

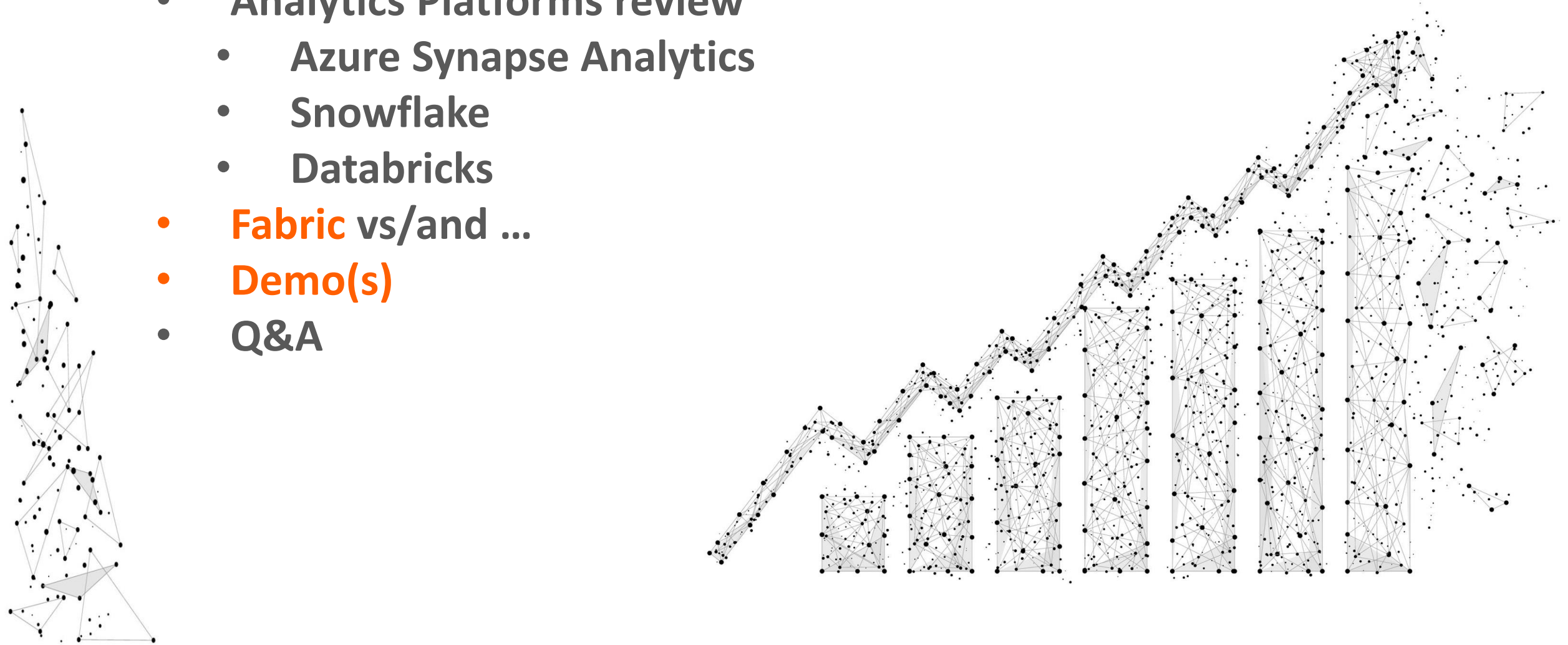
Po **MSBuild 2023** - czyli co **Fabric** zmienia w podejściu do tworzenia rozwiązań analitycznych?

Tomasz Krawczyk
Future Processing - Data Solutions



PLAN

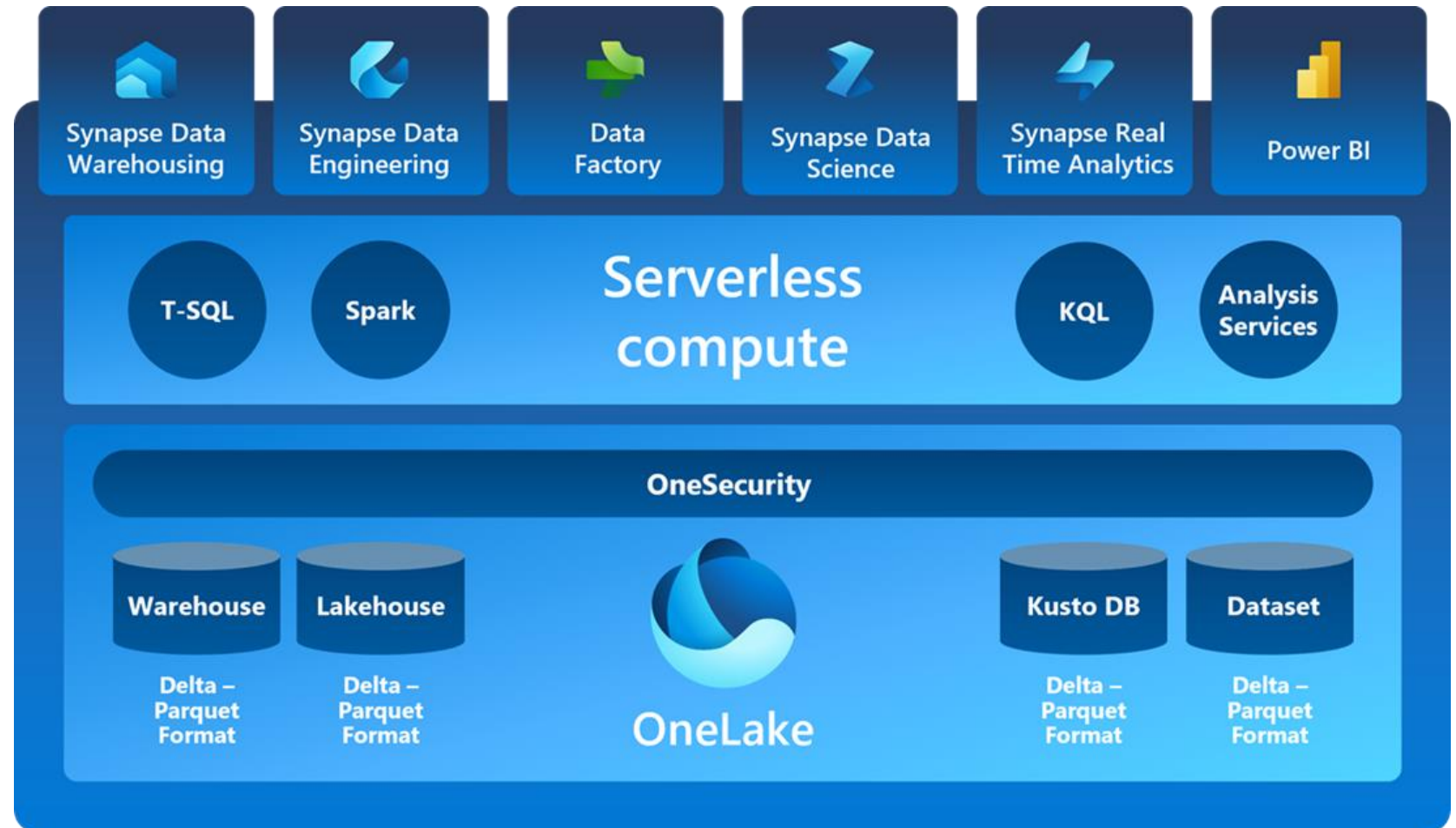
- **Microsoft Fabric** –quick recap
- Analytics Platforms review
 - Azure Synapse Analytics
 - Snowflake
 - Databricks
- **Fabric** vs/and ...
- **Demo(s)**
- Q&A



Microsoft Fabric

Microsoft Fabric is an all-in-one enterprise analytics solution that covers everything from data movement to data science, real-time analytics, and business analytics.

Microsoft Fabric combines new and existing components from Power BI, Azure Synapse, and Azure Data Explorer into one integrated environment.



