

Azure Synapse e DataBricks: due approcci al Data LakeHouse

Roberto Messori

@robymes

Head of Business Integration & Architectures @ Jakala



Agenda

- Data Warehouse vs Data Lake
- The best of both worlds: Data Lakehouse
- Il cuore del Data Lakehouse: Delta Lake
- Azure Data Lakehouse:
 - Synapse vs Databricks
 - Synapse & Databricks

Dati, Dati ovunque

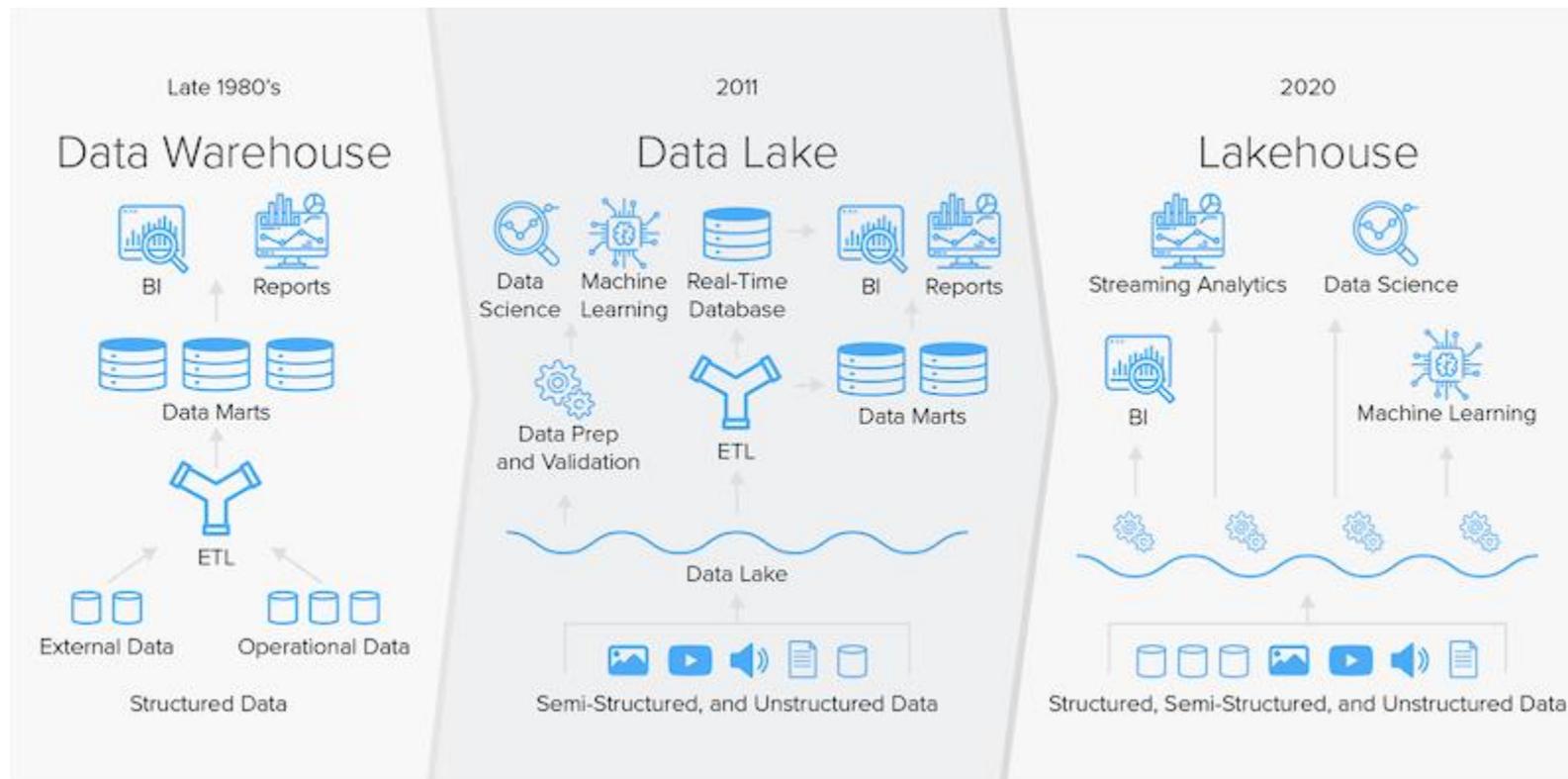


Viviamo nell'era del dato in cui dobbiamo gestire diversità di

- Volumi
- Tipologia
 - Strutturato (tabelle con schema)
 - Non strutturato (immagini, testo libero, log, ...)
 - Semi-Strutturato (JSON, XML, ...)

