

Running Pipeline

Microsoft Azure | Data Factory | adf-netflix-vass

Factory Resources

- Pipelines: 1
 - pipeline1
- Datasets: 3
 - Dataset_Github
 - Dataset_Sink_Datalake
 - Dataset_Validation
- Data flows: 0
- Power Query: 0

Activities

- General
 - Append variable
 - Set variable

Web: Github_Metadata

Set variable: (X) Set variable1

Validation: Validation_Raw_Container

ForEach: ForEachFile

Output

Activity name	Activity status	Activity type	Run start	Duration	Integration runtime
Copy Github data	Queued	Copy data	9/4/2025, 11:44:26 PM	6s	
Copy Github data	Queued	Copy data	9/4/2025, 11:44:26 PM	6s	
Copy Github data	Queued	Copy data	9/4/2025, 11:44:26 PM	6s	
Copy Github data	Queued	Copy data	9/4/2025, 11:44:26 PM	6s	
ForEachFile	In progress	ForEach	9/4/2025, 11:44:25 PM	7s	
Validation_Raw_Container	Succeeded	Validation	9/4/2025, 11:44:18 PM	5s	
Set variable1	Succeeded	Set variable	9/4/2025, 11:44:16 PM	Less than 1s	
Github_Metadata	Succeeded	Web	9/4/2025, 11:44:10 PM	5s	AutoResolveIntegrationRuntime (East US)

After Pipeline run

Microsoft Azure | Data Factory | adf-netflix-vass

Factory Resources

- Pipelines: 1
 - pipeline1
- Datasets: 3
 - Dataset_Github
 - Dataset_Sink_Datalake
 - Dataset_Validation
- Data flows: 0
- Power Query: 0

Activities

- General
 - Append variable
 - Set variable

Web: Github_Metadata

Set variable: (X) Set variable1

Validation: Validation_Raw_Container

ForEach: ForEachFile

Output

Pipeline run ID: 68e55353-2523-4f42-83af-5e972f3e616b

Pipeline status: Succeeded

Activity name	Activity status	Activity type	Run start	Duration	Integration runtime
Copy Github data	Succeeded	Copy data	9/4/2025, 11:44:26 PM	14s	AutoResolveIntegrationRuntime (Canada Centr)
Copy Github data	Succeeded	Copy data	9/4/2025, 11:44:26 PM	14s	AutoResolveIntegrationRuntime (Canada Centr)
Copy Github data	Succeeded	Copy data	9/4/2025, 11:44:26 PM	14s	AutoResolveIntegrationRuntime (Canada Centr)
Copy Github data	Succeeded	Copy data	9/4/2025, 11:44:26 PM	13s	AutoResolveIntegrationRuntime (Canada Centr)
ForEachFile	Succeeded	ForEach	9/4/2025, 11:44:25 PM	18s	
Validation_Raw_Container	Succeeded	Validation	9/4/2025, 11:44:18 PM	5s	
Set variable1	Succeeded	Set variable	9/4/2025, 11:44:16 PM	Less than 1s	

After running pipeline checking if the data is stored in the sink

The first screenshot shows the 'bronze' container overview. The authentication method is 'Access key'. The container contains four items:

Name	Last modified	Access tier	Blob type	Size	Lease state
netflix_cast	9/4/2025, 11:44:37 PM				...
netflix_category	9/4/2025, 11:44:38 PM				...
netflix_countries	9/4/2025, 11:44:37 PM				...
netflix_directors	9/4/2025, 11:44:37 PM				...

The second screenshot shows the 'netflix_cast' directory selected. It contains one item:

Name	Last modified	Access tier	Blob type	Size	Lease state
netflix_cast.csv	9/4/2025, 11:44:37 PM	Hot (Inferred)	Block blob	1.02 MiB	Available

Azure Databricks deployment

The deployment is in progress. The deployment name is 'RG-NetflixDataset_adb-netflix-vass'. The start time is 9/4/2025, 4:16:31 PM. The correlation ID is 884ed6a1-43f0-404d-8236-b654f5c82e2c. The resource group is 'RG-NetflixDataset'.

