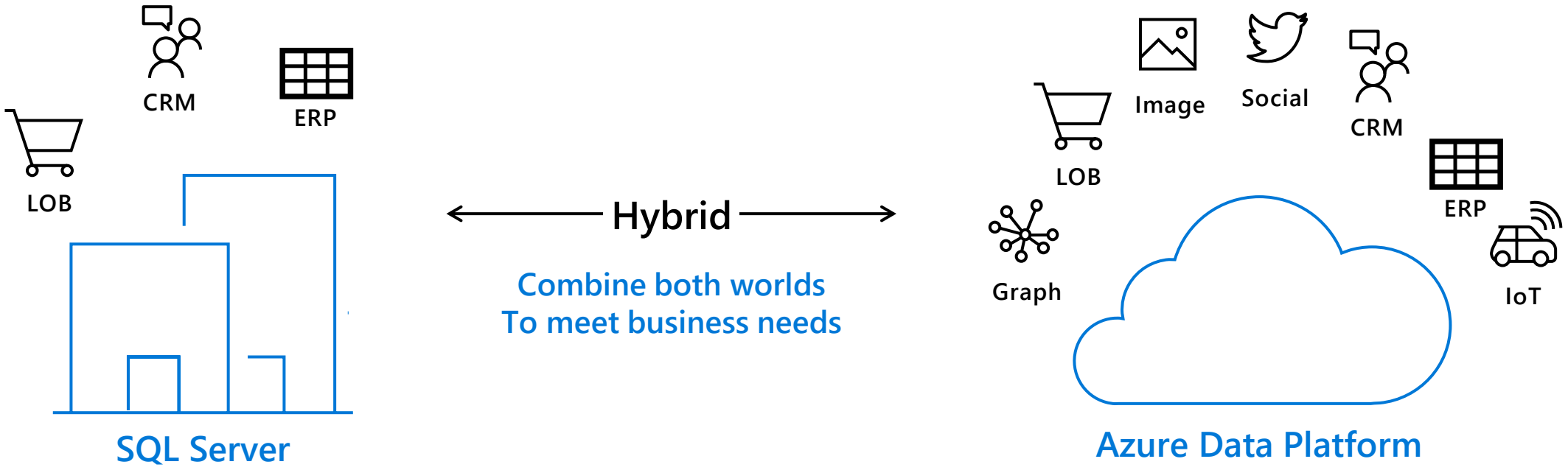
Microsoft

Delivering a Modern Data Warehouse in Azure

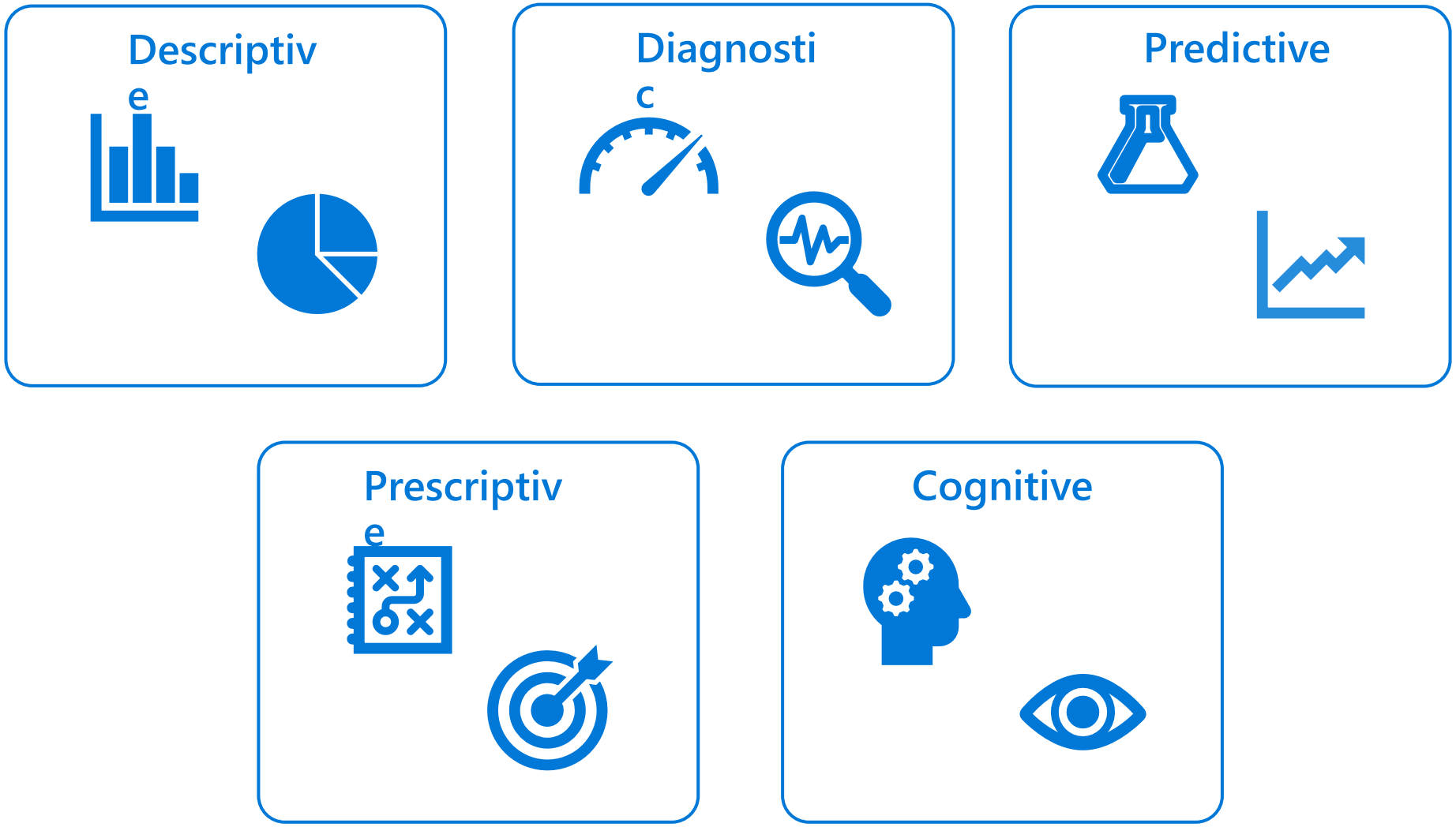
Junghwan Lee
Infinov

Why modernize?

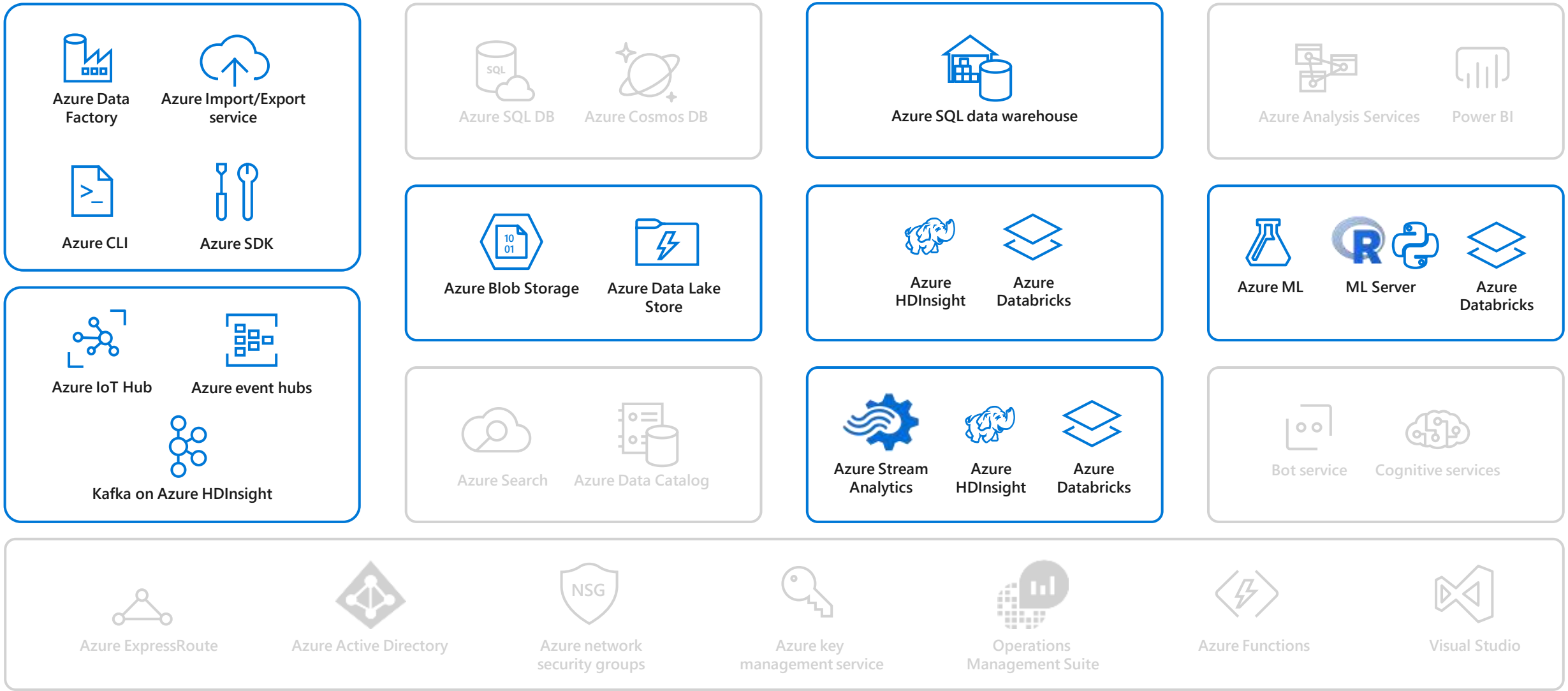
The evolving world of data



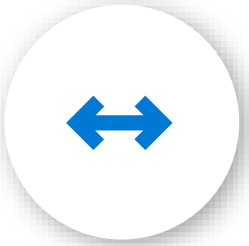
The evolving world of analytics



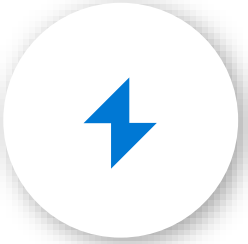
The Azure Big Data Landscape



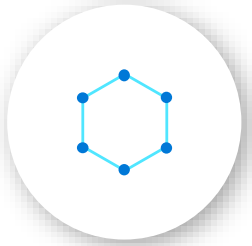
Azure Modern Data Warehouse benefits



Elastic Architectures

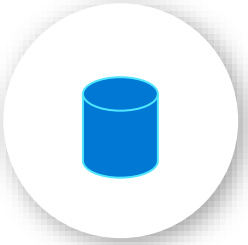


Workload Optimized Compute



Analyze All Data

Hybrid



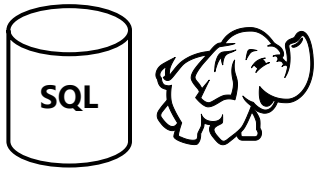
No Data Silos

Governed Self-Service

Solution scenarios

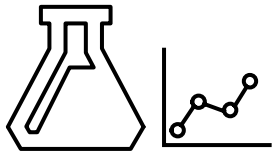
Solution scenarios

Big Data and advanced analytics



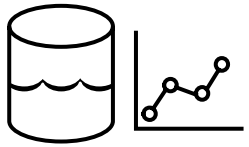
Modern data warehousing

"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"

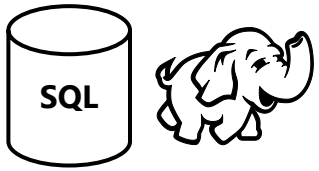


Real-time analytics

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

Azure 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"

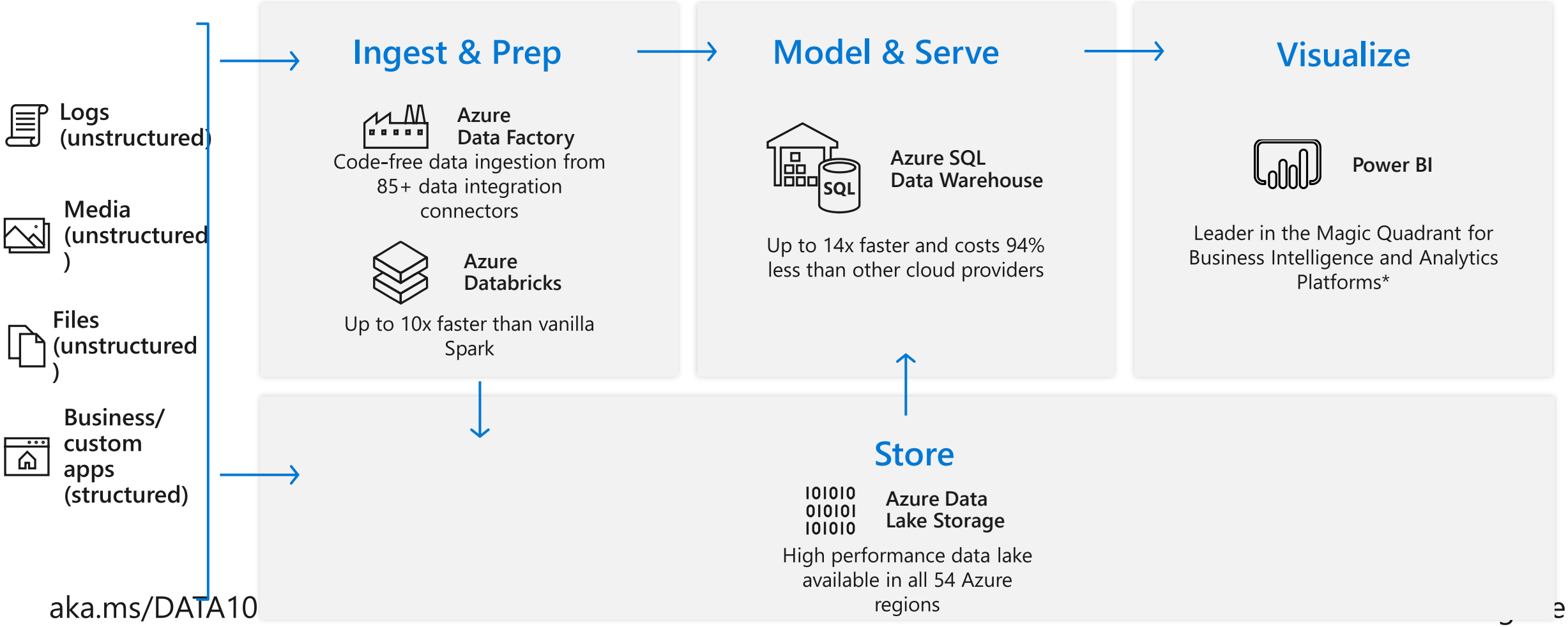
Real-time analytics

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

Azure Modern Data Warehouse Processes

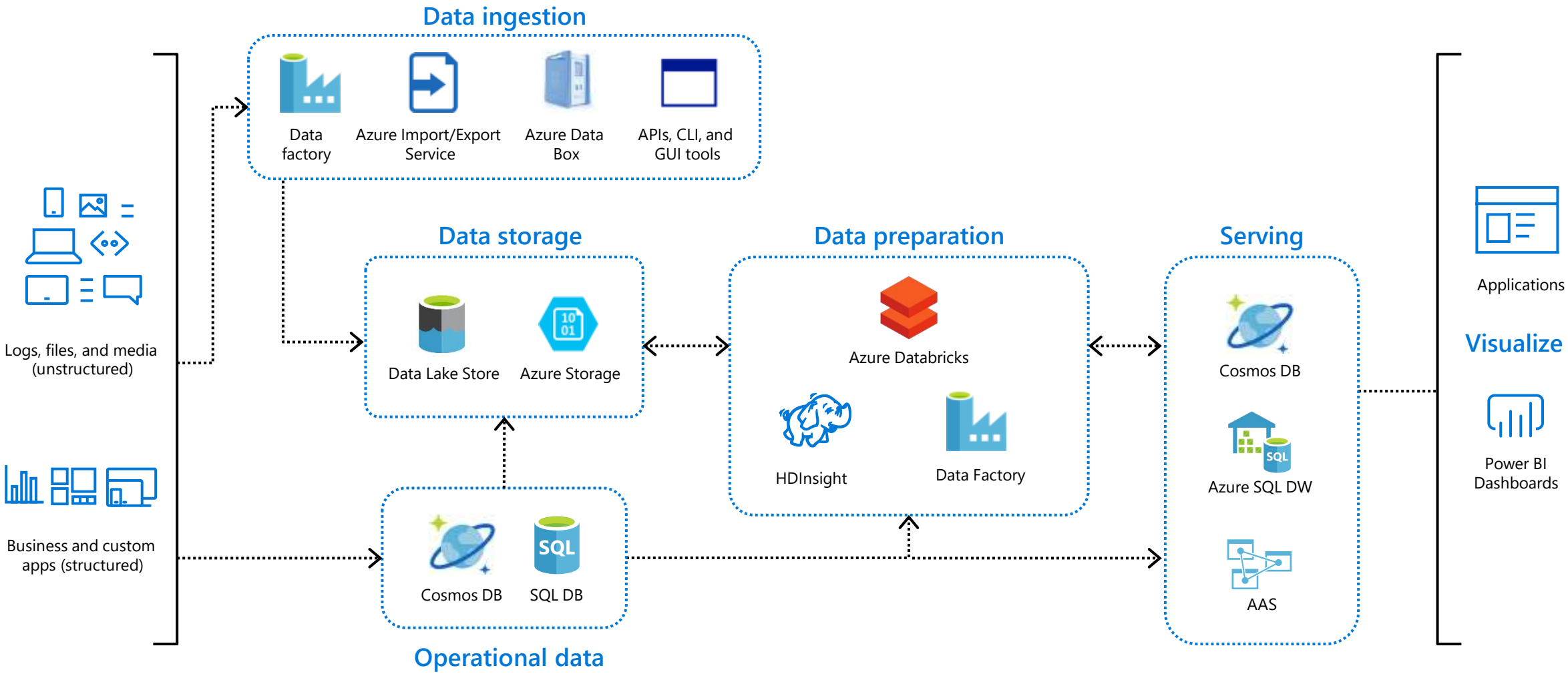
Best end-to-end ecosystem to turn your data into actionable insights