Dati, Data Strategy, Data Architecture...

- Non esiste una architettura/strategia dati buona per tutte le stagioni
- La tecnologia però aiuta a consolidare strumenti che facilitano la gestione del dato in tutte le sue forme



Ma anche:

- Lambda architecture
- Kappa architecture
- Data Mesh
- ...



Data Warehouse

Pro

- Esistono da decenni
- Sono affidabili, ben conosciuti e godono di ampio support
- Se ben progettati forniscono letture e scritture veloci e consistenti
- Design pattern consolidate per schema e governance del dato

Contro

- Difficili da scalare
- Tradizionalmente archiviano solo dati strutturati
- Vendor lock-in
- Costosi da implementare, mantenere e da acquistare
- Scarso support per data science e streaming

Data Lake

Pro

- Flessibili, possono archiviare tutti i tipi di dato
- Poco costosi, è possibile archiviare/storicizzare il dato anche senza un preciso scopo immediato
- Facili da scalare
- Basati su tecnologia open source

Contro

- Difficile integrazione con alcune tecnologie consolidate (BI, query engine, data discovery, ...)
- Difficile gestire buone performance, alta probabilità di accesso al dato lento
- Difficile evitare l'effetto "data swamp" nella governance del dato

Data Lakehouse: key features

- **Transaction support:** support for ACID transactions ensures consistency as multiple parties concurrently read or write data
- **Schema enforcement and governance:** the lakehouse should have a way to support schema enforcement and evolution, supporting DW schema architectures such as star/snowflake-schemas
- **BI support:** lakehouses enable using BI tools directly on the source data
- **Storage is decoupled from compute:** storage and compute use separate clusters, thus these systems are able to scale to many more concurrent users and larger data sizes
- **Openness:** the storage formats they use are open and standardized, such as Parquet, and they provide an API
- **Support for diverse data type:** ranging from unstructured to structured data
- **Support for diverse workloads:** including data science, machine learning, and SQL and analytics
- **End-to-end streaming:** support for streaming eliminates the need for separate systems dedicated to serving real-time data applications

