Microsoft Azure Capstone Project

By Bhavana Jaiswal

INTRODUCTION

- AZURE SYNAPSE ANALYTICS IS A CLOUD-BASED ANALYTICS SERVICE OFFERED BY MICROSOFT. IT PROVIDES A WORKSPACE FOR DATA PROFESSIONALS TO INGEST, PREPARE, MANAGE, AND SERVE DATA FOR IMMEDIATE BUSINESS INTELLIGENCE AND MACHINE LEARNING NEEDS. WITH SYNAPSE, YOU CAN EXPLORE, ANALYZE, AND CREATE REPORTS ON YOUR DATA USING A UNIFIED EXPERIENCE, LEVERAGING A POWERFUL COMBINATION OF BIG DATA AND DATA WAREHOUSING TECHNOLOGIES. IT ALLOWS YOU TO CONNECT TO VARIOUS DATA SOURCES, INCLUDING ON-PREMISES DATA STORES, CLOUD DATA STORES, AND OTHER EXTERNAL SOURCES, AND ANALYZE THEM WITH A WIDE RANGE OF TOOLS SUCH AS APACHE SPARK, SQL, AND MACHINE LEARNING MODELS. SYNAPSE ANALYTICS IS A POWERFUL PLATFORM THAT CAN HELP YOU SCALE UP OUR DATA ANALYTICS CAPABILITIES, IMPROVE DECISION-MAKING, AND GAIN VALUABLE INSIGHTS INTO OUR BUSINESS.

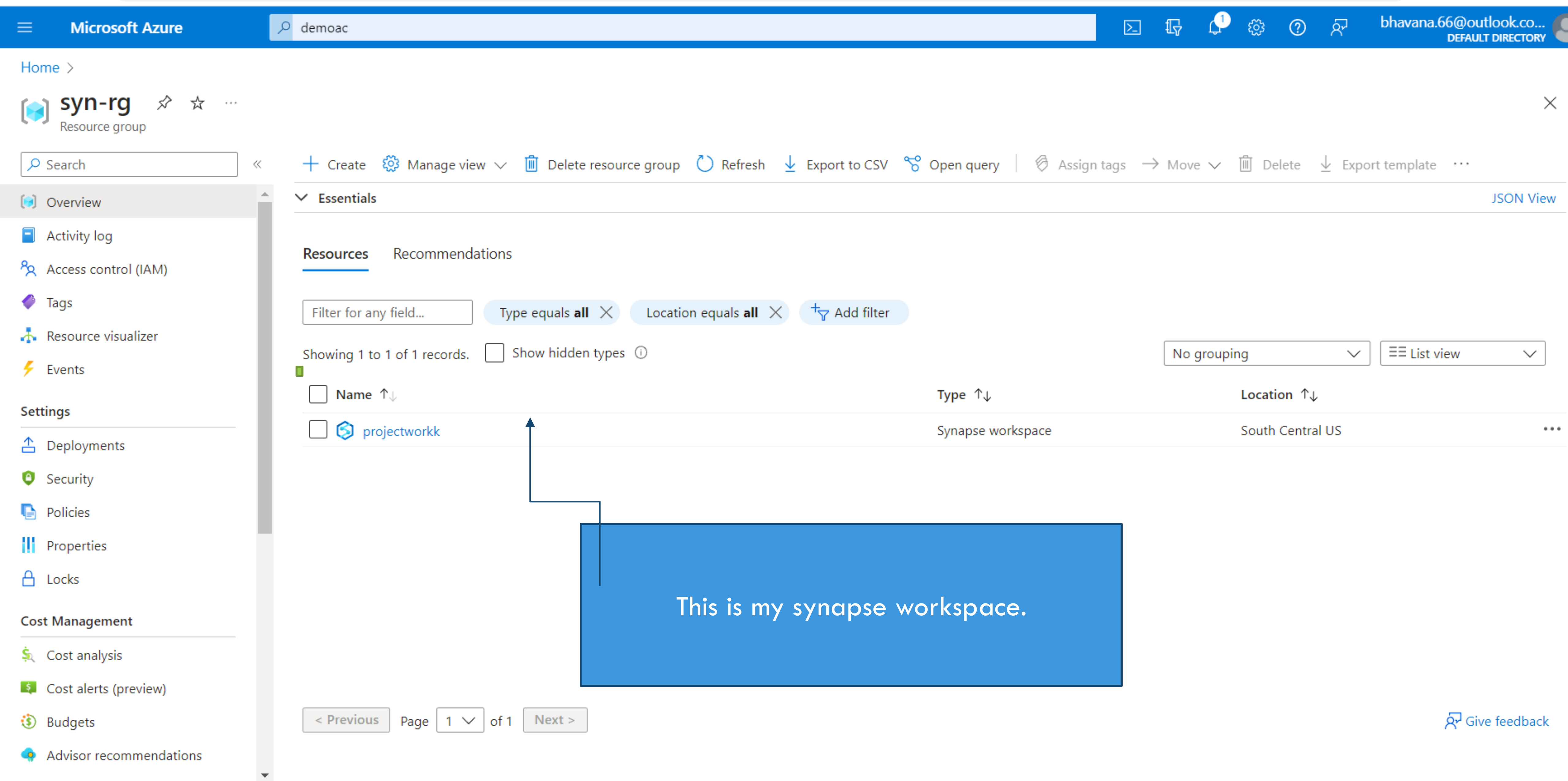
PROBLEM FIRST

- **Problem statement 1:** The task is to explore data analytics workspace by using Azure Synapse Analytics. You will create ADLS Gen 2 accounts and define the pipelines in Azure Synapse Analytics to transfer data from various data sources into the workspace for analysis. The data will be ingested in Azure Synapse with Built-in copy task option, and you will query the uploaded data.

INTEGRATED WITH AZURE SYNAPSE ANALYTICS WORKSPACE

AZURE SYNAPSE ANALYTICS: A CLOUD-BASED ANALYTICS SERVICE THAT BRINGS TOGETHER BIG DATA AND DATA WAREHOUSING. IT USES APACHE SPARK AND PROVIDES INTEGRATION WITH OTHER AZURE SERVICES SUCH AS AZURE DATA FACTORY, AZURE DATA LAKE STORAGE, AND POWER BI.

PRE-REQUISITES: MICROSOFT ACCOUNT, RESOURCE GROUP, SYNAPSE WORKSPACE



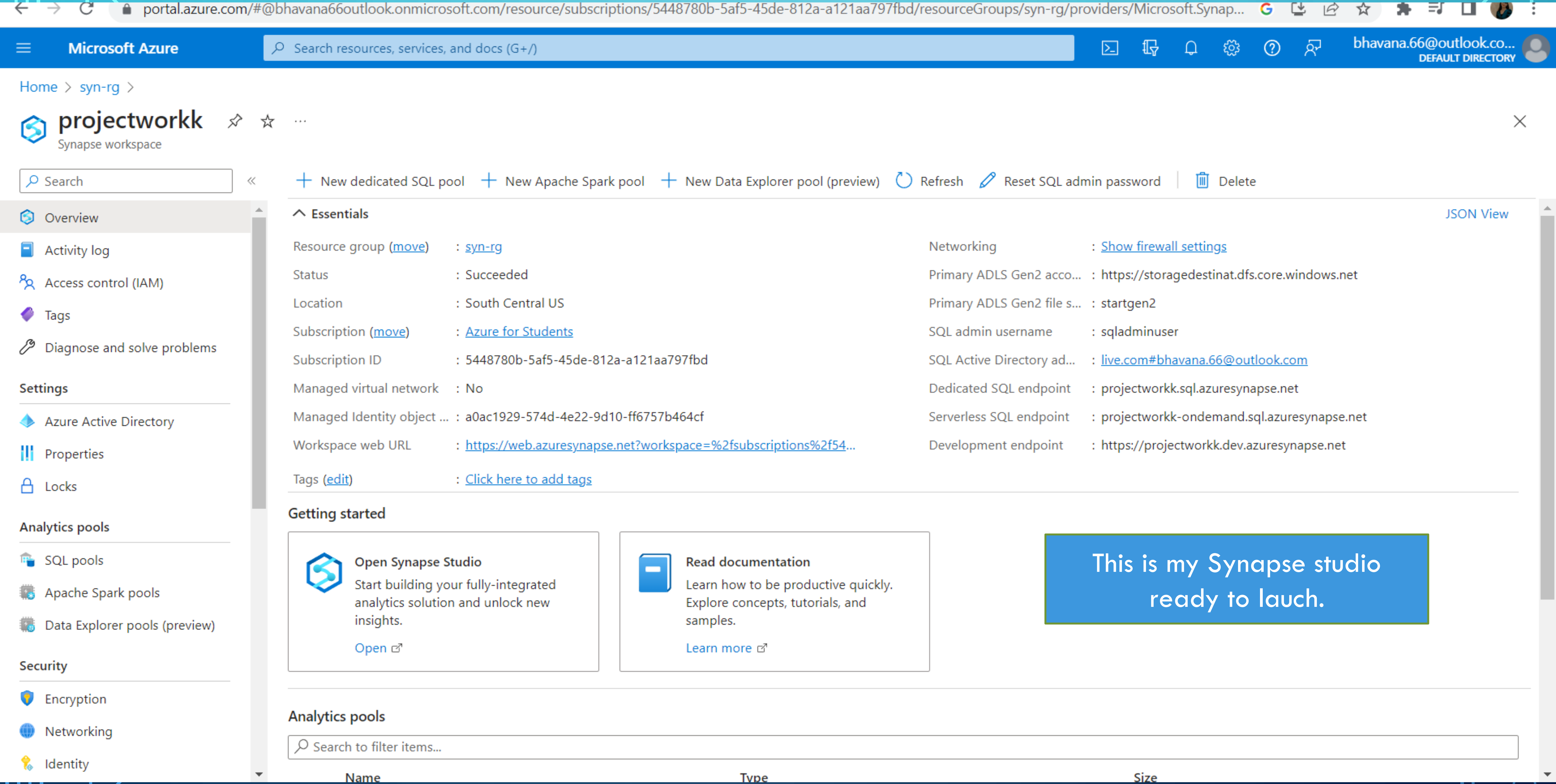
syn-rg

Resource group

Showing 1 to 1 of 1 records. □ Show hidden types

□	Name	Type	Location
□	projectworkk	Synapse workspace	South Central US

This is my synapse workspace.



Resource group (move)	: syn-rg
Status	: Succeeded
Location	: South Central US
Subscription (move)	: Azure for Students
Subscription ID	: 5448780b-5af5-45de-812a-a121aa797fbd
Managed virtual network	: No
Managed Identity object ...	: a0ac1929-574d-4e22-9d10-ff6757b464cf
Workspace web URL	: https://web.azuresynapse.net?workspace=%2fsubscriptions%2f54...
Tags (edit)	: Click here to add tags

Networking	: Show firewall settings
Primary ADLS Gen2 acco...	: https://storagedestinat.dfs.core.windows.net
Primary ADLS Gen2 file s...	: startgen2
SQL admin username	: sqladminuser
SQL Active Directory ad...	: live.com#bhavana.66@outlook.com
Dedicated SQL endpoint	: projectworkk.sql.azuresynapse.net
Serverless SQL endpoint	: projectworkk-ondemand.sql.azuresynapse.net
Development endpoint	: https://projectworkk.dev.azuresynapse.net

Getting started

Open Synapse Studio

Start building your fully-integrated analytics solution and unlock new insights.

