# Microsoft and Accenture GBHD Solution Design Session

July 30, 2020

# AGENDA

- GBHD Background
- Sequence of Events and Timeline
- Current State Review
- Proposed GBHD Future State in Azure
- Draft Bill of Materials (aka Commercial Pricing)
- Next Steps

# EXEC SUMMARY OF GLOBAL BROADBAND HEALTH DASHBOARD (GBHD)

**Background**

- Global Broadband Health Dashboard (GBHD) is a series of Qlik reports of broadband/WAN/wireless usage for the Accenture global offices
- GBHD is one of the three applications running in CAP under Yakuta's management; SNow Change Request/Config Management Reports and IP Address Mapping are the other two.
- Business purpose is to identify what offices are running hot and need extra capacity/new contracts with the WAN providers and what offices are below capacity
- Today GBHD is historical reporting of what happened. Accenture Team wants to evolve to ML as a future state to get predictive. Predictive analytics is important because it takes 4-5 weeks to renegotiate a contract so they don't want to be paying overage penalties.

**Current State Tech Stack & Consumers**

- There are a dozen or so Qlik dashboards running in AWS/Hortonworks/Hive
- Audience for the reports is less than 100 and 2-3 admins spend 30+ hours/week tweaking and managing the reports.
- Geoffray is running a model in Python for daily metrics.

**Key Date**

- AWS/CAP environment will be shutdown on April 1 so replacement solution needs to be running and performance tested in March to ensure there's a smooth transition.

**Objectives and End Goal**

- We should treat this as an opportunity to modernize an improve GBHD vs. a lift and shift to Azure (i.e replace IaaS with Azure PaaS offerings)
- Knowledge transfer is critical. Along the process MSFT and migration partner will assist in skilling up the Accenture team so they can take more ownership in the migration process and management of the new Azure platform.
- Incorporate ML so Accenture can evolve from historical reporting to predictive analytics for broadband usage forecasting

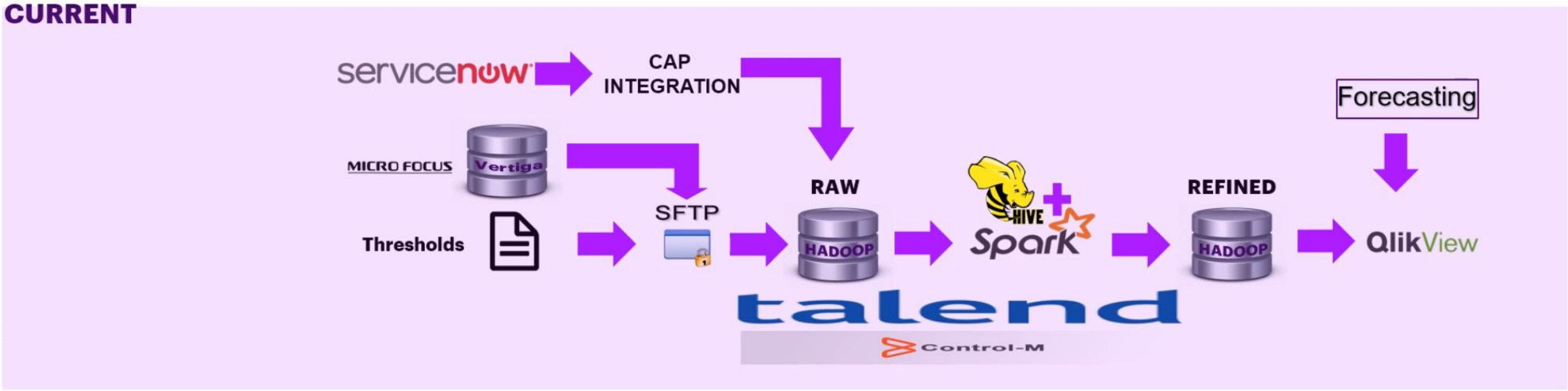
# SEQUENCE OF EVENTS FOR GBHD

| TOPIC | COMMENTS | TIMING | STATUS |
|---|---|---|---|
| **1) *Discovery*** | Series of meetings to understand the current landscape for the entire environment and desired outcomes for the MVP. | June | **Done** |
| ***2) Azure Design Session (ADS)*** | Azure Design Session is a workshop where we reconfirm our findings from the Discovery session and start whiteboarding out a solution architecture and approach. Outputs are:<br>**a) Solution Design:** Jon and Alex will have a draft solution design and we'll walk through the current state and proposed Azure future state with Yakuta and team.<br>**b) Azure Rough Order of Magnitude Bill of Materials**<br>**c) High Level Migration Timeline:** MSFT and SI partner, Neudesic created a draft for Yakuta's review<br>**d) First draft of Accenture roles and resources required for the migration** | a) (today,7/30)<br>b) 7/15<br>c) 8/3<br>d) 8/3 | |
| ***3) Skilling*** | Assumption that there will be Accenture practitioners who are new to Azure and will need skilling on the relevant Azure services | TBD | Open |
| **4) *MVP/POC*** | PoC key capability and migration areas (i.e. Hortonworks HQL → Databricks/Spark) | October? | Open |
| ***5) Detailed Migration Schedule*** | Roles and responsibilities between Microsoft, migration partner and Accenture stakeholders | TBD | Open |
| **6) *Production Rollout*** | Need to have Azure platform running at full capacity for X # of weeks before CAP shutdown in April (April 1? April 30?) shutdown of CAP. | March? | Open |

# Solution Architecture Current State and Proposed Azure Future State

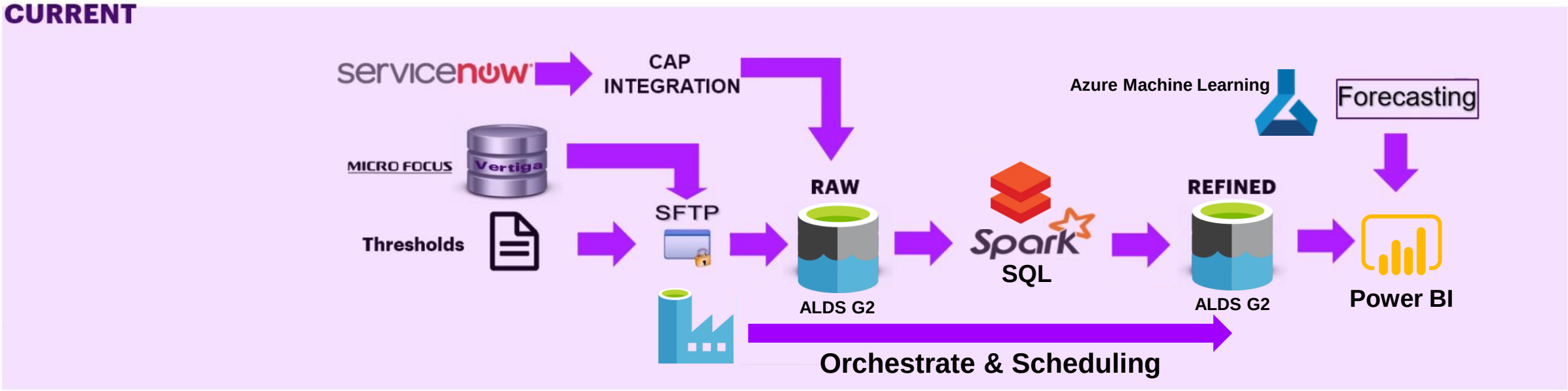
# CURRENT STATE

## GBHD MIGRATION MVP



# FUTURE STATE AZURE

## GBHD MIGRATION MVP

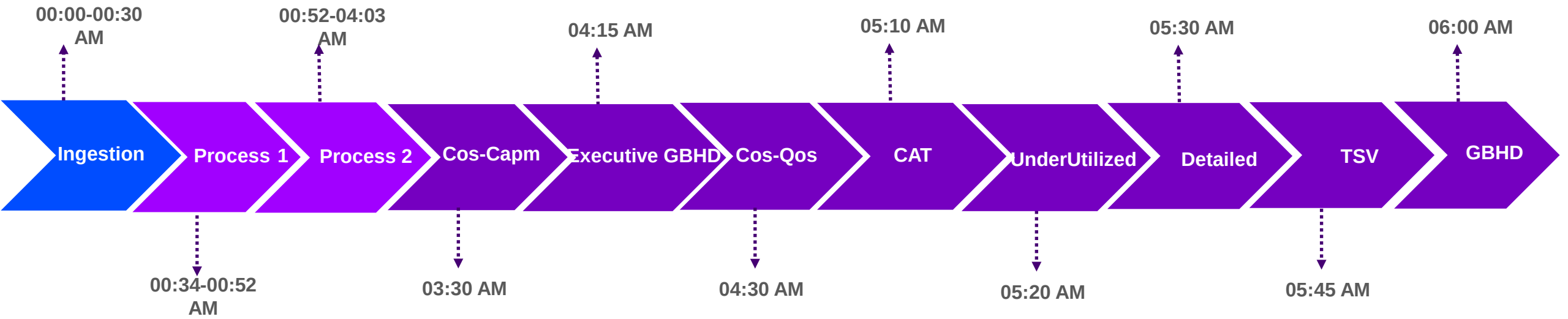


# GBHD DATA SOURCES

| | Details | Integration | Frequency | Field Info | Data Granularity | Data Volume |
|---|---|---|---|---|---|---|
| **CMT** | Link and costing details | Via EBI as a csv file in sftp folder | Once in a day | Microsoft Excel Worksheet | Gets total of 966 records - one per circuit | 280.2 KB |
| **Link Utilization** | Information related to devices and metadata | Via EBI as a csv file in sftp folder | Once in a day | Microsoft Excel Worksheet | 1 row for every 5 min interval per circuit | 152.6 MB |
| **Qos** | Information related to bandwidth used by each traffic class | Via EBI as a csv file in sftp folder | Once in a day | Microsoft Excel Worksheet | Aggregated matrix for traffic class per circuit | 119.6 KB |
| **MRDR** | Location hierarchy informati on | Already in Data Lake | | | | 143.6 KB |

Note: Currently around **500 GB** of CRAFT data is present in Data Lake spanning across 2 years

# GBHD DATA PROCESSING SCHEDULE




Ingestion

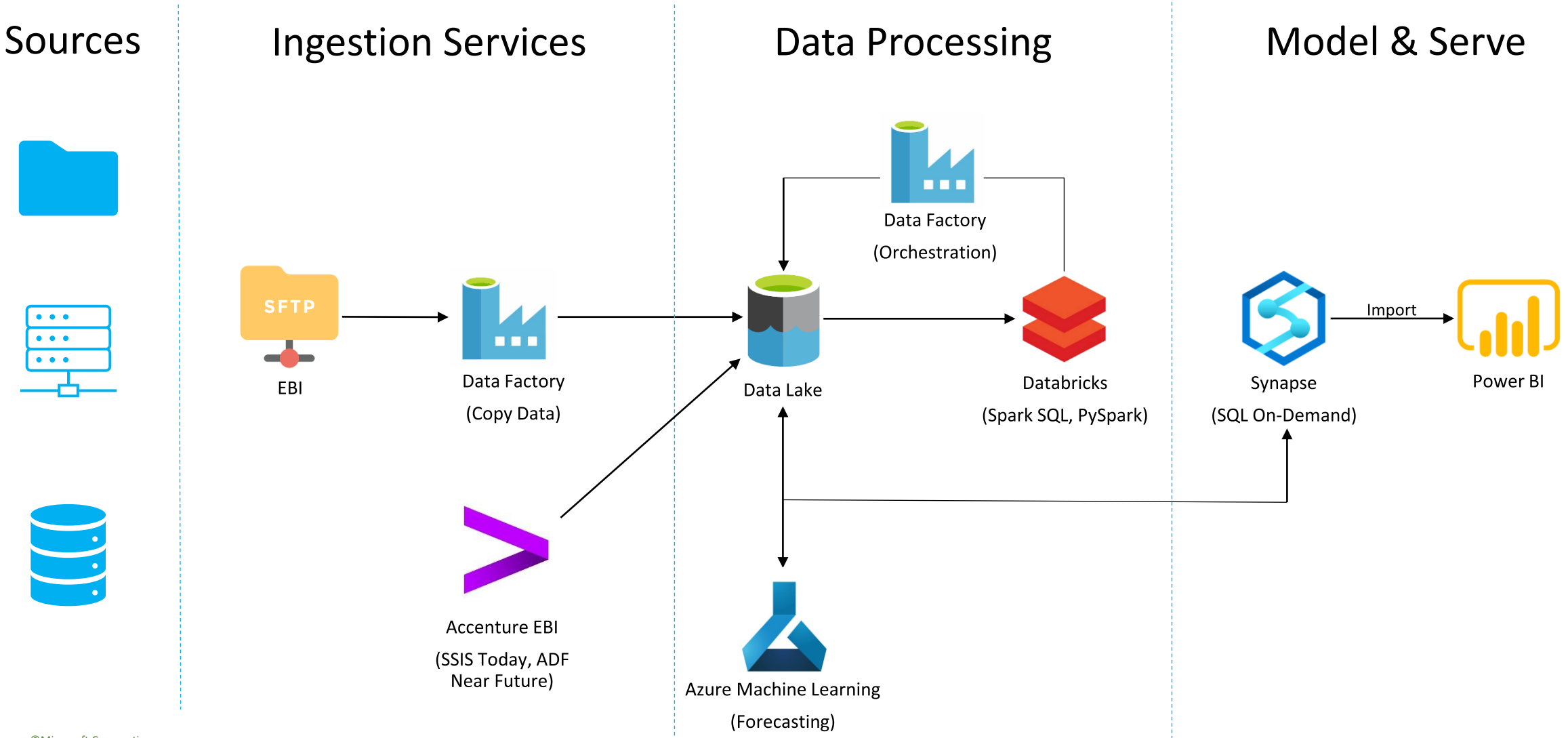
Process

Qlik Dashboard Refresh

# Current State *Tech* to Future Azure Services

| Layer in Architecture | Functions/Capability | Current State | Azure Future State |
|---|---|---|---|
| Ingestion Services | Scheduling<br>Orchestration<br>ETL | Talend | Azure Data Factory  |
| | SFTP<br>Data Copying | Accenture EBI<br>Talend<br>Control M | Accenture **EBI**<br>Azure Data Factory  |
| Data Processing | Data Lake | HortonWorks HDFS | Azure Data Lake  |
| | Data Processing | Talend<br>Spark<br>Hive QL | Azure Data Factory  <br>Azure Databricks (Spark SQL)  <br>Azure Synapse  |
| Reporting | Semantic modeling<br>Formulas<br>Data visualization | Qlik | Power BI  |
| Forecasting | Machine Learning<br>Data Science<br>Predictive Analytics | Local Jupyter Notebooks<br>Excel<br>Manual data sourcing | Azure Machine Learning  |

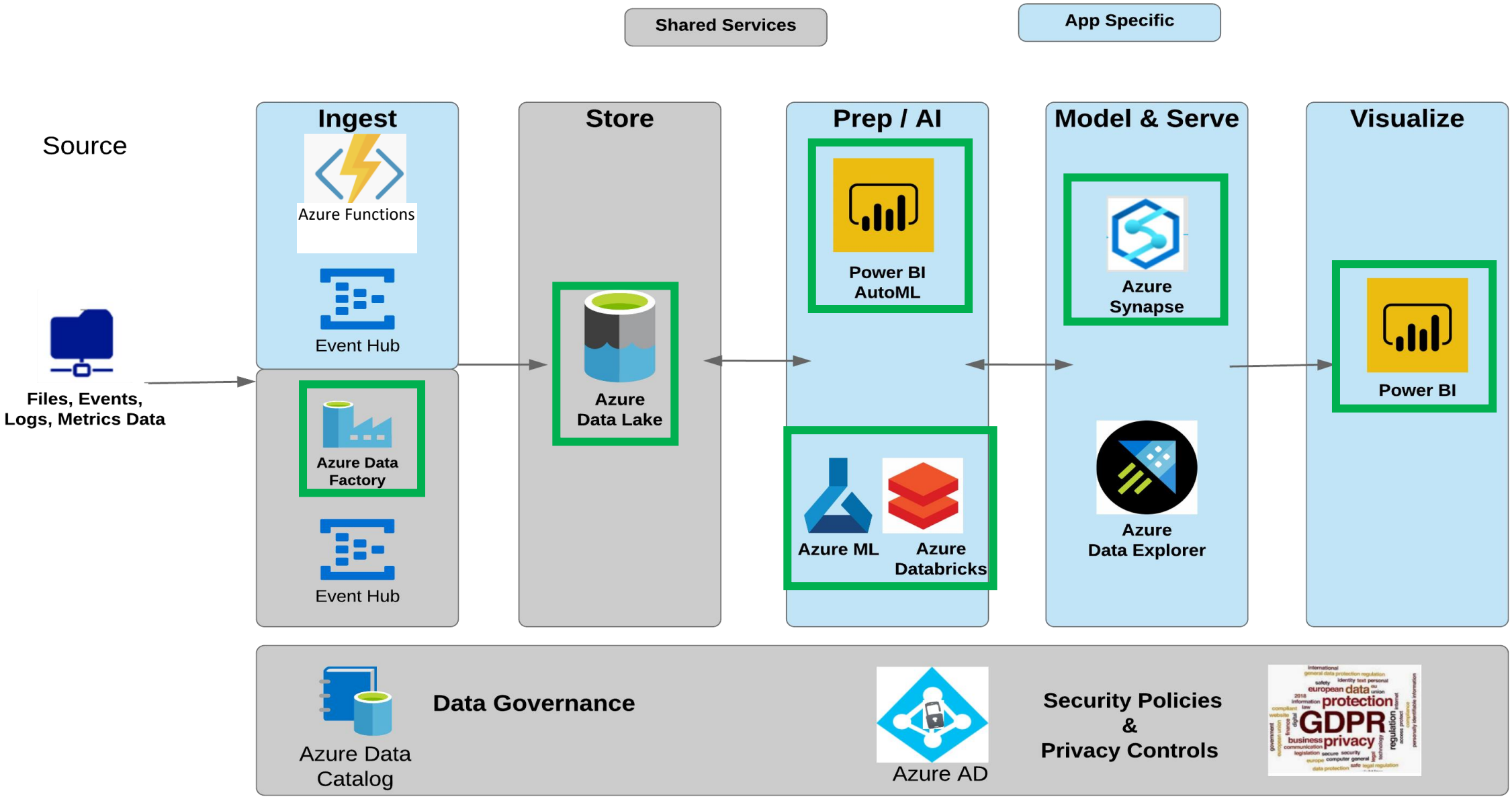
# Modern Data Warehouse for GBHD



# Accenture Enterprise IT Ops Lake Solution Architecture

All components below are certified or in process of being certified

= Proposed Azure Service for GBHD



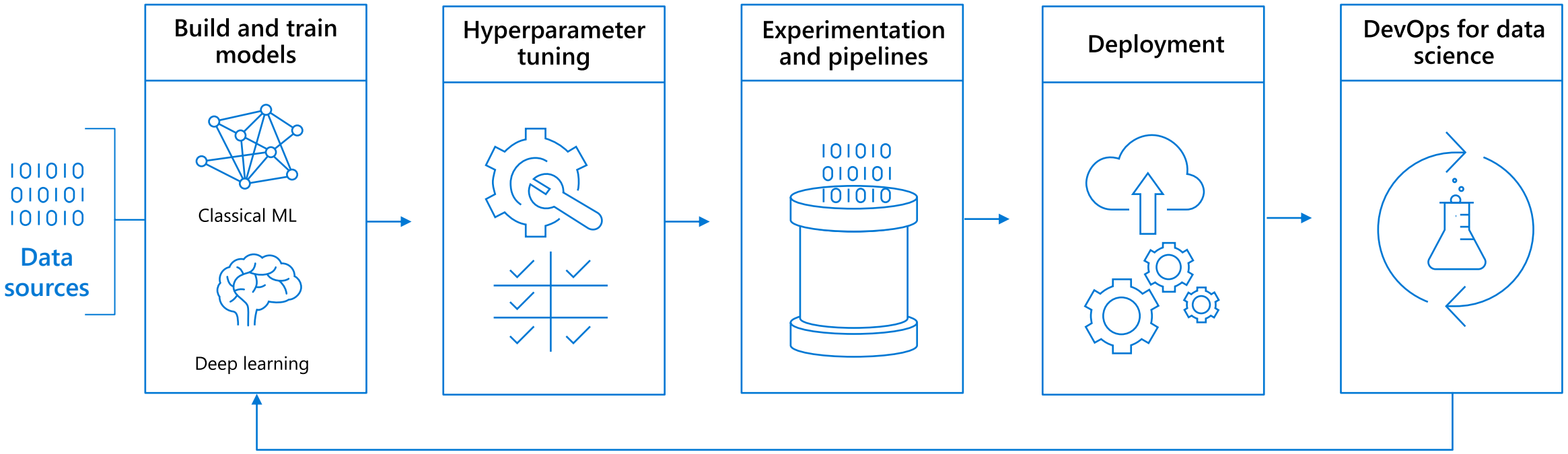
# Draft Bill of Materials for GBHD

| Service type | Custom name | Region | Description | Estimated monthly cost |
|---|---|---|---|---|
| Data Factory | Required | East US | Azure Data Factory V2 Type, Data Pipeline Service Type, Azure Integration Runtime: 250 Activity Run(s), 0 Data movement unit(s), 0 Pipeline activities, 1 Pipeline activities – External; Self-hosted Integration Runtime: 250 Activity Run(s), 0 Data movement unit(s), 0 Pipeline activities, 0 Pipeline activities – External, 0 x 8 Compute Optimized vCores, 0 x 8 General Purpose vCores, 0 x 8 Memory Optimized vCores, 1 Read/Write operation(s), 1 Monitoring operation(s) | $625.75 |
| Azure Data Lake v2 (aka Storage Accounts) | Required | East US | Data Lake Storage Gen2, Standard, GRS Redundancy, Hot Access Tier, Hierarchical Namespace File Structure, 5 TB Capacity - Pay as you go, Write operations: 4 MB x 100,000 operations, 100,000 List and Create Container Operations, Read operations: 4 MB x 100,000 operations, 10,000 Iterative write operations, 100,000 Other operations. 1,000 GB Data Retrieval, 1,000 GB Data Write, 1,000 GB Meta-data storage | $315.70 |
| Azure Databricks | Required | West US | Data Engineering Workload, Premium Tier, 8 DS12V2 (4 vCPU(s), 28 GB RAM) x 200 Hours, Pay as you go, 8 DBU x 200 Hours | $1,072.00 |
| Azure Synapse Analytics | Optional | East US | Tier: Compute Optimized Gen2, Synapse SQL (Provisioned): Compute: DWU 500 x 40 Hours, Storage: 0 TB with Geo-redundant disaster recovery | $302.00 |
| Azure Machine Learning | Optional | East US 2 | Enterprise, 2 DS5 v2 (16 Core(s), 56 GB RAM) x 400 Hours, Pay as you go | $732.80 |
| Support | | | **Support** | $0.00 |
| | | | **Licensing Program** | **Microsoft Online Services Agreement** |
| | | | **Total** | **$3,048.25** |
| | | | | |
| Disclaimer | | | | |
| *All prices shown are in US Dollar ($). This is a summary estimate, not a quote. For up to date pricing information please visit https://azure.microsoft.com/pricing/calculator/* | | | | |
| *This estimate was created at 7/13/2020 6:38:46 PM UTC.* | | | | |

- BOM above is for ***one*** Production environment
- BOM above is ***list price*** (Accenture will need to get EA discounted pricing from Accenture Procurement

# APPENDIX

# Building blocks for a Data Science Project



# Azure Machine Learning service

Set of Azure Cloud Services

Python SDK

That enables you to:

- ✓ Prepare Data
- ✓ Build Models
- ✓ Train Models
- ✓ Manage Models
- ✓ Track Experiments
- ✓ Deploy Models

# Building your own AI models

Transforming data into intelligence

SQL DB

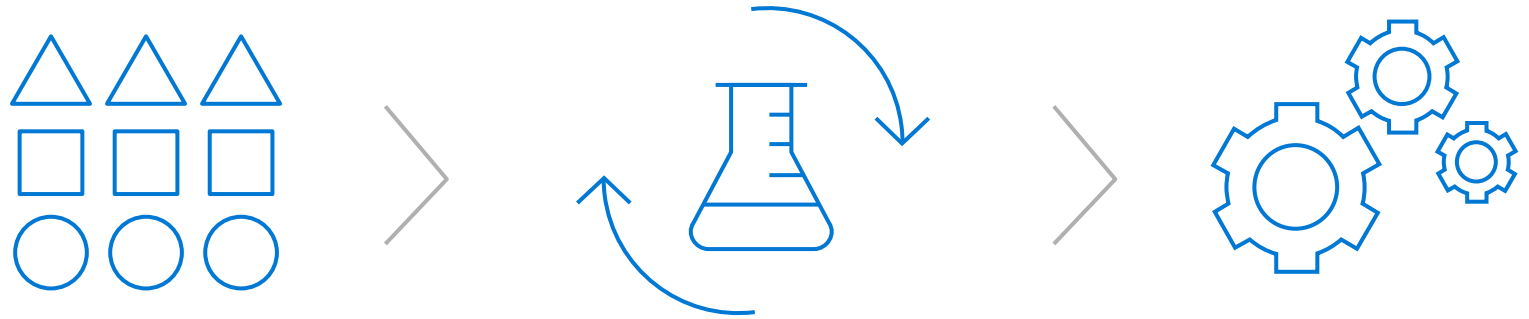
Cosmos DB

Datawarehouse

Data lake

Blob storage

...



