

Data Vault Conference Nürnberg

Data Vault with Spark

Marco Amhof
September 2021

2 HALLO, GRÜEZI, HI!



MARCO AMHOF

- since 01/2001 @Trivadis
- Data Platforms with Microsoft Technology
- Spare Time – hiking, skiing, travelling

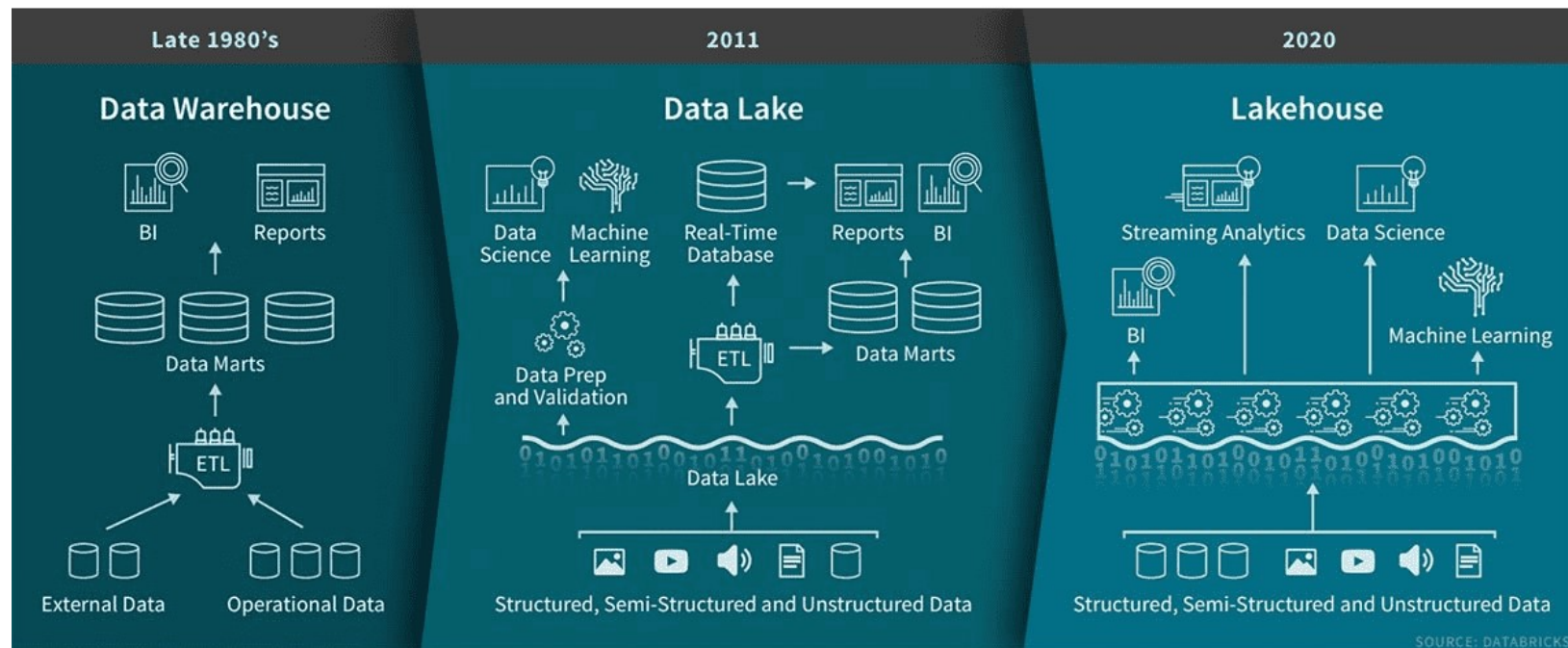


trivadis Part of Accenture

AGENDA

- Data Warehouse vs Data Lake vs Lakehouse
- The modern Data Warehouse Pattern
- Azure Synapse Analytics
- Databricks
- Delta.io (the ACID Data Lake)

4 DATA WAREHOUSE vs DATA LAKE vs DATA LAKEHOUSE



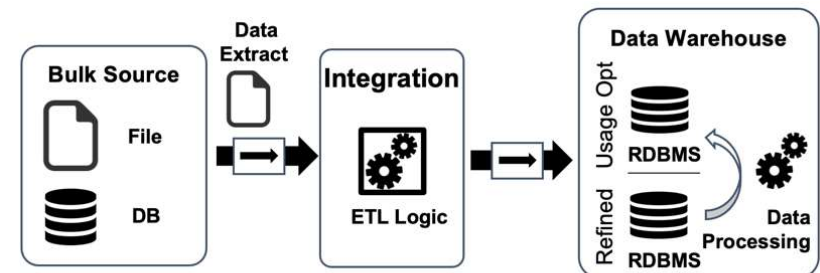
5 DATA WAREHOUSE vs DATA LAKE vs DATA LAKEHOUSE

