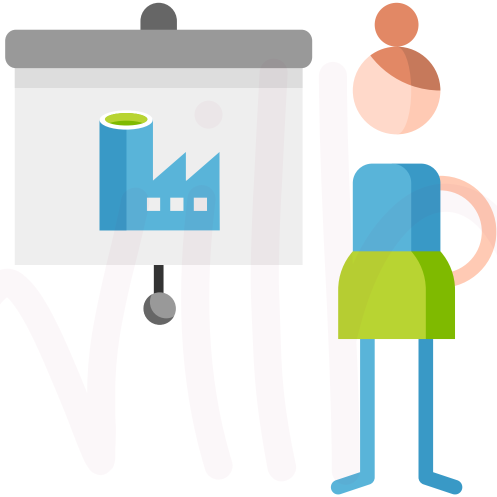
February 6th-7th

nic
20/20 VISION

Oslo Spektrum

Understanding Azure Data Factory

The What, When, and Why



Cathrine Wilhelmsen

NIC · February 6th, 2019

Understanding Azure Data Factory

What is at the core of every Business Intelligence, Data Science, and Machine Learning project?

Data.

You need data to understand what has happened in the past, to predict what may happen in the future, to discover patterns and anomalies, and to gain the insight necessary for making faster and better decisions.

But before you can do any of those things, you need to collect, store, transform, integrate, and prepare your data. Azure Data Factory (ADF) is a service that enables you to quickly and efficiently create automated data pipelines – without having to write any code!

In this session, we will go through the fundamentals of Azure Data Factory and see how easy it is to build solutions that can work with all your data on-premises and in the cloud. We will explore some key features such as Mapping Data Flows for visual data transformations and Wrangling Data Flows for visual data preparation, as well as how to schedule and orchestrate your finished data pipelines. Throughout the session, we will discuss different use cases and scenarios, as well as when and why you should use Azure Data Factory for your projects.

cathrine WILHELMSEN

inmeta
EVANGELIST
MICROSOFT
AZURE

MVP
Microsoft®
Most Valuable
Professional

@cathrinew

CW cathrinew.net

Big Data and Analytics

Data Warehousing

Business Intelligence

Artificial Intelligence

Machine Learning

Data Science

Big Data and Analytics

Data Warehousing

Business Intelligence

Artificial Intelligence

Machine Learning

Data Science

What has happened?

What?

What will happen?

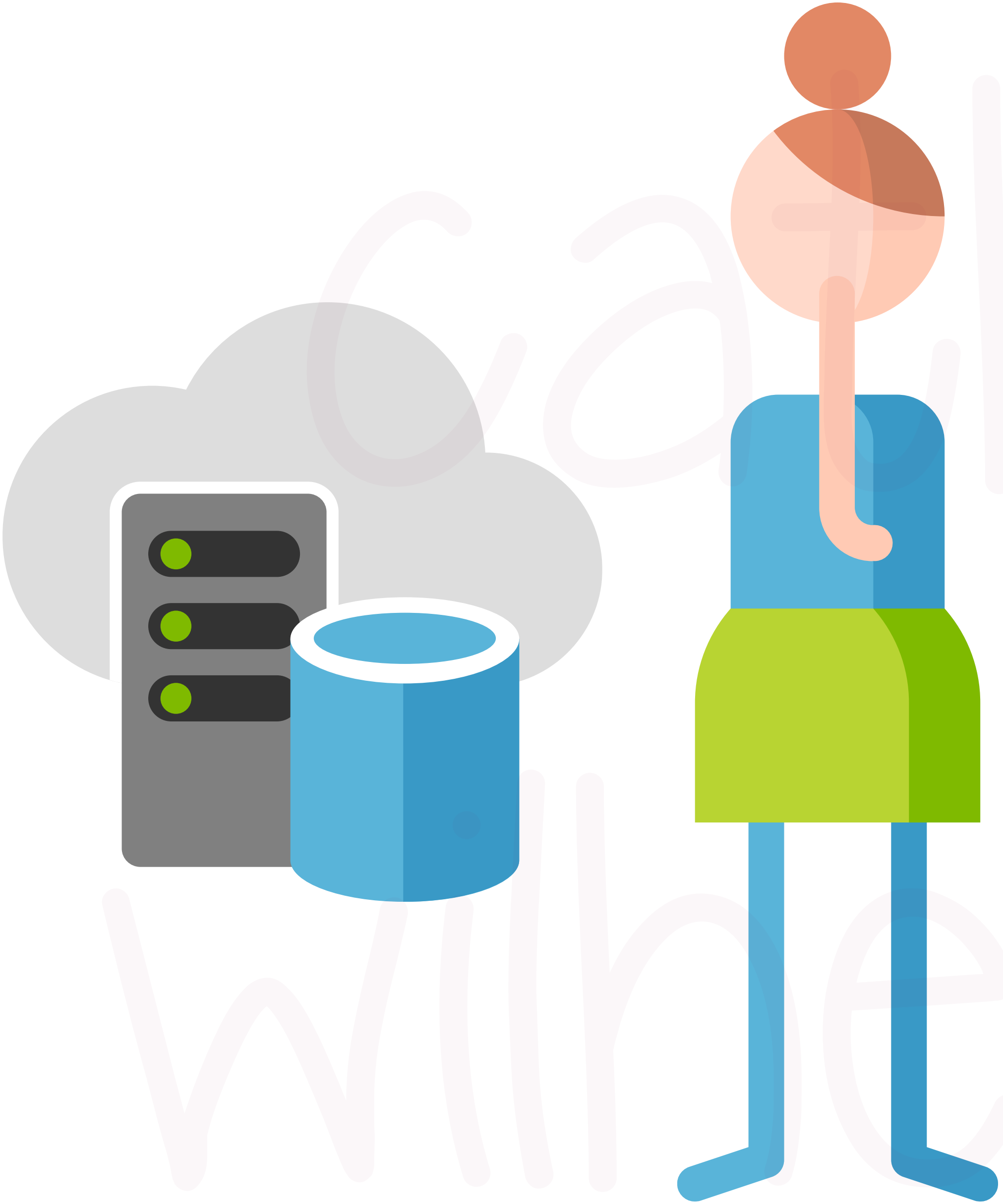
When did it happen?

When?

When will it happen?

Why?

Why did it happen?