Open

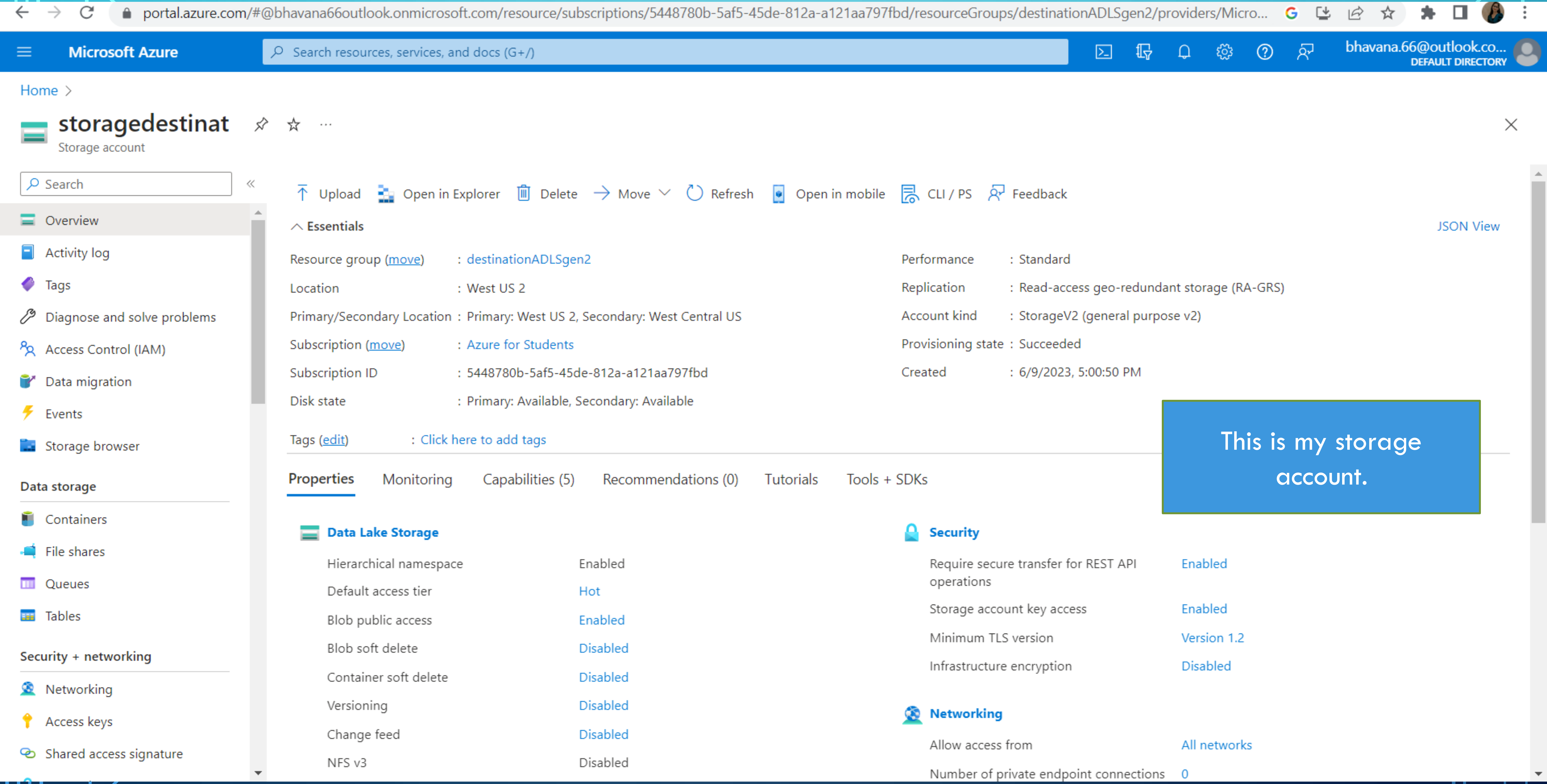
Read documentation

Learn how to be productive quickly. Explore concepts, tutorials, and samples.

Learn more

This is my Synapse studio ready to lauch.

Analytics pools



storagedestinat

Storage account

Essentials

Resource group (move)	: destinationADLSgen2
Location	: West US 2
Primary/Secondary Location	: Primary: West US 2, Secondary: West Central US
Subscription (move)	: Azure for Students
Subscription ID	: 5448780b-5af5-45de-812a-a121aa797fbd
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	: 6/9/2023, 5:00:50 PM
Tags (edit)	: Click here to add tags

Data Lake Storage

Hierarchical namespace	Enabled
Default access tier	Hot
Blob public access	Enabled
Blob soft delete	Disabled
Container soft delete	Disabled
Versioning	Disabled
Change feed	Disabled
NFS v3	Disabled

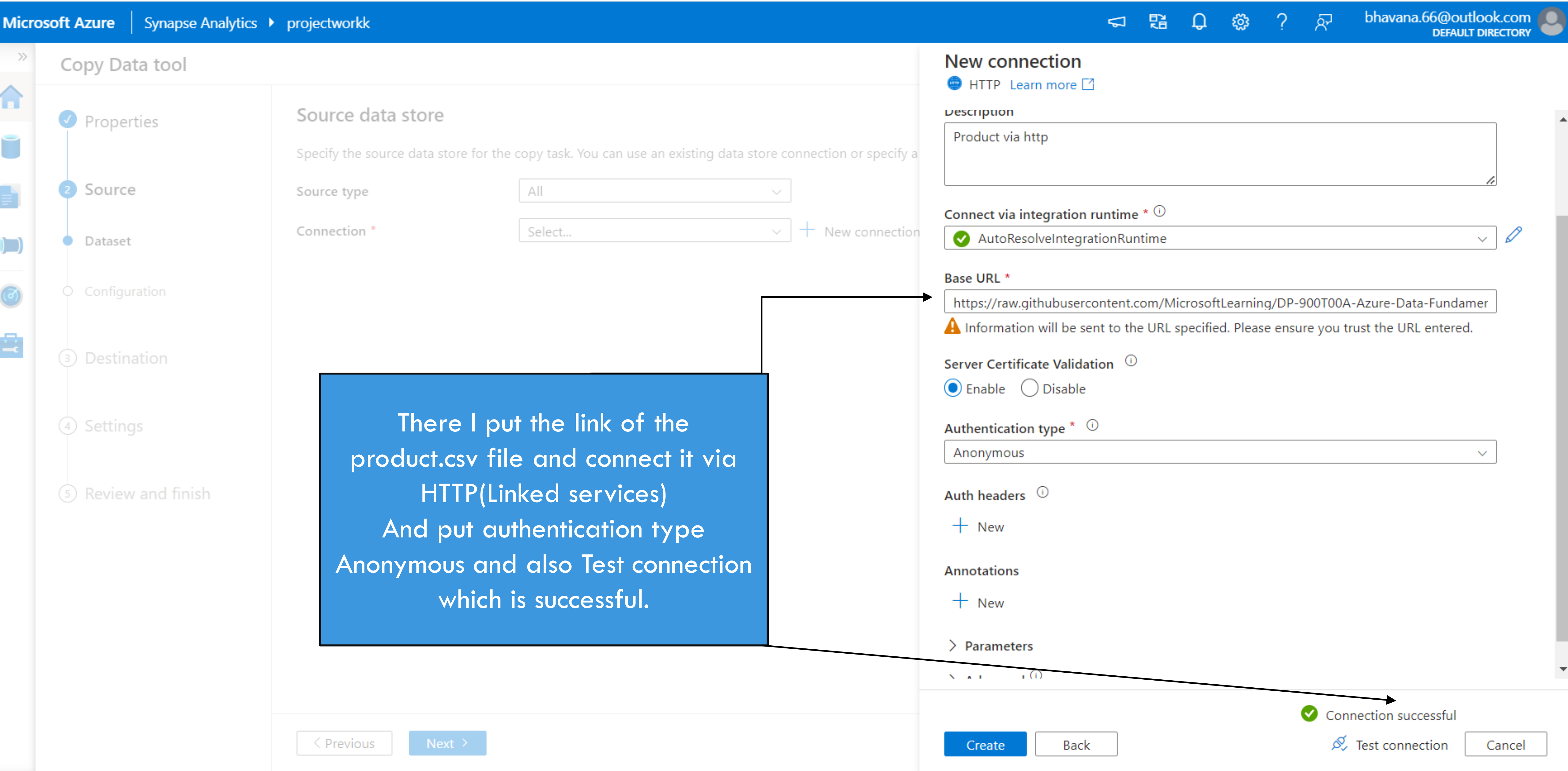
Security

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

This is my storage account.



Source data store

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

Source type: All

Connection *: Select...

+ New connection

New connection

HTTP Learn more

Description: Product via http

Connect via integration runtime *: AutoResolveIntegrationRuntime

Base URL *: https://raw.githubusercontent.com/MicrosoftLearning/DP-900T00A-Azure-Data-Fundamer

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

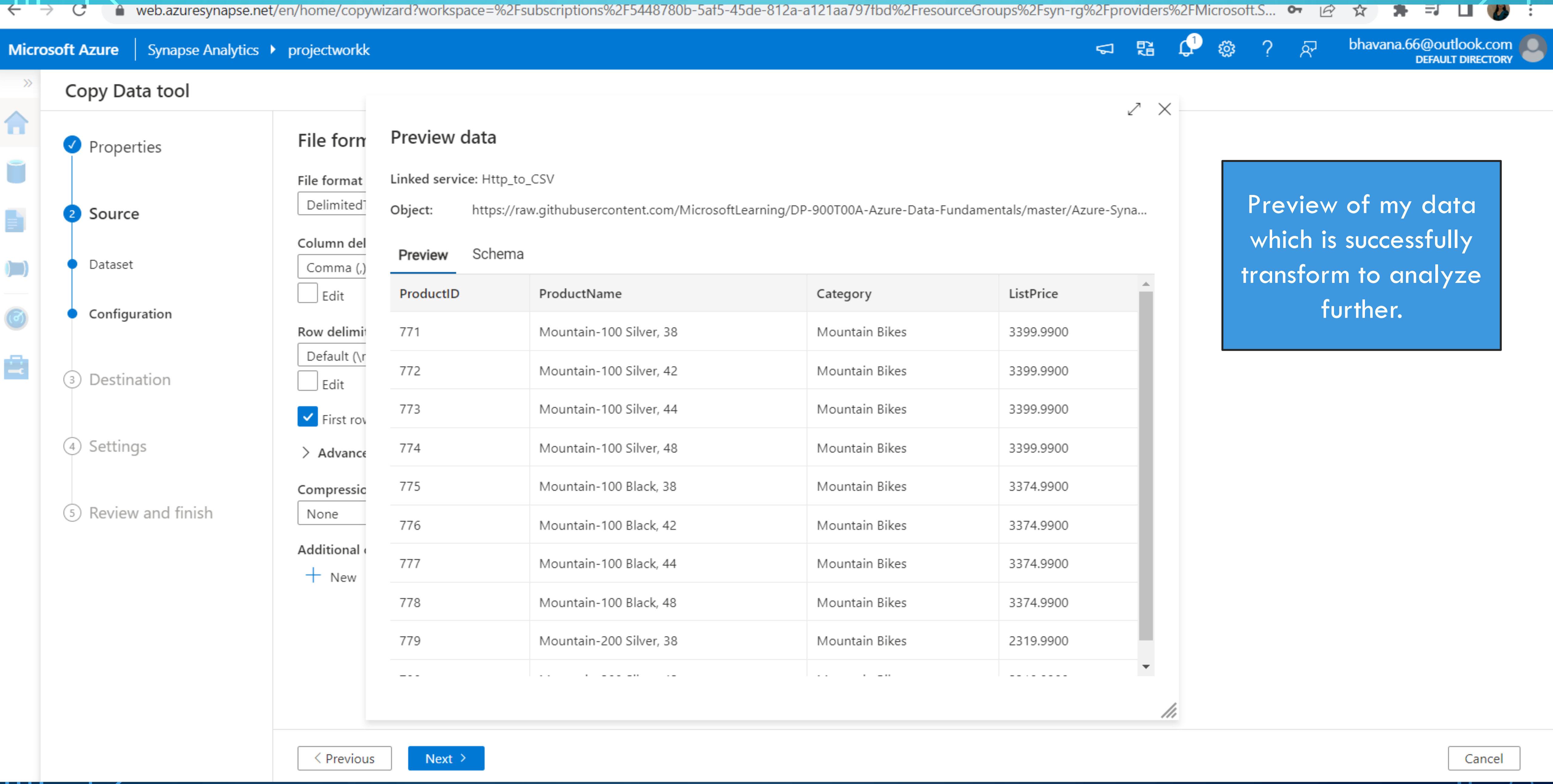
Annotations

+ New

> Parameters

Connection successful

There I put the link of the product.csv file and connect it via HTTP(Linked services)
And put authentication type Anonymous and also Test connection which is successful.


