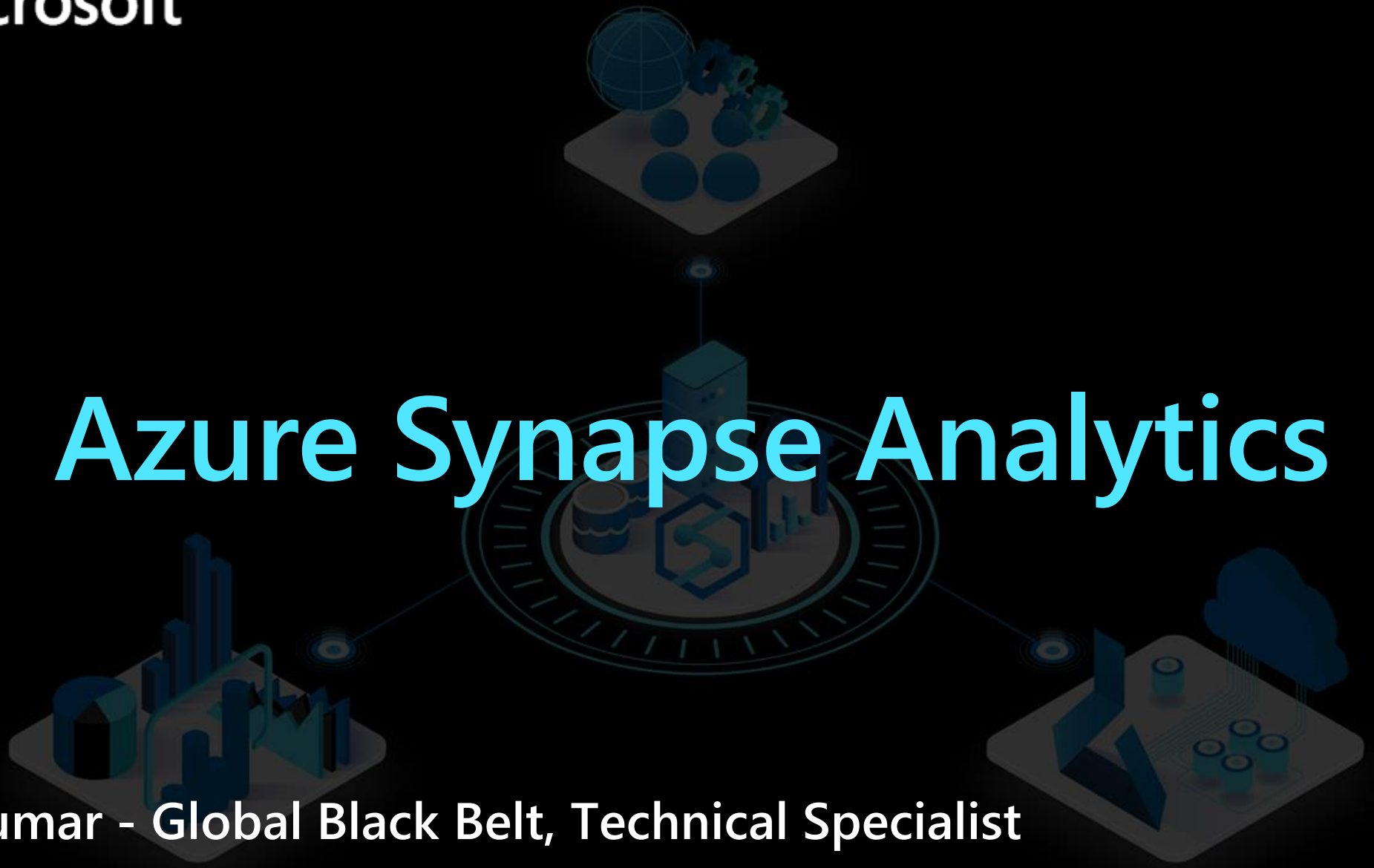
Microsoft

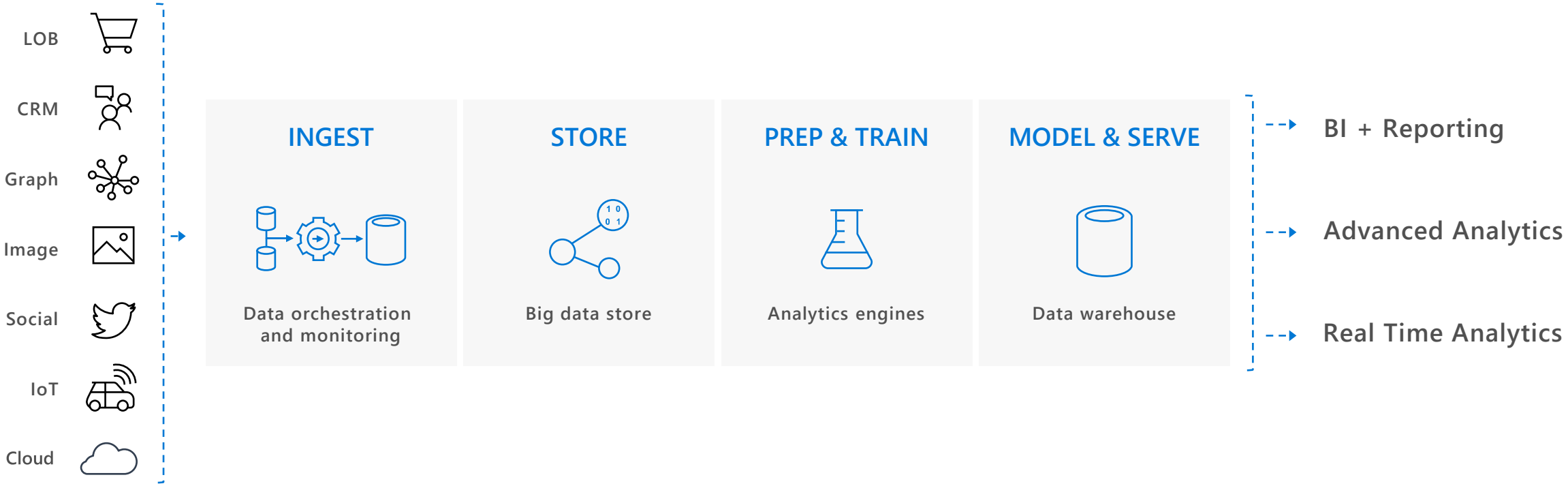


Azure Synapse Analytics

Ashish Kumar - Global Black Belt, Technical Specialist

17th June 2020

Data Platform



Data lake

A data lake is a collection of data, not a platform for data

Hadoop is the preferred platform for data lakes

A data lake handles large volumes of diverse data…

Semi- and un-structured data formats, possibly Exabytes of data

…ingest it quickly…

Straight from data source, no wrangling/ETL

…and persist it in its original, raw and refined formats

Detailed source data as basis for data engineering/science

Benefits of a data lake

Flexible…

…choose and work with whatever tool you prefer

No vendor lock-in…

…full control on your data (data sovereignty)

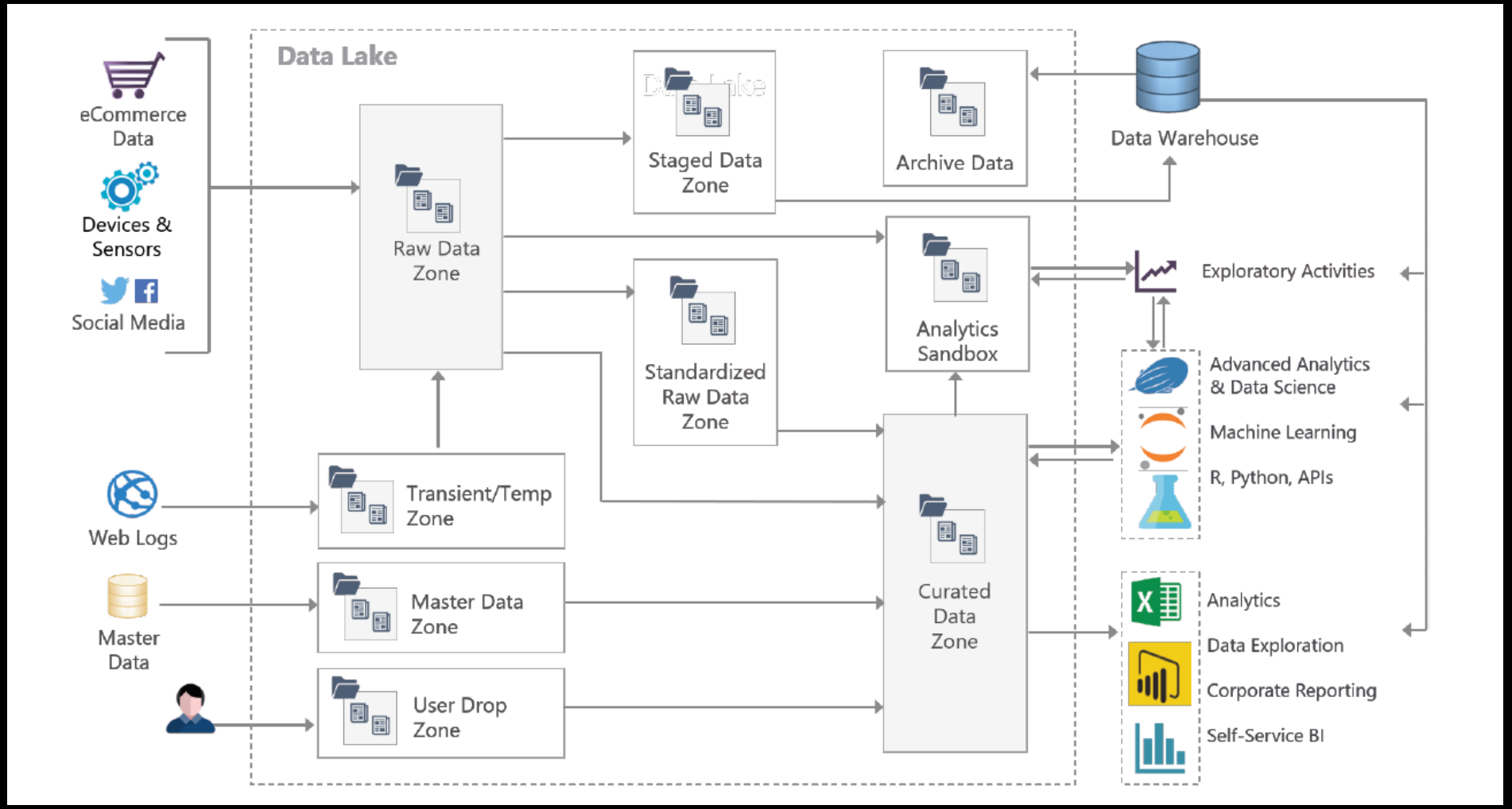
…take your data and move it somewhere else

…you are **NOT** binding to a specific technology or tool

…and that means for you:

NO MORE MIGRATIONS

Typical Data Lake Architecture

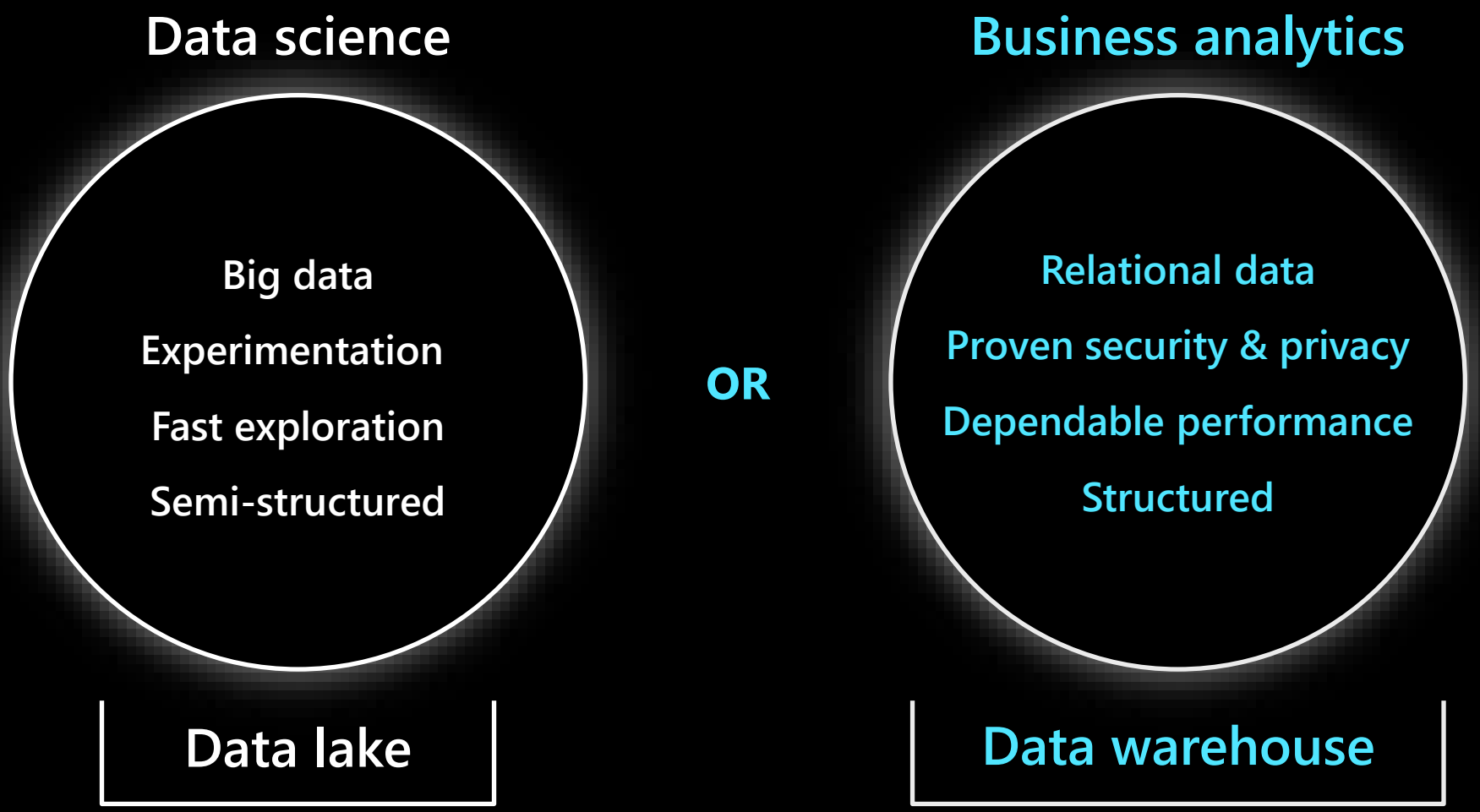


Microsoft

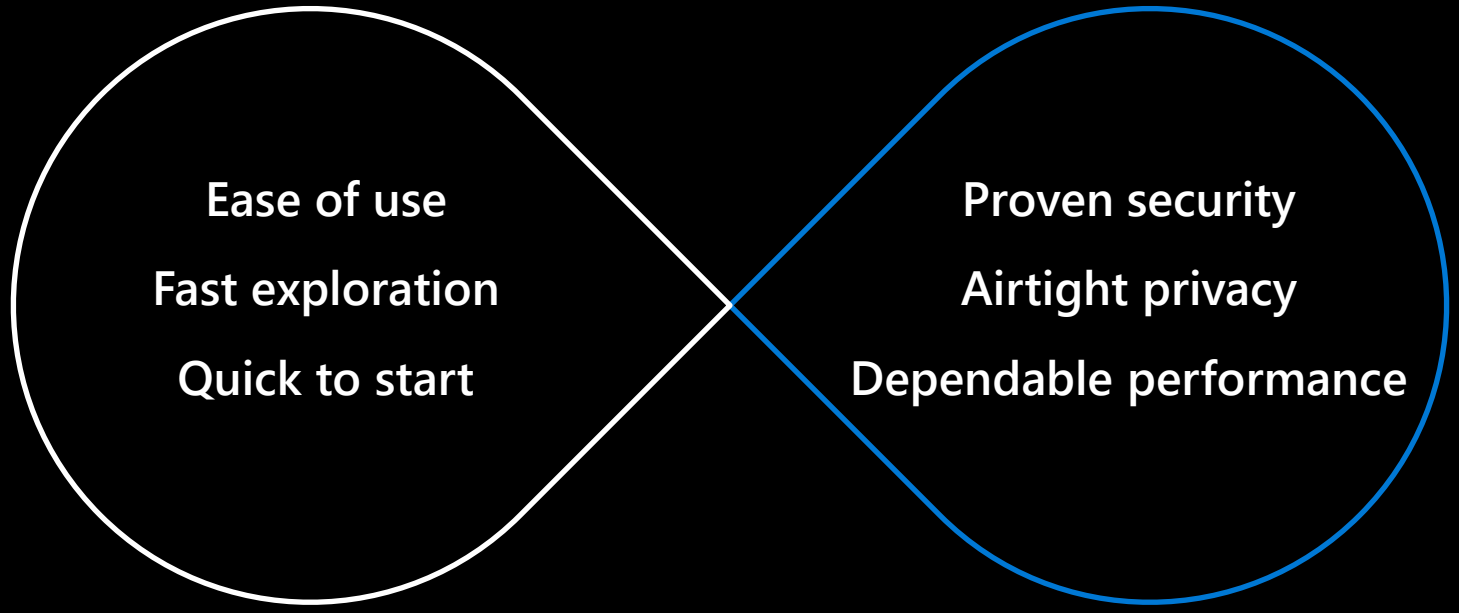
Azure Synapse Analytics

Overview

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



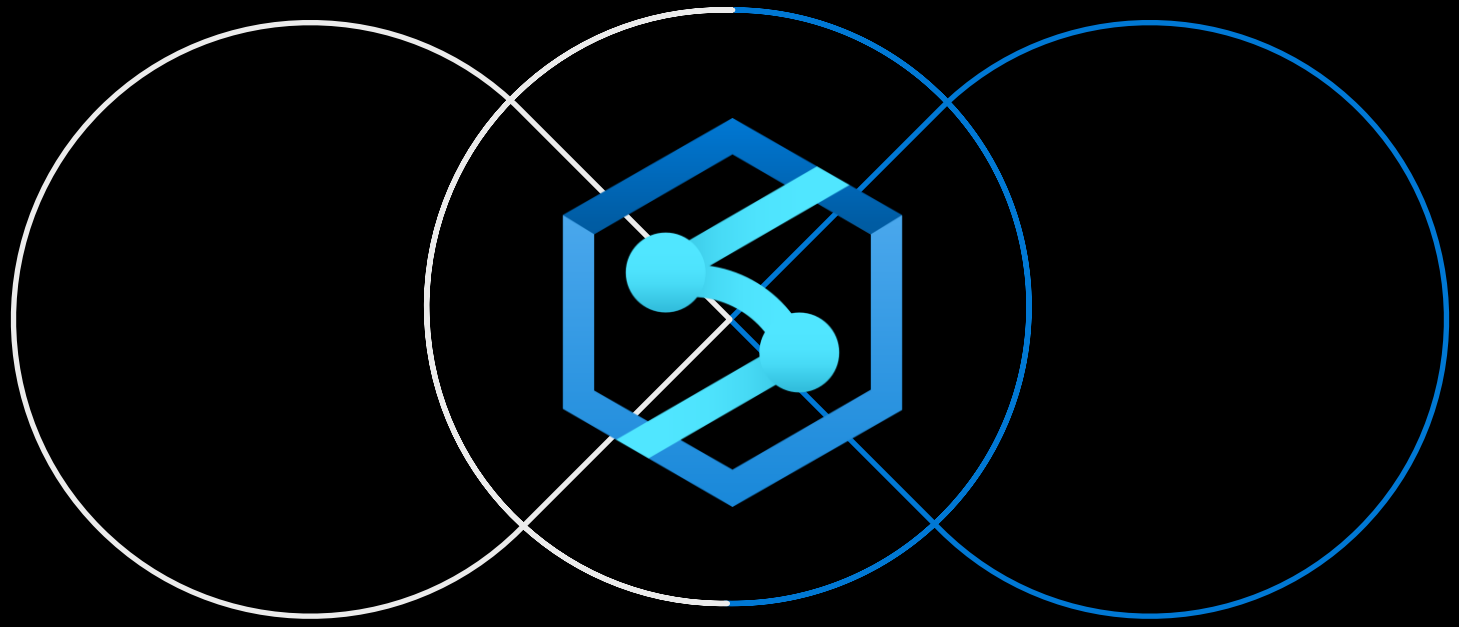
Azure meets these challenges with a single service to provide limitless analytics



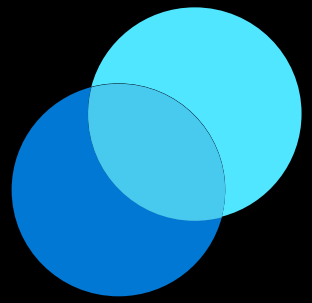
Welcome to limitless

Data warehousing & big data analytics—all in one service

Azure meets these challenges with a single service to provide limitless analytics



Azure Synapse Analytics



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** - blending big data, data warehousing, and data integration into **a single service** for end-to-end analytics at cloud scale

Azure Synapse Analytics Customers

Azure Synapse Analytics Roadmap

May 2020

New GA features
- Resultset caching
- Materialized Views
- Ordered Columnstore
- JSON support
- Dynamic Data Masking
- SSDT support
- Workload Isolation
- Simple ingestion with COPY
- Private LINK support
- Updatable Hash Key

Preview features
- Synapse Studio
- Synapse Link
- SQL Serverless
- Data sharing with Azure Data Share
- Streaming ingestion & analytics in DW
- Native Predict/Scoring
- Bulk Load Wizard
- FROM clause with joins
- Managed Virtual Networks

Synapse Analytics (GA)

Synapse Analytics (GA)
(formerly SQL DW)

Synapse Analytics (PREVIEW)

Limited Preview features
- Multi-cluster warehouse
- Online/auto scaling
- Multi-column Hash Distribution
- SQL MERGE support, DML JOINs
- Column encryption
- Cross-database queries
- Data Flow CDM Support

Synapse SQL

Query and analyze data with T-SQL using both provisioned and serverless models

Synapse Spark

Quickly create notebooks with your choice of Python, Scala, SparkSQL, and .NET for Spark

