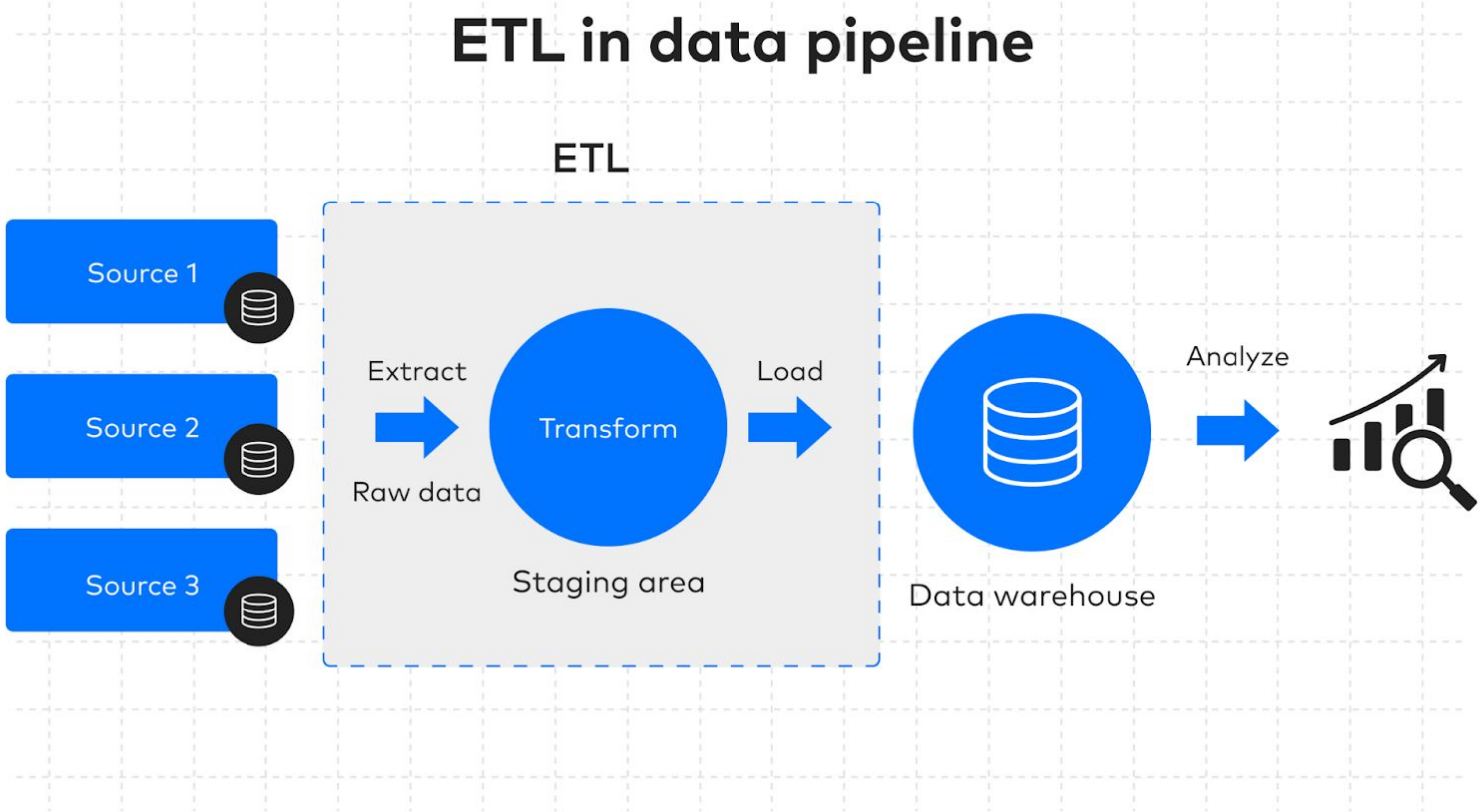
Visual ETL with Mapping Data Flows in Azure Data Factory

Antony Prince J

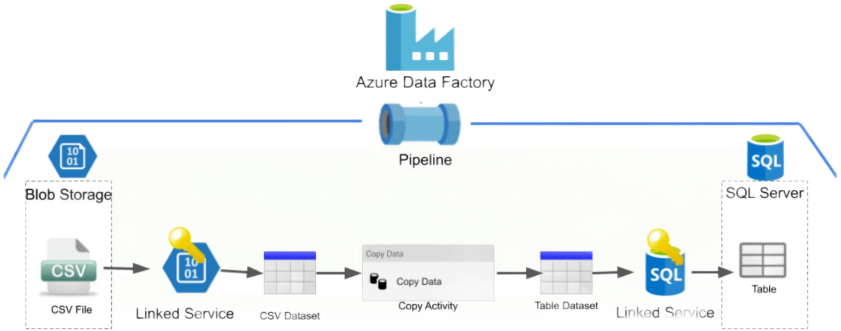


Reference - https://www.fivetran.com/learn/data-pipeline-vs-etl

Azure Data Factory

Azure Data Factory (ADF) is a cloud-based ETL and data integration service data integration service provided by Microsoft as part of its Azure cloud platform.

Create, schedule, and orchestrate data workflows and pipelines for moving, transforming, and integrating data from various sources to desired destinations.



Reference
https://www.productiveedge.com/blog/azure-data-factory-capabilities

CodeFree ETL as a service

Ingest

- Multi-cloud and on premise hybrid copy data
- 100+ native connectors
- Serverless and auto scale
- Use wizard for quick copy jobs

Control Flow

- Design codefree data pipelines
- Generate pipelines via SDK
- Utilize workflow constructs: loops, branches, conditional execution, variables, parameters, …

Data Flow

- Codefree data transformations that execute in Spark
- Scaleout with Azure Integration Runtimes
- Generate data flows via SDK
- Designers for data engineers and data analysts

