Load and upload the files to Azure File Share using Ingestion Framework

Azure Data Lake Storage Gen2 is a highly scalable and cost-effective data lake solution for big data analytics. Data Lake Storage Gen2 extends Azure Blob Storage capabilities and is optimized for analytics workloads.

Azure Data Lake Storage Gen2 provides 4 styles of data storage which are

1.Containers

2.File shares

3.Tables and

4.Queues

Every component has their unique feature to support the requirements,

For e.g. Containers are for Big data analytics, databricks etc and

File shares extend your servers to Azure with Sync for on-premises performance and capability

*Working with Azure files using Ingestion Framework:*

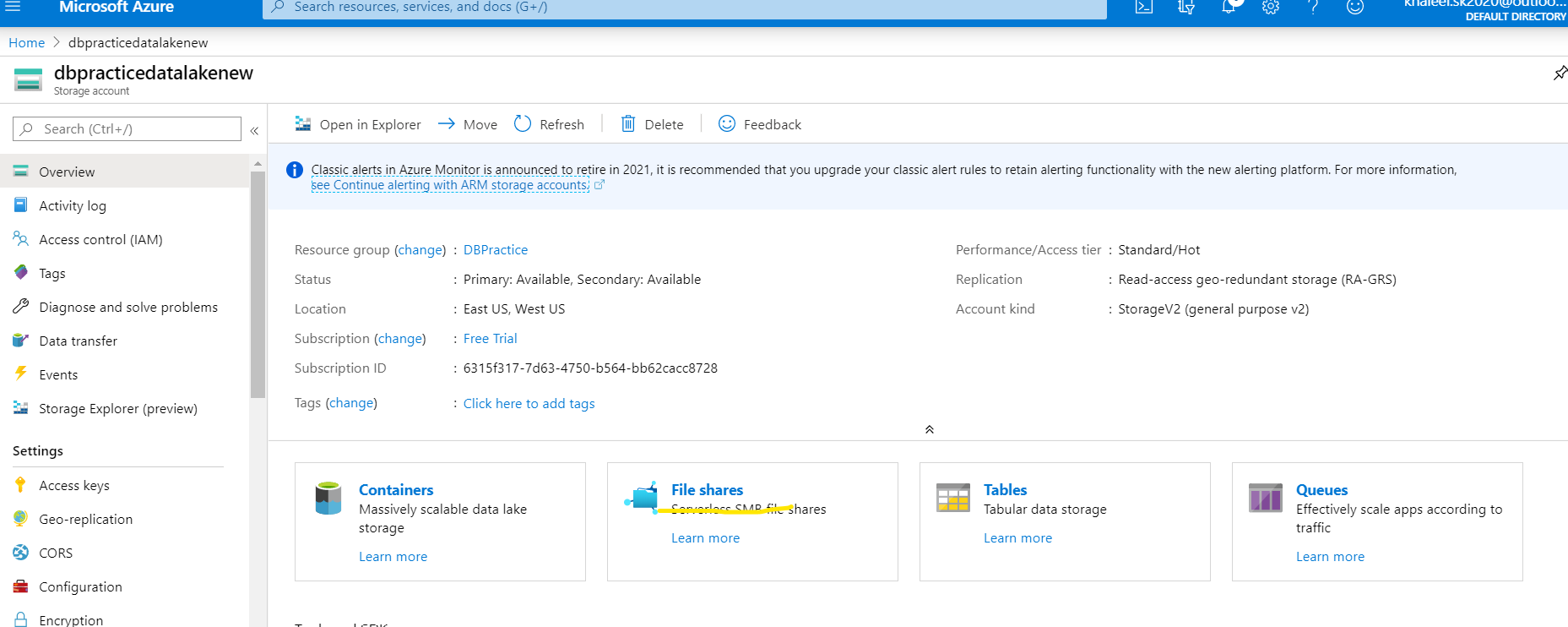
Ingestion Framework use Azure File shares to load and upload transformed files to Azure

