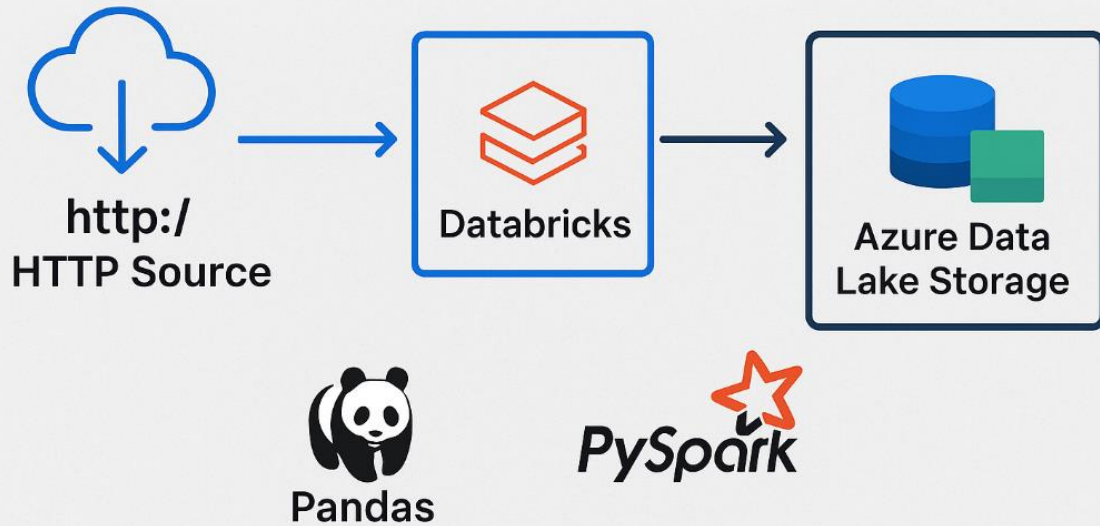


# INCREMENTAL INGESTION FROM PUBLIC HTTP SOURCE TO AZURE DATA LAKE USING DATABRICKS



Run all (1000) adb\_uda\_dev default New SQL editor: OFF

```
1 USE CATALOG pricing_analytics;
2
3 CREATE SCHEMA IF NOT EXISTS processrunlogs;
4
5 CREATE TABLE IF NOT EXISTS processrunlogs.DELTALAKEHOUSE_PROCESS_RUNS (
6   PROCESS_NAME STRING,
7   PROCESSED_FILE_TABLE_DATE DATE,
8   PROCESS_STATUS STRING
9 );
10
11
12 SELECT * FROM pricing_analytics.processrunlogs.deltalakehouse_process_runs;
```

Raw results +

	PROCESS_NAME	PROCESSED_FILE_TABLE_DATE	PROCESS_STATUS
1	dailyPricingSourceIngest	2023-01-02	Completed
2	dailyPricingSourceIngest	2023-01-01	Completed

prm\_processName  
dailyPricingSourceIngest

## Databricks Notebook: Daily Pricing Data Pipeline

12:01 AM (<1s) 2: Import Spark and Pandas Libraries

```
from datetime import datetime
import pandas as pds
from pyspark.sql.functions import *
from pyspark.sql.types import *
```

▶

✓ 12:10 AM (1s)

3: Fetch Next Source File Date

```
processName = dbutils.widgets.get("prm_processName")

nextSourceFileDateSql = f"""
SELECT NVL(MAX(PROCESSED_FILE_TABLE_DATE)+1,'2023-01-01') AS NEXT_SOURCE_FILE_DATE
FROM pricing_analytics.processrunlogs.DELTALAKEHOUSE_PROCESS_RUNS
WHERE PROCESS_NAME = '{processName}' AND PROCESS_STATUS = 'Completed'
"""

nextSourceFileDateDF = spark.sql(nextSourceFileDateSql)
```

▶ nextSourceFileDateDF: pyspark.sql.connect.dataframe.DataFrame = [NEXT\_SOURCE\_FILE\_DATE: date]

▶

✓ 12:01 AM (<1s)

4: Extract Daily Pricing Source and Sink Details

Python

```
#source
dailyPricingSourceBaseUrl = 'https://retailpricing.blob.core.windows.net/'
dailyPricingSourceFolder = 'daily-pricing/'
dailyPricingSourceFileDate = datetime.strptime(str(nextSourceFileDateDF.select('NEXT_SOURCE_FILE_DATE').collect()[0]['NEXT_SOURCE_FILE_DATE']), '%Y-%m-%d').
strftime('%d/%m/%Y')
dailyPricingSourceFileName = f"PW_MW_DR_{dailyPricingSourceFileDate}.csv"
# print(dailyPricingSourceFileName)

#sink
dailyPricingSinkLayerName = 'bronze'
dailyPricingSinkStorageAccountName = 'adlssivadatalakehousedev'
dailyPricingSinkFolderName = 'daily-pricing'
```

