



Microsoft Azure Virtual Training Day: Delivering the modern data warehouse

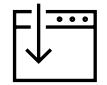


Delivering a modern data warehouse

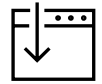
Nicholas Moore

Cloud Solutions Architect

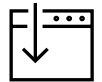
Agenda



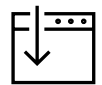
Why Modern Data Warehousing?



Building the Modern Datawarehouse



Advanced Analytics Patterns



The evolution of Cloud Scale Analytics

Why modern data warehousing?

Digital transformation

91% of business leaders see Digital Transformation as a way of sparking innovation and **finding efficiencies**

68% say Digital Transformation is **increasing profits**

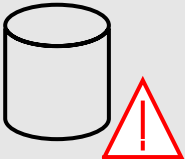
85% say they must offer digital services or **become irrelevant**

64% say they have less than 4 years to complete a Digital Transformation or they may **go out of business**

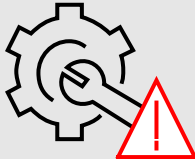


Common challenges with on-premises solutions

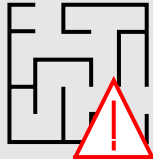
Data Silos



Performance Constraints



Solution Complexity



Escalating Costs

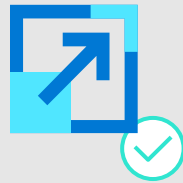


Derive real value from your data in the cloud

One hub for
all data



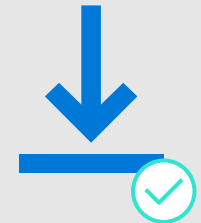
Unlimited
data scale



Common Platform



Lower TCO



Common customer use cases

Modern data warehouse



“Integrate all our data—including Big Data—with our data warehouse for analytics and reporting”

Advanced analytics



“Predict next best offer and customer churn”

Real-time analytics



“Derive insights from our devices and data streams in real-time”

Building the modern data warehouse

Modern data warehouse patterns

Modern data warehouse



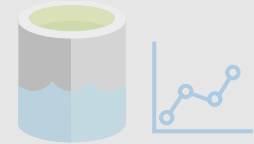
“Integrate all our data—including Big Data—with our data warehouse for analytics and reporting”

Advanced analytics



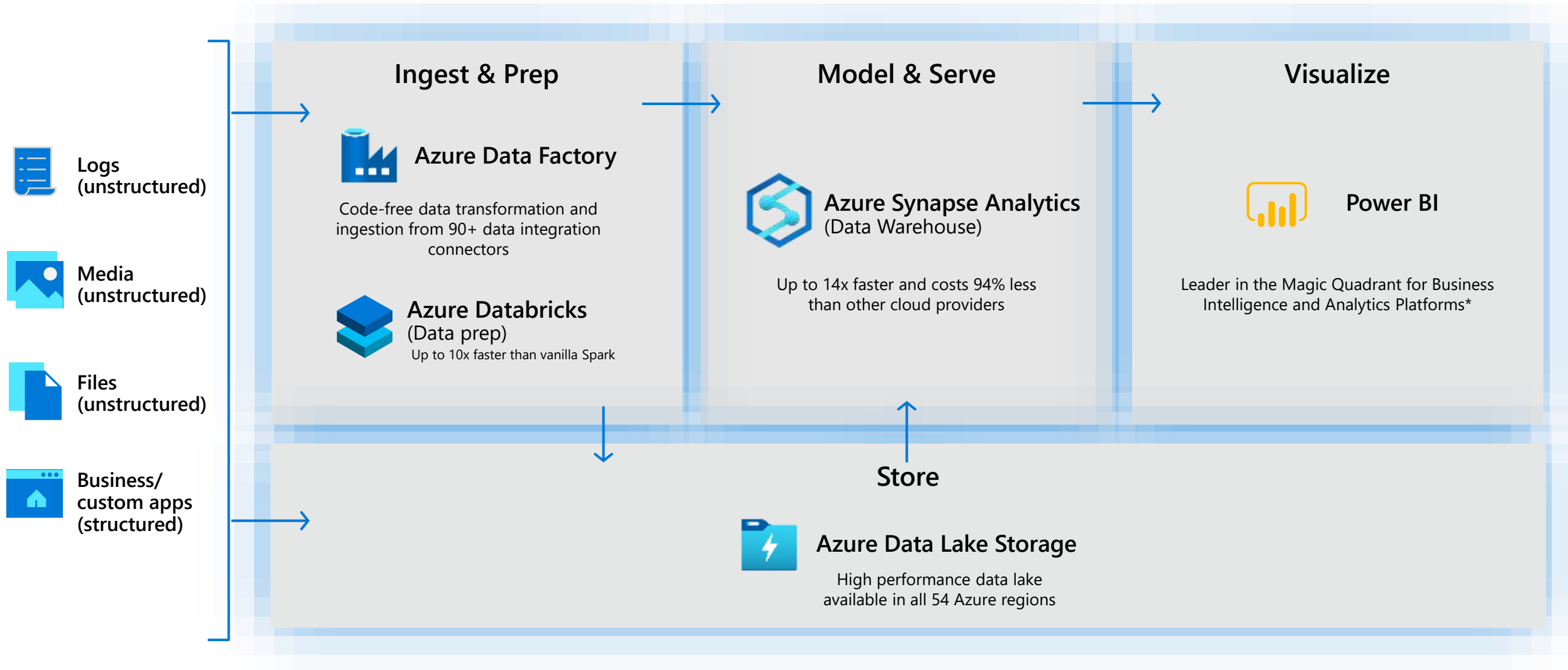
“Predict next best offer and customer churn”

