

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.

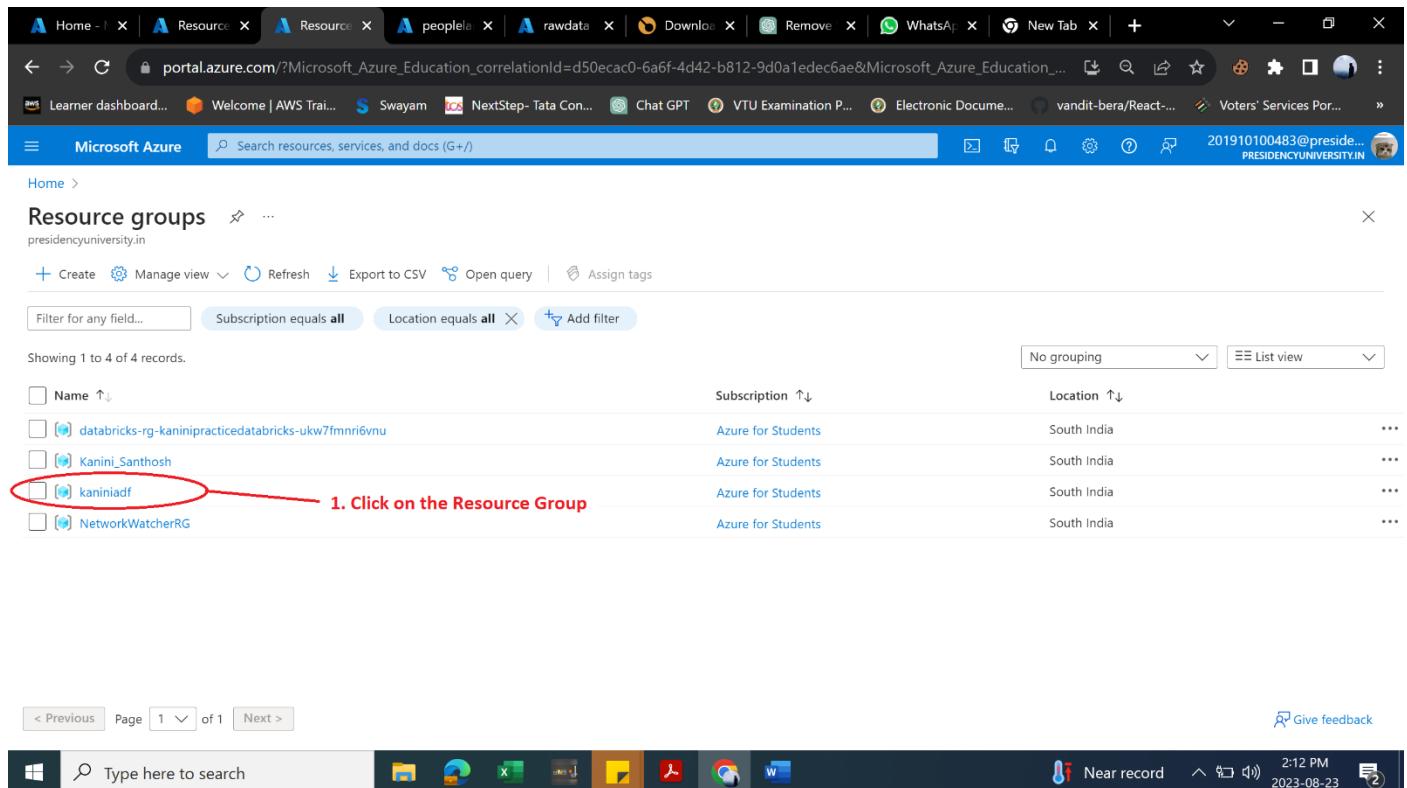
Tasks To be Done:

- 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.

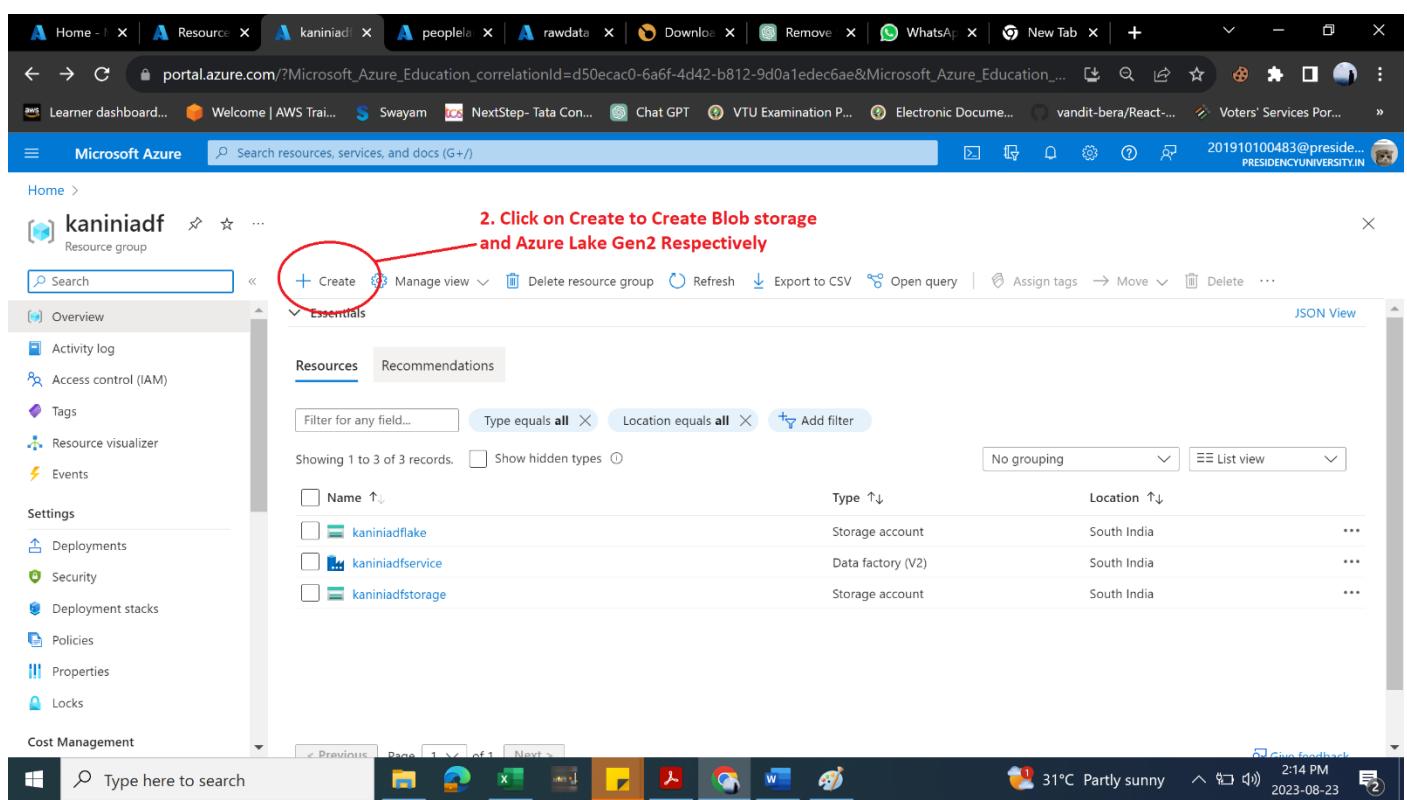
General Steps to Start the Project:

1: Create Two Storages in the Resource Group

1. BLOB STORAGE.
2. DATA LAKE.



The screenshot shows the Microsoft Azure Resource groups page. It lists four resource groups: 'databricks-rg-kaninipracticedatabricks-ukw7fmnri6vnu', 'Kanini_Santhosh', 'kaniniadf', and 'NetworkWatcherRG'. Each group is associated with 'Azure for Students' subscription and located in South India. A red circle highlights the 'kaniniadf' group, with the text '1. Click on the Resource Group' pointing to it.



The screenshot shows the Microsoft Azure Resource group details page for 'kaniniadf'. The left sidebar shows various options like Overview, Activity log, Access control (IAM), Tags, Resource visualizer, Events, Settings, Deployments, Security, Deployment stacks, Policies, Properties, and Locks. The main area shows the 'kaniniadf' resource group with three resources listed: 'kaniniadflake' (Storage account), 'kaniniadfservice' (Data factory (V2)), and 'kaniniadfstorage' (Storage account). A red circle highlights the '+ Create' button at the top, with the text '2. Click on Create to Create Blob storage and Azure Lake Gen2 Respectively' pointing to it.

Get Started

Service Providers

Management

Private Marketplace

Private Offer Management

My Marketplace

Favorites

Recently created

Private plans

Categories

Storage (77)

Compute (47)

storage account

Pricing : All X Operating System : All X Publisher Type : All X Product Type : All X

Azure services only

Showing 1 to 20 of 172 results for 'storage account': [Clear search](#)

3. Search for Blob storage

4. Select this and create the blob storage

Storage account

Storage Account Using ARM Template

Azure Storage Mover

APEX Protection Storage for Microsoft Azure (DDVE)

Storage Account Using ARM

Type here to search

31°C Partly sunny 2:17 PM 2023-08-23

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 .

Home > Resource groups >

kaniniadf Resource group

Search

+ Create Manage view Delete resource group Refresh Export to CSV Open query Assign tags Move Delete ...

Overview

Activity log

Access control (IAM)

Tags

Resource visualizer

Events

Settings

Deployments

Security

Deployment stacks

Policies

Properties

Locks

Cost Management

Essentials

Resources Recommendations

Filter for any field... Type equals all X Location equals all X Add filter

Showing 1 to 3 of 3 records. Show hidden types

No grouping List view

Name	Type	Location	...
kaniniadflake	Data Lake Storage	Storage account	South India
kaniniadfservice	Data Factory	Data factory (V2)	South India
kaniniadfstorage	Blob Storage	Storage account	South India

Type here to search

31°C Partly sunny 2:22 PM 2023-08-23

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

The screenshot shows the Microsoft Azure portal interface. The user is in the 'Containers' section of the 'kaniniadfstorage' storage account. A red circle highlights the '+ Container' button in the top-left corner of the main content area. Another red circle highlights the 'Containers' link under the 'Data storage' heading in the left sidebar. A callout text '8. click here inside the Storage container' points to the 'Containers' link. A callout text '9. Once Container page is opened, Click here to create new container' points to the '+ Container' button. A callout text '10. Name the container as per the choice' points to the 'Name' input field in the 'New container' dialog, which has 'landing' typed into it. A callout text '11. Click on the "Create" Button' points to the 'Create' button in the bottom right of the dialog.

kaniniadfstorage | Containers

New container

Name * landing

Public access level Private (no anonymous access)

10. Name the container as per the choice

11. Click on the "Create" Button

Name	Last modified	Public access level
\$logs	8/17/2023, 2:42:37 PM	Private
landing	8/17/2023, 2:45:18 PM	Private
organizationlanding	8/23/2023, 1:39:29 AM	Private
peoplelanding	8/23/2023, 11:15:42 AM	Private

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.

