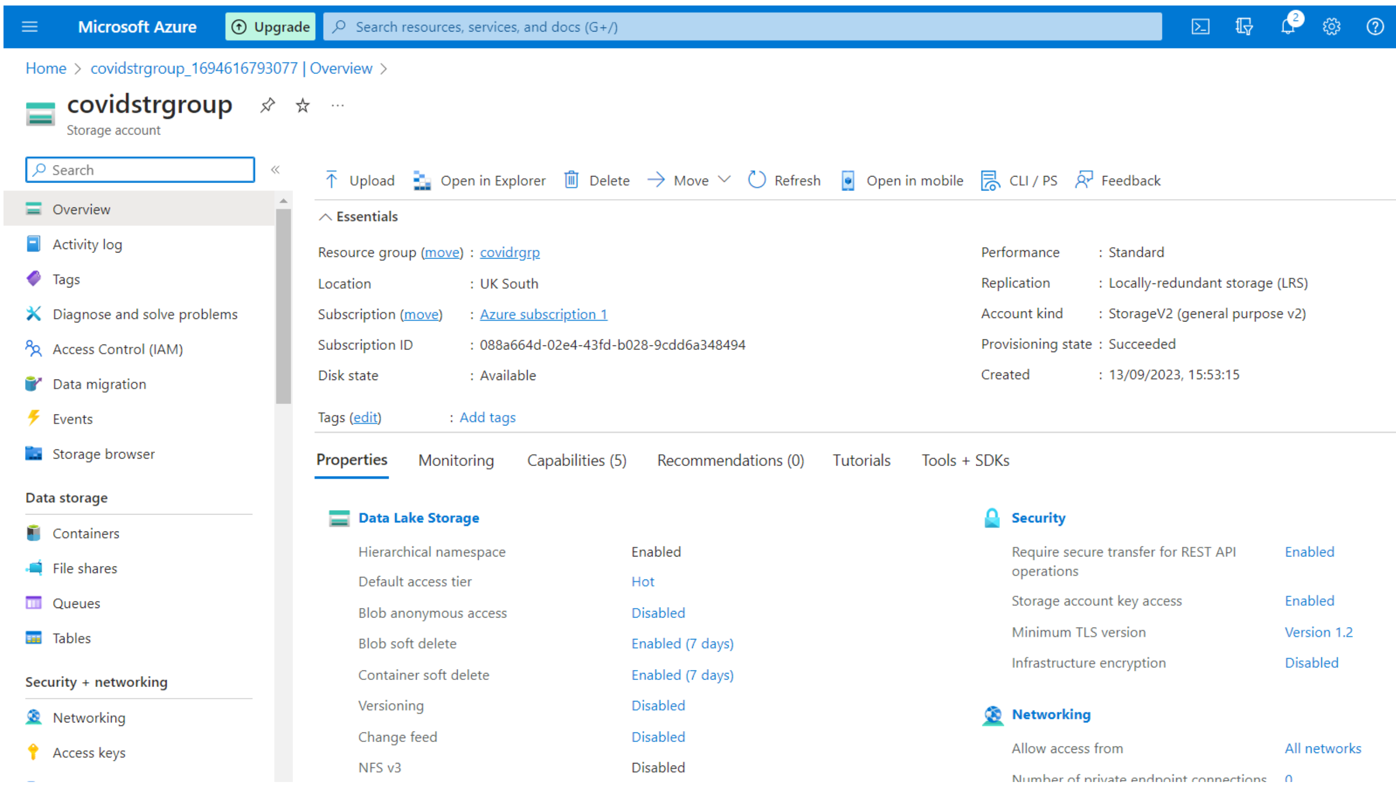
Imran Iqbal Microsoft Azure Profile Project

This is a Microsoft Azure Profile Project created by me and the idea is to upload Covid.xlsx in the datalake storage gen 2 and create a pipeline in Azure synapse to convert xlsx format to parquet.

(Parquet is optimized to work with complex data in bulk and features different ways for efficient data compression and encoding types. This approach is best especially for those queries that need to read certain columns from a large table)

Then using Azure synapse, Apache spark pool is created that would allocate resource to notebook which would be used to read the parquet file , create database and tables to perform aggregation functions on the file which is then saved on lake database.

First created a storage account and enabled it to be Azure data lake Gen 2.



covidstrgroup
Storage account

Essentials

Resource group (move)	: covidrgrp
Location	: UK South
Subscription (move)	: Azure subscription 1
Subscription ID	: 088a664d-02e4-43fd-b028-9cdd6a348494
Disk state	: Available
Performance	: Standard
Replication	: Locally-redundant storage (LRS)
Account kind	: StorageV2 (general purpose v2)
Provisioning state	: Succeeded
Created	: 13/09/2023, 15:53:15
Tags (edit)	: Add tags

Data Lake Storage

Hierarchical namespace	Enabled
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