Deployment details:

Resource	Type	Status	Operation details
adb-netflix-vass	Azure Databricks Service	Created	Operation details

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

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 >](#)

After deployment

The screenshot shows the Microsoft Azure portal interface. At the top, there's a navigation bar with 'Microsoft Azure', an 'Upgrade' button, a search bar, and a 'Copilot' button. The user's profile is 'vsk230045@utdallas.edu THE UNIVERSITY OF TEXAS AT D...'. The main heading is 'Home > RG-NetflixDataset_adb-netflix-vass | Overview'. Below this, there's a 'Deployment' section with a search bar and buttons for 'Delete', 'Cancel', 'Redeploy', 'Download', and 'Refresh'. The 'Overview' tab is selected, showing a green checkmark and the text 'Your deployment is complete'. It lists the deployment name as 'RG-NetflixDataset_adb-netflix-vass', the subscription as 'Azure subscription 1', and the resource group as 'RG-NetflixDataset'. The start time is '9/4/2025, 4:18:29 PM' and the correlation ID is '884ed6a1-43f0-404d-8236-b654f5c82e2c'. There are sections for 'Deployment details' and 'Next steps', with a 'Go to resource' button. A 'Give feedback' link is also present. On the right, there are three informational cards: 'Cost management' (Get notified to stay within your budget...), 'Microsoft Defender for Cloud' (Secure your apps and infrastructure...), and 'Free Microsoft tutorials' (Start learning today...).

Dataricks Autoloader

Reading data

The screenshot shows the Databricks workspace interface. The top bar includes 'Microsoft Azure', 'databricks', a search bar, and a 'CTRL + P' shortcut. The user's profile is 'adb-netflix-vass'. The left sidebar shows a 'New' button and a 'Workspace' section with a 'Catalog' tab. The 'Catalog' tab shows a search bar and a list of catalogs: 'My organization', 'system', 'catalog_netflix', 'main', 'Delta Shares Received', 'samples', and 'Legacy'. The 'Legacy' catalog is selected, showing a 'hive_metastore'. The main area displays a notebook titled 'Incremental Data Loading using AutoLoader'. The notebook has five cells. The first cell is a SQL statement: 'CREATE SCHEMA catalog_netflix.schema;'. The second cell is a Python statement: 'checkpoint_path = "abfss://silver@netflixprojstorageacc.dfs.core.windows.net/checkpoints"'. The third cell is a Python statement: 'df = spark.readStream().format("cloudFiles").option("cloudFiles.format", "csv").option("cloudFiles.schemaLocation", checkpoint_path).load("abfss://raw@netflixprojstorageacc.dfs.core.windows.net")'. The fourth cell is a Python statement: 'display(df)'. The fifth cell is a Python statement: 'display_query_1 (id: 818bea72-6a7a-436c-9a10-f007607ab612)'. The notebook is running on a cluster named 'Vasanth S's Cluster 202...'. The bottom status bar shows 'Last updated: 10 seconds ago'.

Microsoft Azure databricks

Search data, notebooks, recents, and more... CTRL + P

Workspace

Catalog

Type to search...

For you All

- My organization
 - system
 - catalog_netflix
 - main
- Delta Shares Received
- samples
- Legacy
 - hive_metastore

1_Autoloader x +

File Edit View Run Help Python Tabs: ON v Last e... Interrupt Vasanth S's Cluster 202... Schedule Share

Interrupt 5 Python

display(df)

(1) Spark Jobs

display_query_1 (id: 818bea72-6a7a-436c-9a10-f007607ab612) Last updated: 10 seconds ago

Dashboard Raw Data

Input vs. Processing Rate
records per second

0 rec/s 0 rec/s
Input rate Processing rate

Batch Duration
in milliseconds

922.1 ms 1 ms
Average Latest

Table	duration_minutes	duration_seasons	type	title
1	90	null	Movie	Norm of the North: King Sized Adventure
2	94	null	Movie	Jandino: Whatever it Takes
3	null	1	TV Show	Transformers Prime
4	null	1	TV Show	Transformers: Robots in Disguise
5	99	null	Movie	#realityhigh
6	null	1	TV Show	Apaches

Writing data

Microsoft Azure databricks

Search data, notebooks, recents, and more... CTRL + P

Workspace

Catalog

Type to search...

