

Azure Data Engineer Associate Certification Guide, Second Edition

Preface:

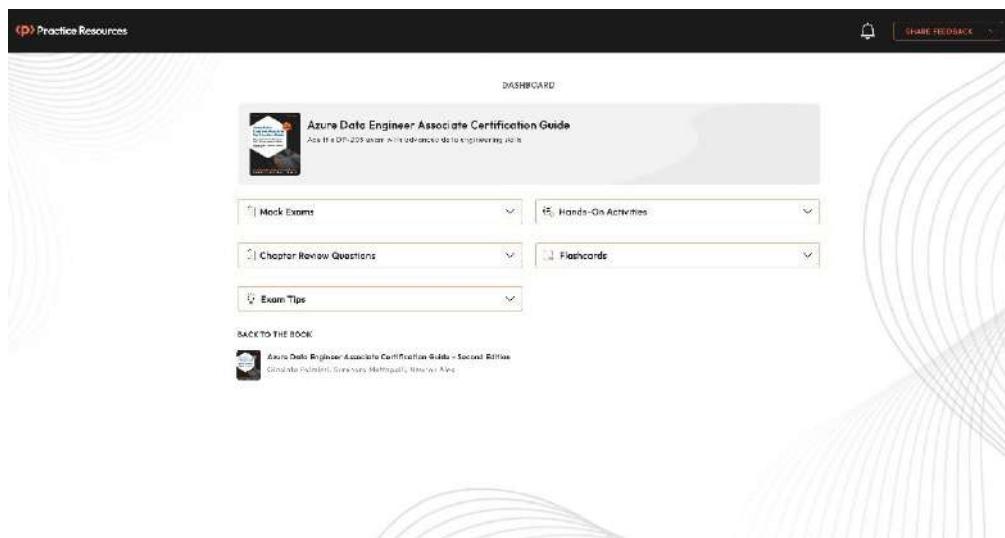


Figure 0.1 – Exam DP-203: Azure data engineer Associate online exam-prep platform

PacktPublishing / **DP-203-Azure-Data-Engineer-Associate-Certification-Guide-Second-Edition** Public

Code Issues Pull requests Actions Security Insights

main Go to file Code About

	SurendraMettapalli	76 Commits
	Chapter01	Create Readme.md 3 weeks ago
	Chapter02	Create Readme.md 3 weeks ago
	Chapter03	Create Readme.md 3 weeks ago
	Chapter04	Update Readme.md last week
	Chapter05	Update Readme.md last week
	Chapter06	Update ASA-Transformations.sql last week
	Chapter07	Create Readme.md 2 weeks ago
	Chapter08	Update RowLevelSecurity-Syna... 4 days ago
	Chapter09	Update readme.md last week
	Chapter10	Update Readme.md 5 days ago
	LICENSE	Initial commit 9 months ago
	README.md	Add files via upload 3 weeks ago

No description, website provided.

Readme MIT license Activity Custom properties 1 star 2 watching 0 forks Report repository

Releases No releases published

Packages No packages published

Contributors 3

SurendraMettapalli

Figure 0.2 – GitHub code files for DP-203 certification

Chapter 1: Introducing Azure Basics

Before You Proceed

To learn how to access these resources, head over to *Chapter 11, Accessing the Online Resources*, at the end of the book.

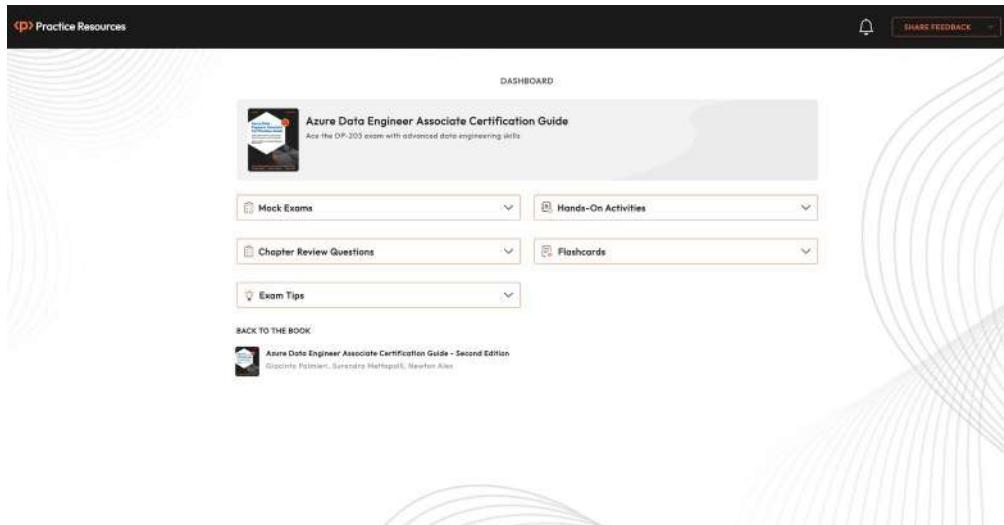


Figure 1.1: Dashboard interface of the online practice resources

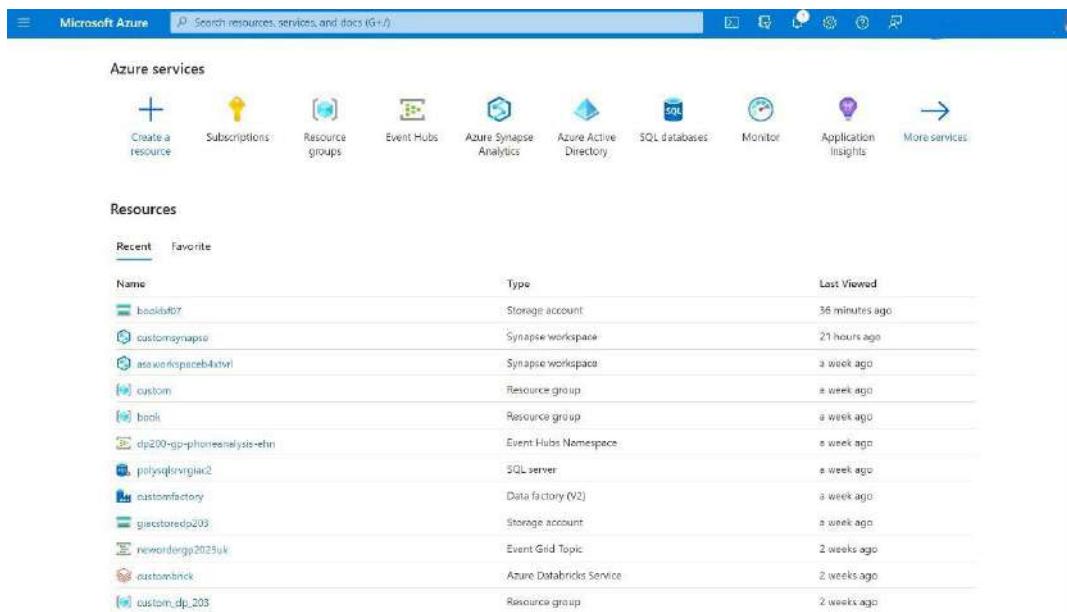


Figure 1.2 – The Azure portal home page

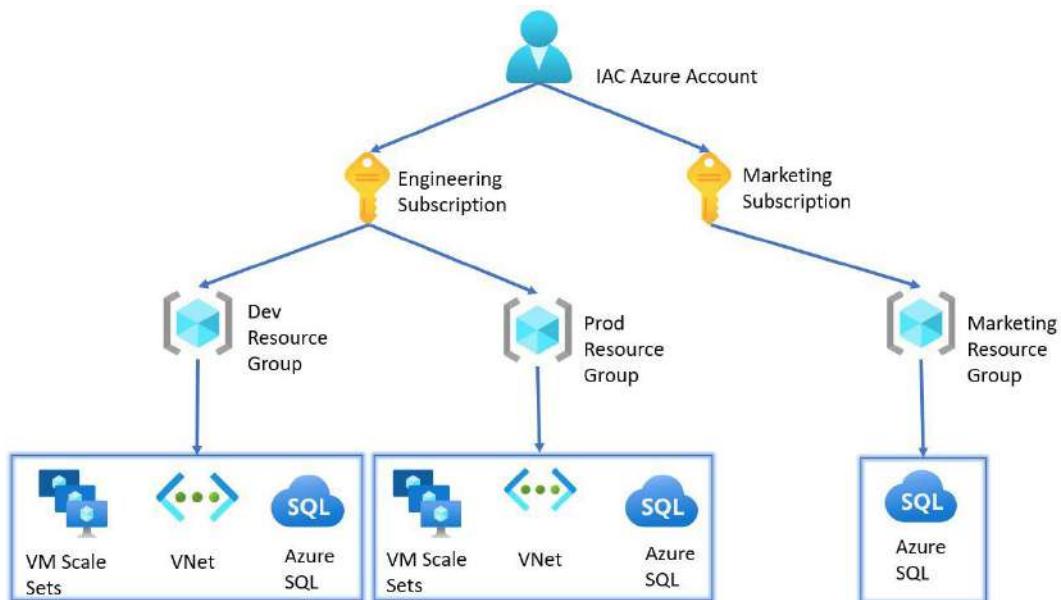


Figure 1.3 – Relationship between accounts, subscriptions, resource groups, and resources

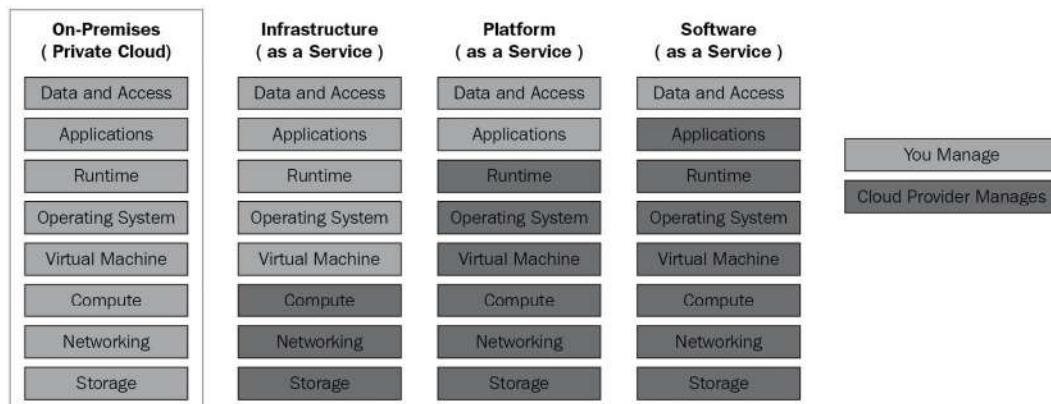


Figure 1.4 – Breakdown of Azure services

Create a virtual machine

Basics Disks Networking Management Advanced Tags Review + create

Create a virtual machine that runs Linux or Windows. Select an image from Azure marketplace or use your own customized image. Complete the Basics tab then Review + create to provision a virtual machine with default parameters or review each tab for full customization. [Learn more](#)

Project details

Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources.

Subscription *	<input type="text" value="Free Trial"/>
Resource group *	<input type="text" value="(New) DP203-Sandbox"/> Create new

Instance details

Virtual machine name *	<input type="text" value="samplevm"/>
Region *	<input type="text" value="(US) East US"/>
Availability options	<input type="text" value="No infrastructure redundancy required"/>
Image *	<input checked="" type="checkbox"/> Ubuntu Server 18.04 LTS - Gen1 See all images
Azure Spot instance	<input type="checkbox"/>
Size *	<input type="text" value="Standard_D2s_v3 - 2 vcpus, 8 GiB memory (₹5,048.93/month)"/> See all sizes

Administrator account

Authentication type [\(i\)](#)

SSH public key
 Password

Azure now automatically generates an SSH key pair for you and allows you to store it for future use. It is a fast, simple, and secure way to connect to your virtual machine.

[Review + create](#)

< Previous

Next : Disks >

Figure 1.5 – Creating VMs using the Azure portal

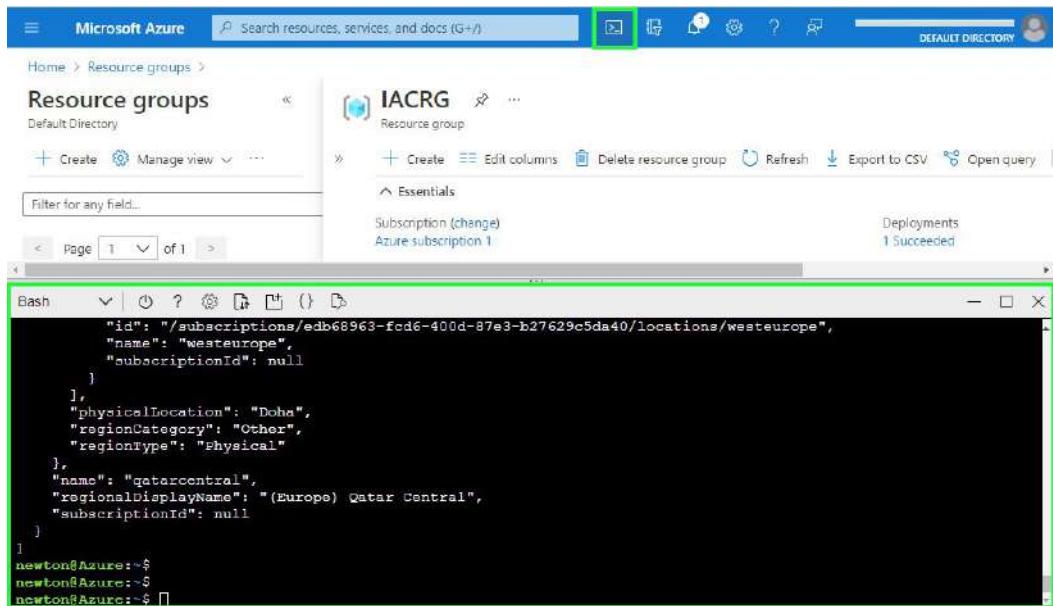


Figure 1.6 – Using the CLI directly from the Azure portal

Create a storage account

Basics Advanced Networking Data protection Tags Review + create

Azure Storage is a Microsoft-managed service providing cloud storage that is highly available, secure, durable, scalable, and redundant. Azure Storage includes Azure Blobs (objects), Azure Data Lake Storage Gen2, Azure Files, Azure Queues, and Azure Tables. The cost of your storage account depends on the usage and the options you choose below. [Learn more about Azure storage accounts](#)

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 *	Free Trial
Resource group *	(New) DP203-Sandbox
	Create new

Instance details

If you need to create a legacy storage account type, please click [here](#).

Storage account name ⓘ *	dp203blobstore
Region ⓘ *	(US) East US
Performance ⓘ *	<input checked="" type="radio"/> Standard: Recommended for most scenarios (general-purpose v2 account) <input type="radio"/> Premium: Recommended for scenarios that require low latency.
Redundancy ⓘ *	Geo-redundant storage (GRS)
	<input checked="" type="checkbox"/> Make read access to data available in the event of regional unavailability.

[Review + create](#)

< Previous

Next : Advanced >

Figure 1.7 – Creating a storage account using the Azure portal

Home > Storage accounts >

Create a storage account

Basics ● Advanced ● Networking Data protection Tags

Review + create

Data Lake Storage Gen2

The Data Lake Storage Gen2 hierarchical namespace accelerates big data analytics workloads and enables file-level access control lists (ACLs). [Learn more](#)

Enable hierarchical namespace



Figure 1.8 – Enable hierarchical namespace in Azure Storage for an ADLS Gen2 account

The screenshot shows the Azure Storage account overview for 'dp203book'. The left sidebar has 'Access keys' selected under 'Security + networking'. The main pane displays two access keys: 'key1' and 'key2'. Each key includes a 'Rotate key' button, a timestamp ('Last rotated: 04/03/2024 (0 days ago)'), and a 'Show' button to reveal the key value. Below each key is a 'Connection string' input field and another 'Show' button.

Storage account name: dp203book

key1 Rotate key
Last rotated: 04/03/2024 (0 days ago)
Key
[REDACTED] Show

key2 Rotate key
Last rotated: 04/03/2024 (0 days ago)
Key
[REDACTED] Show

Figure 1.9 – Locating access keys for a storage account

The screenshot shows the Practice Resources dashboard. At the top, there's a header with the logo and navigation links. Below the header, the breadcrumb navigation shows 'DASHBOARD > CHAPTER 1'. The main content area has a title 'Introducing Azure Basics' and a section titled 'Summary'. The summary text discusses the completion of the first chapter, mentioning Azure's overview, its relationship with accounts, subscriptions, resource groups, and resources, and how to create new VMs, storage instances, VNets, and so on. It also notes the availability of major compute services like Azure portal and CLI. The next chapter will explore partition strategies and implementation. To the right, a sidebar titled 'Chapter Review Questions' is visible, showing the 'The Azure Data Engineer Associate Certification Guide - Second Edition' by Giacinto Palmieri, Surendra Mettapalli, and Newton Alex. It includes a 'Select Quiz' section with 'Quiz 1' and a 'START' button.

DASHBOARD > CHAPTER 1

Introducing Azure Basics

Summary

With that, you have completed the first chapter. For those with some experience of Azure, this overview has hopefully offered a useful recap. For those of you who are completely new to Azure, it might have felt a bit overwhelming. Don't worry; as you complete the next few chapters, your confidence will grow.

In this chapter, you learned how to navigate the Azure portal and observed the relationship between Azure accounts, subscriptions, resource groups, and resources. You also learned how to create new VMs, storage instances, VNets, and so on using both the Azure portal and the CLI. You are also aware of the major compute services that are available in Azure. With this foundational knowledge in place, you can move on to more interesting and certification-oriented topics.

In the next chapter, you will explore the concept of a partition strategy and how to implement it.

Chapter Review Questions

The Azure Data Engineer Associate Certification Guide
- Second Edition by Giacinto Palmieri, Surendra Mettapalli, Newton Alex

Select Quiz

Quiz 1

SHOW QUIZ DETAILS

START

Figure 1.11 – Chapter Review Questions for Chapter 1

Chapter 2: Implementing a Partition Strategy

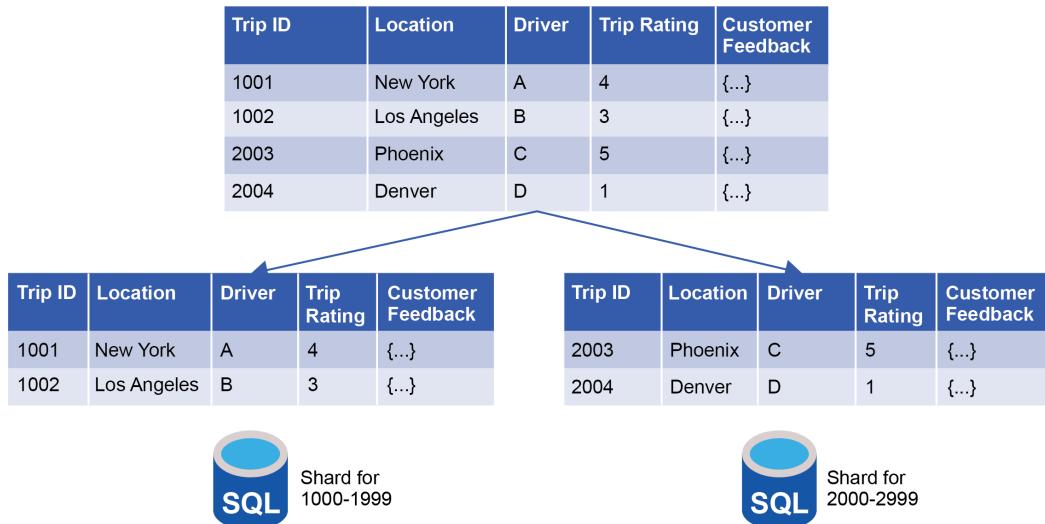


Figure 2.1 – An illustration of horizontal partitioning in action

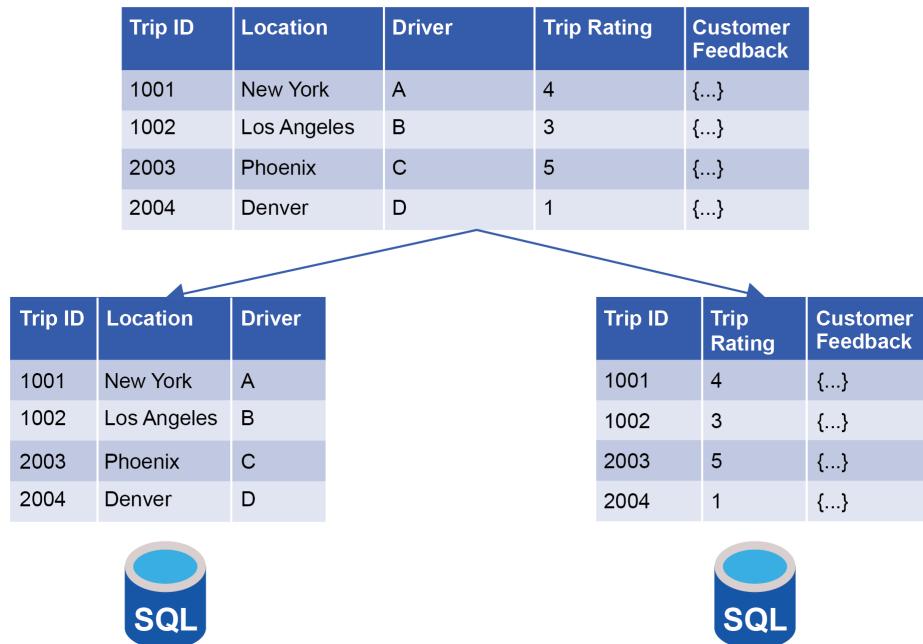


Figure 2.2 – An illustration of Vertical partitioning in action

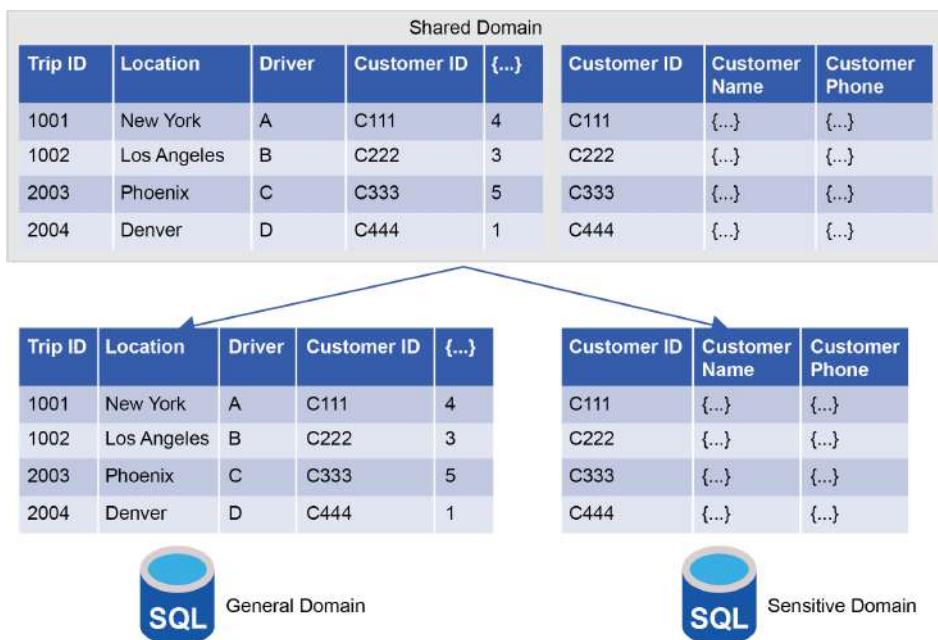


Figure 2.3 – Example of a functional partition

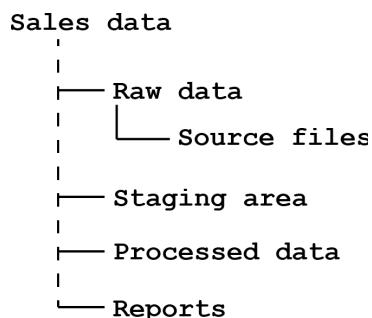


Figure 2.4 – The directory structure within ADLS Gen2

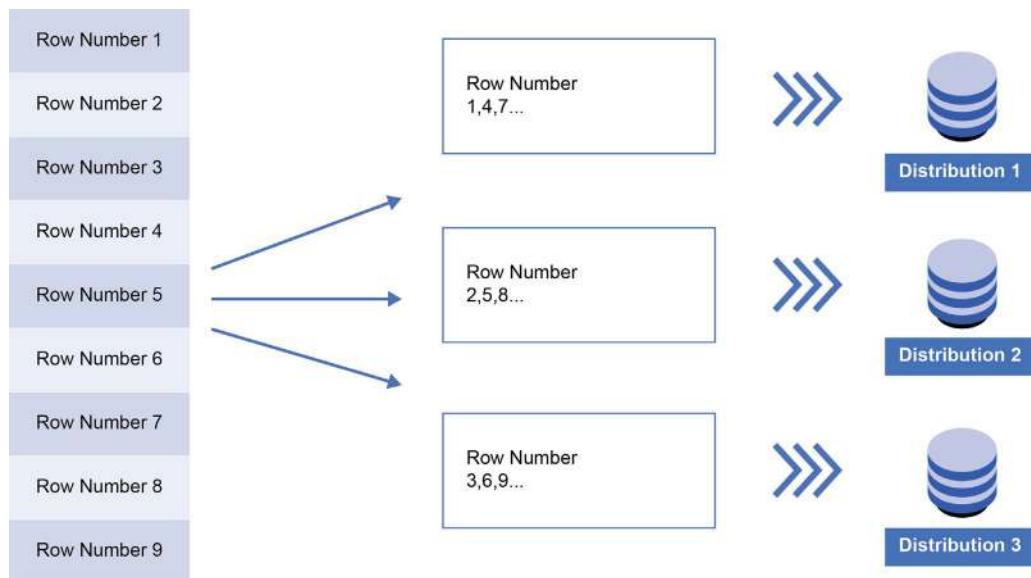


Figure 2.5 – Round-robin table distribution showing the distribution of data

Practice Resources

DASHBOARD > CHAPTER 2

Implementing a Partition Strategy

Summary

With that, you have come to the end of the second chapter. In this chapter, you learned about the different partitioning techniques available in Azure. You started with the basics of partitioning and its benefits of partitioning, before moving on to partitioning techniques for storage and analytical workloads. You next explored the best practices to improve partitioning efficiency and performance and reviewed the concept of distribution tables and how they impact the partitioning of Azure Synapse Analytics. Finally, you studied storage limitations, which play an important role in deciding when to partition for ADLS Gen2.

Chapter Review Questions

The Azure Data Engineer Associate Certification Guide
- Second Edition by Giacinto Palmieri, Surendra Metlapalli, Newton Alex

Select Quiz

Quiz 1 [SHOW QUIZ DETAILS](#) [START](#)

Figure 2.7 – Chapter Review Questions for Chapter 2

Chapter 3: Designing and Implementing the Data Exploration Layer

The screenshot shows a SQL editor interface with the following details:

- Toolbar:** Run, Undo, Publish, Query plan, Connect to (Built-in), ...
- Code Area:** A multi-line text input showing T-SQL code for creating an external table named "TripsExtTable". The code includes columns for tripId, driverId, customerId, cabId, tripDate, startLocation, and endLocation. It uses a WITH clause to define the table's location as a Parquet file in the "/parquet/trips/" directory, specifies the data source as "Dp203DataSource", and the file format as "Dp203ParquetFormat".
- Results Area:** A table view showing five rows of trip data. The columns are tripId, driverId, customerId, cabId, tripDate, startLocation, and endLocation.
- View Options:** Table (selected), Chart, Export results.

tripId	driverId	customerId	cabId	tripDate	startLocation	endLocation
107	207	307	407	20220203	Los Angeles	San Diego
108	208	308	408	20220301	Phoenix	Las Vegas
100	200	300	400	20220101	New York	New Jersey
101	201	301	401	20220102	Tempe	Phoenix
111	211	311	411	20220303	New York	New Jersey

Figure 3.1 – Creating a SQL pool external table

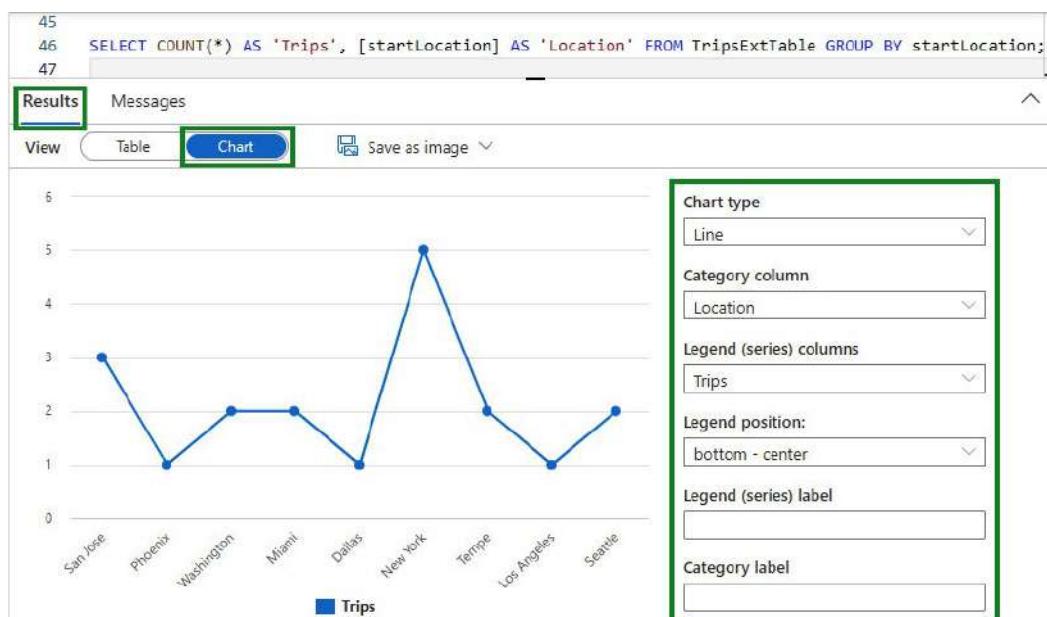


Figure 3.2 – Sample query on Parquet data visualized as a chart

```
1 df = spark.read.load('abfss://[REDACTED].dfs.core.windows.net/parquet/trips/*.parquet'
2   format='parquet')
3 spark.sql("CREATE DATABASE IF NOT EXISTS TripsDatabase")
4 df.write.mode("overwrite").option("overwriteSchema", "true").saveAsTable("TripsTable")
5 sqldf = spark.sql("""
6   SELECT COUNT(*) AS Trips,
7     startLocation AS Location
8   FROM TripsTable
9   GROUP BY startLocation """)
10 display(sqldf)
```

✓ 3 min 54 sec - Apache Spark session started in 2 min 58 sec 614 ms. Command executed in 55 sec 882 ms by newton.pac...

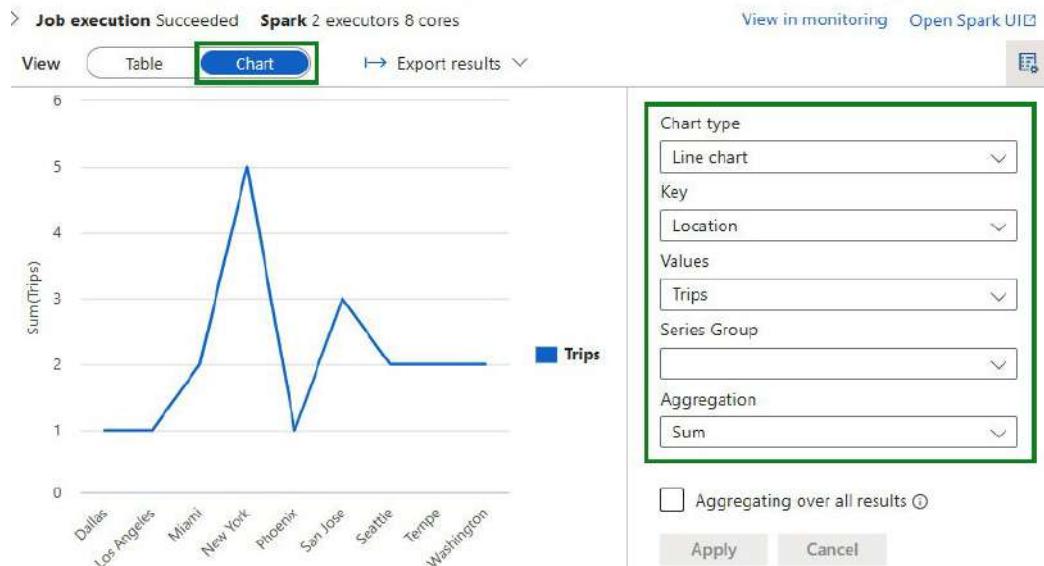


Figure 3.3 – Using Spark to query Parquet files

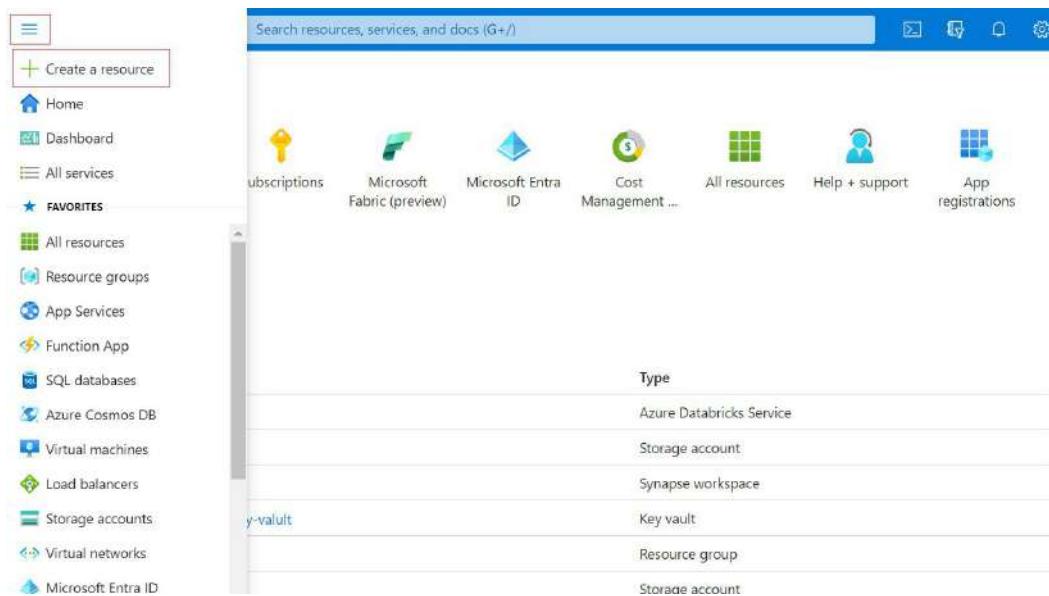


Figure 3.4 – Creating a resource in the Azure portal

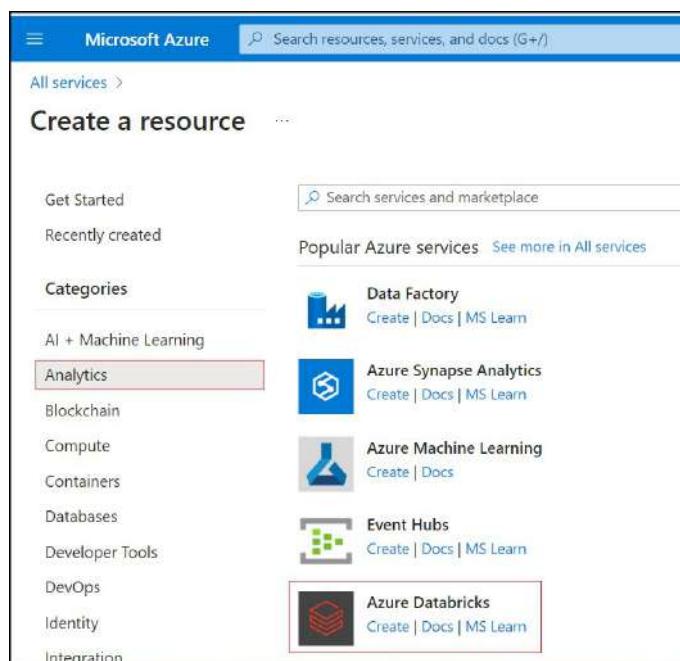


Figure 3.5 – Azure Databricks resource creation

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

Home > Create a resource > Marketplace > Azure Databricks > Create an Azure Databricks workspace ...

Basics Networking Encryption Security & compliance Tags Review + create

Project Details

Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources.

Subscription * ⓘ Azure subscription (New) rg-dp203-databricks Create new

Resource group * ⓘ (New) rg-dp203-databricks Create new

Instance Details

Workspace name * DP-203-databricks-workspace

Region * UK South

Pricing Tier * ⓘ Standard (Apache Spark, Secure with Microsoft Entra ID)

Managed Resource Group name Enter name for managed resource group

Review + create < Previous Next : Networking >

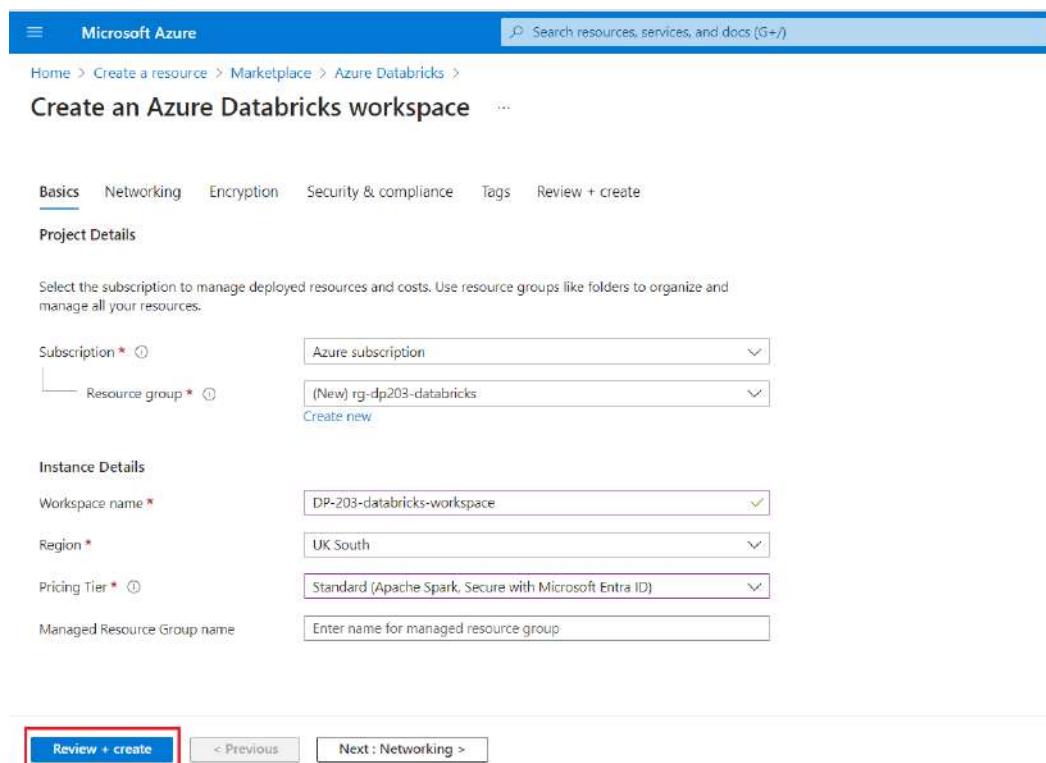


Figure 3.6 – Creating an Azure Databricks workspace

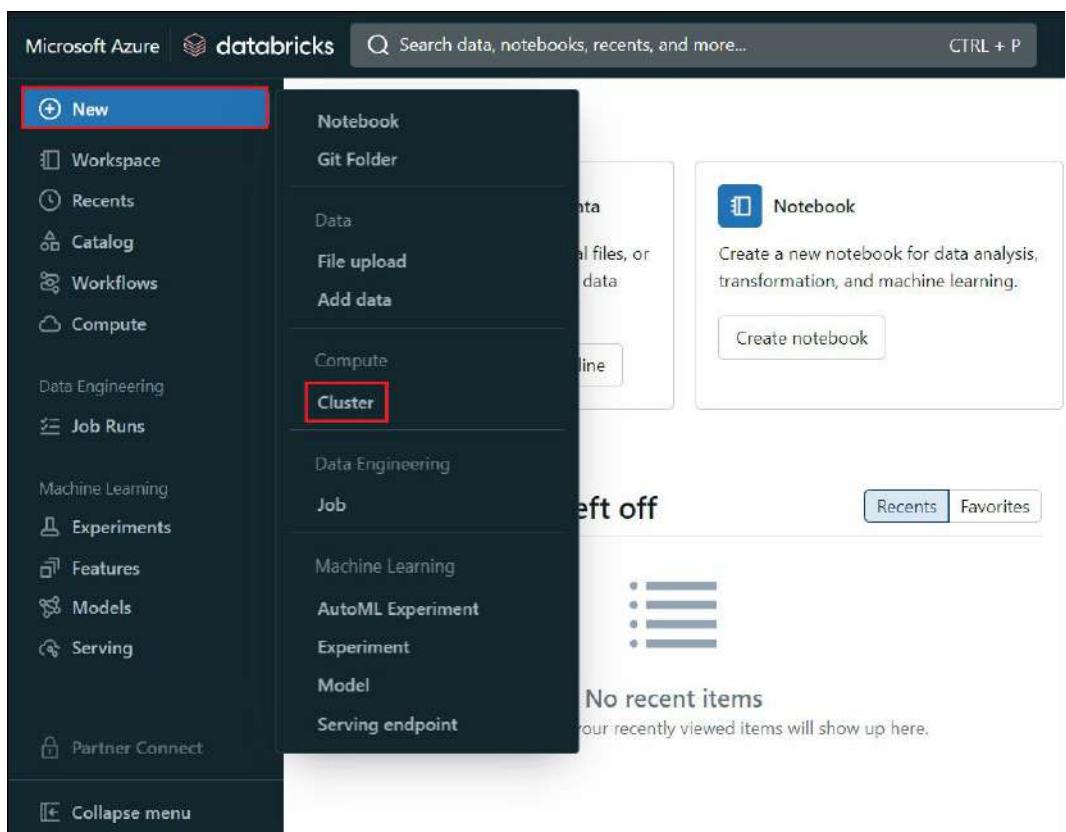


Figure 3.7 – Cluster creation in an Azure Databricks workspace

Microsoft Azure |  databricks | Search data, notebooks, recents, and more... | CTRL + P

+ New

Compute > New compute >

DP-203 Cluster

Access mode  Single user access 

Single user Surendra Mettapalli

Performance

Databricks runtime version 

Runtime: 14.3 LTS (Scala 2.12, Spark 3.5.0)

Use Photon Acceleration 

Worker type 

Standard_DS3_v2 14 GB Memory, 4 Cores

Min workers 2 Max workers 8  Spot instances 

Driver type 

Same as worker 14 GB Memory, 4 Cores

Enable autoscaling 
 Terminate after 120 minutes of inactivity 

Tags 

Add tags

Key Value Add

Create compute  Cancel

Partner Connect

Collapse menu

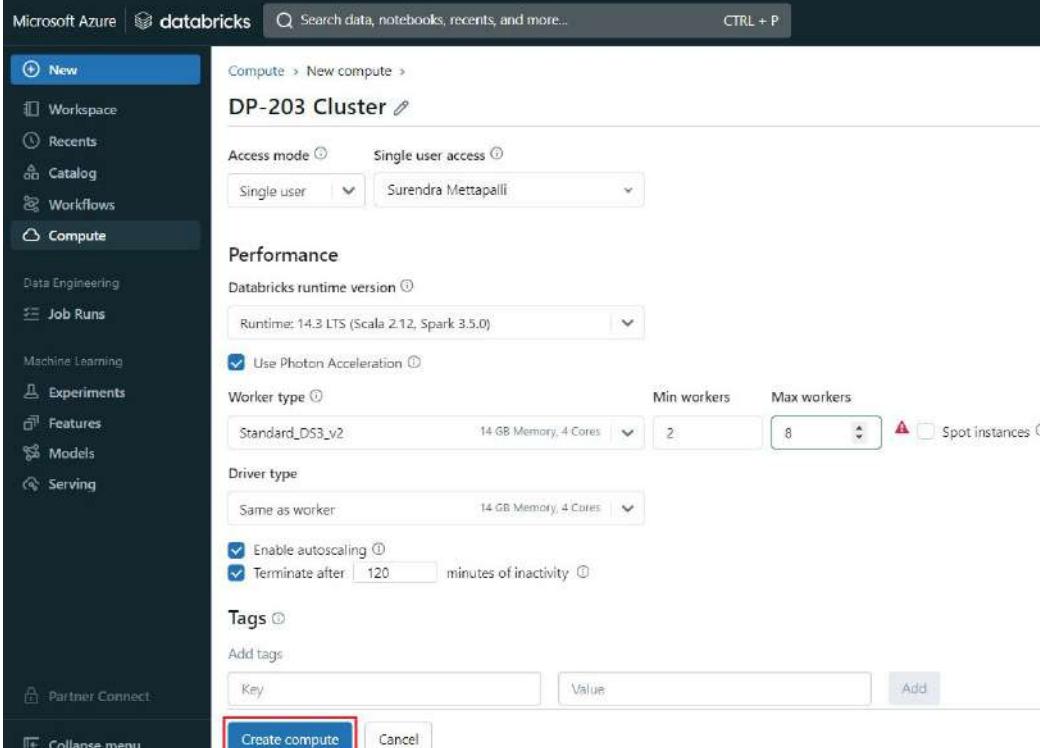


Figure 3.8 – Providing values to create a cluster in Azure Databricks

ParquetWithSpark-C7 Python

Schedule Share

ADBsmall

```
3
4 df =
5 spark.read.load('abfss://[REDACTED].dfs.core.windows.net/
6 parquet/trips/*.parquet',
7 format='parquet')
8 df.show(5)
```

▶ (3) Spark Jobs

▶ df: pyspark.sql.dataframe.DataFrame = [tripId: string, driverId: string ... 5 more fields]

tripId	driverId	customerId	cabId	tripDate	startLocation	endLocation
116	212	312	412	20220405	Washington	Atlanta
117	212	312	412	20220405	Seattle	Portland
118	212	312	412	20220405	Miami	Tampa
114	212	312	412	20220404	San Jose	Oakland
115	212	312	412	20220404	Tempe	Scottsdale

only showing top 5 rows

Figure 3.9 – An Azure Databricks notebook query to read Parquet files

Python

```
1 df = spark.read.load('abfss://[REDACTED].dfs.core.windows.net/parquet/trips/*.parquet'
2 format='parquet')
3 spark.sql("CREATE DATABASE IF NOT EXISTS TripsDatabase")
4 df.write.mode("overwrite").option("overwriteSchema", "true").saveAsTable("TripsTable")
5 sqldf = spark.sql("""
6     SELECT COUNT(*) AS Trips,
7         startLocation AS Location
8     FROM TripsTable
9     GROUP BY startLocation """
10 display(sqldf)
```

▶ (7) Spark.Jobs

▶ df: pyspark.sql.dataframe.DataFrame = [tripId: string, driverId: string ... 5 more fields]

▶ sqldf: pyspark.sql.dataframe.DataFrame = [Trips: long, Location: string]

Chart Data Profile

Location	Trips
Washington	2
Seattle	2
Miami	2
San Jose	4.5
New York	1
Phoenix	1
Los Angeles	1
Dallas	1
Tempe	2

Plot Options...

Figure 3.10 – A simple Spark query on Parquet data in Databricks

The screenshot shows the Microsoft Azure Synapse Analytics interface. At the top, it displays the URL "Microsoft Azure | Synapse Analytics > synapse-az-ws". Below this is a search bar with the placeholder "Search". On the left, there's a vertical sidebar with icons for Home, Databases, Notebooks, SQL scripts, and Pipelines. The main area is titled "Gallery" and has tabs for "Database templates", "Datasets", "Notebooks", "SQL scripts", and "Pipelines". A "Filter by keyword" input field is present. The "Database templates" section contains eight cards, each with an icon and a title followed by a brief description:

- Agriculture**: For companies engaged in growing crops, raising livestock and dairy production.
- Airlines**: For companies operating passenger or cargo airline services.
- Automotive**: For companies manufacturing automobiles, heavy vehicles, tires, and other automotive components.
- Banking**: For companies providing a wide range of banking and related financial services.
- Energy & Commodity Trading**: For traders of energy, commodities, or carbon credits.
- Freight & Logistics**: For companies providing freight and logistics services.
- Fund Management**: For companies managing investment funds on behalf of investors.
- Genomics**: For companies acquiring and analyzing genomic data about human beings or other species.

At the bottom left, there's a "Continue" button.

Figure 3.11 – A Synapse database showing the available templates

The screenshot shows the Microsoft Azure Synapse Analytics workspace interface. At the top, there's a blue header bar with the text "Microsoft Azure | Synapse Analytics > synapse-az-ws". Below the header, there are navigation buttons for "Synapse live", "Validate all", "Publish all", and a "Data" section. The "Data" section has tabs for "Workspace" (which is selected) and "Linked". A search bar labeled "Filter resources by name" is present. Under the "Lake database" heading, there are four entries: "Airlines_b4n", "Banking_bvs" (which is highlighted with a green box), "default", and "nyc_taxi_ldw_spark". Under the "SQL database" heading, there is one entry: "5".

Figure 3.12 – The Synapse workspace showing the lake database templates

The screenshot shows the Microsoft Purview governance portal. The left sidebar contains navigation links for "Data sources", "Collections", "Monitoring", "Metamodel", "Asset types (preview)", "Data sharing", "Shares", "Shared sources", "Share myree", "Source management", "Scan rule sets", "Pattern rules", "Integration runtimes", "Annotation management", "Classifications", and "Classification rules". The main area is titled "Data sources" and shows a "Register data source (Azure Synapse Analytics)" form. The form fields include: "Data source name" (set to "AzureSynapseAnalytics-CTA"), "Azure subscription" (set to "All"), "Workspace name" (with a dropdown menu), "Dedicated SQL endpoint" (empty), "Serverless SQL endpoint" (empty), and "Collection" (set to "root DP203-Purview Account"). A note at the bottom states: "All items in this data source will belong to the collection that you select." At the bottom right of the form are "Register", "Back", and "Cancel" buttons.

Figure 3.13 – Registering a data source in the Microsoft Purview governance portal

The screenshot shows the Microsoft Purview Data Map interface. The left sidebar contains a navigation menu with the following items:

- Data sources
- Collections
- Monitoring
- Metamodel
- Asset types (preview)
- Data sharing
- Shares
- Shared sources
- Share invites
- Source management
- Scan rule sets
- Pattern rules
- Integration runtimes
- Annotation management
- Classifications
- Classification rules

The main area displays a collection named "DP203-Purview-Account" which is described as "The default container". It lists two data sources: "AzureSynapseAnalytics..." (Azure Synapse Analytics) and "AzureDataLakeStorage..." (Azure Data Lake Storage Gen2). Each data source entry includes a "View details" button.

Figure 3.14 – Visualizing a Data Map in a Synapse workspace and Data Lake Storage Gen2 Account



Figure 3.15 – An example of data lineage for Power BI

The screenshot shows the Microsoft Purview interface for 'DP203-Purview-Account'. The left sidebar has a 'Browse assets' section with a 'Data' category expanded, showing options like Dashboard, Data pipeline, Data share, Database, File, Folder, Report, Stored procedure, and Table. Below 'Data' is a 'Business' category with Application service, Business process, Data domain, and Glossary terms. The main pane displays a search result for 'synapseazdl' under the 'Data source type: All' filter. The result is 'synapseazdl' (Azure Data Lake Storage Gen2 Service), which is a 'No description available' entry with a 'Fully qualified name: https://synapseazdl.dfs.core.windows.net'. It was 'Updated 18 minutes ago'. The search bar at the top contains 'Search catalog'.

Figure 3.16 – Filtering search results with facets

This screenshot is identical to Figure 3.16, but it includes a date filter 'Created within: Last 24 hours' in the search facet bar. The search result for 'synapseazdl' remains the same, showing it was updated 21 minutes ago. The rest of the interface, including the sidebar categories and the overall layout, is consistent with Figure 3.16.

Figure 3.17 – Filtering search results with date filters



DASHBOARD > CHAPTER 3

Designing and Implementing the Data Exploration Layer

Summary

This chapter helped you to develop your expertise in navigating Azure Data Services effectively, such as mastering queries with Serverless SQL and Spark clusters and utilizing Azure Synapse Analytics templates for efficient database management, enabling you to focus on data analysis. Additionally, it emphasized the importance of keeping track of data lineage with Microsoft Purview to know about data's origins, transformation, and destination, ensuring transparency and traceability. Furthermore, you now have a sound knowledge of exploring and managing metadata, making it easier for you to locate and govern your datasets using the Microsoft Purview data catalog.

All these topics complete the syllabus for DP203 – Design and implement the data exploration layer, and with the completion of this chapter, you have now learned how to design your own serving layer in Azure.

In Chapter 4, *Ingesting and Transforming Data*, you will develop data processing systems that include reading data using different file formats, encoding, performing data cleansing, handling duplicate and missing data, error handling, and running transformations using services such as Spark, SQL, ADF, and Azure Synapse Analytics.

Chapter Review Questions

The Azure Data Engineer Associate Certification Guide
- Second Edition by Giacinto Palmieri, Surendra Mettapalli, Newton Alex

Select Quiz

Quiz 1

[SHOW QUIZ DETAILS](#) ▾[START](#)

Figure 3.19 – Chapter Review Questions for Chapter 3

Chapter 4: Ingesting and Transforming Data



Figure 4.1 – The Data factory workspace landing screen showing a range of ADF activities

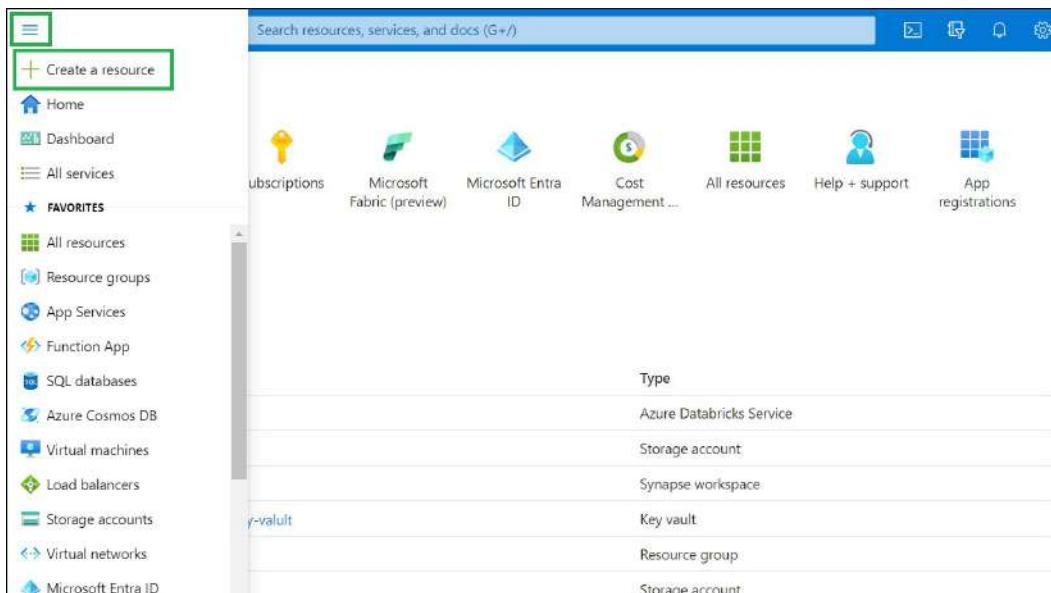


Figure 4.2 – Creating a ADF resource in the Azure portal

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

All services > Create a resource ...

Get Started 

Recently created Popular Azure services See more in All services

Categories

- AI + Machine Learning
- Analytics**
- Blockchain
- Compute
- Containers
- Databases
- Developer Tools
- DevOps
- Identity
- Integration

Data Factory 
[Create](#) | [Docs](#) | [MS Learn](#)

Azure Synapse Analytics 
[Create](#) | [Docs](#) | [MS Learn](#)

Azure Machine Learning 
[Create](#) | [Docs](#)

Event Hubs 
[Create](#) | [Docs](#) | [MS Learn](#)

Azure Databricks 
[Create](#) | [Docs](#) | [MS Learn](#)

Figure 4.3 – ADF resource creation

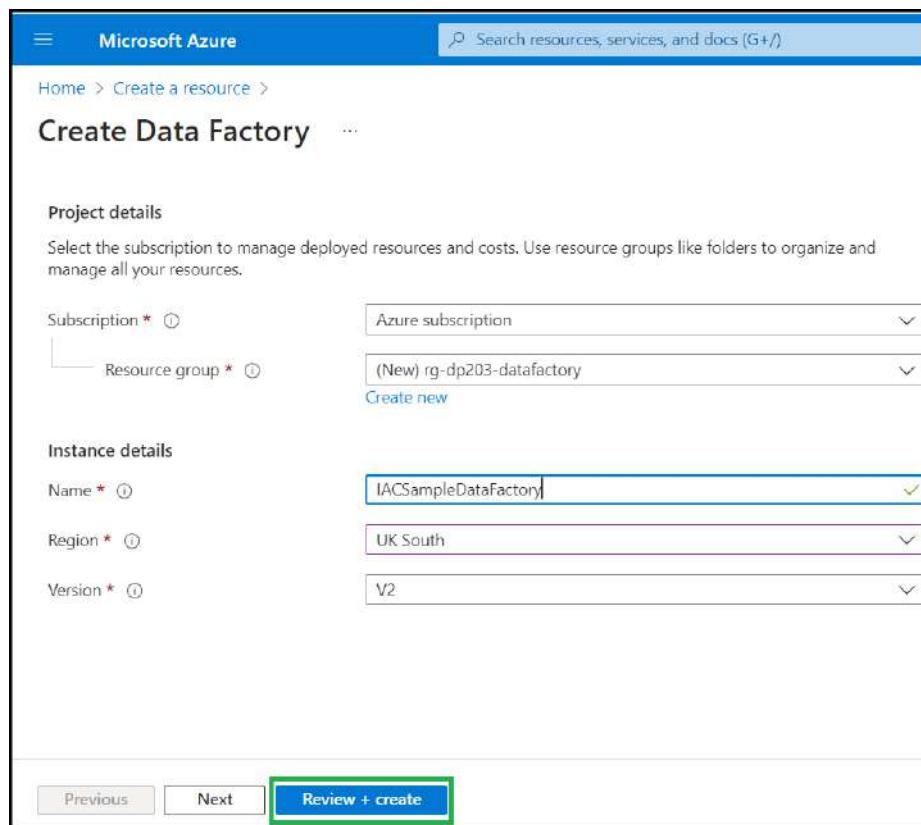


Figure 4.4 – Creating an ADF resource

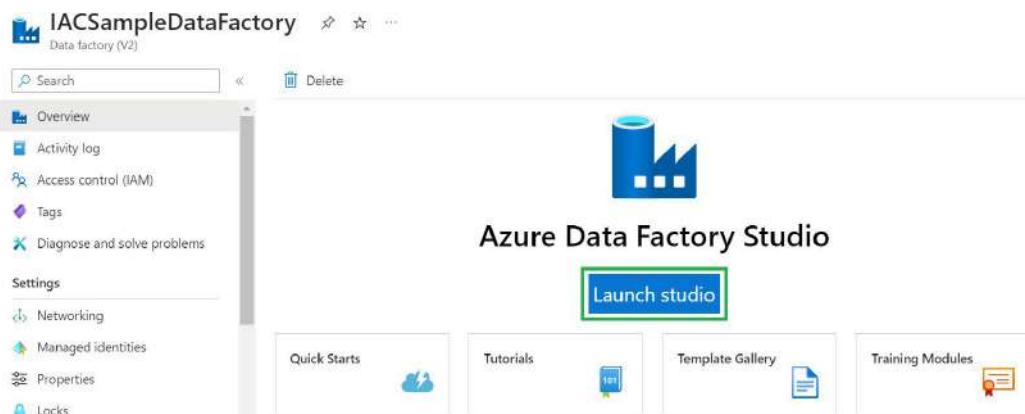


Figure 4.5 – ADF prompting to open the workspace and start building pipelines

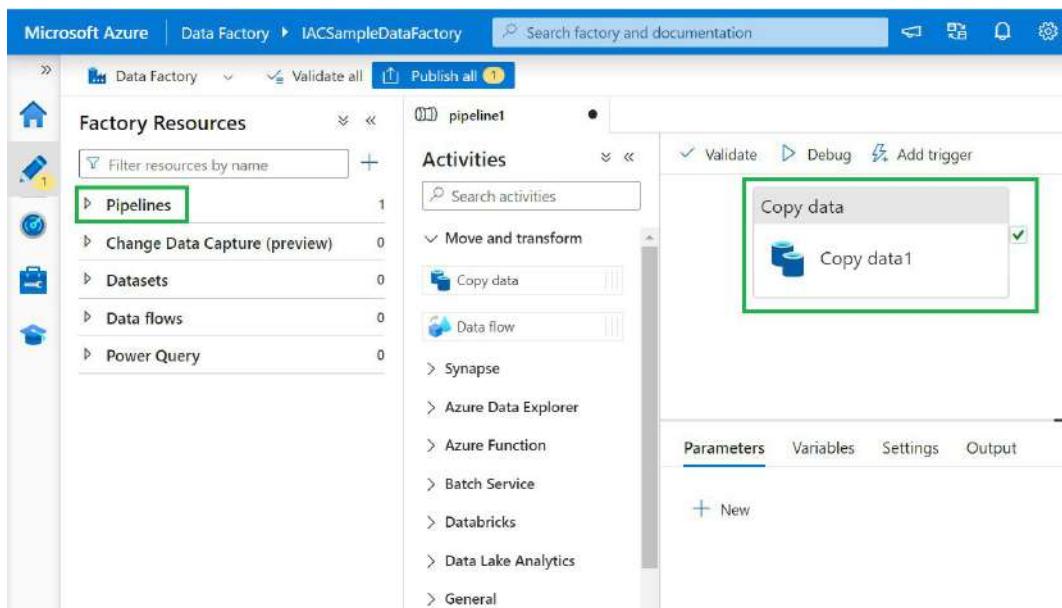


Figure 4.6 – ADF Studio to build the pipeline activities

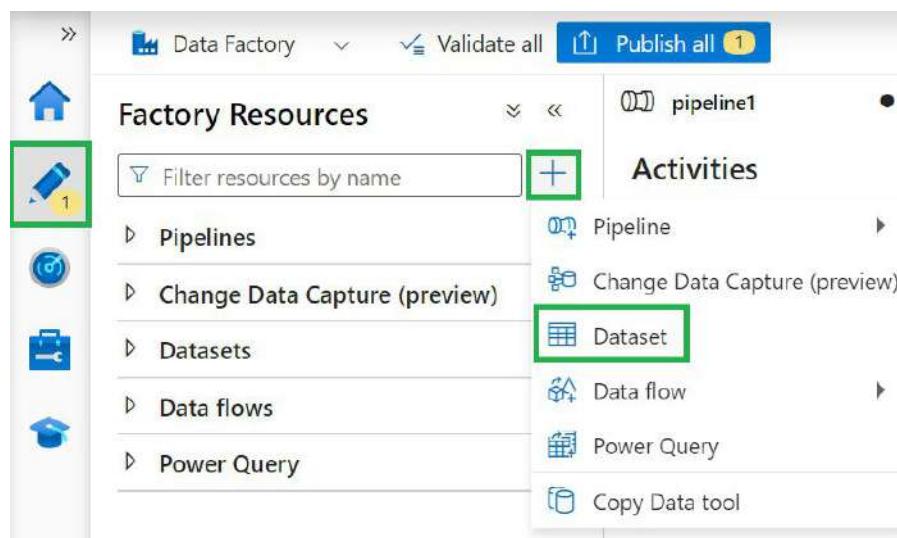


Figure 4.7 – Creating a new dataset in ADF

New dataset

In pipeline activities and data flows, reference a dataset to specify the location and structure of your data within a data store. [Learn more](#)

Select a data store

All **Azure** Database File Generic protocol NoSQL Services and apps

		
Azure AI Search	Azure Blob Storage	Azure Cosmos DB for MongoDB
		
Azure Cosmos DB for NoSQL	Azure Data Explorer (Kusto)	Azure Data Lake Storage Gen2

Continue **Cancel**

Figure 4.8 – Dataset source options in ADF

The screenshot shows the Azure SQL Query editor interface. On the left, the navigation pane includes links for Overview, Activity log, Tags, Diagnose and solve problems, Quick start, Query editor (preview), Power Platform (Power BI, Power Apps, Power Automate), Settings (Compute + storage, Connection strings, Properties, Locks), and a search bar. The main area has tabs for Query 1 (selected), Run, Cancel query, Save query, Export data as, Show only Editor, Results, and Messages. The Results tab displays a table with three rows of data from the FactTrips table.

tripID	customerID	LastModifiedTime
100	200	2021-09-03T16:44:54.1300000
101	201	2021-09-03T16:45:22.0870000
102	202	2021-09-03T16:45:22.0930000

```
3 CREATE TABLE FactTrips (
4     tripID INT,
5     customerID INT,
6     LastModifiedTime DATETIME
7 );
8
9 INSERT INTO [dbo].[FactTrips] values (100, 200, CURRENT_TIMESTAMP);
10 INSERT INTO [dbo].[FactTrips] values (101, 201, CURRENT_TIMESTAMP);
11 INSERT INTO [dbo].[FactTrips] values (102, 202, CURRENT_TIMESTAMP);
```

Figure 4.9 – Creating a simple table using the Query editor in Azure SQL

The screenshot shows the Azure Data Factory (ADF) authoring page. The left sidebar lists Factory Resources: Pipelines (1), Change Data Capture (preview), Datasets, Data flows, and Power Query. The Pipelines item is highlighted with a green box. The main area shows a pipeline named "pipeline1" with an "Activities" section. The "General" tab is selected, showing a list of activities: Append variable, Delete, Execute Pipeline, Execute SSIS package, Fail, Get Metadata, and a "Lookup" activity which is also highlighted with a green box. To the right, the "Lookup1" activity is detailed in a panel with tabs for General, Settings, and User properties. The General tab shows the Name as "Lookup1" and a Description field.

Figure 4.10 – The ADF authoring page

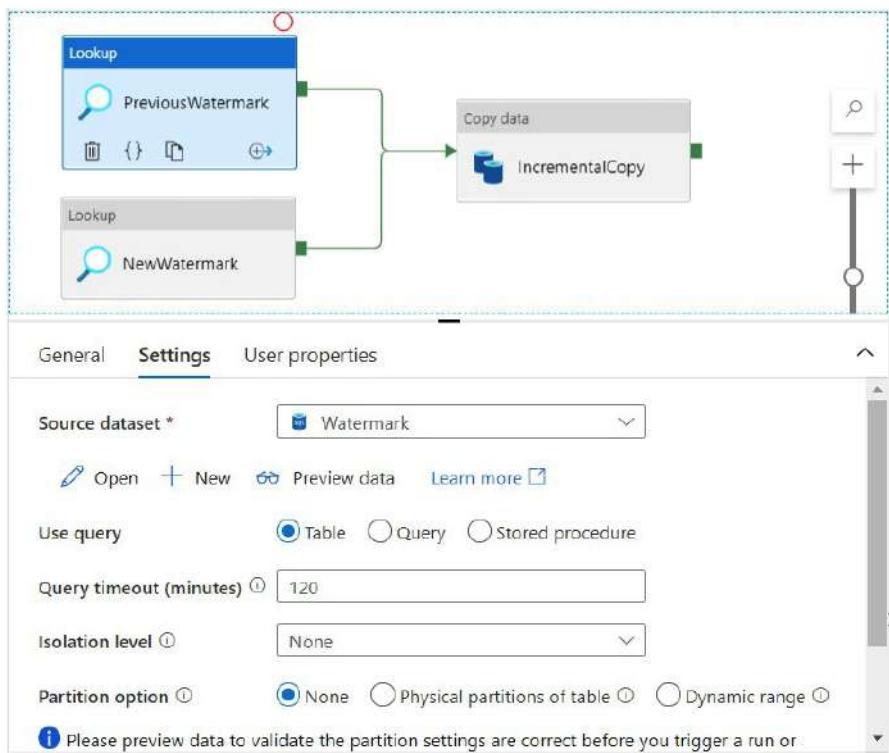


Figure 4.11 – The lookup activity configuration using the Watermark table

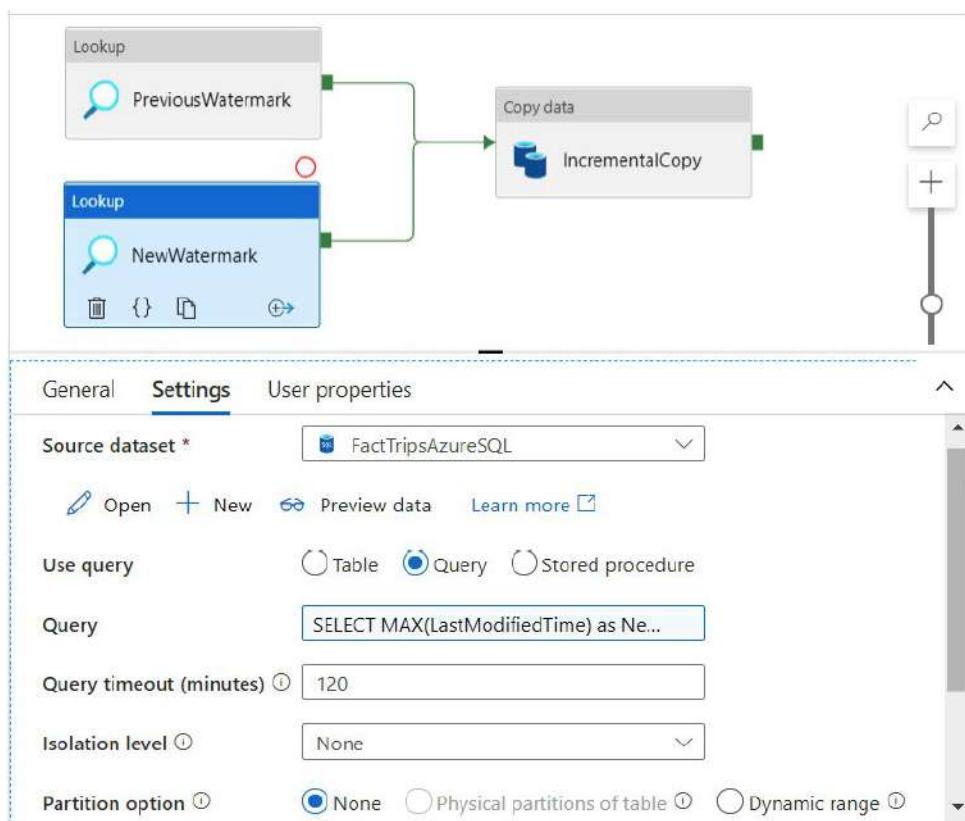


Figure 4.12 – New watermark lookup configuration using LastModifiedTime

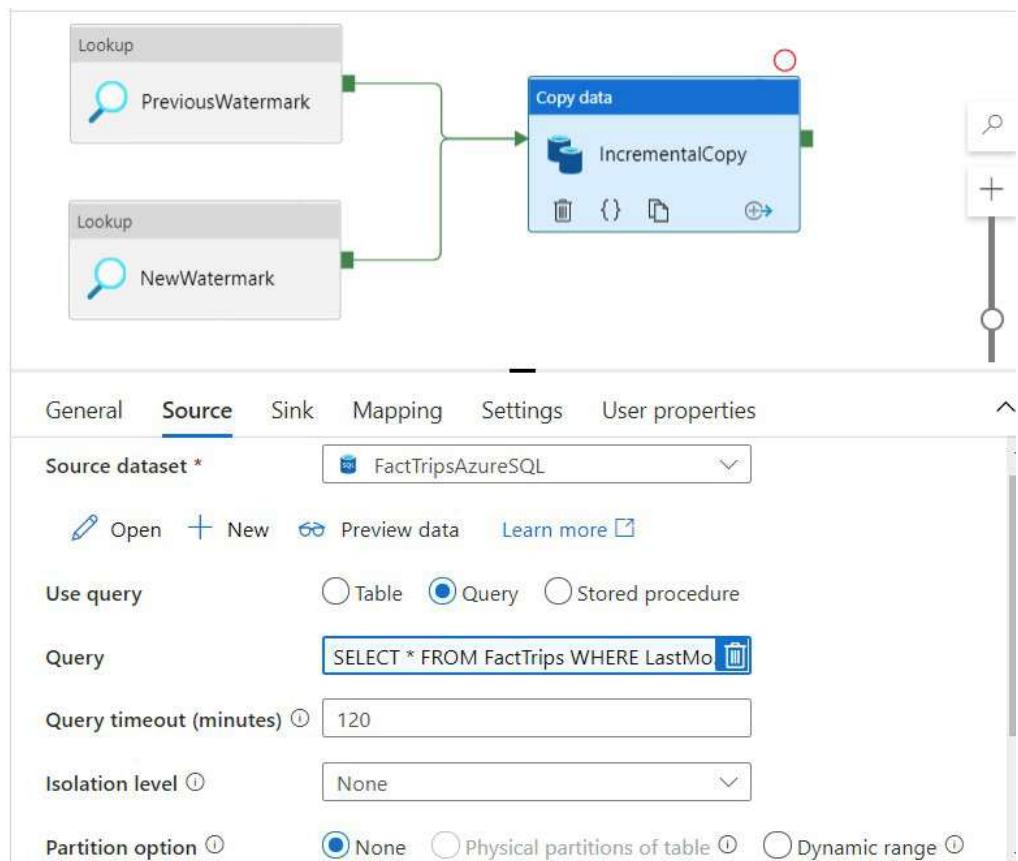


Figure 4.13 – ADF Copy activity with watermark-based delta generation

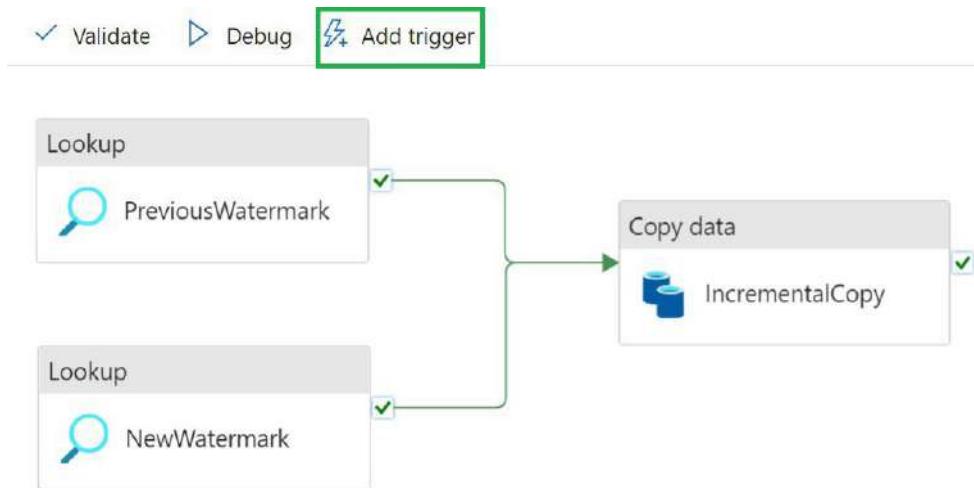


Figure 4.14 – Scheduling the ADF pipeline by clicking Add trigger

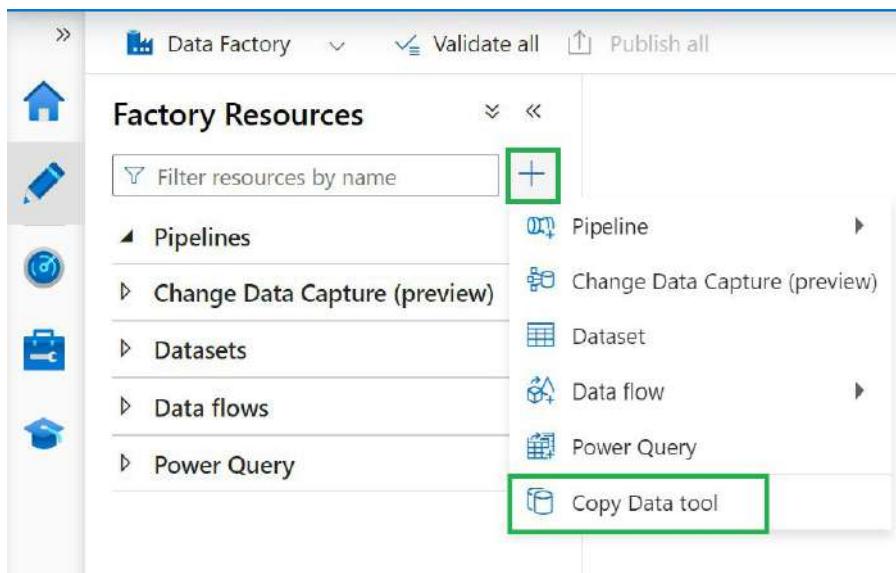


Figure 4.15 – Scanning files at the source based on the LastModifiedDate attribute

Copy Data tool

The screenshot shows the 'Copy Data tool' wizard interface. On the left, a vertical navigation bar lists five steps: 1 Properties (highlighted with a blue circle), 2 Source, 3 Target, 4 Settings, and 5 Review and finish. The main content area is titled 'Properties' and contains the following information:

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 cadence or task schedule *

Run once now Schedule Tumbling window

Start Date (UTC) * [?](#)
06/24/2021 7:06 PM

Recurrence * [?](#)
Every Hour(s) [▼](#)

Specify an end date

[Advanced](#)

Figure 4.16 – Selecting Tumbling window for an incremental load, based on the file-modified time

Copy Data tool

The screenshot shows the 'Copy Data tool' interface with a vertical navigation bar on the left containing five steps: Properties (selected), Source, Dataset, Configuration, Target, Settings, and Review and finish. The 'Source' step is currently active, displaying configuration options for a 'Source data store'. The 'Source type' is set to 'Azure Blob Storage' and the 'Connection' is 'BlobSource'. The 'File or folder' field contains 'testcontainer/' and includes a 'Browse' button. Under 'Options', the 'File loading behavior' dropdown is set to 'Load all files', which is highlighted in grey. Other options include 'Load all files', 'Incremental load: LastModifiedDate', 'Incremental load: time-partitioned folder/file names', 'Recursively' (checked), and 'Delete files after completion'. The 'Max concurrent connections' field is empty. At the bottom are 'Previous' and 'Next >' buttons.

Properties

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

Connection *

File or folder *

If the identity you use to access the data store only has permission to subdirectory instead of the entire account, specify the path to browse.

Options

File loading behavior

Load all files

Incremental load: LastModifiedDate

Incremental load: time-partitioned folder/file names

Recursively ①

Delete files after completion ①

Max concurrent connections

Figure 4.17 – Incremental load with LastModifiedDate timestamp behavior

Copy Data tool

Properties

Source

Dataset

Configuration

Target

Settings

Review and finish

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: Azure Blob Storage

Connection: BlobSource

File or folder: testcontainer/{year}/{month}/{day}

Options

File loading behavior: Incremental load: time-partitioned folder/file names

year format: yy

month format: MM

day format: dd

Time to preview generated file path:

< Previous Next >

The screenshot shows the 'Source' step of the ADF Copy Data tool. The 'Source type' is set to 'Azure Blob Storage' and the 'Connection' is 'BlobSource'. The 'File or folder' field contains the path 'testcontainer/{year}/{month}/{day}'. In the 'Options' section, 'File loading behavior' is set to 'Incremental load: time-partitioned folder/file names'. The 'year format' is 'yy', 'month format' is 'MM', and 'day format' is 'dd'. At the bottom, there are 'Previous' and 'Next' navigation buttons.

Figure 4.18 – ADF incremental load option with time-partitioned folders

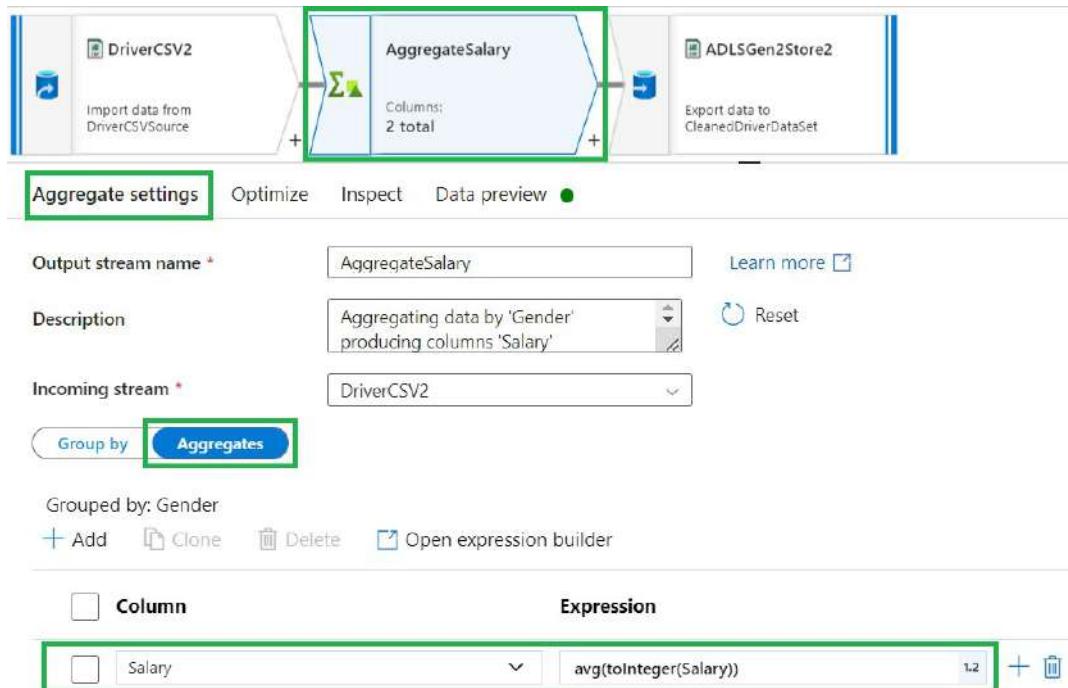


Figure 4.19 – Performing Aggregate transformation in ADF

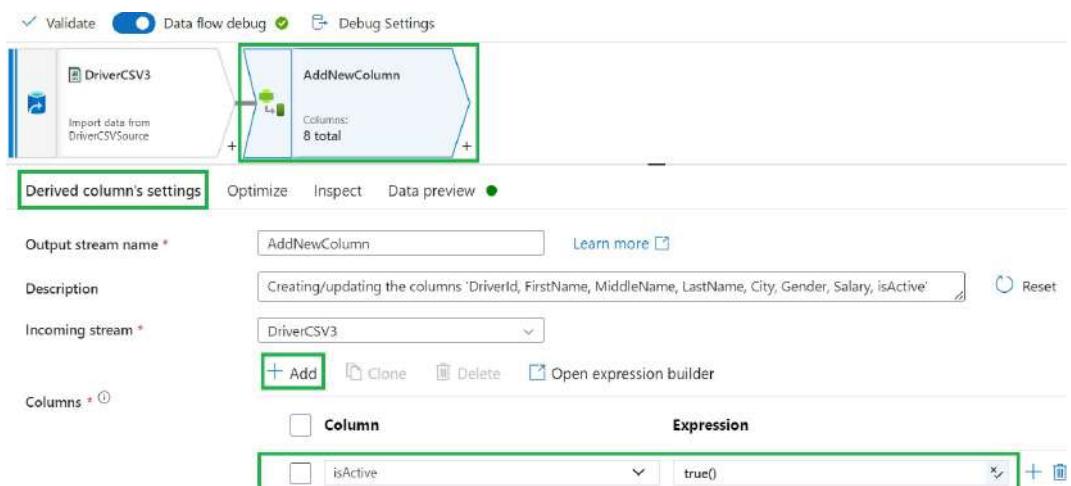


Figure 4.20 – Derived column transformation in ADF showing an added column

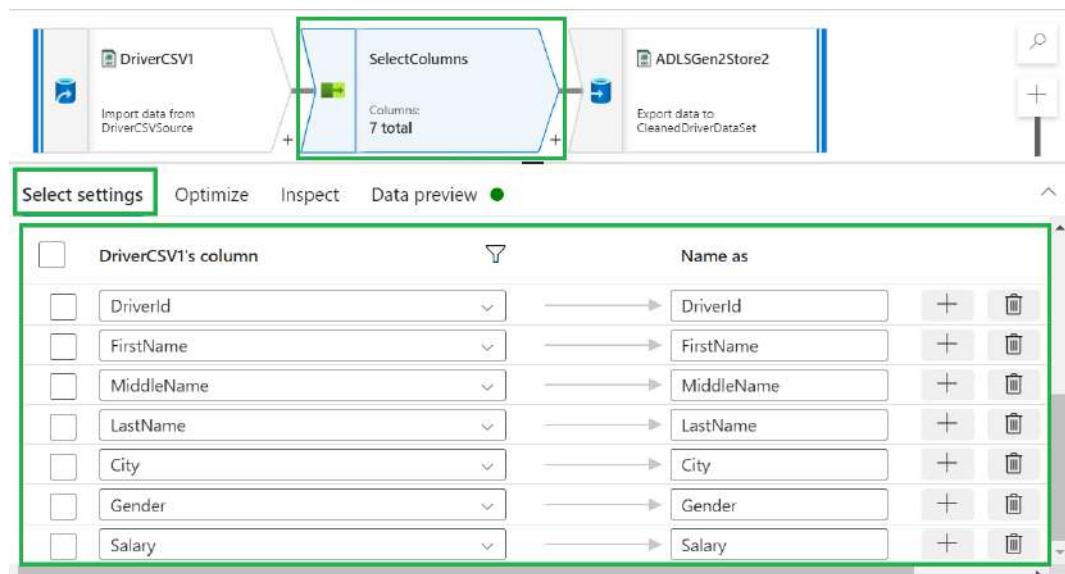


Figure 4.21 – Select transformation in ADF showing selecting required columns

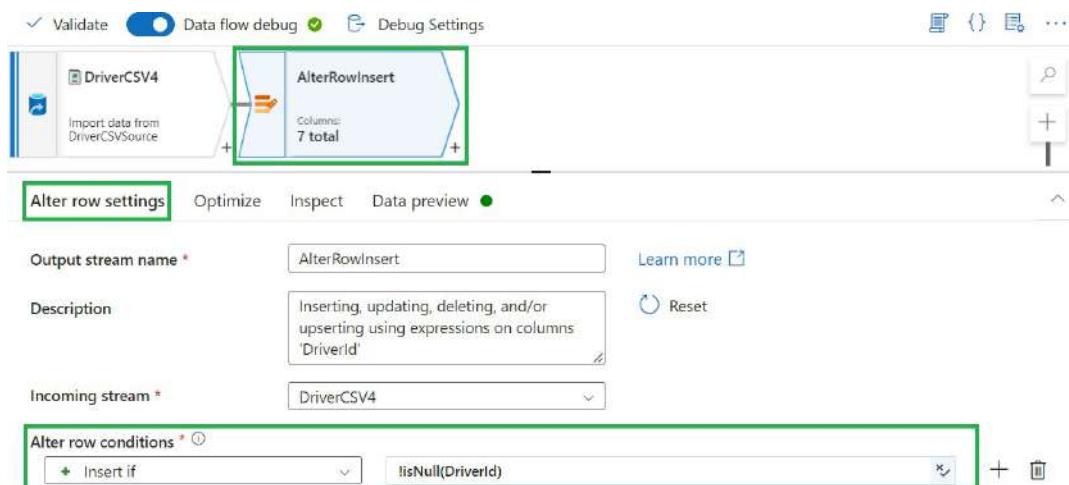


Figure 4.22 – Performing Alter row transformation as per NULL condition

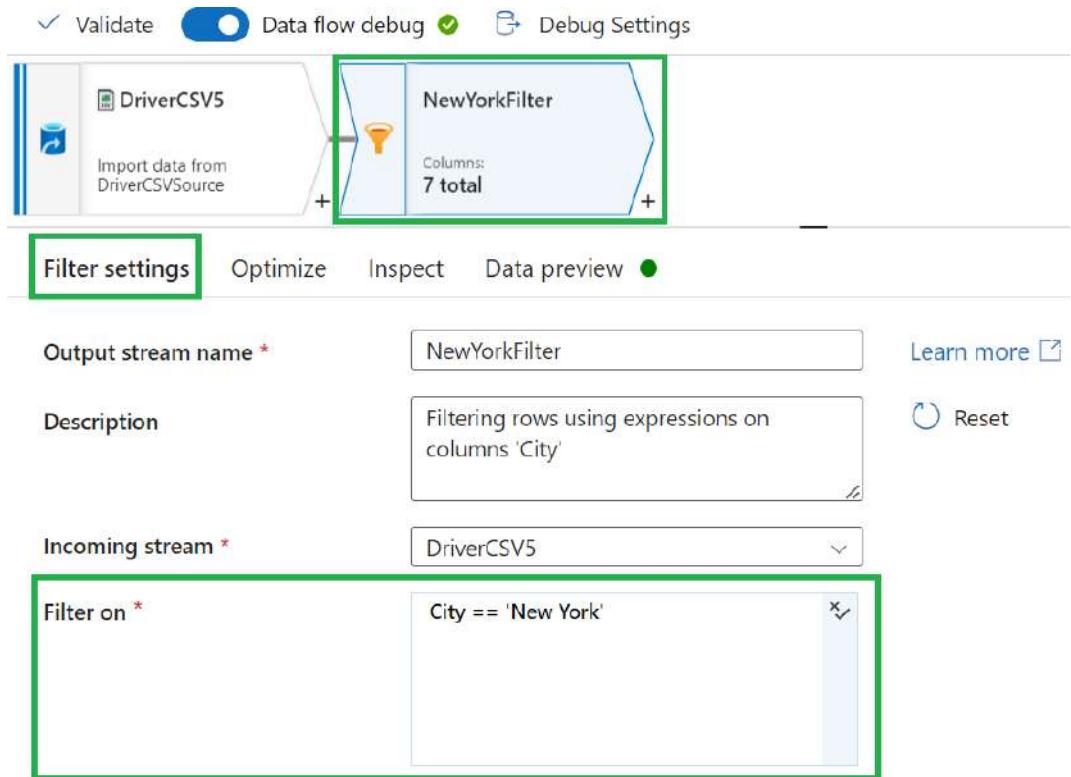


Figure 4.23 – Performing Filter transformation in ADF as per City

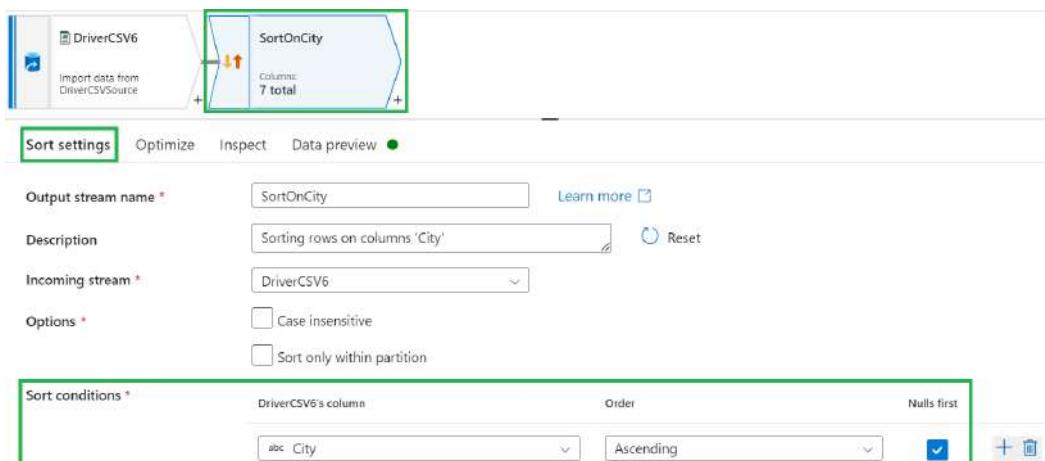


Figure 4.24 – Performing Sort transformation on specific rows

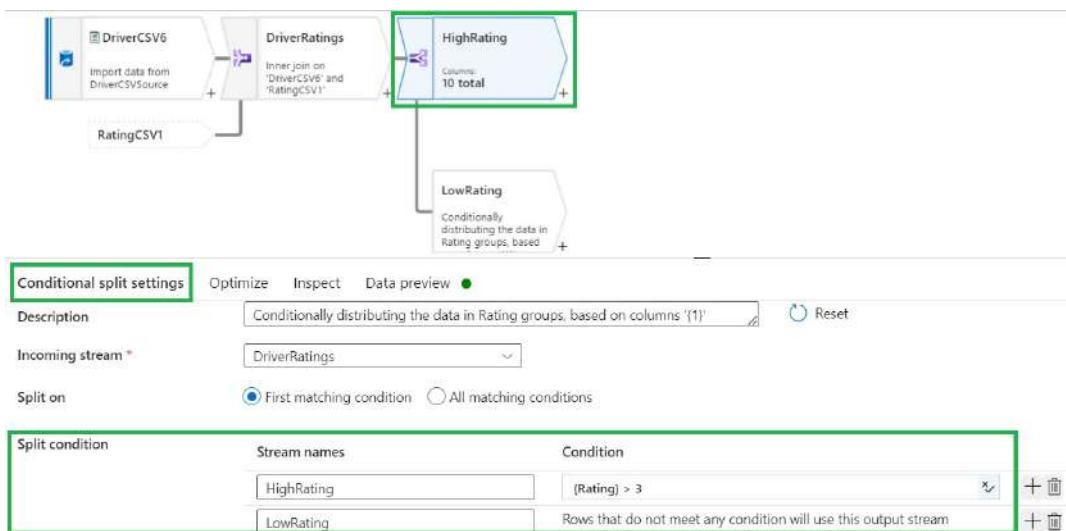


Figure 4.25 – Performing Conditional Split transformation in ADF for output or input stream

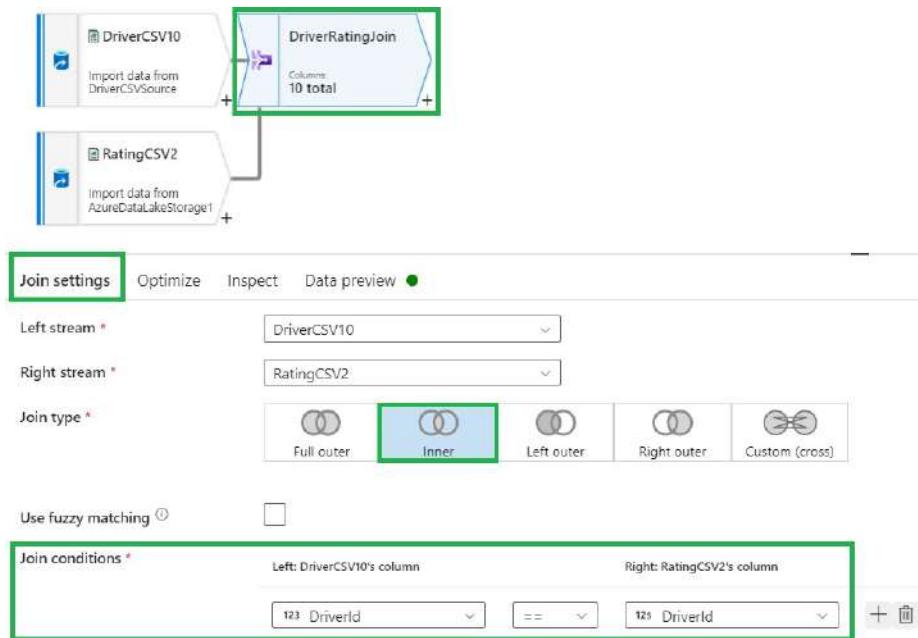


Figure 4.26 – Performing Join transformation in ADF based on join condition

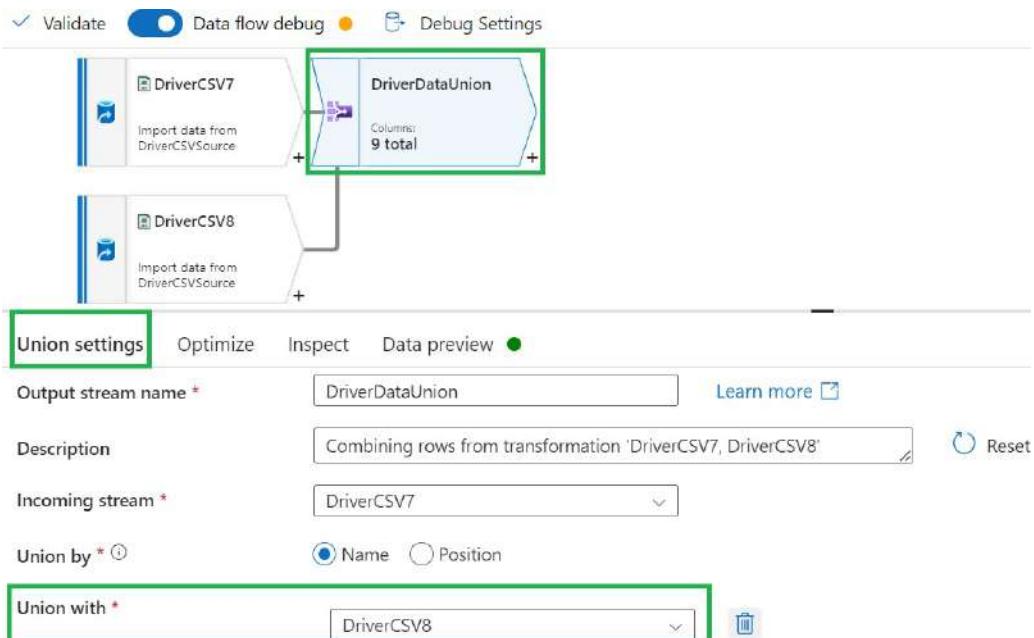


Figure 4.27 – Performing a Union example in ADF

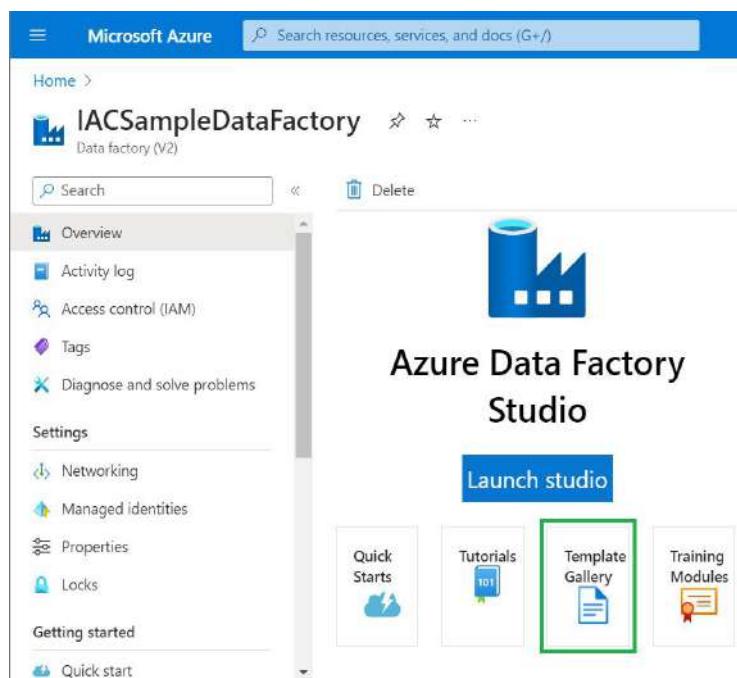


Figure 4.28 – Clicking the Template Gallery link to view the templates in ADF instance

↑ Import pipeline template

 Bulk Copy from Database to Azure Data Explorer Use this template to copy large amount of data in bulk from database like SQL Server, Google BigQuery, etc to Azure Data Explorer (ADX), using...  by Microsoft	 Bulk Copy from Database Use this template to copy data in bulk from database using external control table to store partition list of source tables. ...  by Microsoft
 Bulk Copy from Files to Database Use this template to copy data in bulk from Azure Data Lake Storage Gen2 to Azure Synapse Analytics / Azure Sql Database. If you want to copy data from a small number of...  by Microsoft	 Copy and convert data from Office 365 into Common Data Model for Open Data... Use this template to copy data from your Office 365 organization and convert it into Common Data Model format to be included in the Open Data...  by Microsoft
 Copy data from Google BigQuery to Azure Data Lake Store Use this template to copy data from Google BigQuery to Azure Data Lake Storage. ...  by Microsoft	 Copy data from HDFS to Azure Data Lake Store Use this template to copy data from HDFS (Hadoop Distributed File System) to Azure Data Lake Storage. ...  by Microsoft

Figure 4.29 – ADF template gallery showing copy and transformation activities templates

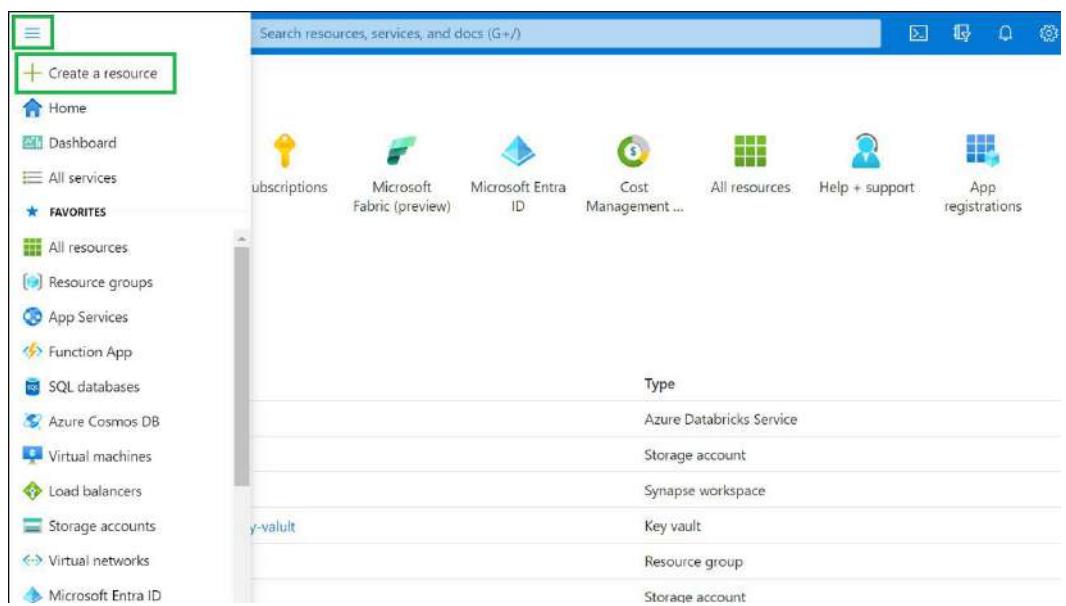


Figure 4.30 – Creating an Azure Synapse Analytics resource in the Azure portal

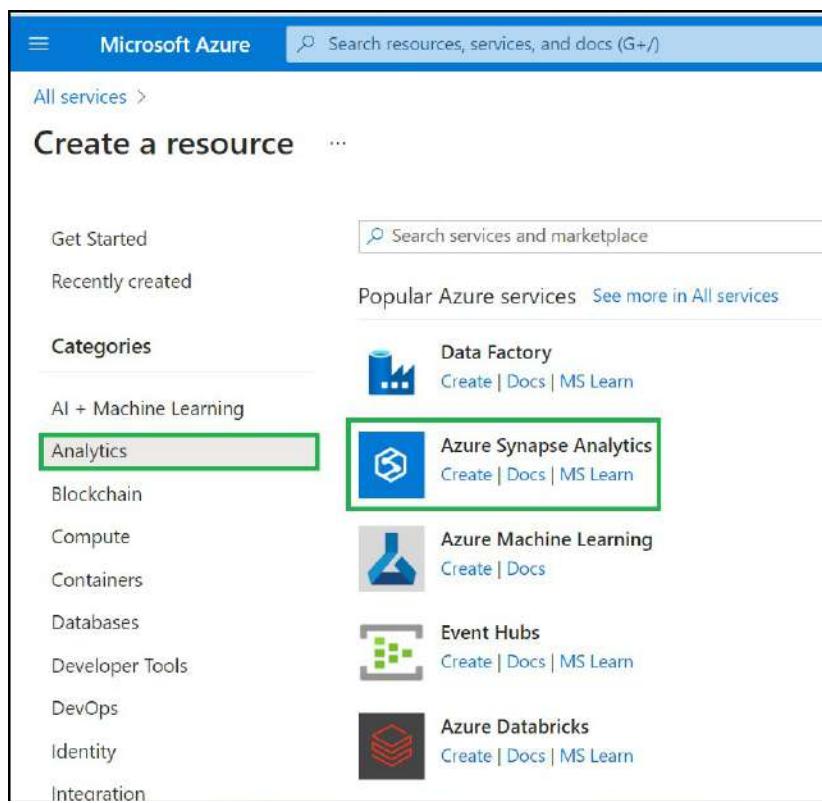


Figure 4.31 – Choosing an Azure Synapse Analytics service in the Azure portal

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

Home > Create a resource >

Create Synapse workspace

Subscription * (New) rg-dp203-synapseanalytics Create new

Resource group * (New) rg-dp203-synapseanalytics Create new

Managed resource group Enter managed resource group name

Workspace details

Name your workspace, select a location, and choose a primary Data Lake Storage Gen2 file system to serve as the default location for logs and job output.

Workspace name * iacsynapsews

Region * UK South

Select Data Lake Storage Gen2 * From subscription Manually via URL

Account name * synapseazdl Create new

File system name * ws-container Create new

Review + create < Previous Next: Security >

Figure 4.32 – Creating an Azure Synapse Analytics workspace

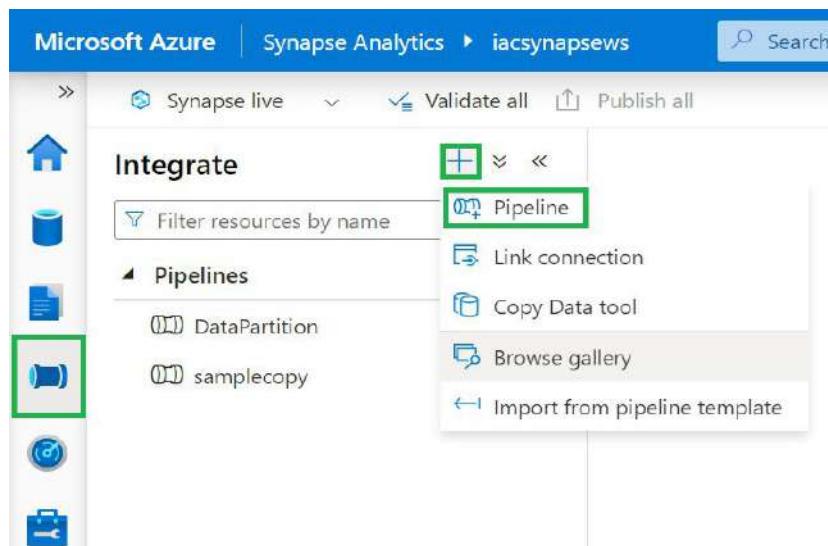


Figure 4.33 – Launching Synapse pipelines to perform transformations

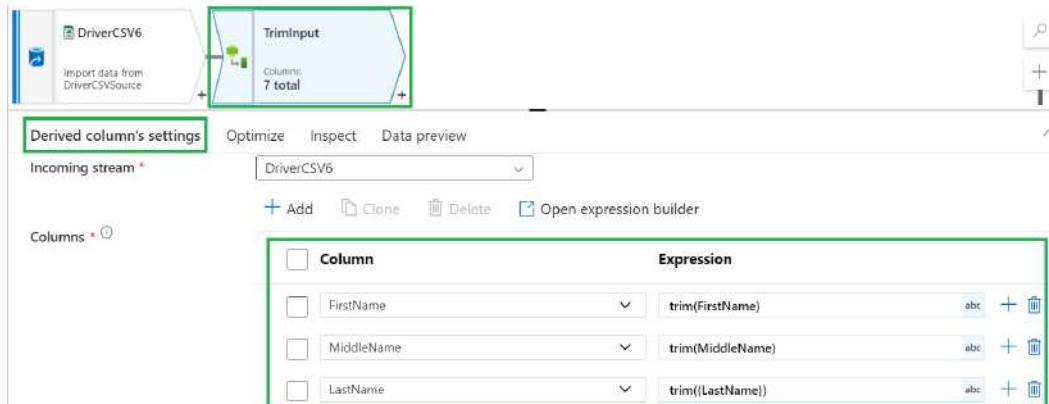


Figure 4.34 – Trimming whitespace in column values

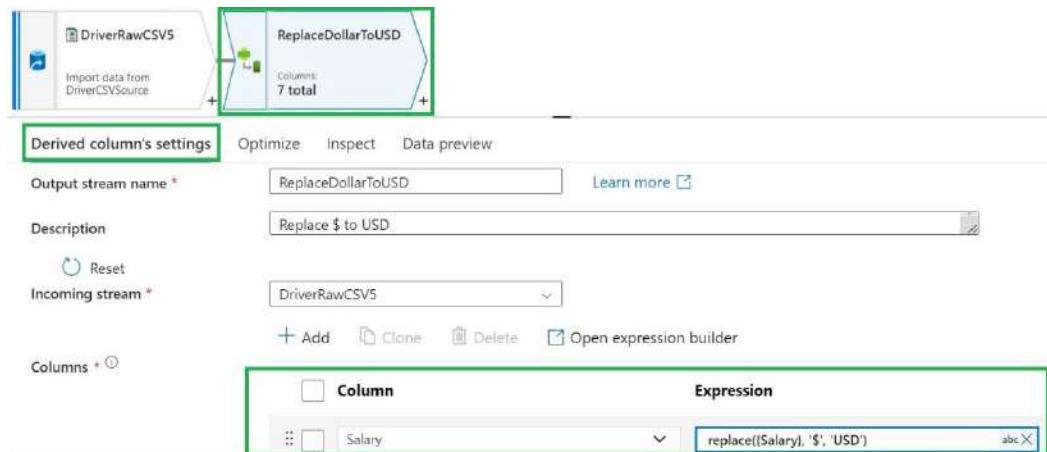


Figure 4.35 – Replacing \$ values in Salary column with USD

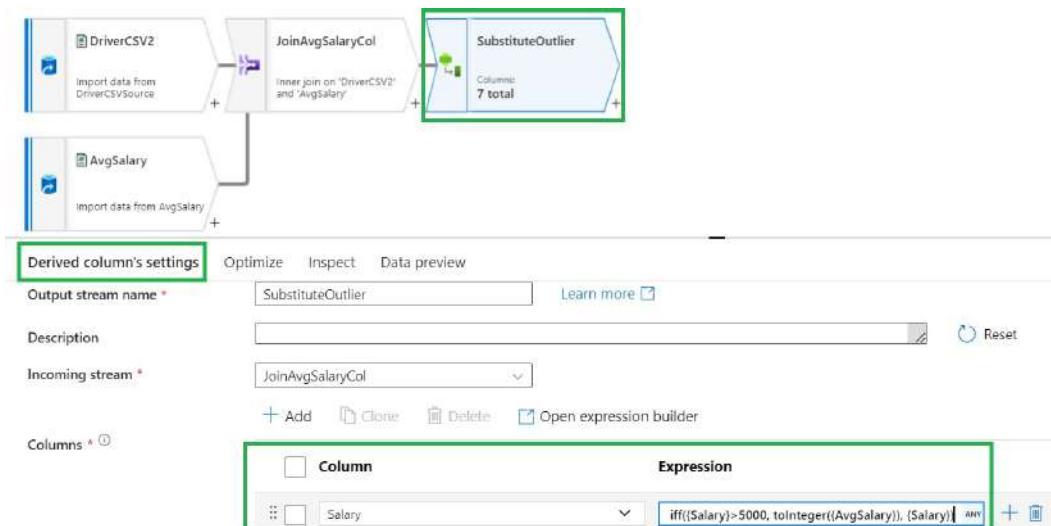


Figure 4.36 – Substituting value with AvgSalary

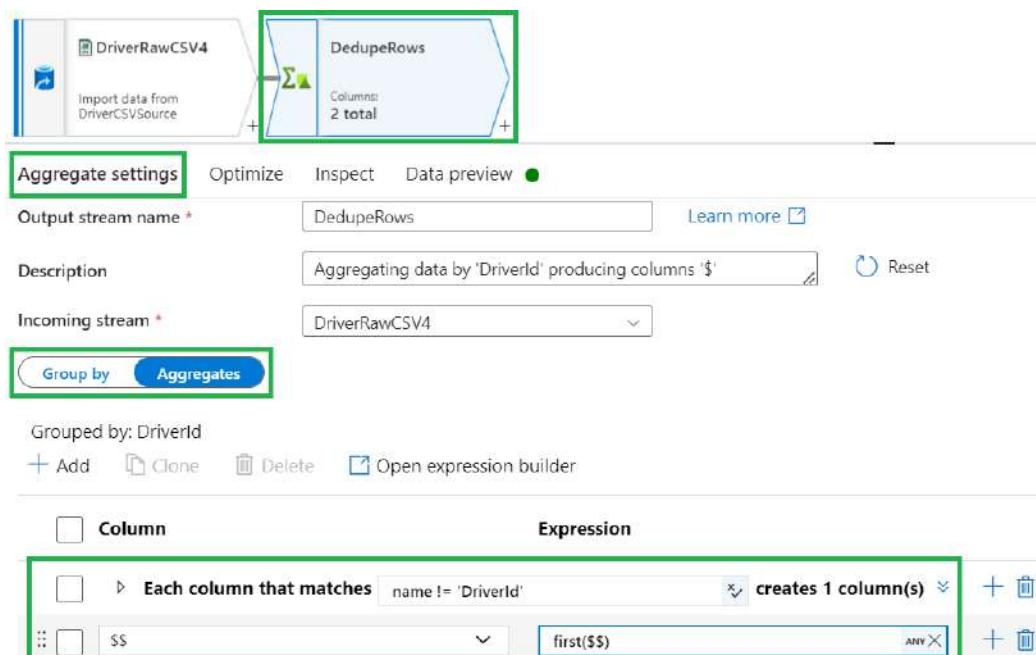


Figure 4.37 – Deduping using the Aggregate transformation

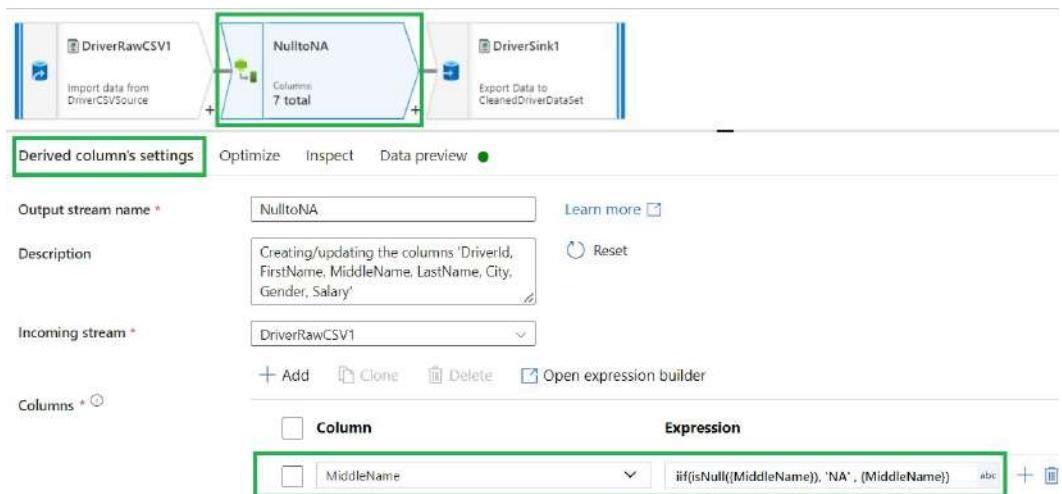


Figure 4.38 – Substituting missing values with default values

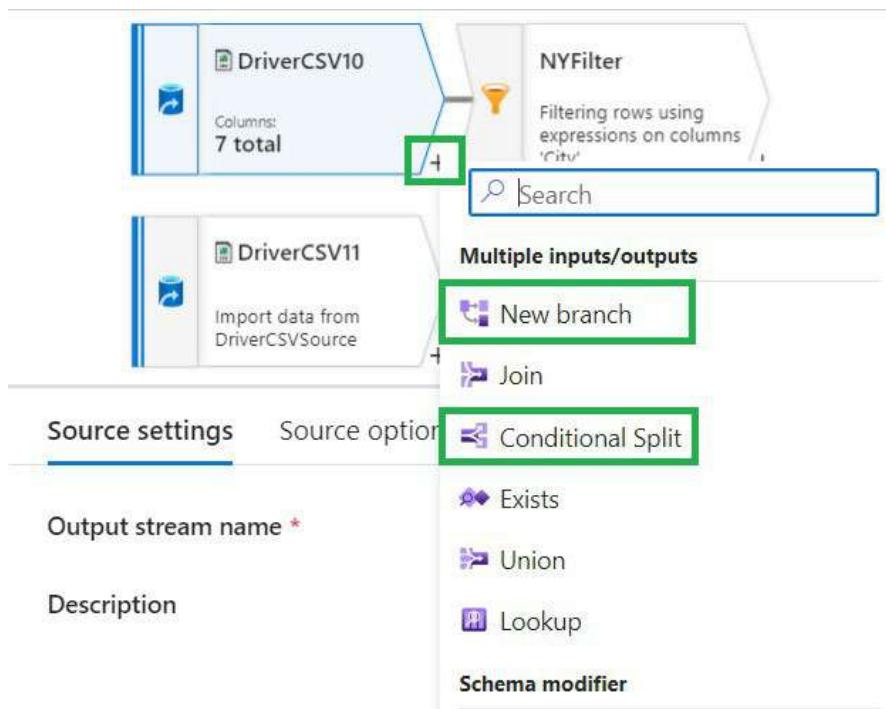


Figure 4.39 – Selecting New branch option to copy the entire dataset for a new execution flow

1.

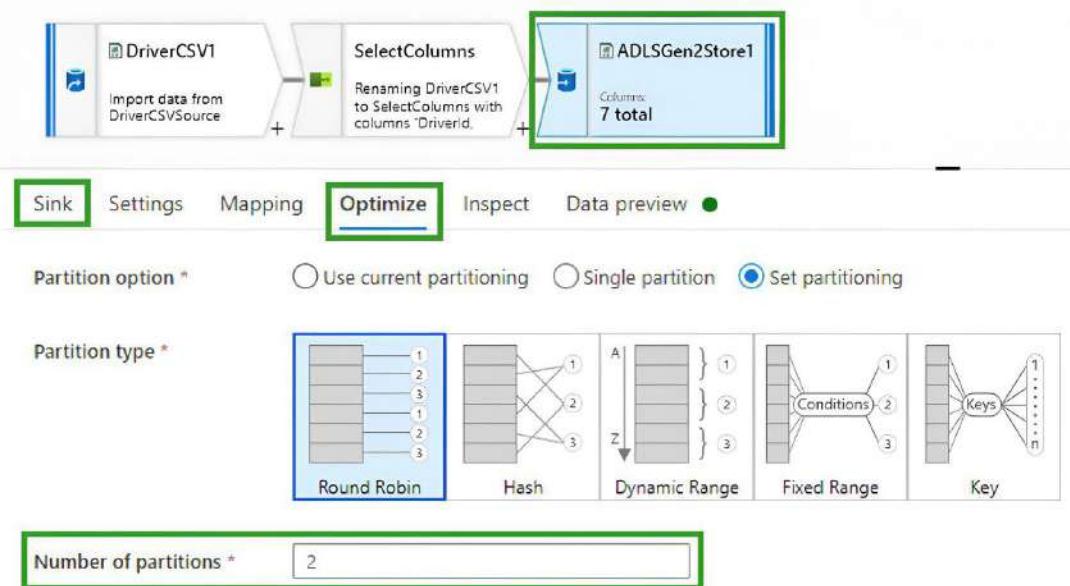


Figure 4.40 – Splitting files using ADF based on the number of partitions

```

root
|-- firstname: string (nullable = true)
|-- gender: string (nullable = true)
|-- id: integer (nullable = true)
|-- lastname: string (nullable = true)
|-- location: string (nullable = true)
|-- middlename: string (nullable = true)
|-- salary: integer (nullable = true)

+-----+-----+-----+-----+-----+
|firstname|gender| id|lastname| location|middlename|salary|
+-----+-----+-----+-----+-----+
|Catherine|Female|102|    NULL|California|   Goodwin|  4300|
|      Jenny|Female|104|  Simons|    Arizona|       Anne|  3400|
|      Bryan|  Male|101|Williams|    New York|          M|  4000|
|      Alice|Female|100|     Hood|    New York|        NULL|  4100|
|      Daryl|  Male|103|    Jones|    Florida|        NULL|  5500|
|      Daryl|  Male|103|    Jones|    Florida|        NULL|  5500|
+-----+-----+-----+-----+-----+

```

Figure 4.41 – The output of JSON operation

Results Messages

View Table **Chart** Export results ▾

Search

firstname	lastname	driverid	salary
Alice	Hood	100	4100
Bryan	Williams	101	4000
Daryl	Jones	103	5500
Daryl	Jones	103	5500
Jenny	Simons	104	3400
Catherine		102	4300

Figure 4.42 – Sample output of parsing JSON using OPENROWSET

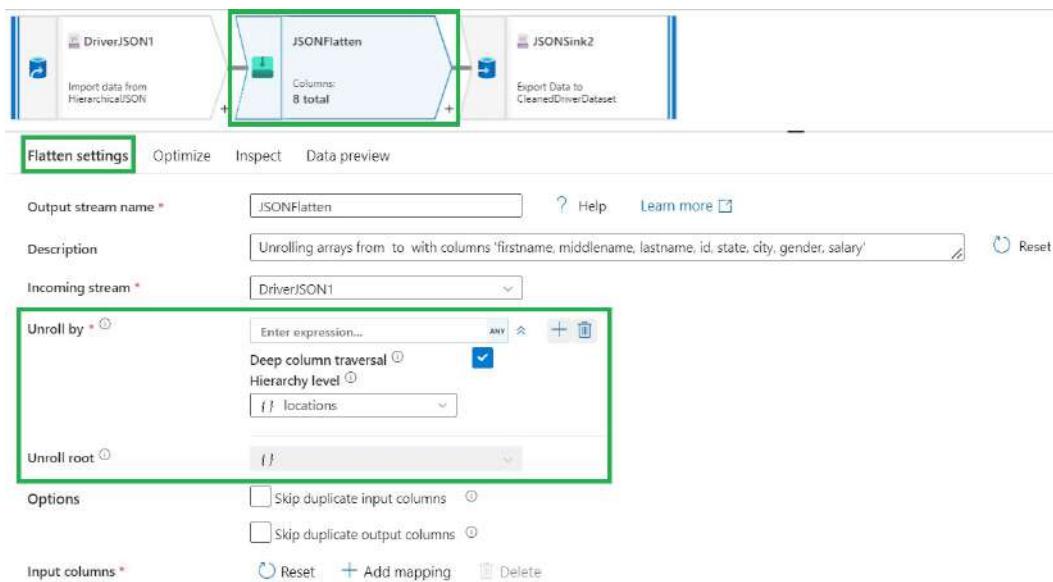


Figure 4.43 – Performing Flatten transformation in ADF

Connection Schema Parameters

Linked service * AzureDataLakeStorage Test connection Edit + New Learn more ↗

File path * users / rawedriver/sample/csv / driver.csv Browse

Compression type Select...

Column delimiter Comma (,)

Row delimiter Default (\r\n, or \n\r)

Encoding Default(UTF-8) **Default(UTF-8)**

Quote character Double quote (")

Escape character Backslash (\)

First row as header

Null value

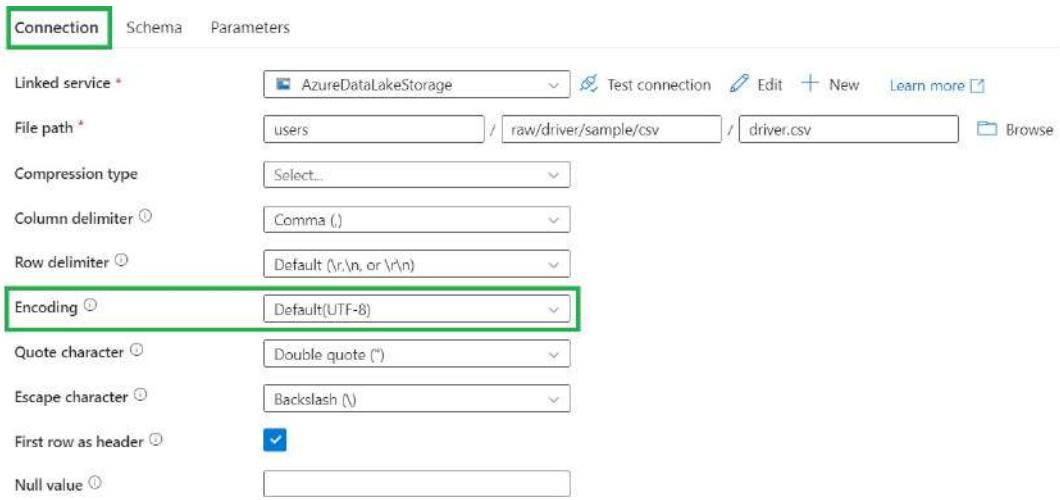


Figure 4.44 – Using the Encoding function in ADF source datasets

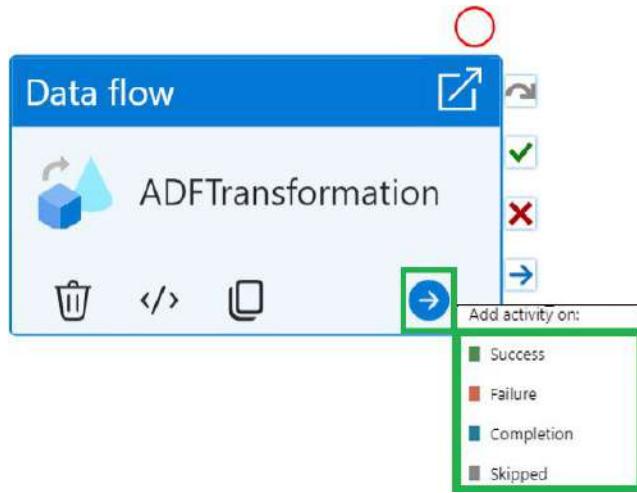


Figure 4.45 – ADF supporting four activity flows

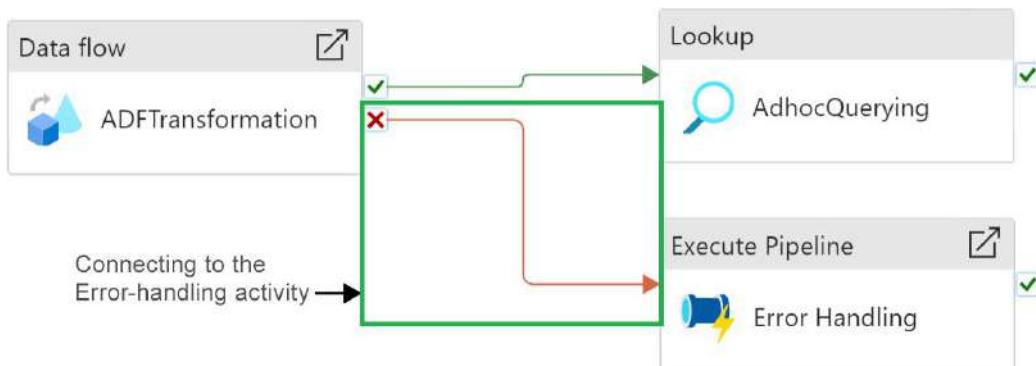


Figure 4.46 – Creating an error-handling pipeline

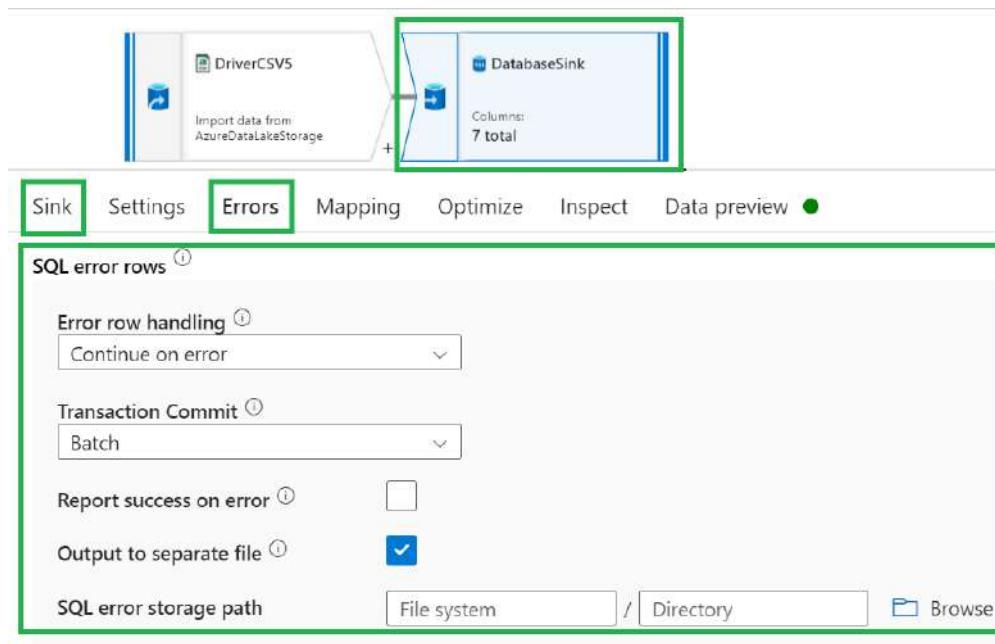


Figure 4.47 – Redirecting error lines to Blob Storage

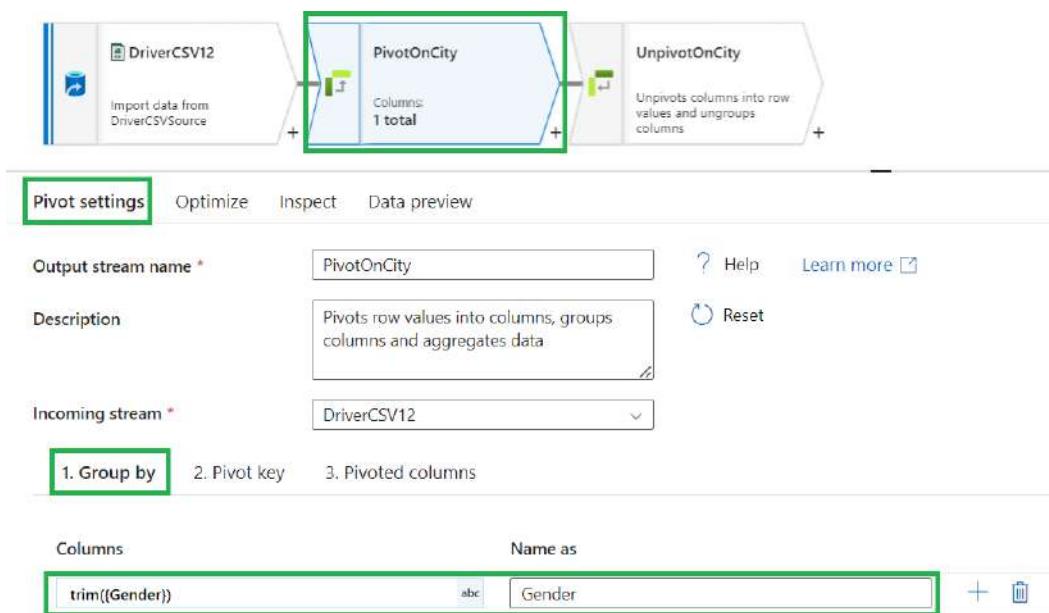


Figure 4.48 – Using Group by settings on Gender for the Pivot operation

1. Group by **2. Pivot key** 3. Pivoted columns

Pivot key * abc City

Value

Enter value (optional)... + -

Null value

Figure 4.49 – Using City for the Pivot operation

Pivot settings Optimize Inspect Data preview

Output stream name * PivotOnCity ? Help Learn more ↗

Description Pivots row values into columns, groups columns and aggregates data ⚡ Reset

Incoming stream * DriverCSV12

1. Group by 2. Pivot key **3. Pivoted columns**

Column name pattern * prefix{expression prefix}middle[Pivot key value]suffix

Prefix Middle Suffix

Column arrangement * Normal Lateral

avg((Salary)) 1.2 Avg + -

Figure 4.50 – Specifying the column name pattern and aggregation for the Pivot function

The screenshot shows a Data preview tab with the following data:

	Gender	AvgArizona	AvgCalifornia	AvgFlorida	AvgNew York
Female	3400.0	4300.0	NULL	4100.0	
Male	NULL	5500.0	4000.0	NULL	

Figure 4.51 – The post-Pivot table showing values

The screenshot shows the Unpivot settings tab with the following configuration:

- Output stream name ***: UnpivotOnCity
- Description**: Unpivots columns into row values and ungroups columns
- Incoming stream ***: PivotOnCity
- 1. Ungroup by** (selected tab)
- 2. Unpivot key**
- 3. Unpivoted columns**

Columns section (highlighted with a green box):

abc	Gender	+ (add)	trash
-----	--------	---------	-------

Figure 4.52 – Using the Ungroup by tab for Unpivoting on a single column Gender

The screenshot shows the 2. Unpivot key tab with the following configuration:

- Unpivot column name ***: Cities
- Unpivot column type ***: 12 double
- Option ***: Pick column names as values Enter values

Figure 4.53 – Specifying Unpivot key for the Unpivot operation

The screenshot shows the 3. Unpivoted columns tab with the following configuration:

- Column arrangement ***: Normal (selected)
- Drop rows with null**:
- Columns ***:

Column name	Type
AvgSalary	12 double

Figure 4.54 – Using the Unpivoted columns tab to calculate AvgSalary by Gender and Cities

Validate Data flow debug Debug Settings

Unpivot settings Optimize Inspect **Data preview**

Number of rows + INSERT 5 * UPDATE 0 × DELETE 0

Refresh | Typecast | Modify | Map drifted | Statistics | Remove

↓	Gender	Cities	AvgSalary
+	Female	AvgArizona	3400.0
+	Female	AvgCalifornia	4300.0
+	Female	AvgNew York	4100.0
+	Male	AvgFlorida	5500.0
+	Male	AvgNew York	4000.0

Figure 4.55 – Displaying an unpivoted table as output

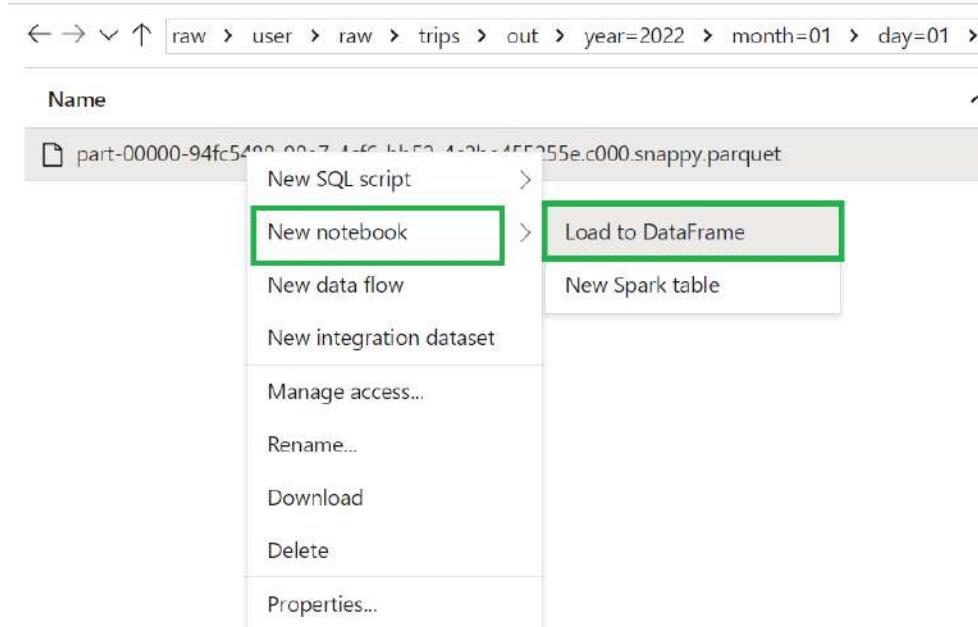


Figure 4.56 – Launching a DataFrame from the Synapse data file

```
%pyspark
df = spark.read.load('abfss://users@synaps.dfs.core.windows.net/user/raw/trips/out/year=2022/month=01/day=01/tripDate=20220101/part-00000-94fc5488-98a7-4cf6-bb52-4c2be455255e.c000.snappy.parquet',
format='parquet')
display(df.limit(10))
```

[23] ✓ 4 sec - Command executed in 3 sec 943 ms by surendramettapalli on 11:32:09 AM, 4/13/24

> Job execution Succeeded Spark 2 executors 8 cores View in monitoring Open Spark UI

View	Table	Chart	Export results
tripId	driverId	customerId	startLocation
100	200	300	New York

Figure 4.57 – Loading data to a DataFrame to explore using Spark

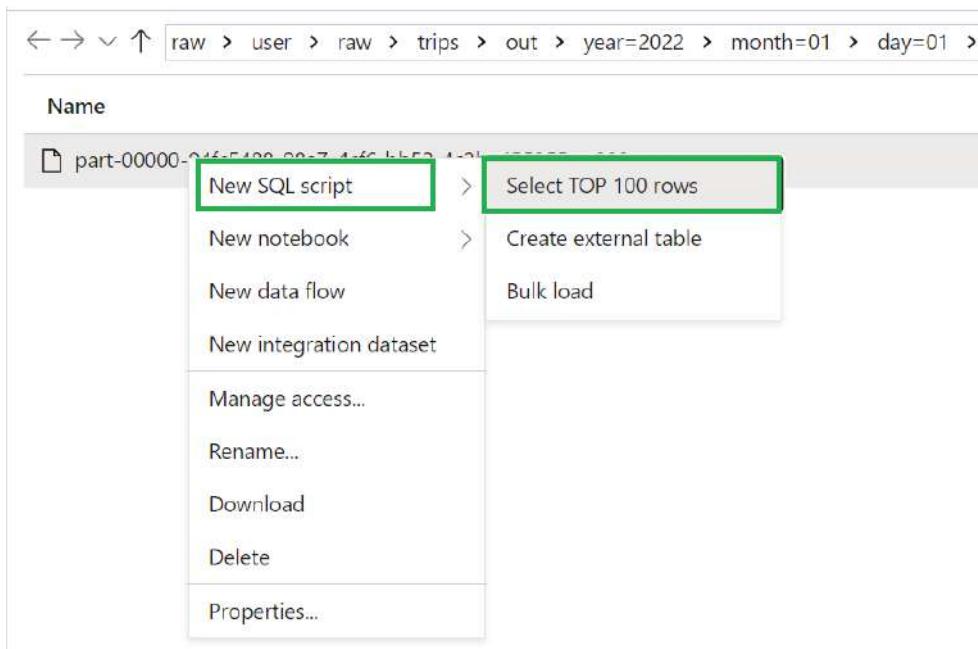


Figure 4.58 – Launching a SQL script to explore data from the Synapse data file

```

Run Undo | Publish Query plan Connect to Built-in Use database master
1 -- This is auto-generated code
2 SELECT
3   TOP 100 *
4 FROM
5   OPENROWSET(
6     BULK 'https://synapseuser.dfs.core.windows.net/raw/user/raw/trips/out/year=2022/month=01/
7     day=01/tripDate=20220101/part-00000-94fc5488-98a7-4cf6-bb52-4c2be455255e.c000.snappy.parquet',
8     FORMAT = 'PARQUET'
9   ) AS [result]
  
```

The screenshot shows the Azure Data Studio interface with the following details:
 - Top bar: Run, Undo, Publish, Query plan, Connect to (set to Built-in), Use database (set to master).
 - Script pane: An auto-generated SQL script to select top 100 rows from a Parquet file located at the specified URL.
 - Results pane: Shows a table with columns: tripId, driverId, customerId, startLocation, endLocation. One row is displayed: tripId 100, driverId 200, customerId 300, startLocation New York, endLocation New Jersey.

Figure 4.59 – SQL script auto-launched by Synapse to help explore data

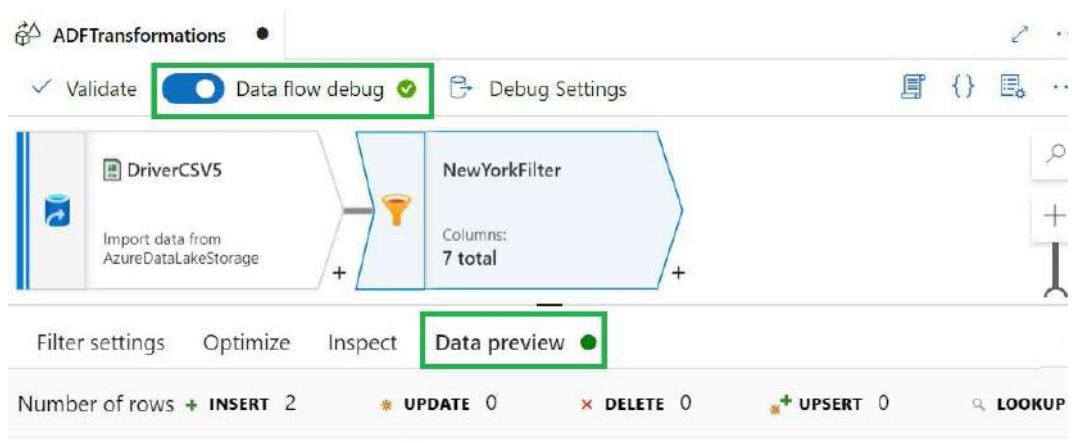


Figure 4.60 – ADF Data preview tab showing fine-tuned transformations and pipelines

The screenshot shows the 'Practice Resources' dashboard. At the top, there's a navigation bar with the 'Practice Resources' logo, a bell icon for notifications, and a 'SHARE FEEDBACK' button. Below the navigation, the path 'DASHBOARD > CHAPTER 4' is visible. The main content area is titled 'Data Processing' and has a 'Summary' section. The summary text reads: 'With that, you have come to the end of this interesting chapter. There were lots of examples and screenshots to help you learn the concepts. It might be overwhelming at times, but the easiest way to follow is to open a live Spark, SQL, or ADF session and try to execute the examples in parallel.' Another paragraph states: 'You covered a lot of details in this chapter, such as performing transformations in Spark, SQL, and ADF, data cleansing techniques, reading, and parsing JSON data, encoding and decoding, error handling during transformations, normalizing and denormalizing datasets, and finally, a bunch of data exploration techniques. This is one of the important chapters in the syllabus. You should now be able to comfortably build data pipelines with transformations involving Spark, SQL, and ADF.' A final note says: 'In the upcoming chapter, you will create resilient batch-processing solutions leveraging Azure's analytics services including Data Lake Storage, Databricks, Synapse Analytics, and Data Factory.' To the right, there's a 'Chapter Review Questions' section. It includes the book title 'The Azure Data Engineer Associate Certification Guide - Second Edition by Giacinto Palmieri, Surendra Mettapalli, Newton Alex', a 'Select Quiz' button, and a 'Quiz 1' section with a 'SHOW QUIZ DETAILS' dropdown and a 'START' button.

Figure 4.62 – Chapter Review Questions for Chapter 4

Chapter 5: Developing Batch-Processing Solutions

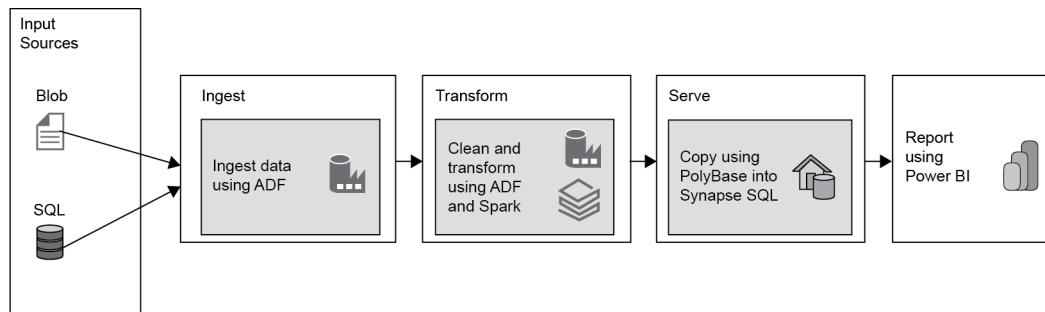


Figure 5.1 – High-level architecture of the batch use case

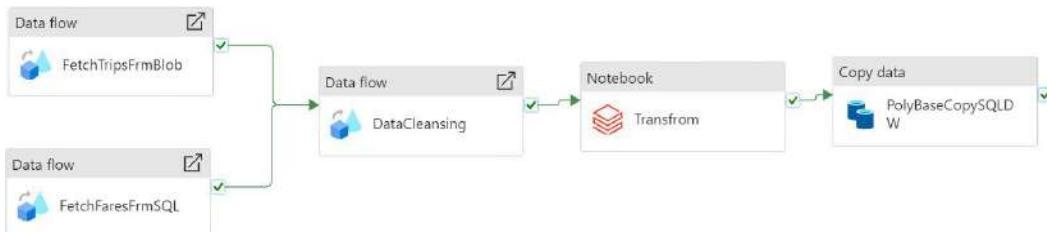


Figure 5.2 – Sample data pipeline in ADF

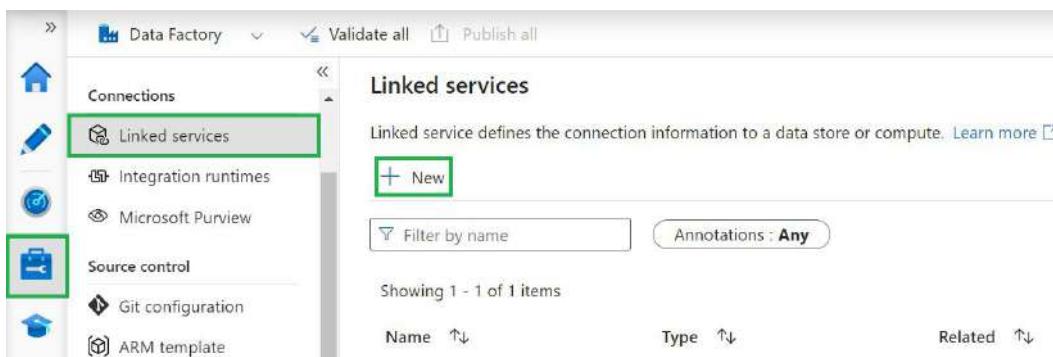


Figure 5.3 – Configuring a linked service in ADF

New linked service (Azure Blob Storage)

Name *

Description

Connect via integration runtime *

Authentication method

Connection string Azure Key Vault

Account selection method From Azure subscription Enter manually

Azure subscription

Storage account name *

Additional connection properties

+ New

Test connection To linked service To file path

Annotations

+ New

Parameters

Advanced

Figure 5.4 – Creating an Azure Blob Storage linked service

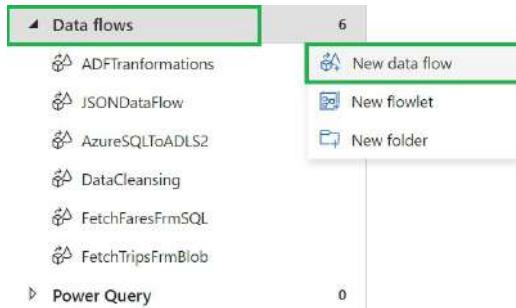


Figure 5.5 – Creating a new data flow in ADF

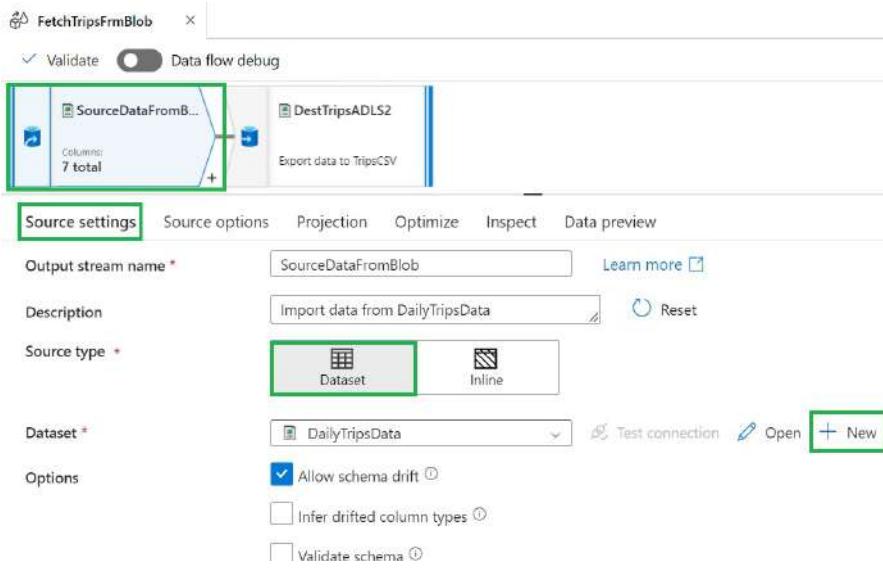


Figure 5.6 – Creating source and destination datasets in a data flow

The screenshot shows the 'CSV' dataset creation screen. The 'DelimitedText' tab is selected. The 'File path' field contains 'dailytrips' and 'File name' is empty. Other fields include 'Compression type' (Select...), 'Column delimiter' (Comma (,), selected), 'Row delimiter' (Default (\r\n, or \n)), 'Encoding' (Default(UTF-8)), and 'Quote character' (Double quote (")).

Figure 5.7 – Sample dataset screen to create a dataset named DailyTripData

Microsoft Azure

Home > Create a resource > Marketplace > Azure Databricks >

Create an Azure Databricks workspace

Project Details

Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources.

Subscription *

Resource group *

Instance Details

Workspace name *

Region *

Pricing Tier *

Figure 5.8 – Sample Azure Databricks workspace creation screen

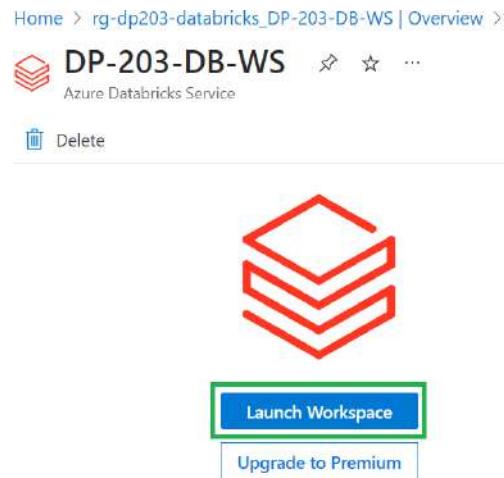


Figure 5.9 – Launching the Azure Databricks workspace

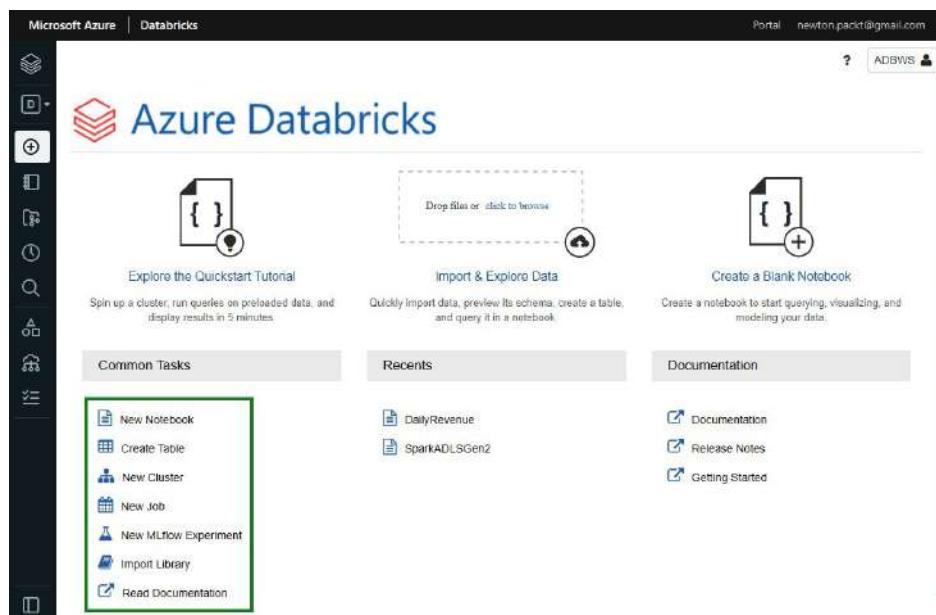


Figure 5.10 – Azure Databricks portal

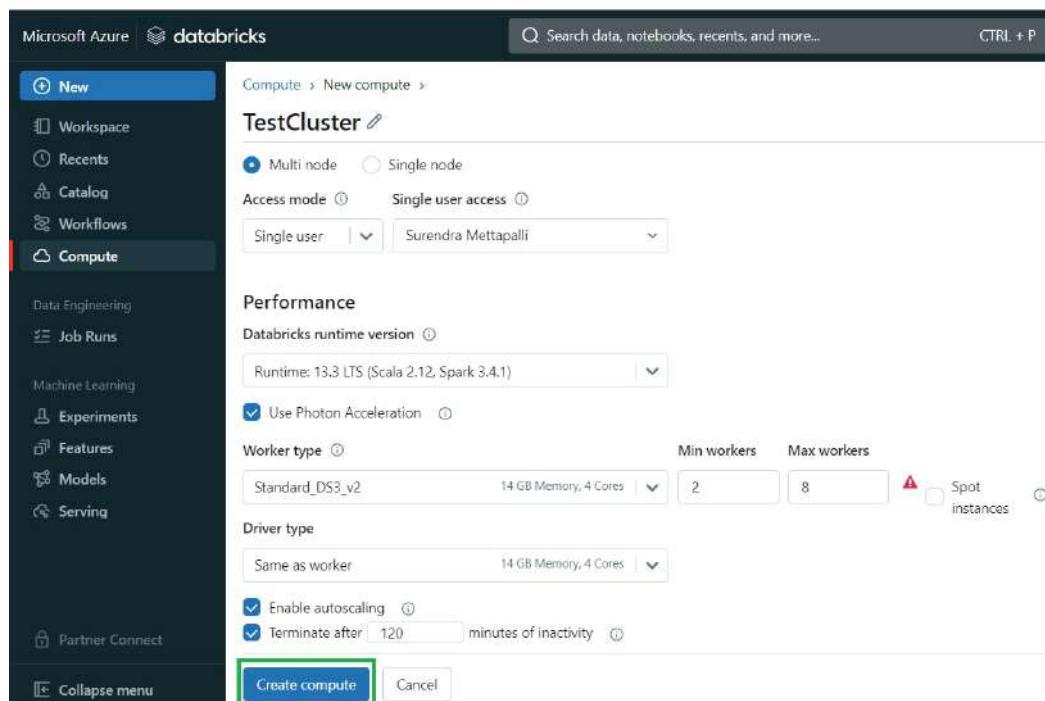


Figure 5.11 – Azure Databricks cluster creation screen

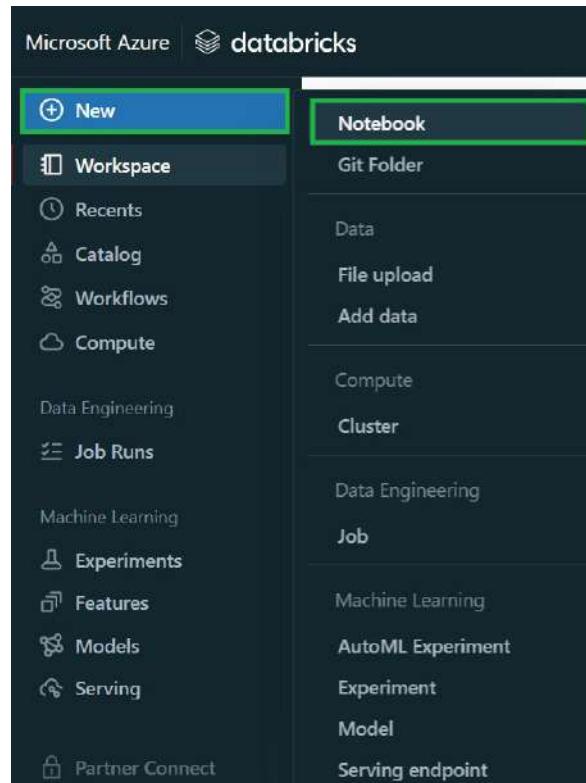


Figure 5.12 – Azure Databricks new notebook creation

A screenshot of an Azure Databricks notebook titled "SampleNotebook". The notebook is set to Scala. The cell contains the following code:

```
print("Hello world")
```

The output of the cell is "Hello world". The notebook interface includes a toolbar with File, Edit, View, Run, Help, Last edit was now, New cell UI: ON, Run all, DP-203 Cluster, and Schedule buttons.

Figure 5.13 – Azure Databricks notebook with Spark code inserted within the Cmd block

A screenshot of the Azure Data Factory (ADF) interface. On the left, there's a sidebar titled "Factory Resources" with sections for Pipelines, Change Data Capture (preview), Datasets, Data flows, and Power Query. In the center, there's a "Batch Pipeline" section with a "Activities" pane. The activities pane shows a search bar with "databricks" and a list of activities: Notebook, Jar, and Python. The "Notebook" activity is highlighted with a green border. At the top of the activities pane, there are buttons for Validate, Debug, and Add trigger. To the right of the activities pane, there's a preview window showing a "Notebook" card with "Notebook1" listed, along with standard card controls like delete, copy, and refresh.

Figure 5.14 – Choosing an Azure Databricks activity in ADF

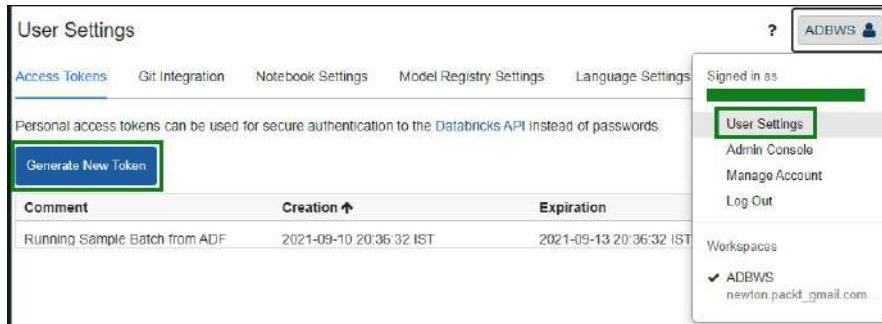


Figure 5.15 – Selecting the access token from the Azure Databricks portal

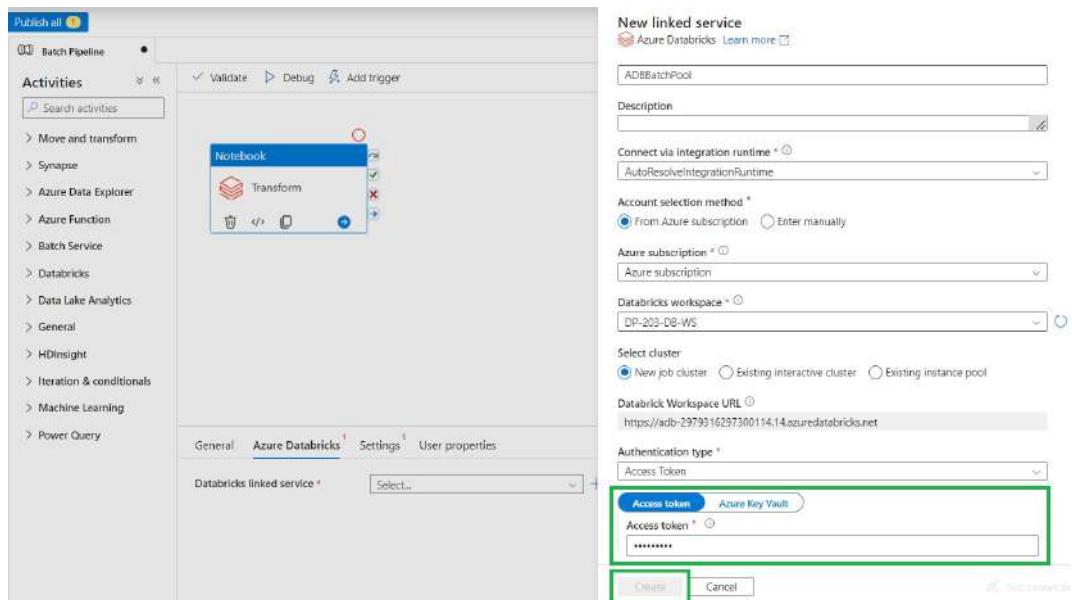


Figure 5.16 – Creating an Azure Databricks linked service

New linked service

Azure SQL Database [Learn more](#)

Connect via integration runtime * ⓘ

AutoResolveIntegrationRuntime ▼ 

Connection string Azure Key Vault

Account selection method ⓘ

From Azure subscription Enter manually

Azure subscription

Azure subscription (2090a2ce-6634-454f-ab7c-5bca1c4dfe27) ▼

Server name *

synapse-az-ws ↻

Database name *

master ↻

Authentication type *

SQL authentication ▼

User name *

admin

Password Azure Key Vault

Password *

Create Back  Test connection Cancel

Figure 5.17 – Creating a linked service for Synapse Link

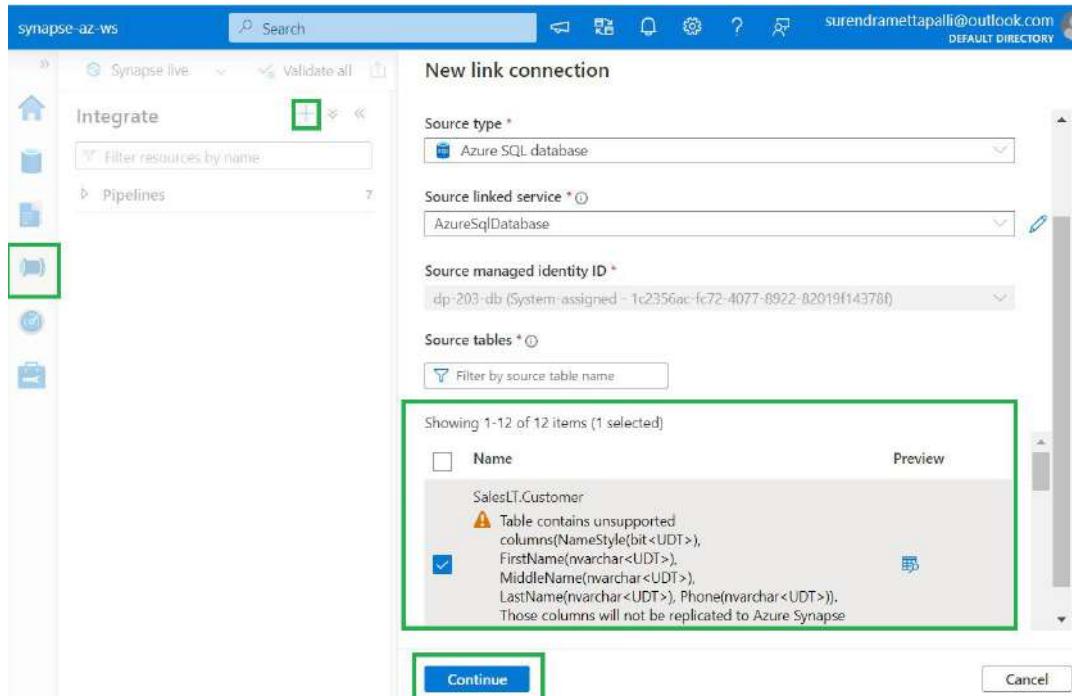


Figure 5.18 – Data replication with Synapse Link

Figure 5.19 – Monitoring replication with Synapse Link

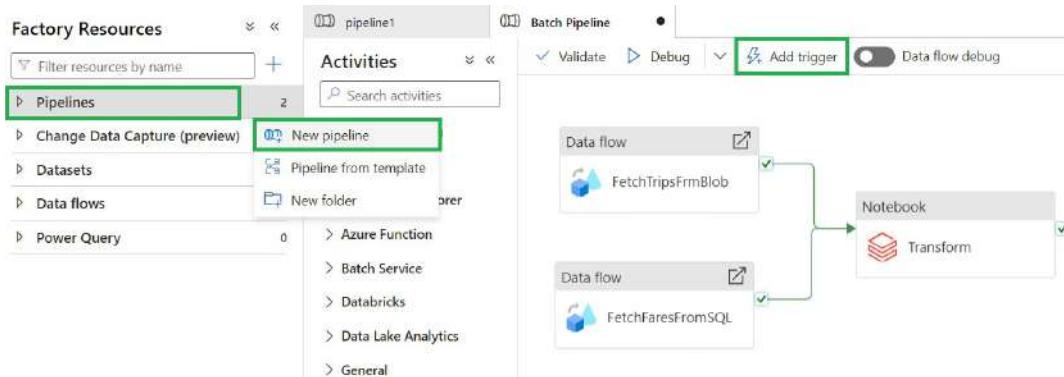


Figure 5.20 – Creating new pipelines in ADF

Figure 5.21 – Specifying data integration units for the Copy activity

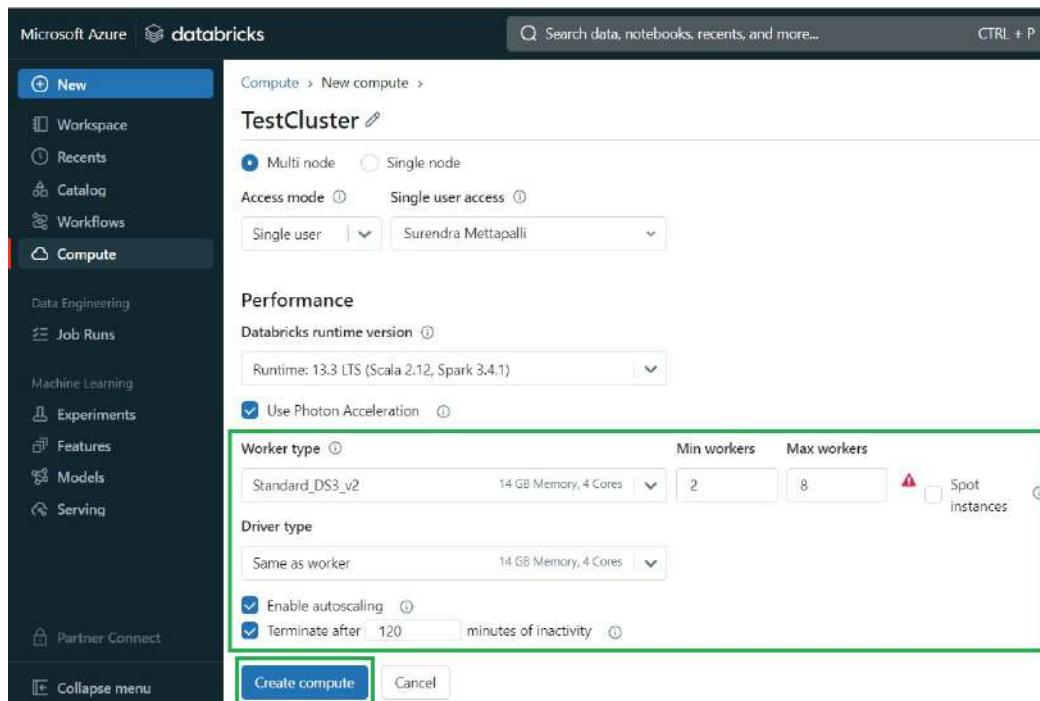


Figure 5.22 – Enabling the Azure Databricks cluster autoscale option

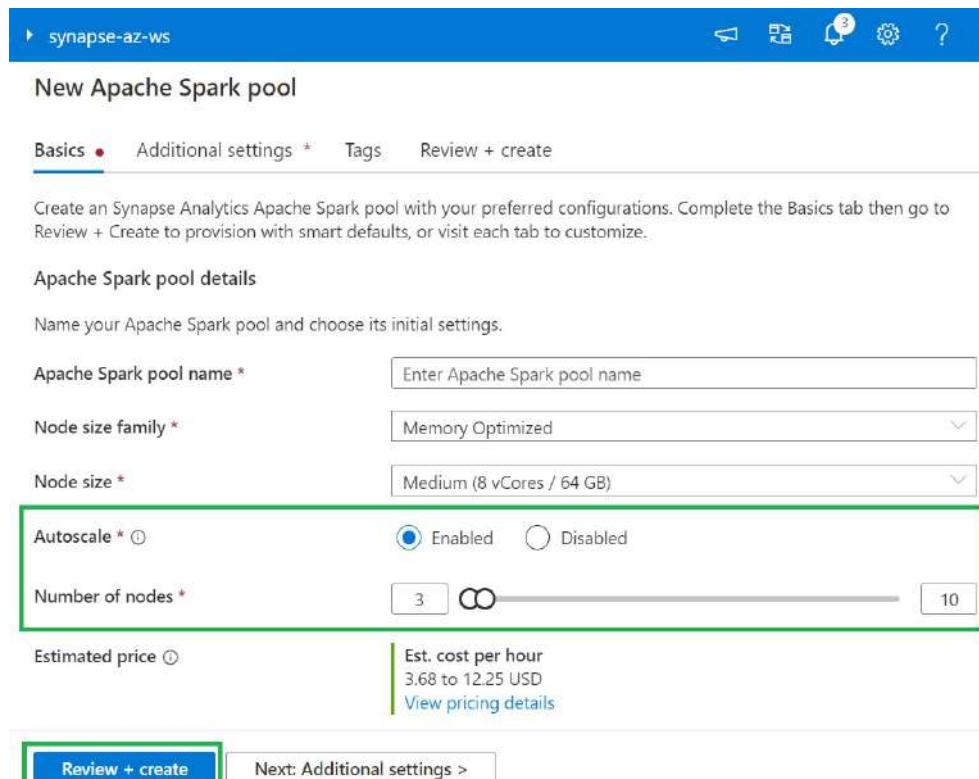


Figure 5.23 – Selecting the Synapse Spark Autoscale option for autoscaling

The serverless SQL pool, Built-in, is immediately available for your workspace. Dedicated SQL pools can be configured and constrained. [Learn more](#)

+ New Refresh

Filter by name

Showing 1-2 of 2 items (1 Serverless, 1 Dedicated)

Name	Type	Status
Built-in	Serverless	Online
MySampleDataWarehouse	Dedicated	Online

...
Pause
Scale
Assign tags
Delete

Figure 5.24 – Selecting the Synapse SQL Scale option

Scale

mySampleDataWarehouse

Scaling can impact workload management settings. Consider using the [workload management scale experience](#) in the Azure portal to configure the settings that best align to your workload needs. [Learn more about performance levels](#)

Performance level



DW200c

Estimated price ⓘ

Est. cost per hour
3.32 USD

[View pricing details](#)

Apply

Cancel

Figure 5.25 – Synapse SQL Scale option to scale out for performance and scale in to save costs

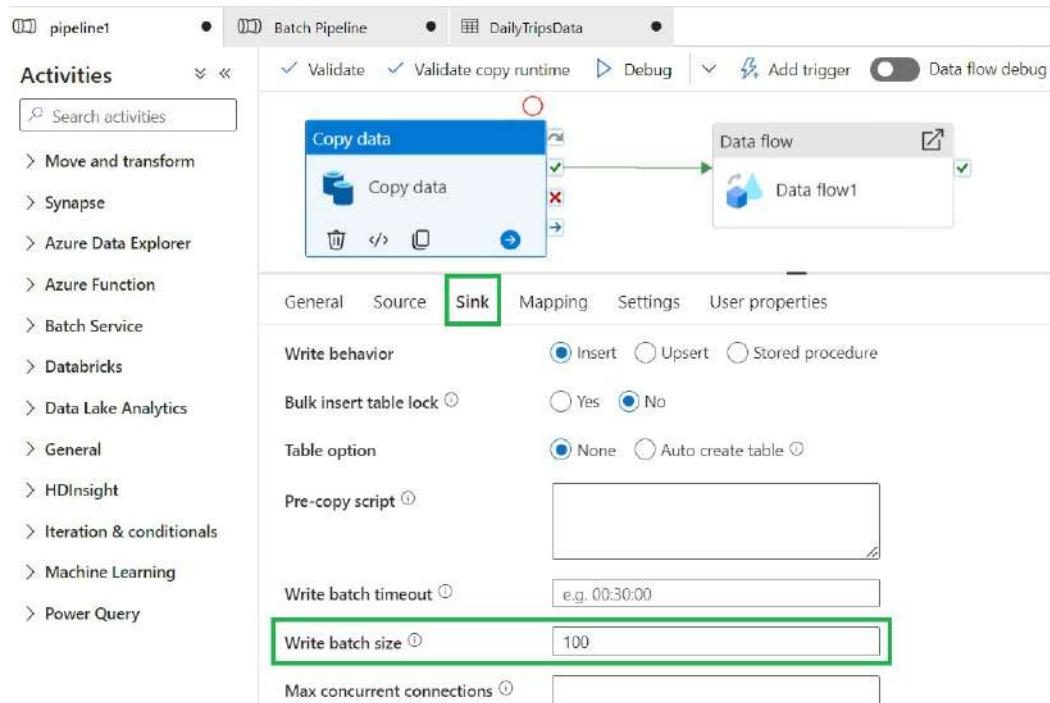


Figure 5.26 – Specifying the batch size for the Copy activity

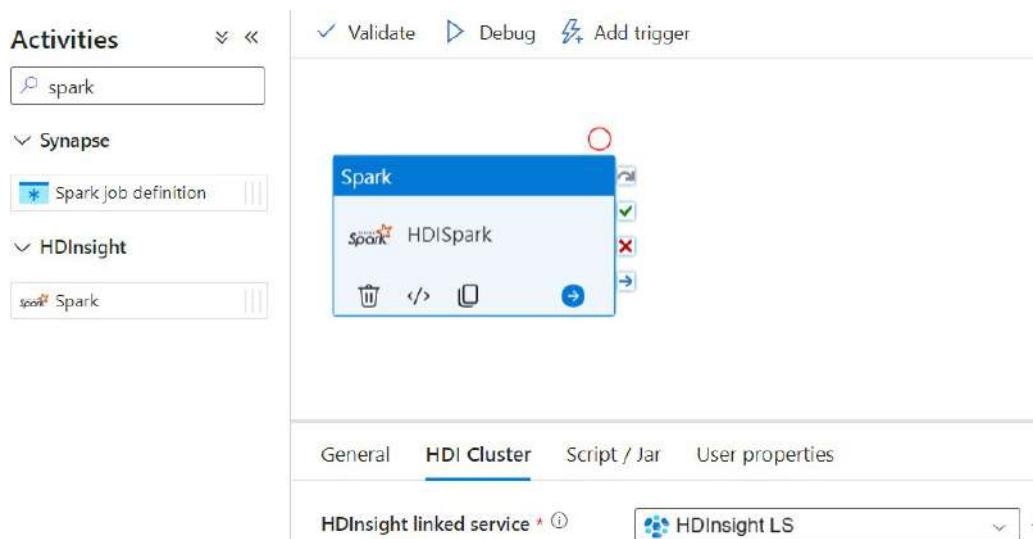


Figure 5.27 – Configuring a Spark activity in ADF

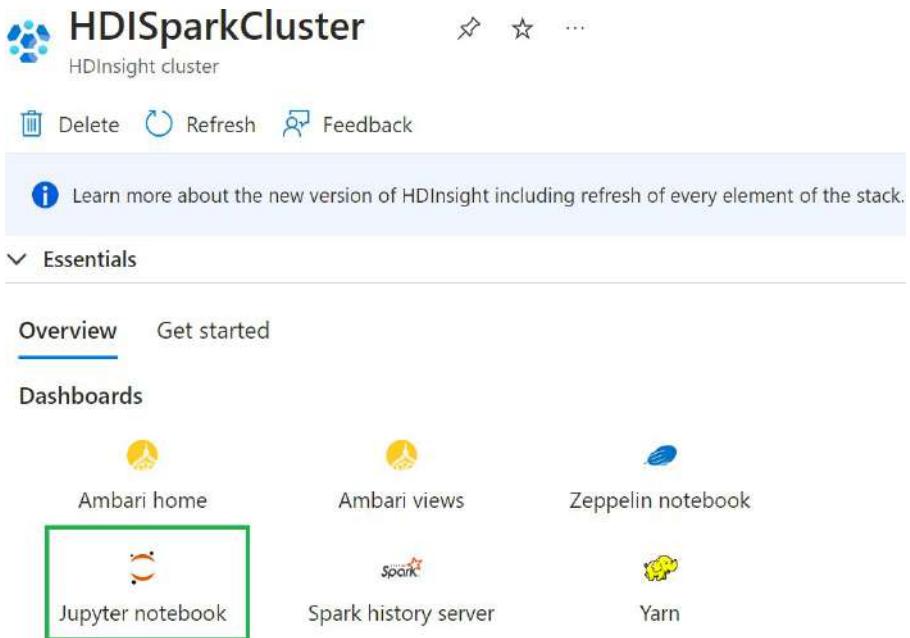


Figure 5.28 – Launching the Jupyter notebook from HDInsight



Figure 5.29 – Launching a PySpark Jupyter notebook from HDInsight

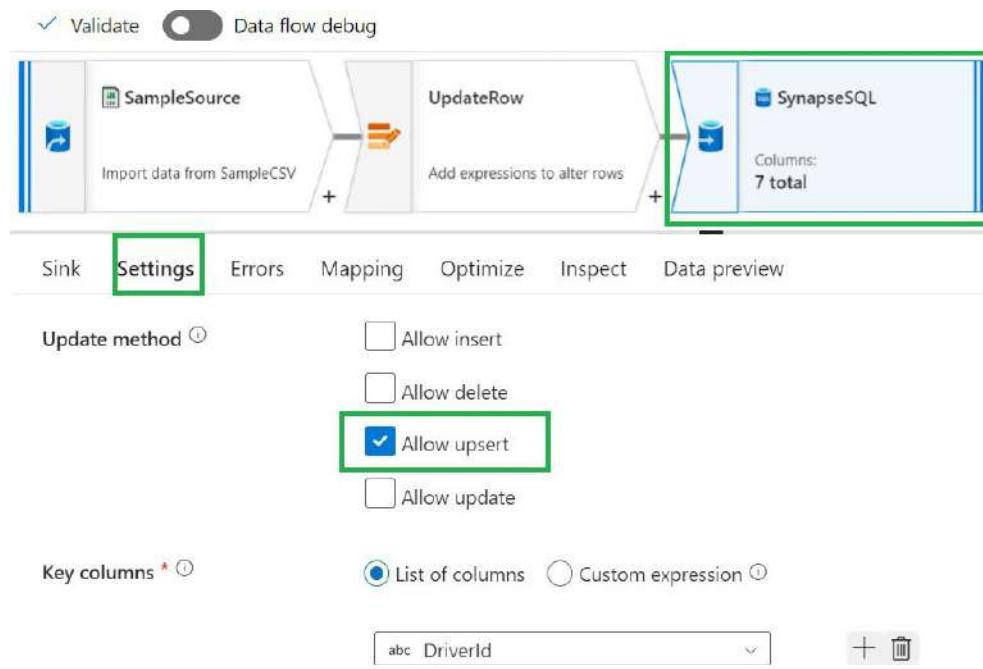


Figure 5.30 – Upsert operation in ADF used for updating and inserting data

General Source Sink Mapping Settings User properties

Settings You will be charged # of used DIUs * copy duration * \$0.25/DIU-hour. Local currency

Data integration unit Auto Edit

Degree of copy parallelism Edit

Data consistency verification

Fault tolerance Select all Skip incompatible rows Skip missing files Skip forbidden files Skip files with invalid names

Enable logging Logging settings

Storage connection name *

Enable staging

Figure 5.31 – Enabling consistency verification and fault tolerance in an ADF Copy activity

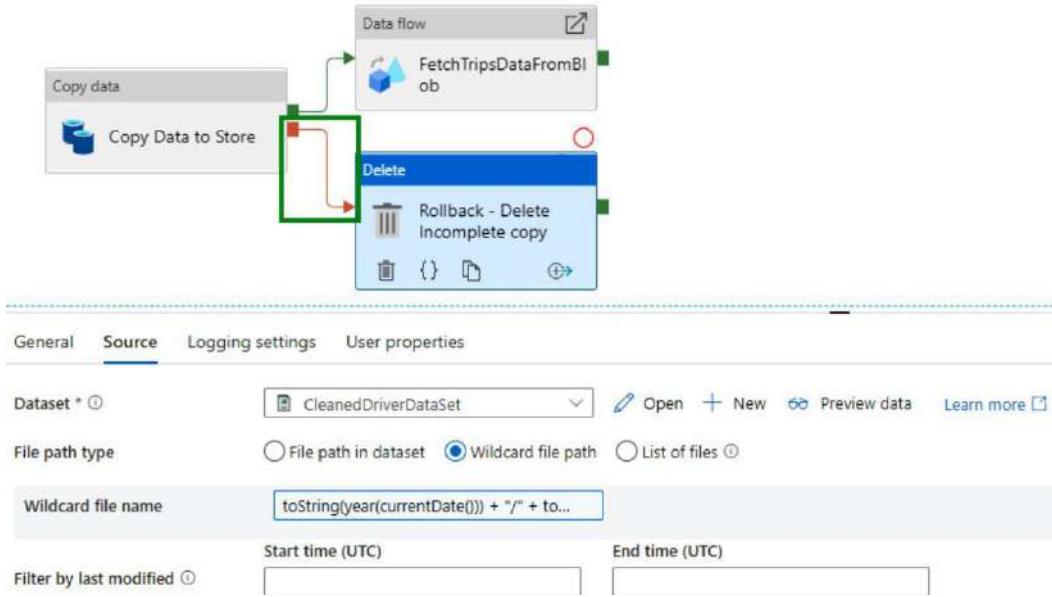


Figure 5.32 – Deleting an incomplete Copy activity

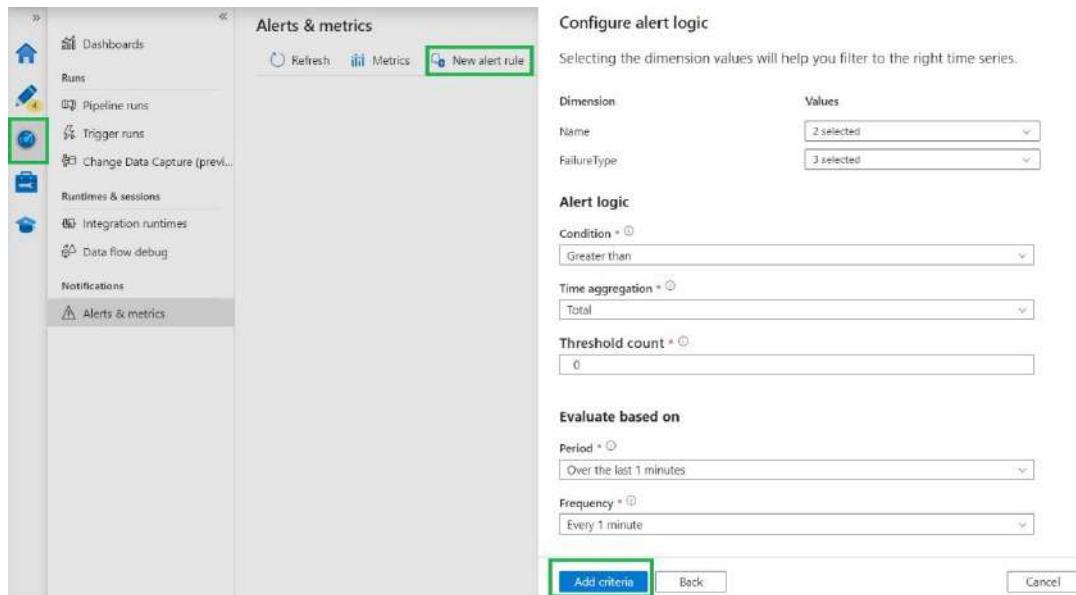


Figure 5.33 – Configuring an alert for pipeline failure

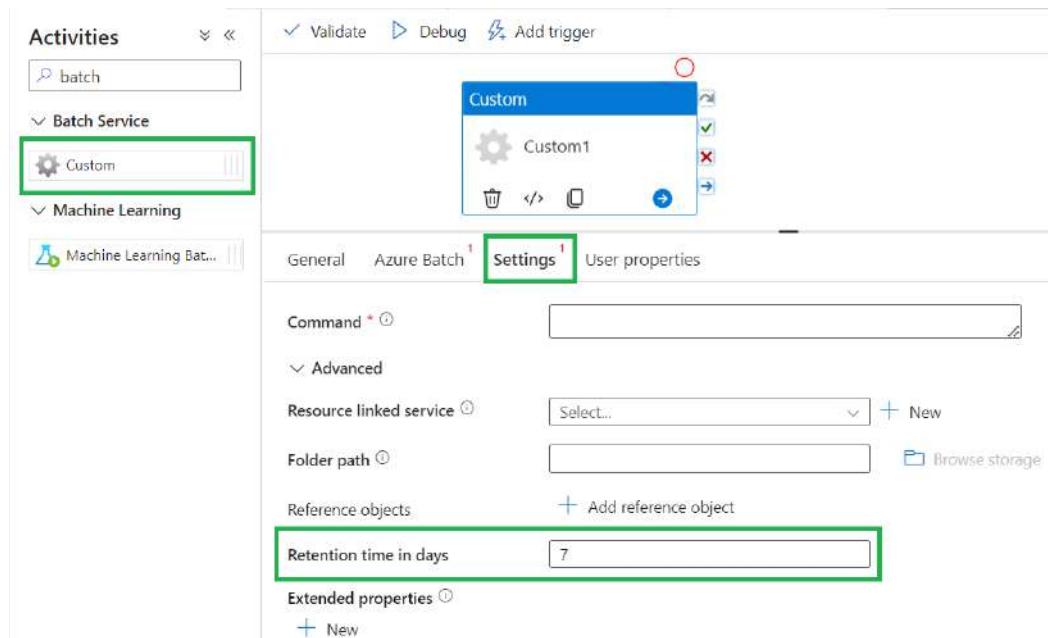


Figure 5.34 – Configuring batch retention for an Azure Batch activity in a pipeline

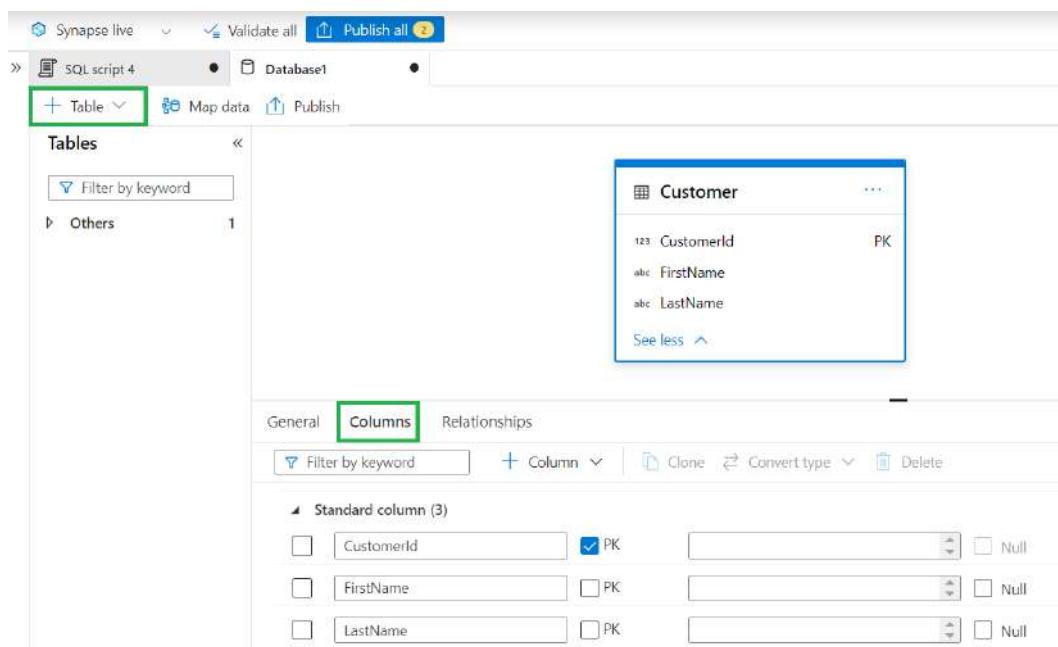


Figure 5.35 – Creating a custom delta table in Synapse

The screenshot shows the Azure Synapse Studio interface. On the left, there's a sidebar with icons for Home, Data, Workspace, and Linked. Under 'Data', 'Lake database' is selected, showing a tree view of tables: Airlines_b4n (Channel, Communication, Customer, Document, Employee, Item, Order, Product). In the main area, a SQL script named 'SQL script 4' is displayed with the following code:

```

1  SELECT TOP (100)
2    locationID
3    ,Borough
4    ,Zone
5    ,service_zone
6   FROM [dbo].[Customer]

```

Below the script, the 'Results' tab is active, showing a table with four columns: LocationID, Borough, Zone, and service_zone. The data is as follows:

LocationID	Borough	Zone	service_zone
1	EWR	Newark Airport	EWR
2	Queens	Jamaica Bay	Boro Zone
3	Bronx	Allerton/Pelha...	Boro Zone

Figure 5.36 – Querying a delta table in Synapse

The screenshot shows the 'Practice Resources' page for Chapter 5. At the top, there's a navigation bar with 'Practice Resources', a bell icon, and 'SHARE FEEDBACK'. Below it, a breadcrumb navigation shows 'DASHBOARD > CHAPTER 5'. The main content area is titled 'Developing Batch-Processing Solutions' and includes a 'Summary' section and a 'Chapter Review Questions' section.

Summary

With this chapter, you have gained a profound knowledge of crafting batch-processing solutions within Azure's ecosystem. You've mastered data integration with PolyBase, queried replicated data using Synapse Link, and orchestrated tasks through data pipelines. Now, you can seamlessly scale resources, optimize batch sizes, and ensure pipeline reliability with testing. Additionally, Python notebooks empower you with advanced data processing, upserting, and data reversion for agile data management. Finally, your expertise in exception handling, retention policies, and Delta Lakes ensures efficient handling of big data.

The next chapter will focus on how you can design and develop a stream-processing solution. You will explore the utilization of Spark structured streaming to efficiently handle data streams. Furthermore, you will gain insights into processing time-series data, managing data across partitions, and optimizing processing within a single partition. Scalability is another focus, where you will learn to scale resources.

Chapter Review Questions

The Azure Data Engineer Associate Certification Guide
- Second Edition by Giacinto Palmieri, Surendra Mettapalli, Newton Alex

Select Quiz

Quiz 1 START

SHOW QUIZ DETAILS ▾

Figure 5.38 – Chapter Review Questions for Chapter 5

Chapter 6: Developing a Stream Processing Solution

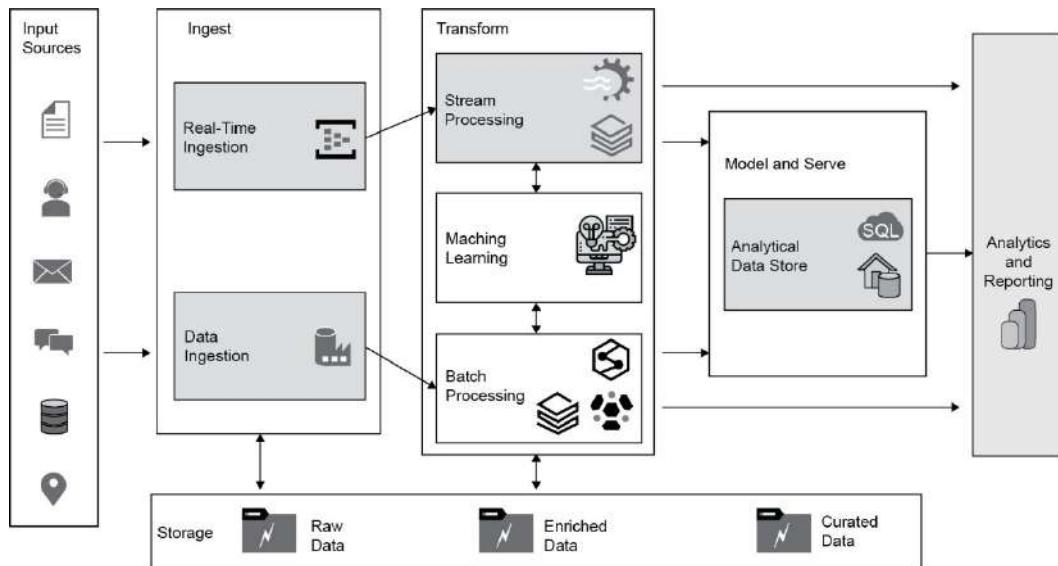


Figure 6.1 – Stream processing architecture with several key stages

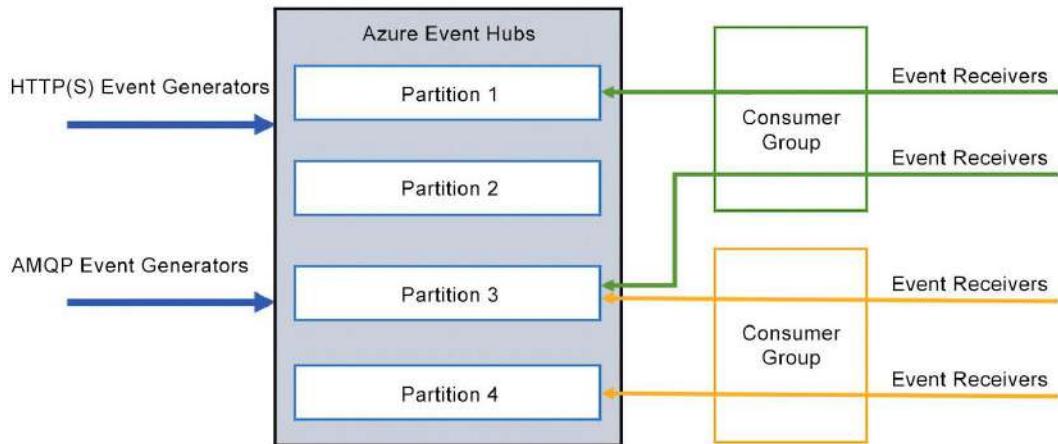


Figure 6.2 – Azure Event Hubs architecture

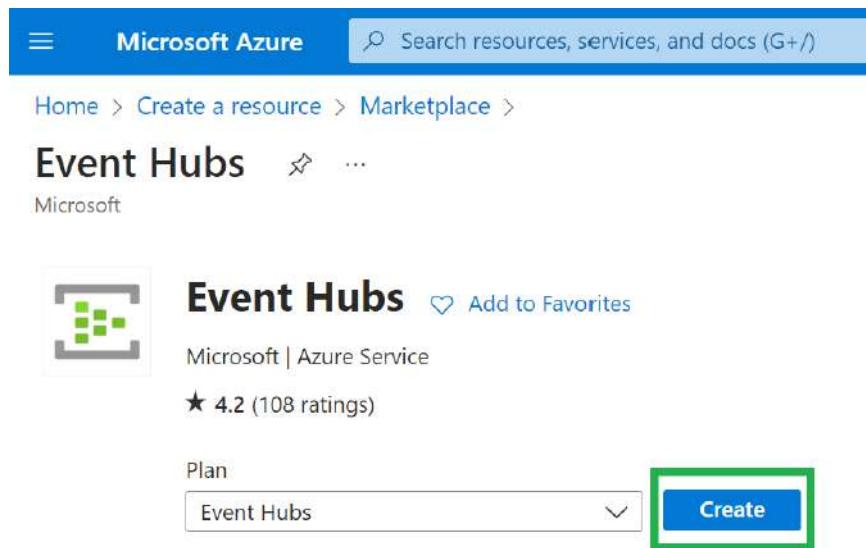


Figure 6.3 – The Event Hubs creation screen

This screenshot shows the 'Create Namespace' step of the Azure Event Hubs creation wizard. It includes fields for 'Subscription*', 'Resource group*', 'Namespace name*', 'Location*', 'Pricing tier*', and 'Throughput Units*'. The 'Review + create' button at the bottom left is highlighted with a green box.

Subscription *	Azure subscription
Resource group *	(New) rg-dp203-eh Create new
Namespace name *	DP203EHNS .servicebus.windows.net
Location *	UK South <small>The region selected supports Availability zones. Your namespace will have Availability Zones enabled. Learn more.</small>
Pricing tier *	Basic (~\$11 USD per TU per Month) <small>Browse the available plans and their features</small>
Throughput Units *	1

Figure 6.4 – Creating the Event Hubs namespace

The screenshot shows the Microsoft Azure Event Hubs Namespace overview page for the workspace 'DP203EHNS'. At the top, there's a navigation bar with 'Microsoft Azure' and a search bar. Below it, the workspace name 'DP203EHNS' is displayed along with a green 'Event Hubs Namespace' icon. A green button labeled '+ Event Hub' is highlighted. To its right are 'Delete', 'Refresh', and 'Give feedback' buttons. A dropdown menu 'Essentials' is open, showing 'NAMESPACE CONTENTS' (0 EVENT HUBS), 'KAFKA SURFACE' (NOT SUPPORTED), and 'ZONE REDUNDANCY' (ENABLED). Below this, a time range selector shows 'Show data for the last: 1 hour' (selected), 6 hours, 12 hours, 1 day, 7 days, and 30 days.

Figure 6.5 – Creating an Event Hub from within the Event Hubs workspace

The screenshot shows the 'Create Event Hub' configuration screen. At the top, the breadcrumb navigation is 'Home > DP203EHNS | Overview > DP203EHNS > Create Event Hub'. The main title is 'Create Event Hub' with a '...' button. Below it, the section 'Event Hub Details' is shown with the sub-instruction 'Enter required settings for this event hub, including partition count and message retention.' A 'Name' field is set to 'asaeh' with a green checkmark. A 'Partition count' slider is set to '1'. In the 'Retention' section, 'Configure retention settings for this Event Hub' is mentioned, with 'Learn more' link. 'Cleanup policy' is set to 'Delete' and 'Retention time (hrs)' is set to '1'. Below these, a note says 'min. 1 hour, max. 24 hours (1day)'. At the bottom, there are three buttons: 'Review + create' (highlighted with a green border), '< Previous', and 'Next: Capture >'.

Figure 6.6 – The Create Event Hub screen

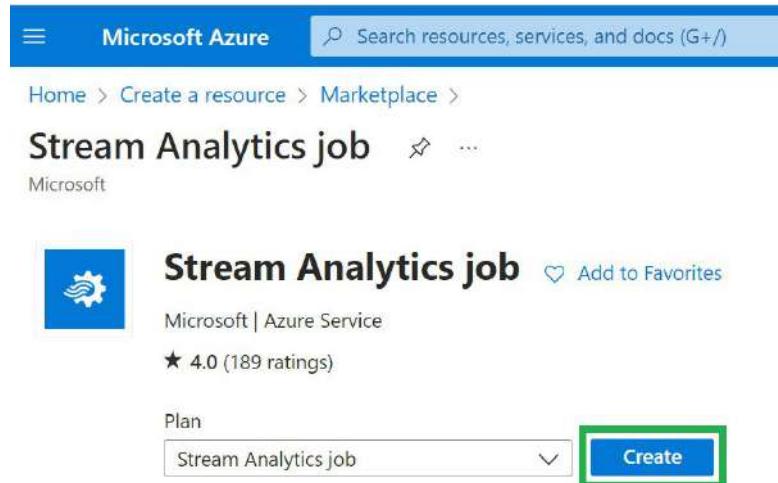


Figure 6.7 – Creating an ASA job

A screenshot of the 'New Stream Analytics job' creation screen in the Microsoft Azure portal. The top navigation bar shows 'Microsoft Azure' and a search bar. The breadcrumb navigation shows 'Home > Create a resource > Marketplace > Stream Analytics job >'. The main title is 'New Stream Analytics job' with three dots. Below the title, there are four tabs: 'Basics' (which is selected and underlined), 'Storage', 'Tags', and 'Review + create'.
Instance details
Name *: SampleASAJob
Region *: (Europe) UK South
Hosting environment *:
 Cloud
 Edge
Streaming unit details
Streaming units (SUs) represents the computing resources that are allocated to execute a Stream Analytics job. The higher the number of SUs, the more CPU and memory resources are allocated for your job. The number of SUs can be modified once you create the job. You will be charged for the job's Streaming Units only when the job runs. [Learn more about](#)

Previous

Figure 6.8 – The ASA job creation screen

The screenshot shows the Azure portal interface for managing Event Hubs. On the left, a sidebar lists 'Shared access policies' under 'Settings'. In the center, a modal window titled 'Add SAS Policy' is open, showing a form to create a new policy. The 'Policy name' field is filled with 'EH-ASA-Access'. Under the 'Policy' section, three checkboxes are checked: 'Manage', 'Send', and 'Listen'. A green box highlights the 'Policy name' field and the checked permission checkboxes. At the bottom right of the modal is a blue 'Create' button.

Figure 6.9 – Creating shared access policies in Event Hubs

The screenshot shows the configuration details for the 'EH-ASA-Access' SAS policy. At the top, there are standard save and delete buttons. Below them, three checkboxes are shown: 'Manage' (checked), 'Send' (unchecked), and 'Listen' (checked). Under 'Primary key', the value 'rqSnyv5R8rk' is displayed. Under 'Secondary key', the value 'QElbKzj+Pny' is displayed. A large green box highlights the 'Connection string-primary key' and 'Connection string-secondary key' sections. Each section contains an 'Endpoint=sb:' field with a copy icon. Below these is a 'SAS Policy ARM ID' field containing '/subscriptions' with a copy icon. At the bottom right of the page is a blue 'Save' button.

Figure 6.10 – Accessing the connection string for Event Hubs

The screenshot shows the Microsoft Azure Stream Analytics job configuration interface. At the top, there's a search bar and a navigation bar with 'Home > SampleASAJob'. The main title is 'SampleASAJob | Inputs'. Below the title, it says 'Stream Analytics job'. On the left, there's a sidebar with 'Job topology' sections: 'Inputs' (highlighted with a green box), 'Functions', 'Query', 'Outputs', and 'No-code editor (preview)'. Under 'Settings', there are 'Environment' and 'Storage account settings'. At the top right, there's a search bar, a 'Refresh' button, and a 'Add input' button (also highlighted with a green box). A dropdown menu titled 'Stream input' lists options like 'Event Hub' (which is also highlighted with a green box). Other options in the 'Stream input' list include 'Blob storage/ADLS Gen2', 'IoT Hub', and 'Kafka (preview)'. Below this, under 'Reference input', are 'Blob storage/ADLS Gen2' and 'SQL Database'.

Figure 6.11 – Selecting an Event Hub as input for the ASA job

The screenshot shows the 'Event Hub' configuration dialog. It starts with a 'New input' header. The first section is 'Input alias *' with a dropdown containing 'EH-ASA-Stream' (highlighted with a green box). Below this are two radio buttons: 'Provide Event Hub settings manually' (unchecked) and 'Select Event Hub from your subscriptions' (checked). The next section is 'Subscription' with a dropdown set to 'Azure subscription'. The third section is 'Event Hub namespace *' with a dropdown containing 'DP203EHNS' (highlighted with a green box). Below this is 'Event Hub name *' with a dropdown containing 'asaeh' (highlighted with a green box). To the right, there are sections for 'Event Hub consumer group *': 'Create new' (unchecked) and 'Use existing' (checked, with a dropdown set to '\$Default'). There's also an 'Authentication mode' dropdown set to 'Connection string'. The final section on the right is 'Event Hub policy name *': 'Create new' (unchecked) and 'Use existing' (checked, with a dropdown set to 'EH-ASA-Access'). Below this is 'Event Hub policy key *' with a masked input field. At the bottom right is a large blue 'Save' button (highlighted with a green box).

Figure 6.12 – Linking the Event Hub as an input in ASA

Home > SampleASAJob

SampleASAJob | Outputs

Stream Analytics job

Search Add output

Job topology

- Inputs
- Functions
- Query
- Outputs
- No-code editor (preview)

Alias ↑

- Blob storage/ADLS Gen 2
- Cosmos DB
- Data Lake Storage Gen1
- Event Hub
- Kafka (preview)
- PostgreSQL database
- Power BI
- Service Bus queue
- Service Bus topic
- SQL Database
- Table storage

Figure 6.13 – Selecting Power BI as the output for the ASA job

Power BI

New output

Output alias *

Provide Power BI settings manually
 Select Power BI from your subscriptions

Group workspace *

Authentication mode

Dataset name *

Table name *

Authorize connection
You'll need to authorize with Power BI to configure your output settings.

Figure 6.14 – Configuring the Power BI sink details for the ASA job

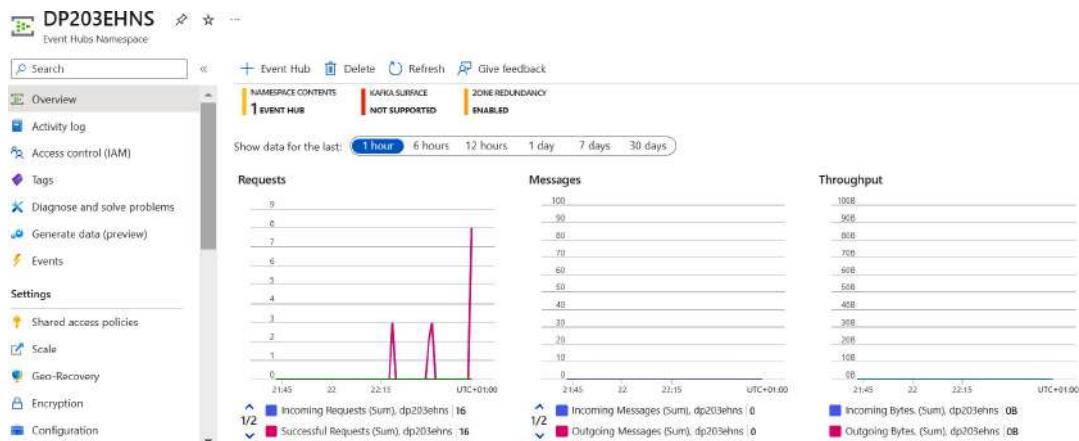


Figure 6.15 – The Event Hubs overview screen showing the event metrics

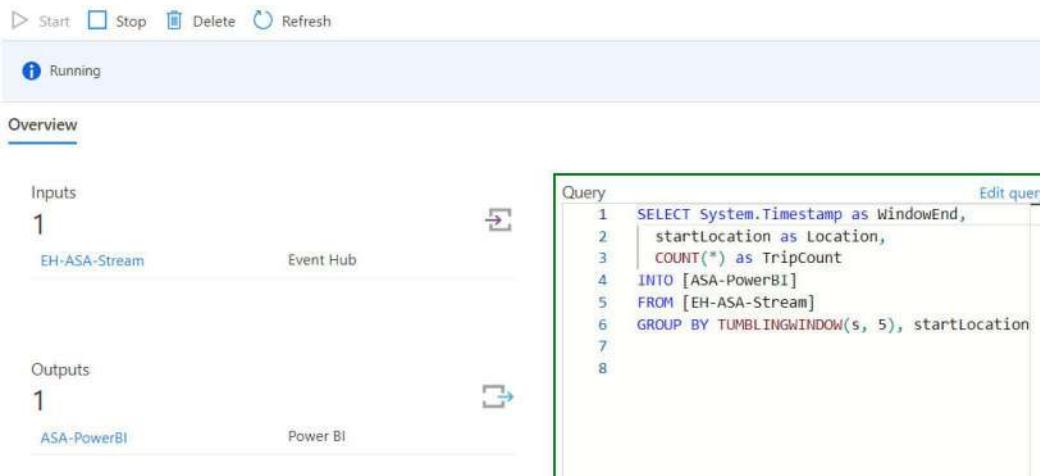


Figure 6.16 – ASA sample code reading input event from Event Hub and outputs to Power BI

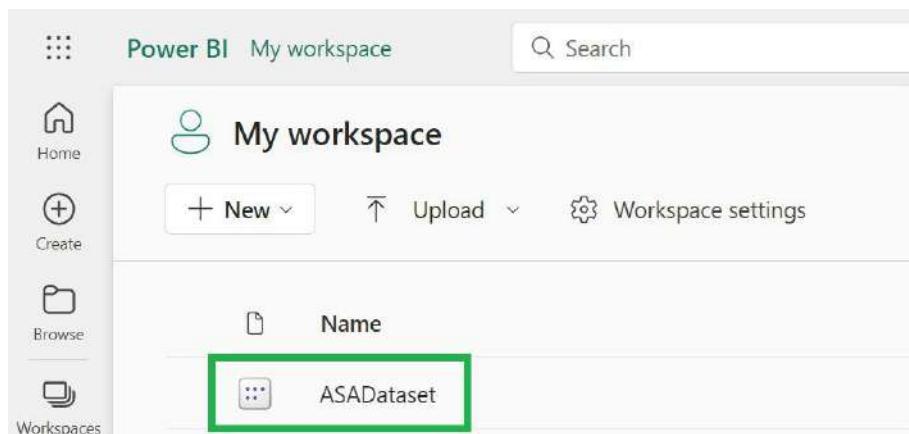


Figure 6.17 – My workspace showing the configured ASA dataset in the Power BI dataset

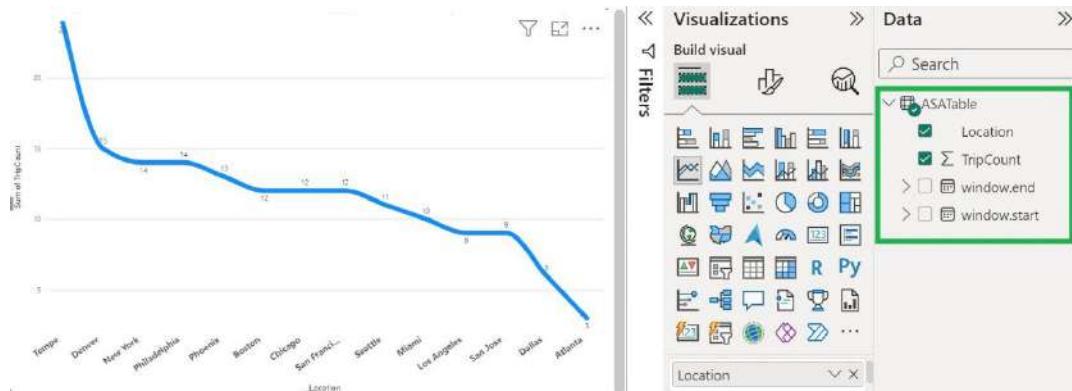


Figure 6.18 – Creating the Power BI dashboard from the ASA streaming output

Table		Data Profile	New result table: ON	Search
	window	startLocation	count	
1	> ("end": "2024-12-06T11:53:00+0000", "start": "2024-12-06T11:52:00+0...")	San Francisco	4	
2	> ("end": "2024-12-06T11:53:00+0000", "start": "2024-12-06T11:52:00+0...")	Dallas	2	
3	> ("end": "2024-12-06T11:53:00+0000", "start": "2024-12-06T11:52:00+0...")	Atlanta	1	
4	> ("end": "2024-12-06T11:53:00+0000", "start": "2024-12-06T11:52:00+0...")	Tempe	8	
5	> ("end": "2024-12-06T11:53:00+0000", "start": "2024-12-06T11:52:00+0...")	San Jose	3	
6	> ("end": "2024-12-06T11:53:00+0000", "start": "2024-12-06T11:52:00+0...")	Denver	5	
7	> ("end": "2024-12-06T11:53:00+0000", "start": "2024-12-06T11:52:00+0...")	Chicago	6	

14 rows | 1.08 seconds runtime Refreshed 2 minutes ago

Figure 6.19 – Viewing the results of the streaming query

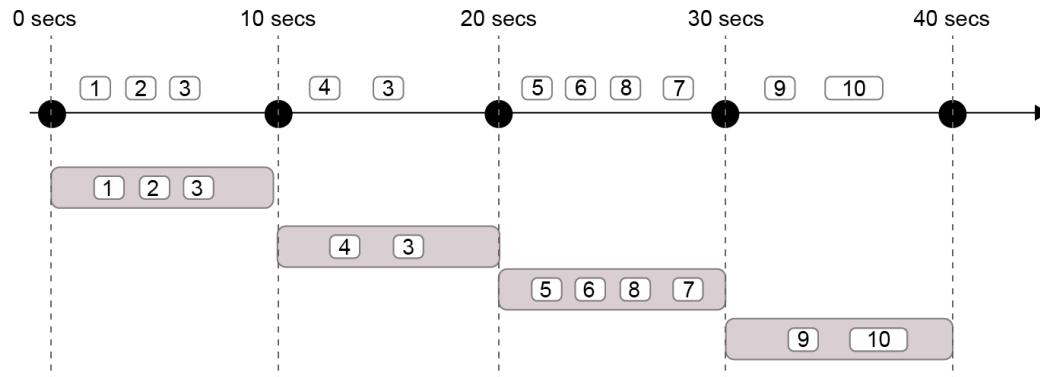


Figure 6.20 – A Tumbling Window showing a stream segmented into 10-second time windows

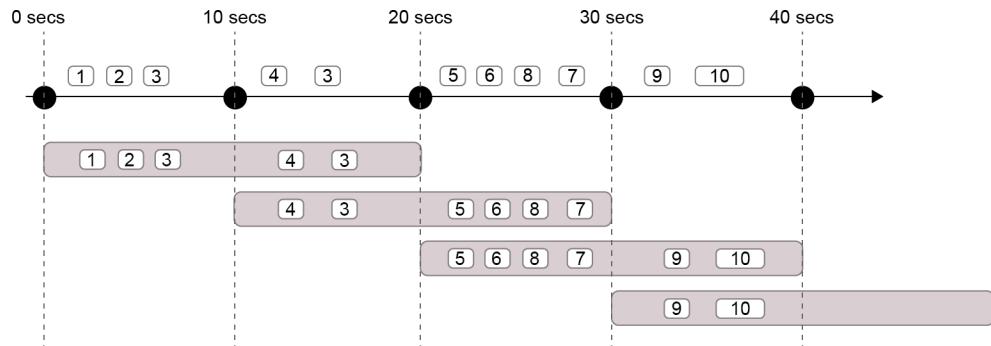


Figure 6.21 – A hopping window showing a fixed-size overlap with the previous window

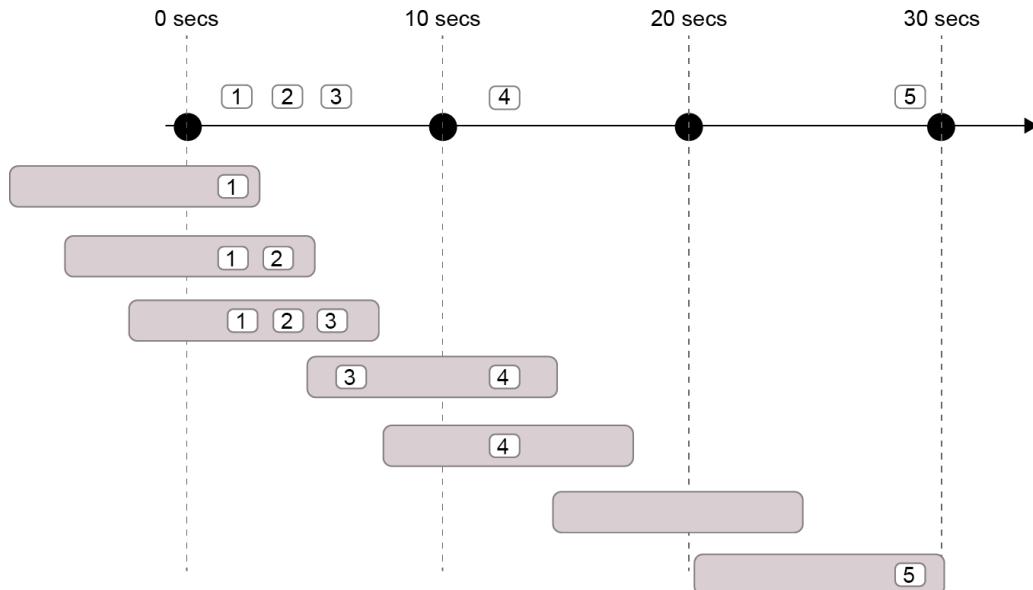


Figure 6.22 – A sliding window reporting when an event is added or removed

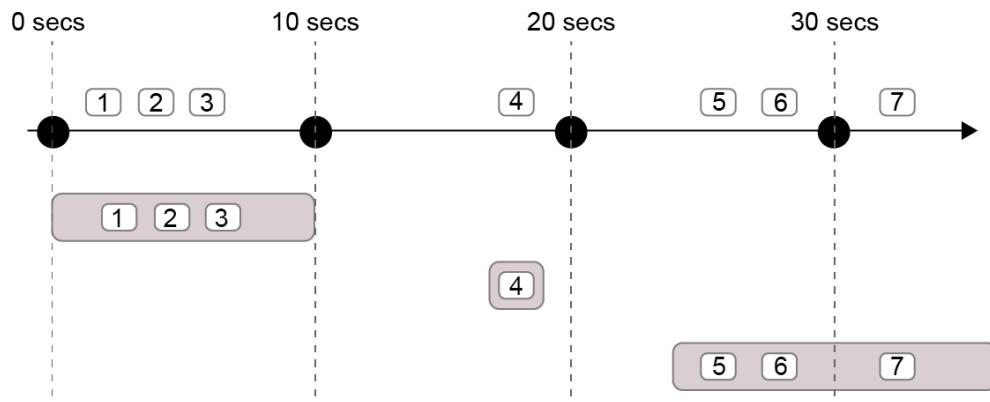


Figure 6.23 – An example of a session window with no fixed sizes

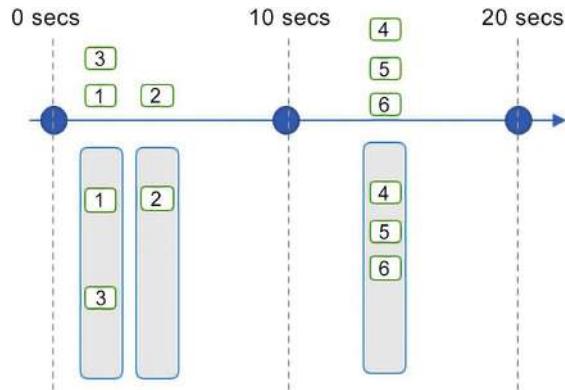


Figure 6.24 – A snapshot window showing an overview of an event at a particular time

Home > Event Hubs >

Create Namespace

Event Hubs

Instance Details

Enter required settings for this namespace, including a price tier and configuring the number of units (capacity).

Namespace name *	asaeh	.servicebus.windows.net
Location *	UK South	<small>i The region selected supports Availability zones. Your namespace will have Availability Zones enabled. Learn more.</small>
Pricing tier *	Standard (~\$22 USD per TU per Month)	<small>Browse the available plans and their features</small>
Throughput Units *	<input type="range"/> 5	
Enable Auto-Inflate	<input checked="" type="checkbox"/>	
Auto-Inflate Maximum Throughput Units	<input type="range"/> 5	

Review + create < Previous Next: Advanced >

Figure 6.25 – Enabling auto-inflate feature in Event Hubs

Home > [SampleASAJob](#)

SampleASAJob | Scale

Stream Analytics job

Search Save Discard Refresh Logs Feedback

Settings

- Environment
- Scale**
- Locale
- Event ordering
- Networking (preview)
- Error policy
- Compatibility level
- Managed Identity
- Schema registry (preview)

Choose how to scale your resource

- Manual scale**
 Manually adjust your job's fixed Streaming Units
- Custom autoscale
 Dynamically scale Streaming Units based on load or schedule

Manual scale

Override condition

Streaming unit	1
----------------	---

Figure 6.26 – Scaling an ASA job

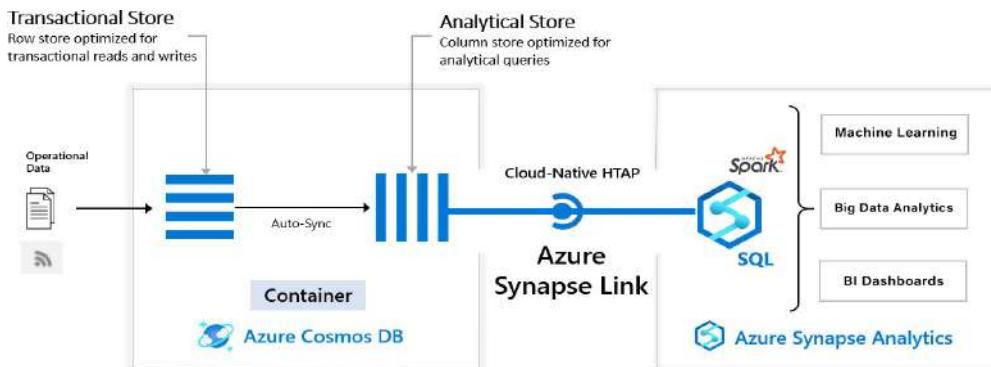


Figure 6.27 – Azure Synapse Link for Cosmos DB architecture

Home >

Azure Cosmos DB

Default Directory

Create Restore Manage view Refresh Export to CSV

Filter for any field... Subscription equals all Type equals all

Showing 0 to 0 of 0 records.

Name ↑↓	Status ↑↓
---------	-----------

Figure 6.28 – Creating a new Cosmos DB instance

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

Home > Create a resource > Marketplace > Azure Cosmos DB >

Create Azure Cosmos DB Account - Azure Cosmos DB...

Subscription * Azure subscription

Resource Group * rg-dp203-eh Create new

Instance Details

Account Name * dp203samplecosmosdb

Configure availability zone settings for your account. You cannot change these settings once the account is created.

Enable Disable

Availability Zones ⓘ

Location * ⓘ (Europe) UK South

Available locations are determined by your subscription's access and availability zone support (if that is enabled). If you don't see or cannot select your desired location, please open a support request for region access.
[Click here for more details on how to create a region access request](#)

Provisioned throughput Serverless
[Learn more about capacity mode](#)

With Azure Cosmos DB free tier, you will get the first 1000 RU/s and 25 GB of storage for free in an account. You can enable free tier on up to one account per subscription. Estimated \$64/month discount per account.

Apply Do Not Apply

Review + create Previous Next: Global Distribution Feedback

Figure 6.29 – Cosmos DB creation screen for building a NoSQL account

Home > Microsoft.Azure.CosmosDB-20240421145409 | Overview > dp203samplecosmosdb

dp203samplecosmosdb | Azure Synapse Link ⚡ ...

Azure Cosmos DB account

Search

Integrations

- Power BI
- Azure Synapse Link**
- Add Azure AI Search
- Add Azure Function

Containers

- Browse

Enable Azure Synapse Link Select Workspace Link Synapse Analytics

Enable Azure Synapse Link

Enable Azure Synapse Link to run near real-time analytics over operational data in Azure Cosmos DB. [Learn more](#)
 Here are a few analytics use cases you can build on your Azure Cosmos DB data.

Enable Azure Synapse Link for your containers

After Synapse Link is enabled on your Azure Cosmos DB database account, you can choose which containers will be enabled for analytics with Synapse Link. Enabling Synapse Link on your containers will have cost implications. [Learn more](#)

Enable

Figure 6.30 – Enabling Synapse Link in Cosmos DB

New Container

- Create new Use existing
- SampleDB
- Share throughput across containers
- Database throughput (400 - unlimited RU/s)**
- Autoscale Manual
- Estimate your required RU/s with capacity calculator.
- Database Required RU/s** 400
- Estimated cost (USD) **\$0.032 hourly / \$0.77 daily / \$23.36 monthly** (1 region, 400RU/s, \$0.00008/RU)

Container id Container1

Indexing Automatic

Figure 6.31 – Configuring a new container by selecting Analytical store

Linked services

Linked services are much like connection strings, which define the external data source.

New

Name	Type
AzureSqlDatabase	Azure SQL Database
bing-covid-19-data	Azure Blob Storage

Azure

Azure Cosmos DB for MongoDB
Azure Cosmos DB for NoSQL

Figure 6.32 – Creating a linked service to Cosmos DB

New linked service

 Azure Cosmos DB for NoSQL [Learn more](#) 

SampleCosmosDb

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 

Azure Cosmos DB account name *

dp203samplecosmosdb 

Database name *

SampleDB 

[Create](#)

[Back](#)

 Test connection

[Cancel](#)

Figure 6.33 – Cosmos DB linked service showing Cosmos DB for Synapse Link

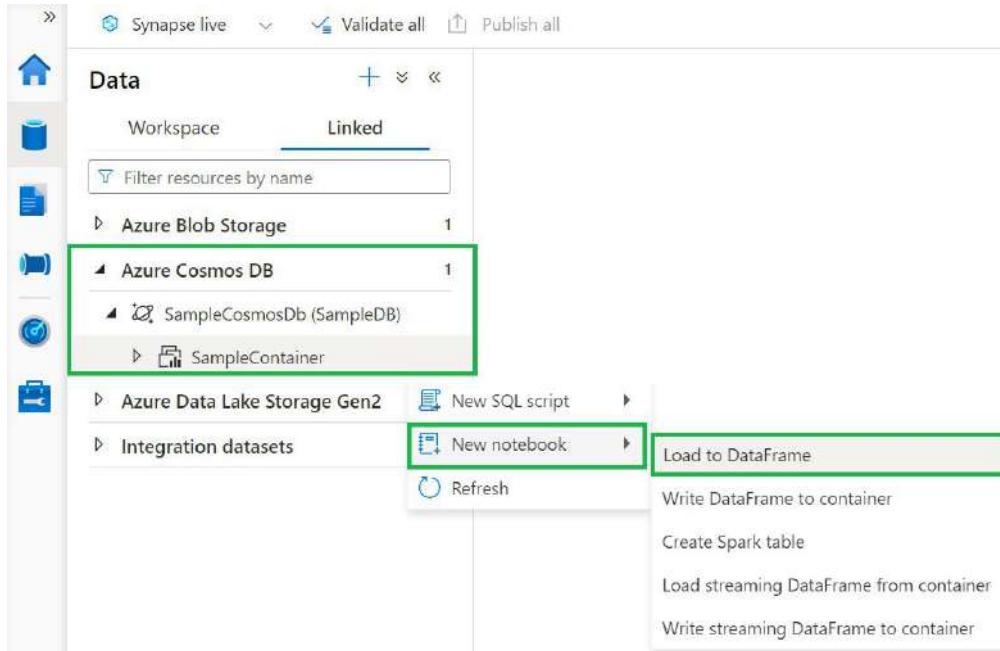


Figure 6.34 – Loading Cosmos DB data using the Synapse Link

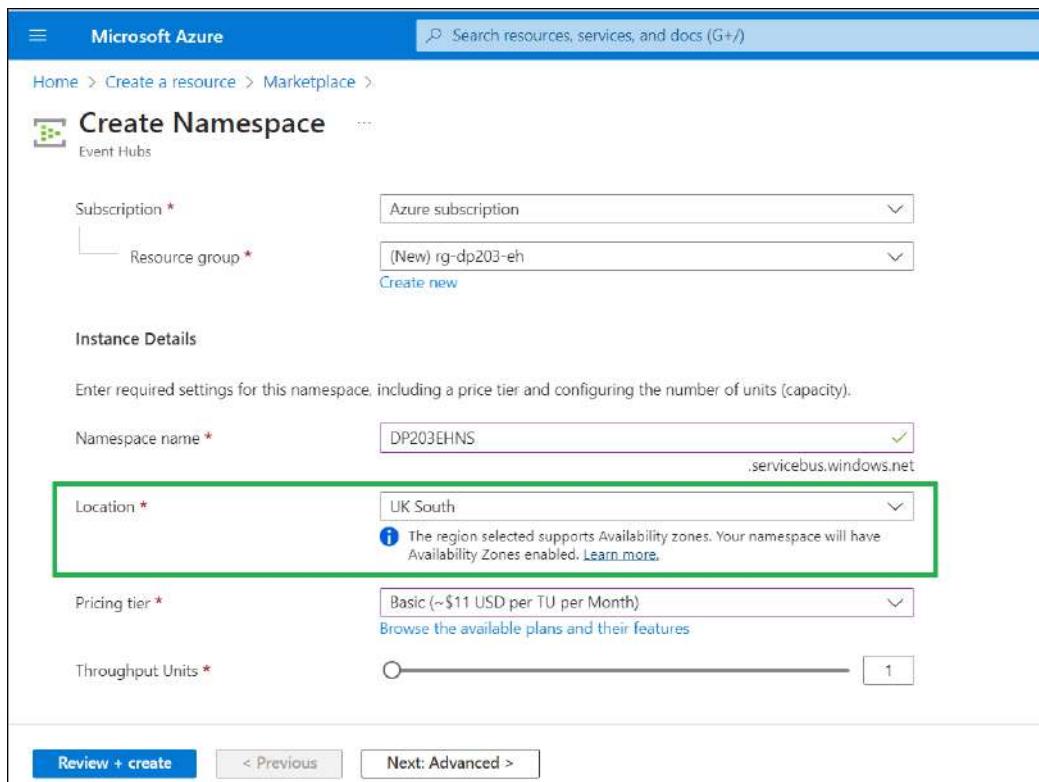


Figure 6.35 – Choosing locations with availability zones in Event Hubs

The screenshot shows the 'SampleASAJob' settings page in the Azure portal. The left sidebar lists various settings like Environment, Storage account settings, Scale, Locale, Event ordering, Networking (preview), Error policy, Compatibility level, Managed Identity, and Schema registry (preview). The 'Compatibility level' section is highlighted with a green box. A dropdown menu is open, showing options 1.2, 1.0, 1.1, and 1.2 again (the current selection). At the bottom of the page are 'Apply' and 'Discard changes' buttons.

Figure 6.36 – Updating the compatibility level in ASA to insert or update within a document

The screenshot shows the 'Practice Resources' dashboard for 'CHAPTER 6'. The main content area is titled 'Developing a Stream Processing Solution' and has a 'Summary' section. It discusses the completion of the chapter, mentioning Event Hubs, ASA, and Spark Streaming, and provides links for further learning. To the right, there's a 'Chapter Review Questions' section for 'The Azure Data Engineer Associate Certification Guide - Second Edition'. It includes a 'Select Quiz' button, a 'Quiz 1' section with a 'SHOW QUIZ DETAILS' link, and a 'START' button.

Figure 6.38 – Chapter Review Questions for Chapter 6

Chapter 7: Managing Batches and Pipelines

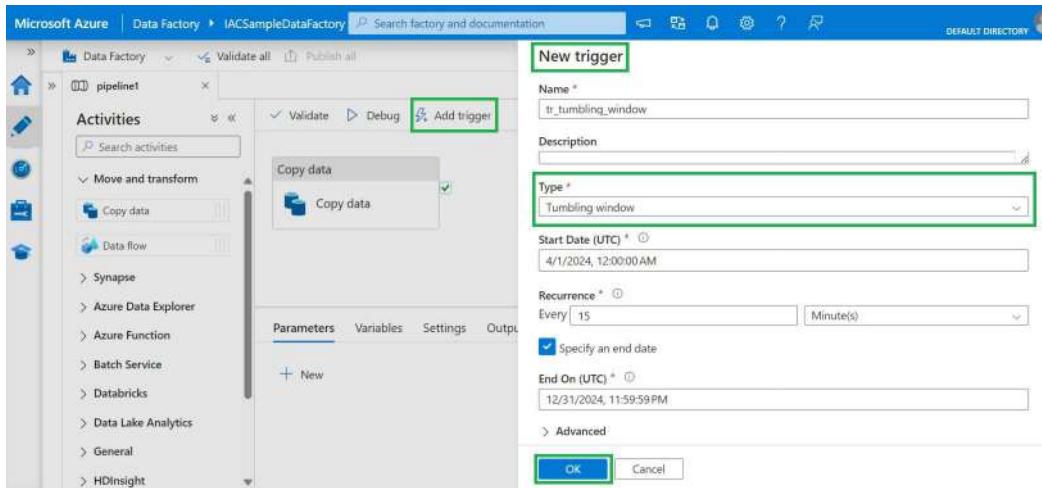


Figure 7.1 – Defining a Tumbling Window trigger for a pipeline

New trigger

Name *
tr_storge_events

Description

Type *
Storage events

Account selection method * ⓘ
 From Azure subscription
 Enter manually

Azure subscription ⓘ
Azure subscription

Storage account name * ⓘ
synapseazdl

Container name * ⓘ
raw

Blob path begins with ⓘ
2024

Blob path ends with ⓘ
.csv

Event * ⓘ
 Blob created
 Blob deleted

Ignore empty blobs * ⓘ
 Yes No

Annotations
+ New

Continue Cancel

Figure 7.2 – Defining a Storage Event trigger for a pipeline

New trigger

Name *
tr_custom_event

Description
Custom Event Trigger

Type *
Custom events

Account selection method * ⓘ
 From Azure subscription
 Enter manually

Azure subscription ⓘ
Azure subscription

Event grid topic name *
customEventDemo

Subject filters * ⓘ

Subject begins with ⓘ
factories

Subject ends with ⓘ

Event types * ⓘ

+ New

Value

CopyCompleted

CopySucceeded

Advanced filters ⓘ

OK Cancel

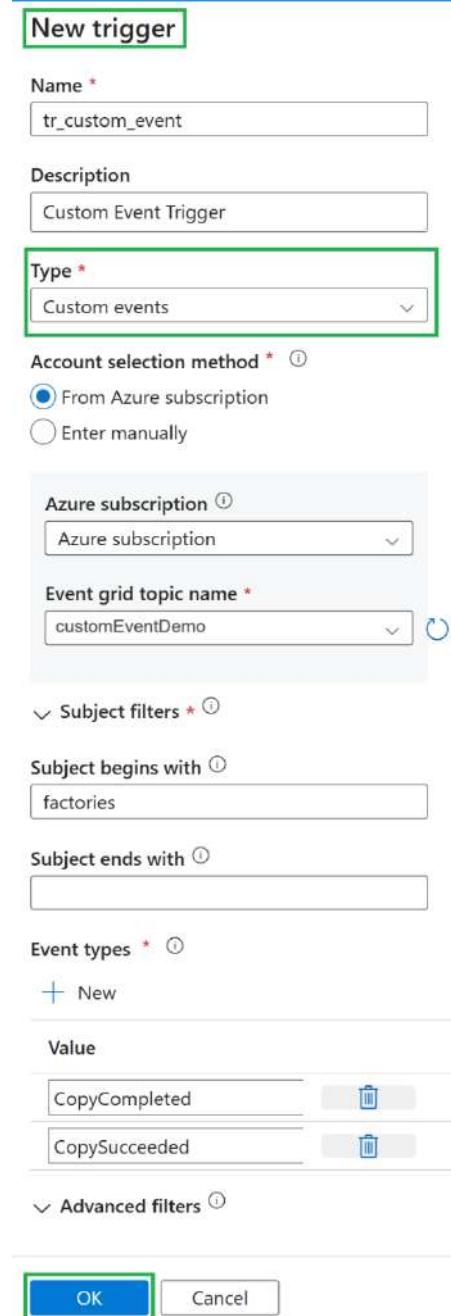


Figure 7.3 – Defining a custom Event trigger for a pipeline

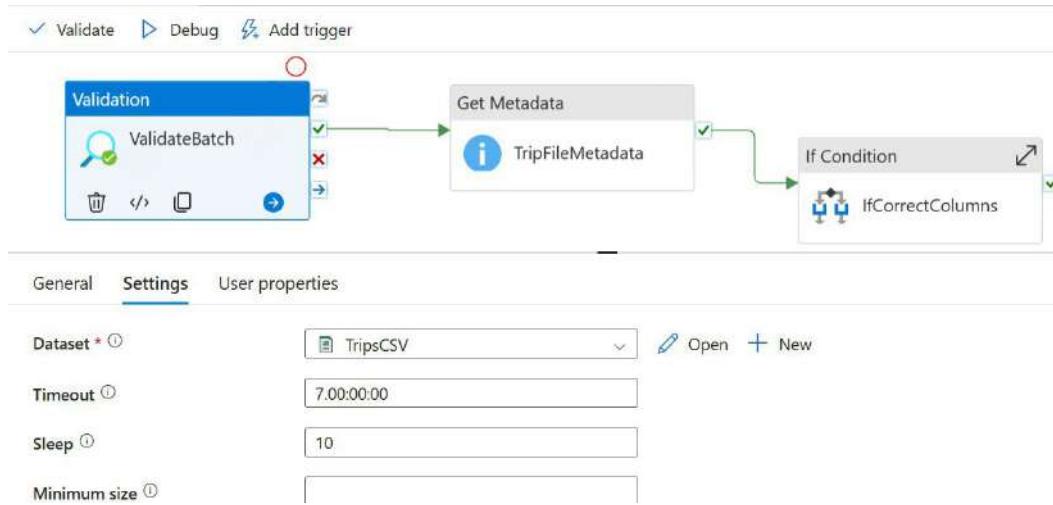


Figure 7.4 – Using the ADF Validation activity to check for a file's existence

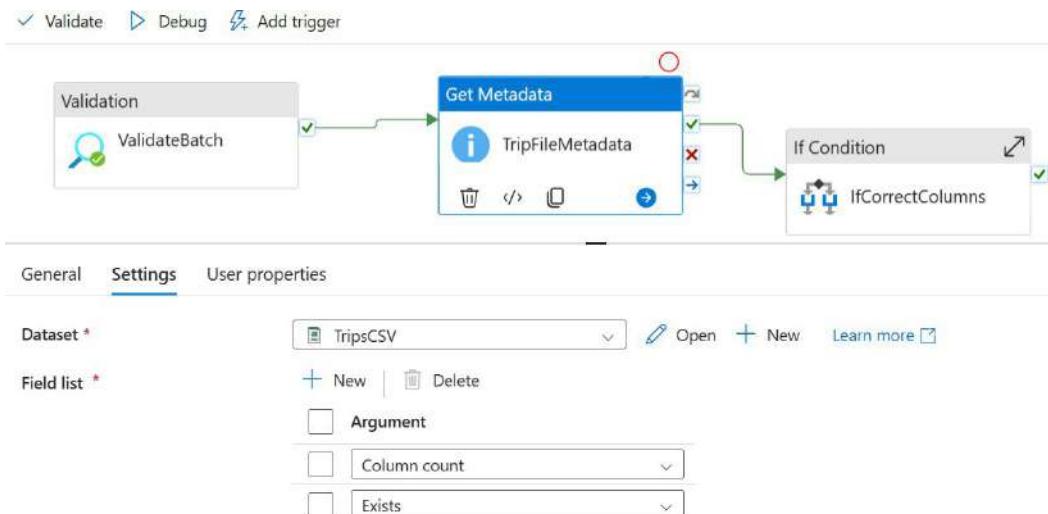


Figure 7.5 – Configuring the Get Metadata activity to publish the column count

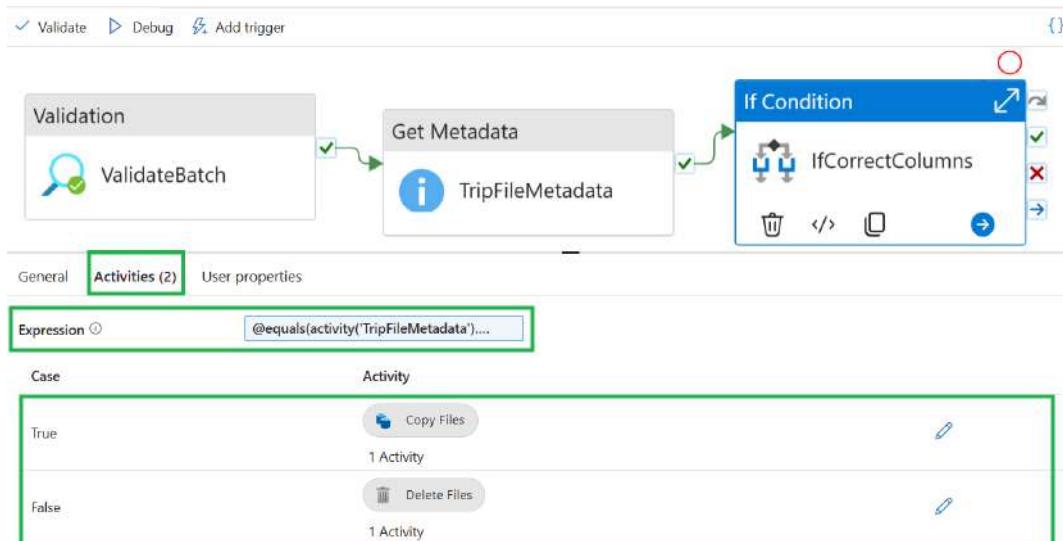


Figure 7.6 – Using the metadata from the Get Metadata activity to make a decision

The screenshot shows the Azure Data Factory Manage screen. The left sidebar contains the following navigation options:

- General
- Connections
- Linked services
- Integration runtimes
- Microsoft Purview
- Source control
- Git configuration
- ARM template
- Author
- Triggers
- Global parameters
- Data flow libraries
- Security
- Credentials
- Customer managed key
- Outbound rules
- Managed private endpoints
- Workflow orchestration manager
- Apache Airflow

The main area is titled "Linked services" and defines the connection information to a data store or compute. It shows three items:

Name	Type	Related
AzureBlobConnection	Azure Data Lake Storage Gen2	1
AzureSqlDatabase1	Azure SQL Database	0
ls_AzureDataLakeStorage	Azure Data Lake Storage Gen2	9

Figure 7.7 – The Manage screen of ADF showing the settings menu options to manage various components

The screenshot shows the 'Integration runtimes' blade in the Azure Data Factory portal. On the left, a navigation menu includes 'Connections', 'Linked services', 'Integration runtimes' (which is highlighted with a green box), 'Microsoft Purview', 'Source control', 'Git configuration', 'ARM template', 'Author', 'Triggers', and 'Global parameters'. The main area is titled 'Integration runtimes' and contains a table with one item:

Name	Type	Sub-type	Status	Related	Region	Version
AutoResolveIntegrationR...	Azure	Public	Running	0	Auto Resolve	---

Figure 7.8 – Creating a new IR to run data pipelines and data flows

The screenshot shows the 'Pipeline runs' blade in the Azure Data Factory portal. On the left, a navigation menu includes 'Dashboards', 'Runs' (which is highlighted with a green box), 'Pipeline runs' (highlighted with a green box), 'Trigger runs', 'Change Data Capture (previous)', 'Runtimes & sessions', 'Integration runtimes', 'Data flow debug', 'Notifications', and 'Alerts & metrics'. The main area is titled 'Pipeline runs' and displays a table of recent runs for an 'HDI Spark Pipeline':

Pipeline name	Run start	Run end	Run	Status
HDI Spark Pipeline	4/25/2024, 4:54:19 PM	4/25/2024, 4:54:37 PM	Original	Succeeded
HDI Spark Pipeline	4/25/2024, 4:51:11 PM	4/25/2024, 4:51:28 PM	Original	Failed
HDI Spark Pipeline	4/25/2024, 4:43:08 PM	4/25/2024, 4:43:33 PM	Original	Succeeded
HDI Spark Pipeline	4/25/2024, 4:41:40 PM	4/25/2024, 4:42:00 PM	Original	Succeeded

Figure 7.9 – Monitoring ADF pipelines

The screenshot shows the 'Activity runs' tab for the 'HDI Spark Pipeline' in the Azure Data Factory portal. On the left, a navigation menu includes 'Dashboards', 'Runs', 'Pipeline runs' (highlighted with a green box), 'Trigger runs', 'Change Data Capture (previous)', 'Runtimes & sessions', 'Integration runtimes', 'Data flow debug', 'Notifications', and 'Alerts & metrics'. The main area is titled 'All pipeline runs > HDI Spark Pipeline - Activity runs' and displays a table of activity runs:

Activity name	Activity status	Activity type	Run start	Duration
Transform	Failed	Copy data	4/25/2024, 5:31:51 PM	13s
FetchTripsFro... (highlighted with a green box)	Succeeded	Data flow	4/25/2024, 5:31:41 PM	9s
FetchTripsFrmSQL	(Data flow details)	Data flow	4/25/2024, 5:28:39 PM	3m 2s

Figure 7.10 – Activity runs tab showing the Data flow details

The screenshot shows the 'Add trigger' dialog box in the Azure Data Factory or Synapse Studio interface. The dialog has a green border and contains two options:

- 'Trigger now'
- 'New/Edit'

Figure 7.11 – Adding a trigger from ADF/Synapse pipelines

New trigger

Name *
trigger1

Description

Type *
Schedule

Start date * ⓘ
4/25/2024, 3:00:52 PM

Time zone * ⓘ
Coordinated Universal Time-11 (UTC-11)

Recurrence * ⓘ
Every 15 Minute(s)
 Specify an end date

Annotations
+ New

Start trigger ⓘ
 Start trigger on creation

OK **Cancel**

Figure 7.12 – Defining the trigger in ADF or Synapse pipelines

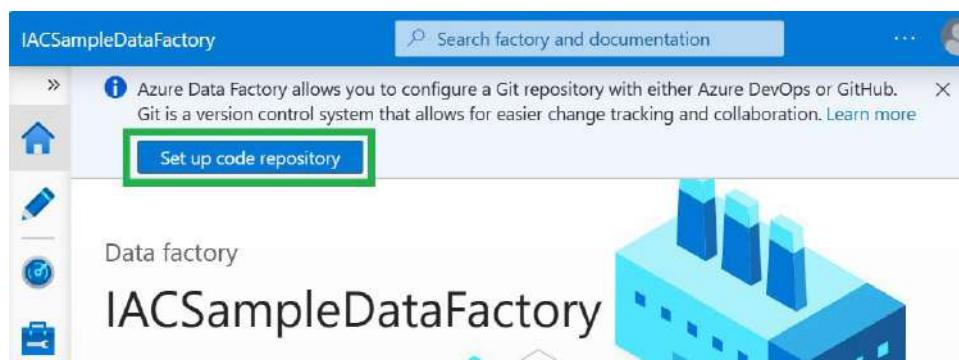


Figure 7.13 – Configuring the Git repository using the Set up code repository button

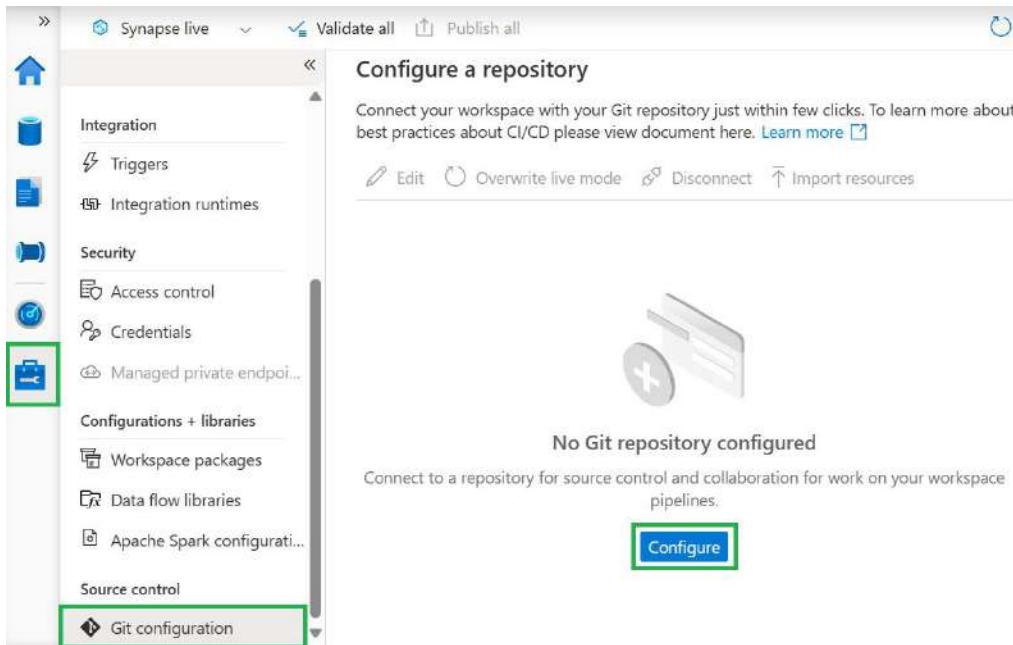


Figure 7.14 – Launching Git configuration from Synapse's Manage tab (the same in ADF)



Figure 7.15 – Creating a new DevOps organization for managing and organizing projects

Configure a repository

Specify the settings that you want to use when connecting to your repository.

Repository type * ⓘ

Select...



Azure DevOps Git



Continue

Cancel

Figure 7.16 – Configuring a repository to select the Azure DevOps Git repository type

Configure a repository

Specify the settings that you want to use when connecting to your repository.

Select repository Use repository link

Azure DevOps organization name * ⓘ

DevOps

Project name * ⓘ

ADF

Repository name * ⓘ

DEV

Collaboration branch * ⓘ

main

Publish branch * ⓘ

adf_publish

Root folder * ⓘ

/

Custom comment

Use custom comment

Apply

Back

Cancel

Figure 7.17 – Configuring Azure DevOps Git as the source control for ADF

Configure a repository



Specify the settings that you want to use when connecting to your repository.

Select repository

Use repository link

Repository name * ⓘ

adf

Collaboration branch * ⓘ

main

Publish branch * ⓘ

adf_publish

Root folder * ⓘ

/

Import existing resource

Import existing resources to repository

Figure 7.18 – Configuring GitHub as the source control for ADF or Synapse

New linked service

Azure HDInsight [Learn more](#)

From Azure subscription Enter manually

Azure subscription

Hdi Cluster *

DP203HDISpark

Storage accounts associated with cluster ⓘ

Storage	Type
dp203hdisparkhdstorage	Blob Storage

Azure Storage linked service

Blob Storage ADLS Gen 2

Azure Storage linked service *

HDIBlobStorageLS

Connection successful

Test connection

Create

Back

Figure 7.19 – Creating an HDInsight linked service

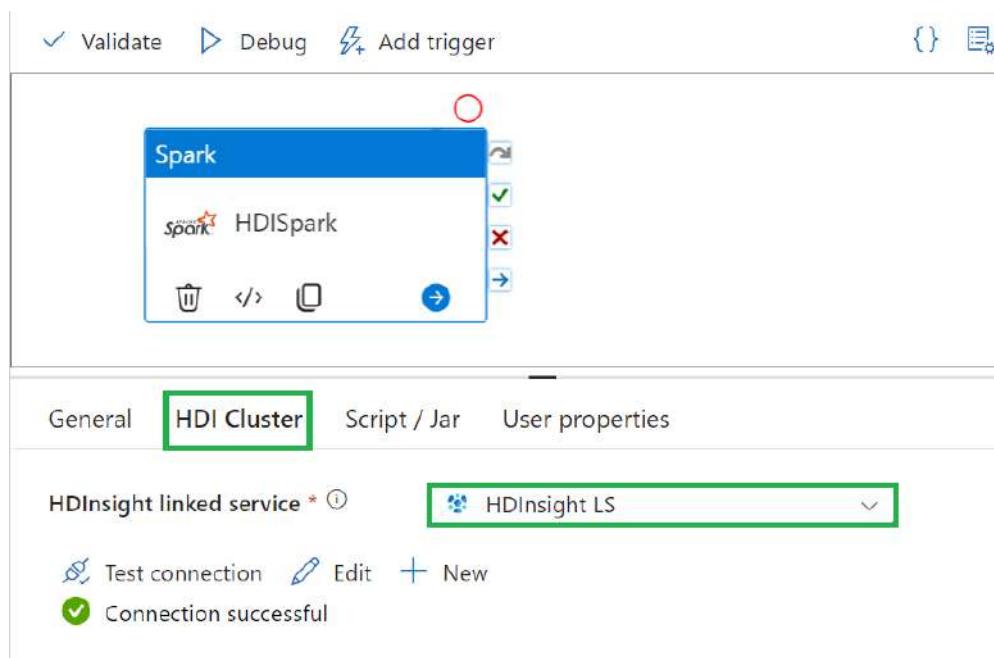


Figure 7.20 – Configuring the HDInsight Spark cluster in the ADF pipeline

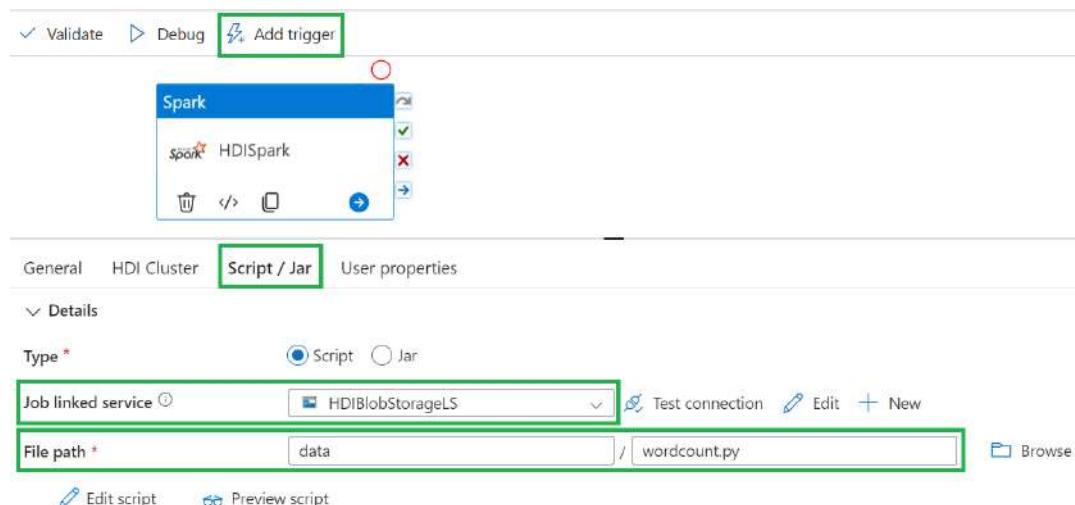


Figure 7.21 – Creating a pipeline with Spark

The screenshot shows a dark-themed web application interface. At the top left is the 'Practice Resources' logo. On the right are a bell icon and a 'SHARE FEEDBACK' button. Below the header, a breadcrumb navigation shows 'DASHBOARD > CHAPTER 7'. The main content area has a light gray background and contains the following text:

Managing Batches and Pipelines

Summary

With that, you have come to the end of this chapter. In the preceding sections, you learned how to define event triggers with ADF and Synapse pipelines, set managing and monitoring pipelines, run Spark pipelines, and configure version control in ADF and Synapse pipelines. With all this knowledge, you should now be confident in creating and managing data batch workloads and pipelines.

This chapter marks the end of the **Develop Data Processing** domain, which accounts for about 40–45% of the certification goals.

From the next chapter onward, you will move on to the **Secure, Monitor, and Optimize Data Storage and Processing** domain, where you will be focusing on the security aspects of data processing—that is, implementing data security, monitoring data storage, and optimizing resource management.

To the right of the main content is a dark sidebar titled 'Chapter Review Questions' with the following text:

Chapter Review Questions

The Azure Data Engineer Associate Certification Guide
- Second Edition by Giacinto Palmieri, Surendra Mettapalli, Newton Alex

Select Quiz

Quiz 1 [SHOW QUIZ DETAILS](#) ▾ [START](#)

Figure 7.23 – Chapter Review Questions for Chapter 7

Chapter 8: Implementing Data Security

The screenshot shows the Azure portal interface for a Dedicated SQL pool named 'mySampleDataWarehouse'. The left sidebar has a 'Dynamic Data Masking' section highlighted. The main area displays masking rules for the 'TempDriver' table. It shows two columns being masked: 'driverId' and 'firstName'. Each column has a 'Mask Function' dropdown set to 'Add mask', which is highlighted with a green box.

Figure 8.1 – Setting up DDM from the Azure portal

This screenshot shows the 'Add masking rule' dialog box. The 'Mask name' field contains 'dbo_CustomerContact_Email'. The 'Schema *' dropdown is set to 'dbo'. The 'Table *' dropdown is set to 'CustomerContact'. The 'Column *' dropdown is set to 'Email (varchar)'. The 'Masking field format' dropdown is expanded, showing options: 'Email (aXXX@XXXX.com)', 'Default value (0, xxxx, 01-01-1900)', 'Credit card value (xxxx-xxxx-xxxx-1234)', 'Email (aXXX@XXXX.com)', 'Number (random number range)', and 'Custom string [prefix [padding] suffix]'. The 'Email (aXXX@XXXX.com)' option is highlighted with a green box.

Figure 8.2 – Creating an email mask

Encryption selection

Enable support for customer-managed keys Blobs and files only

Infrastructure encryption Disabled

Encryption type

Microsoft-managed keys
 Customer-managed keys

Key selection

Encryption key Select from key vault
 Enter key URI

Figure 8.3 – Enabling CMKs in Azure Storage

Security

Auditing

Data Discovery & Classification

Dynamic Data Masking

Microsoft Defender for Cloud

Transparent data encryption

Data encryption

ON **OFF**

Encryption status
Unencrypted

Figure 8.4 – Enabling TDE using Azure Synapse SQL

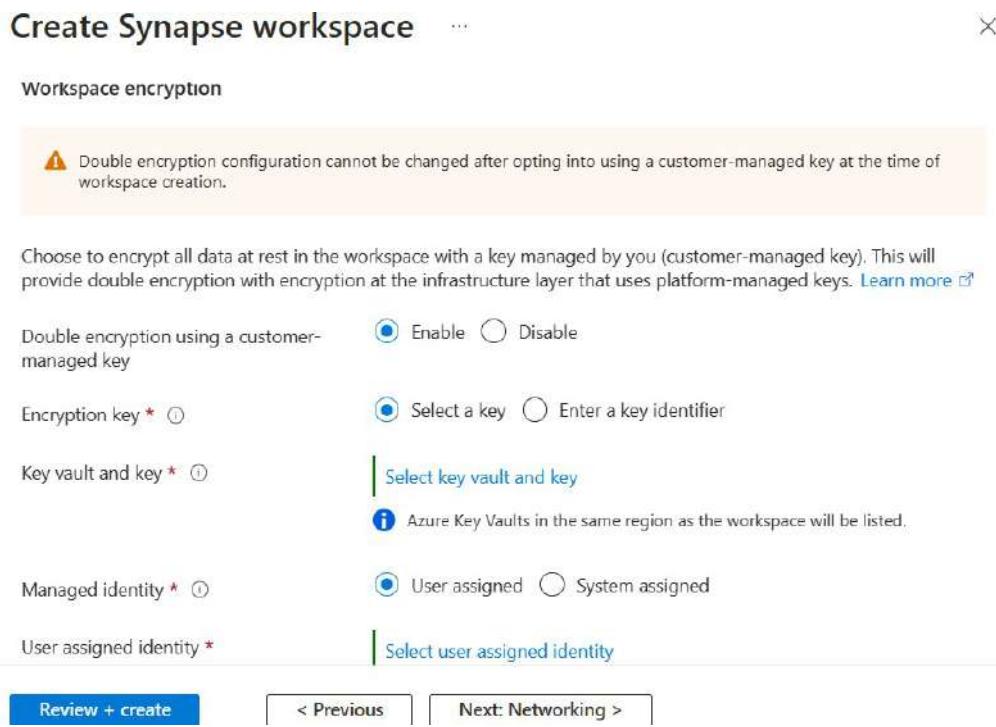


Figure 8.5 – Configuring a CMK in Azure Synapse SQL

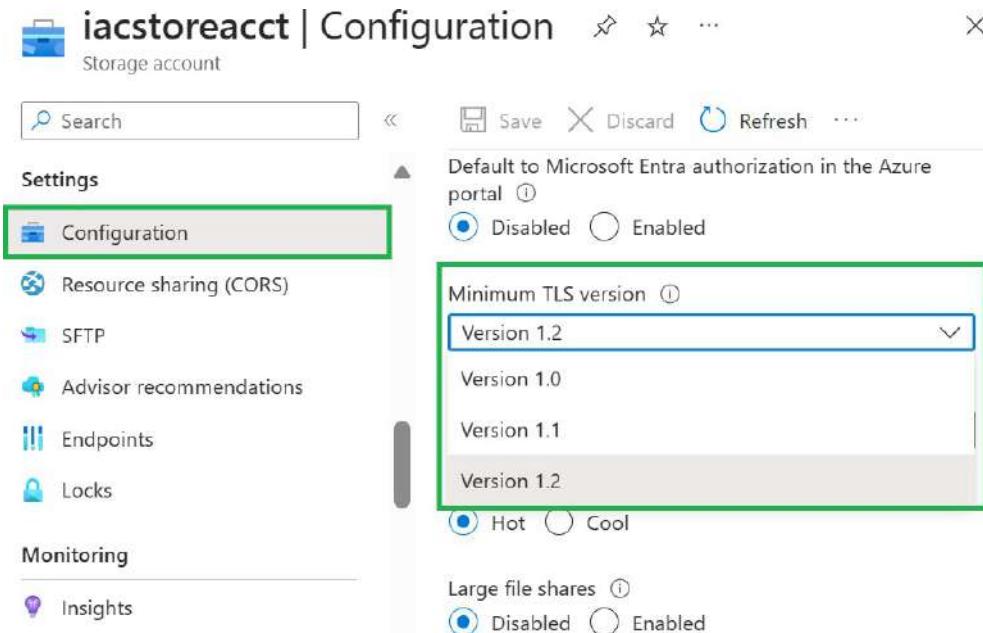


Figure 8.6 – Enabling TLS in Azure Storage

The screenshot shows a database interface with the following details:

- Toolbar:** Run, Undo, Publish, Query plan, Connect to mySampleDataWarehouse.
- Query:** 1 EXECUTE AS USER = 'HiPriv_User'; 2 SELECT * from dbo.TripTable
- Results View:** Results (selected), Messages, View (Table, Chart), Export results.
- Table Data:**

tripId	driverId	customerId	tripDate	startLocation	endLocation
111	201	301	20240101	New York	New Jersey
114	204	303	20240204	LA	San Jose
900	299	399	20240301	Pre-Launch	Pre-Launch
112	202	302	20240101	Miami	Dallas
- Message Bar:** 00:00:00 Query executed successfully.

Figure 8.7 – Displaying all rows including pre-launch for user HiPriv_User

The screenshot shows a database interface with the following details:

- Toolbar:** Run, Undo, Publish, Query plan, Connect to mySampleDataWarehouse, ...
- Query:** 1 EXECUTE AS USER = 'LowPriv_User'; 2 SELECT * from dbo.TripTable
- Results View:** Results (selected), Messages, View (Table, Chart), Export results.
- Table Data:**

tripId	driverId	customerId	tripDate	startLocation	endLocation
106	206	306	20240301	Atlanta	Chicago
114	204	303	20240204	LA	San Jose
102	202	302	20240101	Miami	Dallas
103	203	303	20240102	Phoenix	Tempe

Figure 8.8 – Row-Level Security (RLS) blocking the pre-launch location rows

The screenshot shows the Microsoft Azure interface for managing access control in a storage account named 'iacstoreacct'. The left sidebar lists several options: Overview, Activity log, Tags, Diagnose and solve problems, Access Control (IAM), Data migration, and Events. The 'Access Control (IAM)' option is highlighted with a green box. The main content area displays the 'Add role assignment' section, which includes a search bar, a 'Download role assignments' button, and a 'View my access' button. A dropdown menu is open, showing 'Add role assignment' and 'Add co-administrator' options. The 'My access' section is visible below.

Figure 8.9 – Adding role assignment in storage account through Access Control IAM

The screenshot shows the 'Add role assignment' page for Azure Data Lake Storage Gen2 (ADLS Gen2). The 'Members' tab is selected. The 'Selected role' is set to 'Storage Blob Data Owner'. The 'Assign access to' field has 'User, group, or service principal' selected. The 'Members' section shows a 'Select members' button, which is highlighted with a green box. A modal window titled 'Select members' is open, showing a search bar and a list with one item: 'Surendra Mettapalli'. Below the list, it says 'Selected members: No members selected. Search for and add one or more members you want to assign to the role for this resource.' At the bottom of the modal are 'Select' and 'Close' buttons, with 'Select' highlighted with a green box.

Figure 8.10 – Configuring the RBAC role assignment in ADLS Gen2

The screenshot shows the 'Containers' section of the Azure Storage Account 'iacstoreacct'. On the left, there's a sidebar with sections like 'Data storage' (highlighted with a green box), 'File shares', 'Queues', 'Tables', 'Security + networking', 'Data management', and 'Storage tasks (preview)'. The main area lists containers: '\$logs', 'customers' (selected and highlighted with a green box), 'drivers', 'fares', 'raw', and 'trips'. A context menu is open over the 'customers' container, listing options: 'Container properties', 'Generate SAS', 'Manage ACL' (highlighted with a green box), 'Access policy', 'Acquire lease', 'Break lease', 'Change access level', 'Edit metadata', and 'Delete'.

Figure 8.11 – Selecting Manage ACL to provide required access permissions

The screenshot shows the 'Manage ACL' page for the 'customers' container. At the top, it says 'Set and manage permissions for: root directory'. Below that is a link 'Learn more about access control lists (ACLs)'. The main area has tabs 'Access permissions' (selected, highlighted with a green box) and 'Default permissions'. Under 'Access permissions', there are buttons for '+ Add principal' and '+ Add mask'. The table below shows permissions for three security principals:

Security principal	Read	Write	Execute
Owner: \$superuser	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Owning group: \$superuser	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Other	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

A note at the bottom states: 'Read and write permissions will only work for a security principal if the security principal also has execute permissions on all parent directories, including the container (root directory.)'. At the bottom are 'Save' and 'Discard' buttons.

Figure 8.12 – Configuring the ACL in ADLS Gen2 to provide the required access to users

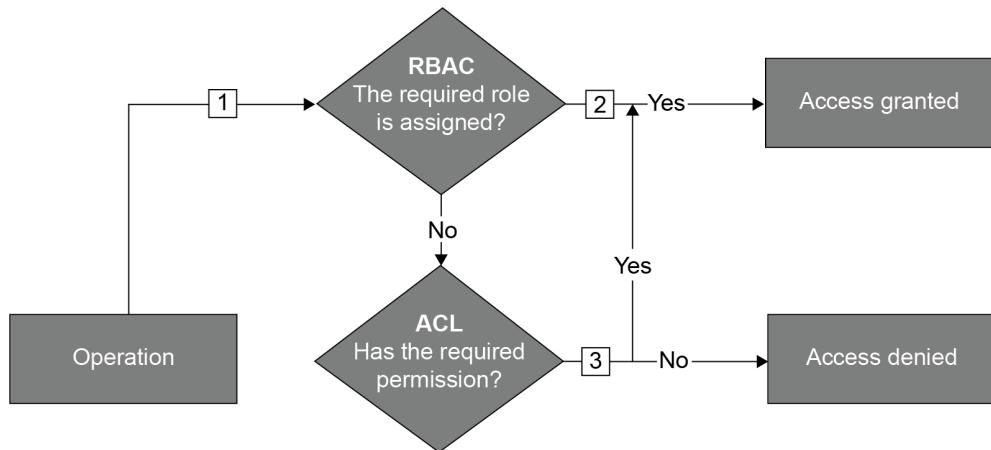


Figure 8.13 – RBAC and ACL evaluation sequence

Add a rule

...

Details

2 Base blobs

Lifecycle management uses your rules to automatically move blobs to cooler tiers or to delete them. If you create multiple rules, the associated actions must be implemented in tier order (from hot to cool storage, then archive, then deletion).

If

Base blobs were *

Last modified

Created

More than (days ago) *

365

Then

Delete the blob

+ Add conditions

Previous **Add**

Figure 8.14 – Configuring data life cycle management

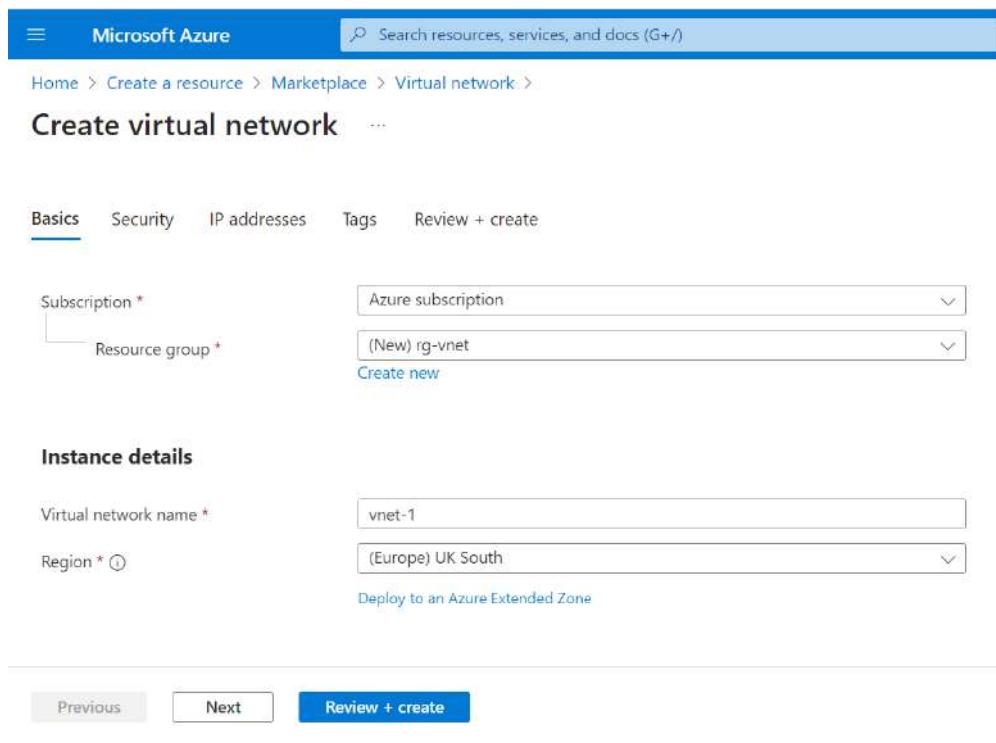


Figure 8.15 – Creating a new VNet

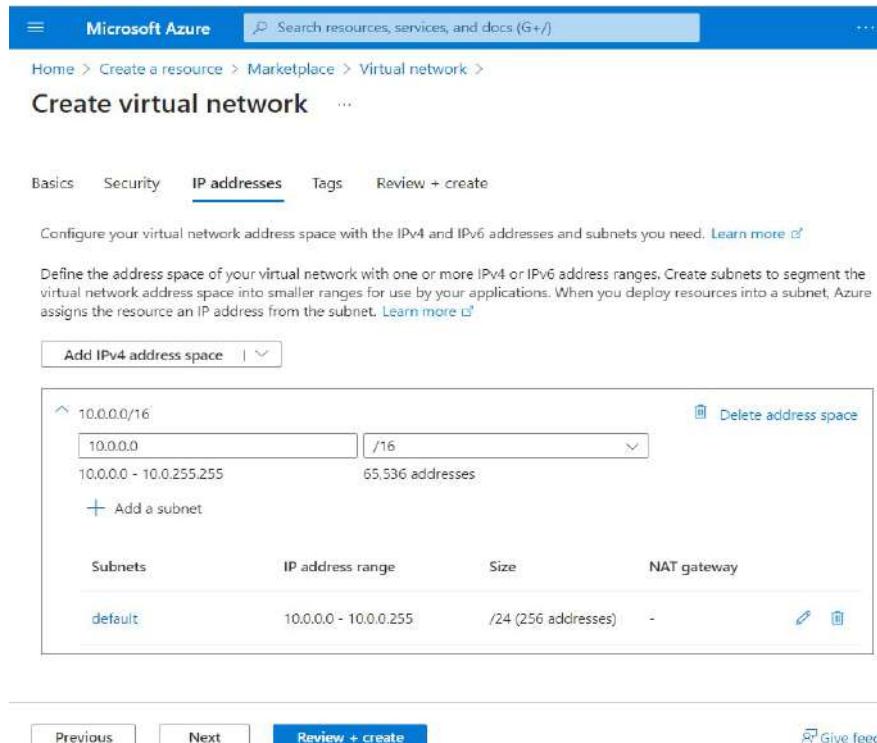


Figure 8.16 – Configuring IP details for the new VNet

The screenshot shows the 'Private Link Center' interface. The 'Private endpoints' section is highlighted with a green box. The 'Create' button is also highlighted with a green box. Other visible sections include 'Overview', 'Pending connections', 'Private link services', 'Azure Arc private link scopes', and 'Azure Monitor private link scopes'. A search bar, a 'Manage view' dropdown, and a 'Refresh' button are at the top right. A filter bar indicates 'Subscription equals all'. A message at the bottom says 'Showing 0 to 0 of 0 records.'

Figure 8.17 – Creating a private endpoint from Private Link Center

Home > Private Link Center | Private endpoints >

Create a private endpoint

X

✓ Basics **2 Resource** ③ Virtual Network ④ DNS ⑤ Tags ⑥ Review + create

Private Link offers options to create private endpoints for different Azure resources, like your private link service, a SQL server, or an Azure storage account. Select which resource you would like to connect to using this private endpoint. [Learn more](#)

Connection method ①	<input checked="" type="radio"/> Connect to an Azure resource in my directory. <input type="radio"/> Connect to an Azure resource by resource ID or alias.
Subscription * ①	Azure subscription ✓
Resource type * ①	Microsoft.Synapse/workspaces ✓
Resource * ①	synapse-az-ws ✓
Target sub-resource * ①	Select a target sub-resource ✓
	Sql SqlOnDemand Dev

< Previous Next : Virtual Network >

Figure 8.18 – Configuring a service for creating the private endpoint

Home > Private Link Center | Private endpoints >

Create a private endpoint

...

✓ Basics ✓ Resource **3 Virtual Network** ④ DNS ⑤ Tags ⑥ Review + create

Networking

To deploy the private endpoint, select a virtual network subnet. [Learn more](#)

Virtual network ①	vnet-1 (rg-vnet) ✓
Subnet * ①	default ✓

Network policy for private endpoints Disabled (edit)

Private IP configuration

Dynamically allocate IP address
 Statically allocate IP address

< Previous **Next : DNS >**

Figure 8.19 – Configuring VNet for private endpoint

Create Synapse workspace

* Basics * Security Networking Tags Review + create

Configure networking options for your workspace.

Managed virtual network

Choose whether to set up a dedicated Azure Synapse-managed virtual network for your workspace.

[Learn more](#)

Managed virtual network [?](#) Enable Disable

Create managed private endpoint to primary storage account [?](#) Yes No

Allow outbound data traffic only to approved targets [?](#) Yes No

[Review + create](#)

[< Previous](#)

[Next: Tags >](#)

Figure 8.20 – Enabling a Synapse-managed VNet to create a private endpoint

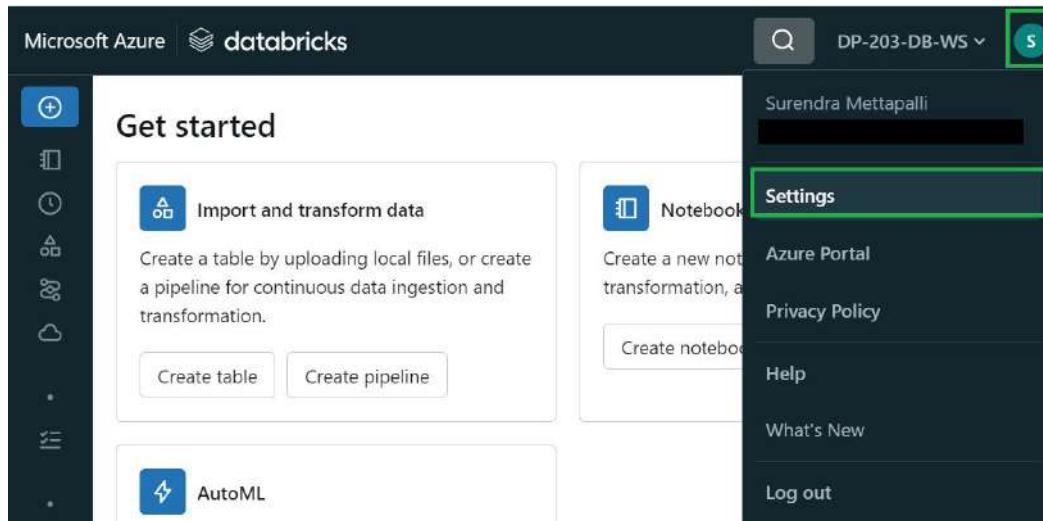


Figure 8.21 – Accessing User Settings in Azure Databricks for developer settings

The screenshot shows the 'Developer' settings page. On the left, there's a sidebar with 'User' selected. In the main area, 'Access tokens' is highlighted with a green box. It contains a 'Manage' button. Below it, 'Editor settings' and 'Spark tips' sections are shown, each with a toggle switch set to 'On'.

Figure 8.22 – Accessing Developer settings to manage access tokens

The screenshot shows the 'Access tokens' generation page. The 'Developer' section in the sidebar is highlighted. The main area displays a message about using personal access tokens for secure authentication to the Databricks API. A 'Generate new token' button is highlighted with a green box. Below it, there are fields for 'Comment', 'Creation', and 'Expiration'.

Figure 8.23 – Generating New Token in Azure Databricks

The screenshot shows the 'Generate new token' dialog. It has fields for 'Comment' (containing 'What's this token for?') and 'Lifetime (days)' (set to 90). At the bottom are 'Cancel' and 'Generate' buttons, with 'Generate' highlighted with a green box.

Figure 8.24 – Creating a new Azure Databricks PAT

Generate New Token

Your token has been created successfully.

dapi417b46c3



A Make sure to copy the token now. You won't be able to see it again.

Done

Figure 8.25 – Copying the successfully generated token

	Name	SSN	email
1	Adam Smith	111-11-1111	james@james.com
2	Brenda Harman	222-22-2222	brenda@brenda.com
3	Carmen Pinto	333-33-3333	carmen@carmen.com

↓ 3 rows | 9.49 seconds runtime

Figure 8.26 – Output displaying PII, such as SSN

	Name	SSN
1	Adam Smith	45aK90bV3!@#5fTg^&*LpZxQwE4rVtBnUjMkO10p9o8i7u6y5t4r3Ml0pkejMEkOx4rVtBnUjkO10p9o8i7u6y52w1q
2	Brenda Harman	8fjR56#eD90bV3!@c4\$7vBnUjMkOxQwE4rVtBnUjMkO10p9o8i7u6y0bV5t4r3e2i7u6y5te2w1qZxQwE4rVtB^&*L
3	Carmen Pinto	XsP12vQz5&*1aK90bVZxQwE43!@#5fTg^&*LpZxQwE4rVtBkO10p9o8i7nUjMkO10p9o8i7u6y5t4r3e2i7u6y5t4r3e

↓ 3 rows | 0.44 seconds runtime

Figure 8.27 – Displaying output with encrypted PII

▶ decrypted: pyspark.sql.DataFrame = [Name: string, SSN: string]

	Name	SSN	email
1	Adam Smith	111-11-1111	james@james.com
2	Brenda Harman	222-22-2222	brenda@brenda.com
3	Carmen Pinto	333-33-3333	carmen@carmen.com

Figure 8.28 – Displaying output with decrypted PII

The screenshot shows the 'mySampleDataWarehouse' database in the Azure Synapse Analytics portal. The left sidebar includes sections for Security, Auditing, Data Discovery & Classification (which is selected and highlighted in green), Dynamic Data Masking, Microsoft Defender for Cloud, and Transparent data encryption. Under Common Tasks, there are links for Open in Visual Studio, Monitoring, Query activity, and Alerts. The main content area displays a table titled '8 columns with classification recommendations'. The table has columns for Schema, Table, Column, Information type, and Sensitivity label. The data shown is:

Select all	Schema	Table	Column	Information type	Sensitivity label
<input type="checkbox"/>	dbo	TempDriver	driveId	National ID	Confidential - GDPR
<input type="checkbox"/>	dbo	TempDriver	firstName	Name	Confidential - GDPR
<input type="checkbox"/>	dbo	TempDriver	lastName	Name	Confidential - GDPR

Figure 8.29 – Displaying Data Discovery & Classification in Synapse SQL

The screenshot shows the 'Practice Resources' page for 'DASHBOARD 2: CHAPTER 8'. The title is 'Azure Governance and Compliance'. The 'Summary' section states: 'This chapter included complete coverage of the following AZ-900 Azure Fundamentals exam objectives: Describe Azure management and governance.' It also notes that the chapter introduced the purpose of Microsoft Purview in Azure, the purpose of Azure Policy, resource locks, and tags, the Azure service lifecycle, Microsoft Cloud Adoption Framework, and the Port Center. A note says, 'Further knowledge beyond required exam content was provided to prepare for real-world, day-to-day Azure usage.' Below this is a section titled 'Chapter Review Questions' with a link to 'The Microsoft Azure Fundamentals Certification and Beyond - Second Edition by Trish Miller'. A 'Select Quiz' button is present, with 'Quiz 1' and 'HOW IT ALL DETAILS...' options.

Figure 8.31 – Chapter Review Questions for Chapter 8

Chapter 9: Monitoring Data Storage and Data Processing

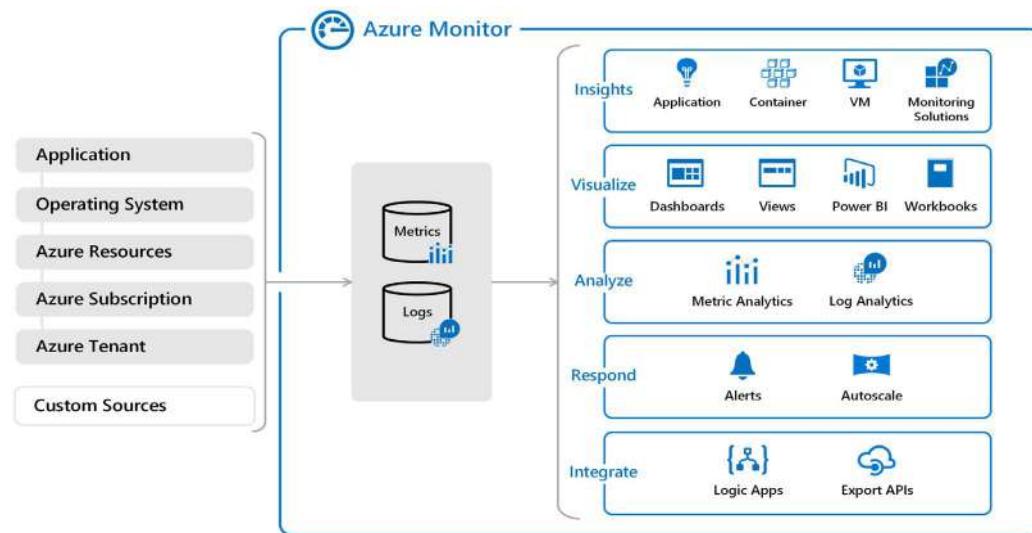


Figure 9.1 – Azure Monitor architecture showing the core components of Azure Monitor, Metrics, and Logs

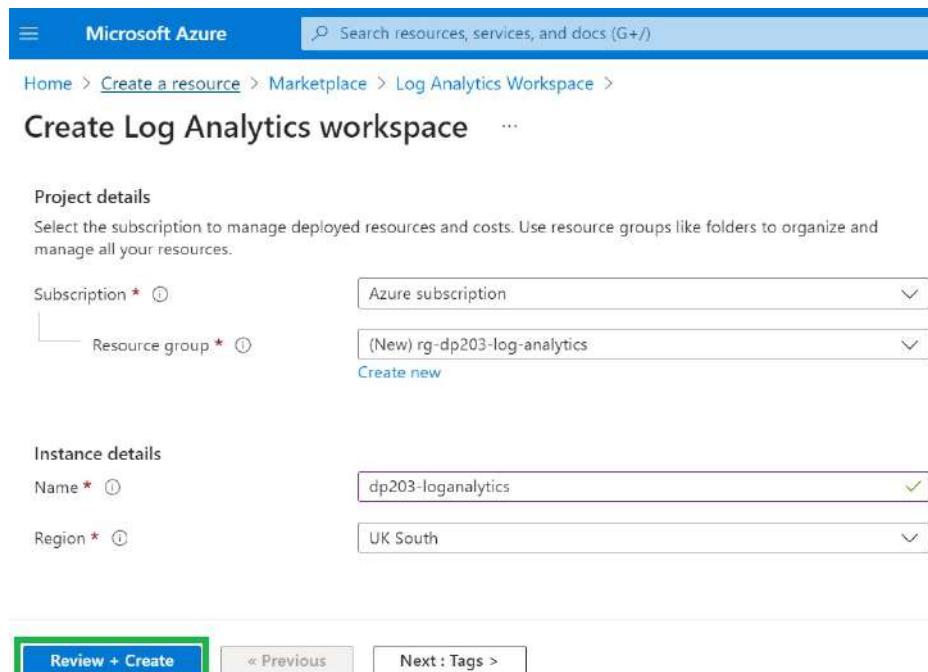


Figure 9.2 – Choosing a subscription to deploy resources during Log Analytics workspace creation

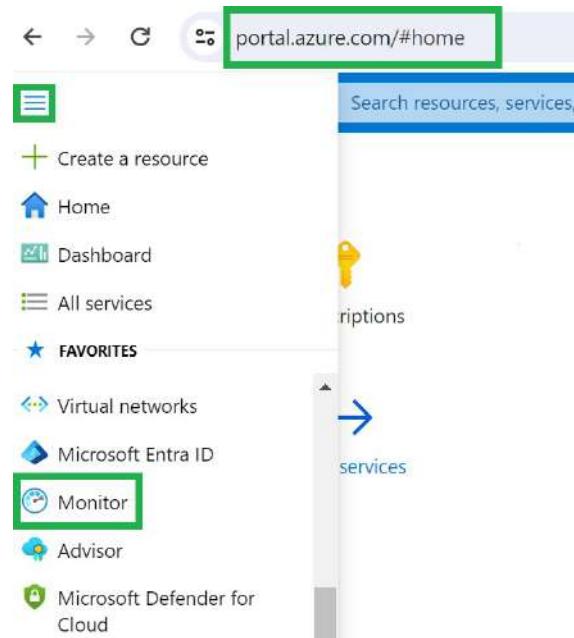


Figure 9.3 – Selecting the Monitor service from the Azure portal

A screenshot of the 'Monitor | Activity log' screen in the Azure portal. The top navigation bar shows 'Microsoft Azure' and a search bar. Below it, the breadcrumb navigation shows 'Home > Monitor'. The main area has a title 'Monitor | Activity log' with a back arrow and three dots. On the left is a sidebar with links: Overview, Activity log (which is highlighted with a green box), Alerts, Metrics, Logs, Change Analysis, Service health, Workbooks, Insights, Applications, and Virtual Machines. The main content area shows a table of activity logs with 14 items. The columns are: Operation name, Status, Event categ..., Time, Time stamp, and Subscription. The first few rows show operations like 'Delete existing', 'List changes of', and 'Create new OM', all with 'Succeeded' status and timestamped as 'an hour ago' or '2 hours ago' on 'Thu Apr 18 ...'. There are also buttons for 'Search', 'Activity', 'Edit columns', 'Refresh', 'Export Activity Logs' (which is highlighted with a green box), and 'Download as CSV'.

Figure 9.4 – Azure Monitor Activity log screen showing all the activity logs

Diagnostic settings

The screenshot shows the 'Diagnostic settings' screen in the Azure portal. At the top, there's a 'Subscription' dropdown set to 'Azure subscription'. Below it, a note explains that diagnostic settings allow streaming export of logs and metrics to various destinations. A table lists existing diagnostic settings by name, storage account, event hub, log analytics workspace, partner solution, and edit setting. A green box highlights the '+ Add diagnostic setting' button. Below the table, a note says to click 'Add Diagnostic setting' to configure data collection for categories like Administrative, Security, ServiceHealth, Alert, Recommendation, Policy, Autoscale, and ResourceHealth.

Figure 9.5 – Configuring logs in the Diagnostic settings screen

Diagnostic setting

The screenshot shows the 'Diagnostic setting' configuration screen. It has tabs for 'Logs' and 'Metrics'. Under 'Logs', a 'Diagnostic setting name' is set to 'LogManagement'. The 'Logs' section shows a list of categories: Administrative, Security, ServiceHealth, Alert, Recommendation, Policy, Autoscale, and ResourceHealth, all of which are checked. The 'Destination details' section is also highlighted with a green box. It includes a checkbox for 'Send to Log Analytics workspace' (which is checked), a 'Subscription' dropdown set to 'Azure subscription', a 'Log Analytics workspace' dropdown set to 'dp203-loganalytics (ulysouth.)', and three unchecked options: 'Archive to a storage account', 'Stream to an event hub', and 'Send to partner solution'.

Figure 9.6 – Selecting a list of categories for the Log Analytics workspace in Diagnostic setting

The screenshot shows the Azure Log Analytics workspace titled "dp203-loganalytics | Logs". The left sidebar has a "Logs" section highlighted with a green box. The main area displays a query titled "New Query 1" with the table "AzureActivity" selected. The "Run" button is highlighted with a blue box. The results pane shows log entries with columns: TimeGenerated [UTC], OperationNameValue, Level, and Activity. The log entries are:

TimeGenerated [UTC]	OperationNameValue	Level	Activity
18/04/2024, 15:25:15.452	MICROSOFT.INSIGHTS/DIAGNOSTICSETTINGS/WRITE	Error	Failure
18/04/2024, 15:25:15.358	MICROSOFT.INSIGHTS/DIAGNOSTICSETTINGS/WRITE	Information	Start
18/04/2024, 15:24:31.264	MICROSOFT.INSIGHTS/DIAGNOSTICSETTINGS/WRITE	Information	Success
18/04/2024, 15:24:29.733	MICROSOFT.INSIGHTS/DIAGNOSTICSETTINGS/WRITE	Information	Start
18/04/2024, 15:24:06.092	MICROSOFT.INSIGHTS/DIAGNOSTICSETTINGS/DELETE	Information	Success
18/04/2024, 15:24:04.858	MICROSOFT.INSIGHTS/DIAGNOSTICSETTINGS/DELETE	Information	Start

Figure 9.7 – The AzureActivity table displaying various log results in the Log Analytics workspace

The screenshot shows the Azure Storage account metrics configuration page for "iacstoreacct". The left sidebar has a "Metrics" section highlighted with a green box. The main area shows a line chart for the "Egress" metric over the last 24 hours. The chart has a Y-axis from 0 to 5000B and an X-axis from Thu 18 to UTC+01:00. The legend includes CAPACITY, TRANSACTION, and EGRESS metrics. The "Scope" dropdown is set to "iacstoreacct" and the "Metric Namespace" dropdown is set to "Account". The "Metric" dropdown is set to "Egress" and the "Aggregation" dropdown is set to "Sum".

Figure 9.8 – Configuring metrics for Azure Storage

The screenshot shows the Azure Monitor Storage accounts dashboard. The left sidebar has an "Insights" section highlighted with a green box. The main area displays metrics for "Azure subscription (2)" storage accounts: "iacstoreacct" and "synapseazdl". The metrics shown are Transactions, Transactions Timeline, E2E Latency, Server Latency, and ClientOtherError/Errors. The "Transactions" card shows values: 51, 26, 25. The "E2E Latency" card shows values: 11.06 ms, 8.67 ms, 8.58 ms. The "ClientOtherError/Errors" card shows values: 1, 1.

Figure 9.9 – The Monitor Storage account showing storage service metrics from the Azure Monitor service

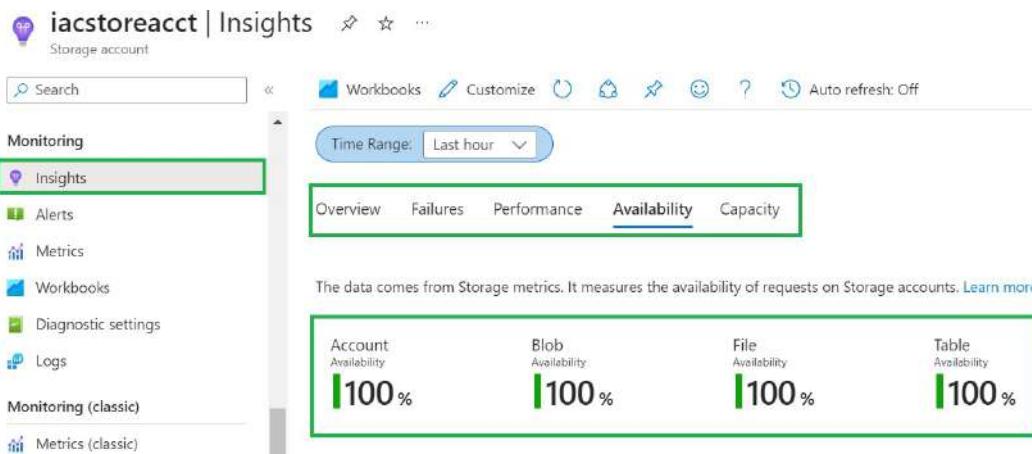


Figure 9.10 – Checking storage availability and status with Azure Monitor

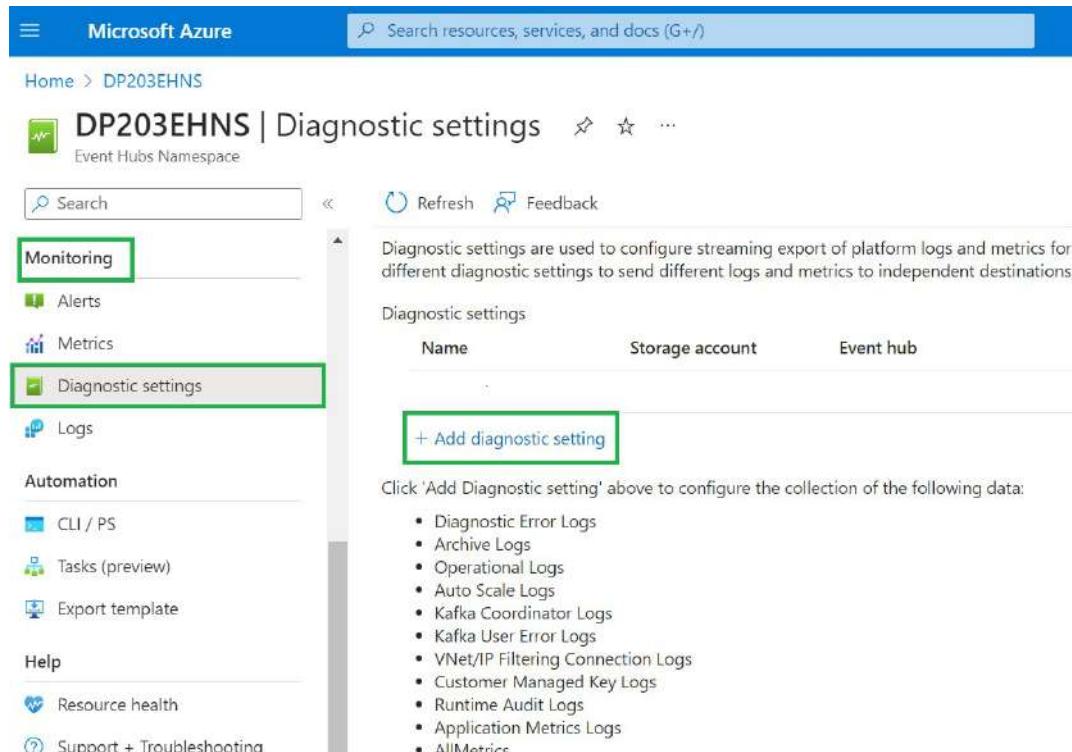


Figure 9.11 – Configuring new diagnostics for Event Hub

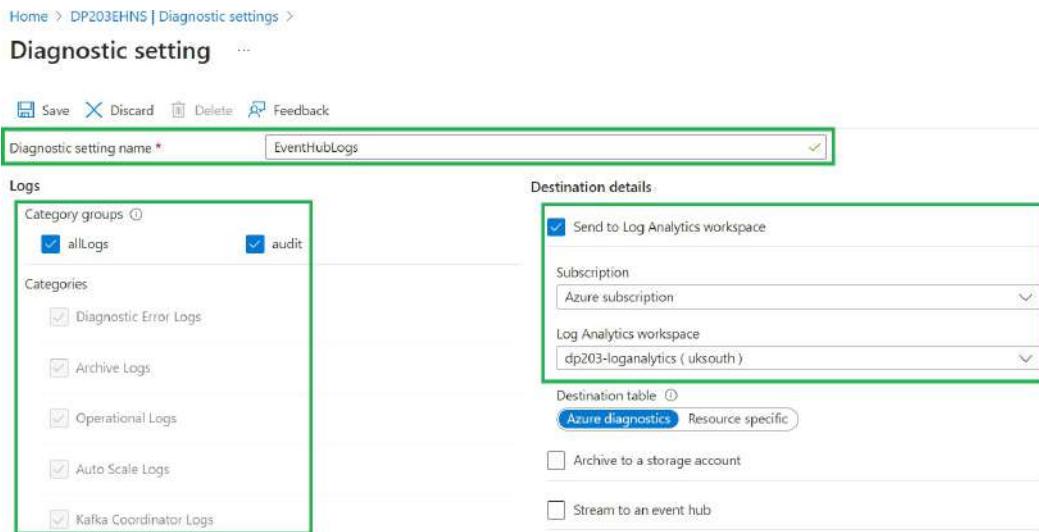


Figure 9.12 – Adding diagnostic settings for Event Hubs

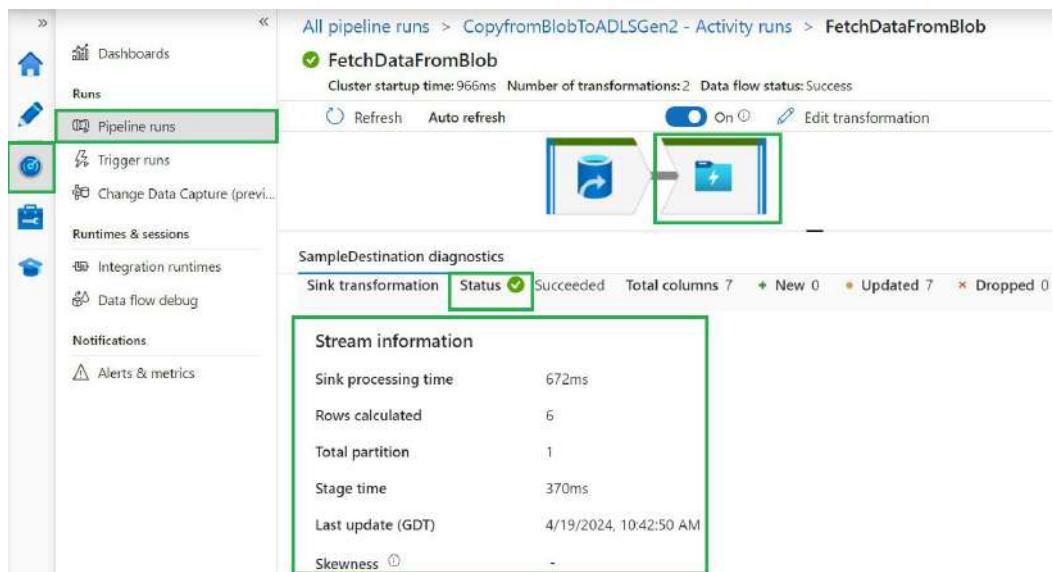


Figure 9.13 – Displaying data movement performance details

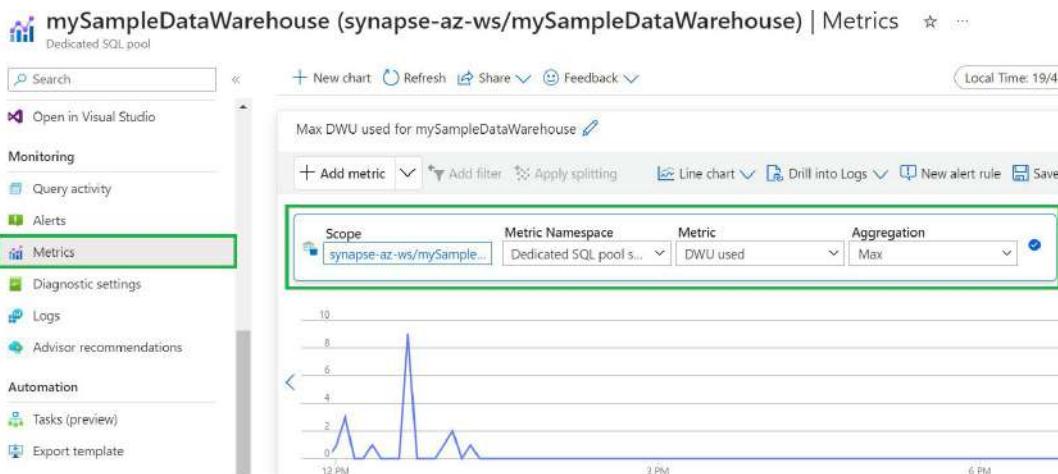


Figure 9.14 – Displaying metrics identifying performance regression in Synapse SQL pool

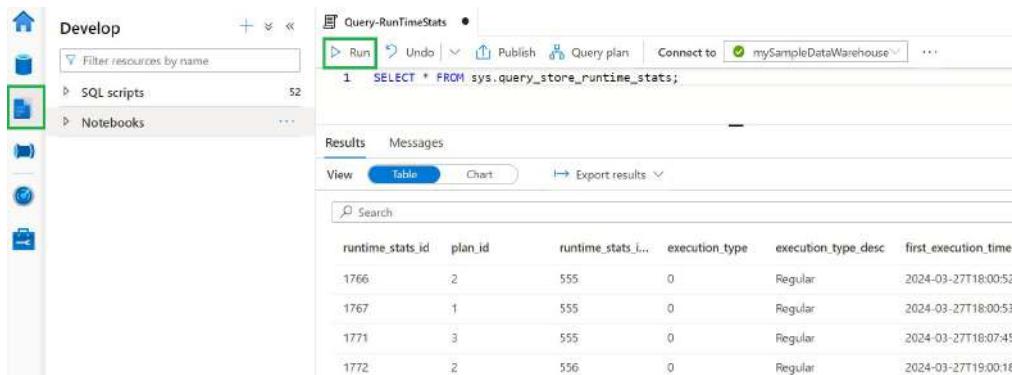


Figure 9.15 – Displaying a sample Query Store runtime statistics

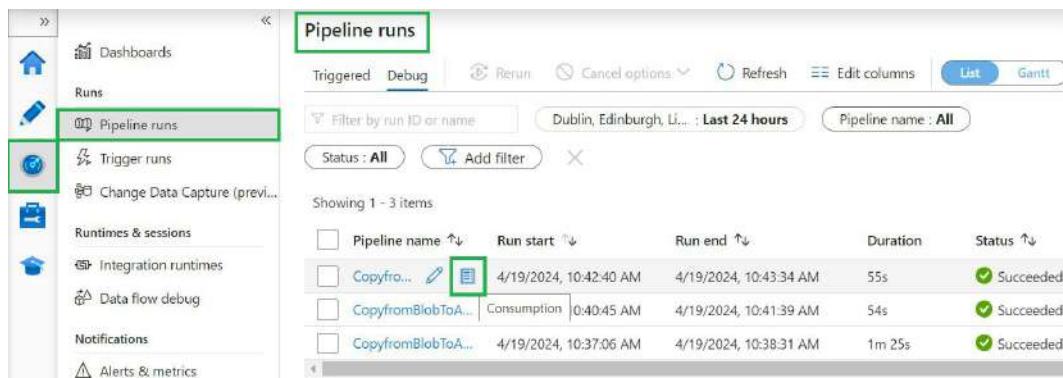


Figure 9.16 – The pipeline name showing a Consumption icon to display the details

Pipeline run consumption

Name

HDISSparkPipeline

Status

✓ Succeeded

Run ID

dd73e5a7-e7f4-405a-9eab-4519950abf05

	Quantity	Unit
Pipeline orchestration		
Activity runs	3	Activity runs
Pipeline execution		
Azure integration runtime	0.0667	DIU-hour
Data movement activities		
Data flow		
Data Flow	0.4330	vCore-hour

[Close](#)

Figure 9.17 – Resource consumption details screen showing the status of the resources



Figure 9.18 – Displaying additional pipeline details in the Gantt chart screen

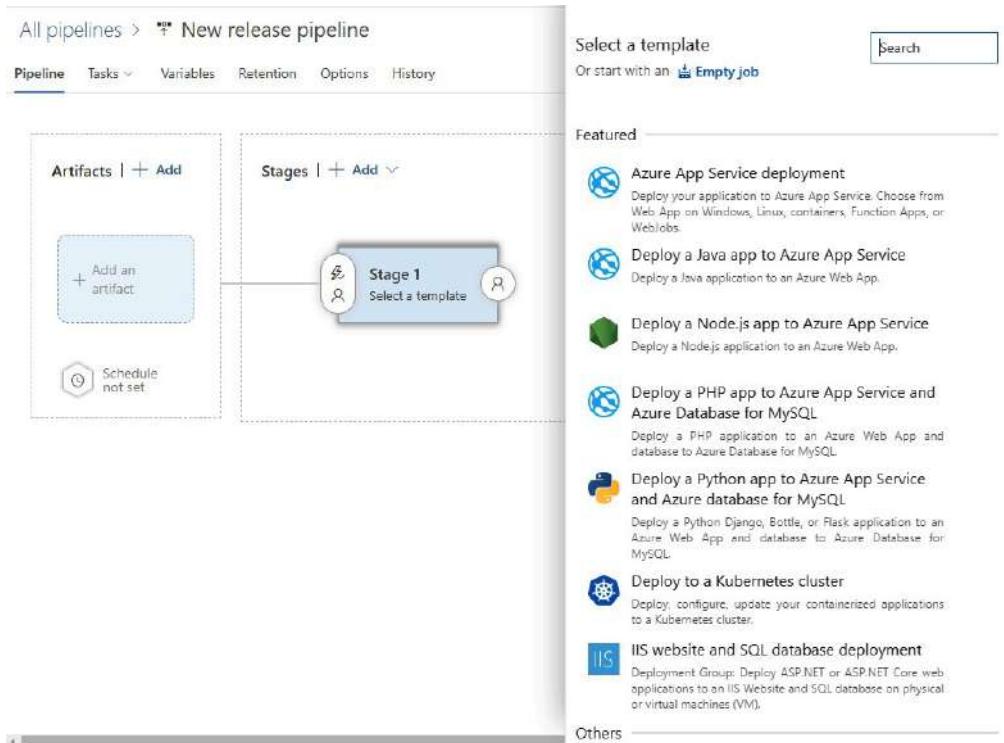


Figure 9.19 – The Azure release pipeline creation wizard creating an empty job

The screenshot shows the 'Add an artifact' interface in Azure Pipelines. On the left, there's a sidebar with 'Artifacts | + Add' and a button 'Add an artifact' which is highlighted with a green box. Below it is a note 'Schedule not set'. The main area has tabs for 'Source type' (selected), 'Build', 'GitHub', and 'TFVC'. Under 'Source type', there are fields for 'Project' (dp203), 'Source (repository)' (adf-dev), 'Default branch' (main), 'Default version' (Latest from the default branch), and checkboxes for 'Checkout submodules' and 'Checkout files from LFS'. There's also a 'Shallow fetch depth' field and a 'Source alias' field containing '_adf-dev'.

