A Short Story about Using Data Factory

A “little bit” about why first…

“Well begun is half done.”

Before the well begun – define the problem

- Customers do not always know their own problems well, though they feel something is wrong.

Understand the Situation

Problem: We don’t any insight.

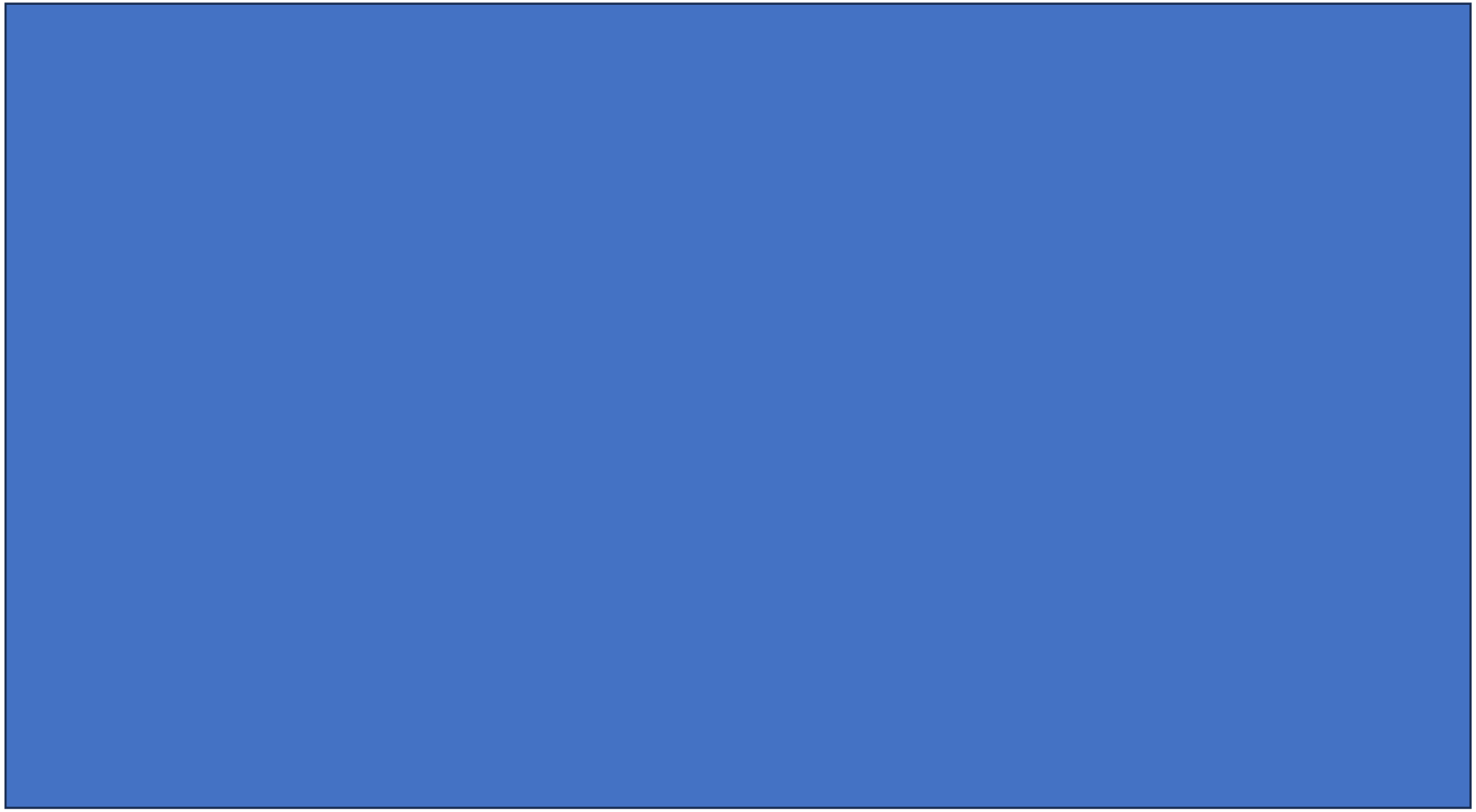
Similar as going to doctor, before treatment, an important question is:

- What is the root cause to the problem?

