Exercise 1

Thursday, June 2, 2022

5:06 PM

In this lab you will explore data using the engine of your choice (SQL or Spark).

Understanding data through data exploration is one of the core challenges faced today by data engineers and data scientists as well. Depending on the data's underlying structure and the specific requirements of the exploration process, different data processing engines will offer varying degrees of performance, complexity, and flexibility.

In Azure Synapse Analytics, you have the possibility of using either the SQL Serverless engine, the big-data Spark engine, or both.

* + Explore the Data Lake with serverless SQL Pool and Spark
    - Task 1 - Explore the Data Lake with Azure Synapse serverless SQL pool
    - Task 2 - Explore the Data Lake with Azure Synapse Spark

Task 1 - Explore the Data Lake with Azure Synapse serverless SQL pool

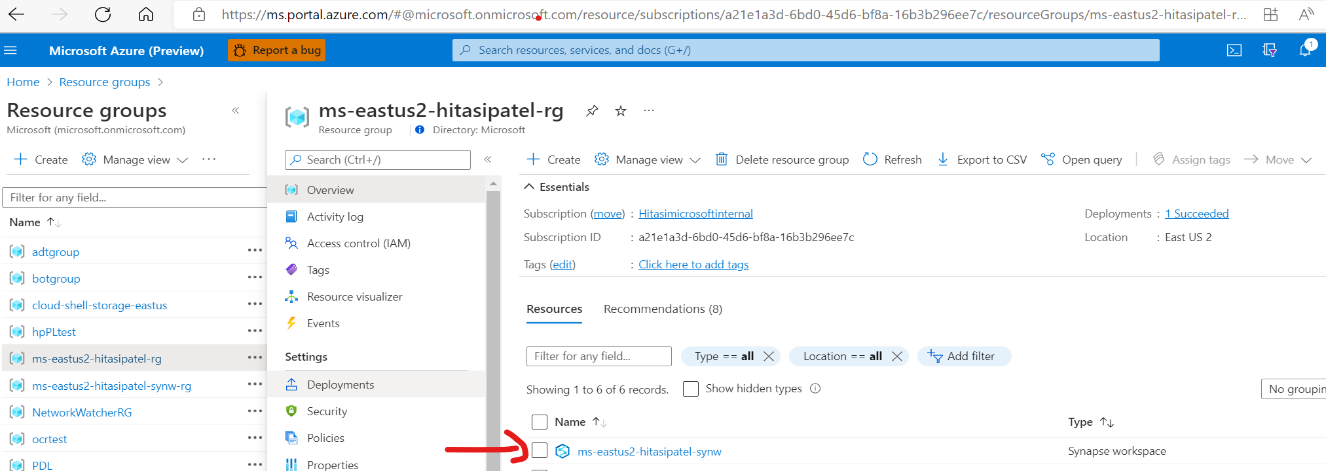
* + In a Microsoft Edge web browser, navigate to the [Azure portal](https://portal.azure.com/) (<https://portal.azure.com>) and login with your credentials. Then select Resource groups.

