

---

# Capstone Project

INTEGRATED ANALYTICS WITH AZURE  
SYNAPSE

Shashank S,  
Batch :DS19M

---

# Project abstract

Exploring data analytics workspace by using Azure Synapse Analytics and using Azure Synapse Analytics Workspace as a resource in Azure Synapse Analytics for Integrated Data Analytics.

Finding out solution for Three problem statements which included,

- Ingesting dataset from an external resource, transforming the dataset and storing the transformed data into ADLS Gen-2(Azure Data-lake Storage Generation 2) storage.
- Exploring datahub and performing SQL queries on top of transformed data and visualize the result set.
- Creating a Pyspark environment(Apache Spark Pool) and performing operations on the same set of data.

# Project Procedure

1. Exploring data analytics workspace by using Azure synapse analytics.
2. Creating a ADLS Gen 2 account in the Azure synapse.
3. By using Built-in Copy task option Ingesting the data from a source through HTML link.
4. After ingestion, creating a connection and transforming the data and load the transformed data into ADLS gen storage.

Problem statement 1

1. Analyse and query the data which has been ingested in azure synapse workspace.
2. Through SQL query performing AGGREGATION and GROUPING the data as per the requirement
3. Analysing the result set through Chart view.

Problem statement 2

1. Creating a Apache spark pool with given inputs.
2. Loading the data into dataframe by using built-in code.
3. Analysing the data in spark pool by setting the language as pyspark.
4. Performing the same query and getting insights from the help of chart view.

Problem statement 3

# Project Outcomes

1. **Created the Azure synapse workspace, Ingested the data choosing HTTP and Transformed the data as per the requirements and stored it in ADLS gen2 storage.**
2. **SQL and SPARK pool has been used to analyse the data.**
3. **With the help of chart view getting visual information and proper insights about the data.**

—

# SCREENSHOTS

Cancel

# PIPELINE creation

## 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.

Finish

Edit pipeline

Monitor



Type here to search



11:59 AM  
13-04-2023

3

# View of the UPDATED QUERY

DAZ1002: Capstone Project

shashankcapstone - Azure Synap

fileadlsgen2 - Microsoft Azure

+

web.azuresynapse.net/en/authoring/explore/linked/sqlscripts/SQL%20script%201?workspace=%2Fsubscriptions%2F522db352-b560-4a36-9aa9-e36d03d5c4ee%2FresourceGroups%2Fca...

Microsoft Azure | shashankcapstone

Search

Synapse live

Validate all

Publish all 1

»

Home

Data

Workspace

Linked

Filter resources by name

Azure Data Lake Storage Gen2 4

shashankcapstone (Primary - ca...

adls2Connection (capstoneadls...

(Attached Containers)

AzureDataLakeStorageCONNEC...

fileadlsgen2

Integration datasets 6

fileadlsgen2

SQL script 1

Run

Undo

Publish

Query plan

Connect to Built-in

Use database master

Refresh

More

```
1 -- This is auto-generated code
2 SELECT
3     category,count(productid)as productnumbers
4 FROM
5     OPENROWSET(
6         BULK 'https://capstoneadlsgen2.dfs.core.windows.net/fileadlsgen2/product.csv',
7         FORMAT = 'CSV',
8         PARSER_VERSION = '2.0',
9         HEADER_ROW=True
10    ) AS [result]
11 GROUP BY category;
12
```

Results

Messages

View

Table

Chart

Export results

Search

category	productnumbers
Bib-Shorts	3
Bike Racks	1
Bike Stands	1

00:00:03 Query executed successfully.

Properties

General

Related (0)

Name \*

SQL script 1

Description

Type

.sql script

Size

236 bytes

Results settings per query

☒ First 5000 rows (default)

☐ All rows

Type here to search

12:34 PM

13-04-2023



# CHART VIEW of the updated query

DAZ1002: Capstone Project

shashankcapstone - Azure Synap

fileadlsgen2 - Microsoft Azure

+

web.azure.synapse.net/en/authoring/analyze/sqlscripts/Count%20Products%20by%20category?workspace=%2Fsubscriptions%2F522db352-b560-4a36-9aa9-e36d03d5c4ee%2Fresource...

Microsoft Azure | shashankcapstone

Search

Synapse live

Validate all

Publish all

Develop

Filter resources by name

SQL scripts

Count Products by category

fileadlsgen2

Count Products by c...

