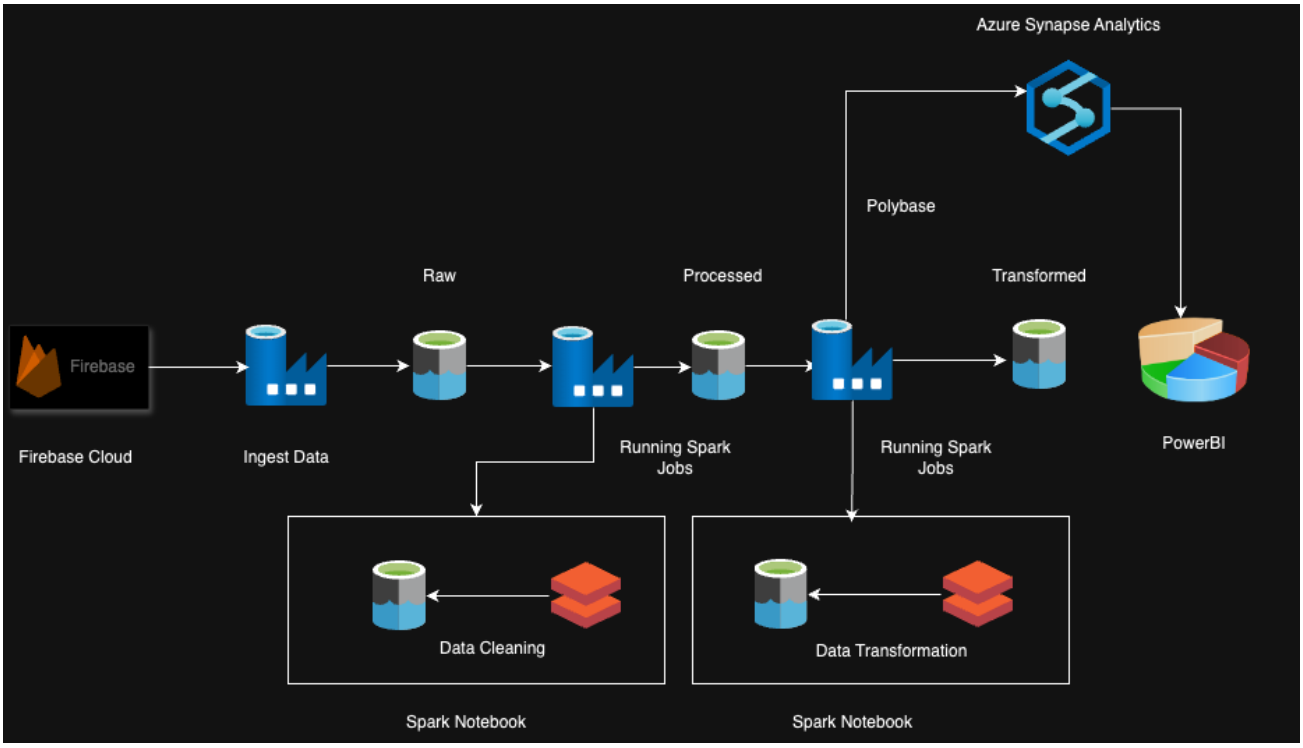
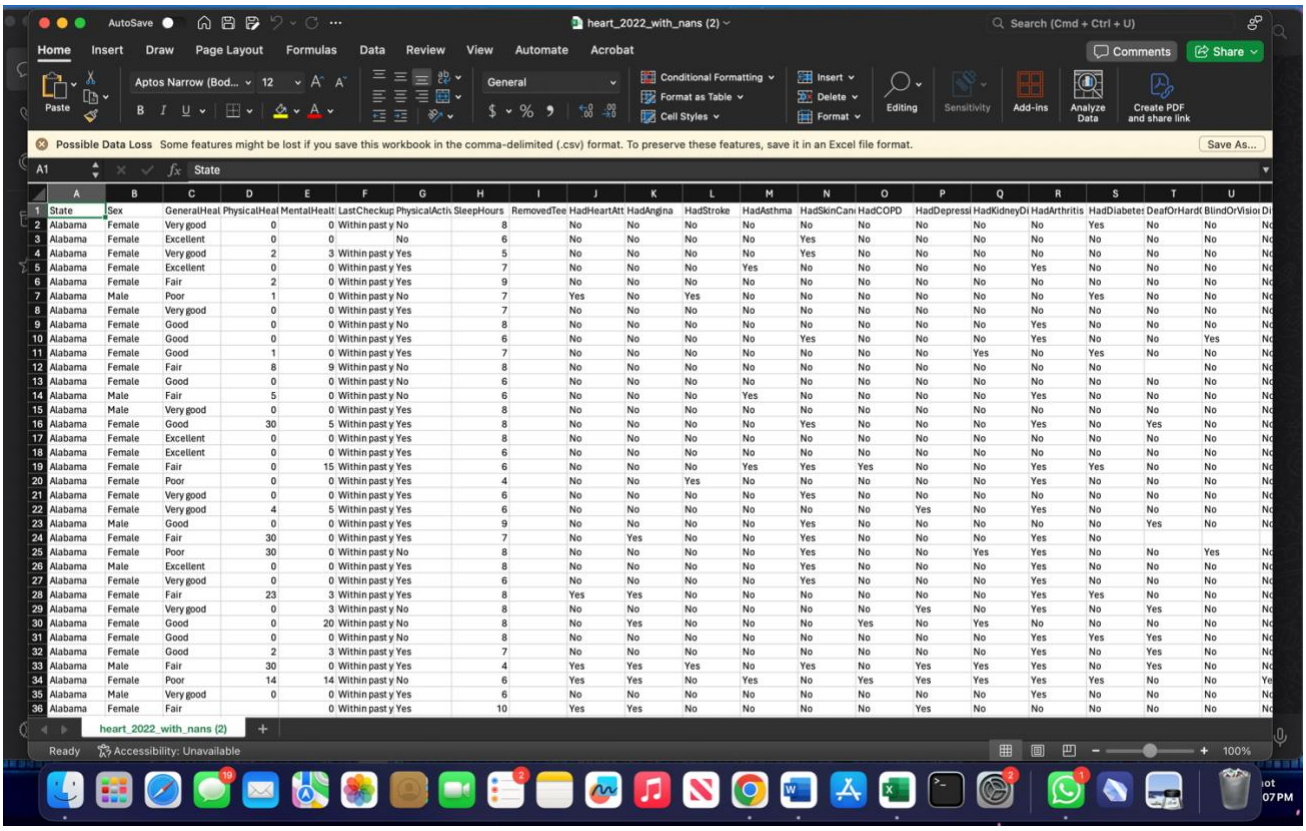
PROJECT – HEALTH DATA ANALYSIS

