

NAME : AKULA SHARATH CHANDRA
BATCH: DATA ENGINEERING
DATE :21-02-2024

AZURE CODING CHALLENGE

- 1)Exploratory data analysis (EDA) in Databricks &Visualizing data in Databricks.**
- 2)Explain Overview of 3 level namespace and creating Unity Catalog objects.**
- 3)Execute & explain Azure datafactory and its copy activity.**

ANSWERS:

- 1)Exploratory data analysis (EDA) in Databricks &Visualizing data in Databricks?**

EDA:

Exploratory Data Analysis is a data analysis approach that involves summarizing, visualizing, and understanding the main characteristics of a dataset to uncover patterns, trends, relationships.

Here are some common steps involved in EDA using Databricks,

Data Loading: Databricks allows users to load data from various sources such as databases, data lakes, or cloud storage into Spark DataFrames.

Data Profiling: Users can generate summary statistics and descriptive analytics to understand the structure and distribution of data. Databricks offers tools for generating summary statistics, histograms, box plots, and other visualizations.

Data Visualization: Databricks provides integration with popular visualization libraries like Matplotlib, Seaborn, and Plotly, allowing users to create insightful plots and charts to visualize patterns and relationships within the data.

Data Cleaning: EDA often involves identifying and handling missing values, outliers, and inconsistencies in the data. Databricks offers powerful data manipulation and transformation capabilities using Spark SQL and DataFrame APIs.

Feature Engineering: Databricks enables users to create new features or derive insights from existing features to improve model performance. This may involve tasks such as feature scaling, encoding categorical variables, or creating new variables through transformations.

Statistical Analysis: Databricks allows users to perform statistical tests and analyses to validate assumptions, test hypotheses, and uncover relationships between variables.

Interactive Exploration: Databricks notebooks provide an interactive environment where users can iterate on their analysis, visualize results, and collaborate with team members.

Scalability: Databricks leverages the distributed computing capabilities of Apache Spark, enabling EDA on large-scale datasets that may not fit into memory on a single machine.

Microsoft Azure | Overview

Your deployment is complete

Deployment name : resourcegroup1069_1069codingchallenge Start time : 21/02/2024, 10:29:51
Subscription : Azure subscription 1 Correlation ID : 9357812a-a030-4d51-9b66-9...
Resource group : resourcegroup1069

Deployment 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.

26°C Haze

Search

Home > resourcegroup1069_1069codingchallenge | Overview >

1069codingchallenge | Overview

resourcegroup1069

Overview

Activity log

Access control (IAM)

Tags

Diagnose and solve problems

Virtual Network Peerings

Encryption

Networking

Properties

Locks

Diagnostic settings

CLI / PS

Launch Workspace

Documentation

Getting Started

Import Data from File

Import Data from Azure Storage

26°C Haze

Microsoft Azure | Overview

Home > resourcegroup1069_1069codingchallenge | Overview >

1069codingchallenge | Overview

resourcegroup1069

Overview

Activity log

Access control (IAM)

Tags

Diagnose and solve problems

Virtual Network Peerings

Encryption

Networking

Properties

Locks

Monitoring

Diagnostic settings

Automation

CLI / PS

Launch Workspace

Documentation

Getting Started

Import Data from File

Import Data from Azure Storage

26°C Haze

Microsoft Azure  databricks Search data, notebooks, recents, and more... CTRL + P 1069codingchallenge azuser1069_mml.local@ihtl.onmicrosoft.com's Cluster More ... Terminate Edit

New Workspace Recents Catalog Workflows Compute SQL Editor Queries Dashboards Alerts Query History SQL Warehouses Data Engineering Job Runs Data Ingestion Delta Live Tables Machine Learning Experiments 26°C Haze

Compute > Preview  Send feedback Configuration Notebooks (0) Libraries Event log Spark UI Driver logs Metrics Apps Spark compute UI - Master

Single user azuser1069_mml.local@ihtl.onmicrosoft.com's Cluster

1 Driver 14 GB Memory, 4 Cores
Runtime 14.3.x-cpu-nl-scala2.12
Standard_DS3_v2 0.75 DBU/h

Performance
Databricks Runtime Version 14.3 LTS ML (includes Apache Spark 3.5.0, Scala 2.12)
 Use Photon Acceleration
Node type Standard_DS3_v2 14 GB Memory, 4 Cores
 Terminate after 100 minutes of inactivity

Tags
No custom tags
Automatically added tags

Advanced options

Search              ENG US 10:45 AM 2024-02-21

New Workspace Recents Catalog Workflows Compute SQL Editor Queries Dashboards Alerts Query History SQL Warehouses Data Engineering Job Runs Data Ingestion Delta Live Tables Machine Learning Experiments NH163 / Uppal R... Construction