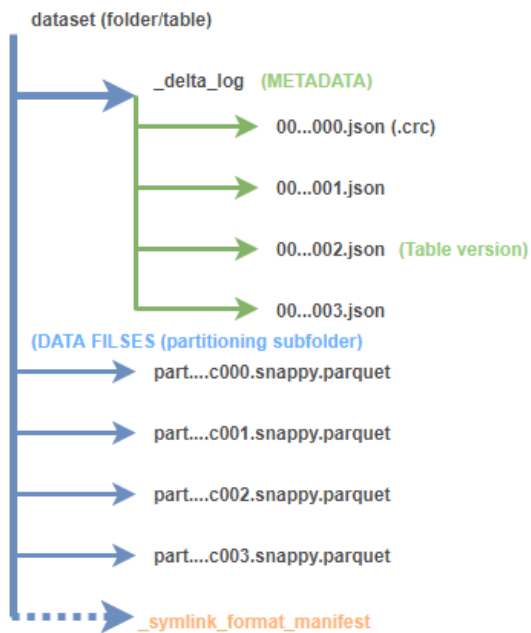
Microsoft Fabric - key features

OneLake is open at every level. Built on top of Azure Data Lake Storage Gen2, OneLake can support any type of file, structured or unstructured.

All Fabric data items like data warehouses and lakehouses store their data automatically in OneLake in **delta parquet format**.



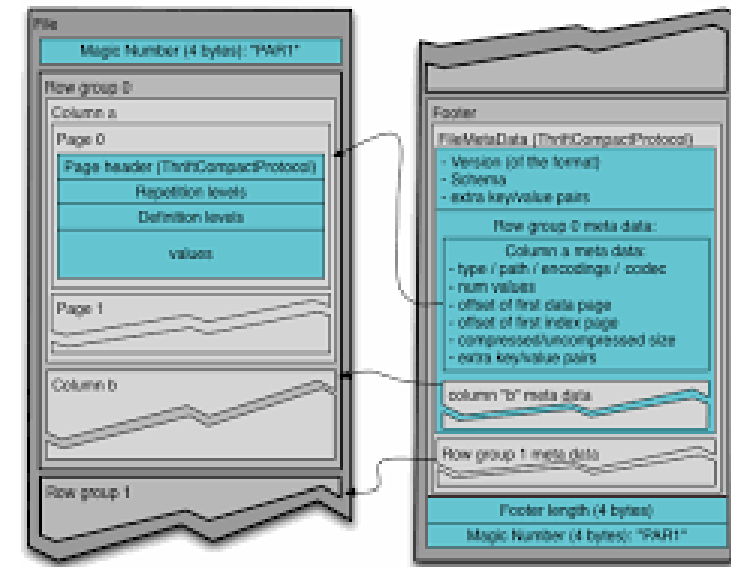
Microsoft Fabric - key features Delta Format



Location: data / DemoDWH / DemoDWH / DWH / part

Search blobs by prefix (case-...

Name		
<input type="checkbox"/>	[-.]	...
<input type="checkbox"/>	_delta_log	...
<input type="checkbox"/>	_symlink_format_manifest	...
<input type="checkbox"/>	Par=1	...
<input type="checkbox"/>	Par=2	...
<input type="checkbox"/>	Par=3	...
<input type="checkbox"/>	Par=4	...

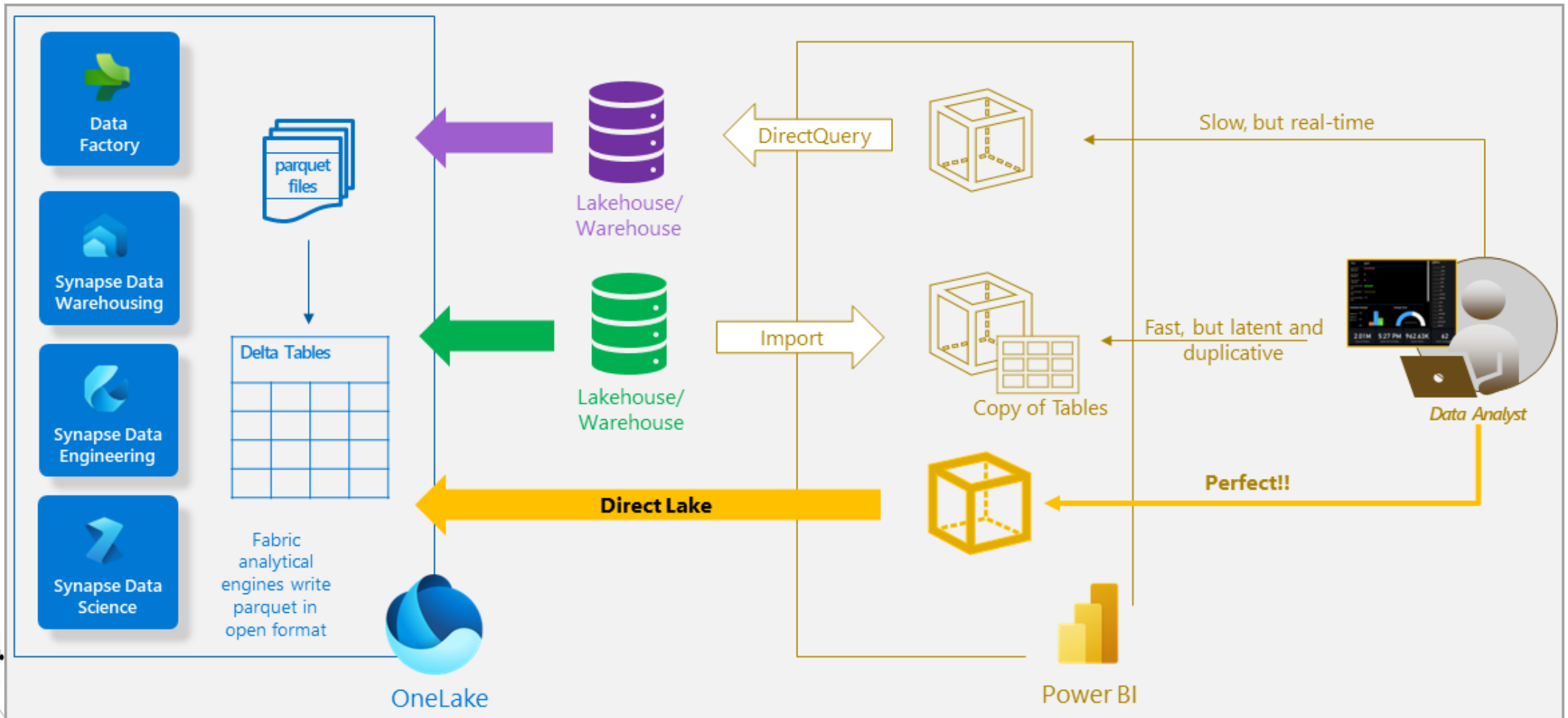


Parquet Format

```
{ "commitInfo": { "timestamp": 1603978502578, "userId": "3792484595433007", "userName": "tkrawczyk@future-processing.com", "operation": "UPDATE", "operationParameters": { "predicate": { "remove": { "path": "part-00000-d4d9f7da-2720-410e-8fd1-1f1f93ffc6f6-c000.snappy.parquet", "deletionTimestamp": 1603978501061, "dataChange": true } }, "add": { "path": "part-00000-91042731-ac59-401f-acb2-59b2bfad21f0-c000.snappy.parquet", "partitionValues": {}, "size": 3698, "modificationTime": 1603978502000, "dataChange": true, 
```

<https://delta.io/>

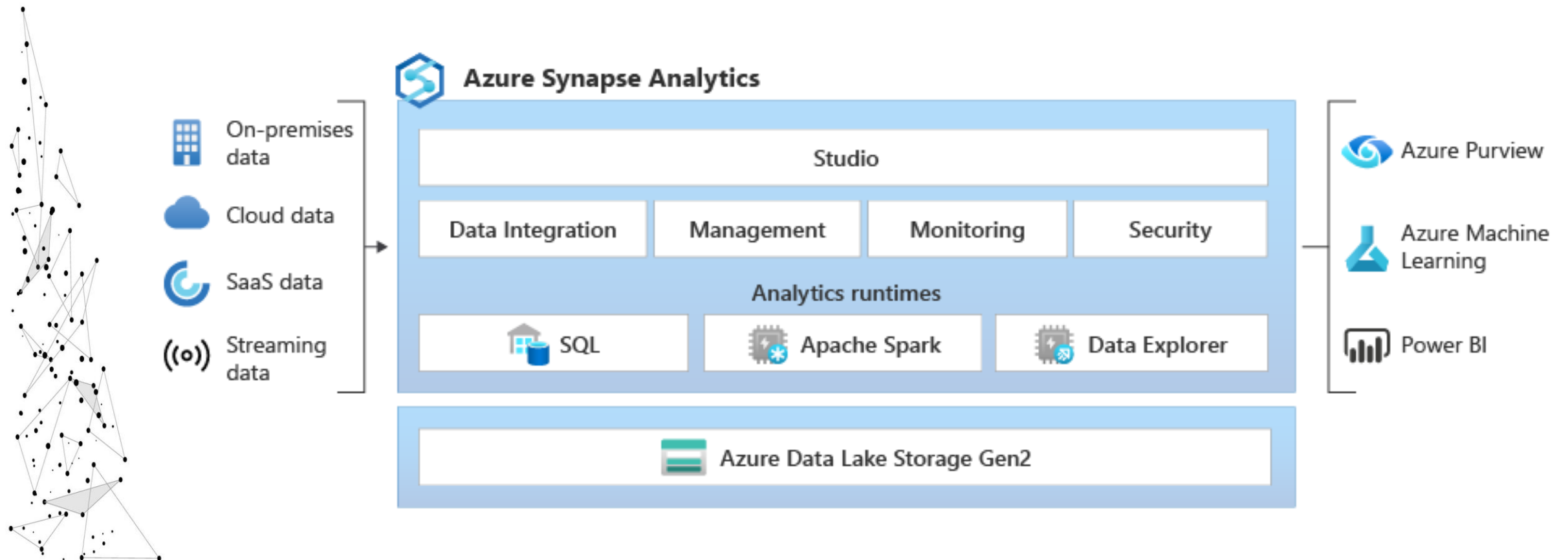
Microsoft Fabric – key features Direct Lake



<https://learn.microsoft.com/en-us/power-bi/enterprise/directlake-overview>

Microsoft Synapse Analytics

Azure Synapse is an enterprise analytics service that accelerates time to insight across data warehouses and big data systems.

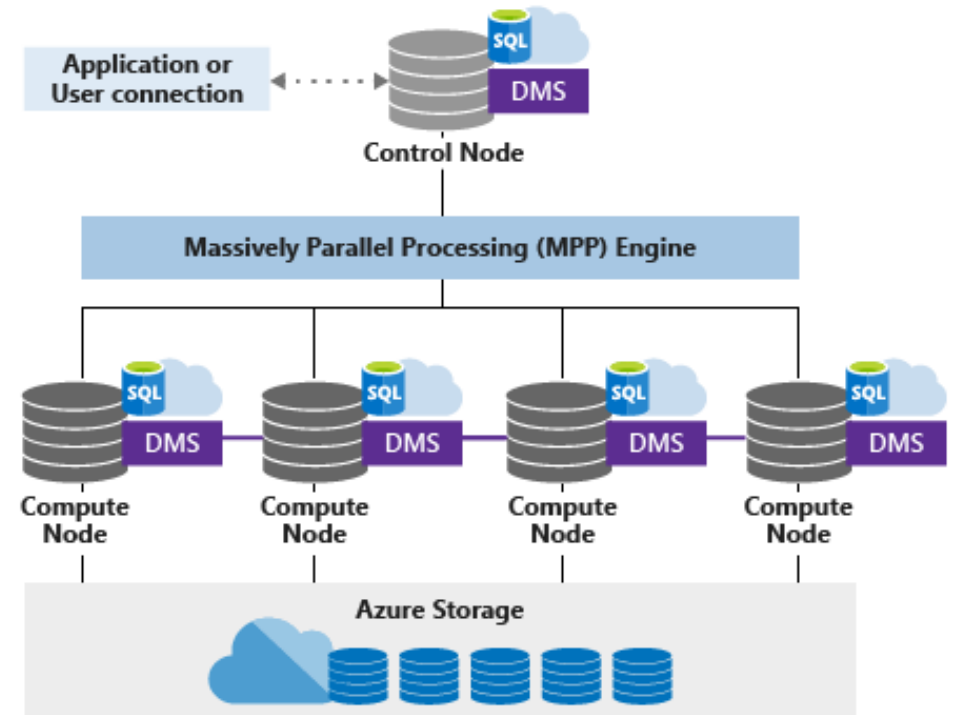


Microsoft Synapse Analytics Dedicated Pool

Azure Synapse Dedicated SQL Pool is a distributed, parallel processing data warehouse that is designed to handle large-scale analytical workloads.

Dedicated SQL Pool provides a massively parallel processing (MPP) architecture that distributes and processes data across multiple compute nodes. This distributed processing capability enables high-performance query execution and scalability to handle large volumes of data. It allows you to load, store, and analyze data from various sources using familiar T-SQL (Transact-SQL) language and tools.

Dedicated SQL pool

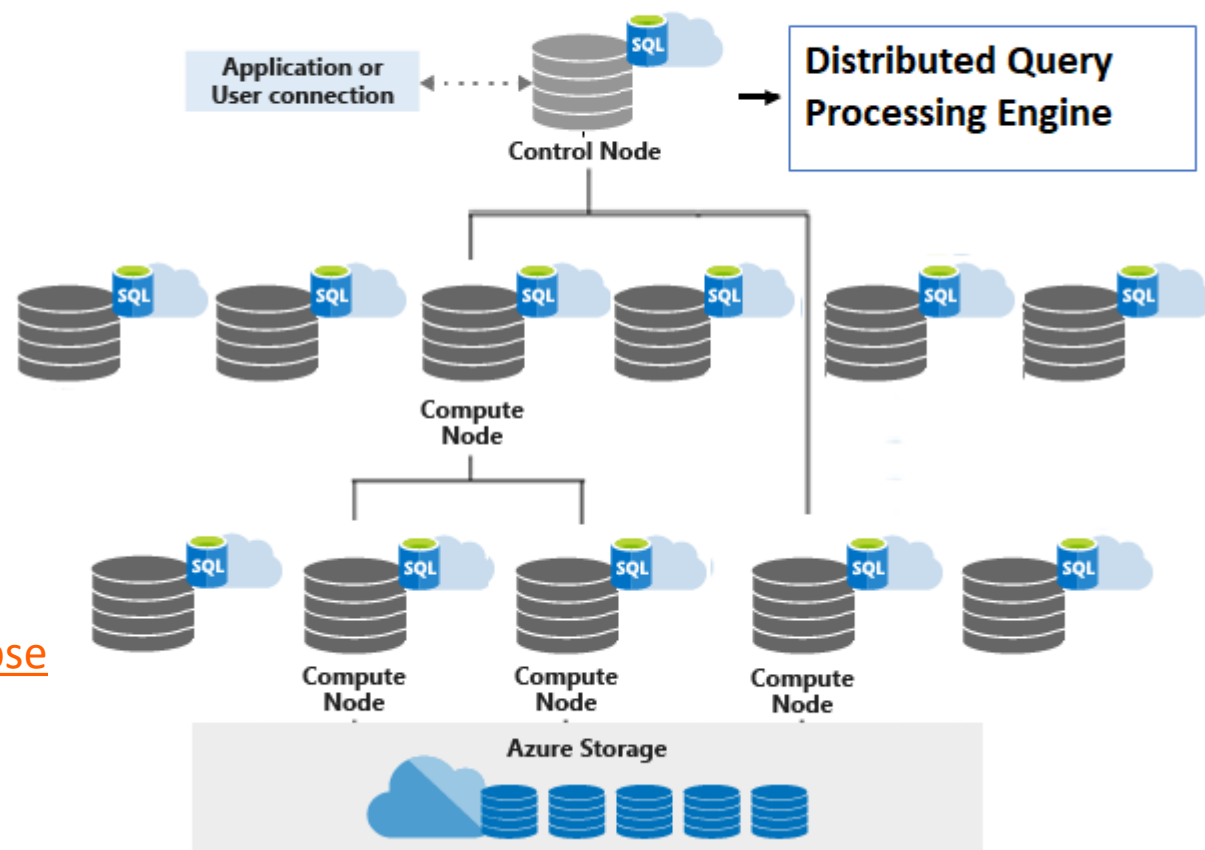


Microsoft Synapse Analytics Serverless Pool

The **serverless SQL** pool Control node utilizes Distributed Query Processing (DQP) engine to optimize and orchestrate distributed execution of user query by splitting it into smaller queries that will be executed on Compute nodes. Each small query is called task and represents distributed execution unit. It reads file(s) from storage, joins results from other tasks, groups, or orders data retrieved from other tasks.


[Polaris - The Distributed SQL Engine in Azure Synapse](#)

Serverless SQL pool



Microsoft Synapse Analytics Pools

DEDICTED POOL

- 
- Requires upfront provisioning of resources
 - Manual scaling based on provisioned resources
 - Provisioned pricing model
 - Fixed set of allocated compute resources
 - Pay for provisioned resources
 - **Predictable and steady workloads**
 - Limited flexibility in resource allocation
 - Determined by provisioned resources
 - Can store and query large volumes of data
 - More control and management overhead
 - Data stored in internal data warehouse
 - Supports caching of query results
 - Supports all “traditional” T-SQL features

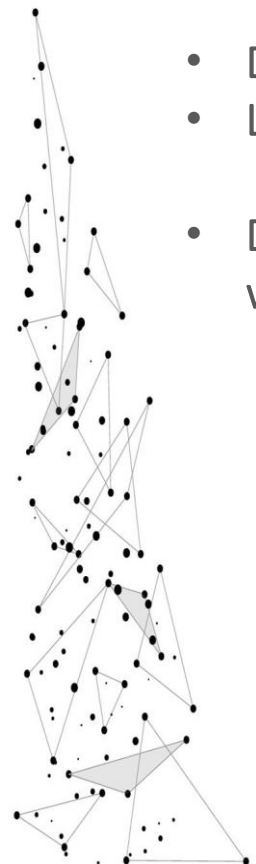
SERVERLESS POOL

- No upfront provisioning required
- Auto-scaling based on workload demand
- Consumption-based pricing model
- Automatically allocated resources
- Pay for data processed during queries
- **Sporadic or unpredictable workloads**
- Flexible scaling and resource allocation
- Cost optimized based on data processed
- Can query data stored in various formats
- Less management overhead, automatic scaling
- Data stored outside Synapse (ADLS) in open formats
- Doesn't support result set caching
- Supports most all “traditional” T-SQL features
 - Doesn't support UPDATE, DELETE, INSERT(*)


Microsoft Synapse Analytics Pools

LIMITATIONS

DEDICATED POOL

- 
- Doesn't support cross databases queries
 - Limit support for external sources
 - EXTERNAL TABLES (no support for DELTA format)
 - Doesn't support "multi warehouses" –problem with workload management
 - Workload Classification
 - Workload Importance
 - Workload Isolation

SERVERLESS POOL

- 
- Doesn't support UPDATE, DELETE, INSERT(*)
 - Unpredictable workloads - nondeterministic queries execution time
 - Limited optimization mechanisms
 - (no indexes)

Databricks and Snowflake



- 2013 (Spark 2009)
- PaaS
- **Databricks** is built on Apache Spark's distributed computing framework, making management of infrastructure easier. Databricks is a data lake rather than a data warehouse, with emphasis more on use cases such as streaming, machine learning, and data science-based analytics. Databricks can be used to handle raw unprocessed data in large volume, and can run on AWS, Azure, and Google clouds.
- Analysts, data scientists and data engineers



- 2012
- SaaS (DaaS)
- **Snowflake** uses a SQL engine to manage information stored in the database. It processes queries against virtual warehouses, each one in its own independent cluster node. On top of that can sit cloud services for authentication, infrastructure management, queries, and access controls. Snowflake enables users to analyze and store data using Amazon S3 or Azure resources.
- Data analysts

Snowflake

Snowflake is an analytic data warehouse provided as Software-as-a-Service (SaaS). Snowflake provides a data warehouse that is faster, easier to use, and far more flexible than traditional data warehouse offerings.

Unique Offerings

- Scalability (Storage and Compute)
- No complex management (performance)
 - No indexing
 - No performance tuning
 - No partitioning
 - No Physical Storage Design

■ Pay as you use

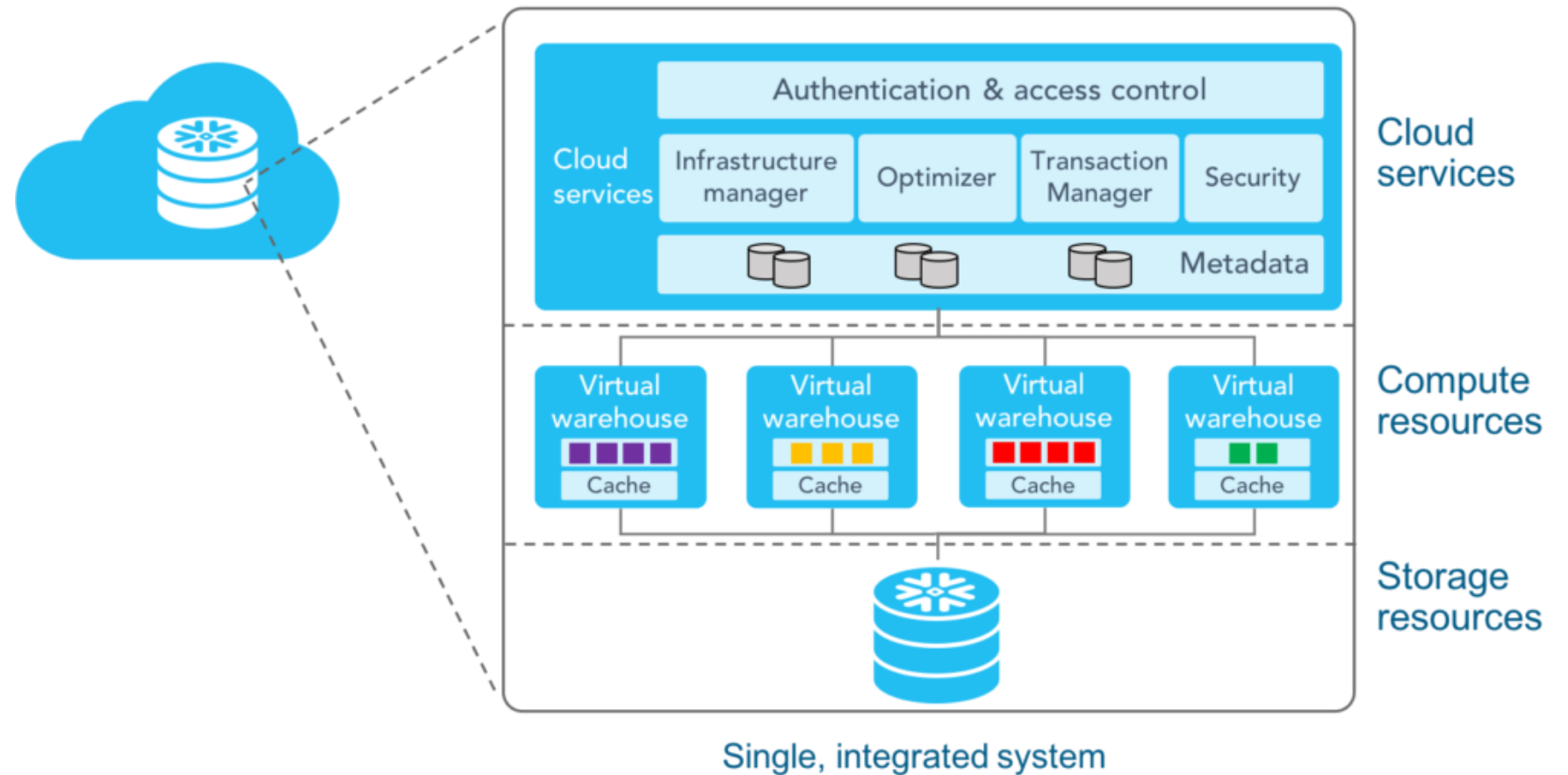
- **Azure/AWS/GCP (hosting environment)**
- **Azure/AWS/GCP integration (stage, snowpipe, ..)**
- **ODBC/JDBC support**
- **Scale up/out (online)**
- **Zero Cloning Copy**
- **Time Travel**
- **Result Caching**
- **Data sharing**
- **Security (RBAC, Data encryption)**
- **Semi-structured data support**
- **SQL (Merge, Copy Into, Materialized view, SP)**
- **External Tables**



<https://docs.snowflake.net/manuals/user-guide/intro-key-concepts.html>

Snowflake Architecture

Snowflake Multi-Cluster Shared Data Architecture



Snowflake Architecture Storage

Snowflake automatically converts all data stored into an optimized **immutable compressed columnar format** (Micro-Partitions –one partition contains 50M -500MB uncompressed data) and encrypts it using AES-256 strong encryption.

Partition Metadata:

- The range of values for each of the columns in the micro-partition.
- The number of distinct values.
- Additional properties used for both optimization and efficient query processing.

Logical Structure

type	name	country	date
2	A	UK	11/2
4	C	SP	11/2
3	C	DE	11/2
2	B	DE	11/2
3	A	FR	11/2
2	C	SP	11/2
3	Z	DE	11/2
2	B	UK	11/2
4	C	NL	11/2
5	X	FR	11/3
1	A	NL	11/3
5	A	FR	11/3
2	X	FR	11/2
4	Z	NL	11/2
2	Y	SP	11/2
1	B	SP	11/3
5	X	DE	11/3
3	A	UK	11/4
1	C	FR	11/3
4	Z	NL	11/4
5	Y	SP	11/4
5	B	SP	11/5
3	X	DE	11/5
2	Z	UK	11/5

Table: t1

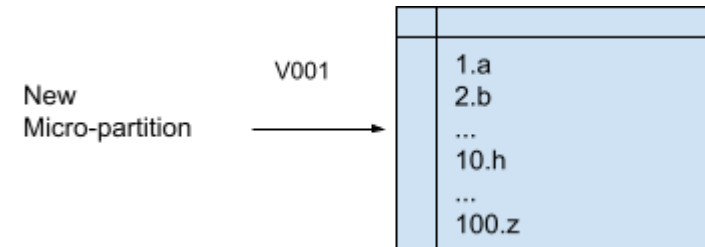
Physical Structure

	Micro-partition 1 (rows 1-6)	Micro-partition 2 (rows 7-12)	Micro-partition 3 (rows 13-18)	Micro-partition 4 (rows 19-24)
type	2 4 3 2 3 2	3 2 4 5 1 5	2 4 2 1 5 3	1 4 5 5 3 2
name	A C C B A C	Z B C X A A	X Z Y B X A	C Z Y B X Z
country	UK SP DE DE FR SP	DE UK NL FR NL FR	FR NL SP SP DE UK	FR NL SP SP DE UK
date	11/2 11/2 11/2 11/2 11/2 11/2	11/2 11/2 11/2 11/3 11/3 11/3	11/2 11/2 11/2 11/3 11/3 11/4	11/3 11/4 11/4 11/5 11/5 11/5

Snowflake Architecture Storage

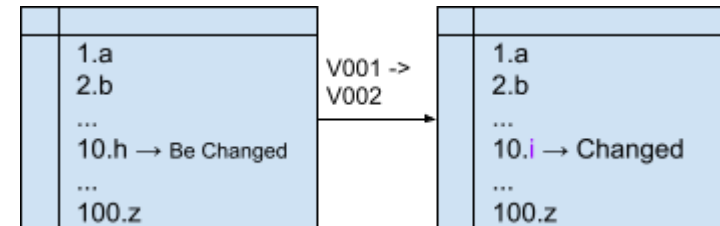
INSERT/COPY

INSERT and COPY into table operations only create new micro-partitions.



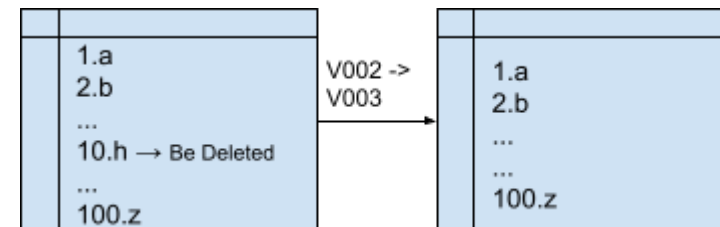
UPDATE

UPDATE operations keep the old MPs (before the change) and create new MPs with the change. Each MP will have its own unique version IDs.



DELETE

DELETE operations keep the old MPs (before the delete) and create new MPs with the change by removing the record(s). Each MP will have its own unique version IDs.



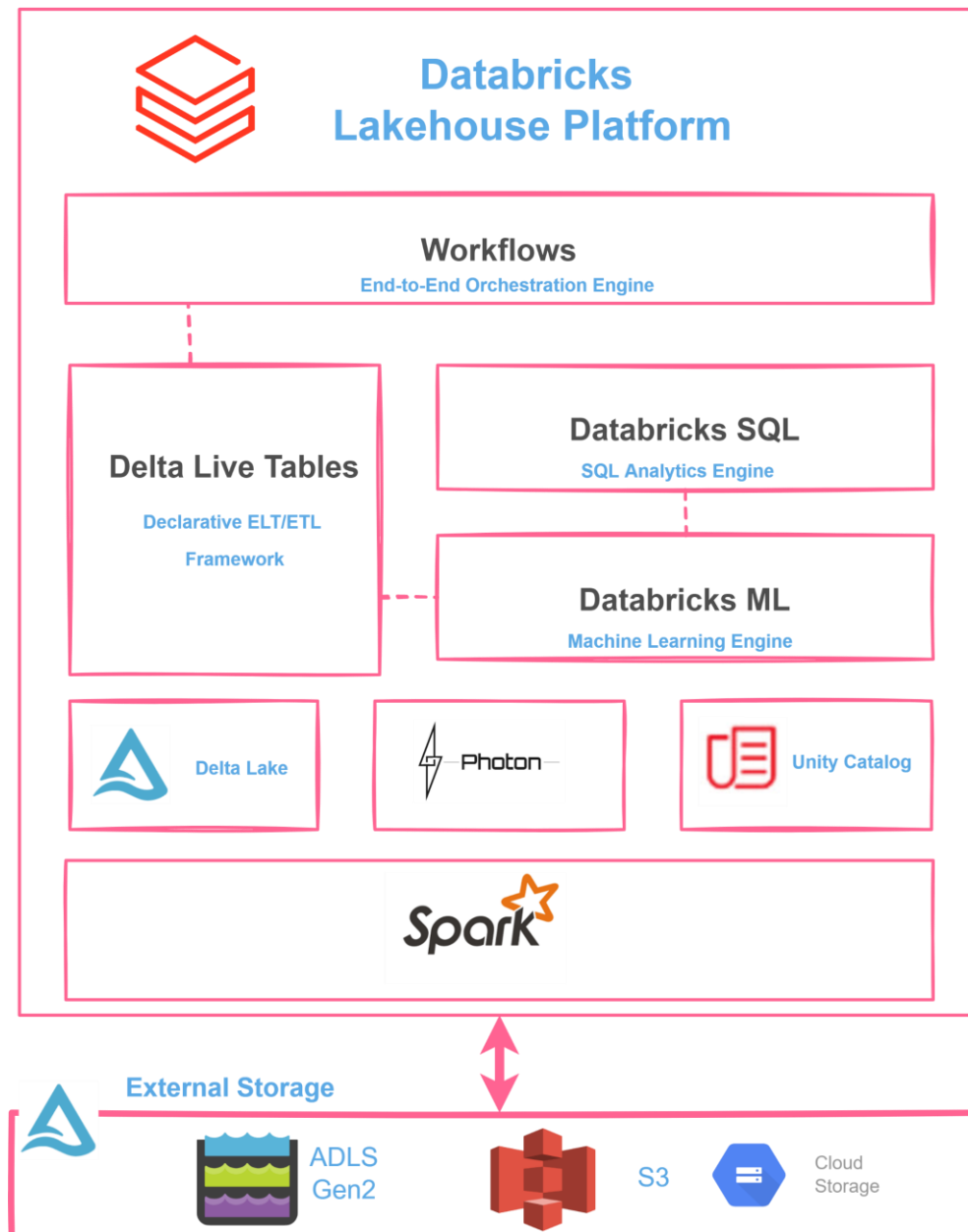
Databricks

DATA LAKEHOUSE = DATA WAREHOUSE + DATA LAKE

One data architecture for **BI**, **SQL**, **Streaming** and **Machine Learning**

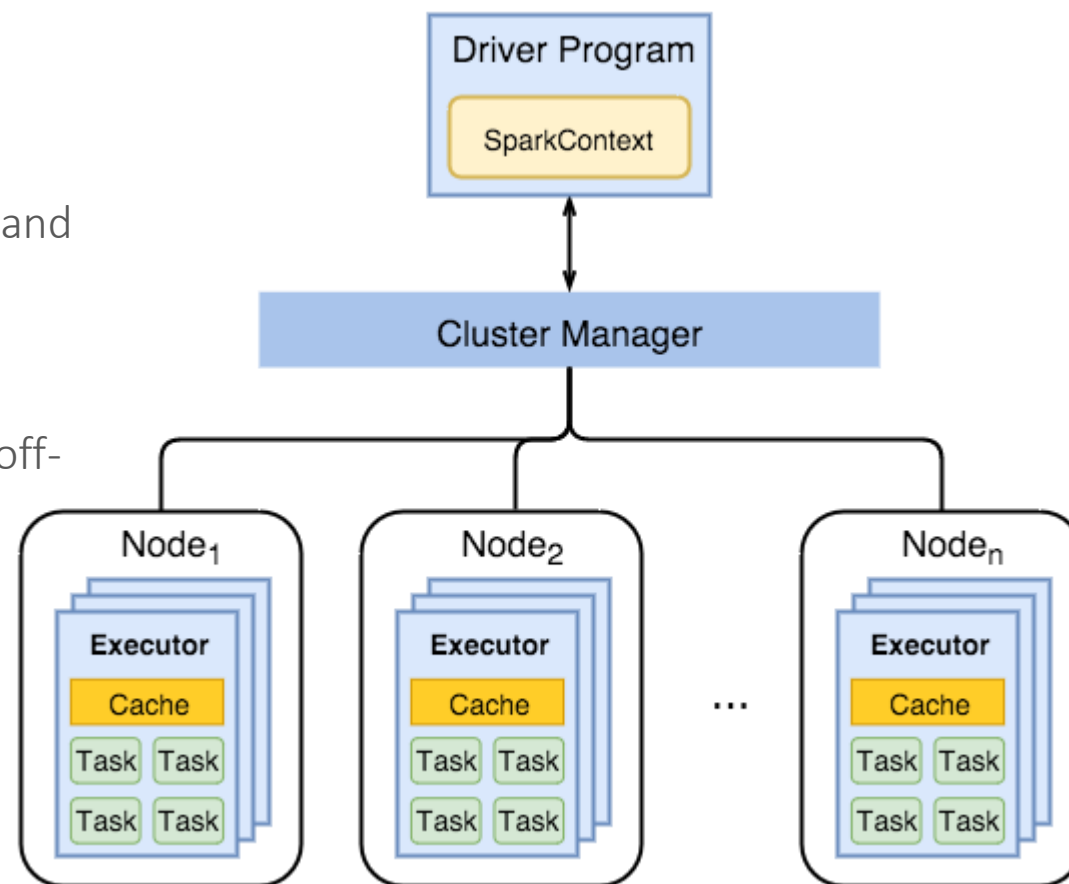
Why Lakehouse:

- Unifies **data warehousing** and **machine learning** (ML)
- Build on open source and open standards
- Reduces cost
- Simplifies data governance
- Simplifies ETL jobs
- Removes data redundancy
- **Enables direct data access**
- Connects directly to BI tools

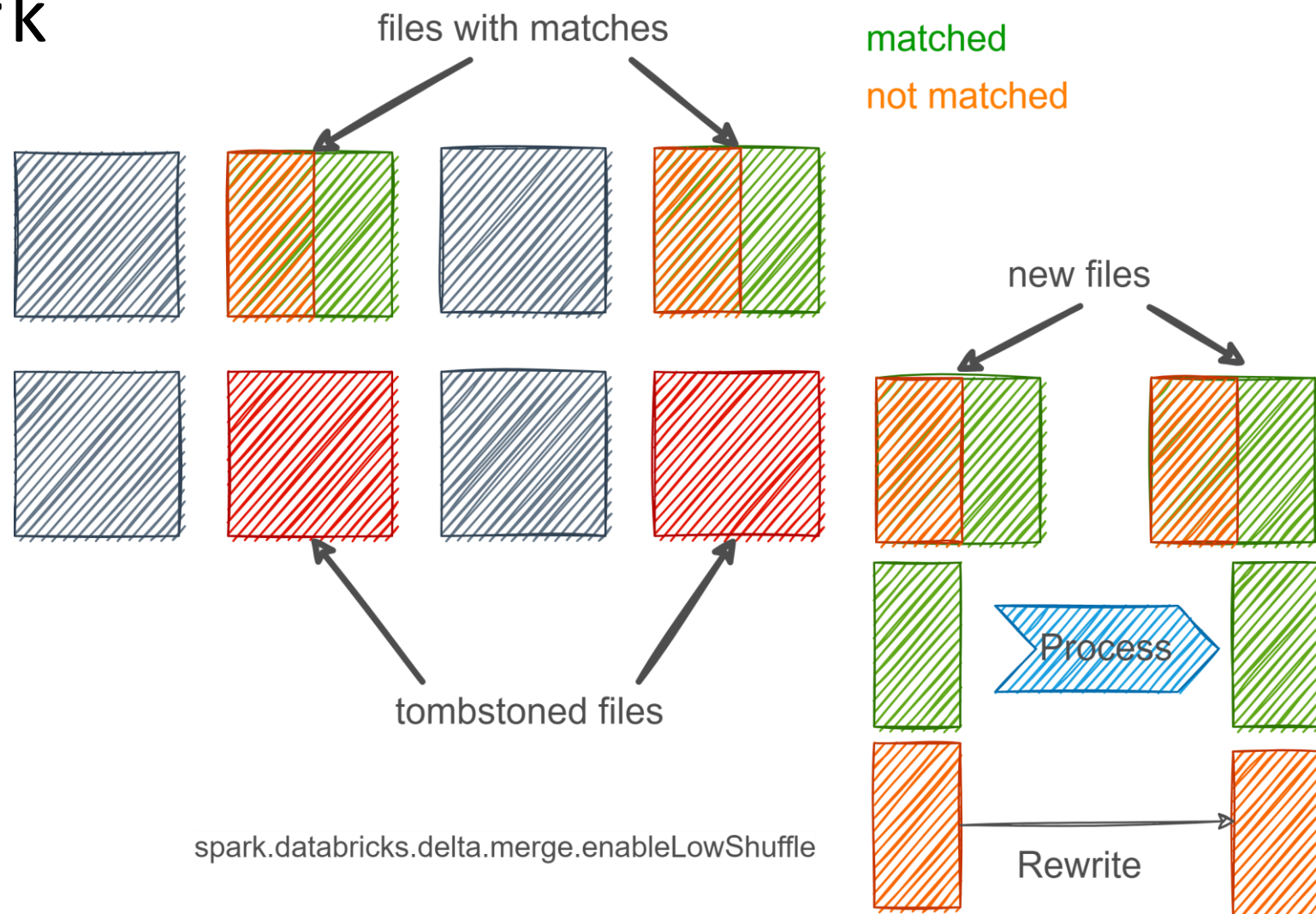
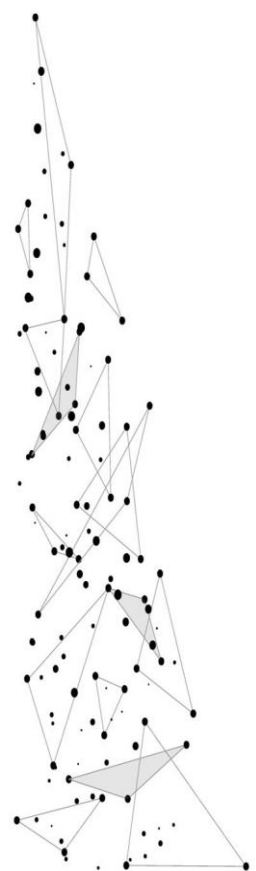


Spark Architecture

- Spark Driver
 - separate process to execute user applications
 - creates SparkContext to schedule jobs execution and negotiate with cluster manager
- Executors
 - run tasks scheduled by the driver
 - store computation results in memory, on disk or off-heap
 - interact with storage systems
- Cluster Manager
 - Mesos
 - YARN
 - Spark Standalone



Databricks Architecture Storage Merge Spark

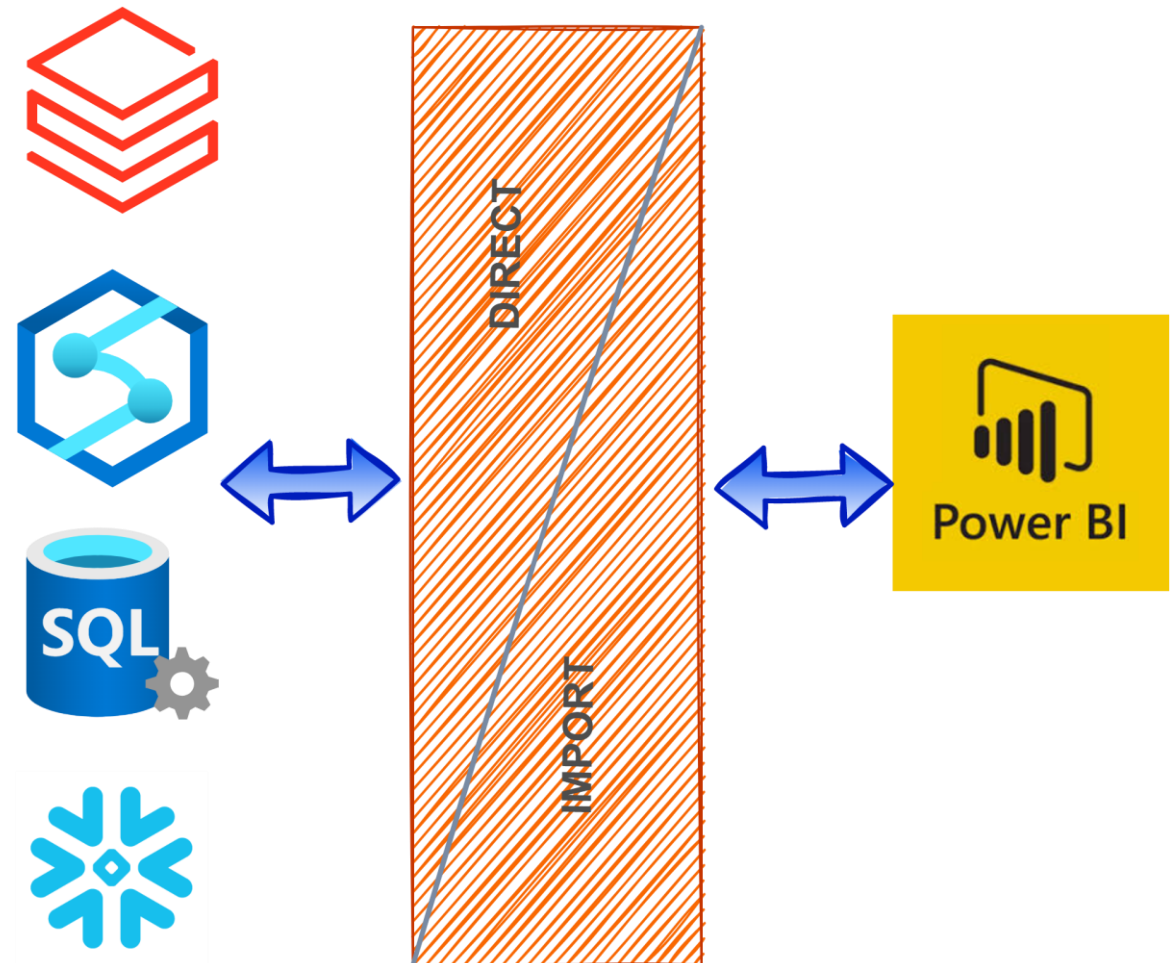


Reporting Layer

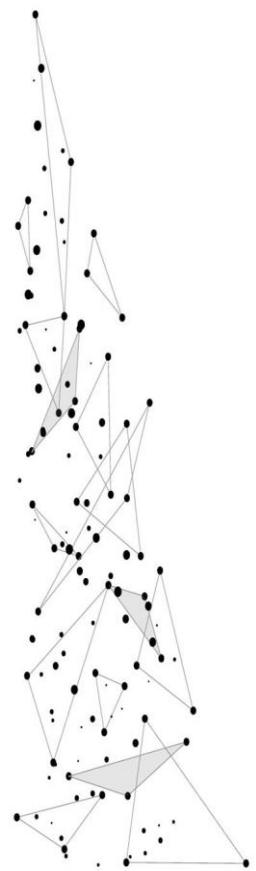
Direct Lake is based on loading parquet-formatted files **directly from a data lake** without having to query a Lakehouse endpoint, and without having to import or **duplicate data** into a **Power BI** dataset.

Direct Lake is a **fast-path** to load the data from the lake straight into the **Power BI** engine, ready for analysis.

Current State:



What's next?



Open format wins

Jun 9, 2023

1D 5D 1M 6M YTD 1Y 5Y MAX

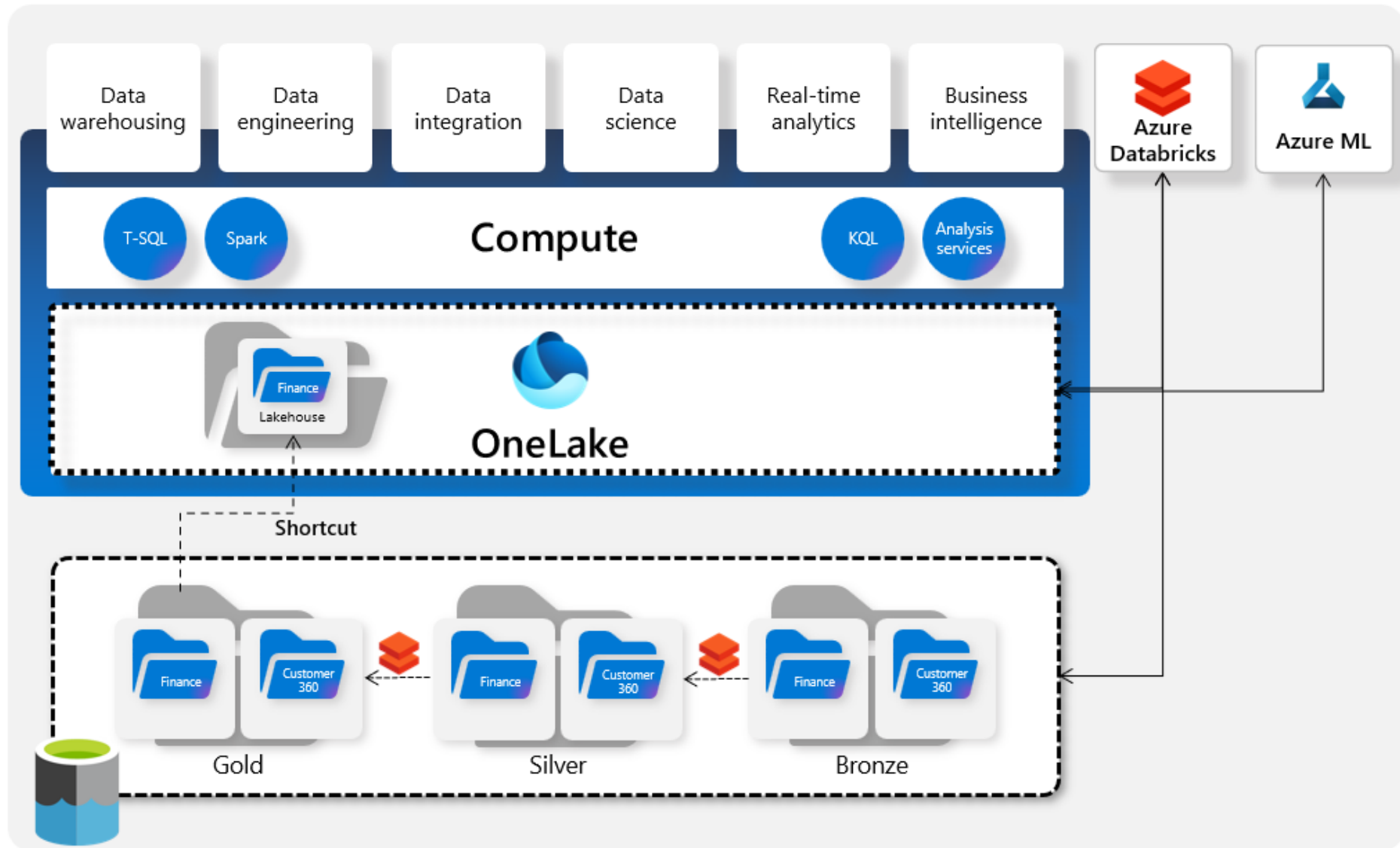


Snowflake

Microsoft Build

2023

Fabric And Databricks



Fabric Access to Storage

Shortcuts are objects in **OneLake** that point to other **storage locations**. The location can be **internal** or **external** to OneLake.

External shortcuts:

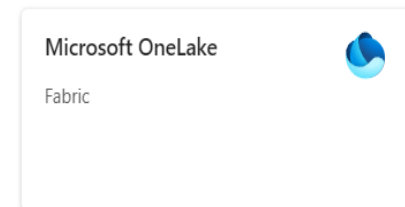
- ADLS shortcuts
- S3 shortcuts

OneLake access via APIs

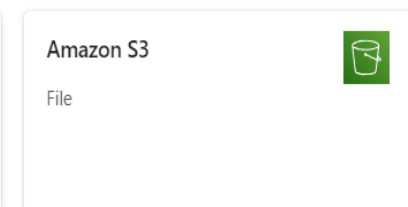
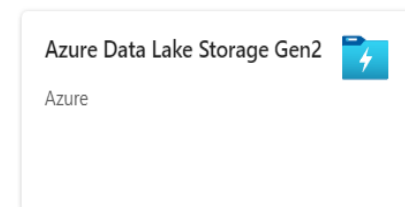
Microsoft OneLake provides open access to all of your Fabric items through existing **ADLS Gen2 APIs** and SDKs.

<https://onelake.dfs.fabric.microsoft.com/<workspace>/<item>.<itemtype>/<path>/<fileName>>

Internal sources



External sources



Fabric – Power BI Direct Lake Mode

Direct Lake is supported on **Power BI Premium P** and Microsoft Fabric F SKUs only. It's not supported on Power BI Pro, Premium Per User, or Power BI Embedded A/EM SKUs.

Refresh

By default, data changes in OneLake are automatically reflected in a Direct Lake dataset. You can change this behavior by disabling **Keep your Direct Lake data up to date** in the dataset's settings.

Refresh

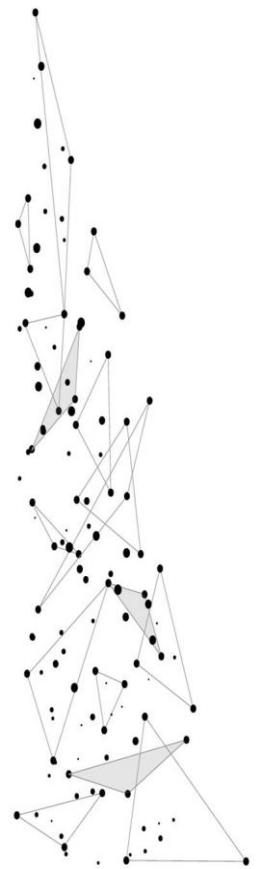
Keep your Direct Lake data up to date

Configure Power BI to detect changes to the data in OneLake and automatically update the Direct Lake tables that are included in this dataset. [Learn more](#)



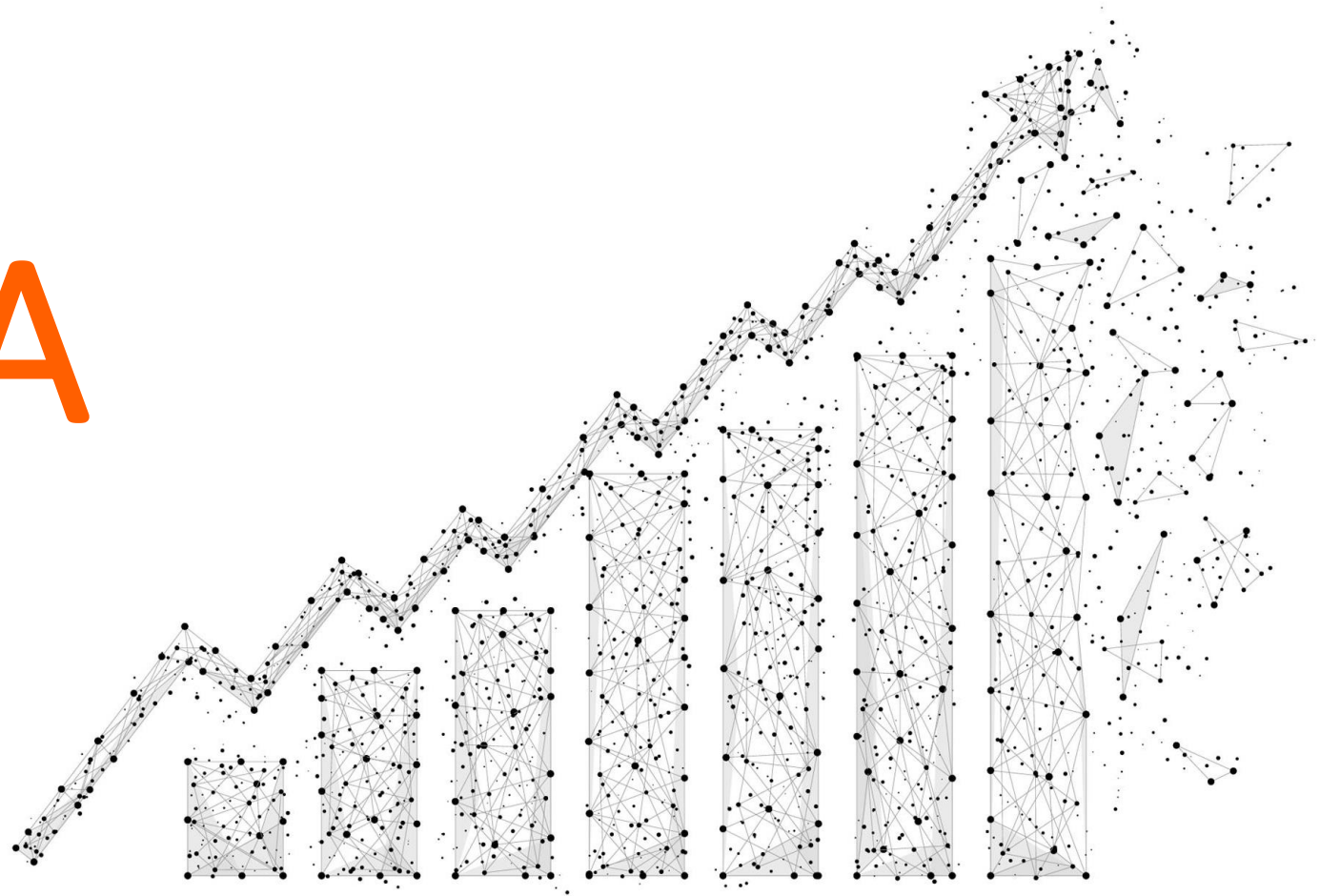
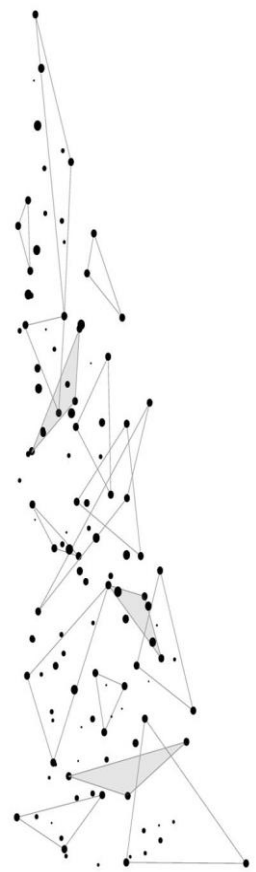
You may want to disable if, for example, you need to allow completion of data preparation jobs before exposing any new data to consumers of the dataset. When disabled, you can invoke refresh manually or by using the refresh APIs. Invoking a refresh for a Direct Lake dataset is a low cost operation where the dataset analyzes the metadata of the latest version of the Delta Lake table and is updated to reference the latest files in the OneLake.

<https://learn.microsoft.com/en-us/power-bi/enterprise/directlake-overview#refresh>



DEMO TIME

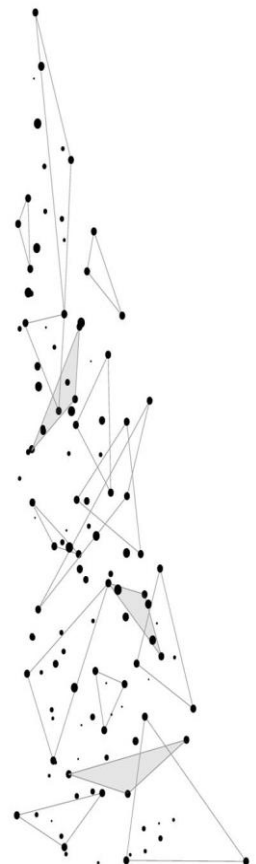
Q & A



Resources

Examples (and demos)

- <https://www.youtube.com/watch?v=ZYfAhNjmaOQ>
- **Blogs, pages, documentation, articles**
- <https://www.microsoft.com/en-us/microsoft-fabric>
- <https://learn.microsoft.com/en-us/power-bi/enterprise/directlake-overview>



THANK YOU!

tkrawczyk@future-processing.com

tomasz.k.krawczyk@gmail.com

