Name – Sujay Chavan

Mobile - +91 8600588141

Email – sujay.chavan@cognizant.com

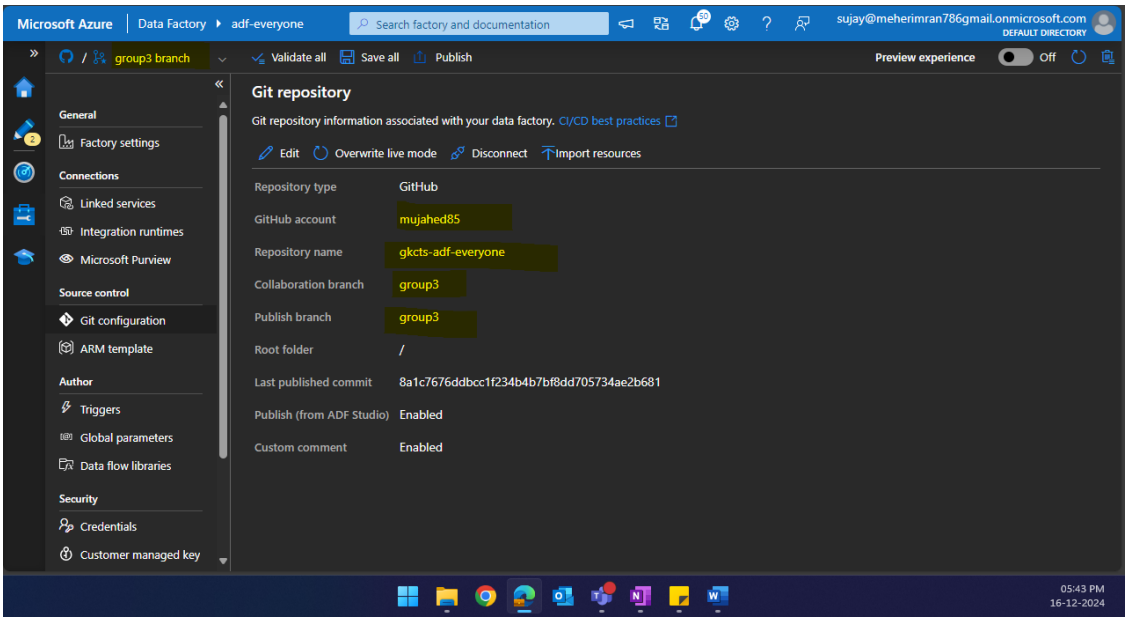
Group No. - 3

Initial Setup

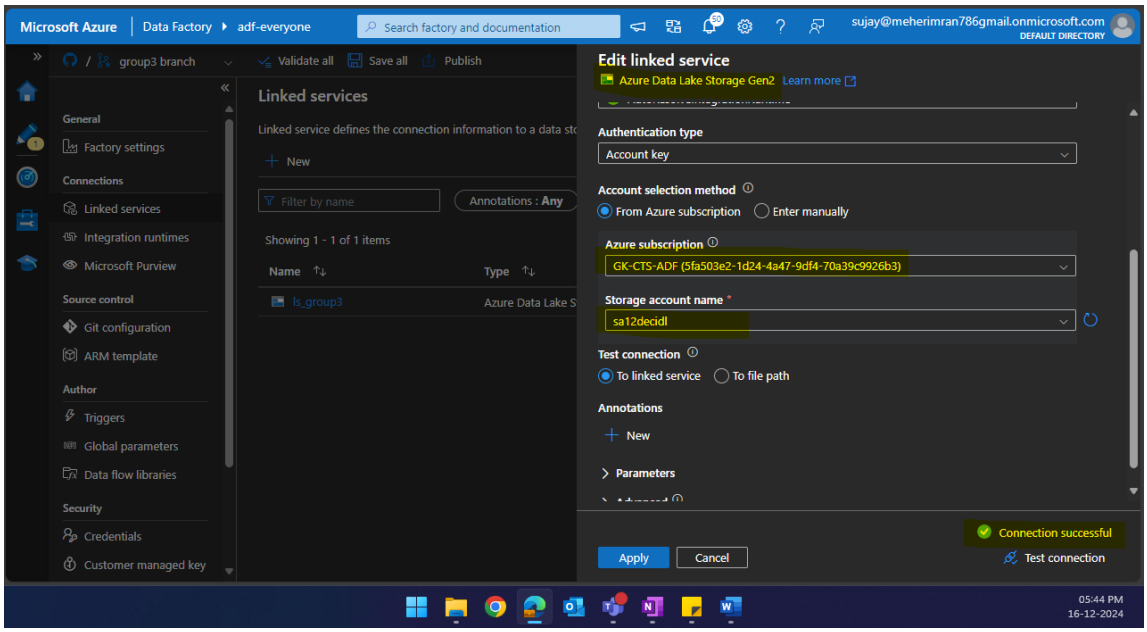
A) Git Configuration:

Collaboration Branch: group3

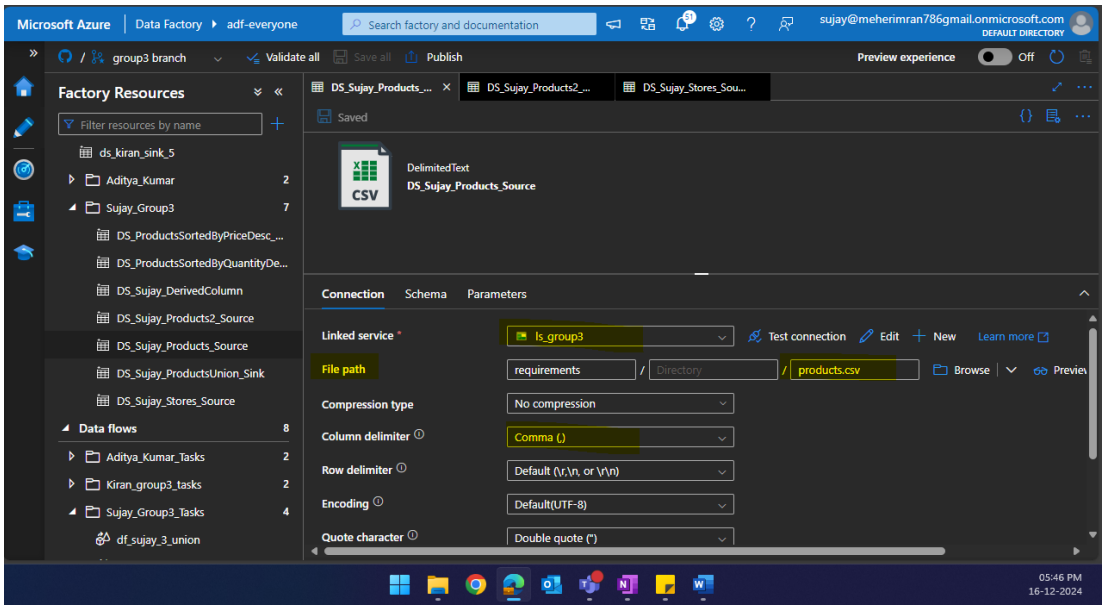
Pulish Branch: group3



B) Creating Linked Service to 'sa12decidl':



C) Creating Datasets:



Similarly created datasets for source as well as target files.