Synapse Studio

Build end-to-end data-driven workflows for your data movement and data processing scenario

Synapse Pipelines

Execute all data tasks with a simple UI and unified environment

Customer Migration Path

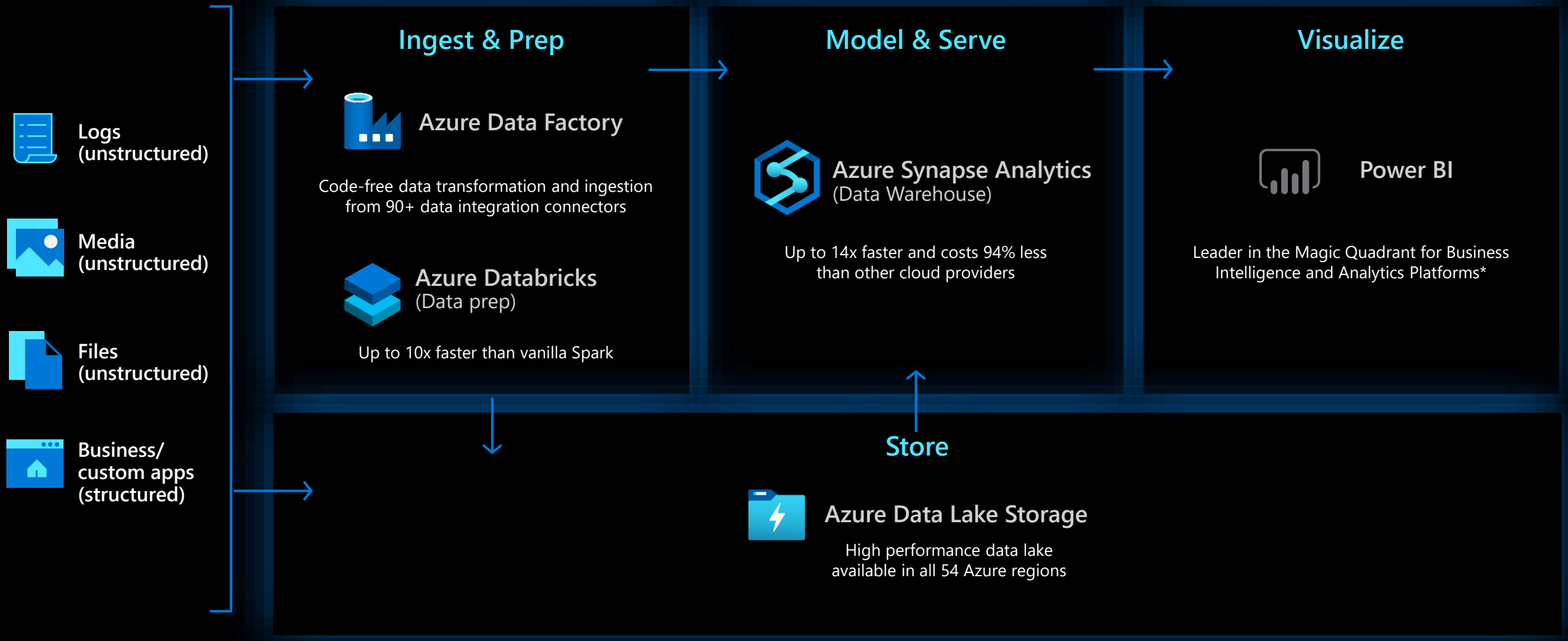
SQL DW/Synapse SQL – Features that were generally available in SQL DW (workload management, row/column security, materialized views etc.) continue to be in GA today. Businesses can continue running their existing DW workloads in production today with Azure Synapse and will automatically benefit from the new capabilities in preview (web studio, query-as-a-service, built-in data integration, integrated Apache Spark etc.) once they are GA and can use them in production if they choose to do so. **Customers will not have to migrate any workloads as SQL DW will simply be moved under a Synapse workspace. Use SSMS** to connect to both SQL Serverless SQL and provisioned SQL.

Azure Data Factory - Continue using ADF. When Synapse Pipeline within Azure Synapse becomes generally available, import your ADF pipelines into Azure Synapse workspace. Existing ADF artefacts will work with Azure Synapse if customers choose not to import them into the Azure Synapse workspace. Note that Azure-SSIS Integration Runtime (IR) will not be supported in Synapse.

Power BI – Link to a Power BI workspace within Azure Synapse Studio so no migration needed

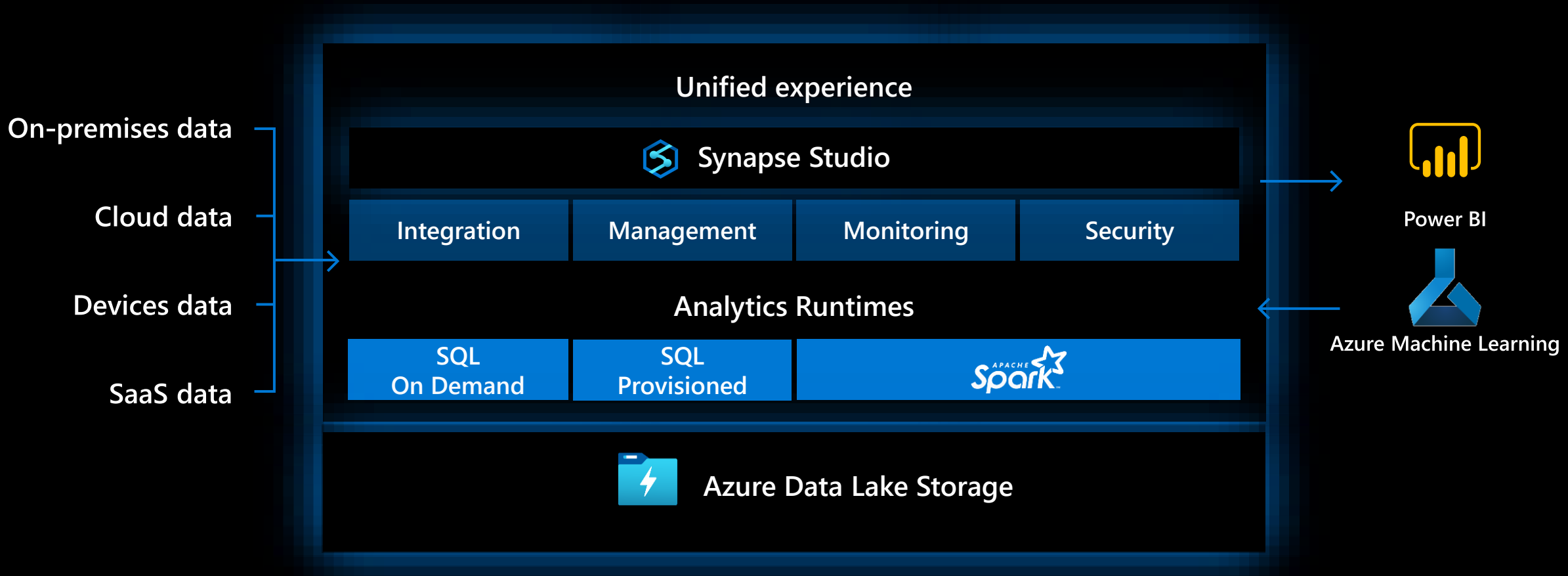
Azure Databricks – ADB notebooks can be exported as .ipynb files and then imported into Synapse Spark, that part is easy. The hard part is if any code dependencies exist in the user code on features that are unique to ADB like dbutils or behaviors that are unique to ADB like ML Runtime, GPU support etc.

