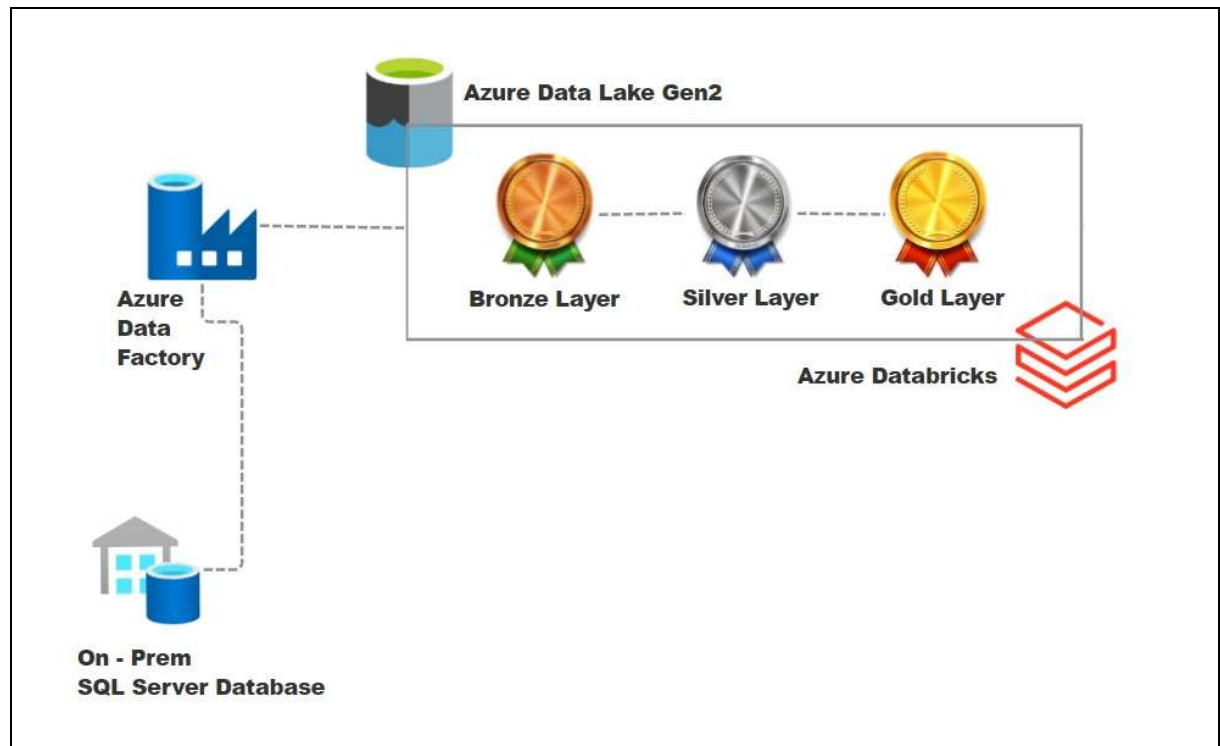
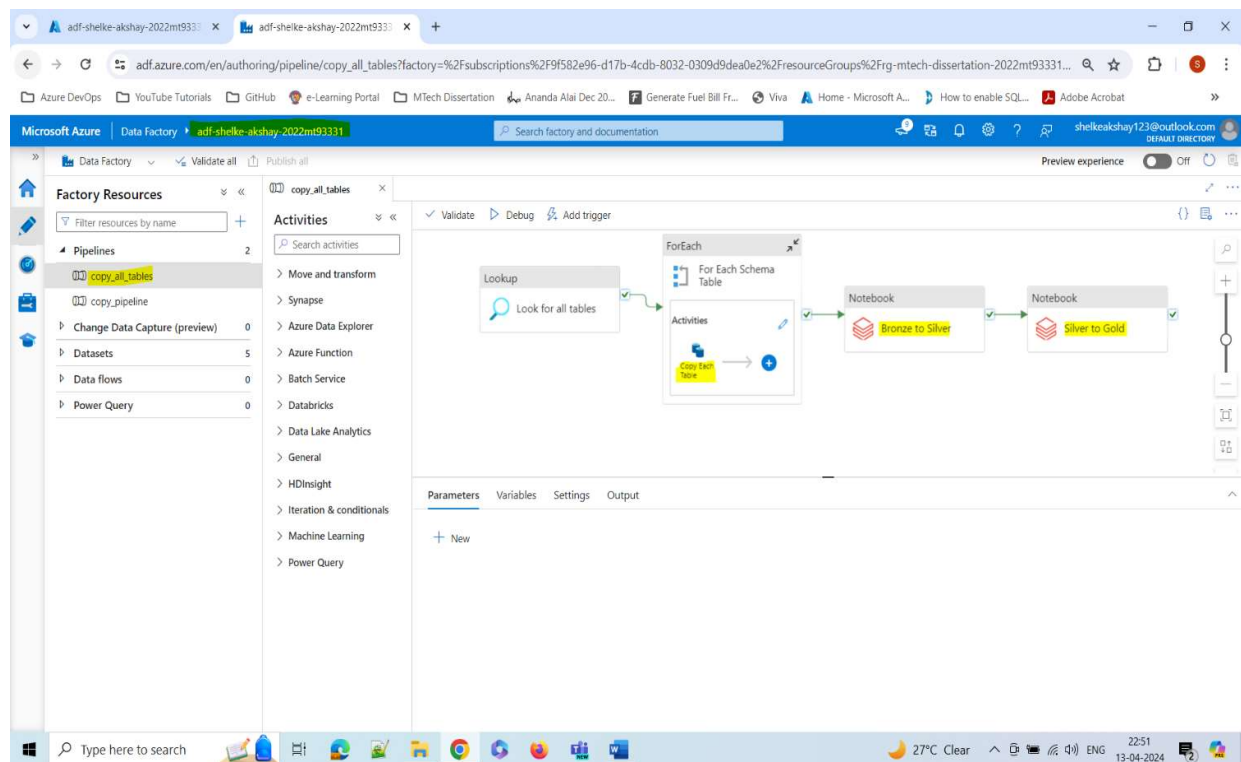
