STATEMENT 4

Create a pipeline to read the Customer table data from SQL and Customer Address data from CSV, join both of them, and then save the result where customer id> 1000 & Customer id <2000 in ascending order as a Parquet file.

In this statement/problem, we need to combine data from two different datasets, customer and customer address. Records where customer id is between 1000 and 2000 will be joined and copied into parquet file.

Create Datasets:

1. Create Customer Table
   We had created customer table for statement 1 and statement 2. The same table will be used in this statement as well.
   The SQL scripts for creating and inserting data into the table can be found in **customerTable.sql** and **insertCustomers.sql**.

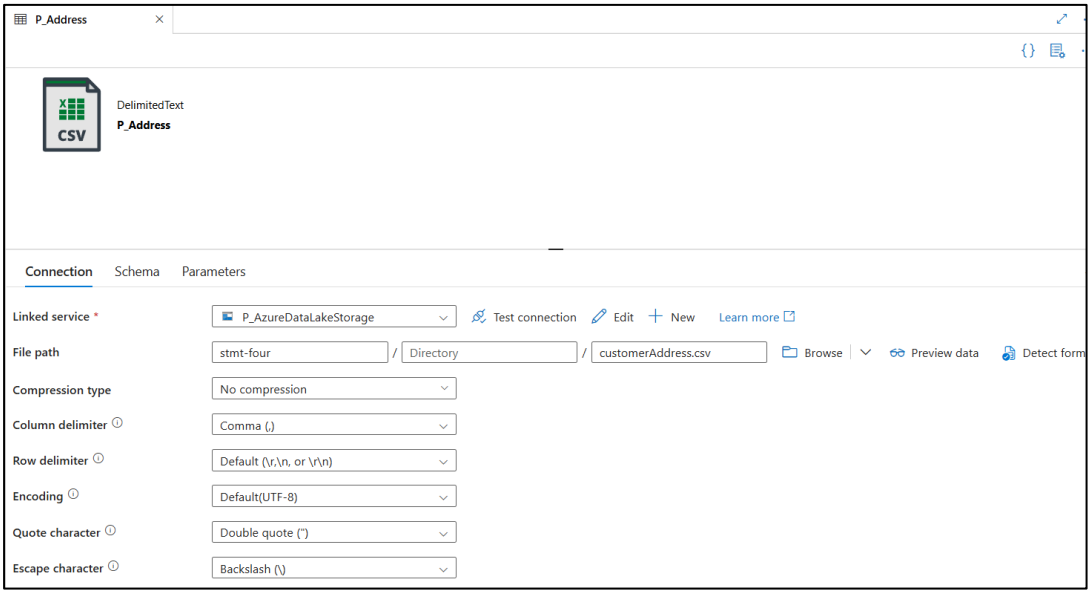
2. Create Customer Address CSV file
   The Customer Address CSV file contains the address of some of the cutomers along with their customer IDs. The data can be found in **customerAddress.csv**

customerAddress.csv U ×

customerAddress.csv

```
CustomerID,Street,City,State,PostalCode
1001,12 MG Road,Bangalore,Karnataka,560001
1002,45 Anna Salai,Chennai,Tamil Nadu,600002
1003,78 Park Street,Kolkata,West Bengal,700016
1004,21 FC Road,Pune,Maharashtra,411004
1005,88 Banjara Hills,Hyderabad,Telangana,500034
1006,15 Marine Drive,Mumbai,Maharashtra,400020
1007,33 Nehru Place,Delhi,Delhi,110019
1008,56 IT Park,Chandigarh,Chandigarh,160101
1009,102 Ashram Road,Ahmedabad,Gujarat,380009
1010,23 VIP Road,Raipur,Chhattisgarh,492001
```