2024-02-21 - DBFS Example (1) Python 

File Edit View Run Help Last edit was now New cell UI: OFF

```
11 df = spark.read.format(file_type) \
12 .option("inferSchema", infer_schema) \
13 .option("header", first_row_is_header) \
14 .option("sep", delimiter) \
15 .load(file_location)
16
17 display(df)
```

(2) Spark Jobs

df: pyspark.sql.dataframe.DataFrame = [c0: string, c1: string ... 3 more fields]

Table + New result table: OFF

_c0	_c1	_c2	_c3	_c4
1	Product	Category	Price	Quantity
2	Laptop	Electronics	1200	10
3	Smartphone	Electronics	800	15
4	Headphones	Electronics	100	50
5	Jacket	Apparel	150	20
6	Sneakers	Apparel	80	30
7	Coffee Maker	Appliances	200	5

9 rows | 1.09 seconds runtime

Refreshed now

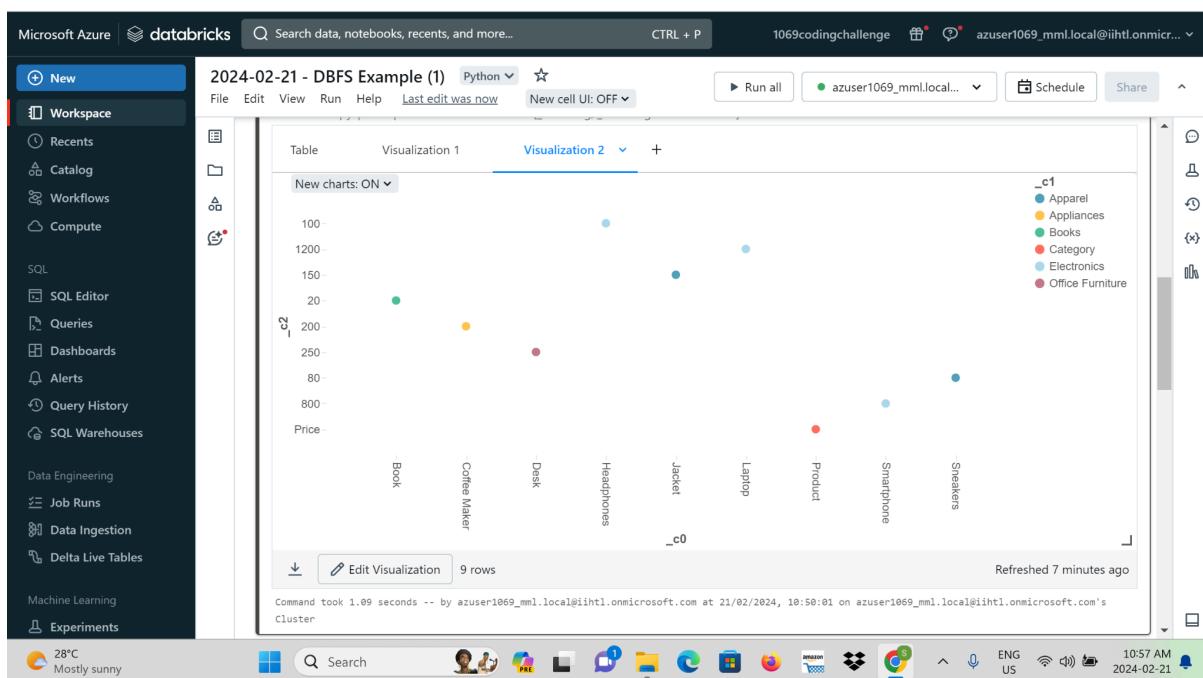
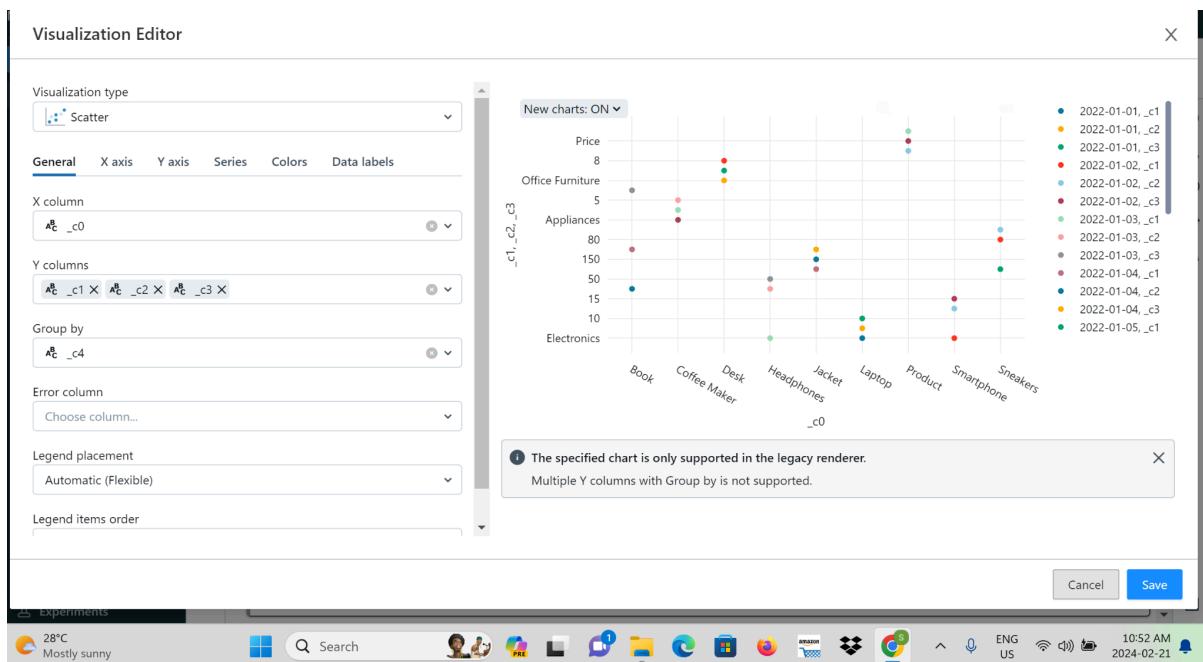
Command took 1.09 seconds -- by azuser1069_mml.local@ihtl.onmicrosoft.com at 21/02/2024, 10:50:01 on azuser1069_mml.local@ihtl.onmicrosoft.com's Cluster

