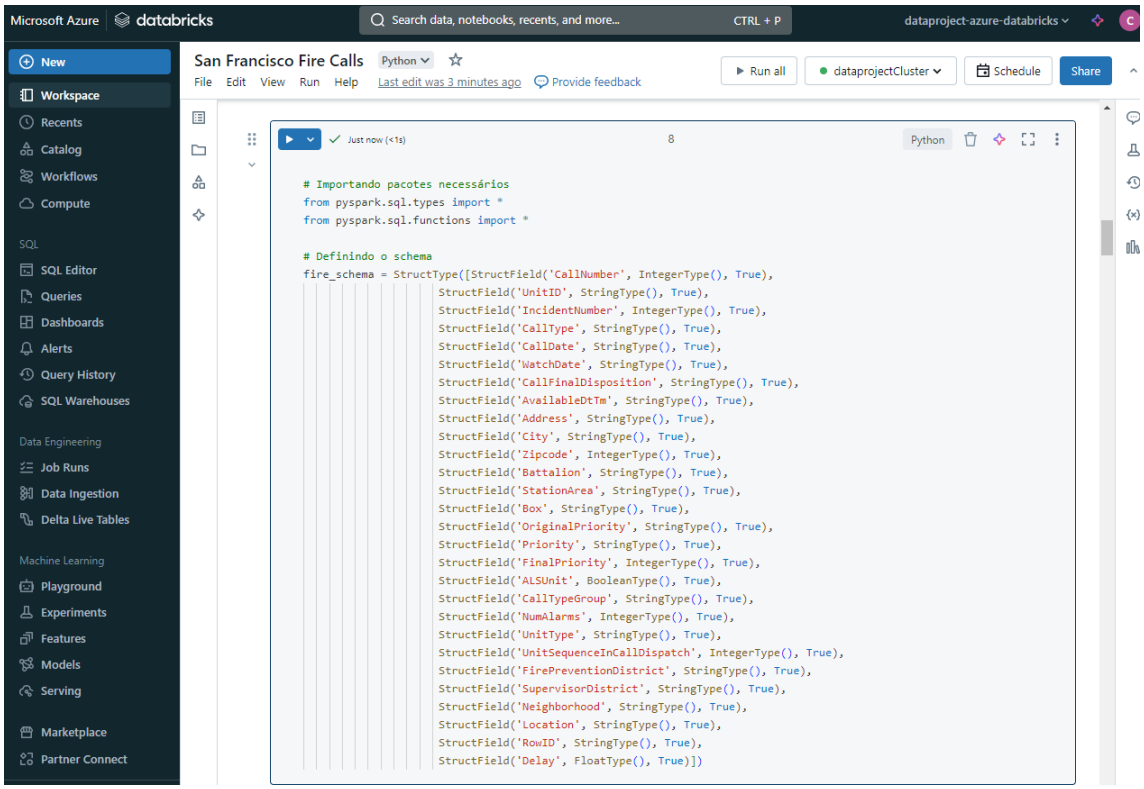
# 8 Data Extraction Layer

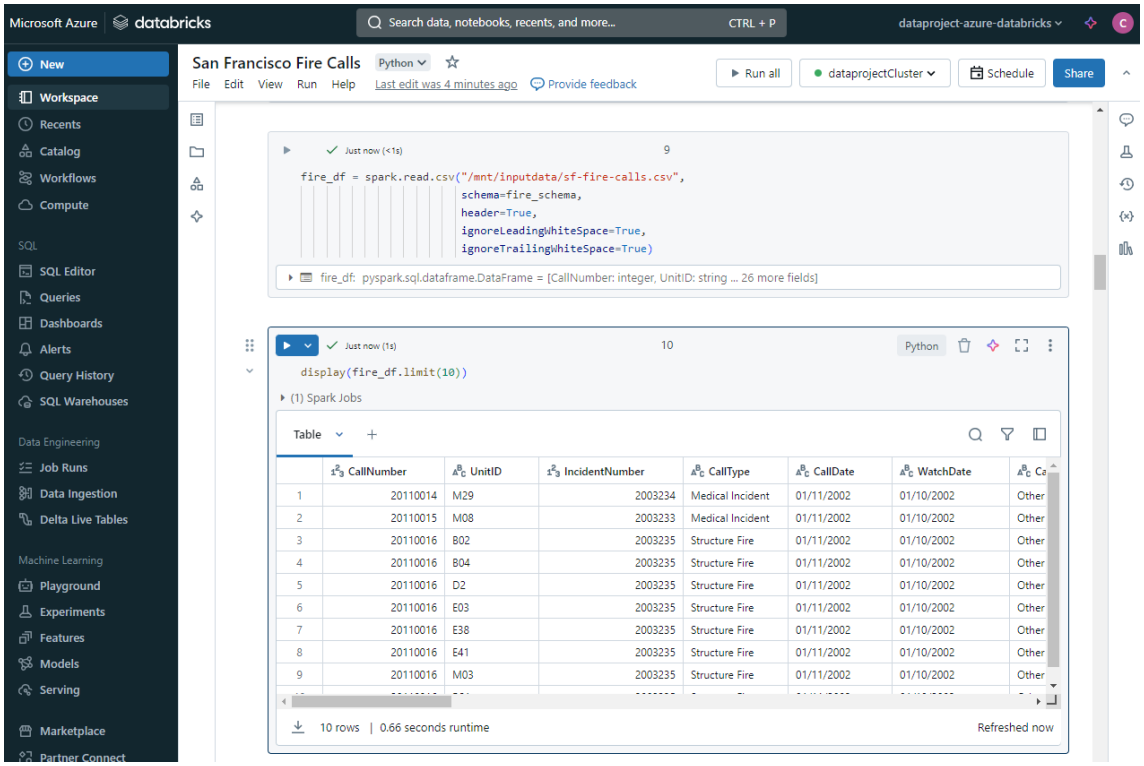



Microsoft Azure
databricks
Search data, notebooks, recents, and more...
CTRL + P
dataproject-azure-databricks
New
Workspace
Recents
Catalog
Workflows
Compute
SQL
SQL Editor
Queries
Dashboards
Alerts
Query History
SQL Warehouses
Data Engineering
Job Runs
Data Ingestion
Delta Live Tables
Machine Learning
Playground
Experiments
Features
Models
Serving
Marketplace
Partner Connect
Collapse menu
San Francisco Fire Calls
Python
File Edit View Run Help
Last edit was 3 minutes ago
Provide feedback
Run all
dataprojectCluster
Schedule
Share
1. Data Sourcing/Extraction
Dataset has been downloaded from San Francisco Fire Department Calls
To load your data into DBFS, please refer to Databricks Guide > Importing Data.
Auto-fix error in Chat: ON
04:57 PM (22s)
3
dbutils.fs.unmount('/mnt/inputdata')
dbutils.fs.mount(
source = 'wasbs://inputdataset@dataprojectst.blob.core.windows.net',
mount_point = '/mnt/inputdata',
extra_configs=
{'fs.azure.account.key.dataprojectst.blob.core.windows.net':dbutils.secrets.get('dataproject-databricks-scope',
'dataproject-databricksnewkey')}
)
/mnt/inputdata has been unmounted.
True
2 minutes ago (<1s)
4
Python
%fs ls dbfs:/mnt/inputdata
Table
path
name
size
modificationTime
1
dbfs:/mnt/inputdata/sf-fire-calls.csv
sf-fire-calls.csv
1137925359
1721851374000
1 row
0.35 seconds runtime
Refreshed 2 minutes ago


Microsoft Azure
databricks
Search data, notebooks, recents, and more...
CTRL + P
dataproject-azure-databricks
New
Workspace
Recents
Catalog
Workflows
Compute
SQL
SQL Editor
Queries
Dashboards
Alerts
Query History
SQL Warehouses
Data Engineering
Job Runs
Data Ingestion
Delta Live Tables
Machine Learning
Playground
Experiments
Features
San Francisco Fire Calls
Python
File Edit View Run Help
Last edit was 2 minutes ago
Provide feedback
Run all
dataprojectCluster
Schedule
Share
Inspect the data looks like before defining a schema
Just now (<1s)
6
Python
%fs head /mnt/inputdata/sf-fire-calls.csv
[Truncated to first 65536 bytes]
Call Number,Unit ID,Incident Number,CallType,Call Date,Watch Date,Call Final Disposition,Available DtTm,Address,City,Zipcode of Incident,Battalion,Station Area,Box,OrigPriority,Priority,Final Priority,ALS Unit,Call Type Group,NumAlarms,UnitType,Unit sequence in call dispatch,Fire Prevention District,Supervisor District,Neighborhood,Location,RowID,Delay
20110014,M29,2003234,Medical Incident,01/11/2002,01/10/2002,Other,01/11/2002 01:58:43 AM,10TH ST/MARKET ST,SF,94103,B02,36,2338,1,1,2,true,"",1,MEDIC,1,2,6,Tenderloin,"(37.7765408927183, -122.417501464907)",020110014-M29,5.233333333333333
20110015,M08,2003233,Medical Incident,01/11/2002,01/10/2002,Other,01/11/2002 02:10:17 AM,300 Block of 5TH ST,SF,94107,B03,08,2243,1,1,2,true,"",1,MEDIC,1,3,6,South of Market,"(37.7792841462441, -122.402061300134)",020110015-M08,3.083333333333335
20110016,B02,2003235,Structure Fire,01/11/2002,01/10/2002,Other,01/11/2002 01:47:00 AM,2000 Block of CALIFORNIA ST,SF,94109,B04,38,3362,3,3,3,false,"",1,CHIEF,6,4,5,Pacific Heights,"(37.7895840679362, -122.428071912459)",020110016-B02,3.05
20110016,B04,2003235,Structure Fire,01/11/2002,01/10/2002,Other,01/11/2002 01:51:54 AM,2000 Block of CALIFORNIA ST,SF,94109,B04,38,3362,3,3,3,false,"",1,CHIEF,3,4,5,Pacific Heights,"(37.7895840679362, -122.428071912459)",020110016-B04,2.3166666666666667
20110016,D2,2003235,Structure Fire,01/11/2002,01/10/2002,Other,01/11/2002 01:47:00 AM,2000 Block of CALIFORNIA ST,SF,94109,B04,38,3362,3,3,3,false,"",1,CHIEF,4,4,5,Pacific Heights,"(37.7895840679362, -122.428071912459)",020110016-D2,3.0166666666666666
20110016,E03,2003235,Structure Fire,01/11/2002,01/10/2002,Other,01/11/2002 01:47:00 AM,2000 Block of CALIFORNIA ST,SF,94109,B04,38,3362,3,3,3,false,"",1,ENGINE,7,4,5,Pacific Heights,"(37.7895840679362, -122.428071912459)",020110016-E03,2.

