Azure Data Engineer Associate Certification Guide, Second Edition

Preface:

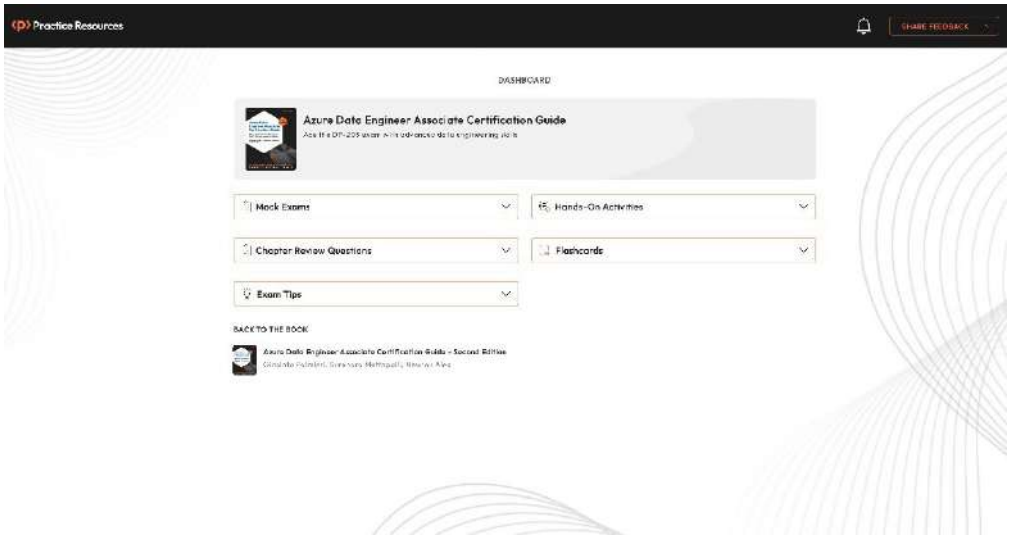

Figure 0.1 – Exam DP-203: Azure data engineer Associate online exam-prep platform

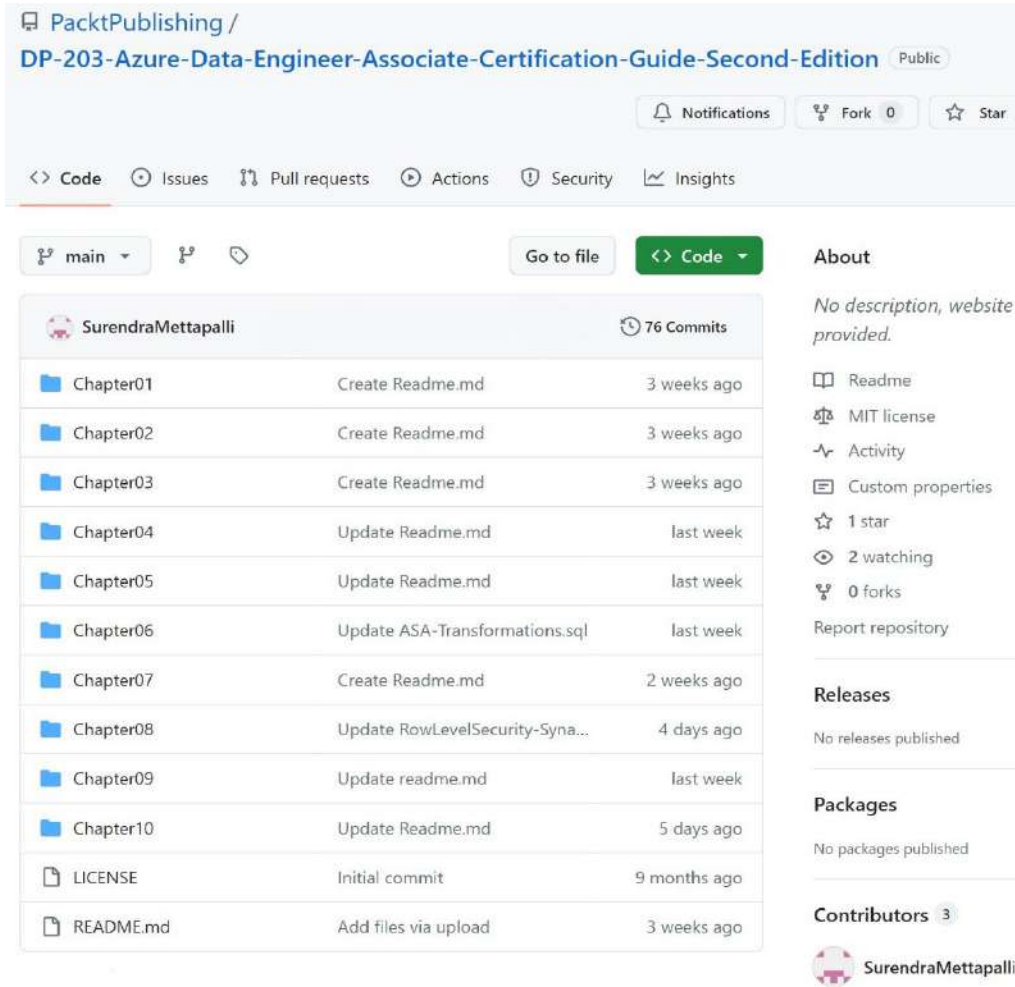

Figure 0.2 – GitHub code files for DP-203 certification

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.

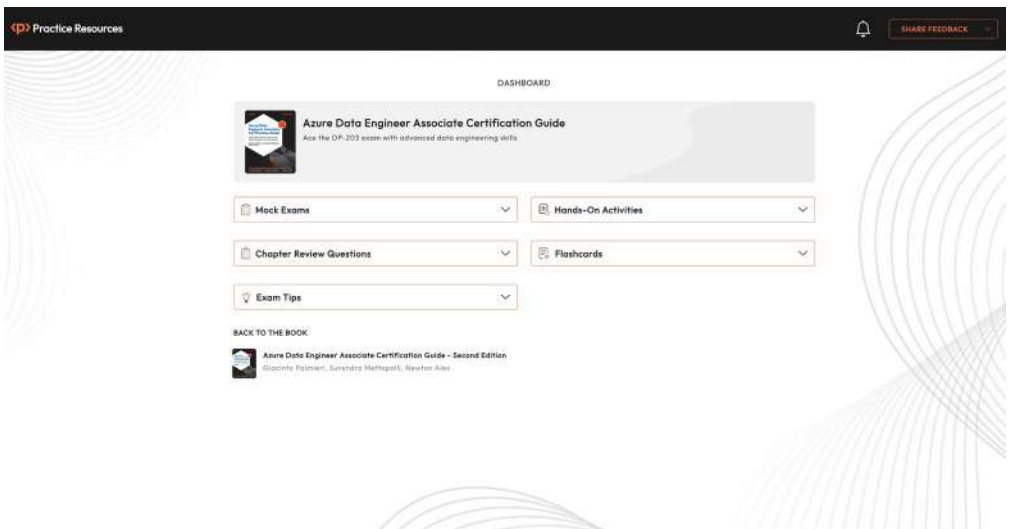

Figure 1.1: Dashboard interface of the online practice resources

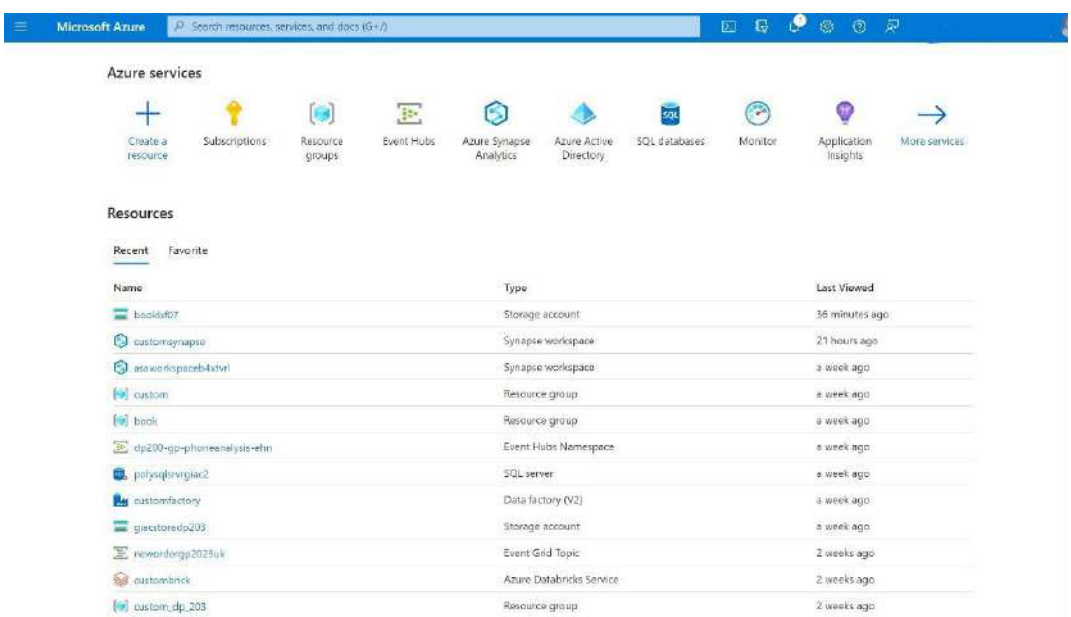

Figure 1.2 – The Azure portal home page

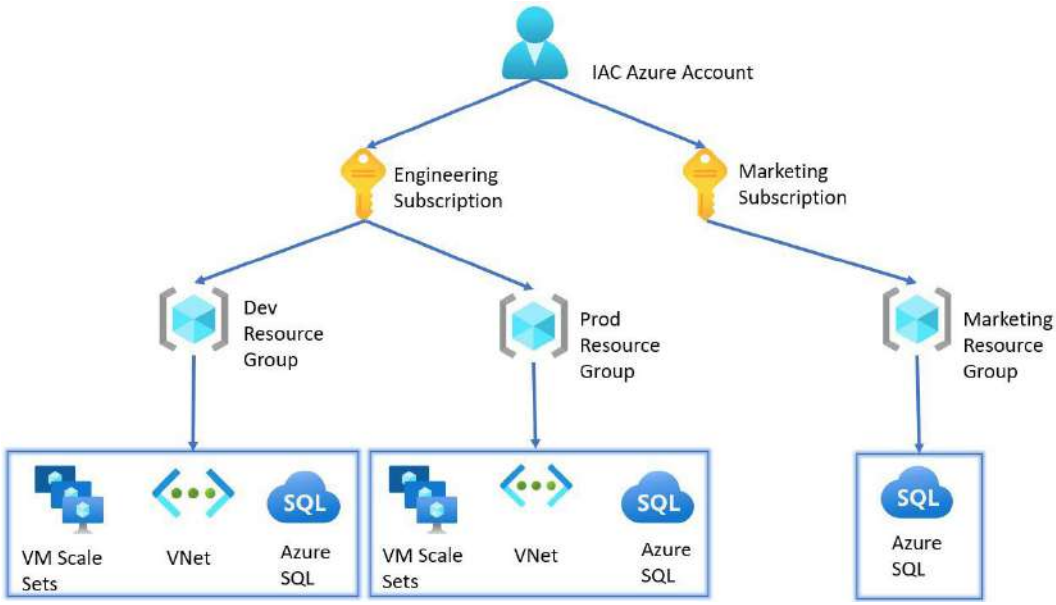

Figure 1.3 – Relationship between accounts, subscriptions, resource groups, and resources

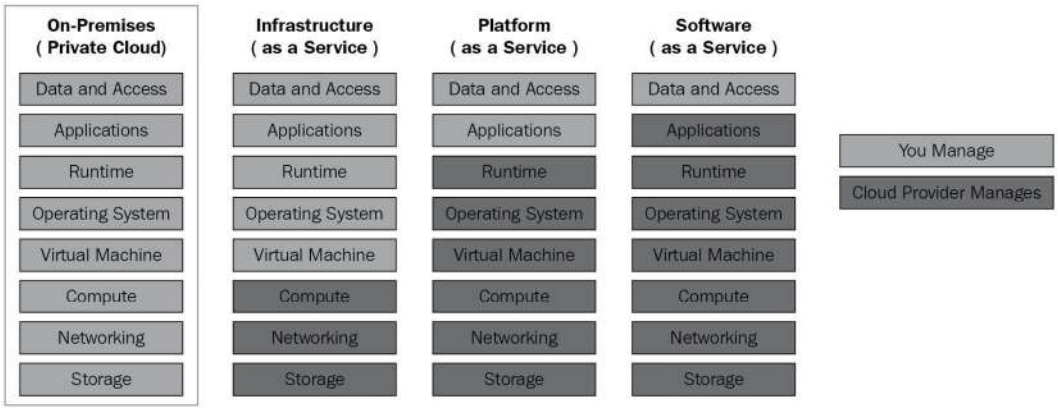

Figure 1.4 – Breakdown of Azure services

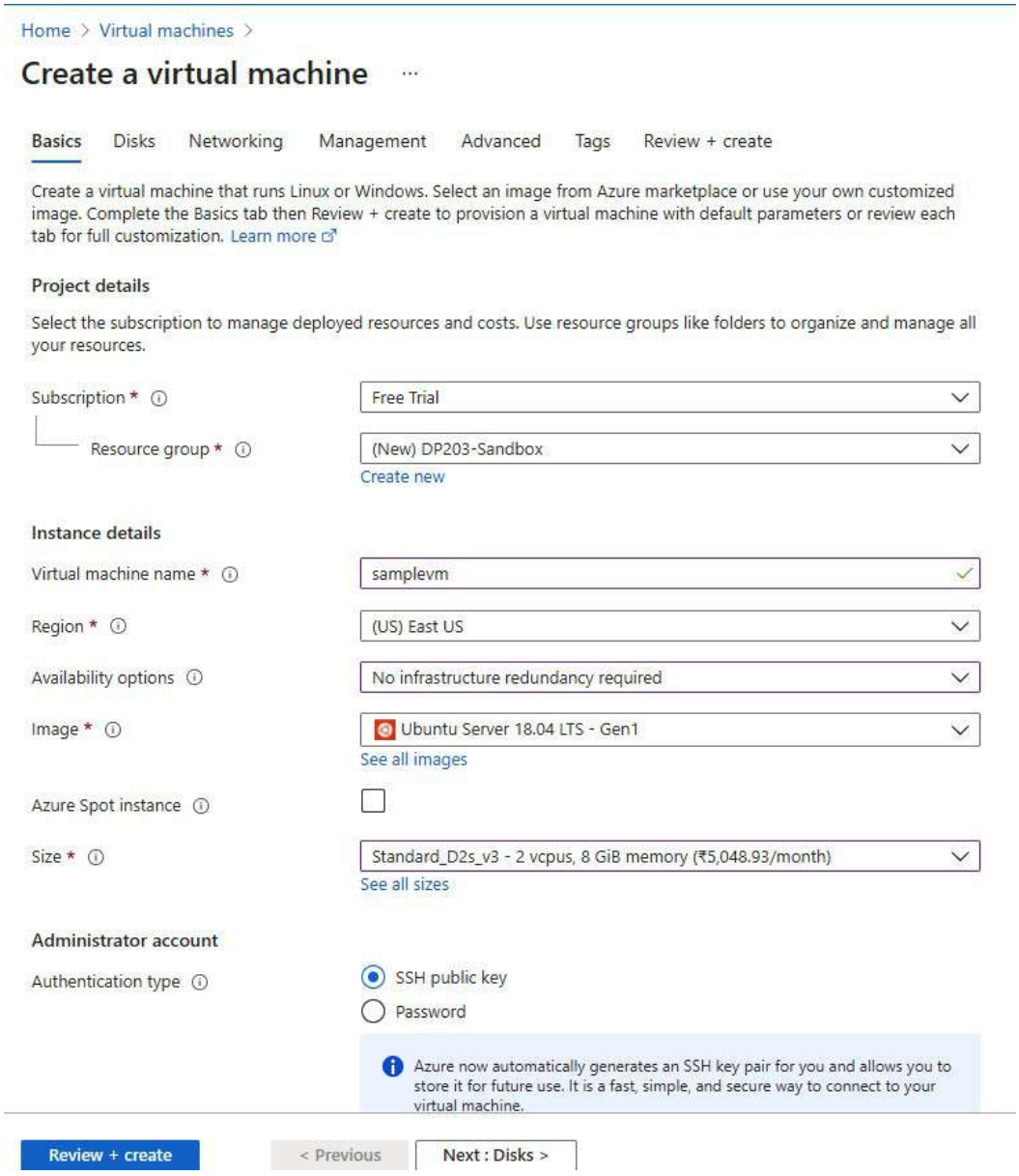

Figure 1.5 – Creating VMs using the Azure portal

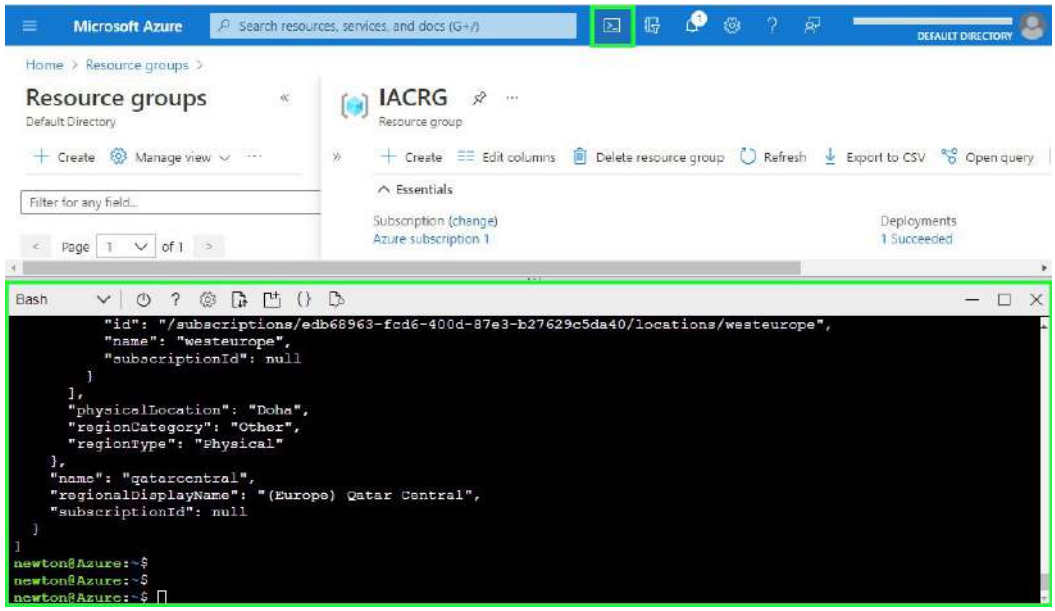

Figure 1.6 – Using the CLI directly from the Azure portal

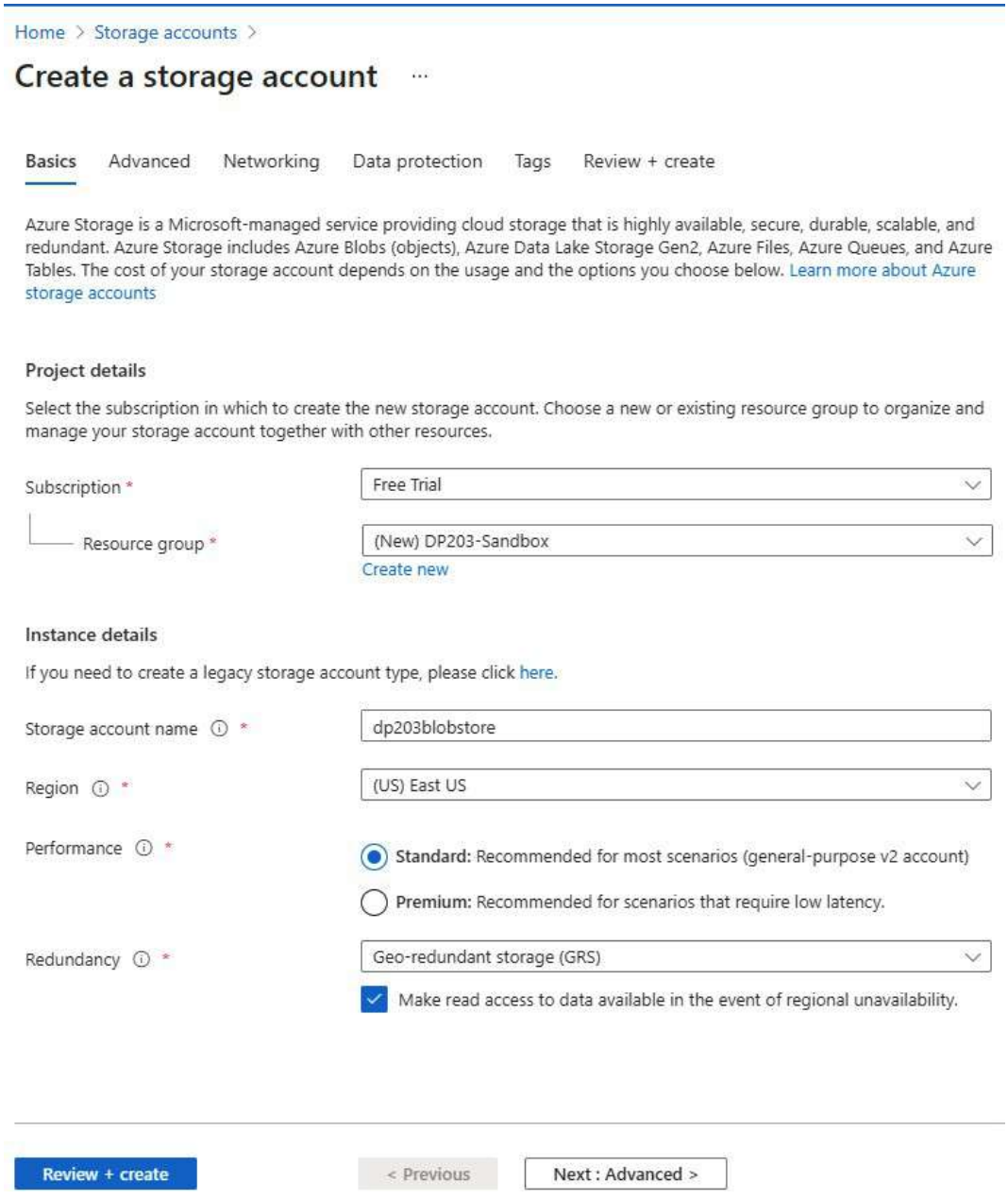

Figure 1.7 – Creating a storage account using the Azure portal

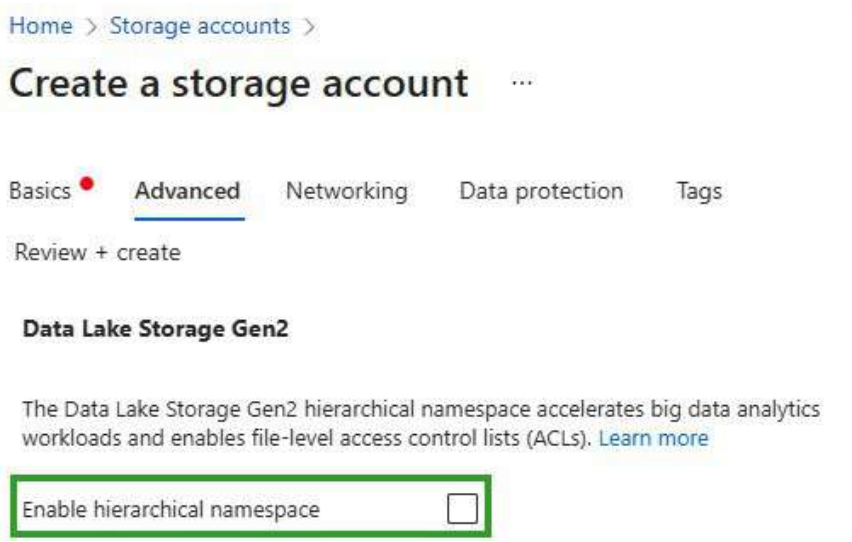

Figure 1.8 – Enable hierarchical namespace in Azure Storage for an ADLS Gen2 account

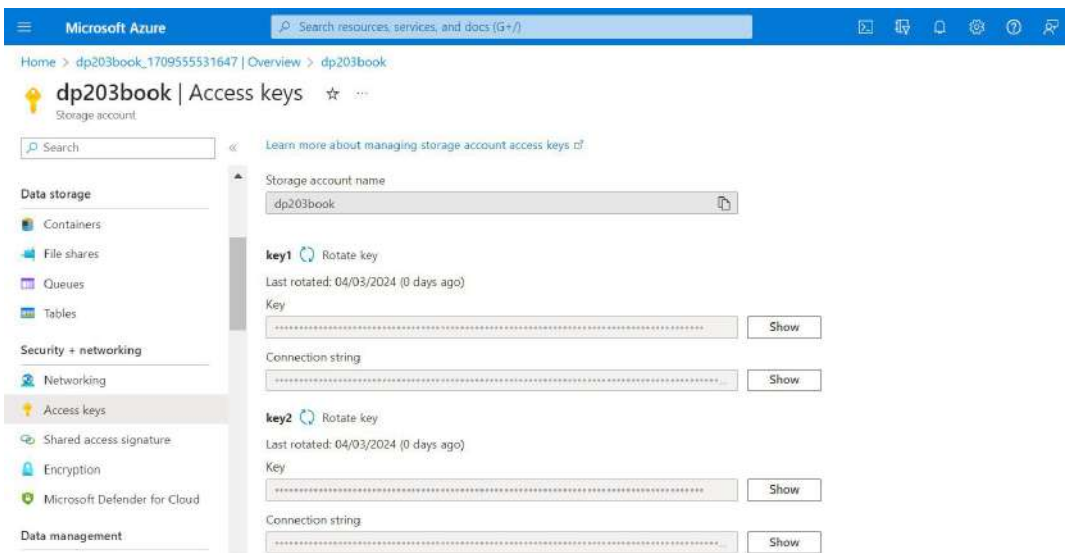

Figure 1.9 – Locating access keys for a storage account

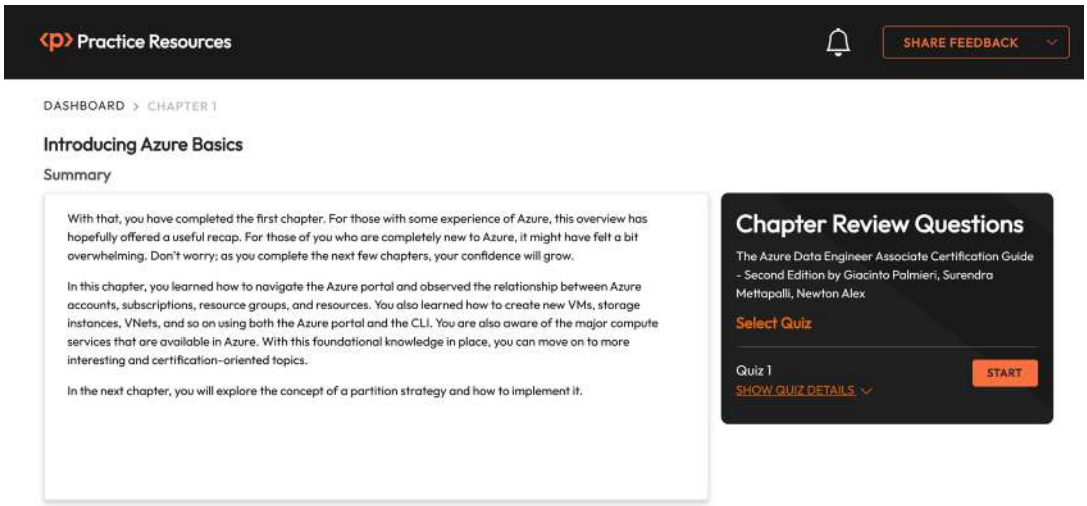

Figure 1.11 – Chapter Review Questions for Chapter 1

Chapter 2: Implementing a Partition Strategys

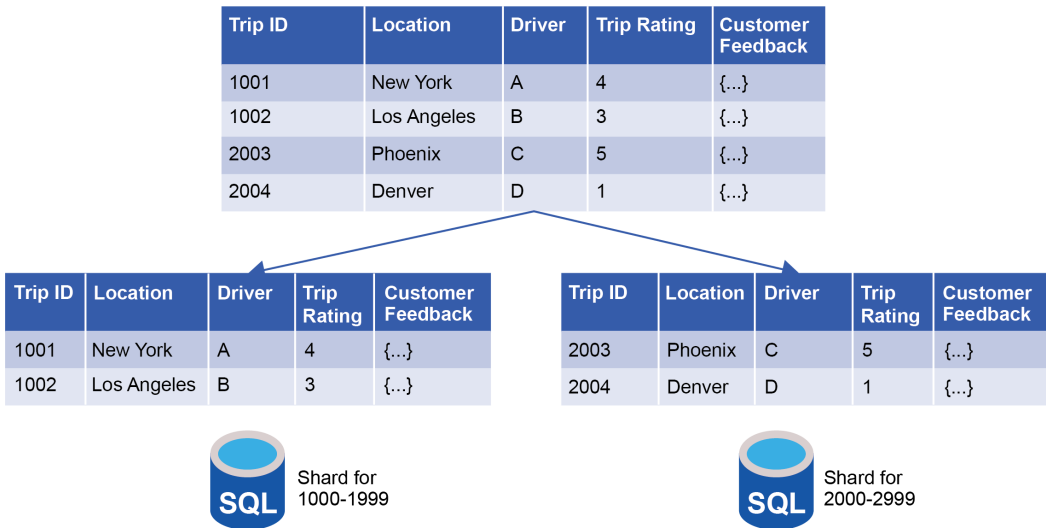

Figure 2.1 – An illustration of horizontal partitioning in action

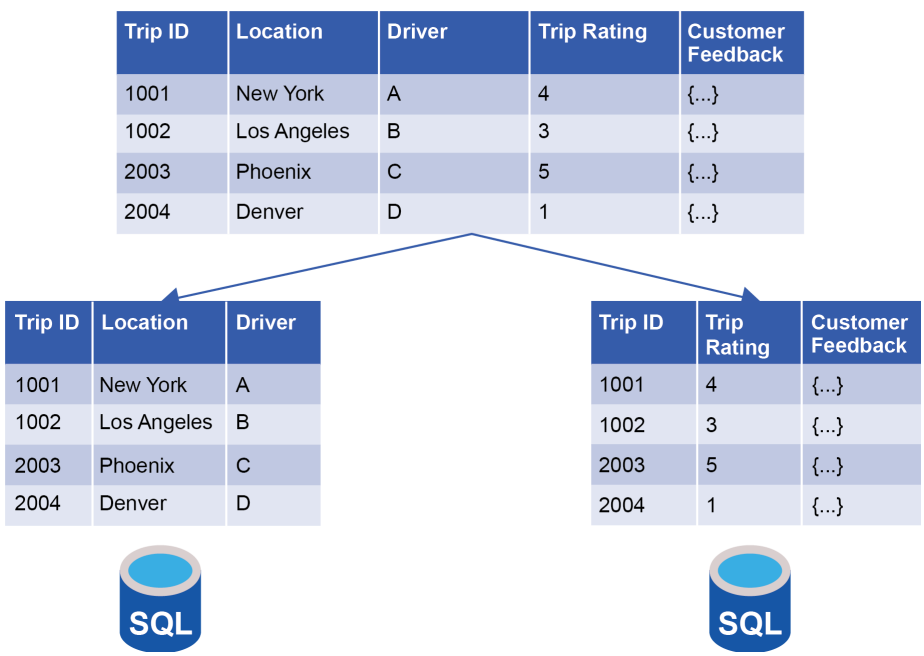

Figure 2.2 – An illustration of Vertical partitioning in action

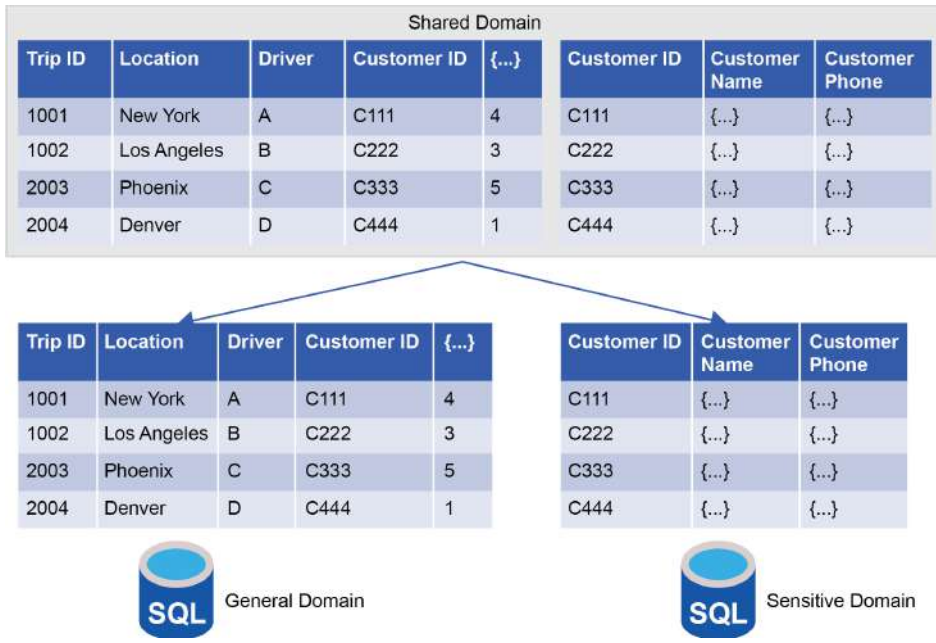

Figure 2.3 – Example of a functional partition

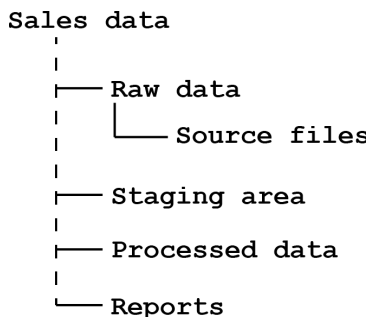

Figure 2.4 – The directory structure within ADLS Gen2

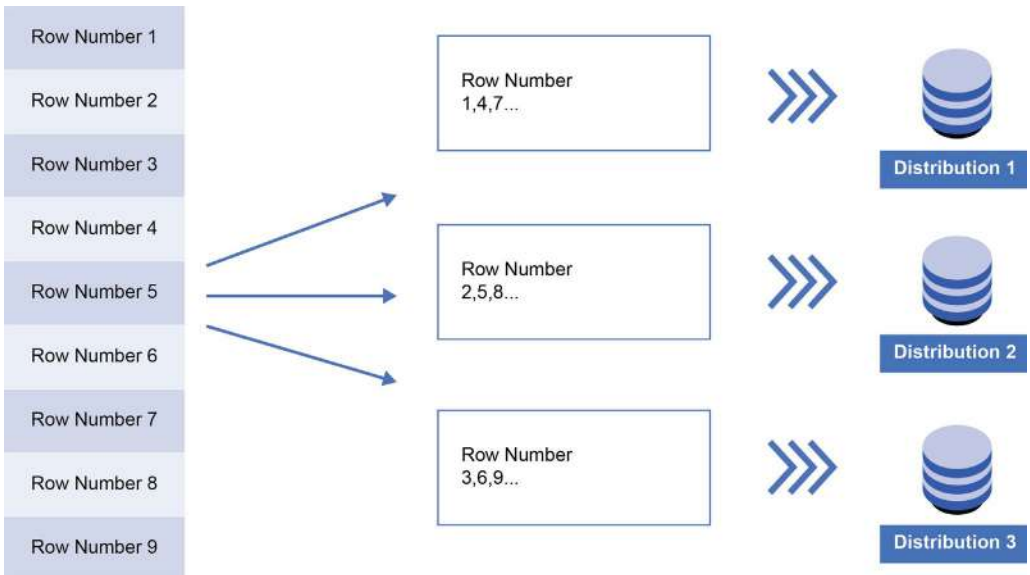

Figure 2.5 – Round-robin table distribution showing the distribution of data

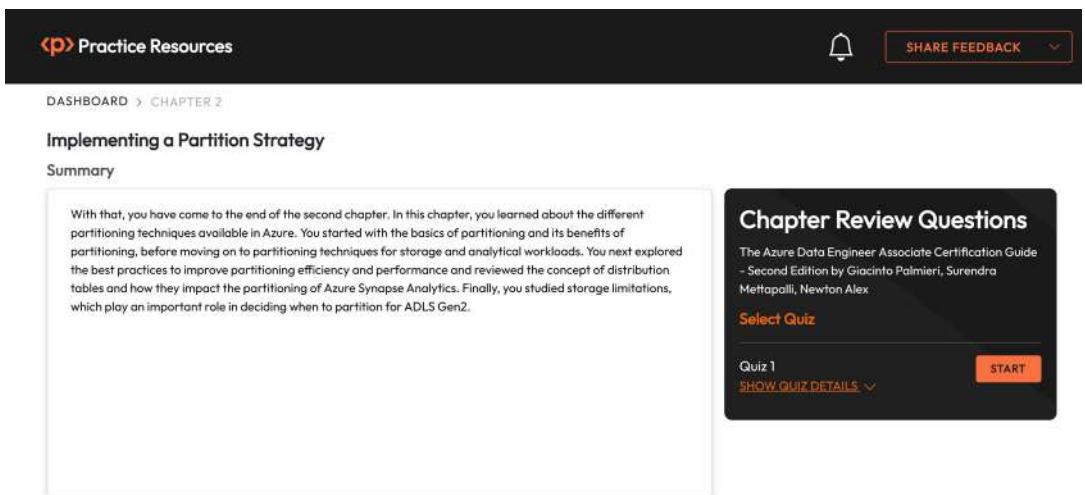

Figure 2.7 – Chapter Review Questions for Chapter 2

Chapter 3: Designing and Implementing the Data Exploration Layer

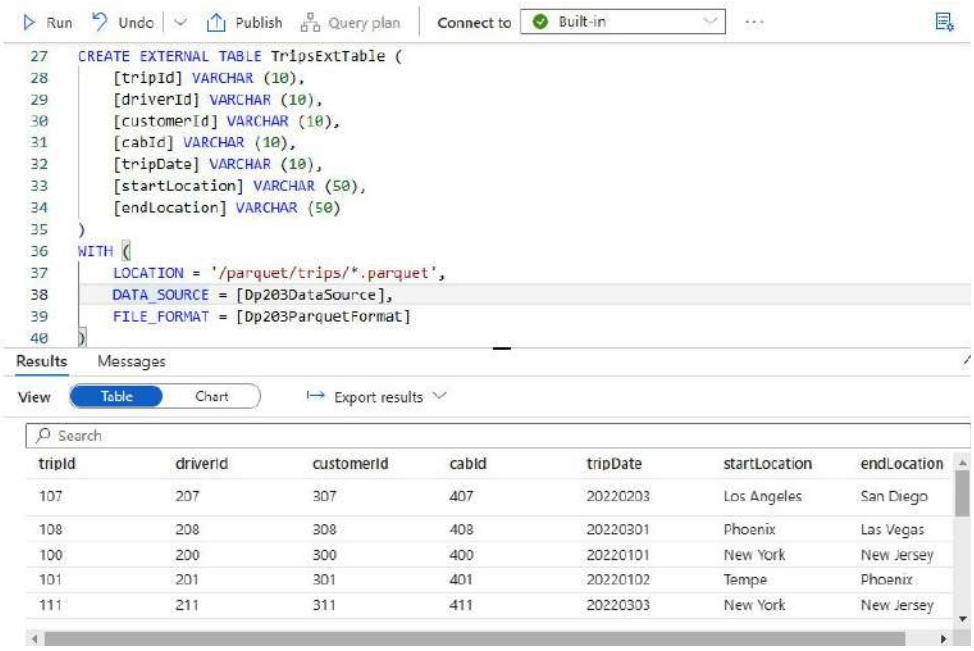

Figure 3.1 – Creating a SQL pool external table

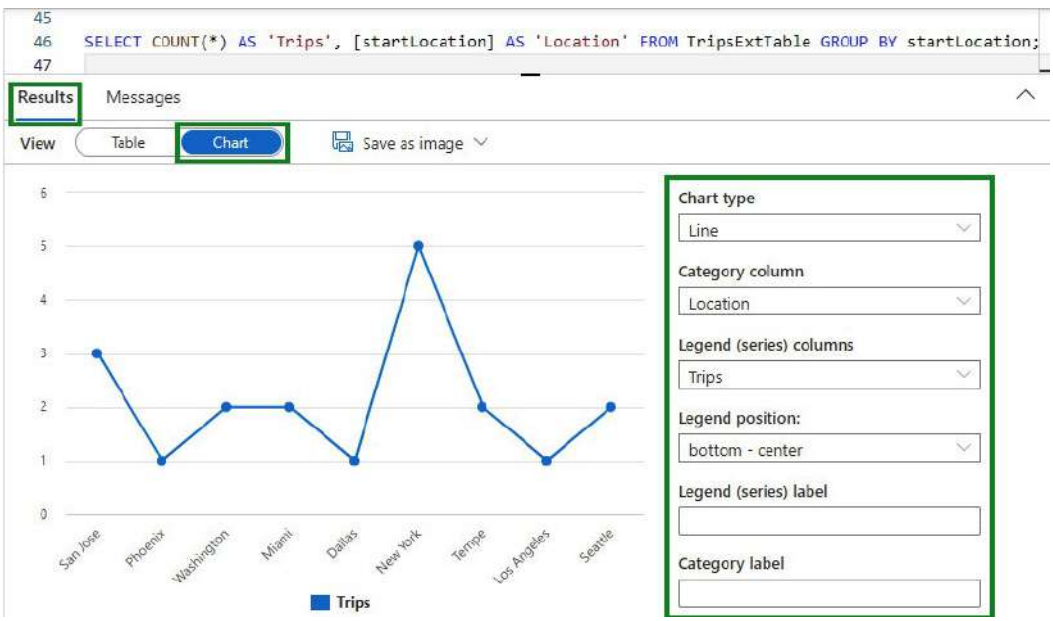

Figure 3.2 – Sample query on Parquet data visualized as a chart

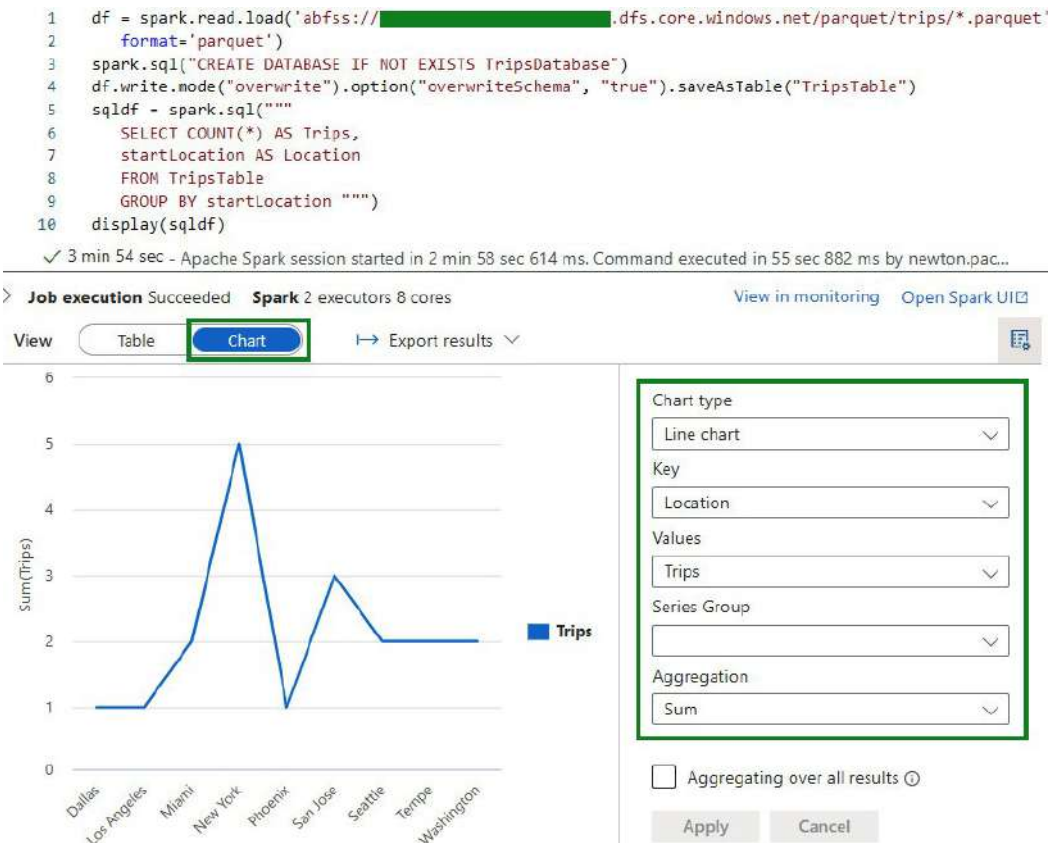

Figure 3.3 – Using Spark to query Parquet files

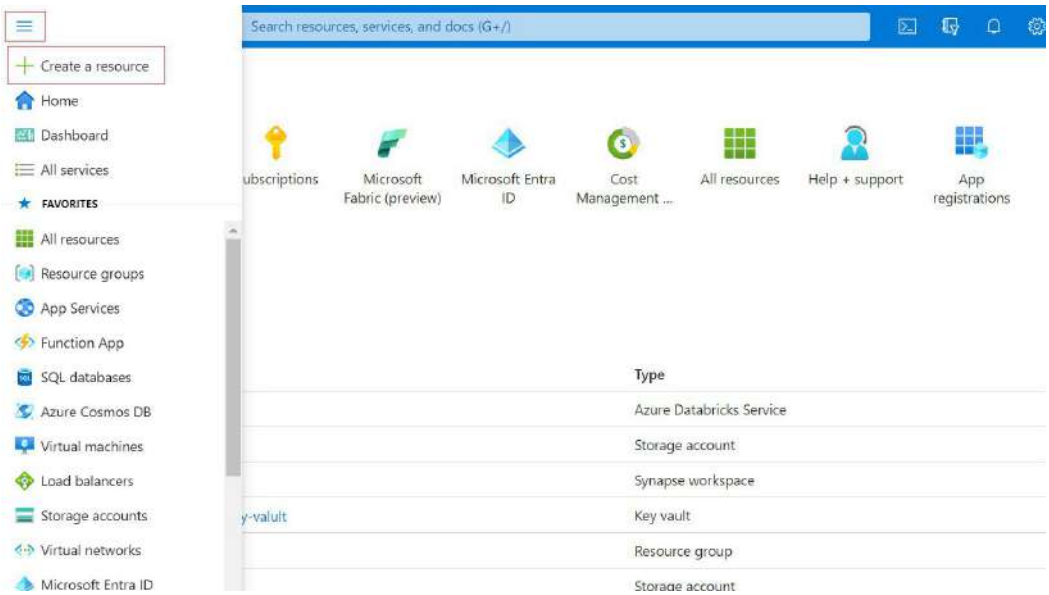

Figure 3.4 – Creating a resource in the Azure portal

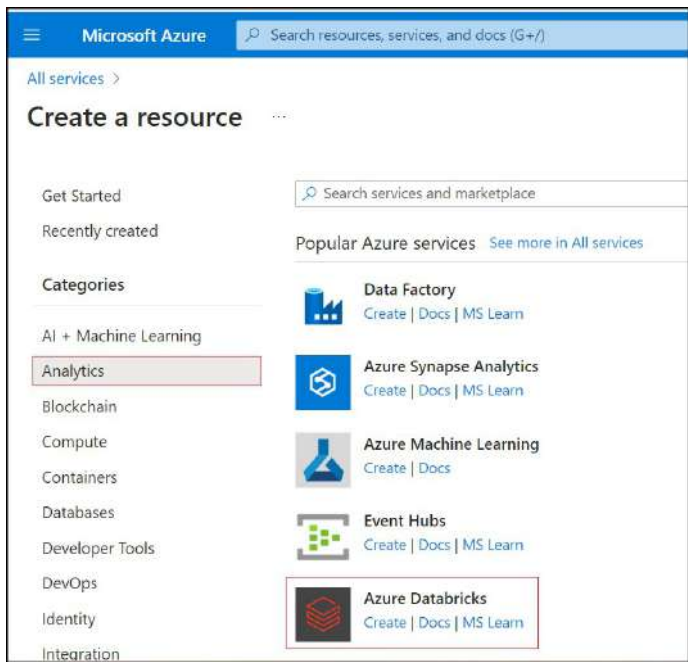

Figure 3.5 – Azure Databricks resource creation

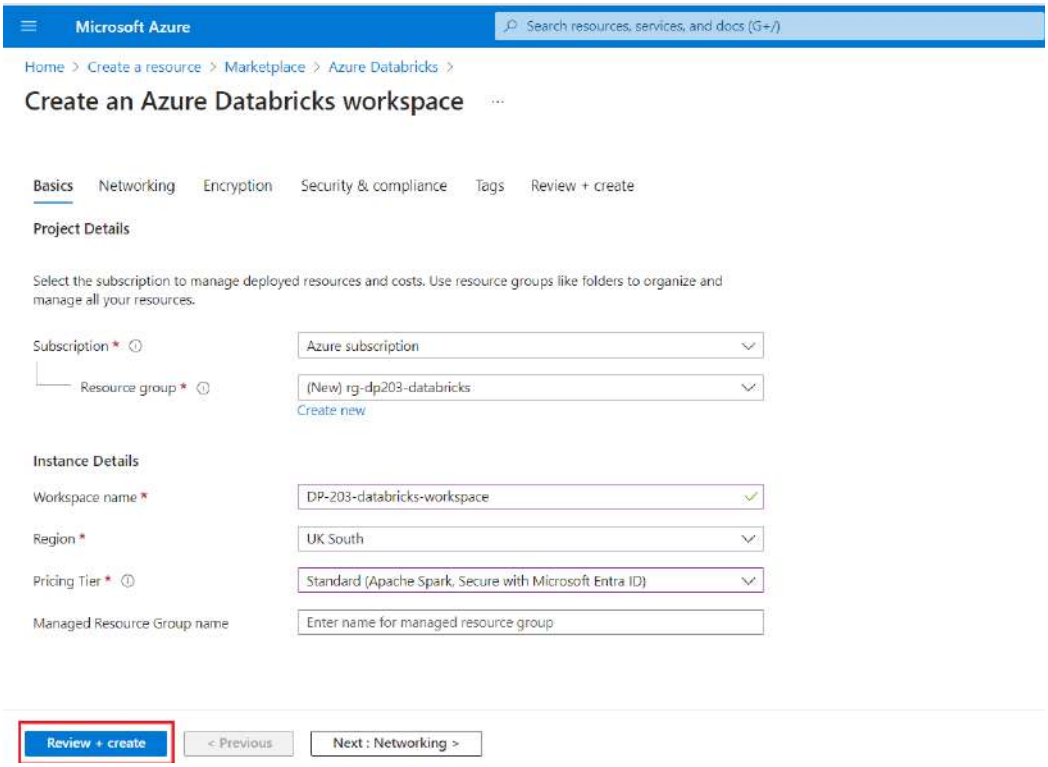

Figure 3.6 – Creating an Azure Databricks workspace

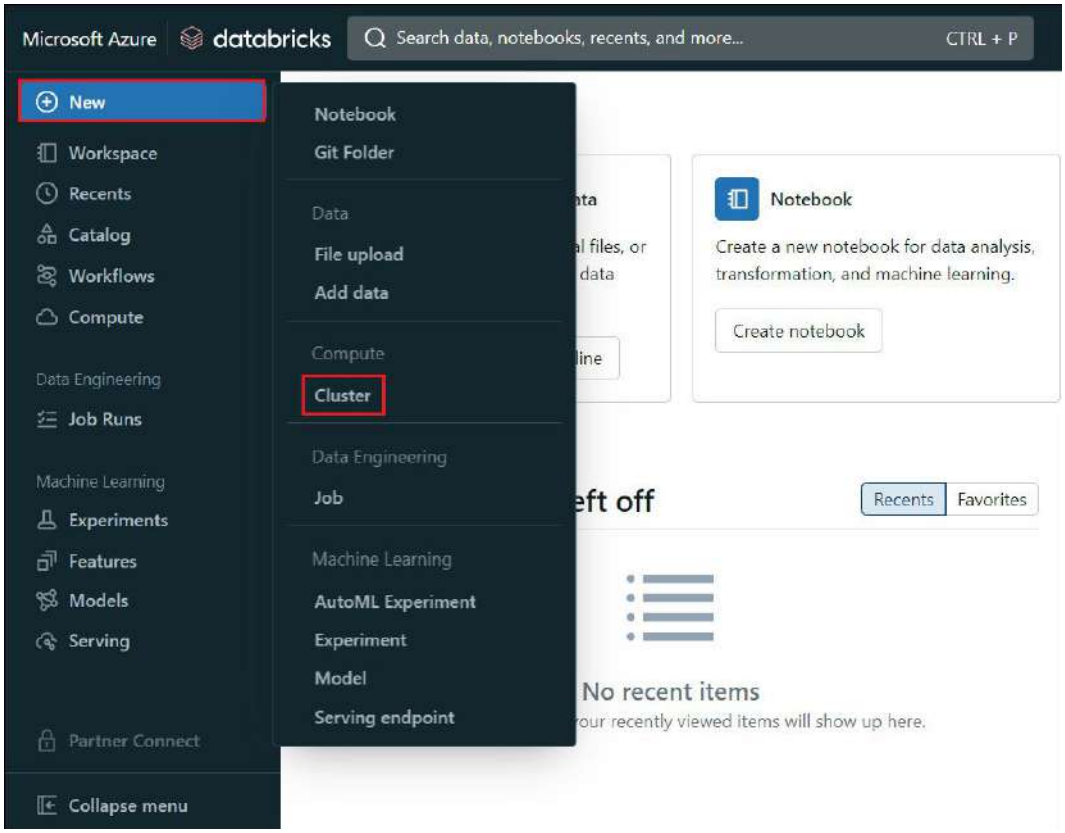

Figure 3.7 – Cluster creation in an Azure Databricks workspace

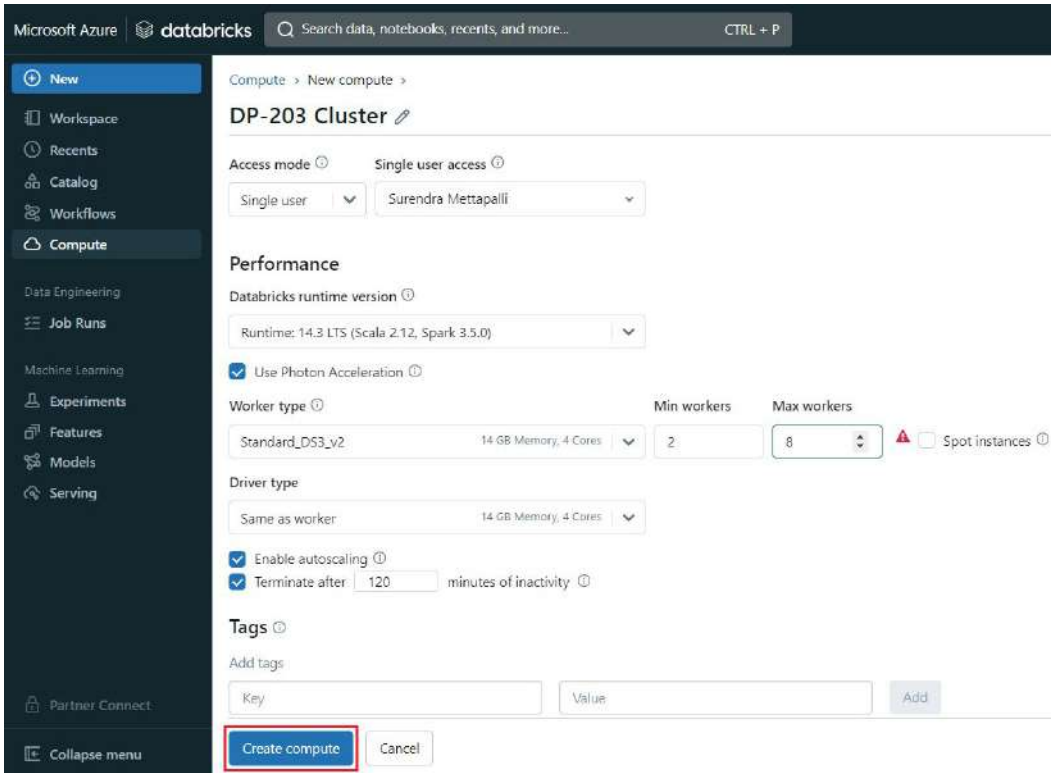

Figure 3.8 – Providing values to create a cluster in Azure Databricks

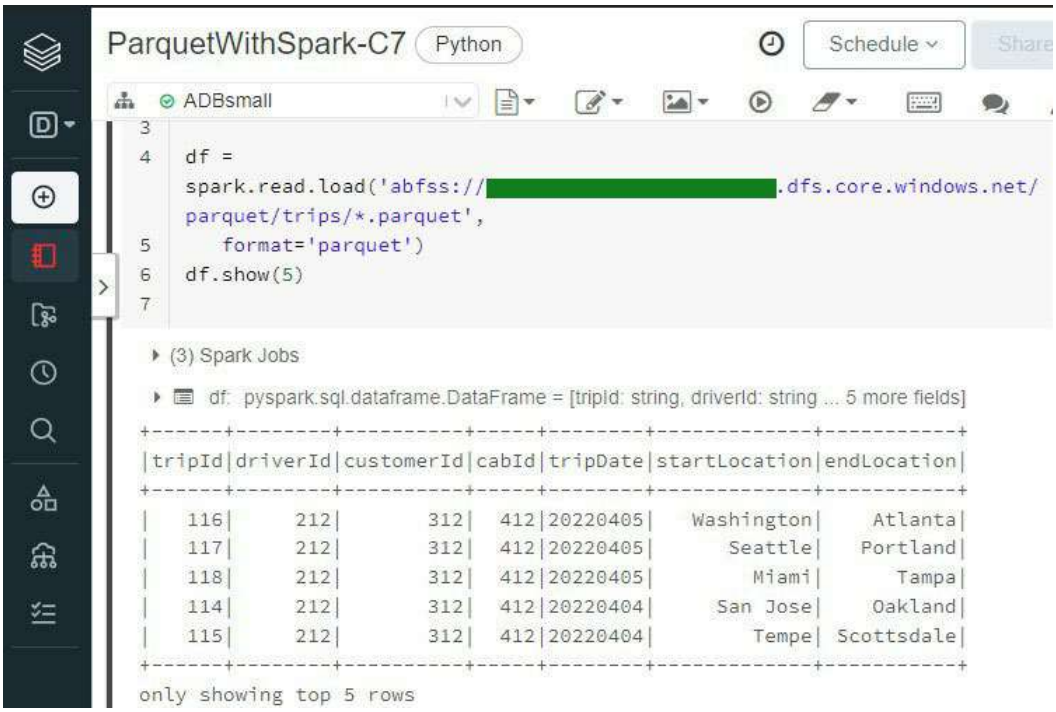

Figure 3.9 – An Azure Databricks notebook query to read Parquet files

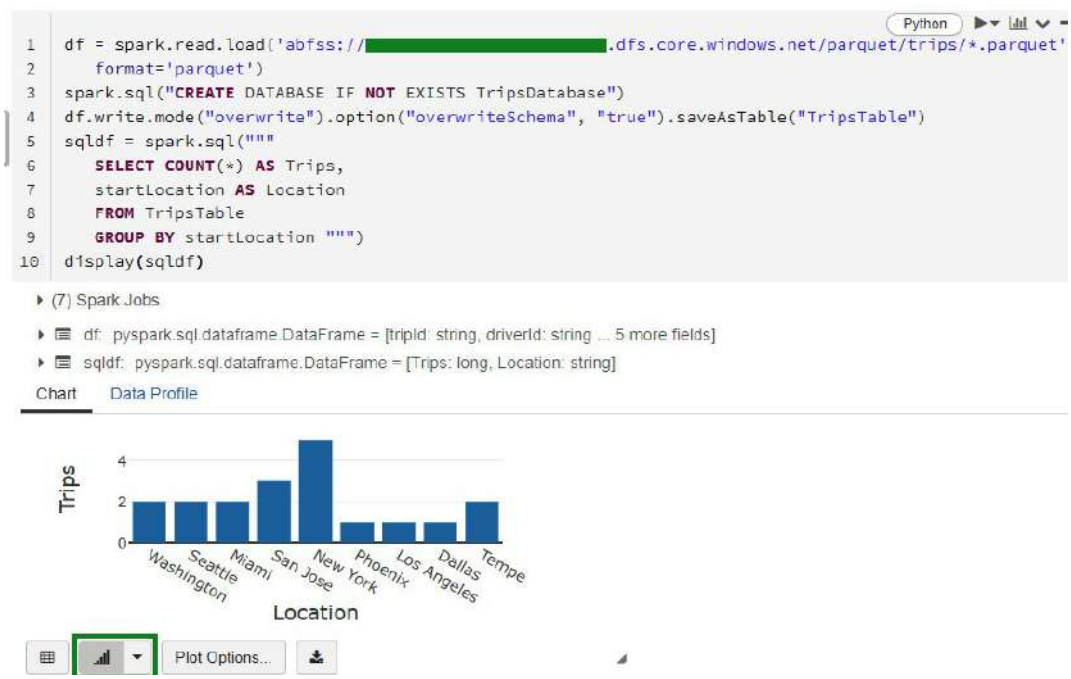

Figure 3.10 – A simple Spark query on Parquet data in Databricks

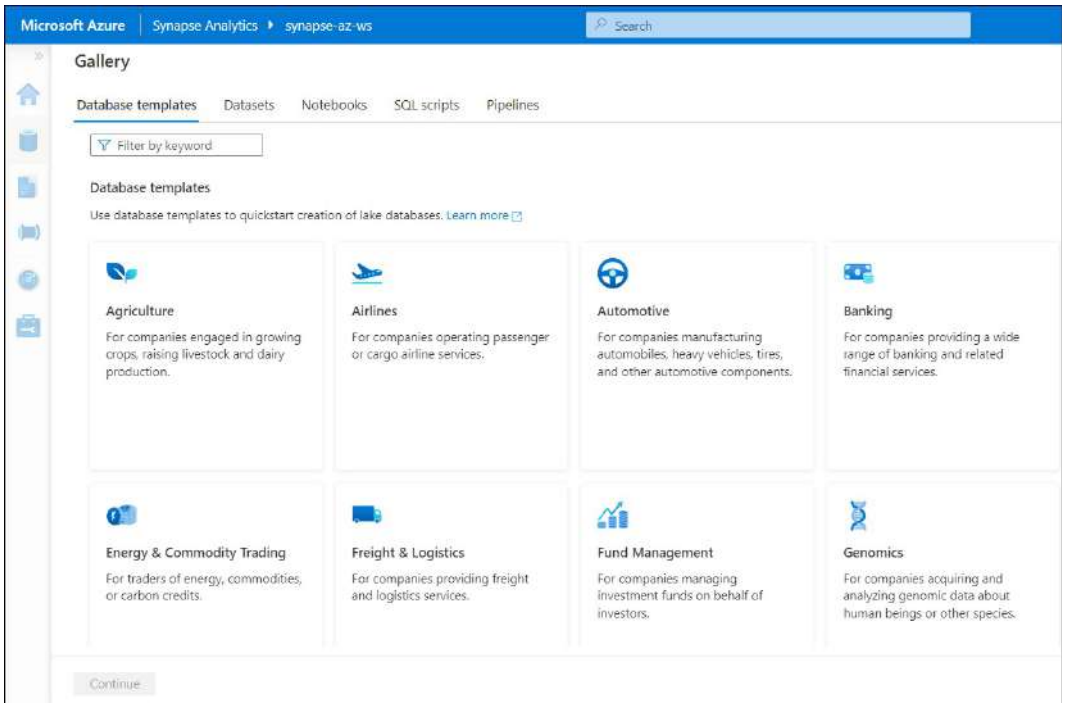

Figure 3.11 – A Synapse database showing the available templates

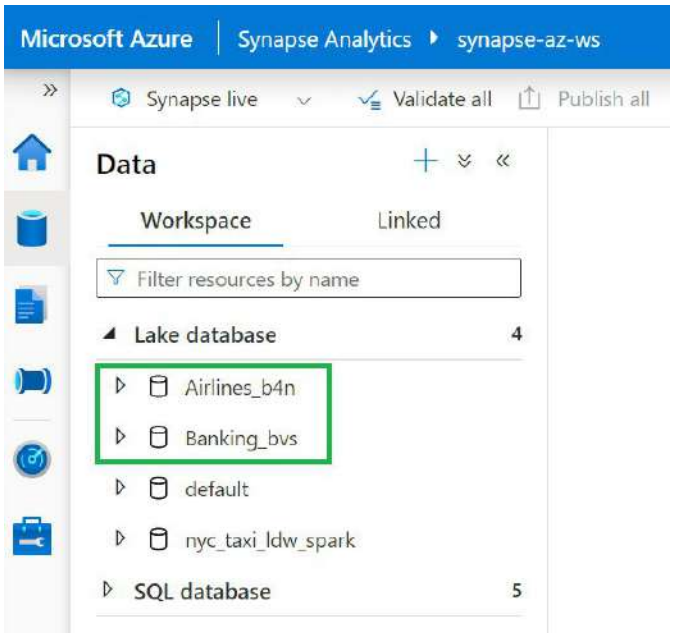

Figure 3.12 – The Synapse workspace showing the lake database templates

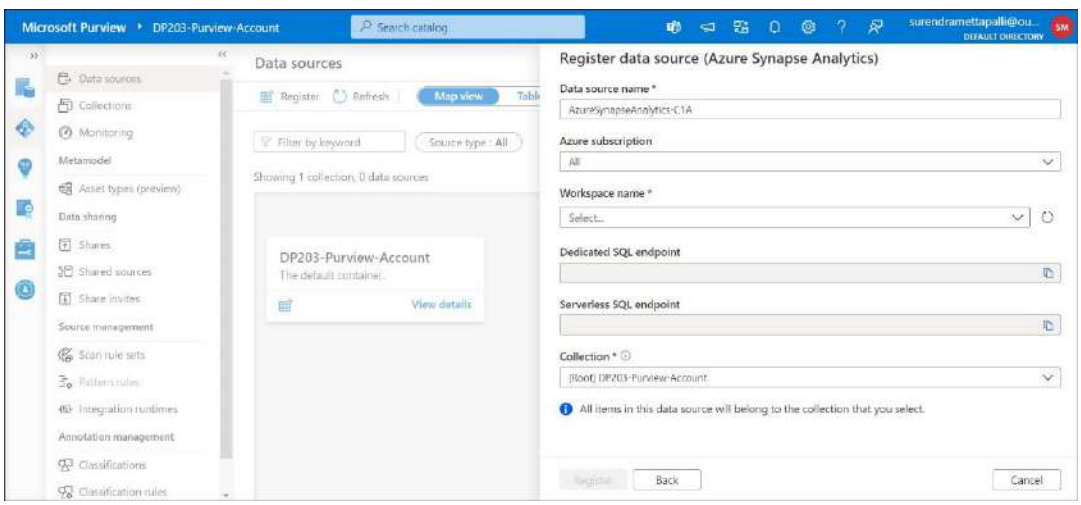

Figure 3.13 – Registering a data source in the Microsoft Purview governance portal