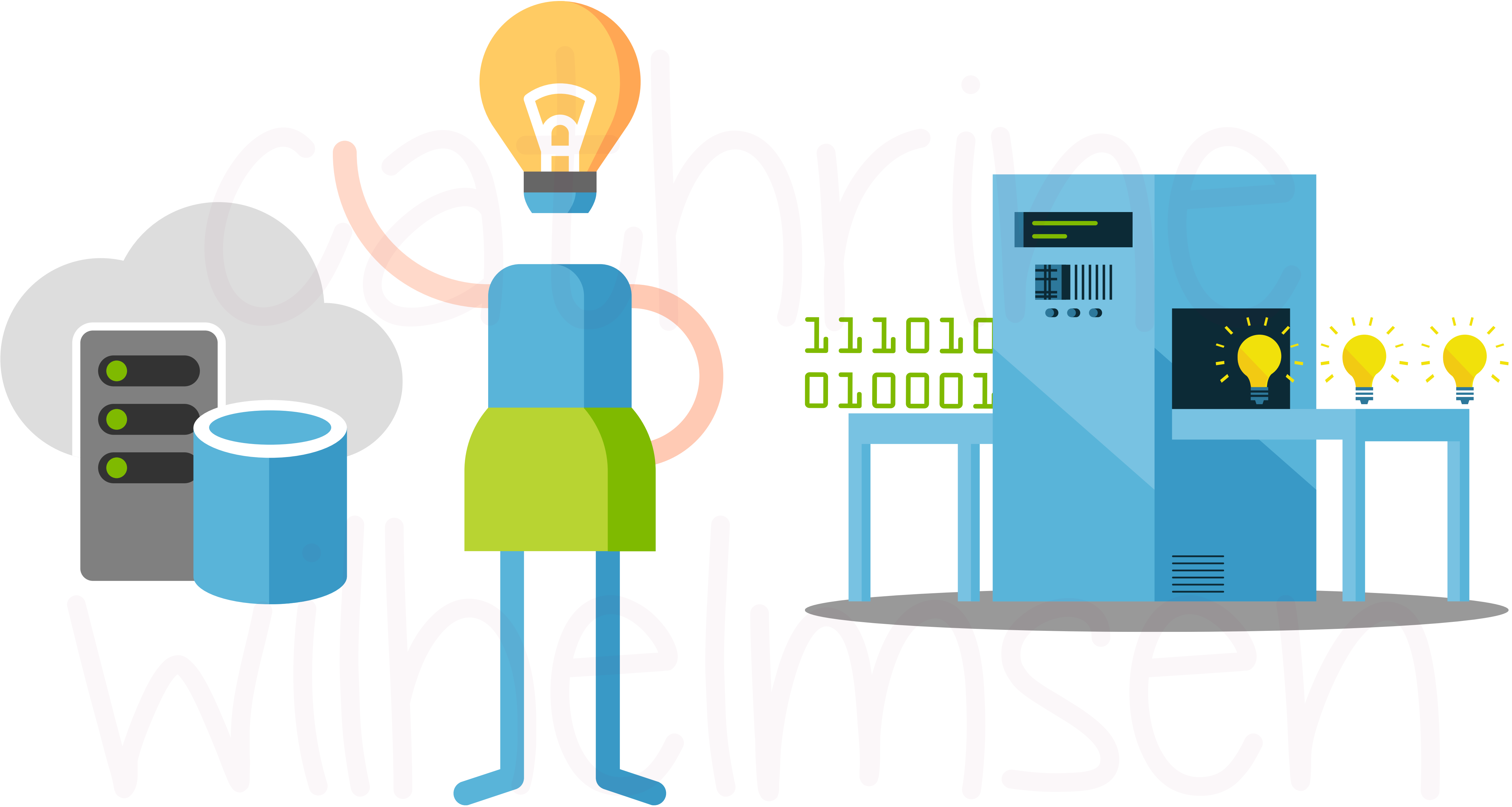
Collect

Store

Transform

Integrate

Prepare


111010
010001

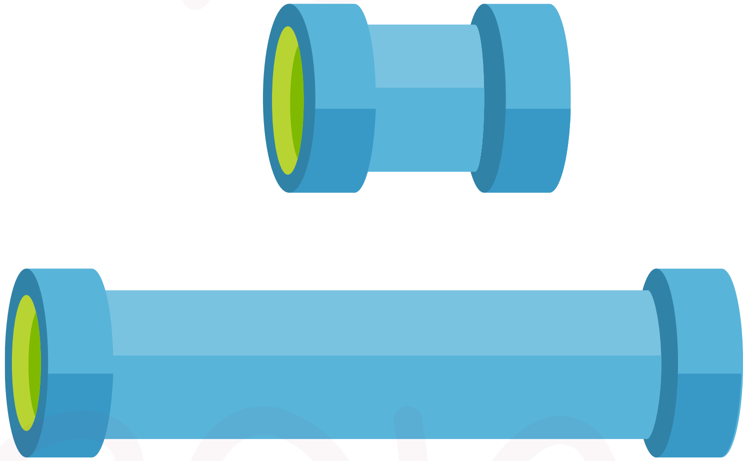
Azure Data Factory

What is Azure Data Factory?

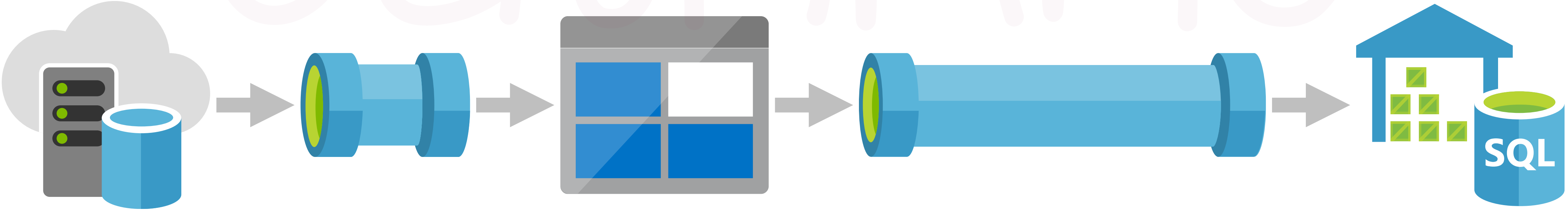
Hybrid data integration service

Complex and scalable pipelines

No-code ETL/ELT data flows



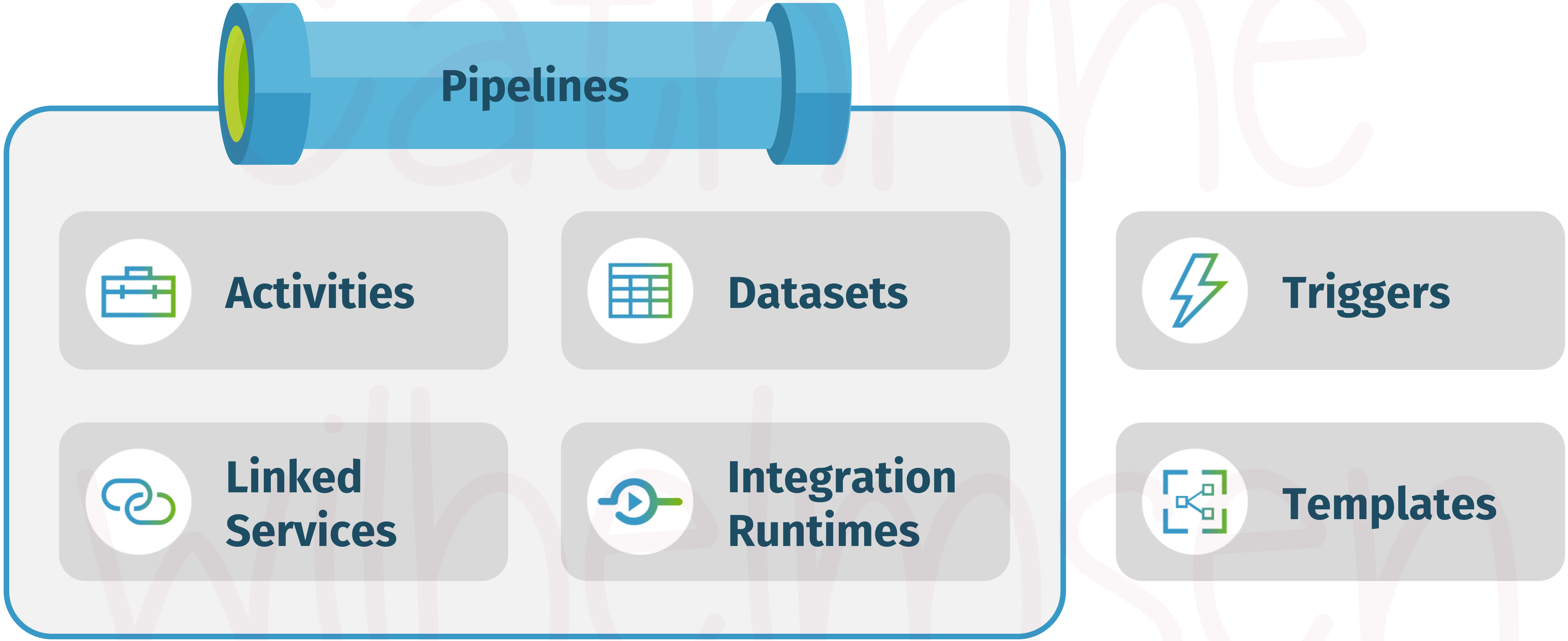
What can you do in Azure Data Factory?