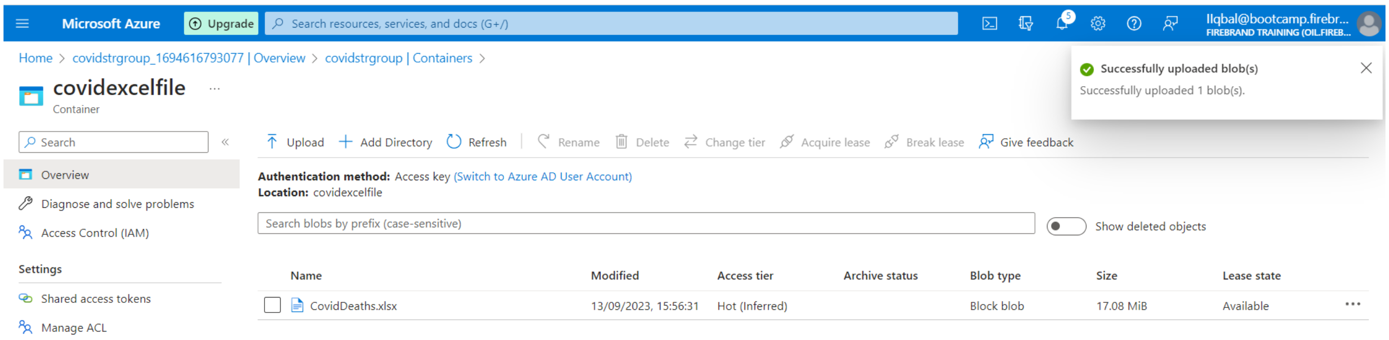
Security

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

Uploaded Covid.xlsx which is excel format into a folder



Microsoft Azure | Upgrade | Search resources, services, and docs (G+/) | Ilqbal@bootcamp.firebr... FIREBRAND TRAINING (OIL.FIREB...

Home > covidstrgroup_1694616793077 | Overview > covidstrgroup | Containers >

covidexcelfile
Container

Successfully uploaded blob(s)
Successfully uploaded 1 blob(s).

Search | Overview | Diagnose and solve problems | Access Control (IAM)

Settings: Shared access tokens | Manage ACL | Access policy | Properties | Metadata

Upload | Add Directory | Refresh | Rename | Delete | Change tier | Acquire lease | Break lease | Give feedback

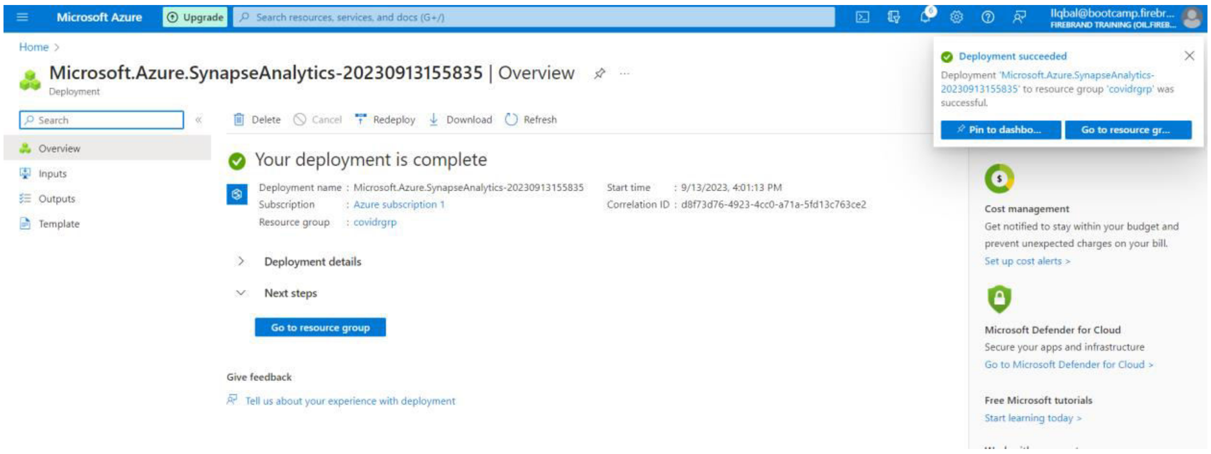
Authentication method: Access key (Switch to Azure AD User Account)
Location: covidexcelfile

Search blobs by prefix (case-sensitive) | Show deleted objects

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
CovidDeaths.xlsx	13/09/2023, 15:56:31	Hot (Inferred)		Block blob	17.08 MiB	Available

Then created a synapse space to run a pipeline which converts excel format of a file to parquet as parquet is columnar based format and best used for query purpose.

To accomplish this Synapse workspace is created and assigned myself (the Storage Blob Data Contributor role on the Data Lake Storage Gen2 account to interactively query it in the workspace).



Microsoft Azure | Upgrade | Search resources, services, and docs (G+/) | Ilqbal@bootcamp.firebr... FIREBRAND TRAINING (OIL.FIREB...