Unparalleled performance

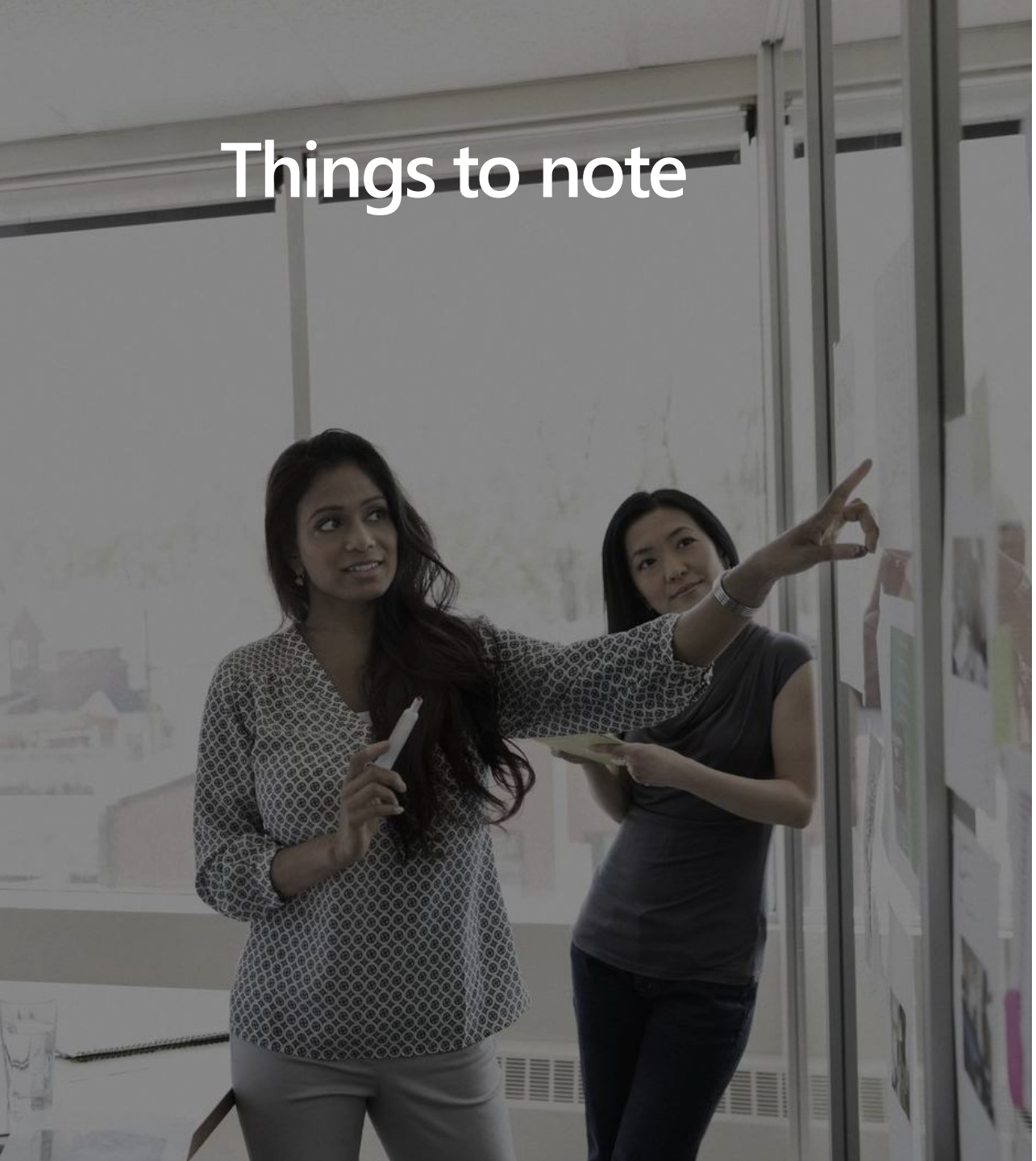


Data warehousing pattern in Azure

Loading and preparing data for analysis with a data warehouse

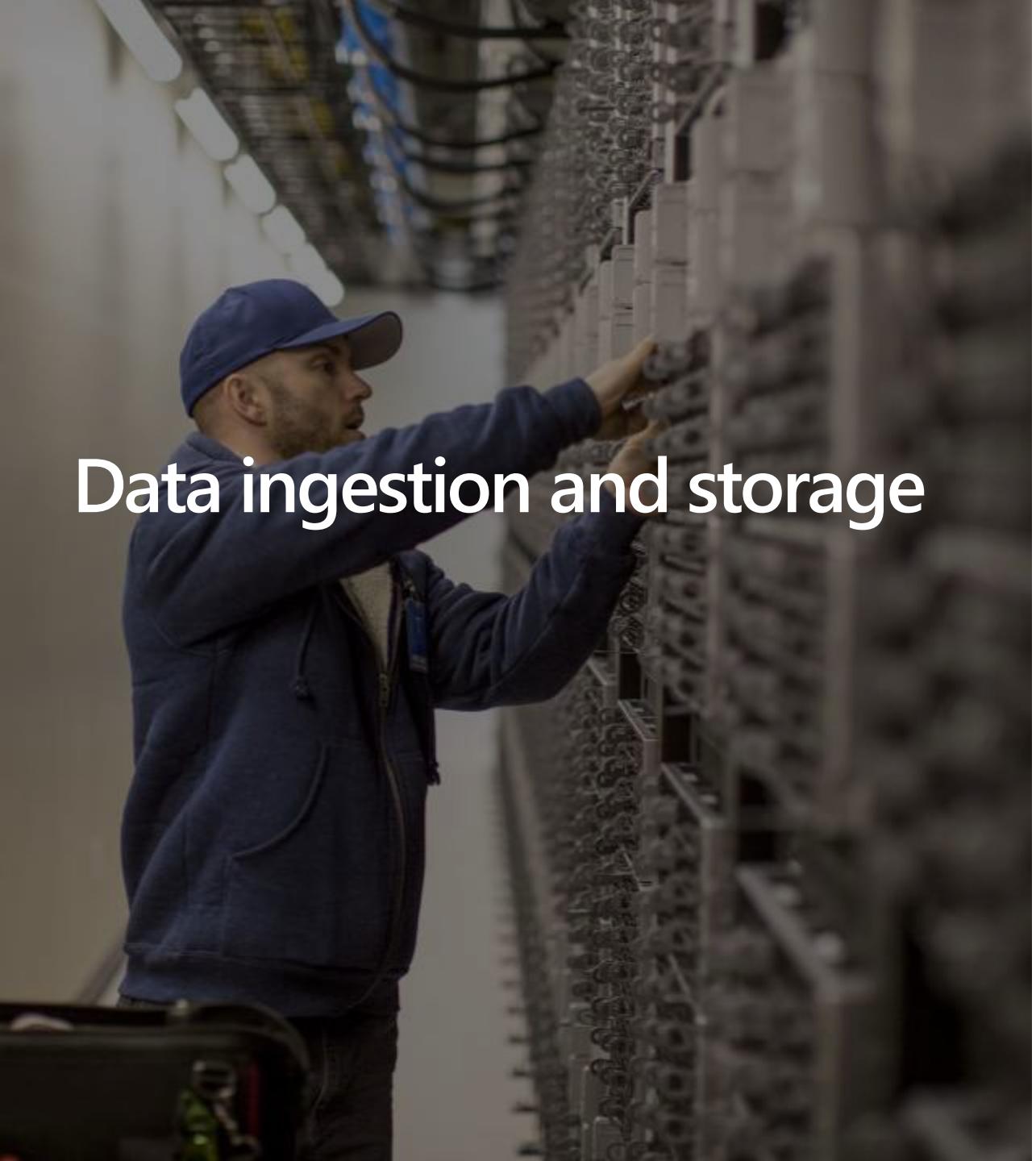


Example solution architecture



Things to note

- There are no right or wrong solutions, only optimal solutions
- Lead with certain solutions and customize based on customer scenarios
- Customer voice and product and service maturity govern lead solutions
- Consider price and performance, ease of use, and ecosystem acceptance as factors
- Everything is fluid - a lead solution today might be non-optimal tomorrow, based on the factors above and new releases

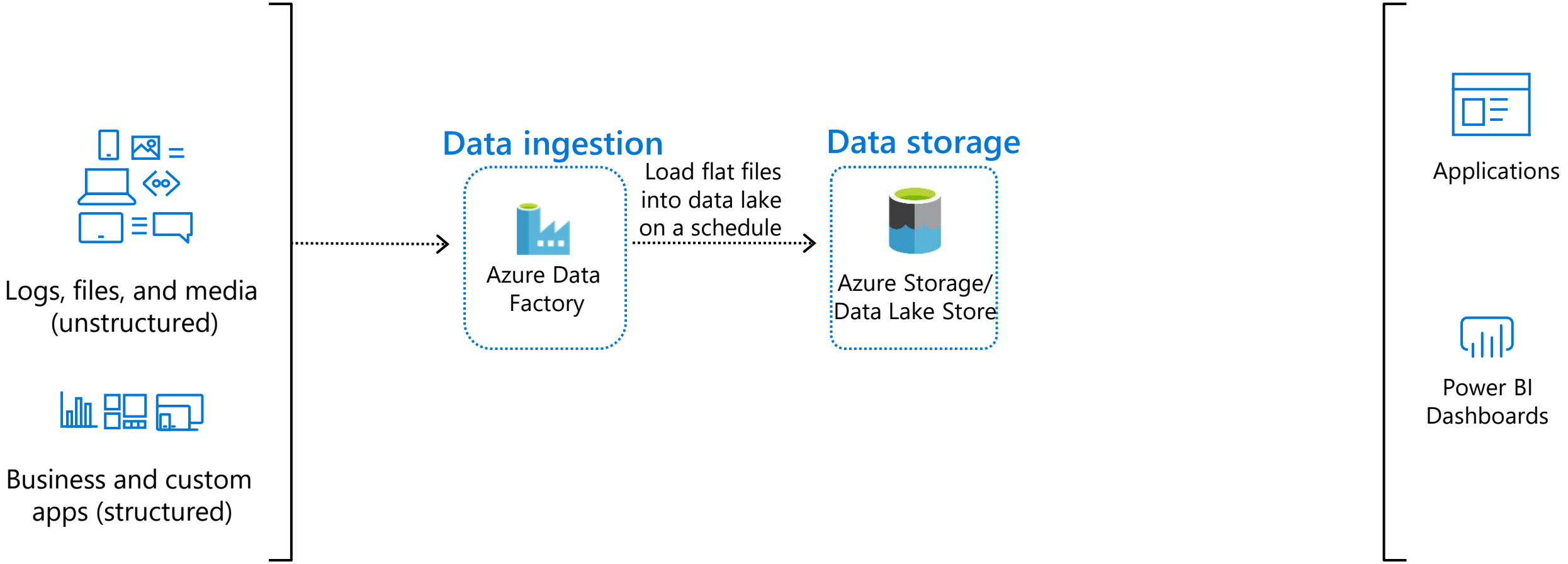


Data ingestion and storage

The storage that persists the transferred data that will be consumed by subsequent processing

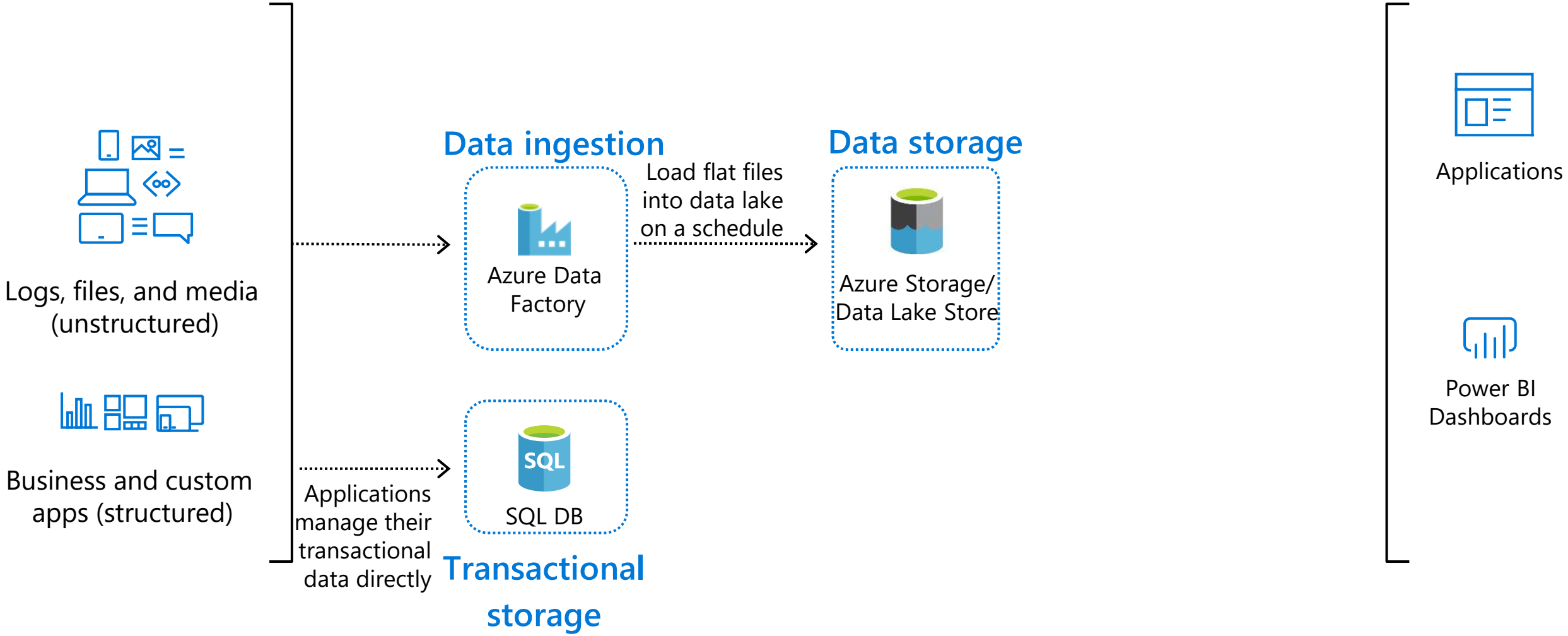
Data warehousing pattern in Azure

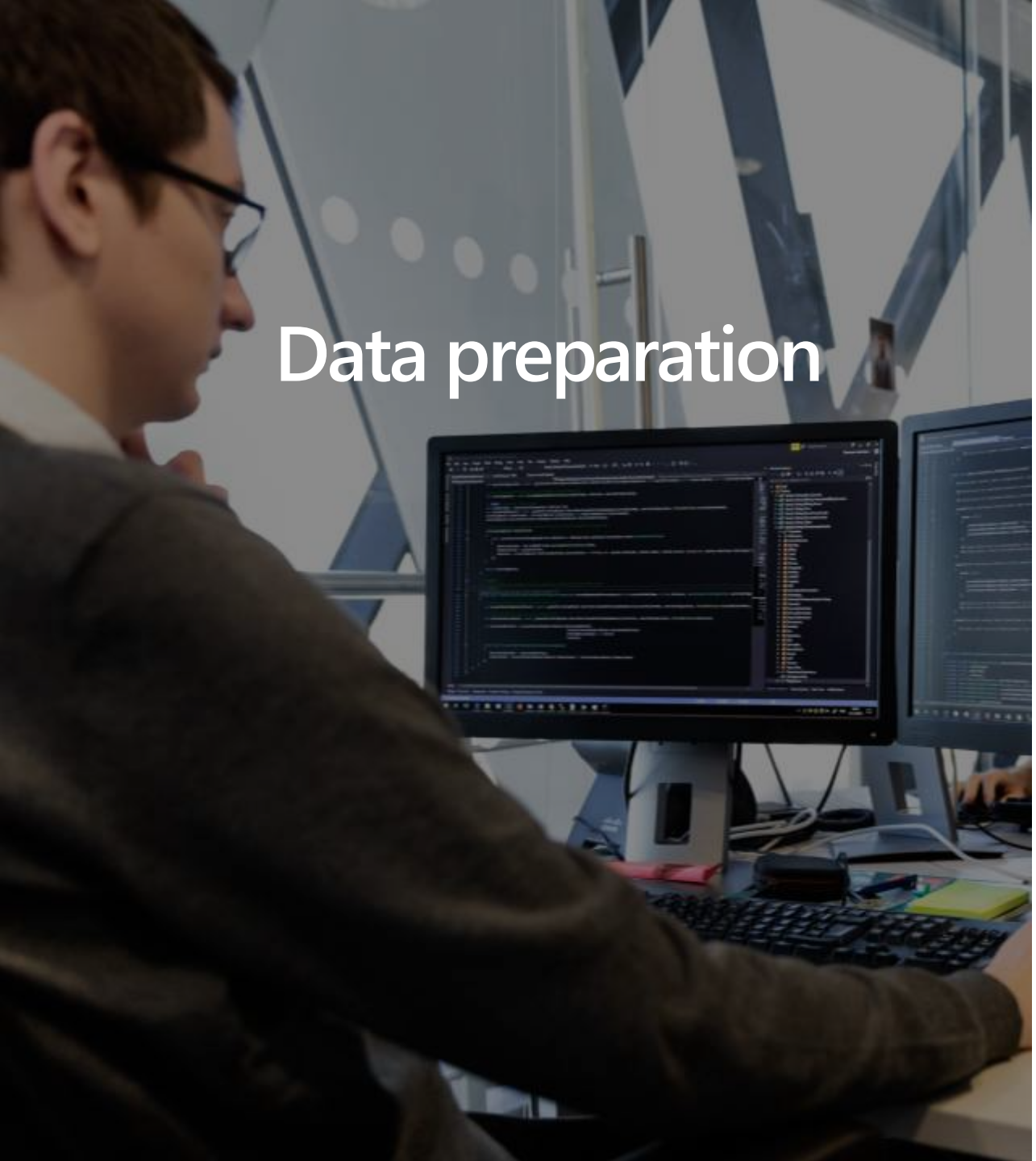
Ingesting data into Azure Data Lake with Azure Data Factory



Data warehousing pattern in Azure

Ingesting data into Azure Data Lake with Azure Data Factory





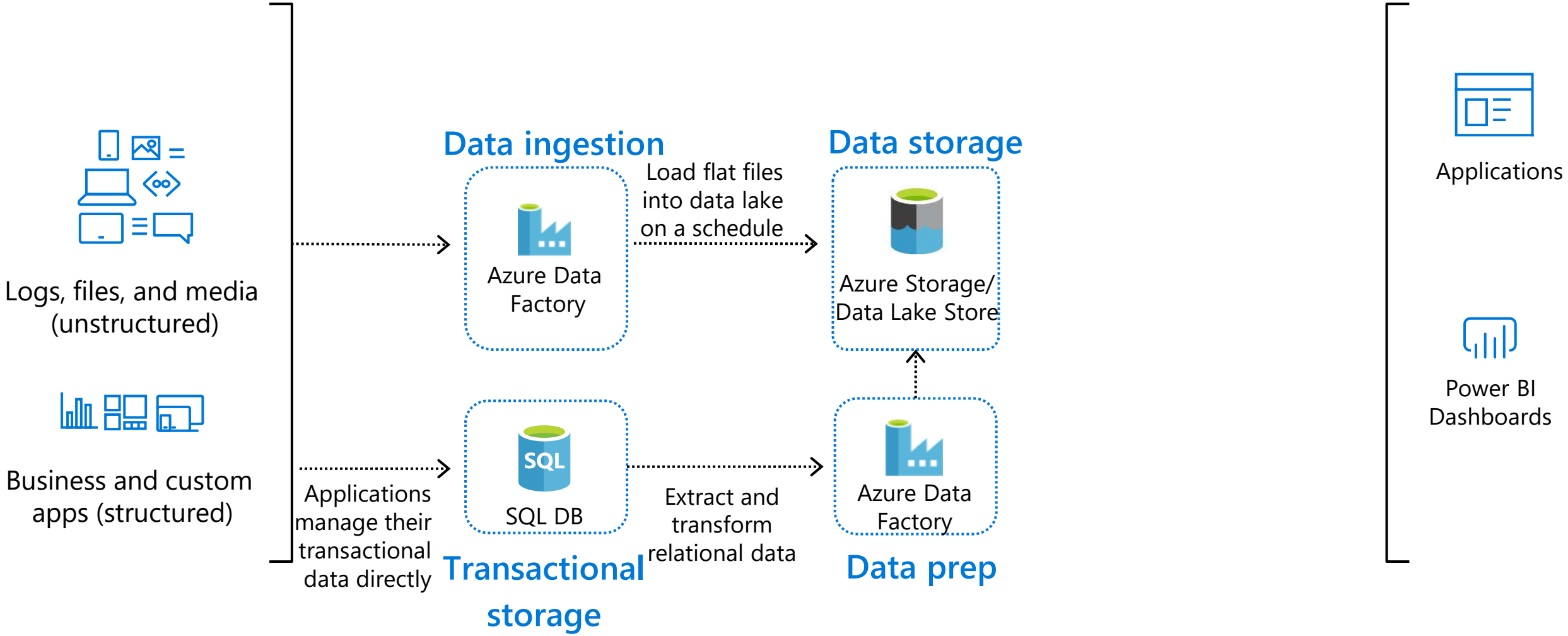
Data preparation

Is data cleansing, structuring, curation, and aggregation in data warehousing.

The data is batch processed in preparation for loading into a data warehouse

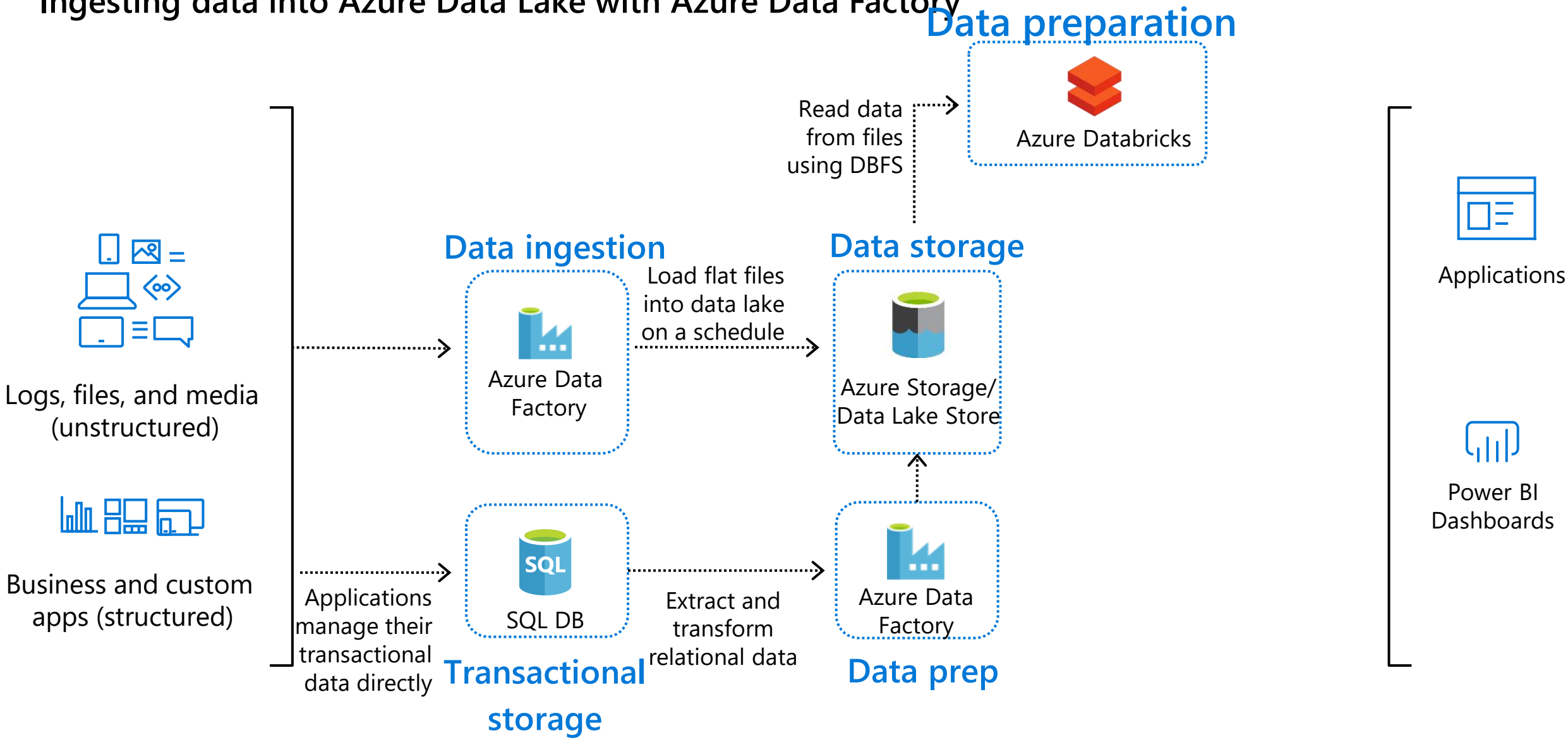
Data warehousing pattern in Azure

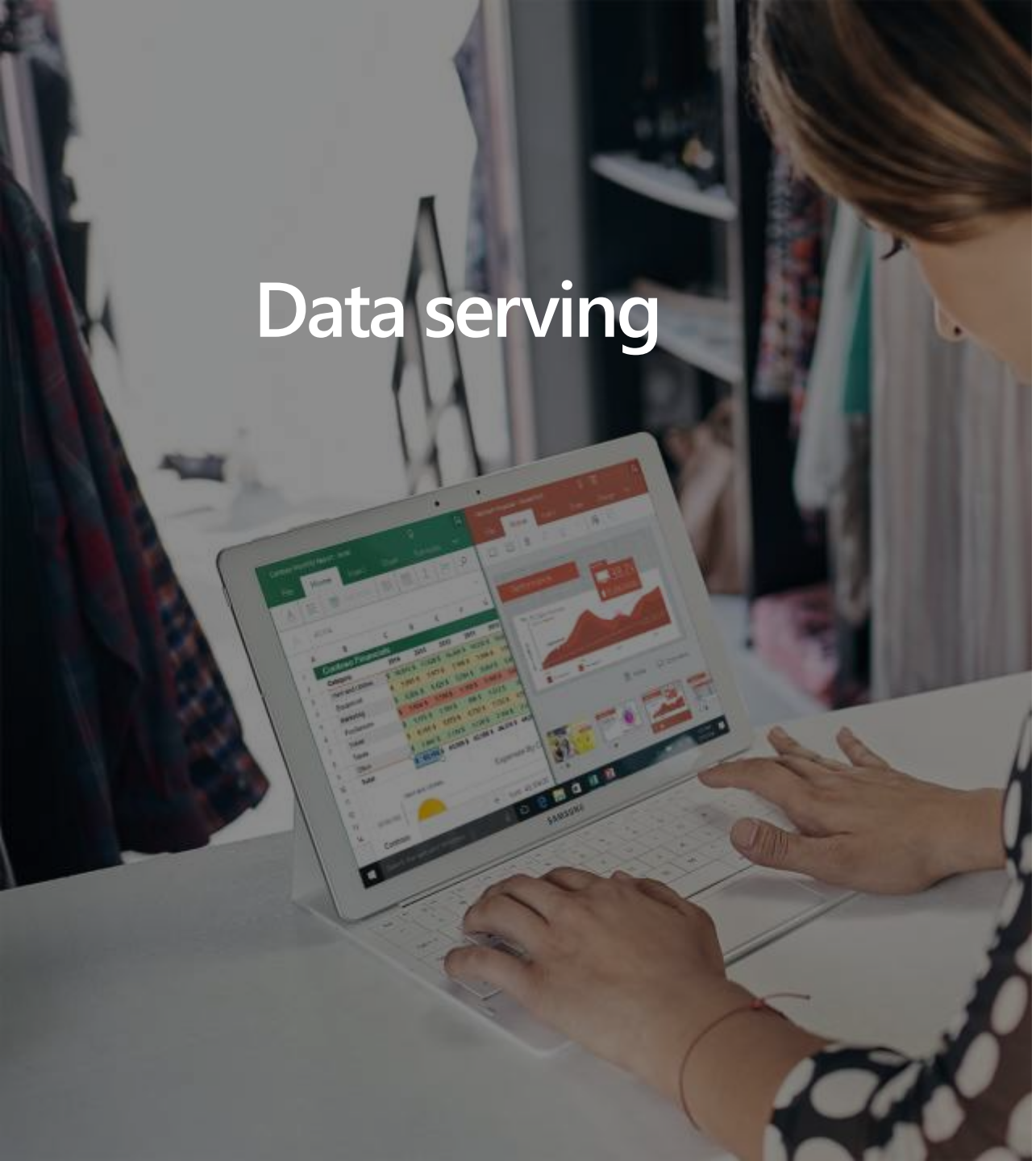
Ingesting data into Azure Data Lake with Azure Data Factory



