Fellowmind

Microsoft Fabric
Guidebooks

Data Engineering

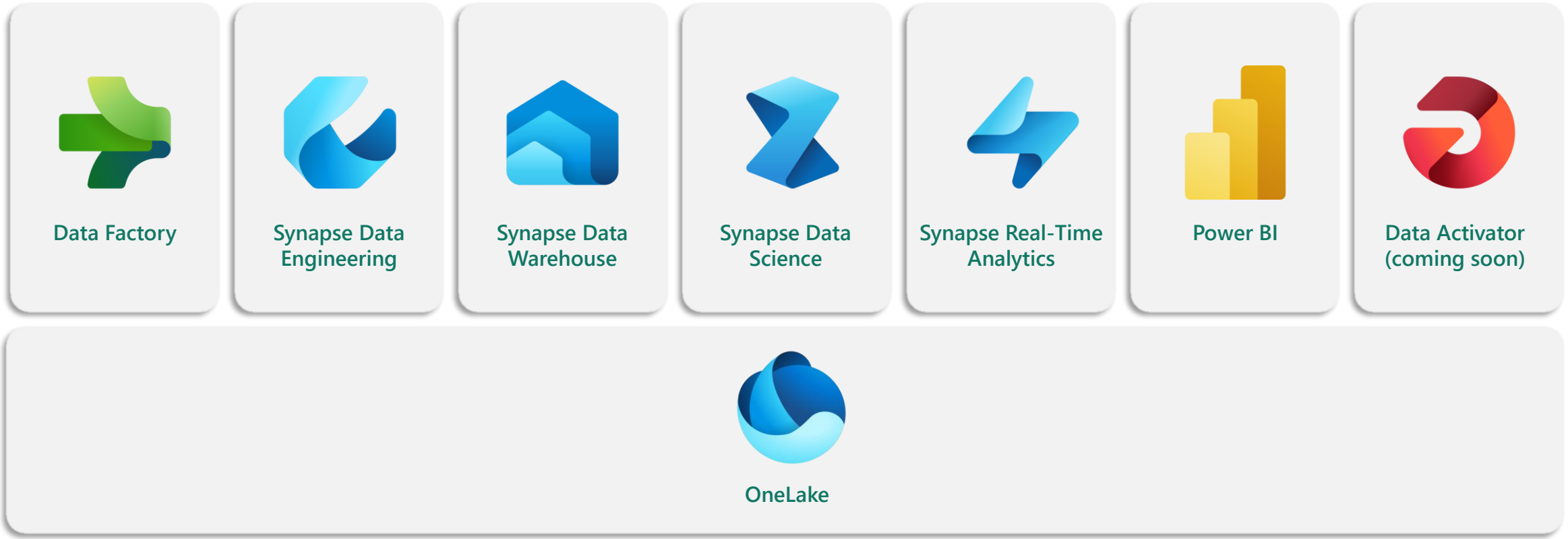
"Microsoft Fabric is a new paradigm in how we work with data – it goes beyond BI as we know it."

"It is probably the biggest innovation in data analytics since Power BI"

Mathias Halkjær
Principal Architect

MVP Microsoft® Most Valuable Professional

Microsoft Fabric



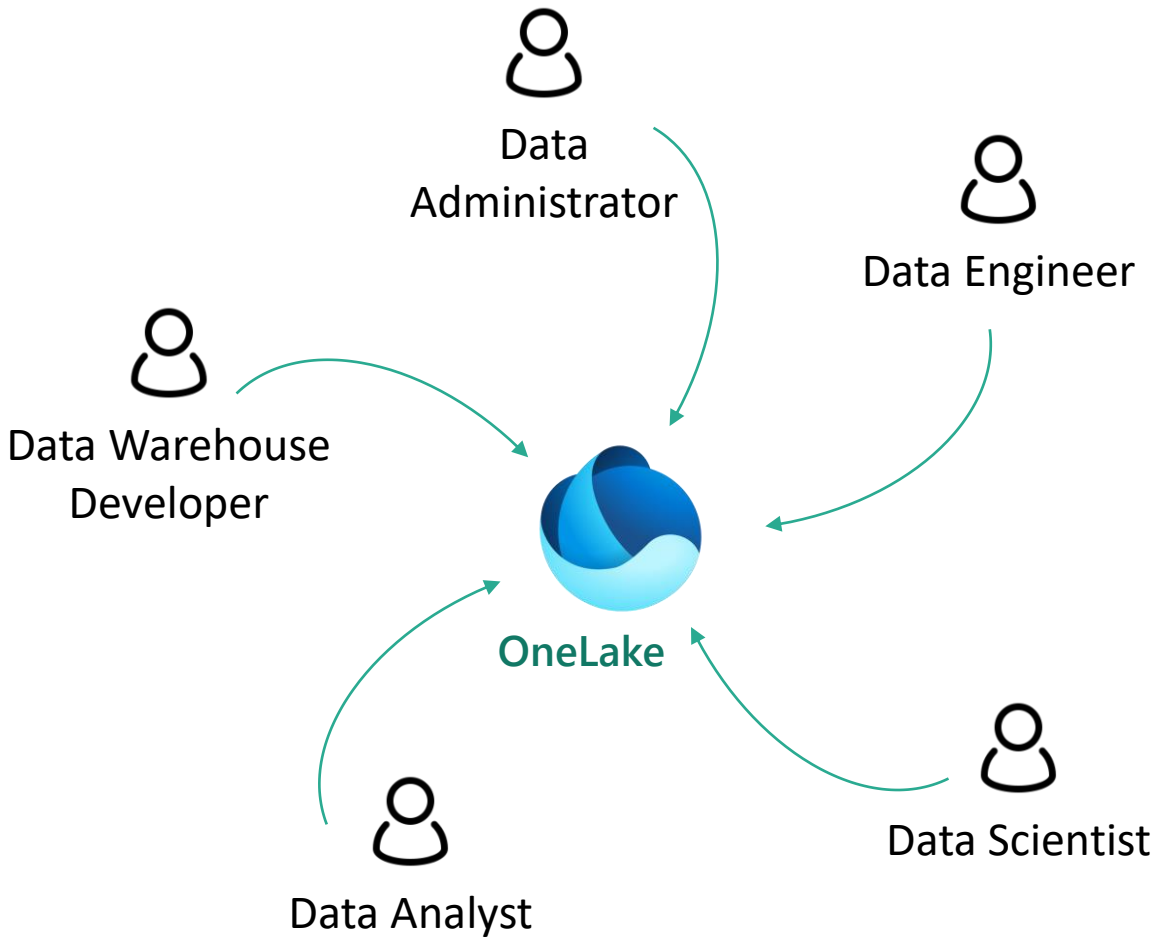
Collaboration

ONELAKE

With OneLake at its core, Microsoft Fabric unifies data disciplines and enhance collaboration across all data professionals.

OneLake both ties together all the tools, experiences and technologies – and by doing so the people working in it.

Never has it been as easy to share ones important and impactful work instantly with the right colleagues.

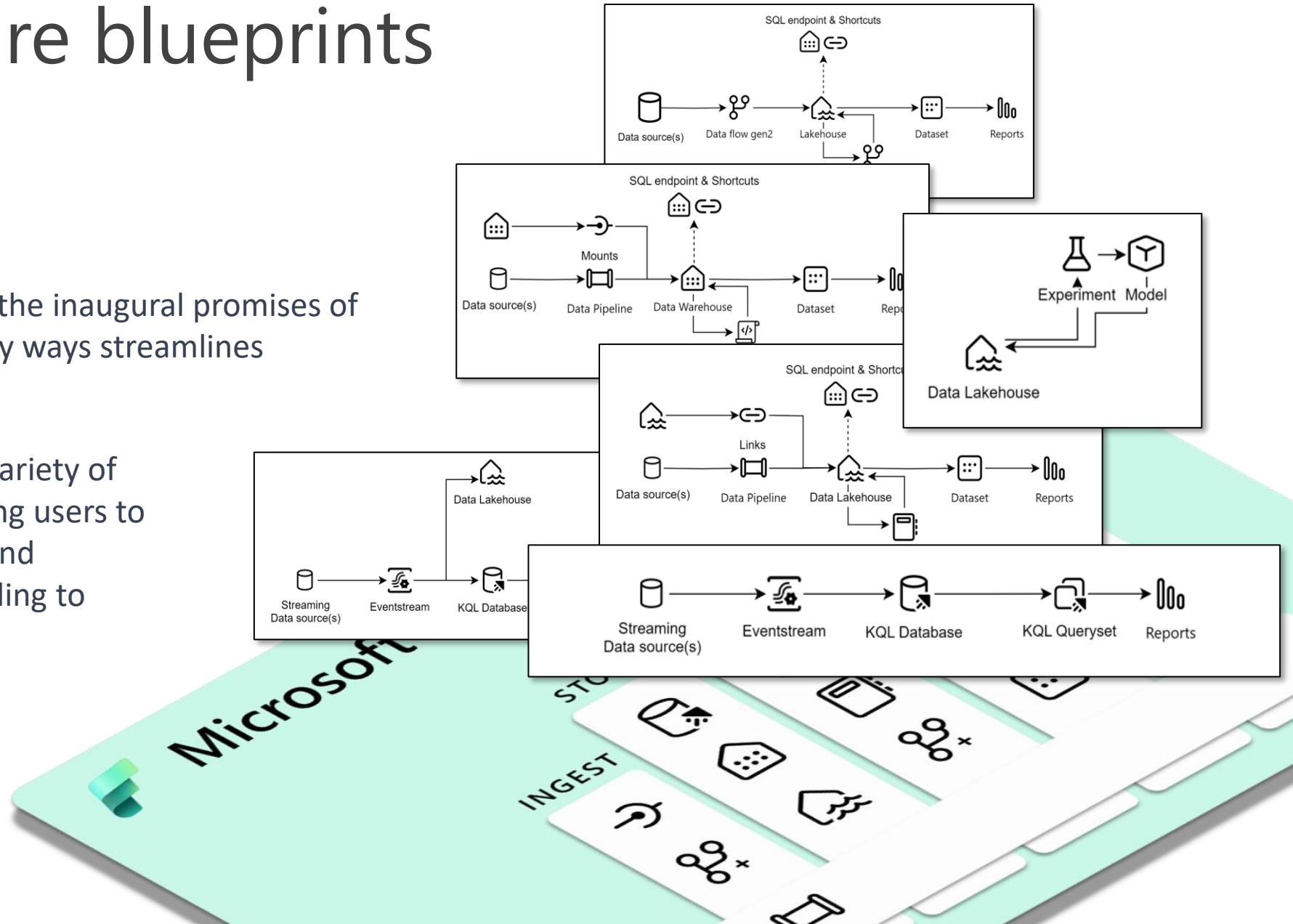


Architecture blueprints

ALL

“One Architecture” is one of the inaugural promises of Microsoft Fabric, that in many ways streamlines architectural complexities.

It does, however, present a variety of options and patterns, enabling users to customize their experience and maximize its potential according to their needs.



Capacity & pricing

CAPACITY | DOMAIN | WORKSPACE

Microsoft Fabric offers a variety of purchasable capabilities divided into SKUs, each providing unique computing power quantified by Capacity Units (CU).

Fabric features two SKU types:

- **Azure** – Billed per second with no commitment.
- **Microsoft 365** – Billed monthly or yearly, with a monthly commitment

SKU*	Capacity Units (CU)	Power BI SKU	Power BI v-cores
F2	2	-	0.25
F4	4	-	0.5
F8	8	EM/A1	1
F16	16	EM2/A2	2
F32	32	EM3/A3	4
F64	64	P1/A4	8
F128	128	P2/A5	16
F256	256	P3/A6	32
F512	512	P4/A7	64
F1024	1024	P5/A8	128
F2048	2048	-	256

Organization

CAPACITY

DOMAIN

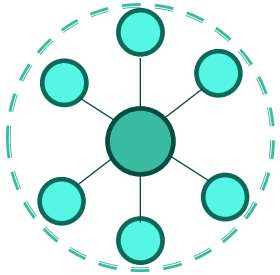
WORKSPACE

Warehouses, Lakehouses, Data Marts, Pipelines and Notebooks. Microsoft Fabrics launched with more new gadgets and technologies than we could have ever dreamed of.

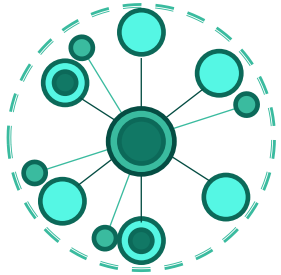
However, it's essential that as organizations, we don't overlook the foundational aspects such as our internal structure, objectives, and strategic planning.

One common organizational decision to consider when deploying a data platform like Fabric, is to choose between a centralized, decentralized, or hybrid implementation approach.

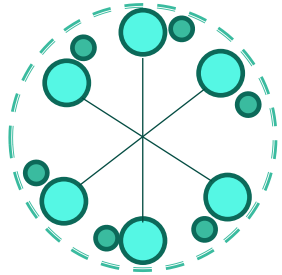
ENTERPRISE



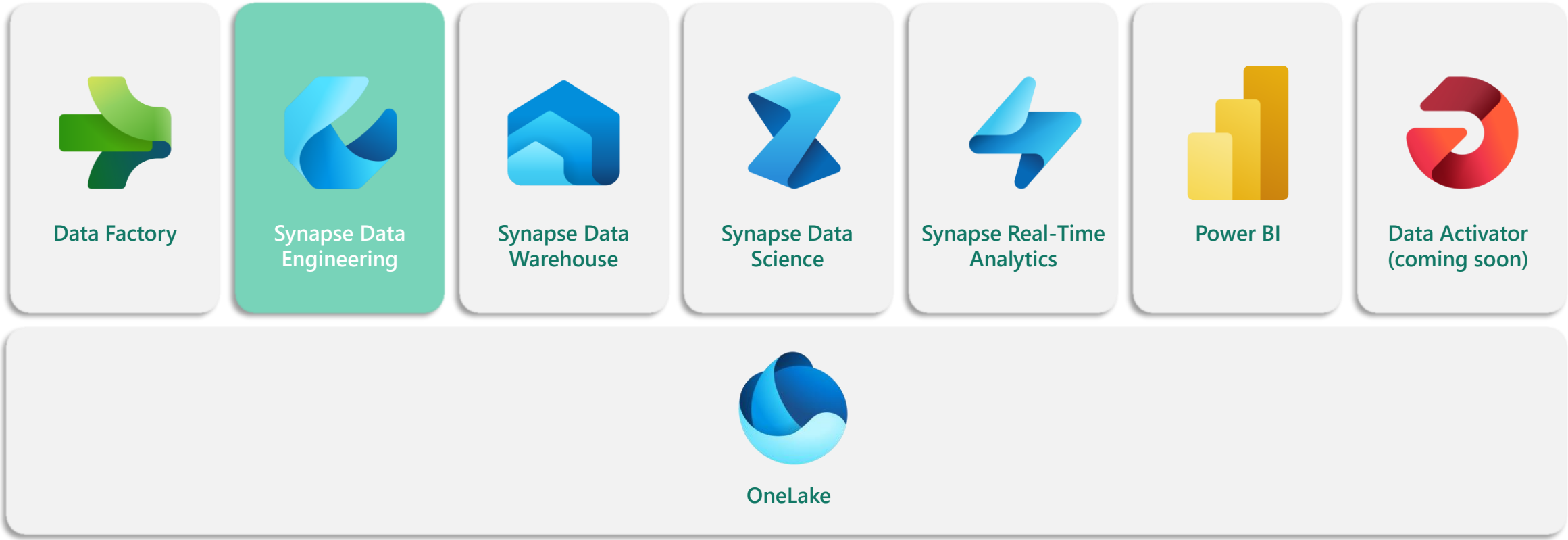
HYBRID



SELF-SERVICE



Microsoft Fabric



Synapse Data Engineering

Empower data engineers to transform data at scale and build a lakehouse architecture as easy as 1-2-3.

Streamline data processes by removing integration complexities. Microsoft Fabric enables the creation of a "lakehouse"—using Delta format—combining data lake and warehouse capabilities.

Users can then interact with the data using SQL endpoint, Power BI, or powerful Spark pools, with optimized performance.

TOOLS

- LAKEHOUSE
- NOTEBOOK
- SPARK JOB DEFINITION
- DATA PIPELINES

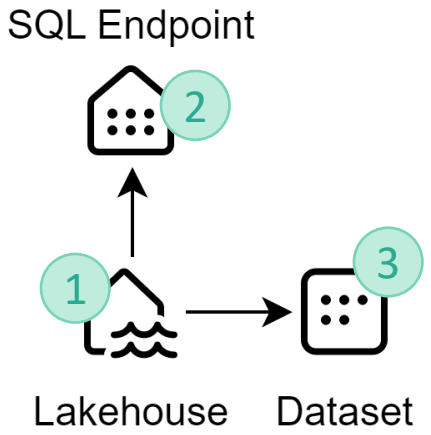
Data Lakehouse in Microsoft Fabric

LAKEHOUSE

The Lakehouse in Microsoft Fabric, introduced as part of Synapse Data Engineering, is a unified data platform that integrates the best of data lakes and warehouses.

This fusion makes it easier for data engineers to ingest, transform, and share data in an open Delta Lake format. The Lakehouse is a versatile storage solution where data can be brought in through dataflow, pipelines, or even shortcuts to create virtual folders and tables.

The Lakehouse facilitates collaboration among different data professionals by providing a SQL endpoint for data warehousing functions and a semantic dataset for building BI reports. It even allows Power BI to connect directly to the lakehouse data for efficient data reading.



Deploying a lakehouse, creates 3 linked artifacts:

1. *The lakehouse itself*
2. *SQL endpoint*
3. *Auto-generated dataset*

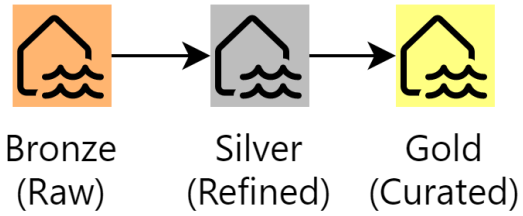
Medallion architecture

LAKEHOUSE

NOTEBOOK

A **Medallion architecture** is a design pattern in data architecture for logically organizing data into layers with a specified step-wise improvement of data quality and structure.

A medallion architecture is typically divided into 3-4 layers:



"By using a medallion architecture, understanding and troubleshooting failing data pipelines becomes much easier"

Lars Kristensen
Architect, Fellowmind

Data quality libraries

WORKSPACE | NOTEBOOK | LIBRARIES

With workspace-level library management, using libraries like **Great Expectations** or **PyDeequ** could enhance data quality management.

Using Python libraries readily available to us, we ensure that our decisions are based on reliable and trustworthy data that meet the criteria of the six dimensions of data quality:

```
from pydeequ.checks import *
from pydeequ.verification import *

check = Check(spark, CheckLevel.Warning, "Review Check")

checkResult = VerificationSuite(spark) \
    .onData(df) \
    .addCheck(
        check.hasSize(lambda x: x >= 3) \
        .hasMin("b", lambda x: x == 0) \
        .isComplete("c")  \
        .isUnique("a")  \
        .isContainedIn("a", ["foo", "bar", "baz"]) \
        .isNonNegative("b")) \
    .run()

checkResult_df = VerificationResult.checkResultsAsDataFrame(spark, checkResult)
checkResult_df.show()
```

Constraint verification in PyDeequ

1. **Accurate** – Accurate data mirrors the real world, like correctly recording names or factual data
2. **Complete** – Complete data refers to having all essential data for a specific use
3. **Unique** – Uniqueness relates to the absence of duplicate data
4. **Consistent** – Consistency is about ensuring data values don't contradict within or across datasets
5. **Timely** – Timeliness indicates data availability when needed
6. **Valid** – Validity means data meets the expected format and range

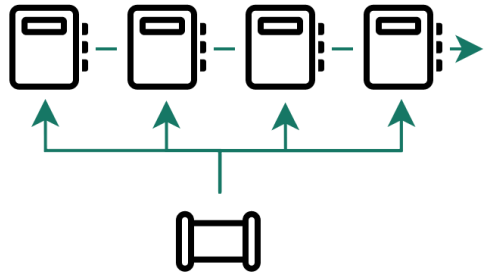
Notebooks in Data Engineering

NOTEBOOK

Notebooks empower users to leverage Apache Spark, the most potent data processing framework to date, thus maximizing the potential of data engineering tasks.

Notebooks offer a versatile environment for data engineering and data analysis, allowing users to install libraries, encouraging unbounded exploration beyond the UI. This code-first approach, while offering full flexibility, does call for structured use to avoid clutter. The tool aids transparency and promotes good commenting practices, providing a clear view of data pipeline operations.

They encourage chunking code into modular, atomic, and auditable steps, endorsing efficient management and traceability. This balance between flexibility and structure enhances readability and cultivates good coding habits.



*Orchestrating **notebooks** with **data pipelines** is a great way to add more structure and control to a complex workflow.*

Fact:
Like in Microsoft Office, notebooks auto-save, making sure valuable work isn't lost

Get started today

Try Microsoft Fabric

Try Fabric (microsoft.com)

Watch Fellowmind's monthly Power BI Update

Power BI Update (fellowmindcompany.com)

Connect with our Microsoft Data Platform MVPs

https://www.linkedin.com/in/mhalkjaer

https://www.linkedin.com/in/brianbonk