Prepare data

Build and train

Deploy

# Azure Machine Learning service

Bring AI to everyone with an end-to-end, scalable, trusted platform

## Boost your data science productivity

## Increase your rate of experimentation

## Deploy and manage your models everywhere

## Built with your needs in mind

- Automated machine learning
- Managed compute
- Simple deployment
- DevOps for machine learning
- Support for open source frameworks
- Tool agnostic Python SDK

Seamlessly integrated with the Azure Portfolio

# Azure Machine Learning services

**Experience**
SDK, Notebooks, Drag-n-drop, Wizard

**MLOps**
Reproducible, Automatable, GitHub, CLI, REST

**Datasets**
Profiling, Drift, Labeling

**Training**
Experiments, Runs

**Model Registry**
Models, Images

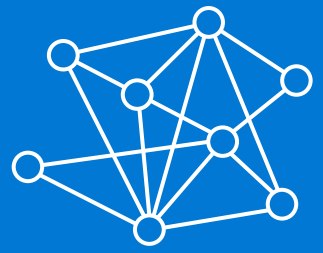
**Inferencing**
Batch, Realtime

**Compute**
Jobs, Clusters, Instances

**Azure IoT Edge**
Security, Mgmt., Deployment

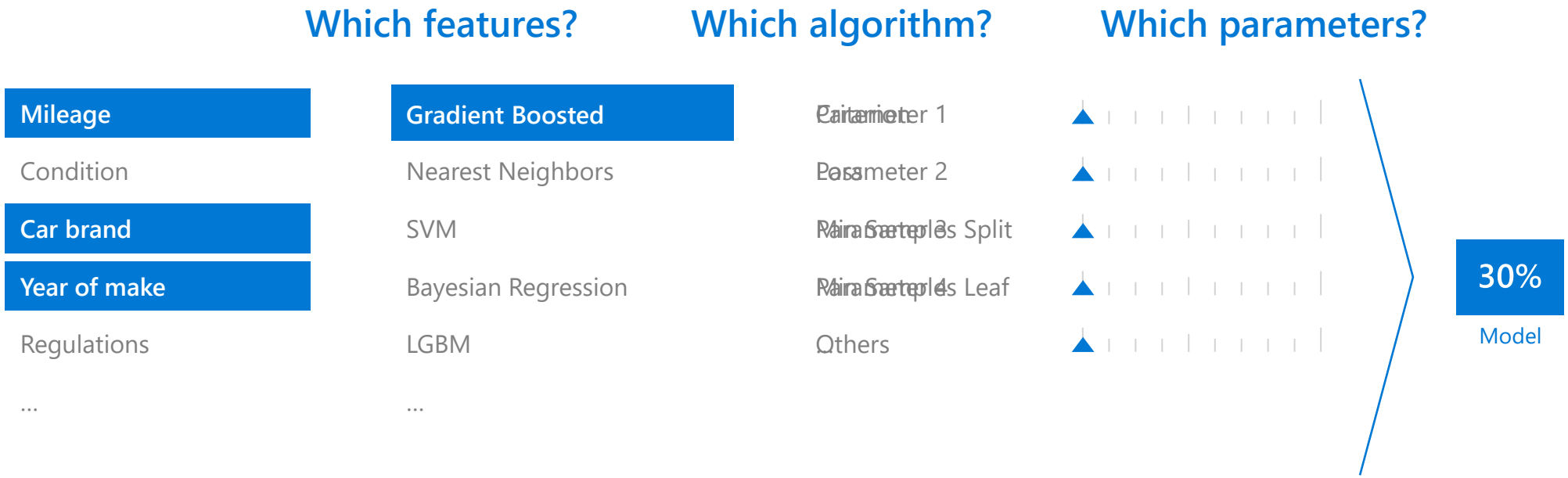
**Cloud**
CPU, GPU, FPGA

**Edge**
CPU, GPU, NPU



# Automated machine learning

# Model creation is typically a time consuming process



# Model creation is typically a time consuming process

## Which features?

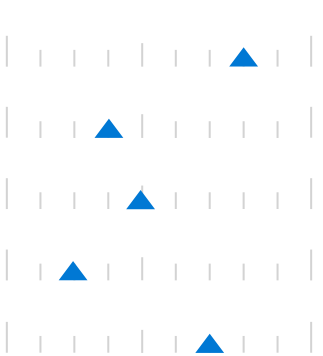
- Mileage
- Condition
- Car brand
- Year of make
- Regulations
- ...

## Which algorithm?

- Gradient Boosted
- Nearest Neighbors
- SGD
- Bayesian Regression
- LGBM
- ...

## Which parameters?

- Number of Neighbors
- Weights
- Min Samples Split
- Min Samples Leaf
- XYZ



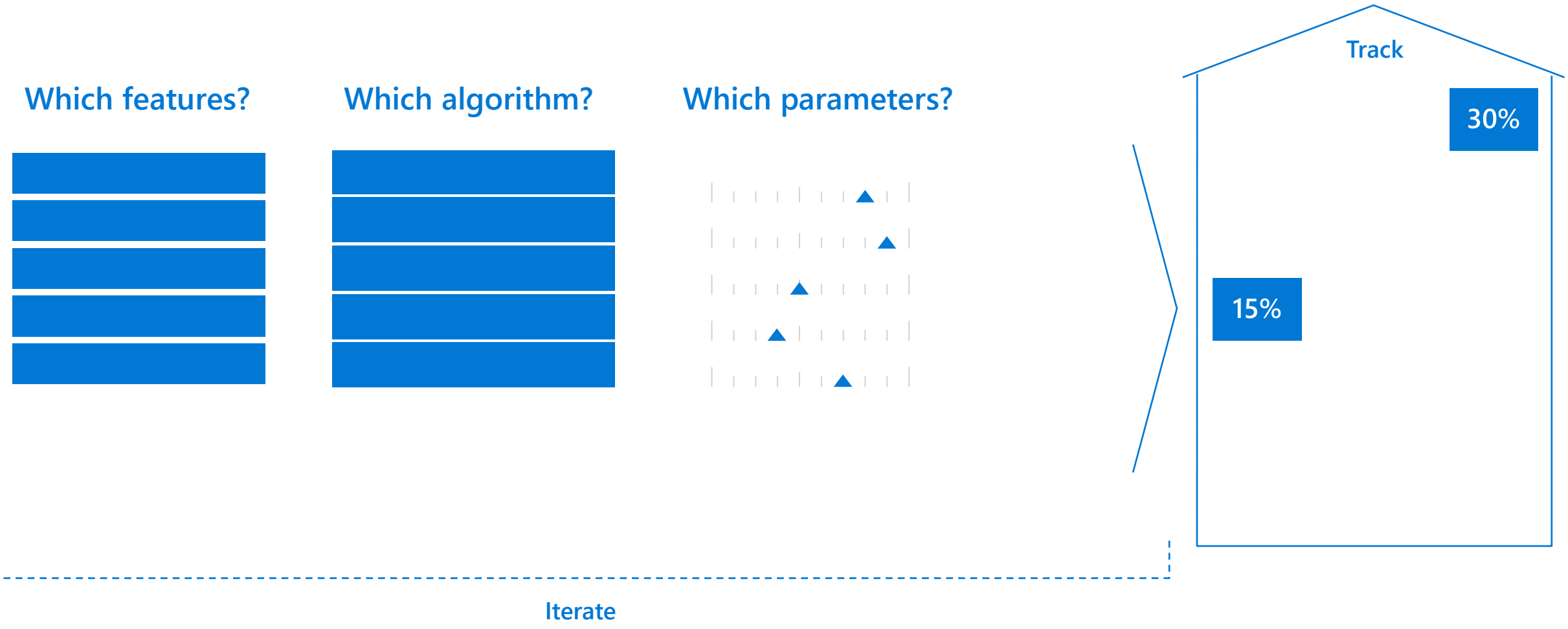
Track

30%

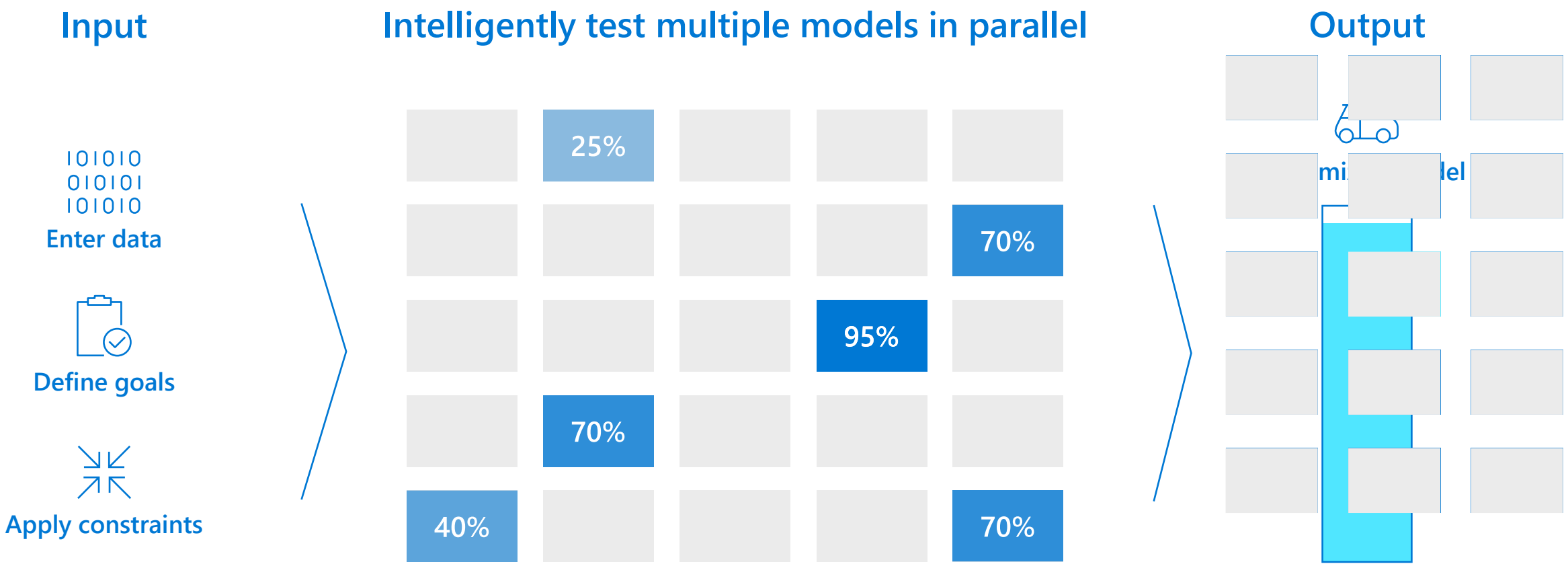
Model

Iterate

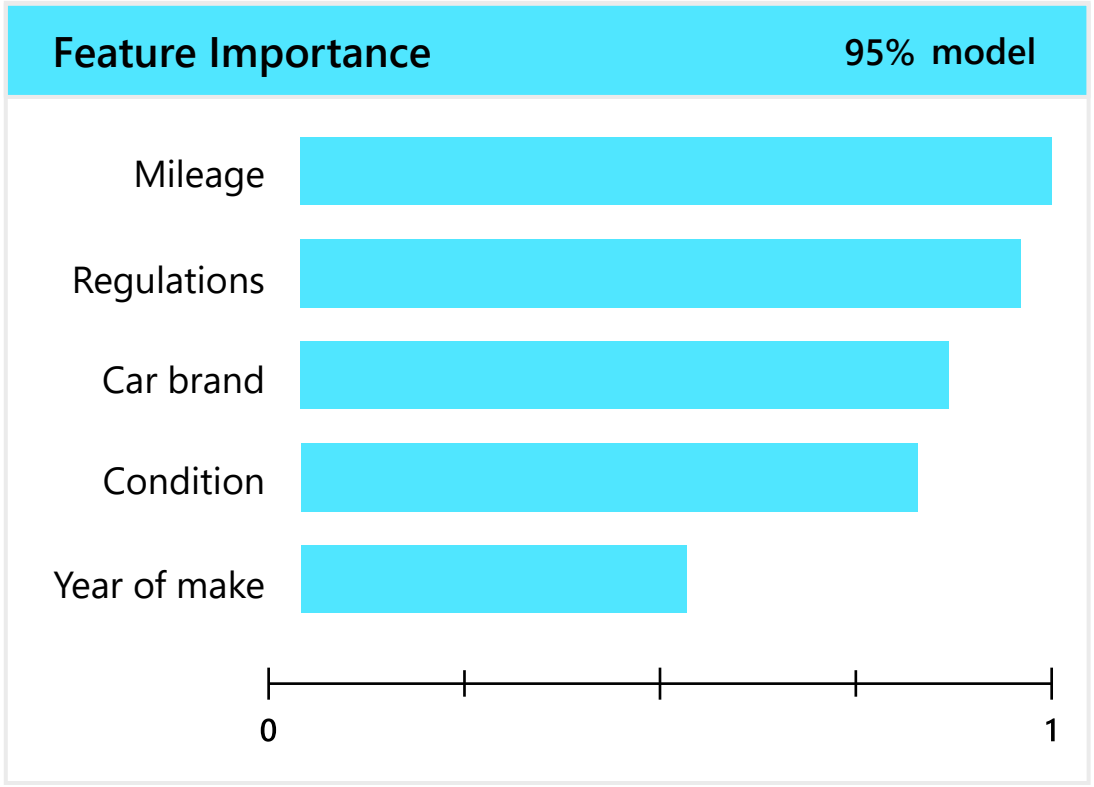
# Model creation is typically a time consuming process

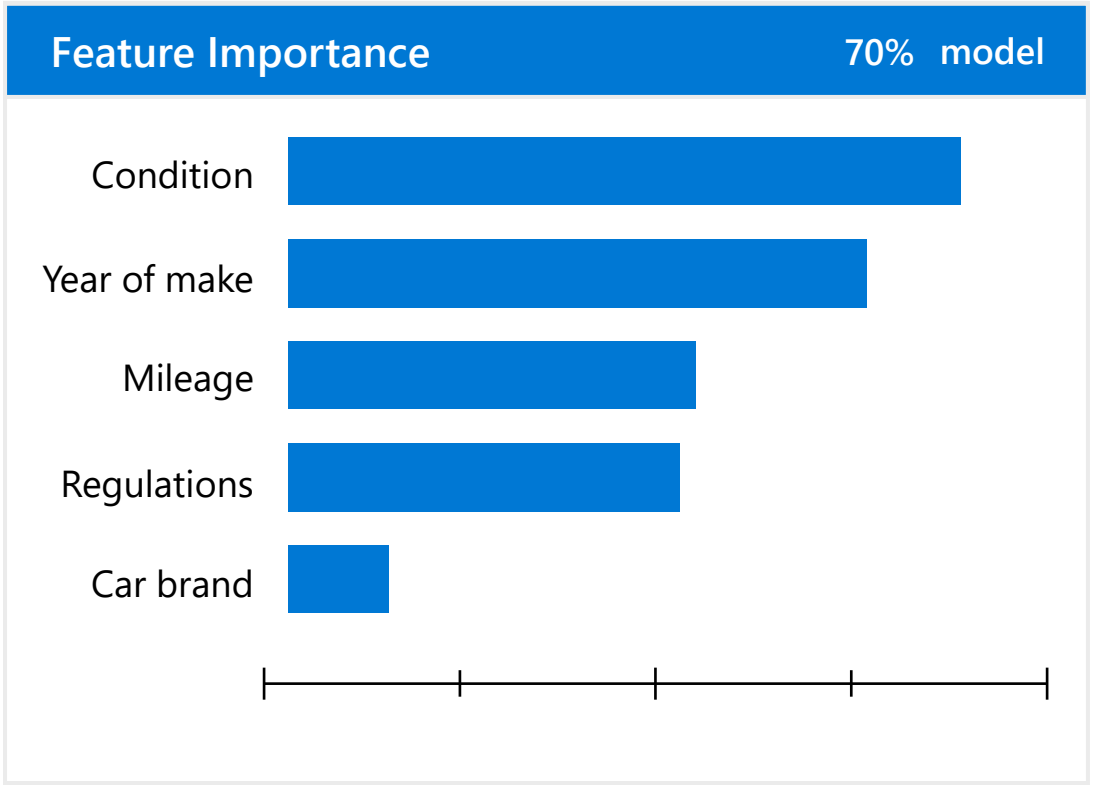


# Automated Machine Learning accelerates model development



# Understand the inner workings of ML by analyzing feature importance




Enable model explain-ability for every automated ML iteration, not just the optimal model

# Why converge enterprise & self-service BI?

Customers must choose Azure AS or Power BI

- Once chosen, difficult to change
- Migration cost from Azure AS to Power BI Premium

Difficult to promote datasets to centralized BI

- Reduced business agility

Customer confusion around messaging

**Enterprise BI**

Azure
Analysis Services

**Self-service BI users**

Power BI

# Why converge enterprise & self-service BI?

One-stop shop for all BI requirements

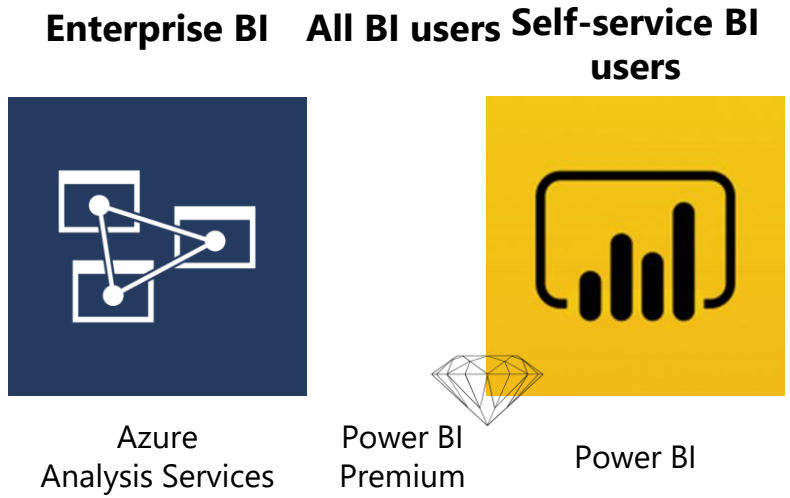
- Colocation of BI artifacts
- Native integration with Power BI eco-system
- Simplified messaging

Easy access to next-gen features

- Start in Power BI, no migration required

Promotion of datasets to centralized BI

- Improved collaboration between the business & IT



# Feature disparity

** In development / preview*

## Azure AS

- ✔ Large models
- ✔ XMLA endpoint:
  - ✔ SSDT/SSMS/community tools
  - ✔ Source control/DevOps
- ✔ AS engine semantic modeling
  - ✔ Calculation groups
- ✔ Azure Log Analytics
- Backup restore
- Reliable processing REST API
- Query scale out

Tabular models

## Power BI Premium

- Paginated reports, dataflows, AI workloads
- Unlimited consumers
- Incremental refresh policies
- Composite models and aggregations
- BYOK, VNETs & private links *
- Dynamic memory management / eviction
- Live connect composite models *
- Interactive reports in service
  - Metadata translations
  - App sharing
  - Shared/certified datasets
- Power BI Desktop authoring
- Deployment Pipelines *
- Azure Synapse integration *
- Data protection
- Data lineage
- Azure Data Catalog integration *
- Lots of data sources / connectors
- Scheduled refresh in the PBI service

# Power BI as a superset of AS

** In development / preview*

**Power BI Premium ***

**Azure AS**

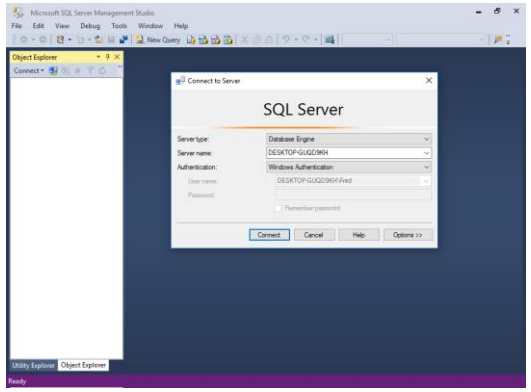
- Large models
- XMLA endpoint:
  - SSDT/SSMS/community tools
  - Source control/DevOps
- AS engine semantic modeling
  - Calculation groups
- Azure Log Analytics
- Backup restore
- Reliable processing REST API
- Query scale out

Tabular models

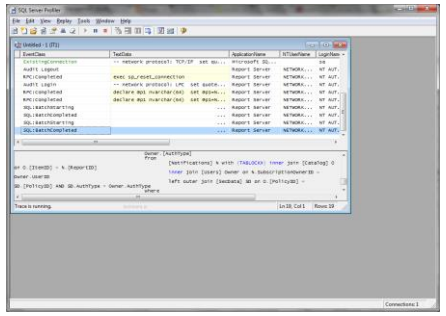
- Paginated reports, dataflows, AI workloads
- Unlimited consumers
- Incremental refresh policies
- Composite models and aggregations
- BYOK, VNETs & private links *
- Dynamic memory management / eviction
- Live connect composite models *
- Interactive reports in service
  - Metadata translations
  - App sharing
  - Shared/certified datasets
- Power BI Desktop authoring
- Deployment Pipelines *
- Azure Synapse integration *
- Data protection
- Data lineage
- Azure Data Catalog integration *
- Lots of data sources / connectors
- Scheduled refresh in the PBI service

# XMLA endpoint for Power BI Premium

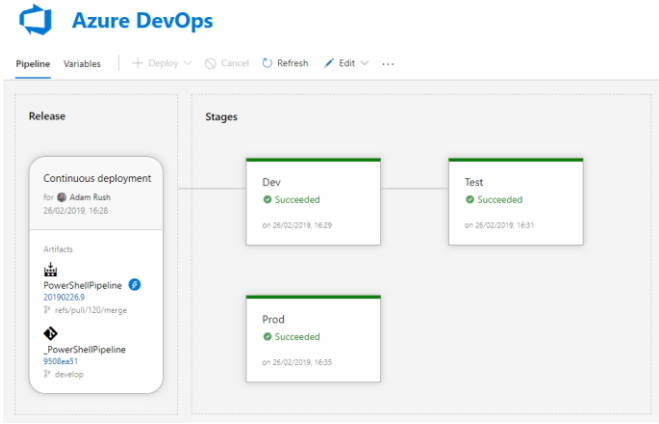
Backwards compatibility with Analysis Services tools and processes



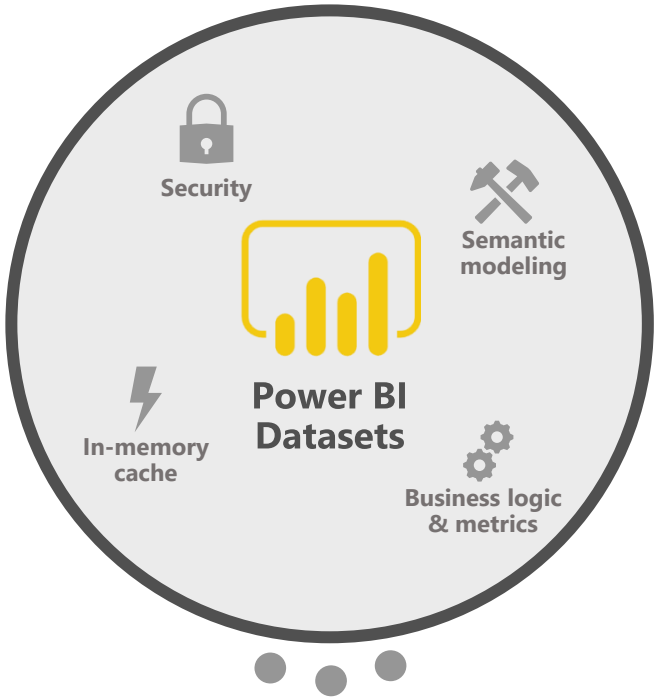

SQL Server Management Studio for **management**

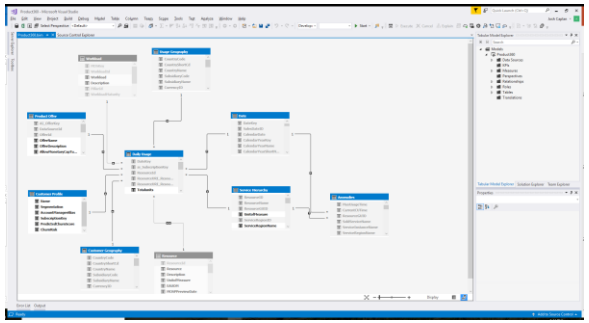


SQL Server Profiler for **debugging**

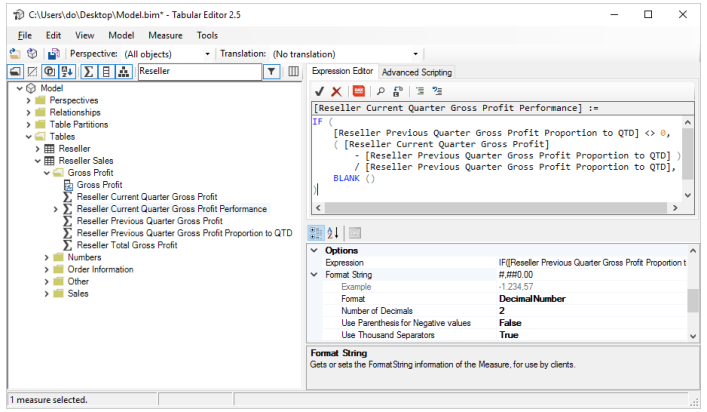



Azure DevOps for **source control & CI/CD**



SQL Server Data Tools for **modeling**

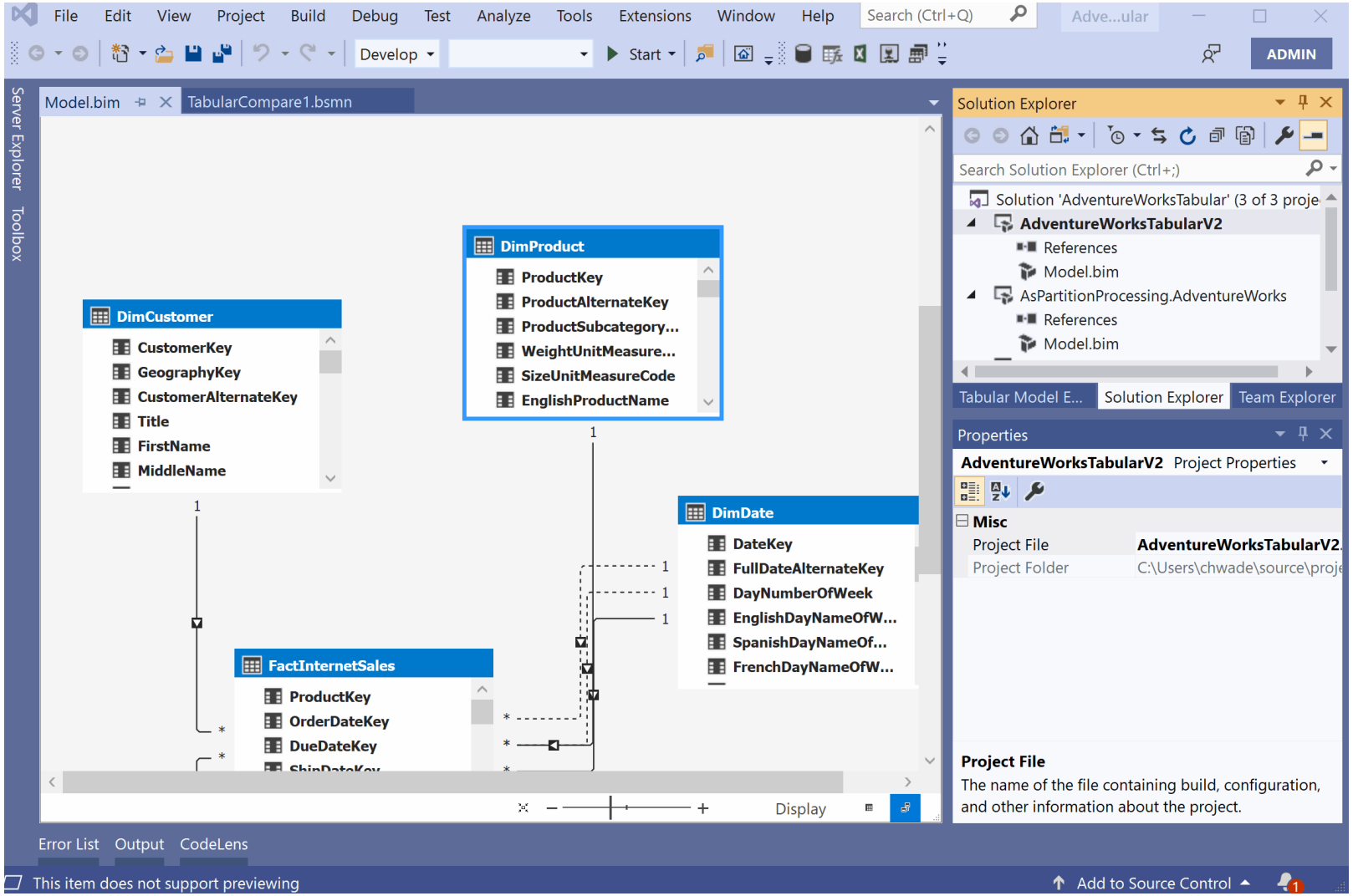



Community tools and **non-Microsoft** data-viz tools

# Compatibility with Analysis Services tools

Deploy from SSDT with the **XMLA endpoint**

SSDT integrates with Azure DevOps



# Auto ML for Power BI dataflows

## Auto ML

Self-service ML models for BI Pros & data analysts

**Simplified wizard flow**
Creates ML models for Binary Predictions, Classifications & Forecasting with just a few clicks