web.azuresynapse.net/en/home/copywizard?workspace=%2Fsubscriptions%2F5448780b-5af5-45de-812a-a121aa797fbd%2FresourceGroups%2Fsyn-rg%2Fproviders%2FMicrosoft.S...
Microsoft Azure
Synapse Analytics
projectworkk
bhavana.66@outlook.com
DEFAULT DIRECTORY

Copy Data tool

1. Properties
2. Source
   - Dataset
   - Configuration
3. Destination
4. Settings
5. Review and finish

File form

File format
Delimited

Column del
Comma (,)
□ Edit

Row delimi
Default (\r
□ Edit

☑ First row

> Advance

Compressio
None

Additional
+ New

Preview data

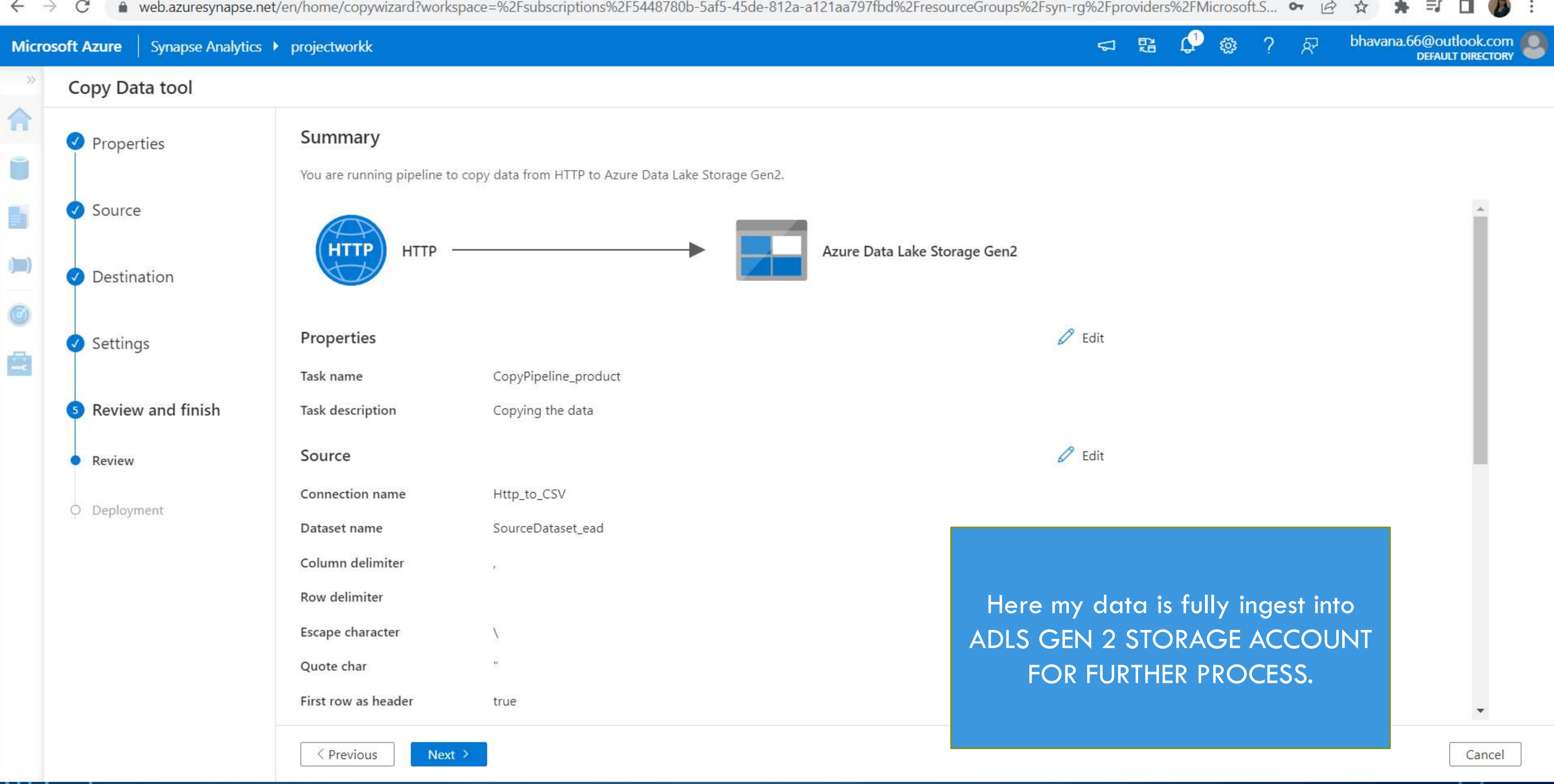
Linked service: Http_to_CSV

Object: https://raw.githubusercontent.com/MicrosoftLearning/DP-900T00A-Azure-Data-Fundamentals/master/Azure-Syna...

Preview | Schema

ProductID	ProductName	Category	ListPrice
771	Mountain-100 Silver, 38	Mountain Bikes	3399.9900
772	Mountain-100 Silver, 42	Mountain Bikes	3399.9900
773	Mountain-100 Silver, 44	Mountain Bikes	3399.9900
774	Mountain-100 Silver, 48	Mountain Bikes	3399.9900
775	Mountain-100 Black, 38	Mountain Bikes	3374.9900
776	Mountain-100 Black, 42	Mountain Bikes	3374.9900
777	Mountain-100 Black, 44	Mountain Bikes	3374.9900
778	Mountain-100 Black, 48	Mountain Bikes	3374.9900
779	Mountain-200 Silver, 38	Mountain Bikes	2319.9900

Preview of my data which is successfully transform to analyze further.



Summary

You are running pipeline to copy data from HTTP to Azure Data Lake Storage Gen2.

Properties

Task name CopyPipeline_product

Task description Copying the data

Source

Connection name Http_to_CSV

Dataset name SourceDataset_ead

Column delimiter ,

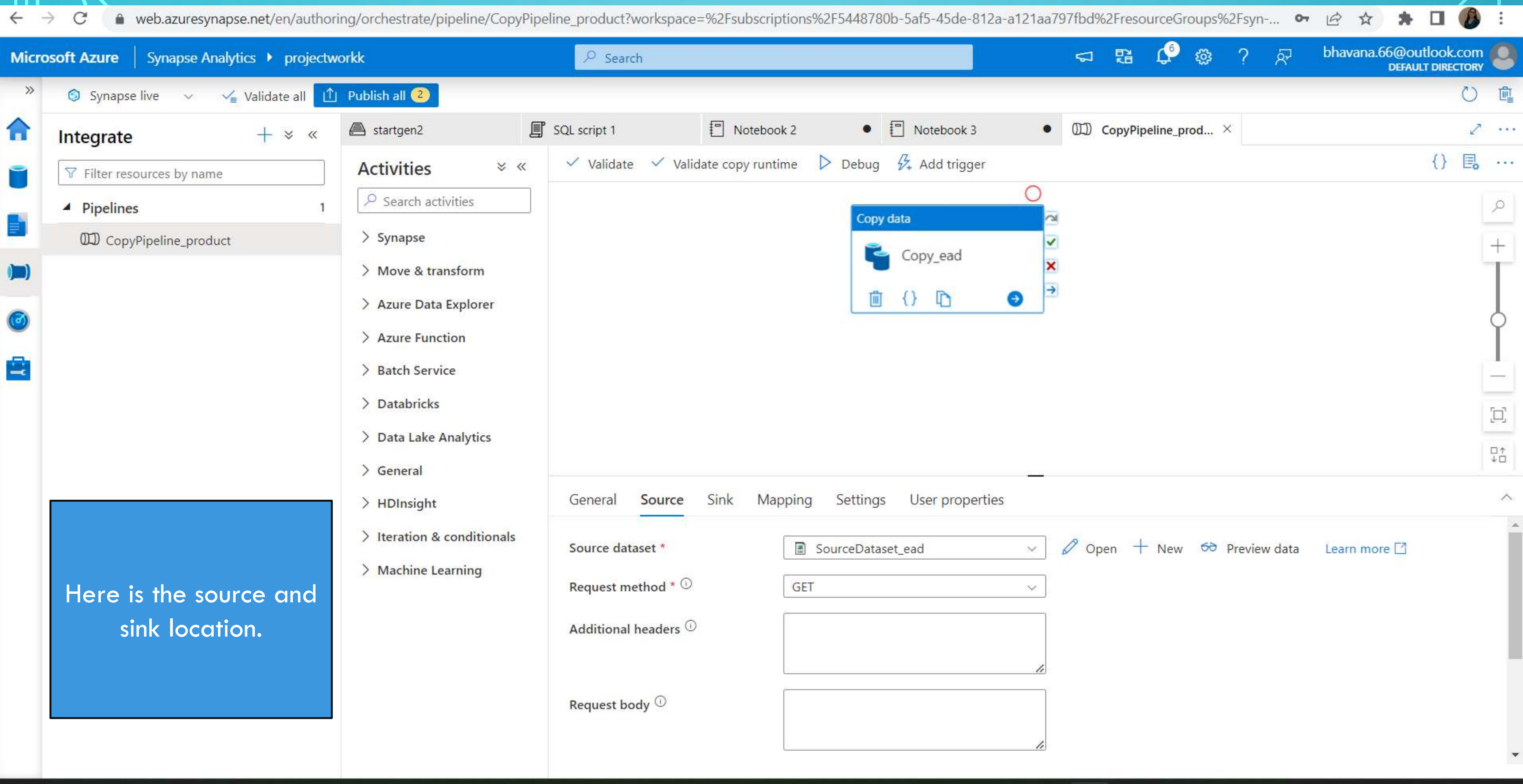
Row delimiter

Escape character \

Quote char "

First row as header true

Here my data is fully ingest into ADLS GEN 2 STORAGE ACCOUNT FOR FURTHER PROCESS.



