PROJECT - 1

Extract, Transform, Load

Submitted by:
Santhosh N
(KB23021)

Dataset : 1.Users.parquet (1000records of male and female details both)

2.Organizations.csv (100 records of organization details)

3.People.csv (1000 records of People and theirdetails)

Problems:

1. Finding the Average,Min,Max, Salary of Female registered application users of a bank according to the Professions using Dataset 1.
2. Find the Total no of industries opened country wise and total number of employees working in those industries using Dataset 2.
3. Find the Male and Female Born in from 1940 -2023 according to each decade as per the Dataset 3.

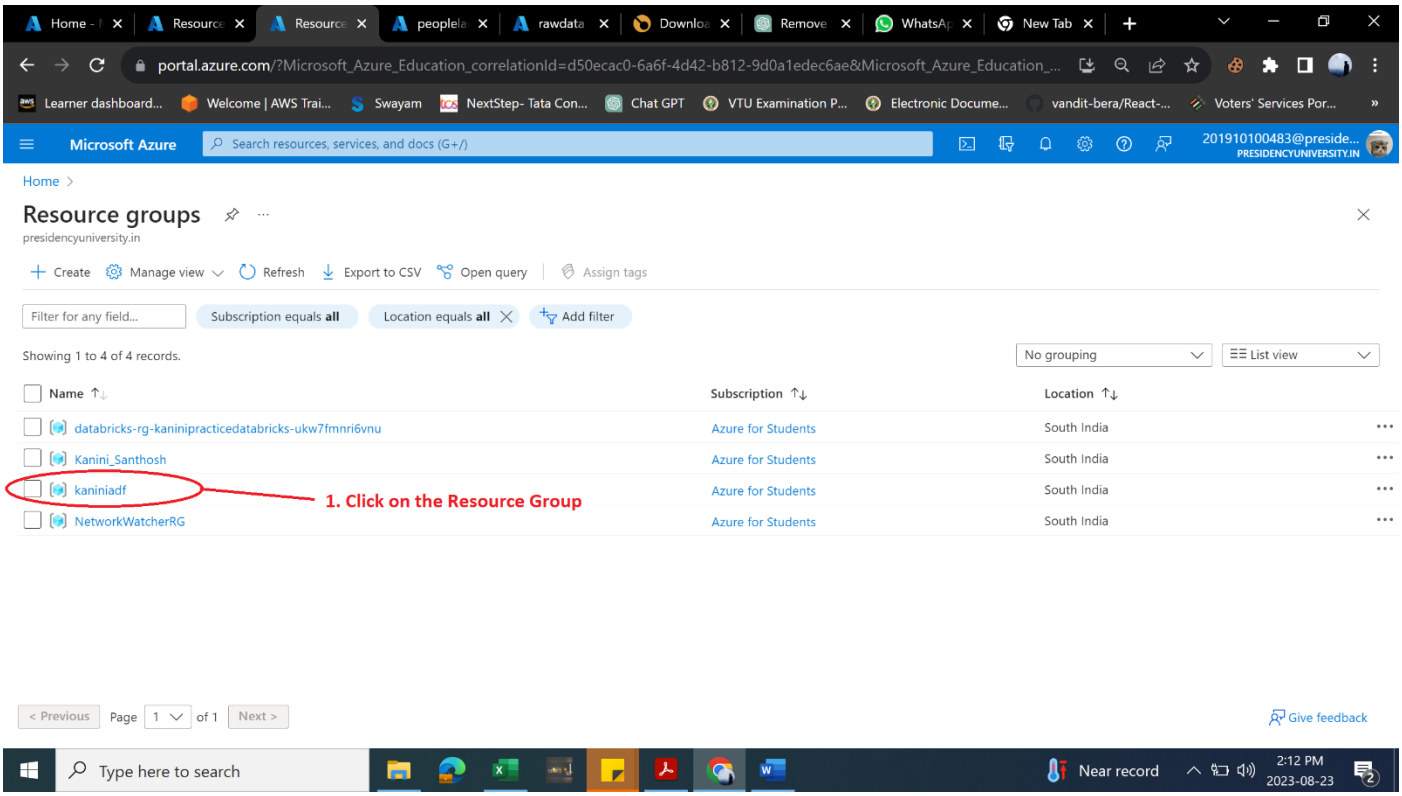
Tasks To be Done:

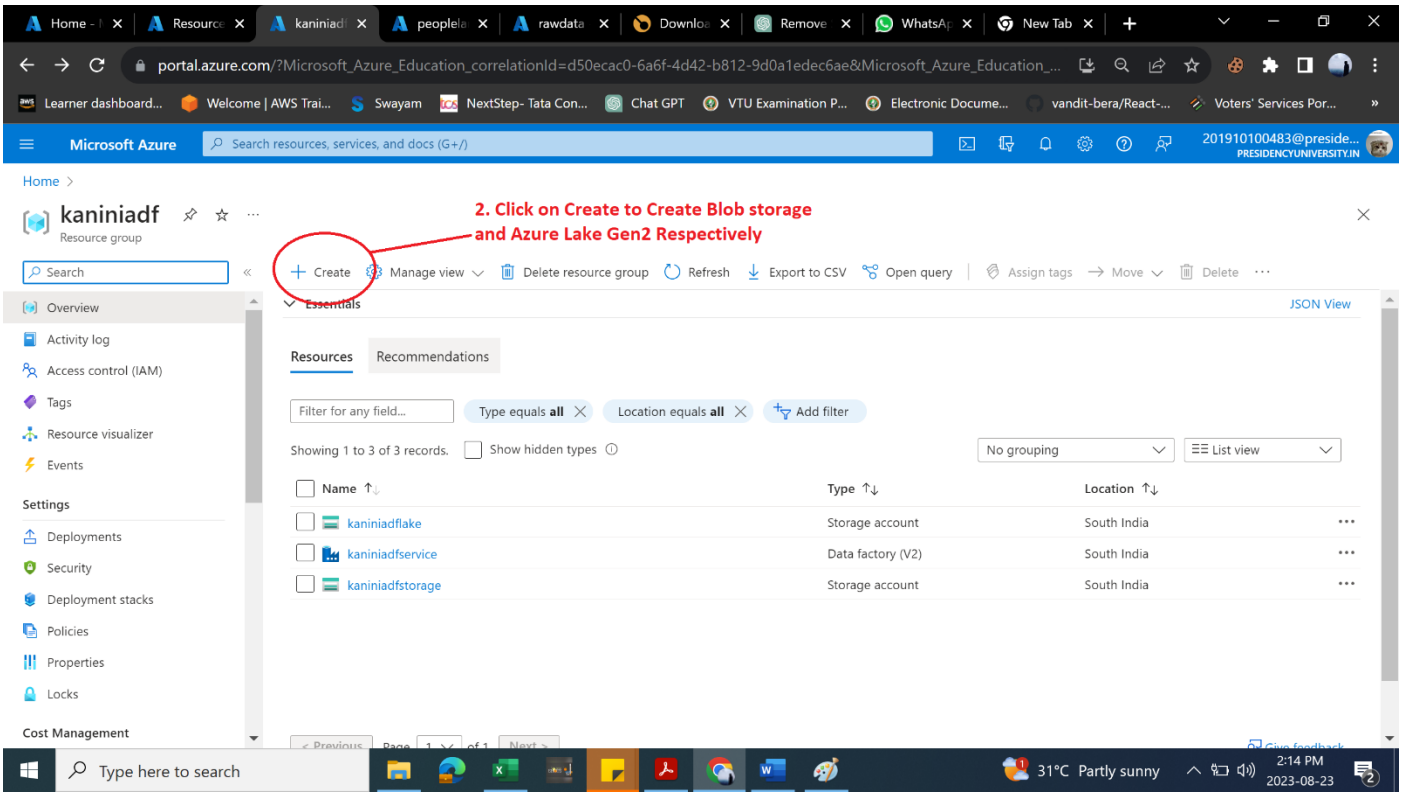
- Performing as much as transformation Possible.
- Need to use Medallion Architecture.
- Removing special Characters without Affecting any data.
- Removing Null values and replacing it with Suitable Strings.
- Formatting the Date format.
- Formatting the Mobile No format.
- Referencing the data from the Raw Datasets.
- Missed values of Name Column should be replaced with alternative data
- Adding the IP address of the device from one dataset to the Other Datasets .
- Should have to add current Timestamps.
- Storing the Cleaned and Structured data to the respective Container.
- Using Proper Modelling and Naming Conventions for all the Datasets , Transformations, Pipeline, and Dataflows.

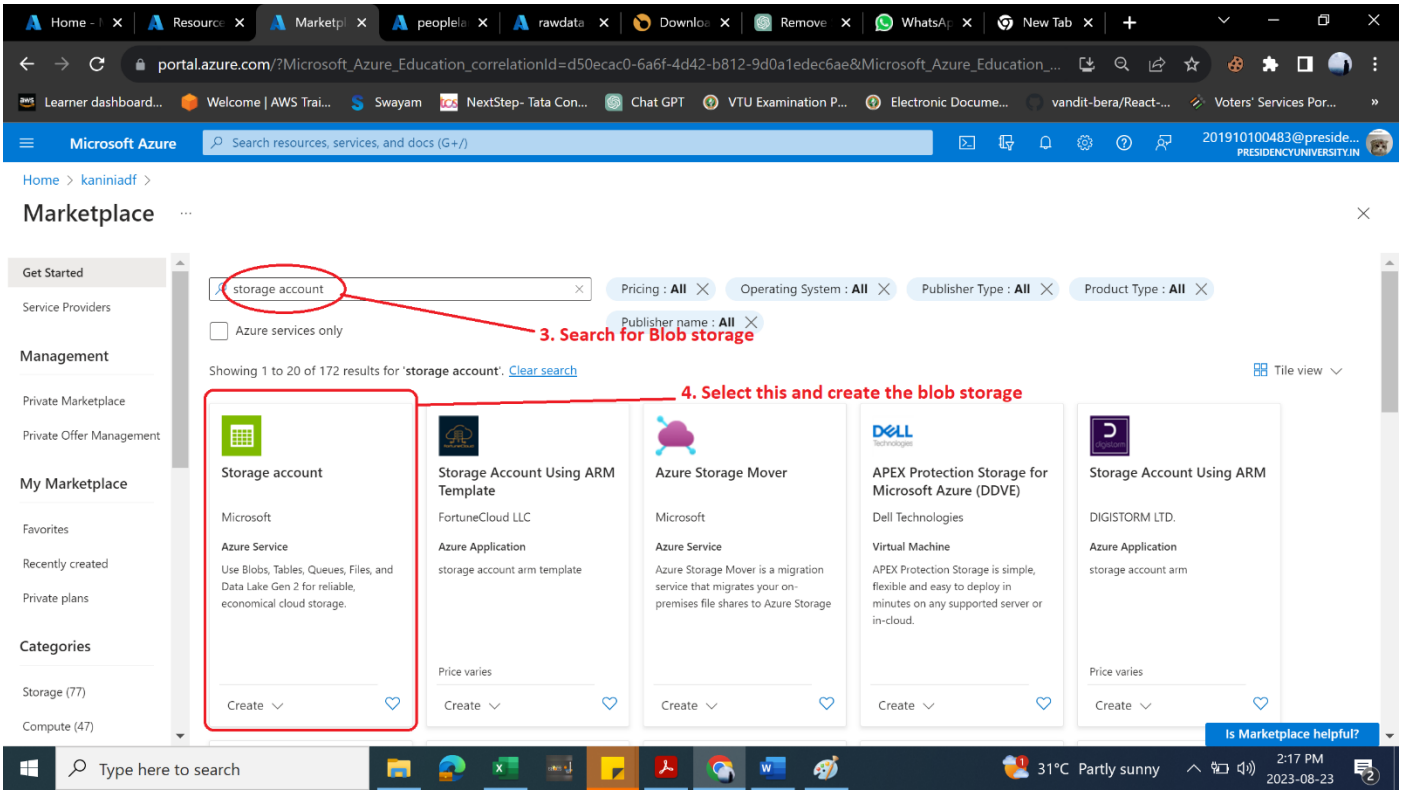
General Steps to Start the Project:

1: Create Two Storages in the Resource Group

1. BLOB STORAGE.
2. DATA LAKE.



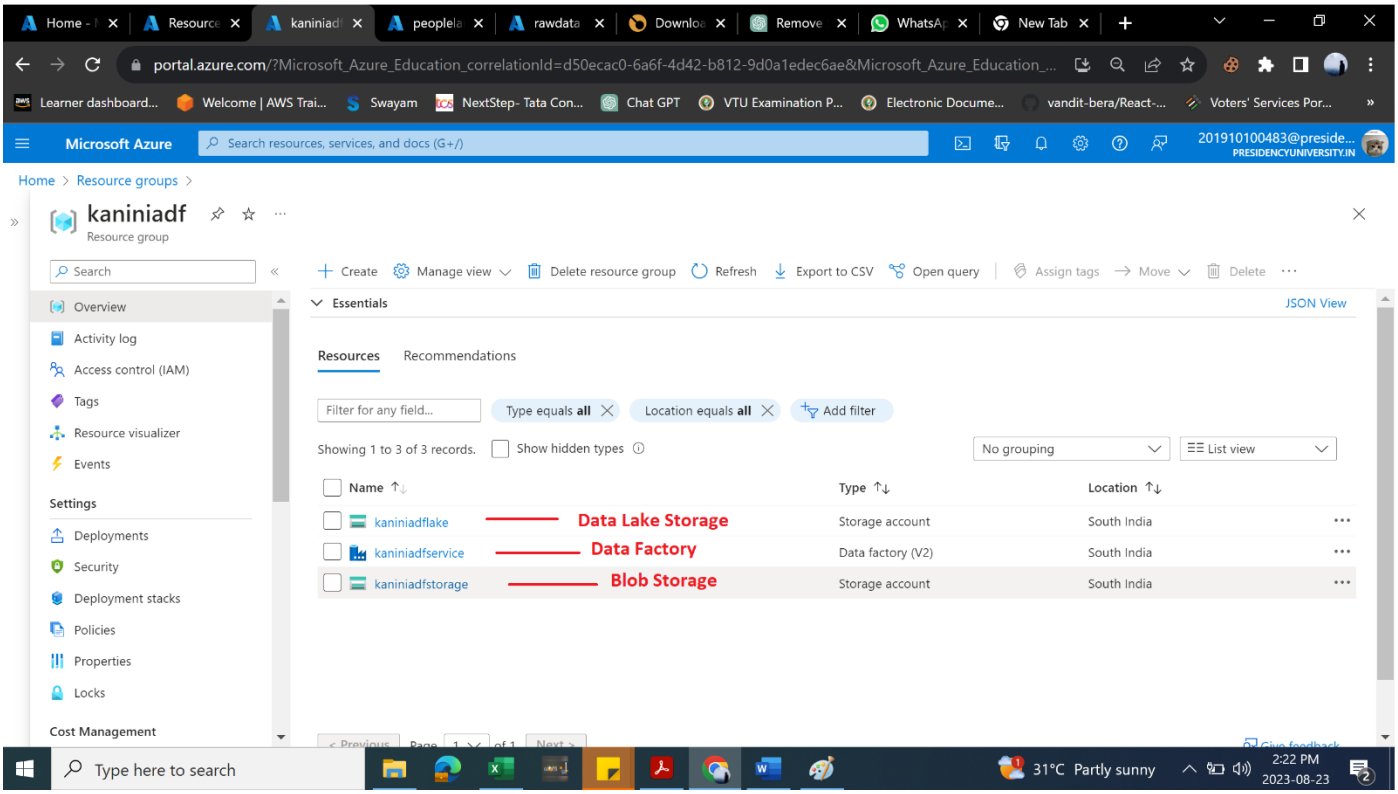




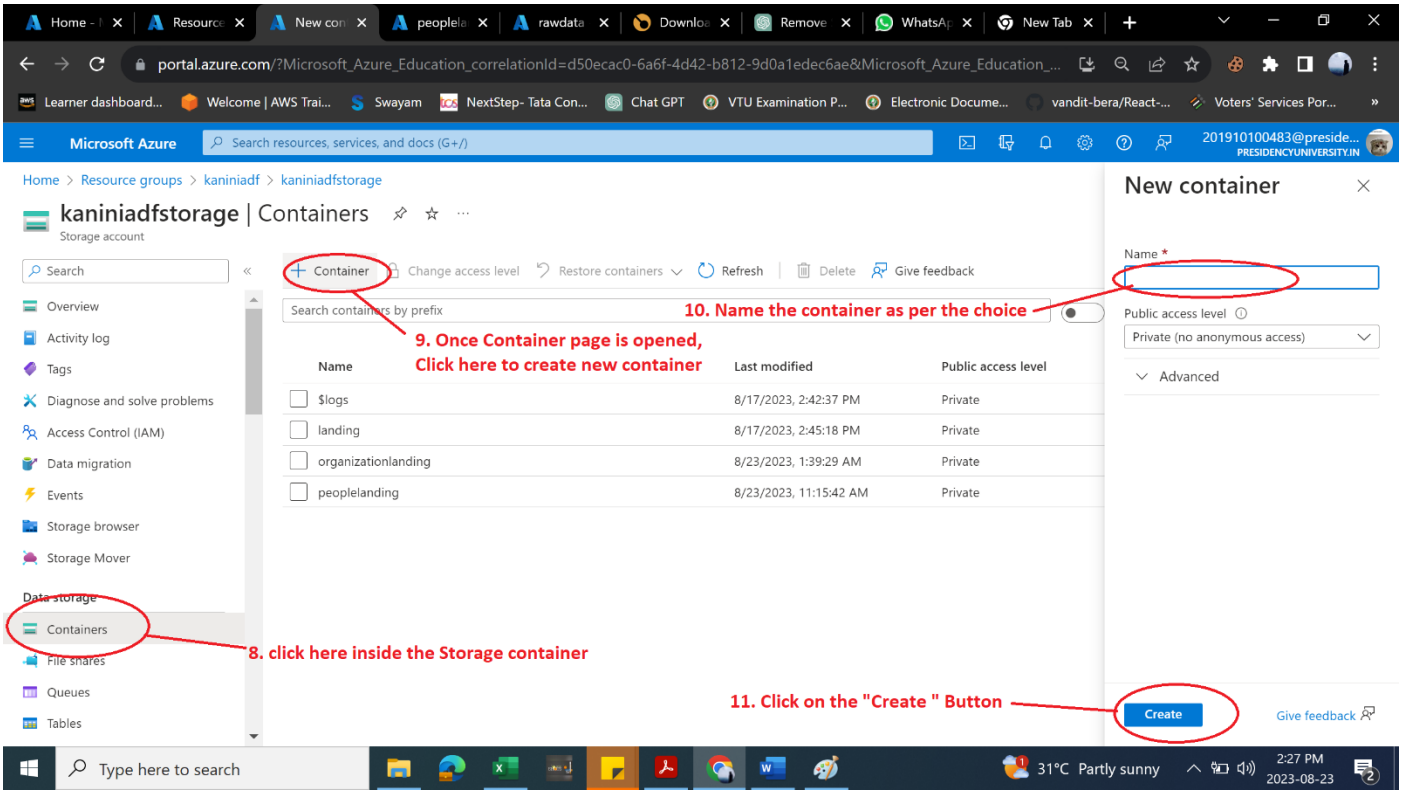
5. For Data Lake Storage just create one more Blob storage with Enabling the "Hierarchical Namespace ".

6. Next Create Azure Data Factory in the Resource Group .

7.Now you have all the resource needed for the project .



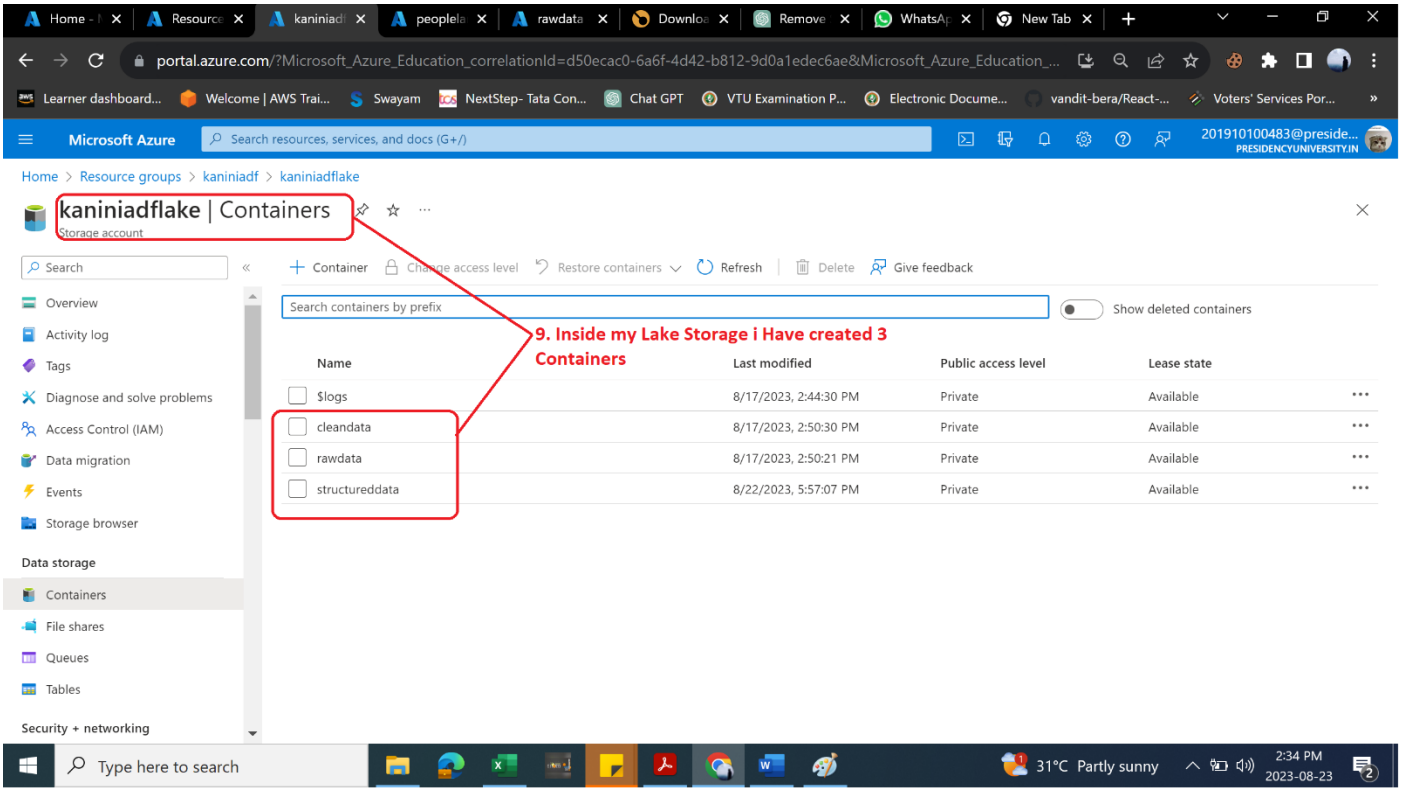
8. Now in Blob Storage , create a container called Landing (Note : Use separate Container for Each dataset used) .



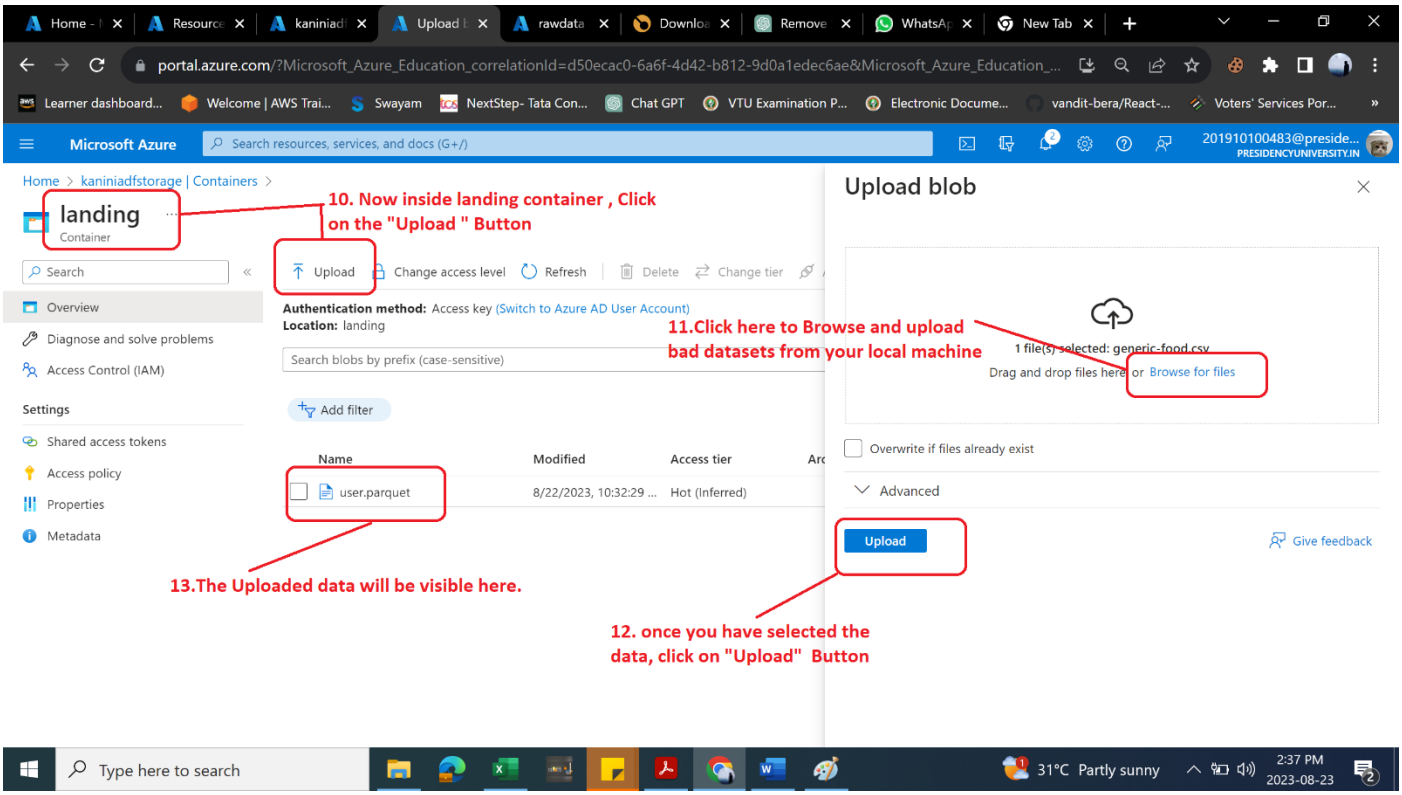
9. Likewise Create 3 Containers Inside the Data Lake as per Medallion Architecture ,

Here I am using Mainly 3 containers named :

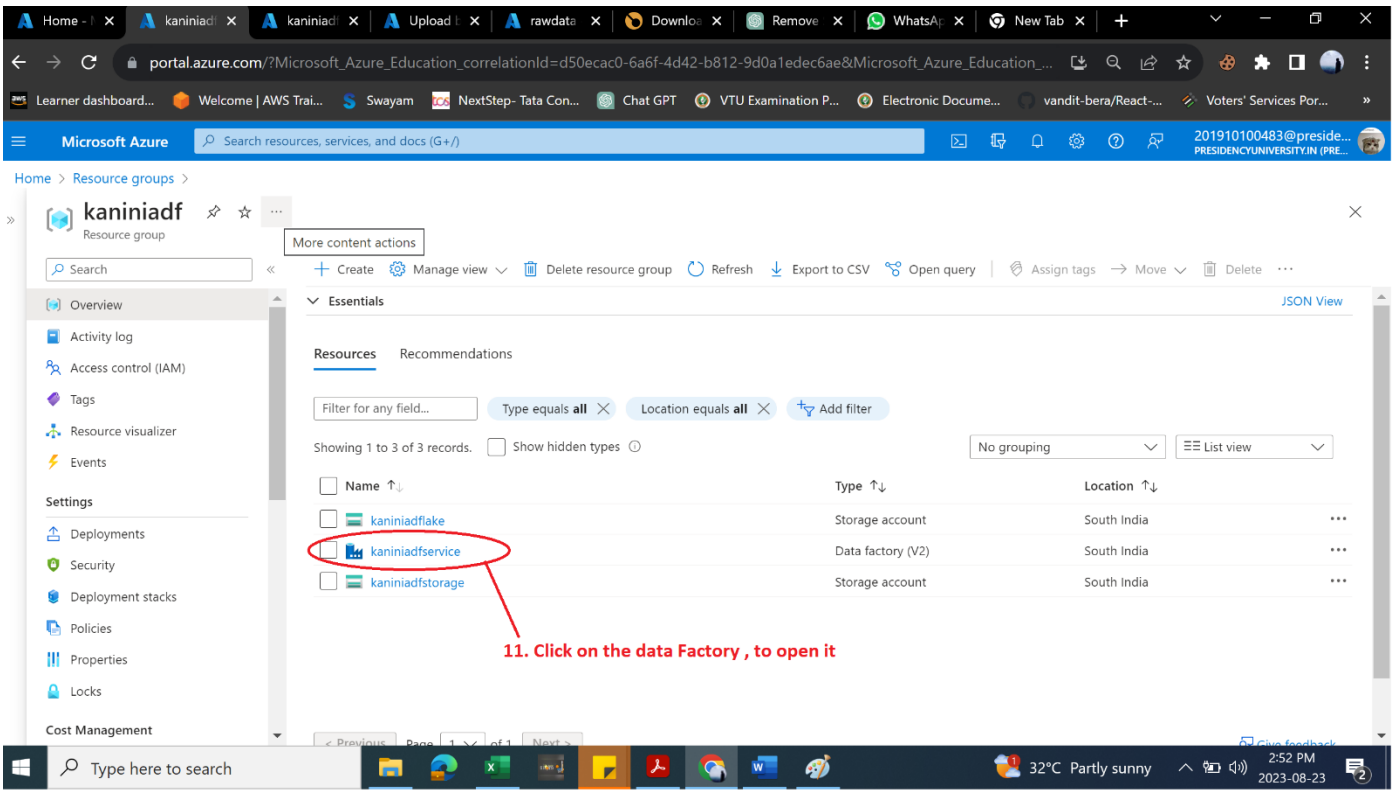
- **Raw Data – Bronze – Injecting the Data**
- **Clean Data – Silver – Cleaning The Data**
- **Structured Data – Gold – Performing the Logics.**

