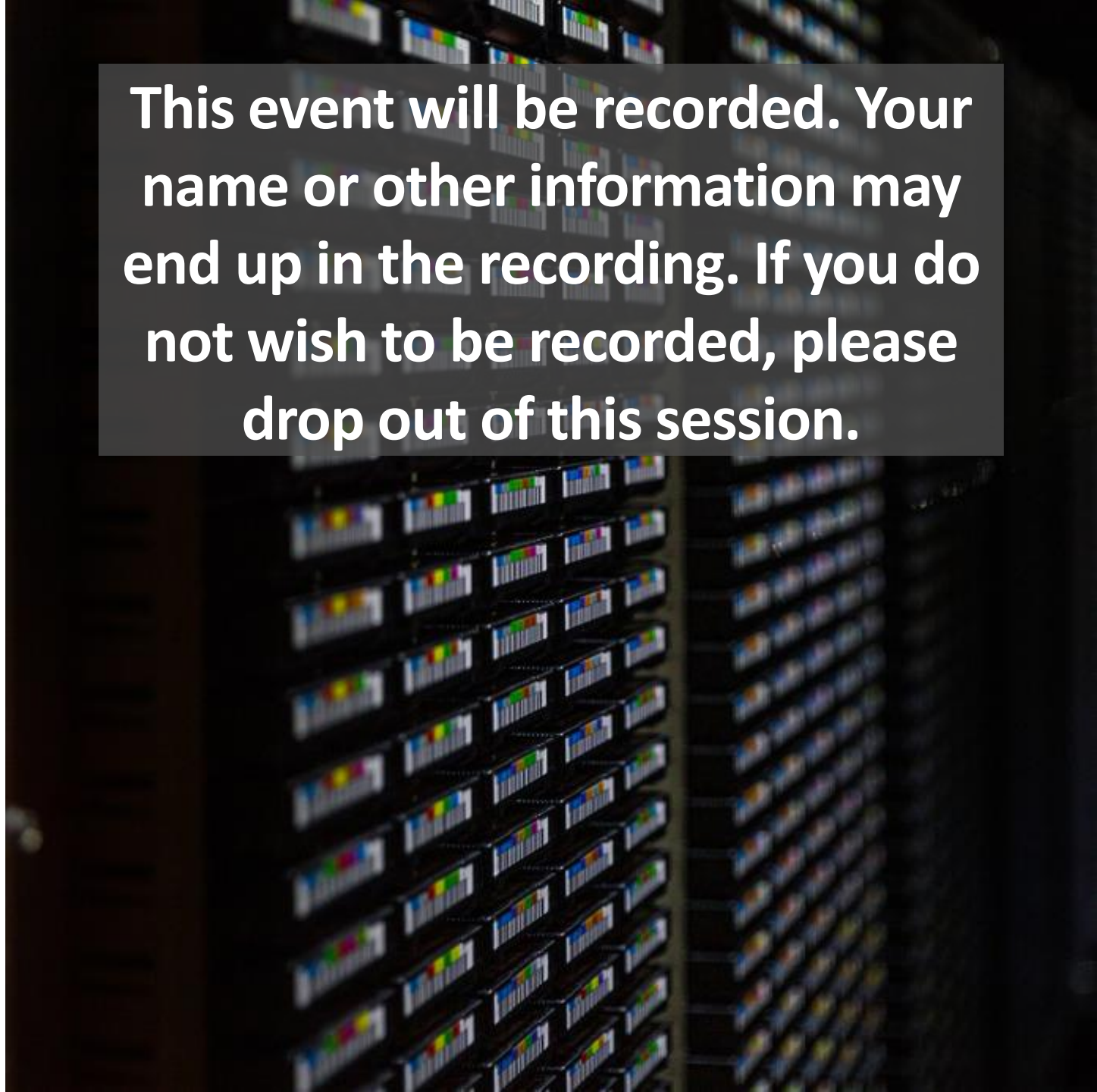
Microsoft

Welcome to the

SaaS Lab Program

Session 6

Innovation with Data & Analytics



Hello, meet your session presenters

Daphne Choong

Partner Technology Strategist

About: I strategize with Microsoft ISV Partners in the APAC region to build technologies in Azure Cloud. My passion is to enable partners in the region to innovate and be globally competitive.

