

Mallu Sravana Sandhya
Assessment-27
05/01/24

Data Lakehouse:

It is a combination or integration of Delta lake which is an open source storage that gather all the information stored in the cloud storage ,process it and Data Warehouse which contains huge amount of databases.

The Lakehouse platform is built on top of Delta Lake, which is an open-source storage layer that adds reliability, scalability, and performance to data lakes. It provides ACID transactions, schema enforcement, and indexing to data lakes, making them more suitable for data warehousing workloads.

Key features of Lakehouse:

- It provides transactional updates on data updation, deletion and transfer of data which indicates the ACID properties
- Data lakehouse can contain data fthat belongs to different version and maintains history of data which helps in analysis of version control
- Scalabitlity and performance: it allows any oraganization to process, maintain and analyze huge amount of data.it uses various distributed computing techniques to process and analysis
- Schema enforcement: it provides strong governance and auditing mechanisms by maintaing data quality and consistency
- It works on wide range of data sources and formats irrespective of their structured forms
- Lake house provide a unified way to manage metadata across multiple data sources, making it easier to discover, access, and query data.

Architectural component of Lakehouse:

Delta Lake —

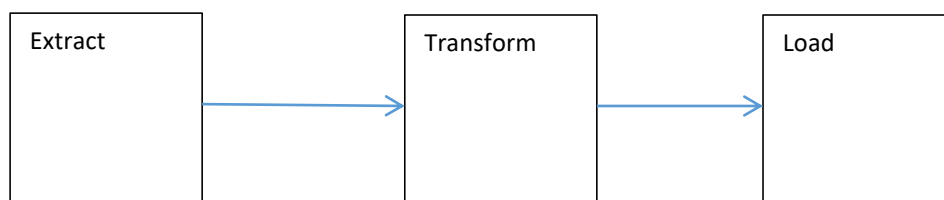
Delta Lake provides a powerful set of features for managing data in a lakehouse architecture. It is a optimized storage layer that allows users to store and manage data in various file formats.Delta Lake also provides a transaction log, which enables ACID transactions and allows for data versioning and rollbacks.

Unity Catalog —

Unity Catalog is a Unified Governance for All Data & AI Assets including files, tables, ML models and dashboards in our lakehouse on any cloud. It is designed to help organizations manage and organize their data in a scalable and efficient manner. Unity catalog is known for -

- Centralized governance for data and AI
- Provide enhanced query performance and boosts productivity
- Transparency of creating automated lineage for all workloads
- Secure data sharing across organizations

Databricks Delta Live Tables provide one of the key solution to build and manage, reliable and robust data engineering pipelines that can load the Streaming and batch data and deliver high-quality data on the Lakehouse Platform.



In data ingestion there are three layers they are

Bronze layer-it provides the raw data which is collected from data sources .it maintains all the history of organizations and provide data with no filters

Silver layer- it provides filter data it maintains data in semistructured form.It handles the missing data and also does the data type conversation if required. It converts the nested objects into the flat structures .It is responsible for cleaning and filtering bronze data

Gold layer-It contains the well structured, aggregated and modeled data which helps for the business queries

Complexity of Data Delivery

The following problems can be faced while maintaining data quality and consistency with large scale solution data pipelines:

- Difficult to maintain table dependencies and to switch between streaming and batch load.
- Error handling and recovery is time consuming
- Pipeline failure may impact the downstream system and the team relying on it.

Data engineers have to focus on tools instead of doing the development because operational complexity dominates.

To overcome this data delivery complexity we use DLT Delta Live Tables which helps to solve the complexity issues

Delta Live Tables or DLT is one of the best way to do ETL process in Lakehouse.It is solely responsible for performing data transformation and managing the task orchestration, cluster management, monitoring, data quality, and error handling.