	Data warehouse	Data lake	Data lakehouse
Data format	Closed, proprietary format	Open format	
Types of data	Structured data, with limited support for semi-structured data	All types: Structured data, semi-structured data, textual data, unstructured (raw) data	
Data access	SQL-only, no direct access to file Schema-on-write	Open APIs for direct access to files with SQL, R, Python and other languages Schema-on-read	
Reliability	High quality, reliable data with ACID transactions	Low quality, data swamp	High quality, reliable data with ACID transactions
Governance and security	Fine-grained security and governance for row/columnar level for tables	Poor governance as security needs to be applied to files	Fine-grained security and governance for row/columnar level for tables
Performance	High	Low	High
Scalability	Scaling becomes exponentially more expensive	Scales to hold any amount of data at low cost, regardless of type	
Use case support	Limited to BI, SQL applications and decision support	Limited to machine learning	One data architecture for BI, SQL and machine learning

6 DATA WAREHOUSE VS DATA LAKE(HOUSE)

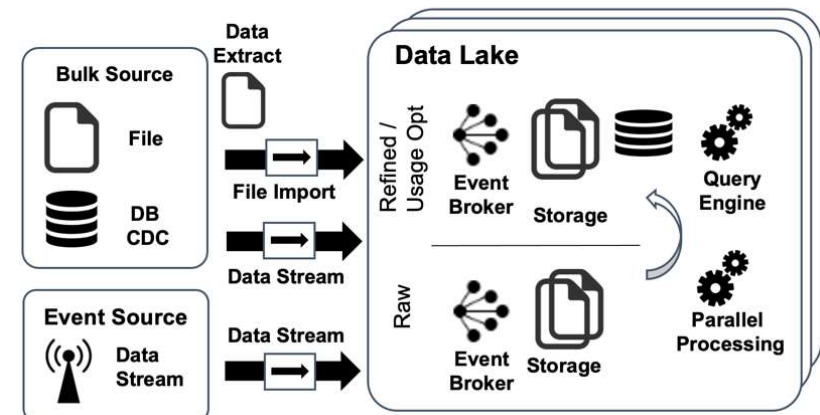
■ Data Warehouse

- ETL Pipelines bring data in a structured unified form before they get stored
- Data model is pre-defined (Schema-on-write)
- Hard to ingest from new data sources



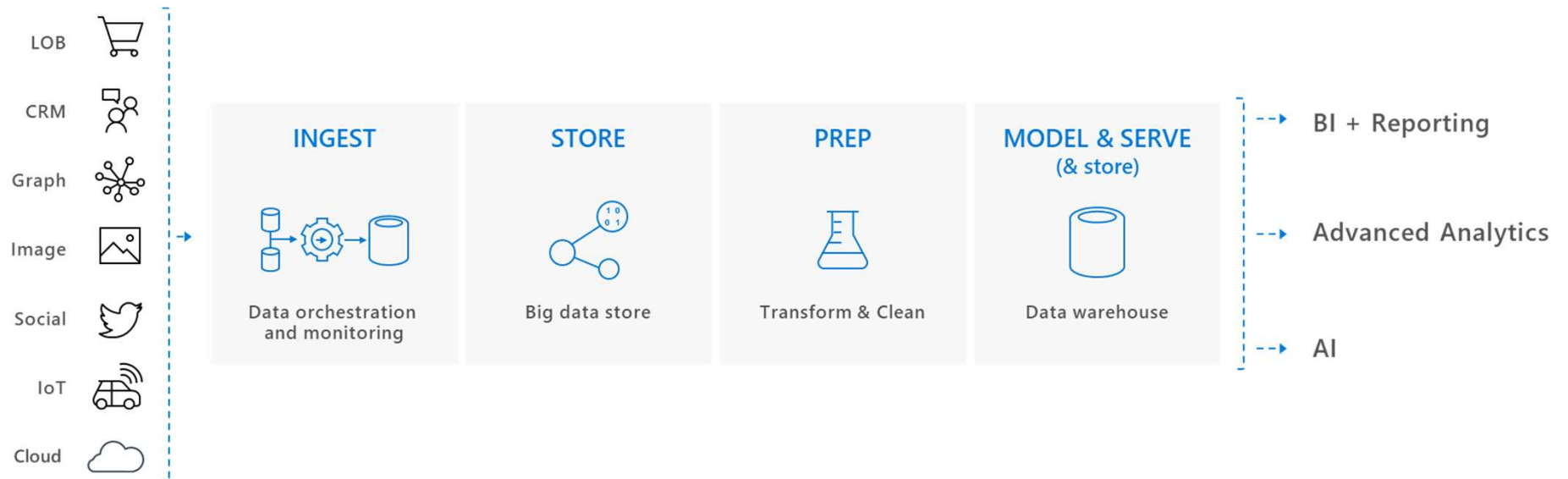
■ Data Lake

- No unification / transformation before storing data (raw-data is stored)
- Data model during read (Schema-on-read)
- Easy to ingest from new data sources
- Good for Data Scientists, Advanced Analytics and Machine Learning