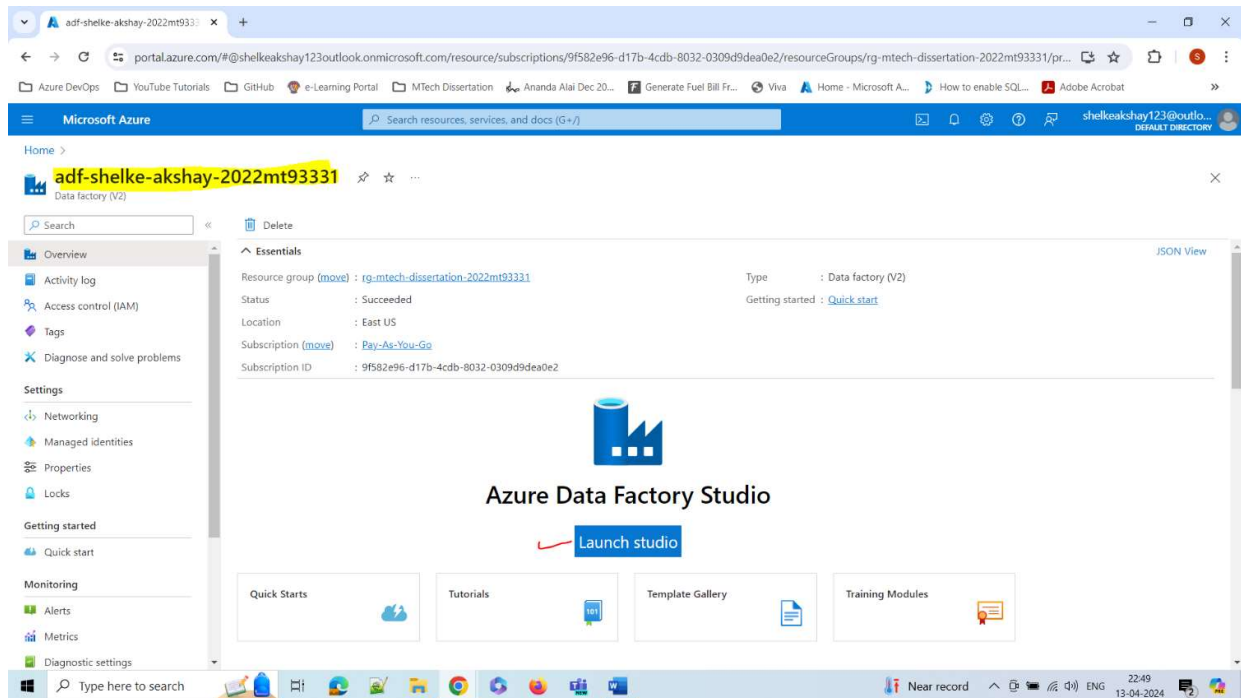
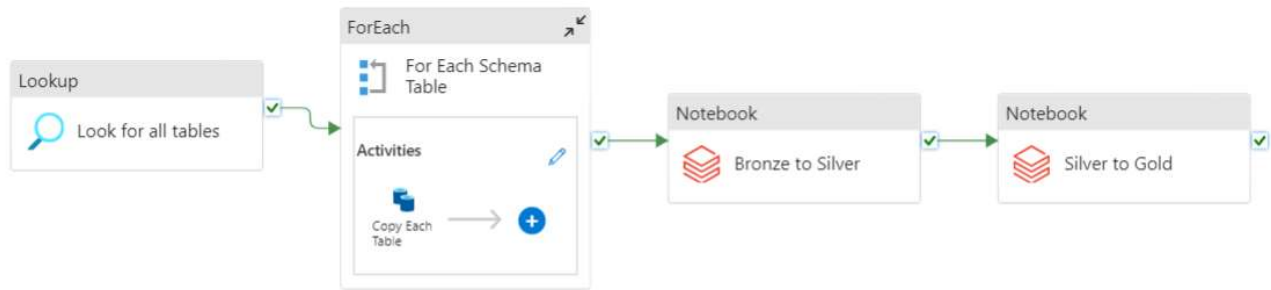


