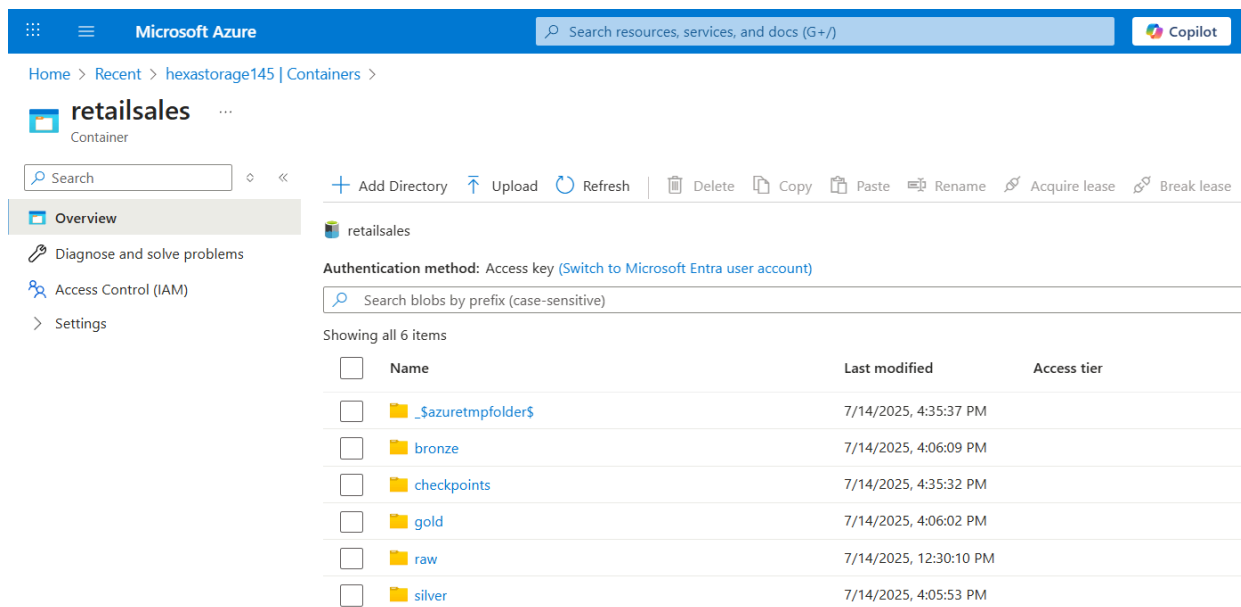


Retail Sales Management - Documentation

This document outlines the steps and components of a **Retail Sales ETL and ML Pipeline** implemented using **Azure Databricks, Delta Lake, MLflow, and Azure DevOps**. Screenshots are included to visualize each stage of the process.

