

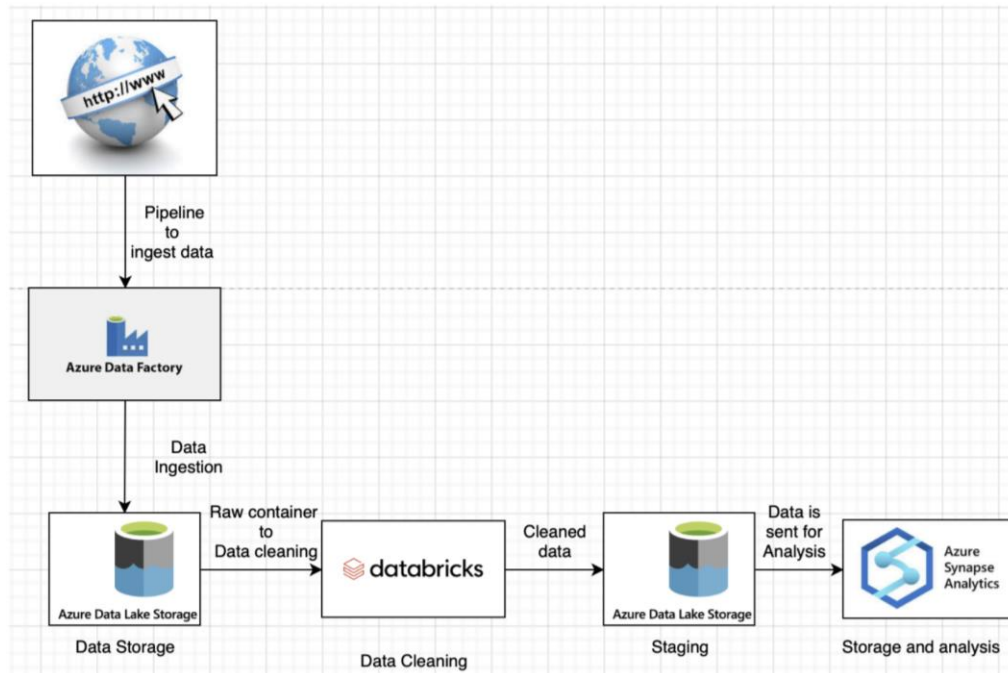
PROJECT-1

Sleep, Health and Life Style.

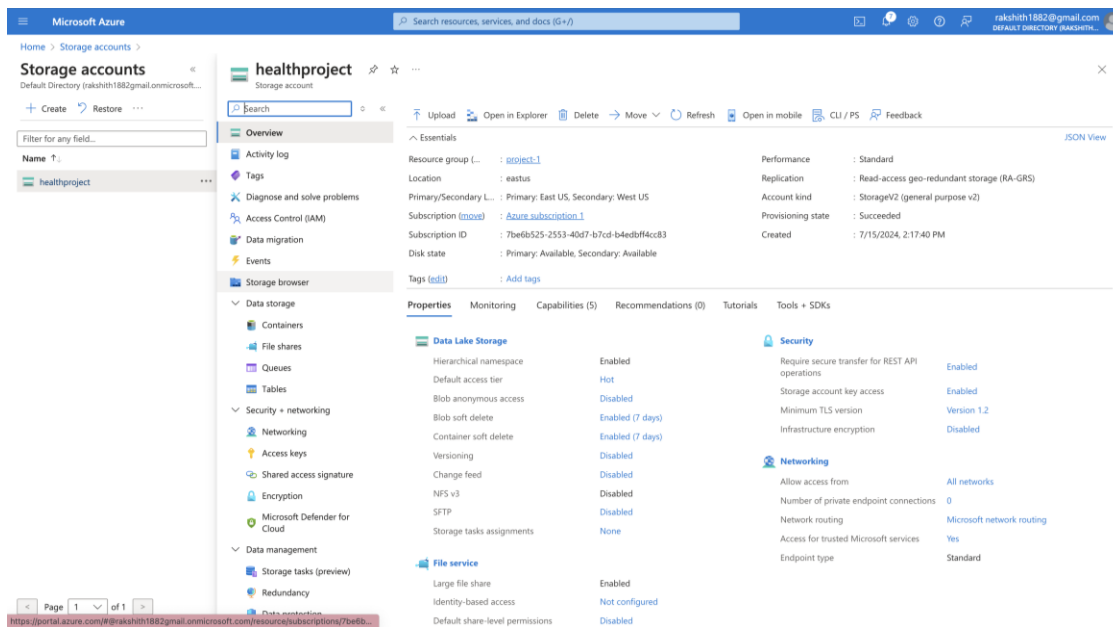
Steps followed:

1. Upload data to Storage account - Raw Container
2. Setup Databricks
 - a. Create data bricks workspace
 - b. Create Azure Service Principal
 - c. Grant access to ADLS Gen2
 - d. Mount ADLS Gen2 in Databricks
3. Read the file to data bricks and perform any transformation
4. Load the data to Synapse Analytics
5. Perform Analysis using SQL in Synapse Analytics
6. Visualize using Power BI

Architectural Diagram:



Create ADLS account:



Create Data factory Account:

Microsoft Azure

Search resources, services, and docs (G+)

Home > Data factories >

Data factories
Default Directory (rakshith1882@gmail.com@microsoft.com)

+ Create Manage view ...

Filter for any field...

Name ↑

datafach

datafach Data factory (V2)

Search

Overview

Activity log

Access control (IAM)

Tags

Diagnose and solve problems

Settings

Getting started

Monitoring

Automation

Help

Delete

Essentials

Resource group (move) : project-1

Status : Succeeded

Location : East US

Subscription (move) : Azure subscription 1

Subscription ID : 7be6b525-2553-40d7-b7cd-b4edbff4cc83

Type : Data factory (V2)

Getting started : [Quick start](#)

JSON View

Azure Data Factory Studio

[Launch studio](#)

Quick Starts

Tutorials

Template Gallery

Training Modules

Monitoring

PipelineRuns

ActivityRuns

Page 1 of 1

Creating a pipeline to copy data from GIT to ADLS:

Details Refresh

[Learn more on copy performance details from here.](#)

Activity run id: 5295712b-aff6-466e-adc7-a4be636d4285

HTTP Succeeded **Azure Data Lake Storage Gen2**
Region: East US

Data read: 24.137 KB
Files read: 1
Peak connections: 1

Data written: 24.137 KB
Files written: 1
Peak connections: 1

Copy duration: 00:00:10
Throughput: 12.068 KB/s

▼ HTTP → Azure Data Lake Storage Gen2

Start time: 7/15/2024, 2:27:44 PM
Used DIUs: 4
Used parallel copies: 1

▼ Duration: 00:00:10

Details	Working duration	Total duration
Queue		00:00:06
Transfer		00:00:02
Listing source	00:00:00	
Reading from source	00:00:00	
Writing to sink	00:00:00	

Data consistency verification: Unsupported

Ingested data:

Microsoft Azure

Search resources, services, and docs (G+/)

Home > Storage accounts > healthproject | Containers > health >

health
Container

Search

Upload + Add Directory ...

Overview

Diagnose and solve problems

Access Control (IAM)

Settings

Authentication method: Access key (Switch to Microsoft Entra user account)

Location: health / ruthvikaa / Project-1 / main

Search blobs by prefix (case-...)