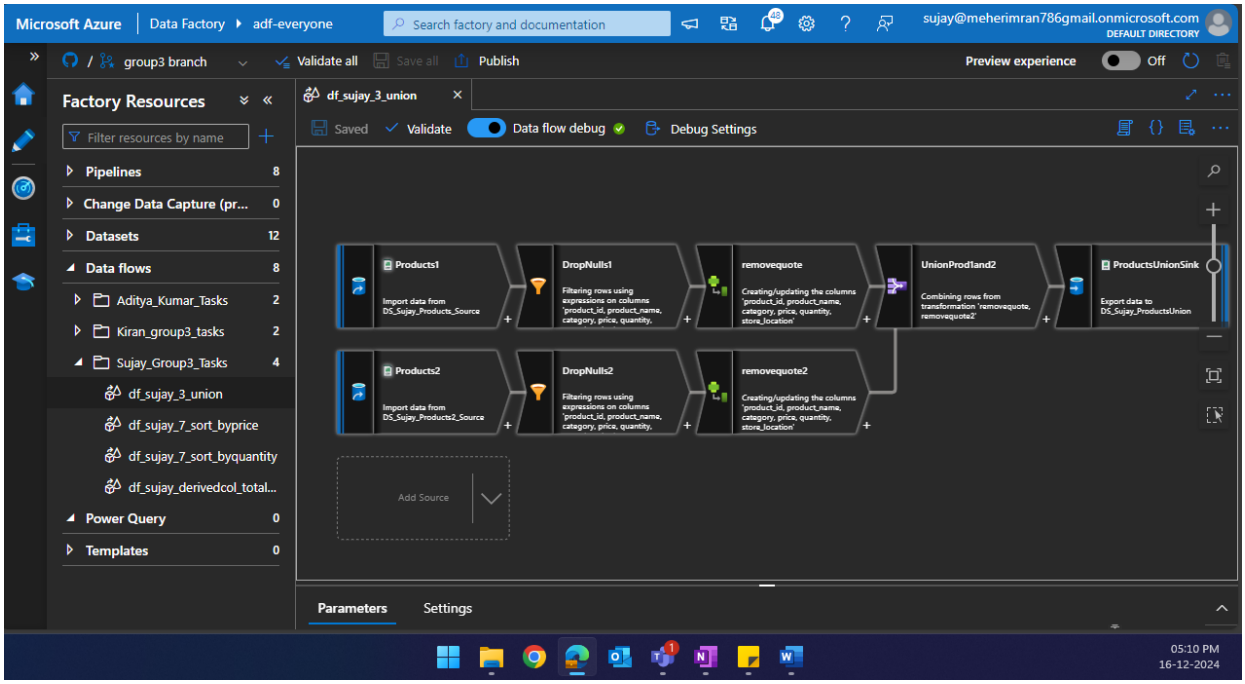
1. Union Transformation:

- **Use Case**: Combine two datasets (e.g., products from two different CSV files) with the same structure.
- **Example:** If we have two datasets with same schema (products.csv, products2.csv), we can use a **Union Transformation** to combine both datasets into a single output.

Product Dataset is stored in 'DS_Sujay_Products_Source'

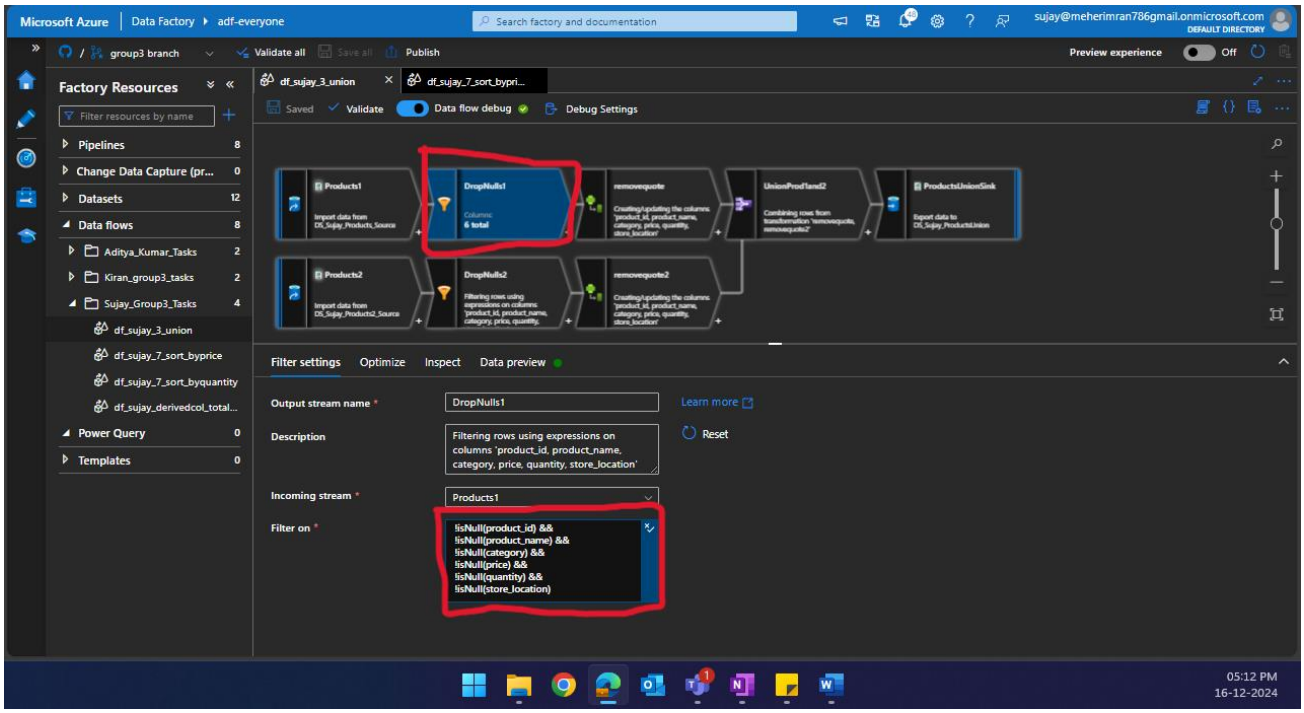
Product2 Dataset is stored in 'DS_Sujay_Products2_Source'

