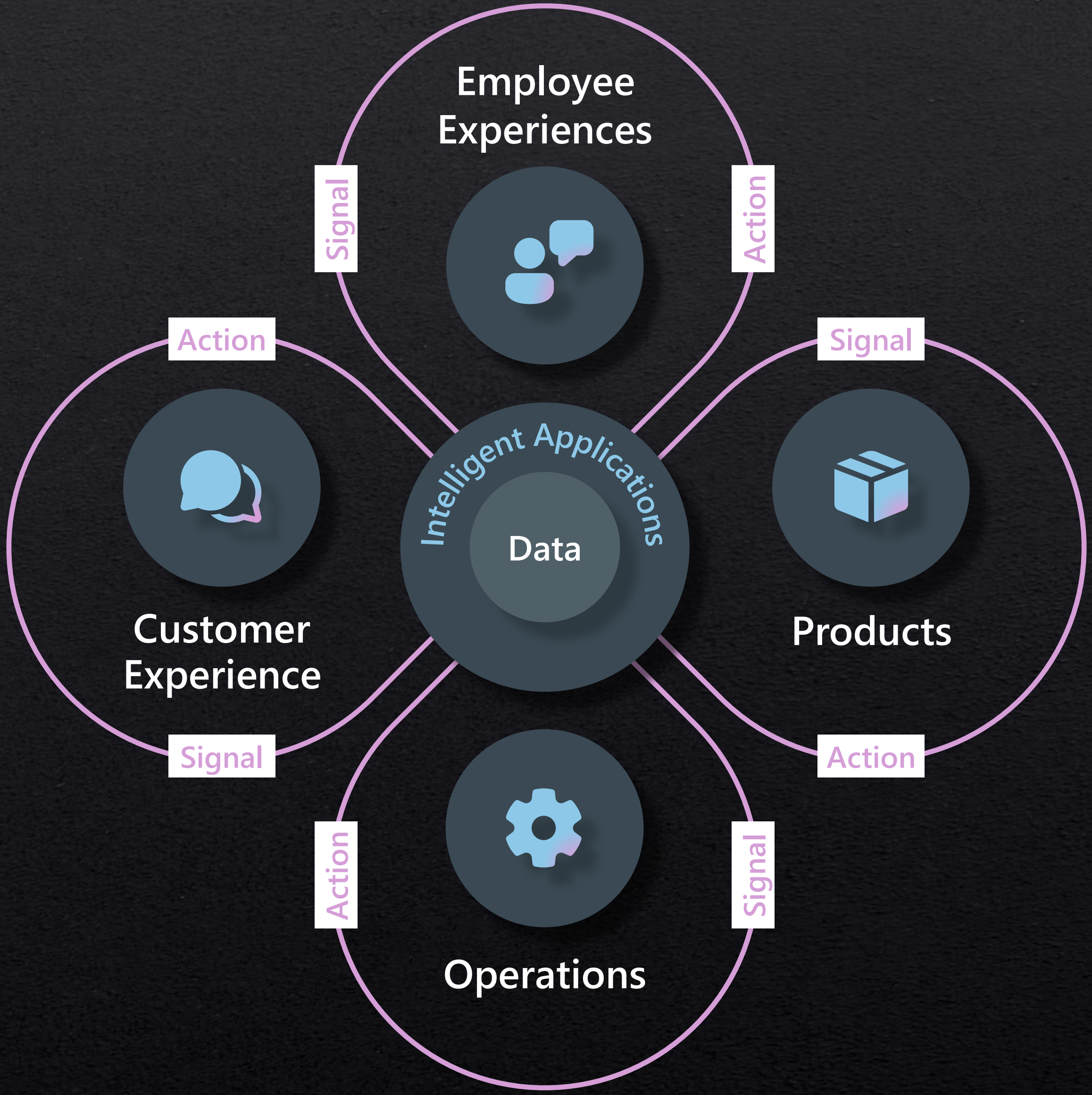
Cloud Scale Analytics with Azure Databricks and Microsoft Fabric Hands-on Workshop

Owais Hashmi, Principal - Global Black Belt Team
Microsoft
https://www.linkedin.com/in/owaish/