Data Lakehouse: gap tecnologico

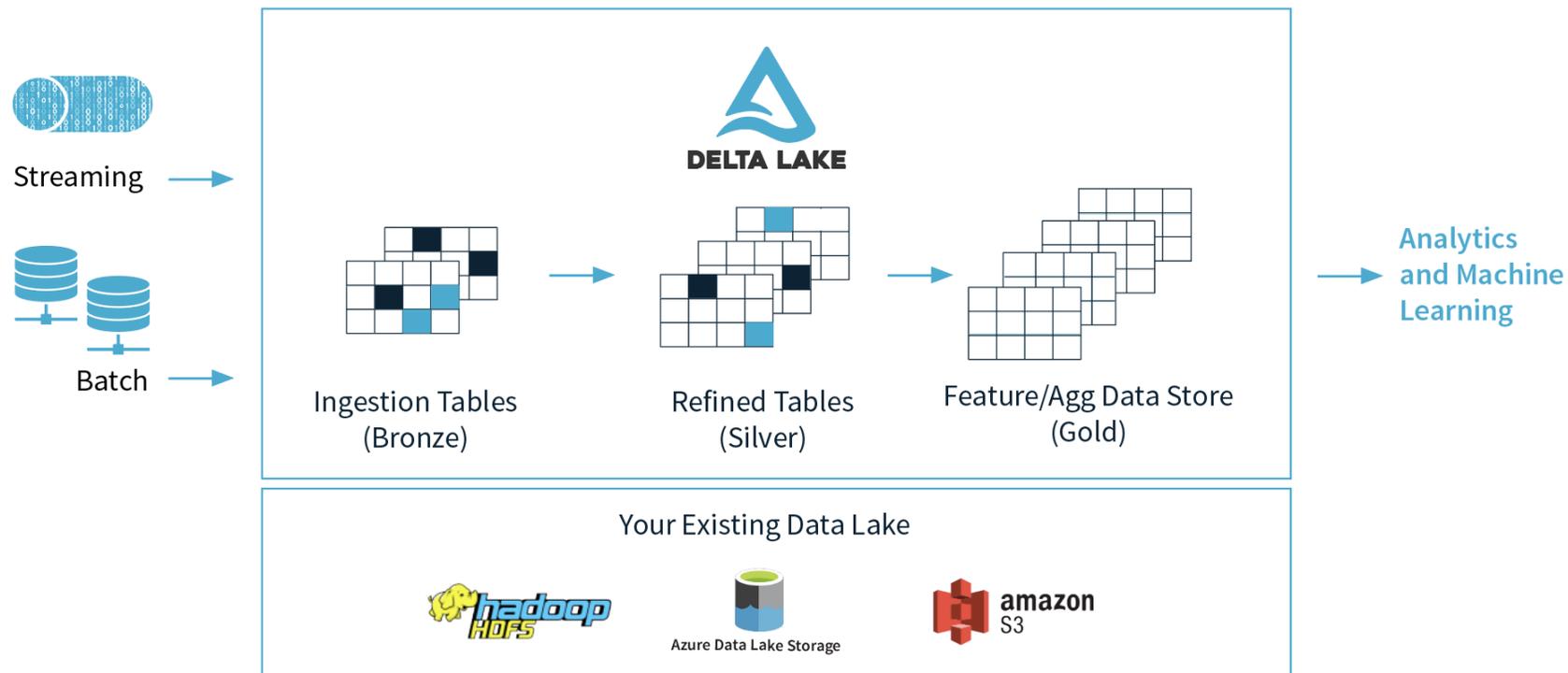
«...directly on the kind of low cost storage used for data lakes»

Basare un approccio lakehouse su data lake comporta una serie di sfide:

- Difficile aggiungere dati senza incorrere in errori di lettura
- Modificare dati esistenti è molto difficile soprattutto su dati granulari
- Mixare batch e streaming spesso porta a inconsistenze nel dato
- Molto costoso mantenere versioni storicizzate del dato
- Difficile gestire il metadato di grandi volumi di dato
- Difficile gestire grandi quantità di file molto piccoli
- Difficile mantenere la data quality

Data Lakehouse: filling the gap with Delta Lake

Delta Lake è un framework **open source** che consente di costruire un'architettura Data Lakehouse su sistemi di storage esistenti come AWS S3, Azure Data Lake Storage, Google Cloud Storage e HDFS