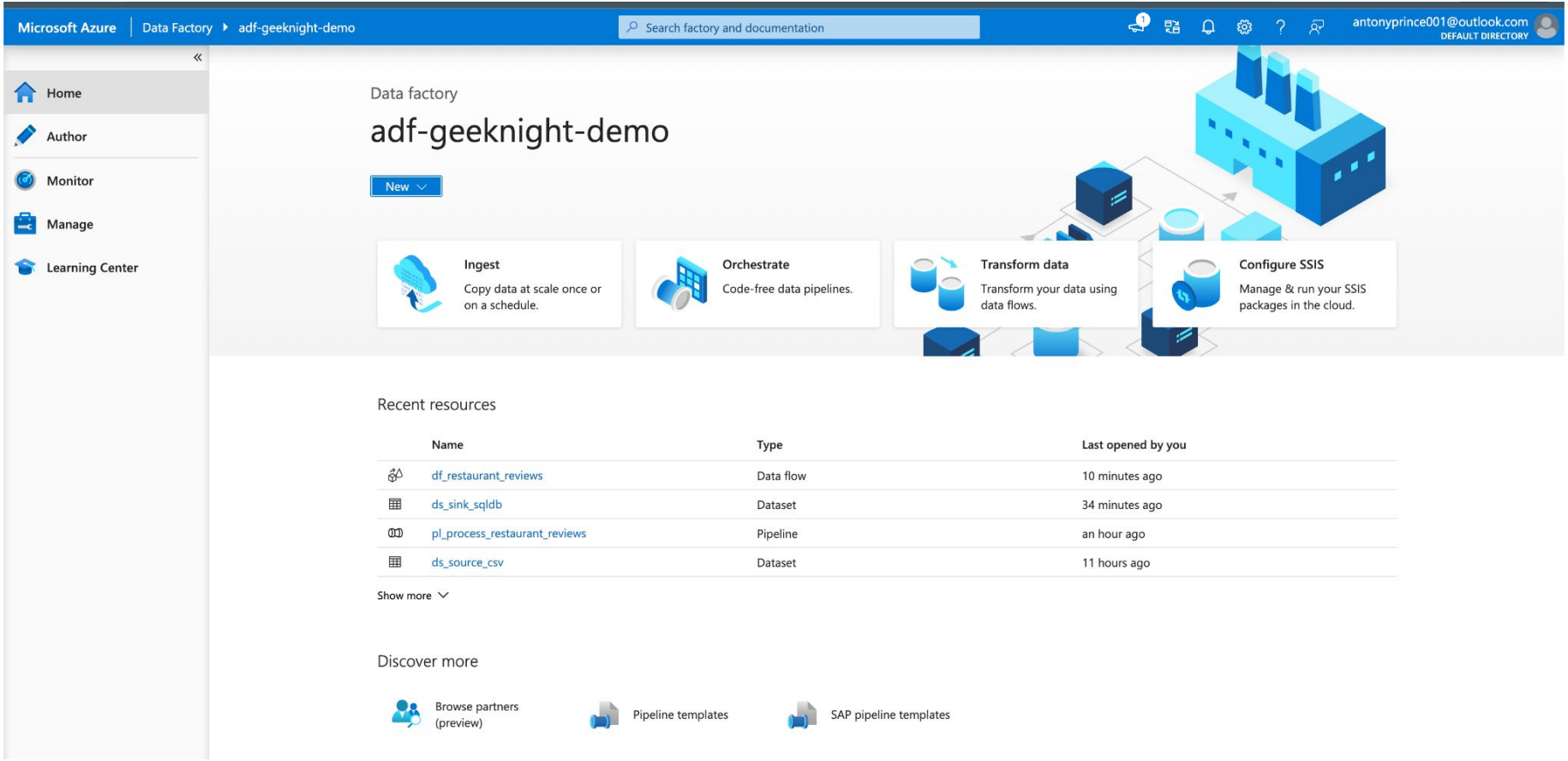
Schedule

- Build and maintain operational schedules for your data pipelines
- Wall clock, eventbased, tumbling windows, chained

Monitor

- View active executions and pipeline history
- Detail activity and data flow executions
- Establish alerts and notifications

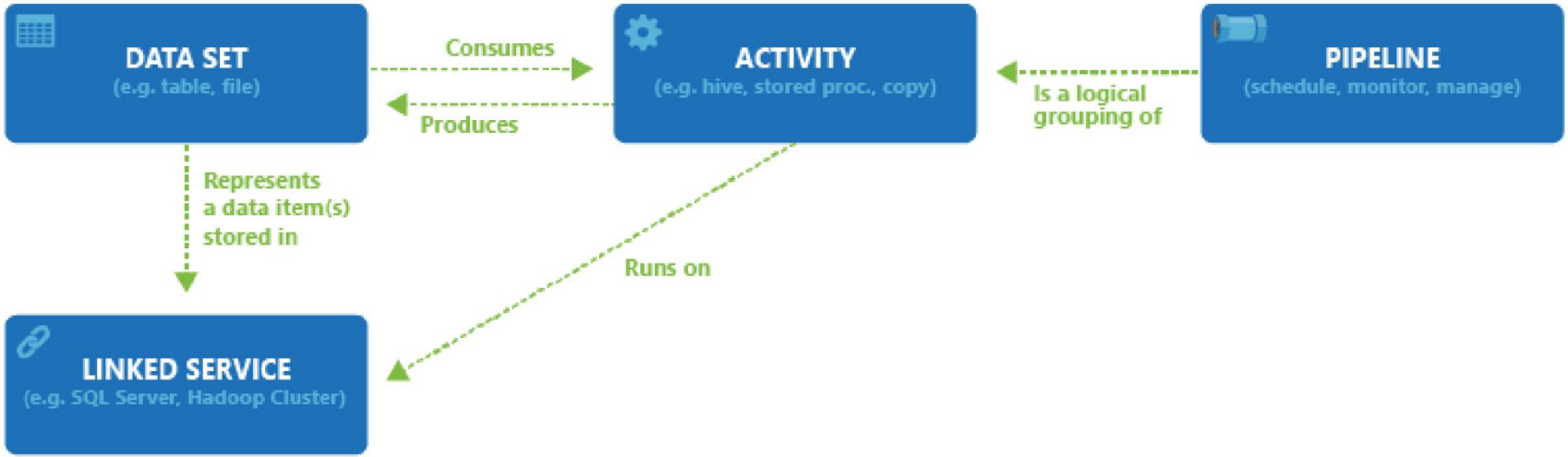
Reference - https://learn.microsoft.com/en-us/azure/data-factory/introduction


Microsoft Azure
Data Factory
adf-geeknight-demo
Search factory and documentation
antonyprince001@outlook.com
DEFAULT DIRECTORY
Home
Author
Monitor
Manage
Learning Center
Data factory
adf-geeknight-demo
New
Ingest
Copy data at scale once or on a schedule.
Orchestrate
Code-free data pipelines.
Transform data
Transform your data using data flows.
Configure SSIS
Manage & run your SSIS packages in the cloud.
Recent resources
Name
Type
Last opened by you
df_restaurant_reviews
Data flow
10 minutes ago
ds_sink_sqldb
Dataset
34 minutes ago
pl_process_restaurant_reviews
Pipeline
an hour ago
ds_source_csv
Dataset
11 hours ago
Show more
Discover more
Browse partners (preview)
Pipeline templates
SAP pipeline templates