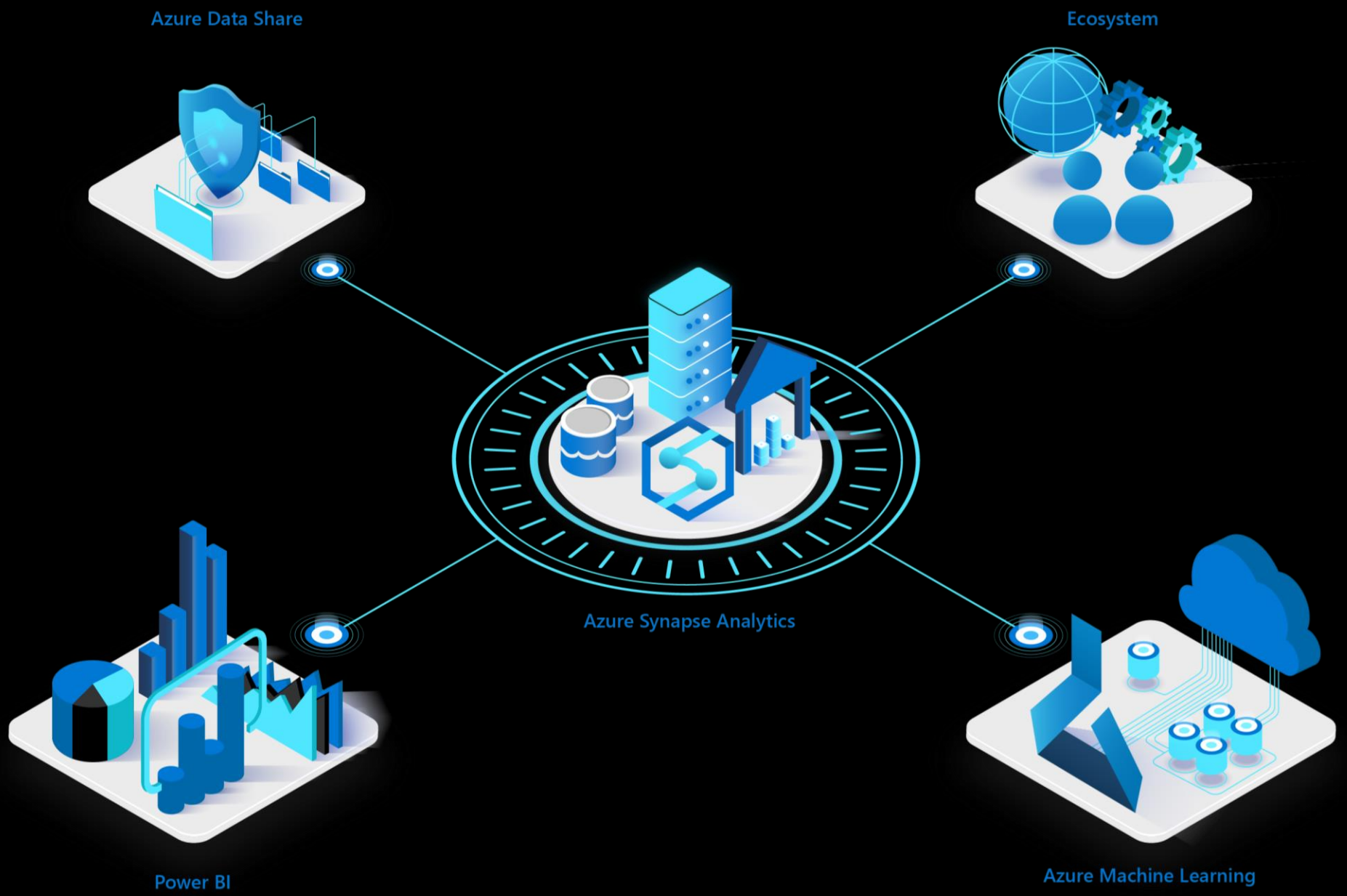
The existing Modern Data Warehouse……

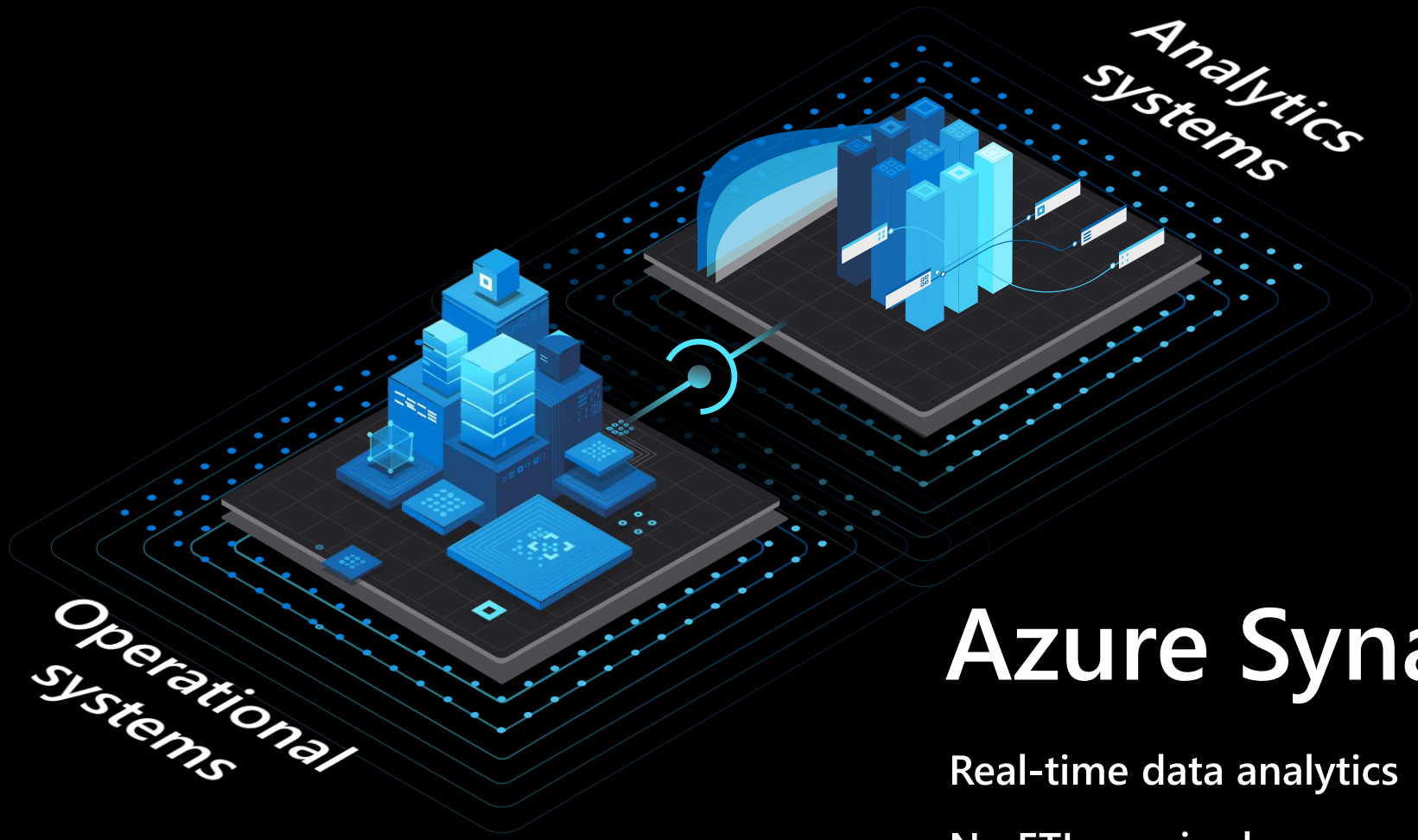


Azure Synapse Analytics

Limitless data warehouse with unmatched time to insights




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



Azure Synapse Link

Real-time data analytics

No ETL required

No performance impact on transactions

Azure Synapse Analytics

Limitless scale

Powerful insights

Unified experience

Instant clarity

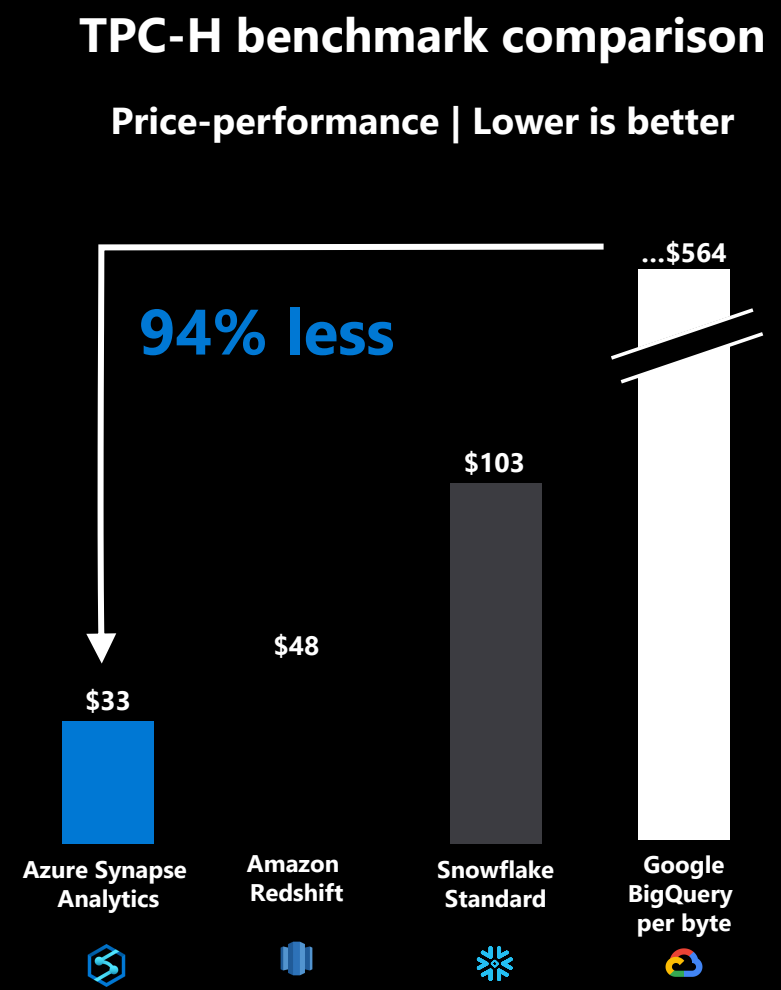
Unmatched security

Price Performance

A breakthrough in the cost of enterprise analytics

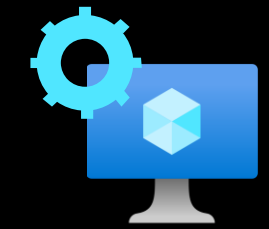
With the best price-performance in the business

Up to 14x faster and costs 94% less than other cloud providers



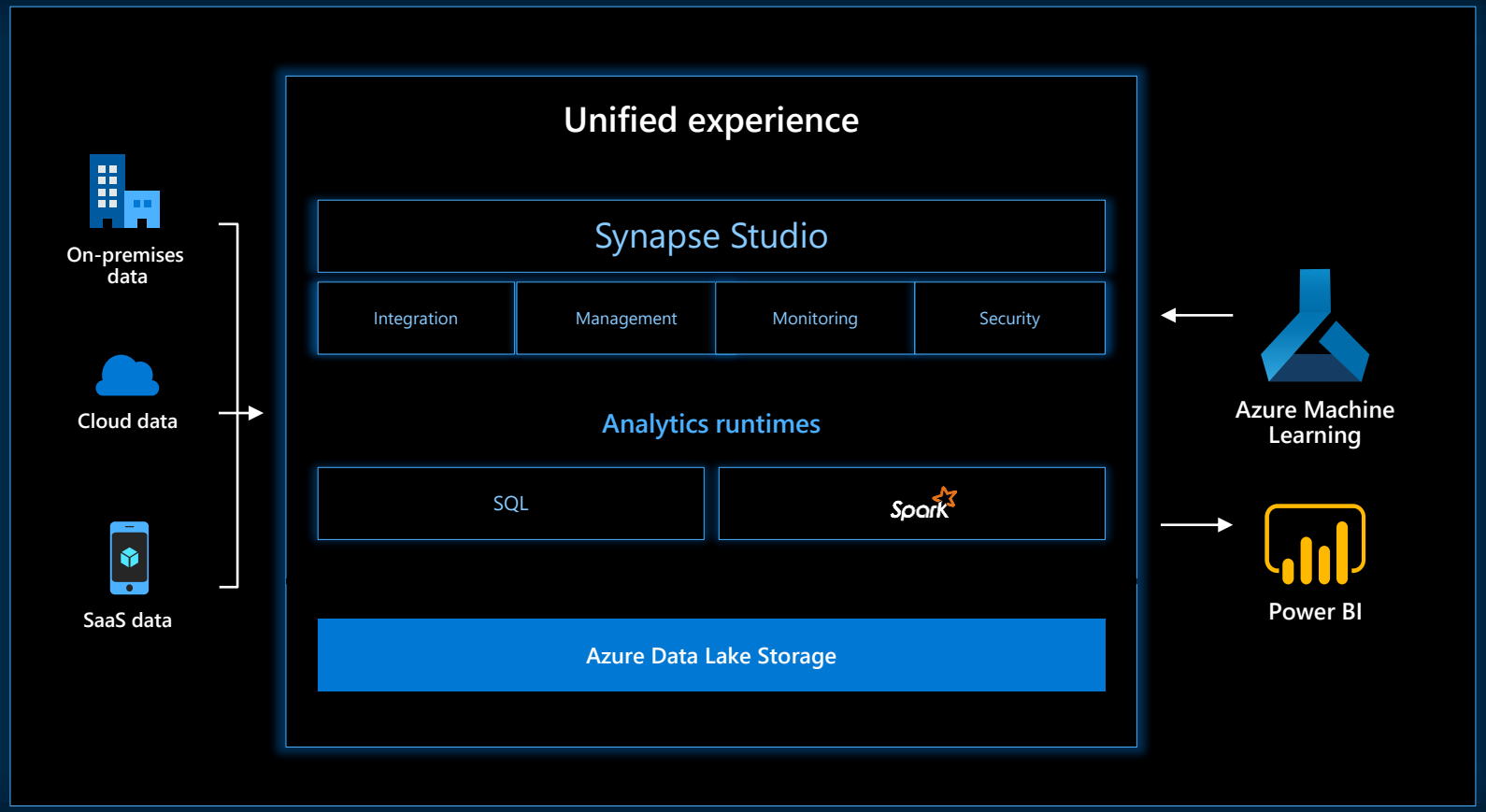
* GigaOm TPC-H benchmark report, January 2019, "GigaOm report: Data Warehouse in the Cloud Benchmark

Setup and Administration



Azure Synapse is easier to setup and administer

Synapse Studio provides a unified end-to-end experience that simplifies and automates setup and administration



Security, Privacy and Compliance

Rigorous assurance of safe-keeping with the most advanced security and privacy features

We protect sensitive data in real time, monitoring and responding to threats as they arise, with industry-leading security and privacy features at no extra cost to you.

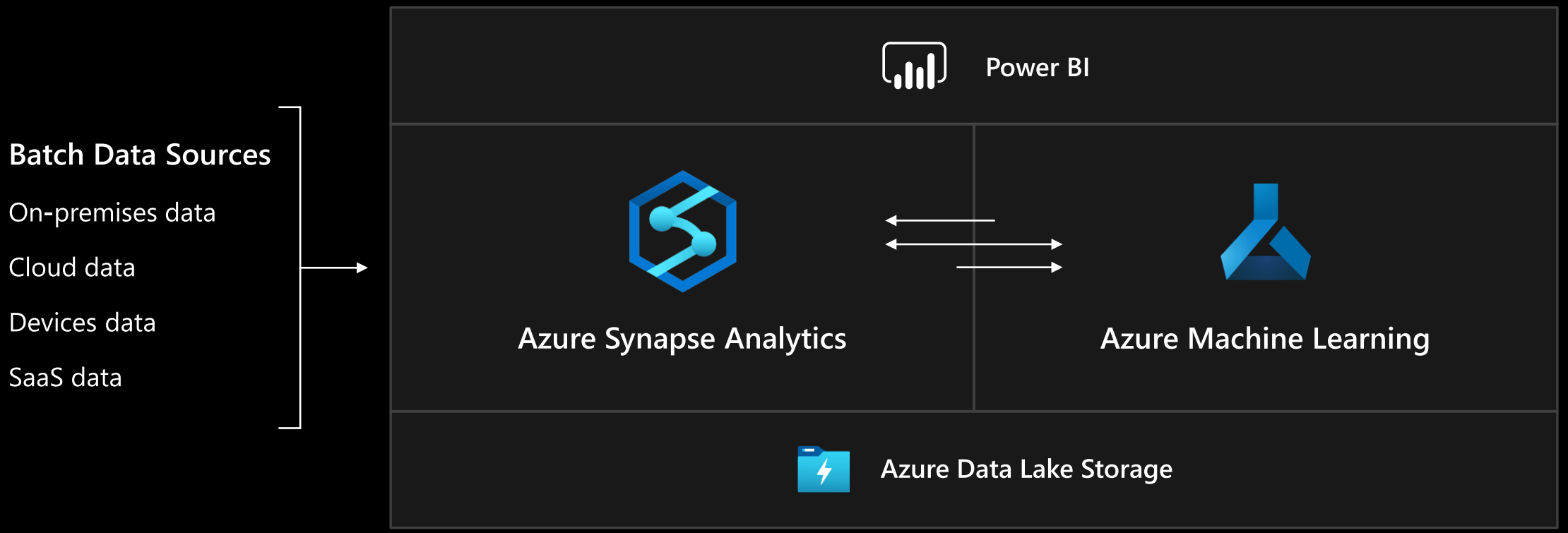
Access control for complete security

Category	Feature	Azure Synapse Analytics
Data Protection	Data In Transit	Yes
	Data encryption at rest (Service & User Managed Keys)	Yes
	Data Discovery and Classification	Yes
Access Control	Native Row Level Security	Yes
	Table and View Security (GRANT / DENY)	Yes
	Column Level Security	Yes
	Dynamic Data Masking	Yes
Authentication	SQL Authentication	Yes
	Native Azure Active Directory	Yes
	Integrated Security	Yes
	Multi-Factor Authentication	Yes
Network Security	Virtual Network (VNET)	Yes
	SQL Firewall (server)	Yes
	Integration with ExpressRoute	Yes
Threat Protection	SQL Threat Detection	Yes
	SQL Auditing	Yes
	Vulnerability Assessment	Yes

Azure has more compliance certifications than any other vendor

Machine Learning

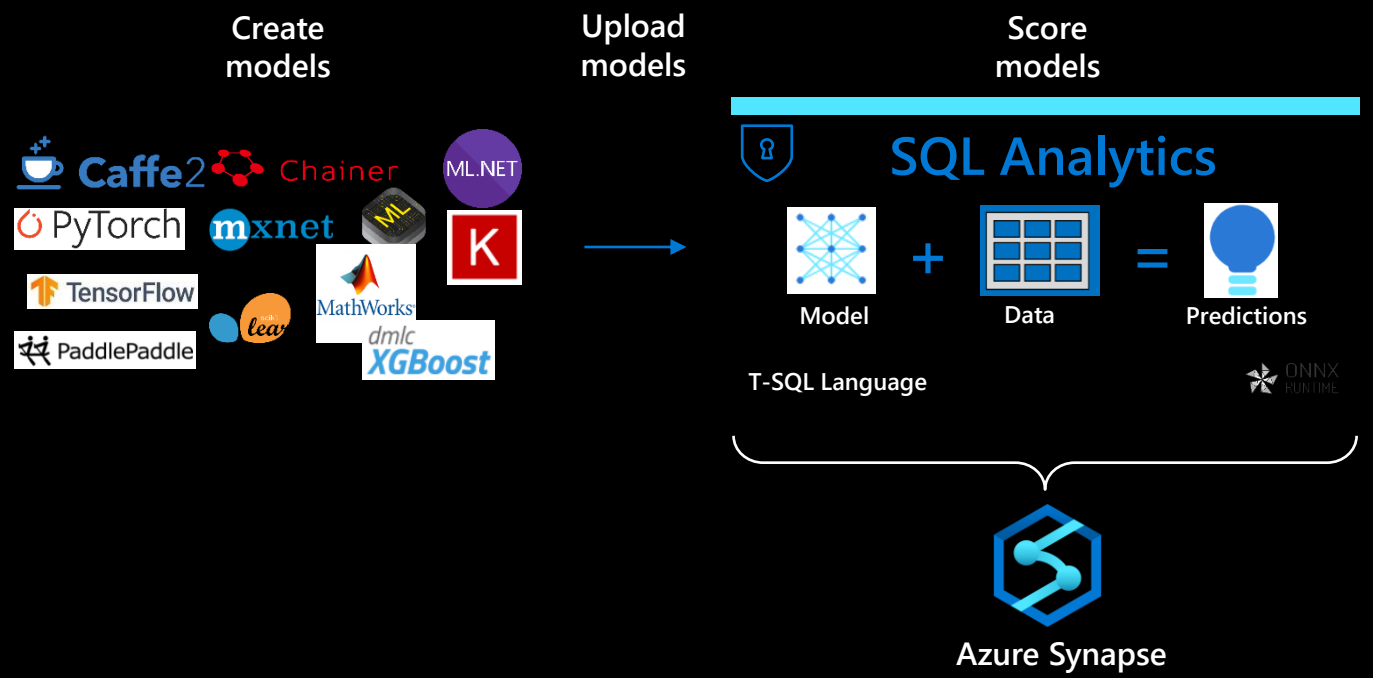
Azure Synapse + Power BI unlock the door to seamlessly incorporating AI & machine learning



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

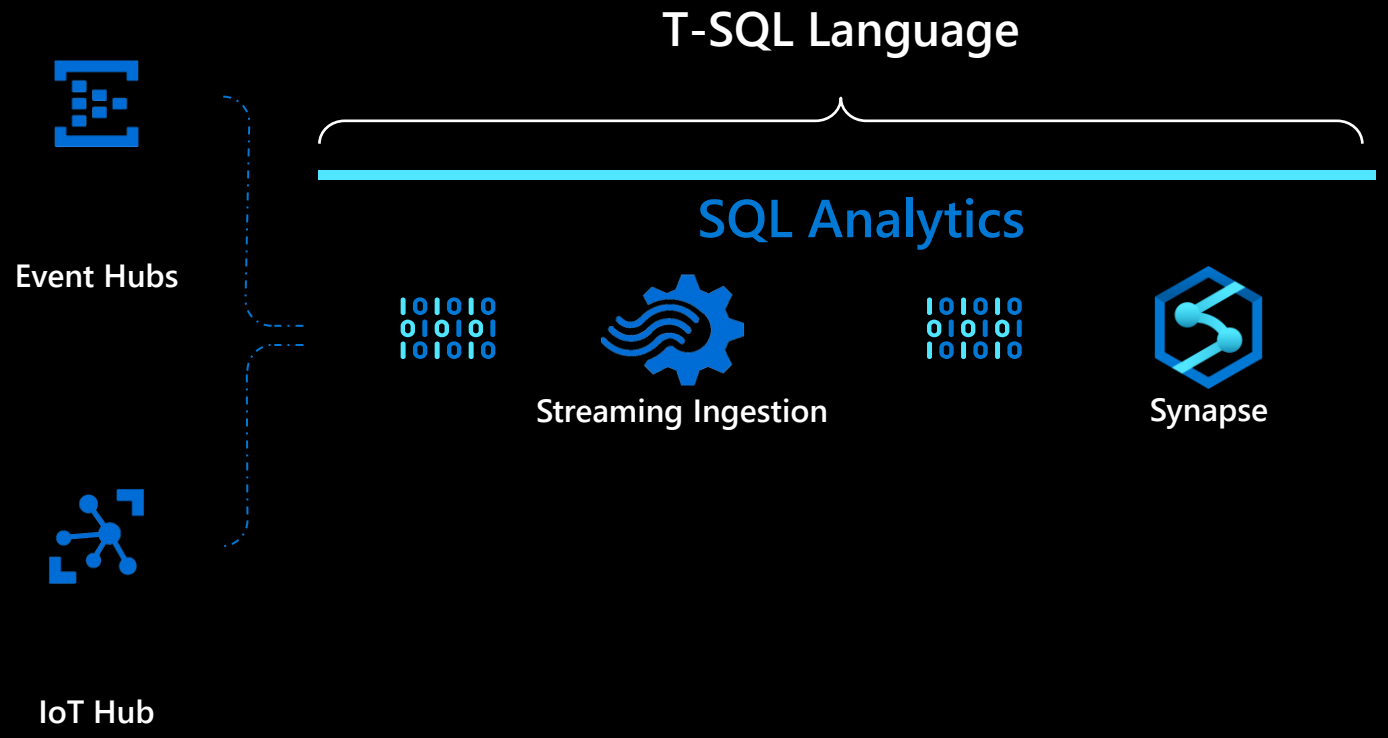



```
--T-SQL syntax for scoring data in SQL DW
SELECT d.*, p.Score
FROM PREDICT(MODEL = @onnx_model, DATA = dbo.mytable AS d)
WITH (Score float) AS p;
```

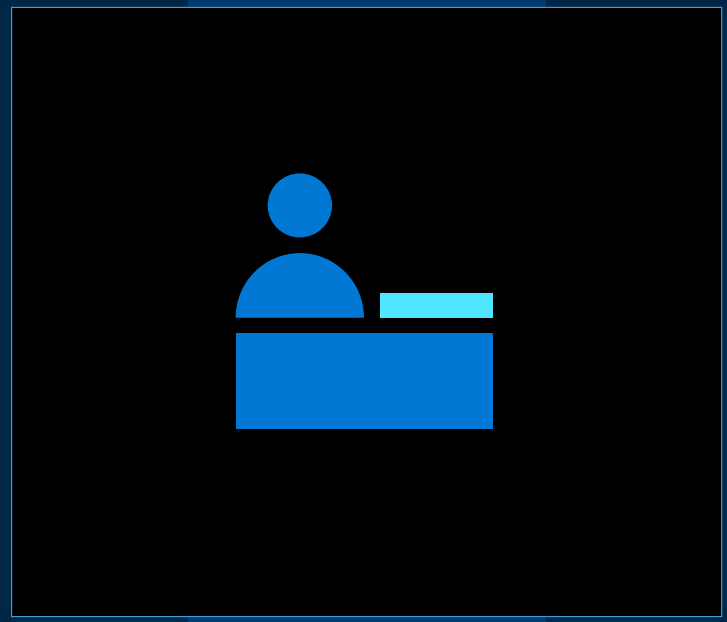
Heterogenous Data Preparation & Ingestion

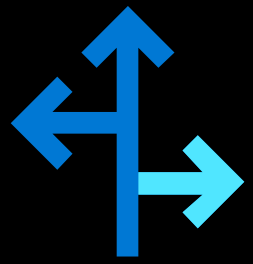
Native SQL Streaming

- High throughput ingestion (up to 200MB/sec)
- Delivery latencies in seconds
- Ingestion throughput scales with compute scale
- Analytics capabilities (SQL-based queries for joins, aggregations, filters)



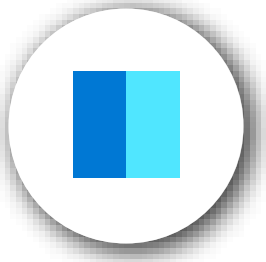
Workload Management





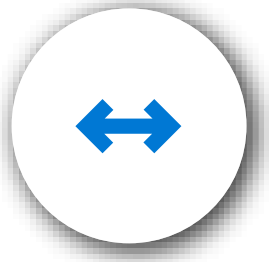
Prioritize your workloads using workload management

Workload Importance

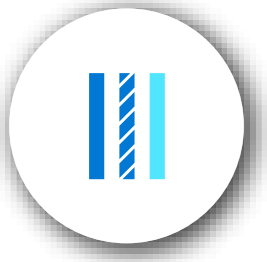


Intra-Cluster Isolation

Workload Classification

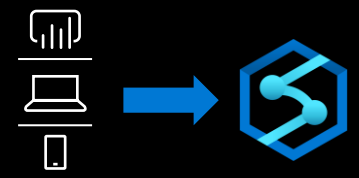


Elasticity

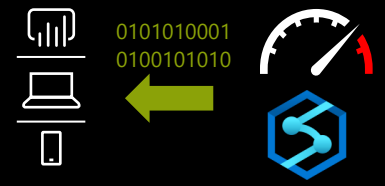


Multi-Cluster Isolation

Result-set caching

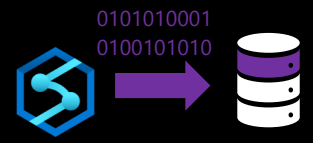


1 Client sends query to SQL pool

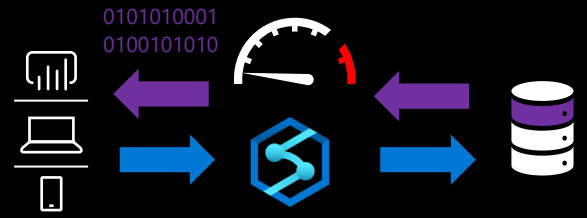


2 Query is processed using compute nodes which pull data from remote storage, process query and output back to client app

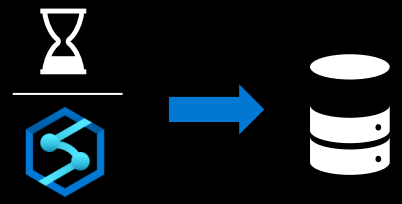
+



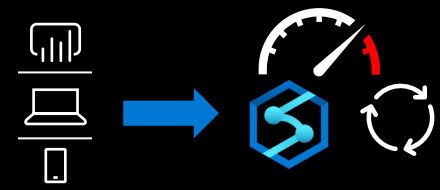
Query results are cached in remote storage so subsequent requests can be served immediately



3 Subsequent executions for the same query bypass compute nodes and can be fetched instantly from persistent cache in remote storage



4 Remote storage cache is evicted regularly based on time, cache usage, and any modifications to underlying table data.

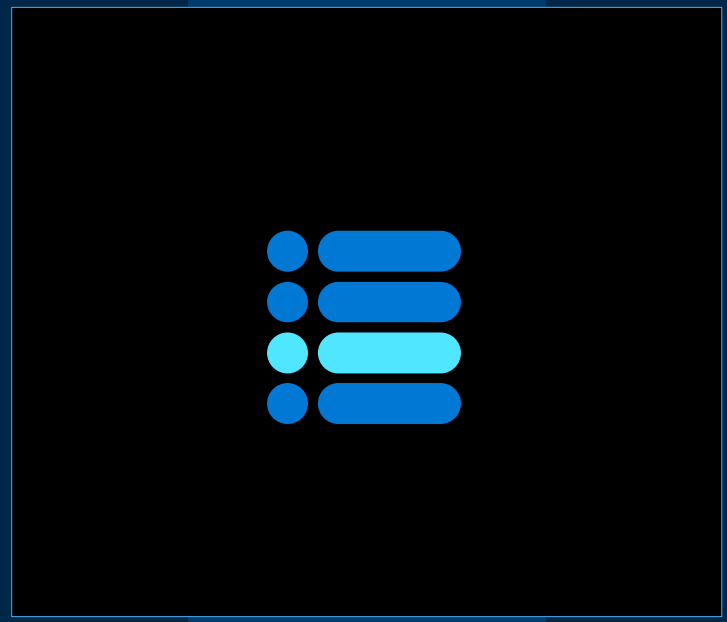


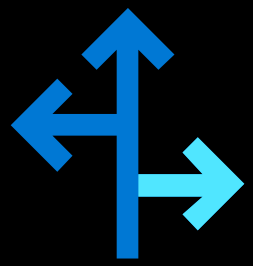
5 Cache will need to be regenerated if query results have been evicted from cache

Azure

Developer Productivity

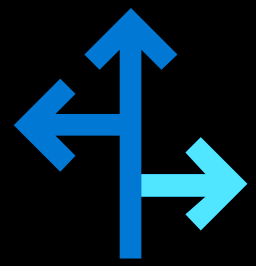
Agile development, CI/CD, DataOps



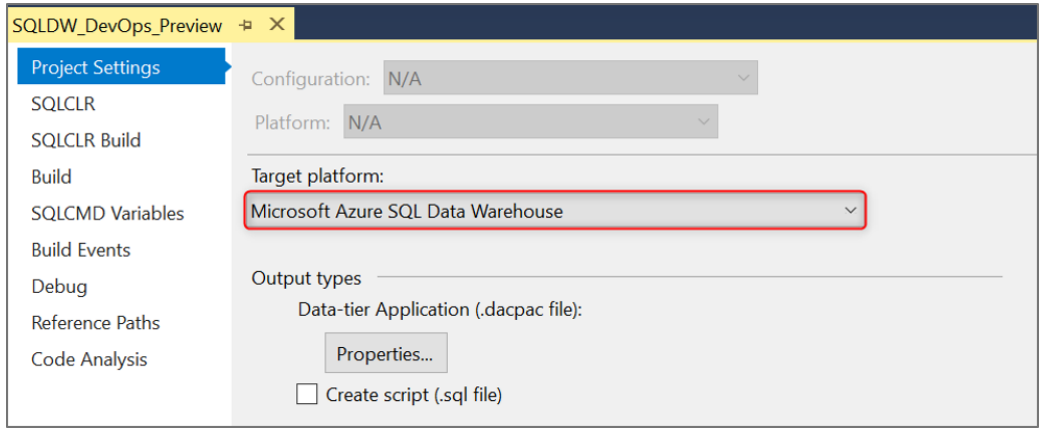


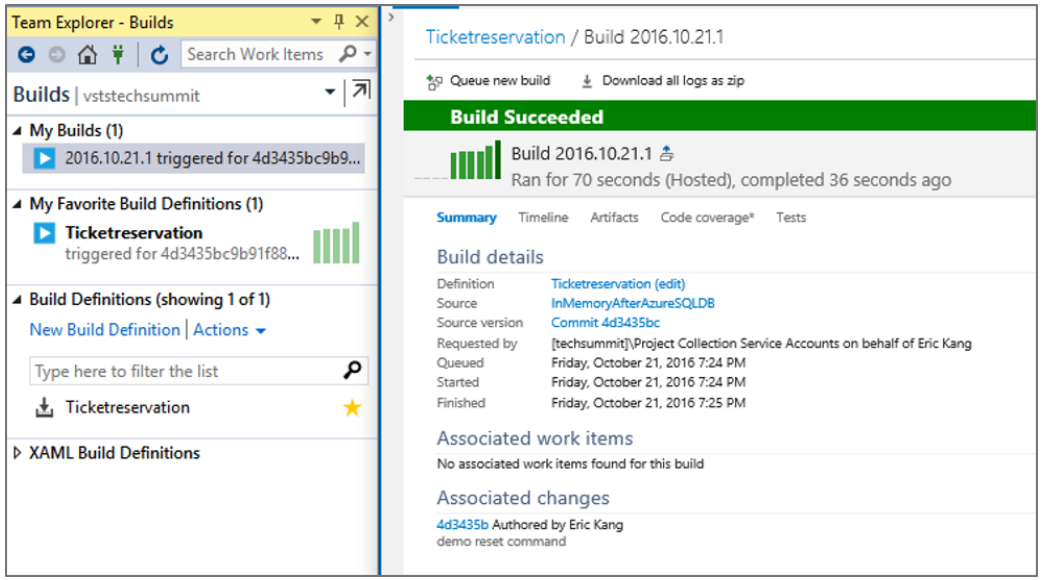
Use preferred tools for Azure Synapse Analytics development

- Familiar SQL experience with SQL Server Management Studio
- Track, apply, and deploy changes with Azure DevOps in Visual Studio
- Cross platform functionality with Azure Data Studio and VSCode

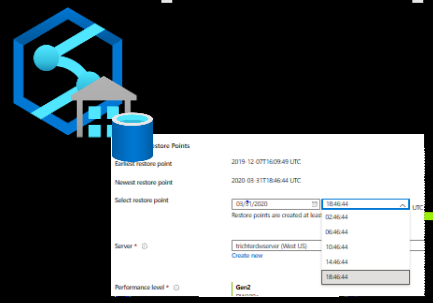


Use preferred tools for Azure Synapse Analytics development





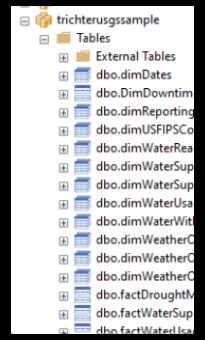
DevOps: Rapidly provision non-Prod environments

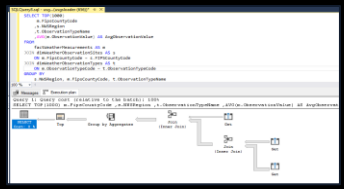


Select a restore point or create a new one from your source data warehouse

Recover using that restore point to a new data warehouse in your non prod environment