Show deleted objects

Name

[-]

Sleep_health_and_lifestyle_dat...

ruthvikaa/Project-1/main/Sleep_health_and_lifestyle_dataset.csv

Blob

Save Discard Download Refresh Delete

Overview Versions Edit Generate SAS

Person ID	Gender	Age	Occupation	Sleep Duration	Quality of Sleep	Physical Activity Level	Stress Level	BMI Category	Blood Pressure	Heart Rate	Daily Steps	Sleep Disorder
1	Male	27	Software Engineer	6.1	6	42	6	Overweight	126/83	77	4200	None
2	Male	28	Doctor	6.2	6	60	8	Normal	125/80	75	10000	None
3	Male	28	Doctor	6.2	6	60	8	Normal	125/80	75	10000	None
4	Male	28	Sales Representative	5.9	4	30	8	Obese	140/90	85	3000	Sleep Apnea
5	Male	28	Sales Representative	5.9	4	30	8	Obese	140/90	85	3000	Sleep Apnea
6	Male	28	Software Engineer	5.9	4	30	8	Obese	140/90	85	3000	Insomnia
7	Male	29	Teacher	6.3	6	40	7	Obese	140/90	82	3500	Insomnia
8	Male	29	Doctor	7.8	7	75	6	Normal	120/80	70	8000	None
9	Male	29	Doctor	7.8	7	75	6	Normal	120/80	70	8000	None
10	Male	29	Doctor	7.8	7	75	6	Normal	120/80	70	8000	None
11	Male	29	Doctor	6.1	6	30	8	Normal	120/80	70	8000	None
12	Male	29	Doctor	7.8	7	75	6	Normal	120/80	70	8000	None
13	Male	29	Doctor	6.1	6	30	8	Normal	120/80	70	8000	None
14	Male	29	Doctor	6	6	30	8	Normal	120/80	70	8000	None
15	Male	29	Doctor	6	6	30	8	Normal	120/80	70	8000	None
16	Male	29	Doctor	6	6	30	8	Normal	120/80	70	8000	None
17	Female	29	Nurse	6.5	5	40	7	Normal Weight	132/87	80	4000	Sleep Apnea
18	Male	29	Doctor	6	6	30	8	Normal	120/80	70	8000	Sleep

Creating Secrets in Key vault:

Microsoft Azure Upgrade

Search resources, services, and docs (G+/)

Home > key-health-vault

key-health-vault | Secrets

Key vault

Search

+ Generate/Import Refresh Restore Backup View sample code Manage deleted secrets

Overview

Activity log

Access control (IAM)

Tags

Diagnose and solve problems

Access policies

Events

Objects

Keys

Secrets

Certificates

Settings

Access configuration

Networking

Microsoft Defender for Cloud

Properties

Locks

Monitoring

Alerts

Metrics

Diagnostic settings

Logs

Insights

Give feedback

Creating the secret 'SecretID'.
The secret 'SecretID' has been successfully created.

The secret 'SecretID' has been successfully created.

Name	Type	Status	Expiration date
SecretID		✓ Enabled	
ClientID		✓ Enabled	
TenantID		✓ Enabled	

Creating secret scope:

Microsoft Azure

databricks

Q

Search data, notebooks, recents, and more...

+

P

Hdbricks

New

Workspace

Recents

Catalog

Workflows

Compute

SQL

SQL Editor

Queries

Dashboards

Alerts

Query History

SQL Warehouses

Data Engineering

Job Runs

Data Ingestion

Delta Live Tables

Machine Learning

Playground

Experiments

Features

Models

Serving

Marketplace

Partner Connect

Collapse menu

HomePage / Create Secret Scope

Create Secret Scope

Cancel

Create

A store for secrets that is identified by a name and backed by a specific store type. [Learn more](#)

Scope Name

secret-scope

Manage Principal

All workspace users

Azure Key Vault

DNS Name

https://key-health-vault.vault.azure.net/

Resource ID

7be6b525-2553-40d7-b7cd-b4edbf4cc83

Configuration:

Microsoft Azure

databricks

Q

Search data, notebooks, recents, and more...