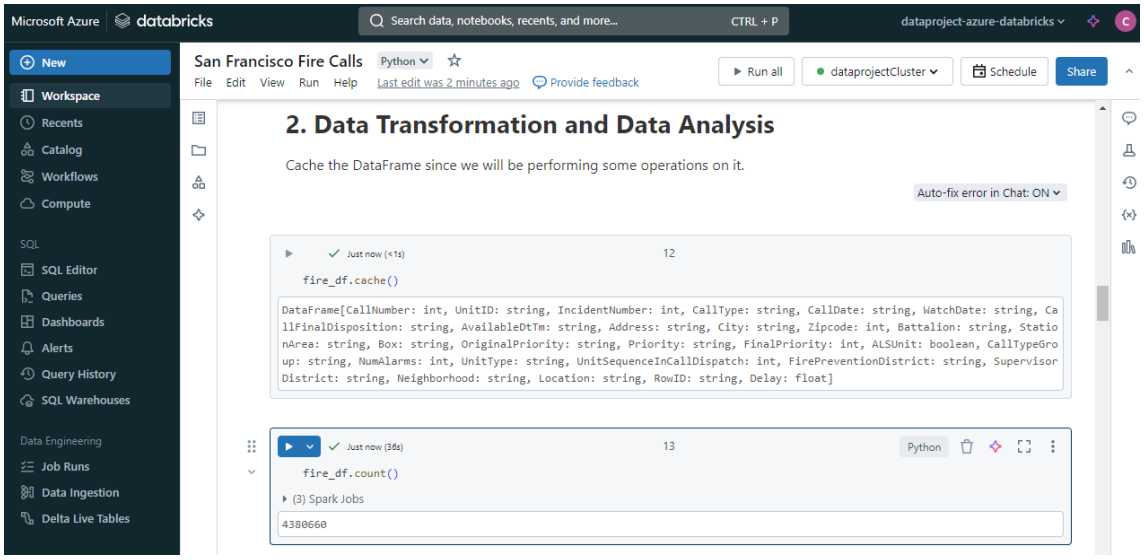


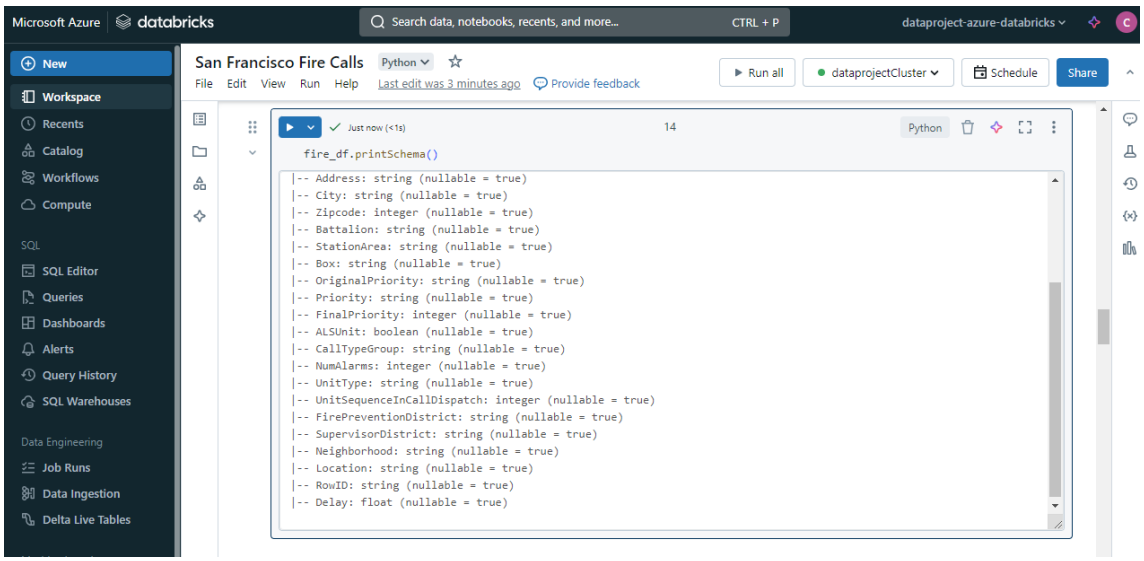



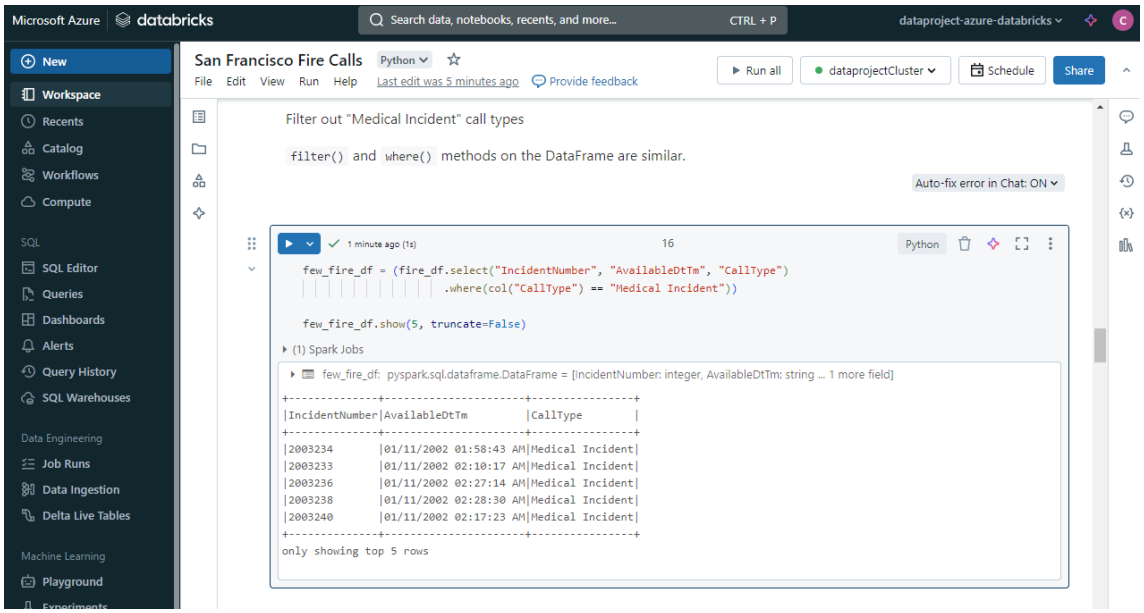



Cleber Zumba
Data Engineer

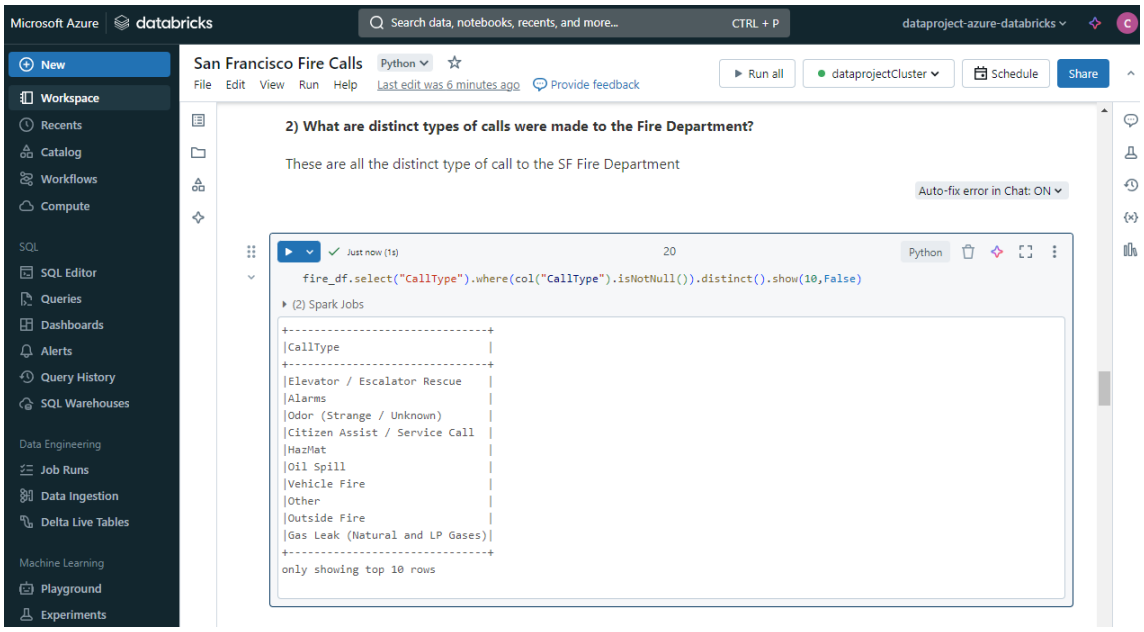
# 9 Data Transformation and Analysis Layer