The screenshot shows the Microsoft Azure portal interface. The user is in the 'Containers' section of the 'kaniniadflake' storage account. A red circle highlights the 'Containers' link under the 'Data storage' heading in the left sidebar. A callout text '9. Inside my Lake Storage i Have created 3 Containers' points to the list of containers in the main content area. The list shows four containers: '\$logs', 'cleandata', 'rawdata', and 'structureddata'. The 'rawdata' and 'structureddata' entries are highlighted with a red box.

kaniniadflake | Containers

9. Inside my Lake Storage i Have created 3 Containers

Name	Last modified	Public access level	Lease state
\$logs	8/17/2023, 2:44:30 PM	Private	Available
cleandata	8/17/2023, 2:50:30 PM	Private	Available
rawdata	8/17/2023, 2:50:21 PM	Private	Available
structureddata	8/22/2023, 5:57:07 PM	Private	Available

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

The screenshot shows the Microsoft Azure Blob Storage interface. On the left, a sidebar lists 'Overview', 'Diagnose and solve problems', 'Access Control (IAM)', 'Properties', and 'Metadata'. The main area shows a table with one row: 'user.parquet' (Modified: 8/22/2023, 10:32:29 ...). A red box highlights the 'landing Container' link in the breadcrumb navigation. Another red box highlights the 'Upload' button. A callout '10. Now inside landing container, Click on the "Upload" Button' points to the 'Upload' button. A red box highlights the 'Browse for files' button in the 'Upload blob' dialog, with a callout '11. Click here to Browse and upload bad datasets from your local machine'. A red box highlights the 'generic-fraud.csv' file in the dialog, with a callout '12. once you have selected the data, click on "Upload" Button'. A callout '13. The Uploaded data will be visible here.' points to the table listing the uploaded file.

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

The screenshot shows the Microsoft Azure Resource Groups interface. The left sidebar includes 'Activity log', 'Access control (IAM)', 'Tags', 'Resource visualizer', 'Events', 'Deployments', 'Security', 'Deployment stacks', 'Policies', 'Properties', and 'Locks'. The main area displays a table of resources under the 'kaniniadf' resource group. The table has columns for 'Name', 'Type', and 'Location'. Three resources are listed: 'kaniniadflake' (Storage account, South India), 'kaniniadbservice' (Data factory (V2), South India), and 'kaniniadstorage' (Storage account, South India). A red box highlights the 'kaniniadbservice' entry, with a callout '11. Click on the data Factory , to open it'.

A screenshot of the Microsoft Azure portal showing the 'kaniniadfservice' Data factory (V2) details. The 'Essentials' section shows the resource group is 'kaniniadfservice', status is 'Succeeded', location is 'South India', and it's associated with 'Azure for Students'. A large blue icon of a factory building is displayed. Below it, the text 'Azure Data Factory Studio' is centered, with a red circle highlighting the 'Launch studio' button. A tooltip says 'Click here to Launch the Data factory'.

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

A screenshot of the Azure Data Factory 'Manage' page. On the left sidebar, under the 'Data Factory' section, the 'Manage' option is highlighted with a red circle. The main area shows the 'Linked services' section, which lists three existing linked services: 'ls_blob_landing' (Azure Blob Storage), 'ls_lake_cleandata' (Azure Data Lake Storage Gen2), and 'ls_lake_rawdata' (Azure Data Lake Storage Gen2). A red circle highlights the '+ New' button, with the text '14. Click on "New" to create a linked services' positioned above it. Another red circle highlights the 'Linked services' link in the navigation bar, with the text '13. Now Click on "Linked Services"' positioned below it.

The screenshot shows the Microsoft Azure Data Factory interface. On the left, the navigation menu includes Home, Author, Monitor, Manage, and Learning Center. Under the 'Connections' section, 'Linked services' is selected. The main area displays a list of existing linked services:

Name	Type
ls_blob_landing	Azure Blob Storage
ls_lake_cleandata	Azure Data Lake Storage Gen2
ls_lake_rawdata	Azure Data Lake Storage Gen2

To the right, a 'New linked service' dialog is open, titled 'New linked service'. It shows a search bar with 'blob' typed in. Below it, under the 'Data store' tab, a list of options is shown, with 'Azure Blob Storage' highlighted and circled in red. A red arrow points from the text 'Search "Blob"' to this circled item.

The screenshot shows the Microsoft Azure Data Factory interface, similar to the previous one. The 'Manage' section of the navigation menu is selected. The 'Linked services' section is highlighted. A red callout box with the text 'Fill up the Required Fields required to create the linked service and create linked service' points to the 'Create' button at the bottom right of the 'New linked service' dialog.

The 'New linked service' dialog contains the following fields:

- Name:** AzureBlobStorage1 (circled in red)
- Description:** (empty)
- Connect via integration runtime:** AutoResolveIntegrationRuntime
- Authentication type:** Account key
- Connection string:** (selected)
- Account selection method:** From Azure subscription (radio button selected)
- Azure subscription:** Select all
- Storage account name:** (empty)
- Additional connection properties:** (empty)

At the bottom of the dialog, there are 'Create' and 'Back' buttons, with 'Create' also circled in red.

The screenshot shows the Microsoft Azure Data Factory interface. On the left, the navigation pane is open with 'Data Factory' selected. Under 'Connections', 'Linked services' is also selected. The main area displays a list of linked services with three items listed: 'ls_blob_landing', 'ls_lake_cleandata', and 'ls_lake_rawdata'. A red arrow points from the text 'Follow the same steps to create linked services for "Azure lake storage"' to the 'ls_lake_cleandata' item. Another red circle highlights the 'Continue' button at the bottom right of the dialog.