Home >

Microsoft.Azure.SynapseAnalytics-20230913155835 | Overview
Deployment

Search | Overview | Inputs | Outputs | Template

Delete | Cancel | Redeploy | Download | Refresh

Deployment succeeded
Deployment 'Microsoft.Azure.SynapseAnalytics-20230913155835' to resource group 'covidrgrp' was successful.
Pin to dashbo... | Go to resource gr...

Your deployment is complete

```
Deployment name : Microsoft.Azure.SynapseAnalytics-20230913155835    Start time      : 9/13/2023, 4:01:13 PM
Subscription    : Azure subscription 1                               Correlation ID  : d8f73d76-4923-4cc0-a71a-5fd13c763ce2
Resource group  : covidrgrp
```

Deployment details

Next steps

Go to resource group

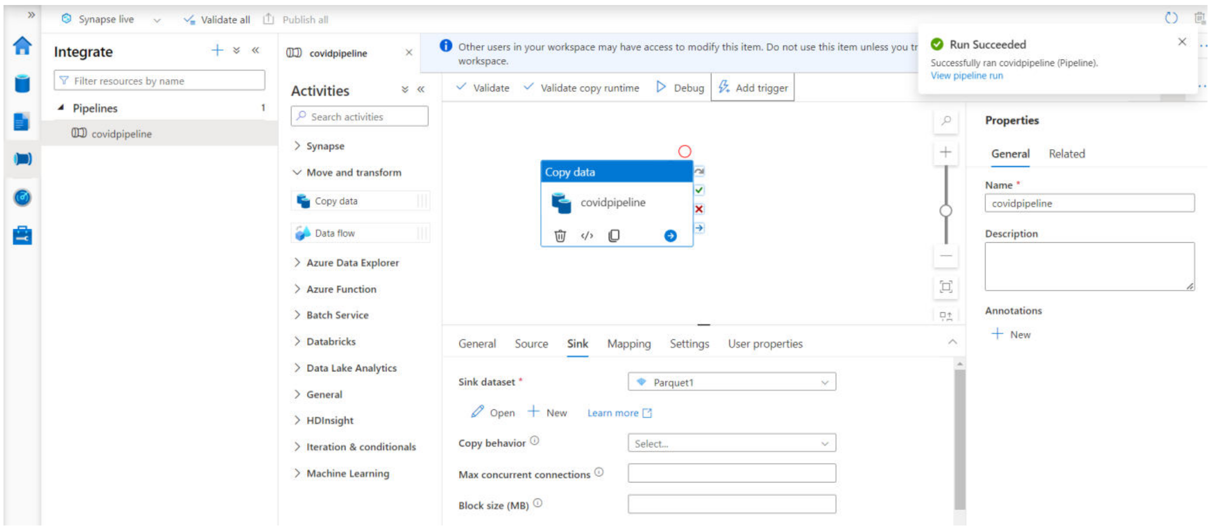
Give feedback
Tell us about your experience with deployment

Cost management
Get notified to stay within your budget and prevent unexpected charges on your bill.
Set up cost alerts >

Microsoft Defender for Cloud
Secure your apps and infrastructure
Go to Microsoft Defender for Cloud >

Free Microsoft tutorials
Start learning today >

Furthermore, run a pipeline successfully to change excel file format to parquet

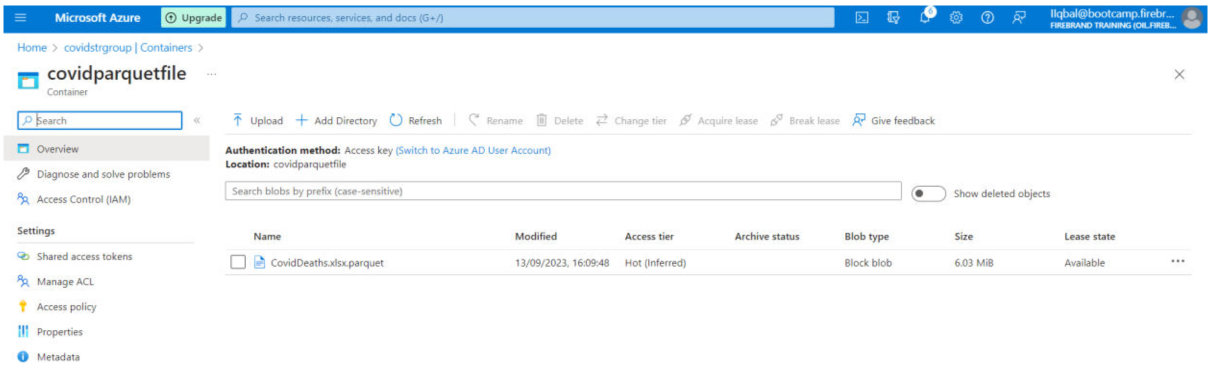

Run Succeeded
Successfully ran covidpipeline (Pipeline).
View pipeline run

Other users in your workspace may have access to modify this item. Do not use this item unless you tr workspace.

Sink dataset * Parquet1
Copy behavior Select...
Max concurrent connections
Block size (MB)

Properties
Name * covidpipeline
Description
Annotations

Parquet file is loaded in parquet folder as a result of pipeline trigger

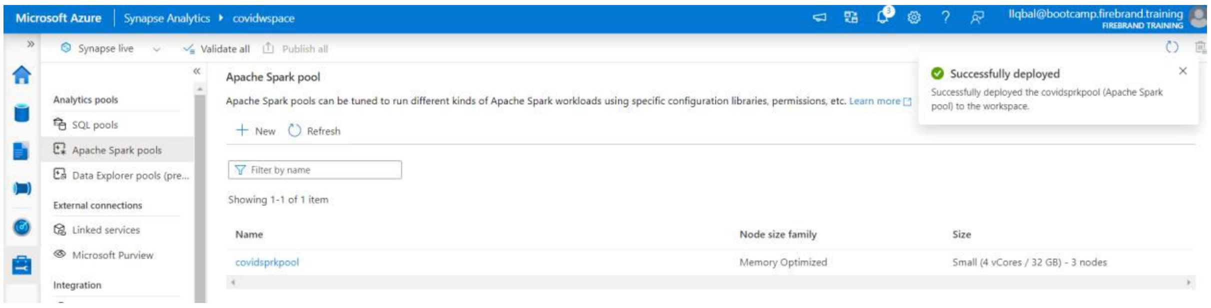

Home > covidstrgroup | Containers >

covidparquetfile
Container

Authentication method: Access key (Switch to Azure AD User Account)
Location: covidparquetfile

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
CovidDeaths.xlsx.parquet	13/09/2023, 16:09:48	Hot (Inferred)		Block blob	6.03 MiB	Available

Memory optimized Apache spark pool created in order to use the notebook



Apache Spark pool

Apache Spark pools can be tuned to run different kinds of Apache Spark workloads using specific configuration libraries, permissions, etc. Learn more

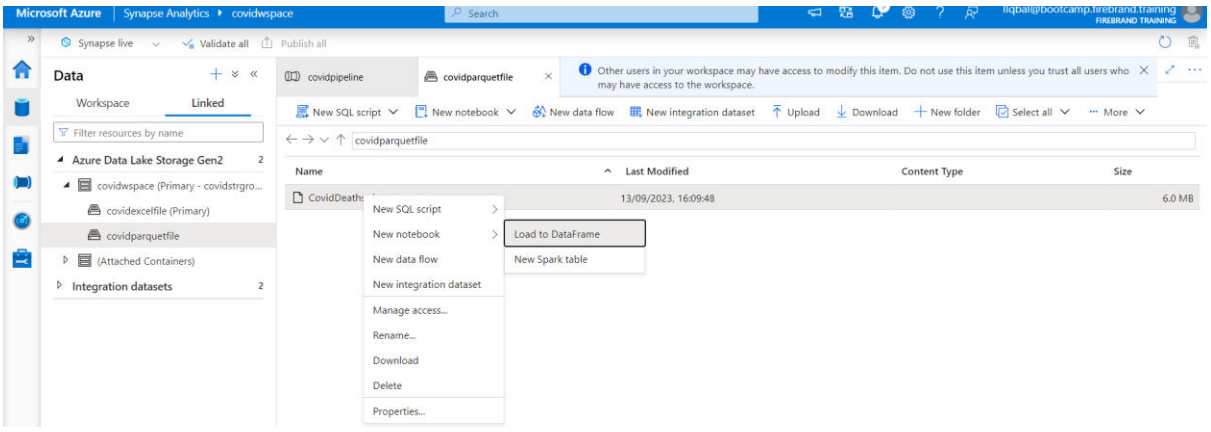
Successfully deployed

Successfully deployed the covidsprkpool (Apache Spark pool) to the workspace.

Showing 1-1 of 1 item

Name	Node size family	Size
covidsprkpool	Memory Optimized	Small (4 vCores / 32 GB) - 3 nodes

