Leveraging Microsoft Fabric for Advanced Data Engineering Solutions

Philip Goldman

Field Sr. Data Solution Architect
EPAM

Introducing Microsoft Fabric

Deep Dive into Data Engineering

DEMO

Introducing Microsoft Fabric for Data Engineering

Microsoft Fabric

End-to-end analytics data fabric
From the data lake to the business user

Complete Analytics Platform

- Everything, unified
- Unified SaaS Solution
- Low Code Plus Pro Dev

Lake-centric and Open

- OneLake
- One Copy
- Always Synced

Empower Every Office User

- Familiar and Intuitive
- Built Into Office 365
- Insight to Action

Persistent Security and Governance

- End-to-End Visibility
- Always Governed
- Secure by Default

Introducing Microsoft Fabric for Data Engineering

- **Complete analytics platform**

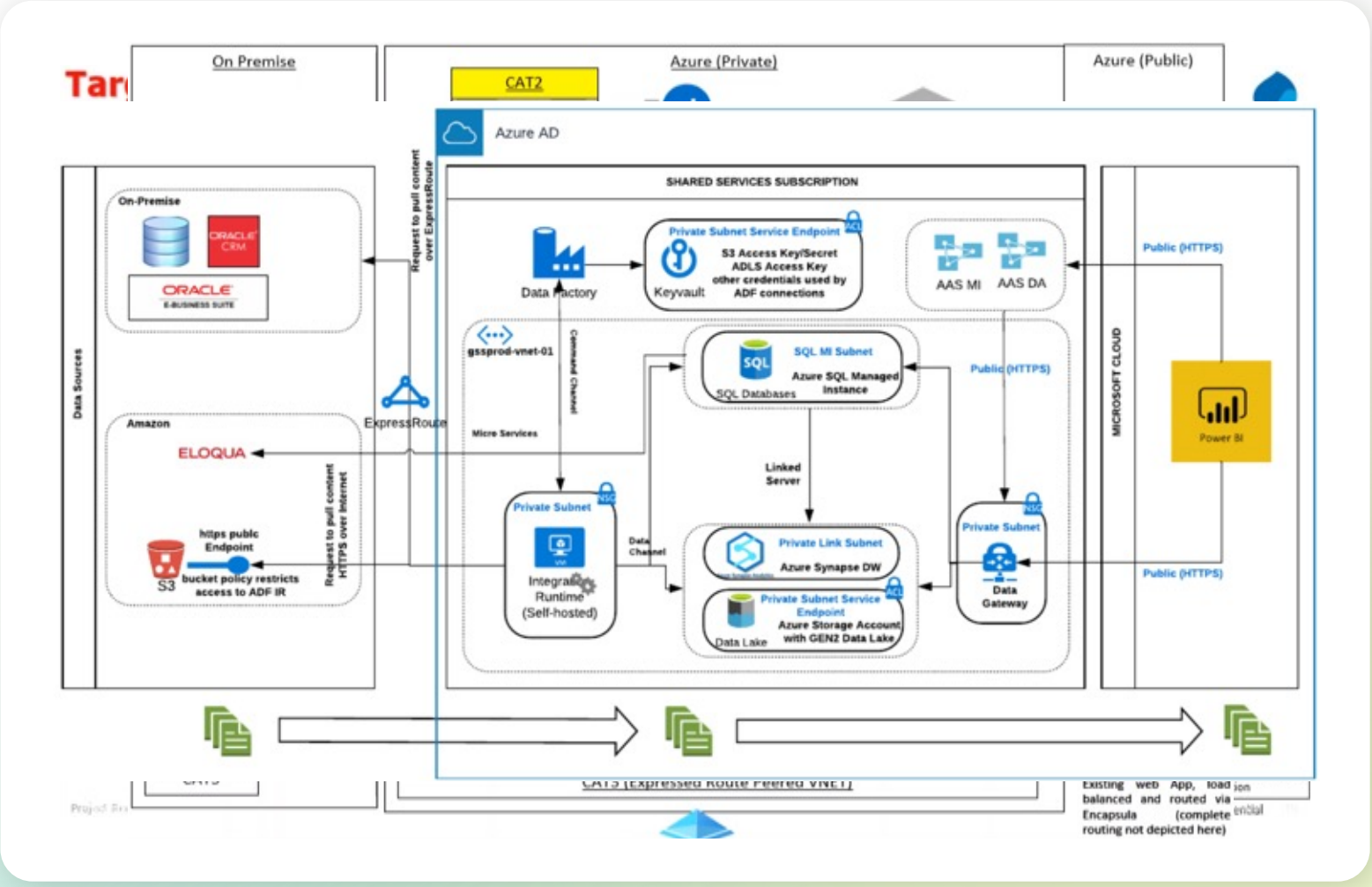
Scalable analytics are complex and fragmented

Every analytics project has many subsystems

Every subsystem need a different class of product

Products often come from multiple vendors

Integration at scale across products is complex, fragile, and expensive



Scalable analytics are complex and fragmented

Every analytics project has many subsystems

Every subsystem need a different class of product

Products often come from multiple vendors

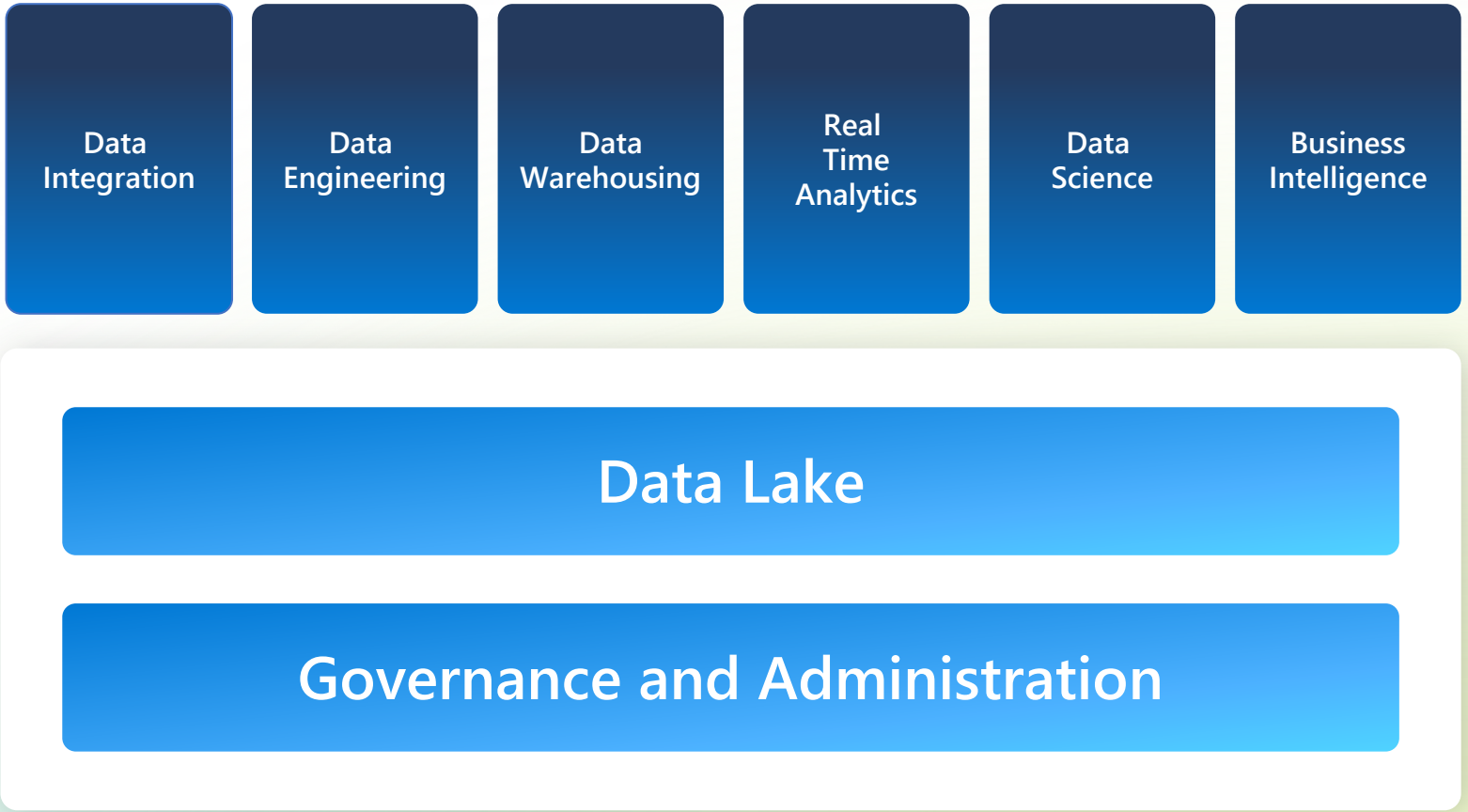
Integration at scale across products is complex, fragile and expensive

> **Simplify,**
> I am the Chief Data Officer and don't want to be the Chief Integration Officer."
>
> Every CDO, Every Enterprise

A silver lining?

Analytics systems have very predictable patterns

Microsoft has all the products with the right scale needed to build a complete analytics system



A silver lining?

Analytics systems have very predictable patterns

Microsoft has all the products with the right scale needed to build a complete analytics system

Still far too complex

- Many Products
- Different Experiences
- Proprietary and Open
- Dedicated and Serverless
- PaaS and SaaS
- Different Business Models
- Steep Learning Curves
- Deep Expertise Needed
- High Integration Effort

Power BI

Synapse

Kusto

Azure AI

Data Factory

Spark

Microsoft Fabric

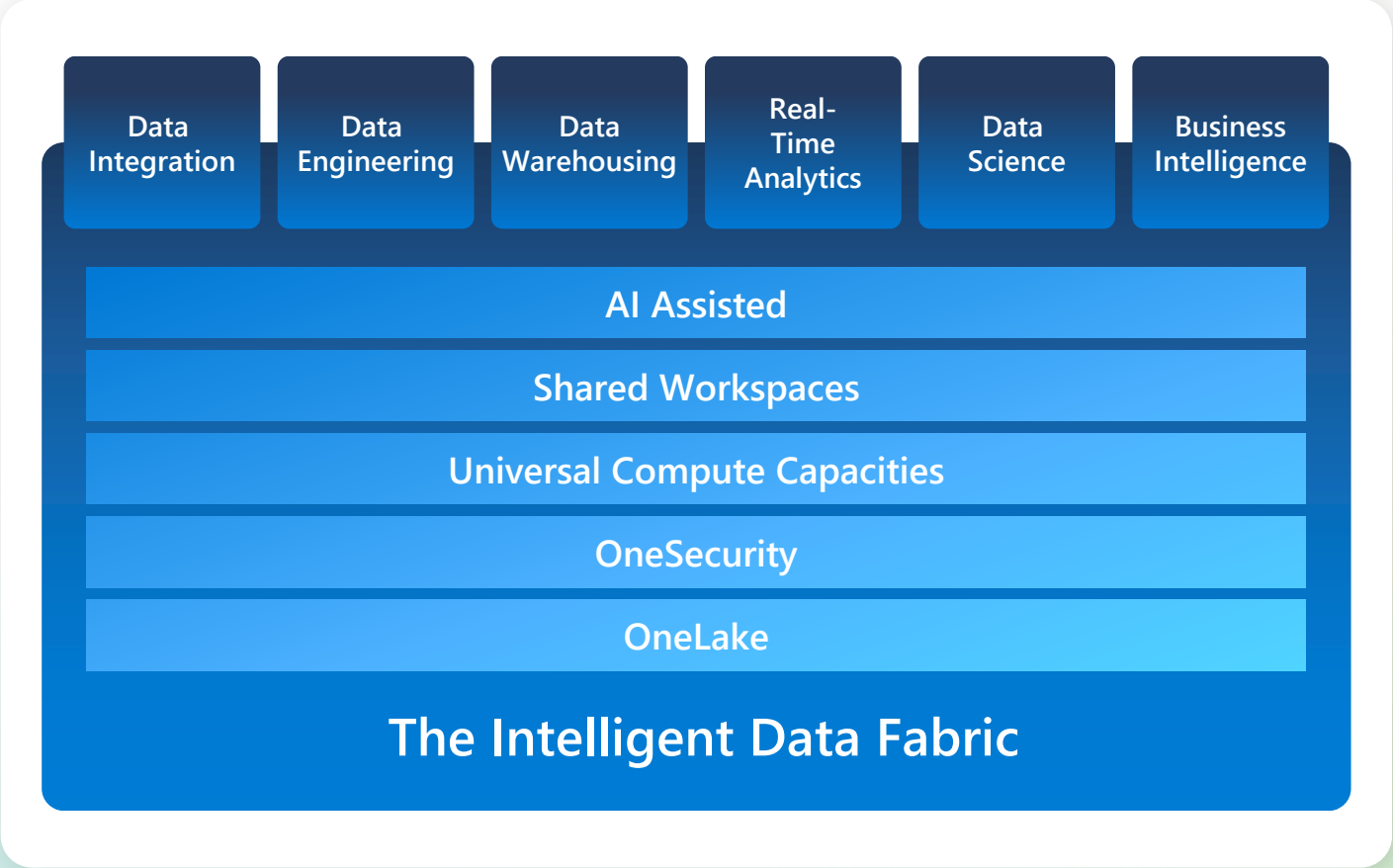
Data Integration | Data Lake | Spark Engines | Data Warehouse | Real Time Analytics | Data Science | Business Intelligence | Governance

Unified analytics fabric

End-to-end analytics data fabric

From the data lake to the business user

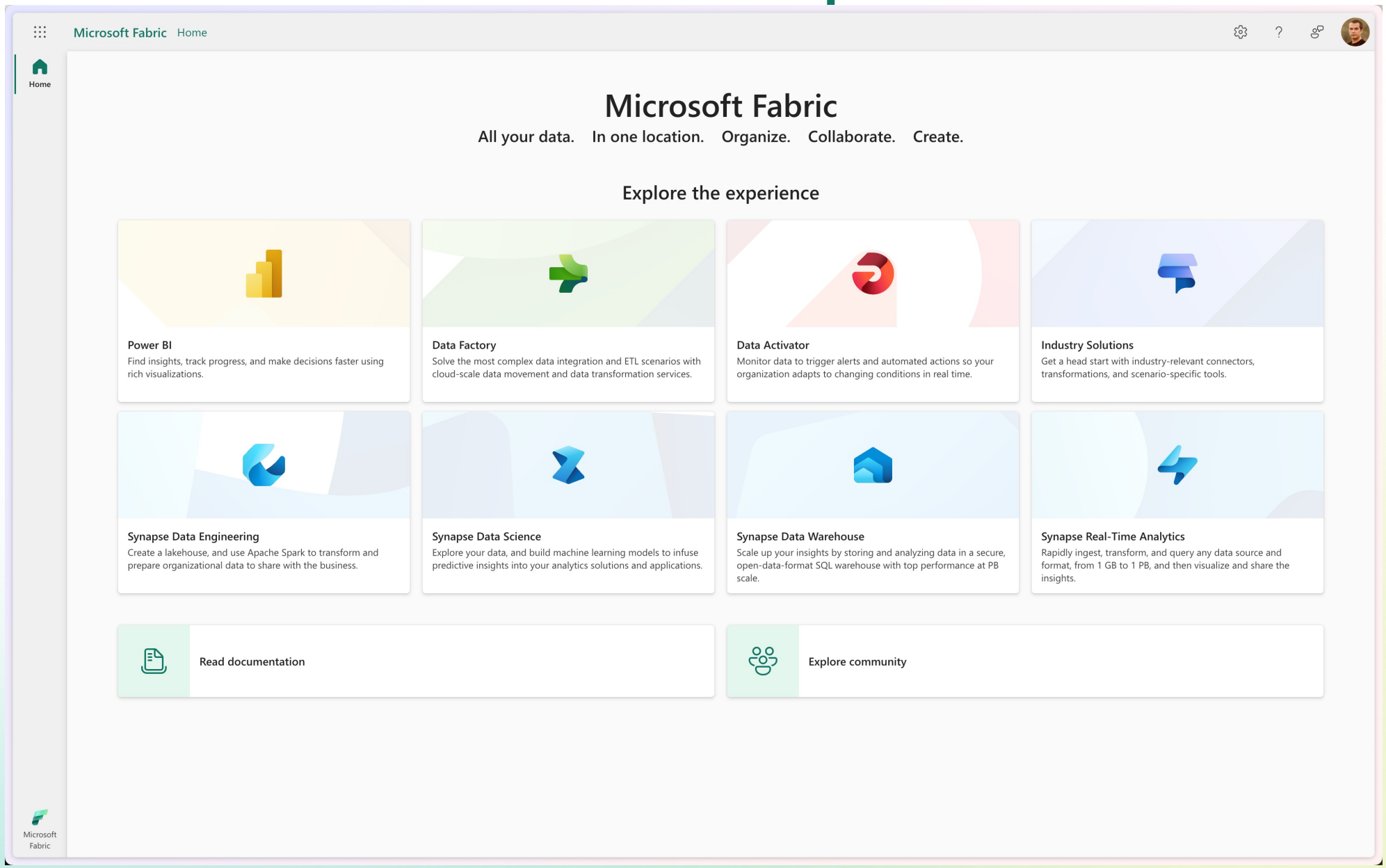
Microsoft Fabric



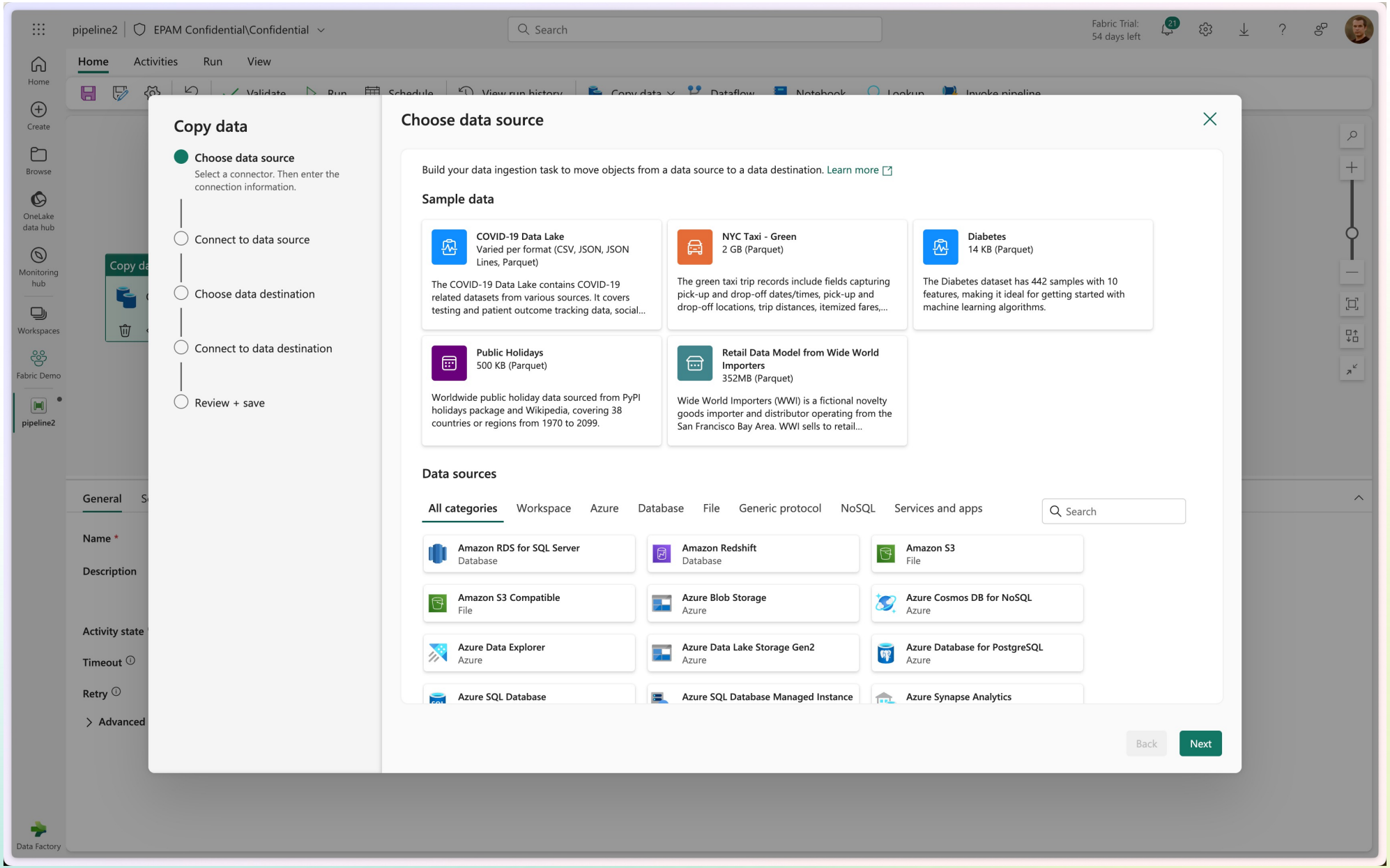
Single...

Onboarding and trials
Sign-on
Navigation model
UX model
Workspace organization
Collaboration experience
Data Lake
Storage format
Data copy for all engines
Security model
CI/CD
Monitoring hub
Data Hub
Governance & compliance

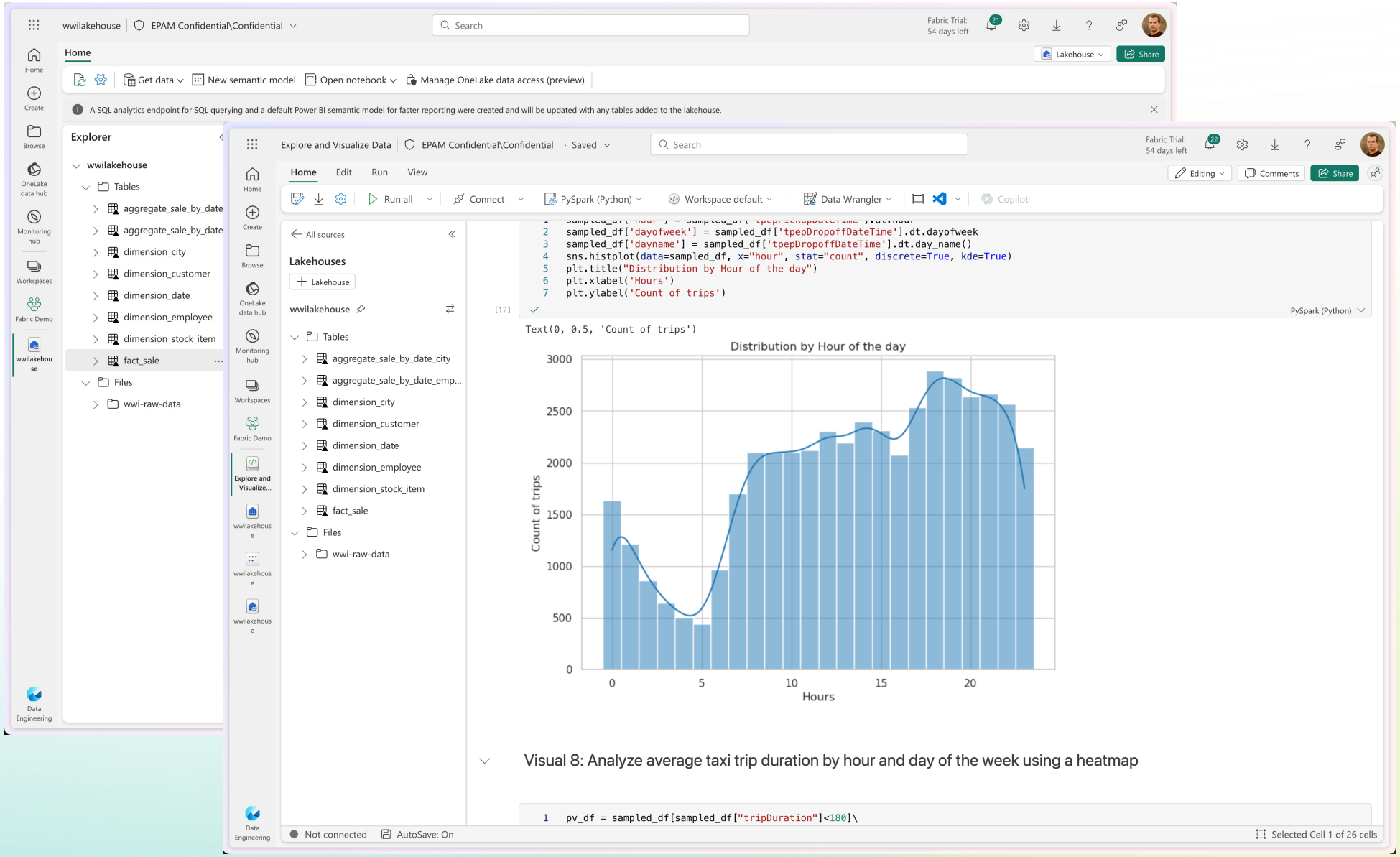
Persona Centric Experiences



Data Integration



Data Engineering



Data Warehouse

wwilakehouse | EPAM Confidential\Confidential

Search

Fabric Trial: 54 days left

Home

SQL analytics endpoint

New SQL query | New visual query

Explorer

Warehouses

- wwilakehouse
  - Schemas
    - dbo
      - Tables
        - aggregate_sal...
        - aggregate_sal...
        - dimension_city
        - dimension_cu...
        - dimension_da...
        - dimension_e...
        - dimension_st...
        - fact_sale
      - Views
      - Functions
      - Stored Procedur...
    - guest
    - INFORMATION_SCHE...
    - queryinsights
    - sys
  - Security
- Queries

SQL query 1

Run | Save as view

```
SELECT BuyingGroup, Count(*) AS Total
FROM dimension_customer
GROUP BY BuyingGroup
```

Messages | Results | Open in Excel | Explore this data (preview)

Search

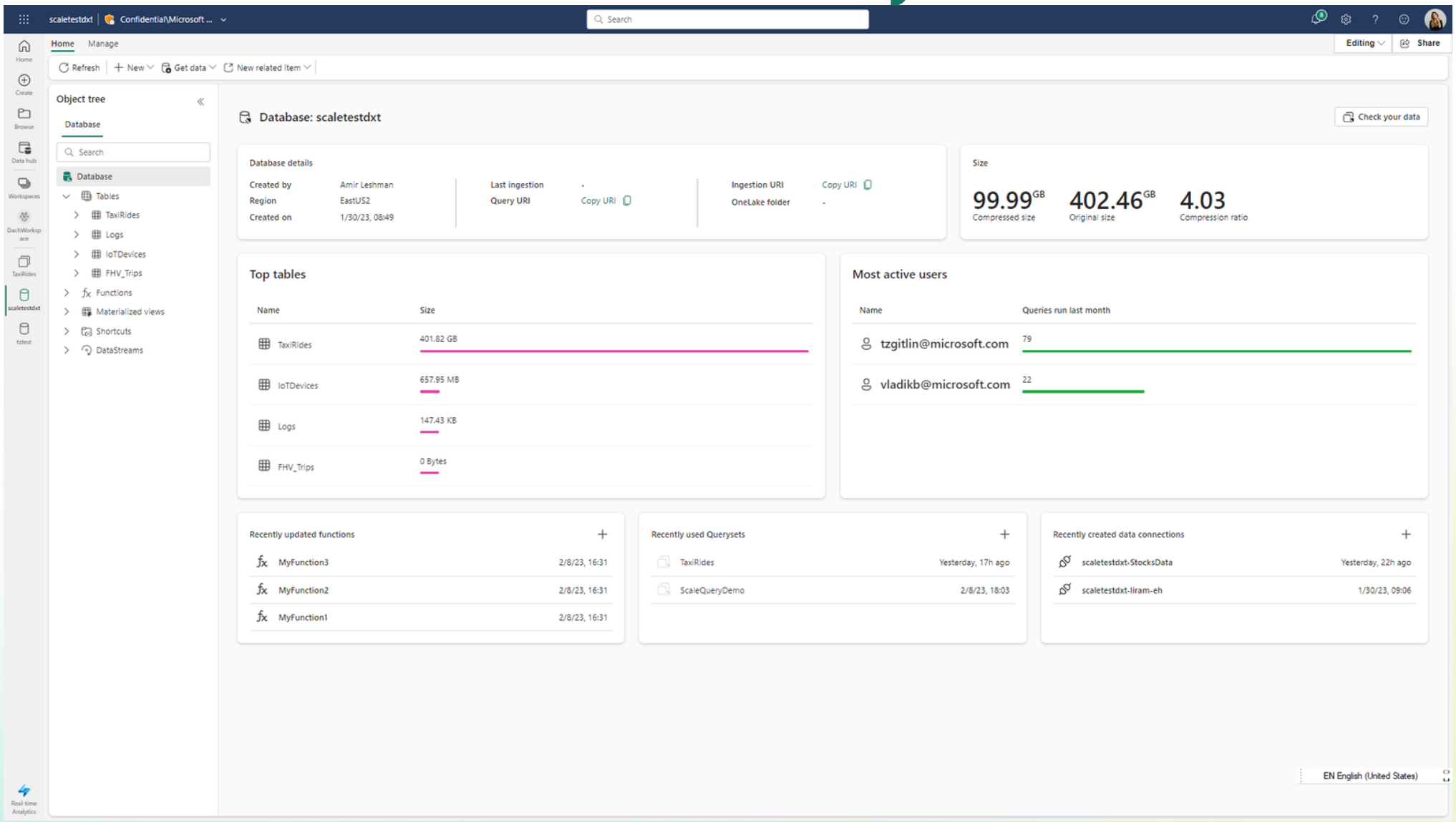
	BuyingGroup	Total
1	Toddler Toys	23
2	N/A	1
3	Wingtip Toys	201
4	Kids Toys	15
5	Tailspin Toys	163

Succeeded (1 sec 713 ms)

Columns: 2 Rows: 5

Data | Query | Model

Real-Time Analytics



Data Science

Train Evaluate | EPAM Confidential\Confidential · Saved

Home Edit Run View

Run all | Connect | PySpark (Python) | Workspace default | Data Wrangler | Copilot

```python
X_res, y_res = sm.fit_resample(X_train, y_train)
new_train = pd.concat([X_res, y_res], axis=1)
```

PySpark (Python)

Model Training

- Train the model using Random Forest with maximum depth of 4 and 4 features

```python
mlflow.sklearn.autolog(registered_model_name='rfc1_sm') # Register the trained model with autologging
rfc1_sm = RandomForestClassifier(max_depth=4, max_features=4, min_samples_split=3, random_state=1) # Pass hyperparameters
with mlflow.start_run(run_name="rfc1_sm") as run:
    rfc1_sm_run_id = run.info.run_id # Capture run_id for model prediction later
    print("run_id: {}; status: {}".format(rfc1_sm_run_id, run.info.status))
    # rfc1.fit(X_train,y_train) # Imbalanaced training data
    rfc1_sm.fit(X_res, y_res.ravel()) # Balanced training data
    rfc1_sm.score(X_val, y_val)
    y_pred = rfc1_sm.predict(X_val)
    cr_rfc1_sm = classification_report(y_val, y_pred)
    cm_rfc1_sm = confusion_matrix(y_val, y_pred)
    roc_auc_rfc1_sm = roc_auc_score(y_res, rfc1_sm.predict_proba(X_res)[:, 1])
```

PySpark (Python)

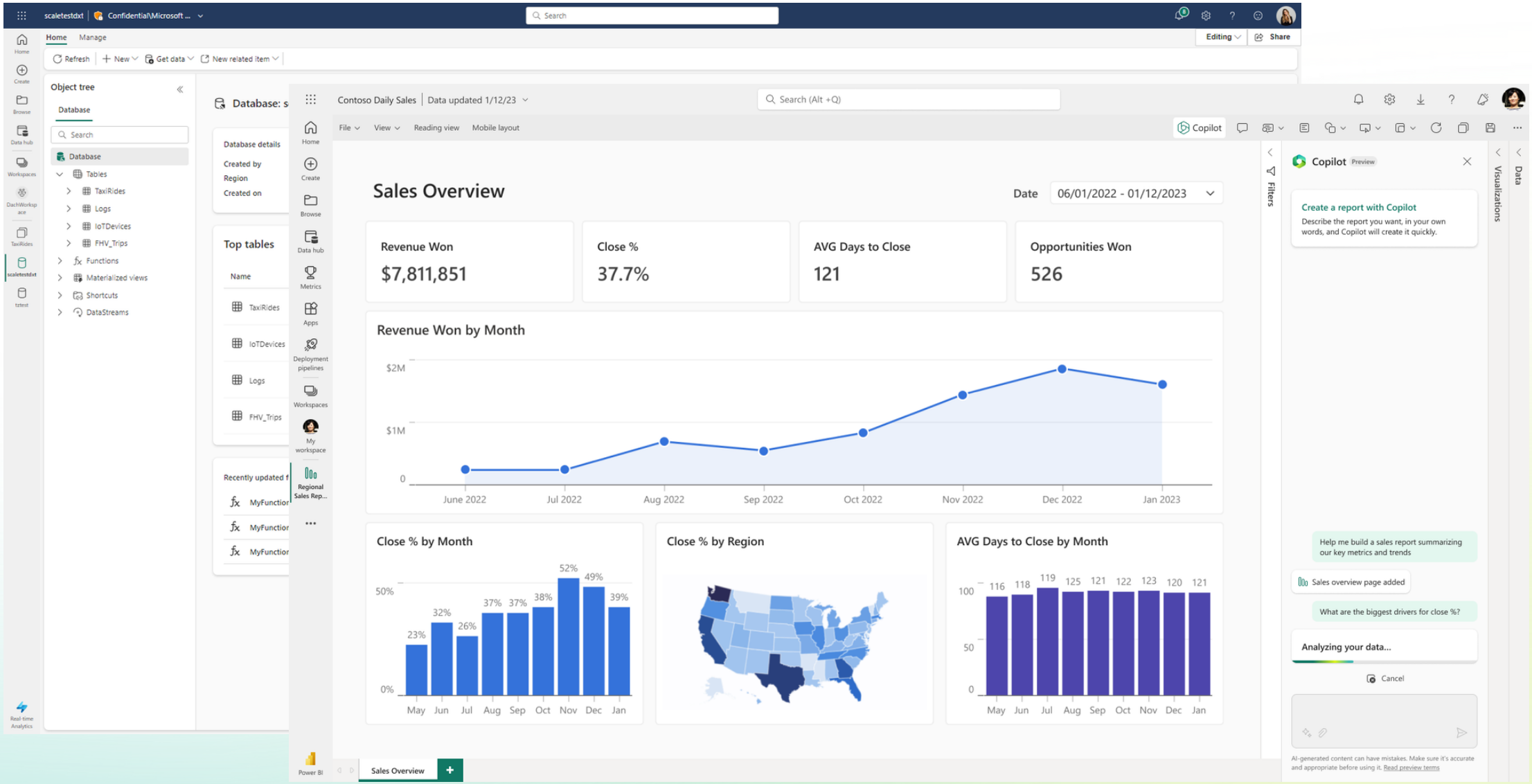
- Train the model using Random Forest with maximum depth of 8 and 6 features

```python
mlflow.sklearn.autolog(registered_model_name='rfc2_sm') # Register the trained model with autologging
rfc2_sm = RandomForestClassifier(max_depth=8, max_features=6, min_samples_split=3, random_state=1) # Pass hyperparameters
with mlflow.start_run(run_name="rfc2_sm") as run:
    rfc2_sm_run_id = run.info.run_id # Capture run_id for model prediction later
    print("run_id: {}; status: {}".format(rfc2_sm_run_id, run.info.status))
    # rfc2.fit(X_train,y_train) # Imbalanced training data
    rfc2_sm.fit(X_res, y_res.ravel()) # Balanced training data
    rfc2_sm.score(X_val, y_val)
    y_pred = rfc2_sm.predict(X_val)
    cr_rfc2_sm = classification_report(y_val, y_pred)
    cm_rfc2_sm = confusion_matrix(y_val, y_pred)
    roc_auc_rfc2_sm = roc_auc_score(y_res, rfc2_sm.predict_proba(X_res)[:, 1])
```

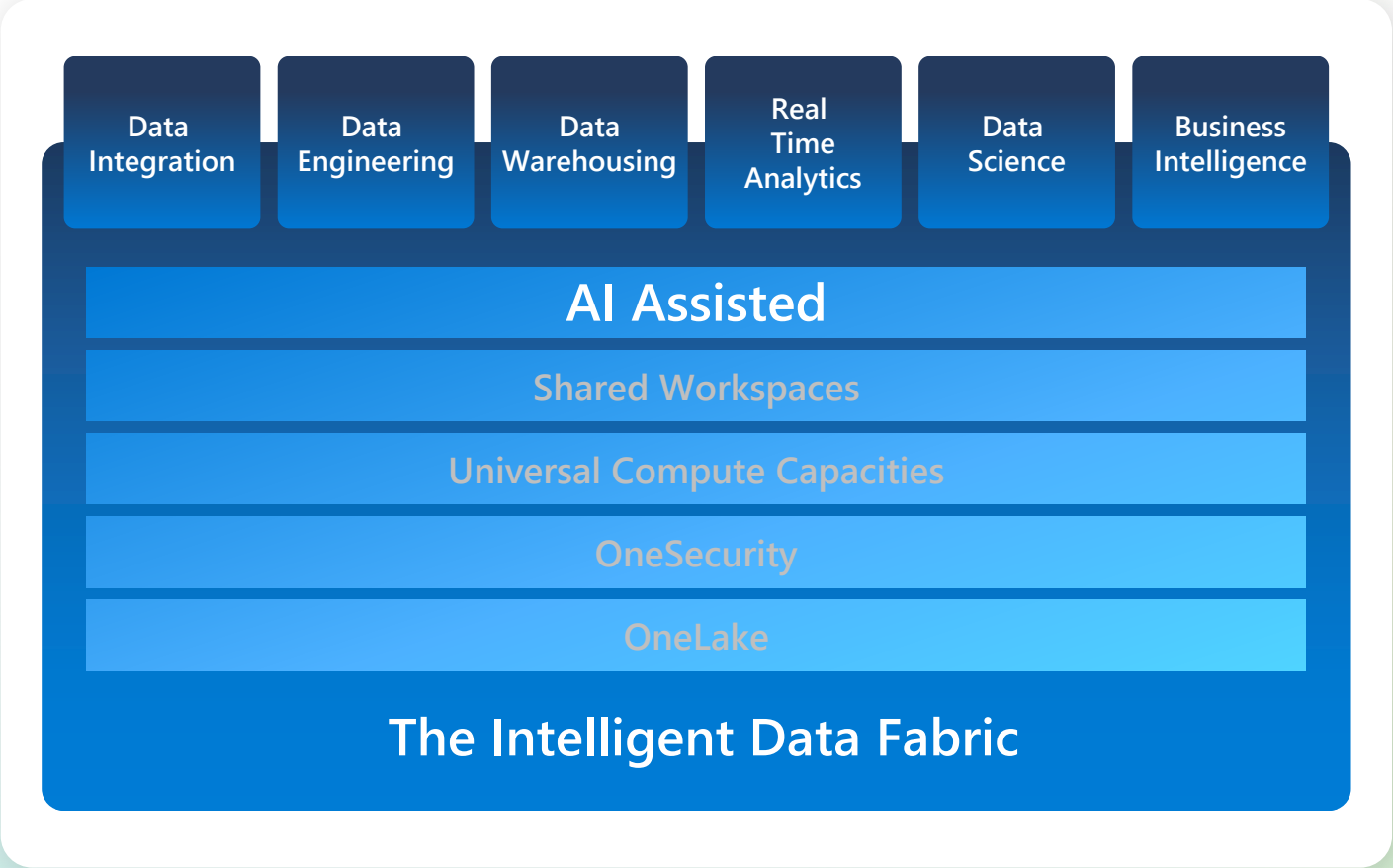
PySpark (Python)

Not connected | AutoSave: On | Selected Cell 1 of 42 cells

Power BI


Contoso Daily Sales | Data updated 1/12/23
Sales Overview
Date 06/01/2022 - 01/12/2023
Revenue Won
$7,811,851
Close %
37.7%
AVG Days to Close
121
Opportunities Won
526
Revenue Won by Month
$2M
$1M
0
June 2022
Jul 2022
Aug 2022
Sep 2022
Oct 2022
Nov 2022
Dec 2022
Jan 2023
Close % by Month
50%
0%
23% 32% 26% 37% 37% 38% 52% 49% 39%
May Jun Jul Aug Sep Oct Nov Dec Jan
Close % by Region
AVG Days to Close by Month
100
50
0
116 118 119 125 121 122 123 120 121
May Jun Jul Aug Sep Oct Nov Dec Jan
Copilot Preview
Create a report with Copilot
Describe the report you want, in your own words, and Copilot will create it quickly.
Help me build a sales report summarizing our key metrics and trends
Sales overview page added
What are the biggest drivers for close %?
Analyzing your data...
Cancel
Sales Overview

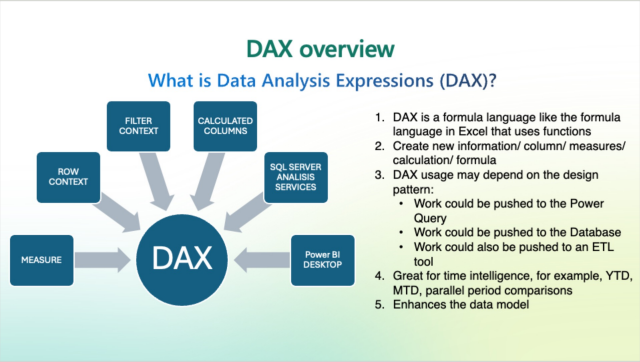
AI Assisted Creation in Microsoft Fabric



The Fabric platform will include built in Azure Open AI based assistant that will serve all the workloads

First GPT-based feature is already shipping in Power BI - NL2DAX – DAX calculation creation based on natural language prompts

Ongoing major ramp-up for pervasive AOAI based product-wide AI assistance

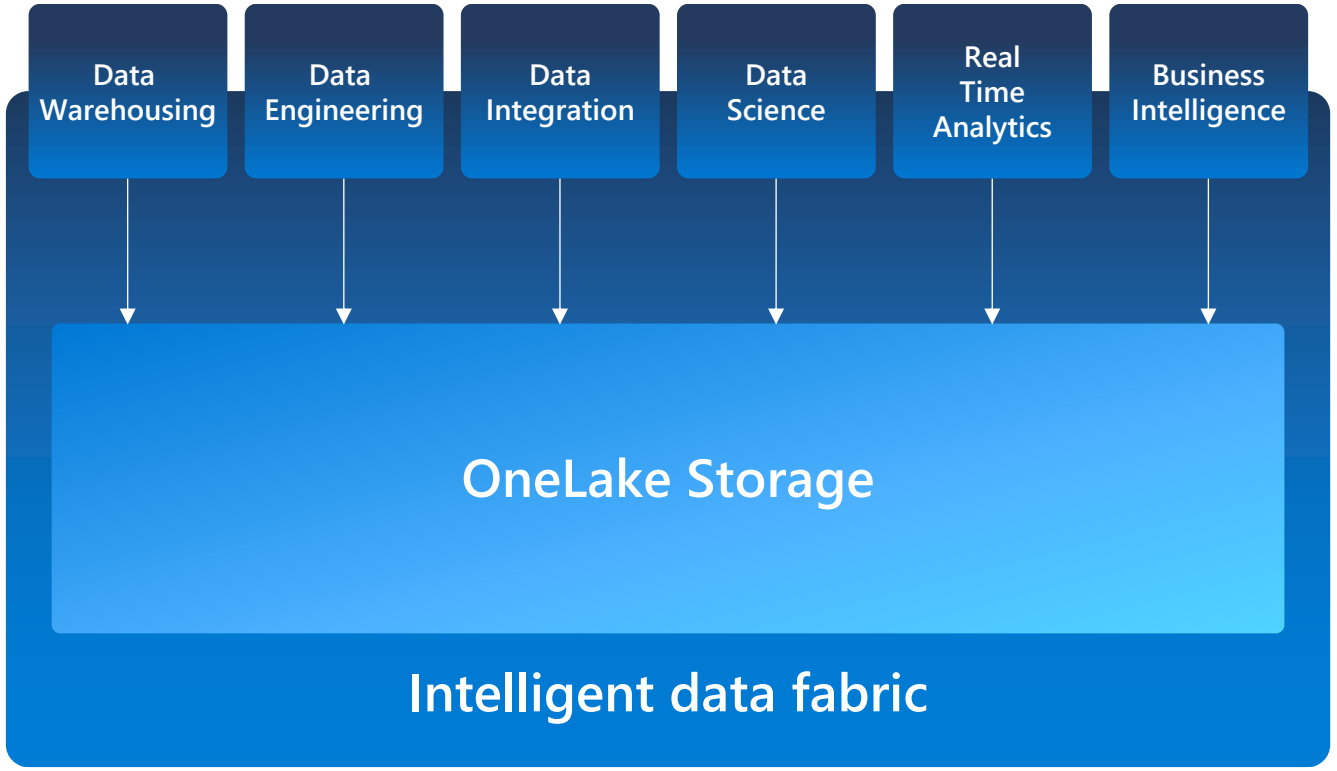


Introducing Microsoft Fabric for Data Engineering

- Complete analytics platform
- **Lake-centric and open architecture**

OneLake for all Data

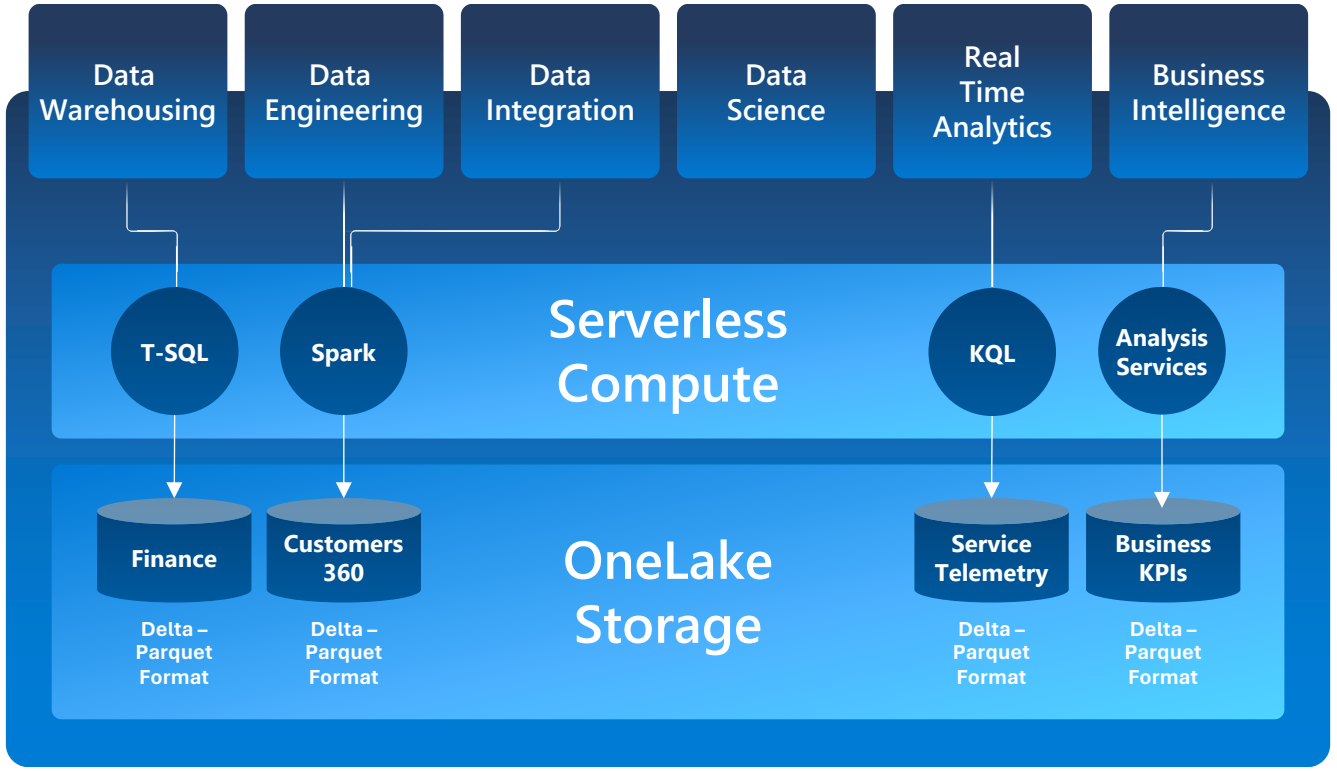
"The OneDrive for Data"



- A single SaaS lake for the whole organization
- Provisioned automatically with the tenant
- All workloads automatically store their data in the OneLake workspace folders
- All the data is organized in an intuitive hierarchical namespace
- The data in OneLake is automatically indexed for discovery, MIP labels, lineage, PII scans, sharing, governance and compliance

One Copy for all computes

Real separation of compute and storage



All the compute engines store their data automatically in OneLake

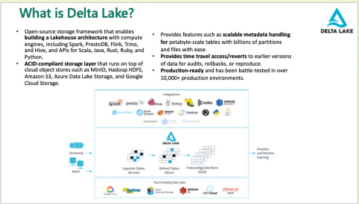
The data is stored in a single common format

Delta – Parquet, an open standards format, is the storage format for all tabular data in Fabric

Once data is stored in the lake, it is directly accessible by all the engines without needing any import/export

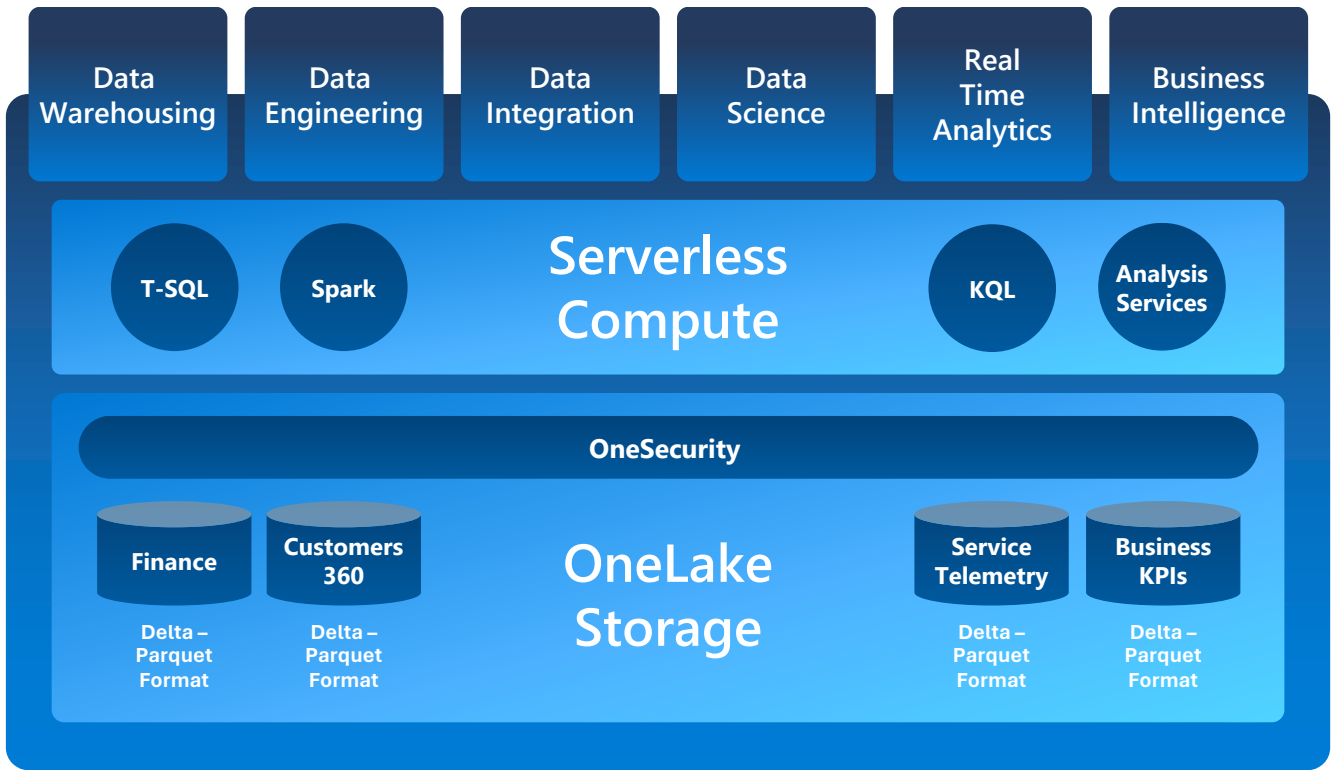
All the compute engines have been fully optimized to work with Delta Parquet as their native format

Shared universal security model is enforced across all the engines



One Copy for all computes

Universal security makes it real



All the compute engines store their data automatically in OneLake

The data is stored in a single common format

Delta – Parquet, an open standards format, is the storage format for all tabular data in Fabric

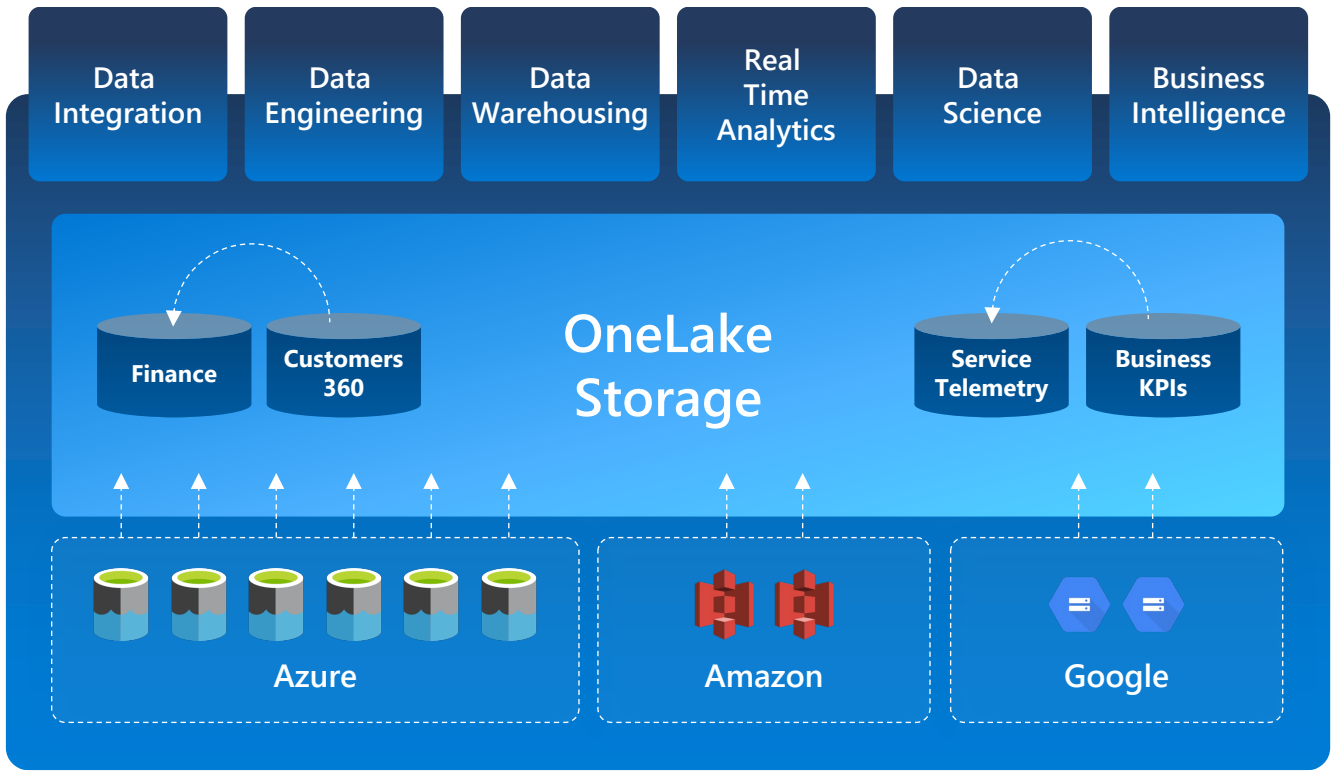
Once data is stored in the lake, it is directly accessible by all the engines without needing any import/export

All the compute engines have been fully optimized to work with Delta Parquet as their native format

Shared universal security model is enforced across all the engines

Taking One Copy to the next level

Shortcuts



Sharing data in OneLake is as easy as sharing files in OneDrive, removing the needs for data duplication

With **shortcuts**, data throughout OneLake can be composed together without any data movement

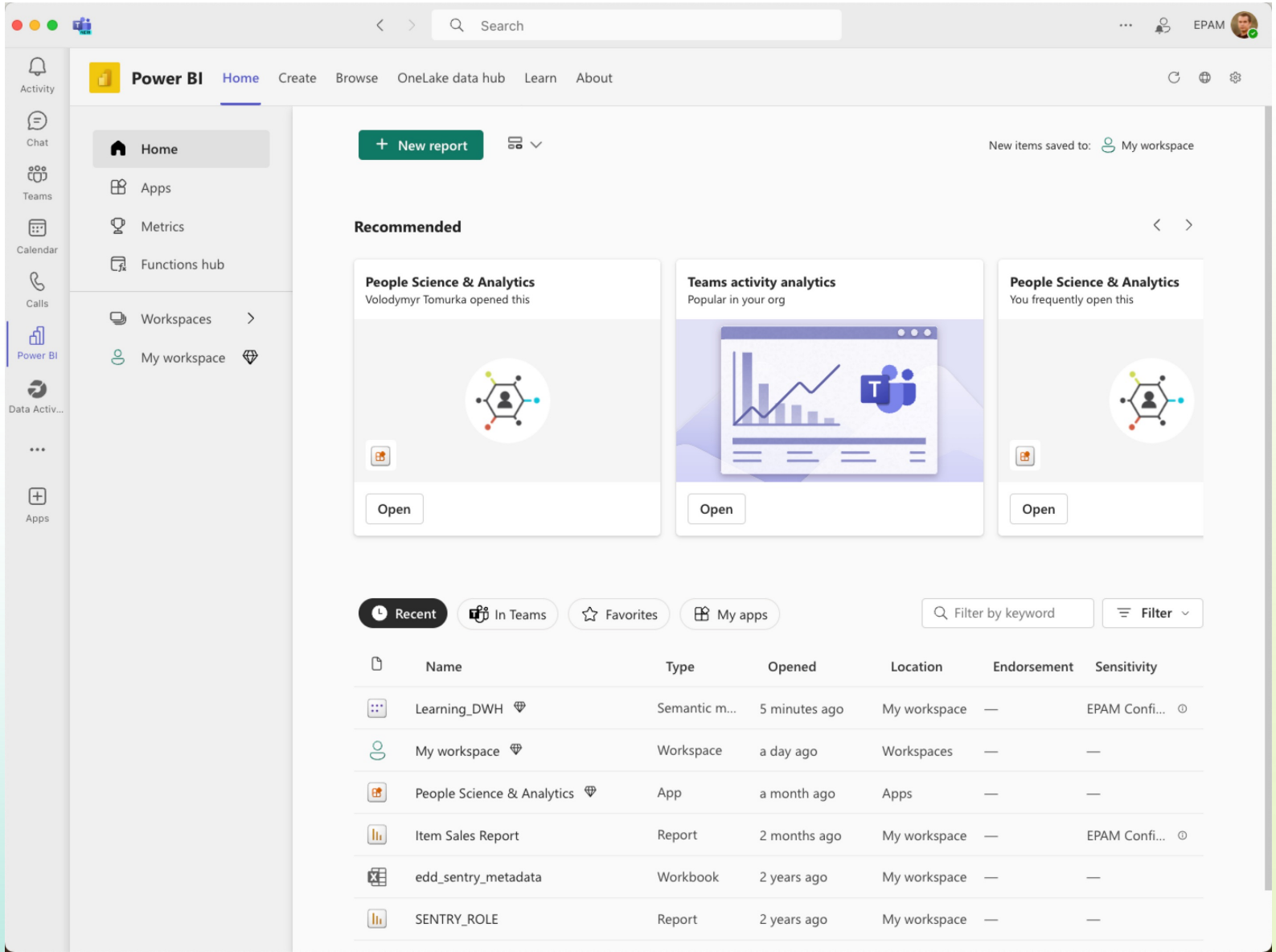
Shortcuts also allow instant linking of data already existing in Azure and in other clouds, without any data duplication and movement, making **OneLake the first multi-cloud data lake**

With support for industry standard APIs, OneLake data can be directly accessed by any application or service

Introducing Microsoft Fabric for Data Engineering

- Complete analytics platform
- Lake-centric and open architecture
- **Empower every office user**

Office Integration



Introducing Microsoft Fabric for Data Engineering

- Complete analytics platform
- Lake-centric and open architecture
- Empower every office user
- **Persistent security and governance**

Regulatory compliance

Data residency

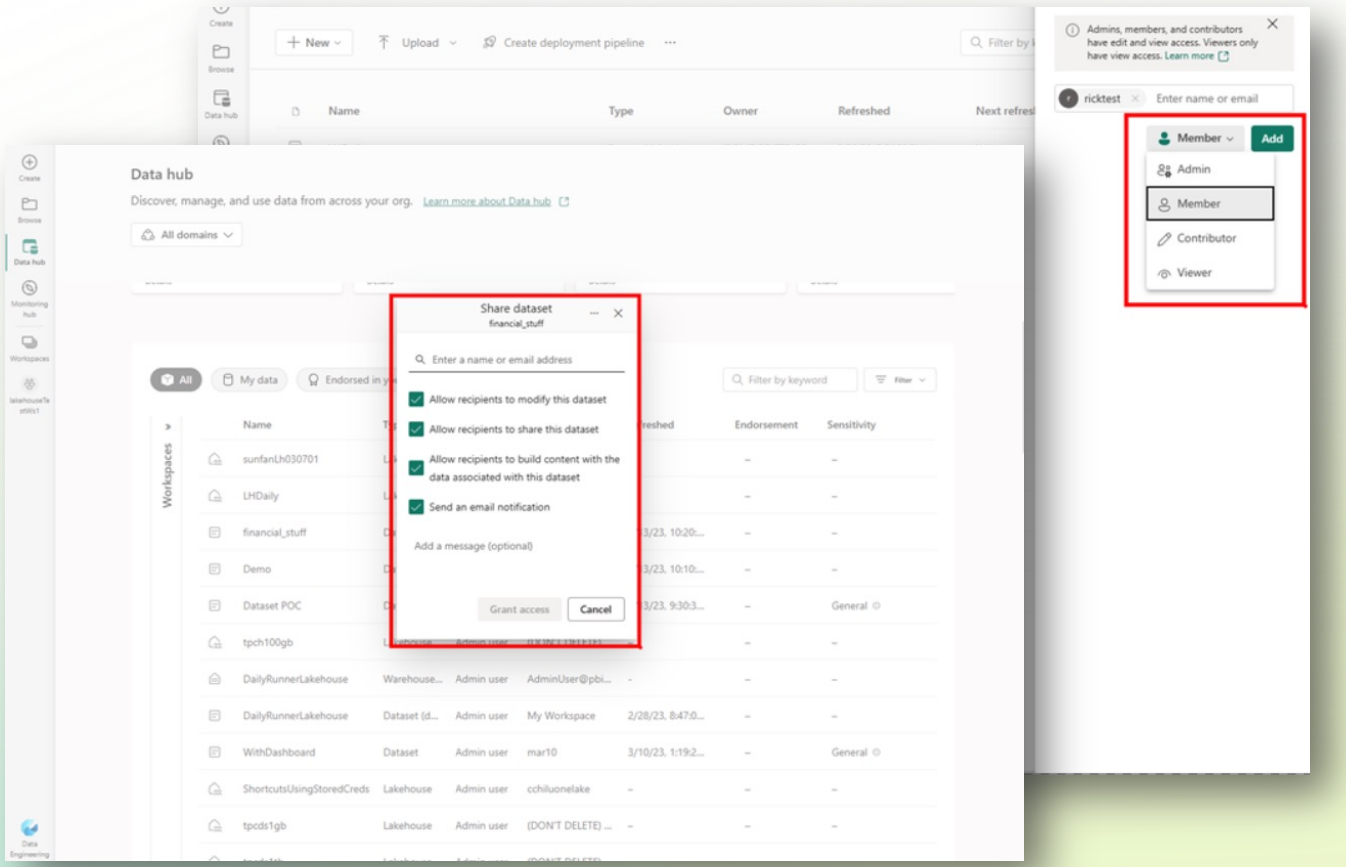
- Fabric will be available in every Azure region
- Data at rest: compliant with EUDB and other single-geo data residency regulations
- Multi-geo capacities allow control over content storage location in most Azure data centers world-wide

Microsoft pledges support for EU Data Boundary

Access control

Workspace roles and artifact permissions

- Fabric workspace roles define default permissions on workload items on the Control Plane
- Workload item permissions can be modified and managed via sharing
- On the Data Plane, Universal Security defines access policies on the delta tables directly and all workload compute engines will respect such policies



Defining user access via workspace roles and sharing

Microsoft Fabric

End-to-end analytics data fabric
From the data lake to the business user

Complete Analytics Platform

- Everything, unified
- SaaS Solution
- Low Code Plus Pro Dev

Lake-centric and Open

- OneLake
- One Copy
- Always Synced

Empower Every Office User

- Familiar and Intuitive
- Built Into Office 365
- Insight to Action

Persistent Security and Governance

- End-to-End Visibility
- Always Governed
- Secure by Default

Introducing Microsoft Fabric for Data Engineering

- Complete analytics platform
- Lake-centric and open architecture
- Empower every office user
- Persistent security and governance

Deep Dive into Data Engineering

- **Data pipelines and movement**

The choice of tools for your data transformations

Notebooks

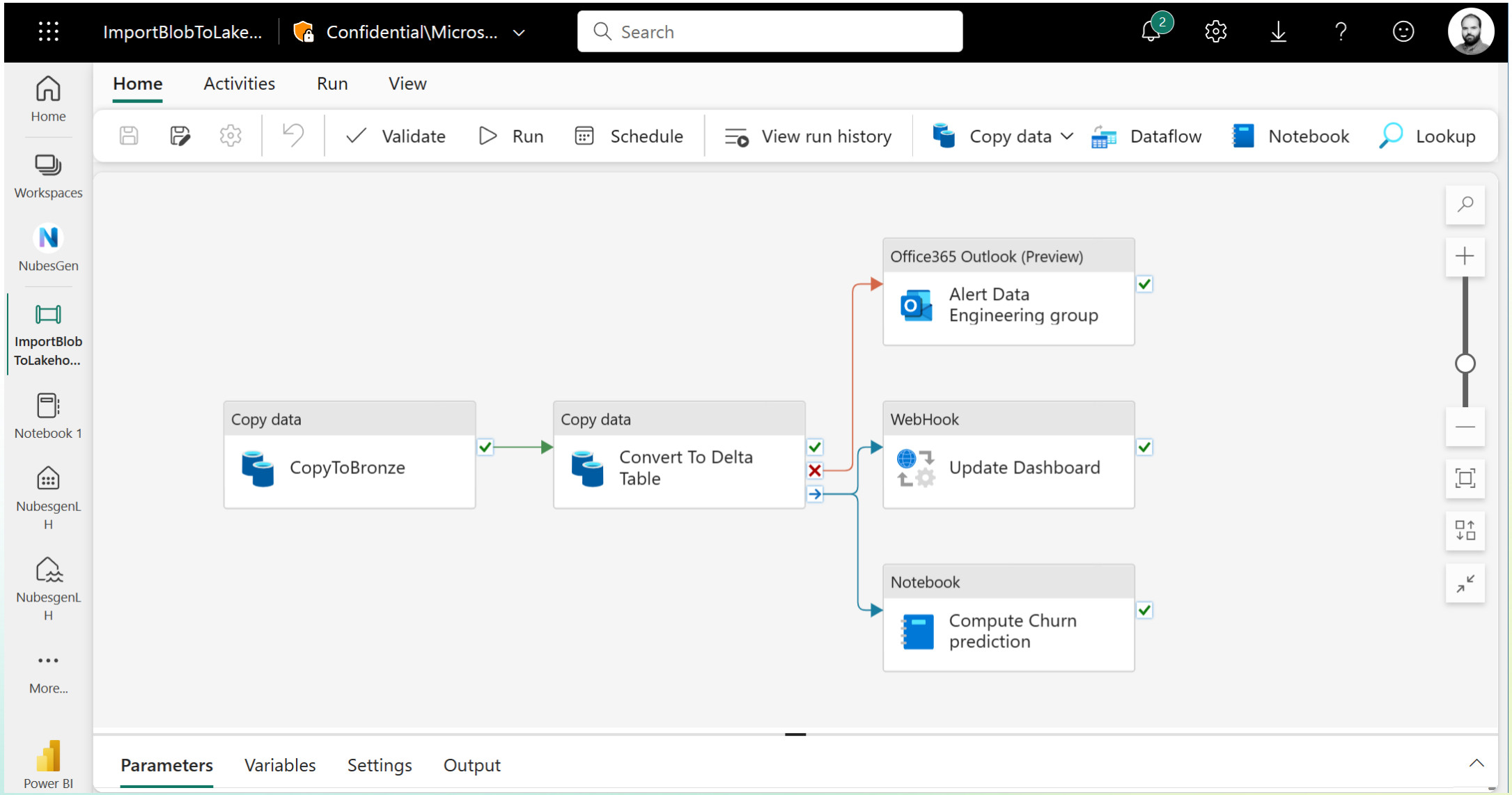
Spark Job Definitions

Data pipelines

Data Flows

Code-based pipelines

UI-based pipelines


ImportBlobToLake...
Confidential\Micros...
Search
Home
Activities
Run
View
Validate
Run
Schedule
View run history
Copy data
Dataflow
Notebook
Lookup
Home
Workspaces
NubesGen
ImportBlob ToLakeho...
Notebook 1
NubesgenLH
NubesgenLH
More...
Power BI
Office365 Outlook (Preview)
Alert Data Engineering group
Copy data
CopyToBronze
Copy data
Convert To Delta Table
WebHook
Update Dashboard
Notebook
Compute Churn prediction
Parameters
Variables
Settings
Output

Home | Transform | Add column | View | Help

Get data | Enter data | Options | Manage parameters | Refresh | Properties | Advanced editor | Manage | Add data destination | Choose columns | Remove columns | Keep rows | Remove rows | Filter rows | Split column | Group by | Data type: Whole number | Use first row as headers | Replace values | Merge queries | Append queries | Combine files | Map to entity | Export template

New query | Options | Parameters | Query | Manage columns | Reduce rows | Sort | Transform | Combine | CDM | Share

Queries [2]

- DimCustomers
- SalesLT Customer

DimCustomers: SQL Server datab... → Navigation → Replace values → Filter rows

SalesLT Customer: SQL Server datab... → Navigation → Merge (1) → Remove columns → Remove duplicates → Remove columns

100%

Query settings

	CustomerID	NameStyle	Title	FirstName	MiddleName	LastName	Suffix	CompanyName	SalesPerson	EmailAddress	Phone	ModifiedDate
Valid	100%	100%	99%	100%	60%	100%	6%	100%	100%	100%	100%	100%
Error	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Empty	0%	0%	1%	0%	40%	0%	94%	0%	0%	0%	0%	0%
	847 distinct, 847 uni...	1 distinct, 0 uni...	5 distinct, 0 uni...	315 distinct, 18 ...	29 distinct, 1 uniq...	398 distinct, 31 ...	6 distinct, 0 uni...	439 distinct, 33 unique	9 distinct, 0 unique	440 distinct, 33 unique	430 distinct, 33 un...	29 distinct, 0 unique
1	1	FALSE	Mr.	Orlando	N.	Gee	null	A Bike Store	adventure-works\pamela0	orlando0@adventure-works.com	245-555-0173	8/1/2005, 12:00:00 AM
2	2	FALSE	Mr.	Keith	null	Harris	null	Progressive Sports	adventure-works\david8	keith0@adventure-works.com	170-555-0127	8/1/2006, 12:00:00 AM
3	3	FALSE	Ms.	Donna	F.	Carreras	null	Advanced Bike Components	adventure-works\jillian0	donna0@adventure-works.com	279-555-0130	9/1/2005, 12:00:00 AM
4	4	FALSE	Ms.	Janet	M.	Gates	null	Modular Cycle Systems	adventure-works\jillian0	janet1@adventure-works.com	710-555-0173	7/1/2006, 12:00:00 AM
5	5	FALSE	Mr.	Lucy	null	Harrington	null	Metropolitan Sports Supply	adventure-works\shu0	lucy0@adventure-works.com	828-555-0186	9/1/2006, 12:00:00 AM
6	6	FALSE	Ms.	Rosmarie	J.	Carroll	null	Aerobic Exercise Company	adventure-works\linda3	rosmarie0@adventure-works.com	244-555-0112	9/1/2007, 12:00:00 AM
7	7	FALSE	Mr.	Dominic	P.	Gash	null	Associated Bikes	adventure-works\shu0	dominic0@adventure-works.com	192-555-0173	7/1/2006, 12:00:00 AM
8	10	FALSE	Ms.	Kathleen	M.	Garza	null	Rural Cycle Emporium	adventure-works\josé1	kathleen0@adventure-works.com	150-555-0127	9/1/2006, 12:00:00 AM
9	11	FALSE	Ms.	Katherine	null	Harding	null	Sharp Bikes	adventure-works\josé1	katherine0@adventure-works.com	926-555-0159	8/1/2005, 12:00:00 AM
10	12	FALSE	Mr.	Johnny	A.	Caprio	Jr.	Bikes and Motorbikes	adventure-works\garrett1	johnny0@adventure-works.com	112-555-0191	8/1/2006, 12:00:00 AM
11	16	FALSE	Mr.	Christopher	R.	Beck	Jr.	Bulk Discount Store	adventure-works\jae0	christopher1@adventure-works.c...	1 (11) 500 555-0132	9/1/2006, 12:00:00 AM
12	18	FALSE	Mr.	David	J.	Liu	null	Catalog Store	adventure-works\michael9	david20@adventure-works.com	440-555-0132	8/1/2005, 12:00:00 AM
13	19	FALSE	Mr.	John	A.	Beaver	null	Center Cycle Shop	adventure-works\pamela0	john8@adventure-works.com	521-555-0195	4/1/2007, 12:00:00 AM
14	20	FALSE	Ms.	Jean	P.	Handley	null	Central Discount Store	adventure-works\david8	jean1@adventure-works.com	582-555-0113	9/1/2005, 12:00:00 AM
15	21	FALSE	null	Jinghao	null	Liu	null	Chic Department Stores	adventure-works\jillian0	jinghao1@adventure-works.com	928-555-0116	9/1/2006, 12:00:00 AM
16	22	FALSE	Ms.	Linda	E.	Burnett	null	Travel Systems	adventure-works\jillian0	linda4@adventure-works.com	121-555-0121	8/1/2005, 12:00:00 AM
17	23	FALSE	Mr.	Kerim	null	Hanif	null	Bike World	adventure-works\shu0	kerim0@adventure-works.com	216-555-0122	9/1/2006, 12:00:00 AM
18	24	FALSE	Mr.	Kevin	null	Liu	null	Eastside Department Store	adventure-works\linda3	kevin5@adventure-works.com	926-555-0164	9/1/2006, 12:00:00 AM
19	25	FALSE	Mr.	Donald	L.	Blanton	null	Coalition Bike Company	adventure-works\shu0	donald0@adventure-works.com	357-555-0161	9/1/2006, 12:00:00 AM
20	28	FALSE	Ms.	Jackie	E.	Blackwell	null	Commuter Bicycle Store	adventure-works\josé1	jackie0@adventure-works.com	972-555-0163	8/1/2007, 12:00:00 AM
21	29	FALSE	Mr.	Bryan	null	Hamilton	null	Cross-Country Riding Supplies	adventure-works\josé1	bryan2@adventure-works.com	344-555-0144	8/1/2005, 12:00:00 AM
22	30	FALSE	Mr.	Todd	R.	Logan	null	Cycle Merchants	adventure-works\garrett1	todd0@adventure-works.com	783-555-0110	9/1/2006, 12:00:00 AM

Notebook 1 | EPAM Confidential\Confidential · Saved

Search

Fabric Trial: 52 days left

Home | Edit | Run | View

Editing | Comments | Share

Run all | Standard session | PySpark (Python) | Workspace default | Data Wrangler | Copilot

All sources

Lakehouses

+ Lakehouse

wwilakehouse

- Tables
  - aggregate_sale_by_date_city
  - aggregate_sale_by_date_emp...
  - dimension_city
  - dimension_customer
  - dimension_date
  - dimension_employee
  - dimension_stock_item
  - fact_sale
- Files

```python
df = spark.read.parquet("Files/wwi-raw-data/full/fact_sale_1y_full/part-00000-ced648ca-e8c8-46e5-8526-5ca85d56e67e-c000.snappy.parquet")
df.printSchema()
```

PySpark (Python)

Session ready | AutoSave: On | Selected Cell 1 of 1 cells

Home | Create | Browse | OneLake data hub | Monitoring hub | Workspaces | Fabric DEMO | Notebook 1 | Untitled report | wwilakehouse | wwilakehouse | wwilakehouse | Data Engineering

TRIDENT: SQLTHREADS DEV

- Notebook
  - Notebook 1
  - 00 - Initial Setup
  - Notebook 2
  - Compute Visitor fingerprint
  - Notebook 3
- Spark job definition
  - TestSparkJob
  - TestSparkJob2
- Lakehouse
  - SQLThreadsLakehouse
    - Tables
    - Files
      - DemoTable.csv File
  - wwilakehouse

00 - Initial Setup.ipynb

C: > src > trident-root > 138a0420-3e20-471c-a436-5df0a515d57e > 00 - Initial Setup > 00 - Initial Setup.ipynb > Sample data copy from public storage to your lakehouse

+ Code + Markdown | Run All | Clear All Outputs | Outline ··· .NET Interactive

```python
# Copyright (c) Microsoft Corporation.
# Licensed under the MIT License.

import os

bronze_zone = "//lakehouse/default/Files/bronze/"
# silver_zone = "//lakehouse/default/Tables/silver/"
# gold_zone = "//lakehouse/default/Tables/gold/"

zones = [bronze_zone] # [bronze_zone,silver_zone,gold_zone]
for path in zones:
    if not os.path.exists(path):
        print(path, "zone doesn't exist, creating it now!")
        os.mkdir(path)
    else:
        print(path, "zone exists already, skipping it now!")
```

[ ] Python

```
StatementMeta(, 78fe1cca-f90e-484a-83d4-34beb734248d, 4, Finished, Available)

//lakehouse/default/Files/bronze/ zone doesn't exist, creating it now!
```

Sample data copy from public storage to your lakehouse

```python
source_wasb = "wasbs://sampledata@azuresynapsestorage.blob.core.windows.net/WideWorldImportersDW"
```

ImportBlobToLakehouse
Data pipeline

Search

About

Sensitivity label

Endorsement

Schedule

Last success is in
April 13, 2023 at 4:34:14 PM
(UTC+01:00) Brussels, Copenhagen, Madrid, Paris

Next refresh in
15 hour(s) 47 minute(s)

Run

Schedule

Scheduled run
On Off

Repeat
Daily

Time
08:00 AM

Add a time

Start mm/dd/yyyy --:-- --

End mm/d

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

Compute Churn prediction
Notebook

Search

About

Sensitivity label

Endorsement

Schedule

Last success is in
April 13, 2023 at 2:38:08 PM
(UTC) Coordinated Universal Time

The scheduled refresh is turned off

Schedule

Scheduled run
On Off

Repeat
Hourly

Every
5 hour(s)

Start mm/dd/yyyy --:-- --

End mm/dd/yyyy --:-- --

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

Apply Discard

Introducing Microsoft Fabric for Data Engineering

- Complete analytics platform
- Lake-centric and open architecture
- Empower every office user
- Persistent security and governance

Deep Dive into Data Engineering

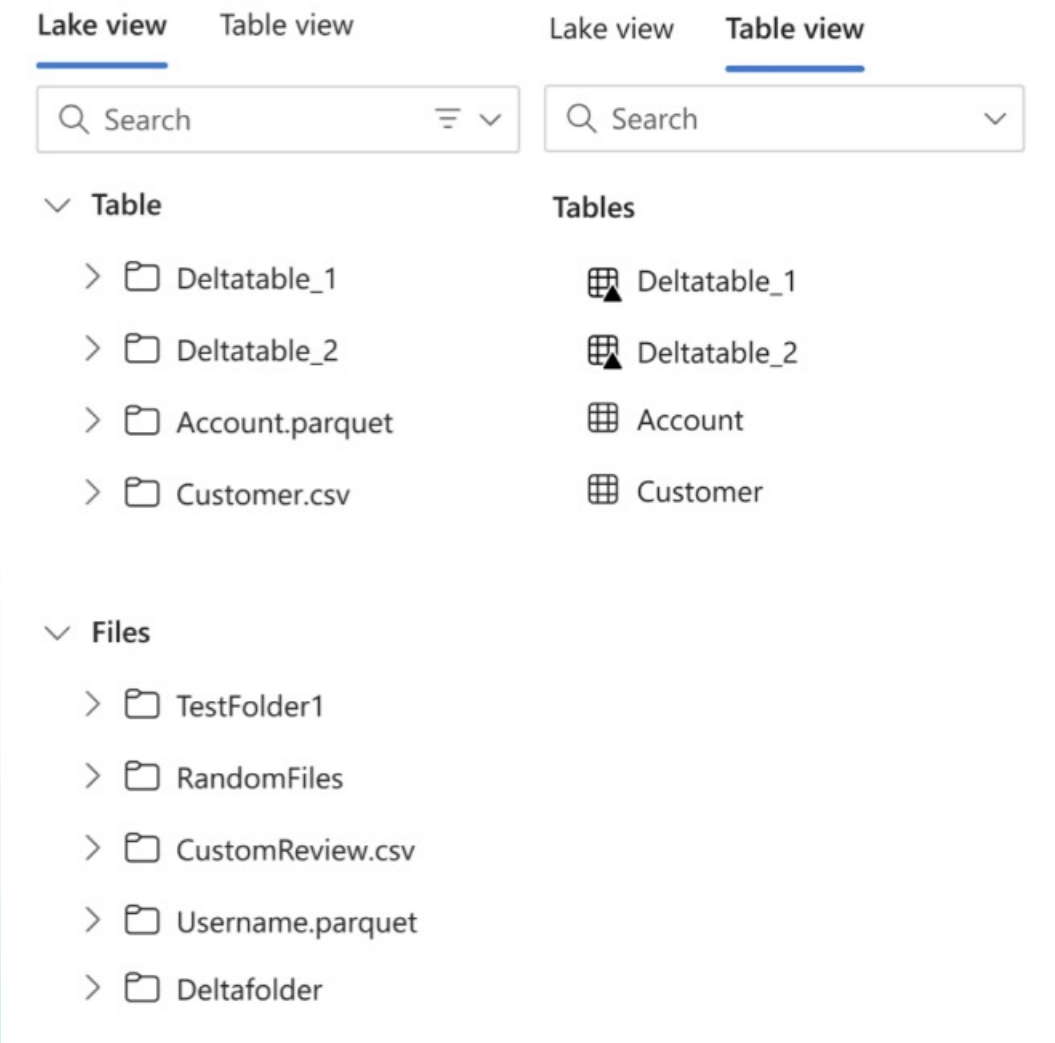
- Data pipelines and movement
- **Data storage and architecture**

One Lake

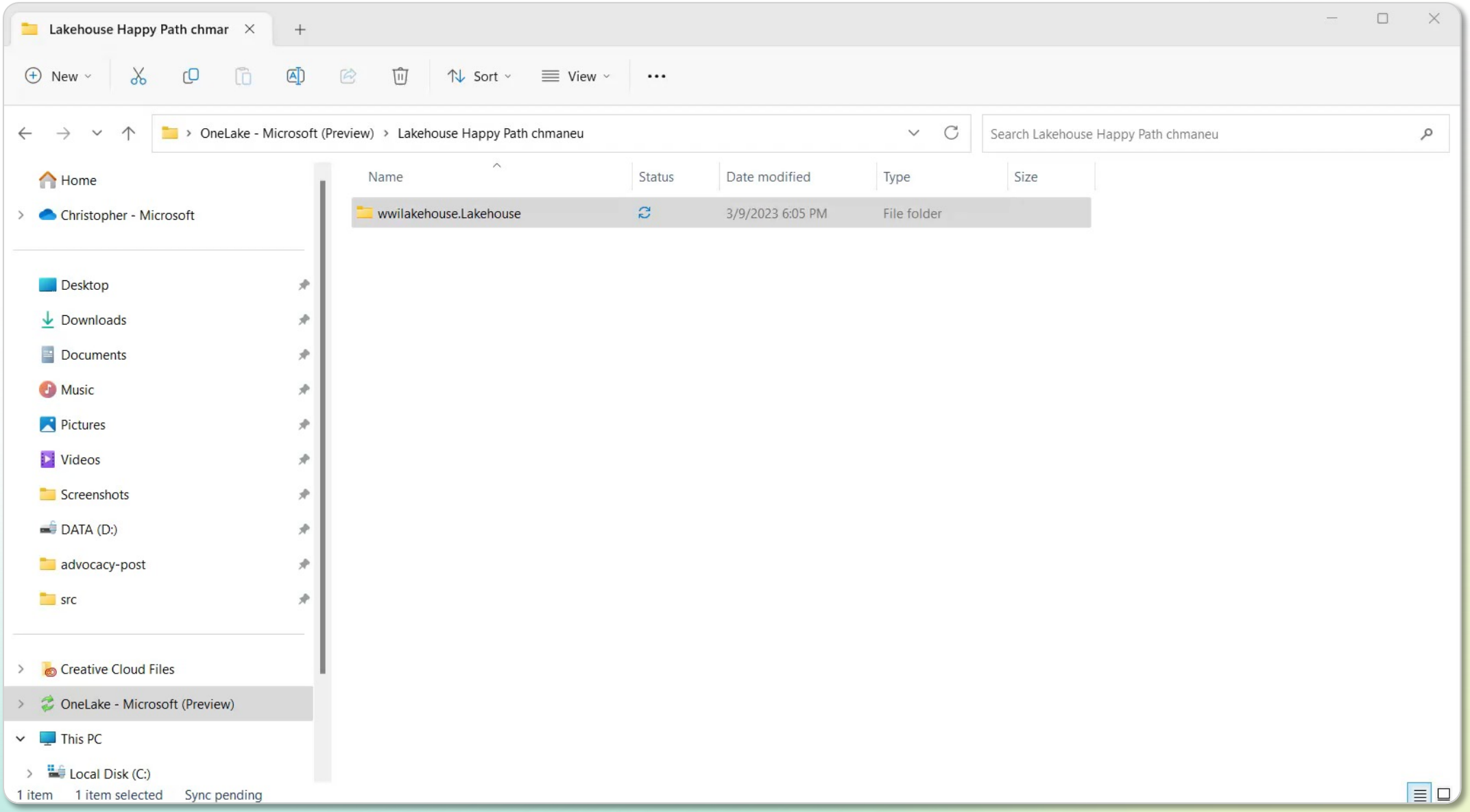
Unified lake house

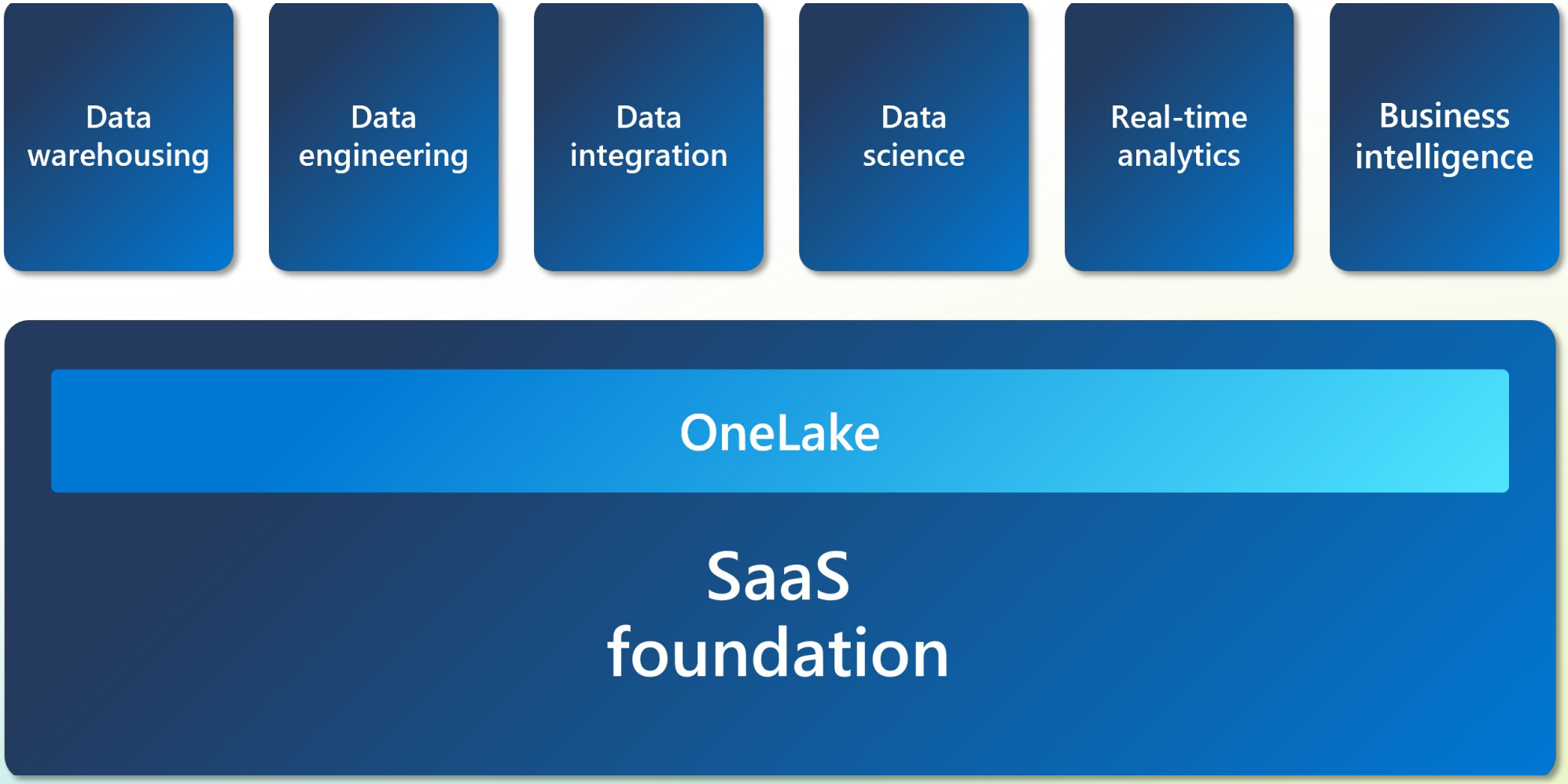
Based on open standards

Accessible from any workload


Lake view
Table view
Search
Table
Deltatable_1
Deltatable_2
Account.parquet
Customer.csv
Files
TestFolder1
RandomFiles
CustomReview.csv
Username.parquet
Deltafolder
Lake view
Table view
Search
Tables
Deltatable_1
Deltatable_2
Account
Customer


Delta Lake ACID Implementation
A
C
I
D


Lakehouse Happy Path chmar
New
Sort
View
OneLake - Microsoft (Preview) › Lakehouse Happy Path chmaneu
Search Lakehouse Happy Path chmaneu
Home
Christopher - Microsoft
Desktop
Downloads
Documents
Music
Pictures
Videos
Screenshots
DATA (D:)
advocacy-post
src
Creative Cloud Files
OneLake - Microsoft (Preview)
This PC
Local Disk (C:)
Name
Status
Date modified
Type
Size
wwilakehouse.Lakehouse
3/9/2023 6:05 PM
File folder
1 item
1 item selected
Sync pending


Data warehousing
Data engineering
Data integration
Data science
Real-time analytics
Business intelligence
OneLake
SaaS foundation

All | Azure | Database | File | Generic protocol | Services and apps

Search

Amazon RDS for SQL Server Database	**Amazon Redshift** Database	**Amazon S3** File	**Amazon S3 Compatible** File
Apache Impala Database	**Azure Blob Storage** Azure	**Azure Cosmos DB for NoSQL** Azure	**Azure Data Explorer (Kusto)** Azure
Azure Data Lake Storage Gen1 Azure	**Azure Data Lake Storage Gen2** Azure	**Azure Database for PostgreSQL** Azure	**Azure SQL Database** Azure
Azure SQL Database Managed Instance Azure	**Azure Synapse Analytics** Azure	**Azure Table Storage** Azure	**Dataverse** Services and apps
Dynamics CRM Services and apps	**Google Cloud Storage** File	**HTTP** Generic protocol	**Hive** Database
Microsoft 365 Services and apps	**OData** Generic protocol	**PostgreSQL** Database	**REST** Generic protocol,Services and apps
SQL server Database	**SharePoint Online List** Services and apps	**Snowflake** Services and apps	**Spark** Database

Choose data source

Select a connector or directly drag a file from your computer.

All categories | File | Database | Power Platform | Azure | Online services | Other

Search

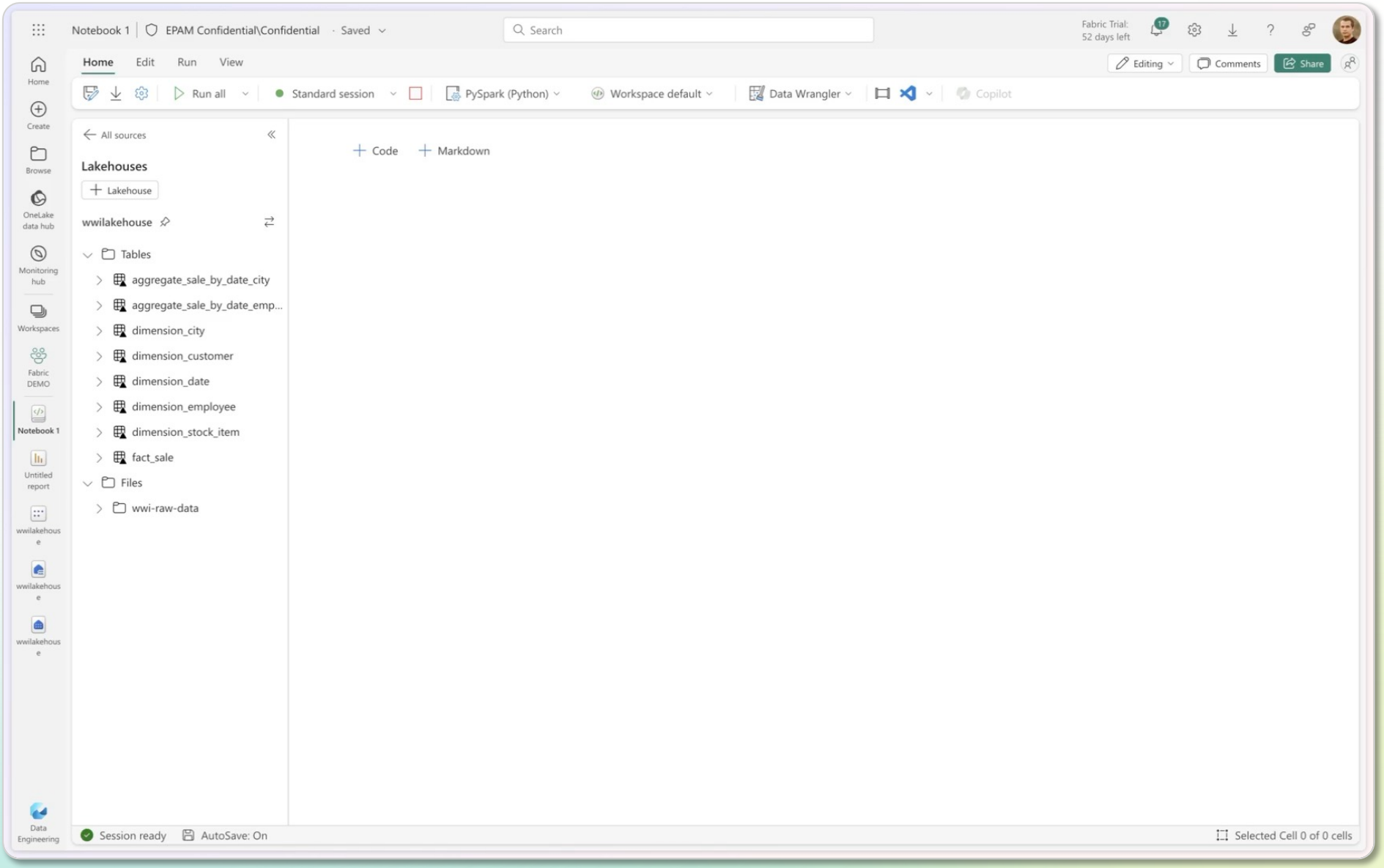
Excel workbook File	Text/CSV File	XML File	JSON File
Folder File	PDF File	Parquet File	SharePoint folder File
SQL Server database Database	Access Database	SQL Server Analysis Services Database	Oracle database Database
IBM Db2 database Database	MySQL database Database	PostgreSQL database Database	Teradata database Database
SAP HANA database Database	SAP BW Application Server Database	SAP BW Message Server Database	Snowflake Database
Google BigQuery Database	Amazon Redshift Database	Impala Database	Vertica Database
Dataflows Power Platform	Power BI dataflows (Legacy) Power Platform	Dataverse Power Platform	Common Data Service (Legacy) Power Platform
Azure SQL database Azure	Azure Synapse Analytics (SQL DW) Azure	Azure Analysis Services Azure	Azure Blobs Azure
Azure Tables Azure	Azure Data Explorer (Kusto) Azure	Azure Data Lake Storage Gen2 Azure	Azure HDInsight Spark Azure
SharePoint Online list	Microsoft Exchange Online	Salesforce objects	Salesforce reports

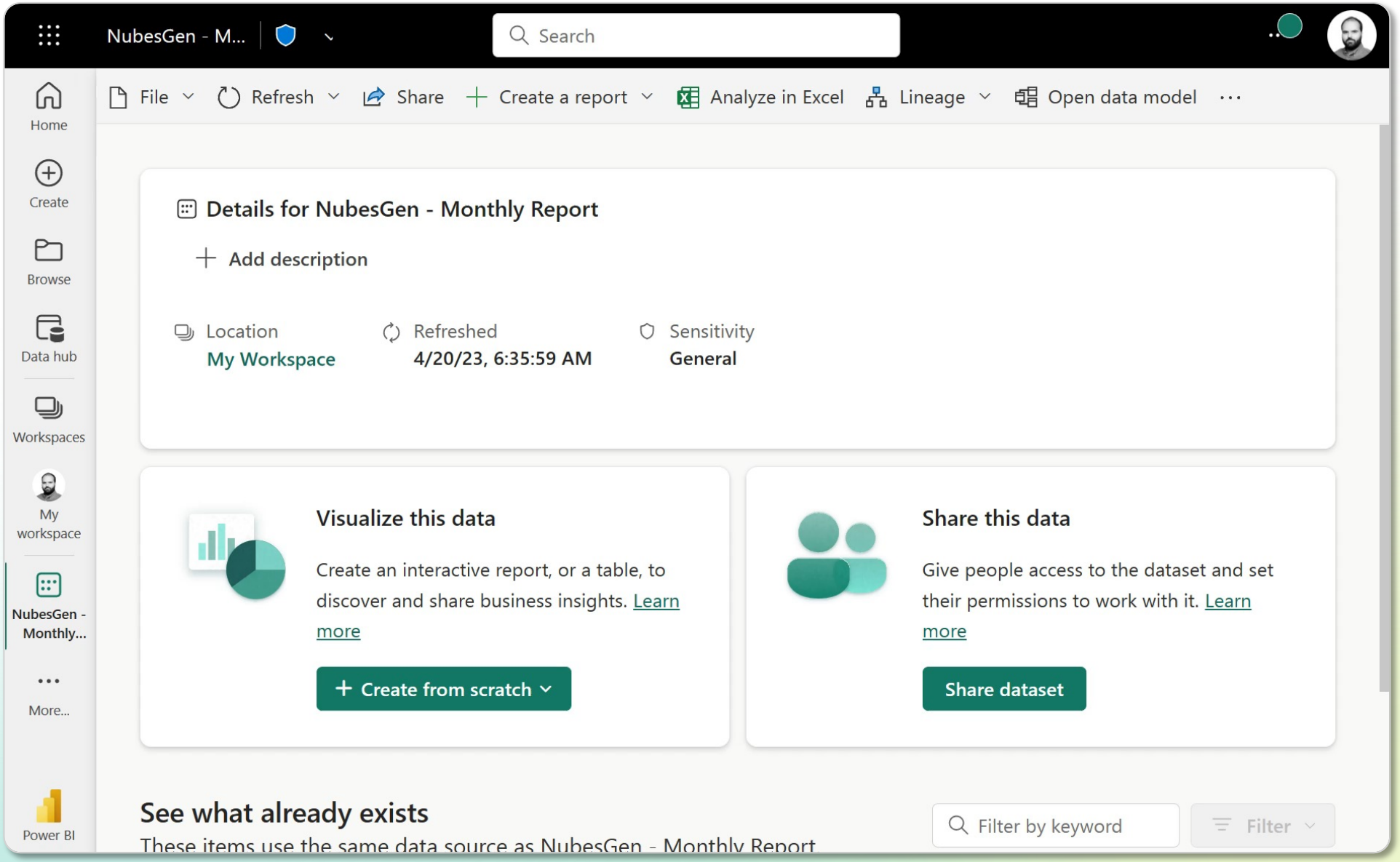
Introducing Microsoft Fabric for Data Engineering

- Complete analytics platform
- Lake-centric and open architecture
- Empower every office user
- Persistent security and governance

Deep Dive into Data Engineering

- Data pipelines and movement
- Data storage and architecture
- **Delivering data to data analysts and data scientists**


Notebook 1
EPAM Confidential\Confidential · Saved
Search
Fabric Trial: 52 days left
Home
Edit
Run
View
Editing
Comments
Share
Run all
Standard session
PySpark (Python)
Workspace default
Data Wrangler
Copilot
Home
Create
Browse
OneLake data hub
Monitoring hub
Workspaces
Fabric DEMO
Notebook 1
Untitled report
wwilakehouse
wwilakehouse
wwilakehouse
Data Engineering
All sources
Lakehouses
Lakehouse
wwilakehouse
Tables
aggregate_sale_by_date_city
aggregate_sale_by_date_emp...
dimension_city
dimension_customer
dimension_date
dimension_employee
dimension_stock_item
fact_sale
Files
wwi-raw-data
Code
Markdown
Session ready
AutoSave: On
Selected Cell 0 of 0 cells


NubesGen - M...
Search
Home
Create
Browse
Data hub
Workspaces
My workspace
NubesGen - Monthly...
More...
Power BI
File
Refresh
Share
Create a report
Analyze in Excel
Lineage
Open data model
Details for NubesGen - Monthly Report
Add description
Location
My Workspace
Refreshed
4/20/23, 6:35:59 AM
Sensitivity
General
Visualize this data
Create an interactive report, or a table, to discover and share business insights. Learn more
Create from scratch
Share this data
Give people access to the dataset and set their permissions to work with it. Learn more
Share dataset
See what already exists
These items use the same data source as NubesGen - Monthly Report
Filter by keyword
Filter

wwilakehouse | EPAM Confidential\Confidential

Search

Fabric Trial: 52 days left

Home

SQL analytics endpoint

New SQL query | New visual query

Explorer

Warehouses

- wwilakehouse
  - Schemas
    - dbo
      - Tables
        - aggregate_sale_by_dat...
        - aggregate_sale_by_dat...
        - dimension_city
        - dimension_customer
        - dimension_date
        - dimension_employee
          - EmployeeKey
          - WWIEmployeeID
          - Employee
          - PreferredName
          - IsSalesperson
          - ValidFrom
          - ValidTo
          - LineageKey

SQL query 1 | SQL query 2

Run | Save as view

```
SELECT TOP (100) [EmployeeKey]
            ,[WWIEmployeeID]
            ,[Employee]
            ,[PreferredName]
            ,[IsSalesperson]
            ,[ValidFrom]
            ,[ValidTo]
            ,[LineageKey]
FROM [wwilakehouse].[dbo].[dimension_employee]
```

Get the full current results in an Excel worksheet.

Messages | Results | Open in Excel | Explore this data (preview)

Search

	EmployeeKey	WWIEmployeeID	Employee	PreferredName	IsSalesperson	ValidFrom
1	0	0	Unknown	N/A	0	2013-01-01 00:00:00.000000
2	175	20	Jack Potter	Jack	1	2016-05-31 23:13:00.000000
3	176	19	Jai Shand	Jai	0	2016-05-31 23:13:00.000000
4	177	18	Katie Darwin	Katie	0	2016-05-31 23:13:00.000000
5	178	17	Piper Koch	Piper	0	2016-05-31 23:13:00.000000
6	179	16	Archer Lamble	Archer	1	2016-05-31 23:13:00.000000
7	180	15	Taj Shand	Taj	1	2016-05-31 23:13:00.000000
8	181	14	Lily Code	Lily	1	2016-05-31 23:13:00.000000
9	182	13	Hudson Hollinworth	Hudson	1	2016-05-31 23:13:00.000000
10	183	12	Henry Forlonge	Henry	0	2016-05-31 23:13:00.000000
11	184	11	Ethan Onslow	Ethan	0	2016-05-31 23:13:00.000000

File Home Insert Draw Page Layout Formulas Data Review View Automate Help Acrobat Table Design Query

Comments Share

Edit Properties Delete Refresh Load To Duplicate Reference Merge Append Export Connection File

Edit Load Reuse Combine Share

B3 Yasmine55@gmail.com

	A	B	C	D	E	F
1	AccountCreationDate	Email	FirstName	Id	LastName	Phone
2	1/8/2023 21:11	Yasmine_Wunsch99@yahoo.com	Yasmine	4156	Wunsch	3228813783
3	6/2/2022 2:03	Yasmine55@gmail.com	Yasmine	6527	Roberts	3279177223
4	3/31/2022 4:30	Yasmine_Lynch76@gmail.com	Yasmine	7351	Lynch	8121975420
5	2/2/2022 7:51	Yasmine_Jast@hotmail.com	Yasmine	8454	Jast	5680932006
6	11/23/2022 13:48	Yasmine_Crooks78@gmail.com	Yasmine	8867	Crooks	4530626066
7	3/2/2022 9:45	Yasmine_Daniel15@hotmail.com	Yasmine	9687	Daniel	3505196520
8	7/8/2022 21:59	Yasmine.Doyle@yahoo.com	Yasmine	11182	Doyle	0265165530
9	6/8/2022 12:40	Yasmine23@yahoo.com	Yasmine	12003	Conn	3667236525
10	2/16/2022 4:23	Yasmine78@gmail.com	Yasmine	12344	O'Hara	7778866216
11	5/27/2022 2:15	Yasmine59@yahoo.com	Yasmine	15347	Klein	7183792198
12	3/27/2022 4:21	Yasmine_Koelpin@hotmail.com	Yasmine	17023	Koelpin	4855619223
13	6/27/2022 0:42	Yasmine66@yahoo.com	Yasmine	19151	Kub	7707600661
14	8/11/2022 12:33	Yasmine.Smith@yahoo.com	Yasmine	19549	Smith	6494922664
15	7/4/2022 0:22	Yasmine_Ebert6@gmail.com	Yasmine	21910	Ebert	5162541843
16	5/12/2022 3:20	Yasmine.Howell86@yahoo.com	Yasmine	24419	Howell	5600443383
17	3/23/2022 15:39	Yasmine25@yahoo.com	Yasmine	25997	Schamberger	9396420143
18	9/19/2022 5:12	Yasmine62@yahoo.com	Yasmine	27860	Rempel	6688823727
19	7/18/2022 20:14	Yasmine.Williamson47@gmail.com	Yasmine	28558	Williamson	6160756694
20	1/30/2022 21:00	Yasmine.Reynolds29@gmail.com	Yasmine	29892	Reynolds	8579706764
21	9/7/2022 14:09	Yasmine32@gmail.com	Yasmine	36574	Pollich	4927669913

DBeaver 24.0.3 - dimension_customer

Database Navigator | Projects

- AclaraOne - jdbc:sqlserver://mdmdev.database.windows.net:1433;databa
- clickhouse.apps.dataplatform.westus2.aroapp.io - clickhouse.apps.dat
- clickhouse.apps.dataplatformqa2.westus2.aroapp.io - clickhouse.apps
- dev-cis-hub-db - jdbc:sqlserver://dev-cis-hub-db.database.windows.ne
- InsightsMetaData - mdmdev.database.windows.net:1433
- master - 2bzbxne7jytezctj7fe7gz6jdu-tou4qy6fjnruvm6docsqk3x3pi.dat
  - Databases
    - DataflowsStagingLakehouse
    - DataflowsStagingWarehouse
    - wwilakehouse
      - Schemas
        - _rsc
        - db_accessadmin
        - db_backupoperator
        - db_datareader
        - db_datawriter
        - db_ddladmin
        - db_denydatareader
        - db_denydatawriter
        - db_owner
        - db_securityadmin
        - dbo
          - Tables
            - dimension_customer 72K
          - External Tables
          - Views
          - Indexes
          - Procedures
          - Synonyms
          - Data Types
        - guest
        - queryinsights

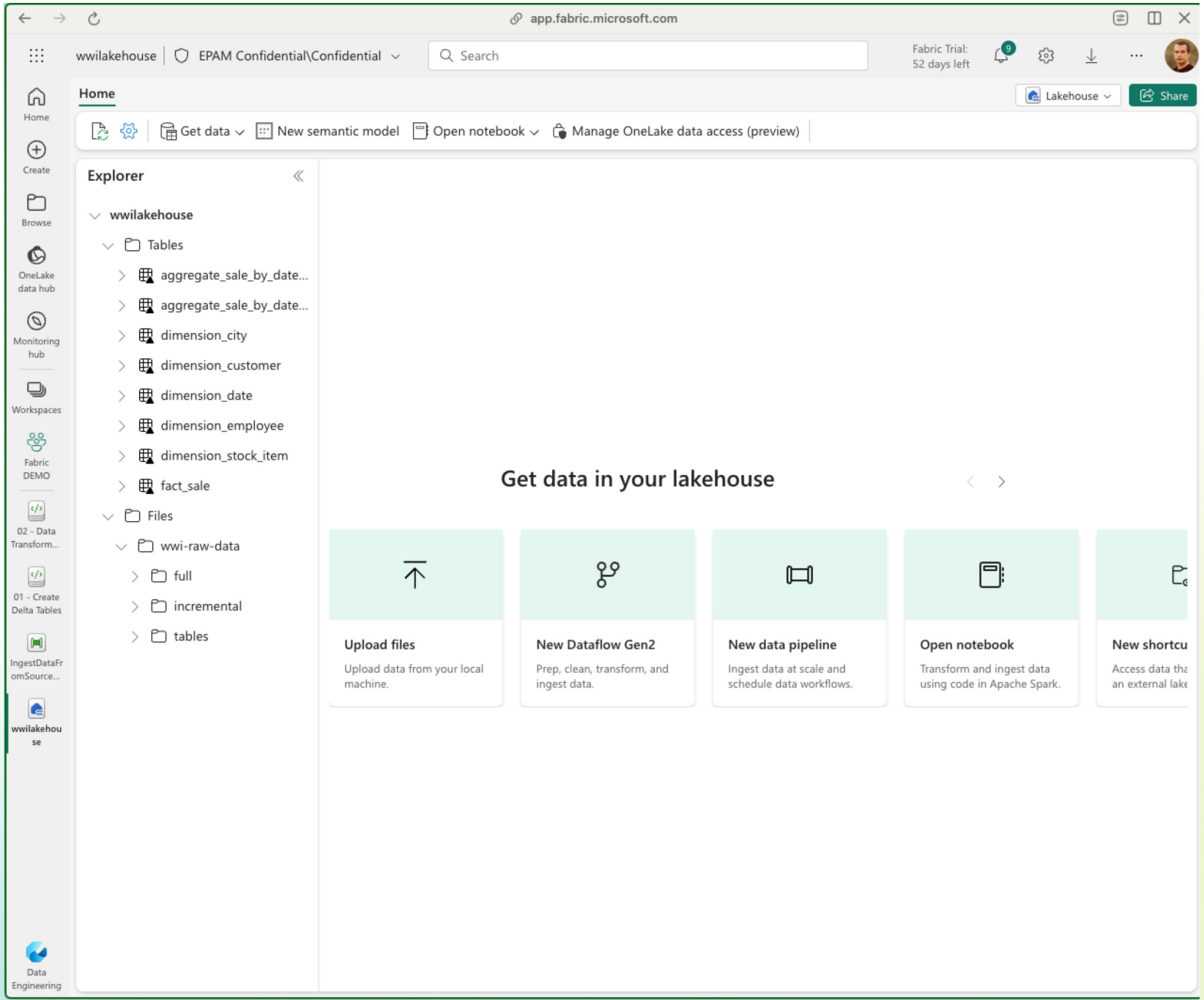
dimension_customer

Properties | Data | ER Diagram

master | Databases | wwilakehouse | Schemas | dbo | Tables

dimension_customer | Enter a SQL expression to filter results (use Ctrl+Space)

	123 CustomerKey	123 WWICustomerID	ABC Customer	ABC BillToCustomer	ABC Category	ABC BuyingGroup	ABC
3	103	103	Tailspin Toys (Kalvesta, KS)	Tailspin Toys (Head Office)	Novelty Shop	Kids Toys	Nas
4	104	104	Tailspin Toys (Wallagrass, ME)	Tailspin Toys (Head Office)	Novelty Shop	Kids Toys	Lab
5	105	105	Tailspin Toys (Tomnolen, MS)	Tailspin Toys (Head Office)	Novelty Shop	Kids Toys	Sur
6	106	106	Tailspin Toys (Tumacacori, AZ)	Tailspin Toys (Head Office)	Novelty Shop	Kids Toys	Shi
7	107	107	Tailspin Toys (Glen Avon, CA)	Tailspin Toys (Head Office)	Novelty Shop	Kids Toys	Kar
8	108	108	Tailspin Toys (Bernie, MO)	Tailspin Toys (Head Office)	Novelty Shop	Kids Toys	Bha
9	109	109	Tailspin Toys (South Laguna, CA)	Tailspin Toys (Head Office)	Novelty Shop	Kids Toys	Ae-
10	110	110	Tailspin Toys (North Crows Nest, IN)	Tailspin Toys (Head Office)	Novelty Shop	Kids Toys	Din
11	111	111	Tailspin Toys (Oriole Beach, FL)	Tailspin Toys (Head Office)	Novelty Shop	Kids Toys	Ada
12	112	112	Tailspin Toys (Sallyards, KS)	Tailspin Toys (Head Office)	Novelty Shop	Kids Toys	Ing
13	113	113	Tailspin Toys (Dahlia, NM)	Tailspin Toys (Head Office)	Novelty Shop	Kids Toys	Jae
14	114	114	Tailspin Toys (Cherry Grove Beach, SC)	Tailspin Toys (Head Office)	Novelty Shop	Kids Toys	Duc
15	115	115	Tailspin Toys (Bethania, NC)	Tailspin Toys (Head Office)	Novelty Shop	Kids Toys	Jag
16	116	116	Tailspin Toys (Rafael Capó, PR)	Tailspin Toys (Head Office)	Novelty Shop	Kids Toys	Ton
17	79	79	Tailspin Toys (Page City, KS)	Tailspin Toys (Head Office)	Novelty Shop	Toddler Toys	Mat
18	80	80	Tailspin Toys (Valdese, NC)	Tailspin Toys (Head Office)	Novelty Shop	Toddler Toys	Bha
19	81	81	Tailspin Toys (Big Moose, NY)	Tailspin Toys (Head Office)	Novelty Shop	Toddler Toys	Ver
20	82	82	Tailspin Toys (La Cueva, NM)	Tailspin Toys (Head Office)	Novelty Shop	Toddler Toys	Rak
21	83	83	Tailspin Toys (Absecon, NJ)	Tailspin Toys (Head Office)	Novelty Shop	Toddler Toys	Sar
22	84	84	Tailspin Toys (Aceitunas, PR)	Tailspin Toys (Head Office)	Novelty Shop	Toddler Toys	Eek
23	85	85	Tailspin Toys (Andrix, CO)	Tailspin Toys (Head Office)	Novelty Shop	Toddler Toys	Ton
24	86	86	Tailspin Toys (New Lexington, OH)	Tailspin Toys (Head Office)	Novelty Shop	Toddler Toys	And
25	87	87	Tailspin Toys (Sauquoit, NY)	Tailspin Toys (Head Office)	Novelty Shop	Toddler Toys	Isab
26	88	88	Tailspin Toys (Dracut, MA)	Tailspin Toys (Head Office)	Novelty Shop	Toddler Toys	Am
27	89	89	Tailspin Toys (Victory Gardens, NJ)	Tailspin Toys (Head Office)	Novelty Shop	Toddler Toys	Aak
28	90	90	Tailspin Toys (Tolna, ND)	Tailspin Toys (Head Office)	Novelty Shop	Toddler Toys	Sta


app.fabric.microsoft.com
wwilakehouse
EPAM Confidential\Confidential
Search
Fabric Trial:
52 days left
Home
Lakehouse
Share
Get data
New semantic model
Open notebook
Manage OneLake data access (preview)
Explorer
wwilakehouse
Tables
aggregate_sale_by_date...
aggregate_sale_by_date...
dimension_city
dimension_customer
dimension_date
dimension_employee
dimension_stock_item
fact_sale
Files
wwi-raw-data
full
incremental
tables
Get data in your lakehouse
Upload files
Upload data from your local machine.
New Dataflow Gen2
Prep, clean, transform, and ingest data.
New data pipeline
Ingest data at scale and schedule data workflows.
Open notebook
Transform and ingest data using code in Apache Spark.
New shortcu
Access data tha
an external lake
Home
Create
Browse
OneLake data hub
Monitoring hub
Workspaces
Fabric DEMO
02 - Data Transform...
01 - Create Delta Tables
IngestDataFromSource...
wwilakehouse
Data Engineering

File | Create a report | Analyze in Excel | Lineage | Open data model | ...

Details for NubesgenLH

Add description

Location	Endorsement	Refreshed	Sensitivity
NubesGen	Promoted	4/12/23, 2:41:42 PM	Confidential\Microsoft Exter

Introducing Microsoft Fabric for Data Engineering

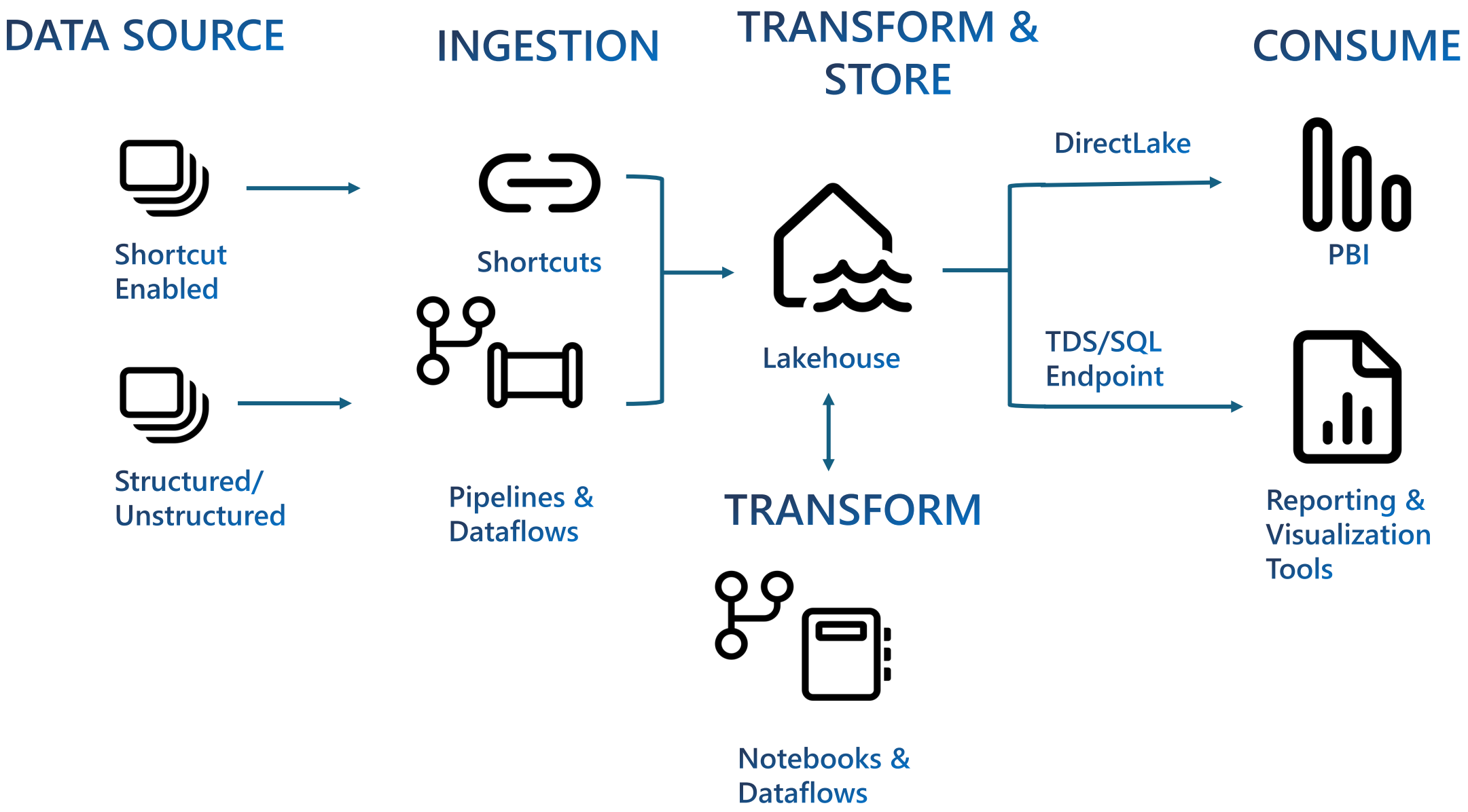
- Complete analytics platform
- Lake-centric and open architecture
- Empower every office user
- Persistent security and governance

Deep Dive into Data Engineering

- Data pipelines and movement
- Data storage and architecture
- Delivering data to data analysts and data scientists

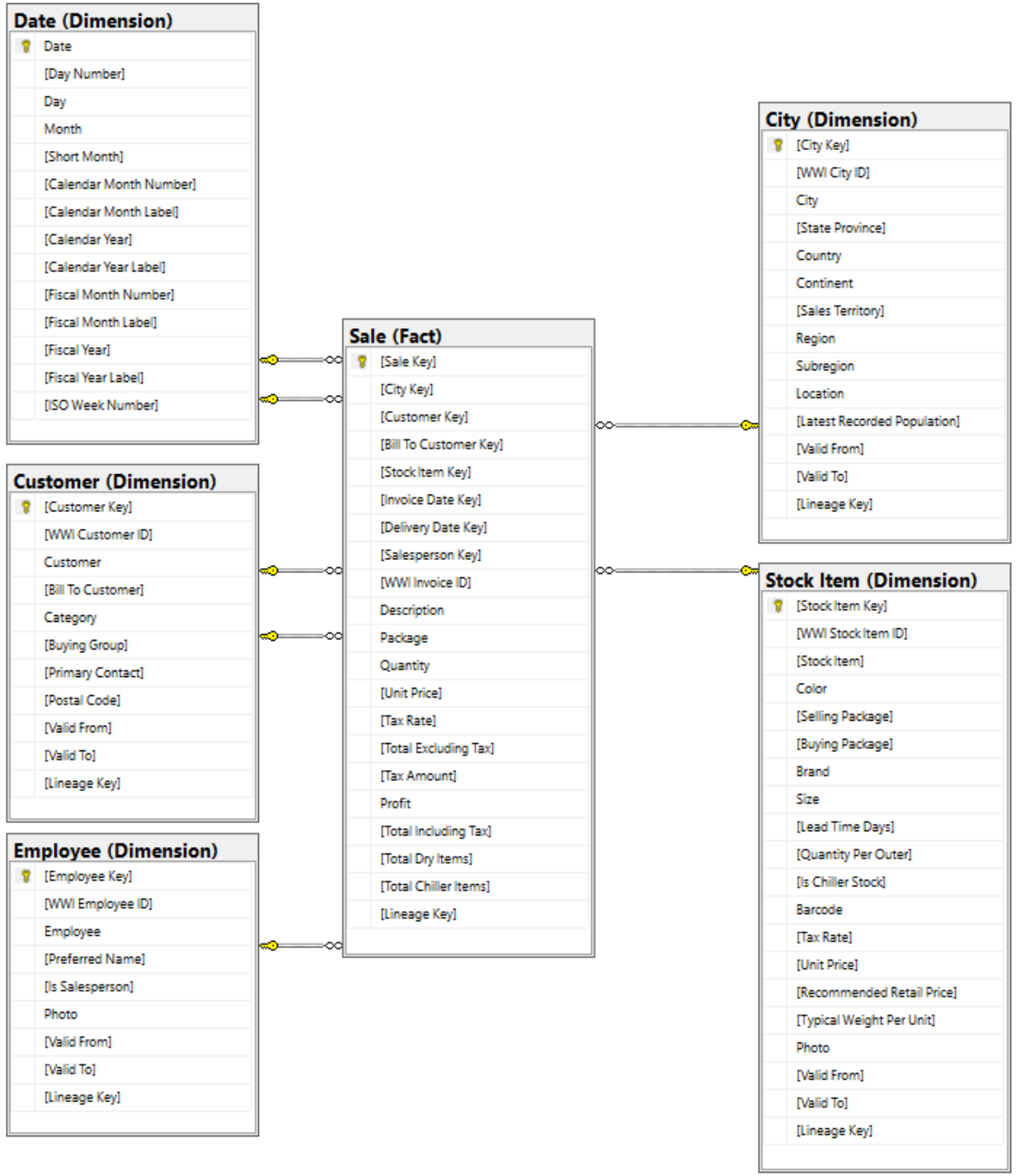
DEMO

Architecture



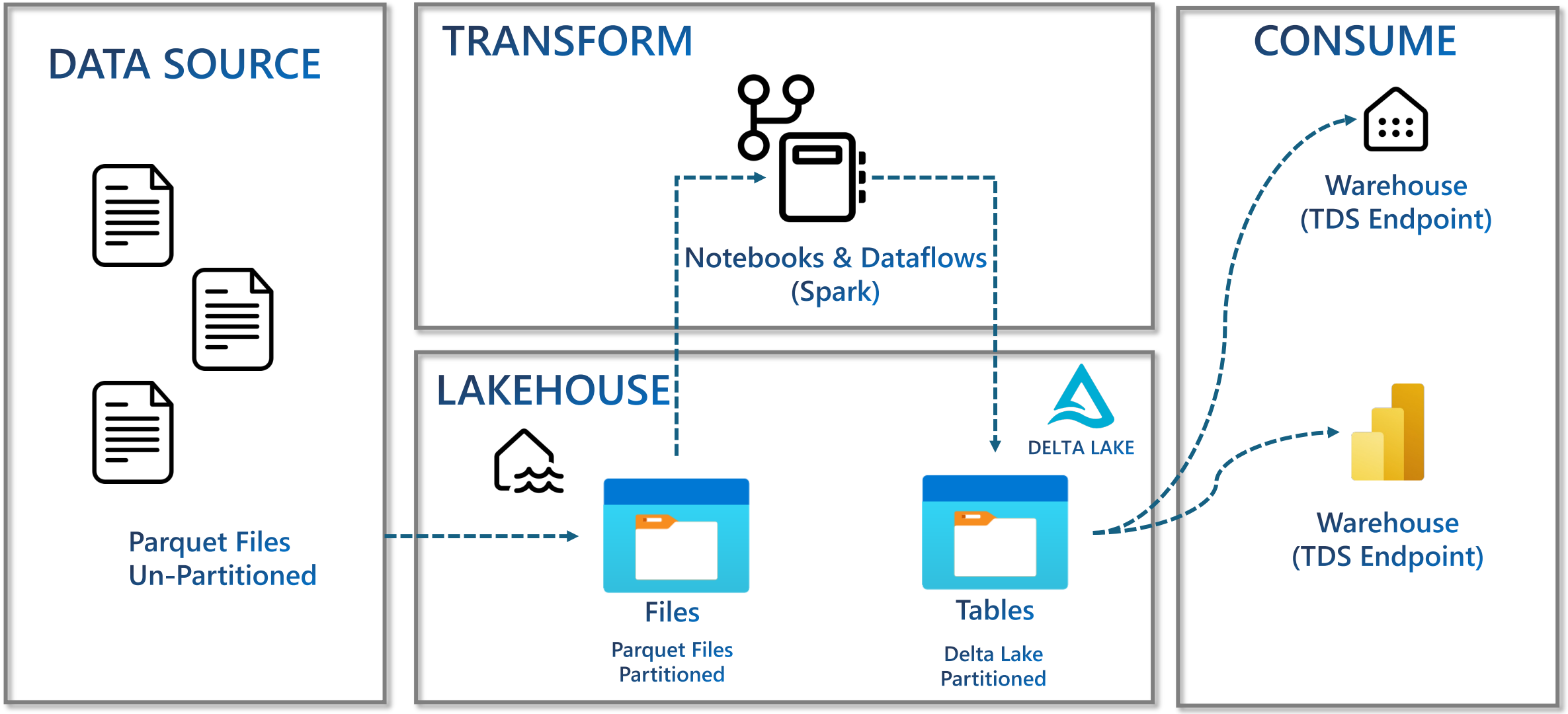
Data model

Wide World Importers (WWI) data model



See Wide World Importers sample databases for Microsoft SQL.

Data and Transformation flow



References

Learning Fabric

- Introduction End-to-End Analytics in Microsoft Fabric
- Lakehouse Get Started Lakehouses
- Spark on Lakehouse : Use Apache Spark in Lakehouse
- Work with Delta Delta Lake Tables in Microsoft Fabric
- Data Factory Pipelines Pipelines, Activities, Templates
- Data Warehouses Get started with Data Warehouses
- Real-Time Analytics Analyze real-time data
- Data Science Get started with data science in Microsoft Fabric
- Administer Administration, Security, and Govern data in Microsoft Fabric
- Medallion Architecture Design Fabric Medallion Architecture with Bronze, Silver and Gold layers of Lakehouse
- DataFlow Gen2 Ingest with Dataflows in Microsoft Fabric
- Data Analysis with Kusto Query Language Explore the fundamentals of data analysis
- Azure Data Engineer - free online training from Azure

Q&A AND SOCIAL ADS

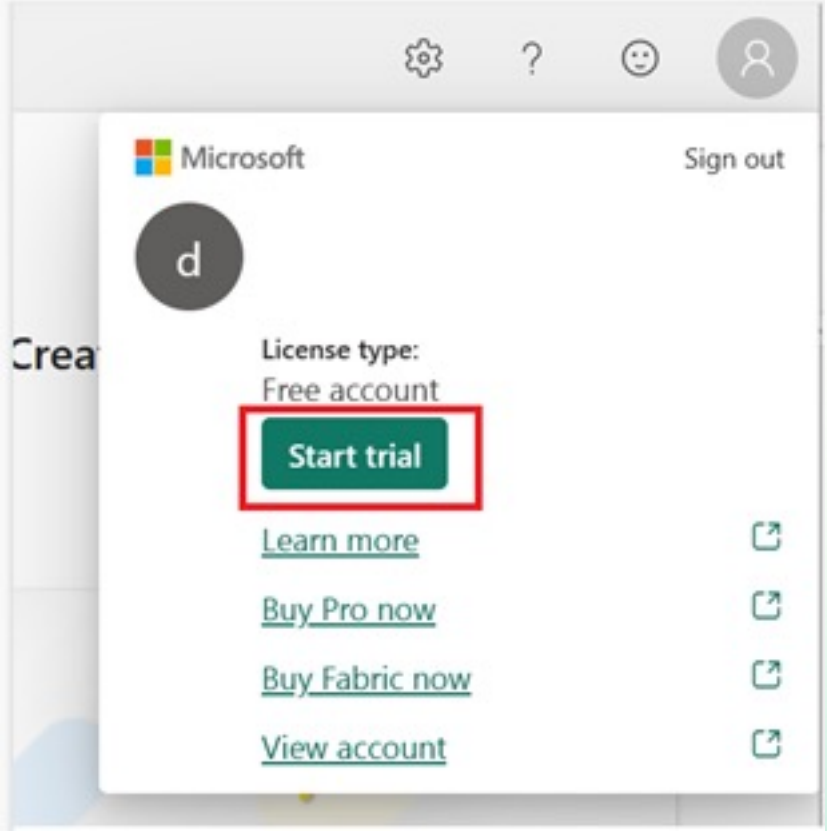
Slides

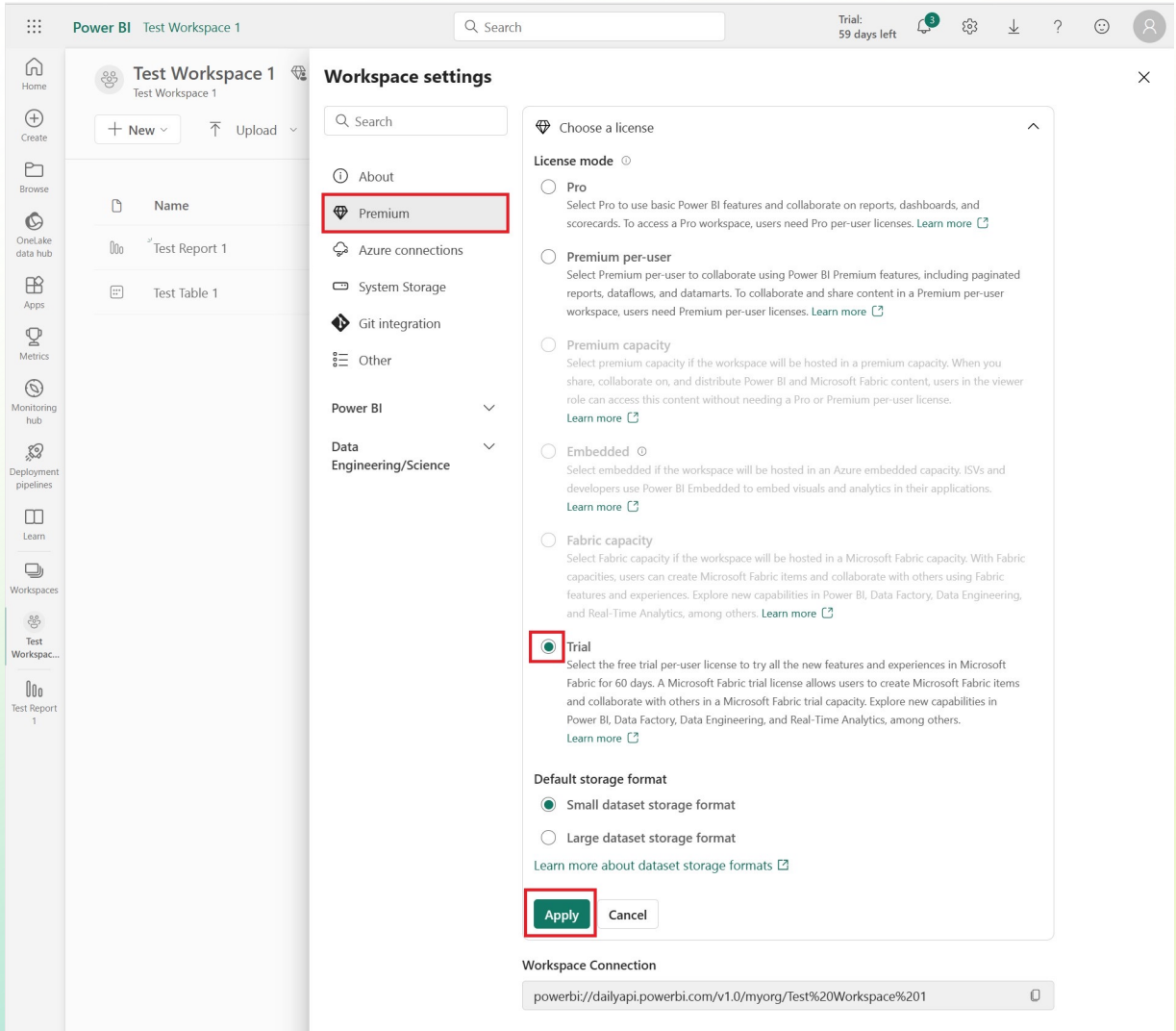
X @philg0ld #NashvilleDataEngineering

http://linkedin.com/in/philg0ld

APPENDIX

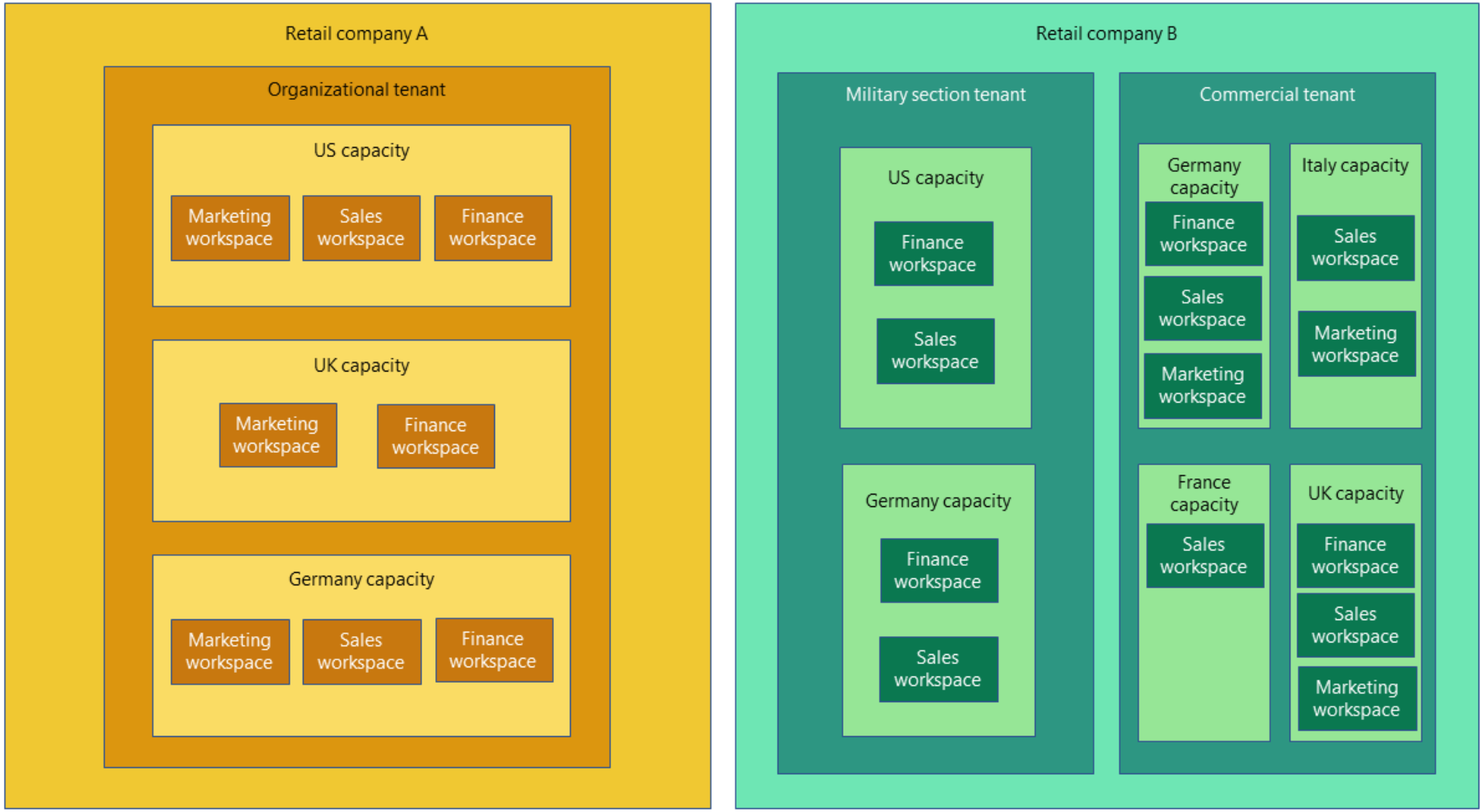
Sign up for a free Fabric trial





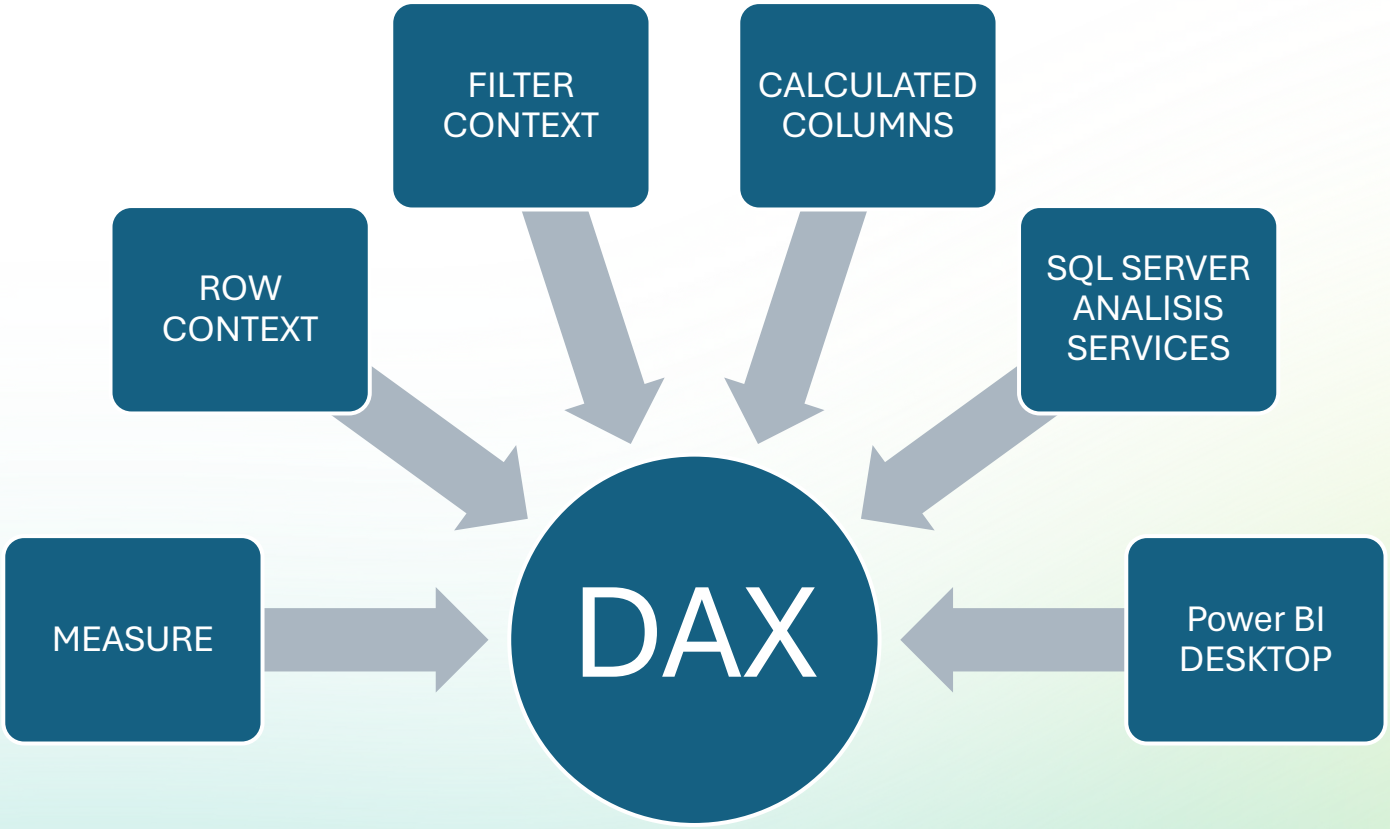
Fabric trial - Microsoft Fabric | Microsoft Learn

Microsoft Fabric concepts and licenses



DAX overview

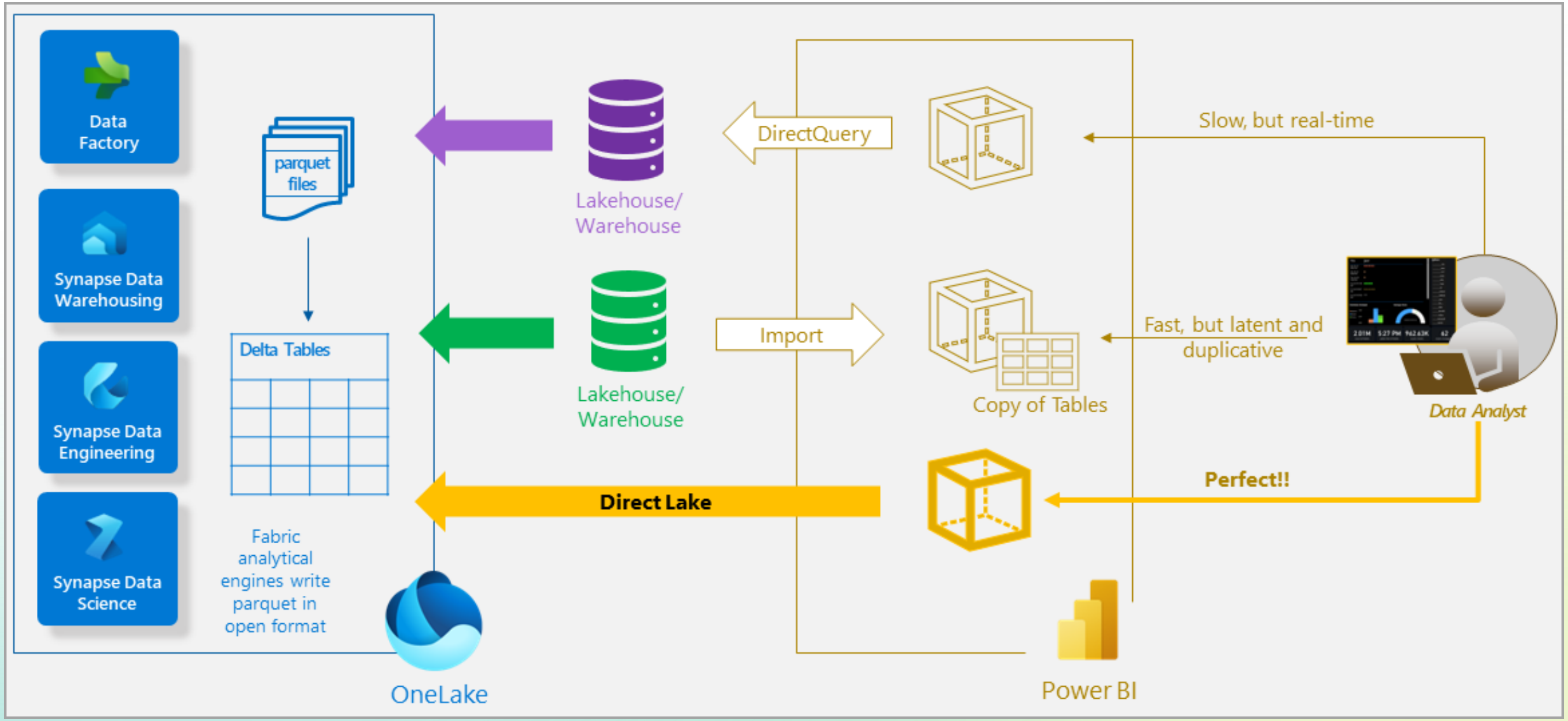
What is Data Analysis Expressions (DAX)?



1. DAX is a formula language like the formula language in Excel that uses functions
2. Create new information/ column/ measures/ calculation/ formula
3. DAX usage may depend on the design pattern:
   - Work could be pushed to the Power Query
   - Work could be pushed to the Database
   - Work could also be pushed to an ETL tool
4. Great for time intelligence, for example, YTD, MTD, parallel period comparisons
5. Enhances the data model

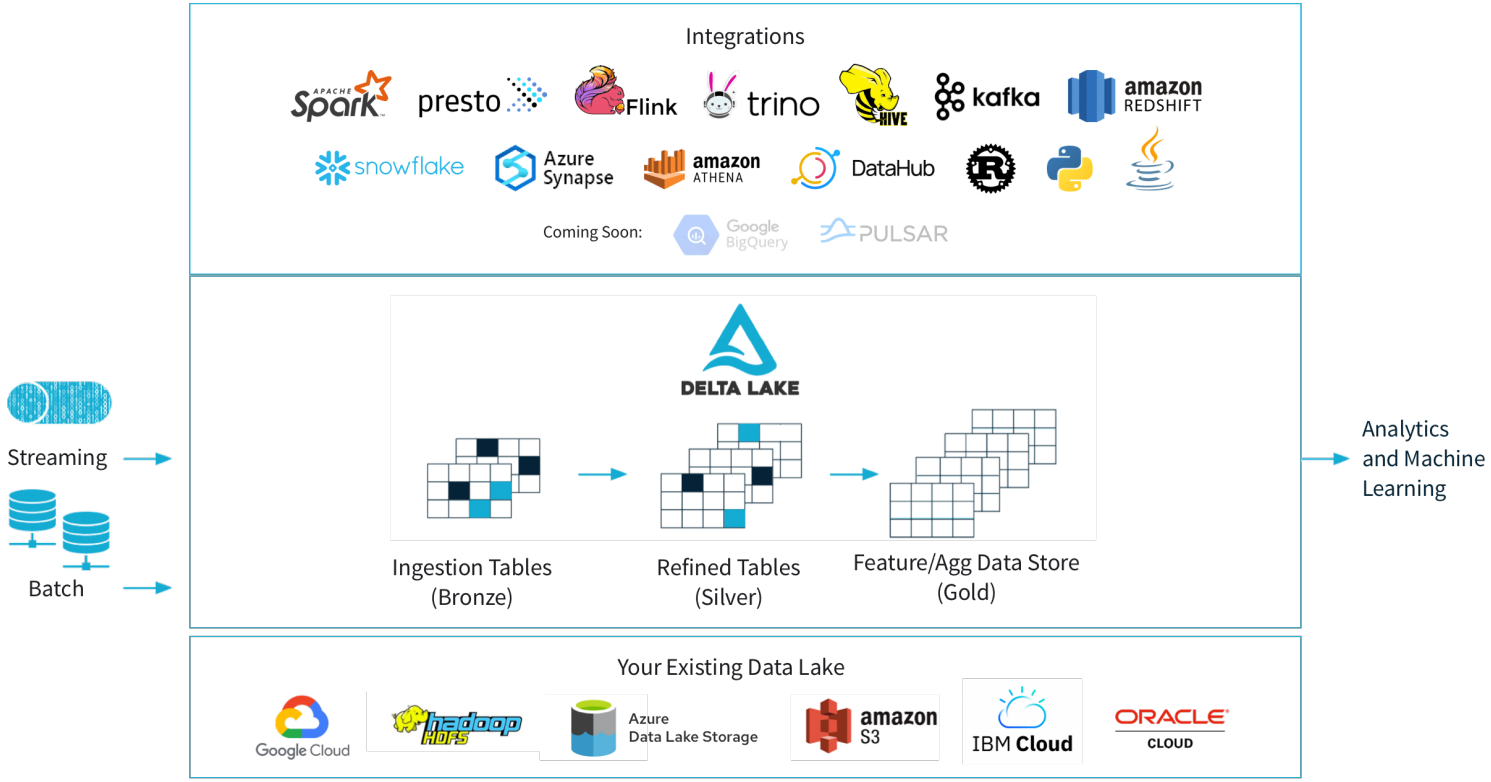
Power BI Architecture - Fabric

Understanding Direct Lake in Microsoft Fabric



What is Delta Lake?

- Open-source storage framework that enables **building a Lakehouse architecture** with compute engines, including Spark, PrestoDB, Flink, Trino, and Hive, and APIs for Scala, Java, Rust, Ruby, and Python.
- **ACID-compliant storage layer** that runs on top of cloud object stores such as MinIO, Hadoop HDFS, Amazon S3, Azure Data Lake Storage, and Google Cloud Storage.
- Provides features such as **scalable metadata handling for** petabyte-scale tables with billions of partitions and files with ease.
- **Provides time travel access/reverts** to earlier versions of data for audits, rollbacks, or reproduce.
- **Production-ready** and has been battle-tested in over 10,000+ production environments.



Delta Lake ACID Implementation