- **Data Flow Architecture for Union Transformation:**

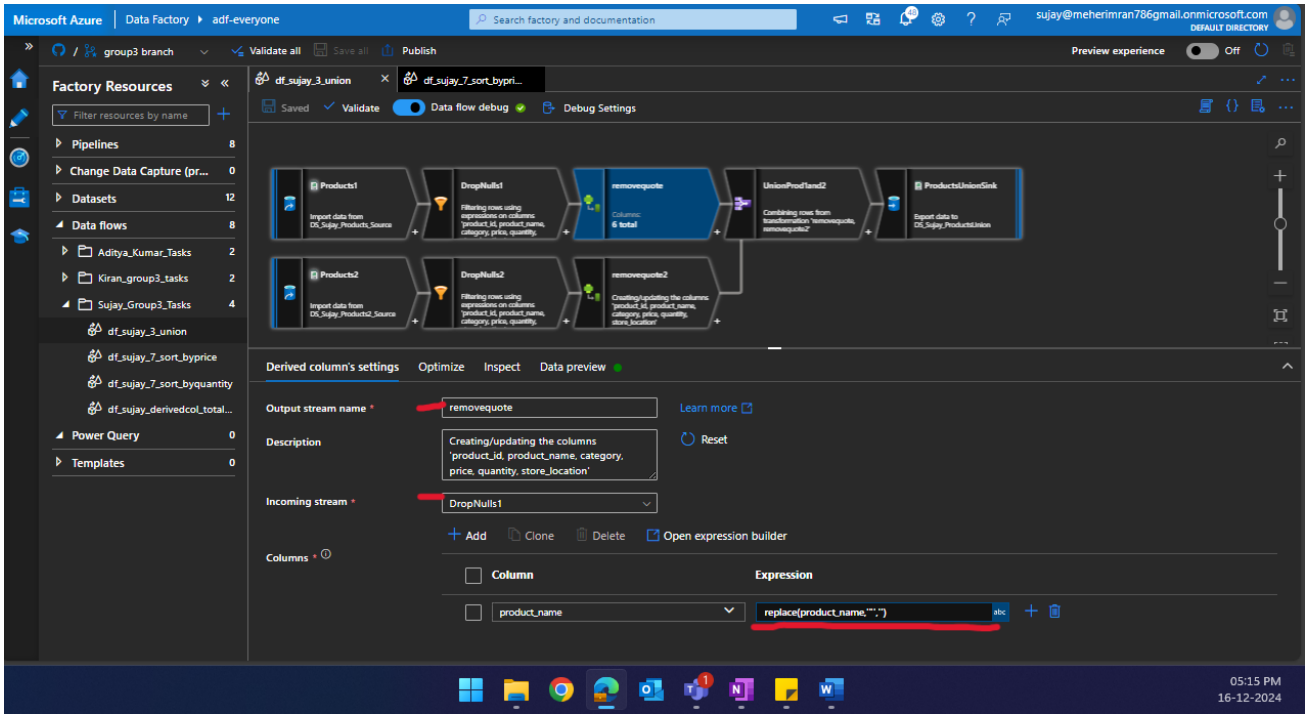


- To join Products and Products2, two source datasets are chosen in 1st step. Later nulls values are removed in DropNulls step. In further step, data is cleaned (extra double quotes are replaced with empty string). Union is performed in UnionProd1and2 and result is exported to DS_Sujay_ProductUnion dataset in Sink step.
- **DropNulls**:

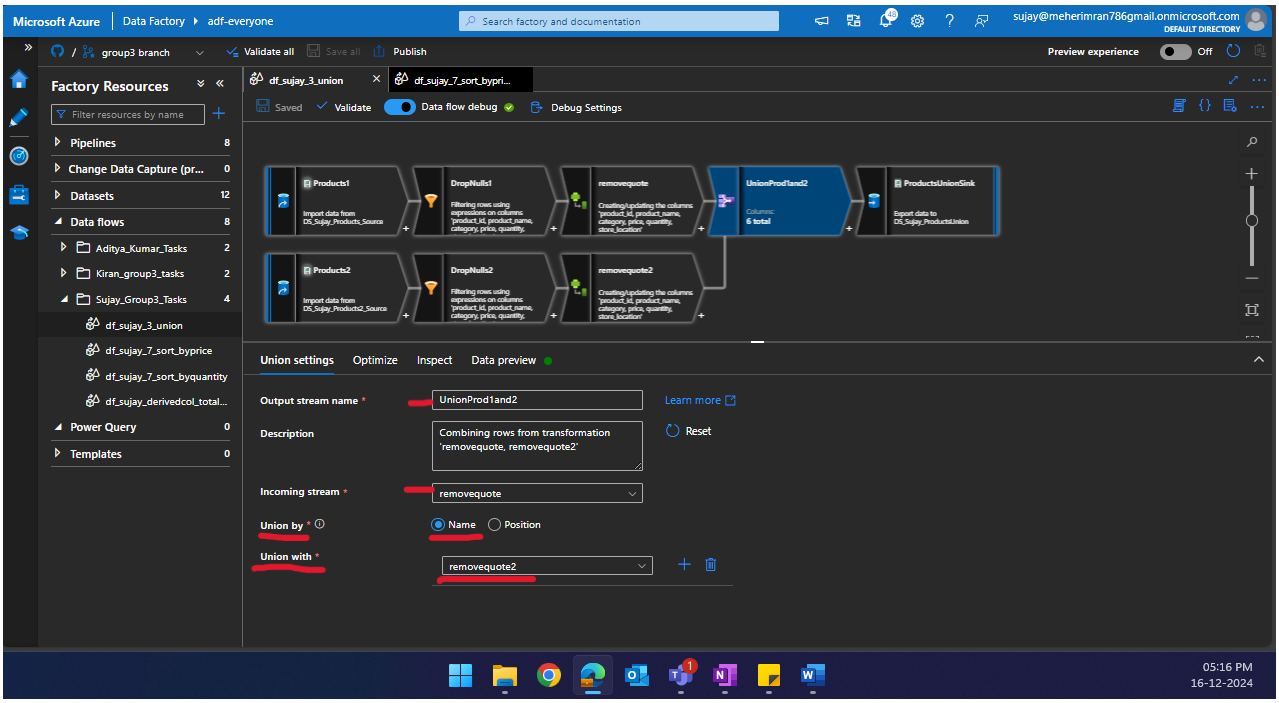
!isNull(product_id) &&
!isNull(product_name) &&
!isNull(category) &&
!isNull(price) &&
!isNull(quantity) &&
!isNull(store_location)



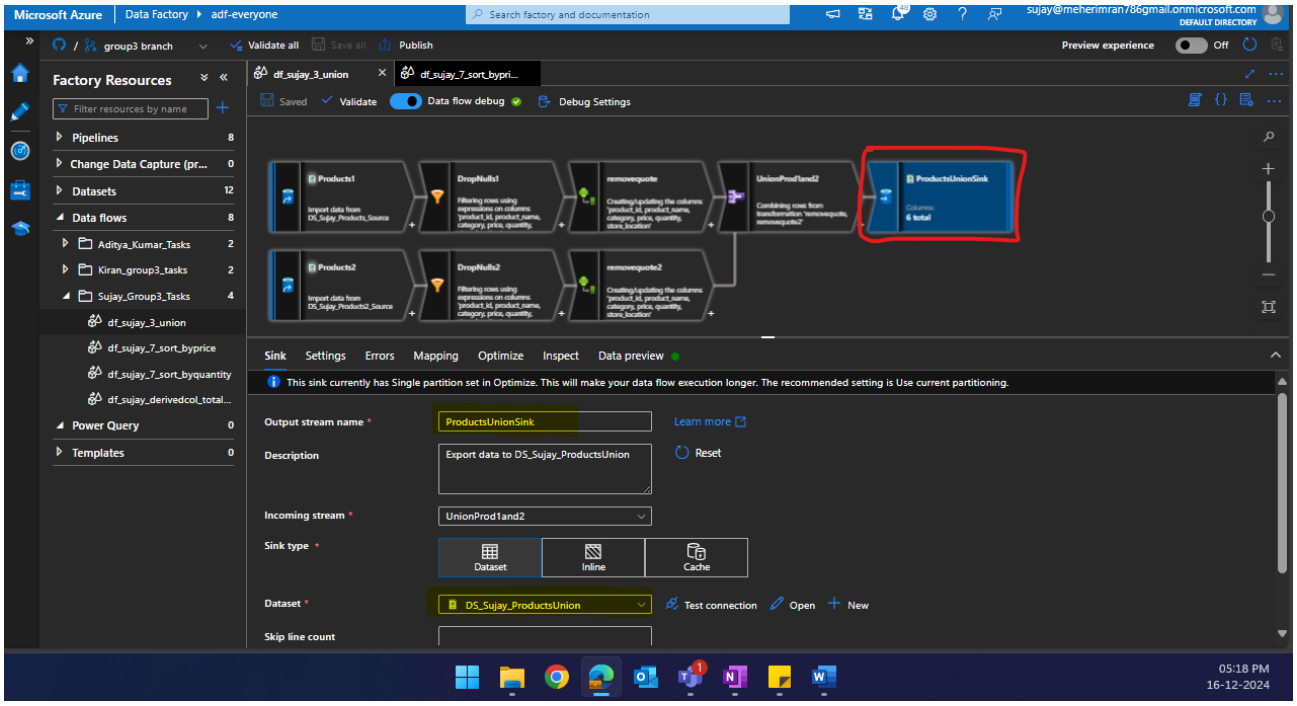
- **RemoveQuotes:** replace(product_name,"","") --Replace quotes with empty string.



- **Union:** Stream from Products1 is merged with stream from Products2 and passed to sink.



- **Export to Sink:** Result is exported to 'DS_Sujay_ProductsUnion' dataset.

