Microsoft Azure

# Azure Data Platform End-to-End

## Implement a Modern Data Platform Architecture

**<your name>**
<your role>
*<your email>*

# Begin with the end in mind

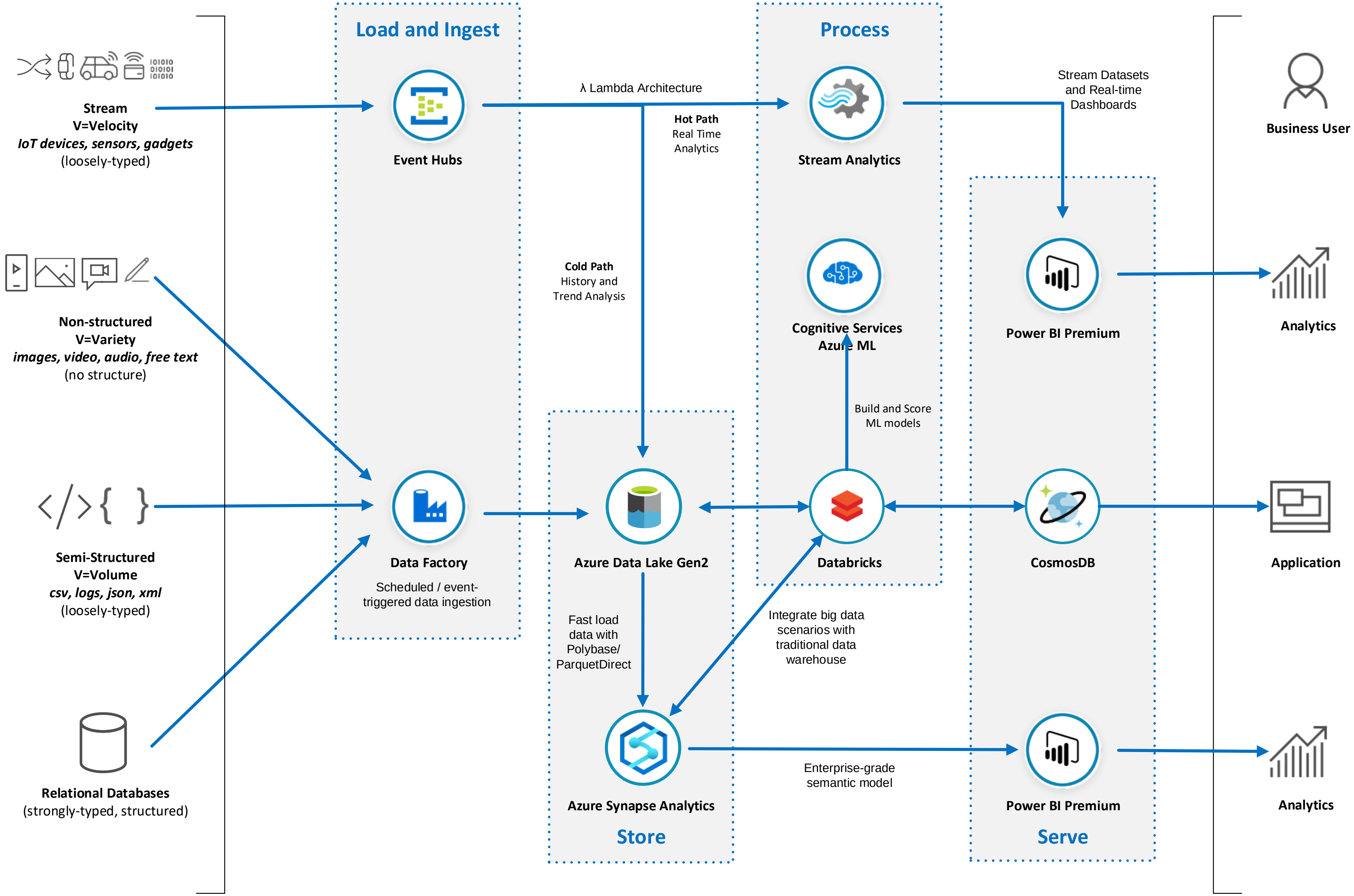
# Course Objectives

- We will understand Cloud and Big Data concepts and technologies used to solve the **most common** advanced analytics problems
- We will understand the role of Microsoft Azure data services in a modern data platform architecture
- We will look at individual Azure Data Services and use them to implement a modern data platform reference architecture
- We will have a ARM template of a data platform that will enable us to solve most of our data challenges

# Important Reminder

- The modern data platform architecture proposed in this course aims to help with your technology decisions when architecting data solutions in Azure.
- The Azure services covered in this course are only a subset of a much larger family of data services. Some real-world data scenarios may require the use of services not included in this course.
- This course does not replace in-depth training on each Azure service covered today.
- Some concepts presented in this course can be quite complex and you may need to seek for more information from different sources.

# Modern Data Platform Reference Architecture



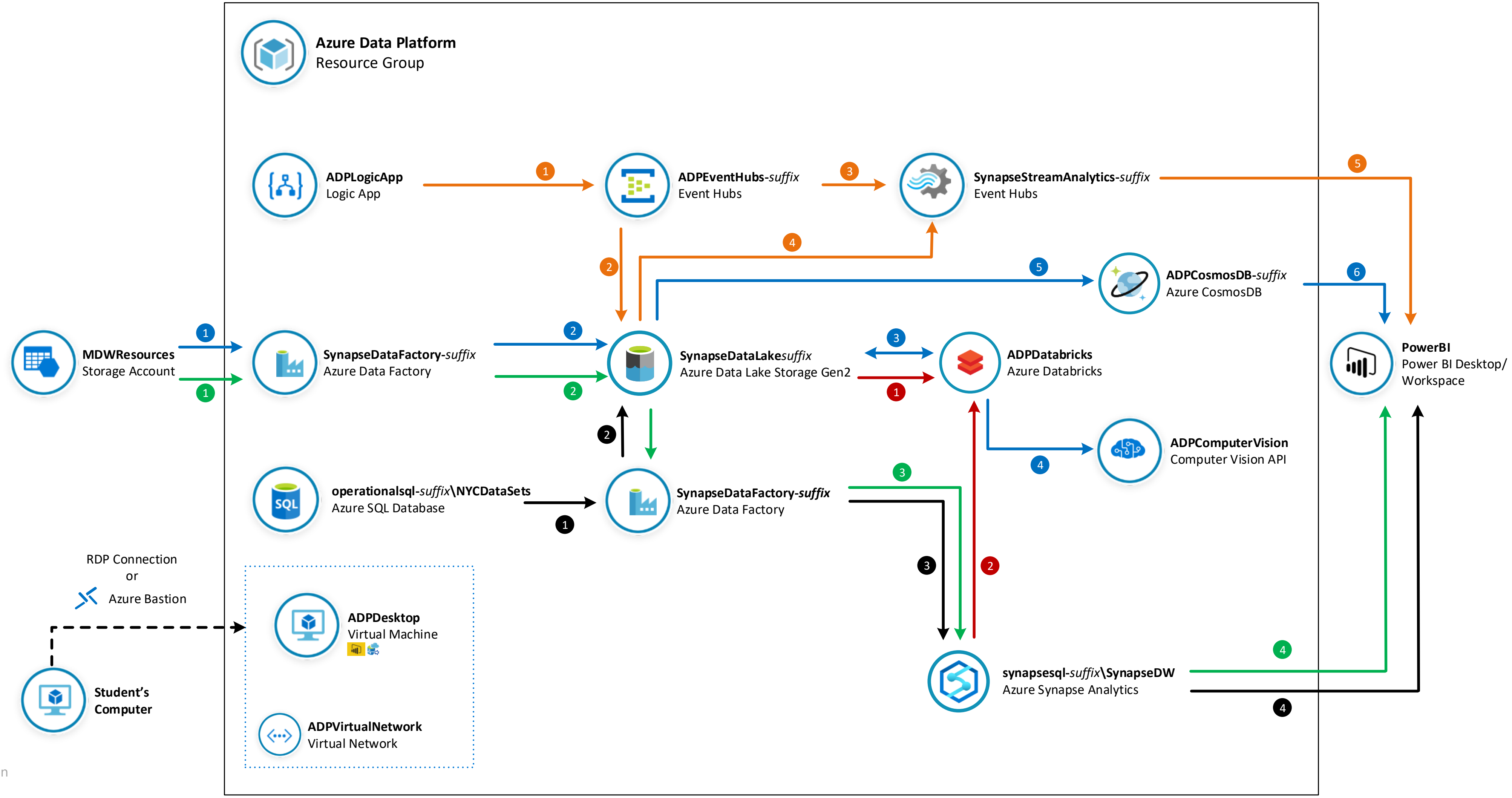
# Lab Guide

Microsoft

## Azure Data Platform End2End

Lab Architecture

- → Lab 1: Load Data into Azure Synapse Analytics using Azure Data Factory Pipelines
- → Lab 2: Transform Big Data using Azure Data Factory Mapping Data Flows
- → Lab 3: Explore Big Data with Azure Databricks
- → Lab 4: Add AI to your Big Data pipeline with Cognitive Services
- → Lab 5: Ingest and Analyse Real-Time Data with Event Hubs and Stream Analytics



**Azure Data Platform**
Resource Group

**ADPLogicApp**
Logic App

**ADPEventHubs**-*suffix*
Event Hubs

**SynapseStreamAnalytics**-*suffix*
Event Hubs

**ADPCosmosDB**-*suffix*
Azure CosmosDB

**MDWResources**
Storage Account

**SynapseDataFactory**-*suffix*
Azure Data Factory

**SynapseDataLake***suffix*
Azure Data Lake Storage Gen2

**ADPDatabricks**
Azure Databricks

**PowerBI**
Power BI Desktop/
Workspace

**ADPComputerVision**
Computer Vision API

**operationalsql**-*suffix***\NYCDataSets**
Azure SQL Database

**SynapseDataFactory**-***suffix***
Azure Data Factory

RDP Connection
or
Azure Bastion

**ADPDesktop**
Virtual Machine

**Student's**
**Computer**

**ADPVirtualNetwork**
Virtual Network

**synapsesql**-*suffix***\SynapseDW**
Azure Synapse Analytics

# The modern data world out there

# I tried to understand it, but…

No-SQL

Databricks

Storm

Data Catalog

IoT

PaaS vs IaaS

Hadoop

Power BI

Streaming

Deep Learning

Machine Learning

SMP vs MPP

Predictive

Data Mart

ETL vs ELT

Prescriptive

Data Visualisation

Data Warehouse

Data Lake

Master Data

Data Factory

Cloud vs On-prem

Big Data

Data Quality

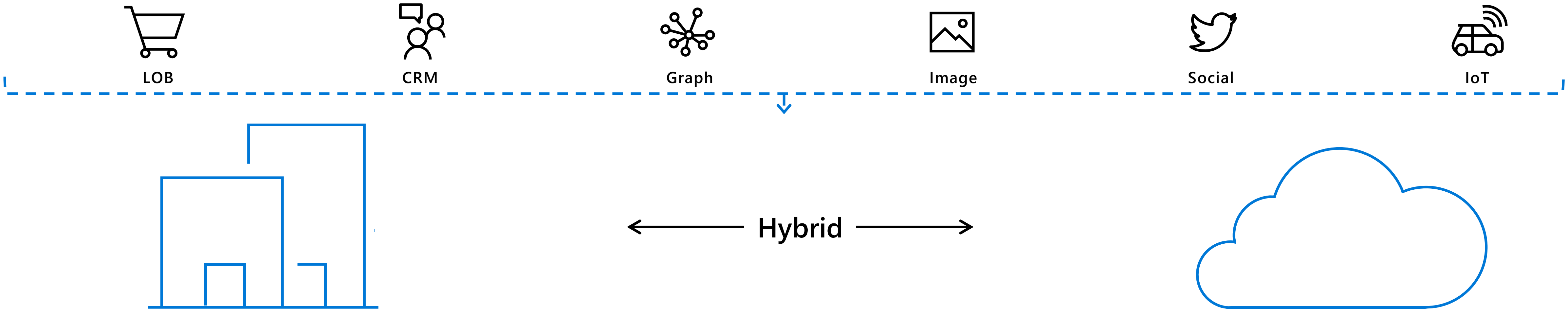
Velocity, Variety and Volume

AI

Semantic Layer

Spark

# The modern data estate



LOB

CRM

Graph

Image

Social

IoT

Hybrid

Operational databases

Data warehouses

Data lakes

Operational databases

Data warehouses

Data lakes

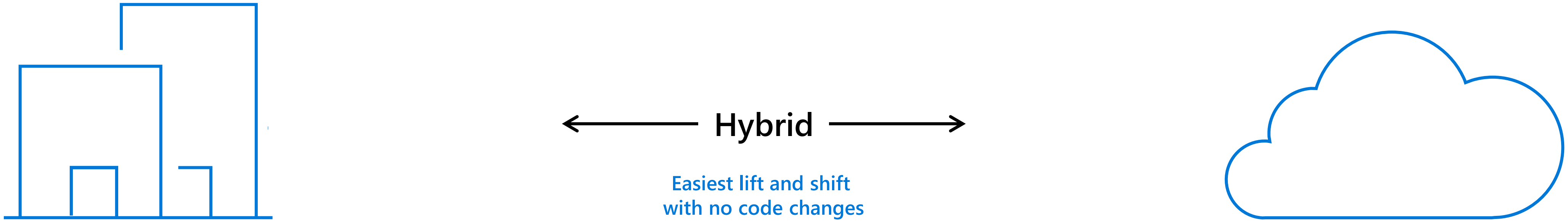
Reason over any data, anywhere

Flexibility of choice

Security and performance

# The Microsoft offering

LOB | CRM | Graph | Image | Social | IoT



Easiest lift and shift
with no code changes

## SQL Server

| | |
|---|---|
| Industry leader 4 years in a row | Operational databases |
| #1 TPC-H performance | Data warehouses |
| T-SQL query over any data | Data lakes |

## Azure Data Services

| | |
|---|---|
| Operational databases | 70% faster |
| Data warehouses | 2x the global reach |
| Data lakes | 99.9% SLA |

AI built-in | Most secure | Lowest TCO

Reason over any data, anywhere

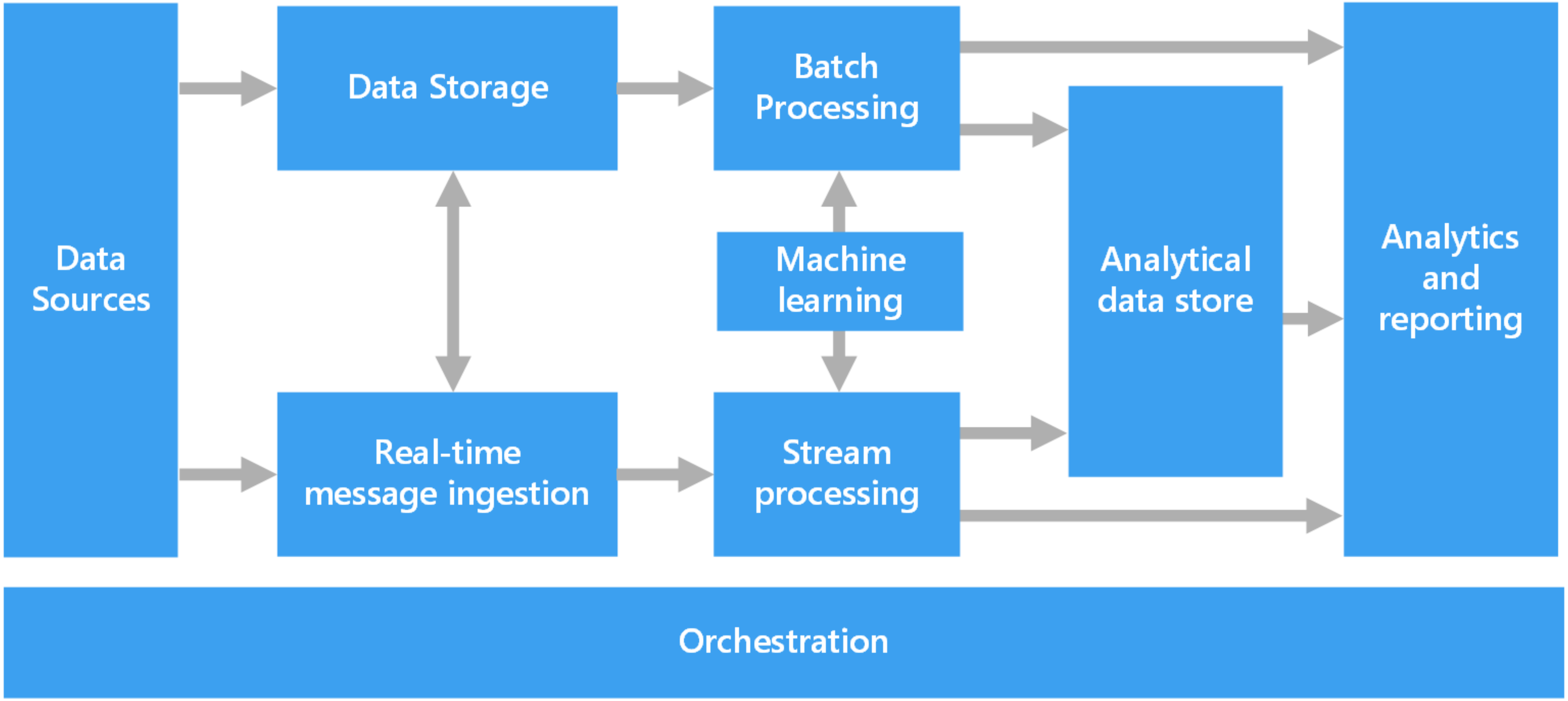
Flexibility of choice

Security and performance

# Azure Data Architecture Guide

Valuable collection of architecture principles to help you with your technology choices

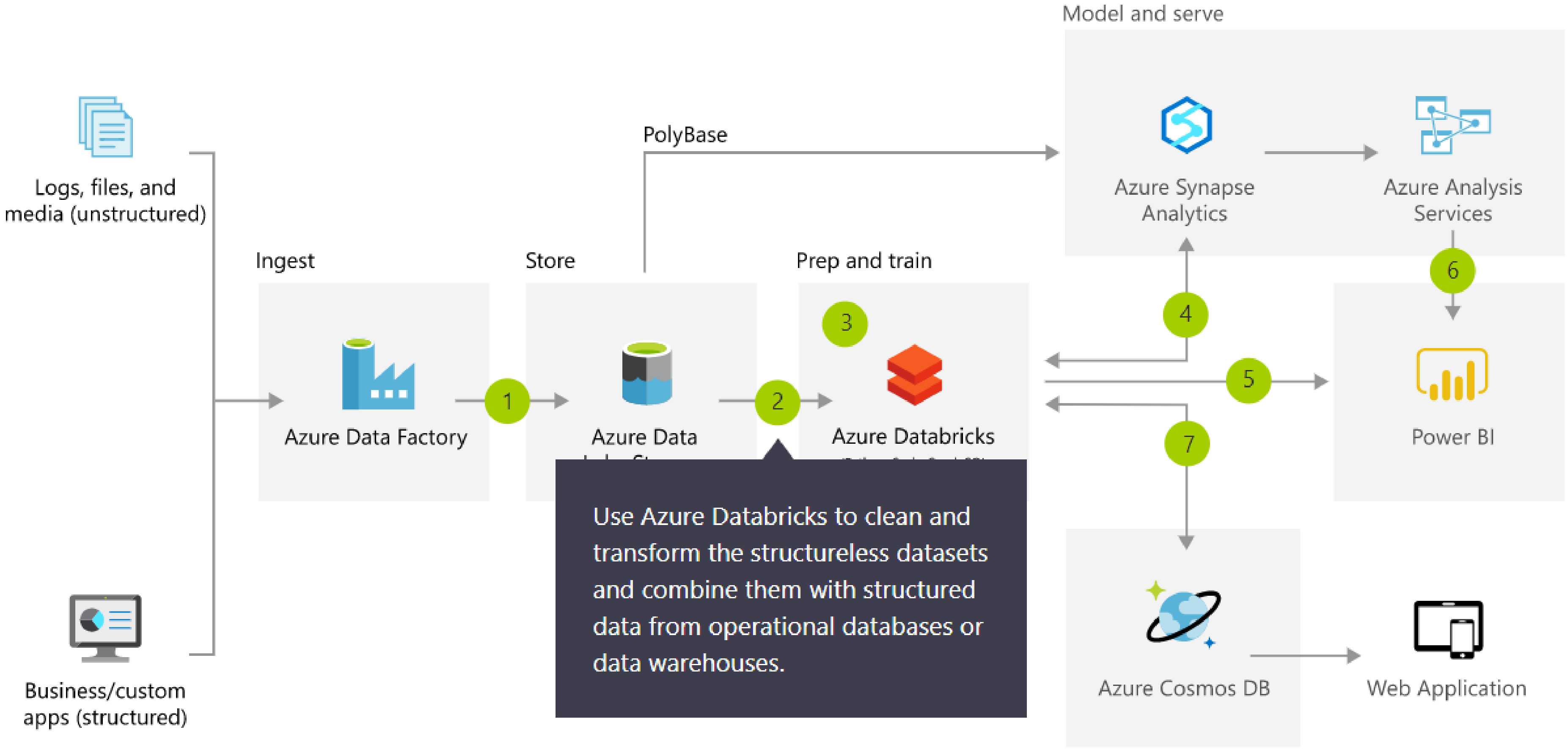
https://aka.ms/adag



# Azure Architecture Solutions

## Collection of reference architectures for most common challenges

[https://azure.microsoft.com/en-us/solutions/architecture/](https://azure.microsoft.com/en-us/solutions/architecture/)



Use Azure Databricks to clean and transform the structureless datasets and combine them with structured data from operational databases or data warehouses.