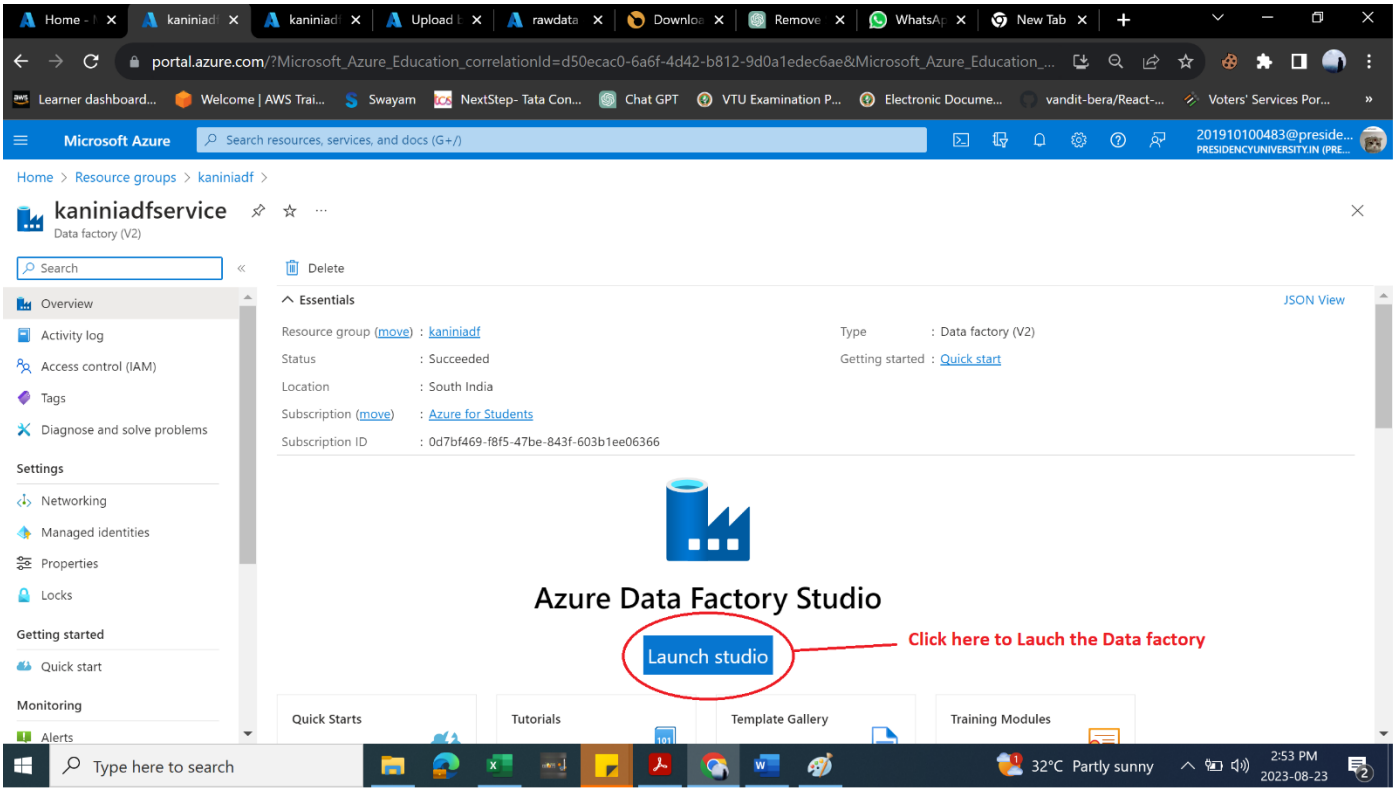


10. Now Upload the Bad Data set inside your Landing Container of the Blob Storage.

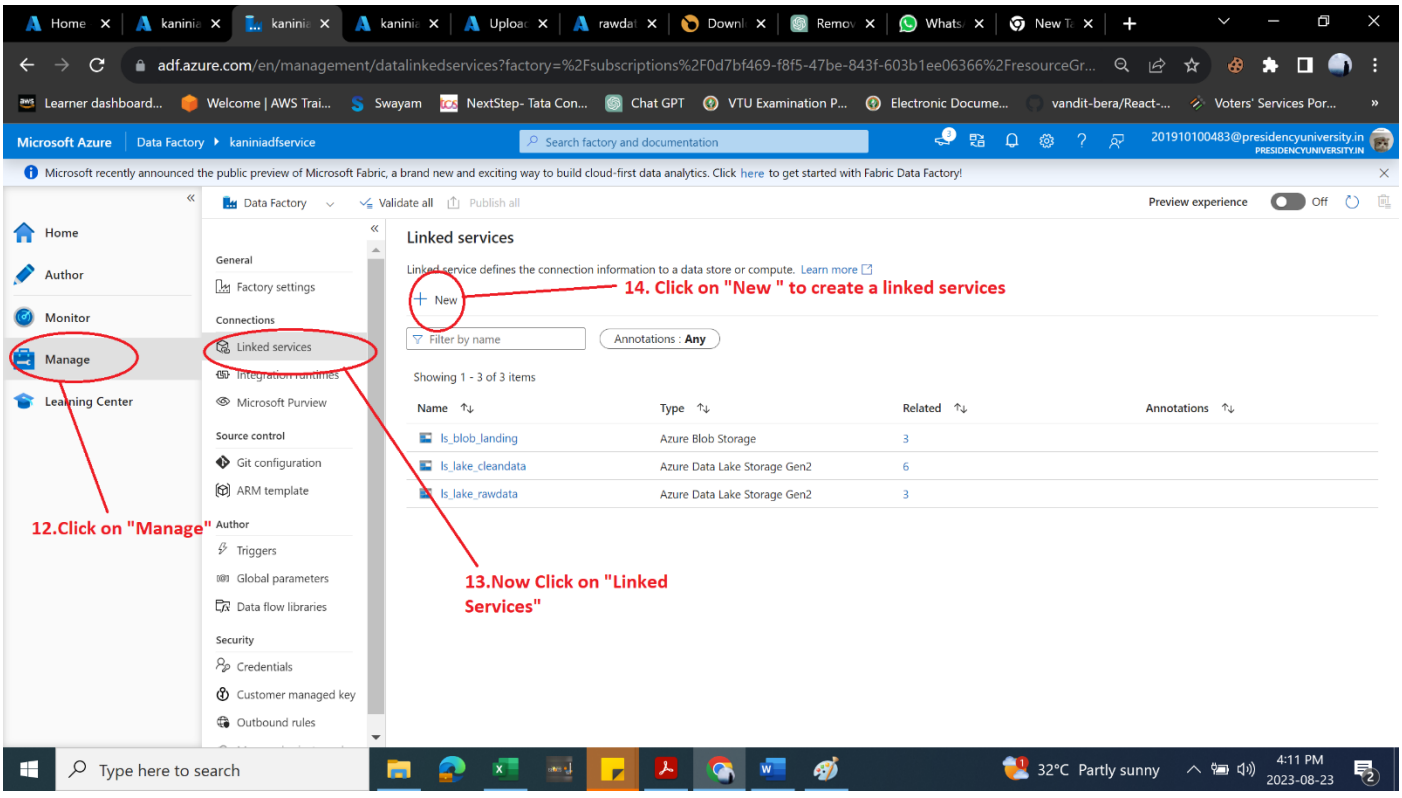


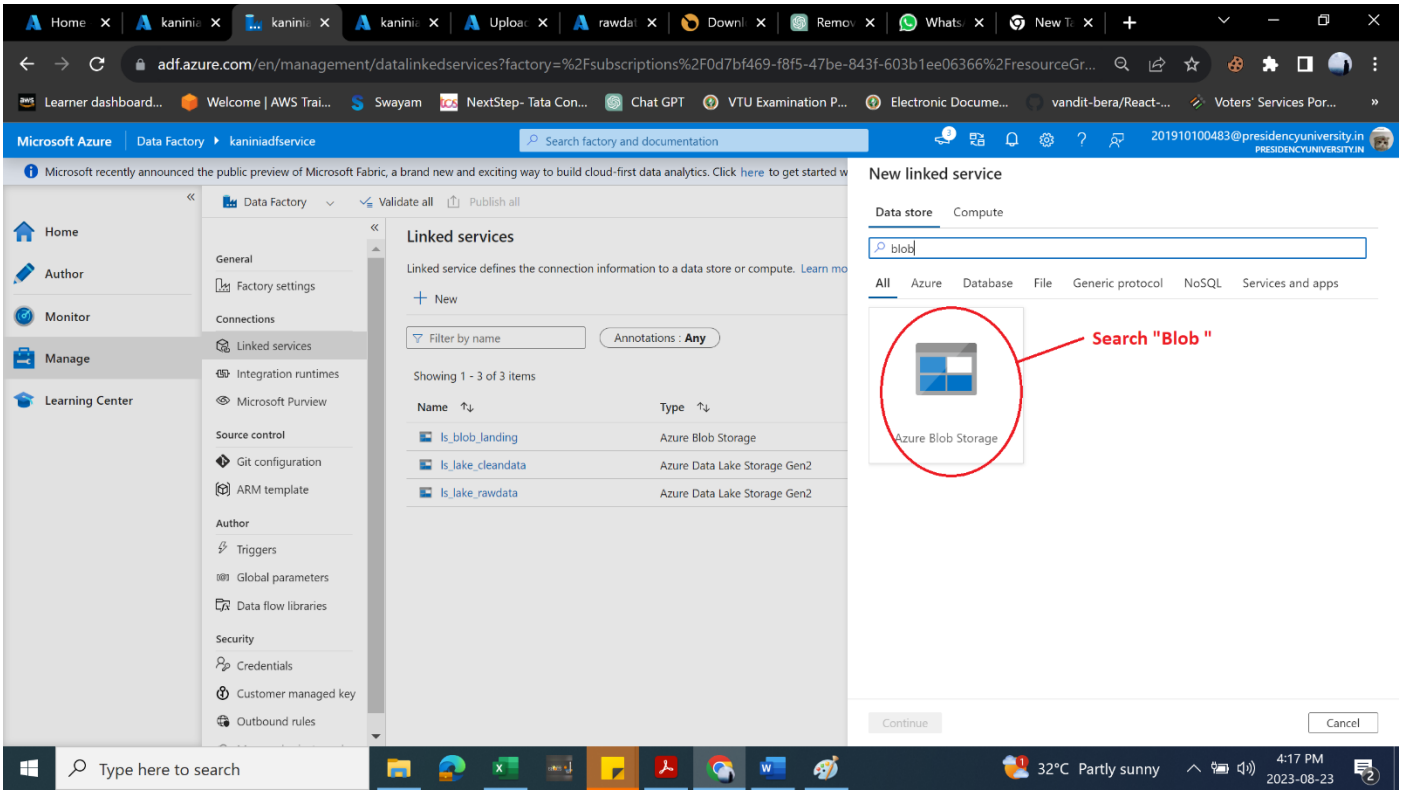
11. Now Lauch the Azure Data Factory for Further Processing of data.