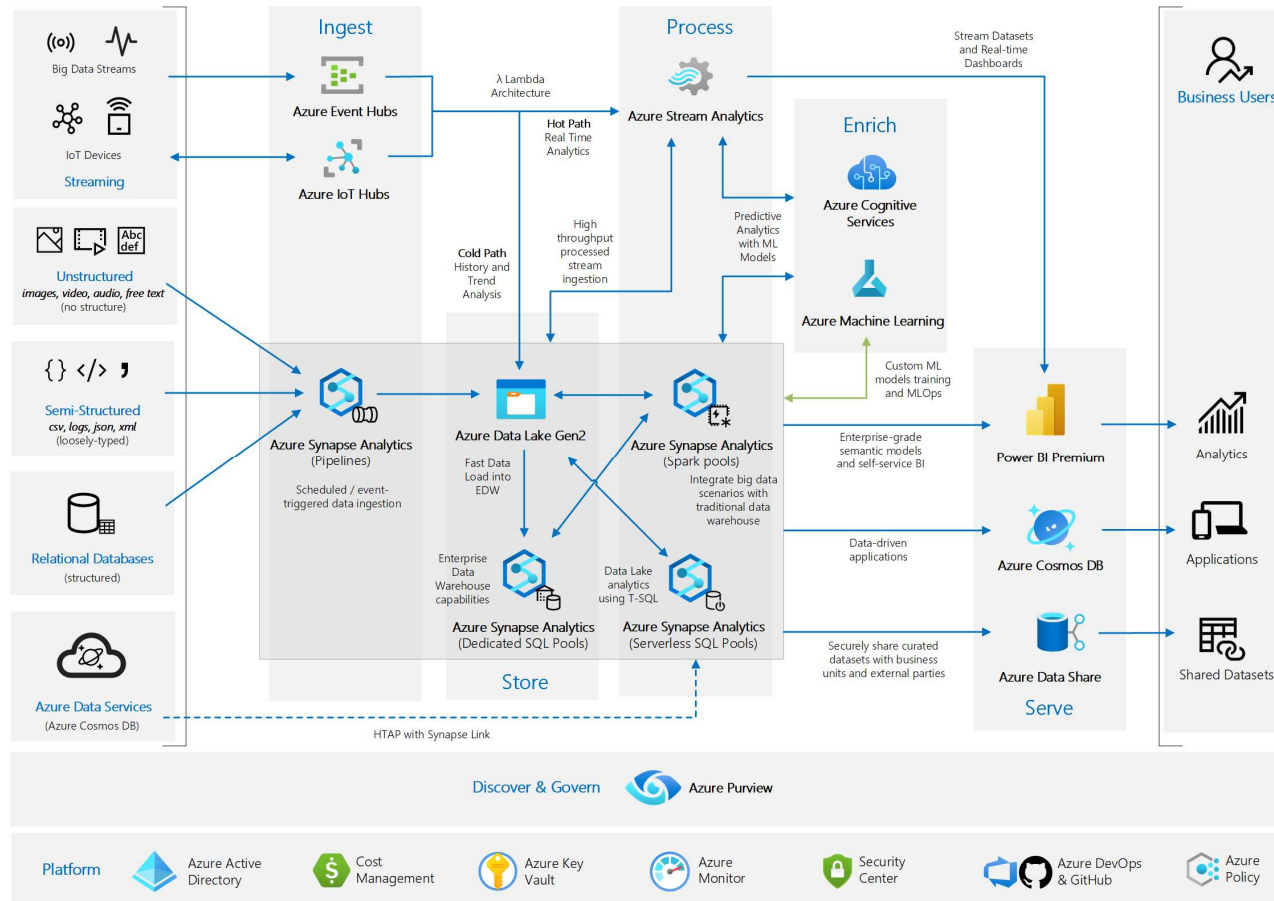


“...Data warehouses are for questions, whereas data lakes are for when organizations don’t know what questions to ask...”