daphnechoong@microsoft.com
https://www.linkedin.com/in/daphnecys/

Agenda

Data-Driven digital strategy

How to start Data & Analytics in the workplace

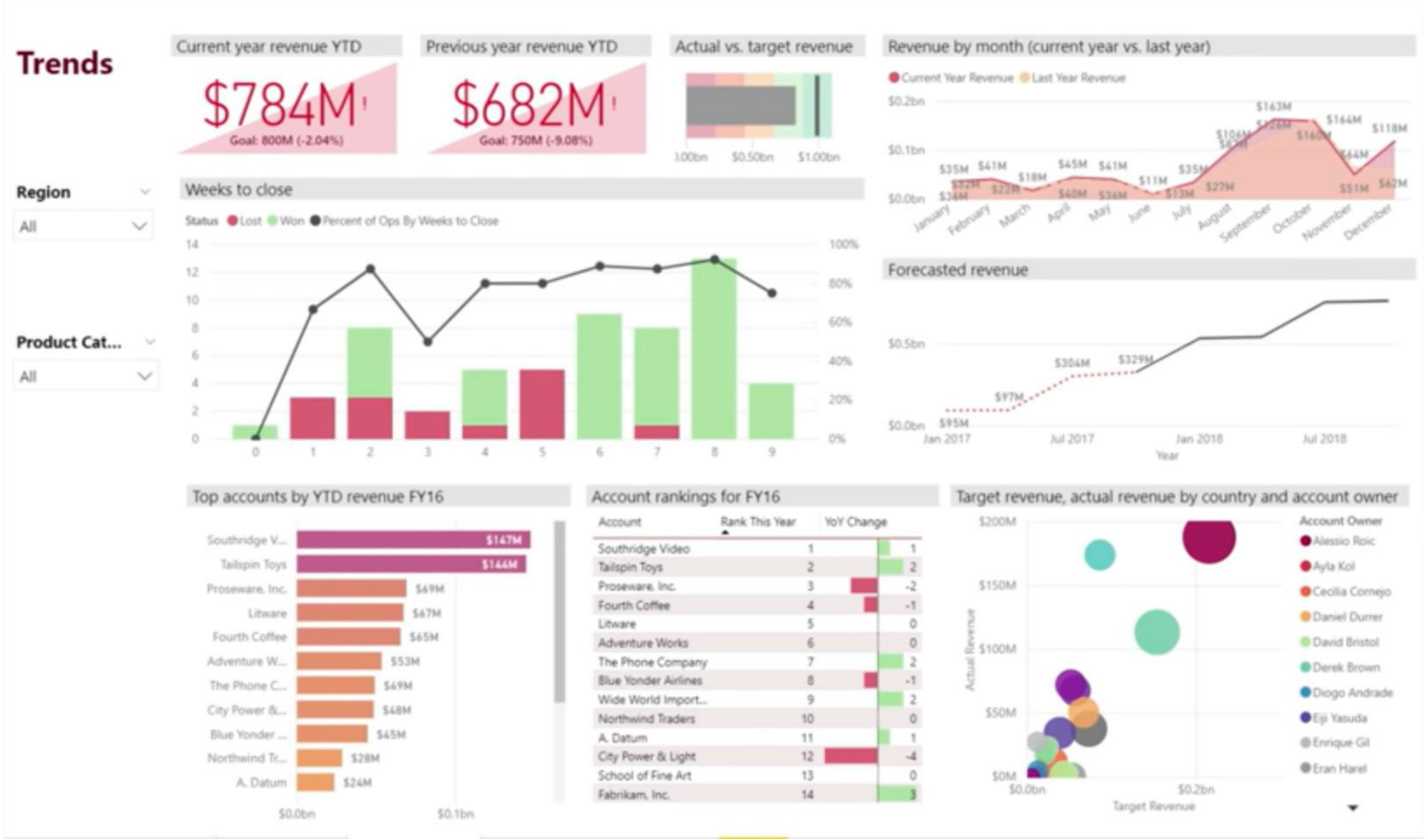
Modern Data Warehouse Architecture

Data DevOps

Azure Synapse Analytics

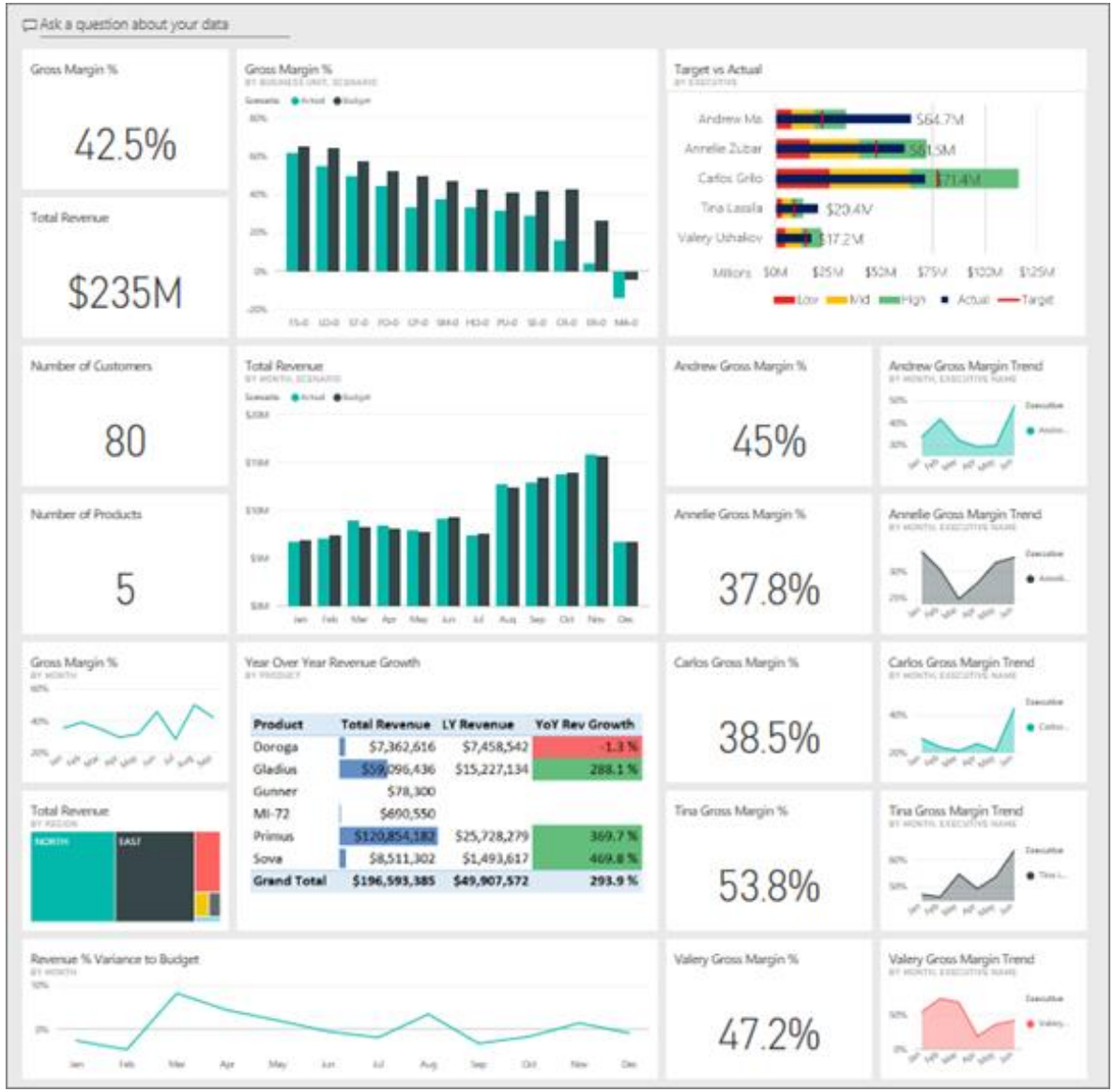
Azure Synapse Analytics & Power BI demo

Data-Driven Digital Strategy



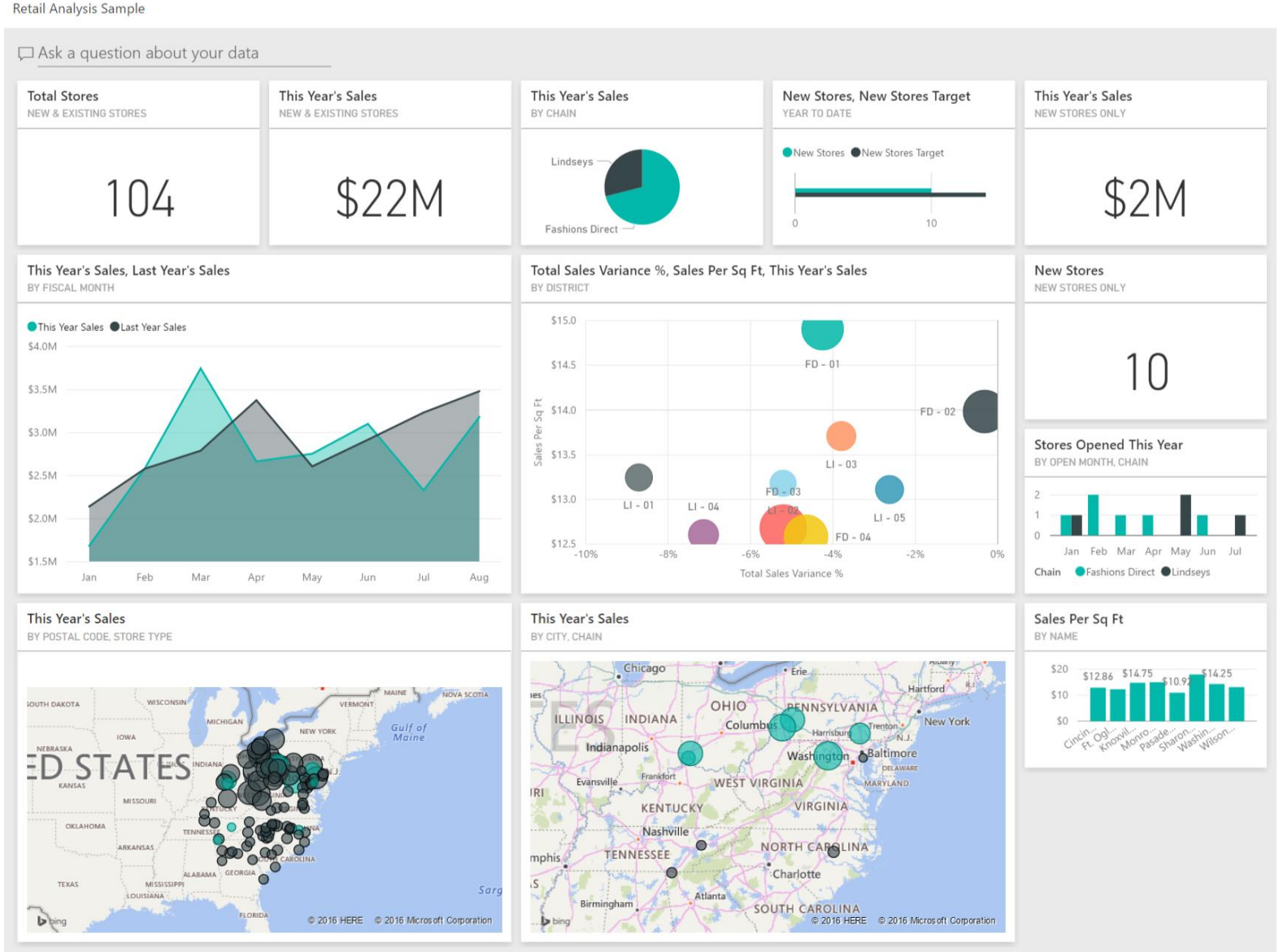
- Data tells a story
- Making decisions backed by data
- Respond to market and trend changes
- Gain customer insights
- Understanding our competitors
- Understanding our strengths and weaknesses
- Increase top-line and bottom-line growth

Example: Customer Profitability



- Factors impacting profitability
- Key Metrics
- Business unit managers
- Products
- Customers
- Gross margins

Retail Analysis Sample



Other examples:

- Sales & Marketing
- Supplier Quality Analysis
- IT Spend Analysis
- HR
- Opportunity Analysis
- Procurement Analysis

Problem Statement: We have a lot of data, we don’t know what to do with it

- Understand your data
- Find out where they are
- Collect them into one place (Azure Synapse)
- Build dashboards
- Revisit – clean / collect more data

How to start Data Analytics in the workplace

Understand business goals

Observe / Ask Business Questions

Get close to the user

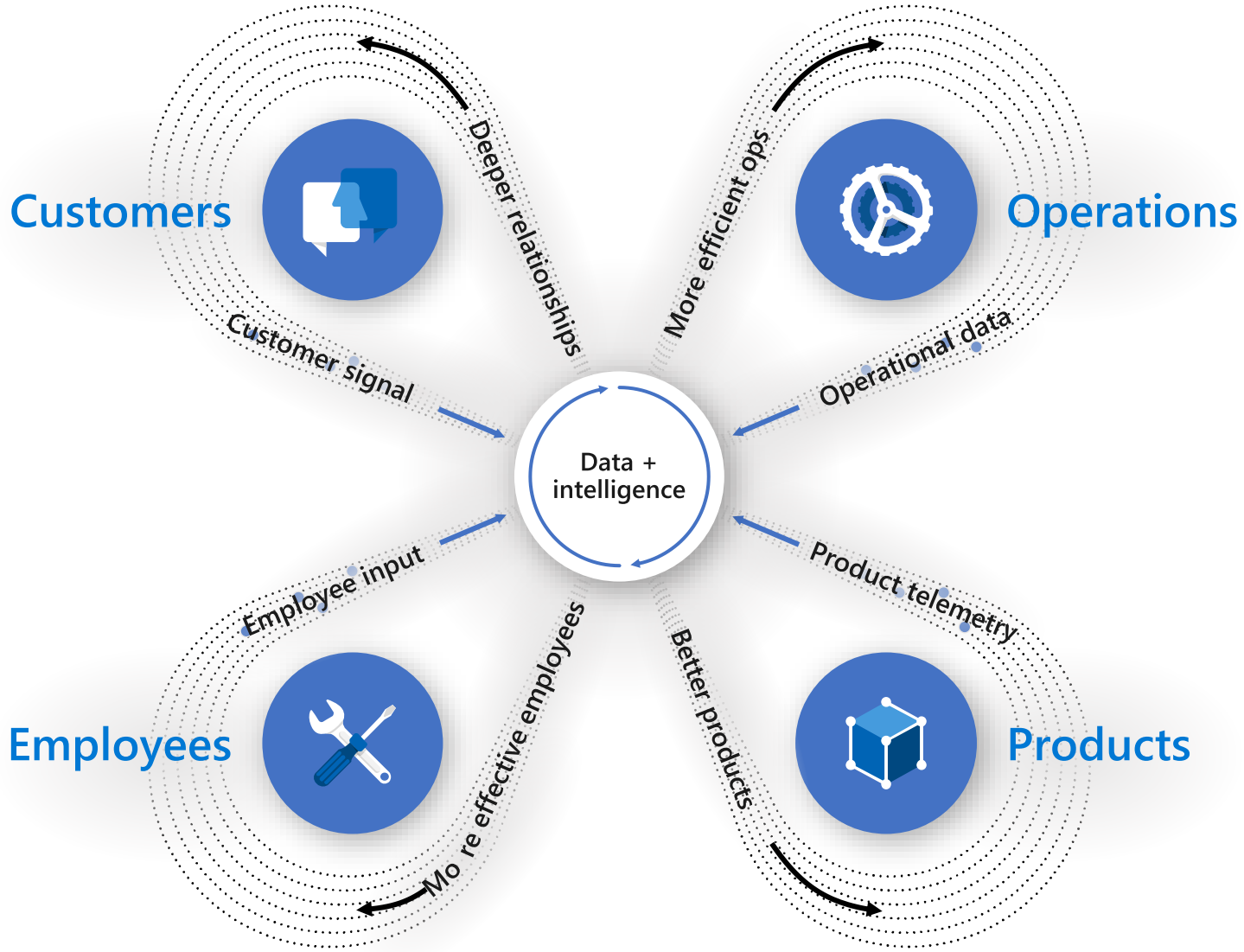
Research the tools

Practice getting data, presenting data to the user, probe more questions (gently), repeat

Success: when users ask their own Business Questions

The digital feedback loop

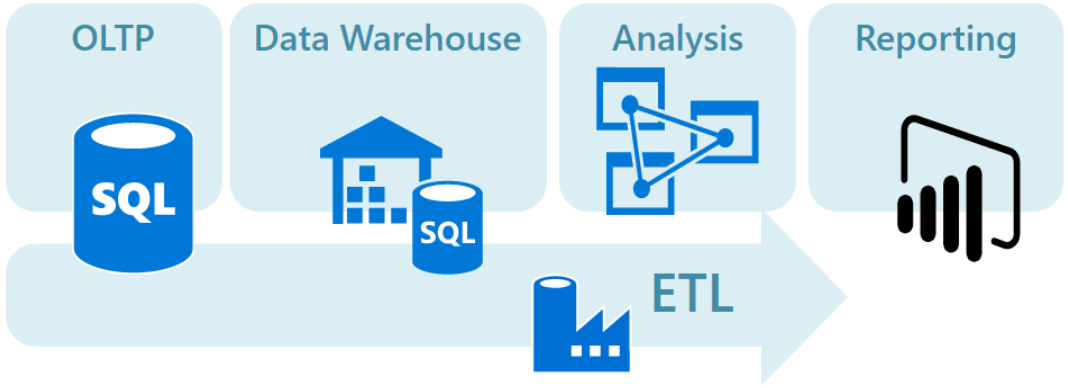
1. Data: Capture digital signal across business
2. Insight: Connect and synthesize data
3. Action: Improve business outcomes



Modern Data Warehouse Architecture

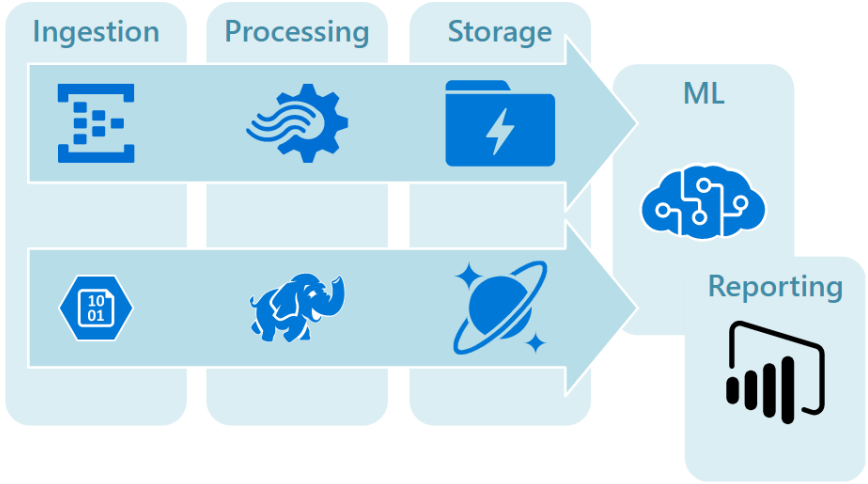
Azure Data Architecture Guide

Traditional RDBMS workloads vs Big Data Solutions



Traditional RDBMS workloads
Include OLTP & OLAP
Predefined schema and constraints.
Consolidated into a data warehouse

Big Data Solutions
Data too large or complex for traditional DB systems
Data processed in batch or in real time
Non-relational data, key-value data, JSON documents or time series data
NoSQL – "Not only SQL"