14/09/2023

Data is the oxygen of digital transformation



Driving analytics modernization is the key to unlocking value from your data

Key Data Trends

Data & AI will add $13 trillion in global GDP by 2030

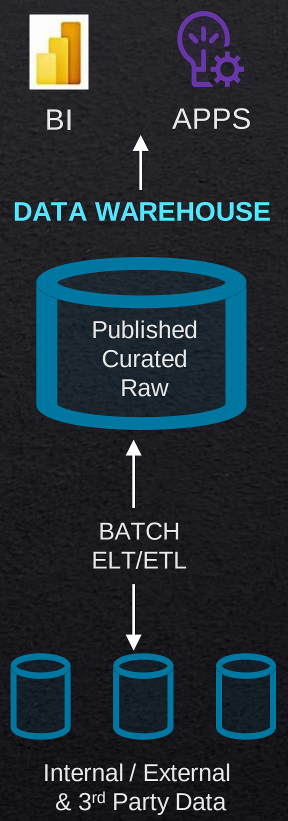
Data silos = Lack a robust data estate

Fragmented tool sets for Analytics & AI

Source: McKinsey & Co,

Lakehouse Evolution

Late 1980s – Mid 2000s



BI

APPS

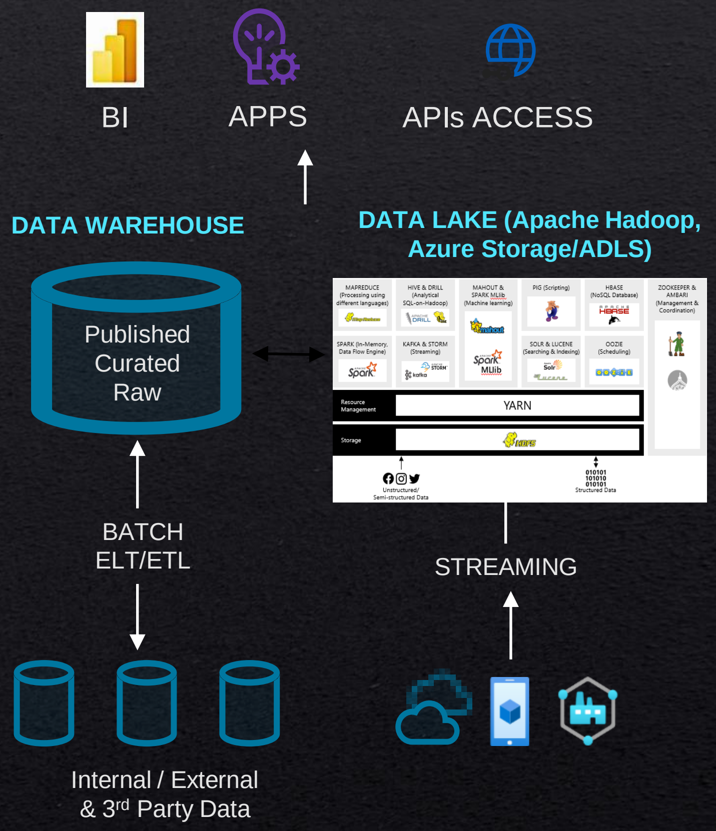
DATA WAREHOUSE

Published
Curated
Raw

BATCH
ELT/ETL

Internal / External
& 3rd Party Data

Mid 2000s – 2020



BI

APPS

APIs ACCESS

DATA WAREHOUSE

Published
Curated
Raw

BATCH
ELT/ETL

Internal / External
& 3rd Party Data

DATA LAKE (Apache Hadoop, Azure Storage/ADLS)

MAPREDUCE (Processing using different languages)

HIVE & DRILL (Analytical SQL-on-Hadoop)

MAHOUT & SPARK MLlib (Machine learning)

PIG (Scripting)

HBASE (NoSQL Database)

ZOOKEEPER & AMBARI (Management & Coordination)

SPARK (In-Memory, Data Flow Engine)

KAFKA & STORM (Streaming)

SOLR & LUCENE (Searching & Indexing)

OOZIE (Scheduling)

Resource Management: YARN

Storage: HDFS

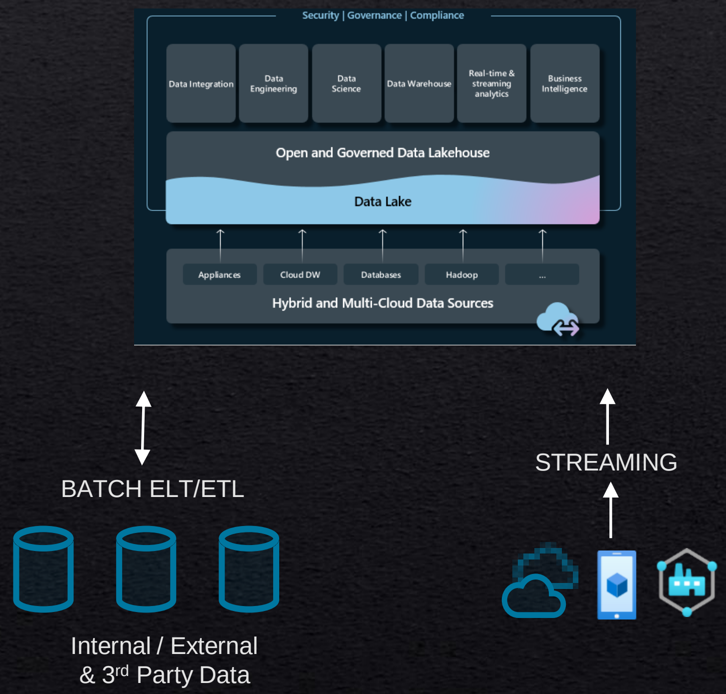
Unstructured/ Semi-structured Data

010101
101010
010101
Structured Data

STREAMING

2021 - Present

LAKEHOUSE – Powered by Delta Lake, Iceberg, and Hudi



Security | Governance | Compliance

Data Integration

Data Engineering

Data Science

Data Warehouse

Real-time & streaming analytics

Business Intelligence

Open and Governed Data Lakehouse

Data Lake

Appliances

Cloud DW

Databases

Hadoop

...