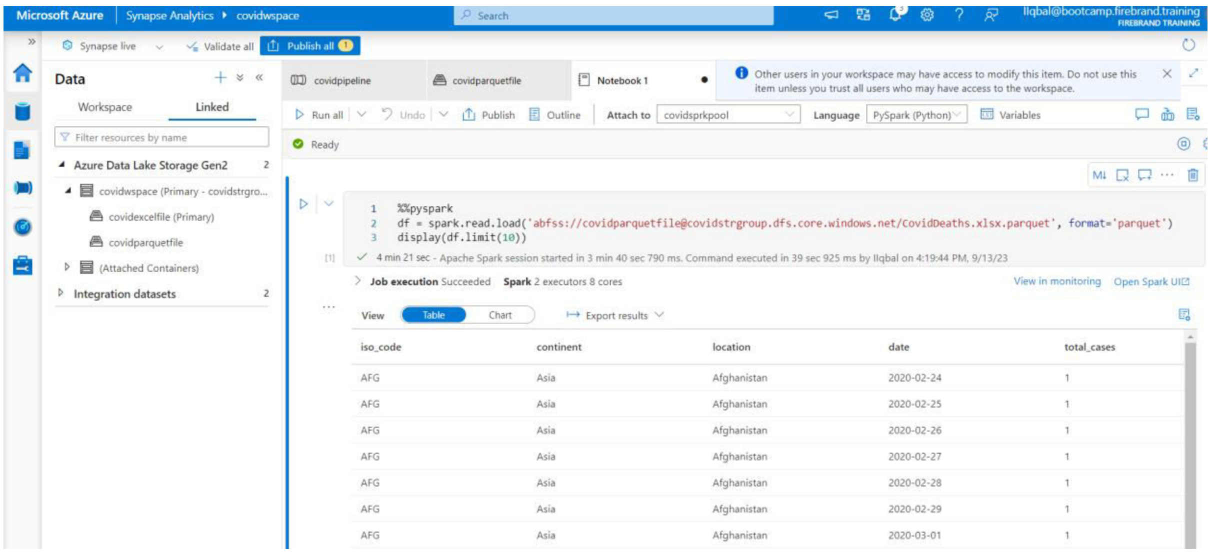
Parquet file loaded to dataframe in order to explore and use aggregate function



Other users in your workspace may have access to modify this item. Do not use this item unless you trust all users who may have access to the workspace.

Name	Last Modified	Content Type	Size
CovidDeaths	13/09/2023, 16:09:48		6.0 MB

All the data from CovidDeaths parquet file is loaded and displayed into the table using %pyspark


Microsoft Azure | Synapse Analytics ▸ covidwspace
Search
Iiqbal@bootcamp.firebrand.training
FIREBRAND TRAINING
Synapse live
Validate all
Publish all 1
Data
Workspace
Linked
Filter resources by name
Azure Data Lake Storage Gen2 2
covidwspace (Primary - covidstrgro...
covidexcelfile (Primary)
covidparquetfile
(Attached Containers)
Integration datasets 2
covidpipeline
covidparquetfile
Notebook 1
Other users in your workspace may have access to modify this item. Do not use this item unless you trust all users who may have access to the workspace.
Run all
Undo
Publish
Outline
Attach to covidsprkpool
Language PySpark (Python)
Variables
Ready

```
%%pyspark
df = spark.read.load('abfss://covidparquetfile@covidstrgroup.dfs.core.windows.net/CovidDeaths.xlsx.parquet', format='parquet')
display(df.limit(10))
```

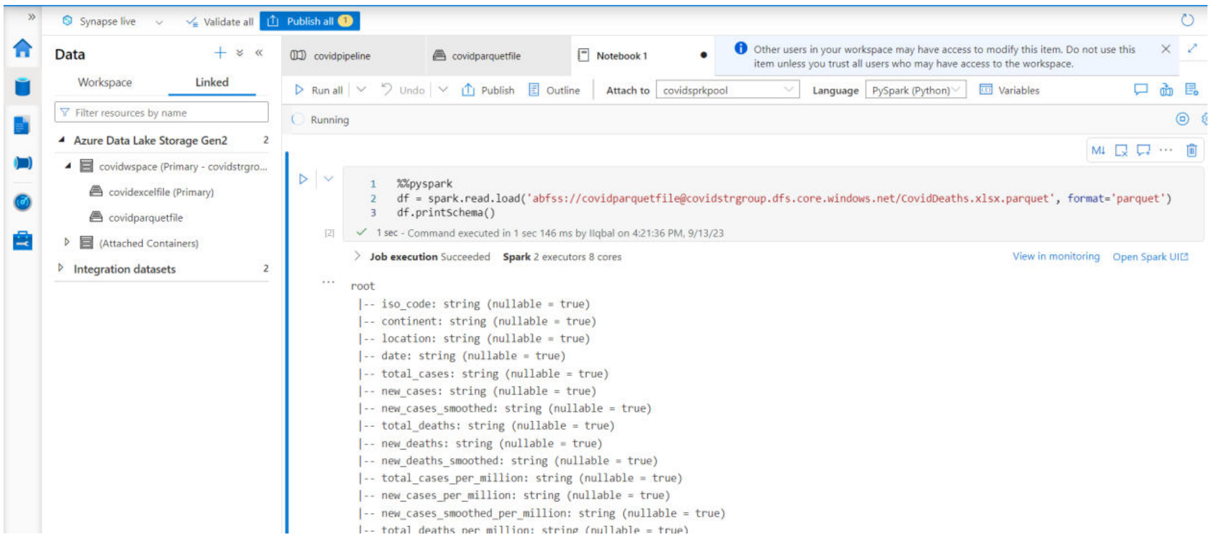
[1] ✓ 4 min 21 sec - Apache Spark session started in 3 min 40 sec 790 ms. Command executed in 39 sec 925 ms by IIqbal on 4:19:44 PM, 9/13/23

Job execution Succeeded Spark 2 executors 8 cores View in monitoring Open Spark UI