Key benefits of working with Delta Live tables are:

- Accelerates the ETL Process
- Automatically manage your infrastructure
- Ensure high data quality
- Unify batch and streaming

Creating Azure Data Factory

This screenshot shows the 'Create a resource' page in the Microsoft Azure portal. The browser address bar displays 'portal.azure.com/#create/hub'. The page features a search bar at the top with the text 'Search resources, services, and docs (G+)'. Below the search bar, there are sections for 'Get Started', 'Recently created', and 'Categories'. The 'Categories' section lists various services including AI + Machine Learning, Analytics, Blockchain, Compute, Containers, Databases, Developer Tools, DevOps, Identity, Integration, and Internet of Things. The 'Popular Azure services' section highlights Data Factory, Azure Synapse Analytics, Azure Machine Learning, Event Hubs, Azure Databricks, and Data Lake Storage Gen1. The 'Popular Marketplace products' section lists CloudOps - Monitoring-as-a-Service, MongoDB Atlas (pay-as-you-go), Azure Cost Management plan, Azure SQL Edge Developer, Data Science Virtual Machine - Ubuntu 20.04, and azure_sen_managed_siem_prem. The bottom of the page shows a Windows taskbar with a search bar and various application icons.

This screenshot shows the 'Create Data Factory' page in the Microsoft Azure portal. The browser address bar displays 'portal.azure.com/#create/Microsoft.DataFactory'. The page features a search bar at the top with the text 'Search resources, services, and docs (G+)'. Below the search bar, there are tabs for 'Basics', 'Git configuration', 'Networking', 'Advanced', 'Tags', and 'Review + create'. The 'Basics' tab is selected, showing a section for 'Project details' with a description: 'One-click to create data factory with sample pipeline and datasets. Try it'. Below this, there are fields for 'Subscription' (Azure subscription 1), 'Resource group' (rg-azuser921_mml.local-yVpvZ), 'Instance details' (Name: SravanaSandhya, Region: East US, Version: V2), and a 'Review + create' button. The bottom of the page shows a Windows taskbar with a search bar and various application icons.

Microsoft Azure portal interface showing the "Create Data Factory" wizard. The wizard is at the "Basics" step, where the user is prompted to agree to the legal terms and privacy statement(s) associated with the Marketplace offering(s). The "Basics" section displays the following configuration:

Section	Configuration
Subscription	Azure subscription 1
Resource group	rg-azuser921_mml.local-yVpvZ
Name	SravanaSandhya
Region	East US
Version	V2
Networking	Connect via: Public endpoint

Navigation buttons: Previous, Next, Create. A "Give feedback" link is also present.

Microsoft Azure portal interface showing the "Microsoft.DataFactory-20240106114258 | Overview" page. The page displays the deployment status as "Your deployment is complete".

Deployment details:

- Deployment name: Microsoft.DataFactory-20240106114258
- Start time: 06/01/2024, 11:44:14
- Subscription: Azure subscription 1
- Correlation ID: 124cf5f1-630c-45d7-bddd-2b0937cac...
- Resource group: rg-azuser921_mml.local-yVpvZ

Next steps:

- Go to resource

Give feedback:

- Tell us about your experience with deployment

Deployment succeeded notification:

Deployment 'Microsoft.DataFactory-20240106114258' to resource group 'rg-azuser921_mml.local-yVpvZ' was successful.

Cost management:

- Get notified to stay within your budget and prevent unexpected charges on your bill.
- Set up cost alerts >

Microsoft Defender for Cloud:

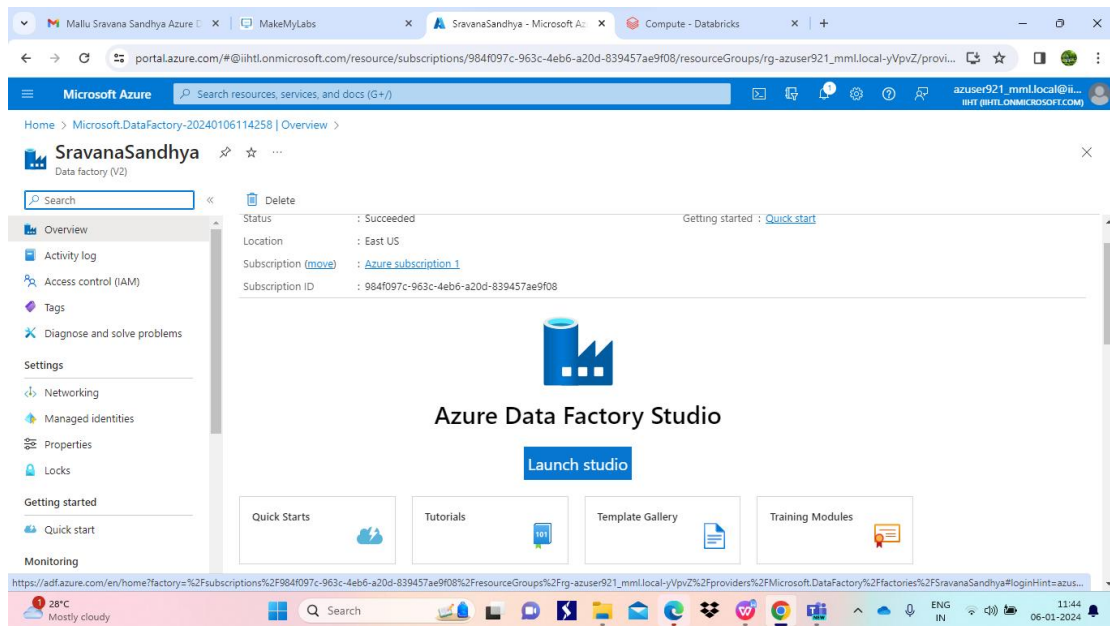
- Secure your apps and infrastructure
- Go to Microsoft Defender for Cloud >