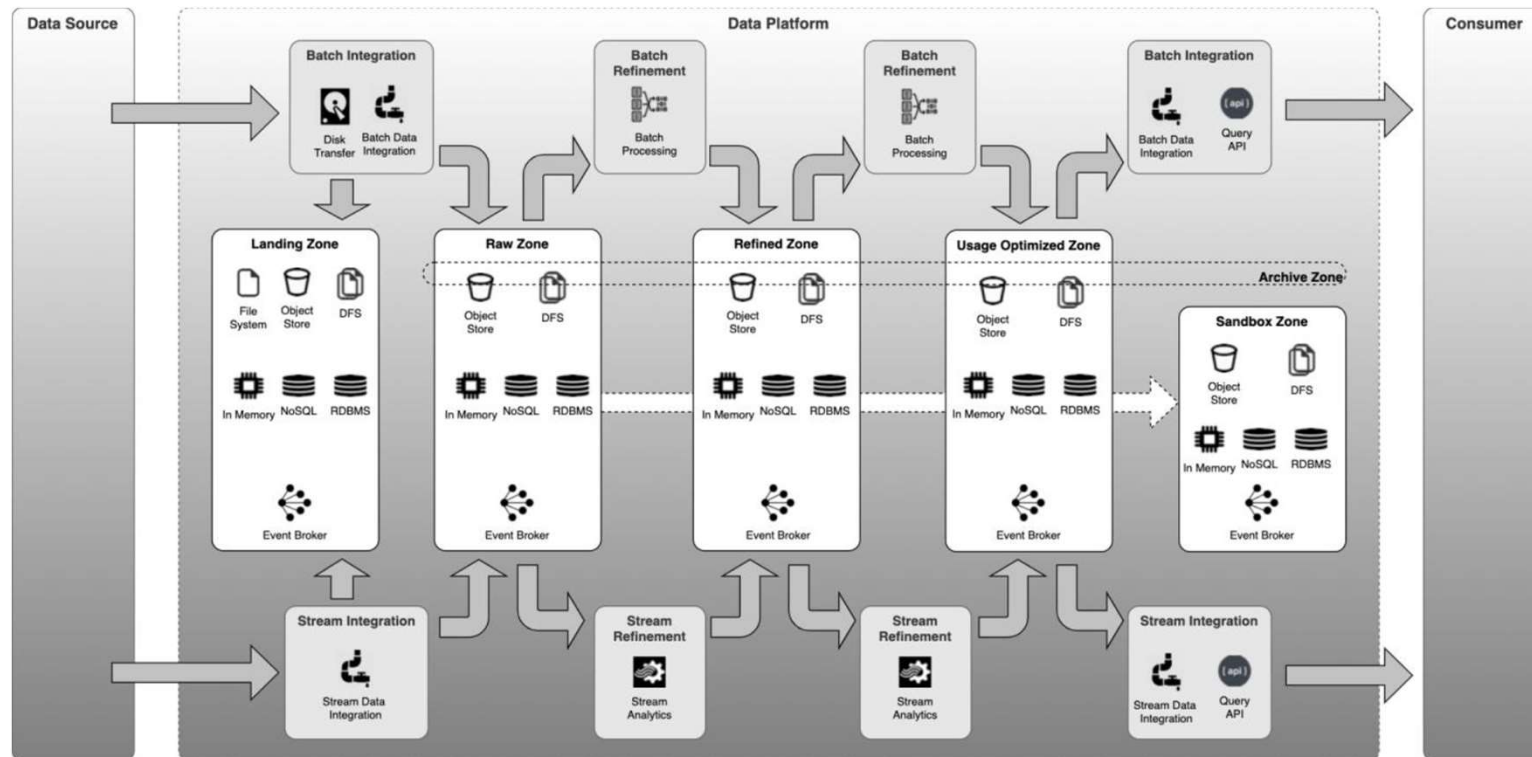
7 THE MODERN DATA WAREHOUSING PATTERN



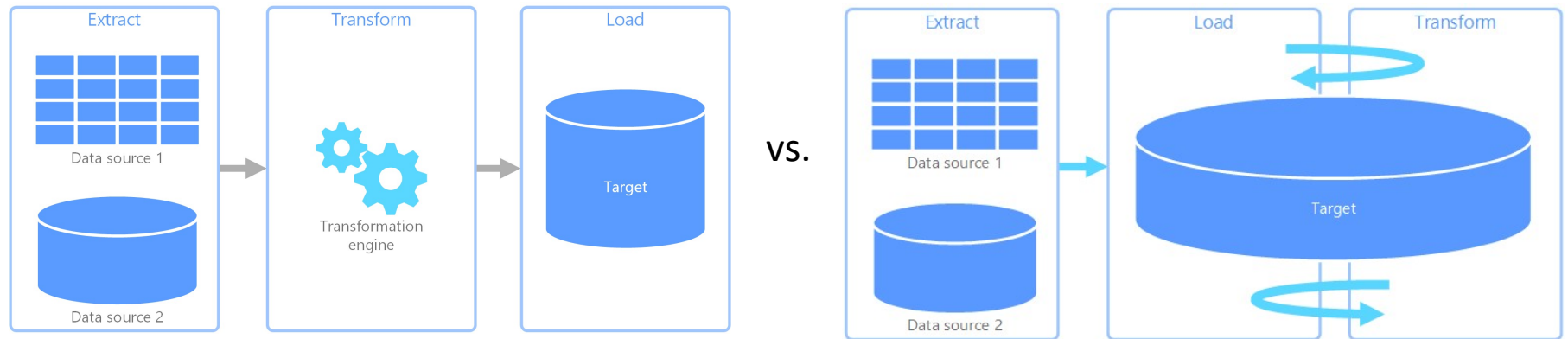
8 ANALYTICS END-TO-END WITH AZURE SYNAPSE