Search              ENG US 10:50 AM 2024-02-21

Table +

Visualization

_c0	_c1	_c2	_c3	_c4
1	Product	Category	Price	Quantity
	Data Profile			SalesDate



3) Execute & explain Azure datafactory and its copy activity?

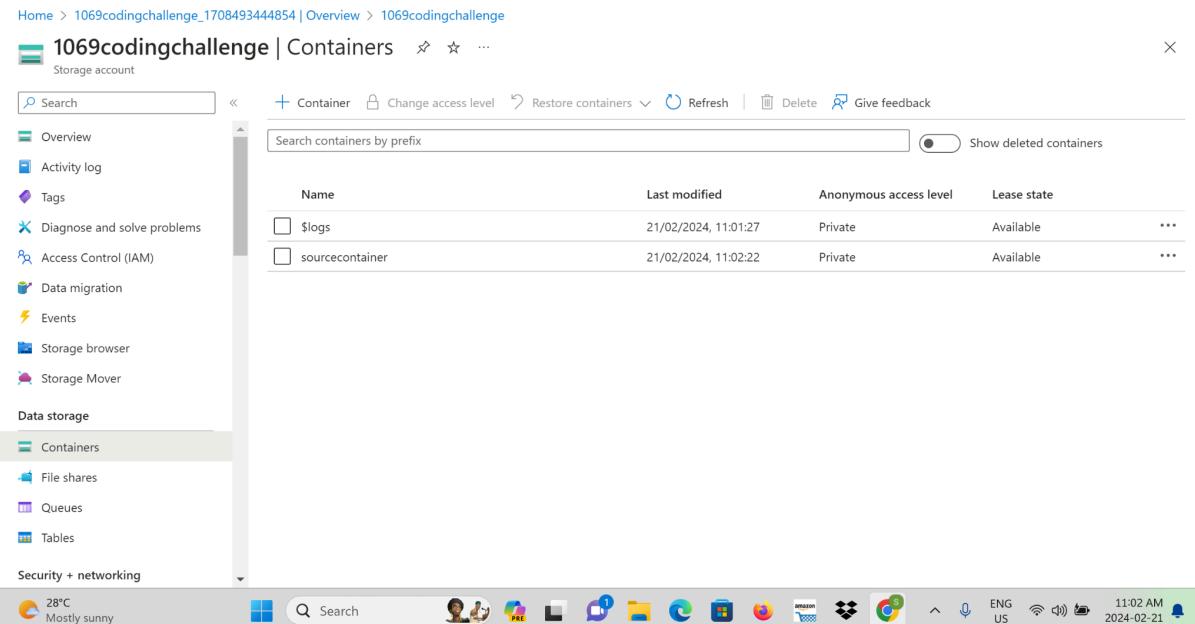
A) DATA FACTORY: Azure Data Factory is a cloud-based data integration service that allows you to create data-driven workflows in the cloud for orchestrating and automating data movement and data transformation. ADF does not store any data itself. It allows you to create data-driven workflows to orchestrate the movement of data between supported data stores and then process the data using compute services in other regions or in an on-premise environment. It also allows you to monitor and manage workflows using both programmatic and UI mechanisms.

