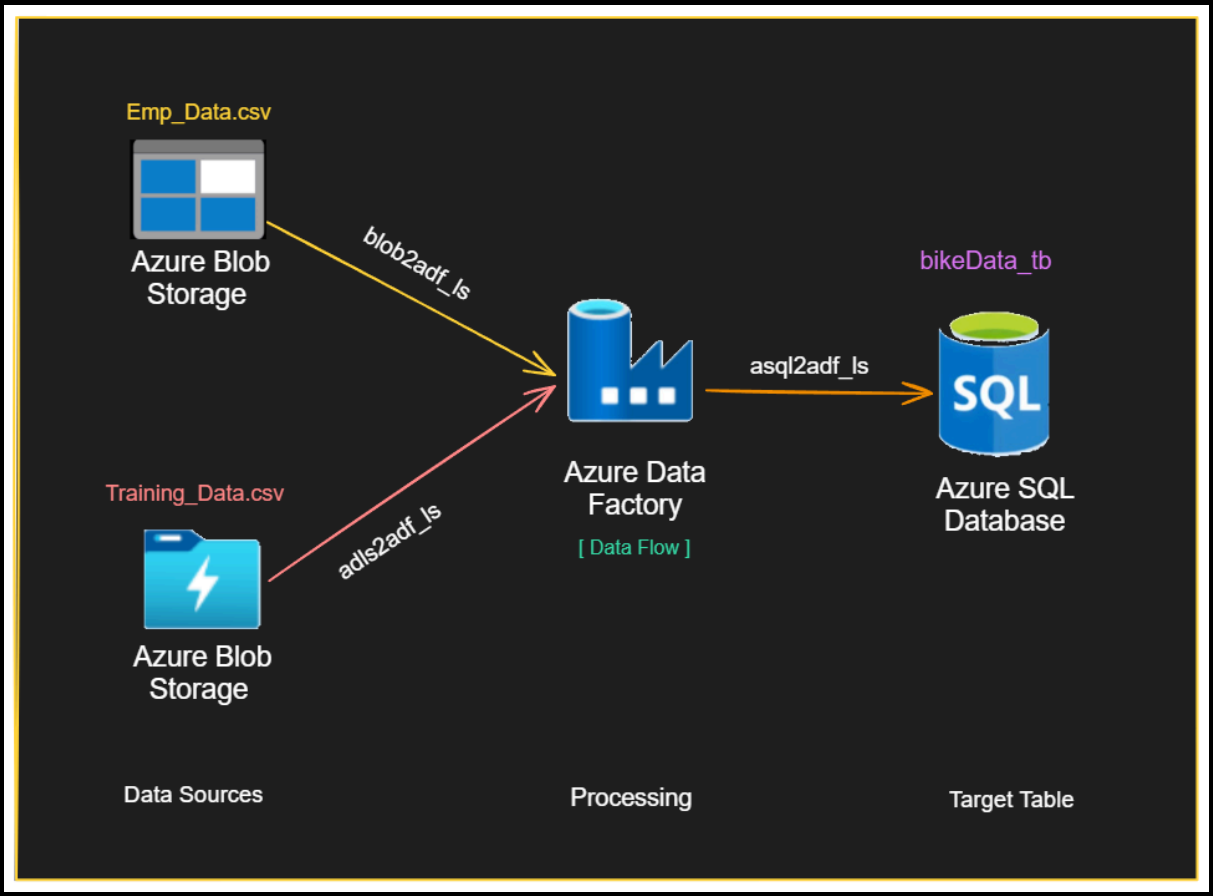
4th pipeline (blob,datalake2asql) using dataflow