Configure developer/Test/UAT access





Start using your new dev/test/UAT environment

Run environments with smaller DWU to manage costs

Delete environments when you are done with them

Pause environments when you are not using them

Azure Advisor recommendations

Suboptimal Table Distribution

Reduce data movement by replicating tables

Data Skew

Choose new hash-distribution key

Slowest distribution limits performance

Cache Misses

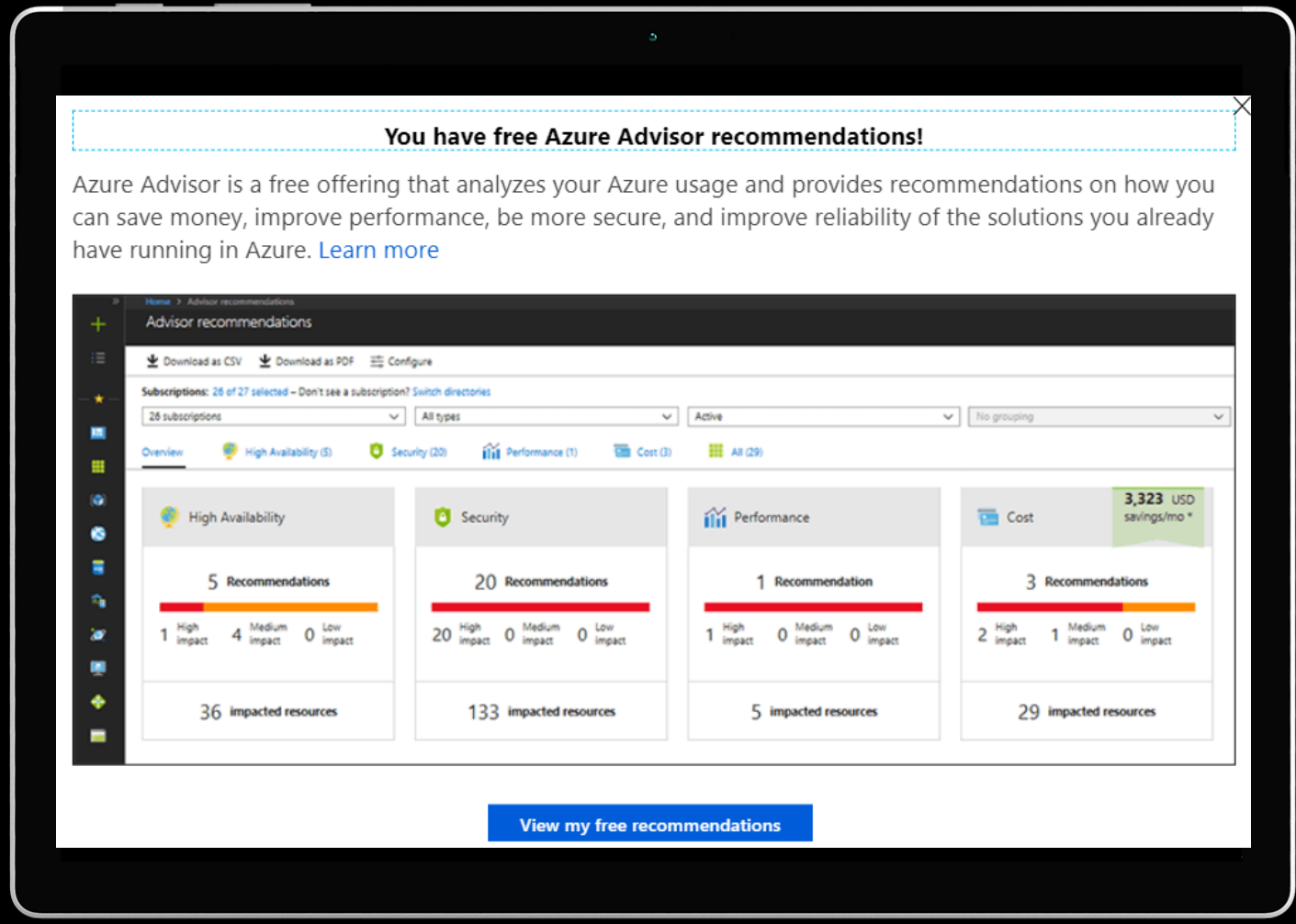
Provision additional capacity

Tempdb Contention

Scale or update user resource class

Suboptimal Plan Selection

Create or update table statistics



Build Data Warehouse on Demand

Speed to value

Azure Synapse SQL on-demand scenarios

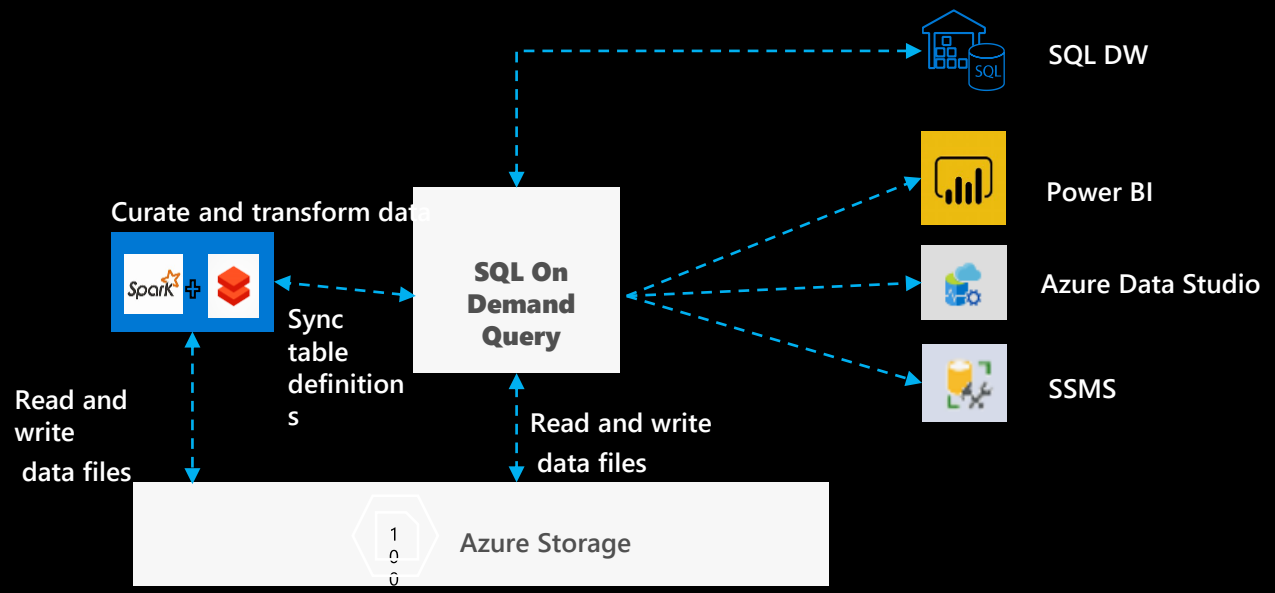
Discovery and exploration	What's in this file? How many rows are there? What's the max value? **SQL On-demand reduces data lake exploration to the right-click!**
Logical Data Warehouse	How to create a data warehouse? **Model raw files as virtual tables & views, implement security and use any SQL based tools to analyze data**
Data transformation	How to convert CSVs to Parquet quickly? How to transform the raw data? **Use the full power of T-SQL to transform the data in the data lake**

SQL On-Demand

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

Benefits

- Use SQL to work with files on Azure storage
  - Directly query files on Azure storage using T-SQL
  - Logical Data Warehouse on top of Azure storage
  - Easy data transformation of Azure storage files
- Supports any tool or library that uses T-SQL to query data
- Automatically synchronize tables from Spark
- Serverless
  - No infrastructure/upfront cost, no resource reservation
  - Pay only for query execution (per data processed)



Microsoft

Azure Synapse Analytics

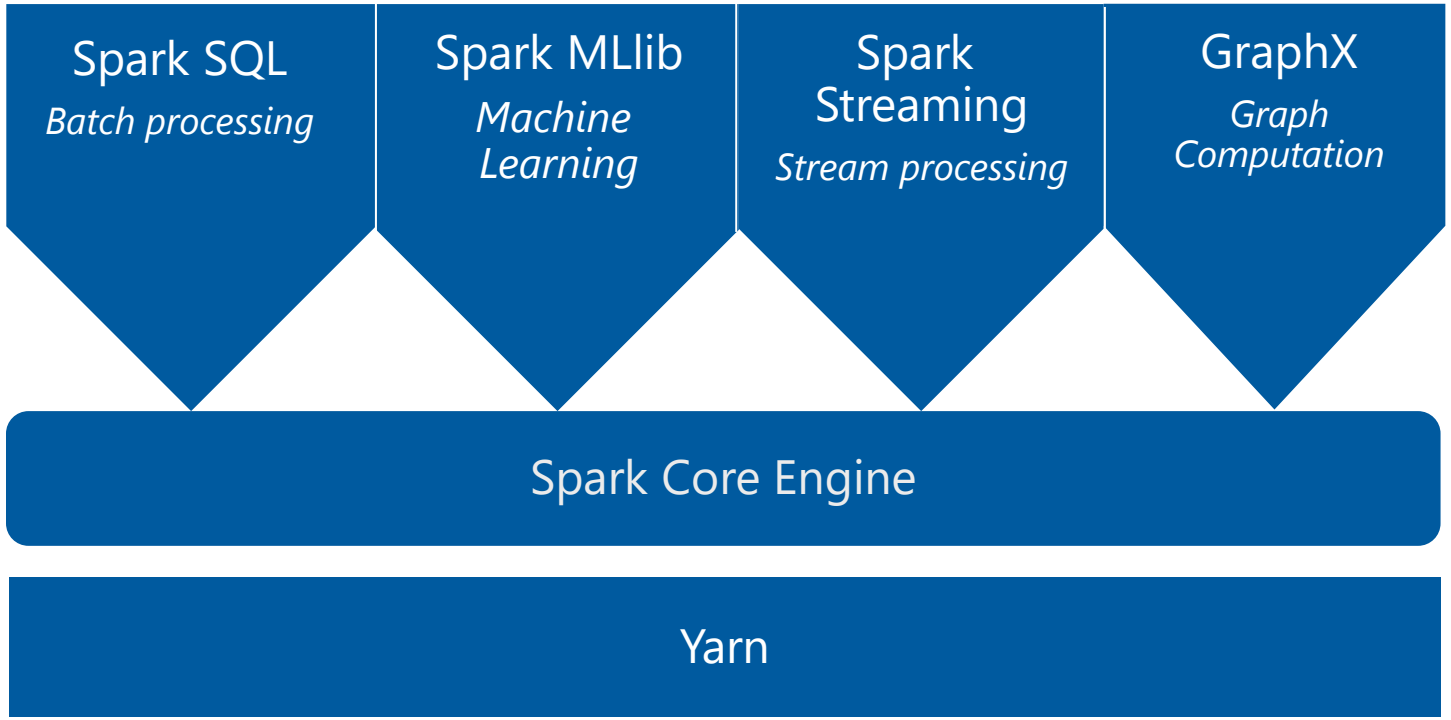
Spark

Apache Spark

An unified, open source, parallel, data processing framework for Big Data Analytics