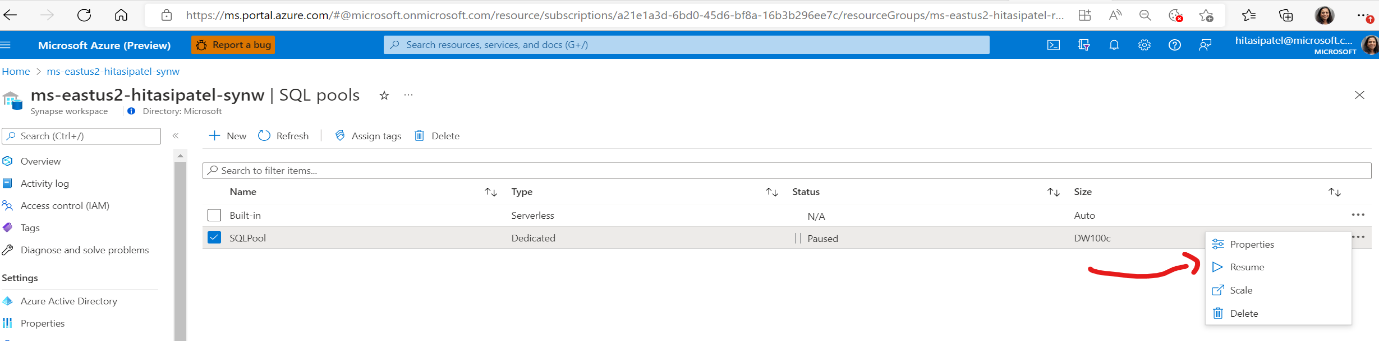
Azure services 
Create a 
resource 
Resource 
groups 
Azure Synapse 
Analytics 
Cost analysis 
(preview) 
o 
Azure Synapse 
Analytics... 
Sto rage 
accounts 
Managed 
Identities 
Virtual 
Key vaults 
More services 

* + select the Synapse Analytics resource group.

* + (O) ms-eastus2-hitasipatel-rg 
    Resource group 
    a few seconds ago 

* + Verify SQLPool is running by going into synapse workspase.

* 1. 

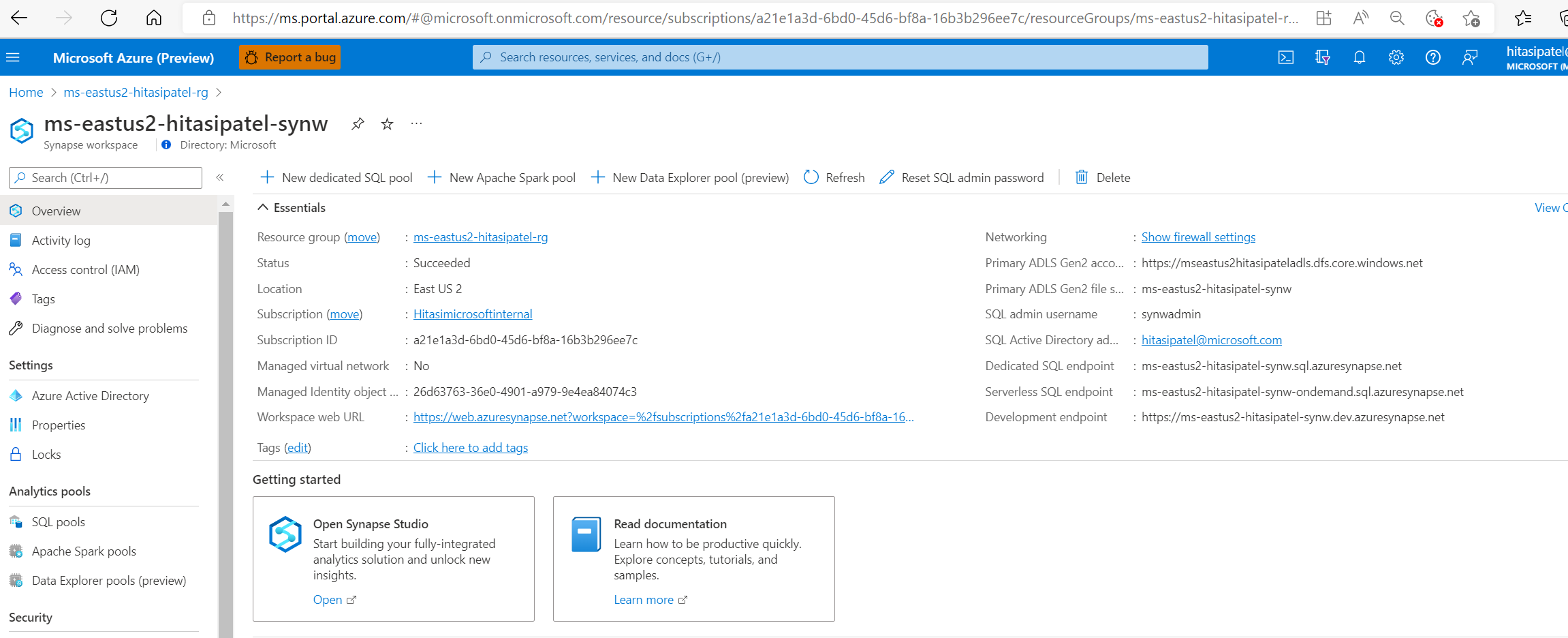
* 1. 