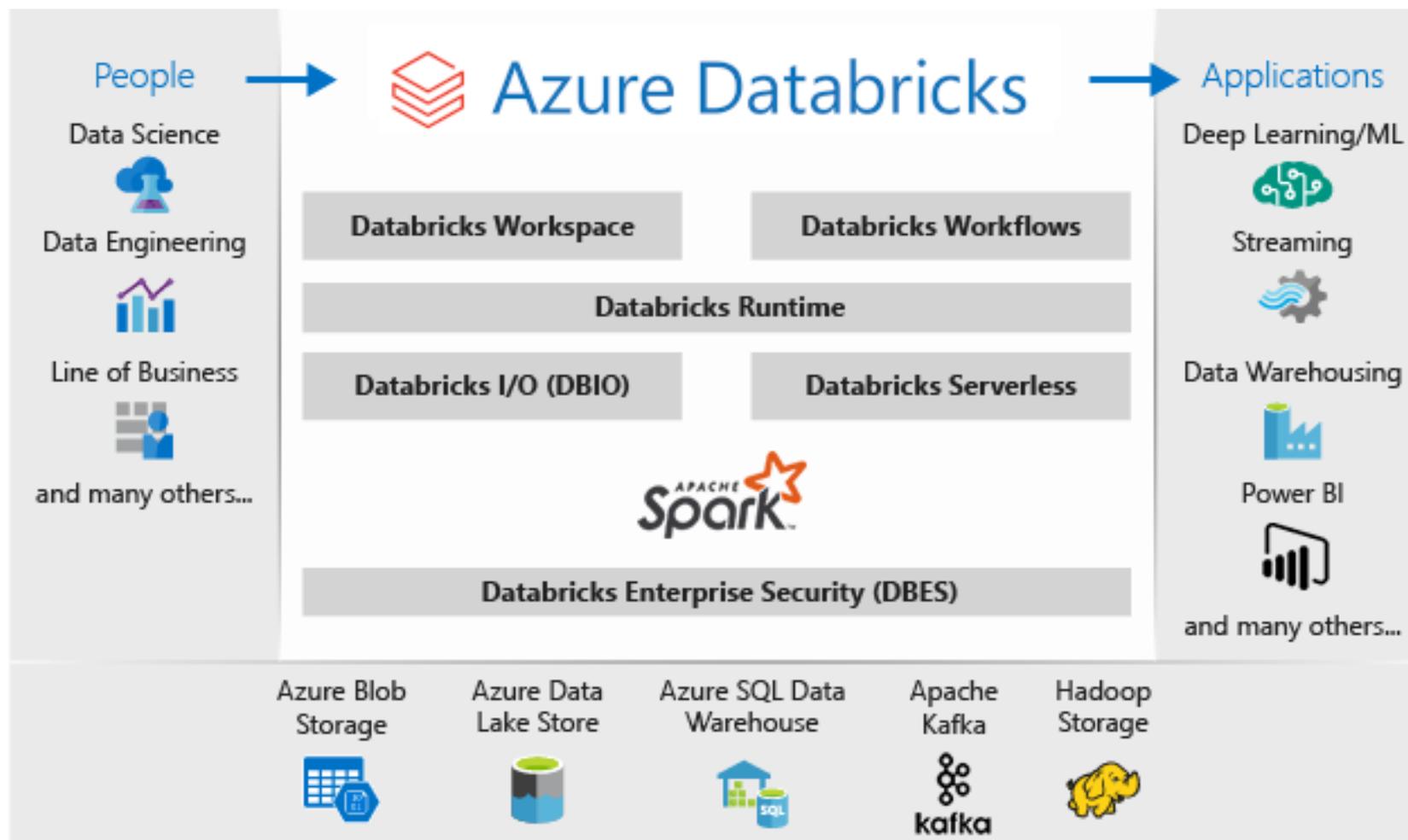


Delta Lake: in pillole

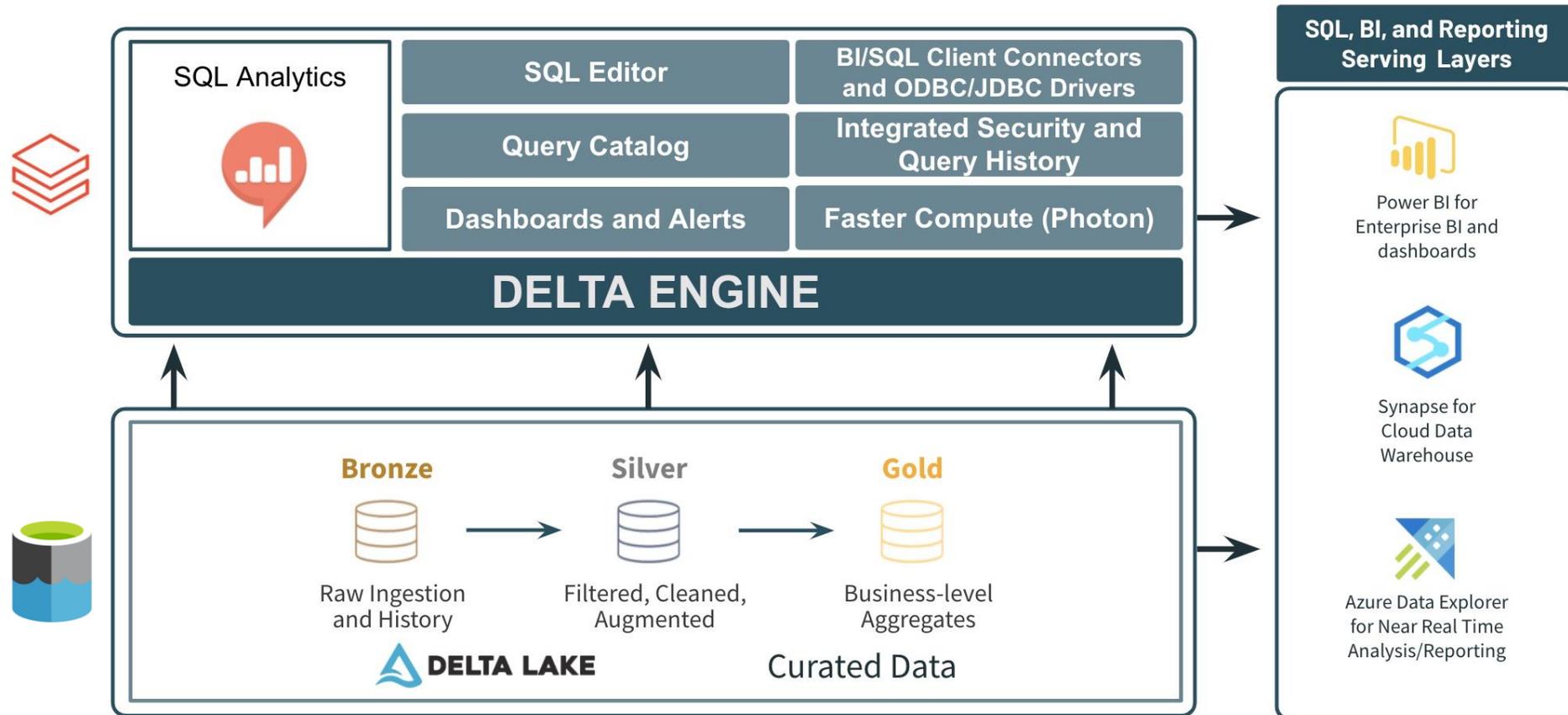
- Transazioni ACID
- Gestione del metadato scalabile
- Versionamento e storico di audit
- Formato open (Parquet)
- Supporto a batch e streaming
- Imposizione ed evoluzione dello schema
- Supporto a updates e delete
- Compatibile con Apache Spark