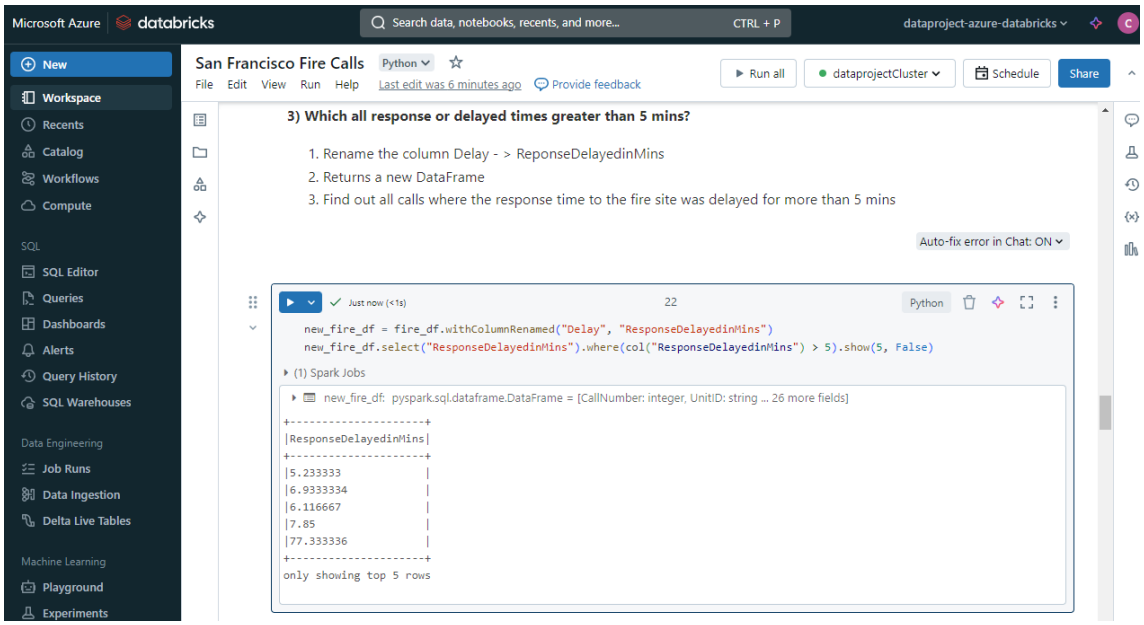


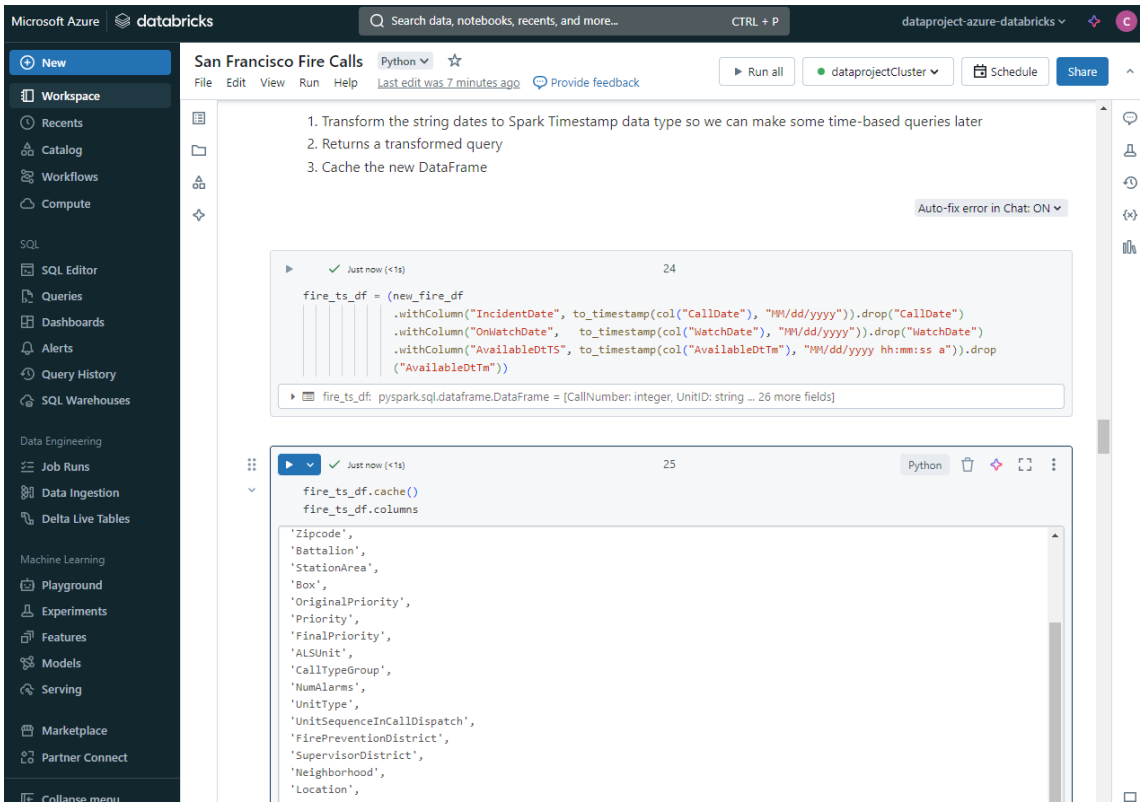



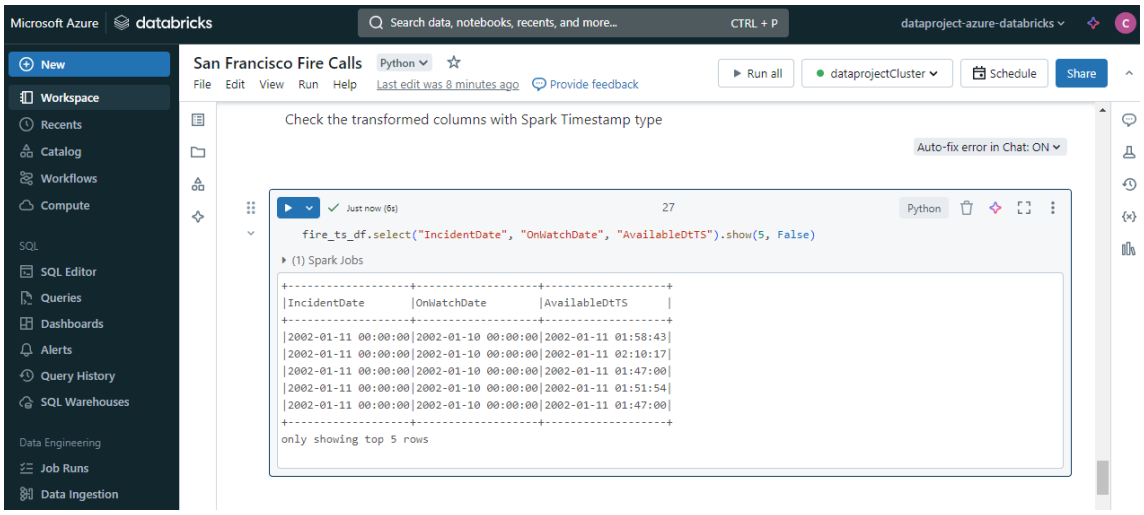

Microsoft Azure
databricks
Search data, notebooks, recents, and more... CTRL + P
dataproject-azure-databricks
New
Workspace
Recents
Catalog
Workflows
Compute
SQL
SQL Editor
Queries
Dashboards
Alerts
Query History
SQL Warehouses
Data Engineering
Job Runs
Data Ingestion
Delta Live Tables
Machine Learning
Playground
San Francisco Fire Calls
Python
File Edit View Run Help
Last edit was 5 minutes ago
Provide feedback
Run all
dataprojectCluster
Schedule
Share
Filter out "Medical Incident" call types
filter() and where() methods on the DataFrame are similar.
Auto-fix error in Chat: ON
1 minute ago (1s)
16
Python
few_fire_df = (fire_df.select("IncidentNumber", "AvailableDtTm", "CallType")
.where(col("CallType") == "Medical Incident"))
few_fire_df.show(5, truncate=False)
(1) Spark Jobs
few_fire_df: pyspark.sql.dataframe.DataFrame = [IncidentNumber: integer, AvailableDtTm: string ... 1 more field]
+--------------+----------------------+---------------+
|IncidentNumber|AvailableDtTm |CallType |
+--------------+----------------------+---------------+
|2003234 |01/11/2002 01:58:43 AM|Medical Incident|
|2003233 |01/11/2002 02:10:17 AM|Medical Incident|
|2003236 |01/11/2002 02:27:14 AM|Medical Incident|
|2003238 |01/11/2002 02:28:30 AM|Medical Incident|
|2003240 |01/11/2002 02:17:23 AM|Medical Incident|
+--------------+----------------------+---------------+
only showing top 5 rows

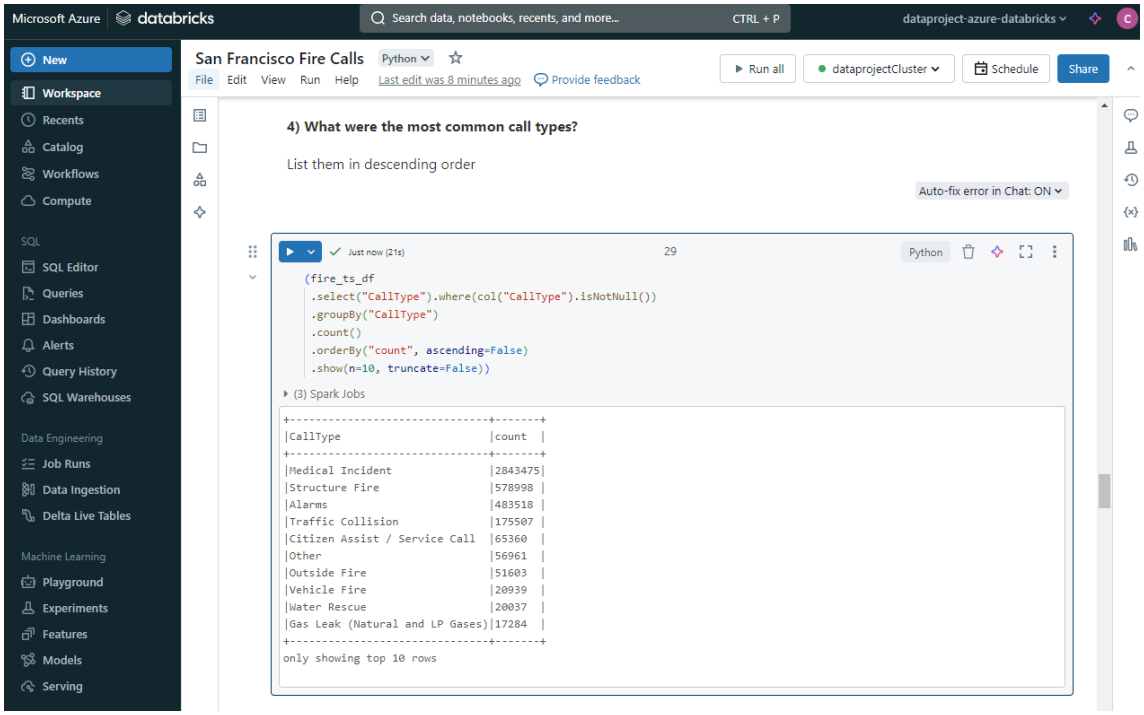

Microsoft Azure
databricks
Search data, notebooks, recents, and more... CTRL + P
dataproject-azure-databricks
New
Workspace
Recents
Catalog
Workflows
Compute
SQL
SQL Editor
Queries
Dashboards
Alerts
Query History
San Francisco Fire Calls
Python
File Edit View Run Help
Last edit was 5 minutes ago
Provide feedback
Run all
dataprojectCluster
Schedule
Share
1) How many distinct types of calls were made to the Fire Department?
To be sure, let's not count "null" strings in that column.
Auto-fix error in Chat: ON
Just now (1s)
18
Python
fire_df.select("CallType").where(col("CallType").isNotNull()).distinct().count()
(3) Spark Jobs
32

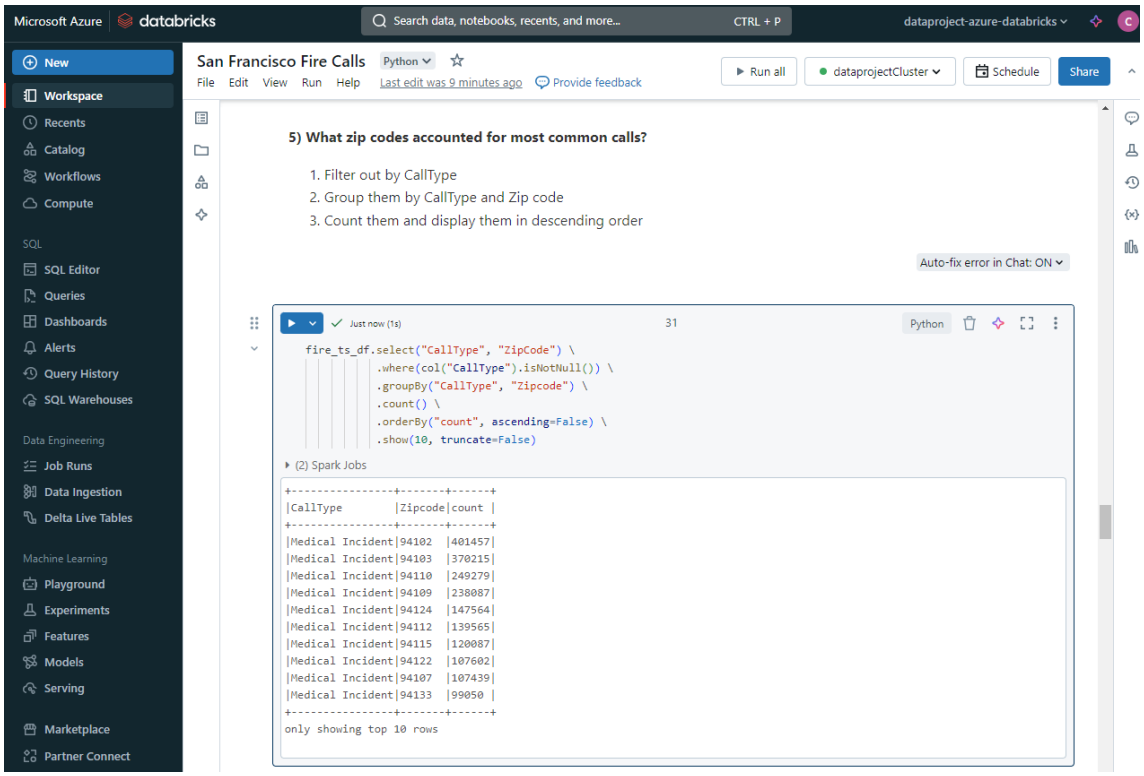

Microsoft Azure
databricks
Search data, notebooks, recents, and more... CTRL + P
dataproject-azure-databricks
New
Workspace
Recents
Catalog
Workflows
Compute
SQL
SQL Editor
Queries
Dashboards
Alerts
Query History
SQL Warehouses
Data Engineering
Job Runs
Data Ingestion
Delta Live Tables
Machine Learning
Playground
Experiments
San Francisco Fire Calls
Python
File Edit View Run Help
Last edit was 6 minutes ago
Provide feedback
Run all
dataprojectCluster
Schedule
Share
2) What are distinct types of calls were made to the Fire Department?
These are all the distinct type of call to the SF Fire Department
Auto-fix error in Chat: ON
Just now (1s)
20
Python
fire_df.select("CallType").where(col("CallType").isNotNull()).distinct().show(10,False)
(2) Spark Jobs
+--------------------------------+
|CallType |
+--------------------------------+
|Elevator / Escalator Rescue |
|Alarms |
|Odor (Strange / Unknown) |
|Citizen Assist / Service Call |
|HazMat |
|Oil Spill |
|Vehicle Fire |
|Other |
|Outside Fire |
|Gas Leak (Natural and LP Gases)|
+--------------------------------+
only showing top 10 rows