- It works in 3 steps
- i) Connect and collect
- ii) Transform and enrich
- iii) Publish

COPY ACTIVITY: Copy Activity in Azure Data Factory copies data from a source data store to a sink data store.

→ to create a data factory account firstly we need to create two storage accounts one is for source group and one is for destination group.

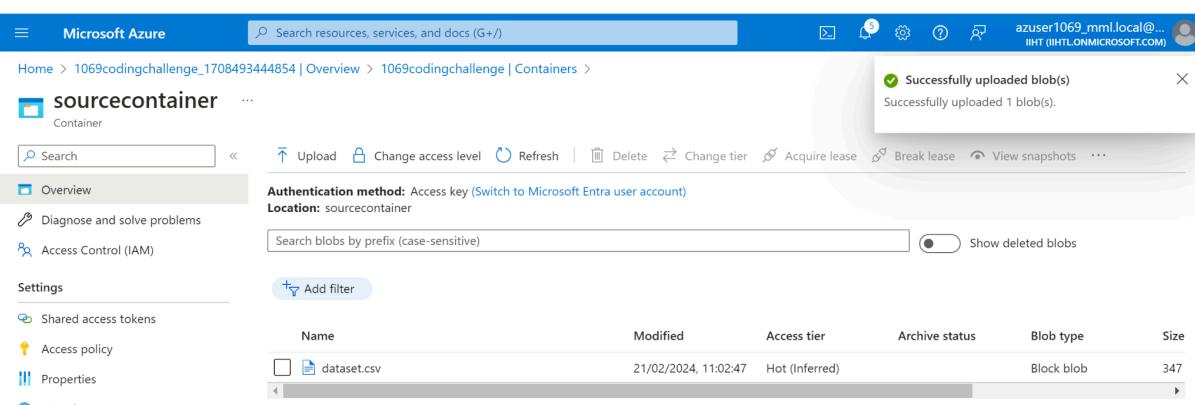
→ I have created a source storage account and then created a container and uploaded a file in it .



The screenshot shows the 'Containers' page for the '1069codingchallenge' storage account. The left sidebar includes links for Overview, Activity log, Tags, Diagnose and solve problems, Access Control (IAM), Data migration, Events, Storage browser, Storage Mover, Data storage (Containers selected), File shares, Queues, and Tables. The main area displays a table of containers:

Name	Last modified	Anonymous access level	Lease state
\$logs	21/02/2024, 11:01:27	Private	Available
sourcecontainer	21/02/2024, 11:02:22	Private	Available

At the bottom right, there is a success message: "Successfully uploaded blob(s) Successfully uploaded 1 blob(s)".



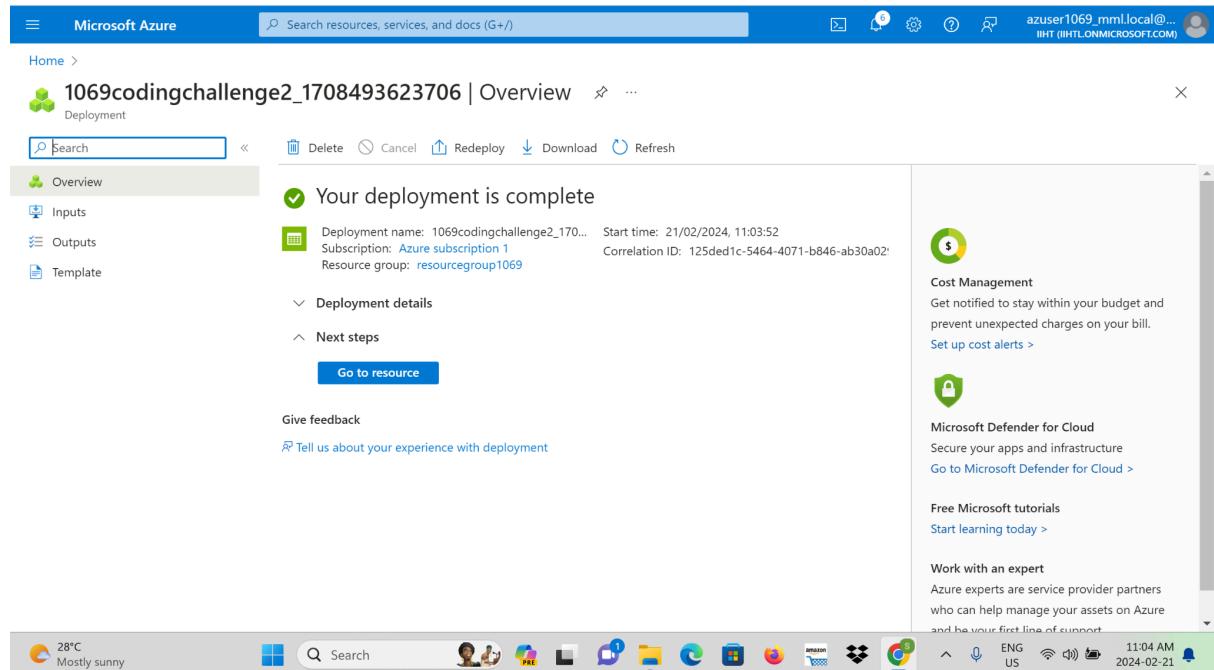
The screenshot shows the 'Containers' page for the 'sourcecontainer' storage account. The left sidebar includes links for Overview (selected), Diagnose and solve problems, Access Control (IAM), Settings (Shared access tokens, Access policy, Properties, Metadata), and Add filter. The main area displays a table of blobs:

Name	Modified	Access tier	Archive status	Blob type	Size
dataset.csv	21/02/2024, 11:02:47	Hot (Inferred)		Block blob	347

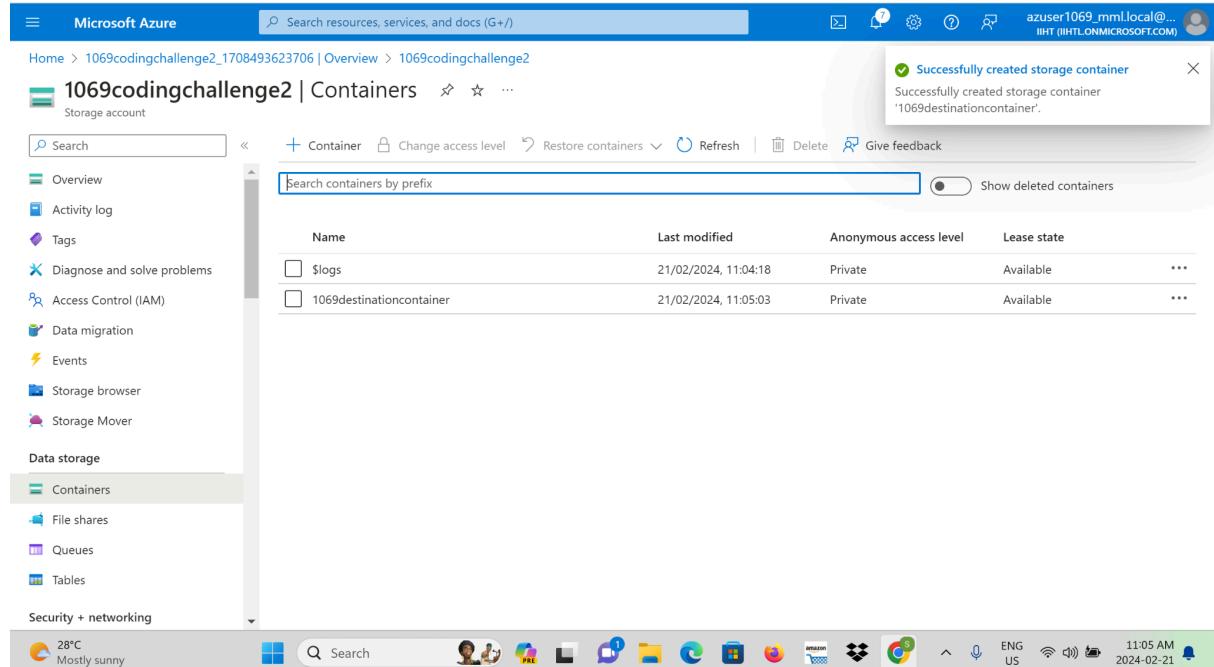
At the bottom right, there is a success message: "Successfully uploaded blob(s) Successfully uploaded 1 blob(s)".



→ I have created another storage account for the destination group and then I have created a container without upload a file in it.



The screenshot shows the Microsoft Azure Deployment Overview page for a deployment named "1069codingchallenge2_1708493623706". The status is "Your deployment is complete". Deployment details include a name of "1069codingchallenge2_170...", a subscription of "Azure subscription 1", and a resource group of "resourcegroup1069". The deployment started at 21/02/2024, 11:03:52. A "Go to resource" button is present. The page also includes links for "Give feedback" and "Tell us about your experience with deployment". A sidebar on the right provides links to "Cost Management", "Microsoft Defender for Cloud", "Free Microsoft tutorials", and "Work with an expert". The bottom of the screen shows a taskbar with various icons and system status.



The screenshot shows the Microsoft Azure Storage Containers page for the "1069codingchallenge2" storage account. It displays a list of containers, including "\$logs" and "1069destinationcontainer", both of which were successfully created. A success message box is visible stating "Successfully created storage container '1069destinationcontainer'". The page includes a search bar, filter options, and a table header for "Name", "Last modified", "Anonymous access level", and "Lease state". The bottom of the screen shows a taskbar with various icons and system status.

