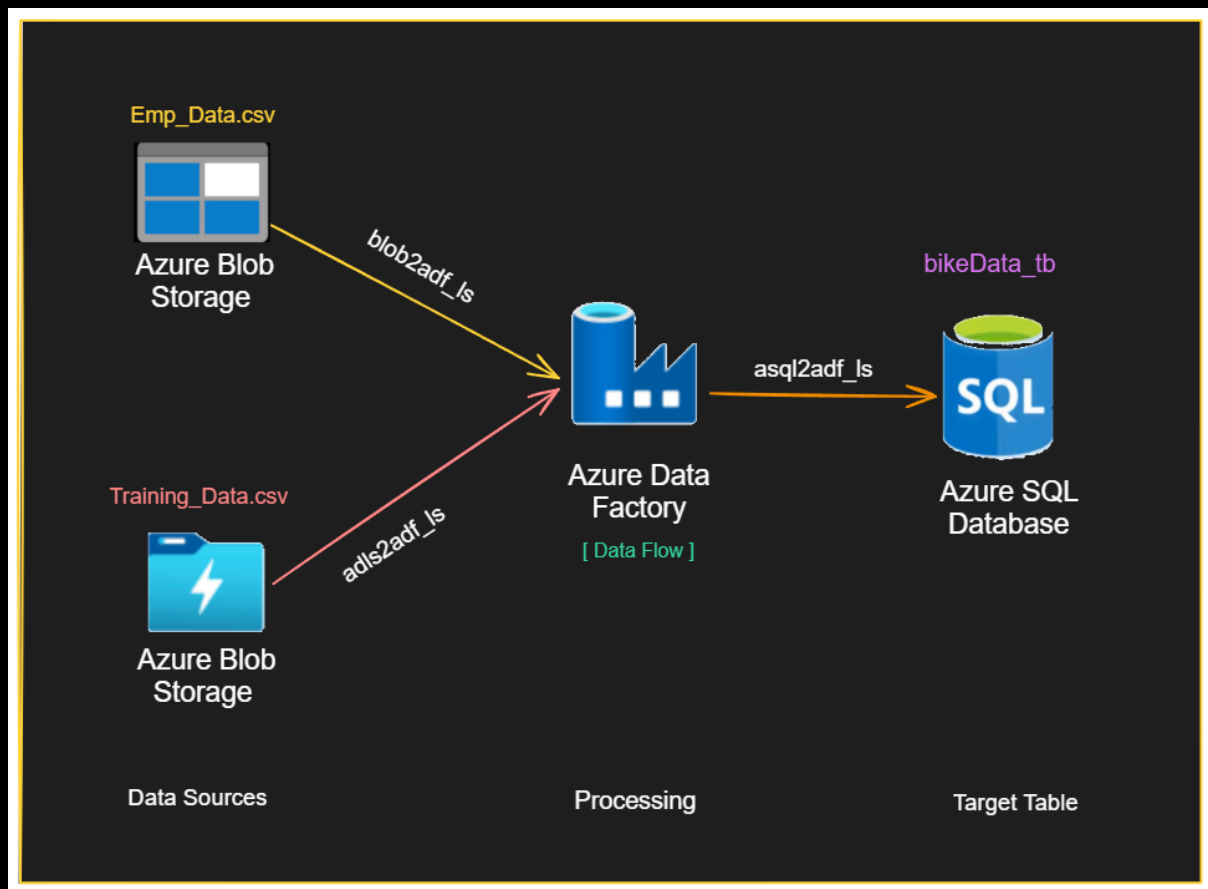


4th pipeline (blob,datalake2asql) using dataflow

Pipeline Architecture :



Azure Services Required :

1. Azure blob storage
2. Azure Data Lake Storage
3. Azure DataFactory
4. Azure SQL Database

Azure Blob Storage Creation :

- Creation of azure blob storage can be shown in the document below .
- Parameters used while creating azure blob storage.
- Storage account name : **blobbsa**
- https://docs.google.com/document/d/1vZCMfM9ieALTQm6Jdwl2JzxcMkwnyvkSJHCrry2kVlo/edit?usp=drive_link
- Create a container in blobbsa (name : **Emp Dataset**).
- Upload **emp_data** in **Emp Dataset**

Azure Data Lake Storage creation :

- The Azure data lake storage creation is shown in the below document.
- Parameters used while creating azure data lake storage.
- Storage account name : **addlsa**
- https://docs.google.com/document/d/1Gyz7yN9HDF7d_i6wM_i9ph0Z0PNiKP_FLgPfRJshs_A/edit?usp=sharing
- Create a container in addlsa (name : trainingdataset)
- Upload **Training_data** in **trainingdataset**

Azure SQL Database Creation :