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

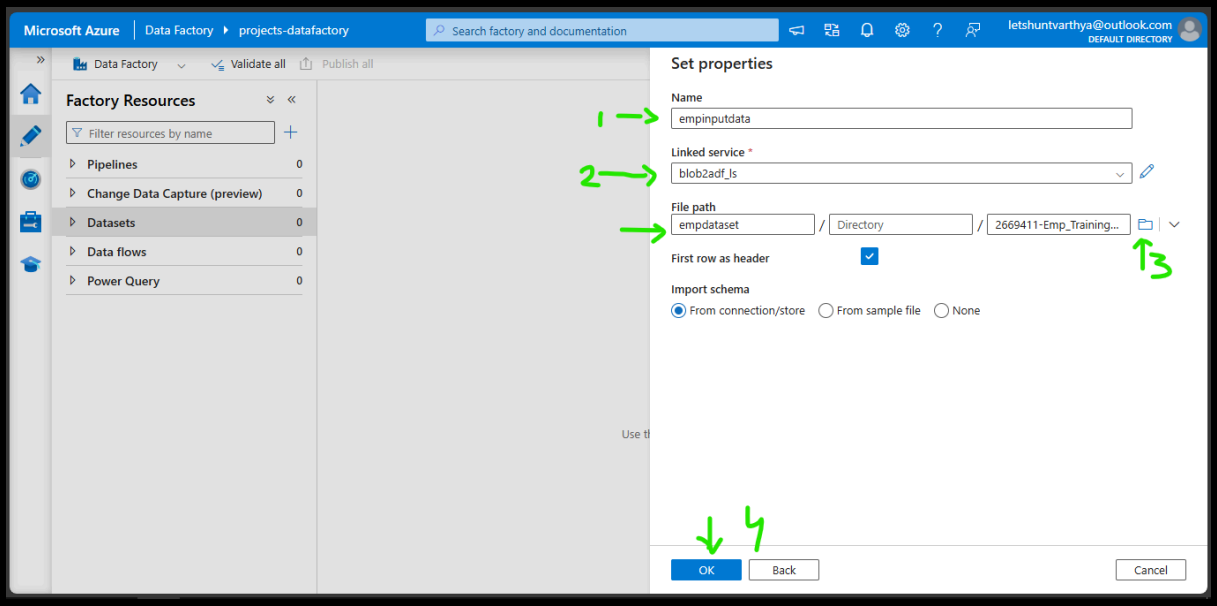
Azure SQL Database Creation :

- 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

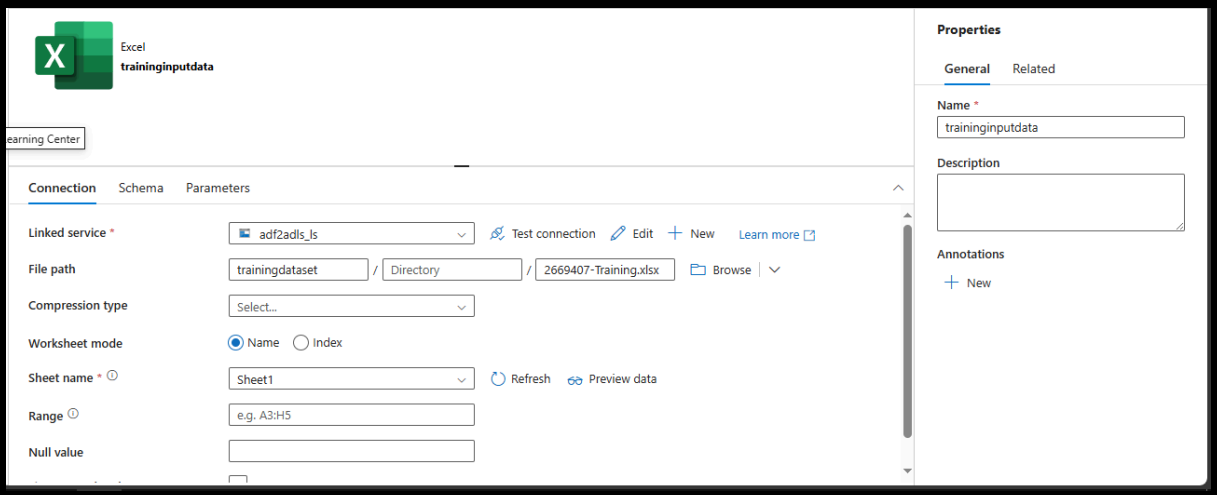
Azure Data Factory Creation :

- 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/1IpvA7XumJjbIP0wPWWUlfd_gdvh6jTn_0a12HWv0jcM/edit?usp=sharing
- Click on Go to resource.
- Click on launch workspace.

