



Independent Study

Building a Scalable Data Migration and Transformation Pipeline Using Azure

Sakshi Basapure
200534164

Problem Statement



Develop a cloud-based solution that automates the ingestion of on-premise data, transforms it into a clean and structured format, and stores it in a scalable cloud environment.

Objectives and BRs



BR 1: Ensure secure automated **data migration** between on-premise systems and cloud.

BR 2: Develop a **transformation** layer to convert raw data into clean, structured formats.

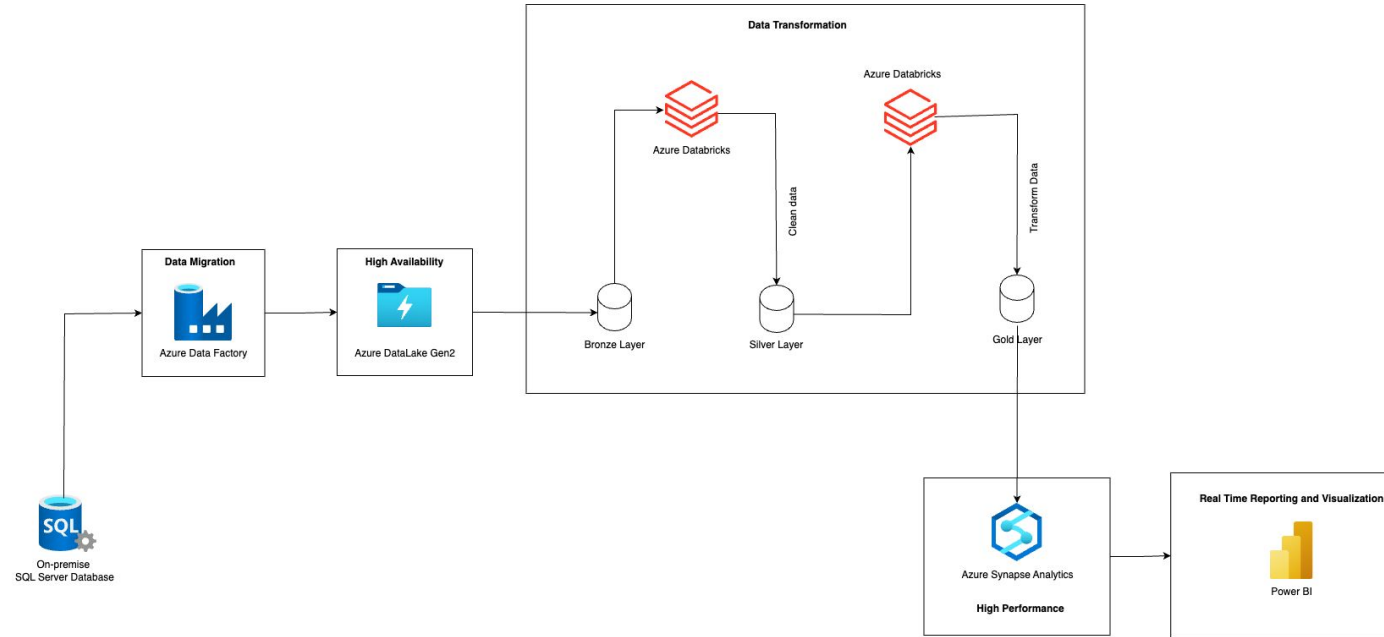
BR 3: Maintain **high availability** throughout the data pipeline.

BR 4: Maintain **high performance** throughout the data pipeline.

BR 5: Enforce **governance and compliance** with **data security** regulations.

BR 6: Implement **real-time reporting and visualization** capabilities, providing business users with actionable insights through intuitive dashboards.

Architecture



Implementation



BR 1: Ensure secure **automated data migration** between on-premise systems and cloud.

TR 1.2: Utilize a secure connection method to connect to on-premise data sources.

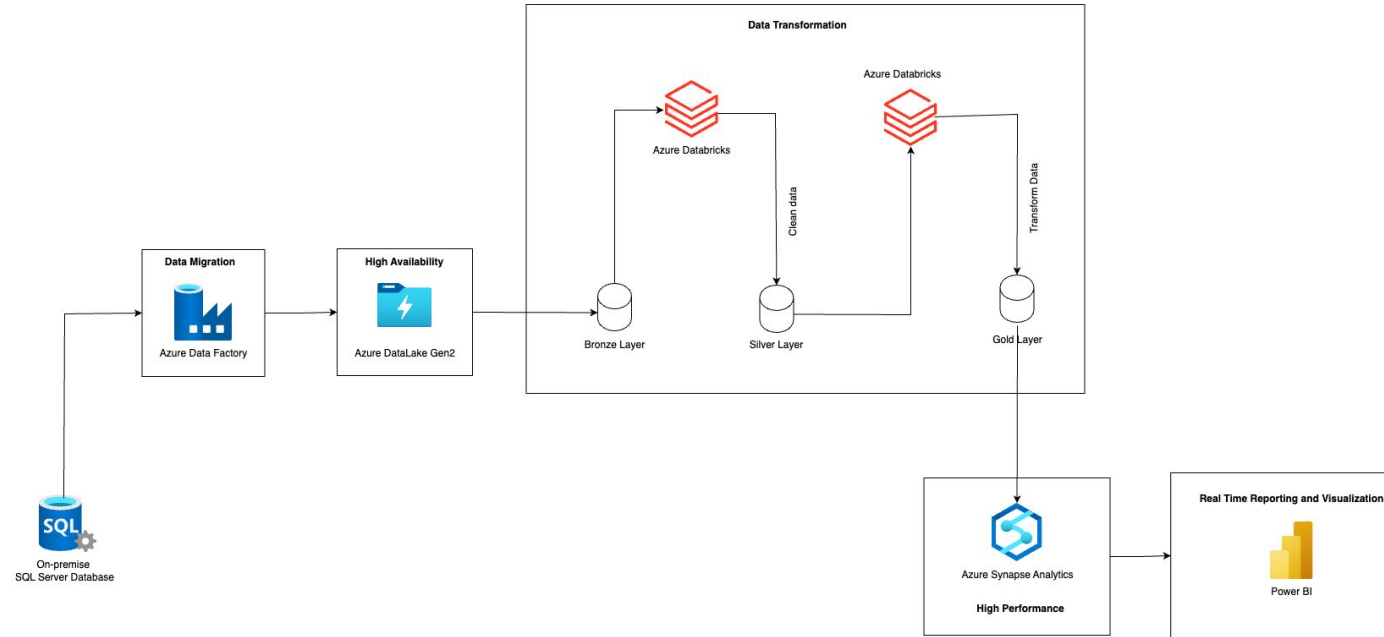
BR 2: Develop a **transformation** layer to convert raw data into clean, structured formats.

TR 2.1: Design and implement ETL (Extract, Transform, Load) processes for data cleansing and normalization.

BR 3: Maintain **high availability** throughout the data pipeline.

TR 3.1: Configure redundant storage and automatic failover mechanisms to ensure continuous data availability.

