Azure Mini Project: End-to-end ETL

Yeh, Margaret

Questions:

1. **Why should one use Azure Key Vault when working in the Azure environment? What are the alternatives to using Azure Key Vault? What are the pros and cons of using Azure Key Vault?**

   Azure key vault is an open-source password manager used for security. Some benefits to using it is that apps and/or people cannot have direct access to the passwords or keys. However, this can have increased costs for APIs and result in higher processing times.

2. **How do you achieve the loop functionality within an Azure Data Factory pipeline? Why would you need to use this functionality in a data pipeline?**

   Loops allow for a specific activity to be repeated until the set condition is met. This can be useful for many reasons, such as needing to iterate through a list. They can be created by using the ForEach Activity.

3. **What are expressions in Azure Data Factory? How are they helpful when designing a data pipeline (please explain with an example)?**

   Expressions take some parameterized input and convert it to some output. It can be used for mapping data flow in a data pipeline, by taking some parameter and mapping it to a set type. This allows data to be regulated so that it fits a specific schema. It can also be used to remove null values in the data. Another use of expressions is to create dynamic paths for the pipeline that would take a parameterized input as part of the path.

4. **What are the pros and cons of parametrizing a dataset in Azure Data Factory pipeline's activity?**

Pros – Parameterizing the dataset changes the attributes of the dataset into parameters. This can allow parameters from many different data sources to become standardized, which can help avoid the need to create multiple different datasets. With parameterization, different datasets can work in the pipeline as long as their differences are not too great.

Cons – If differences are too great, such as in cases where the schemas or file formats of the datasets do not match, then problems can occur when trying to work with them in the pipeline.

5. **What are the different supported file formats and compression codecs in Azure Data Factory? When will you use a Parquet file over an ORC file? Why would you choose an AVRO file format over a Parquet file format?**

Azure Data Factory supports:
Avro format
Binary format
Delimited text format
Excel format
JSON format
ORC format
Parquet format
XML format

Comparing Parquet and ORC, parquet uses a column-wise data storage method, while ORC is arranged as indexed stripes of data. ORC is good at reading large data quickly and has better compression ability than the Parquet format, so it is often used for read-heavy applications. Parquet is also good at compression, but not to the extent of ORC.

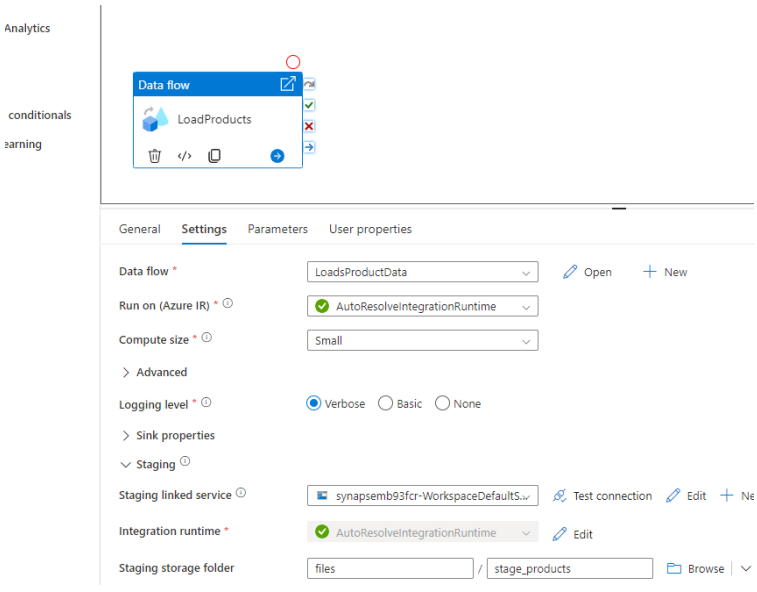
However, Parquet is open-source and accessible in many languages, which contributes to its high popularity. Parquet is heavily used for OLAP cases with OLTP databases.