> [See performance \(1\)](#) Optimize

▶

✓ 12:01 AM (2s)

5: Load Daily Pricing Data into Pandas DataFrame

```
dailyPricingSourceURL = dailyPricingSourceBaseUrl + dailyPricingSourceFolder + dailyPricingSourceFileName
# print(dailyPricingSourceURL)

dailyPricingPandasDF = pds.read_csv(dailyPricingSourceURL)
# print(dailyPricingPandasDF)
```

▶ dailyPricingPandasDF: pandas.core.frame.DataFrame = [DATE\_OF\_PRICING: object, ROW\_ID: int64 ... 10 more fields]

▶

✓ 12:01 AM (<1s)

6: Convert Pandas DataFrame to Spark DataFrame

Python

```
dailyPricingSparkDF = spark.createDataFrame(dailyPricingPandasDF)
# print(dailyPricingSparkDF)
```

▶ dailyPricingSparkDF: pyspark.sql.connect.dataframe.DataFrame = [DATE\_OF\_PRICING: string, ROW\_ID: long ... 10 more fields]

▶

✓ 12:01 AM (1s)

7: Save Daily Pricing Data to Azure Data Lake

```
dailyPricingSinkFolderPath = f"abfss://{dailyPricingSinkLayerName}@{dailyPricingSinkStorageAccountName}.dfs.core.windows.net/{dailyPricingSinkFolderName}"
# print(dailyPricingSinkFolderPath)

(
    dailyPricingSparkDF
    .withColumn("source_file_load_date", current_timestamp())
    .write
    .mode("append")
    .option("header", "true")
    .csv(dailyPricingSinkFolderPath)
)
```

> [See performance \(1\)](#) Optimize

▶

✓ 12:01 AM (1s)

8: Insert Completed Process Run Log

```
processFileDate = nextSourceFileDateDF.select('NEXT_SOURCE_FILE_DATE').collect()[0]['NEXT_SOURCE_FILE_DATE']
processStatus = "Completed"

processInsertSql = f"""
INSERT INTO pricing_analytics.processrunlogs.DELTALAKEHOUSE_PROCESS_RUNS VALUES('{processName}', '{processFileDate}', '{processStatus}')
"""

spark.sql(processInsertSql)
```

> [See performance \(2\)](#)

DataFrame[num\_affected\_rows: bigint, num\_inserted\_rows: bigint]

```
▶ ✓ 12:01 AM (<1s) 9: Remove Daily Pricing Sink Folder Path
# dbutils.fs.rm(dailyPricingSinkFolderPath, True)
```

```
▶ ✓ 12:01 AM (<1s) 10: Delete Process Run Logs from DeltaLakehouse
# %sql
# delete from pricing_analytics.processrunlogs.DELTALAKEHOUSE_PROCESS_RUNS
```

[Shift+Enter] to run and move to next cell  
[Ctrl+Shift+P] to open the command palette

JOBs & PIPELINES >

## Job-Ingest-Daily-Pricing-HTTP-Source-Data ☆ Lakeflow Jobs UI: OFF ▾

Runs Tasks

▶ ⏏ ↶ ⏏

dailyPricingSourceIngest  
...gest-Daily-Pricing-HTTP-Source-Data

+ Add task

Task name\* ⓘ

dailyPricingSourceIngest

Type\*

Notebook ▾

Source\* ⓘ

Workspace ▾

Path\* ⓘ

...mail.onmicrosoft.com/01-Ingestion/01-Ingest-Daily-Pricing-HTTP-Source-Data 🔗 ▾ 🔗

Compute\* ⓘ

Serverless ▾ Autoscaling ▾

Dependent libraries ⓘ

+ Add

Parameters ⓘ

UI JSON

Key

Value

prm\_processName

dailyPricingSourceIngest { }

+ Add

Notifications ⓘ

+ Add

Cancel

Save task

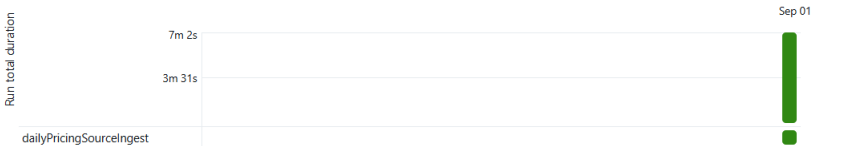
JOBs & PIPELINES >

## Job-Ingest-Daily-Pricing-HTTP-Source-Data ☆ Lakeflow Jobs UI: OFF ▾

Send feedback

Runs Tasks

### Runs



Tasks

Go to the latest successful run

Cancel runs ▾

Start time	Run ID	Launched	Duration	Spark	Status	Error code	Run parameters	
Sep 01, 2025, 12:09...	223131249...	Manually	7m 3s	Logs	✓ Succeed...			⋮