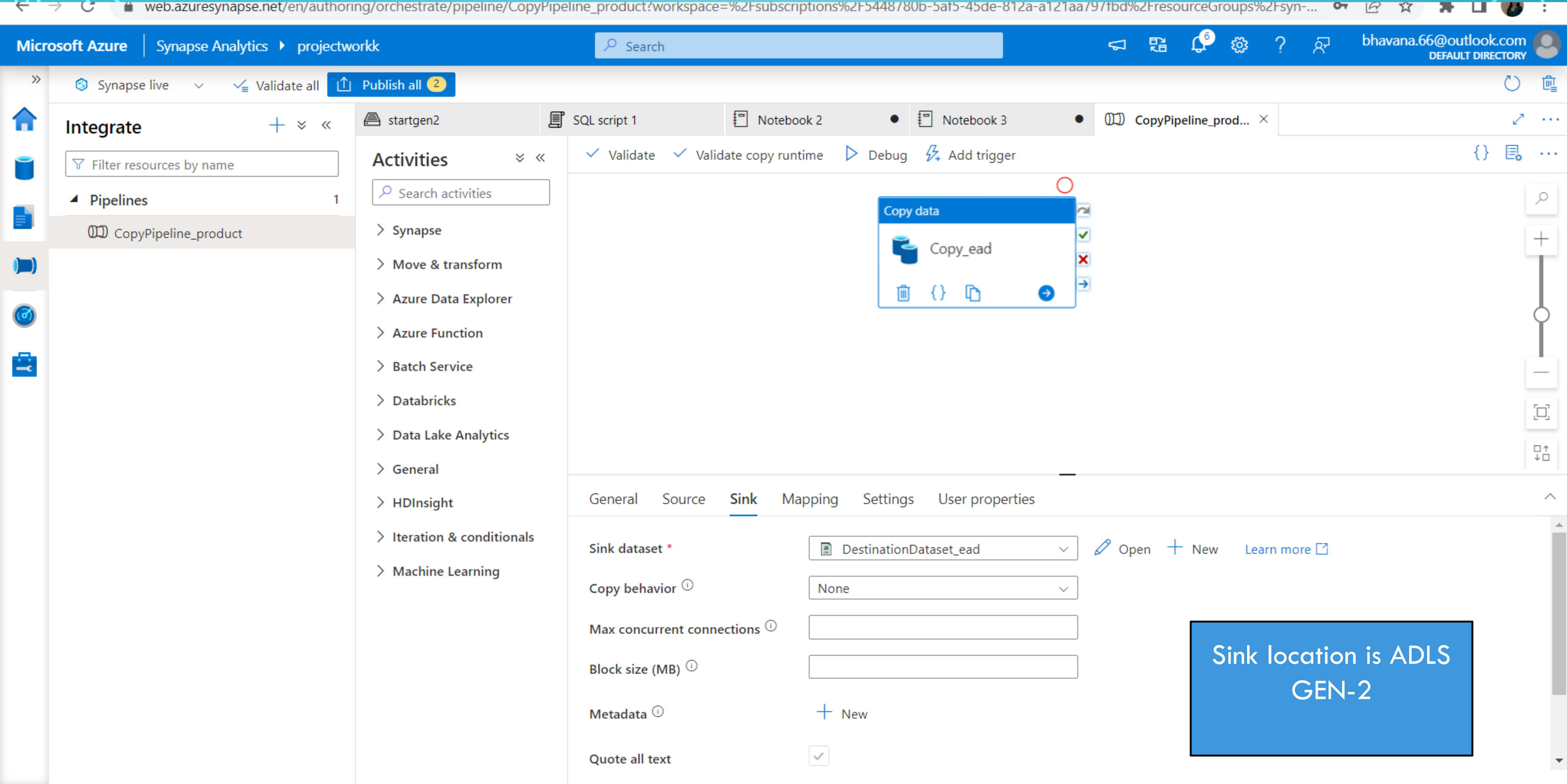
Source dataset * SourceDataset_ead

Request method * GET

Additional headers

Request body

Here is the source and sink location.



Sink dataset * DestinationDataset_ead

Copy behavior None

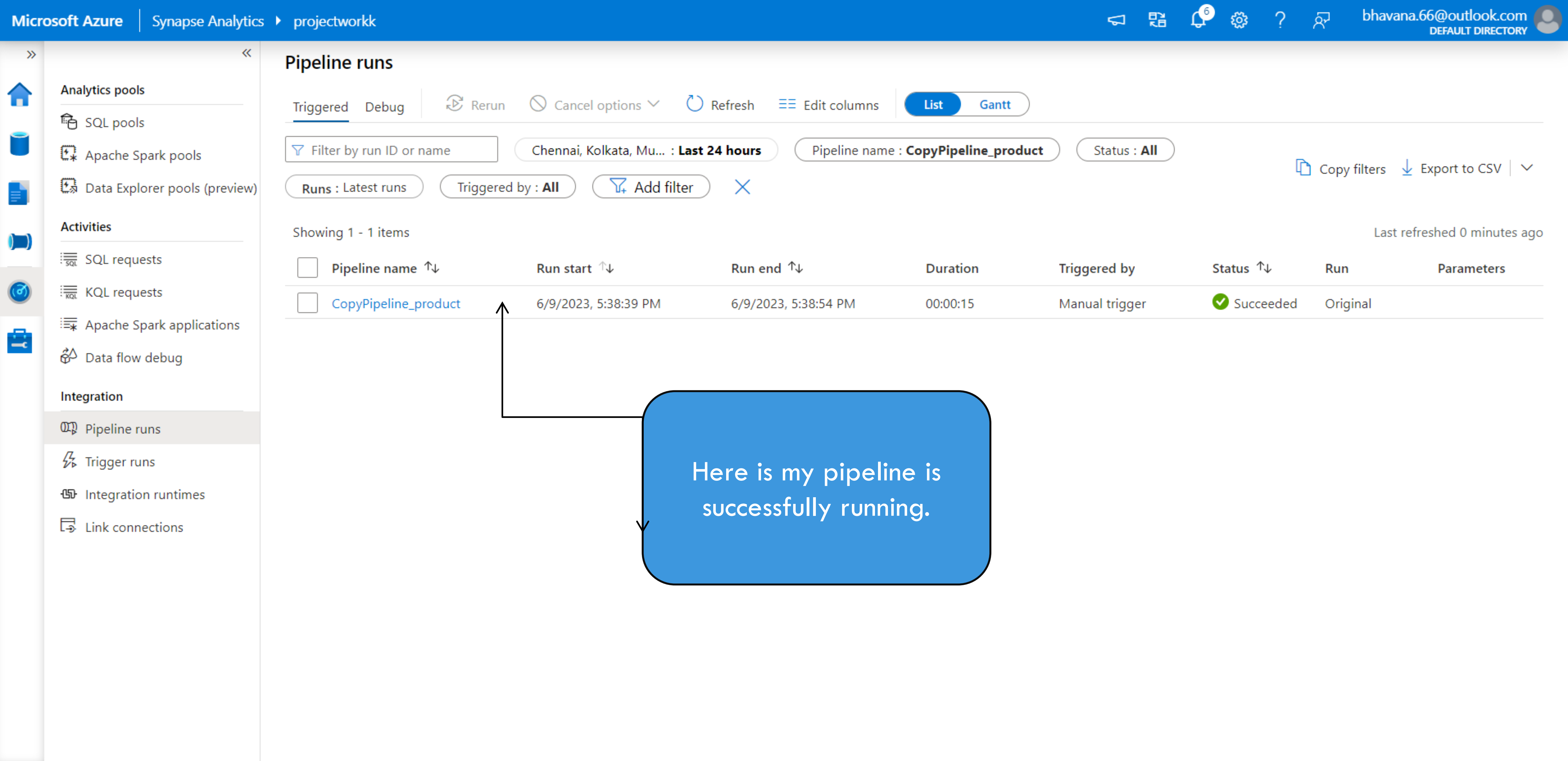
Max concurrent connections

Block size (MB)

Metadata + New

Quote all text ✓

Sink location is ADLS GEN-2



Pipeline runs

Chennai, Kolkata, Mu... : **Last 24 hours**

Pipeline name : **CopyPipeline_product**

Status : **All**

Runs : Latest runs

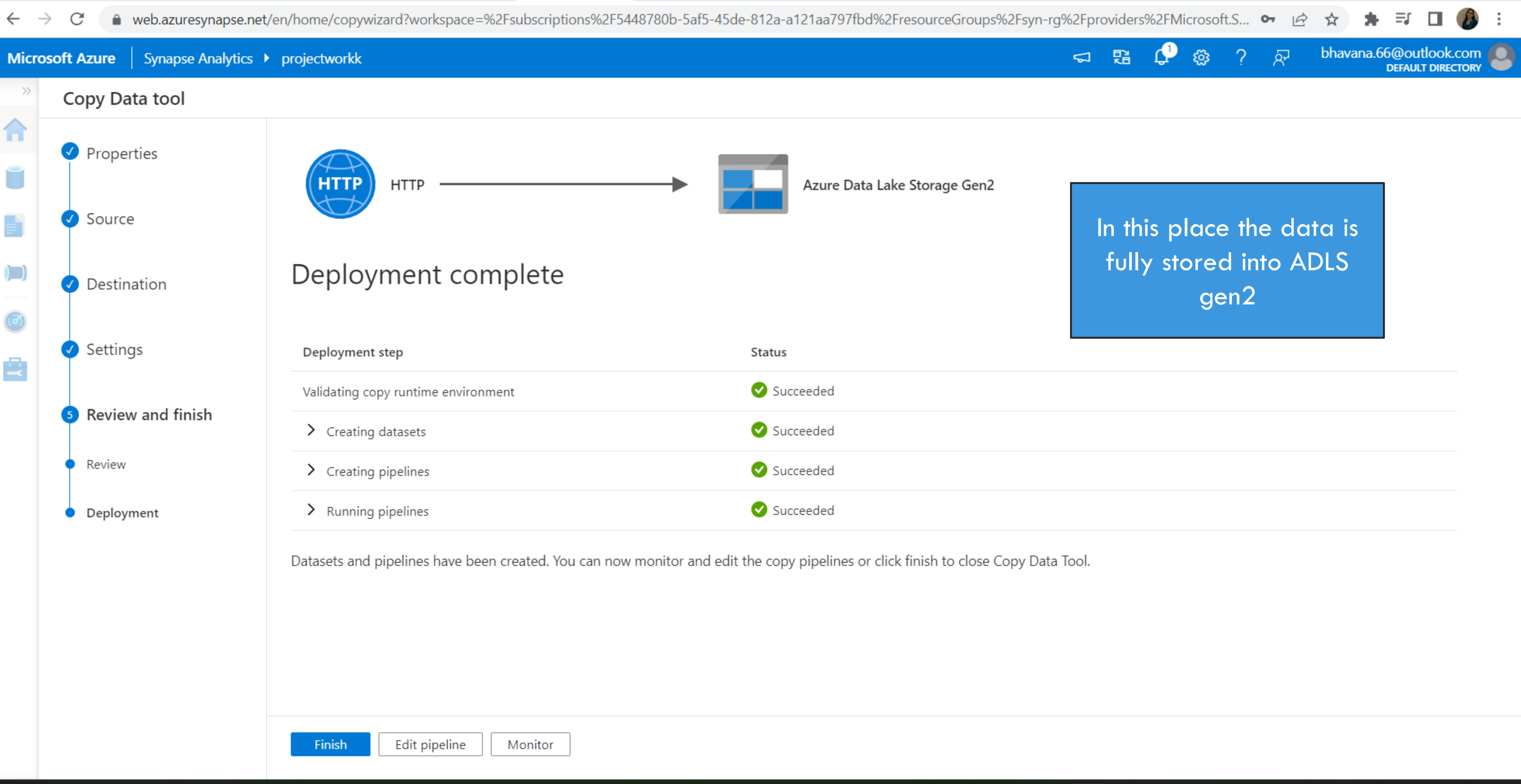
Triggered by : **All**

Showing 1 - 1 items

Last refreshed 0 minutes ago

	Pipeline name	Run start	Run end	Duration	Triggered by	Status	Run	Parameters
□	CopyPipeline_product	6/9/2023, 5:38:39 PM	6/9/2023, 5:38:54 PM	00:00:15	Manual trigger	✓ Succeeded	Original	

Here is my pipeline is successfully running.



web.azuresynapse.net/en/home/copywizard?workspace=%2Fsubscriptions%2F5448780b-5af5-45de-812a-a121aa797fbd%2FresourceGroups%2Fsyn-rg%2Fproviders%2FMicrosoft.S...