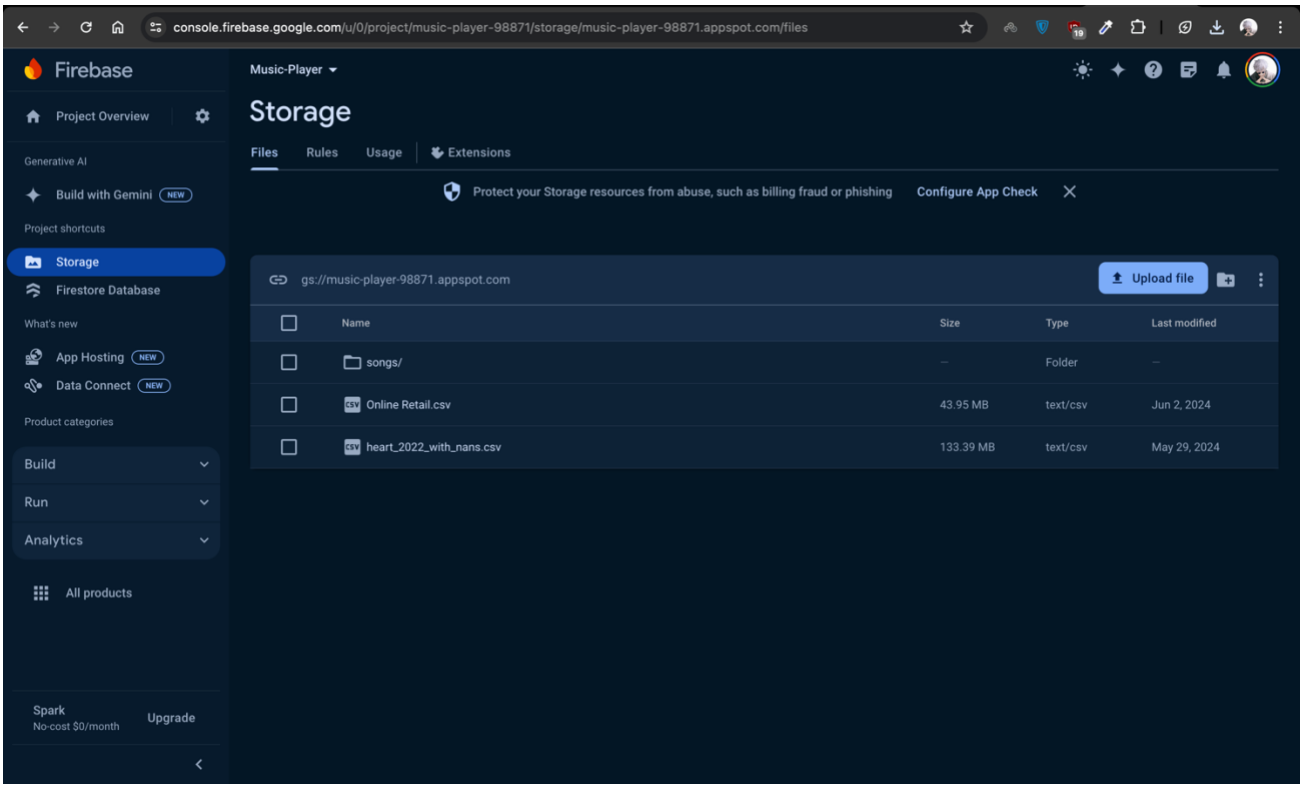
Architecture Diagram



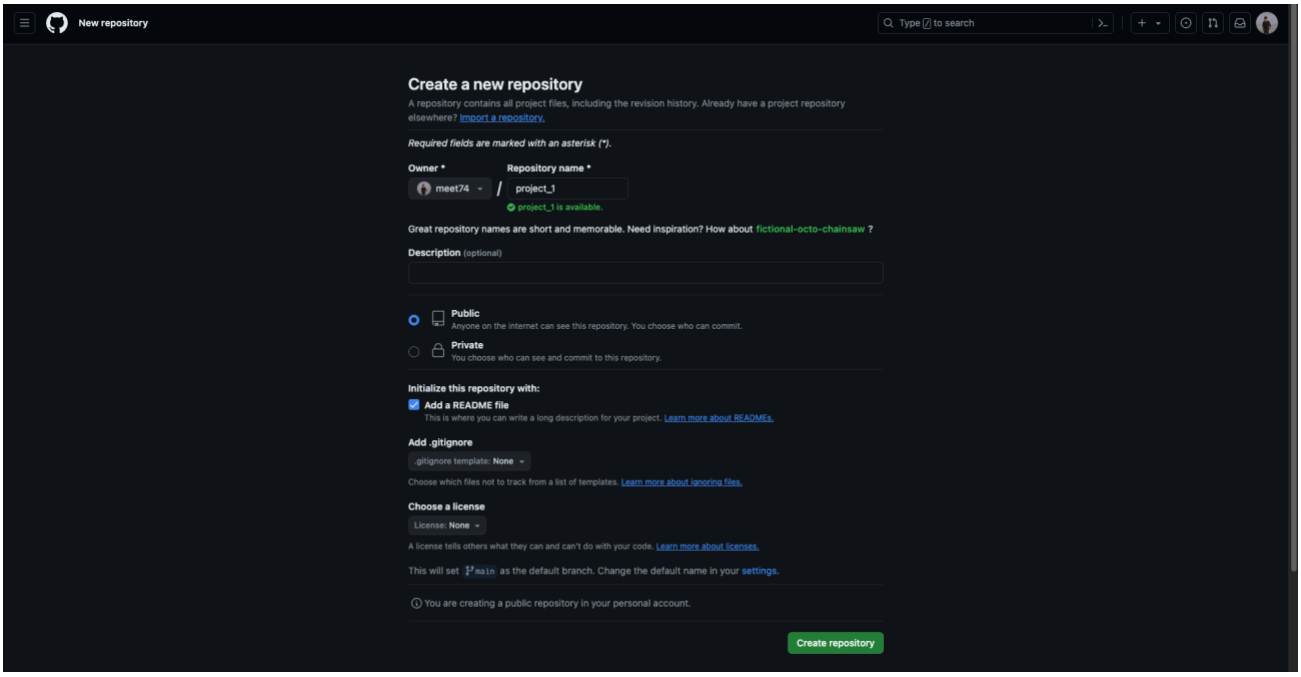
Raw Data of Health Dataset



Data in Firebase cloud

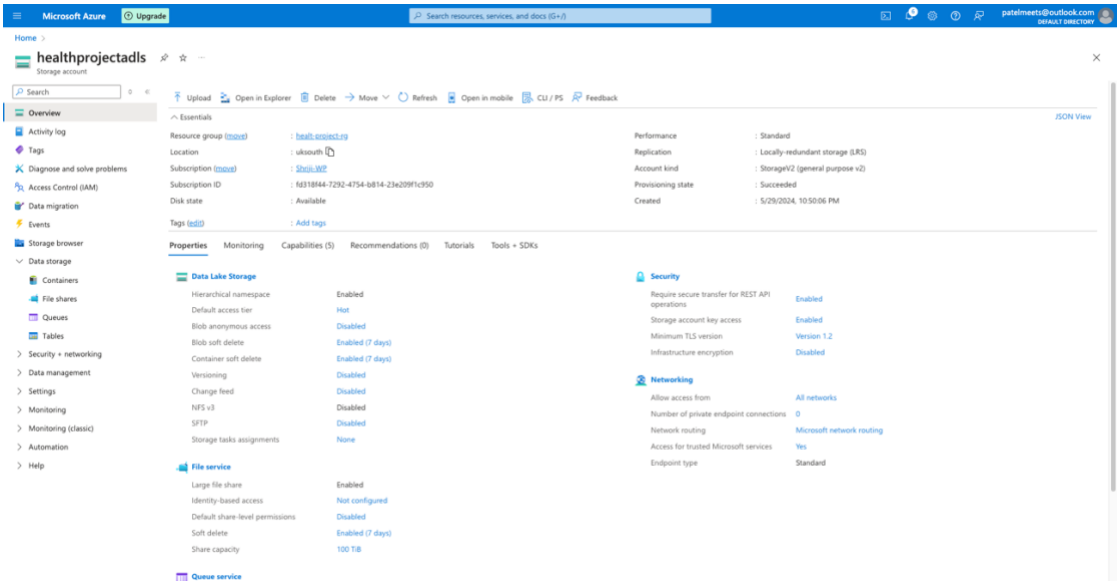


Git Repo

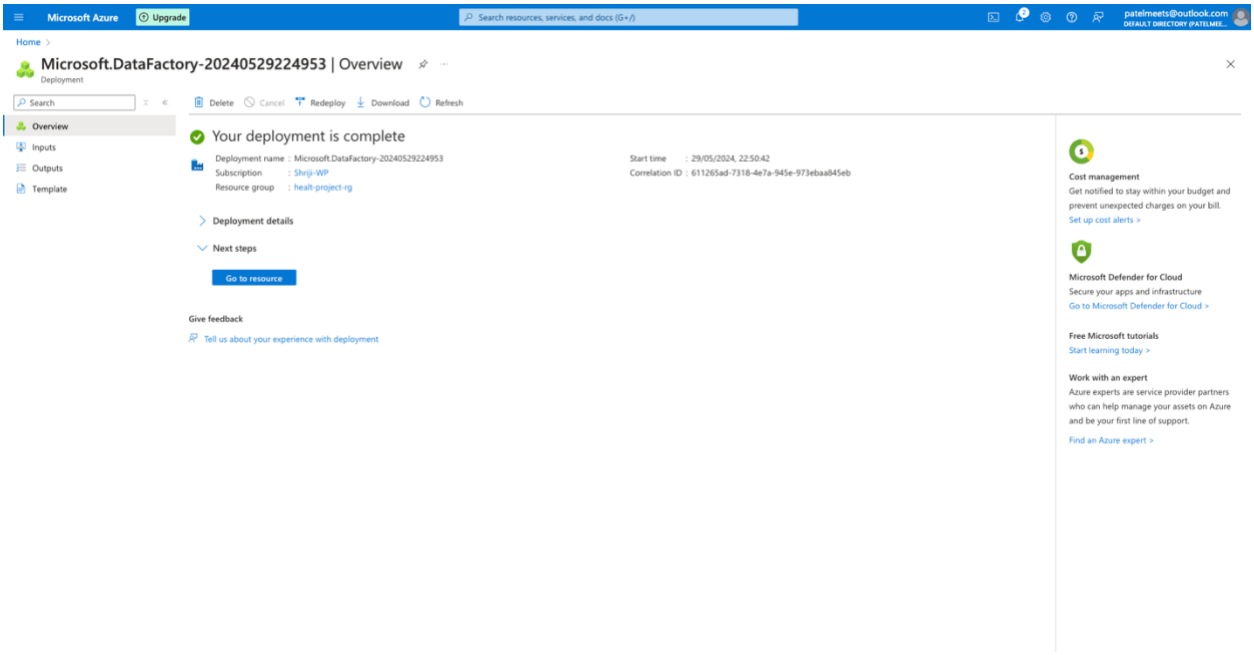


Azure Data Lake Storage and Azure Data Factory

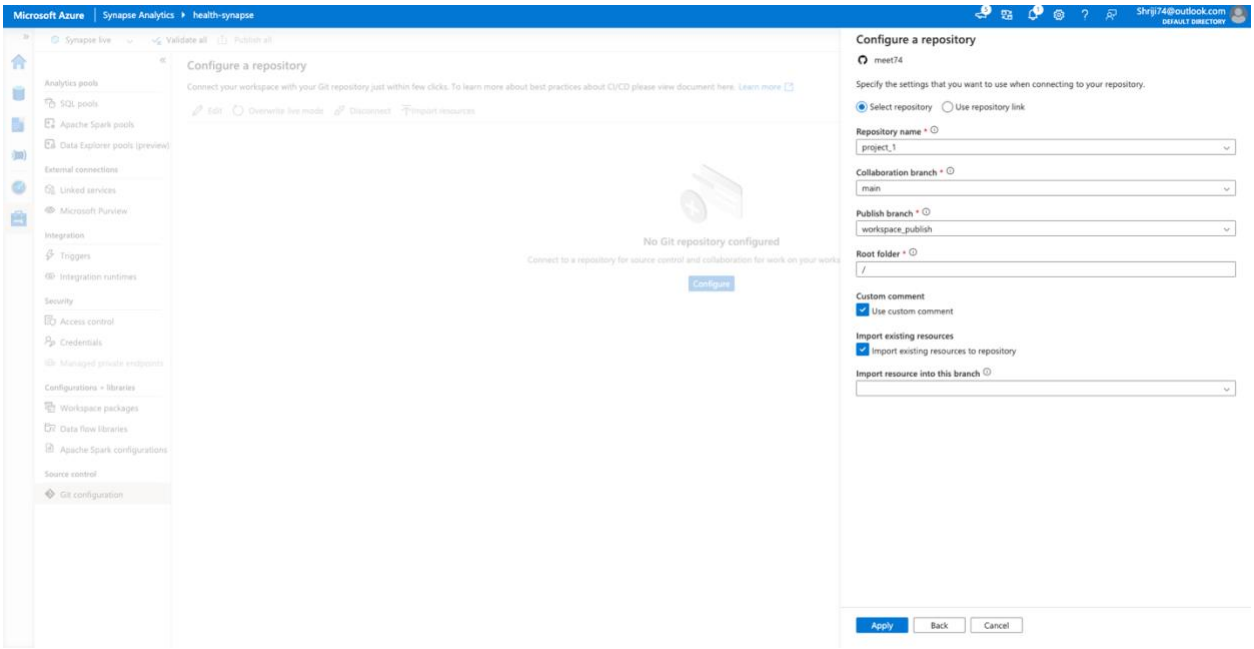
Creating Azure Data Lake Storage

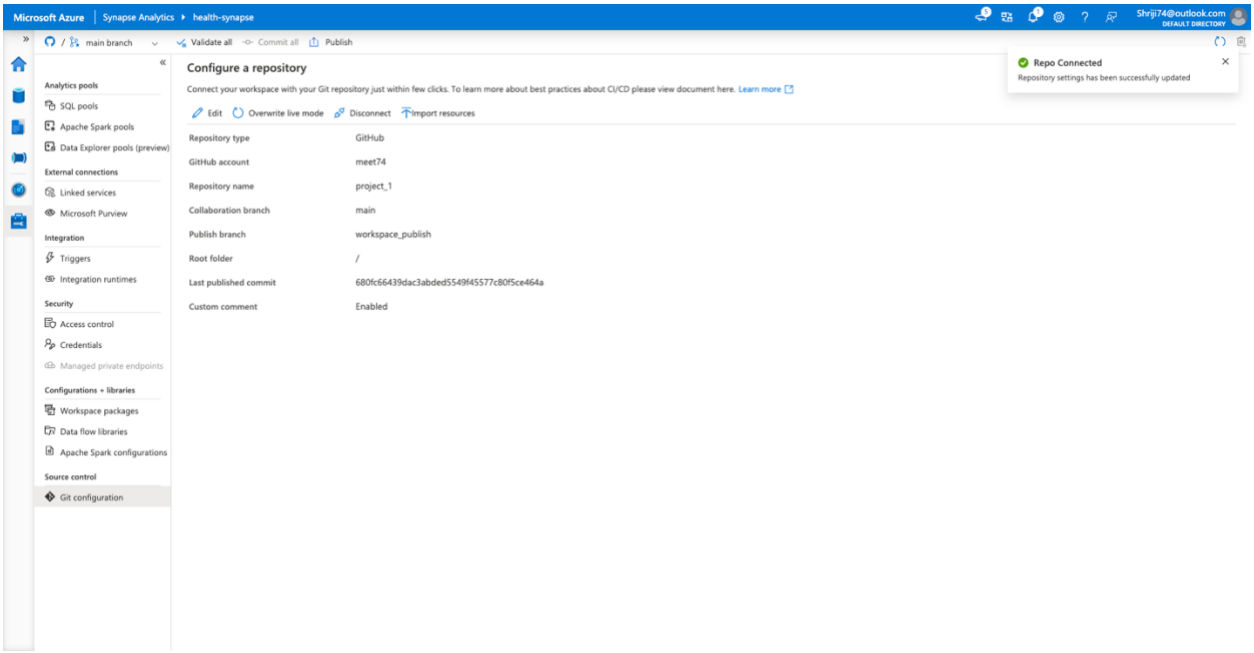


Creating Azure Data Factory



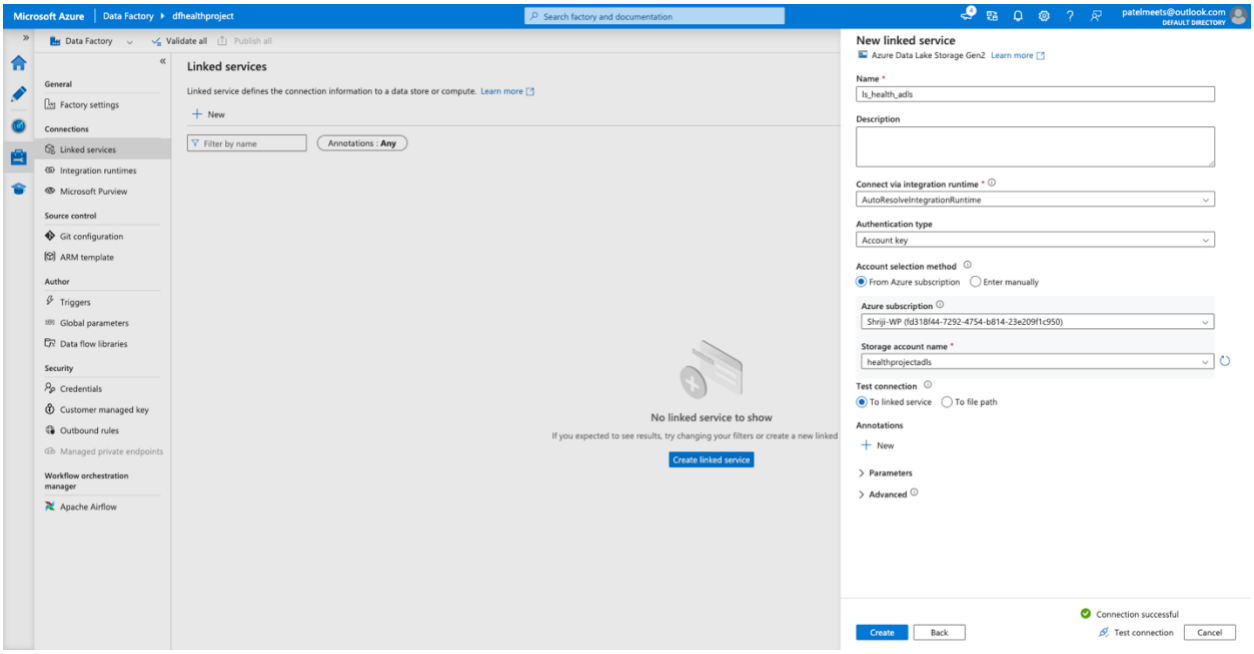
Configuring GitHub in ADF

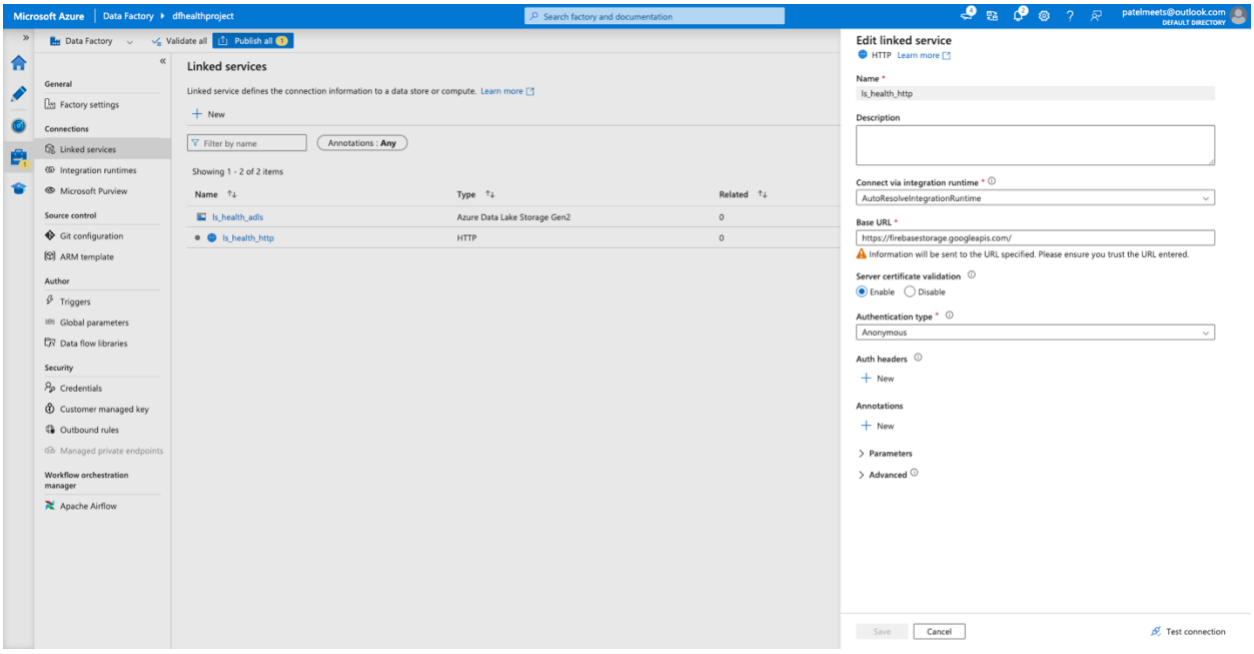




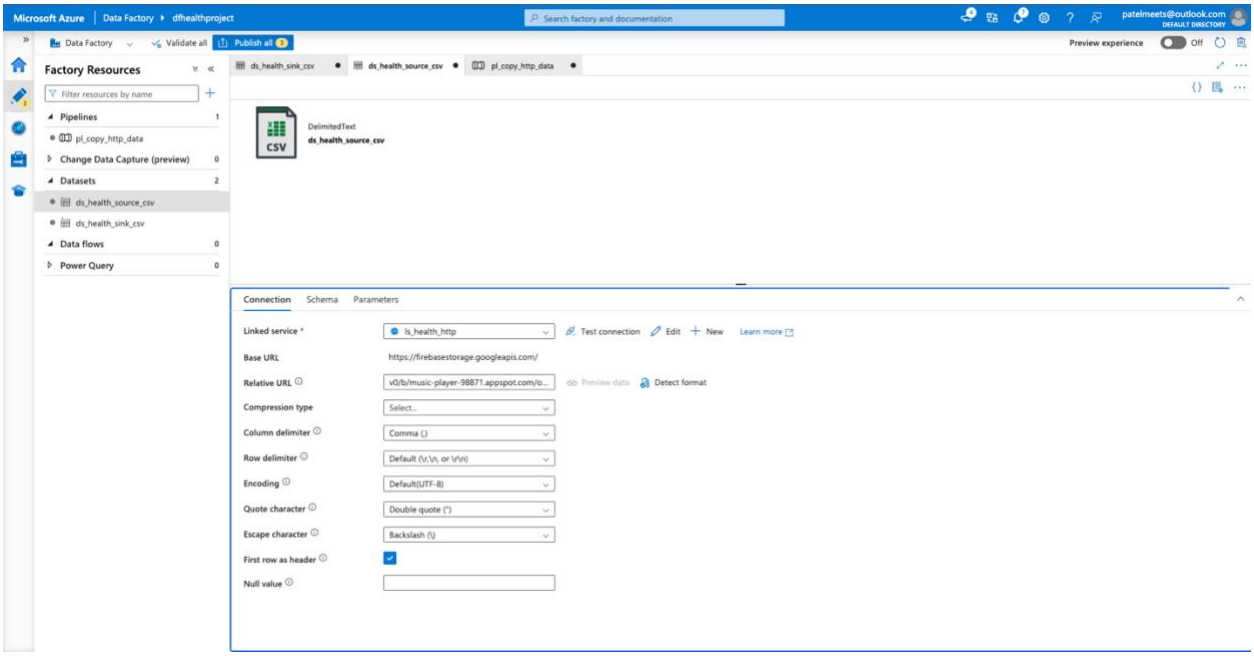
S

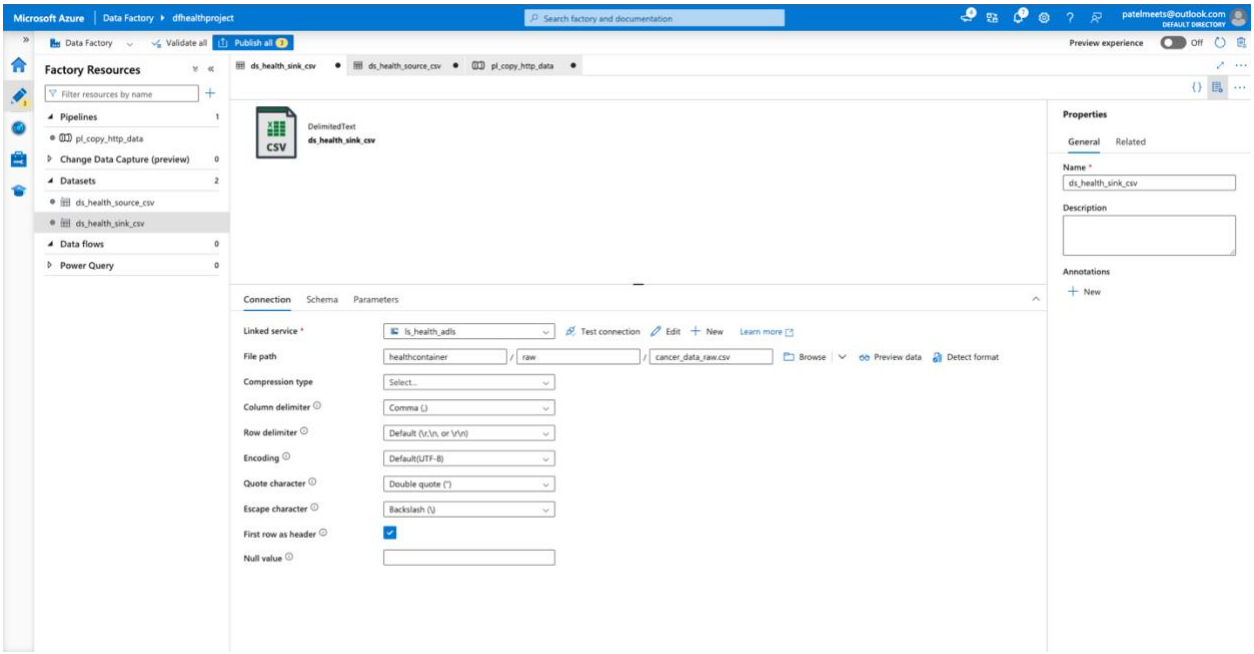
Creating Linked Service of HTTP and ADLS Service in ADF



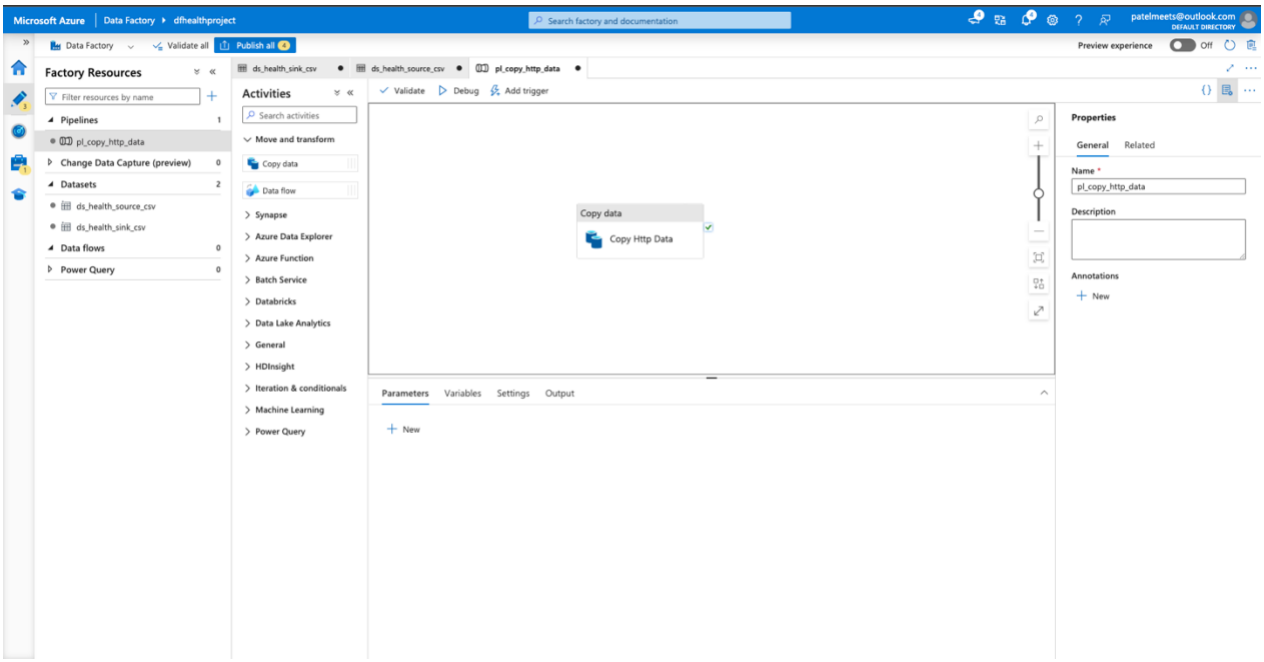


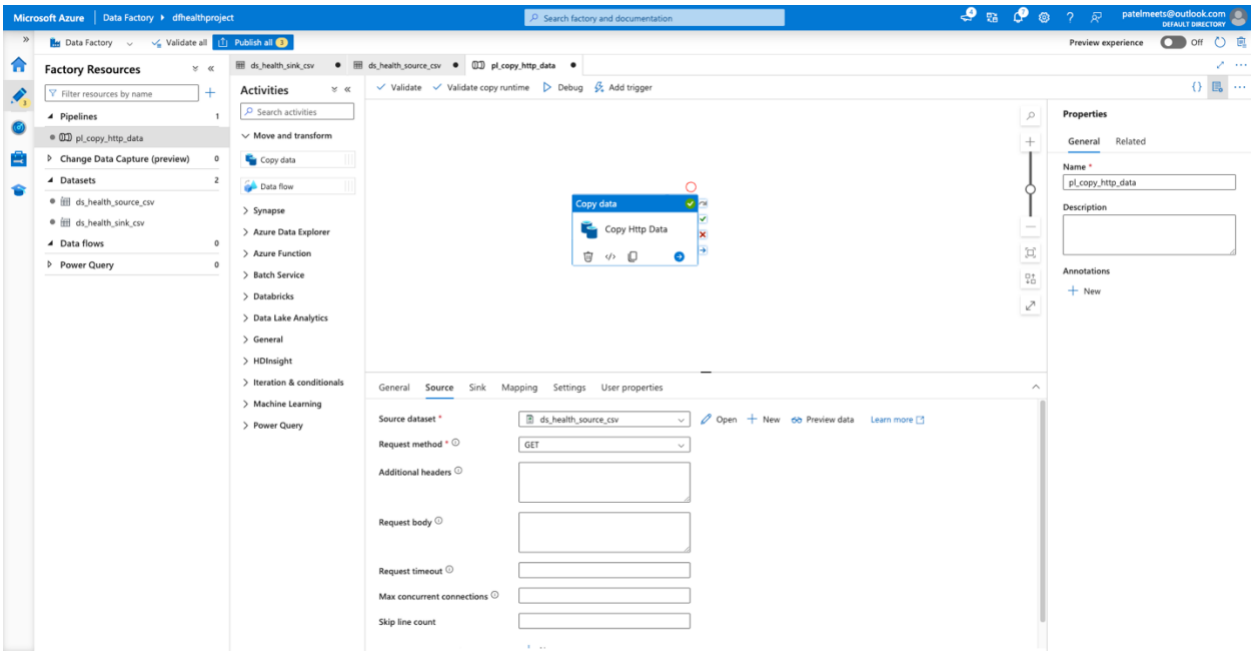
Creating Dataset for HTTP input and ADLS output



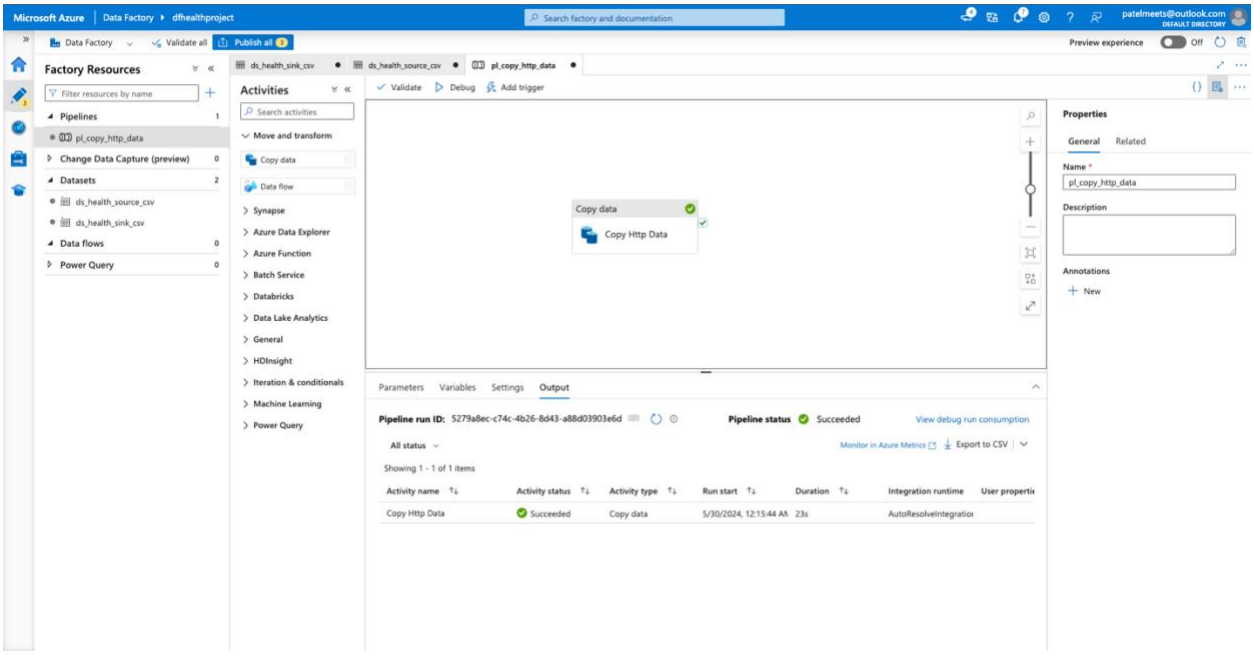


Creating a Pipeline to transfer data from Firebase Cloud to ADLS using ADF

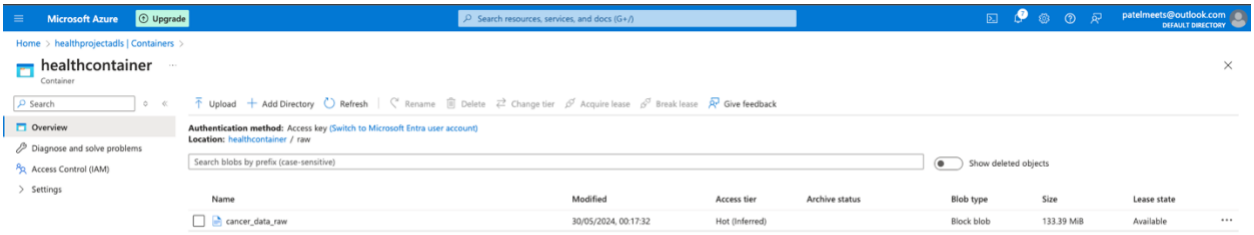

Microsoft Azure
Data Factory
dfhealthproject
Search factory and documentation
patelmeets@outlook.com
DEFAULT DIRECTORY
Data Factory
Validate all
Publish all 4
Preview experience
Off
Factory Resources
Filter resources by name
Pipelines 1
pl_copy_http_data
Change Data Capture (preview) 0
Datasets 2
ds_health_source_csv
ds_health_sink_csv
Data flows 0
Power Query 0
ds_health_sink_csv
ds_health_source_csv
pl_copy_http_data
Activities
Search activities
Move and transform
Copy data
Data flow
Synapse
Azure Data Explorer
Azure Function
Batch Service
Databricks
Data Lake Analytics
General
HDInsight
Iteration & conditionals
Machine Learning
Power Query
Validate
Debug
Add trigger
Copy data
Copy Http Data
Properties
General
Related
Name *
pl_copy_http_data
Description
Annotations
New
Parameters
Variables
Settings
Output
New


Microsoft Azure
Data Factory
dfhealthproject
Search factory and documentation
patelmeets@outlook.com
DEFAULT DIRECTORY
Data Factory
Validate all
Publish all 3
Preview experience
Off
Factory Resources
Filter resources by name
Pipelines 1
pl_copy_http_data
Change Data Capture (preview) 0
Datasets 2
ds_health_source_csv
ds_health_sink_csv
Data flows 0
Power Query 0
ds_health_sink_csv
ds_health_source_csv
pl_copy_http_data
Activities
Search activities
Move and transform
Copy data
Data flow
Synapse
Azure Data Explorer
Azure Function
Batch Service
Databricks
Data Lake Analytics
General
HDInsight
Iteration & conditionals
Machine Learning
Power Query
Validate
Validate copy runtime
Debug
Add trigger
Copy data
Copy Http Data
Properties
General
Related
Name *
pl_copy_http_data
Description
Annotations
New
General
Source
Sink
Mapping
Settings
User properties
Source dataset *
ds_health_source_csv
Open
New
Preview data
Learn more
Request method *
GET
Additional headers
Request body
Request timeout
Max concurrent connections
Skip line count

Pipeline Succeeded

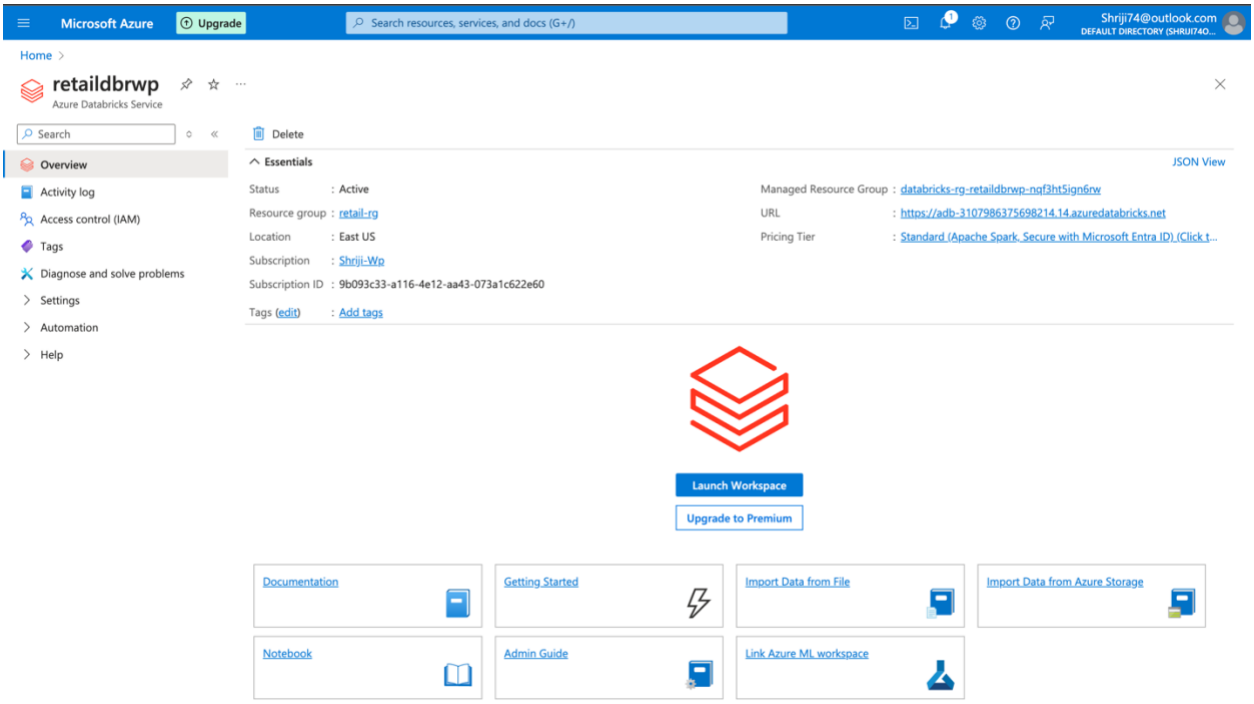


Transferred Data in ADLS

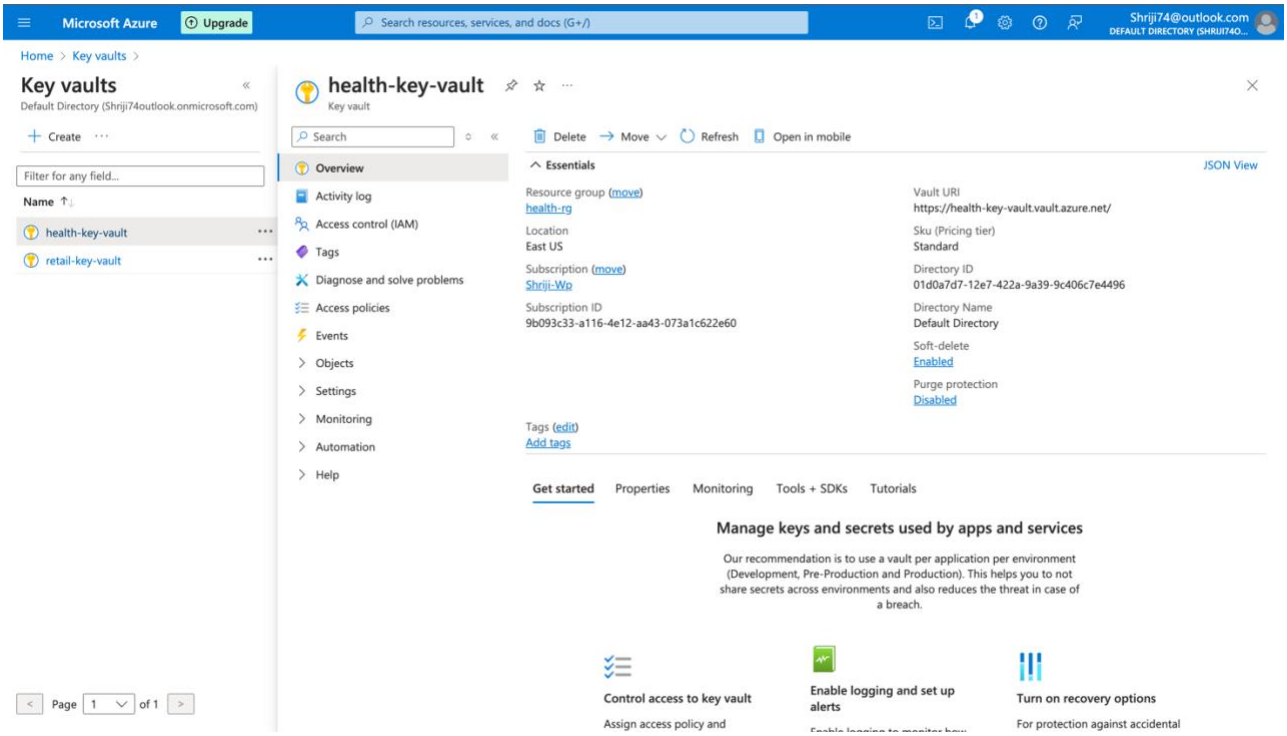

Microsoft Azure
Upgrade
Search resources, services, and docs (G+/)
patelmeets@outlook.com
DEFAULT DIRECTORY
Home > healthprojectadls | Containers >
healthcontainer
Container
Search
Overview
Diagnose and solve problems
Access Control (IAM)
Settings
Upload
Add Directory
Refresh
Rename
Delete
Change tier
Acquire lease
Break lease
Give feedback
Authentication method: Access key (Switch to Microsoft Entra user account)
Location: healthcontainer / raw
Search blobs by prefix (case-sensitive)
Show deleted objects
Name
Modified
Access tier
Archive status
Blob type
Size
Lease state
cancer_data_raw
30/05/2024, 00:17:32
Hot (Inferred)
Block blob
133.39 MiB
Available

Databricks

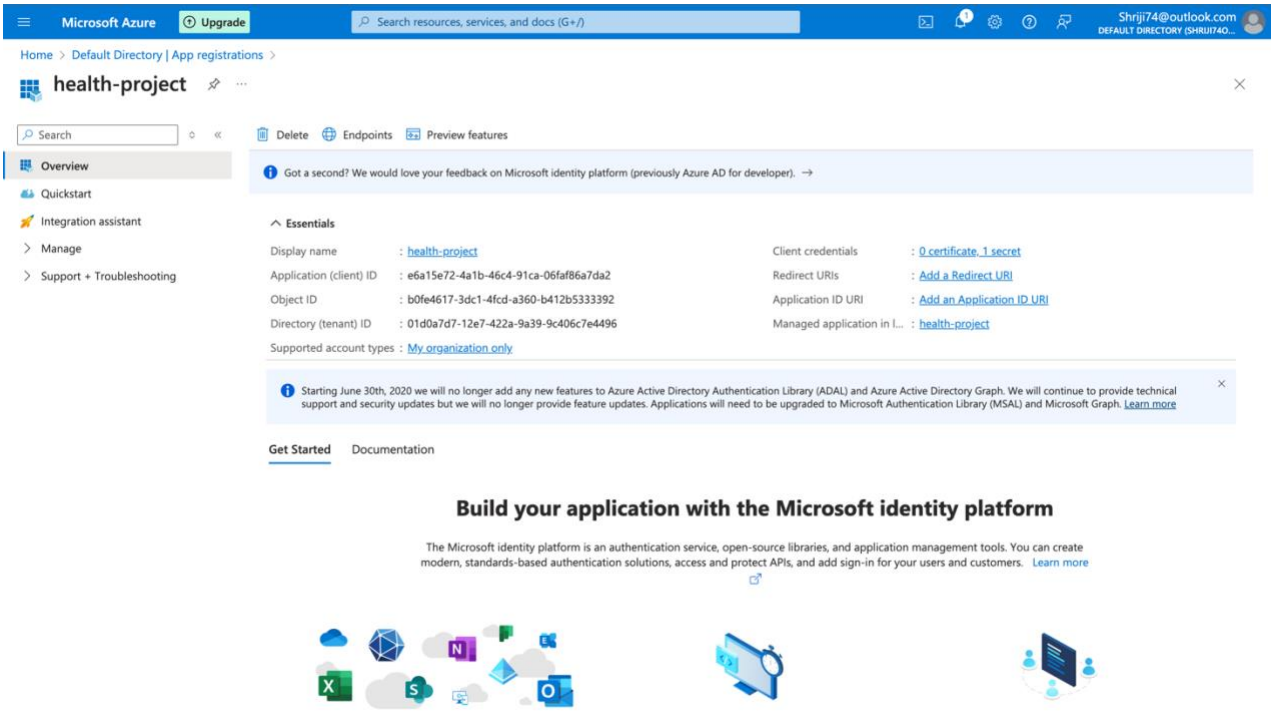
Creating Databricks Workspace



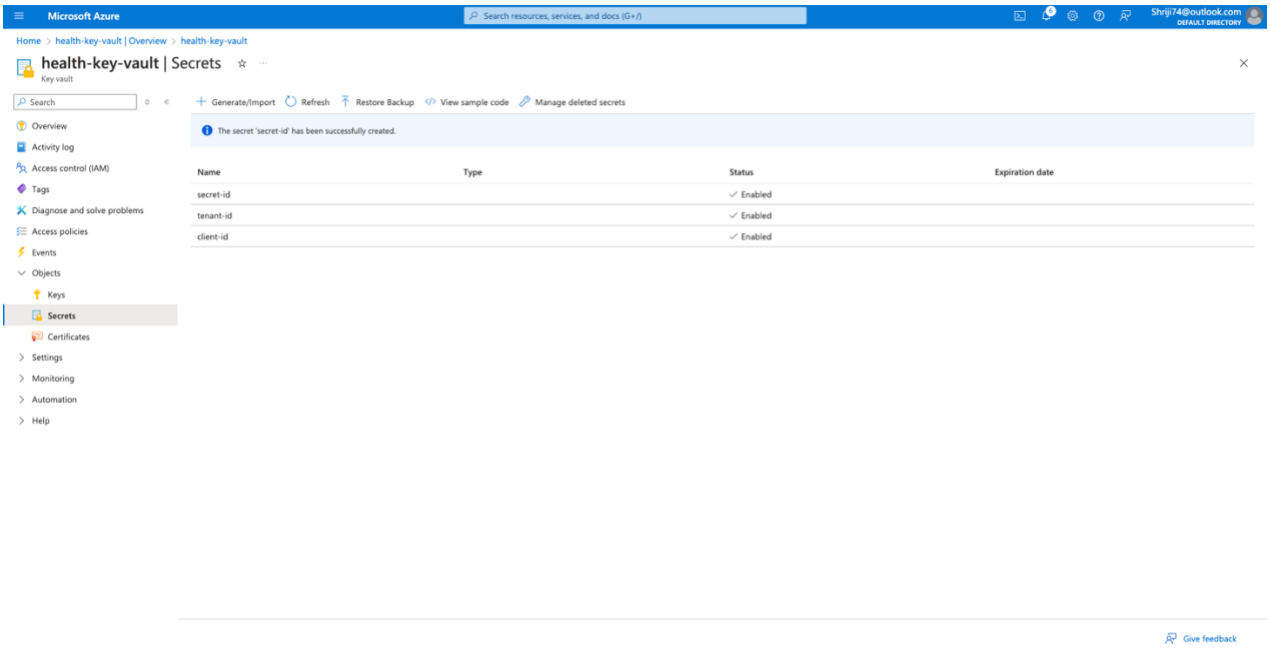
Creating Key vault



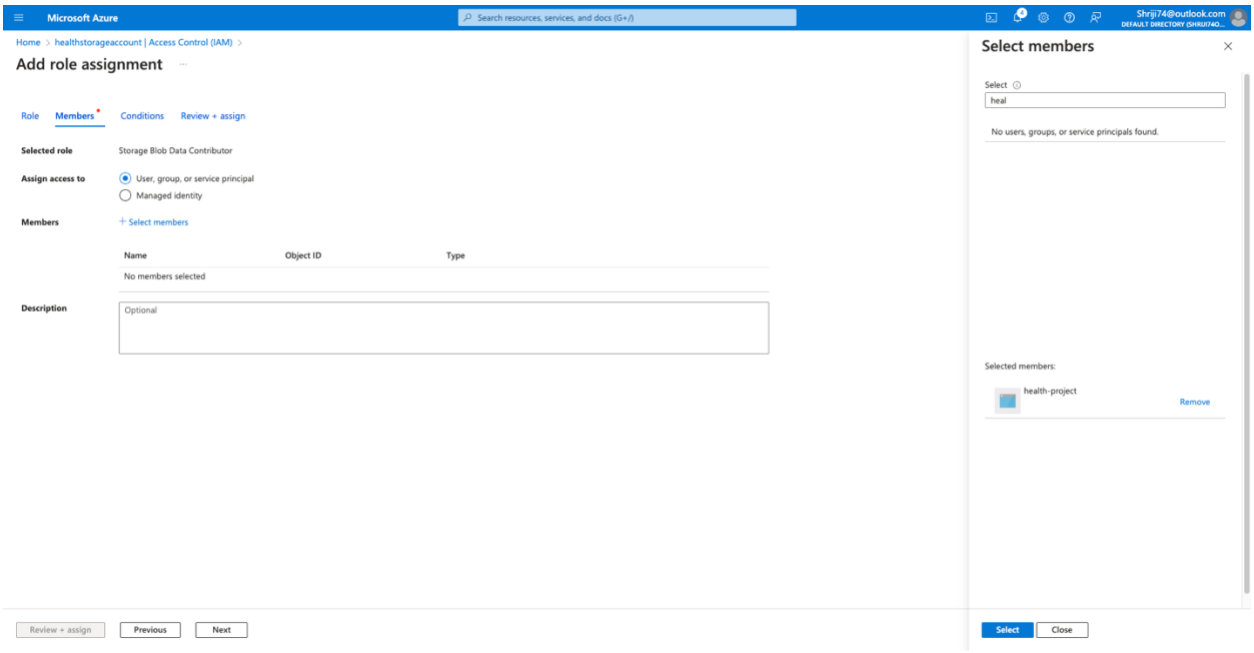
Creating App in Microsoft Intra ID



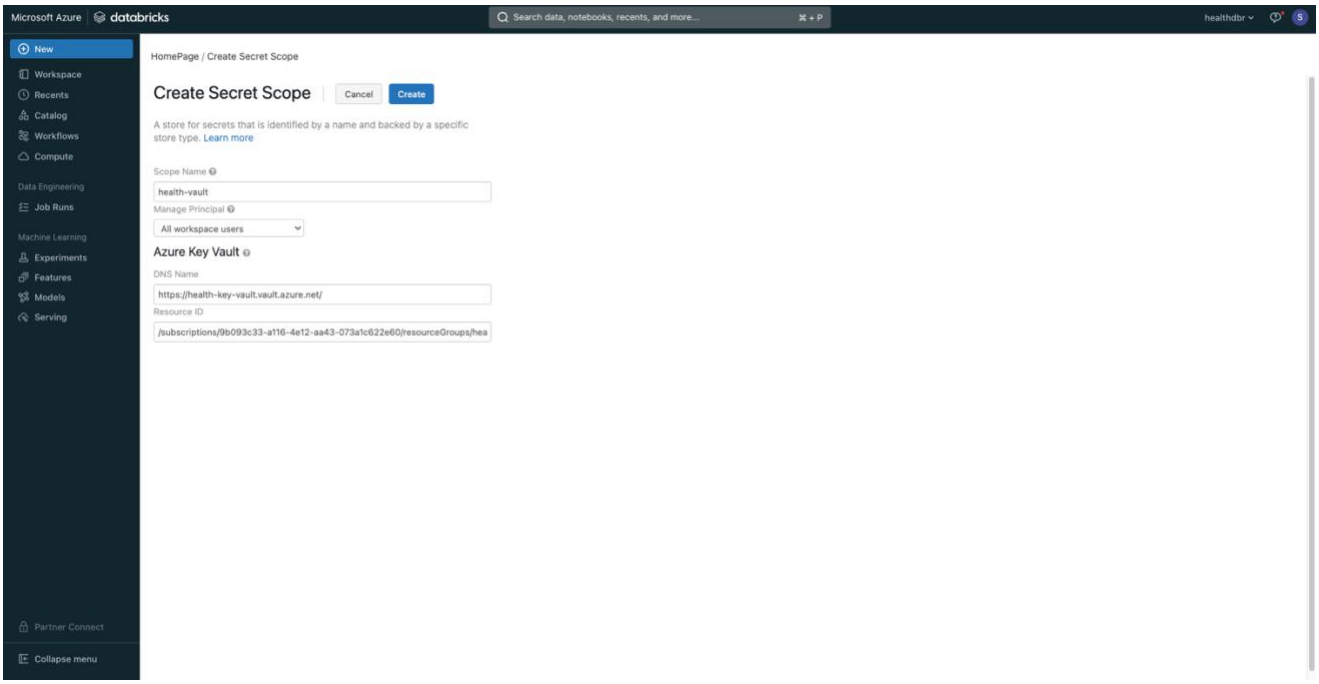
Creating secrets in Key Vault



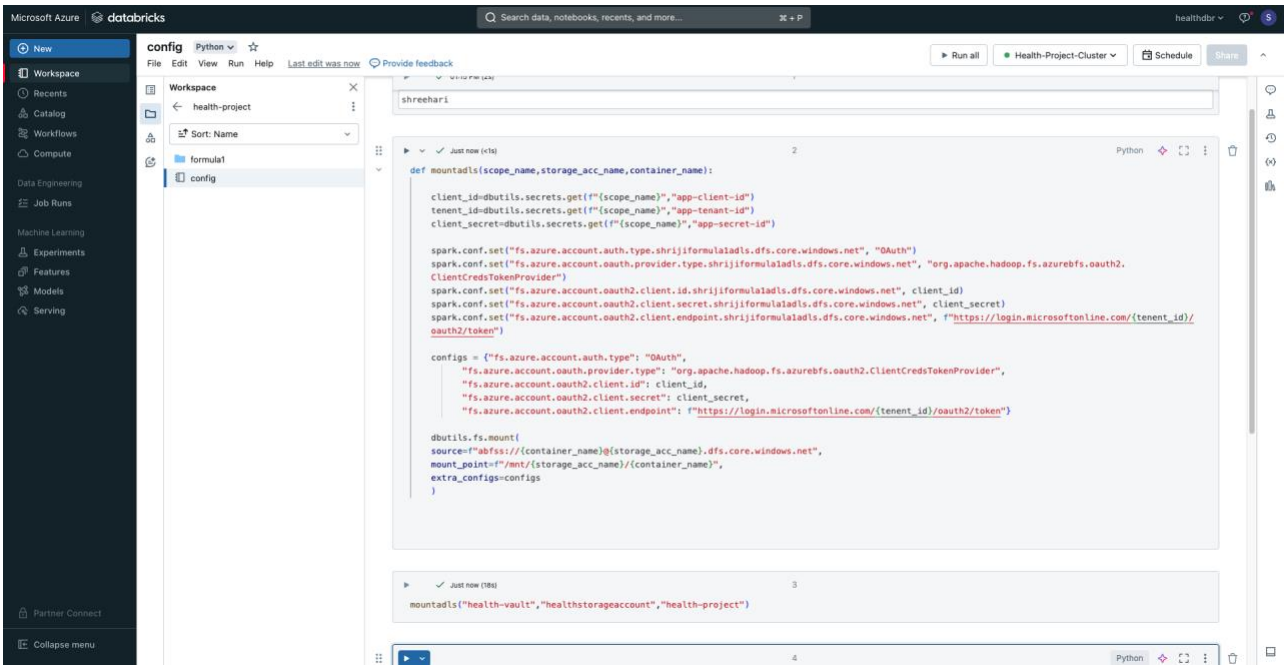
Giving RBAC Permission of Storage Blob Contributor to ADLS



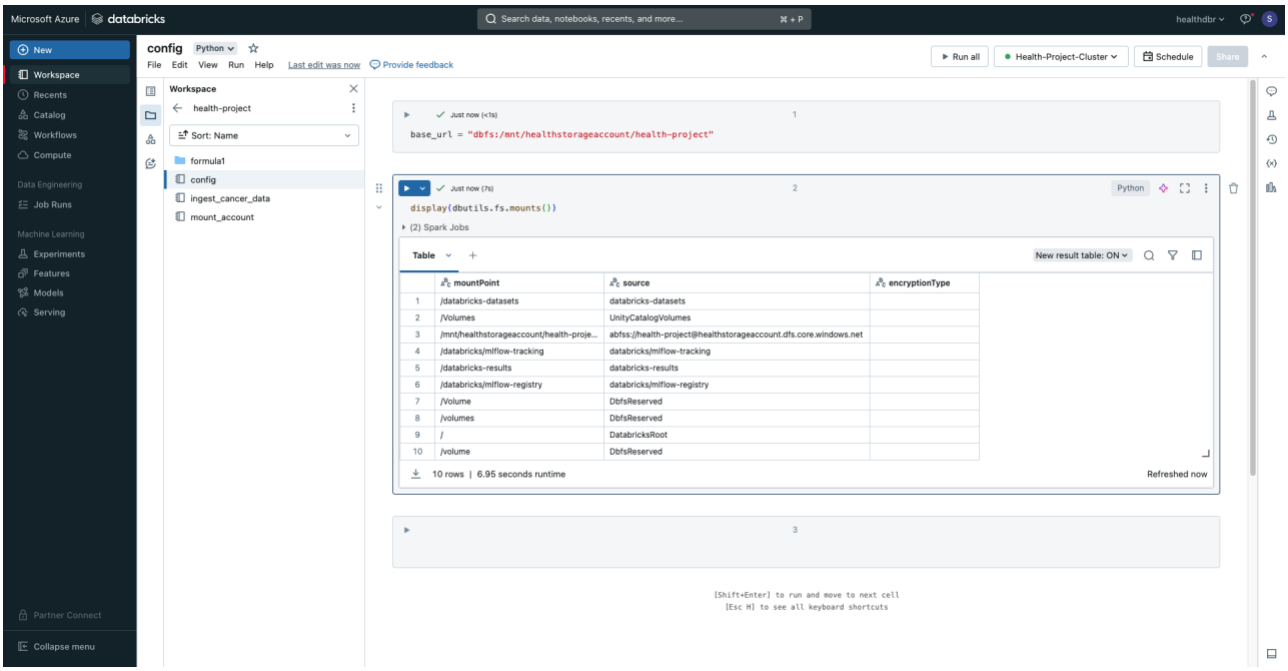
Creating Secret Scope



Mounting ADLS to Databricks Workspace

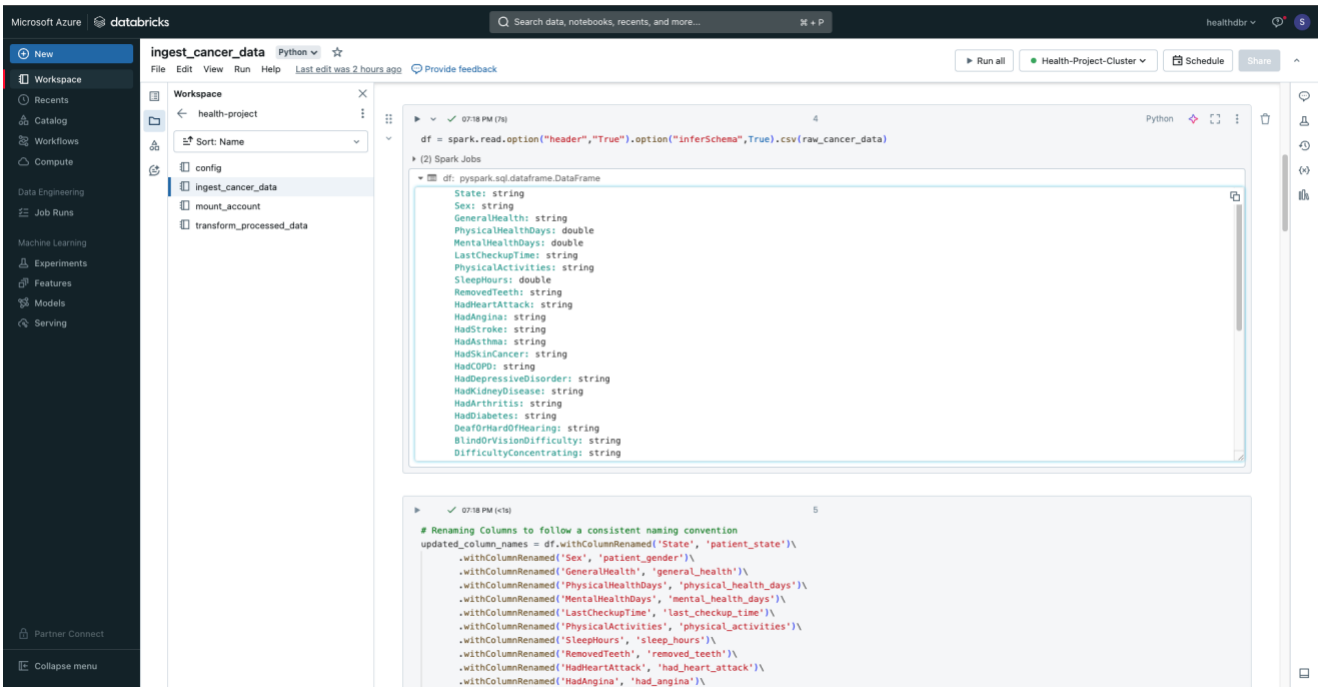


Creating a config file

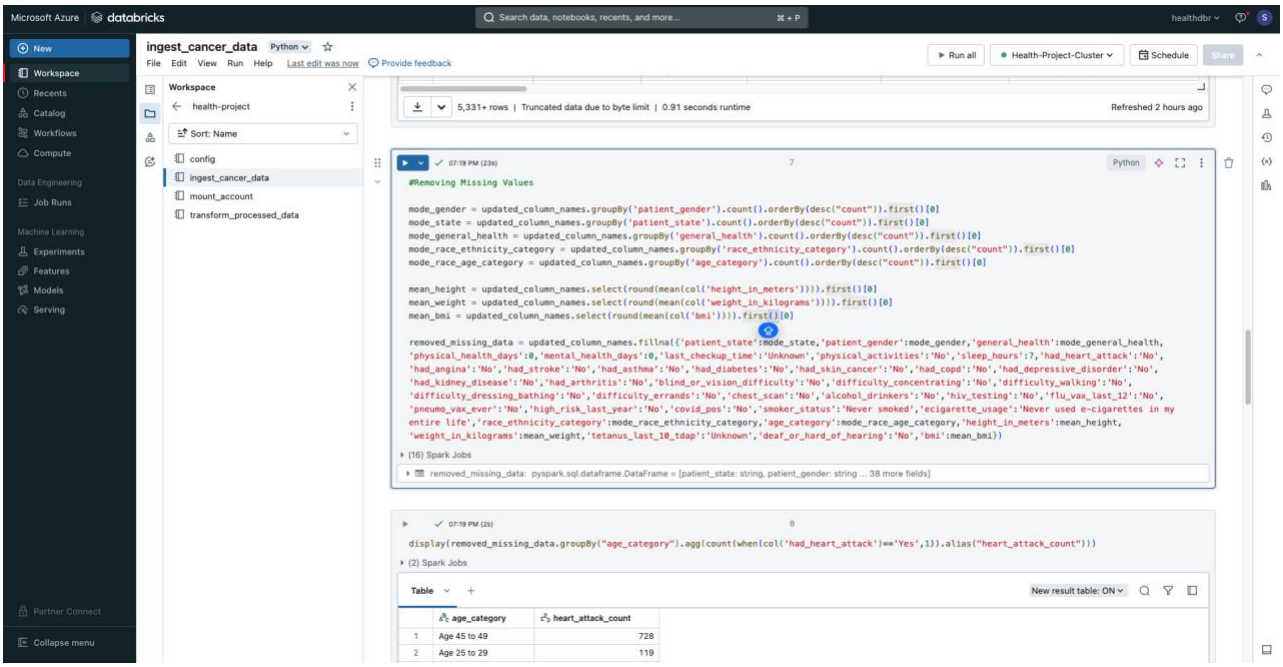

Microsoft Azure
databricks
Search data, notebooks, recents, and more...
healthdbr
New
Workspace
Recents
Catalog
Workflows
Compute
Data Engineering
Job Runs
Machine Learning
Experiments
Features
Models
Serving
Partner Connect
Collapse menu
config
Python
File Edit View Run Help
Last edit was now
Provide feedback
Run all
Health-Project-Cluster
Schedule
Share
Workspace
health-project
Sort: Name
formula1
config
ingest_cancer_data
mount_account
Just now (<1s)
1
base_url = "dbfs:/mnt/healthstorageaccount/health-project"
Just now (7s)
2
Python
display(dbutils.fs.mounts())
(2) Spark Jobs
Table
New result table: ON
mountPoint
source
encryptionType
1 /databricks-datasets databricks-datasets
2 /Volumes UnityCatalogVolumes
3 /mnt/healthstorageaccount/health-proje... abfss://health-project@healthstorageaccount.dfs.core.windows.net
4 /databricks/mlflow-tracking databricks/mlflow-tracking
5 /databricks-results databricks-results
6 /databricks/mlflow-registry databricks/mlflow-registry
7 /Volume DbfsReserved
8 /volumes DbfsReserved
9 / DatabricksRoot
10 /volume DbfsReserved
10 rows | 6.95 seconds runtime
Refreshed now
3
[Shift+Enter] to run and move to next cell
[Esc H] to see all keyboard shortcuts

Data Cleaning

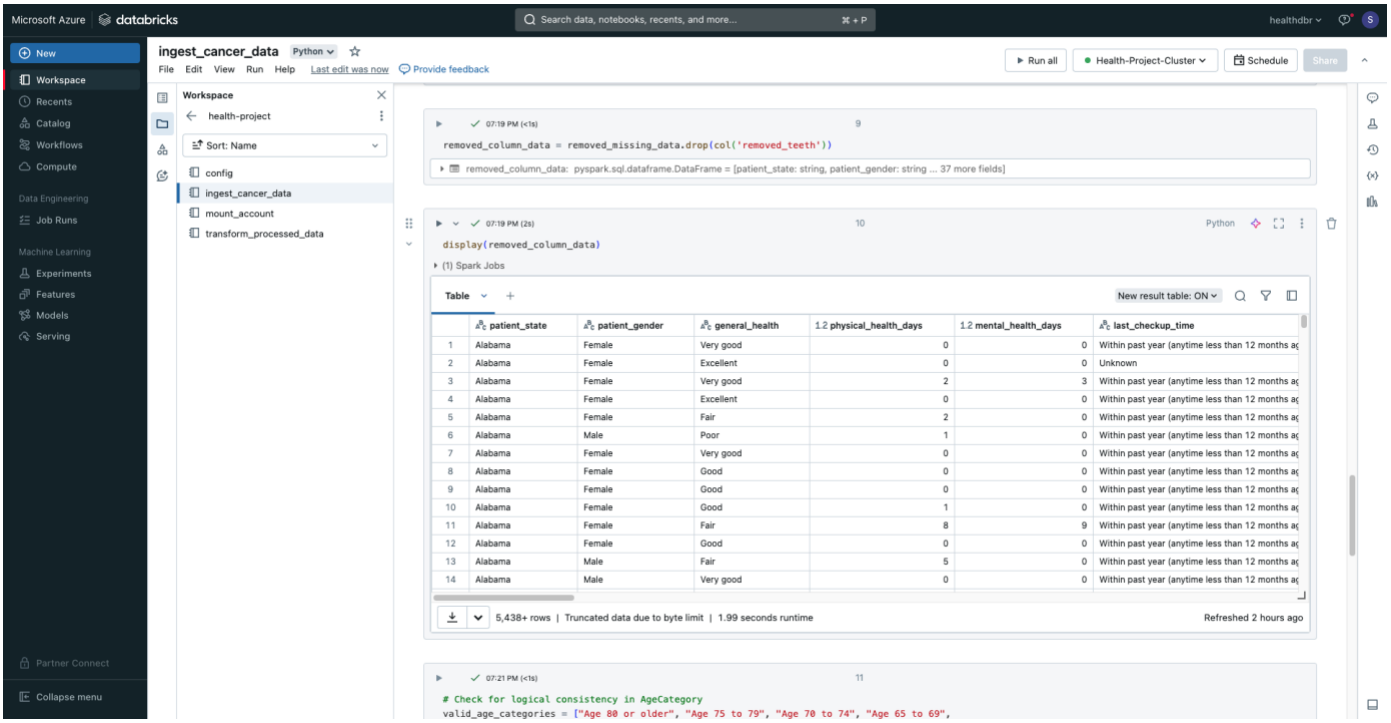
Reading Data through PySpark



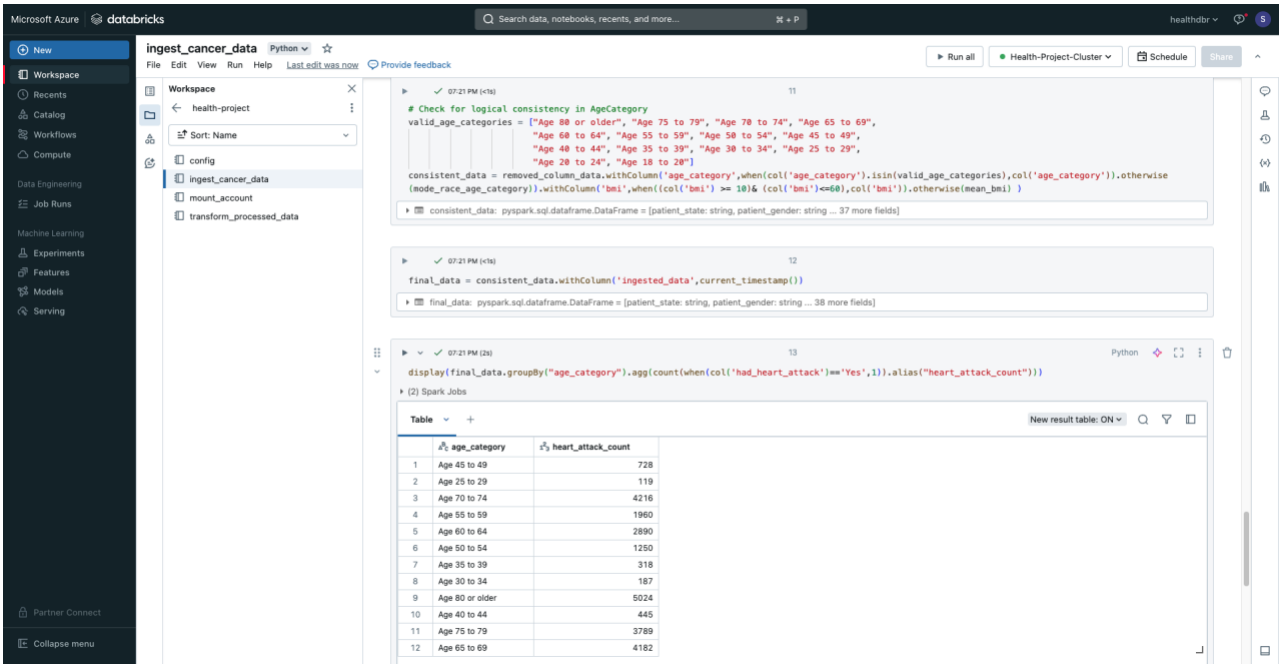
Handling Missing Values


Microsoft Azure
databricks
ingest_cancer_data
Python
File Edit View Run Help Last edit was now Provide feedback
Run all
Health-Project-Cluster
Schedule
Share
Workspace
health-project
Sort: Name
config
ingest_cancer_data
mount_account
transform_processed_data
5,331+ rows | Truncated data due to byte limit | 0.91 seconds runtime
Refreshed 2 hours ago
#Removing Missing Values
mode_gender = updated_column_names.groupBy('patient_gender').count().orderBy(desc("count")).first()[0]
mode_state = updated_column_names.groupBy('patient_state').count().orderBy(desc("count")).first()[0]
mode_general_health = updated_column_names.groupBy('general_health').count().orderBy(desc("count")).first()[0]
mode_race_ethnicity_category = updated_column_names.groupBy('race_ethnicity_category').count().orderBy(desc("count")).first()[0]
mode_race_age_category = updated_column_names.groupBy('age_category').count().orderBy(desc("count")).first()[0]
mean_height = updated_column_names.select(round(mean(col('height_in_meters')))).first()[0]
mean_weight = updated_column_names.select(round(mean(col('weight_in_kilograms')))).first()[0]
mean_bmi = updated_column_names.select(round(mean(col('bmi')))).first()[0]
removed_missing_data = updated_column_names.fillna({'patient_state':mode_state,'patient_gender':mode_gender,'general_health':mode_general_health,
'physical_health_days':0,'mental_health_days':0,'last_checkup_time':'Unknown','physical_activities':'No','sleep_hours':7,'had_heart_attack':'No',
'had_angina':'No','had_stroke':'No','had_asthma':'No','had_diabetes':'No','had_skin_cancer':'No','had_copd':'No','had_depressive_disorder':'No',
'had_kidney_disease':'No','had_arthritis':'No','blind_or_vision_difficulty':'No','difficulty_concentrating':'No','difficulty_walking':'No',
'difficulty_dressing_bathing':'No','difficulty_errands':'No','chest_scan':'No','alcohol_drinkers':'No','hiv_testing':'No','flu_vax_last_12':'No',
'pneumo_vax_ever':'No','high_risk_last_year':'No','covid_pos':'No','smoker_status':'Never smoked','ecigarette_usage':'Never used e-cigarettes in my
entire life','race_ethnicity_category':mode_race_ethnicity_category,'age_category':mode_race_age_category,'height_in_meters':mean_height,
'weight_in_kilograms':mean_weight,'tetanus_last_10_tdap':'Unknown','deaf_or_hard_of_hearing':'No','bmi':mean_bmi})
(16) Spark Jobs
removed_missing_data: pyspark.sql.dataframe.DataFrame = [patient_state: string, patient_gender: string ... 38 more fields]
display(removed_missing_data.groupBy("age_category").agg(count(when(col('had_heart_attack')=='Yes',1)).alias("heart_attack_count")))
(2) Spark Jobs
Table
New result table: ON
age_category
heart_attack_count
1 Age 45 to 49 728
2 Age 25 to 29 119

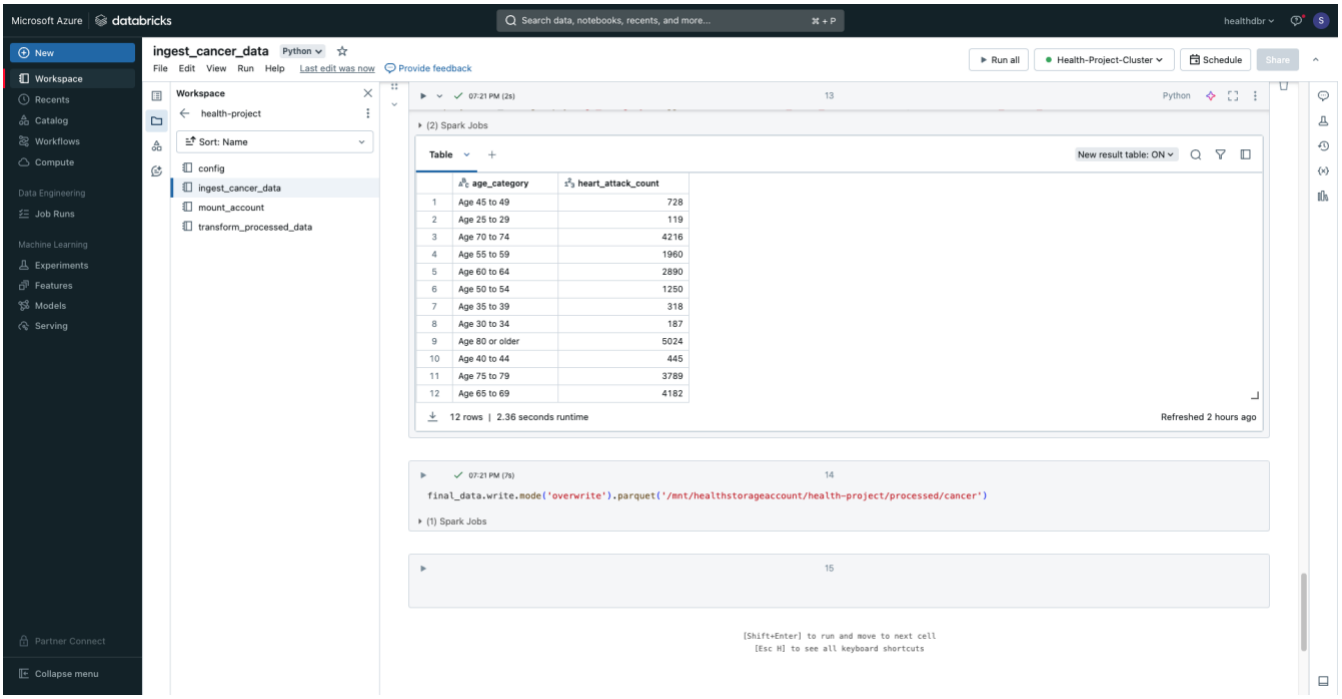
Dropping Unwanted Columns using .drop func



