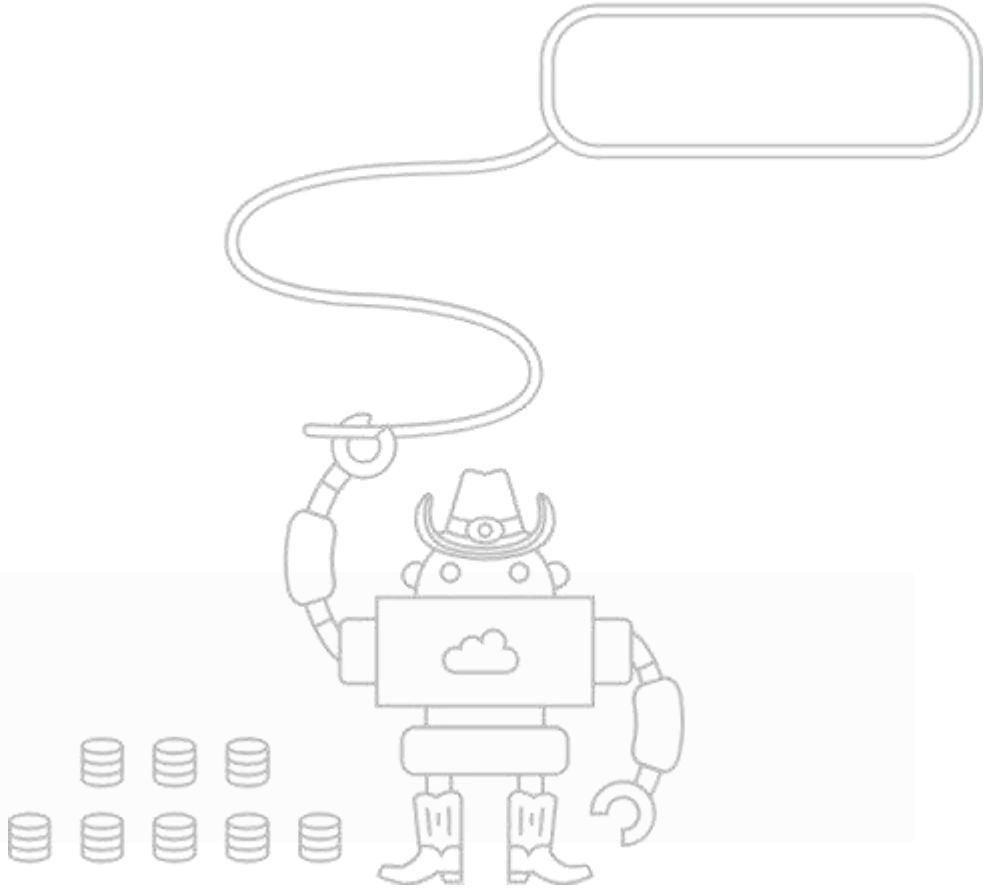
KI group



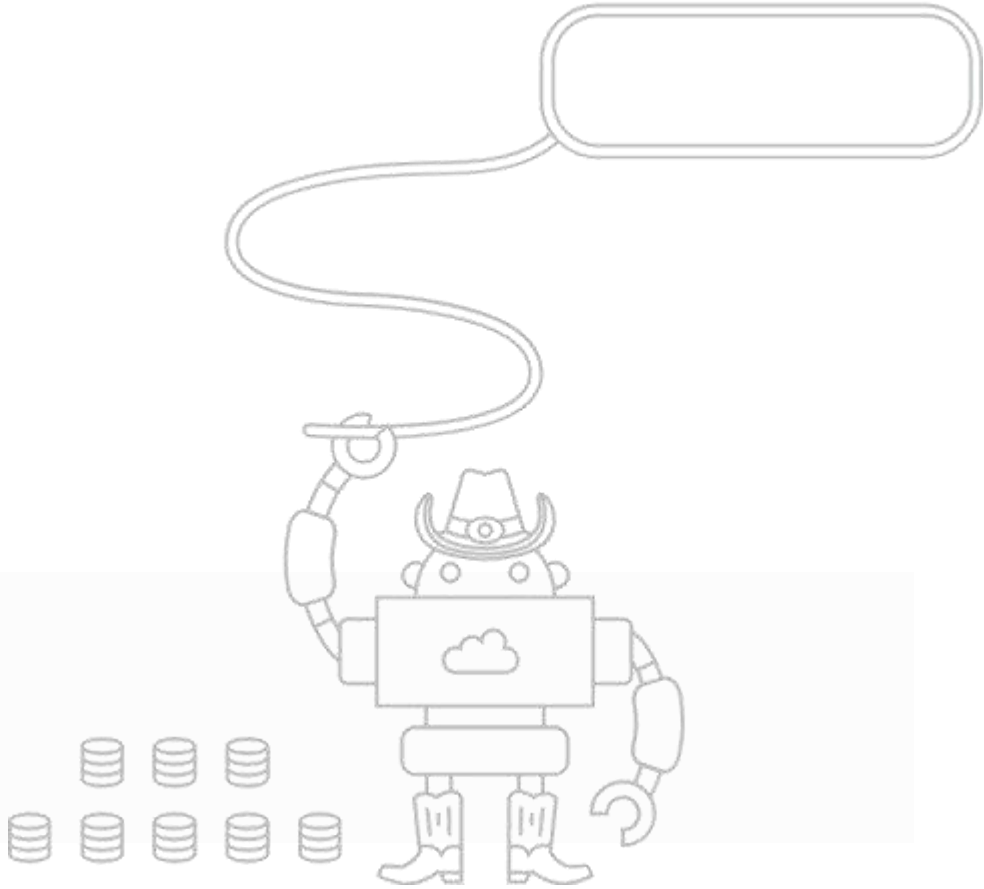
ADF WRANGLING DATA FLOWS

POWER QUERY GOES SPARK

Christoph Seck | KI Group

c.seck@kigroup.de

KI group



~~ADF~~ WRANGLING DATA FLOWS

POWER QUERY GOES ~~SPARK~~ DATA LAKE

Christoph Seck | KI Group

c.seck@kigroup.de

>280
EMPLOYEES

> 150 ENGINEERS

FOUNDED

1999

>300
PROJECTS

5 LOCATIONS

COLOGNE

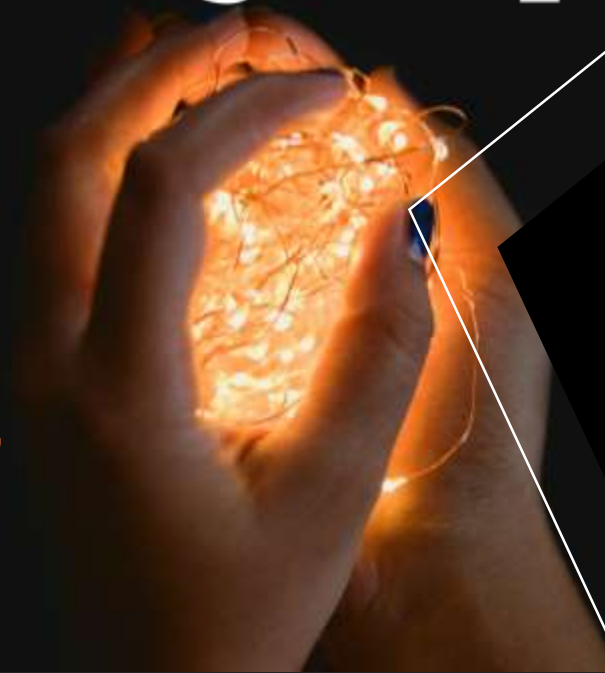
STUTTGART

BERLIN

MUNICH

LISBON

KI group



PASSION FOR TECH AND INNOVATIONS

HOME FOR ENTREPRENEURS, CREATORS AND SOLVERS

KI performance

BUSINESS MODELS
& PROCESSES

KI mobility KI finance
KI challenge.rs KI growth
KI chemicals

EXPERIENCES , DESIGN AND LIFESTYLE

dreaminc GROUP
dreaminc MEDIA
dreaminc SPACES
dreaminc ENTERTAINMENT
Food Experts Group

SOFTWARE
& DATA

KI performance
KI decentralized
KI analytics
KI labs
MOBILAB

KI INTRODUCING group

HOME FOR ENTREPRENEURS,
CREATORS AND SOLVERS

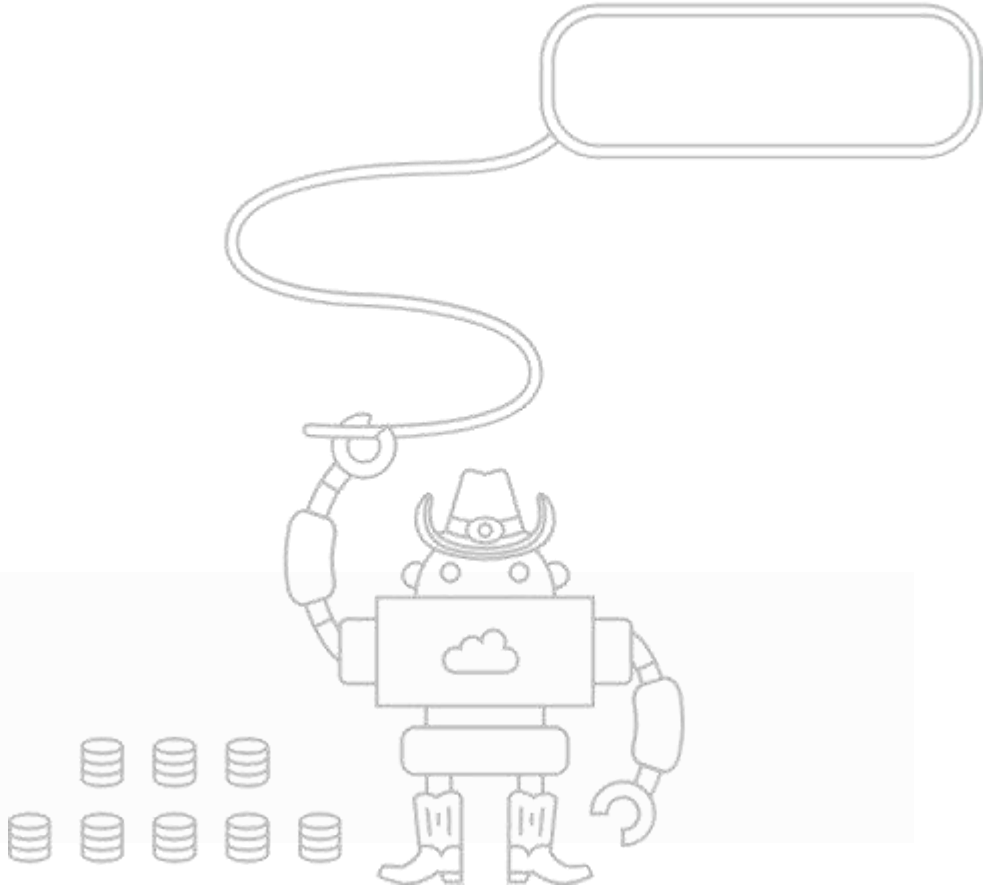
INVESTMENTS,
COMPANYBUILDING
AND ECOSYSTEM

KI capital

HUMAN RESOURCES

KI professionals KI connect

KI group

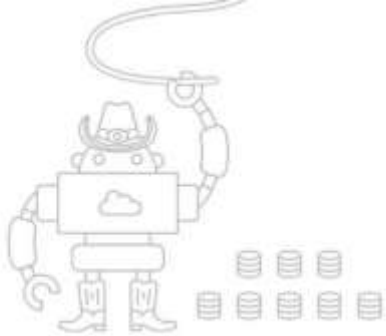


WRANGLING DATA FLOWS

POWER QUERY GOES DATA LAKE

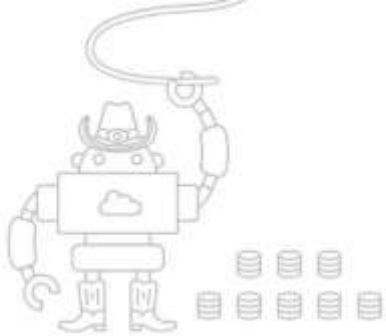
Christoph Seck | KI Group

c.seck@kigroup.de



Agenda

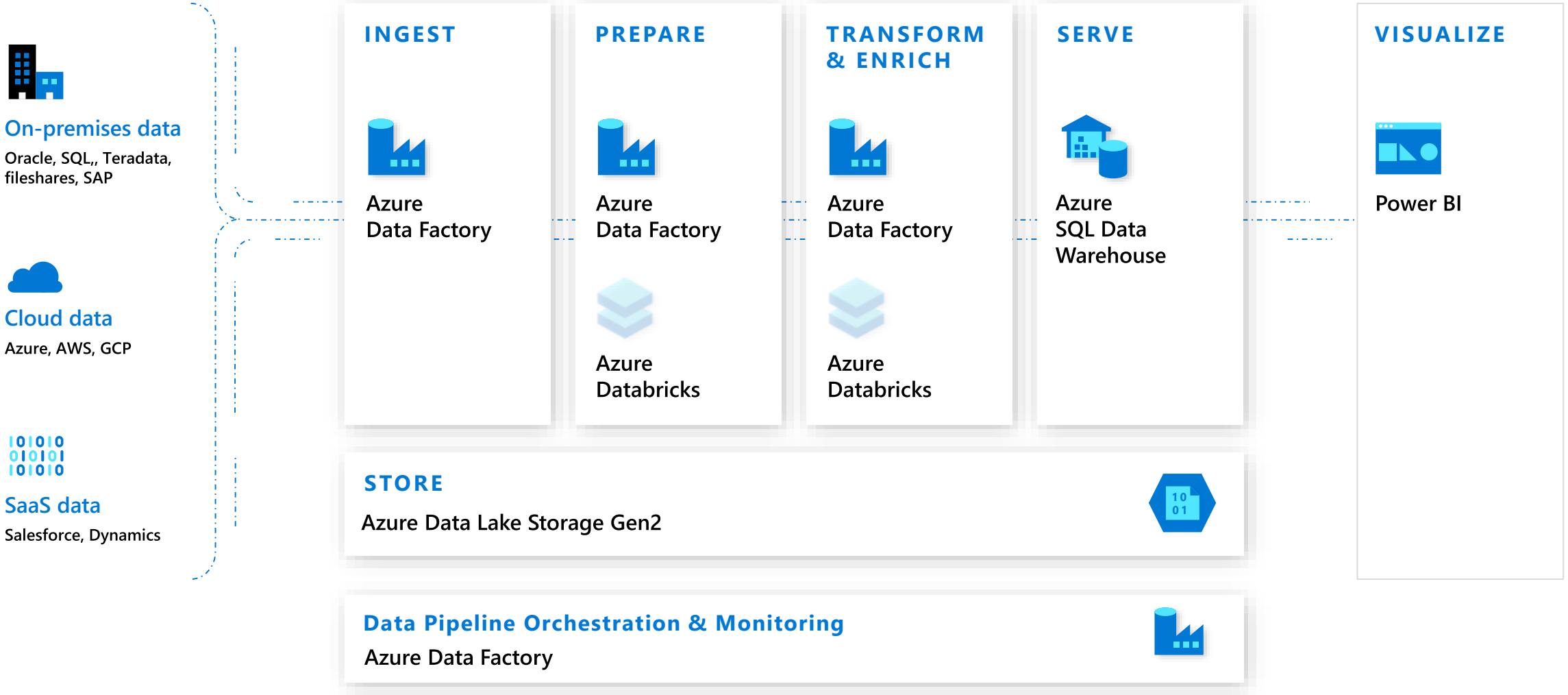
- Wrangling in Context
- Refresher Power Query & M
- ADF Data Wrangling
- The swotty Power Cousin
- Resumé