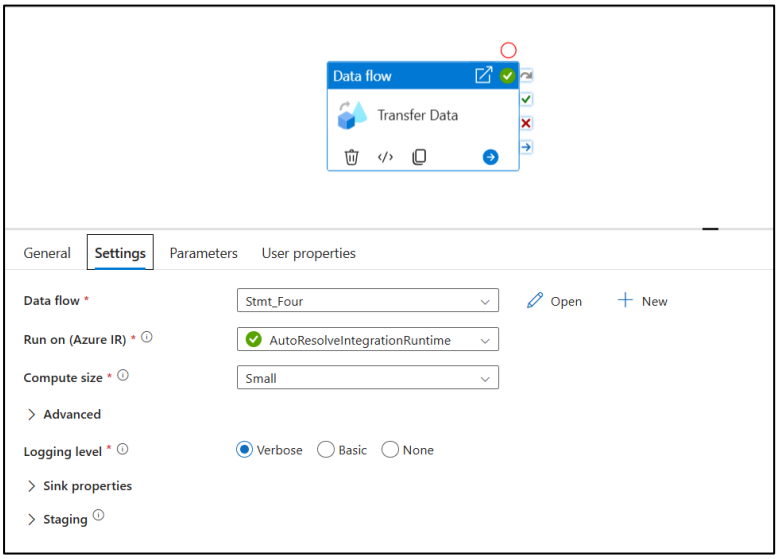
Upload the CSV file in the storage container and create a new dataset for accessing the file.



The dataset for CSV file has been imported from **stmt-four** directory present in the cprojectde storage account as shown in the above snapshot.

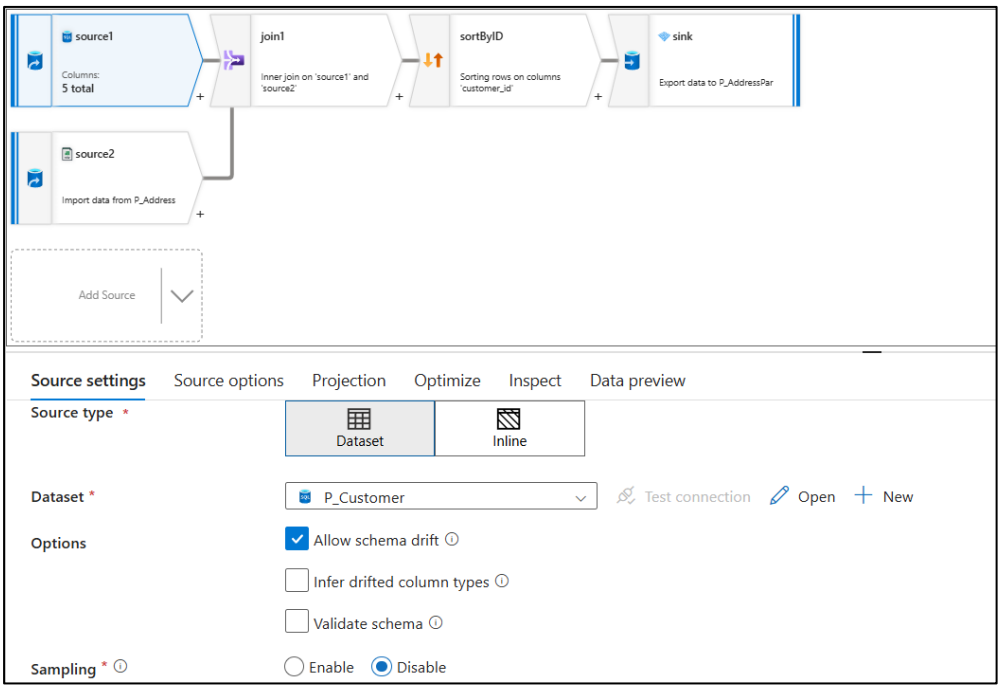
Step-Wise Guidelines:

1. Create a **Data Flow** activity
   Create new Data Flow

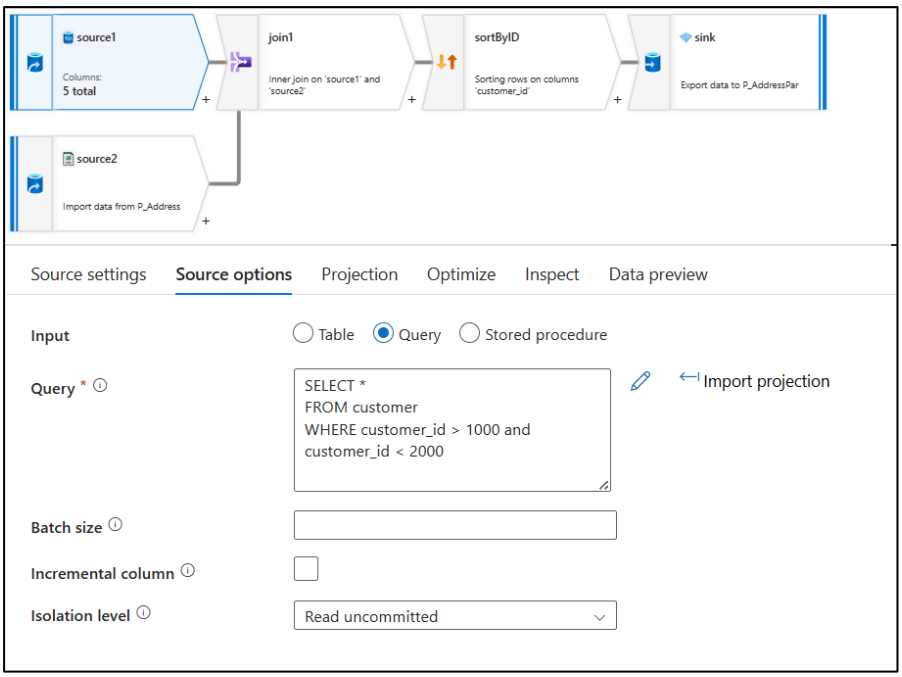


Data Flow:

1. Add Source1



The customer data is present in the customer table, which has been used in statement 1 and statement 2.



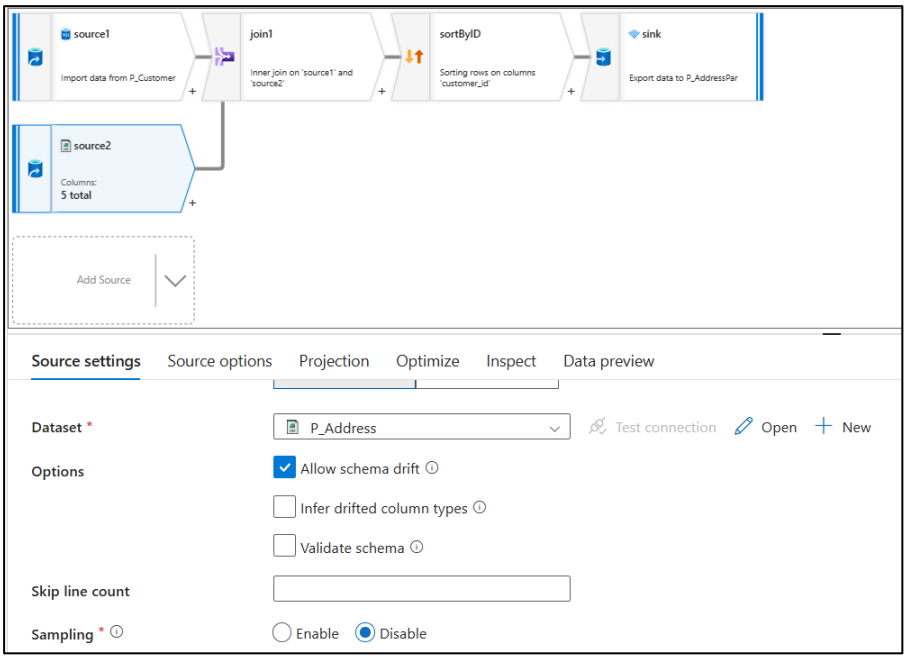
Go to **Source Options** tab and insert the following query to select record where customer id is between 1000 and 2000:

SELECT *
FROM customer
WHERE customer_id > 1000 and customer_id < 2000

The query can be found in **query.sql** script.