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

Home > 1069codingchallenge2_1708493623706 | Overview > 1069codingchallenge2 | Containers >

1069destinationcontainer

Container

Search Upload Change access level Refresh Delete Change tier Acquire lease Break lease View snapshots ...

Overview Diagnose and solve problems Access Control (IAM)

Authentication method: Access key (Switch to Microsoft Entra user account)
Location: 1069destinationcontainer

Search blobs by prefix (case-sensitive) Show deleted blobs

Add filter

Name	Modified	Access tier	Archive status	Blob type	Size
No results					

Shared access tokens
Access policy
Properties
Metadata



→ Then I created a data factory account with the same resource group and clicked on launch studio.

The screenshot shows the Microsoft Azure Data Factory Studio interface. At the top, there's a navigation bar with 'Microsoft Azure' and a search bar. Below the navigation bar, the title is 'Microsoft.DataFactory-20240221110534 | Overview'. On the left, a sidebar menu includes 'Overview', 'Inputs', 'Outputs', and 'Template'. The main content area displays a green checkmark indicating 'Your deployment is complete' with details: Deployment name: Microsoft.DataFactory-20240..., Start time: 21/02/2024, 11:06:11, Subscription: Azure subscription 1, Correlation ID: 23fa0dd7-4c35-4fcc-b0c2-2a..., Resource group: resourcegroup1069. Below this, there are sections for 'Deployment details' and 'Next steps', with a 'Go to resource' button. A feedback section asks for deployment experience. To the right, there are promotional cards for 'Cost management', 'Microsoft Defender for Cloud', 'Free Microsoft tutorials', and 'Work with an expert'. The bottom of the screen shows a taskbar with various icons and system status information.

Microsoft Azure | Microsoft.DataFactory-20240221110534 | Overview

Home > Microsoft.DataFactory-20240221110534 | Overview

Microsoft Data Factory Studio

Your deployment is complete

Deployment name : Microsoft.DataFactory-20240... Start time : 21/02/2024, 11:06:11

Subscription : Azure subscription 1 Correlation ID : 23fa0dd7-4c35-4fcc-b0c2-2a...

Resource group : resourcegroup1069

Deployment 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.

28°C Mostly sunny

Search

ENG US 11:06 AM 2024-02-21

Microsoft Azure | 1069datafactory | Overview

Home > Microsoft.DataFactory-20240221110534 | Overview >

1069datafactory Data factory (V2)

Search Delete

Overview Activity log Access control (IAM) Tags Diagnose and solve problems

Subscription ID
984f097c-963c-4eb6-a20d-839457ae9f08

Azure Data Factory Studio

Launch studio

Quick Starts Tutorials Template Gallery Training Modules

Monitoring

Alerts

https://adf.azure.com/en/home?factory=%2Fsubscriptions%2F984f097c-963...

28°C Mostly sunny

Search

ENG US 11:07 AM 2024-02-21

→ Then I have selected an option called ingestion to set the connection .

The screenshot shows the Microsoft Azure Data Factory portal. At the top, there's a banner about the public preview of Microsoft Fabric. Below it, the main dashboard for the '1069datafactory' resource is displayed. The dashboard features a central illustration of a factory with various data storage and processing components. Below the illustration are four main service cards: 'Ingest' (Copy data at scale once or on a schedule), 'Orchestrate' (Code-free data pipelines), 'Transform data' (Transform your data using data flows), and 'Configure SSIS' (Manage & run your SSIS packages in the cloud). On the left, there's a sidebar with icons for Home, New, and other resources. Below the dashboard, there's a section for 'Recent resources'. The bottom of the screen shows the Windows taskbar with various pinned icons and system status.

→ Then I have given all the properties and clicked on next .

The screenshot shows the 'Copy Data tool' wizard, step 1: Properties. The left sidebar lists steps 1 through 5. Step 1 is currently selected and highlighted. The main pane contains instructions for using the Copy Data Tool to perform one-time or scheduled data loads from over 90 data sources. It includes a 'Properties' section for selecting copy data task type and configuring task schedule. Two options are shown: 'Built-in copy task' (which provides a single pipeline to copy data from 90+ data sources) and 'Metadata-driven copy task' (which provides parameterized pipelines for reading metadata from external stores). Below these, there's a note about getting single pipelines for copying objects between data source and destination. Task cadence or task schedule is set to 'Run once now'. The bottom of the screen shows the Windows taskbar.

→ Then I have created a source data source and then I have selected azure blob storage as source type and created a connection by giving a source storage name and source container in it.

Copy Data tool

Source data store

Specify the source data store for the copy task. You can use a connection or a direct URL.

Source type: Azure Blob Storage

Connection: AzureBlobStorage

File or folder: dataset.csv

If the identity you use to access the data store only has permission to list files, specify the path to browse. Append a slash (/) at the end of the path if you want to browse the contents of a folder.

Options

- Binary copy
- Recursively
- Enable partitions discovery

Max concurrent connections: 10

Filter by last modified

Start time (UTC): 2024-02-21T00:00:00Z

End time (UTC): 2024-02-21T23:59:59Z

Browse

Select a file or folder.

Root folder > sourcecontainer

dataset.csv

Showing 1 item

OK **Cancel**

Copy Data tool

Source data store

Specify the source data store for the copy task. You can use a connection or a direct URL.

Source type: Azure Blob Storage

Connection: Select...

New connection

Azure Blob Storage [Learn more](#)

Name: AzureBlobStorage1

Description

Connect via integration runtime: AutoResolveIntegrationRuntime

Authentication type: Account key

Connection string **Azure Key Vault**

Account selection method

- From Azure subscription
- Enter manually

Azure subscription: Azure subscription 1 (984f097c-963c-4eb6-a20d-839457ae9f08)

Storage account name

Connection successful

Create **Cancel**

The screenshot shows the 'Copy Data tool' interface in Microsoft Azure Data Factory. The left sidebar lists steps: Properties, Source, Dataset, Configuration, Destination, Settings, and Review and finish. The 'Source' step is selected. The main panel is titled 'Source data store' and contains the following configuration:

- Source type: Azure Blob Storage
- Connection: AzureBlobStorage1
- Options:
 - Binary copy (checked)
- Compression type: None
- Recursively (checked)
- Delete files after completion (unchecked)
- Max concurrent connections: (empty input field)
- Filter by last modified: Start time (UTC) and End time (UTC) input fields

At the bottom are 'Previous' and 'Next >' buttons, and a 'Cancel' button.

→ Later I have created a destination data source and then I have selected azure blob storage as destination type and created a connection by giving a destination storage name and destination container in it.

→ In the destination container there is no file present because we are copying source to destination type .So our motive is to copy the data from source and review at destination group . For that we need to give an empty destination container.

The screenshot shows the 'Copy Data tool' interface in Microsoft Azure Data Factory. The left sidebar lists steps: Properties, Source, Destination, Dataset, Configuration, Settings, and Review and finish. The 'Destination' step is selected. The main panel is titled 'Destination data store' and contains the following configuration:

- Destination type: Azure Blob Storage
- Connection: Select...

A large right-hand panel titled 'New connection' is open, showing the configuration for a new Azure Blob Storage connection:

- Name: AzureBlobStorage2
- Description: (empty input field)
- Connect via integration runtime: AutoResolveIntegrationRuntime
- Authentication type: AutoResolveIntegrationRuntime
- Account key: (dropdown menu)
- Connection string: (radio button selected)
- Azure Key Vault: (radio button)
- Account selection method:
 - From Azure subscription (radio button selected)
 - Enter manually (radio button)
- Azure subscription: Azure subscription 1 (984f097c-963c-4eb6-a20d-839457ae9f08)
- Storage account name: (empty input field)

At the bottom are 'Create' and 'Cancel' buttons, and status messages: 'Connection successful' and 'Test connection'.

Microsoft Azure | Data Factory | 1069datafactory | Search factory and documentation | ? | Help | azuser1069_mml.local@iht.onmicrosoft.com

Copy Data tool

Properties
Source
Destination
Dataset
Configuration
Settings
Review and finish

Destination data store

Specify the destination data store for the copy task. You can use an existing data store connection or specify a new data store.

Destination type: Azure Blob Storage
Connection: AzureBlobStorage

Folder path: 1069destinationcontainer

File name:

Compression type: None

Copy behavior: Select...

Max concurrent connections:

Browse

Select a file or folder.

Root folder: 1069destinationcontainer

Showing 1 item

OK Cancel

29°C Haze Search ENG US 11:12 AM 2024-02-21

Microsoft Azure | Data Factory | 1069datafactory | Search factory and documentation | ? | Help | azuser1069_mml.local@iht.onmicrosoft.com

Copy Data tool

Properties
Source
Destination
Dataset
Configuration
Settings
Review and finish

Destination data store

Specify the destination data store for the copy task. You can use an existing data store connection or specify a new data store.

Destination type: Azure Blob Storage
Connection: AzureBlobStorage2

Folder path: 1069destinationcontainer/

File name:

Compression type: None

Copy behavior: Select...

Max concurrent connections:

Browse

OK Cancel

29°C Haze Search ENG US 11:12 AM 2024-02-21

→ Then we need to give the task name as “CopyActivity” for the above process and click on the next .

The screenshot shows the 'Copy Data tool' settings page. On the left, a vertical navigation bar lists steps: Properties, Source, Destination, Settings, Review and finish, and a circled 5 indicating the current step. The main panel is titled 'Settings' with the sub-section 'Task name'. A text input field contains 'CopyActivity'. Below it are sections for 'Task description', 'Data consistency verification' (unchecked), 'Enable logging' (unchecked), and 'Enable staging' (unchecked). At the bottom are 'Previous' and 'Next >' buttons, and a 'Cancel' button on the right.

→ After giving the task name we need to review all the settings and click on finish .

The screenshot shows the 'Copy Data tool' review and finish page. The left navigation bar shows steps 1 through 5, with 'Review and finish' highlighted. The main panel has a 'Summary' section showing a flow from 'Azure Blob Storage' to 'Azure Blob Storage'. Below it is a 'Properties' section with 'Task name' set to 'CopyActivity'. Under 'Source', details are listed: 'Connection name' (AzureBlobStorage1), 'Dataset name' (SourceDataset_lyv), 'File name' (dataset.csv), and 'Container' (sourcecontainer). There are 'Edit' buttons for each source detail. At the bottom are 'Previous' and 'Next >' buttons, and a 'Cancel' button on the right.

Microsoft Azure | Data Factory > 1069datafactory

Search factory and documentation

azuser1069_mml.local@iht.onmicrosoft.com

Copy Data tool

Properties → Source → Destination → Settings → Review and finish → Review → Deployment

Azure Blob Storage → Azure Blob Storage

Deployment complete

Deployment step Status

- Validating copy runtime environment Succeeded
- > Creating datasets Succeeded
- > Creating pipelines Succeeded
- > Running pipelines Succeeded

Datasets and pipelines have been created. You can now monitor and edit the copy pipelines or click finish to close Copy Data Tool.

Finish Edit pipeline Monitor

29°C Haze Search ENG US 11:13 AM 2024-02-21

The screenshot shows the Microsoft Azure Data Factory interface. A 'Copy Data tool' is open, displaying a flow from 'Azure Blob Storage' to 'Azure Blob Storage'. The 'Deployment complete' section shows a summary of deployment steps: 'Validating copy runtime environment' (Succeeded), 'Creating datasets' (Succeeded), 'Creating pipelines' (Succeeded), and 'Running pipelines' (Succeeded). Below this, a message states: 'Datasets and pipelines have been created. You can now monitor and edit the copy pipelines or click finish to close Copy Data Tool.' At the bottom, there are buttons for 'Finish', 'Edit pipeline', and 'Monitor'. The status bar at the bottom right shows the date and time: '11:13 AM 2024-02-21'.

→ Then we can see that datasets and pipelines are successfully created and running successfully.
→ Thereby we can see that in the destination container the file has been successfully copied from the source account .

Microsoft Azure

Search resources, services, and docs (G+)

Home > 1069codingchallenge2 | Containers >

1069destinationcontainer Container

Search Overview Diagnose and solve problems Access Control (IAM) Settings Shared access tokens Access policy Properties Metadata

Upload Change access level Refresh Delete Change tier Acquire lease Break lease View snapshots ...

Authentication method: Access key (Switch to Microsoft Entra user account)
Location: 1069destinationcontainer

Search blobs by prefix (case-sensitive) Show deleted blobs

Name	Modified	Access tier	Archive status	Blob type	Size
dataset.csv	21/02/2024, 11:13:26	Hot (Inferred)		Block blob	347

https://portal.azure.com/#

29°C Haze Search ENG US 11:13 AM 2024-02-21

The screenshot shows the Microsoft Azure Storage Explorer interface. It displays a container named '1069destinationcontainer'. Under the 'Overview' tab, a single blob named 'dataset.csv' is listed. The blob was modified on 21/02/2024, 11:13:26, is in the 'Hot (Inferred)' access tier, and is a 'Block blob' with a size of 347. The status bar at the bottom right shows the date and time: '11:13 AM 2024-02-21'.

2) Explain Overview of 3 level namespace and creating Unity Catalog objects?

A)

In Unity Catalog, the hierarchy of primary data objects flows from metastore to table or volume:

Metastore: The top-level container for metadata. Each metastore exposes a three-level namespace (catalog.schema.table) that organizes your data.

Catalog: The first layer of the object hierarchy, used to organize your data assets.

Schema: Also known as databases, schemas are the second layer of the object hierarchy and contain tables and views.

Tables, views, and volumes: At the lowest level in the data object hierarchy are tables, views, and volumes. Volumes provide governance for non-tabular data.

Models: Although they are not, strictly speaking, data assets, registered models can also be managed in Unity Catalog and reside at the lowest level in the object hierarchy.