Business Wishes – which exactly “insights”are missing and how much they are wanted

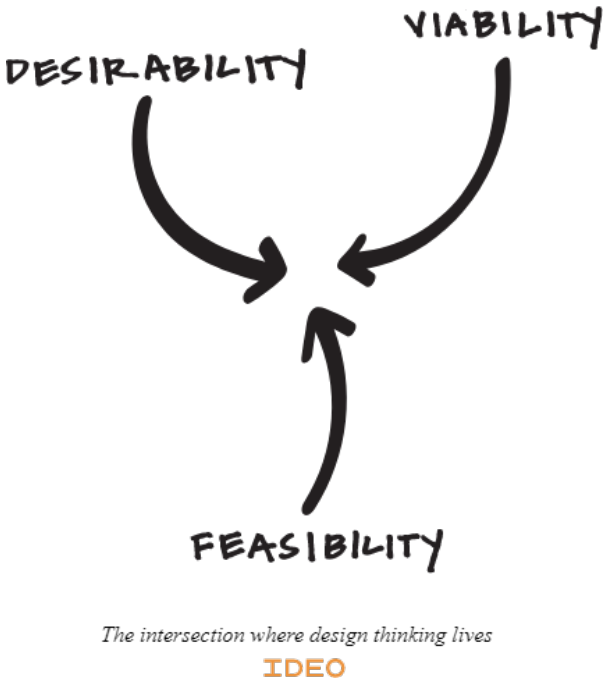
		Business value								
Wishes		Reputation Impacts (0 -5)	Commercial impacts (0 - 5)	Value for prodcut development	Necessary for continuing daily work (0-5)	Manpower Saved (0-5)	Employee satification	Overall Score	Weights	
Number of li		4	5	1	5	1	5	180	Reputation Impacts	10
What is "nor		1	3	3	2	0	2	87	Commercial Impacts	10
Usages of Te		3	3	1	3	0	5	129	Value for prodcut development	5
Usage of boo		1	3	5	2	2	4	123	Necessary for continuing daily work	8
Segementati		1	5	4	2	0	3	120	Manpower Saved	5
Churn analys		3	5	5	5	3	5	200	Employee satification	8
CRM automa		2	5	0	5	5	5	175		
CRM automa		1	3	4	0	0	4	92		
Generates a		1	4	3	4	4	4	149		
Calculate the		2	4	4	3	0	3	128		
Automate ca		0	4	0	4	5	4	129		
Clustering documents based on links and usage separately, and compare		2	2	5	1	0	3	97		

Where the Data is – Tech. Feasibility



- Figured out how data flow between systems
- Read evey column in the backend database

Project Selection



Business viability

Make (long-term) plan

as long as budget and resources fit it…

Technical feasibility

Make multiple selection matrics as needed

- I was doing similar thing here

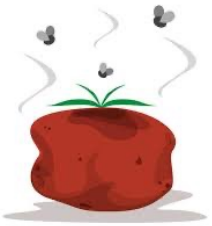
Now come back to the topic: the root cause was identified during the exercises too.

We don’t any insight because:

- ~~We don’t have any data.~~

- Data are in different systems - but no one collected, connected, and **stored** them

- There is no fridge or freezer, materials (data) are rotten and casted away



Common Practices in Industries

Systems living in present

serving **daily business**

Data Warehouse

Systems collect and keeps history from systems living in present

serving **analytics**

“It is almost impossible to have a sense of vision without a sense of history. “ – Kofi Annan

Didn’t exist (but now we have both)

What is a Data Warehouse

In computing, a **data warehouse** (**DW** or **DWH**), also known as an **enterprise data warehouse** (**EDW**), is a system used for reporting and data analysis and is considered a core component of business intelligence.[1] Data warehouses are central repositories of integrated data from one or more disparate sources. They store current and historical data in one single place[2] that are used for creating analytical reports for workers throughout the enterprise.[3] This is beneficial for companies as it enables them to interrogate and draw insights from their data and make decisions.[4]

Sure, imagine a data warehouse as a giant, organized library for information that a business uses to make important decisions. Think of it like a magic storage space where all the data from different parts of the company gets neatly organized and kept in one place.

In a regular library, books are arranged by categories, like fiction, history, science, and so on. In a data warehouse, information from different areas of the business, like sales, customer info, inventory, and more, is stored in a way that makes it easy to find and use.