Copy Data

Transform Data

What is inside Azure Data Factory?



Pipelines

Activities

Datasets

Linked Services

Integration Runtimes

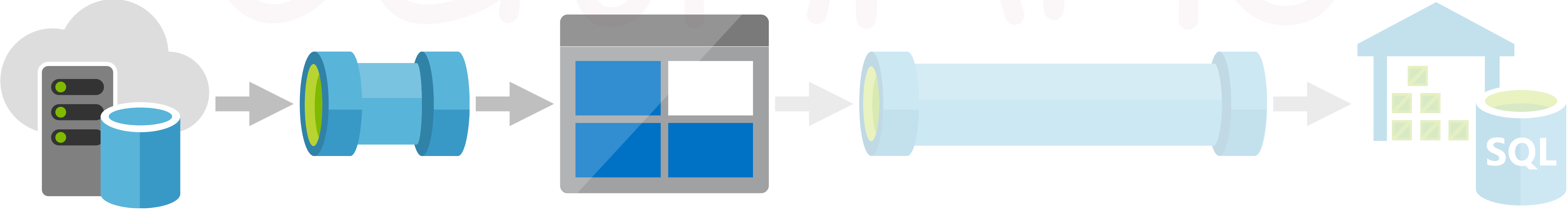
Triggers

Templates

DEMO

Let's look inside Azure Data Factory!

What can you do in Azure Data Factory?



Copy Data

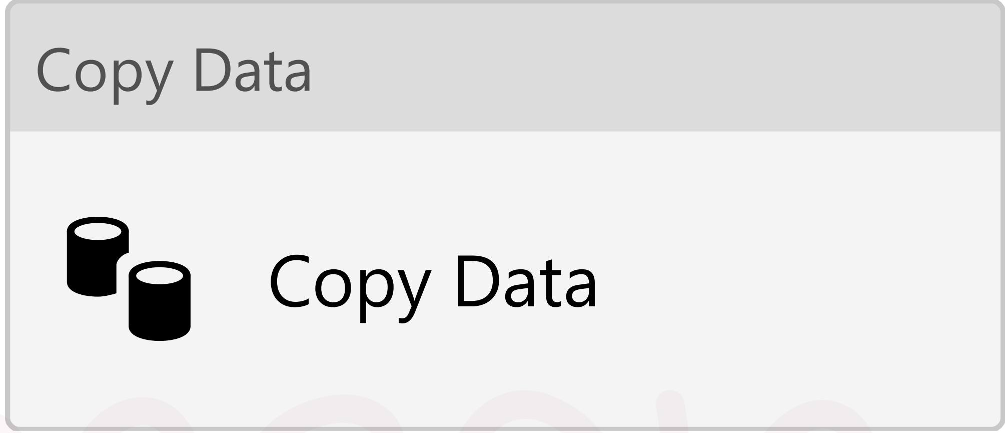
Transform Data

What is the Copy Data Activity?