**Advanced configuration with Power Query**
Model customization with feature selection, filters, and more

**Model management**
Model validation, retraining, etc..

**Explanability**
Model interpretation of key influencers targeted at business users

**Composition with ETL**
Apply for data ingestion and refresh

# Current AI investments in Power BI


End users


Analysts


BI Professionals

Data Scientists

| Explanations | Cognitive AI | Scripting | Auto ML | Azure ML |
|---|---|---|---|---|
| Natural Language exploration<br>Explore influencers<br><br>Capabilities:<br>Quick Insights<br>Q&A<br>Explorations with Curie | Out of box AI in PQ<br>One-click AI transformations<br><br>Capabilities:<br>Sentiment analysis<br>Key phrase extraction<br>OCR, Text Translations | Transforms & Visualizations<br>Custom Python models<br><br>Languages:<br>R<br>Python | Generate ML models in clicks<br>Turnkey model deployments<br><br>Capabilities:<br>Predictions, Classifications, Forecasting, Clustering, Recommendations | PQ integration for Azure ML<br>Other Azure hosted models<br><br>Integration:<br>Azure ML<br>Azure Frameworks |


Azure Synapse Analytics

# 2019

Figure 1. Magic Quadrant for Data Management Solutions for Analytics


Source: Gartner (January 2019)

***Azure SQL Data Warehouse (aka Synapse SQL Provisioned)***

# Analytics & AI is the #1 investment for business leaders, however they struggle to maximize ROI

**80%**

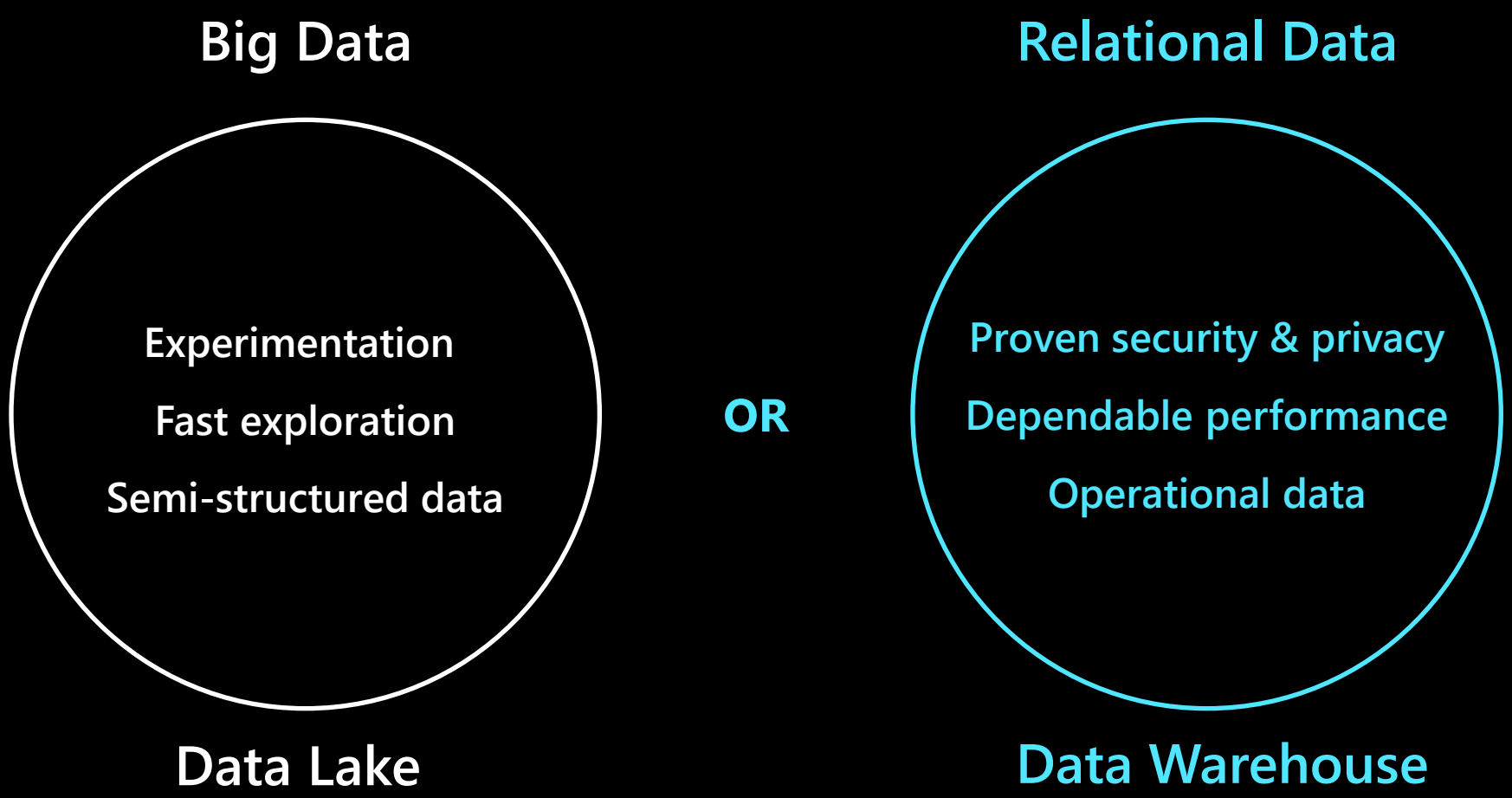
report struggling to become mature users of data*

**55%**

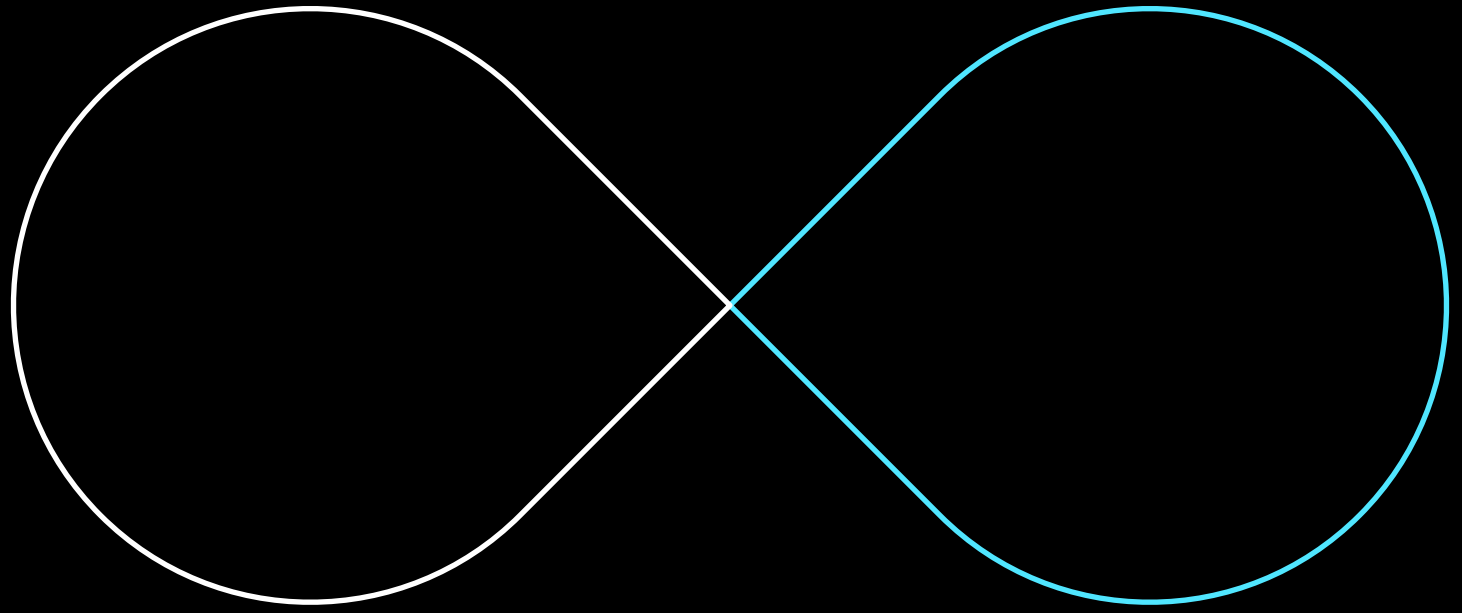
report data silos and data management difficulties as roadblocks*

* Harvard Business Review (2019), Understanding why analytics strategies fall short for some, but not for others

# Businesses are forced to maintain two critical, yet independent analytics systems

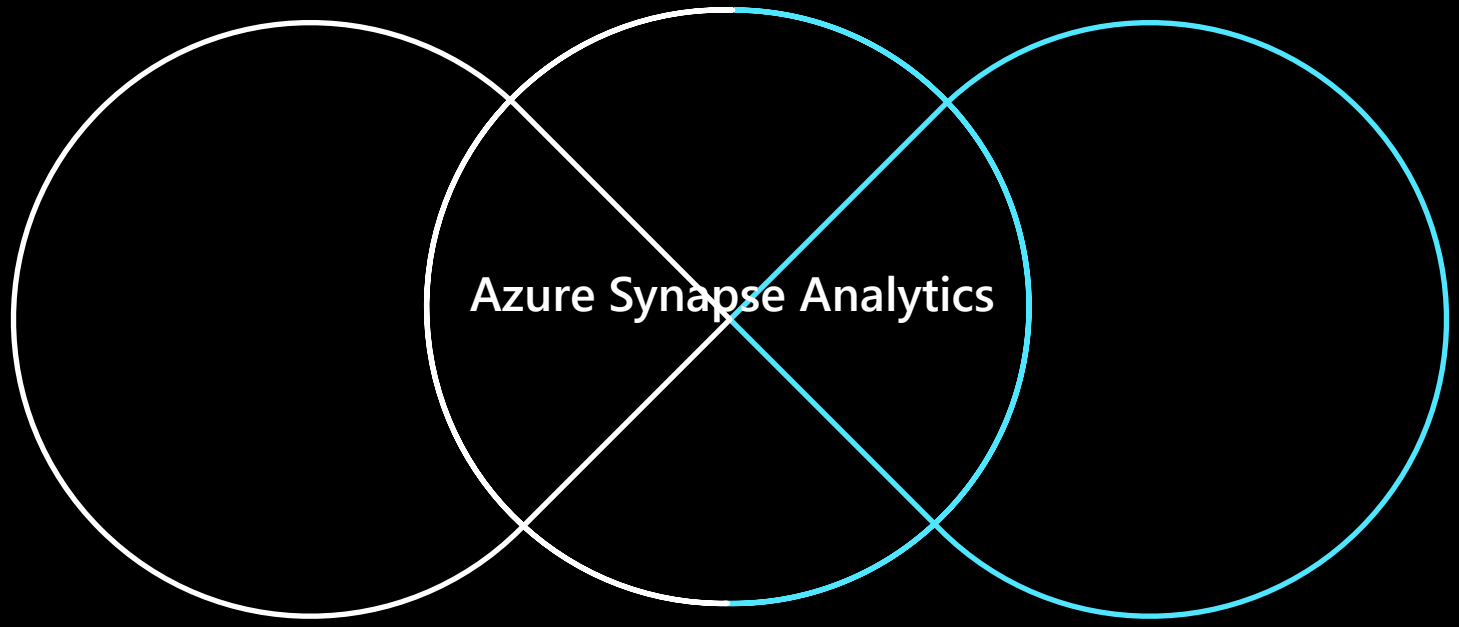


# Azure brings these two worlds together, in a single service, to provide limitless analytics

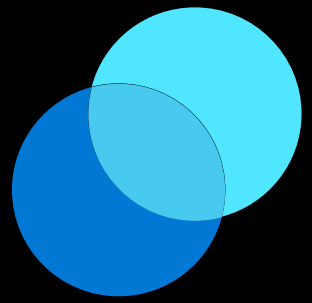


Data warehousing & big data analytics—all in one service

# Azure brings these two worlds together, in a single service, to provide limitless analytics




Data warehousing & big data analytics—all in one service

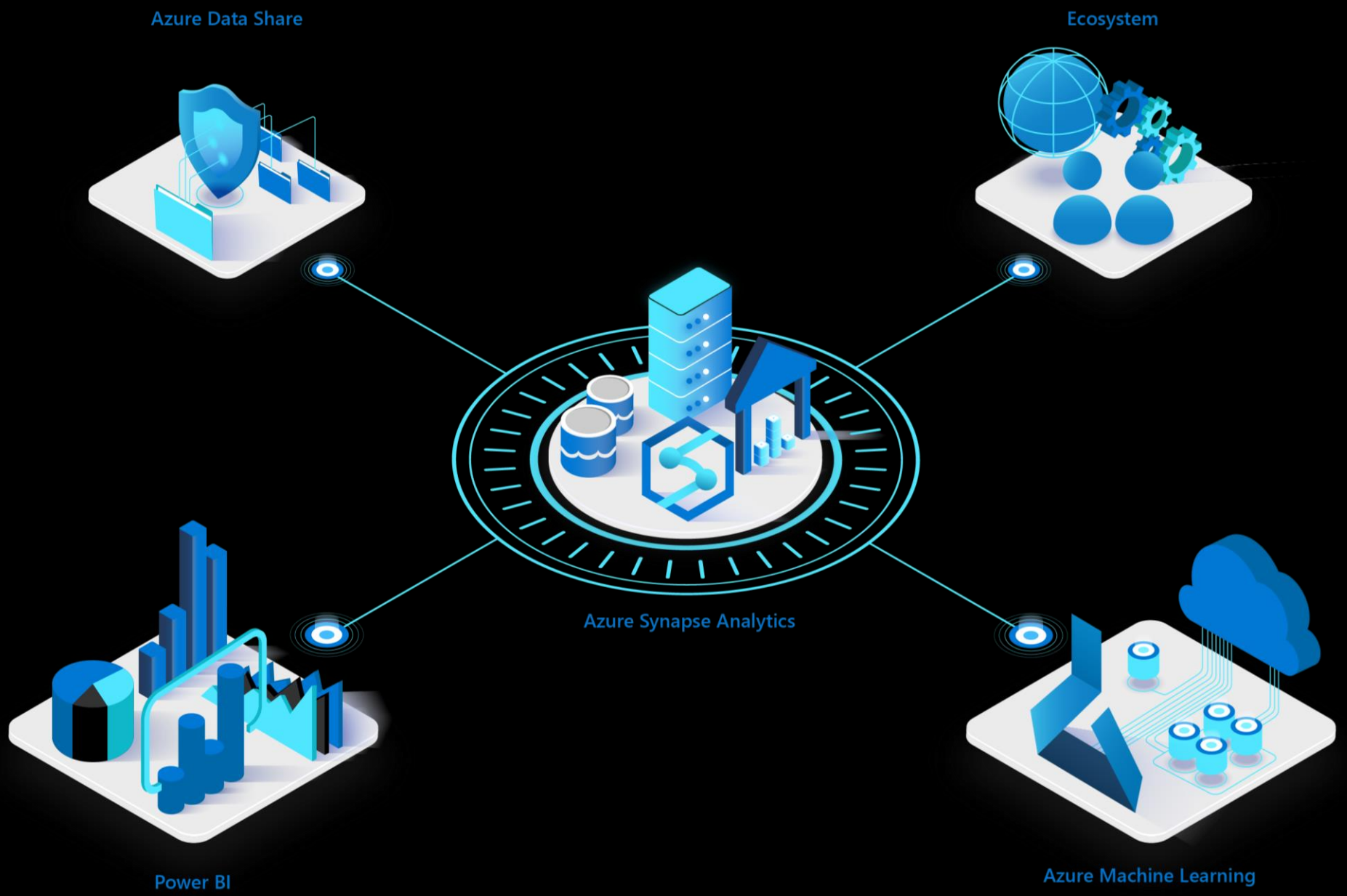


# Introducing Azure Synapse Analytics

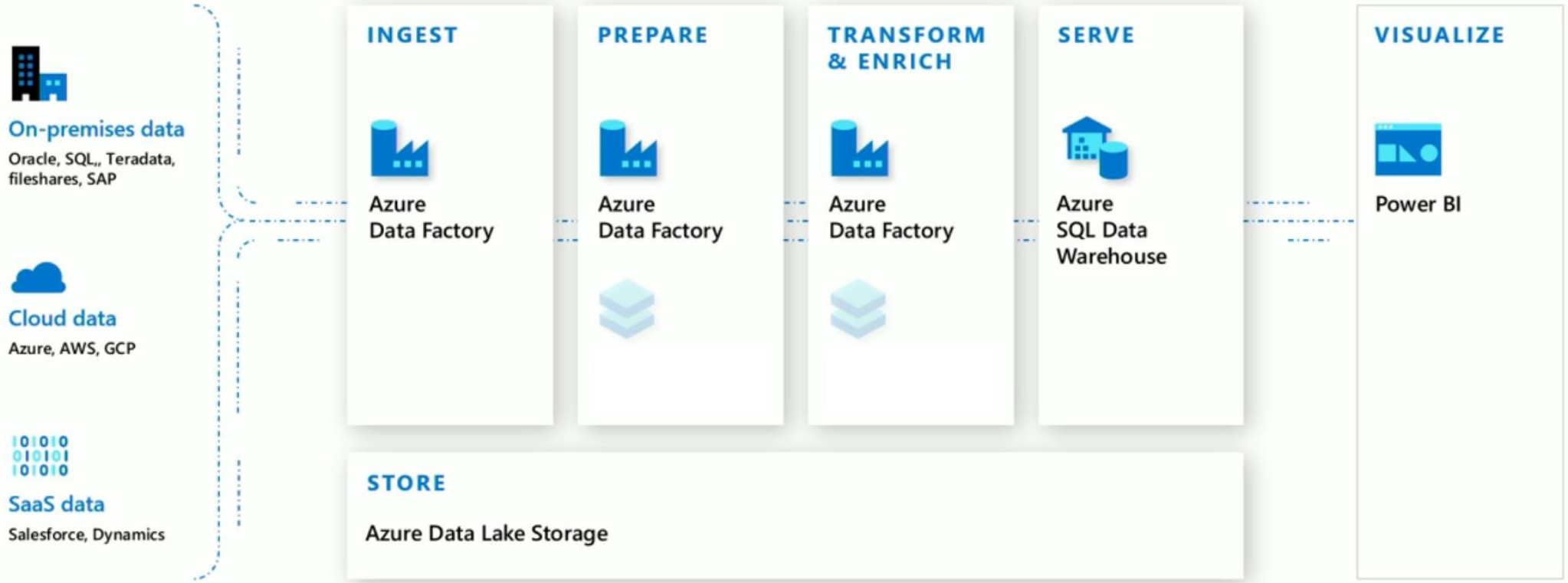
A **limitless** analytics service with **unmatched time to insight**, that delivers insights from all your data, **across data warehouses and big data** analytics systems, **with blazing speed**

Simply put, **Azure Synapse is Azure SQL Data Warehouse evolved**

We have taken the same industry leading data warehouse and elevated it to a whole new level of performance and capabilities


Azure Data Share
Ecosystem
Azure Synapse Analytics
Power BI
Azure Machine Learning

# Azure Modern Data Warehouse 1.0

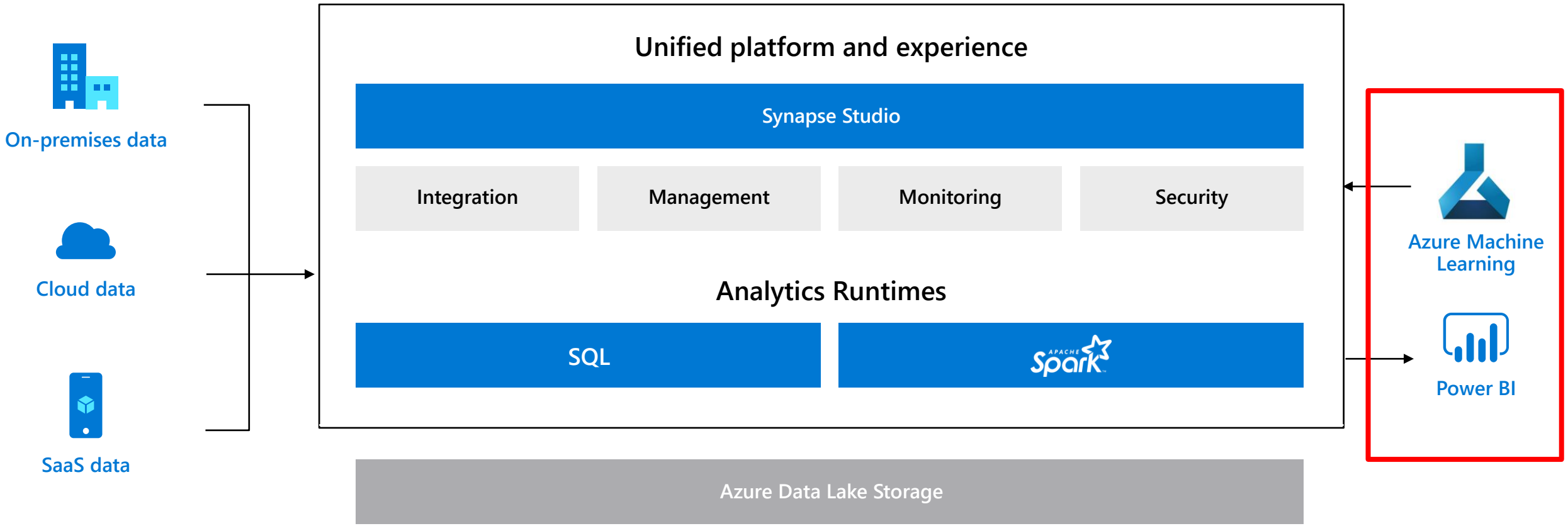


# Azure Synapse Unified Analytics (Modern Datawarehouse 2.0)

# Azure Synapse Analytics

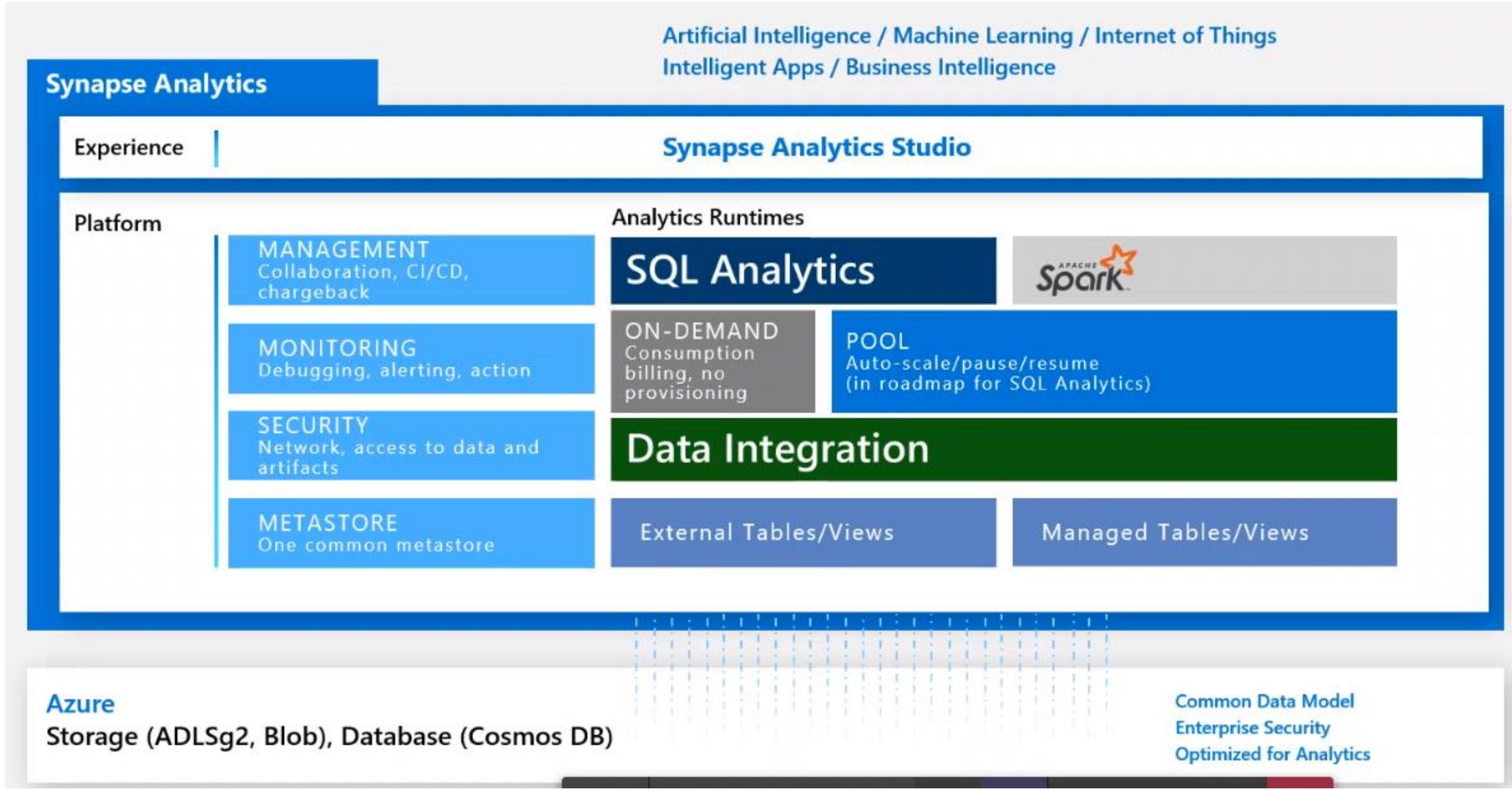
Data Lakehouse

## Limitless analytics service with unmatched time to insight



# Azure Synapse Analytics

Integrated data platform for BI, AI and continuous intelligence



**Synapse Analytics (GA)**

Add new capabilities
to the GA service

**New GA features**
- Resultset caching
- Materialized Views
- Ordered columnstore
- JSON support
- Dynamic Data Masking
- SSDT support
- Read committed snapshot isolation
- Private LINK support

**Preview features**
- Workload Isolation
- Simple ingestion with COPY
- Share DW data with Azure Data Share

**Private preview features**
- Streaming ingestion & analytics in DW
- Native Prediction/Scoring
- Fast query over Parquet files
- FROM clause with joins