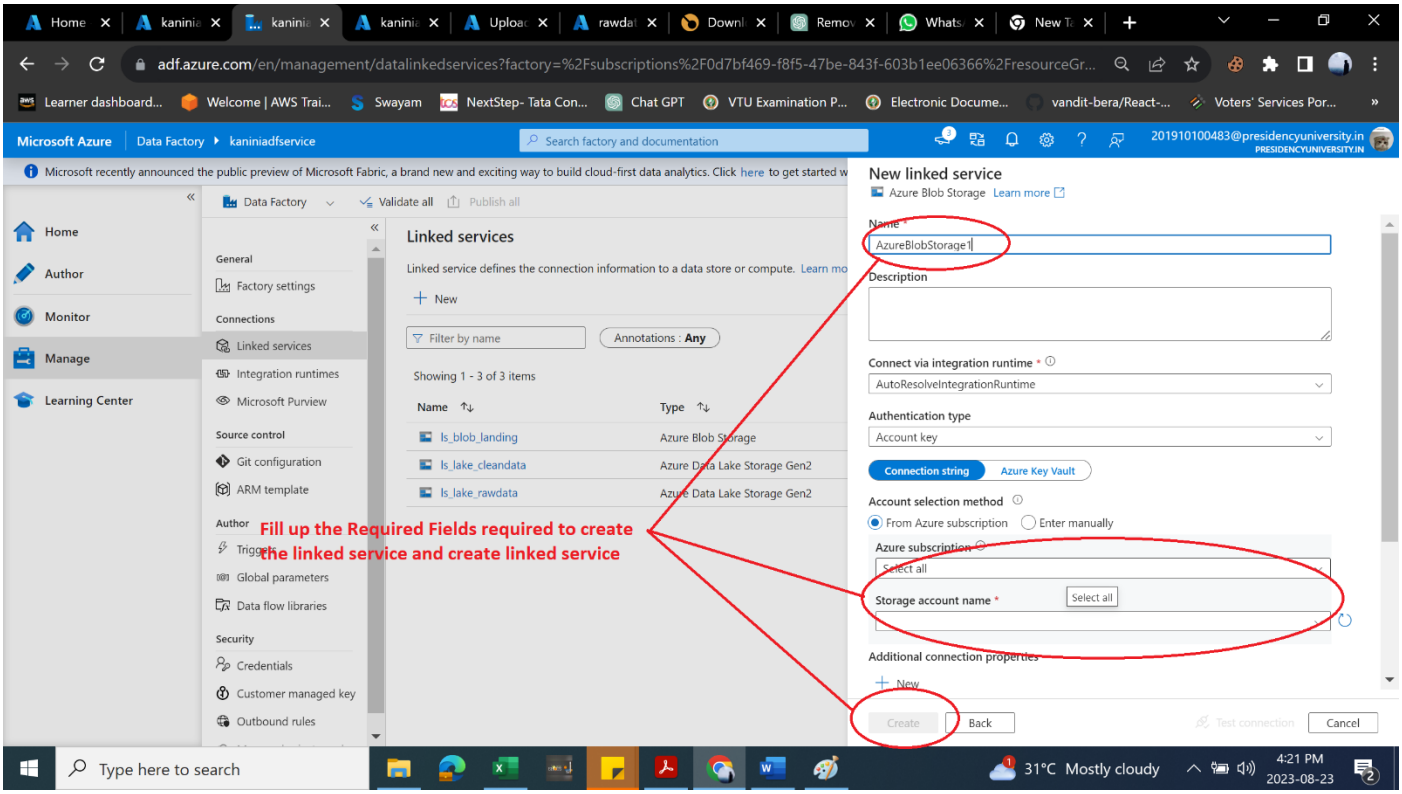


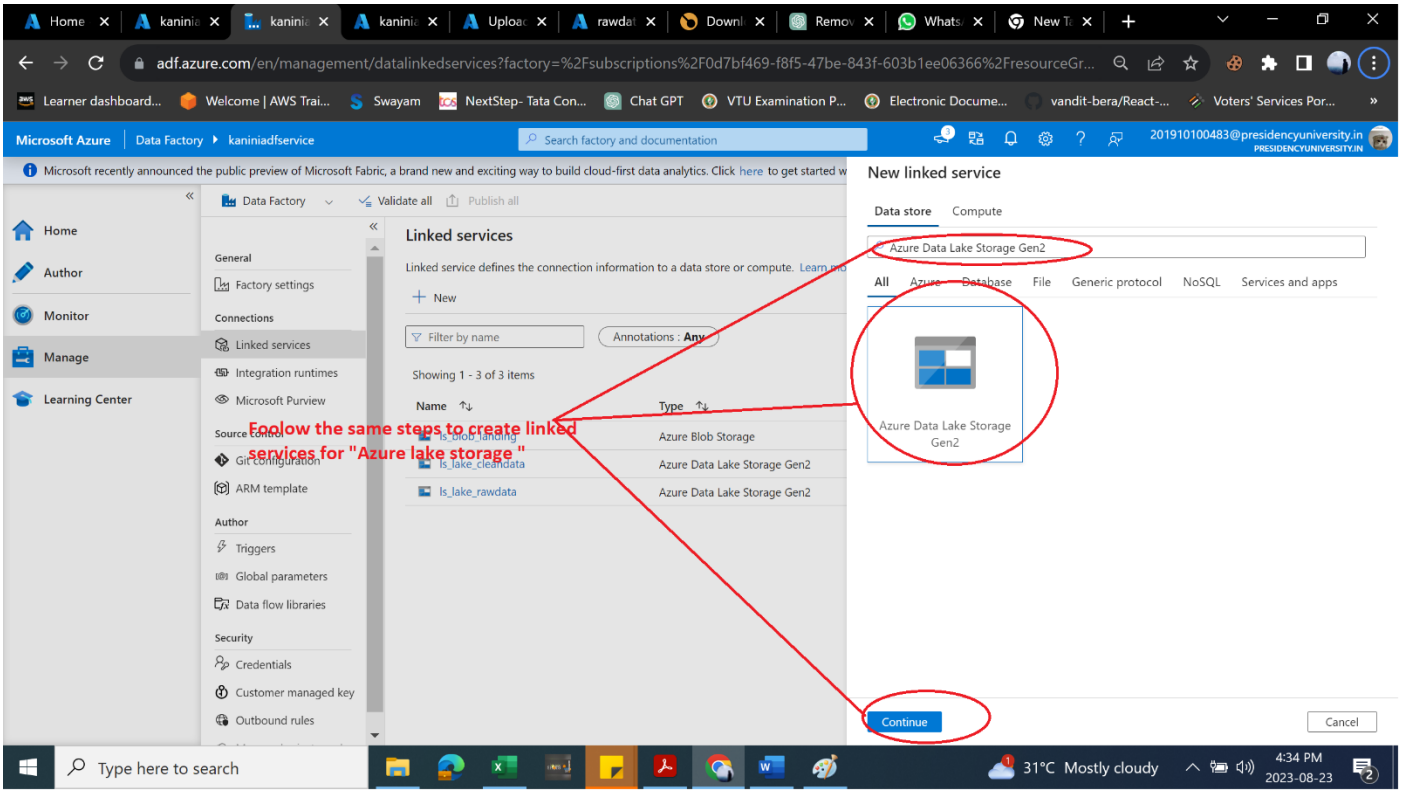


12. In Azure Data Factory(ADF) ,Create the Linked services for Both blob storage and lake storage.

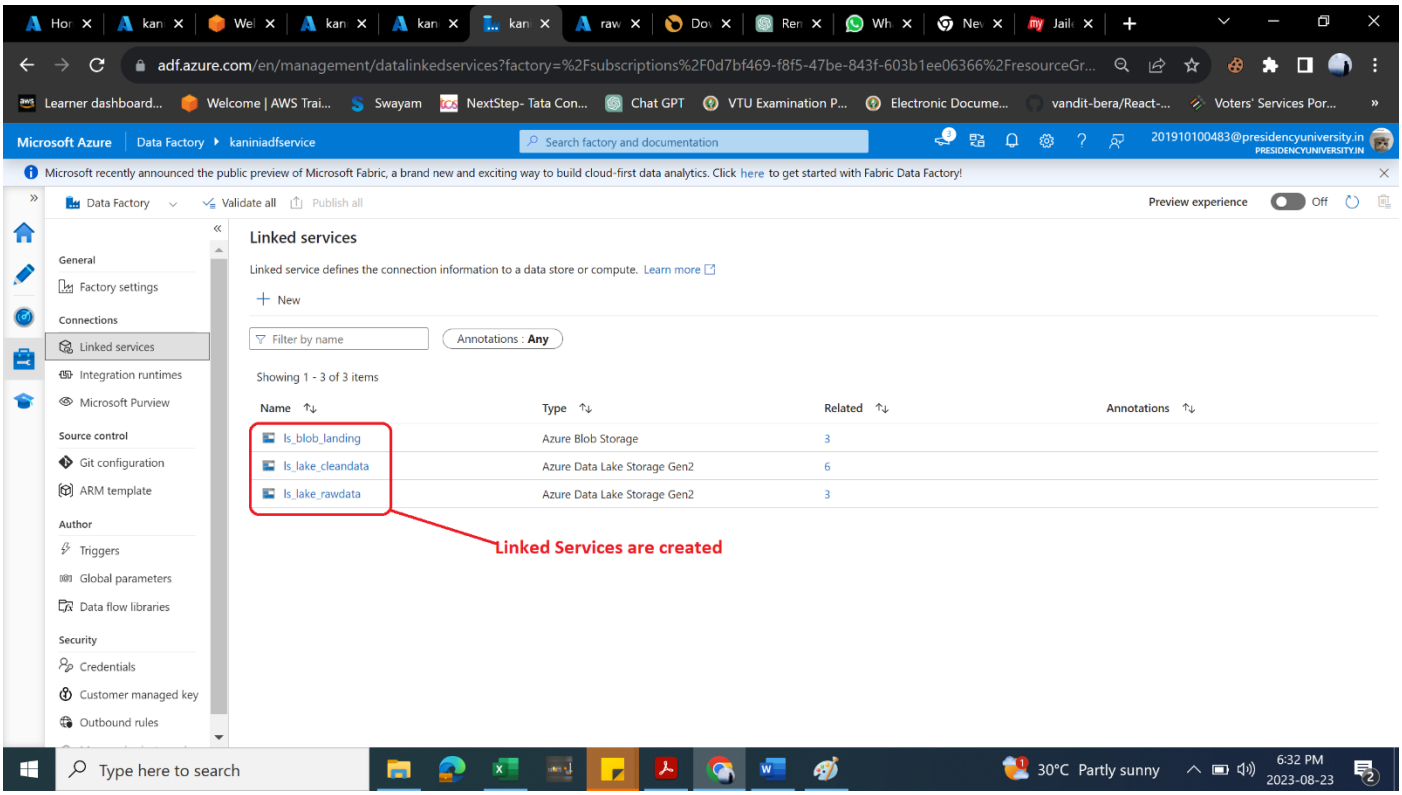



adf.azure.com/en/management/datalinkedservices?factory=%2Fsubscriptions%2F0d7bf469-f8f5-47be-843f-603b1ee06366%2FresourceGr...
Microsoft Azure
Data Factory
kaniniadfservice
Search factory and documentation
201910100483@presidencyuniversity.in
PRESIDENCYUNIVERSITY.IN
Microsoft recently announced the public preview of Microsoft Fabric, a brand new and exciting way to build cloud-first data analytics. Click here to get started w
Home
Author
Monitor
Manage
Learning Center
Data Factory
Validate all
Publish all
General
Factory settings
Connections
Linked services
Integration runtimes
Microsoft Purview
Source control
Git configuration
ARM template
Author
Triggers
Global parameters
Data flow libraries
Security
Credentials
Customer managed key
Outbound rules
Linked services
Linked service defines the connection information to a data store or compute. Learn mo
New
Filter by name
Annotations : Any
Showing 1 - 3 of 3 items
Name
Type
ls_blob_landing
Azure Blob Storage
ls_lake_cleandata
Azure Data Lake Storage Gen2
ls_lake_rawdata
Azure Data Lake Storage Gen2
New linked service
Data store
Compute
blob
All
Azure
Database
File
Generic protocol
NoSQL
Services and apps
Azure Blob Storage
Search "Blob "
Continue
Cancel
Type here to search
32°C Partly sunny
4:17 PM
2023-08-23

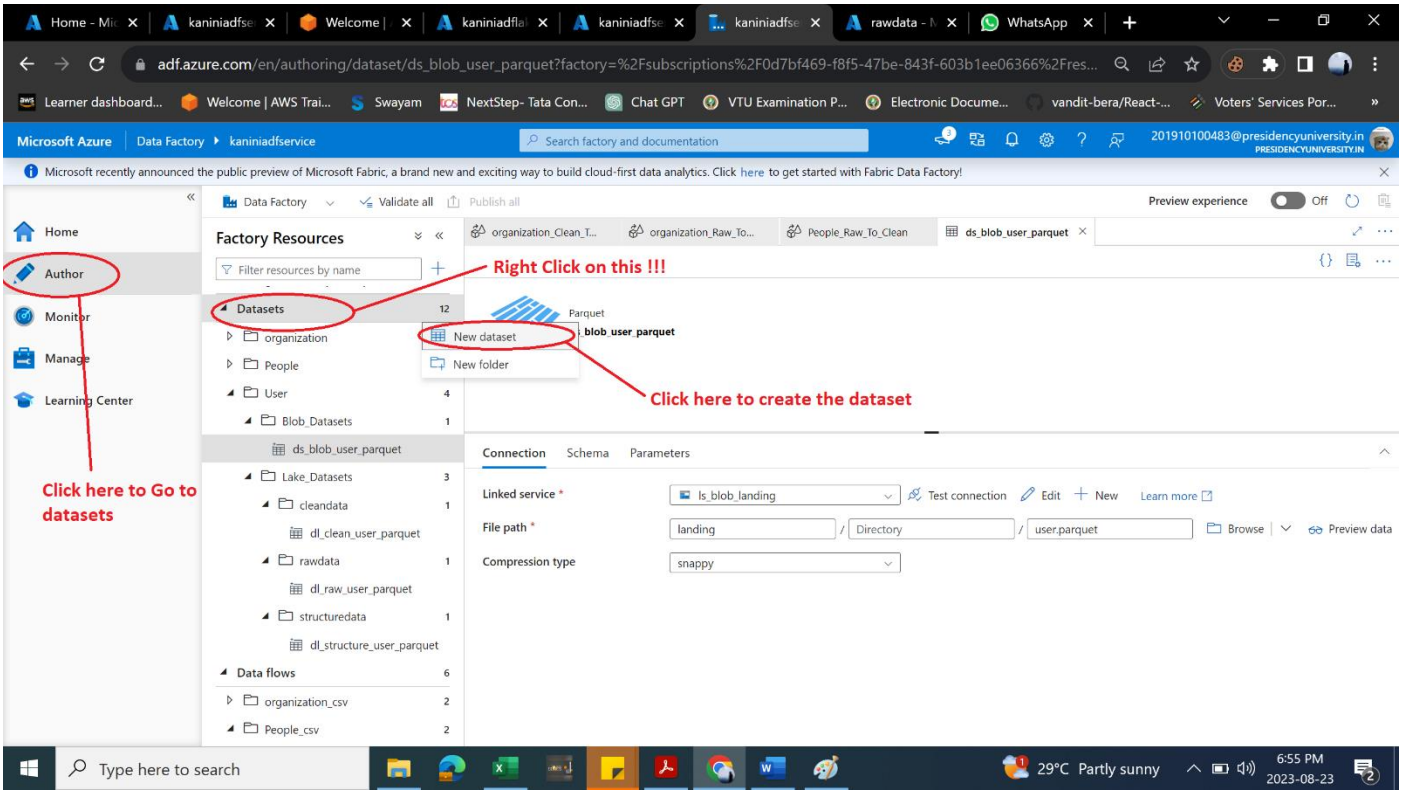

adf.azure.com/en/management/datalinkedservices?factory=%2Fsubscriptions%2F0d7bf469-f8f5-47be-843f-603b1ee06366%2FresourceGr...
Microsoft Azure
Data Factory
kaniniadfservice
Search factory and documentation
201910100483@presidencyuniversity.in
PRESIDENCYUNIVERSITY.IN
Microsoft recently announced the public preview of Microsoft Fabric, a brand new and exciting way to build cloud-first data analytics. Click here to get started w
Home
Author
Monitor
Manage
Learning Center
Data Factory
Validate all
Publish all
General
Factory settings
Connections
Linked services
Integration runtimes
Microsoft Purview
Source control
Git configuration
ARM template
Author
Global parameters
Data flow libraries
Security
Credentials
Customer managed key
Outbound rules
Linked services
Linked service defines the connection information to a data store or compute. Learn mo
New
Filter by name
Annotations : Any
Showing 1 - 3 of 3 items
Name
Type
ls_blob_landing
Azure Blob Storage
ls_lake_cleandata
Azure Data Lake Storage Gen2
ls_lake_rawdata
Azure Data Lake Storage Gen2
Fill up the Required Fields required to create the linked service and create linked service
New linked service
Azure Blob Storage
Learn more
Name *
AzureBlobStorage1
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
Select all
Storage account name *
Select all
Additional connection properties
New
Create
Back
Test connection
Cancel
Type here to search
31°C Mostly cloudy
4:21 PM
2023-08-23



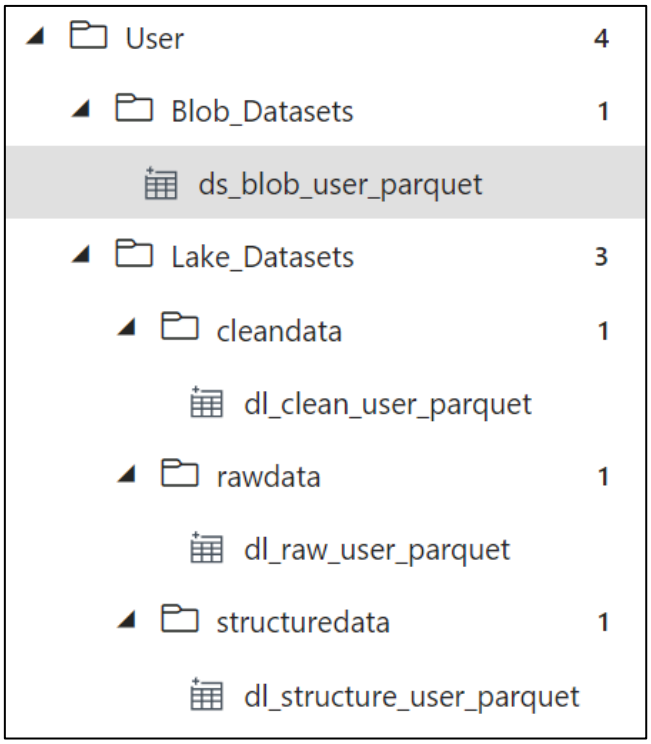
Linked Services are Succesfully created.