Use cases

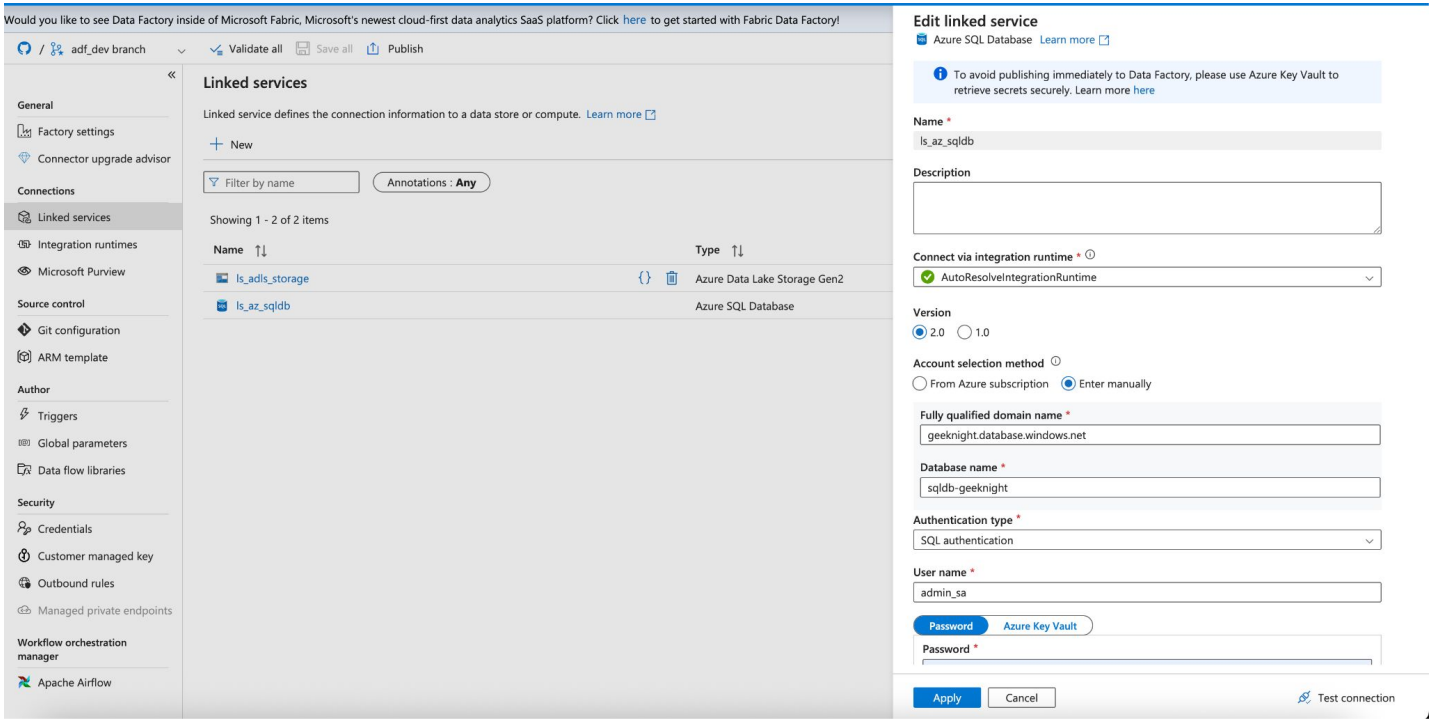
Banking & Finance	Consolidate transaction, credit score, and loan data for risk assessment and regulatory compliance.
Retail & Ecommerce	Merge online and in-store sales data to analyze customer buying patterns and optimize inventory.
Healthcare	Aggregate insurance claims, and clinical trial data to support research, and comply with health data regulations.
Manufacturing	Combine production line sensor data, supply chain logistics, and inventory records to optimize operations and predict equipment failures.

Azure Data Factory Components



Linked Services

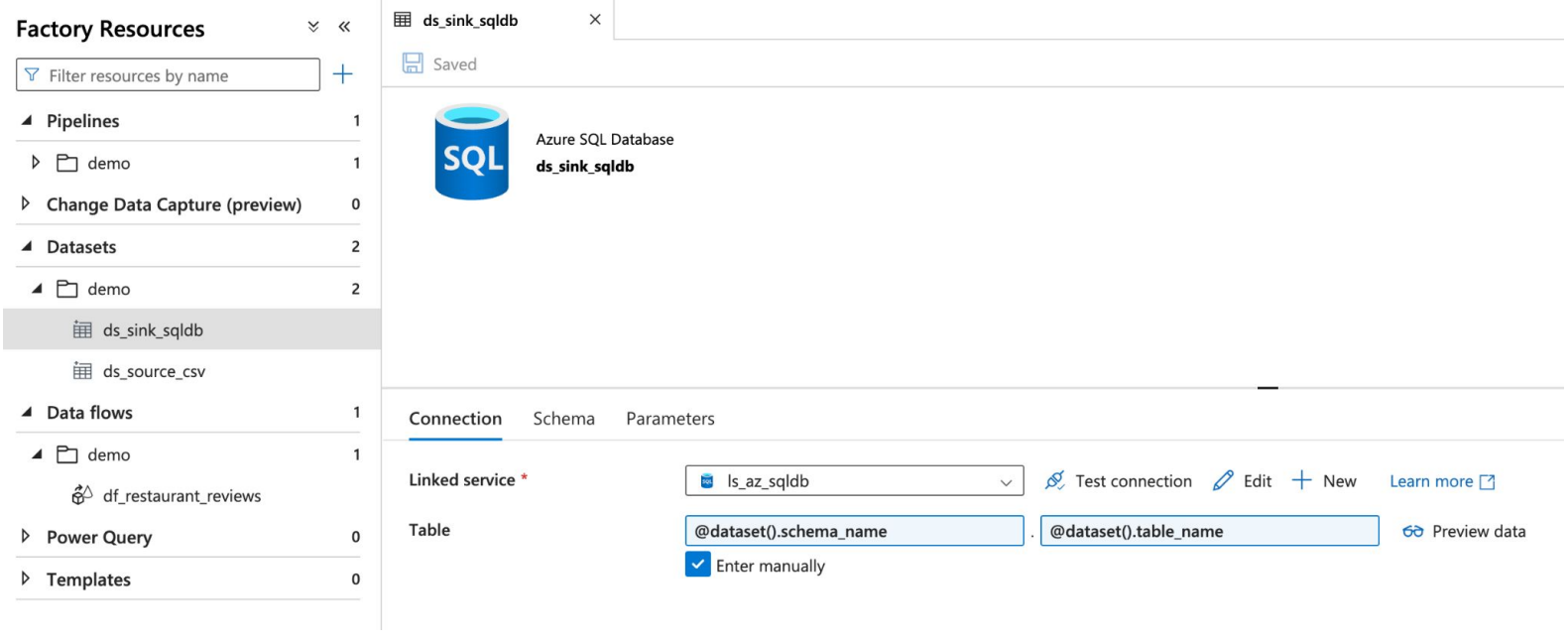
Linked services are connections to external data stores or compute resources. They define the connection information needed to connect to external sources.