1. Storage Setup

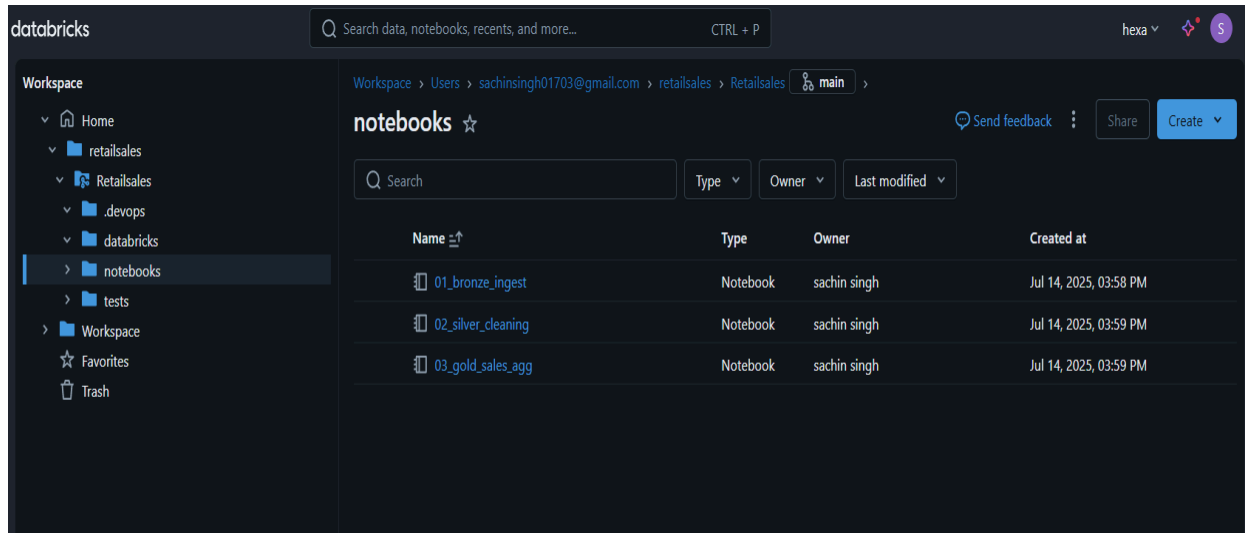


The screenshot displays the Microsoft Azure portal interface for the 'retailsales' container. The top navigation bar shows 'Microsoft Azure' and a search bar. The breadcrumb trail indicates the path: Home > Recent > hexastorage145 | Containers > retailsales. The left sidebar contains the 'Overview' section, which includes links to 'Diagnose and solve problems', 'Access Control (IAM)', and 'Settings'. The main content area shows the 'retailsales' container details, including the authentication method (Access key) and a search bar for blobs. Below this, a table lists six items, each with a checkbox, name, last modified timestamp, and access tier.

	Name	Last modified	Access tier
<input type="checkbox"/>	._azuretmpfolder\$	7/14/2025, 4:35:37 PM	
<input type="checkbox"/>	bronze	7/14/2025, 4:06:09 PM	
<input type="checkbox"/>	checkpoints	7/14/2025, 4:35:32 PM	
<input type="checkbox"/>	gold	7/14/2025, 4:06:02 PM	
<input type="checkbox"/>	raw	7/14/2025, 12:30:10 PM	
<input type="checkbox"/>	silver	7/14/2025, 4:05:53 PM	

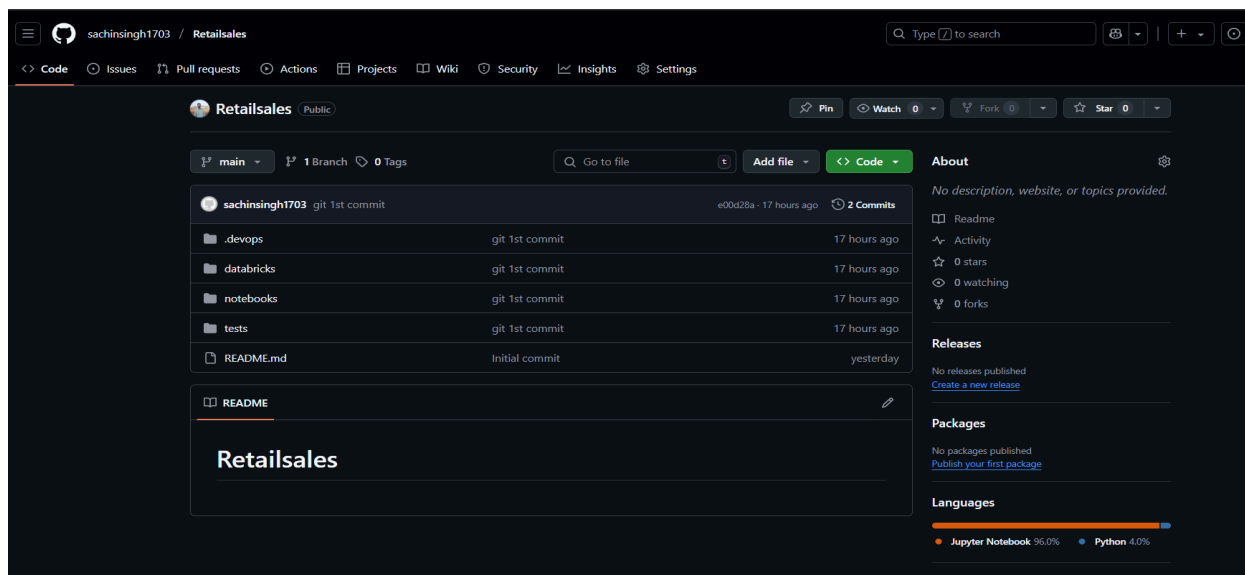
The foundational step is setting up storage for raw and processed data. ADLS Gen2 is used with directories like **/bronze**, **/silver**, and **/gold**.