Follow the same steps to create linked services for "Azure lake storage"

New linked service

Azure Data Lake Storage Gen2

Azure Data Lake Storage Gen2

Continue

Linked Services are Successfully created.

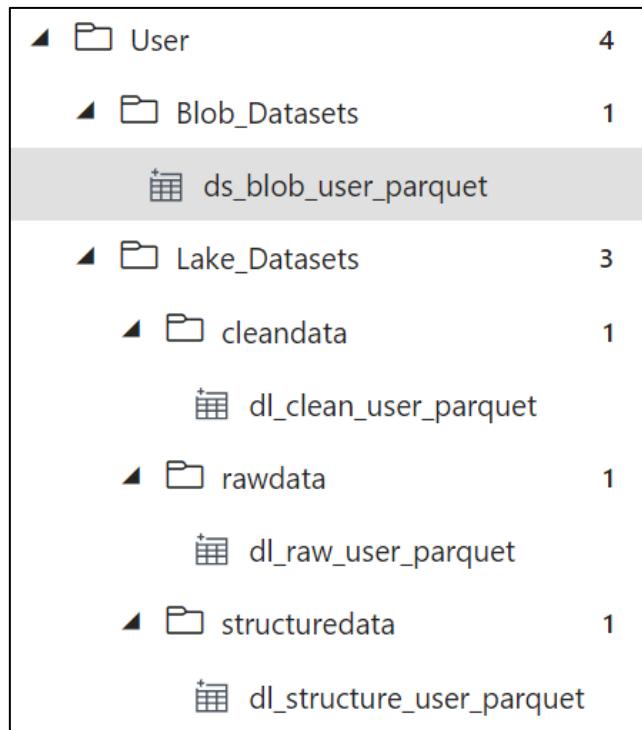
The screenshot shows the Microsoft Azure Data Factory interface after the linked services have been created. The 'Linked services' section now lists three items: 'ls_blob_landing', 'ls_lake_cleandata', and 'ls_lake_rawdata'. A red box highlights these three items. A red arrow points from the text 'Linked Services are created' to the highlighted items.

Linked Services are created

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

The screenshot shows the Microsoft Azure Data Factory interface. On the left sidebar, under 'Author', there is a link labeled 'Click here to Go to datasets'. In the main 'Factory Resources' pane, there is a section titled 'Datasets' which contains a list of datasets. One dataset, 'blob_user_parquet', is highlighted with a red circle and a red arrow pointing to it from the text 'Right Click on this !!!'. Another red arrow points to a 'New dataset' option within the 'Datasets' list, with the text 'Click here to create the dataset' next to it. The interface also shows other sections like 'Lake_Datasets', 'Data flows', and 'User'. At the bottom, there is a taskbar with various icons and a system tray showing the date and time.

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



Now create a Dataflow for adding Transformations:

The screenshot shows the Microsoft Azure Data Factory Author interface. On the left sidebar, the 'Author' option is selected and highlighted with a red box. A red arrow points from the text 'Now again click here to open Author' to the 'Author' button. In the main workspace, under 'Data flows', a red box highlights 'New data flow'. Another red arrow points from the text 'Right click on Data flows' to the 'Data flows' section. A third red arrow points from the text 'Click on "New data flow" to create data flow' to the 'New data flow' button.

The screenshot shows the Microsoft Azure Data Factory Author interface. The 'Author' option is selected and highlighted with a red box. A red arrow points from the text 'Use the data flow debug for Preview the content' to the 'Data flow debug' toggle switch. Another red arrow points from the text 'Click on Add source to add the Entry Point' to the 'Add Source' button, which is highlighted with a red box.

At the End add the Sink to export the dataflow

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.

Right click on the Pipelines

Click on New Pipeline to create a pipeline

The screenshot shows the Microsoft Azure Data Factory pipeline editor. On the left, the 'Factory Resources' sidebar lists 'Pipelines' (including 'pipeline9') and 'Datasets'. In the center, the 'Activities' pane shows a list of activities: 'Move and transform' (with 'Copy data' and 'Data flow' options), 'Synapse', 'Azure Data Explorer', 'Azure Function', 'Batch Service', 'Databricks', 'Data Lake Analytics', 'General', 'HDInsight', 'Iteration & conditionals', 'Machine Learning', and 'Power Query'. A red circle highlights the 'Data flow' option under 'Move and transform'. A red arrow points from this circle to a callout text: 'Select the Dataflow ,drag and drop option on this sheet'. To the right, the 'Properties' panel for 'pipeline9' is open, showing the 'General' tab with 'Name' set to 'pipeline9'. A red circle highlights the 'Debug' button in the top toolbar. A red arrow points from this circle to a callout text: 'Click on this option'.

The screenshot shows the Microsoft Azure Data Factory pipeline editor. The 'Activities' pane now displays a 'Data flow' activity named 'Data flow1'. The 'Properties' panel on the right shows the 'Settings' tab for 'Data flow1', which includes sections for 'User_Raw_To_Clean', 'User_Clean_To_Structure', 'organization_Raw_To_Clean', 'organization_Clean_To_Structure', 'Logging level' (set to 'Verbose'), 'Sink properties', and 'Staging'. A red circle highlights the 'User_Clean_To_Structure' section. A red arrow points from this circle to a callout text: 'In Settings select the dataflow created and click on debug option'. The 'Debug' button in the top toolbar is also highlighted with a red circle.