For this we mount Azure File shares as a network drive in on-premise system

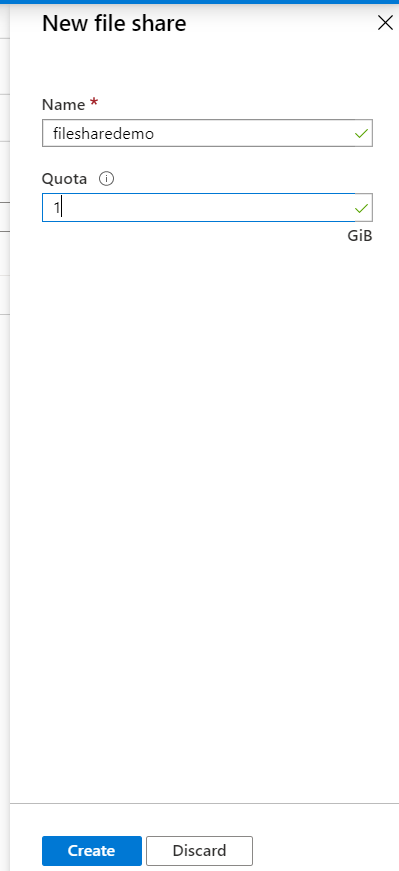
* Crate a new Azure Data lake account in azure portal

<https://docs.microsoft.com/en-in/azure/storage/common/storage-account-create?tabs=azure-portal>

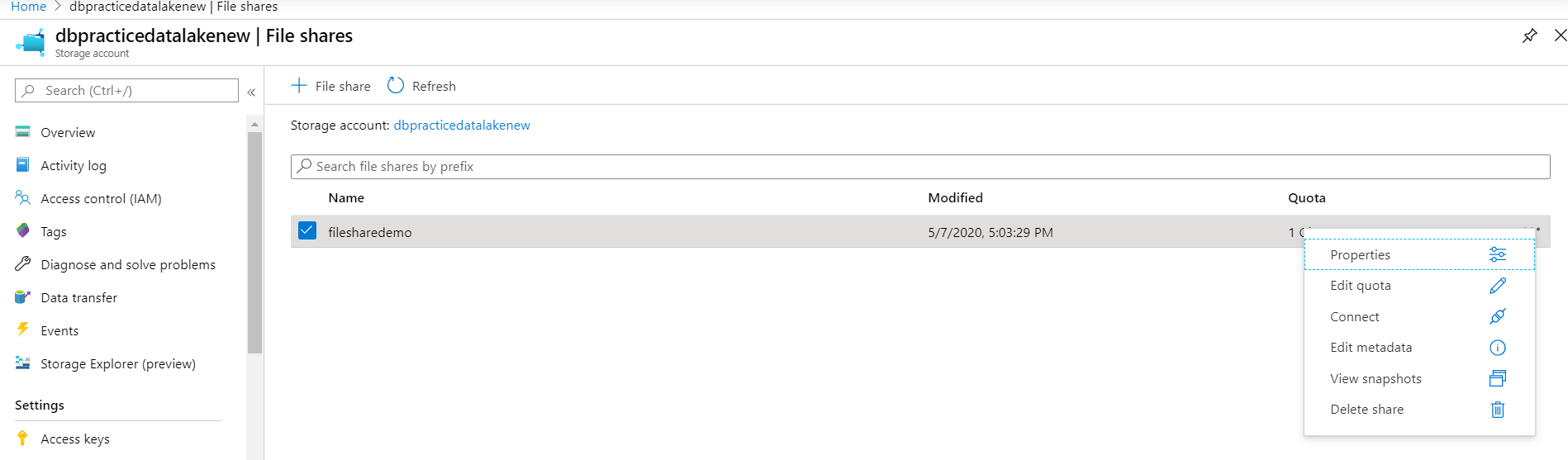
* Go to newly created ADL account click **File shares**