Microsoft Azure | Synapse Analytics ▸ projectworkk

bhavana.66@outlook.com
DEFAULT DIRECTORY

Copy Data tool

- ✓ Properties
- ✓ Source
- ✓ Destination
- ✓ Settings
- 5 Review and finish
  - Review
  - Deployment

HTTP ⟶ Azure Data Lake Storage Gen2

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.

In this place the data is fully stored into ADLS gen2

PROBLEM SECOND

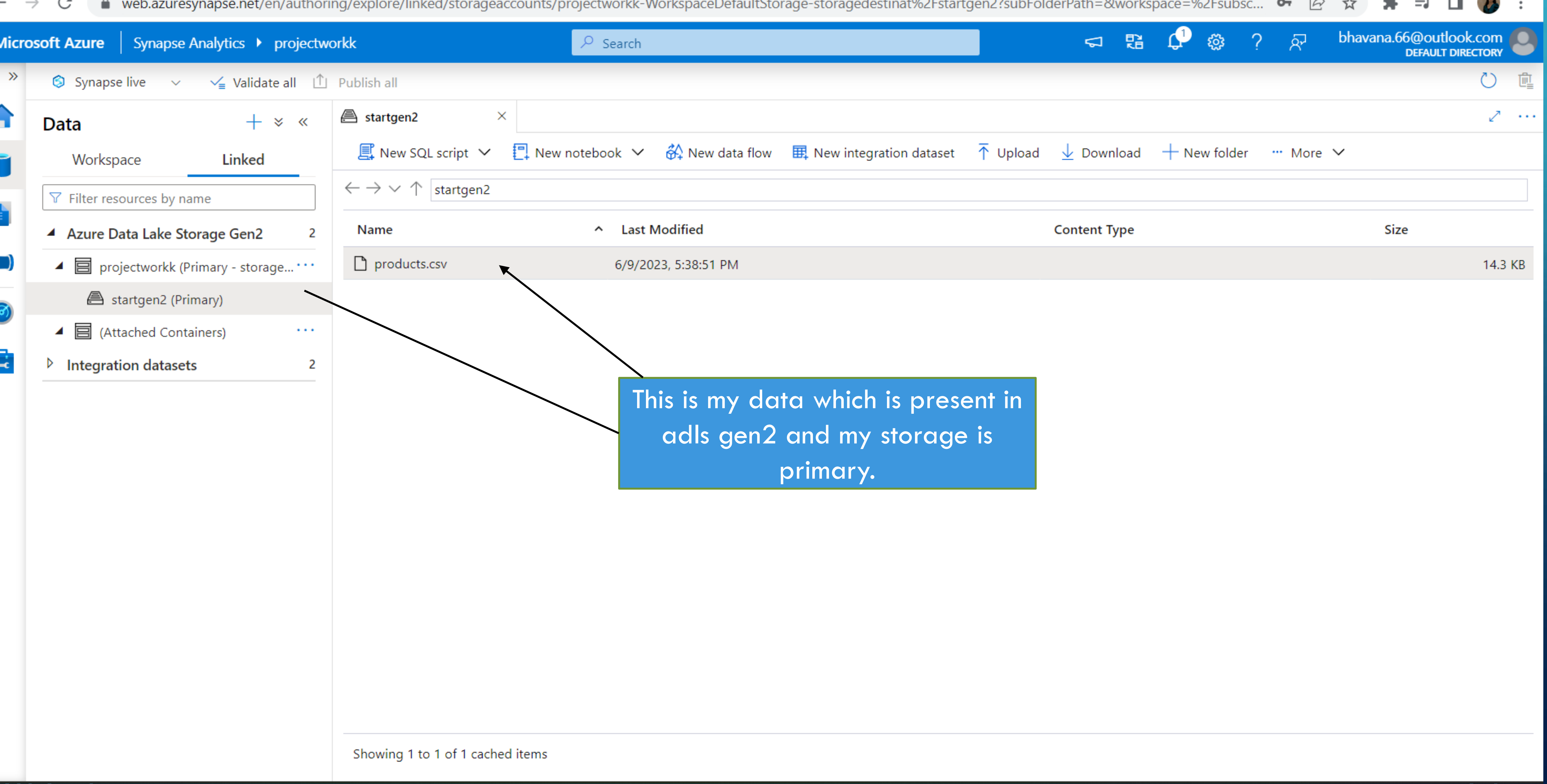
- PROBLEM STATEMENT 2: YOU NEED TO FIND THE TOP 100 ROWS FROM NEW SQL SCRIPT FROM YOUR DATA IN THE WORKSPACE, RUN THE CODE AND CHECK THE RESULT DATASETS. THEN, UPDATE THE QUERY BY SELECTING THE CATEGORY AND COUNT AS PRODUCT NUMBERS. FINALLY, MAKE NECESSARY CHANGES TO THE CHART VIEW.

AZURE DATA ANALYTICS WITH SQL POOL ENVIRONMENT

- CLOUD-BASED DATA WAREHOUSING SOLUTION FOR STORING AND MANAGING LARGE AMOUNTS OF DATA.
- HIGHLY SCALABLE AND SECURE ENVIRONMENT FOR ANALYZING LARGE DATASETS.
- PROVIDES ADVANCED SECURITY FEATURES AND SEAMLESS INTEGRATION WITH OTHER AZURE SERVICES.

TECH STACK USED IN 2ND PROBLEM

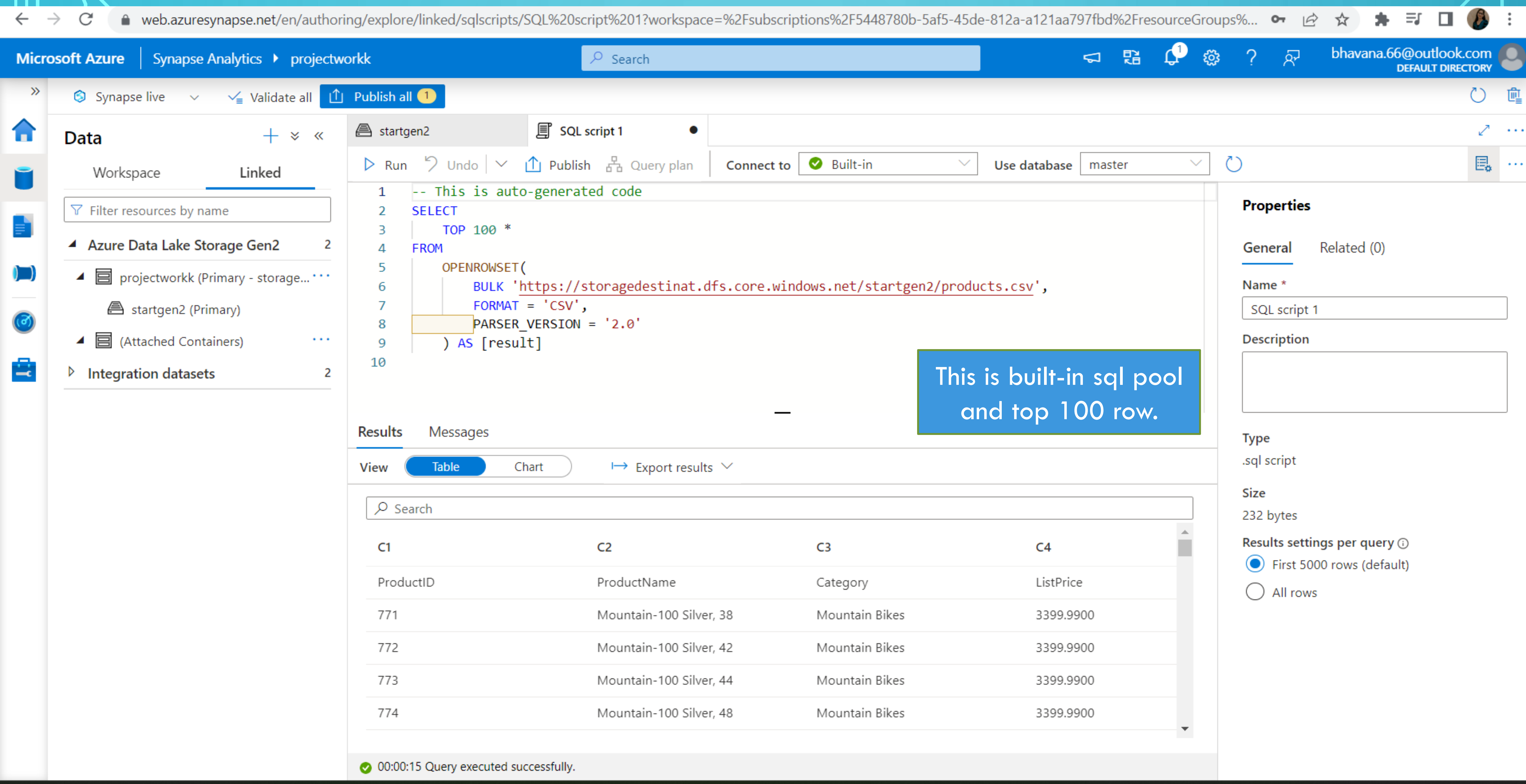
- NEW SQL SCRIPT: A SCRIPT WRITTEN IN SQL (STRUCTURED QUERY LANGUAGE) USED TO QUERY DATA FROM A DATABASE.
- WORKSPACE: A DATA ANALYTICS WORKSPACE IN AZURE SYNAPSE ANALYTICS WHERE DATA CAN BE INGESTED, TRANSFORMED, AND ANALYZED.
- CHART VIEW: A GRAPHICAL REPRESENTATION OF DATA THAT CAN BE CREATED IN AZURE SYNAPSE ANALYTICS.



Name	Last Modified	Content Type	Size
products.csv	6/9/2023, 5:38:51 PM		14.3 KB

Showing 1 to 1 of 1 cached items

This is my data which is present in adls gen2 and my storage is primary.



```
-- This is auto-generated code
SELECT
    TOP 100 *
FROM
    OPENROWSET(
        BULK 'https://storagedestinat.dfs.core.windows.net/startgen2/products.csv',
        FORMAT = 'CSV',
        PARSER_VERSION = '2.0'
    ) AS [result]
```

This is built-in sql pool and top 100 row.

C1	C2	C3	C4
ProductID	ProductName	Category	ListPrice
771	Mountain-100 Silver, 38	Mountain Bikes	3399.9900
772	Mountain-100 Silver, 42	Mountain Bikes	3399.9900
773	Mountain-100 Silver, 44	Mountain Bikes	3399.9900
774	Mountain-100 Silver, 48	Mountain Bikes	3399.9900

00:00:15 Query executed successfully.