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 :

Screenshot of Microsoft Data Factory Data Flow preview showing raw data from 'User_Raw_To_Clean' dataset:

regis...	id	first...	last...	email	gender	ip_ad...	cc	coun...	birth...	salary	12...	title	abc	cor
2016...	1	Aman...	Jordan	ajord...	Female	1.197...	6759...	Indon...	3/8/1...	4975...	Intern...	1E+		
2016...	2	Albert	Free...	afree...	Male	218.1...		Canada	1/16/...	1502...	Acco...			
2016...	3	Evelyn	Morg...	emor...	Female	7.161...	6767...	Russia	2/1/1...	1449...	Struct...			
2016...	4	Denise	Riley	driley...	Female	140.3...	3576...	China	4/8/1...	9026...	Senio...			
2016...	5	Carlos	Burns	cburn...		169.1...	5602...	South...		NULL				
2016...	6	Kathr...	White	kwhit...	Female	195.1...	3583...	Indon...	2/25/...	6922...	Acco...			
2016...	7	Samuel	Holmes	shol...	Male	232.2...	3582...	Portu...	12/18...	1424...	Senio...			
2016...	8	Harry	Howell	hhow...	Male	91.23...		Bosni...	3/1/1...	1864...	Web ...			
2016...	9	Jose	Foster	jfoste...	Male	132.3...	3574...	South...	3/27/...	2310...	Softw...	1E+		
2016...	10	Emily	Stewart	estew...	Female	143.2...		Nigeria	1/28/...	2723...	Healt...			
2016...	11	Susan	Perkins	sperki...	Female	180.8...	3573...	Russia		2100...				

Raw to Clean data Transformations Used :



Clean to Structured data Transformations Used :



Final Output :

Screenshot of Microsoft Data Factory Data Flow preview showing the final output from 'User_Clean_To_Struct...' dataset:

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	205883.956666666667	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 Practitioner	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	166215.19000000003	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		X DELETE 0		+ UPSERT 0		Q LOOKUP 0		X ERROR 0		TOTAL 100			
↻ Refresh ▼	Typecast ▼	✖ Modify ▼	☒ Map drifted ☒	Statistics X	Remove X	⬇️ Export to CSV ▾											
↑ ↓ Index	12s ↑ ↓	Organization...	abc ↑ ↓	Name	abc ↑ ↓	Website	abc ↑ ↓	Country	abc ↑ ↓	Description	abc ↑ ↓	Founded	12s ↑ ↓	Industry	abc ↑ ↓	Number of ...	12s ↑ ↓
+ 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		Synchroniz...		1991		Transportati...		7870			
+ 6	cC75116fe...	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		X DELETE N/A		+ UPSERT N/A		Q LOOKUP N/A					
↻ Refresh ▼	Statistics ⬇️	Export to CSV ▾													
Country	abc ↑ ↓	Industry	abc ↑ ↓	Number of employ...	abc ↑ ↓	UpdatedOn	abc ↑ ↓								
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 :

The screenshot shows the Microsoft Azure Data Factory Data Preview interface. The top navigation bar includes links for Home, kanini, Welcome | AWS Train..., Swayam, NextStep- Tata Con..., Chat GPT, VTU Examination P..., Electronic Docume..., vandit-bera/React..., and Voters' Services Por... The URL is adf.azure.com/en/authoring/dataflow/People_Raw_To_Clean?factory=%2Fsubscriptions%2F0d7bf469-f8f5-47be-843f-603b1ee06366.... The main area displays a table of raw data with 100 rows. The columns include User Id, First Name, Last Name, Sex, Email, Phone, Date of birth, and Job Title. The data preview tab is selected, showing various operations like INSERT, UPDATE, DELETE, UPSERT, LOOKUP, and ERROR counts. The bottom status bar shows the date as 2023-08-23 and the time as 7:50 PM.

Raw to Clean data Transformations Used :



Clean to Structured data Transformations Used :



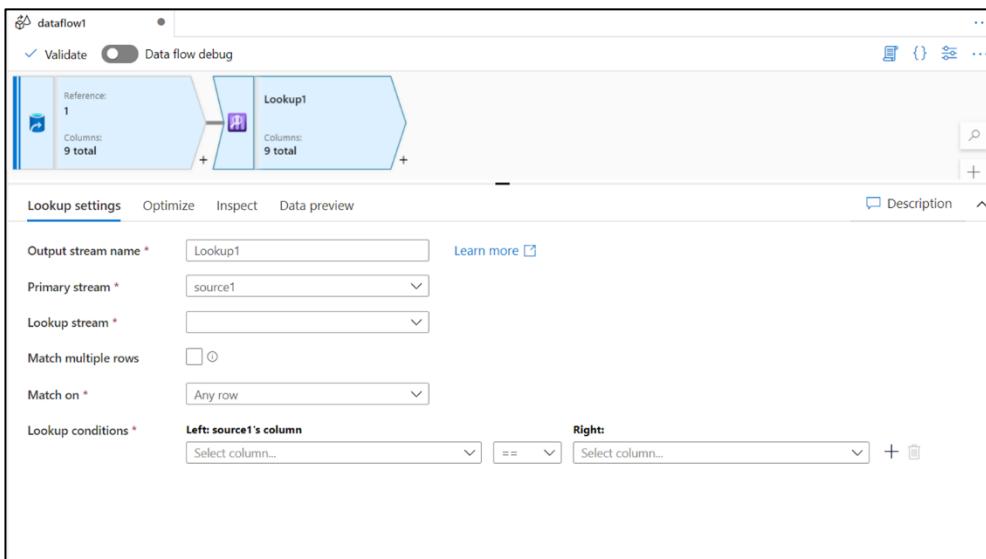
Final Output :

The screenshot shows the Microsoft Azure Data Factory Data Preview interface. The top navigation bar includes links for Home, kanini, Welcome | AWS Train..., Swayam, NextStep- Tata Con..., Chat GPT, VTU Examination P..., Electronic Docume..., vandit-bera/React..., and Voters' Services Por... The URL is adf.azure.com/en/authoring/dataflow/People_Clean_To_Structured?factory=%2Fsubscriptions%2F0d7bf469-f8f5-47be-843f-603b1ee06366.... The main area displays a table of structured data with 2 rows. The columns include Gender, 1941-1950, 1951-1960, 1961-1970, 1971-1980, 1981-1990, 1991-2000, 2001-2010, 2011-2020, and 2021-2023. The data preview tab is selected, showing various operations like INSERT, UPDATE, DELETE, UPSERT, LOOKUP, and ERROR counts. The bottom status bar shows the date as 2023-08-23 and the time as 7:50 PM.