**Synapse Analytics (GA)**
(formerly SQL DW)
*"v1"*

**Synapse Analytics (PREVIEW)**
*"v2"*

**Preview features**
- Synapse Studio
- Collaborative *workspaces*
- Distributed T-SQL Query service
- SQL Script editor
- Unified security model
- Notebooks
- Apache Spark
- On-demand T-SQL
- Code-free data flows
- Orchestration Pipelines
- Data movement
- Integrated Power BI

Far future:
Gen3
*"v3"*


SQL ANALYTICS

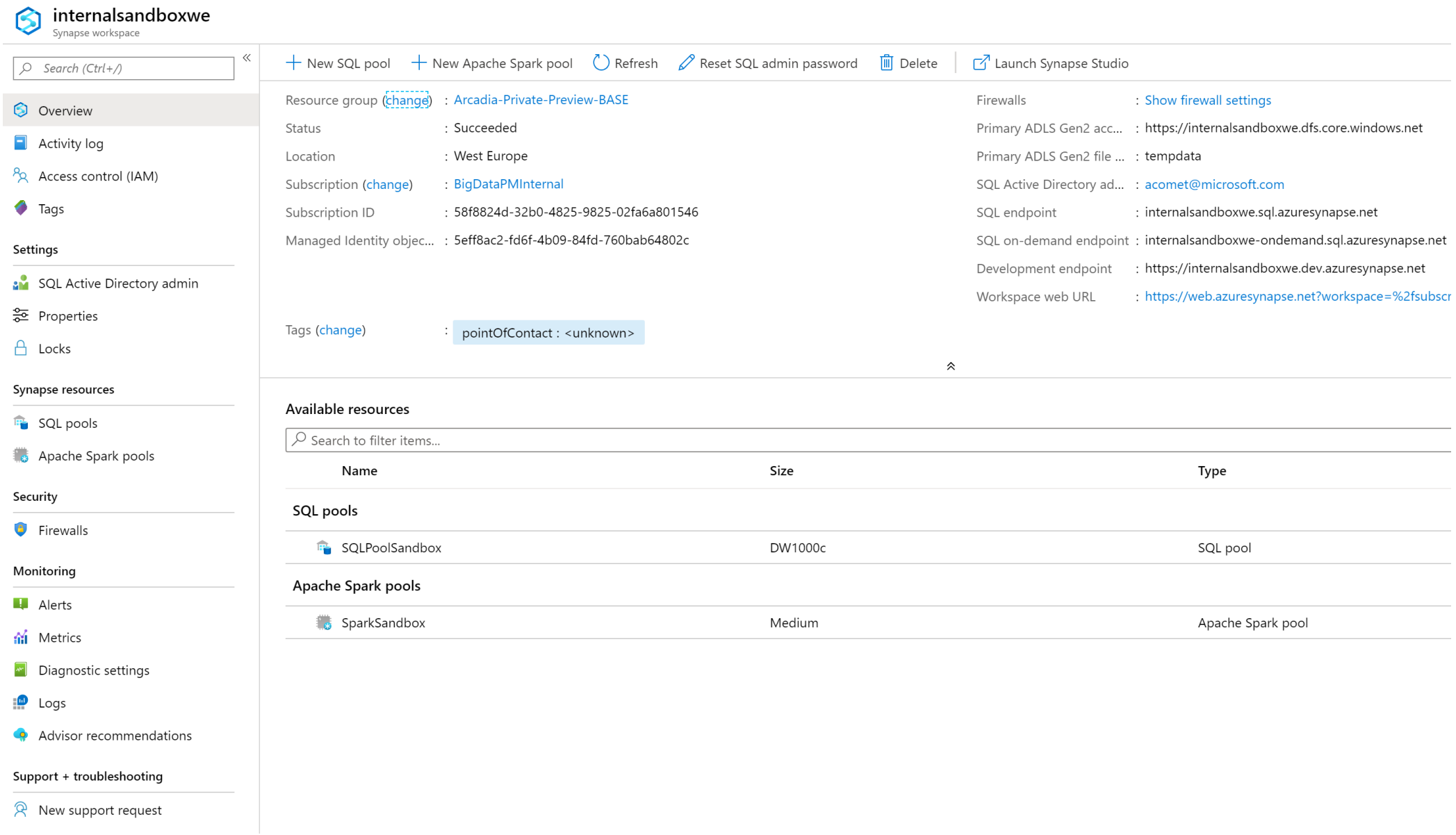
APACHE SPARK

STUDIO

DATA INTEGRATION

*World class unified analytics service*

# Synapse workspace



# SQL Analytics

*World class unified analytics service*

# SQL pools

New Refresh

Search to filter items...

| Name | Type | Status | Size |
|---|---|---|---|
| SQL on-demand | SQL Analytics on-demand | N/A | N/A |
| SQLPoolSandbox | SQL Analytics pool | Online | DW1000c |
| SQLSandboxLarge | SQL Analytics pool | Online | DW2000c |
| SQLSandboxSmall | SQL Analytics pool | Online | DW100c |

## Create SQL pool

Synapse

Basics * | Additional settings * | Tags | Review + create

Create a SQL pool with your Preferred Configuration. Complete the basics tab then go to Review + create provision with smart defaults. Learn more

### SQL pool Details

Name your SQL pool and choose its initial settings.

SQL pool Name * Enter SQL pool Name

Performance level DW1000c

Basics * | Additional settings * | Tags | Review + create

Customize additional configuration parameters including collation & sample data.

### Data source

Start with a blank SQL pool, restore from a backup or select sample data to populate your new SQL pool.

Use existing data * None Backup

### SQL pool collation

Collation defines the rules that sort and compare data, and cannot be changed after SQL pool creation. The default collation is SQL_Latin1_General_CP1_CI_AS. Learn more

Collation * SQL_Latin1_General_CP1_CI_AS

SQL Analytics pool = SQL Data Warehouse

# Synapse SQL on-demand scenarios

| | |
|---|---|
| **Discovery and exploration** | What's in this file? How many rows are there? What's the max value?<br><br>**SQL On-demand reduces data lake exploration to the right-click!** |
| **Data transformation** | How to convert CSVs to Parquet quickly? How to transform the raw data?<br><br>**Use the full power of T-SQL to transform the data in the data lake** |

# SQL On-Demand Pools

*Single Pane of Glass Data Discovery*

## Overview

An interactive query service that provides T-SQL queries over high scale data in Azure Storage.

## Benefits

Serverless

No infrastructure

Pay only for query execution

No ETL

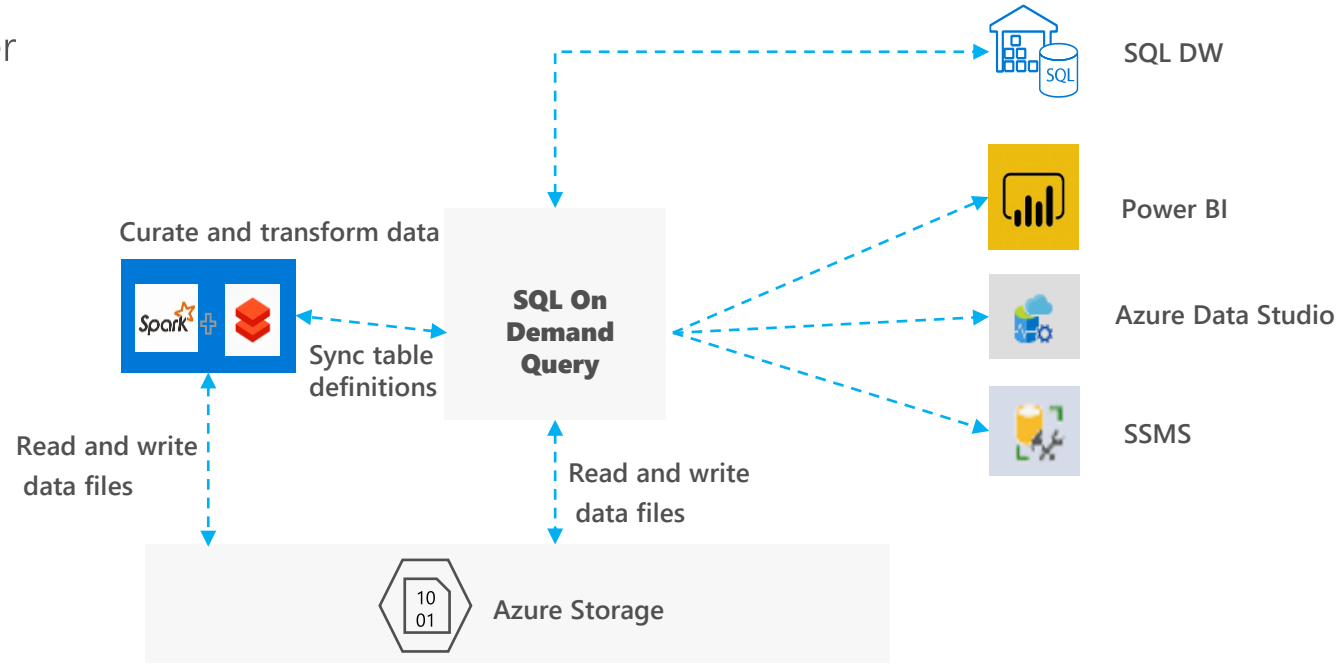
Offers security

Data integration with Databricks, HDInsight

T-SQL syntax to query data

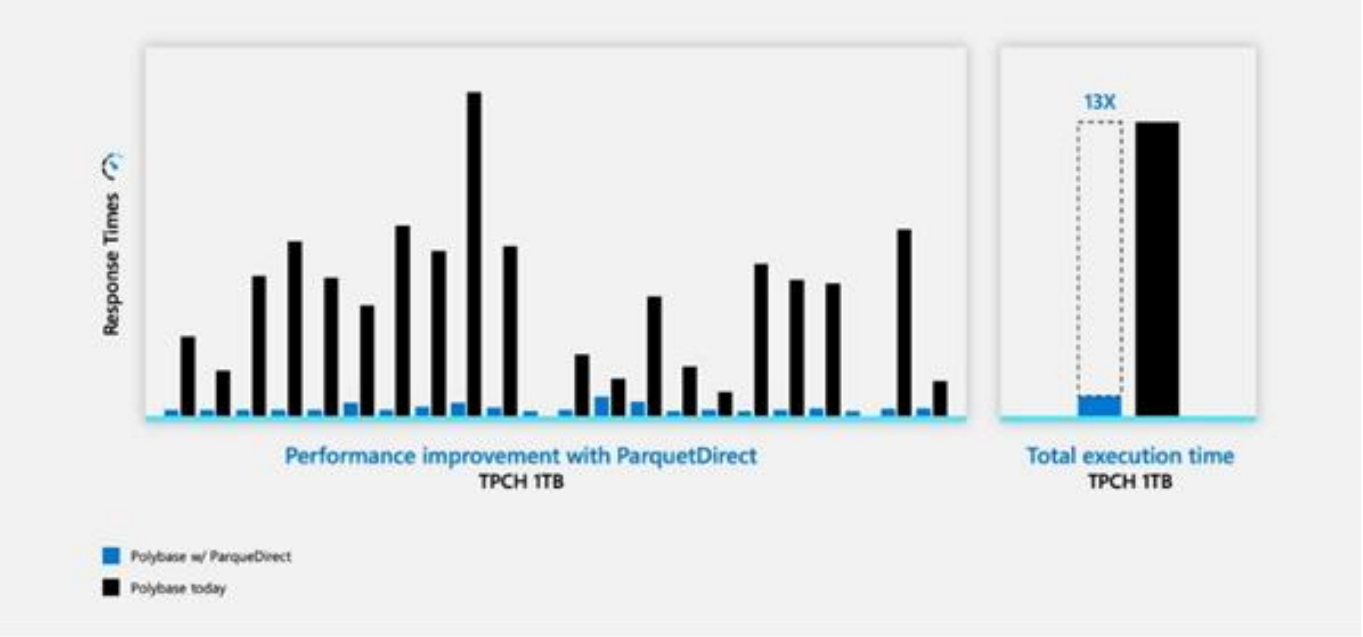
Supports data in various formats (Parquet, CSV, JSON)

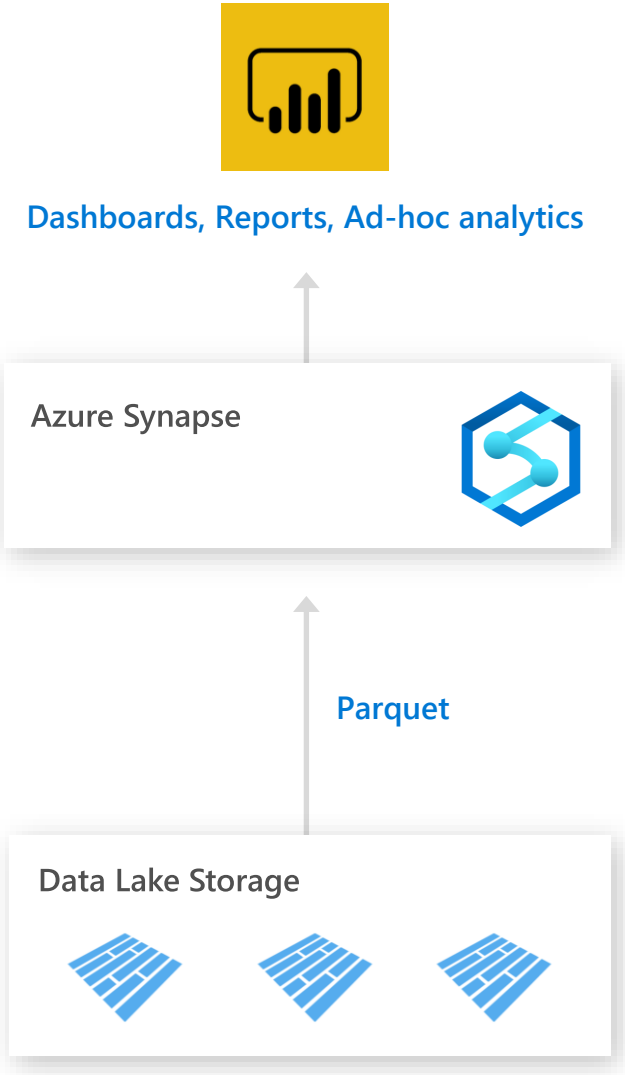
Support for BI ecosystem



# Data Flexibility – Parquet Direct

## Overview



Microsoft

# Azure Synapse Analytics

## Spark

*World class unified analytics service*

# Apache Spark pools

New Refresh

Search to filter items...

| Name | Size |
|---|---|
| SparkSandbox | Medium (8 vCPU / 64 GB) - 3 to 20 nodes |
| SparkSmall | Small (4 vCPU / 32 GB) - 3 to 20 nodes |
| SparkLarge | Large (16 vCPU / 128 GB) - 3 to 80 nodes |

## Create Apache Spark pool

Basics * | Additional settings * | Tags | Summary

Create a Synapse Analytics Apache Spark pool with your preferred configurations. Complete the Basics tab then go to Review + create to provision with smart defaults, or visit each tab to customize.

### Apache Spark pool details

Name your Apache Spark pool and choose its initial settings.

Apache Spark pool name *: Enter Apache Spark pool name

Node size family: MemoryOptimized

Node size *: Medium (8 vCPU / 64 GB)

Autoscale *: Enabled | Disabled

Number of nodes *: 3 — 40

Note: There are no on-demand pools for Spark

Basics * | Additional settings * | Tags | Summary

Customize additional configuration parameters including autoscale and component versions.

**Auto-pause**

Enter required settings for this Apache Spark pool, including setting auto-pause and picking versions.

Auto-pause *: Enabled | Disabled

Number of minutes idle *: 15

**Component versions**

Select the Apache Spark version for your Apache Spark pool.

| | |
|---|---|
| Apache Spark * | 2.4 |
| Python | 3.6.1 |
| Scala | 2.11.12 |
| Java | 1.8.0_222 |
| .NET Core | 3.0 |
| .NET for Apache Spark | 0.6.0 |
| Delta Lake | 0.4.0 |

**Packages**

Upload environment configuration file ("PIP freeze" output).

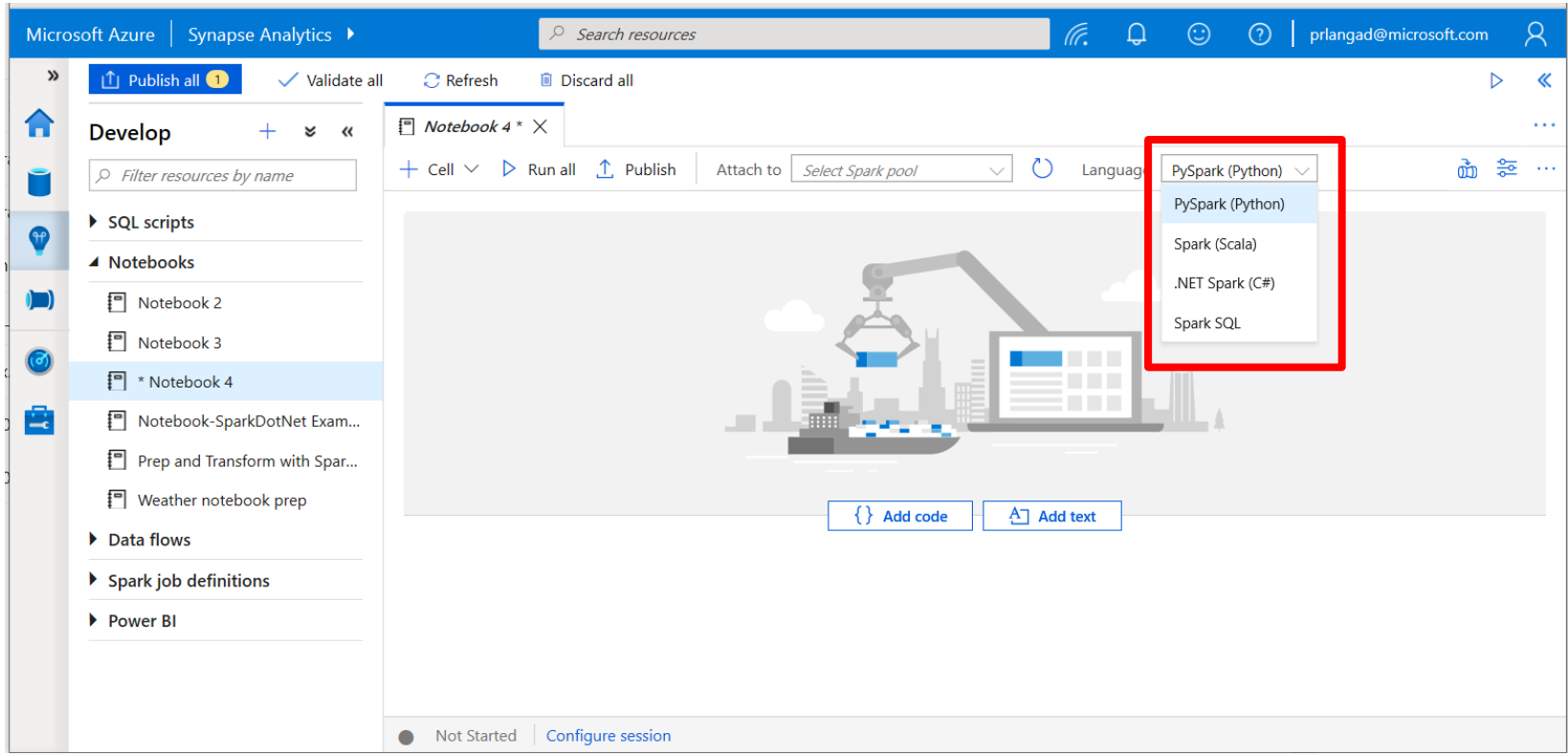
File upload: Select a file

# Languages

## Overview

Supports multiple languages to develop notebook

- PySpark (Python)
- Spark (Scala)
- .NET Spark (C#)
- Spark SQL
- Java
- R (early 2020)




## Benefits

Allows to write multiple languages in one notebook

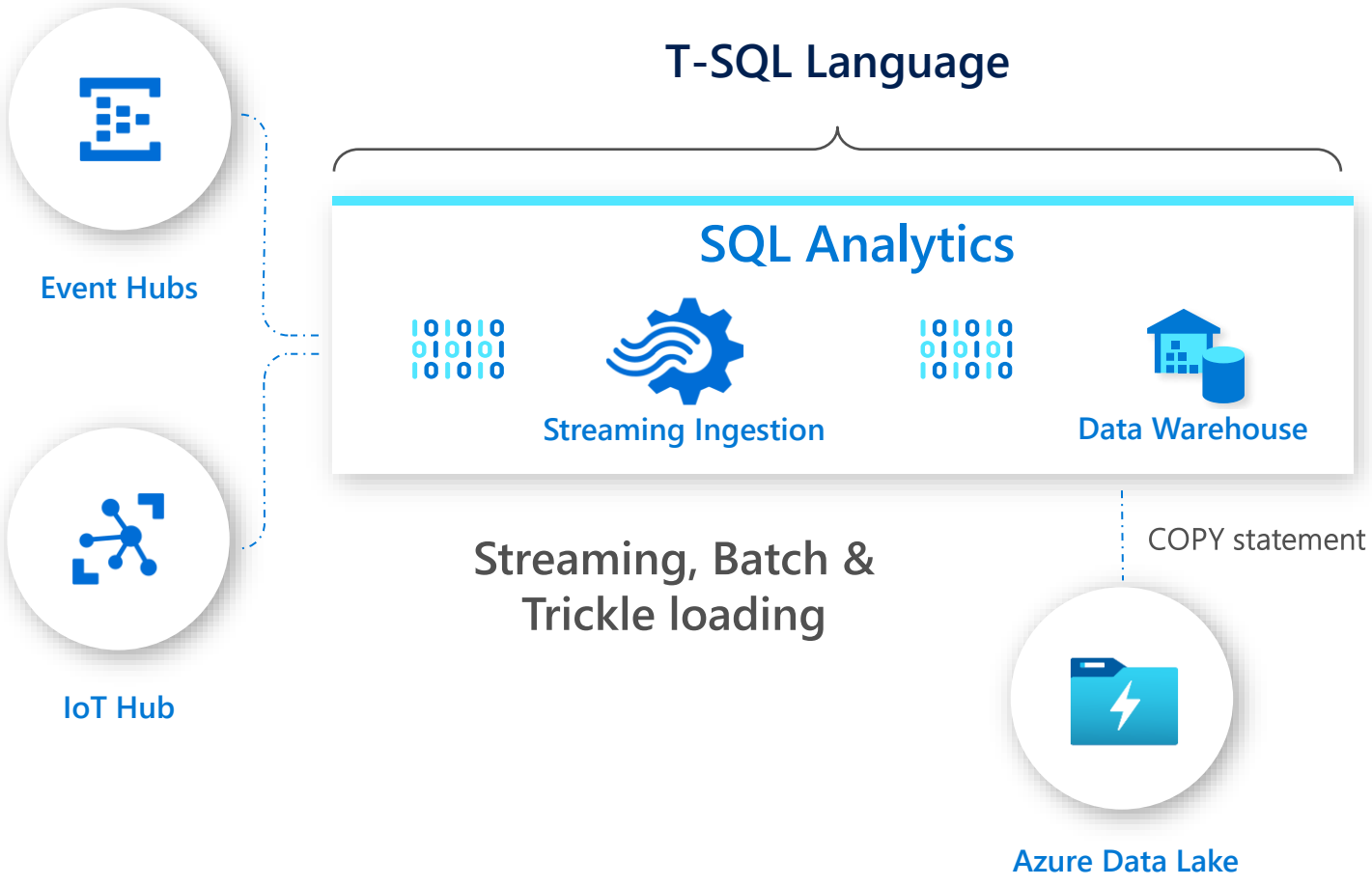
%%<Name of language>

Offers use of temporary tables across languages

# Heterogenous Data Preparation & Ingestion

## COPY statement

- Simplified permissions (no CONTROL required)
- No need for external tables
- Standard CSV support (i.e. custom row terminators, escape delimiters, SQL dates)
- User-driven file selection (wild card support)

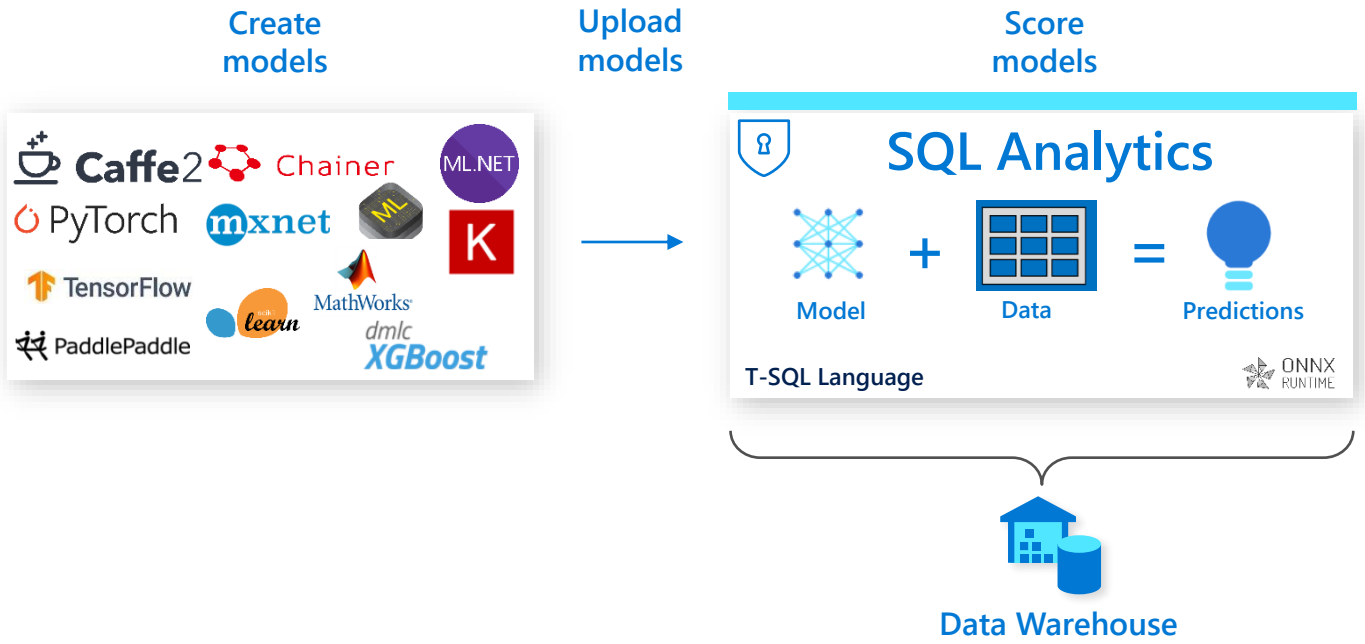



```
--Copy files in parallel directly into data warehouse table
COPY INTO [dbo].[weatherTable]
FROM 'abfss://<storageaccount>.blob.core.windows.net/<filepath>'
WITH (
FILE_FORMAT = 'DELIMITEDTEXT',
SECRET = CredentialObject);
```

# Machine Learning enabled DW

## Native PREDICT-ion

- T-SQL based experience (interactive./batch scoring)
- Interoperability with other models built elsewhere
- Execute scoring where the data lives




```
--T-SQL syntax for scoring data in SQL DW
SELECT d.*, p.Score
FROM PREDICT(MODEL = @onnx_model, DATA = dbo.mytable AS d)
WITH (Score float) AS p;
```

# Azure Machine Learning

***Integrated Spark and ML for ELT and machine learning***

## Overview

Data Scientists can use Azure ML notebooks to do (distributed) data preparation on Synapse Spark compute.