The *core* activity *

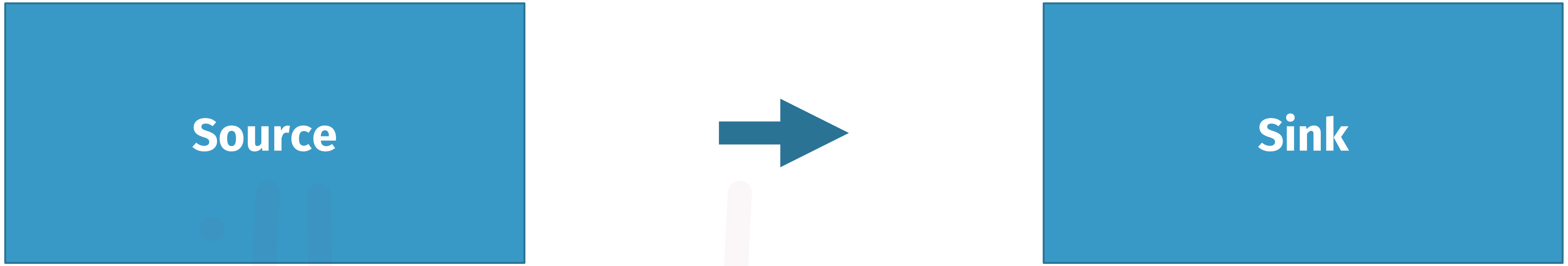
Supports 80+ connectors

Copy from *Source* to *Sink*



* *Cathrine's opinion :)*

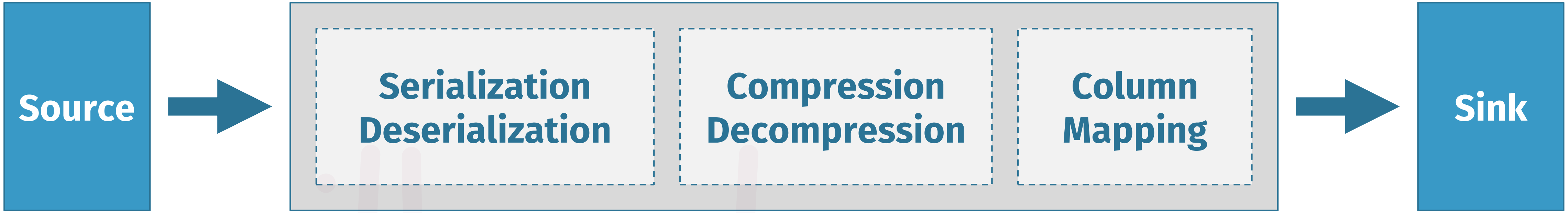
Copy Data Process: Binary Files



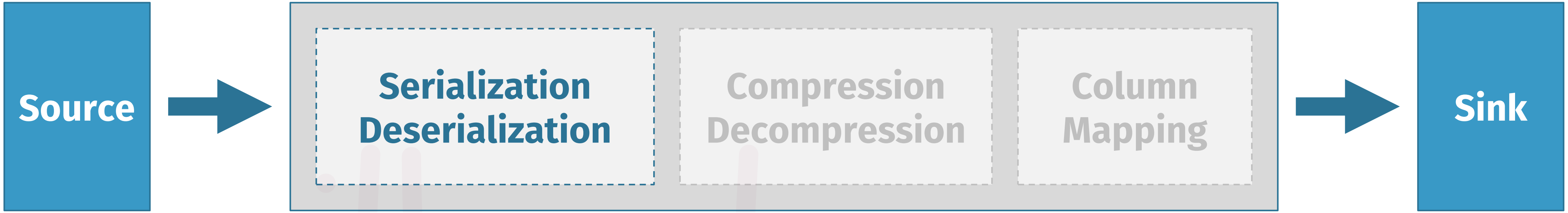
Source

Sink

Copy Data Process: Complex Files

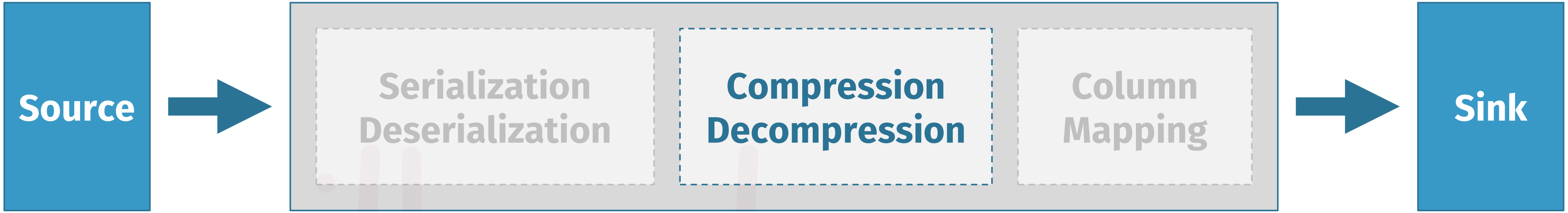


Copy Data Process: Complex Files



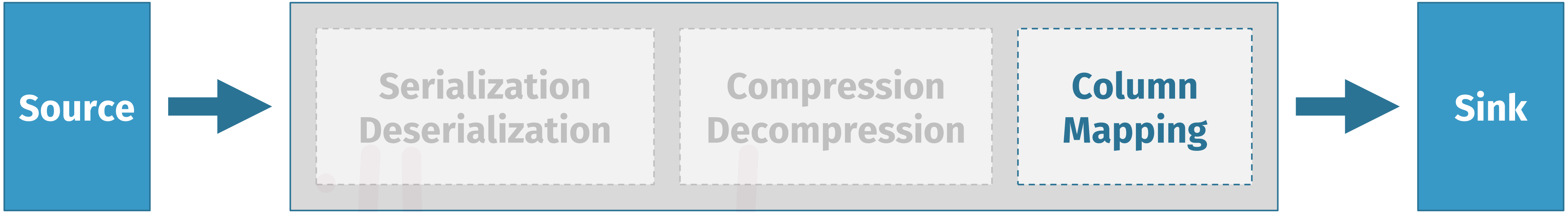
Convert file formats

Copy Data Process: Complex Files



Zip or unzip files

Copy Data Process: Complex Files



Map columns implicitly or explicitly

DEMO

Let's copy some data!

what if my systems are on-premises?

Hybrid Azure Data Factory

Azure excels at cloud data integrations, but can also work with your on-premises systems!

What are Integration Runtimes?

Azure Integration Runtime

Self-Hosted Integration Runtime

Azure Integration Runtime

Restrict to specific Azure regions

- Data does not leave that specific region

Fully managed compute infrastructure

- Scale up by specifying Data Integration Units (DIUs)

Self-Hosted Integration Runtime

Acts like a gateway

- Get access to on-premises system within the network

Bring your own compute infrastructure

- Scale out by installing up to 4 nodes

Copy Data Scenarios

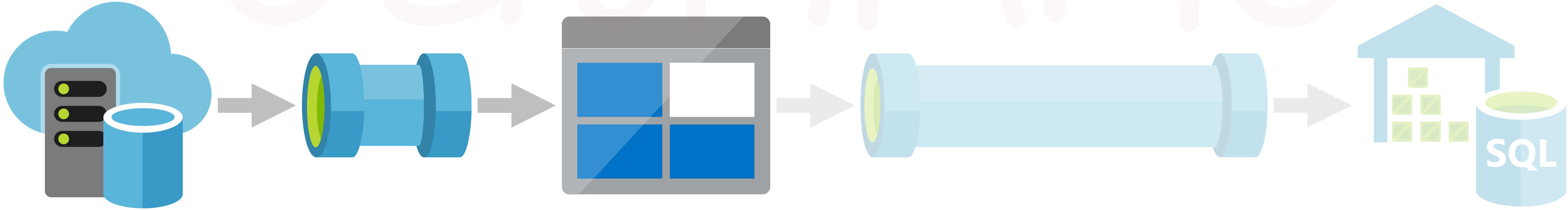
Use Azure Data Factory to copy data between:

- Cloud Stores
- Cloud and On-premises Stores
- On-premises Stores

DEMO

Let's connect to an *on-prem* SQL Server!

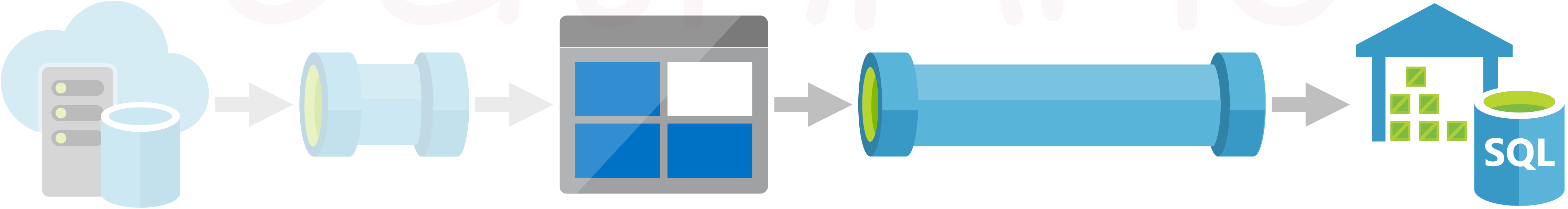
Ok, so we can copy data...



Copy Data

Transform Data

...what about transforming data?



Copy Data

Transform Data

Mapping or Wrangling

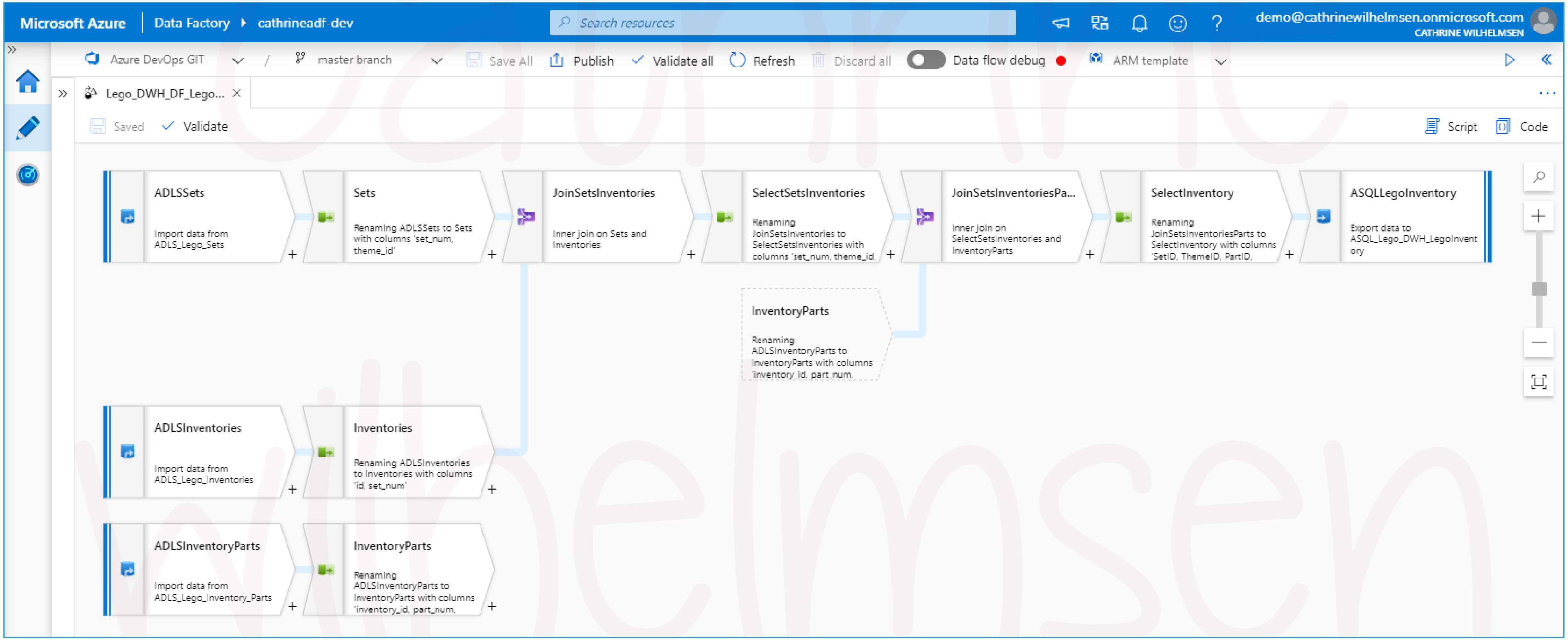
What are Mapping Data Flows?