Data warehousing pattern in Azure

Ingesting data into Azure Data Lake with Azure Data Factory





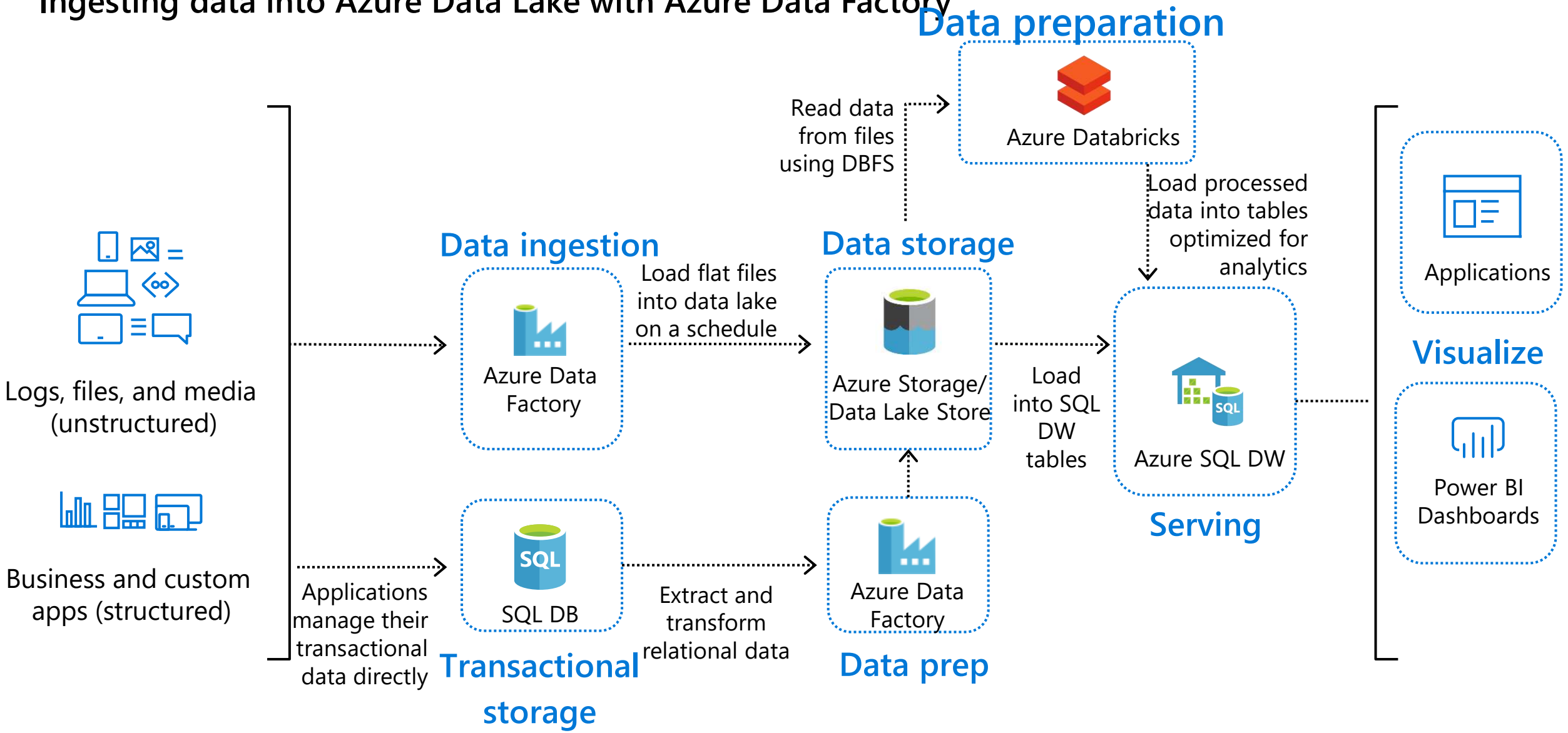
Data serving

Processed data served by a data warehouse to analytic clients and reporting tools

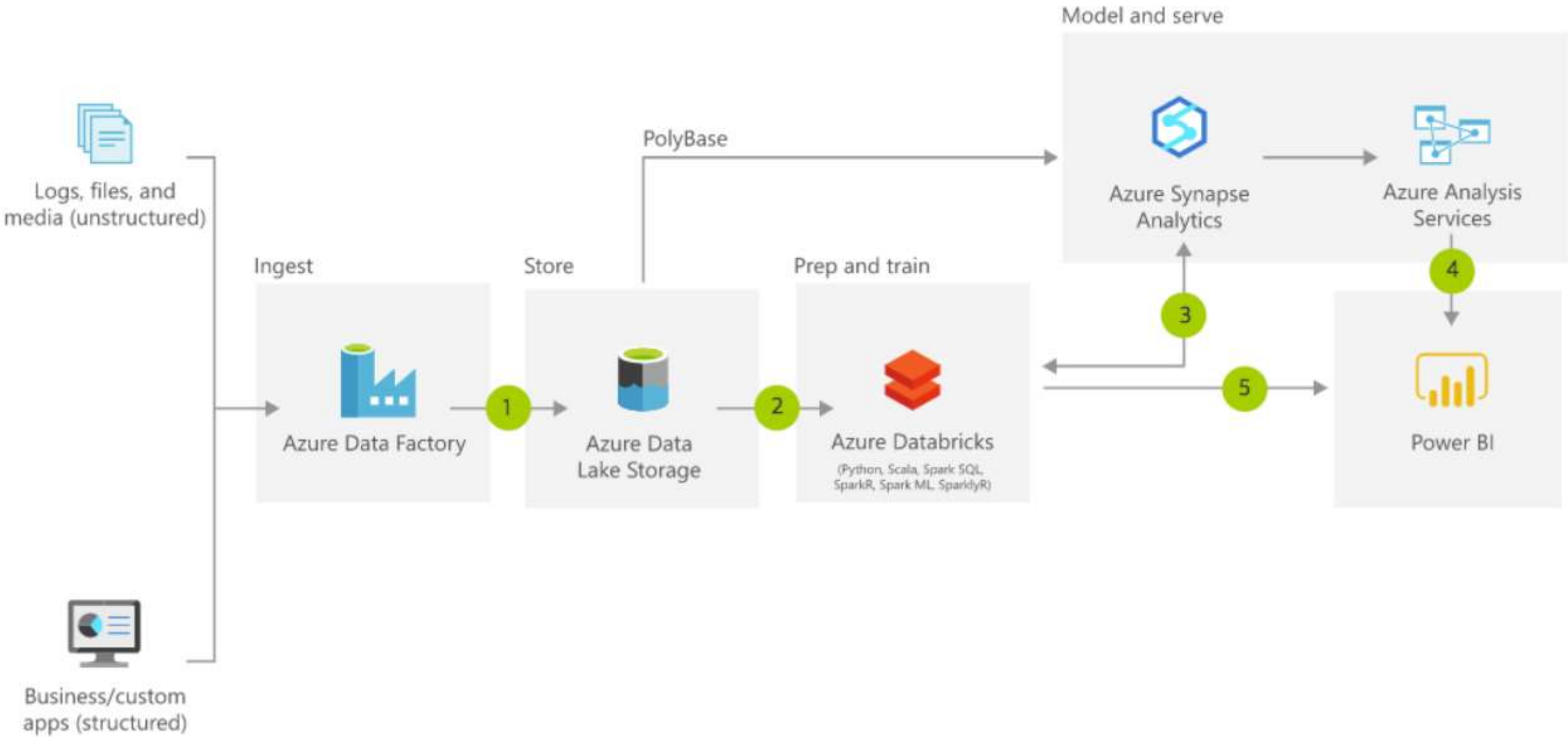
The data warehouse provides increased query flexibility and reduced query latency in comparison to batch data processing options

Data warehousing pattern in Azure

Ingesting data into Azure Data Lake with Azure Data Factory



Architecture



Modern Data Warehouse considerations

Big Data and advanced analytics

Security

Enables the modern data warehouse to control access in order to protect sensitive data and maintain desired compliance

Automation

Enables all components of the modern data warehouse solution to be controlled, deployed, and monitored programmatically

Monitoring

Provides insights into the status and health of the data warehouse solution

Microsoft

Ingesting Data with Azure Data Factory

What is Azure Data Factory?

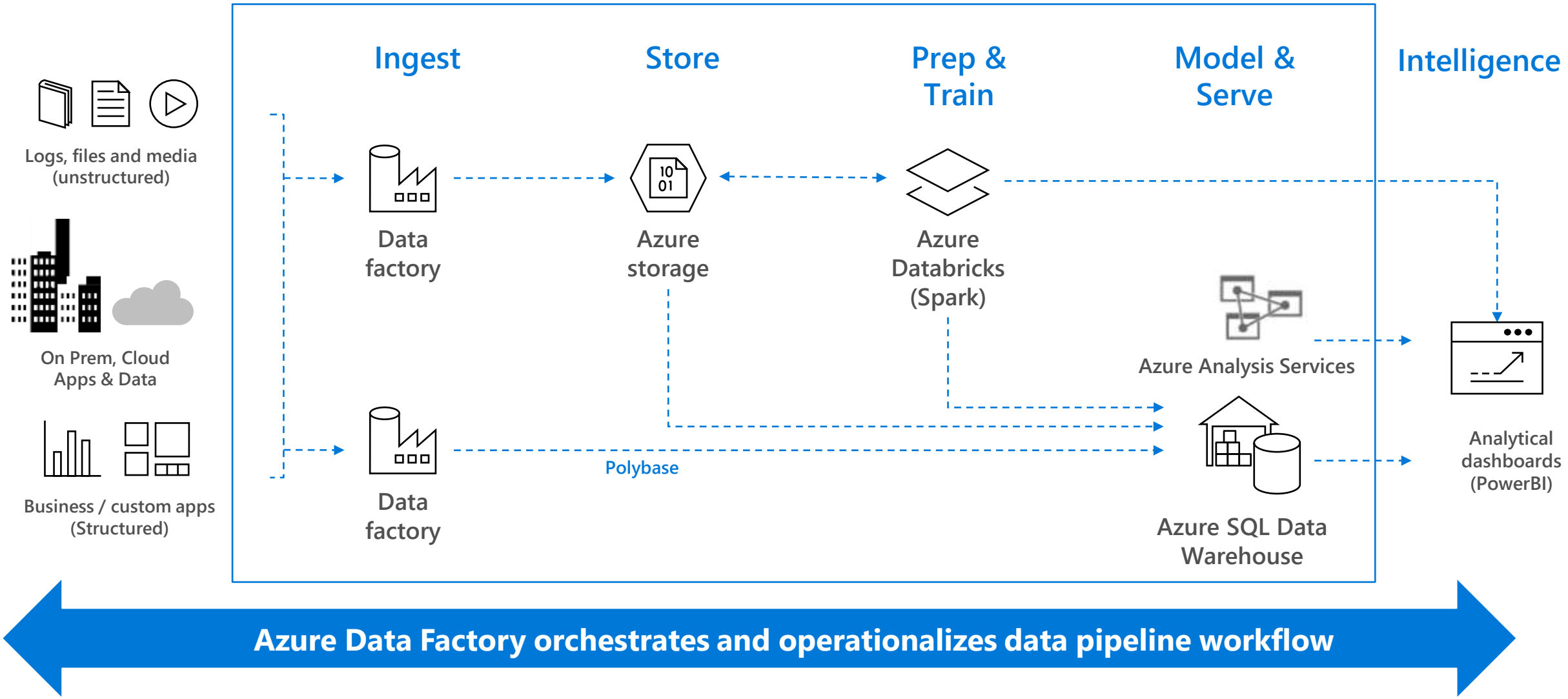
AZURE DATA FACTORY

A cloud-based data integration service that allows you to orchestrate and automate data movement and data transformation.

AZURE DATA FACTORY는

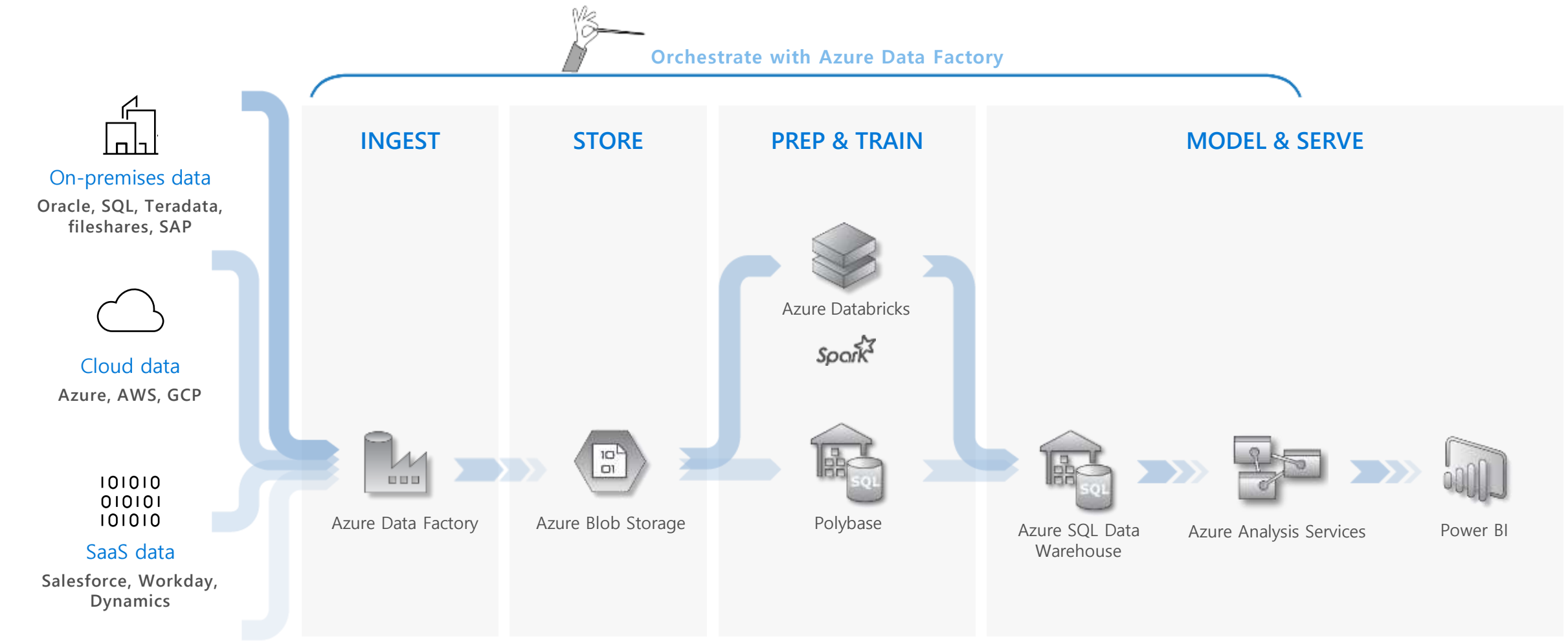
코딩이 필요 없는 하이브리드 데이터 통합 서비스 솔루션
- 다양한 데이터 소스에서 데이터를 추출하고
- 원하는 분석 엔진 또는 비즈니스 인텔리전스 도구에 게시
- 데이터 파이프라인을 모니터링 및 관
- 데이터가 클라우드와 온-프레미스 중 어디에 있든, 엔터프라이즈급 보안으로 작업
- 80개 이상의 데이터 원본 커넥터를 사용.
- 그래픽 사용자 인터페이스를 사용하여 데이터 파이프라인을 빌드하고, 모니터링하고, 관리

Modern DW for BI



AZURE DATA FACTORY

Modernize your enterprise data warehouse at scale



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

Hybrid and Multi-Cloud Data Integration

Azure Data Factory
PaaS Data Integration

Author, orchestrate and monitor with Azure Data Factory

On-Prem

SAP
ORACLE
cloudera
TERADATA
Microsoft SQL Server
HORTONWORKS

SaaS Apps

salesforce
Microsoft Dynamics
workday
Adobe

Public Cloud

Azure
amazon web services
Google Cloud Platform

DATA DRIVEN APPLICATIONS

DATA SCIENCE AND MACHINE LEARNING MODELS

ANALYTICAL DASHBOARDS USING POWER BI

Copy Data

1 Properties
One time copy

2 Source
- Connection
- Dataset

3 Destination

4 Settings
Fault tolerance

5 Summary

6 Deployment

Source data store

Specify the source data store for the copy task. You can use an existing data store connection or specify a new data store. Click HERE to suggest new copy sources or give comments.