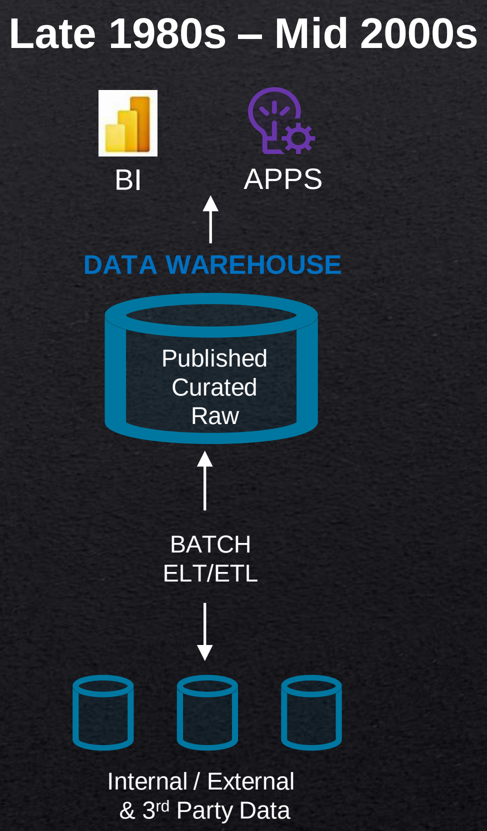
Hybrid and Multi-Cloud Data Sources

BATCH ELT/ETL

STREAMING

Internal / External
& 3rd Party Data

Lakehouse Evolution



Use Cases

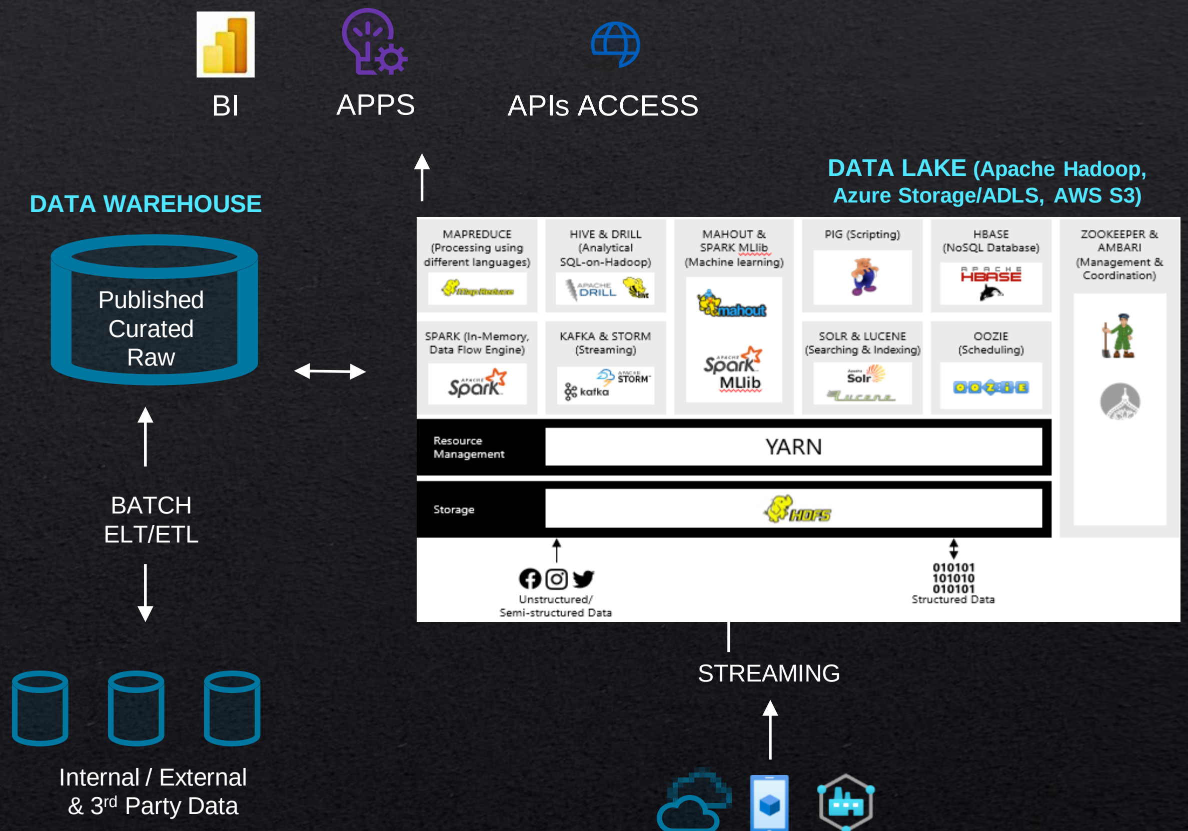
- Aggregate data from OLTP databases and other sources
- Single source of truth
- Business Intelligence applications
- High concurrency / Low latency requirements
- Supports ACID transactions
- Optimize queries (Indexes, Materialized Views, etc.)
- Governance and Auditing

Limitations

- Limited support for ML/AI use cases
- Limited support for Streaming use cases
- Limited support for Data Types/formats
- Expensive to store Petabyte scale data
- Inelastic compute (solved by cloud-based DW)
- Vendor lock-in

Lakehouse Evolution

Mid 2000s – 2020



BI

APPS

APIs ACCESS

DATA WAREHOUSE

Published
Curated
Raw

BATCH
ELT/ETL

Internal / External
& 3rd Party Data

DATA LAKE (Apache Hadoop, Azure Storage/ADLS, AWS S3)

MAPREDUCE (Processing using different languages)	HIVE & DRILL (Analytical SQL-on-Hadoop)	MAHOUT & SPARK MLlib (Machine learning)	PIG (Scripting)	HBASE (NoSQL Database)	ZOOKEEPER & AMBARI (Management & Coordination)
SPARK (In-Memory, Data Flow Engine)	KAFKA & STORM (Streaming)		SOLR & LUCENE (Searching & Indexing)	OOZIE (Scheduling)	

Resource Management: YARN

Storage: HDFS

Unstructured/
Semi-structured Data

010101
101010
010101
Structured Data

STREAMING

Use Cases

- Enabled ML/AI use cases
- Enabled Streaming use cases
- Support for all Data Types
- Cheap storage for Petabyte scale data
- Distributed computing & Fault-tolerant
- Decouple Compute from Storage