# Data Transformation



## 1) Azure Data Factory(ADF) Pipeline





- a. Look for all tables.
- b. Copy Each table into BRONZE container.
- c. Data Transformation – BRONZE to SILVER
- d. Data Transformation – SILVER to GOLD

## 2) Azure Databricks(ADB) – Code Notebooks for Data Transformation

### Azure Databricks Instance

The screenshot displays the Azure Databricks instance overview page. The browser address bar shows the URL: `portal.azure.com/#@shelkeakashay123outlook.onmicrosoft.com/resource/subscriptions/9f582e96-d17b-4cdb-8032-0309d9dea0e2/resourceGroups/rg-mtech-dissertation-2022mt93331/pr...`. The Microsoft Azure portal header is visible, including the search bar and user profile. The left sidebar contains navigation options: Overview, Activity log, Access control (IAM), Tags, Diagnose and solve problems, Settings, Virtual Network Peering, Encryption, Networking, Security & compliance, Properties, Locks, Monitoring, Diagnostic settings, and Automation. The main content area shows the instance details for **az-databricks-2022mt93331**. The **Essentials** section lists: Status: Active, Managed Resource Group: `databricks-rg-az-databricks-2022mt93331-5s6ckzox7zoh6`, Resource group: `rg-mtech-dissertation-2022mt93331`, Location: East US, URL: `https://adb-1395394472918857.17.azuredatabricks.net`, Pricing Tier: Premium (Role-based access control), Subscription: [Pay-As-You-Go](#), and Subscription ID: `9f582e96-d17b-4cdb-8032-0309d9dea0e2`. A large red Databricks logo is centered, with a **Launch Workspace** button below it. At the bottom, there are four tiles: [Documentation](#), [Getting Started](#), [Import Data from File](#), and [Import Data from Azure Storage](#). The Windows taskbar at the bottom shows the system clock as 22:55 on 13-04-2024.

## Azure Databricks Compute Cluster

The screenshot shows the Azure Databricks web interface for configuring a compute cluster. The browser address bar shows the URL: `adb-1395394472918857.17.azuredatabricks.net/compute/clusters/0413-084505-hxp8y4n?o=1395394472918857`. The left sidebar contains navigation options: New, Workspace, Recents, Catalog, Workflows, Compute (selected), SQL, SQL Editor, Queries, Dashboards, Alerts, Query History, SQL Warehouses, Data Engineering, Job Runs, Data Ingestion, Delta Live Tables, Machine Learning, Playground, and Experiments. The main content area is titled 'Akshay Shelke's Cluster' and includes tabs for Configuration, Notebooks (0), Libraries, Event log, Spark UI, Driver logs, Metrics, Apps, and Spark compute UI - Master. The Configuration tab is active, showing settings for Policy (Unrestricted), Access mode (Single user), and Performance (Databricks Runtime Version 13.3 LTS). A Summary box on the right displays cluster details: 1 Driver, 14 GB Memory, 4 Cores, Runtime 13.3.x-scala2.12, and Standard\_DS3\_v2 node type with 0.75 DBU/h. The bottom of the screen shows a Windows taskbar with the search bar and system tray.

## Python Code To create Mount point for Azure Blob Storage Container – For Bronze Container

```
configs = {
    "fs.azure.account.auth.type": "CustomAccessToken",
    "fs.azure.account.custom.token.provider.class":
spark.conf.get("spark.databricks.passthrough.adls.gen2.tokenProviderClassName")
}

try:
    dbutils.fs.mount(
        source = "abfss://bronze@storageaccmtechdemo.dfs.core.windows.net/",
        mount_point = "/mnt/bronze",
        extra_configs = configs)
    print("Mount Point created successfully")
except:
    print("Mount Point already exists")
```

#### Python Code To create Mount point for Azure Blob Storage Container – For Silver Container

```
configs = {
    "fs.azure.account.auth.type": "CustomAccessToken",
    "fs.azure.account.custom.token.provider.class":
spark.conf.get("spark.databricks.passthrough.adls.gen2.tokenProviderClassName")
}

try:
    dbutils.fs.mount(
        source = "abfss://silver@storageacmtechdemo.dfs.core.windows.net/",
        mount_point = "/mnt/silver",
        extra_configs = configs)
    print("Mount Point created successfully")
except:
    print("Mount Point already exists")
```

#### Python Code To create Mount point for Azure Blob Storage Container – For Gold Container

```
configs = {
    "fs.azure.account.auth.type": "CustomAccessToken",
    "fs.azure.account.custom.token.provider.class":
spark.conf.get("spark.databricks.passthrough.adls.gen2.tokenProviderClassName")
}

try:
    dbutils.fs.mount(
        source = "abfss://gold@storageacmtechdemo.dfs.core.windows.net/",
        mount_point = "/mnt/gold",
        extra_configs = configs)
    print("Mount Point created successfully")
except:
    print("Mount Point already exists")
```

## Python Code for Data Transformation – BRONZE to SILVER

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

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

table_name = []

for i in dbutils.fs.ls("/mnt/bronze/SalesLT/"):
    table_name.append(i.name.split('/')[0])

for i in table_name:
    path = "/mnt/bronze/SalesLT/" + i + "/" + i + ".parquet"
    df = spark.read.format("parquet").load(path)
    column = df.columns

    for col in column:
        if "Date" in col or "date" in col:
            df = df.withColumn(col,
date_format(from_utc_timestamp(df[col].cast(TimestampType()), "UTC"), "yyyy-MM-dd"))

    output_path = "/mnt/silver/SalesLT/" + i + "/"
    df.write.format("delta").mode("overwrite").save(output_path)

display(df)
```

## Python Code for Data Transformation – SILVER to GOLD

```
from pyspark.sql import SparkSession
from pyspark.sql.functions import col, regexp_replace

dbutils.fs.ls("/mnt/silver/SalesLT/")

table_name = []

for i in dbutils.fs.ls("/mnt/silver/SalesLT/"):
    table_name.append(i.name.split('/')[0])

for name in table_name:
    path = "/mnt/silver/SalesLT/" + name
    print(path)
    df = spark.read.format("delta").load(path)

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

    for old_col_name in column_names:
        # Convert column name from ColumnName to Column_Name format
        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)]).rstrip("_")