Real-time analytics

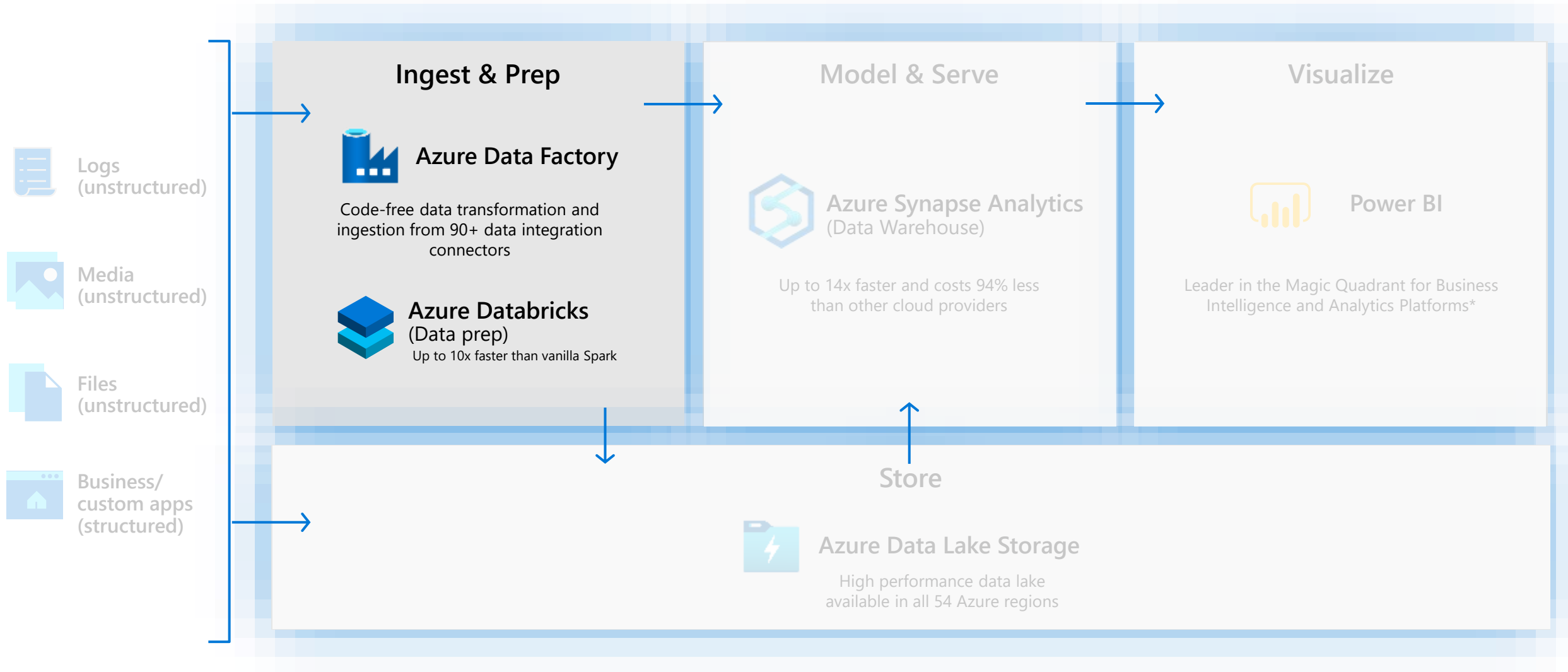


“Derive insights from our devices and data streams in real-time”