# Modern Data Platform Solution Scenarios

## Big Data and advanced analytics

### Modern data warehousing

“We want to integrate all our data—including Big Data—with our data warehouse”

### Advanced analytics

“We’re trying to predict when our customers churn”

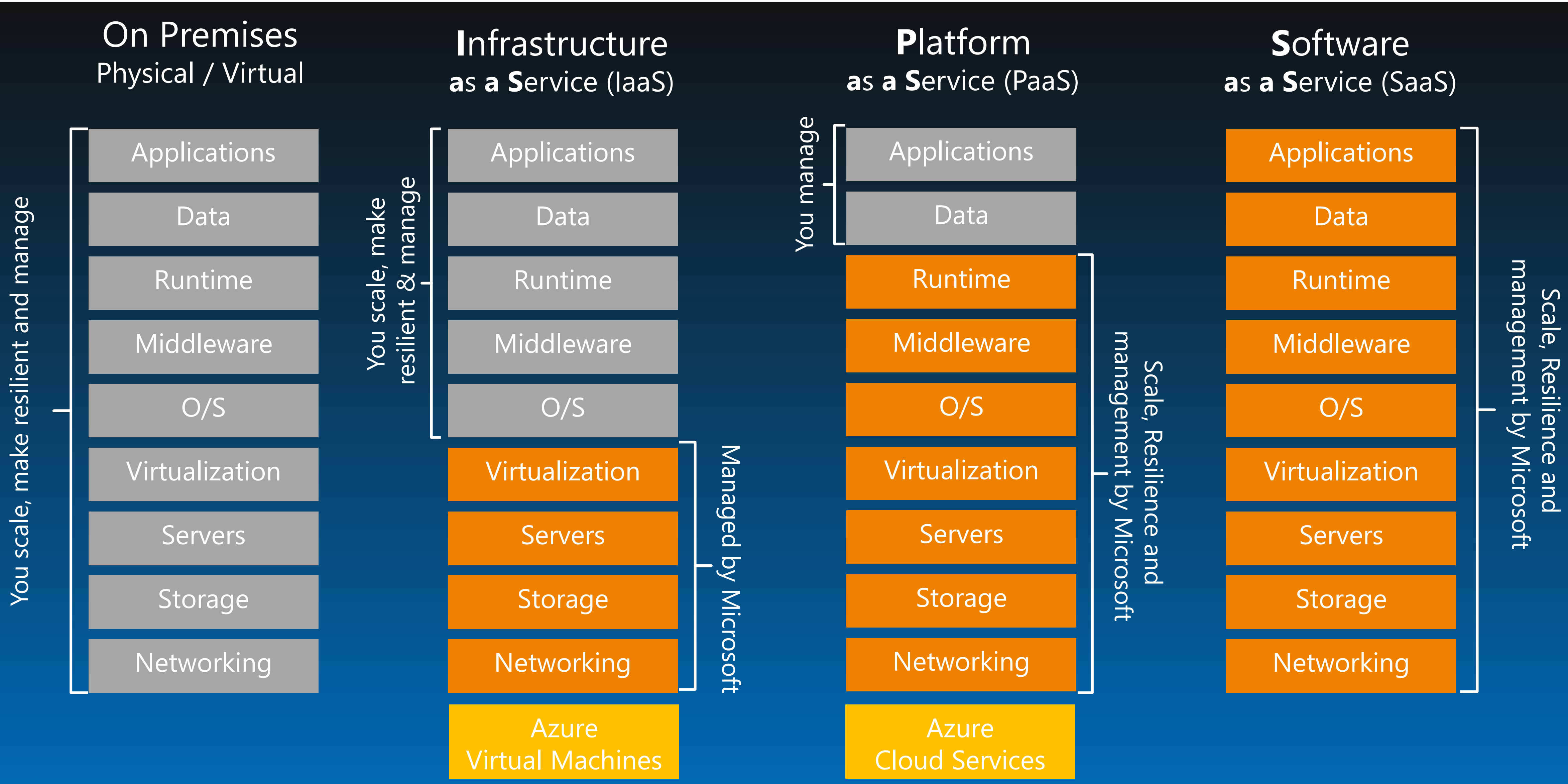
### Real-time analytics

“We’re trying to get insights from our devices in real-time”

# Modern Data Platform Concepts

## Part I

# IaaS vs PaaS vs SaaS



| On Premises<br>Physical / Virtual | Infrastructure<br>as a Service (IaaS) | Platform<br>as a Service (PaaS) | Software<br>as a Service (SaaS) |
|---|---|---|---|
| Applications | Applications | Applications | Applications |
| Data | Data | Data | Data |
| Runtime | Runtime | Runtime | Runtime |
| Middleware | Middleware | Middleware | Middleware |
| O/S | O/S | O/S | O/S |
| Virtualization | Virtualization | Virtualization | Virtualization |
| Servers | Servers | Servers | Servers |
| Storage | Storage | Storage | Storage |
| Networking | Networking | Networking | Networking |
|  | Azure<br>Virtual Machines | Azure<br>Cloud Services |  |

You scale, make resilient and manage

You scale, make resilient & manage

Managed by Microsoft

You manage

Scale, Resilience and management by Microsoft

Scale, Resilience and management by Microsoft

# What is a Data Warehouse?

A data warehouse is a large collection of business data used to help an organization make decisions. Data in the Data Warehouse has been identified as valuable to specifically defined business cases and is stored in a structured way readily available for reporting and data analysis.

**It is not an Operational Database**

Different workload types: transactional (DB) versus analytics (DW)

**It is not a Data Lake**

These are different concepts, they can co-exist and they compliment each other

**It is not a Data Mart**

A data mart is a subject-oriented database populated from a subset of the Data Warehouse

# Modern Data Warehousing

# Modern data warehousing

**The modern data warehouse extends the scope of the data warehouse to serve Big Data that's prepared with techniques beyond relational ETL**

## Modern data warehousing

"We want to integrate all our data—including Big Data—with our data warehouse"

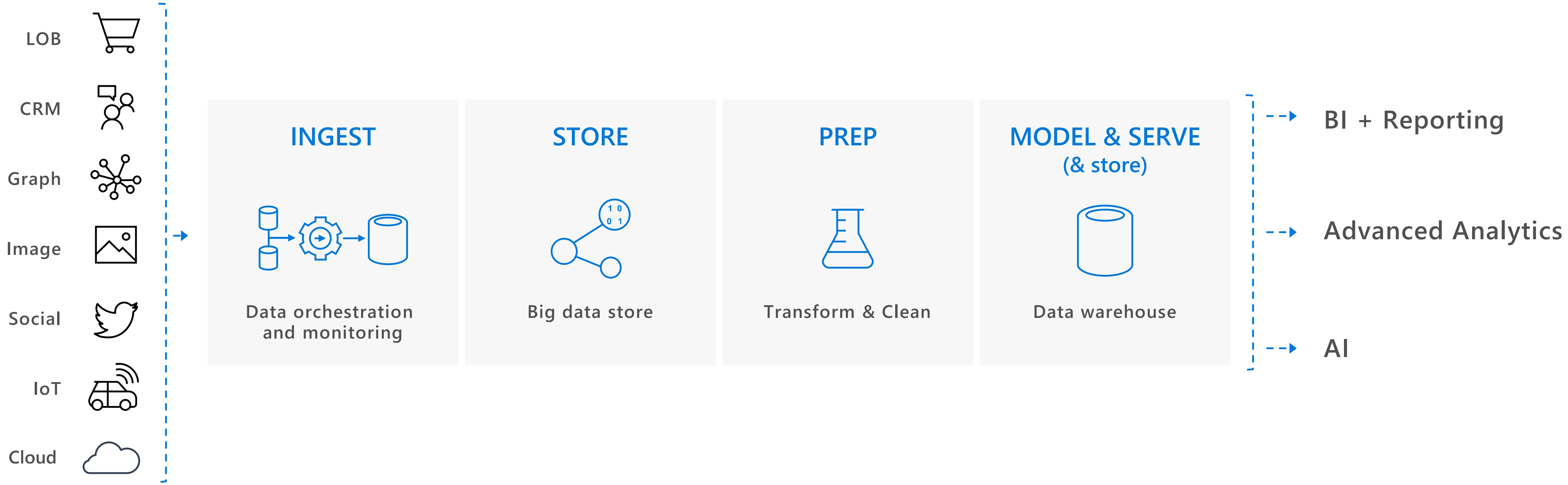
## Advanced analytics

"We're trying to predict when our customers churn"

## Real-time analytics

"We're trying to get insights from our devices in real-time"

# Modern data warehousing pattern



LOB

CRM

Graph

Image

Social

IoT

Cloud

INGEST

Data orchestration and monitoring

STORE

Big data store

PREP

Transform & Clean

MODEL & SERVE (& store)

Data warehouse

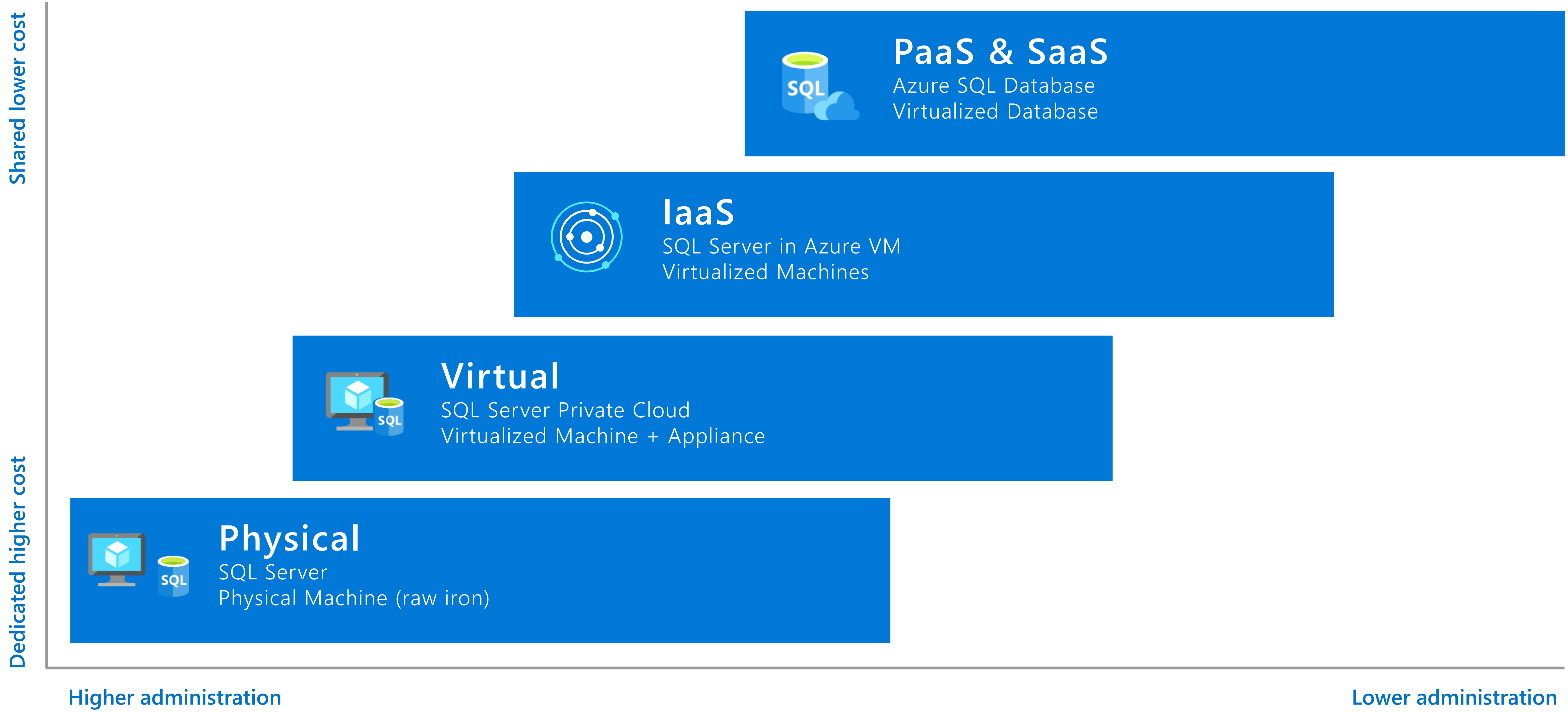
BI + Reporting

Advanced Analytics

AI

# SQL Server and Azure SQL Database

# Data platform continuum



Shared lower cost

Dedicated higher cost

**PaaS & SaaS**
Azure SQL Database
Virtualized Database

**IaaS**
SQL Server in Azure VM
Virtualized Machines

**Virtual**
SQL Server Private Cloud
Virtualized Machine + Appliance

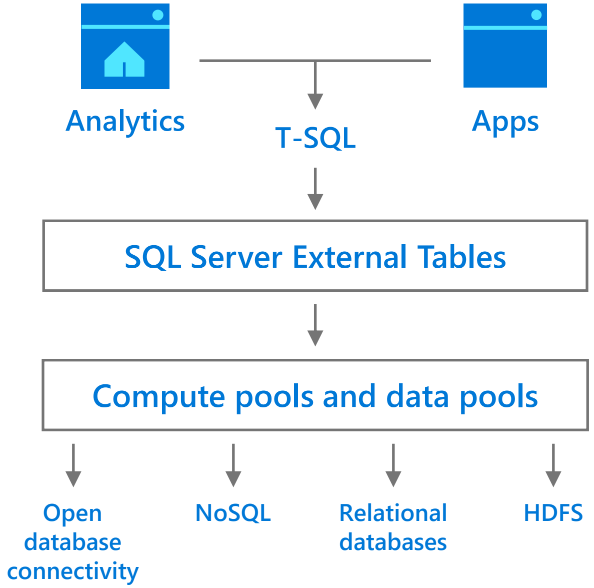
**Physical**
SQL Server
Physical Machine (raw iron)

Higher administration

Lower administration

# SQL Server 2019 big data, analytics, and AI

## Data virtualization



Analytics

T-SQL

Apps

SQL Server External Tables

Compute pools and data pools

Open database connectivity

NoSQL

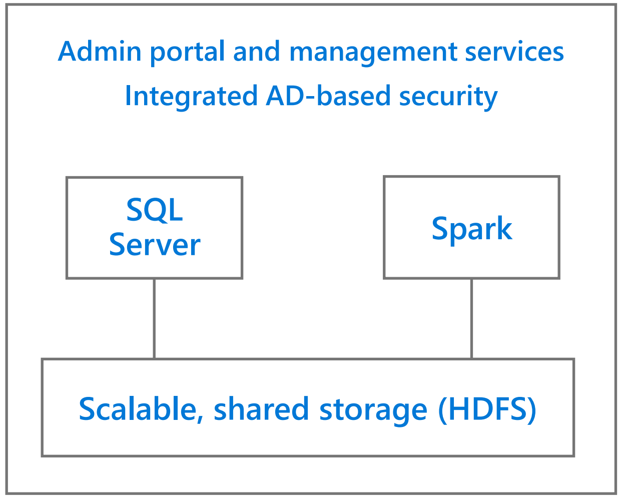
Relational databases

HDFS

Combine data from many sources without moving or replicating it

Scale out compute and caching to boost performance

## Managed SQL Server, Spark, and data lake



Admin portal and management services

Integrated AD-based security

SQL Server

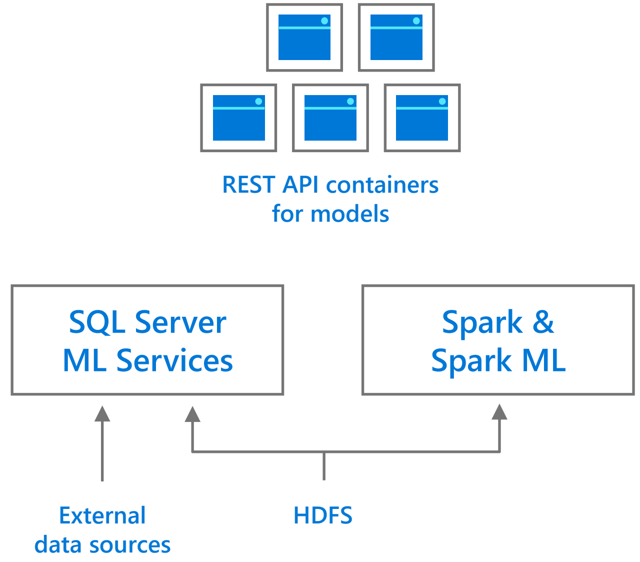
Spark

Scalable, shared storage (HDFS)

Store high volume data in a data lake and access it easily using either SQL or Spark

Management services, admin portal, and integrated security make it all easy to manage

## Complete AI platform



REST API containers for models

SQL Server ML Services

Spark & Spark ML

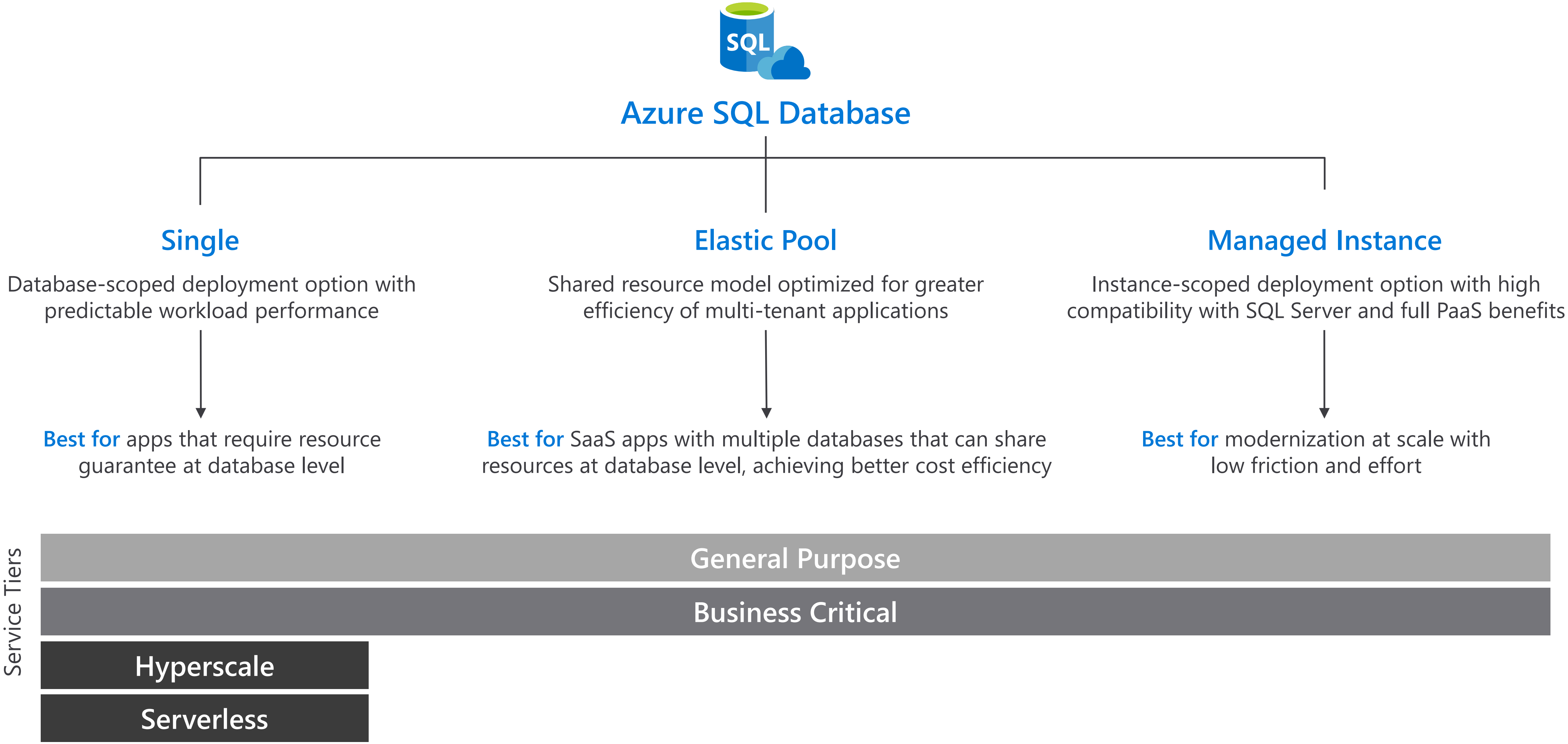
External data sources

HDFS

Easily feed integrated data from many sources to your model training

Ingest and prep data and then train, store, and operationalize your models all in one system

# Azure SQL Database deployment option



**Azure SQL Database**

**Single**

Database-scoped deployment option with predictable workload performance

**Best for** apps that require resource guarantee at database level

**Elastic Pool**

Shared resource model optimized for greater efficiency of multi-tenant applications

**Best for** SaaS apps with multiple databases that can share resources at database level, achieving better cost efficiency

**Managed Instance**

Instance-scoped deployment option with high compatibility with SQL Server and full PaaS benefits