- Creation of an azure sql database is shown in the document below.
- Parameters used while creating azure sql database.
- Database name : **sqldb**
- Server name : **projectsserver**
- Server admin login : **project_admin**
- https://docs.google.com/document/d/16iB1EsGKHc6-bcgTPSfqkK6BVf3n8_fpbat42uNOvXc/edit?usp=drive_link

Azure Data Factory Creation :

- Creation of the data factory is shown in the document below.
- Parameters used while creating azure data factory.
- Datafactory name : **projects-datafactory**
- https://docs.google.com/document/d/1lpvA7XumJjbIP0wPWWUlf_d_gdvh6iTn_0a12HWv0jcM/edit?usp=sharing
- Click on **Go to resource**.
- Click on **launch workspace**.

- Now let's do the transformations on data
- Open data factory and create datasets
- Pencil icon → datasets → new dataset → Azure blob Storage → csv → continue →

Microsoft Azure | Data Factory | projects-datafactory

Search factory and documentation

lets huntvarthya@outlook.com
DEFAULT DIRECTORY

Factory Resources

- Pipelines 0
- Change Data Capture (preview) 0
- Datasets 0
- Data flows 0
- Power Query 0

Set properties

1 → Name: empinputdata

2 → Linked service: blob2adf_js

File path: empdataset / Directory / 2669411-Emp_Training... 3

First row as header: ☒

Import schema: ☒ From connection/store ☐ From sample file ☐ None

4 ↓ OK Back Cancel

- Pencil icon → datasets → new dataset → Azure data lakeStorage → excel → continue →

Excel
traininginputdata

learning Center

Connection Schema Parameters

Linked service: adf2adls_js Test connection Edit + New Learn more

File path: trainingdataset / Directory / 2669407-Training.xlsx Browse

Compression type: Select...

Worksheet mode: ☒ Name ☐ Index

Sheet name: Sheet1 Refresh Preview data

Range: e.g. A3:H5

Null value:

Properties

General Related

Name: traininginputdata

Description:

Annotations
+ New

- Publish all → publish

Factory Resources

Filter resources by name

Pipelines 0

Change Data Capture (preview) 0

Datasets 2

- empinputdata
- traininginputdata

Data flows 1

- EmpTrain_DF

Power Query

empinputdata

traininginputdata

EmpTrain_DF

Source settings

New data flow

New flowlet

New folder

Dataset *

empinputdata

Options

- ☒ Allow schema drift
- ☐ Infer drifted column types
- ☐ Validate schema

Properties

General

Related

Name *

EmpTrain_DF

Description

source1

Import data from empinputdata

source2

Columns: 6 total

Source settings

Source type *

Dataset *

traininginputdata

Options

- ☒ Allow schema drift

Properties

General

Related

Name *

EmpTrain_DF

Description

- Joining empinputdata with traininginputdata

getEmployeeData

14 Columns

getTrainingData

Join settings

Left stream *

getEmployeeData

Right stream *

getTrainingData

Join type *

Full outer

Inner

Left outer

Right outer

Custom (cross)

Use fuzzy matching

Join conditions *

Left: getEmployeeData's column

EmployeeID

Right: getTrainingData's column

Employee_ID

Properties

General

Related

Name *

EmpTrain_DF

Description

Join settings Optimize Inspect **Data preview** ←

↑↓	Empl...	123	↑↓	Manag...	abc	↑↓	Title	abc	↑↓	Marital...	abc	↑↓	Gender	abc	↑↓	HireDate	abc	↑↓	Dept	abc	↑↓
+	1			16			Gustav...			M			M			2/2/20...			Sales		
+	1			16			Gustav...			M			M			2/2/20...			Sales		
+	1			16			Gustav...			M			M			2/2/20...			Sales		
+	1			16			Gustav...			M			M			2/2/20...			Sales		
+	1			16			Gustav...			M			M			2/2/20...			Sales		
+	1			16			Gustav...			M			M			2/2/20...			Sales		
+	1			16			Gustav...			M			M			2/2/20...			Sales		
+	2			6			Catheri...			S			M			8/31/2...			Sales		
+	2			6			Catheri...			S			M			8/31/2...			Sales		
+	2			6			Catheri...			S			M			8/31/2...			Sales		
+	2			6			Catheri...			S			M			8/31/2...			Sales		
+	2			6			Catheri...			S			M			8/31/2...			Sales		

Properties

General Related

Name *
EmpTrain_DF

Description

getEmployeeDa... join1 14 Columns ← select

getTrainingData

Select settings Optimize Inspect **Data preview**

joint's column	Name as
129 EmployeeID	EmployeeID
abc ManagerID	ManagerID
abc Title	Title
abc MaritalStatus	MaritalStatus
abc Gender	Gender
abc HireDate	HireDate
abc Dept	Dept
abc Job Grade	Job Grade
abc Start Date	Start Date
129 Course Code	Course Code
abc Course Name	Course Name
129 Employee_ID	Employee_ID
129 Cost	Cost
abc Supplier	Supplier

Properties

General Related

Name *
EmpTrain_DF

Description

Data Factory Validate all Publish all 3

empinputdata traininginputdata EmpTrain_DF

✓ Validate Data flow debug Debug Settings

getEmployeeDa... join1 select1 13 Columns

getTrainingData

Sink Settings Errors Mapping Optimize Inspect Data preview

Output stream name *
sink1 Learn more

Description
Add sink dataset Reset

Incoming stream *
select1

Sink type *
Dataset Inline Cache

Dataset *
Select... + New

Options
☒ Allow schema drift
☐ Validate schema

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

Search

All Azure Database File Generic protocol NoSQL Services and apps

Amazon S3 Azure Blob Storage Azure Cosmos DB for NoSQL

Azure Data Explorer (Kusto) Azure Data Lake Storage Gen2 Azure Database for MySQL

Continue Cancel

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

Search

All Azure Database File Generic protocol NoSQL Services and apps

Azure Database for PostgreSQL Azure SQL Database Azure SQL Database Managed Instance

Azure Synapse Analytics Dataaverse (Common Data Service for Apps) Dynamics 365

Continue Cancel

Handwritten notes: 1 sink, T2, 3, 4

Set properties

Name: EmpTraining_TB

Linked service: adf2asql_ls

Select from existing table New table

Schema and table name

Schema name: emptraining_tb Table name: emptraining_tb

Advanced

OK Back Cancel

Handwritten notes: 1, 2, 3, 4, 5

- Publish all → publish
- Pencil icon → new pipeline →
- Pipeline name : EmpTraining_PL
- Drag & drop data flow activity to the field

Factory Resources

Pipelines: EmpTraining_PL

Datasets: empinputdata, EmpTraining_TB, traininginputdata

Data flows: EmpTrain_DF

Power Query

Activities

Move and transform: Copy data, Data flow

Synapse: Azure Data Explorer, Azure Function, Batch Service, Databricks, Data Lake Analytics

General: HDInsight, Iteration & conditionals, Machine Learning, Power Query

Data flow

Data flow1

Properties

General: Name: EmpTraining_PL, Description:

Annotations: + New

Settings: Data flow: EmpTrain_DF, Run on (Azure IR): AutoResolveIntegrationRuntime, Compute size: Small, Logging level: Verbose

Handwritten notes: 1, 2, 3

- Publish all → publish
- Add trigger → trigger now

Dashboard > All pipeline runs > EmpTraining_PL - Activity runs

Rerun Cancel Refresh Update pipeline List Gantt

Data flow Data flow1

Activity runs

Pipeline run ID 76301d1f-6ee4-471d-b3f5-f9c0c3922631

All status Monitor in Azure Metrics Export to CSV

Showing 1 - 1 items

Activity name	Activity status	Activity type	Run start	Duration	Integration runtime	User properties	Activity run ID
Data flow1	Succeeded	Data flow	5/28/2024, 5:54:03 PM	3m 26s	AutoResolveIntegration		2bec814b-dcd1-4708-f

Home > Resource groups > projects_rg > sqladb (projectsserver/sqladb)

sqladb (projectsserver/sqladb) | Query editor (preview)

Search Login New Query Open query Feedback Getting started

Overview Activity log Tags Diagnose and solve problems Query editor (preview) Settings Compute + storage Connection strings Properties Locks Data management Replicas Sync to other databases Integrations

Welcome to SQL Database Query Editor

SQL server authentication

Login * project_admin

Password *

OK

Microsoft Entra authentication

Continue as letshuntvarthya@outlook.c...

Home > Resource groups > projects_rg > sqladb (projectsserver/sqladb)

sqladb (projectsserver/sqladb) | Query editor (preview)

Login New Query Open query Feedback Getting started

sqladb (project_admin)

Showing limited object explorer here. For full capability please click here to open Azure Data Studio.

Tables Views Stored Procedures

dbo.emptraining_tb

Select Top 1000 Rows Edit Data (Preview) Rename

Query 1 Query 2

Run Cancel query Save query Export data as Show only Editor

1 SELECT TOP (1000) * FROM [dbo].[emptraining_tb]

ManagerID	Title	MaritalStatus	Gender	HireDate	Dept	Job Grade	Start Date	Co
1	Gustavo Achong	M	M	2/2/2013 0:00	Sales	Admin	2017-09-17	1
1	Gustavo Achong	M	M	2/2/2013 0:00	Sales	Admin	2017-06-25	16

Query succeeded | 1s