View Table Chart Export results

iso_code	continent	location	date	total_cases
AFG	Asia	Afghanistan	2020-02-24	1
AFG	Asia	Afghanistan	2020-02-25	1
AFG	Asia	Afghanistan	2020-02-26	1
AFG	Asia	Afghanistan	2020-02-27	1
AFG	Asia	Afghanistan	2020-02-28	1
AFG	Asia	Afghanistan	2020-02-29	1
AFG	Asia	Afghanistan	2020-03-01	1

Schema is displayed using pyspark for CovidDeath file



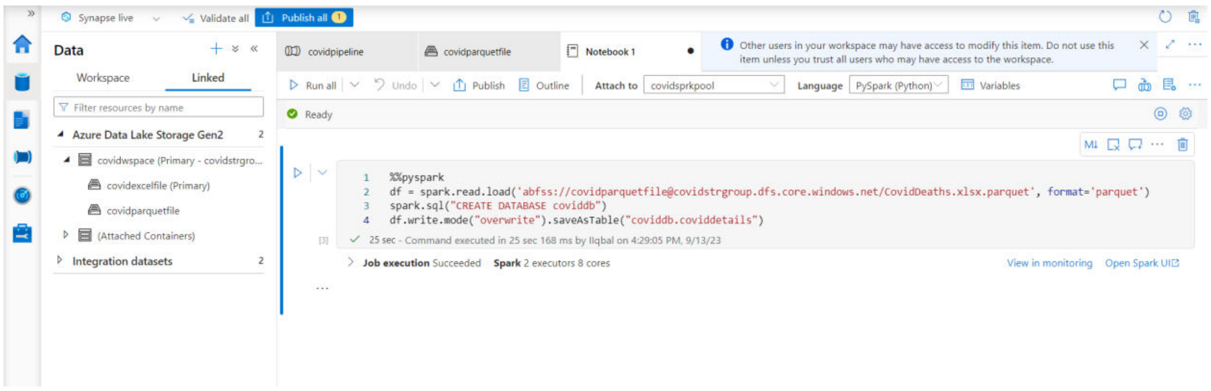
```
%%pyspark
df = spark.read.load('abfss://covidparquetfile@covidstrgroup.dfs.core.windows.net/CovidDeaths.xlsx.parquet', format='parquet')
df.printSchema()
```

[2] ✓ 1 sec - Command executed in 1 sec 146 ms by IIqbal on 4:21:36 PM, 9/13/23

Job execution Succeeded Spark 2 executors 8 cores View in monitoring Open Spark UI

```
root
 |-- iso_code: string (nullable = true)
 |-- continent: string (nullable = true)
 |-- location: string (nullable = true)
 |-- date: string (nullable = true)
 |-- total_cases: string (nullable = true)
 |-- new_cases: string (nullable = true)
 |-- new_cases_smoothed: string (nullable = true)
 |-- total_deaths: string (nullable = true)
 |-- new_deaths: string (nullable = true)
 |-- new_deaths_smoothed: string (nullable = true)
 |-- total_cases_per_million: string (nullable = true)
 |-- new_cases_per_million: string (nullable = true)
 |-- new_cases_smoothed_per_million: string (nullable = true)
 |-- total_deaths_per_million: string (nullable = true)
```

This data from dataframe is loaded into spark database by creating a database called “coviddb” and a table called coviddetails is created and data is loaded into the table



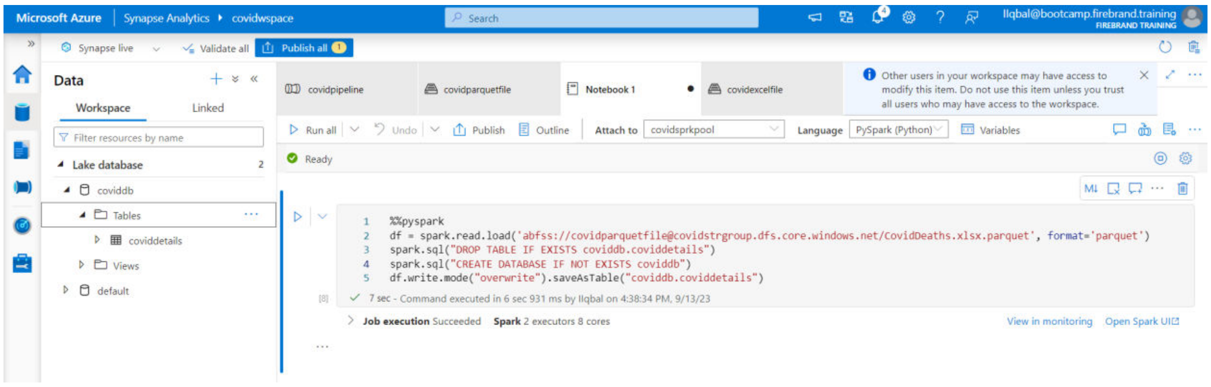
Other users in your workspace may have access to modify this item. Do not use this item unless you trust all users who may have access to the workspace.

```
%%pyspark
df = spark.read.load('abfss://covidparquetfile@covidstrgroup.dfs.core.windows.net/CovidDeaths.xlsx.parquet', format='parquet')
spark.sql("CREATE DATABASE coviddb")
df.write.mode("overwrite").saveAsTable("coviddb.coviddetails")
```