**Best for** modernization at scale with low friction and effort

Service Tiers

General Purpose

Business Critical

Hyperscale

Serverless

# Azure Data Factory

# Azure Data Factory

## Hybrid data integration service for enabling code-free ETL

Industry leading
data ingestion

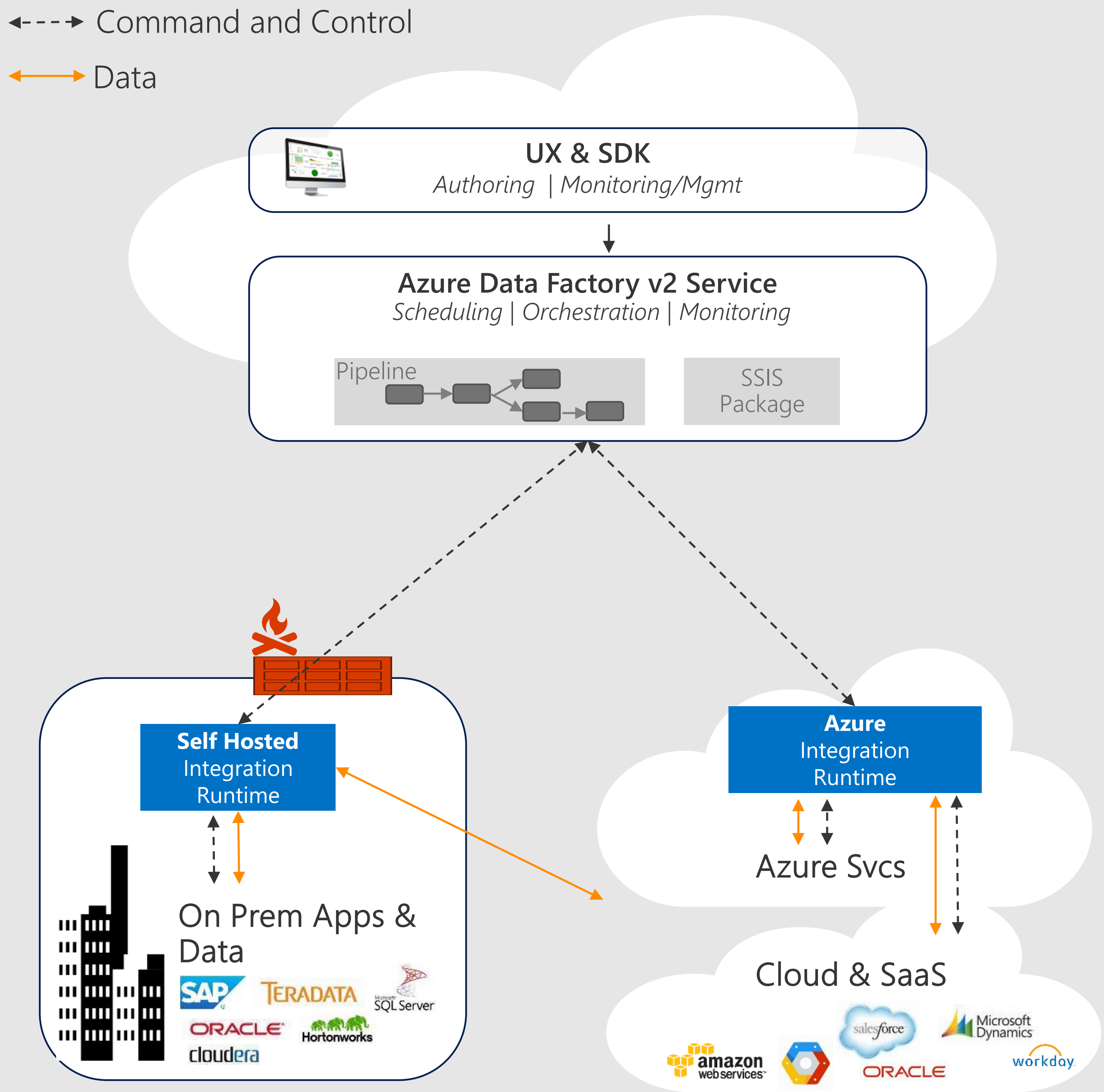
Visual
No Code

Hybrid

Pay only for what
you use

Managed SSIS

**Productive & trusted hybrid data integration service
that simplifies ETL with any data, from any source, at scale.**



## Data Factory

A data integration account.

Location of orchestration, service metadata

## Integration Runtime (IR)

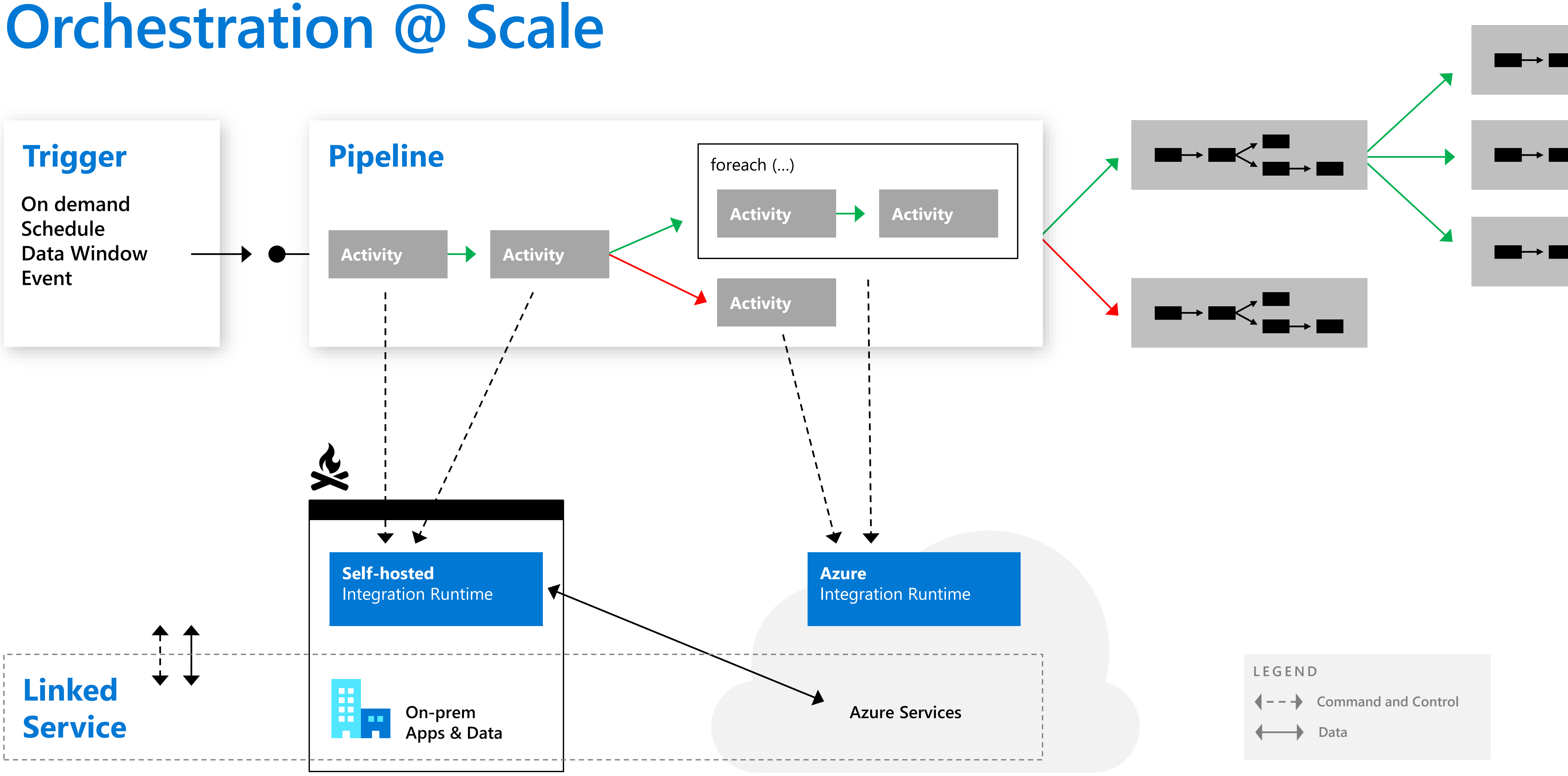
ADF's execution engine

- Azure Integration Runtime
- Self-Hosted Integration Runtime
- SSIS Integration Runtime

Three core capabilities:

- data movement
- pipeline activity execution
- SSIS package execution

# Orchestration @ Scale



## Trigger

On demand
Schedule
Data Window
Event

## Pipeline

Activity
Activity
foreach (...)
Activity
Activity
Activity

Self-hosted Integration Runtime

Azure Integration Runtime

## Linked Service

On-prem Apps & Data

Azure Services

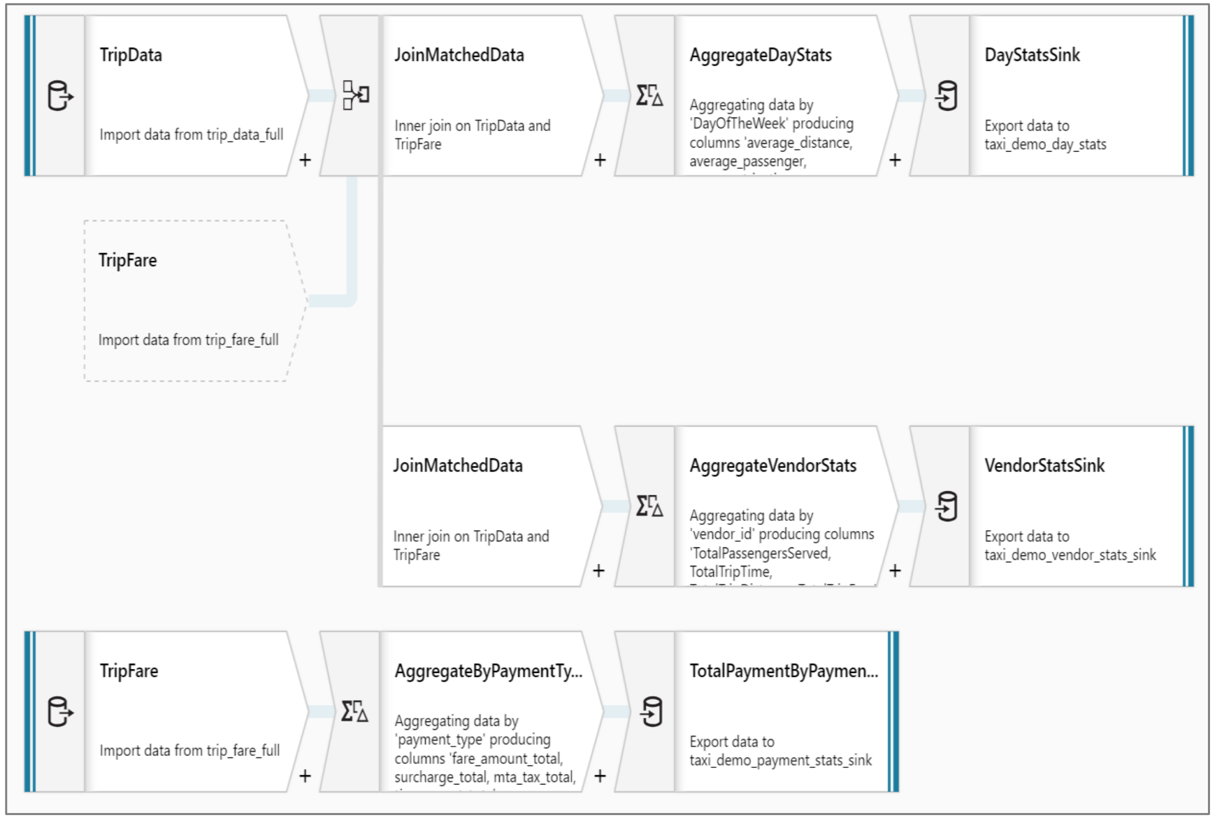
LEGEND
Command and Control
Data

# Azure Data Factory Data Flows

**No-code data transformation and preparation @ scale**

## Mapping Dataflow

Code free data transformation @scale



TripData

Import data from trip_data_full

JoinMatchedData

Inner join on TripData and TripFare

AggregateDayStats

Aggregating data by 'DayOfTheWeek' producing columns 'average_distance, average_passenger,

DayStatsSink

Export data to taxi_demo_day_stats

TripFare

Import data from trip_fare_full

JoinMatchedData

Inner join on TripData and TripFare

AggregateVendorStats

Aggregating data by 'vendor_id' producing columns 'TotalPassengersServed, TotalTripTime,

VendorStatsSink

Export data to taxi_demo_vendor_stats_sink

TripFare

Import data from trip_fare_full

AggregateByPaymentTy...

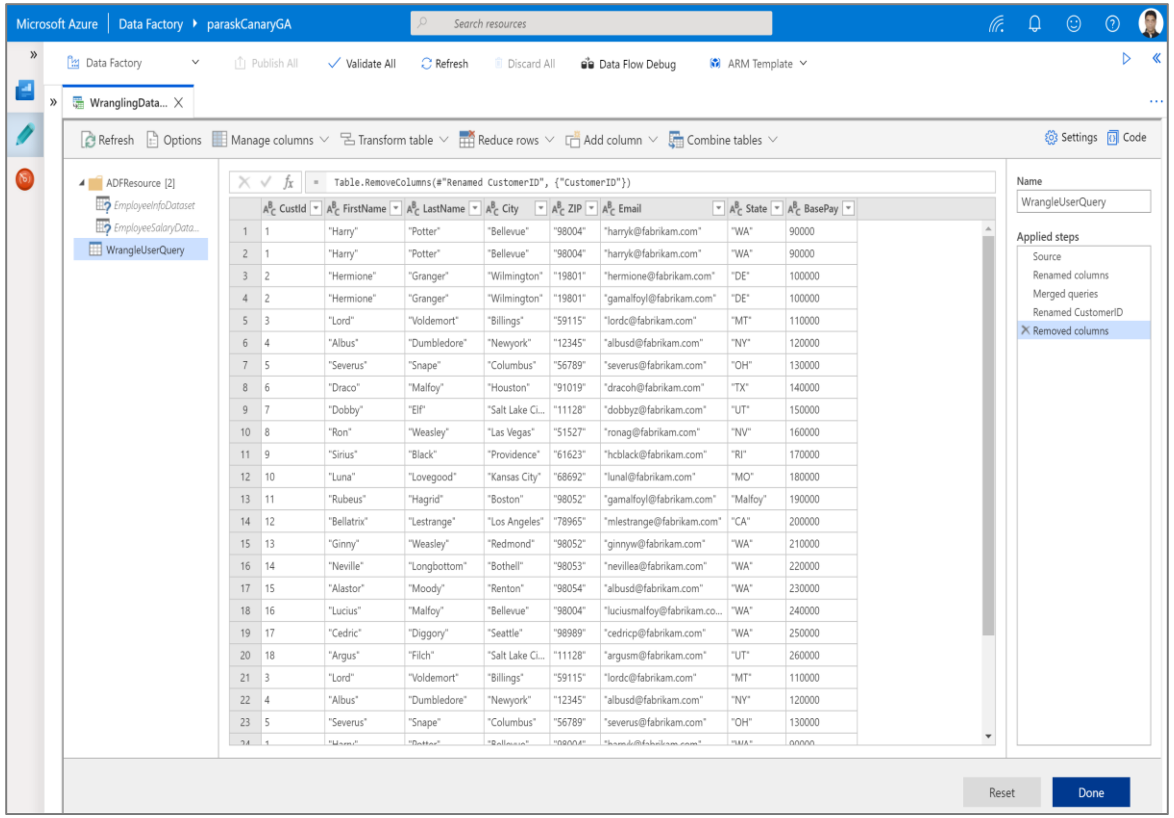
Aggregating data by 'payment_type' producing columns 'fare_amount_total, surcharge_total, mta_tax_total,

TotalPaymentByPaymen...

Export data to taxi_demo_payment_stats_sink

## Wrangling Dataflow

Code free data preparation @scale



```
= Table.RemoveColumns(#"Renamed CustomerID", {"CustomerID"})
```

|  | CustId | FirstName | LastName | City | ZIP | Email | State | BasePay |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | "Harry" | "Potter" | "Bellevue" | "98004" | "harryk@fabrikam.com" | "WA" | 90000 |
| 2 | 1 | "Harry" | "Potter" | "Bellevue" | "98004" | "harryk@fabrikam.com" | "WA" | 90000 |
| 3 | 2 | "Hermione" | "Granger" | "Wilmington" | "19801" | "hermione@fabrikam.com" | "DE" | 100000 |
| 4 | 2 | "Hermione" | "Granger" | "Wilmington" | "19801" | "gamalfoyl@fabrikam.com" | "DE" | 100000 |
| 5 | 3 | "Lord" | "Voldemort" | "Billings" | "59115" | "lordc@fabrikam.com" | "MT" | 110000 |
| 6 | 4 | "Albus" | "Dumbledore" | "Newyork" | "12345" | "albusd@fabrikam.com" | "NY" | 120000 |
| 7 | 5 | "Severus" | "Snape" | "Columbus" | "56789" | "severus@fabrikam.com" | "OH" | 130000 |
| 8 | 6 | "Draco" | "Malfoy" | "Houston" | "91019" | "dracoh@fabrikam.com" | "TX" | 140000 |
| 9 | 7 | "Dobby" | "Elf" | "Salt Lake Ci... | "11128" | "dobbyz@fabrikam.com" | "UT" | 150000 |
| 10 | 8 | "Ron" | "Weasley" | "Las Vegas" | "51527" | "ronag@fabrikam.com" | "NV" | 160000 |
| 11 | 9 | "Sirius" | "Black" | "Providence" | "61623" | "hcblack@fabrikam.com" | "RI" | 170000 |
| 12 | 10 | "Luna" | "Lovegood" | "Kansas City" | "68692" | "lunal@fabrikam.com" | "MO" | 180000 |
| 13 | 11 | "Rubeus" | "Hagrid" | "Boston" | "98052" | "gamalfoyl@fabrikam.com" | "Malfoy" | 190000 |
| 14 | 12 | "Bellatrix" | "Lestrange" | "Los Angeles" | "78965" | "mlestrange@fabrikam.com" | "CA" | 200000 |
| 15 | 13 | "Ginny" | "Weasley" | "Redmond" | "98052" | "ginnyw@fabrikam.com" | "WA" | 210000 |
| 16 | 14 | "Neville" | "Longbottom" | "Bothell" | "98053" | "nevillea@fabrikam.com" | "WA" | 220000 |
| 17 | 15 | "Alastor" | "Moody" | "Renton" | "98054" | "albusd@fabrikam.com" | "WA" | 230000 |
| 18 | 16 | "Lucius" | "Malfoy" | "Bellevue" | "98004" | "luciusmalfoy@fabrikam.co... | "WA" | 240000 |
| 19 | 17 | "Cedric" | "Diggory" | "Seattle" | "98989" | "cedricp@fabrikam.com" | "WA" | 250000 |
| 20 | 18 | "Argus" | "Filch" | "Salt Lake Ci... | "11128" | "argusm@fabrikam.com" | "UT" | 260000 |
| 21 | 3 | "Lord" | "Voldemort" | "Billings" | "59115" | "lordc@fabrikam.com" | "MT" | 110000 |
| 22 | 4 | "Albus" | "Dumbledore" | "Newyork" | "12345" | "albusd@fabrikam.com" | "NY" | 120000 |
| 23 | 5 | "Severus" | "Snape" | "Columbus" | "56789" | "severus@fabrikam.com" | "OH" | 130000 |
| 24 | 1 | "Harry" | "Potter" | "Bellevue" | "98004" | "harryk@fabrikam.com" | "WA" | 90000 |

Name: WrangleUserQuery

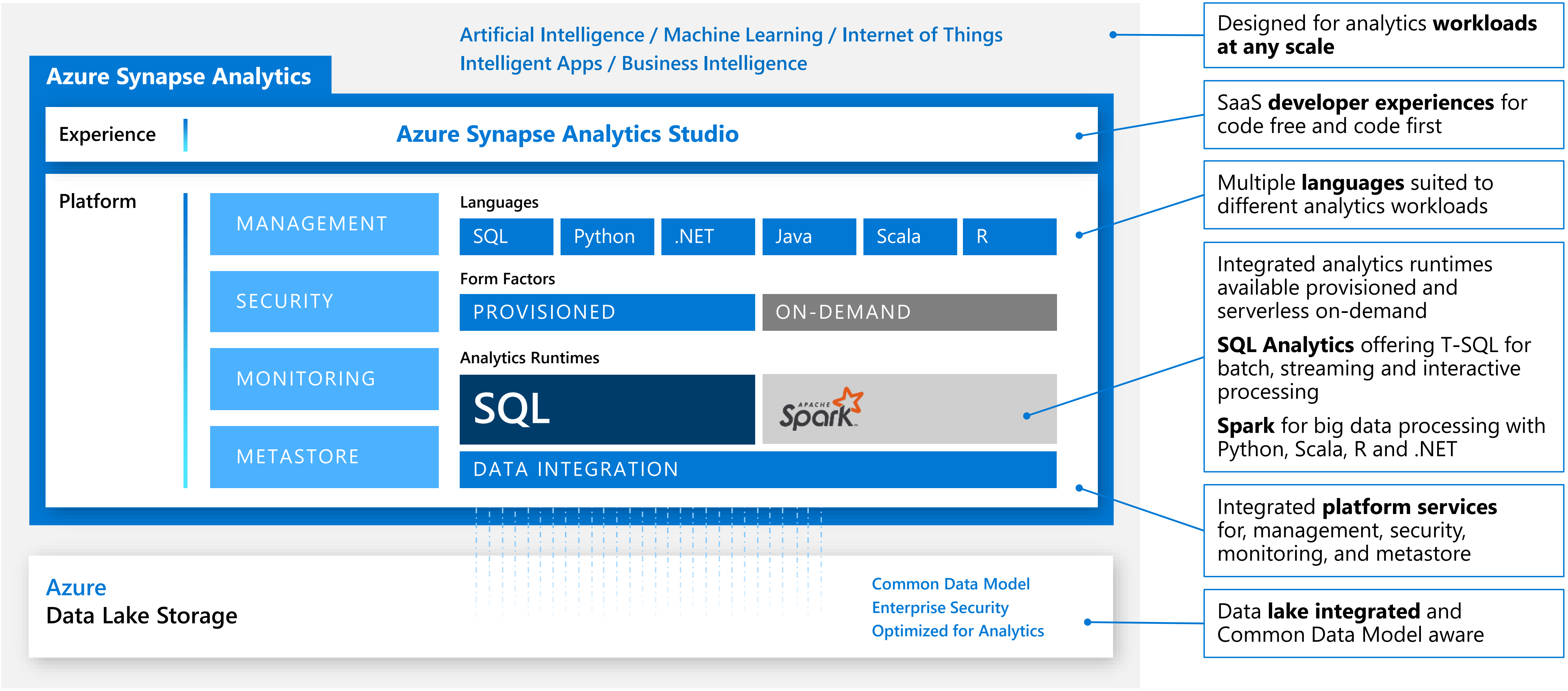
Applied steps

- Source
- Renamed columns
- Merged queries
- Renamed CustomerID
- Removed columns