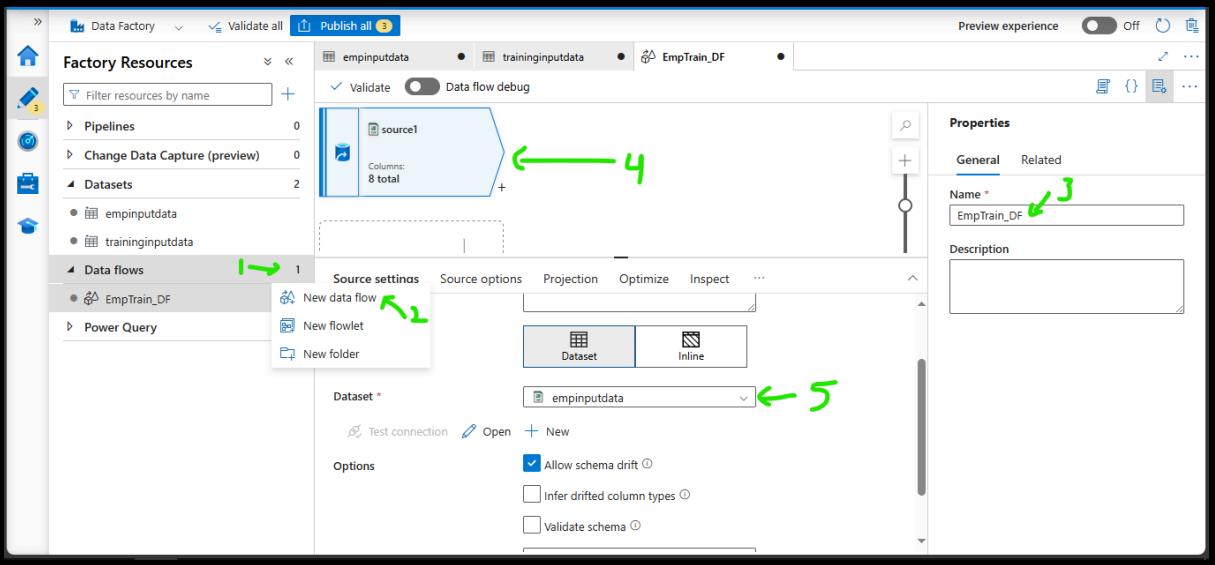
- Now let's do the transformations on data
- Open data factory and create datasets
- Pencil icon → datasets → new dataset → Azure blob Storage → csv → continue →