For you All

- My organization
 - system
 - catalog_netflix
 - main
- Delta Shares Received
- samples
- Legacy
 - hive_metastore

1_Autoloader x +

File Edit View Run Help Python Tabs: ON v Last e... Interrupt Vasanth S's Cluster 202... Schedule Share

Interrupt 6 Python

```

1 df.writeStream\
2   .option("checkpointLocation", checkpoint_path)\
3   .trigger(processingTime="10 seconds")\
4   .start("abfss://bronze@netflixprojstorageacc.dfs.core.windows.net/netflix_titles")

```

(1) Spark Jobs

a82bb863-4108-4d4a-80e4-703d1a3263e4 Last updated: 5 seconds ago

Dashboard Raw Data

Input vs. Processing Rate
records per second

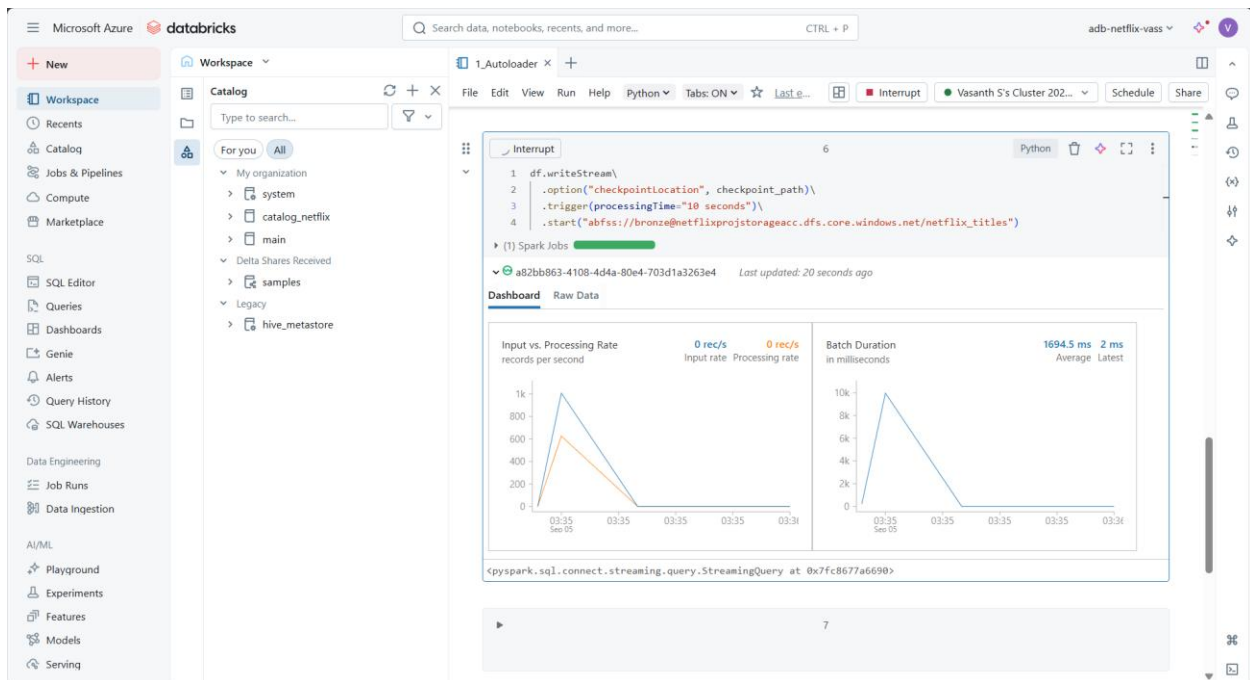
0 rec/s 0 rec/s
Input rate Processing rate