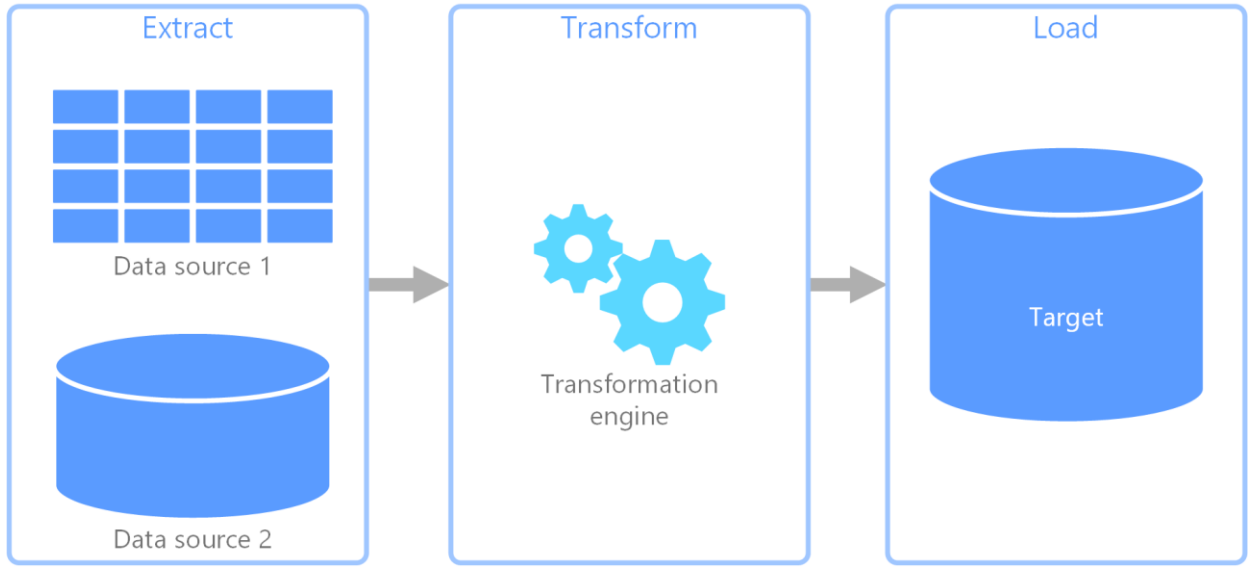


Relational Data

[**E**xtract] Data comes in multiple sources, multiple formats

[**T**ransform] Need to shape & clean

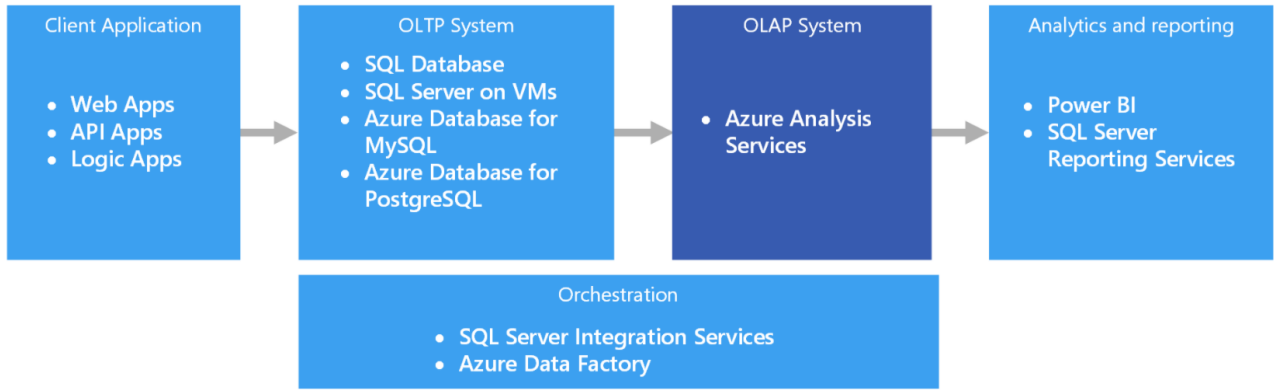
[**L**oad] Moving data to destination – a data warehouse – Hadoop cluster (Hive or Spark) or Azure Synapse Analytics



OLTP vs OLAP

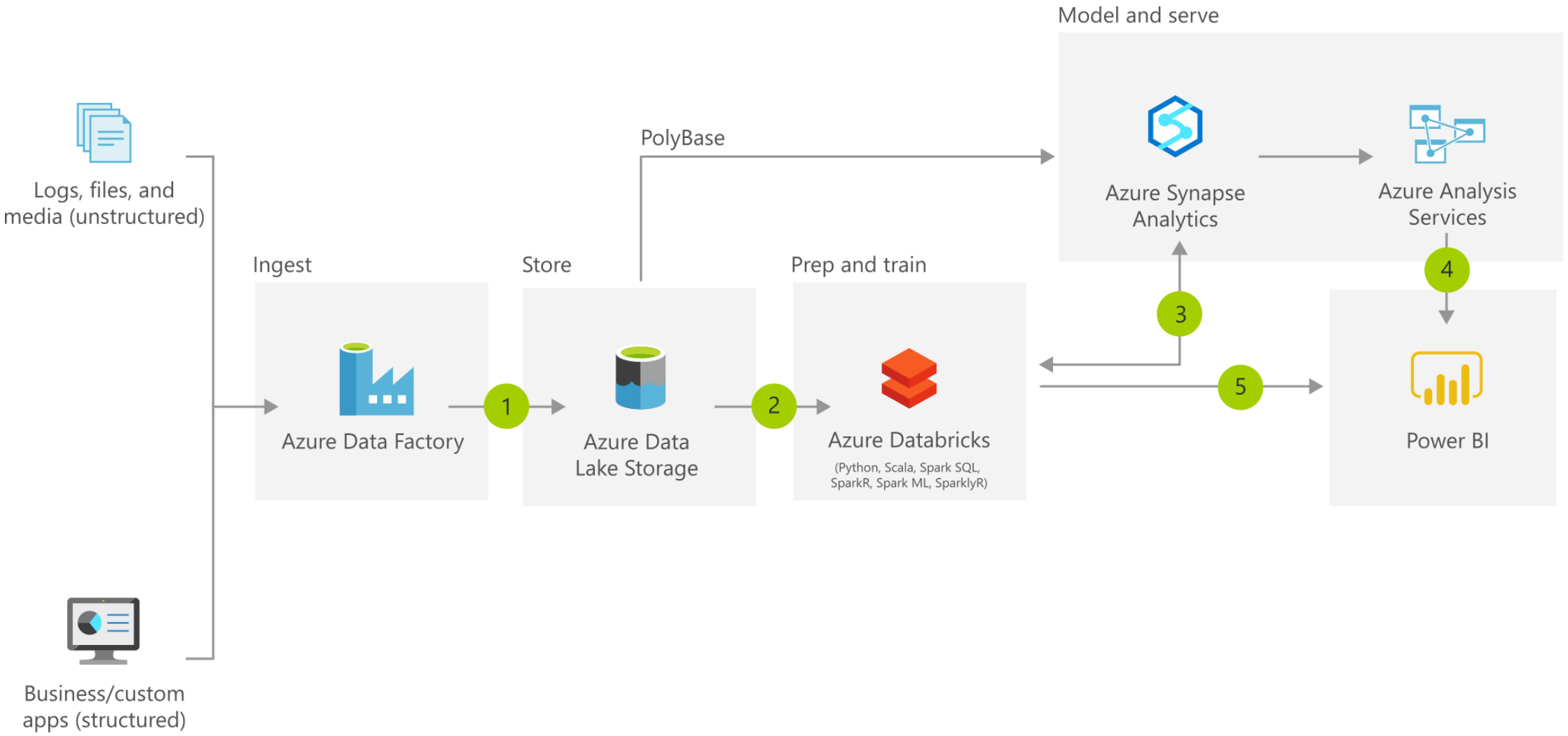
Online Transaction Processing vs Online Analytical Processing

OLTP	OLAP
Atomic and consistent	Optimized for heavy read, low write workloads
Transactional	Not transactional
Locking strategies	No locking strategies
Not good for aggregation	Good for aggregation, calculations, time-oriented calculations



Modern Data Warehouse Architecture

[Modern Data Warehouse Architecture - Azure Solution Ideas | Microsoft Docs]