2. Notebooks Overview



Azure Databricks notebooks were used for ETL and ML tasks. Each notebook handles a specific stage—Bronze, Silver, and Gold layers.

3. 📁 Files Uploaded to GitHub



All notebooks and scripts were version-controlled and stored in GitHub for CI/CD via Azure DevOps.

4. ❌ Failed DevOps Pipeline

Azure DevOps | sachinsingh01703 / hexawareproject / Pipelines / hexawareproject / 20250715.1

hexawareproject

Overview
Boards
Repos
Pipelines
Pipelines
Environments
Library
Test Plans
Artifacts
Project settings

#20250715.1 • Set up CI with Azure Pipelines
hexawareproject

Rerun failed jobs Run new

This run will be cleaned up after 1 month based on your project settings.

Summary Code Coverage

Manually run by sachin singh

Repository and version
hexawareproject
main 7e52f20e

Time started and elapsed
Just now
<1s

Related
0 work items
0 artifacts

Tests and coverage
Get started

View 3 changes

Errors 1

No hosted parallelism has been purchased or granted. To request a free parallelism grant, please fill out the following form <https://aka.ms/azpipelines-parallelism-request>

View documentation for troubleshooting failed runs

Jobs

Name	Status	Duration
Job	Failed	

Early pipeline runs failed due to config or code issues, which were later resolved.

5. Successful DevOps Pipeline

Azure DevOps | sjlakshan2004 / Retail Sales DT / Pipelines / Retail Sales DT / 20250714.1

Retail Sales DT

Overview
Boards
Repos
Pipelines
Pipelines
Environments
Library
Test Plans
Artifacts
Project settings

#20250714.1 • Set up CI with Azure Pipelines
Retail Sales DT

Run new

This run is being retained as one of 3 recent runs by main (Branch).