It acctually sounds more like core layer of DWH

Current Data Warehouse Structure

- Data marts

- denomarlized, ready for being consumed

- Core Layer

- organized library as mentioned by ChatGPT
- normalized, single source reference

- Staging Layer

- tables imported from different systems, as they are
- extracted information from one single event/steaming source (e.g. event hub)

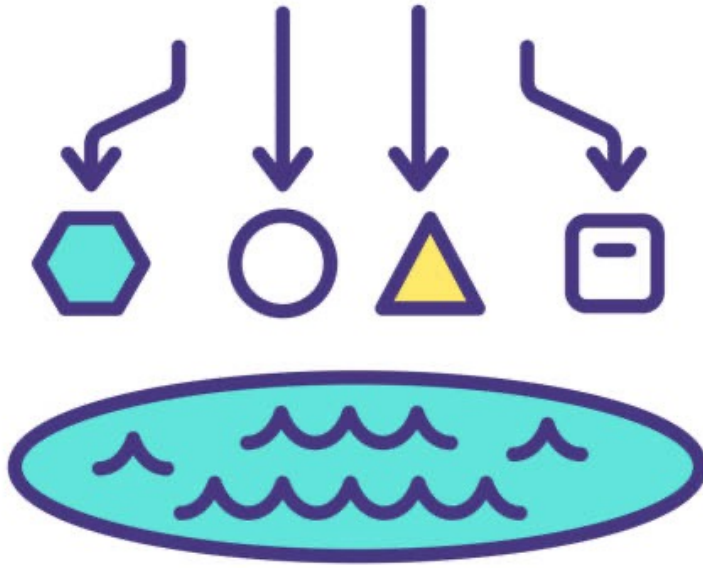
What is Data Lake

Data Lake:

Imagine a data lake as a vast, unstructured reservoir where the business stores all kinds of data, from raw files to structured information, without much organization. It's like a big lake where data from various sources, like customer logs, website clicks, social media mentions, and more, are all dumped in one place. This makes it easy to throw data into the lake without worrying about organizing it first.

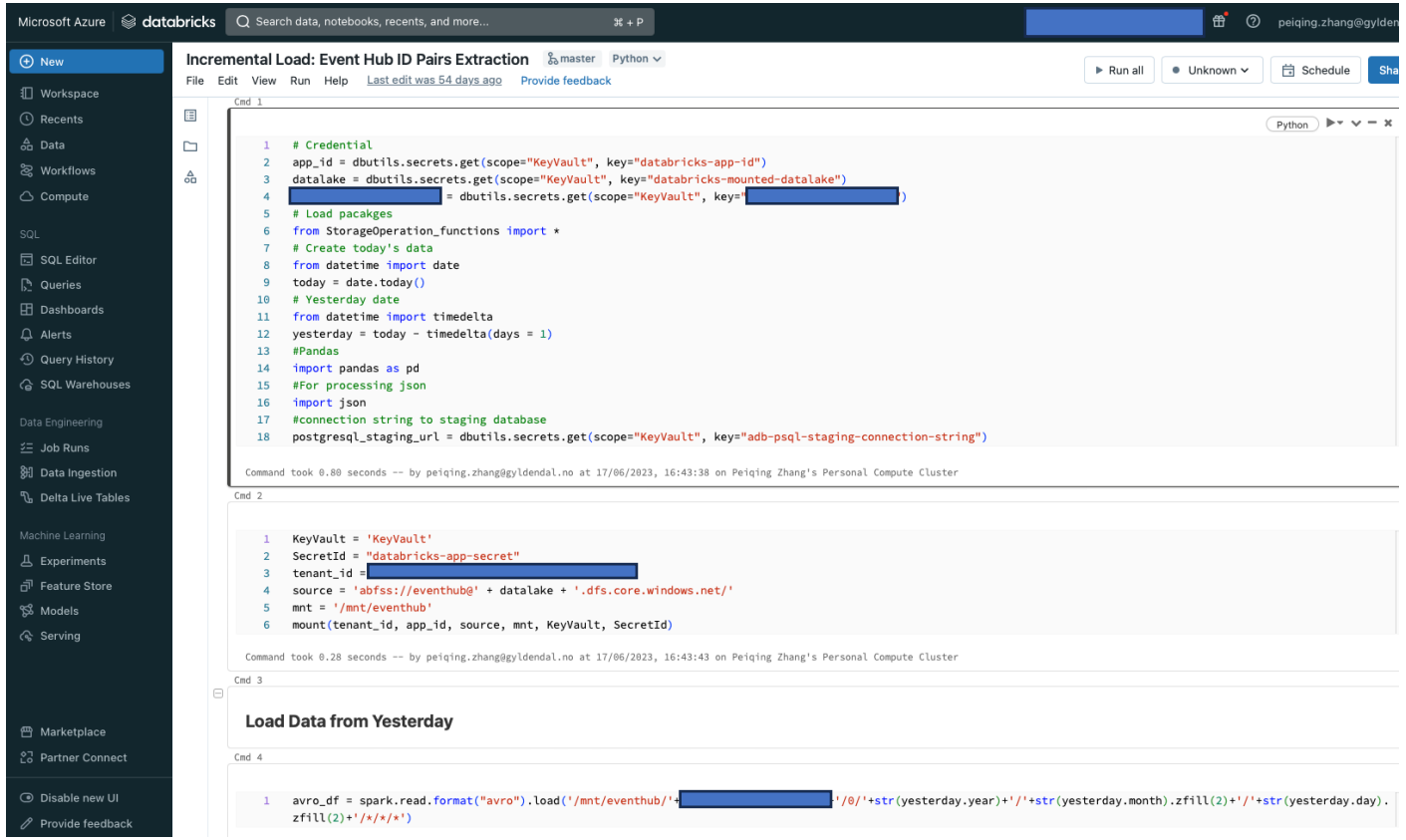
4. **Use Cases:**
   - Data Warehouse: Best for structured data analysis, reporting, and business intelligence, where the data is already well-understood.
   - Data Lake: Great for storing large volumes of raw data that might have future use, especially for big data analytics, machine learning, and exploratory analysis.