Free Microsoft tutorials:

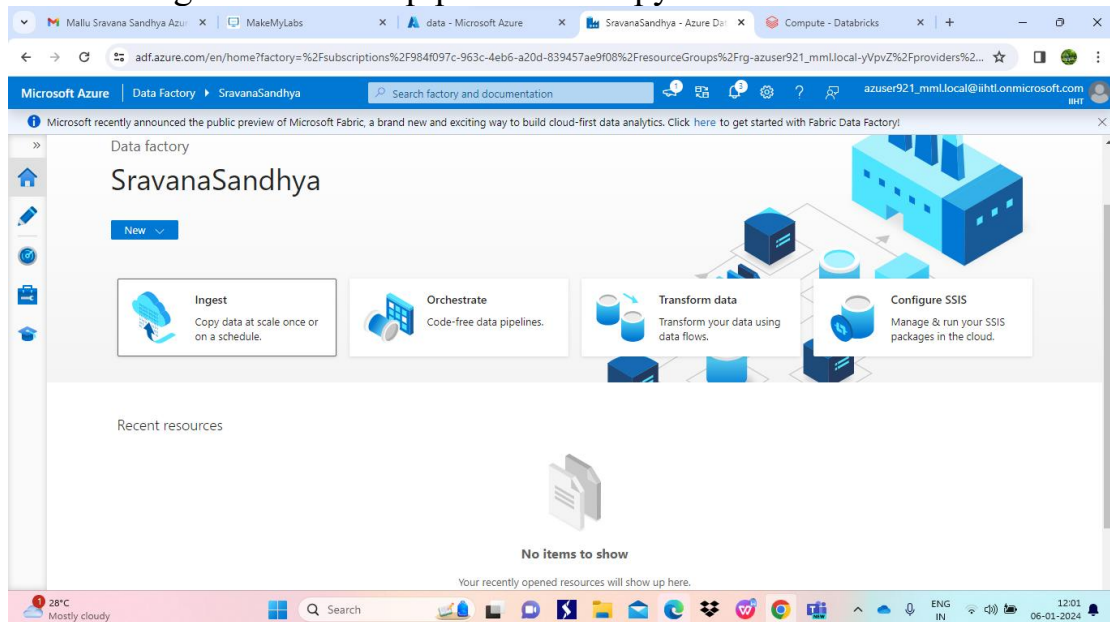
- Start learning today >

Work with an expert:

- Azure experts are service provider partners



Click on ingest to create a pipeline and copy the file



Click on next

Microsoft Azure | Data Factory | SravanaSandhya

Search factory and documentation

Copy Data tool

1 Properties

2 Source

3 Destination

4 Settings

5 Review and finish

Use Copy Data Tool to perform a one-time or scheduled data load from 90+ data sources. Follow the wizard experience to specify your data loading settings, and let the Copy Data Tool generate the artifacts for you, including pipelines, datasets, and linked services. [Learn more](#)

Properties

Select copy data task type and configure task schedule

Task type

Built-in copy task

You will get single pipeline to copy data from 90+ data source easily.

Metadata-driven copy task

You will get parameterized pipelines which can read metadata from an external store to load data at a large scale.

You will get single pipeline to quickly copy objects from data source store to destination in a very intuitive manner.