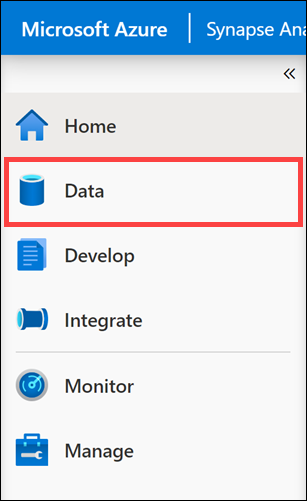
* 1. Download the sample file [RetailSales.csv](https://github.com/Azure-Samples/Synapse/blob/main/Notebooks/PySpark/Synapse%20Link%20for%20Cosmos%20DB%20samples/Retail/RetailData/RetailSales.csv) and upload it to the container.
     1. Select storage account created by scripts --> select bronze container -->select upload file and upload RetailSales.csv

Switch to the Linked tab (1). Under Azure Data Lake Storage Gen2 (2), expand the primary data lake storage account, and then select the bronze file system (3).

Home > ms-eastus2-hitasipatel-rg > mseastus2hitasipateladls > 
bronze 
Overview 
Diagnose and solve problems 
Access Control (I AM) 
Settings 
Shared access tokens 
Manage ACL 
Access oolicv 
Upload Add Directory Refresh Rename 
Authentication method: Access key (Switch to Azure AD User Account) 
Location: bronze 
Search blobs by prefix (case-sensitive) 
data. csv 
Sales.Orders.csv 
Delete 
Change tier 
Modified 
Acquire lease 
Break lease 
Access tier 
Hot (Inferred) 
Hot (Inferred) 
Upload blob 
bronze/ 
Files @ 
Select a file 
Overwrite if files already exist 
v Advanced 
Show deleted Objec 
Archive status 
310b type 
Block blob 
Block blob 
Size 
Upload 
6/3/2022, PM 
6/3/2022, PM 