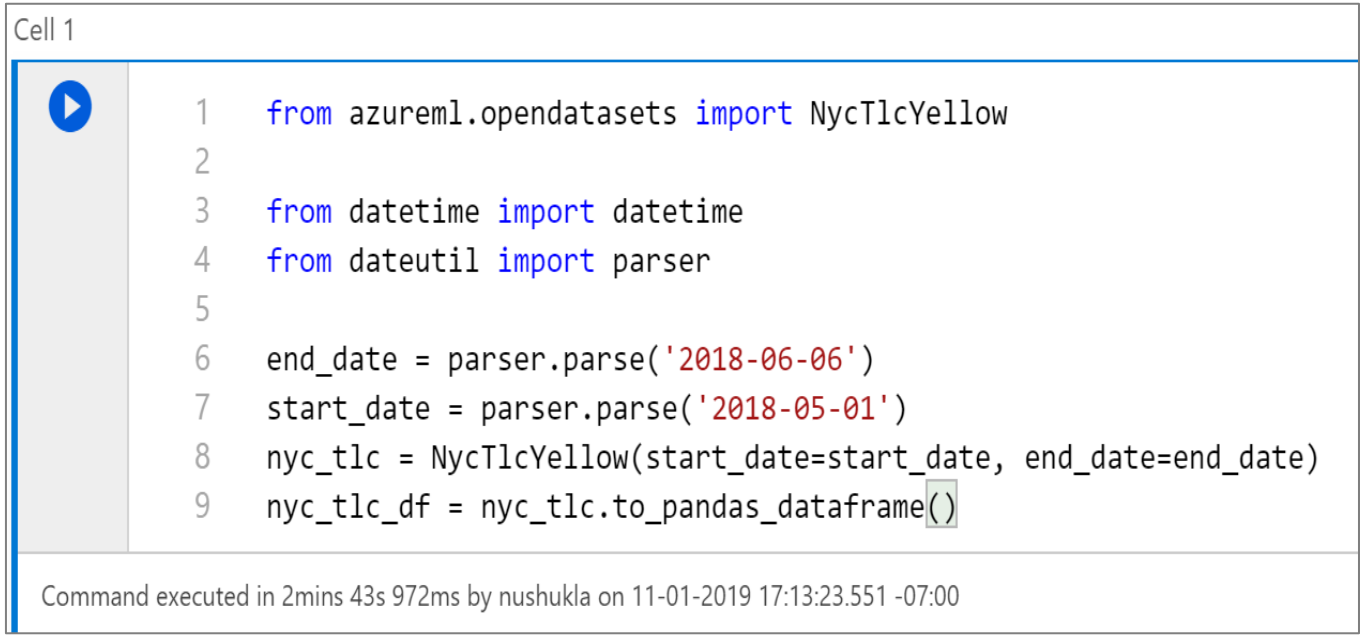
## Benefits

Connect to your existing Azure ML workspace and project

Use the AutoML Classifier for classification or regression problem

Train the model

Access open datasets



# Azure Data Share

## Enterprise data sharing

- Share from DW to DW/DB/other systems
- Choose data format to receive data in (CSV, Parquet)
- One to many data sharing
- Share a single or multiple datasets

### Any Azure Data Sources

Share data from any Azure regions and data stores

### Single Pane of Glass

Manage and monitor data sharing with multiple organizations

### Rich Analytics Tools

Use Azure analytics tools to prepare data and derive insights

| Feature | Azure Data Share |
|---|---|
| **Multiple Data Store Support**<br>Sharing from Azure Data Lake, Azure Storage, Azure SQL Data Warehouse, Azure SQL DB | Yes |
| **Heterogenous Data Sharing**<br>Flexible sharing from/to heterogenous data stores | Yes |
| **Single pane of glass**<br>Centrally managed data sharing experience | Yes |
| **Governed data sharing**<br>Customer can specify terms of use | Yes |
| **Snapshot based sharing**<br>Perform analytics on data for unrestricted computation & no compromise on performance | Yes |

### Governance

Control data access governed by enterprise policies

### Monetization

Charge for data or cost of data curation and access

# Power BI

## Overview

Power BI is a business analytics service that delivers insights to enable fast, informed decisions

## Benefits

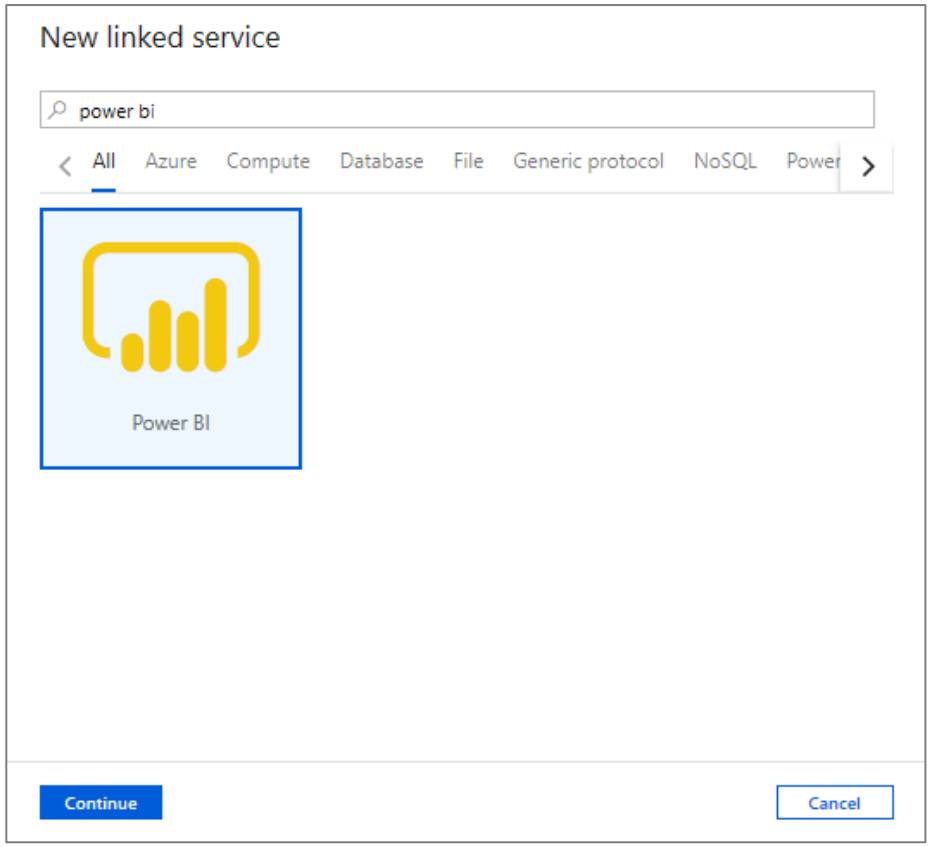
Create Power BI reports in the workspace

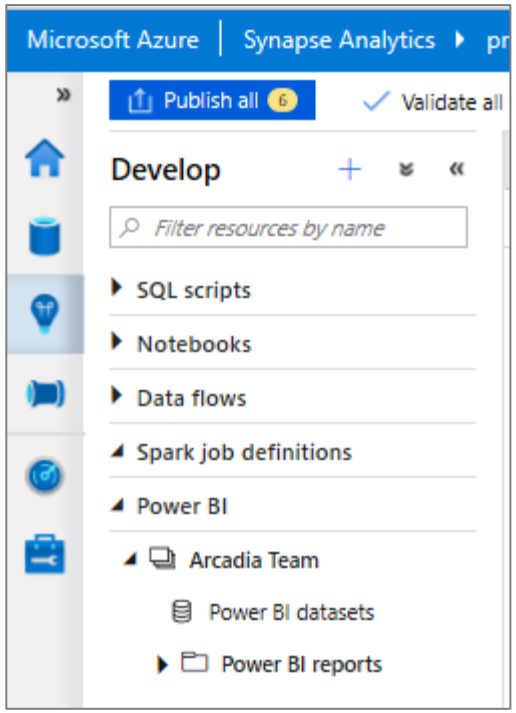
Have access to published reports in workspace

Update reports real time from Synapse workspace to get it reflected on Power BI service

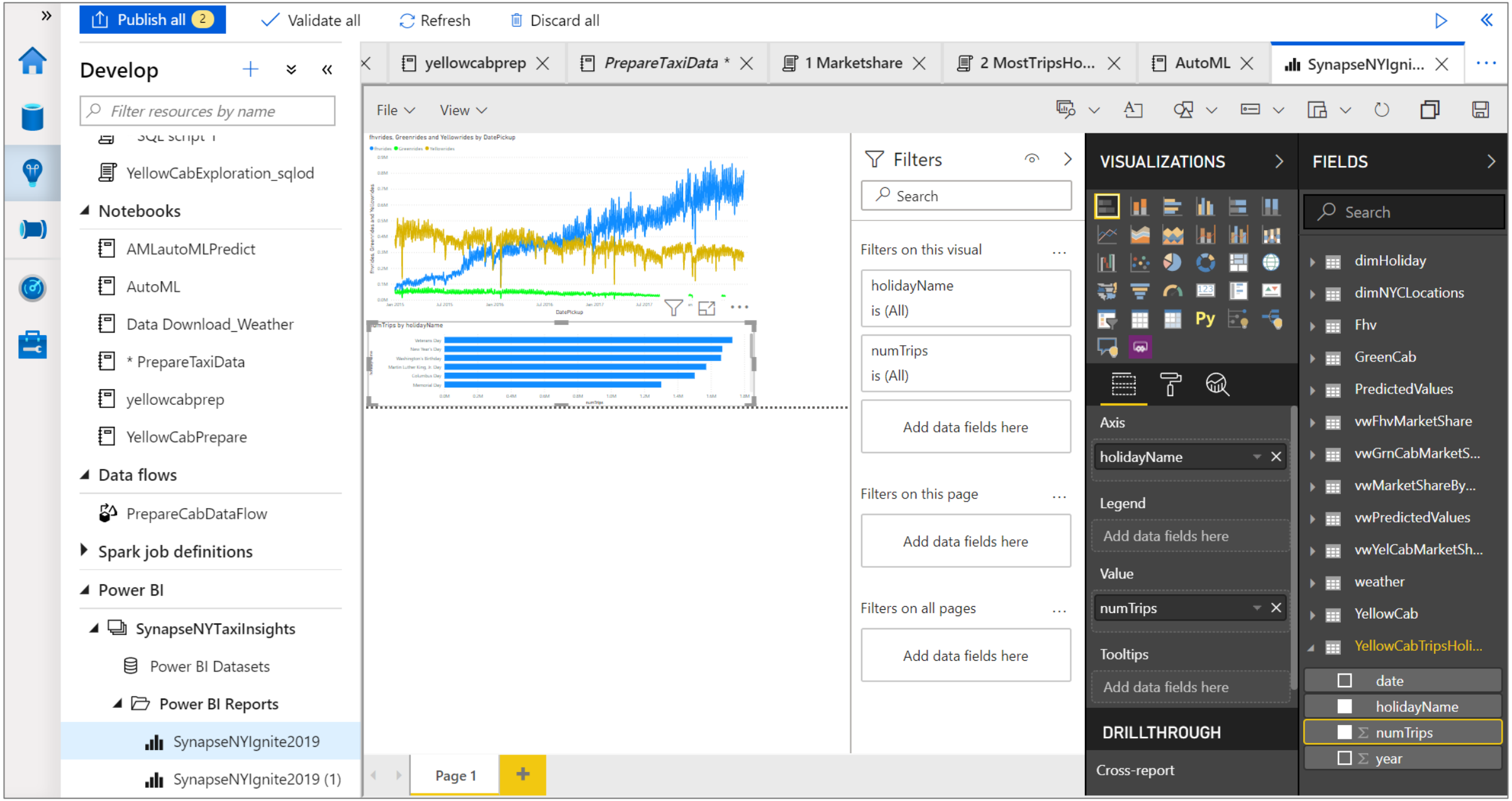
Visually explore and analyze data

***Integrated Power BI Developer Experience***



# *Integrated Power BI Developer Experience*



# Create an engine for business-changing insights with seamless ecosystem integration

Data integration | Data warehousing | Big data processing | Analytics | Build & train AI models | Insights apps | Monitor & manage assets | Service assets | Visualize & interact with assets

Azure Synapse Analytics

Power BI

Azure Machine Learning

D365 Customer Insights

Azure IoT Central

D365 Field Service

Layout, Simulation Remote Assistance

## Azure Data Lake Storage + Common Data Model

**Financial data**

SQL Server

**Enriched data**

Office 365

**Sales data**

Office 365

**Customer profile data**

**Asset & product data**

# Create an engine for business-changing insights with seamless ecosystem integration

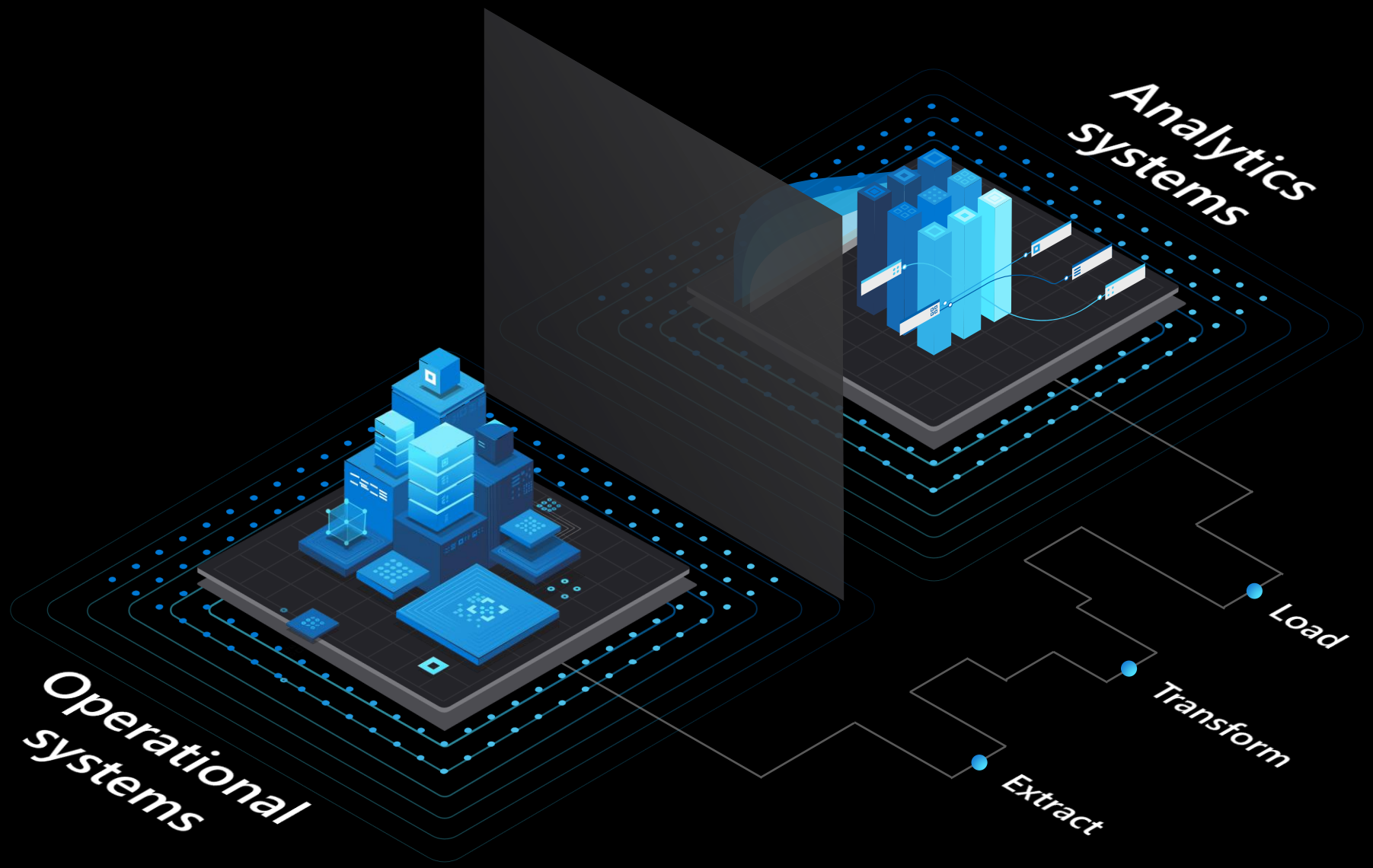
Data integration

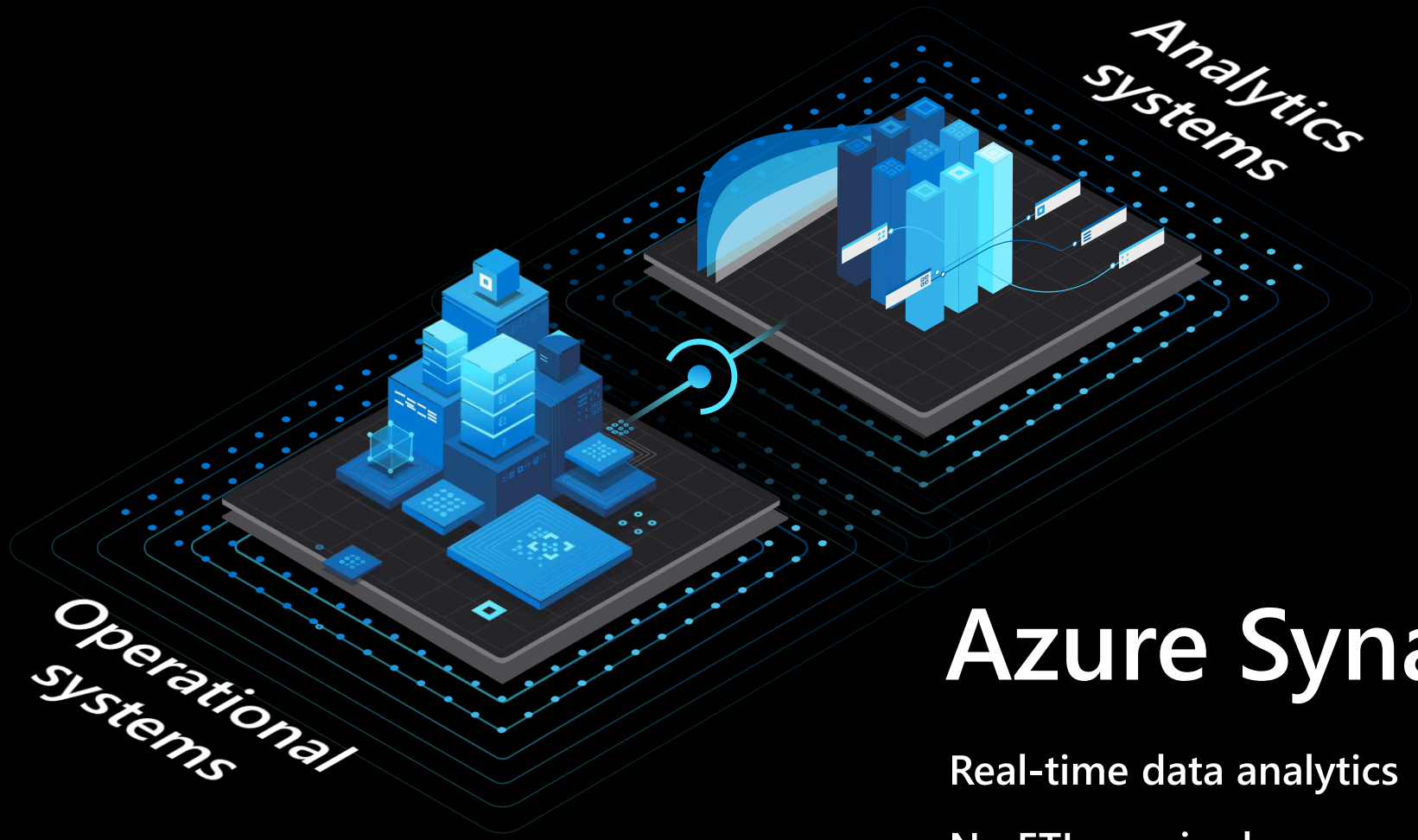
Data warehousing

Big data processing

Azure Synapse Analytics

Azure Data Lake Storage + Common Data Model


Analytics systems
Load
Transform
Extract
Operational systems

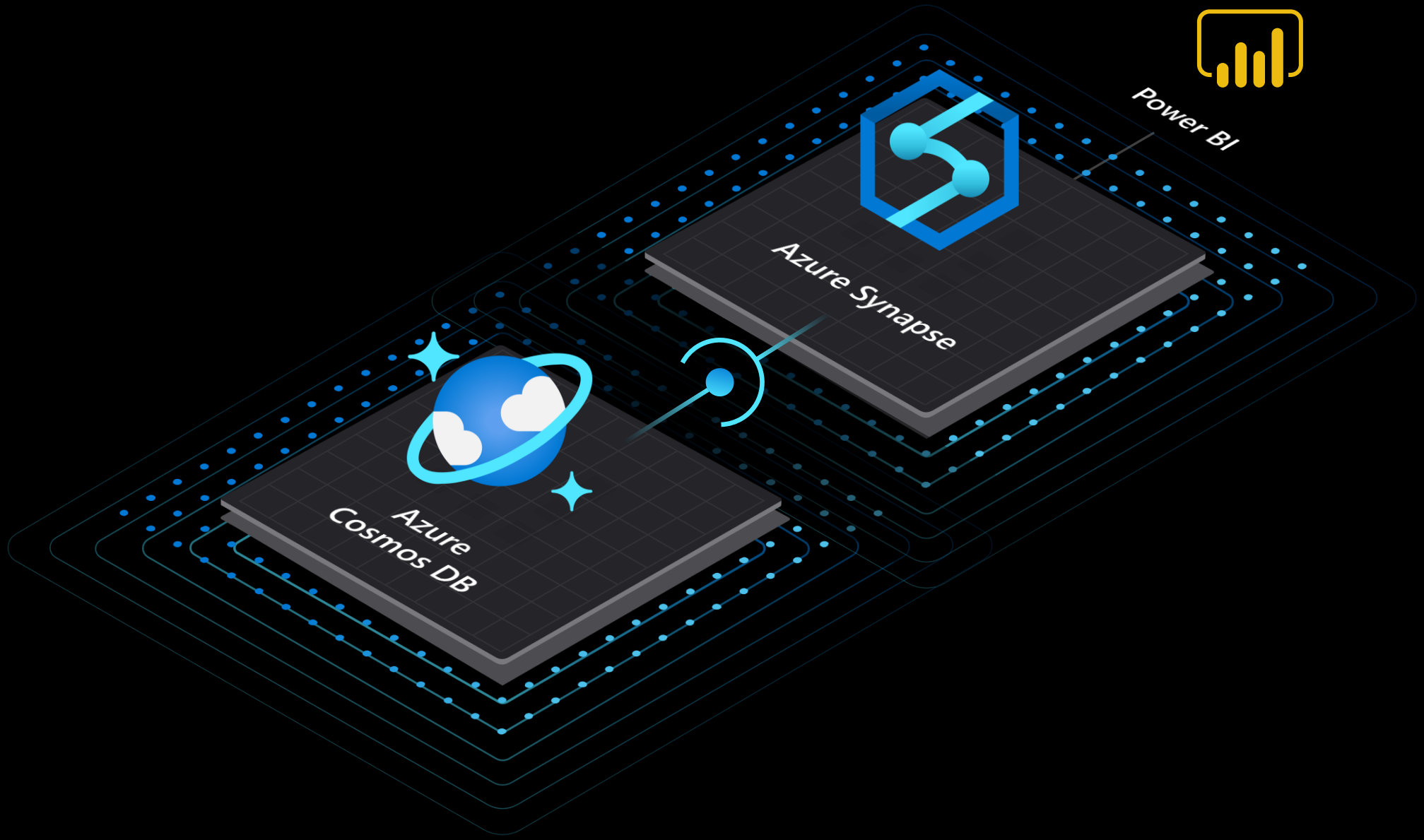



# Azure Synapse Link

Real-time data analytics

No ETL required

No performance impact on transactions


Power BI
Azure Synapse
Azure
Cosmos DB

Microsoft

# Azure Databricks

# Why Spark?

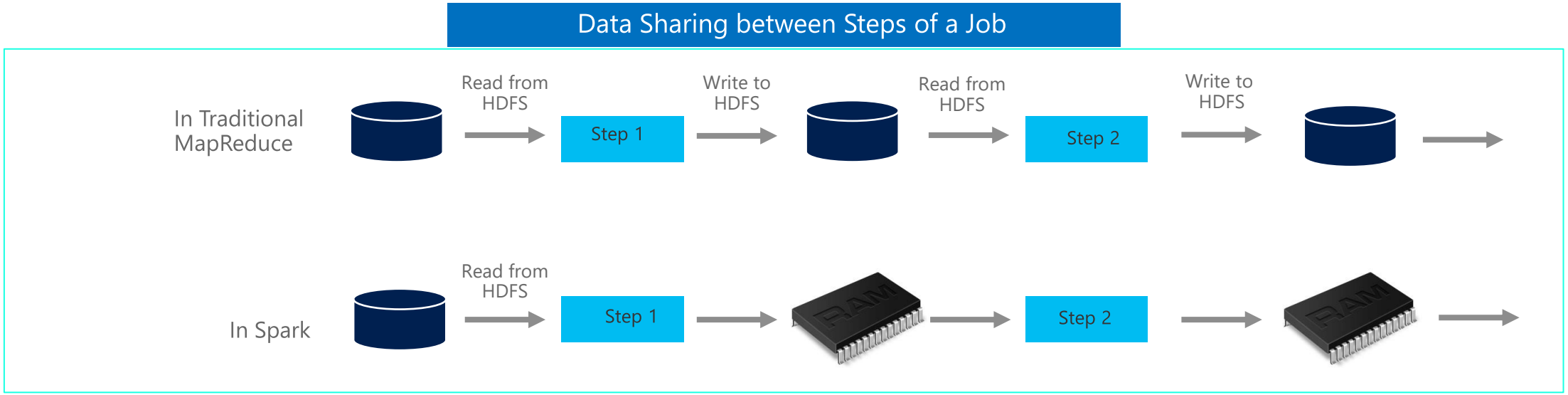

- Open-source data processing engine built around **speed, ease of use, and sophisticated analytics**
- In memory engine that is up to **100 times faster than Hadoop**
- **Largest open-source data project** with 1000+ contributors
- **Highly extensible** with support for Scala, Java and Python alongside Spark SQL, GraphX, Streaming and Machine Learning Library (MLlib)

# Why Databricks?

- Databricks is the **premium version of Spark** available in the market
- **Spark founders created Databricks**
- Spark is the dominant workload in Hadoop
- **Databricks commits 75% of the code to Open Source Spark**

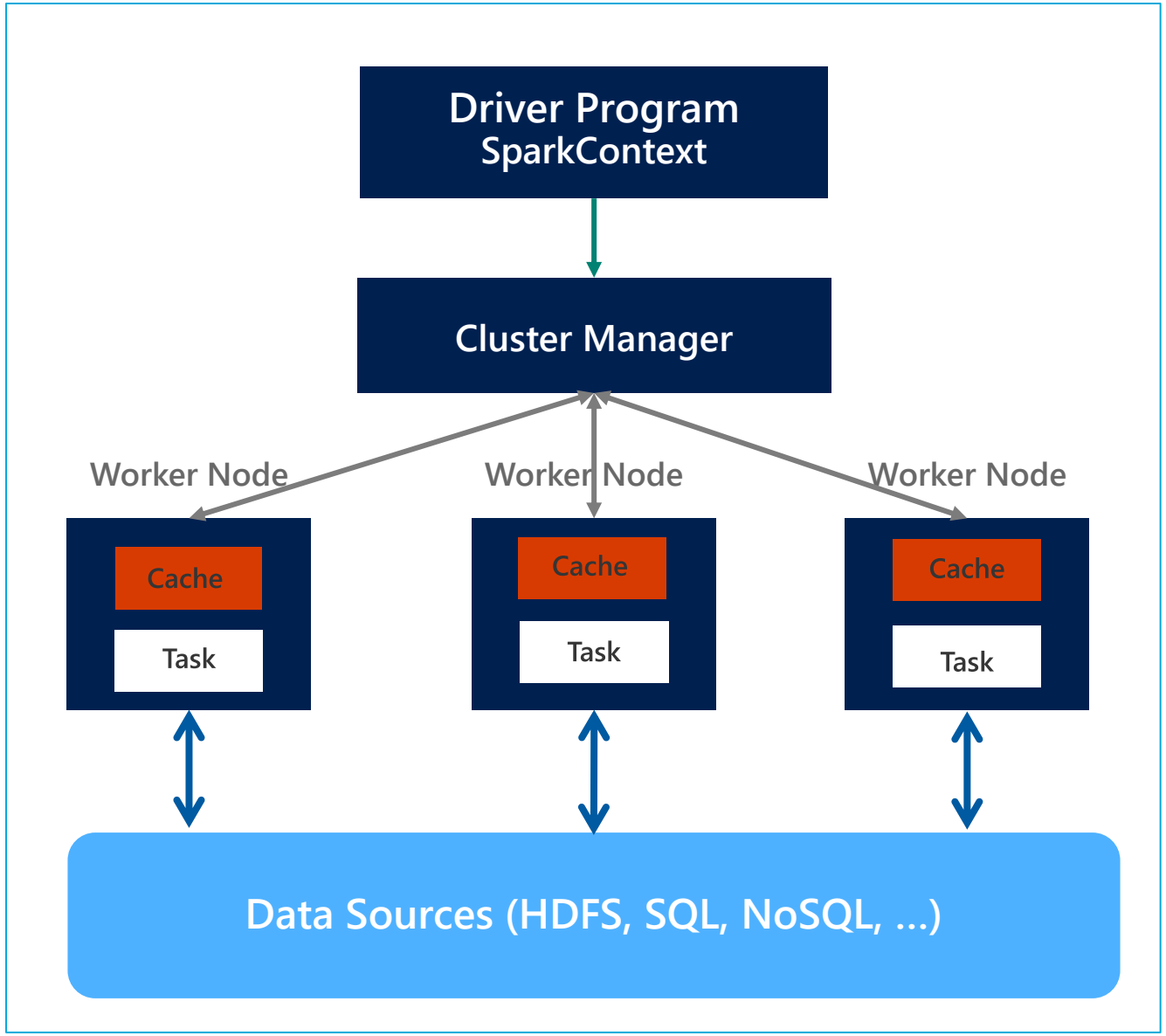
# WHAT MAKES SPARK FAST?

