

MICROSOFT AZURE CAPSTONE PROJECT

BY RAHUL KUMAR

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

Microsoft Azure

demoac

1

Home >

syn-rg

Resource group

Search

<<

Create

Manage view

Delete resource group

Refresh

Export to CSV

Open query

Assign tags

Move

Delete

Export template

JSON View

Overview

Activity log

Access control (IAM)

Tags

Resource visualizer

Events

Settings

Deployments

Security

Policies

Properties

Locks

Cost Management

Cost analysis

Cost alerts (preview)

Budgets

Advisor recommendations

Resources

Recommendations

Filter for any field...

Type equals all

Location equals all

Add filter

Showing 1 to 1 of 1 records. ☐ Show hidden types ⓘ

☐ Name ↑↓

Type ↑↓

Location ↑↓

☐ projectworkkk

Synapse workspace

South Central US

< Previous

Page 1 of 1

Next >

Give feedback

This is my synapse workspace.

Home > syn-rg >

projectworkk

Synapse workspace

...

Search

- Overview
- Activity log
- Access control (IAM)
- Tags
- Diagnose and solve problems
- Settings
- Azure Active Directory
- Properties
- Locks
- Analytics pools
- SQL pools
- Apache Spark pools
- Data Explorer pools (preview)
- Security
- Encryption
- Networking
- Identity

+ New dedicated SQL pool

+ New Apache Spark pool

+ New Data Explorer pool (preview)

Refresh

Reset SQL admin password

Delete

Essentials

JSON View

Resource group (move)	: syn-rg	Networking	: Show firewall settings
Status	: Succeeded	Primary ADLS Gen2 acco...	: https://storagedestinat.dfs.core.windows.net
Location	: South Central US	Primary ADLS Gen2 file s...	: startgen2
Subscription (move)	: Azure for Students	SQL admin username	: sqladminuser
Subscription ID	: 5448780b-5af5-45de-812a-a121aa797fbd	SQL Active Directory ad...	: live.com#bhavana.66@outlook.com
Managed virtual network	: No	Dedicated SQL endpoint	: projectworkk.sql.azuresynapse.net
Managed Identity object ...	: a0ac1929-574d-4e22-9d10-ff6757b464cf	Serverless SQL endpoint	: projectworkk-ondemand.sql.azuresynapse.net
Workspace web URL	: https://web.azuresynapse.net?workspace=%2fsubscriptions%2f54...	Development endpoint	: https://projectworkk.dev.azuresynapse.net
Tags (edit)	: Click here to add tags		

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 launch.

Analytics pools

Search to filter items...

Name	Type	Size
------	------	------

Home >

storagedestinat
Storage account

Search

- Overview
- Activity log
- Tags
- Diagnose and solve problems
- Access Control (IAM)
- Data migration
- Events
- Storage browser
- Data storage
- Containers
- File shares
- Queues
- Tables
- Security + networking
- Networking
- Access keys
- Shared access signature

Upload Open in Explorer Delete Move Refresh Open in mobile CLI / PS Feedback

Essentials

Resource group (move)	: destinationADLSgen2	Performance	: Standard
Location	: West US 2	Replication	: Read-access geo-redundant storage (RA-GRS)
Primary/Secondary Location	: Primary: West US 2, Secondary: West Central US	Account kind	: StorageV2 (general purpose v2)
Subscription (move)	: Azure for Students	Provisioning state	: Succeeded
Subscription ID	: 5448780b-5af5-45de-812a-a121aa797fbd	Created	: 6/9/2023, 5:00:50 PM
Disk state	: Primary: Available, Secondary: Available		
Tags (edit)	: Click here to add tags		

Properties

Monitoring

Capabilities (5)

Recommendations (0)

Tutorials

Tools + SDKs

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.