FROM EXISTING CONNECTIONS | CONNECT TO A DATA STORE

Amazon Redshift	Amazon S3	Azure Blob Storage	Azure Cosmos DB	Azure Data Lake Store	Azure Database for MySQL
Azure Database for PostgreSQL	Azure File Storage	Azure SQL Data Warehouse	Azure SQL Database	Azure Table Storage	Cassandra

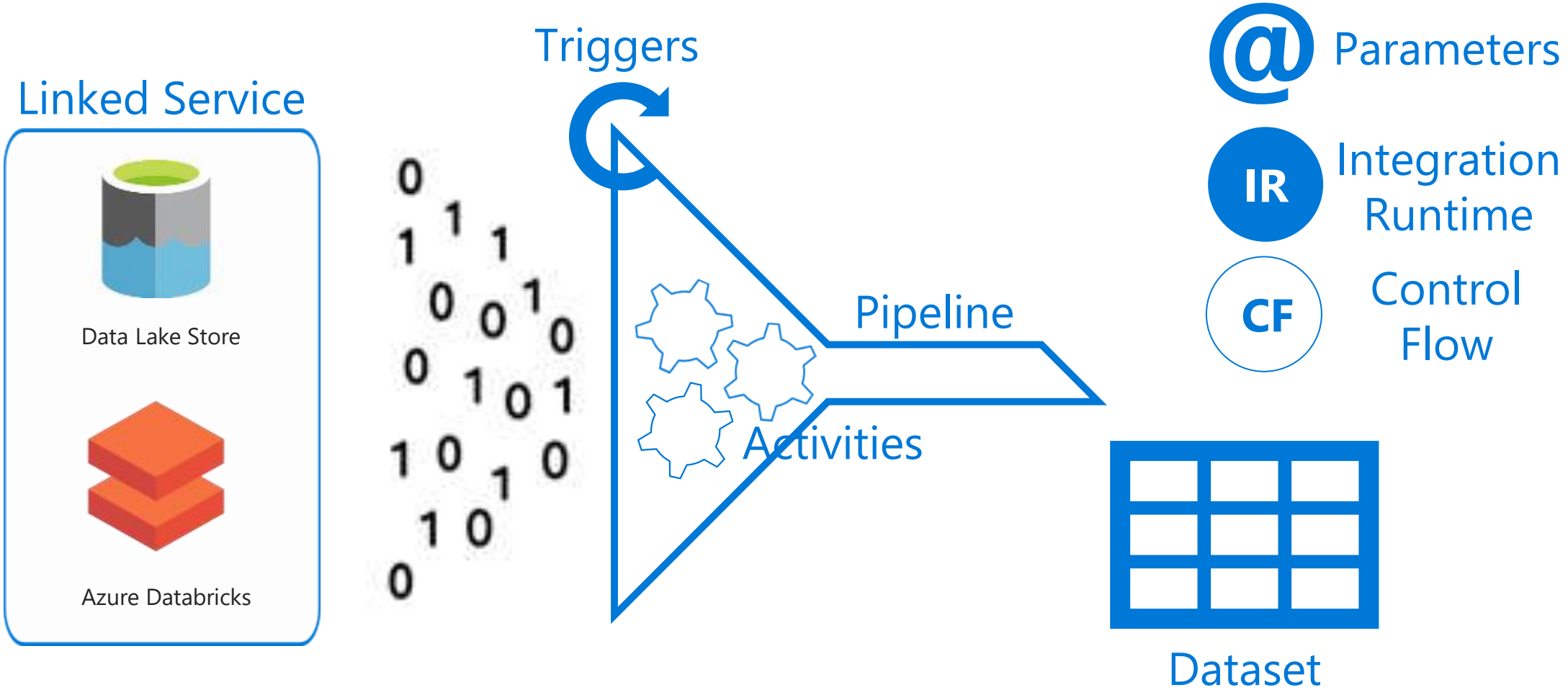
Previous | Next

Access all your data – 90+ connectors & growing

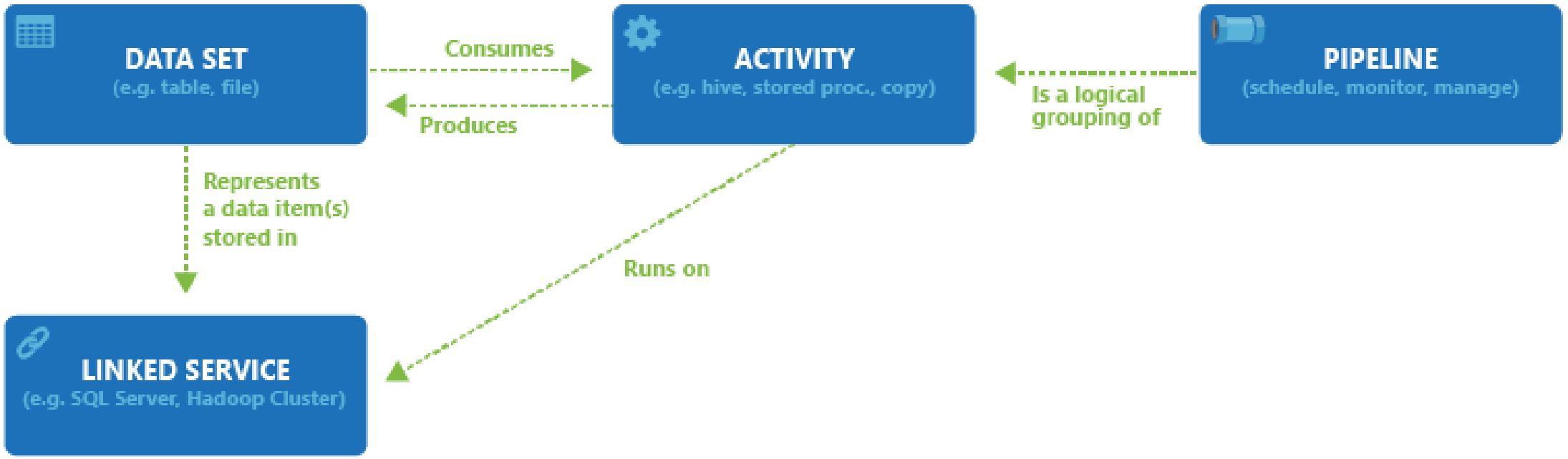
Azure (12)	Database (24)		File Storage (5)	NoSQL (3)	Services and Apps (28)		Generic (3)
Blob Storage	Amazon Redshift	Netezza	Amazon S3	Cassandra	Amazon MWS	Oracle Service Cloud	HTTP
Cosmos DB (SQL API)	DB2	Oracle	File System	Couchbase	Common Data Service for Apps	Paypal	OData
Data Lake Storage Gen1	Drill	Phoenix	FTP	MongoDB	Concur	QuickBooks	ODBC
Data Lake Storage Gen2	Google BigQuery	PostgreSQL	HDFS		Dynamics 365	Salesforce	
DB for MySQL	Greenplum	Presto	SFTP		Dynamics CRM	Salesforce Marketing Cloud	
DB for PostgreSQL	HBase	SAP BW			GE Historian	Salesforce Service Cloud	
File Storage	Hive	SAP HANA			Google AdWords	SAP C4C	
SQL DB	Impala	Spark			HubSpot	SAP ECC	
SQL DB Managed Instance	Informix	SQL Server			Jira	ServiceNow	
SQL DW	MariaDB	Sybase			Magento	Shopify	
Search Index	Microsoft Access	Teradata			Marketo	Square	
Table Storage	MySQL	Vertica			Office 365	Web table	
					Oracle Eloqua	Xero	
					Oracle Responsys	Zoho	

* Supported file formats: CSV, Parquet, AVRO, ORC, JSON

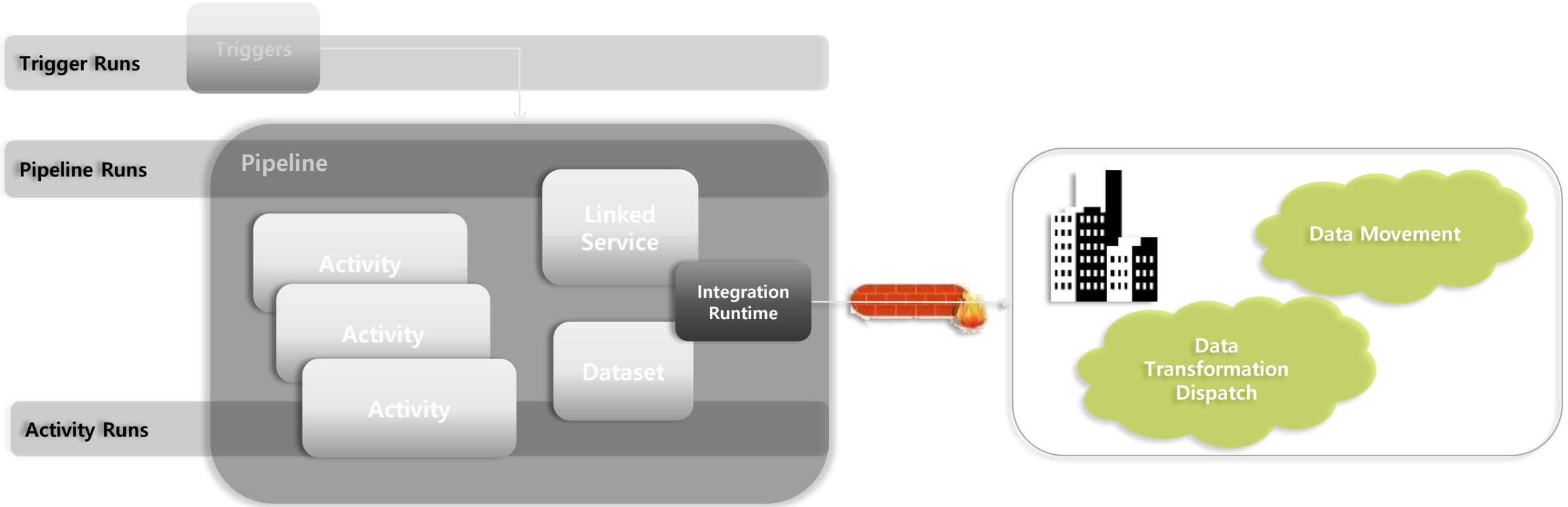
AZURE DATA FACTORY COMPONENTS



COMPONENT DEPENDENCIES

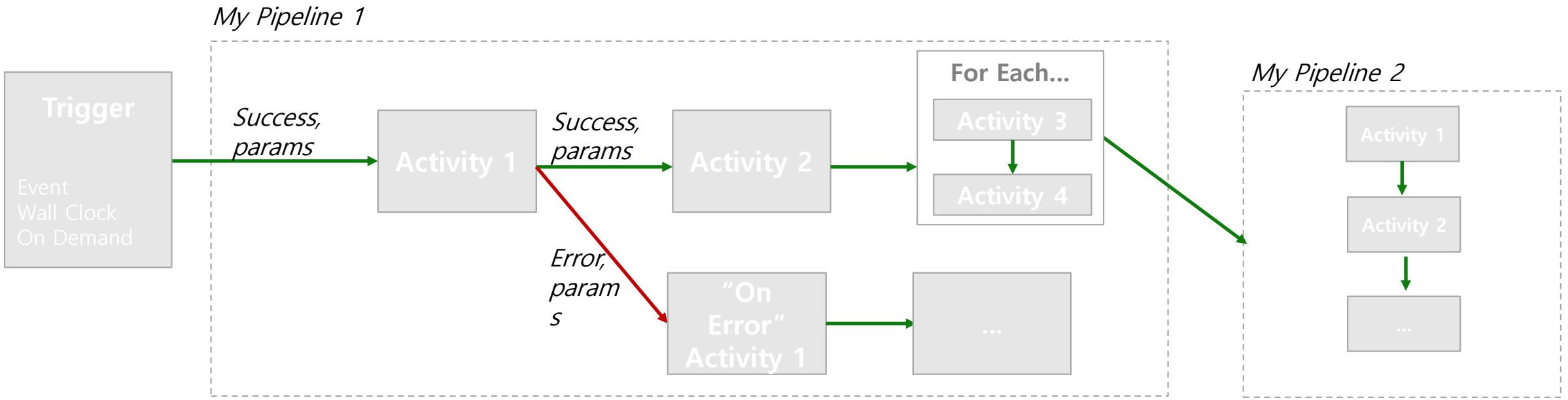


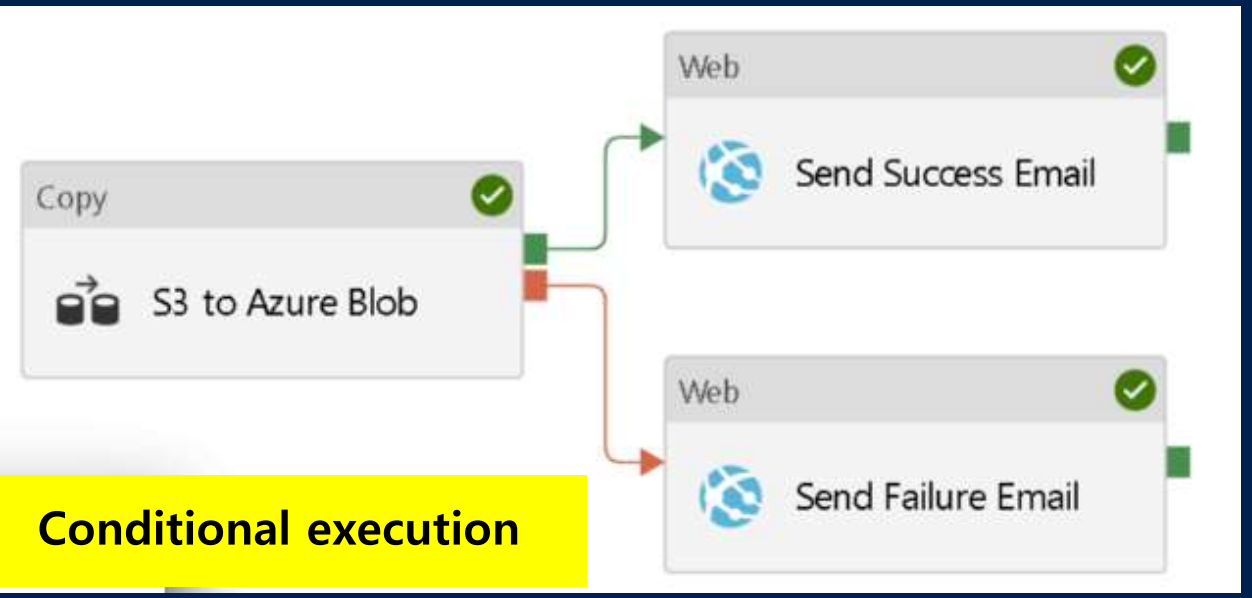
Azure Data Factory Updated Flexible Application Model



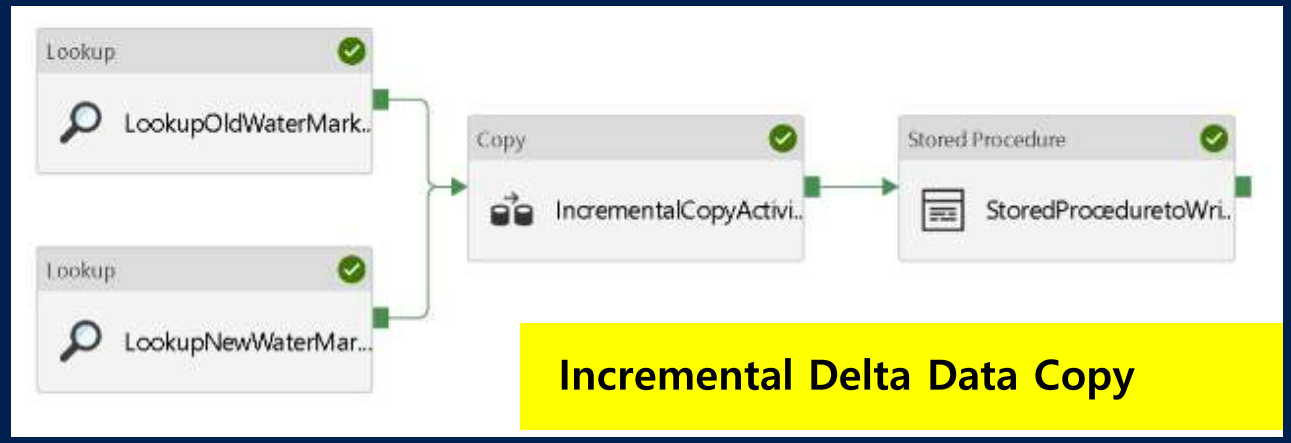
Control Flow Introduced in Azure Data Factory

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

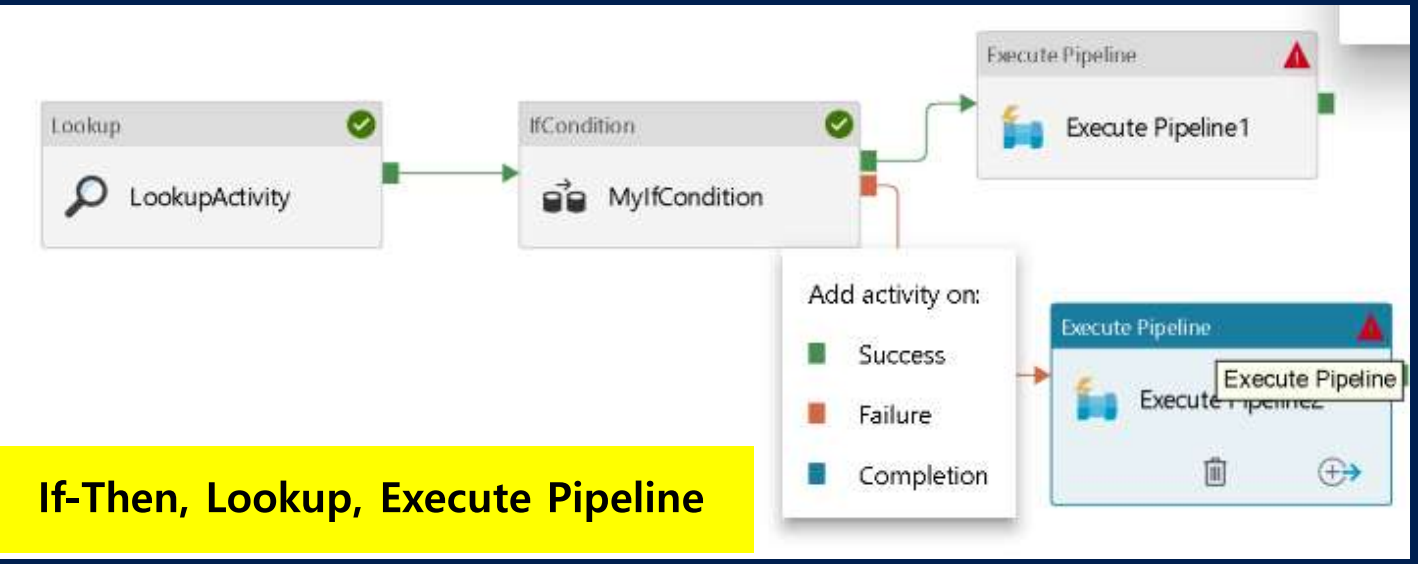




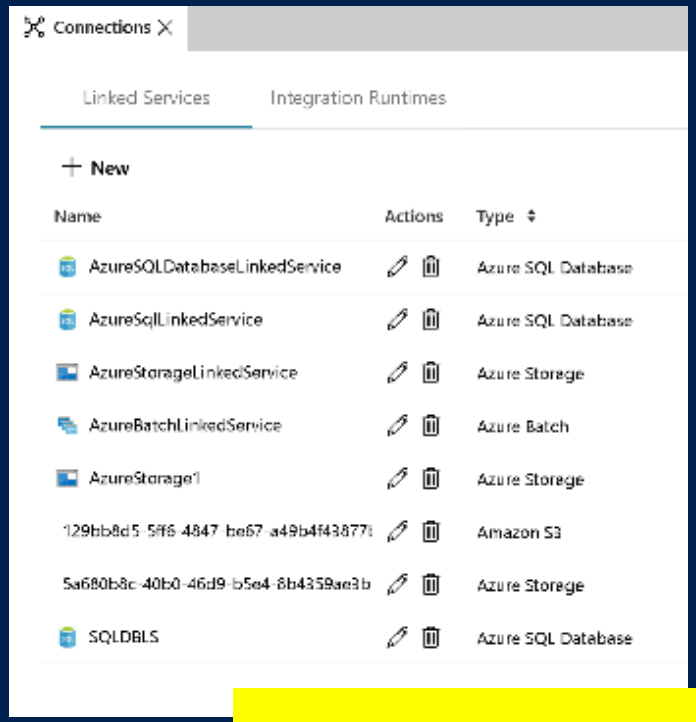
Conditional execution



Incremental Delta Data Copy



If-Then, Lookup, Execute Pipeline



Connection Managers

myadfv2 | Monitor Pipeline Runs

Operationalize – Monitor your data pipelines

Refresh

Custom Range 11/01/2017 9:00 AM - 12/23/2017 9:00 AM

Time Zone (UTC-08:00) Los Angeles

All Succeeded In Progress Failed

Pipeline Name	Actions	Run Start	Duration	Triggered By	Status	Parameters	Error	RunID
LookupPipeline		12/04/2017, 4:59:33 PM	00:00:49	Manual trigger	Succeeded			8fd7c2e1-440c-45d7-aff0-21dc8552c207
LookupPipeline		12/04/2017, 4:56:24 PM	00:00:53	Manual trigger	Succeeded			ecd6bec4-b7b8-47b0-aaac-c32ba199a5ff
LookupPipeline		12/04/2017, 4:53:34 PM	00:00:33	Manual trigger	Failed			c272ebf7-f784-4d8c-9b82-c5e10f06250b
LookupPipeline		12/04/2017, 4:20:25 PM	00:00:29	Manual trigger	Failed			6018a772-81c8-4ec0-ab18-24424c25195c
LookupPipeline		12/04/2017, 4:10:50 PM	00:00:33	Manual trigger	Failed			06c7db30-d77b-47d2-917a-935244f1c2c5
pipeline4__7e0990af-c...		11/27/2017, 11:12:27 AM	00:00:05	Manual trigger	Failed			c3aa1144-ebdc-448b-a1b8-9f1b5d65cb40
MyWebActivityPipeline		11/26/2017, 9:37:02 PM	00:00:10	Manual trigger	Failed			23c5e44c-a191-4a1f-ac21-ff276b7da43b
batchpipe		11/17/2017, 3:24:19 PM	00:00:38	Manual trigger	Succeeded			b2ef549a-b5cf-4786-9ffd-f9f71948c6d9
batchpipe		11/17/2017, 3:20:12 PM	00:00:00	Manual trigger	Failed			a3dec17f-a370-4e8b-9a3e-285483680fde
ifconditionpipeline2		11/16/2017, 6:00:20 PM	00:00:04	Manual trigger	Failed			07b7812d-0af0-4f67-a0b8-ec64ddd38fc9
ifconditionpipeline		11/16/2017, 6:00:11 PM	00:00:05	Manual trigger	Failed			8ac7565d-eefd-4831-92c5-33bfebdf2c60
ifconditionpipeline		11/15/2017, 4:58:45 PM	00:00:07	Manual trigger	Succeeded			dcff3e04-6158-40e7-b21d-70d417ae646f
ifconditionpipeline		11/15/2017, 4:52:36 PM	00:00:06	Manual trigger	Failed			f1d615ca-f4d9-47bf-930b-0bc47dbb3430
pipeline3__9a1f3c55-e...		11/10/2017, 2:52:13 PM	00:00:05	Manual trigger	Failed			052056da-9cd6-48c8-8441-4d11feb911a4
IncrementalCopyPipeli...		11/01/2017, 2:02:16 PM	00:01:36	Manual trigger	Succeeded			f176d4e0-1535-4aec-8eca-25dc7a4b0e80
IncrementalCopyPipeli...		11/01/2017, 1:56:06 PM	00:01:13	Manual trigger	Succeeded			1f3d9bc2-9b30-4245-9489-786ca77796ca
IncrementalCopyPipeli...		11/01/2017, 1:49:30 PM	00:00:36	Manual trigger	Failed			7824bd16-9e72-4409-ae80-238faf861a5c

Copy Data

1 Properties
One time copy

2 Source
- Connection
- Dataset

3 Destination

4 Settings
Fault tolerance

5 Summary

6 Deployment

Source data store

Specify the source data store for the copy task. You can use an existing data store connection or specify a new data store. Click HERE to suggest new copy sources or give comments.