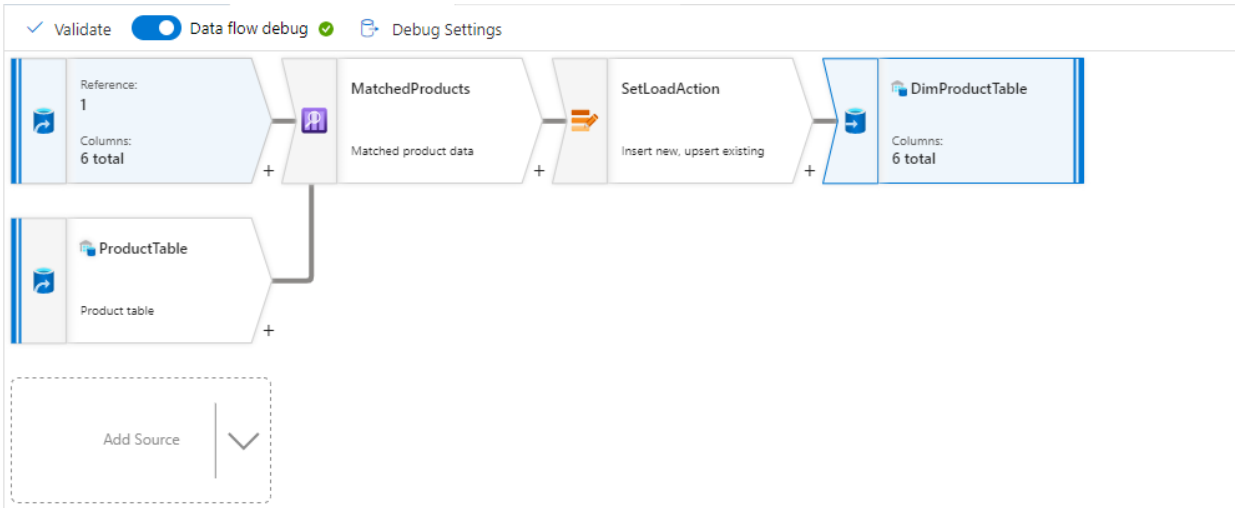
Comparing Parquet and AVRO, unlike Parquet, AVRO uses a row-based format to store data. Therefore, it is good for dynamic schemas that change over time, since it can adjust to additional fields or missing values. AVRO works well in the Hadoop ecosystem and is useful for write-heavy applications.

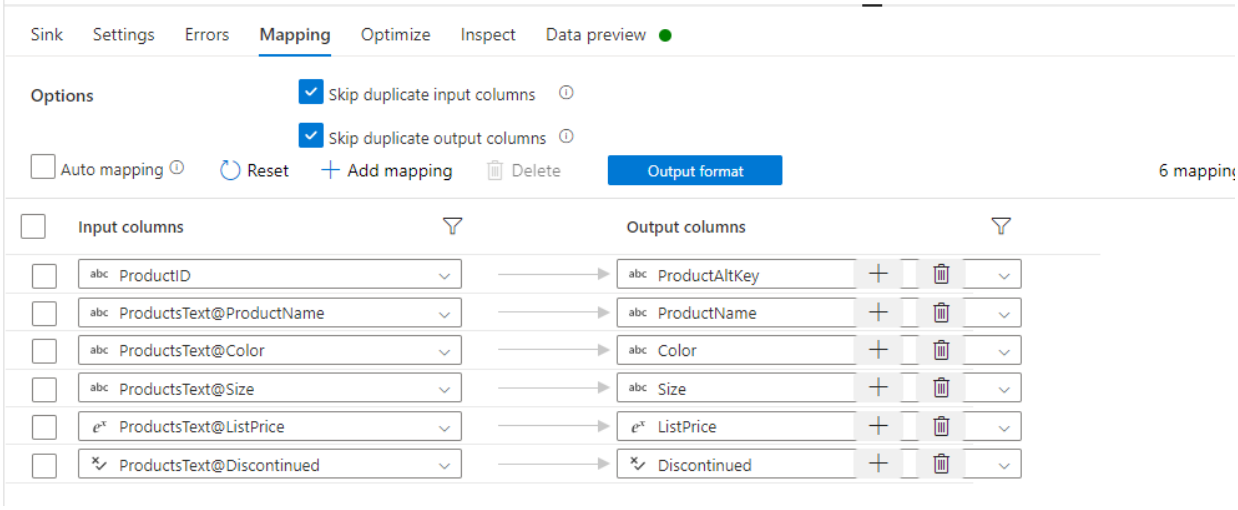
Screenshots:

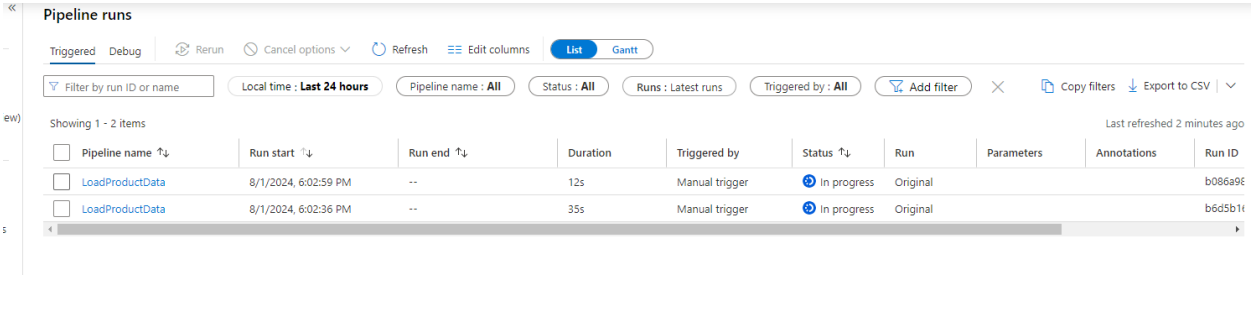
Below are some screenshots of my final result for the lab. While following the instructions, I met a problem where the option for legacy instead of recommended needed to be selected during the creation of the Data_Warehouse linked service. If not, there will be an error at the end similar to this: https://learn.microsoft.com/en-us/answers/questions/1253535/dataset-is-using-azuresqldatabase-linked-service-t

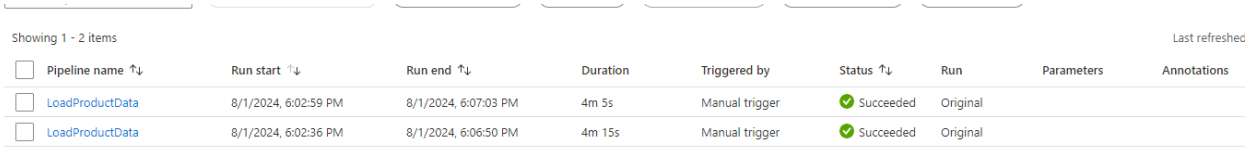
Deleting Data_Warehouse and recreating it using the legacy setting fixed my problem.