In essence, a data warehouse is like a carefully curated library, whereas a data lake is more like a catch-all storage lake for all kinds of data, which can be useful when you're dealing with large, diverse, and unstructured datasets, but it requires more processing and structuring when you want to extract meaningful insights.



Databricks

- Azure Databricks is used to process data in Data Lake and transfer them into DHW
- All needed are:
  - Ensure interpretation from data to real world events is correct
  - Resource



Eliminate the Root Cause

to build a data warehouse and a data lake

and logistics, and other necessary supporting facilities

i.e. a platform to host them, and pipelines to transfer data

Deliver some work to build confidence and trust (directly profits will be a plus).

Status

- Kitchen is built up, fridge and freezer is there, part of the logistics is built
- Continue with building logistics
- Continue with organizing storage
- By the way making salads (PowerBI Dashboards)

Principles

Data Insight Platform

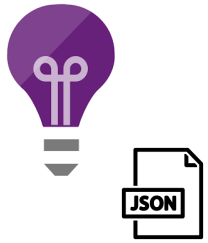
PZ

Owned by Peiqing Zhang •••

Last updated: Mar 02, 2023 • 2 min read • 3 people viewed

Principles:

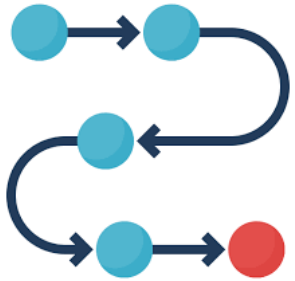
1. Cloud based
2. Flexible - possible to replace building blocks with new ones later
3. Low cost based on not sacrificing security and robustness
4. Serverless as possible
5. Implement IaC (Infrastructure as Code)