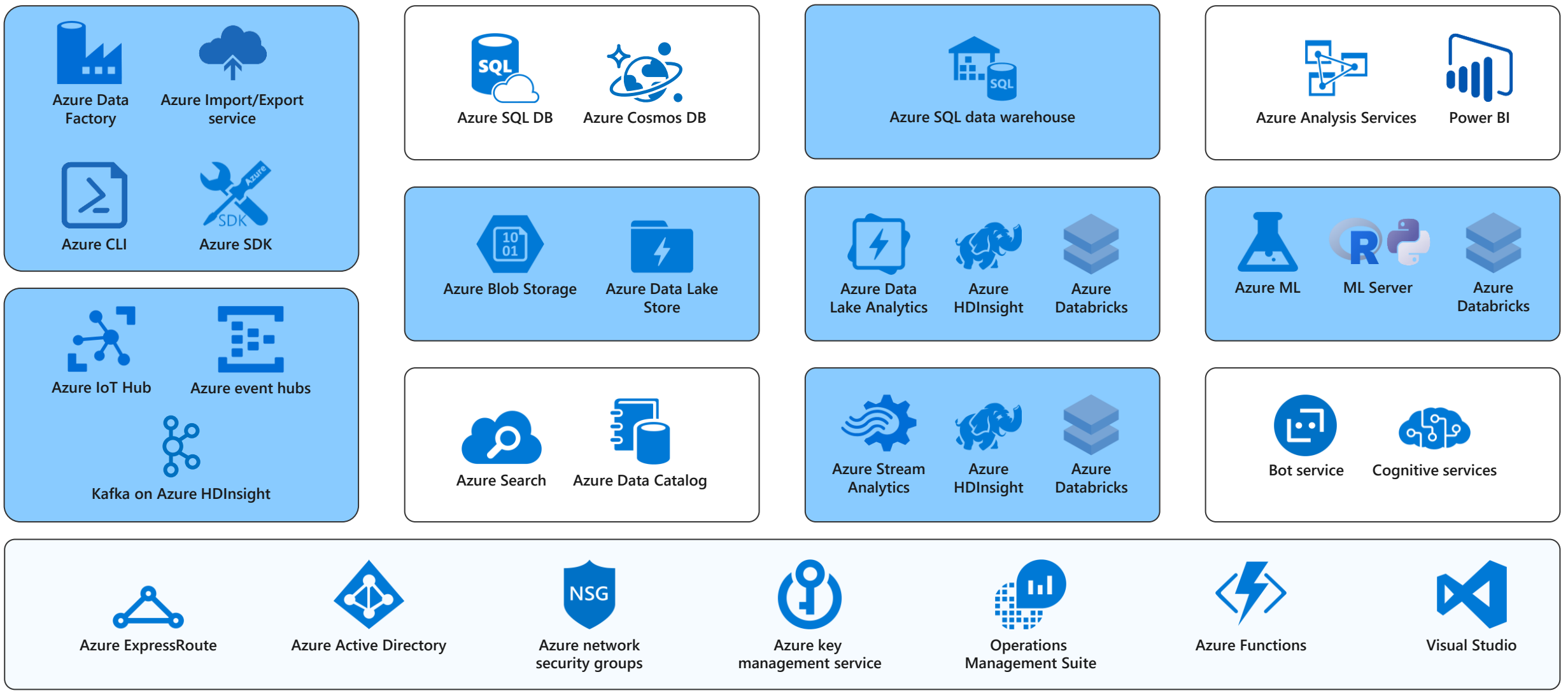
Non-relational data stores

Non-relational data and NoSQL (Not Only SQL)

- Document data stores – Azure Cosmos DB
- Columnar data stores – Azure Cosmos DB Cassandra API, Hbase in HDInsight
- Key/value data stores – Azure Cosmos DB , Cache for Redis, Table Storage
- Graph data stores – Azure Cosmos DB Graph API
- Time Series data stores – Azure Time Series Insights, OpenTSDB with Hbase on HDInsight
- Object data stores – Azure Blob Storage, Data Lake Store, File Storage
- External index data stores – Azure Search

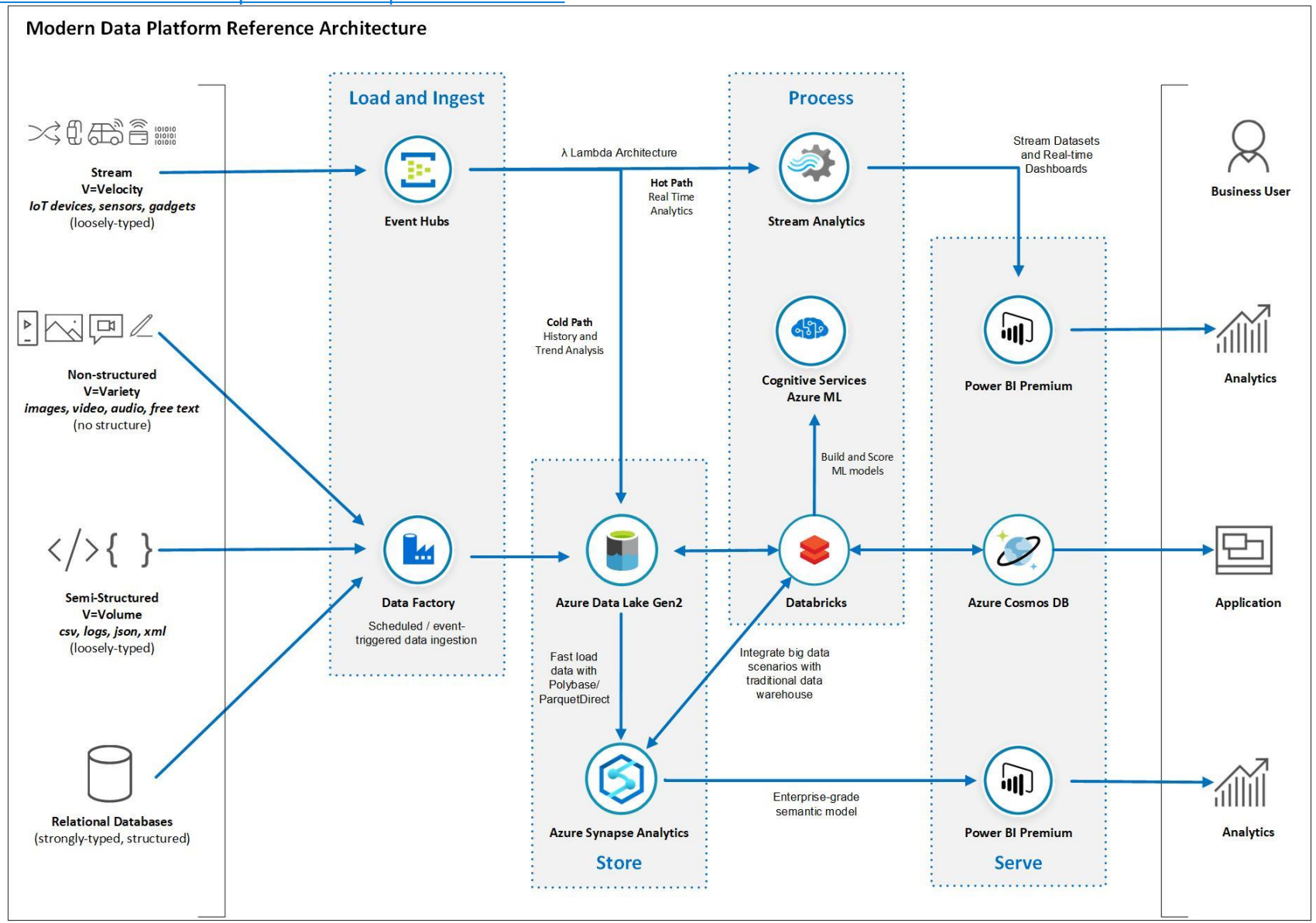
Understanding the Azure portfolio for Big Data & Advanced Analytics