- **In-memory cluster computing:** Spark provides primitives for *in-memory* cluster computing. A Spark job can *load and cache* data into memory and query it repeatedly (iteratively) much quicker than disk-based systems.
- **Scala Integration:** Spark integrates into the Scala programming language, letting you manipulate distributed datasets like local collections. No need to structure everything as map and reduce operations
- **Faster Data-sharing:** Data-sharing between operations is faster as data is in-memory:
  - In (traditional) Hadoop data is shared through HDFS which is expensive. HDFS maintains three replicas.
  - Spark stores data in-memory *without any replication*.



# GENERAL SPARK CLUSTER ARCHITECTURE

- 'Driver' runs the user's 'main' function and executes the various parallel operations on the worker nodes.
- The results of the operations are collected by the driver
- The worker nodes read and write data from/to Data Sources including HDFS.
- Worker node also cache transformed data in memory as RDDs (Resilient Data Sets).
- Worker nodes and the Driver Node execute as VMs in public clouds (AWS, Google and Azure).



# Spark 3.0 Announcement


## Performance

- Adaptive Query Execution
- Dynamic Partition Pruning
- Query Compilation Speedup
- Join Hints

## Built-in Data Sources

- Parquet/ORC Nested Column Pruning
- CSV Filter Pushdown
- Parquet: Nested Column Filter Pushdown
- New Binary Data Source

## Richer APIs

- Accelerator-aware Scheduler
- Built-in Functions
- pandas UDF Enhancements
- DELETE/UPDATE/ MERGE in Catalyst

## SQL Compatibility

- Overflow Checking
- ANSI Store Assignment
- Proleptic Gregorian Calendar
- Reserved Keywords

## Extensibility and Ecosystem

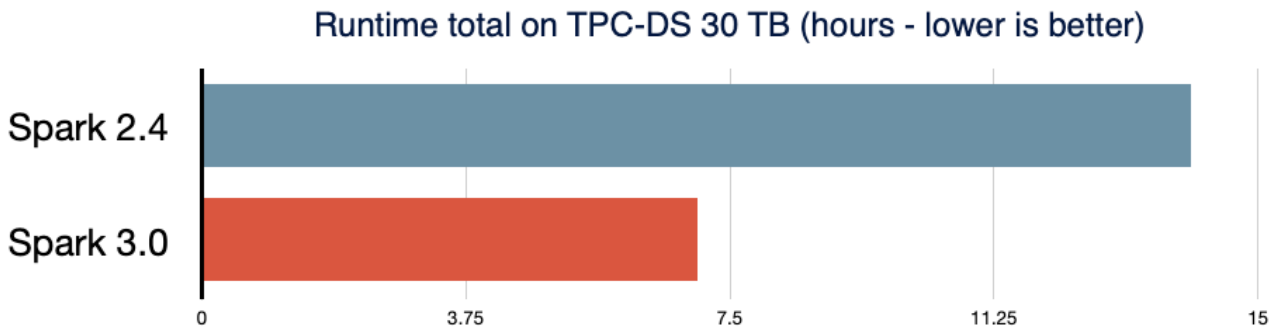
- Data Source V2 API + Catalog Support
- Hadoop 3 Support
- Hive 3.x Metastore Hive 2.3 Execution
- Java 11 Support

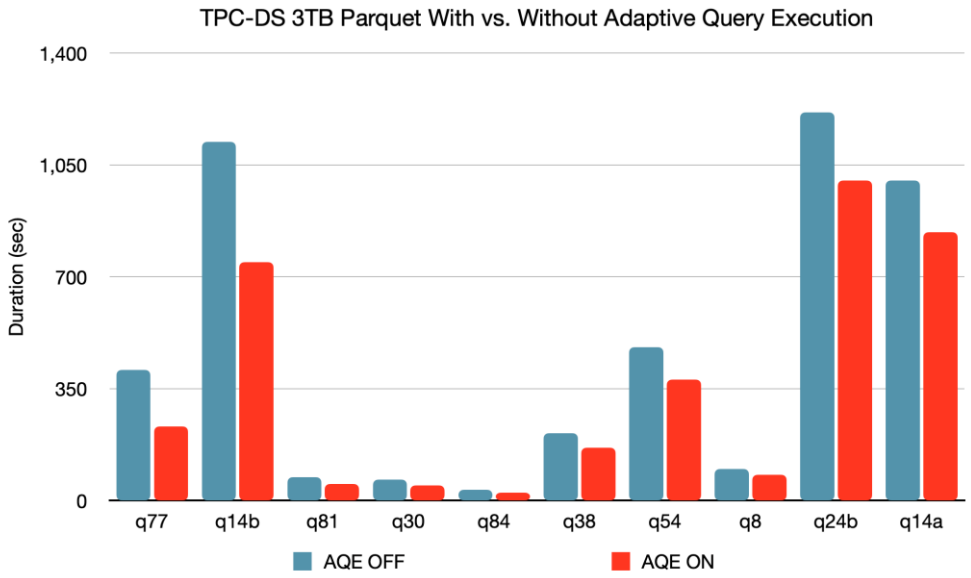
## Monitoring and Debuggability

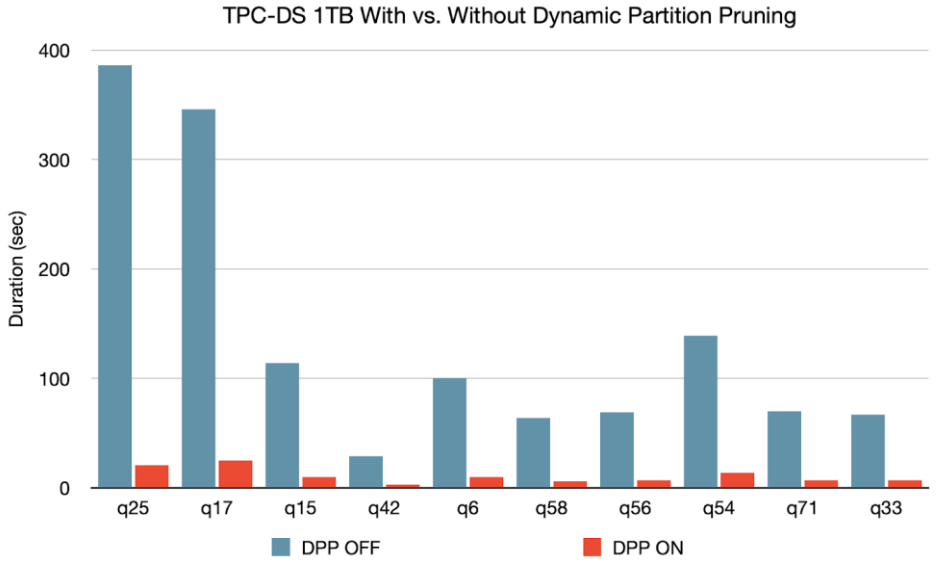
- Structured Streaming UI
- DDL/DML Enhancements
- Observable Metrics
- Advanced Instrumentation

# Spark 3.0 Performance

2x performance improvement on TPC-DS over Spark 2.4, enabled by adaptive query execution, dynamic partition pruning and other optimizations

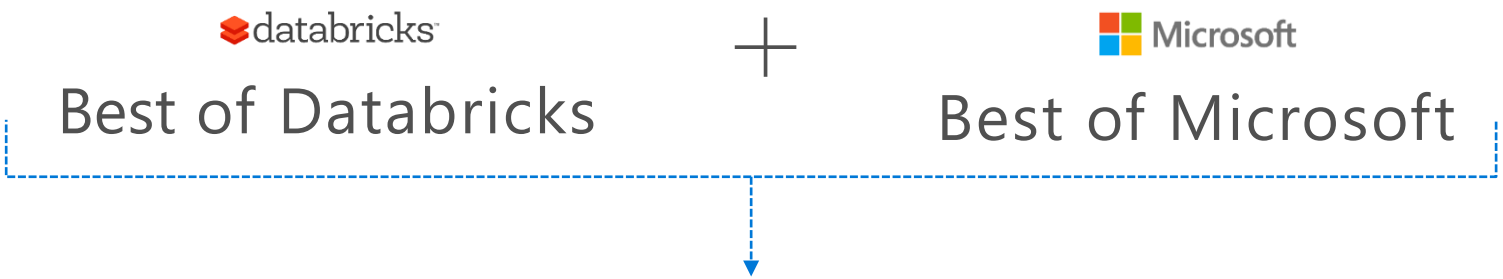


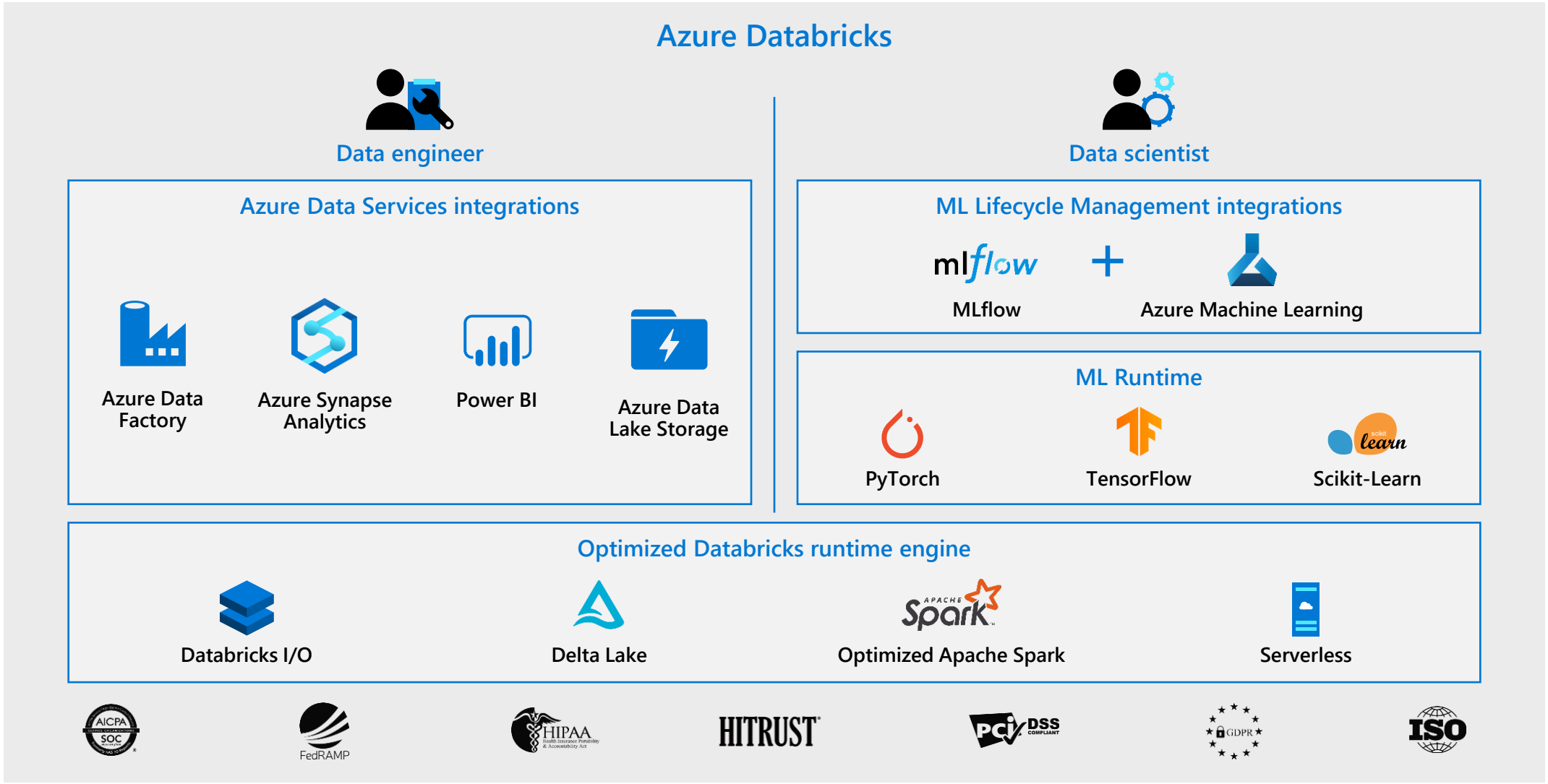
# What is Azure Databricks?

A fast, easy and collaborative Apache® Spark™ based analytics platform optimized for Azure

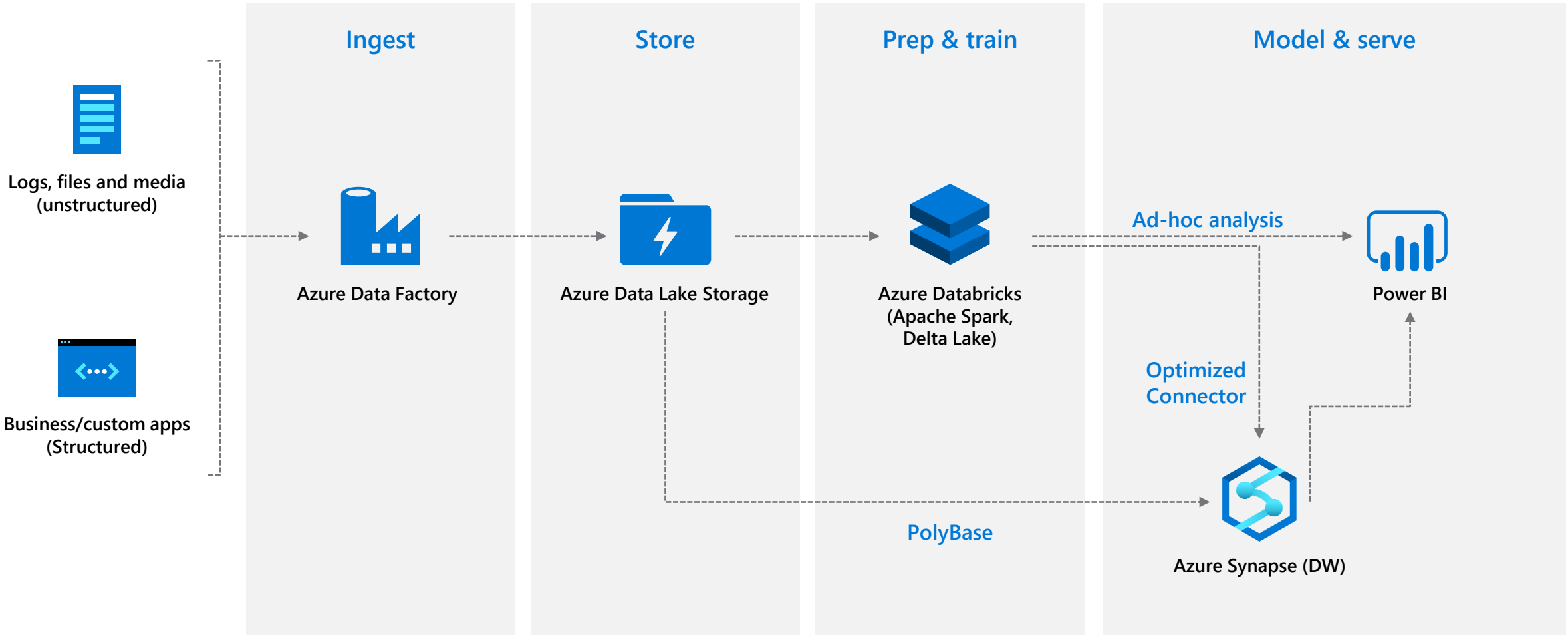



- Designed in collaboration with the founders of Apache Spark
- One-click set up; streamlined workflows
- Interactive workspace that enables collaboration between data scientists, data engineers, and business analysts.
- Native integration with Azure services (Power BI, SQL DW, Cosmos DB, Blob Storage)
- Enterprise-grade Azure security (Active Directory integration, compliance, enterprise-grade SLAs)

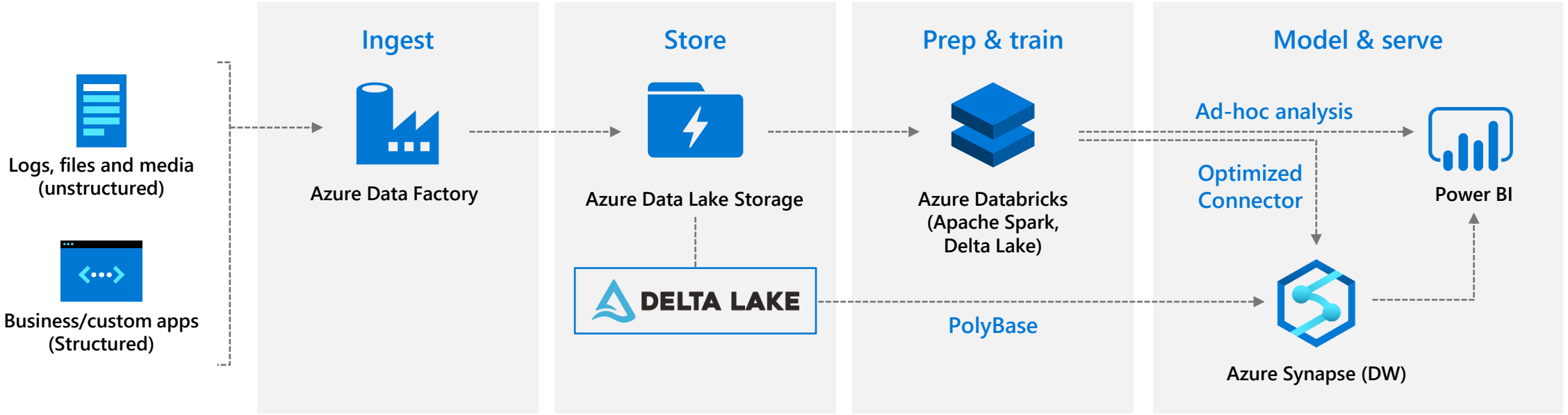
# Azure Databricks



# The modern data warehouse



# The modern data warehouse




**ACID Transaction Guarantees**
Atomic, Consistent, Isolated, Durable

**Versioned parquet files**
Delta transaction log keeps track of all operations

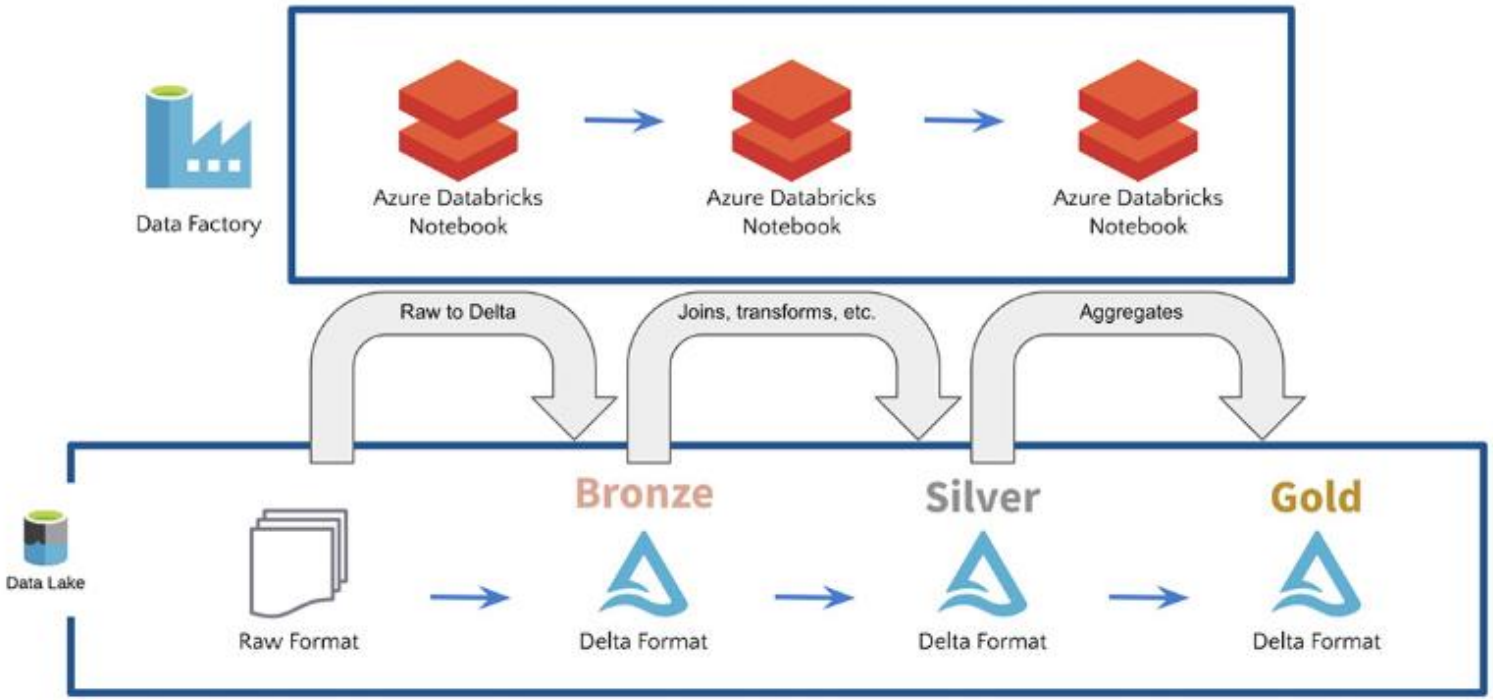
**Efficient Upserts**
MERGE, DELETE, UPDATE

**Time Travel**
Audit history, pipeline debugging, data reproducibility

**Small file compaction w/ no interrupt to availability**
OPTIMIZE and VACUUM

**Z-Order partitioning w/ up to 100x perf**
New multidimensional partitioning enables data skipping

# Azure Delta Lake




https://databricks.com/blog/2020/03/06/connect-90-data-sources-to-your-data-lake-with-azure-databricks-and-azure-data-factory.html

DELTA LAKE

# Bringing Data Reliability & Performance to Data Lakes

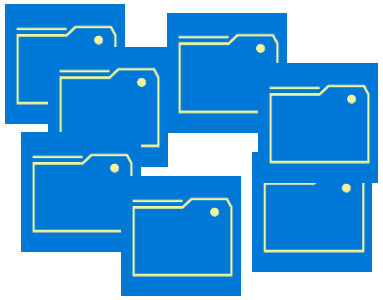
# Data reliability challenges with data lakes

**Failed production jobs** leave data in corrupt state requiring tedious recovery

**Lack of schema enforcement** creates inconsistent and low quality data

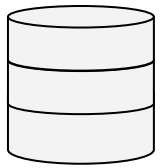
**Lack of consistency** makes it almost impossible to mix appends and reads, batch and streaming

# Performance challenges with data lakes



**Too many small or very big files** - more time opening & closing files rather than reading contents (worse with streaming).

**Partitioning aka "poor man's indexing"-** breaks down if you picked the wrong fields or when data has many dimensions, high cardinality columns.