        # Change the column name using withColumnRenamed and regexp_replace
        df = df.withColumnRenamed(old_col_name, new_col_name)

    output_path = "/mnt/gold/SalesLT/" + name + "/"
    df.write.format("delta").mode("overwrite").save(output_path)

display(df)
```



### 3) Azure Data Factory pipeline RUNS

The screenshot shows the Azure Data Factory Pipeline Runs page. The left sidebar contains navigation options: Dashboards, Runs, Pipeline runs, Trigger runs, Change Data Capture (previ...), Runtimes & sessions, Integration runtimes, Data flow debug, Notifications, and Alerts & metrics. The main area displays the 'Pipeline runs' for the 'copy\_all\_tables' pipeline. The table shows the following data:

Pipeline name	Run start	Run end	Duration	Triggered by	Status	Run	Parameters	Annotations	Run ID
copy_all_tables	4/13/2024, 4:23:47 PM	4/13/2024, 4:29:45 PM	5m 59s	Manual trigger	Succeeded	Original			ce
copy_all_tables	4/13/2024, 4:23:01 PM	4/13/2024, 4:27:11 PM	4m 10s	Manual trigger	Succeeded	Original			07

The screenshot shows the Azure Data Factory Pipeline Runs page for the 'copy\_all\_tables' pipeline. The left sidebar contains navigation options: Dashboards, Runs, Pipeline runs, Trigger runs, Change Data Capture (previ...), Runtimes & sessions, Integration runtimes, Data flow debug, Notifications, and Alerts & metrics. The main area displays the 'All pipeline runs' for the 'copy\_all\_tables' pipeline. The diagram shows a 'Lookup' activity followed by a 'For Each' loop containing 'Copy Each Table' activities, which then lead to 'Bronze to Silver' and 'Silver to Gold' notebook activities. The table below shows the details of the pipeline runs:

Activity	Status	Activity Type	Run start	Run end	Duration	Triggered by	Run	Parameters	Annotations	Run ID
For Each Schema Table	Succeeded	ForEach	4/13/2024, 4:24:06 PM	4/13/2024, 4:24:06 PM	3m 49s					7b6eae58-a6eb-b46b-4489-9b22f81d71e7
Copy Each Table	Succeeded	Copy data	4/13/2024, 4:24:07 PM	4/13/2024, 4:24:07 PM	33s	SHIR-06-03-2024				8bd6555b-9e8c-4181-a92d-519b6a1cf9f4
Copy Each Table	Succeeded	Copy data	4/13/2024, 4:24:07 PM	4/13/2024, 4:24:07 PM	23s	SHIR-06-03-2024				5cf56efb-776b-4f51-aa9d-ee0a48a8be1c
Copy Each Table	Succeeded	Copy data	4/13/2024, 4:24:07 PM	4/13/2024, 4:24:07 PM	21s	SHIR-06-03-2024				5c3c325f-869f-484a-a367-ade20cd9d7a1
Copy Each Table	Succeeded	Copy data	4/13/2024, 4:24:07 PM	4/13/2024, 4:24:07 PM	1m 41s	SHIR-06-03-2024				71c162fc-0ccf-49d3-be5f-6f56ce9e16c3
Copy Each Table	Succeeded	Copy data	4/13/2024, 4:24:07 PM	4/13/2024, 4:24:07 PM	25s	SHIR-06-03-2024				3cae638f-d4a0-4e32-b8eb-12110770bce3
Copy Each Table	Succeeded	Copy data	4/13/2024, 4:24:07 PM	4/13/2024, 4:24:07 PM	24s	SHIR-06-03-2024				e352b737-98ad-49f1-b98b-2f15904228be
Copy Each Table	Succeeded	Copy data	4/13/2024, 4:24:07 PM	4/13/2024, 4:24:07 PM	1m 49s	SHIR-06-03-2024				5a12b380-a4ca-4ee9-b68c-cf0d8095bd94
Copy Each Table	Succeeded	Copy data	4/13/2024, 4:24:07 PM	4/13/2024, 4:24:07 PM	3m 44s	SHIR-06-03-2024				e424c5d-c316-4ab8-b156-5c6018552210
Copy Each Table	Succeeded	Copy data	4/13/2024, 4:24:07 PM	4/13/2024, 4:24:07 PM	22s	SHIR-06-03-2024				a886253d-a99b-44f6-a9cf-34b7932ec2c2
Copy Each Table	Succeeded	Copy data	4/13/2024, 4:24:07 PM	4/13/2024, 4:24:07 PM	23s	SHIR-06-03-2024				8b4a97c2-a53f-4f09-9a0c-7f8f108a6ac
Bronze to Silver	Succeeded	Notebook	4/13/2024, 4:27:55 PM	4/13/2024, 4:27:55 PM	59s	AutoResolveIntegration				2ad50eef-c229-4105-a087-5c349f1c7a53
Silver to Gold	Succeeded	Notebook	4/13/2024, 4:28:55 PM	4/13/2024, 4:28:55 PM	49s	AutoResolveIntegration				77899a17-290e-4605-846a-8718b72b95f



adf-sheike-akshay-2022mt9331 x adf-sheike-akshay-2022mt9331 x ADF\_adf-sheike-akshay-2022mt9331 x ADF\_adf-sheike-akshay-2022mt9331 x

adb-1395394472918857.17.azuredatabricks.net/jobs/196373656445241/runs/1679873586090557o=1395394472918857

Azure DevOps YouTube Tutorials GitHub e-Learning Portal MTech Dissertation Ananda Alai Dec 20... Generate Fuel Bill Fr... Viva Home - Microsoft A... How to enable SQL... Adobe Acrobat

Microsoft Azure databricks Search data, notebooks, recent, and more... CTRL + P az-databricks-2022mt9331

Workflows > Runs > ADF\_adf-sheike-akshay-2022mt9331\_copy\_all\_tables\_Silver to Gold\_77899a17-290e-4605-846a-8718b7f2b951 run [Delete job run](#)

Output

✓ 17.68 seconds