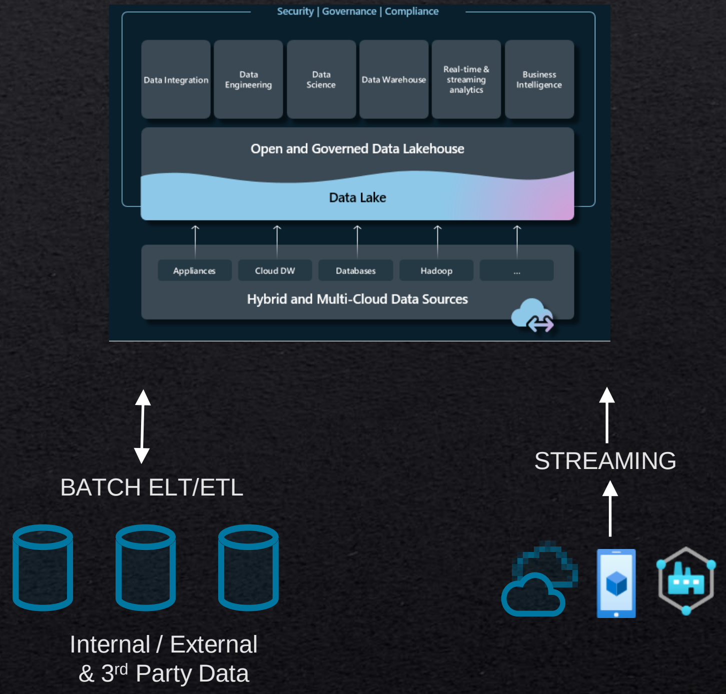
Limitations

- Limited support for ACID transactions
- Slow query response time
- Data quality issues
- Too many files
- Large metadata handling issues
- Inelastic (solved by cloud object stores)
- Complex to manage and monitor

Lakehouse Evolution

2021 - Present

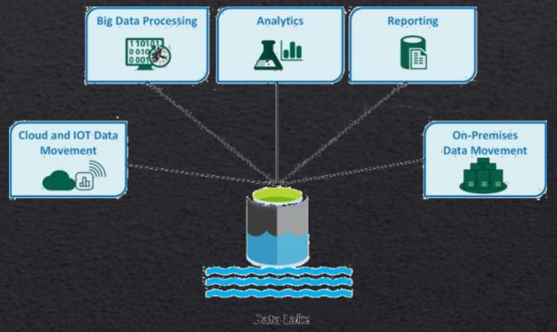
LAKEHOUSE – Powered by Delta Lake, Iceberg, and Hudi



Use Cases

- Aggregate data from OLTP databases and other sources
- Single source of truth
- Business Intelligence applications
- Supports ACID transactions
- Optimize queries (Indexes, Materialized Views, etc.)
- Governance and Auditing

Data Lake - Defined



- A **storage repository** designed to hold **large amounts of data** in its original, raw format.
- **Optimized for scalability** to handle data in the **terabytes and petabytes** range.
- Data originates from **diverse sources**, including **structured, semi-structured, or unstructured** data.
- Contrasts with traditional **data warehouses** that process data upon ingestion.
- It serves as a **single data store**, containing both **raw source data** and transformed data.
- Transformed data is utilized for tasks such as **reporting, visualization, advanced analytics, and machine learning**.
- It encompasses various types of data, including **structured, semi-structured, unstructured**, and even **binary data** like images or audio.
- Inadequately managed data lakes are humorously referred to as **data swamps**.

Data Lake – Pros & Cons

Pros

- Building a **staging area** for your data warehouse
- Audit log of all data ever ingested into your data ecosystem thanks to the **immutable staging** area
- Increase the **time-to-value** and **time-to-insights**
- A single data platform for **real-time** and **batch analytics**
- **Cost effectiveness**, **Convenience**, **Future proofing**

Cons

- Lack of a **schema** or descriptive metadata
- Lack of **semantic consistency** across the data
- It can be hard to guarantee the quality of the data
- **Governance**, **access controls** and **privacy** issues can be problems
- Integration of **relational** data
- Integrated or **holistic views** across the organisation
- **Dumping ground** for data that is never actually analysed or mined for insights
- → As a result, most of the data lakes in the enterprise have become **data swamps**

Azure Data Lake Best Practices

Premium Storage Consideration
If needed, use premium block blob storage for low latency and high I/O operations.

SSDs provide better performance and reduced transaction costs.

Optimizing Data Ingest

Choose appropriate hardware and network connectivity for data ingestion.

Configure tools like DistCp, Azure Data Factory, and Sqoop for parallelization.

Structuring Data Sets

Optimize performance and costs by considering file formats (e.g., Avro, Parquet, ORC) and larger file sizes.

Directory Structure
Different directory structures for IoT, batch jobs, and time-series data are crucial.

Organize directories logically for efficient data processing.

Security Setup

Prioritize security by following Azure AD-based identity management.

Use Azure RBAC roles and ACLs to control access.

Telemetry Monitoring

Monitor performance using Azure Storage logs in Azure Monitor. Gain insights into service performance, operations, and latency.

Data Lake – Considerations in Azure

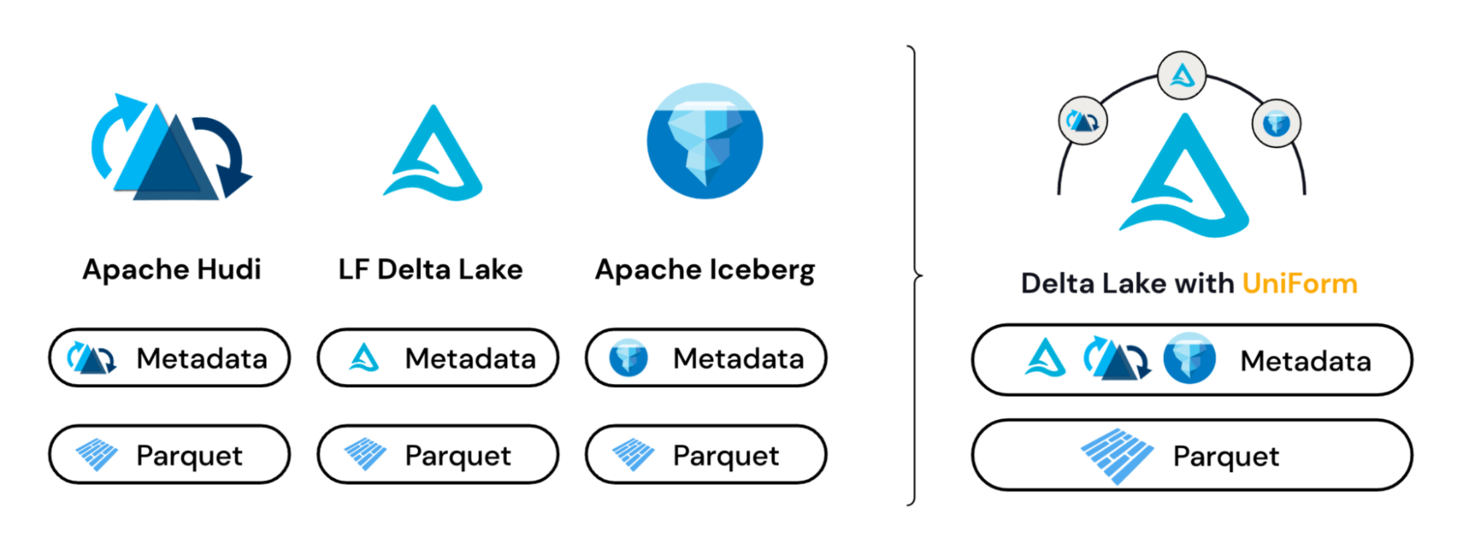
For Azure Data Lake Storage Gen2, check the supported features, noting limitations like no customer fail-over for DR testing.