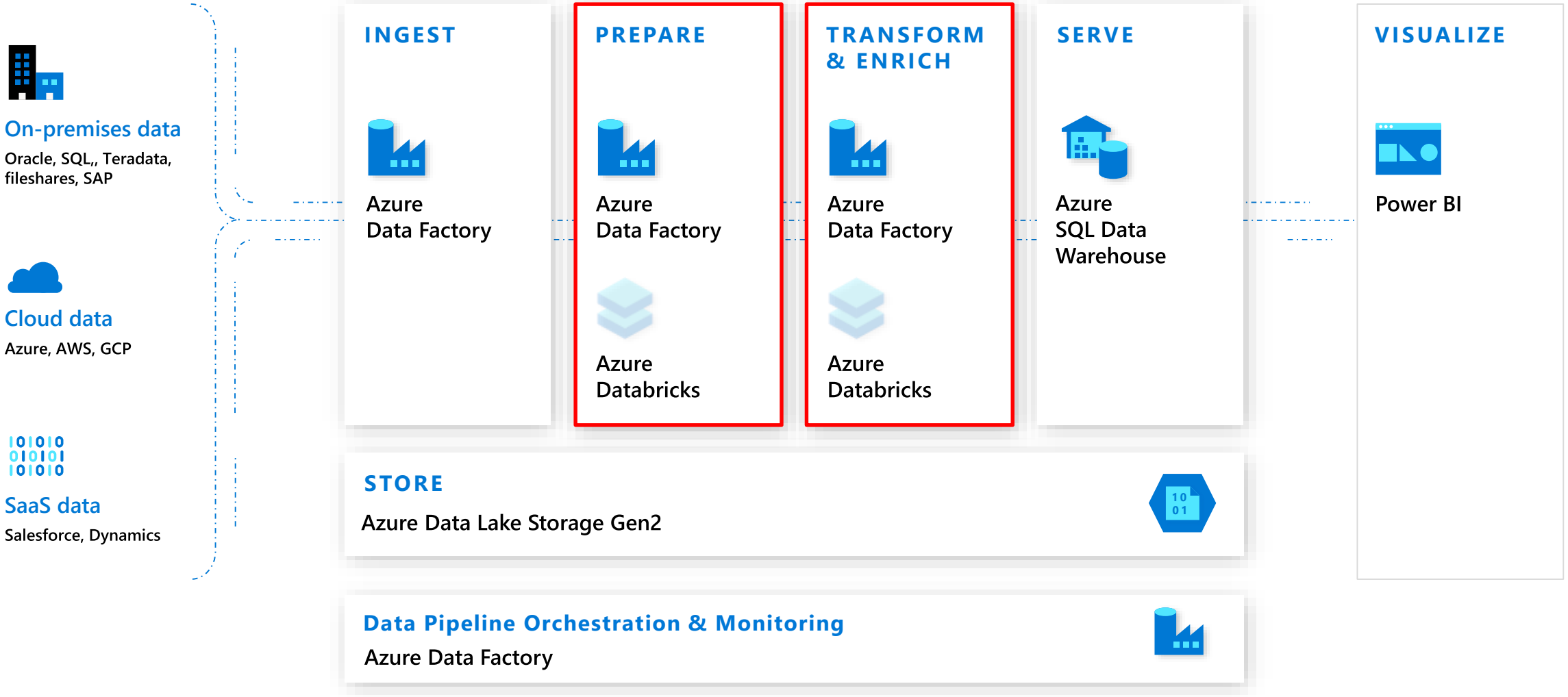
Agenda

- Wrangling in Context
- Refresher Power Query & M
- ADF Data Wrangling
- The swotty Power Cousin
- Resumé

Modern Data Warehouse (MDW)



Modern Data Warehouse (MDW)

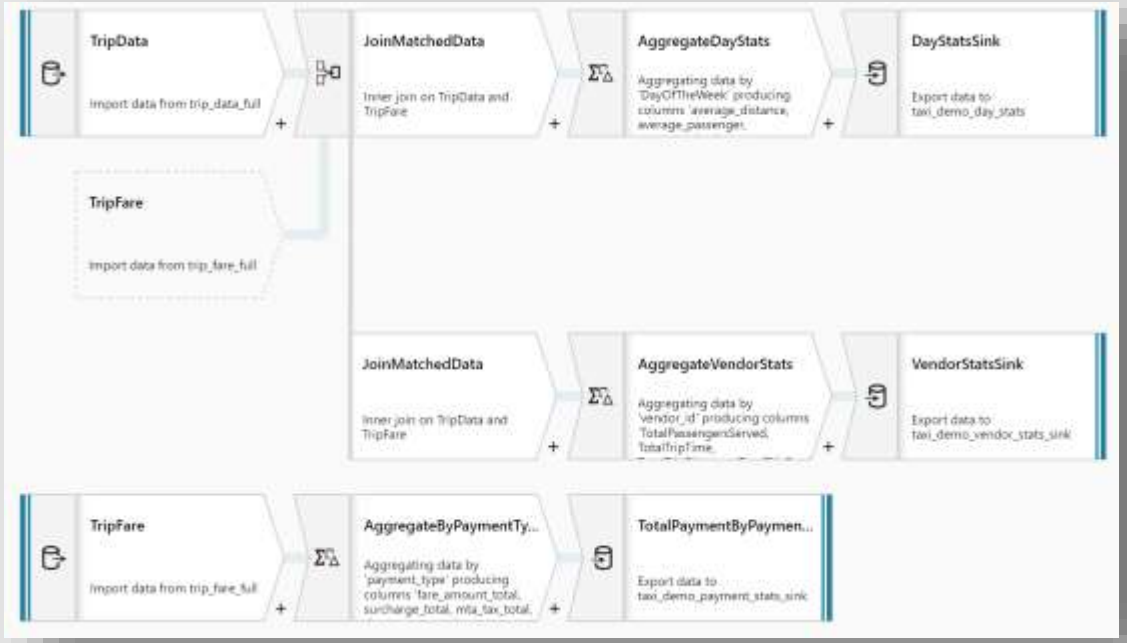


MAPPING DATAFLOW

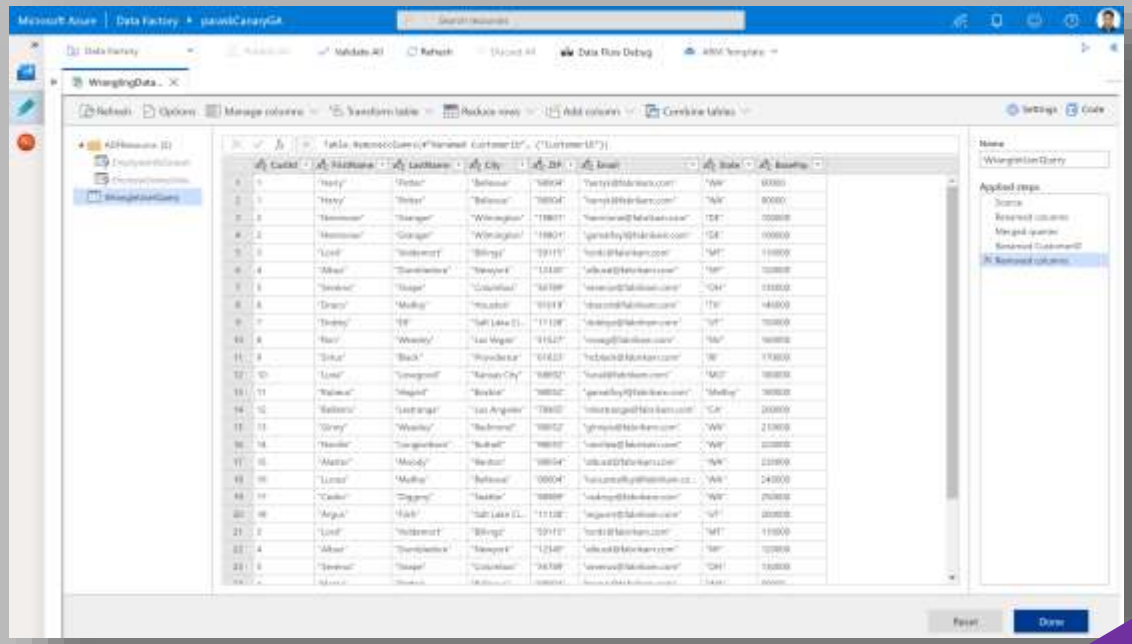
Code-free data transformation @scale

PUBLIC PREVIEW

MAPPING DATAFLOW
Code-free data transformation @scale

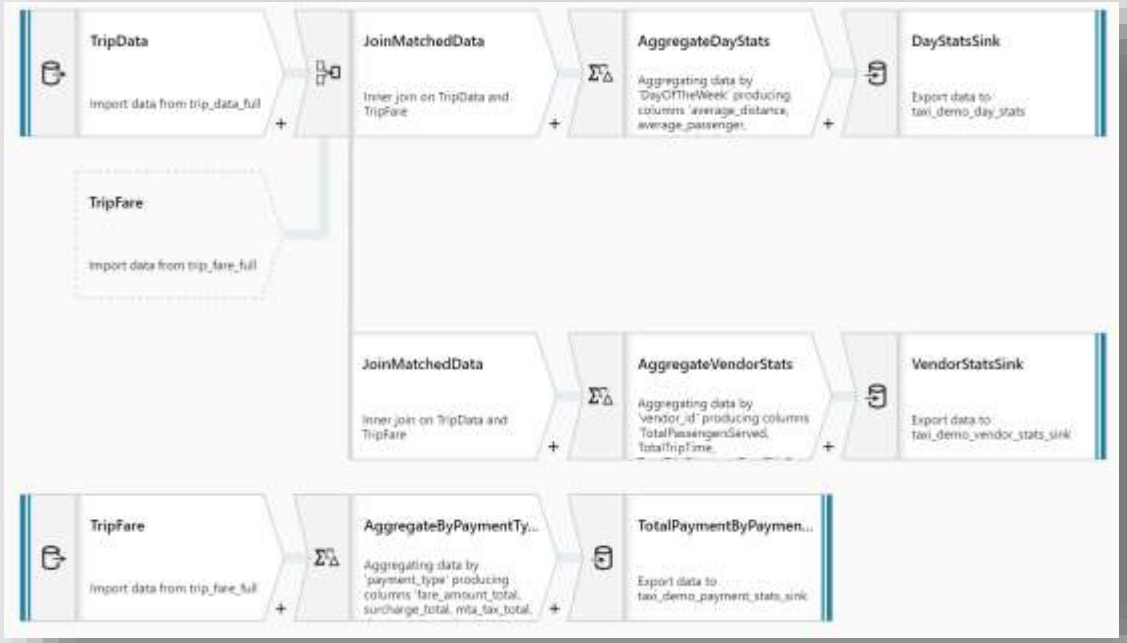


WRANGLING DATAFLOW
Code-free data preparation @scale

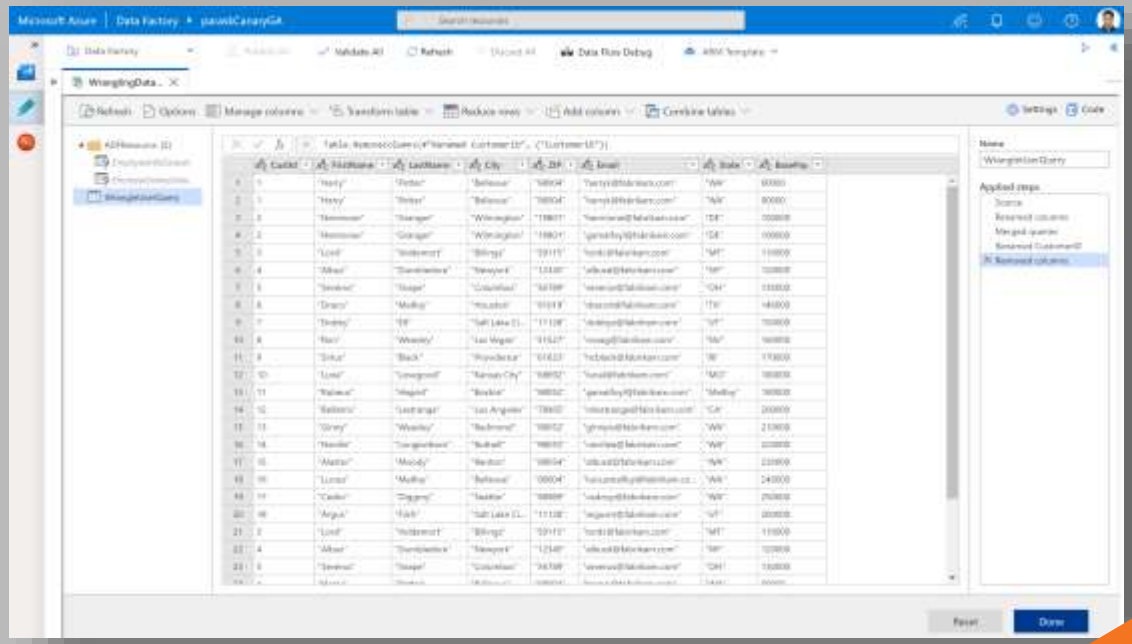


LIMITED PREVIEW

MAPPING DATAFLOW
Code-free data transformation @scale



WRANGLING DATAFLOW
Code-free data preparation @scale



PUBLIC PREVIEW

Data Preparation in today's BI world

- **Finding** & **Connecting** to data is easy for users
- Experiences for data connectivity are **common across different tools**
- Data is **always in the desired shape** for analysis/consumption
- **Reshaping data is easy and quick**, so can be done multiple times without effort
- **Combining** data from multiple sources is straightforward
- **Data Volumes are small and manageable**, data usually comes from a **single type of data source**, **data refresh** is never a problem