Architecture



BR1: Data Migration

Microsoft Azure

Search resources, services, and docs (G+)

Copilot

sbasespu@ncsu.edu
NORTH CAROLINA STATE UNIV...

Home > Microsoft.DataFactory-20241124201425 | Overview >

data-factory-independent-study Data factory (V2)

manage Delete

Overview

- Activity log
- Access control (IAM)
- Tags
- Diagnose and solve problems
- Settings
 - Networking
 - Managed identities
 - Properties
 - Locks
- Getting started
 - Quick start
- Monitoring
 - Alerts
 - Metrics
 - Diagnostic settings
 - Logs
- Automation
 - CLI / PS

Essentials

Resource group (move) : [rg-independent-study](#)

Status : Succeeded

Location : East US

Subscription (move) : [Azure for Students](#)

Subscription ID : fe9c7010-47b1-419c-b7b2-b334792f0bec

Type : Data factory (V2)

Getting started : [Quick start](#)

Azure Data Factory Studio

[Launch studio](#)

Quick Starts Tutorials Template Gallery Training Modules

Monitoring

BR1: Data Migration

Microsoft Azure | Data Factory | data-factory-independent-study

Search factory and documentation

Would you like to see Data Factory inside of Microsoft Fabric, Microsoft's newest cloud-first data analytics SaaS platform? Click [here](#) to get started with Fabric Data Factory!

» Data Factory | Validate all | Publish all

General

Factory settings

Connections

Linked services

Integration runtimes

Microsoft Purview

Source control

Git configuration

ARM template

Author

Triggers

Global parameters

Data flow libraries

Security

Credentials

Customer managed key

Outbound rules

Managed private endpoi...

Workflow orchestration manager

Integration runtimes

The integration runtime (IR) is the compute infrastructure to provide the following data integration capabilities across different network...

+ New Refresh

Filter by name

Showing 1 - 1 of 1 items

Name	Type	Sub-type	Status
AutoResolveIntegrationR...	Azure	Public	Running

Integration runtime setup

Settings Nodes Auto update Sharing Links

Install integration runtime on Windows machine or add further nodes using the Authentication Key.

Name SHIR

Option 1: Express setup

[Click here to launch the express setup for this computer](#)

Option 2: Manual setup

Step 1: [Download and install integration runtime](#)

Step 2: Use this key to register your integration runtime

Name	Authentication key
Key1	IR@a67976f2-b1ef-43d5-89a6-7ba5bc10f6bd@data-factory-indepeder
Key2	IR@a67976f2-b1ef-43d5-89a6-7ba5bc10f6bd@data-factory-indepeder

Close

BR1: Data Migration

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

Copilot

124201425 | Overview >

Credential-study

Delete

Essentials

Resource group (move) : rg-i

Status : Suc

Location : East

Subscription (move) : Azu

Subscription ID : fe9c

Microsoft Integration Runtime Express Setup

Integration Runtime (Self-hosted) Express Setup

Installing and registering the Integration Runtime (Self-hosted) node.

- ✓ Loading configuration
- ✓ Downloading Integration Runtime (Self-hosted)
- ✓ Installing Integration Runtime (Self-hosted)
- ✓ Registering Integration Runtime (Self-hosted)

Integration Runtime (Self-hosted) "SHIR" is successfully installed on your computer.

Note: Credentials for on-premises data sources are stored locally on this machine. Use the Settings page to regularly back up credentials to a file. You can use this file to restore or recover the Integration Runtime (Self-hosted) in case of a failure. See Integration Runtime article for details.

Close

Quick Starts

Tutorials

Template Gallery

Training Modules

BR1: Data Migration





Integration runtimes

The integration runtime (IR) is the compute infrastructure to provide the following data integration capabilities across different network environment. [Learn more](#)

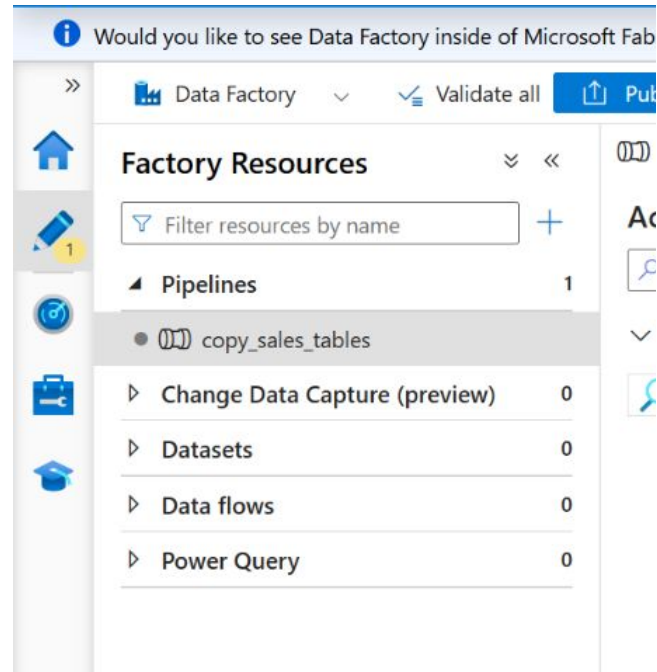
[+](#) New [↻](#) Refresh

[🔍](#) Filter by name

Showing 1 - 2 of 2 items

Name ↑↓	Type ↑↓	Sub-type ↑↓	Status ↑↓	Related ↑↓	Region ↑↓	Version ↑↓
 AutoResolveIntegrationR...	Azure	Public	 Running	0	Auto Resolve	---
 SHIR	Self-Hosted	---	 Running (Limited)	0	---	5.46.9020.1

BR1: Data Migration



BR1: Data Migration

The screenshot displays the Microsoft Fabric Data Factory interface. At the top, a banner asks if the user wants to see Data Factory inside of Microsoft Fabric. The main interface is divided into several sections:

- Factory Resources:** A sidebar on the left showing a tree view of resources. Under 'Pipelines', the 'copy_sales_tables' pipeline is selected, showing a count of 1. Other resources like 'Change Data Capture (preview)', 'Datasets', 'Data flows', and 'Power Query' are listed with counts of 0.
- Activities:** A central pane showing the 'copy_sales_tables' pipeline. It contains a 'Lookup' activity, which is highlighted. The activity is named 'Look for Sales tables' and has a description 'Look for Sales tables'. It is currently in the 'General' tab.
- Properties:** A pane on the right showing the properties of the selected 'Lookup' activity. It includes fields for 'Name' (set to 'copy_sales_tables'), 'Description', and 'Annotations' (with a '+ New' button).
- General Settings:** A pane at the bottom showing the 'General' tab for the 'Lookup' activity. It includes fields for 'Name' (set to 'Look for Sales tables'), 'Description', 'Activity state' (set to 'Activated'), 'Timeout' (set to '0:12:00:00'), 'Retry' (set to '0'), and 'Retry interval (sec)' (set to '30').