+

P

Untitled Notebook 2024-07-18 15:07:17

Python

File

Edit

View

Run

Help

Last edit was 4 minutes ago

Provide feedback

Run all

rakshith L's Cluster

Schedule

Share

New

Workspace

Recents

Catalog

Workflows

Compute

SQL

SQL Editor

Queries

Dashboards

Alerts

Query History

SQL Warehouses

Data Engineering

Job Runs

Data Ingestion

Delta Live Tables

Machine Learning

Playground

Experiments

Features

Models

Serving

Marketplace

Partner Connect

Collapse menu

4 minutes ago (2s)

1

client_id = dbutils.secrets.get(scope="secret-scope",key="ClientID")

tenant_id = dbutils.secrets.get(scope="secret-scope",key="TenantID")

client_secret = dbutils.secrets.get(scope="secret-scope",key="Clientsecret")

3 minutes ago (1s)

2

Python

configs = {"fs.azure.account.auth.type": "OAuth",

"fs.azure.account.oauth.provider.type": "org.apache.hadoop.fs.azurebfs.oauth2.ClientCredsTokenProvider",

"fs.azure.account.oauth2.client.id": client_id,

"fs.azure.account.oauth2.client.secret": client_secret,

"fs.azure.account.oauth2.client.endpoint": f"https://login.microsoftonline.com/{tenant_id}/oauth2/token"}

2 minutes ago (1s)

3

spark.conf.set("fs.azure.account.auth.type.healthproject.dfs.core.windows.net", "OAuth")

spark.conf.set("fs.azure.account.oauth.provider.type.healthproject.dfs.core.windows.net", "org.apache.hadoop.fs.azurebfs.oauth2.ClientCredsTokenProvider")

spark.conf.set("fs.azure.account.oauth2.client.id.healthproject.dfs.core.windows.net", client_id)

spark.conf.set("fs.azure.account.oauth2.client.secret.healthproject.dfs.core.windows.net", client_secret)

spark.conf.set("fs.azure.account.oauth2.client.endpoint.healthproject.dfs.core.windows.net", f"https://login.microsoftonline.com/{tenant_id}/oauth2/token")

1 minute (14s)

4

dbutils.fs.mount(

source = "abfss://health@healthproject.dfs.core.windows.net/",

mount_point = "/mnt/healthproject/health",

extra_configs = configs)

True

Define schema

health-project Python File Edit View Run Help Last edit was 1 minute ago Provide feedback Run all rakshith L's Cluster Schedule Share

```

6
from pyspark.sql import SparkSession
from pyspark.sql.functions import split
from pyspark.sql.types import StructType, StructField, IntegerType, StringType, DoubleType

7
health_schema = StructType([
    StructField("Person ID", IntegerType(), False),
    StructField("Gender", StringType(), False),
    StructField("Age", IntegerType(), False),
    StructField("Occupation", StringType(), False),
    StructField("Sleep Duration", DoubleType(), False),
    StructField("Quality of Sleep", IntegerType(), False),
    StructField("Physical Activity Level", IntegerType(), False),
    StructField("Stress Level", IntegerType(), False),
    StructField("BMI Category", StringType(), False),
    StructField("Blood Pressure", DoubleType(), False),
    StructField("Heart Rate", IntegerType(), False),
    StructField("Daily Steps", IntegerType(), False),
    StructField("Sleep Disorder", StringType(), False),
    StructField("Systolic", IntegerType(), False)
])

8
health_df = spark.read.option("Header", True).schema(health_schema).csv("/mnt/healthproject/health/ruthvikaa/Project-1/main/Sleep_health_and_Lifestyle_dataset/Sleep_health_and_Lifestyle_dataset (1).csv")
health_df: pyspark.sql.dataframe.DataFrame = [Person ID: integer, Gender: string ... 12 more fields]

```

Read CSV File to Data bricks :

health-project Python File Edit View Run Help Last edit was 2 minutes ago Provide feedback Run all rakshith L's Cluster Schedule Share

```

9
display(health_df)