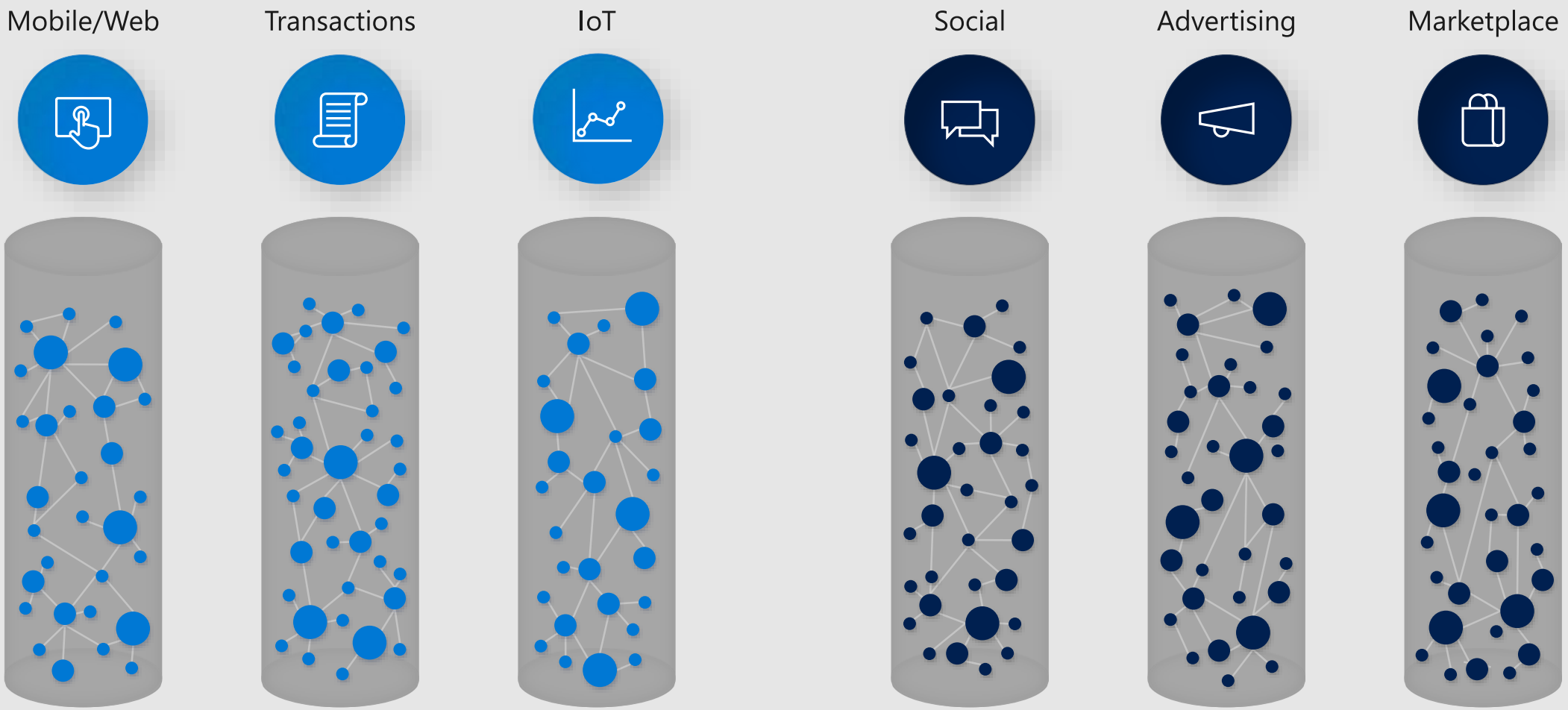
In Reality…

- **Finding** & **Connecting** to data is too difficult
- Experiences for data connectivity are too fragmented
- Data often needs to be **reshaped** before consumption
- Any shaping is one-off and not **repeatable**
- **Combining** data from multiple sources is difficult
- Volume, Velocity & Variety

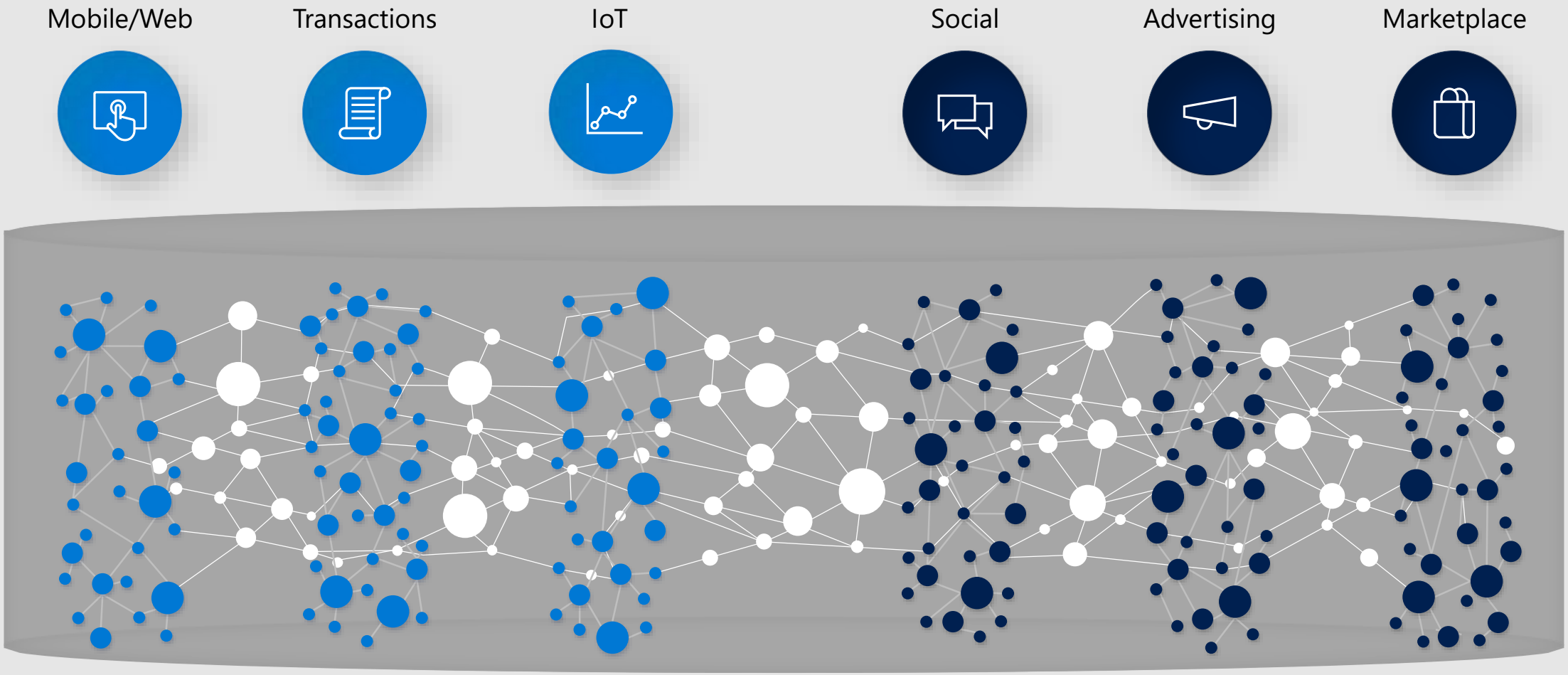
"Analysts spend up to 80% of their time on data preparation delaying the time to analysis and decision making."

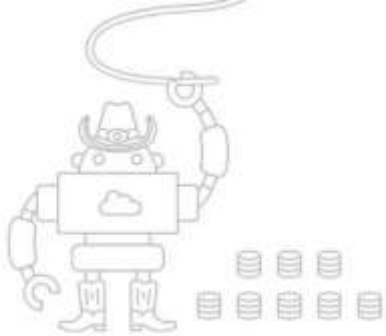
- Gartner

Data is stored within internal and external silos



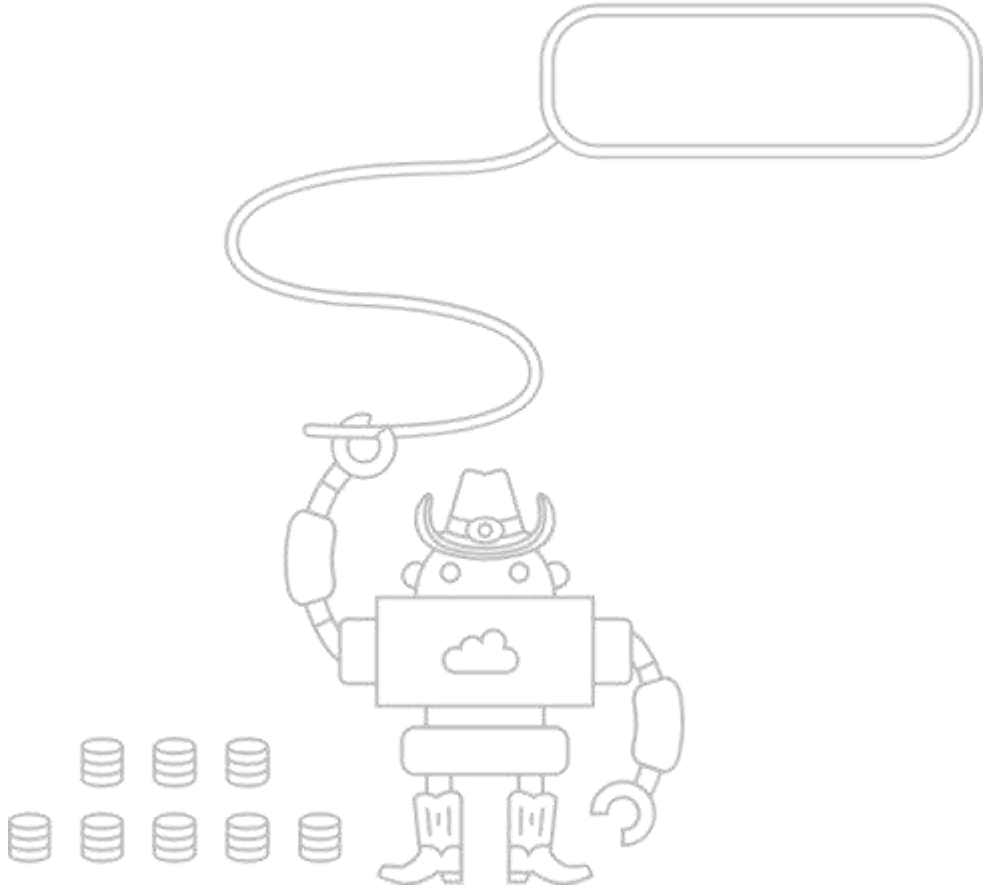
Need a shared data model with rich semantics





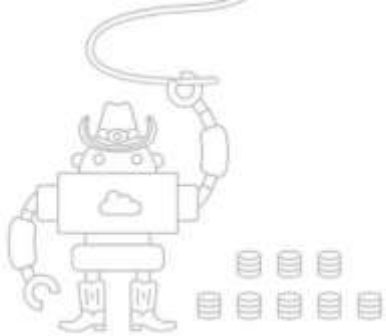
Agenda

- Wrangling in Context
- **Refresher Power Query & M**
- ADF Data Wrangling
- The swotty Power Cousin
- Resumé



DEMO

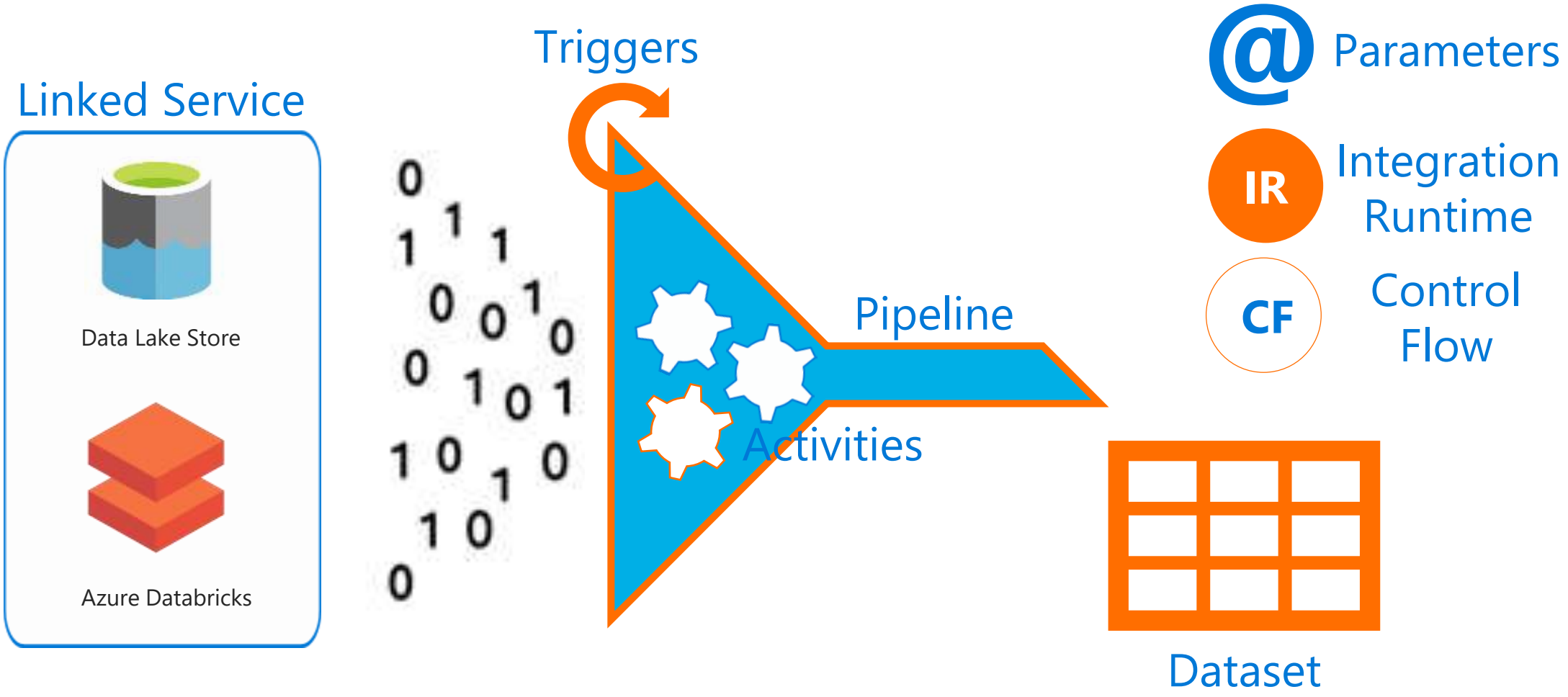
Power Query & M



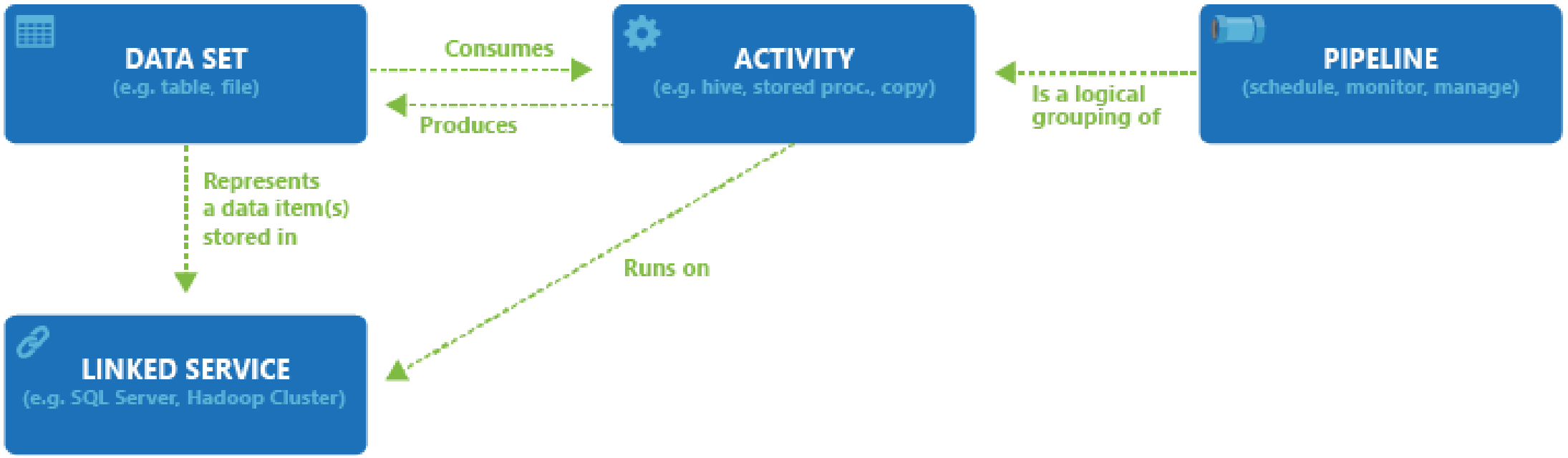
Agenda

- Wrangling in Context
- Refresher Power Query & M
- **ADF Data Wrangling**
- The swotty Power Cousin
- Resumé