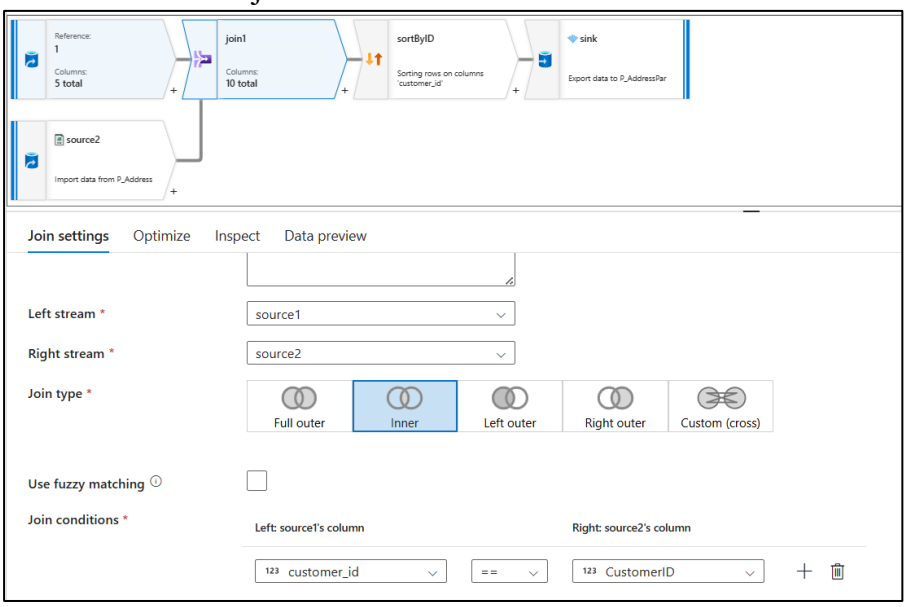
Go to **Projection** tab and import projection.

2. Add Source2



Source2 is the Customer Address CSV file.
Go to **Projection** tab and import the schema.

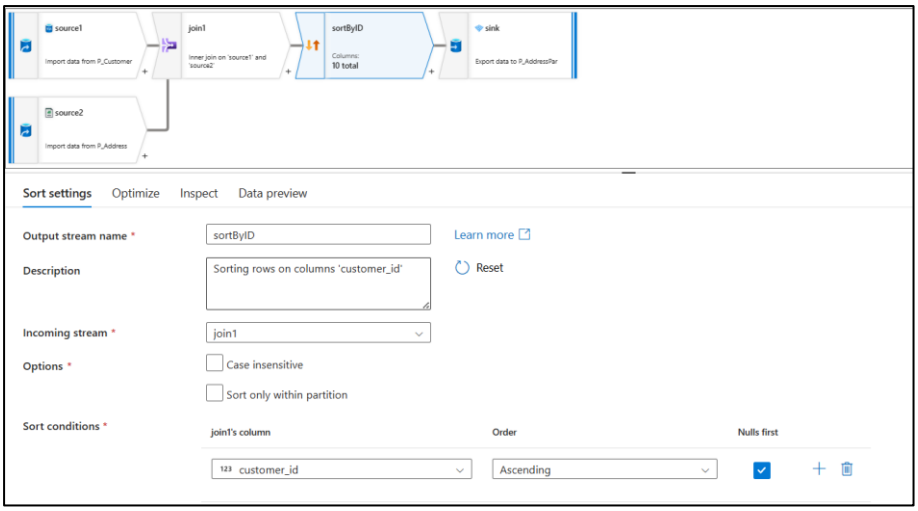
3. Click on + to add a join between the two datasets.



Insert the Left Stream, i.e, left table and Right Stream, i.e, right table from dropdown.