```

(1) Spark Jobs

	Activity Level	Stress Level	BMI Category	Blood Pressure	Heart Rate	Daily Steps	Sleep Disorder	Systolic
6	30	8	Obese	na	85	3000	Insomnia	na
7	40	7	Obese	na	82	3500	Insomnia	na
8	75	6	Normal	na	70	8000	None	na
9	75	6	Normal	na	70	8000	None	na
10	75	6	Normal	na	70	8000	None	na
11	30	8	Normal	na	70	8000	None	na
12	75	6	Normal	na	70	8000	None	na
13	30	8	Normal	na	70	8000	None	na
14	30	8	Normal	na	70	8000	None	na
15	30	8	Normal	na	70	8000	None	na
16	30	8	Normal	na	70	8000	None	na
17	40	7	Normal Weight	na	80	4000	Sleep Apnea	na
18	30	8	Normal	na	70	8000	Sleep Apnea	na
19	40	7	Normal Weight	na	80	4000	Insomnia	na
20	75	6	Normal	na	70	8000	None	na

374 rows | 1.22 seconds runtime Refreshed 32 minutes ago

Selected Required Columns:

health-project Python ☆

File Edit View Run Help Last edit was 3 minutes ago Provide feedback

Run all rakshith L's Cluster Schedule Share

111 8 60 4 Normal 68 7000 None null

112 8 60 5 Normal 68 8000 None null

113 8 60 4 Normal 68 7000 None null

374 rows | 1.15 seconds runtime

Refreshed 10 minutes ago

```

from pyspark.sql.functions import col
df1 = health_df.select(col("Person ID"), col("Gender"), col("Age"), col("Sleep Duration"), col("Physical Activity Level"), col("Stress Level"),
col("BMI Category"), col("Heart Rate"), col("Sleep Disorder"))

df1: pyspark.sql.dataframe.DataFrame = [Person ID: integer, Gender: string ... 7 more fields]

```

```

df1.write.mode("overwrite").parquet("/mnt/healthproject/health-processed/health")

```

(1) Spark Jobs

```

display(spark.read.parquet("/mnt/healthproject/health-processed/health"))

```

(2) Spark Jobs

id	Age	Sleep Duration	Physical Activity Level	Stress Level	BMI Category	Heart Rate	Sleep Disorder
4	28	5.9	30	8	Obese	85	Sleep Apnea
5	28	5.9	30	8	Obese	85	Sleep Apnea
6	28	5.9	30	8	Obese	85	Insomnia
7	29	6.3	40	7	Obese	82	Insomnia
8	29	7.8	75	6	Normal	70	None
9	29	7.8	75	6	Normal	70	None
10	29	7.8	75	6	Normal	70	None
11	29	6.1	30	8	Normal	70	None

Wrote the file to adls processed container in parquet format :

health-project Python ☆

File Edit View Run Help Last edit was 4 minutes ago Provide feedback

Run all rakshith L's Cluster Schedule Share

```

df1.write.mode("overwrite").parquet("/mnt/healthproject/health-processed/health")

```

(1) Spark Jobs

```

display(spark.read.parquet("/mnt/healthproject/health-processed/health"))

```

(2) Spark Jobs

id	Age	Sleep Duration	Physical Activity Level	Stress Level	BMI Category	Heart Rate	Sleep Disorder
4	28	5.9	30	8	Obese	85	Sleep Apnea
5	28	5.9	30	8	Obese	85	Sleep Apnea
6	28	5.9	30	8	Obese	85	Insomnia
7	29	6.3	40	7	Obese	82	Insomnia
8	29	7.8	75	6	Normal	70	None
9	29	7.8	75	6	Normal	70	None
10	29	7.8	75	6	Normal	70	None
11	29	6.1	30	8	Normal	70	None
12	29	7.8	75	6	Normal	70	None
13	29	6.1	30	8	Normal	70	None
14	29	6	30	8	Normal	70	None
15	29	6	30	8	Normal	70	None
16	29	6	30	8	Normal	70	None
17	29	6.5	40	7	Normal Weight	80	Sleep Apnea
18	29	6	30	8	Normal	70	Sleep Apnea