[3] 25 sec - Command executed in 25 sec 168 ms by Ilqbal on 4:29:05 PM, 9/13/23

Job execution Succeeded Spark 2 executors 8 cores View in monitoring Open Spark UI

Table coviddetails created successfully in lakedatabase



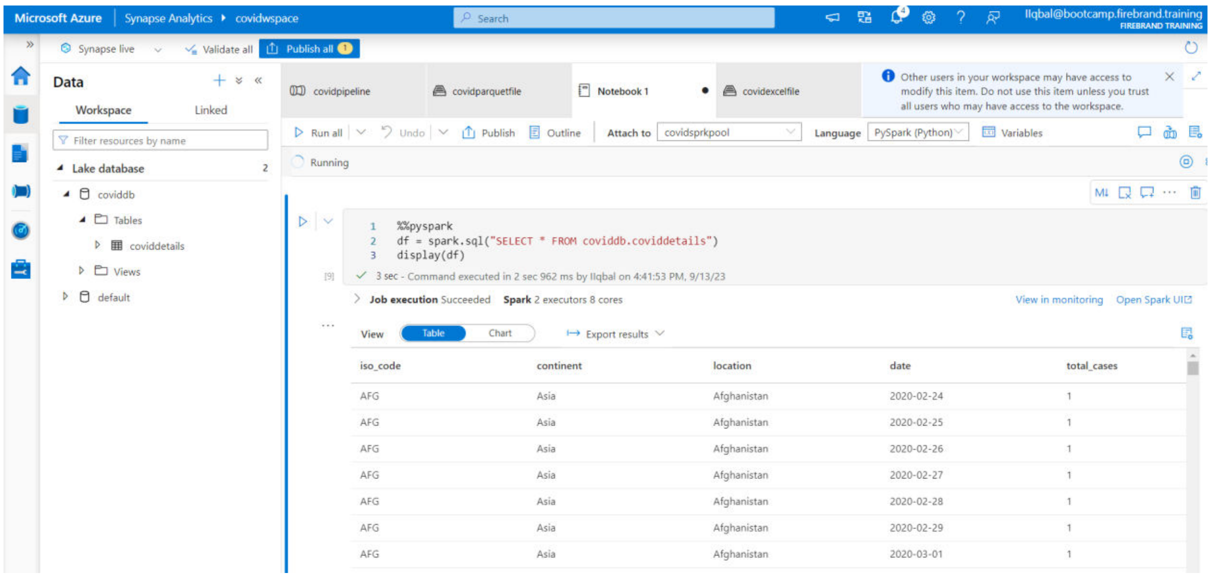
Other users in your workspace may have access to modify this item. Do not use this item unless you trust all users who may have access to the workspace.

```
%%pyspark
df = spark.read.load('abfss://covidparquetfile@covidstrgroup.dfs.core.windows.net/CovidDeaths.xlsx.parquet', format='parquet')
spark.sql("DROP TABLE IF EXISTS coviddb.coviddetails")
spark.sql("CREATE DATABASE IF NOT EXISTS coviddb")
df.write.mode("overwrite").saveAsTable("coviddb.coviddetails")
```

[8] 7 sec - Command executed in 6 sec 931 ms by Ilqbal on 4:38:34 PM, 9/13/23

Job execution Succeeded Spark 2 executors 8 cores View in monitoring Open Spark UI

Coviddetails table in the coviddb database is loaded using in notebook



Other users in your workspace may have access to modify this item. Do not use this item unless you trust all users who may have access to the workspace.