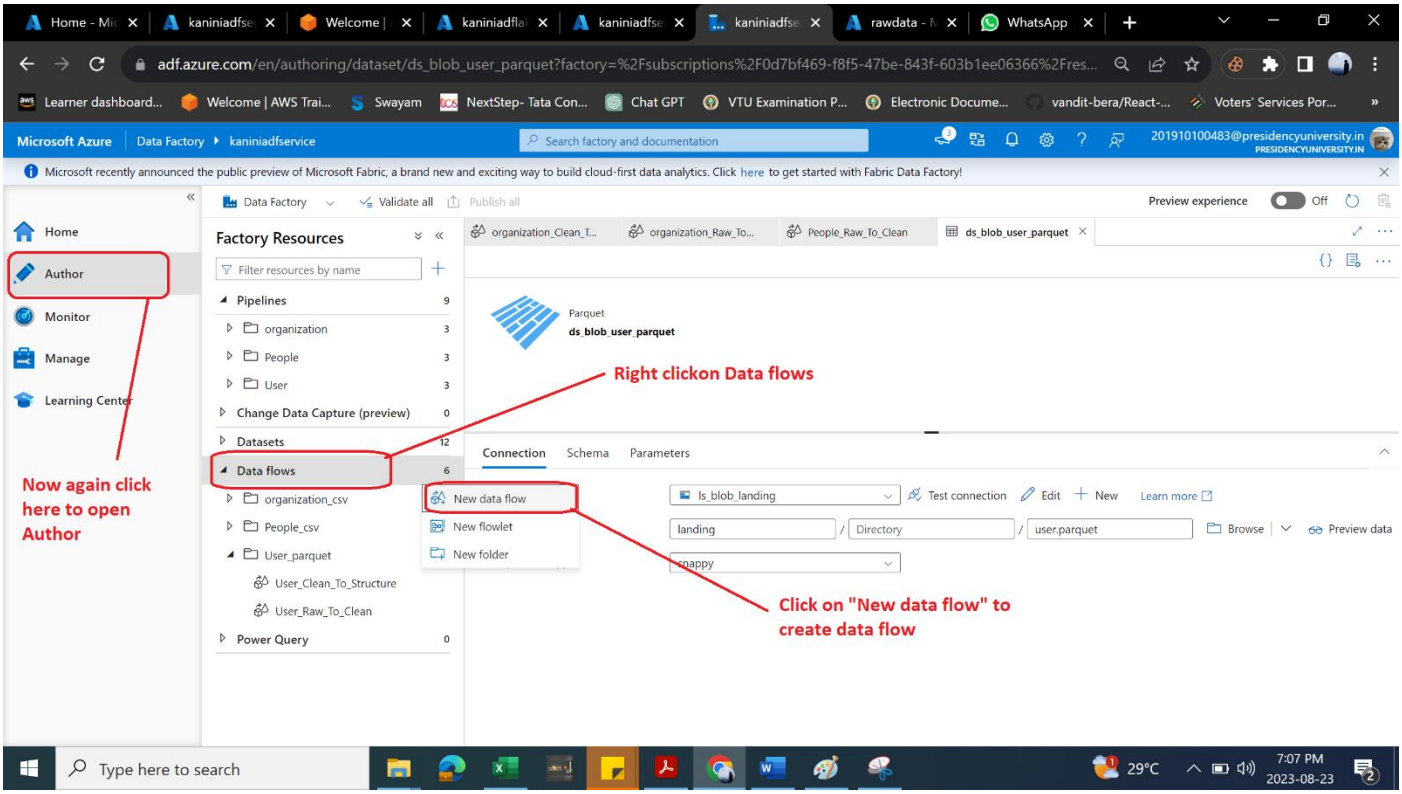
For all the Containers create a Datasets to fetch and save the data for processing :

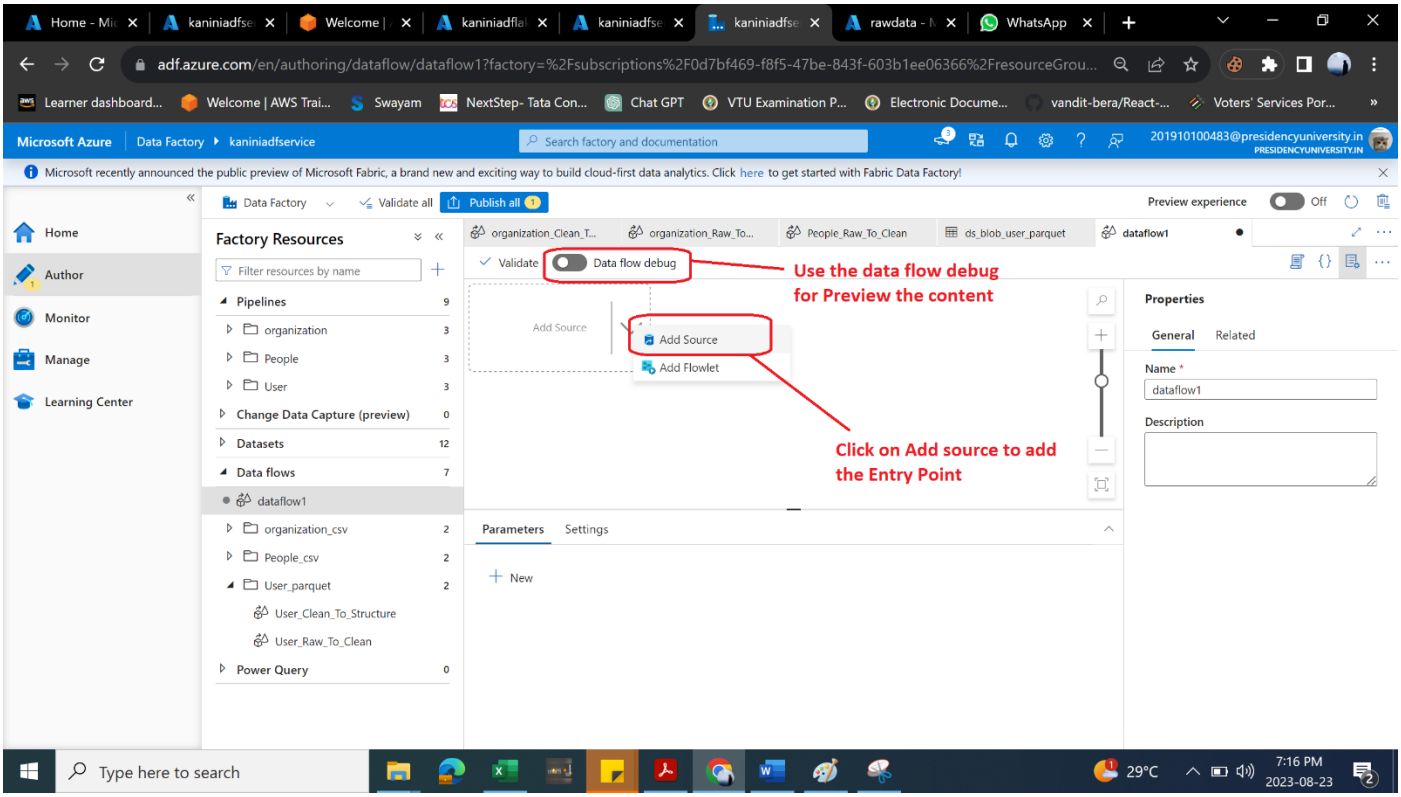


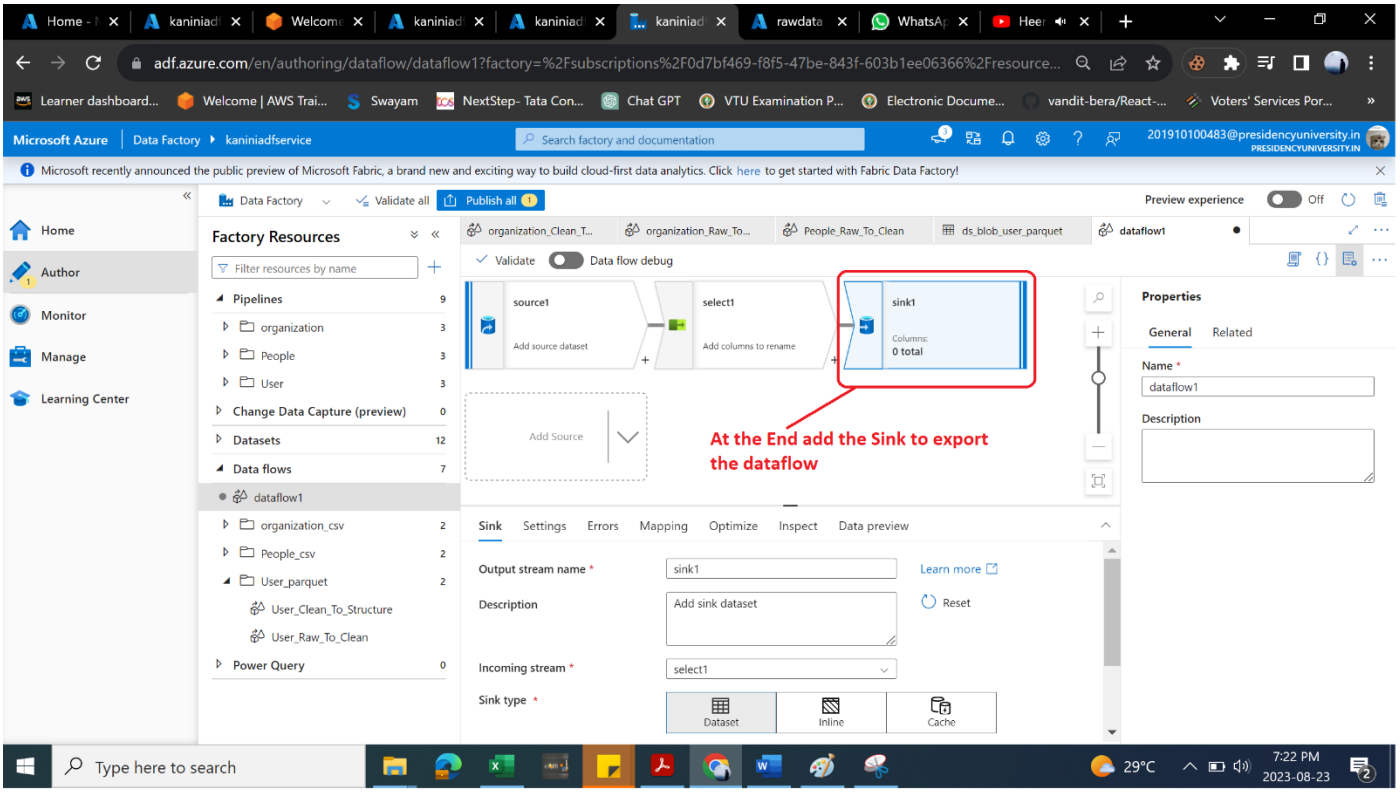
Create the datasets for all the containers which is created like Blob, Raw, clean , structured containers.



Now create a Dataflow for adding Transformations:

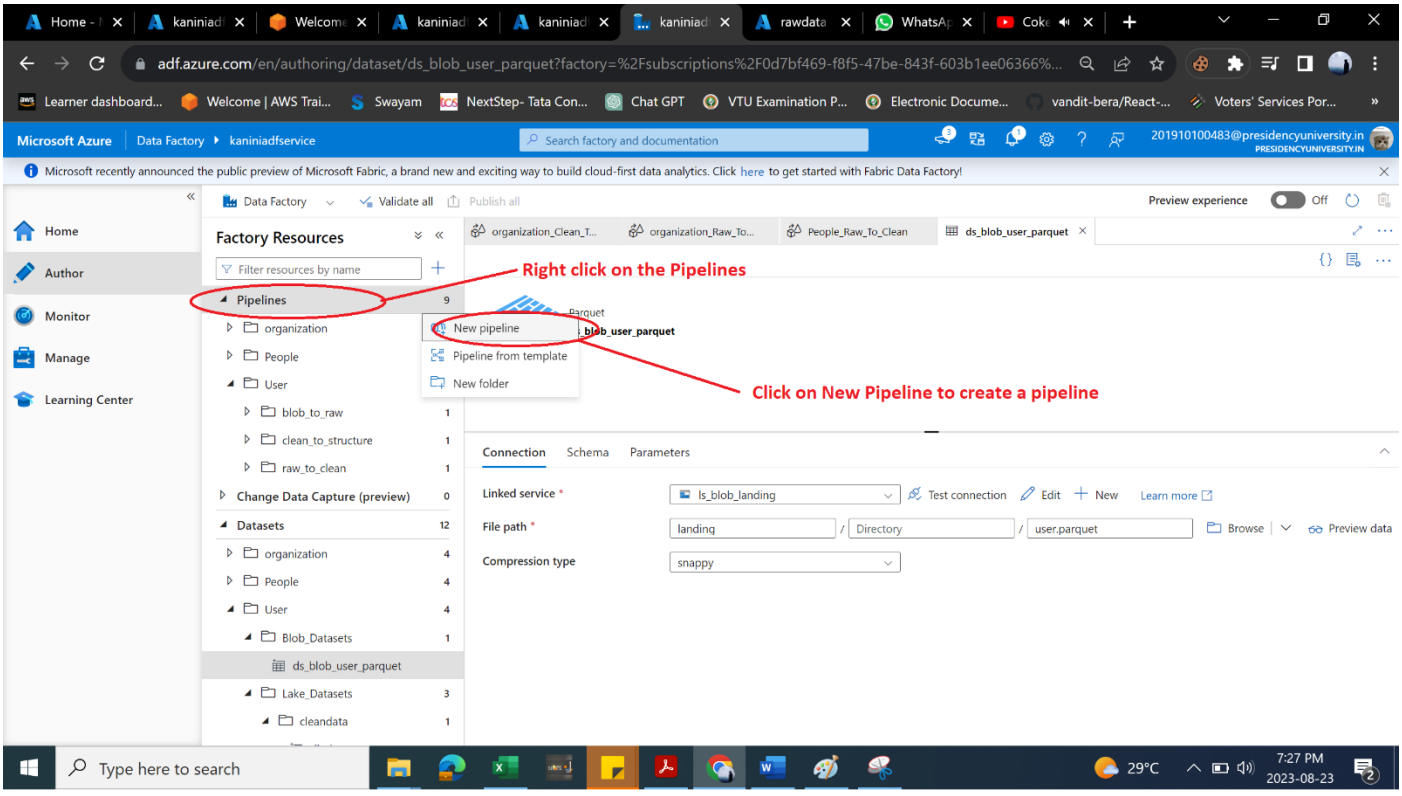


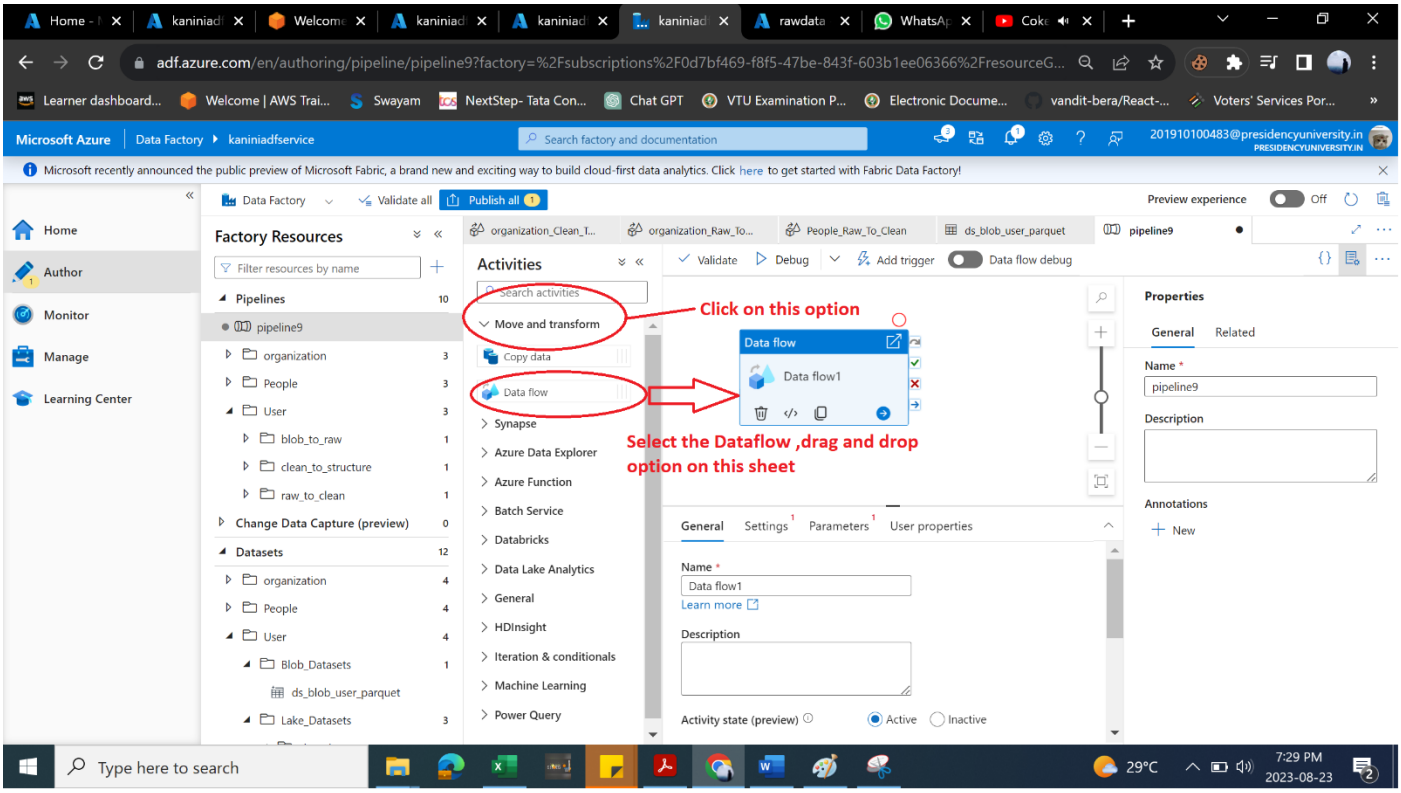


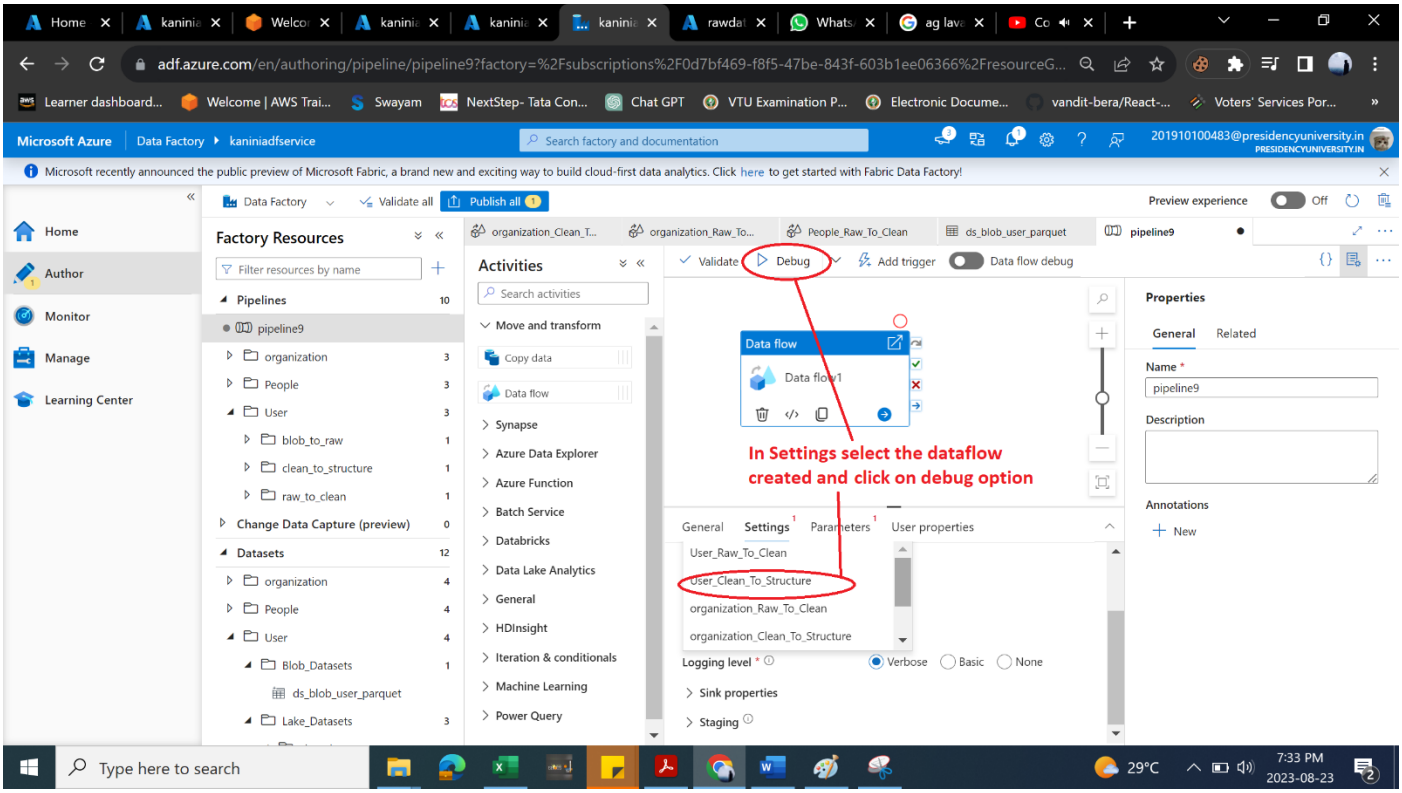


Do the same process to apply transformations on Cleansed Data to Structured Data.

13. Now add the Pipeline to export the Results to the Container.





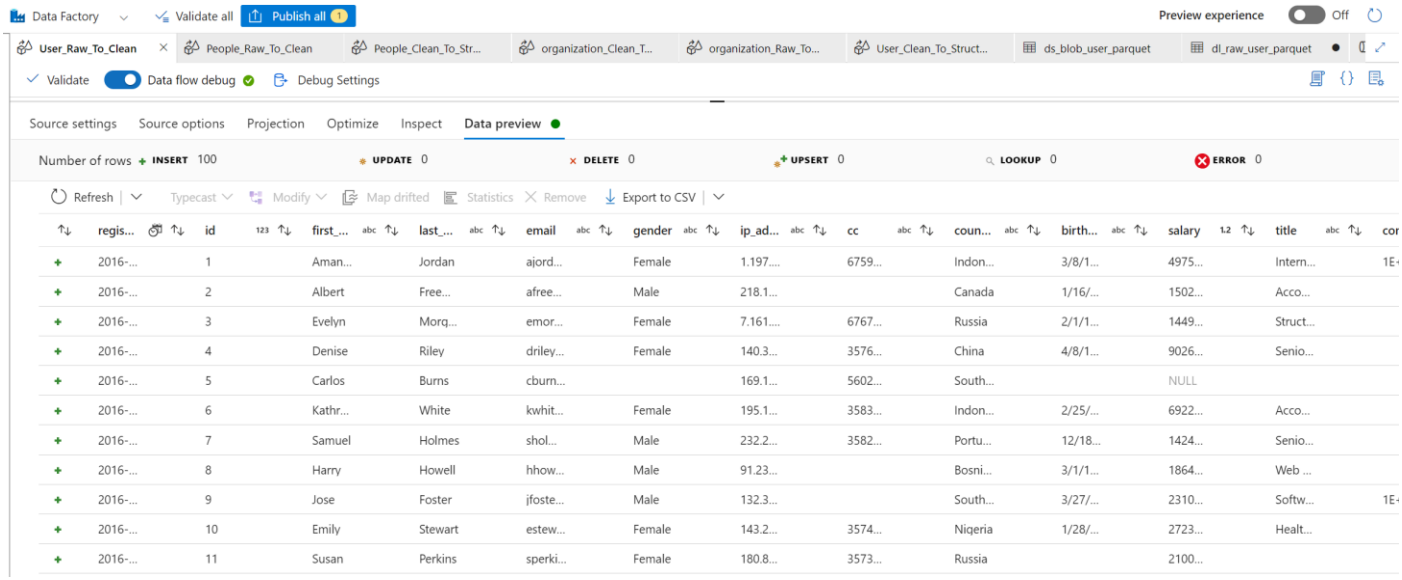


Do this for all the data flows to store the processed result inside the containers!!!

Outputs of Transformations used in My Projects Over 3 Datasets :

Data set 1 :

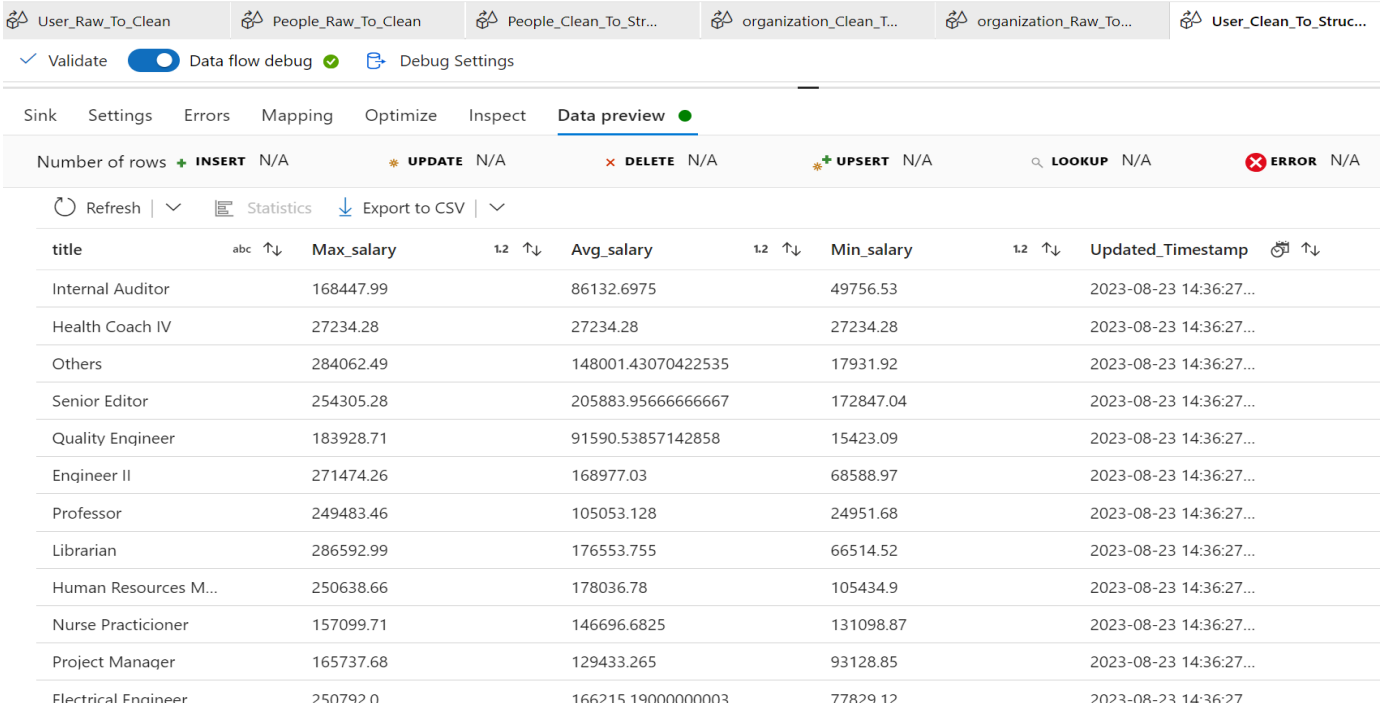
Raw data :



Raw to Clean data Transformations Used :

Clean to Structured data Transformations Used :

Final Output :



title	Max_salary	Avg_salary	Min_salary	Updated_Timestamp
Internal Auditor	168447.99	86132.6975	49756.53	2023-08-23 14:36:27...
Health Coach IV	27234.28	27234.28	27234.28	2023-08-23 14:36:27...
Others	284062.49	148001.43070422535	17931.92	2023-08-23 14:36:27...
Senior Editor	254305.28	205883.95666666667	172847.04	2023-08-23 14:36:27...
Quality Engineer	183928.71	91590.53857142858	15423.09	2023-08-23 14:36:27...
Engineer II	271474.26	168977.03	68588.97	2023-08-23 14:36:27...
Professor	249483.46	105053.128	24951.68	2023-08-23 14:36:27...
Librarian	286592.99	176553.755	66514.52	2023-08-23 14:36:27...
Human Resources M...	250638.66	178036.78	105434.9	2023-08-23 14:36:27...
Nurse Practicioner	157099.71	146696.6825	131098.87	2023-08-23 14:36:27...
Project Manager	165737.68	129433.265	93128.85	2023-08-23 14:36:27...
Electrical Engineer	250792.0	166215.190000000003	77829.12	2023-08-23 14:36:27

Data set 2 :

Raw data :

Source settings | Source options | Projection | Optimize | Inspect | Data preview

Number of rows + INSERT 100 | UPDATE 0 | DELETE 0 | UPSERT 0 | LOOKUP 0 | ERROR 0 | TOTAL 100

Refresh | Typecast | Modify | Map drifted | Statistics | Remove | Export to CSV

Index	Organizatio...	Name	Website	Country	Description	Founded	Industry	Number of ...
1	FAB0d41d5b...	Ferrell LLC	https://price....	Papua New ...	Horizontal e...	1990	Plastics	3498
2	6A7EdDEA9...	Mckinney, Ri...	http://www....	Finland	User-centric ...	2015	Glass / Cera...	4952
3	0bFED1ADA...	Hester Ltd	http://sulliva...	China	Switchable s...	1971	Public Safety	5287
4	2bFC1Be8a4...	Holder-Sellers	https://beck...	Turkmenistan	De-engineer...	2004	Automotive	921
5	9eE8A6a4Eb...	Mayer Group	http://www....	Mauritius	Synchronize...	1991	Transportati...	7870
6	cC757116fe...	Henry-Thom...	http://morse...	Bahamas	Face-to-face...	1992	Primary / Se...	4914
7	219233e8aF...	Hansen-Ever...	https://www....	Pakistan	Seamless dis...	2018	Publishing I...	7832
8	ccc93DCF81...	Mcintosh-M...	https://www....	Heard Island...	Centralized ...	1970	Import / Exp...	4389
9	0B4F93aA06...	Carr Inc	http://ross.c...	Kuwait	Distributed i...	1996	Plastics	8167
10	738b5aDe6B...	Gaines Inc	http://sando...	Uzbekistan	Multi-lateral...	1997	Outsourcing...	9698

Raw to Clean data Transformations Used :

Clean to Structured data Transformations Used :

Final Output :

Sink | Settings | Errors | Mapping | Optimize | Inspect | Data preview

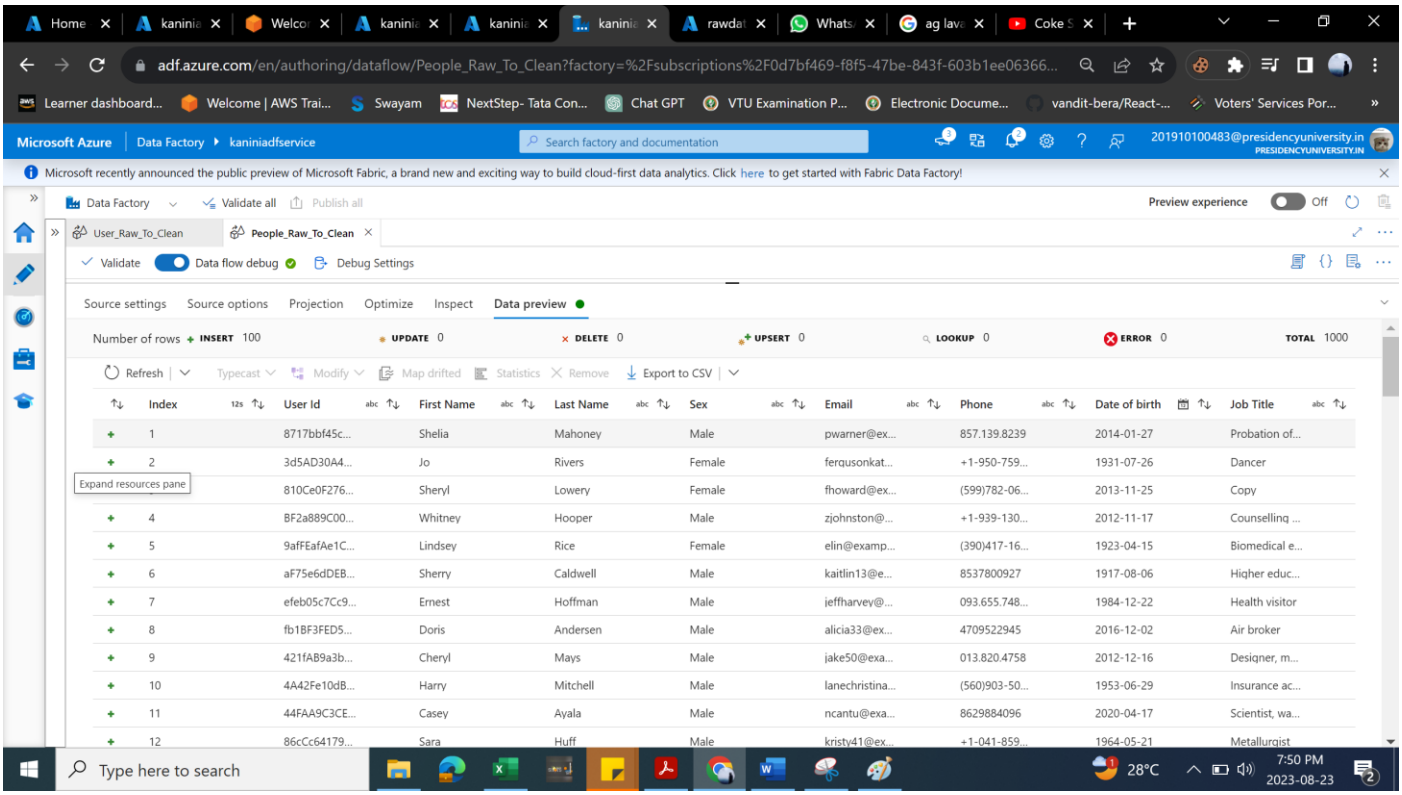
Number of rows + INSERT N/A | UPDATE N/A | DELETE N/A | UPSERT N/A | LOOKU

Refresh | Statistics | Export to CSV

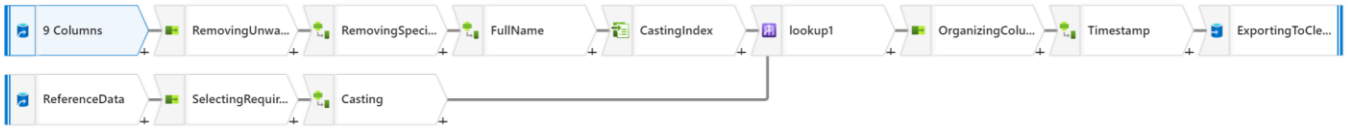
Country	Industry	Number of employ...	UpdatedOn
Anquilla	1	4292	2023-08-23 14:26:03
Australia	1	4155	2023-08-23 14:26:03
Bahamas	1	4914	2023-08-23 14:26:03
Belarus	1	3715	2023-08-23 14:26:03
Belgium	1	5038	2023-08-23 14:26:03
Benin	3	11611	2023-08-23 14:26:03
Bolivia	1	1312	2023-08-23 14:26:03
Botswana	1	7961	2023-08-23 14:26:03
Bouvet Island (Bouv...	1	7473	2023-08-23 14:26:03
Brazil	1	9315	2023-08-23 14:26:03
Burundi	1	1927	2023-08-23 14:26:03

Data set 3 :

Raw data :

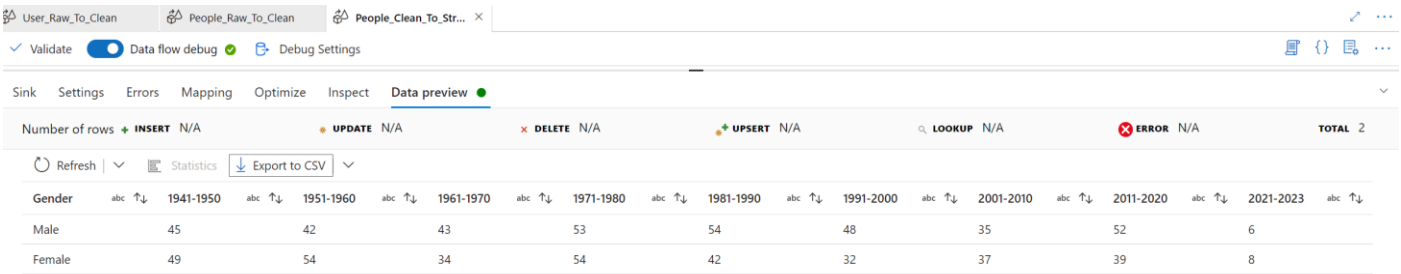


Raw to Clean data Transformations Used :



Clean to Structured data Transformations Used :

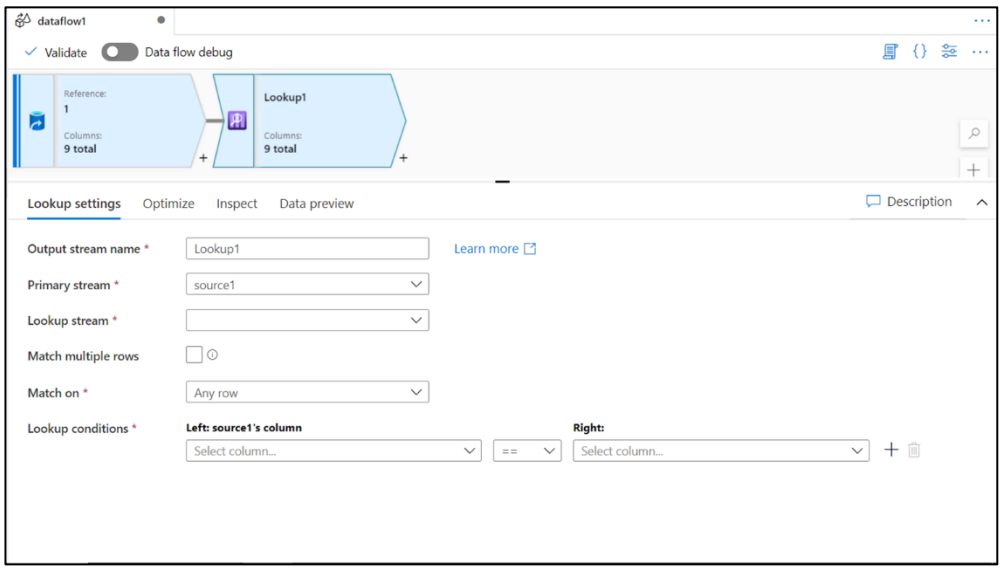
Final Output :



About Transformations I used :

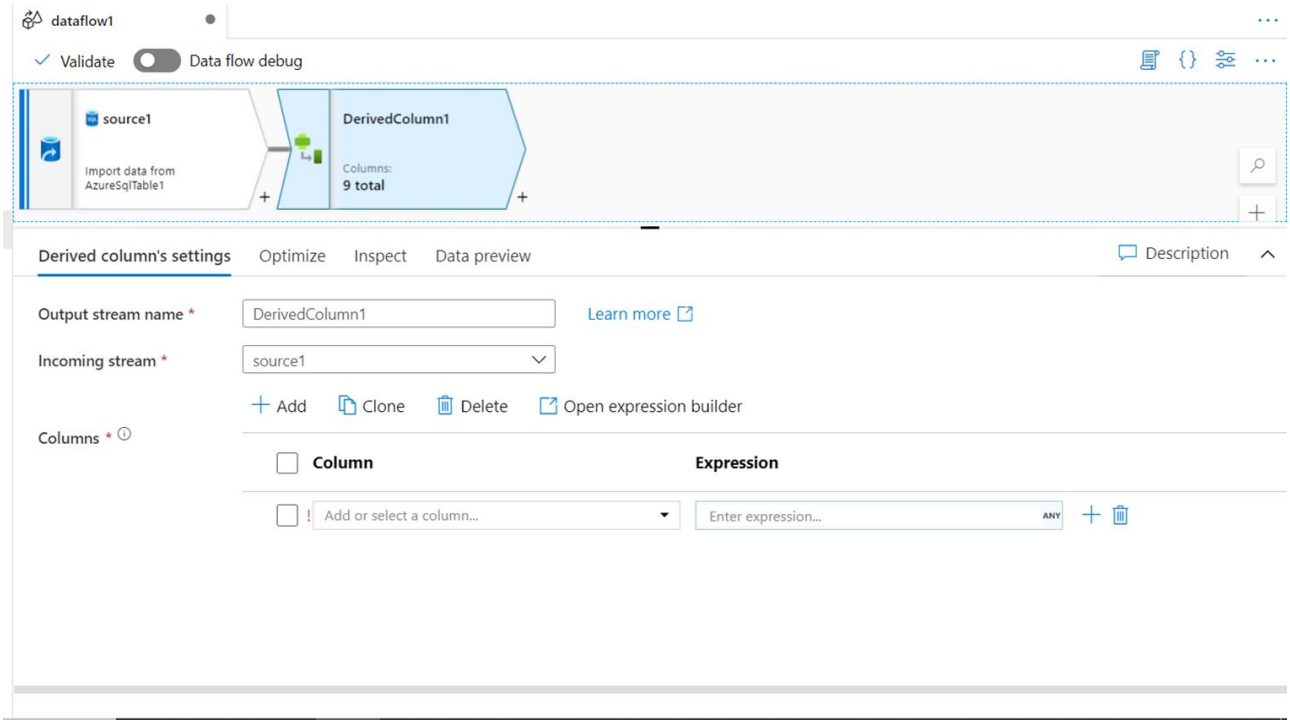
LOOKUP

In Azure Data Factory, a LookUp transformation is a data processing activity that retrieves a set of values from a specified dataset. This transformation is commonly used to perform reference data enrichment by querying external sources or datasets to augment existing data during ETL processes. LookUp transformations help in populating missing or additional information, enhancing the accuracy and context of the transformed data.



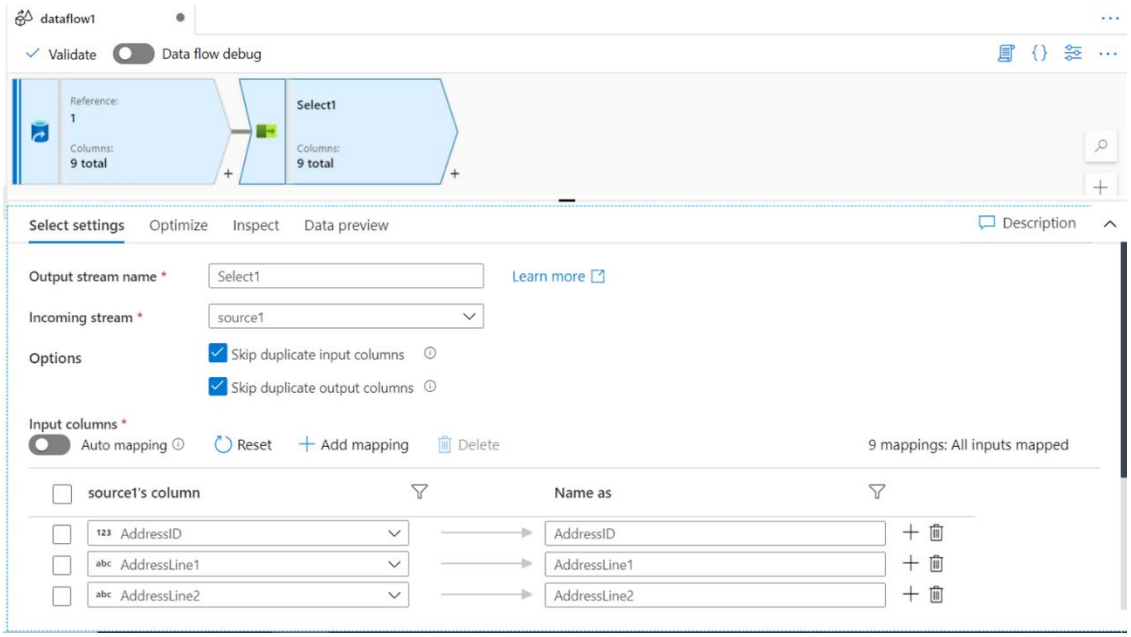
DERIVED COLUMN

The Derived Column transformation in Azure Data Factory is a data manipulation operation that creates new or modified columns based on expressions defined by the user. This transformation is used to transform data within a pipeline by applying calculations, string manipulations, conditional logic, and other computations to existing columns or constants. It enables the creation of derived attributes or data modifications before loading the data into the target destination, enhancing the data's quality and usefulness for downstream analytics or storage.