+ New connection

Cancel

Create

[Back](#)

Copy Data tool

- ✓ Properties
- 2 Source**
- Dataset
- Configuration
- 3 Destination
- 4 Settings
- 5 Review and finish

File format

File format

Delimited

Column del

Comma (,)

☐ Edit

Row delimit

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.**

< Previous

Next >

Cancel

Copy Data tool

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

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

Edit

Source

Connection name	Http_to_CSV
Dataset name	SourceDataset_ead
Column delimiter	,
Row delimiter	
Escape character	\
Quote char	"
First row as header	true

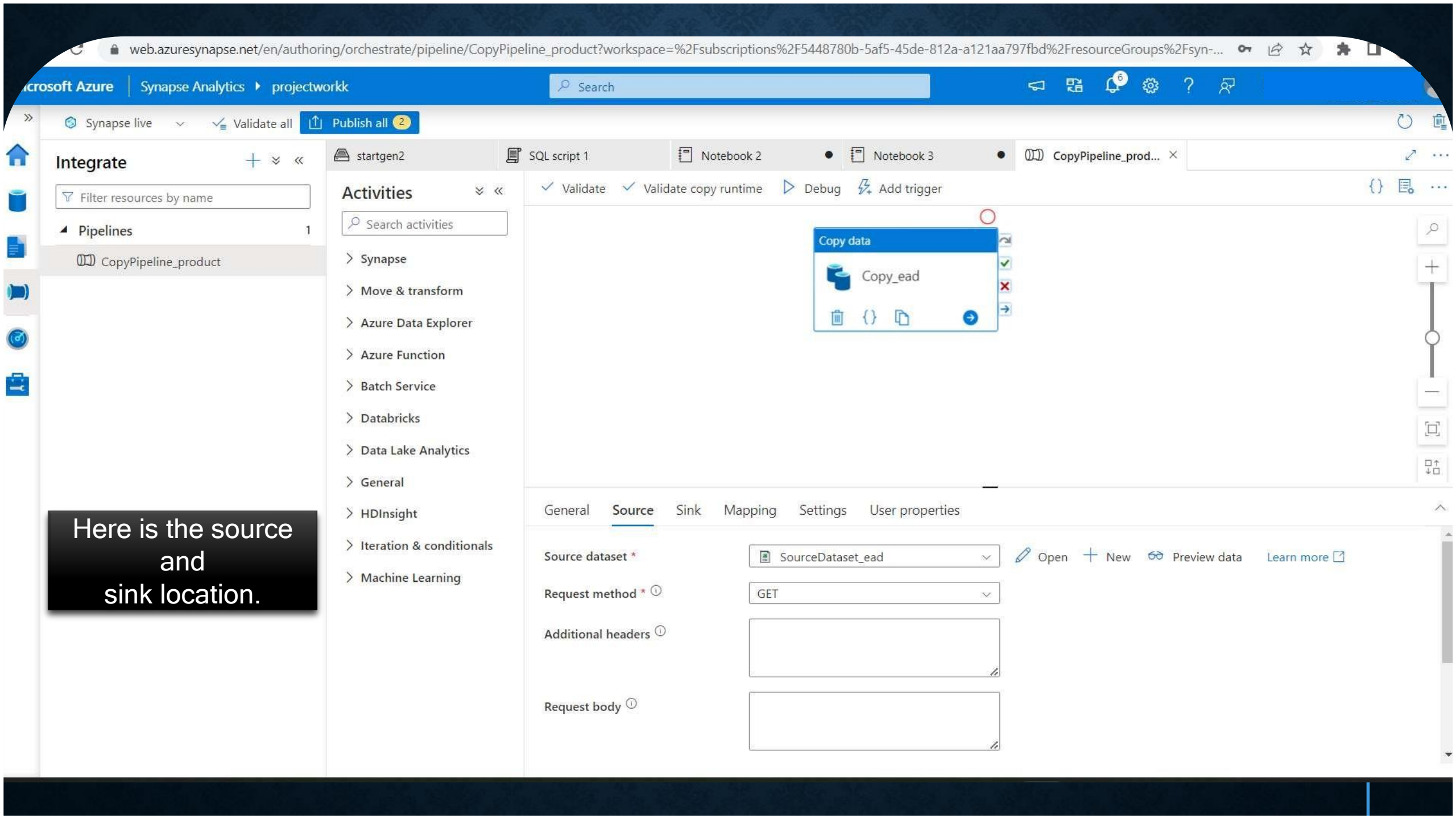
Edit

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

< Previous

Next >

Cancel



Here is the source
and
sink location.