Data *transformation* at scale

Visual editor, no-code experience

Runs on Spark clusters

How do Mapping Data Flows work?



ADLSSets
Import data from ADLS_Lego_Sets

Sets
Renaming ADLSSets to Sets with columns 'set_num, theme_id'

JoinSetsInventories
Inner join on Sets and Inventories

SelectSetsInventories
Renaming JoinSetsInventories to SelectSetsInventories with columns 'set_num, theme_id,

JoinSetsInventoriesPa...
Inner join on SelectSetsInventories and InventoryParts

SelectInventory
Renaming JoinSetsInventoriesParts to SelectInventory with columns 'SetID, ThemeID, PartID,

ASQLLegoInventory
Export data to ASQL_Lego_DWH_LegoInventory

InventoryParts
Renaming ADLSInventoryParts to InventoryParts with columns 'inventory_id, part_num,

ADLSInventories
Import data from ADLS_Lego_Inventories

Inventories
Renaming ADLSInventories to Inventories with columns 'id, set_num'

ADLSInventoryParts
Import data from ADLS_Lego_Inventory_Parts

InventoryParts
Renaming ADLSInventoryParts to InventoryParts with columns 'inventory_id, part_num,

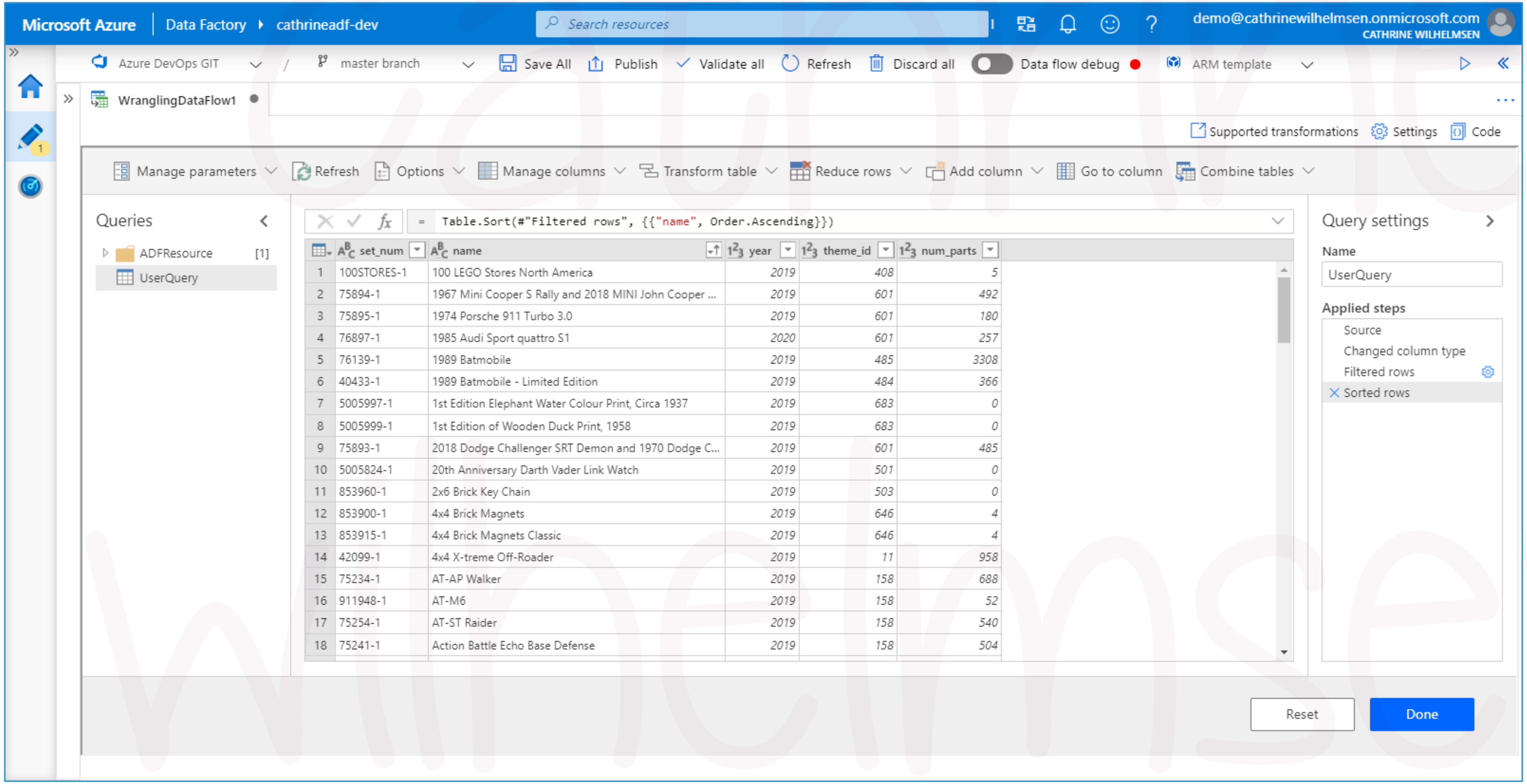
What are Wrangling Data Flows?

Data *preparation* at scale

Visual editor, no-code experience

Runs Power Query Online

How do Wrangling Data Flows work?



```
= Table.Sort(#"Filtered rows", {{"name", Order.Ascending}})
```

	set_num	name	year	theme_id	num_parts
1	100STORES-1	100 LEGO Stores North America	2019	408	5
2	75894-1	1967 Mini Cooper S Rally and 2018 MINI John Cooper ...	2019	601	492
3	75895-1	1974 Porsche 911 Turbo 3.0	2019	601	180
4	76897-1	1985 Audi Sport quattro S1	2020	601	257
5	76139-1	1989 Batmobile	2019	485	3308
6	40433-1	1989 Batmobile - Limited Edition	2019	484	366
7	5005997-1	1st Edition Elephant Water Colour Print, Circa 1937	2019	683	0
8	5005999-1	1st Edition of Wooden Duck Print, 1958	2019	683	0
9	75893-1	2018 Dodge Challenger SRT Demon and 1970 Dodge C...	2019	601	485
10	5005824-1	20th Anniversary Darth Vader Link Watch	2019	501	0
11	853960-1	2x6 Brick Key Chain	2019	503	0
12	853900-1	4x4 Brick Magnets	2019	646	4
13	853915-1	4x4 Brick Magnets Classic	2019	646	4
14	42099-1	4x4 X-treme Off-Roader	2019	11	958
15	75234-1	AT-AP Walker	2019	158	688
16	911948-1	AT-M6	2019	158	52
17	75254-1	AT-ST Raider	2019	158	540
18	75241-1	Action Battle Echo Base Defense	2019	158	504

Query settings

Name

UserQuery

Applied steps

- Source
- Changed column type
- Filtered rows
- Sorted rows

Mapping Data Flows

Data Transformation

Similar to SSIS

Wrangling Data Flows

Data Preparation

Power Query Online

DEMO

Let's transform some data!

how do we schedule data pipelines?

Trigger pipelines...