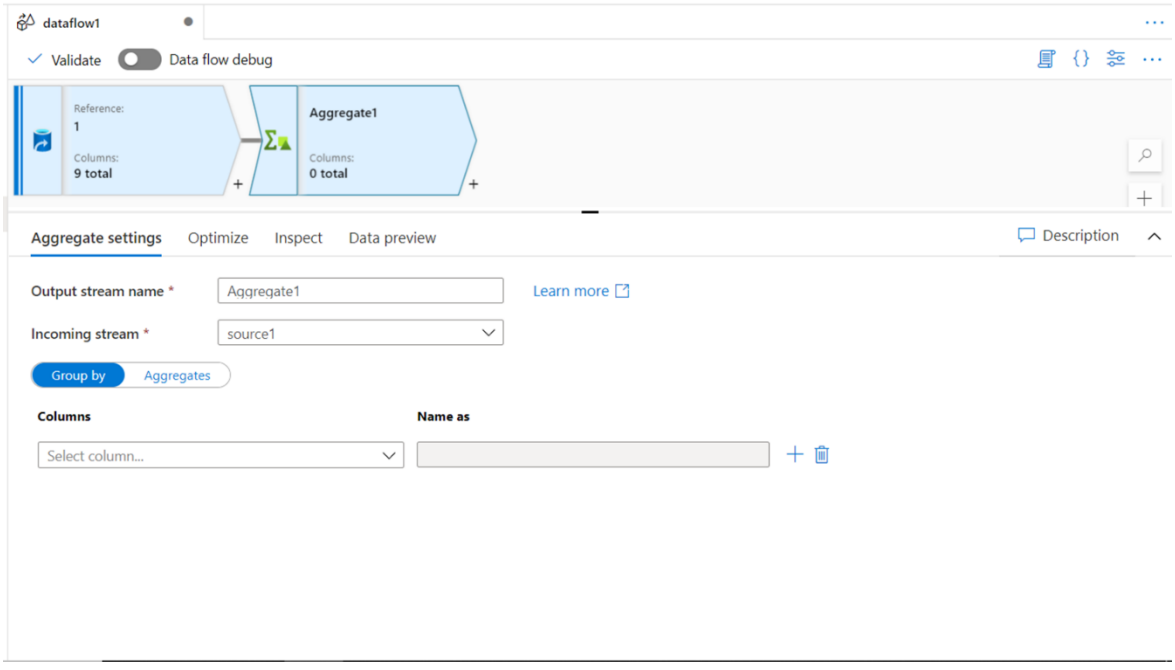
SELECT

The Select transformation in Azure Data Factory is a data projection operation that allows you to specify a subset of columns from a source dataset to be included in the output. This transformation is used to streamline data by reducing the number of columns being transferred or transformed in an ETL pipeline, optimizing performance and reducing storage costs. By selecting only the relevant attributes, the Select transformation enhances data processing efficiency and ensures that only necessary information is moved to the target destination for further analysis or storage.



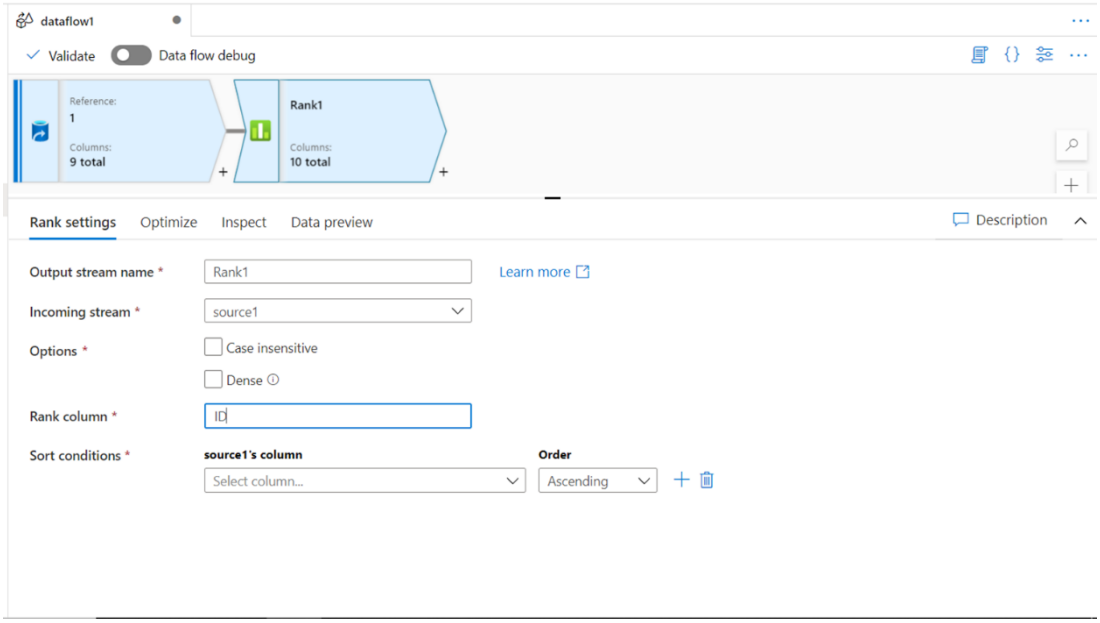
AGGREGATE

The Aggregate transformation in Azure Data Factory is a data processing step that groups and summarizes data based on specific criteria. This operation is commonly used to calculate summary statistics like sums, averages, counts, and more, on selected columns within a dataset. By condensing data into meaningful insights, the Aggregate transformation helps in data analysis and reporting, enabling better decision-making. This transformation is particularly useful for creating consolidated views and compact representations of data before it is loaded into target destinations for further processing or analysis.



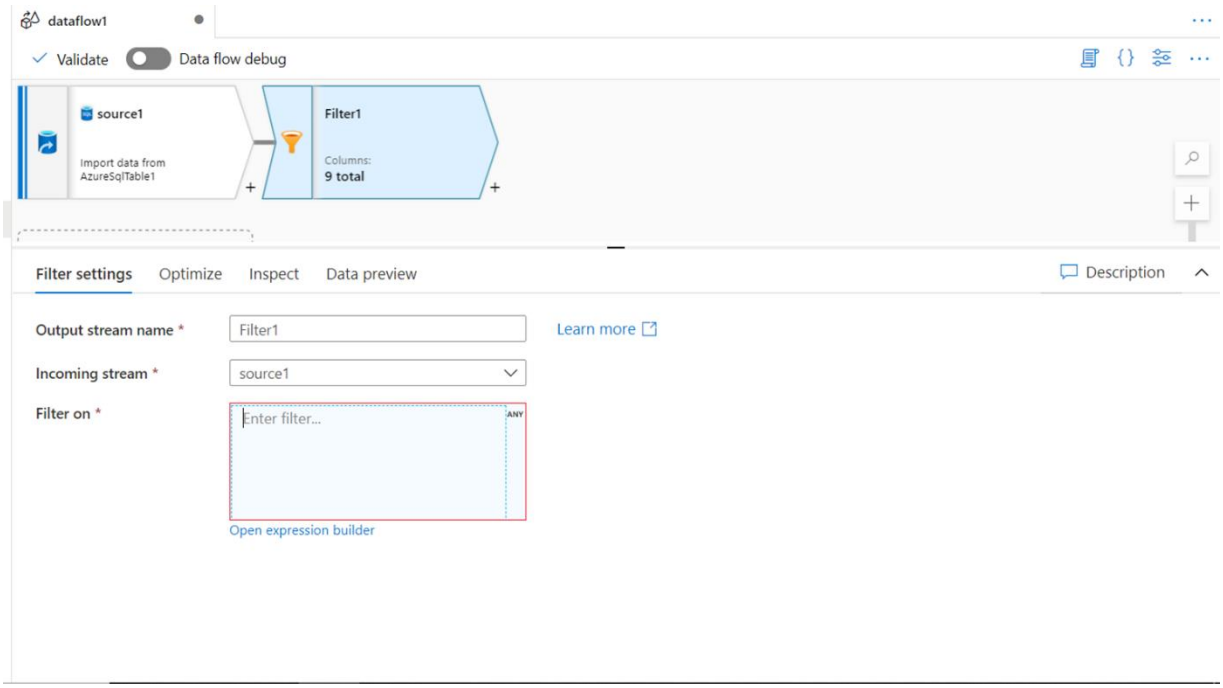
RANK

The Rank transformation in Azure Data Factory is a data processing operation that assigns a ranking to rows within a dataset based on specified criteria. This transformation is used to determine the relative position of rows in terms of a particular attribute, such as ordering sales data by revenue. Ranks can be assigned in ascending or descending order, and ties can be handled using different strategies. The Rank transformation is valuable for identifying top performers, outliers, or trends within datasets, aiding in decision-making and data analysis. It is often employed in scenarios where data needs to be prioritized or sorted before further processing or loading into target systems.



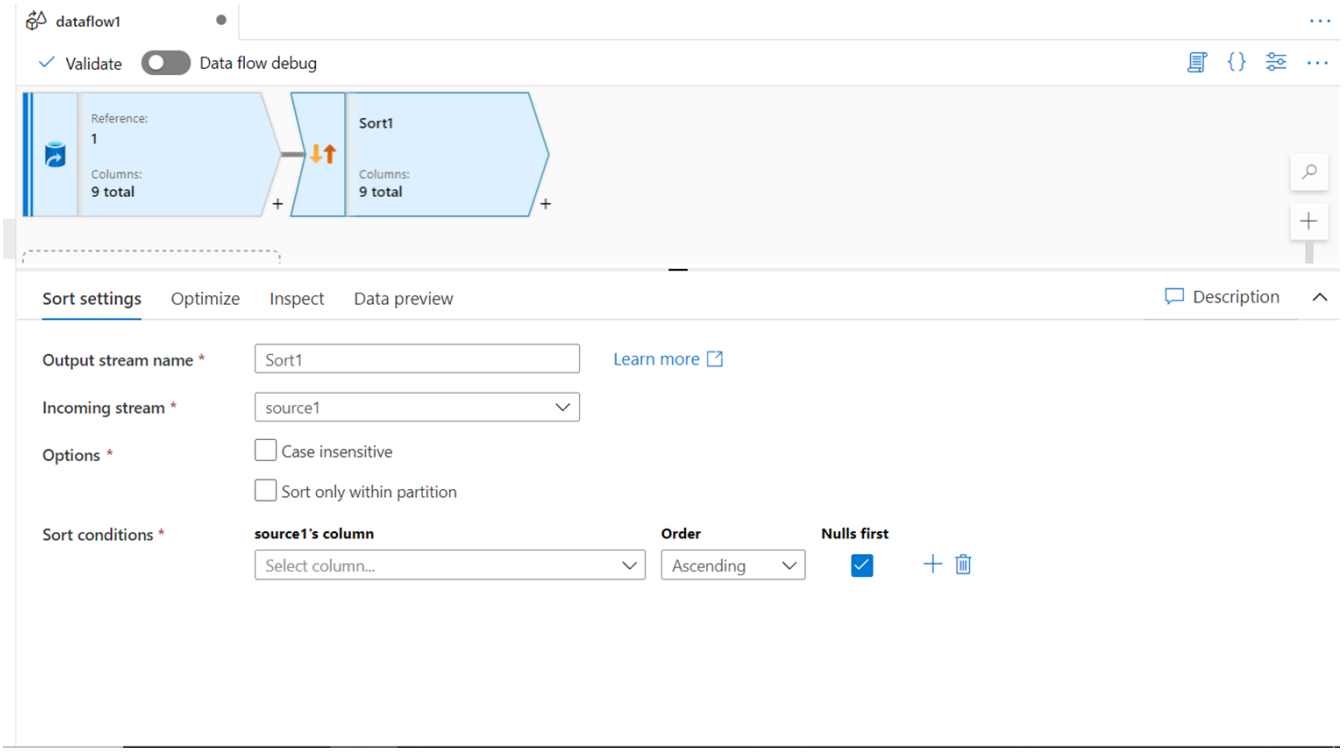
FILTER

The Filter transformation in Azure Data Factory is a data processing operation that allows you to selectively include or exclude rows from a dataset based on specified conditions. This transformation is used to refine data by applying logical expressions, comparisons, and other criteria to determine which rows meet certain requirements. Filtering helps in data quality improvement, reduction of unnecessary data transfer, and preparation of focused datasets for downstream analytics or storage. The Filter transformation is particularly useful for extracting relevant information and eliminating noise before loading data into target destinations or performing subsequent transformations and analyses.



SORT

The Sort transformation in Azure Data Factory is a data processing operation that arranges rows within a dataset in a specified order based on one or more columns. This transformation is used to establish a particular sequence for data, such as organizing records chronologically or alphabetically. Sorting aids in data organization, improves data presentation, and supports efficient querying and analysis. By arranging data in a defined order, the Sort transformation ensures consistency and enhances the usability of data when loading it into target systems.



CAST

The Cast transformation in Azure Data Factory is a data type conversion operation that changes the data type of a column in a dataset to a different type. This transformation is used to ensure compatibility between data types when transferring or processing data in ETL pipelines. Casting helps prevent data type conflicts and enables accurate calculations, comparisons, and transformations. By converting data to the appropriate format, the Cast transformation enhances the quality and integrity of data during movement between source and target systems, ensuring consistent and reliable data processing for downstream analytics or storage.

SINK

In Azure Data Factory, a Sink refers to the destination or target where data is loaded or written to as part of an ETL (Extract, Transform, Load) process. A Sink defines the location, format, and settings for storing transformed data, such as a database table, data lake, blob storage, or any other supported data repository. It represents the endpoint where data is delivered after undergoing transformations and processing in the pipeline. Sinks in Data Factory enable efficient and controlled data storage, ensuring that data is appropriately formatted and organized for downstream analytics, reporting, or further integration with other systems.