# Azure Synapse Analytics

# Azure Synapse Analytics

Integrated data platform for BI, AI and continuous intelligence



Designed for analytics **workloads at any scale**

SaaS **developer experiences** for code free and code first

Multiple **languages** suited to different analytics workloads

Integrated analytics runtimes available provisioned and serverless on-demand

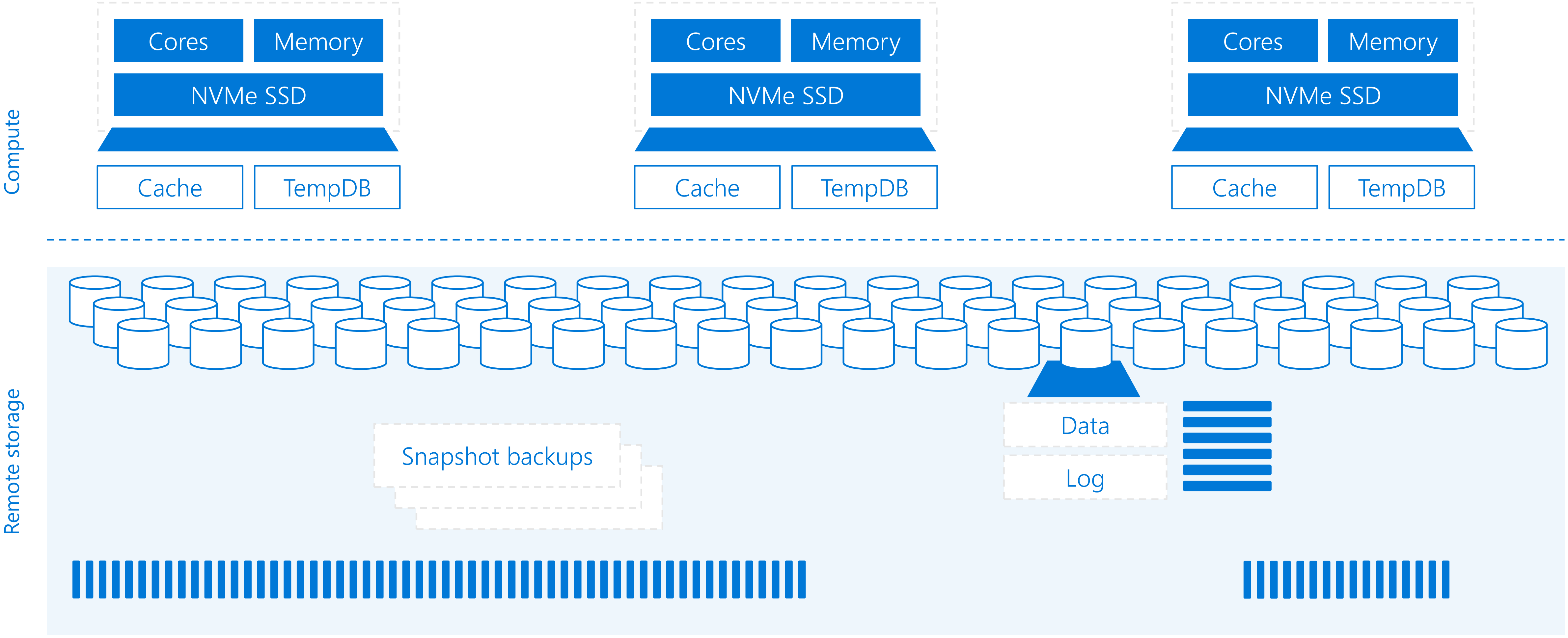
**SQL Analytics** offering T-SQL for batch, streaming and interactive processing

**Spark** for big data processing with Python, Scala, R and .NET

Integrated **platform services** for, management, security, monitoring, and metastore

Data **lake integrated** and Common Data Model aware

# Azure Synapse Analytics MPP Architecture



# Table Distributions

## Round-robin distributed

Distributes table rows evenly across all distributions at random.

## Hash distributed

Distributes table rows across the Compute nodes by using a deterministic hash function to assign each row to one distribution.

## Replicated

Full copy of table accessible on each Compute node.

```
CREATE TABLE [dbo].[FactInternetSales]
(
[ProductKey]            int          NOT NULL,
[OrderDateKey]          int          NOT NULL,
[CustomerKey]           int          NOT NULL,
[PromotionKey]          int          NOT NULL,
[SalesOrderNumber]      nvarchar(20) NOT NULL,
[OrderQuantity]         smallint     NOT NULL,
[UnitPrice]             money        NOT NULL,
[SalesAmount]           money        NOT NULL
)
WITH
(
CLUSTERED COLUMNSTORE INDEX,
DISTRIBUTION = HASH([ProductKey]) |
                ROUND ROBIN |
                REPLICATED
);
```

# Polybase

## Data ingestion using external data sources

### Overview

Polybase supports querying files stored in a Hadoop File System (HDFS), Azure Blob storage, or Azure Data Lake Store.

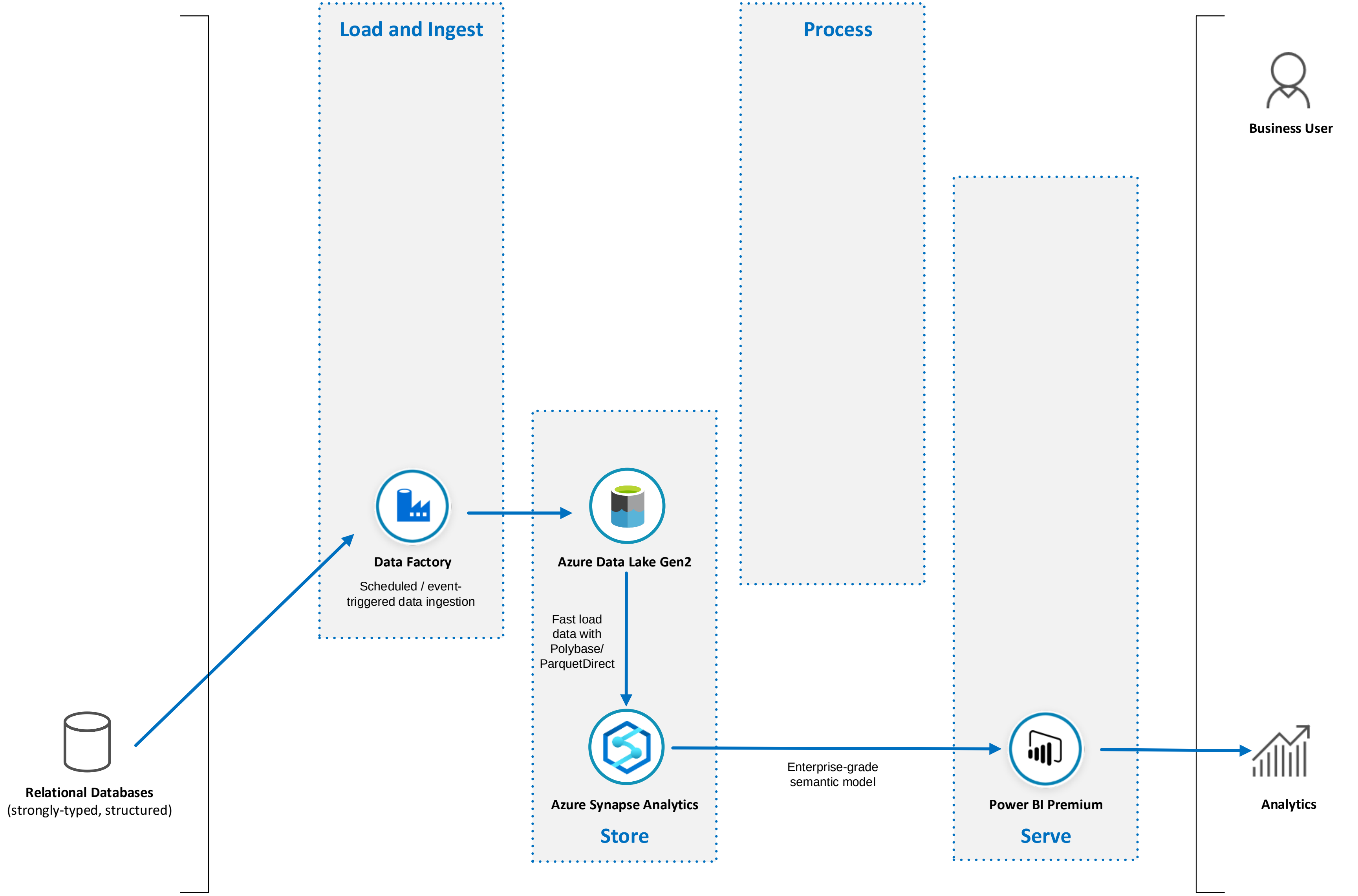
To query files, users create three objects: External data source, external file format, external table.

```
-- Create Azure DataLake Gen2 Storage reference
CREATE EXTERNAL DATA SOURCE AzureStorage with
(
TYPE = HADOOP,
LOCATION='abfss://<container>@<storageaccnt>.blob.core.windows.net',
CREDENTIAL = AzureStorageCredential -- not required if using
managed identity
);
-- Type of format in Hadoop (CSV, RCFILE , ORC, PARQUET).
CREATE EXTERNAL FILE FORMAT TextFileFormat WITH
(
FORMAT_TYPE = DELIMITEDTEXT,
FORMAT_OPTIONS (FIELD_TERMINATOR ='|', USE_TYPE_DEFAULT =
TRUE)
)
-- LOCATION: path to file or directory that contains data
CREATE EXTERNAL TABLE [dbo].[CarSensor_Data]
(
[SensorKey] int NOT NULL,
[Speed] float NOT NULL,
[YearMeasured] int NOT NULL
)
WITH (LOCATION='/Demo/’, DATA_SOURCE = AzureStorage,
FILE_FORMAT = TextFileFormat
);
```

# Lab 1: Load data into Azure Synapse Analytics using Azure Data Factory Pipelines

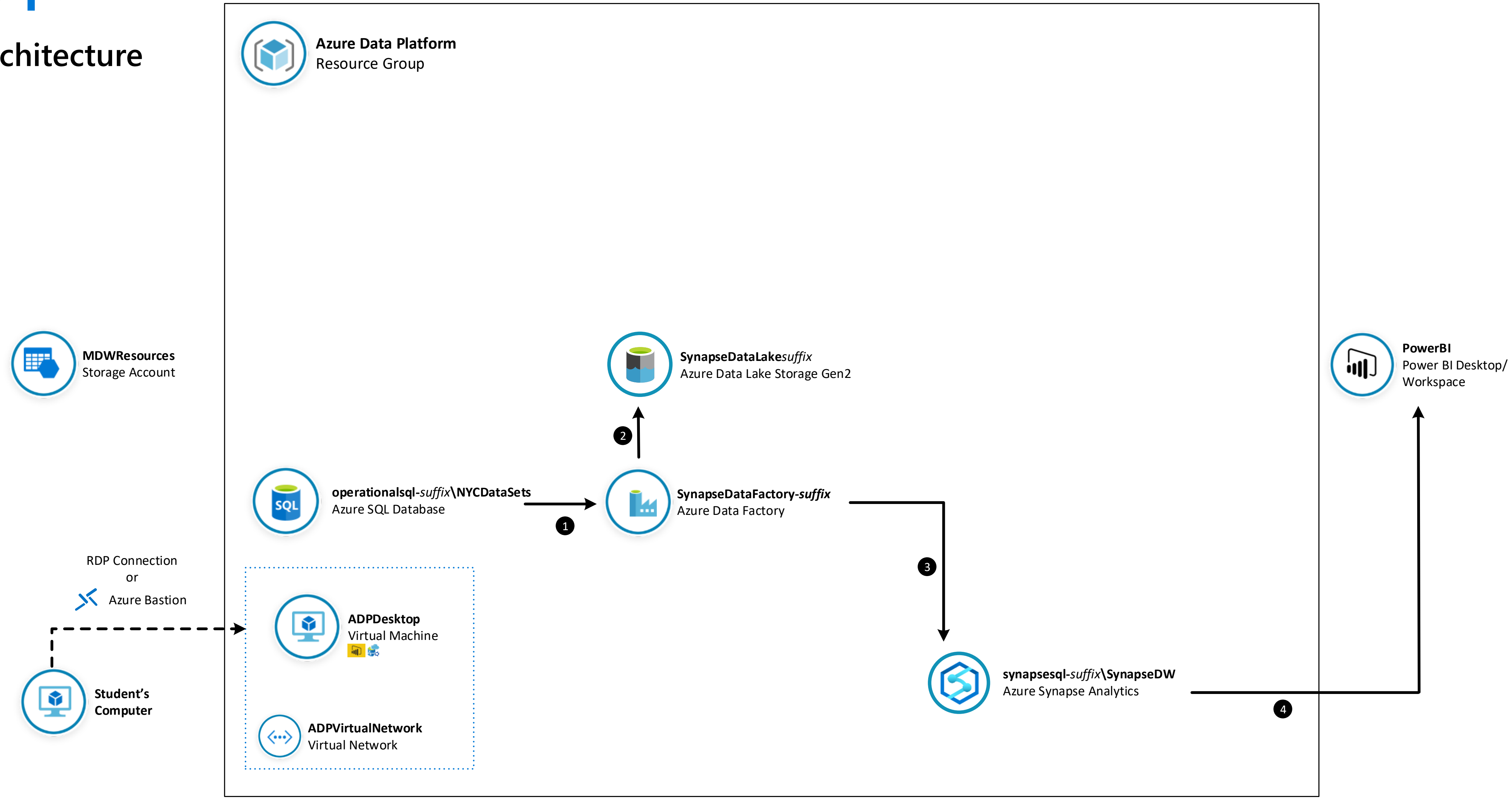
# Lab 1

## Load data into Azure Synapse Analytics using Azure Data Factory Pipelines



# Lab 1

## Lab Architecture



# Modern Data Platform Concepts

## Part II

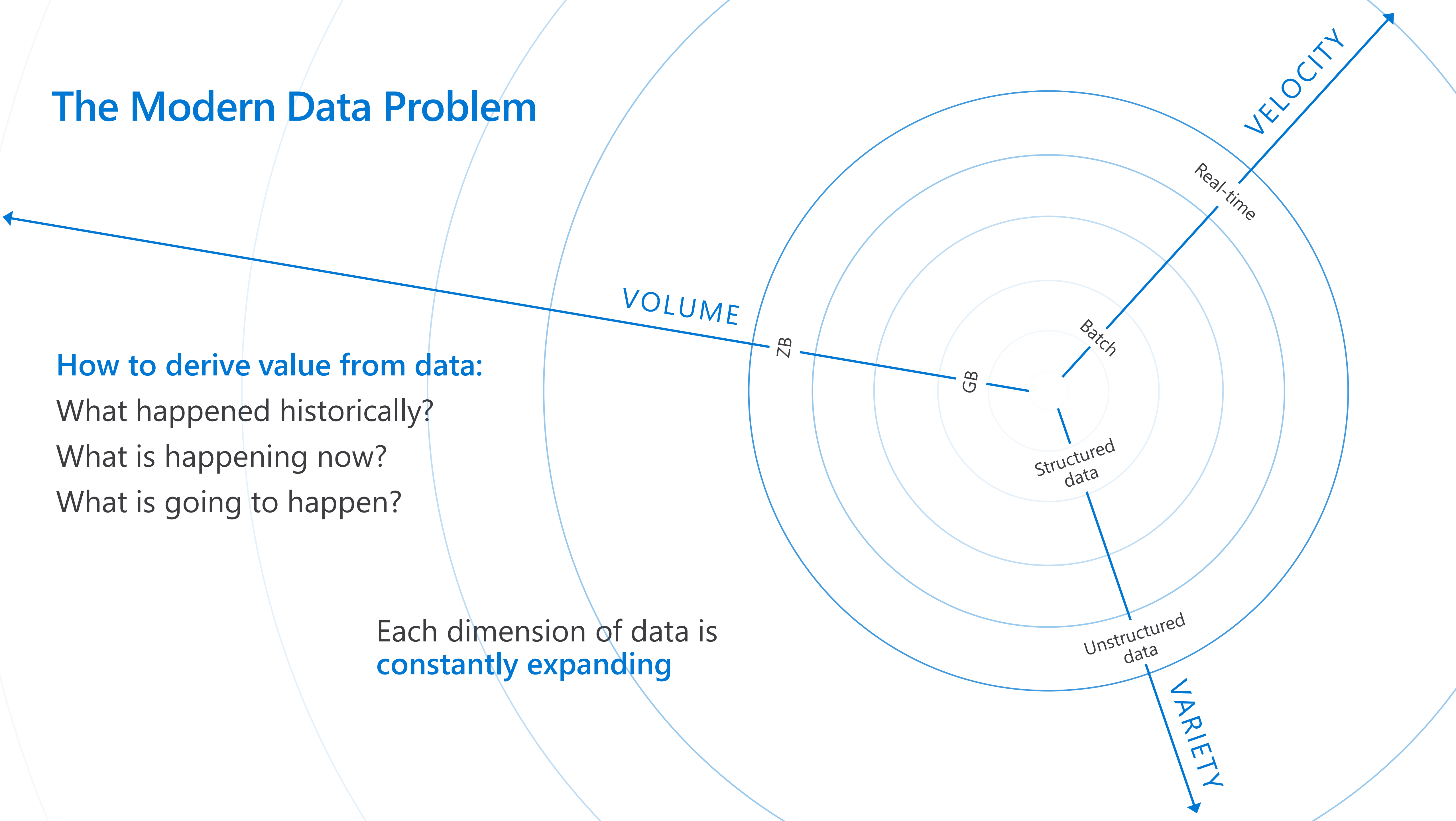
# The Modern Data Problem

**How to derive value from data:**

What happened historically?

What is happening now?

What is going to happen?



Each dimension of data is **constantly expanding**

# What is a Data Lake?

It is a central storage repository that holds data coming from many sources in a raw, granular format. It can store **structured, semi-structured, or unstructured data**, which means data ingested quickly and can be kept in a more flexible format for future use cases.

## Characteristics

- Schema-on-read (ELT)
- Collection of data, not a platform
- Perfect place for evolving data

## Benefits

- Quickly ingest high volumes of diverse data structures
- Enable advanced analytics and data exploration
- Scalability and storage cost reduction

## Best Practices

- Data Governance needed to avoid Data Swamp
- Security considerations
- Design your Data Lake
- Metadata management

# Data Warehouse or Data Lake?

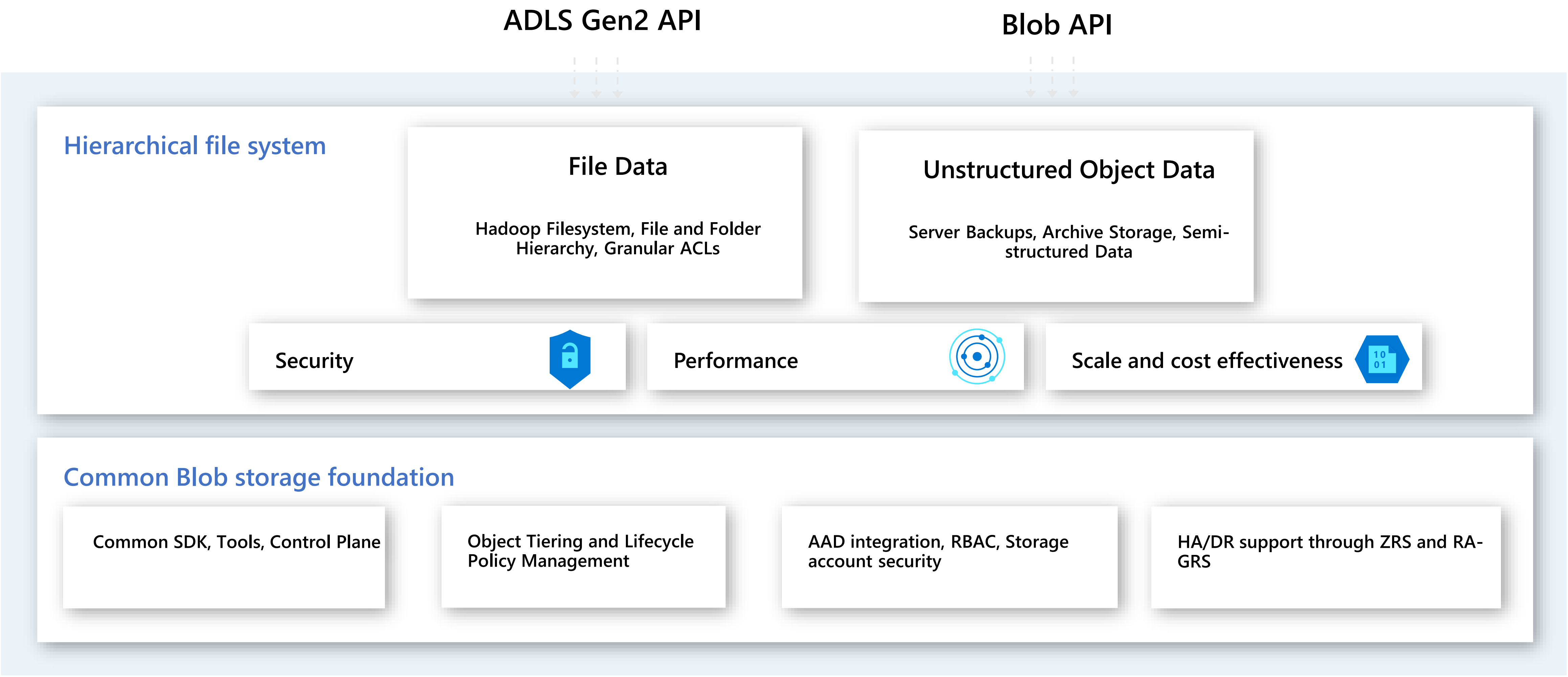
Answer: both.