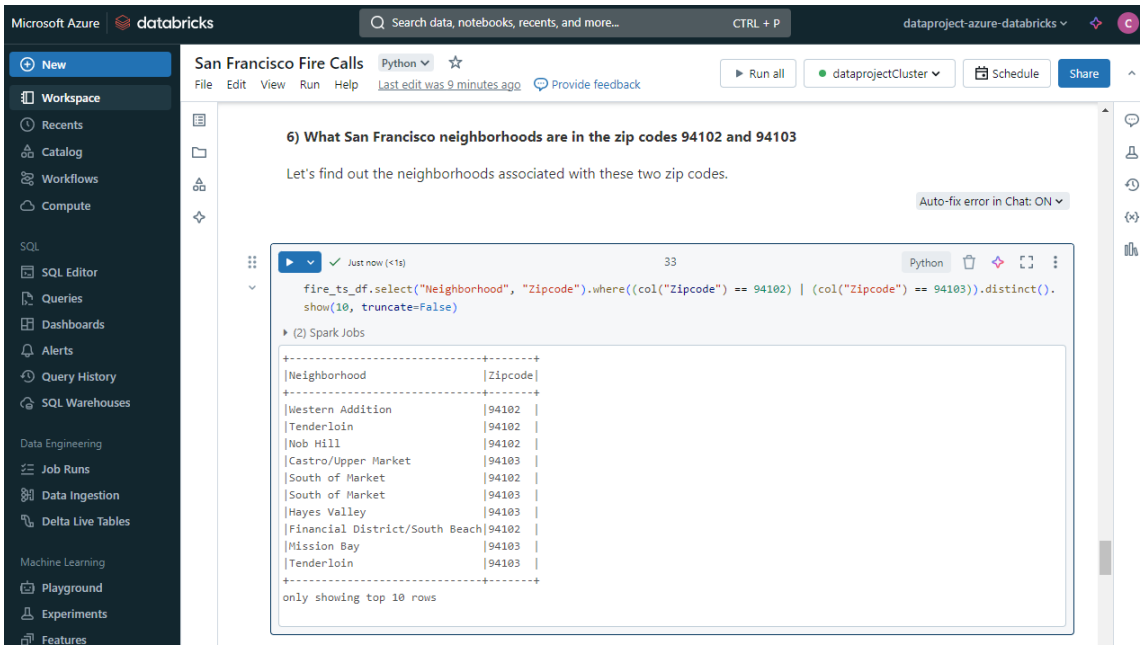

Microsoft Azure
databricks
Search data, notebooks, recents, and more... CTRL + P
dataproject-azure-databricks
New
Workspace
Recents
Catalog
Workflows
Compute
SQL
SQL Editor
Queries
Dashboards
Alerts
Query History
SQL Warehouses
Data Engineering
Job Runs
Data Ingestion
Delta Live Tables
Machine Learning
Playground
Experiments
San Francisco Fire Calls
Python
File Edit View Run Help
Last edit was 6 minutes ago
Provide feedback
Run all
dataprojectCluster
Schedule
Share
3) Which all response or delayed times greater than 5 mins?
1. Rename the column Delay - > ReponseDelayedinMins
2. Returns a new DataFrame
3. Find out all calls where the response time to the fire site was delayed for more than 5 mins
Auto-fix error in Chat: ON
Just now (<1s)
22
Python
new_fire_df = fire_df.withColumnRenamed("Delay", "ResponseDelayedinMins")
new_fire_df.select("ResponseDelayedinMins").where(col("ResponseDelayedinMins") > 5).show(5, False)
(1) Spark Jobs
new_fire_df: pyspark.sql.dataframe.DataFrame = [CallNumber: integer, UnitID: string ... 26 more fields]
+--------------------+
|ResponseDelayedinMins|
+--------------------+
|5.233333 |
|6.9333334 |
|6.116667 |
|7.85 |
|77.333336 |
+--------------------+
only showing top 5 rows

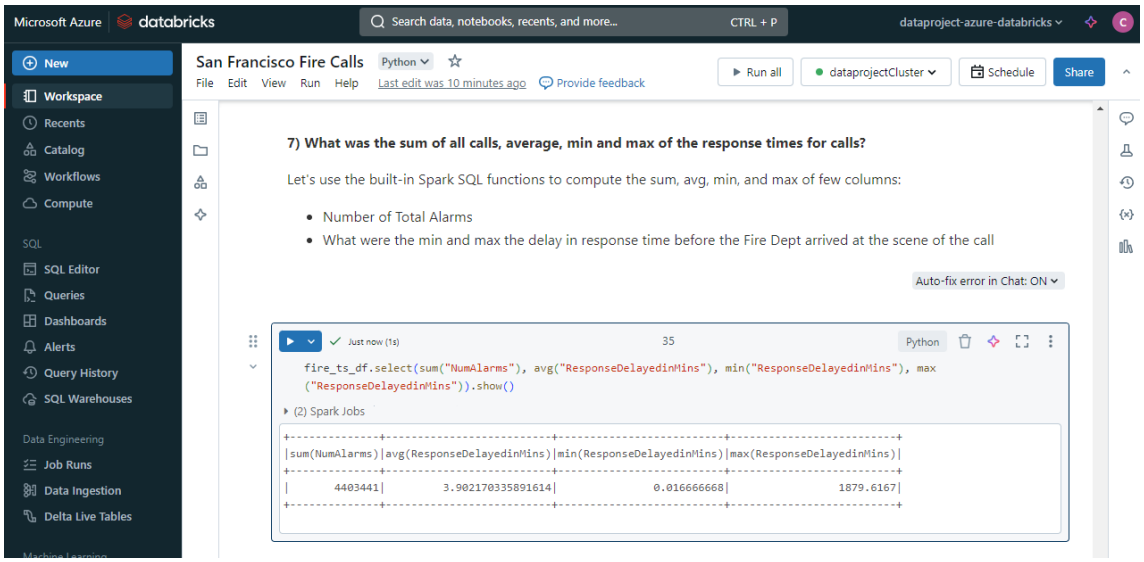

Microsoft Azure
databricks
Search data, notebooks, recents, and more... CTRL + P
dataproject-azure-databricks
New
Workspace
Recents
Catalog
Workflows
Compute
SQL
SQL Editor
Queries
Dashboards
Alerts
Query History
SQL Warehouses
Data Engineering
Job Runs
Data Ingestion
Delta Live Tables
Machine Learning
Playground
Experiments
Features
Models
Serving
Marketplace
Partner Connect
San Francisco Fire Calls
Python
File Edit View Run Help
Last edit was 7 minutes ago
Provide feedback
Run all
dataprojectCluster
Schedule
Share
1. Transform the string dates to Spark Timestamp data type so we can make some time-based queries later
2. Returns a transformed query
3. Cache the new DataFrame
Auto-fix error in Chat: ON
Just now (<1s)
24
fire_ts_df = (new_fire_df
.withColumn("IncidentDate", to_timestamp(col("CallDate"), "MM/dd/yyyy")).drop("CallDate")
.withColumn("OnWatchDate", to_timestamp(col("WatchDate"), "MM/dd/yyyy")).drop("WatchDate")
.withColumn("AvailableDtTS", to_timestamp(col("AvailableDtTm"), "MM/dd/yyyy hh:mm:ss a")).drop
("AvailableDtTm"))
fire_ts_df: pyspark.sql.dataframe.DataFrame = [CallNumber: integer, UnitID: string ... 26 more fields]
Just now (<1s)
25
Python
fire_ts_df.cache()
fire_ts_df.columns
'Zipcode',
'Battalion',
'StationArea',
'Box',
'OriginalPriority',
'Priority',
'FinalPriority',
'ALSUnit',
'CallTypeGroup',
'NumAlarms',
'UnitType',
'UnitSequenceInCallDispatch',
'FirePreventionDistrict',
'SupervisorDistrict',
'Neighborhood',
'Location',

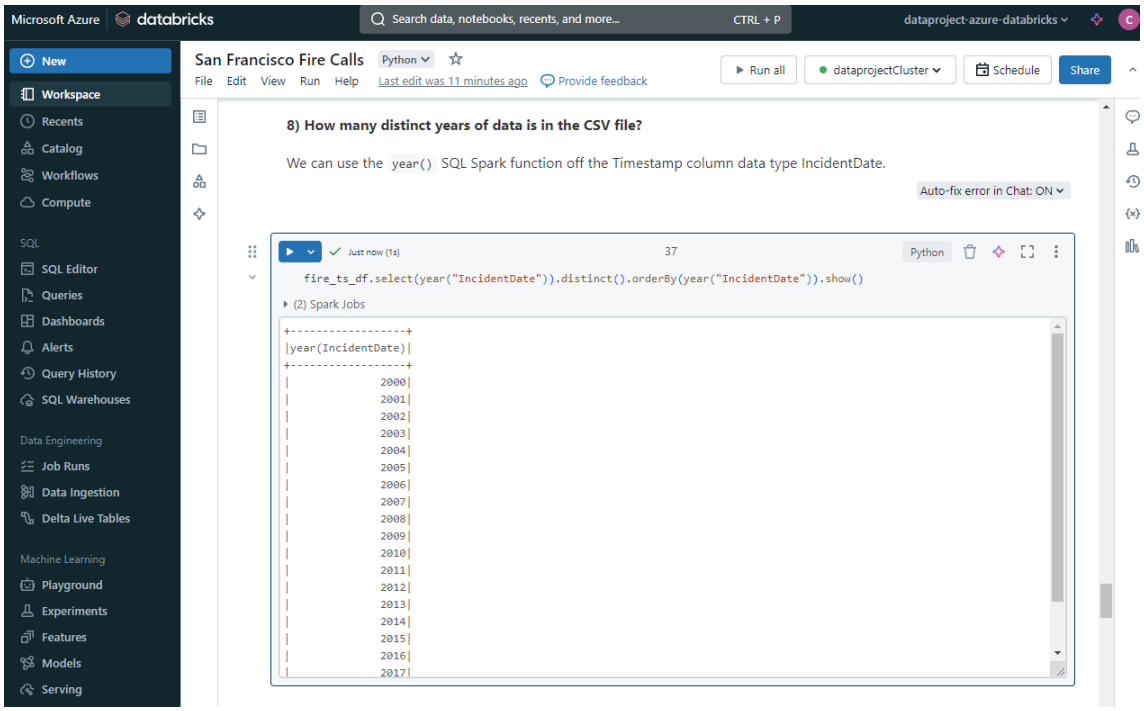

Microsoft Azure
databricks
Search data, notebooks, recents, and more...
CTRL + P
dataproject-azure-databricks
New
Workspace
Recents
Catalog
Workflows
Compute
SQL
SQL Editor
Queries
Dashboards
Alerts
Query History
SQL Warehouses
Data Engineering
Job Runs
Data Ingestion
San Francisco Fire Calls
Python
File Edit View Run Help
Last edit was 8 minutes ago
Provide feedback
Run all
dataprojectCluster
Schedule
Share
Check the transformed columns with Spark Timestamp type
Auto-fix error in Chat: ON
Just now (6s)
27
Python
fire_ts_df.select("IncidentDate", "OnWatchDate", "AvailableDtTS").show(5, False)
(1) Spark Jobs
+-------------------+-------------------+-------------------+
|IncidentDate |OnWatchDate |AvailableDtTS |
+-------------------+-------------------+-------------------+
|2002-01-11 00:00:00|2002-01-10 00:00:00|2002-01-11 01:58:43|
|2002-01-11 00:00:00|2002-01-10 00:00:00|2002-01-11 02:10:17|
|2002-01-11 00:00:00|2002-01-10 00:00:00|2002-01-11 01:47:00|
|2002-01-11 00:00:00|2002-01-10 00:00:00|2002-01-11 01:51:54|
|2002-01-11 00:00:00|2002-01-10 00:00:00|2002-01-11 01:47:00|
+-------------------+-------------------+-------------------+
only showing top 5 rows