Azure Data Factory Components



Component dependencies

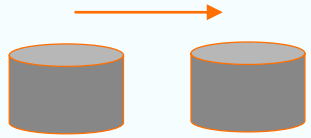


Methods for transforming in Azure Data Factory

Compute resources

SSIS Packages

Data Flow



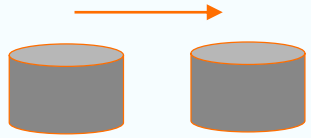
Methods for transforming data in Azure Data Factory

Code free data transformation at scale

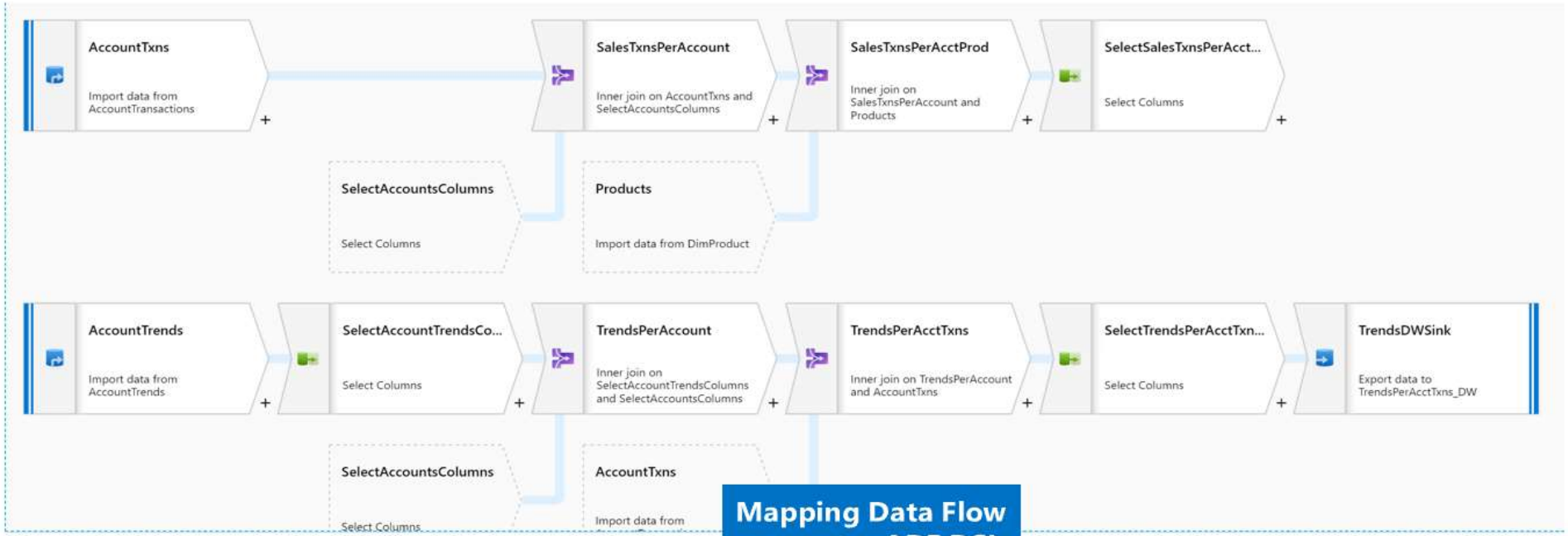
Compute resources

SSIS Packages

Data Flow



Mapping Data Flow: High Level Architecture



Mapping Data Flow generates ADF DSL

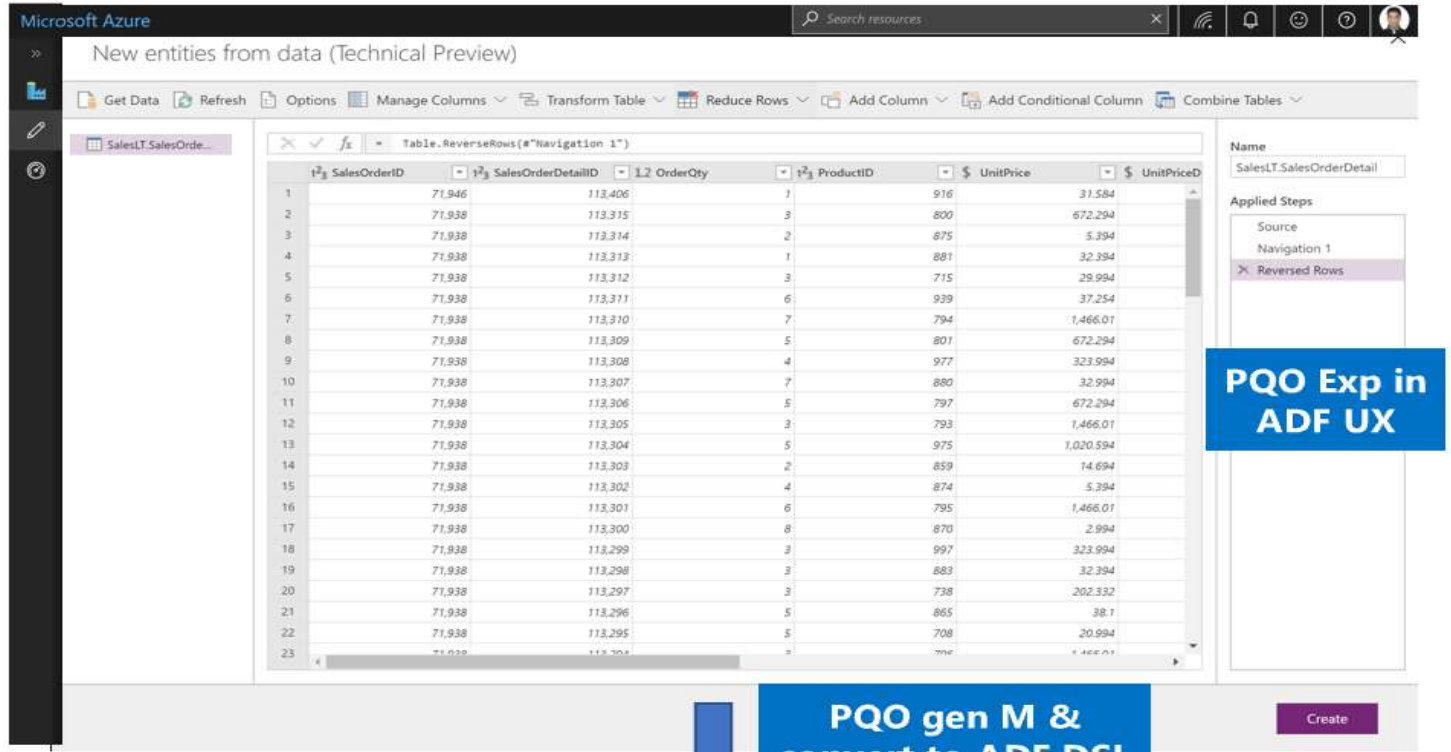
ADF DSL->Spark

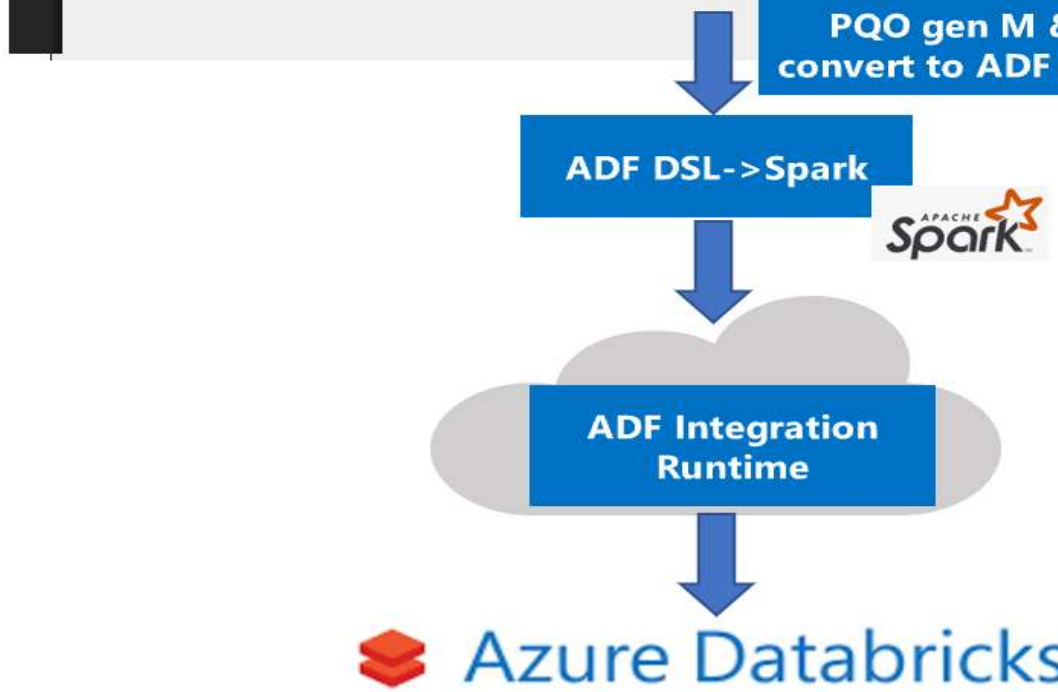
Apache Spark

ADF Integration Runtime

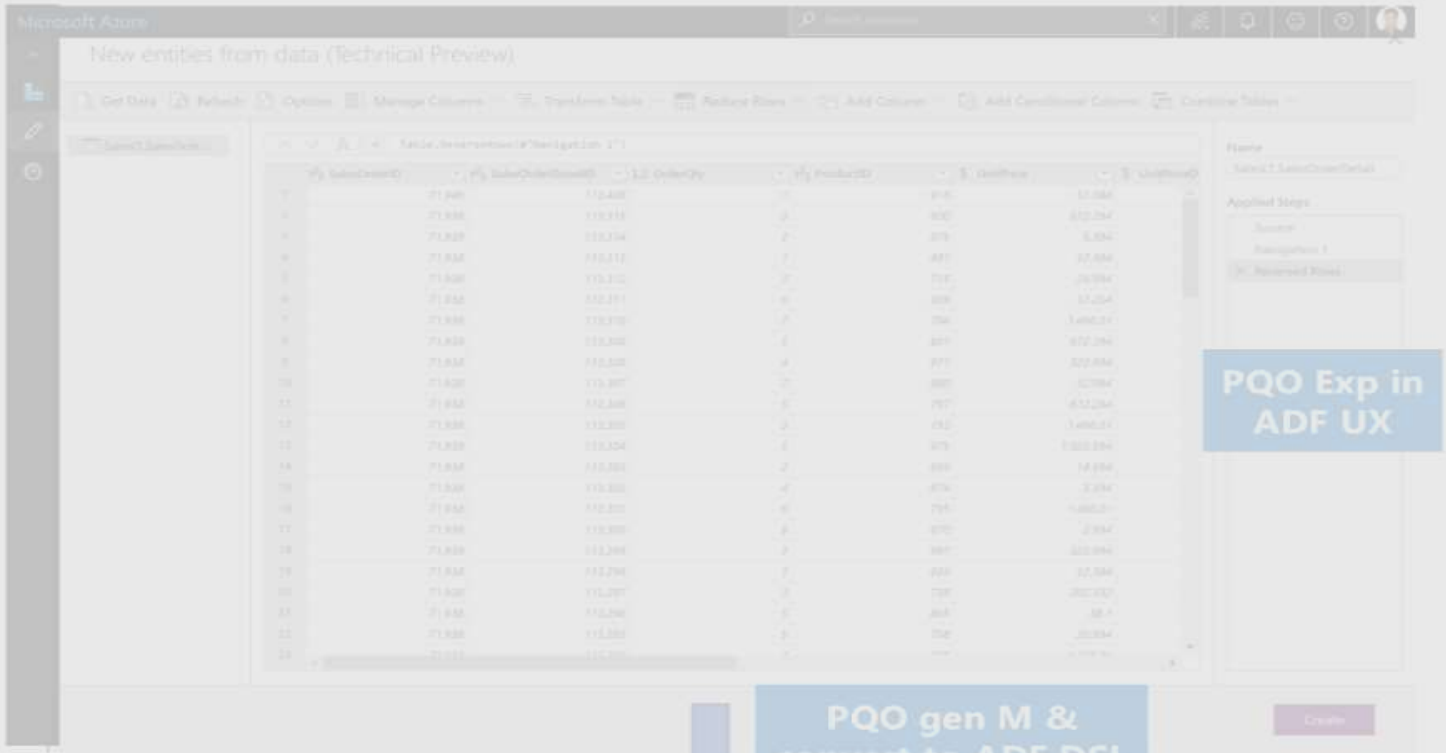
Azure Databricks

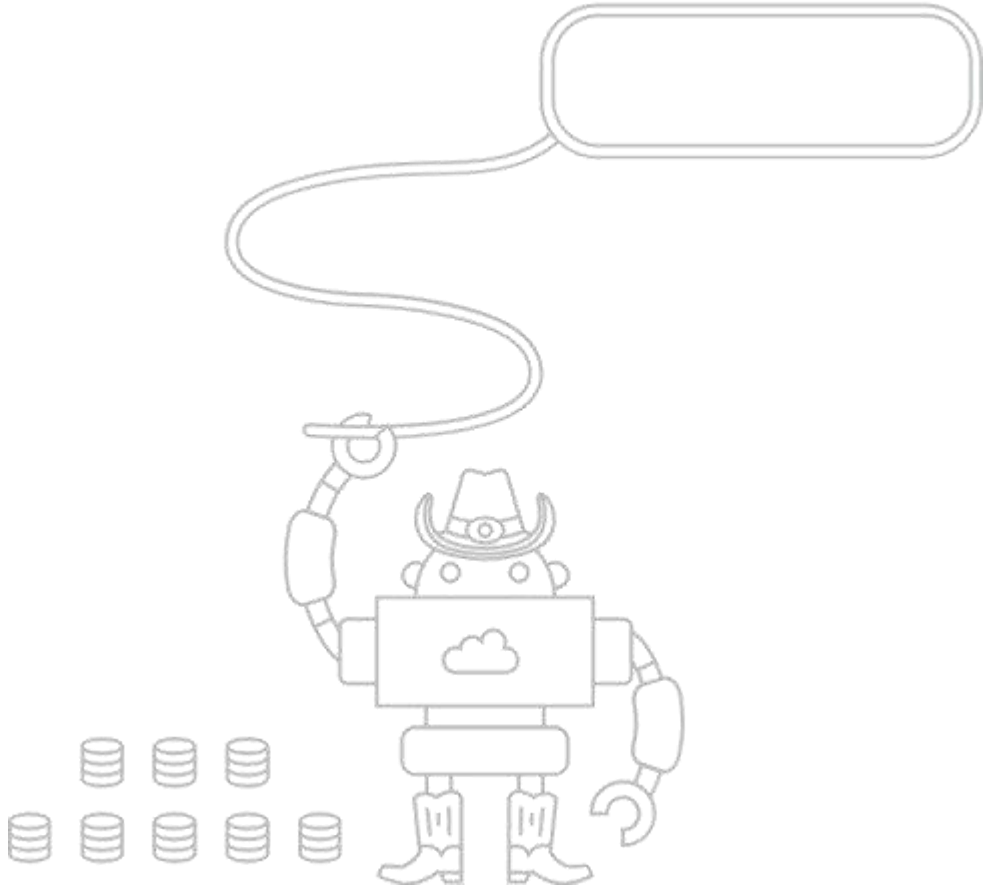
Wrangling Data Flow: High Level Architecture





Wrangling Data Flow: High Level Architecture





Demo

ADF Wrangling Data Flows

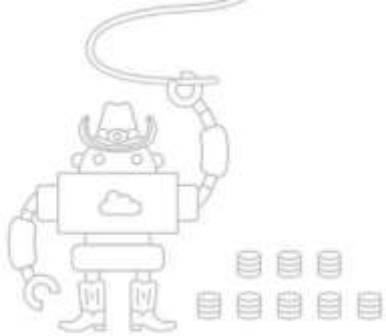
Not supportet (yet)

- Merge columns (can be achieved with AddColumn)
- Split column
- Append queries
- 'Use first row as headers' and 'Use headers as first row’

Details at https://docs.microsoft.com/en-us/azure/data-factory/wrangling-data-flow-functions

Data Sources

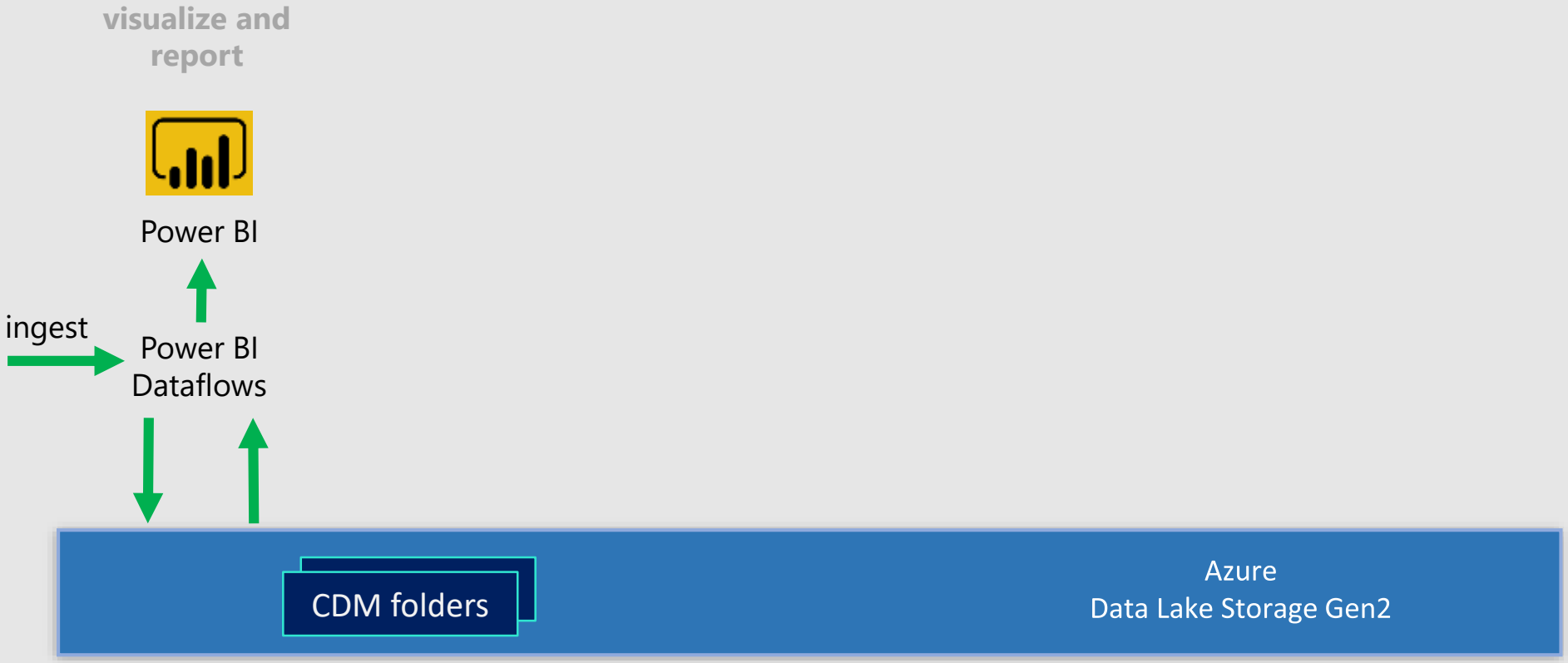
- ADL Gen2
- CSV
- Azure SQL Server



Agenda

- Wrangling in Context
- Refresher Power Query & M
- ADF Data Wrangling
- **The swotty Power Cousin**
- Resumé

The idea …

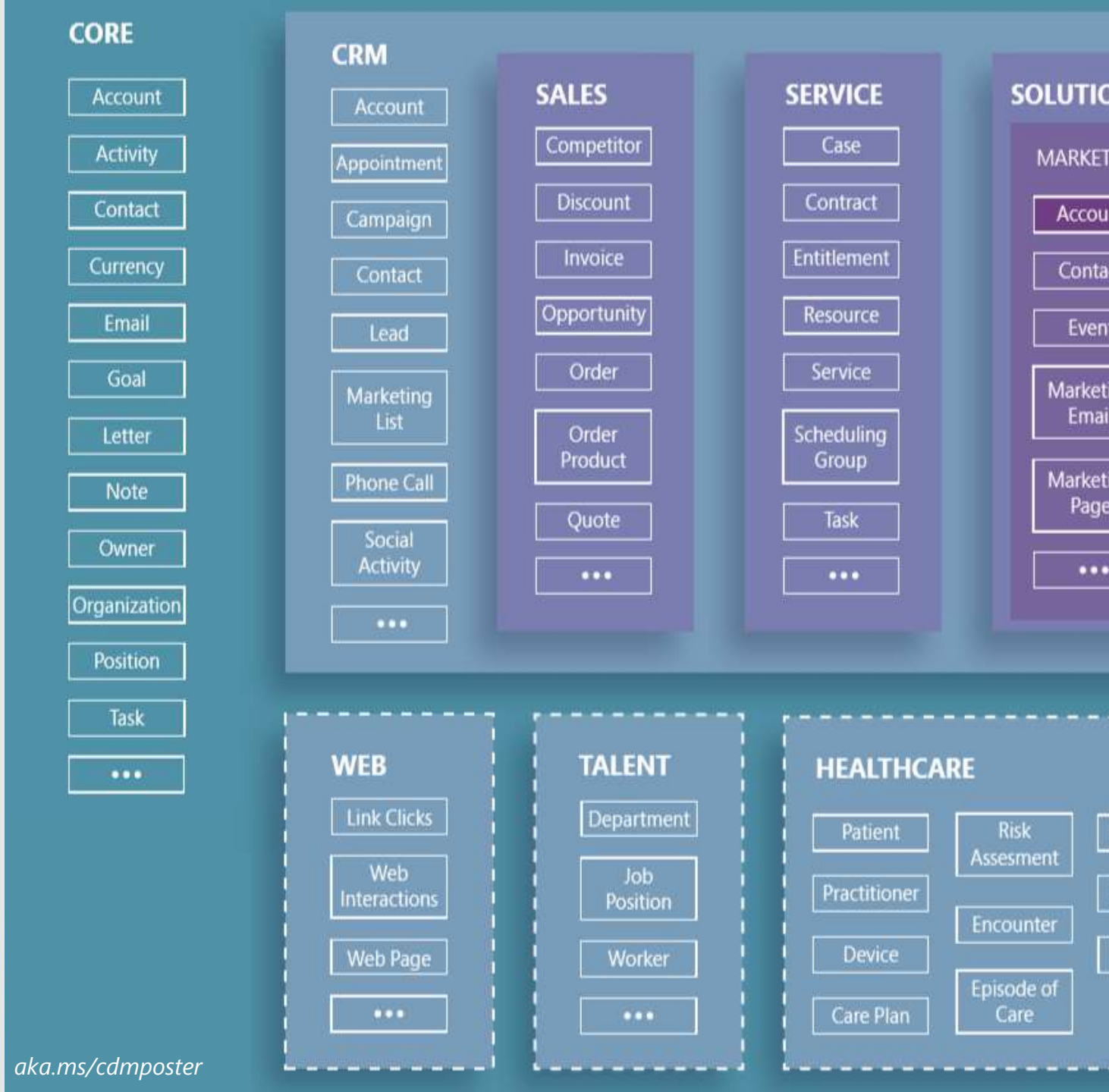

visualize and report
Power BI
ingest
Power BI Dataflows
CDM folders
Azure
Data Lake Storage Gen2

Common Data Model (CDM)

The **Common Data Model (CDM)** provides modular, and extensible business entities (Account, Lead, Opportunity, etc.)

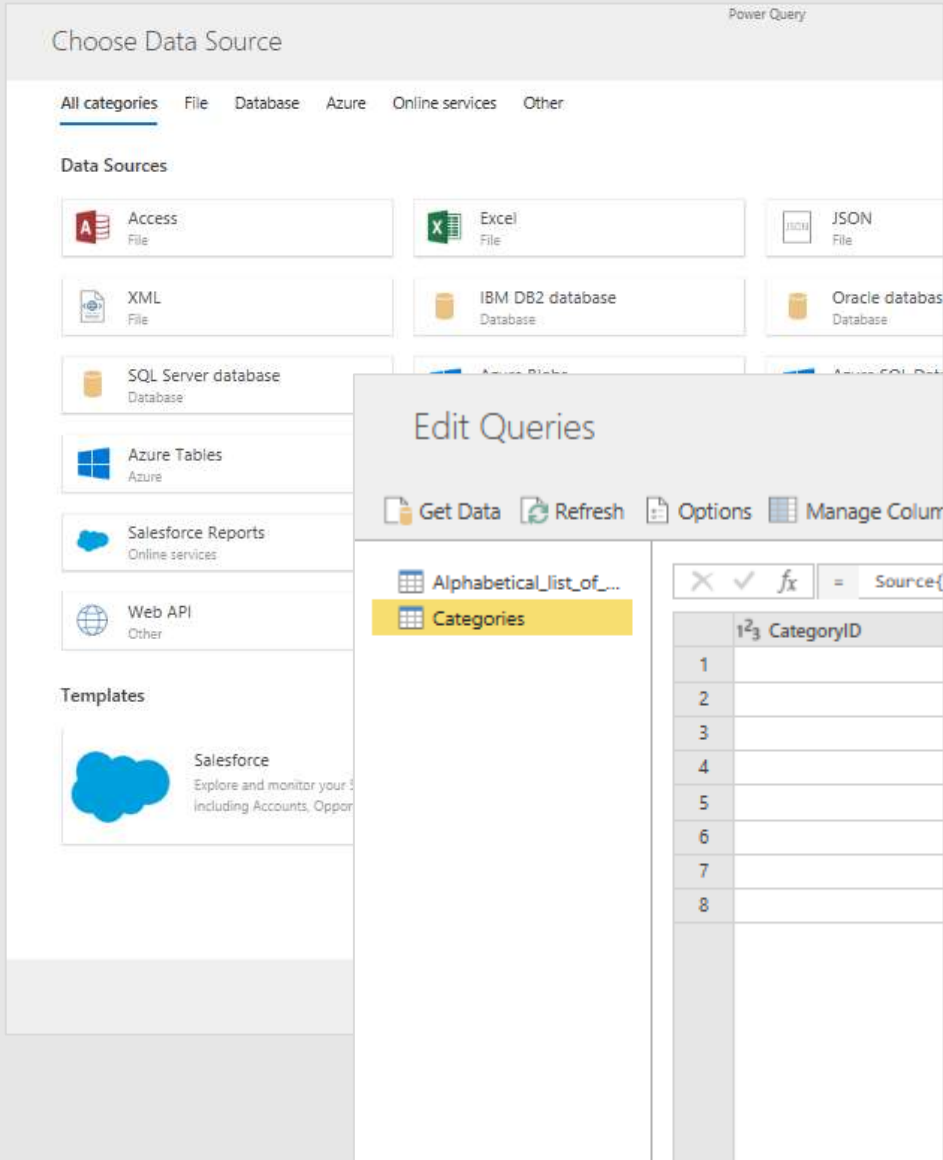
It **unifies data in a well-known schema** with semantic consistency across applications and deployments.

Customers, system integrators, and ISVs can develop turnkey business and intelligence applications, that **share data** based on these CDM entities. They can also extend entities to capture additional business-specific ideas.



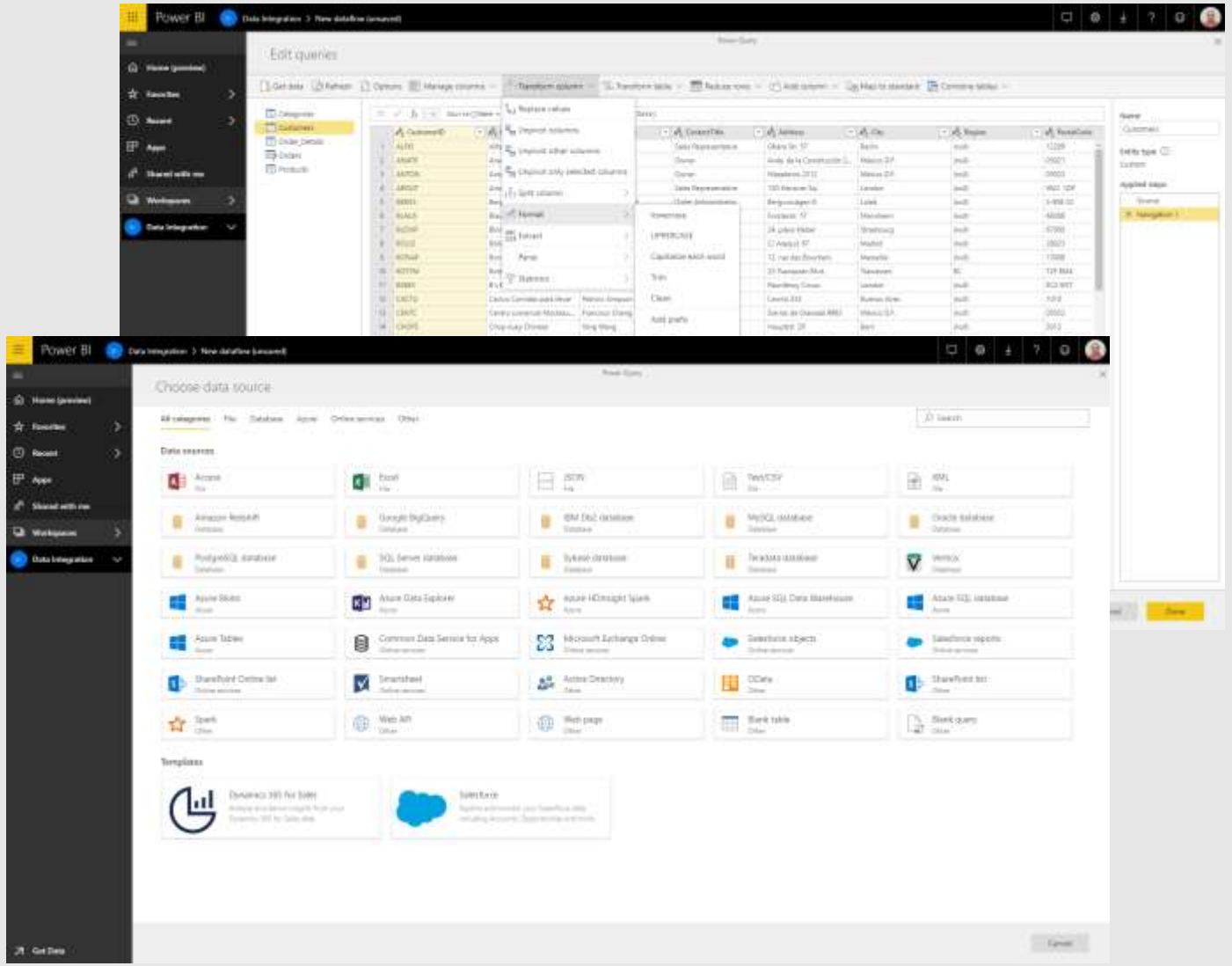
CDM in dataflows

- Dataflows allow customers to bring in data from a variety of disparate sources
- Leveraging **Power Query Online**, same technology in Excel and Power BI, visually transform and combine data sources in a no-code/low-code experience
- **Map** the resulting data **to a standard Common Data Model entity**, **or create a new entity** that can be related back to the standard



