16 APRILE 2021

VIRTUAL EVENT

#GLOBALAZURE ITALY 2021

AZURE SYNAPSE: DATA LAKE & MODERN DATA WAREHOUSE DALLA A ALLA Z

R. MESSORA – M. SHAHEDI

@ROBYMES - @MHMD_BURTON

SPONSOR

managed/designs

il partner tecnologico per chi ha idee ambiziose. Innovazione pratica da 15 anni.

Microsoft

empower every person and every organization on the planet to achieve more.

#GLOBALAZURE

Swag and more

- Claim your attendee Learner Badge here:
- 30 Days to learn it: aka.ms/global-azure/30D2L
- Virtual background and ANOTHER Badge: blog.globalazure.net/Swag

BUON POMERIGGIO

CHI SIAMO

- **Roberto Messora**
  - Responsabile, Enterprise Architect
    - area Business Integration & Architectures
- **Mohammad Shahedi**
  - Big Data Engineer
    - area Business Integration & Architectures

JAKALA
DATA DRIVEN PERFORMANCE

DI CHE COSA PARLEREMO

TOPICS

- Modern Data Warehouse (per gli amici Data Lakehouse)
- Azure Synapse Analytics
  - Pipelines (ETL/ELT/Orchestration)
  - Spark Pools & Delta Lake
  - SQL pools serverless & dedicated

16 APRILE 2021

VIRTUAL EVENT

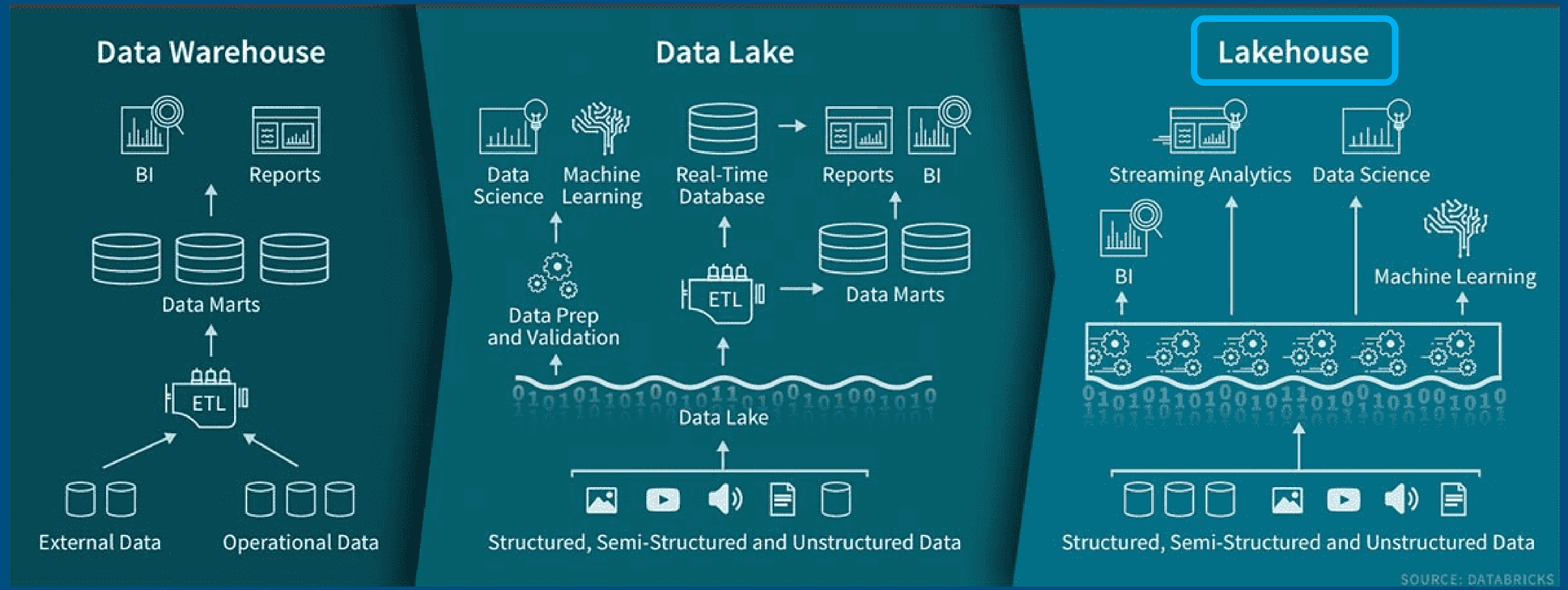
MODERN DATA WAREHOUSE

DATA LAKEHOUSE

#GLOBALAZURE

ARCHITETTURE DATI

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



- *Lambda architecture*
- *Kappa architecture*
- *Data Mesh*
- *…*

https://databricks.com/blog/2020/01/30/what-is-a-data-lakehouse.html

DATA LAKEHOUSE

THE BEST OF BOTH WORLDS

A **Data Lakehouse** is a new, open architecture that combines the best elements of **Data Lakes** and **Data Warehouses**

Lakehouses are enabled by a new open and standardized system design

- Implementing similar data structures and data management features to those in a data warehouse
- Directly on the kind of low cost storage used for data lakes
- They are what you would get if you had to redesign data warehouses in the modern world

https://databricks.com/blog/2020/01/30/what-is-a-data-lakehouse.html

DATA LAKEHOUSE

KEY FEATURE

- **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**: In practice this means 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

https://databricks.com/blog/2020/01/30/what-is-a-data-lakehouse.html

16 APRILE 2021

VIRTUAL EVENT

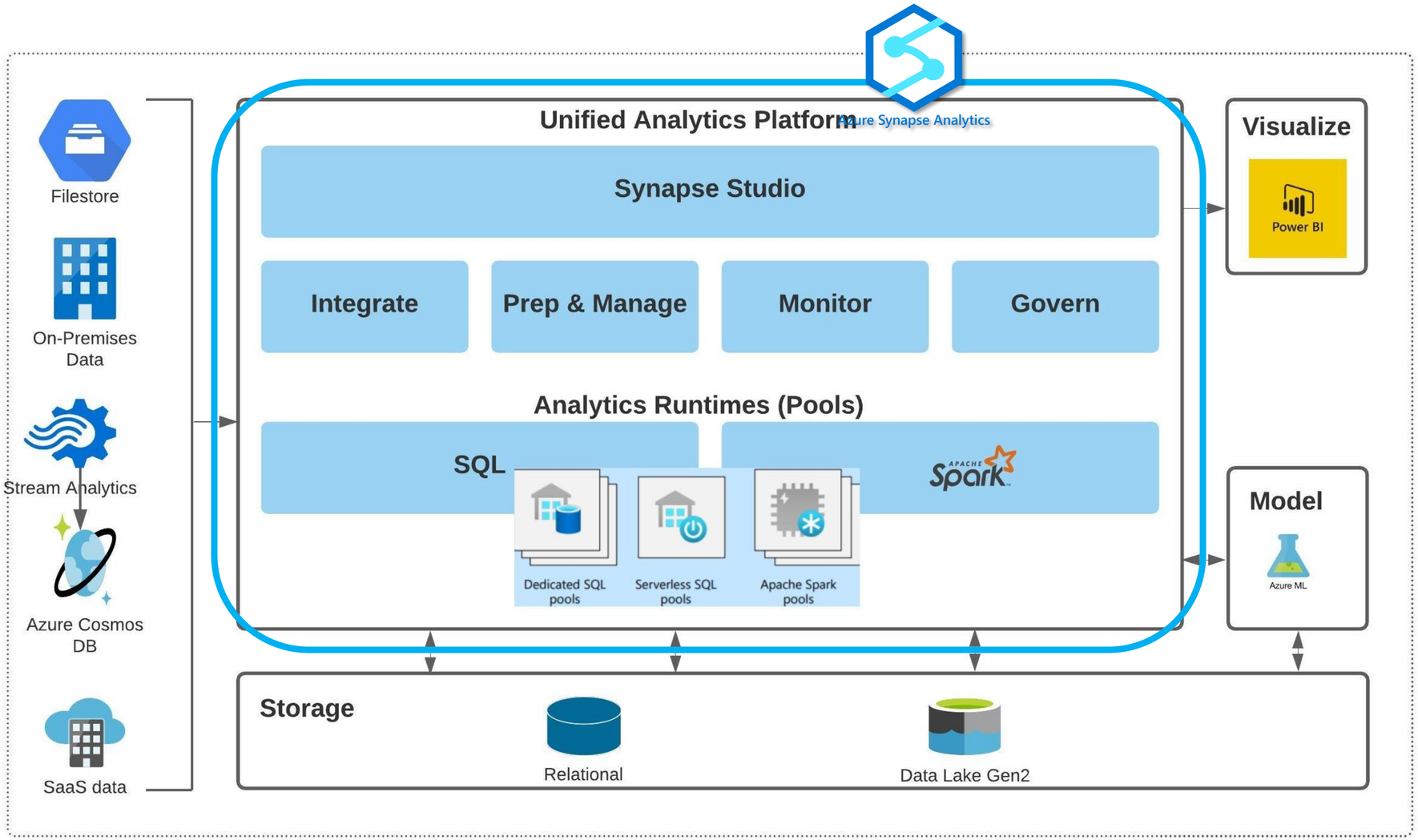
AZURE SYNAPSE ANALYTICS

OVERVIEW

#GLOBALAZURE

AZURE SYNAPSE ANALYTICS

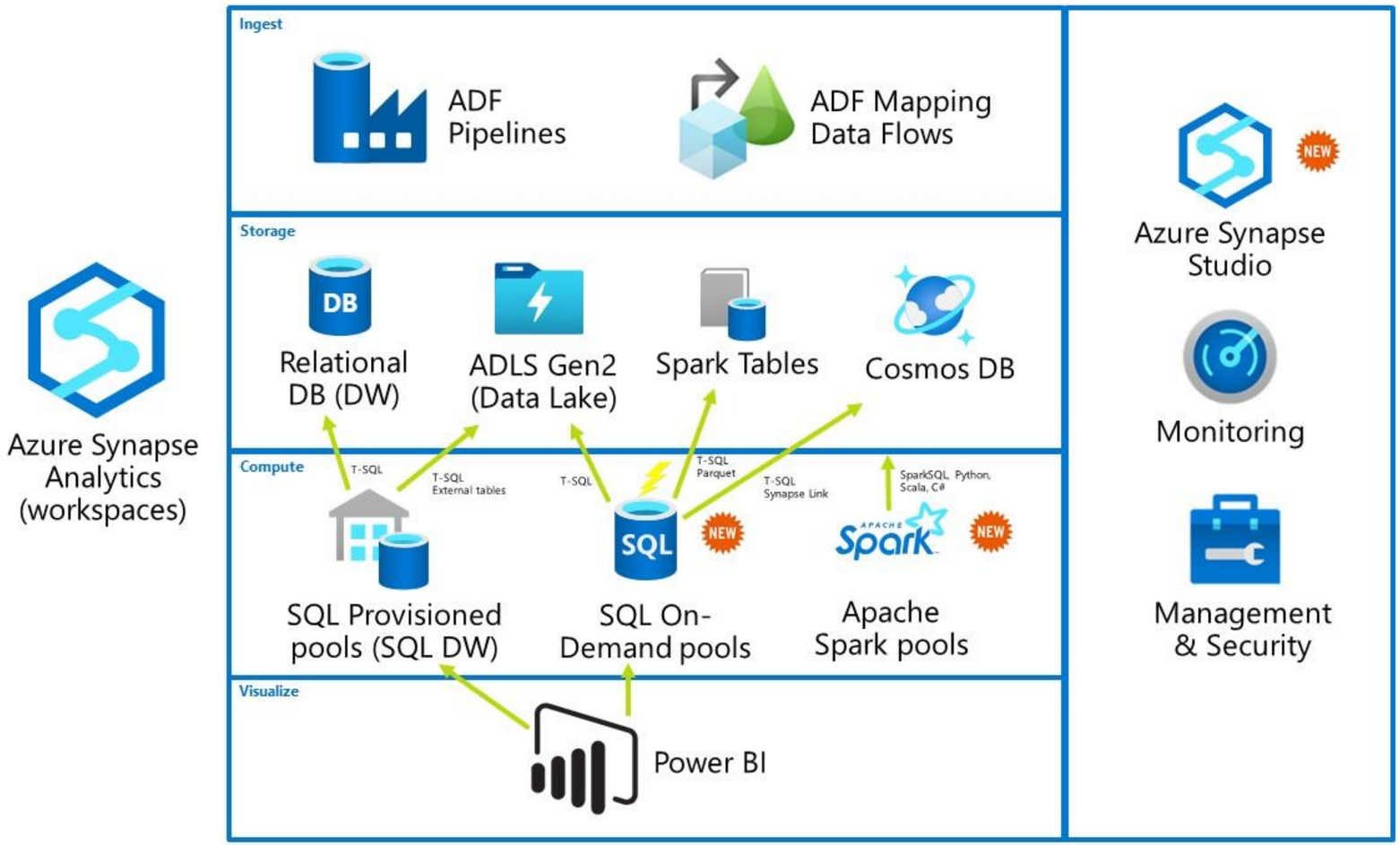
ARCHITETTURA



#GLOBALAZURE

AZURE SYNAPSE ANALYTICS

COMPONENTI



#GLOBALAZURE

AZURE SYNAPSE ANALYTICS

IN PILLOLE

- Razionalizzazione e miglioramento di una serie di servizi di data platform pre-esistenti all'interno di un unico **Workspace** (Azure Data Factory, Azure SQL DWH)
- Disponibilità di un ambiente di lavoro unificato, **Azure Synapse Studio**
- Possibilità di creare e scalare diversi workload di tipo DWH (**SQL Pools**) e Spark (**Spark Pools**)
- Ambiente di monitoraggio e alerting disponibile per ognuno dei componenti (data integration, DWH, Spark)
- Integrazione con Git in modalità Pull Request
- Classico modello di security Azure basato su RBAC

16 APRILE 2021

VIRTUAL EVENT

AZURE SYNAPSE PIPELINES

ETL/ELT/ORCHESTRATION

#GLOBALAZURE

SYNAPSE PIPELINES

MI È PARSO DI VEDERE DATA FACTORY…

- Integrazione di Azure Data Factory all’interno del Synapse Workspace
- Motore di ETL/ELT in modalità low-code/no-code dotato di una grande varietà di connettori verso fonti dati di prodotto e di protocollo
- Synapse Studio è di fatto l’evoluzione dell’ambiente di sviluppo e monitoraggio di Data Factory
- È disponibile un motore Spark serverless low-code/no-code denominato Data Flow

16 APRILE 2021

VIRTUAL EVENT

AZURE SYNAPSE SPARK POOLS

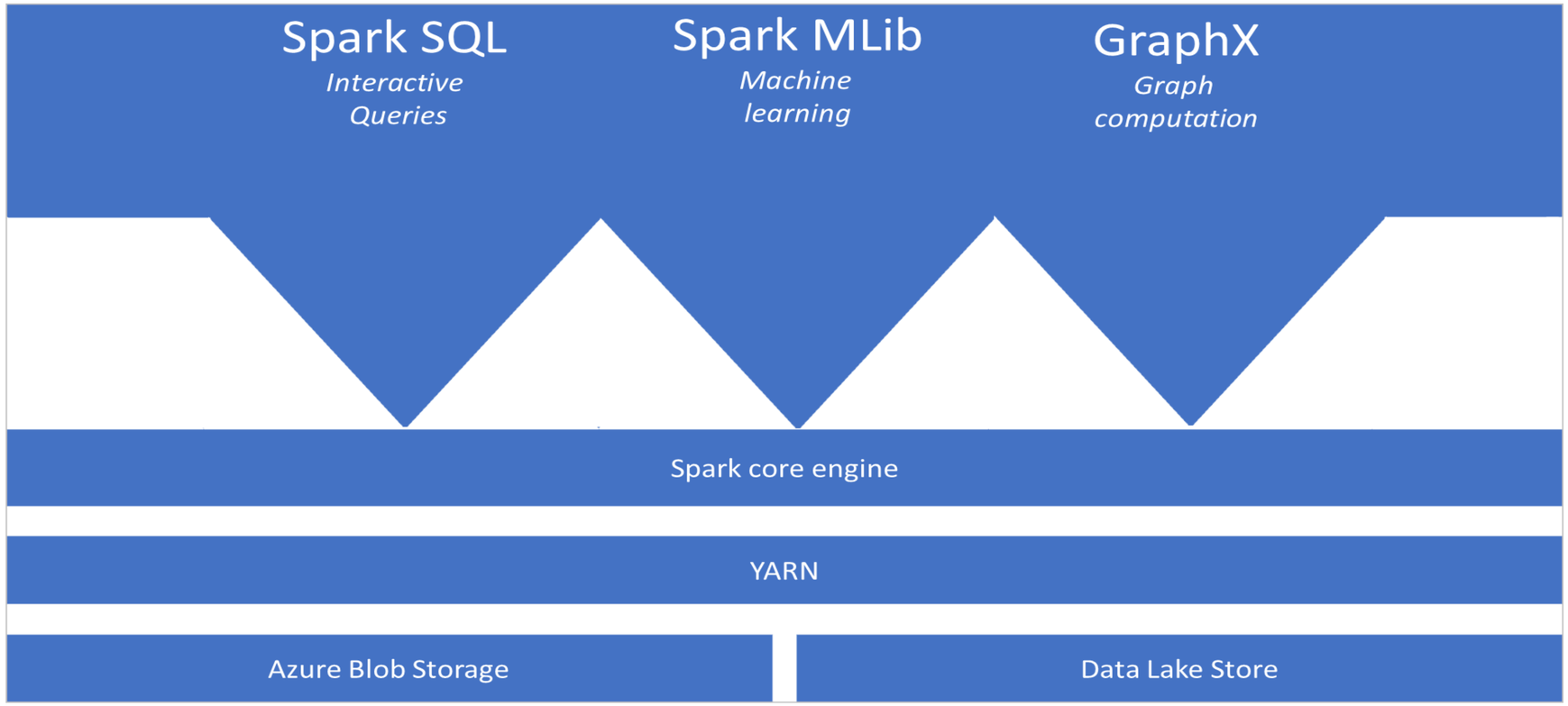
APACHE SPARK & DELTA LAKE

#GLOBALAZURE

SYNAPSE SPARK POOLS

- Azure Synapse permette di creare e configurare cluster Apache Spark serverless in diversi dimensionamenti e capacità di calcolo

- Gli Spark Pool in Azure Synapse sono compatibili con Azure Blob Storage e Azure Data Lake Storage Gen2



SYNAPSE SPARK POOLS

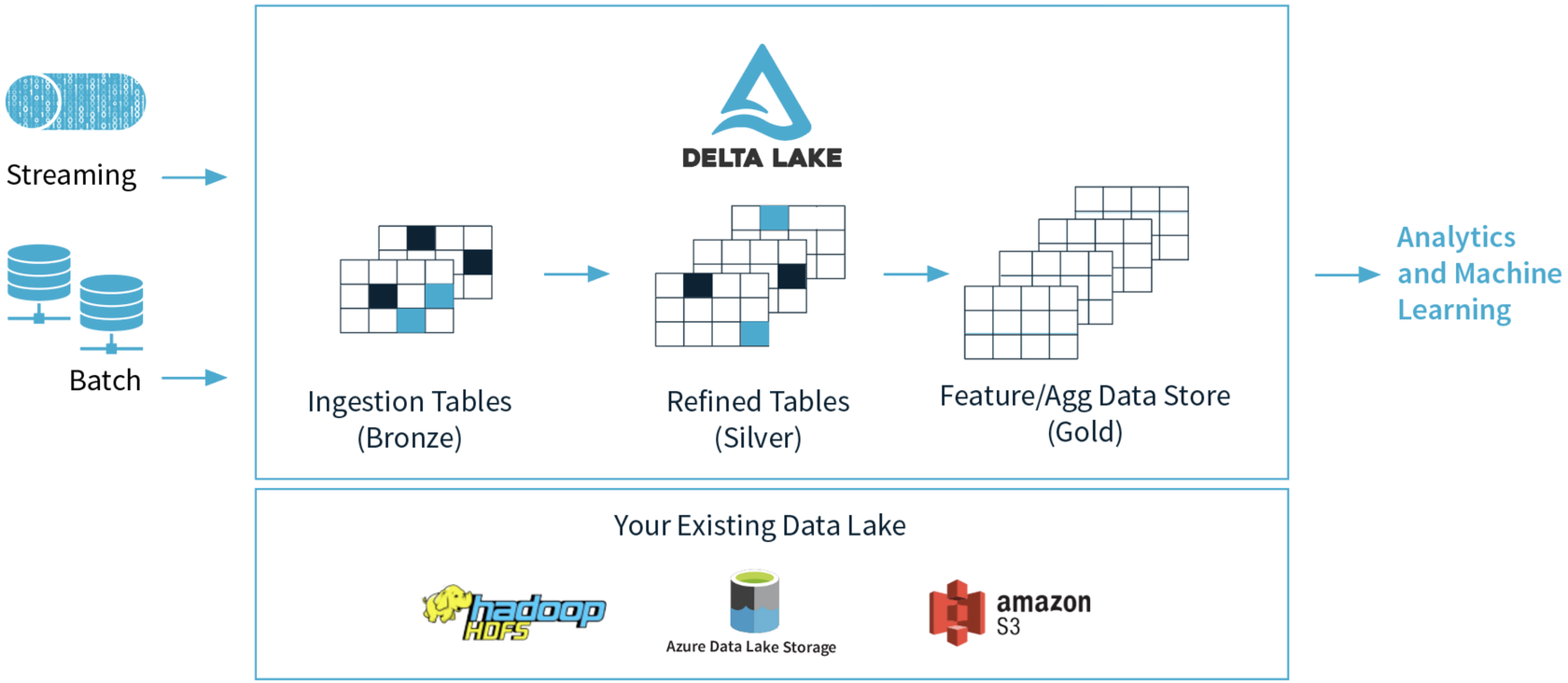
IN PILLOLE

- **Efficienza**: tempo di del cluster startup di circa 2 minuti sotto i 60 nodi, 5 minuti sopra i 60 nodi
- **Notebook integrati**: esperienza di sviluppo in Synapse Studio basata su ***Nteract*** (https://nteract.io/)
- **REST API**: disponibilità di uno Spark job server che espone una API per interagire con il cluster basata su ***Apache Livy*** (http://livy.incubator.apache.org./)
- **Librerie Anaconda precaricate**
- **Supporto per Delta Lake** (https://delta.io/)
- **Scalabilità**: i cluster possono essere abilitati all'auto-scaling

DELTA LAKE

DATA LAKE ON STEROIDS

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

16 APRILE 2021

VIRTUAL EVENT

AZURE SYNAPSE SQL POOLS

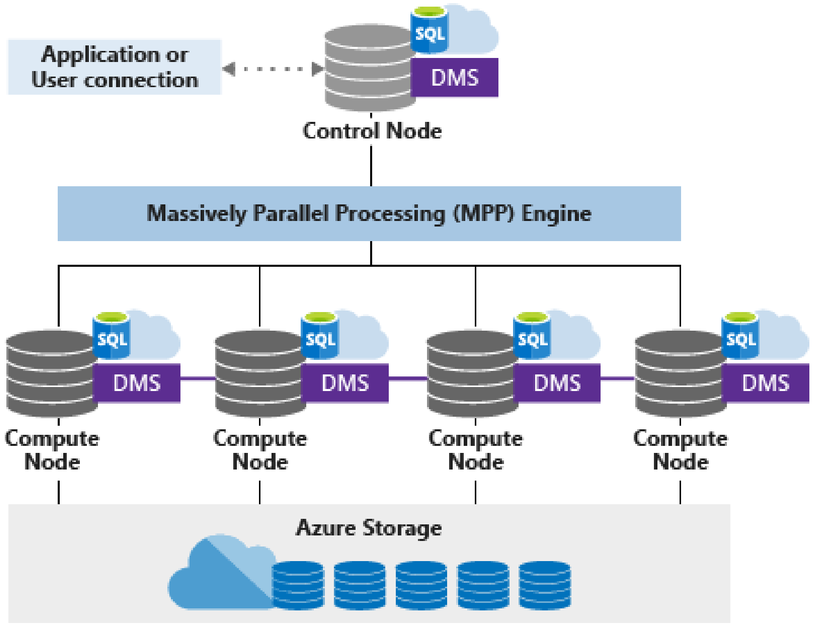
SERVERLESS & DEDICATED

#GLOBALAZURE

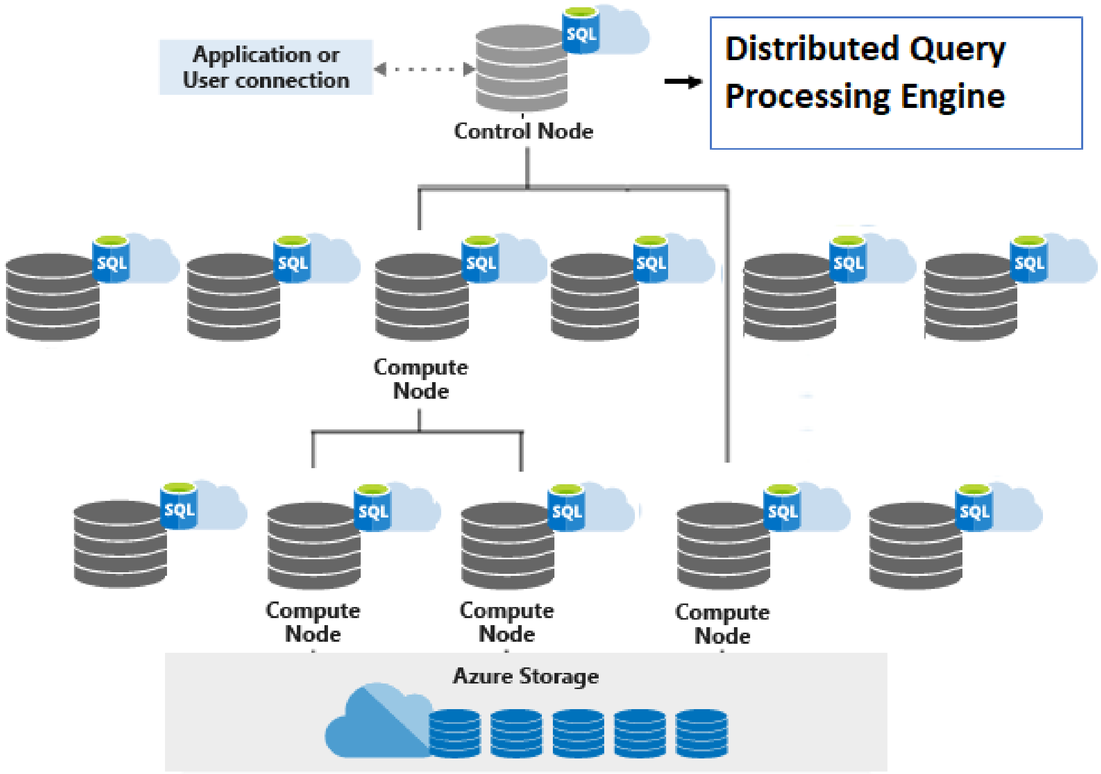
SQL POOL ARCHITECTURE

OLAP MPP: NON È UN DB TRANSAZIONALE!

Dedicated SQL pool



Serverless SQL pool



#GLOBALAZURE

SERVRLESS SQL POOL

SQL QUERY ENGINE PER IL DATA LAKE

- NON è un DWH e NON ha storage, si possono creare solo external table che puntano ai file sul Data Lake
- Permette di utilizzare una buona parte del linguaggio T-SQL per interrogare al volo i dati che si trovano all'interno di Azure Data Lake Storage mediante una tecnologia denominata **Polybase** che non necessita di driver o connettori addizionali
- Permette di accedere anche alle **Spark Table** create negli Spark Pool
- Permette di essere agganciato a **Power BI**
- Il pricing model è per quantità di dati analizzati in una query con un taglio minimo di 10 MB

DEDICATED SQL POOL

OLAP DWH (RIPETIAMO: NON È UN DB TRANSAZIONALE…)

- **Data Warehouse Units** (DWU): taglio di overall performance/risorse del singolo SQL Pool che può variare da 1 a 60 Compute Node
- **Tabelle**:
  - **External**: come per i SQL Pool Serverless permettono di eseguire query T-SQL sui file del Data Lake mediante Polybase
  - **Regular**: le vere e proprie tabelle del DWH
  - **Temporary**: durano il solo lo spazio della sessione di lavoro
- A differenza degli Spark Pool non ha meccanismo automatico di sospensione e scalabilità, sono operazioni che possono essere fatte solo manualmente

DEDICATED SQL POOL

STORAGE E DISTRIBUZIONE

- **Data Movement Service** (DMS): tecnologia di trasporto che coordina la movimentazione del dato fra i compute node
- Tipologie di **distribuzione delle tabelle** sui compute node:
  - **Hash**: distribuzione delle righe tramite funzione di hash, maggior performance delle query per join e aggregazioni su grandi volumi
  - **Round-Robin**: distribuzione delle righe in maniera uniforme sui nodi, logica di distribuzione semplice e non ottimizzata, ideale per tabelle di staging
  - **Replication**: distribuzione della copia della tabella su ogni nodo, maggior performance per piccoli volumi

DEDICATED SQL POOL

INDICIZZAZIONE E PARTIZIONAMENTO

- Tipologie di **indici** applicabili alle Regular Table
  - **Clustered Columnstore**: indicizzazione di default se non specificata, miglior compromesso fra compressione e performance, ideale per tabelle con più di 60 milioni di righe
  - **Heap**: ideale per tabelle di landing perché velocizza i tempi di inserimento e per tabelle con meno di 60 milioni di righe
  - **Clustered e Nonclustered**: ideale per lookup table da cui ottenere poche righe in risposta ad una query con filtri molto selettivi
- Le Regular Table possono essere **partizionate** (tipicamente per campi datetime) per migliorare le performance di query e delete

16 APRILE 2021

VIRTUAL EVENT

AZURE SYNAPSE GOODIES

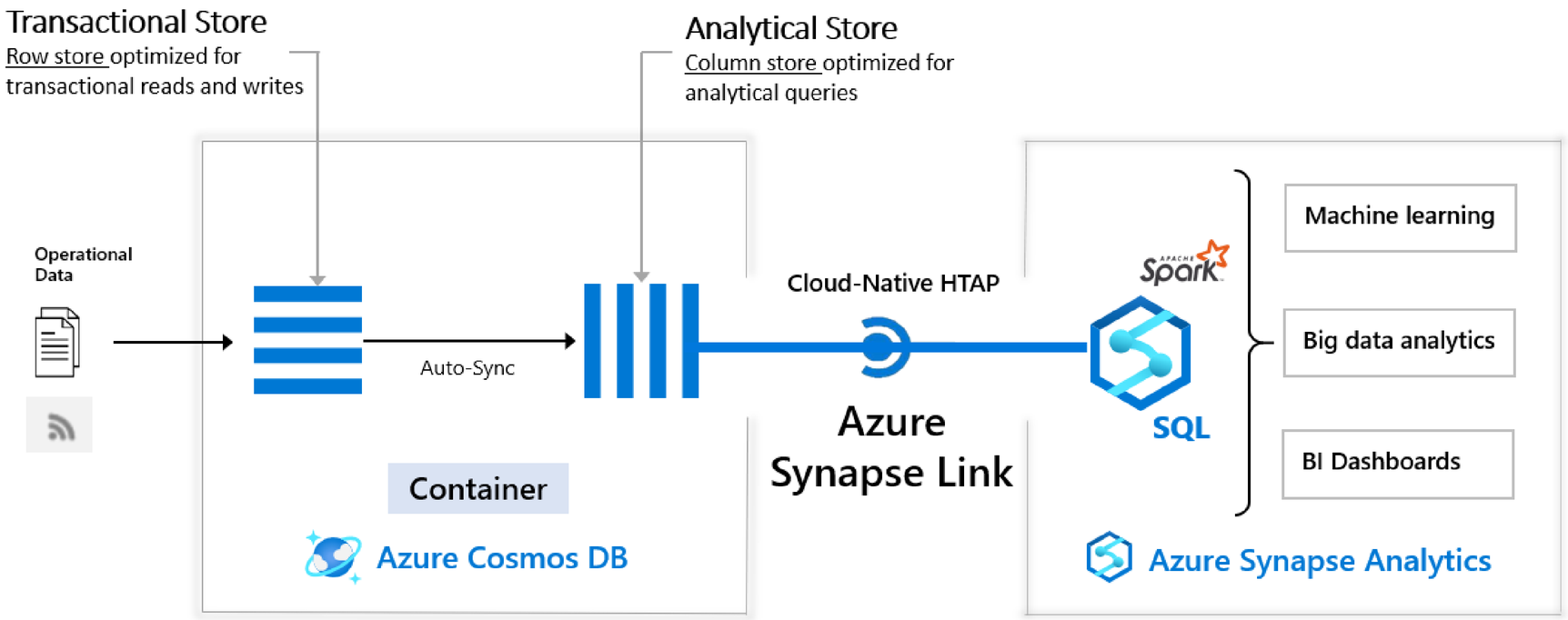
COSMOS DB LINK

#GLOBALAZURE

SYNAPSE LINK

REAL TIME ANALYTICS PER COSMOS DB

Synapse Link è un servizio cloud di Hybrid Transactional and Analtyical Processing (HTAP) per Cosmos DB che sfrutta una sua recente nuova feature: l'Analytical Store



Grazie

- Claim your attendee Learner Badge here:
- 30 Days to learn it: aka.ms/global-azure/30D2L
- Virtual background and ANOTHER Badge: blog.globalazure.net/Swag