Data Lakehouse in Azure: Databricks Spark Cluster



Azure Data Lakehouse: Databricks SQL Analytics



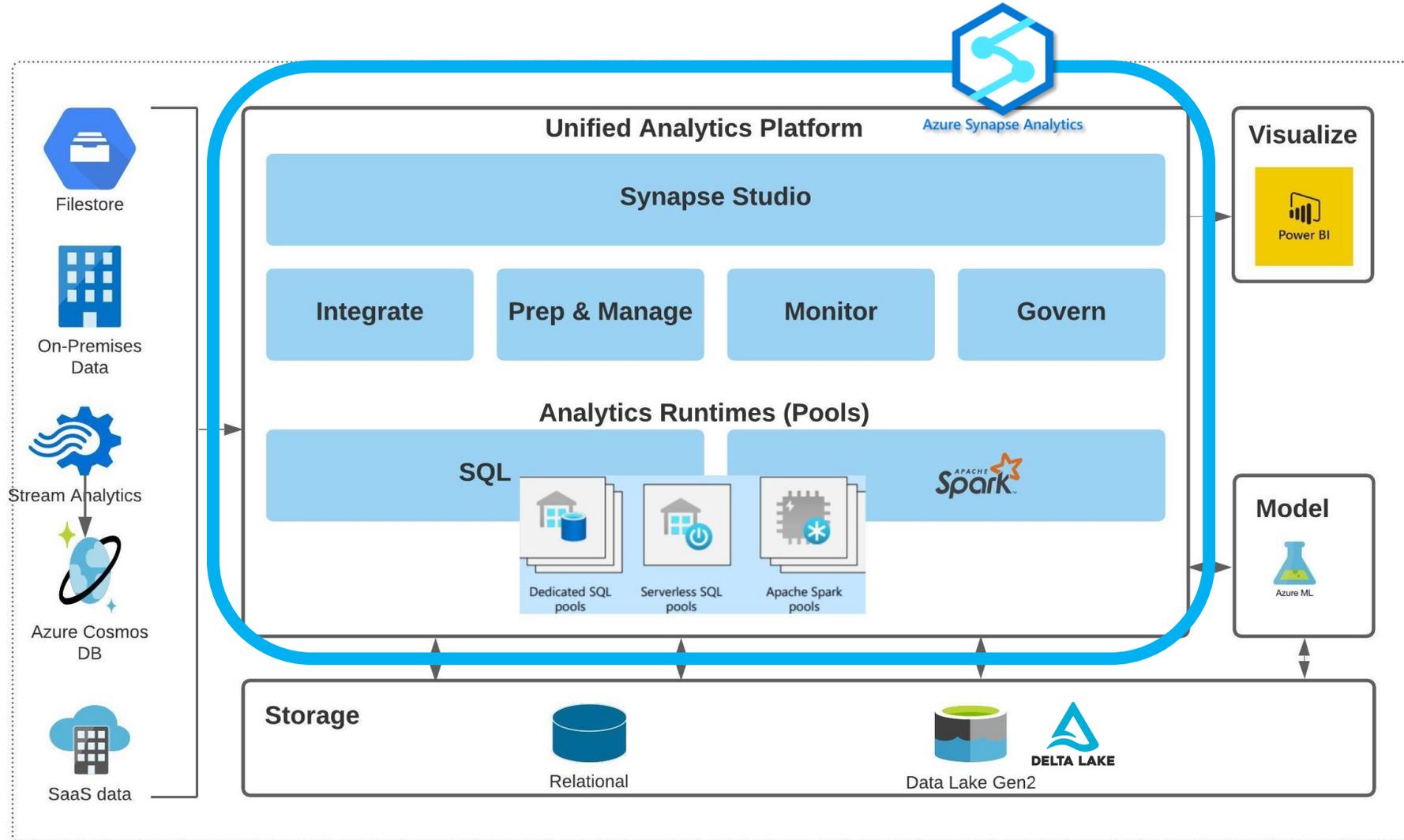
Attualmente in preview in US, disponibile fra qualche mese in EU



Demo

Azure Databricks overview

Azure Data Lakehouse: Synapse Spark + SQL pools



Demo

Azure Synapse overview

Azure Data Lakehouse: Synapse vs Databricks

Entrambe le soluzioni supportano Delta Lake che assume il ruolo di layer architetturale di data governance e data quality in Azure Data Lake Storage Gen2

Synapse

- Spark pool basato su libreria versione 2.4 (3.0 in preview), supporta Python, Scala, SQL, C#
- SQL pool basato su engine SQL Server DWH e Polybase (sia serverless che dedicato), non supporta l'autosospensione
- Integra anche motore di Data Factory (ora Synapse Pipeline) per scheduling/orchestration/data integration
- Completo supporto per tool di BI esterni

Databricks

- Spark engine basato su libreria versione 3.1.1, supporta Python, Scala, R, SQL, Java
- SQL Analytics engine basato Redash.io e Delta Lake Engine/Photon (in preview solo in US), supporta l'autosospensione
- Non è presente un motore di data integration nativo (si utilizza direttamente Spark)
- Completo supporto per tool di BI esterni



Azure Data Lakehouse: Synapse & Databricks

Non siamo affatto costretti a scegliere una soluzione o l'altra, possiamo utilizzarle entrambe (e consiglio di farlo):

- **Il motore Spark di Databricks ad oggi è superiore** a quello di Synapse (anche se Synapse supporta C# per coloro che non hanno skill su Python o Scala)
- **Synapse Pipeline (ex Data Factory) è un plus non presente su Databricks**, supporta l'orchestrazione di notebook Databricks
- **I Synapse SQL pool offrono un ambiente consolidato** e ben conosciuto basato su T-SQL e MPP (Massive Parallel Processing), anche nella versione pool dedicato (quindi con storage dedicato, strategie di distribuzione del dato e indicizzazione)
- **Synapse offer anche il Synapse Link**, servizio cloud di Hybrid Transactional and Analytical Processing (HTAP) **per Cosmos DB** che sfrutta una sua recente nuova feature: l'Analytical Store



Grazie!

- Il materiale sarà online nei prossimi giorni su <http://www.communitydays.it>