**No caching** - cloud storage throughput is low

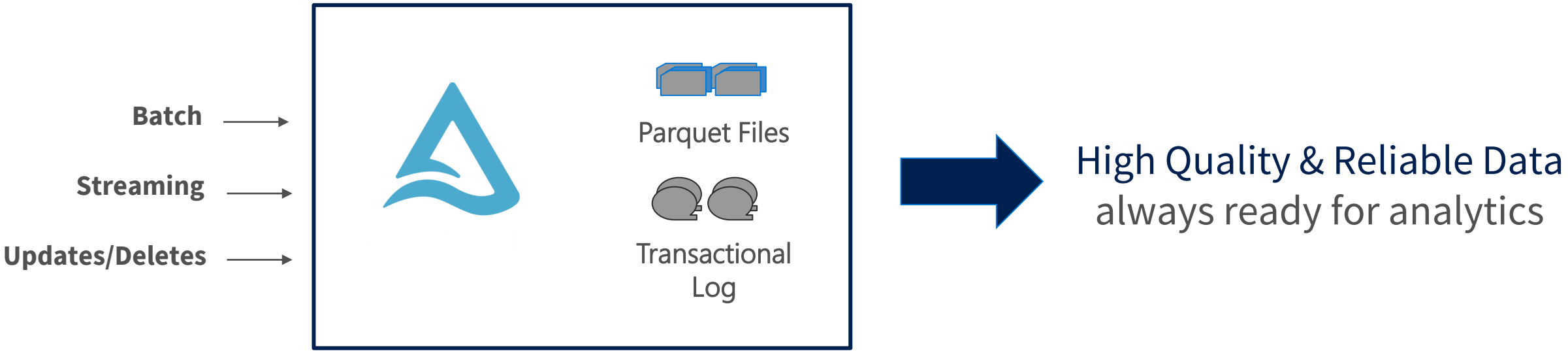
# Delta - A New Standard for Building Data Lakes

Open Format Based on Parquet

With Transactions

Apache Spark API's

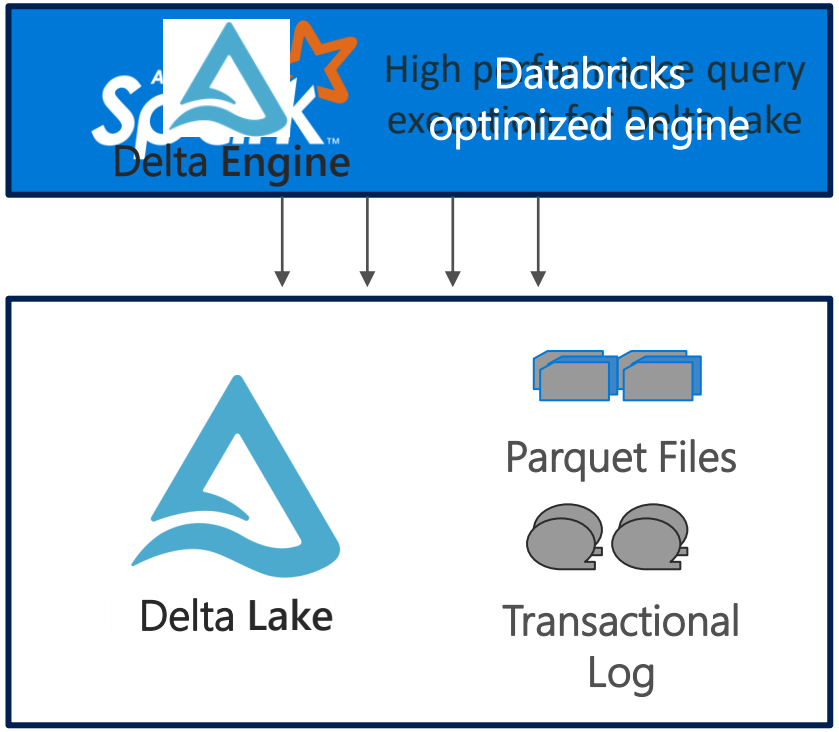
# Delta Lake ensures data reliability




**Key Features**

- ACID Transactions
- Schema Enforcement
- Unified Batch & Streaming
- Time Travel/Data Snapshots

# Delta Lake optimizes performance




Highly Performant
queries at scale

**Key Features**

- Indexing
- Compaction
- Data skipping
- Caching

# Get Started with Delta using Spark APIs

Instead of **parquet**

```
CREATE TABLE ...
USING parquet
...

dataframe
    .write
    .format("parquet")
    .save("/data")
```

... simply say **delta**

```
CREATE TABLE ...
USING delta
…

dataframe
    .write
    .format("delta")
    .save("/data")
```

# Using Delta with your Existing Parquet Tables

## Step 1: Convert `Parquet` to `Delta` Tables

```
CONVERT TO DELTA parquet.`path/to/table` [NO STATISTICS]
[PARTITIONED BY (col_name1 col_type1, col_name2 col_type2, ...)]
```

## Step 2: Optimize Layout for Fast Queries

```
OPTIMIZE events
WHERE date >= current_timestamp() - INTERVAL 1 day
ZORDER BY (eventType)
```

# Upsert/Merge: Fine-grained Updates

```
MERGE INTO customers     -- Delta table
USING updates
ON customers.customerId = source.customerId
WHEN MATCHED THEN
      UPDATE SET address = updates.address
WHEN NOT MATCHED THEN
      INSERT (customerId, address) VALUES (updates.customerId, updates.address)
```

# The Alternative in Hive QL…

```
--Step 1: Create temporary tables to hold merge records
CREATE TABLE merge_demo1wmmergeupdate LIKE merge_demo1;

--Step 2: Insert records when condition is MATCHED
INSERT INTO table merge_demo1WMMergeUpdate
SELECT A.id        AS ID,
       A.firstname AS FirstName,
       CASE
         WHEN B.id IS NOT NULL THEN B.lastname
         ELSE A.lastname
       END         AS LastName
FROM   merge_demo1 AS A
       LEFT OUTER JOIN merge_demo2 AS B
                    ON A.id = B.id;

--Step 3: Insert records when condition is NOT MATCHED
INSERT INTO merge_demo1wmmergeupdate
SELECT B.id        AS ID,
       B.firstname AS FirstName,
       B.lastname  AS LastName
merge_demo2 AS B
       FROM LEFT OUTER JOIN merge_demo1wmmergeupdate AS A
                    ON A.id = B.id
WHERE  A.id IS NULL;
```

```
--Step 4: Drop origianal table
DROP TABLE IF EXISTS merge_demo1;

--Step 5: Rename temp table to origianal table
ALTER TABLE merge_demo1wmmergeupdate
  RENAME TO merge_demo1;

--Step 6: Drop temp table if exists
DROP TABLE IF EXISTS merge_demo1wmmergeupdate;
```

# Time Travel (Data Versioning)

## Reproduce experiments & reports

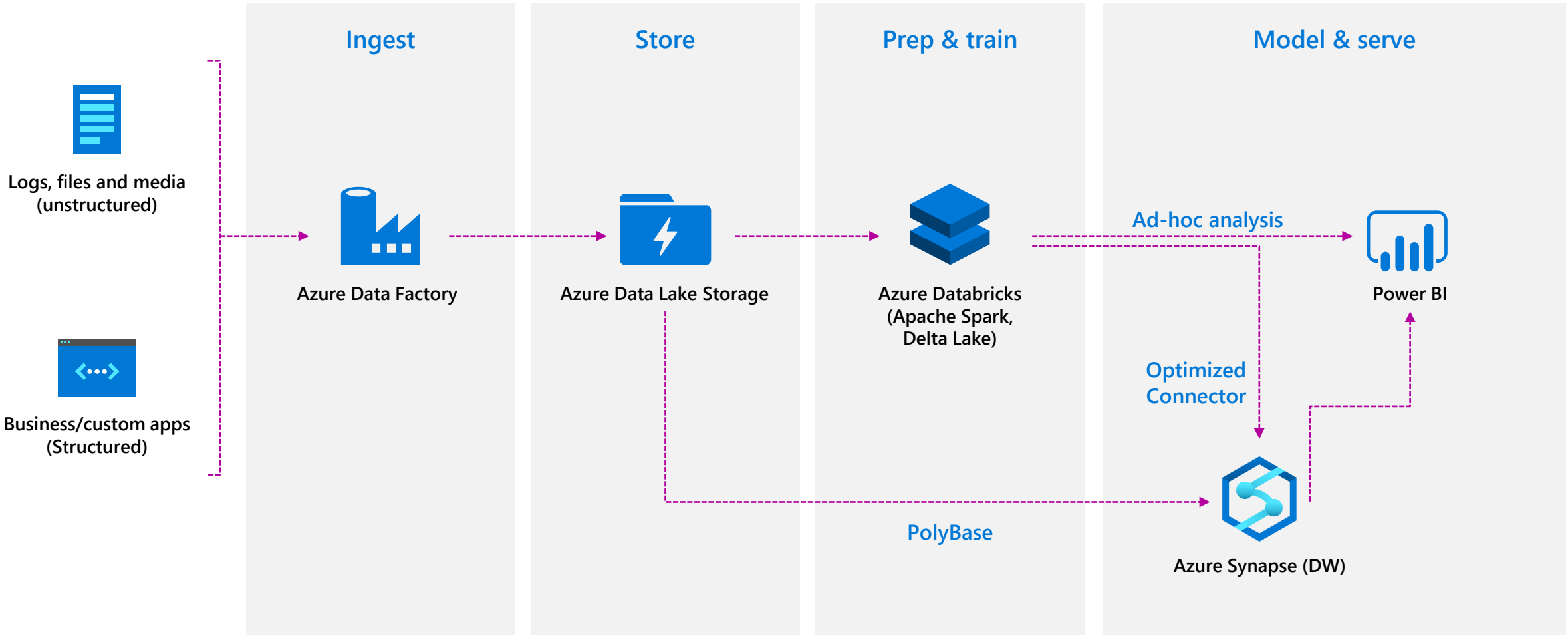
```
SELECT count(*) FROM events
TIMESTAMP AS OF timestamp

SELECT count(*) FROM events
VERSION AS OF version

spark.read.format("delta").option("timestampAsOf",
timestamp_string).load("/events/")
```

## Rollback accidental bad writes

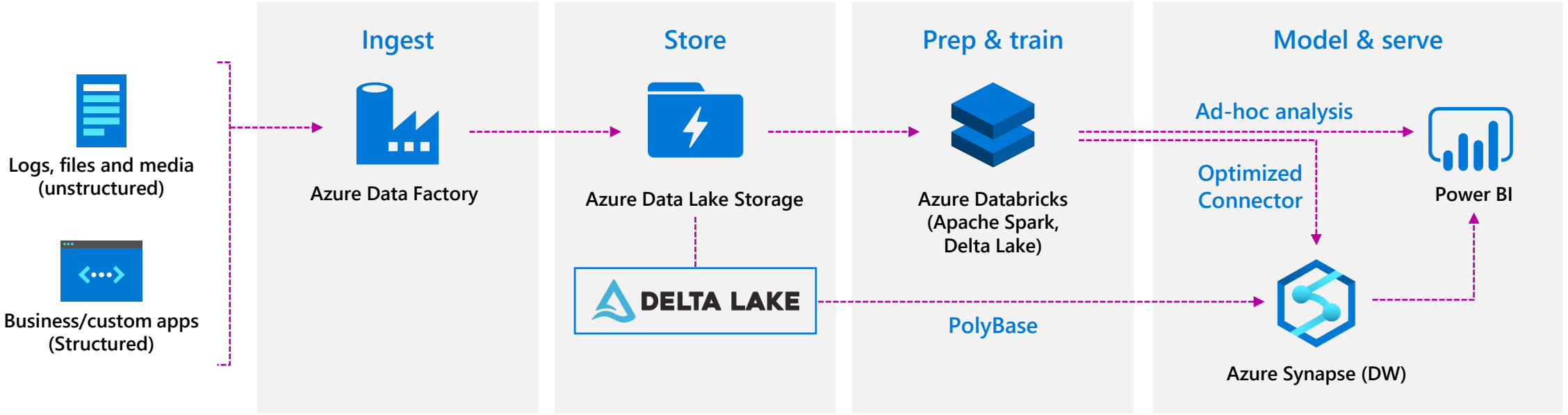
```
INSERT INTO my_table
    SELECT * FROM my_table TIMESTAMP AS OF
    date_sub(current_date(), 1)
```

Azure Databricks

# The modern data warehouse



# The modern data warehouse




**ACID Transaction Guarantees**
Atomic, Consistent, Isolated, Durable

**Versioned parquet files**
Delta transaction log keeps track of all operations

**Efficient Upserts**
MERGE, DELETE, UPDATE

**Time Travel**
Audit history, pipeline debugging, data reproducibility

**Small file compaction w/ no interrupt to availability**
OPTIMIZE and VACUUM

**Z-Order partitioning w/ up to 100x perf**
New multidimensional partitioning enables data skipping

# AZURE DATA FACTORY

Hybrid data integration, at global scale

PRODUCTIVE

- ✓ Drag & Drop UI
- ✓ Codeless Data Movement

HYBRID

- ✓ Orchestrate where your data lives
- ✓ Lift SSIS packages to Azure

SCALABLE

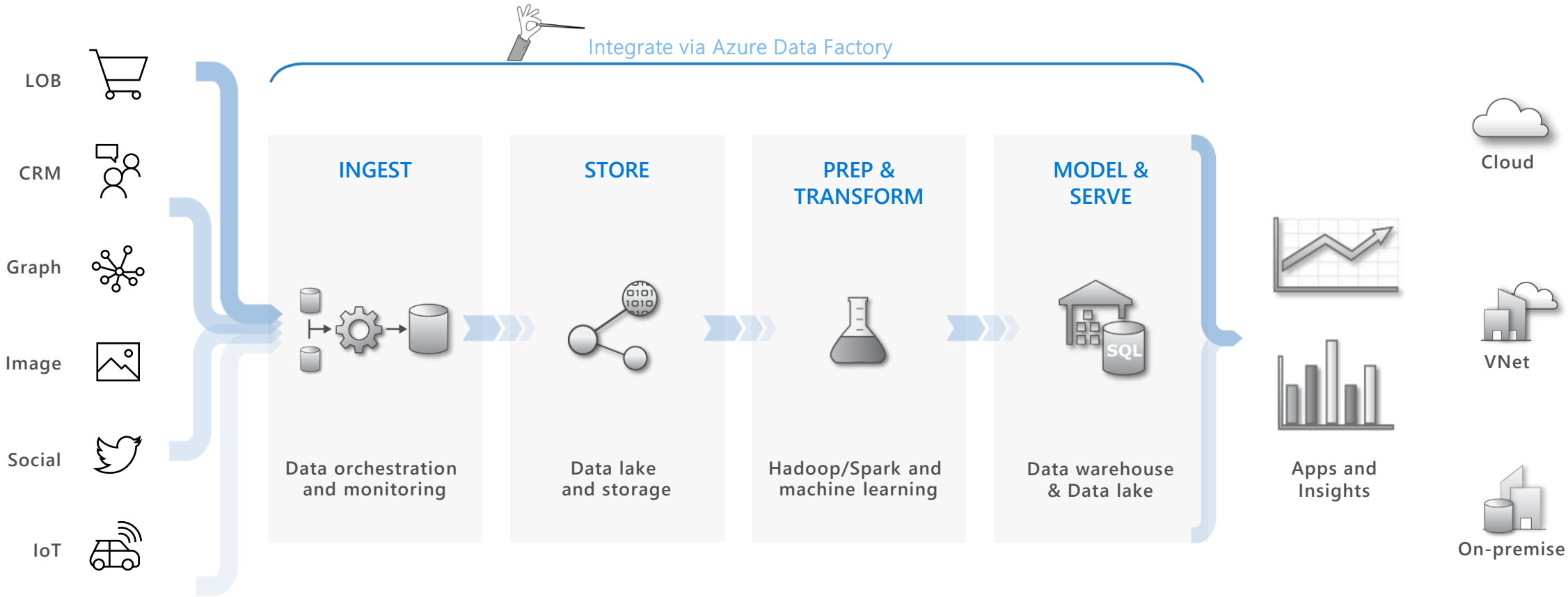
- ✓ Serverless scalability with no infrastructure to manage

TRUSTED

- ✓ Certified compliant Data Movement

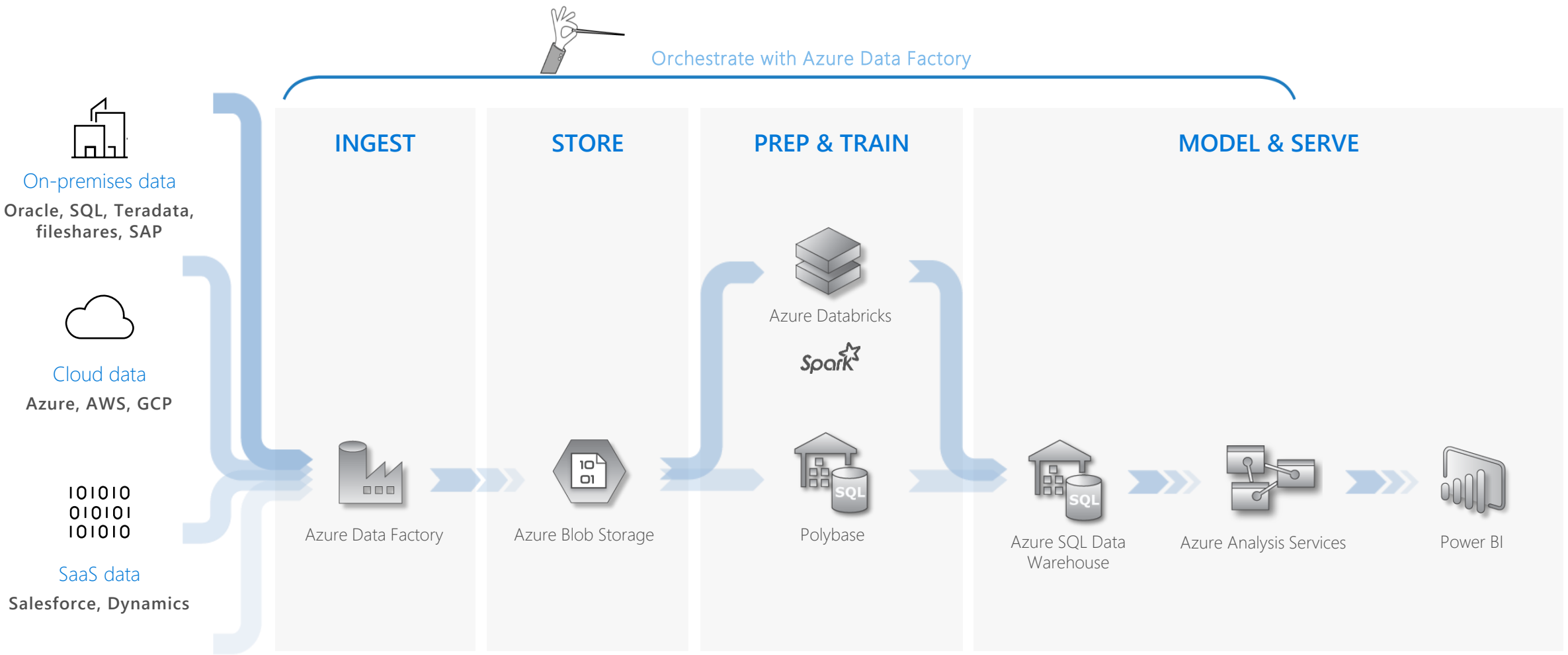
# AZURE DATA FACTORY

## Hybrid data integration, at global scale



# AZURE DATA FACTORY

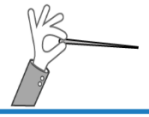
## Modernize your enterprise data warehouse at scale



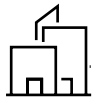

Microsoft Azure also supports other **Big Data** services like **Azure HDInsight**, **Azure SQL Database** and **Azure Data Lake** to allow customers to tailor the above architecture to meet their unique needs.

# AZURE DATA FACTORY

## Create intelligent data driven applications



Orchestrate with Azure Data Factory

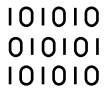


On-premises data

**Oracle, SQL, Teradata, fileshares, SAP**

Cloud data

**Azure, AWS, GCP**



SaaS data

**Salesforce, Dynamics**

DATA TIER

APPLICATION TIER

### INGEST

Azure Data Factory

### STORE

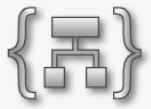
Azure Blob Storage

### PREP & TRAIN


Azure Databricks

### MODEL & SERVE

Azure Cosmos DB

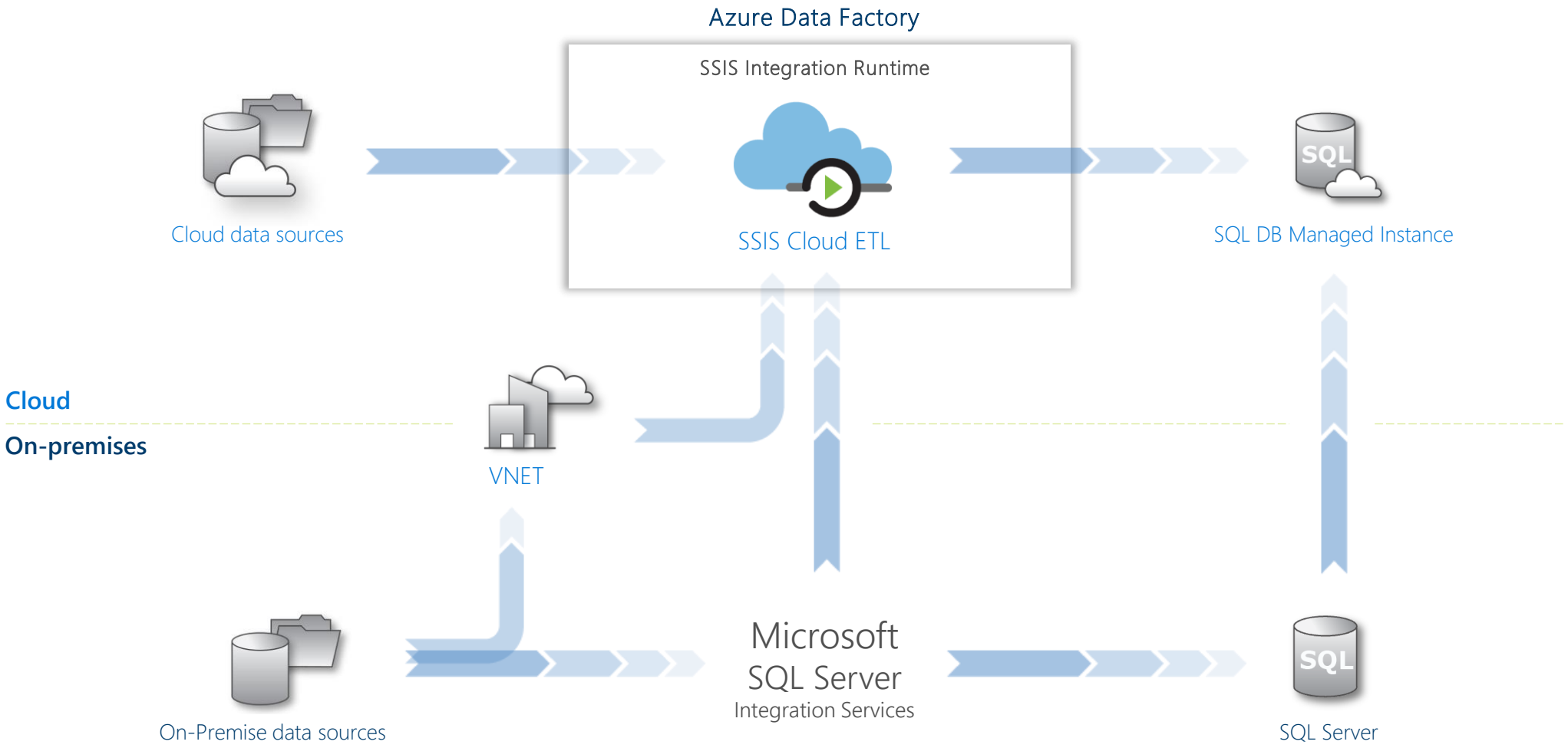


Application Logic

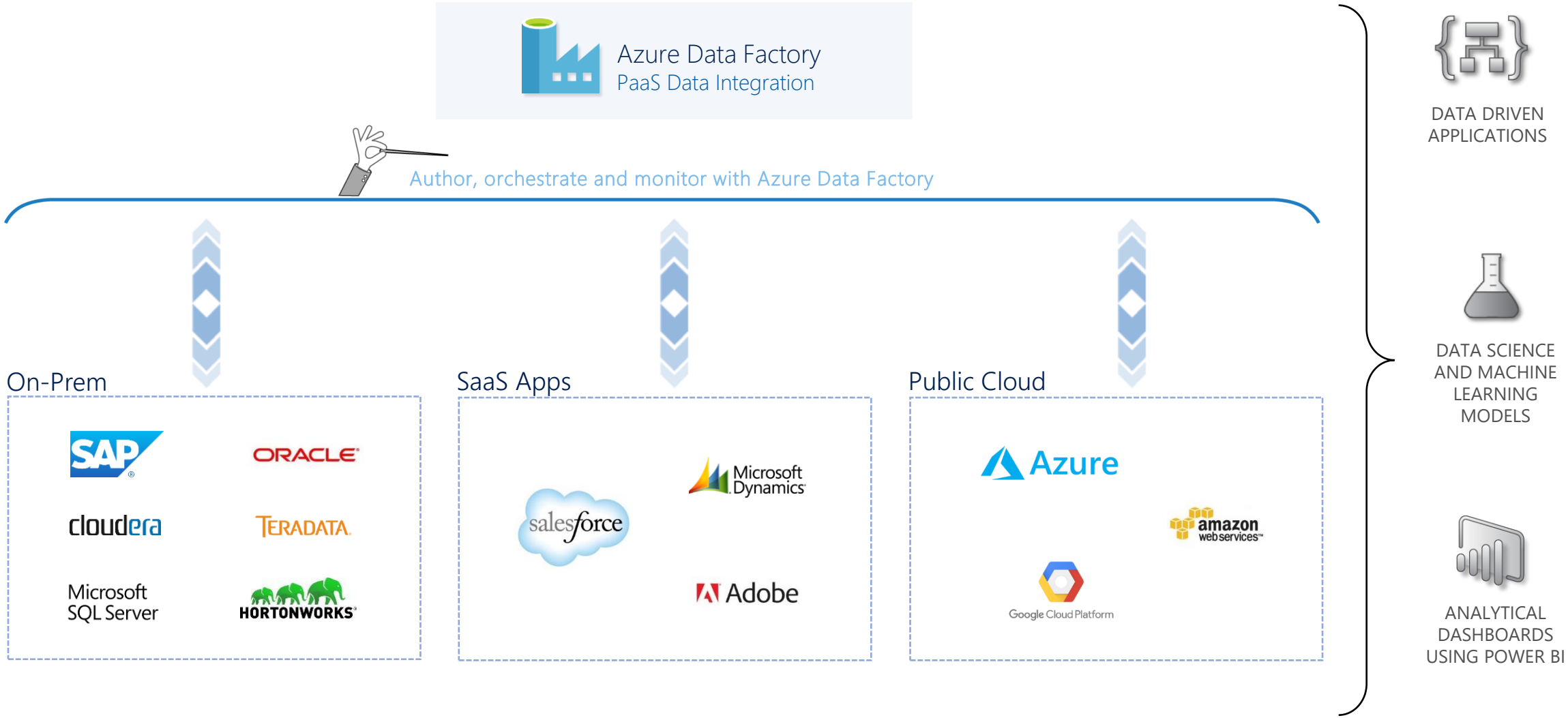
Microsoft Azure also supports other **Big Data** services like **Azure HDInsight**, **Azure SQL Database** and **Azure Data Lake** to allow customers to tailor the above architecture to meet their unique needs.