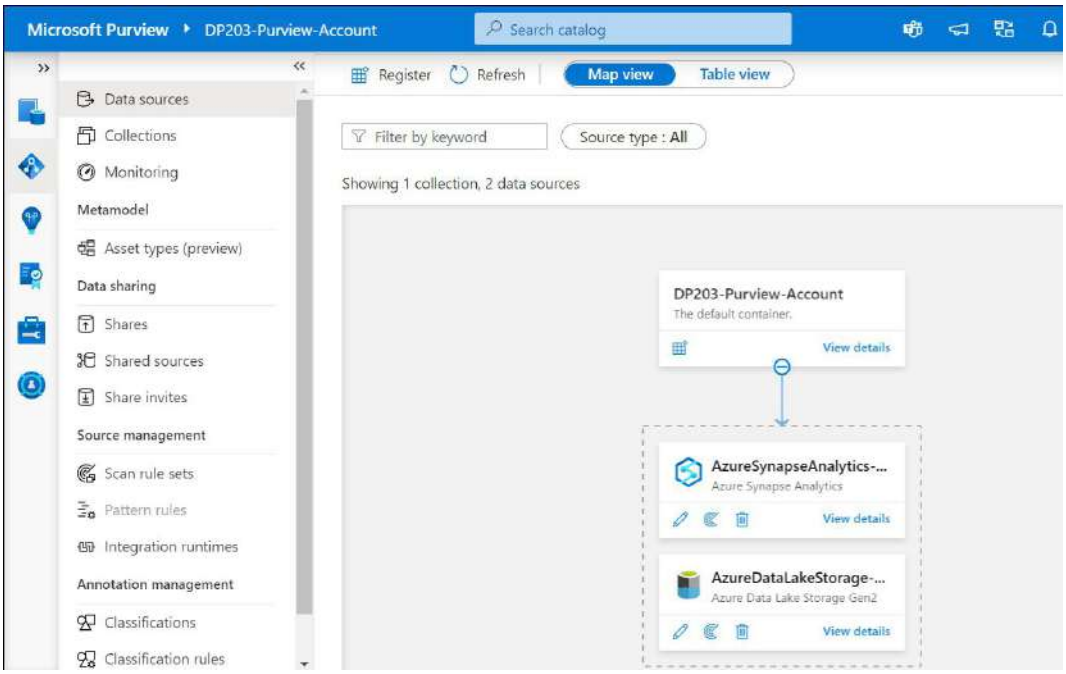

Figure 3.14 – Visualizing a Data Map in a Synapse workspace and Data Lake Storage Gen2 Account

Figure 3.15 – An example of data lineage for Power BI

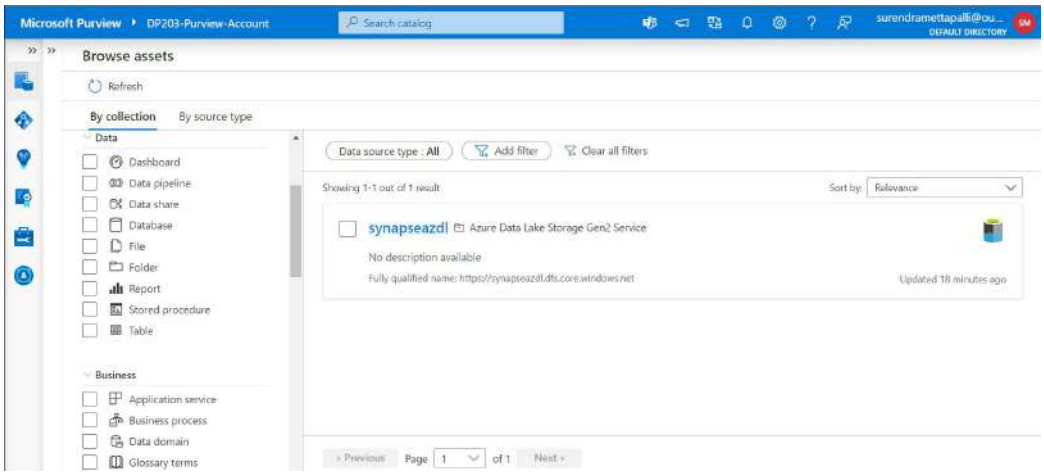

Figure 3.16 – Filtering search results with facets

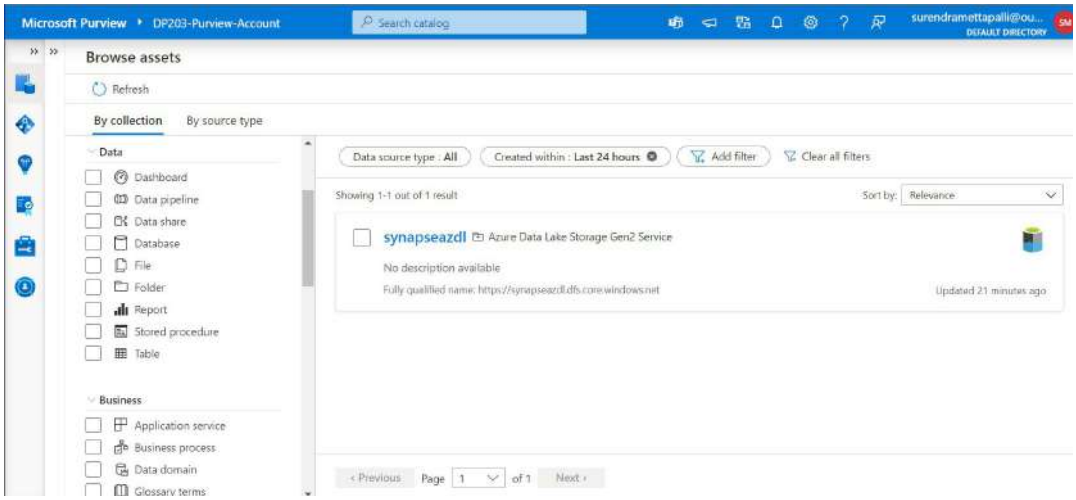

Figure 3.17 – Filtering search results with date filters

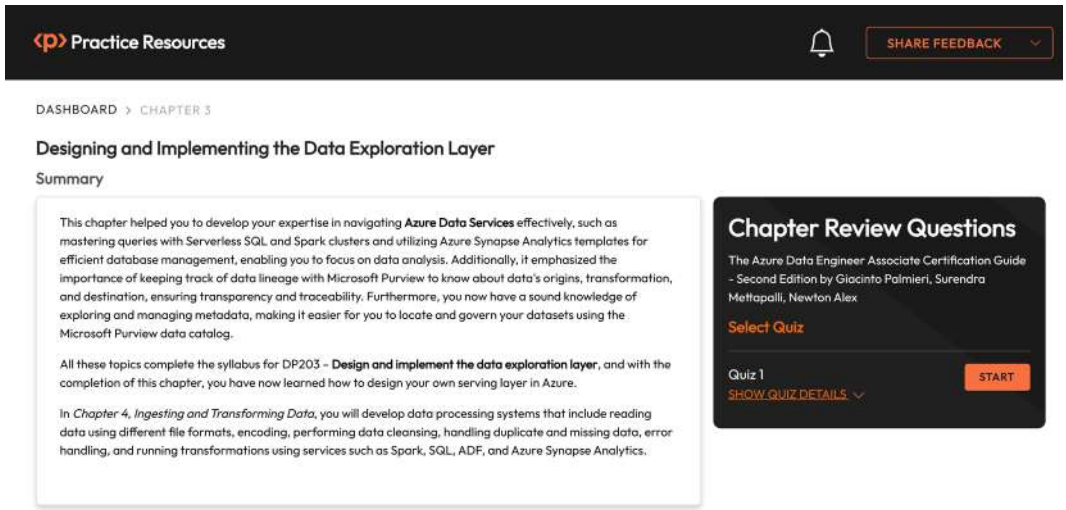

Figure 3.19 – Chapter Review Questions for Chapter 3

Chapter 4: Ingesting and Transforming Data

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

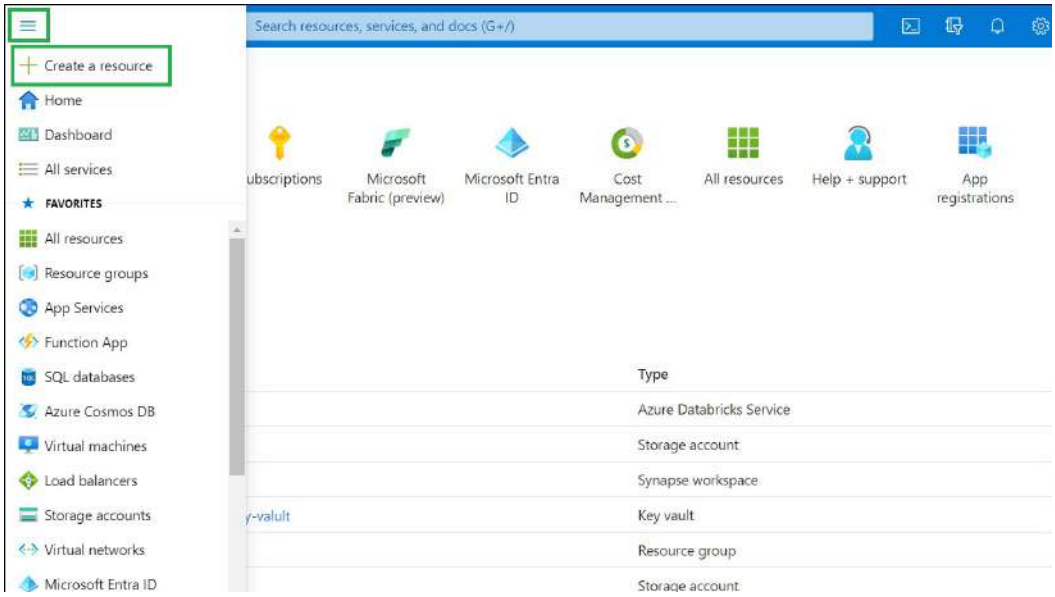

Figure 4.2 – Creating a ADF resource in the Azure portal

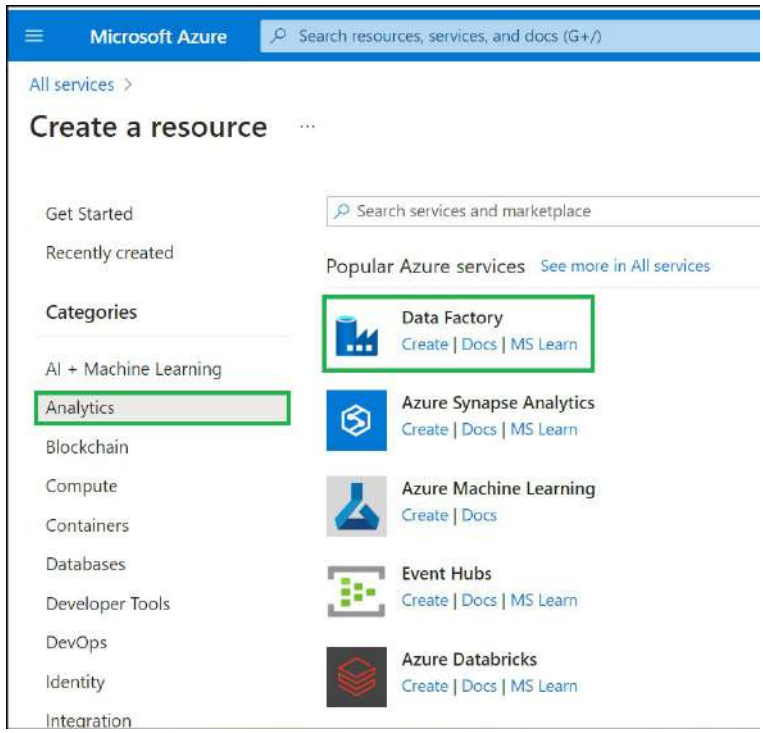

Figure 4.3 – ADF resource creation

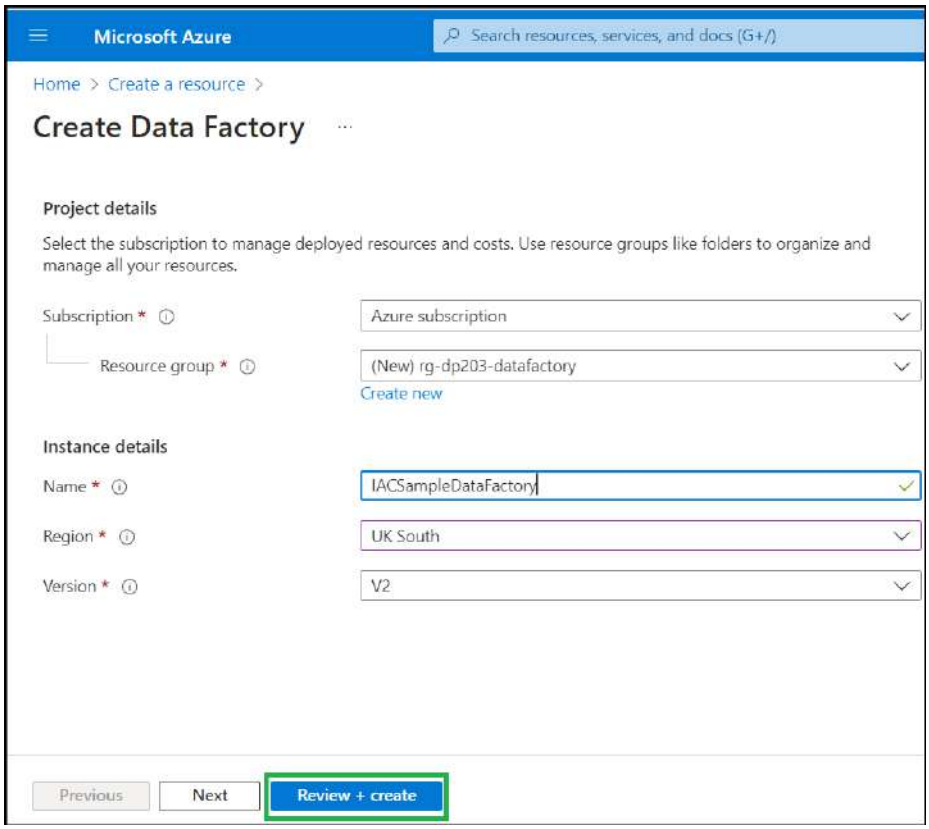

Figure 4.4 – Creating an ADF resource

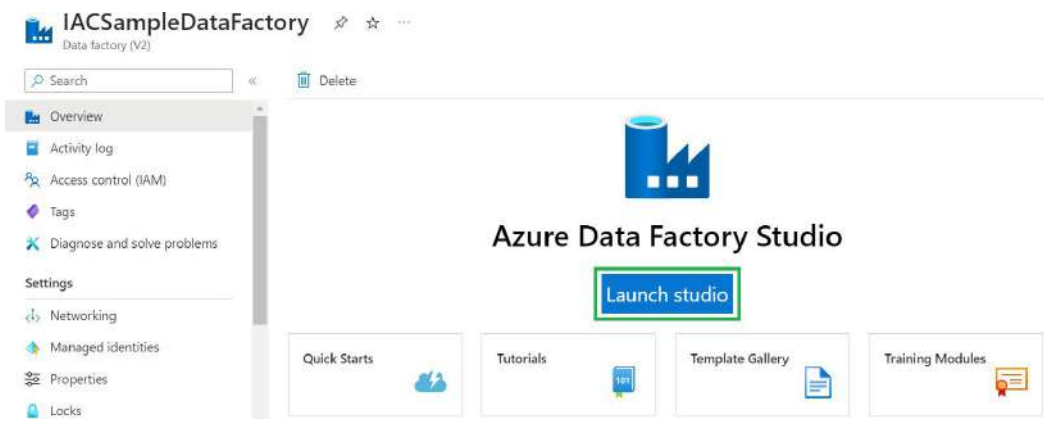

Figure 4.5 – ADF prompting to open the workspace and start building pipelines

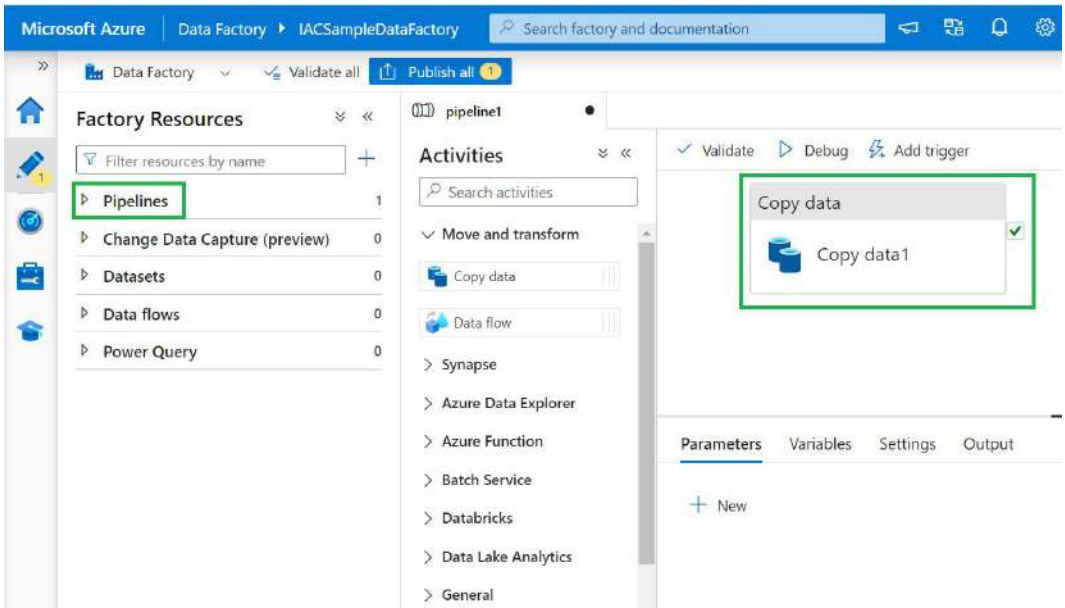

Figure 4.6 – ADF Studio to build the pipeline activities

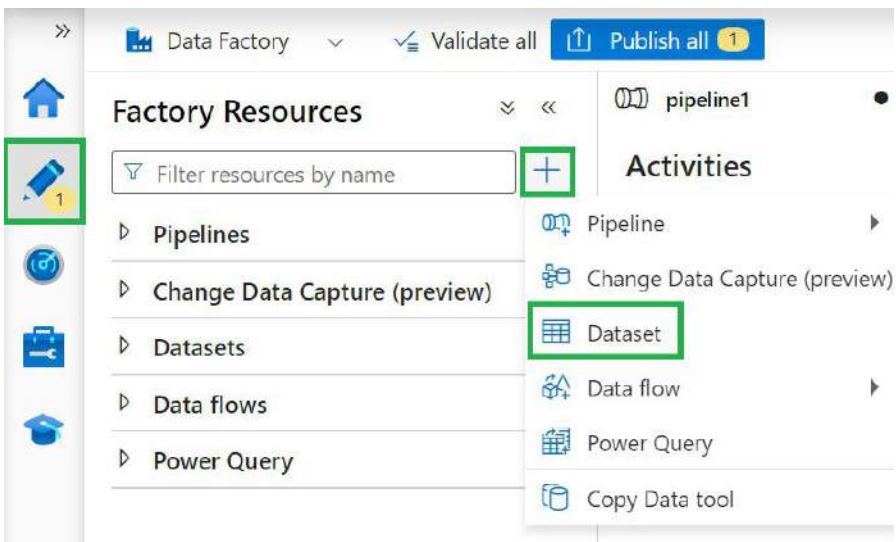

Figure 4.7 – Creating a new dataset in ADF

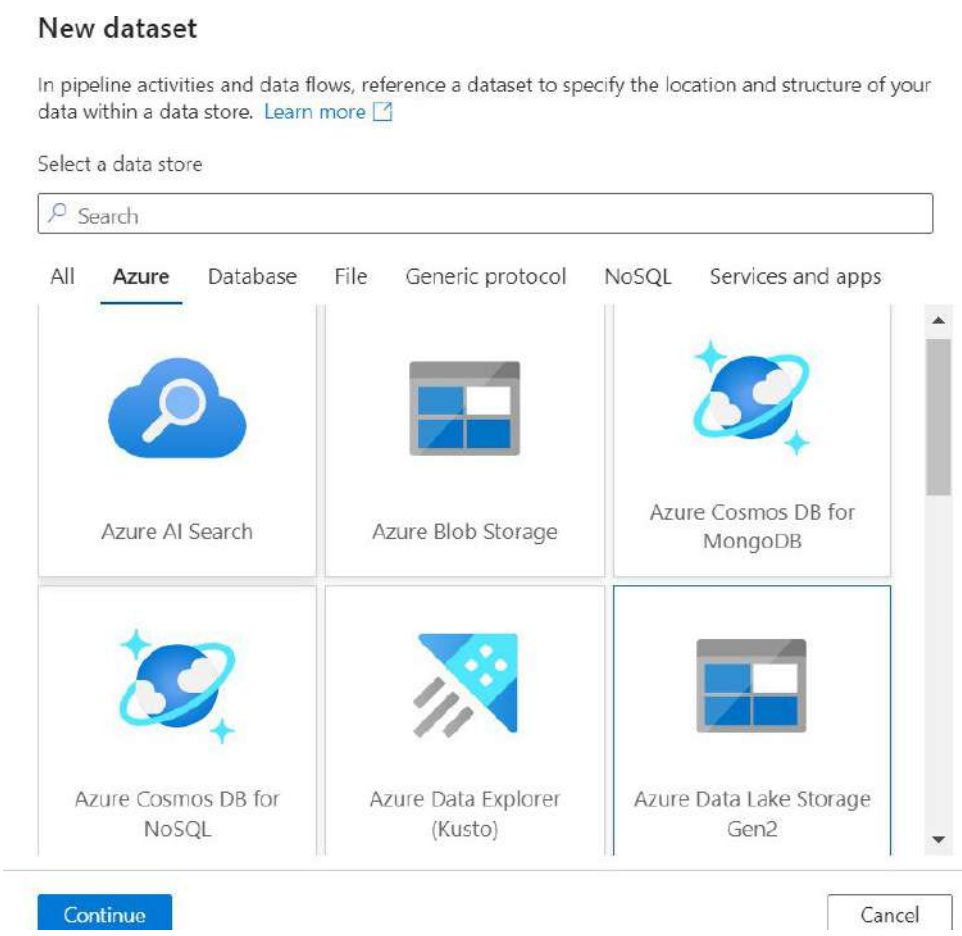

Figure 4.8 – Dataset source options in ADF

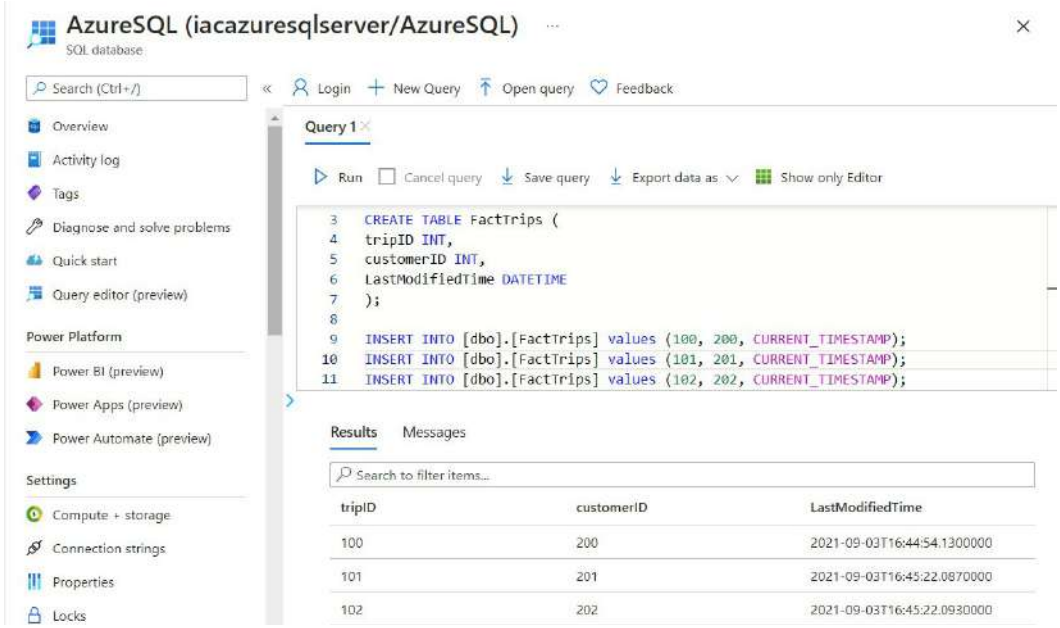

Figure 4.9 – Creating a simple table using the Query editor in Azure SQL

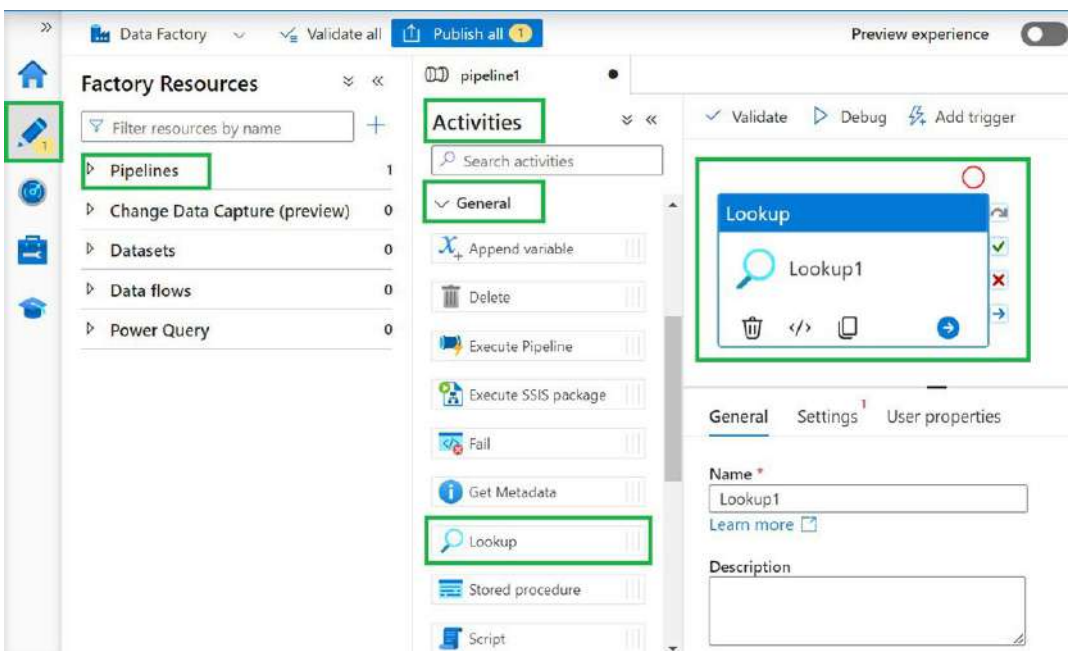

Figure 4.10 – The ADF authoring page

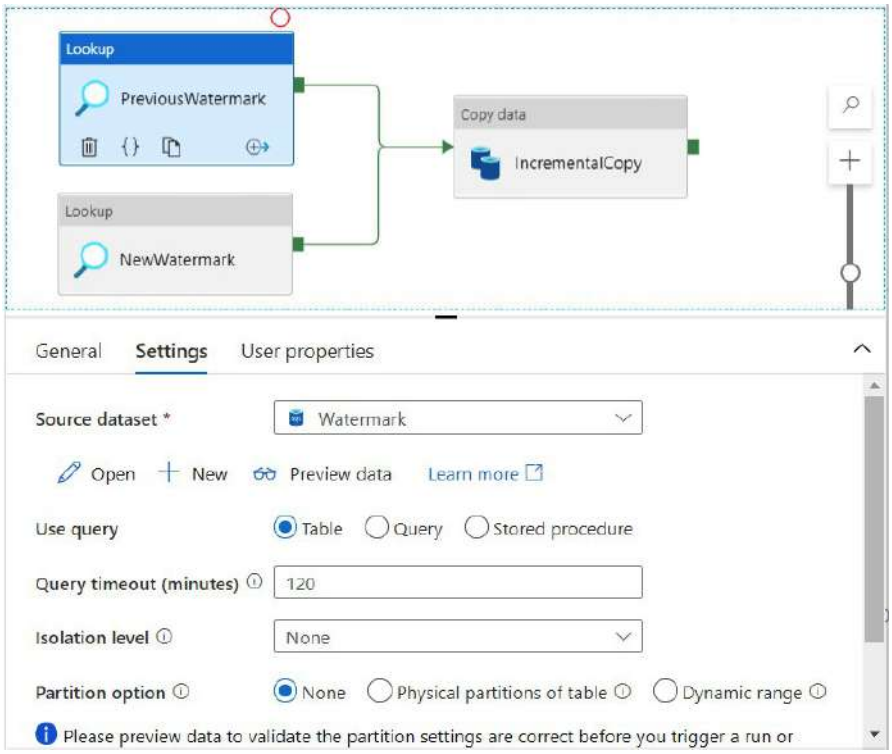

Figure 4.11 – The lookup activity configuration using the Watermark table

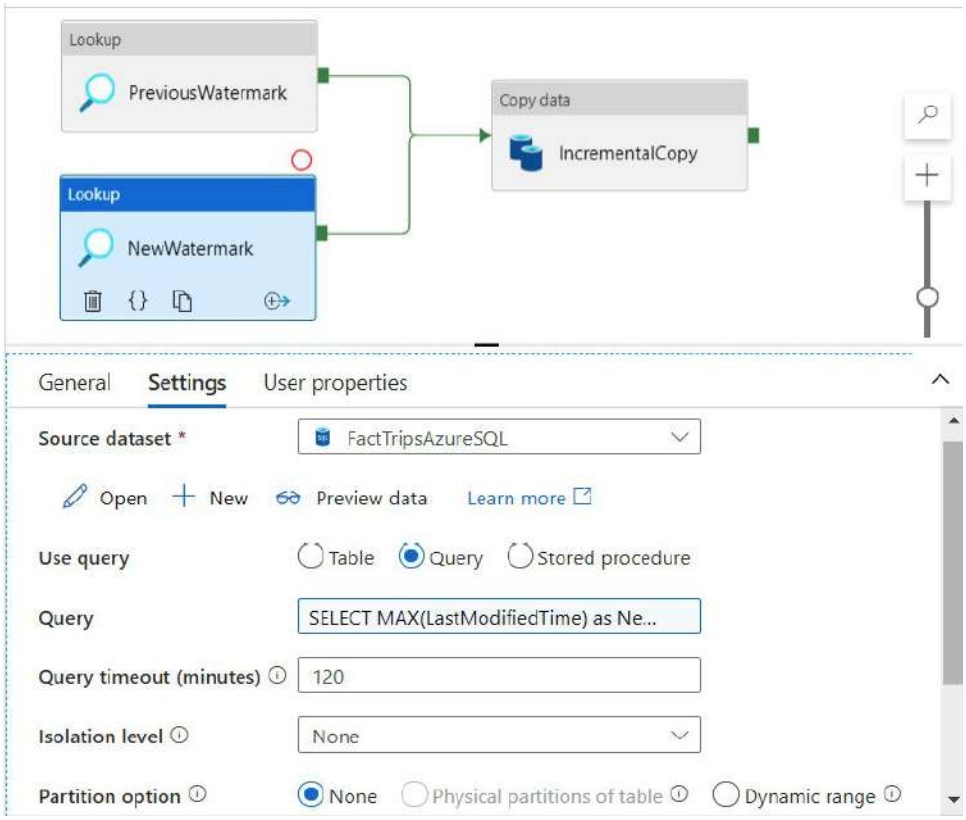

Figure 4.12 – New watermark lookup configuration using LastModifiedTime

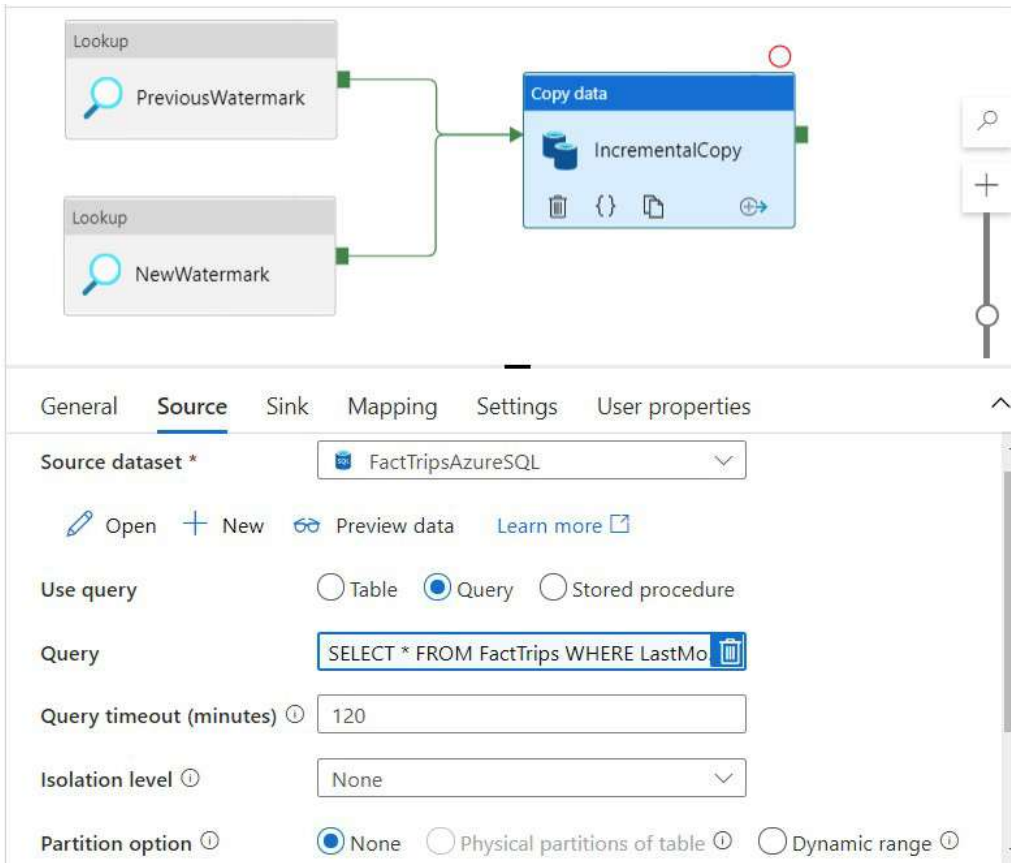

Figure 4.13 – ADF Copy activity with watermark-based delta generation

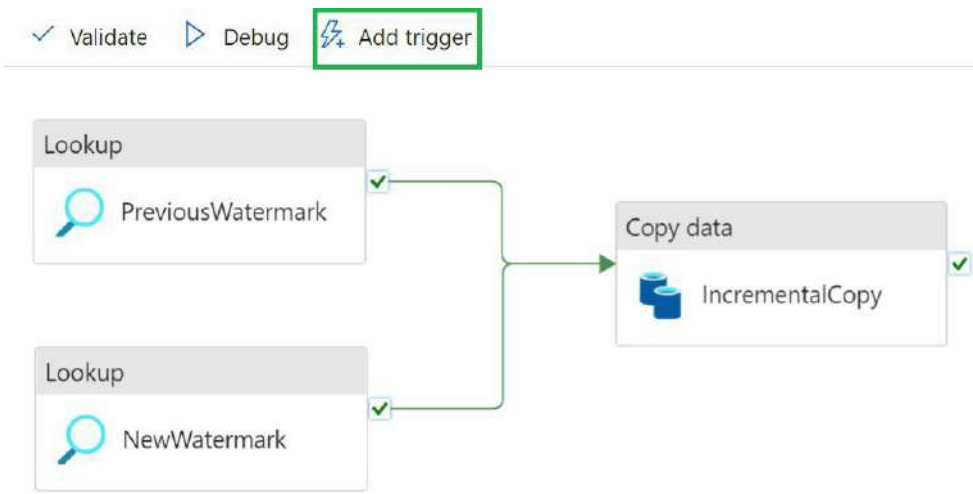

Figure 4.14 – Scheduling the ADF pipeline by clicking Add trigger

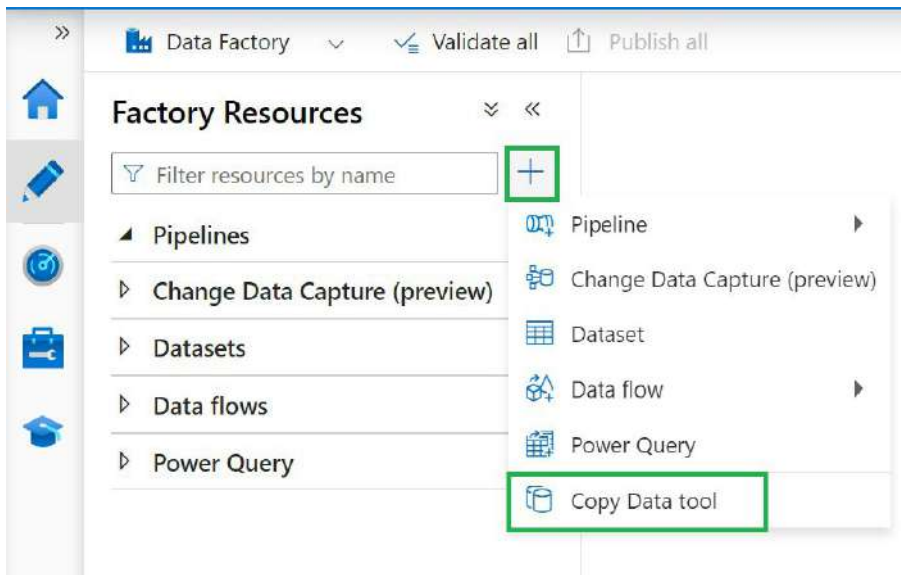

Figure 4.15 – Scanning files at the source based on the LastModifiedDate attribute

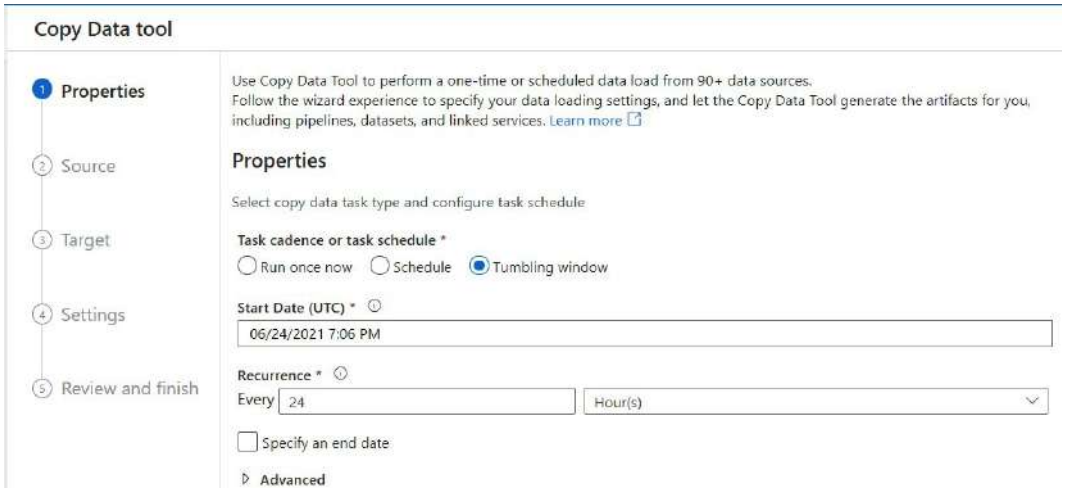

Figure 4.16 – Selecting Tumbling window for an incremental load, based on the file-modified time

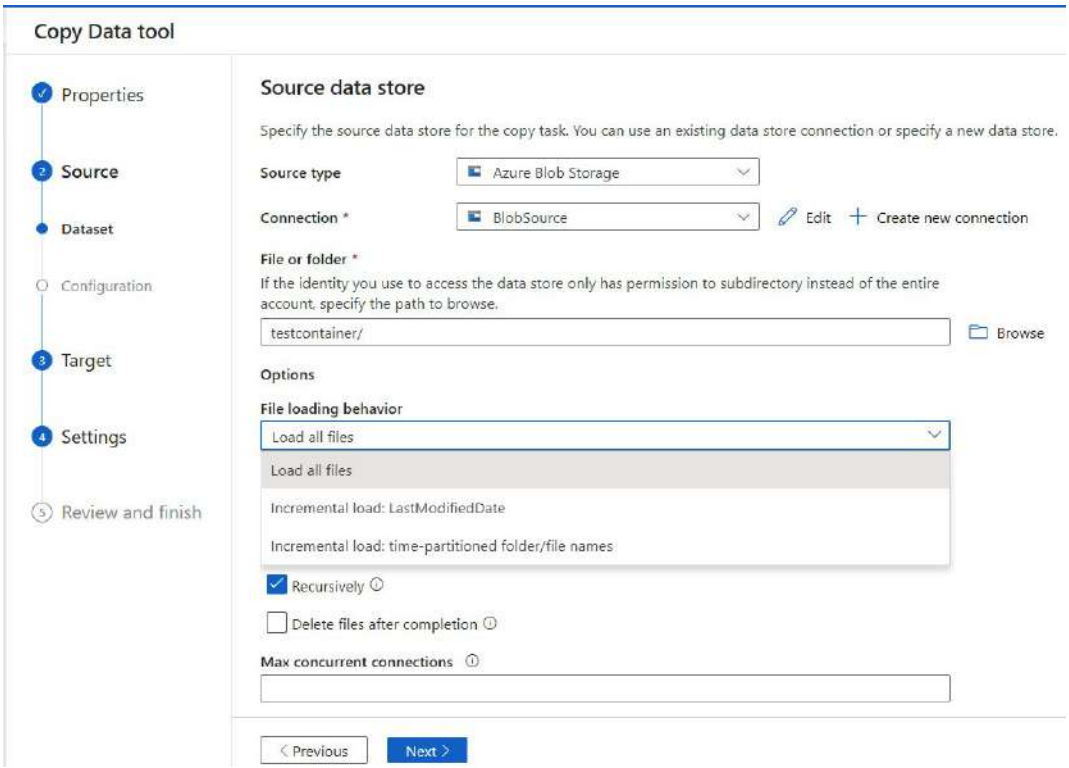

Figure 4.17 – Incremental load with LastModifiedDate timestamp behavior

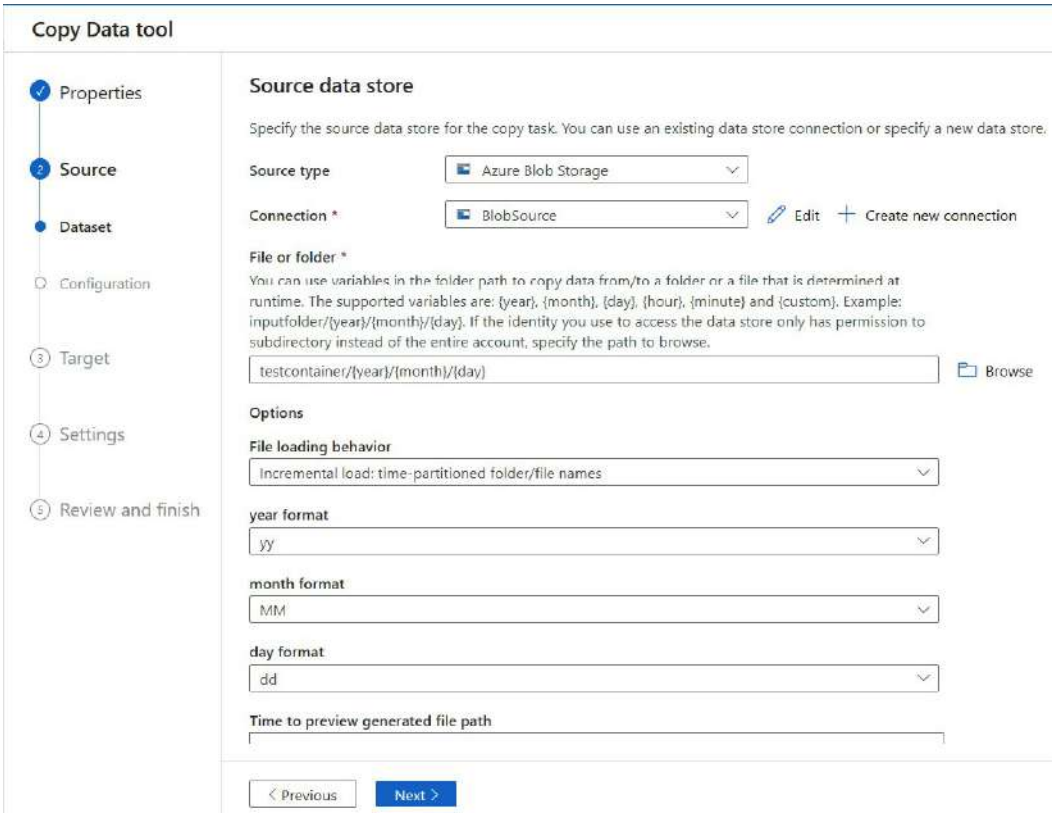

Figure 4.18 – ADF incremental load option with time-partitioned folders

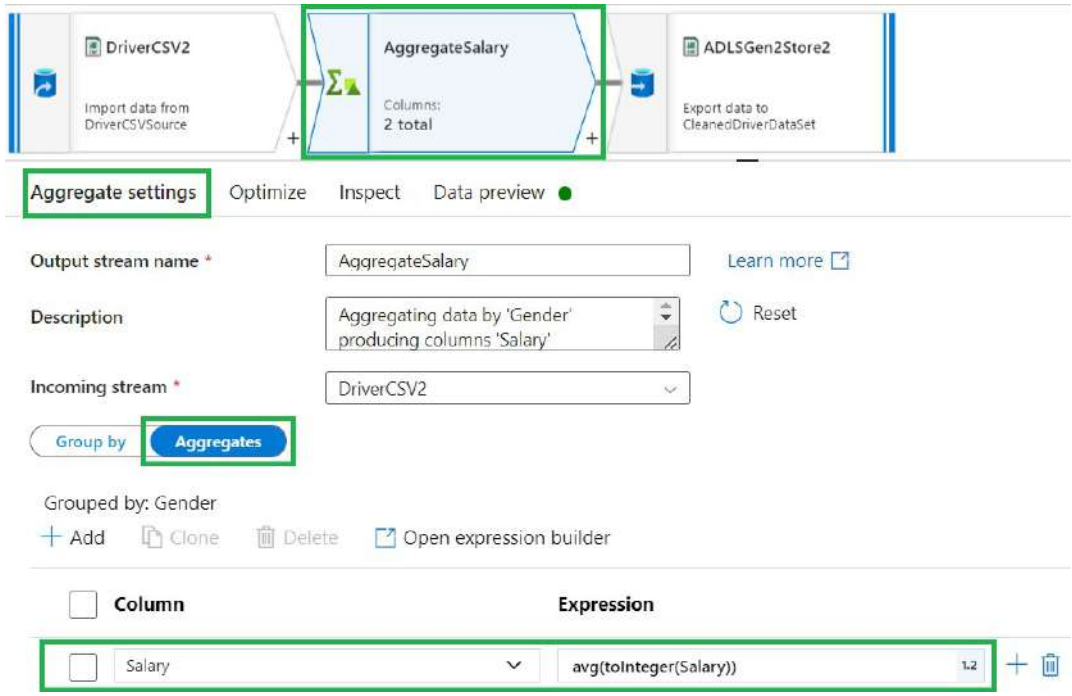

Figure 4.19 – Performing Aggregate transformation in ADF

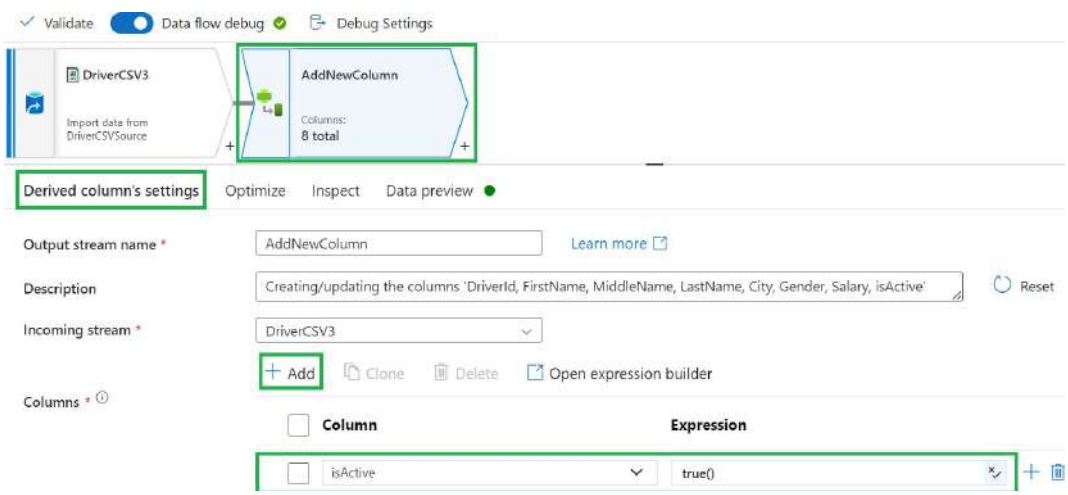

Figure 4.20 – Derived column transformation in ADF showing an added column

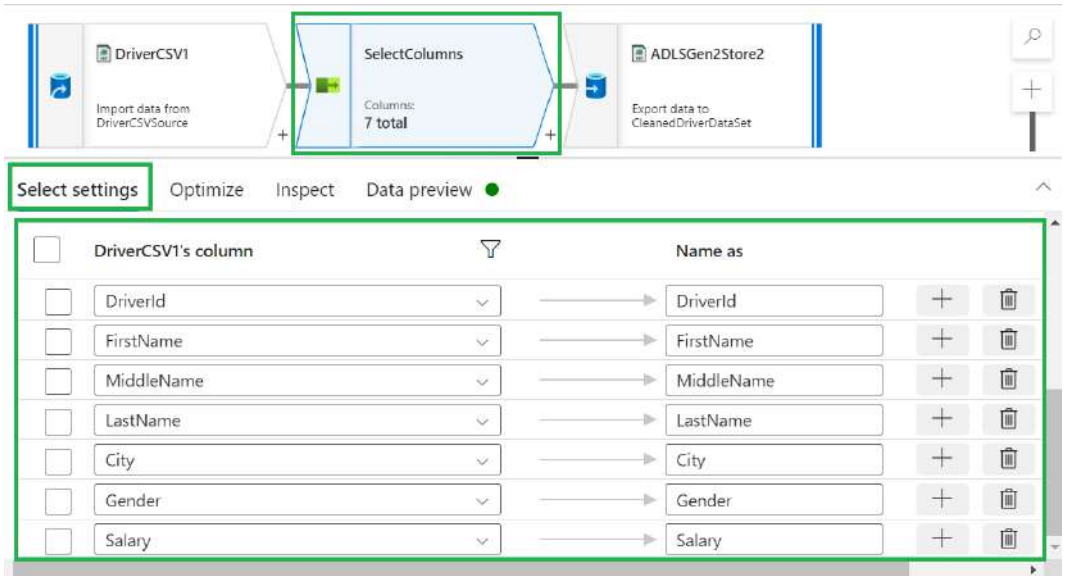

Figure 4.21 – Select transformation in ADF showing selecting required columns

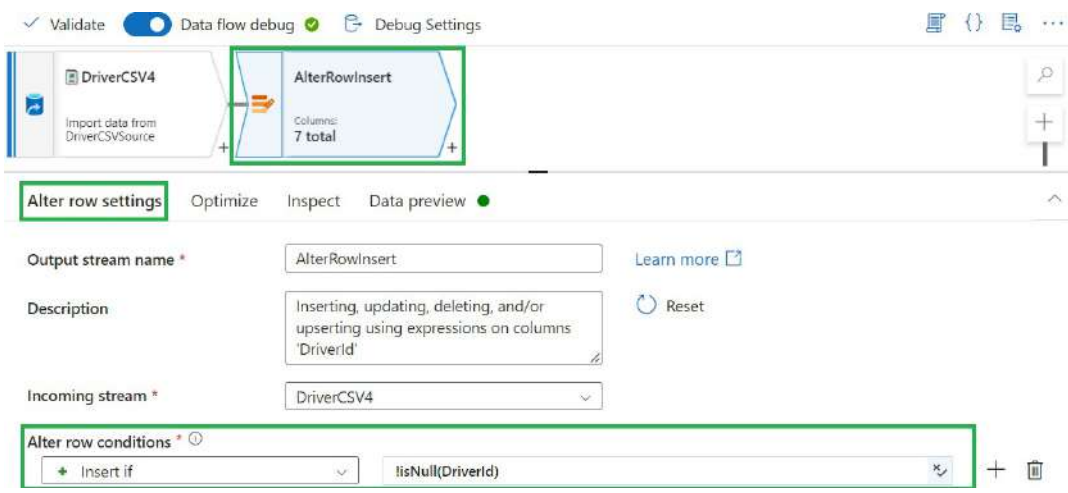

Figure 4.22 – Performing Alter row transformation as per NULL condition

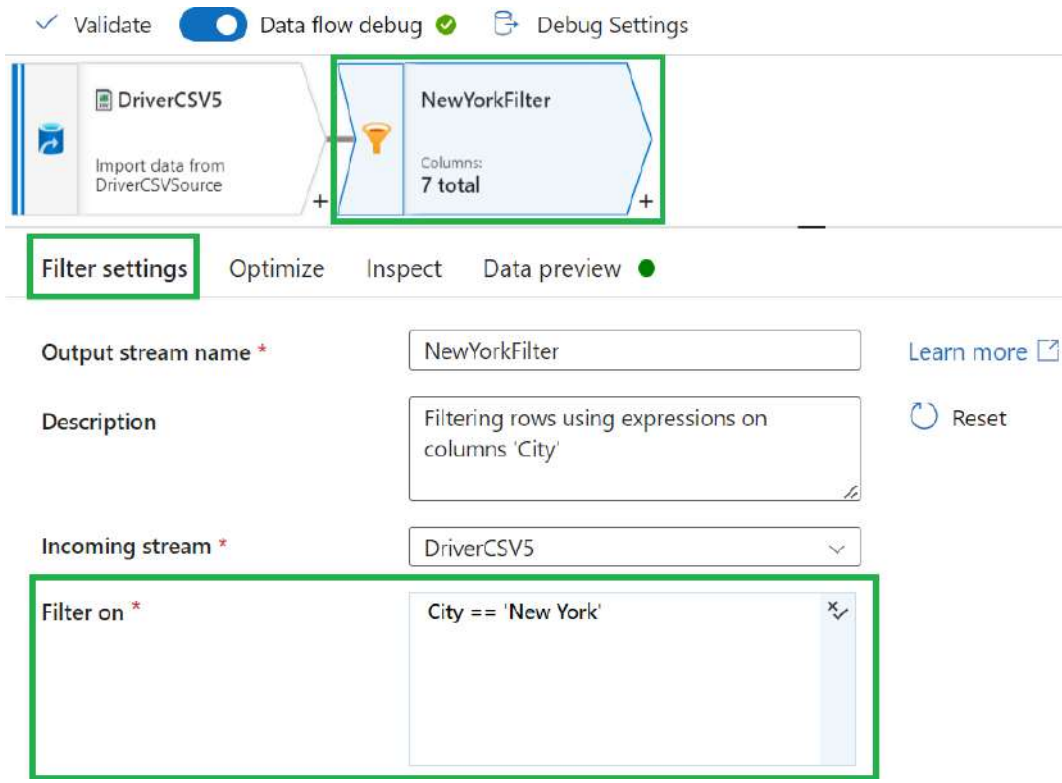

Figure 4.23 – Performing Filter transformation in ADF as per City

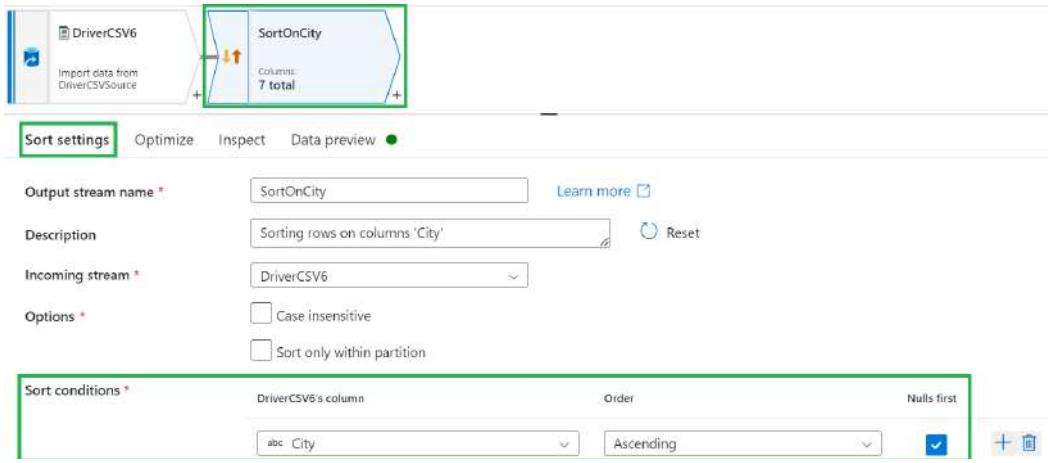

Figure 4.24 – Performing Sort transformation on specific rows

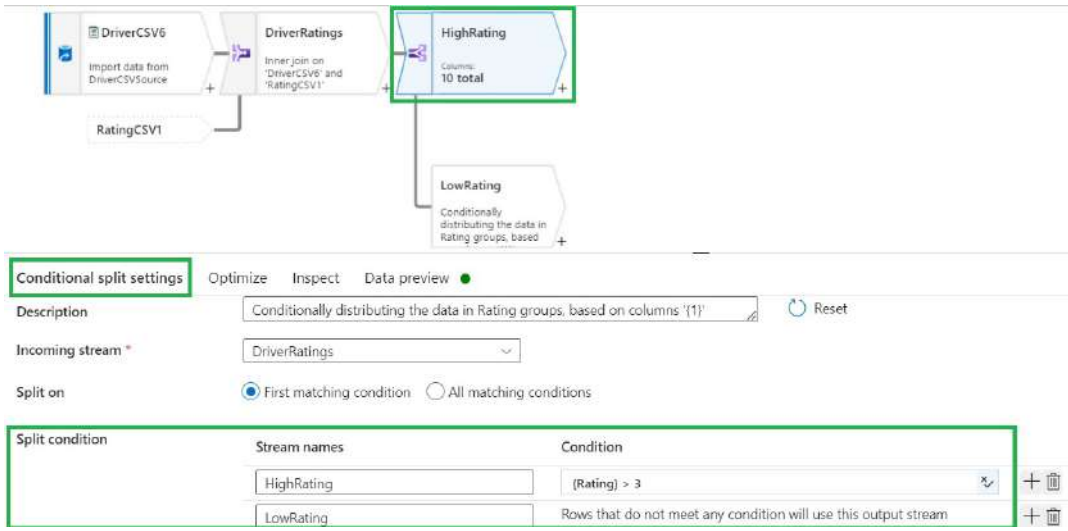

Figure 4.25 – Performing Conditional Split transformation in ADF for output or input stream