Data Lake Storage has certain limits:
- Maximum of 10 accounts per subscription (request increase if needed).
- Maximum of 32 access and default ACLs per file or folder (hard limit).
- Storage capacity limits: 2 PB for US and Europe, 500 TB for other regions, including the UK.
- Maximum request rate: 20,000 per second per storage account.
- ADLS does not have a true inheritance model; use default ACLs.
- Changing default ACL on a parent doesn't affect access or default ACL of existing child items.
- Scripting permission changes on existing folders requires recursion.
- RBAC (filesystem level) takes precedence over ACLs in evaluation.
- The Storage Blob Data Owner built-in role and SAS authentication grant super-user access.

Origin of Delta Lake

Parquet
DELTA LAKE

- Delta Lake file format uses the Apache Parquet file format, initially designed for Hadoop File System (HDFS).
- Apache Parquet is a columnar data store, utilizing efficient compression and encoding techniques.
- Apache Parquet is most suitable for read operations, optimizing query performance.
- While write operations are possible with parquet files, intensive writes can lead to the "small files" issue.
- The Delta file format builds on top of the great features of the parquet format and introduced additional benefits.
- Additionally, the recent introduction of Delta UniForm, enables translation from Delta to Iceberg and Hudi formats.

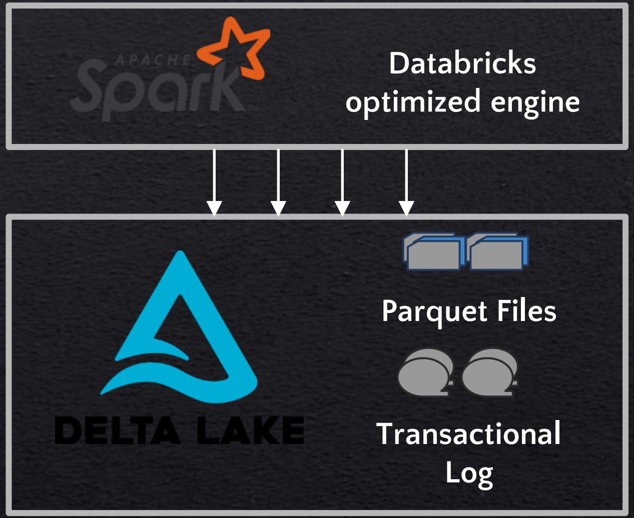


Delta Lake Features

Delta Lake is an open-source storage format that brings ACID transactions to Apache Spark™ and big data workloads

That works with your existing ADLS

- OPEN SOURCE, OPEN FORMAT
- FULL ACID TRANSACTIONS
- UNIFIED STREAMING AND BATCH
- SCHEMA ENFORCEMENT
- TIME TRAVEL/DATA SNAPSHOTS
- NATIVE SUPPORT FOR UPDATE/DELETE/MERGE



Databricks optimized engine

Parquet Files

Transactional Log

- Z/V-ORDER INDEXING AND STATS
- COMPACTION TO OPTIMIZE FILE SIZES
- DATA SKIPPING READS ONLY THE RELEVANT DATA
- CACHING

Lakehouse - Defined

- **Overcomes Data Lakes' limitations**.
- Introduces a new data management architecture that streamlines enterprise data systems and boosts innovation amid machine learning's transformative potential.
- Adds a transactional storage layer.
- Adopts data structures and management features from data warehouses, directly applied to cloud data lakes.
- Enables coexistence of conventional analytics, data science, and machine learning within a unified, open system.