Dataset

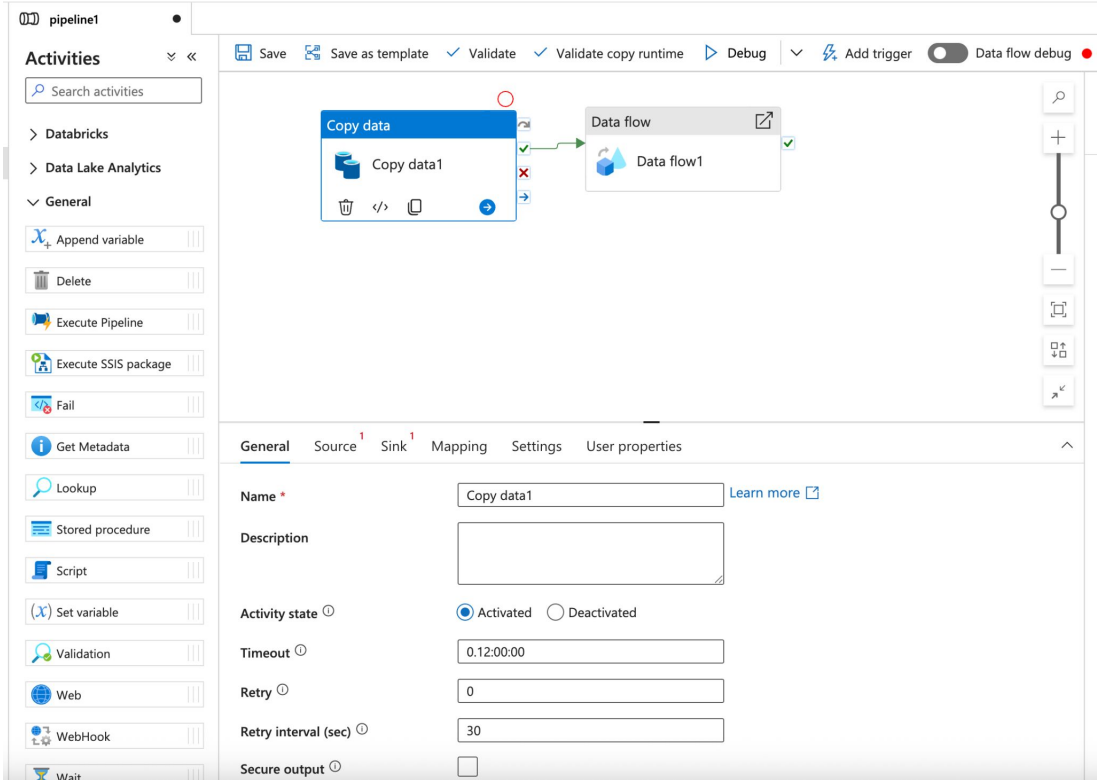
Dataset is a representation of the data you intend to work with in your data pipelines.

- Acts as a bridge between the source or destination data store and the ADF pipeline activities.
- Doesn't contain the data itself but references the data stored in various locations (e.g., databases, file systems, cloud storage).



Activity

Activities are the individual tasks that get executed within a pipeline.



Pipeline

A pipeline is a logical grouping of activities that together perform a task. A pipeline allows you to manage and schedule workflows of data movement and transformation

Trigger

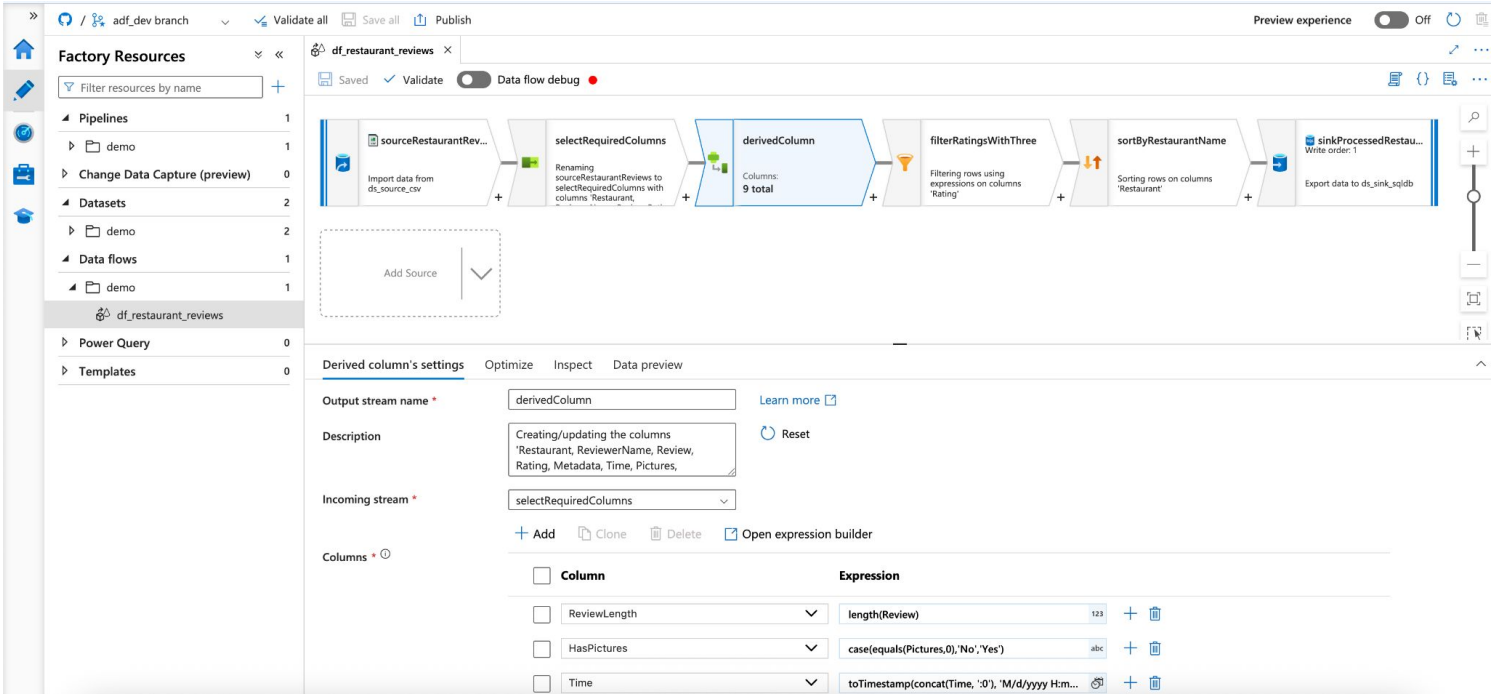
Triggers allow you to automatically initiate the execution of a pipeline.

Integration Runtime

The Integration Runtime is the compute infrastructure used by ADF to perform data movement, data transformation, and other activities.