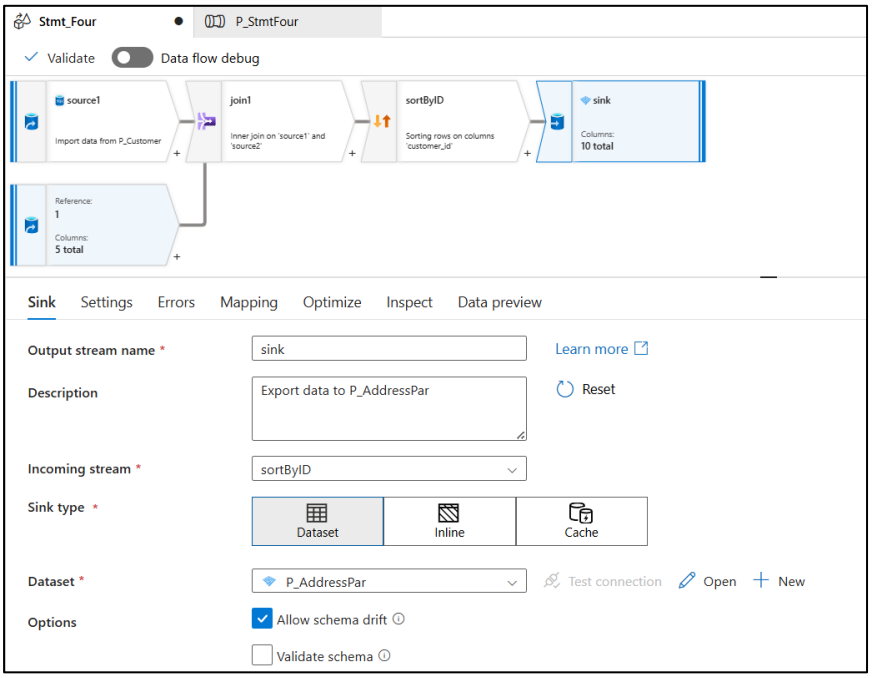
Specify the join condition, in this case, data has to matched according to Customer ID.

4. Add **Sort** Row Modifier (Optional)



Specify the column according to which data has to be sorted, customer_id in this case and sorting order, i.e, ascending.