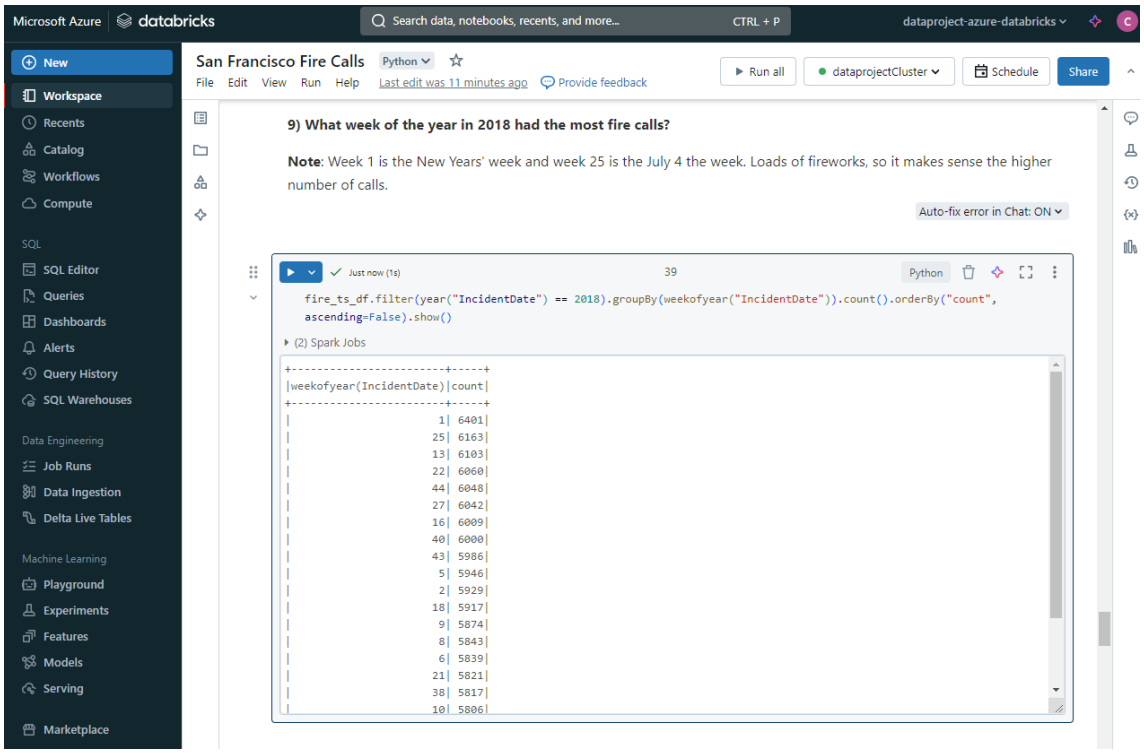

Microsoft Azure
databricks
Search data, notebooks, recents, and more...
CTRL + P
dataproject-azure-databricks
New
Workspace
Recents
Catalog
Workflows
Compute
SQL
SQL Editor
Queries
Dashboards
Alerts
Query History
SQL Warehouses
Data Engineering
Job Runs
Data Ingestion
Delta Live Tables
Machine Learning
Playground
Experiments
Features
Models
Serving
San Francisco Fire Calls
Python
File Edit View Run Help
Last edit was 8 minutes ago
Provide feedback
Run all
dataprojectCluster
Schedule
Share
4) What were the most common call types?
List them in descending order
Auto-fix error in Chat: ON
Just now (21s)
29
Python
(fire_ts_df
.select("CallType").where(col("CallType").isNotNull())
.groupBy("CallType")
.count()
.orderBy("count", ascending=False)
.show(n=10, truncate=False))
(3) Spark Jobs
+---------------------------------+-------+
|CallType |count |
+---------------------------------+-------+
|Medical Incident |2843475|
|Structure Fire |578998 |
|Alarms |483518 |
|Traffic Collision |175507 |
|Citizen Assist / Service Call |65360 |
|Other |56961 |
|Outside Fire |51603 |
|Vehicle Fire |20939 |
|Water Rescue |20037 |
|Gas Leak (Natural and LP Gases)|17284 |
+---------------------------------+-------+
only showing top 10 rows

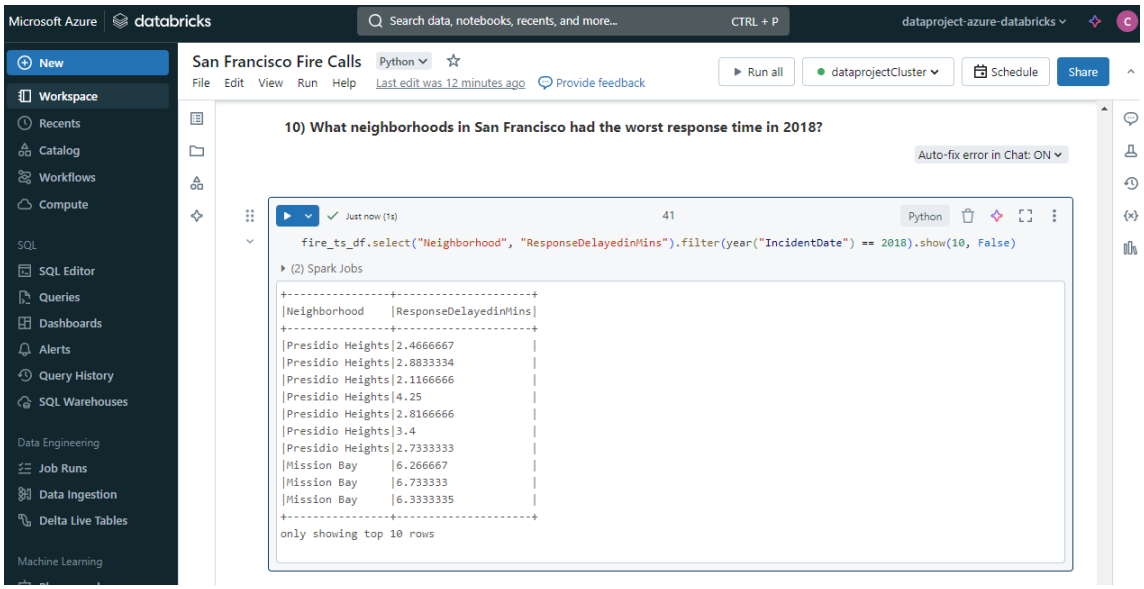

Microsoft Azure
databricks
Search data, notebooks, recents, and more... CTRL + P
dataproject-azure-databricks
New
Workspace
Recents
Catalog
Workflows
Compute
SQL
SQL Editor
Queries
Dashboards
Alerts
Query History
SQL Warehouses
Data Engineering
Job Runs
Data Ingestion
Delta Live Tables
Machine Learning
Playground
Experiments
Features
Models
Serving
Marketplace
Partner Connect
San Francisco Fire Calls
Python
File Edit View Run Help
Last edit was 9 minutes ago
Provide feedback
Run all
dataprojectCluster
Schedule
Share
5) What zip codes accounted for most common calls?
1. Filter out by CallType
2. Group them by CallType and Zip code
3. Count them and display them in descending order
Auto-fix error in Chat: ON
Just now (1s)
31
Python
fire_ts_df.select("CallType", "ZipCode") \
.where(col("CallType").isNotNull()) \
.groupBy("CallType", "Zipcode") \
.count() \
.orderBy("count", ascending=False) \
.show(10, truncate=False)
(2) Spark Jobs
+----------------+-------+------+
|CallType |Zipcode|count |
+----------------+-------+------+
|Medical Incident|94102 |401457|
|Medical Incident|94103 |370215|
|Medical Incident|94110 |249279|
|Medical Incident|94109 |238087|
|Medical Incident|94124 |147564|
|Medical Incident|94112 |139565|
|Medical Incident|94115 |120087|
|Medical Incident|94122 |107602|
|Medical Incident|94107 |107439|
|Medical Incident|94133 |99050 |
+----------------+-------+------+
only showing top 10 rows


Microsoft Azure
databricks
Search data, notebooks, recents, and more... CTRL + P
dataproject-azure-databricks
New
Workspace
Recents
Catalog
Workflows
Compute
SQL
SQL Editor
Queries
Dashboards
Alerts
Query History
SQL Warehouses
Data Engineering
Job Runs
Data Ingestion
Delta Live Tables
Machine Learning
Playground
Experiments
Features
San Francisco Fire Calls
Python
File Edit View Run Help
Last edit was 9 minutes ago
Provide feedback
Run all
dataprojectCluster
Schedule
Share
6) What San Francisco neighborhoods are in the zip codes 94102 and 94103
Let's find out the neighborhoods associated with these two zip codes.
Auto-fix error in Chat: ON
Just now (<1s)
33
Python
fire_ts_df.select("Neighborhood", "Zipcode").where((col("Zipcode") == 94102) | (col("Zipcode") == 94103)).distinct().
show(10, truncate=False)
(2) Spark Jobs
+------------------------------+-------+
|Neighborhood |Zipcode|
+------------------------------+-------+
|Western Addition |94102 |
|Tenderloin |94102 |
|Nob Hill |94102 |
|Castro/Upper Market |94103 |
|South of Market |94102 |
|South of Market |94103 |
|Hayes Valley |94103 |
|Financial District/South Beach|94102 |
|Mission Bay |94103 |
|Tenderloin |94103 |
+------------------------------+-------+
only showing top 10 rows