- On a set Schedule
- In a Tumbling Window
- When Event Happens
- Now

Triggers: Schedule

Execute one or more pipelines on a set schedule

- Every Wednesday at 06:00
- Last day of the month at 18:00
- Every Monday at 04:00 and Friday at 20:00

Triggers: Tumbling Window

Execute a single pipeline for each time slice

- For every 15 minutes
- For every 1 hour
- For every 24 hours

Triggers: Event Based

Execute one or more pipelines when event happens

- Blob is Created
- Blob is Deleted
- Blob is Created or Deleted

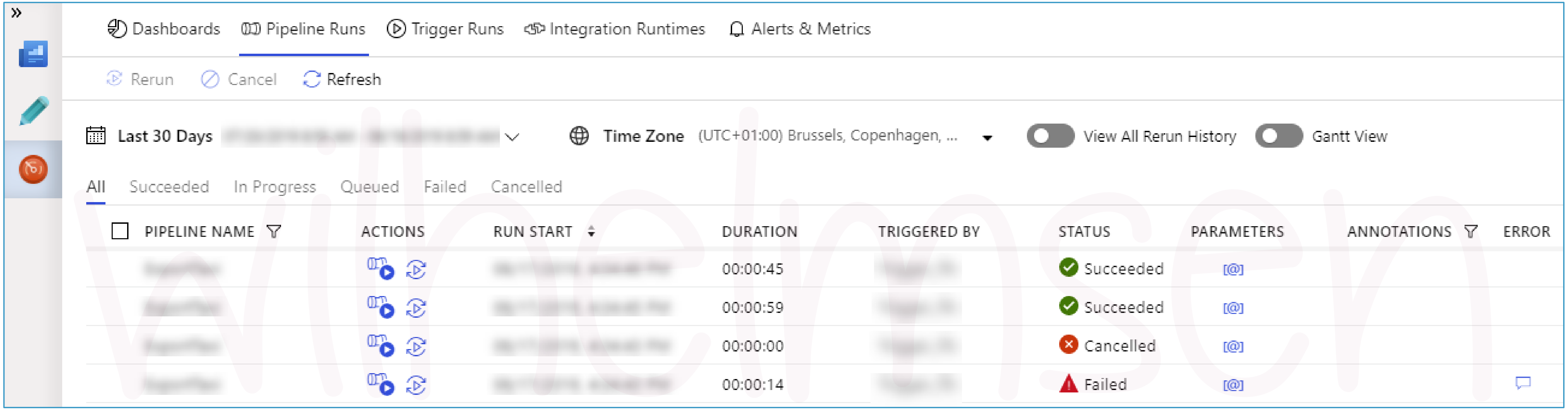
Triggers: Now

Execute a single pipeline immediately

Monitoring Triggers

Triggers save and log execution information

Information is available on the Monitor page



Last 30 Days [redacted]

Time Zone (UTC+01:00) Brussels, Copenhagen, ...

○ View All Rerun History ○ Gantt View

All Succeeded In Progress Queued Failed Cancelled

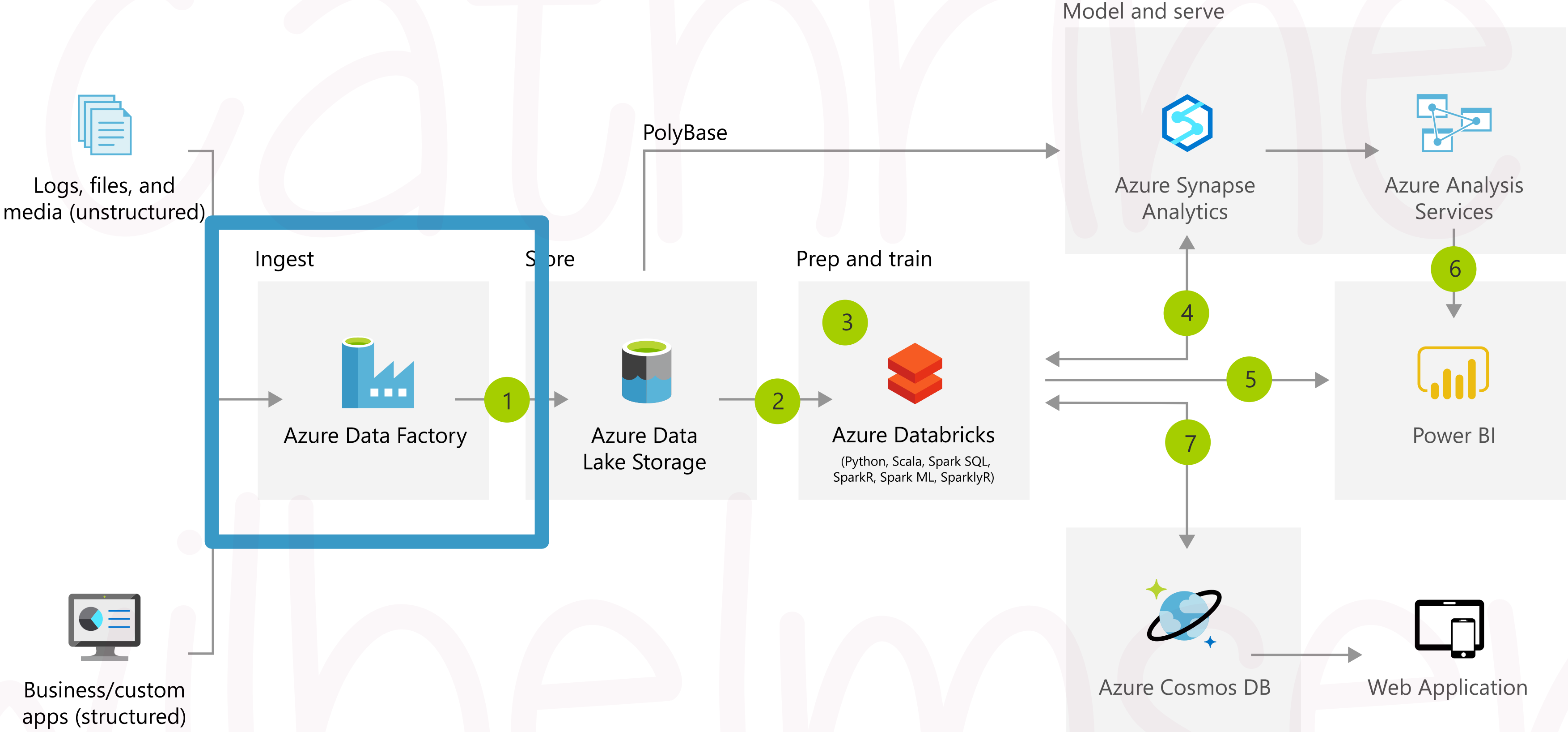
□	PIPELINE NAME	ACTIONS	RUN START	DURATION	TRIGGERED BY	STATUS	PARAMETERS	ANNOTATIONS	ERROR
	[redacted]		[redacted]	00:00:45	[redacted]	✓ Succeeded	[@]		
	[redacted]		[redacted]	00:00:59	[redacted]	✓ Succeeded	[@]		
	[redacted]		[redacted]	00:00:00	[redacted]	✗ Cancelled	[@]		
	[redacted]		[redacted]	00:00:14	[redacted]	⚠ Failed	[@]		

DEMO

Let's schedule some pipelines!