About Transformations I used :

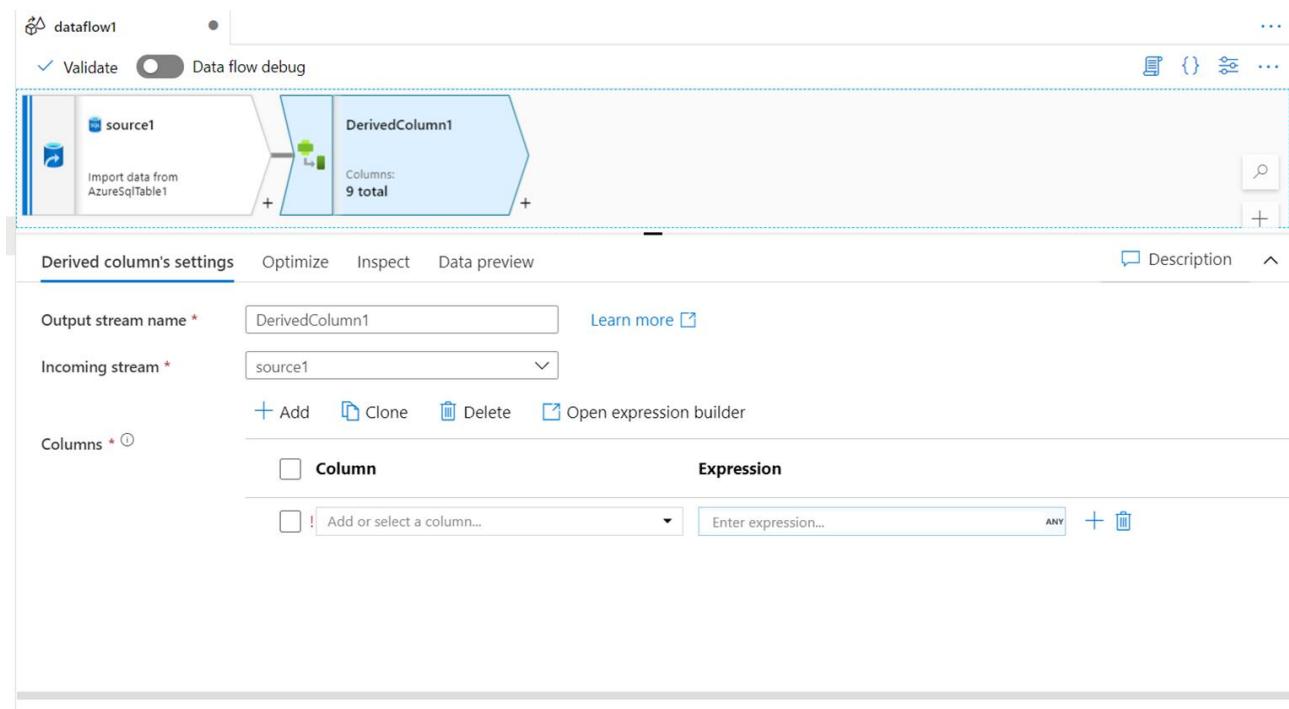
LOOKUP

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.



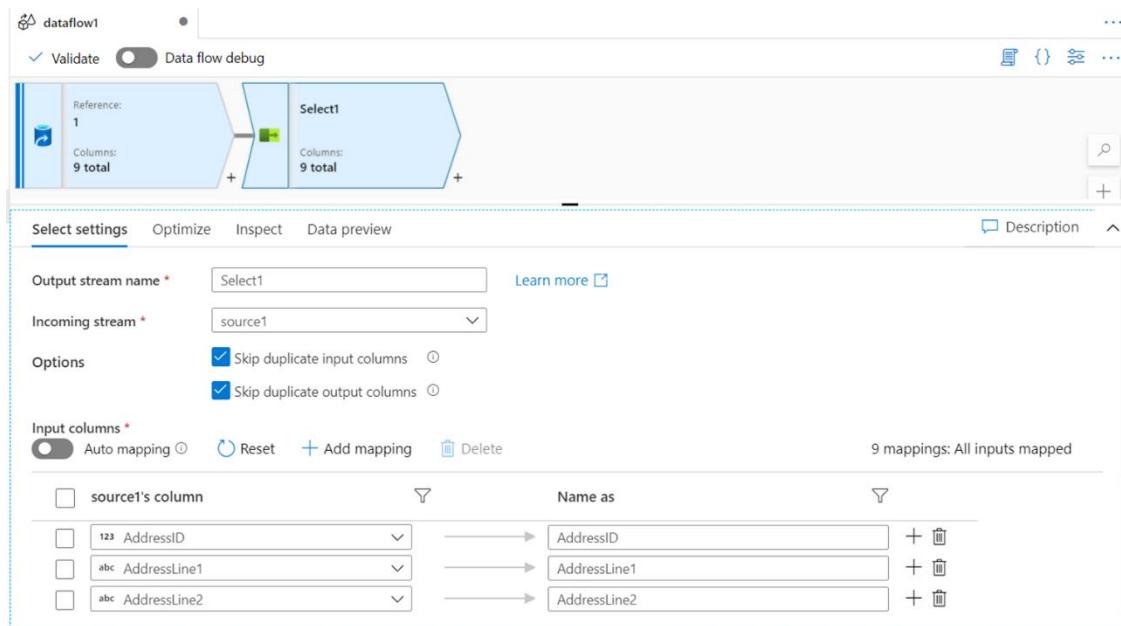
DERIVED COLUMN

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.