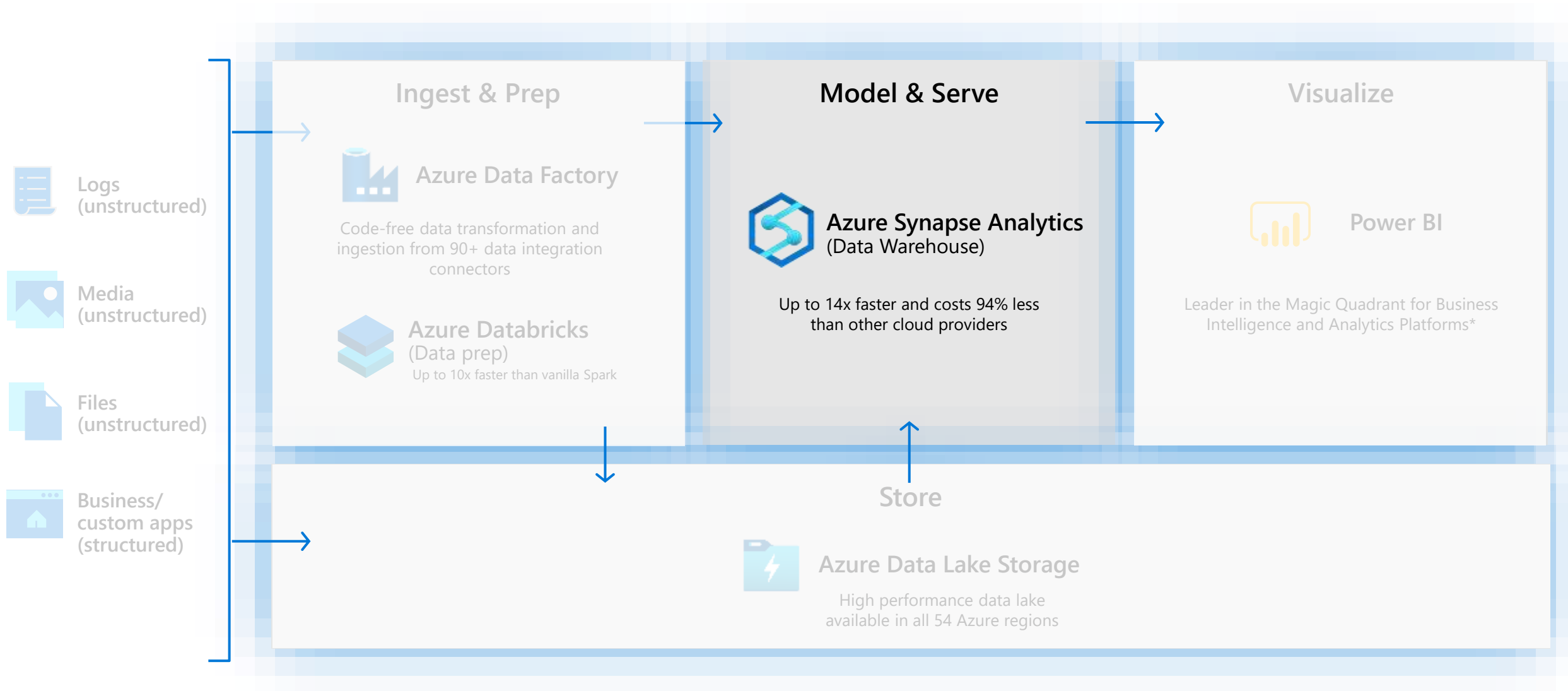
Modern data warehousing patterns



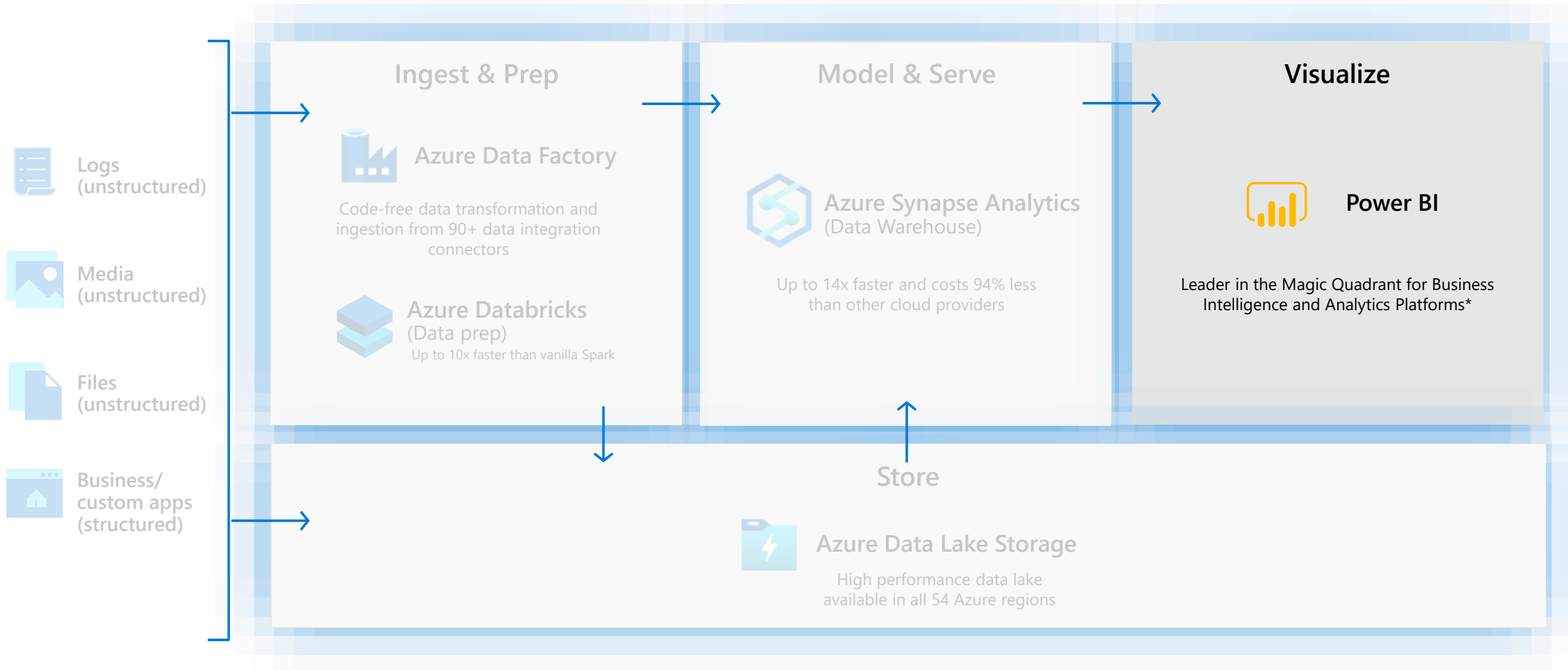
Ingest and Prep



Model and Serve



Visualize



Advanced Analytics patterns

The evolving world of Analytics

Descriptive



Diagnostic



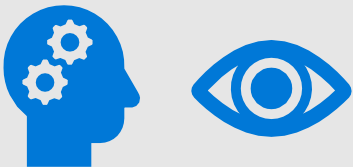
Predictive



Prescriptive



Cognitive



Advanced Analytics patterns

Modern data warehouse



“Integrate all our data—including Big Data—with our data warehouse for analytics and reporting”



Advanced analytics



“Predict next best offer and customer churn”

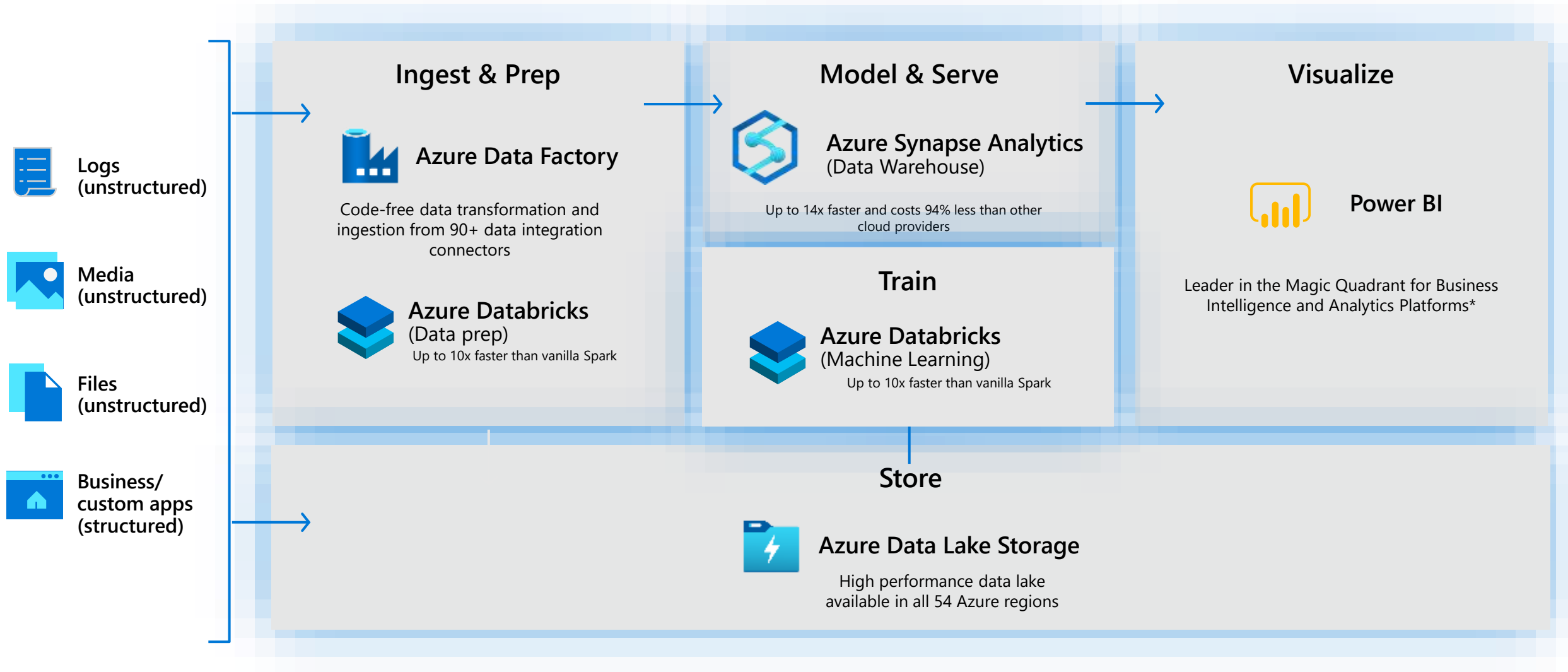


Real-time analytics



“Derive insights from our devices and data streams in real-time”