Task cadence or task schedule *

☒ Run once now ☐ Schedule ☐ Tumbling window

[< Previous](#) [Next >](#) [Cancel](#)

28°C Mostly cloudy

Search

ENG IN 11:54 06-01-2024

Fill the details in source from where we have to copy file

Copy Data tool

Source data store

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

Source type: All

Connection: Select...

New connection

Azure Blob Storage [Learn more](#)

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-839457aef08)

Storage account name: mallu

Additional connection properties: + New

Test connection: ☒ To linked service ☐ To file path

Annotations

Create Back Testing connection... Cancel

Copy Data tool

Source data store

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

Source type: All

Connection: AzureBlobStorage1 Edit + New connection

Options:

- ☐ Binary copy
- ☒ Recursively
- ☐ Enable partitions discovery

Max concurrent connections:

Filter by last modified:

Start time (UTC): End time (UTC):

Previous Next Cancel

Select the file which we have to copy

Microsoft Azure | Data Factory | SravanaSandhya

Copy Data tool

Source data store

Specify the source data store for the copy task. You can use an existing data store or create a new one.

Source type: Azure Blob Storage

Connection: AzureBlobStorage1

File or folder:

If the identity you use to access the data store only has permission to subid account, specify the path to browse.

Options

☒ Binary copy

Compression type: None

☒ Recursively

☐ Delete files after completion

Please input or choose a folder or file

Previous Next

Browse

Select a file or folder.