BR1: Data Migration

Microsoft Azure | Data Factory | data-factory-independent-study | Search factory and documentation

Would you like to see Data Factory inside of Microsoft Fabric, Microsoft's newest cloud-first data analytics SaaS platform? Click [here](#) to get started with Fabric Data Factory!

» Data Factory | Validate all | Publish all

Factory Resources

- Pipelines 1
 - copy_sales_tables
- Datasets 0
- Data flows 0
- Power Query 0

Activities

- look
- Lookup

Lookup

- Look for Sales tables

Settings

Source dataset * Select...

First row only ☒

New dataset

In pipeline activities and data flows, reference a dataset to specify the location and structure of your data within a data store. [Learn more](#)

Select a data store

sql


All | Azure | Database | File | Generic protocol

- Amazon RDS for SQL Server
- Azure SQL Database
- Azure SQL Database Managed Instance
- SQL server

Continue Cancel

BR1: Data Migration

Edit linked service

 SQL server [Learn more](#)

Version

☒ Recommended ☐ Legacy

Server name *


Database name *

Authentication type


User name *


☒ Password ☐ Azure Key Vault

Password *

Always encrypted 

☐


Encrypt 


Trust server certificate 

☒

Additional connection properties

[+ New](#)

 Connection successful

 Test connection

BR1: Data Migration

The screenshot displays the Azure Data Factory (ADF) console. On the left, the 'Factory Resources' pane shows the hierarchy: Pipelines > copy_sales_tables. The main canvas shows the pipeline diagram with a 'Lookup' activity connected to a 'ForEach' activity. The 'Lookup' activity is labeled 'Look for Sales tables' and has a green checkmark. The 'ForEach' activity is labeled 'ForEach Schema Table' and contains an 'Activities' container with a plus sign. Below the canvas, the 'Output' tab shows the pipeline run details.

Pipeline run ID: 56403b28-7dd2-495e-982e-a3e9eef43e

Pipeline status: Succeeded

Activity status: Succeeded

Activity name	Activity status	Activity type	Run start	Duration	In
Look for Sales tables	Succeeded	Lookup	11/25/2024, 1:03:17 PM	10s	St

BR1: Data Migration

The screenshot displays the Microsoft Azure Data Factory interface for a pipeline named 'copy_sales_tables'. The left sidebar shows the 'Factory Resources' tree with 'Pipelines' expanded, showing 'copy_sales_tables'. The main canvas shows the pipeline design with a 'Lookup' activity named 'Look for Sales tables'. The 'Settings' tab is active, showing 'Sequential' execution and 'Batch count' of 1. The 'Items' property is set to 'This property should be parameterized.' with a hint to 'Add dynamic content [Alt+Shift+D]'. The 'Pipeline expression builder' is open on the right, showing the expression '@activity('Look for Sales tables').output'.

Microsoft Azure | Data Factory | data-factory-independent-study | Search factory and documentation

Would you like to see Data Factory inside of Microsoft Fabric, Microsoft's newest cloud-first data analytics SaaS platform? Click [here](#) to get started with Fabric Data Factory!

Factory Resources

- Filter resources by name
- Pipelines (1)
 - copy_sales_tables
- Datasets (1)
 - SqlServerTable1
- Data flows (0)
- Power Query (0)

Activities

- for
- Move and transform
 - Copy data
 - Data flow
- Iteration & conditionals
 - ForEach

Lookup

- Look for Sales tables

Settings

- Sequential ☐
- Batch count
- Items

Add dynamic content [Alt+Shift+D]

Pipeline expression builder

Add dynamic content below using any combination of expressions, functions and system variables.

```
@activity('Look for Sales tables').output
```

Clear contents

Activity outputs | Parameters | System variables | Functions | Variables

Search

- Look for Sales tables
 - Look for Sales tables activity output
- Look for Sales tables count
 - Count of the rows
- Look for Sales tables value array
 - Array of row data

OK Cancel

BR3: High Availability

The screenshot displays the Azure Data Factory (ADF) interface for a pipeline named 'copy_sales_tables'. The pipeline is currently in a 'Succeeded' state, as indicated by the green checkmark and the 'Pipeline status' section. The pipeline consists of two activities: a 'Lookup' activity named 'Look for Sales tables' and a 'ForEach' activity named 'ForEach Schema Table'. The 'Lookup' activity is connected to the 'ForEach' activity via a green arrow, indicating a sequential flow. The 'ForEach' activity is currently empty, showing 'No activities' inside its container. The 'Properties' pane on the right shows the 'General' tab with the pipeline name 'copy_sales_tables' and a description field. The 'Output' tab at the bottom shows the 'Pipeline run ID' as '56403b28-7dd2-495e-982e-a3e96eeef43e' and the 'Pipeline status' as 'Succeeded'. A table below shows the activity details for the 'Lookup' activity, which was successful on 11/25/2024 at 1:03:17 PM, taking 10 seconds to complete.

Activity name	Activity status	Activity type	Run start	Duration	In
Look for Sales tables	✓ Succeeded	Lookup	11/25/2024, 1:03:17 PM	10s	Sp

BR3: High Availability

The screenshot displays the Microsoft Azure Data Factory portal. The top navigation bar includes 'Microsoft Azure', 'Data Factory', and the specific factory name 'data-factory-independent-study'. A search bar is available for finding factory resources and documentation. A notification banner at the top left suggests viewing Data Factory within Microsoft Fabric.

The left sidebar, titled 'Factory Resources', lists various components: Pipelines (1), Datasets (1), Data flows (0), and Power Query (0). The 'Pipelines' section is expanded, showing a pipeline named 'copy_sales_tables'. Below it, the 'Activities' section is expanded, showing a 'Lookup' activity named 'Look for Sales tables'.

The main canvas shows the 'copy_sales_tables' pipeline with a 'Lookup' activity. The activity is configured with the expression '@activity(''Look for Sales tables'').output'.

The right sidebar, titled 'Pipeline expression builder', provides a text area for adding dynamic content. Below this, there are tabs for 'Activity outputs', 'Parameters', 'System variables', 'Functions', and 'Variables'. The 'Activity outputs' tab is selected, showing a search bar and a list of activity outputs for the 'Look for Sales tables' activity:

- Look for Sales tables activity output
- Look for Sales tables count (Count of the rows)
- Look for Sales tables value array (Array of row data)

At the bottom of the right sidebar, there are 'OK' and 'Cancel' buttons.

BR3: High Availability

The screenshot displays the Microsoft Azure Data Factory portal interface. The top navigation bar shows the user is logged in as 'sbasapu@ncsu.edu' from 'NORTH CAROLINA STATE UNIVERSITY'. The main header indicates the current environment is 'Data Factory' for the workspace 'data-factory-independent-study'. A notification banner at the top suggests using Data Factory inside Microsoft Fabric.

The left sidebar, titled 'Factory Resources', lists various components: Pipelines (1), Change Data Capture (0), Datasets (2), and Data flows (0). Under 'Datasets', 'SqlServerCopy' and 'SqlServerTable1' are listed. The 'Activities' section on the right lists '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', and 'Power Query'.