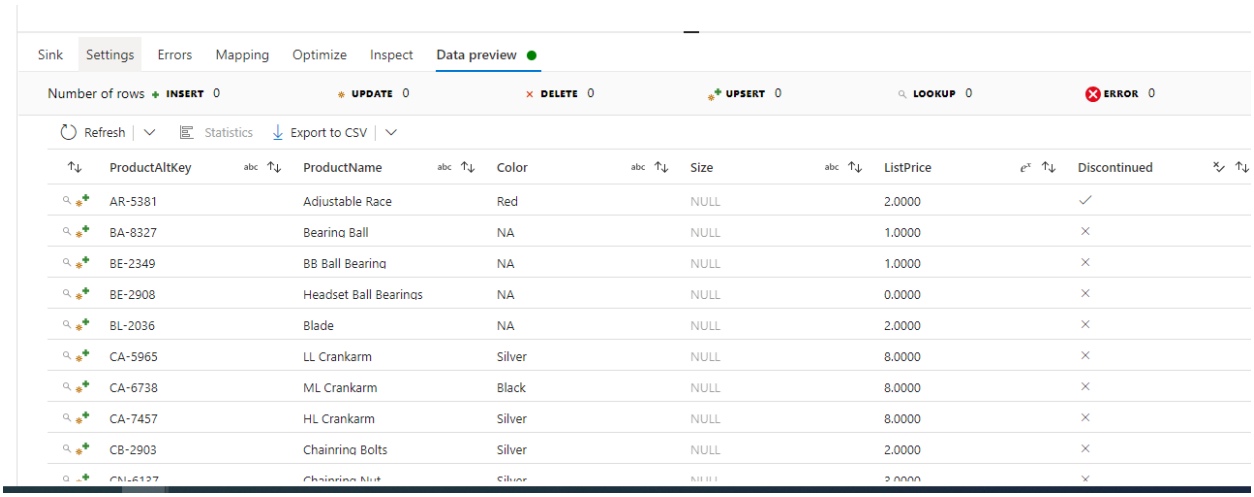



Validate
Data flow debug
Debug Settings
Reference:
1
Columns:
6 total
MatchedProducts
Matched product data
SetLoadAction
Insert new, upsert existing
DimProductTable
Columns:
6 total
ProductTable
Product table
Add Source


Sink
Settings
Errors
Mapping
Optimize
Inspect
Data preview
Options
Skip duplicate input columns
Skip duplicate output columns
Auto mapping
Reset
Add mapping
Delete
Output format
6 mapping
Input columns
Output columns
ProductID
ProductAltKey
ProductsText@ProductName
ProductName
ProductsText@Color
Color
ProductsText@Size
Size
ProductsText@ListPrice
ListPrice
ProductsText@Discontinued
Discontinued