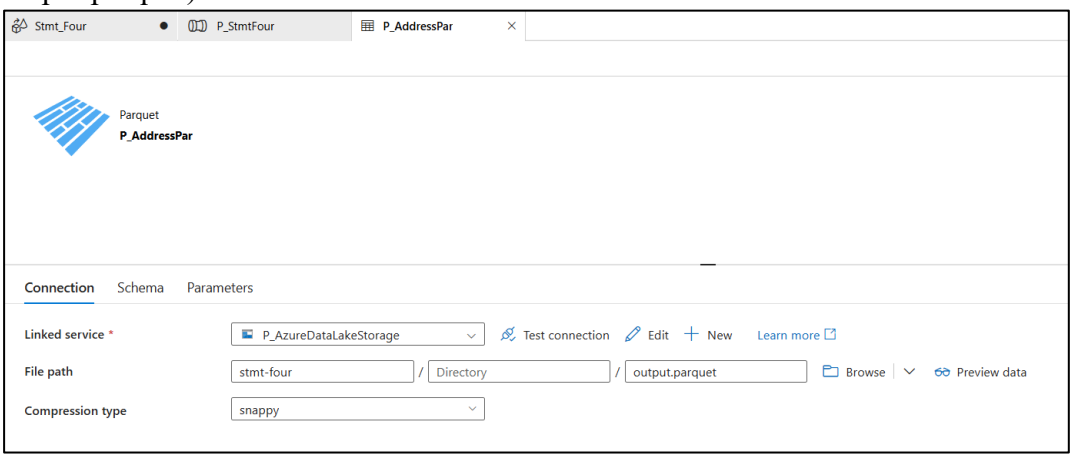
5. Add **Sink**



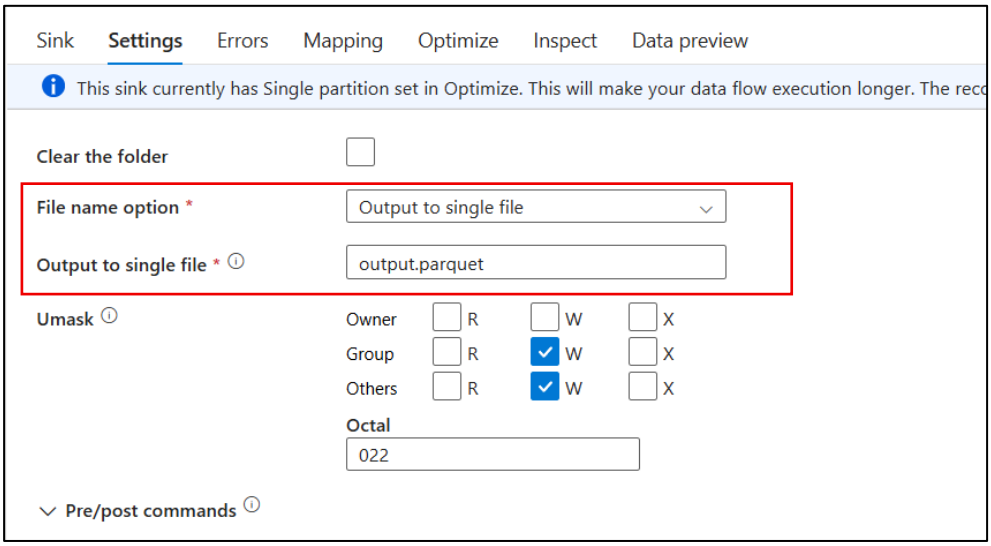
Create new dataset, using **New** option beside dataset.

Select Azure Data Lake Storage Gen2

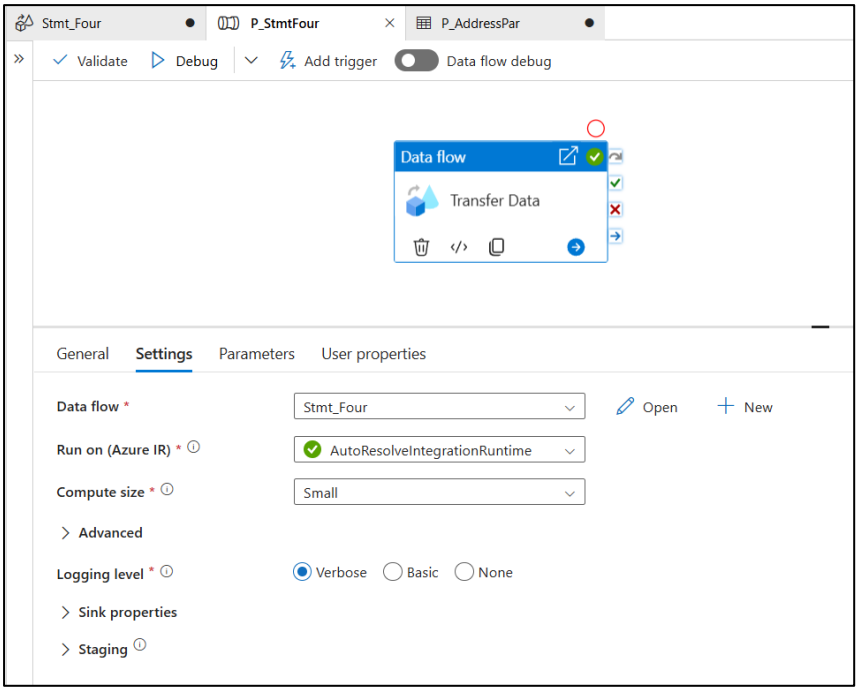
Parquet file format, name the dataset, select the ADLS linked service and specify the file system(in this scenario, stmt-four) and file name(in this scenario, output.parquet)



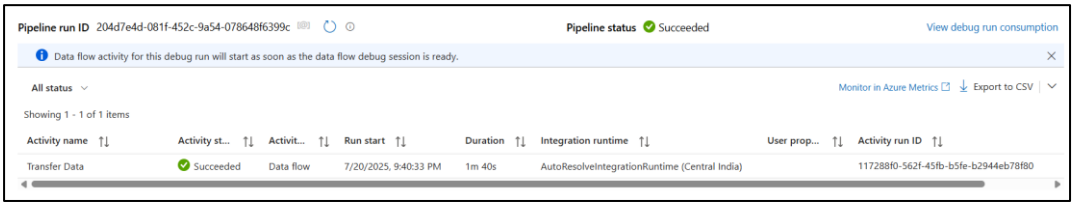
Go to **Settings** tab and select the following options:



Final Pipeline:



Pipeline Execution:



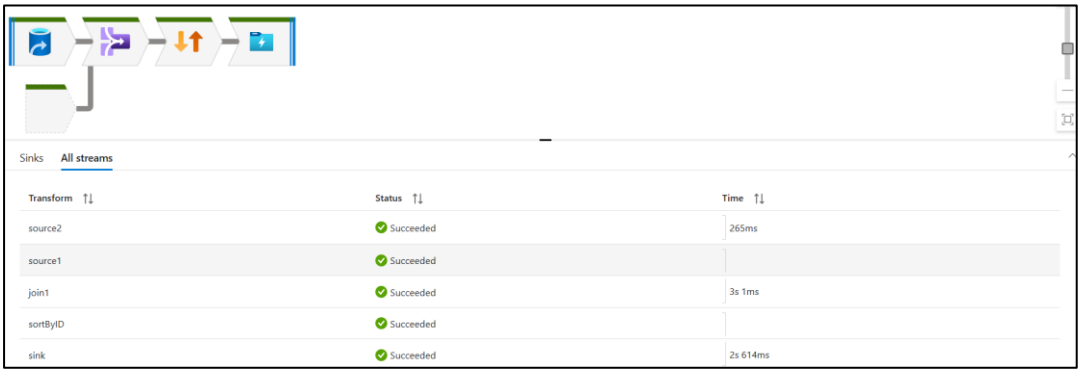
Pipeline run ID 204d7e4d-081f-452c-9a54-078648f6399c Pipeline status Succeeded View debug run consumption

Data flow activity for this debug run will start as soon as the data flow debug session is ready.

All status Monitor in Azure Metrics Export to CSV

Showing 1 - 1 of 1 items

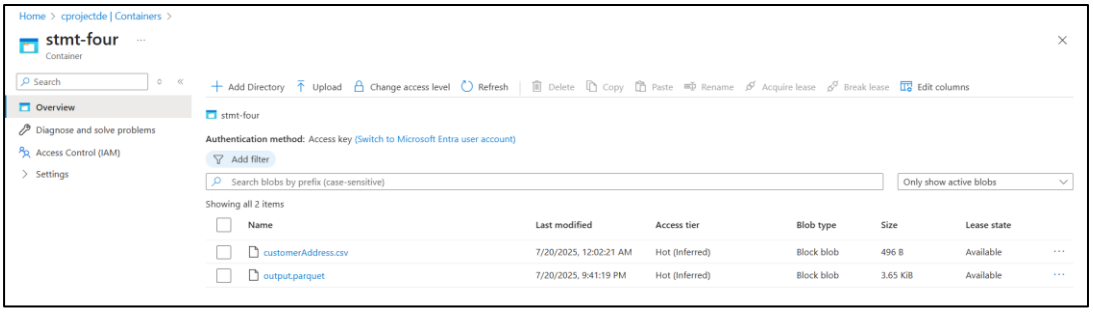
Activity name	Activity st...	Activit...	Run start	Duration	Integration runtime	User prop...	Activity run ID
Transfer Data	Succeeded	Data flow	7/20/2025, 9:40:33 PM	1m 40s	AutoResolveIntegrationRuntime (Central India)		117288f0-562f-45fb-b5fe-b2944eb78f80



Sinks All streams

Transform	Status	Time
source2	Succeeded	265ms
source1	Succeeded	
join1	Succeeded	3s 1ms
sortByID	Succeeded	
sink	Succeeded	2s 614ms

Parquet File



Home > cprojectde | Containers >

stmt-four
Container

Search
Overview
Diagnose and solve problems
Access Control (IAM)
Settings

Add Directory Upload Change access level Refresh Delete Copy Paste Rename Acquire lease Break lease Edit columns

stmt-four

Authentication method: Access key (Switch to Microsoft Entra user account)