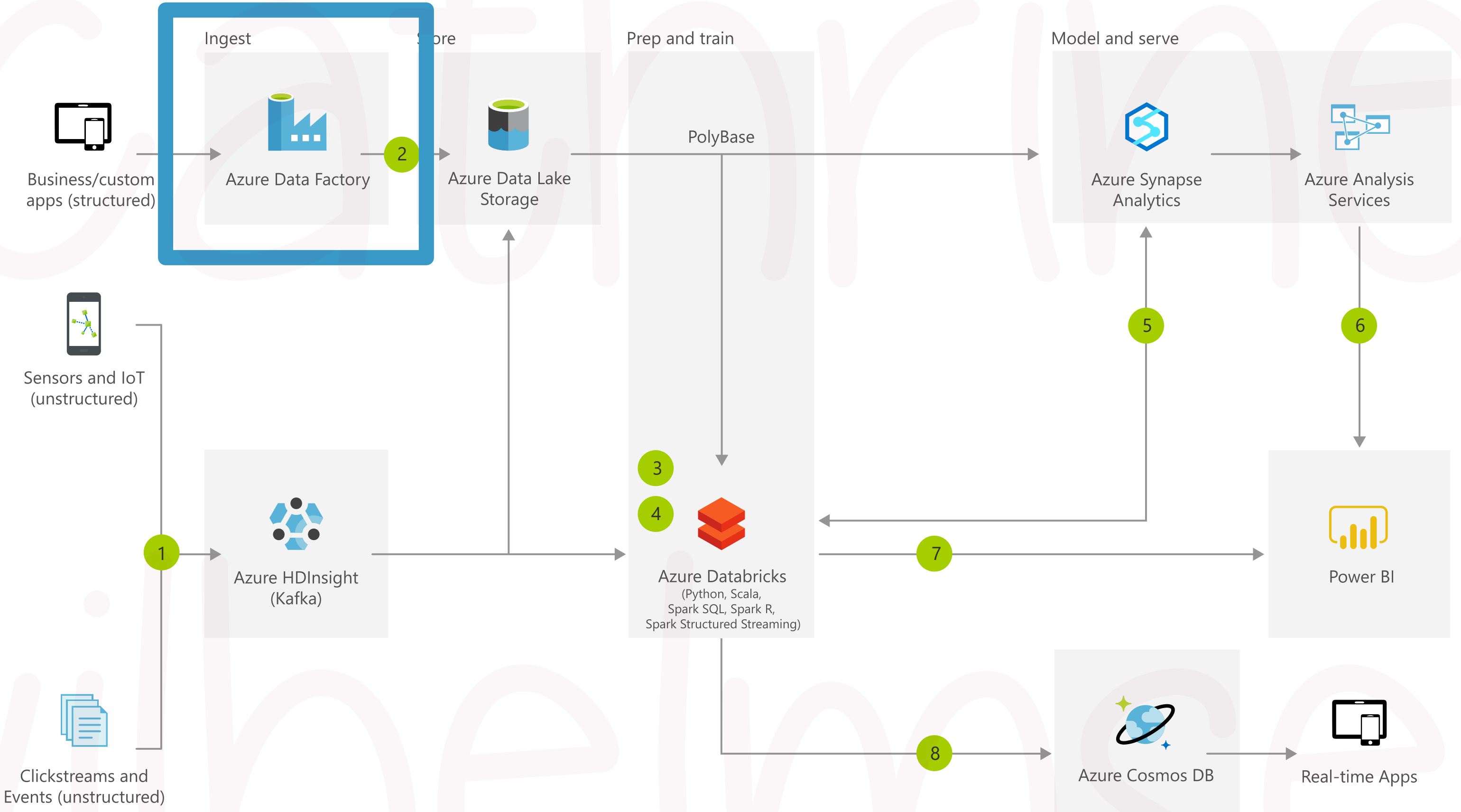
Azure Data Architectures

Advanced Analytics on Big Data



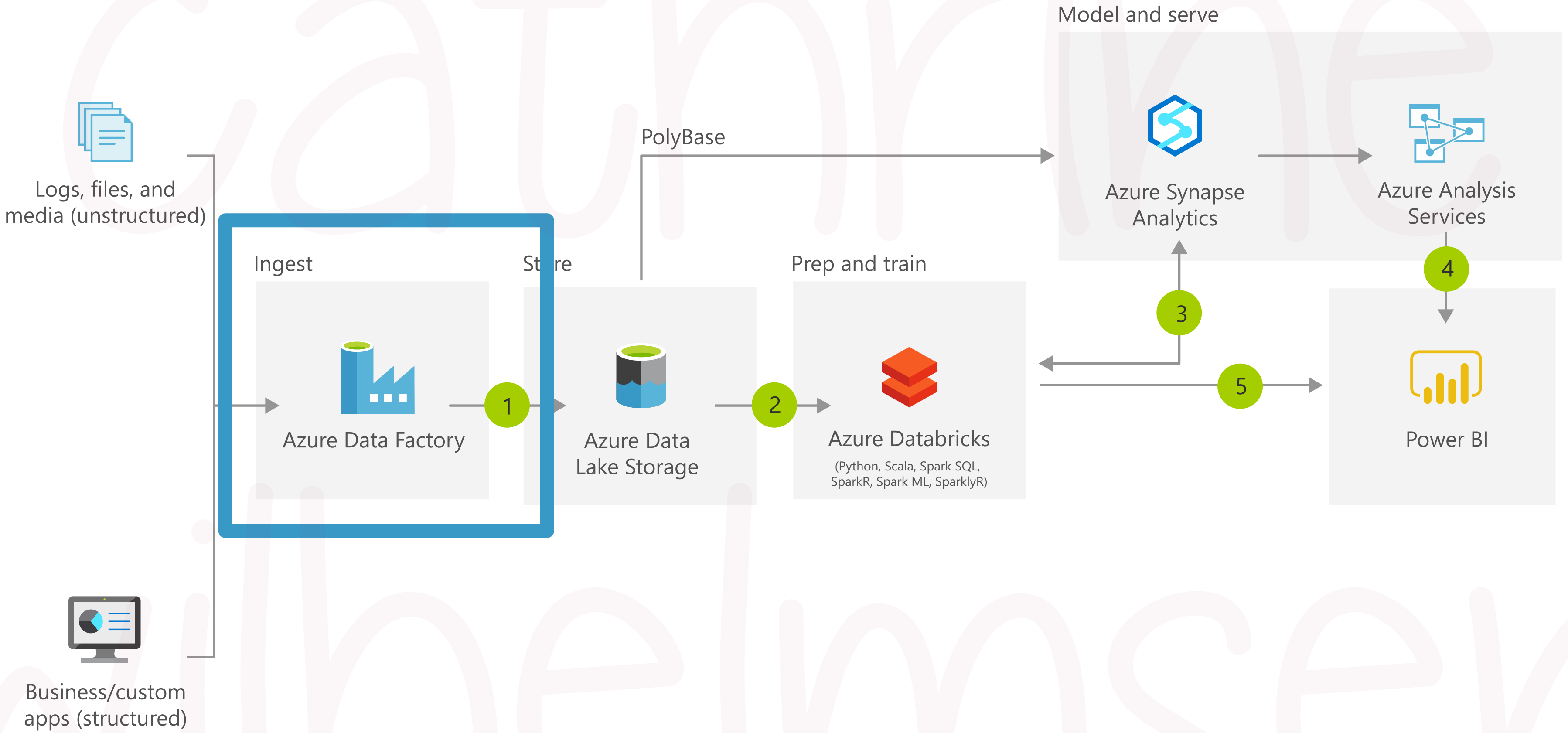
https://azure.microsoft.com/en-us/solutions/architecture/advanced-analytics-on-big-data/