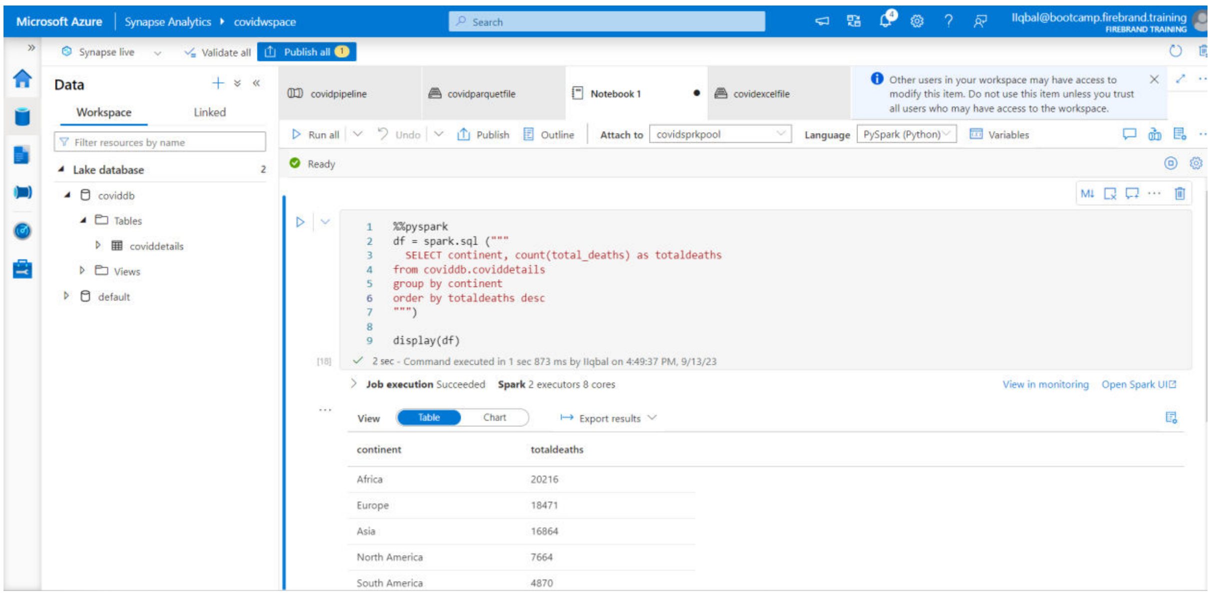
```
%%pyspark
df = spark.sql("SELECT * FROM coviddb.coviddetails")
display(df)
```

[9] ✓ 3 sec - Command executed in 2 sec 962 ms by IIqbal on 4:41:53 PM, 9/13/23

Job execution Succeeded Spark 2 executors 8 cores

iso_code	continent	location	date	total_cases
AFG	Asia	Afghanistan	2020-02-24	1
AFG	Asia	Afghanistan	2020-02-25	1
AFG	Asia	Afghanistan	2020-02-26	1
AFG	Asia	Afghanistan	2020-02-27	1
AFG	Asia	Afghanistan	2020-02-28	1
AFG	Asia	Afghanistan	2020-02-29	1
AFG	Asia	Afghanistan	2020-03-01	1

Aggregation is applied to display which continents have most deaths and Africa is the top continent which was displayed using group by and order by function.



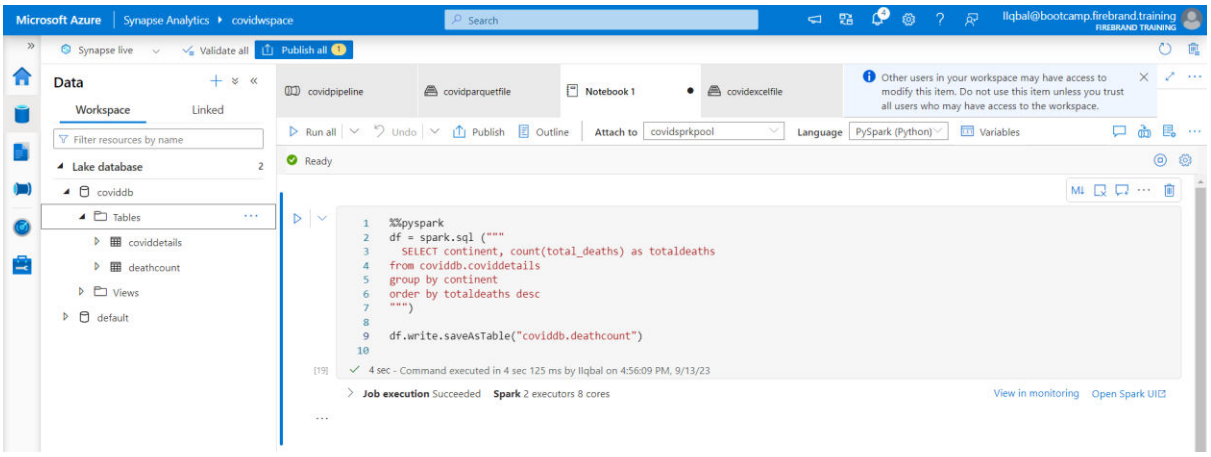
```
%%pyspark
df = spark.sql ("""
  SELECT continent, count(total_deaths) as totaldeaths
from coviddb.coviddetails
group by continent
order by totaldeaths desc
""")

display(df)
```

2 sec - Command executed in 1 sec 873 ms by IIqbal on 4:49:37 PM, 9/13/23

Job execution Succeeded Spark 2 executors 8 cores

continent	totaldeaths
Africa	20216
Europe	18471
Asia	16864
North America	7664
South America	4870

Aggregated table deathcount is created and saved in the database as shown



```
%%pyspark
df = spark.sql ("""
  SELECT continent, count(total_deaths) as totaldeaths
from coviddb.coviddetails
group by continent
order by totaldeaths desc
""")

df.write.saveAsTable("coviddb.deathcount")
```

4 sec - Command executed in 4 sec 125 ms by IIqbal on 4:56:09 PM, 9/13/23

Job execution Succeeded Spark 2 executors 8 cores