```
!strip("")

# Change the column name using withColumnRenamed and regexp_replace
df = df.withColumnRenamed(old_col_name, new_col_name)

output_path = "/mnt/gold/SalesLT/" + name + "/"
df.write.format("delta").mode("overwrite").save(output_path)
```

• d1: pyspark.sql.DataFrame = (Sales\_Order\_ID: integer, Revision\_Number: integer ... 20 more fields)

```
/mnt/silver/SalesLT/Address
/mnt/silver/SalesLT/Customer
/mnt/silver/SalesLT/CustomerAddress
/mnt/silver/SalesLT/Product
/mnt/silver/SalesLT/ProductCategory
/mnt/silver/SalesLT/ProductDescription
/mnt/silver/SalesLT/ProductModel
/mnt/silver/SalesLT/ProductModelProductDescription
/mnt/silver/SalesLT/SalesOrderDetail
/mnt/silver/SalesLT/SalesOrderHeader
```

✓ 0.22 seconds

display(df)

Table

New result table: OFF

	Sales_Order_ID	Revision_Number	Order_Date	Due_Date	Ship_Date	Status	Online_Order_Flag	Sales_Order_Number	Purcha
1	71774	2	2008-06-01	2008-06-13	2008-06-08	5	false	5071774	PC9481

Task run details

Job ID 196373586445241

Task run ID 167987358609055

Run as Akshay Sheike

Launched By runs submit app

Started 04/13/2024, 04:28:57 PM

Ended 04/13/2024, 04:29:31 PM

Duration 34s

Queue duration -

Status Succeeded

Lineage No lineage information for this job. [Learn more](#)

Notebook