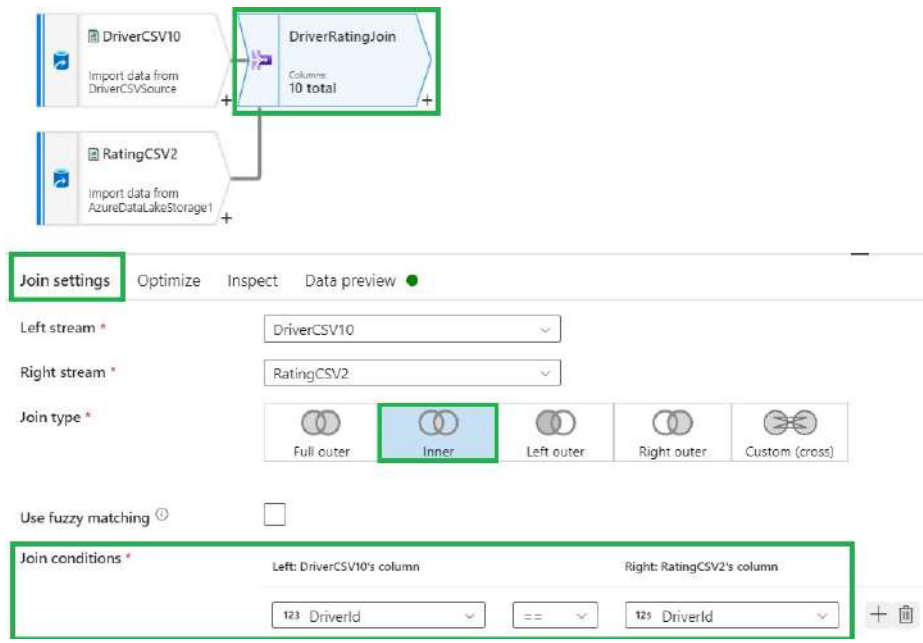

Figure 4.26 – Performing Join transformation in ADF based on join condition

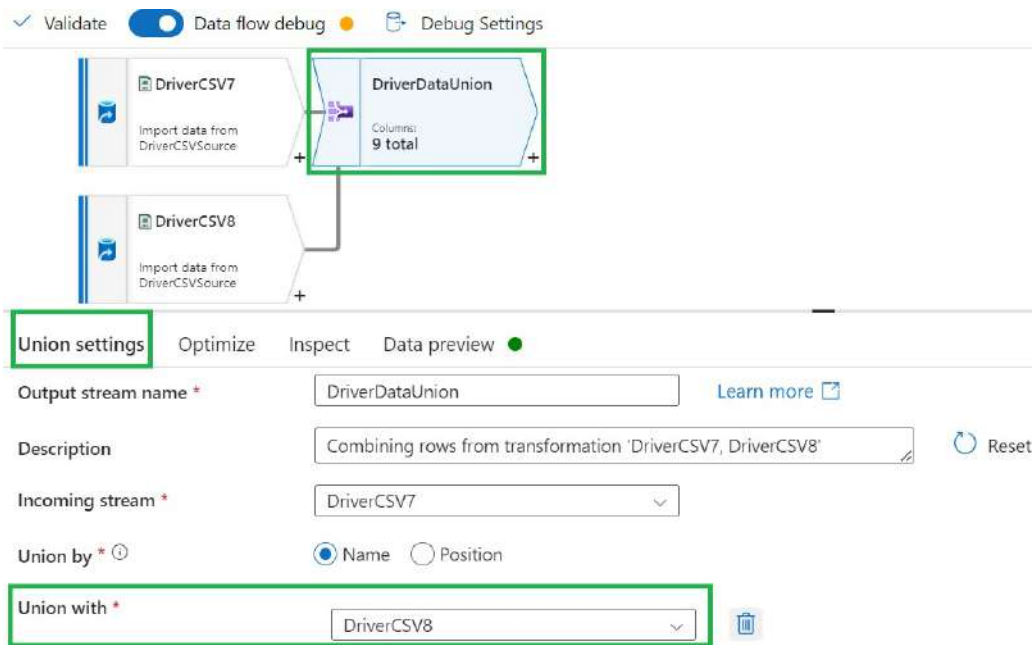

Figure 4.27 – Performing a Union example in ADF

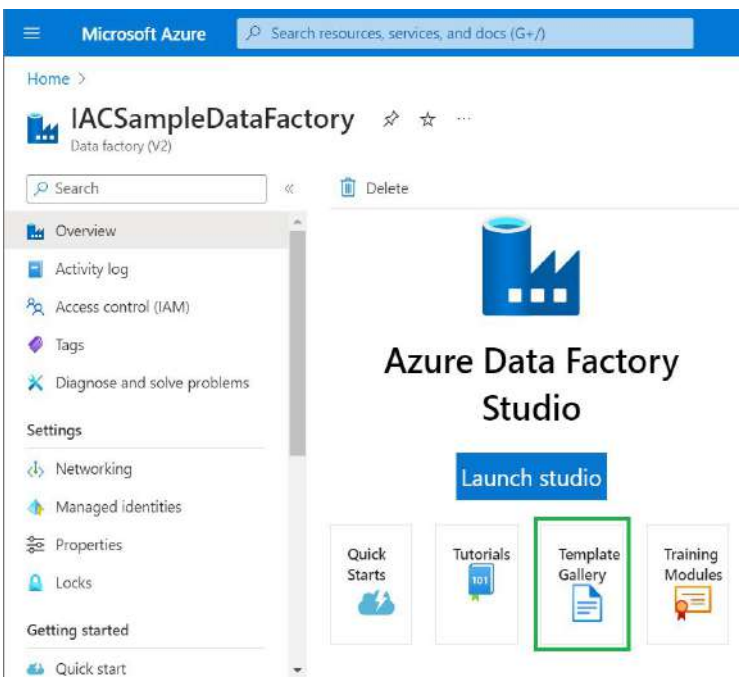

Figure 4.28 – Clicking the Template Gallery link to view the templates in ADF instance

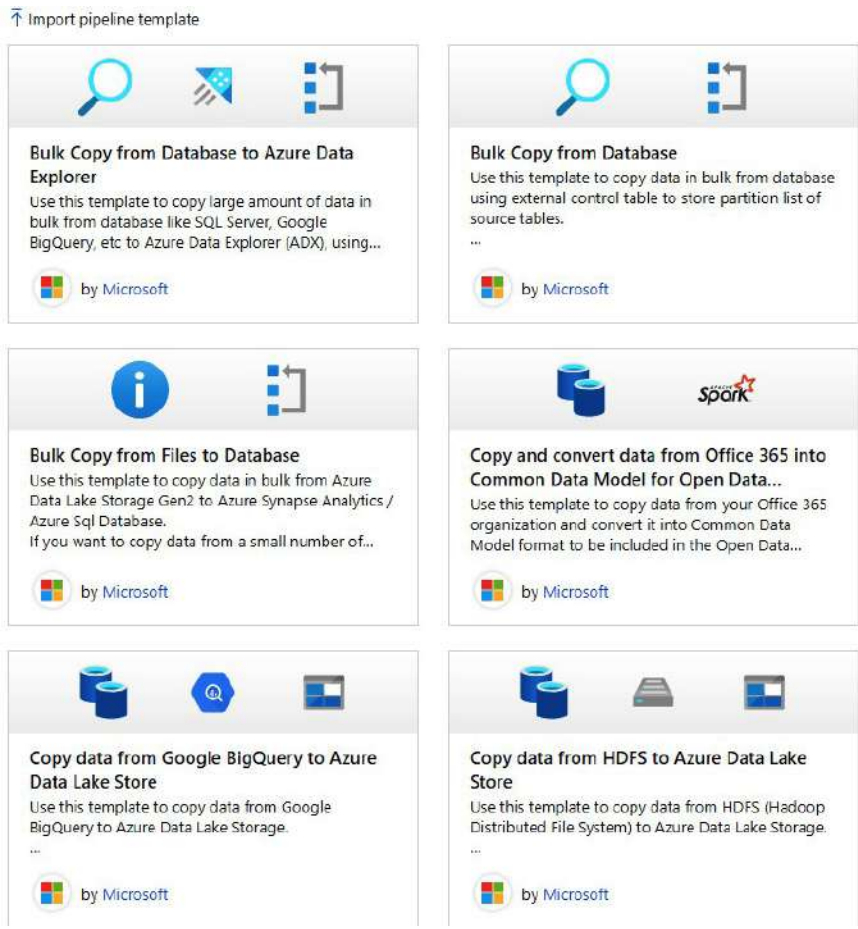

Figure 4.29 – ADF template gallery showing copy and transformation activities templates

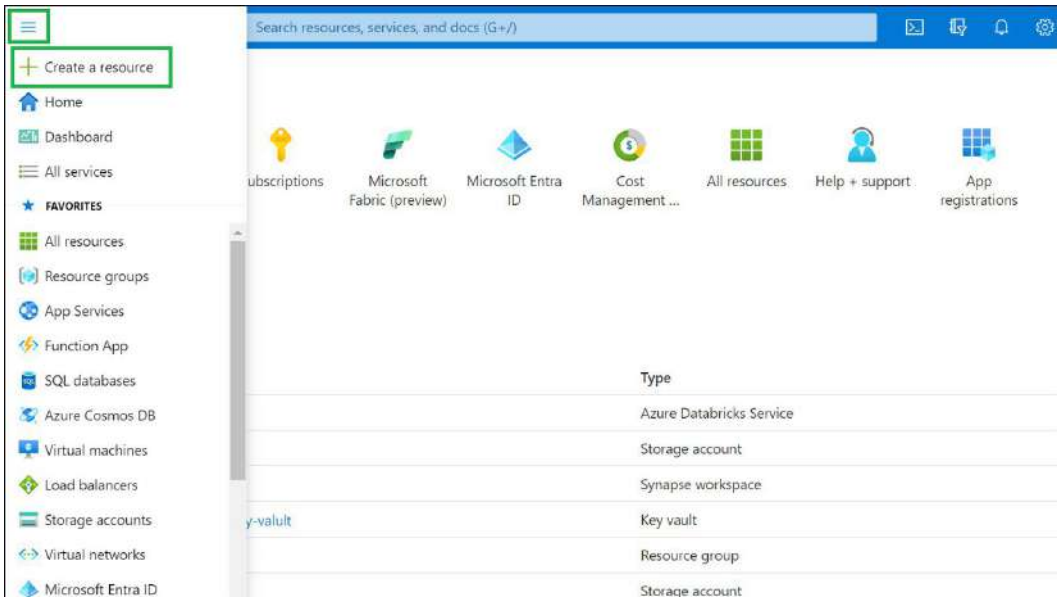

Figure 4.30 – Creating an Azure Synapse Analytics resource in the Azure portal

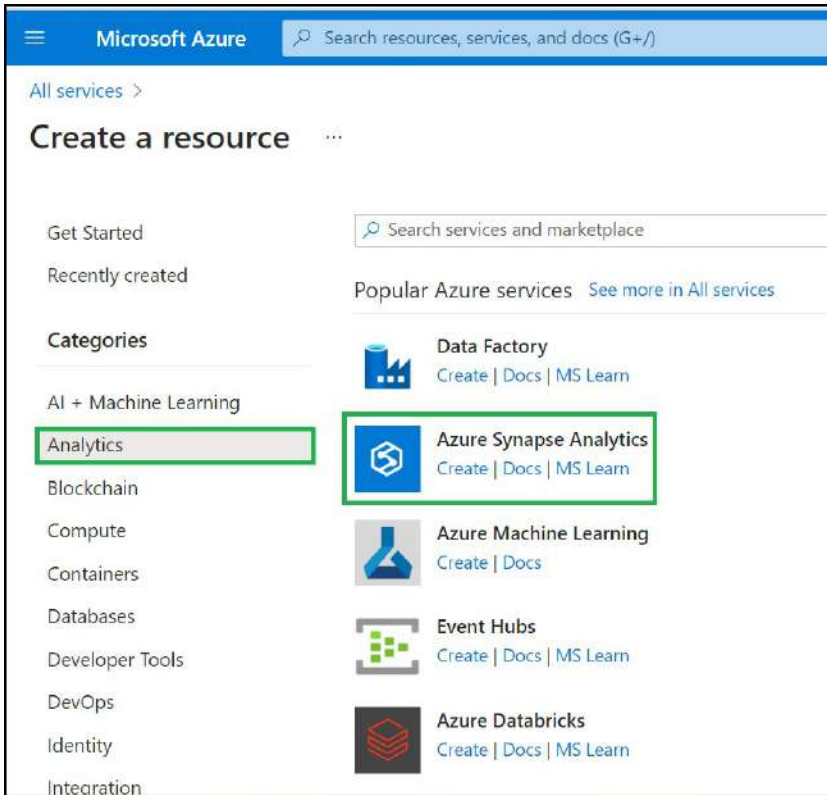

Figure 4.31 – Choosing an Azure Synapse Analytics service in the Azure portal

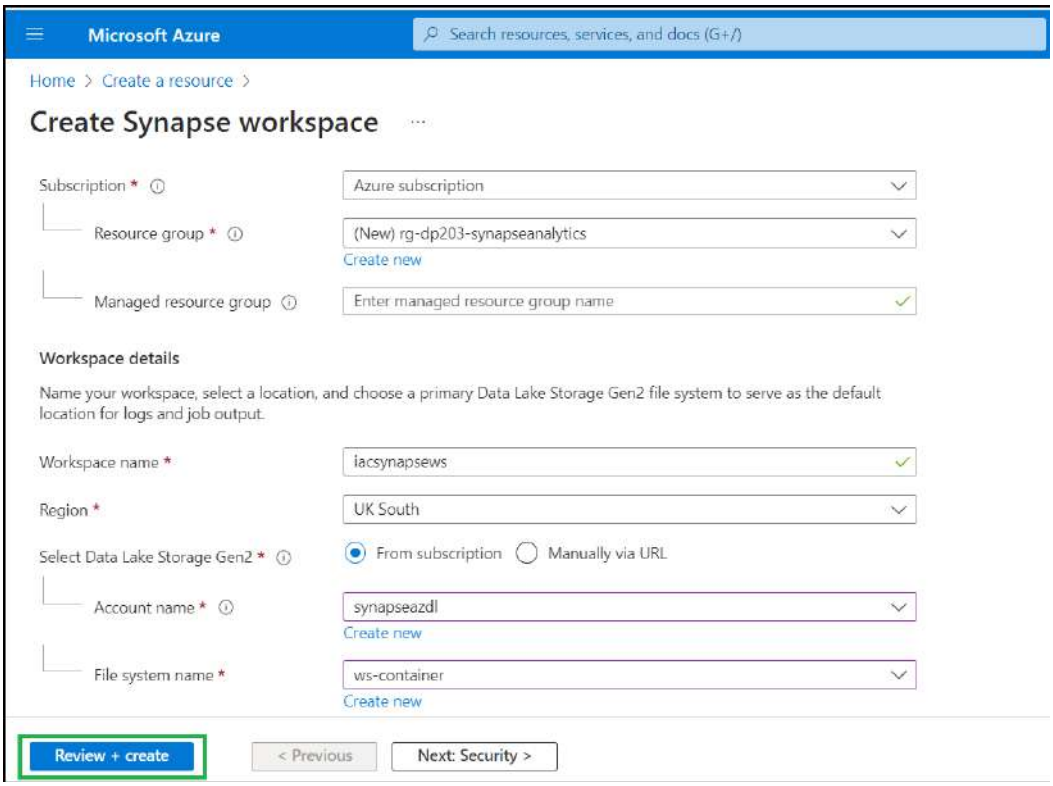

Figure 4.32 – Creating an Azure Synapse Analytics workspace

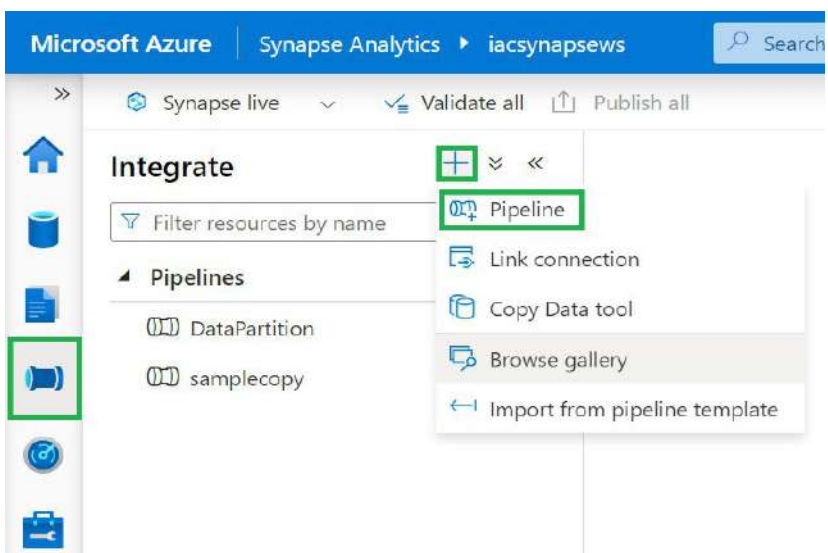

Figure 4.33 – Launching Synapse pipelines to perform transformations

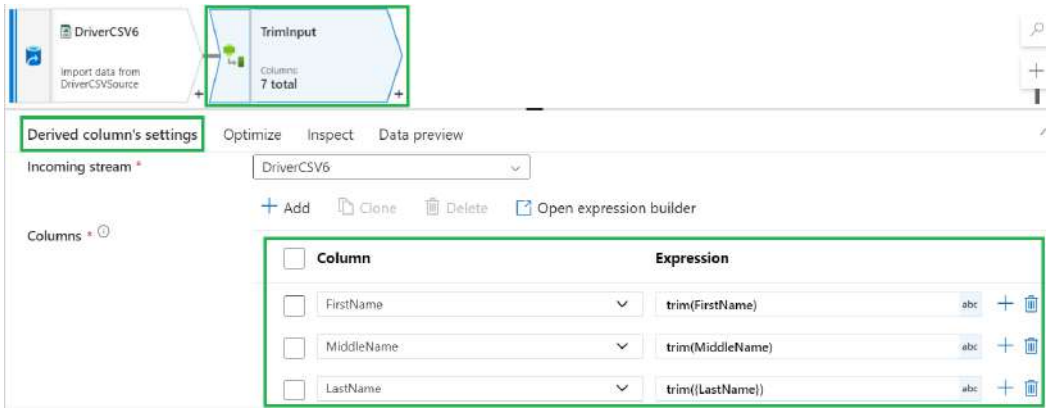

Figure 4.34 – Trimming whitespace in column values

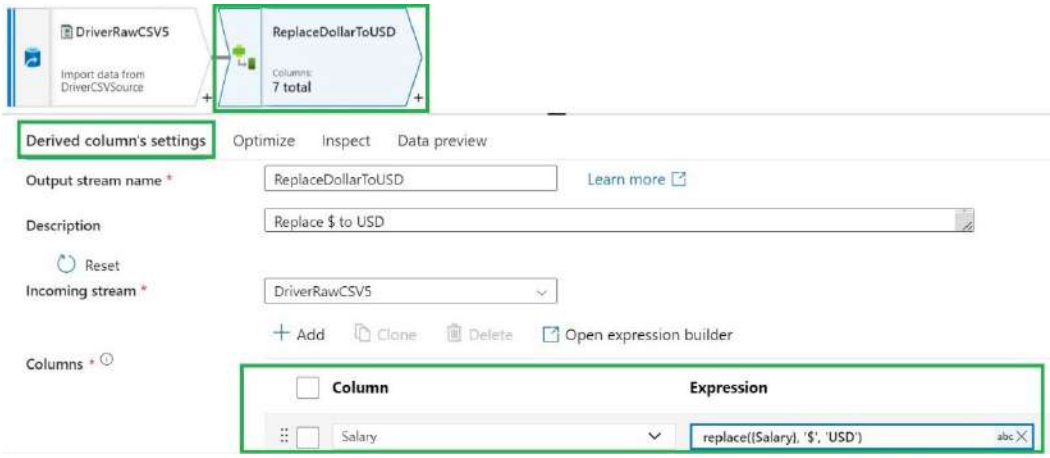

Figure 4.35 – Replacing $ values in Salary column with USD

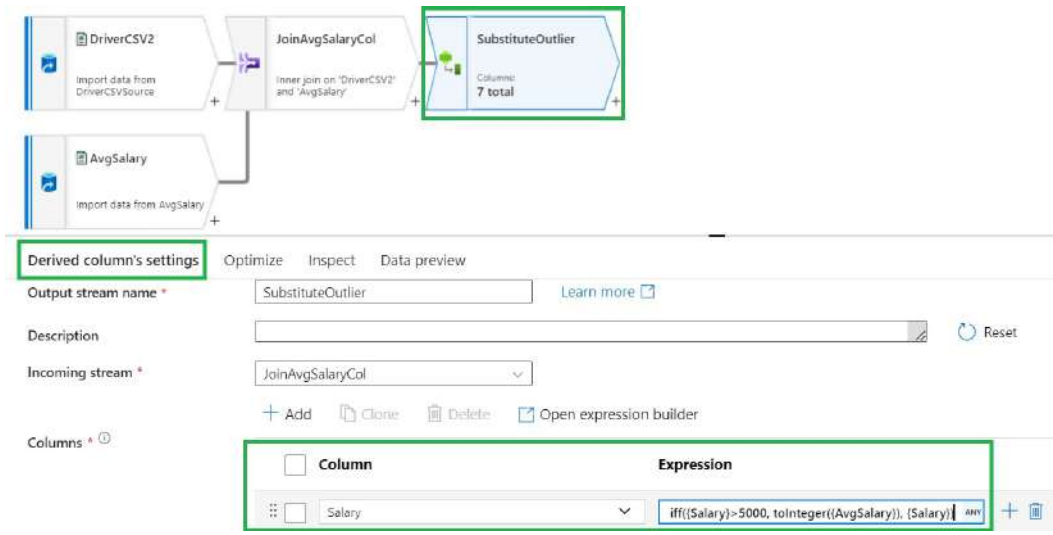

Figure 4.36 – Substituting value with AvgSalary

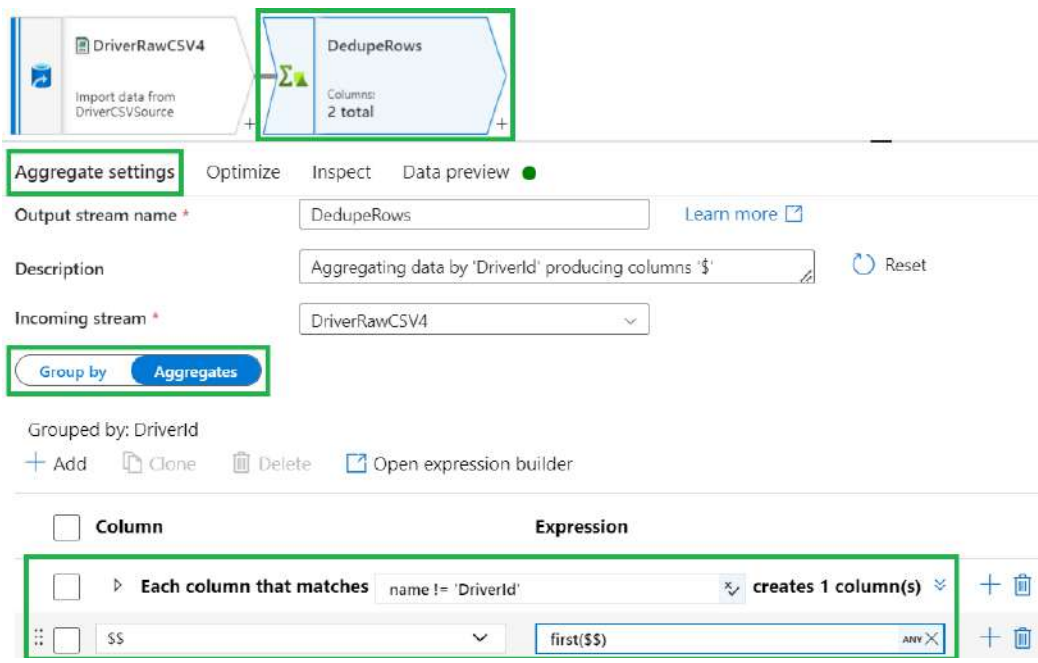

Figure 4.37 – Deduping using the Aggregate transformation

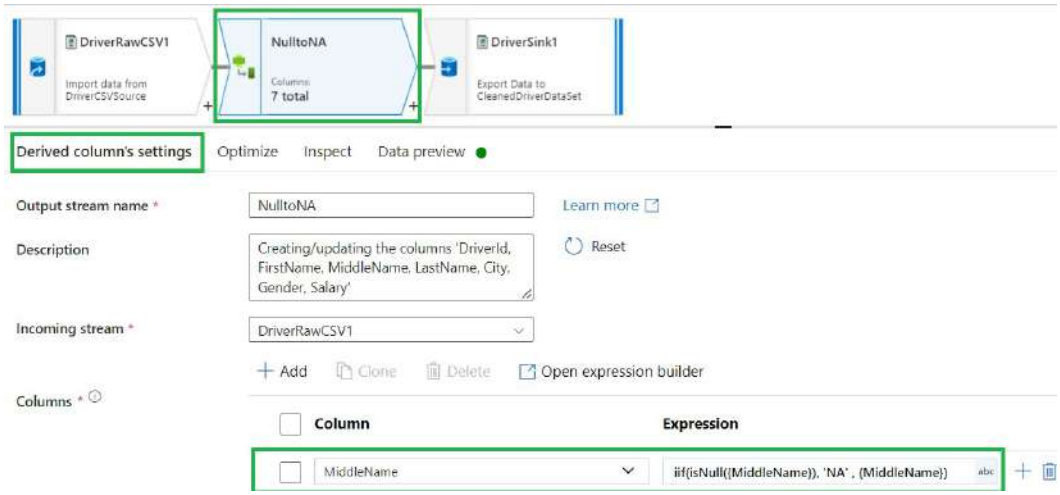

Figure 4.38 – Substituting missing values with default values

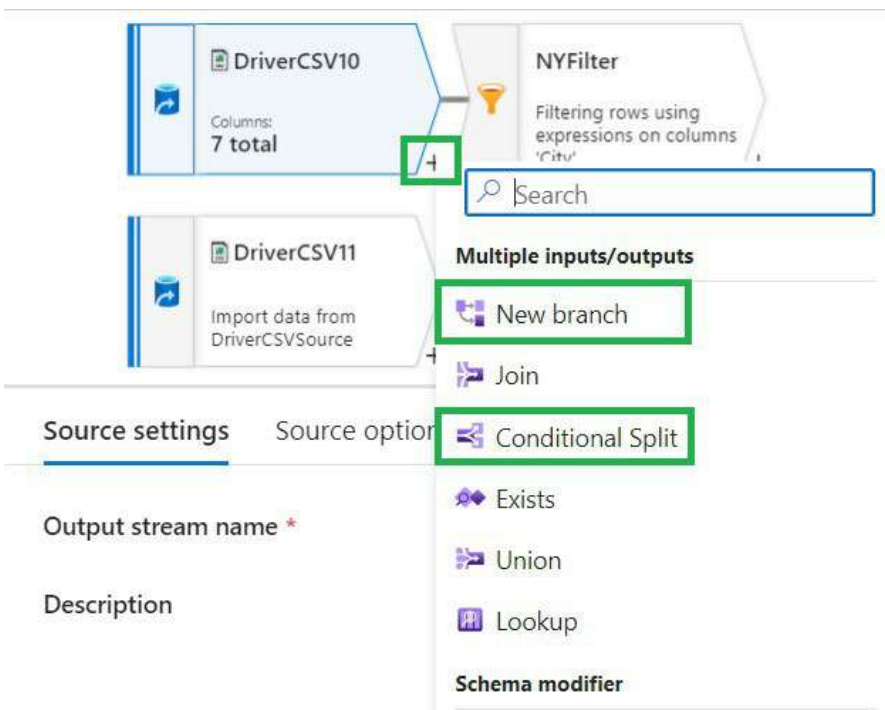

Figure 4.39 – Selecting New branch option to copy the entire dataset for a new execution flow

1.