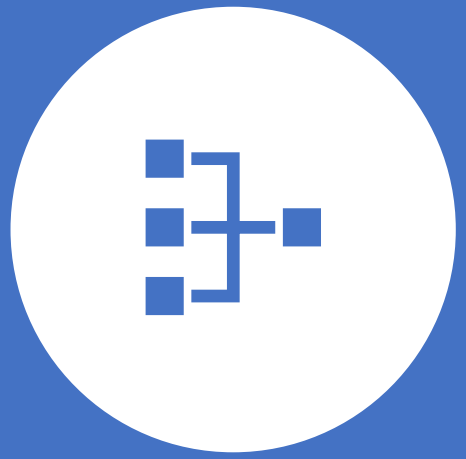
The Azure **big** data landscape



Big Data Analytics – Real Time Processing

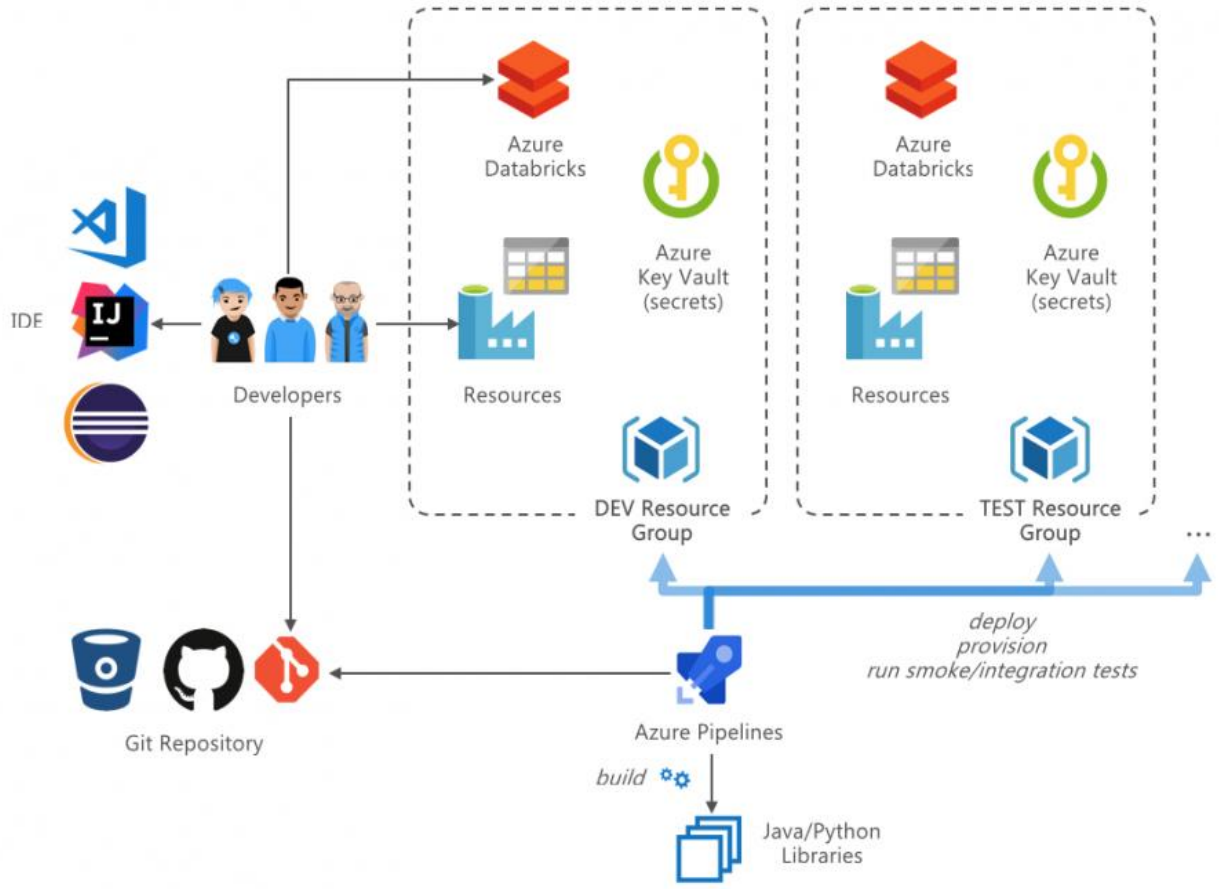
[Azure data platform end-to-end - Azure Example Scenarios | Microsoft Docs]





Data DevOps

Data Devops

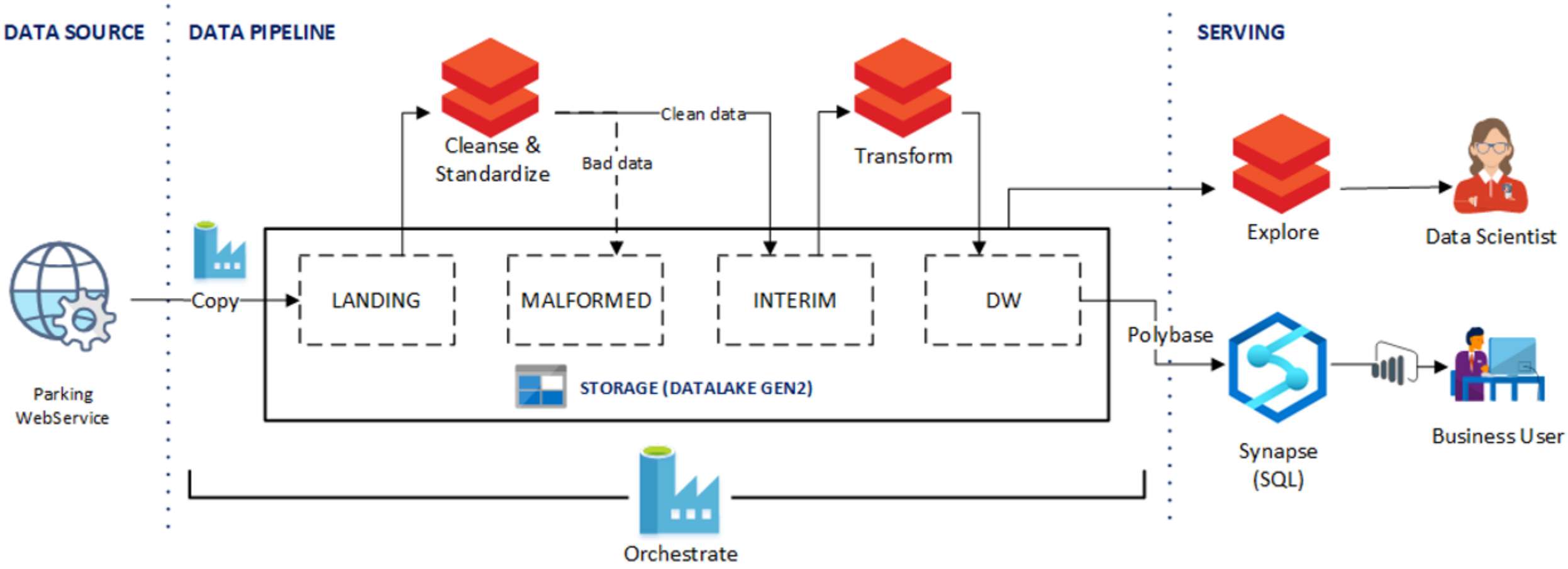


- Integrate the deployment of an entire Azure environment (comprising for example storage accounts, data factories and databases) within a single pipeline, or a coherent set of interdependent pipelines
- Fully provision environments including resources and notebooks
- Manage service identities as well as credentials
- Run integration and smoke tests