/shared/\_silver\_to\_gold

Compute

• Akshay Sheikes Cluster

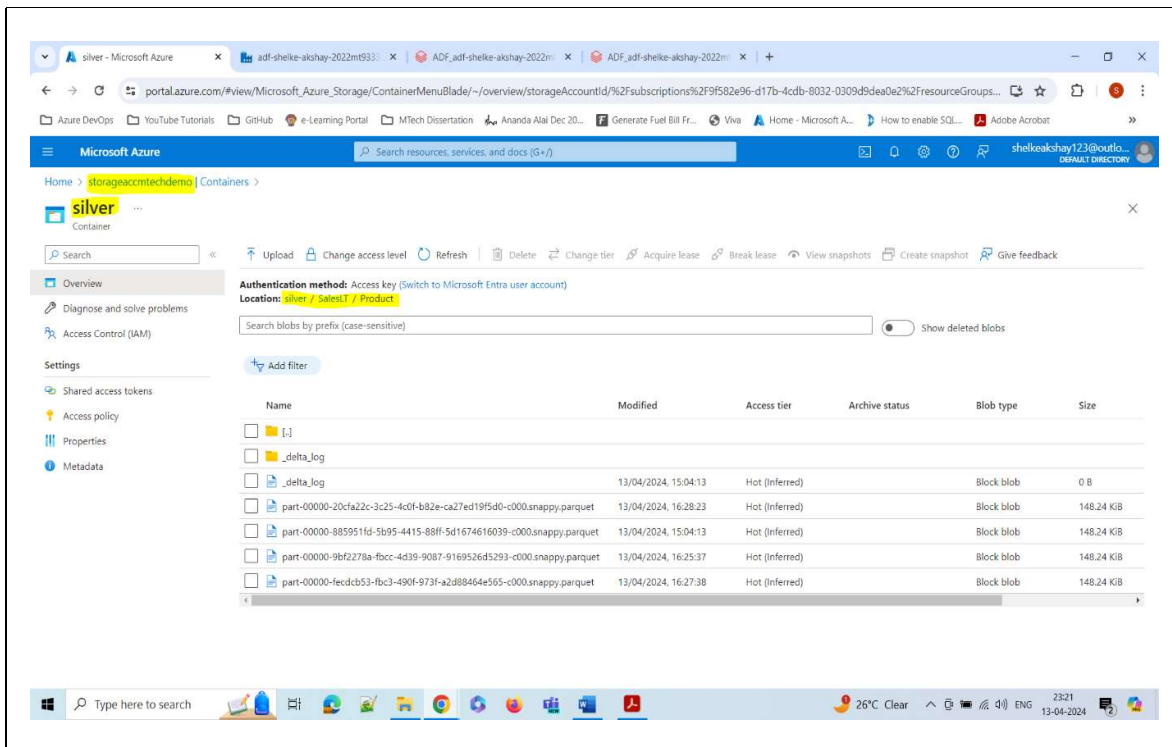
Single node: Standard\_DS3\_v2 - 13.3 UTs (includes Apache Spark 3.4.1, Scala 2.12)

[View details](#) [Spark UI](#) [Logs](#) [Metrics](#)

Permissions

26°C Clear 23:18 13-04-2024

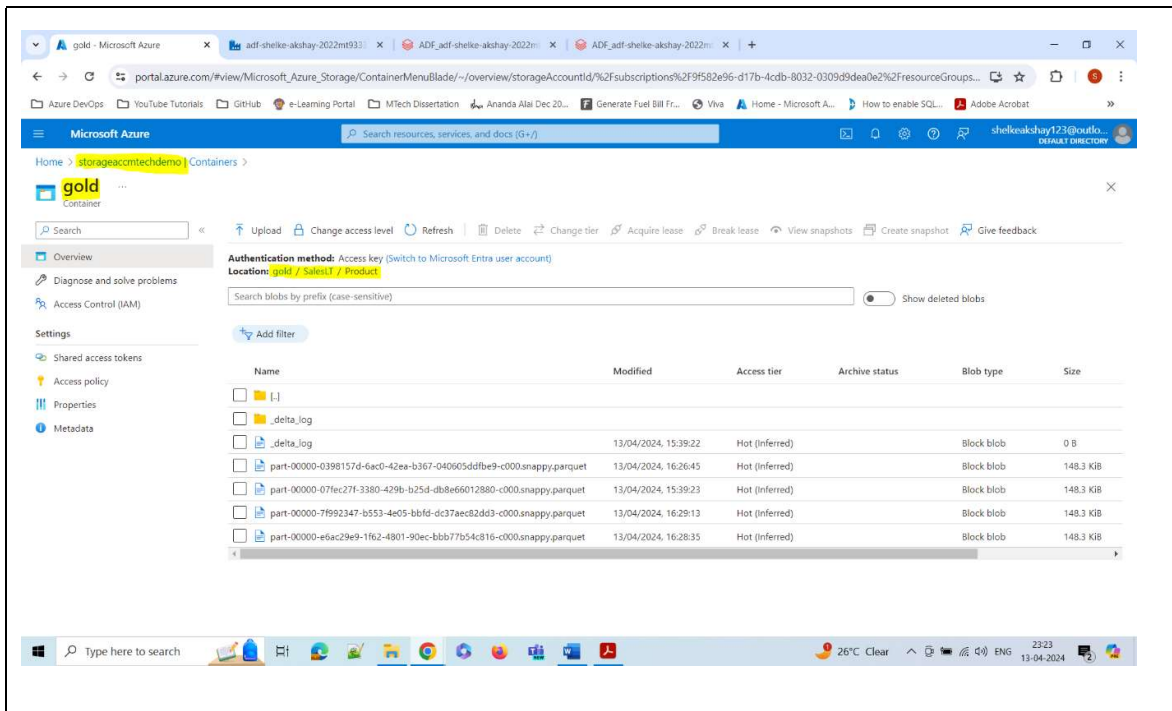
#### 4) Delta tables stored in Silver Container



The screenshot shows the Microsoft Azure portal interface for a Silver Container. The container is named 'silver' and is located in the 'SalesLT / Product' hierarchy. The authentication method is 'Access key (Switch to Microsoft Entra user account)'. The location is 'silver / SalesLT / Product'. The search filter is 'Search blobs by prefix (case-sensitive)'. The table lists the following blobs:

Name	Modified	Access tier	Archive status	Blob type	Size
[-]					
[-]					
_delta_log					
_delta_log	13/04/2024, 15:04:13	Hot (Inferred)		Block blob	0 B
part-00000-20fa22c-3c25-4c0f-b82e-ca7ed19f5d0-c000.snappy.parquet	13/04/2024, 16:28:23	Hot (Inferred)		Block blob	148.24 KiB
part-00000-885951fd-5b95-4415-88ff-5d1674616039-c000.snappy.parquet	13/04/2024, 15:04:13	Hot (Inferred)		Block blob	148.24 KiB
part-00000-9bf227ba-fbcc-4d39-9087-9169526d5293-c000.snappy.parquet	13/04/2024, 16:25:37	Hot (Inferred)		Block blob	148.24 KiB