The main canvas shows a data flow activity named 'Copy data' within a pipeline 'copy_sales_tables'. The activity is configured with the following settings:

- Source dataset:** SqlServerCopy
- Use query:** Query (selected)
- Query:** @(concat('SELECT * FROM ', item()).Sc...
- Query timeout (minutes):** 120
- Isolation level:** Select...
- Partition option:** None (selected)

The 'Properties' panel on the right shows the 'General' tab with the following details:

- Name:** copy_sales_tables
- Description:** (empty)
- Annotations:** + New

BR3: High Availability

Connection

Schema

Parameters

Linked service *

AzureDataLakeStorage1

Test connection

Edit

New

Learn more

File path

bronze

/

@{concat(dataset().schemaname, '/', ...}

/

@{concat(dataset().tablename, '.parq...

Compression type

snappy

BR3: High Availability

Microsoft Azure | Data Factory | data-factory-independent-study

Search factory and documentation

Would you like to see Data Factory inside of Microsoft Fabric, Microsoft's newest cloud-first data analytics SaaS platform? Click [here](#) to get started with Fabric Data Factory!

Renew Cancel Refresh Update pipeline List Gantt

Activity runs

Pipeline run ID: ed4593cb-f037-497a-b8ef-251fff55821d

All status List Monitor in Azure Metrics Export to CSV

Showing 1 - 21 items

Activity name	Activity status	Activity type	Run start	Duration	Integration runtime	User properties	Activity run ID	Log
Look for Sales tables	Succeeded	Lookup	11/28/2024, 3:53:19 PM	12s	SHIR		5deb458c-177b-4397-8732-1350f03d9a90	
ForEach Scchma Table	Succeeded	ForEach	11/28/2024, 3:53:31 PM	1m 51s			7b59ec6a-bc8c-4b95-b8a0-0b18d26c5882	
Copy Each Table	Succeeded	Copy data	11/28/2024, 3:53:32 PM	1m 24s	SHIR		65a3db97-f9ea-450e-92a5-b551f982d2da	
Copy Each Table	Succeeded	Copy data	11/28/2024, 3:53:32 PM	44s	SHIR		12f3ea6c-1d92-4f01-bdcb-66368d663c19	
Copy Each Table	Succeeded	Copy data	11/28/2024, 3:53:32 PM	49s	SHIR		85a961f0-2b2b-42f9-9025-5e1937177f5e	
Copy Each Table	Succeeded	Copy data	11/28/2024, 3:53:32 PM	52s	SHIR		e76c6758-eb61-42d1-a2f9-1c2cb443038f	
Copy Each Table	Succeeded	Copy data	11/28/2024, 3:53:32 PM	48s	SHIR		6ae2727e-3469-4b06-bbed-76d86cc7e00a	
Copy Each Table	Succeeded	Copy data	11/28/2024, 3:53:32 PM	48s	SHIR		dfdbfd32-9299-41c4-a809-be49eb45d31b	
Copy Each Table	Succeeded	Copy data	11/28/2024, 3:53:32 PM	48s	SHIR		3406d430-733c-4e85-8b6d-f5b47c862b91	
Copy Each Table	Succeeded	Copy data	11/28/2024, 3:53:32 PM	46s	SHIR		af713751-3bcd-42a4-b8a9-5125dd540c5c	
Copy Each Table	Succeeded	Copy data	11/28/2024, 3:53:32 PM	46s	SHIR		f3a4ad08-9757-4e17-9676-aeeb6f6ccb80	
Copy Each Table	Succeeded	Copy data	11/28/2024, 3:53:32 PM	1m 22s	SHIR		961353e4-b74a-4b7c-b5c1-a2fe43059613	
Copy Each Table	Succeeded	Copy data	11/28/2024, 3:53:32 PM	47s	SHIR		e3bbc72e-b3f4-45cd-a0bf-fc8c8fa859a	
Copy Each Table	Succeeded	Copy data	11/28/2024, 3:53:32 PM	1m 23s	SHIR		e6b8d1d4-6509-4a33-8b09-7b96e14adef4	
Copy Each Table	Succeeded	Copy data	11/28/2024, 3:53:32 PM	49s	SHIR		4091127b-6bce-4376-8b25-a94050317d43	

BR3: High Availability

The screenshot displays the SQL Server Enterprise Manager interface. On the left, the Object Explorer shows the database structure for 'MS\SQLEXPRESS (SQL Server 16.0.1000 - MS\Niles)'. The 'AdventureWorks2022' database is expanded, showing various system and user tables. The main window shows a SQL query in the 'SQLQuery2.sql' file, which is a query to find tables within the 'Sales' schema. The query is as follows:

```
SELECT s.name AS Schemaname, t.name As Tablelname
FROM sys.tables t
INNER JOIN sys.schemas s
ON t.schema_id = s.schema_id
WHERE s.name = 'Sales'
```

The query results are displayed in the 'Results' pane, showing a list of tables within the 'Sales' schema. The results are as follows:

SchemaName	TableName
Sales	SalesTaxRate
Sales	PersonCreditCard
Sales	SalesTerritory
Sales	SalesTerritoryHistory
Sales	ShoppingCartItem
Sales	SpecialOffer
Sales	SpecialOfferProduct
Sales	Store
Sales	CountryRegionCurrency
Sales	CreditCard
Sales	Currency
Sales	CurrencyRate
Sales	Customer
Sales	SalesOrderDetail
Sales	SalesOrderHeader
Sales	SalesOrderHeaderSalesReason
Sales	SalesPerson
Sales	SalesPersonQuotaHistory
Sales	SalesReason

The status bar at the bottom indicates that the query was executed successfully, returning 19 rows in 00:00:00. The interface also shows the status of the SQL Server instance as 'MS\SQLEXPRESS (16.0 RTM) MS\Niles (62) AdventureWorks2022 00:00:00 19 rows'.

BR3: High Availability

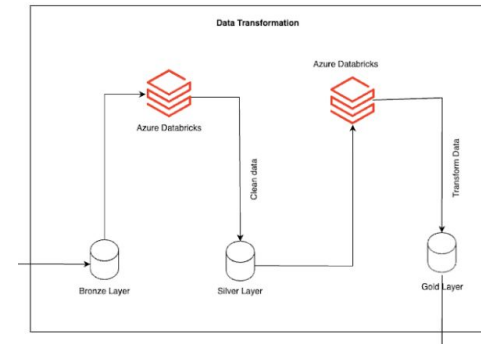
The screenshot displays the Microsoft Azure portal interface. At the top, the navigation bar includes the Microsoft Azure logo, a search bar, the Copilot icon, and user information for 'sbasapu@ncsu.edu'. The breadcrumb trail indicates the path: Home > indstudstorageaccount | Containers >. The main content area is titled 'bronze Container'. On the left, a sidebar shows the 'Overview' tab selected, with links for 'Diagnose and solve problems' and 'Access Control (IAM)'. The main pane shows the 'Overview' details for the 'bronze' container, including the authentication method (Access key) and location (bronze / Sales). A search bar for blobs is present, along with a toggle for 'Show deleted objects'. Below this is a table listing the contents of the container.