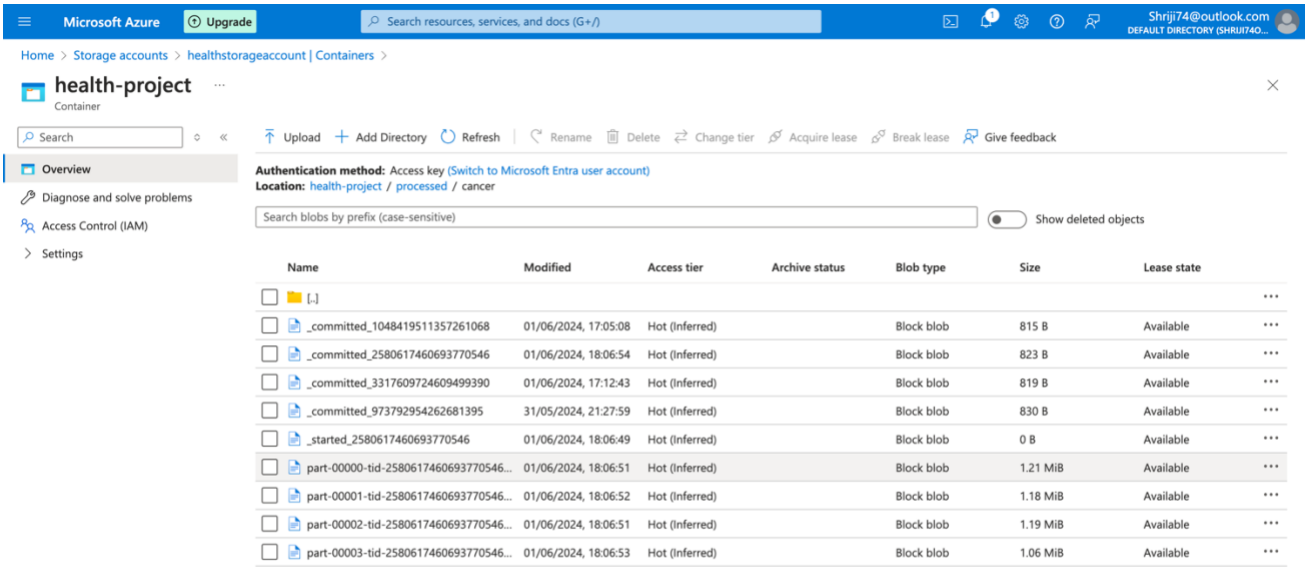
Making Data Consistent



Writing Data to ADLS in Parquet Format

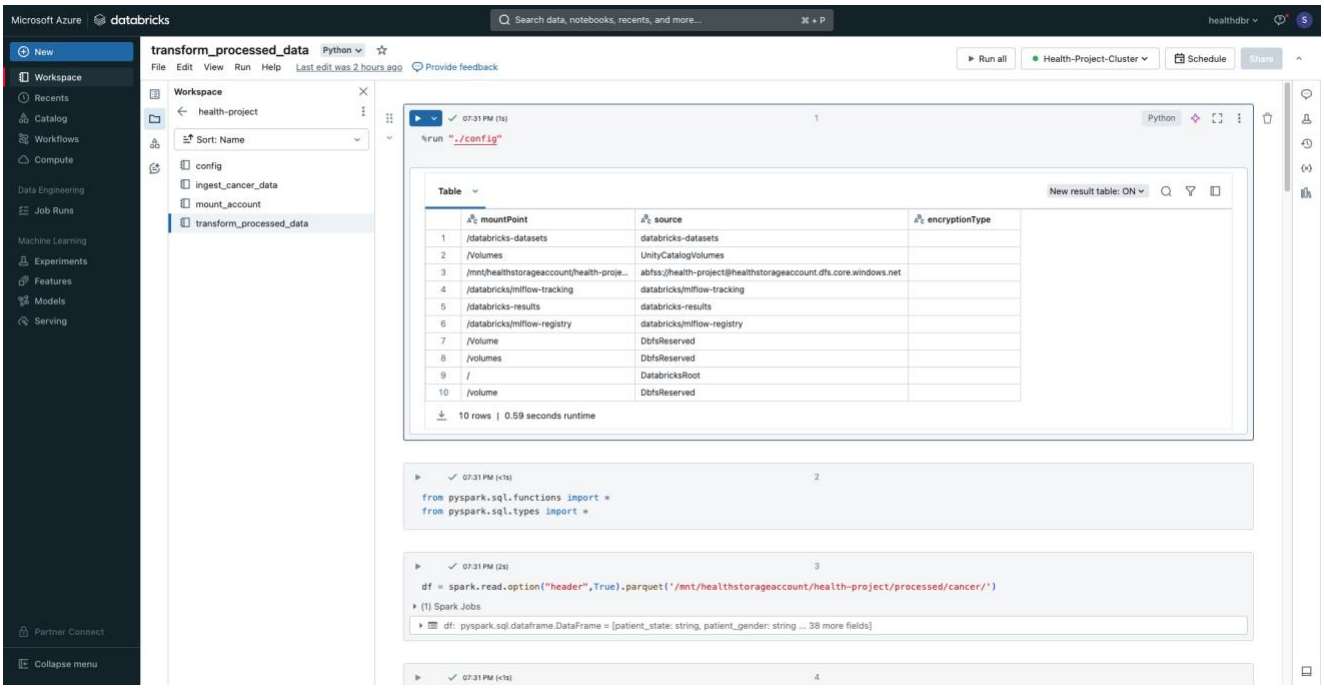


Parquet Processed Data in ADLS

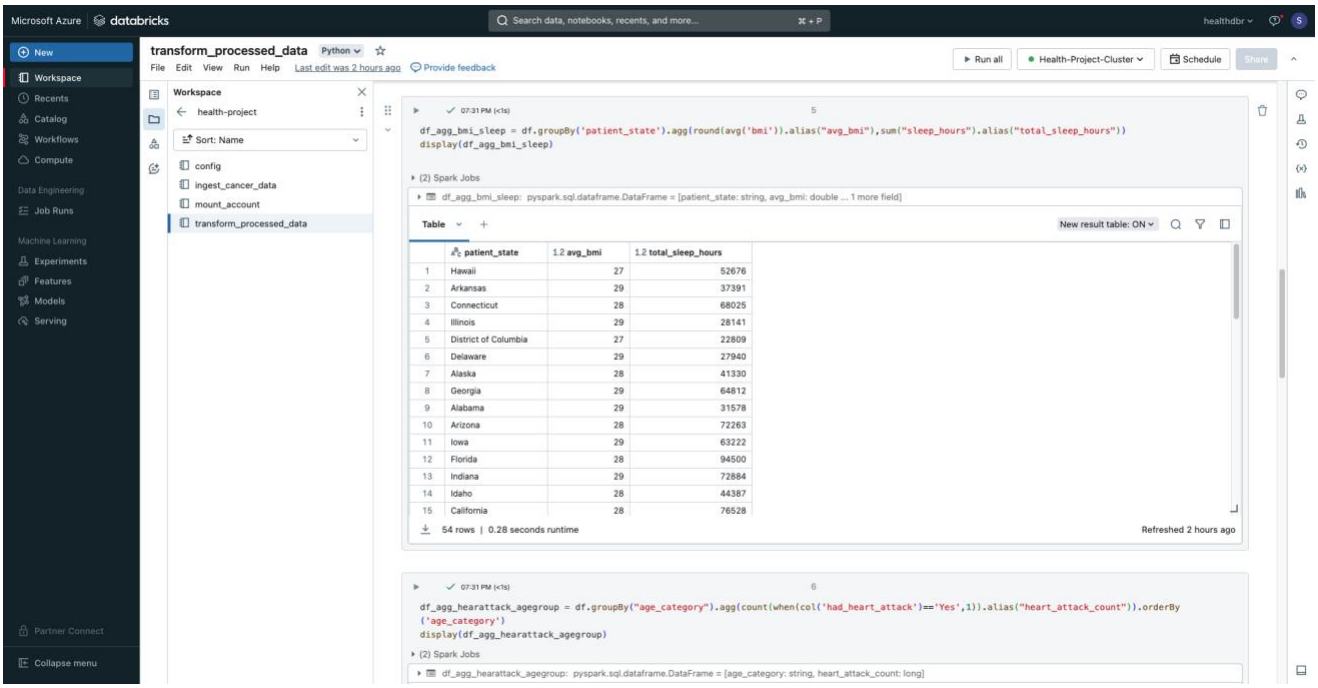


Data Transformation

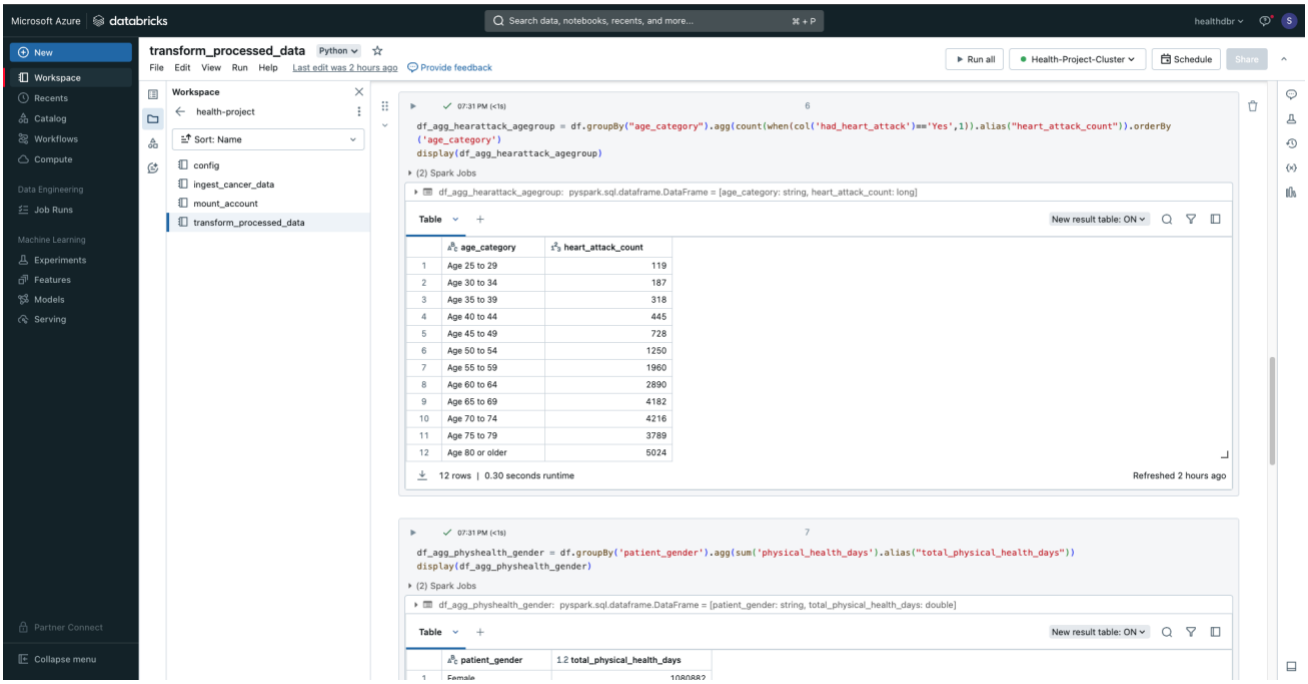
Reading Parquet Data using PySpark



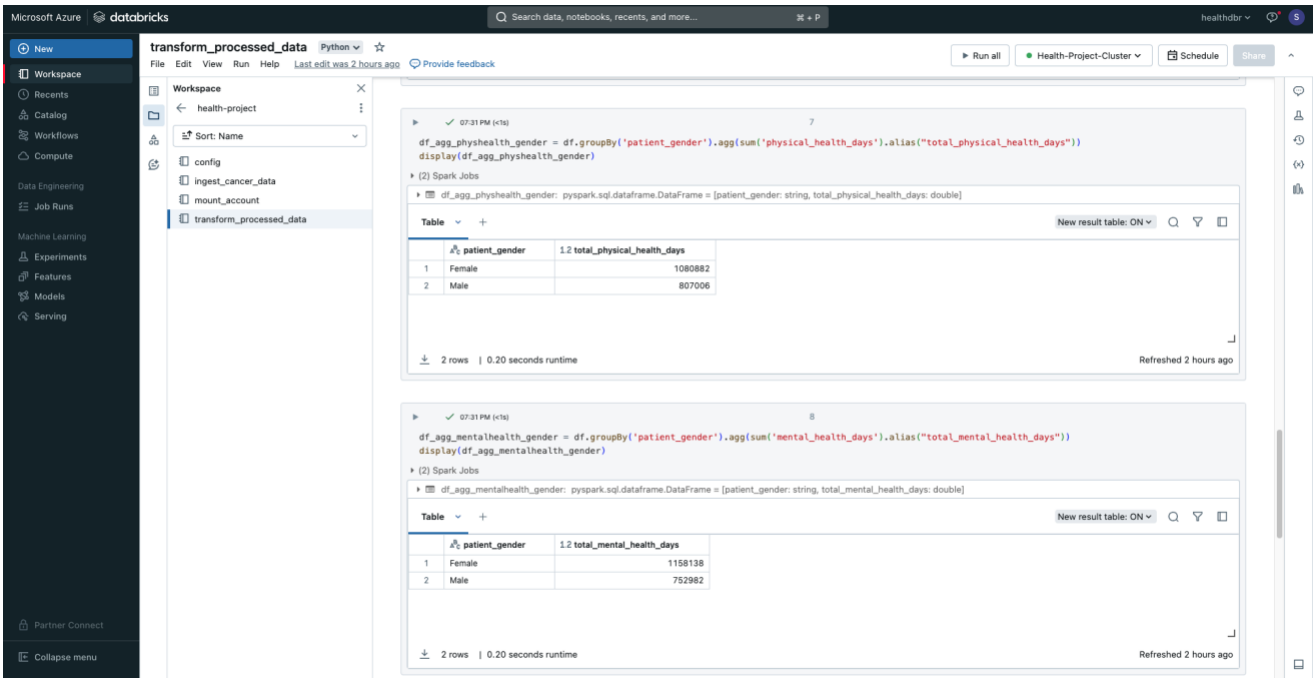
Aggregating Data of BMI and Sleep Hours



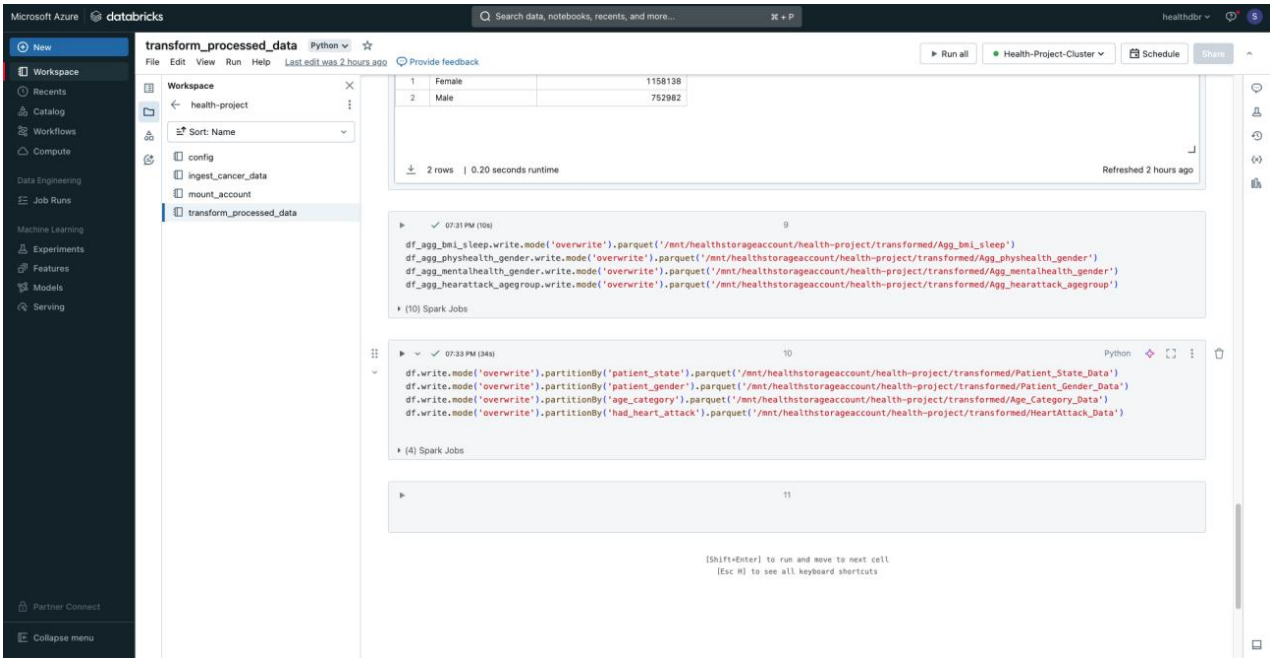
Aggerating Data of Heart Attack and Age group



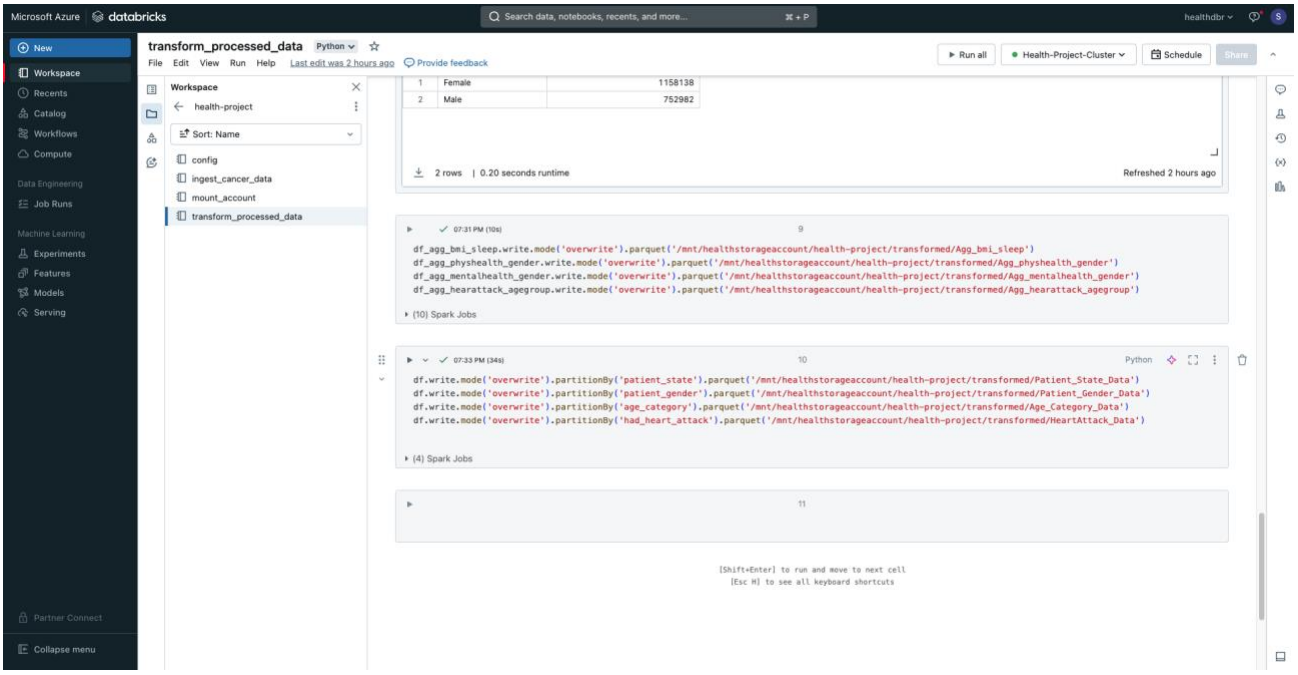
Aggregating Data of Gender physical days



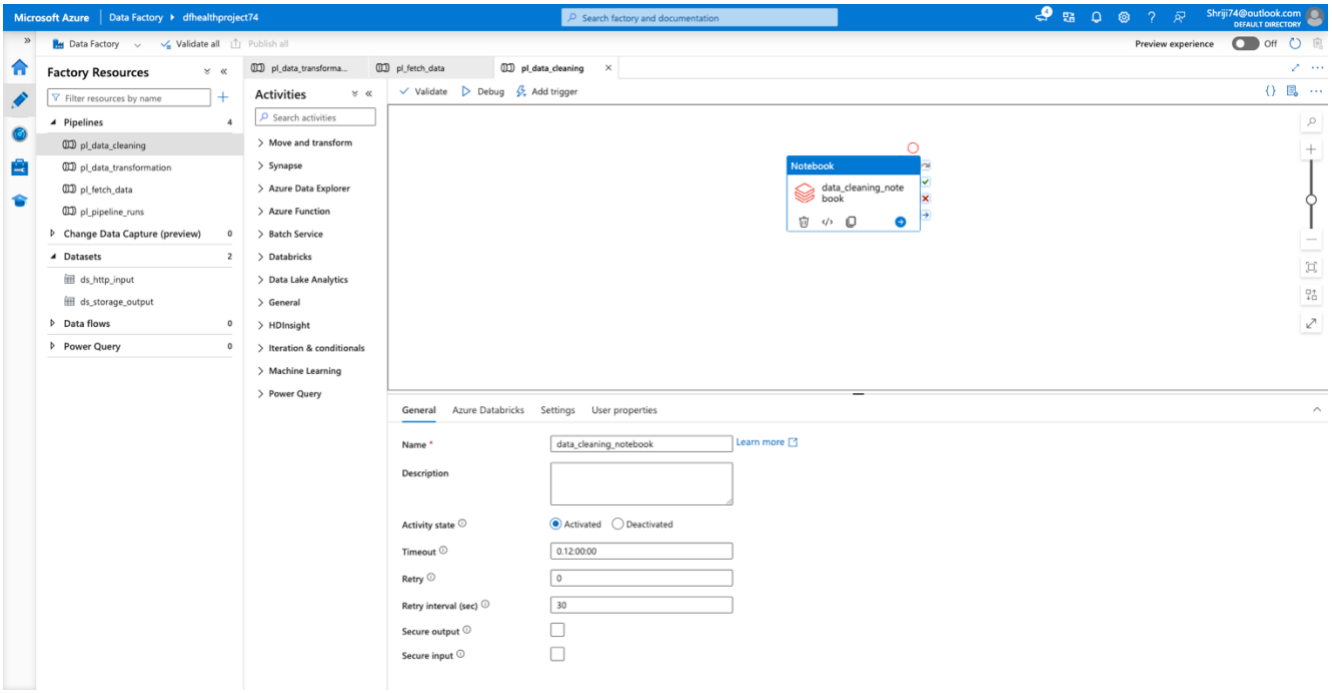
Writing this Aggregated Data in different Directory