part-00000-fedcb53-fbc3-490f-973f-a2d89464e565-c000.snappy.parquet	13/04/2024, 16:27:38	Hot (Inferred)		Block blob	148.24 KiB

#### 5) Delta tables stored in Gold Container – Data Transformation Completed



The screenshot shows the Microsoft Azure portal interface for a Gold Container. The container is named 'gold' and is located in the 'SalesLT / Product' hierarchy. The authentication method is 'Access key (Switch to Microsoft Entra user account)'. The location is 'gold / SalesLT / Product'. The search filter is 'Search blobs by prefix (case-sensitive)'. The table lists the following blobs:

Name	Modified	Access tier	Archive status	Blob type	Size
[-]					
[-]					
_delta_log					
_delta_log	13/04/2024, 15:39:22	Hot (Inferred)		Block blob	0 B
part-00000-0398157d-6ac0-42ea-b367-040605dd1be9-c000.snappy.parquet	13/04/2024, 16:26:45	Hot (Inferred)		Block blob	148.3 KiB
part-00000-07fec27f-3380-429b-b25d-db8e66012880-c000.snappy.parquet	13/04/2024, 15:39:23	Hot (Inferred)		Block blob	148.3 KiB
part-00000-7f992347-b553-4e05-bbf6-dc37aec2d8d3-c000.snappy.parquet	13/04/2024, 16:29:13	Hot (Inferred)		Block blob	148.3 KiB
part-00000-e6ac29e9-1f62-4801-90ec-bbb77b54c816-c000.snappy.parquet	13/04/2024, 16:28:35	Hot (Inferred)		Block blob	148.3 KiB