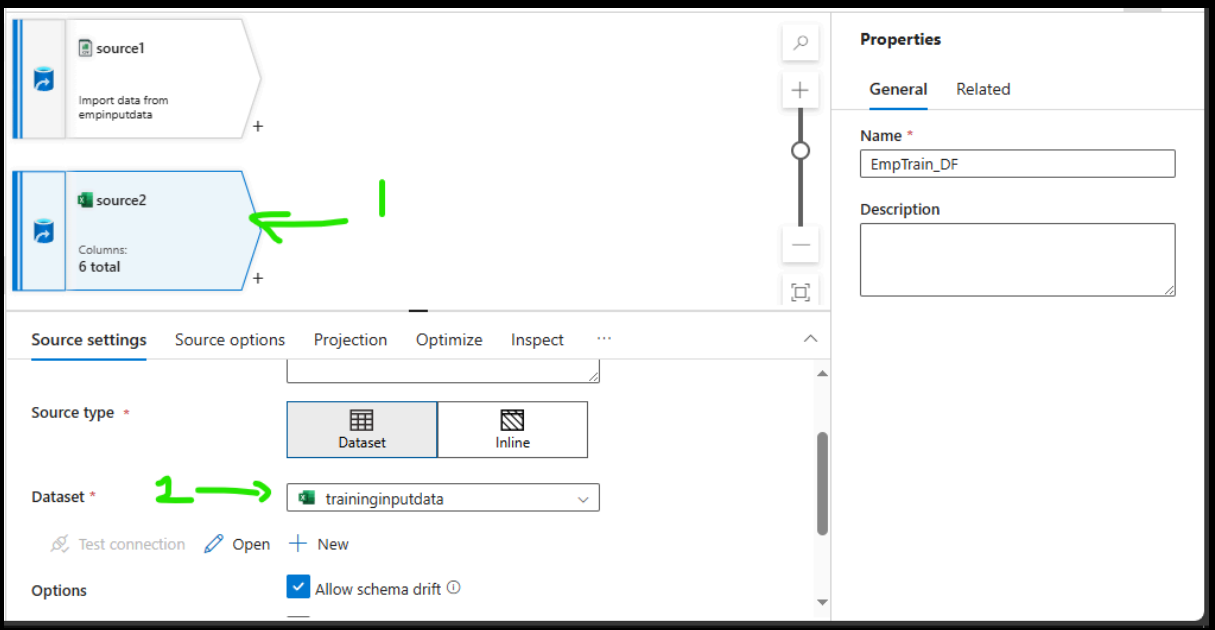


- Pencil icon → datasets → new dataset → Azure data lakeStorage → excel → continue →