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state	
<input type="checkbox"/> [.]							...
<input type="checkbox"/> CountryRegionCurrency	11/28/2024, 3:54:14 PM					-	...
<input type="checkbox"/> CreditCard	11/28/2024, 3:54:16 PM					-	...
<input type="checkbox"/> Currency	11/28/2024, 3:54:12 PM					-	...
<input type="checkbox"/> CurrencyRate	11/28/2024, 3:54:16 PM					-	...
<input type="checkbox"/> Customer	11/28/2024, 3:54:22 PM					-	...
<input type="checkbox"/> PersonCreditCard	11/28/2024, 3:54:16 PM					-	...
<input type="checkbox"/> SalesOrderDetail	11/28/2024, 3:54:22 PM					-	...
<input type="checkbox"/> SalesOrderHeader	11/28/2024, 3:54:19 PM					-	...
<input type="checkbox"/> SalesOrderHeaderSalesReason	11/28/2024, 3:54:17 PM					-	...
<input type="checkbox"/> SalesPerson	11/28/2024, 3:55:18 PM					-	...
<input type="checkbox"/> SalesPersonQuotaHistory	11/28/2024, 3:54:17 PM					-	...
<input type="checkbox"/> SalesReason	11/28/2024, 3:54:14 PM					-	...
<input type="checkbox"/> SalesTaxRate	11/28/2024, 3:54:53 PM					-	...
<input type="checkbox"/> SalesTerritory	11/28/2024, 3:54:51 PM					-	...

BR2: Data Transformation

The data will be processed through **three layers**:

- **Bronze Layer:** Raw, unprocessed data stored in its original form.
- **Silver Layer:** Cleaned and transformed data, where data quality and consistency improvements are applied.
- **Gold Layer:** Fully refined and aggregated data, ready for reporting, advanced analytics, and business insights.



BR2: Data Transformation

[Home >](#)

rg-independent-study_ind-study-databricks | Overview

Deployment

[Delete](#) [Cancel](#) [Redeploy](#) [Download](#) [Refresh](#)

Overview

Inputs

Outputs

Template



Your deployment is complete



Deployment name : rg-independent-study_ind-study-databricks

Subscription : [Azure for Students](#)Resource group : [rg-independent-study](#)

Start time : 11/28/2024, 6:06:55 PM

Correlation ID : a53ff3dc-1117-450e-8420-109bdf61ea5d



Deployment details

Resource	Type	Status	Operation details
ind-study-databricks	Azure Databricks Service	OK	Operation details



Next steps

[Go to resource](#)

Give feedback

[Tell us about your experience with deployment](#)

Cost management

Get notified to stay within your budget and prevent unexpected charges on your bill.

[Set up cost alerts >](#)

Microsoft Defender for Cloud

Secure your apps and infrastructure

[Go to Microsoft Defender for Cloud >](#)

Free Microsoft tutorials

[Start learning today >](#)

Work with an expert

Azure experts are service provider partners who can help manage your assets on Azure and be your first line of support.

[Find an Azure expert >](#)

BR2: Data Transformation

The screenshot displays the Databricks web interface for configuring a cluster. The top navigation bar includes the Microsoft Azure logo, the Databricks logo, a search bar, and the user profile 'ind-study-databricks'. The left sidebar contains a 'New' button and a list of navigation items: Workspace, Recents, Catalog, Workflows, Compute (selected), SQL, SQL Editor, Queries, Dashboards, Genie, Alerts, Query History, SQL Warehouses, Data Engineering, Job Runs, Data Ingestion, Delta Live Tables, Machine Learning, Playground, Experiments, Features, and Models.

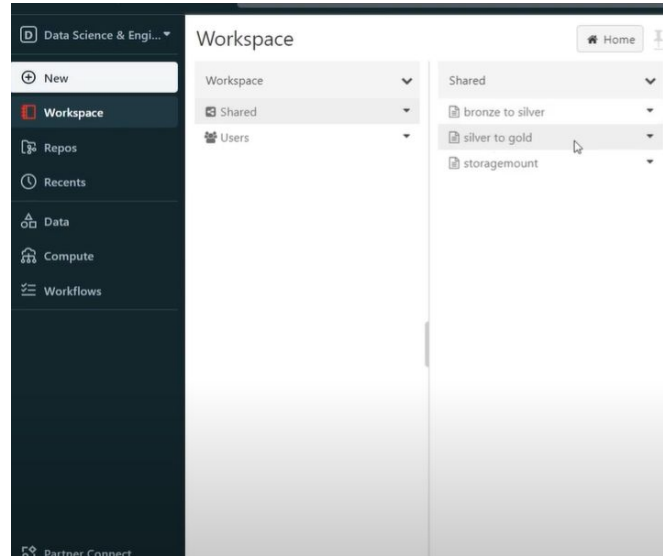
The main content area is titled 'Sakshi Basapure's Cluster' and shows the 'Configuration' tab. The cluster is currently in a 'Stopped' state, indicated by a green circle with an 'S'. The configuration details are as follows:

- Policy:** Unrestricted
- Multi node / Single node:** Single node (selected)
- Access mode:** Single user access (selected)
- Single user:** Sakshi Basapure
- Performance:** Databricks Runtime Version 15.4 LTS (includes Apache Spark 3.5.0, Scala 2.12)
- Use Photon Acceleration:** Checked
- Node type:** Standard_D4ds_v5 (16 GB Memory, 4 Cores)
- Terminate after:** 10 minutes of inactivity (checked)
- Tags:** No custom tags. Automatically added tags are visible under 'Advanced options'.