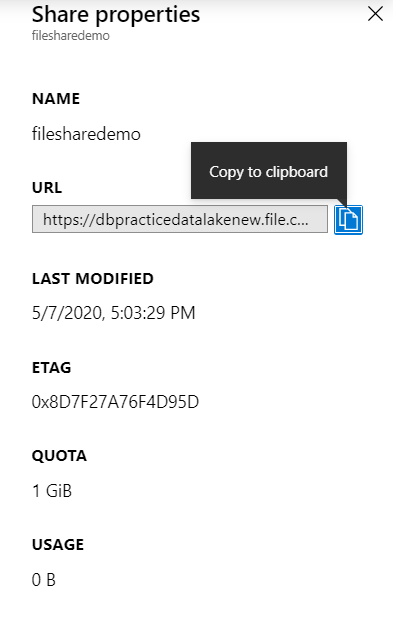


* Select + File share on top to create a new File share
* Provide the name and Quota in GB then click Create

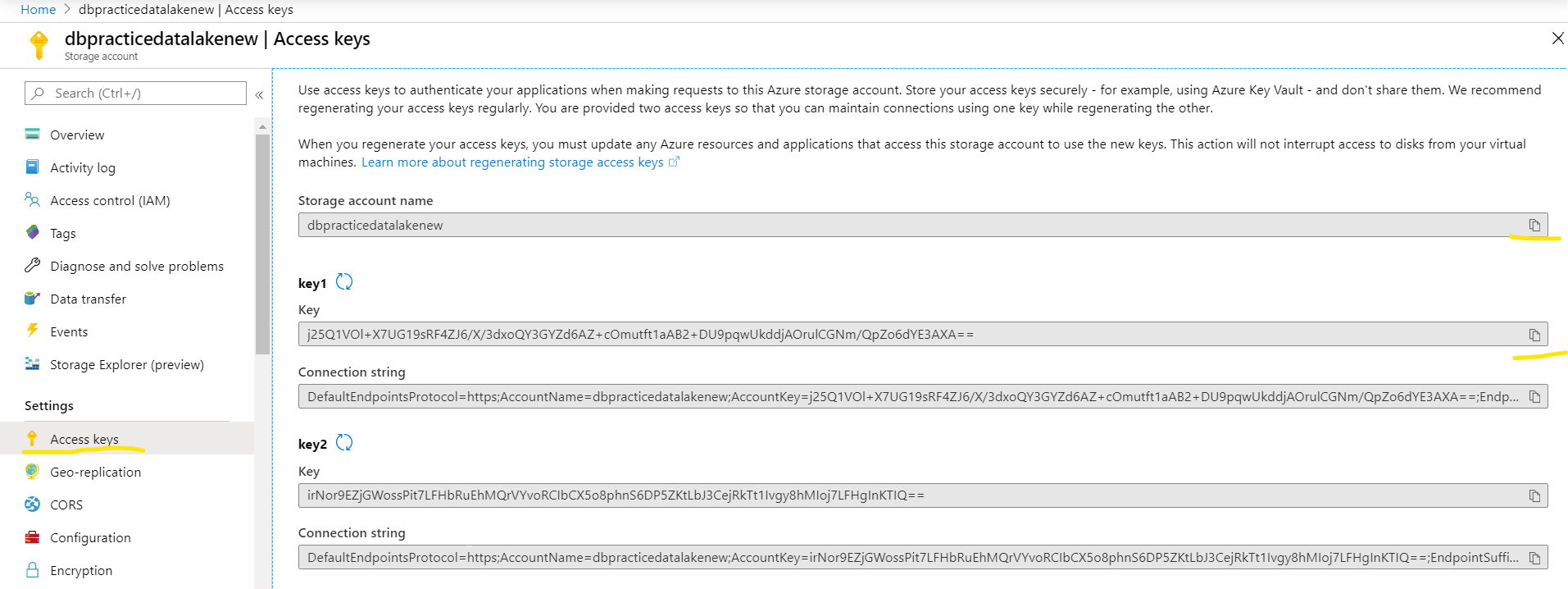


* Click on ellipse (…) of created file share, select properties and copy the URL



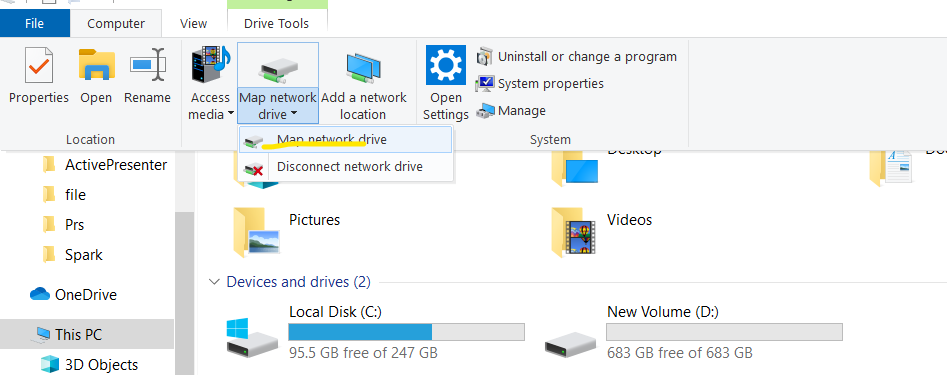


* From left middle corner Settings -> Access keys copy storage account name and key1.



*Now create a Mount point in windows machine*

* File explorer -> This PC -> Top corner Computer->Map network drive->Map network drive