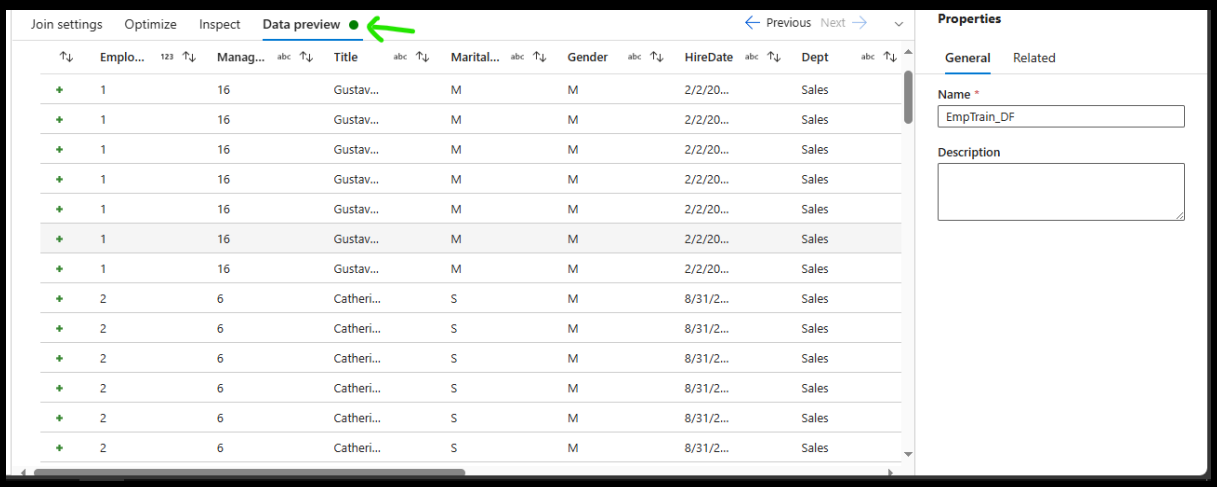


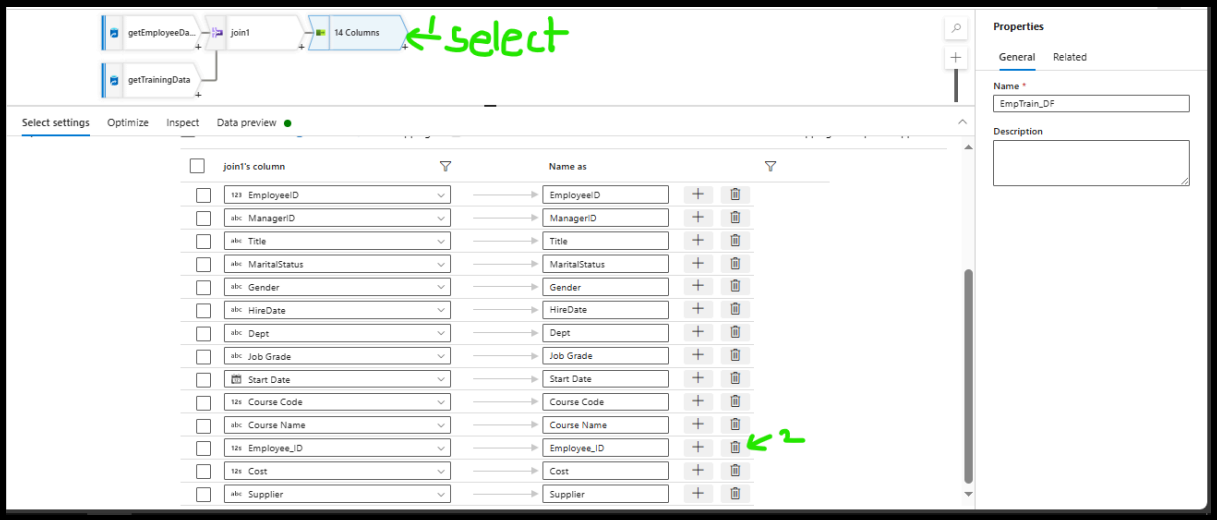
- Publish all —> publish

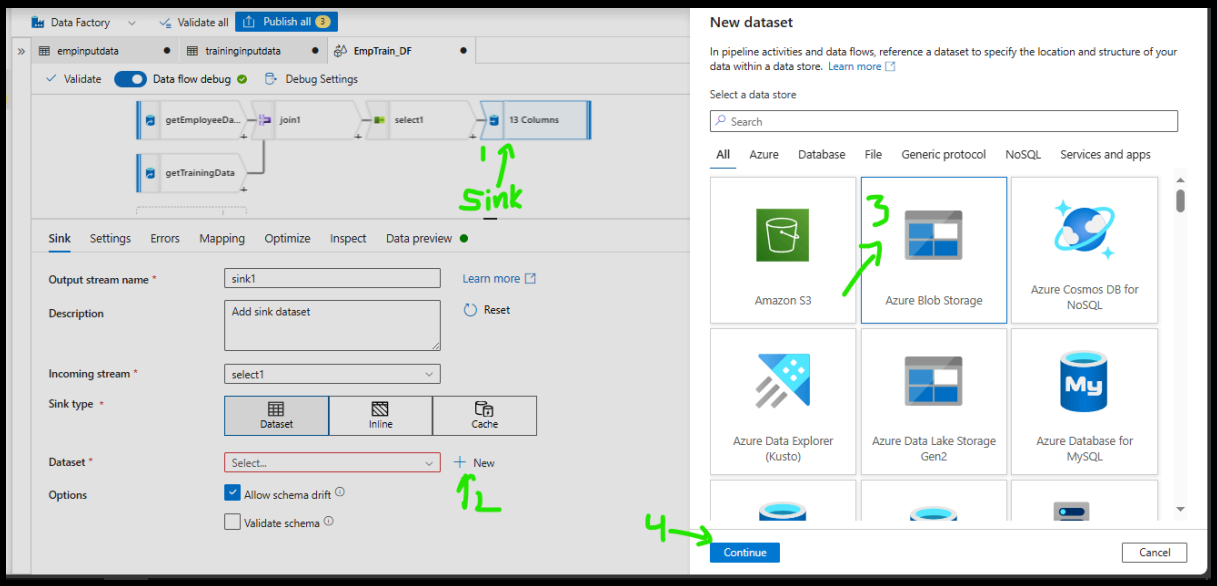