Batch Duration
in seconds

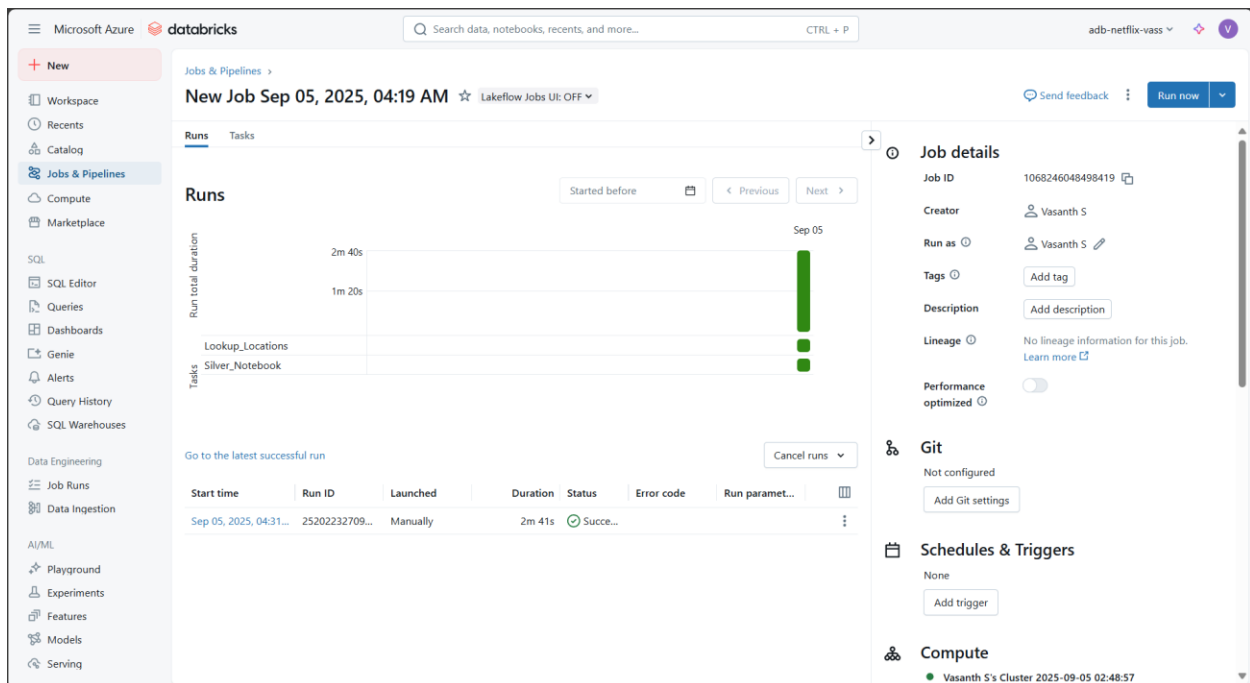
3.4 s 0 s
Average Latest

<pyspark.sql.connect.streaming.query.StreamingQuery at 0x7fc8677a6690>

7



Notebook pipeline run successfully in Databricks



Microsoft Azure databricks

Search data, notebooks, recents, and more... CTRL + P

adb-netflix-vass

+ New

- Workspace
- Recents
- Catalog
- Jobs & Pipelines
- Compute
- Marketplace

SQL

- SQL Editor
- Queries
- Dashboards
- Genie
- Alerts
- Query History
- SQL Warehouses

Data Engineering

- Job Runs
- Data Ingestion

AI/ML

- Playground
- Experiments
- Features
- Models
- Serving

Jobs & Pipelines >

New Job Sep 05, 2025, 04:19 AM ☆ Lakeflow Jobs UI: OFF

Runs Tasks

Lookup_Locations

.../Project_Netflix/3_Lookup_Notebook

Vasanth S's Cluster 2025-09-05 02:48:...