Mapping Data Flows in Azure Data Factory

Mapping Data Flows in Azure Data Factory (ADF) allows you to transform and move data in a no-code environment. Data flows support a variety of transformation activities to achieve tasks like aggregations, filtering, joins, and more. It can support - Data cleaning, aggregation, transformation and enrichment.

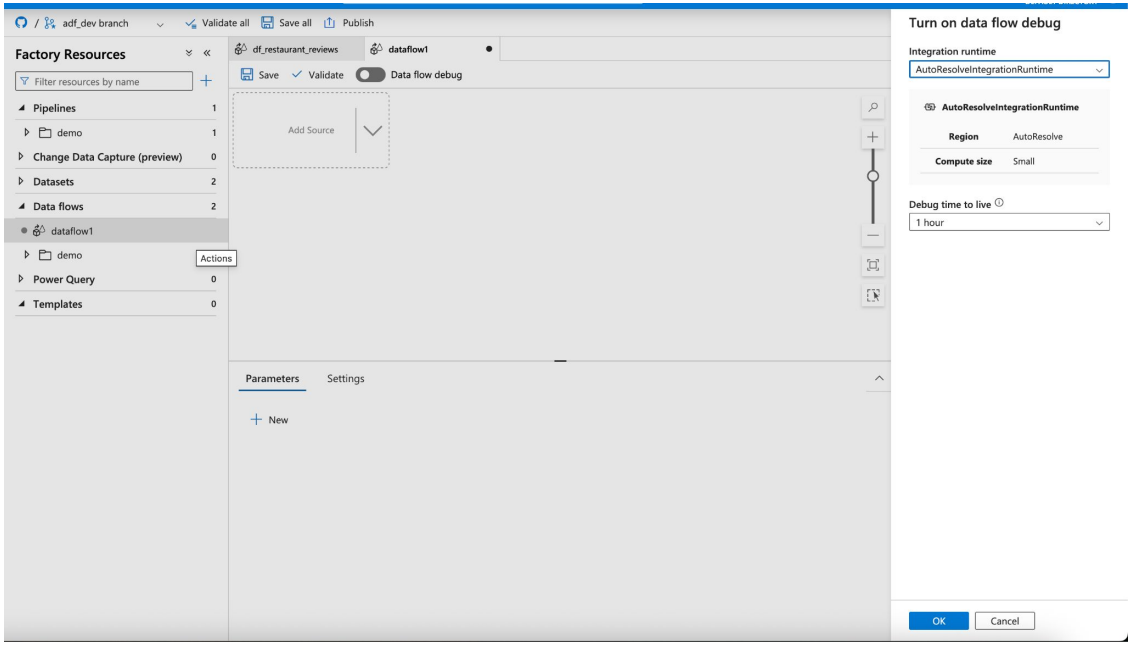


1. Preparing the environment

Turn the Data Flow Debug slider located at the top of the authoring module on

Add a Data Flow activity.

Create the Required Linked Services and Datasets for source and destination sink

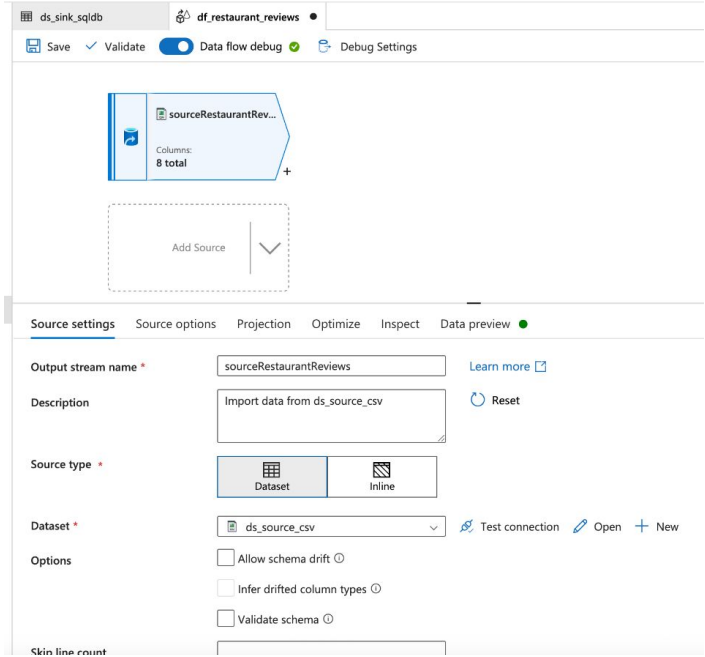


2. Add source data to the Mapping Data Flow

When you design data flows, your first step is always configuring a source.

Every data flow requires at least one source transformation, but you can add as many sources as necessary to complete your data transformations.

The sources can be combined using a join, lookup, or a union transformation.



3. Transformations in the mapping data flow

Mapping Data Flows provides a number of different transformations types that enable you to modify data.