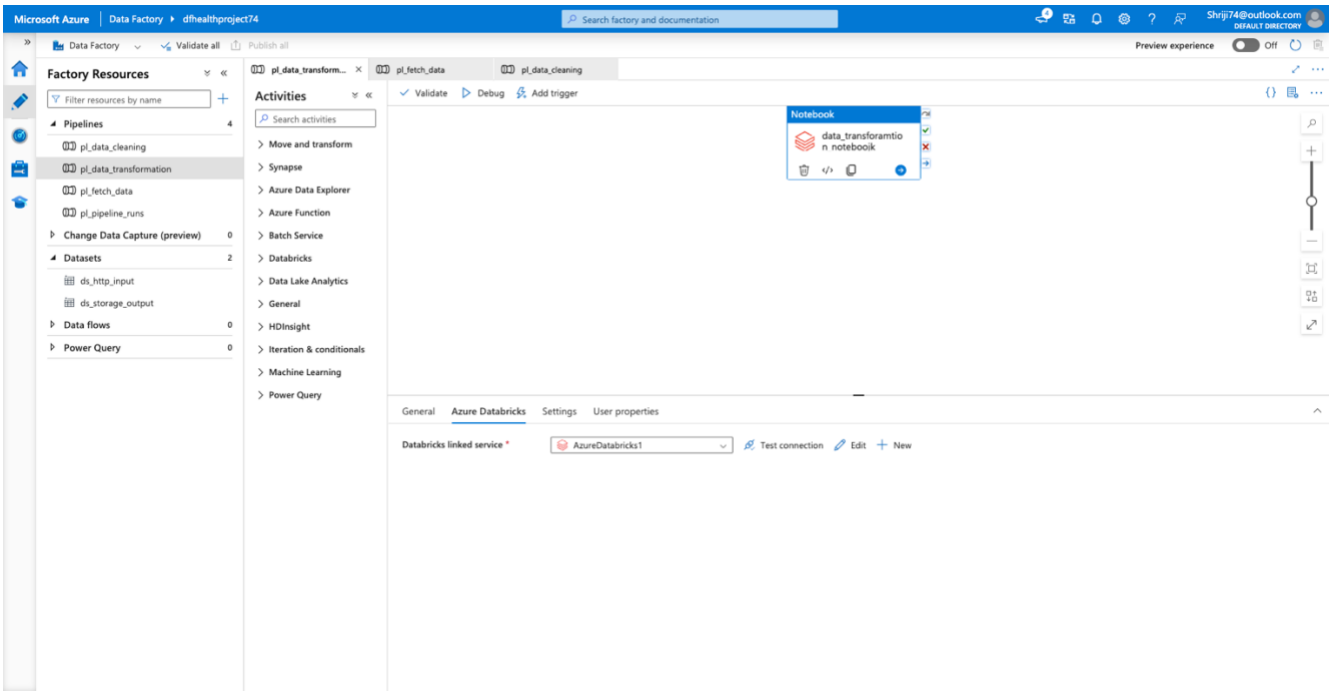


Partitioning the Data and Writing in ADLS in Parquet Format

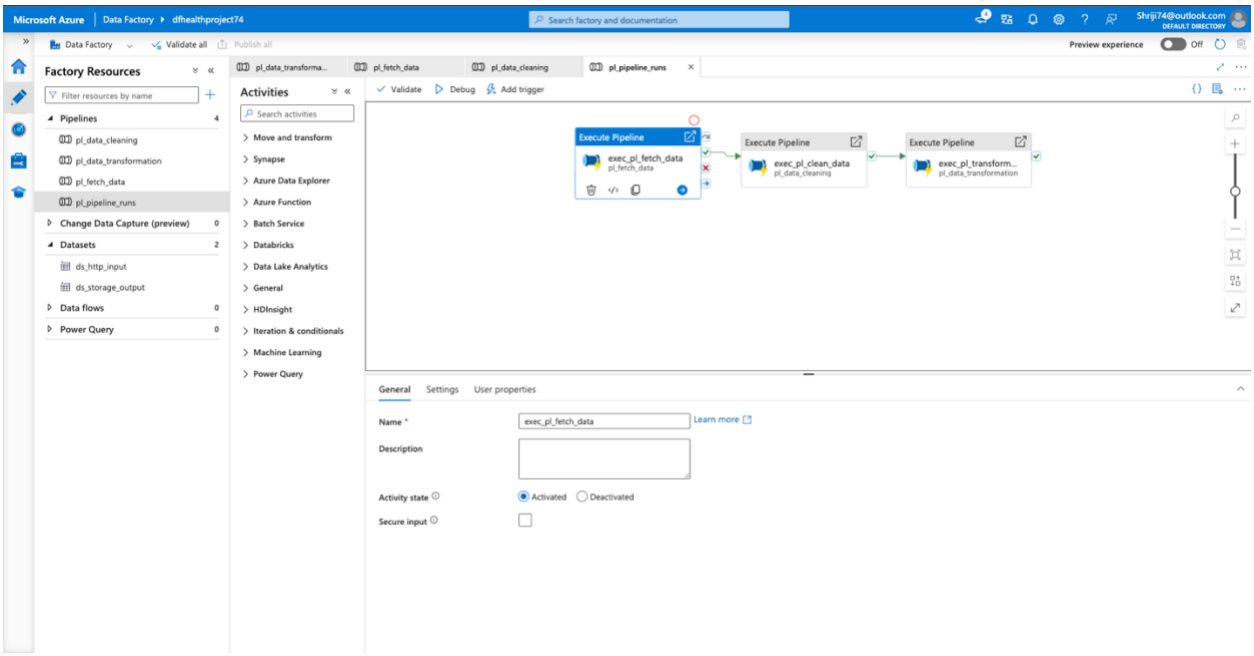

Microsoft Azure | databricks
Search data, notebooks, recents, and more...
transform_processed_data
Python
File Edit View Run Help
Last edit was 2 hours ago
Provide feedback
Run all
Health-Project-Cluster
Schedule
Share
Workspace
health-project
Sort: Name
config
ingest_cancer_data
mount_account
transform_processed_data
1 Female 1158138
2 Male 752982
2 rows | 0.20 seconds runtime
Refreshed 2 hours ago
df_agg_bmi_sleep.write.mode('overwrite').parquet('/mnt/healthstorageaccount/health-project/transformed/Agg_bmi_sleep')
df_agg_physhealth_gender.write.mode('overwrite').parquet('/mnt/healthstorageaccount/health-project/transformed/Agg_physhealth_gender')
df_agg_mentalhealth_gender.write.mode('overwrite').parquet('/mnt/healthstorageaccount/health-project/transformed/Agg_mentalhealth_gender')
df_agg_hearattack_agegroup.write.mode('overwrite').parquet('/mnt/healthstorageaccount/health-project/transformed/Agg_hearattack_agegroup')
(10) Spark Jobs
df.write.mode('overwrite').partitionBy('patient_state').parquet('/mnt/healthstorageaccount/health-project/transformed/Patient_State_Data')
df.write.mode('overwrite').partitionBy('patient_gender').parquet('/mnt/healthstorageaccount/health-project/transformed/Patient_Gender_Data')
df.write.mode('overwrite').partitionBy('age_category').parquet('/mnt/healthstorageaccount/health-project/transformed/Age_Category_Data')
df.write.mode('overwrite').partitionBy('had_heart_attack').parquet('/mnt/healthstorageaccount/health-project/transformed/HeartAttack_Data')
(4) Spark Jobs
[Shift+Enter] to run and move to next cell
[Esc H] to see all keyboard shortcuts

Creating Databricks Notebook Pipeline of Data cleaning and Data Transformation

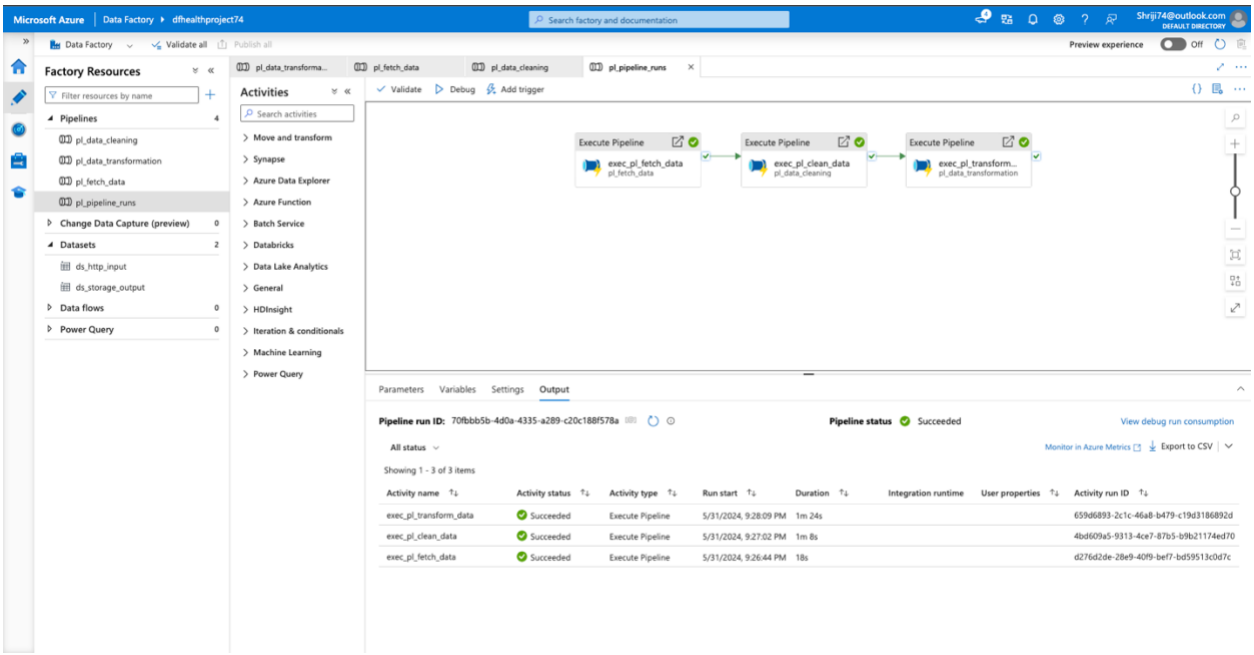




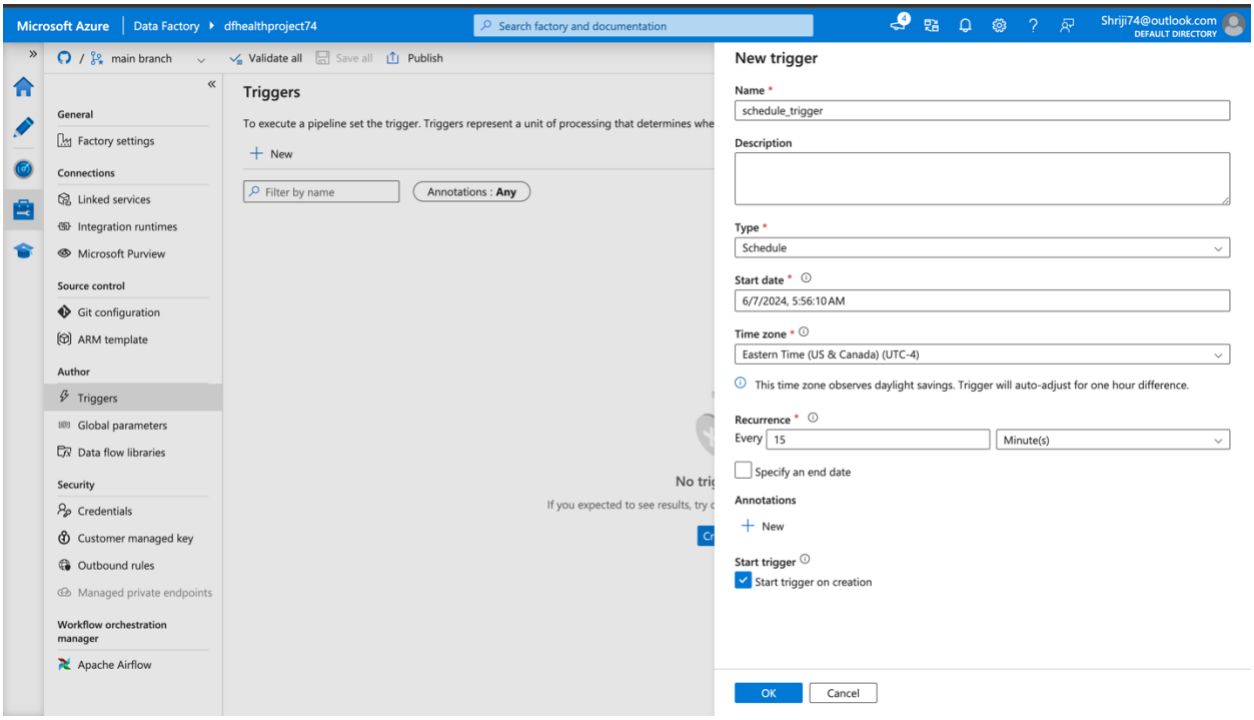
Creating a Pipeline to Execute all three pipelines.