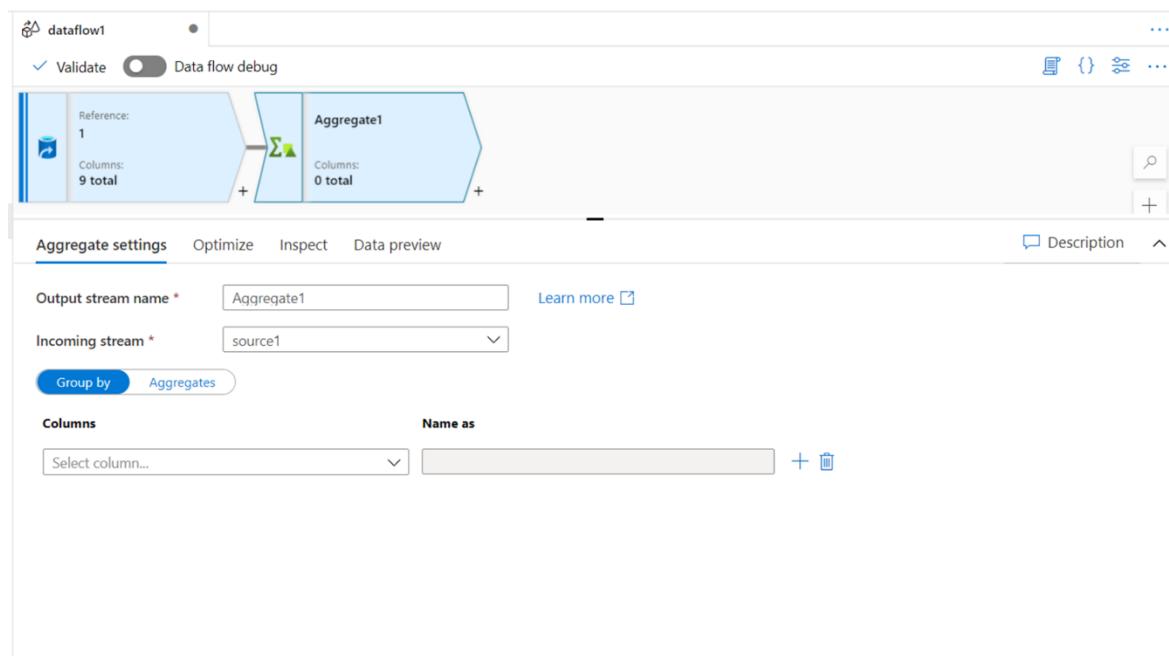
SELECT

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.



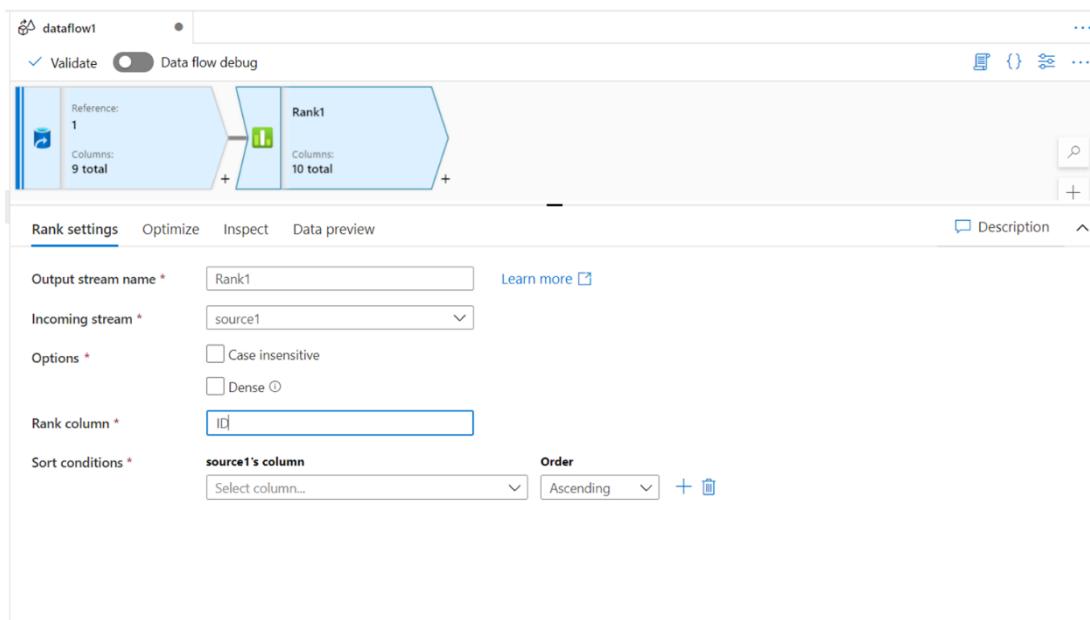
AGGREGATE

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.