Easy-to-use Wizard for Copying Data at Scale

FROM EXISTING CONNECTIONS | CONNECT TO A DATA STORE

Amazon Redshift	Amazon S3	Azure Blob Storage	Azure Cosmos DB	Azure Data Lake Store	Azure Database for MySQL
Azure Database for PostgreSQL	Azure File Storage	Azure SQL Data Warehouse	Azure SQL Database	Azure Table Storage	Cassandra

Previous | Next

ADF Certifications

HIPAA/HITECH

ISO/IEC 27001

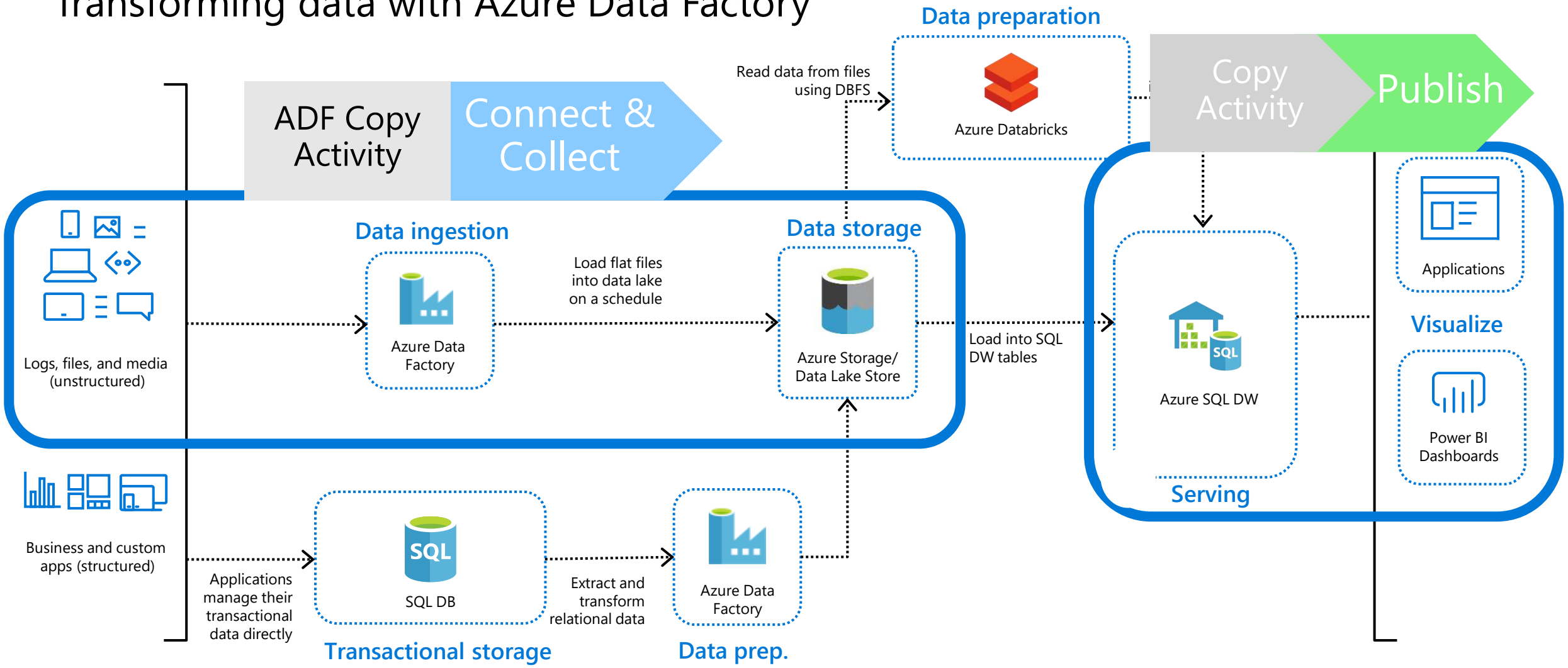
ISO/IEC 27018

CSA STAR

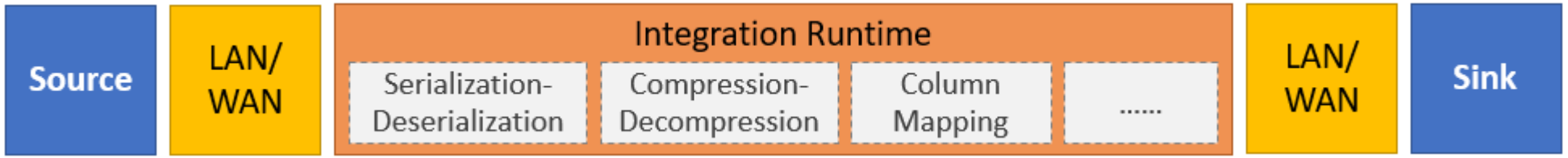
Ingesting data

Data transformation in Azure

Transforming data with Azure Data Factory

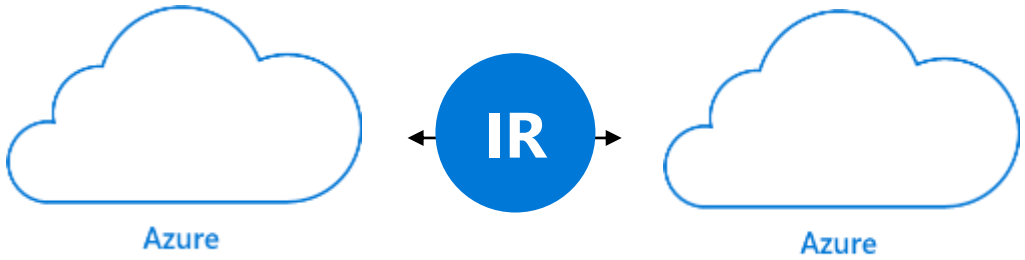


COPY ACTIVITY PROCESS

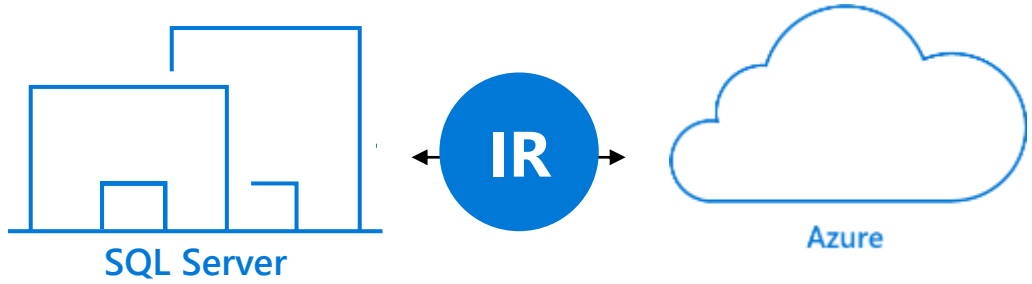


- Reads data from a source data store.
- Performs serialization/deserialization, compression/decompression, column mapping, and so on. It performs these operations based on the configuration of the input dataset, output dataset, and Copy activity.
- Writes data to the sink/destination data store

INTEGRATION RUNTIME

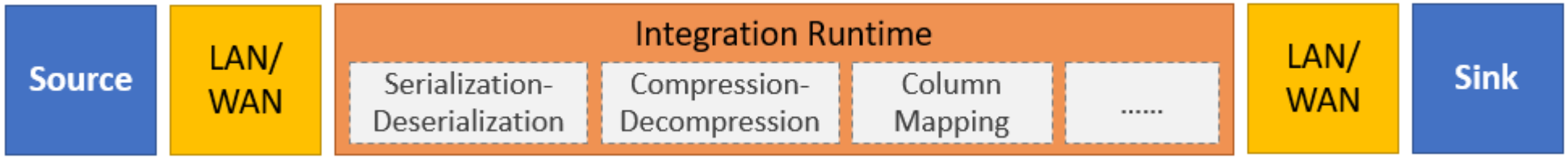


Azure Integration Runtime



Self-hosted Integration Runtime

COPY FILES WITH THE COPY ACTIVITY



Supported file formats:

- Text
- JSON
- Avro
- ORC
- Parquet

Copy activity can compress and decompress files with The following codecs:

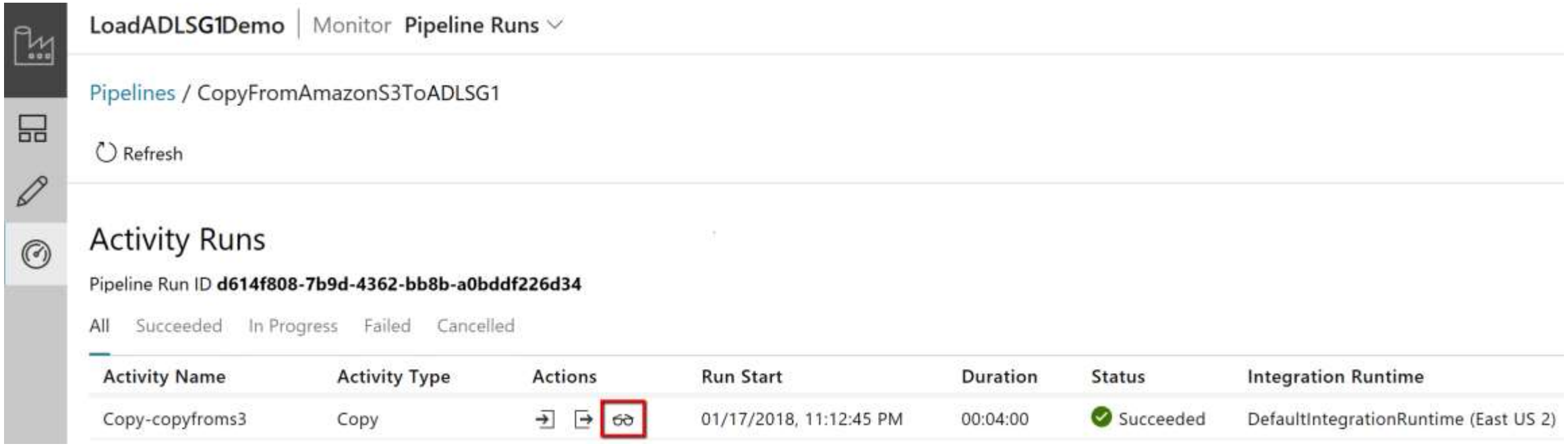
- Gzip
- Deflate
- Bzip2
- ZipDeflate

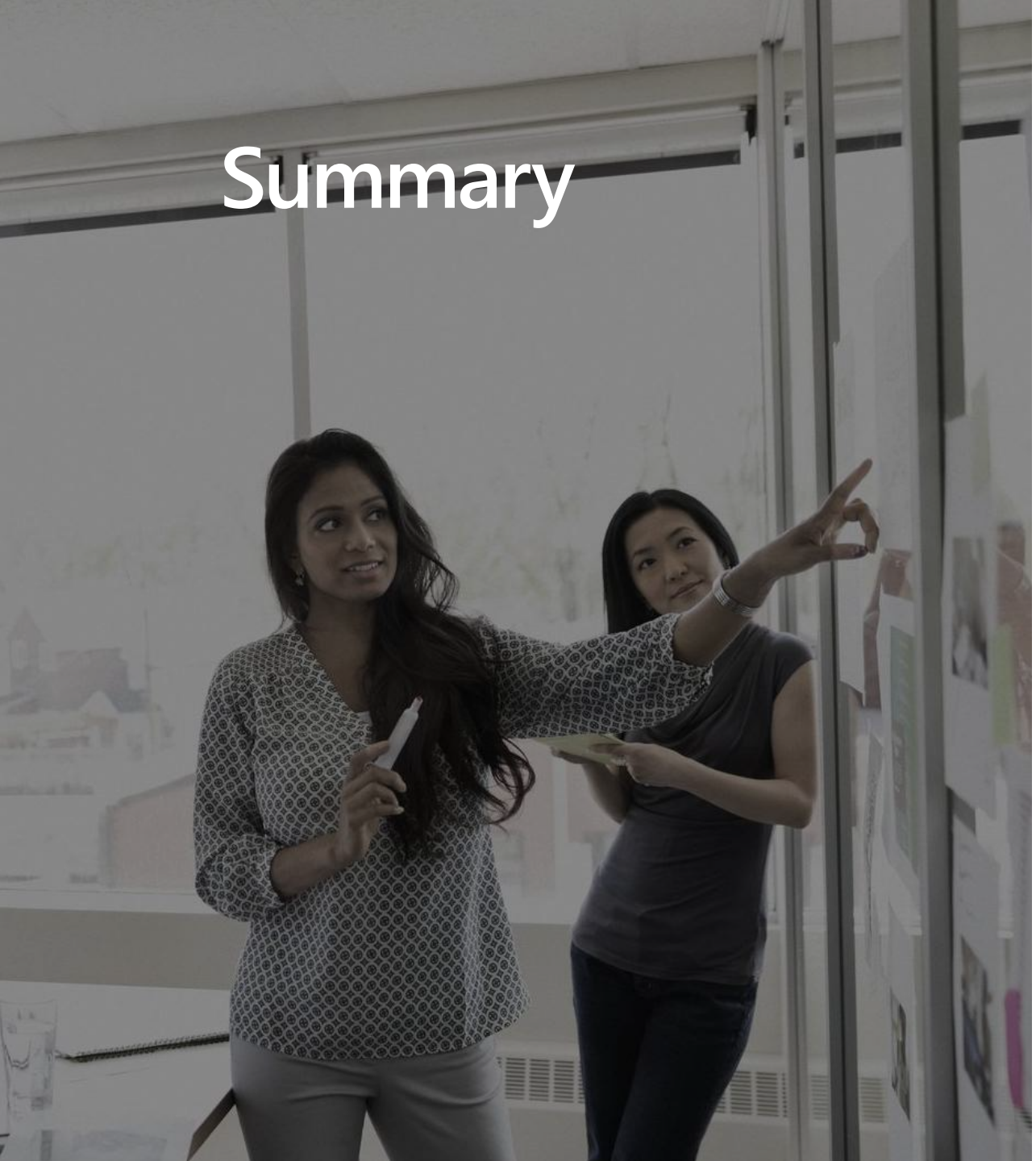
Monitoring data ingestion

Monitoring

Pipeline runs

Activity runs





Summary

- Azure Data Factory (ADF) is a cloud-based data integration service that allows you to orchestrate and automate data movement and data transformation.
- Ingesting data can be performed by the ADF Copy Activity
- The ADF Copy Activity can be used to connect and collect data for ingestion, and to publish data to BI tools and applications.
- Different Integration Runtimes are required for different ingestion scenarios
- File copy are very efficient using the ADF Copy Activity
- You can monitor the performance of the ADF Copy Activity both visually and programmatically

Microsoft

Transform your Data with Azure Data Factory

Speaker name

Title

RESOURCES

Session Resources Hub

aka.ms/DATA30

Session Code on GitHub

aka.ms/DATA30Repo

All Event Session Resources

aka.ms/mymsignitethetour

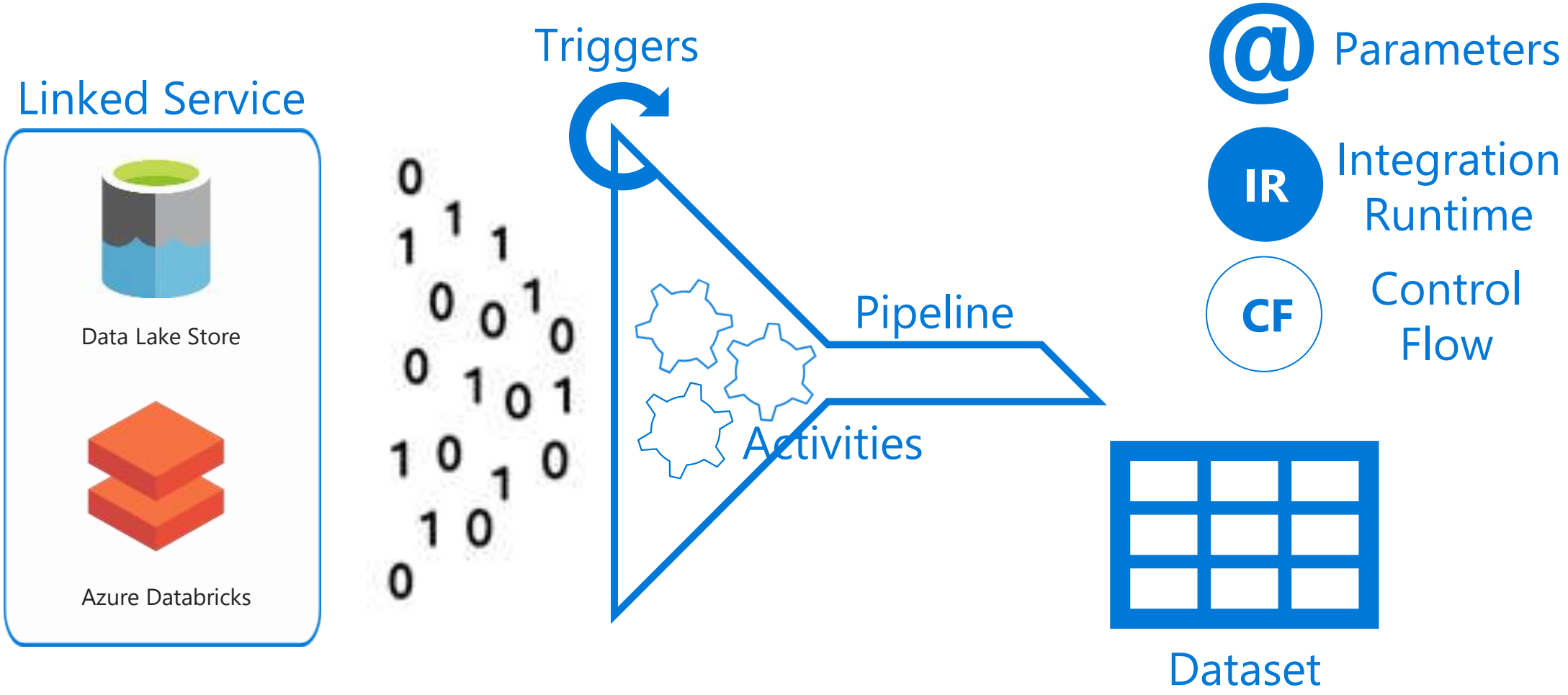
What is Azure Data Factory?

AZURE DATA FACTORY

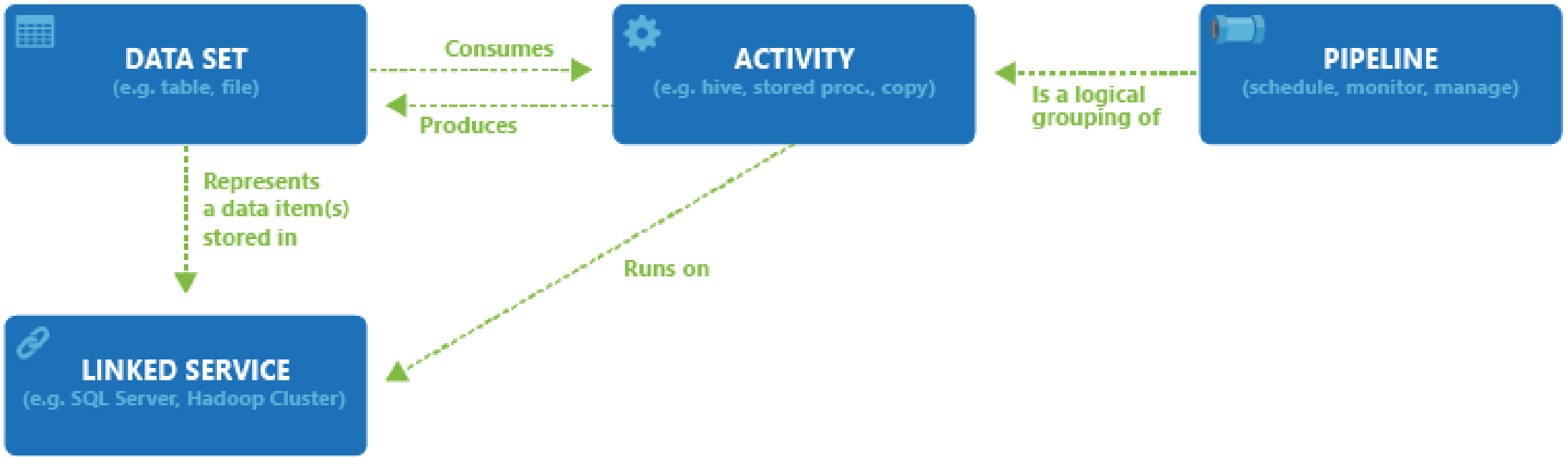
A cloud-based data integration service that allows you to orchestrate and automate data movement and data transformation.

AZURE DATA FACTORY PROCESS

AZURE DATA FACTORY COMPONENTS



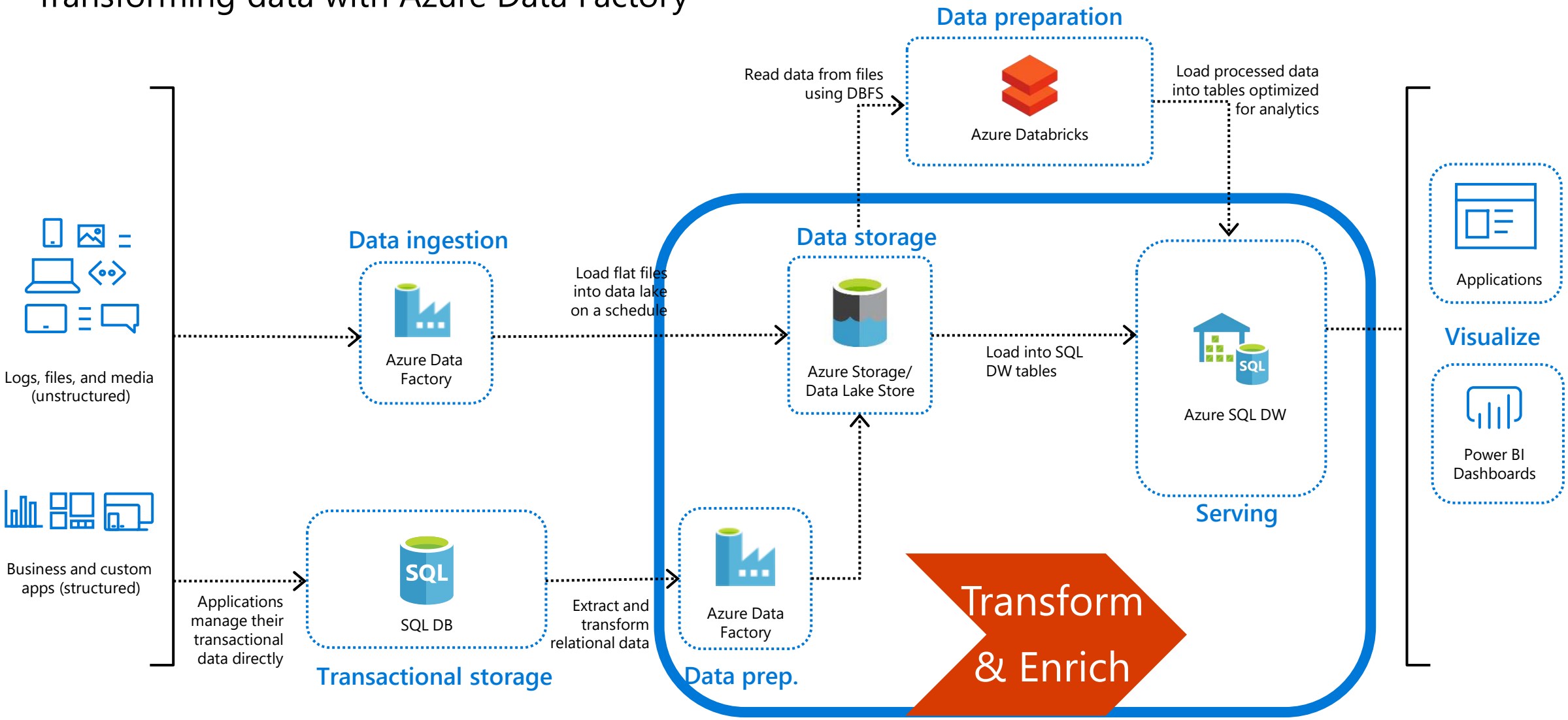
COMPONENT DEPENDENCIES



Transforming data with the ADF Mapping Data Flow

DATA TRANSFORMATION IN AZURE

Transforming data with Azure Data Factory

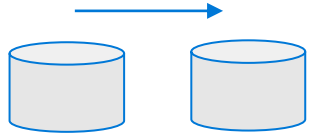


METHODS FOR TRANSFORMING IN AZURE DATA FACTORY

Compute resources

SSIS Packages

Mapping Data Flow



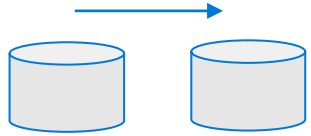
METHODS FOR TRANSFORMING DATA IN AZURE DATA FACTORY

- Code free data transformation at scale

Compute resources

SSIS Packages

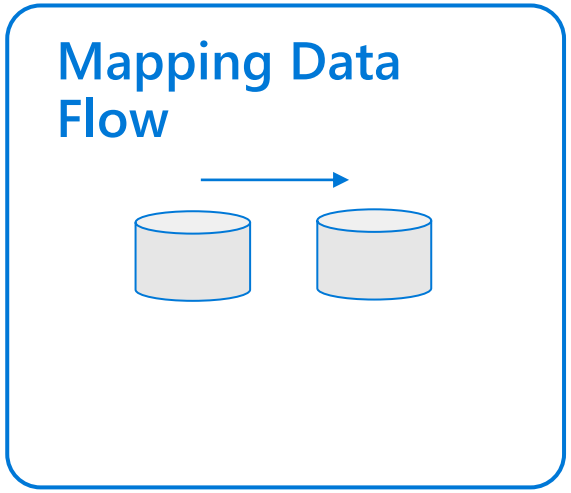
Mapping Data Flow



BENEFITS OF MAPPING DATA FLOW

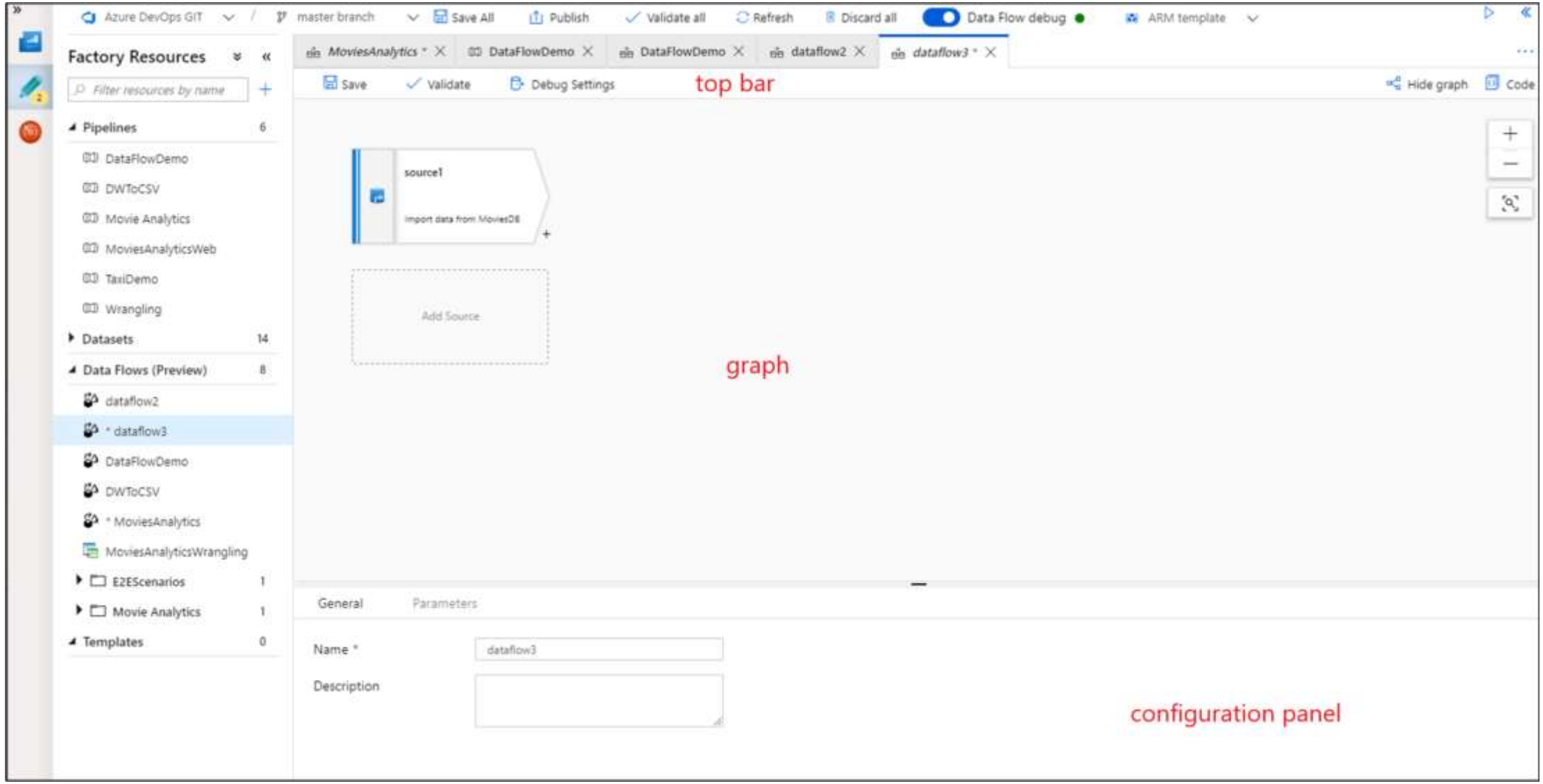
Code free data transformation at scale

- Perform data cleansing, transformation, aggregations, etc.
- Enables you to build resilient data flows in a code free environment
- Enable you to focus on building business logic and data transformation
- Underlying infrastructure is provisioned automatically with cloud scale via Spark execution



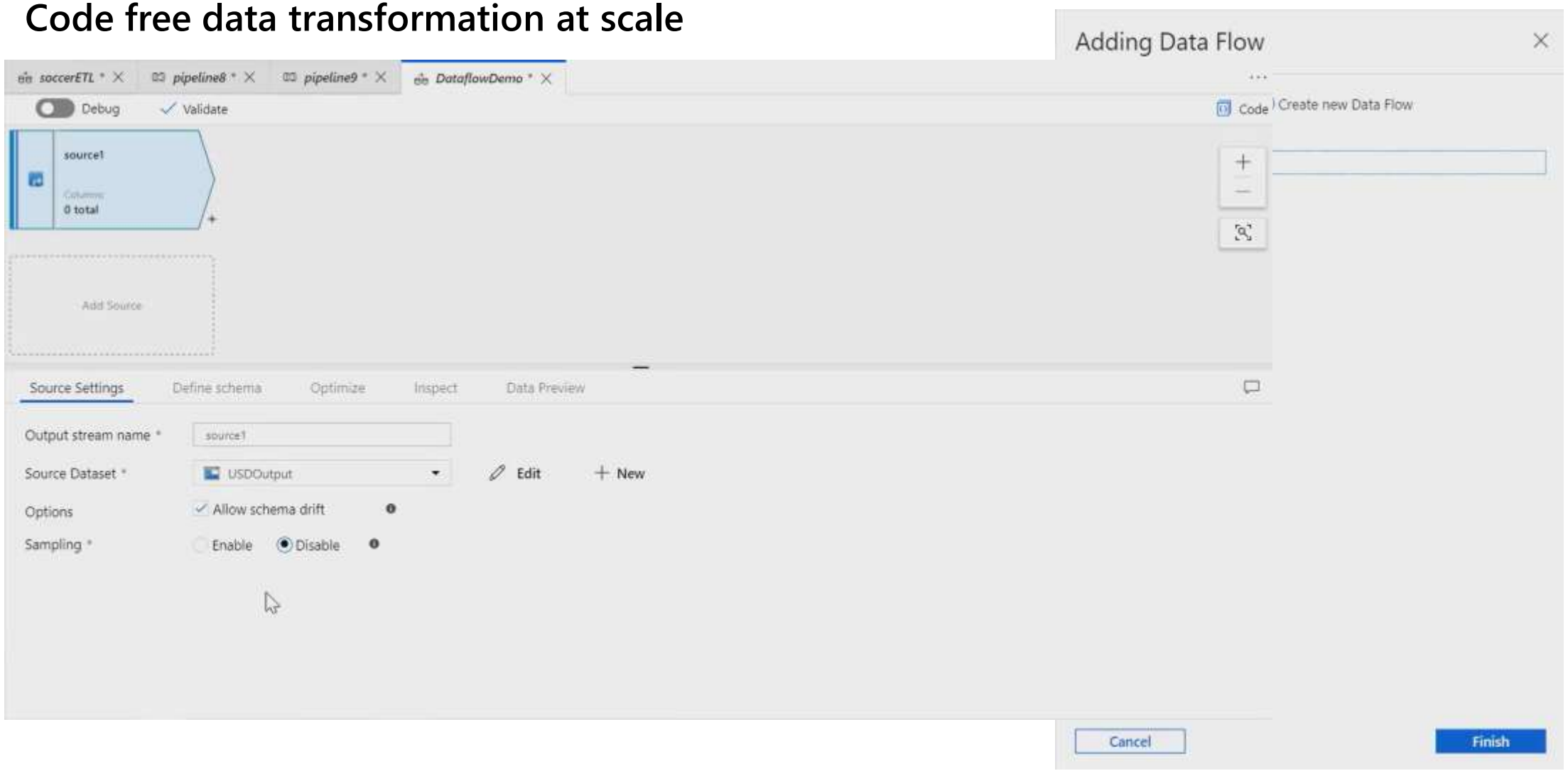
USING THE MAPPING DATA FLOW

Code free data transformation at scale

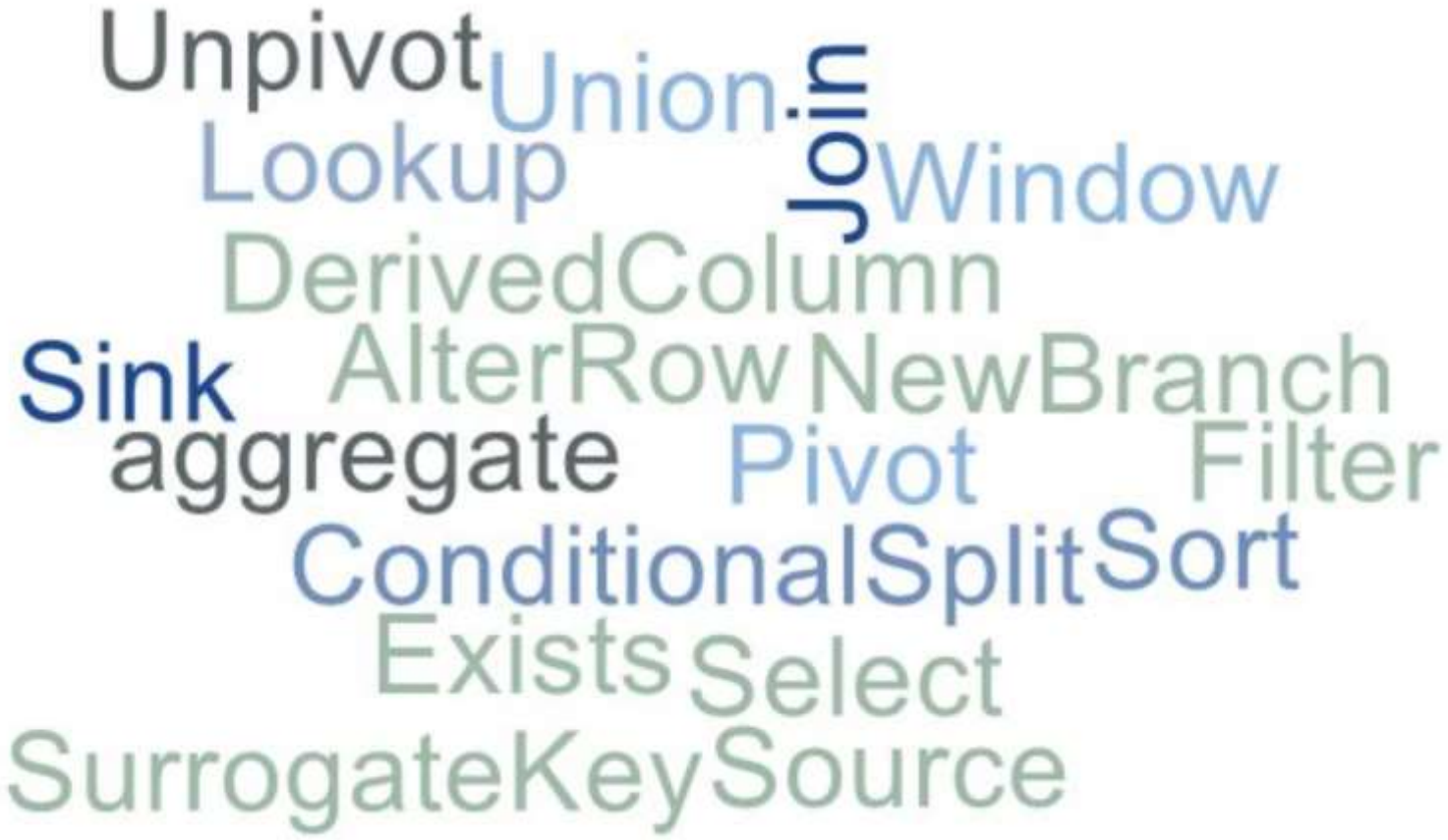


STARTING THE MAPPING DATA FLOW

Code free data transformation at scale



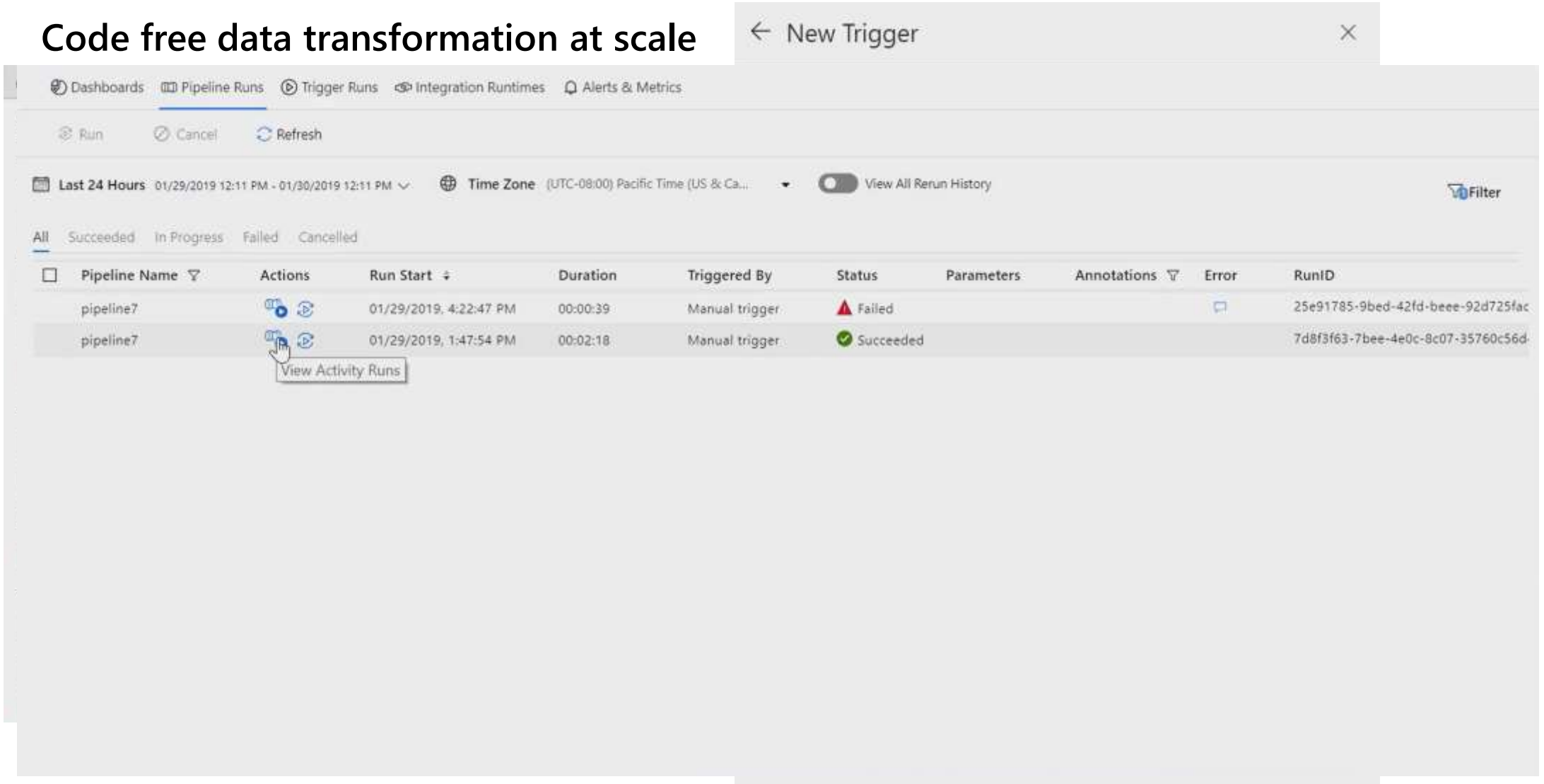
TRANSFORMATION OPTIONS IN THE MAPPING DATA FLOW

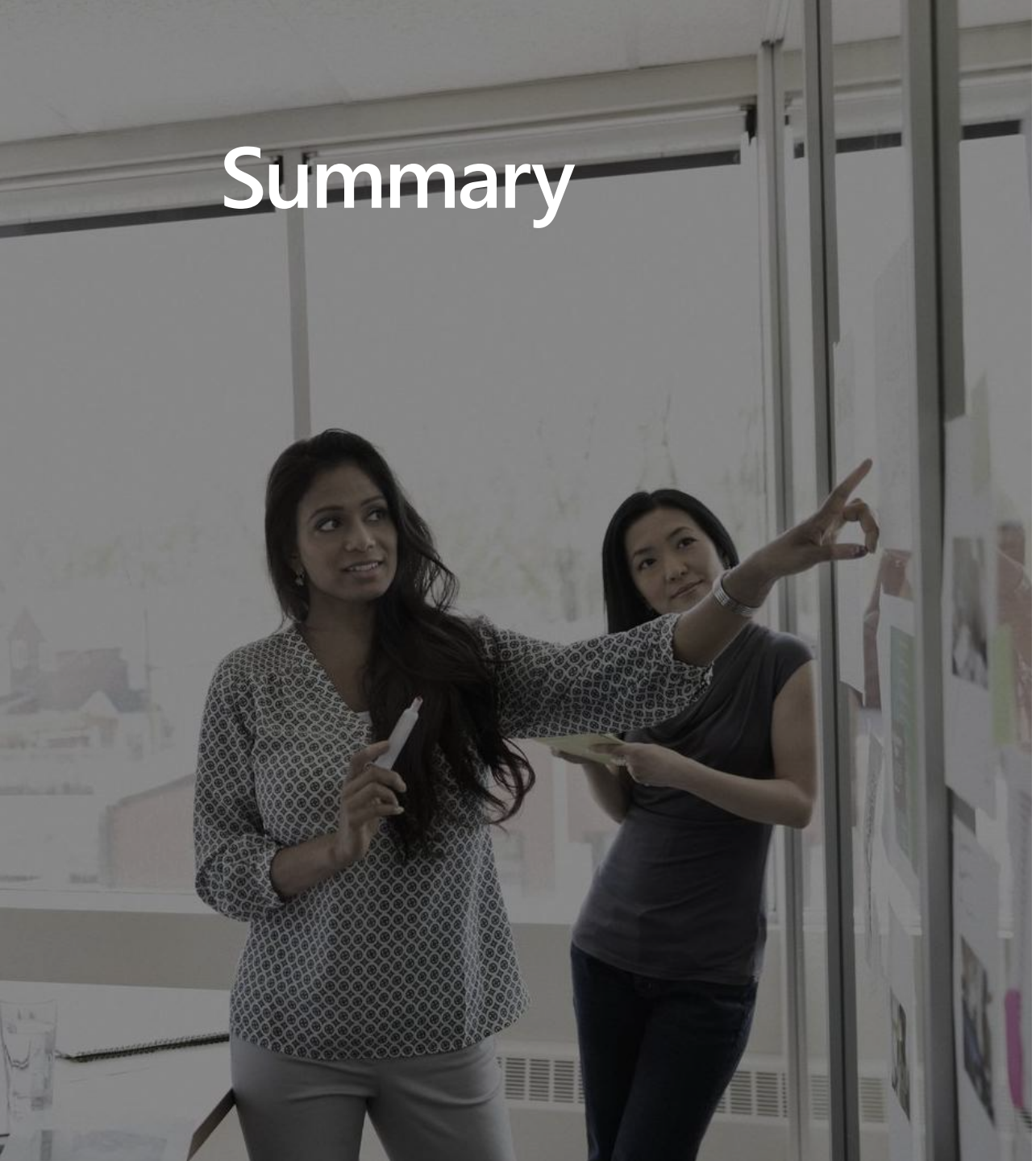


Triggering and monitoring

TRIGGERING THE MAPPING DATA FLOW

Code free data transformation at scale





Summary

- Azure Data Factory (ADF) is a cloud-based data integration service that allows you to orchestrate and automate data movement and data transformation.
- Transforming data can be performed in ADF by orchestrating a compute resource, calling an SSIS package or using the Mapping Data Flow feature
- The Mapping Data Flow feature enables code free data transformation at scale
- Enable you to focus on building business logic and data transformation
- It is added to an ADF Pipeline, and can be scheduled or triggered
- You can monitor the Mapping Data Flow both visually and programmatically

Microsoft

Data Loading Best Practices

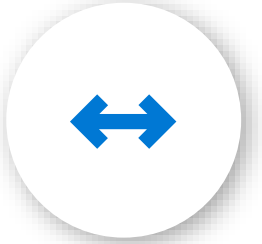
Speaker name

Title

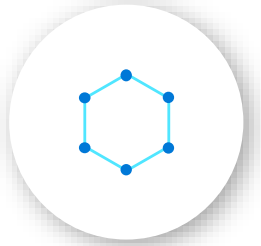
What is Azure SQL Data Warehouse?

AZURE SQL DATA WAREHOUSE

PaaS



Elastic Scale



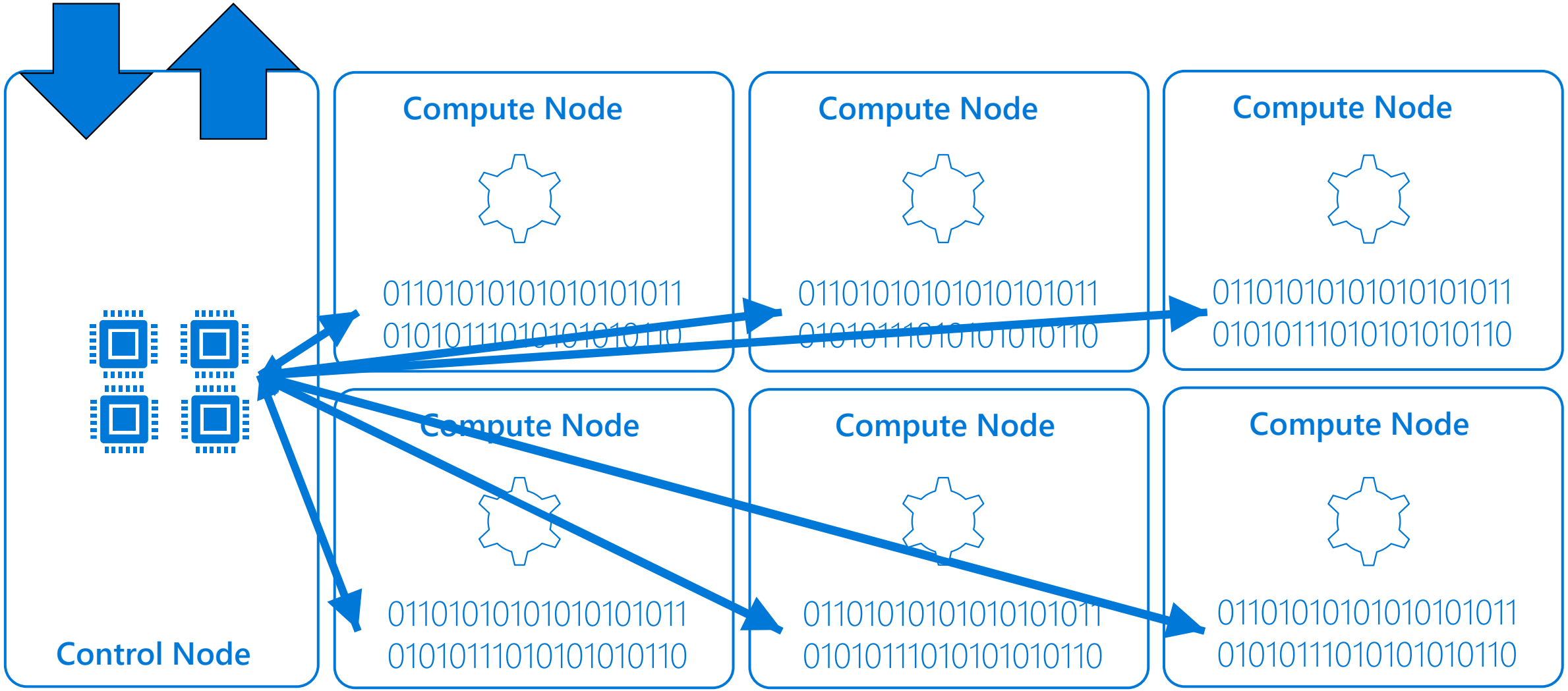
Big Data

Pause/Resume

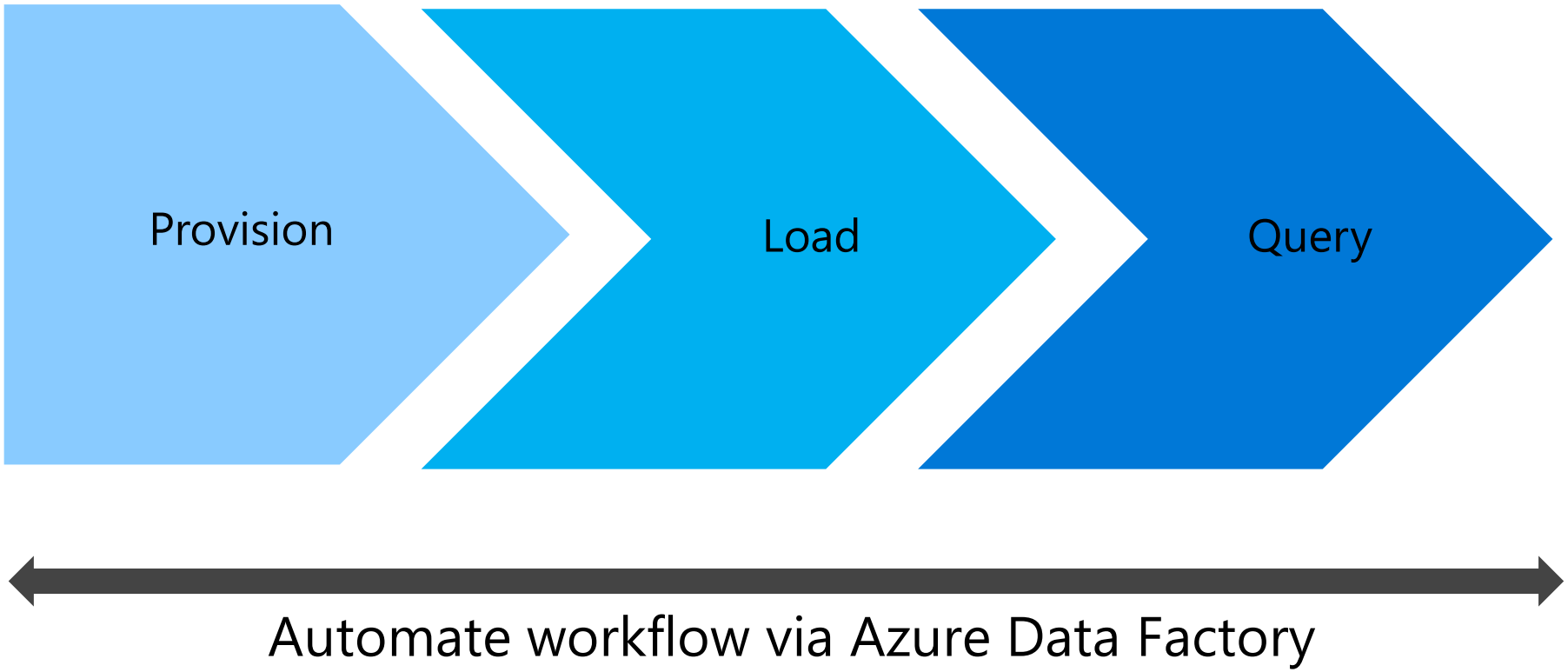
Separate Storage/Compute

Workload Management

SQL Data Warehouse Architecture



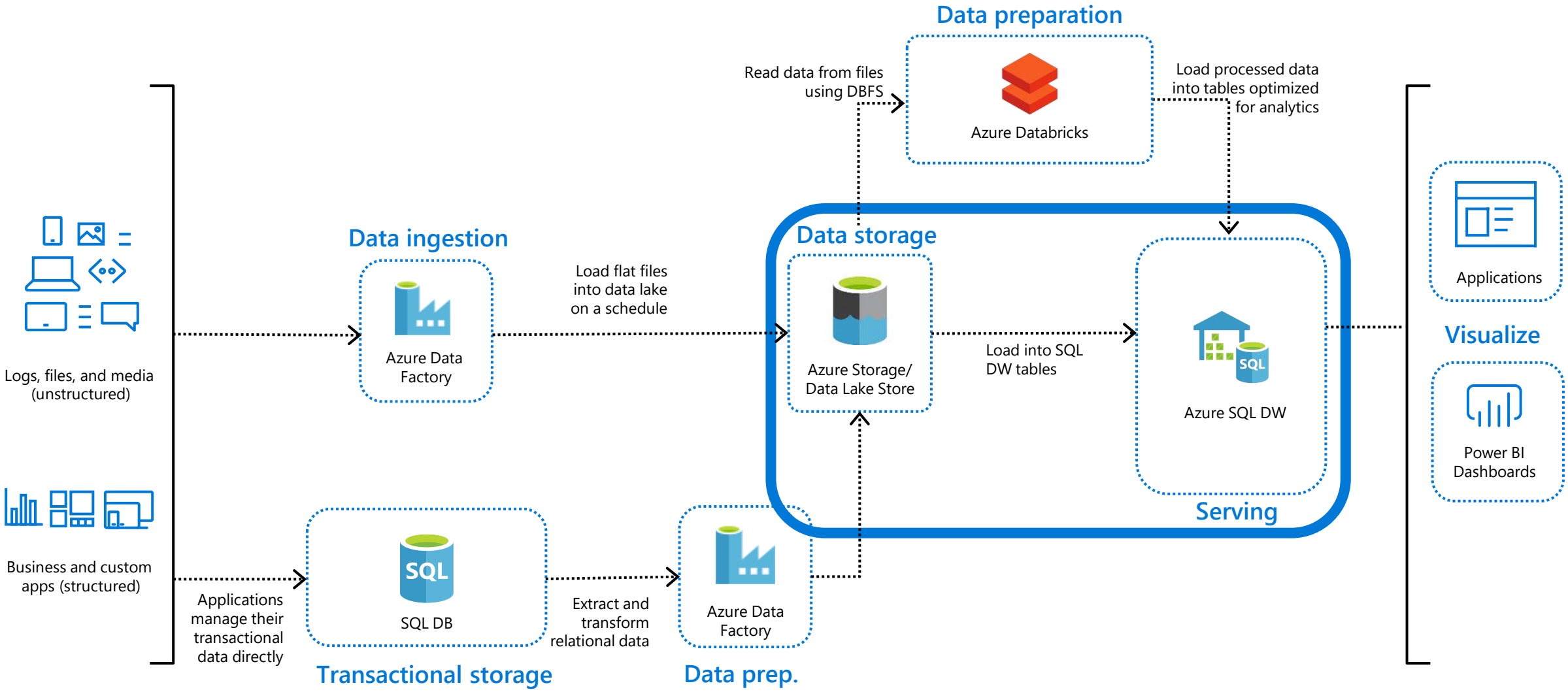
AZURE SQL DATA WAREHOUSE PROCESSES



Loading design goals

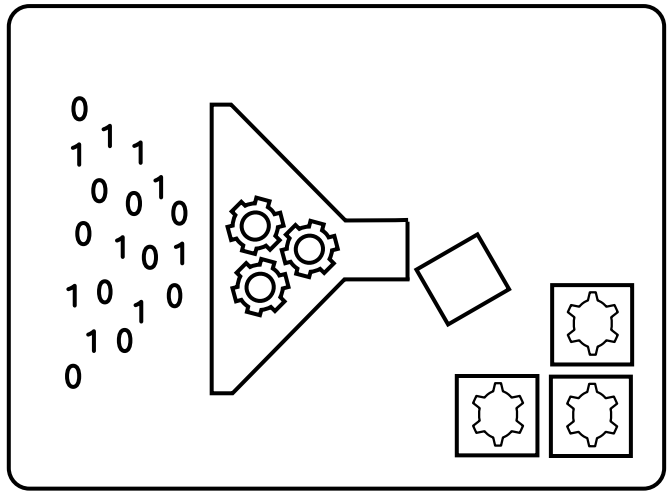
Data warehousing loading in Azure

Loading data into Azure SQL Data Warehouse



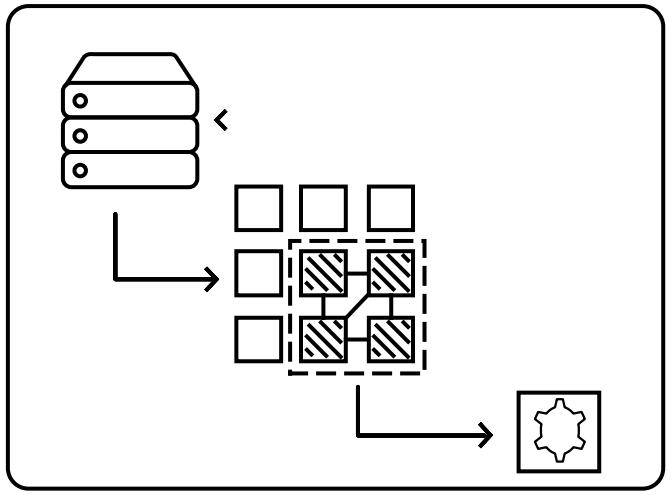
Loading Methods

BCP



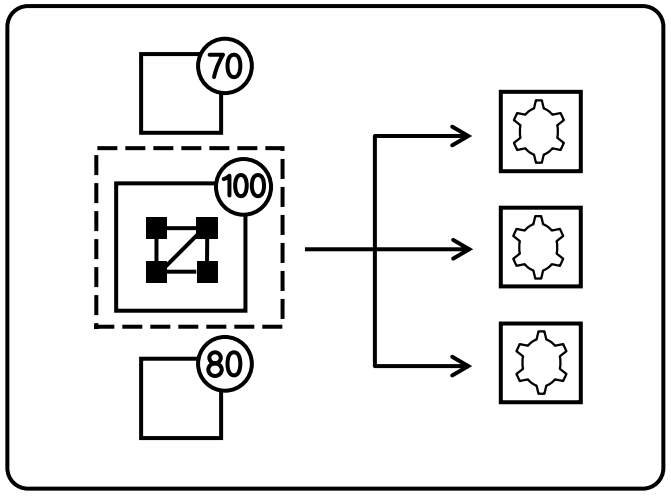
File based

SSIS



Heterogenous

PolyBase



File based

Loading Methods

For large amounts of data, there is only one choice

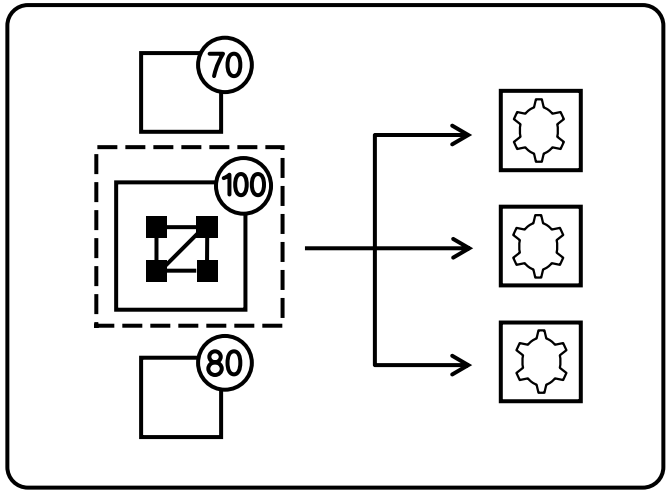
BCP

File based

SSIS

Heterogenous

PolyBase



File based

PolyBase benefits

The best practice for loading large amount of data

Leverages MPP architecture

PolyBase is designed to leverage the MPP (Massively Parallel Processing) architecture of SQL Data Warehouse and will therefore load and export data magnitudes faster than any other tool.

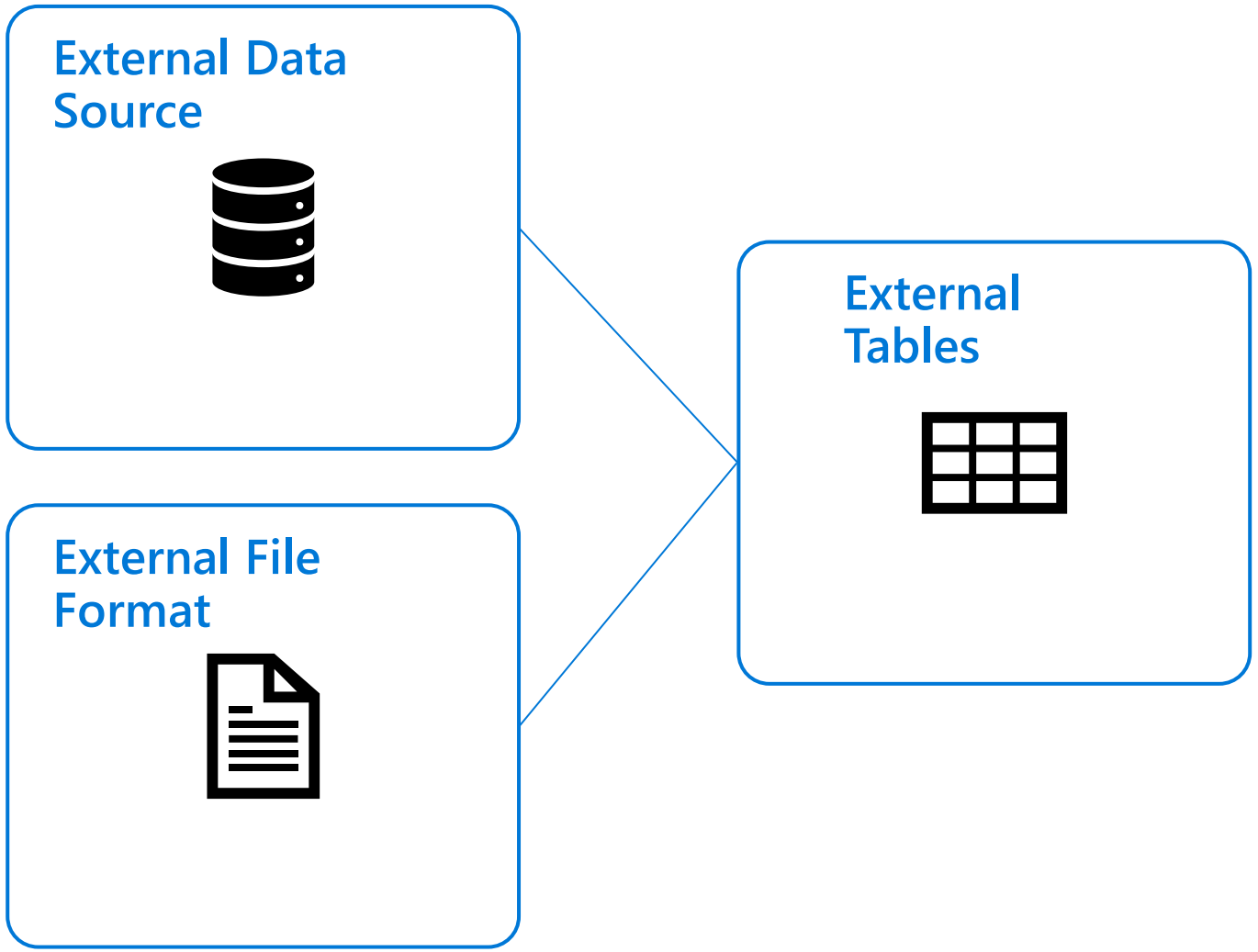
Azure Data Factory support

Azure Data Factory also supports PolyBase loads and can achieve similar performance to running PolyBase manually

Variety of file formats

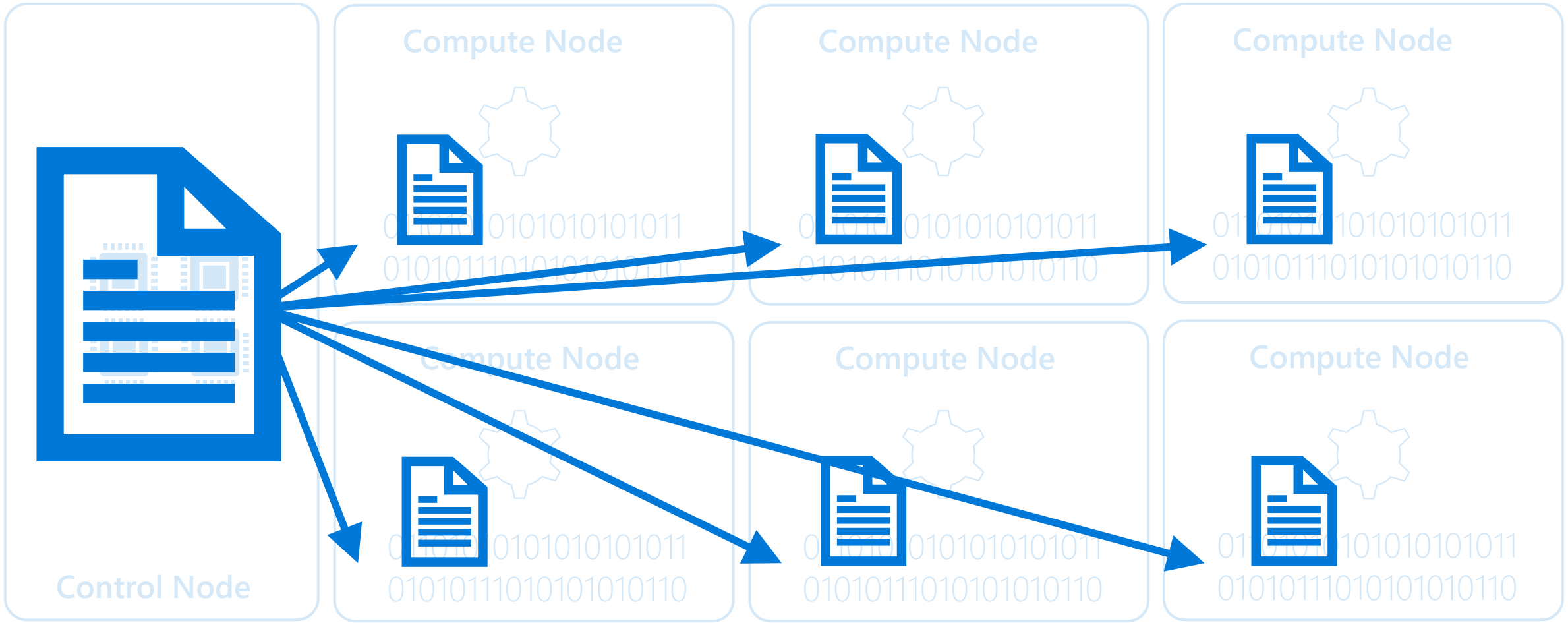
PolyBase supports a variety of file formats including RC, ORC and Gzip files.

Components of PolyBase

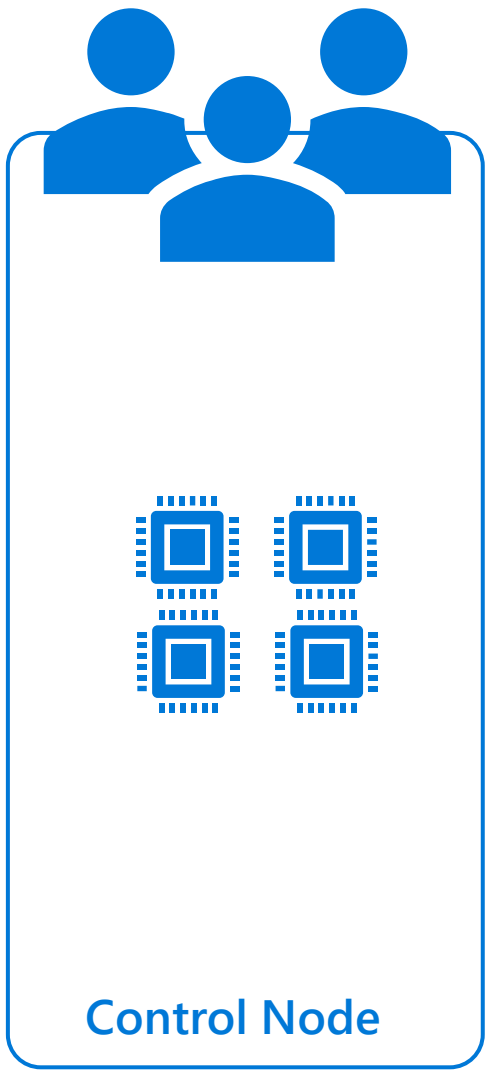


Loading best practices

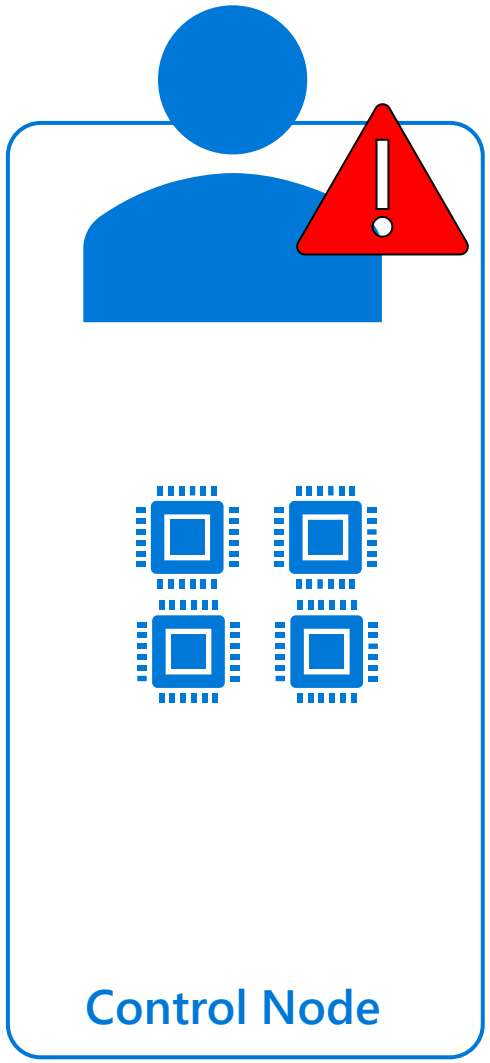
Manage your files



Reduce concurrent access



Create a dedicated load user account



Manage singleton updates

Control Node

Compute Node

01101010101010101011
01010111010101010110

Compute Node

01101010101010101011
01010111010101010110

Compute Node

01101010101010101011
01010111010101010110

Compute Node

01101010101010101011
01010111010101010110

Compute Node

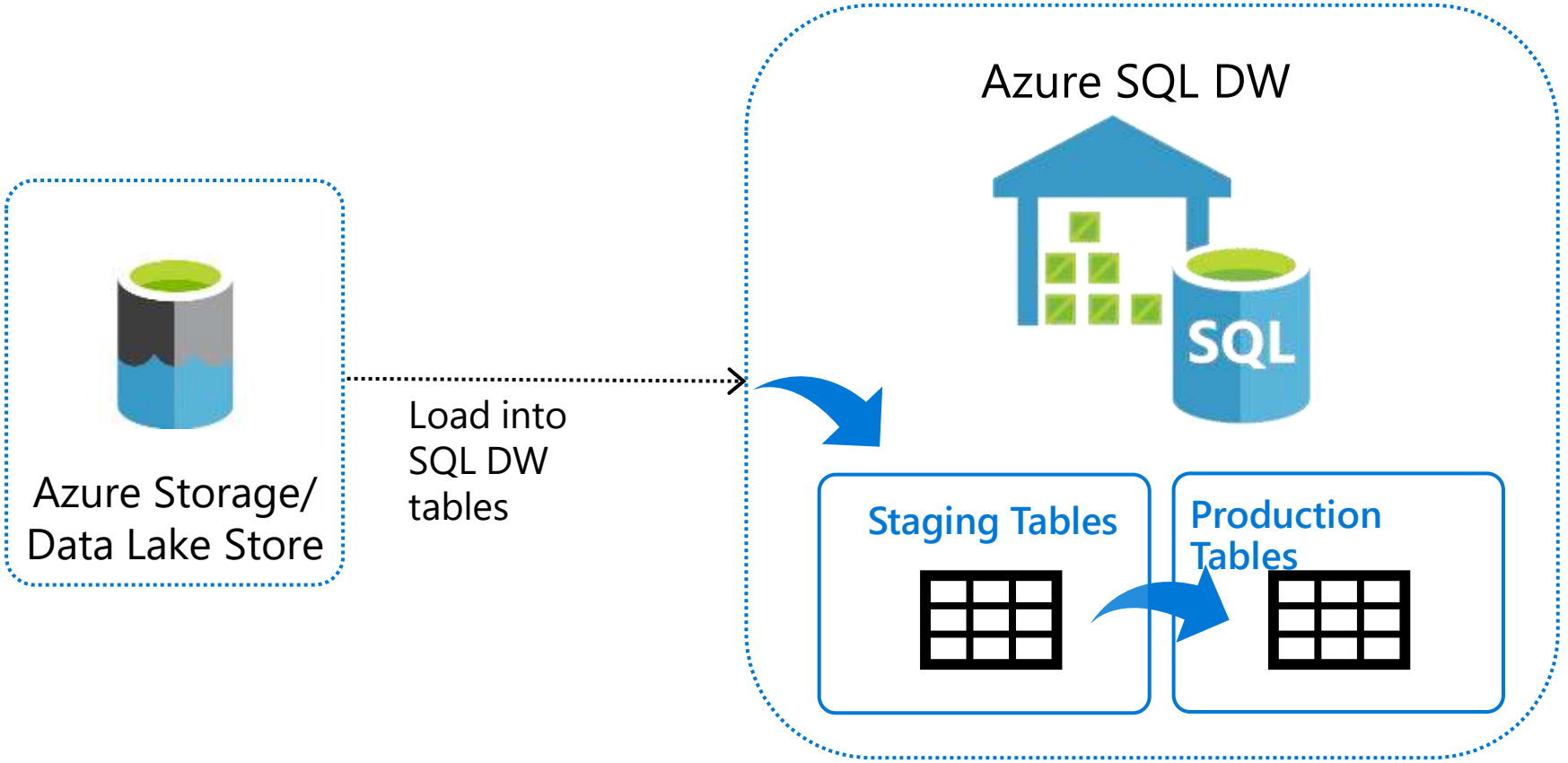
01101010101010101011
01010111010101010110

Compute Node

01101010101010101011
01010111010101010110

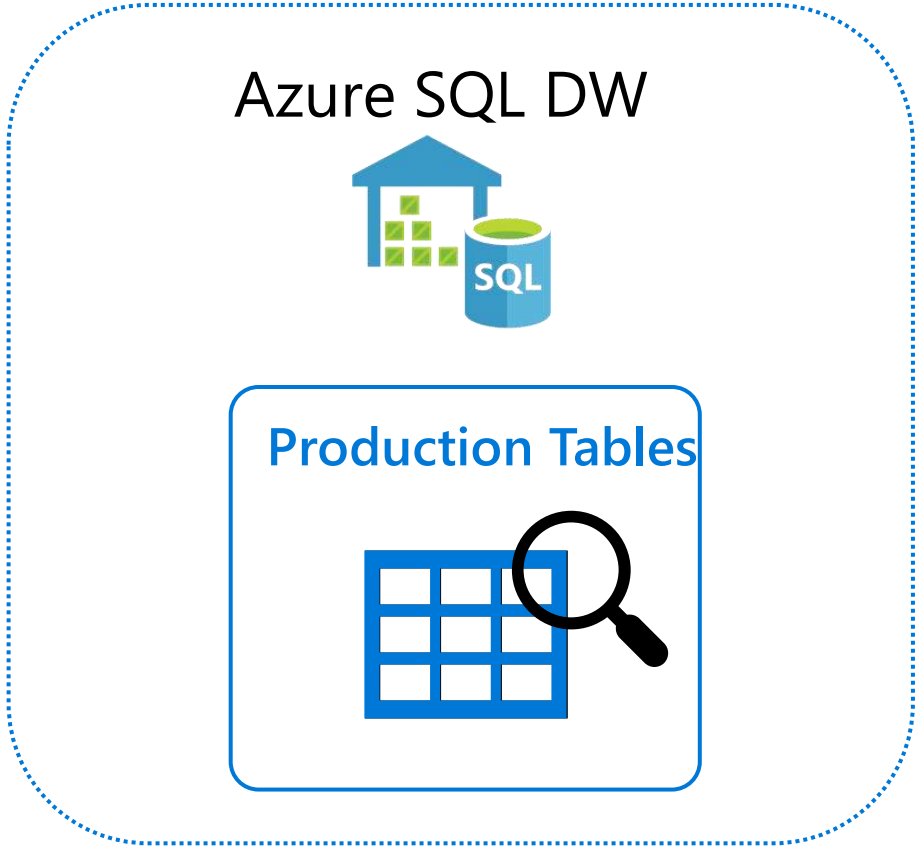
Optimize your loads

View it as a two-stage process



Create statistics after loading

Improve the query performance for users

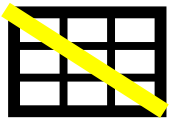


Maximizing Performance

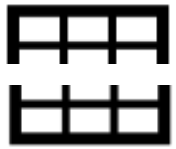
Maximizing Query Performance

Table distribution

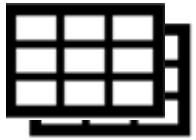
Round Robin Tables



Hash Distributed Tables

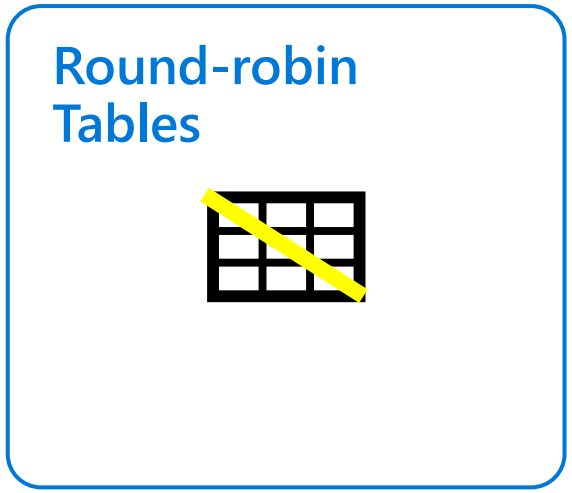


Replicated Tables



Maximizing Query Performance

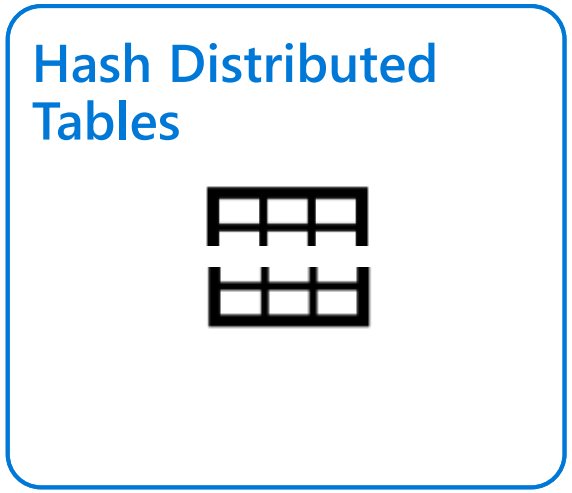
Round-robin distribution



- Is the default option for newly created tables
- Evenly distributes the data across the available compute nodes in a random manner, giving an even distribution of data across all nodes
- Loading into Round-robin tables is fast
- Queries on Round-robin tables may require more data movement as data is "reshuffled" to organize the data for the query
- Great to use for loading staging tables

Maximizing Query Performance

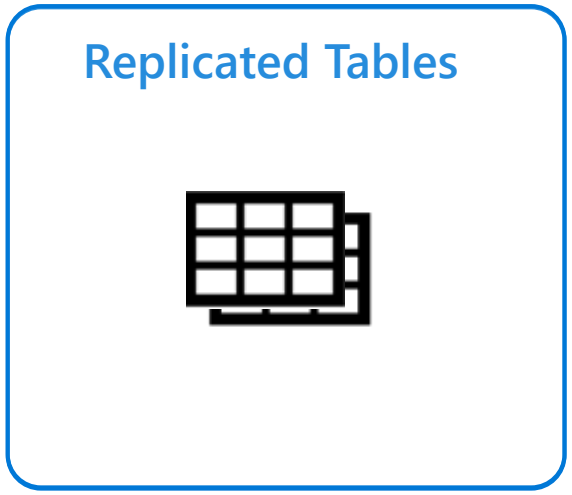
Hash distribution



- Distributes rows based on the value in the distribution column, using a deterministic hash function to assign each row to one distribution.
- Is designed to achieve high performance for queries that run against large fact tables in a star schema.
- Choosing a good distribution column is important to ensure the hash distribution performs well
- As a starting point, use on tables that are greater than 2GB in size and has frequent inserts, updates and deleted
- But don’t choose a volatile column for the hash distributed column

Maximizing Query Performance

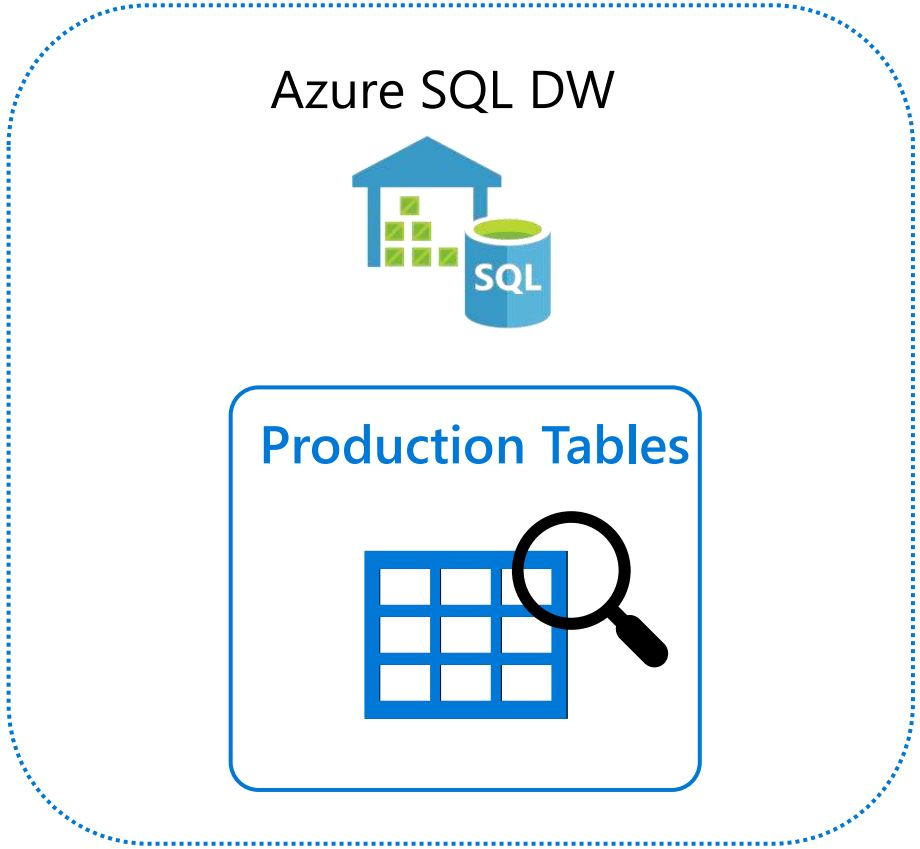
Replicated Table

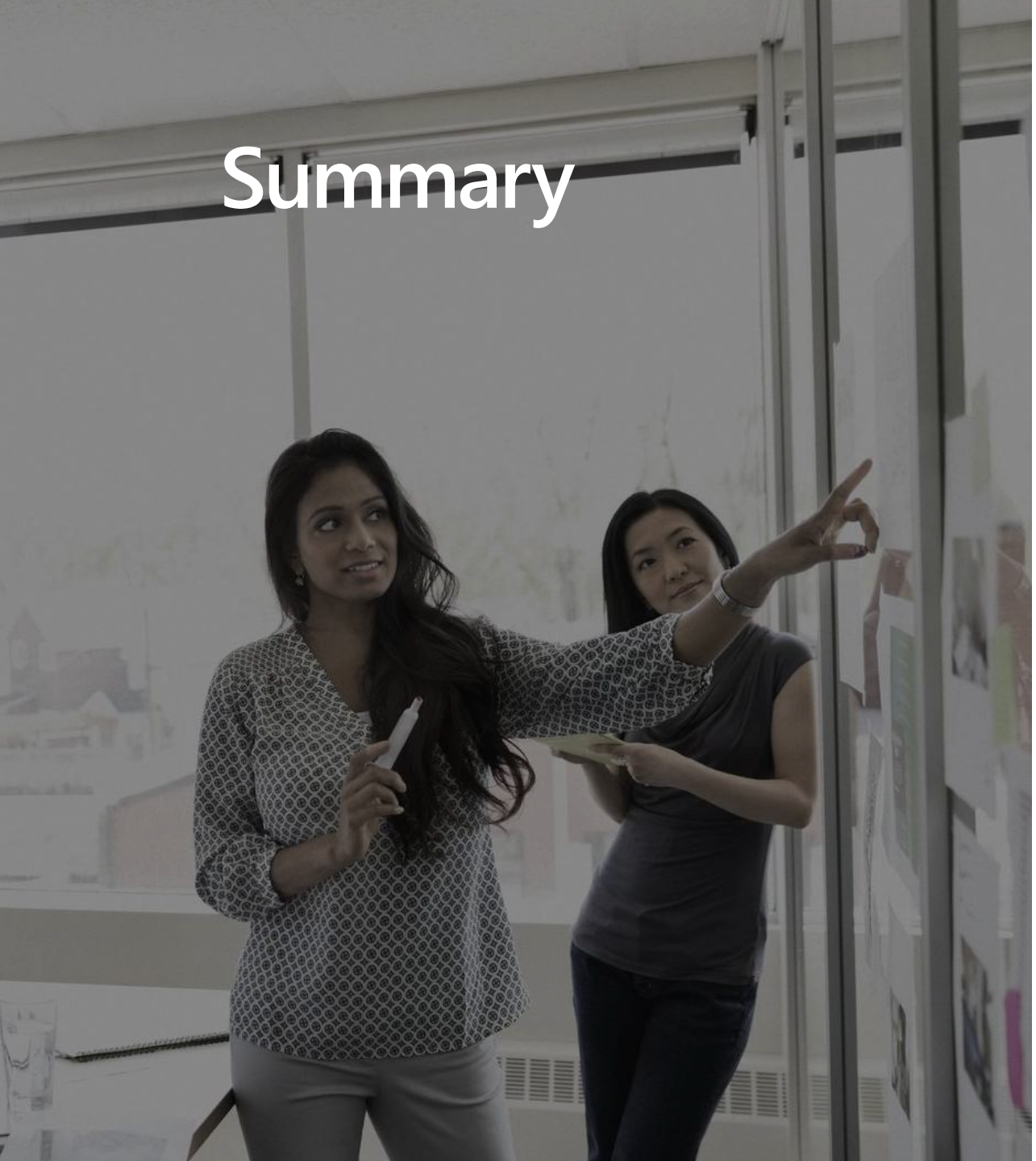


- A full copy of a table is placed on every single compute node to minimize data movement
- Works well for dimension tables in a star schema that are less than 2GB in size and are used regularly in queries with simple predicates
- Should not be used on dimension tables that are updated on a regular basis

Create statistics after loading

Improve the query performance for users





Summary

- Azure Data Factory (ADF) is a cloud-based data integration service that allows you to orchestrate and automate data movement and data transformation.
- Enable you to focus on building business logic and data transformation
- It is added to an ADF Pipeline, and can be scheduled or triggered
- You can monitor the Mapping Data Flow both visually and programmatically
- Load data efficiently
- Multiple methods of loading


Microsoft Azure