Spark Unifies:

- Batch Processing
- Interactive SQL
- Real-time processing
- Machine Learning
- Deep Learning
- Graph Processing



http://spark.apache.org

Automatic syncing of Spark tables

Overview

Tables created in Spark pool are automatically created as external tables that reference external files in your SQL serverless Logical Data Warehouse

Benefits

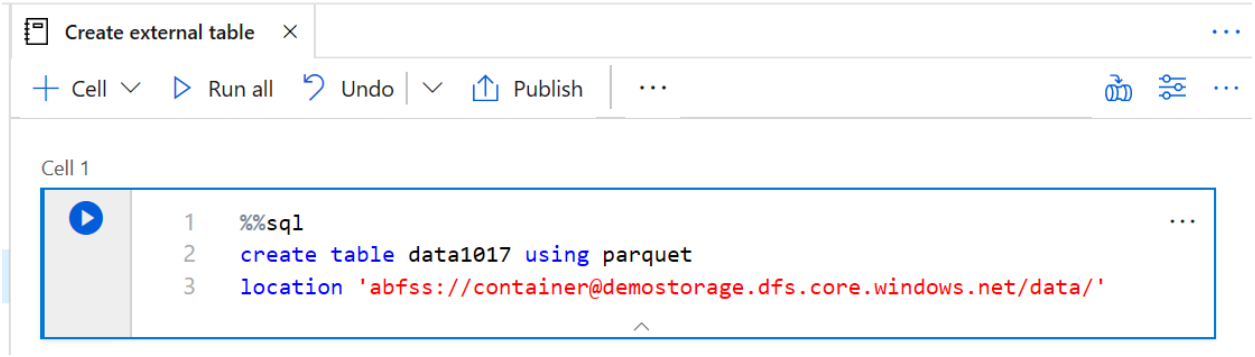
Tables designed using Spark languages are immediately available in SQL serverless.

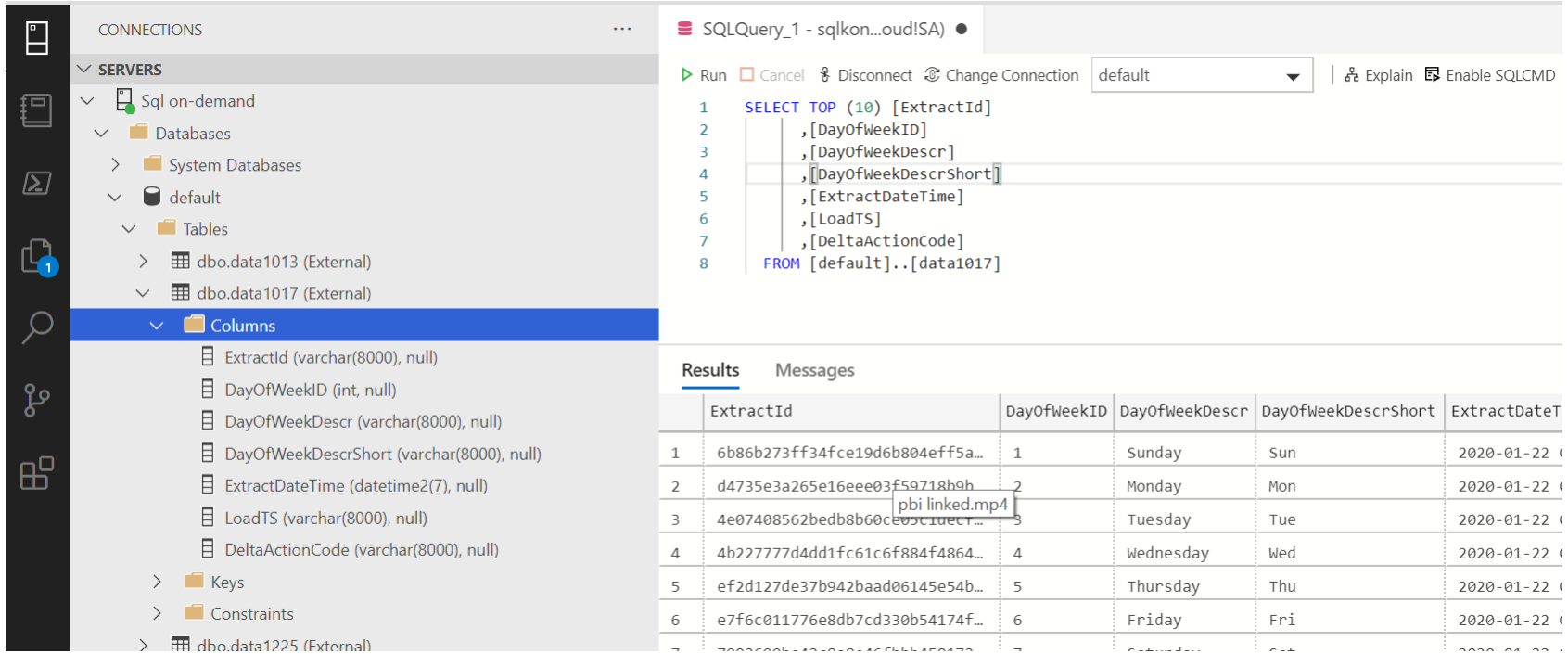
Schema definition matches original

Spark table updates are applied in SQL serverless

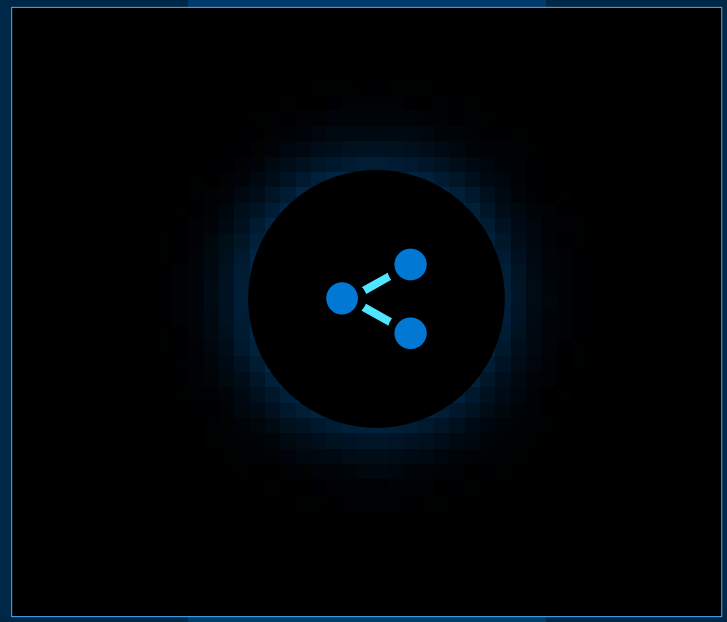
No need to manually create SQL tables that match Spark tables

Spark and SQL serverless tables references the same external files.

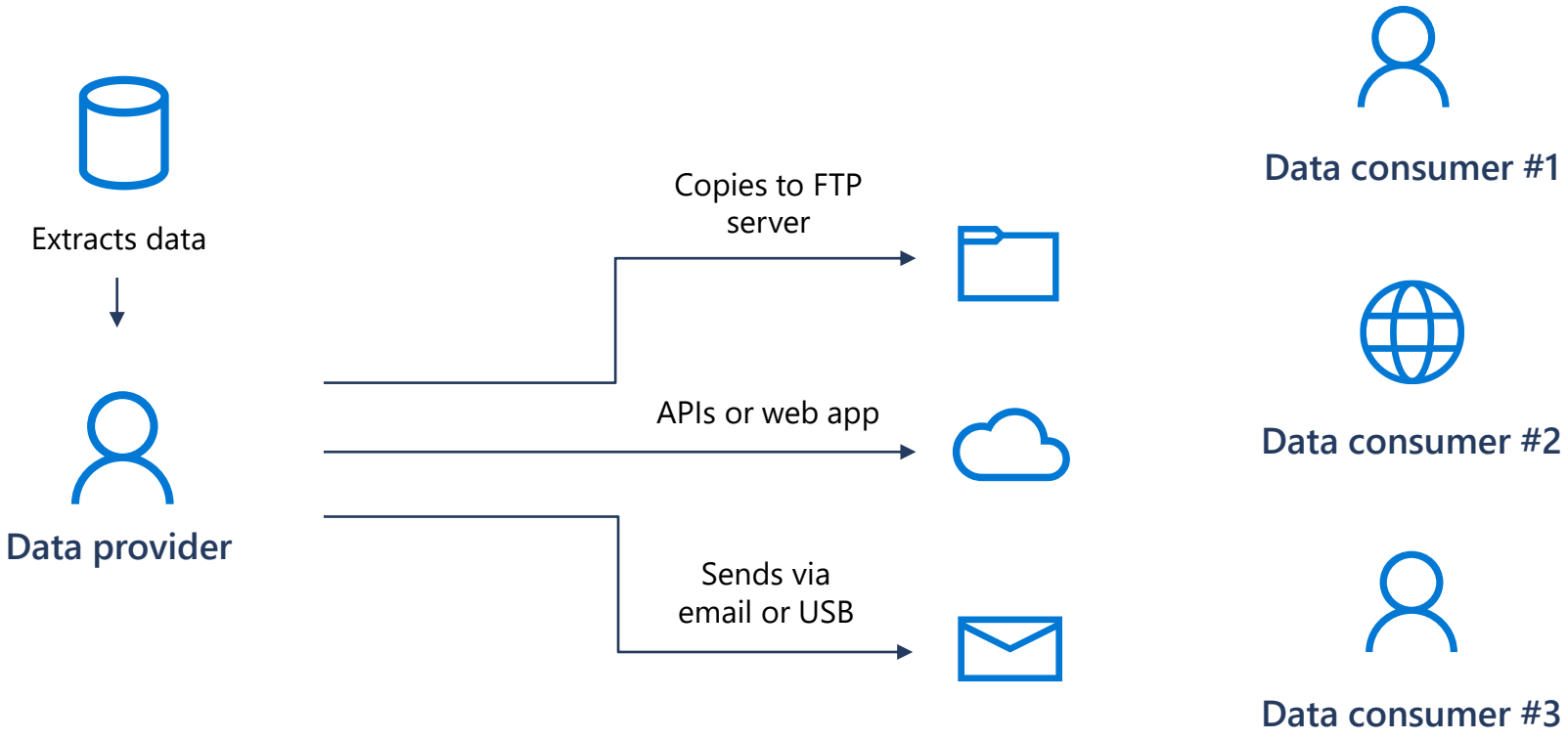




Azure Data Share



How data is shared today

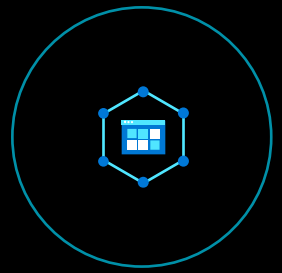


Difficult to manage, track, and not suitable for big data

Azure Data Share vision

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

Governance

Control data access governed by enterprise policies

Monetization

Charge for data or cost of data curation and access

Microsoft Azure

Be future ready | Build on your terms | Operate hybrid seamlessly | Trust your cloud