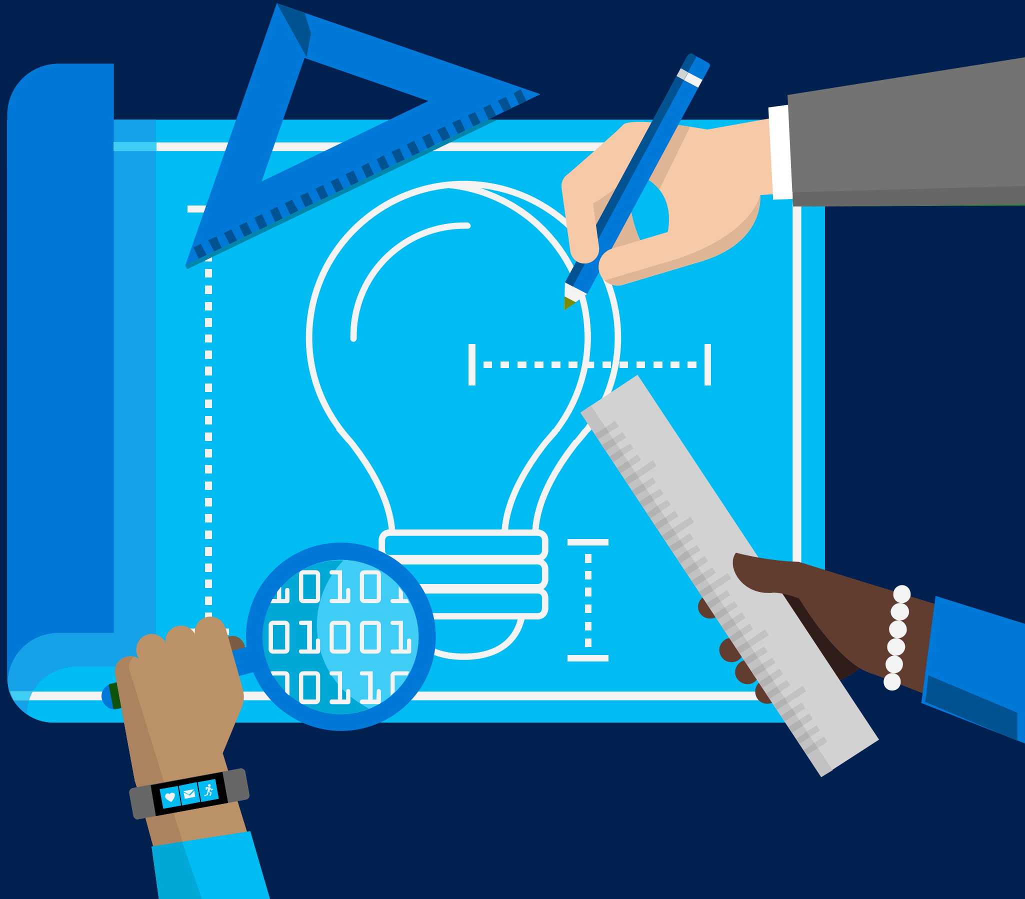
Microsoft

Big Data and Visualization Briefing Deck

December 6, 2017

Hong Kong



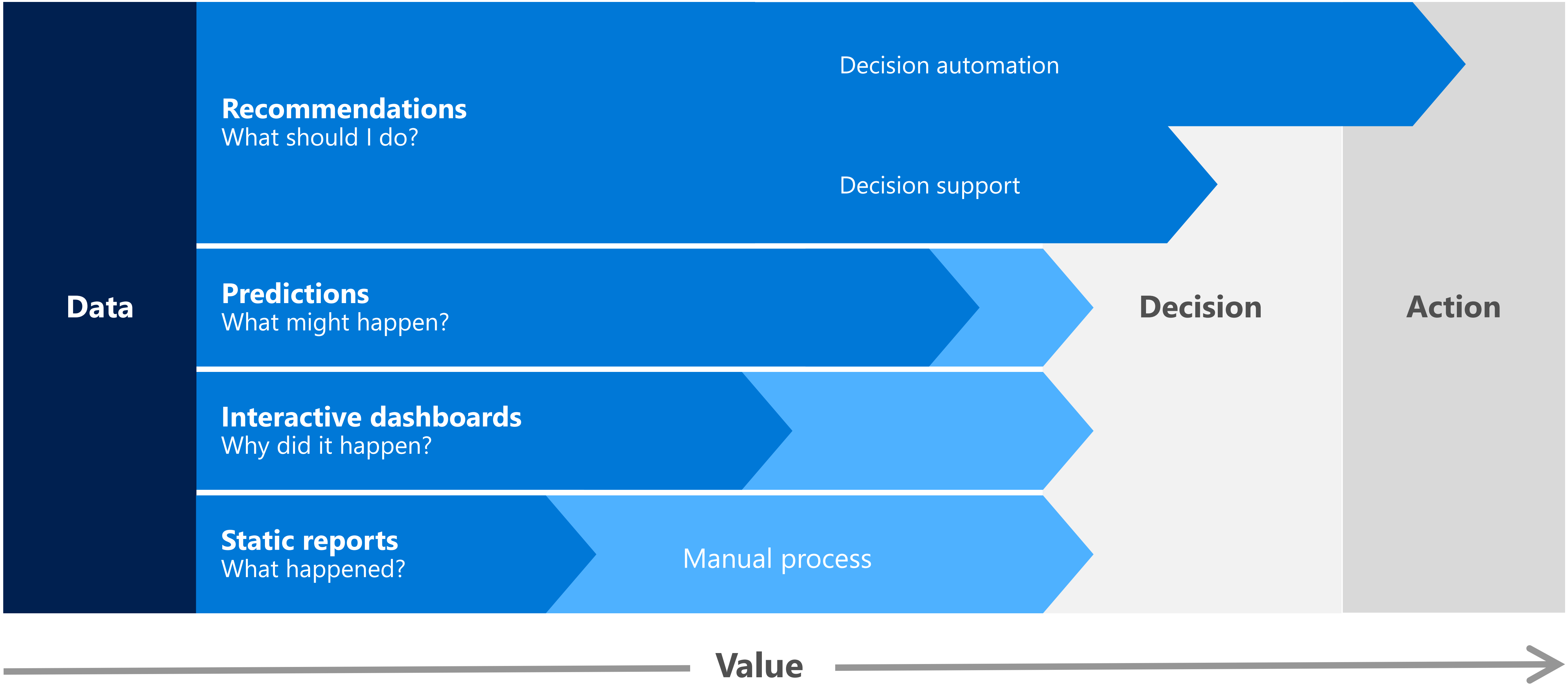
Microsoft

Azure Big Data
Patterns and Tools

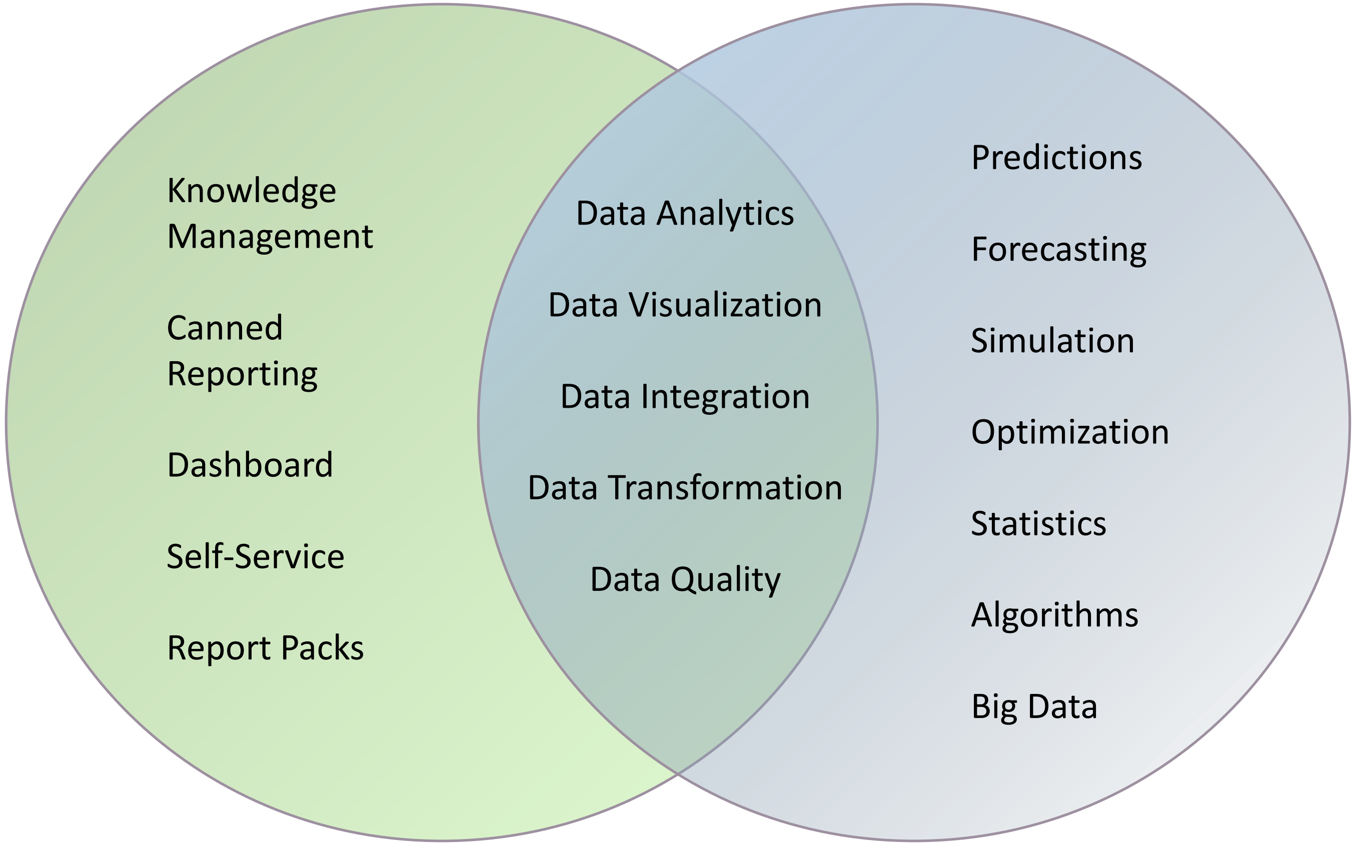
Julius Chen

Global Technology Manager
Greater Asia

From data to decisions and actions



Traditional BI vs Data Science



Business Intelligence

Knowledge Management

Canned Reporting

Dashboard

Self-Service

Report Packs

Data Analytics

Data Visualization