View retention leases

Summary Code Coverage

Individual CI by Lakshan S J

Repository and version
Retail Sales DT
main 811ddcb4

Time started and elapsed
Just now
27s

Related
0 work items
0 artifacts

Tests and coverage
Get started

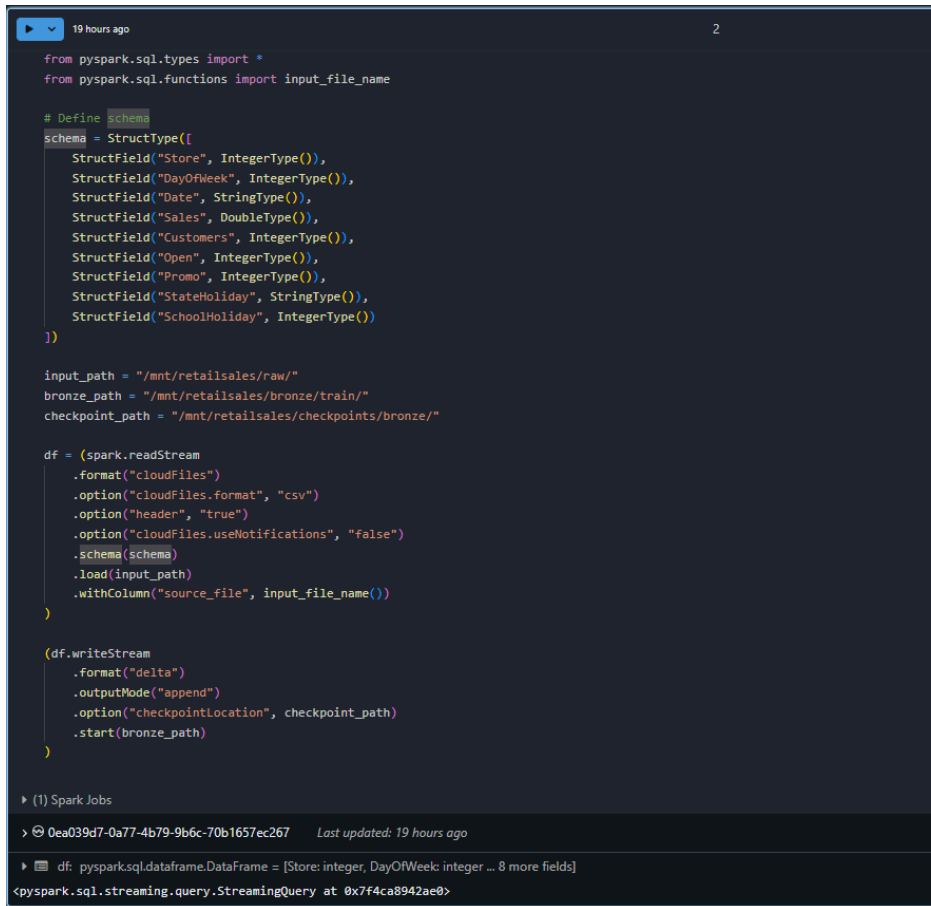
View 4 changes

Jobs

Name	Status	Duration
Job	Success	15s

After resolving errors, pipelines were successfully triggered to deploy notebooks into Databricks.

6. Bronze Ingestion Code



```
from pyspark.sql.types import *
from pyspark.sql.functions import input_file_name

# Define schema
schema = StructType([
    StructField("Store", IntegerType()),
    StructField("DayOfWeek", IntegerType()),
    StructField("Date", StringType()),
    StructField("Sales", DoubleType()),
    StructField("Customers", IntegerType()),
    StructField("Open", IntegerType()),
    StructField("Promo", IntegerType()),
    StructField("StateHoliday", StringType()),
    StructField("SchoolHoliday", IntegerType())
])

input_path = "/mnt/retailsales/raw/"
bronze_path = "/mnt/retailsales/bronze/train/"
checkpoint_path = "/mnt/retailsales/checkpoints/bronze/"

df = (spark.readStream
    .format("cloudFiles")
    .option("cloudFiles.format", "csv")
    .option("header", "true")
    .option("cloudFiles.useNotifications", "false")
    .schema(schema)
    .load(input_path)
    .withColumn("source_file", input_file_name())
)

(df.writeStream
    .format("delta")
    .outputMode("append")
    .option("checkpointLocation", checkpoint_path)
    .start(bronze_path)
)
```

▶ (1) Spark Jobs

> 0ea039d7-0a77-4b79-9b6c-70b1657ec267 Last updated: 19 hours ago

▶ df: pyspark.sql.dataframe.DataFrame = [Store: integer, DayOfWeek: integer ... 8 more fields]

<pyspark.sql.streaming.query.StreamingQuery at 0x7f4ca8942ae0>

Initial ingestion of raw CSVs from storage into Delta Lake using Databricks Autoloader.

7. Silver Cleaning Code

```
19 hours ago (55s) 1

from pyspark.sql.functions import col, to_date

bronze_df = spark.read.format("delta").load("/mnt/retailsales/bronze/train/")

silver_df = (bronze_df
    .filter(col("Open") == 1)
    .withColumn("Date", to_date("Date", "yyyy-MM-dd"))
    .withColumn("StateHoliday", col("StateHoliday").cast("string"))
)

silver_df.write.format("delta").mode("overwrite").save("/mnt/retailsales/silver/train/")

▶ (2) Spark Jobs
▶ bronze_df: pyspark.sql.dataframe.DataFrame = [Store: integer, DayOfWeek: integer ... 8 more fields]
▶ silver_df: pyspark.sql.dataframe.DataFrame = [Store: integer, DayOfWeek: integer ... 8 more fields]
```

Data cleaning and transformations were applied in this stage, including null handling and formatting.