Name *: SQL script 1

Description

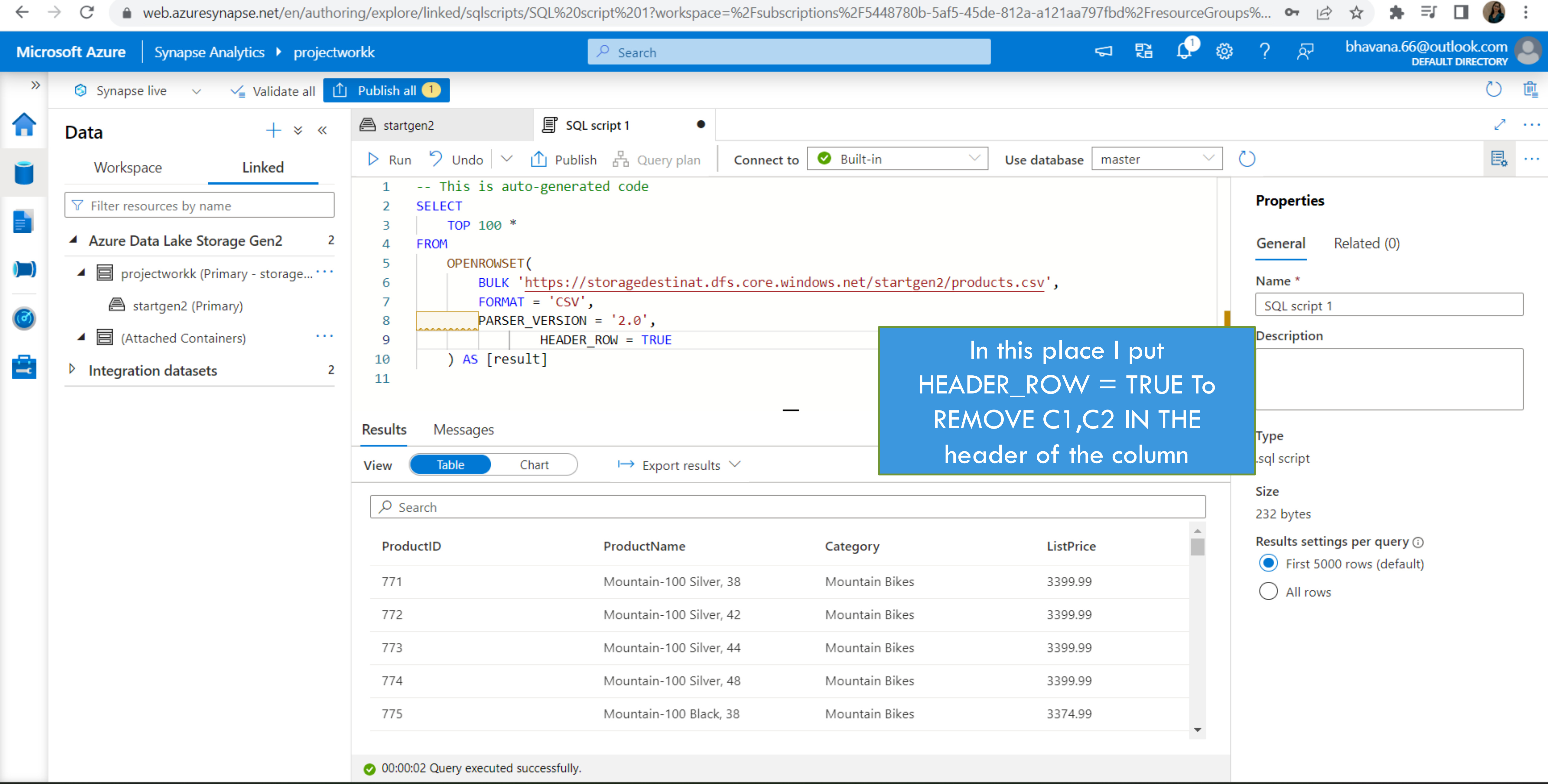
Type: .sql script

Size: 232 bytes

Results settings per query

● First 5000 rows (default)

○ All rows



Connect to: Built-in

Use database: master

```
-- This is auto-generated code
SELECT
    TOP 100 *
FROM
    OPENROWSET(
        BULK 'https://storagedestinat.dfs.core.windows.net/startgen2/products.csv',
        FORMAT = 'CSV',
        PARSER_VERSION = '2.0',
                HEADER_ROW = TRUE
    ) AS [result]
```

ProductID	ProductName	Category	ListPrice
771	Mountain-100 Silver, 38	Mountain Bikes	3399.99
772	Mountain-100 Silver, 42	Mountain Bikes	3399.99
773	Mountain-100 Silver, 44	Mountain Bikes	3399.99
774	Mountain-100 Silver, 48	Mountain Bikes	3399.99
775	Mountain-100 Black, 38	Mountain Bikes	3374.99

00:00:02 Query executed successfully.

Name *: SQL script 1

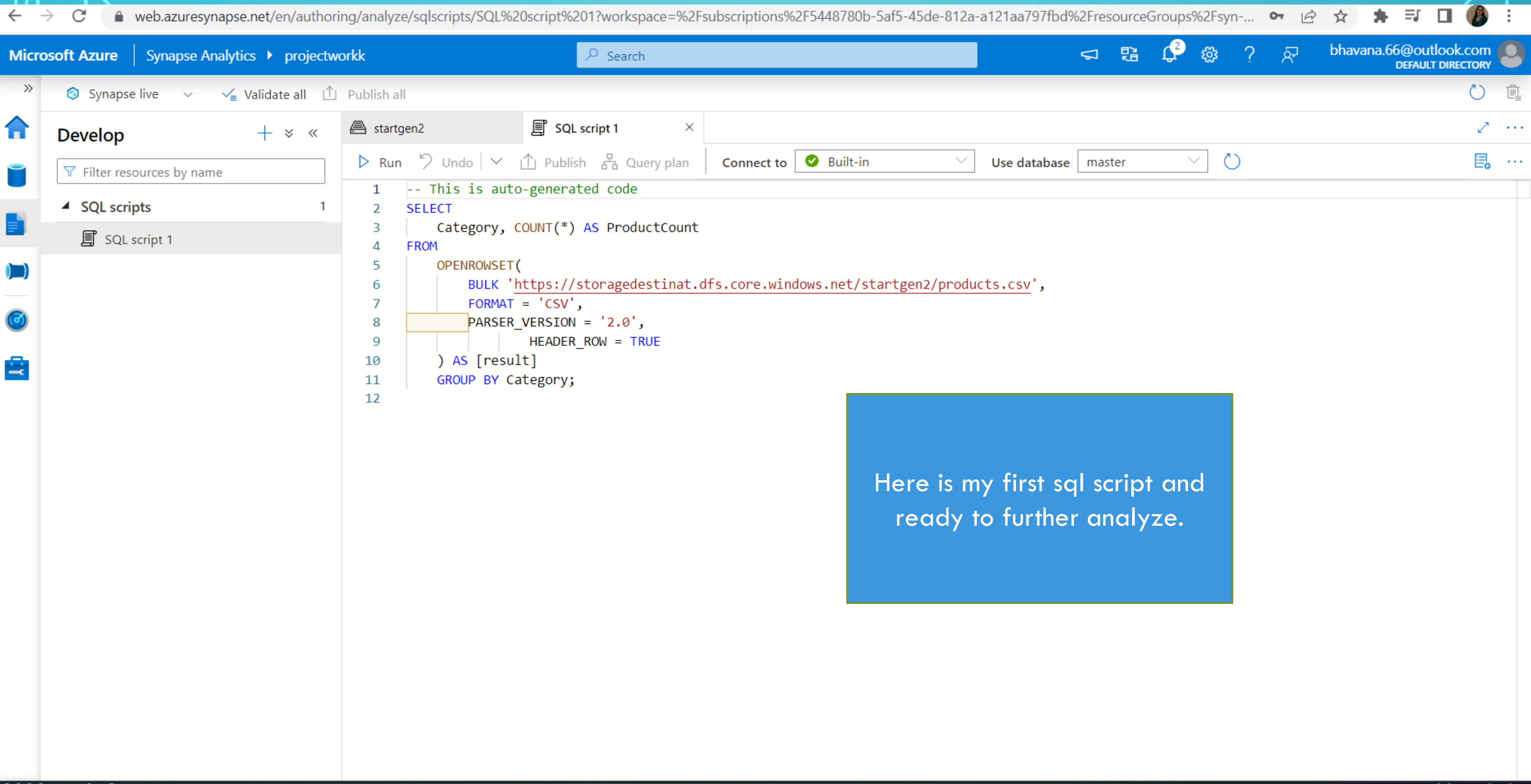
Description:

Type: .sql script

Size: 232 bytes

Results settings per query
● First 5000 rows (default)
○ All rows

In this place I put HEADER_ROW = TRUE To REMOVE C1,C2 IN THE header of the column



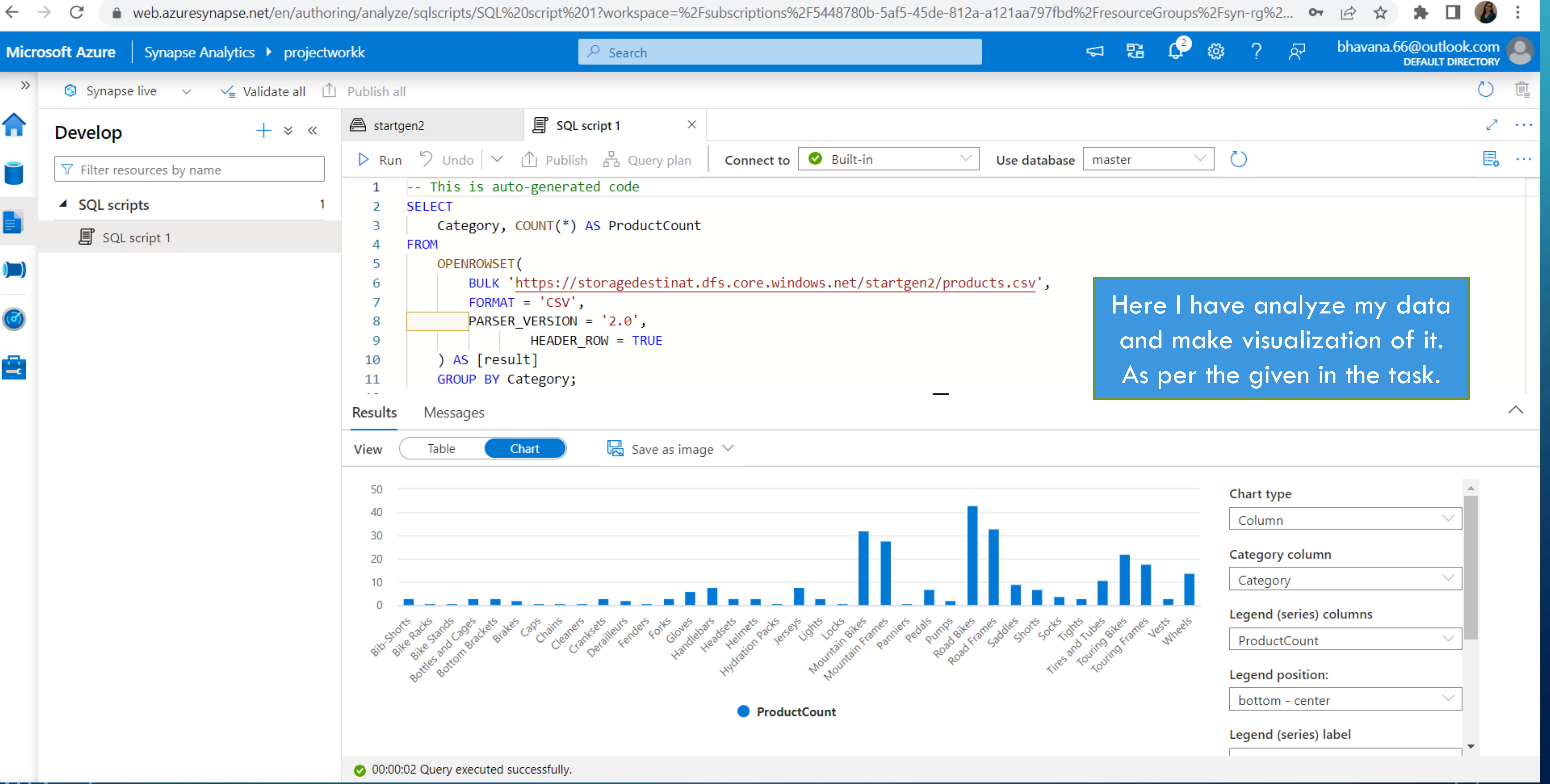
web.azuresynapse.net/en/authoring/analyze/sqlscripts/SQL%20script%201?workspace=%2Fsubscriptions%2F5448780b-5af5-45de-812a-a121aa797fbd%2FresourceGroups%2Fsyn-...