- Schema modifier transformations

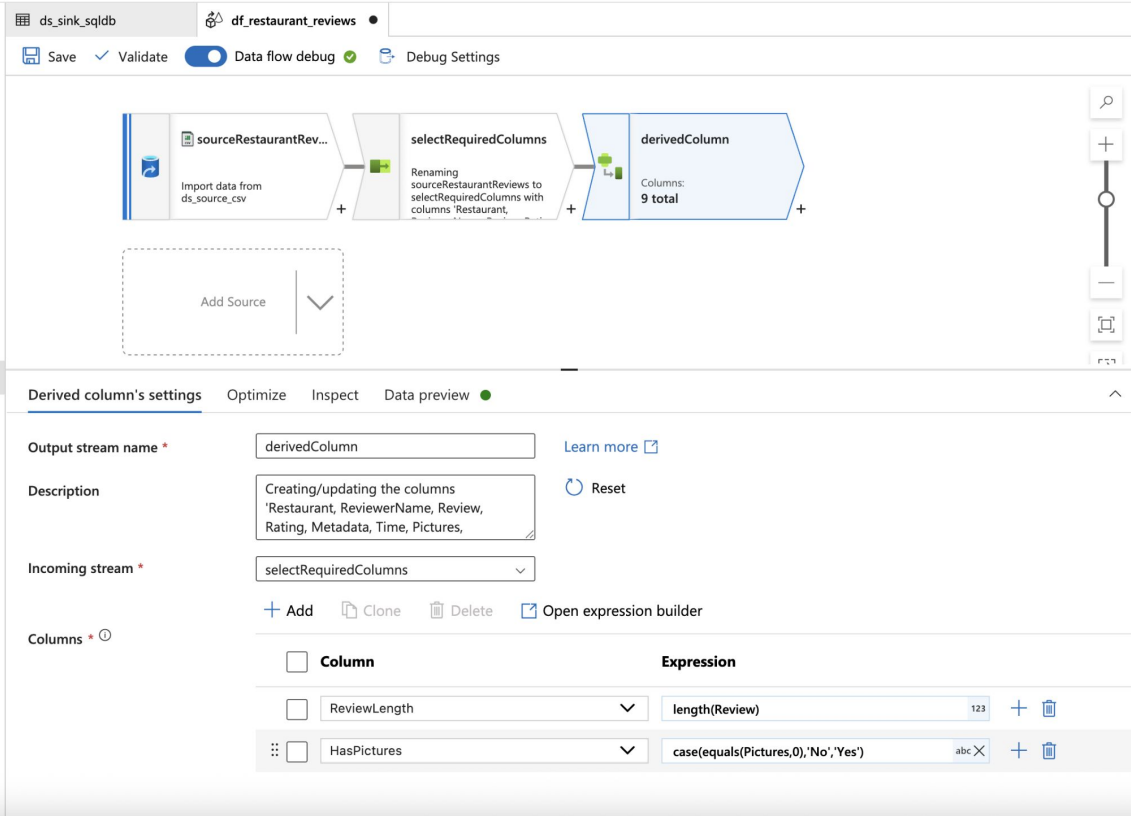
  Ex: Derived Column

- Row modifier transformations

  Ex : Sort

- Multiple inputs/outputs transformations

  Ex : Join



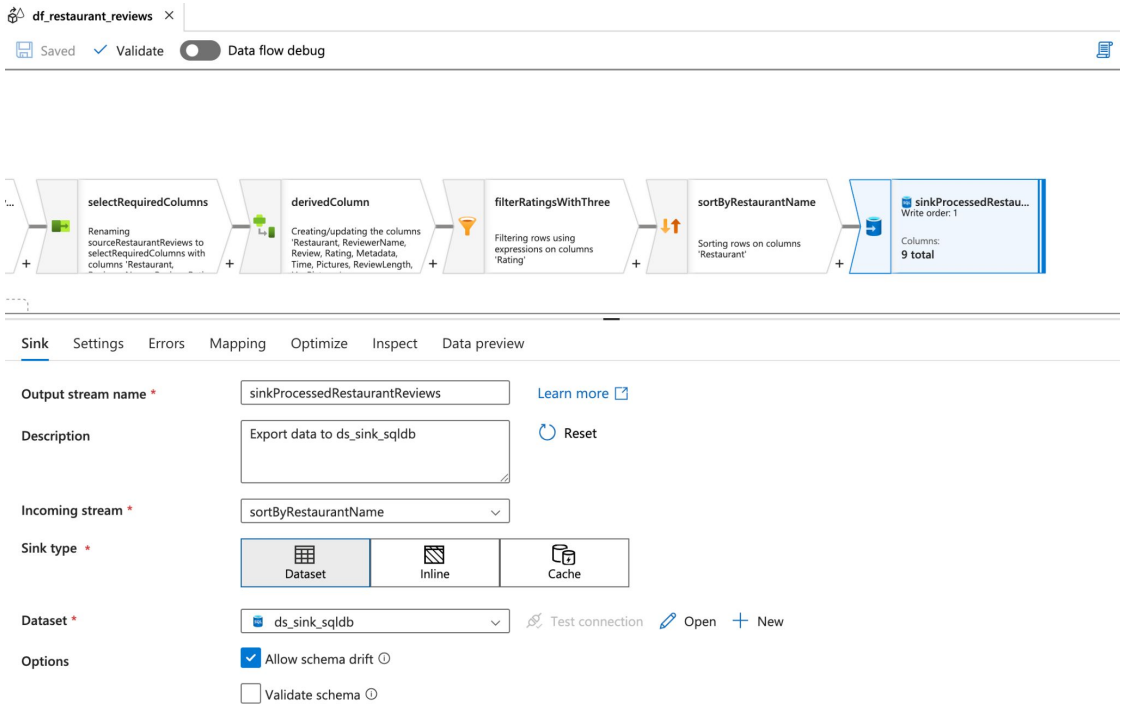
https://learn.microsoft.com/en-us/training/modules/code-free-transformation-scale/3-describe-transformation-types

4. Writing to Data Sink

The Sink transformation is the endpoint in a data flow where transformed data is written to a target destination.

It supports various sinks like Azure Blob Storage, Azure SQL Database, Data Lake, Cosmos DB, etc.

This is where data is stored after all processing is completed.



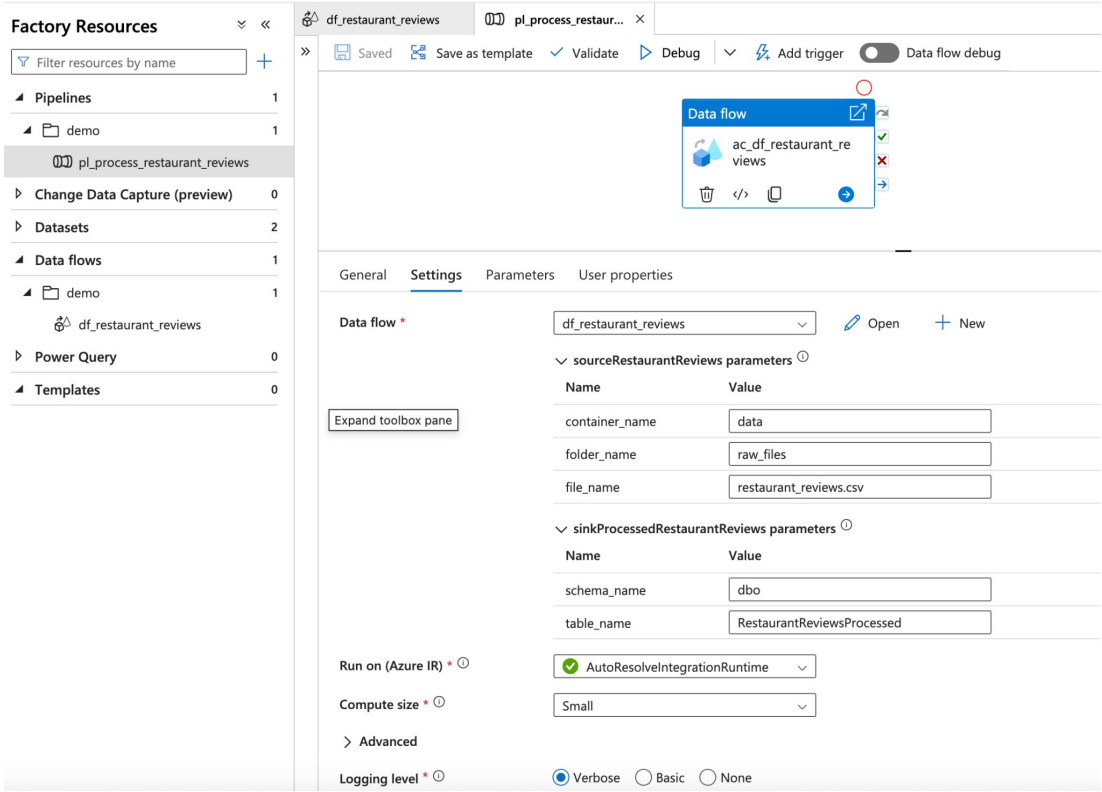
5. Add Dataflow activity to pipeline and trigger it

Create a new pipeline.