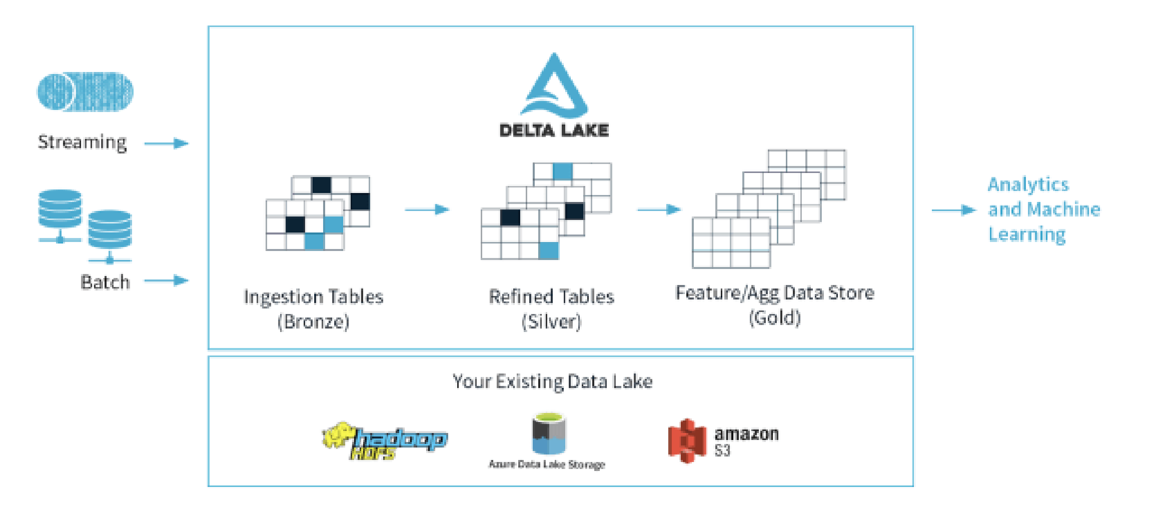
Lakehouse Advantages

This advancement expands the scope for cross-functional enterprise-scale analytics, BI, and machine learning (ML) projects, unlocking significant business value.

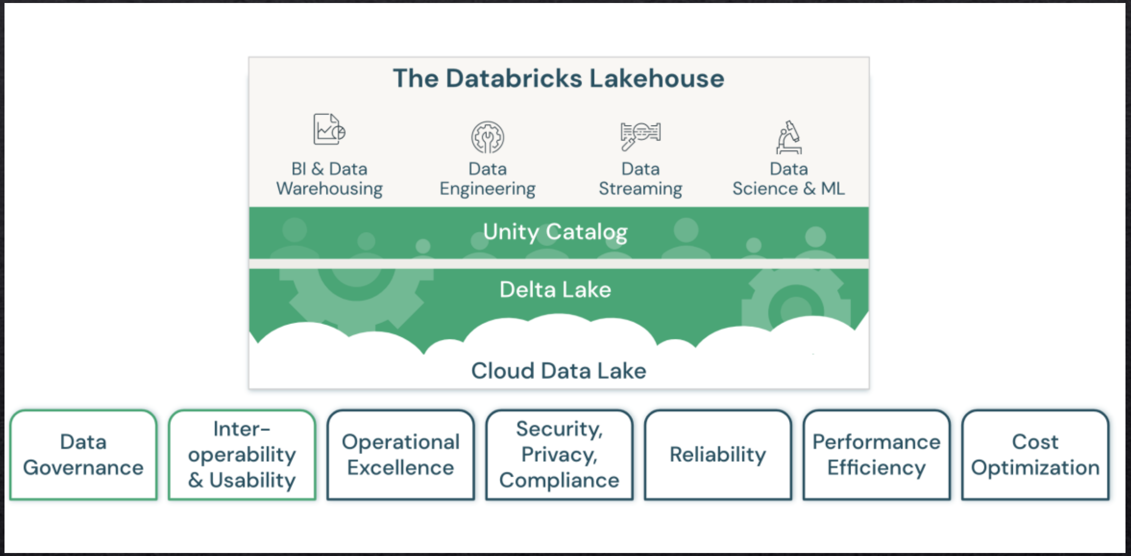
- **Data analysts** can extract valuable insights through SQL queries from the data lake.
- **Data Scientists** can enhance ML models by combining & enriching data sets.
- **Data engineers** can establish automated ETL pipelines.
- **BI analysts** can create visual dashboards and reporting tools more efficiently.
- These applications are all feasible simultaneously on the data lake, without data migration, even during real-time data streaming.

Building a Lakehouse with Delta Lake

- To build a successful lakehouse, organizations have turned to Delta Lake, an open format data management and governance layer that combines the best of both data lakes and data warehouses.
- Across industries, enterprises are leveraging Delta Lake to power collaboration by providing a reliable, single source of truth.
- By delivering quality, reliability, security and performance on your data lake — for both streaming and batch operations — Delta Lake eliminates data silos and makes analytics accessible across the enterprise.
- With Delta Lake, customers can build a cost-efficient, highly scalable lakehouse that eliminates data silos and provides self-serving analytics to end users.



Lakehouse – Considerations in Azure



The *well-architected lakehouse* consists of 7 pillars which describe different areas of concern for the implementation of a data lakehouse in the cloud:

Data governance
The oversight to ensure that data brings value and supports your business strategy.

Interoperability and usability
The ability of the lakehouse to interact with users and other systems.

Operational excellence
All operations processes that keep the lakehouse running in production.

Security, privacy, compliance
Protect the Azure Databricks application, customer workloads and customer data from threats.

Reliability
The ability of a system to recover from failures and continue to function.