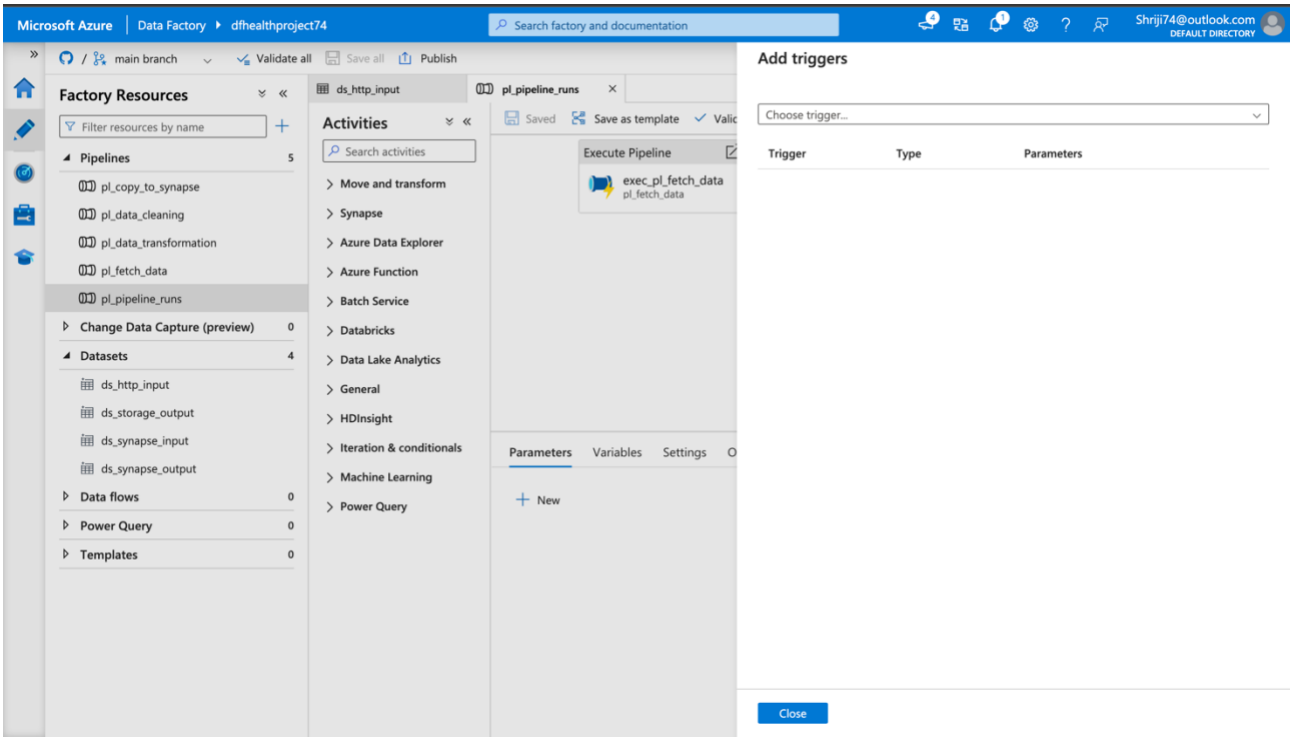
Pipeline Succeeded

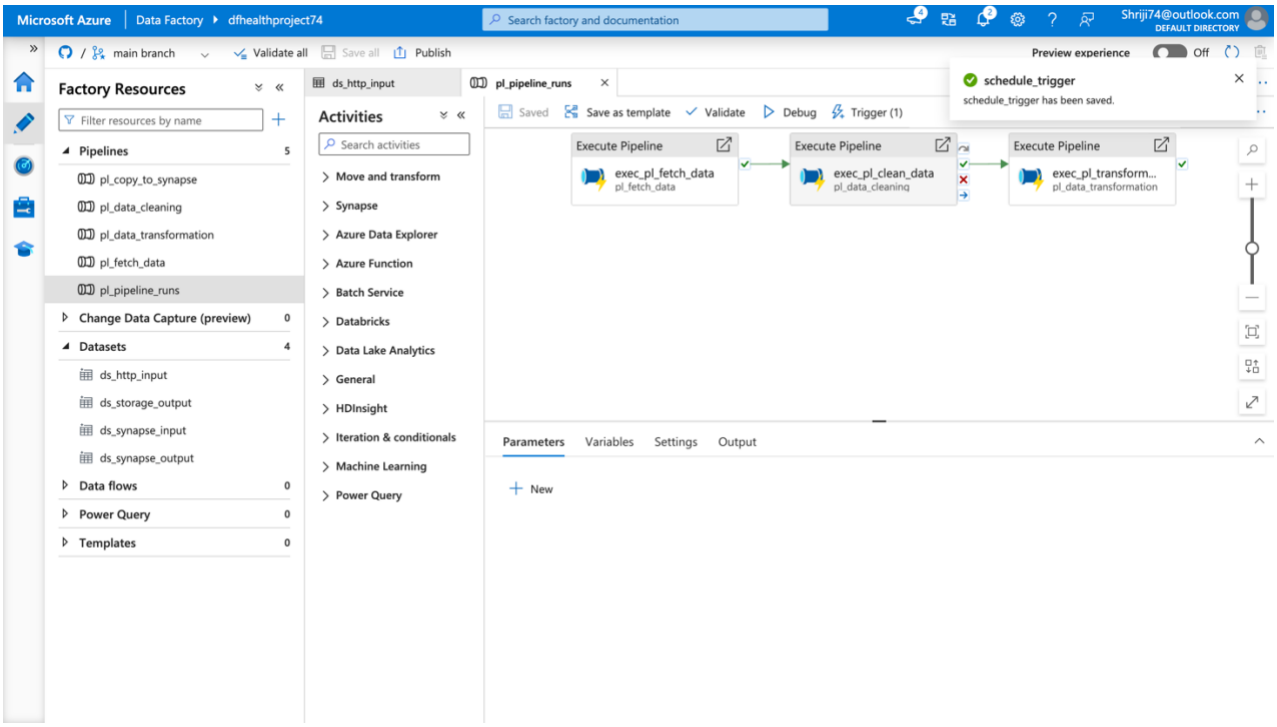


Creating Schedule Trigger



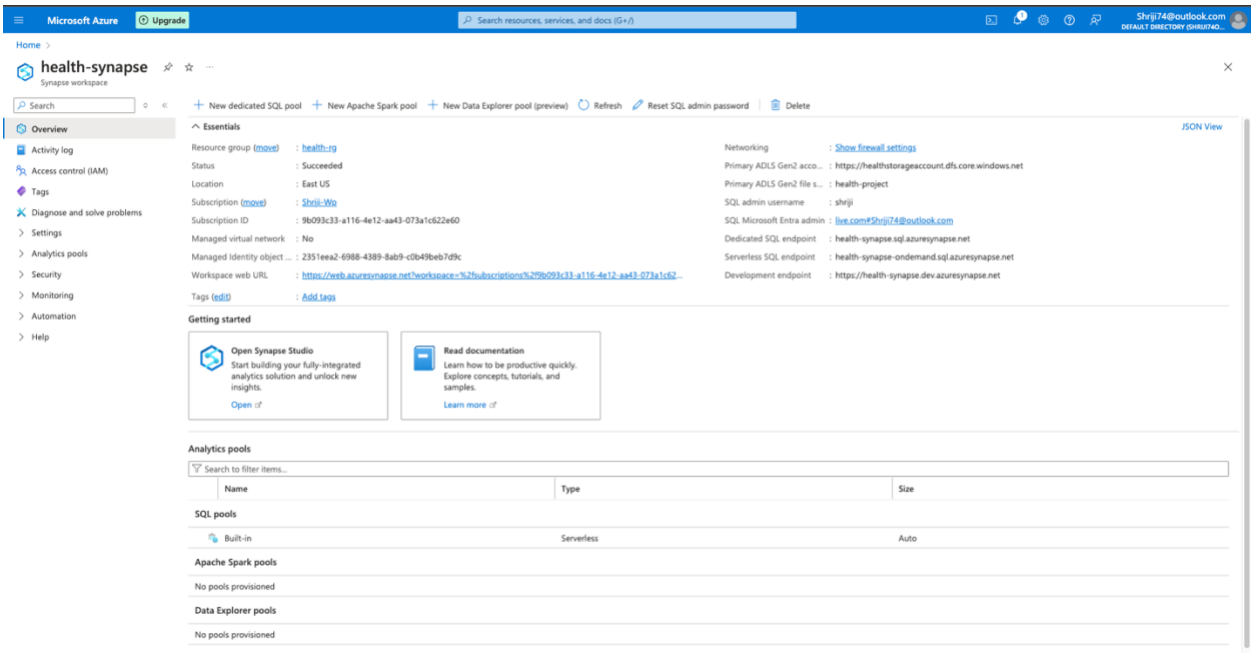
Adding Schedule Trigger to Pipeline



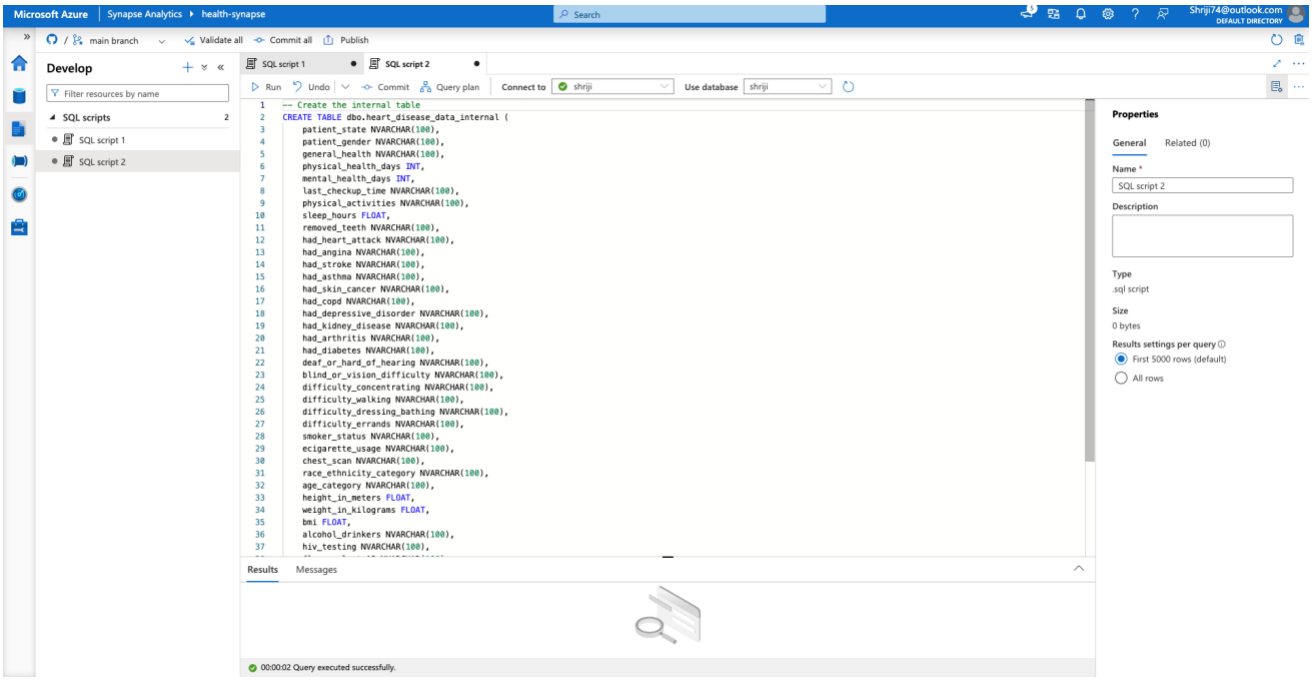


Azure Synapse Anaytics

Creating Synapse Workspace

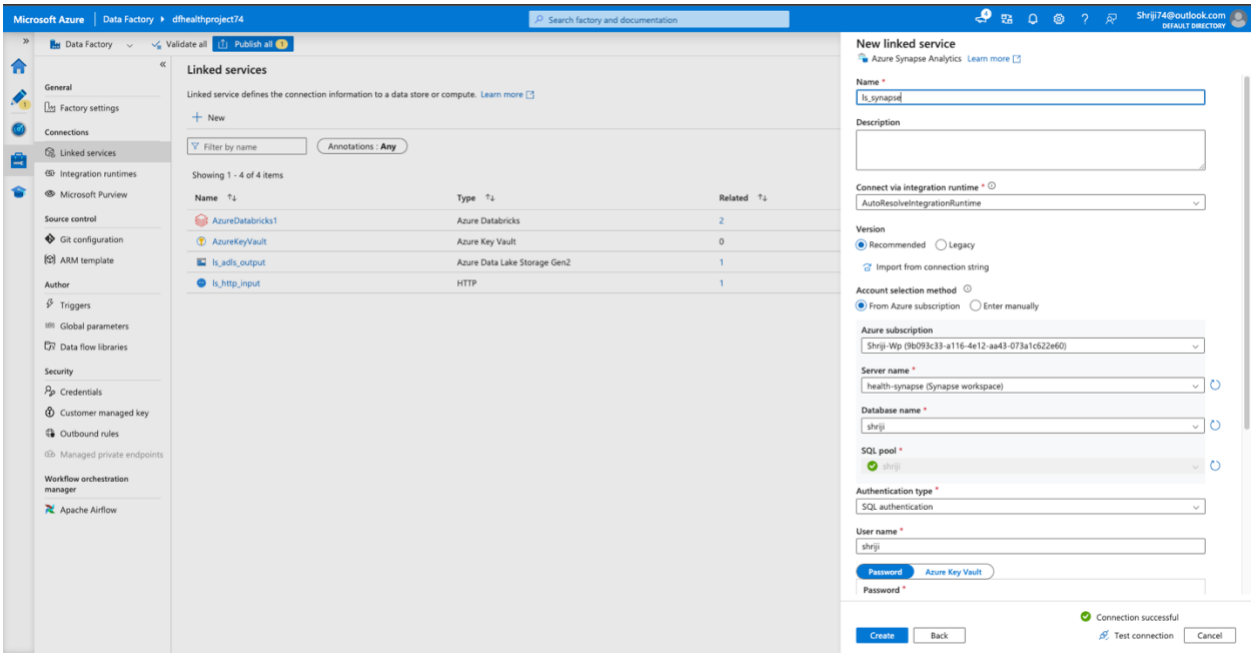


Creating Table in Dedicated SQL Pool

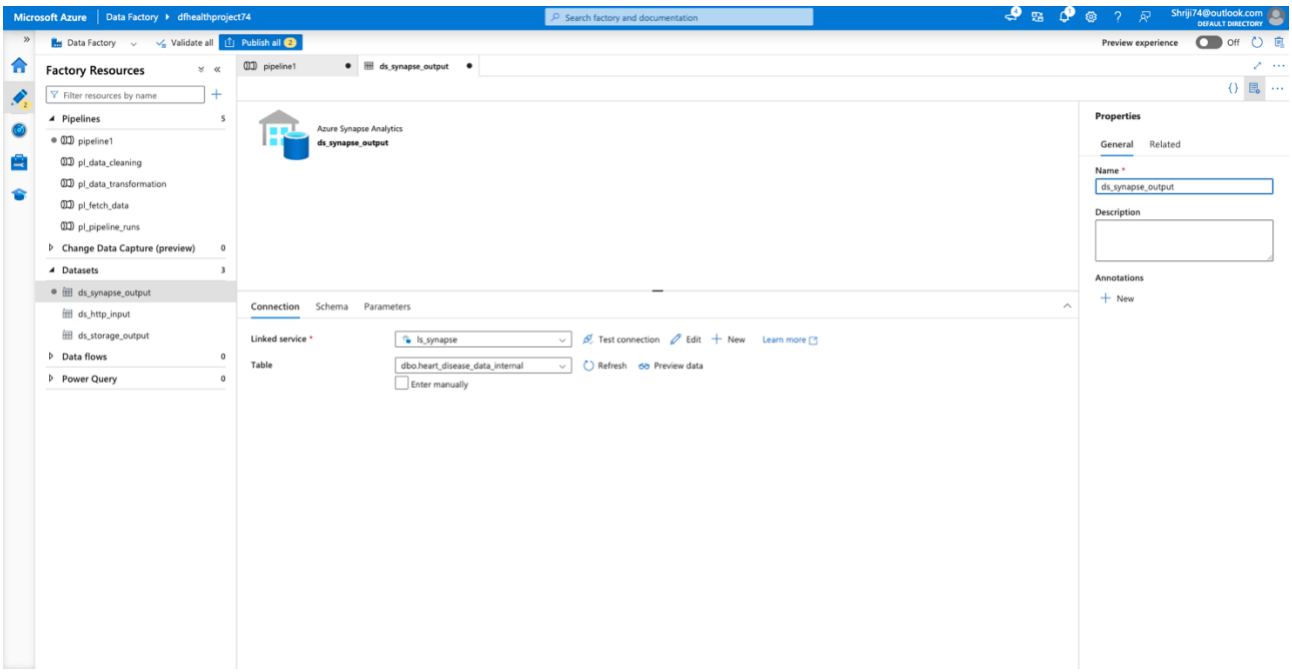



```sql
-- Create the internal table
CREATE TABLE dbo.heart_disease_data_internal (
    patient_state NVARCHAR(100),
    patient_gender NVARCHAR(100),
    general_health NVARCHAR(100),
    physical_health_days INT,
    mental_health_days INT,
    last_checkup_time NVARCHAR(100),
    physical_activities NVARCHAR(100),
    sleep_hours FLOAT,
    removed_teeth NVARCHAR(100),
    had_heart_attack NVARCHAR(100),
    had_angina NVARCHAR(100),
    had_stroke NVARCHAR(100),
    had_asthma NVARCHAR(100),
    had_skin_cancer NVARCHAR(100),
    had_copd NVARCHAR(100),
    had_depressive_disorder NVARCHAR(100),
    had_kidney_disease NVARCHAR(100),
    had_arthritis NVARCHAR(100),
    had_diabetes NVARCHAR(100),
    deaf_or_hard_of_hearing NVARCHAR(100),
    blind_or_vision_difficulty NVARCHAR(100),
    difficulty_concentrating NVARCHAR(100),
    difficulty_walking NVARCHAR(100),
    difficulty_dressing_bathing NVARCHAR(100),
    difficulty_errands NVARCHAR(100),
    smoker_status NVARCHAR(100),
    ecigarette_usage NVARCHAR(100),
    chest_scan NVARCHAR(100),
    race_ethnicity_category NVARCHAR(100),
    age_category NVARCHAR(100),
    height_in_meters FLOAT,
    weight_in_kilograms FLOAT,
    bmi FLOAT,
    alcohol_drinkers NVARCHAR(100),
    hiv_testing NVARCHAR(100),
```