EventHub

SQL server on Premise

Data Sources



- **Connect**
- **Process and Transfer**
- **Orchestration and Automation**

Data Warehouse

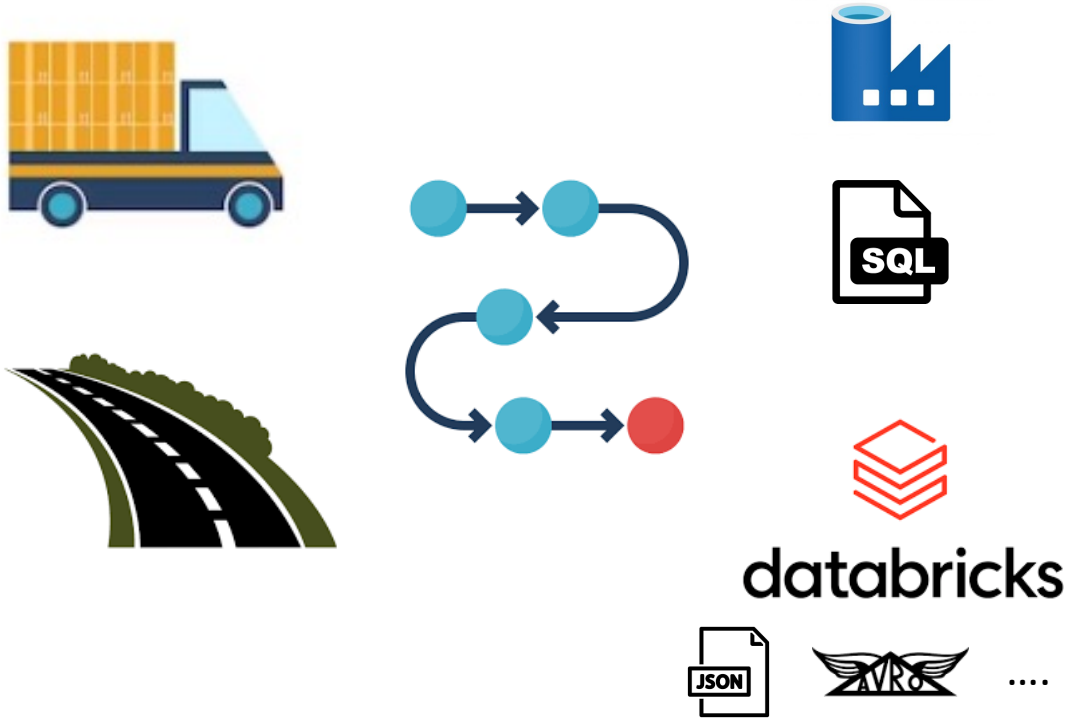
Postgres SQL – the cheapst, Prod upgrade to 4 cores when summer student was here

Storage account Datalake Gen2.

Data Storage

Needs Identified

- Able to connect to different data sources, especially ERP, CRM, on-premise.
- Tool to process and transfer unstructured/**semi-structured** data to unstructured/semi-structured/**structured data** (bold is what is being executed now)
- Tool to process and transfer structured data to structured data.
- Tool for automation and orchestration

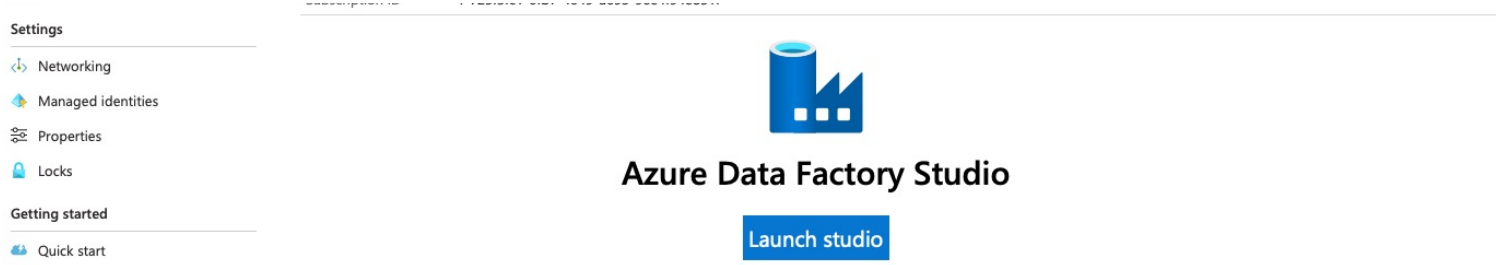


The second point is covered by Databricks (Python notebooks in Databricks, called by Data Factory), rest are covered by Data Factory.