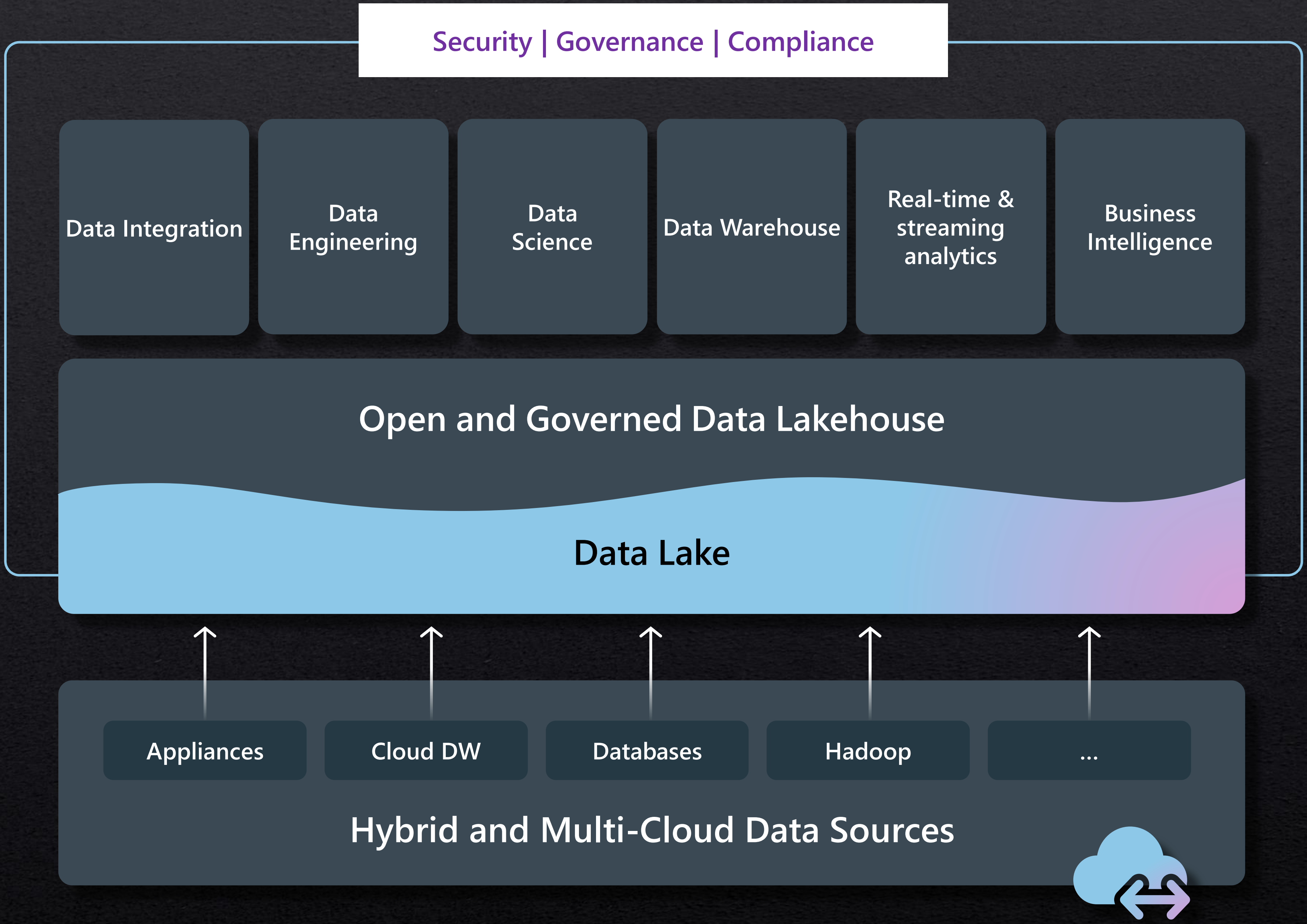
Performance efficiency
The ability of a system to adapt to changes in load.

Cost optimization
Managing costs to maximize the value delivered.
effective and efficient lakehouse.

Data Lake vs Data Lakehouse vs Data warehouse

	Data lake	Data warehouse	Data Lakehouse
Data Types	Structured, semi-structured, unstructured Relational, non-relational	Structured Relational	Structured, semi-structured, unstructured Relational, non-relational
Schema	Schema on read	Schema on write	Schema on read, schema on write
Format	Open, Raw, unfiltered,	Proprietary ,Processed, vetted	Open, Raw, unfiltered, processed, curated, delta format files
Sources	Big data, IoT, social media, streaming data	Application, business, transactional data, batch reporting	Big data, IoT, social media, streaming data, application, business, transactional data, batch reporting
Scalability	Easy to scale at a low cost	Difficult and expensive to scale	Easy to scale at a low cost
Users	Data scientists, data engineers	Data warehouse professionals, business analysts	Business analysts, data engineers, data scientists
Cost	$	$$$	$
Use cases	Machine learning, predictive analytics, real-time analytics	Core reporting, BI	Core reporting, BI, machine learning, predictive analytics
Reliability, Performance	Low quality, data swamp, Poor performance	High quality, reliable data , High performance	High quality, reliable data , High performance