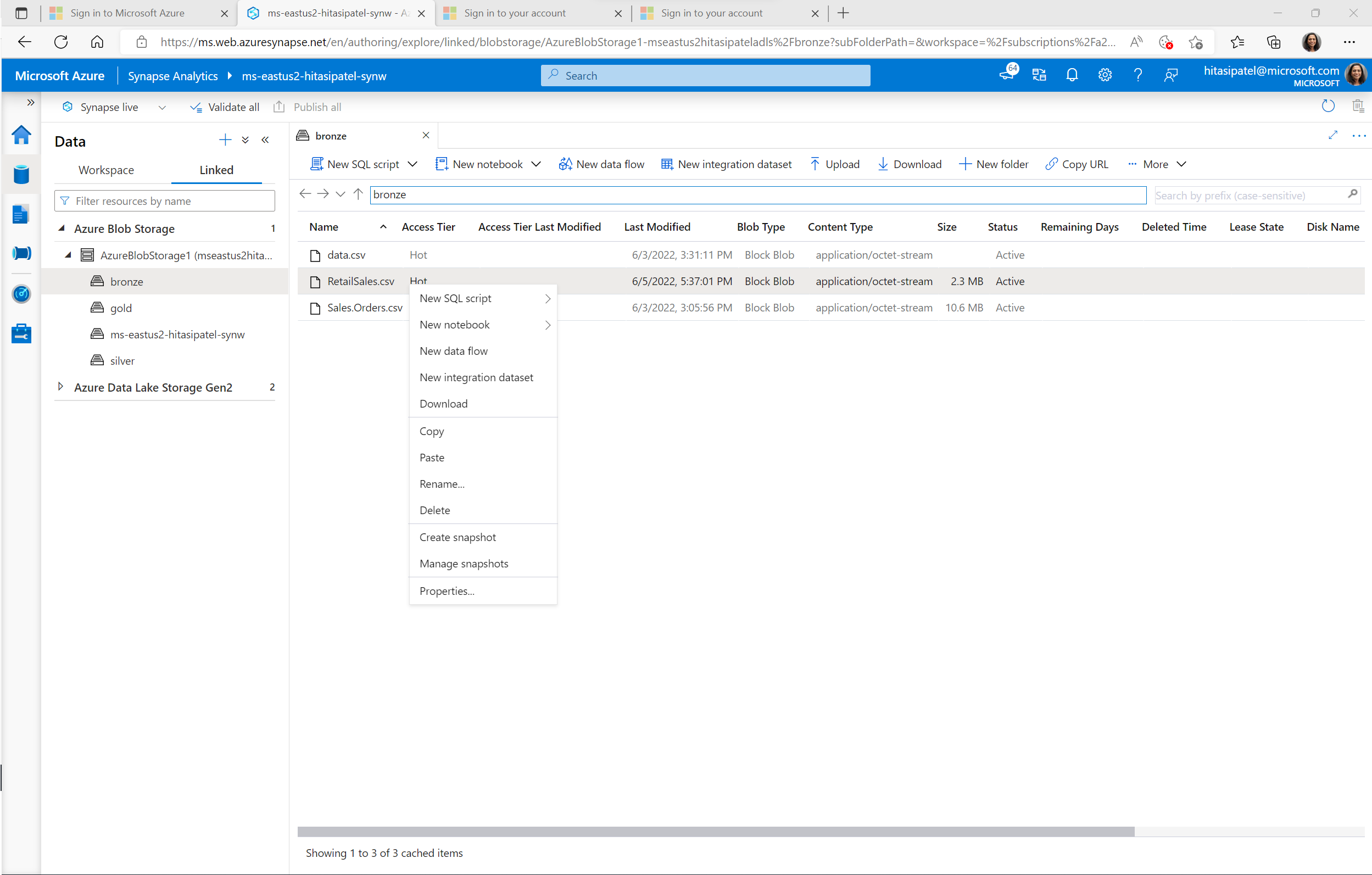
* 1. On the Synapse workspace blade, open Synapse Analytics Studio by navigating to the Workspace web URL from the overview page



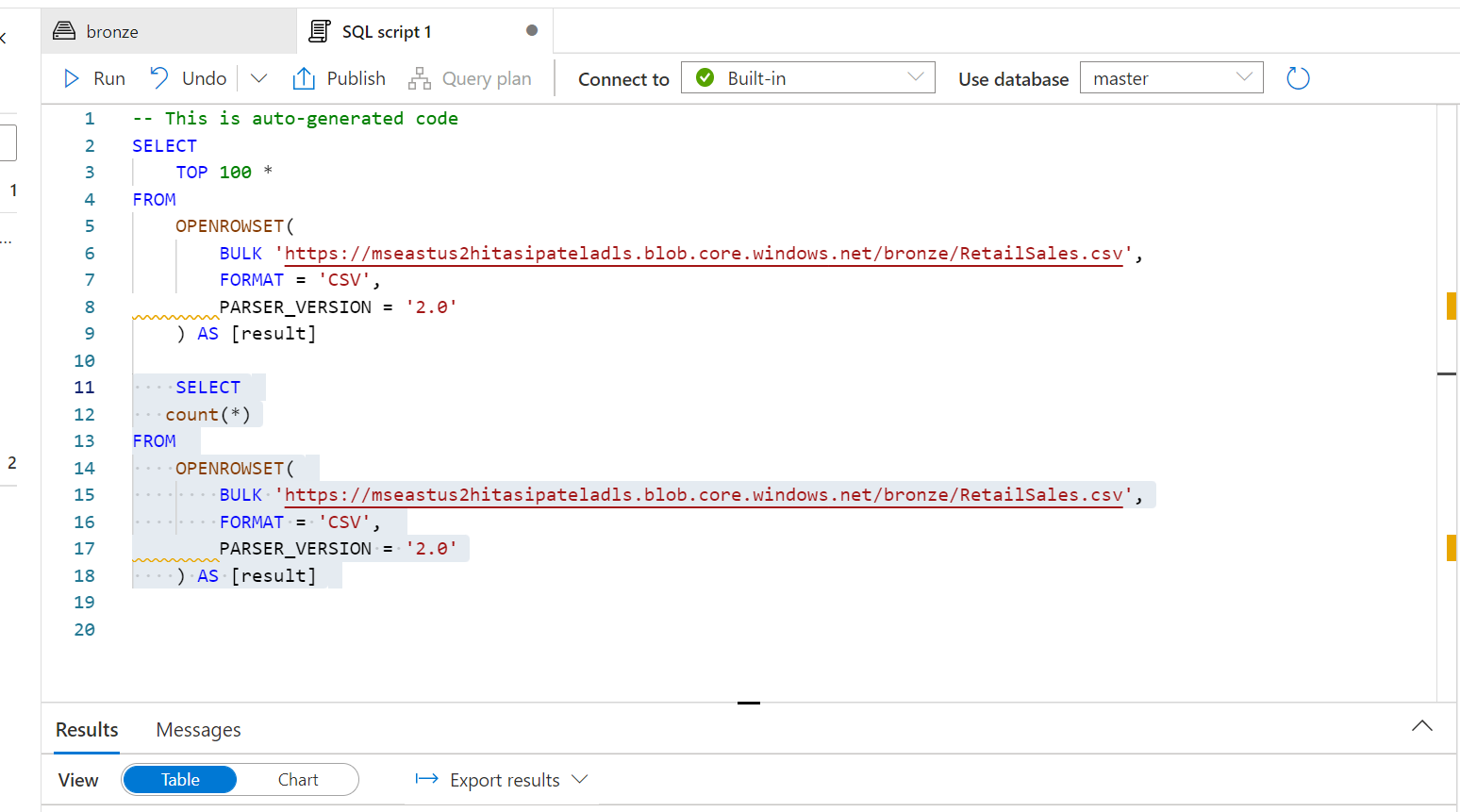
* 1. In Synapse Analytics Studio, navigate to the Data hub and locate Linked service and double click Bronze.
* [](https://github.com/solliancenet/azure-synapse-analytics-day/blob/master/media/data-hub.png)

Microsoft Azure ms-eastus2-hitasipatel-synw 
Validate all 
O Synapse live 
Data 
p Search 
Publish all 
e bronze 
New SQL script v New notebook v New data flow New integration dataset 
v bronze 
hitasi 
Upload 
Content Type 
Download New folder 
Workspace 
Linked 
C.P copy G Select all v 
Y 
Access Tier Last Modified 
Last Modified 
Blob Type 
Size 
D) 
Filter resources by name 
Azure Blob Storage 
AzureBIobStorage1 (mseastus2hita... 
bronze 
gold 
silver 
Azure Data Lake Storage Gen2 
Name 
data.csv 
A Access Tier 
Hot 
Status 
Active 
Remaining Days 
Deleted Time 
6/3/2022, PM Block Blob 
6/3/2022, PM Block Blob 
application/octet-stream 
Sales.Orders.csv Hot 
application/octet-stream 10.6 MB Active 

Once you are inside folder click on 'New SQL Script'

* 1. 

* 1. Let us change the initial script to load multiple Parquet files at once.
     + In line 2, replace TOP 100 \* with COUNT(\*).

* 1. 