374 rows | 1.91 seconds runtime

Refreshed 4 minutes ago

[Shift+Enter] to run and move to next cell
[Esc H] to see all keyboard shortcuts

Created a table in synapse

Microsoft Azure | Synapse Analytics | synapsehealth

We use optional cookies to provide a better experience. [Learn more](#)

Accept Reject More options

Synapse live Validate all Publish all

Data

Workspace Linked

Filter resources by name

- Azure Data Lake Storage Gen2 3
- synapsehealth (Primary - healthpro...
- health-processed (Primary)
- health
- (Attached Containers)
- healthstorage (healthproject)

health-processed SQL script 1

Other users in your workspace may have access to modify this item. Do not use this item unless you trust all users who may have access to the workspace.

Connect to sqlpool Use database sqlpool

```

1 IF NOT EXISTS (SELECT * FROM sys.external_file_formats WHERE name = 'SynapseParquetFormat')
2 CREATE EXTERNAL FILE FORMAT [SynapseParquetFormat]
3 WITH (FORMAT_TYPE = PARQUET)
4 GO
5
6 IF NOT EXISTS (SELECT * FROM sys.external_data_sources WHERE name = 'health-processed_healthproject_dfs_core_windows_net')
7 CREATE EXTERNAL DATA SOURCE [health-processed_healthproject_dfs_core_windows_net]
8 WITH (
9     LOCATION = 'abfss://health-processed@healthproject.dfs.core.windows.net'
10 )
11 GO
12
13 CREATE EXTERNAL TABLE dbo.healthexternaltable (
14     [Person ID] int,
15     [Gender] nvarchar(4000),
16     [Age] int,
17     [Sleep Duration] float,
18     [Physical Activity Level] int,
19     [Stress Level] int,
20     [BMI Category] nvarchar(4000),
21     [Heart Rate] int
22 )
23 WITH (
24     LOCATION = 'abfss://health-processed@healthproject.dfs.core.windows.net'
25 )
26
27 SELECT * FROM dbo.healthexternaltable
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99

```

Results Messages

View Table Chart Export results

Search

Person ID	Gender	Age	Sleep Duration	Physical Activity Level	Stress Level	BMI Category	Heart Rate
1	Male	27	6.1	42	6	Overweight	77
2	Male	28	6.2	60	8	Normal	75
3	Male	28	6.2	60	8	Normal	75
4	Male	28	5.9	30	8	Obese	85
5	Male	28	5.9	30	8	Obese	85

00:00:03 Query executed successfully.

Properties

General Related (0)

Name * SQL script 1

Description

Type .sql script

Size 943 bytes

Results settings per query

First 5000 rows (default)

All rows

AGGREGATIONS:

Synapse live Validate all Publish all

Data

Workspace Linked

Filter resources by name

- Azure Data Lake Storage Gen2 3
- synapsehealth (Primary - healthpro...
- health-processed (Primary)
- health
- (Attached Containers)
- healthstorage (healthproject)

health-processed SQL script 1

Other users in your workspace may have access to modify this item. Do not use this item unless you trust all users who may have access to the workspace.

Connect to sqlpool Use database sqlpool

```

39 SELECT
40     CASE
41         WHEN Age BETWEEN 28 AND 29 THEN '28-29'
42         WHEN Age BETWEEN 30 AND 39 THEN '30-39'
43         WHEN Age BETWEEN 40 AND 49 THEN '40-49'
44         ELSE '50+'
45     END AS Age_Group,
46     AVG([Heart Rate]) AS Average_Heart_Rate
47 FROM dbo.healthexternaltable
48 GROUP BY Age;
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99

```

Results Messages

View Table Chart Export results

Search

Person ID	Gender	Age	Sleep Duration	Physical Activity Level	Stress Level	BMI Category	Heart Rate
1	Male	27	6.1	42	6	Overweight	77
2	Male	28	6.2	60	8	Normal	75
3	Male	28	6.2	60	8	Normal	75
4	Male	28	5.9	30	8	Obese	85
5	Male	28	5.9	30	8	Obese	85

00:00:01 Query executed successfully.