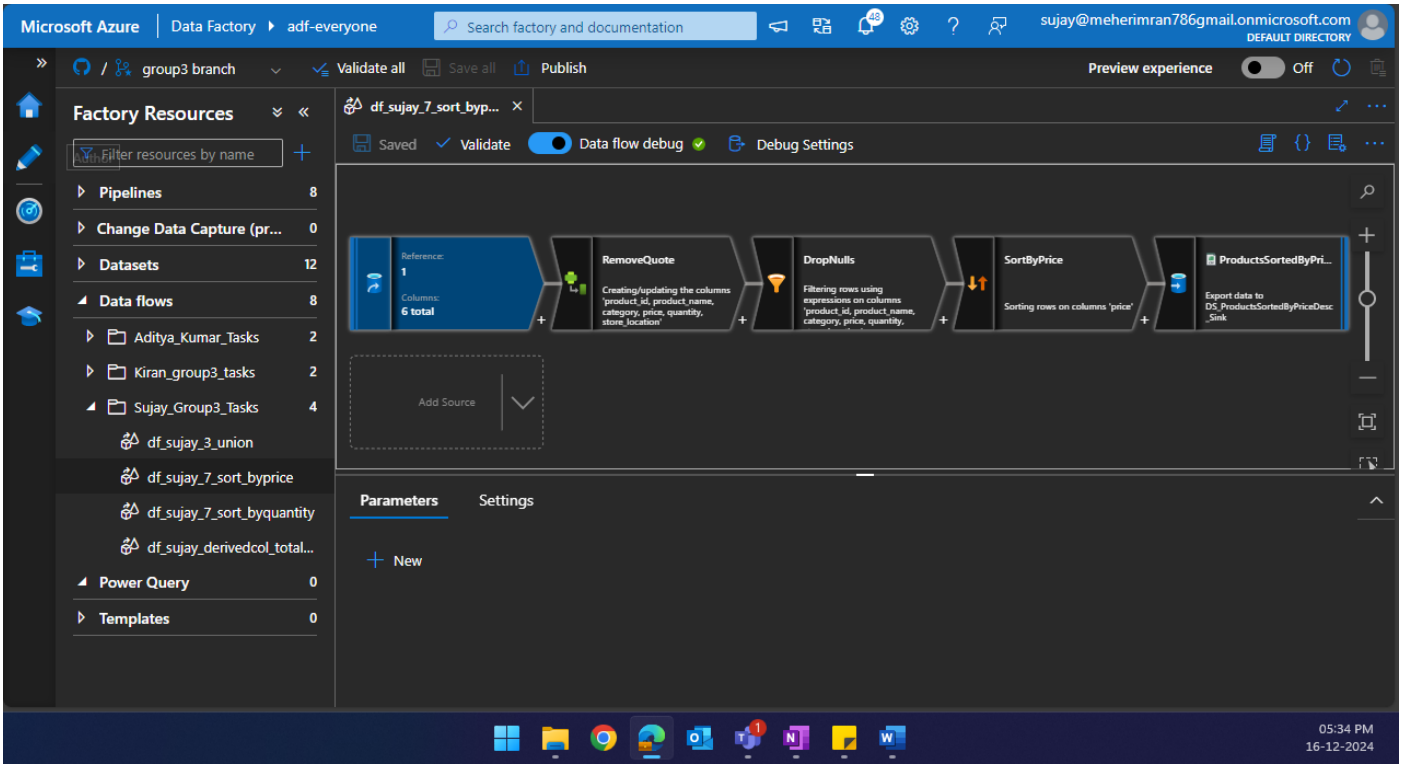


2. Sort Transformation:

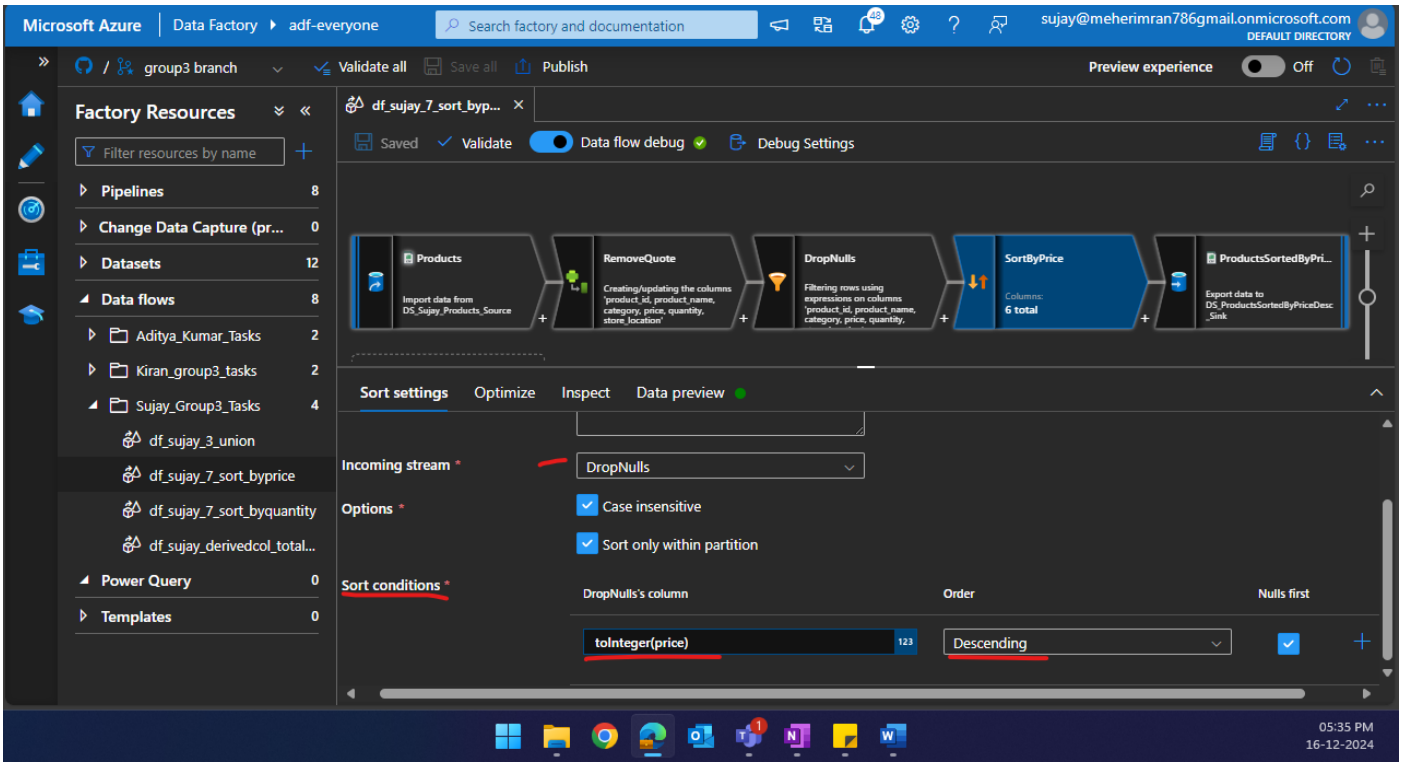
Use Case: Sort the data based on a specific column, such as sorting products by price or quantity.

Example: Sort products by price/quantity in descending order.

- **Data Flow Architecture for Sort Transformation:**



- **Sort by Price:** Here datatype of price column is converted from string to integer first and sorted in descending order.