Root folder: firstcontainer

ADB Topics.docx

Showing 1 item

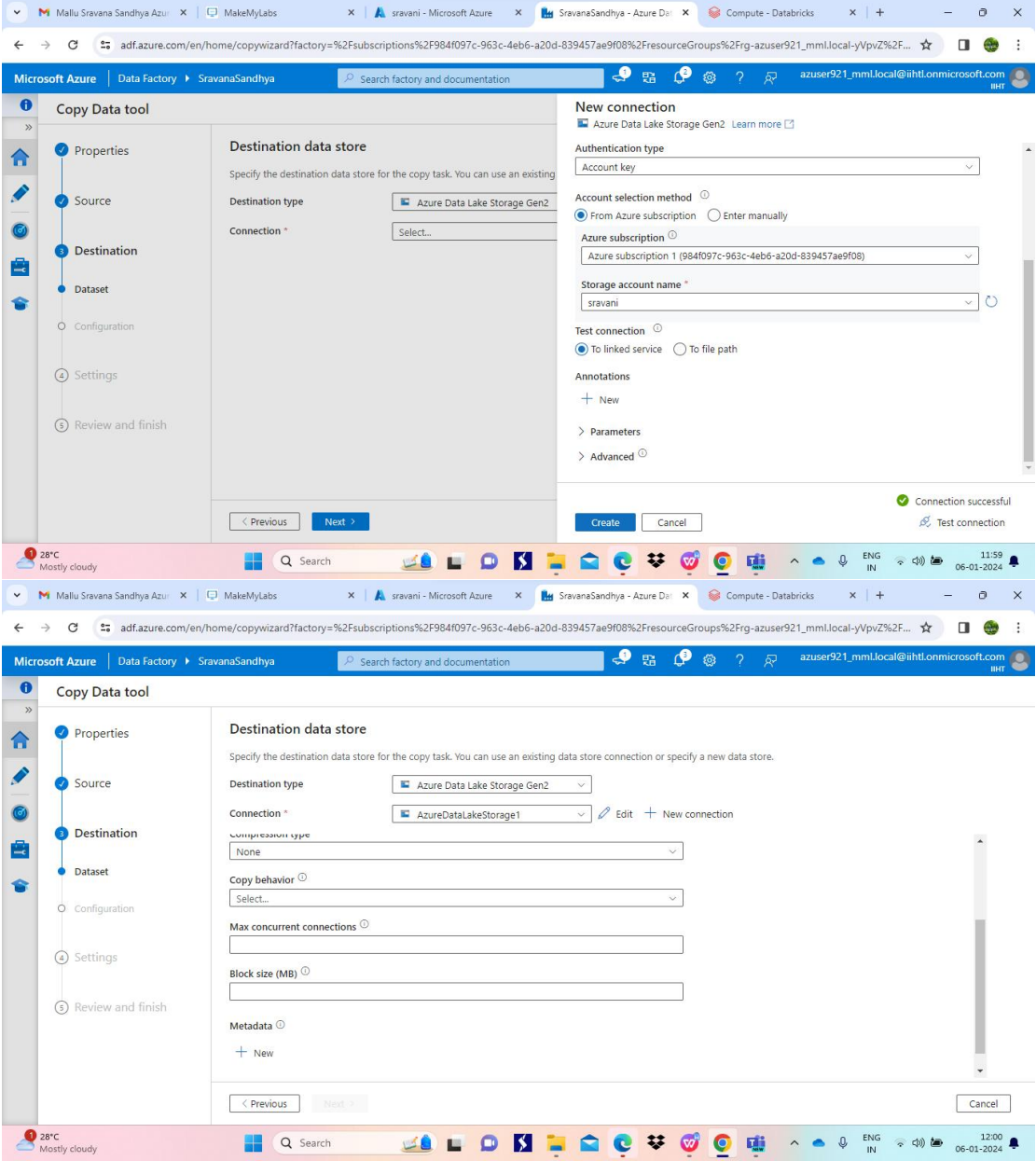
OK Cancel

28°C Mostly cloudy

Search

11:58 06-01-2024

Give the destination location



Click on next

Microsoft Azure | Data Factory | SravanaSandhya

Copy Data tool

Settings

Enter name and description for the copy data task, more options for data movement

Task name *

Task description

Data consistency verification ☐

Enable logging ☐

Enable staging ☐

> Advanced

< Previous Next > Cancel

Review the souce and destination and click next

Microsoft Azure | Data Factory | SravanaSandhya

Copy Data tool

Summary

You are running pipeline to copy data from Azure Blob Storage to Azure Data Lake Storage Gen2.

Properties

Task name CopyPipeline_J01 Edit

Task description Edit

Source

Connection name AzureBlobStorage1 Edit

Dataset name SourceDataset_J01

File name ADB_Topics.docx

Container firstcontainer

Destination

Connection name AzureDataLakeStorage1 Edit

< Previous Next > Cancel

after completion of deployment click on finish

The screenshot shows the Microsoft Azure Copy Data tool interface. The left sidebar contains a navigation menu with the following items: Properties, Source, Destination, Settings, Review and finish (highlighted), Review, and Deployment. The main content area displays a diagram showing data flow from 'Azure Blob Storage' to 'Azure Data Lake Storage Gen2'. Below the diagram, the text 'Deployment complete' is shown. A table lists the deployment steps and their status:

Deployment step	Status
Validating copy runtime environment	Succeeded
> Creating datasets	Succeeded
> Creating pipelines	Succeeded
> Running pipelines	Succeeded

Below the table, 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 of the main content area, there are three buttons: 'Finish', 'Edit pipeline', and 'Monitor'.

After successfully copying of file we can see the file in destination folder

The screenshot shows the Microsoft Azure portal interface. The top navigation bar includes the Microsoft Azure logo, a search bar, and the user's profile. The main content area displays the 'data' container. The left sidebar contains a navigation menu with the following items: Overview (highlighted), Diagnose and solve problems, Access Control (IAM), Settings, Shared access tokens, Manage ACL, Access policy, Properties, and Metadata. The main content area shows the 'Overview' tab for the 'data' container. It includes a search bar, a list of actions (Upload, Add Directory, Refresh, Rename, Delete, Change tier, Acquire lease, Break lease, Give feedback), and a table of blobs. The table has the following columns: Name, Modified, Access tier, Archive status, Blob type, Size, and Lease state. The table contains one row with the file 'ADB Topics.docx'.

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
ADB Topics.docx	06/01/2024, 12:01:15	Hot (Inferred)		Block blob	4.92 MiB	Available

Creating source location which is a storage account

Microsoft Azure

Search resources, services, and docs (G+)

Home > Storage accounts >

Create a storage account

BasicsAdvancedNetworkingData protectionEncryptionTagsReview

Project details

Select the subscription in which to create the new storage account. Choose a new or existing resource group to organize and manage your storage account together with other resources.