| | Data Warehouse | Data Lake |
|---|---|---|
| Requirements | Relational requirements | Diverse data, scalability, low cost |
| Data Value | Data of recognised high value | Candidate data of potential value |
| Data Processing | Mostly refined calculated data | Mostly detailed source data |
| Business Entities | Known entities, tracked over time | Raw material for discovering entities and facts |
| Data Standards | Data conforms to enterprise standards | Fidelity to original format and condition |
| Data Integration | Data integration upfront | Data prep on demand |
| Transformation | Data transformed, in principle | Data repurposed later, as needs arise |
| Schema Definition | Schema-on-write | Schema-on-read |
| Metadata Management | Metadata improvement | Metadata developed on read |

# Azure Data Lake Storage Gen2

# Azure Data Lake Storage Gen2

## High performance HDFS Endpoint to Azure Blob Storage



ADLS Gen2 API

Blob API

Hierarchical file system

File Data

Hadoop Filesystem, File and Folder Hierarchy, Granular ACLs

Unstructured Object Data

Server Backups, Archive Storage, Semi-structured Data

Security

Performance

Scale and cost effectiveness

Common Blob storage foundation

Common SDK, Tools, Control Plane

Object Tiering and Lifecycle Policy Management

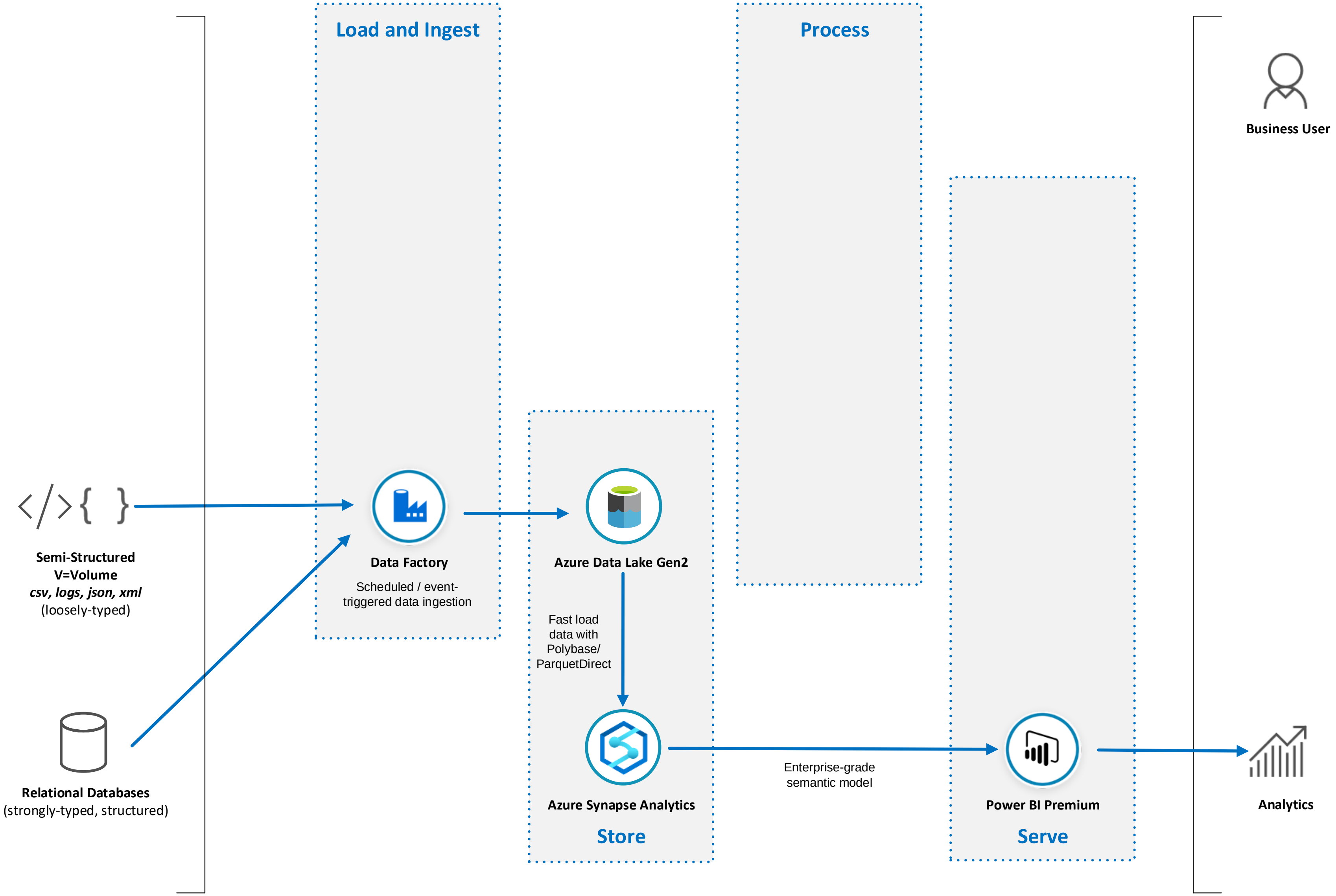
AAD integration, RBAC, Storage account security

HA/DR support through ZRS and RA-GRS

# Lab 2: Transform Big Data using Azure Data Factory Mapping Data Flows

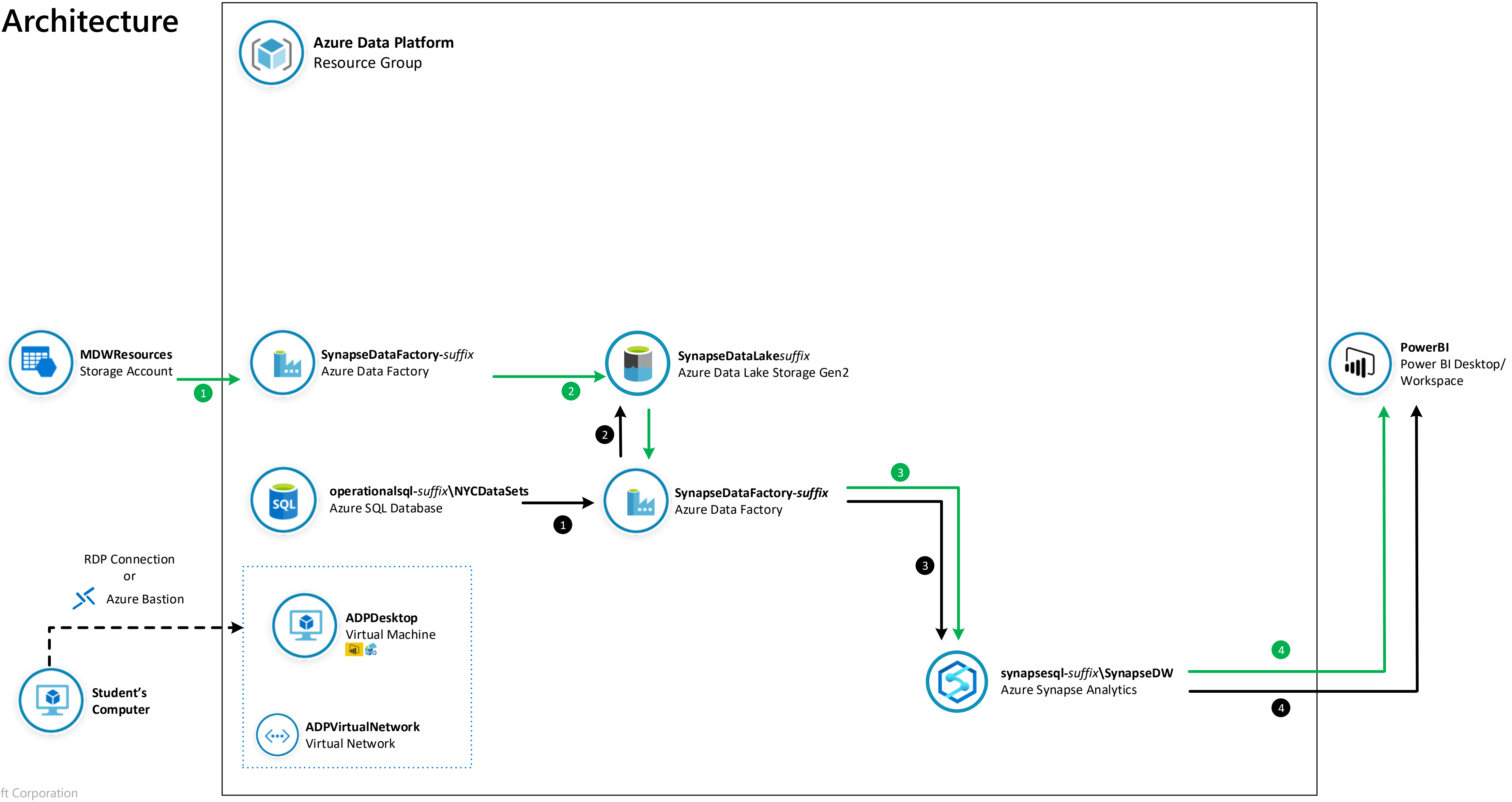
# Lab 2

## Transform Big Data using Azure Data Factory Mapping Data Flows



# Lab 2

## Lab Architecture



# Advanced Analytics

# Advanced analytics

**Advanced analytics goes beyond the traditional business intelligence (BI) and uses mathematical, probabilistic, and statistical modeling techniques to enable predictive processing and automated decision making.**

## Modern data warehousing

"We want to integrate all our data—including Big Data—with our data warehouse"

## Advanced analytics

"We're trying to predict when our customers churn"

## Real-time analytics

"We're trying to get insights from our devices in real-time"

# Modern Data Platform Concepts

## Part III

# Hadoop and Spark in Azure

## Open Source Apache Projects for Big Data Compute

| hadoop | APACHE Spark |
|---|---|
| It was the original open-source framework for distributed processing and analysis of big data sets on clusters. | Effective, fast, general-purpose unified cluster computing framework with high-level APIs in Java, Scala, Python and R. |
| Read/write from disk. | In-memory processing. |
| Economical batch mode. | Fast, interactive data processing. |
| Linear processing of huge datasets. | Streaming and Machine Learning Support |

## Azure HDInsight is a managed, full-spectrum, open-source analytics service for enterprises

What comes with HDInsight?

Apache Hadoop

Apache Spark

Apache Kafka

Apache HBase

Apache Hive LLAP

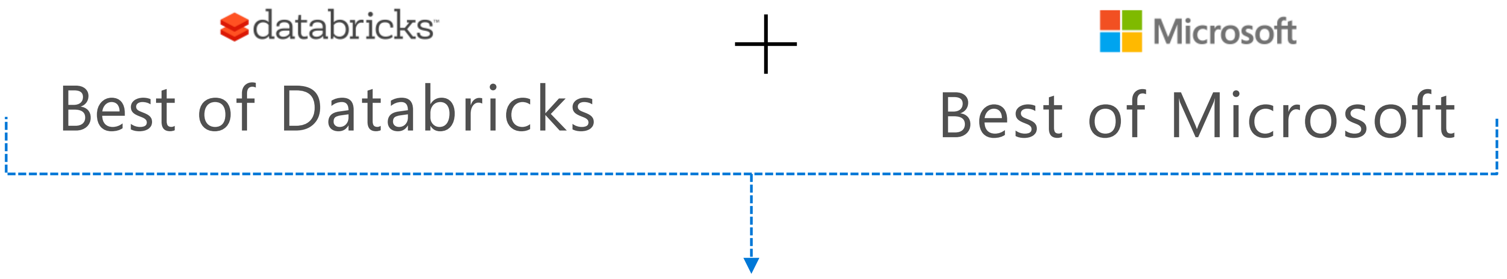
Apache Storm

Machine Learning

# Azure Databricks

# Azure Databricks

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



databricks

Best of Databricks

+

Microsoft

Best of Microsoft

Spark Designed in collaboration with the founders of Apache Spark

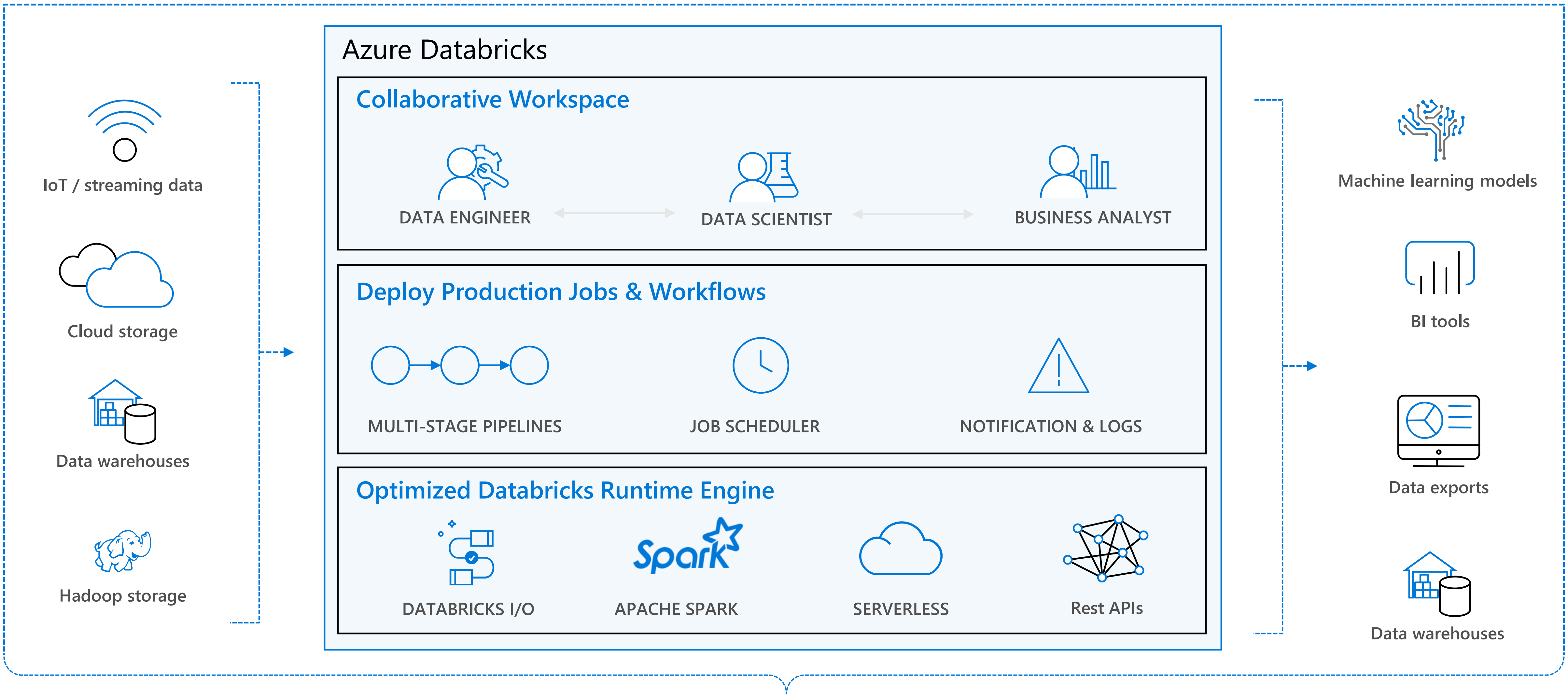
One-click set up; streamlined workflows

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

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

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

# Azure Databricks



IoT / streaming data

Cloud storage

Data warehouses

Hadoop storage

Azure Databricks

## Collaborative Workspace

DATA ENGINEER ↔ DATA SCIENTIST ↔ BUSINESS ANALYST

## Deploy Production Jobs & Workflows

MULTI-STAGE PIPELINES

JOB SCHEDULER

NOTIFICATION & LOGS

## Optimized Databricks Runtime Engine

DATABRICKS I/O

APACHE SPARK

SERVERLESS

Rest APIs

Machine learning models

BI tools

Data exports

Data warehouses

Enhance Productivity

Build on secure & trusted cloud

Scale without limits

# Azure Databricks Notebooks

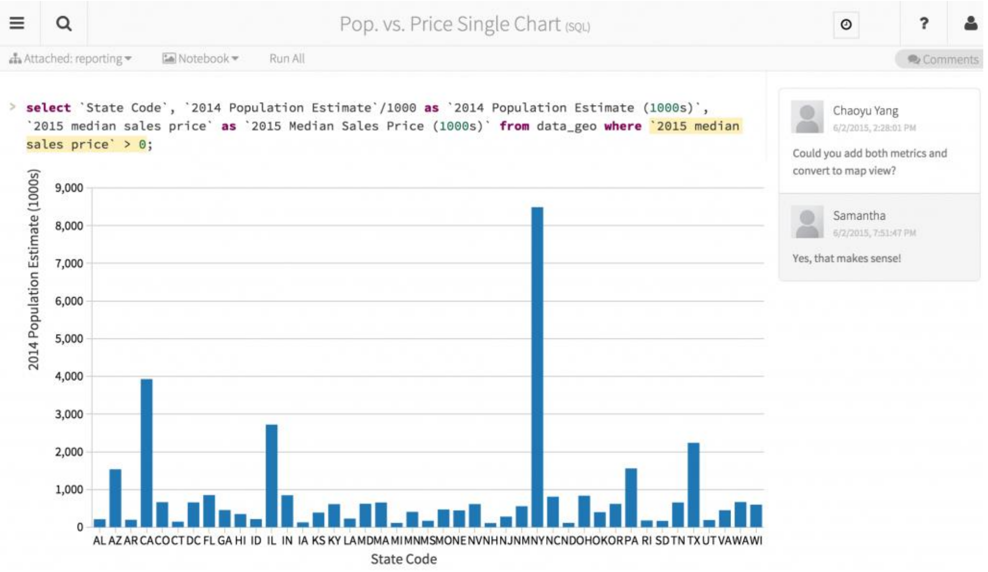
## Notebooks are a popular way to develop, and run, Spark Applications

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

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

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

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


☰
Pop. vs. Price Single Chart (SQL)
?
Attached: reporting ▾
Notebook ▾
Run All
Comments
2014 Population Estimate (1000s)
9,000
8,000
7,000
6,000
5,000
4,000
3,000
2,000
1,000
0
AL AZ AR CA CO CT DC FL GA HI ID IL IN IA KS KY LA MD MA MI MN MS MO NE NV NH NJ NM NY NC ND OH OK OR PA RI SD TN TX UT VA WA WI
State Code

```
> select `State Code`, `2014 Population Estimate`/1000 as `2014 Population Estimate (1000s)`,
`2015 median sales price` as `2015 Median Sales Price (1000s)` from data_geo where `2015 median
sales price` > 0;
```

Chaoyu Yang
6/2/2015, 2:28:01 PM
Could you add both metrics and convert to map view?

Samantha
6/2/2015, 7:51:47 PM
Yes, that makes sense!

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

`%python` Allows you to execute python code in a notebook (even if that notebook is not python)

`%sql` Allows you to execute sql code in a notebook (even if that notebook is not sql).

`%r` Allows you to execute r code in a notebook (even if that notebook is not r).

`%scala` Allows you to execute scala code in a notebook (even if that notebook is not scala).

`%sh` Allows you to execute shell code in your notebook.

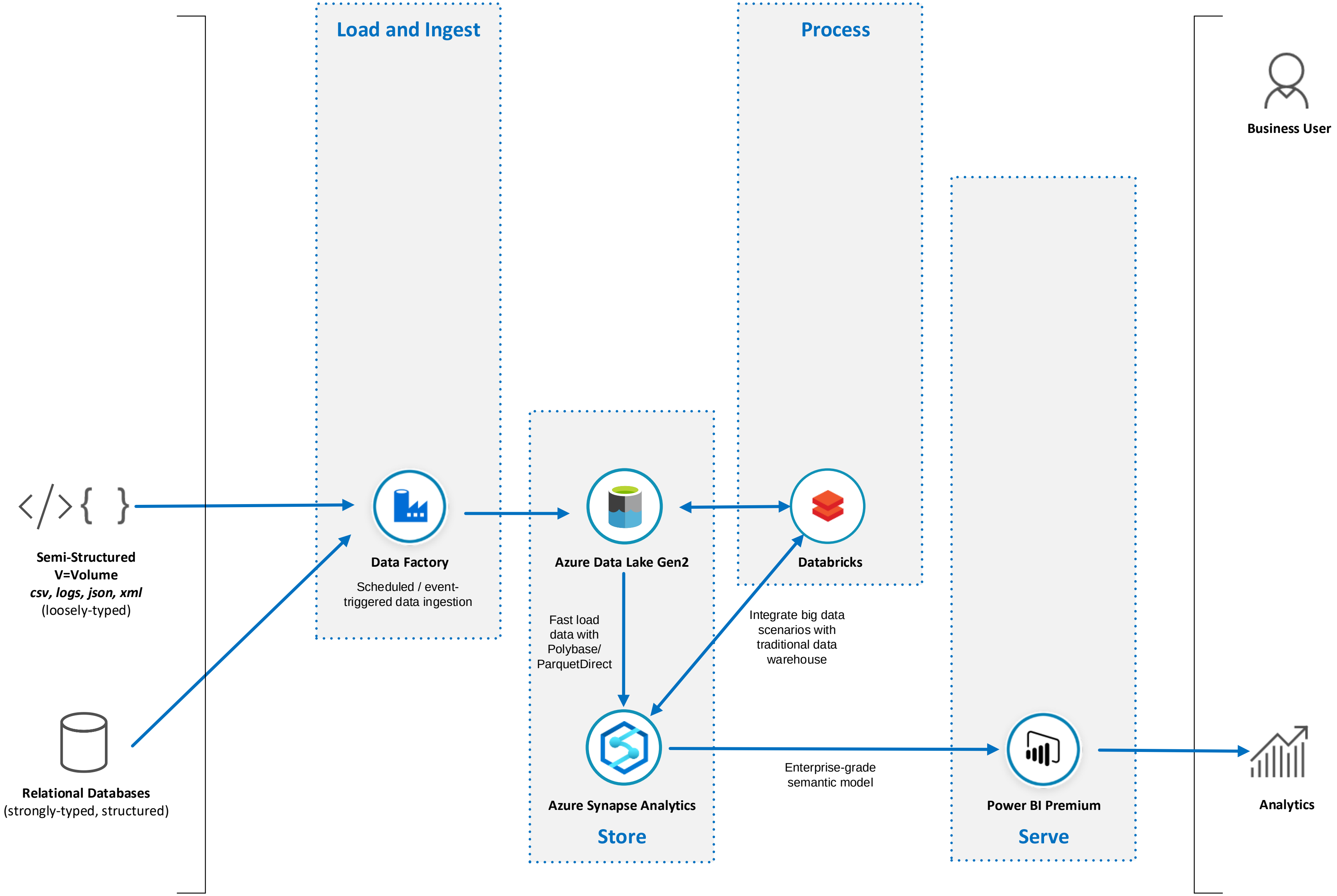
`%fs` Allows you to use Databricks Utilities - dbutils filesystem commands.