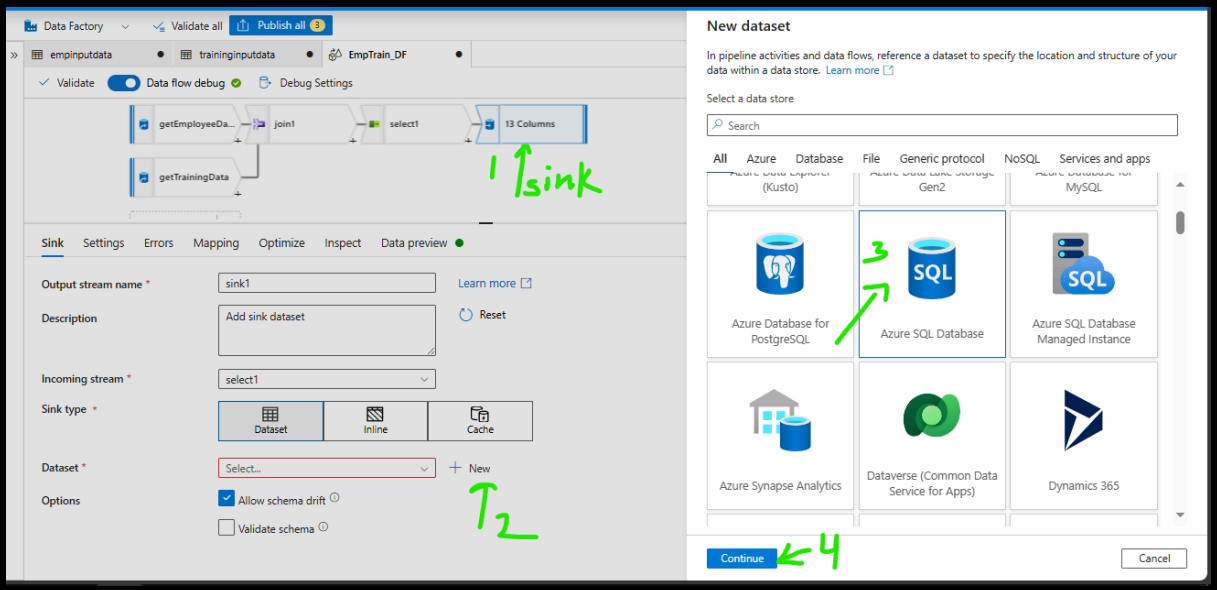


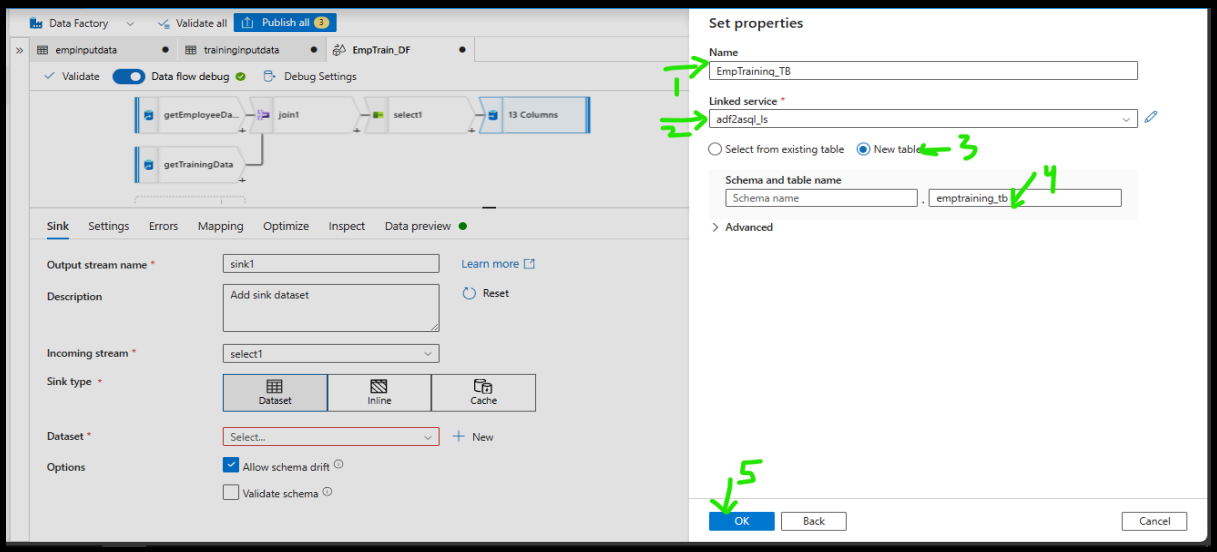
- Joining empinputdata with traininginputdata