Silver_Notebook

/Workspace/Project_Netflix/2_Silver

Vasanth S's Cluster 2025-09-05 02:48:...

+ Add task

No task selected
Choose a task from the graph to edit its properties

Job details

Job ID 1068246048498419

Creator Vasanth S

Run as Vasanth S

Tags Add tag

Description Add description

Lineage No lineage information for this job. Learn more

Performance optimized

Git Not configured
Add Git settings

Schedules & Triggers None
Add trigger

Compute Vasanth S's Cluster 2025-09-05 02:48:57

After running silver notebook 4

Microsoft Azure

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

Copilot

vasanth.utilities2@gmail...
DEFAULT DIRECTORY

Home > Resource_Group_Netflix > netflixprojstorageacc | Containers >

silver
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

- Shared access tokens
- Manage ACL
- Access policy
- Properties
- Metadata

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

Search blobs by prefix (case-sensitive)

Only show active objects

Showing all 6 items

Name	Last modified	Access tier	Blob type	Size	Lease state
checkpoints	9/5/2025, 3:28:06 AM				...
netflix_cast	9/5/2025, 4:33:20 AM				...
netflix_category	9/5/2025, 4:34:22 AM				...
netflix_countries	9/5/2025, 4:33:49 AM				...
netflix_directors	9/5/2025, 4:32:47 AM				...
netflix_titles	9/5/2025, 5:33:57 AM				...

Add or remove favorites by pressing Ctrl+Shift+F

Final Silver Pipeline

The screenshot displays the Databricks Jobs & Pipelines interface. The left sidebar shows the navigation menu with 'Jobs & Pipelines' selected. The main area shows a 'New Job' for 'Sep 05, 2025, 05:50 AM'. The 'Runs' tab is active, showing a bar chart of run durations. The 'Job details' panel on the right provides information about the job, including its ID, creator, and tags. The 'Schedules & Triggers' panel shows no triggers are configured. The 'Compute' panel indicates the job is running on 'Vasanth S's Cluster'.

Runs

Task	Run total duration
Lookup_Weekday	1m 37s
If_Week_Day	49s
False_Notebook	
Silver_Master_Data	

Job details

- Job ID: 876189549601642
- Creator: Vasanth S
- Run as: Vasanth S
- Tags: Add tag
- Description: Add description
- Lineage: No lineage information for this job. [Learn more](#)
- Performance optimized: ☐

Schedules & Triggers

None

Compute

Vasanth S's Cluster 2025-09-05 02:48:57

Creating the DLT Pipeline

The screenshot displays the Databricks Jobs & Pipelines interface for creating a new DLT pipeline. The left sidebar shows the navigation menu with 'Jobs & Pipelines' selected. The main area shows the 'DLT_GOLD' pipeline in the 'Development' state. The 'Run' dropdown is set to 'Sep 05, 2025, 06:49 AM'. The 'Pipeline details' panel on the right provides information about the pipeline, including its ID, type, and source code. The 'Event log' at the bottom shows the progress of the pipeline creation, including the 'Creating update' and 'Waiting for resources' steps.

DLT_GOLD

Run: Sep 05, 2025, 06:49 AM

Pipeline details

- Pipeline ID: 7c66877e-ddf3-4bbc-9289-7626ed6c8afe
- Pipeline type: ETL pipeline
- Source code: /Project_Netflix7_DLT
- Run as: Vasanth S
- Tags: None

Event log

Time	Event	Message
11 seconds ago	user_action	User vasanth.utilities2@gmail.com started an update.
11 seconds ago	create_update	Update 398f4a started by USER_ACTION.
10 seconds ago	update_progress	Update 398f4a is WAITING_FOR_RESOURCES.