`%md` To include rendered markdown

# Lab 3: Explore Big Data with Azure Databricks

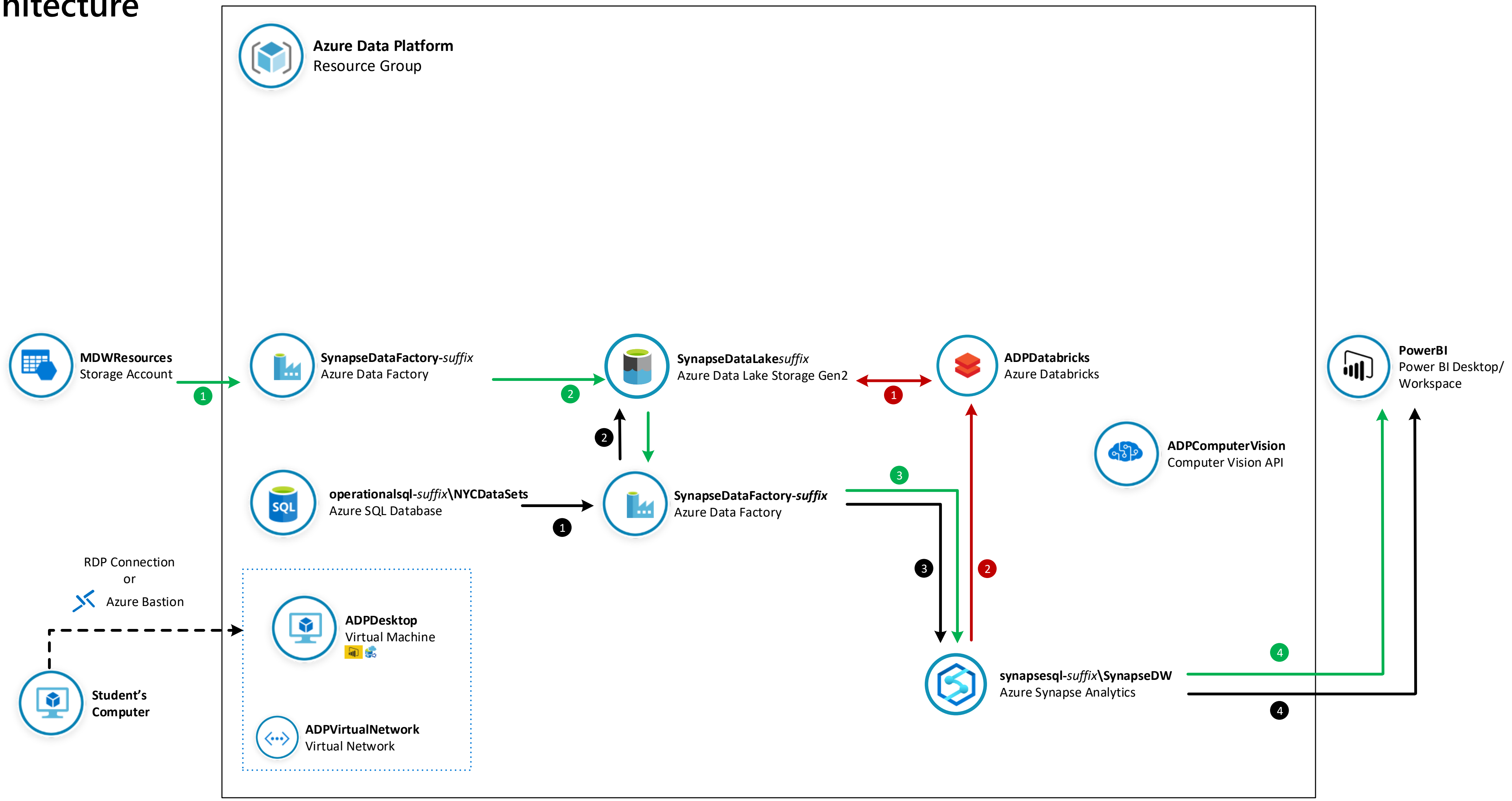
# Lab 3

## Explore Big Data with Azure Databricks



# Lab 3

## Lab Architecture

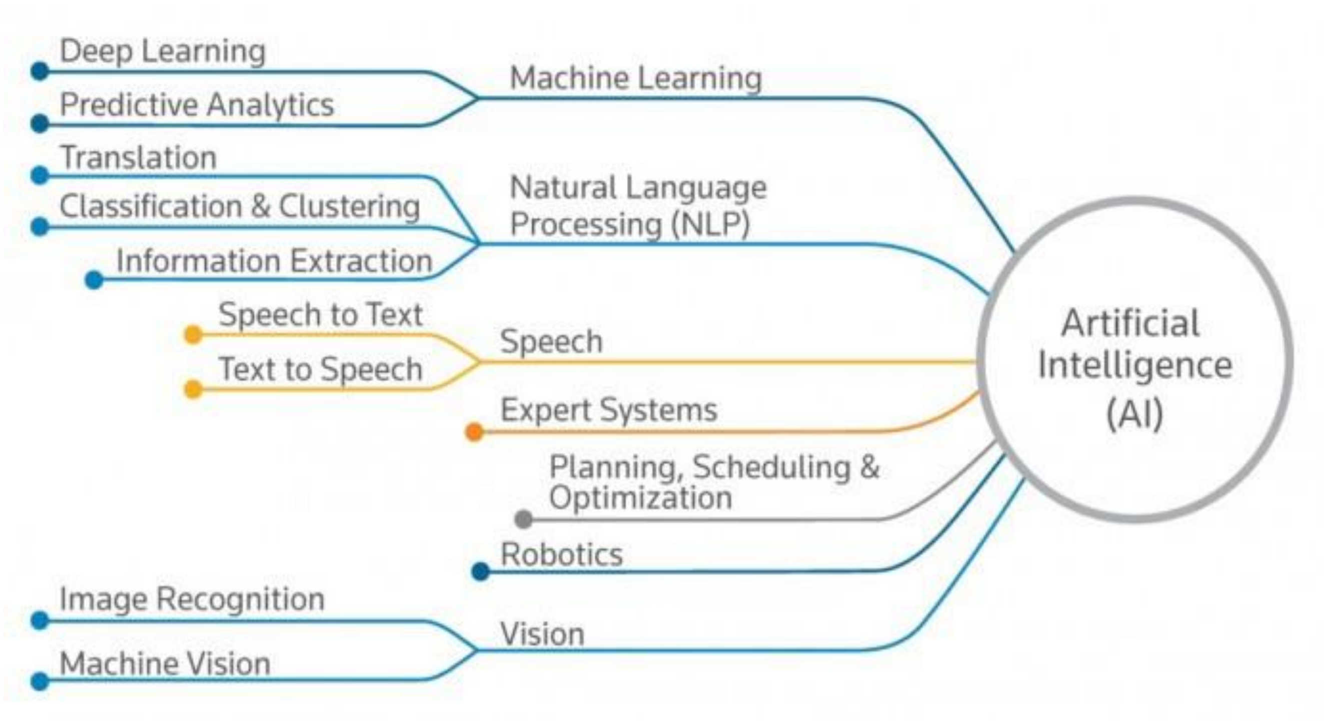


# Modern Data Platform Concepts

## Part IV

# Artificial Intelligence

**“The ability of a digital computer or computer-controlled robot to perform tasks commonly associated with intelligent beings.” – Encyclopedia Britannica**



## Machine Learning

**Supervised Learning**

- Regression
- Classification

**Unsupervised Learning**

- Cluster Analysis

**Application Examples**

- Weather Forecast
- Fraud Detection
- Customer Churn
- Insurance Premium

# What’s No-SQL?

**Term coined in 2009 for a developer meetup – ”Not Only SQL” -> “NoSQL”.**

**Databases that allow you to store and retrieve data in various structures, formats, and models other than tabular relational model.**

## There’s a time and a place for everything

Sometimes a relational store is the right choice

Sometimes a NoSQL store is the right choice

Sometimes you need more than one store for an app -> polyglot persistence

## Data Structures

**Key-Value Databases**

Cosmos DB, Redis Cache, Azure Table

**Column Family Stores**

Cosmos DB, Cassandra, HBase

**Graph Databases**

Cosmos DB, Neo4j, Gremlin

**Document Databases**

Cosmos DB, MongoDB

# Azure AI

# Azure AI

## Solution Areas

### AI apps and agents

Azure Cognitive Services

Azure Bot Service

### Knowledge mining

Azure Search

### Machine learning

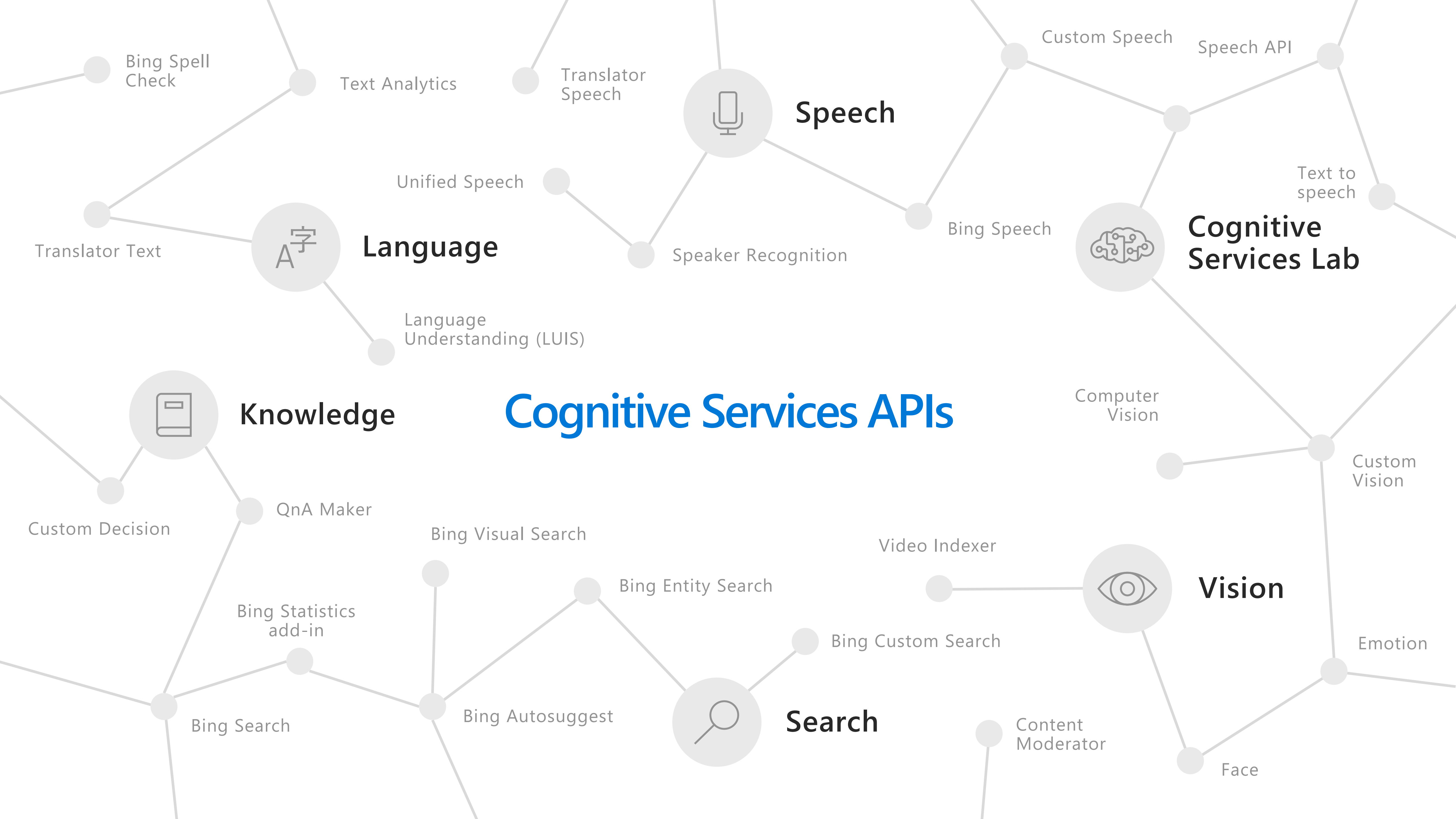
Azure Databricks

Azure Machine Learning

Azure AI Infrastructure

Productive Built for enterprises Trusted

# Cognitive Services APIs



Bing Spell Check

Text Analytics

Translator Speech

Speech

Custom Speech

Speech API

Unified Speech

Translator Text

Language

Speaker Recognition

Bing Speech

Text to speech

Cognitive Services Lab

Language Understanding (LUIS)

Knowledge

Computer Vision

Custom Vision

Custom Decision

QnA Maker

Bing Visual Search

Video Indexer

Bing Entity Search

Vision

Bing Statistics add-in

Bing Custom Search

Emotion

Bing Search

Bing Autosuggest

Search

Content Moderator

Face

# Cognitive Services capabilities

## Infuse your apps, websites, and bots with human-like intelligence

### Vision

Object, scene, and activity detection

Face recognition and identification

Celebrity and landmark recognition

Emotion recognition

Text and handwriting recognition (OCR)

Customizable image recognition

Video metadata, audio, and keyframe extraction and analysis

Explicit or offensive content moderation

### Speech

Speech transcription (speech-to-text)

Custom speech models for unique vocabularies or complex environment

Text-to-speech

Custom Voice

Real-time speech translation

Customizable speech transcription and translation

Speaker identification and verification

### Language

Language detection

Named entity recognition

Key phrase extraction

Text sentiment analysis

Multilingual and contextual spell checking

Explicit or offensive text content moderation

PII detection for text moderation

Text translation

Customizable text translation

Contextual language understanding

### Knowledge

Q&A extraction from unstructured text

Knowledge base creation from collections of Q&As

Semantic matching for knowledge bases

Customizable content personalization learning

### Search

Ad-free web, news, image, and video search results

Trends for video, news

Image identification, classification and knowledge extraction

Identification of similar images and products

Named entity recognition and classification

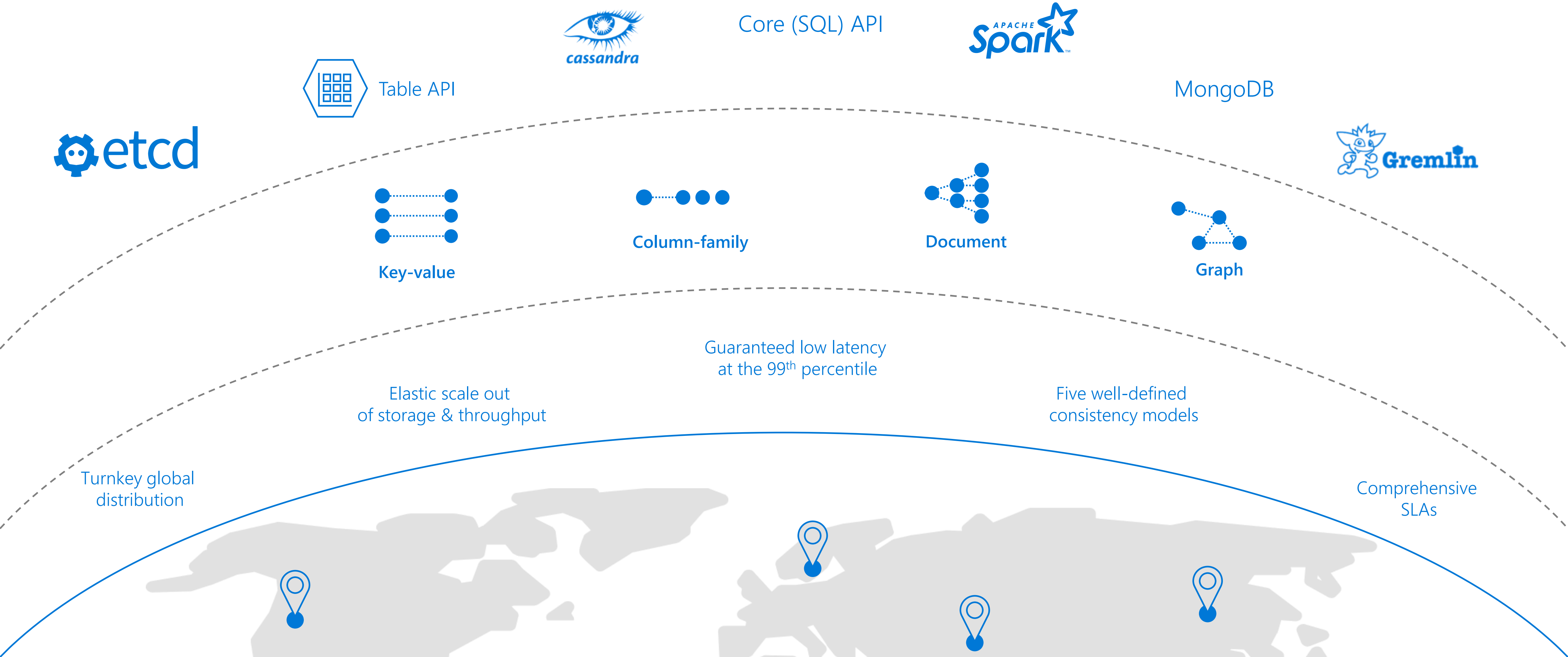
Knowledge acquisition for named entities

Search query autosuggest

Ad-free custom search engine creation

# Cosmos DB

# Azure Cosmos DB



Guaranteed low latency at the 99th percentile

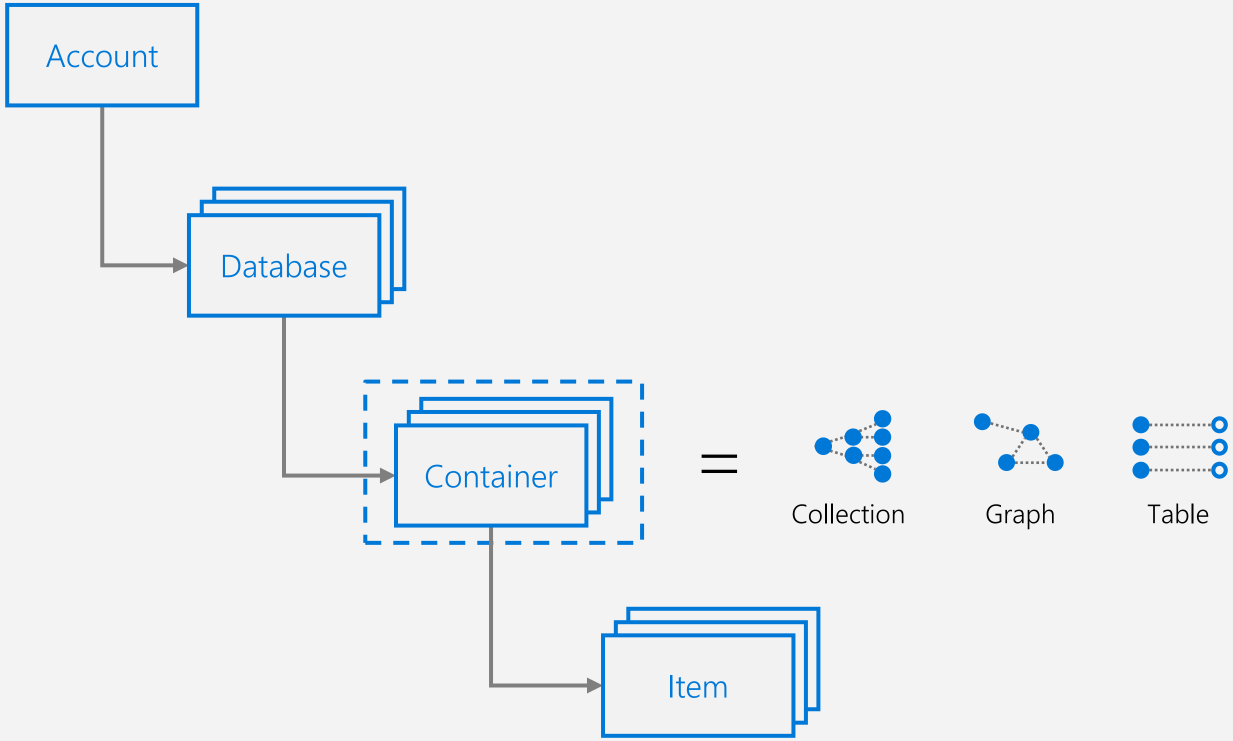
Elastic scale out of storage & throughput