Data Integration

Data Transformation

Data Quality

Predictions

Forecasting

Simulation

Optimization

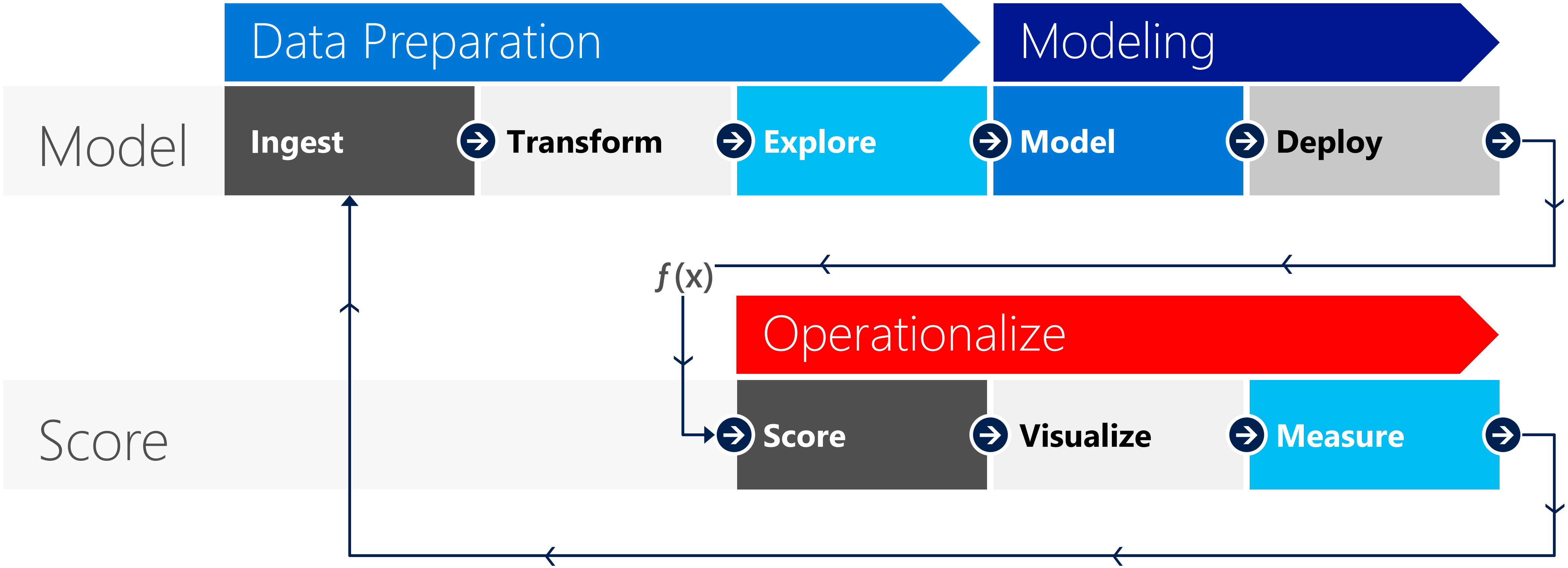
Statistics

Algorithms

Big Data

Data Science

Data Science Workflow



Data Preparation

Modeling

Model

Ingest

Transform

Explore

Model

Deploy

f(x)

Operationalize

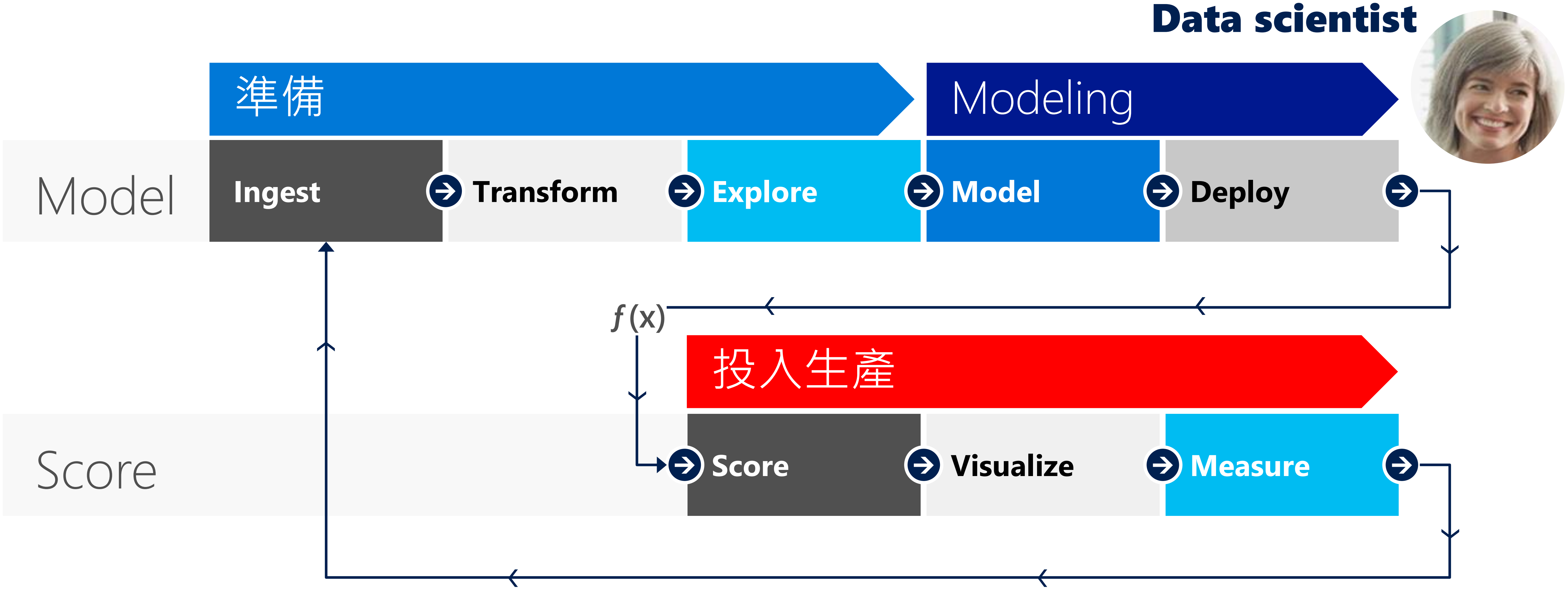
Score

Score

Visualize

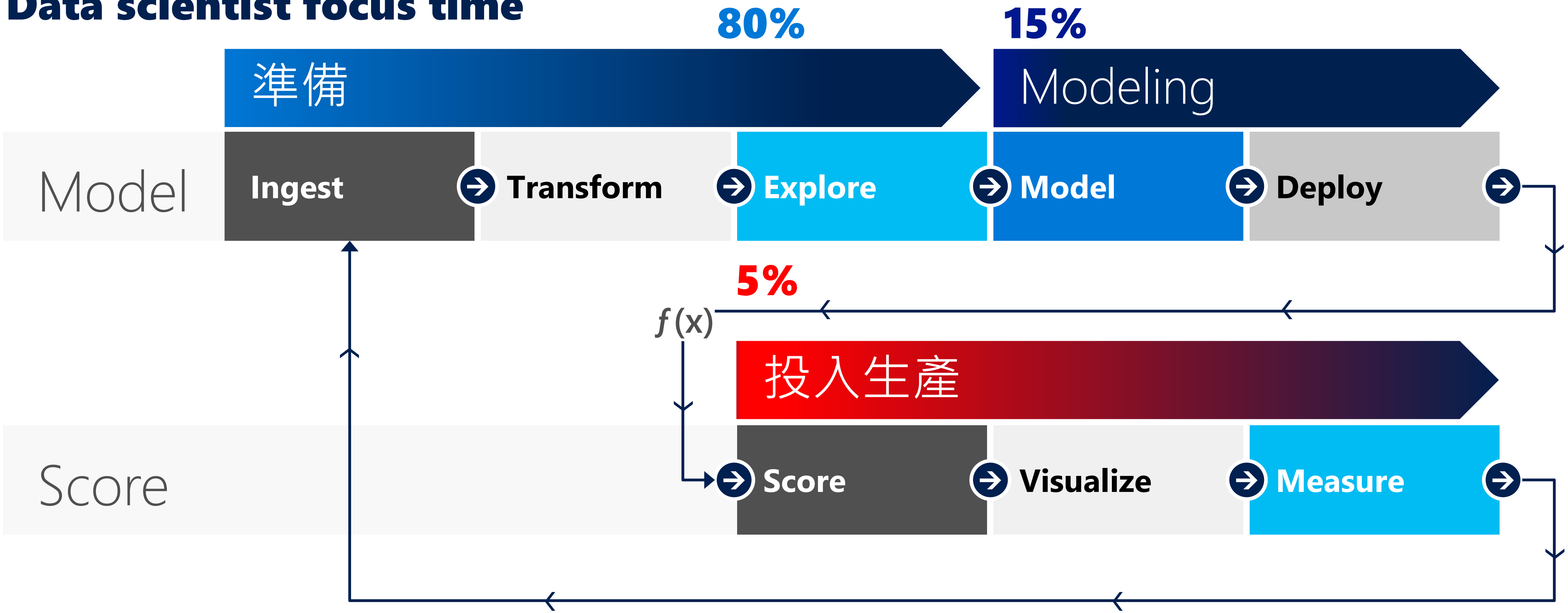
Measure

Data Scientist Primary Focus is Modeling



But In Reality….

Data scientist focus time



80%

15%

準備

Modeling

Model

Ingest

Transform

Explore

Model

Deploy

5%

f (x)

投入生產

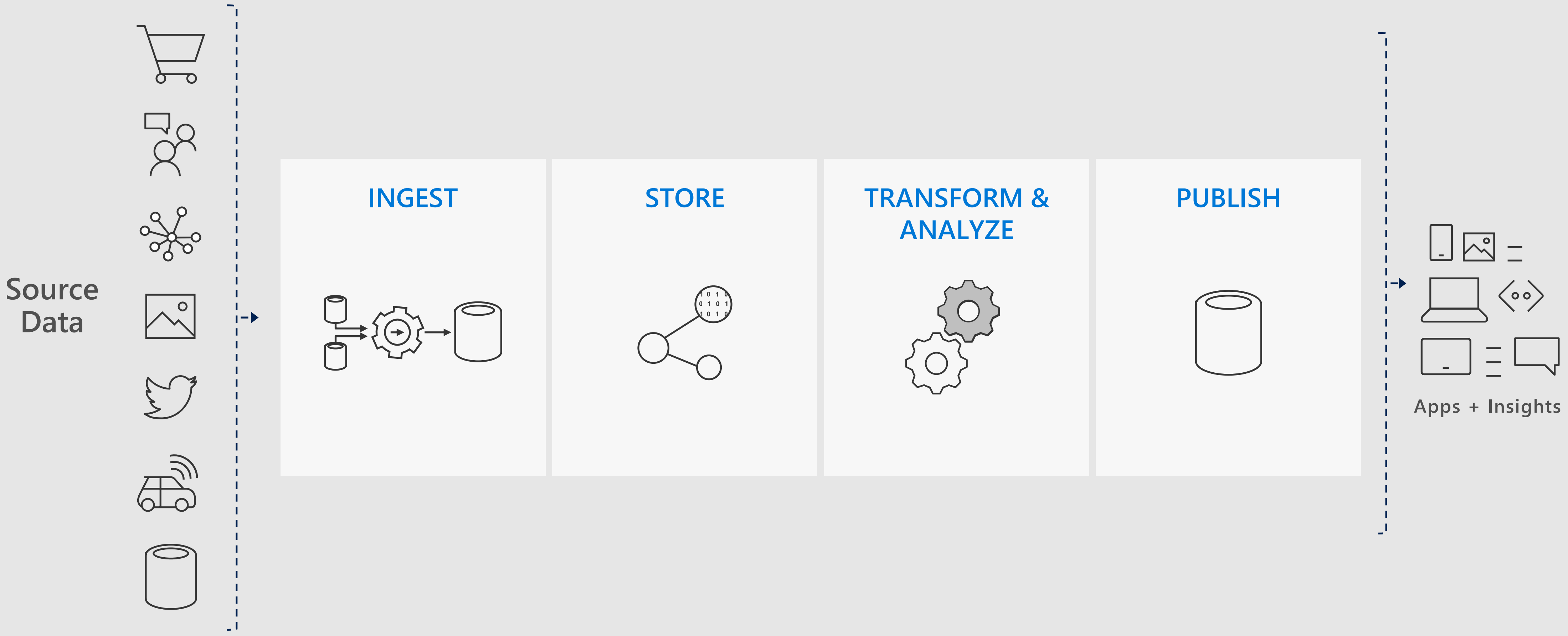
Score

Score

Visualize

Measure

Very High-Level Big Data Architecture



Solution scenarios

Let’s walk through these scenarios to see the architecture in action…

Modern DW

“We want to incorporate all of our data including ‘big data” with our data warehouse”

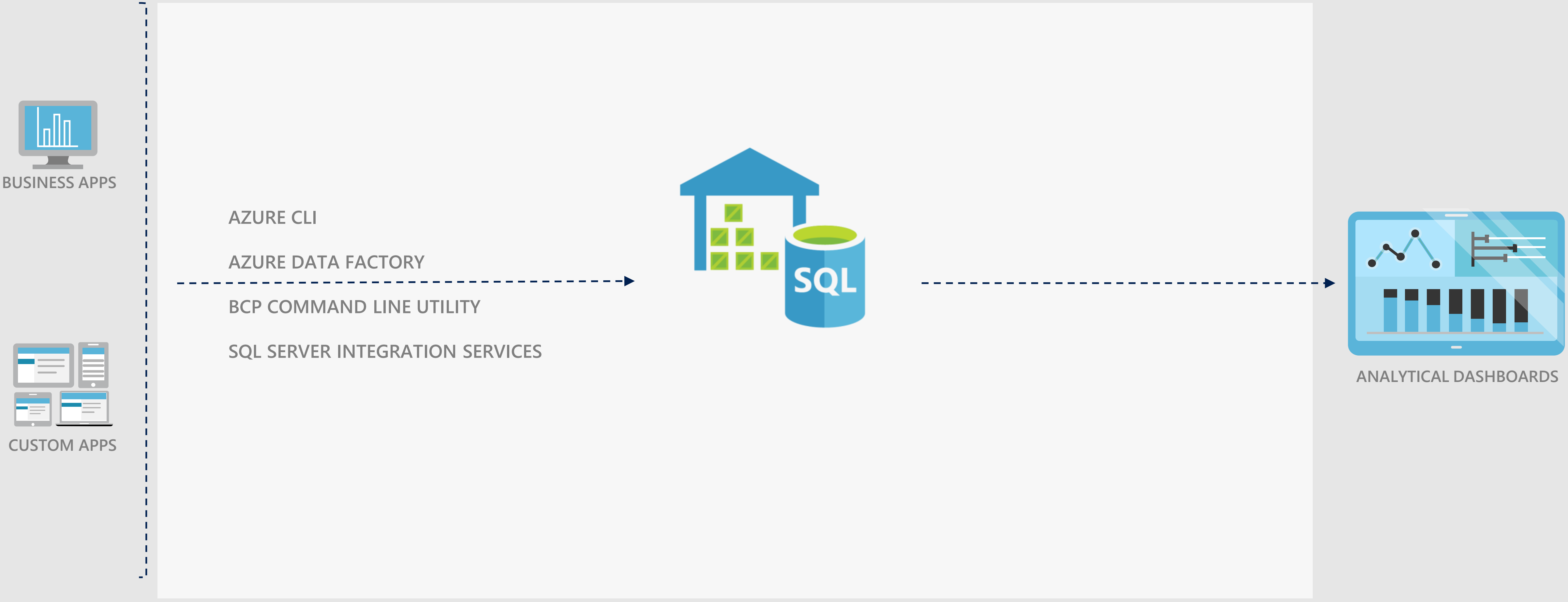
Advanced Analytics

“We are trying to predict when our customers churn.”

Internet of Things (IoT)

“We are trying to get insights from our devices in real-time, etc.”

Traditional Data Warehouse



BUSINESS APPS

CUSTOM APPS

AZURE CLI

AZURE DATA FACTORY

BCP COMMAND LINE UTILITY

SQL SERVER INTEGRATION SERVICES

ANALYTICAL DASHBOARDS

Azure SQL Data Warehouse

Elastic data warehouse as a service with enterprise-class features

Enterprise-class cloud data warehouse that can grow, shrink, and pause in seconds

Full SQL Server experience

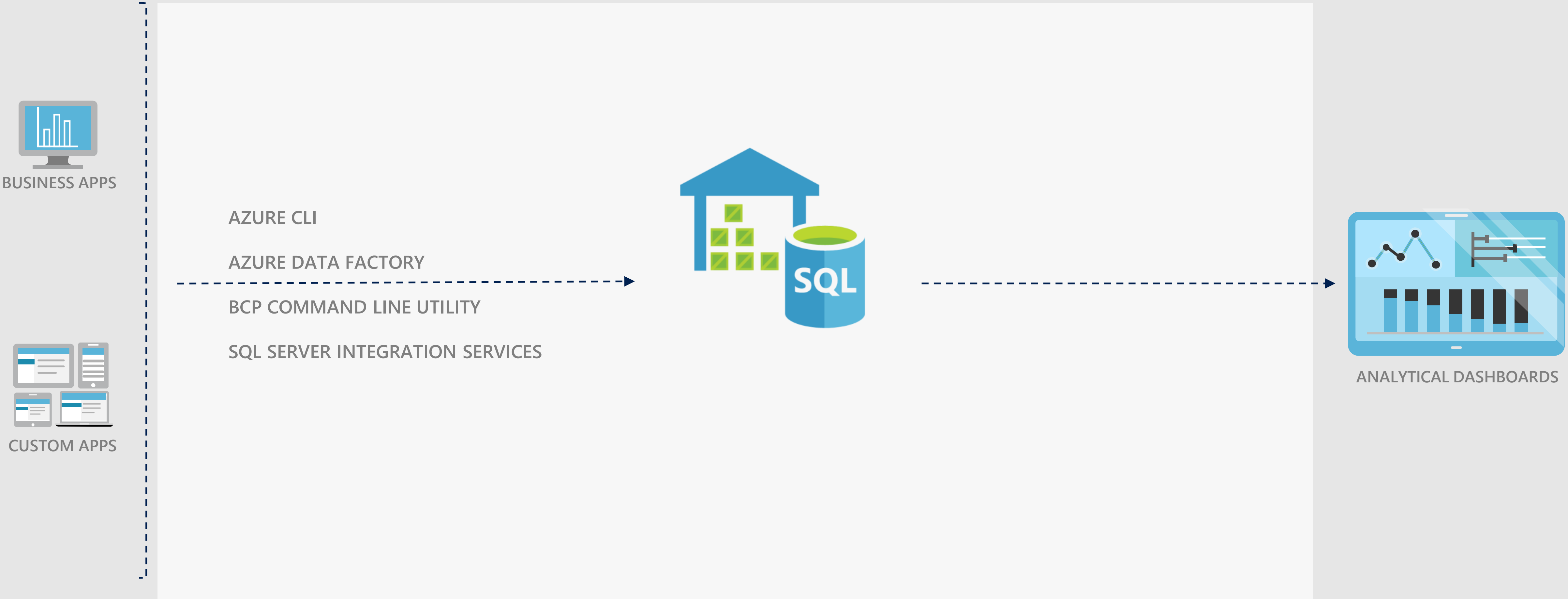
Seamless compatibility with Power BI, Azure Machine Learning, HDInsight, and Azure Data Factory

Petabyte scalability with massive parallel processing

Independent scale of compute and storage in seconds

Transaction of SQL queries across relational and non-relational data in Hadoop with PolyBase

Traditional Data Warehouse



BUSINESS APPS

CUSTOM APPS

AZURE CLI

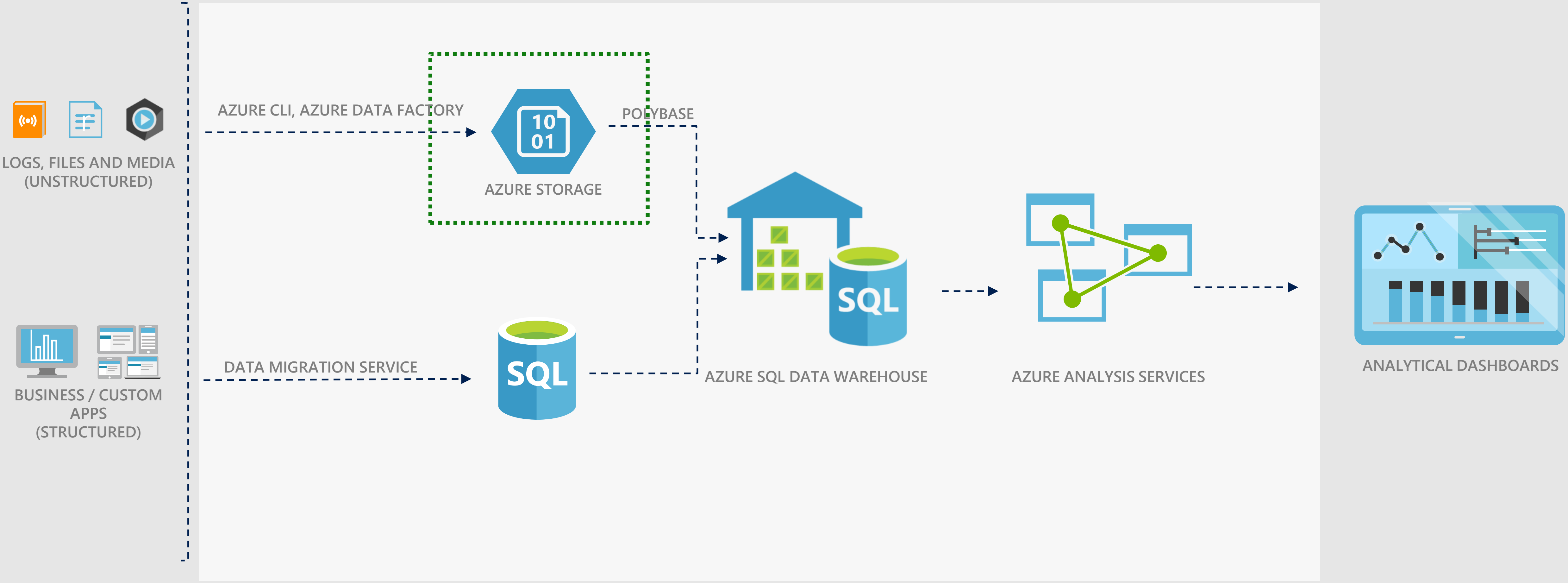
AZURE DATA FACTORY

BCP COMMAND LINE UTILITY

SQL SERVER INTEGRATION SERVICES

ANALYTICAL DASHBOARDS

Cloud Data Warehouse



Azure Storage Options

AZURE BLOB STORAGE

- **Purpose:** General purpose object store for a wide variety of storage scenarios
- **Use Cases:** Any type of text or binary data, such as application back end, backup data, media storage for streaming and general purpose data
- **Key Concepts:** Storage account has containers, which in turn has data in the form of blobs
- **Structure:** Object store with flat namespace
- **Limit:** Specific limits documented here

AZURE DATA LAKE STORE

- **Purpose:** Optimized storage for big data analytics workloads
- **Use Cases:** Batch, interactive, streaming analytics and machine learning data such as log files, IoT data, click streams, large datasets
- **Key Concepts:** Data Lake Store account contains folders, which in turn contains data stored as files
- **Structure:** Hierarchical file system
- **Limit:** No limits on account sizes, file sizes or number of files

Azure Analysis Services

Enterprise grade analytics engine as a service

Build semantic models

Transform complex data into business user friendly semantic models

Proven technology

Based on SQL Server Analysis Services

In-Memory Cache

Gain instant insights with in-memory cache using your preferred visualization tools

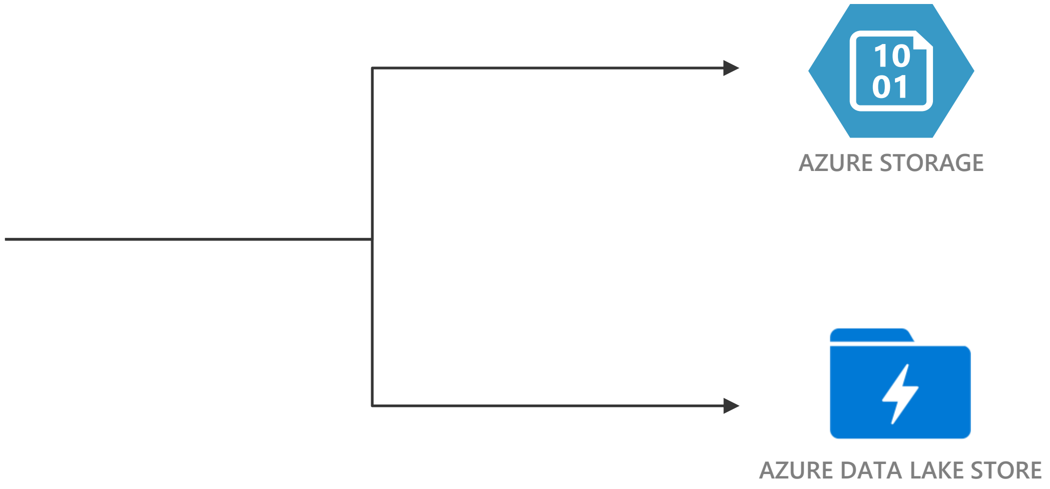
Provision and scale

Easy to deploy, scale, and manage as platform-as-a-service

Polybase

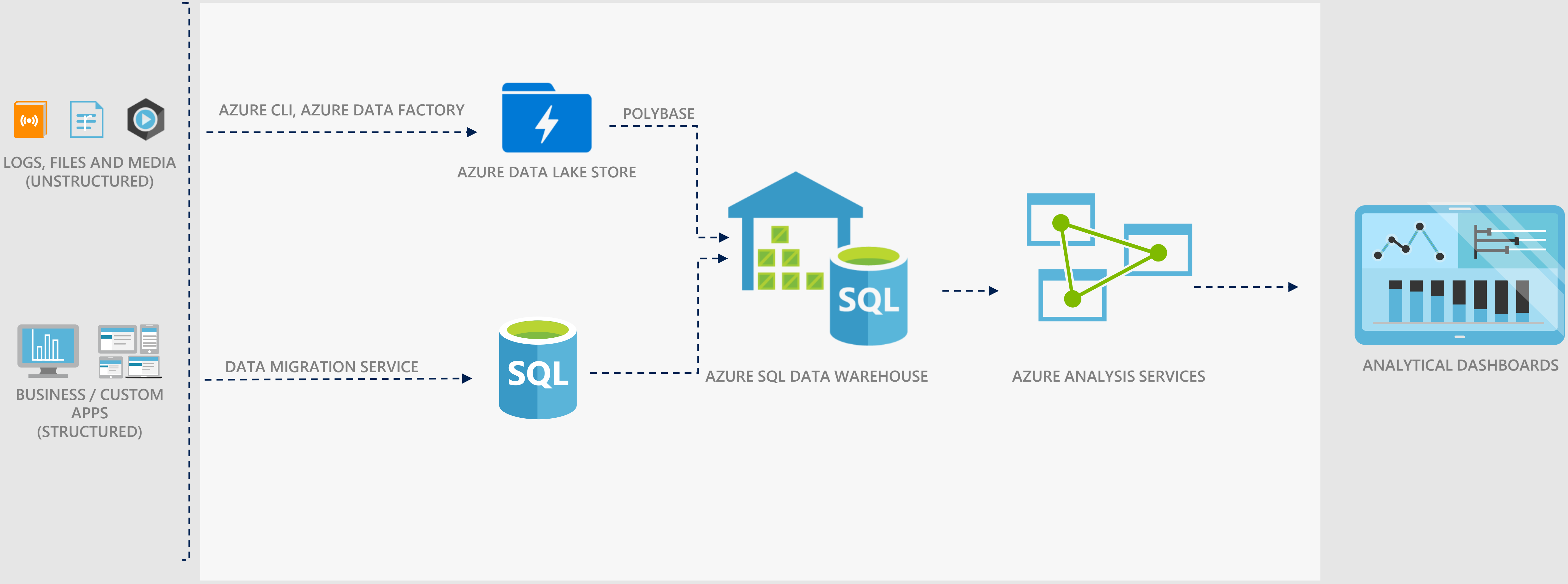
Available in SQL Server 2016+ and Azure SQL DW

```
CREATE EXTERNAL TABLE [dbo].[CarSensor_Data] (
       [SensorKey] int NOT NULL,
       [CustomerKey] int NOT NULL,
       [GeographyKey] int NULL,
       [Speed] float NOT NULL,
       [YearMeasured] int NOT NULL
)
WITH (LOCATION='/Demo/',
       DATA_SOURCE = MyAzureStorage,
       FILE_FORMAT = TextFileFormat
);
```

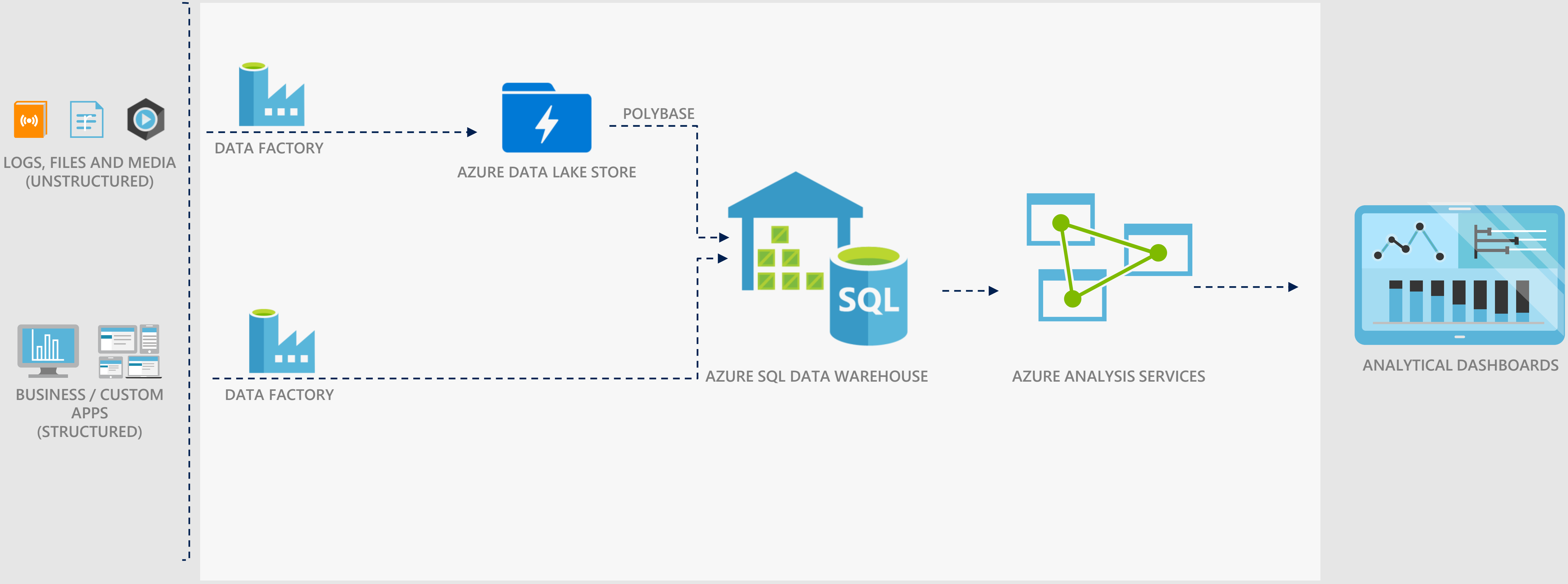


```
SELECT * FROM [dbo].[CarSensor_Data];
```

Cloud Data Warehouse



Cloud Data Warehouse



Azure Data Factory

Cloud-Scale Data Orchestration Tool

Manage Data Pipelines

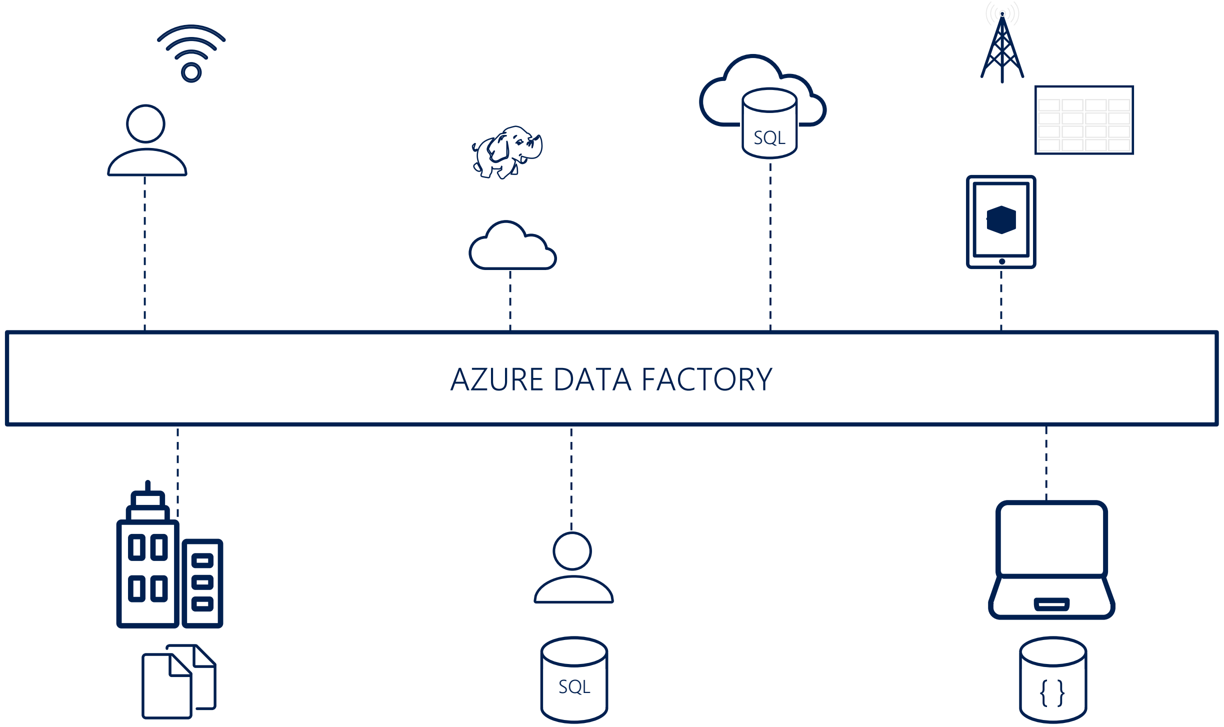
Create, schedule, orchestrate, and manage data pipelines

Hybrid Data Movement

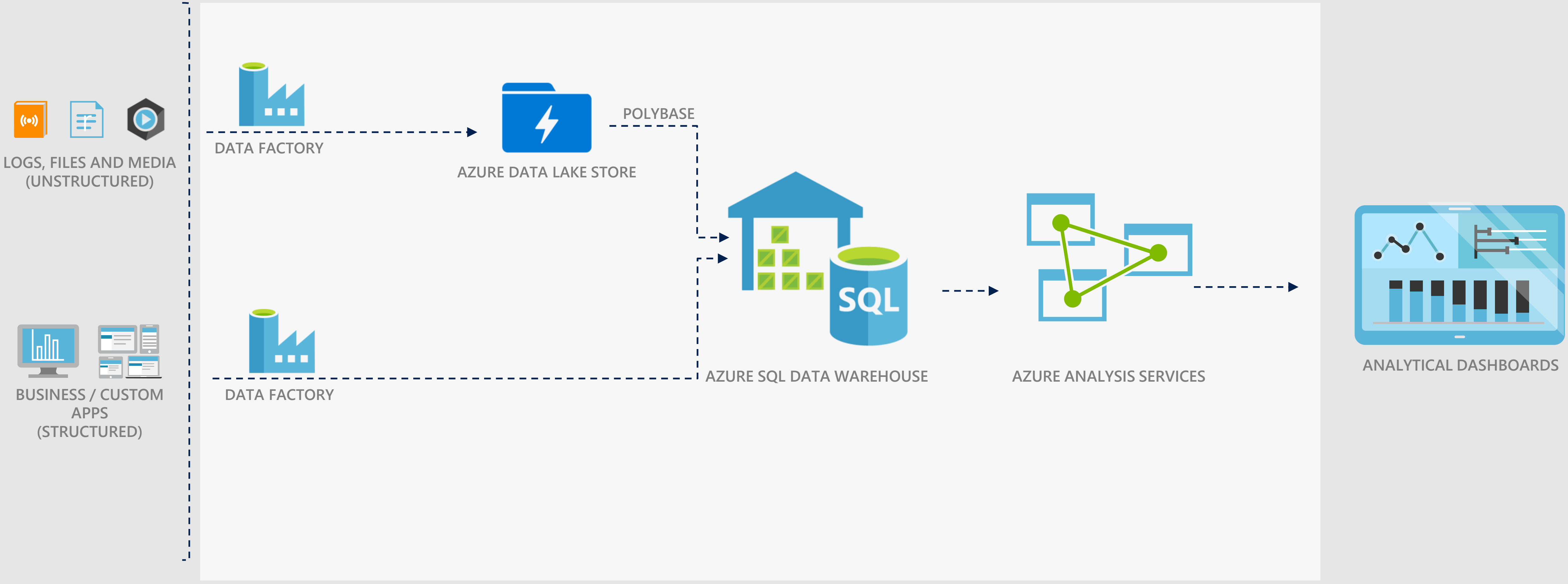
Connect to on-premises and cloud data sources

Provision Resources

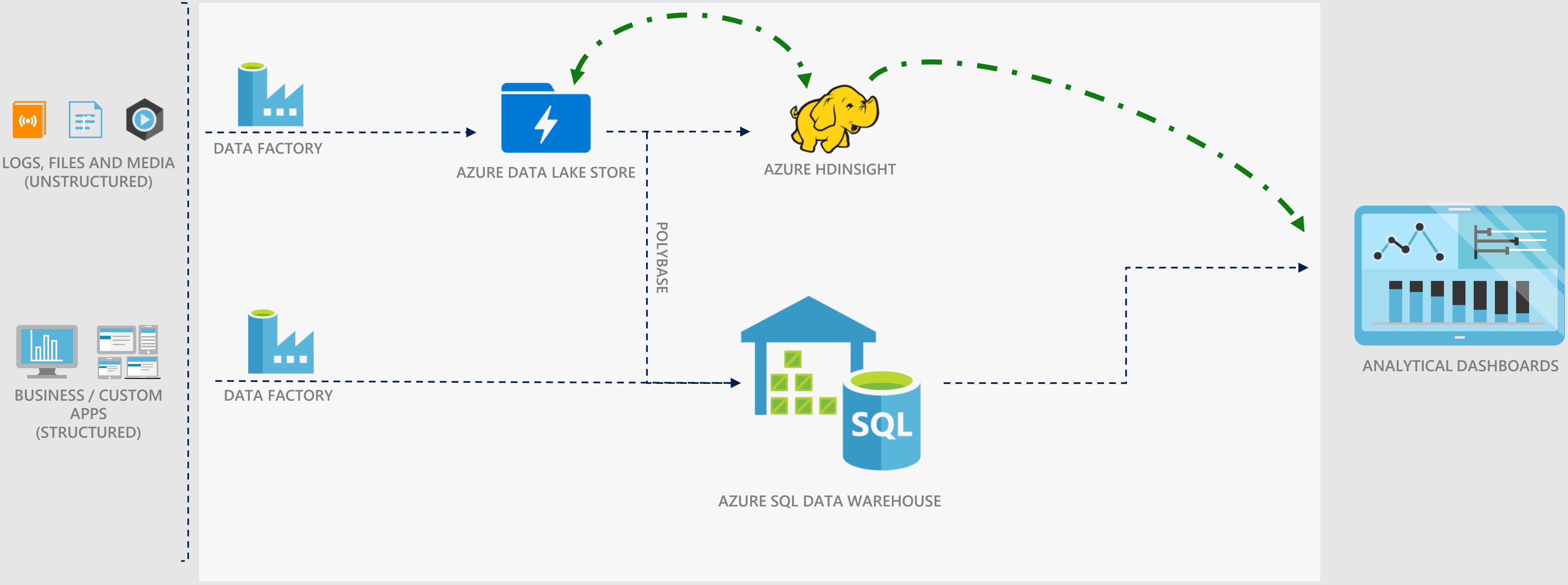
Manage transient resources to run your data pipelines



Cloud Data Warehouse



Cloud Data Warehouse



Azure HDInsight

Managed Hadoop Cluster – Built on Hortonworks

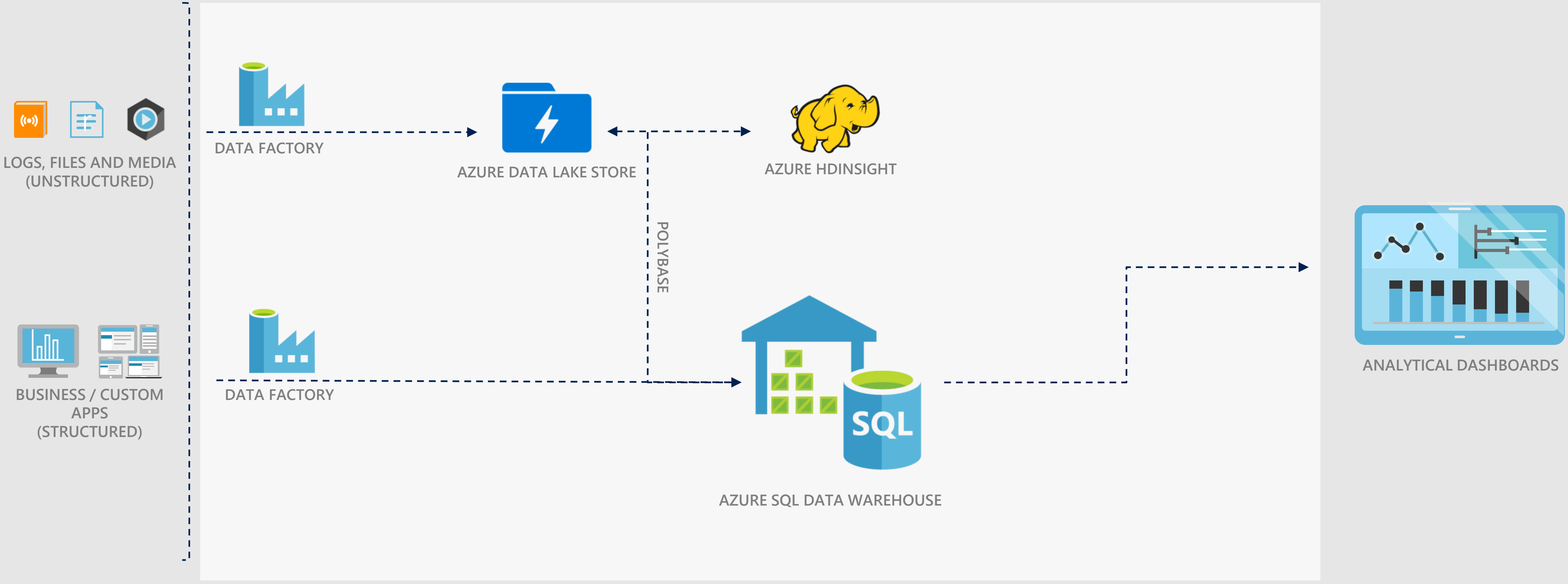
Microsoft

+

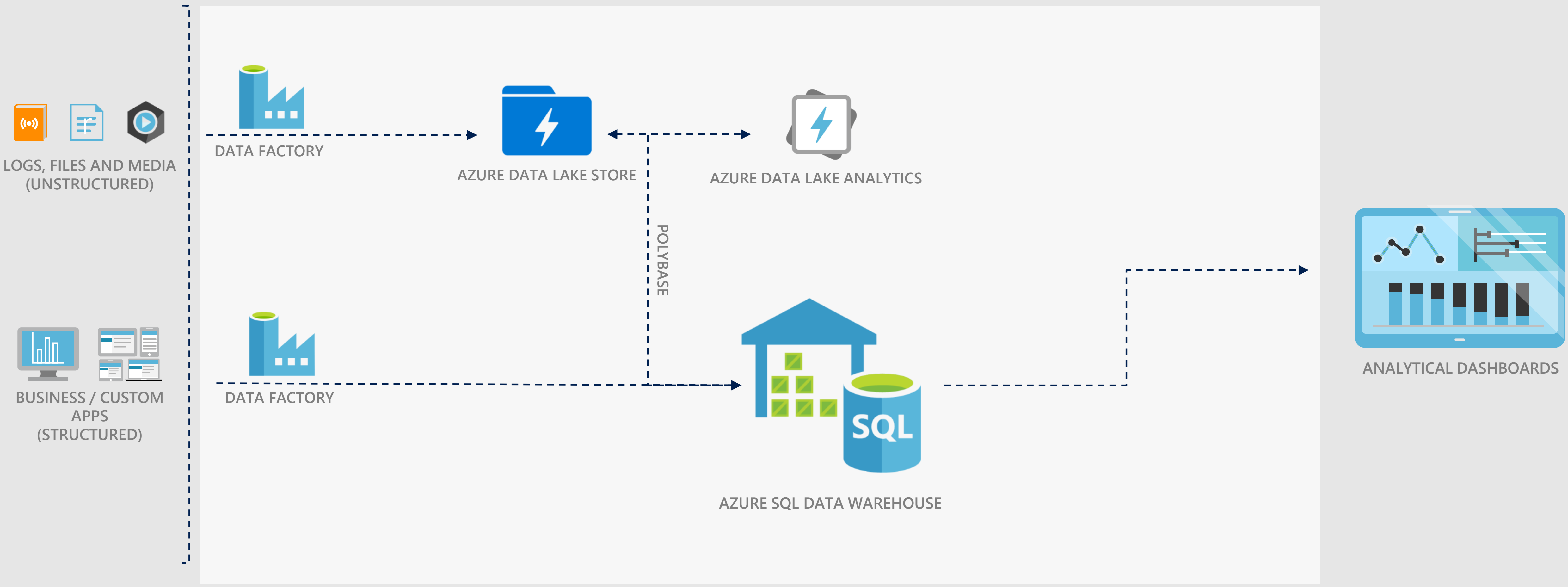
HORTONWORKS
POWERING THE FUTURE OF DATA

- **Cost-effectively scale workloads** up or down through decoupled compute and storage.
- **Rich productivity suites for Hadoop and Spark** – such as such as Visual Studio, Eclipse, and IntelliJ for Scala, Python, R, Java, and .NET support, Jupyter notebook, Microsoft Machine Learning Server.
- Managed service open source analytics with an **Industry-leading 99.9% SLA**
- Available in >**25 regions** globally
- **Secure and compliant**: HIPPA, PCI, SOC compliance.

Cloud Data Warehouse



Cloud Data Warehouse



Azure Data Lake Analytics

Big Data Compute as-a-Service

Data Lake Store
Storage service optimized for big data analytics

Data Lake Analytics
Big data as a service

HDInsight
Clusters as a service

- Easily develop and run massively parallel data transformation and processing programs in U-SQL, R, Python and .NET over petabytes of data.
- No infrastructure to manage.
- Process data on demand.
- Scale instantly.
- Only pay per job
- Enterprise-grade Support and Security

3 Different Big Data Compute Options

	HDP | CDH | MapR (Azure Marketplace) Any OSS Analytics technology	**HDInsight** Workload-optimized, managed clusters	**Data Lake Analytics** Specific apps in a multi-tenant form factor
	IaaS Clusters	Managed Clusters	Big Data Compute as-a-service
Best for...	Lifting and shifting existing Hadoop workloads to the cloud without changes, full control	Spinning up HDInsight (PaaS) in minutes, fully managed by Microsoft with some control	Easiest way to get started on big data - Leverage SQL + C# skills, no infrastructure administration needed
Workloads	Full Hadoop distribution and projects	Most Hadoop distribution: batch, streaming, interactive and machine learning with ability to customize cluster	No Hadoop distribution: Batch processing supported currently (U-SQL)
Administrative	Will need Hadoop admin experience – everything done yourself. Still need to manage clusters.	Easier to use—Make admin jobs easier: OS upgrades, patching, Hadoop version upgrades done for you. Still need to manage clusters.	Easiest to use—minimal admin functions needed. No cluster notion. Instantly, scales elastically per job.
Developer	Use familiar Hadoop tooling (Hive, Spark, etc.).	Use familiar Hadoop tooling (Hive, Spark, etc.). Microsoft provides some Visual Studio and IntelliJ integration	Deep Visual Studio integration for coding, debugging, optimizing (.NET – C# / SQL)
Control & configuration	**Full control** of managing and running your clusters. Spin up VMs as needed	**Some control and some configuration.** Fully managed and monitored by Microsoft with 99.9% SLA, scale nodes on demand, control # of VMs on Azure	No need to control or configure Instantly scales elastically per job
Service Level Agreement	Only on VM network connectivity	99.9% on both network connectivity and Hadoop bits are running in VMs	99.9% SLA at GA
TCO	Lowest cost per query Higher TCO	Low cost per query Low TCO from balanced resourcing	Highest cost per query Lowest TCO

Solution scenarios

Modern DW

“We want to incorporate all of our data including ‘big data” with our data warehouse”

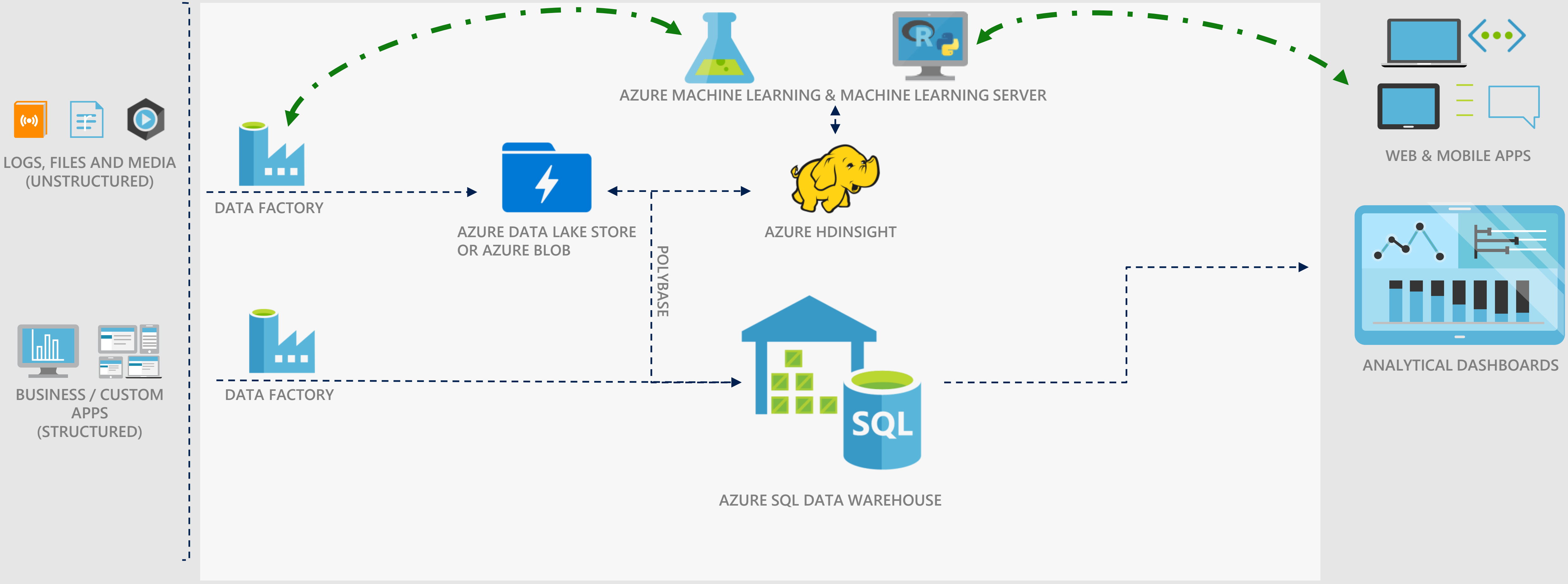
Advanced Analytics

“We are trying to predict when our customers churn.”