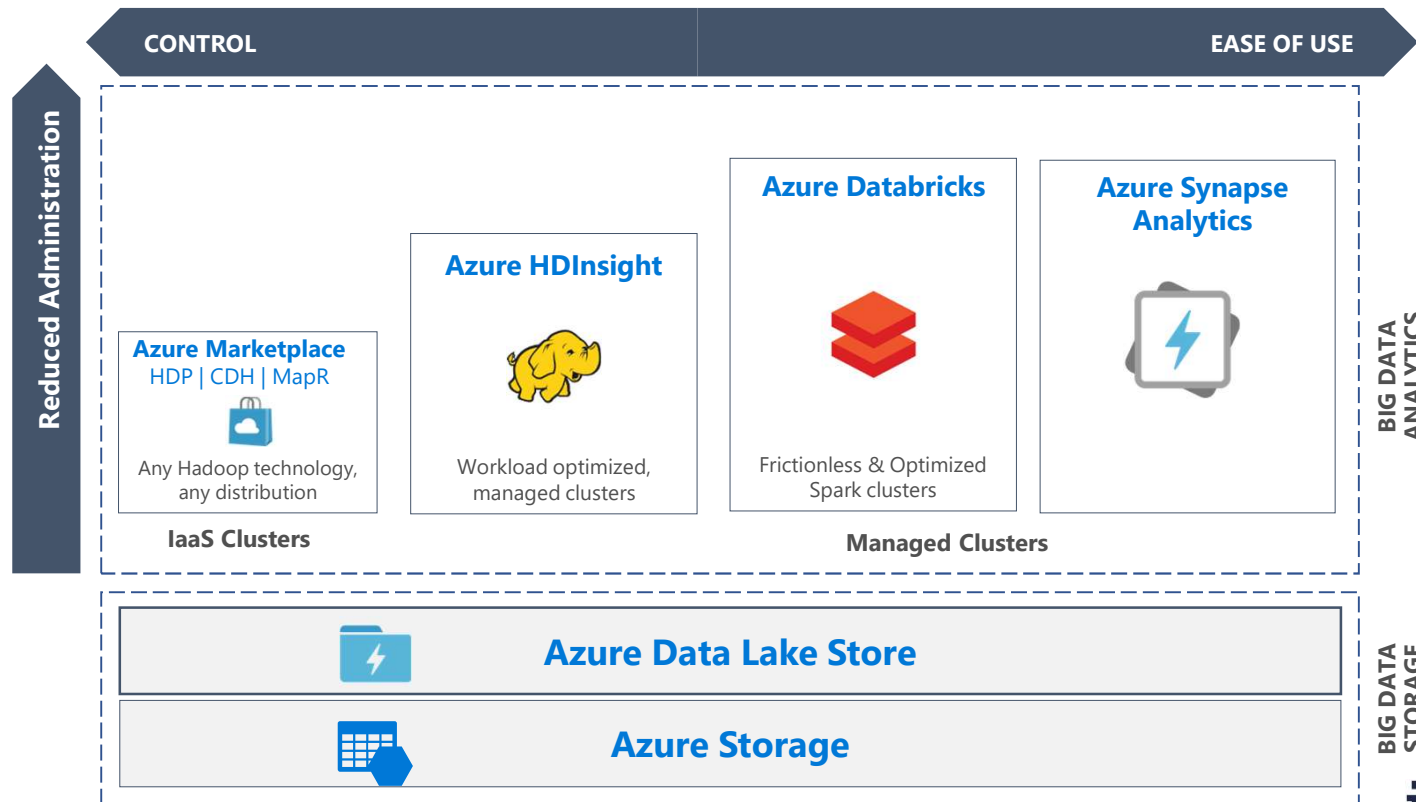
9 DATA LAKE ZONES



ETL vs. ELT

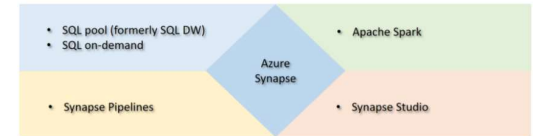


KNOWING THE VARIOUS BIG DATA SOLUTIONS

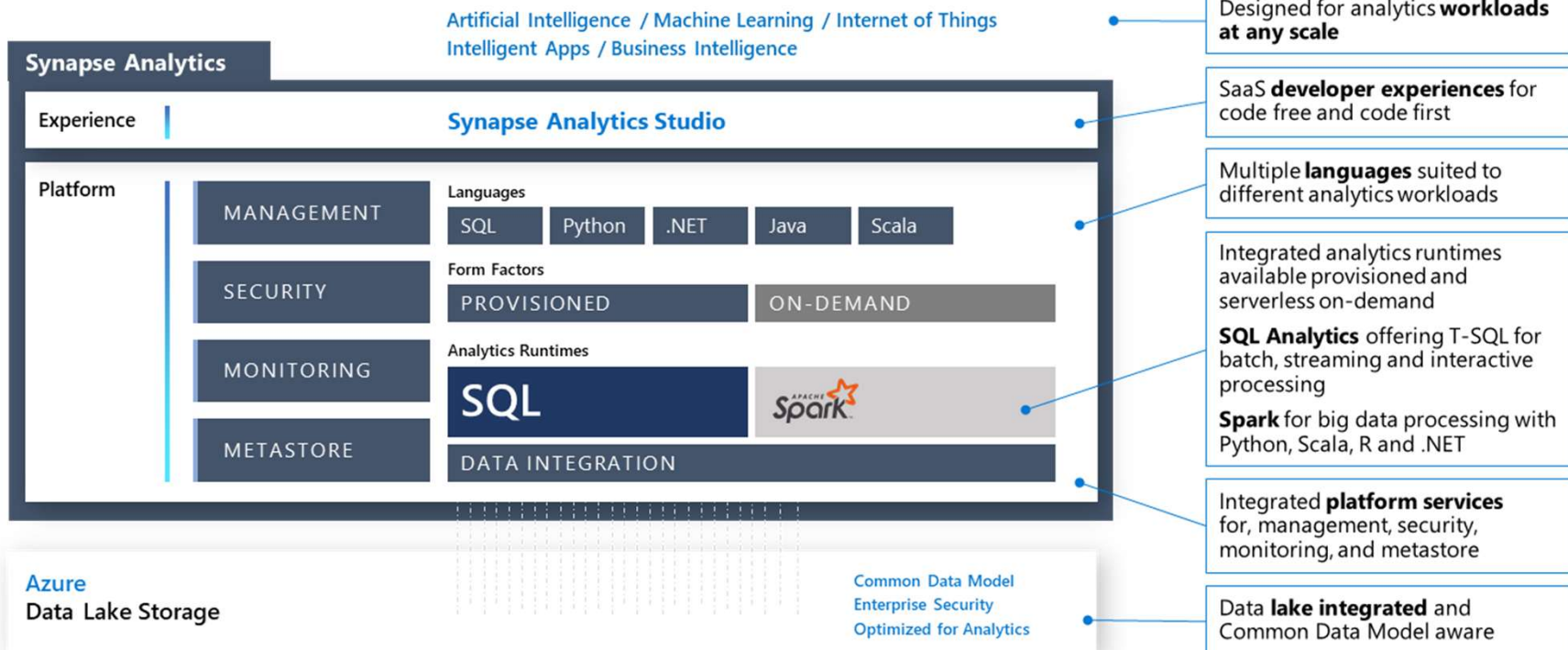


12 AZURE SYNAPSE ANALYTICS

Integrated data platform for BI, AI and continuous intelligence

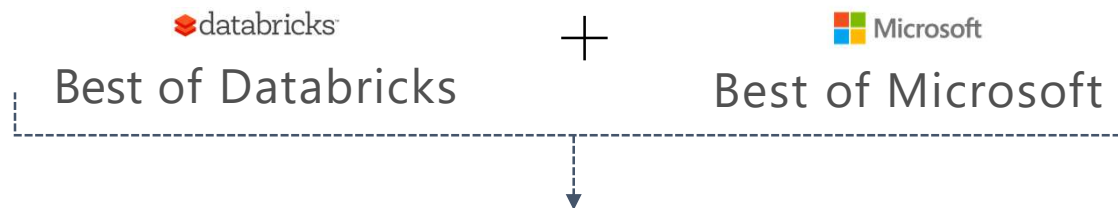



The four main components of Azure Synapse.