Analytics in Lakehouse



Security | Governance | Compliance

Data Integration

Data Engineering

Data Science

Data Warehouse

Real-time & streaming analytics

Business Intelligence

Open and Governed Data Lakehouse

Data Lake

Appliances

Cloud DW

Databases

Hadoop

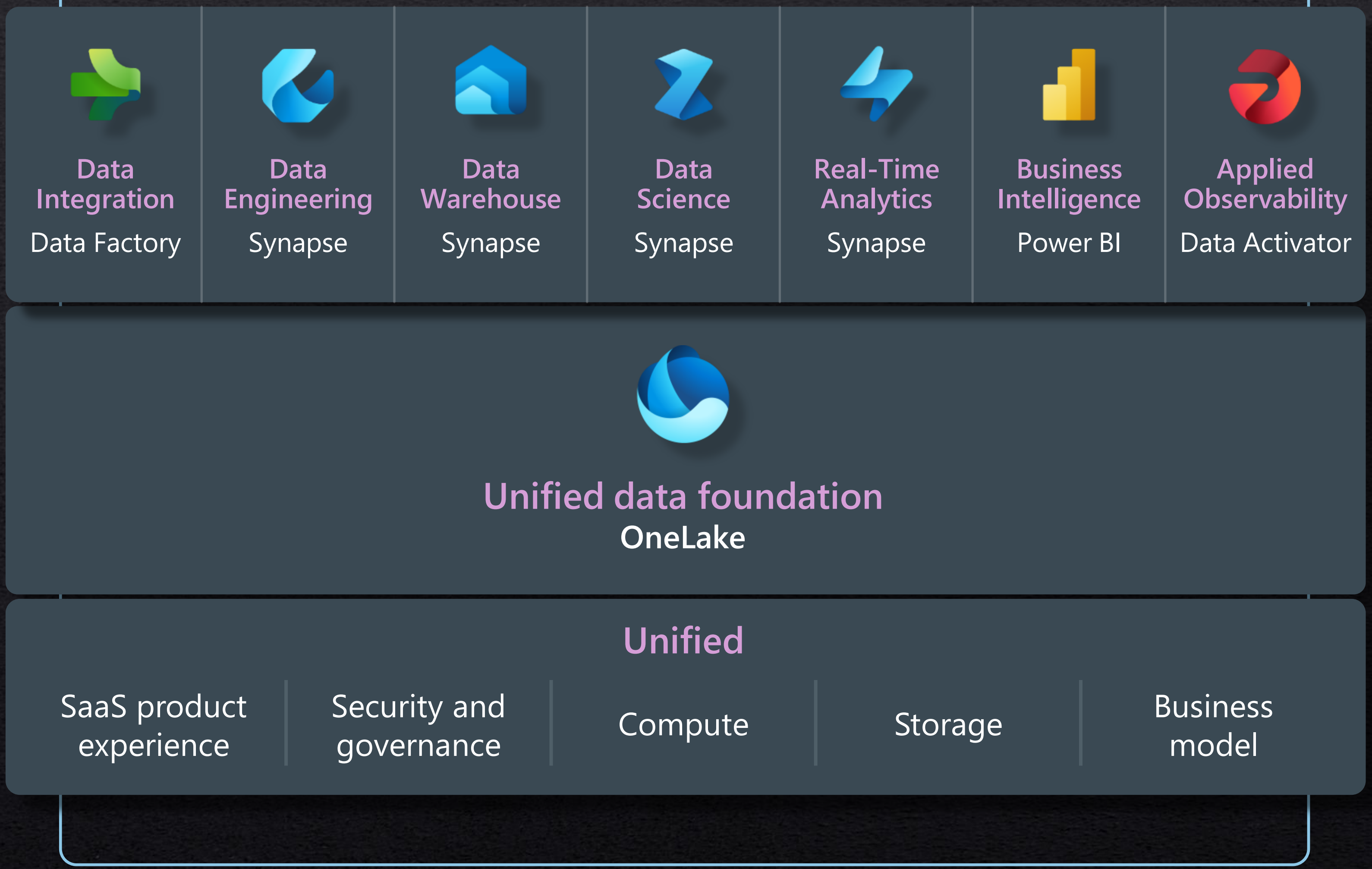
...

Hybrid and Multi-Cloud Data Sources

Cloud Scale Analytics

- Azure Databricks
- Azure Data Factory
- Azure Synapse Analytics
- Microsoft Power BI

Microsoft Fabric



Data Integration
Data Factory

Data Engineering
Synapse

Data Warehouse
Synapse

Data Science
Synapse

Real-Time Analytics
Synapse

Business Intelligence
Power BI

Applied Observability
Data Activator

Unified data foundation
OneLake

Unified

SaaS product experience

Security and governance

Compute

Storage

Business model

Seven key experiences for end-to-end analytics

Experiences are designed to target specific personas and tasks, yet work together seamlessly in a unified platform via OneLake to enable creators to collaboratively do their best work

- Combines the ease of use of Power Query with the scale and Power of Azure Data Factory to leverage 200+ native connectors to data sources on premises and in cloud
- World-class Spark platform with great authoring experiences to empower data engineers to transform data at scale
- Providing industry leading SQL performance and scale, fully separating compute from storage for independently scaling and natively storing data in open Delta Lake
- Build, deploy, and operationalize machine learning models directly within Fabric to empower data scientists and analysts with predictive insights
- Best-in-class engine for observational data analytics to create actionable insights from real-time data
- The world's leading Business Intelligence platform empowers users to quickly and intuitively to make better decisions with data
- No-code Microsoft Fabric experience that empowers the business analyst to drive actions automatically from your data.

Cloud Scale Analytics Evolution

Microsoft Fabric

Azure Databricks

Open and Governed Lakehouse

Available now

Public preview

- Data Factory
- Synapse Data Engineering
- Synapse Data Science
- Synapse Data Warehousing
- Synapse Real Time analytics
- Copilot for Power BI (DAX)
- OneLake

Generally available

- Power BI

Private Preview

- Data Activator
- Copilot for Microsoft Fabric
- Copilot for Power BI (full)

Upgrade to Fabric at your own pace

Continue building on Synapse Gen2, Azure Data Factory, Azure Data Explorer, Azure Databricks

1

Mount existing Synapse Gen2, Azure Data Factory, Azure Data Explorer to Microsoft Fabric, at zero cost/risk

2

Upgrade to full Microsoft Fabric experience with tooling and support from Microsoft

3

Resources

Lakehouse Well Architected Framework : https://learn.microsoft.com/en-us/azure/databricks/lakehouse-architecture/
Microsoft Fabric Trial : Microsoft Fabric free trial.
Guided tour: Microsoft Fabric
Getting started with Fabric e-book : https://aka.ms/fabric-get-started-ebook
Fabric webinar series Microsoft Fabric Webinar Series
Fabric Documentation : https://aka.ms/fabric-docs
Data Lakes :Data lakes - Azure Architecture Center | Microsoft Learn
Azure Data Lake Best Practices :https://learn.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-best-practices
Data Lake considerations in Azure:Scalability and performance targets for standard storage accounts - Azure Storage | Microsoft Learn
Introduction to Delta Lake : https://learn.microsoft.com/en-us/azure/databricks/delta/
Delta Lake Project : Introduction — Delta Lake Documentation
Delta Universal Format : https://docs.delta.io/3.0.0rc1/delta-uniform.html
Data Lakehouse Evolution : https://www.databricks.com/blog/2021/05/19/evolution-to-the-data-lakehouse.html

For any questions please reach out to Owais Hashmi – owhashmi@microsoft.com