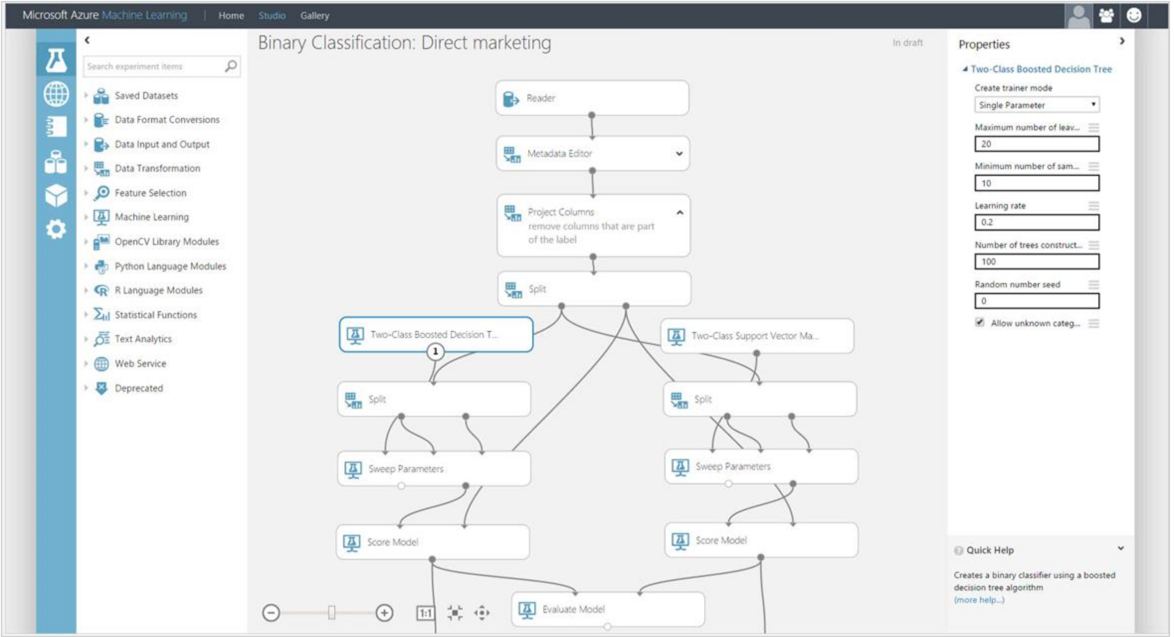
Internet of Things (IoT)

“We are trying to get insights from our devices in real-time, etc.”

Advanced Analytics on Big Data


LOGS, FILES AND MEDIA (UNSTRUCTURED)
BUSINESS / CUSTOM APPS (STRUCTURED)
DATA FACTORY
DATA FACTORY
AZURE DATA LAKE STORE OR AZURE BLOB
POLYBASE
AZURE MACHINE LEARNING & MACHINE LEARNING SERVER
AZURE HDINSIGHT
AZURE SQL DATA WAREHOUSE
SQL
WEB & MOBILE APPS
ANALYTICAL DASHBOARDS

Azure Machine Learning



Properties

Two-Class Boosted Decision Tree

Create trainer mode: Single Parameter

Maximum number of leav...: 20

Minimum number of sam...: 10

Learning rate: 0.2

Number of trees construct...: 100

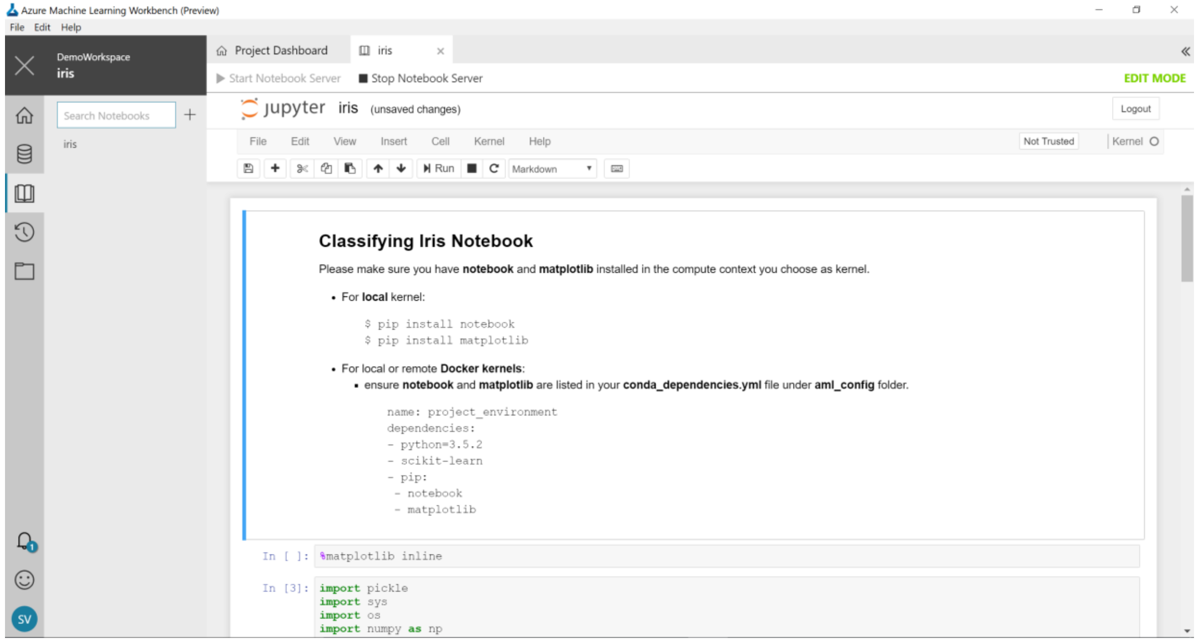
Random number seed: 0

☑ Allow unknown categ...

Quick Help

Creates a binary classifier using a boosted decision tree algorithm (more help...)

VISUAL DRAG-AND-DROP



Classifying Iris Notebook

Please make sure you have **notebook** and **matplotlib** installed in the compute context you choose as kernel.

- For **local** kernel:

```
$ pip install notebook
$ pip install matplotlib
```

- For local or remote **Docker kernels**:
  - ensure **notebook** and **matplotlib** are listed in your **conda_dependencies.yml** file under **aml_config** folder.

```
name: project_environment
dependencies:
- python=3.5.2
- scikit-learn
- pip:
 - notebook
 - matplotlib
```

```
In [ ]: %matplotlib inline
```

```
In [3]: import pickle
        import sys
        import os
        import numpy as np
```

CODE-FIRST

Microsoft ML Server

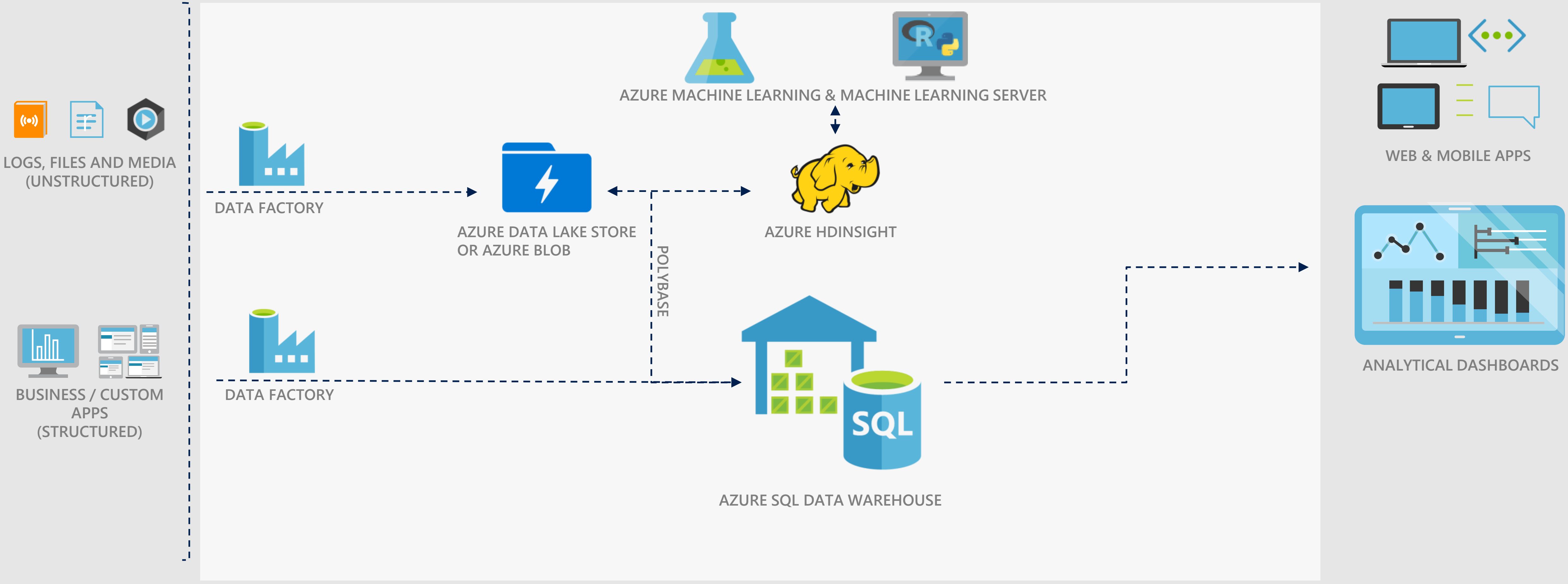
Extend beyond open source R and Python, and transform business with Enterprise-grade analytics

Create smarter apps with **industry-leading artificial intelligence (AI) and leading machine learning** capabilities, in addition to open source R and Python.

Simplify deployment of your analytics models. Integrate analytics faster with apps written in any language and score easily across data platforms using web services and your preferred development environment.

When your data stores grow, Machine Learning Server can be deployed to **perform at scale wherever your big data lives**—including databases such as SQL Server 2016, Hadoop clusters, data warehouses, and even data stores in the cloud.