Real-time Analytics

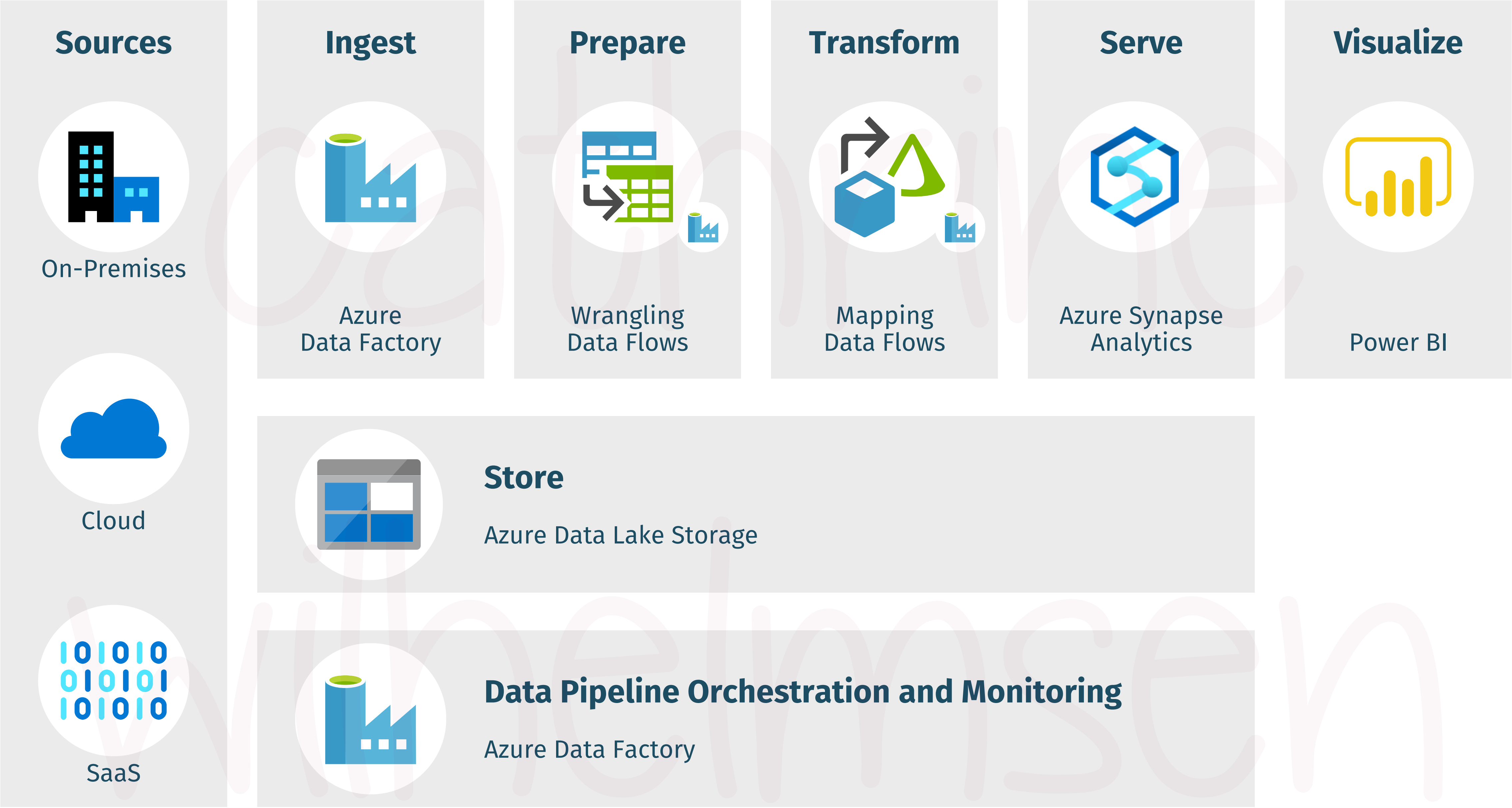


https://azure.microsoft.com/en-us/solutions/architecture/real-time-analytics/

Modern Data Warehouse



https://azure.microsoft.com/en-us/solutions/architecture/modern-data-warehouse/



Sources

On-Premises

Cloud

SaaS

Ingest

Azure Data Factory

Prepare

Wrangling Data Flows

Transform

Mapping Data Flows

Serve

Azure Synapse Analytics

Visualize

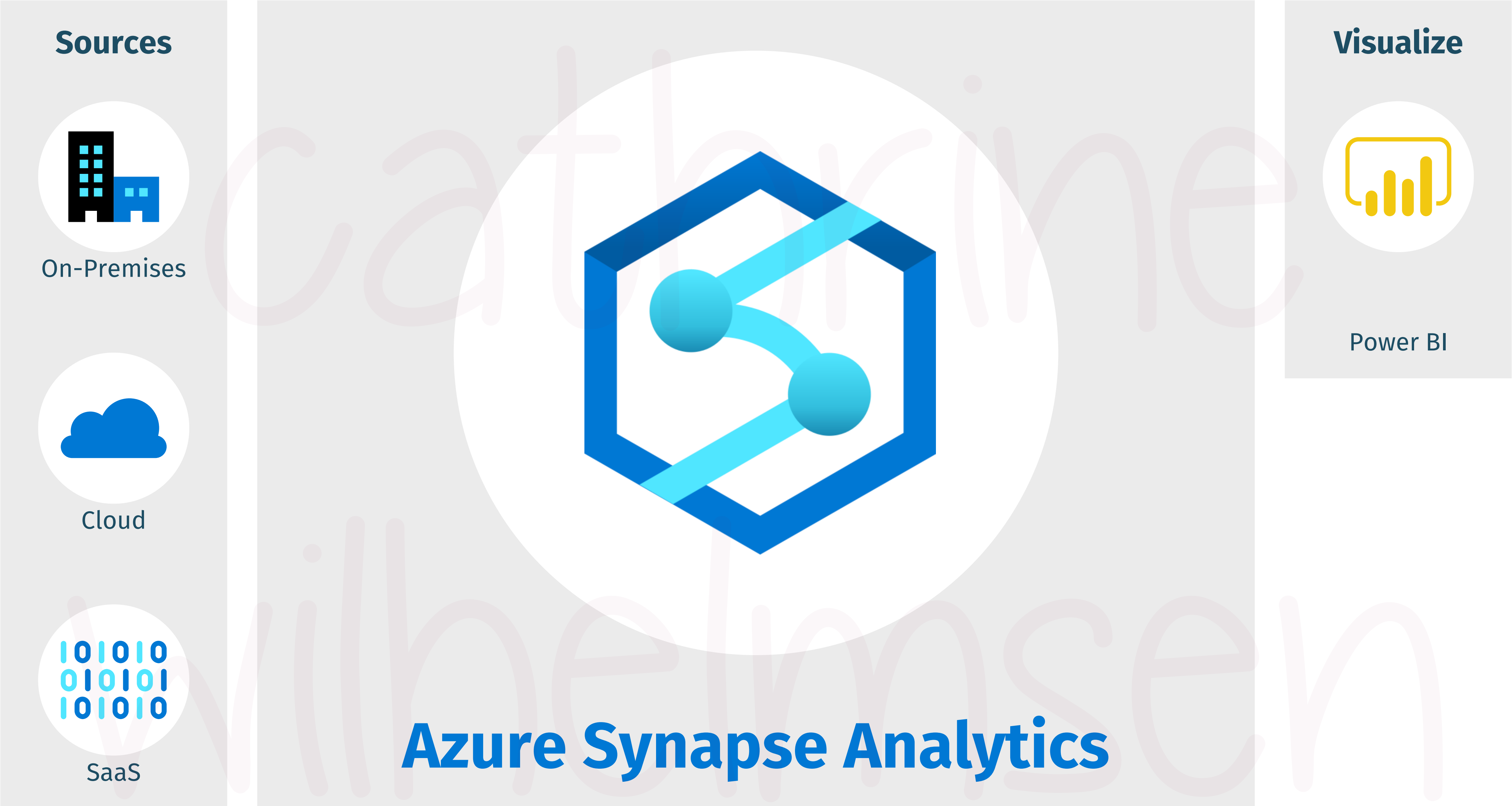
Power BI

Store

Azure Data Lake Storage

Data Pipeline Orchestration and Monitoring

Azure Data Factory



Sources

On-Premises

Cloud

SaaS

Azure Synapse Analytics

Visualize

Power BI

Good luck!

thank you!

hi@cathrinew.net

@cathrinew

CW cathrinew.net

inmeta EVANGELIST MICROSOFT AZURE