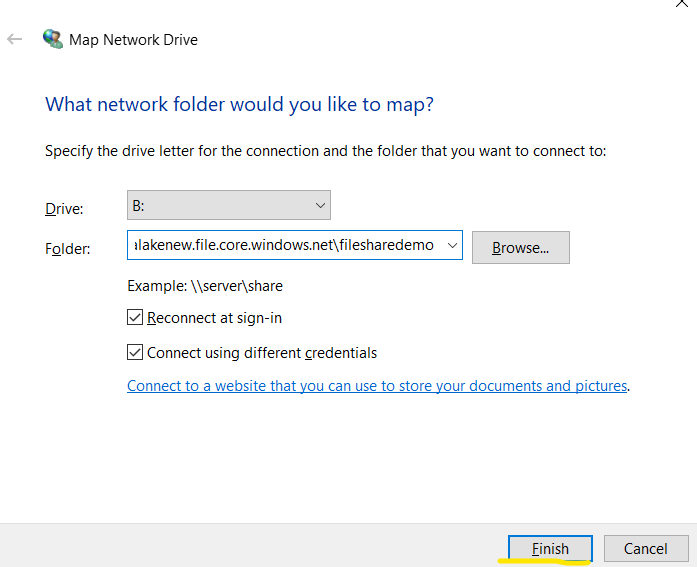
* Select Drive and in Folder paste the copied share folder URL in accepted format

E.g. Drive -> A

Folder -> https://dbpracticedatalakenew.file.core.windows.net/filesharedemo

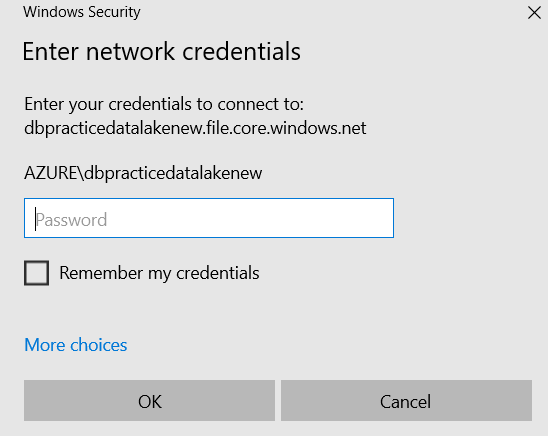
E.g. \\dbpracticedatalakenew.file.core.windows.net\filesharedemo

* Select both the check boxes and Finish



* In the Windows security window (give ‘AZURE\storage account name’ as user id if prompted) and Key value as Password.