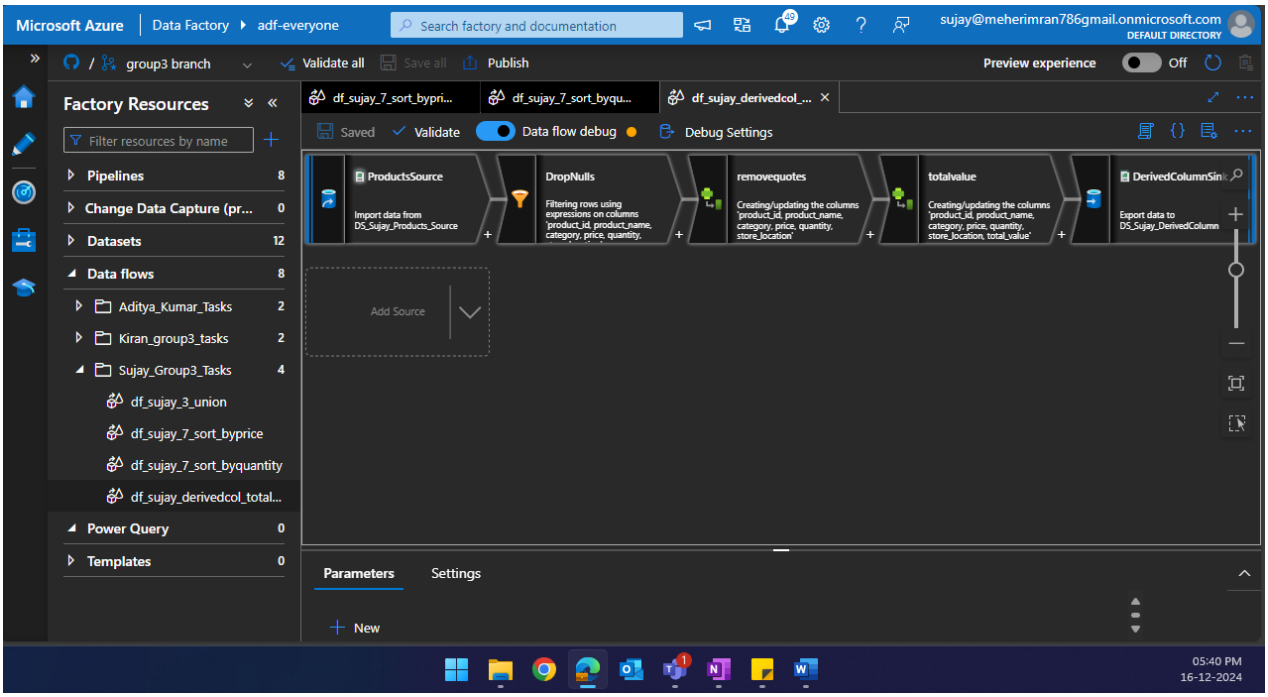
Same data flow is followed to sort data by product quantity. Only change is in above, instead of price column, quantity column is used.

3. Derived Column Transformation:

Use Case: Create new calculated columns from existing columns.

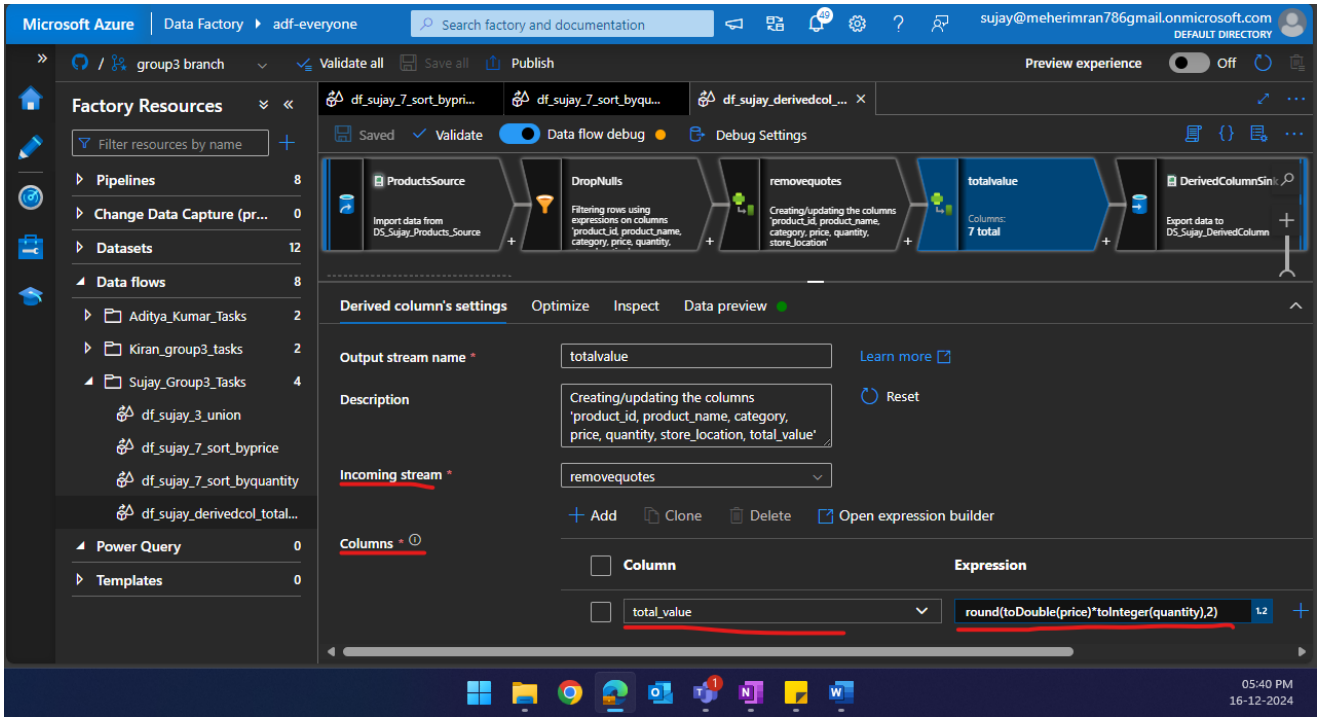
Example: Create a new column `total_value` by multiplying `price` and `quantity` (i.e., the total value of each product).

- **Data Flow Architecture for Derived Column Transformation:**



- **Creating derived column 'total_value':**

'total_value' column is created by multiplying price with quantity column. To do so, datatype of price and quantity is first converted to double and int respectively and result is rounded upto 2 decimal places.