### Job details

Job ID

717454217166517 🔗

Creator

👤 siva sanagondla

Run as ⓘ

👤 siva sanagondla ✎

Tags ⓘ

Add tag

Description

Add description

Lineage ⓘ

1 upstream table, 1 downstr

Performance optimized ⓘ

🔘

### Git

Not configured

Add Git settings

### Schedules & Triggers

None

Home > Resource groups > RG-UGA-DEV > adriavidatalakehousedev > Containers >

bronze

Container

Search

+ Add Directory Upload Refresh Delete Copy Paste Rename Acquire lease Break lease Edit columns

Overview

Diagnose and solve problems

Access Control (IAM)

Settings

bronze > daily-pricing

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

Search blobs by prefix (case-sensitive)

Only show active objects

Showing all 21 items

Name	Last modified	Access tier	Blob type	Size	Lease state
...					...
_SUCCESS	01/09/2025, 02:16:25	Hot (Inferred)	Block blob	0	Available
_committed_2623440797420723808	01/09/2025, 02:01:11	Hot (Inferred)	Block blob	758 B	Available
_committed_6892886767974528227	01/09/2025, 02:16:25	Hot (Inferred)	Block blob	758 B	Available
_startat_2623440797420723808	01/09/2025, 02:01:11	Hot (Inferred)	Block blob	0	Available
_startat_6892886767974528227	01/09/2025, 02:16:24	Hot (Inferred)	Block blob	0	Available
par-00000-e6-2923440797420723808-8f85da69-3a95-4133-82d8-c3a296475179-221-1-c000.csv	01/09/2025, 02:01:11	Hot (Inferred)	Block blob	181.71 KiB	Available
par-00000-e6-4892886767974528227-6887802a-0986-4862-9253-ca07b6da767b-126-1-c000.csv	01/09/2025, 02:16:25	Hot (Inferred)	Block blob	195.4 KiB	Available
par-00001-e6-2923440797420723808-8f85da69-3a95-4133-82d8-c3a296475179-221-1-c000.csv	01/09/2025, 02:01:11	Hot (Inferred)	Block blob	180.2 KiB	Available
par-00001-e6-4892886767974528227-6887802a-0986-4862-9253-ca07b6da767b-135-1-c000.csv	01/09/2025, 02:16:25	Hot (Inferred)	Block blob	193.86 KiB	Available
par-00003-e6-2923440797420723808-8f85da69-3a95-4133-82d8-c3a296475179-221-1-c000.csv	01/09/2025, 02:01:11	Hot (Inferred)	Block blob	181.22 KiB	Available
par-00002-e6-4892886767974528227-6887802a-0986-4862-9253-ca07b6da767b-127-1-c000.csv	01/09/2025, 02:16:25	Hot (Inferred)	Block blob	194.94 KiB	Available
par-00003-e6-2923440797420723808-8f85da69-3a95-4133-82d8-c3a296475179-222-1-c000.csv	01/09/2025, 02:01:11	Hot (Inferred)	Block blob	178.8 KiB	Available
par-00003-e6-4892886767974528227-6887802a-0986-4862-9253-ca07b6da767b-131-1-c000.csv	01/09/2025, 02:16:25	Hot (Inferred)	Block blob	182.27 KiB	Available
par-00004-e6-2923440797420723808-8f85da69-3a95-4133-82d8-c3a296475179-223-1-c000.csv	01/09/2025, 02:01:11	Hot (Inferred)	Block blob	187.22 KiB	Available
par-00004-e6-4892886767974528227-6887802a-0986-4862-9253-ca07b6da767b-132-1-c000.csv	01/09/2025, 02:16:25	Hot (Inferred)	Block blob	194.64 KiB	Available
par-00005-e6-2923440797420723808-8f85da69-3a95-4133-82d8-c3a296475179-227-1-c000.csv	01/09/2025, 02:01:11	Hot (Inferred)	Block blob	178.96 KiB	Available
par-00005-e6-4892886767974528227-6887802a-0986-4862-9253-ca07b6da767b-133-1-c000.csv	01/09/2025, 02:16:25	Hot (Inferred)	Block blob	182.29 KiB	Available
par-00006-e6-2923440797420723808-8f85da69-3a95-4133-82d8-c3a296475179-228-1-c000.csv	01/09/2025, 02:01:11	Hot (Inferred)	Block blob	180.19 KiB	Available
par-00006-e6-4892886767974528227-6887802a-0986-4862-9253-ca07b6da767b-128-1-c000.csv	01/09/2025, 02:16:25	Hot (Inferred)	Block blob	193.87 KiB	Available
par-00007-e6-2923440797420723808-8f85da69-3a95-4133-82d8-c3a296475179-224-1-c000.csv	01/09/2025, 02:01:11	Hot (Inferred)	Block blob	174.69 KiB	Available
par-00007-e6-4892886767974528227-6887802a-0986-4862-9253-ca07b6da767b-129-1-c000.csv	01/09/2025, 02:16:25	Hot (Inferred)	Block blob	188.29 KiB	Available