Five well-defined consistency models

Turnkey global distribution

Comprehensive SLAs

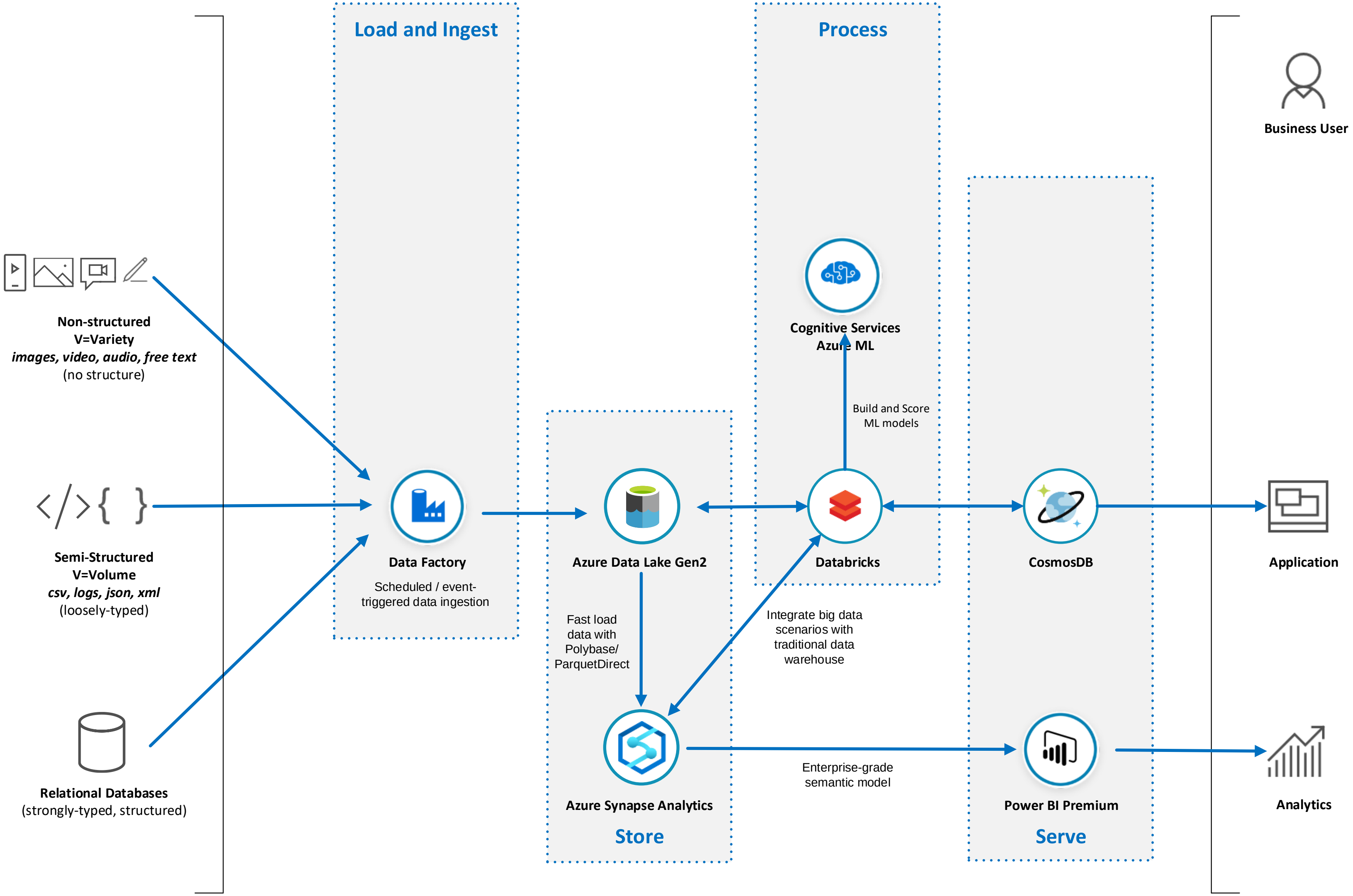
# RESOURCE MODEL



# Lab 4: Add AI to your Big Data Pipeline with Cognitive Services

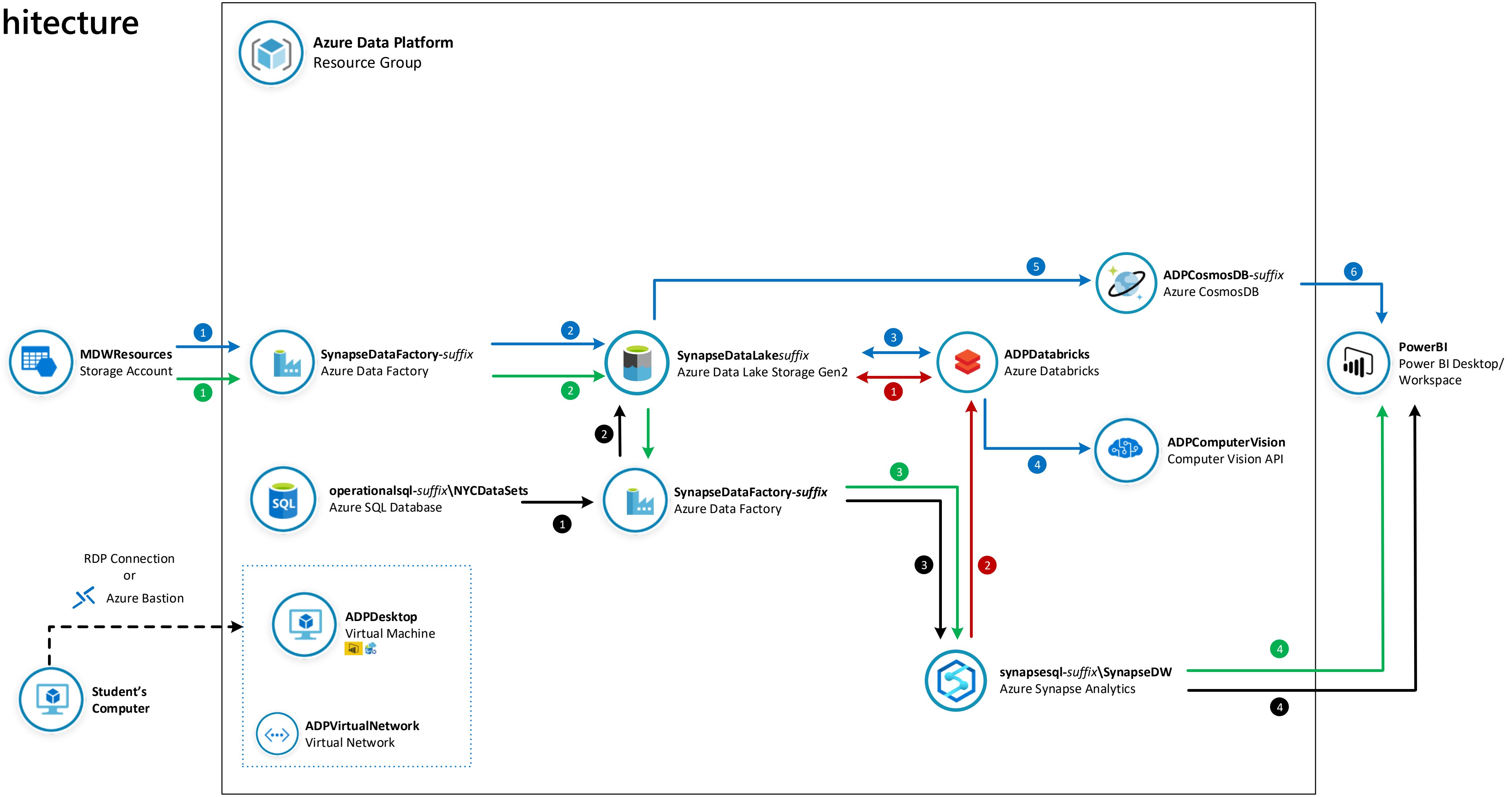
# Lab 4

## Add AI to your Big Data Pipeline with Cognitive Services



# Lab 4

## Lab Architecture



© Microsoft Corporation

# Real-time Analytics

# Real-time analytics

**Deals with streams of data that are captured in real-time and processed with minimal latency to generate real-time (or near-real-time) reports or automated responses.**

## Modern data warehousing

“We want to integrate all our data—including Big Data—with our data warehouse”

## Advanced analytics

“We’re trying to predict when our customers churn”

## Real-time analytics

“We’re trying to get insights from our devices in real-time”

# Modern Data Platform Concepts

## Part V

# Streaming Use Cases

## Retail

**CONSUMER ENGAGEMENT**

### Real-time Pricing Optimization

- Demand-Elasticity
- Personal Pricing Schemes
- Promotion events
- Multi-channel engagement

## Financial

**RISK AND REVENUE MANAGEMENT**

### Risk and Fraud, Threat Detection

- Real-time anomaly detection
- Card Monitoring and Fraud Detection
- Risk Aggregation

## Oil/Gas & Energy

**GRID OPS, ASSET OPTIMIZATION**

### Industrial IoT

- Preventive Maintenance
- Smart Grids and Microgrids
- Asset performance as a Service
- UAV image analysis

## Security

**ACTIONABLE THREAT INTELLIGENCE**

### Security Intelligence

- Real-time firewall, network, and auth log correlation
- Anomaly detection
- Security context, enrichment
- Security Orchestration

## Healthcare

**SENSOR DATA**

### IoT DEVICE ANALYTICS

- Aggregation of streaming events
- Predictive Maintenance
- Anomaly Detection

## Advertising

**RECOMMENDATION ENGINE**

### Next Best and Personalized Offers

- Right product, promotion, at right time
- Real time Ad bidding platform
- Personalized Ad Targeting

## Media Entertainment

**CONSUMER ENGAGEMENT ANALYSIS**

### Sentiment Analysis

- Demand-Elasticity
- Social Network Analysis
- Promotion events
- Multi-channel Attribution

## And Much More!

# Unlocking Real-time Insights

## Time to Insight is Critical

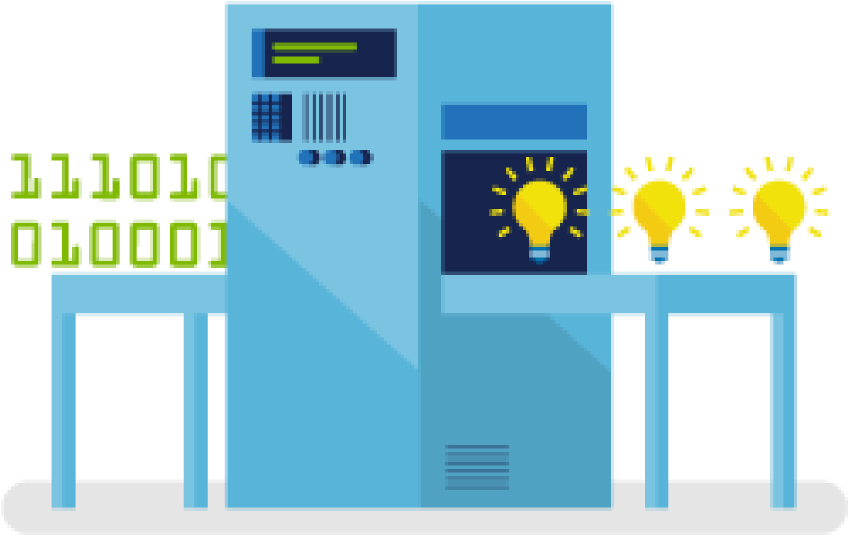
Reducing decision latency can unlock business value

## Insights are Perishable

Window of opportunity for insights to be actionable

## Ask Questions to Data in Motion

Can’t wait for data to get to rest before running computation



# Scenario Types

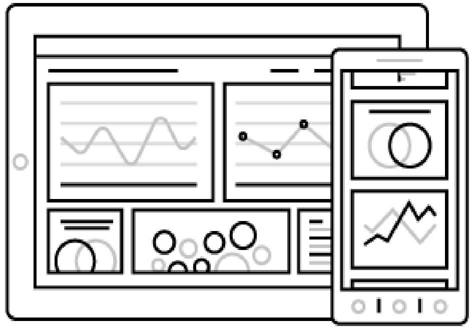
## Actions by Human Actors

- “See and seize” insights
- Live visualization
- Alerts and alarms
- Dynamic aggregation

## Machine to Machine Interactions

- Data movement with enrichment
- Kick-off workflows for automation

Dashboarding

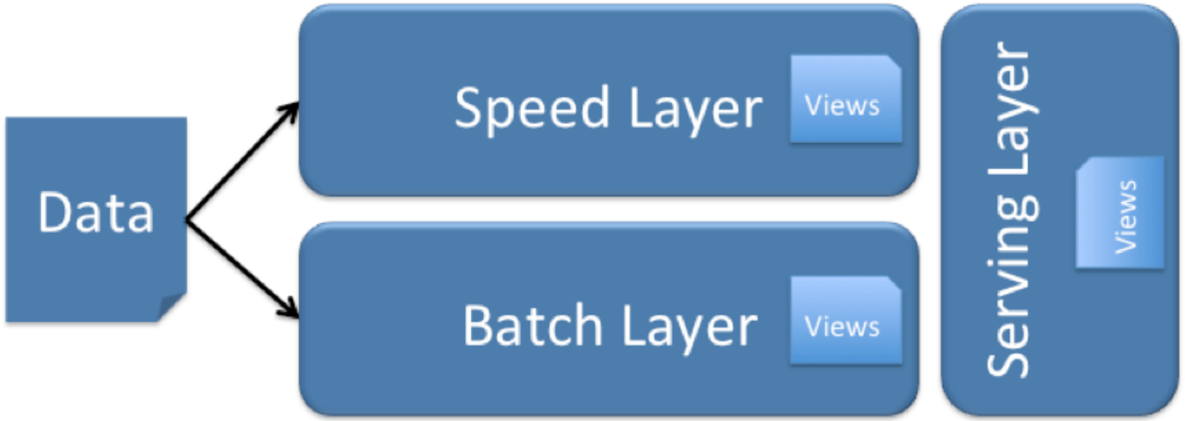


Automation

Enriched Data Movement

# Lambda (λ) Architecture

**Designed to handle Big Data use cases by taking advantage of both batch and stream-processing methods**



1. All **data** entering the system is dispatched to both the batch layer and the speed layer for processing.

2. The **batch layer** has two functions:
   I. manage the master dataset (an immutable, append-only set of raw data)
   II. pre-compute the batch views.

3. The **serving layer** indexes the batch views so that they can be queried in low-latency, ad-hoc way.

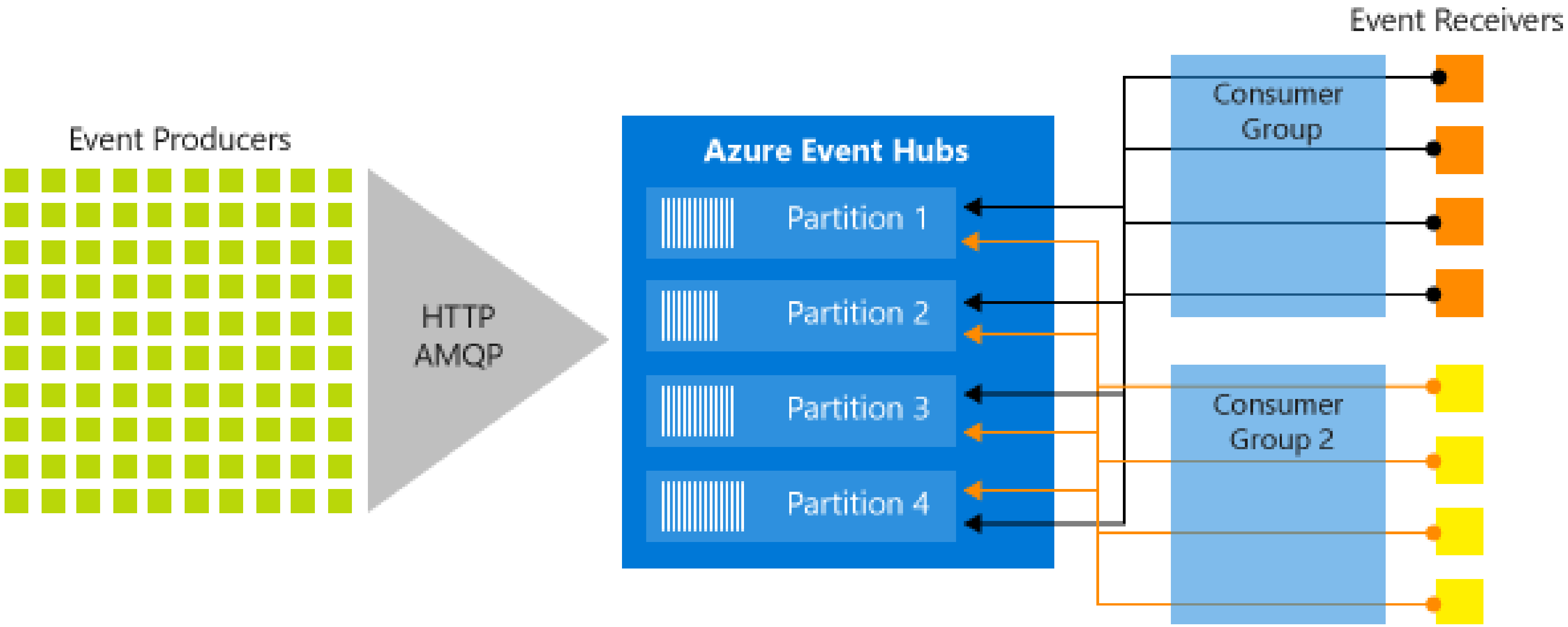
4. The **speed layer** compensates for the high latency of updates to the serving layer and deals with recent data only.

5. Any incoming **query** can be answered by merging results from batch views and real-time

# Event Hubs

# Event Hubs

**Big data streaming platform and event ingestion service capable of receiving and processing millions of events per second.**



# Event Hubs Capture

## Batch on stream

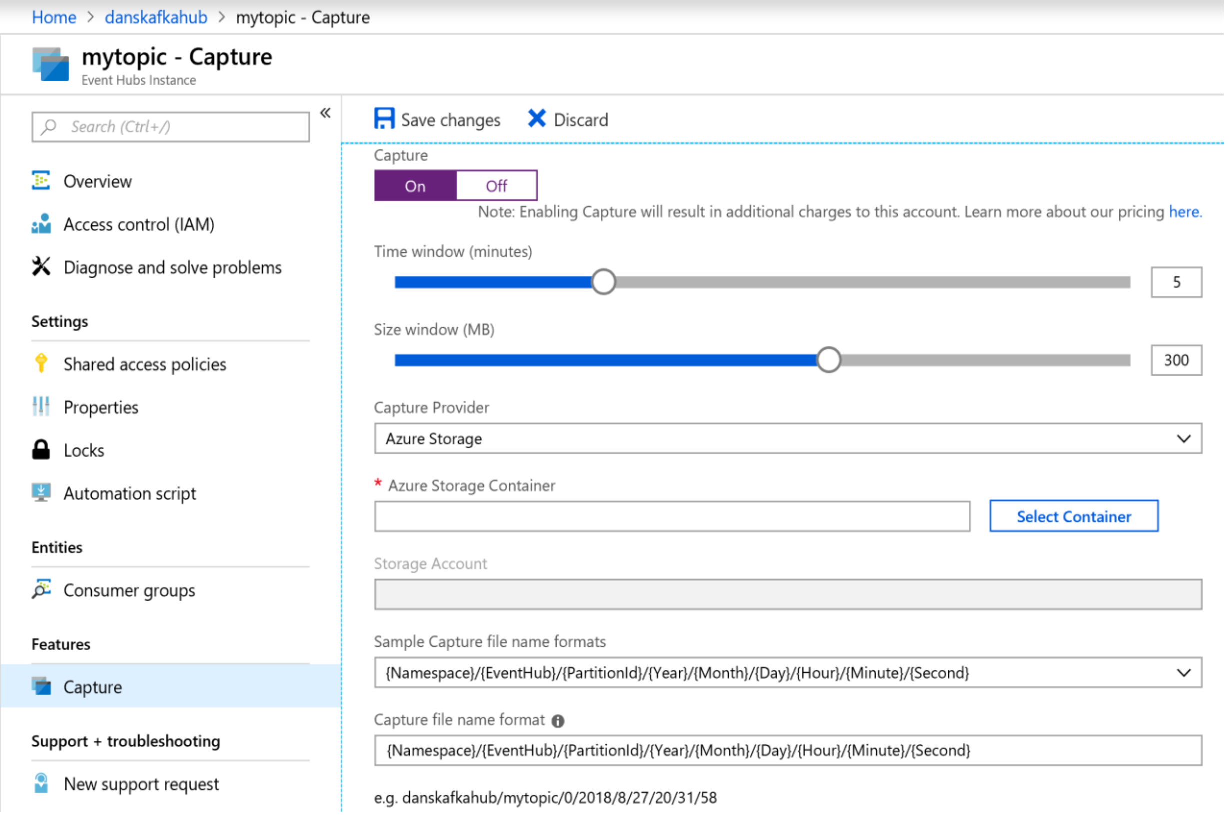
Policy based push to your own storage

Uses Avro format

Raises Event Grid events – connect to Functions, ACI, or whatever you like

Does not impact throughput

Offloads batch processing from your real-time stream



Capture: On | Off

Note: Enabling Capture will result in additional charges to this account. Learn more about our pricing here.