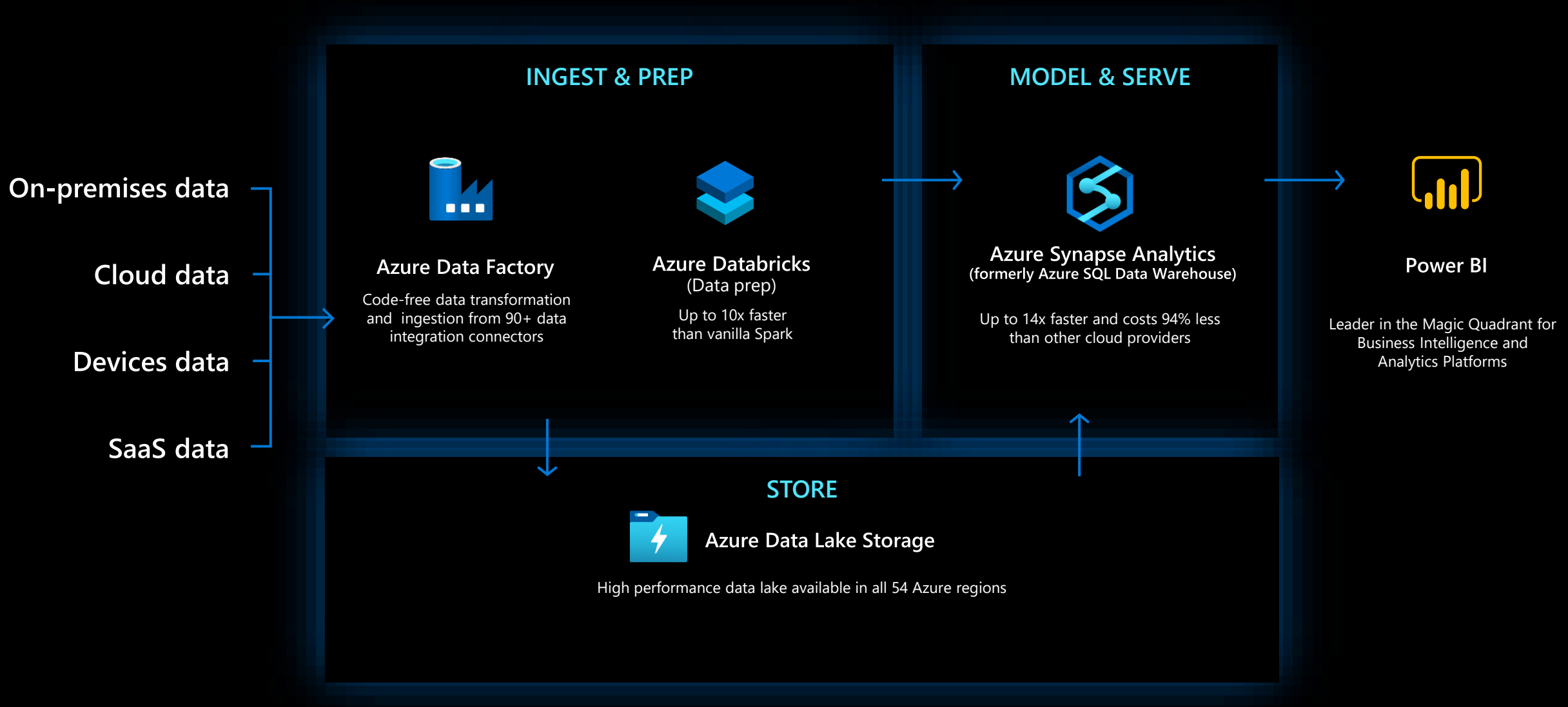
DataOps for the modern data warehouse

- Automate collection of data from various sources
- Reduce risk of errors
- Infrastructure as Code
- CI/CI, deployment gates
- Pipeline as Code
- Integration tests
- Row-level / object-level security
- Monitoring
- Centralized configuration in Azure Key Vault

Example architecture of a data pipeline



Azure Analytics



Analytics in Azure is simply unmatched

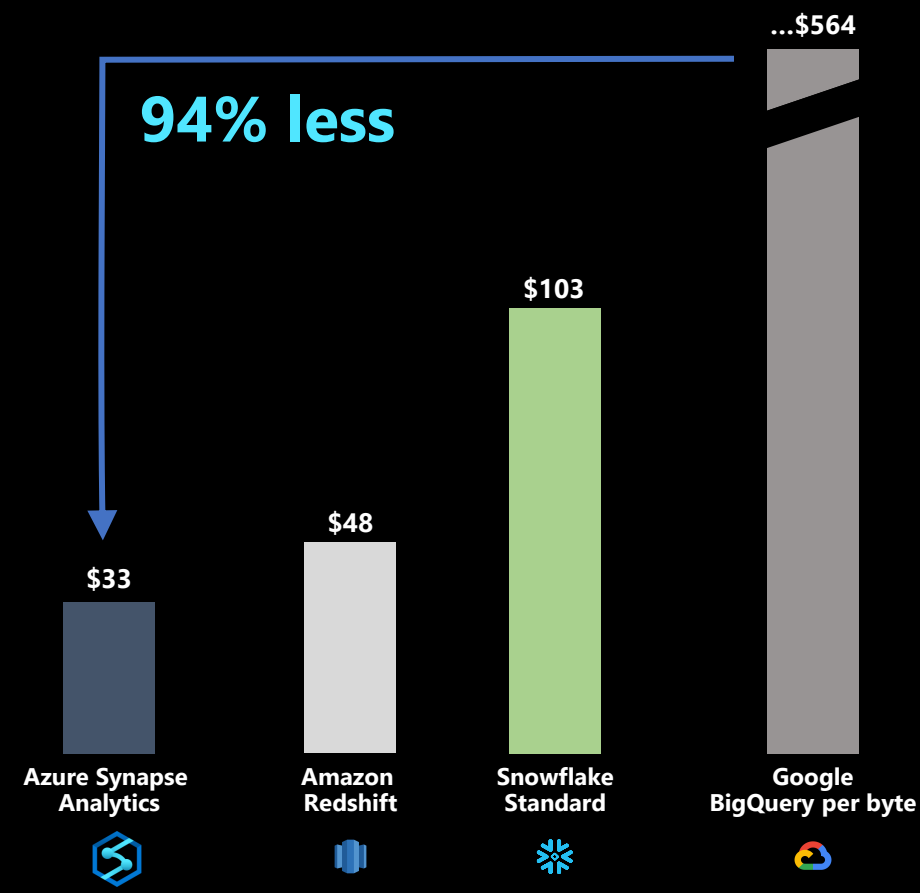
Analytics in Azure is up **14x faster and costs 94% less** than other cloud providers

Azure offers the most **comprehensive security and privacy** capabilities on the market