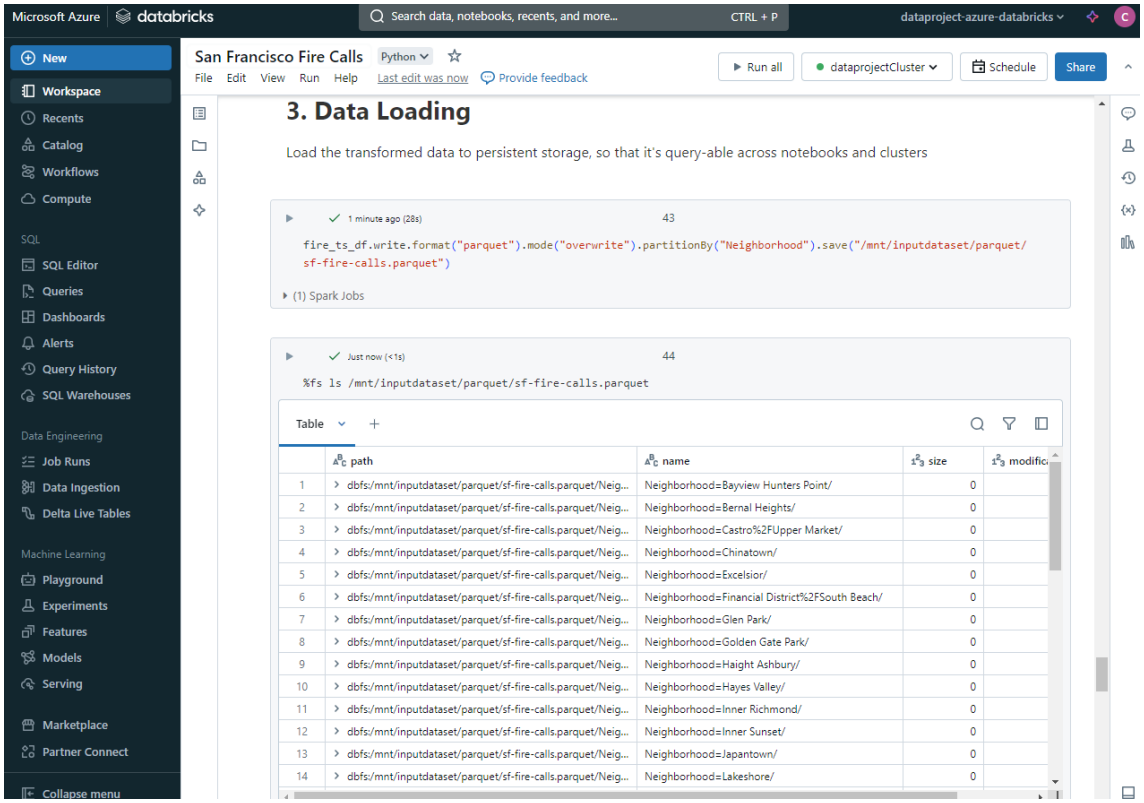

Microsoft Azure
databricks
Search data, notebooks, recents, and more... CTRL + P
dataproject-azure-databricks
New
Workspace
Recents
Catalog
Workflows
Compute
SQL
SQL Editor
Queries
Dashboards
Alerts
Query History
SQL Warehouses
Data Engineering
Job Runs
Data Ingestion
Delta Live Tables
San Francisco Fire Calls
Python
File Edit View Run Help
Last edit was 10 minutes ago
Provide feedback
Run all
dataprojectCluster
Schedule
Share
7) What was the sum of all calls, average, min and max of the response times for calls?
Let's use the built-in Spark SQL functions to compute the sum, avg, min, and max of few columns:
Number of Total Alarms
What were the min and max the delay in response time before the Fire Dept arrived at the scene of the call
Auto-fix error in Chat: ON
Just now (1s)
35
Python
fire_ts_df.select(sum("NumAlarms"), avg("ResponseDelayedinMins"), min("ResponseDelayedinMins"), max("ResponseDelayedinMins")).show()
(2) Spark Jobs
+-------------+-------------------------+-------------------------+-------------------------+
|sum(NumAlarms)|avg(ResponseDelayedinMins)|min(ResponseDelayedinMins)|max(ResponseDelayedinMins)|
+-------------+-------------------------+-------------------------+-------------------------+
| 4403441| 3.902170335891614| 0.016666668| 1879.6167|
+-------------+-------------------------+-------------------------+-------------------------+


Microsoft Azure
databricks
Search data, notebooks, recents, and more... CTRL + P
dataproject-azure-databricks
New
Workspace
Recents
Catalog
Workflows
Compute
SQL
SQL Editor
Queries
Dashboards
Alerts
Query History
SQL Warehouses
Data Engineering
Job Runs
Data Ingestion
Delta Live Tables
Machine Learning
Playground
Experiments
Features
Models
Serving
San Francisco Fire Calls
Python
File Edit View Run Help
Last edit was 11 minutes ago
Provide feedback
Run all
dataprojectCluster
Schedule
Share
8) How many distinct years of data is in the CSV file?
We can use the year() SQL Spark function off the Timestamp column data type IncidentDate.
Auto-fix error in Chat: ON
Just now (1s)
37
Python
fire_ts_df.select(year("IncidentDate")).distinct().orderBy(year("IncidentDate")).show()
(2) Spark Jobs
+------------------+
|year(IncidentDate)|
+------------------+
| 2000|
| 2001|
| 2002|
| 2003|
| 2004|
| 2005|
| 2006|
| 2007|
| 2008|
| 2009|
| 2010|
| 2011|
| 2012|
| 2013|
| 2014|
| 2015|
| 2016|
| 2017|

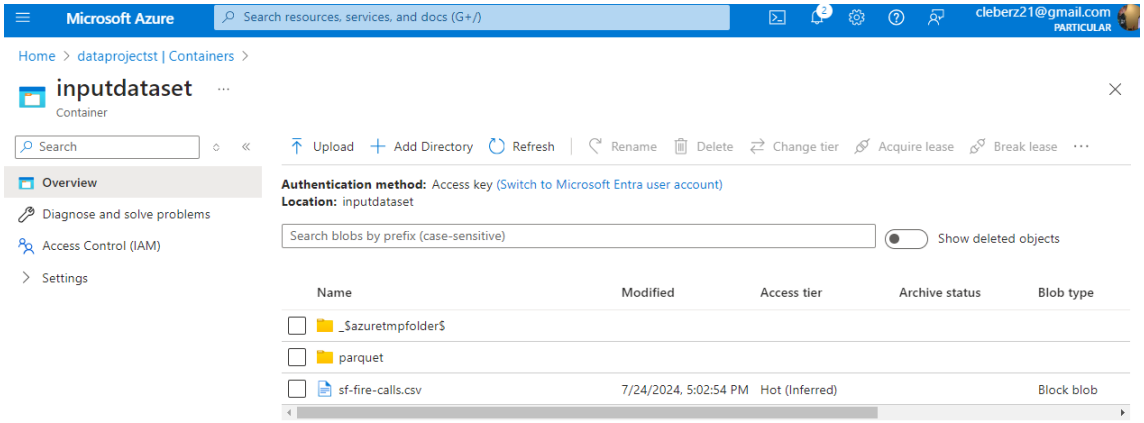

Microsoft Azure
databricks
Search data, notebooks, recents, and more... CTRL + P
dataproject-azure-databricks
New
Workspace
Recents
Catalog
Workflows
Compute
SQL
SQL Editor
Queries
Dashboards
Alerts
Query History
SQL Warehouses
Data Engineering
Job Runs
Data Ingestion
Delta Live Tables
Machine Learning
Playground
Experiments
Features
Models
Serving
Marketplace
San Francisco Fire Calls
Python
File Edit View Run Help
Last edit was 11 minutes ago
Provide feedback
Run all
dataprojectCluster
Schedule
Share
9) What week of the year in 2018 had the most fire calls?
Note: Week 1 is the New Years' week and week 25 is the July 4 the week. Loads of fireworks, so it makes sense the higher number of calls.
Auto-fix error in Chat: ON
Just now (1s)
39
Python
fire_ts_df.filter(year("IncidentDate") == 2018).groupBy(weekofyear("IncidentDate")).count().orderBy("count", ascending=False).show()
(2) Spark Jobs
+------------------------+-----+
|weekofyear(IncidentDate)|count|
+------------------------+-----+
| 1| 6401|
| 25| 6163|
| 13| 6103|
| 22| 6060|
| 44| 6048|
| 27| 6042|
| 16| 6009|
| 40| 6000|
| 43| 5986|
| 5| 5946|
| 2| 5929|
| 18| 5917|
| 9| 5874|
| 8| 5843|
| 6| 5839|
| 21| 5821|
| 38| 5817|
| 10| 5806|


Microsoft Azure
databricks
Search data, notebooks, recents, and more... CTRL + P
dataproject-azure-databricks
New
Workspace
Recents
Catalog
Workflows
Compute
SQL
SQL Editor
Queries
Dashboards
Alerts
Query History
SQL Warehouses
Data Engineering
Job Runs
Data Ingestion
Delta Live Tables
Machine Learning
San Francisco Fire Calls
Python
File Edit View Run Help
Last edit was 12 minutes ago
Provide feedback
Run all
dataprojectCluster
Schedule
Share
10) What neighborhoods in San Francisco had the worst response time in 2018?
Auto-fix error in Chat: ON
Just now (1s)
41
Python
fire_ts_df.select("Neighborhood", "ResponseDelayedinMins").filter(year("IncidentDate") == 2018).show(10, False)
(2) Spark Jobs
+---------------+--------------------+
|Neighborhood |ResponseDelayedinMins|
+---------------+--------------------+
|Presidio Heights|2.4666667 |
|Presidio Heights|2.8833334 |
|Presidio Heights|2.1166666 |
|Presidio Heights|4.25 |
|Presidio Heights|2.8166666 |
|Presidio Heights|3.4 |
|Presidio Heights|2.7333333 |
|Mission Bay |6.266667 |
|Mission Bay |6.733333 |
|Mission Bay |6.3333335 |
+---------------+--------------------+
only showing top 10 rows

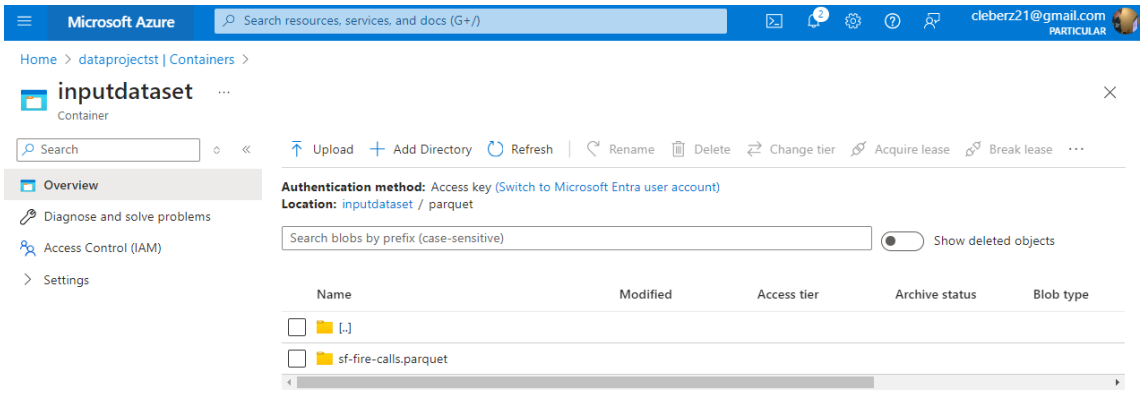
# 10 Data Storage Layer



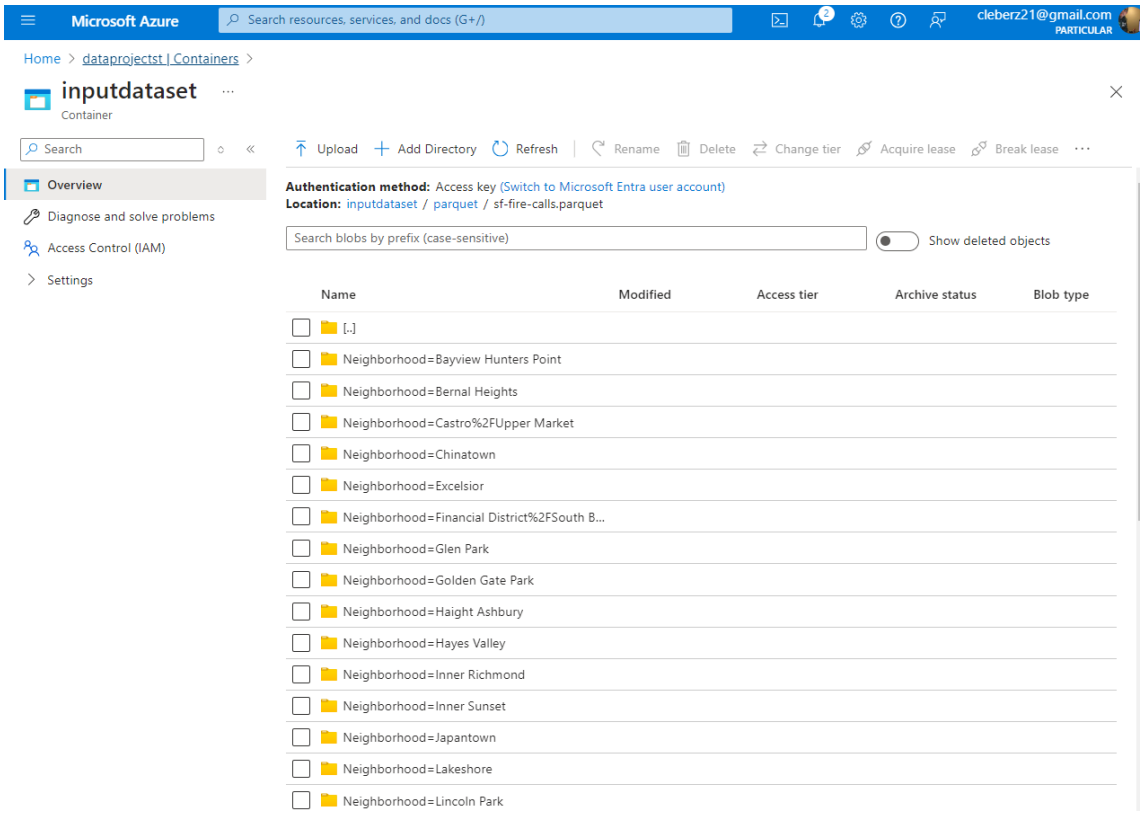

Saving data in parquet format to the Databricks file system.