Advanced Analytics patterns



Real-time analytics patterns

Modern data warehouse



“Integrate all our data—including Big Data—with our data warehouse for analytics and reporting”



Advanced analytics



“Predict next best offer and customer churn”

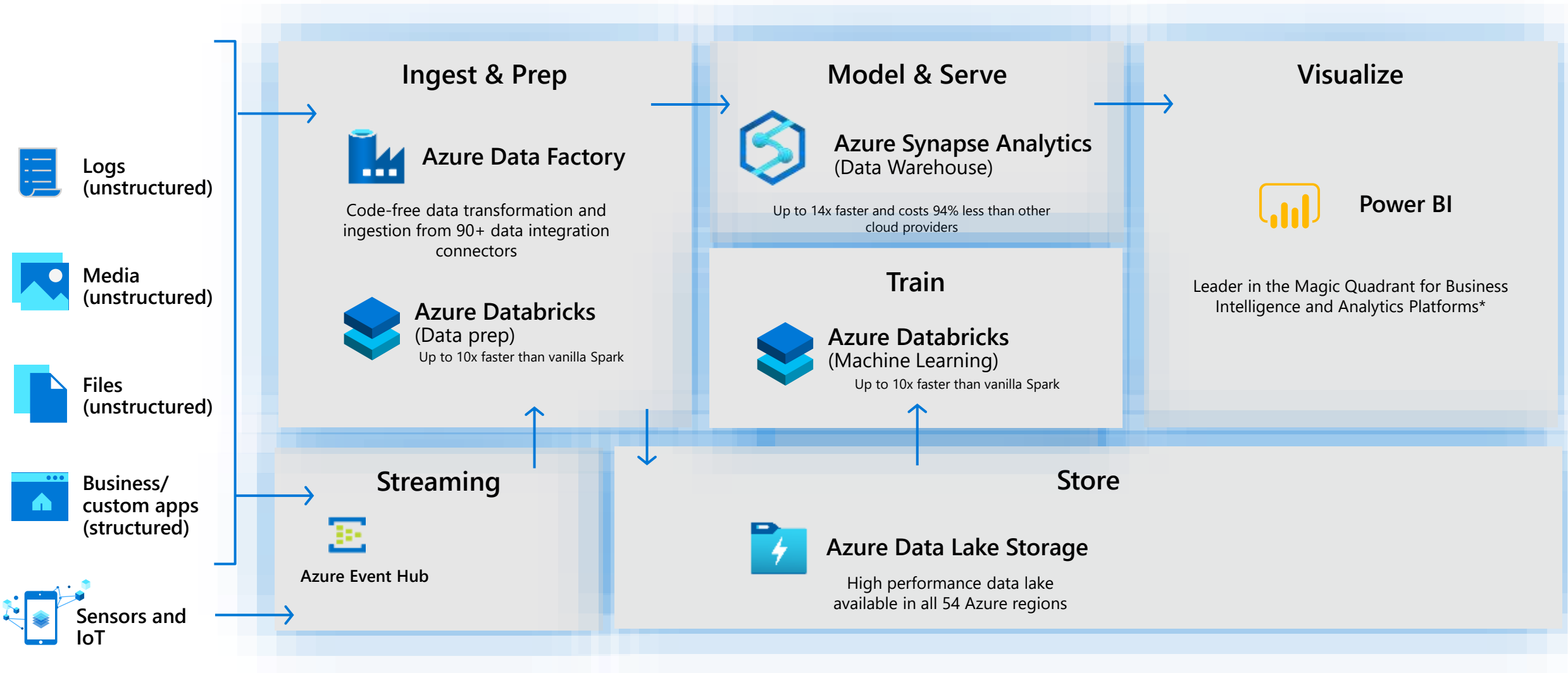


Real-time analytics



“Derive insights from our devices and data streams in real-time”