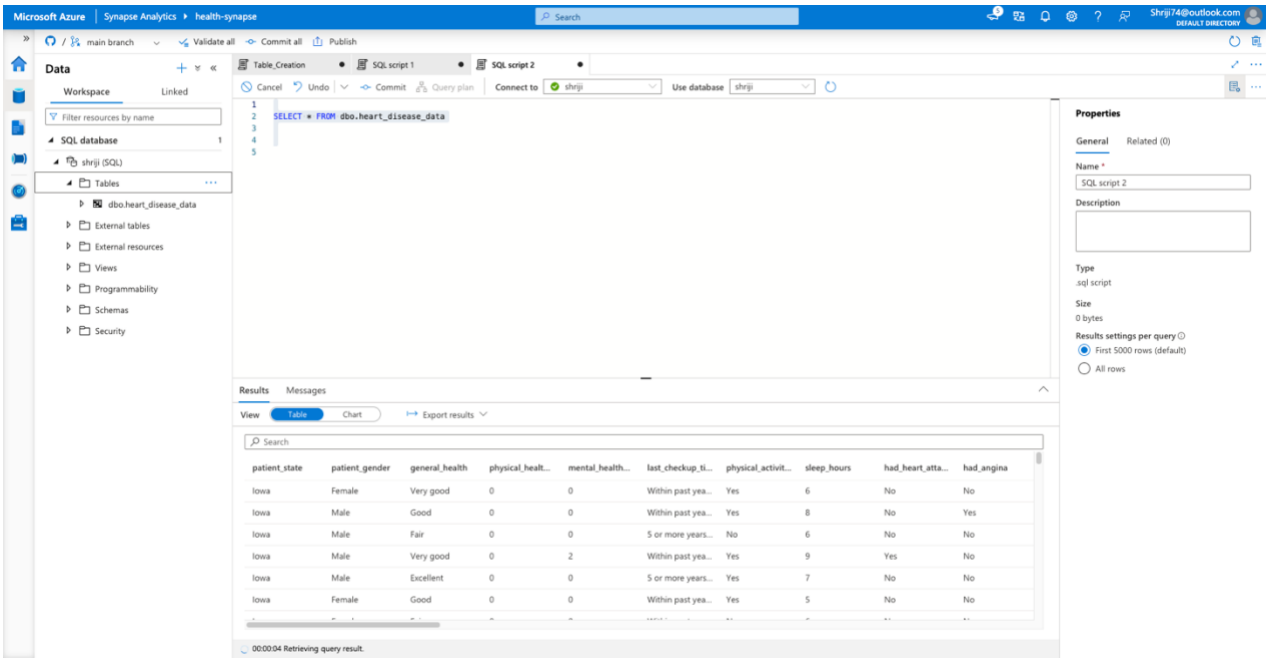
Creating Linked Service in Synapse



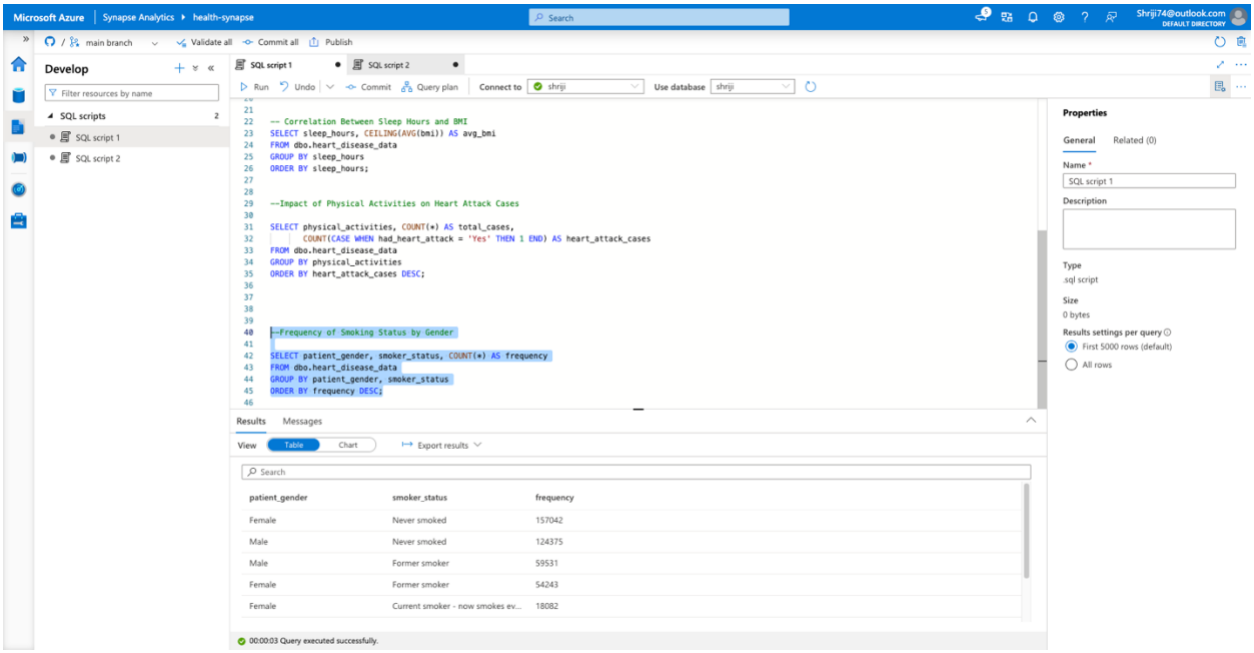
Creating Dataset for Synapse

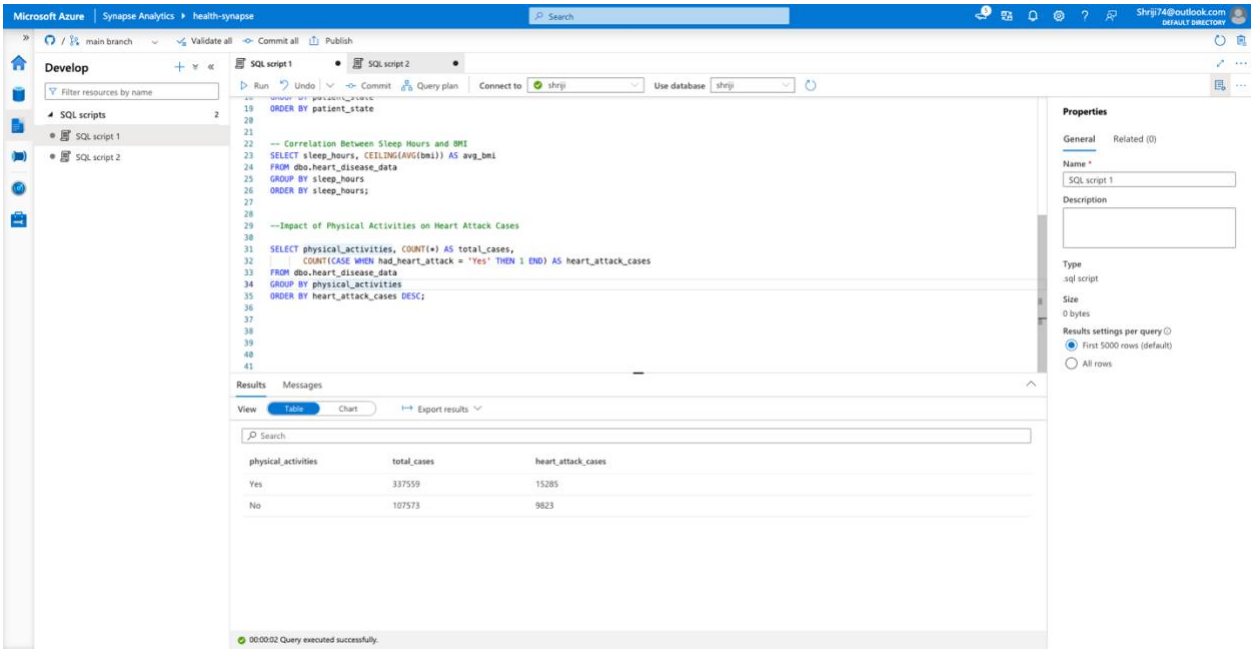


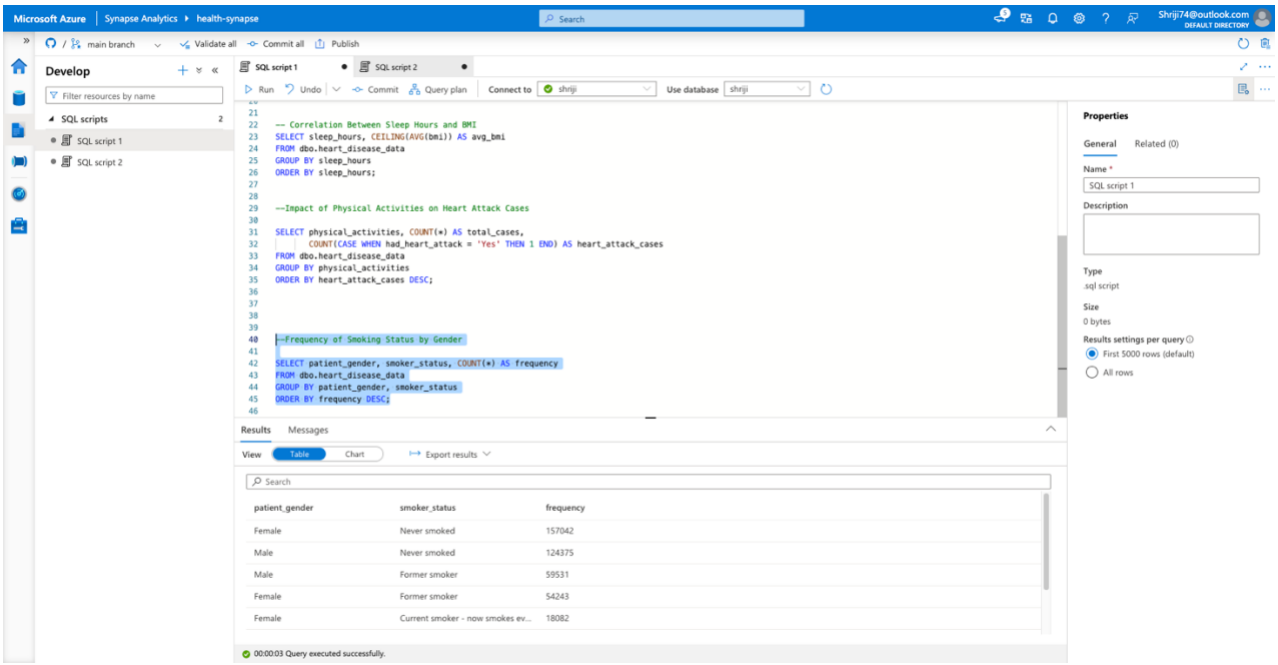
Transferred Data in Synapse

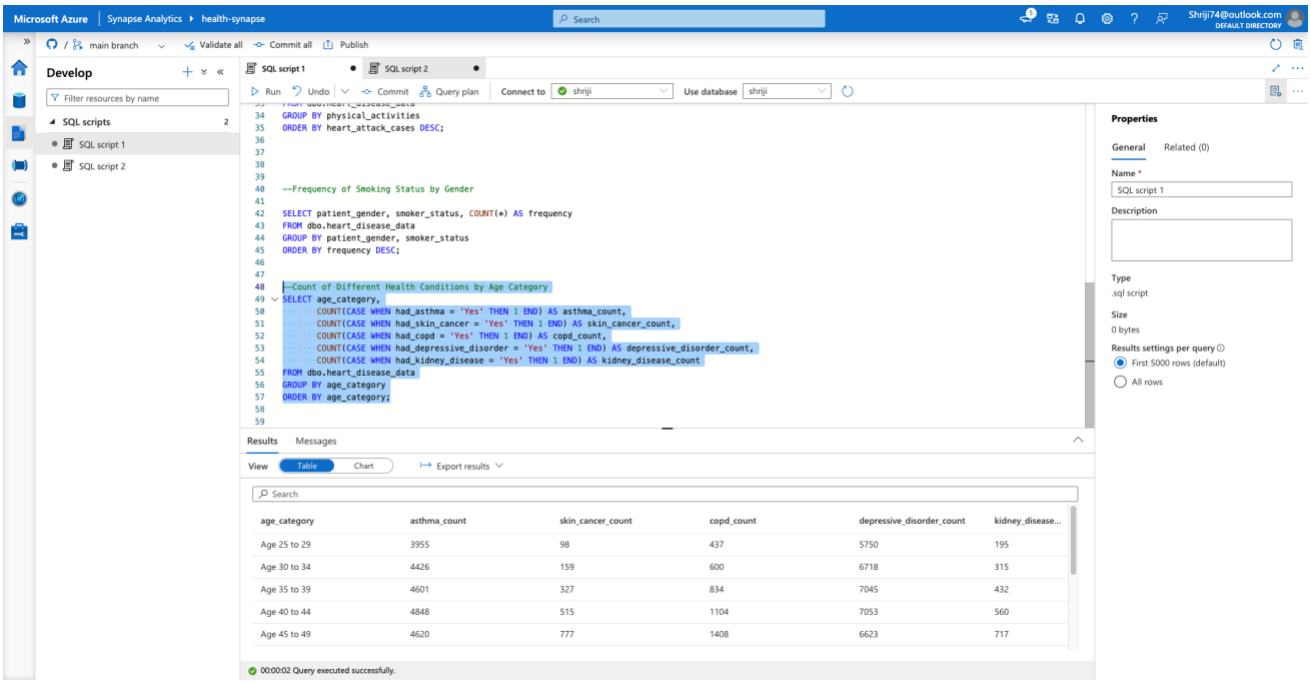


Performing Different Analysis in Synapse

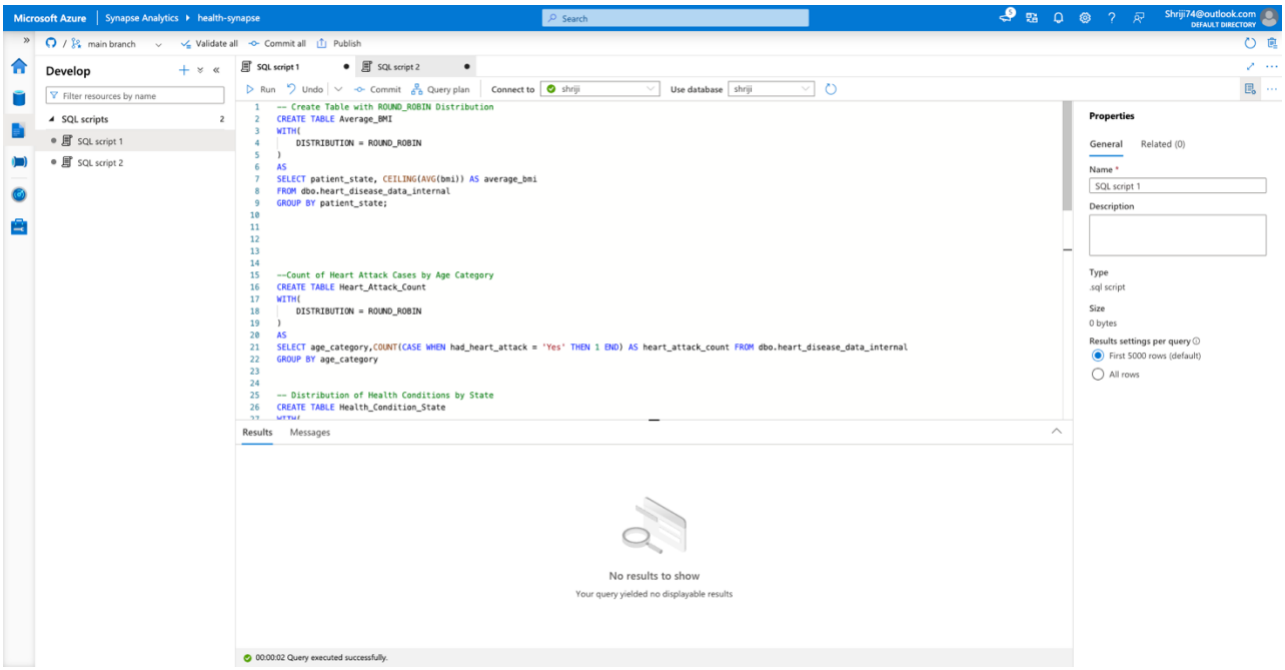


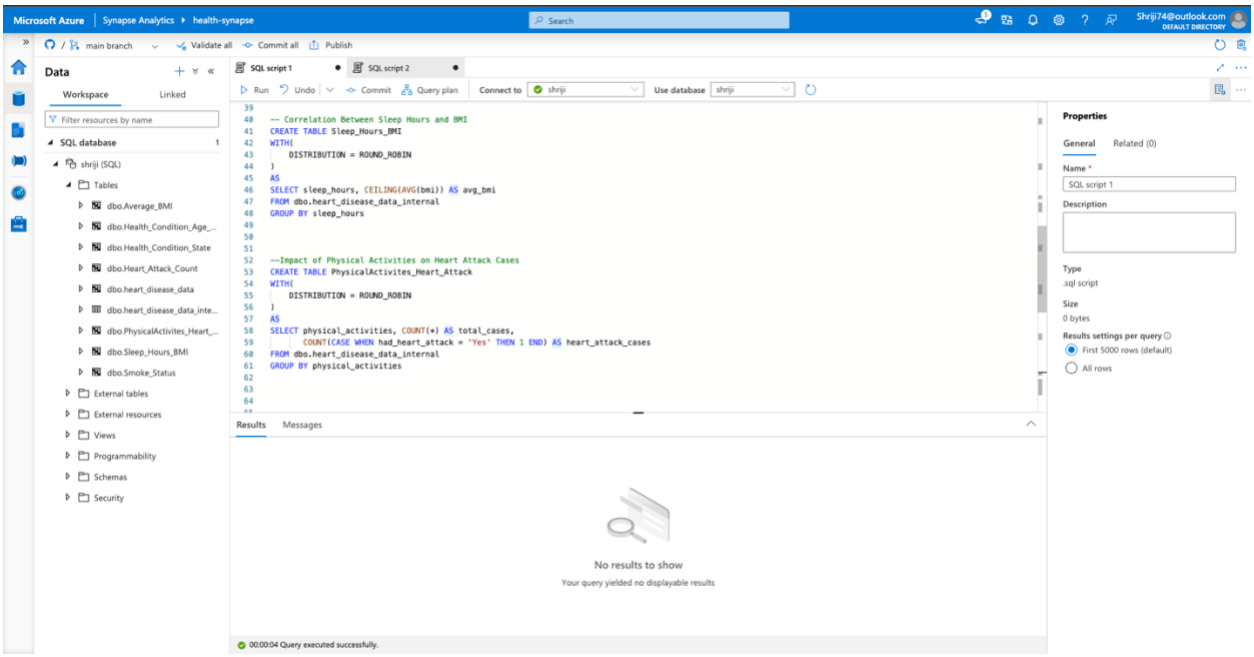



Microsoft Azure | Synapse Analytics | health-synapse
Develop
SQL scripts
SQL script 1
SQL script 2
-- Correlation Between Sleep Hours and BMI
SELECT sleep_hours, CEILING(AVG(bmi)) AS avg_bmi
FROM dbo.heart_disease_data
GROUP BY sleep_hours
ORDER BY sleep_hours;
--Impact of Physical Activities on Heart Attack Cases
SELECT physical_activities, COUNT(*) AS total_cases,
COUNT(CASE WHEN had_heart_attack = 'Yes' THEN 1 END) AS heart_attack_cases
FROM dbo.heart_disease_data
GROUP BY physical_activities
ORDER BY heart_attack_cases DESC;
--Frequency of Smoking Status by Gender
SELECT patient_gender, smoker_status, COUNT(*) AS frequency
FROM dbo.heart_disease_data
GROUP BY patient_gender, smoker_status
ORDER BY frequency DESC;
Properties
Results
Messages
patient_gender smoker_status frequency
Female Never smoked 157042
Male Never smoked 124375
Male Former smoker 59531
Female Former smoker 54243
Female Current smoker - now smokes ev... 18082
00:00:03 Query executed successfully.


Microsoft Azure | Synapse Analytics ▸ health-synapse
Search
Shriji74@outlook.com
DEFAULT DIRECTORY
main branch
Validate all
Commit all
Publish
Develop
Filter resources by name
SQL scripts 2
SQL script 1
SQL script 2
SQL script 1
SQL script 2
Run
Undo
Commit
Query plan
Connect to
shriji
Use database
shriji
GROUP BY physical_activities
ORDER BY heart_attack_cases DESC;
--Frequency of Smoking Status by Gender
SELECT patient_gender, smoker_status, COUNT(*) AS frequency
FROM dbo.heart_disease_data
GROUP BY patient_gender, smoker_status
ORDER BY frequency DESC;
--Count of Different Health Conditions by Age Category
SELECT age_category,
COUNT(CASE WHEN had_asthma = 'Yes' THEN 1 END) AS asthma_count,
COUNT(CASE WHEN had_skin_cancer = 'Yes' THEN 1 END) AS skin_cancer_count,
COUNT(CASE WHEN had_copd = 'Yes' THEN 1 END) AS copd_count,
COUNT(CASE WHEN had_depressive_disorder = 'Yes' THEN 1 END) AS depressive_disorder_count,
COUNT(CASE WHEN had_kidney_disease = 'Yes' THEN 1 END) AS kidney_disease_count
FROM dbo.heart_disease_data
GROUP BY age_category
ORDER BY age_category;
Properties
General
Related (0)
Name *
SQL script 1
Description
Type
.sql script
Size
0 bytes
Results settings per query
First 5000 rows (default)
All rows
Results
Messages
View
Table
Chart
Export results
Search
age_category
asthma_count
skin_cancer_count
copd_count
depressive_disorder_count
kidney_disease...
Age 25 to 29
3955
98
437
5750
195
Age 30 to 34
4426
159
600
6718
315
Age 35 to 39
4601
327
834
7045
432
Age 40 to 44
4848
515
1104
7053
560
Age 45 to 49
4620
777
1408
6623
717
00:00:02 Query executed successfully.

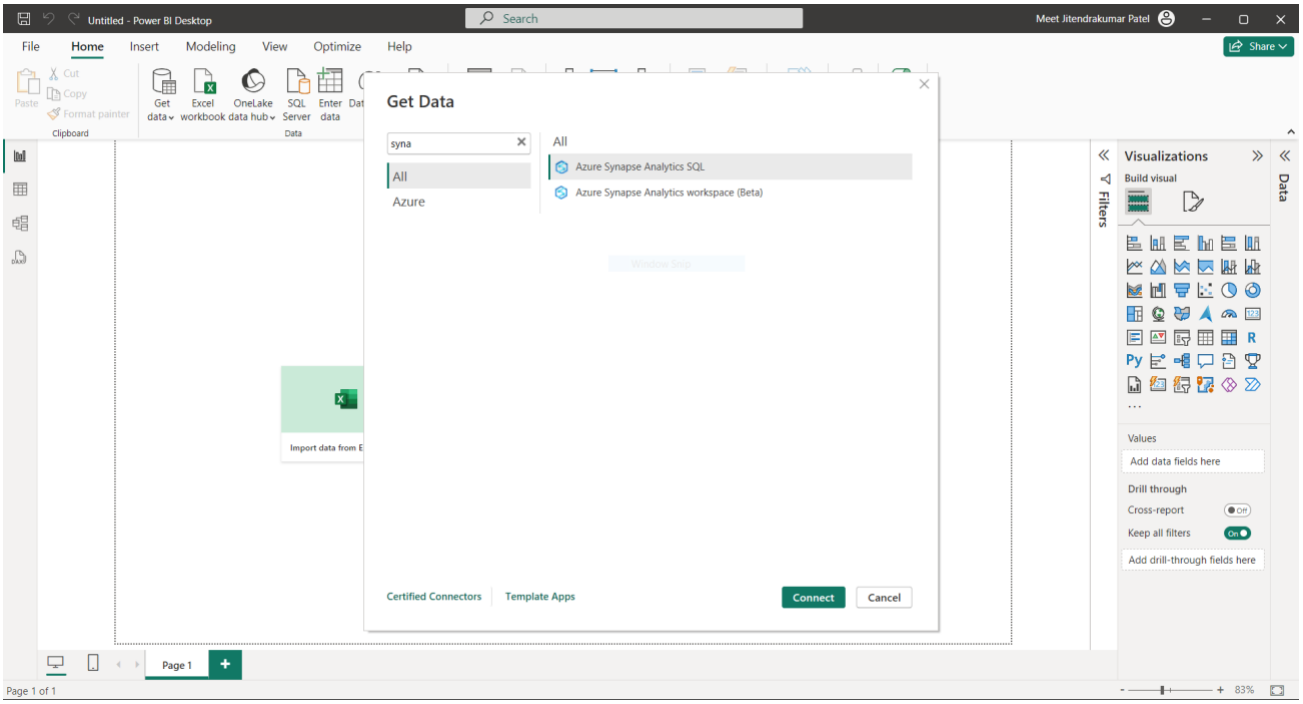
Creating Table in Dedicated SQL of those Analysis for Visualization

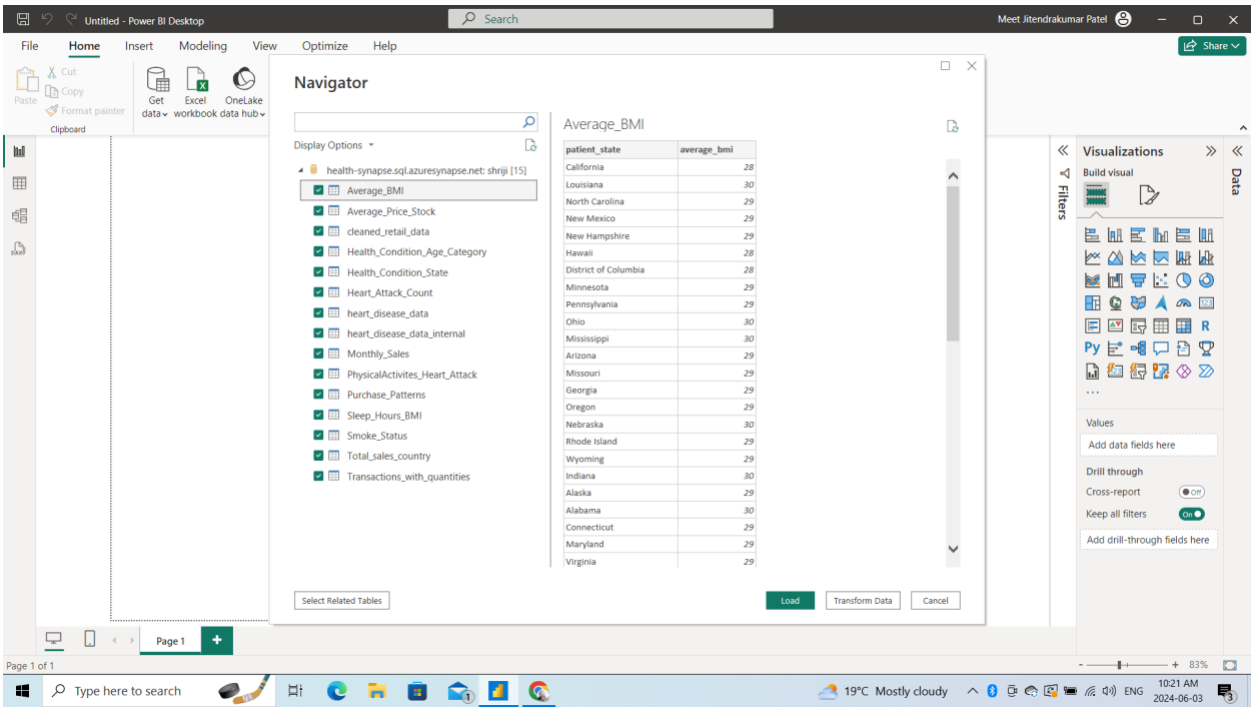




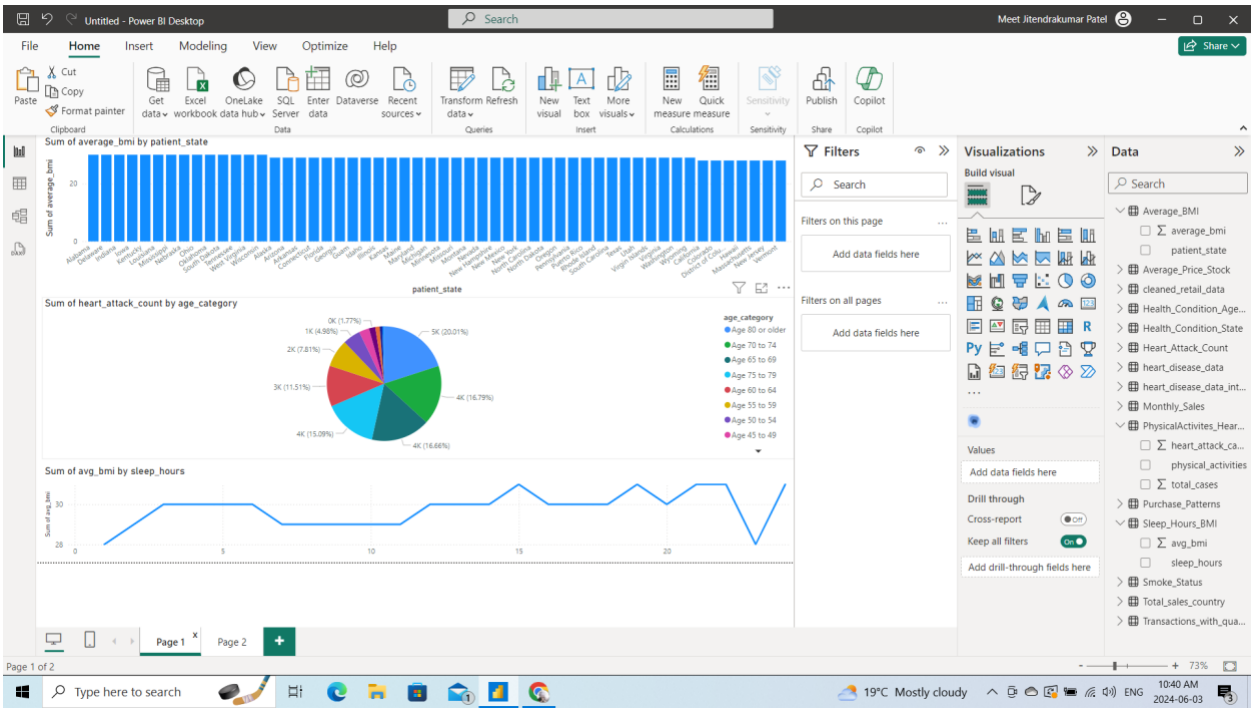
PowerBI

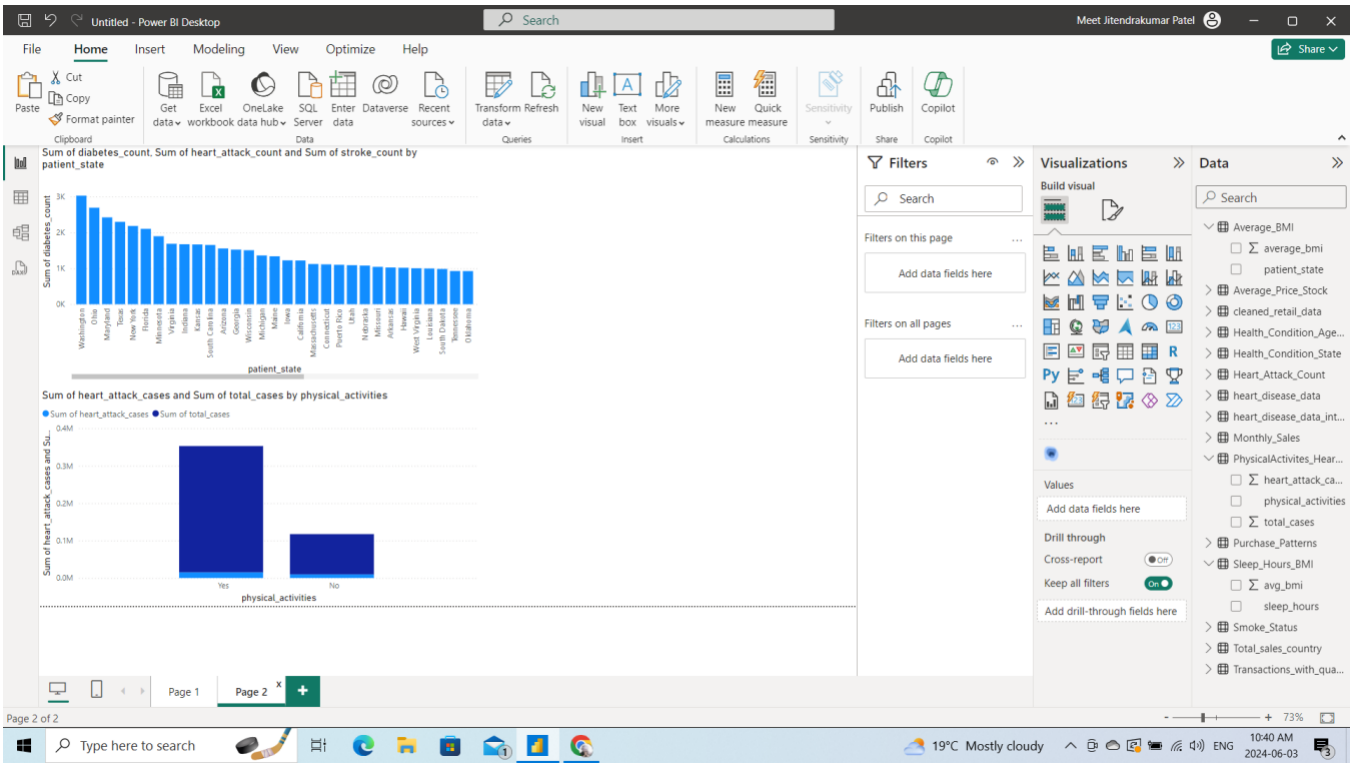
Getting Data from Azure Synapse Analytics




Untitled - Power BI Desktop
Search
Meet Jitendrakumar Patel
File Home Insert Modeling View Optimize Help
Share
Navigator
Display Options
health-synapse.sql.azuresynapse.net: shriji [15]
Average_BMI
Average_Price_Stock
cleaned_retail_data
Health_Condition_Age_Category
Health_Condition_State
Heart_Attack_Count
heart_disease_data
heart_disease_data_internal
Monthly_Sales
PhysicalActivites_Heart_Attack
Purchase_Patterns
Sleep_Hours_BMI
Smoke_Status
Total_sales_country
Transactions_with_quantities
Average_BMI
patient_state average_bmi
California 28
Louisiana 30
North Carolina 29
New Mexico 29
New Hampshire 29
Hawaii 28
District of Columbia 28
Minnesota 29
Pennsylvania 29
Ohio 30
Mississippi 30
Arizona 29
Missouri 29
Georgia 29
Oregon 29
Nebraska 30
Rhode Island 29
Wyoming 29
Indiana 30
Alaska 29
Alabama 30
Connecticut 29
Maryland 29
Virginia 29
Select Related Tables
Load
Transform Data
Cancel
Visualizations
Build visual
Values
Add data fields here
Drill through
Cross-report
Keep all filters
Add drill-through fields here
Filters
Data
Page 1
Page 1 of 1
83%
Type here to search
19°C Mostly cloudy
ENG
10:21 AM
2024-06-03

Creating Different Visualization of the Table in Dedicated SQL pool




Untitled - Power BI Desktop
Search
Meet Jitendrakumar Patel
File Home Insert Modeling View Optimize Help
Share
Paste Cut Copy Format painter
Get data Excel workbook OneLake data hub SQL Server Enter data Dataverse Recent sources
Transform data Refresh
New visual Text box More visuals
New measure Quick measure
Sensitivity
Publish
Copilot
Clipboard Data Queries Insert Calculations Sensitivity Share Copilot
Sum of diabetes_count, Sum of heart_attack_count and Sum of stroke_count by patient_state
Sum of diabetes_count
patient_state
Sum of heart_attack_cases and Sum of total_cases by physical_activities
Sum of heart_attack_cases
Sum of total_cases
physical_activities
Yes
No
Filters
Search
Filters on this page
Add data fields here
Filters on all pages
Add data fields here
Visualizations
Build visual
Values
Add data fields here
Drill through
Cross-report
Keep all filters
Add drill-through fields here
Data
Search
Average_BMI
average_bmi
patient_state
Average_Price_Stock
cleaned_retail_data
Health_Condition_Age...
Health_Condition_State
Heart_Attack_Count
heart_disease_data
heart_disease_data_int...
Monthly_Sales
PhysicalActivites_Hear...
heart_attack_ca...
physical_activities
total_cases
Purchase_Patterns
Sleep_Hours_BMI
avg_bmi
sleep_hours
Smoke_Status
Total_sales_country
Transactions_with_qua...
Page 1
Page 2
Page 2 of 2
73%
Type here to search
19°C Mostly cloudy
ENG
10:40 AM
2024-06-03

Conclusion

My project demonstrated the effectiveness of using data engineering techniques to analyze health related. By cleaning, transforming, and analyzing the data, the project provided valuable insights that can support healthcare decision-making, improve patient outcomes, and inform public health strategies. The use of Azure services, including Data Factory, Data Lake Storage, Databricks, Synapse Analytics, and Power BI, ensured the pipeline's scalability, efficiency, and effectiveness.

GitHub Link

Github