>>

Synapse live

Validate all

Publish all 2

Integrate

Filter resources by name

Pipelines 1

CopyPipeline_product

Activities

Search activities

Synapse

Move & transform

Azure Data Explorer

Azure Function

Batch Service

Databricks

Data Lake Analytics

General

HDInsight

Iteration & conditionals

Machine Learning

startgen2

SQL script 1

Notebook 2

Notebook 3

CopyPipeline_prod...

Validate

Validate copy runtime

Debug

Add trigger

Copy data

Copy_ead

General

Source

Sink

Mapping

Settings

User properties

Sink dataset *

DestinationDataset_ead

Open

New

Learn more

Copy behavior ⓘ

None

Max concurrent connections ⓘ

Block size (MB) ⓘ

Metadata ⓘ

New

Quote all text

Sink location is ADLS GEN-2

>> <<

Analytics pools

SQL pools

Apache Spark pools

Data Explorer pools (preview)

Activities

SQL requests

KQL requests

Apache Spark applications

Data flow debug

Integration

Pipeline runs

Trigger runs

Integration runtimes

Link connections

Pipeline runs

Triggered Debug Rerun Cancel options Refresh Edit columns List Gantt

Filter by run ID or name

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

Pipeline name : CopyPipeline_product

Status : All

Copy filters Export to CSV

Runs : Latest runs Triggered by : All Add filter

Showing 1 - 1 items

Last refreshed 0 minutes ago

<input type="checkbox"/> Pipeline name ↑↓	Run start ↑↓	Run end ↑↓	Duration	Triggered by	Status ↑↓	Run	Parameters
<input type="checkbox"/> 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.

Copy Data tool

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



HTTP



Azure Data Lake Storage Gen2

Deployment complete

**IN THIS PLACE THE DATA IS
FULLY STORED INTO ADLS
GEN2**

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.

[Finish](#) [Edit pipeline](#) [Monitor](#)

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.

Microsoft Azure

Synapse Analytics

projectworkk

Search

Synapse live

Validate all

Publish all

Data

Workspace

Linked

Filter resources by name

Azure Data Lake Storage Gen2

2

projectworkk (Primary - storage...

...

startgen2 (Primary)

(Attached Containers)

...

Integration datasets

2

startgen2

New SQL script

New notebook

New data flow

New integration dataset

Upload

Download

New folder

More

startgen2

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.

Microsoft Azure | Synapse Analytics | projectworkk

Search

Synapse live | Validate all | Publish all 1

Data

Workspace | Linked

Filter resources by name

Azure Data Lake Storage Gen2 2

- projectworkk (Primary - storage...
 - startgen2 (Primary)
- (Attached Containers)

Integration datasets 2

startgen2 | SQL script 1

Run | Undo | Publish | Query plan

Connect to Built-in | Use database master

```
1 -- This is auto-generated code
2 SELECT
3     TOP 100 *
4 FROM
5     OPENROWSET(
6         BULK 'https://storagedestinat.dfs.core.windows.net/startgen2/products.csv',
7         FORMAT = 'CSV',
8         PARSE_VERSION = '2.0'
9     ) AS [result]
10
```

Results | Messages

View Table | Chart | Export results

Search

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.

Properties

General | Related (0)

Name *
SQL script 1

Description

Type
.sql script

Size
232 bytes

Results settings per query

- ☒ First 5000 rows (default)
- ☐ All rows

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

Microsoft Azure | Synapse Analytics | projectworkk

Search

Synapse live | Validate all | Publish all 1

Data

Workspace | Linked

Filter resources by name

Azure Data Lake Storage Gen2 2

projectworkk (Primary - storage...)

startgen2 (Primary)

(Attached Containers)

Integration datasets 2

startgen2 | SQL script 1

Run | Undo | Publish | Query plan

Connect to Built-in | Use database master

```
1 -- This is auto-generated code
2 SELECT
3     TOP 100 *
4 FROM
5     OPENROWSET(
6         BULK 'https://storagedestinat.dfs.core.windows.net/startgen2/products.csv',
7         FORMAT = 'CSV',
8         PARSE_VERSION = '2.0',
9         HEADER_ROW = TRUE
10    ) AS [result]
11
```

Results | Messages

View | Table | Chart | Export results

Search

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.