Real-time analytics patterns

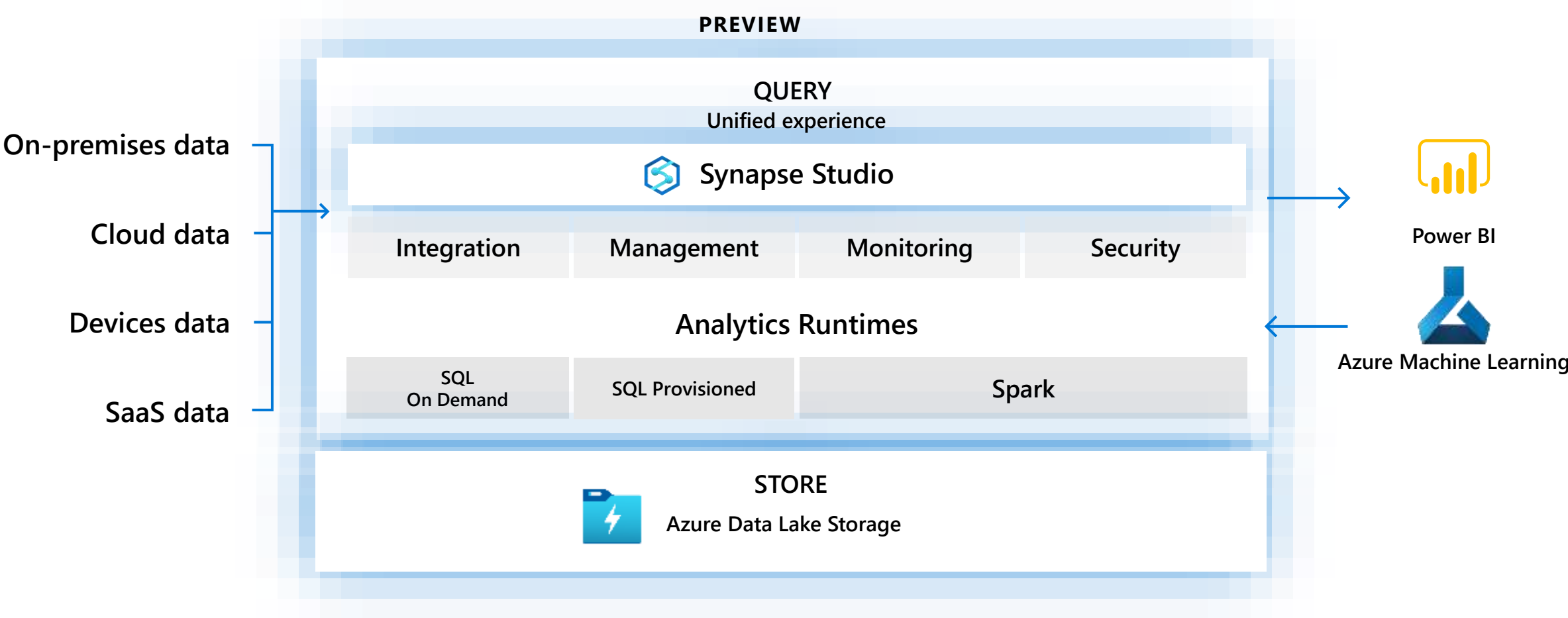


Demo:
**Modern Data Warehousing
and Cloud Scale Analytics**

The evolution of Cloud Scale Analytics

Azure Synapse Analytics

Limitless data warehouse with unmatched time to insights



Azure Synapse Analytics



**Limitless
scale**



**Powerful
insights**



**Unified
experience**



**Unmatched
security**



Ingesting data for analytics workloads

Nicholas Moore

Cloud Solutions Architect

Agenda

 What is Azure Data Factory?