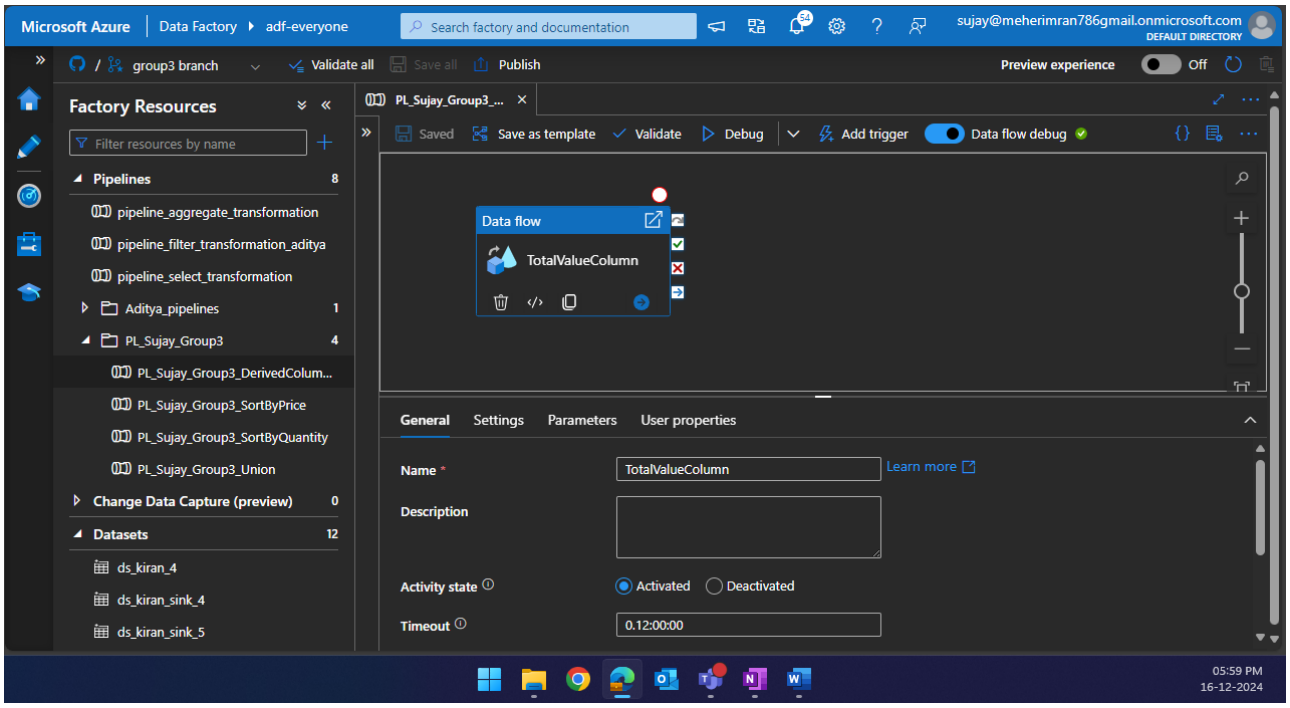


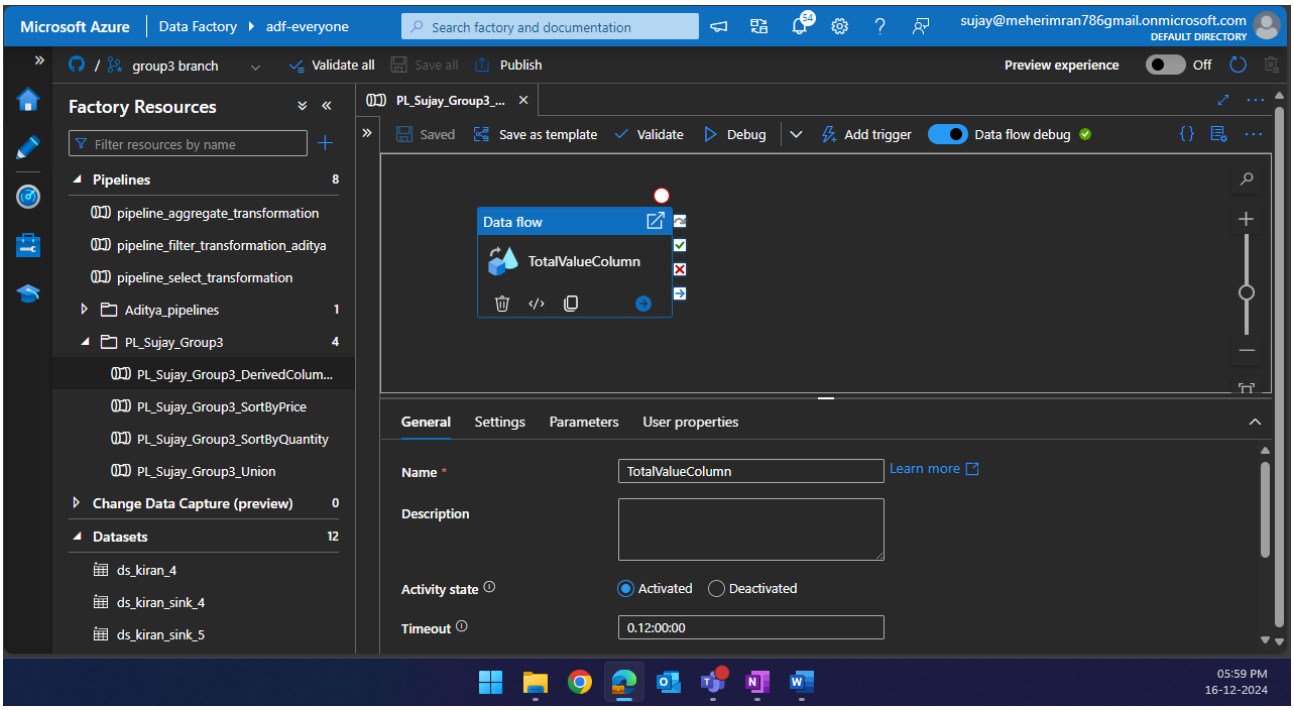
The steps: Products dataset as source, Dropping null values, removing extra double quotes and exporting results to respective dataset are same for all data flows. After creating data flows, the next step is to create a pipeline to execute data flow. This can be done using pipelines option in factory resources.

4. Creating Pipeline:

Pipeline is created using pipelines in factory resources. Required options are included using drag and drop feature and corresponding settings are made as per functionality. To create a data flow pipeline, data flow option is selected, and respective data flow option is chosen for required pipeline.

- Creating data flow pipeline for derived column transformation:

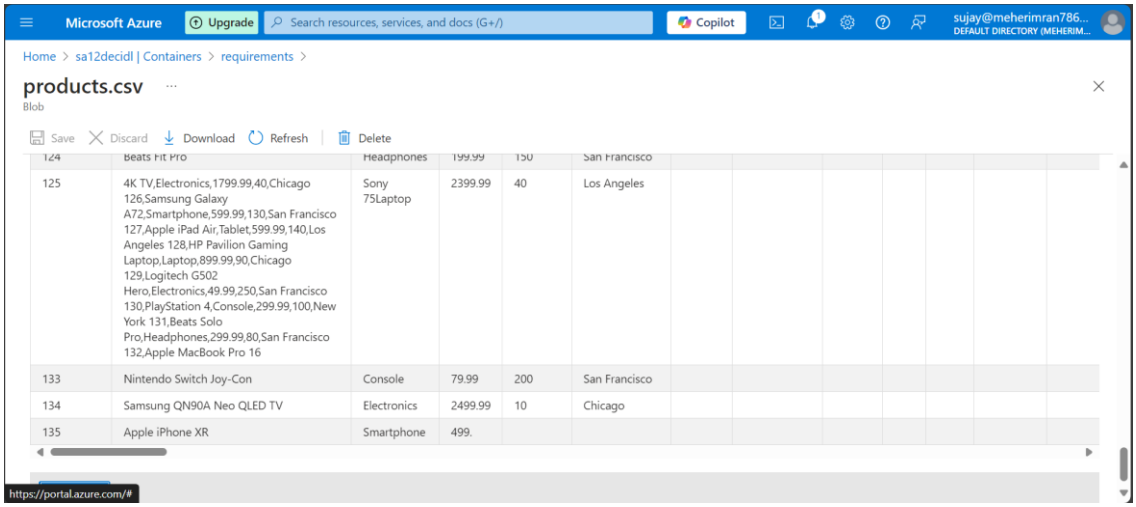




Similar pipelines are created for all the tasks and run using 'Debug' option. As per settings configured, outputs for all pipelines are stored in Azure Data Lake Gen2 storage.