Properties

General | Related (0)

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

Synapse live

Validate all

Publish all

Develop

Filter resources by name

SQL scripts

SQL script 1

startgen2

SQL script 1

Run

Undo

Publish

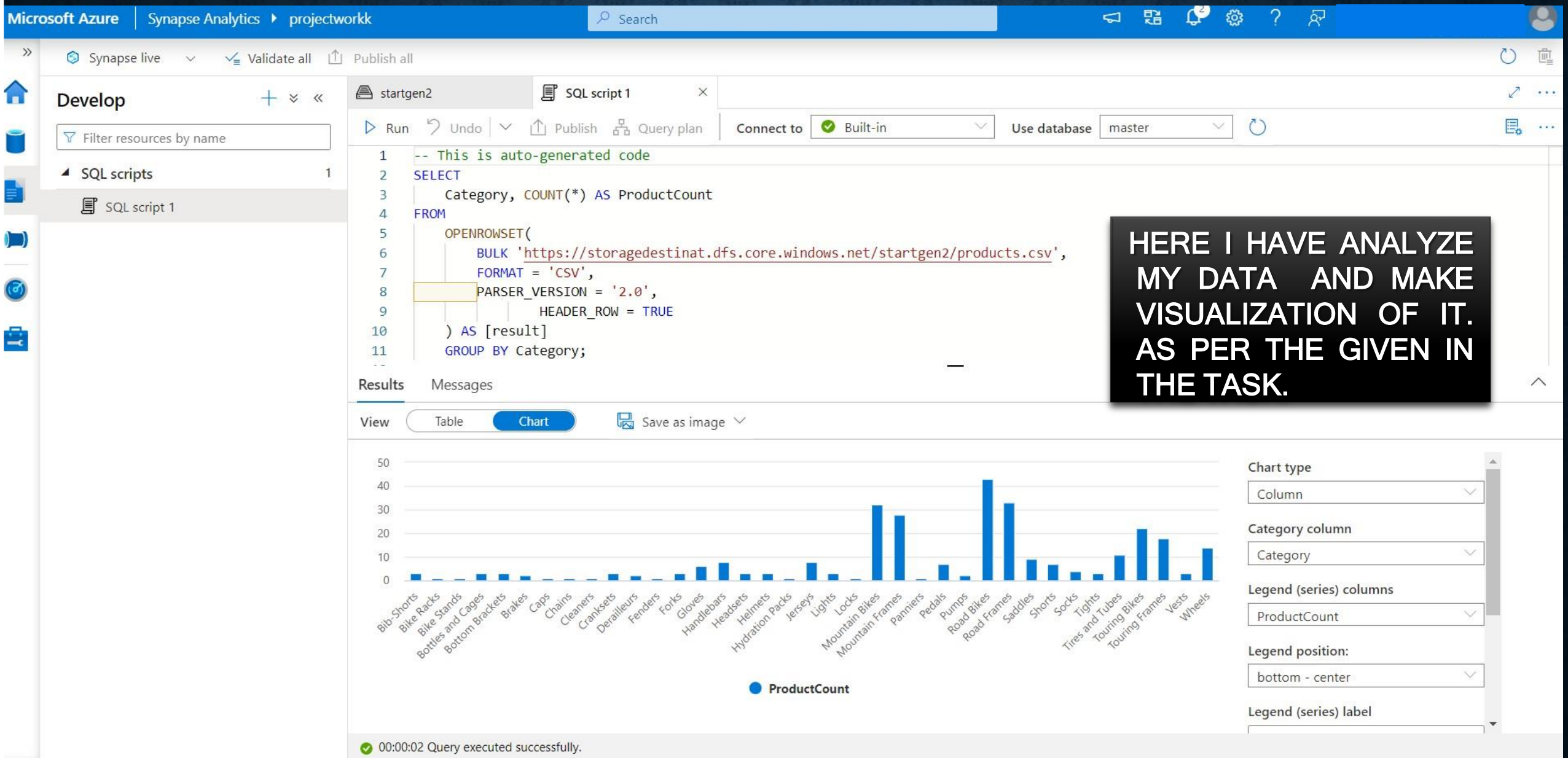
Query plan

Connect to Built-in

Use database master

```
1 -- This is auto-generated code
2 SELECT
3     Category, COUNT(*) AS ProductCount
4 FROM
5     OPENROWSET(
6         BULK 'https://storagedestinat.dfs.core.windows.net/startgen2/products.csv',
7         FORMAT = 'CSV',
8         PARSE_VERSION = '2.0',
9         HEADER_ROW = TRUE
10    ) AS [result]
11 GROUP BY Category;
```