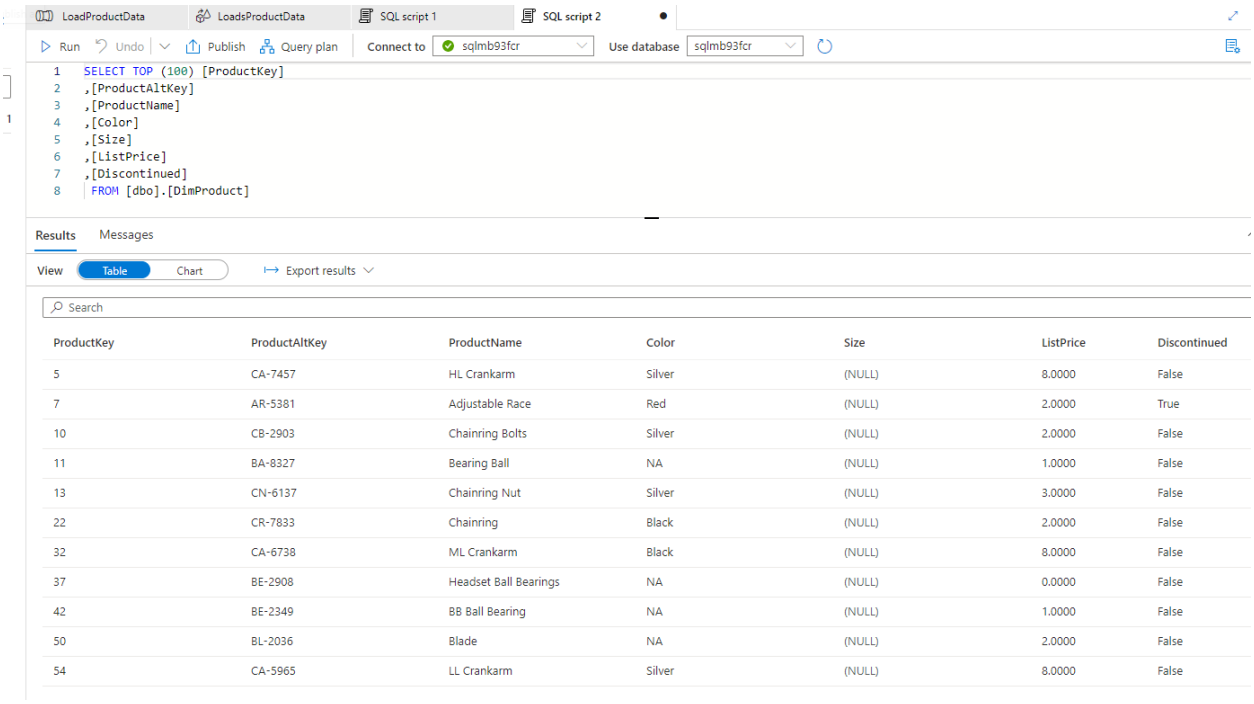

Pipeline runs
Triggered
Debug
Rerun
Cancel options
Refresh
Edit columns
List
Gantt
Filter by run ID or name
Local time : Last 24 hours
Pipeline name : All
Status : All
Runs : Latest runs
Triggered by : All
Add filter
Copy filters
Export to CSV
Showing 1 - 2 items
Last refreshed 2 minutes ago
Pipeline name
Run start
Run end
Duration
Triggered by
Status
Run
Parameters
Annotations
Run ID
LoadProductData
8/1/2024, 6:02:59 PM
--
12s
Manual trigger
In progress
Original
LoadProductData
8/1/2024, 6:02:36 PM
--
35s
Manual trigger
In progress
Original


Showing 1 - 2 items
Last refreshed
Pipeline name
Run start
Run end
Duration
Triggered by
Status
Run
Parameters
Annotations
LoadProductData
8/1/2024, 6:02:59 PM
8/1/2024, 6:07:03 PM
4m 5s
Manual trigger
Succeeded
Original
LoadProductData
8/1/2024, 6:02:36 PM
8/1/2024, 6:06:50 PM
4m 15s
Manual trigger
Succeeded
Original



ProductAltKey	ProductName	Color	Size	ListPrice	Discontinued
AR-5381	Adjustable Race	Red	NULL	2.0000	✓
BA-8327	Bearing Ball	NA	NULL	1.0000	×
BE-2349	BB Ball Bearing	NA	NULL	1.0000	×
BE-2908	Headset Ball Bearings	NA	NULL	0.0000	×
BL-2036	Blade	NA	NULL	2.0000	×
CA-5965	LL Crankarm	Silver	NULL	8.0000	×
CA-6738	ML Crankarm	Black	NULL	8.0000	×
CA-7457	HL Crankarm	Silver	NULL	8.0000	×
CB-2903	Chainring Bolts	Silver	NULL	2.0000	×

In a previous run, AR5381 had been upserted.



ProductKey	ProductAltKey	ProductName	Color	Size	ListPrice	Discontinued
5	CA-7457	HL Crankarm	Silver	(NULL)	8.0000	False
7	AR-5381	Adjustable Race	Red	(NULL)	2.0000	True
10	CB-2903	Chainring Bolts	Silver	(NULL)	2.0000	False
11	BA-8327	Bearing Ball	NA	(NULL)	1.0000	False
13	CN-6137	Chainring Nut	Silver	(NULL)	3.0000	False
22	CR-7833	Chainring	Black	(NULL)	2.0000	False
32	CA-6738	ML Crankarm	Black	(NULL)	8.0000	False
37	BE-2908	Headset Ball Bearings	NA	(NULL)	0.0000	False
42	BE-2349	BB Ball Bearing	NA	(NULL)	1.0000	False
50	BL-2036	Blade	NA	(NULL)	2.0000	False
54	CA-5965	LL Crankarm	Silver	(NULL)	8.0000	False

Final top 100 query after the pipeline finished running