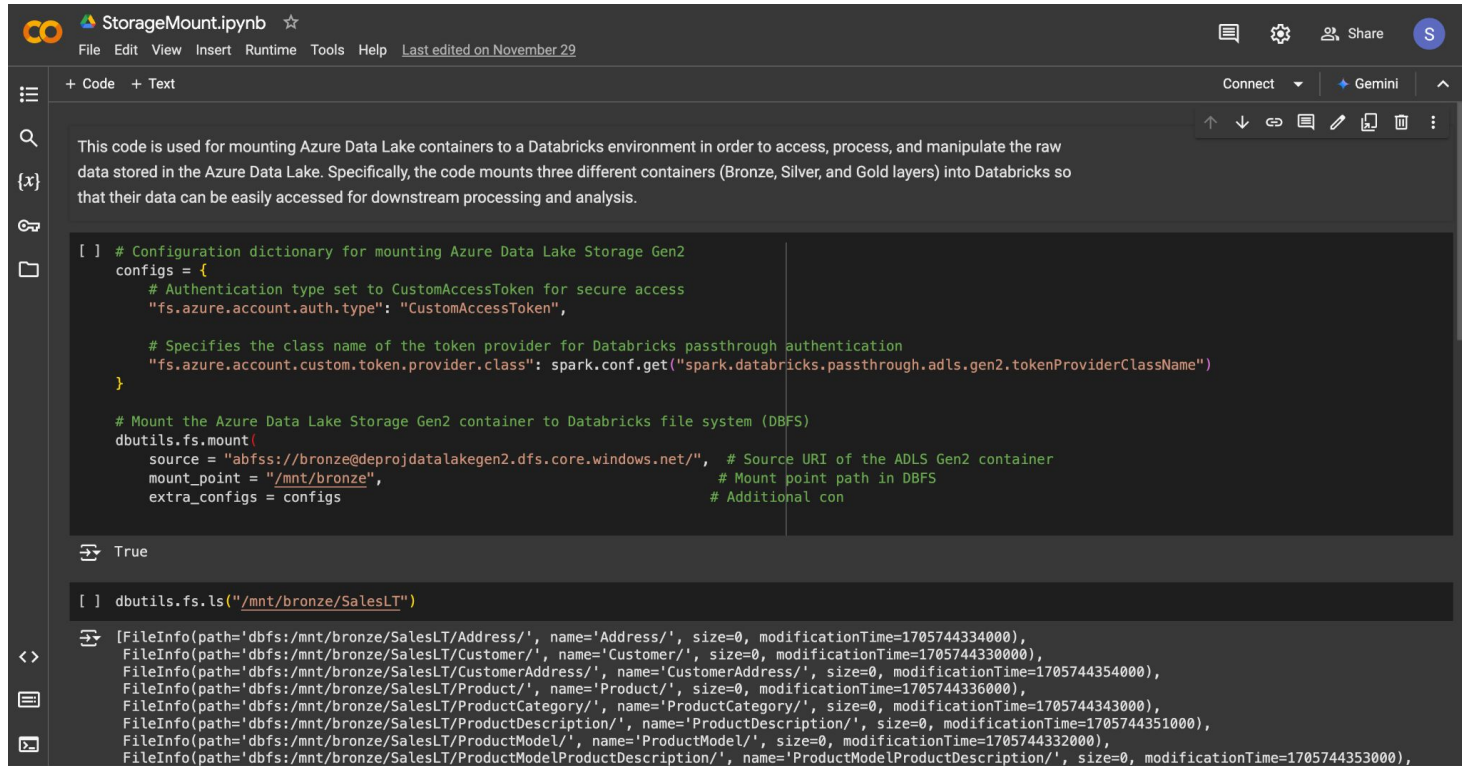
A 'Summary' box on the right provides a quick overview of the cluster's resources:

Summary	
1 Driver	16 GB Memory, 4 Cores
Runtime	15.4x-scala2.12
Photon	Standard_D4ds_v5 2 DBU/h

BR2: Data Transformation



BR2: Data Transformation



The screenshot shows a Jupyter Notebook titled "StorageMount.ipynb" with a menu bar (File, Edit, View, Insert, Runtime, Tools, Help) and a status bar (Last edited on November 29). The notebook is in "Code" mode. The left sidebar contains icons for file explorer, search, and other functions. The main area displays a code cell with the following content:

```
[ ] # Configuration dictionary for mounting Azure Data Lake Storage Gen2
configs = {
    # Authentication type set to CustomAccessToken for secure access
    "fs.azure.account.auth.type": "CustomAccessToken",

    # Specifies the class name of the token provider for Databricks passthrough authentication
    "fs.azure.account.custom.token.provider.class": spark.conf.get("spark.databricks.passthrough.adls.gen2.tokenProviderClassName")
}

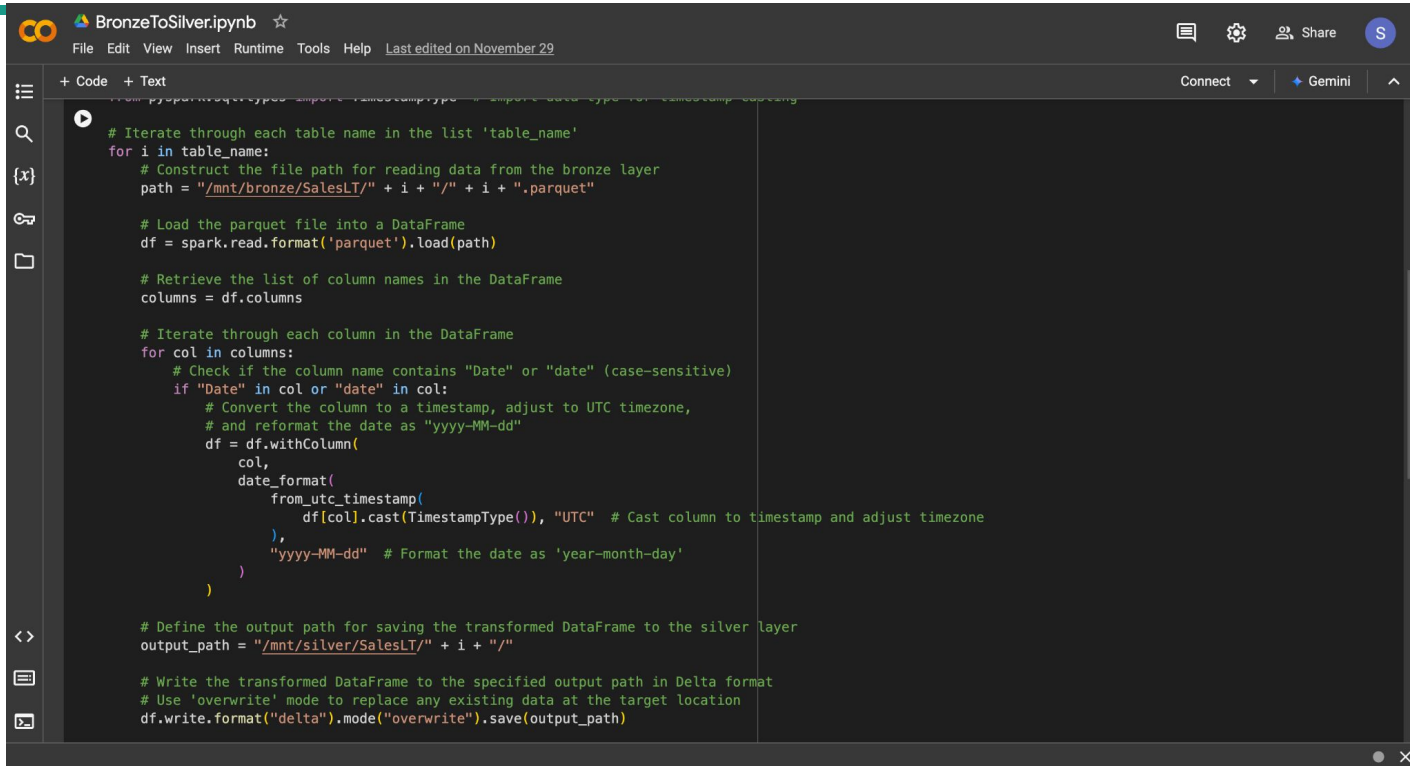
# Mount the Azure Data Lake Storage Gen2 container to Databricks file system (DBFS)
dbutils.fs.mount(
    source = "abfss://bronze@deprojdatalakegen2.dfs.core.windows.net/", # Source URI of the ADLS Gen2 container
    mount_point = "/mnt/bronze", # Mount point path in DBFS
    extra_configs = configs # Additional con

True

[ ] dbutils.fs.ls("/mnt/bronze/SalesLT")

[FileInfo(path='dbfs:/mnt/bronze/SalesLT/Address/', name='Address/', size=0, modificationTime=1705744334000),
 FileInfo(path='dbfs:/mnt/bronze/SalesLT/Customer/', name='Customer/', size=0, modificationTime=1705744330000),
 FileInfo(path='dbfs:/mnt/bronze/SalesLT/CustomerAddress/', name='CustomerAddress/', size=0, modificationTime=1705744354000),
 FileInfo(path='dbfs:/mnt/bronze/SalesLT/Product/', name='Product/', size=0, modificationTime=1705744336000),
 FileInfo(path='dbfs:/mnt/bronze/SalesLT/ProductCategory/', name='ProductCategory/', size=0, modificationTime=1705744343000),
 FileInfo(path='dbfs:/mnt/bronze/SalesLT/ProductDescription/', name='ProductDescription/', size=0, modificationTime=1705744351000),
 FileInfo(path='dbfs:/mnt/bronze/SalesLT/ProductModel/', name='ProductModel/', size=0, modificationTime=1705744332000),
 FileInfo(path='dbfs:/mnt/bronze/SalesLT/ProductModelProductDescription/', name='ProductModelProductDescription/', size=0, modificationTime=1705744353000),
```