Dataflows use Power Query Online

- Power Query available as **web-based self-service data prep experience**
- Supports the same set of **300+ transformations** as desktop Power Query (M Engine parity level)
- Currently **~45 connectors**, including cloud & on-prem data sources via the **On-premises data gateway**



Recently Shipped Features – November 2019

New Data Connectors

- PDF Files
- Local Folder
- SharePoint Folder
- Google BigQuery
- Teradata
- MySQL
- PostgreSQL
- ODBC
- HDInsight Spark
- Apache Spark
- Impala

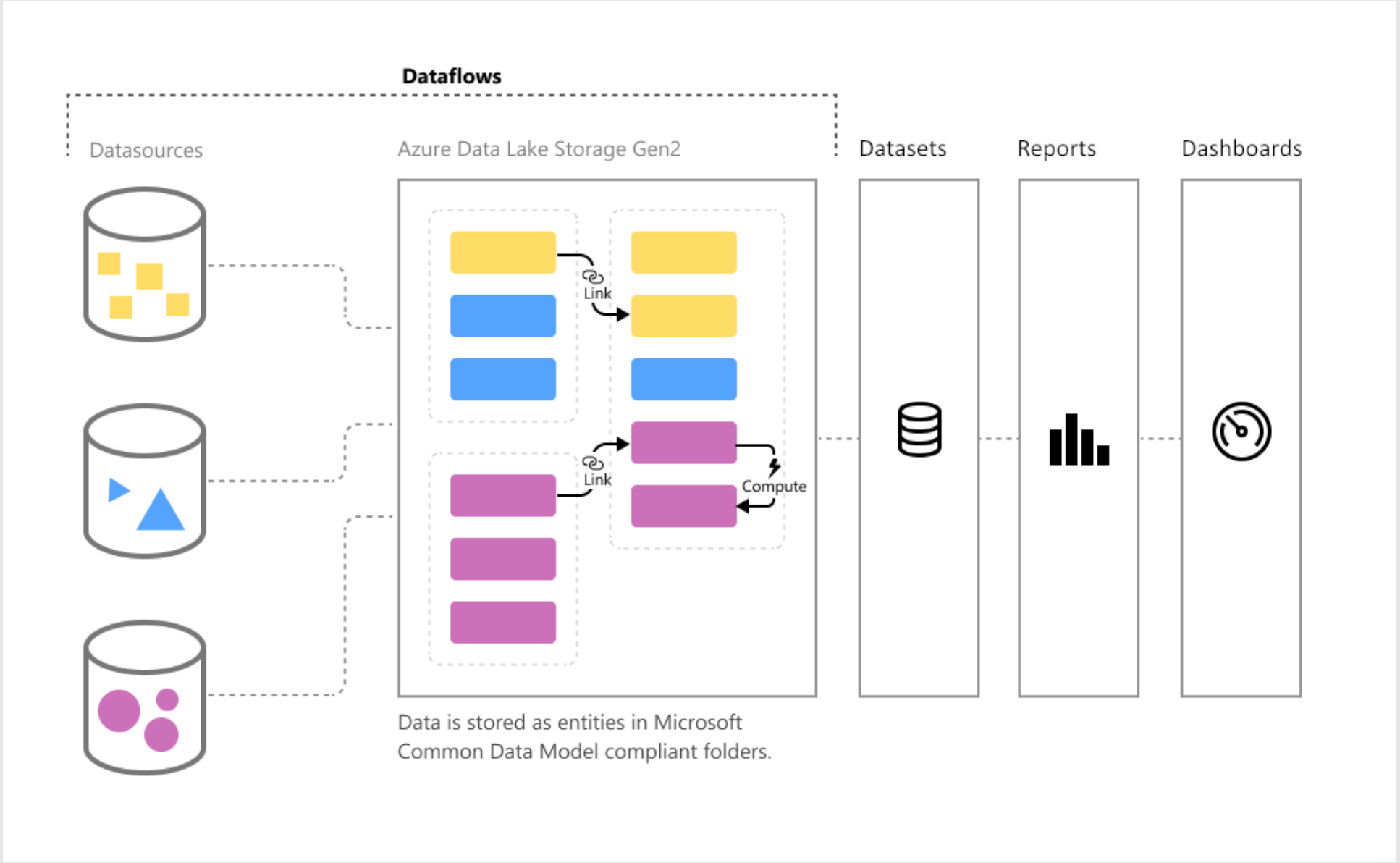
New Data Transformations

- Combine Files UX
- Merge Queries – Visual join kind selection
- Additional Number/Date/DateTime/Duration transformations UX
- List Transforms: Statistics, Sort, Keep/Remove/Reverse items
- Fill Up/Down
- Move Columns left/right/beginning/end
- Replace Errors

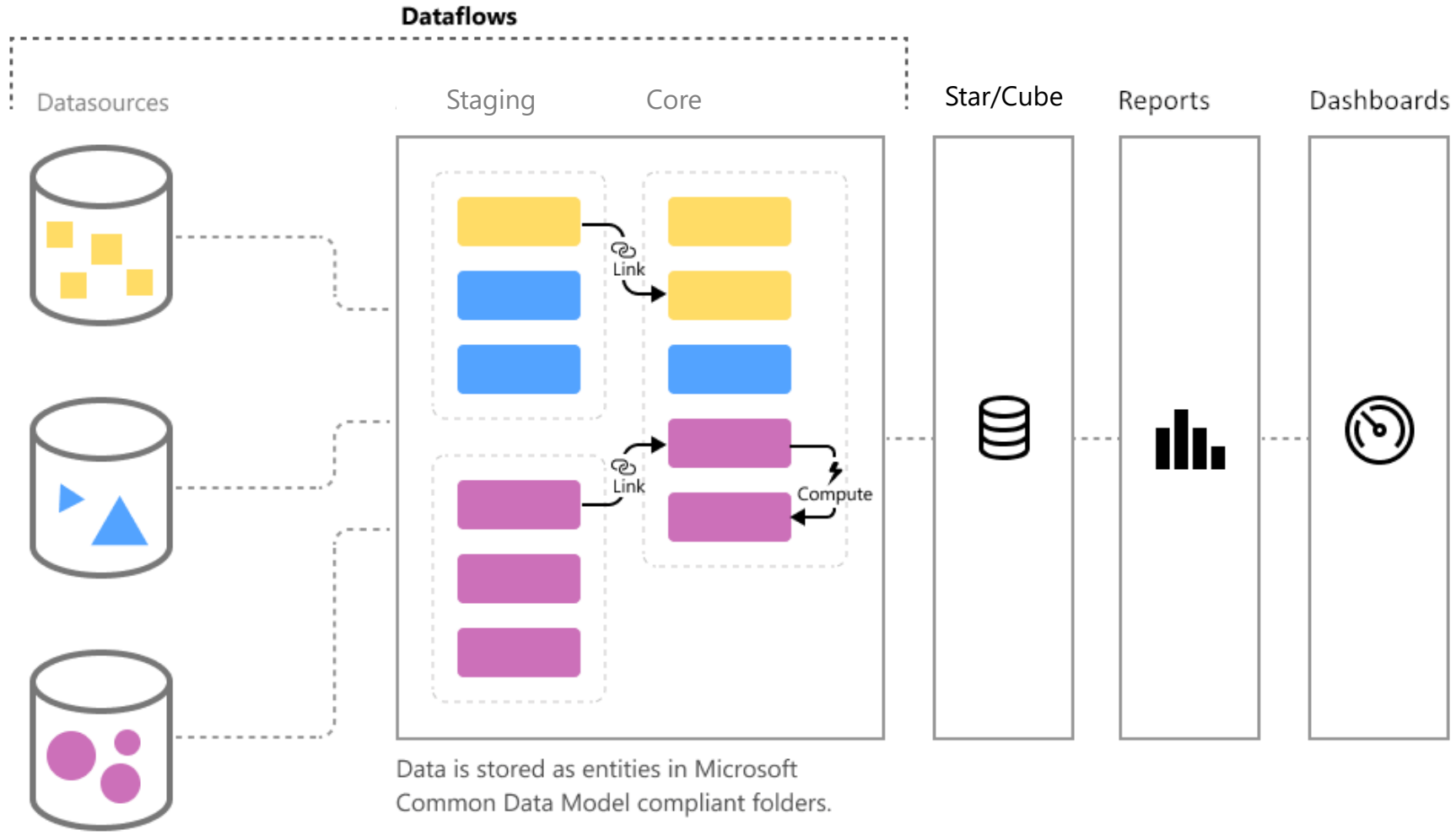
Other Enhancements

- Data Profiling
- Query Parameters
- Function Authoring
- M Intellisense support in Advanced Query Editor & Formula Bar
- Select Related Tables as part of Get Data UX
- More descriptive error messages within Power Query Online

Dataflows are composable, just like Excel



Speaking old architecture …



Power BI dataflows “FAQ”

- A new capability for self-service data preparation in Power BI
- Delivered in a familiar Power Query experience
- Built on the foundation of Azure Data Lake Storage gen2
- Utilize the CDM folder format for data storage & enable mapping to CDM
- A tool for business users to drive data reuse without requiring IT involvement
- Enable Excel-like data lineage and orchestration

- NOT a replacement for datasets
- NOT a replacement for a data warehouse
- NOT a replacement for Azure Data Factory or SSIS
- NOT a Premium-only feature
- NOT an additional cost or fee
- NOT spelled with a space or any capital letters

Power BI dataflows “FAQ”

- A new capability for self-service data preparation in Power BI
- Delivered in a familiar Power Query experience
- Built on the foundation of Azure Data Lake Storage gen2
- Utilize the CDM folder format for data storage & enable mapping to CDM
- A tool for business users to drive data reuse without requiring IT involvement
- Enable Excel-like data lineage and orchestration

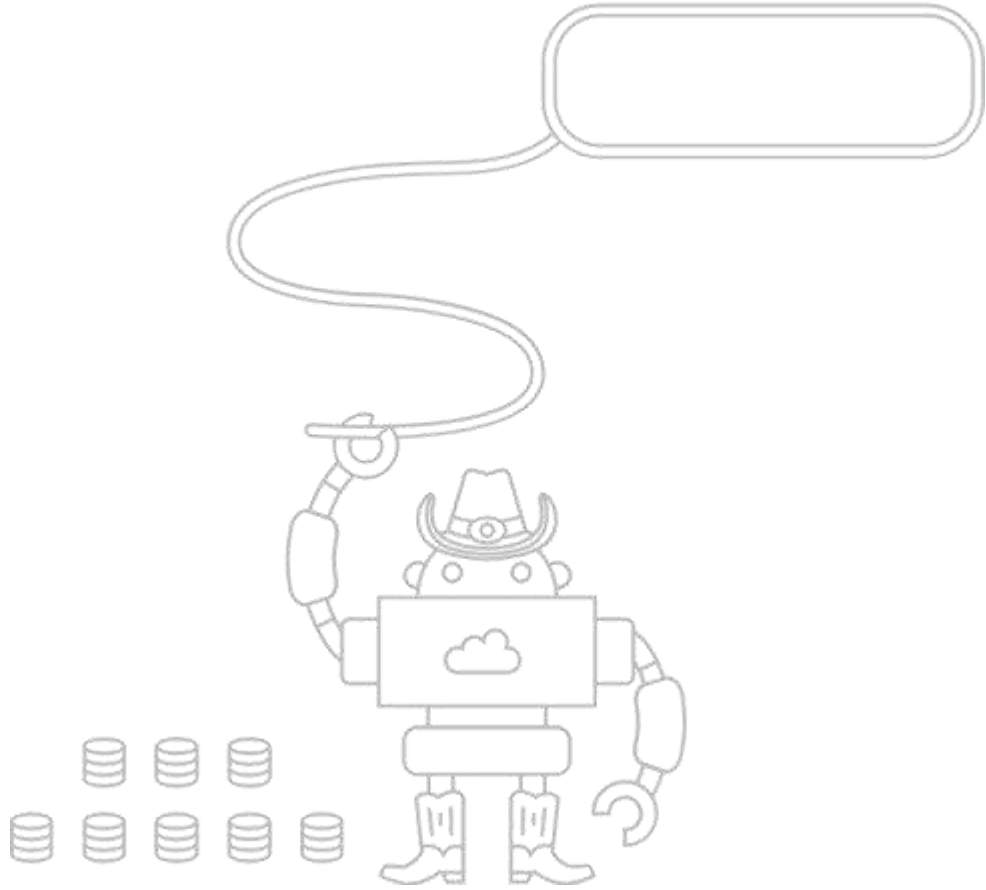
- NOT a replacement for datasets
- NOT a replacement for a data warehouse
- NOT a replacement for Azure Data Factory or SSIS
- NOT a Premium-only feature
- NOT an additional cost or fee
- NOT spelled with a space or any capital letters

Key takeaways from this session

Move your ETL to Dataflows and benefit from reusability and sharing

Try out the new Compute engine inside dataflows to greatly improve refresh times of your ETL Jobs

Have a deeper understanding of how Power BI and dataflows work



Demo

Power BI Dataflows

Use cases

- There are two distinct use cases for Dataflows and ADLS Gen2:

1. Prep data with Dataflows in Power BI, and all dataflow data will be stored in your organization's ADLS Gen2

    - Data will be stored in CDM Folder format.
    - Power BI, Azure Data services or your custom LOB applications can read the data from the lake

Using an organizational ADLSg2 resource with dataflows

Reports & dashboards

Datasets

Dataflows

Azure Data Factory

Azure Databricks

Azure SQL DW

Azure ML

Azure Data Lake Storage Gen2

CDM folder | CDM folder | CDM folder

Business analysts

Low/no code

Data scientists
Data engineers

Use cases

- There are two distinct use cases for Dataflows and ADLS Gen2:

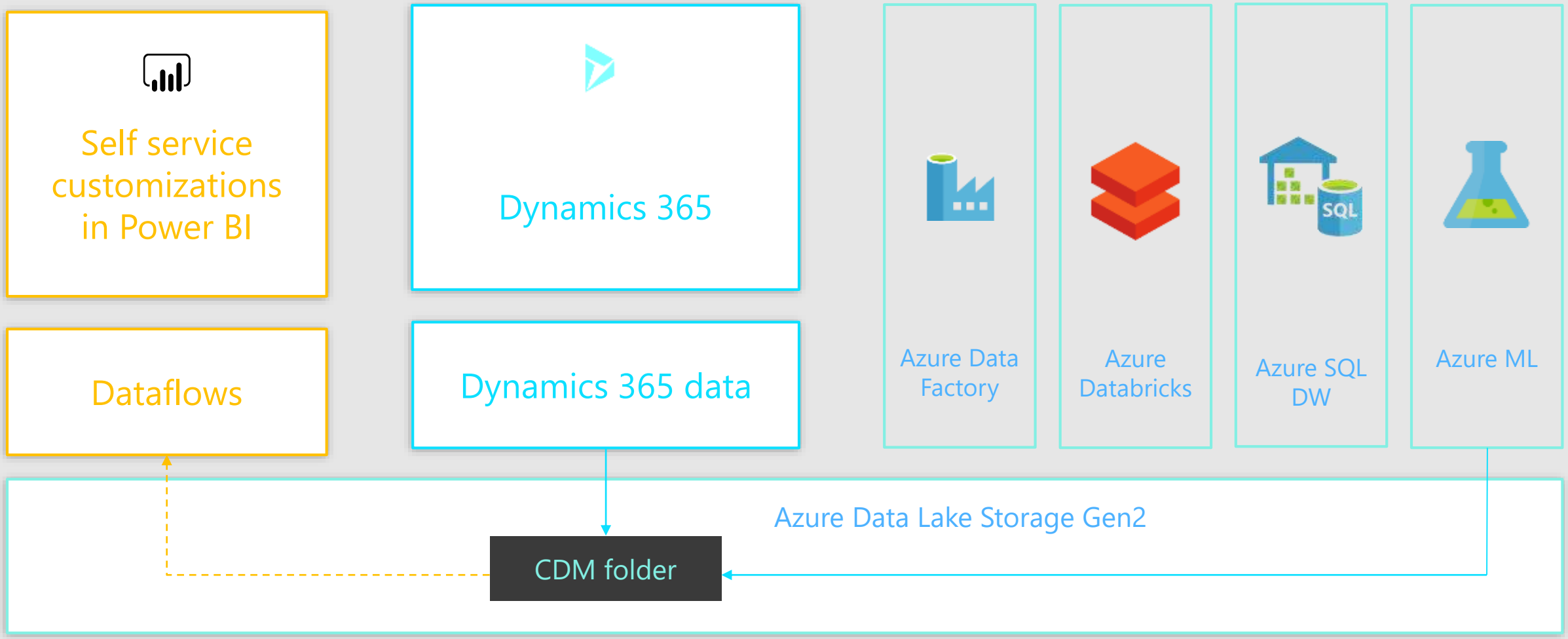
1. Prep data with Dataflows, and all dataflow data will be stored in your organization's ADLS Gen2

    - Data will be stored in CDM Folder format.
    - Power BI, Azure Data services or your custom LOB applications can read the data from the lake

2. Prep data in the lake, and add it as an external dataflow

    - Data stored in CDM Folder format, can easily be added as a dataflow
    - Business Analysts in Power BI Desktop can easily build reports and dashboards using the Dataflows connector

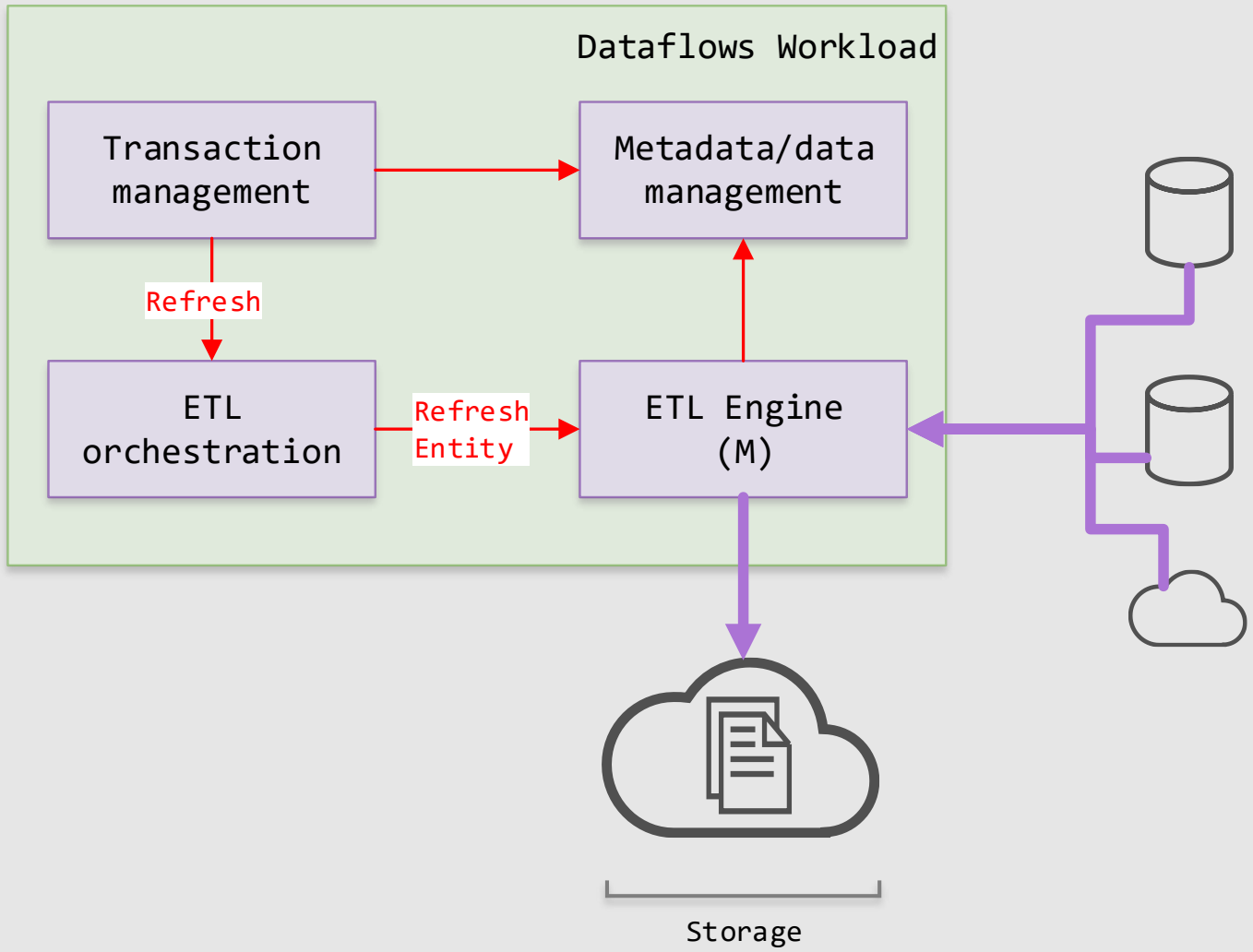
Deliver ready-made insights to Power BI users from Azure



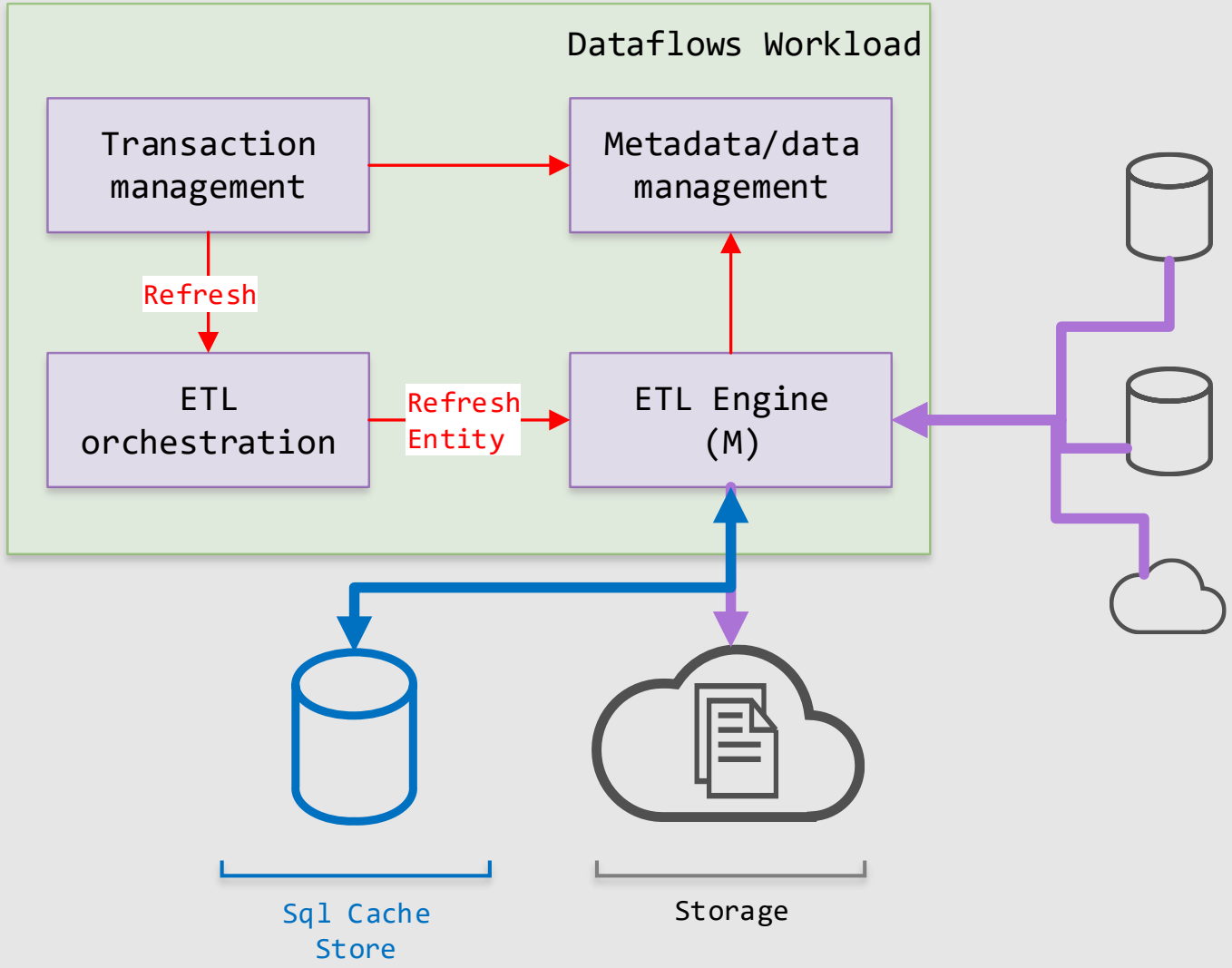
Feature	Pro	Premium
Create dataflows	✓	✓
Connect to dataflows from power bi desktop	✓	✓
Work with on prem and cloud sources	✓	✓
Use ADLS Gen 2 instead of built in storage	✓	✓
Use External CDM folders	✓	✓
Use linked entities (link to other dataflows)		✓
Use computed entities (link to other dataflows and compute over it)		✓
Enhanced compute engine for faster ETL		✓
Direct Query over dataflows		✓
Automated ML and Cognitive Services		✓

Feature	Pro	Premium
Create dataflows	✓	✓
Connect to dataflows from power bi desktop	✓	✓
Work with on prem and cloud sources	✓	✓
Use ADLS Gen 2 instead of built in storage	✓	✓
Use External CDM folders	✓	✓
Use linked entities (link to other dataflows)		✓
Use computed entities (link to other dataflows and compute over it)		✓
Enhanced compute engine for faster ETL		✓
Direct Query over dataflows		✓
Automated ML and Cognitive Services		✓

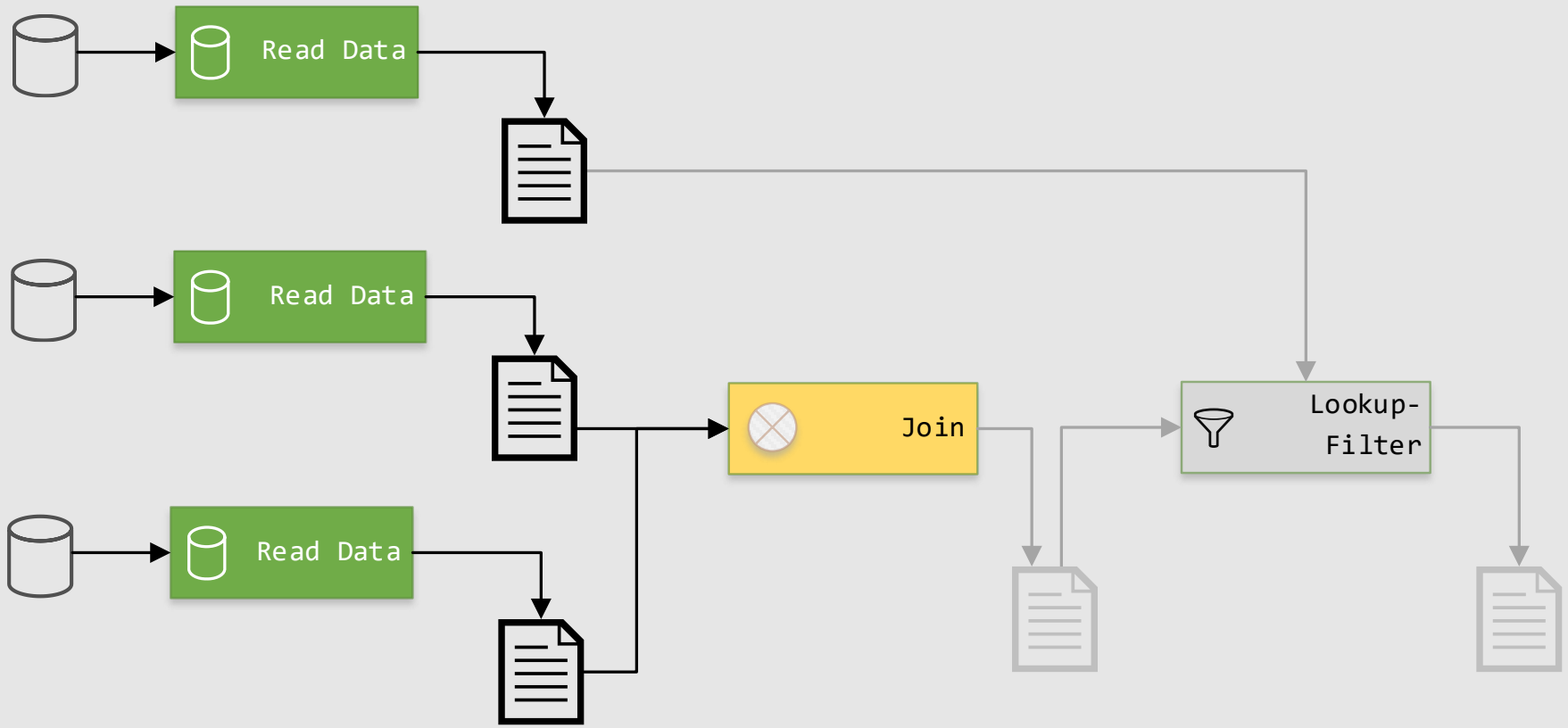
Current Dataflows Compute Engine (GA)