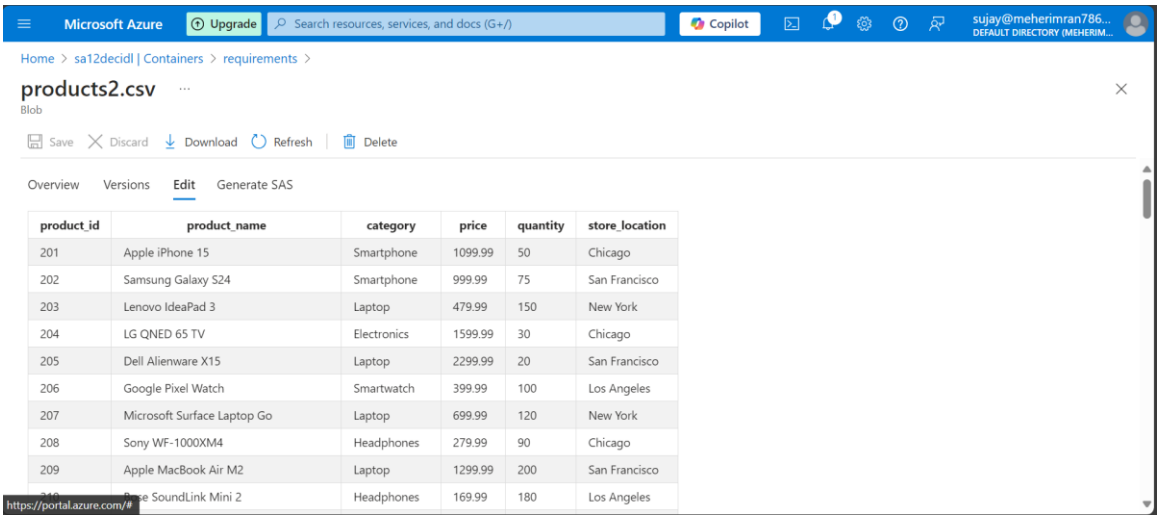
5. Input files:

- products.csv input file:



124	Beats Fit Pro	Headphones	199.99	150	San Francisco
125	4K TV,Electronics,1799.99,40,Chicago 126,Samsung Galaxy A72,Smartphone,599.99,130,San Francisco 127,Apple iPad Air,Tablet,599.99,140,Los Angeles 128,HP Pavilion Gaming Laptop,Laptop,899.99,90,Chicago 129,Logitech G502 Hero,Electronics,49.99,250,San Francisco 130,PlayStation 4,Console,299.99,100,New York 131,Beats Solo Pro,Headphones,299.99,80,San Francisco 132,Apple MacBook Pro 16	Sony 75Laptop	2399.99	40	Los Angeles
133	Nintendo Switch Joy-Con	Console	79.99	200	San Francisco
134	Samsung QN90A Neo QLED TV	Electronics	2499.99	10	Chicago
135	Apple iPhone XR	Smartphone	499.		

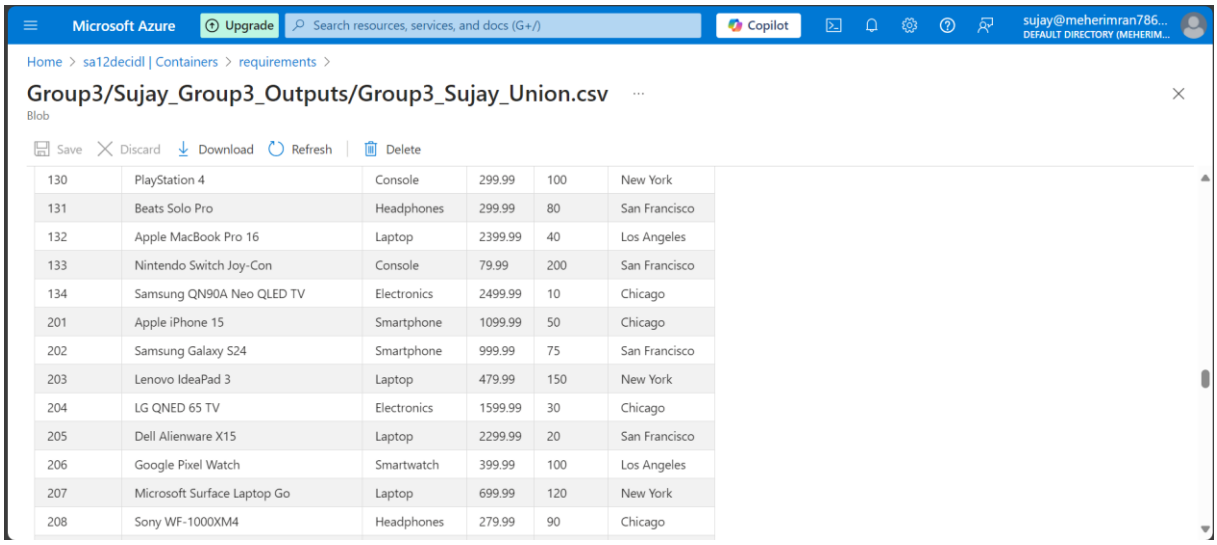
- products2.csv input file:



product_id	product_name	category	price	quantity	store_location
201	Apple iPhone 15	Smartphone	1099.99	50	Chicago
202	Samsung Galaxy S24	Smartphone	999.99	75	San Francisco
203	Lenovo IdeaPad 3	Laptop	479.99	150	New York
204	LG QNED 65 TV	Electronics	1599.99	30	Chicago
205	Dell Alienware X15	Laptop	2299.99	20	San Francisco
206	Google Pixel Watch	Smartwatch	399.99	100	Los Angeles
207	Microsoft Surface Laptop Go	Laptop	699.99	120	New York
208	Sony WF-1000XM4	Headphones	279.99	90	Chicago
209	Apple MacBook Air M2	Laptop	1299.99	200	San Francisco
[illegible]	[illegible] SoundLink Mini 2	Headphones	169.99	180	Los Angeles

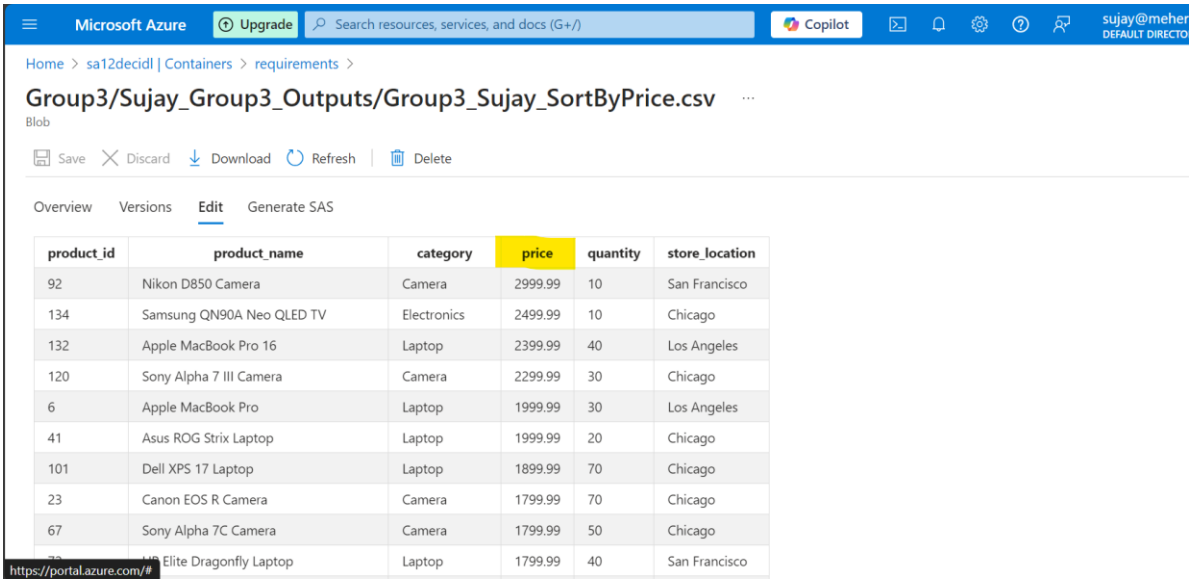
6. Outputs:

- Output of **Union Transformation**: Contains records from both files (except null values).