After creating the DLT pipeline

The screenshot shows the Databricks workspace interface. On the left is a sidebar with navigation options like Workspace, Recents, Catalog, Jobs & Pipelines, Compute, Marketplace, SQL, SQL Editor, Queries, Dashboards, Genie, Alerts, Query History, SQL Warehouses, Data Engineering, Job Runs, Data Ingestion, AI/ML, Playground, Experiments, Features, Models, and Serving. The main area displays the DLT pipeline code in a notebook. The code includes imports for pyspark.sql.functions and defines a streaming table 'gold_stg_netflix_titles' and a table 'gold_netflix_titles'. Below the code, the 'DLT graph' is visible, showing the pipeline's execution flow with nodes for 'gold_netflix_category', 'gold_netflix_category', 'gold_netflix_category', 'gold_netflix_direct', and 'gold_stg_netflix_titles'.

Running the DLT pipeline

The screenshot shows the Databricks Jobs & Pipelines page. The pipeline 'DLT_GOLD' is shown in the 'Development' environment, with a status of 'Running...'. The pipeline details on the right show the Pipeline ID, Pipeline type (ETL pipeline), Source code (/Project_Netflix/7_DLT), Run as (Vasanth S), and Tags (None). The Event log at the bottom shows the pipeline's execution steps, including the creation of streaming tables and the definition of flows.

Event log	Query history
55 seconds ago	dataset_life_cycle
54 seconds ago	dataset_life_cycle
46 seconds ago	flow_definition
46 seconds ago	flow_definition

Completed

Microsoft Azure databricks

Search data, notebooks, recents, and more...

CTRL + P

adb-netflix-vass

New

Workspace

Recents

Catalog

Jobs & Pipelines

Compute

Marketplace

SQL

SQL Editor

Queries

Dashboards

Genie

Alerts

Query History

SQL Warehouses

Data Engineering

Job Runs

Data Ingestion

AI/ML

Playground

Experiments

Features

Models

Serving

Jobs & Pipelines >

DLT_GOLD

Send feedback

Development

Production

Settings

Schedule

Share

Start

Run: Sep 05, 2025, 07:11 AM

Completed

Select tables for refresh

Graph

List

gold_netflix_cat

gold_netflix_catag

gold_netflix_count

gold_netflix_direct

gold_stg_netflixitles

gold_transformed...

gold_netflixitles

Pipeline details

Update details

Pipeline ID7c66877e-ddf3-4bbc-9289-7626ed6c8afe

Pipeline typeETL pipeline

Source code/Project_Netflix/7_DLT

Run asVasanth S

TagsNone

Event log

Query history

All

Info

Warning

Error

Filter...

39 seconds agoflow_progressFlow 'catalog_netflix.dlt_schema.gold_netflixitles' is RUNNING.

35 seconds agoflow_progressCompleted a streaming update of 'catalog_netflix.dlt_schema.gold_netflixitles'.

35 seconds agoflow_progressFlow 'catalog_netflix.dlt_schema.gold_netflixitles' has COMPLETED.

32 seconds agoupdate_progressUpdate cfd2cc is COMPLETED.



Data Quality

Jobs & Pipelines >

DLT_GOLD ☆ Send feedback

Development Production Settings Schedule Share Start

Run: Sep 05, 2025, 07:11 AM Completed Select tables for refresh

Graph List

Completed - 37s
148 0

gold_netflix_count...
Completed - 39s
7.2K 0

gold_netflix_direct...
Completed - 39s
4.9K 0

gold_stg_netflixitles
Completed - 38s
6.2K 0

gold_transformed_...
Completed - 3s
6.2K 0

gold_netflixitles
Completed - 3s
6.2K 4

catalog_netflix.dlt_schema.gold_netflix_titles

Details Expectations Schema Flows

Written > 99.9% (6,232)

Dropped < 0.1% (4)

Expectations All Failures only

Name	Action	Fail %	Failed records
rule 1	DROP	< 0.1%	4