Drag and drop the Data Flow activity onto the pipeline canvas.

Configure the Data Flow Activity with required parameters.

Trigger the Pipeline and monitor the data flow run to see if the data is written to sink after performing the expected transformations.



Demo

You are working for a company that runs a large online restaurant review platform

You are to extract data from the source which has thousands of new reviews from different customers.

Before moving the data to the required destination, you are also expected to perform some transformations

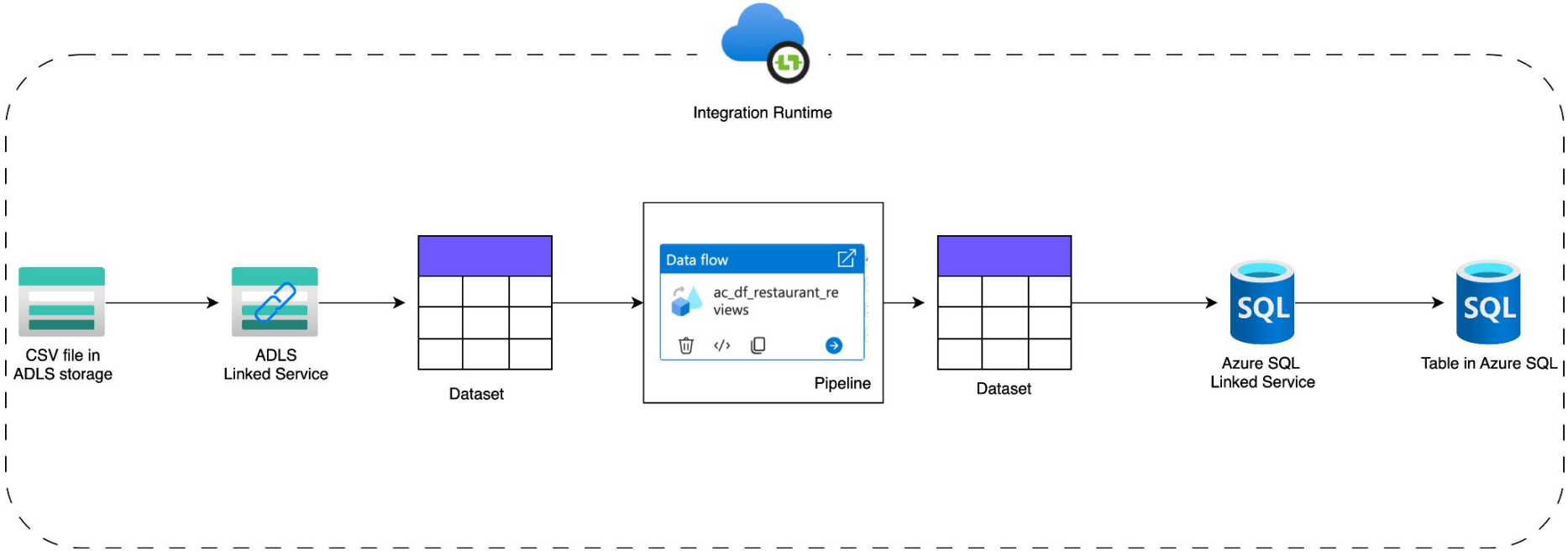
- Drop irrelevant columns
- Create new features like Review Length and HasPictures.
- Select rows with rating not equal to 3

Create a Mapping Data Flow to clean the data, add meaningful columns, summarize key insights, and prepare it for analytics."

DATASET

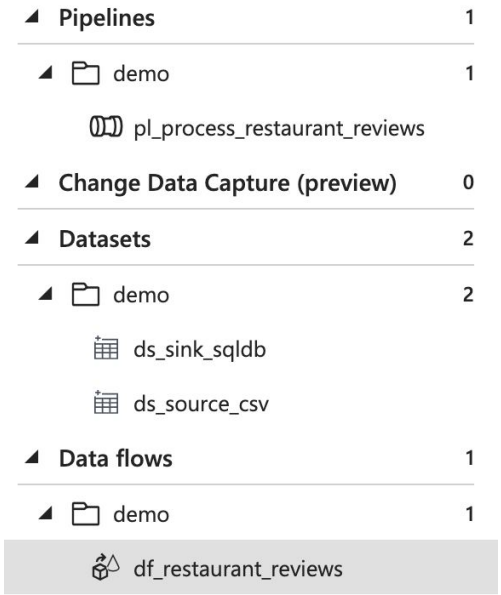
Restaurant	Reviewer	Review	Rating	Metadata	Time	Pictures	DummyColumn
Beyond Flavours	A	The ambience was good, food was quite good . had Saturday lunch , which was cost effective . Good place for a sate brunch. One can also chill with friends and or parents. Waiter Soumen Das was really courteous and helpful.	5	1 Review , 2 Followers	5/25/2019 15:54	0	2447
Beyond Flavours	B	Ambience is too good for a pleasant evening. Service is very prompt. Food is good. Over all a good e	5	3 Reviews , 2 Followers	5/25/2019 14:20	0	
Beyond Flavours	C	A must try.. great food great ambience. Thnx for the service by Pradeep and Subroto. My personal re	5	2 Reviews , 3 Followers	5/24/2019 22:54	0	

https://www.kaggle.com/datasets/joebeachcapital/restaurant-reviews/data

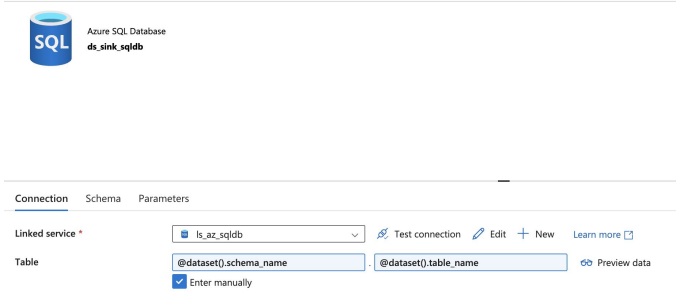

Integration Runtime
CSV file in
ADLS storage
ADLS
Linked Service
Dataset
Data flow
ac_df_restaurant_re
views
Pipeline
Dataset
SQL
Azure SQL
Linked Service
SQL
Table in Azure SQL