Advanced Analytics on Big Data


LOGS, FILES AND MEDIA (UNSTRUCTURED)
BUSINESS / CUSTOM APPS (STRUCTURED)
DATA FACTORY
AZURE DATA LAKE STORE OR AZURE BLOB
AZURE MACHINE LEARNING & MACHINE LEARNING SERVER
AZURE HDINSIGHT
POLYBASE
DATA FACTORY
AZURE SQL DATA WAREHOUSE
SQL
WEB & MOBILE APPS
ANALYTICAL DASHBOARDS

Solution scenarios

Modern DW

“We want to incorporate all of our data including ‘big data” with our data warehouse”

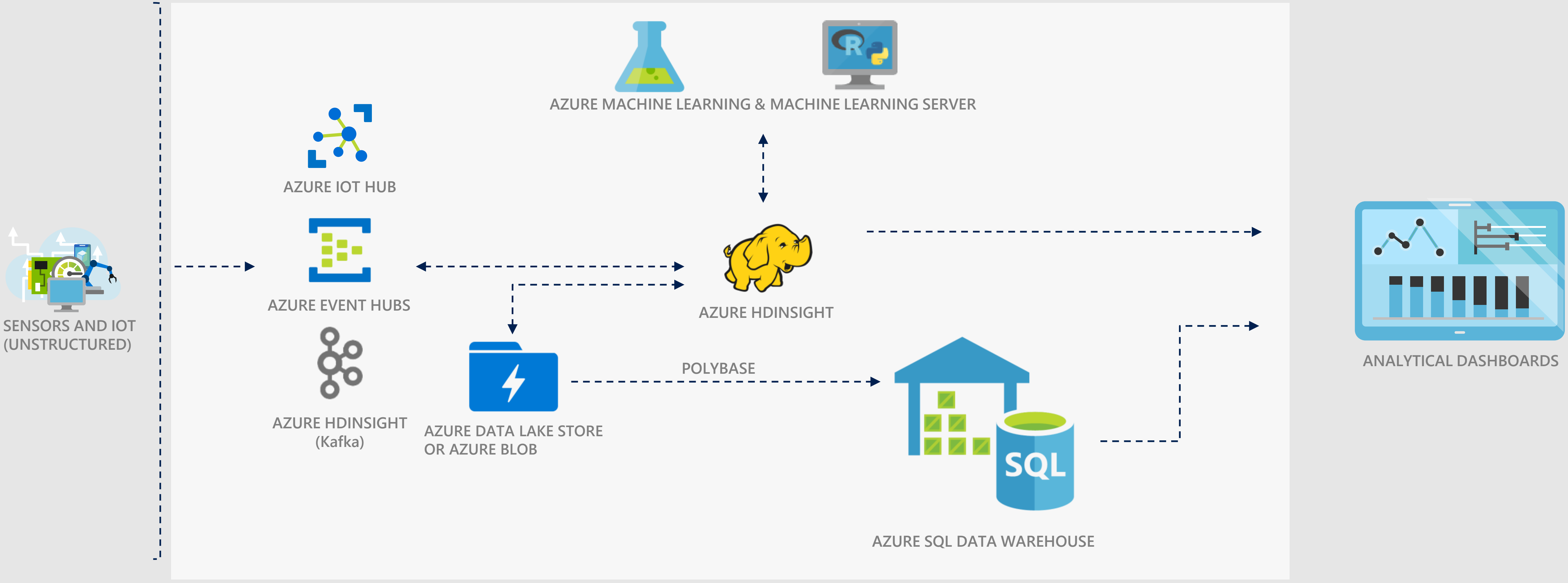
Advanced Analytics

“We are trying to predict when our customers churn.”

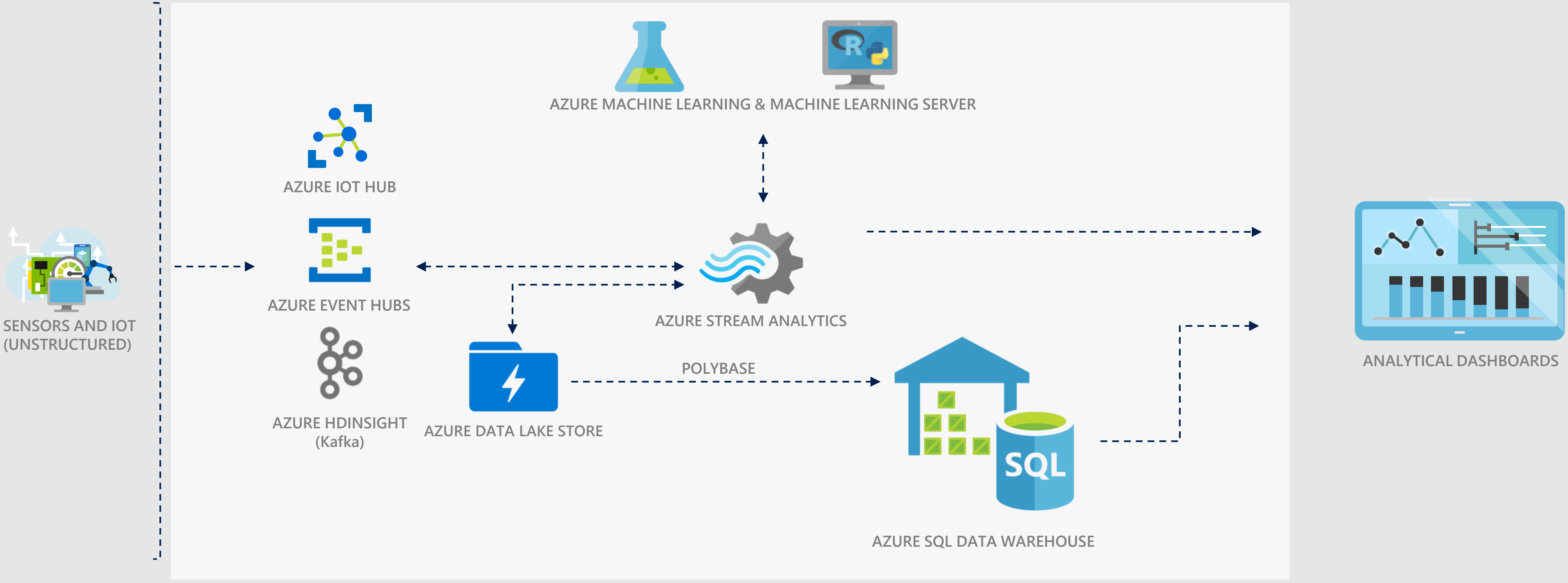
Internet of Things (IoT)

“We are trying to get insights from our devices in real-time, etc.”

Stream Ingestion



Stream Ingestion



Azure Stream Analytics

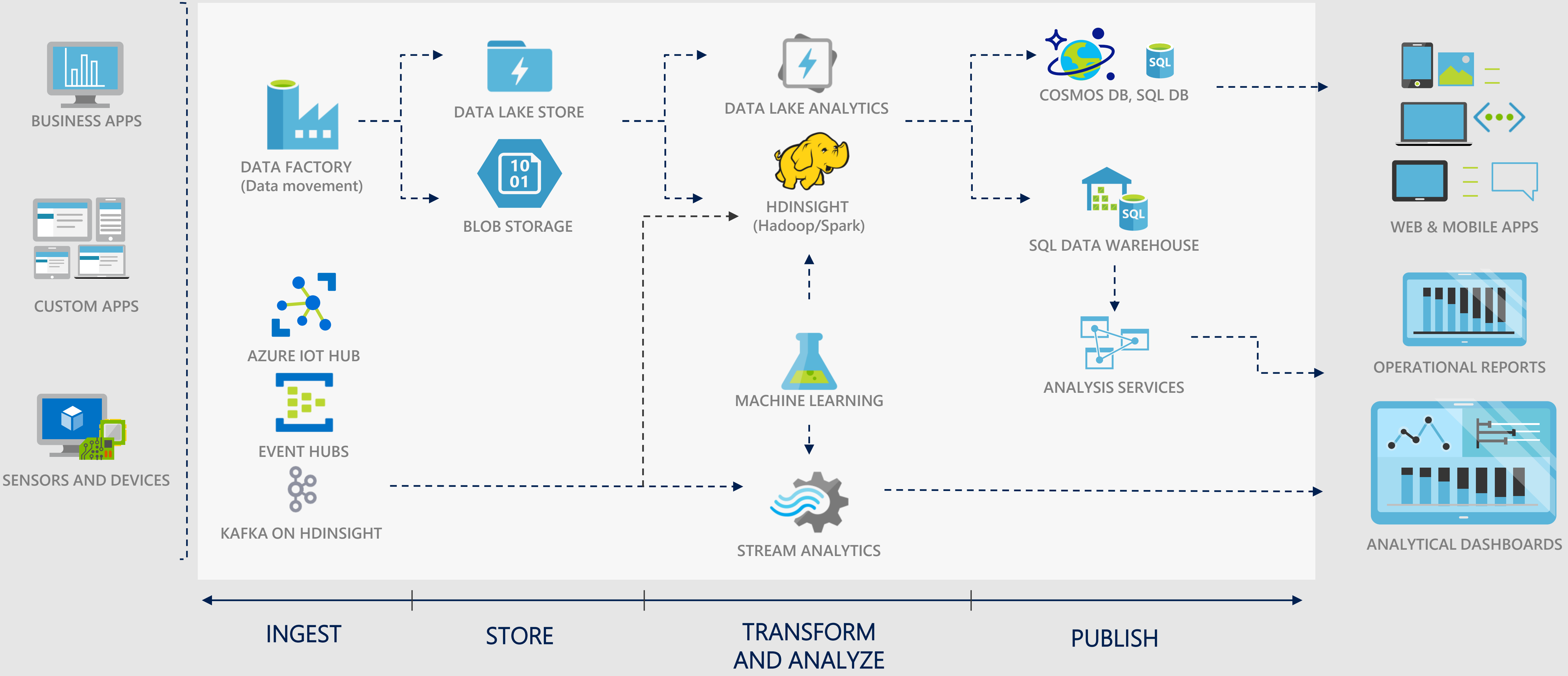
An on-demand real-time analytics service

Develop massively parallel Complex Event Processing (CEP) pipelines with simplicity
Author powerful real-time analytics using very simple declarative SQL like language for more sophisticated analytics such as Pattern detection, Time windows, Joins & correlations

Instantly analyze data from all your IoT devices and gateways
Azure Stream Analytics seamlessly integrates with Azure IoT Hub and Azure IoT Suite to enable powerful real-time analytics on data from your IoT devices and applications.

Build real-time dashboards in minutes
Quickly build real-time dashboards with Power BI for a live command and control view. Real-time dashboards help transform live data into actionable and insightful visuals, and help you focus on what matters to you the most.

All Together Now


BUSINESS APPS
CUSTOM APPS
SENSORS AND DEVICES
DATA FACTORY
(Data movement)
AZURE IOT HUB
EVENT HUBS
KAFKA ON HDINSIGHT
DATA LAKE STORE
BLOB STORAGE
10
01
DATA LAKE ANALYTICS
HDINSIGHT
(Hadoop/Spark)
MACHINE LEARNING
STREAM ANALYTICS
COSMOS DB, SQL DB
SQL
SQL DATA WAREHOUSE
SQL
ANALYSIS SERVICES
WEB & MOBILE APPS
OPERATIONAL REPORTS
ANALYTICAL DASHBOARDS
INGEST
STORE
TRANSFORM
AND ANALYZE
PUBLISH