Azure Analytics + Power BI deliver **insights to all**

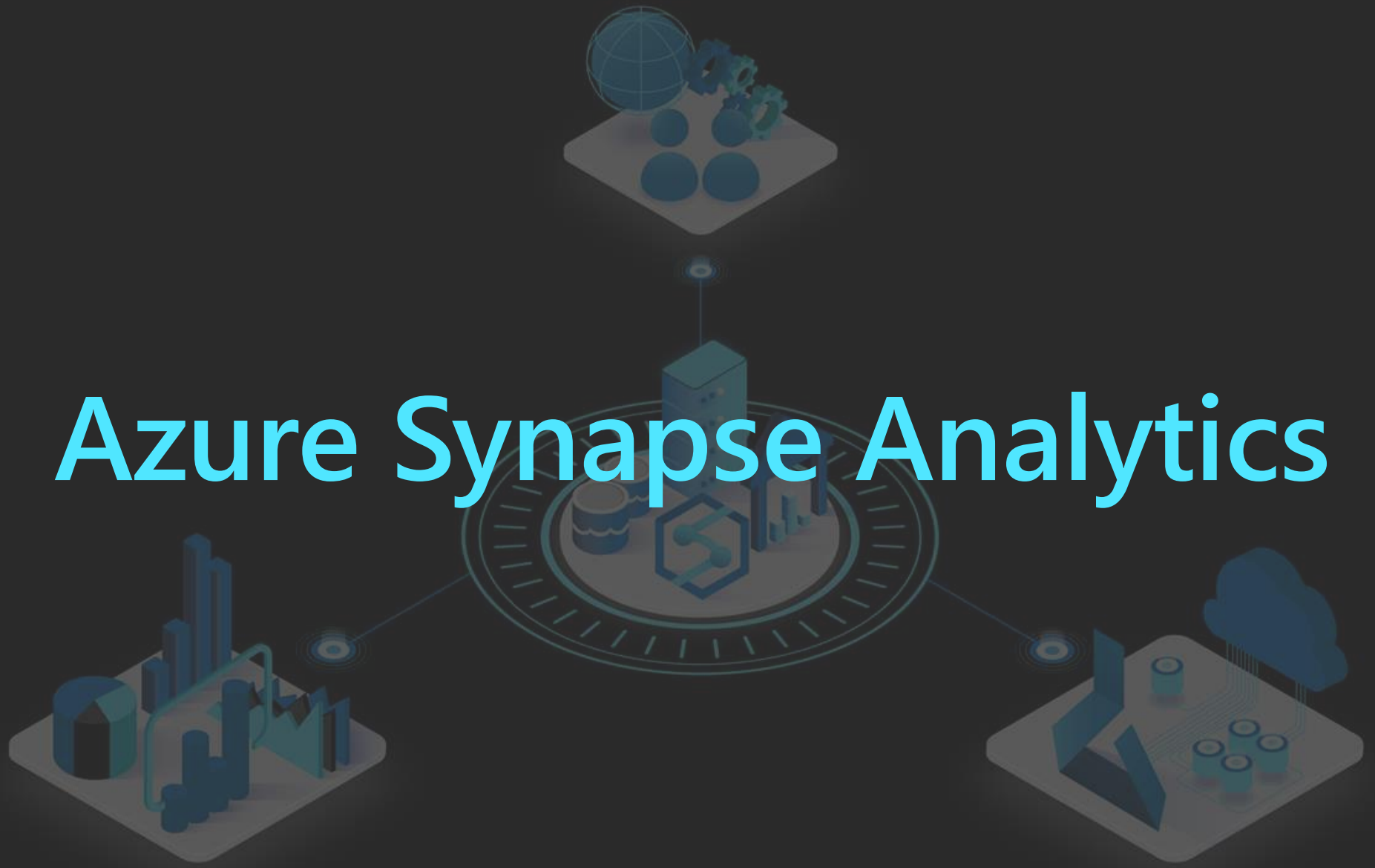
TPC-H Equivalent Benchmark

Price-performance | Lower is better



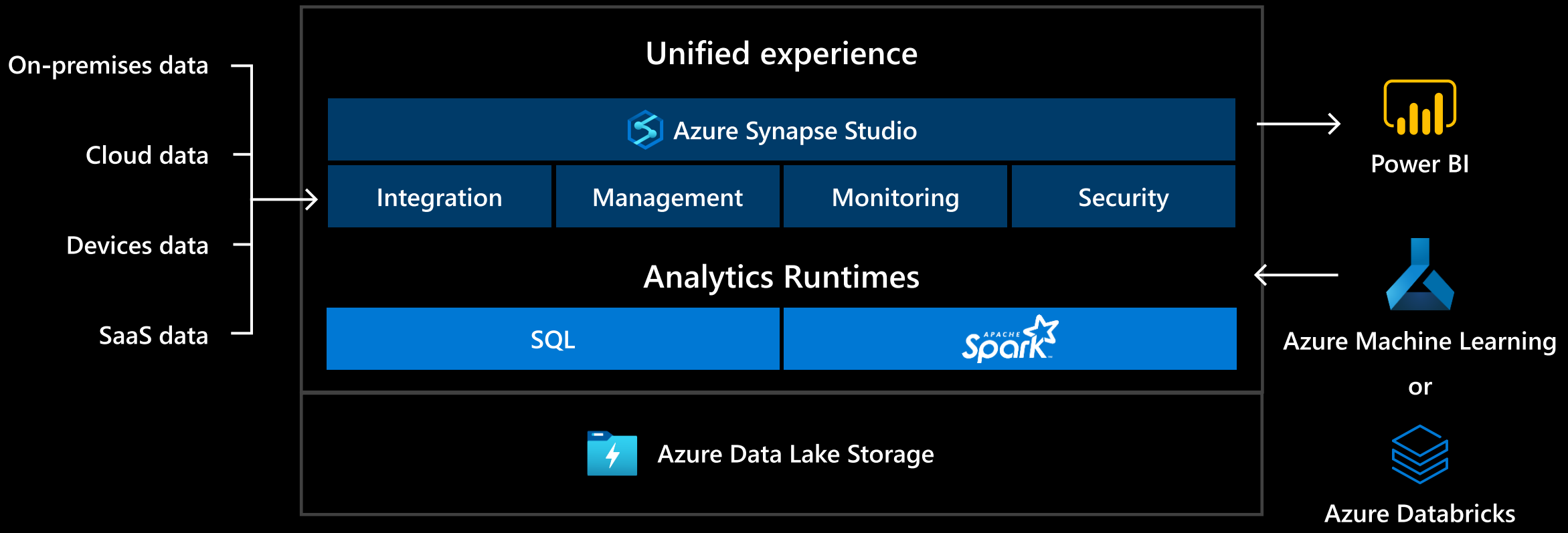
GigaOm TPC-H benchmark report, January 2019, "GigaOm report: Data Warehouse in the Cloud Benchmark"

Azure Synapse Analytics



Azure Synapse Analytics

Limitless data warehouse with unmatched time to insights



Azure Purview

UNIFIED DATA GOVERNANCE

Data Map

- Automate and manage metadata at scale

Data Catalog

- Enable effortless discovery for data consumers

Data Insights

- Assess data usage across your organization

Microsoft

Introducing Azure Synapse Analytics

Tutorial: Get started with Azure Synapse Analytics - Azure Synapse Analytics | Microsoft Docs

- STEP 1 - Create and setup a Synapse workspace
- STEP 2 - Analyze using a serverless SQL pool
- STEP 3 - Analyze using Apache Spark
- STEP 4 - Analyze using a dedicated SQL pool
- STEP 5 - Analyze data in a storage account
- STEP 6 - Orchestrate with pipelines
- STEP 7 - Visualize data with Power BI
- STEP 8 - Monitor activities
- STEP 9 - Explore the Knowledge center

Notes for following the Getting Started Guide 1

Point #1 [Spark]:
Section 3: Analyze using Spark - Load the NYC Taxi data into the Spark nyctaxi database section. Before step #1, you need to create the Spark database named nyctaxi using this command.

```
spark.sql("CREATE DATABASE IF NOT EXISTS nyctaxi")
```

Point #2 [Spark]:
Section 3: Analyze using Spark - Analyze the NYC Taxi data using Spark and notebooks. Step #4: displaying the results of the analysis into a table. The example code used the column TripDistanceMiles, but it is named TripDistance. Correct command looks like this:

```
%%pyspark
df = spark.sql("""
SELECT PassengerCount,
SUM(TripDistance) as SumTripDistance,
AVG(TripDistance) as AvgTripDistance
FROM nyctaxi.trip
WHERE TripDistance > 0 AND PassengerCount > 0
GROUP BY PassengerCount
ORDER BY PassengerCount
""")
display(df)
df.write.saveAsTable("nyctaxi.passengercountstats")
```

This step took a long time for me to execute as it writes more than 500,000,000 records.

Notes for following the Getting Started Guide 2

Point #3 [SQL, Power BI]:
Section 7: Visualize data with Power BI - Overview. From the NYC Taxi data, we created aggregated datasets in two tables:

- `nyctaxi.passengercounats`
- `SQLDB1.dbo.PassengerCountStats`

In the Synapse Analytics Workspace's Data Hub, refresh Databases and expand Databases > SQLPOOL1 > Tables. You will find that you don't have the SQLDB1.dbo.PassengerCountStats table. Run this command in a SQL script to create the table:

```
SELECT PassengerCount,
SUM(TripDistanceMiles) as SumTripDistance,
AVG(TripDistanceMiles) as AvgTripDistance
INTO PassengerCountStats
FROM dbo.Trip
WHERE TripDistanceMiles > 0 AND PassengerCount > 0
GROUP BY PassengerCount
```

Point #4: [General]:
I found that the steps which involve starting up a Spark pool and saving the data frame data into a Spark table took up the most time. If you used a smaller data set, the save time could be reduced. You can complete the tutorial in less time by skipping the Spark sections if you only want to learn about using the SQLPool. If you want to do that, skip sections 2,3,6 and steps 5 and 6 in section 5. See Point #3 above.

Point #5 [General]:
When you are not using your SQLPool and Spark Pool, pause them in Azure Portal. If you're done with your tutorial, I suggest you delete your resources to save on cost. If you created your resources in a new resource group, then deleting the entire resource group makes this easier to manage.

Summary

- Start data-analytics in your workplace to build a data-driven strategy
- Understand what data you have, and what data you need to gather
- Have a close relationship with your users and stakeholders
- Try out Azure Synapse and Power BI
- Reach out / talk to your PDM / PTS to get started

Other Resources

- Azure Data Architecture Guide - Azure Architecture Center | Microsoft Docs
- Choosing an analytical data store - Azure Architecture Center | Microsoft Docs
- Get started with Azure Synapse
- Get samples for Power BI - Power BI | Microsoft Docs

Your feedback is important

Please help us improve this program by completing this short feedback form.

https://aka.ms/saaslabfeedback6


Microsoft
If you'd like more help on your Azure modernization journey, please e-mail the SaaS Lab team
saaslab@microsoft.com
Thank you for being part of the SaaS Lab Program