 Ingesting data

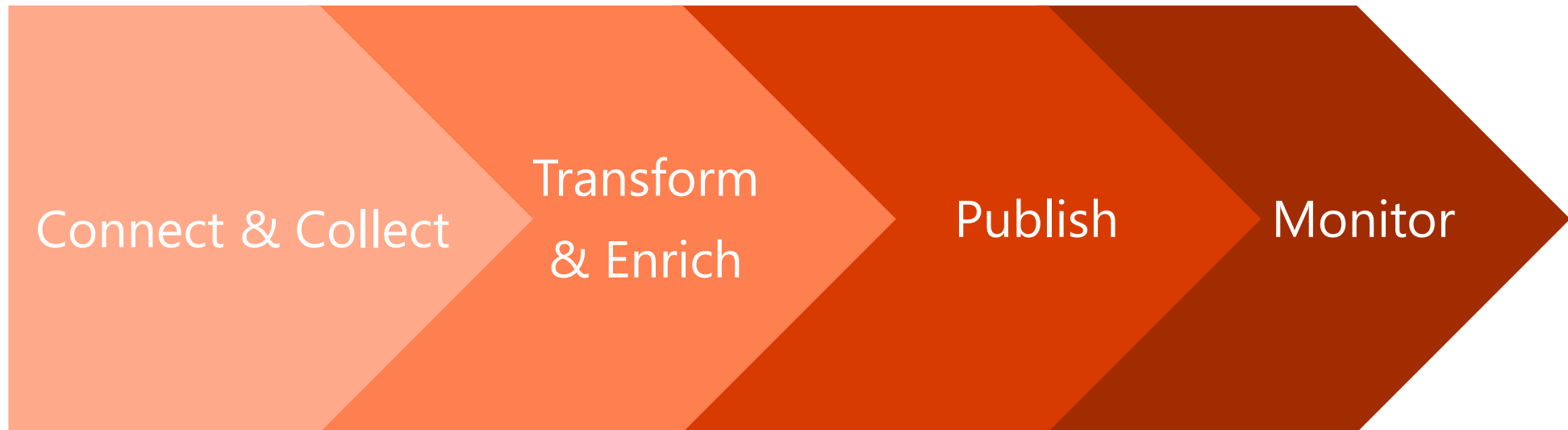
 Monitoring

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

Linked Service



Data Lake Store



Azure Databricks

Triggers



Pipeline

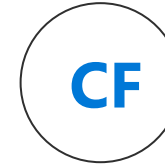
Activities



Parameters