Subscription *

Azure subscription 1

Resource group *

rg-azuser921_mmllocal-yVpvZ

Create new

Instance details

Review

< Previous

Next > Advanced

Give feedback

28°C Mostly cloudy

Search

Mallu Sravana Sandhya AzuMakeMyLabsCreate a storage account - SravanaSandhya - Azure DCompute - Databricks

portal.azure.com/#create/Microsoft.StorageAccount-ARM

Microsoft Azure

Search resources, services, and docs (G+)

Home > Storage accounts >

Create a storage account

BasicsAdvancedNetworkingData protectionEncryptionTagsReview

Basics

Subscription

Azure subscription 1

Resource Group

rg-azuser921_mmllocal-yVpvZ

Location

centralindia

Storage account name

mallu

Deployment model

Resource manager

Performance

Standard

Replication

Read-access geo-redundant storage (RA-GRS)

Advanced

Enable hierarchical namespace

Disabled

Enable network file system v2

Disabled

Create

< Previous

Next >

Download a template for automation

Give feedback

28°C Mostly cloudy

Search

Mallu Sravana Sandhya AzuMakeMyLabsCreate a storage account - SravanaSandhya - Azure DCompute - Databricks

portal.azure.com/#create/Microsoft.StorageAccount-ARM

Microsoft Azure

Search resources, services, and docs (G+)

Home > Storage accounts >

Create a storage account

BasicsAdvancedNetworkingData protectionEncryptionTagsReview

Basics

Subscription

Azure subscription 1

Resource Group

rg-azuser921_mmllocal-yVpvZ

Location

centralindia

Storage account name

mallu

Deployment model

Resource manager

Performance

Standard

Replication

Read-access geo-redundant storage (RA-GRS)

Advanced

Enable hierarchical namespace

Disabled

Enable network file system v2

Disabled

Create

< Previous

Next >

Download a template for automation

Give feedback

Microsoft Azure portal showing the deployment details for the resource group **mallu_1704521795038**. The deployment is in progress.

Deployment details:

- Deployment name: mallu_1704521795038
- Subscription: Azure subscription 1
- Resource group: rg-azuser921_mml.local-yVpvZ
- Start time: 06/01/2024, 11:46:40
- Correlation ID: ded03a0f-019e-4eea-94bc-dec0035b1609

Deployment details table:

Resource	Type	Status	Operation details
No results.			

Deployment in progress...
Deployment to resource group 'rg-azuser921_mml.local-yVpvZ' is in progress.

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 >

creating a folder in the source location

Microsoft Azure portal showing the **Containers** section for the resource group **mallu**. A new container is being created.

Containers section:

- Search containers by prefix
- Table with columns: Name, Last modified
- Container: \$logs (Last modified: 06/01/2024, 11:47:13)

New container dialog:

- Name: firstcontainer
- Anonymous access level: Private (no anonymous access)
- Advanced: (Expanded)
- Create button

Message: The access level is set to private because anonymous access is disabled on this storage account.

Adding file in the source location

The screenshot shows the Microsoft Azure portal interface. The main view is for a container named 'firstcontainer'. An 'Upload blob' dialog box is open on the right side of the screen. The dialog box has a title bar that says 'Upload blob' and a close button. Inside the dialog, there is a message: '*** Uploading on blob(s)... Attempting to upload 1 blob(s)'. Below this, there is a large dashed box with a cloud icon and the text 'Drag and drop files here or Browse for files'. There is a checkbox labeled 'Overwrite if files already exist' which is checked. Below the checkbox, there is an 'Advanced' section with an 'Upload' button and a 'Give feedback' link. At the bottom of the dialog, there is a section for 'Current uploads' showing 'ADB Topics.docx' with a progress bar indicating '0 / 4.92 MiB'. The background shows the 'firstcontainer' overview page with a search bar, 'Authentication method', 'Location', and a table with no results.

Similarly create a destination storage account and the file is copied successfully

The screenshot shows the Microsoft Azure portal interface for a container named 'data'. The 'Overview' tab is selected. The 'Authentication method' is 'Access key' and the 'Location' is 'data'. There is a search bar and a 'Show deleted objects' toggle. Below this, there is a table listing the contents of the container. The table has columns for Name, Modified, Access tier, Archive status, Blob type, Size, and Lease state. One file is listed: 'ADB Topics.docx', modified on '06/01/2024, 12:01:15', with an 'Access tier' of 'Hot (Inferred)', 'Archive status' of 'None', 'Blob type' of 'Block blob', 'Size' of '4.92 MiB', and 'Lease state' of 'Available'. The background shows the 'data' container overview page with various action buttons like 'Upload', 'Add Directory', 'Refresh', 'Rename', 'Delete', 'Change tier', 'Acquire lease', 'Break lease', and 'Give feedback'.