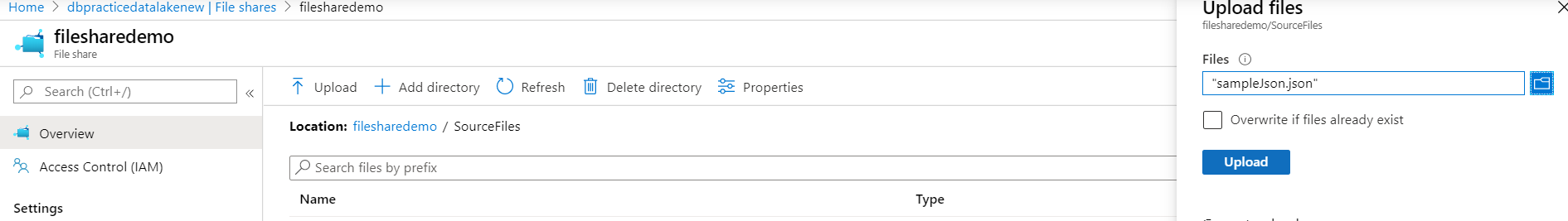
*Note*: If password is asking iteratively use Key2.



*Once Network drive (Mount point) created the uploaded the files are in sync with Azure file share and this folder vice-versa.*

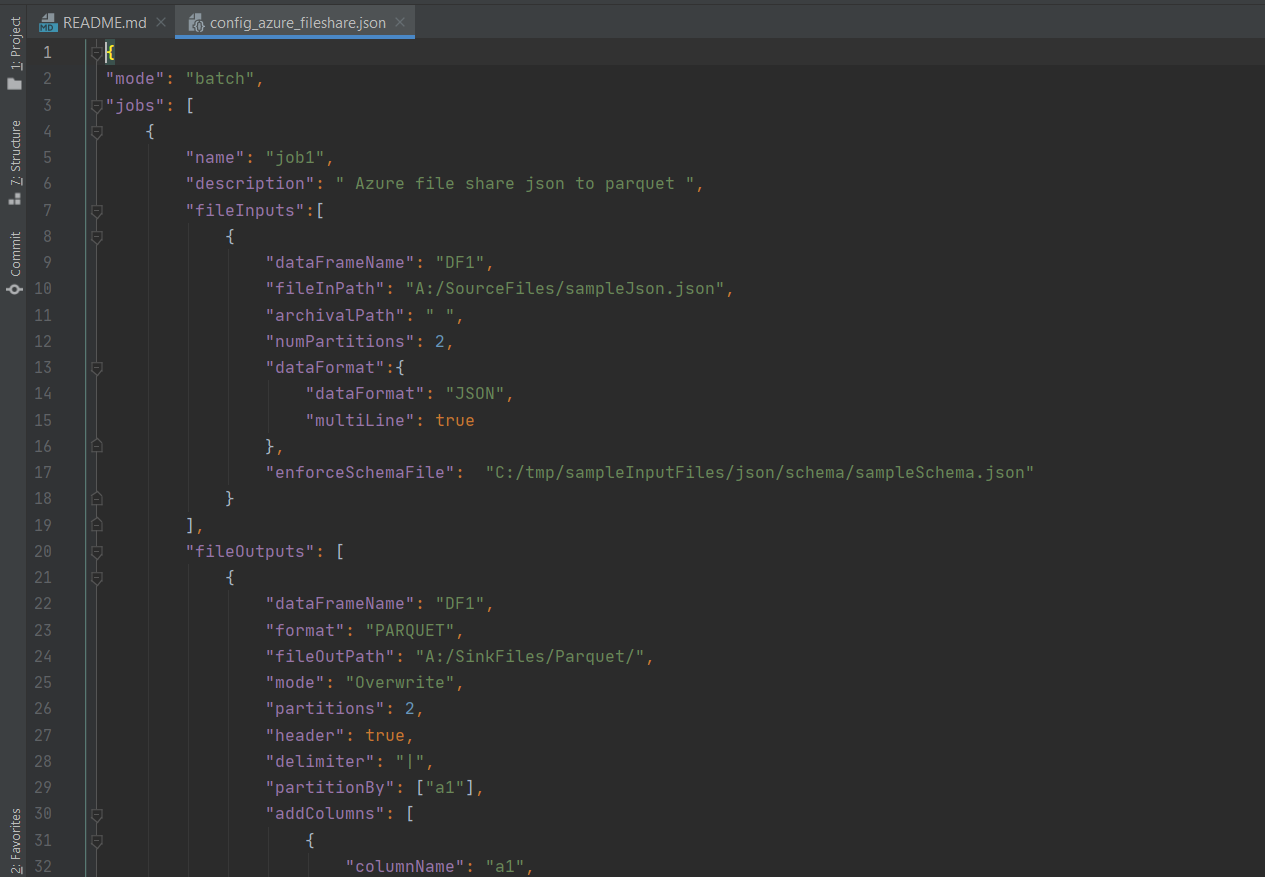
* Create SourceFiles and SinkFiles folders (Automatically sync with Azure File share) in filesharedemo mount drive for load and upload the files using Ingestion Framework.
* Now upload the source file either from Azure Fileshare or Network mount point.