Time window (minutes): 5

Size window (MB): 300

Capture Provider: Azure Storage

* Azure Storage Container:

Storage Account:

Sample Capture file name formats: {Namespace}/{EventHub}/{PartitionId}/{Year}/{Month}/{Day}/{Hour}/{Minute}/{Second}

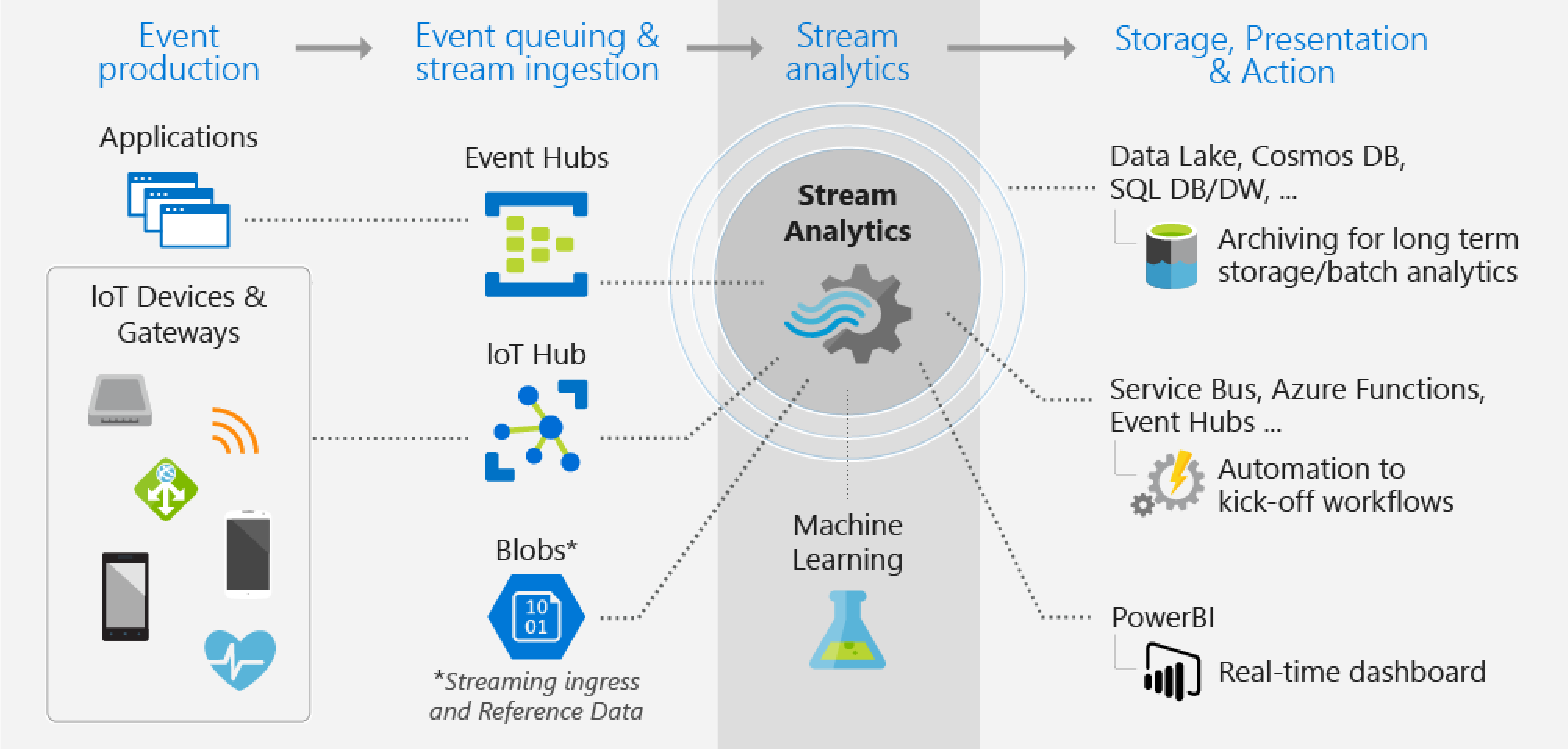
Capture file name format: {Namespace}/{EventHub}/{PartitionId}/{Year}/{Month}/{Day}/{Hour}/{Minute}/{Second}

e.g. danskafkahub/mytopic/0/2018/8/27/20/31/58

# Stream Analytics

# Stream Analytics

Event-processing engine that allows you to examine high volumes of data streaming from devices


Event production
Event queuing & stream ingestion
Stream analytics
Storage, Presentation & Action
Applications
IoT Devices & Gateways
Event Hubs
IoT Hub
Blobs*
*Streaming ingress and Reference Data
Stream Analytics
Machine Learning
Data Lake, Cosmos DB, SQL DB/DW, ...
Archiving for long term storage/batch analytics
Service Bus, Azure Functions, Event Hubs ...
Automation to kick-off workflows
PowerBI
Real-time dashboard

# Stream Analytics Job

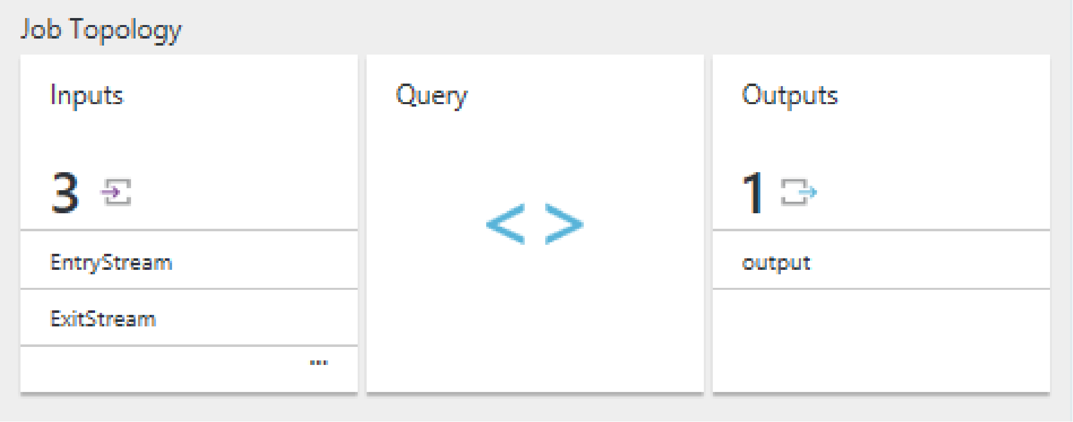
## Users construct and deploy jobs to Azure Stream Analytics

## Job definition includes inputs, a query, and output

Inputs are from where the job reads the data stream

Query runs for perpetuity unless explicitly stopped and transforms the input stream

Output is where the job sends the job results to



Job Topology

Inputs

3

EntryStream

ExitStream

...

Query

< >

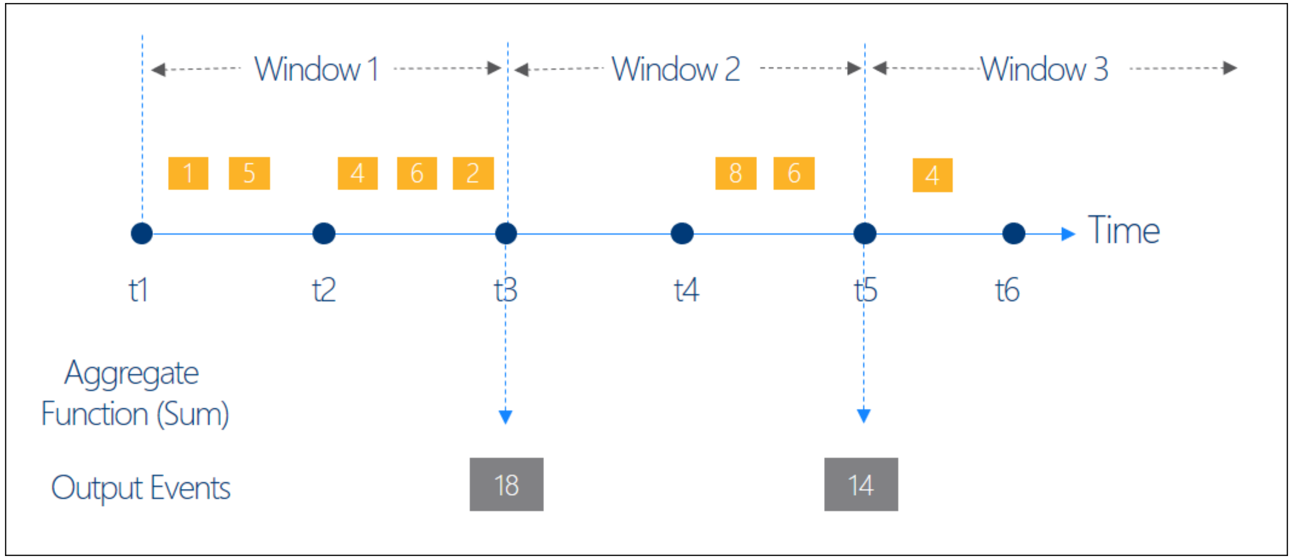
Outputs

1

output

# Windowing Concepts

- Operations on the data contained in temporal windows is a common pattern
- Four types of Temporal Windows:
  - Sliding
  - Tumbling
  - Hopping
  - Session
- Output at the end of each window
- Windows are fixed length
- Used in a GROUP BY clause

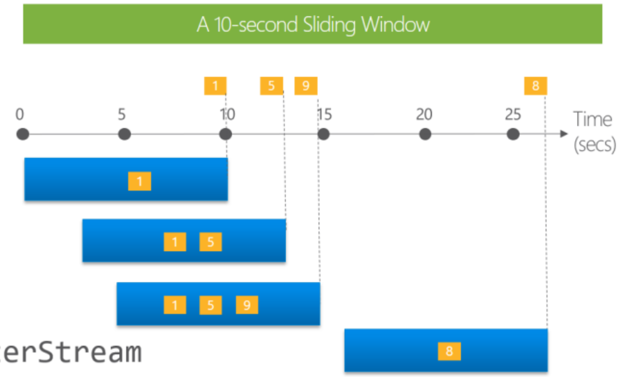


# Windowing Functions

## Sliding Windows and Tumbling Windows

### Sliding Windows

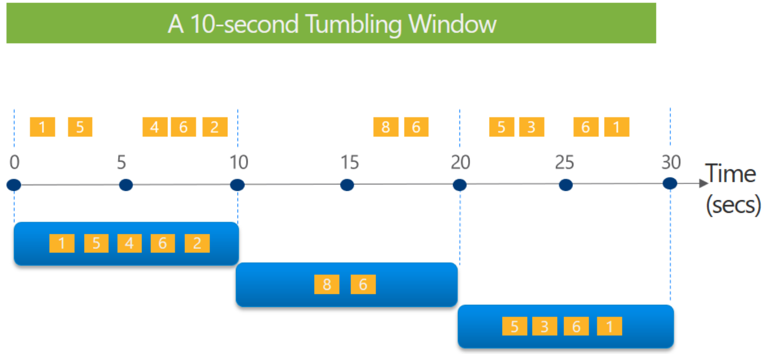
Give me the count of tweets for all topics which are tweeted more than 10 times in the last 10 seconds



```
SELECT Topic, COUNT(*) FROM TwitterStream
TIMESTAMP BY CreatedAt
GROUP BY Topic, SlidingWindow(second, 10)
HAVING COUNT(*) > 10
```

### Tumbling Windows

Tell me the count of tweets per time zone every 10 seconds



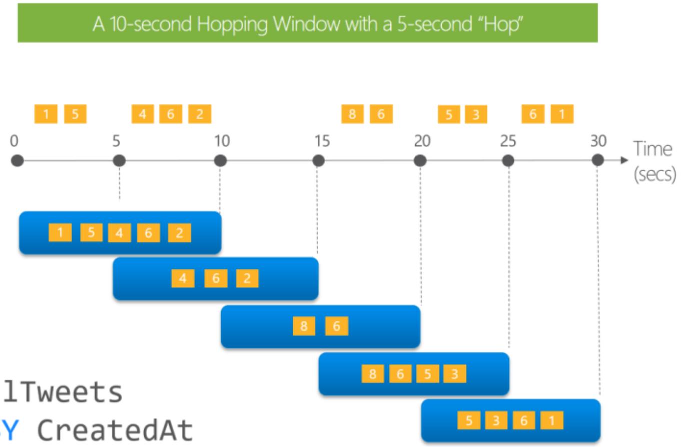
```
SELECT TimeZone, COUNT(*) AS Count
FROM TwitterStream TIMESTAMP BY CreatedAt
GROUP BY TimeZone, TumblingWindow(second,10)
```

# Windowing Functions

## Hopping Windows and Session Windows

### Hopping Windows

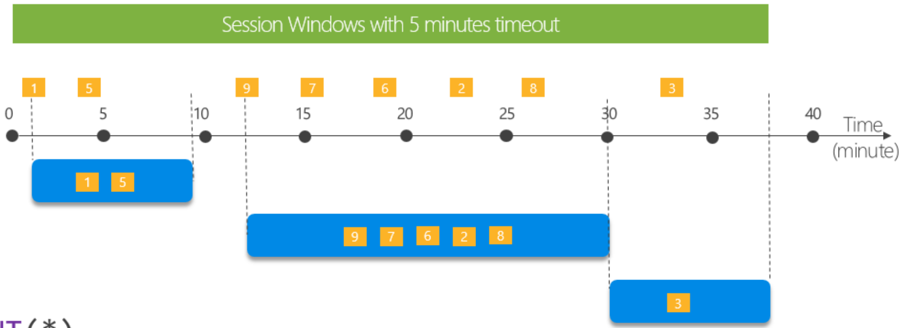
Every 5 seconds give me the count of tweets over the last 10 seconds



```
SELECT Topic, COUNT(*) AS TotalTweets
FROM TwitterStream TIMESTAMP BY CreatedAt
GROUP BY Topic, HoppingWindow(second, 10 , 5)
```

### Session Windows

Tell me the count of tweets that occur within 5 minutes to each other.

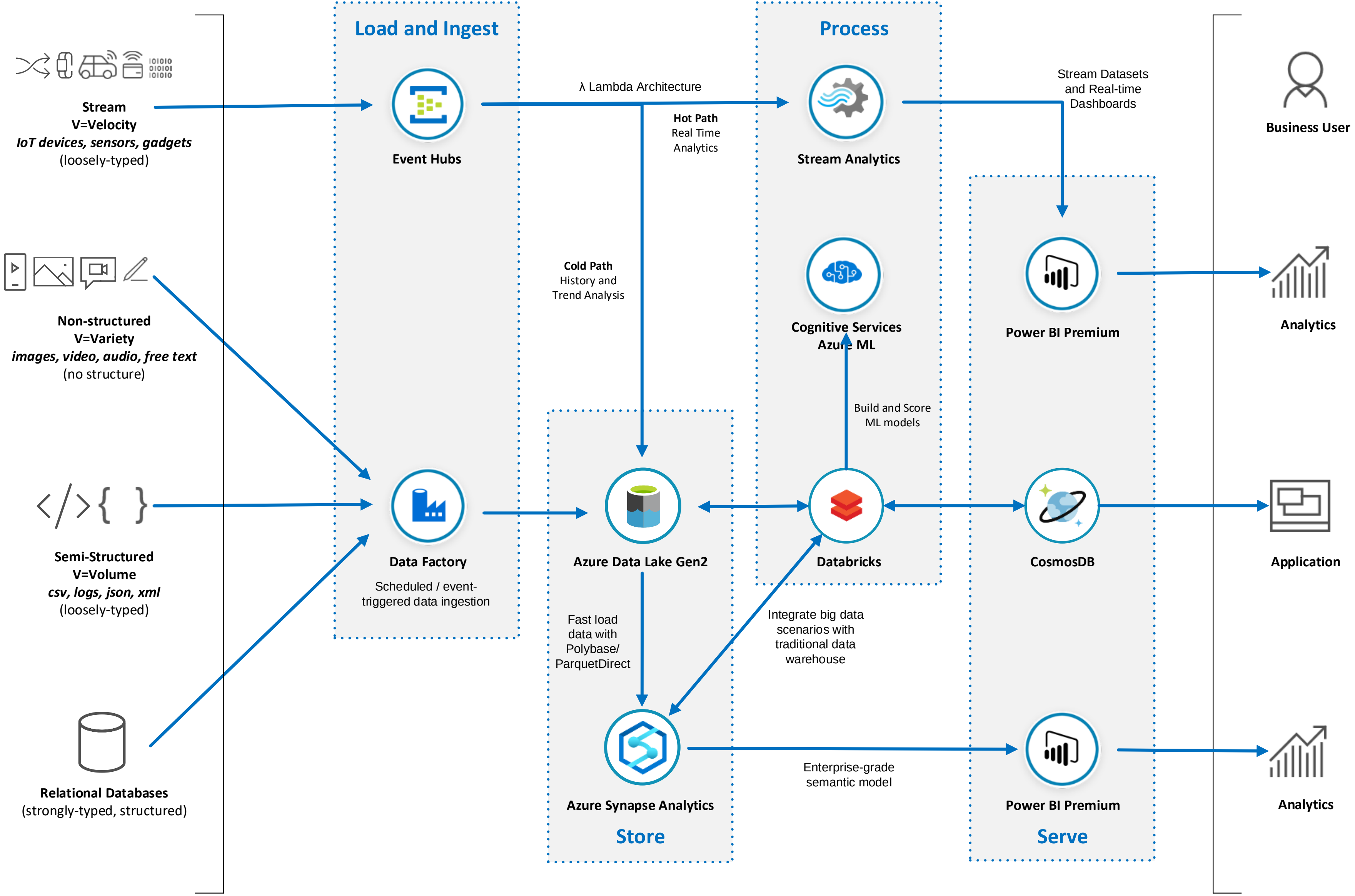


```
SELECT Topic, COUNT(*)
FROM TwitterStream TIMESTAMP BY CreatedAt
GROUP BY Topic, SessionWindow(minute, 5, 10)
```

# Lab 5: Ingest and Analyse real-time data with Event Hubs and Stream Analytics

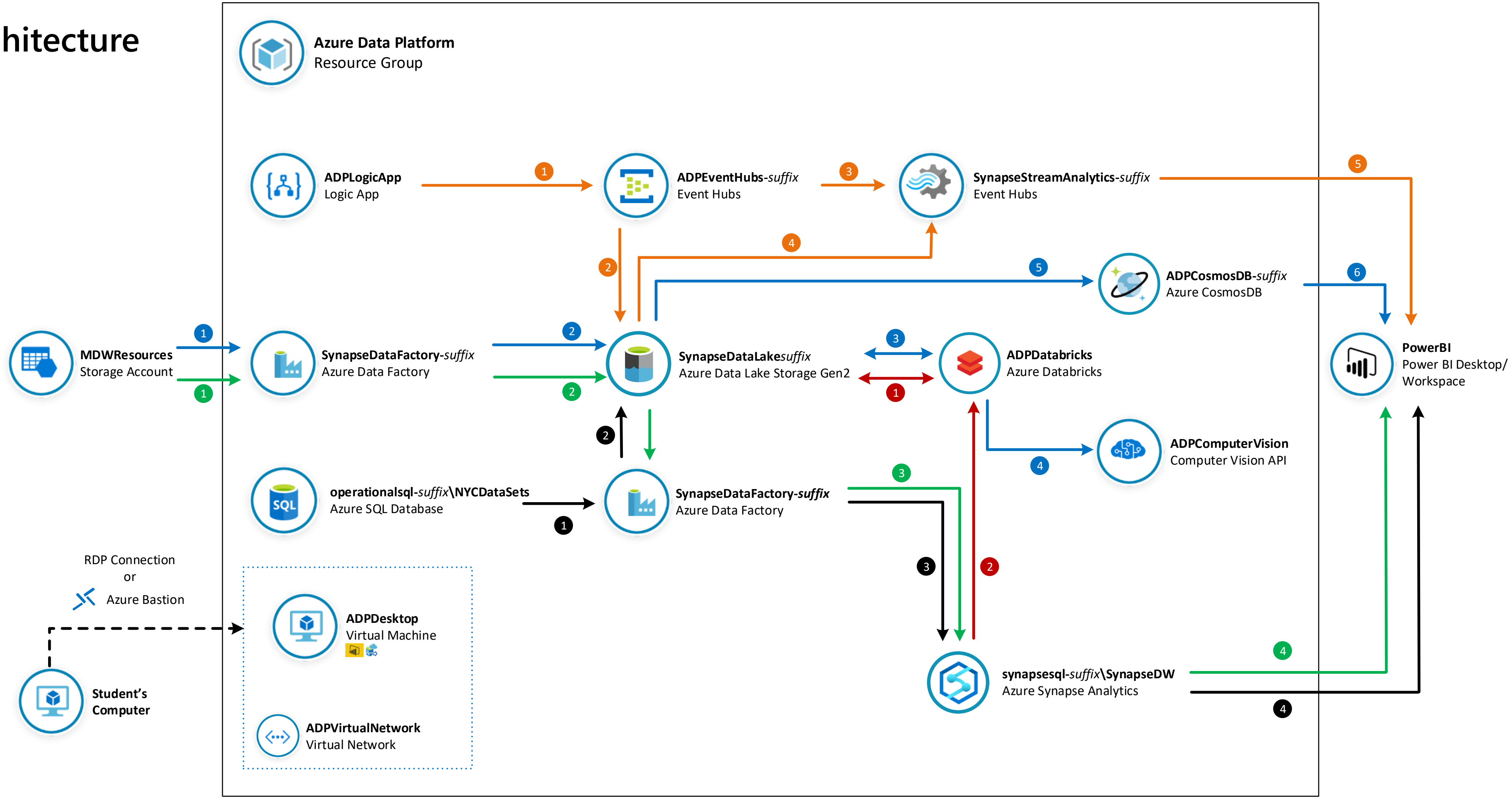
# Lab 5

## Ingest and Analyse real-time data with Event Hubs and Stream Analytics



# Lab 5

## Lab Architecture



# It’s all on

Microsoft Azure

Microsoft