Integration Runtime



Control Flow

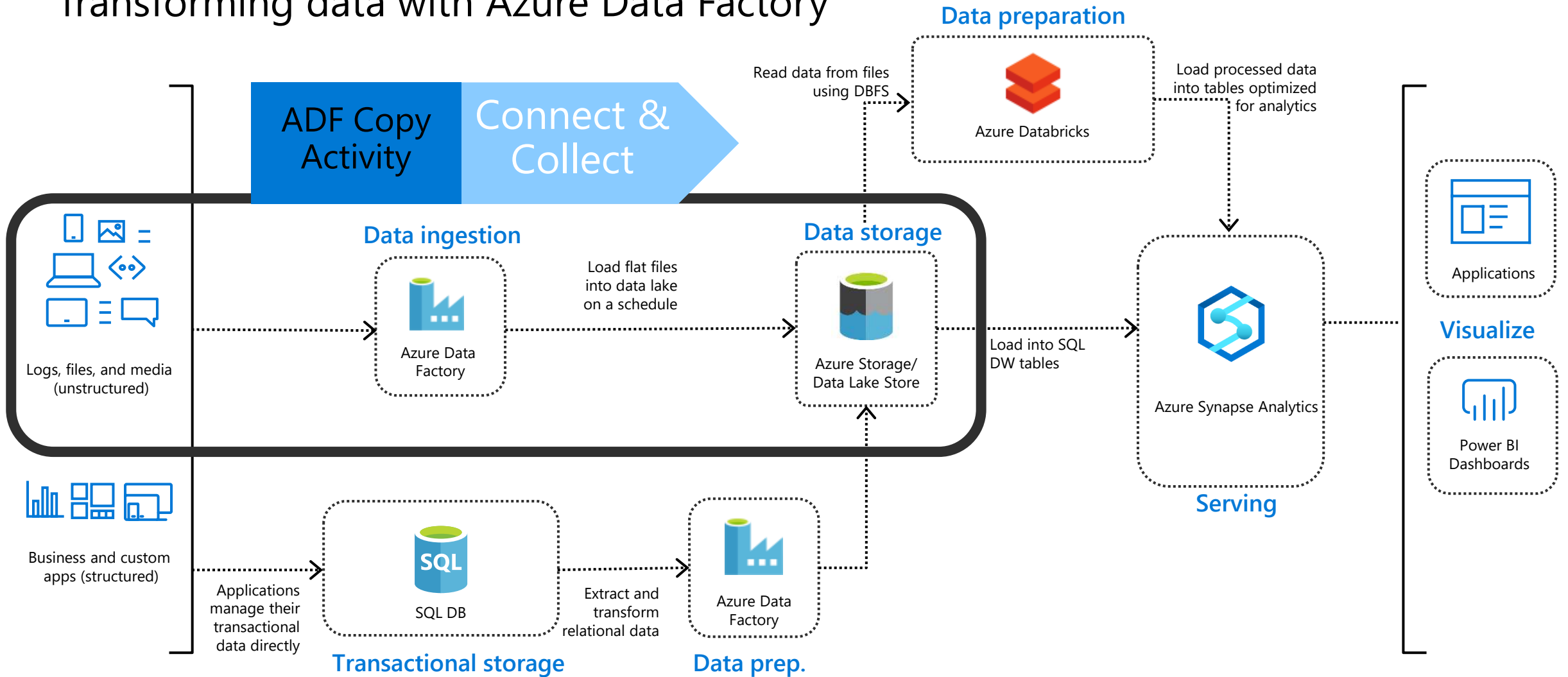


Dataset

Ingesting data

Data transformation in Azure

Transforming data with Azure Data Factory

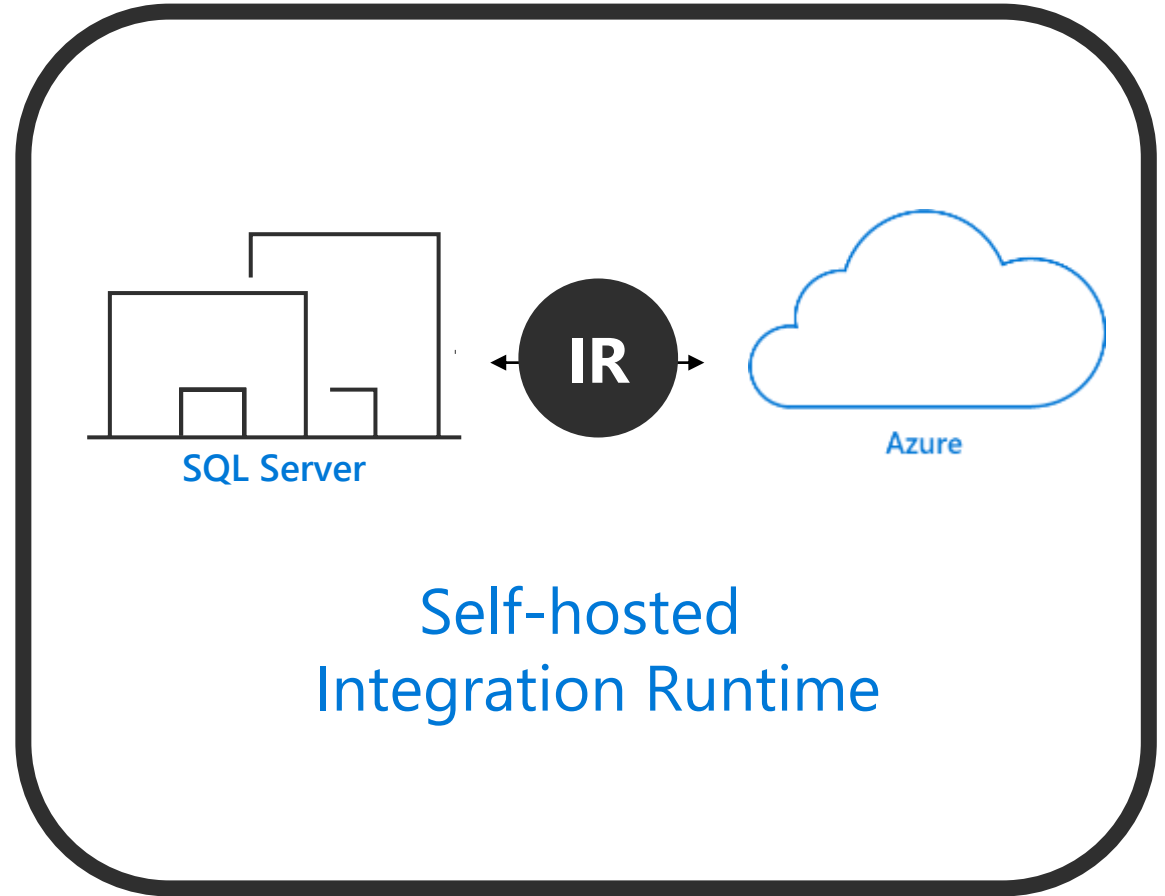
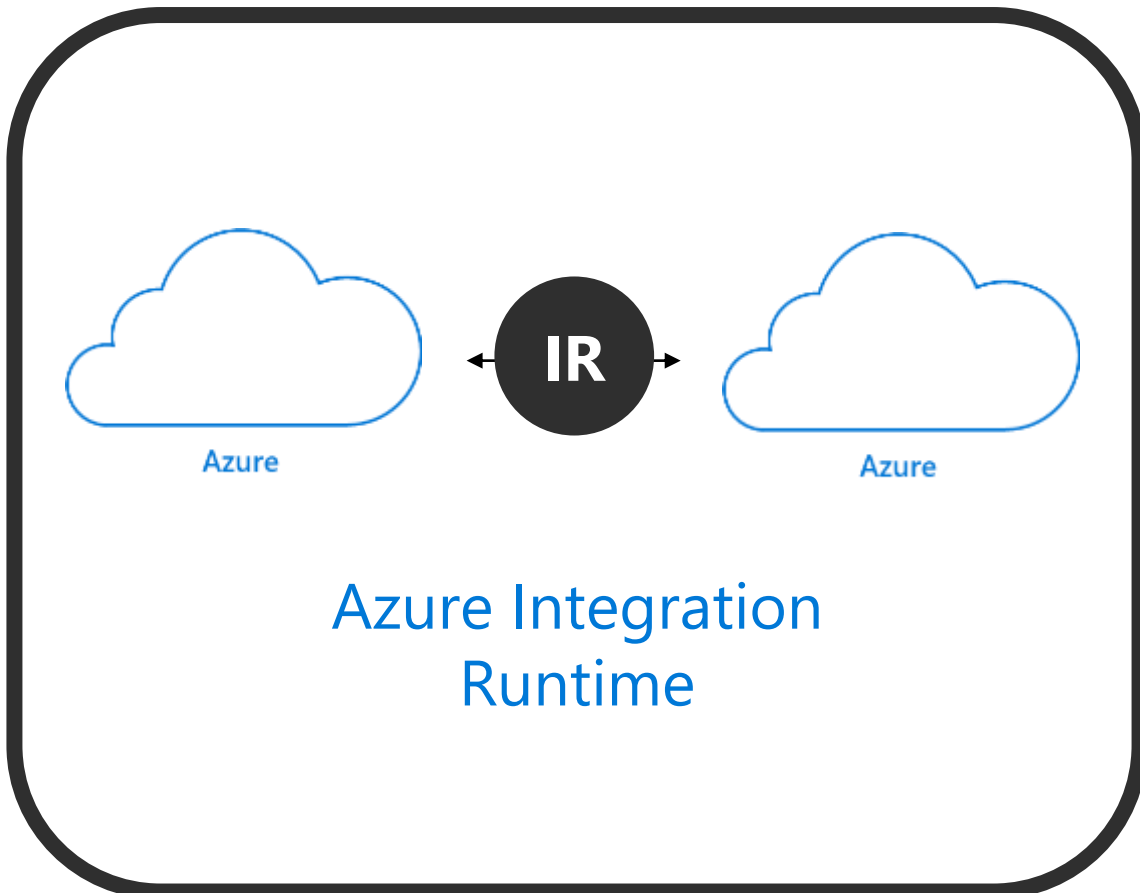