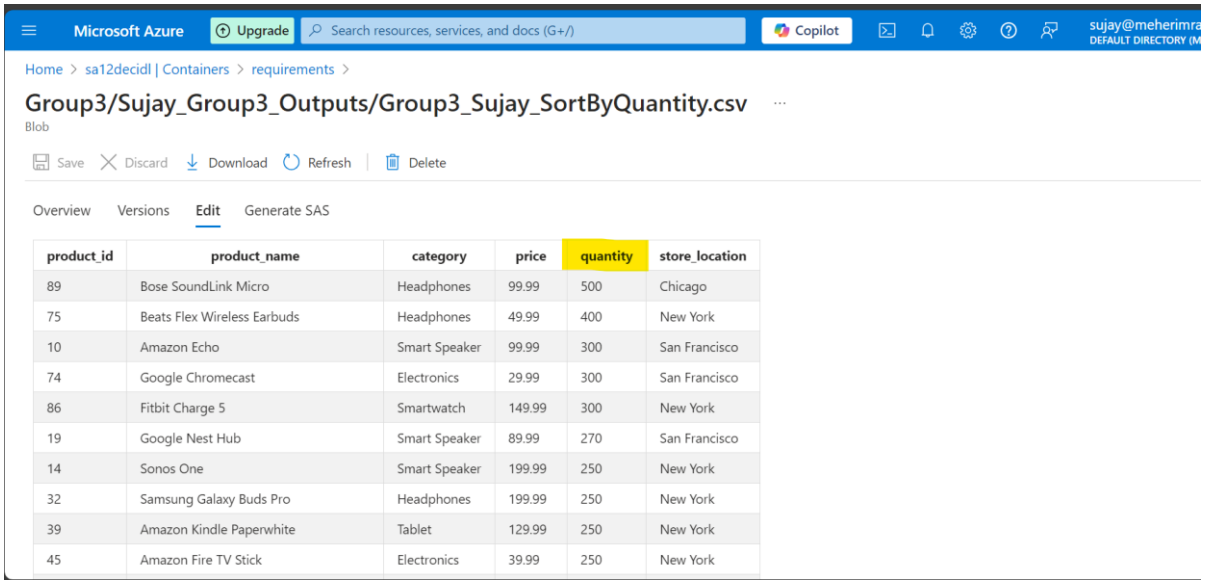
130	PlayStation 4	Console	299.99	100	New York
131	Beats Solo Pro	Headphones	299.99	80	San Francisco
132	Apple MacBook Pro 16	Laptop	2399.99	40	Los Angeles
133	Nintendo Switch Joy-Con	Console	79.99	200	San Francisco
134	Samsung QN90A Neo QLED TV	Electronics	2499.99	10	Chicago
201	Apple iPhone 15	Smartphone	1099.99	50	Chicago
202	Samsung Galaxy S24	Smartphone	999.99	75	San Francisco
203	Lenovo IdeaPad 3	Laptop	479.99	150	New York
204	LG QNED 65 TV	Electronics	1599.99	30	Chicago
205	Dell Alienware X15	Laptop	2299.99	20	San Francisco
206	Google Pixel Watch	Smartwatch	399.99	100	Los Angeles
207	Microsoft Surface Laptop Go	Laptop	699.99	120	New York
208	Sony WF-1000XM4	Headphones	279.99	90	Chicago

- Output of **Sort by Price**: Records are sorted in descending order as per price.



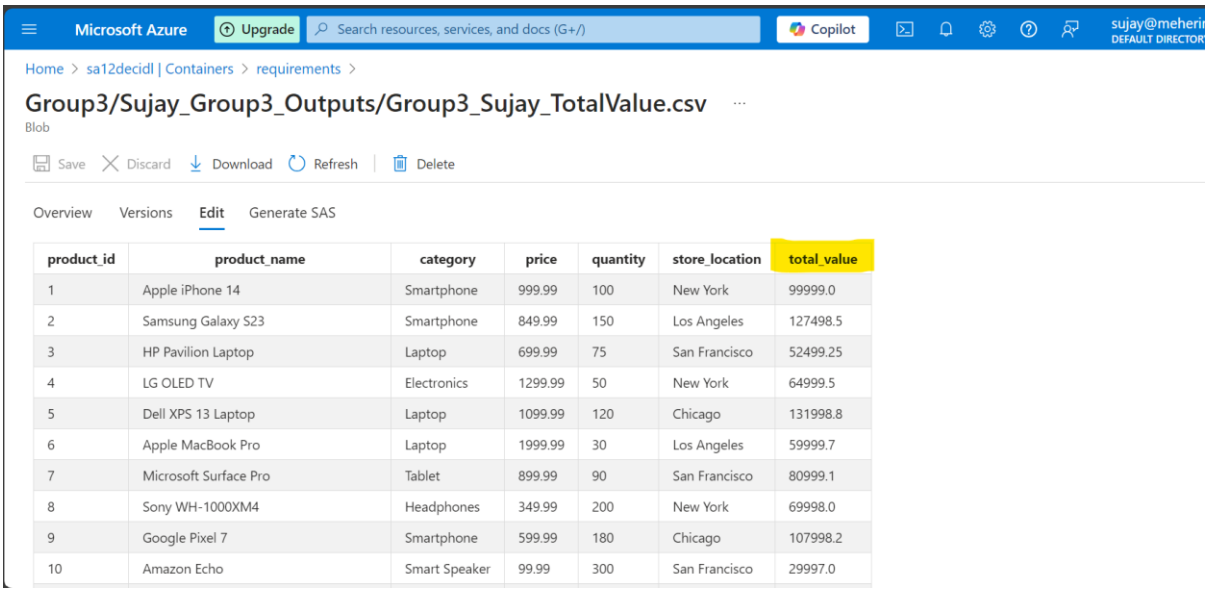
product_id	product_name	category	price	quantity	store_location
92	Nikon D850 Camera	Camera	2999.99	10	San Francisco
134	Samsung QN90A Neo QLED TV	Electronics	2499.99	10	Chicago
132	Apple MacBook Pro 16	Laptop	2399.99	40	Los Angeles
120	Sony Alpha 7 III Camera	Camera	2299.99	30	Chicago
6	Apple MacBook Pro	Laptop	1999.99	30	Los Angeles
41	Asus ROG Strix Laptop	Laptop	1999.99	20	Chicago
101	Dell XPS 17 Laptop	Laptop	1899.99	70	Chicago
23	Canon EOS R Camera	Camera	1799.99	70	Chicago
67	Sony Alpha 7C Camera	Camera	1799.99	50	Chicago
[illegible]	[illegible] Elite Dragonfly Laptop	Laptop	1799.99	40	San Francisco

- Output of **Sort by Quantity**: Records are sorted in descending order as per quantity.



product_id	product_name	category	price	quantity	store_location
89	Bose SoundLink Micro	Headphones	99.99	500	Chicago
75	Beats Flex Wireless Earbuds	Headphones	49.99	400	New York
10	Amazon Echo	Smart Speaker	99.99	300	San Francisco
74	Google Chromecast	Electronics	29.99	300	San Francisco
86	Fitbit Charge 5	Smartwatch	149.99	300	New York
19	Google Nest Hub	Smart Speaker	89.99	270	San Francisco
14	Sonos One	Smart Speaker	199.99	250	New York
32	Samsung Galaxy Buds Pro	Headphones	199.99	250	New York
39	Amazon Kindle Paperwhite	Tablet	129.99	250	New York
45	Amazon Fire TV Stick	Electronics	39.99	250	New York

- Output of **Derived Column**: Column 'total_value' is added.



product_id	product_name	category	price	quantity	store_location	total_value
1	Apple iPhone 14	Smartphone	999.99	100	New York	99999.0
2	Samsung Galaxy S23	Smartphone	849.99	150	Los Angeles	127498.5
3	HP Pavilion Laptop	Laptop	699.99	75	San Francisco	52499.25
4	LG OLED TV	Electronics	1299.99	50	New York	64999.5
5	Dell XPS 13 Laptop	Laptop	1099.99	120	Chicago	131998.8
6	Apple MacBook Pro	Laptop	1999.99	30	Los Angeles	59999.7
7	Microsoft Surface Pro	Tablet	899.99	90	San Francisco	80999.1
8	Sony WH-1000XM4	Headphones	349.99	200	New York	69998.0
9	Google Pixel 7	Smartphone	599.99	180	Chicago	107998.2
10	Amazon Echo	Smart Speaker	99.99	300	San Francisco	29997.0