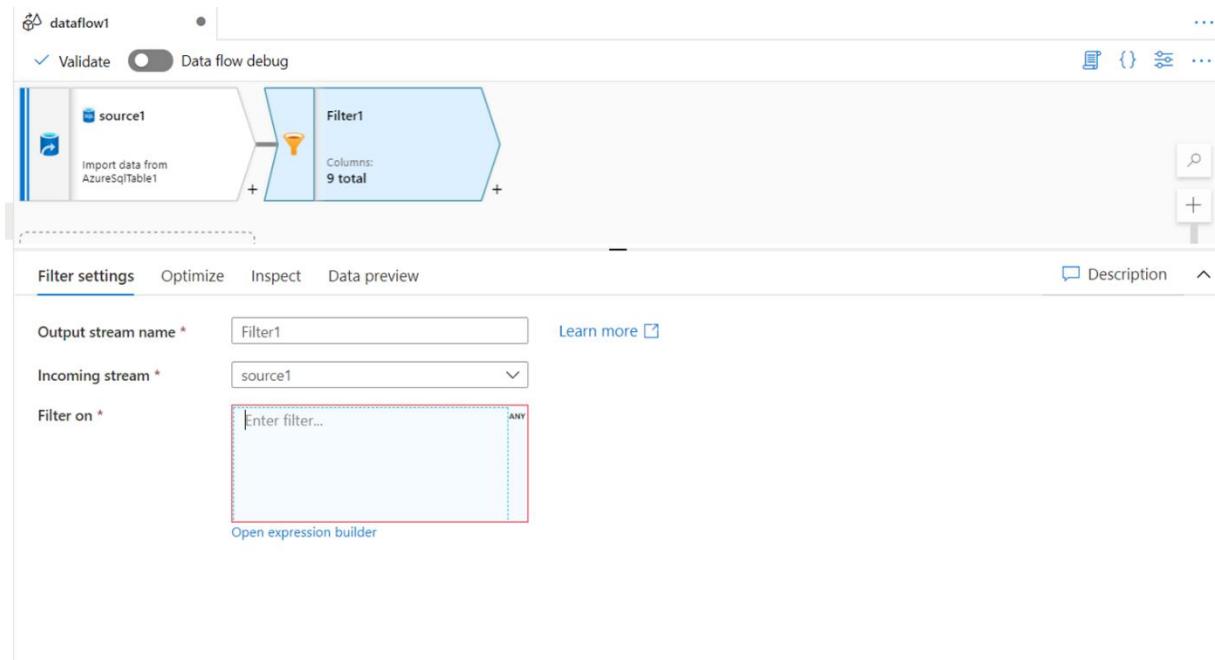
RANK

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.



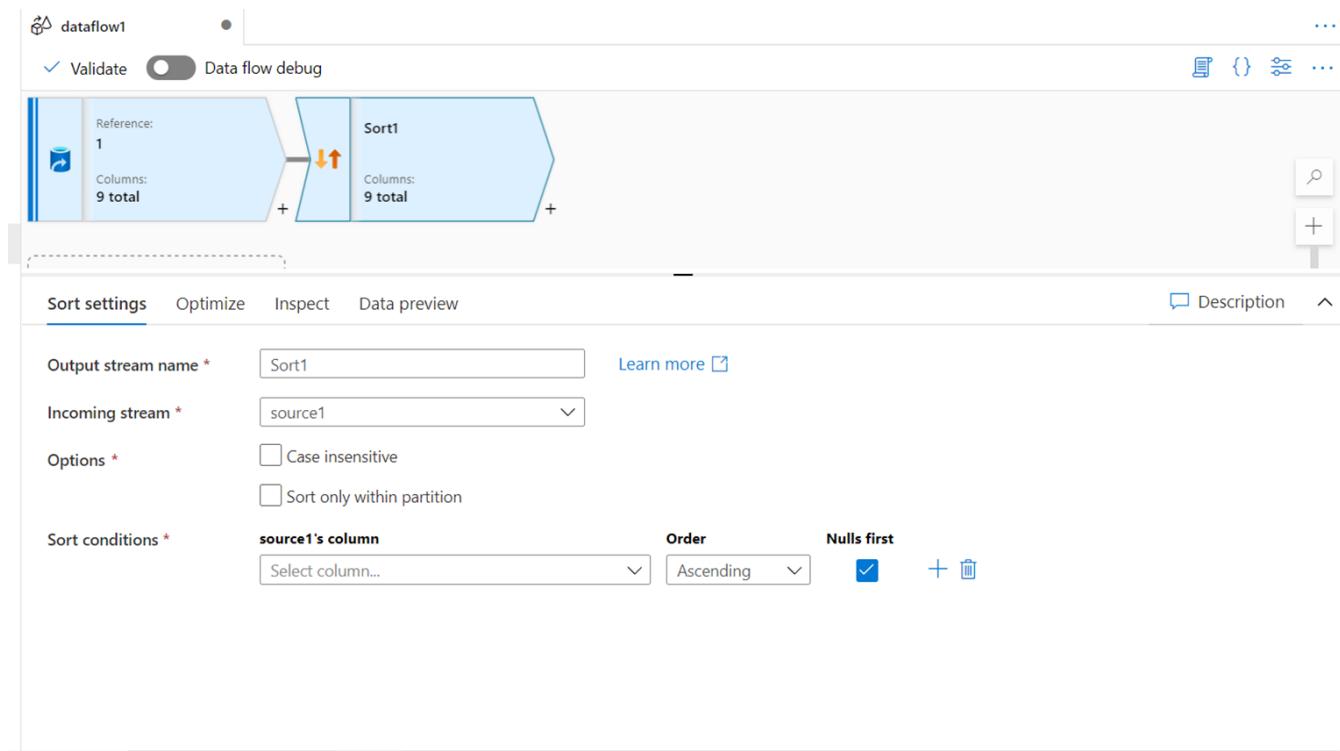
FILTER

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.