13 AZURE DATABRICKS

A FAST, EASY AND COLLABORATIVE APACHE® SPARK™ BASED ANALYTICS PLATFORM OPTIMIZED FOR AZURE



 Designed in collaboration with the founders of Apache Spark

 One-click set up; streamlined workflows

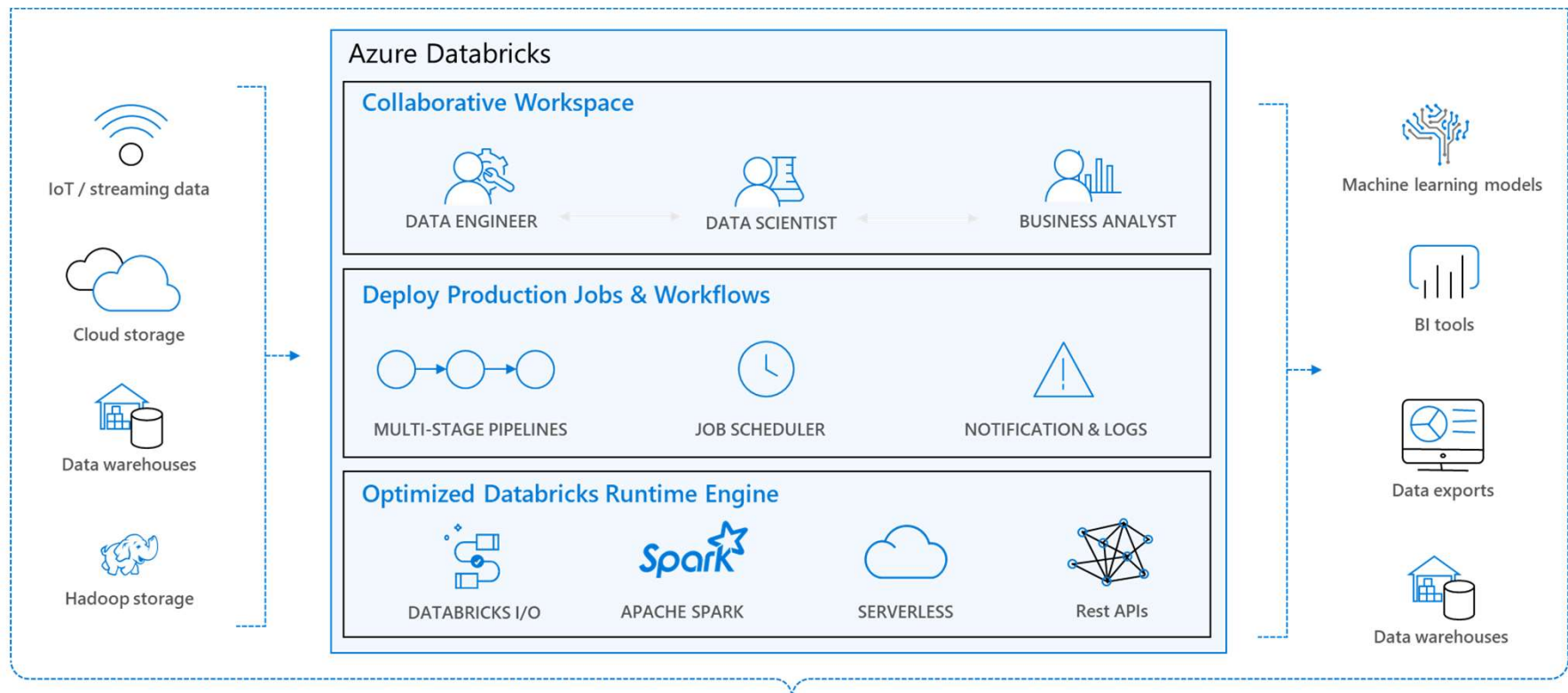
 Interactive workspace that enables collaboration between data scientists, data engineers, and business analysts.

 Native integration with Azure services (Power BI, SQL DW, Cosmos DB, ADLS, Azure Storage, Azure Data Factory, Azure AD, Event Hub, IoT Hub, HDInsight Kafka, SQL DB)

 Enterprise-grade Azure security (Active Directory integration, compliance, enterprise-grade SLAs)

trivadis
Part of **Accenture**

14 (AZURE) DATABRICKS



Enhance Productivity

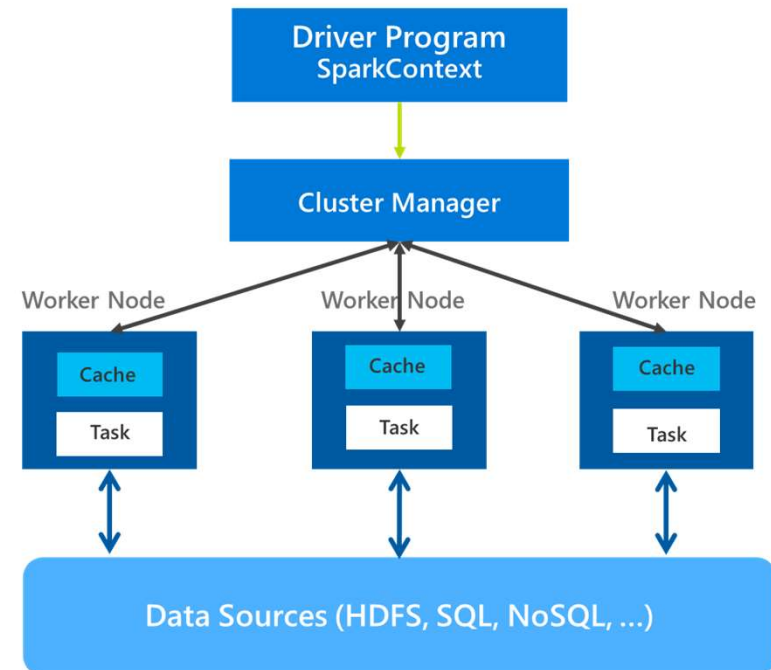
Build on secure & trusted cloud

Scale without limits

trivadis
Part of **Accenture**


15 AZURE DATABRICKS CLUSTERS

- Azure Databricks clusters are the set of Azure Linux VMs that host the Spark Worker and Driver Nodes
- Your Spark application code (i.e. Jobs) runs on the provisioned clusters.
- Azure Databricks clusters are launched in your subscription—but are managed through the Azure Databricks portal.
- Azure Databricks provides a comprehensive set of graphical wizards to manage the complete lifecycle of clusters—from creation to termination.



16 NOTEBOOKS ARE A POPULAR WAY TO DEVELOP, AND RUN, SPARK APPLICATIONS

Notebooks are not only for authoring Spark applications but can be *run/executed directly* on clusters

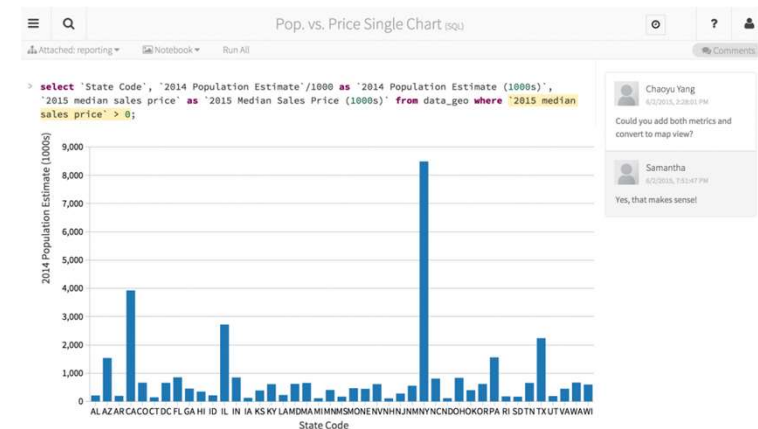
- **Shift+Enter**
- click the  at the top right of the cell in a notebook
- Submit via Job

Fine grained permissions support so they can be *securely shared* with colleagues for collaboration

Notebooks are well-suited for prototyping, rapid development, exploration, discovery and iterative development

With Azure Databricks notebooks you have a default language, but you can mix multiple languages in the same notebook:

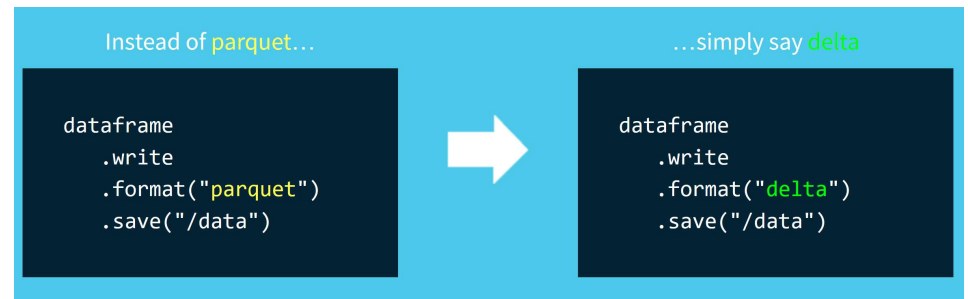
- `%python` Allows you to execute python code in a notebook (even if that notebook is not python)
- `%sql` Allows you to execute sql code in a notebook (even if that notebook is not sql).
- `%r` Allows you to execute r code in a notebook (even if that notebook is not r).
- `%scala` Allows you to execute scala code in a notebook (even if that notebook is not scala).
- `%sh` Allows you to execute shell code in your notebook.
- `%fs` Allows you to use Databricks Utilities - dbutils filesystem commands.
- `%md` To include rendered markdown



17 DELTA.IO

BRING ACID TO DATA LAKE

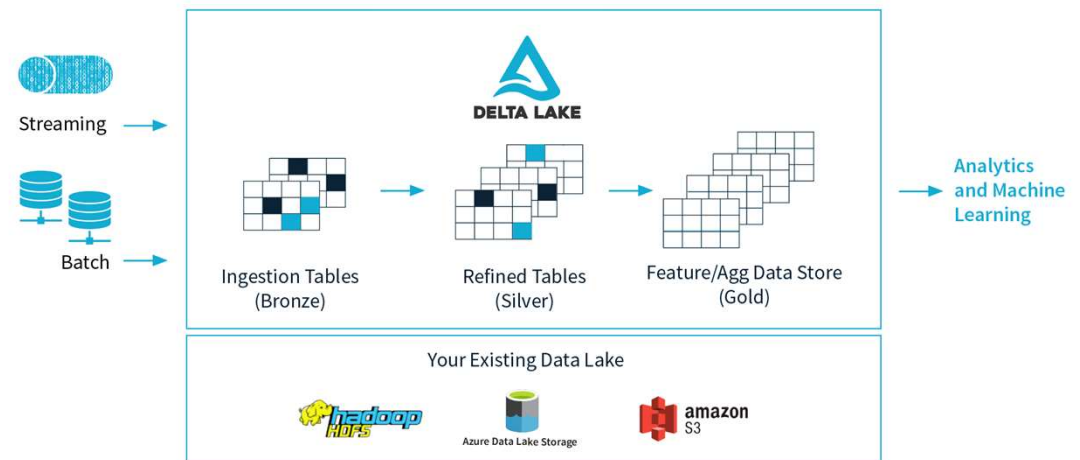
- ACID Transactions
- Separates compute and storage
- Open Format (based on Parquet)
- 100% Compatible with Apache Spark API
- Full DML Support (Update, Delete, Merge (Upsert))



18 DELTA.IO

BRING ACID TO DATA LAKE

- Time Travel (data versioning)
- Unified Batch and Streaming Source and Sink
- Scalable Metadata Handling
- Schema Enforcement
- Schema Evolution
- Audit History
- Compaction and indexing



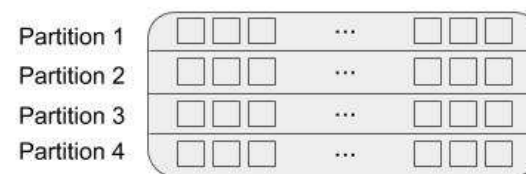
19 DELTA.IO

UPSERTS IN DATA LAKE

Updates to merge



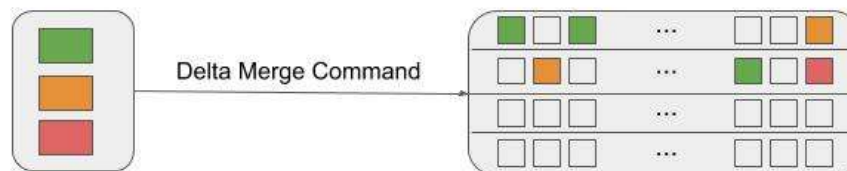
Target Table / Directory in Data Lake



Merging records in data lake **without** Databricks Delta



Merging records in data lake **with** Databricks Delta



20 MONSTER DEMO



FRAGEN...?