Reasoning the Choice on Data Factory

- Allows scheduled trigger (only a daily load is needed for now)
- Provides connections to ERP, CRM, on-premises.
- Able to orchestrate other Azure resources which can be used for handling data, e.g. Databricks, Azure Functions etc.
- Pay as you go

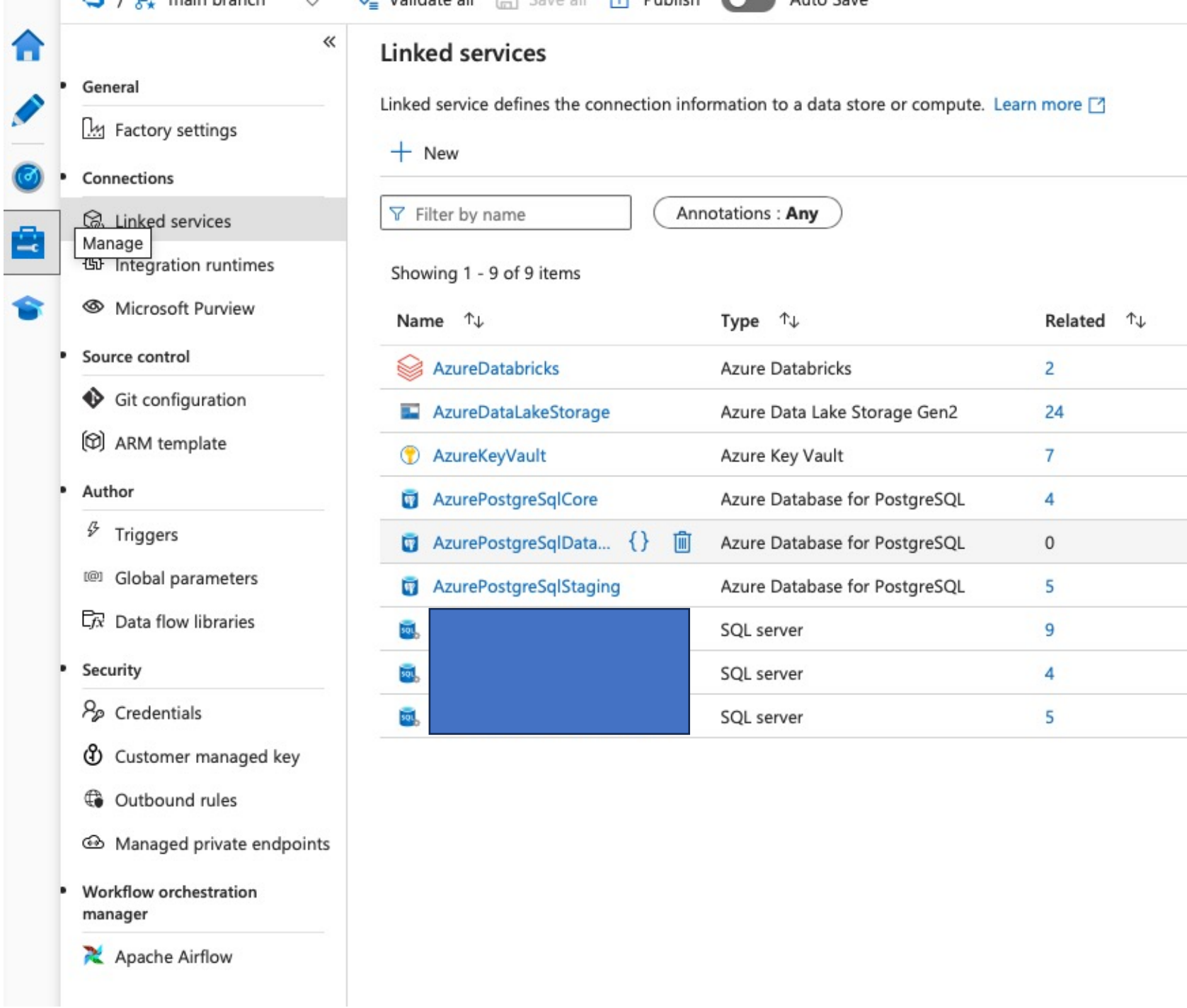
Take a Tour in Azure Data Factory



CI/CD pipeline is in use already:

Linked services: like roads

They are created by Bicep template (IaC)



Integeration Runtimes: like trucks

Integration runtimes

The integration runtime (IR) is the compute infrastructure to provide the following data integration capabilities across different network environment. Learn more

New Refresh

Filter by name

Showing 1 - 3 of 3 items

Name	Type	Sub-type	Status	Related	Region	Version
AutoResolveIntegrationRun...	Azure	Public	Running	0	Auto Resolve	---
integrationRuntime	Azure	Managed Virtual Network	Running	4	West Europe	---
[redacted]	Self-Hosted	Linked	Running (Limited)	3	---	---