BR2: Data Transformation



```
from pyspark.sql.types import TimestampType
from pyspark.sql.functions import col, date_format, from_utc_timestamp

# Iterate through each table name in the list 'table_name'
for i in table_name:
    # Construct the file path for reading data from the bronze layer
    path = "/mnt/bronze/SalesLT/" + i + "/" + i + ".parquet"

    # Load the parquet file into a DataFrame
    df = spark.read.format('parquet').load(path)

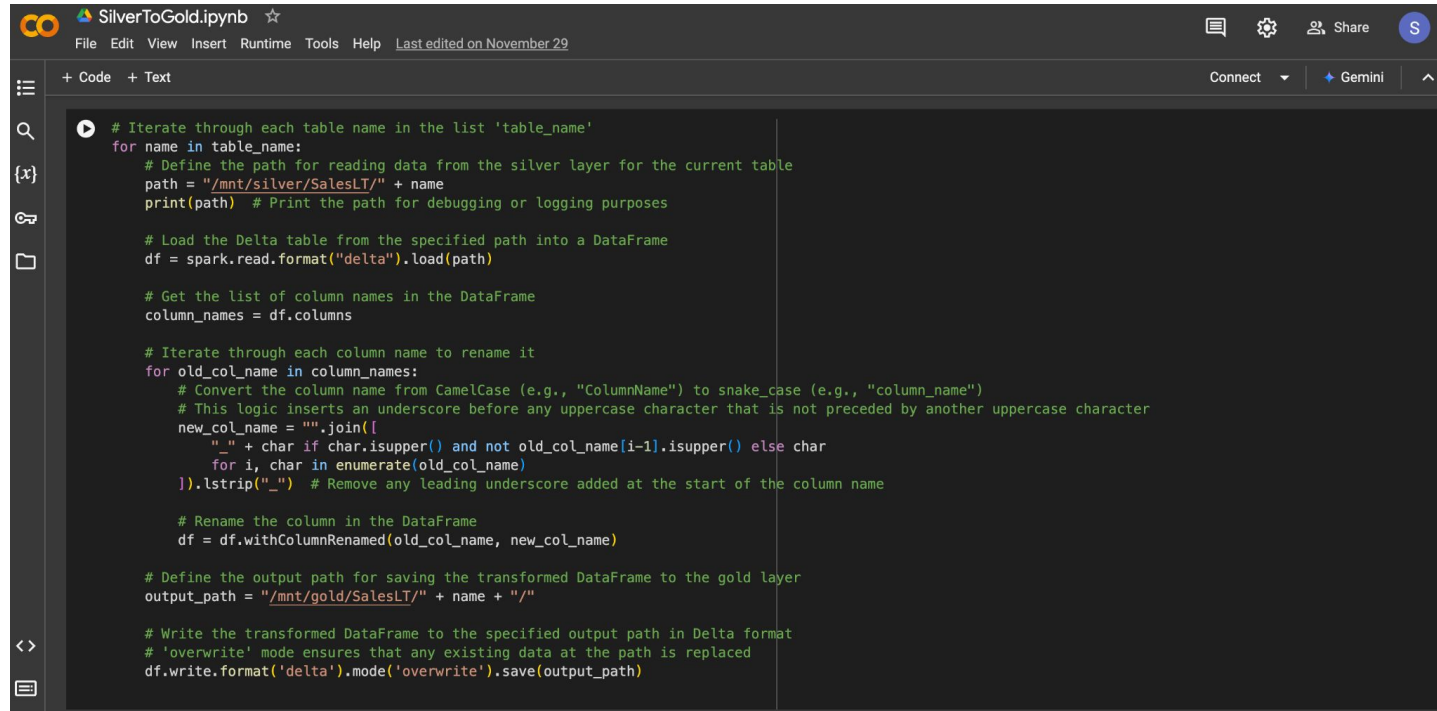
    # Retrieve the list of column names in the DataFrame
    columns = df.columns

    # Iterate through each column in the DataFrame
    for col in columns:
        # Check if the column name contains "Date" or "date" (case-sensitive)
        if "Date" in col or "date" in col:
            # Convert the column to a timestamp, adjust to UTC timezone,
            # and reformat the date as "yyyy-MM-dd"
            df = df.withColumn(
                col,
                date_format(
                    from_utc_timestamp(
                        df[col].cast(TimestampType()), "UTC" # Cast column to timestamp and adjust timezone
                    ),
                    "yyyy-MM-dd" # Format the date as 'year-month-day'
                )
            )

    # Define the output path for saving the transformed DataFrame to the silver layer
    output_path = "/mnt/silver/SalesLT/" + i + "/"

    # Write the transformed DataFrame to the specified output path in Delta format
    # Use 'overwrite' mode to replace any existing data at the target location
    df.write.format("delta").mode("overwrite").save(output_path)
```