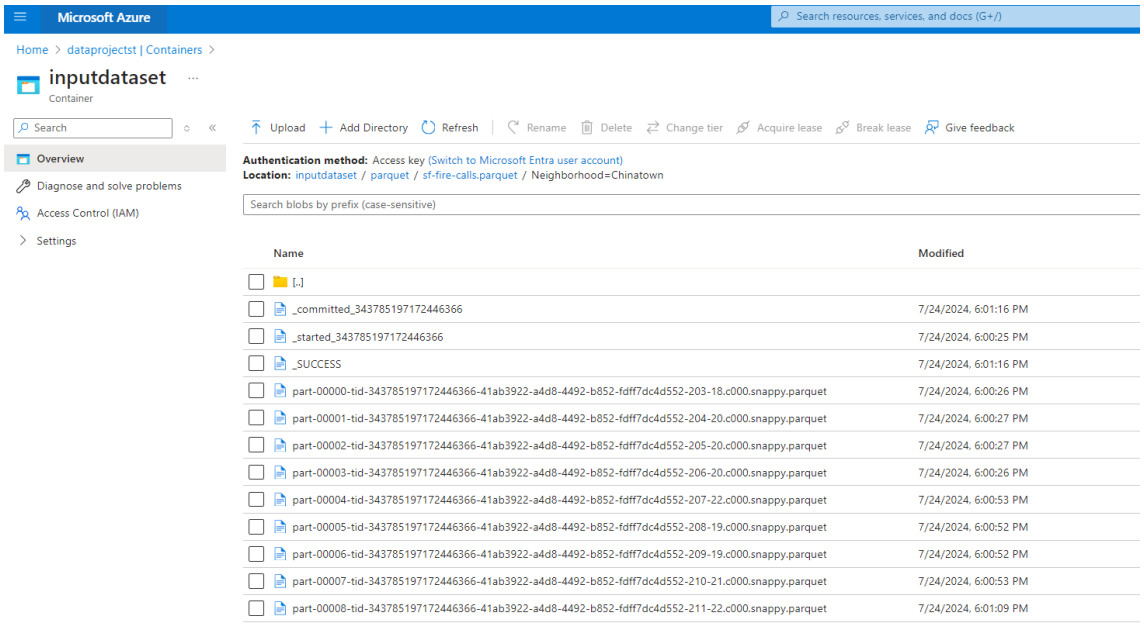

Data saved in parquet format in the Azure Data Lake Storage (ADLS) container.



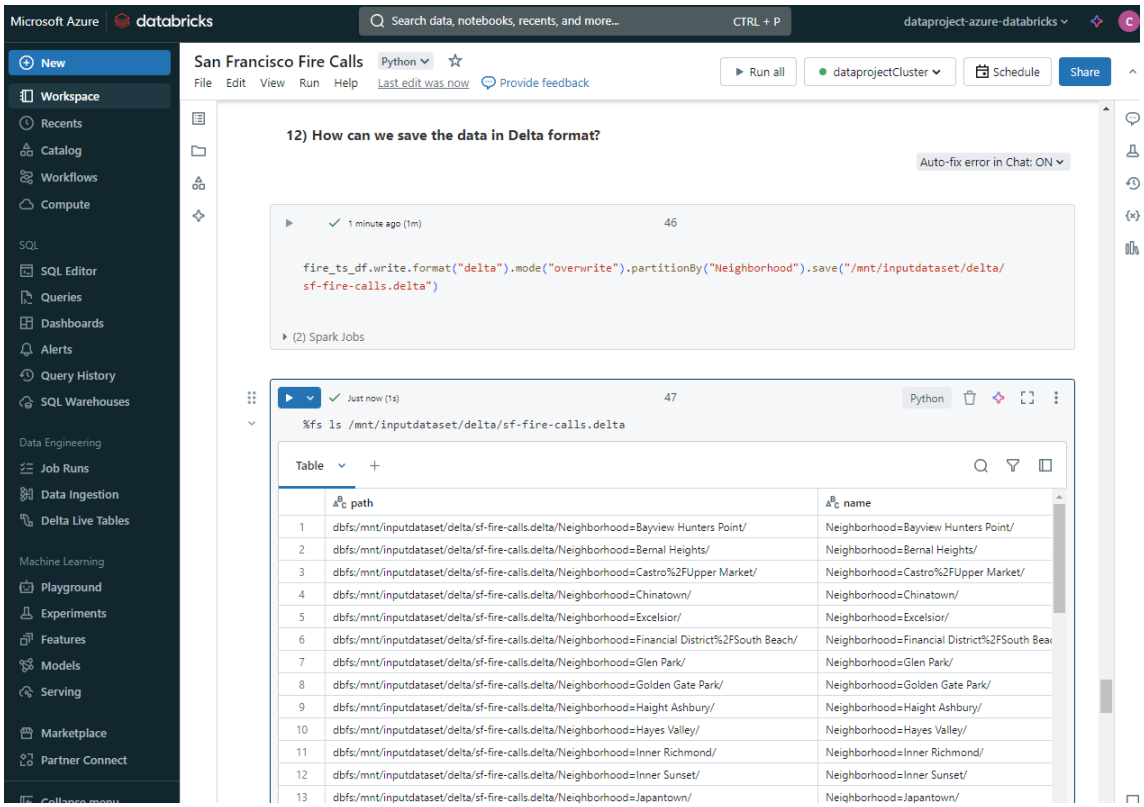

Checking the data saved in the container



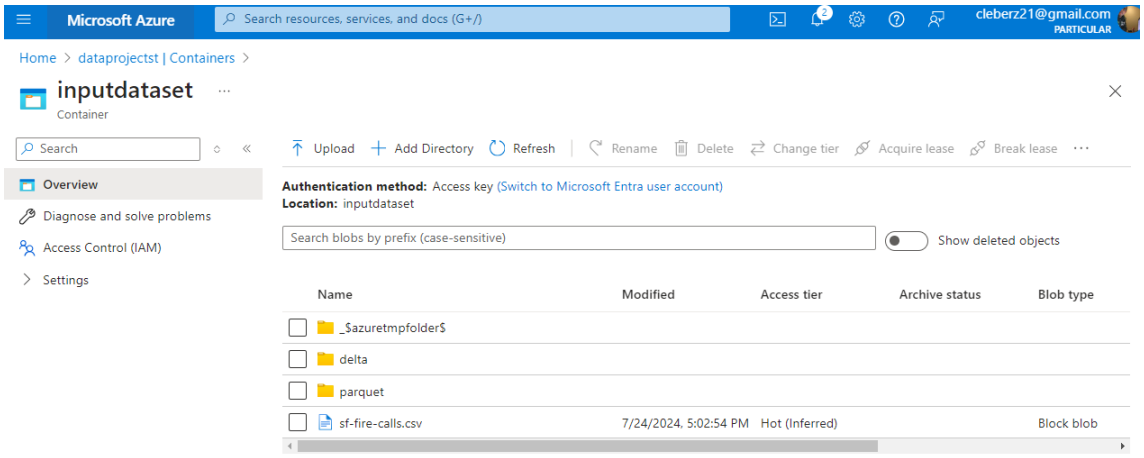

Partitioned data in the container



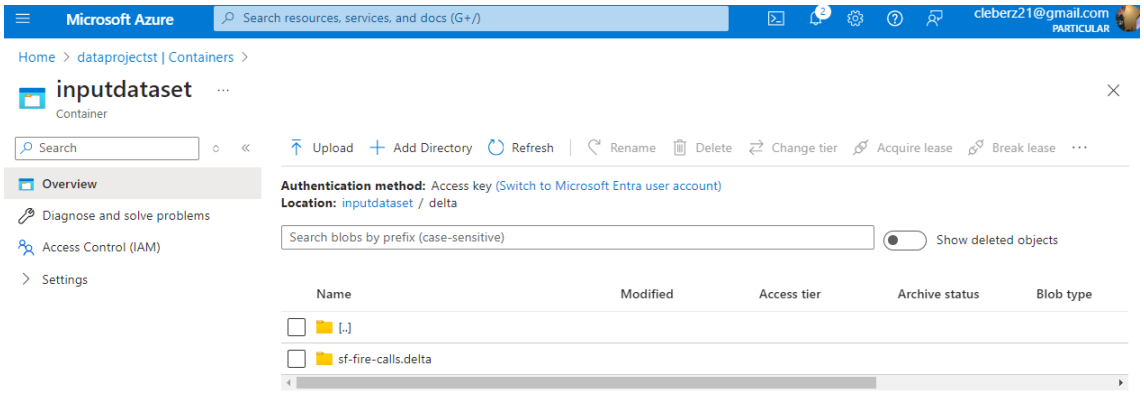

Partitioned data in the container



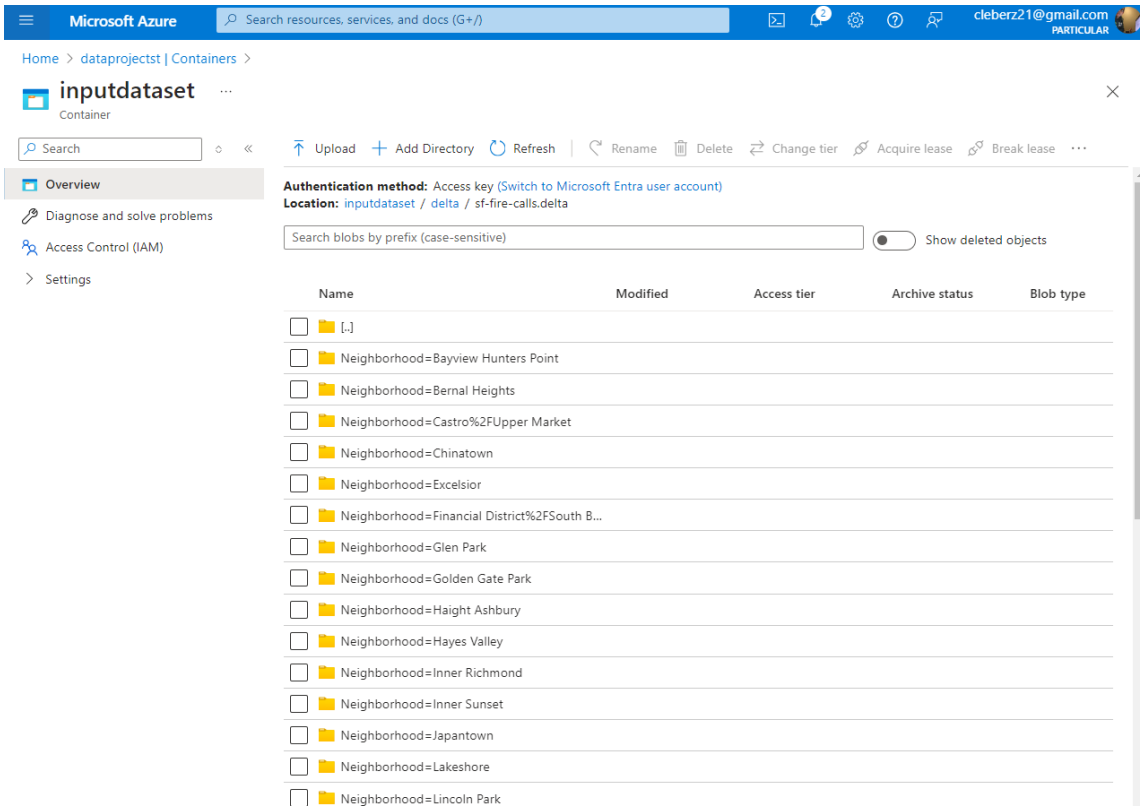

Saving data in delta format to the Databricks file system.