Uploading the source file sampleJson.json in SourceFiles directory



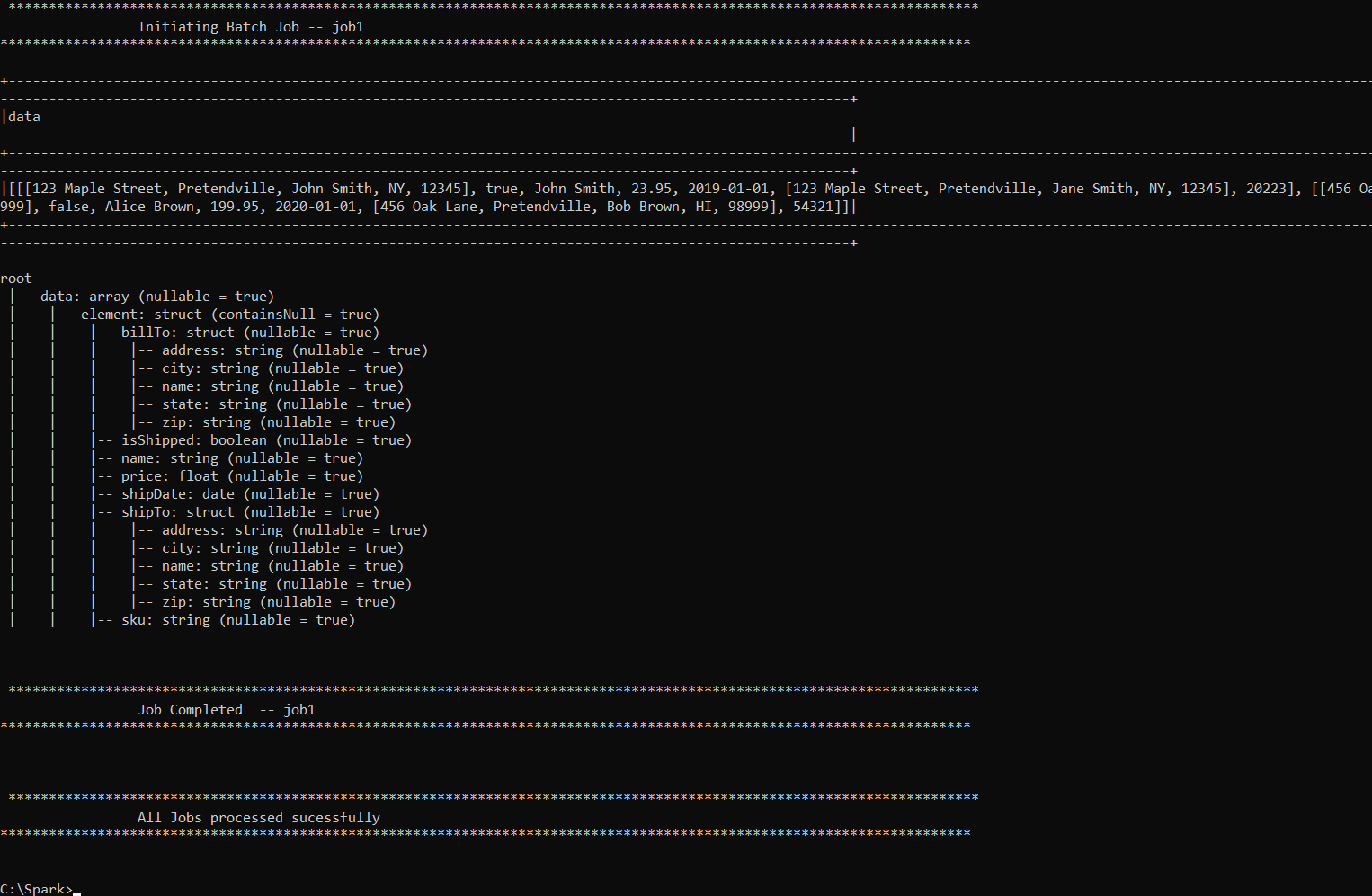
Now using Ingestion Framework load this sampleJson.json file and convert to .parquet format and upload back in Azure Fileshare SinkFiles/Parquet location