Add filter

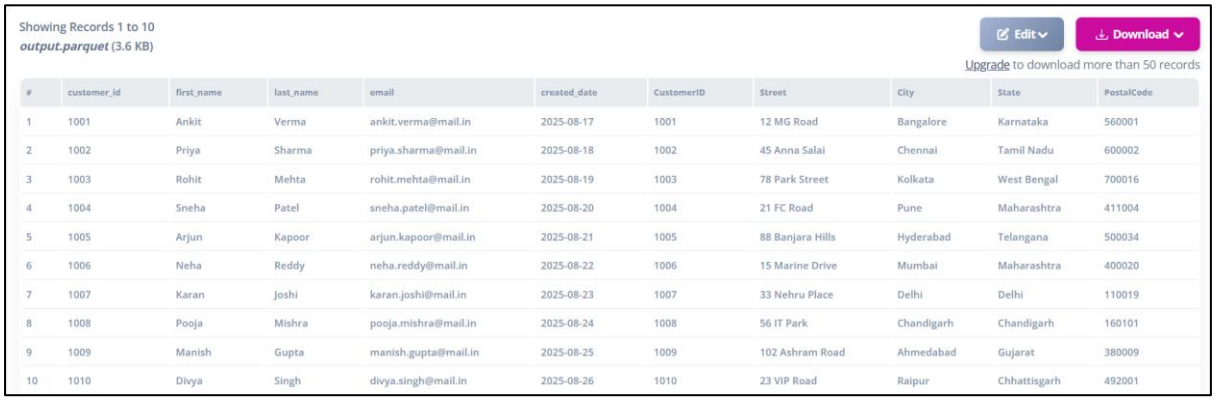
Search blobs by prefix (case-sensitive) Only show active blobs

Showing all 2 items

Name	Last modified	Access tier	Blob type	Size	Lease state
customerAddress.csv	7/20/2025, 12:02:21 AM	Hot (Inferred)	Block blob	496 B	Available
output.parquet	7/20/2025, 9:41:19 PM	Hot (Inferred)	Block blob	3.65 KiB	Available

The file **output.parquet** has been created in **stmt-four** directory of cprojectde storage account..

Parquet File Data



Showing Records 1 to 10
output.parquet (3.6 KB)

Edit Download

Upgrade to download more than 50 records

#	customer_id	first_name	last_name	email	created_date	CustomerID	Street	City	State	PostalCode
1	1001	Ankit	Verma	ankit.verma@mail.in	2025-08-17	1001	12 MG Road	Bangalore	Karnataka	560001
2	1002	Priya	Sharma	priya.sharma@mail.in	2025-08-18	1002	45 Anna Salai	Chennai	Tamil Nadu	600002
3	1003	Rohit	Mehta	rohit.mehta@mail.in	2025-08-19	1003	78 Park Street	Kolkata	West Bengal	700016
4	1004	Sneha	Patel	sneha.patel@mail.in	2025-08-20	1004	21 FC Road	Pune	Maharashtra	411004
5	1005	Arjun	Kapoor	arjun.kapoor@mail.in	2025-08-21	1005	88 Banjara Hills	Hyderabad	Telangana	500034
6	1006	Neha	Reddy	neha.reddy@mail.in	2025-08-22	1006	15 Marine Drive	Mumbai	Maharashtra	400020
7	1007	Karan	Joshi	karan.joshi@mail.in	2025-08-23	1007	33 Nehru Place	Delhi	Delhi	110019
8	1008	Pooja	Mishra	pooja.mishra@mail.in	2025-08-24	1008	56 IT Park	Chandigarh	Chandigarh	160101
9	1009	Manish	Gupta	manish.gupta@mail.in	2025-08-25	1009	102 Ashram Road	Ahmedabad	Gujarat	380009
10	1010	Divya	Singh	divya.singh@mail.in	2025-08-26	1010	23 VIP Road	Raipur	Chhattisgarh	492001