Data Flow

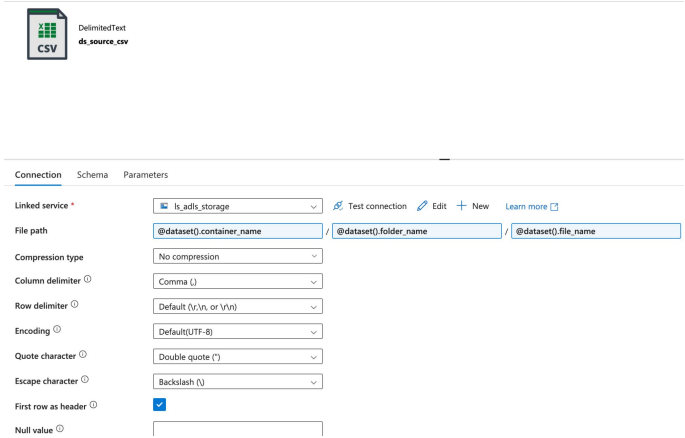
ADF Structure



Azure SQL table dataset



Azure Storage dataset



Linked Services

ls_adls_storage	Azure Data Lake Storage Gen2	1
ls_az_sqldb	Azure SQL Database	1

Azure Data Lake Storage Gen2 Learn more

Name *
ls_adls_storage

Description

Connect via integration runtime * ⓘ
AutoResolveIntegrationRuntime

Authentication type
Account key

Account selection method ⓘ
From Azure subscription / Enter manually

URL *
https://dlsgeeknight.dfs.core.windows.net/

Storage account key / Azure Key Vault

Storage account key *
••••••••••

Test connection ⓘ
To linked service / To file path

Annotations
+ New

Parameters

Advanced ⓘ

Azure SQL Database Learn more

To avoid publishing immediately to Data Factory, please use Azure Key Vault to retrieve secrets securely. Learn more here

Name *
ls_az_sqldb

Description

Connect via integration runtime * ⓘ
AutoResolveIntegrationRuntime

AutoResolveIntegrationRuntime

Version
2.0 / 1.0

Account selection method ⓘ
From Azure subscription / Enter manually

Fully qualified domain name *
geeknight.database.windows.net

Database name *
sqldb-geeknight

Authentication type *
SQL authentication

User name *
admin_sa

Password / Azure Key Vault

Password *

Disadvantages

Not Ideal for Highly Custom Transformations - using custom code in Databricks or other services might be necessary

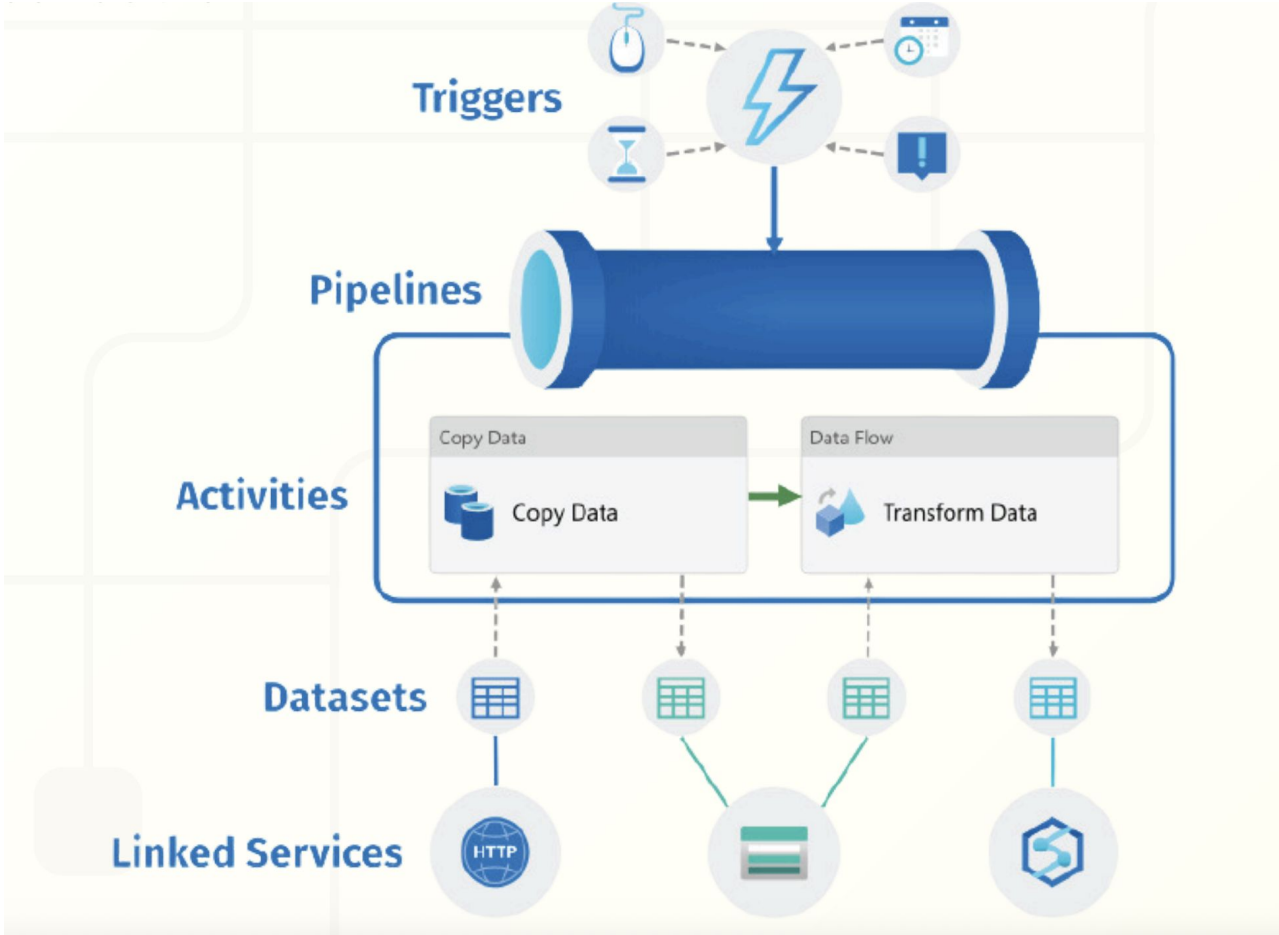
Limited Control Over Spark Environment - Spark environment used in data flows is abstracted, limiting fine tuning and custom configurations.

Higher Costs for Large Data Volumes - Can become costly when processing large volumes of data frequently and also during debugging.

Complex Debugging for Large Flows - Debugging complex data flows with multiple transformations can become challenging

Performance Overhead - Data flows rely on Spark clusters, which introduce startup time and may increase latency.

Limited Real-Time Processing - Primarily designed for batch processing, not real-time or streaming data.


Triggers
Pipelines
Activities
Copy Data
Copy Data
Data Flow
Transform Data
Datasets
Linked Services
HTTP

Thank you