Figure 9.20 – Updating artifact information in Azure Pipelines

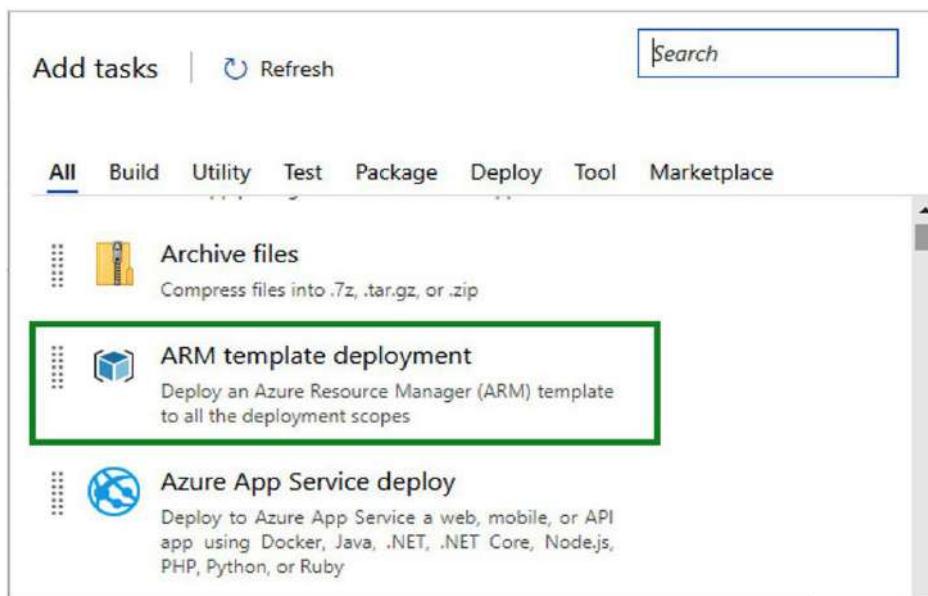


Figure 9.21 – Choosing ARM template deployment from the Add tasks screen

The screenshot shows the 'Template' configuration screen. It includes fields for 'Template location' (set to 'Linked artifact'), 'Template *' (a required field indicated by a red border), 'Template parameters' (a field with a red border), and 'Override template parameters' (another field with a red border). A message at the bottom left states 'This setting is required.'

Figure 9.22 – Specifying the ARM template and template parameters

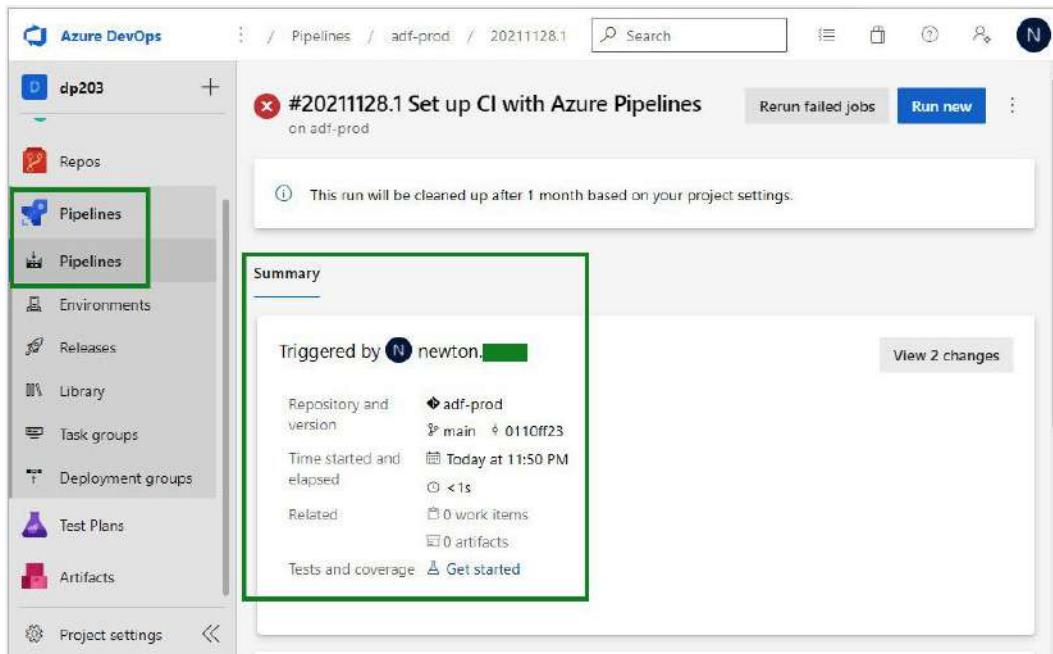


Figure 9.23 – The pipelines CI/CD monitoring screen showing the monitoring process

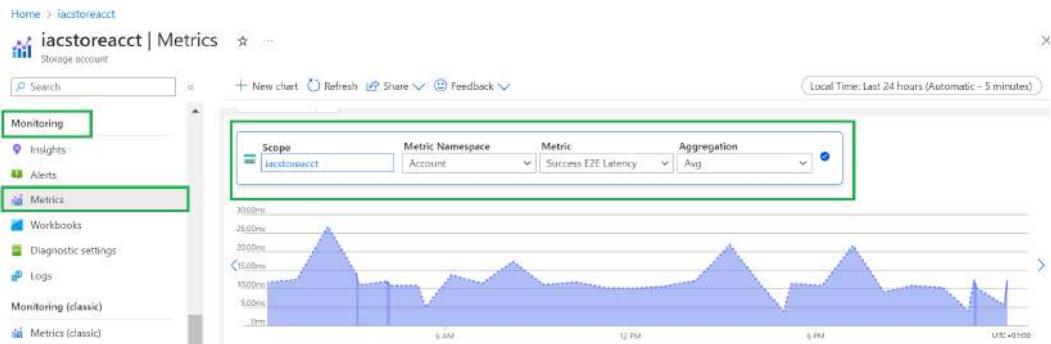


Figure 9.24 – Displaying Metrics data for a storage account collected from Azure resources

dp203-loganalytics | Logs

Log Analytics workspace

Search

Overview

Activity log

Access control (IAM)

Tags

Diagnose and solve problems

Logs

Settings

Tables

Agents

Usage and estimated costs

Data export

Network isolation

Linked storage accounts

SampleQuery*

dp203-loganalytics Select scope

Tables Queries Functions ...

Search

Filter Group by: Solution

Collapse all

Favorites

You can add favorites by clicking on the star icon

LogManagement

- AzureActivity
- AzureDiagnostics
- AzureMetrics
- LAQueryLogs

Figure 9.25 – Running the Log Analytics tables to view the different logs from each one

dp203-loganalytics | Logs

Log Analytics workspace

Search

Overview

Activity log

Access control (IAM)

Tags

Diagnose and solve problems

Logs

Settings

Tables

Agents

Usage and estimated costs

Data export

Network isolation

SampleQuery*

dp203-loganalytics Select scope

Tables Queries Functions ...

Search

Filter Group by: Solution

Run Time range: Last 24 hours

```

1 LAQueryLogs | 
2 take 100 | 
3 where StatsCPUTimeMs > 20 | 
4 sort by ResponseDurationMs | 
5 project TimeGenerated, QueryText, ResponseRowCount, ResponseDurationMs
6

```

Feedback Queries

Results Chart

TimeGenerated [UTC]	QueryText	ResponseRowCount	ResponseDurationMs
20/04/2024, 18:49:36.279	Usage	41	1912
20/04/2024, 18:48:34.747	LAQueryLogs tak...	3	1711
20/04/2024, 18:47:37.936	LAQueryLogs tak...	3	542
20/04/2024, 18:47:36.000	LAQueryLogs tak...	3	402

Figure 9.26 – Kusto query statements displaying the table contents

Home > synapse-az-ws | Alerts > Create an alert rule

Create an alert rule

Scope Condition Actions Details Tags Review +

Configure where the alert rule should trigger by selecting a signal and defining the conditions.

Signal name * Select a signal

Search by signal name Signal type : All Signal source : All

Signal name	Signal source
Log search	
Custom log search	Log Analytics
Resource health	
Resource health	Resource health
Metrics	
Activity runs ended	Platform metrics
Backlogged input events (preview)	Platform metrics
Data conversion errors (preview)	Platform metrics

Figure 9.27 – Creating an alert rule for Synapse pipelines

Home > synapse-az-ws | Alerts > Create an alert rule >

Create action group

An action group invokes a defined set of notifications and actions when an alert is triggered. [Learn more](#)

Project details

Select a subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources.

Subscription <input type="text"/>	Azure subscription <input type="button" value="▼"/>
Resource group * <input type="text"/>	rg-az-synapse <input type="button" value="▼"/> Create new
Region * <input type="text"/>	Global <input type="button" value="▼"/>

Instance details

Action group name * <input type="text"/>	Alert-Action <input type="button" value="✓"/>
Display name * <input type="text"/>	login alert <input type="button" value="✓"/>

Figure 9.28 – Creating an action group for an alert rule in the Synapse workspace

The screenshot shows the Microsoft Learn Practice Resources interface. At the top, there are navigation links for 'Practice Resources', 'Dashboard', 'CHAPTER 9', and 'Azure Governance and Compliance'. On the right side, there are icons for 'SHARE FEEDBACK', a bell, and a user profile.

CHAPTER 9: CHAPTER 9

Azure Governance and Compliance

Summary

This chapter is included in the Microsoft AZ-900 Azure Fundamentals exam. Chapter 9: Azure management and governance.

In this chapter, you'll learn about the purpose of Microsoft Purview in Azure, the purpose of Azure Policy, resource locks, and tags, the Azure service health blade, Microsoft Cloud Health periodic, and the Trust Center. You'll also learn how to use the Microsoft Purview blade to manage data classification and compliance.

Chapter Review Questions

The Microsoft Azure Fundamentals Certification and Beyond - Second Edition by Henry Maki

Select Quiz

Quiz 1 **START**

Figure 9.30 – Chapter Review Questions for Chapter 9

Chapter 10: Optimizing and Troubleshooting Data Storage and Processing

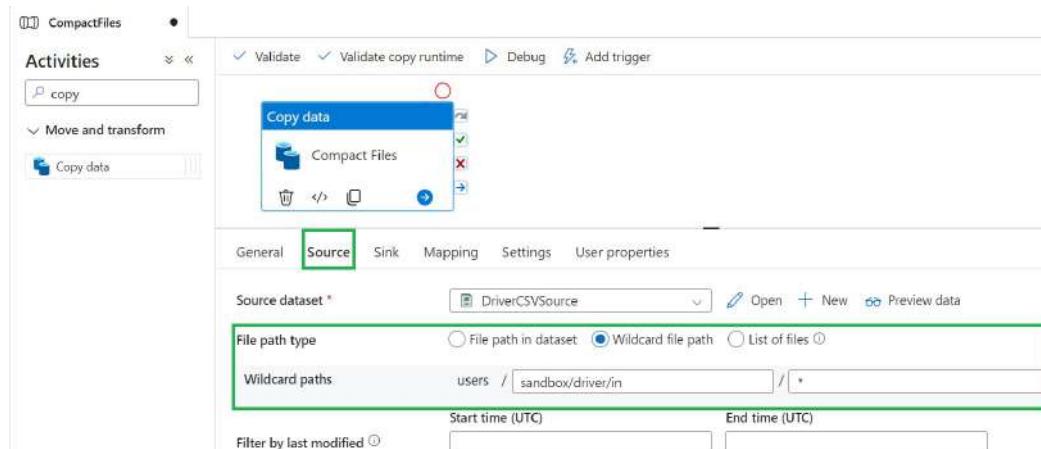


Figure 10.1 – Using wildcards to specify the source folder in Copy activity

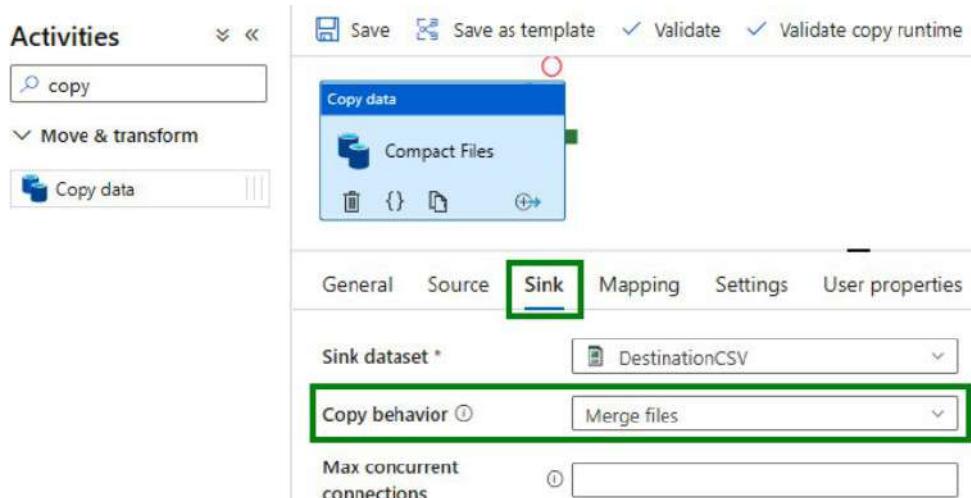


Figure 10.2 – The copying behavior of merging files in the Copy activity of ADF

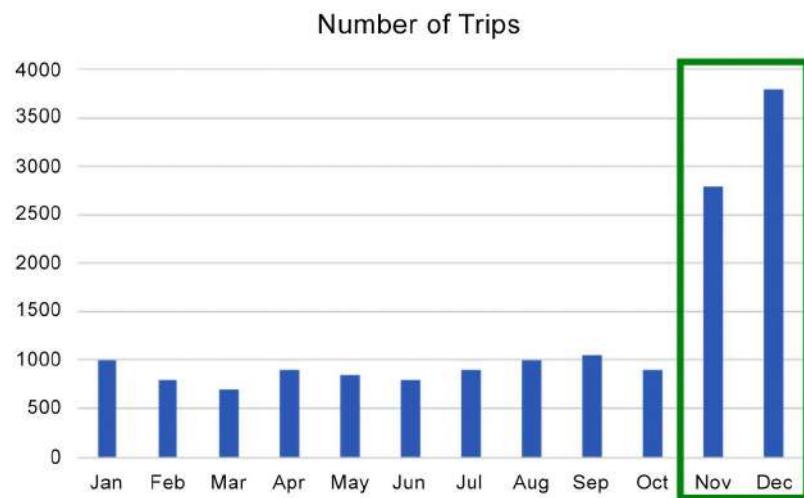


Figure 10.3 – Skewed data example showing the uneven distribution of trips per month

Microsoft Azure | Synapse Analytics > synapse-az-ws

Analytics pools

- SQL pools
- Apache Spark pools
- Data Explorer pools (preview)

Activities

- SQL requests
- KQL requests
- Apache Spark applications
- Data flow debug

Integration

- Pipeline runs
- Trigger runs
- Integration runtimes
- Link connections

Apache Spark applications > Livy ID 6

Read Trip Data

Completed tasks 9 of 9 Status Running Total duration 7m 24s

Cancel Refresh Spark UI

Attempts 0 of 0

All job IDs View Progress Playback

Job 0

Stage 0

Job 0

General

Progress: 100% Duration: 10 sec 625 ms Total tasks: 1

Data

Total rows: 1,000 Read: 128.0 KB Written: 0 bytes

Skew

- Data skew: None detected
- Time skew: None detected

View details

Stage 1

Job 1

General

Progress: 100% Duration: 20 sec 440 ms Total tasks: 8

Data

Rows: 4,609,034 Data read: 199.3 MB Data written: 19.5 MB

View details

Figure 10.4 – Data skew at different stages in Synapse Spark

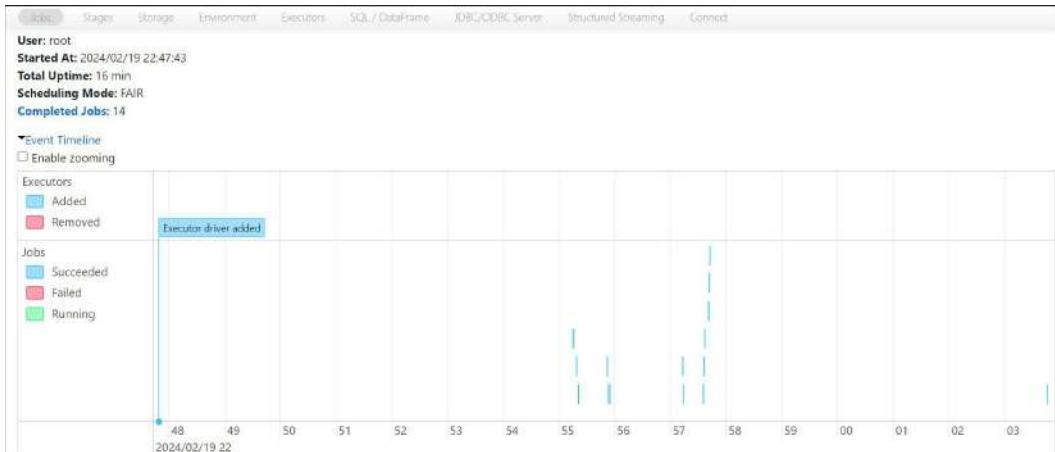


Figure 10.5 – Spark UI screen showing the status of the jobs

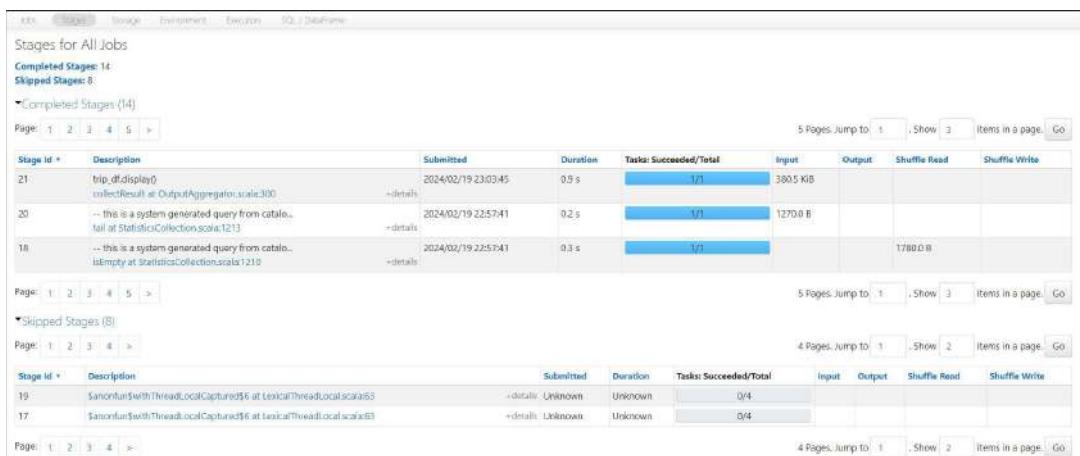


Figure 10.6 – Checking the stages and duration of all job stages

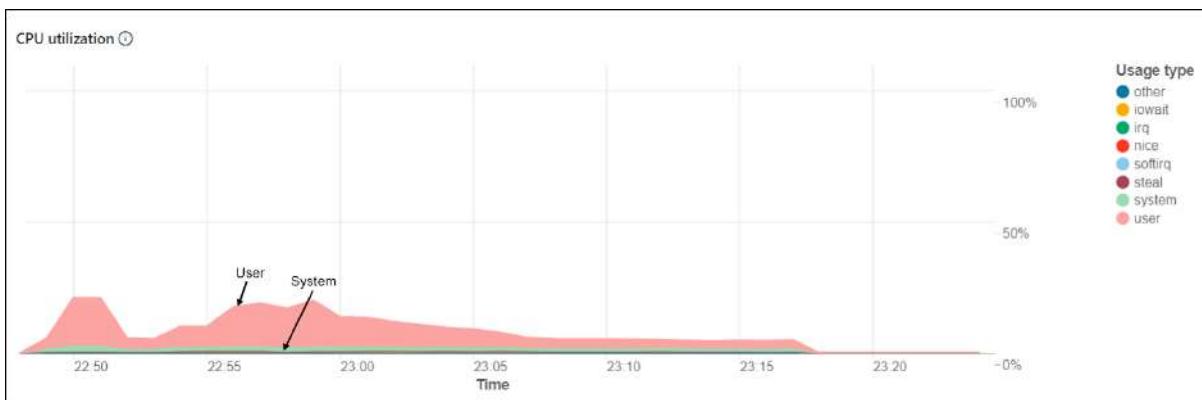


Figure 10.7 – The CPU utilization in cluster metrics

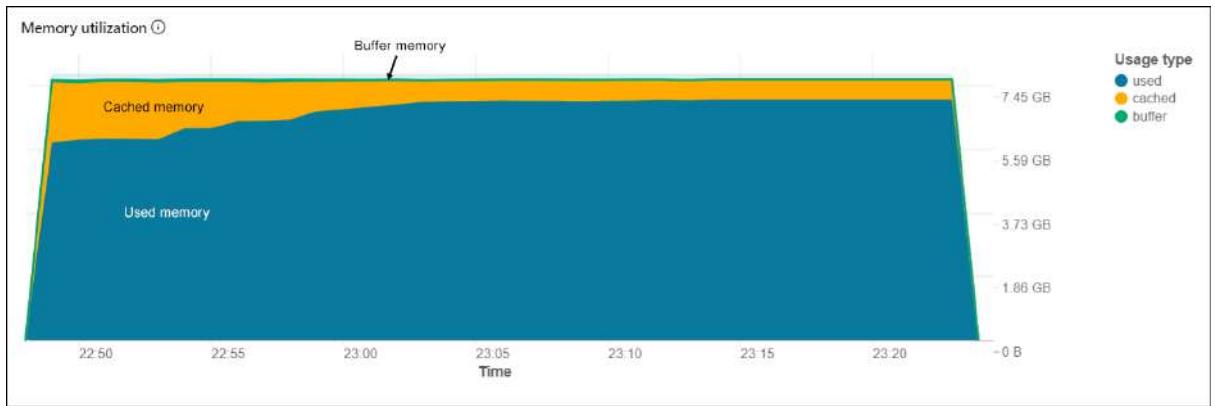


Figure 10.8 – Memory utilization in cluster metrics

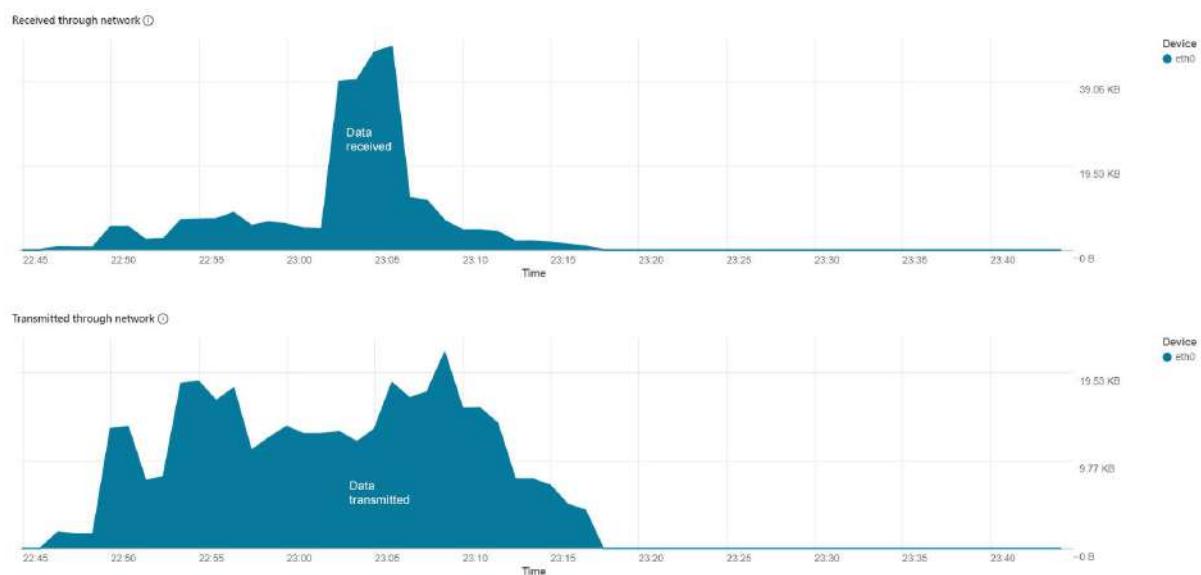


Figure 10.9 – Network usage in cluster metrics

Log file	Log type	Size
stderr--2024-02-19--23-00	Standard error	943 bytes
stderr--2024-02-21--12-00	Standard error	421 bytes
stderr	Standard error	0 bytes
stdout--2024-02-19--23-00	Standard output	104.11 KB
stdout--2024-02-21--12-00	Standard output	96.28 KB
stdout	Standard output	1.94 KB
log4j-2024-02-19-22.log.gz	Log4j output	39.01 KB
log4j-2024-02-21-11.log.gz	Log4j output	13.04 KB

Figure 10.10 – Displaying Driver logs location in an Azure Databricks cluster

Job Id (Job Group)	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
S-600170986151770660218_8249659138900634724_23400abce9547be9900ff33ec05f51	display[abce, df.read.select(explode(df.results))].collectResult at OutputAggregator.scala:90	2024/02/20 16:48:04	0.5 s	1/1	1/1
S-600170986151770660218_82027475285475644022_23400e02ef9547be9900ff133e005f53	display[abce, df].collectResult at OutputAggregator.scala:500	2024/02/20 16:46:03	0.2 s	1/1	1/1
4-6817848213675632351_76005140962580083_9eb5cf10f7540898fc0a93757c4d7886	abce, df = abce, df.read.withColumn("results", expr...).collectResult at OutputAggregator.scala:90	2024/02/20 16:46:03	0.8 s	1/1	1/1
3-600170986151770660218_7052180738371962126_23400a02ef9547be9900ff133e005f51	display[abce, df].collectResult at OutputAggregator.scala:90	2024/02/20 16:45:59	1 s	1/1	1/1
2-6817848213675632331_7070459013654405411_9eb5cf10f7540898fc0a93757c4d7886	display[abce, df.wad].collectResult at OutputAggregator.scala:500	2024/02/20 16:45:58	2 s	1/1	1/1
1-6817848213675632331_49316027593162127517_9eb5cf10f7540898fc0a93757c4d7886	abce, df.read(spark.read.option("inferSchema", ...).json).atNativeMethodAccessingimpl.json()	2024/02/20 16:45:57	0.2 s	1/1	1/1
0-600170986151770660218_8076457526750645161_23400abce9547be9900ff133e005f51	abce, df.read(spark.read.option("inferSchema", ...).json).atNativeMethodAccessingimpl.json()	2024/02/20 16:45:49	2 s	1/1	1/1

Figure 10.11 – Checking the status of a job in Spark job location within the Spark UI

The screenshot shows the Databricks Spark UI interface. On the left, there's a sidebar with various workspace sections like Workspace, Recents, Catalog, Workflows, and Compute, where Compute is selected. The main area has tabs for Configuration, Notebooks (4), Libraries, Event log, Spark UI (which is active and highlighted in green), Driver logs, Metrics, Apps, and Spark compute UI - Master. Below these tabs, there are sub-tabs: Data, Datasets, Storage, Environment, Executors (which is also highlighted in green), SQL / DataFrame, JDBC/JDBC Server, Ingested Datasets, and Cluster.

Executors

Summary

	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write	Excluded
Active(1)	0	17.7 KB / 1.1 GB	0.0 B	4	0	0	3	3	1.7 min (2 s)	9.2 KB	0.0 B	0.0 B	0
Dead(0)	0	0.0 B / 0.0 B	0.0 B	0	0	0	0	0	0.0 ms (0.0 ms)	0.0 B	0.0 B	0.0 B	0
Total(1)	0	17.7 KB / 1.1 GB	0.0 B	4	0	0	3	3	1.7 min (2 s)	9.2 KB	0.0 B	0.0 B	0

Executors

Executor ID	Address	Status	RDD Blocks	Storage Memory	Disk	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write	Add Time	Remove Time
driver	10.139.64.4:46227	Active	0	17.7 KB / 1.1 GB	0.0 B	4	0	0	3	3	1.7 min (2 s)	9.2 KB	0.0 B	0.0 B	2024-02-21 14:19:10	-

Figure 10.12 – Executor log location in the Spark UI

The screenshot shows the Databricks Spark UI interface. The Stages tab is active and highlighted in green. The main area displays completed stages. A table lists the stages with columns: Stage Id, Description, Submitted, Duration, Tasks: Succeeded/Total, Input, Output, Shuffle Read, and Shuffle Write.

Stage Id	Description	Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Read	Shuffle Write
6	display(<code>df = df.select("episode","counts")</code>) collectResult at OutputAggregator.scala:100	2024/02/20 16:45:04	0.4 s	1/1	3.1 KB			
5	display(<code>df2</code>) collectResult at OutputAggregator.scala:100	2024/02/20 16:45:03	0.2 s	1/1	3.1 KB			
4	<code>abc_df = abc_df.readWithColumn("results").exp_</code> <code>ectCollectResult at OutputAggregator.scala:100</code>	2024/02/20 16:45:03	0.4 s	1/1	3.1 KB			
3	display(<code>df1</code>) collectResult at OutputAggregator.scala:100	2024/02/20 16:45:09	1 s	1/1	3.1 KB			
2	display(<code>df.read</code>) collectResult at OutputAggregator.scala:100	2024/02/20 16:45:58	1 s	1/1	3.1 KB			
1	<code>abc_df.read.parquet("zipped_inferSchema")</code> join at NativeMethodAccessorsImpl.java:0	2024/02/20 16:45:57	0.1 s	1/1	3.1 KB			
0	<code>abc_df.read(sparkCreate("zipped_inferSchema"))</code> join at NativeMethodAccessorsImpl.java:0	2024/02/20 16:45:49	1 s	1/1	3.1 KB			

Figure 10.13 – Displaying task in each stage location in the Spark UI

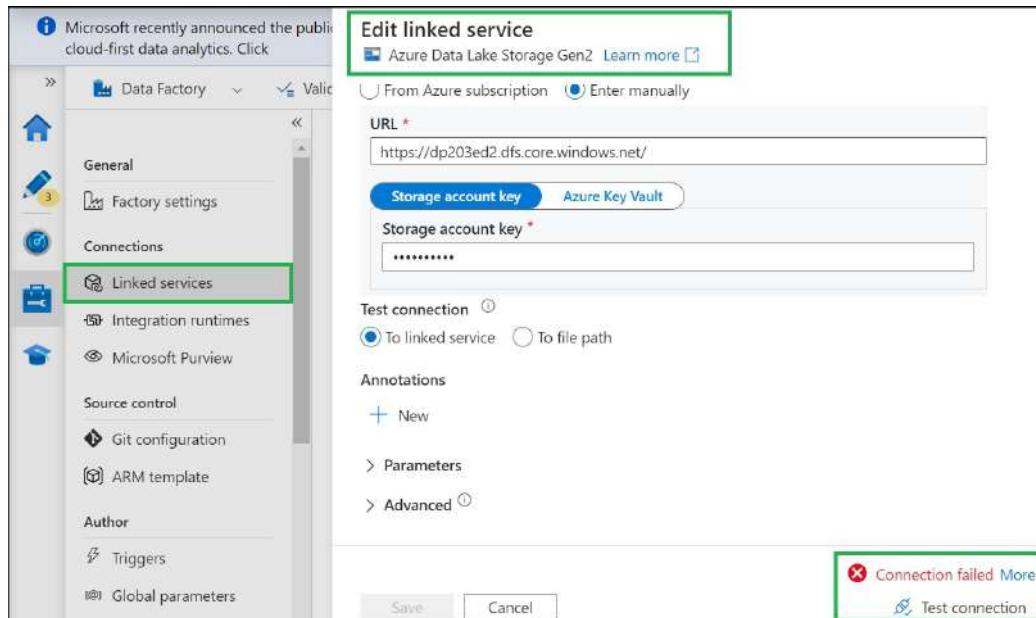


Figure 10.14 – Troubleshooting linked service connection

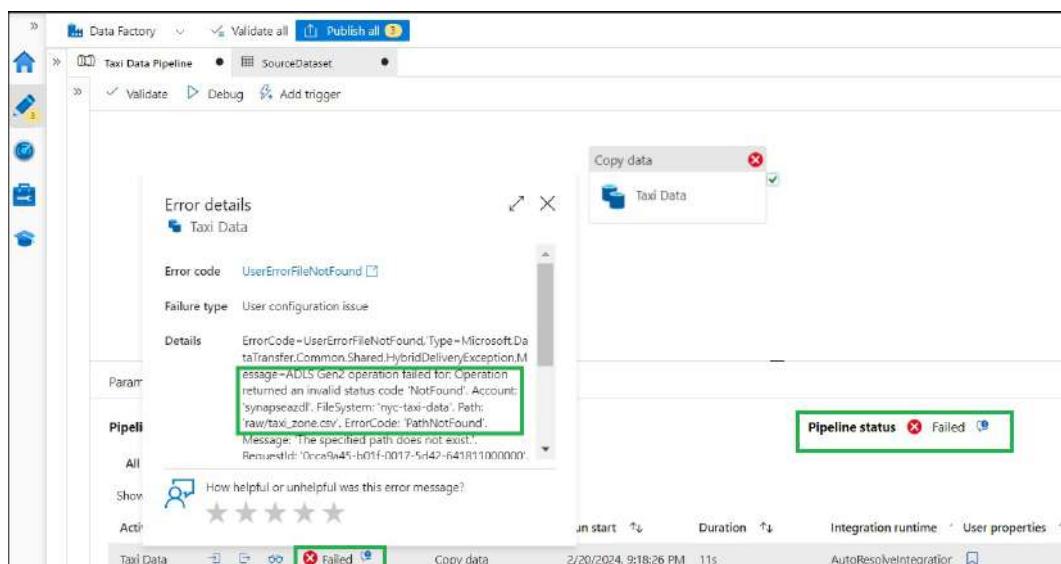


Figure 10.15 – Displaying an error when the Troubleshoot path is not found in the pipeline

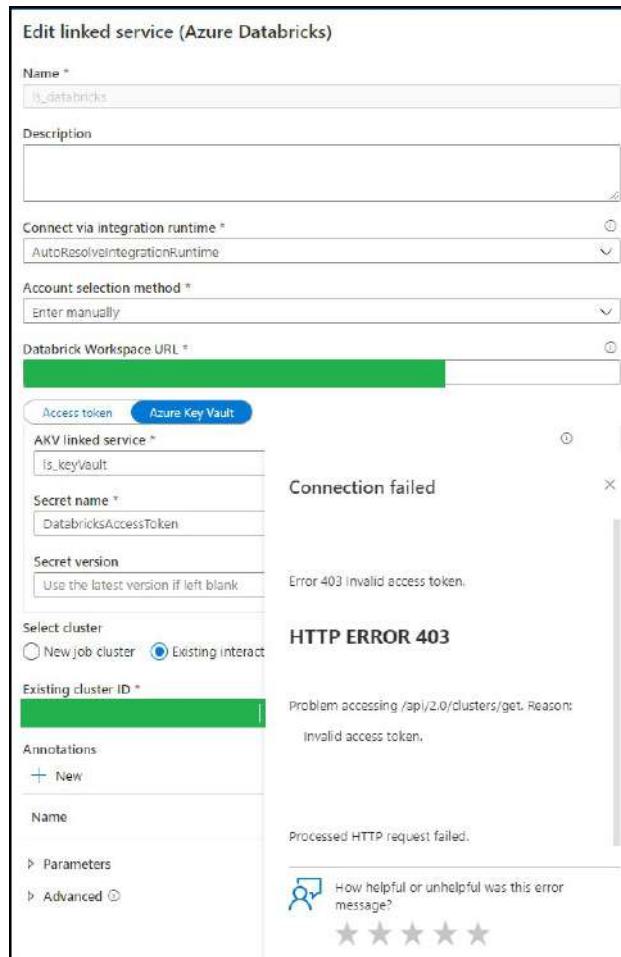


Figure 10.16 – Troubleshooting an access token issue for Azure Databricks service

The screenshot shows the Microsoft Learn Practice Resources dashboard for Chapter 10. The main content area displays the 'Azure Governance and Compliance' chapter summary. The summary includes a brief description of the chapter's purpose and a note about the next chapter. To the right, a 'Chapter Review Questions' sidebar is visible, showing a 'Select Quiz' dropdown menu with 'Quiz 1' and 'HOW IT WORKS' options, along with a 'START' button.

Figure 10.18 – Chapter Review Questions for Chapter 10

Chapter 11: Accessing the Online Practice Resources



Figure 11.2 – Unlock page for the online practice resources

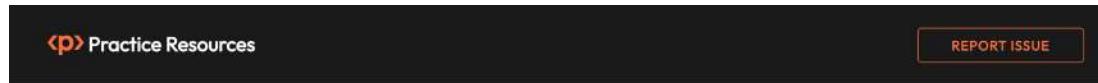


Figure 11.3 – Enter your unique sign-up code to unlock the resources



Figure 11.4 – Page that shows up after a successful unlock

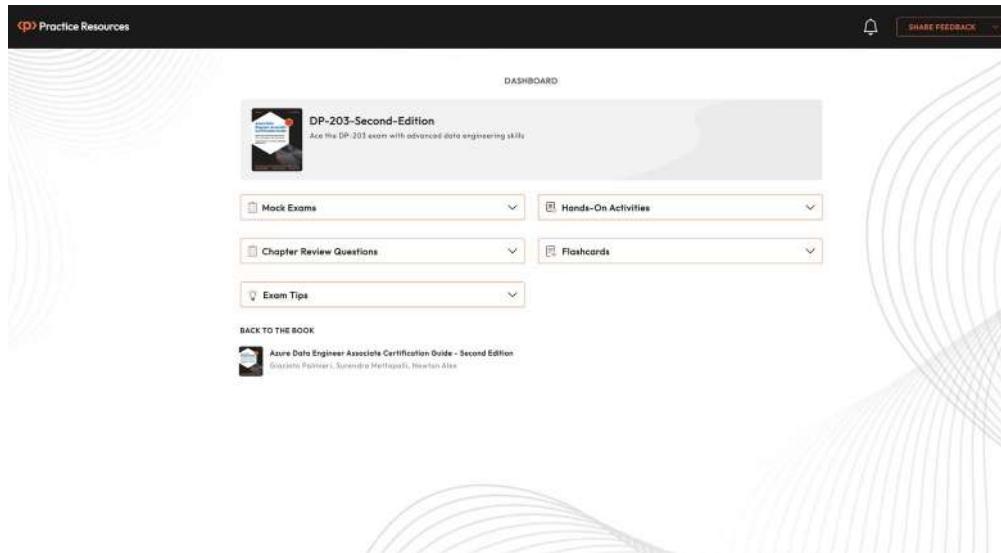


Figure 11.5 – Dashboard page for AZ-900 practice resources

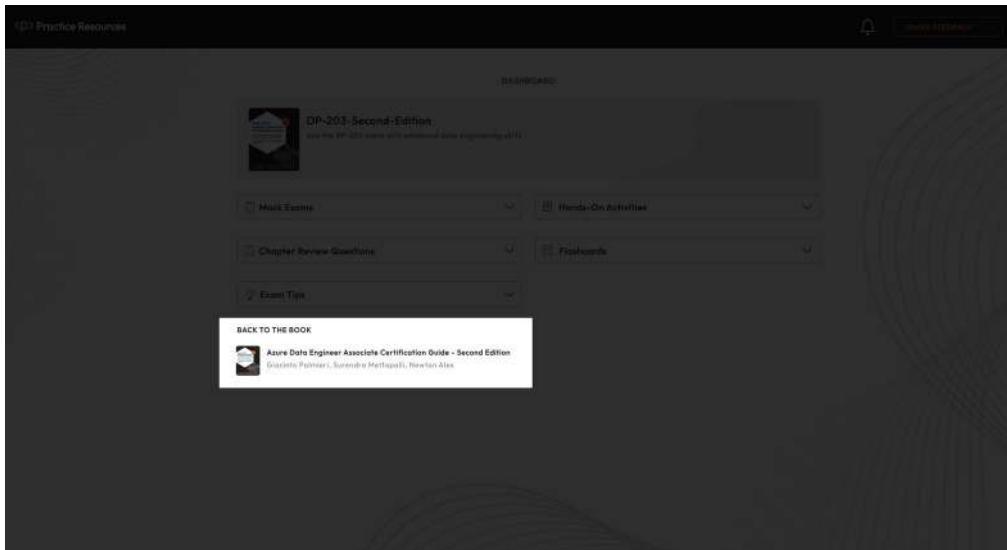


Figure 11.7 – Jump back to the book from the dashboard