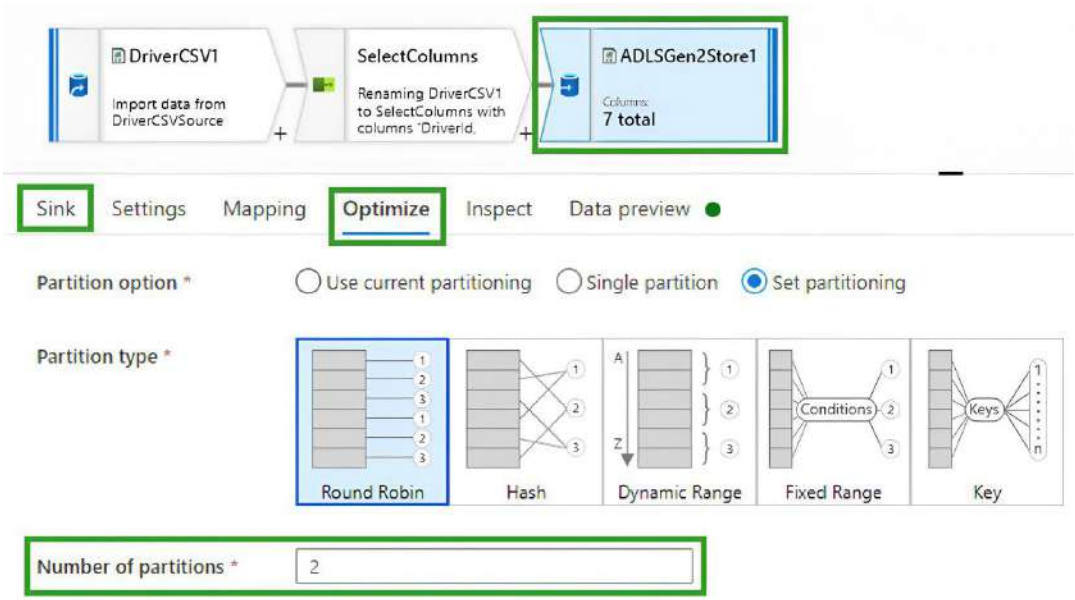

Figure 4.40 – Splitting files using ADF based on the number of partitions

```
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

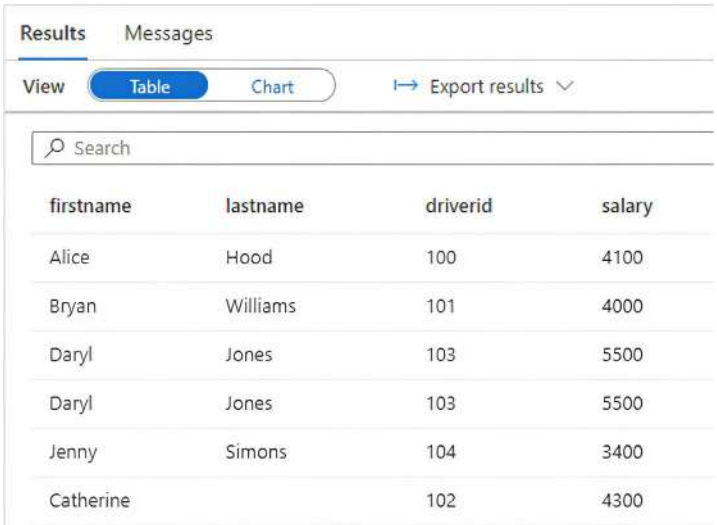

Figure 4.42 – Sample output of parsing JSON using OPENROWSET

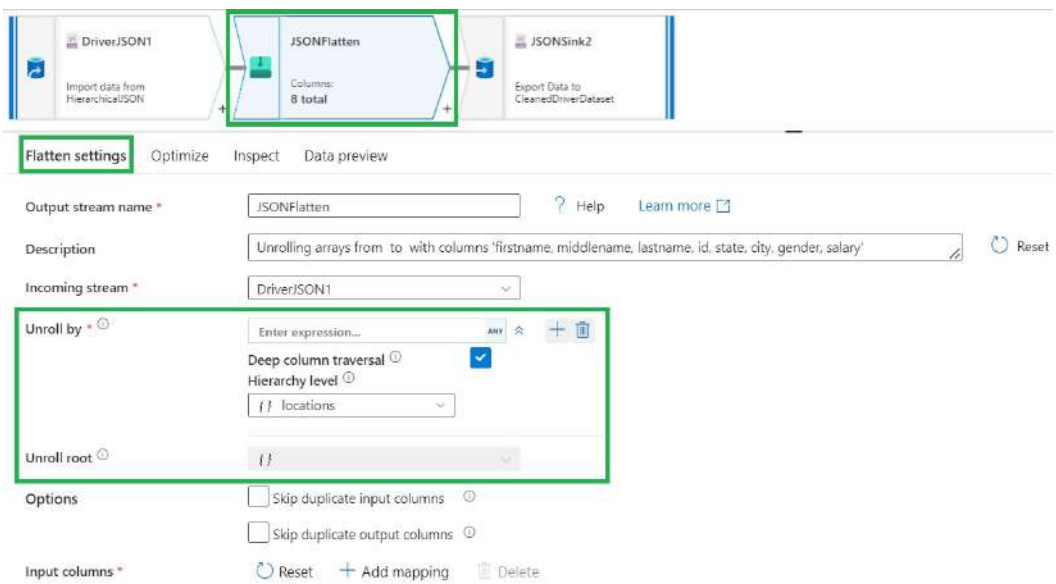

Figure 4.43 – Performing Flatten transformation in ADF

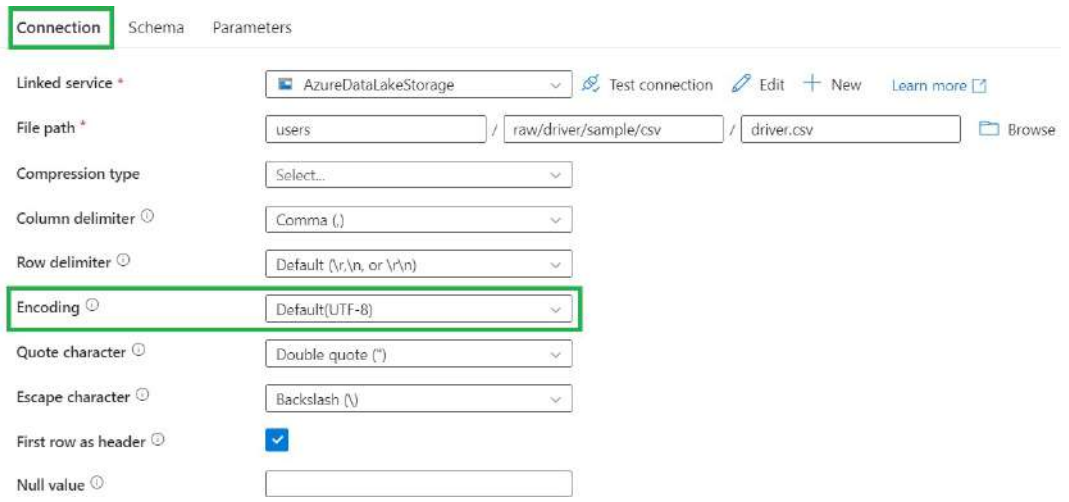

Figure 4.44 – Using the Encoding function in ADF source datasets

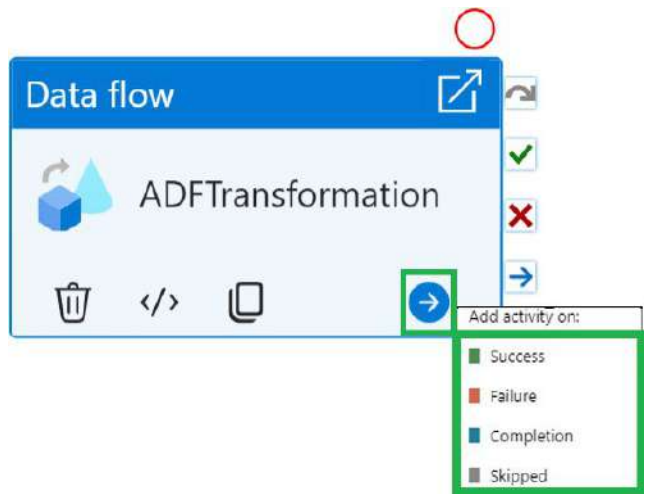

Figure 4.45 – ADF supporting four activity flows

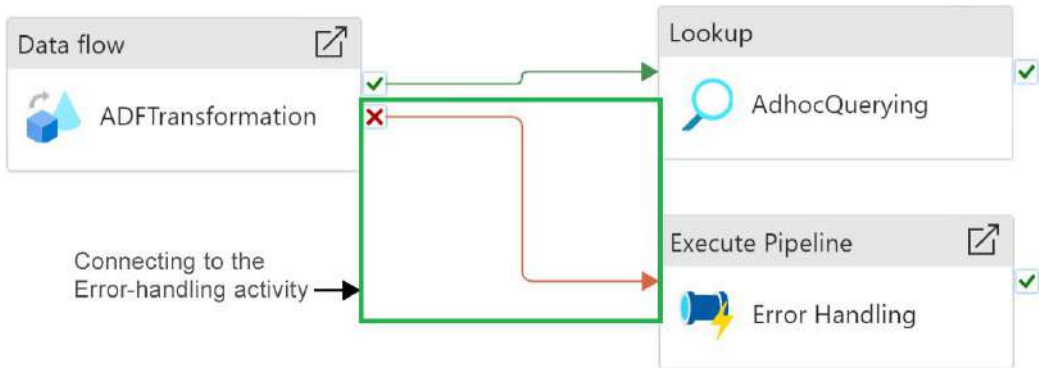

Figure 4.46 – Creating an error-handling pipeline

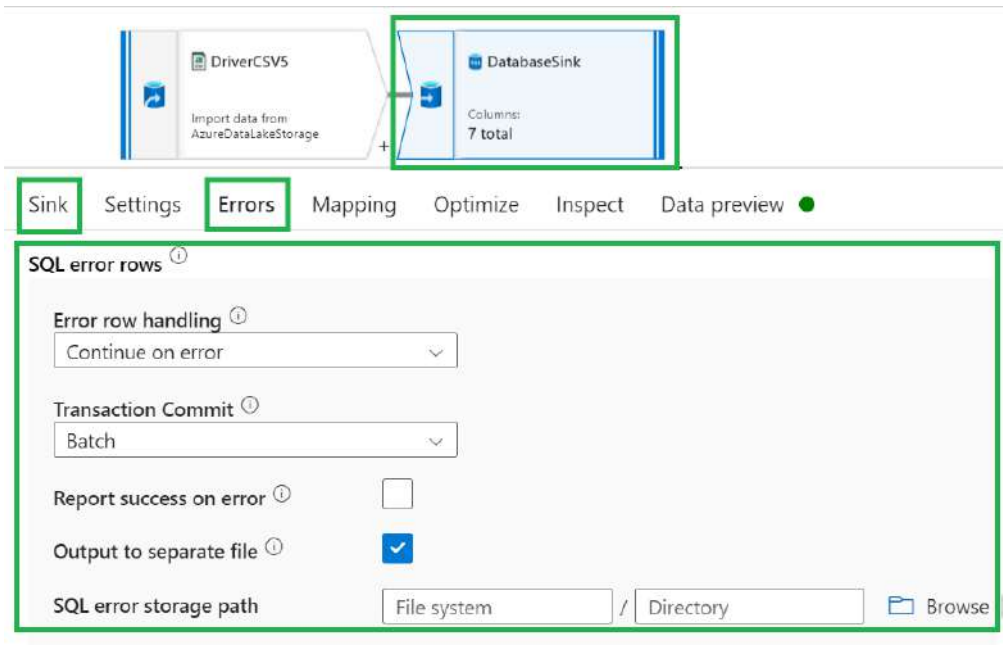

Figure 4.47 – Redirecting error lines to Blob Storage

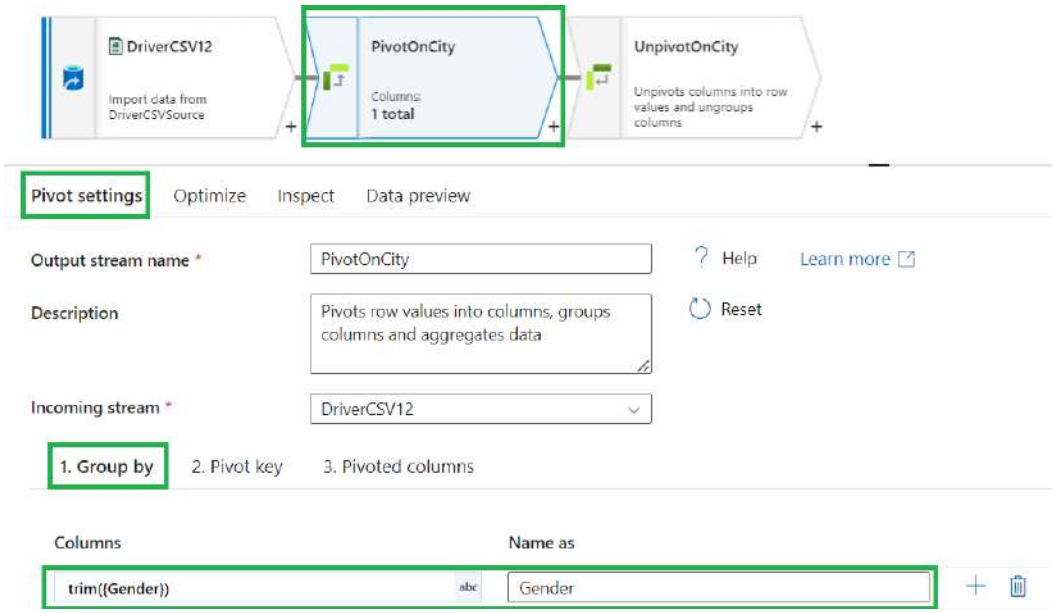

Figure 4.48 – Using Group by settings on Gender for the Pivot operation

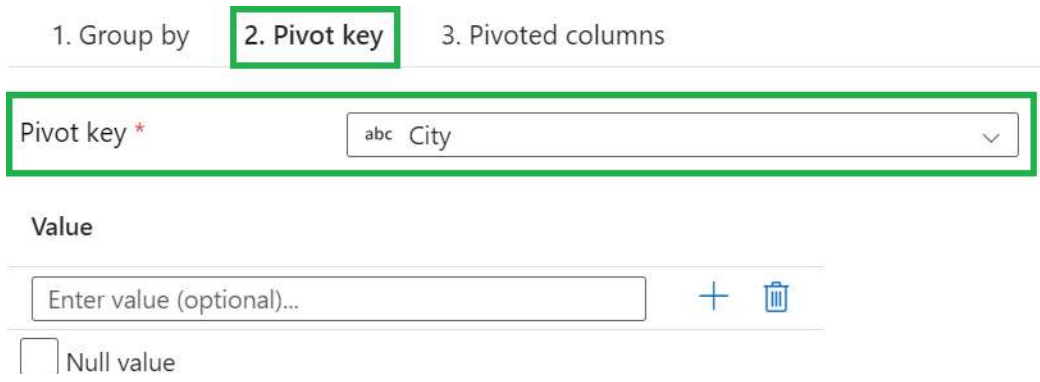

Figure 4.49 – Using City for the Pivot operation

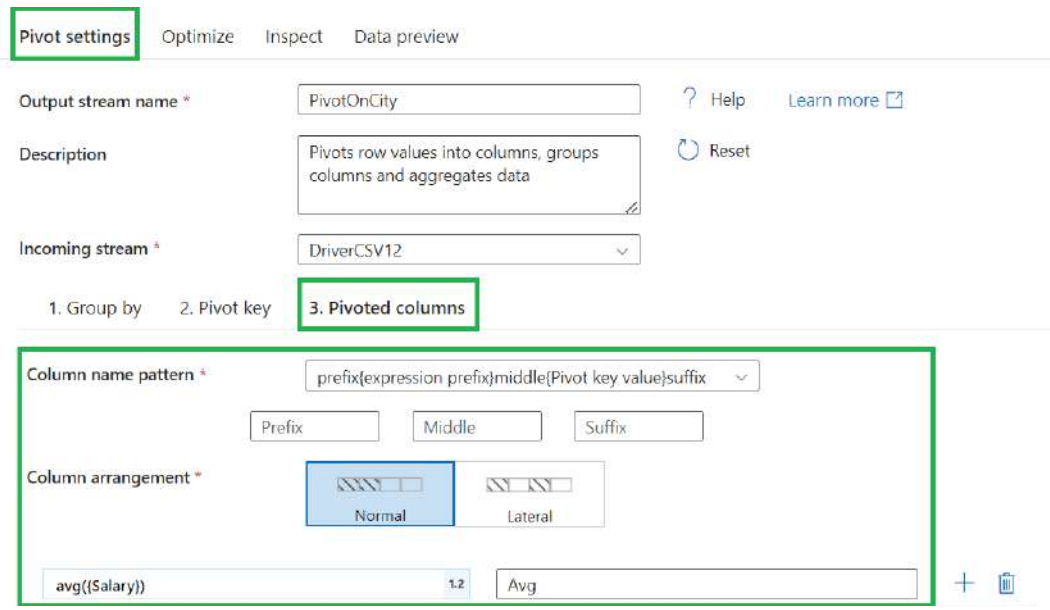

Figure 4.50 – Specifying the column name pattern and aggregation for the Pivot function

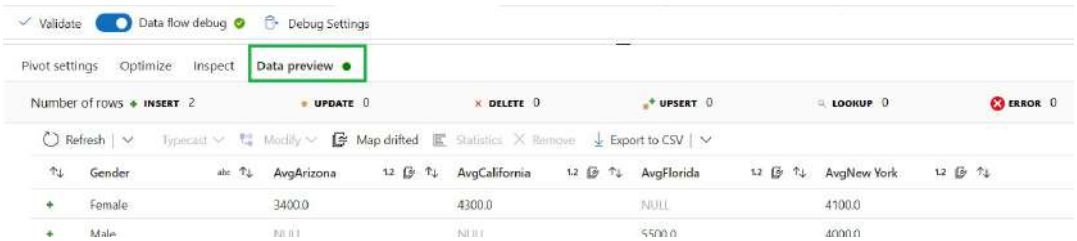

Figure 4.51 – The post-Pivot table showing values

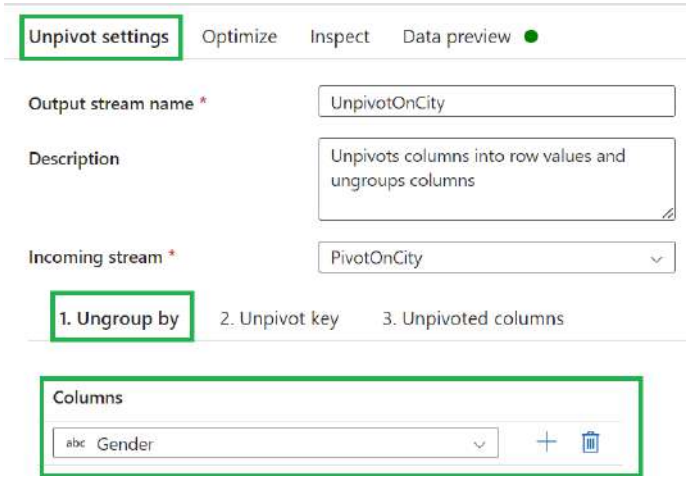

Figure 4.52 – Using the Ungroup by tab for Unpivoting on a single column Gender

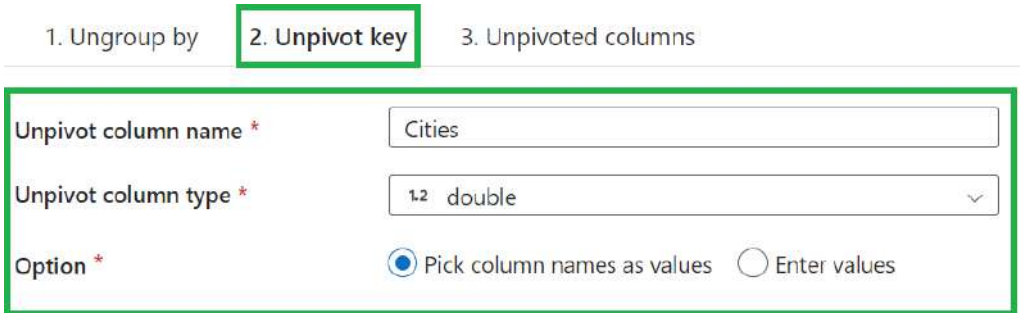

Figure 4.53 – Specifying Unpivot key for the Unpivot operation

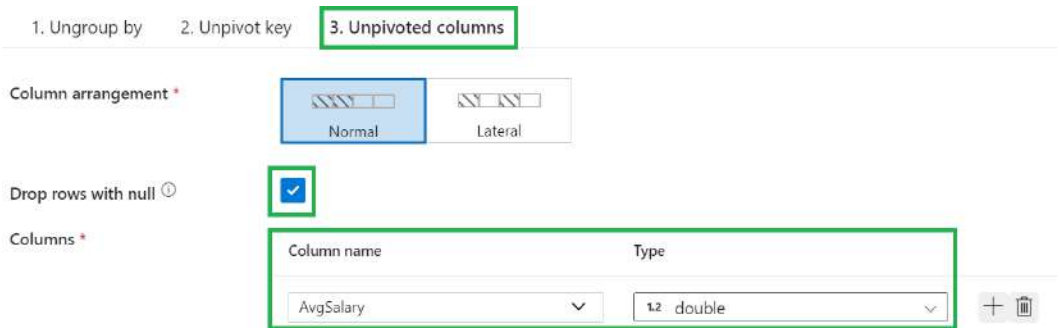

Figure 4.54 – Using the Unpivoted columns tab to calculate AvgSalary by Gender and Cities

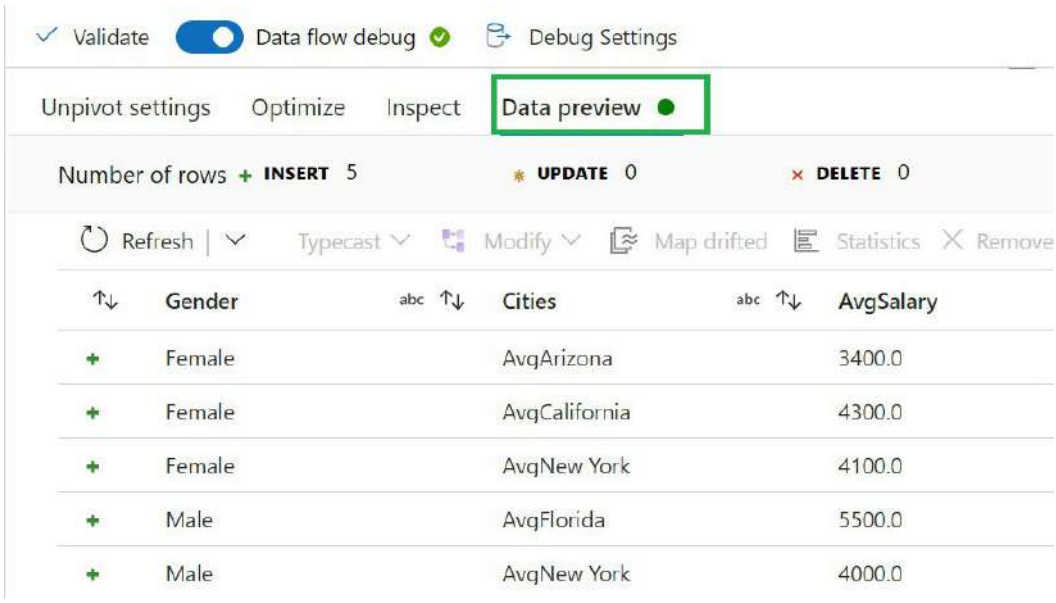

Figure 4.55 – Displaying an unpivoted table as output

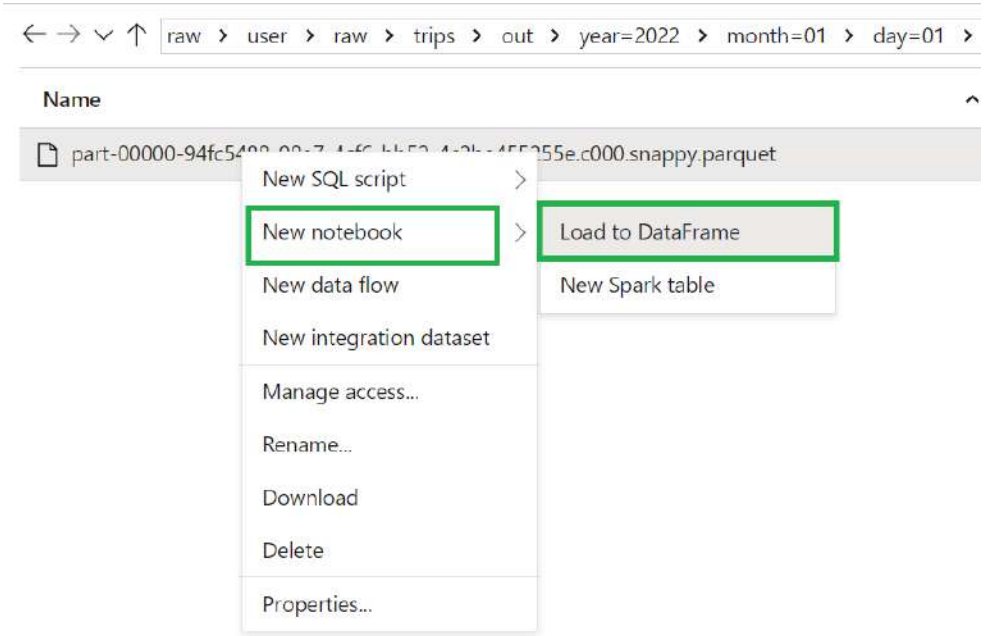

Figure 4.56 – Launching a DataFrame from the Synapse data file

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

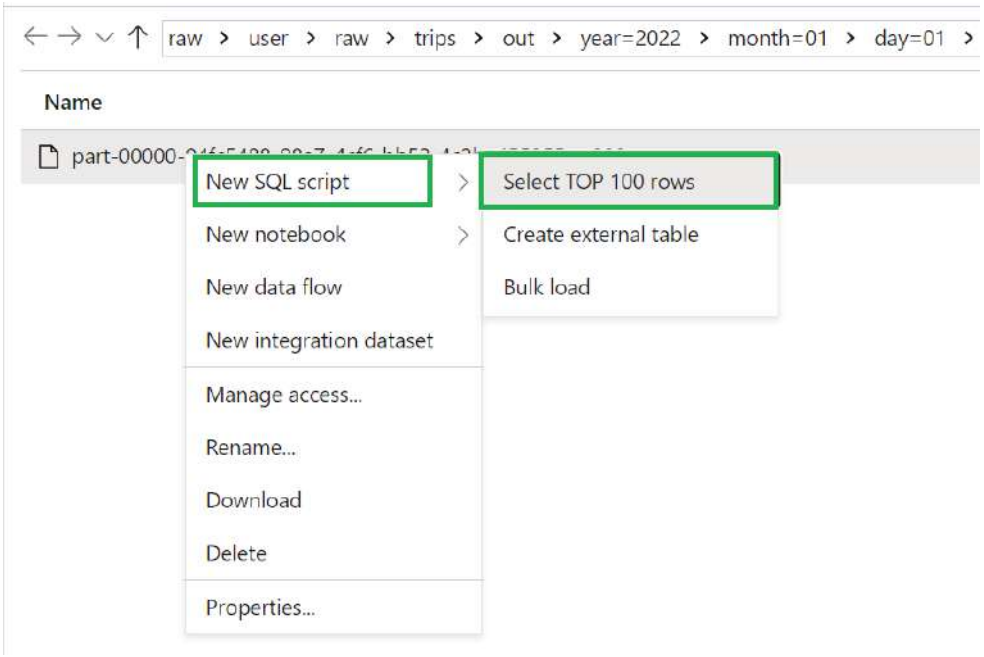

Figure 4.58 – Launching a SQL script to explore data from the Synapse data file

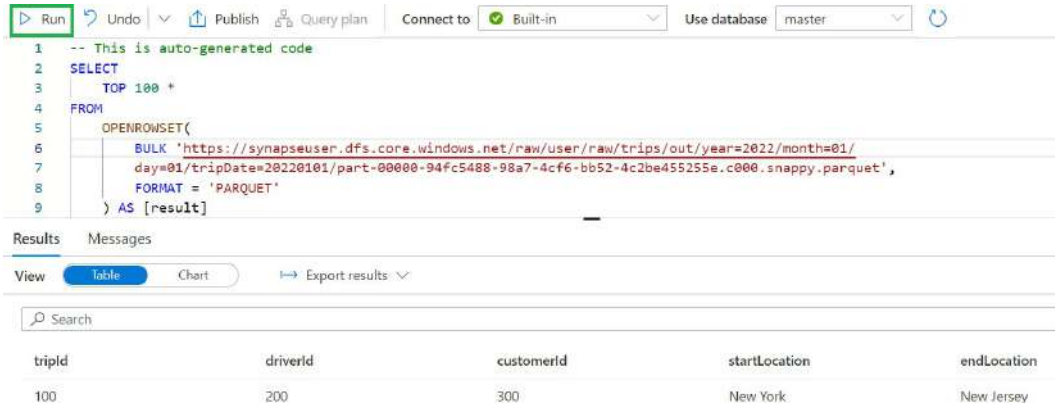

Figure 4.59 – SQL script auto-launched by Synapse to help explore data

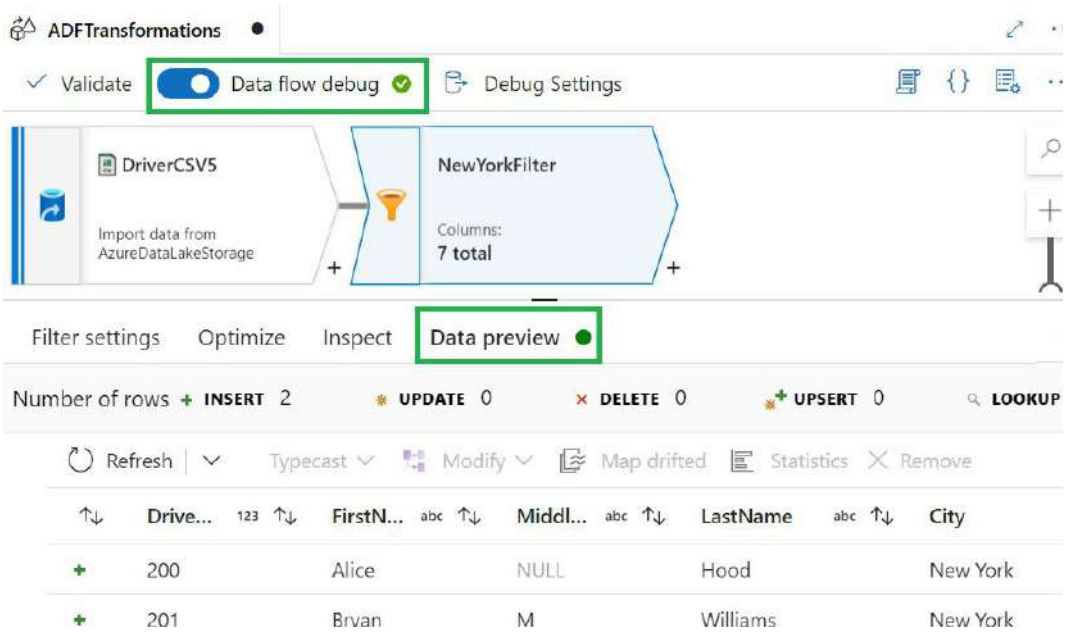

Figure 4.60 – ADF Data preview tab showing fine-tuned transformations and pipelines

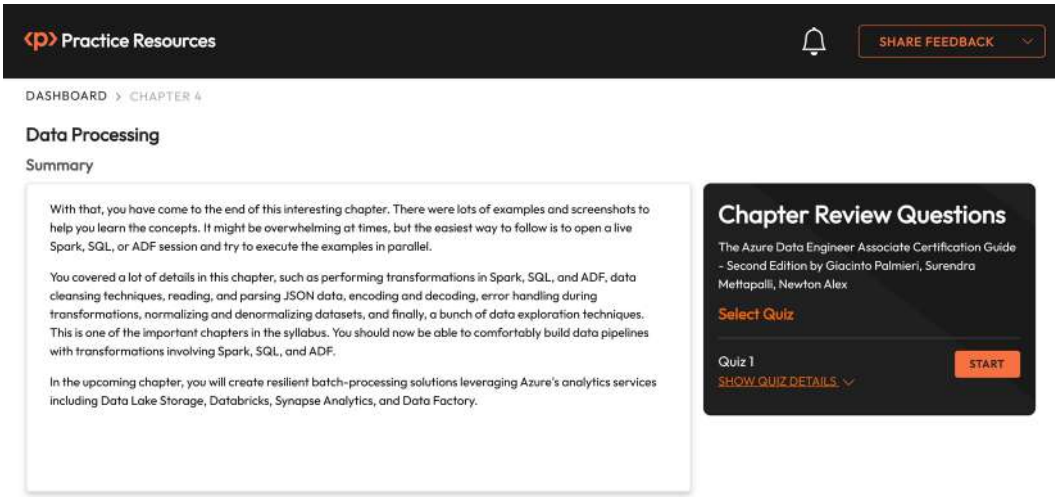

Figure 4.62 – Chapter Review Questions for Chapter 4

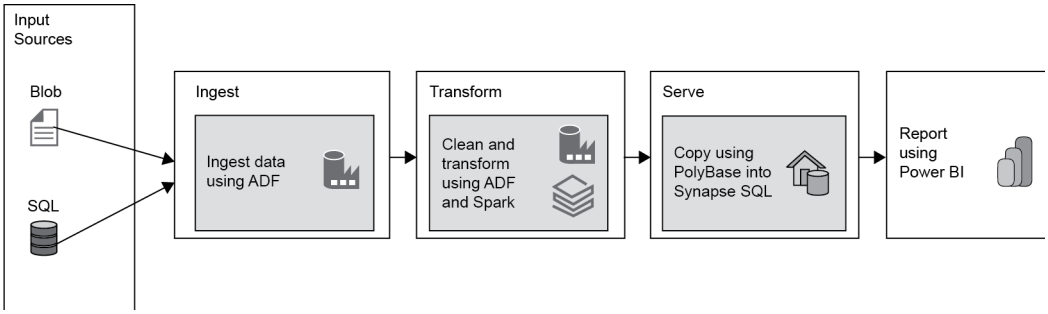

Figure 5.1 – High-level architecture of the batch use case

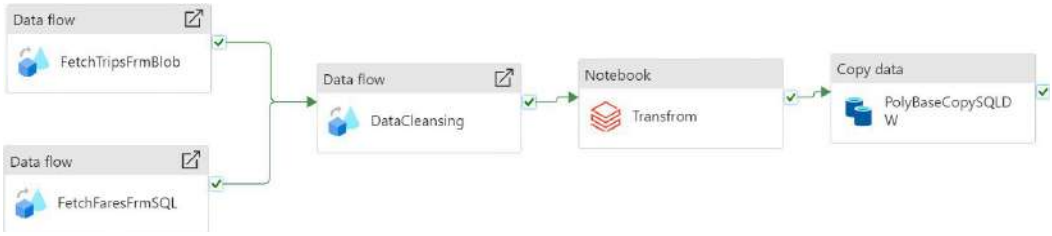

Figure 5.2 – Sample data pipeline in ADF

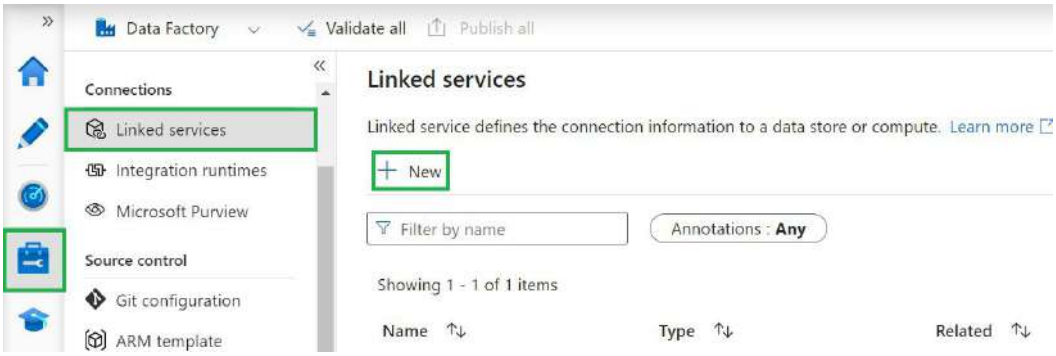

Figure 5.3 – Configuring a linked service in ADF

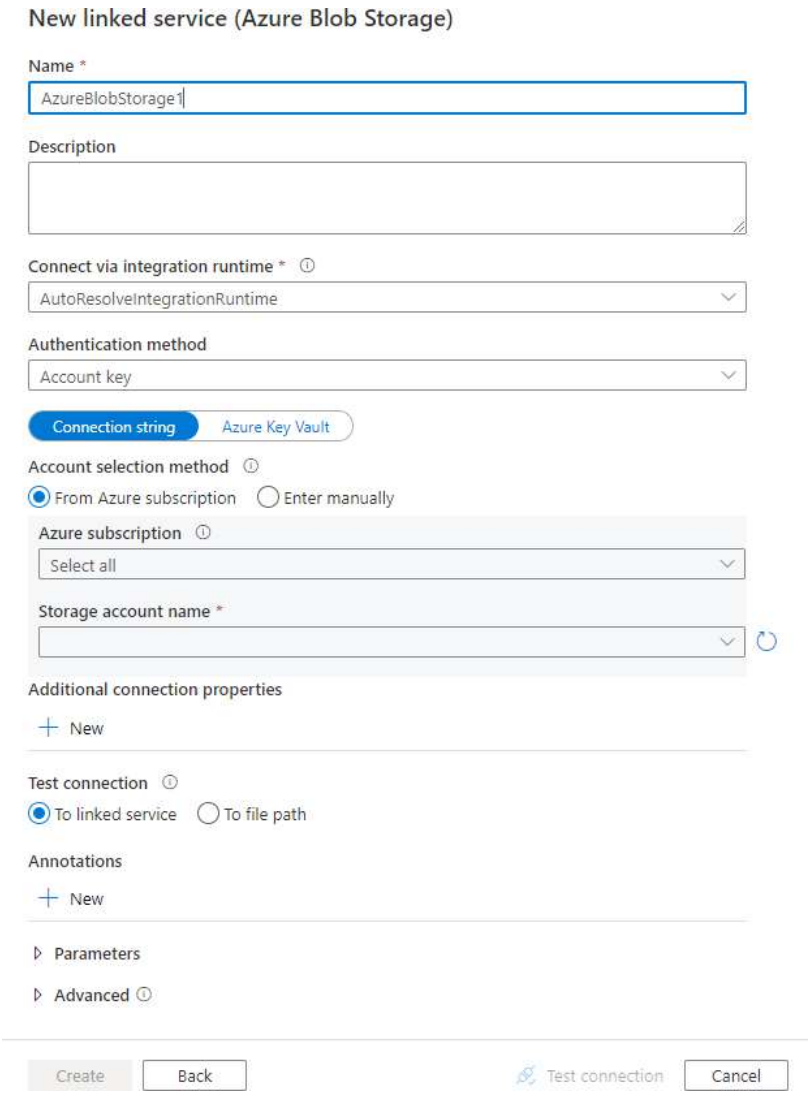

Figure 5.4 – Creating an Azure Blob Storage linked service

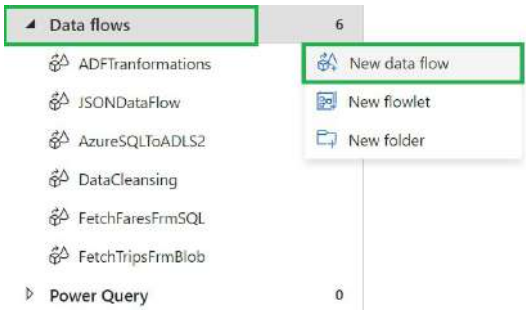

Figure 5.5 – Creating a new data flow in ADF

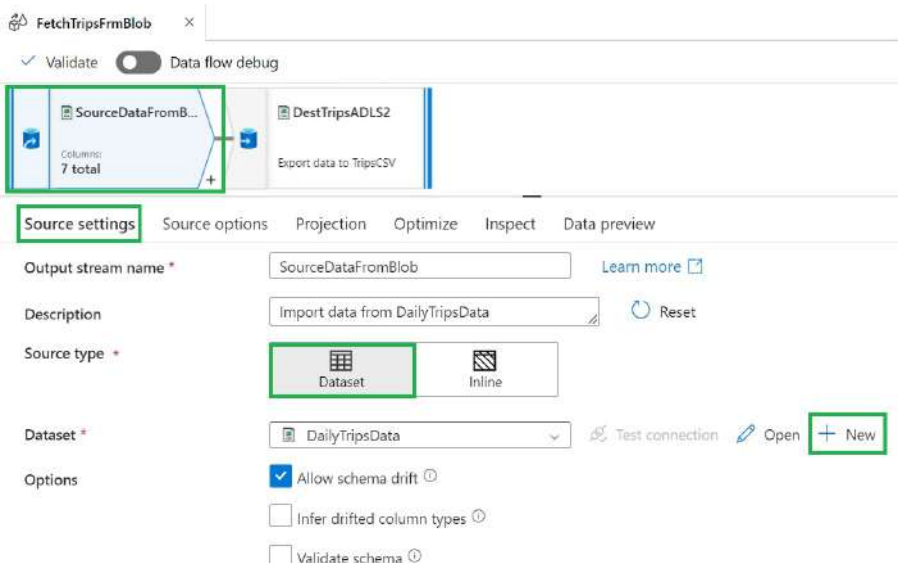

Figure 5.6 – Creating source and destination datasets in a data flow

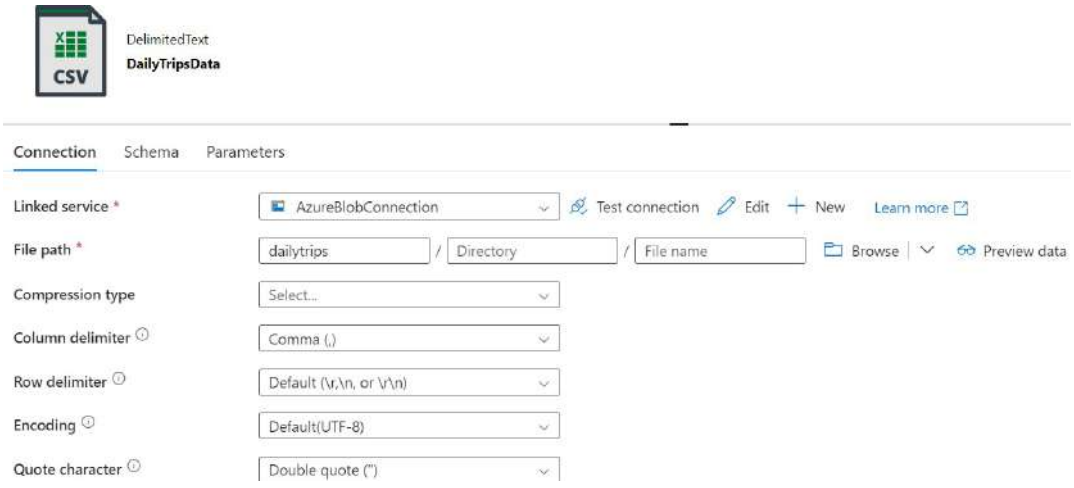

Figure 5.7 – Sample dataset screen to create a dataset named DailyTripData

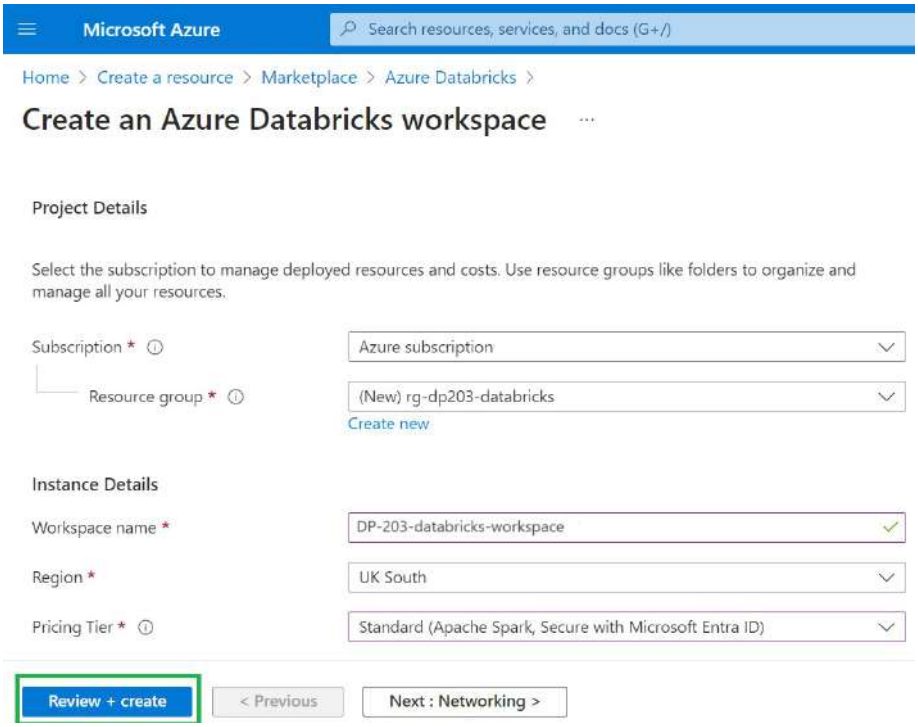

Figure 5.8 – Sample Azure Databricks workspace creation screen

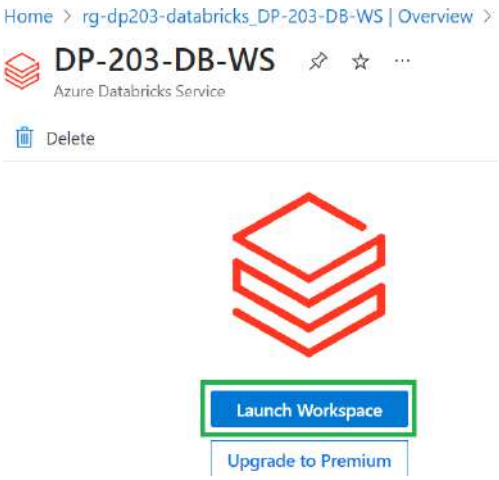

Figure 5.9 – Launching the Azure Databricks workspace

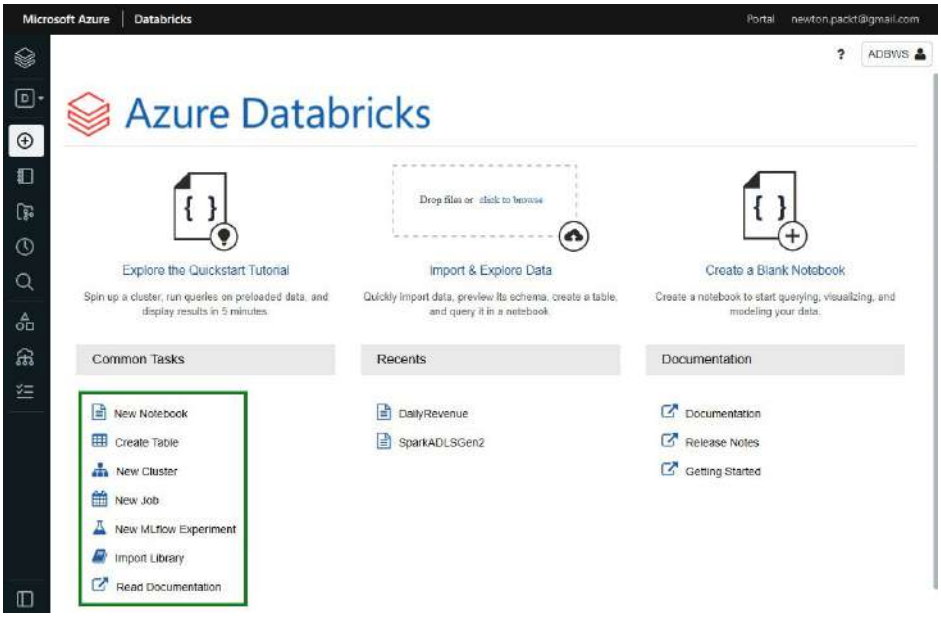

Figure 5.10 – Azure Databricks portal

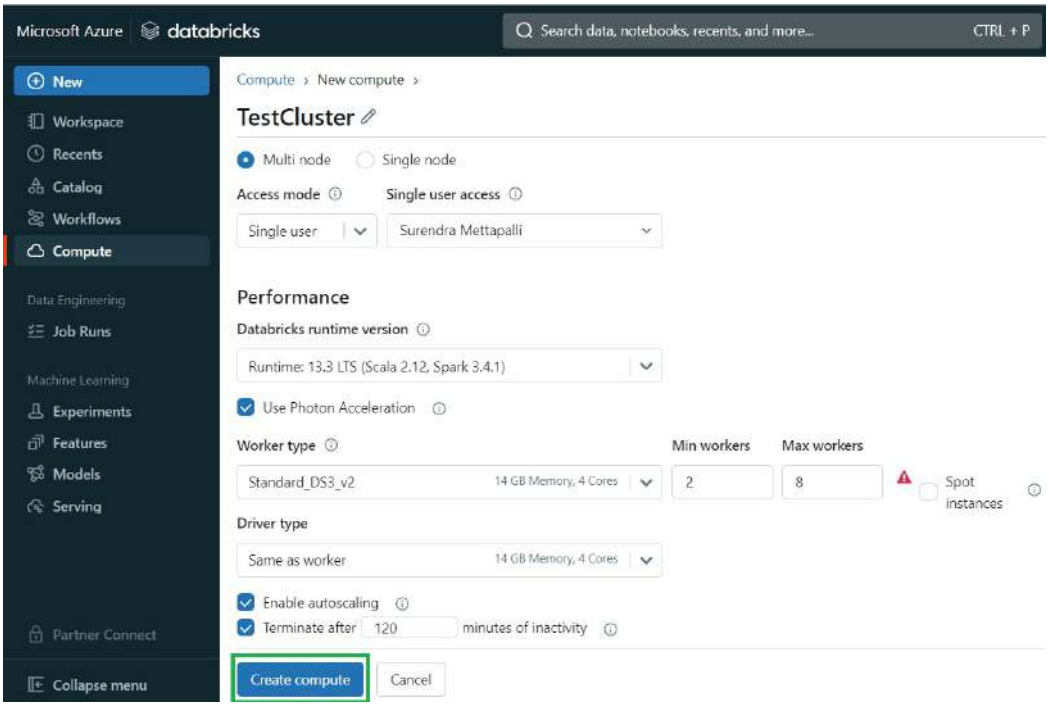

Figure 5.11 – Azure Databricks cluster creation screen

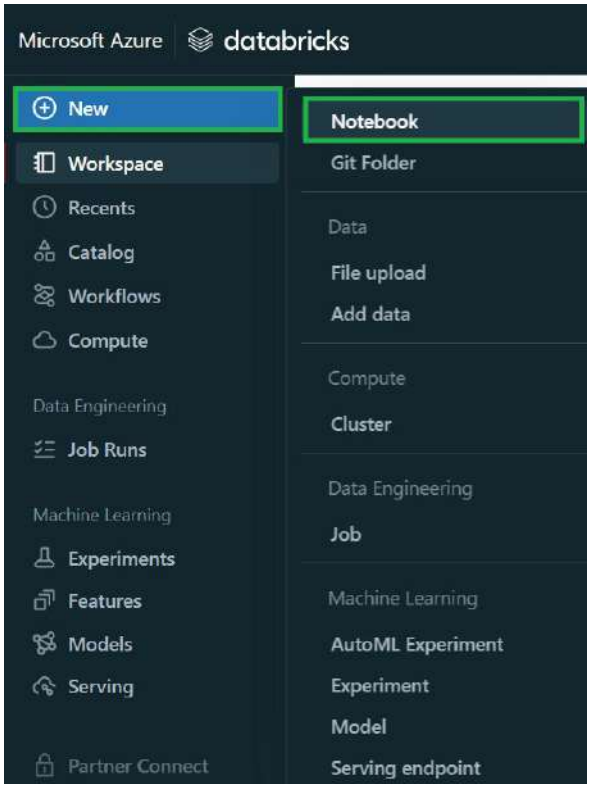

Figure 5.12 – Azure Databricks new notebook creation

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

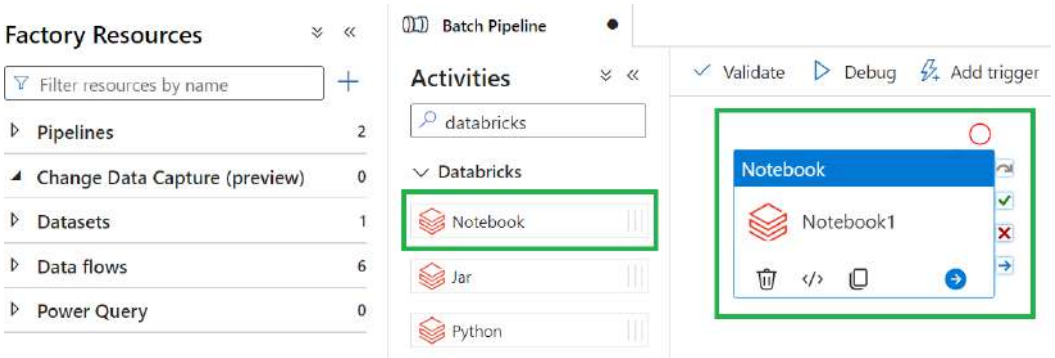

Figure 5.14 – Choosing an Azure Databricks activity in ADF

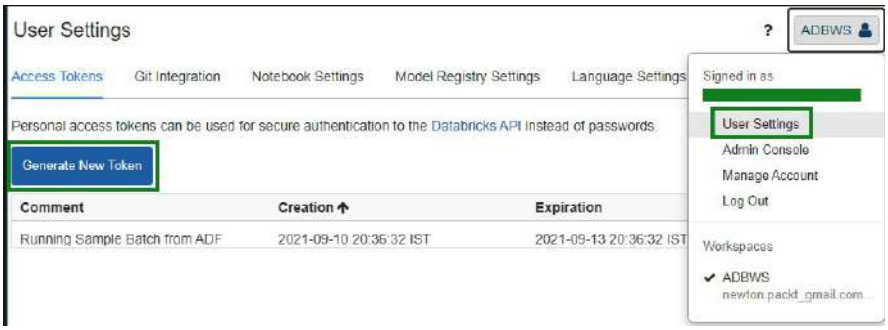

Figure 5.15 – Selecting the access token from the Azure Databricks portal

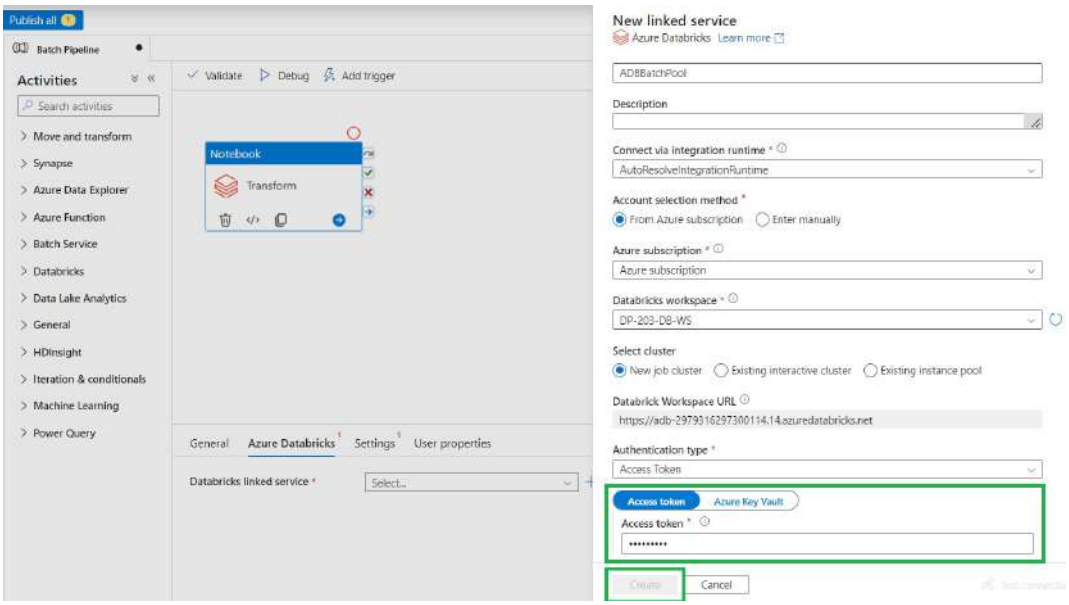

Figure 5.16 – Creating an Azure Databricks linked service

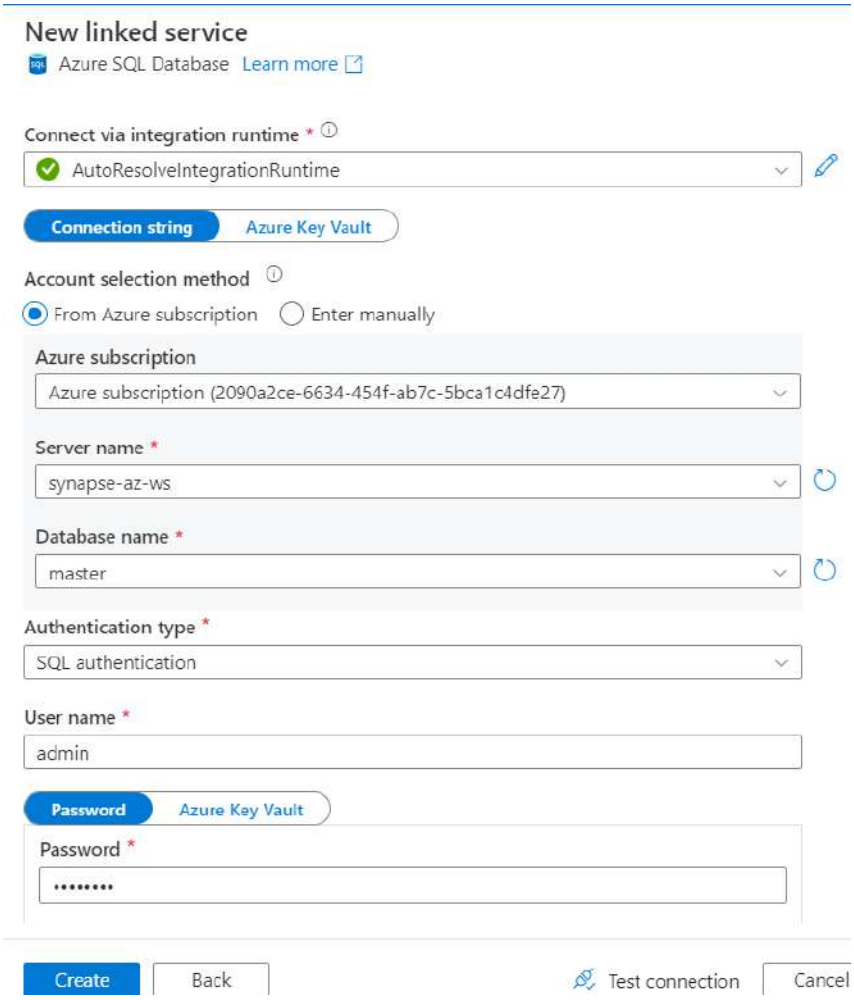

Figure 5.17 – Creating a linked service for Synapse Link

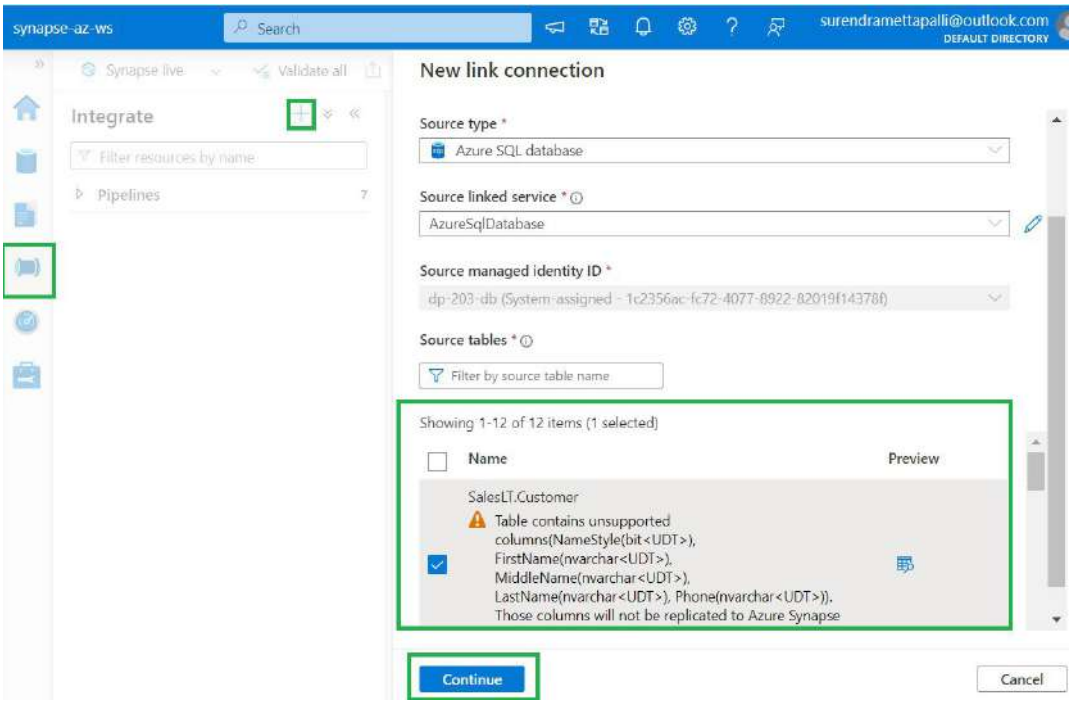

Figure 5.18 – Data replication with Synapse Link

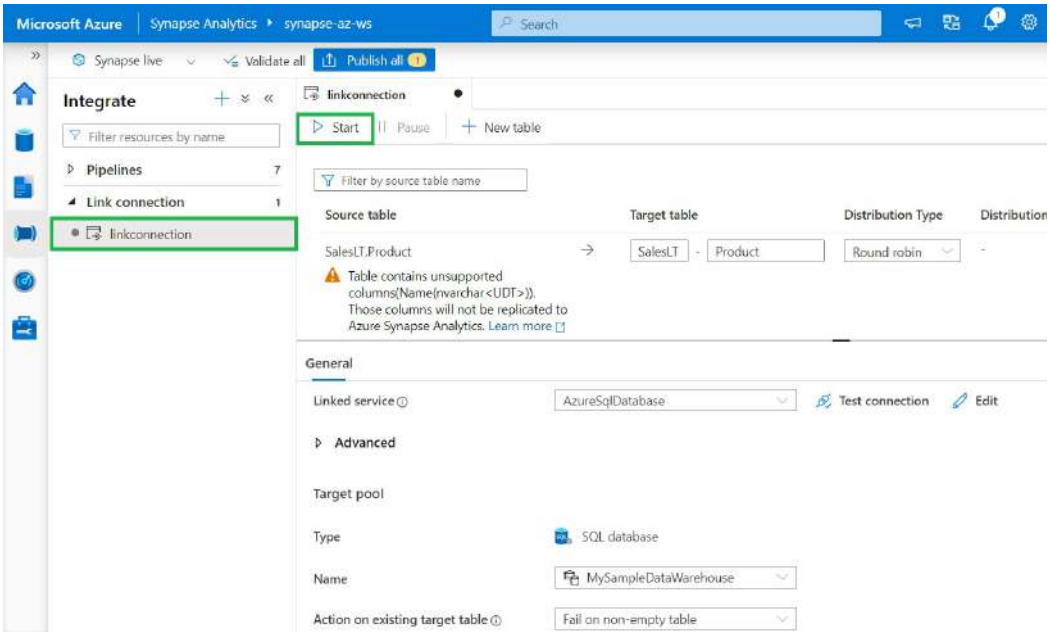

Figure 5.19 – Monitoring replication with Synapse Link

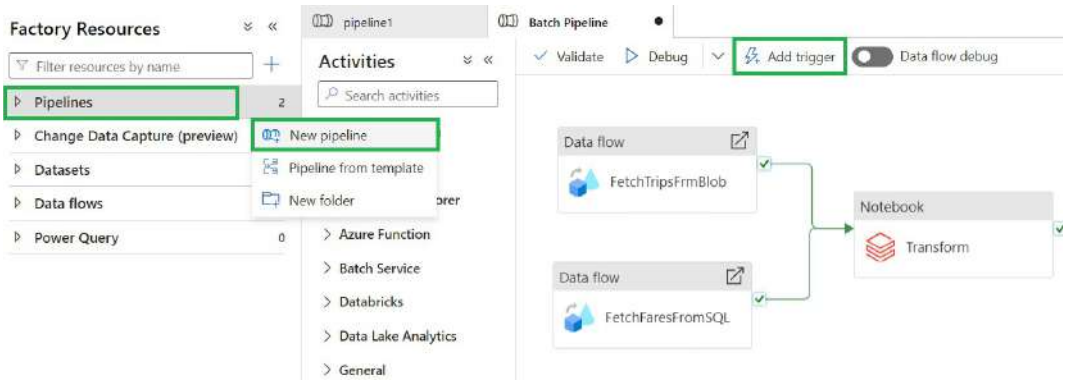

Figure 5.20 – Creating new pipelines in ADF

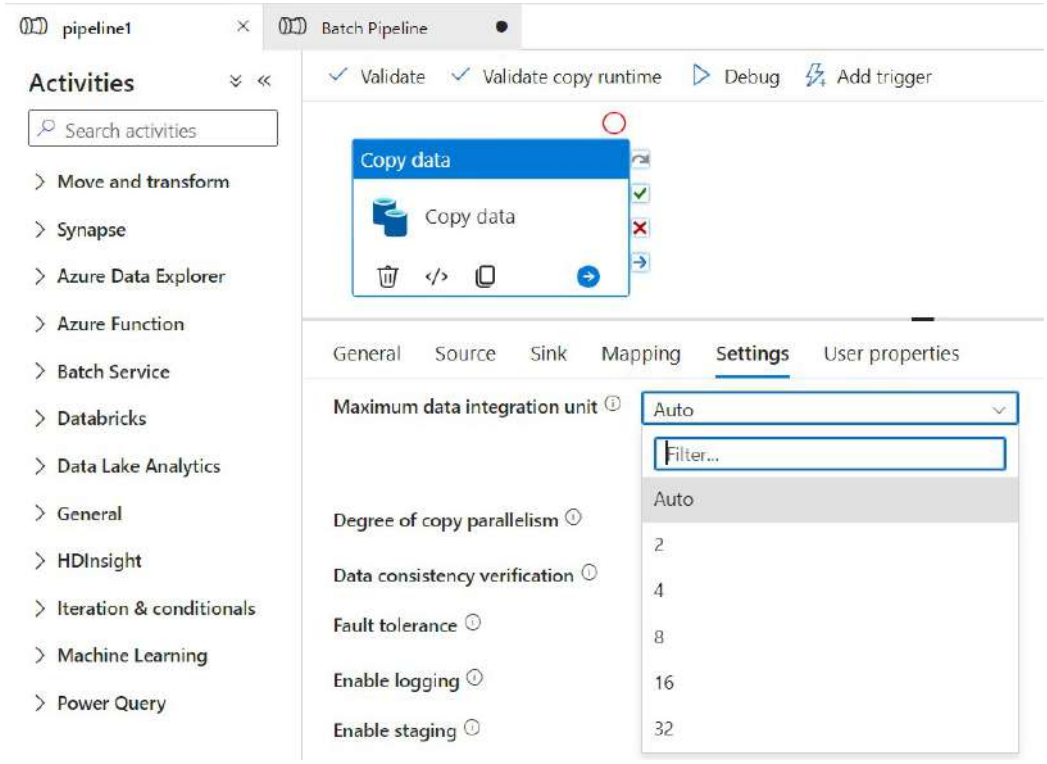

Figure 5.21 – Specifying data integration units for the Copy activity

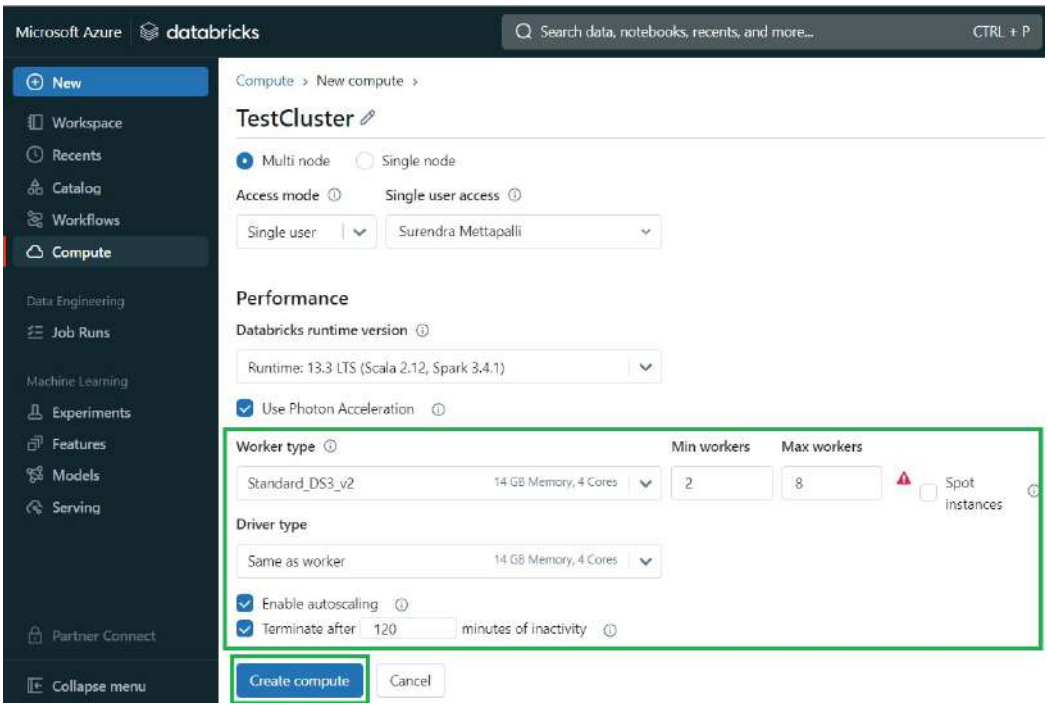

Figure 5.22 – Enabling the Azure Databricks cluster autoscale option

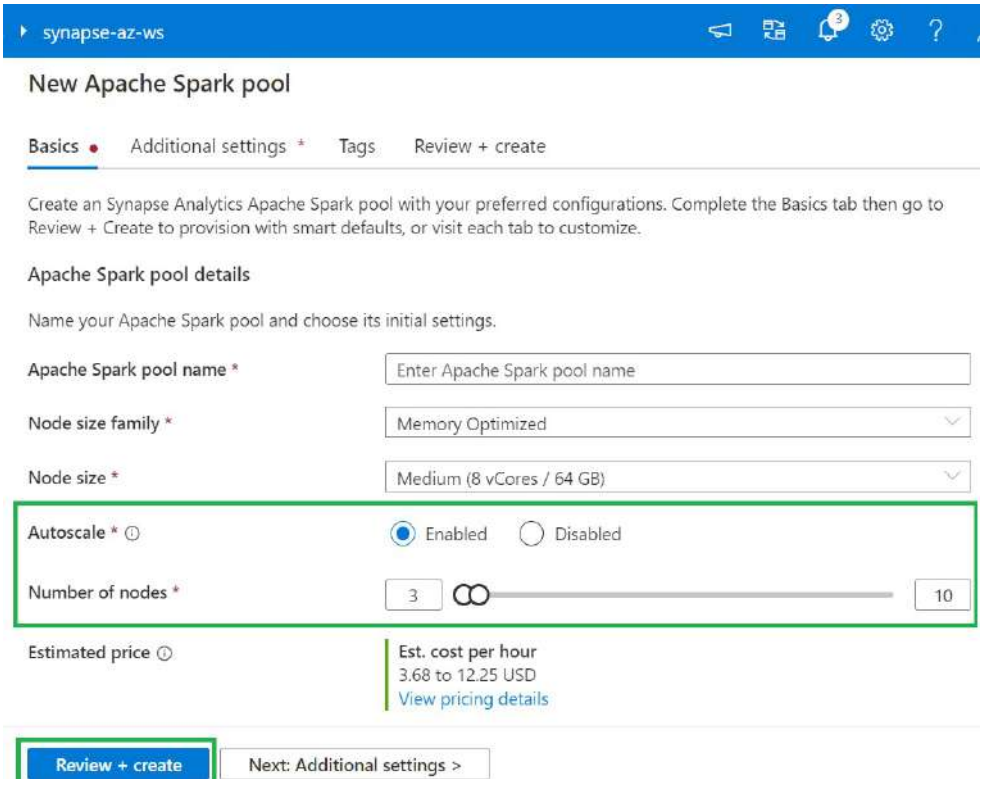

Figure 5.23 – Selecting the Synapse Spark Autoscale option for autoscaling

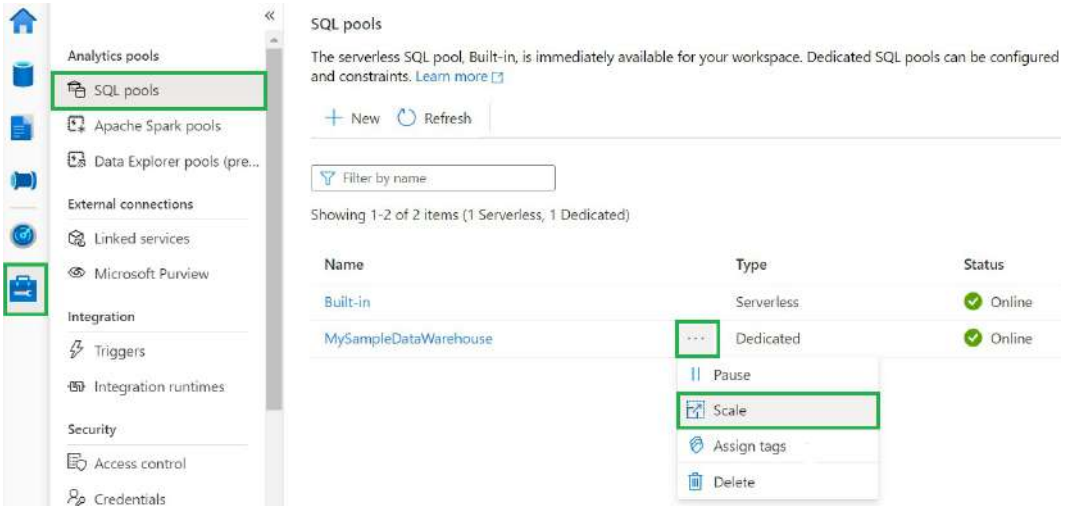

Figure 5.24 – Selecting the Synapse SQL Scale option

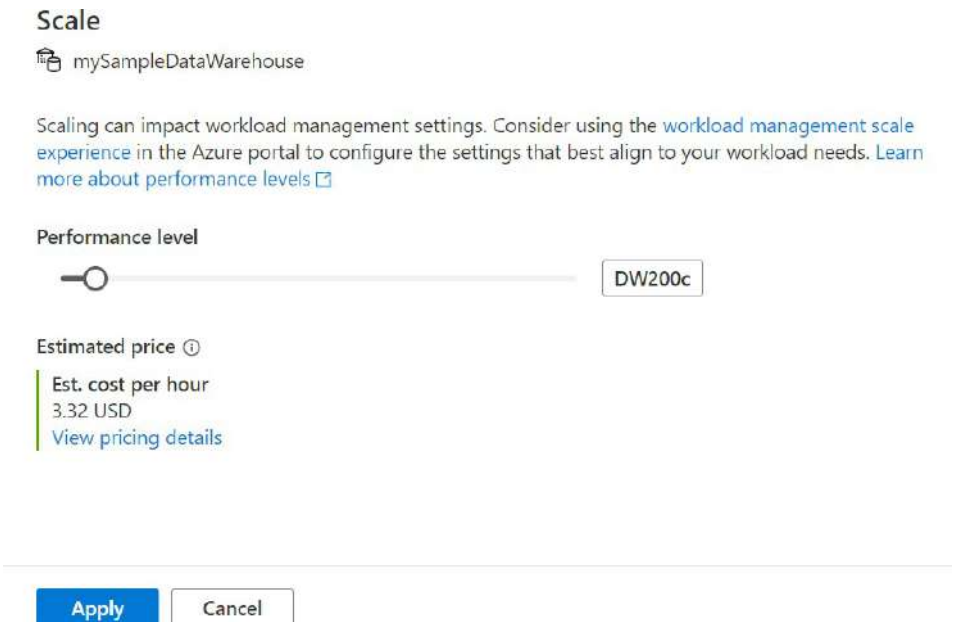

Figure 5.25 – Synapse SQL Scale option to scale out for performance and scale in to save costs

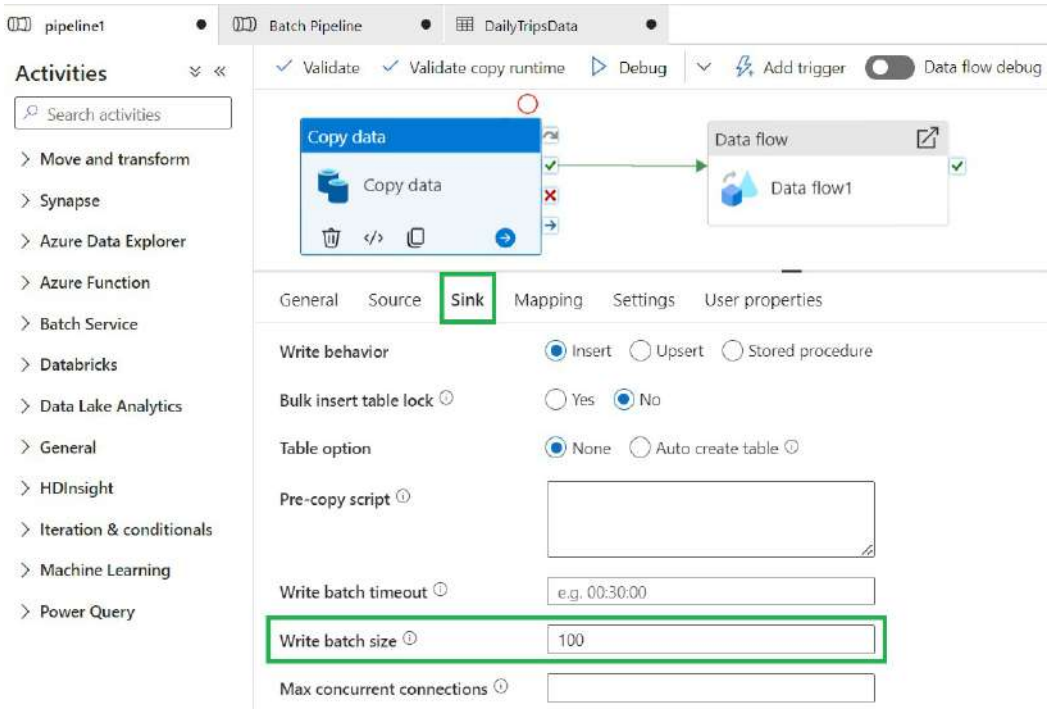

Figure 5.26 – Specifying the batch size for the Copy activity

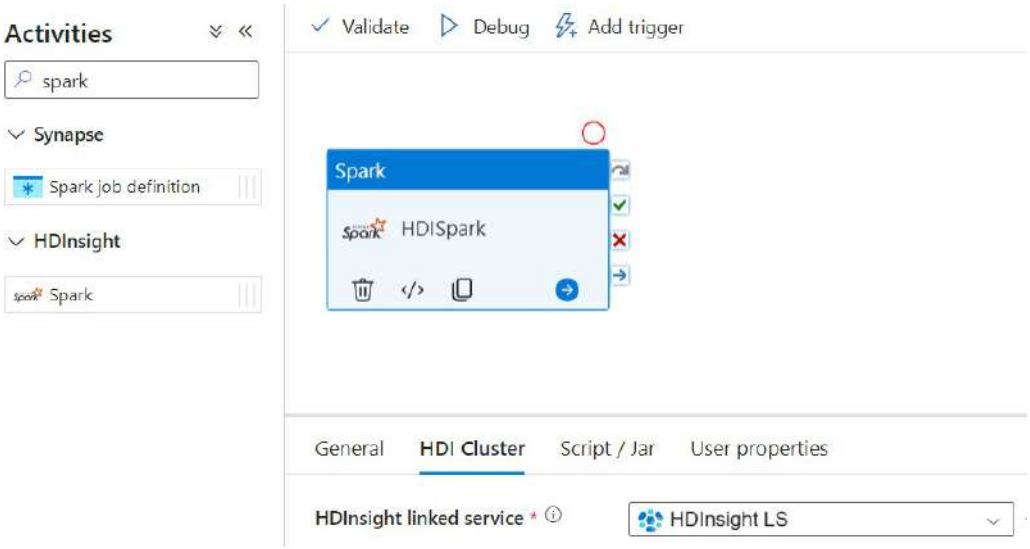

Figure 5.27 – Configuring a Spark activity in ADF

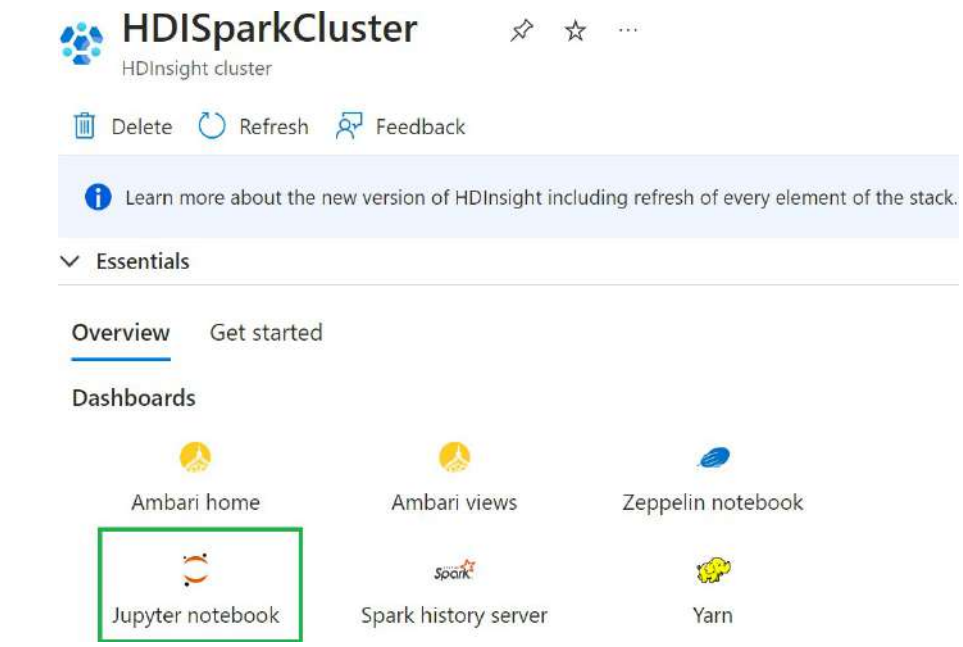

Figure 5.28 – Launching the Jupyter notebook from HDInsight

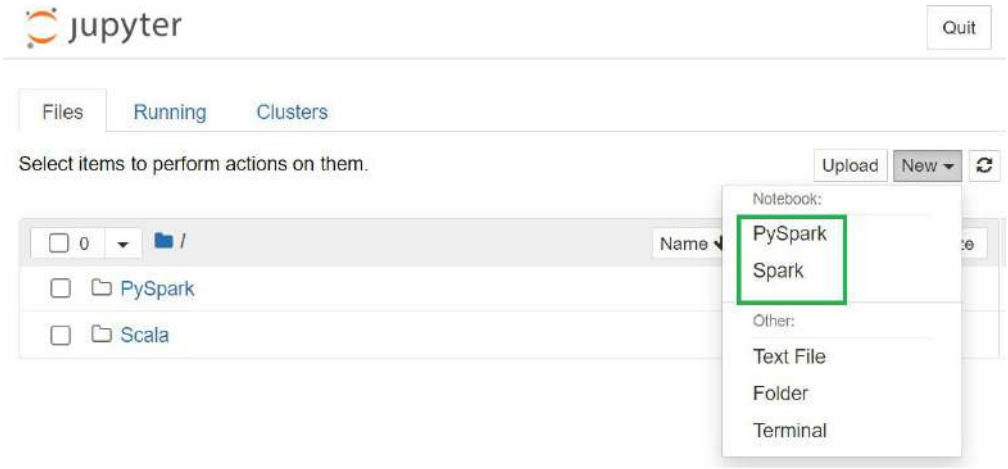

Figure 5.29 – Launching a PySpark Jupyter notebook from HDInsight

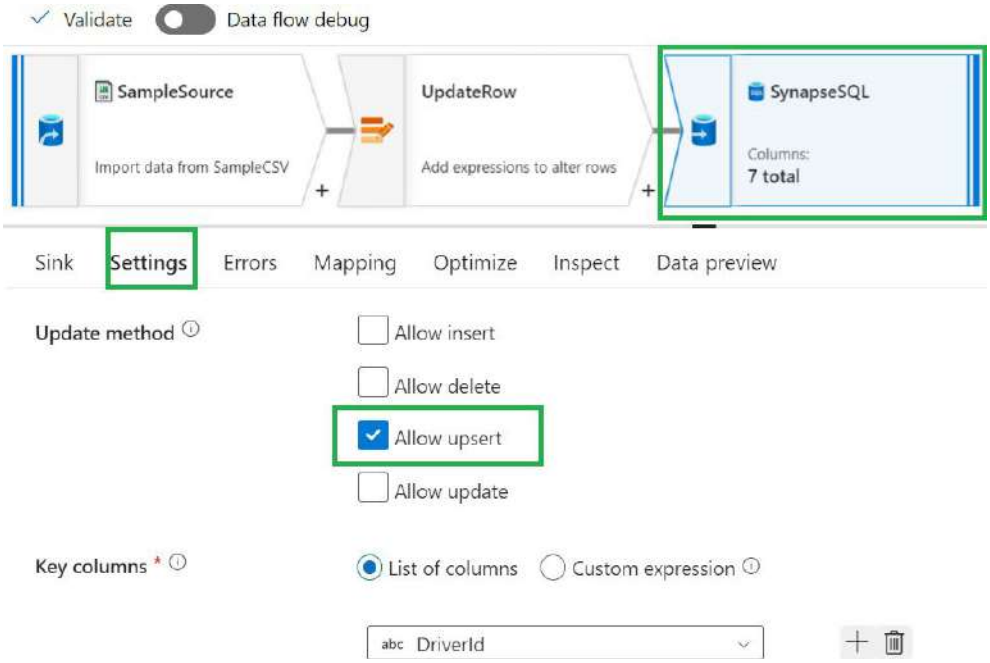

Figure 5.30 – Upsert operation in ADF used for updating and inserting data

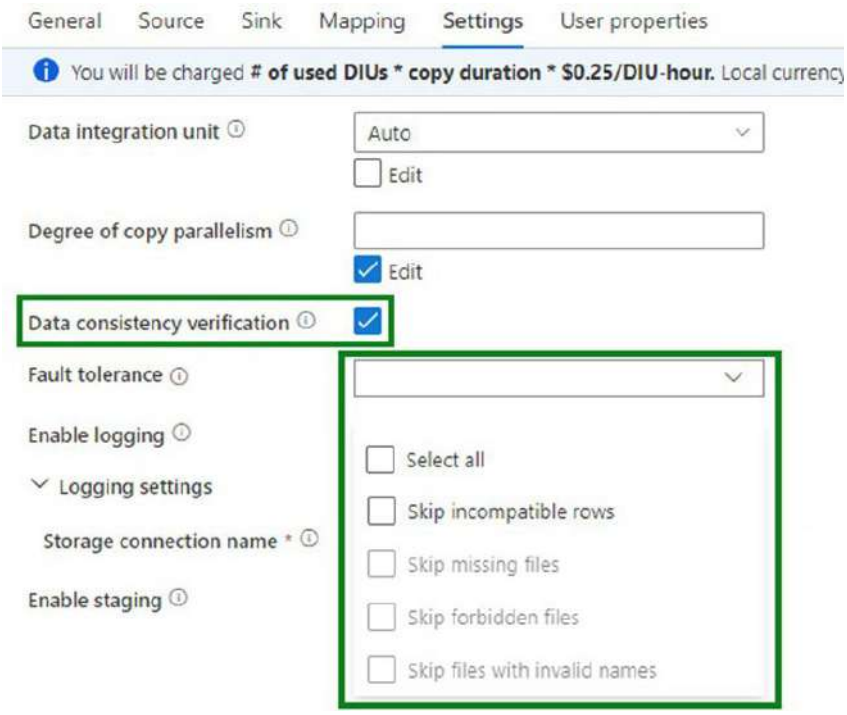

Figure 5.31 – Enabling consistency verification and fault tolerance in an ADF Copy activity

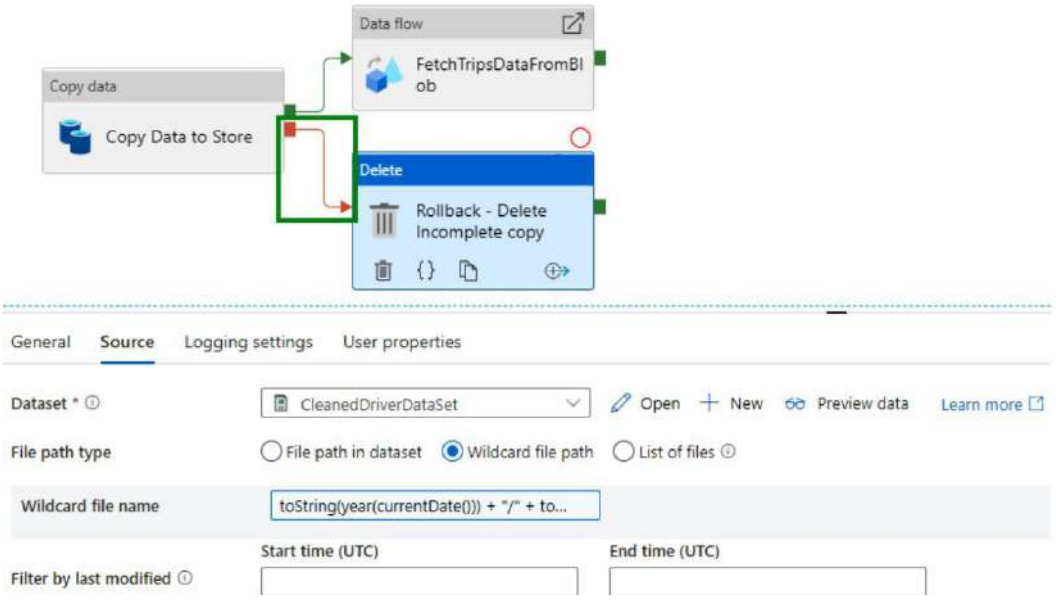

Figure 5.32 – Deleting an incomplete Copy activity

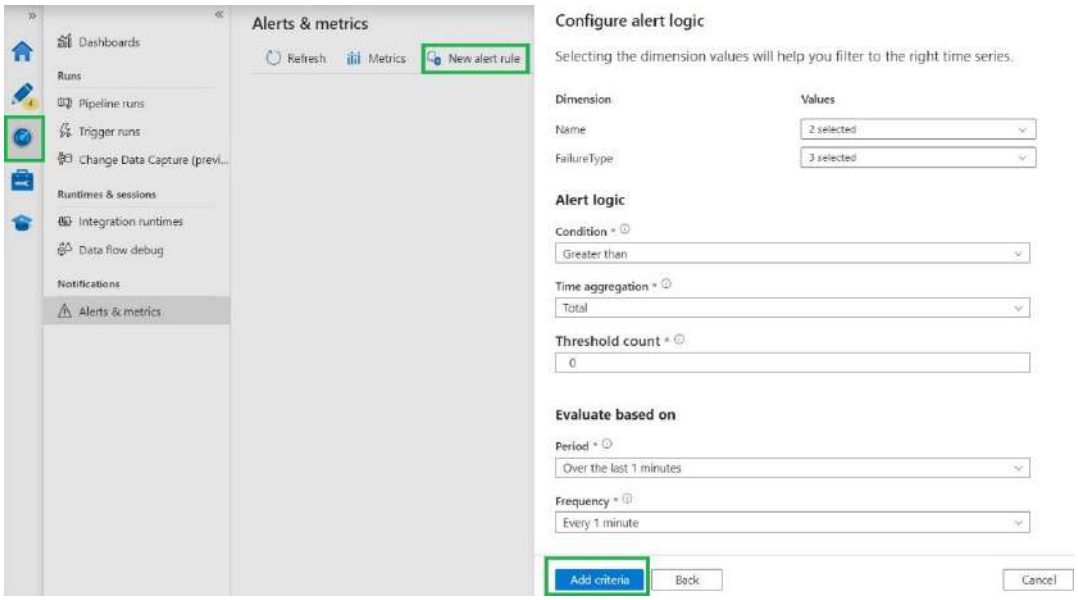

Figure 5.33 – Configuring an alert for pipeline failure

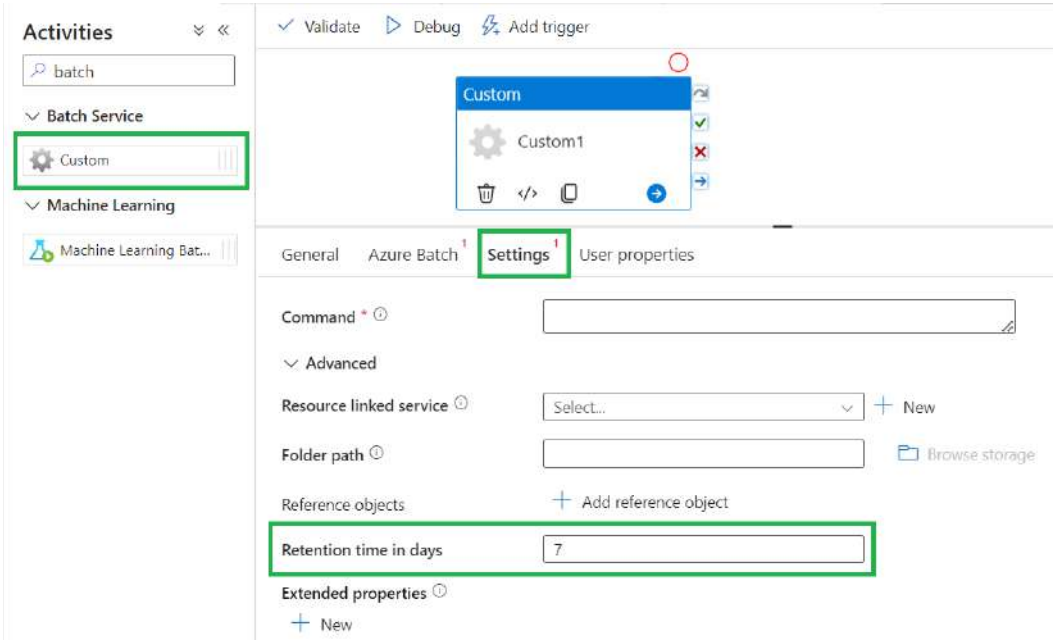

Figure 5.34 – Configuring batch retention for an Azure Batch activity in a pipeline

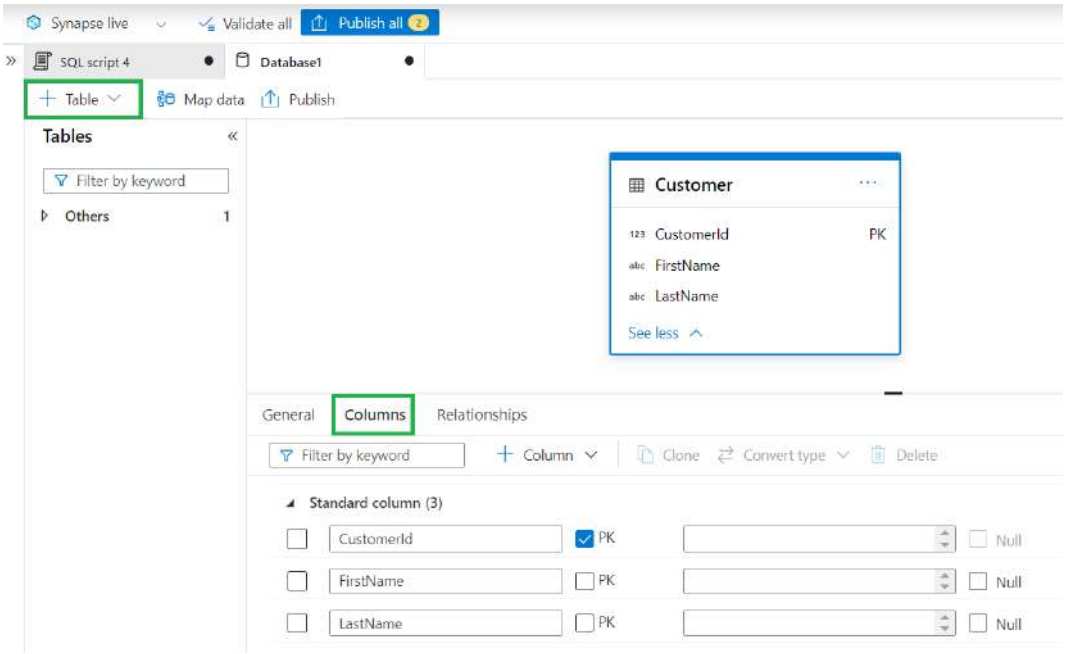

Figure 5.35 – Creating a custom delta table in Synapse

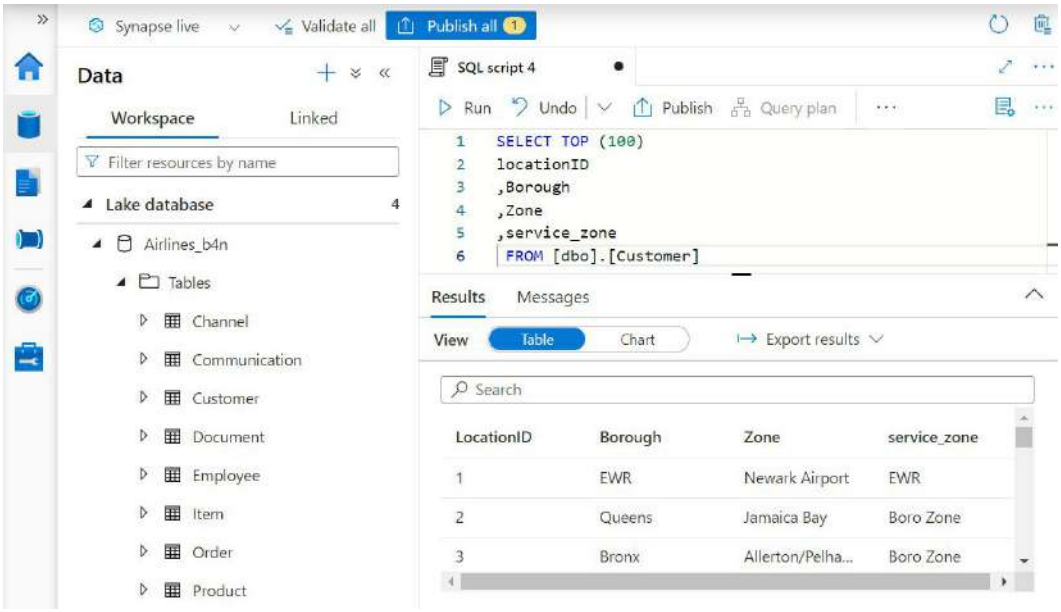

Figure 5.36 – Querying a delta table in Synapse

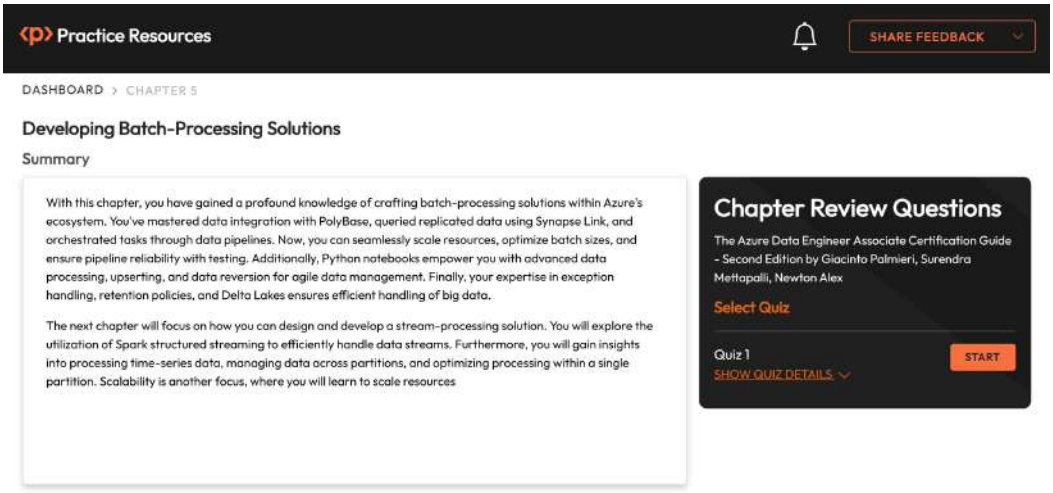

Figure 5.38 – Chapter Review Questions for Chapter 5

Chapter 6: Developing a Stream Processing Solution

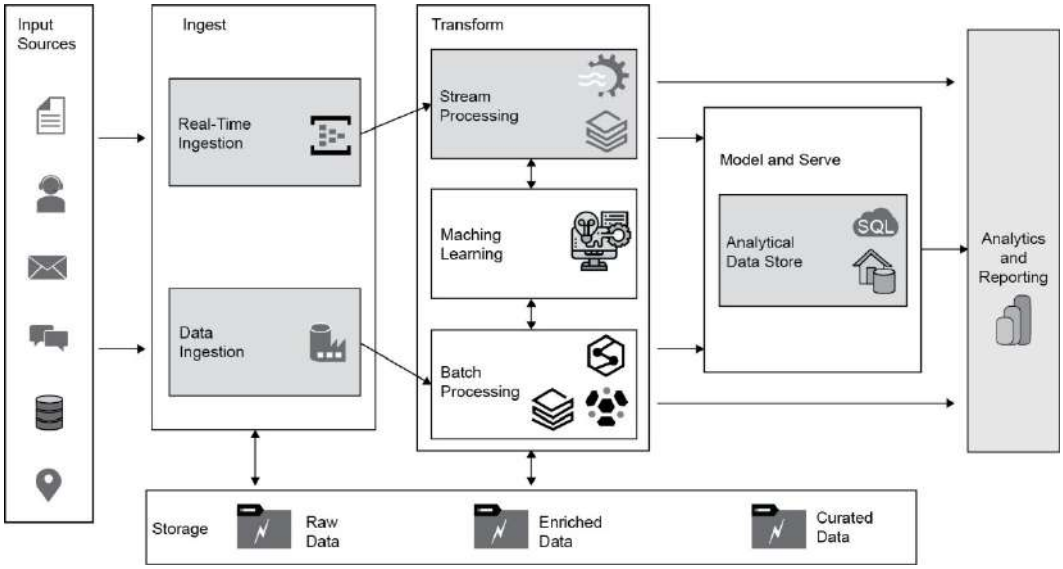

Figure 6.1 – Stream processing architecture with several key stages

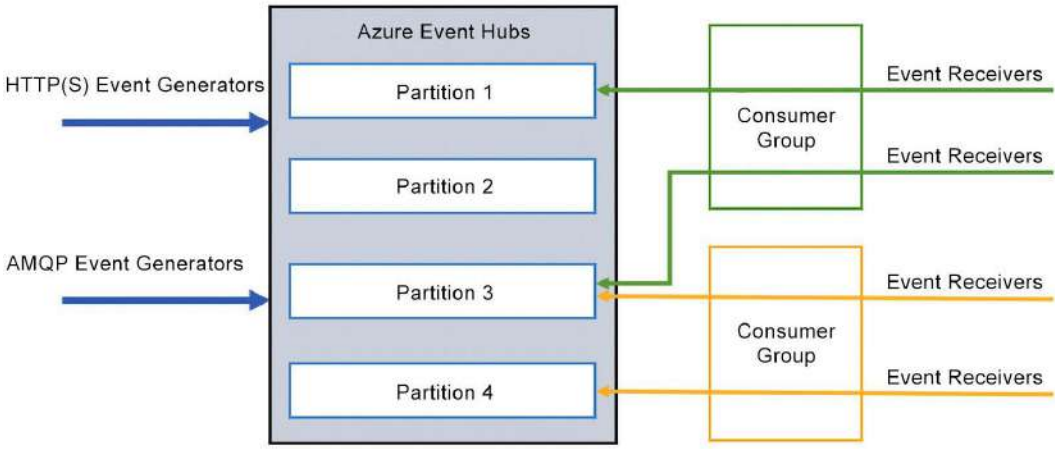

Figure 6.2 – Azure Event Hubs architecture

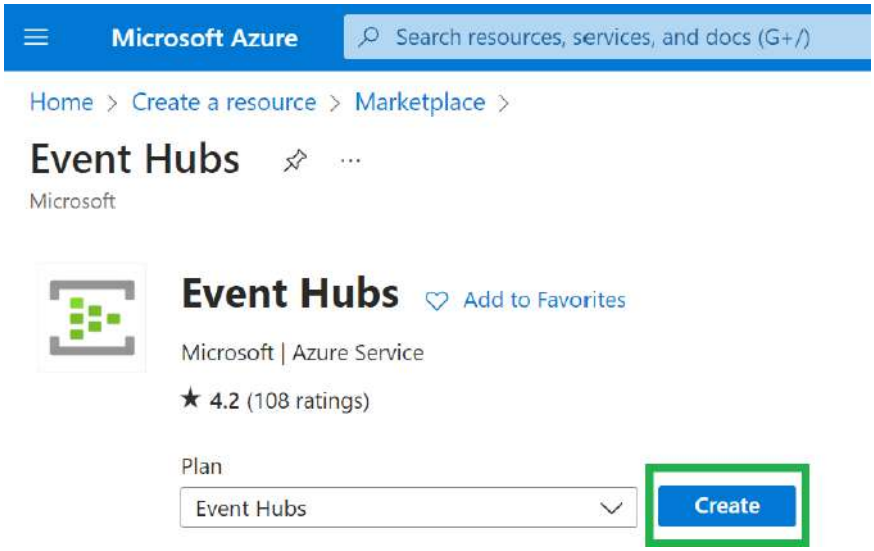

Figure 6.3 – The Event Hubs creation screen

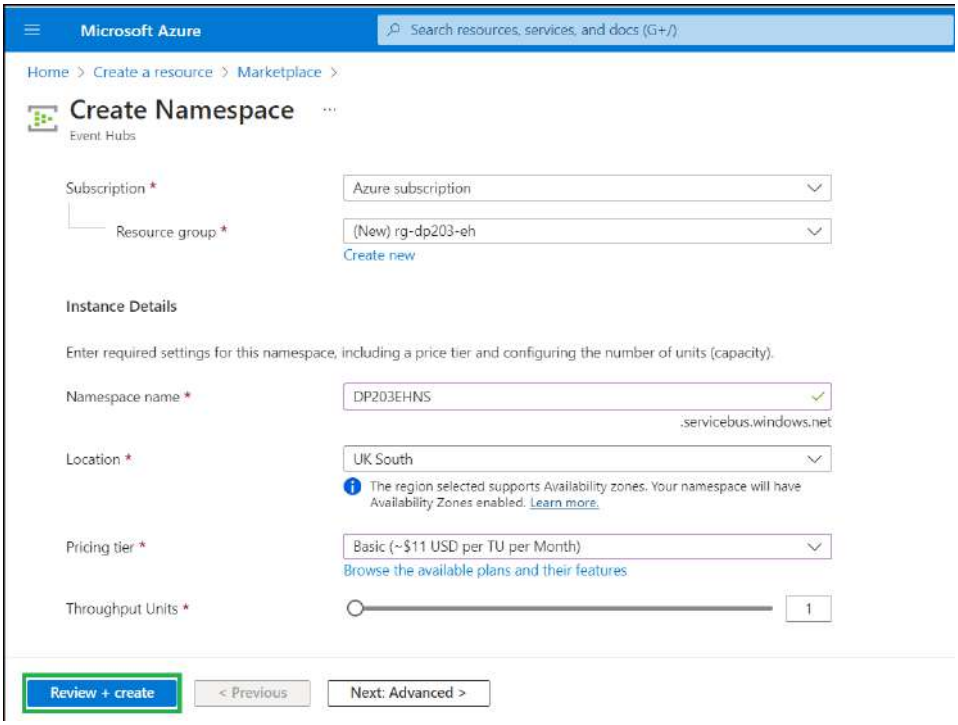

Figure 6.4 – Creating the Event Hubs namespace

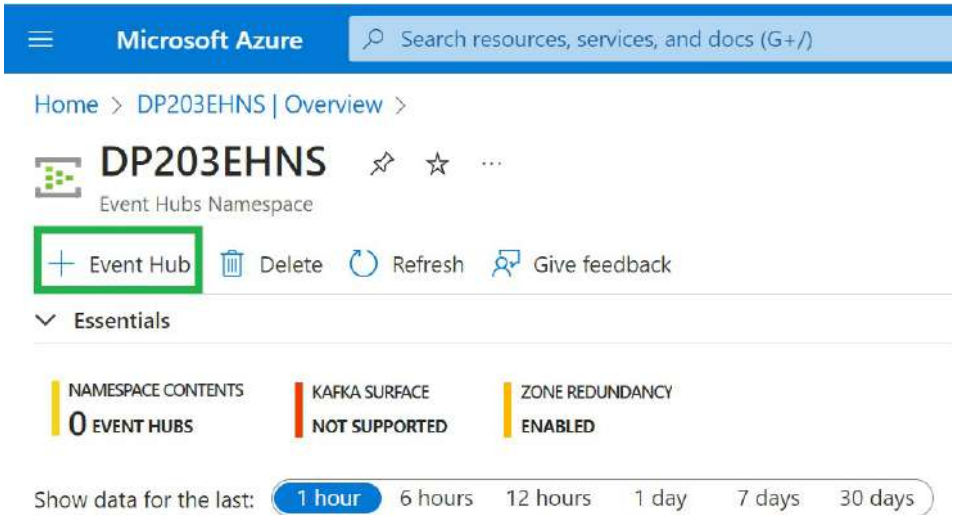

Figure 6.5 – Creating an Event Hub from within the Event Hubs workspace

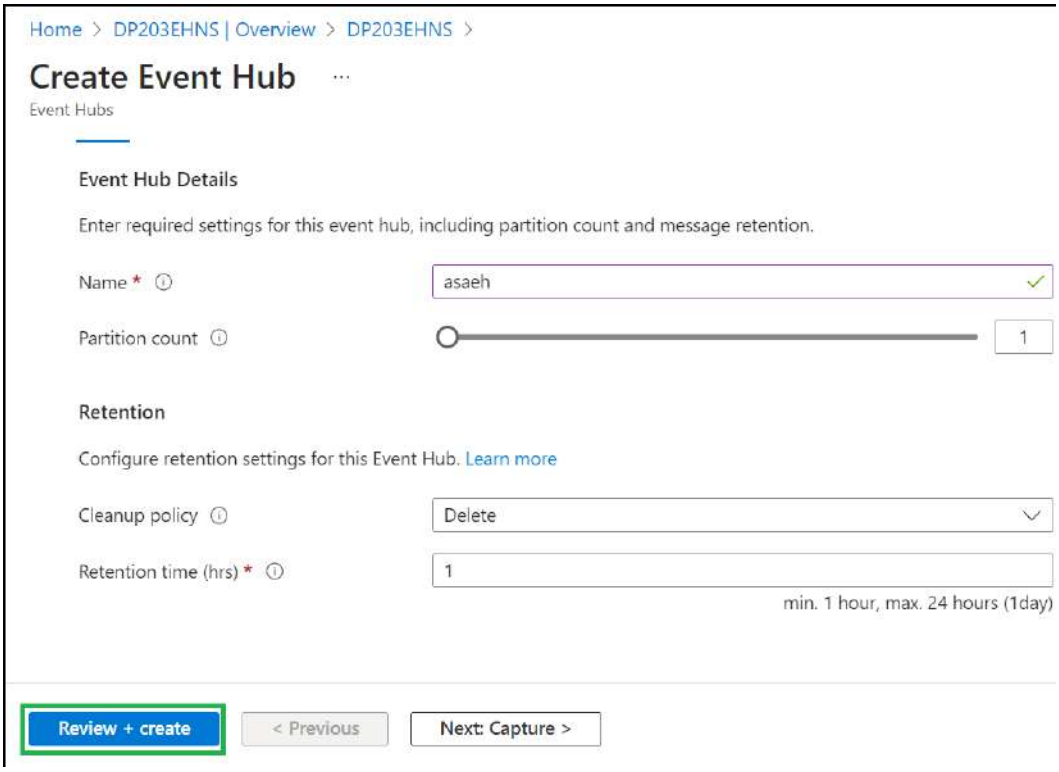

Figure 6.6 – The Create Event Hub screen

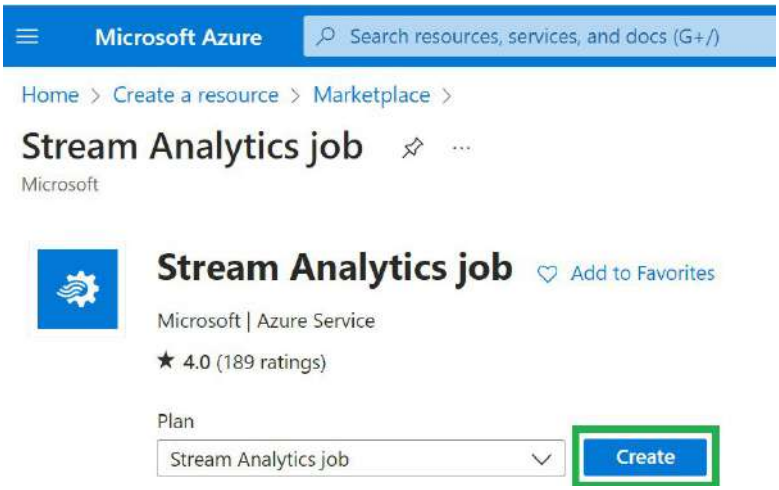

Figure 6.7 – Creating an ASA job

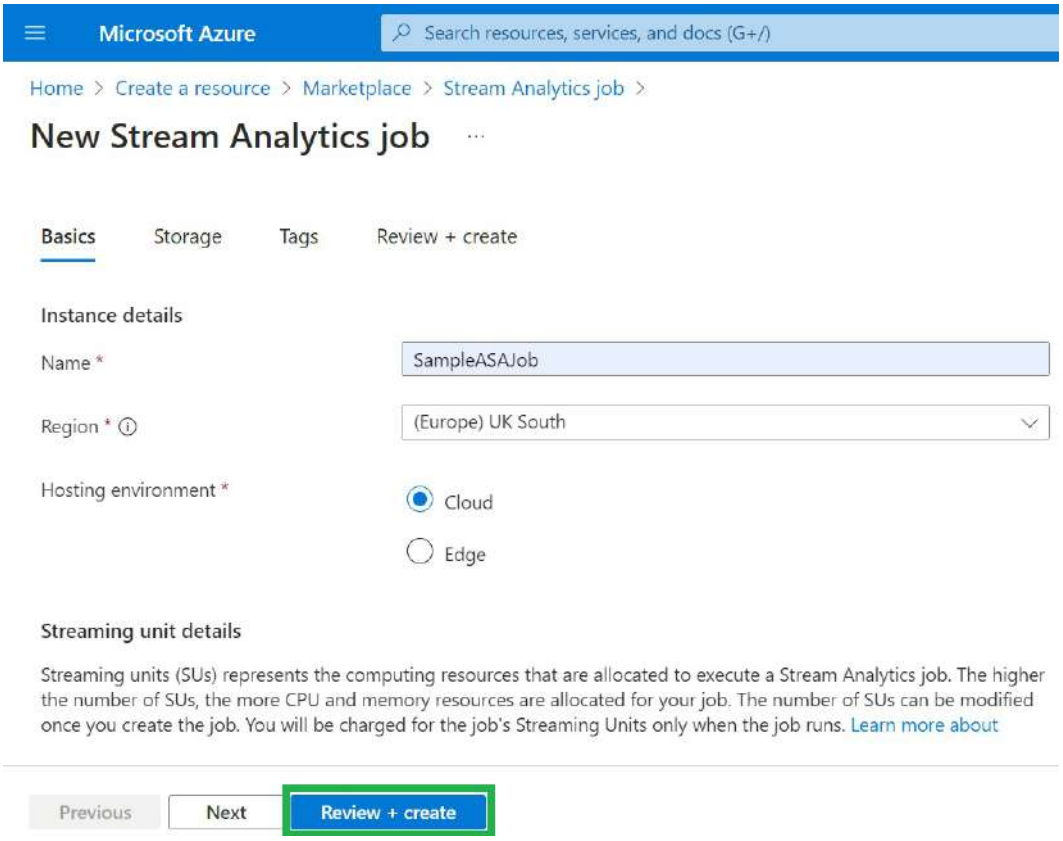

Figure 6.8 – The ASA job creation screen

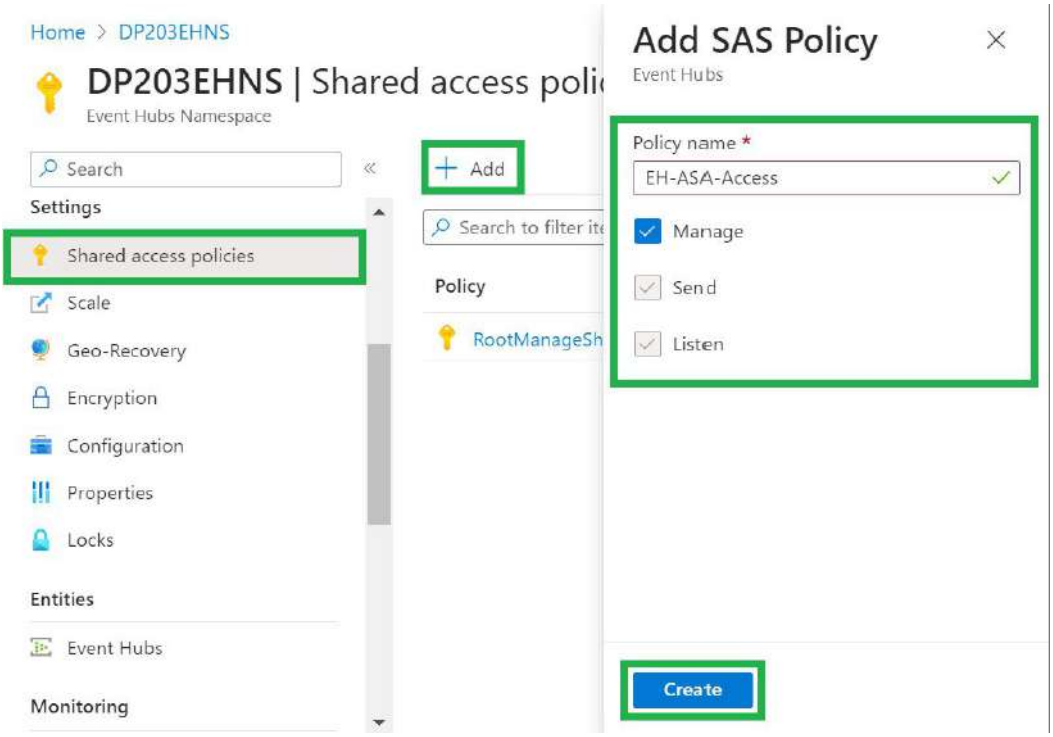

Figure 6.9 – Creating shared access policies in Event Hubs

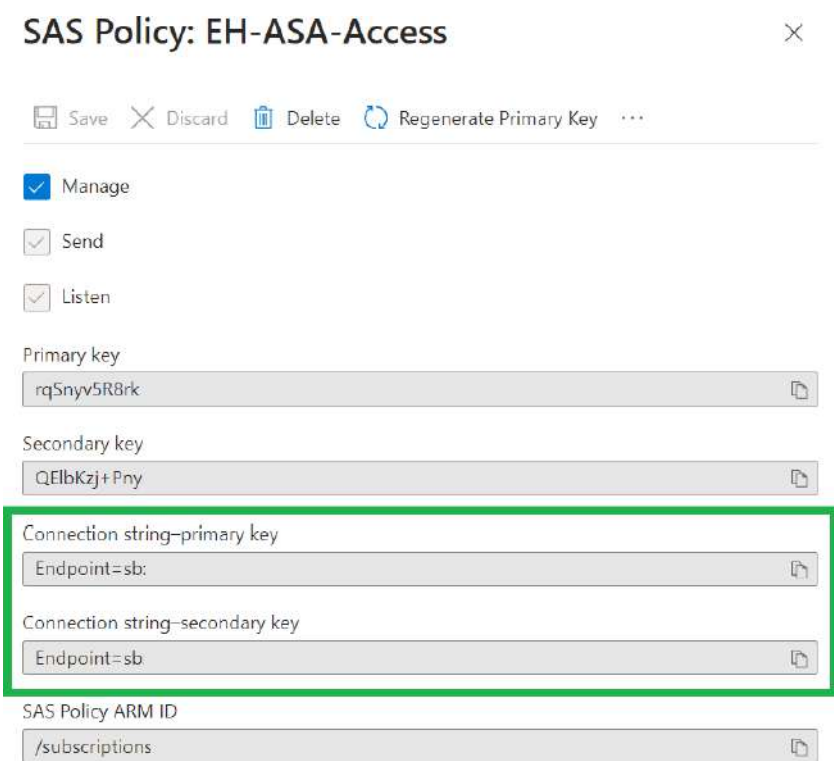

Figure 6.10 – Accessing the connection string for Event Hubs

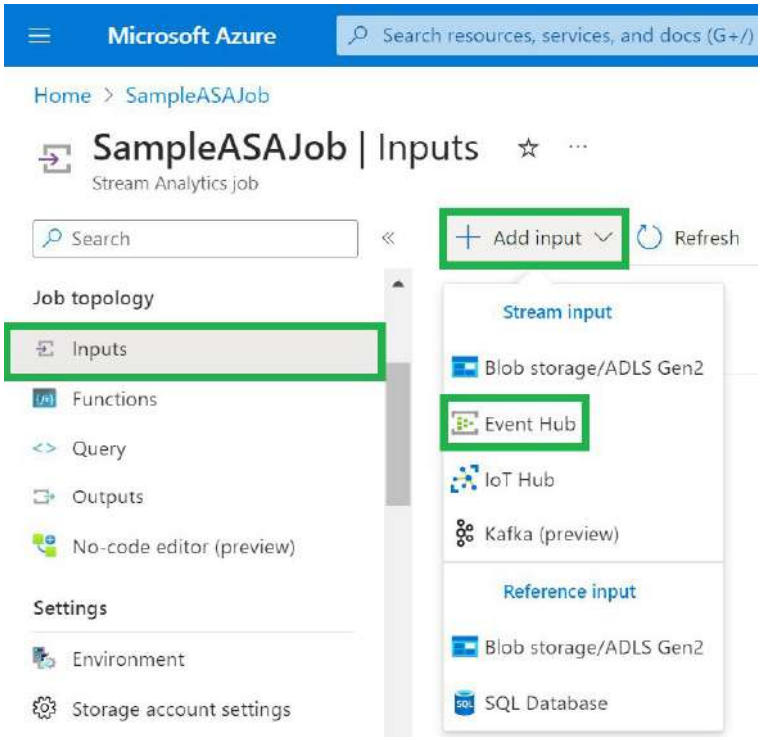

Figure 6.11 – Selecting an Event Hub as input for the ASA job

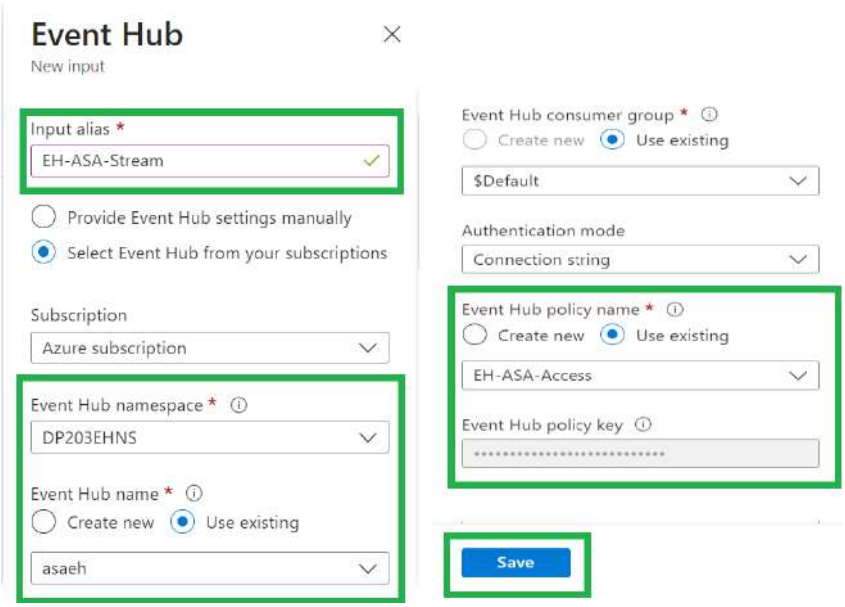

Figure 6.12 – Linking the Event Hub as an input in ASA

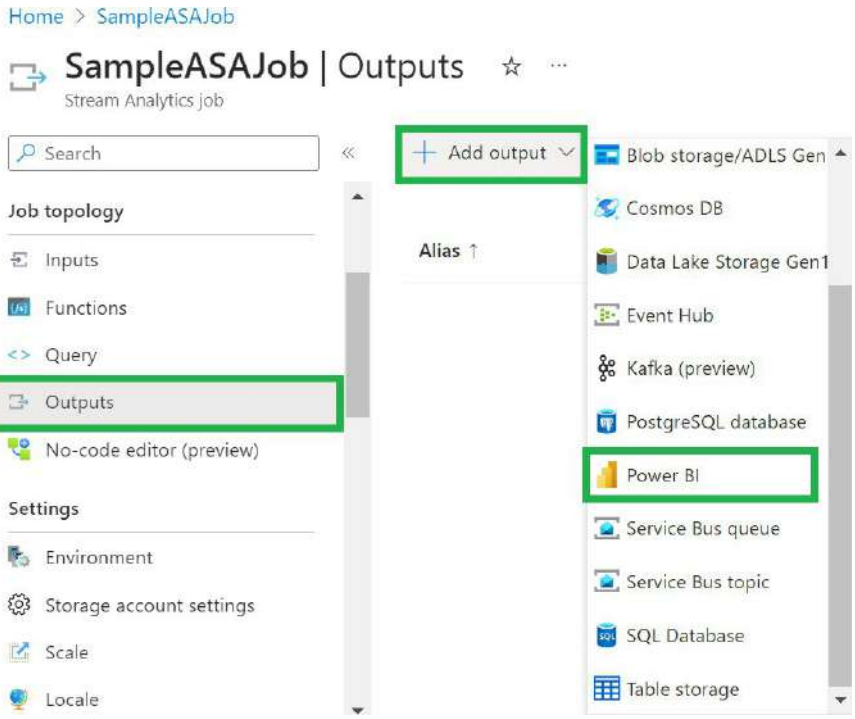

Figure 6.13 – Selecting Power BI as the output for the ASA job

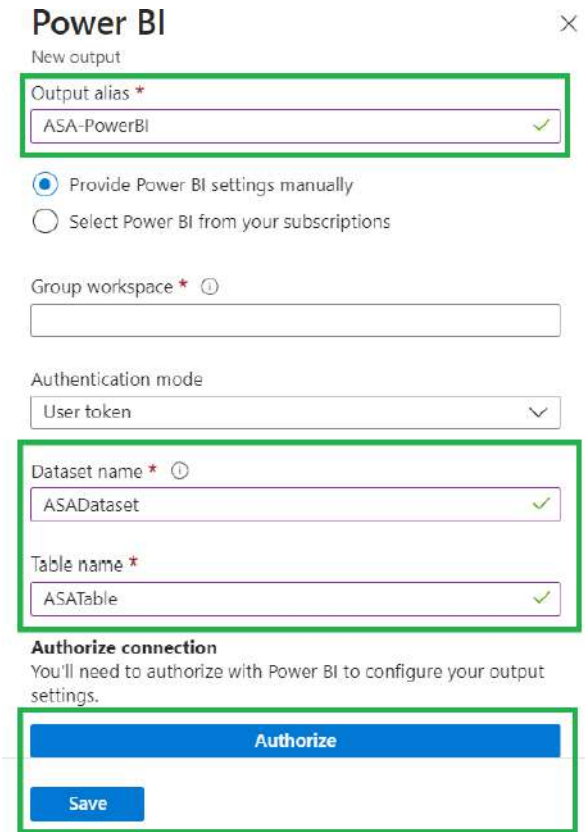

Figure 6.14 – Configuring the Power BI sink details for the ASA job

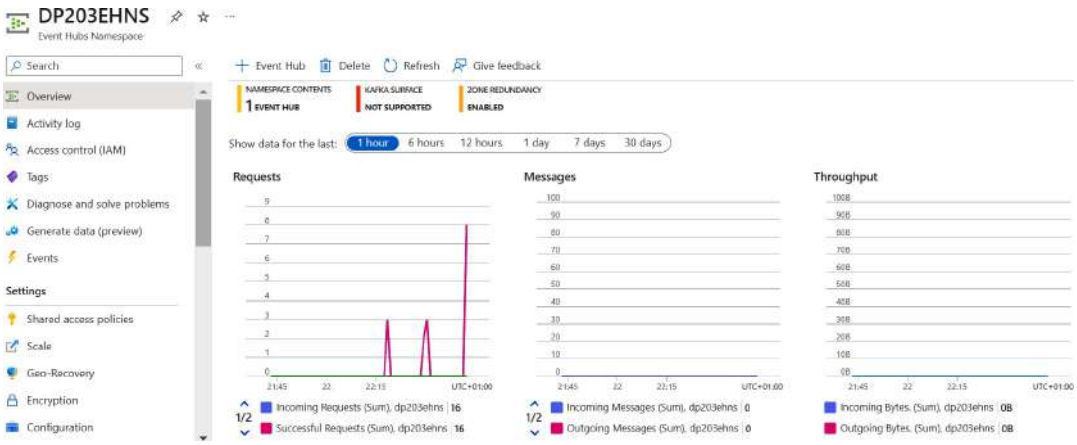

Figure 6.15 – The Event Hubs overview screen showing the event metrics

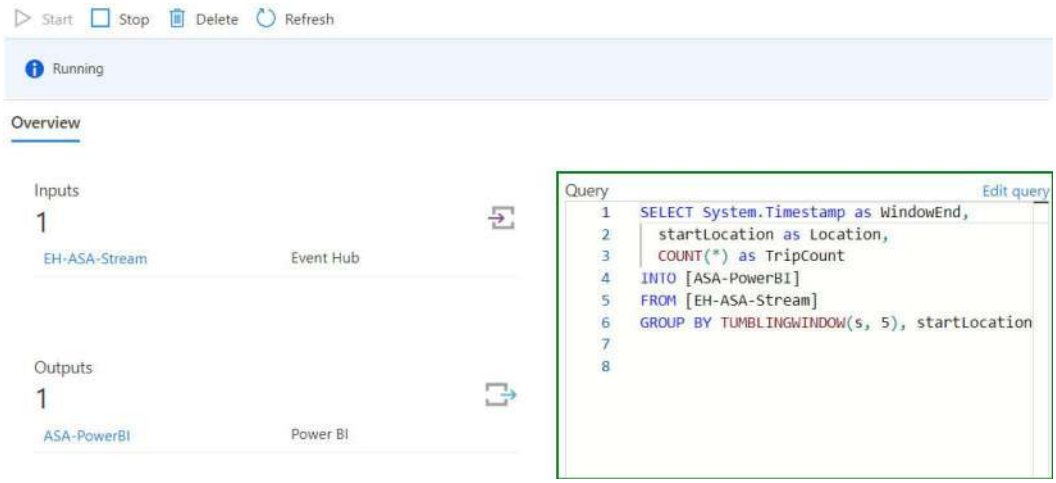

Figure 6.16 – ASA sample code reading input event from Event Hub and outputs to Power BI

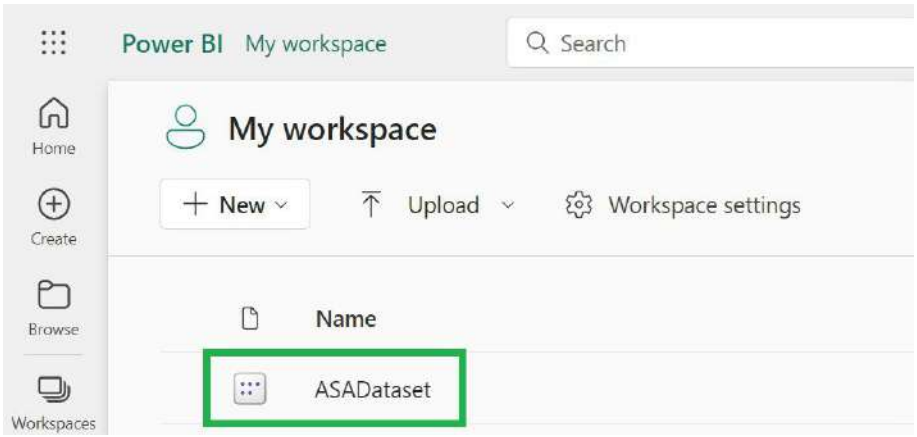

Figure 6.17 – My workspace showing the configured ASA dataset in the Power BI dataset

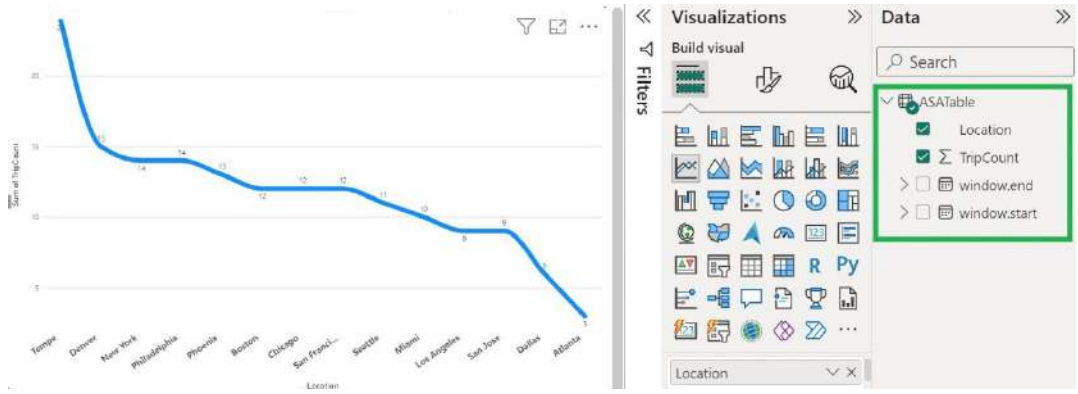

Figure 6.18 – Creating the Power BI dashboard from the ASA streaming output

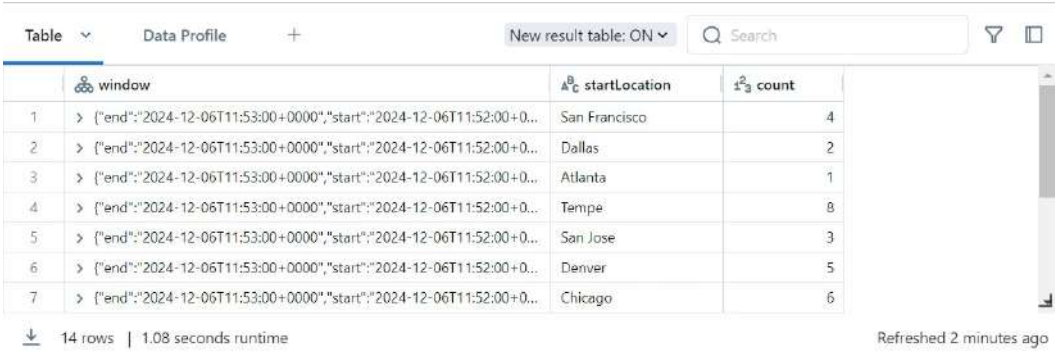

Figure 6.19 – Viewing the results of the streaming query

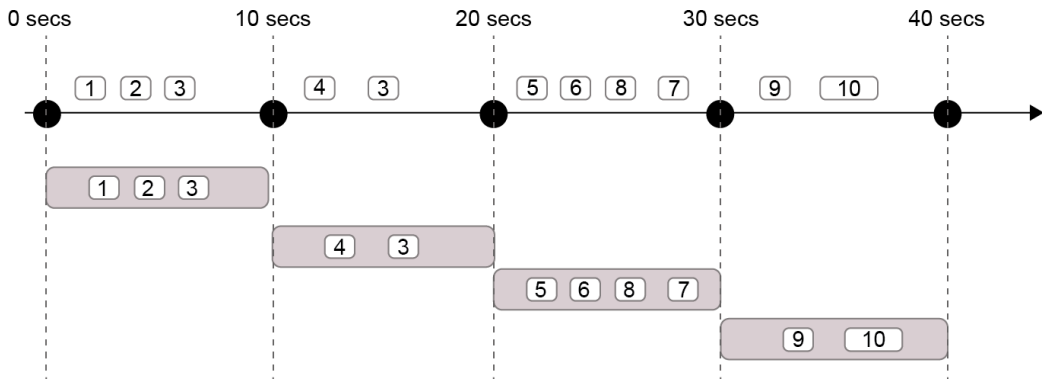

Figure 6.20 – A Tumbling Window showing a stream segmented into 10-second time windows

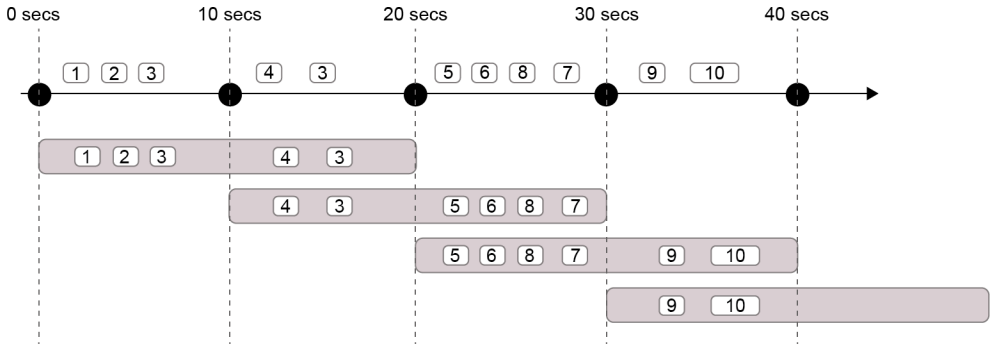

Figure 6.21 – A hopping window showing a fixed-size overlap with the previous window

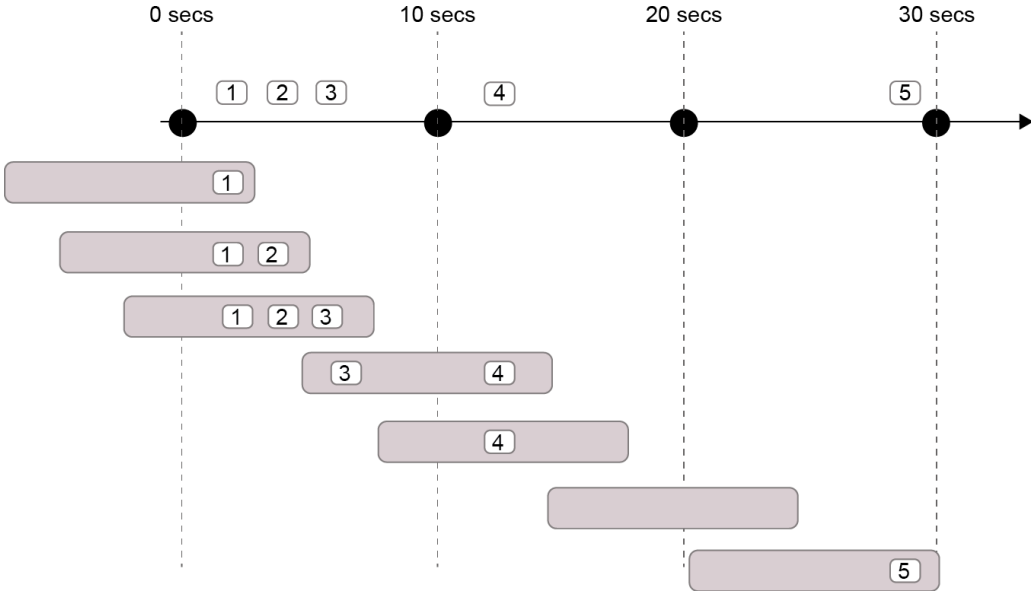

Figure 6.22 – A sliding window reporting when an event is added or removed

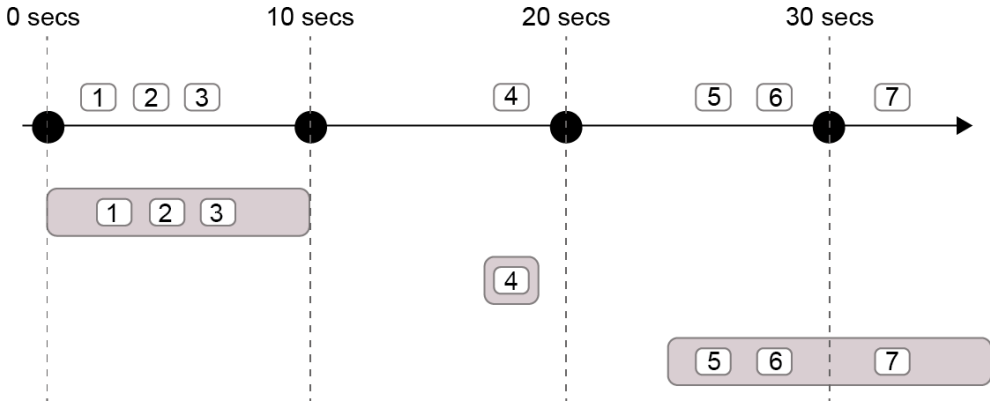

Figure 6.23 – An example of a session window with no fixed sizes

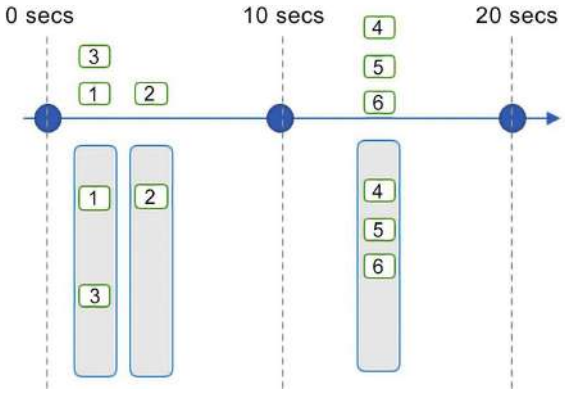

Figure 6.24 – A snapshot window showing an overview of an event at a particular time

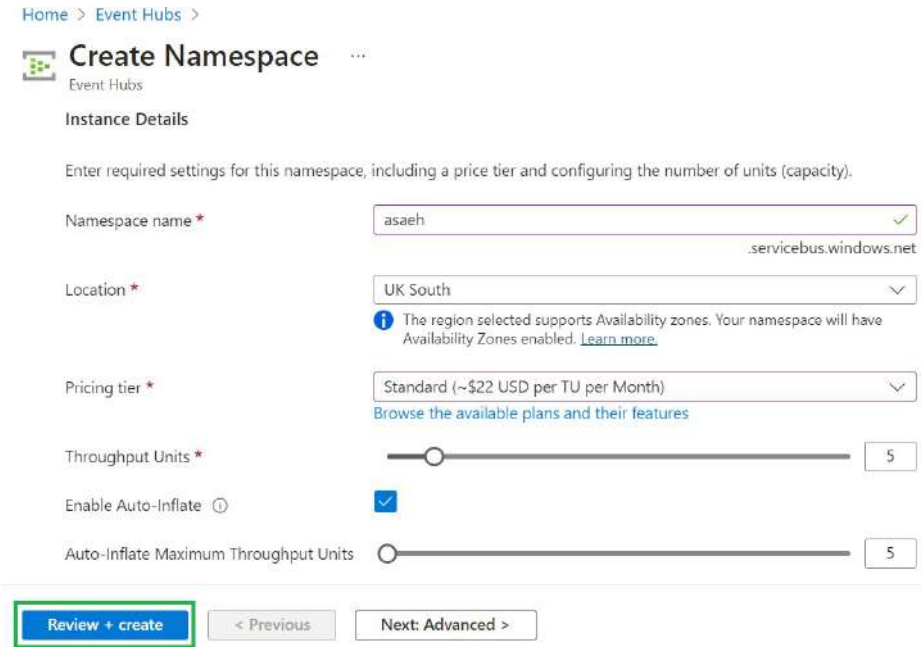

Figure 6.25 – Enabling auto-inflate feature in Event Hubs

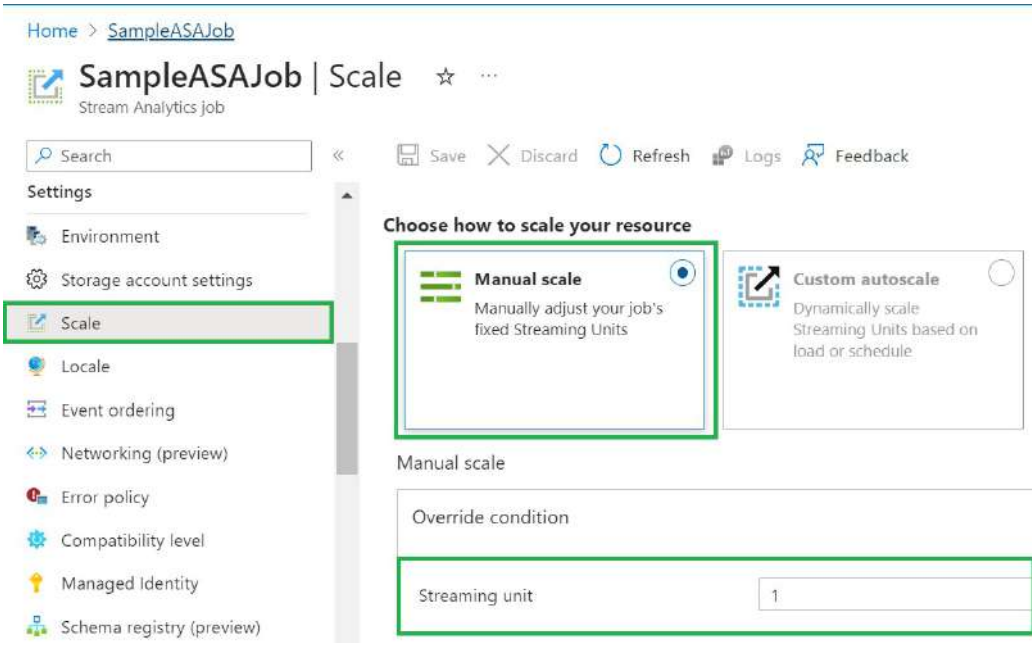

Figure 6.26 – Scaling an ASA job

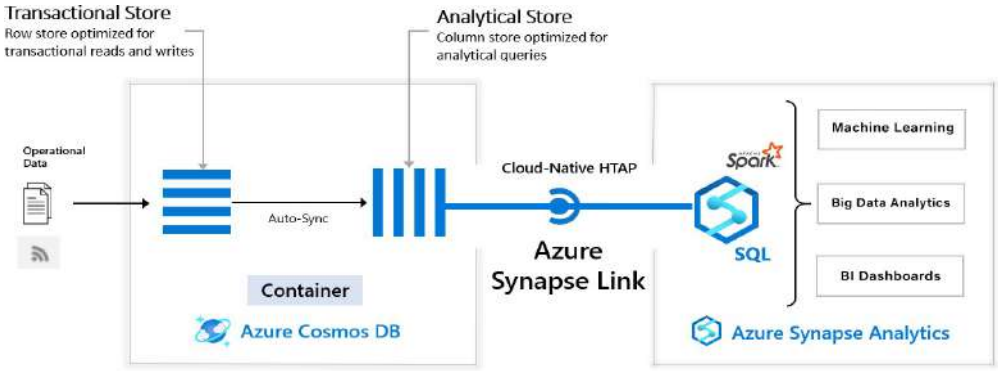

Figure 6.27 – Azure Synapse Link for Cosmos DB architecture

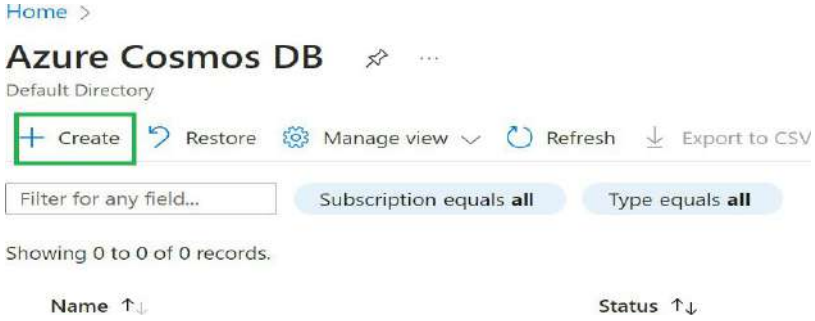

Figure 6.28 – Creating a new Cosmos DB instance

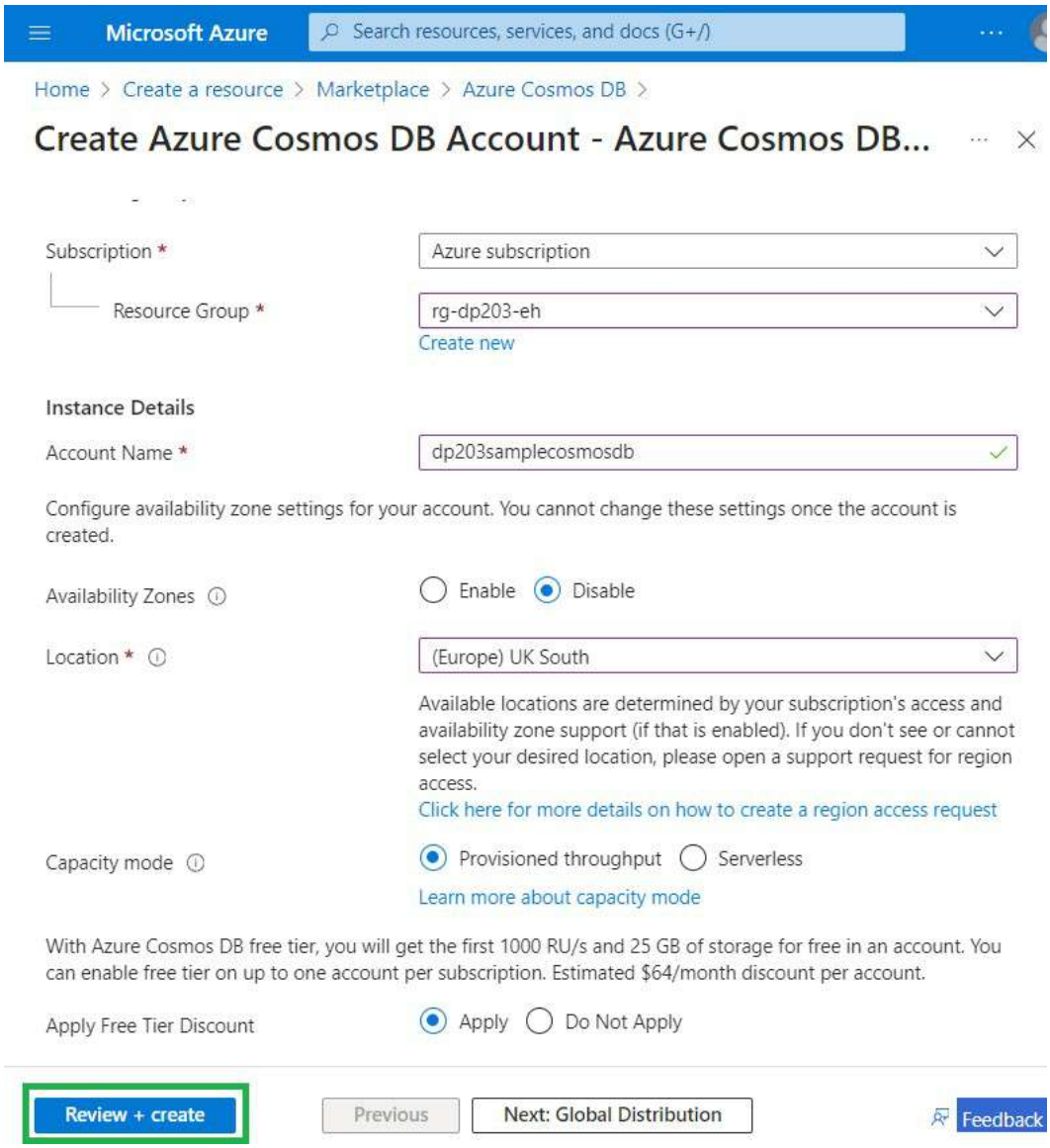

Figure 6.29 – Cosmos DB creation screen for building a NoSQL account

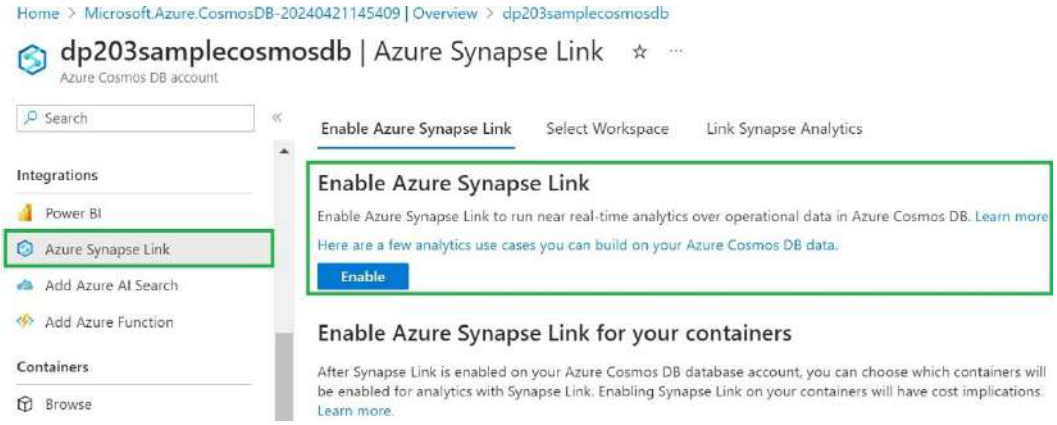

Figure 6.30 – Enabling Synapse Link in Cosmos DB

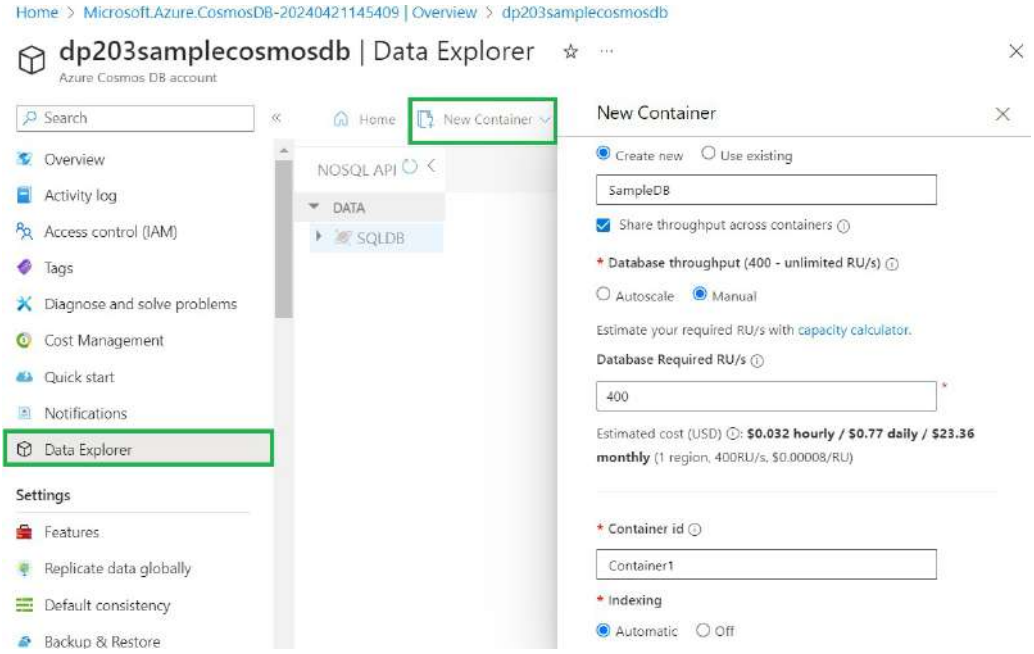

Figure 6.31 – Configuring a new container by selecting Analytical store

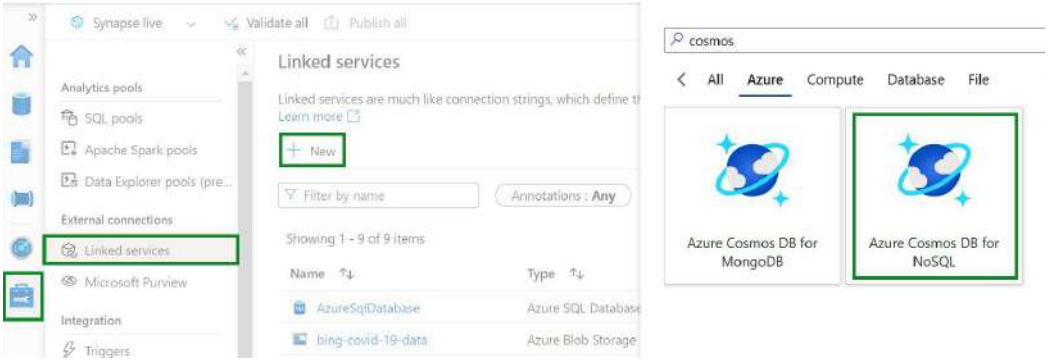

Figure 6.32 – Creating a linked service to Cosmos DB

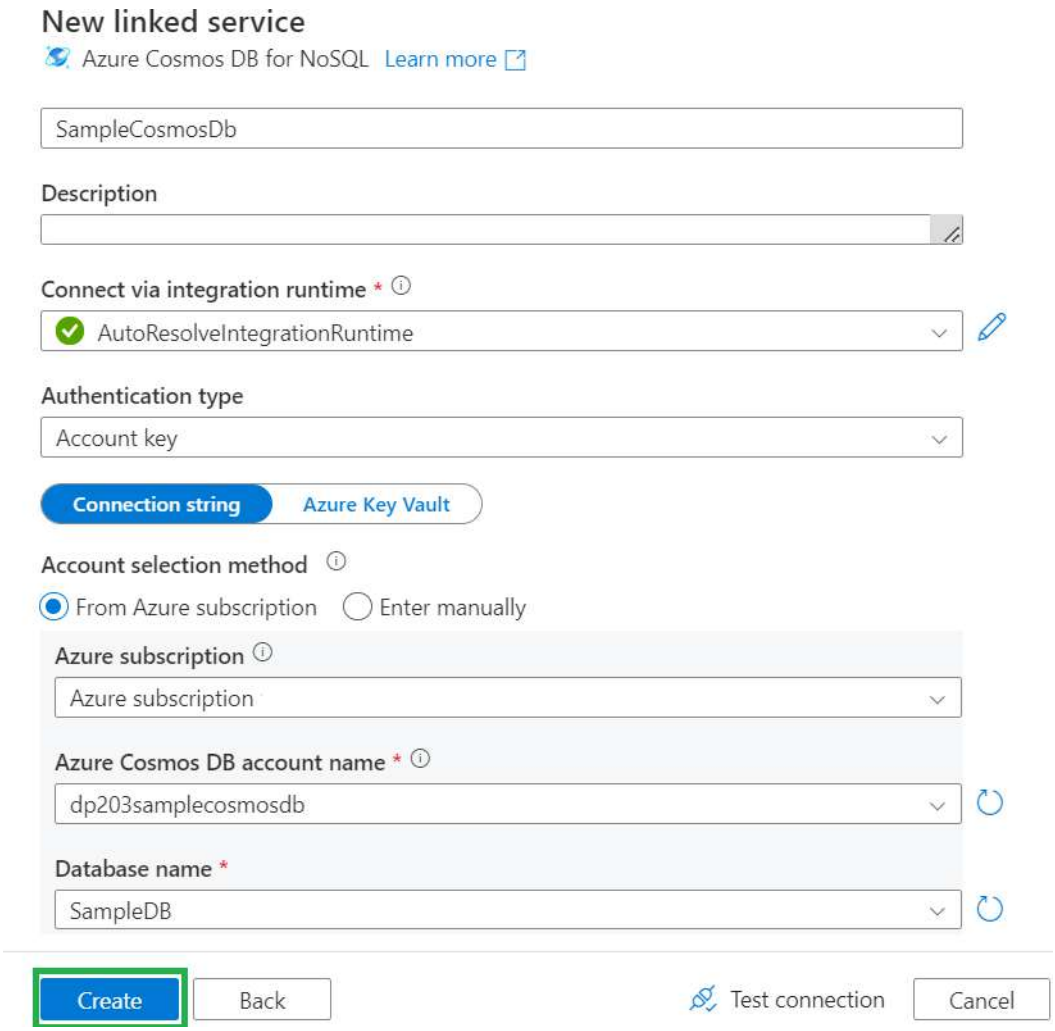

Figure 6.33 – Cosmos DB linked service showing Cosmos DB for Synapse Link

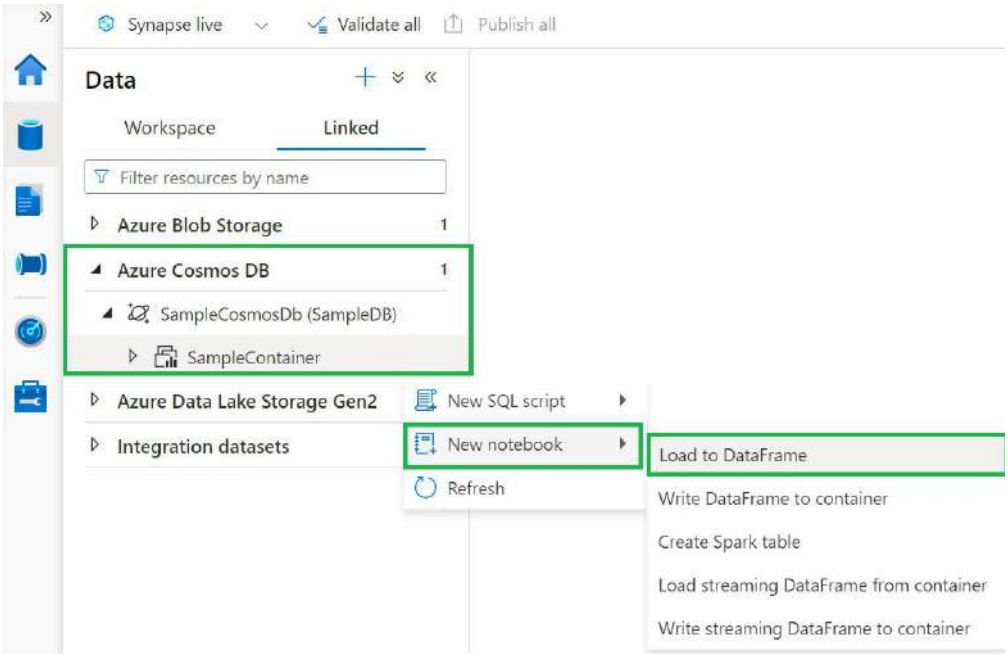

Figure 6.34 – Loading Cosmos DB data using the Synapse Link

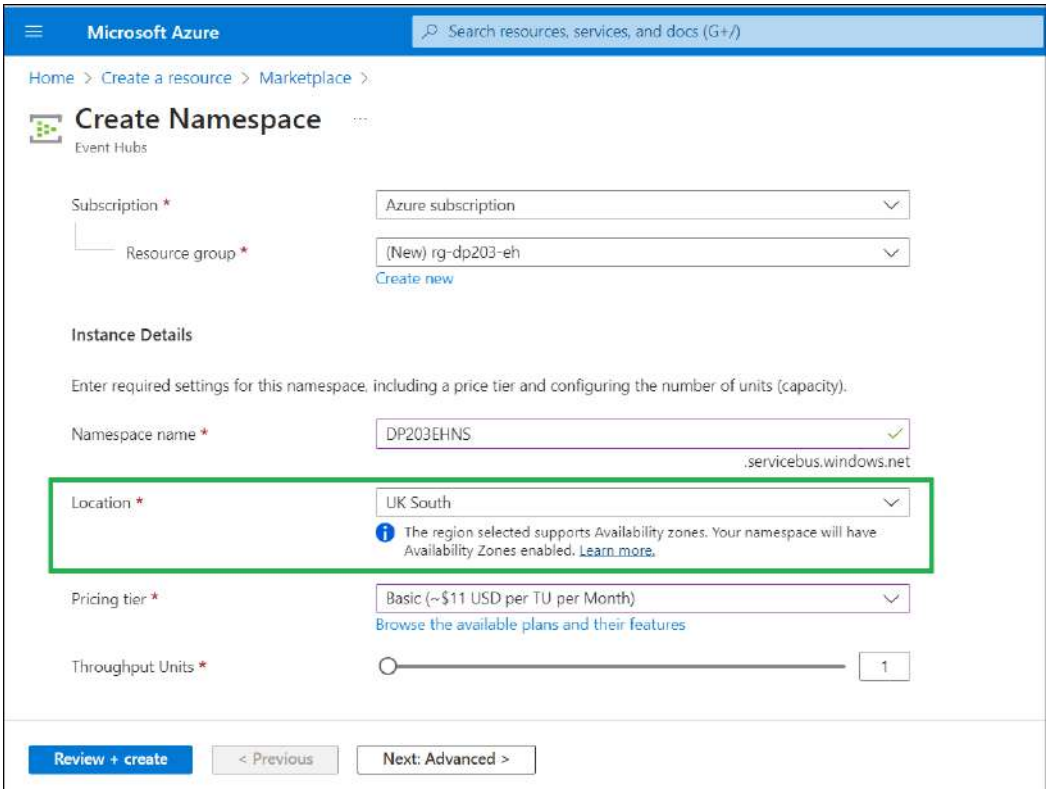

Figure 6.35 – Choosing locations with availability zones in Event Hubs

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

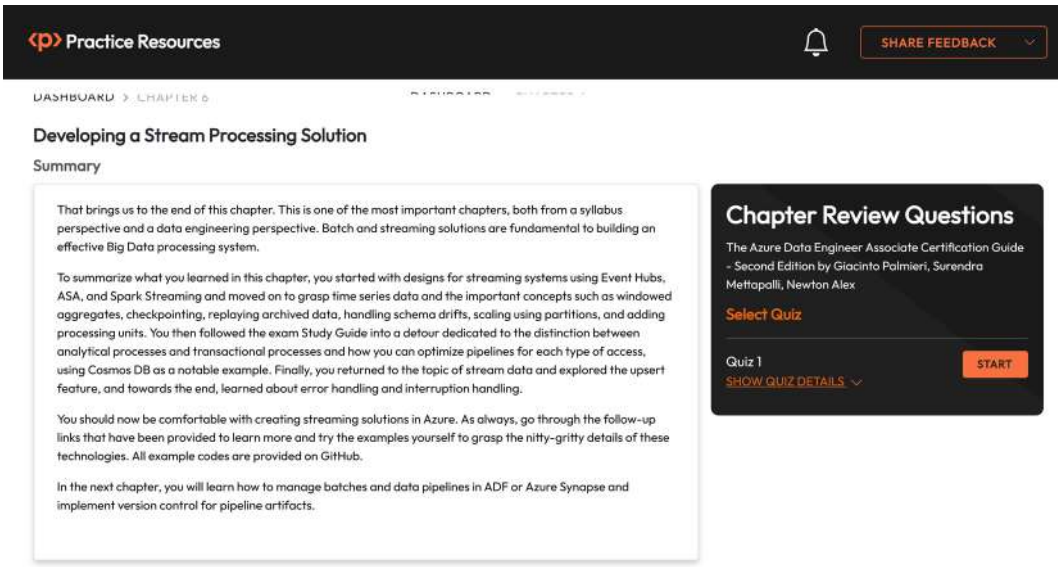

Figure 6.38 – Chapter Review Questions for Chapter 6

Chapter 7: Managing Batches and Pipelines

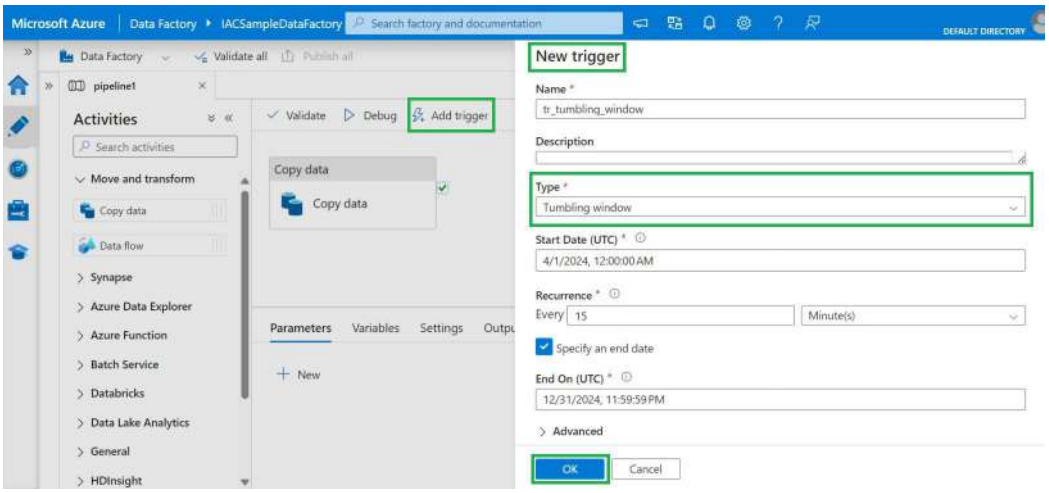

Figure 7.1 – Defining a Tumbling Window trigger for a pipeline

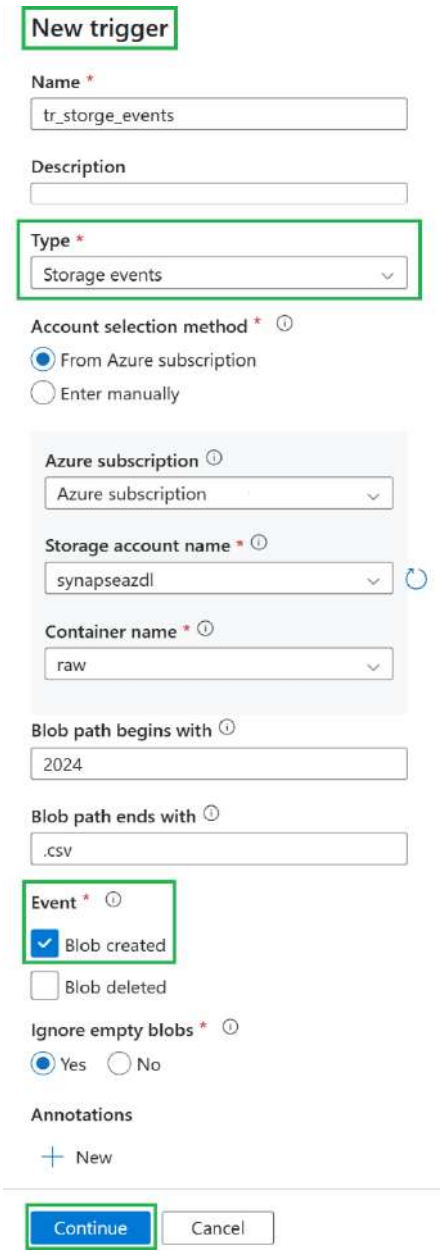

Figure 7.2 – Defining a Storage Event trigger for a pipeline

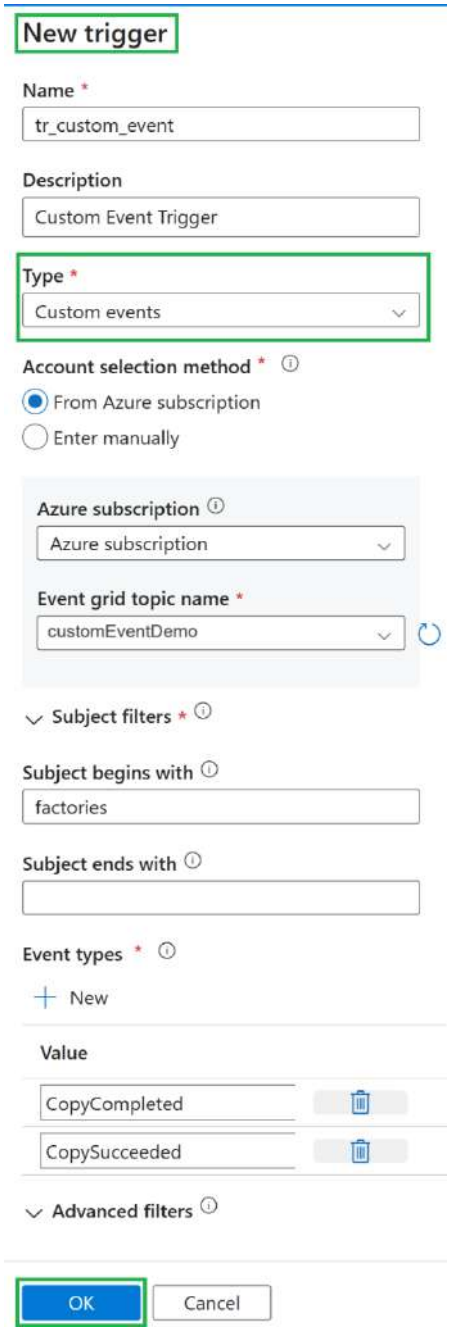

Figure 7.3 – Defining a custom Event trigger for a pipeline

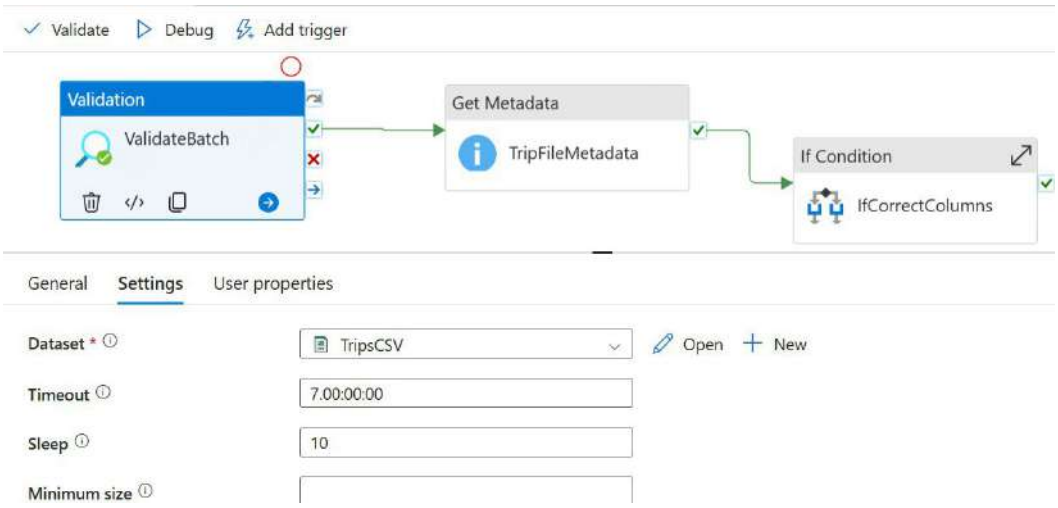

Figure 7.4 – Using the ADF Validation activity to check for a file's existence

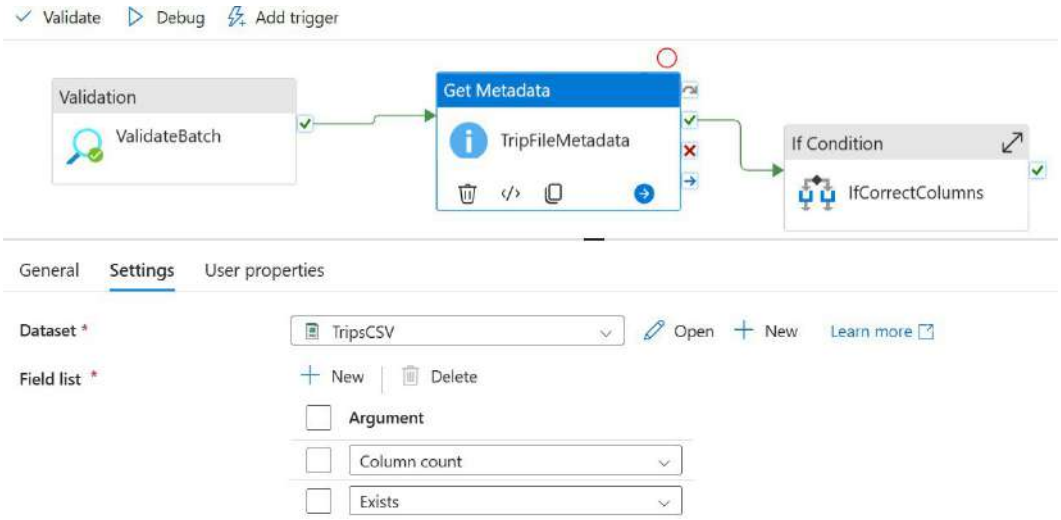

Figure 7.5 – Configuring the Get Metadata activity to publish the column count

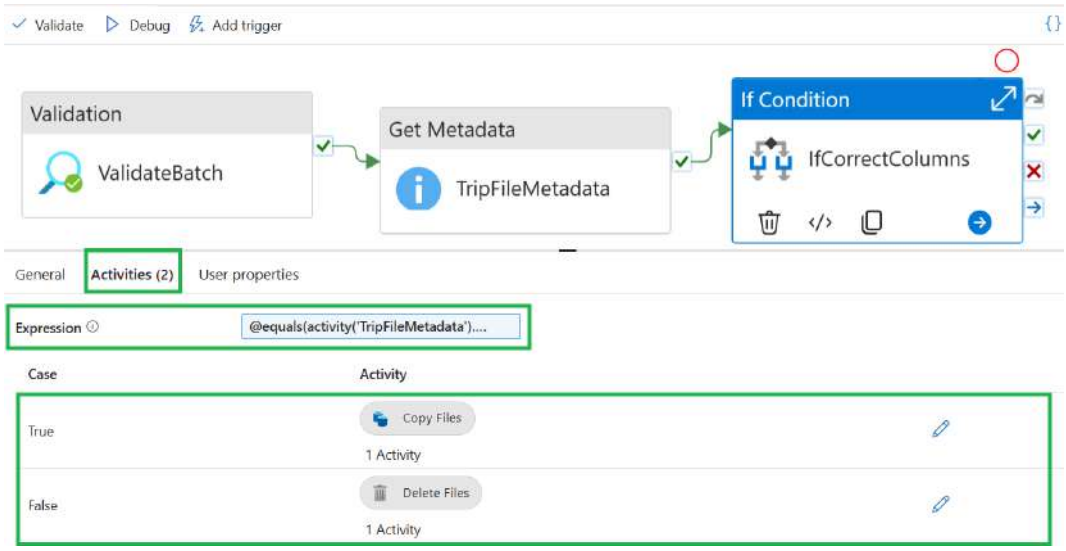

Figure 7.6 – Using the metadata from the Get Metadata activity to make a decision

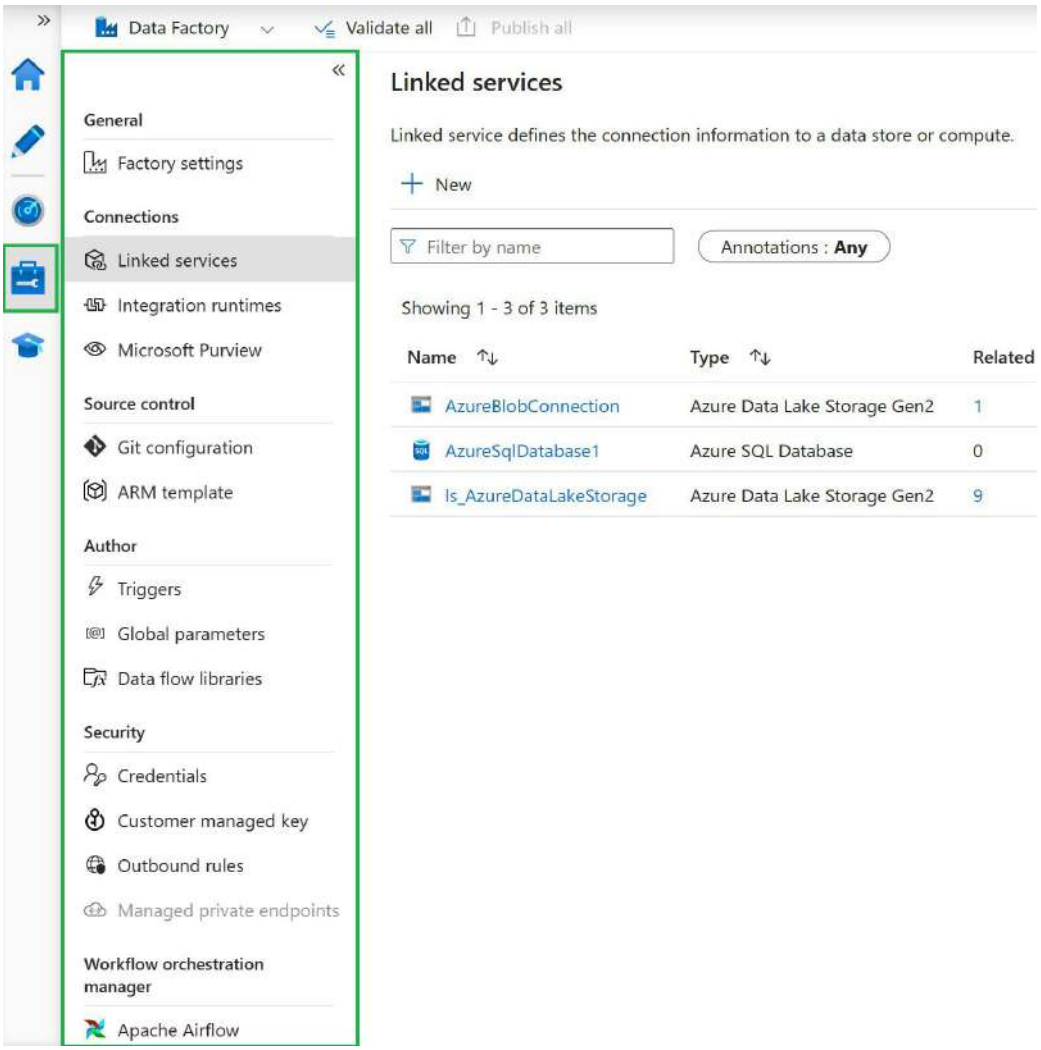

Figure 7.7 – The Manage screen of ADF showing the settings menu options to manage various components

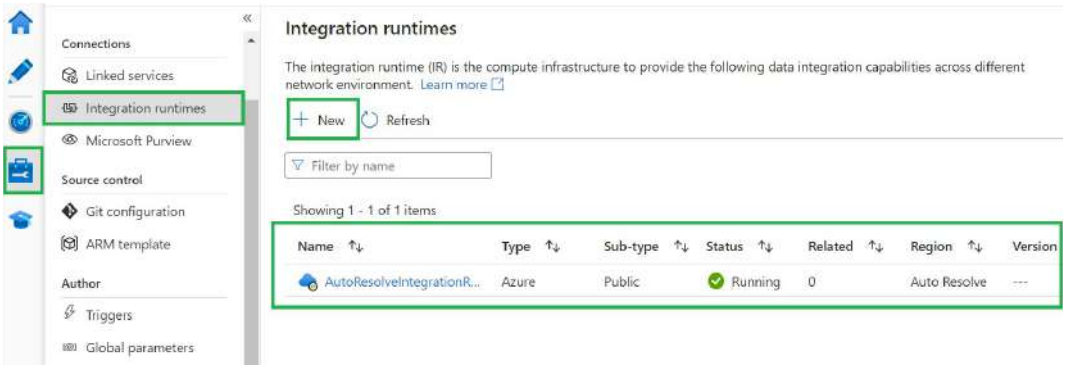

Figure 7.8 – Creating a new IR to run data pipelines and data flows

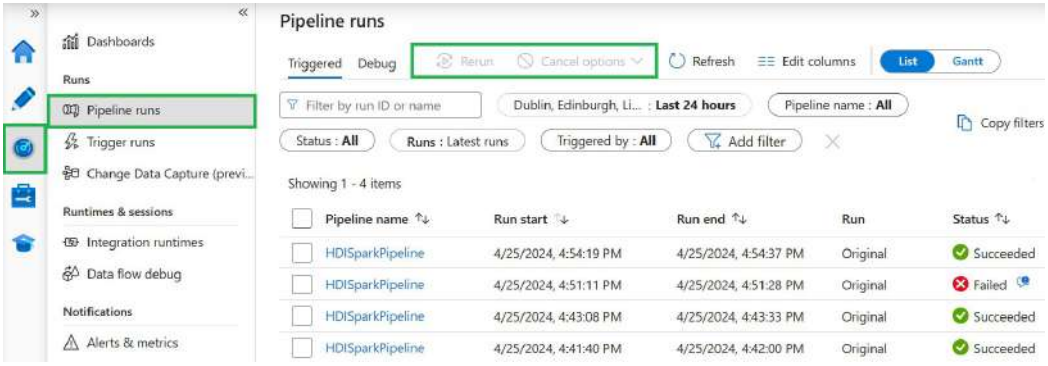

Figure 7.9 – Monitoring ADF pipelines

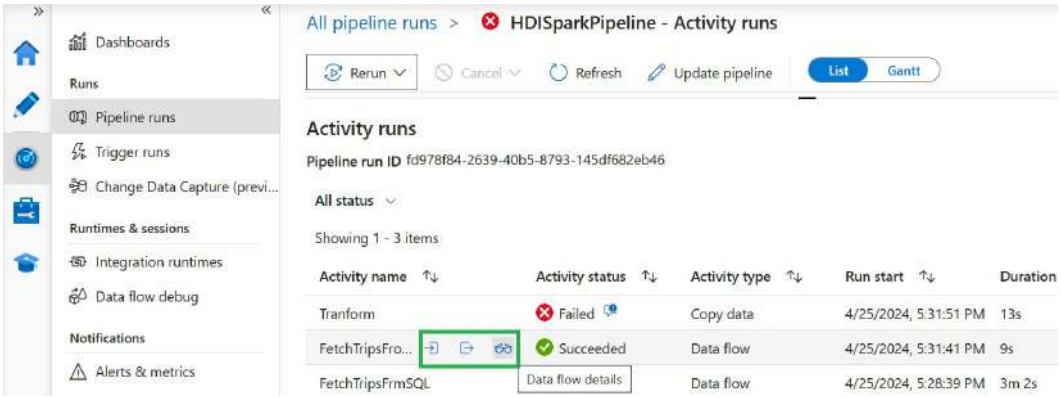

Figure 7.10 – Activity runs tab showing the Data flow details

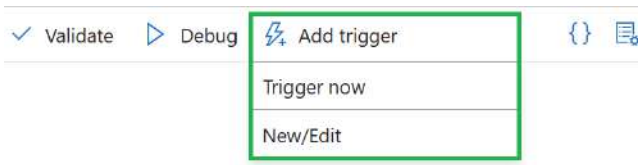

Figure 7.11 – Adding a trigger from ADF/Synapse pipelines

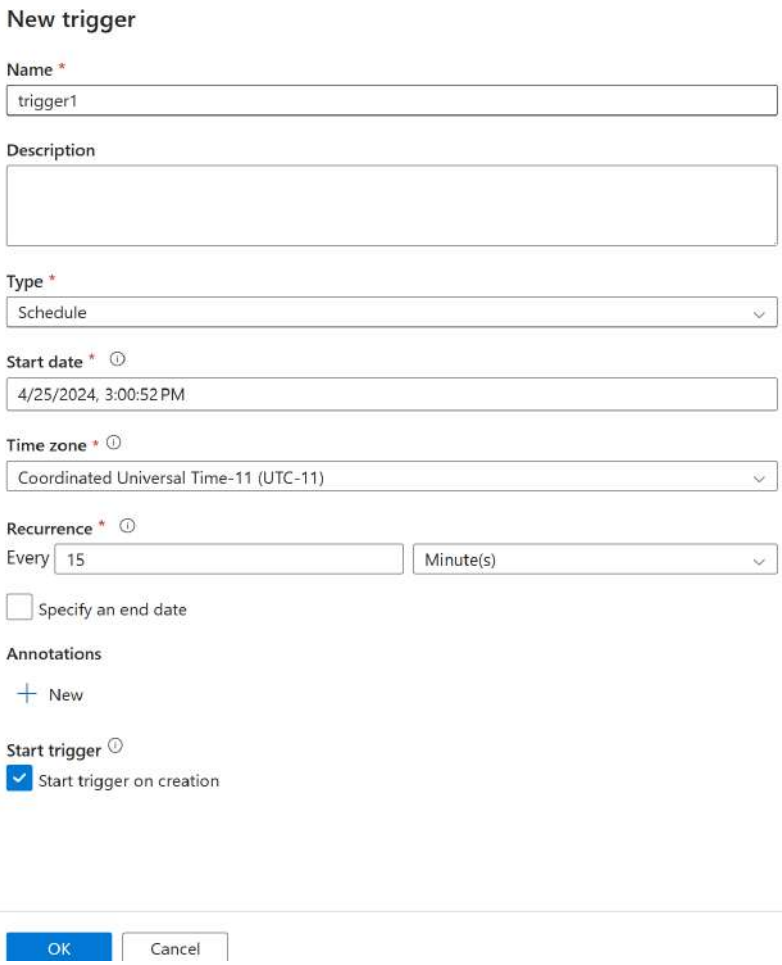

Figure 7.12 – Defining the trigger in ADF or Synapse pipelines

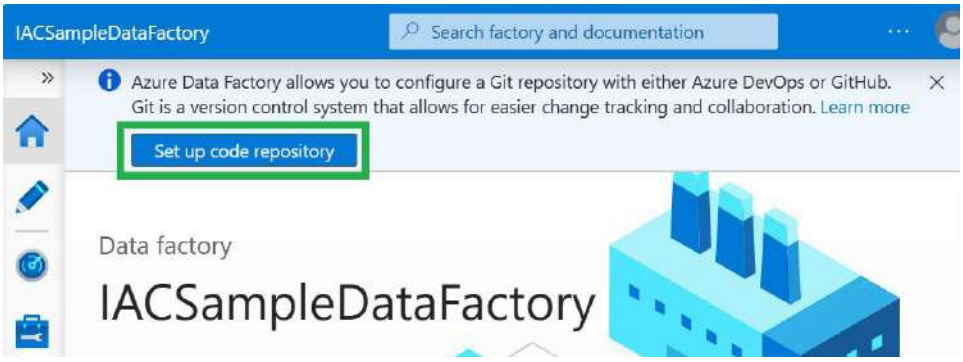

Figure 7.13 – Configuring the Git repository using the Set up code repository button

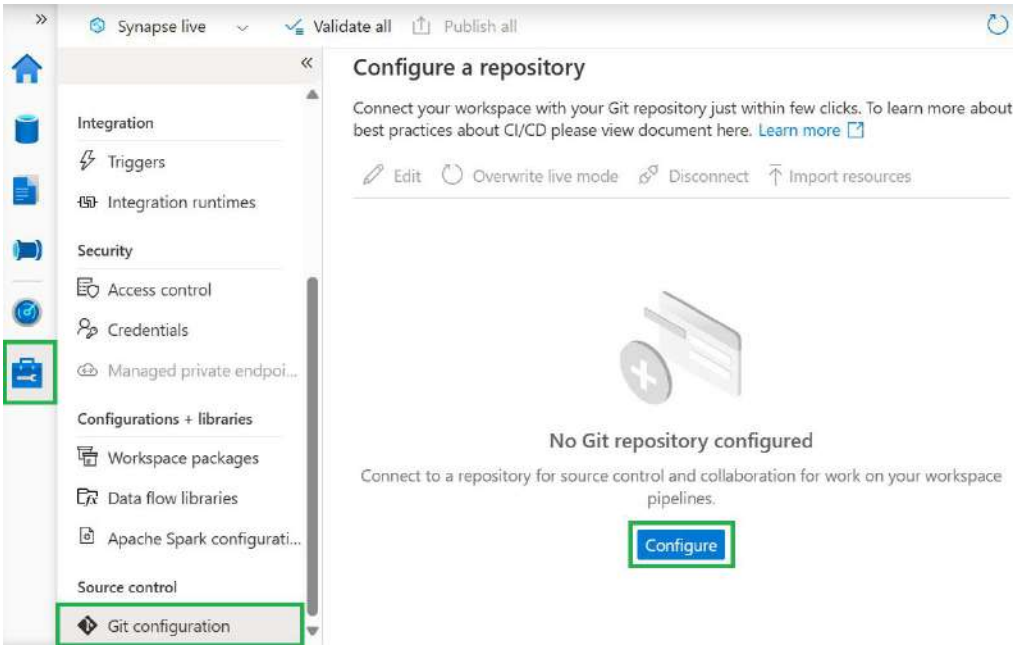

Figure 7.14 – Launching Git configuration from Synapse's Manage tab (the same in ADF)

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

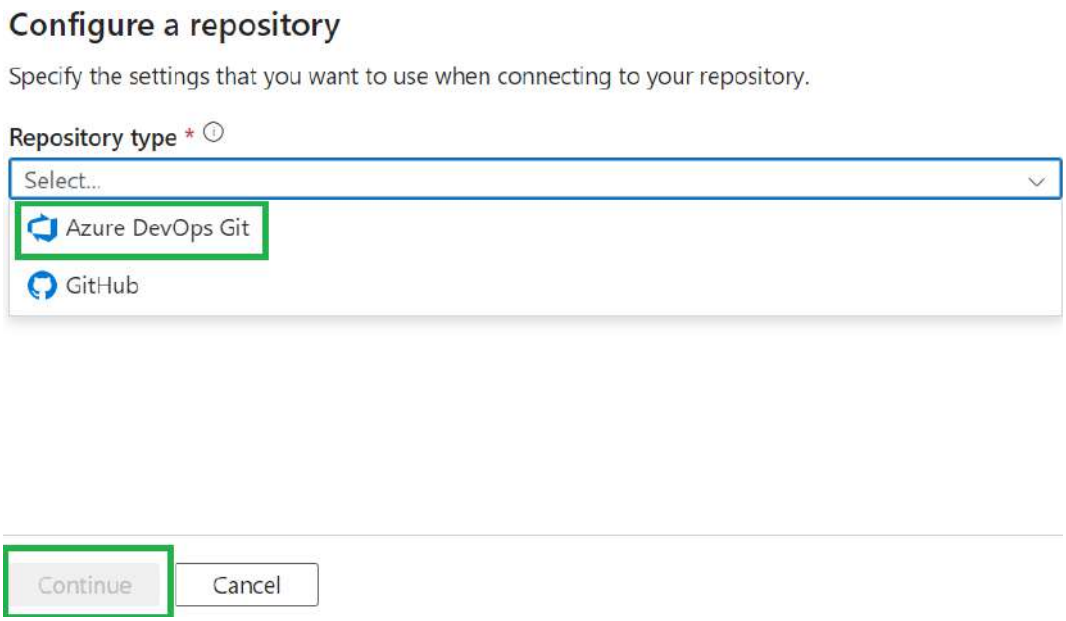

Figure 7.16 – Configuring a repository to select the Azure DevOps Git repository type

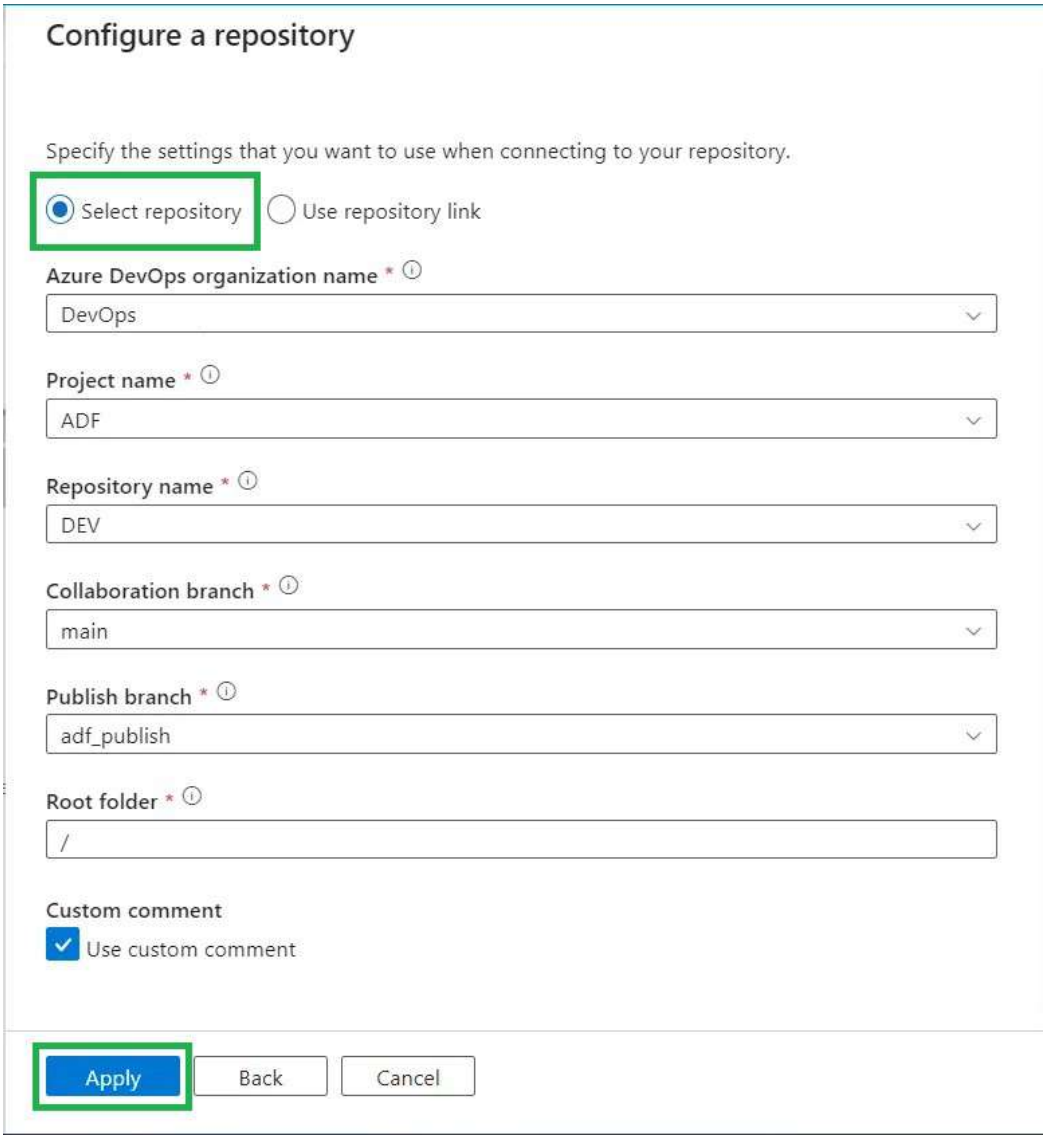

Figure 7.17 – Configuring Azure DevOps Git as the source control for ADF

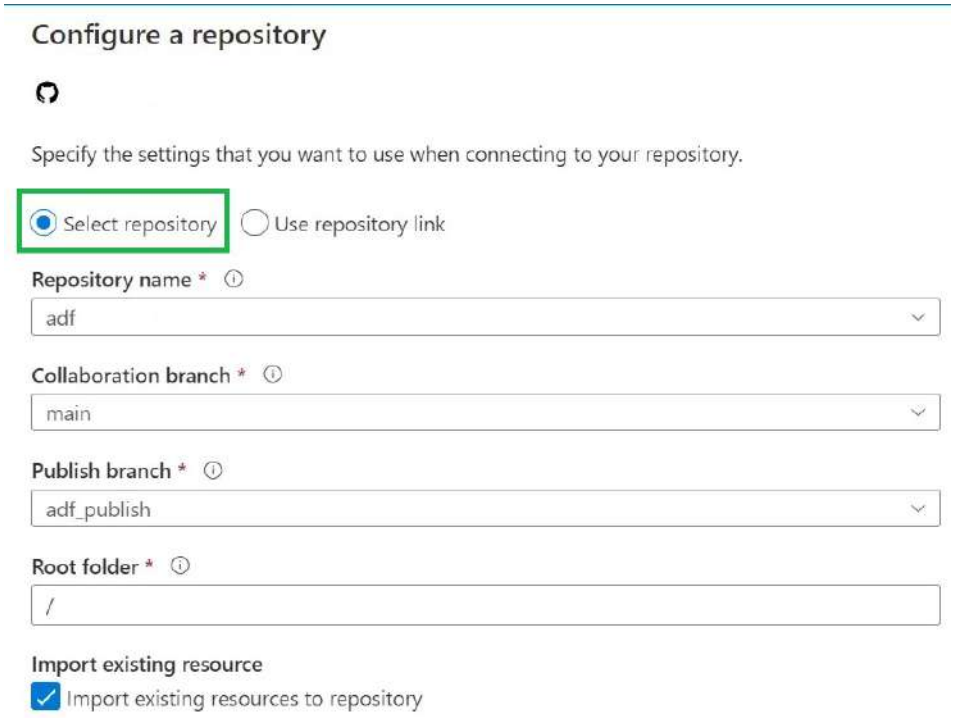

Figure 7.18 – Configuring GitHub as the source control for ADF or Synapse

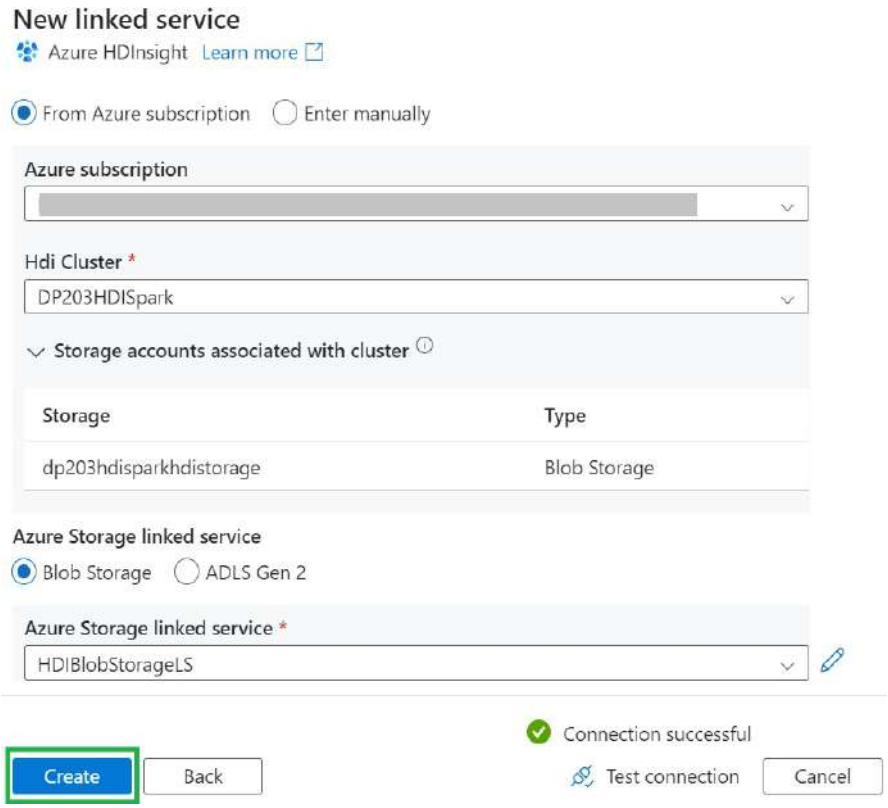

Figure 7.19 – Creating an HDInsight linked service

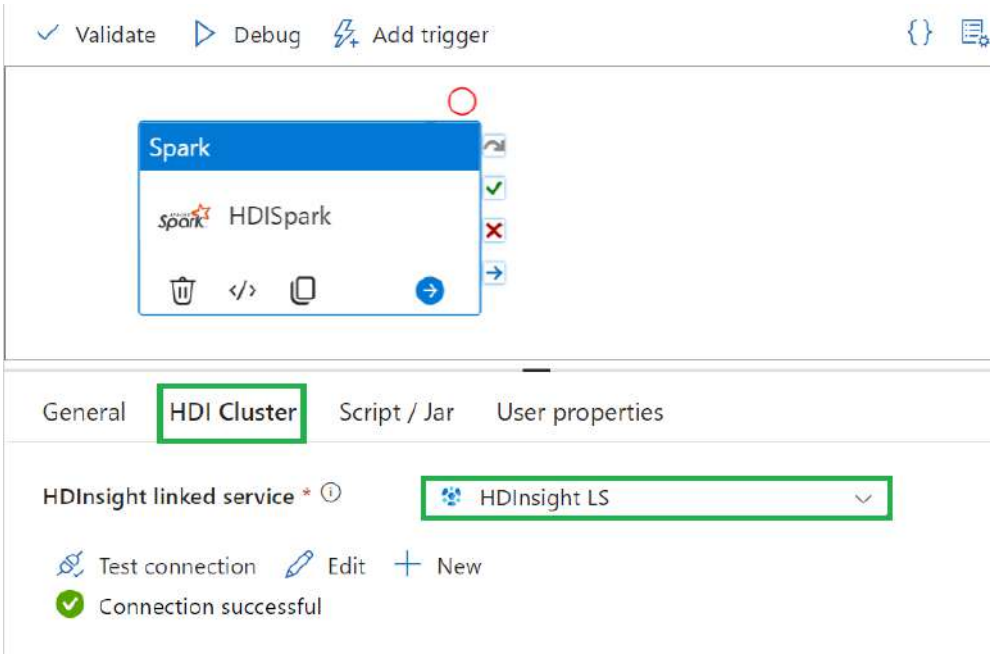

Figure 7.20 – Configuring the HDInsight Spark cluster in the ADF pipeline

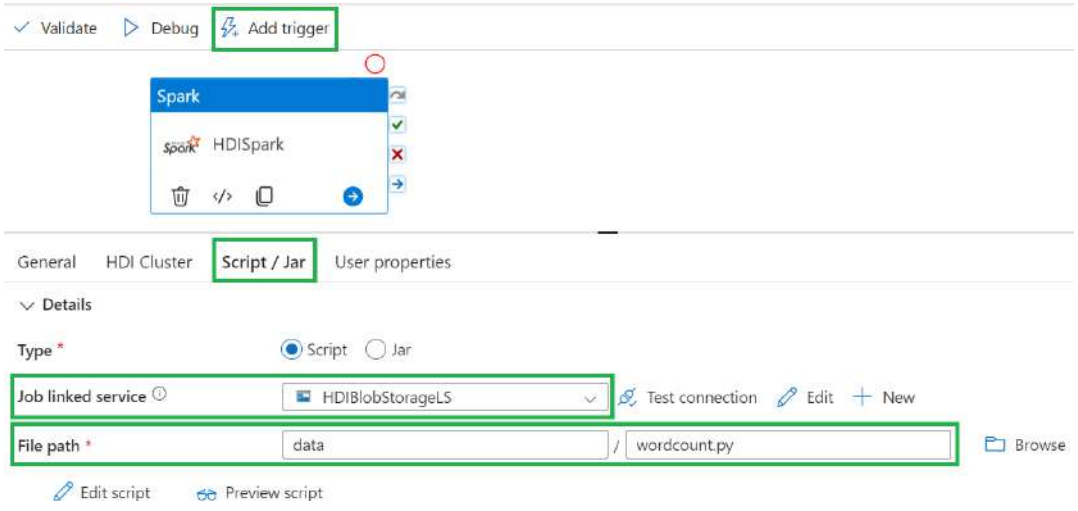

Figure 7.21 – Creating a pipeline with Spark

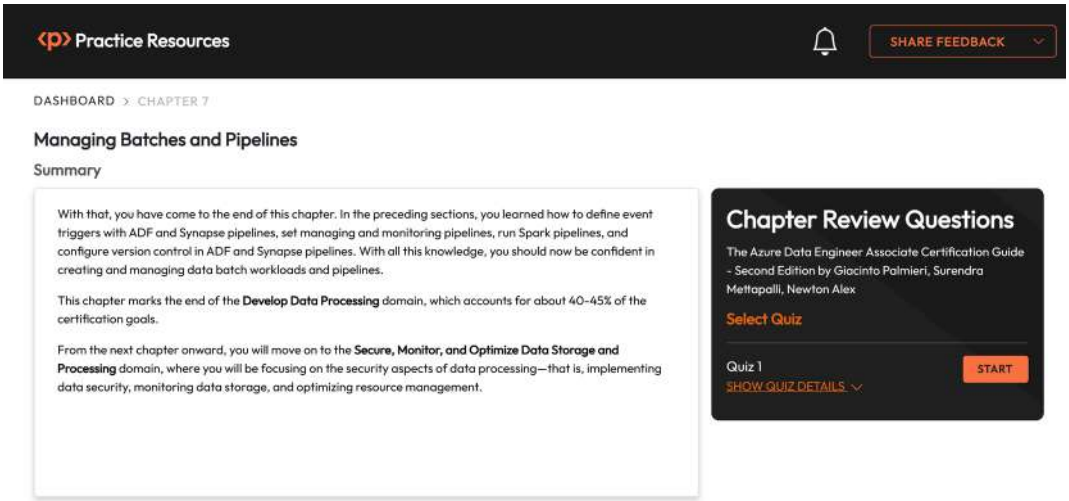

Figure 7.23 – Chapter Review Questions for Chapter 7

Chapter 8: Implementing Data Security

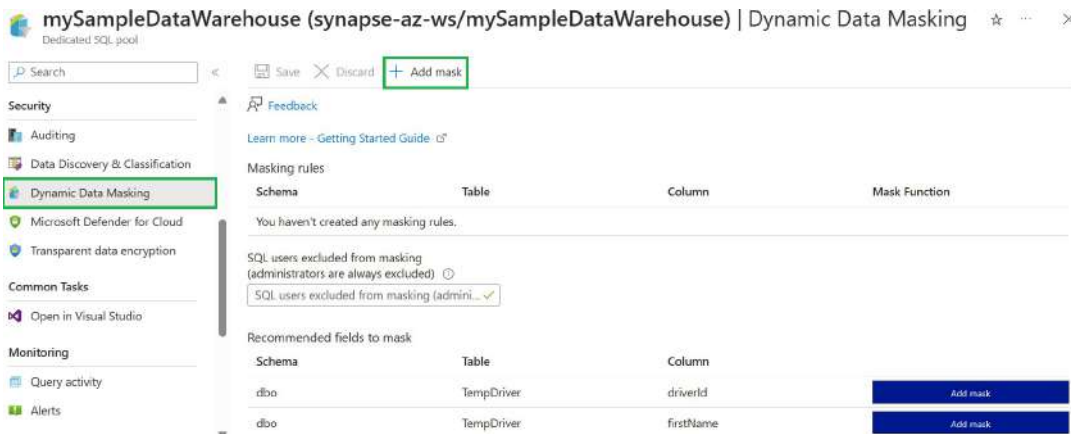

Figure 8.1 – Setting up DDM from the Azure portal

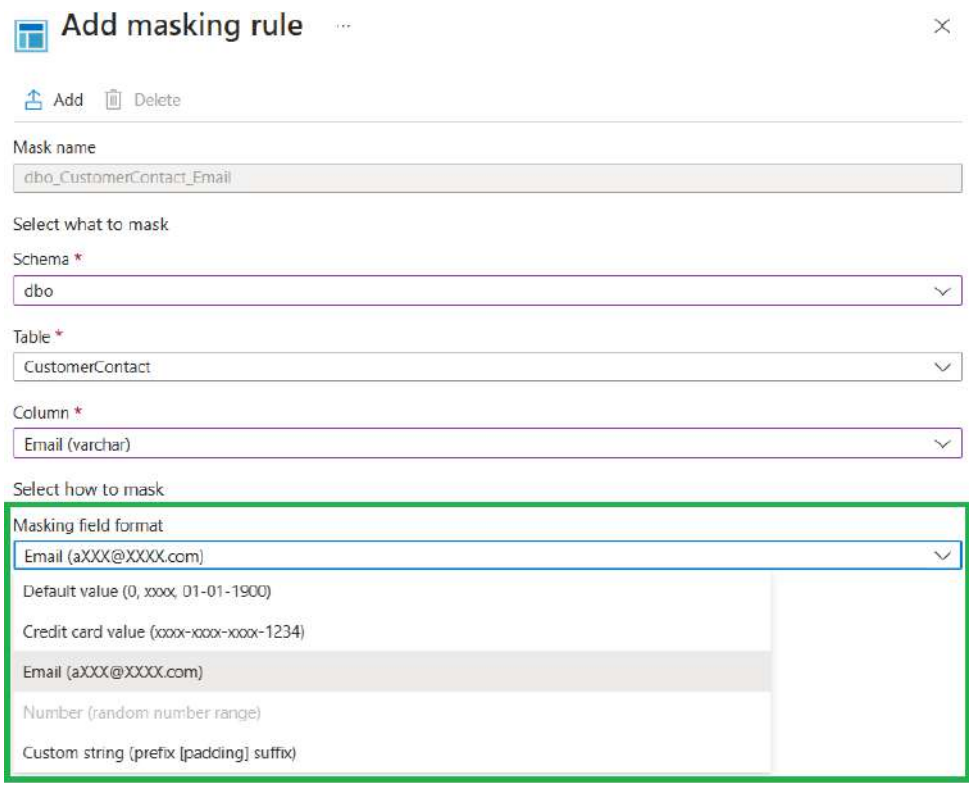

Figure 8.2 – Creating an email mask

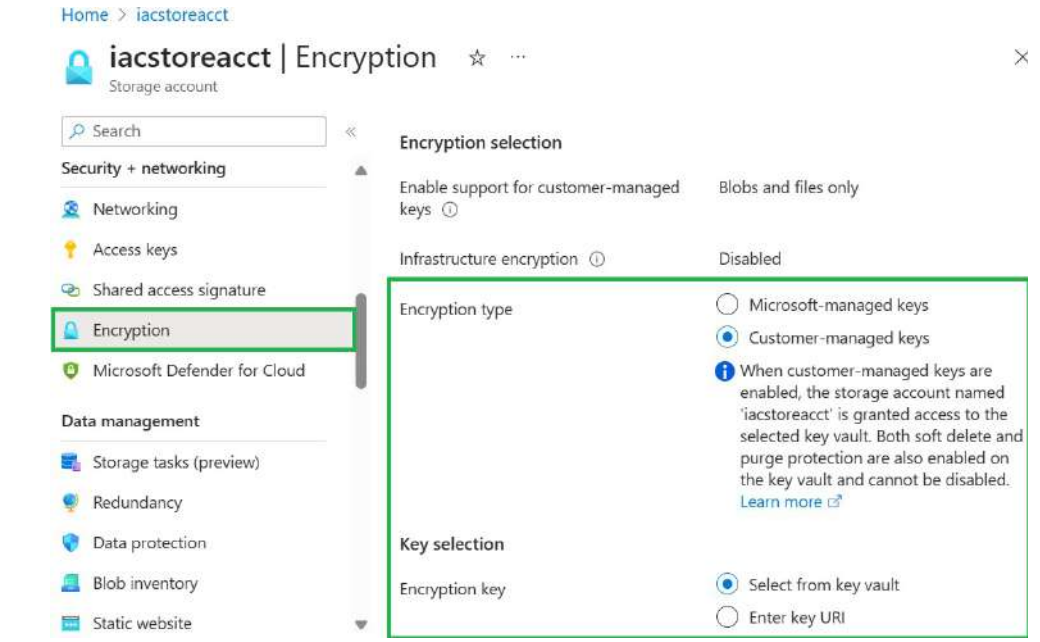

Figure 8.3 – Enabling CMKs in Azure Storage

Figure 8.4 – Enabling TDE using Azure Synapse SQL

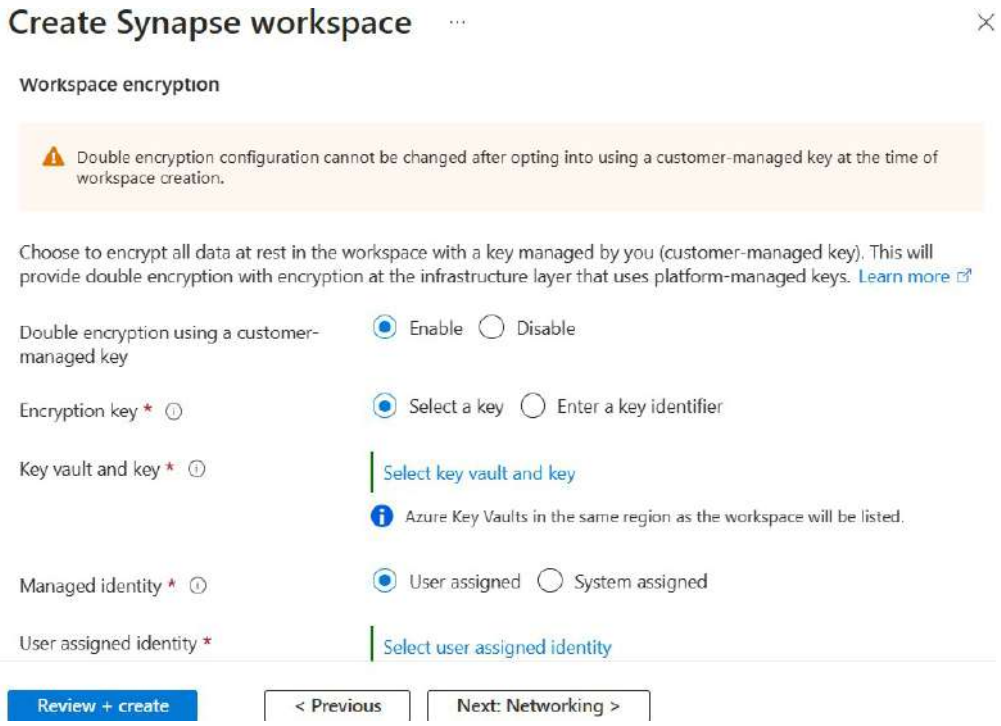

Figure 8.5 – Configuring a CMK in Azure Synapse SQL

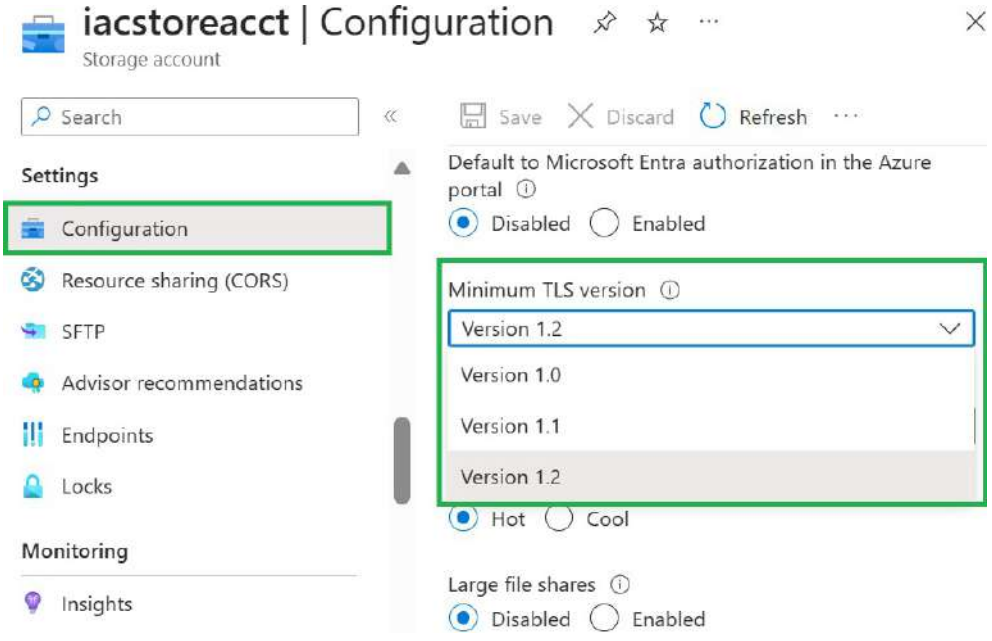

Figure 8.6 – Enabling TLS in Azure Storage

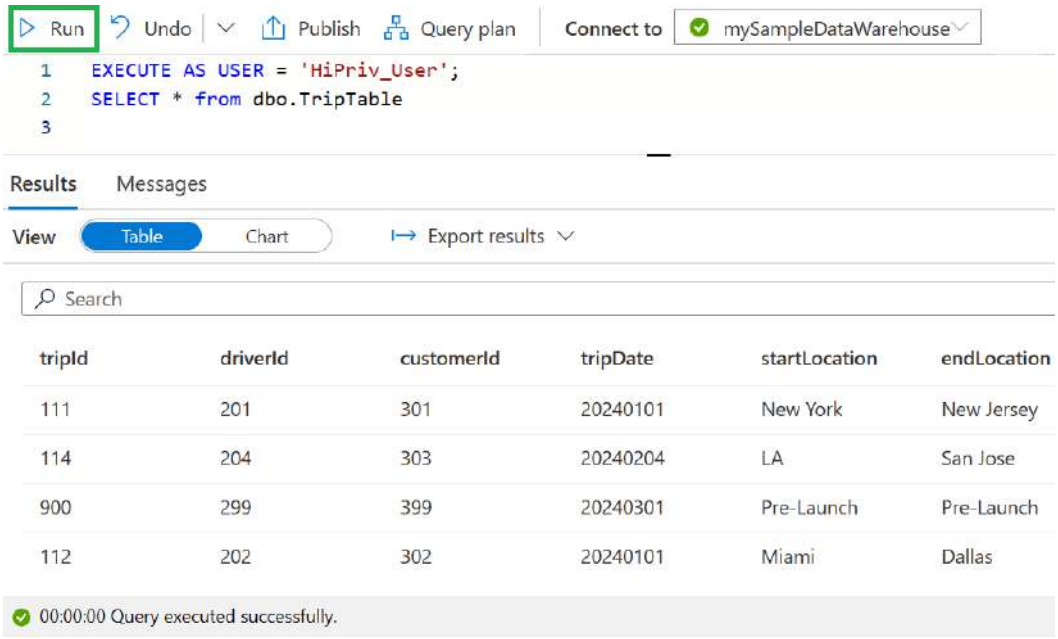

Figure 8.7 – Displaying all rows including pre-launch for user HiPriv_User

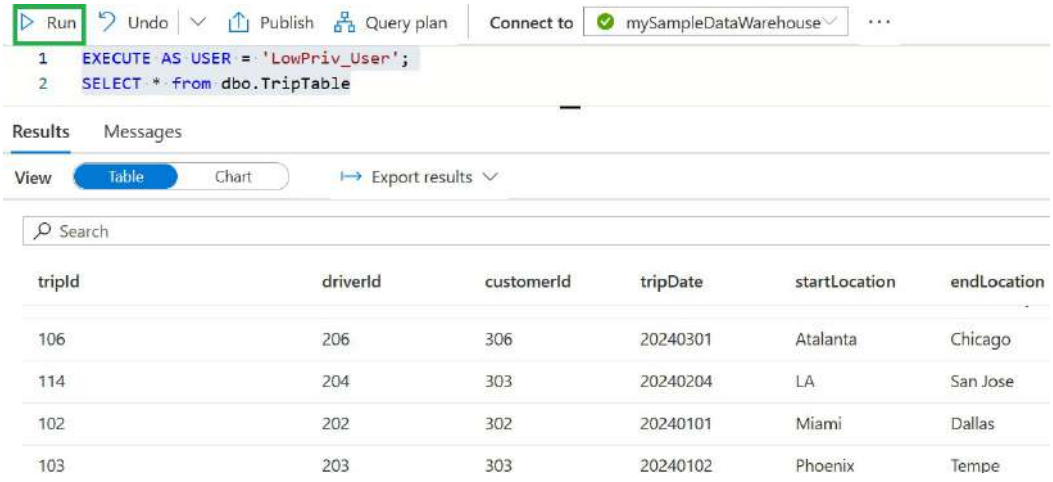

Figure 8.8 – Row-Level Security (RLS) blocking the pre-launch location rows

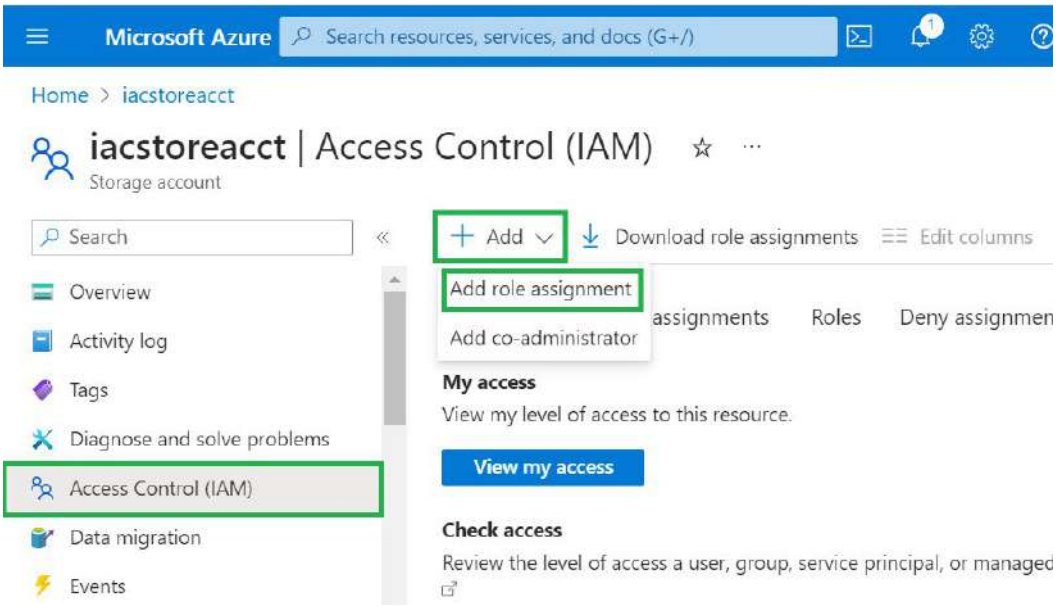

Figure 8.9 – Adding role assignment in storage account through Access Control IAM

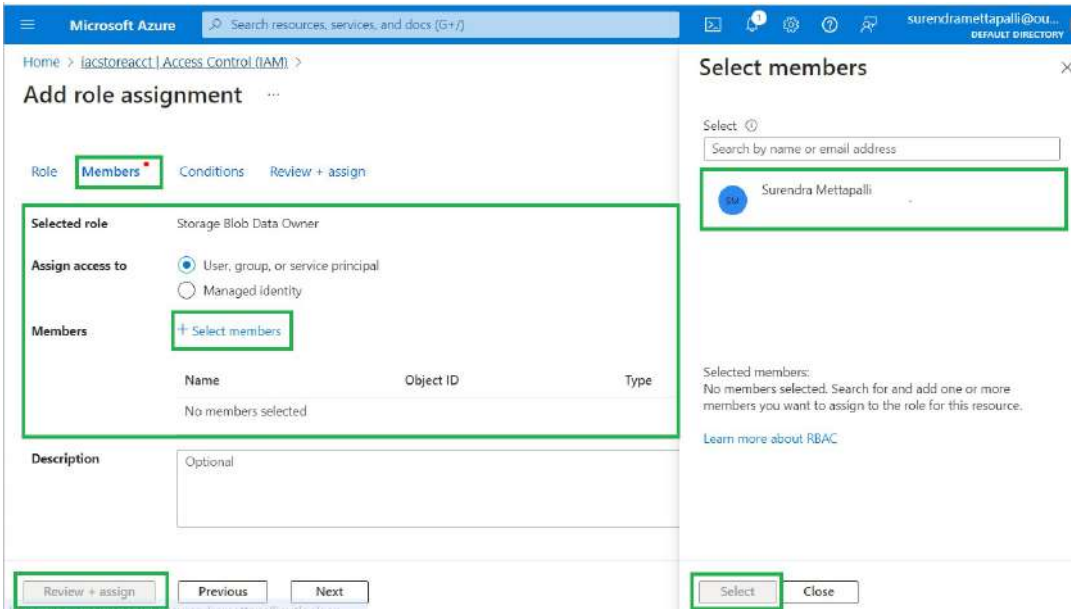

Figure 8.10 – Configuring the RBAC role assignment in ADLS Gen2

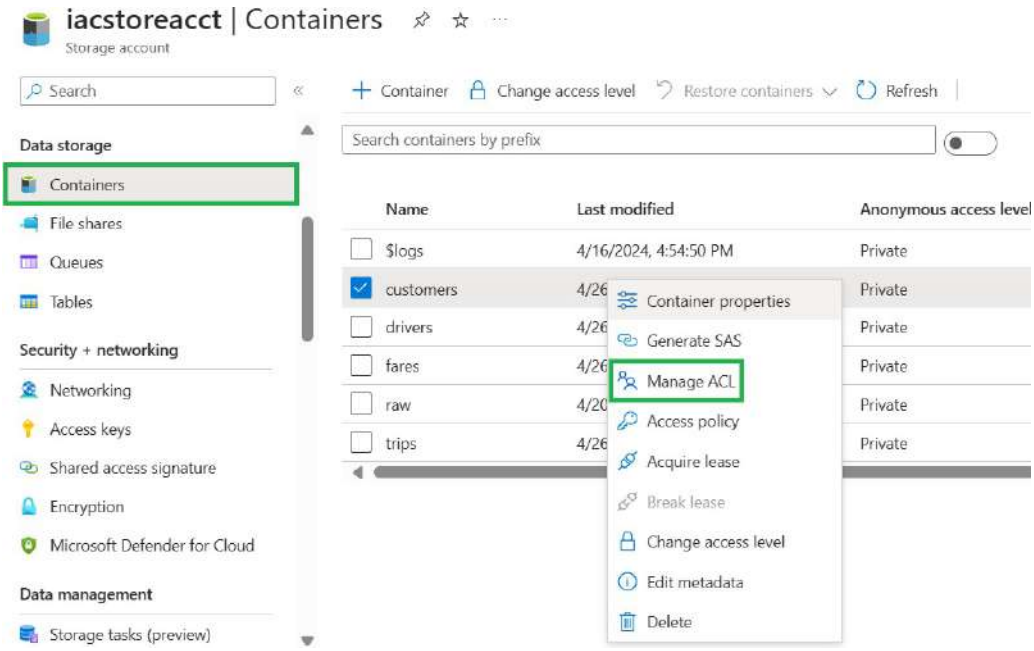

Figure 8.11 – Selecting Manage ACL to provide required access permissions

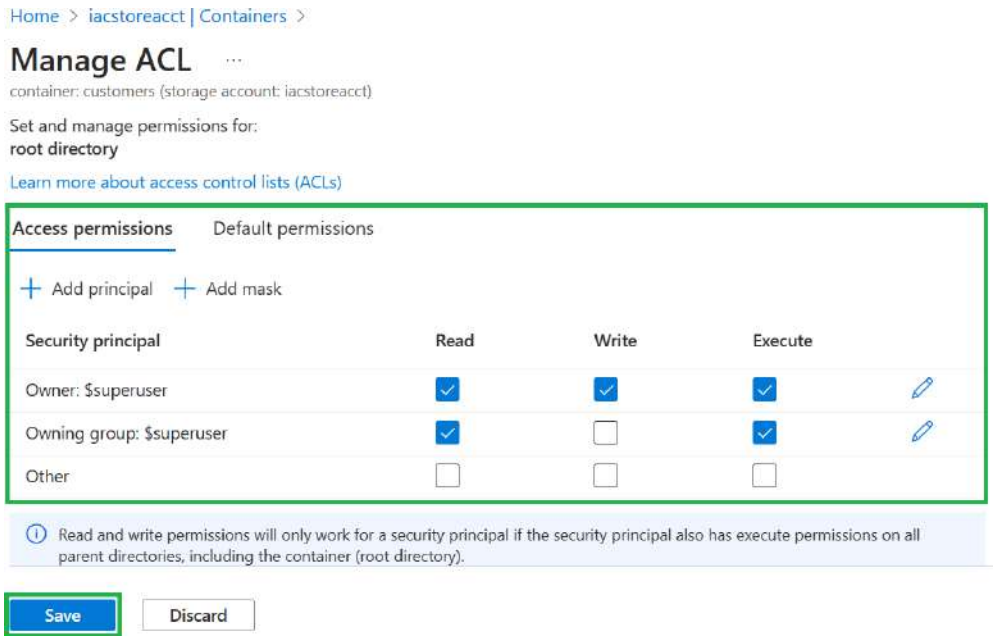

Figure 8.12 – Configuring the ACL in ADLS Gen2 to provide the required access to users

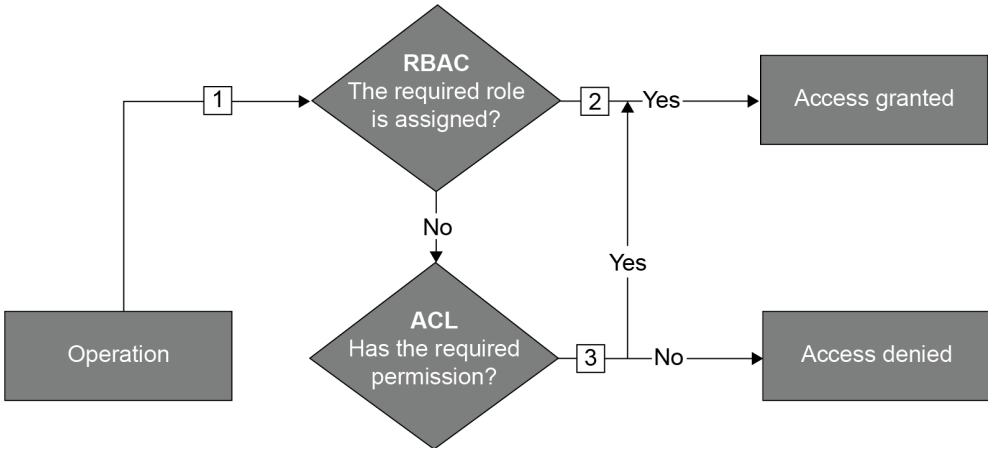

Figure 8.13 – RBAC and ACL evaluation sequence

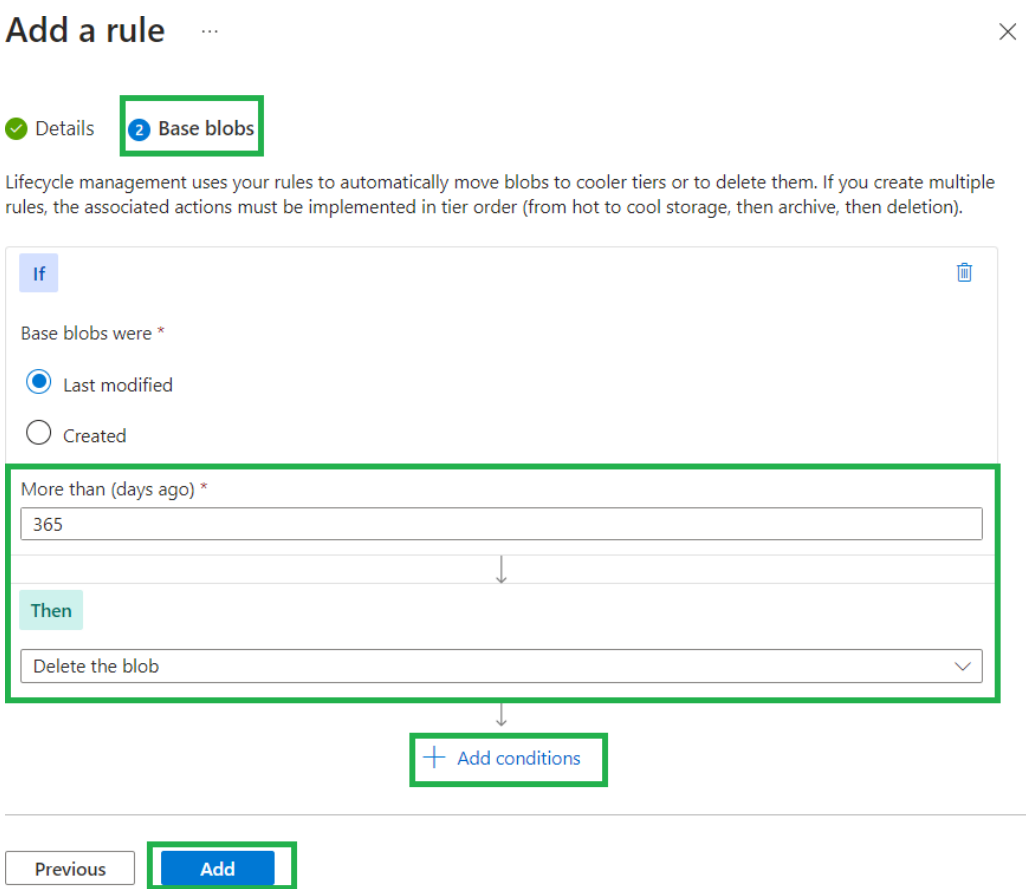

Figure 8.14 – Configuring data life cycle management

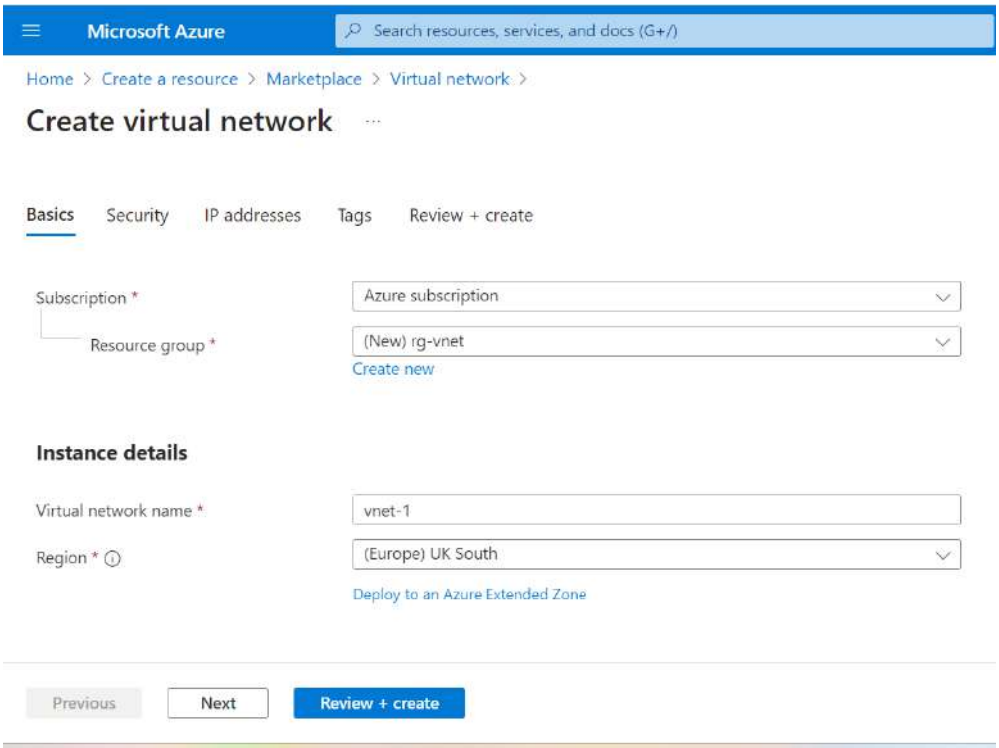

Figure 8.15 – Creating a new VNet

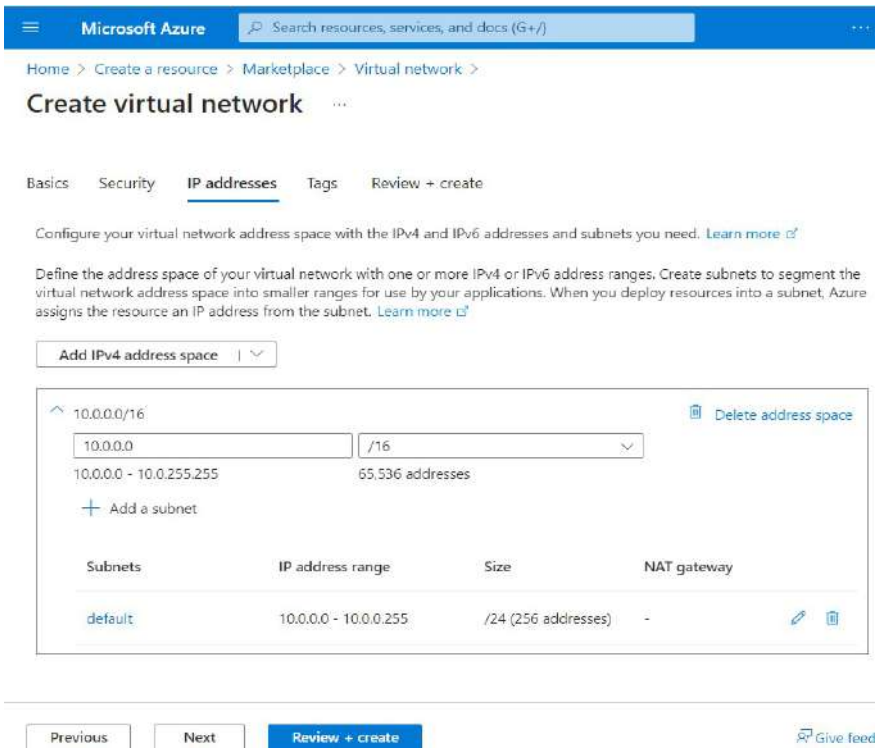

Figure 8.16 – Configuring IP details for the new VNet

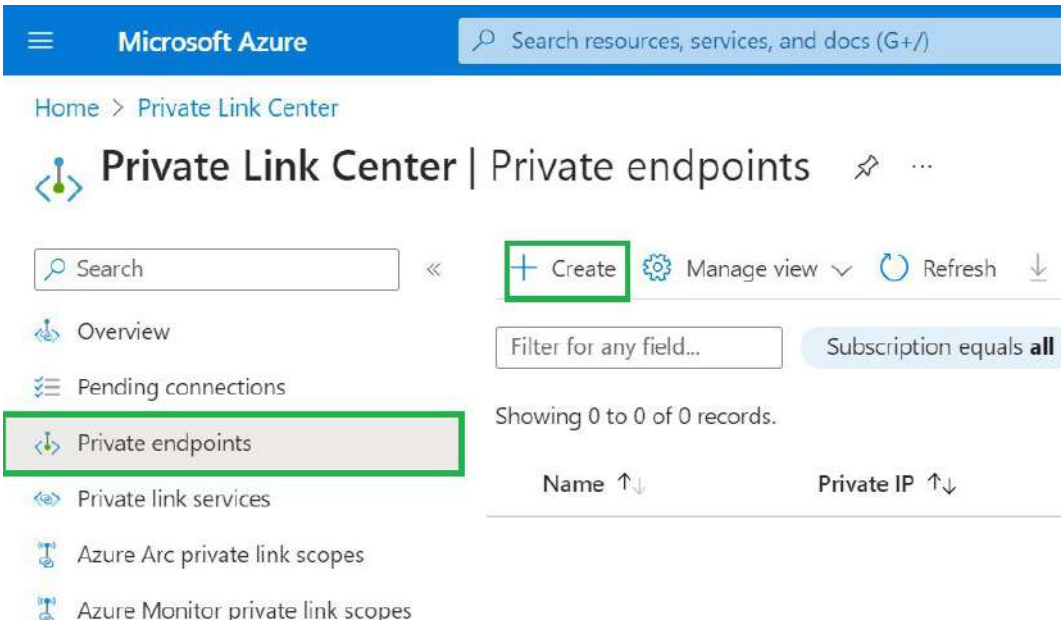

Figure 8.17 – Creating a private endpoint from Private Link Center

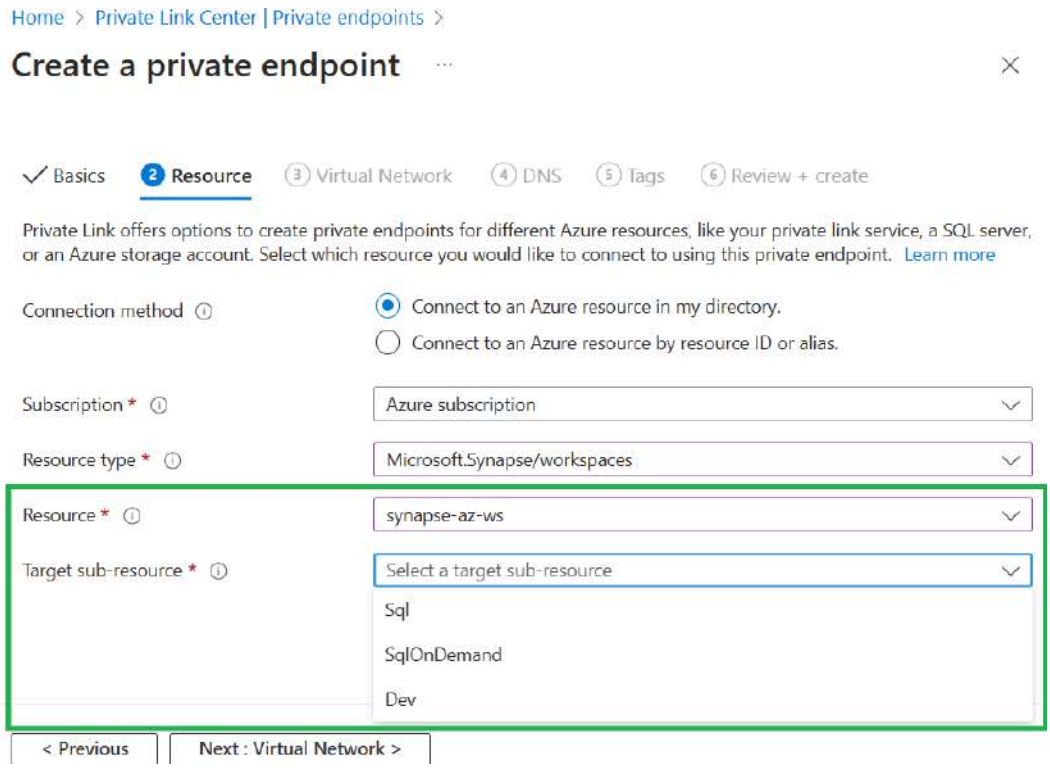

Figure 8.18 – Configuring a service for creating the private endpoint

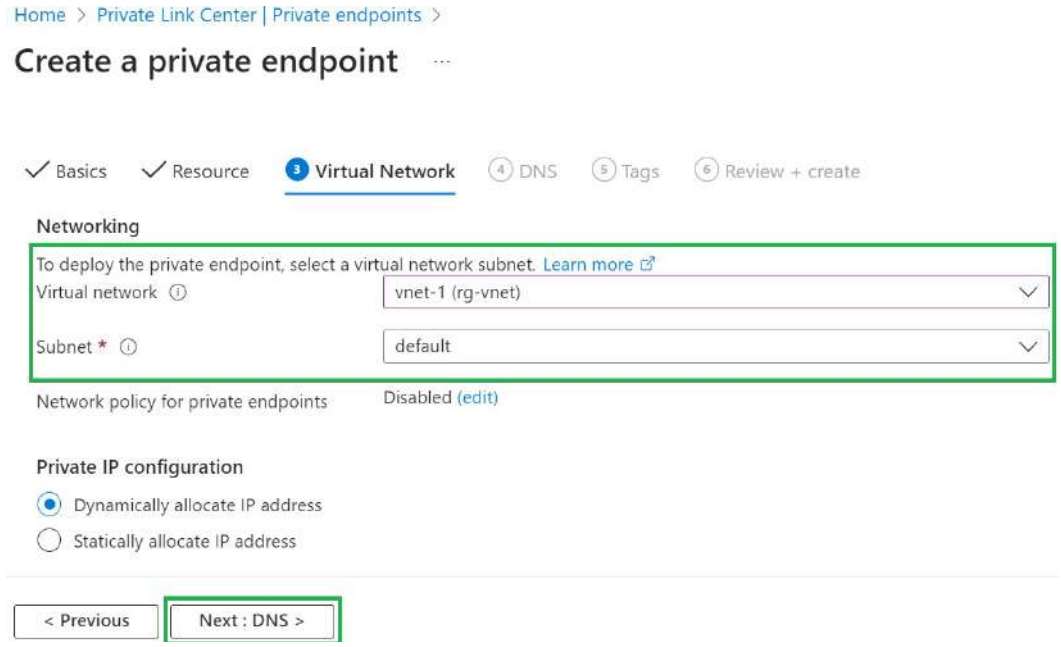

Figure 8.19 – Configuring VNet for private endpoint

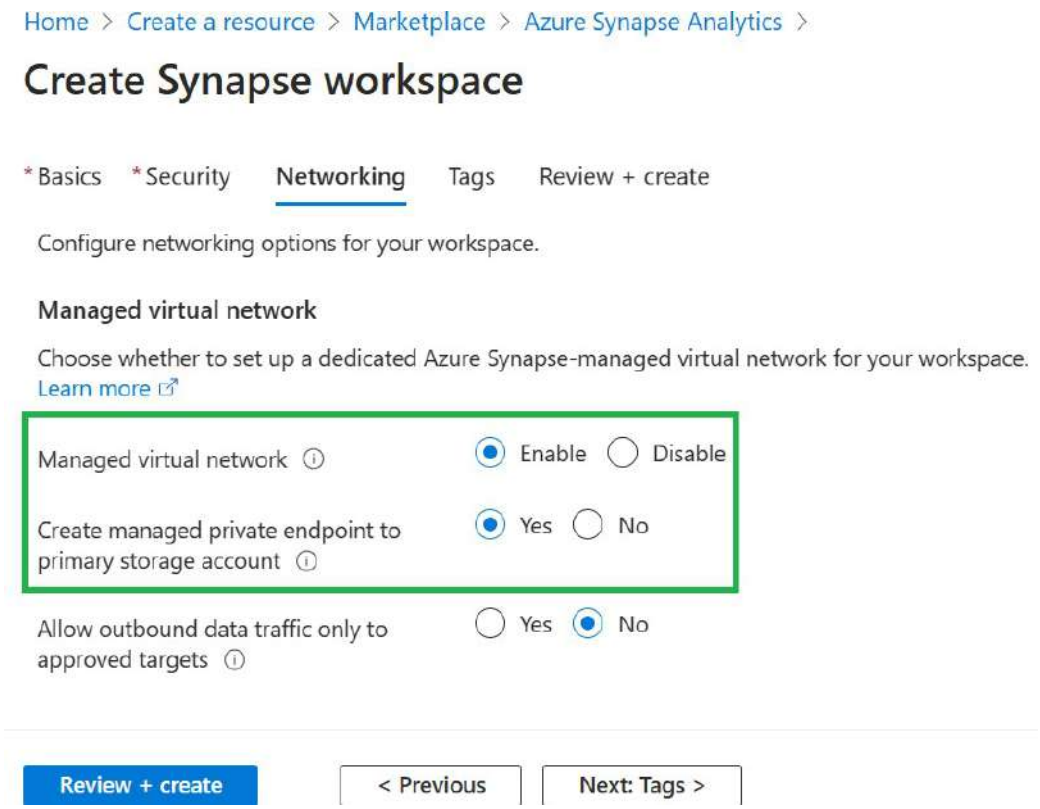

Figure 8.20 – Enabling a Synapse-managed VNet to create a private endpoint

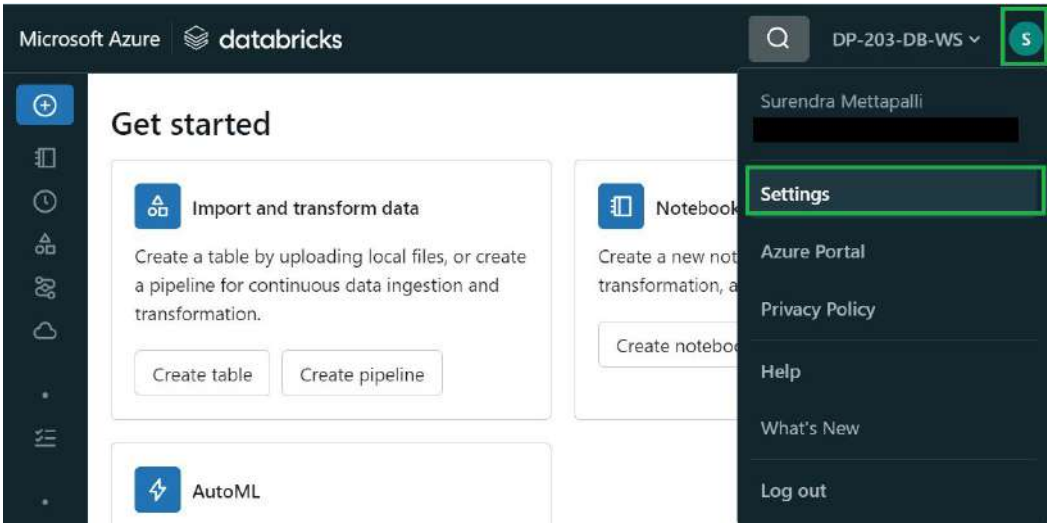

Figure 8.21 – Accessing User Settings in Azure Databricks for developer settings

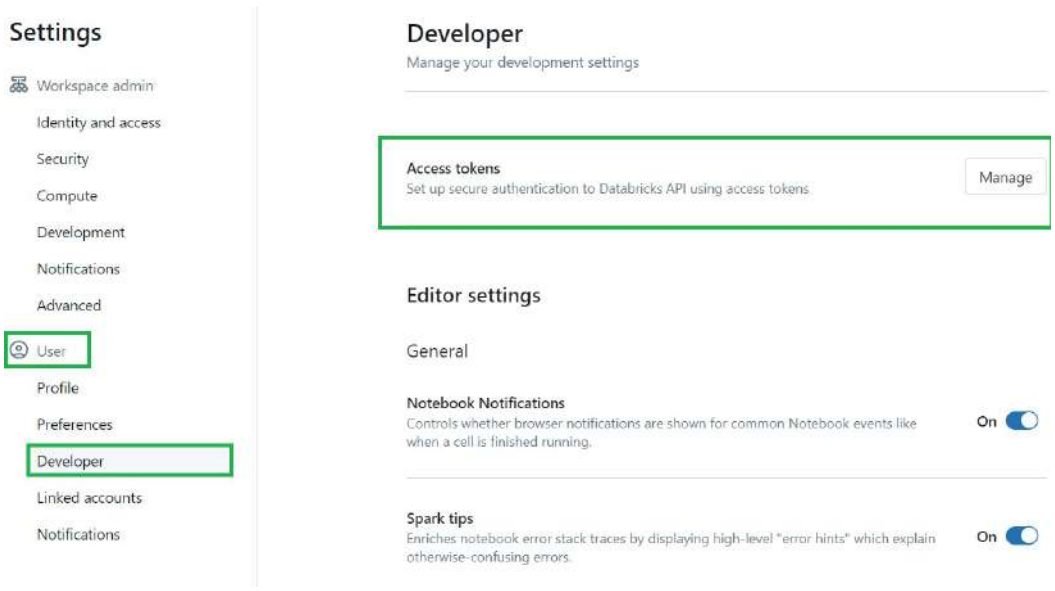

Figure 8.22 – Accessing Developer settings to manage access tokens

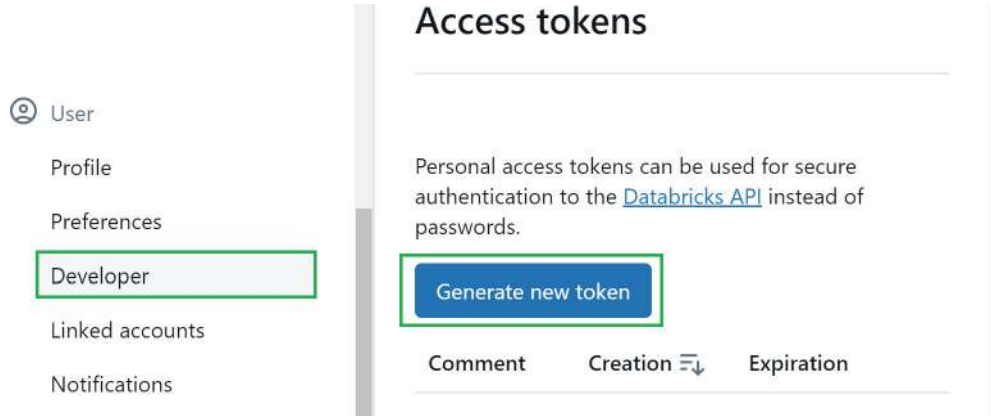

Figure 8.23 – Generating New Token in Azure Databricks

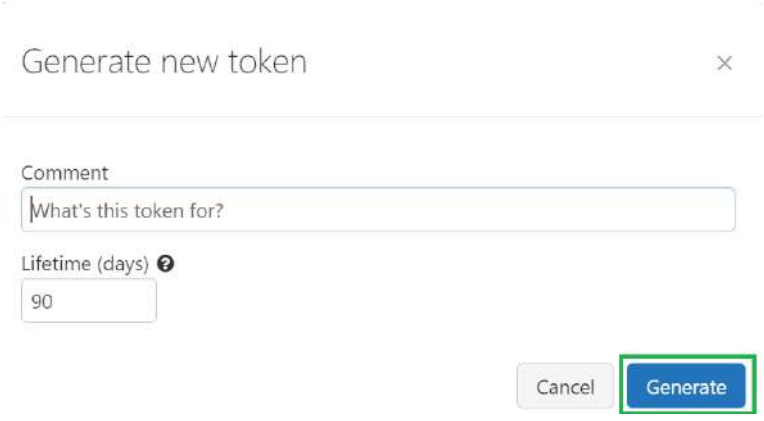

Figure 8.24 – Creating a new Azure Databricks PAT

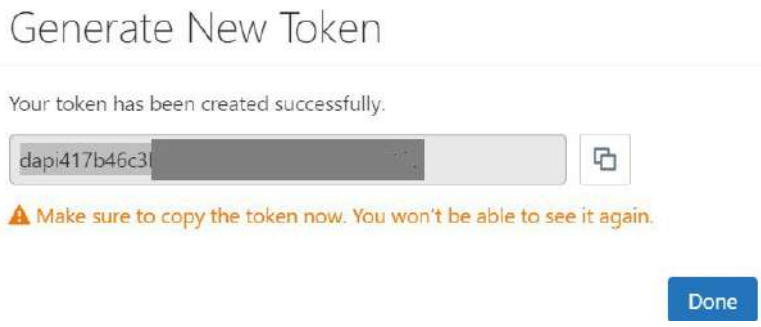

Figure 8.25 – Copying the successfully generated token

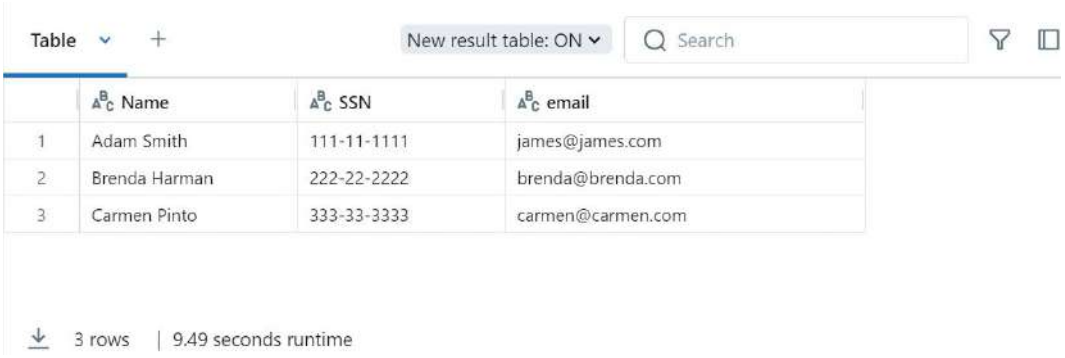

Figure 8.26 – Output displaying PII, such as SSN

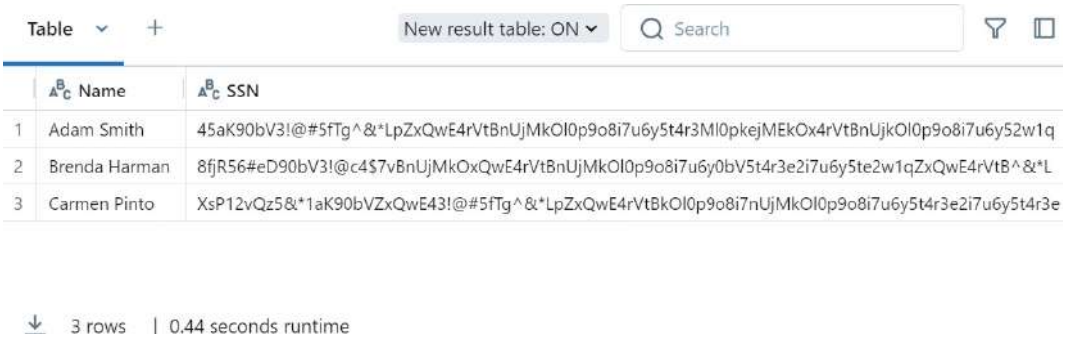

Figure 8.27 – Displaying output with encrypted PII

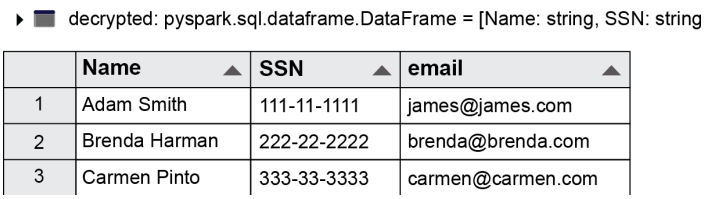

Figure 8.28 – Displaying output with decrypted PII

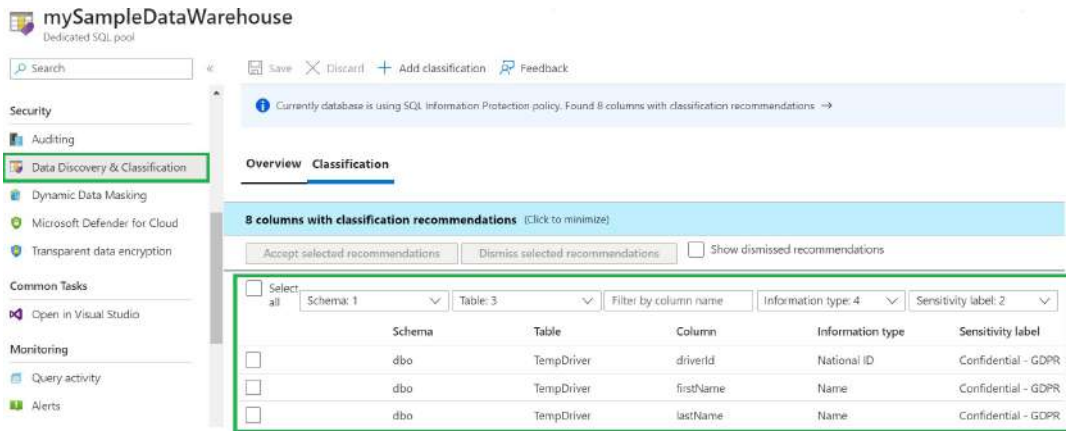

Figure 8.29 – Displaying Data Discovery & Classification in Synapse SQL

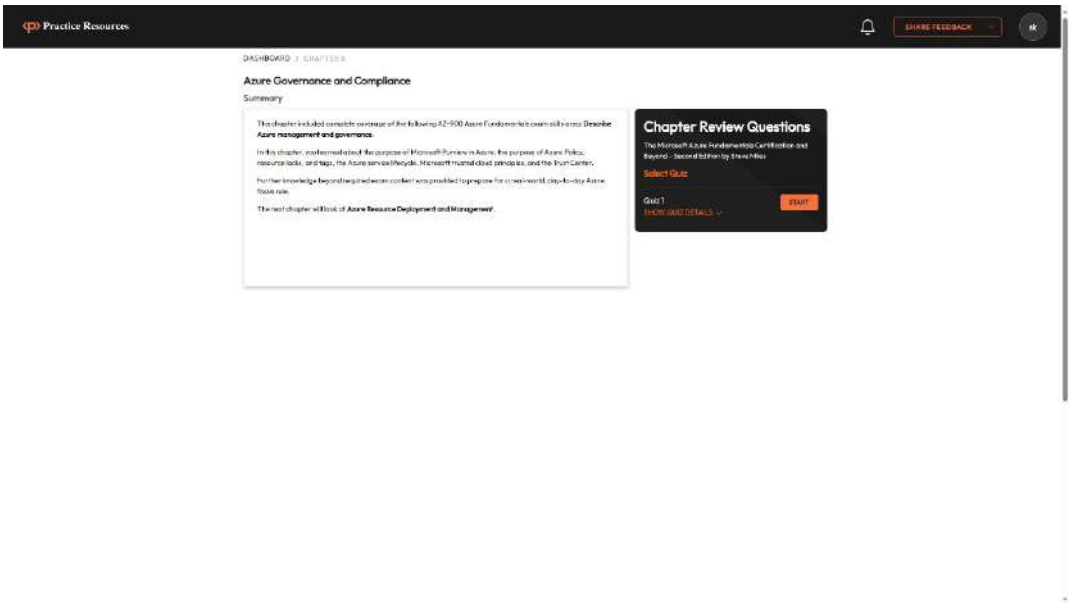

Figure 8.31 – Chapter Review Questions for Chapter 8

Chapter 9: Monitoring Data Storage and Data Processing

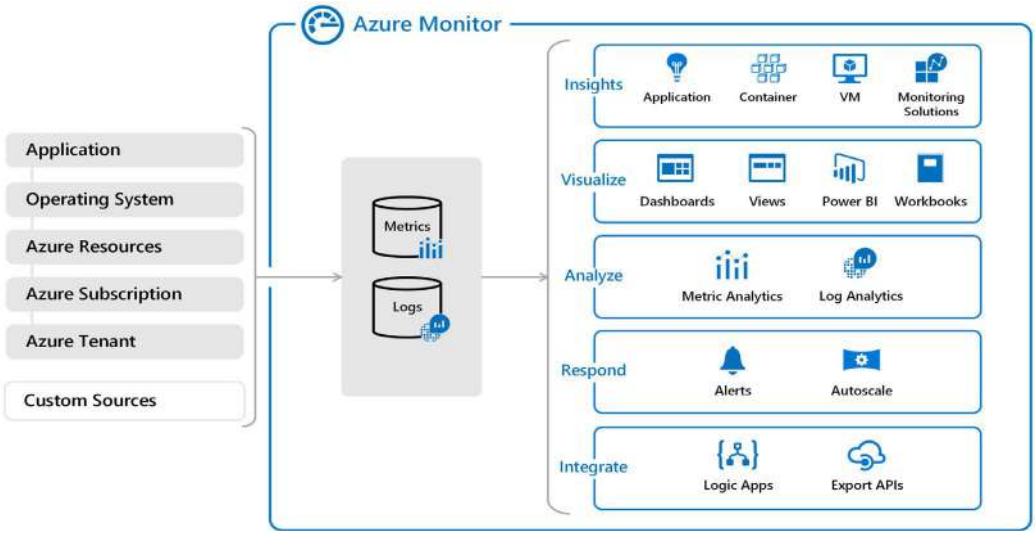

Figure 9.1 – Azure Monitor architecture showing the core components of Azure Monitor, Metrics, and Logs

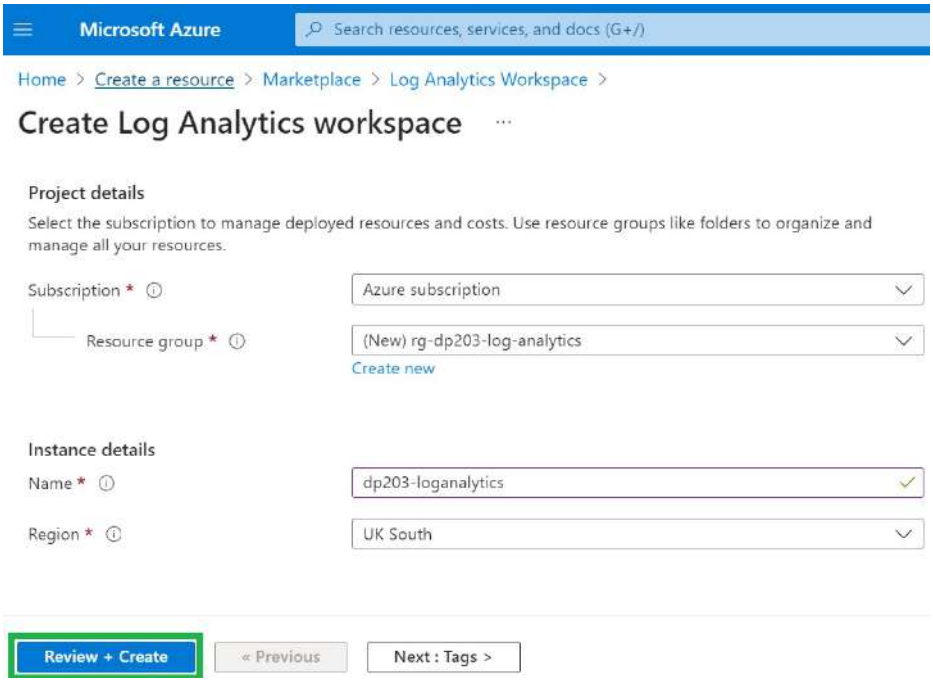

Figure 9.2 – Choosing a subscription to deploy resources during Log Analytics workspace creation

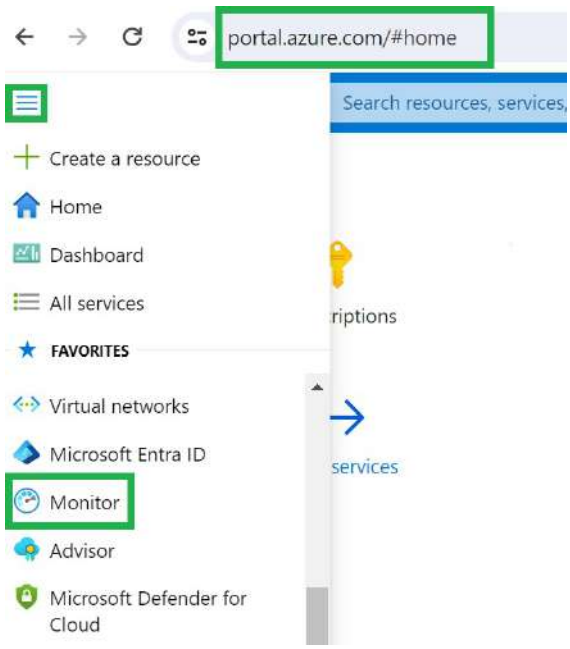

Figure 9.3 – Selecting the Monitor service from the Azure portal

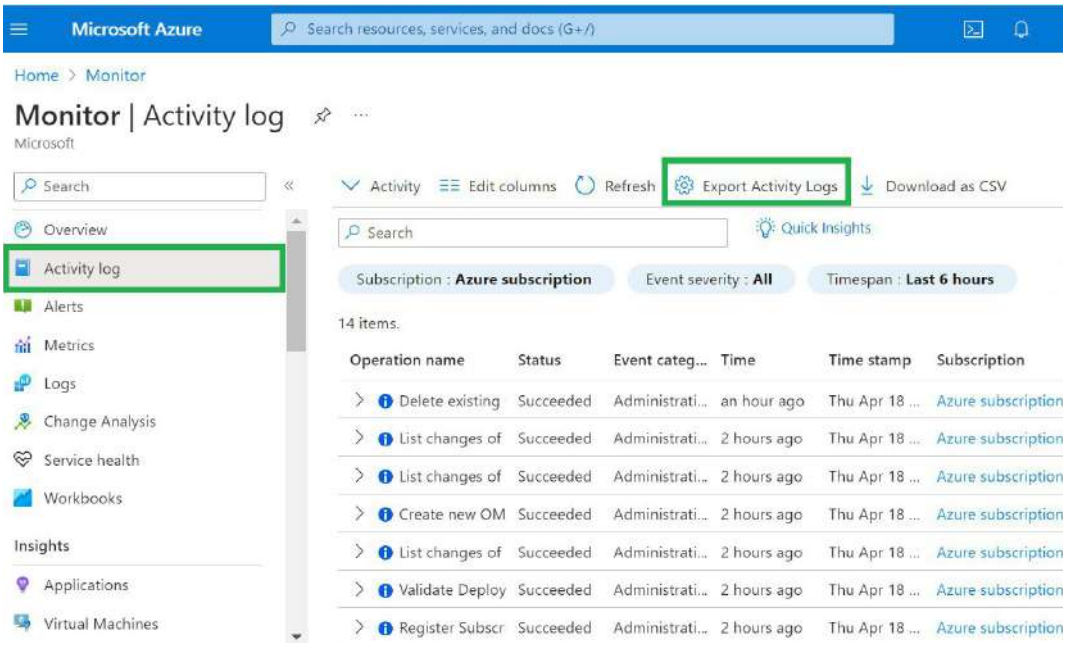

Figure 9.4 – Azure Monitor Activity log screen showing all the activity logs

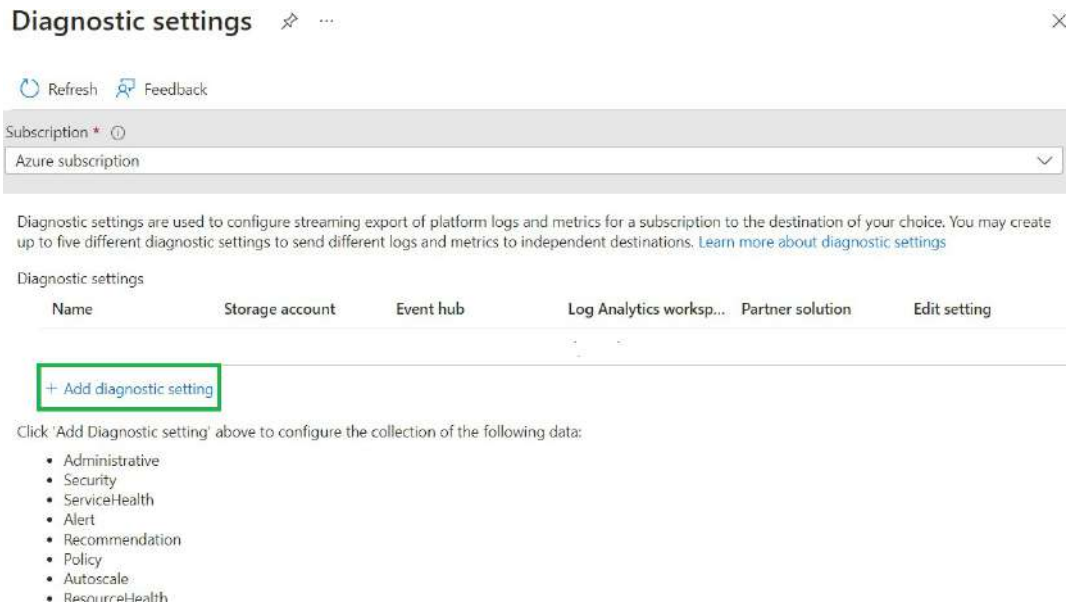

Figure 9.5 – Configuring logs in the Diagnostic settings screen

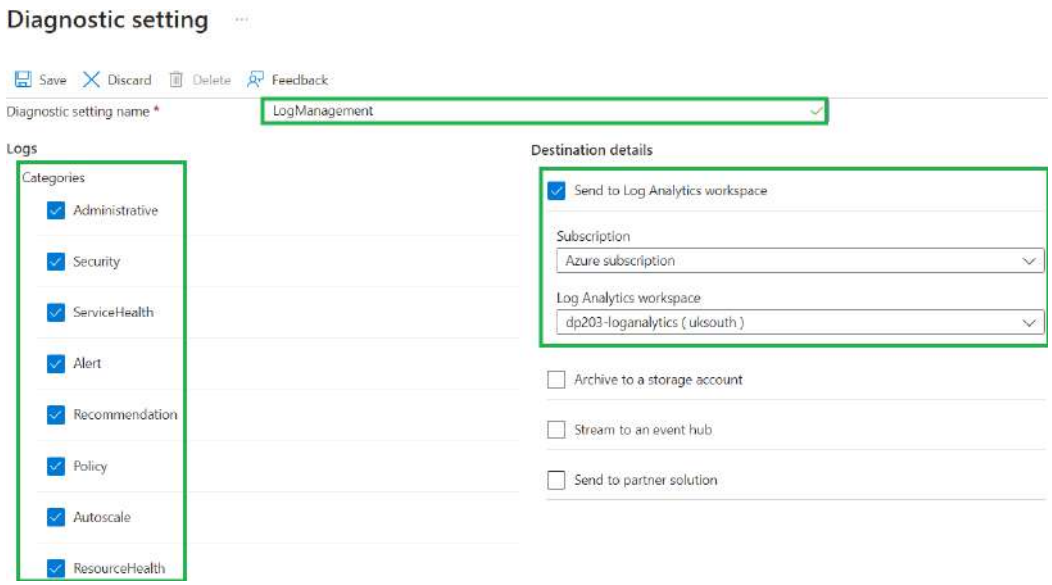

Figure 9.6 – Selecting a list of categories for the Log Analytics workspace in Diagnostic setting

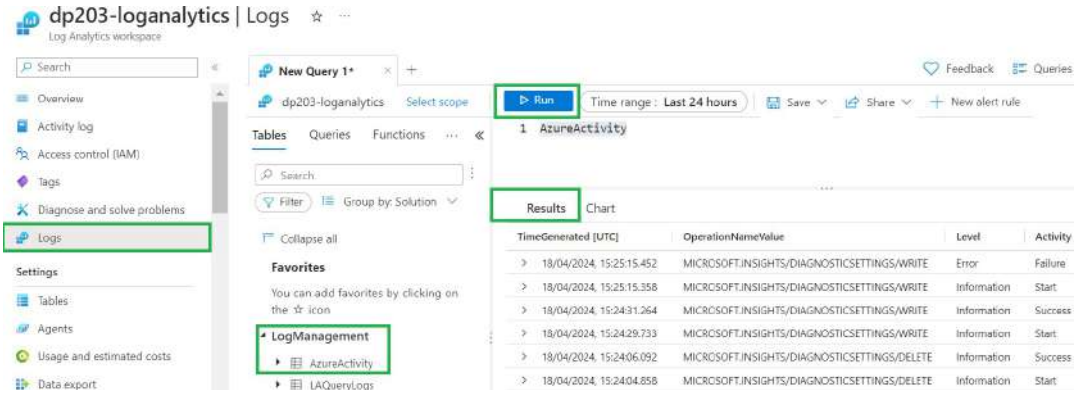

Figure 9.7 – The AzureActivity table displaying various log results in the Log Analytics workspace

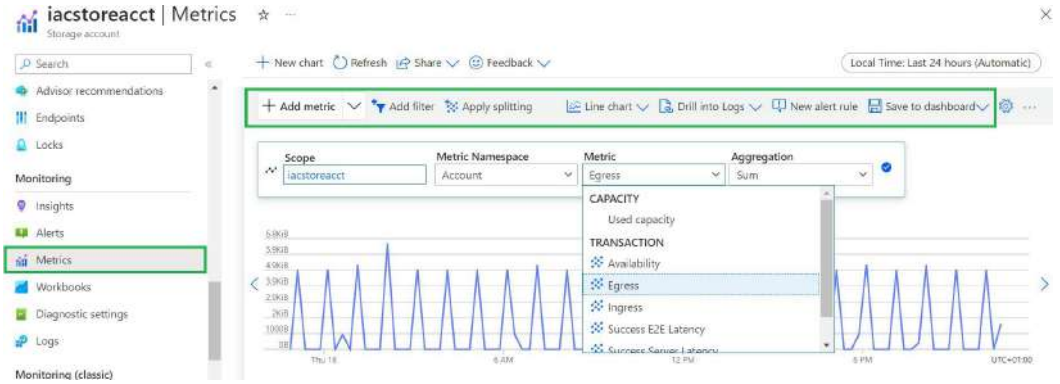

Figure 9.8 – Configuring metrics for Azure Storage

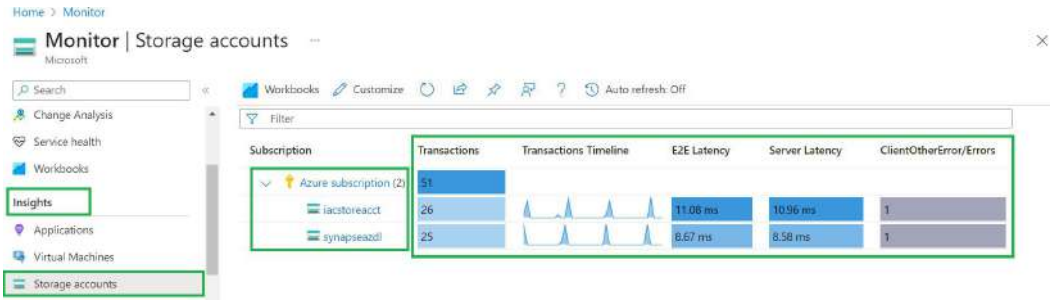

Figure 9.9 – The Monitor Storage account showing storage service metrics from the Azure Monitor service

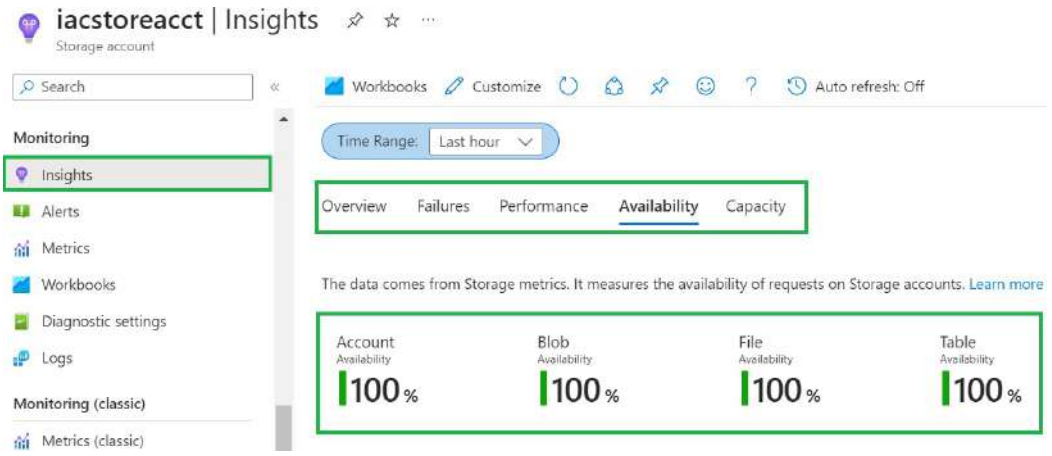

Figure 9.10 – Checking storage availability and status with Azure Monitor

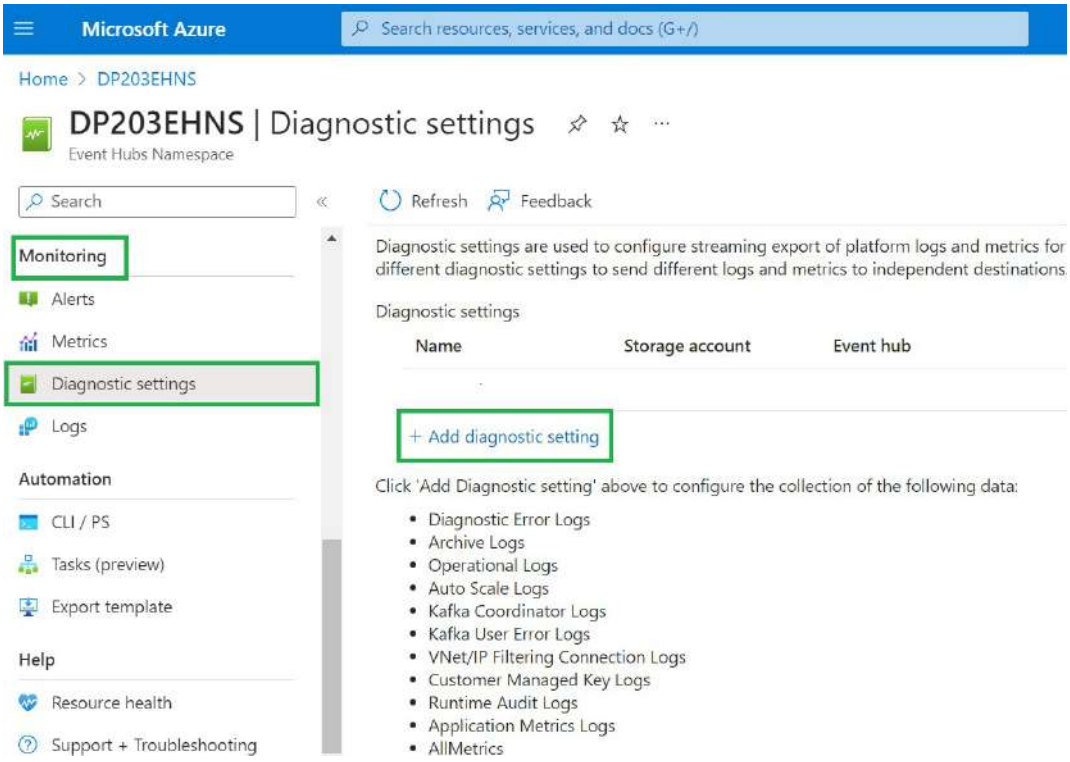

Figure 9.11 – Configuring new diagnostics for Event Hub

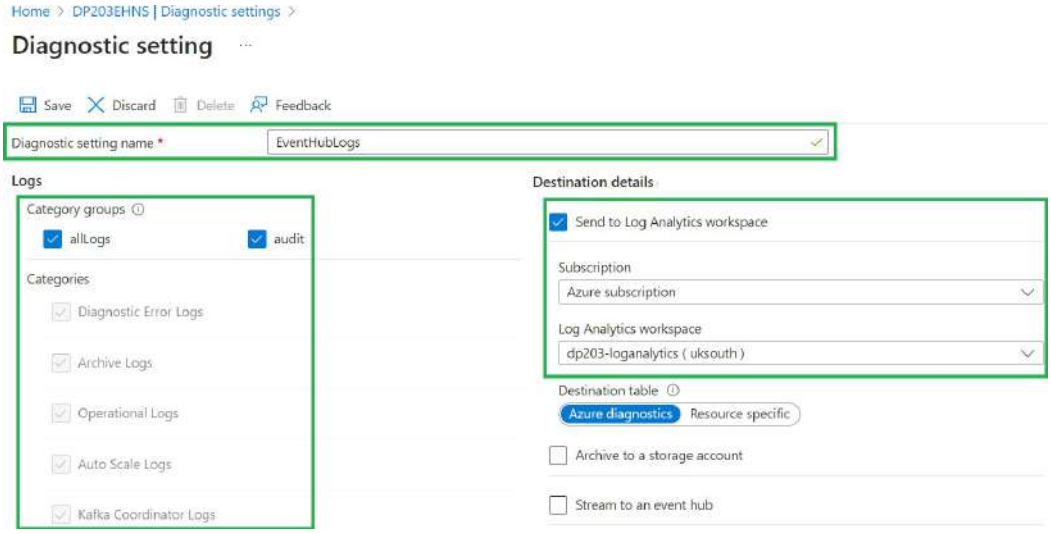

Figure 9.12 – Adding diagnostic settings for Event Hubs

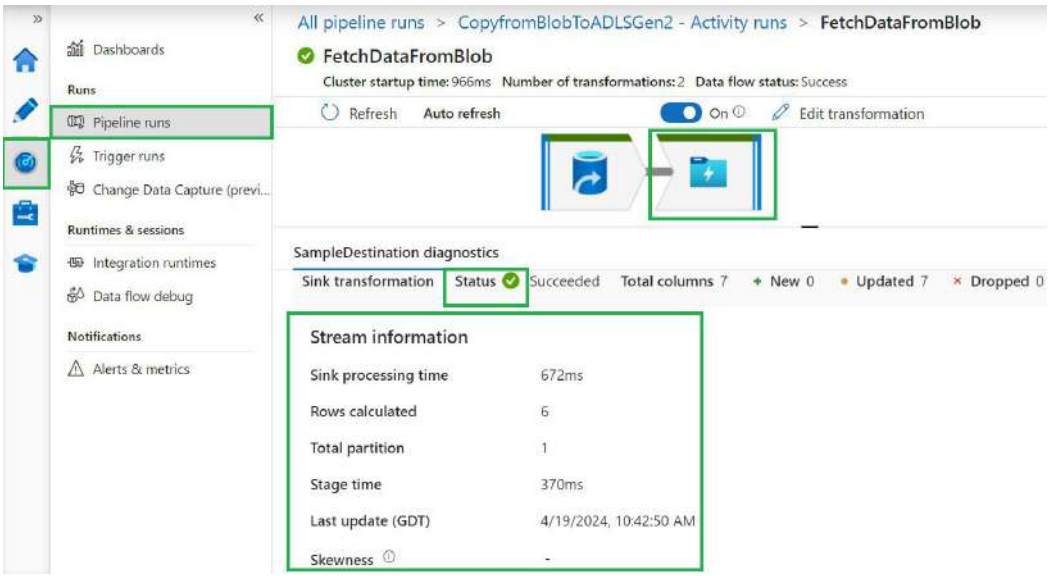

Figure 9.13 – Displaying data movement performance details

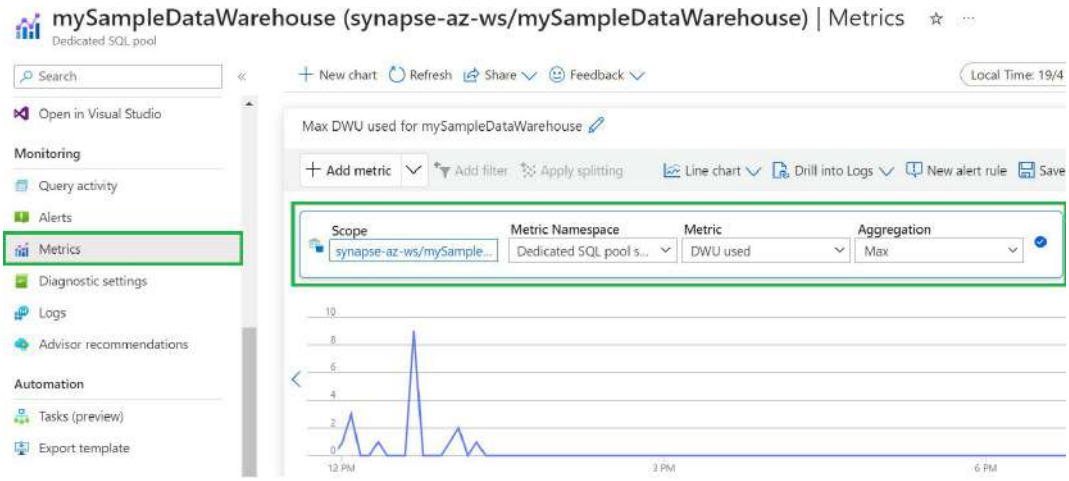

Figure 9.14 – Displaying metrics identifying performance regression in Synapse SQL pool

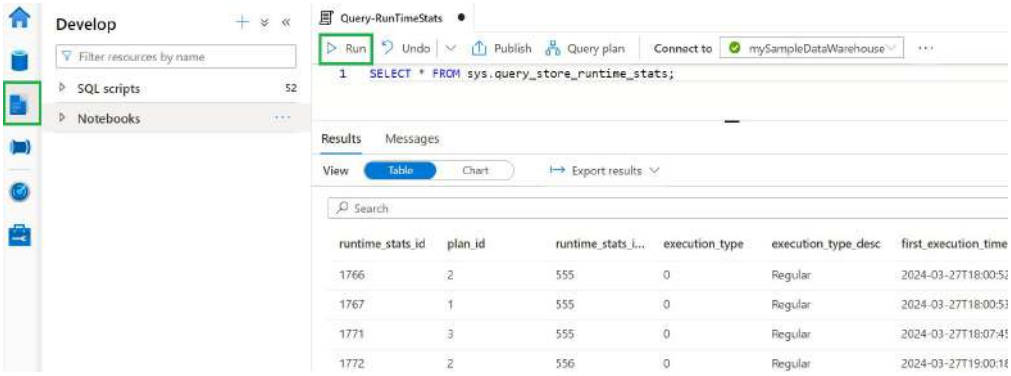

Figure 9.15 – Displaying a sample Query Store runtime statistics

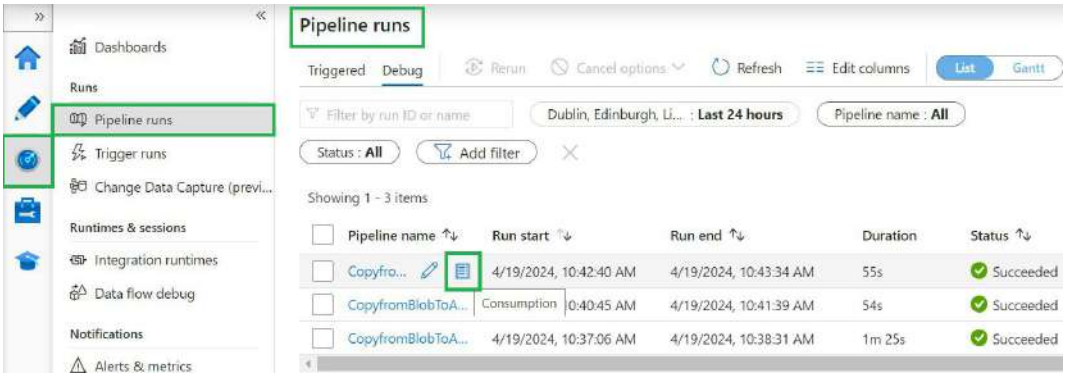

Figure 9.16 – The pipeiline name showing a Consumption icon to display the details

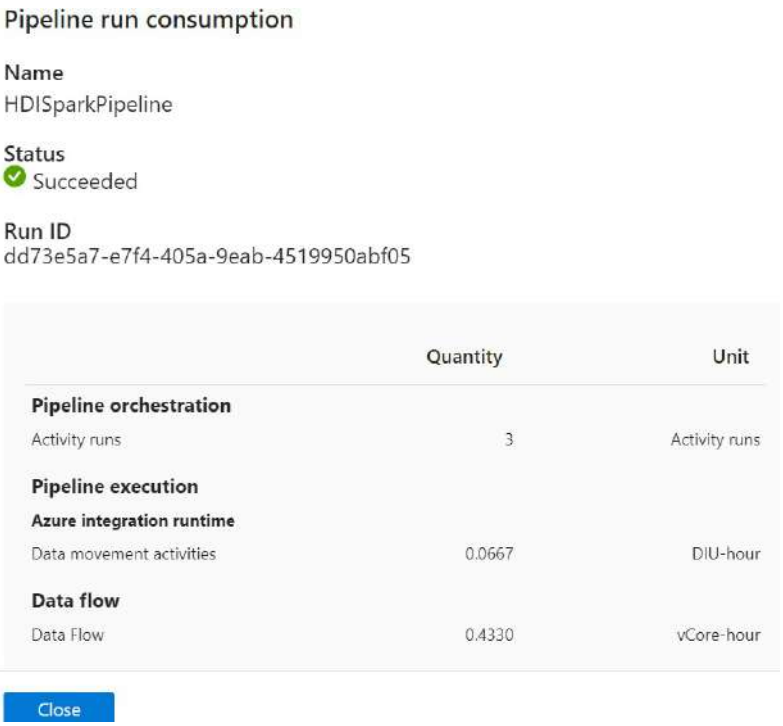

Figure 9.17 – Resource consumption details screen showing the status of the resources

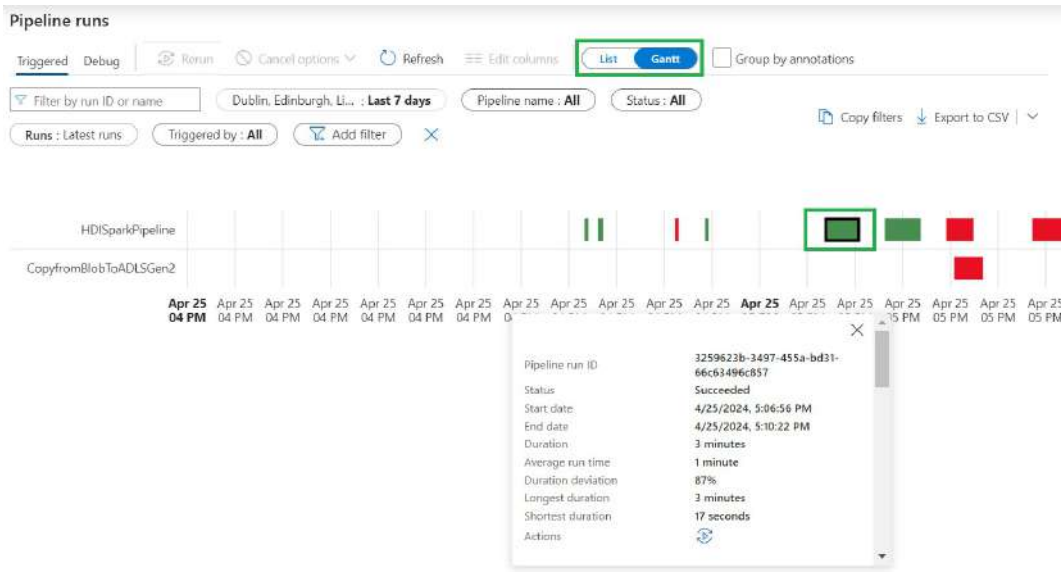

Figure 9.18 – Displaying additional pipeline details in the Gantt chart screen

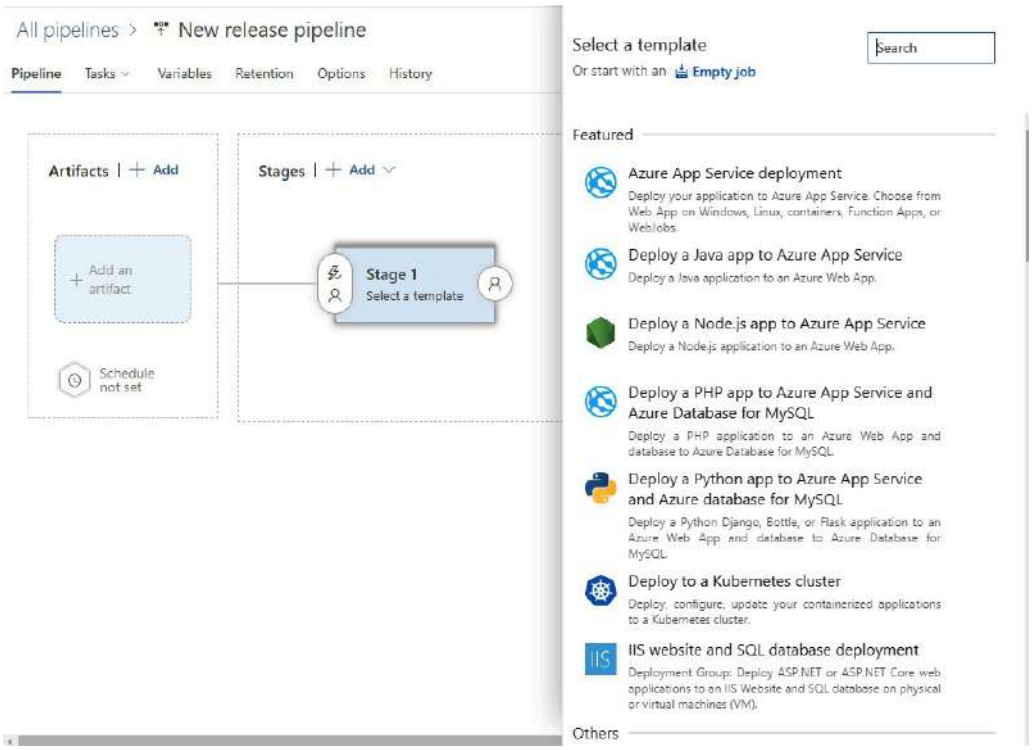

Figure 9.19 – The Azure release pipeline creation wizard creating an empty job

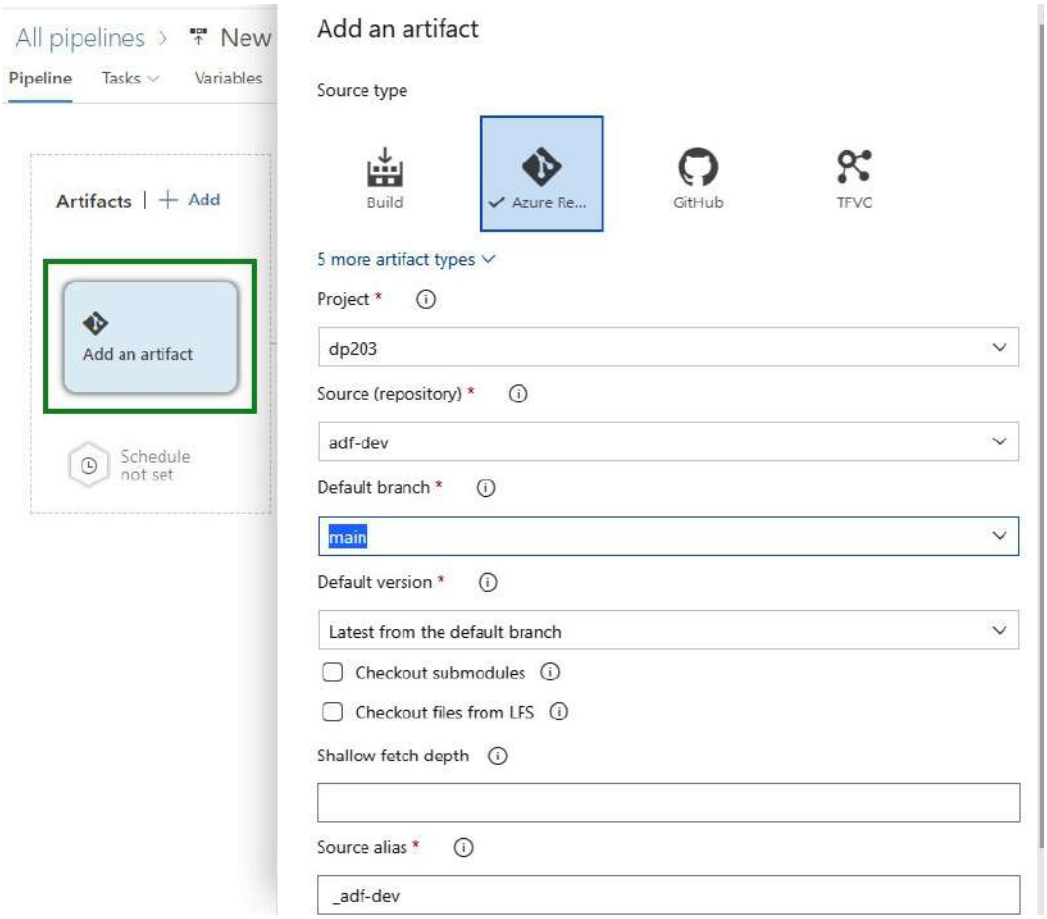

Figure 9.20 – Updating artifact information in Azure Pipelines

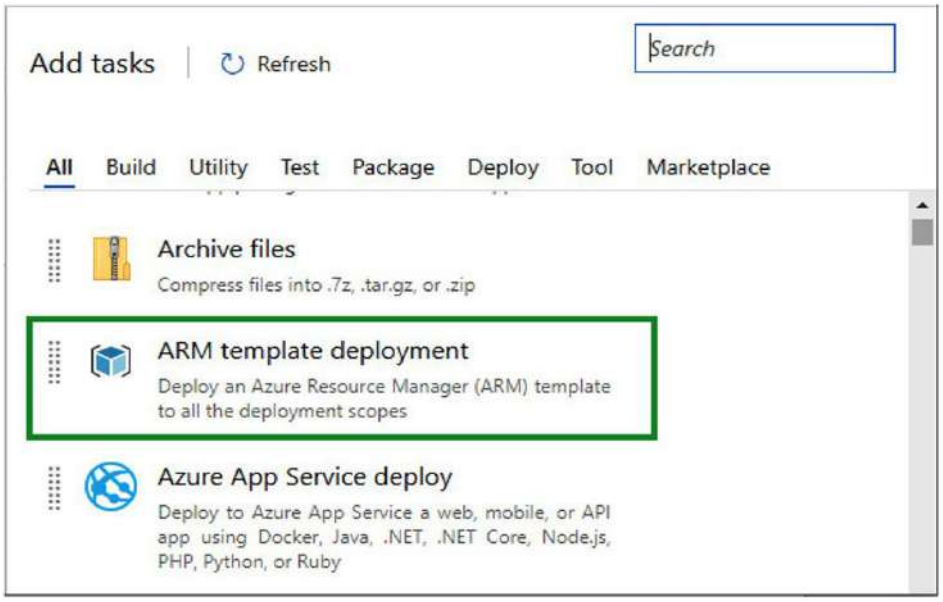

Figure 9.21 – Choosing ARM template deployment from the Add tasks screen

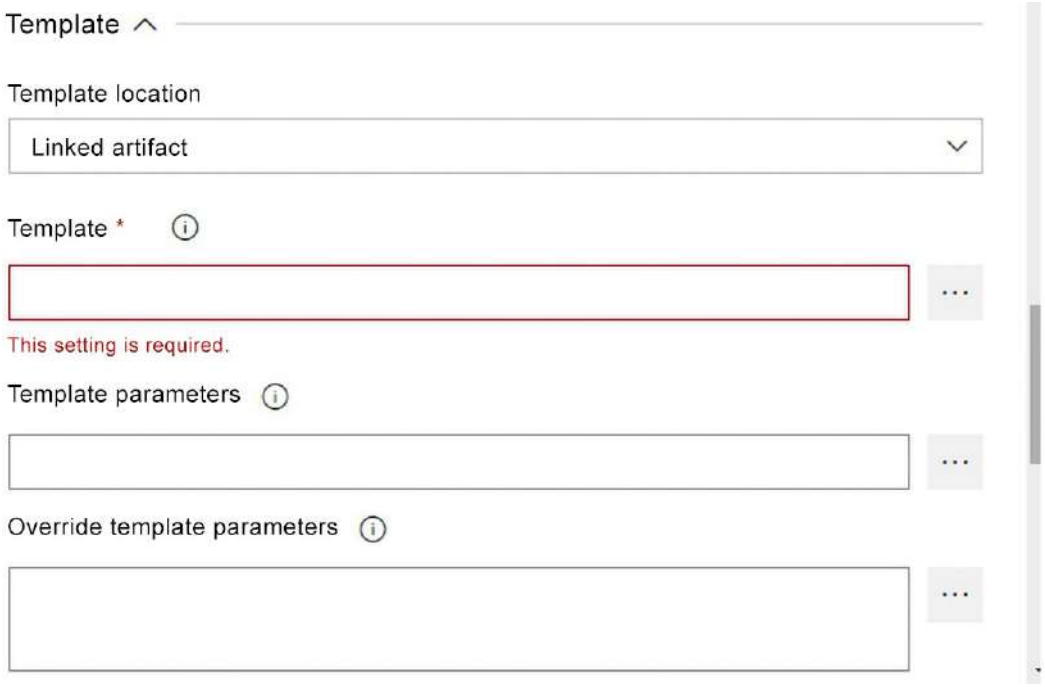

Figure 9.22 – Specifying the ARM template and template parameters

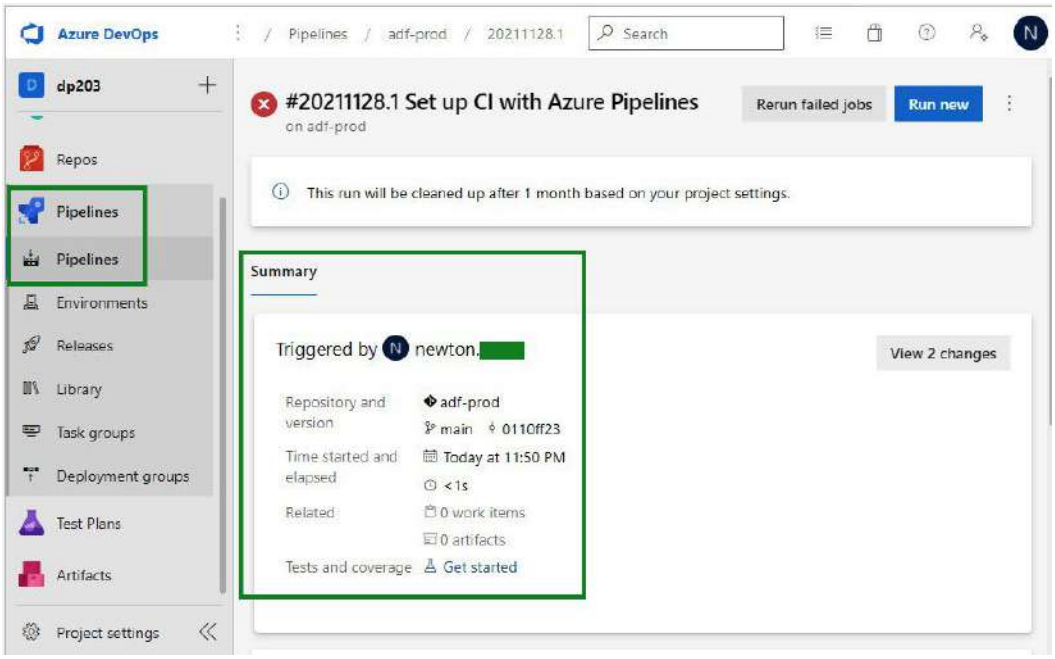

Figure 9.23 – The pipelines CI/CD monitoring screen showing the monitoring process

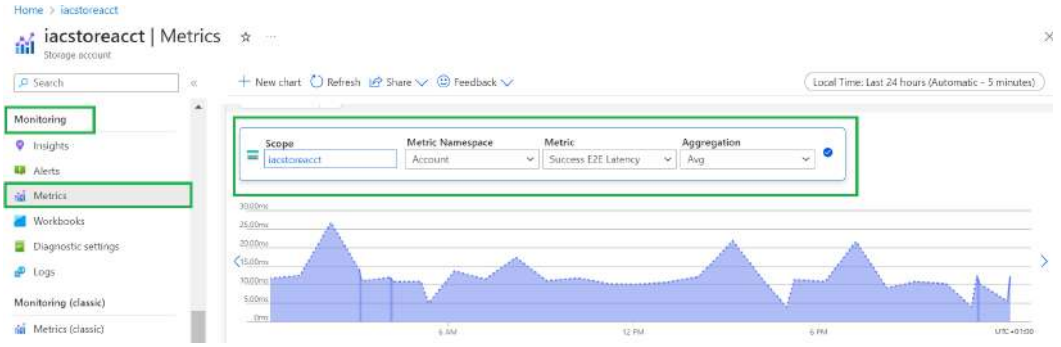

Figure 9.24 – Displaying Metrics data for a storage account collected from Azure resources

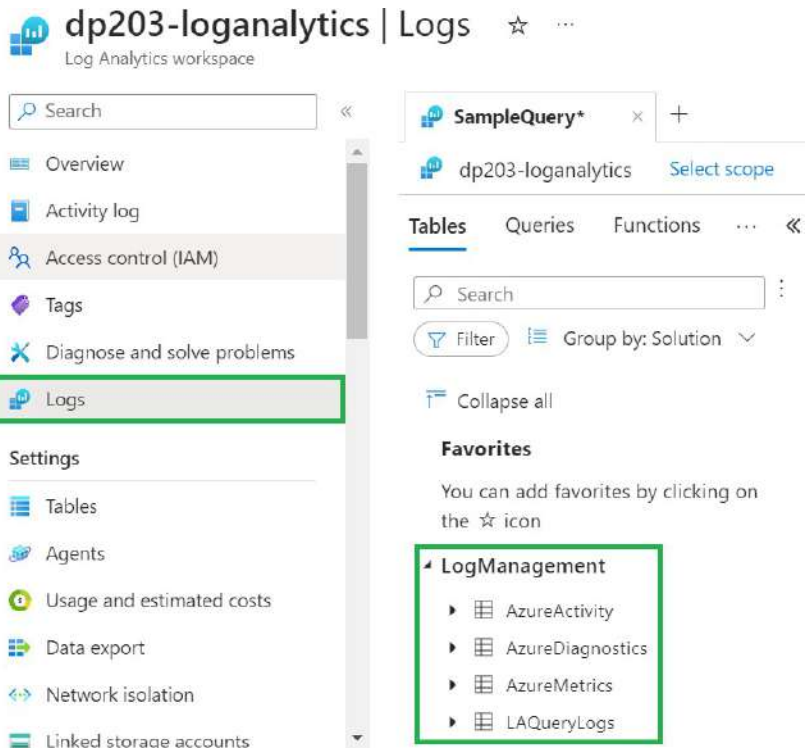

Figure 9.25 – Running the Log Analytics tables to view the different logs from each one

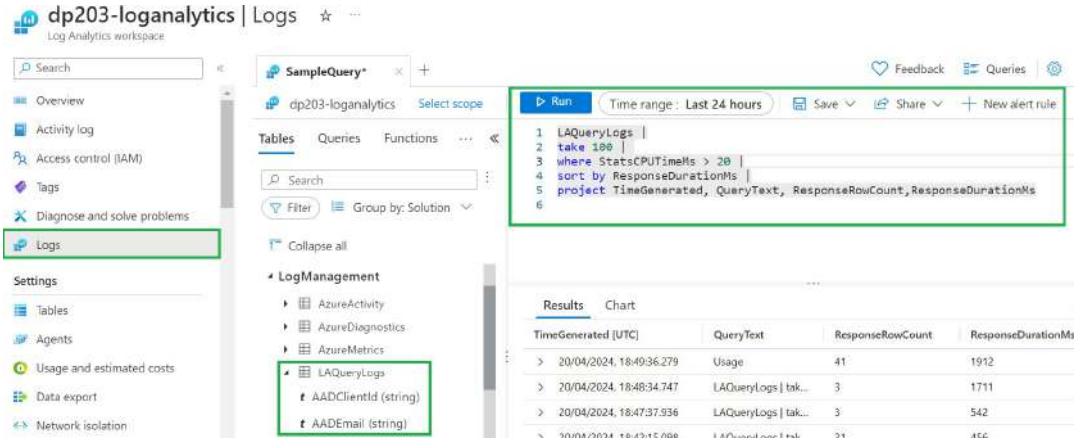

Figure 9.26 – Kusto query statements displaying the table contents

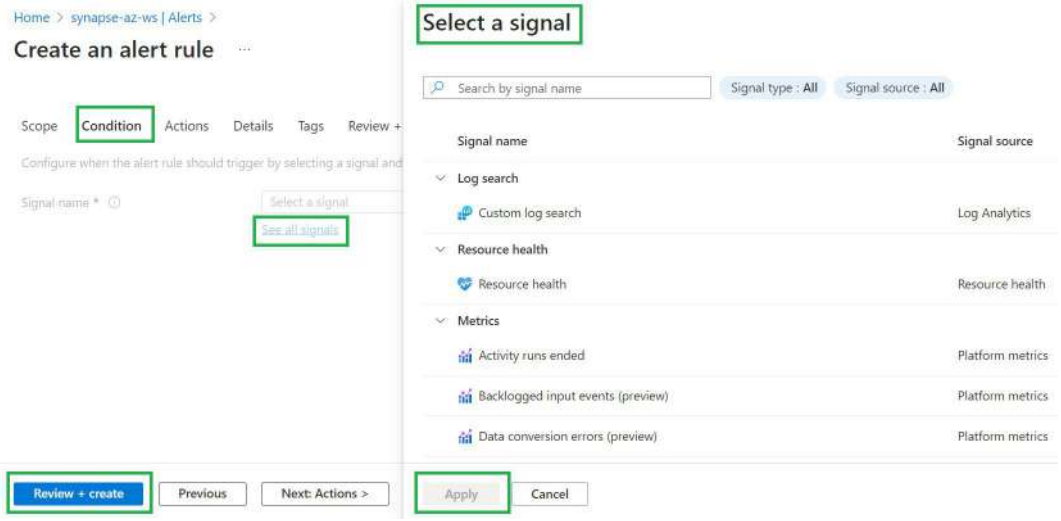

Figure 9.27 – Creating an alert rule for Synapse pipelines

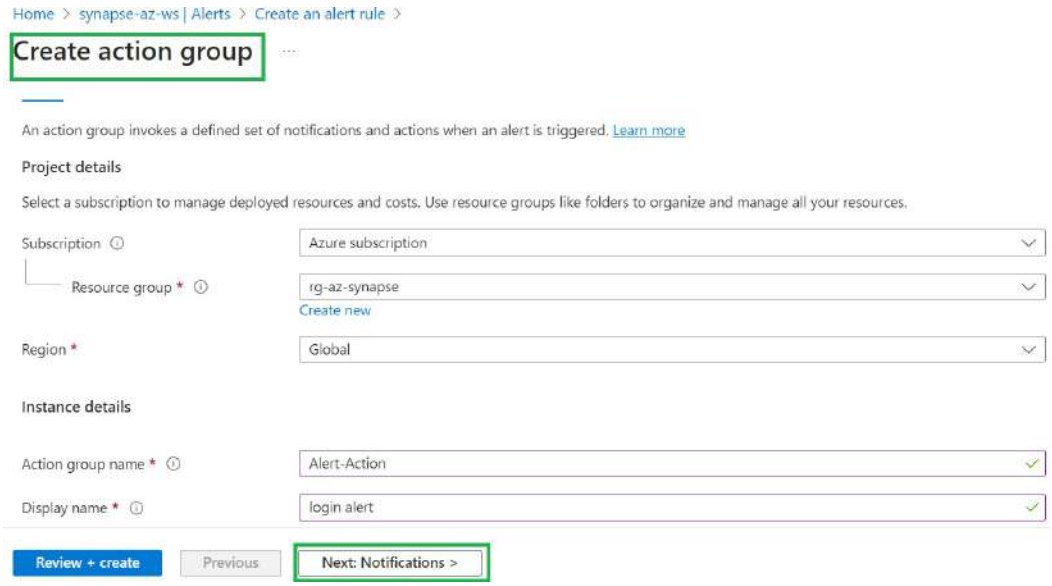

Figure 9.28 – Creating an action group for an alert rule in the Synapse workspace

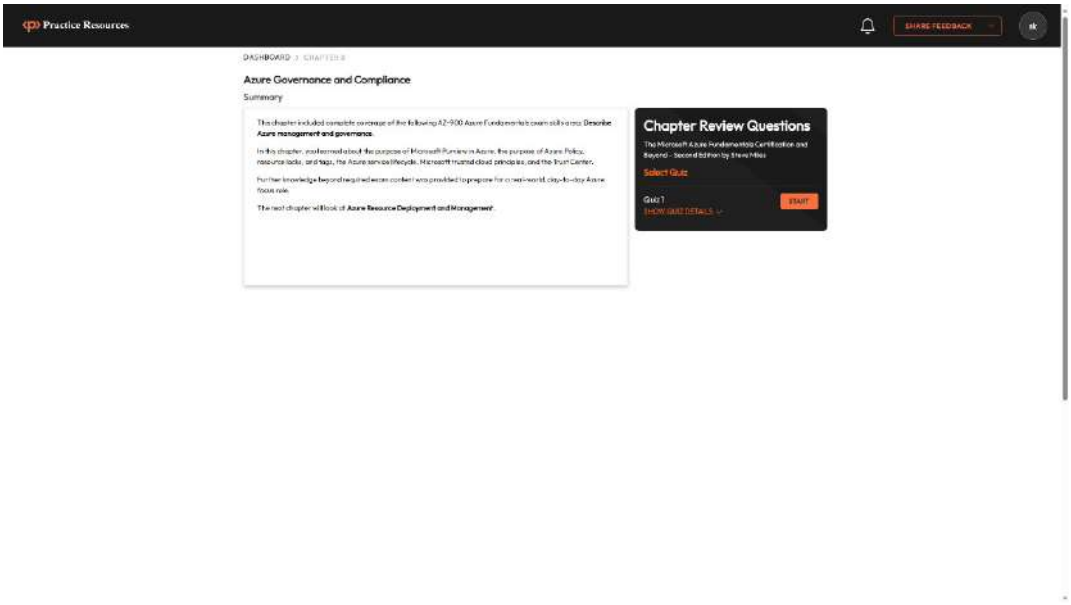

Figure 9.30 – Chapter Review Questions for Chapter 9

Chapter 10: Optimizing and Troubleshooting Data Storage and Processing

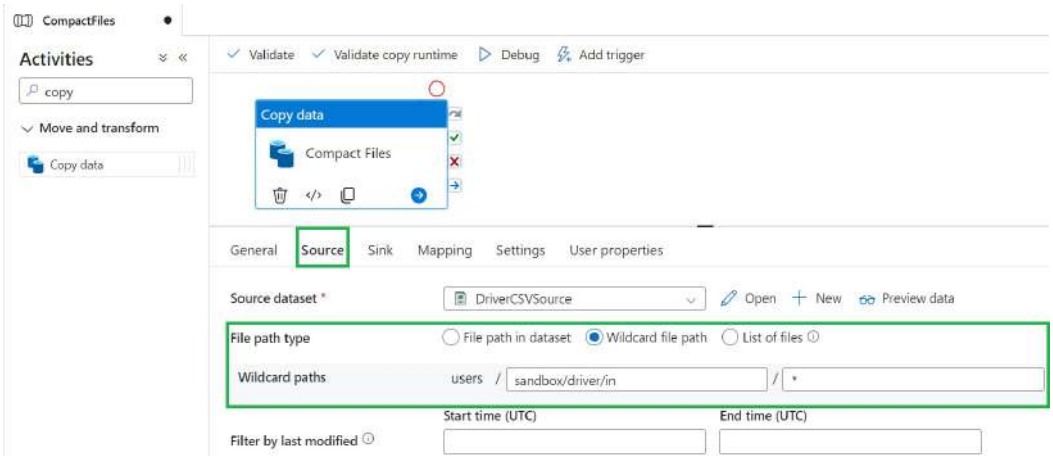

Figure 10.1 – Using wildcards to specify the source folder in Copy activity

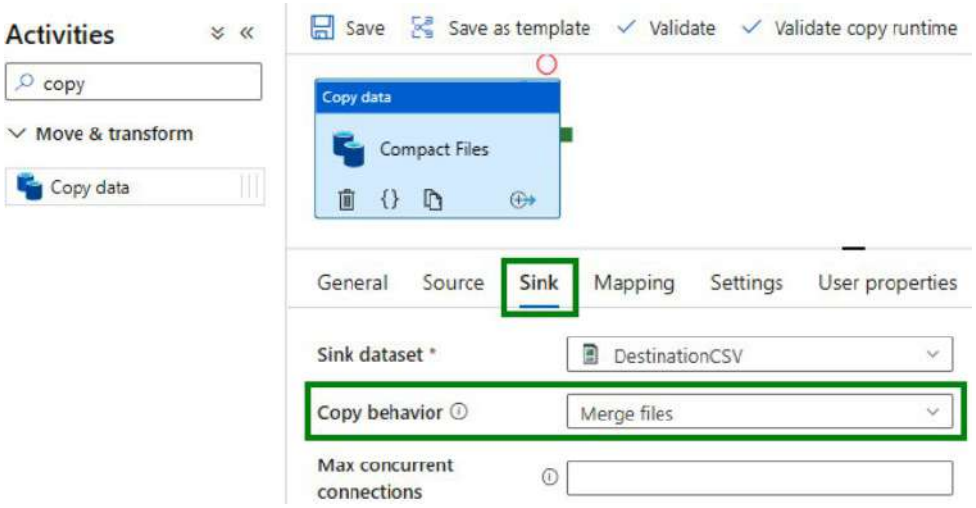

Figure 10.2 – The copying behavior of merging files in the Copy activity of ADF

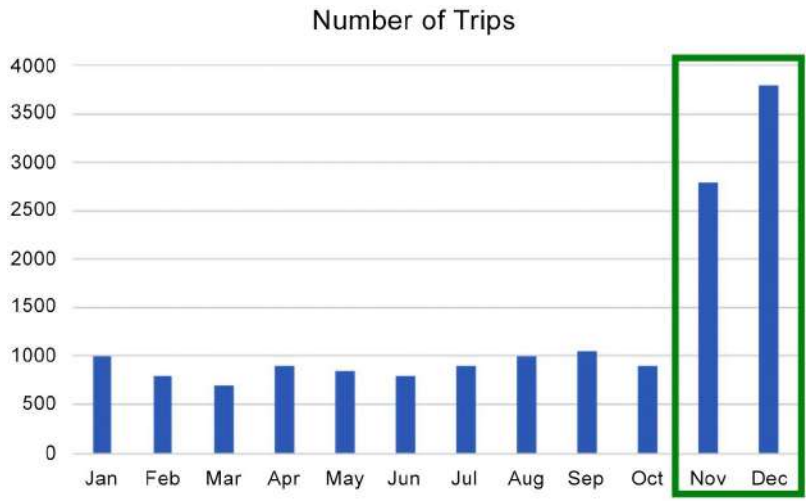

Figure 10.3 – Skewed data example showing the uneven distribution of trips per month

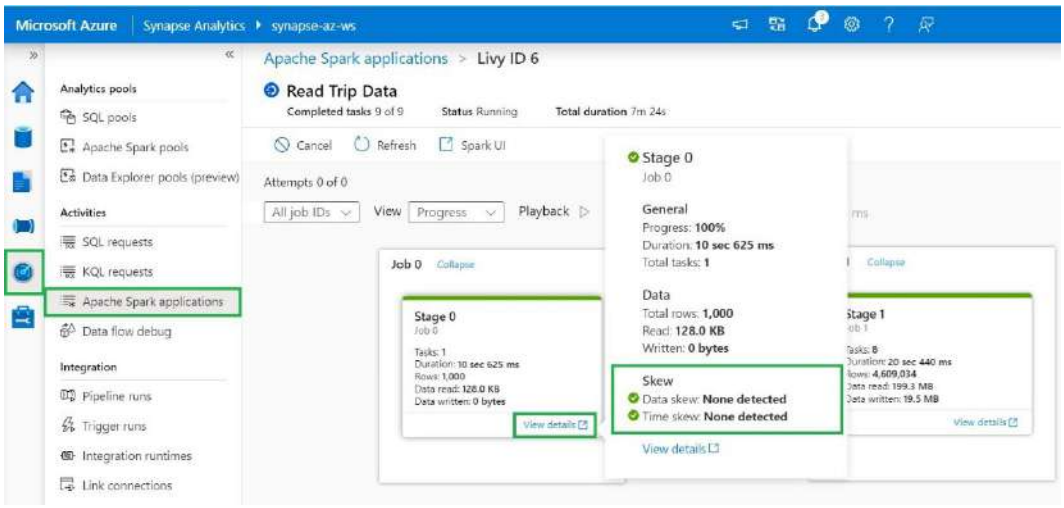

Figure 10.4 – Data skew at different stages in Synapse Spark

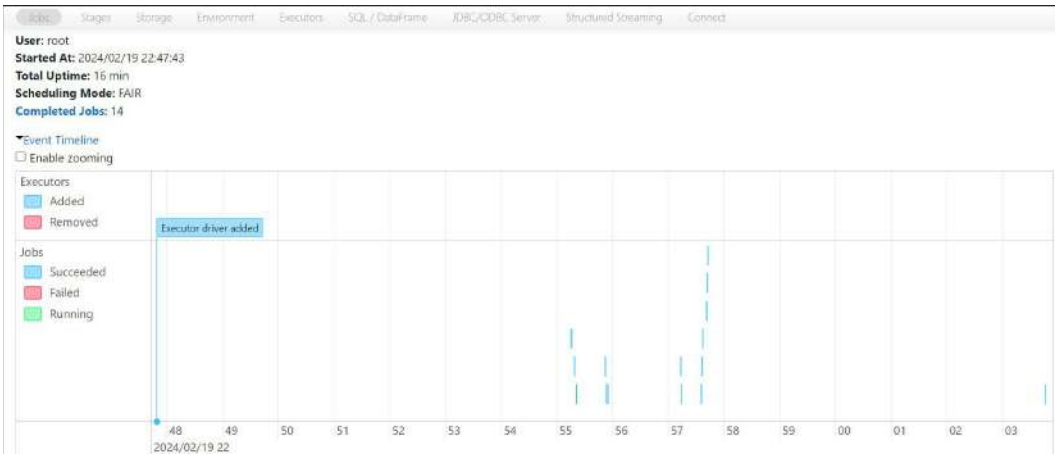

Figure 10.5 – Spark UI screen showing the status of the jobs

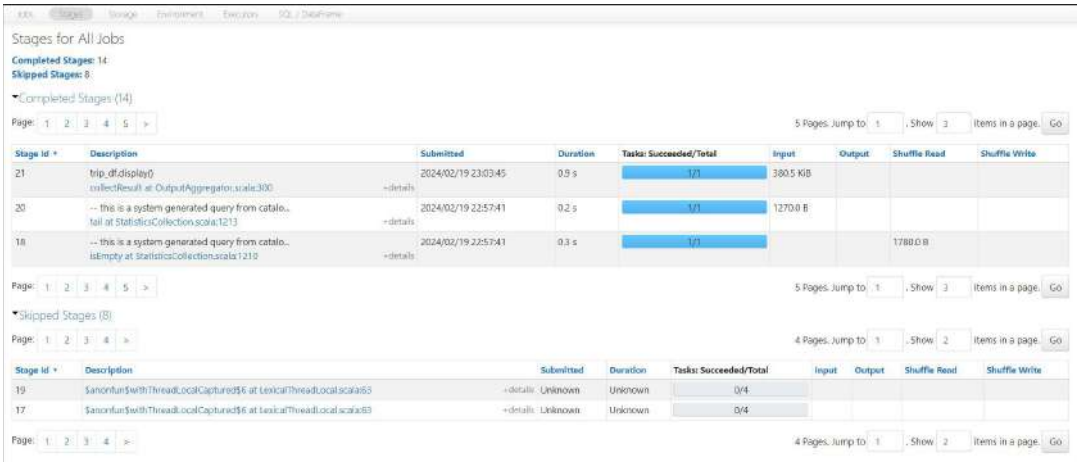

Figure 10.6 – Checking the stages and duration of all job stages

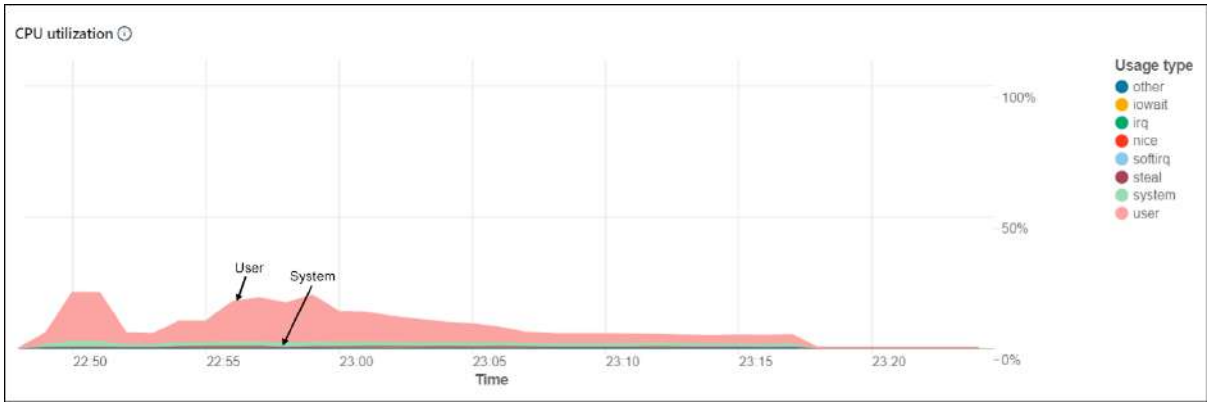

Figure 10.7 – The CPU utilization in cluster metrics

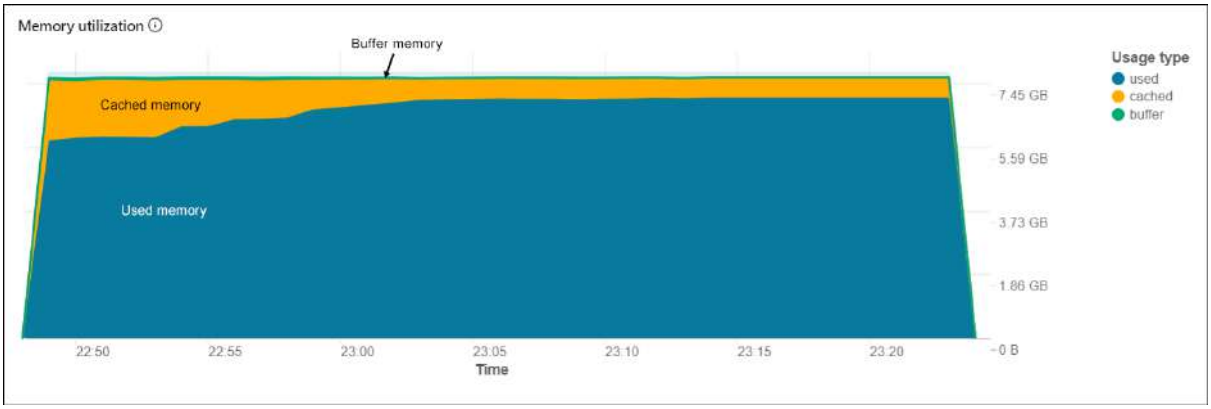

Figure 10.8 – Memory utilization in cluster metrics

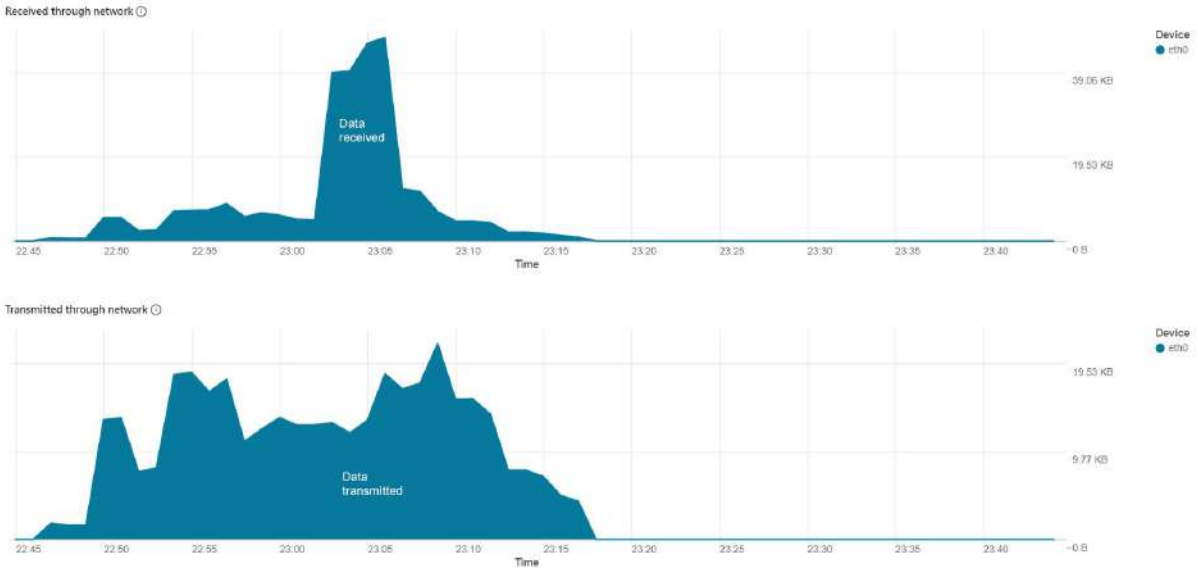

Figure 10.9 – Network usage in cluster metrics

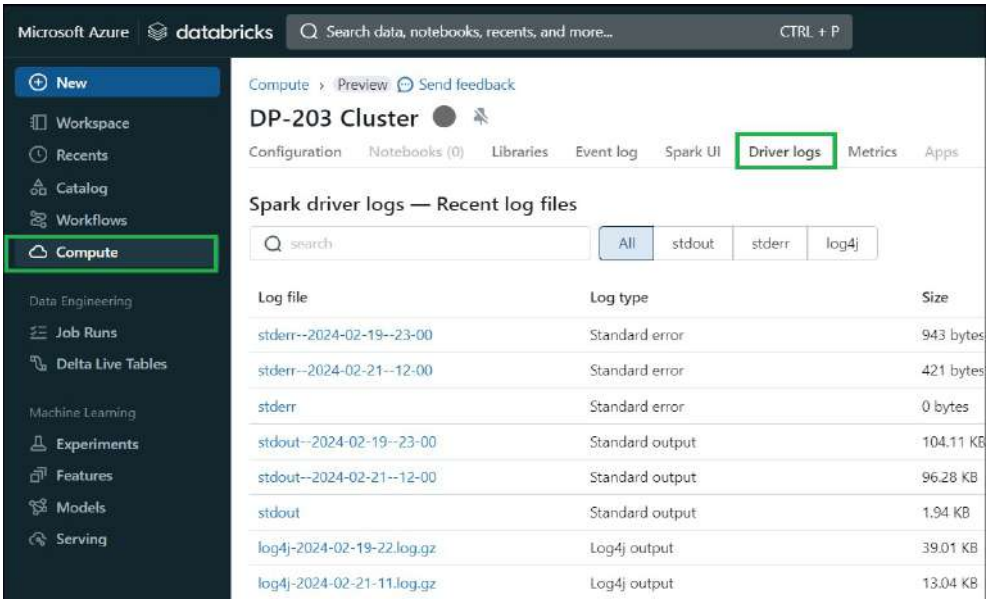

Figure 10.10 – Displaying Driver logs location in an Azure Databricks cluster

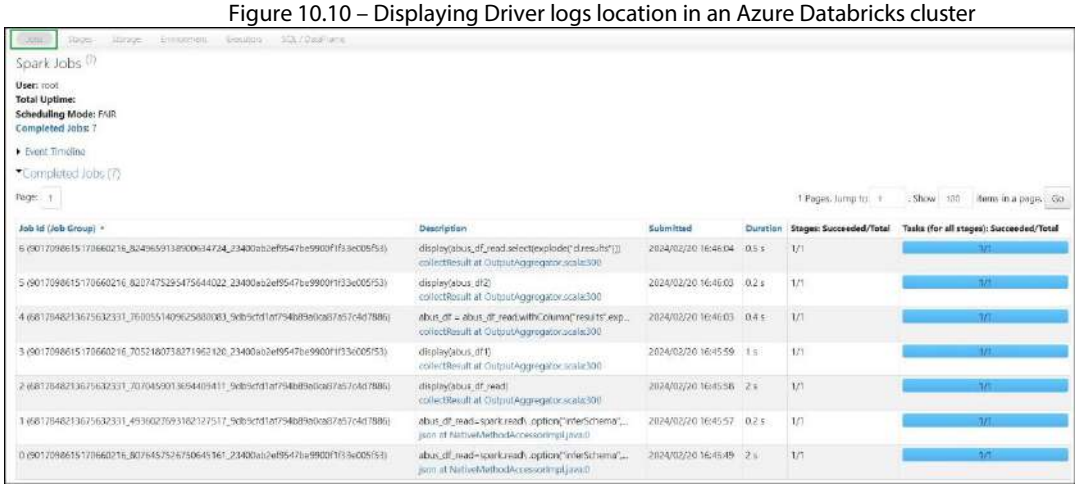

Figure 10.11 – Checking the status of a job in Spark job location within the Spark UI

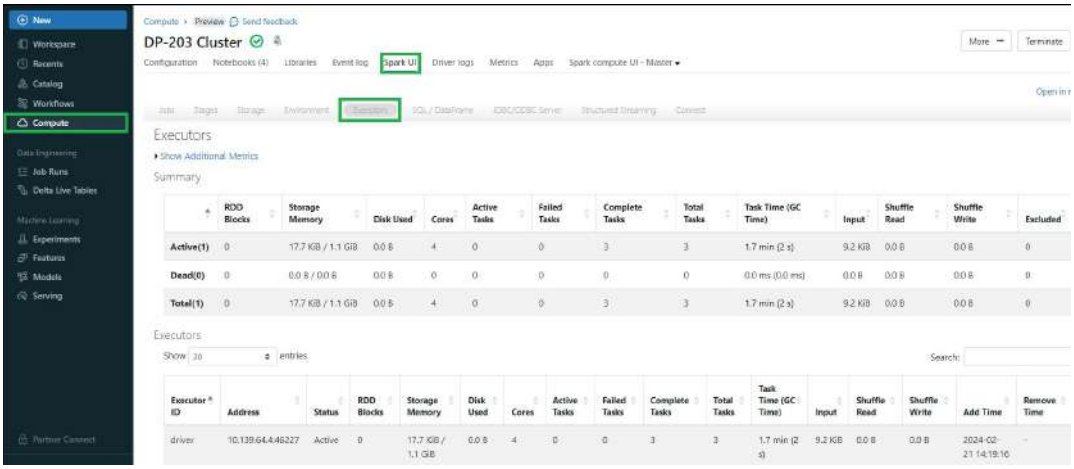

Figure 10.12 – Executor log location in the Spark UI

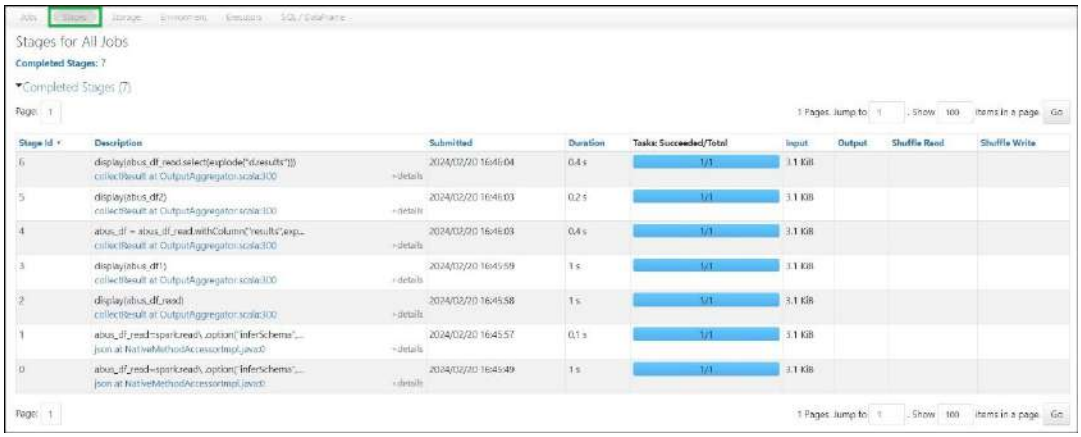

Figure 10.13 – Displaying task in each stage location in the Spark UI

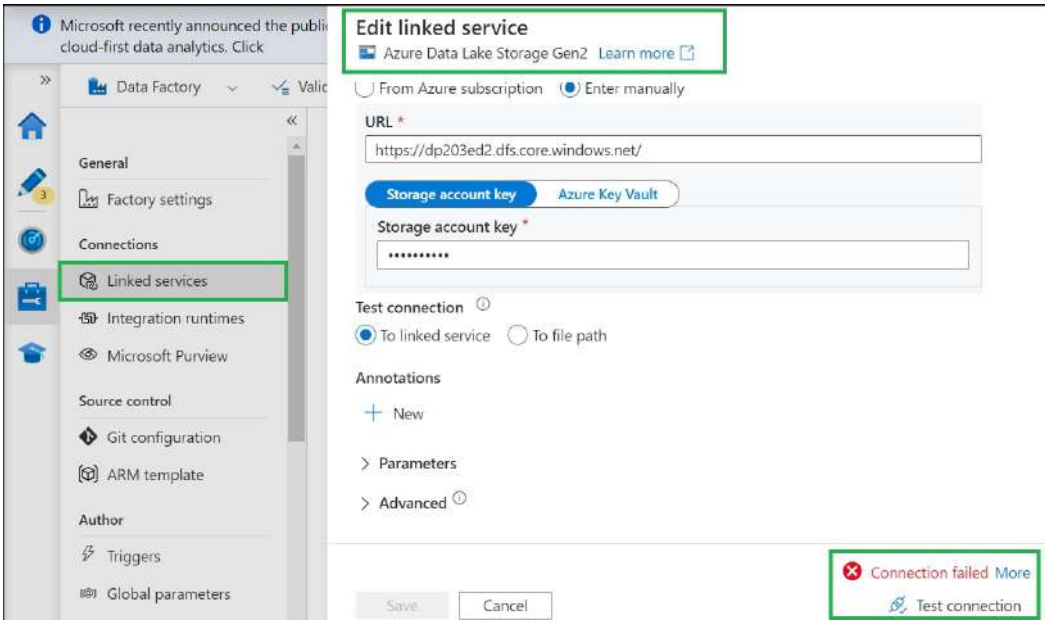

Figure 10.14 – Troubleshooting linked service connection

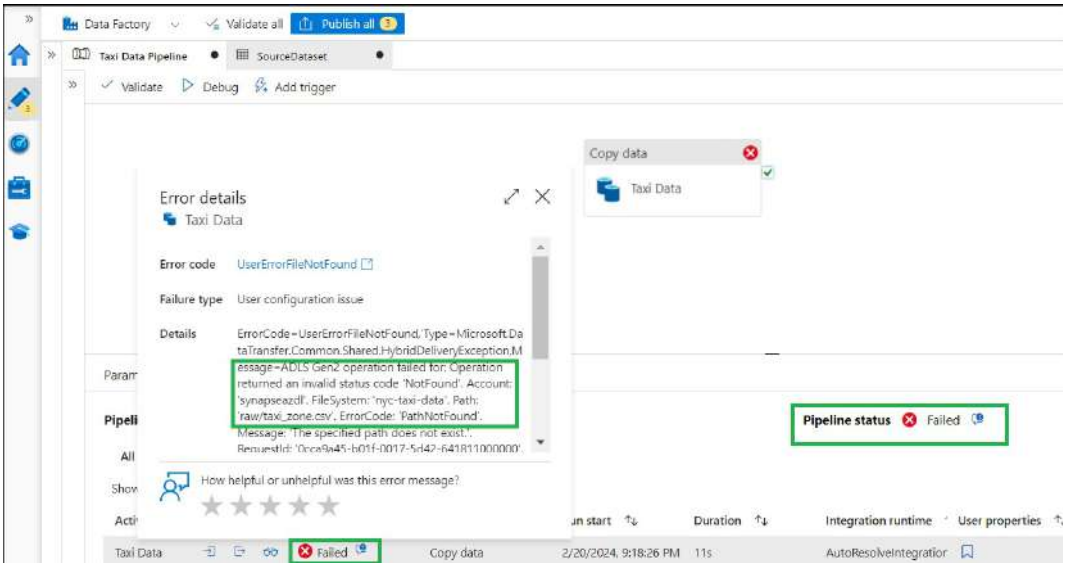

Figure 10.15 – Displaying an error when the Troubleshoot path is not found in the pipeline

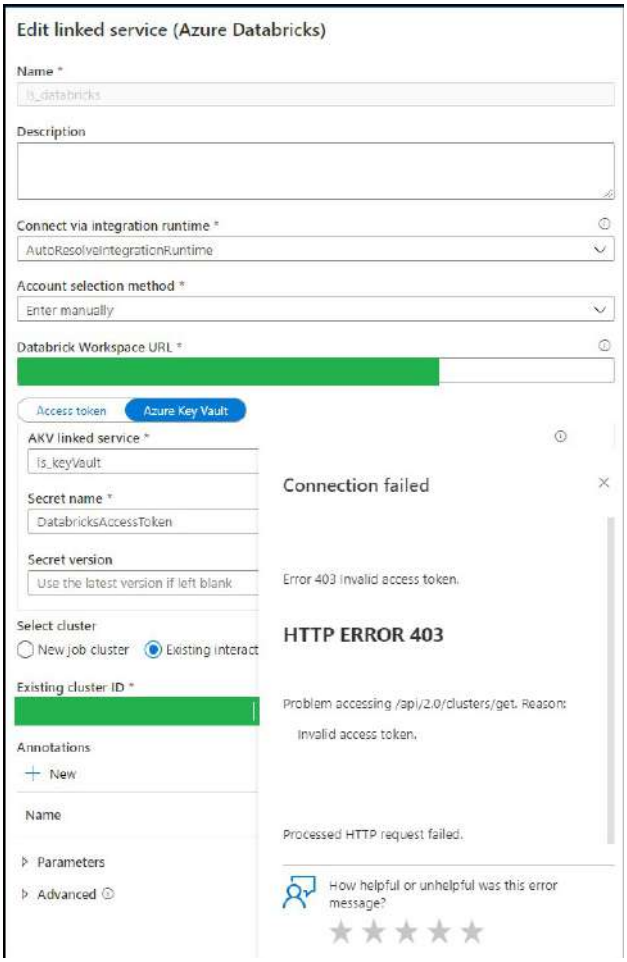

Figure 10.16 – Troubleshooting an access token issue for Azure Databricks service

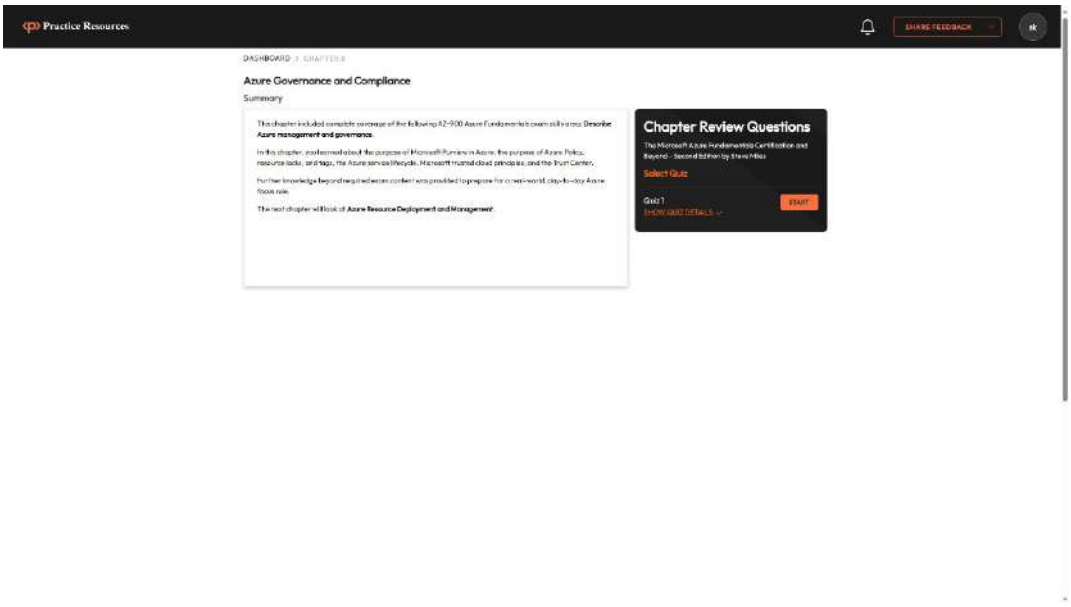

Figure 10.18 – Chapter Review Questions for Chapter 10

Chapter 11: Accessing the Online Practice Resources

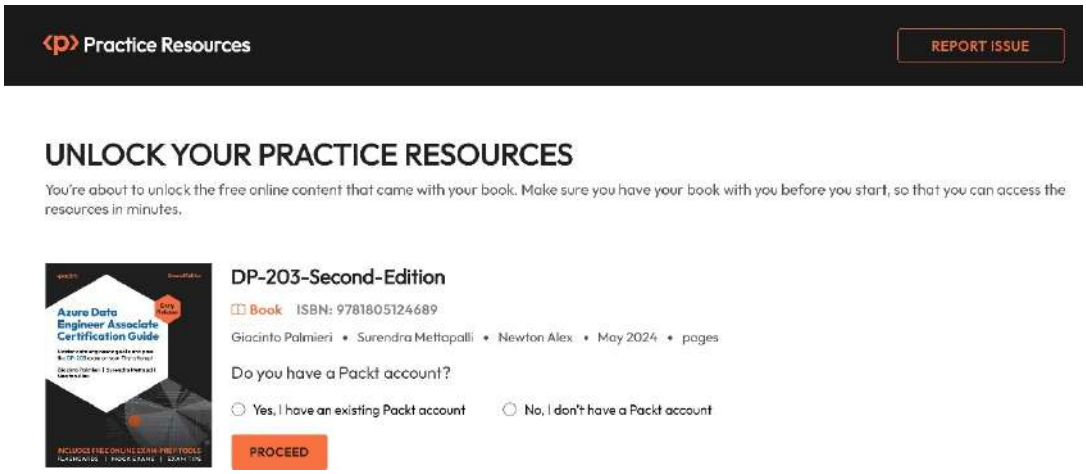

Figure 11.2 – Unlock page for the online practice resources

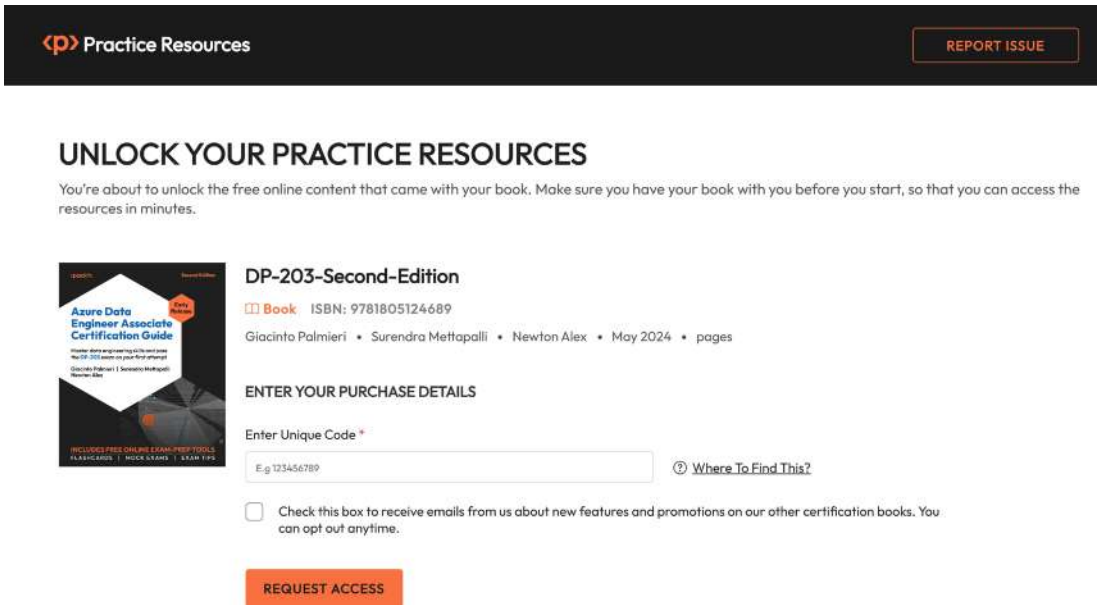

Figure 11.3 – Enter your unique sign-up code to unlock the resources

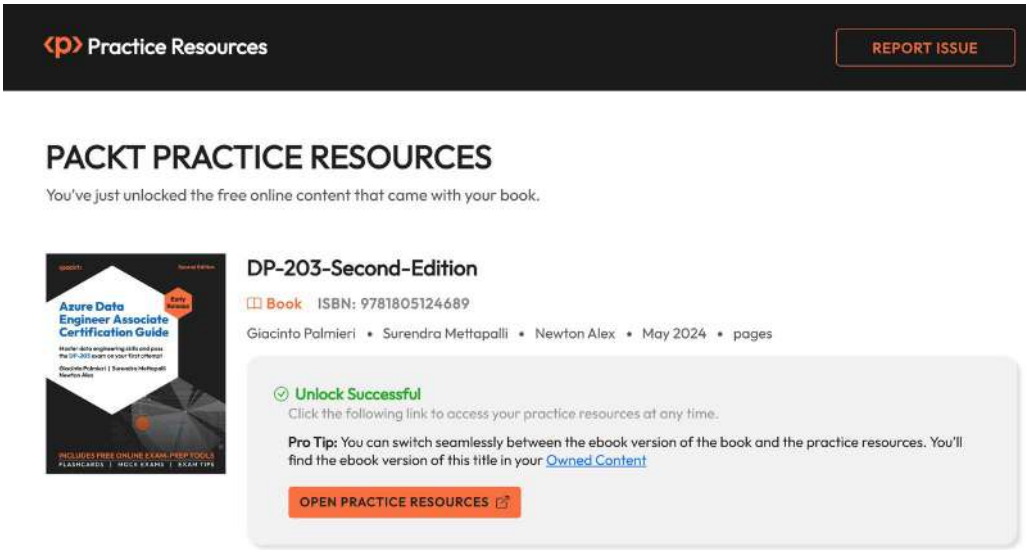

Figure 11.4 – Page that shows up after a successful unlock

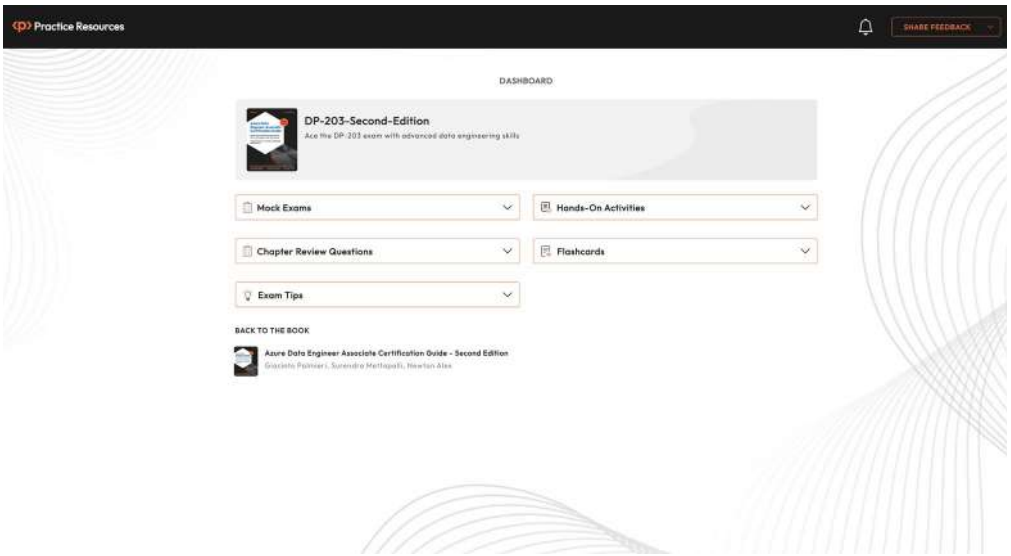

Figure 11.5 – Dashboard page for AZ-900 practice resources

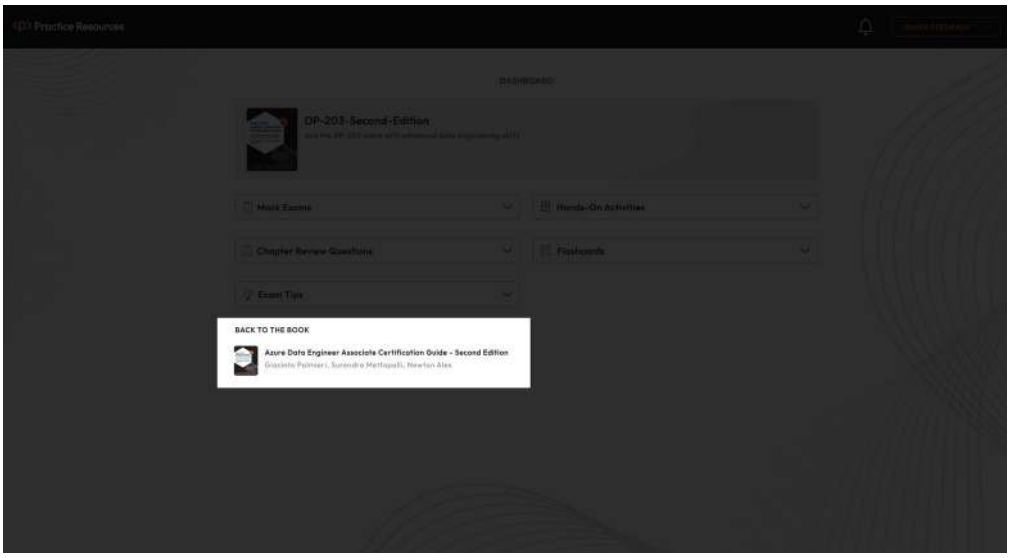

Figure 11.7 – Jump back to the book from the dashboard