* For this create one more directory ‘Parquet’ in ‘SinkFiles’ directory
* Create a new configuration file using above Source and Sink

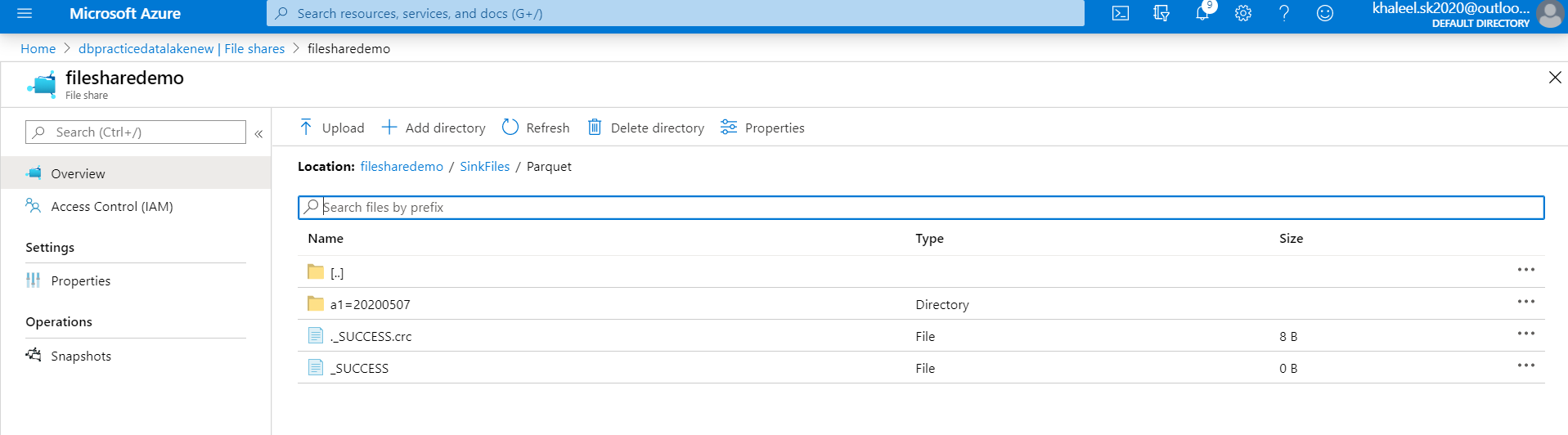


* Run the following command to execute the task

C:\Spark>spark-submit C:\Users\Nisum\Documents\GitHub\ingestionframework\src\main.py C:\Users\Nisum\Documents\GitHub\ingestionframework\src\resources\config\_azure\_fileshare.json



Validation:



Like this we can load the Azure files on fly, apply transformations and upload back to Azure Fileshare using Ingestion Framework.

*Observation*:

Azure Data Lake Containers files also can be used in spark, but it will disturb Ingestion Framework generic code implementation.

Reference:

<https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-directory-file-acl-python>