Jobs & Pipelines >

Job-Ingest-Daily-Pricing-HTTP-Source-Data ☆ Lakelabs Jobs UI: OFF

Send feedback Run now

Runs Tasks

Runs

Started before < Previous Next >

Run total duration

7m 53s

3m 57s

dailyPricingSourceIngest

Sep 01

Tasks

Go to the latest successful run

Cancel runs

Start time	Run ID	Launched	Duration	Spark	Status	Error code	Run parameters
Sep 01, 2025, 07:36 PM	883201097923...	Manually	7m 54s	Logs	Succeeded		
Sep 01, 2025, 12:09 AM	223131249982...	Manually	7m 3s	Logs	Succeeded		

Job details

Job ID

717454217166517

Creator

siva sanagondla

Run as

siva sanagondla

Tags

Add tag

Description

Add description

Lineage

1 upstream table, 1 downstream table

Performance optimized

Git

Not configured

Add Git settings

Schedules & Triggers

None

Add trigger

Compute

Serverless

Swap

```
11
12 | SELECT * FROM pricing_analytics.processrunlogs.deltalakehouse_process_runs;
```

Raw results			
	PROCESS_NAME	PROCESSED_FILE_TABLE_DATE	PROCESS_STATUS
1	dailyPricingSourceIngest	2023-01-02	Completed
2	dailyPricingSourceIngest	2023-01-03	Completed
3	dailyPricingSourceIngest	2023-01-01	Completed

**Resource groups** > RGP-LGA-DEV > adfs\datastorage.azure | Containers

Container

+ Add Directory Upload Refresh | Delete Copy Paste Rename Acquire lease Break lease Edit volume

**Overview**
  
Diagnose and solve problems
   
Access Control (IAM)
   
Settings

Authorization method: Microsoft Entra user account ([Switch to Access key](#))

Search blobs by prefix (name-sensitive)