BR2: Data Transformation



```
# SilverToGold.ipynb
File Edit View Insert Runtime Tools Help Last edited on November 29

+ Code + Text
Connect Gemini

# Iterate through each table name in the list 'table_name'
for name in table_name:
    # Define the path for reading data from the silver layer for the current table
    path = "/mnt/silver/SalesLT/" + name
    print(path) # Print the path for debugging or logging purposes

    # Load the Delta table from the specified path into a DataFrame
    df = spark.read.format("delta").load(path)

    # Get the list of column names in the DataFrame
    column_names = df.columns

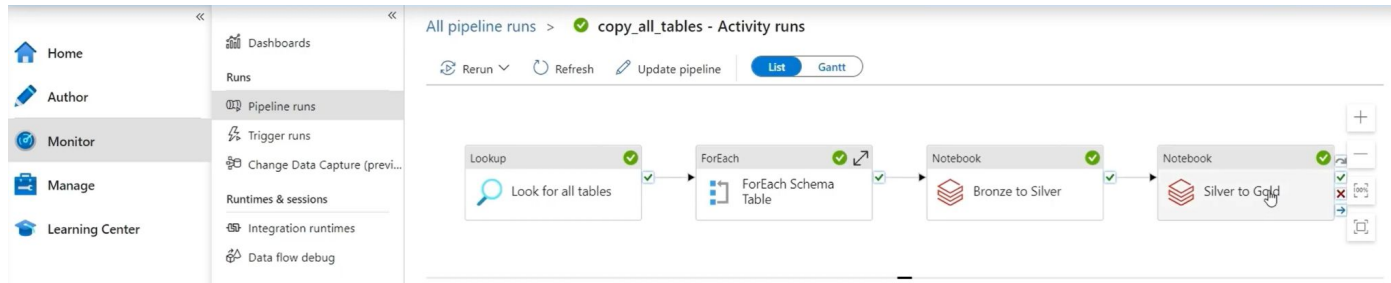
    # Iterate through each column name to rename it
    for old_col_name in column_names:
        # Convert the column name from CamelCase (e.g., "ColumnName") to snake_case (e.g., "column_name")
        # This logic inserts an underscore before any uppercase character that is not preceded by another uppercase character
        new_col_name = "".join([
            "_" + char if char.isupper() and not old_col_name[i-1].isupper() else char
            for i, char in enumerate(old_col_name)
        ]).lstrip("_") # Remove any leading underscore added at the start of the column name

        # Rename the column in the DataFrame
        df = df.withColumnRenamed(old_col_name, new_col_name)

    # Define the output path for saving the transformed DataFrame to the gold layer
    output_path = "/mnt/gold/SalesLT/" + name + "/"

    # Write the transformed DataFrame to the specified output path in Delta format
    # 'overwrite' mode ensures that any existing data at the path is replaced
    df.write.format('delta').mode('overwrite').save(output_path)
```

BR2: Data Transformation



BR2: Data Transformation

Microsoft Azure | Data Factory | data-factory-independent-study | Search factory and documentation

Runs

Pipeline runs

Trigger runs

Change Data Capture (previ...

Runtimes & sessions

Integration runtimes

Data flow debug

Notifications

Alerts & metrics

All pipeline runs > **copy_sales_tables - Activity runs**

Rerun Cancel Refresh Update pipeline List Gantt

Pipeline is modified after this run. The current pipeline configuration is shown.

Copy Each Table

Activity name	Activity status	Activity type	Run start	Duration	Integration runtime	User properties	Activity run ID	Log
MountStorage	Succeeded	Notebook	12/2/2024, 12:15:36 PM	Less than 1s	SHIR		2cc4e8e4-54e6-4efa-891e-1765e85d642e	
SilverToGold	Succeeded	Notebook	12/2/2024, 12:15:36 PM	Less than 1s	SHIR		a113c455-eb66-40bf-9091-f73fdbde495	
BronzeToSilver	Succeeded	Notebook	12/2/2024, 12:15:36 PM	Less than 1s	SHIR		5ac161bd-adf8-4a41-a859-48aafb033d74	
Look for Sales tables	Succeeded	Lookup	12/2/2024, 12:15:36 PM	8s	SHIR		9156f8b8-c1eb-401d-be54-7ad9e1c4b176	
ForEach Schema Table	Succeeded	ForEach	12/2/2024, 12:15:45 PM	1m 48s			f09033d1-b4cb-4cfa-84f2-a6a1a286fc42	
Copy Each Table	Succeeded	Copy data	12/2/2024, 12:15:46 PM	38s	SHIR		29eb6f96-9ea5-4db6-a654-b1e3f7e9b8be	
Copy Each Table	Succeeded	Copy data	12/2/2024, 12:15:46 PM	21s	SHIR		79dc45d0-c4e1-47da-87e4-02909d200b85	
Copy Each Table	Succeeded	Copy data	12/2/2024, 12:15:46 PM	21s	SHIR		9e611a90-013c-428a-849d-c0d7ec1b96dd	
Copy Each Table	Succeeded	Copy data	12/2/2024, 12:15:46 PM	21s	SHIR		4d4d0728-858b-4d80-8451-983608c7a304	
Copy Each Table	Succeeded	Copy data	12/2/2024, 12:15:46 PM	1m 43s	SHIR		395b58b4-0790-410e-8f9e-1c9d71467bf8	
Copy Each Table	Succeeded	Copy data	12/2/2024, 12:15:46 PM	1m 3s	SHIR		753eda62-28ef-4efd-bf4b-1bb5b9c1bf3b	
Copy Each Table	Succeeded	Copy data	12/2/2024, 12:15:46 PM	20s	SHIR		eba5c154-33b8-4945-bf4a-833e29a336d8	
Copy Each Table	Succeeded	Copy data	12/2/2024, 12:15:46 PM	33s	SHIR		1d16872a-5ea4-4bc7-8e22-fc394d902528	
Copy Each Table	Succeeded	Copy data	12/2/2024, 12:15:46 PM	28s	SHIR		842cbac7-6122-42d0-be8c-b3a3016c9c92	
Copy Each Table	Succeeded	Copy data	12/2/2024, 12:15:46 PM	23s	SHIR		7aa0277e-95ae-4ccf-9f26-eb25c9bf2cdf	

BR2: Data Transformation

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: silver / SalesLT / Customer

Search blobs by prefix (case-sensitive)

Show deleted objects

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
<input type="checkbox"/>  [-]						***
<input type="checkbox"/>  _delta_log						***
<input type="checkbox"/>  part-00000-6b145ce2-1de8-4cfe-830f-3ca8696c86...	4/17/2023, 12:23:12 ...	Hot (Inferred)		Block blob	88.33 KiB	Available ***

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: gold / SalesLT / Address

Search blobs by prefix (case-sensitive)

Show deleted objects

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
<input type="checkbox"/>  [-]						***
<input type="checkbox"/>  _delta_log						***
<input type="checkbox"/>  part-00000-bd1df7a5-1ab5-467b-9c67-1cafceea4...	4/17/2023, 12:31:48 ...	Hot (Inferred)		Block blob	34.76 KiB	Available ***

Learnings



1. Importance of Secure Data Handling
2. Scalability in Cloud Architecture
3. Data Transformation and Layered Architecture
4. Real-Time Data Integration

Challenges



1. Design Challenges:
 - a. Architectural Integration
 - b. Balancing Performance and Cost Efficiency
2. Implementation Challenges:
 - a. Setting Up Self-Hosted Integration Runtime (SHIR)
 - b. Debugging Multi-Service Pipelines

Acknowledgement



A heartfelt thanks to Prof. Viniotis for their invaluable guidance and mentorship throughout this project.

The foundational knowledge from courses like "Cloud Computing Architecture" and "Advanced Cloud Computing Architecture" greatly contributed to designing and implementing this architecture.