- Auto Resolve Integration Runtime: trucks which only allowed to be run in public area
- Integration Runtime Connected to Managed Virtual Network: trucks getting access to private territory via a private road and portal (managed private endpoint)
- Self-hosted Integration Runtime: trucks provided by the food supplier (e.g. on-premise SQL-sever)

You don’t own the truck teams 24/7, but you call them as taxis (pay as you go)

Managed private endpoints

Managed private endpoint uses a private IP address from within Managed Virtual Network to connect to an Azure resource or your own private link service. Connections using managed private endpoints listed below provide access to Azure resources or private link services. Learn more

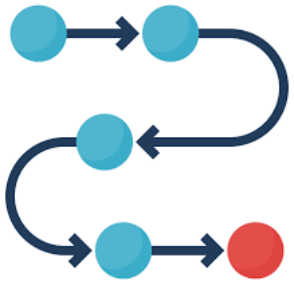
New Refresh

Search to filter items...

Showing 1 - 1 of 1 items

Name	Provisioning ...	Approval state	VNet name	Possible Link...	Linked resource ID
AzurePostgreSqlPrivateEndpoint	Succeeded	Approved	default	0	[redacted]

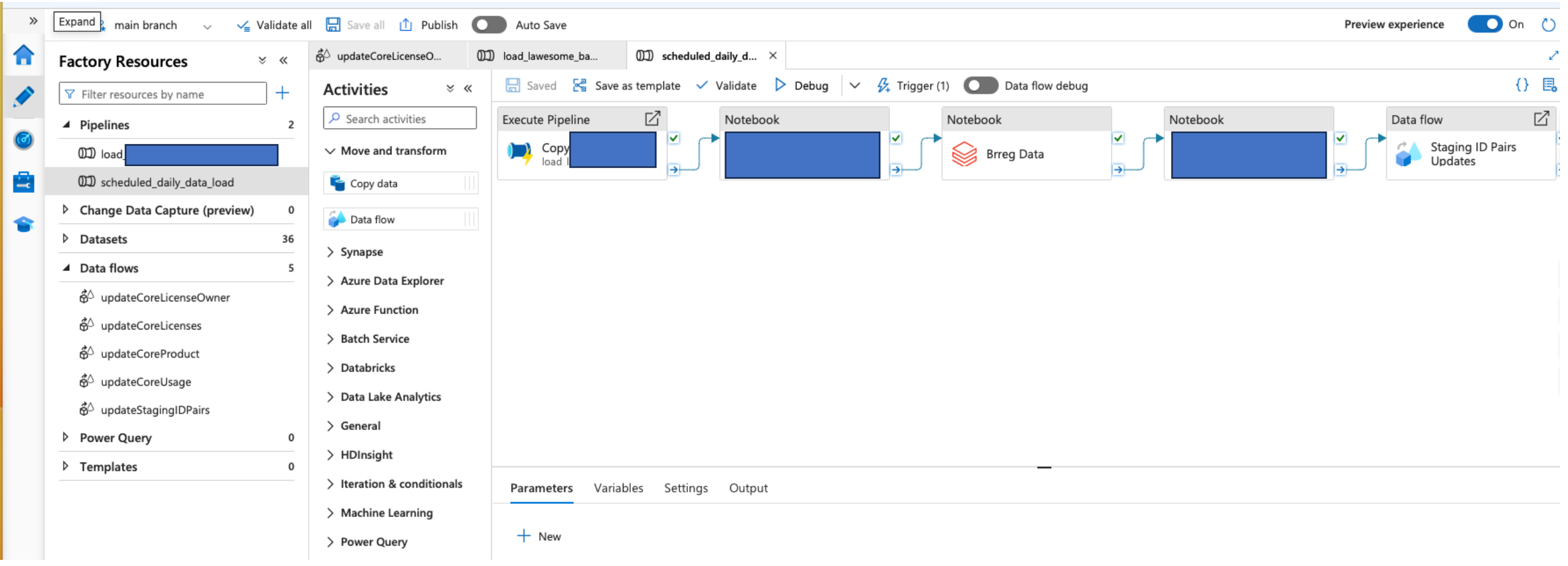
Data flow: food handling



I use it to arrange the fridge

Pipelines and its triggers: Orchestration and Automation - which first, and when

Pipelines can call other Azure resources which can be used for handling data.



Click around and play with it. ☺