8. Gold Sales Code

```
19 hours ago (42s) 1 Py

from pyspark.sql.functions import sum, avg, count

silver_df = spark.read.format("delta").load("/mnt/retailsales/silver/train/")

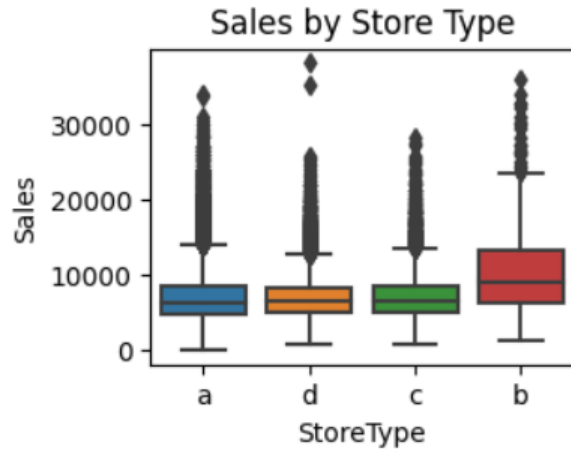
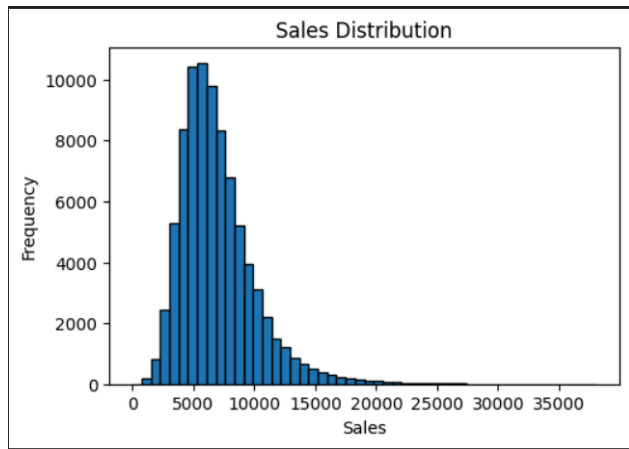
gold_df = (silver_df
    .groupBy("Store")
    .agg(
        sum("Sales").alias("TotalSales"),
        avg("Customers").alias("AvgCustomers"),
        count("Date").alias("DaysOpen")
    )
)

gold_df.write.format("delta").mode("overwrite").save("/mnt/retailsales/gold/store_sales_summary/")

▶ (3) Spark Jobs
▶ gold_df: pyspark.sql.dataframe.DataFrame = [Store: integer, TotalSales: double ... 2 more fields]
▶ silver_df: pyspark.sql.dataframe.DataFrame = [Store: integer, DayOfWeek: integer ... 8 more fields]
```

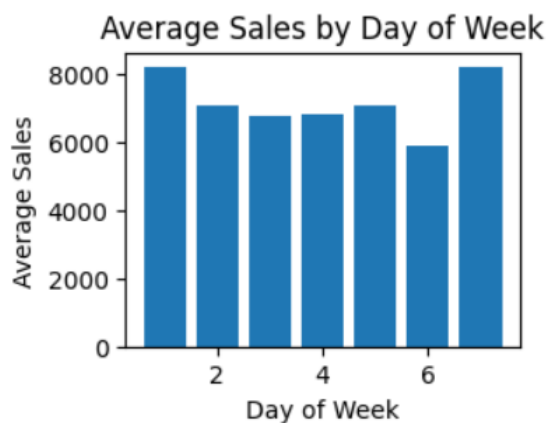
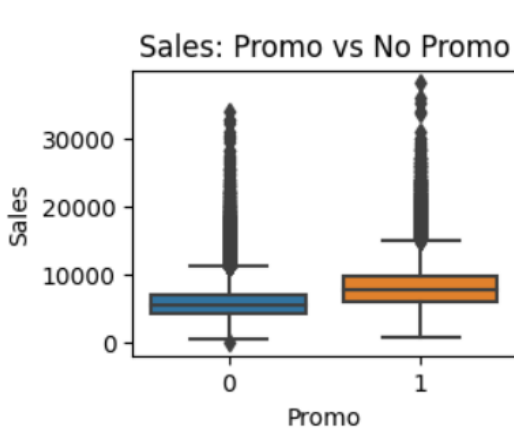
Final aggregations and metrics calculated and saved in gold tables for analytics and ML.