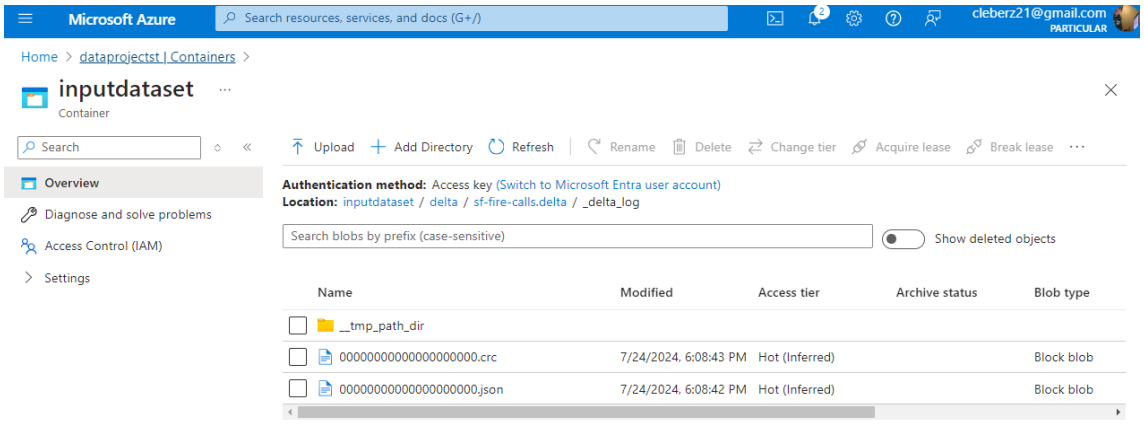

Data saved in delta format in Azure Data Lake Storage (ADLS) container.



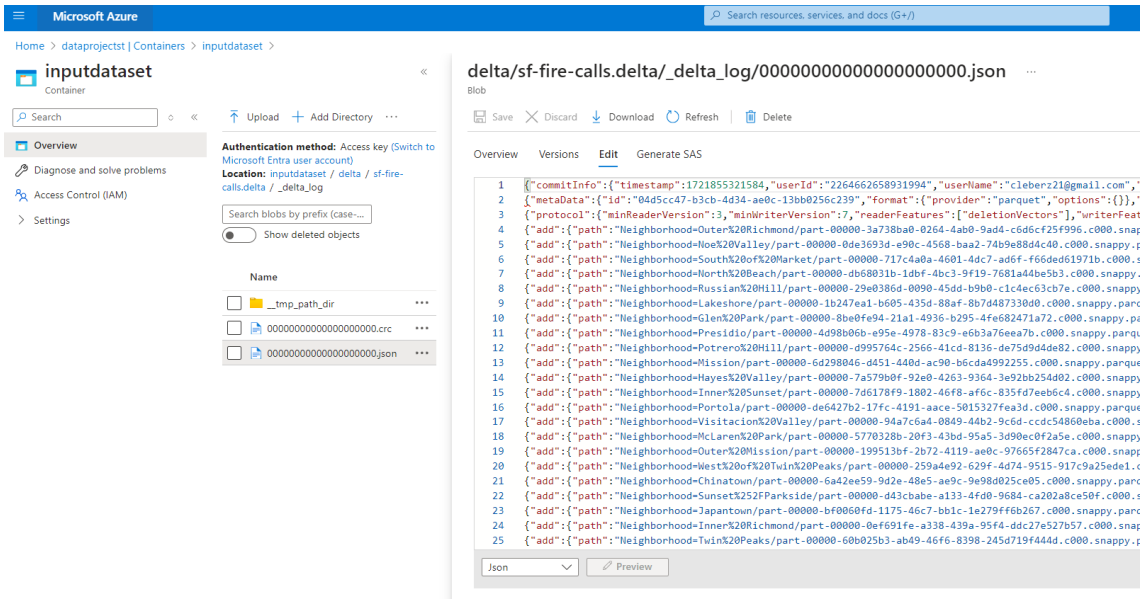

Checking the data saved in the container




Partitioned data in the container

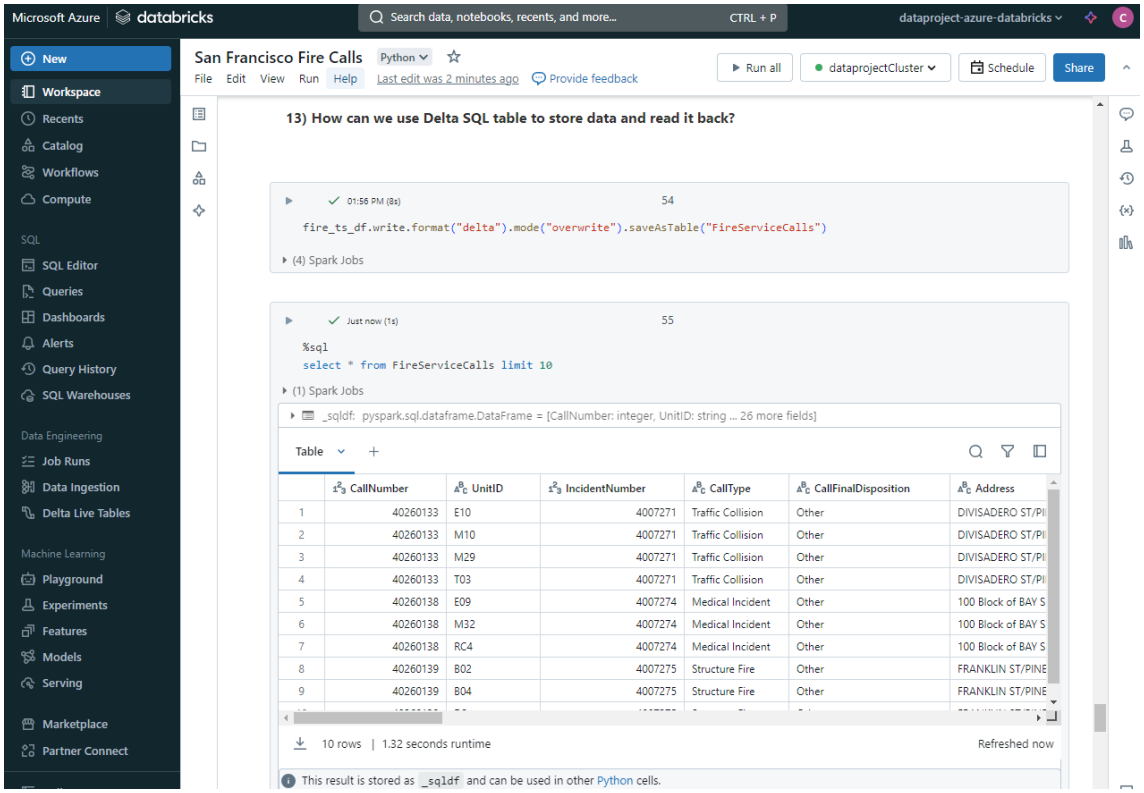



Although the data is saved in Delta format, the underlying data files are in JSON format in the container.

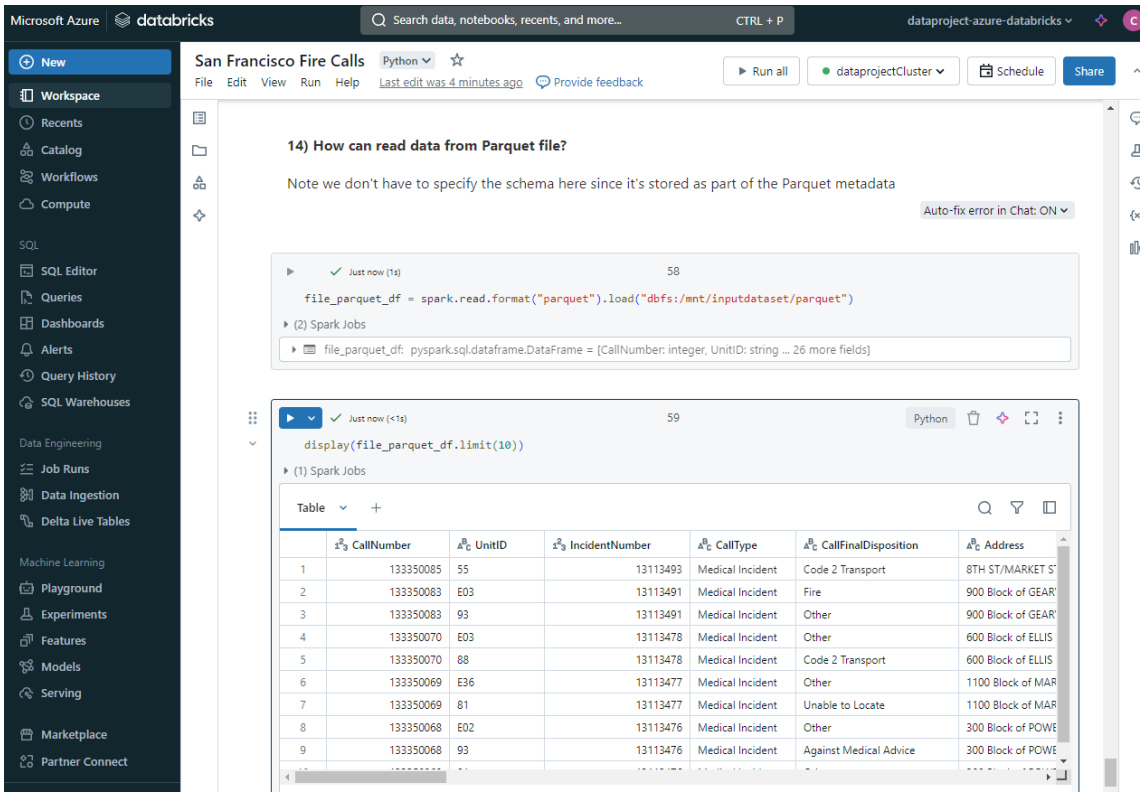



Viewing data in the container.

# 11 Data Query Layer




Querying the result of data transformation and analysis in delta format.




Querying the result of data transformation and analysis in parquet format.

# 12 Conclusion

In this project, I leveraged the powerful capabilities of AWS, Azure, and Databricks to build a robust and scalable data pipeline focused on analyzing and transforming emergency call data from the San Francisco Fire Department.

Performed data extraction, transformation, and loading (ETL) processes, highlighting the seamless integration between AWS S3, Azure Data Lake Storage (ADLS), and Databricks for scalable data processing. We used Azure Data Factory to orchestrate data movement and ensure secure transfer between cloud environments.

In Databricks, I used PySpark and SparkSQL to execute queries and transformations, demonstrating their ability to handle large-scale data analysis with ease. I implemented performance optimization techniques, such as caching, which are essential for achieving high-performance data processing.

I followed best practices using the advanced features of Azure Key Vault for secure secrets management and Databricks for distributed processing.

Cleber Zumba de Souza

# 13 Reference

PARSIAN, Mahmoud. **Data Algorithms with Spark**. Sebastopol, California, United States: O'Reilly Media, 2022.