Connect to: Built-in

Use database: master

```
-- This is auto-generated code
SELECT
    Category, COUNT(*) AS ProductCount
FROM
    OPENROWSET(
        BULK 'https://storagedestinat.dfs.core.windows.net/startgen2/products.csv',
        FORMAT = 'CSV',
        PARSER_VERSION = '2.0',
                HEADER_ROW = TRUE
    ) AS [result]
    GROUP BY Category;
```

Here is my first sql script and ready to further analyze.



Connect to: Built-in

Use database: master

```sql
-- This is auto-generated code
SELECT
    Category, COUNT(*) AS ProductCount
FROM
    OPENROWSET(
        BULK 'https://storagedestinat.dfs.core.windows.net/startgen2/products.csv',
        FORMAT = 'CSV',
        PARSER_VERSION = '2.0',
                HEADER_ROW = TRUE
    ) AS [result]
    GROUP BY Category;
```

Chart type: Column

Category column: Category

Legend (series) columns: ProductCount

Legend position: bottom - center

00:00:02 Query executed successfully.

Here I have analyze my data and make visualization of it. As per the given in the task.

PROBLEM THIRD

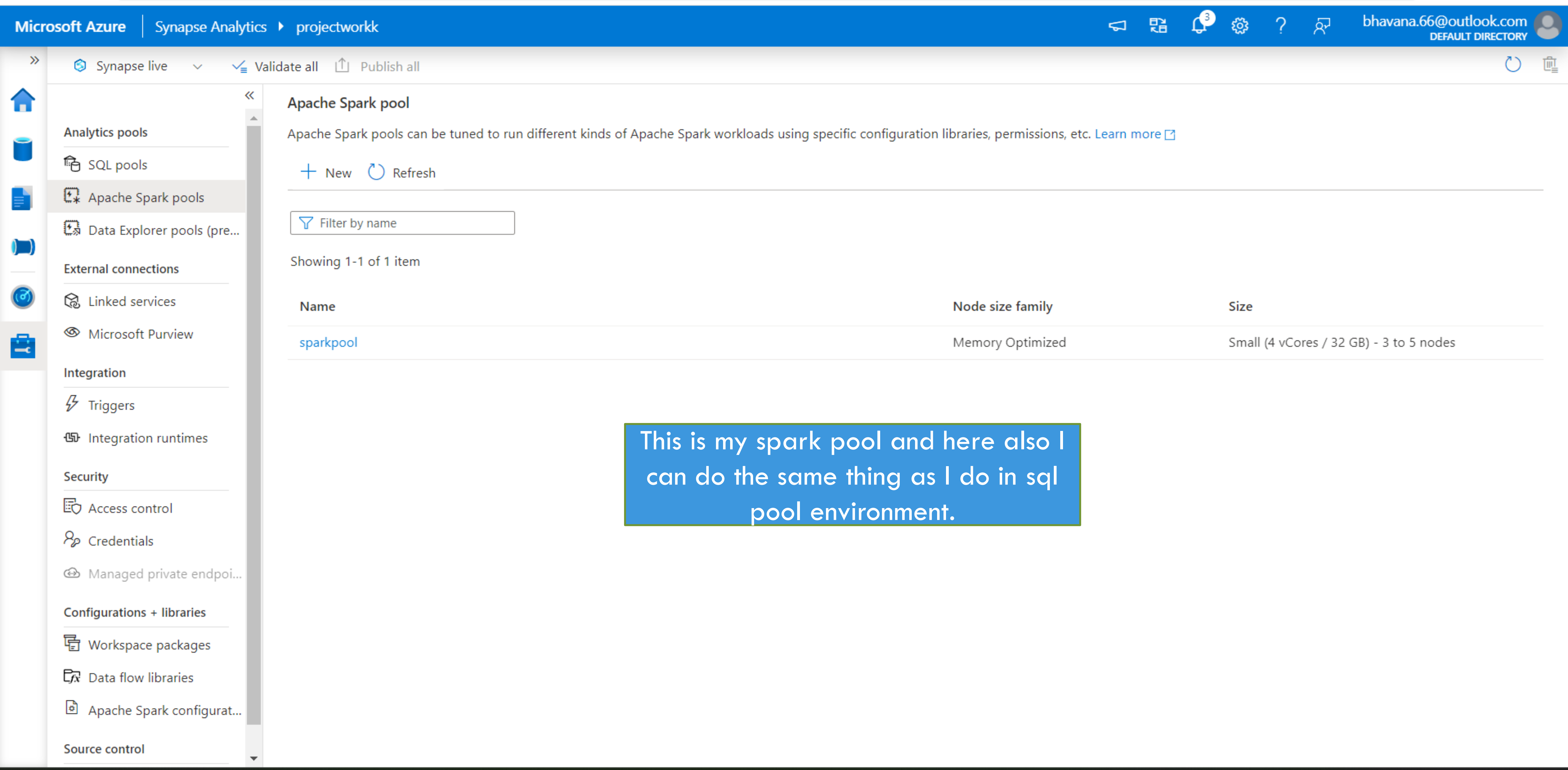
- PROBLEM STATEMENT 3:
- THE TASK INVOLVES FINDING THE TOP 100 ROWS FROM NEW SQL SCRIPT FROM THE DATA IN THE WORKSPACE, RUNNING THE CODE, AND CHECKING THE RESULT DATASETS. THEN, THE QUERY WILL BE UPDATED BY SELECTING THE CATEGORY AND COUNT AS PRODUCT NUMBERS. FINALLY, NECESSARY CHANGES WILL BE MADE TO THE CHART VIEW.

AZURE DATA ANALYTICS WITH SPARK POOL ENVIRONMENT.

- CLOUD-BASED BIG DATA PROCESSING SOLUTION USING APACHE SPARK.
- FULLY MANAGED SPARK ENVIRONMENT FOR PROCESSING LARGE DATASETS.
- HIGHLY SCALABLE AND INTEGRATES SEAMLESSLY WITH OTHER AZURE SERVICES.

TECH STACK USED IN 3RD PROBLEM

- SPARK POOL: A MANAGED APACHE SPARK SERVICE IN AZURE SYNAPSE ANALYTICS THAT ALLOWS YOU TO PROCESS BIG DATA WORKLOADS. IT CAN BE USED FOR DATA TRANSFORMATION, MACHINE LEARNING, AND DATA VISUALIZATION.
- APACHE SPARK: AN OPEN-SOURCE DISTRIBUTED COMPUTING SYSTEM USED FOR PROCESSING LARGE DATA SETS. IT IS DESIGNED TO BE FAST, EASY TO USE, AND SCALABLE.
- MANAGE HUB: A CENTRALIZED MANAGEMENT PORTAL IN AZURE SYNAPSE ANALYTICS WHERE YOU CAN MANAGE RESOURCES SUCH AS SQL POOLS, SPARK POOLS, AND DATA FLOWS.



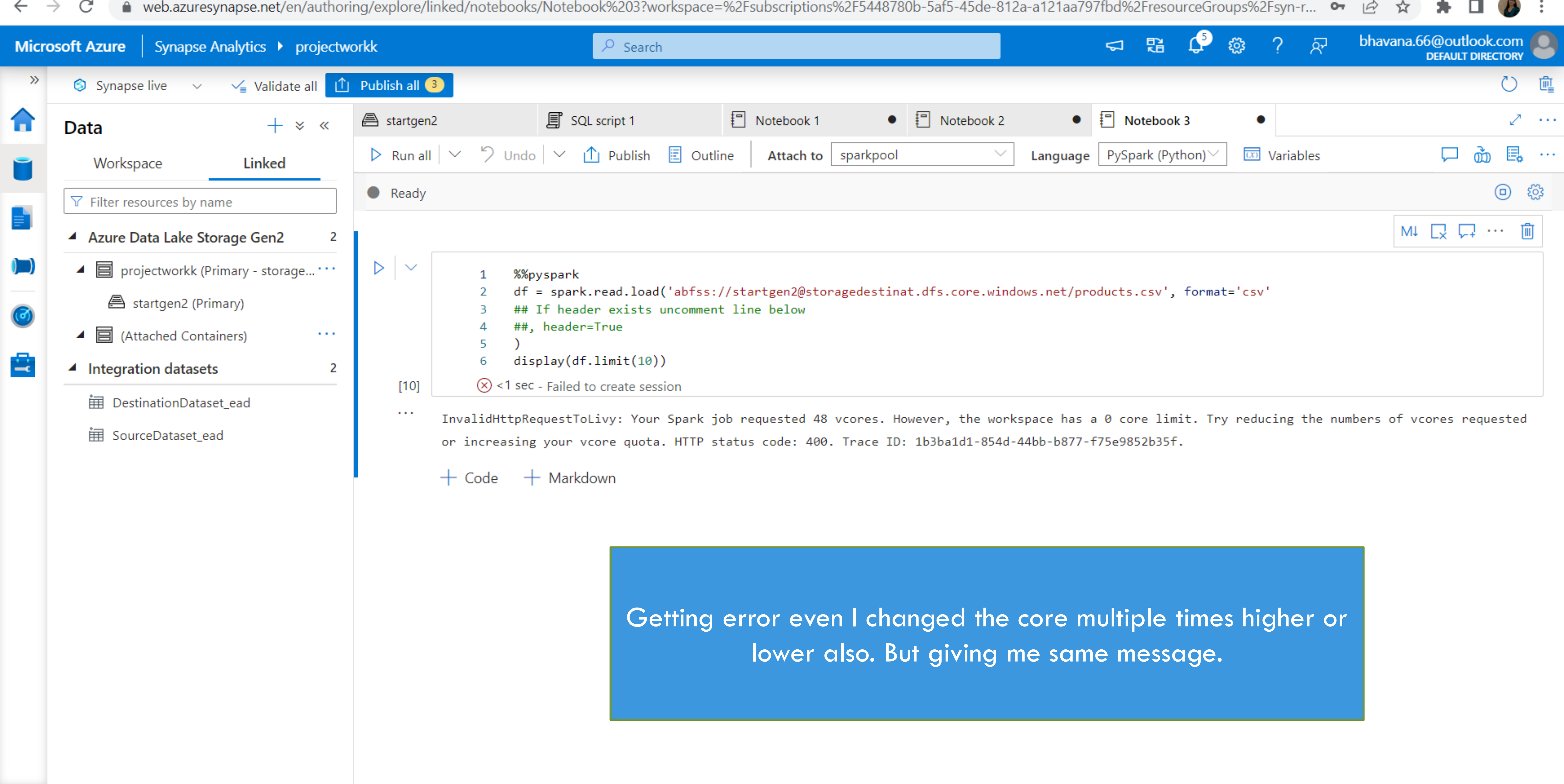
Apache Spark pool

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

Showing 1-1 of 1 item

Name	Node size family	Size
sparkpool	Memory Optimized	Small (4 vCores / 32 GB) - 3 to 5 nodes