* 1. Go back to Linked service --> select bronze --> highlight retail.csv file and click Notebook and then to External table

Workspace 
Linked 
Filter resources by name 
Azure Blob Storage 
AzureBlobStorage1 (mseastus2hita.. 
bronze 
gold 
ms-eastus2-hitasipatel-synw 
silver 
Azure Data Lake Storage Gen2 
ms-eastus2-hitasipatel-synw (Prim... 
e ms-eastus2-hitasipatel-synw (Pri,__ 
e bronze 
e gold 
e silver 
(Attached Containers) 
Na me 
datacsv 
RetailSales.cs- 
Access Tier 
Hot 
Access Tier Last Modified 
Last Modified 
6/3/2022, PM 
6/5/2022, PM 
6/3/2022, PM 
Blob Type 
Block Blob 
Block Blob 
Block Blob 
Content Type 
application/octet-stream 
application/octet-stream 
application/octet-stream 
Size 
2.3 MB 
10.6 MB 
Status 
Active 
Active 
Active 
Rema ining Days 
Deleted Time 
New SQL script 
SalesOrders 
New notebook 
New data flow 
New integration dataset 
Download 
Paste 
Rename„. 
Delete 
Create snapshot 
Manage snapshots 
Properties„. 

* 1. Select SQL Pool from drop down and execute scripts.

* 1. SQL script 1 
     Run all v 
     Not started 
     Undo 
     SQL script 2 
     It-I Publish 
     Outline 
     SQL script 3 
     Attach to 
     Notebook 2 
     SparkP00101 
     Language 
     ms-eastus2-hitasipat„. 
     pySpark (Python) v 
     e bronze 
     Variables 
     format: csv' 
     Notebook 3 
     4 
     8 
     12 
     14 
     15 
     16 
     %%pyspark 
     blob container name = "bronze" 
     from pyspark. sqI import SparkSession 
     sc = builder. getOrCreate() 
     token_library = sc._jvm.com.microsoft.azure.synapse.tokenlibrary.TokenLibrary 
     = token_library. "Azure310bStoragei") 
     spark. conf. set( 
     • core.windows.net• % (blob_container_name, blob_account_name), 
     df spark. read. load( 'wasbs: // bronze—seastus2hitasipateIadIs. blob. core. windows. net/RetaiISaIes. csv' , 
     header-True 
     display(df. limit(lø)) 
     Press shift + enter to run 
     + Code 
     + Markdown 

* 1. Observe External table created in Lake Database under Workspace.

Data 
Workspace 
Linked 
Y Filter resources by name 
Lake database 
default 
t Tables 
retail 
Columns 
storeld (string) 
productCode (string) 
quantity (string) 
logQuantity (string) 
advertising (string) 
price (string) 
weekStarting (string) 
id (string) 
Properties 
yourtablename 
Views 

* 1. Explore Data file with Spark Pool -This time select Notebook and then Load to Dataframe or create new cell and add following code.

' csv' 
7 
15] 
spark. conf. set( 
core.windows.net• % 
df = spark. read. load( 
display(df. ) 
3 sec - Command executed in 2 sec 815 ms by hitasipatel on 3:29:28 PM, 6/15/22 
Table 
storeld 
Chart 
header—True 
advertising 
Export results v 
productCode 
surface.go 
surface-go 
surface.go 
surface-go 
surface.go 
surface-go 
surface.go 
surface-go 
surface.go 
surface-go 
quantity 
105 
80 
68 
28 
16 
253 
107 
66 
65 
17 
logQuantity 
9.264828557 
8987196821 
8.831711918 
7.965545573 
7.377758908 
10.1402973 
9.282847063 
8803273983 
8.793612072 
7.454719949 
pnce 
159 
189 
189 
189 
179 

* 1. Configure Spark Cluster.

Exercise 2

Wednesday, June 15, 2022

5:02 PM

In this lab, you will use a pipeline to copy data from blob storage to Sql DB, transform it, and load it into the Azure SQL DB. You will also monitor the progress of the associated tasks.

1. Access Sales DB from Portal and Management Studio

mseastus2hitasipateIsqIsrvr I SQL databases 
0 
SQL server 
Directory: Microsoft 
P Search (Ctrl*" 
Overview 
Activity log 
Access control (IAM) 
Tags 
Diagnose and solve problems 
Quick Start 
Settings 
Azure Active Directory 
SQL databases 
SQL elastic pools 
OTC' quota 
Properties 
Locks 
Data management 
Search to filter databases.„ 
Database 
SalesDB 
Status 
Online 
Pricing tier 
Standard S2: 50 DTI's 

•i SalesDB (mseastus2hitasipateIsqIsrvr/SaIesDB) I Query editor (preview) 
O Directory: Microsoft 
SQL database 
Search (Ctrl±/) 
Overview 
Activity log 
Tags 
Diagnose and solve problems 
Getting started 
Query editor (preview) 
power Platform 
Power Bl 
Power Apps 
Power Automate 
ngs 
Compute storage 
Connection strings 
Maintenance 
Properties 
manag ement 
Replicas 
Login New Query Open query Feedback 
SQL server authentication 
Login * 
synwadmin 
Password * 
SQL 
Welcome to SQL Database Query Editor 
Active Directory authentication 
Continue as hitasipatel@microsoftcom 

1. Create dbo.Retailtable on query editor

% SalesDB (mseastus2hitasipateIsqIsrvr/SaIesDB) I Query editor (preview) 
o 
Directory: Microsoft 
SQL database 
z' Search (Ctrl +1') 
Overview 
Activity log 
Tags 
Diagnose and solve problems 
Getting started 
Query editor (preview) 
'wer Platform 
Power BI 
Power Apps 
Power Automate 
Compute 4 storage 
Connection strings 
Maintenance 
Properties 
Locks 
management 
« Login -+- New Query Open query Feedback 
SalesDB (synwadmin) 
CD 
Query 1 
Run Cancel query 
Save query 
CREATE TABLE [retailtable] 
[storeid] [int]" 'NULL, 
[productcode] [vat-char] (50) 
(Quantity] [int] 
[logquantity] [vat-char] (50) 
(advertising] [varcharl(50) 
[price] [int] NULL, 
[weekStartingl [date] •NULL 
Export data as v 
.NULL, 
NULL, 
NULL, 
Show only Editor 
Showing limited object explorer here. For 
full capability please open SSDT. 
Tables 
> dbo_BuiIdVersion 
> dbo_ErrorLog 
> dbo.retailtable 
> SalesLT.Address 
> SalesLT.Customer 
> SalesLT.CustomerAddress 
> SalesLT.Product 
> SalesLT.ProductCategory 
> SalesLT.ProductDescription 
> SalesLT.ProductModel 
> SalesLT.ProductModelProductOesc • • 
Results Messages 
> salesLT.salesorderoetail 
> SalesLT.SalesOrderHeader 
Query succeeded: Affected rcnqs: 
> Stored Procedures 

CREATE TABLE [dbo].[retailtable]

(

    [storeid] [int]  NULL,

    [productcode] [varchar](50)  NULL,

    [Quantity] [int]  NULL,

    [logquantity] [varchar](50)  NULL,

    [advertising] [varchar](50)  NULL,

    [price] [int]  NULL,

    [weekStarting] [date]  NULL

)

1. Create Link service from Synapase to connect to SalesDB data base .

New linked service 
Azure SQL Database Learn more 
o 
Choose a name for your linked service. This name cannot be updated later. 
Name * 
Description 
Connect via integration runtime @ 
AutoResolvelntegrationRuntime 
Azure Key Vault 
Connection string 
Account selection method @ 
(D From Azure subscription @ Enter manually 
Fully qualified domain name 
mseastus2hitasipatelsqlstvr.database.windows_net 
Database name 
salesdb 
Authentication type 
SQL authentication 
user name 
synwadmin 
Password 
Password * 
Azure Key Vault 
Always encrypted 
Additional connection properties 
+ New 
Create 
Connection successful 
Test connection 
Cancel 

1. Create New Pipeline to copy data from Blobstorage ( bronze/retail.csv) to Salesdb.dbo.retailtable

O Synapse live 
Integrate 
Y Filter resources by name 
Pipelines 
Validate all 
Publish all 
SourceDataset 
Bulk Copy from File„. 
N !ure Synapse Analytics 
tureSqlDWSinkDataset 
AzureSqlDWSinkDat.. 
x 
D) 
Bulk Copy from Files to Database New pipeline 
New folder 

Source Dataset 
Activities 
P copy 
v Move & transform 
Copy data 
Bulk Copy from Fil... 
V Validate 
AzureSqlDWSinkDat.. 
Validate copy runtime Debug 
( x.py • data 
Lake Database lab2 
Add trigger 
O 
Pipeline I 
General 
Name * 
Source 
Description 
Timeout O 
Retry @ 
Retry interval (sec) O 
Secure output @ 
Secure input O 
Copy data pipeline 
lab2 
Sink Mapping 
Settings User properties 
Learn more c.i 
30 

1. Validate Pipeline and Publish it
2. Select Add trigger --> Trigger Now to execute it
3. Click on Monitor icon and select pipeline from list to check detail.

Microsoft Azure 
ms-eastus2-hitasipatel-synw 
Analytics pools 
SQL pools 
Apache Spark pools 
Data Explorer pools (preview) 
Activities 
SQL requests 
KQL requests 
Apache Spark applications 
Data flow debug 
I ntegration 
Pipeline runs 
Trigger runs 
Pipeline runs 
Triggered Debug 
Filter by run ID or name 
Add filter 
Showing 1 - 5 items 
pipeline name 
Rerun 
S Ca ncel v C_) Refresh 
Local time Last 24 hours 
Run start 
Jun 15, 2022, pm 
Jun 15, 2022, pm 
Jun 15, 2022, pm 
Jun 15, 2022, pm 
Jun 15, 2022, pm 
Edit columns 
Pipeline name All 
Run end 
Jun 15, 2022, 6:20:24 pm 
Jun 15, 2022, pm 
Jun 15, 2022, pm 
Jun 15, 2022, pm 
Jun 15, 2022, 6:13:42 pm 
Gantt 
Status All 
Duration 
00:00:12 
Runs : Latest runs 
Triggered by 
Manual trigger 
Manual trigger 
Manual trigger 
Manual trigger 
Manual trigger 
Triggered by : All 
Status 
e Succeeded 
O Failed 
O Failed 
O Failed 
Failed 
Error 
Bulk Copy from Files to Data... 
Bulk Copy from Files to Data... 
Bulk Copy from Files to Data.. 
Bulk Copy from Files to Data... 
Bulk Copy from Files to 

Exercise 3

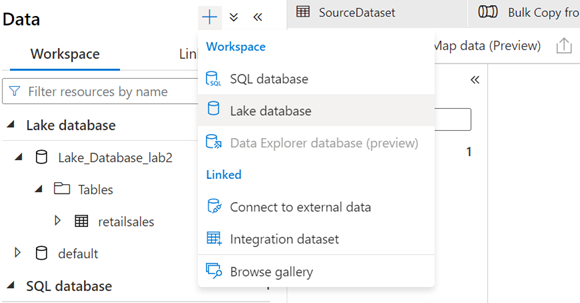
Wednesday, June 15, 2022

10:53 PM

In lab, you will explore the concept of a lake database and you will learn how to use readily available database templates for lake databases.

The lake database in Azure Synapse Analytics enables you to bring together database design, meta information about the data that is stored and a possibility to describe how and where the data should be stored. Lake database addresses the challenge of today's data lakes where it is hard to understand how data is structured.

* 1. Click on Data --> Workspace and Lake Database



* 1. Select Link service --> Blob Storage --> folder and file as below

senice ' 
G•mal R ed § 
Démited 「 : 1 
0 0 n ~ , Database 
until itis - d 
this - 一 & 313W 

* 1. This should create table as below. You can add or remove column and change data types