9. 📊 Sales Distribution



Exploratory data analysis (EDA) was conducted on gold data, visualizing trends and distributions.

10. 🤖 Model Training in Progress



Initial model training phase where regression models are trained to predict sales.

11. ✅ Model Training Completed

Model Summary:

- Algorithm: Linear Regression
- Accuracy: 91.9%
- RMSE: 885
- MAE: 545
- Features: 20
- Training Records: 675,212
- Test Records: 169,180

Successfully trained model is logged and registered using MLflow.

12. 🛠️ Pipeline Creation in Databricks

The screenshot displays the Databricks Jobs & Pipelines interface. At the top, a 'New Job' is being configured for July 15, 2025, at 01:43 PM. The interface is divided into two main sections: 'Runs' and 'Job details'.

Runs Section:

- Runs:** A table showing the status of individual runs. The first run (ID: 32114341799...) is in progress (green arrow icon). The second run (ID: 87143350522...) is canceled (gray circle icon). The third run (ID: 79404081527...) failed (red circle icon).
- Tasks:** A grid showing the progress of tasks within each run. The tasks are labeled 'bronze', 'silver', 'Gold', and 'ML_training'. The 'ML_training' task is currently running (green bar).

Job details Section:

- Job ID:** 369846311647455
- Creator:** sachin singh
- Run as:** sachin singh
- Tags:** Add tag
- Description:** Add description
- Git:** Not configured. Add Git settings
- Schedule:** None. Add trigger
- Compute:** sachin singh's Cluster. Single node: Standard_F4 · Release: 16.4.4. View details, Swap, Spark UI, Logs

Jobs & Pipelines > New Job Jul 15, 2025, 01:43 PM >

New Job Jul 15, 2025, 01:43 PM run Lakeflow UI: OFF

[Send feedback](#) [Cancel job run](#) [Repair run](#)

Graph Timeline List

```
graph LR; bronze[bronze] --> silver[silver]; silver --> Gold[Gold]; Gold --> ML_training[ML_training];
```

Job run details

- Job ID: 369846311647455
- Job run ID: 321143417995075
- Launched: Manually
- Started: Jul 15, 2025, 03:16 PM
- Ended: -
- Duration: 3h 43m 42s
- Queue duration: -
- Status: Running [Cancel](#)

[View run events](#)

Compute

- sachin singh's Cluster
- Single node: Standard_F4 · Release: 16.4.4
- [View details](#) [Spark UI](#) [Logs](#) [Metrics](#)

Workflows and jobs were orchestrated in Databricks for automation.

13. MLflow Tracking

Microsoft Azure **databricks** CTRL + P hexa S

New

- Workspace
- Recents
- Catalog
- Jobs & Pipelines
- Compute
- Data Engineering
- Job Runs
- AI/ML
- Playground
- Experiments**
- Features
- Models
- Serving

Experiments > /Users/sachinsingh01703@gmail.com/retailsales/Retailsales/notebooks/mlflow > Runs >

useful-fly-587 [Send feedback](#)

[Reproduce Run](#) [Model registered](#)

Overview Model metrics System metrics Traces Evaluation results Artifacts

Created at	Jul 15, 2025, 01:00 PM
Created by	sachinsingh01703@gmail.com
Experiment ID	966f67ffe2c9438aac18582808a20980
Status	Finished
Run ID	895d3a65c90e4cc992da4d64d2e2c818
Duration	9.9s
Datasets used	—
Tags	Add tags
Source	mlflow
Logged models	sklearn
Registered models	linear v1

Metrics (2)

Metric	Latest	Min	Max
test_score	0.30043358181377...	0.30043358181377...	0.30043358181377...
train_score	0.92995289636037...	0.92995289636037...	0.92995289636037...

Parameters (1)

Parameter	Value
n_estimators	100

All training runs, parameters, and metrics were tracked using MLflow.