Run

Undo

Publish

Query plan

Connect to Built-in

Use database master

1

-- This is auto-generated code

2 SELECT

3 category, count(productid) as productnumbers

4 FROM

5 OPENROWSET(

6 BULK 'https://capstoneadlsgen2.dfs.core.windows.net/fileadlsgen2/product.csv',

7 FORMAT = 'CSV',

8 PARSER\_VERSION = '2.0',

9 HEADER\_ROW=True

10 ) AS [result]

11 GROUP BY category;

Results

Messages

View

Table

Chart

Save as image

Category	Product Numbers
Bib-Shorts	2
Bike Stands	2
Bottom Brackets	2
Caps	2
Cleaners	2
Derailleurs	2
Forks	2
Handlebars	2
Helmets	2
Jerseys	2
Locks	2
Mountain Frames	30
Pedals	2
Road Bikes	2
Saddles	45
Socks	2
Tires and Tubes	2
Touring Frames	2
Wheels	2

Chart type

Column

Category column

category

Legend (series) columns

productnumbers

Legend position:

bottom - center

Legend (series) label

00:00:03 Query executed successfully.

Properties

General

Related (0)

Name \*

Count Products by category

Description

Type

.sql script

Size

236 bytes

Results settings per query

First 5000 rows (default)

All rows

# Resource creation - MYAPACHE POOL

DAZ1002: Capstone Project

shashankcapstone - Azure Synap

fileadlsgen2 - Microsoft Azure

+

web.azure.synapse.net/en/management/apachesparkpools?workspace=%2Fsubscriptions%2F522db352-b560-4a36-9aa9-e36d03d5c4ee%2FresourceGroups%2Fcp-rc%2Fproviders%2FM...

Microsoft Azure | Synapse Analytics > shashankcapstone

Synapse live Validate all Publish all

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 configurati

New Refresh

Filter by name

Showing 1-1 of 1 item

Name

myapachepool

Properties

Name

myapachepool

URL

/subscriptions/522db352-b560-4a36-9aa9-e36d03d5c4ee/resourceGroups/cp-rc/provide...

Creation date

04/13/2023, 12:50:32 PM

Configuration

Node size family

Memory Optimized

Node size

Small (4 vCores / 32 GB)

Autoscale

Enabled

Number of nodes(min/max)

3 to 4 nodes

Automatic pausing

Enabled

Number of minutes idle

15

Dynamically allocate executors

Disabled

Close

# Query runned in APACHE SPARK POOL

DAZ1002: Capstone Project

shashankcapstone - Azure Synap

fileadlsgen2 - Microsoft Azure

+

web.azure.synapse.net/en/authoring/explore/linked/notebooks/Notebook%201?workspace=%2Fsubscriptions%2F522db352-b560-4a36-9aa9-e36d03d5c4ee%2FresourceGroups%2Fcp-r...

Microsoft Azure | shashankcapstone

Search

shashanksshahi6@gmail.com

DEFAULT DIRECTORY

Synapse live

Validate all

Publish all

Data

Workspace

Linked

Filter resources by name

Azure Data Lake Storage Gen2

shashankcapstone (Primary - ca...)

fileadlsgen2 (Primary)

adls2Connection (capstoneadls...)

(Attached Containers)

AzureDataLakeStorageCONNEC...

Integration datasets

fileadlsgen2

Notebook 1

Run all

Undo

Publish

Outline

Attach to

myapachepool

Language

PySpark (Python)

Variables

Ready

2 sec - Command executed in 2 sec 21 ms by shashanksshahi6 on 12:55:39 PM, 4/13/23

View

Table

Chart

Export results

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

Type here to search

12:55 PM

13-04-2023

# CHART view for the same sql query

Microsoft Azure | Synapse Analytics | shashankcapstone

Develop

Filter resources by name

Notebooks

Notebook 1

fileadlsgen2

Notebook 1

Run all

Undo

Publish

Outline

Attach to: myapachepool

Language: PySpark (Python)

Variables

Ready

```
4 display(df)
```

[6] ✓ 1 sec - Command executed in 1 sec 178 ms by shashanksshahi6 on 1:13:16 PM, 4/13/23

Job execution Succeeded Spark 2 executors 8 cores

View in monitoring Open Spark UI

View: Table Chart Export results

Category: Count(ProductID)

Chart type: Column chart

Key: Category

Values: ProductID

Series Group: Category

Aggregation: Count

☐ Stacked

☐ Aggregating over all results

Apply Cancel

1/2

Type here to search

01:13 PM 13-04-2023

---

**THANK YOU**