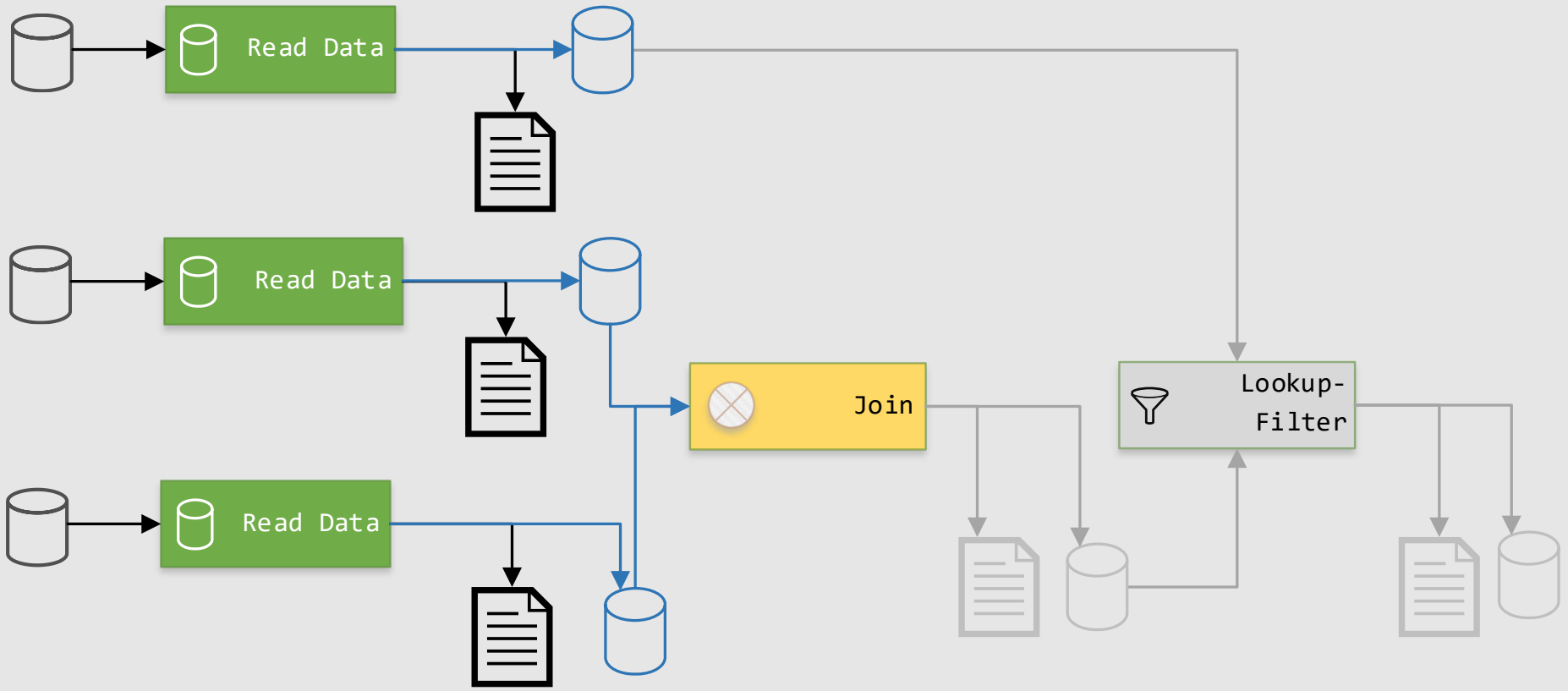
Enhanced Dataflows Compute Engine (Preview)

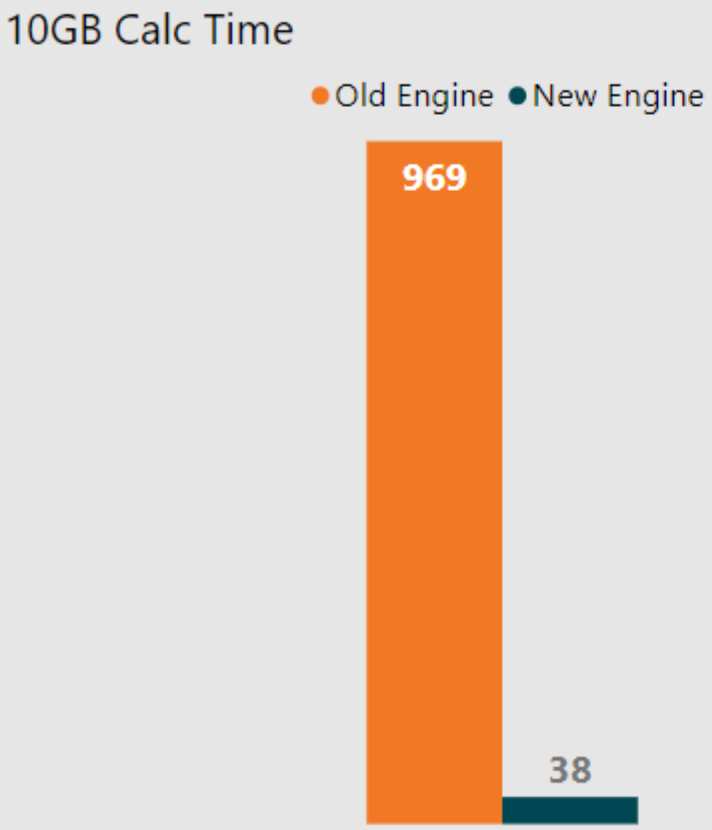


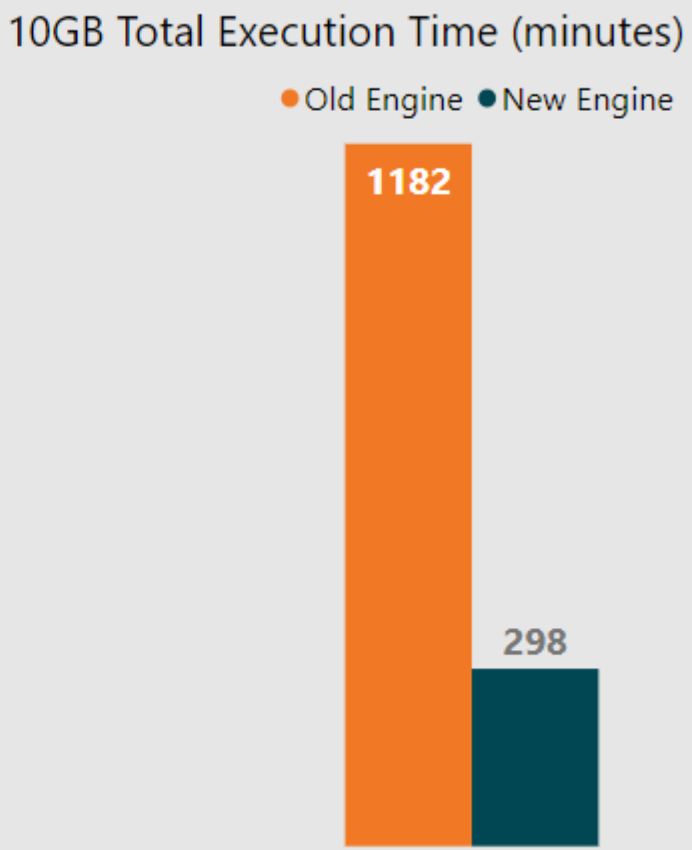
Current Dataflows Compute Engine (GA)



Enhanced Dataflows Compute Engine (Preview)

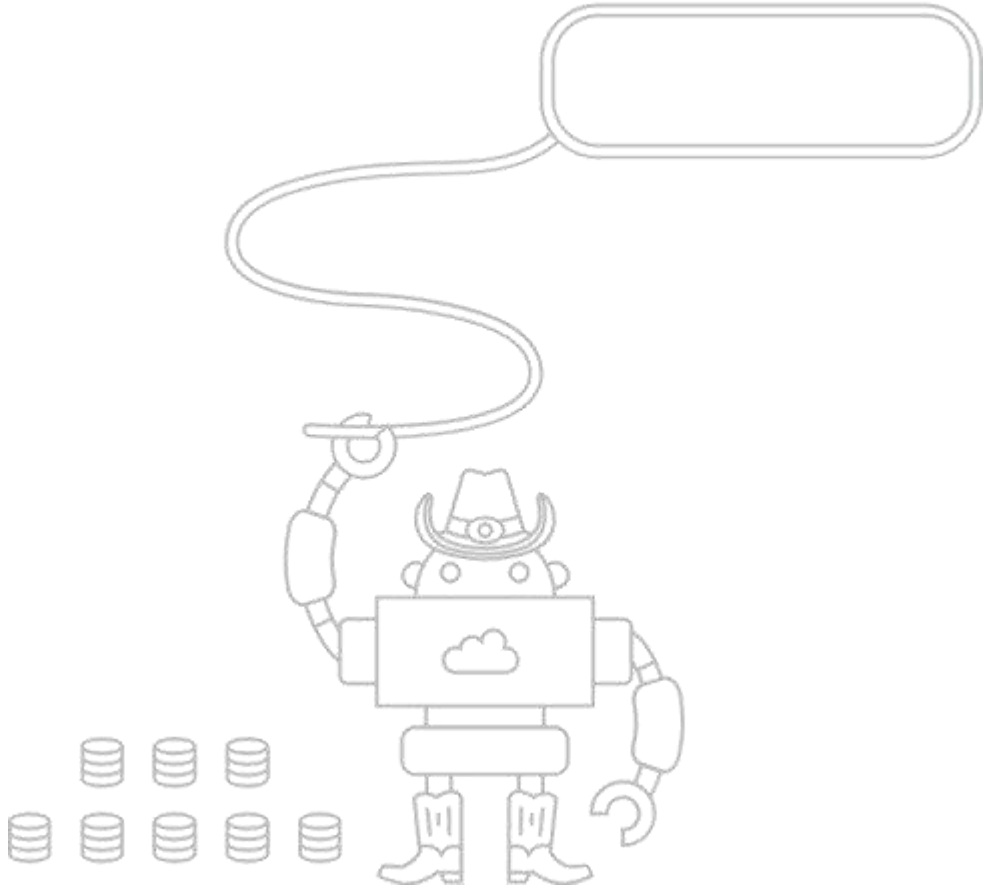






Up to 25X times faster

**Note : Loading data is slower as the data is being written twice*



Demo

Configure the Enhanced Compute Engine

Admin portal

- Usage metrics
- Users
- Audit logs
- Tenant settings
- Capacity settings
- Embed Codes
- Organizational visuals
- Dataflow settings
- Workspaces
- Custom branding

Power BI Premium | Power BI Embedded

CAPACITY NAME	CAPACITY ADMINS	ACTIONS	SKU	REGION	STATUS
[redacted]			A3	West Europe	Active

Set up new capacity in Azure

Admin portal

- Usage metrics
- Users
- Audit logs
- Tenant settings
- Capacity settings
- Embed Codes
- Organizational visuals
- Dataflow settings
- Workspaces
- Custom branding

West Europe

USER PERMISSIONS

▸ Users with assignment permissions

Disabled for the entire organization

MORE OPTIONS

▸ Workloads

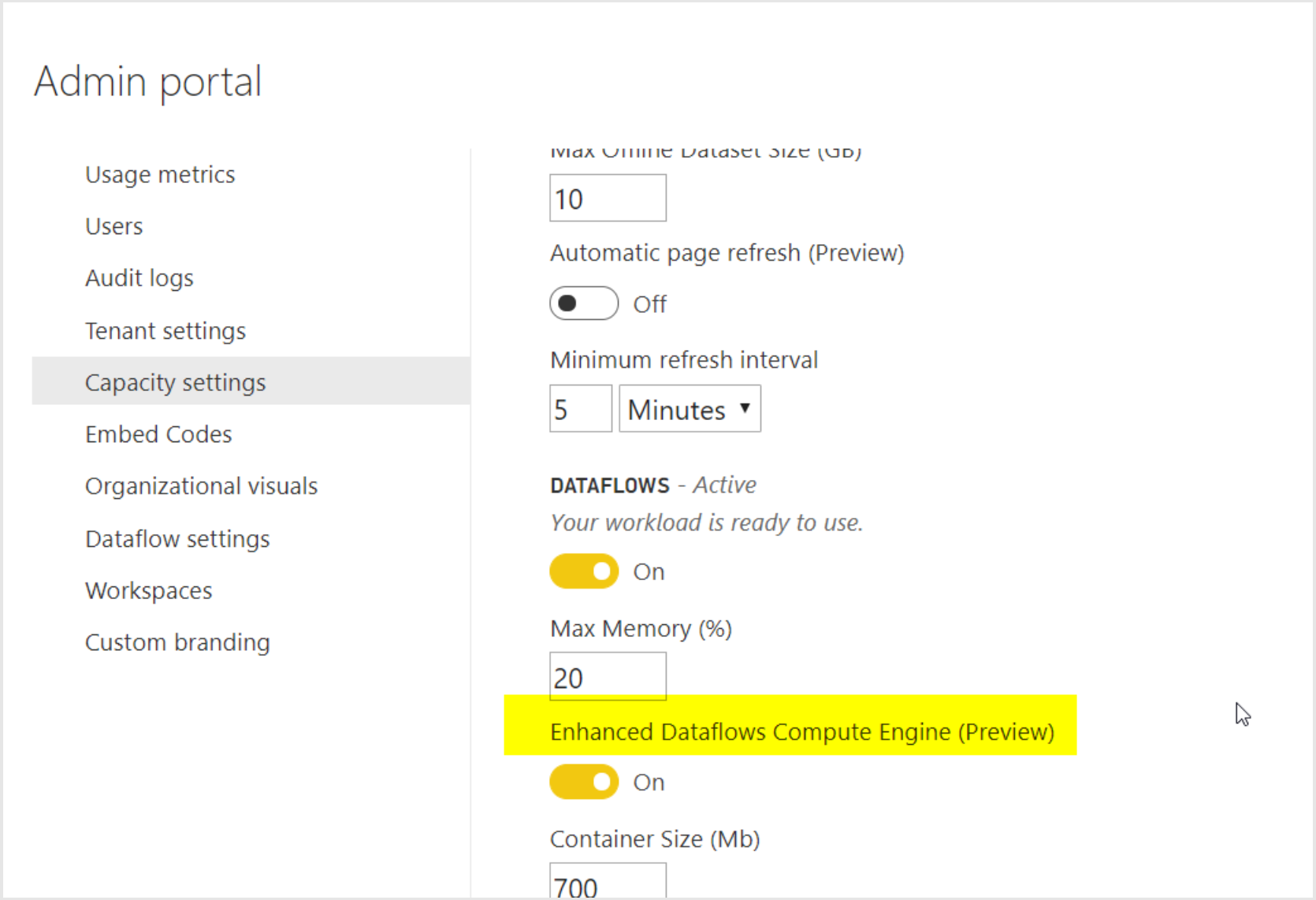
▸ Advanced options

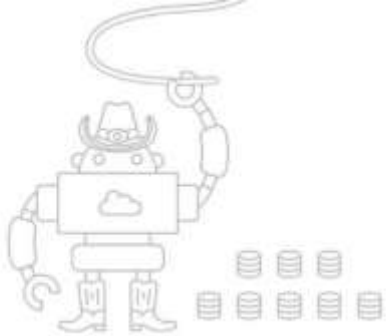
WORKSPACES (3)

Remove all Assign workspaces

Search content... Showing 3 items Sort Workspaces by...

WORKSPACE NAME	WORKSPACE ADMINS	ACTIONS	STATUS


Admin portal
Usage metrics
Users
Audit logs
Tenant settings
Capacity settings
Embed Codes
Organizational visuals
Dataflow settings
Workspaces
Custom branding
Max Offline Dataset Size (GB)
10
Automatic page refresh (Preview)
Off
Minimum refresh interval
5
Minutes
DATAFLOWS - Active
Your workload is ready to use.
On
Max Memory (%)
20
Enhanced Dataflows Compute Engine (Preview)
On
Container Size (Mb)
700



Agenda

- Wrangling in Context
- Refresher Power Query & M
- ADF Data Wrangling
- The swotty Power Cousin
- **Resumé**

Feature	ADF	PBI
Maturity		✓
Source Control	✓	
Modern DWH	✓	
Logical DWH		✓
Agility		✓
Time To Value		✓
Scalability (Data)	✓	
Performance		?
Connectivity		✓
Scalability (Development)	✓	

GET IN TOUCH

!

Christoph Seck
Chief Architect BI
Mail: c.seck@kigroup.de
Mobil: +49 151 12 11 10 95

KI performance GmbH
Mittelstraße 12-14
50672 Köln

KI performance