Join settings
Optimize
Inspect
Data preview
Previous
Next
Properties
General
Related
Name *
EmpTrain_DF
Description
Emplo...
Manag...
Title
Marital...
Gender
HireDate
Dept
1 16 Gustav... M M 2/2/20... Sales
2 6 Catheri... S M 8/31/2... Sales

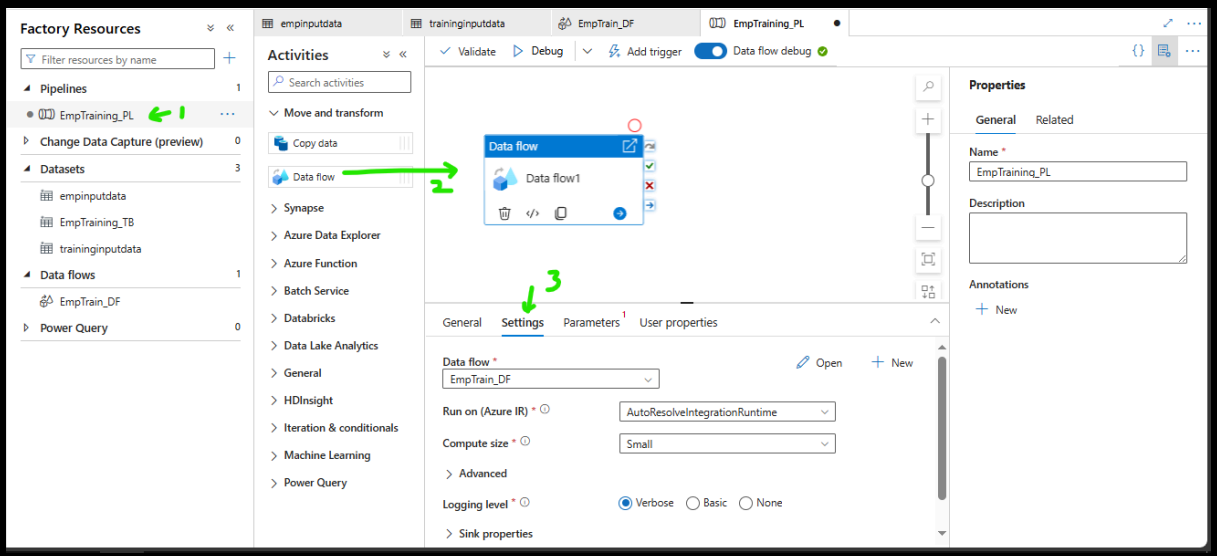

getEmployeeDa...
join1
14 Columns
select
getTrainingData
Select settings
Optimize
Inspect
Data preview
Properties
General
Related
Name *
EmpTrain_DF
Description
join1's column
Name as
EmployeeID
ManagerID
Title
MaritalStatus
Gender
HireDate
Dept
Job Grade
Start Date
Course Code
Course Name
Employee_ID
Cost
Supplier
2


Data Factory
Validate all
Publish all 3
empinputdata
traininginputdata
EmpTrain_DF
Validate
Data flow debug
Debug Settings
getEmployeeDa...
join1
select1
13 Columns
getTrainingData
sink
1
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
2
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
3
Azure Blob Storage
Azure Cosmos DB for NoSQL
Azure Data Explorer (Kusto)
Azure Data Lake Storage Gen2
Azure Database for MySQL
4
Continue
Cancel





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



- Publish all → publish
- Add trigger -> trigger now