# AZURE DATA FACTORY

## Lift your SQL Server Integration Services (SSIS) packages to Azure

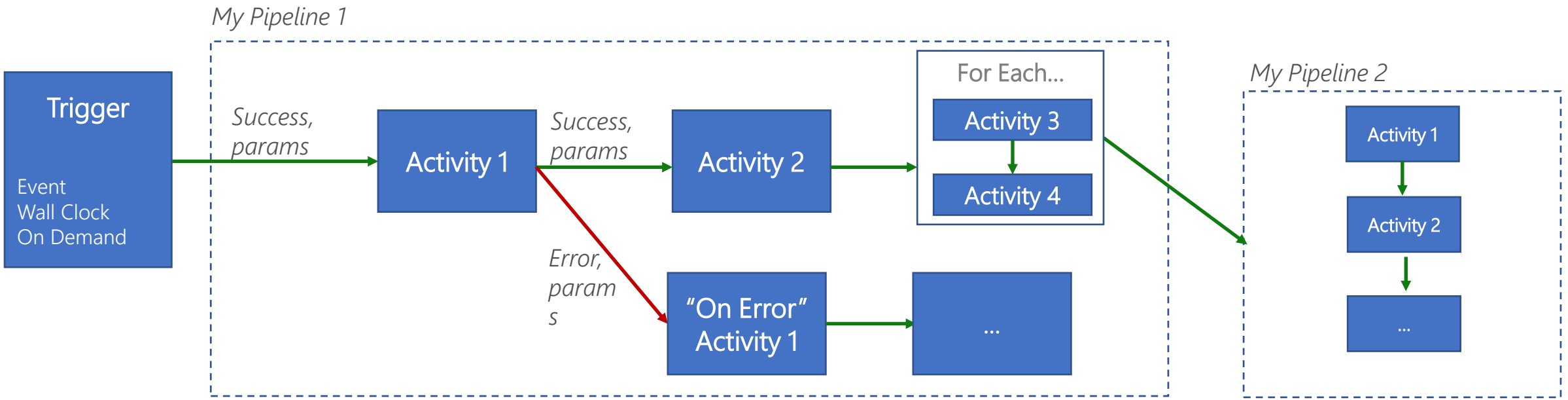


# Hybrid and Multi-Cloud Data Integration

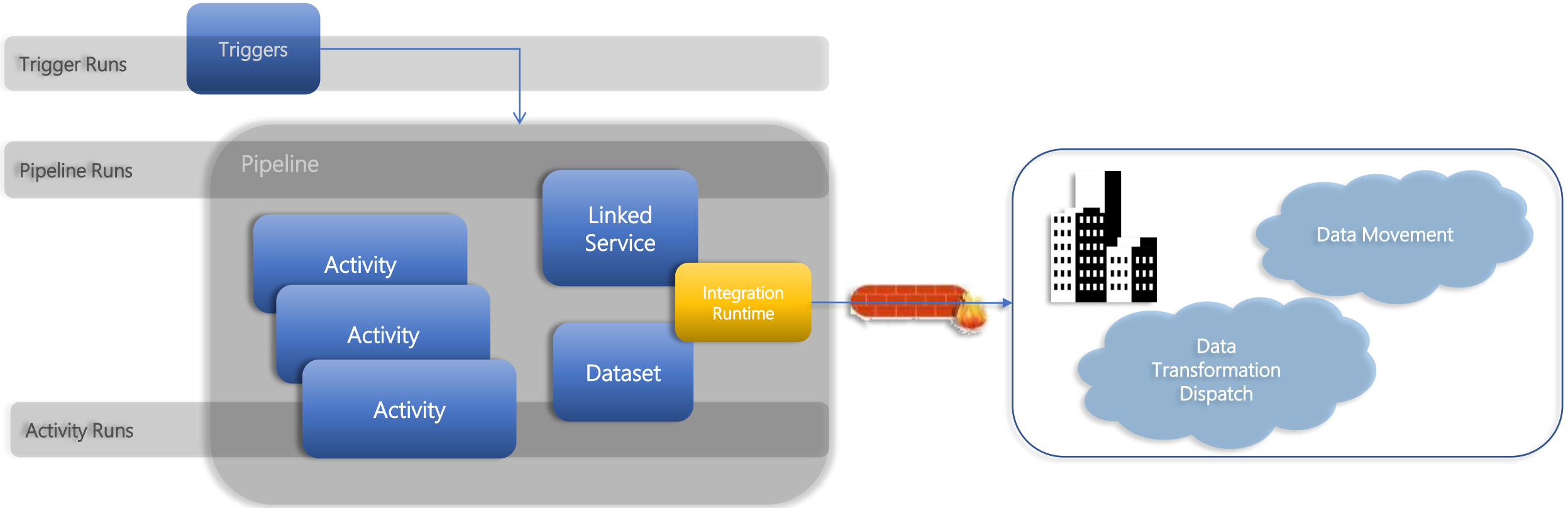


# Control Flow in Azure Data Factory

Coordinate pipeline activities into finite execution steps to enable looping, conditionals and chaining while separating data transformations into individual data flows

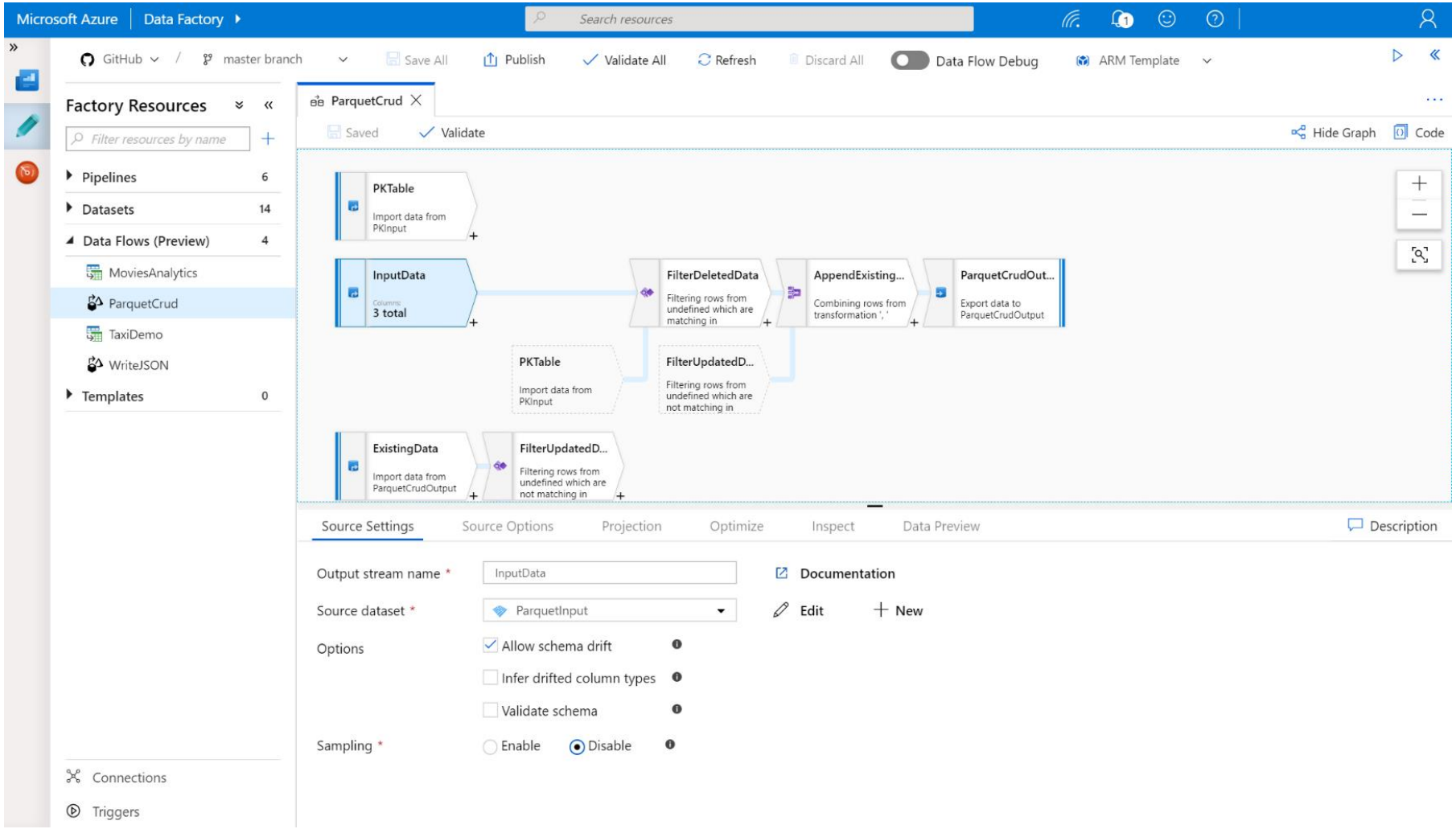


# Azure Data Factory Flexible Application Model



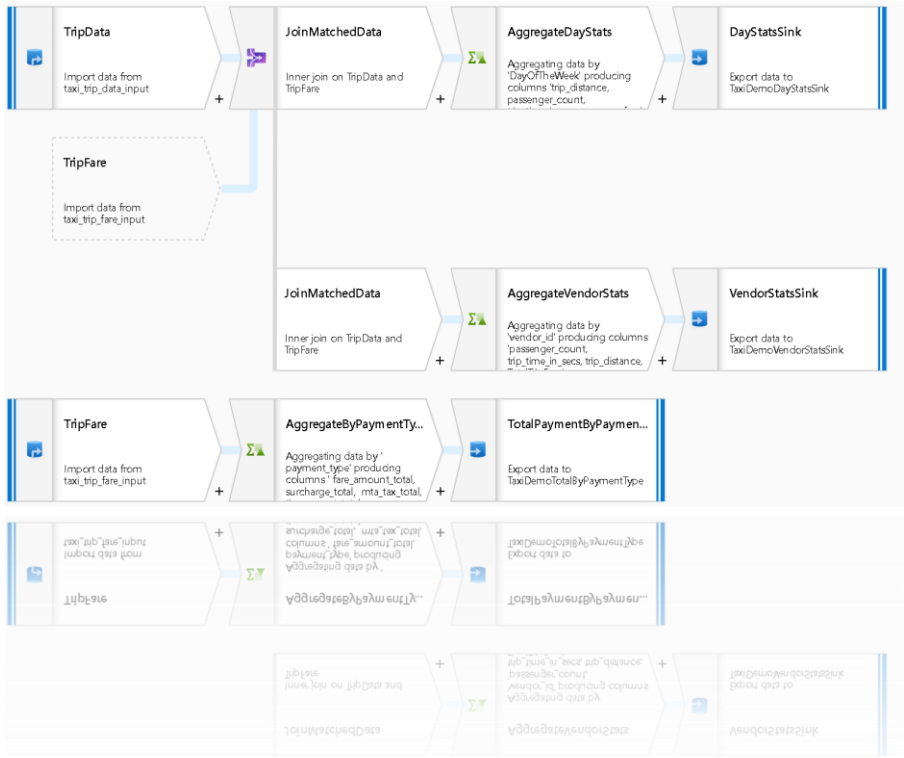
# Data Flows

Data Flow is a new feature of Azure Data Factory that allows you to build data transformations in a visual user interface. **Code-free pipelines**



# Code-free Data Transformation At Scale

- Does not require understanding of Spark, big data execution engines, clusters, Scala, Python, etc
- Focus on building business logic and data transformation
  - Data cleansing
  - Aggregation
  - Data conversions
  - Data prep
  - Data exploration




... not ...

# WHAT MAKES A GREAT DATA LAKE?

| Massive scale | Granular, multi-layered Security | Optimized for Maximum Performance | Integration Friendly | Cost Effectiveness |
|---|---|---|---|---|
| PB Scale, data accessible everywhere, growth on demand | Granular security and protection against accidental data loss | Lightning quick job execution | Supports multiple methods of data ingress, processing, egress and visualization | Cloud economic model with the ability to intelligently manage costs |

**Rich Data Management and Governance**

# Azure Data Lake Storage Gen2

A "no-compromises" Data Lake: secure, performant, massively-scalable Data Lake storage that brings the cost and scale profile of object storage together with the performance and analytics feature set of data lake storage

SECURE

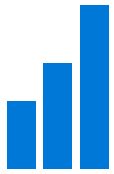
- ✓ Support for fine-grained ACLs, protecting data at the file and folder level
- ✓ Multi-layered protection via at-rest Storage Service encryption and Azure Active Directory integration

MANAGEABLE

- ✓ Automated Lifecycle Policy Management
- ✓ Object Level tiering

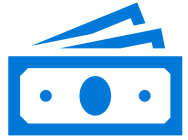
FAST

- ✓ Atomic file operations means jobs complete faster



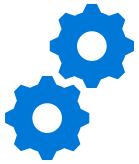
SCALABLE

- ✓ No limits on data store size
- ✓ Global footprint (50 regions)



COST EFFECTIVE

- ✓ Object store pricing levels
- ✓ File system operations minimize transactions required for job completion

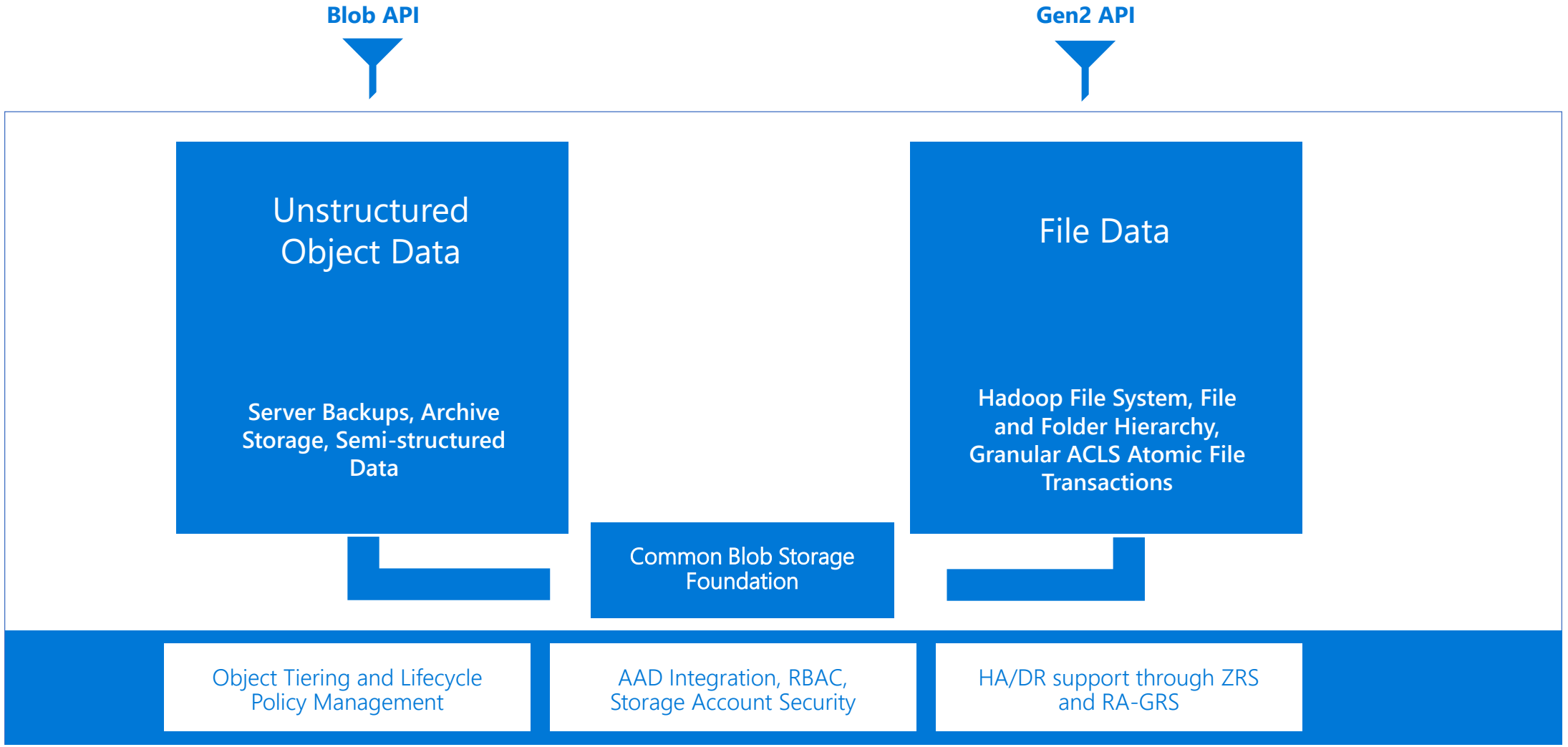


INTEGRATION READY

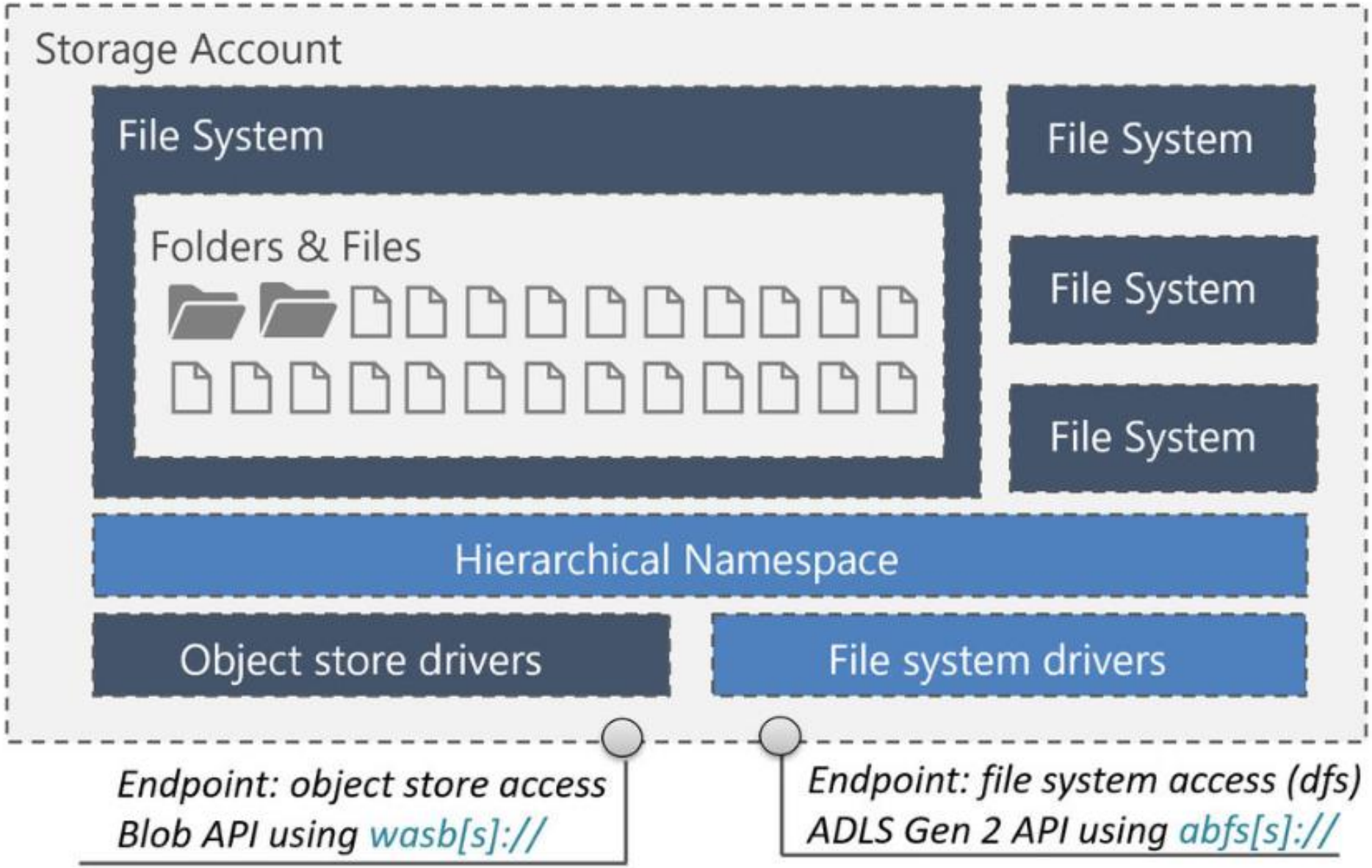
- ✓ Optimized for Spark and Hadoop Analytic Engines
- ✓ Tightly integrated with Azure end to end analytics solutions

# Azure Data Lake Storage Gen2

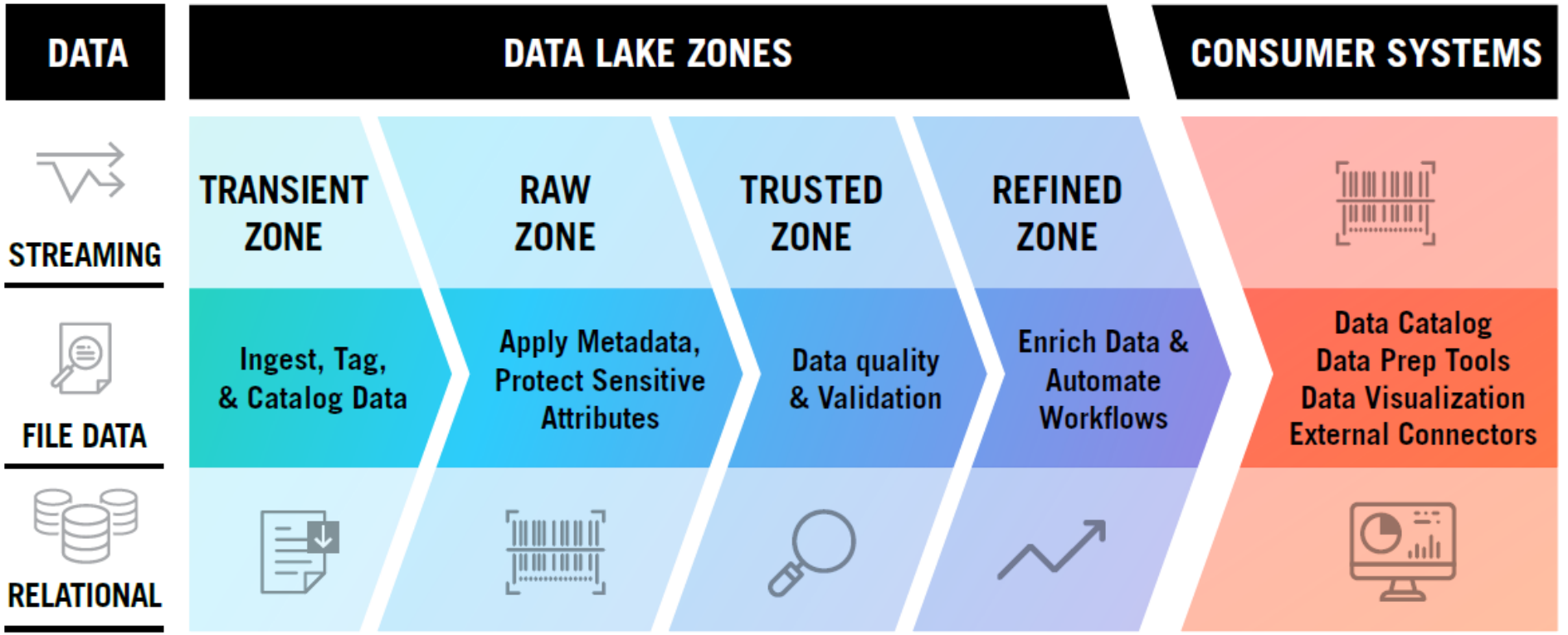
ADLS Gen2 adds a high performance HDFS Endpoint to Azure Blob Storage and inherits the rich feature set of Azure Blob Storage *




** Full Blob and HNS Interoperability available in post-GA*


Storage Account
File System
Folders & Files
File System
File System
File System
Hierarchical Namespace
Object store drivers
File system drivers
Endpoint: object store access
Blob API using wasb[s]://
Endpoint: file system access (dfs)
ADLS Gen 2 API using abfs[s]://

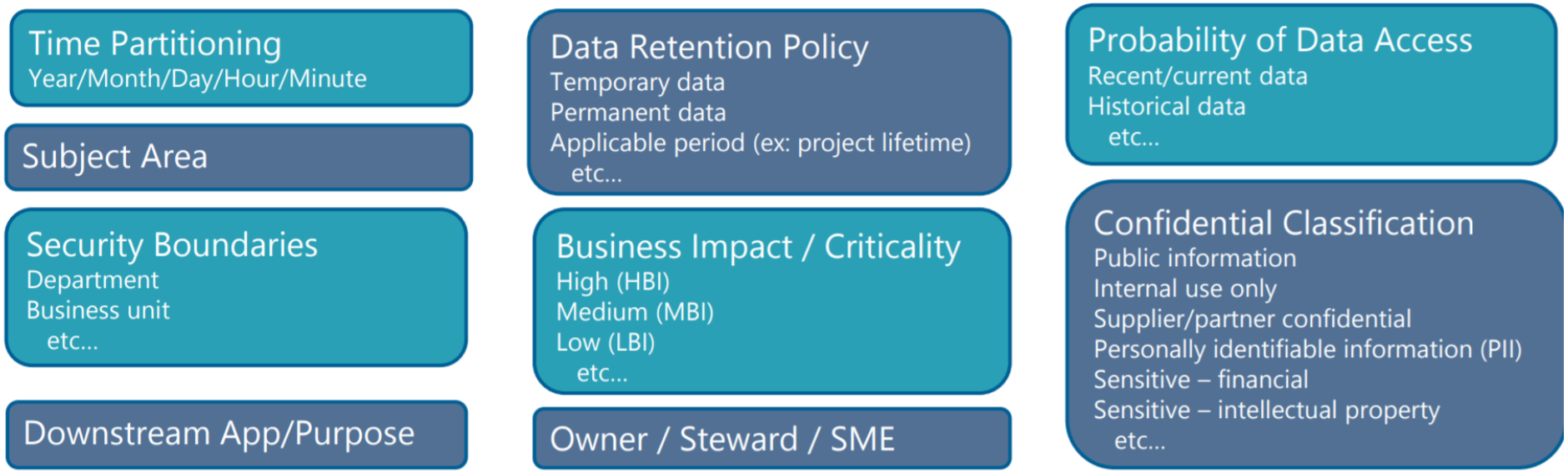
# Data Lake Structure - Example



# 1.2 Data Lake Structure – Considerations

- When organizing the data within a data lake consider:
  - Security boundaries
  - Time partitioning
  - Subject area
  - Confidentiality level
  - Probability of data access (i.e., hot vs. cold data)
  - Data retention policy
  - Owner/steward/subject matter expert

- These different types of data should be separated into various different zones, for better clarity and enforceability of rules and definitions.



Example of date partitioning at year/month level: