Azure Data Factory - how to provision, access and development guide

Azure Data Factory is a cloud-based data integration service offered by Microsoft Azure, designed to empower organizations with the ability to collect, transform, and move data from diverse sources to various destinations. As a fully managed and scalable service, Azure Data Factory facilitates the creation, scheduling, and management of data pipelines, enabling businesses to gain valuable insights and make informed decisions based on their data.

Key Features and Uses:

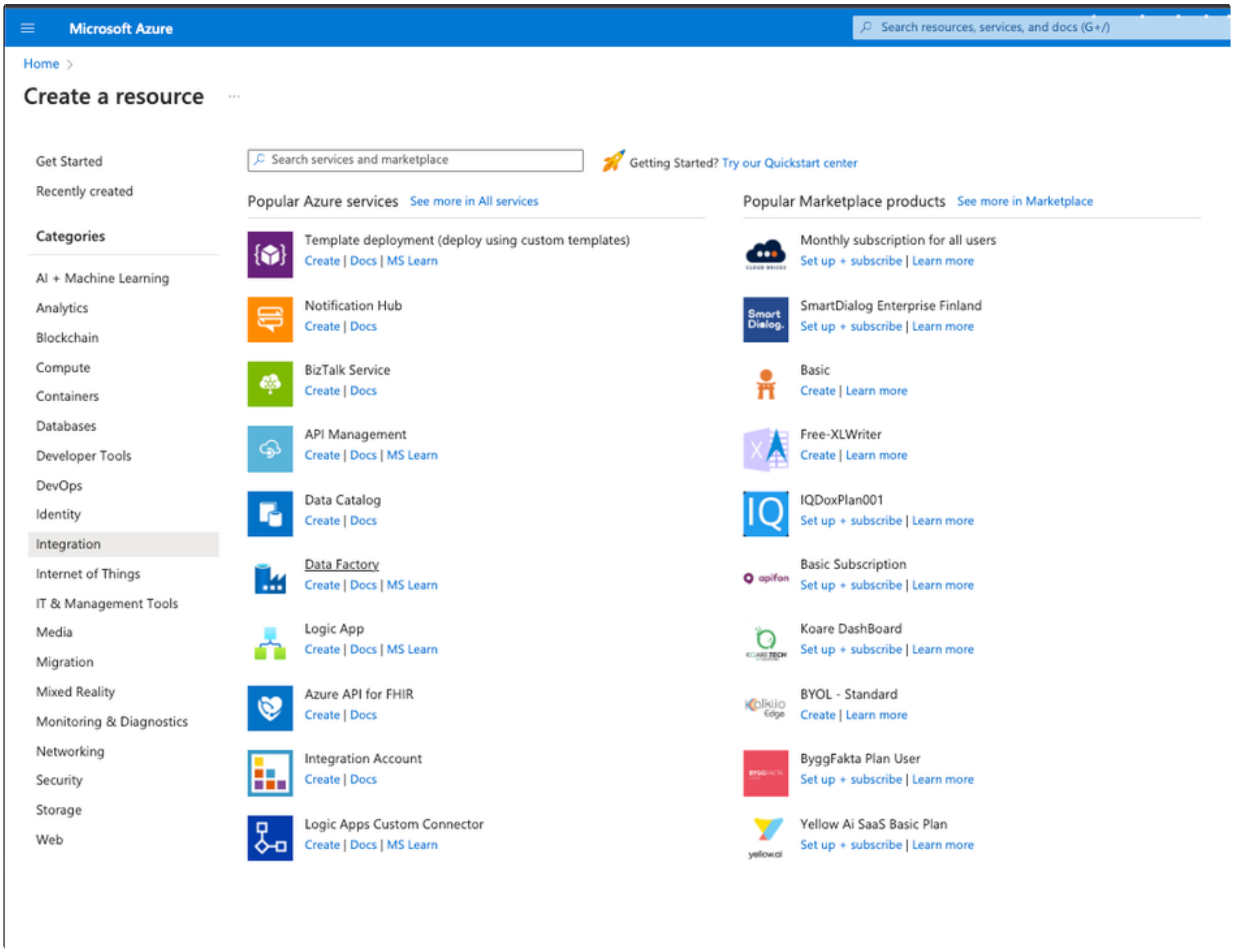
- **Data Orchestration:** Azure Data Factory allows you to orchestrate and automate the movement and transformation of data across on-premises and cloud environments. This facilitates the creation of end-to-end data workflows.
- **Data Transformation:** With built-in data wrangling capabilities, Azure Data Factory supports the transformation of raw data into a usable format for analytics and reporting. This includes cleaning, enriching, and aggregating data as it moves through the pipeline.
- **Connectivity:** Seamlessly connect to a variety of on-premises and cloud-based data sources, such as Azure SQL Database, Azure Blob Storage, Azure Data Lake Storage, and more. This enables comprehensive integration with diverse data ecosystems.
- **Monitoring and Management:** Azure Data Factory provides robust monitoring and management tools, allowing users to track the performance of data pipelines, identify issues, and optimize data integration processes for enhanced efficiency.
- **Hybrid Data Movement:** Support for hybrid data movement ensures that organizations can integrate and synchronize data across both on-premises and cloud environments, providing flexibility in data management strategies.
- **Data Integration with Machine Learning:** Integrate machine learning models into data pipelines for advanced analytics and predictions. Azure Data Factory enables the seamless incorporation of machine learning capabilities to enhance data-driven decision-making.

Benefits:

- **Scalability:** Azure Data Factory offers a serverless and fully managed platform that scales dynamically based on the demands of your data integration processes. This ensures optimal performance and resource utilization without the need for manual intervention.
- **Cost Efficiency:** Pay only for the resources consumed during data integration processes, eliminating the need for upfront investments in infrastructure. The pay-as-you-go model allows organizations to manage costs efficiently and scale as needed.
- **Time Efficiency:** Streamline the development and deployment of data pipelines with a user-friendly visual interface and pre-built connectors. This accelerates the time-to-market for data-driven solutions and reduces the development lifecycle.
- **Data Governance and Security:** Azure Data Factory incorporates robust security measures, including encryption, identity and access management, and compliance certifications. This ensures that sensitive data is protected, and organizations can adhere to regulatory requirements.
- **Business Intelligence and Analytics:** Empower business users with timely and accurate data for analytics and reporting. Azure Data Factory enables organizations to leverage their data assets, driving actionable insights that contribute to informed decision-making and business intelligence.

In order to create your own Azure Data Factory , please follow the steps as mentioned below -

1. Go to Home on the Azure Portal and then search for “Create a Resource”. Once you are in that page, click on “Integration” on the left hand pane. A page similar to what is shown below will come up. From here, select “Data Factory”.



Microsoft Azure

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

Home >

Create a resource ...

Get Started

Recently created

Categories

AI + Machine Learning
Analytics
Blockchain
Compute
Containers
Databases
Developer Tools
DevOps
Identity
Integration
Internet of Things
IT & Management Tools
Media
Migration
Mixed Reality
Monitoring & Diagnostics
Networking
Security
Storage
Web

Search services and marketplace

Getting Started? Try our Quickstart center

Popular Azure services See more in All services

Template deployment (deploy using custom templates)
Create | Docs | MS Learn

Notification Hub
Create | Docs

BizTalk Service
Create | Docs

API Management
Create | Docs | MS Learn

Data Catalog
Create | Docs

Data Factory
Create | Docs | MS Learn

Logic App
Create | Docs | MS Learn

Azure API for FHIR
Create | Docs

Integration Account
Create | Docs

Logic Apps Custom Connector
Create | Docs | MS Learn

Popular Marketplace products See more in Marketplace

Monthly subscription for all users
Set up + subscribe | Learn more

SmartDialog Enterprise Finland
Set up + subscribe | Learn more

Basic
Create | Learn more

Free-XLWriter
Create | Learn more

IQDoxPlan001
Set up + subscribe | Learn more

Basic Subscription
Set up + subscribe | Learn more

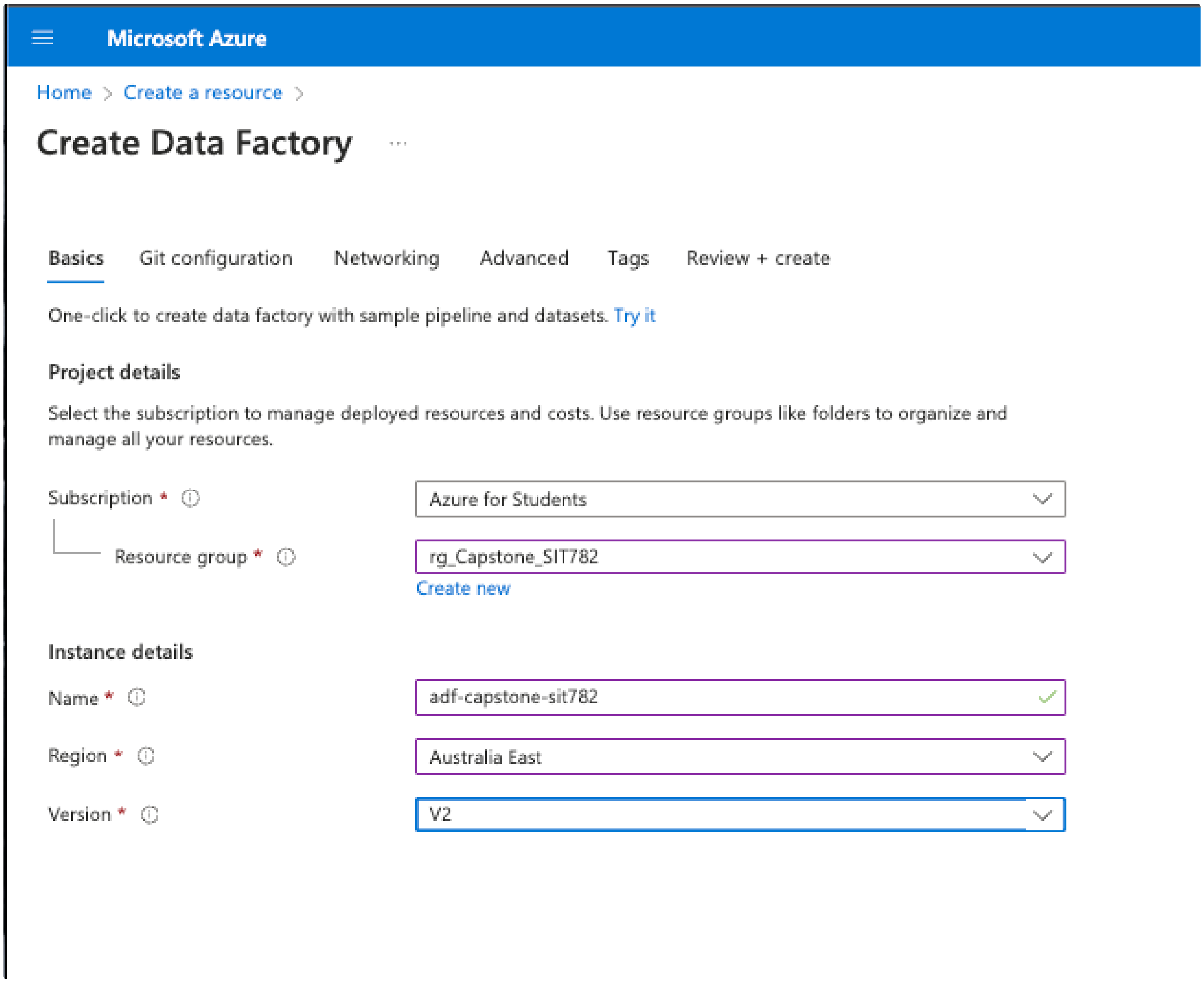
Koare DashBoard
Set up + subscribe | Learn more

BYOL - Standard
Create | Learn more

ByggFakta Plan User
Set up + subscribe | Learn more

Yellow Ai SaaS Basic Plan
Set up + subscribe | Learn more

2. Click on “Create Data Factory”, which will take you to a page like this. Select the resource group which was created earlier in (Azure Blob Storage - how to provision, access and development guide). Name your ADF instance following the naming standards which is mentioned on the creation page and select “Region” as “Australia East”, since that is the closest Azure region for us. Select “Version” as “V2” and click “Create”.



Microsoft Azure

Home > Create a resource >

Create Data Factory ...

Basics | Git configuration | Networking | Advanced | Tags | Review + create

One-click to create data factory with sample pipeline and datasets. Try it

Project details

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

Subscription * : Azure for Students

Resource group * : rg_Capstone_SIT782

Create new

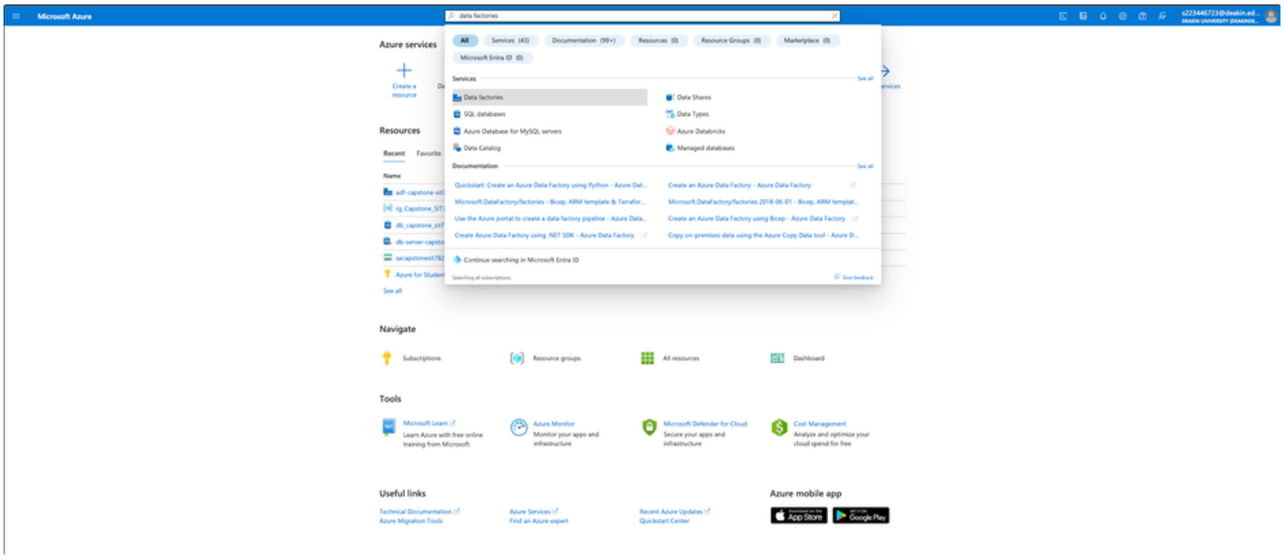
Instance details

Name * : adf-capstone-sit782

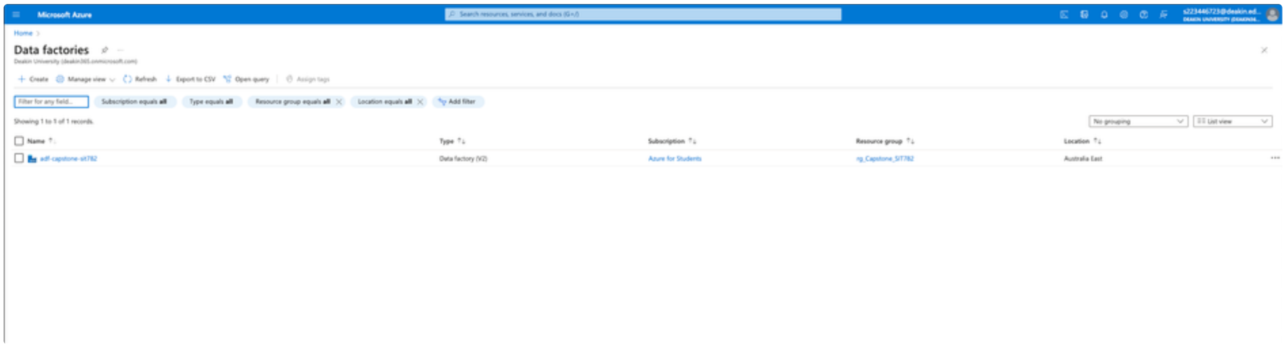
Region * : Australia East

Version * : V2

3. Once the resource has been created, go to Home page and from there in the menu bar, search for “data factories”. Select the Data factories services which comes up, as shown below -



4. Once in the page, you will be able to see all the Data Factories which has been provisioned. Select the ADF which was created in the previous step.

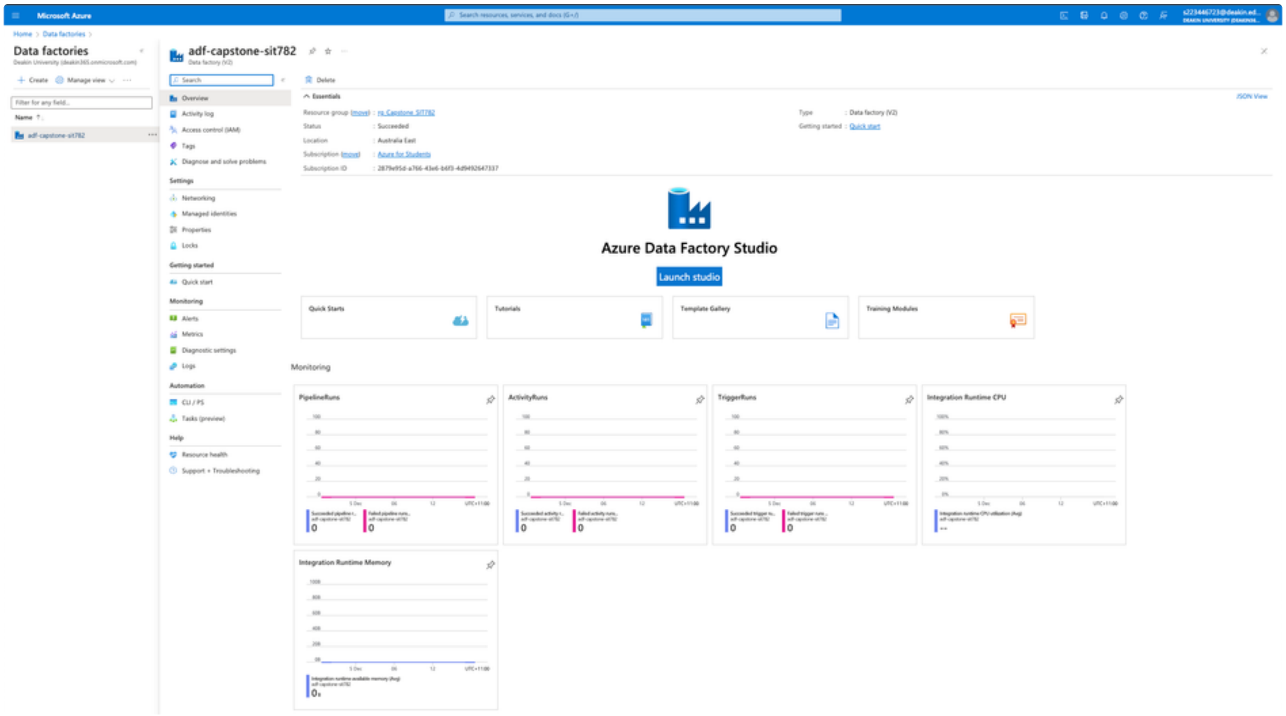


Data factories

Showing 1 to 1 of 1 records.

Name	Type	Subscription	Resource group	Location
adf-capstone-sit782	Data factory (V2)	Azure for Students	rg_Capstone_SIT782	Australia East

5. You will land in a page similar to what is shown below. This is the home page for your Data Factory. From here you can provision access, check logs, set up networking, monitor alerts and check resource health status. Click on “Launch Studio” next.



adf-capstone-sit782

Data factory (V2)

Resource group (move) : rg_Capstone_SIT782

Status : Succeeded

Location : Australia East

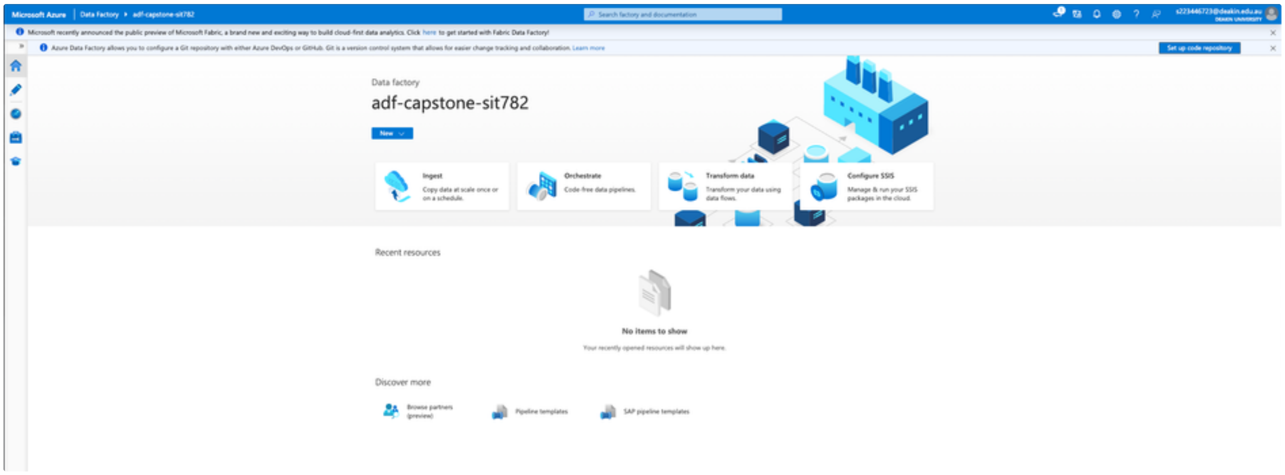
Subscription (move) : Azure for Students

Subscription ID : 2879e95d-a766-43e6-b6f3-4d9492647337

Type : Data factory (V2)

Getting started : Quick start

6. Once you click on “Launch Studio”, it will open a separate page as shown below. This is your homepage for Azure Data Factory instance which you have created earlier. Note that, for each separate Data Factory which has been provisioned, it will open a separate page. From this page, we will be able to perform all data engineering and data warehouse activities. For this example, we will ingest the EPL results file which was uploaded in the Azure Blob Storage earlier. Click on “Orchestrate” on the main page next.



Microsoft recently announced the public preview of Microsoft Fabric, a brand new and exciting way to build cloud-first data analytics. Click here to get started with Fabric Data Factory!

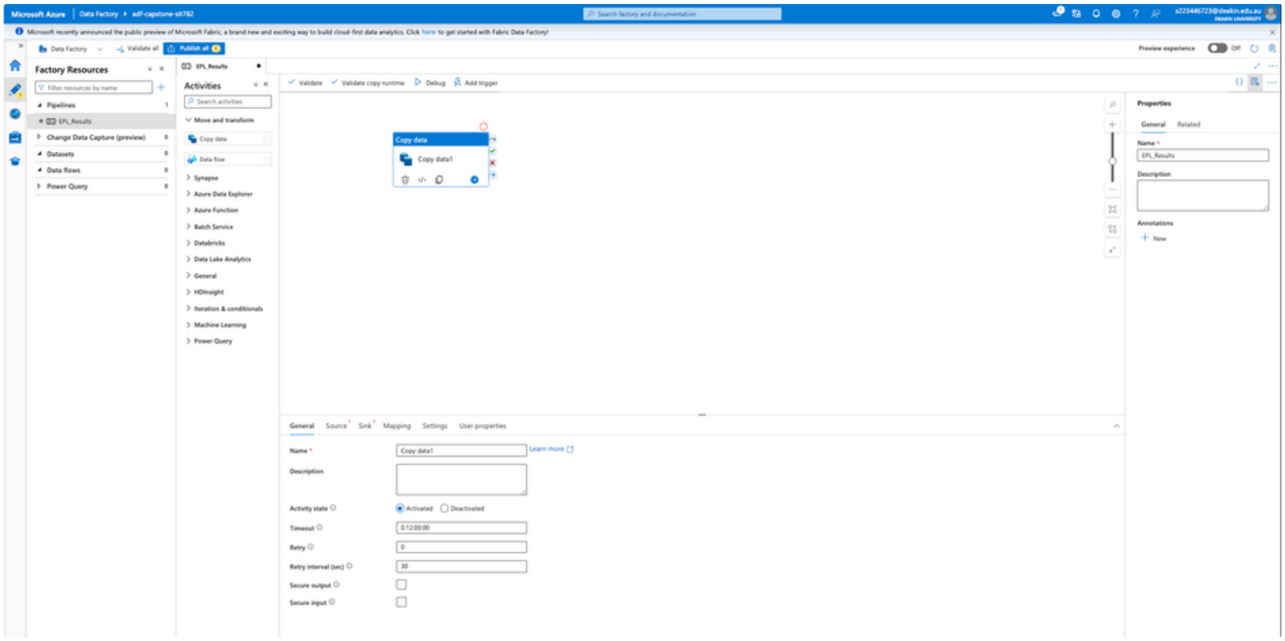
Azure Data Factory allows you to configure a Git repository with either Azure DevOps or GitHub. Git is a version control system that allows for easier change tracking and collaboration. Learn more

Data factory

adf-capstone-sit782

New
Ingest
Copy data at scale once or on a schedule.
Orchestrate
Code-free data pipelines.
Transform data
Transform your data using data flows.
Configure SSIS
Manage & run your SSIS packages in the cloud.
Recent resources
No items to show
Your recently opened resources will show up here.
Discover more
Browse partners (preview)
Pipeline templates
SAP pipeline templates
Set up code repository

7. Once you click on “Orchestrate”, it will open up a blank pipeline. Name the pipeline as “EPL_Results” on the right hand pane. Next click on “Move and transform” to expand it and drag and drop “Copy data” to the main window. You can rename the “Copy data” stage in the “Name” section below. I will keep this as-is for now.



Microsoft Azure | Data Factory ▸ adf-capstone-sit782

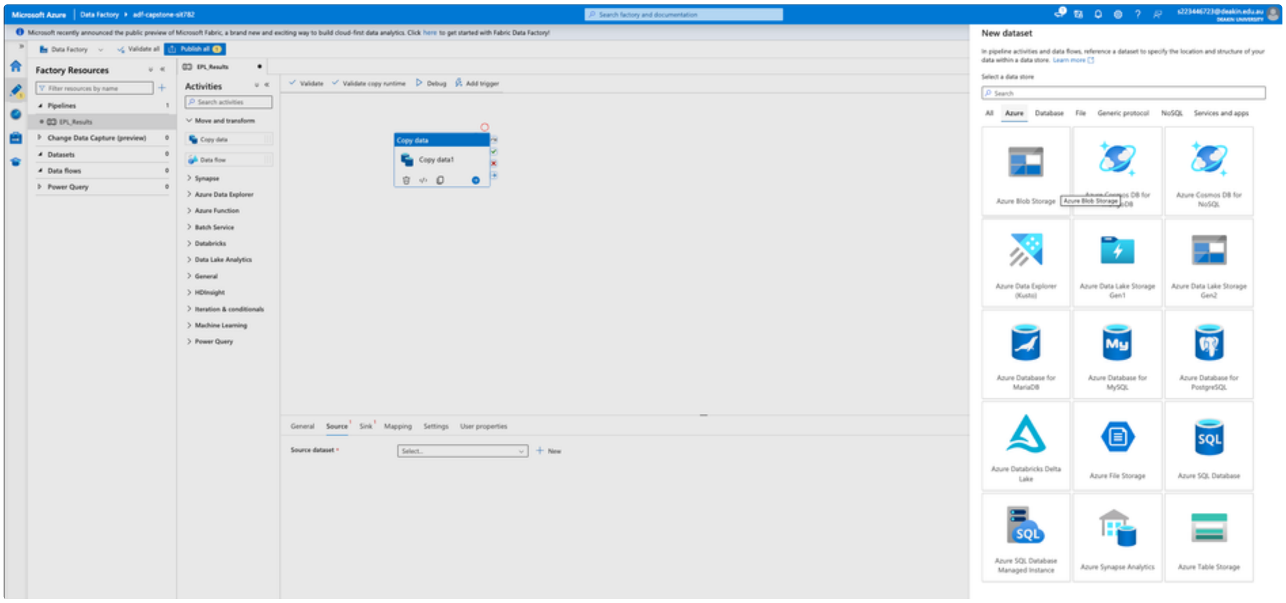
Microsoft recently announced the public preview of Microsoft Fabric, a brand new and exciting way to build cloud-first data analytics. Click here to get started with Fabric Data Factory!

Properties — General
Name *: EPL_Results
Description:

General | Source | Sink | Mapping | Settings | User properties

Name *: Copy data1
Description:
Activity state: ● Activated ○ Deactivated
Timeout: 0.12:00:00
Retry: 0
Retry interval (sec): 30
Secure output: □
Secure input: □

8. Now we will need to define the Source and Sink for this pipeline. Our source is the EPL Results file present in the Blob Storage and our Sink is the table which we created earlier (Azure SQL Database - how to provision, access and development guide). Click on “Source” and then click on “New”. This will open the right hand pane as shown below.



New dataset

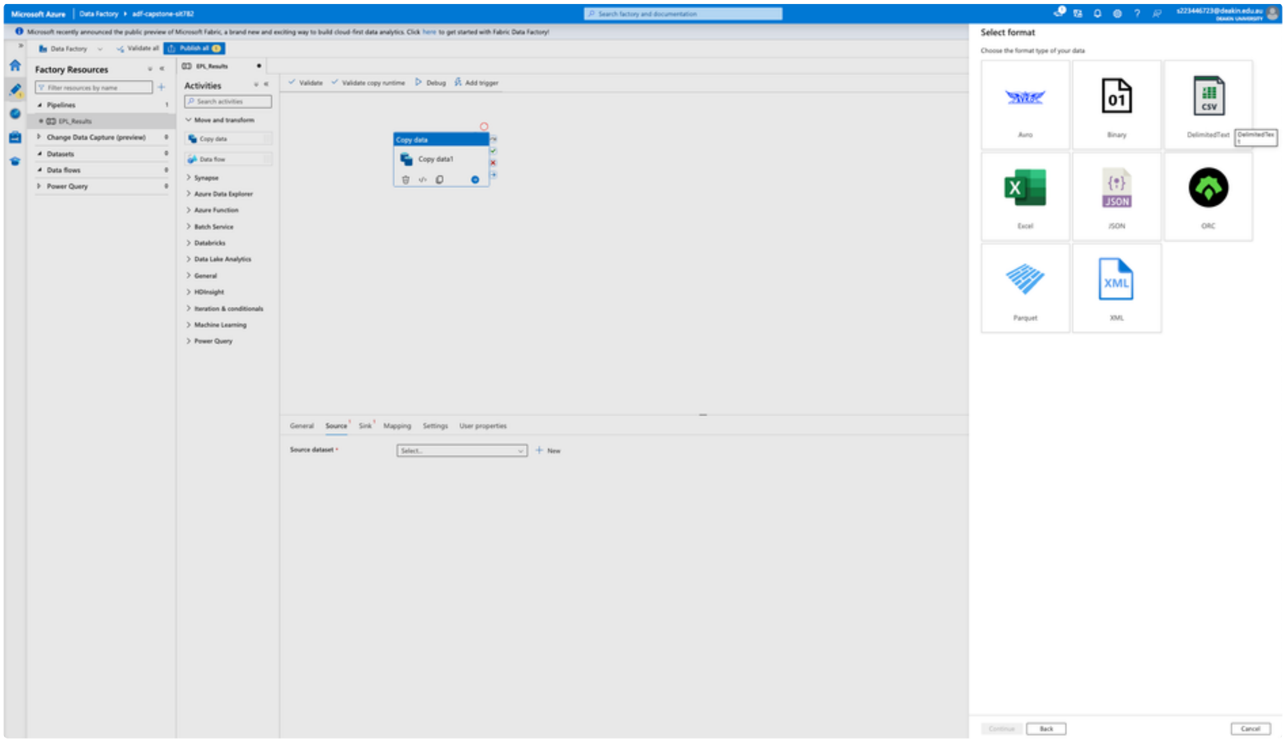
In pipeline activities and data flows, reference a dataset to specify the location and structure of your data within a data store. Learn more

Select a data store

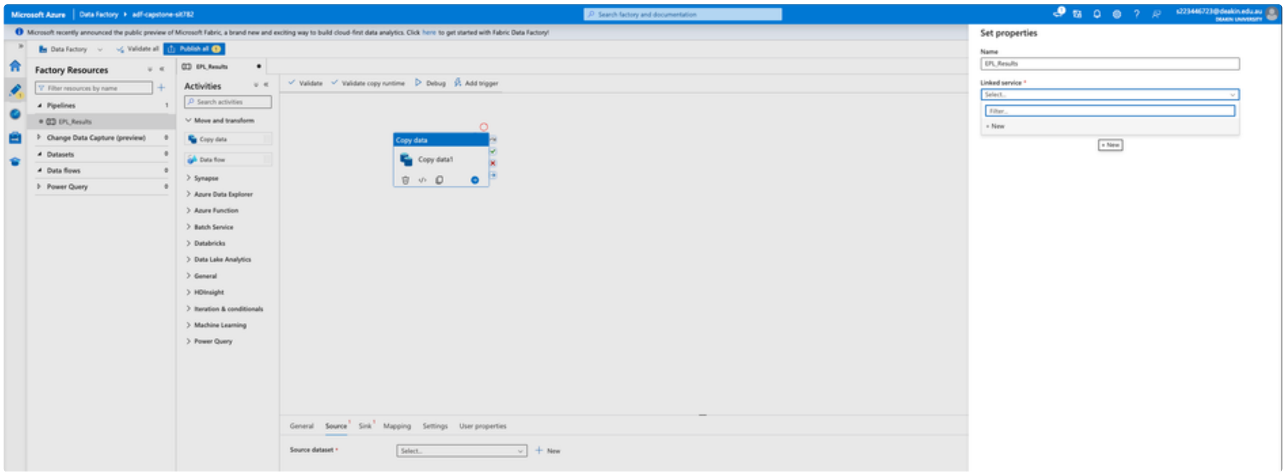
Source dataset *: Select... + New

Microsoft Azure | Data Factory ▸ adf-capstone-sit782
s223446723@deakin.edu.au
DEAKIN UNIVERSITY
Search
All
Azure
Database
File
Generic protocol
NoSQL
Services and apps
Azure Blob Storage
Azure Cosmos DB for MongoDB
Azure Cosmos DB for NoSQL
Azure Data Explorer (Kusto)
Azure Data Lake Storage Gen1
Azure Data Lake Storage Gen2
Azure Database for MariaDB
Azure Database for MySQL
Azure Database for PostgreSQL
Azure Databricks Delta Lake
Azure File Storage
Azure SQL Database
Azure SQL Database Managed Instance
Azure Synapse Analytics
Azure Table Storage
General
Source
Sink
Mapping
Settings
User properties
Copy data
Copy data1

9. Click on “Azure” on the right hand pane and then click on “DelimitedText”, since the file uploaded to Blob Storage is a csv file. If this is any other file, choose the appropriate format from the list below.



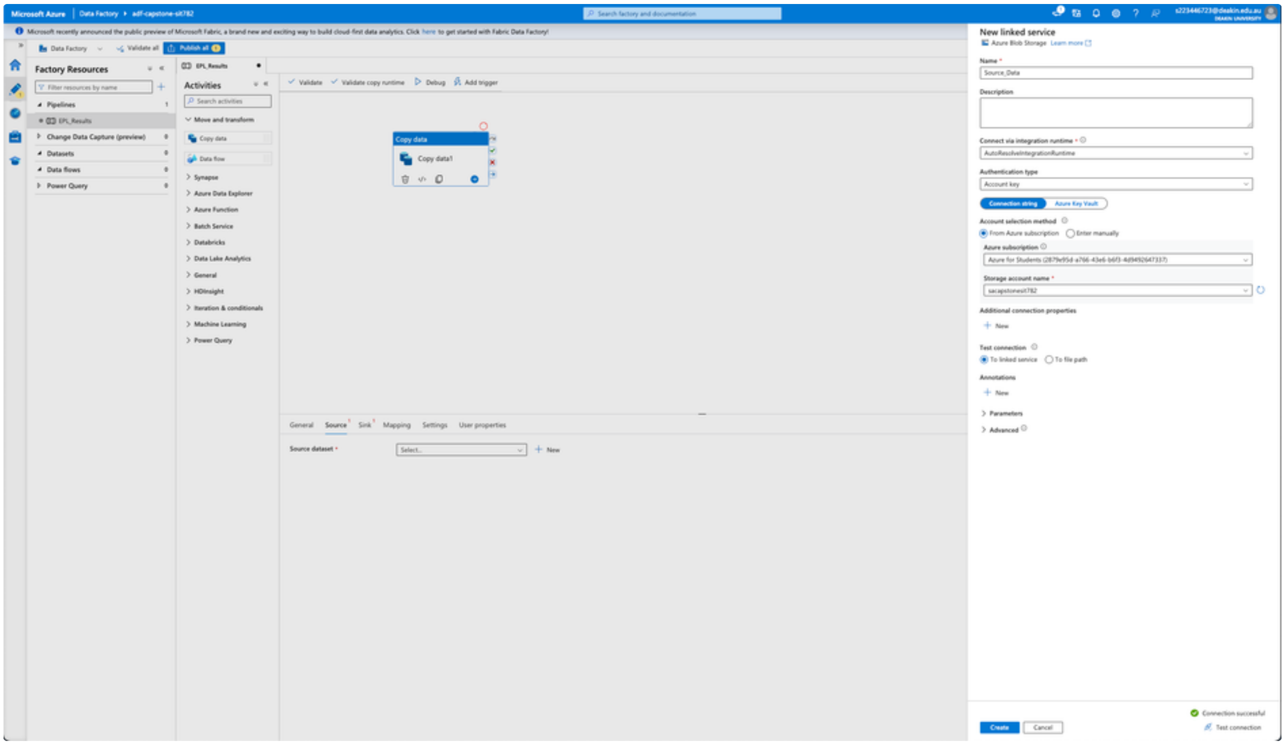
10. Now, name the file. For this POC, I have named it as EPL_Results. Now we need to define the “Linked service” for this Source, which is the connection to Blob Storage. Click on “New” under “Linked service”, since this is the first time we will be creating this Linked service.


Set properties

Name: EPL_Results

Linked service *: Select...

11. Name the Linked service. Under “Integration runtime”, select “AutoResolveIntegrationRuntime”. The other option is Self Hosted Integration Runtime, but that would mean more infrastructure setup. Next, select the Azure subscription and the storage account name. This is the same name which was defined when Azure Blob Storage was provisioned. Click on “Test connection” to test your connectivity and once it is successful, click on “Create”.



New linked service

Azure Blob Storage

Name *: Source_Data

Description:

Connect via integration runtime *: AutoResolveIntegrationRuntime

Authentication type: Account key

Connection string | Azure Key Vault

Account selection method: ● From Azure subscription ○ Enter manually

Azure subscription: Azure for Students (2879e95d-a766-43e6-b6f3-4d9492647337)

Storage account name *: sacapstonesit782

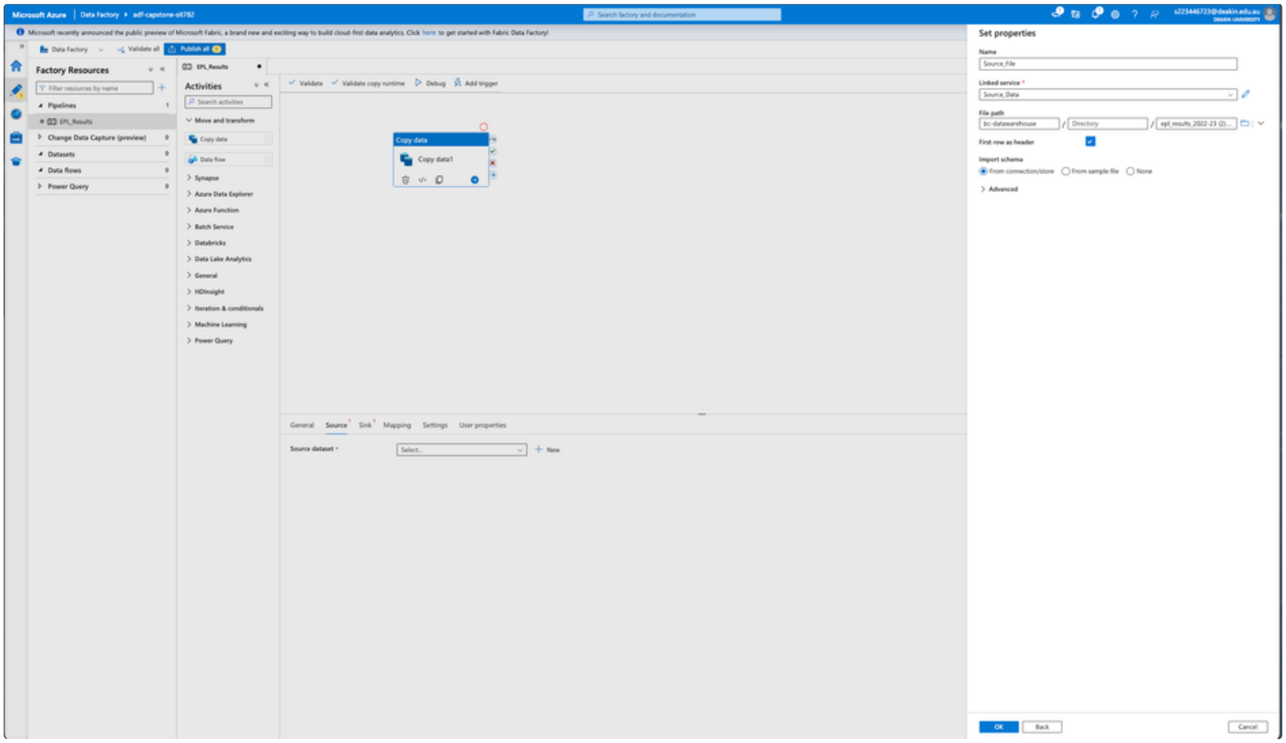
Additional connection properties

Test connection: ● To linked service ○ To file path

Annotations

Connection successful

12. Next we need to define the path to the file. Click on the Folder icon in “File path” and select the file. Under “Import schema”, select “From connection/store” to import the file schema into the pipeline.



Set properties

Name: Source_File

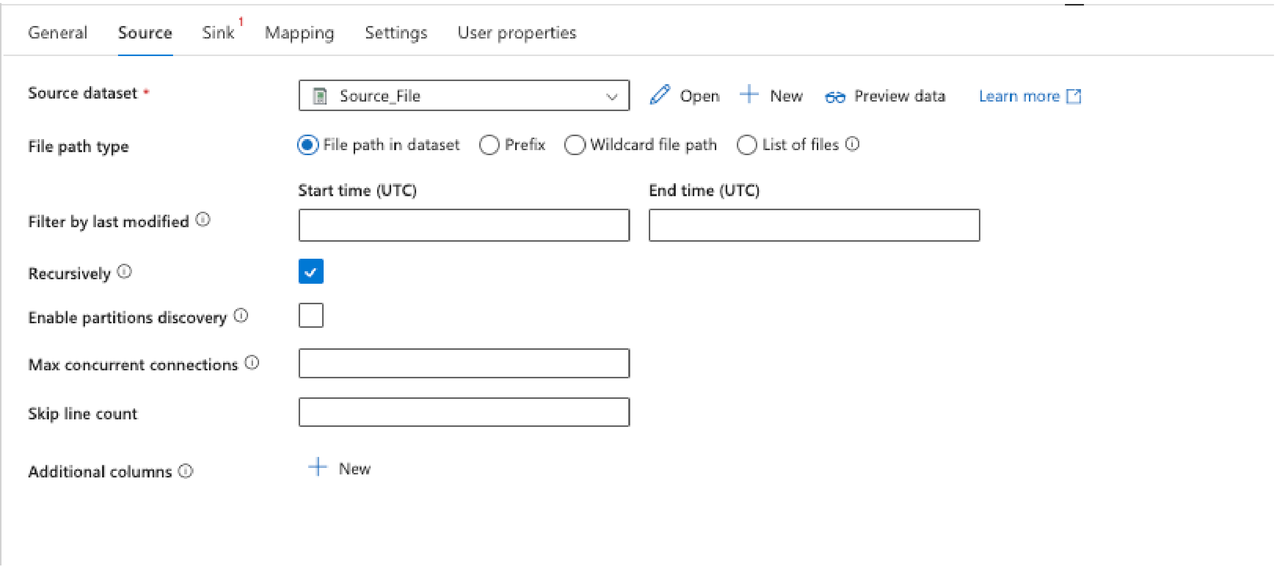
Linked service *: Source_Data

File path: bc-datawarehouse / Directory / epl_results_2022-23 (2)...

First row as header ☑

Import schema: ● From connection/store ○ From sample file ○ None

13. Once everything has been correctly set up by following the steps above, you will see the following under “Source” tab. Here you can see that there are multiple options. Since we have a single fixed file, we have chosen “File path in dataset”. If this involves multiple files, then choose “List of files”.



Source dataset * Source_File

File path type ● File path in dataset ○ Prefix ○ Wildcard file path ○ List of files

Filter by last modified: Start time (UTC) End time (UTC)

Recursively ☑

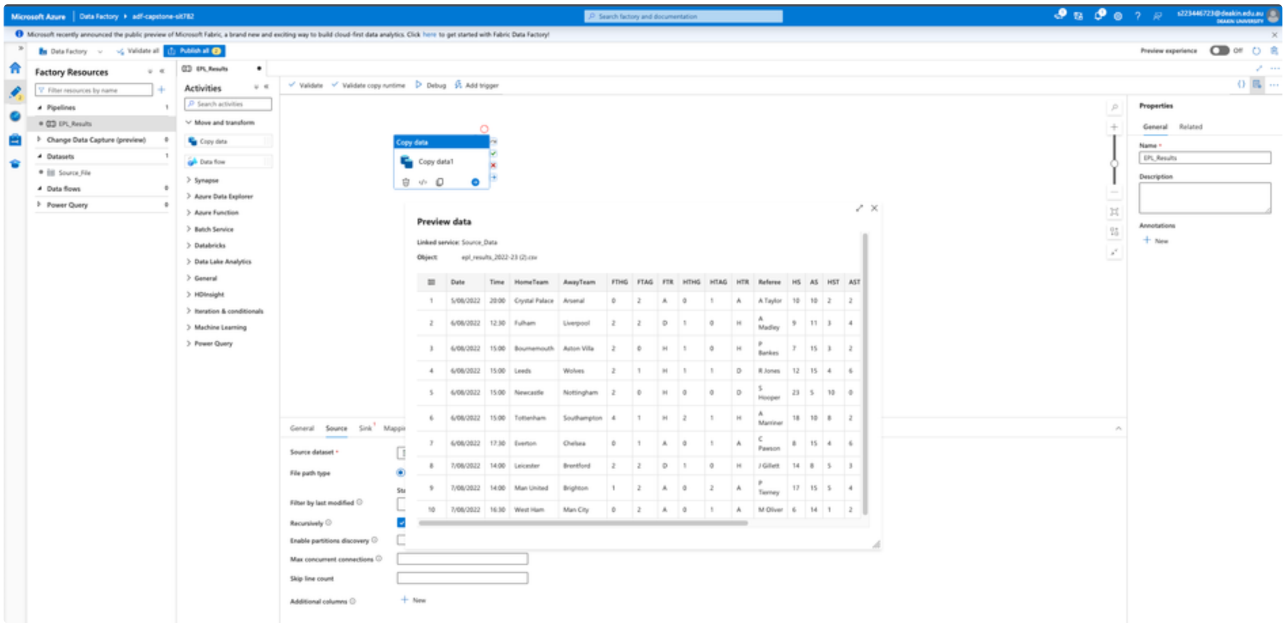
Enable partitions discovery □

Max concurrent connections

Skip line count

Additional columns + New

14. Next, check that the data in the "Source" is coming up correctly. This can be done by selecting "Preview data" on the "Source tab. Check the data in the window which opens up. Ideally, the data should look good, but if there are unicode character issues, then check the source file for any un-translated characters.



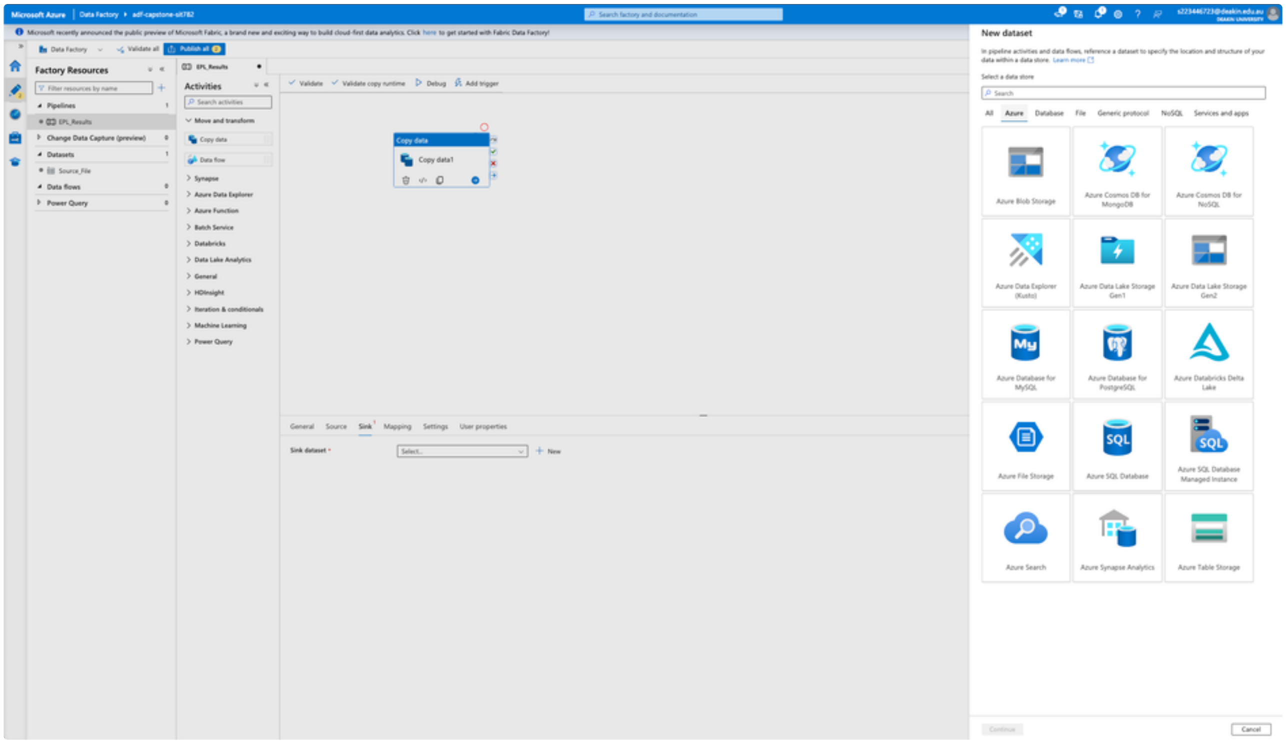
Preview data

Linked service: Source_Data

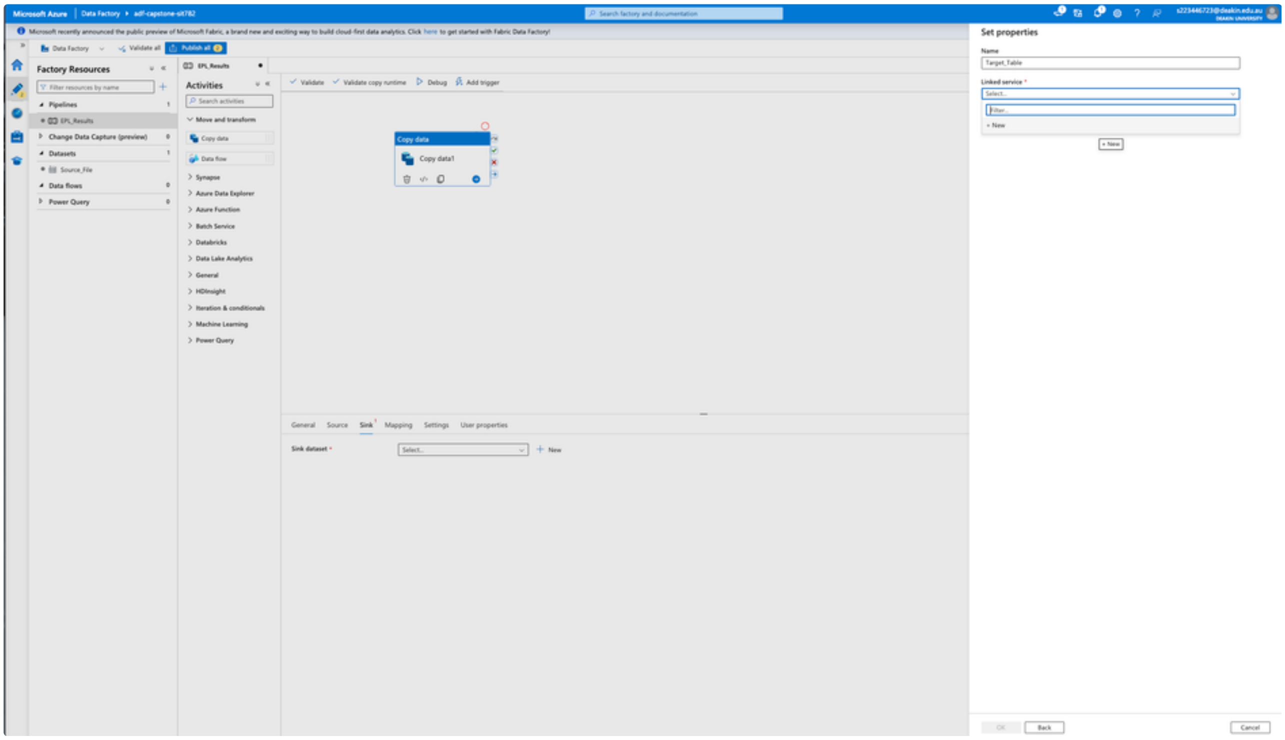
Object: epl_results_2022-23 (2).csv

	Date	Time	HomeTeam	AwayTeam	FTHG	FTAG	FTR	HTHG	HTAG	HTR	Referee	HS	AS	HST	AST
1	5/08/2022	20:00	Crystal Palace	Arsenal	0	2	A	0	1	A	A Taylor	10	10	2	2
2	6/08/2022	12:30	Fulham	Liverpool	2	2	D	1	0	H	A Madley	9	11	3	4
3	6/08/2022	15:00	Bournemouth	Aston Villa	2	0	H	1	0	H	P Bankes	7	15	3	2
4	6/08/2022	15:00	Leeds	Wolves	2	1	H	1	1	D	R Jones	12	15	4	6
5	6/08/2022	15:00	Newcastle	Nottingham	2	0	H	0	0	D	S Hooper	23	5	10	0
6	6/08/2022	15:00	Tottenham	Southampton	4	1	H	2	1	H	A Marriner	18	10	8	2
7	6/08/2022	17:30	Everton	Chelsea	0	1	A	0	1	A	C Pawson	8	15	4	6
8	7/08/2022	14:00	Leicester	Brentford	2	2	D	1	0	H	J Gillett	14	8	5	3
9	7/08/2022	14:00	Man United	Brighton	1	2	A	0	2	A	P Tierney	17	15	5	4
10	7/08/2022	16:30	West Ham	Man City	0	2	A	0	1	A	M Oliver	6	14	1	2

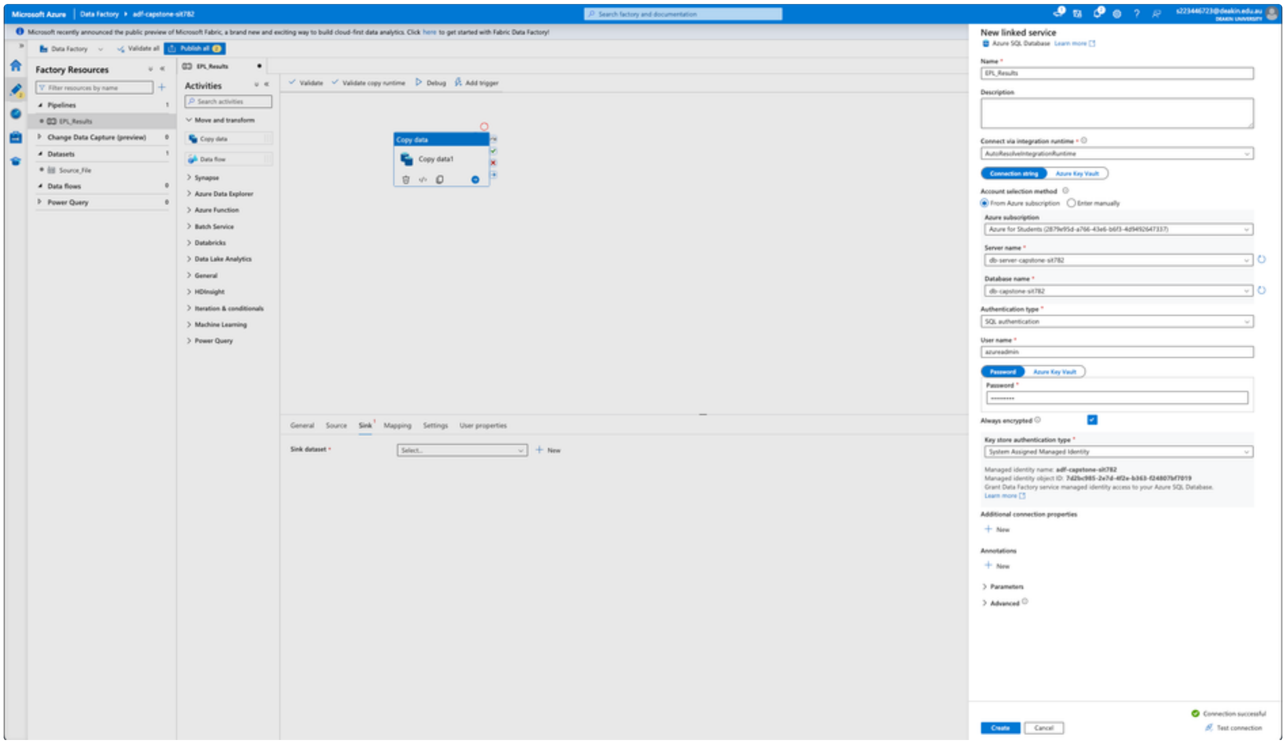
15. Now, we will define the "Sink". Click to "Sink" tab and then click on "New". This will open up a familiar page similar to the below screenshot -



16. Click on “Azure” and then select “Azure SQL Database”. We will next configure the connection to the SQL database which we created earlier (Azure SQL Database - how to provision, access and development guide). Name the Sink/target as appropriate and select “New” under “Linked services”. I will set up a new Linked service for the SQL Database, since this is the first time I am creating the Linked service. If the Linked service already exists, then it will appear in the drop-down list.



17. Name the linked service next and select the subscription from the drop-down list. Next, select the database server name and the database. Select the “Authentication type” as “SQL authentication” and fill in the details of the user id which was created when the database was provisioned. Input the credentials for the user id and select the check box next to “Always encrypted”. This will ensure that the data is always encrypted. Check the connection created by clicking on “Test connection” below. Once it is successful, click on “Create”.



Sink dataset * Select... + New

New linked service
Azure SQL Database Learn more

Name * EPL_Results

Description

Connect via integration runtime * AutoResolveIntegrationRuntime

Connection string | Azure Key Vault

Account selection method: ● From Azure subscription ○ Enter manually

Azure subscription: Azure for Students (2879e95d-a766-43e6-b6f3-4d9492647337)

Server name * db-server-capstone-sit782

Database name * db-capstone-sit782

Authentication type * SQL authentication

User name * azureadmin

Password | Azure Key Vault

Password * ••••••••••

Always encrypted ☑

Key store authentication type * System Assigned Managed Identity

Managed identity name: **adf-capstone-sit782**
Managed identity object ID: **7d2bc985-2e7d-4f2e-b363-f24807bf7019**
Grant Data Factory service managed identity access to your Azure SQL Database.
Learn more

Additional connection properties + New

Annotations + New

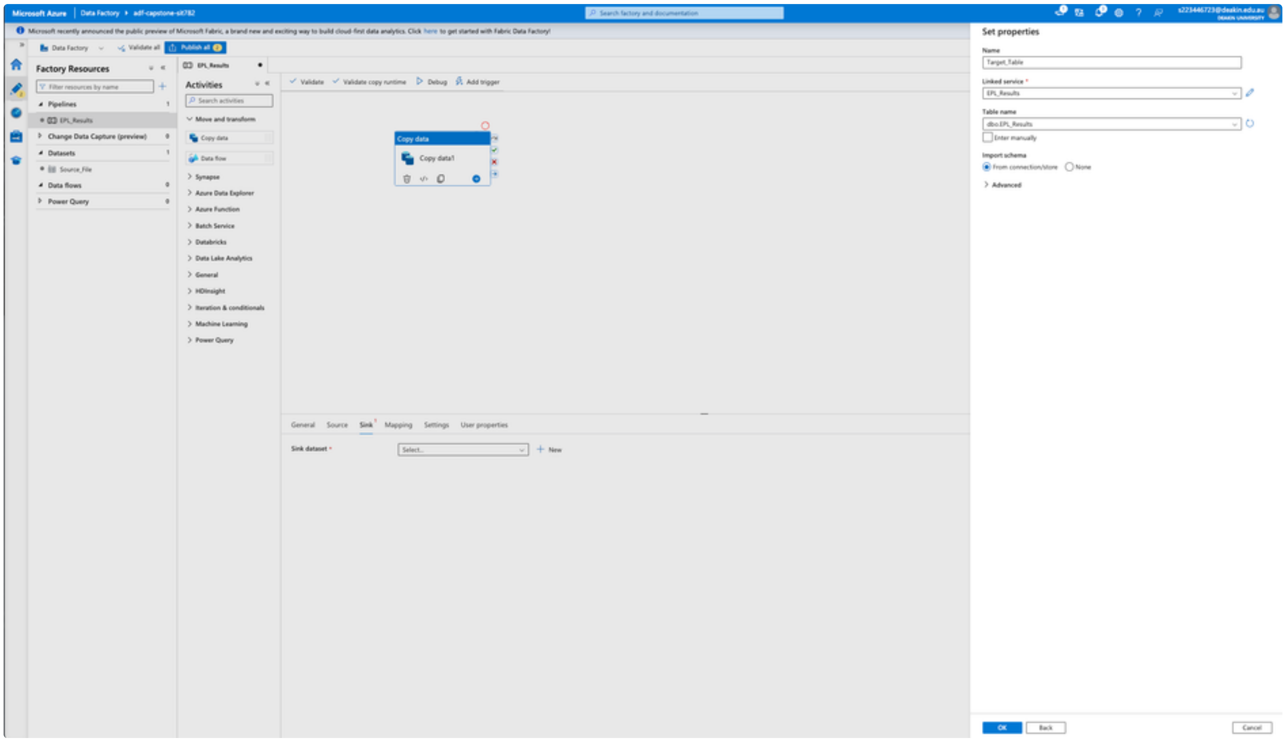
› Parameters

› Advanced

Connection successful

Create | Cancel | Test connection

18. Next, we need to select the table where we want to load the data. Under “Table name”, select “dbo.EPL_Results” and under “Import schema”, select “From connection/store” to get the schema of the target table. This is critical for the next steps. Click on “OK” to create the Sink.



Sink dataset * Select... + New

Set properties

Name: Target_Table

Linked service * EPL_Results

Table name: dbo.EPL_Results

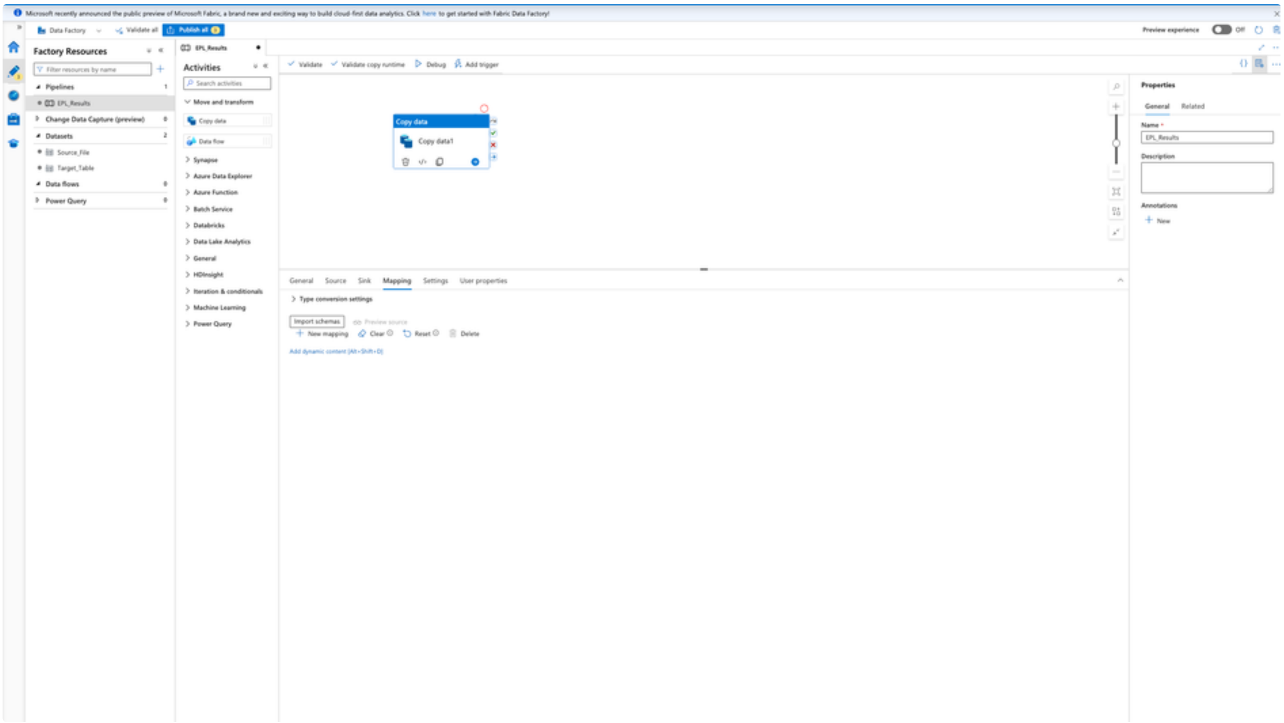
□ Enter manually

Import schema: ● From connection/store ○ None

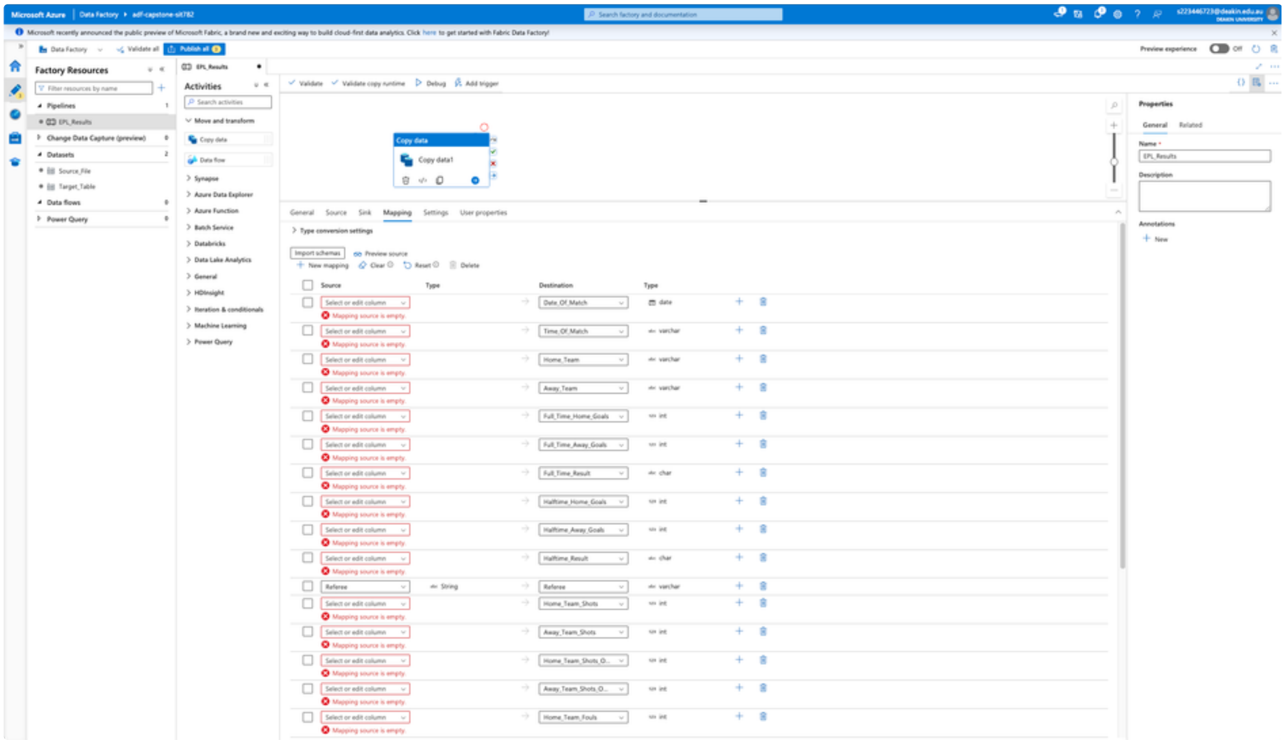
› Advanced

OK | Back | Cancel

19. Now that we have created the Linked services for Source and Sink, and have also imported the source and target schemas, we will next need to map them. Select “Mapping” tab and click on “Import schemas”.

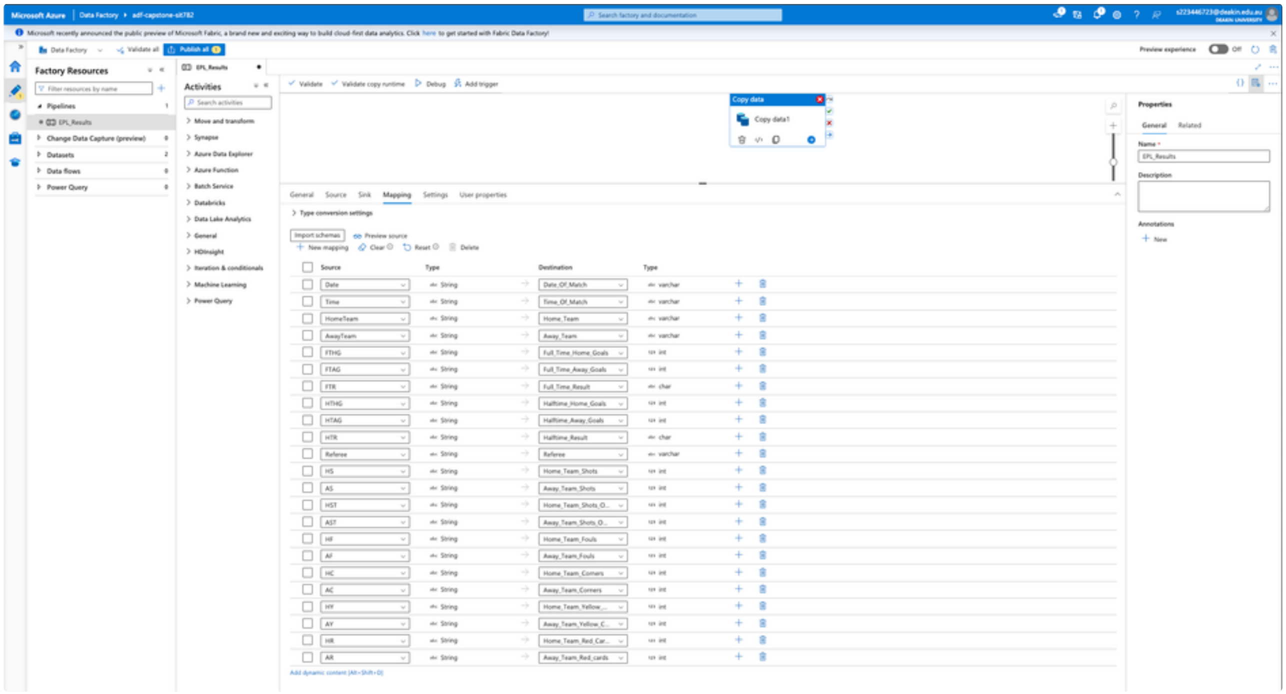


20. Once you click on “Import schemas”, both source and target schemas will appear as shown below. We will need to manually map each source column to target columns for the mapping to be complete. If the name of source and target column is similar, it will auto-map, as can be seen from the “Referee” column below. However, it is necessary to check each and every column to ensure proper mapping has been done.



Source	Type	Destination	Type
Select or edit column Mapping source is empty.		Date_Of_Match	date
Select or edit column Mapping source is empty.		Time_Of_Match	varchar
Select or edit column Mapping source is empty.		Home_Team	varchar
Select or edit column Mapping source is empty.		Away_Team	varchar
Select or edit column Mapping source is empty.		Full_Time_Home_Goals	int
Select or edit column Mapping source is empty.		Full_Time_Away_Goals	int
Select or edit column Mapping source is empty.		Full_Time_Result	char
Select or edit column Mapping source is empty.		Halftime_Home_Goals	int
Select or edit column Mapping source is empty.		Halftime_Away_Goals	int
Select or edit column Mapping source is empty.		Halftime_Result	char
Referee	String	Referee	varchar
Select or edit column Mapping source is empty.		Home_Team_Shots	int
Select or edit column Mapping source is empty.		Away_Team_Shots	int
Select or edit column Mapping source is empty.		Home_Team_Shots_O...	int
Select or edit column Mapping source is empty.		Away_Team_Shots_O...	int
Select or edit column Mapping source is empty.		Home_Team_Fouls	int

21. Once you select all the mappings for source and sink, you will see that the error sign above the “Mapping” tab will disappear. Now, we will first save the pipeline, by clicking on “Publish all” on the top left. Clicking on that starts the validation process and publishing the pipeline saves your work.



General | Source | Sink | **Mapping** | Settings | User properties

> Type conversion settings

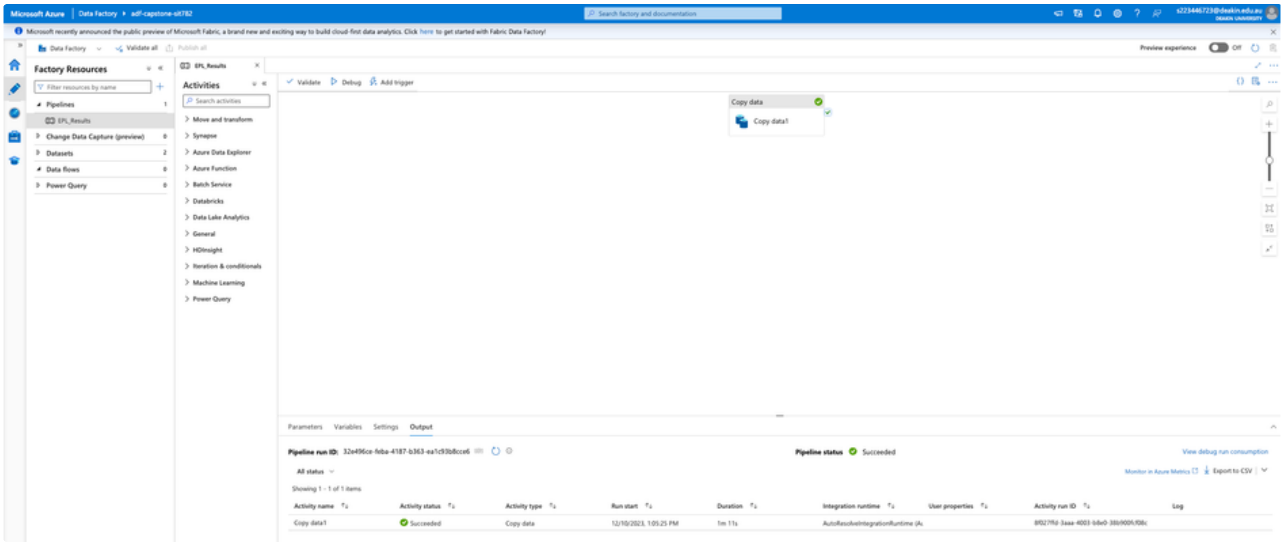
Import schemas · Preview source

+ New mapping · Clear · Reset · Delete

Source	Type	Destination	Type
Date	String	Date_Of_Match	varchar
Time	String	Time_Of_Match	varchar
HomeTeam	String	Home_Team	varchar
AwayTeam	String	Away_Team	varchar
FTHG	String	Full_Time_Home_Goals	int
FTAG	String	Full_Time_Away_Goals	int
FTR	String	Full_Time_Result	char
HTHG	String	Halftime_Home_Goals	int
HTAG	String	Halftime_Away_Goals	int
HTR	String	Halftime_Result	char
Referee	String	Referee	varchar
HS	String	Home_Team_Shots	int
AS	String	Away_Team_Shots	int
HST	String	Home_Team_Shots_O...	int
AST	String	Away_Team_Shots_O...	int
HF	String	Home_Team_Fouls	int
AF	String	Away_Team_Fouls	int
HC	String	Home_Team_Corners	int
AC	String	Away_Team_Corners	int
HY	String	Home_Team_Yellow_...	int
AY	String	Away_Team_Yellow_C...	int
HR	String	Home_Team_Red_Car...	int
AR	String	Away_Team_Red_cards	int

Add dynamic content [Alt+Shift+D]

22. Click on “Debug” next, this will trigger the pipeline. Once the pipeline completes, you will see something similar to this -



Parameters | Variables | Settings | **Output**

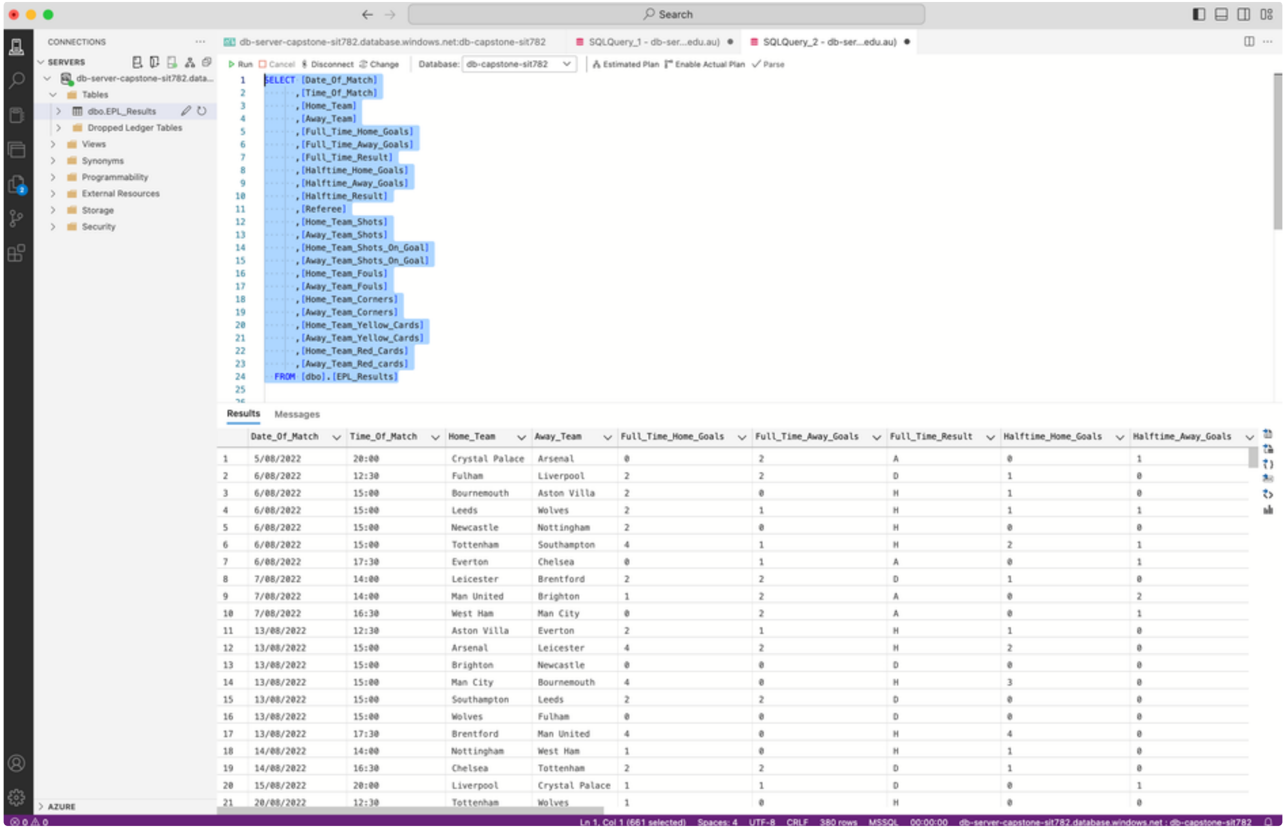
Pipeline run ID: 32e496ce-feba-4187-b363-ea1c93b8cce6

Pipeline status ✓ Succeeded

Showing 1 - 1 of 1 items

Activity name	Activity status	Activity type	Run start	Duration	Integration runtime	User properties	Activity run ID	Log
Copy data1	✓ Succeeded	Copy data	12/10/2023, 1:05:25 PM	1m 11s	AutoResolveIntegrationRuntime (A…		8f027f6d-3aaa-4003-b8e0-38b9005f08c	

23. Now, check the table and confirm if the data has been loaded by logging into Azure Data Studio, as configured before. Selecting all the records from this small table will enable you to see if every field has been loaded properly.


CONNECTIONS
SERVERS
db-server-capstone-sit782.data...
Tables
dbo.EPL_Results
Dropped Ledger Tables
Views
Synonyms
Programmability
External Resources
Storage
Security
AZURE
db-server-capstone-sit782.database.windows.net:db-capstone-sit782
SQLQuery_1 - db-ser...edu.au)
SQLQuery_2 - db-ser...edu.au)
Run
Cancel
Disconnect
Change
Database: db-capstone-sit782
Estimated Plan
Enable Actual Plan
Parse
Search
Results
Messages
Ln 1, Col 1 (661 selected)   Spaces: 4   UTF-8   CRLF   380 rows   MSSQL   00:00:00   db-server-capstone-sit782.database.windows.net : db-capstone-sit782

```sql
SELECT [Date_Of_Match]
      ,[Time_Of_Match]
      ,[Home_Team]
      ,[Away_Team]
      ,[Full_Time_Home_Goals]
      ,[Full_Time_Away_Goals]
      ,[Full_Time_Result]
      ,[Halftime_Home_Goals]
      ,[Halftime_Away_Goals]
      ,[Halftime_Result]
      ,[Referee]
      ,[Home_Team_Shots]
      ,[Away_Team_Shots]
      ,[Home_Team_Shots_On_Goal]
      ,[Away_Team_Shots_On_Goal]
      ,[Home_Team_Fouls]
      ,[Away_Team_Fouls]
      ,[Home_Team_Corners]
      ,[Away_Team_Corners]
      ,[Home_Team_Yellow_Cards]
      ,[Away_Team_Yellow_Cards]
      ,[Home_Team_Red_Cards]
      ,[Away_Team_Red_cards]
  FROM [dbo].[EPL_Results]
```

	Date_Of_Match	Time_Of_Match	Home_Team	Away_Team	Full_Time_Home_Goals	Full_Time_Away_Goals	Full_Time_Result	Halftime_Home_Goals	Halftime_Away_Goals
1	5/08/2022	20:00	Crystal Palace	Arsenal	0	2	A	0	1
2	6/08/2022	12:30	Fulham	Liverpool	2	2	D	1	0
3	6/08/2022	15:00	Bournemouth	Aston Villa	2	0	H	1	0
4	6/08/2022	15:00	Leeds	Wolves	2	1	H	1	1
5	6/08/2022	15:00	Newcastle	Nottingham	2	0	H	0	0
6	6/08/2022	15:00	Tottenham	Southampton	4	1	H	2	1
7	6/08/2022	17:30	Everton	Chelsea	0	1	A	0	1
8	7/08/2022	14:00	Leicester	Brentford	2	2	D	1	0
9	7/08/2022	14:00	Man United	Brighton	1	2	A	0	2
10	7/08/2022	16:30	West Ham	Man City	0	2	A	0	1
11	13/08/2022	12:30	Aston Villa	Everton	2	1	H	1	0
12	13/08/2022	15:00	Arsenal	Leicester	4	2	H	2	0
13	13/08/2022	15:00	Brighton	Newcastle	0	0	D	0	0
14	13/08/2022	15:00	Man City	Bournemouth	4	0	H	3	0
15	13/08/2022	15:00	Southampton	Leeds	2	2	D	0	0
16	13/08/2022	15:00	Wolves	Fulham	0	0	D	0	0
17	13/08/2022	17:30	Brentford	Man United	4	0	H	4	0
18	14/08/2022	14:00	Nottingham	West Ham	1	0	H	1	0
19	14/08/2022	16:30	Chelsea	Tottenham	2	2	D	1	0
20	15/08/2022	20:00	Liverpool	Crystal Palace	1	1	D	0	1
21	20/08/2022	12:30	Tottenham	Wolves	1	0	H	0	0

If you have followed all the steps outlined above, you should be able to successfully provision a new Azure Data Factory, create Linked services and load data from Azure Blob Storage to Azure SQL Database. This can be used as template for any such data engineering/data warehouse activities in your assigned project.