This is my spark pool and here also I can do the same thing as I do in sql pool environment.



```
%%pyspark
df = spark.read.load('abfss://startgen2@storagedestinat.dfs.core.windows.net/products.csv', format='csv'
## If header exists uncomment line below
##, header=True
)
display(df.limit(10))
```

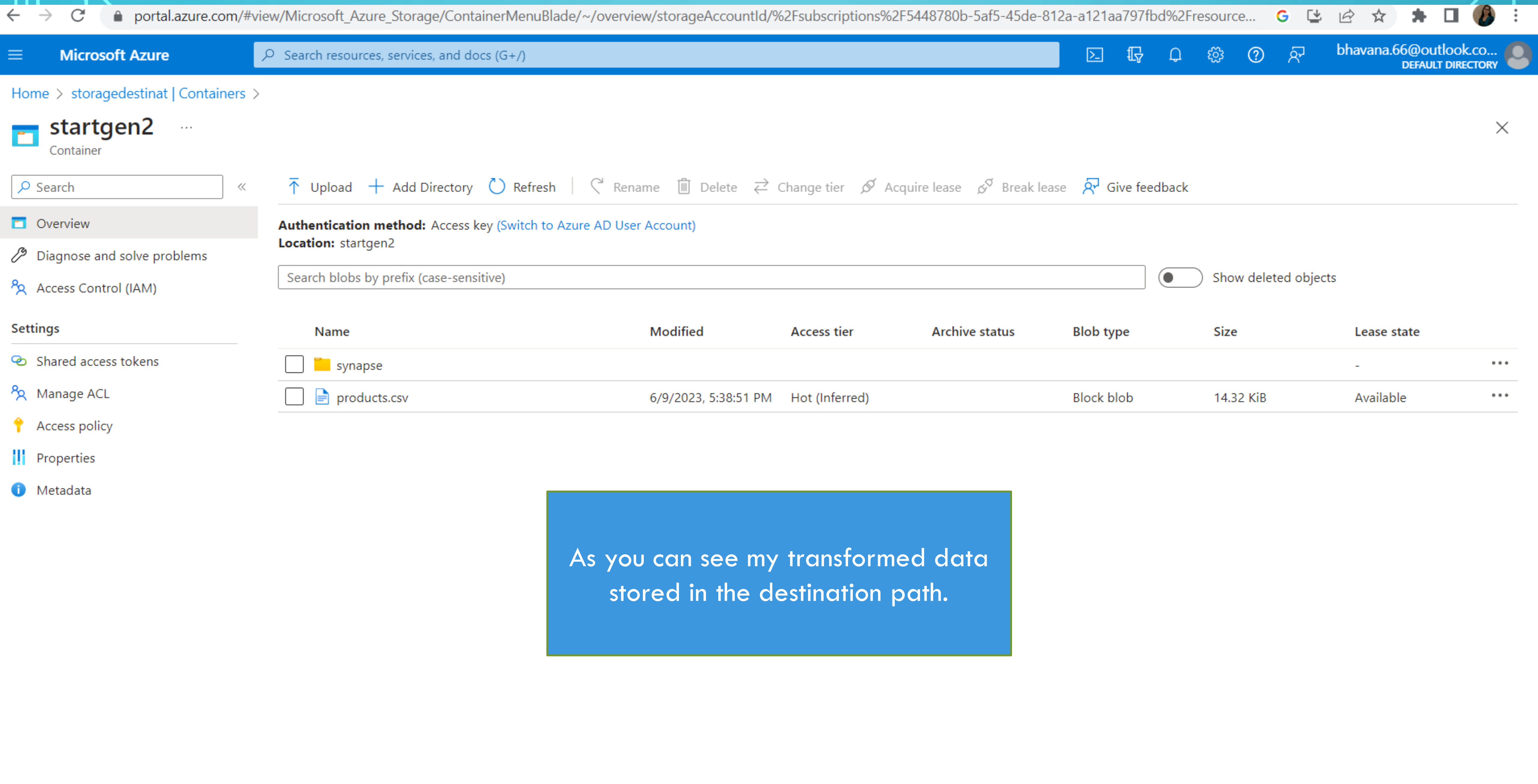
<1 sec - Failed to create session

```
InvalidHttpRequestToLivy: Your Spark job requested 48 vcores. However, the workspace has a 0 core limit. Try reducing the numbers of vcores requested
or increasing your vcore quota. HTTP status code: 400. Trace ID: 1b3ba1d1-854d-44bb-b877-f75e9852b35f.
```

Getting error even I changed the core multiple times higher or lower also. But giving me same message.

FINAL LOCATION OF THE DATA PRESENT IN ADLS_GEN2 STORAGE ACCOUNT

THE DATA IS FINALLY COME TO MY ADLS-GEN2 STORAGE ACCOUNT .


portal.azure.com/#view/Microsoft_Azure_Storage/ContainerMenuBlade/~/overview/storageAccountId/%2Fsubscriptions%2F5448780b-5af5-45de-812a-a121aa797fbd%2Fresource...
Microsoft Azure
Search resources, services, and docs (G+/)
bhavana.66@outlook.co...
DEFAULT DIRECTORY
Home > storagedestinat | Containers >
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
Search blobs by prefix (case-sensitive)
Show deleted objects

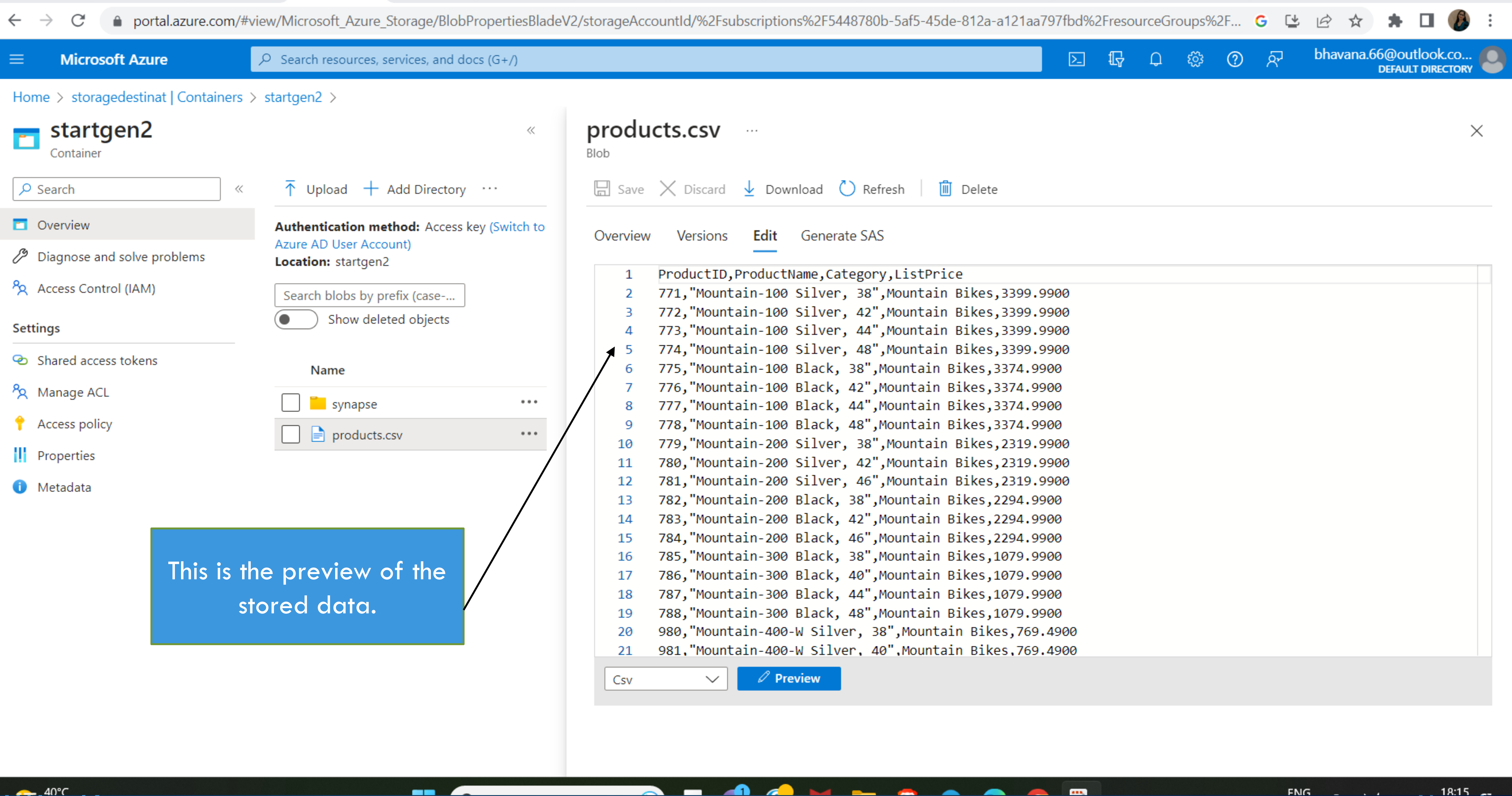
startgen2

Container

Authentication method: Access key (Switch to Azure AD User Account)
Location: startgen2

	Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
□	synapse						-
□	products.csv	6/9/2023, 5:38:51 PM	Hot (Inferred)		Block blob	14.32 KiB	Available

As you can see my transformed data stored in the destination path.



Authentication method: Access key (Switch to Azure AD User Account)
Location: startgen2

```
ProductID,ProductName,Category,ListPrice
771,"Mountain-100 Silver, 38",Mountain Bikes,3399.9900
772,"Mountain-100 Silver, 42",Mountain Bikes,3399.9900
773,"Mountain-100 Silver, 44",Mountain Bikes,3399.9900
774,"Mountain-100 Silver, 48",Mountain Bikes,3399.9900
775,"Mountain-100 Black, 38",Mountain Bikes,3374.9900
776,"Mountain-100 Black, 42",Mountain Bikes,3374.9900
777,"Mountain-100 Black, 44",Mountain Bikes,3374.9900
778,"Mountain-100 Black, 48",Mountain Bikes,3374.9900
779,"Mountain-200 Silver, 38",Mountain Bikes,2319.9900
780,"Mountain-200 Silver, 42",Mountain Bikes,2319.9900
781,"Mountain-200 Silver, 46",Mountain Bikes,2319.9900
782,"Mountain-200 Black, 38",Mountain Bikes,2294.9900
783,"Mountain-200 Black, 42",Mountain Bikes,2294.9900
784,"Mountain-200 Black, 46",Mountain Bikes,2294.9900
785,"Mountain-300 Black, 38",Mountain Bikes,1079.9900
786,"Mountain-300 Black, 40",Mountain Bikes,1079.9900
787,"Mountain-300 Black, 44",Mountain Bikes,1079.9900
788,"Mountain-300 Black, 48",Mountain Bikes,1079.9900
980,"Mountain-400-W Silver, 38",Mountain Bikes,769.4900
981,"Mountain-400-W Silver, 40",Mountain Bikes,769.4900
```

This is the preview of the stored data.

Thank you