Sorting all 30 items

Name	Last modified ↓	Access tier	Blob type	Size	Lease state
_SUCCESSSS	01/08/2025, 19:43:42	Hot (Inferred)	Block blob	0	Available
_commitment_1215546907644410096	01/08/2025, 19:43:42	Hot (Inferred)	Block blob	756 B	Available
_atahaad_123546607644410096	01/08/2025, 19:43:42	Hot (Inferred)	Block blob	0	Available
pam-00000-fcp-123546607644410096-0ctfc380-ewda-dclab-hw5-5378b6mnp123-127-1-000.csv	01/08/2025, 19:43:42	Hot (Inferred)	Block blob	195.42 KiB	Available
pam-00001-fcp-123546607644410096-0ctfc380-ewda-dclab-hw5-5378b6mnp123-128-1-000.csv	01/08/2025, 19:43:42	Hot (Inferred)	Block blob	183.91 KiB	Available
pam-00002-fcp-123546607644410096-0ctfc380-ewda-dclab-hw5-5378b6mnp123-131-1-000.csv	01/08/2025, 19:43:42	Hot (Inferred)	Block blob	184.89 KiB	Available
pam-00003-fcp-123546607644410096-0ctfc380-ewda-dclab-hw5-5378b6mnp123-132-1-000.csv	01/08/2025, 19:43:42	Hot (Inferred)	Block blob	182.42 KiB	Available
pam-00004-fcp-123546607644410096-0ctfc380-ewda-dclab-hw5-5378b6mnp123-129-1-000.csv	01/08/2025, 19:43:42	Hot (Inferred)	Block blob	184.87 KiB	Available
pam-00005-fcp-123546607644410096-0ctfc380-ewda-dclab-hw5-5378b6mnp123-133-1-000.csv	01/08/2025, 19:43:42	Hot (Inferred)	Block blob	182.33 KiB	Available
pam-00006-fcp-123546607644410096-0ctfc380-ewda-dclab-hw5-5378b6mnp123-134-1-000.csv	01/08/2025, 19:43:42	Hot (Inferred)	Block blob	183.92 KiB	Available
pam-00007-fcp-123546607644410096-0ctfc380-ewda-dclab-hw5-5378b6mnp123-130-1-000.csv	01/08/2025, 19:43:42	Hot (Inferred)	Block blob	188.34 KiB	Available
_commitment_888288787878132327	01/08/2025, 00:16:25	Hot (Inferred)	Block blob	758 B	Available
pam-00000-fcp-489288787878132327-68f702ca-099c-4b4d-8253-ca0ff7bd3ca7b0-126-1-000.csv	01/08/2025, 00:16:25	Hot (Inferred)	Block blob	195.4 KiB	Available
pam-00001-fcp-489288787878132327-68f702ca-099c-4b4d-8253-ca0ff7bd3ca7b0-130-1-000.csv	01/08/2025, 00:16:25	Hot (Inferred)	Block blob	183.88 KiB	Available
pam-00002-fcp-489288787878132327-68f702ca-099c-4b4d-8253-ca0ff7bd3ca7b0-127-1-000.csv	01/08/2025, 00:16:25	Hot (Inferred)	Block blob	184.84 KiB	Available
pam-00003-fcp-489288787878132327-68f702ca-099c-4b4d-8253-ca0ff7bd3ca7b0-131-1-000.csv	01/08/2025, 00:16:25	Hot (Inferred)	Block blob	182.37 KiB	Available
pam-00004-fcp-489288787878132327-68f702ca-099c-4b4d-8253-ca0ff7bd3ca7b0-132-1-000.csv	01/08/2025, 00:16:25	Hot (Inferred)	Block blob	184.84 KiB	Available
pam-00005-fcp-489288787878132327-68f702ca-099c-4b4d-8253-ca0ff7bd3ca7b0-133-1-000.csv	01/08/2025, 00:16:25	Hot (Inferred)	Block blob	182.29 KiB	Available
pam-00006-fcp-489288787878132327-68f702ca-099c-4b4d-8253-ca0ff7bd3ca7b0-128-1-000.csv	01/08/2025, 00:16:25	Hot (Inferred)	Block blob	183.87 KiB	Available
pam-00007-fcp-489288787878132327-68f702ca-099c-4b4d-8253-ca0ff7bd3ca7b0-129-1-000.csv	01/08/2025, 00:16:25	Hot (Inferred)	Block blob	188.29 KiB	Available
_commitment_2823440787480728808	01/08/2025, 00:01:11	Hot (Inferred)	Block blob	758 B	Available
pam-00000-fcp-2823440787480728808-0f8a6d69-3a6e-4133-82cb-02d98c7817b8-221-1-000.csv	01/08/2025, 00:01:11	Hot (Inferred)	Block blob	181.71 KiB	Available
pam-00001-fcp-2823440787480728808-0f8a6d69-3a6e-4133-82cb-02d98c7817b8-225-1-000.csv	01/08/2025, 00:01:11	Hot (Inferred)	Block blob	180.2 KiB	Available
pam-00002-fcp-2823440787480728808-0f8a6d69-3a6e-4133-82cb-02d98c7817b8-228-1-000.csv	01/08/2025, 00:01:11	Hot (Inferred)	Block blob	181.22 KiB	Available
pam-00003-fcp-2823440787480728808-0f8a6d69-3a6e-4133-82cb-02d98c7817b8-222-1-000.csv	01/08/2025, 00:01:11	Hot (Inferred)	Block blob	178.8 KiB	Available
pam-00004-fcp-2823440787480728808-0f8a6d69-3a6e-4133-82cb-02d98c7817b8-223-1-000.csv	01/08/2025, 00:01:11	Hot (Inferred)	Block blob	181.22 KiB	Available
pam-00005-fcp-2823440787480728808-0f8a6d69-3a6e-4133-82cb-02d98c7817b8-227-1-000.csv	01/08/2025, 00:01:11	Hot (Inferred)	Block blob	178.66 KiB	Available
pam-00006-fcp-282344078748072880					

A large, bold, 3D white text graphic reading "THANK YOU" against a solid blue background. The letters are thick and blocky, with a slight shadow on the right side, giving them a three-dimensional appearance. The text is centered and occupies most of the frame.