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



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 BIGDATA 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

- 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.
- SPARK POOL: A MANAGED APACHE SPARK SERVICE IN AZURE SYNAPSE

- 🏠
- 🗄️ Analytics pools
- 🗄️ SQL pools
- 🖨️ Apache Spark pools
- 🖨️ Data Explorer pools (pre...
- External connections
- 🔗 Linked services
- 👁️ Microsoft Purview
- Integration
- ⚡ Triggers
- 🔗 Integration runtimes
- Security
- 🔑 Access control
- 👤 Credentials
- 🔒 Managed private endpoi...
- Configurations + libraries
- 📁 Workspace packages
- 📁 Data flow libraries
- 📁 Apache Spark configurat...
- Source control

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](#) 📄

+ New ↻ Refresh

🔍 Filter by name

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.

Microsoft Azure | Synapse Analytics | projectworkk

Search

Synapse live

Validate all

Publish all 3

Data

Workspace

Linked

Filter resources by name

Azure Data Lake Storage Gen2 2

projectworkk (Primary - storage...
startgen2 (Primary)

(Attached Containers)

Integration datasets 2

DestinationDataset_ead

SourceDataset_ead

startgen2

SQL script 1

Notebook 1

Notebook 2

Notebook 3

Run all

Undo

Publish

Outline

Attach to sparkpool

Language PySpark (Python)

Variables

Ready

1 %%pyspark

2 df = spark.read.load('abfss://startgen2@storagedestinat.dfs.core.windows.net/products.csv', format='csv'

3 ## If header exists uncomment line below

4 ##, header=True

5)

6 display(df.limit(10))

[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.

+ Code

+ Markdown

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

startgen2

Container

...

-
- «
- 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: startgen2

☐ Show deleted objects

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state	
<input type="checkbox"/> synapse						-	...
<input type="checkbox"/> 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.

 **startgen2**
Container

Overview

Diagnose and solve problems

Access Control (IAM)

Settings

Shared access tokens

Manage ACL

Access policy

Properties

Metadata

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

Location: startgen2

☒ Show deleted objects

Name	
<input type="checkbox"/>	synapse
<input type="checkbox"/>	products.csv

products.csv
Blob

Save Discard Download Refresh Delete

Overview Versions Edit Generate SAS

	ProductID,ProductName,Category,ListPrice
1	771,"Mountain-100 Silver, 38",Mountain Bikes,3399.9900
2	772,"Mountain-100 Silver, 42",Mountain Bikes,3399.9900
3	773,"Mountain-100 Silver, 44",Mountain Bikes,3399.9900
4	774,"Mountain-100 Silver, 48",Mountain Bikes,3399.9900
5	775,"Mountain-100 Black, 38",Mountain Bikes,3374.9900
6	776,"Mountain-100 Black, 42",Mountain Bikes,3374.9900
7	777,"Mountain-100 Black, 44",Mountain Bikes,3374.9900
8	778,"Mountain-100 Black, 48",Mountain Bikes,3374.9900
9	779,"Mountain-200 Silver, 38",Mountain Bikes,2319.9900
10	780,"Mountain-200 Silver, 42",Mountain Bikes,2319.9900
11	781,"Mountain-200 Silver, 46",Mountain Bikes,2319.9900
12	782,"Mountain-200 Black, 38",Mountain Bikes,2294.9900
13	783,"Mountain-200 Black, 42",Mountain Bikes,2294.9900
14	784,"Mountain-200 Black, 46",Mountain Bikes,2294.9900
15	785,"Mountain-300 Black, 38",Mountain Bikes,1079.9900
16	786,"Mountain-300 Black, 40",Mountain Bikes,1079.9900
17	787,"Mountain-300 Black, 44",Mountain Bikes,1079.9900
18	788,"Mountain-300 Black, 48",Mountain Bikes,1079.9900
19	980,"Mountain-400-W Silver, 38",Mountain Bikes,769.4900
20	981,"Mountain-400-W Silver, 40",Mountain Bikes,769.4900
21	

Csv

Preview

This is the preview of the stored data.

THANK YOU