Copy Activity process



- 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



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



Transforming and enriching data

Nicholas Moore

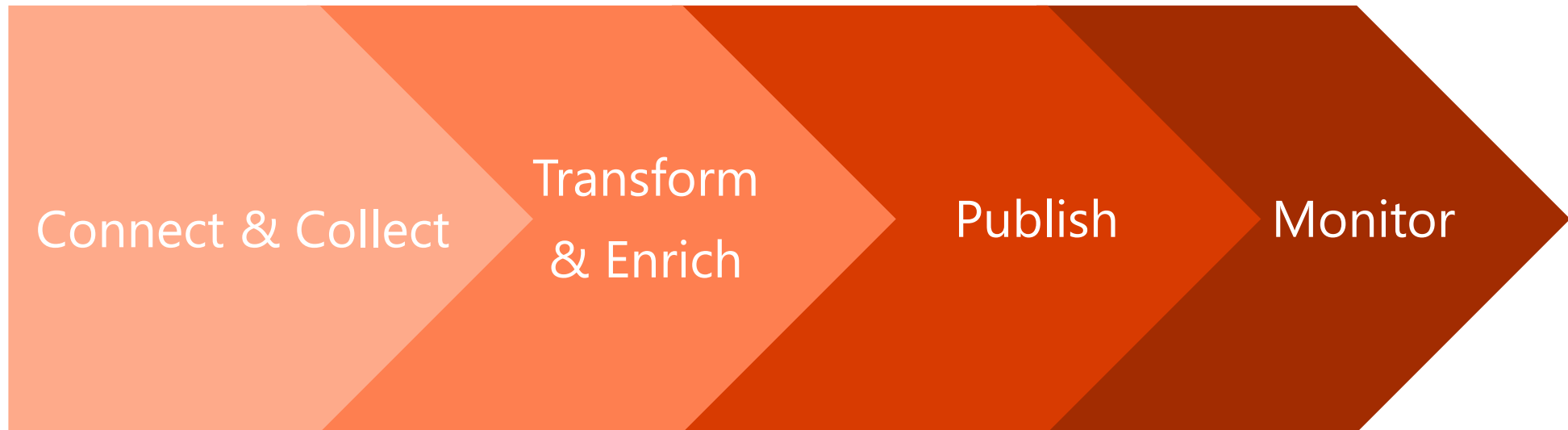
Cloud Solutions Architect

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

Linked Service



Data Lake Store



Azure Databricks

Triggers



Activities

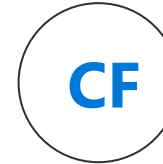
Pipeline



Parameters



Integration Runtime



Control Flow



Dataset

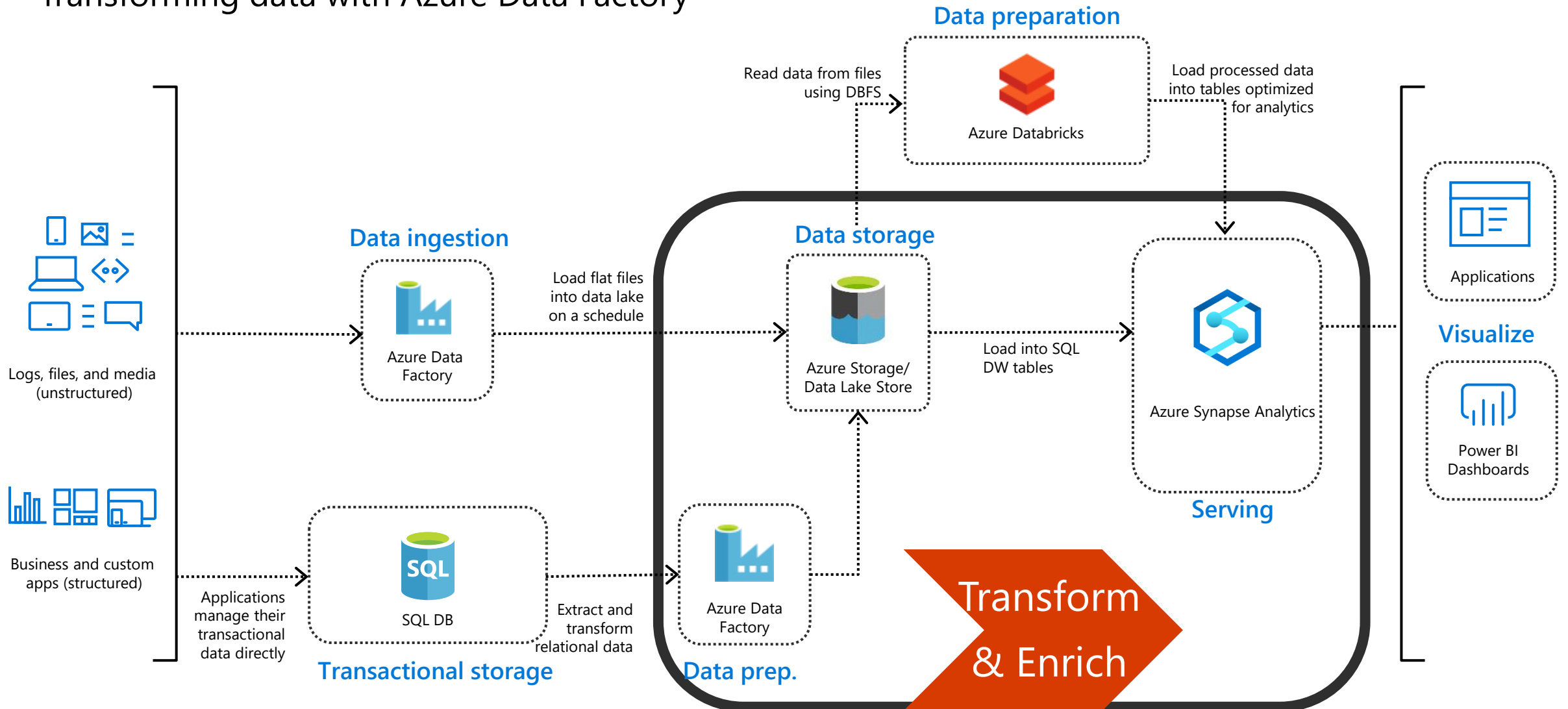
Component dependencies



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

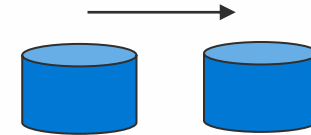
Compute
resources



SSIS Packages



Mapping Data
Flow



Methods for transforming data in Azure Data Factory

Code free data transformation at scale

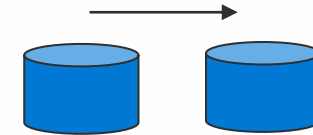
Compute
resources



SSIS Packages



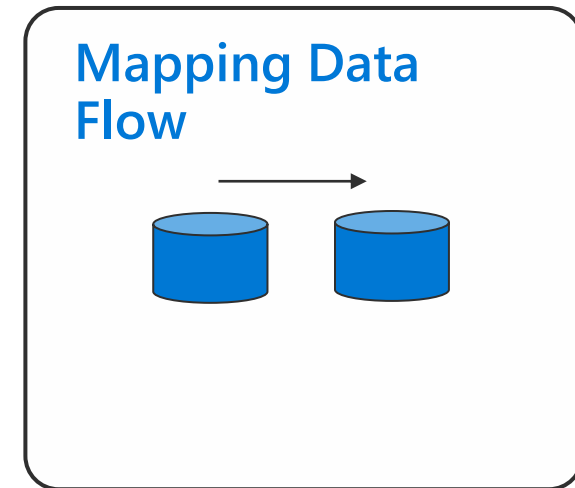
Mapping Data
Flow



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

The screenshot displays the Azure Data Factory Mapping Data Flow interface. The top bar includes navigation and action buttons: Save, Validate, Debug Settings, and Data Flow debug (which is toggled on). The left-hand navigation pane shows a hierarchy of resources under 'Factory Resources', including Pipelines (6), Datasets (14), Data Flows (Preview) (8), and Templates (0). The central graph area, labeled 'graph', shows a data flow with a source node 'source1' (Import data from MoviesDB) and an 'Add Source' placeholder. The bottom configuration panel, labeled 'configuration panel', has tabs for 'General' and 'Parameters'. The 'General' tab is active, showing a 'Name' field with the value 'dataflow3' and an empty 'Description' field.

Starting the Mapping Data Flow

Code free data transformation at scale

The screenshot displays the 'Adding Data Flow' dialog box in a data engineering application. The main workspace shows a data source named 'source1' with '0 total' columns. Below the source, there are tabs for 'Source Settings', 'Define schema', 'Optimize', 'Inspect', and 'Data Preview'. The 'Source Settings' tab is active, showing the following configuration:

- Output stream name: source1
- Source Dataset: USDOutput
- Options: Allow schema drift
- Sampling: Enable Disable

At the bottom of the dialog, there are 'Cancel' and 'Finish' buttons. A 'Code' button is also visible in the top right of the dialog area.

Transformation options in the Mapping Data Flow

Unpivot Union Join
Lookup Window
Derived Column
Sink Alter Row New Branch
aggregate Pivot Filter
Conditional Split Sort
Exists Select
Surrogate Key Source



Triggering and monitoring

Triggering the Mapping Data Flow

Code free data transformation at scale

← New Trigger



Navigation: Dashboards | Pipeline Runs | Trigger Runs | Integration Runtimes | Alerts & Metrics

Actions: Run | Cancel | Refresh

Filters: Last 24 Hours: 01/29/2019 12:11 PM - 01/30/2019 12:11 PM | Time Zone: (UTC-08:00) Pacific Time (US & Ca...) | View All Rerun History | Filter

Filter: All | Succeeded | In Progress | Failed | Cancelled

Pipeline Name	Actions	Run Start	Duration	Triggered By	Status	Parameters	Annotations	Error	RunID
pipeline7		01/29/2019, 4:22:47 PM	00:00:39	Manual trigger	Failed				25e91785-9bed-42fd-beee-92d725fac
pipeline7		01/29/2019, 1:47:54 PM	00:02:18	Manual trigger	Succeeded				7d8f3f63-7bee-4e0c-8c07-35760c56d

Demo:
**Transforming your data
in Azure Data Factory**

A photograph of two women in a meeting room. One woman is pointing at a whiteboard while the other looks on. The image is dimmed to serve as a background for the text.

In Summary:

Transforming Data with Azure Data Factory

- 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

Demo:
Ingesting data with
Azure Data Factory

Monitoring data ingestion

Monitoring Activity runs

LoadADLSG1Demo | Monitor Pipeline Runs ▾





Pipelines / CopyFromAmazonS3ToADLSG1

Refresh

Activity Runs

Pipeline Run ID **d614f808-7b9d-4362-bb8b-a0bddf226d34**

All Succeeded In Progress Failed Cancelled

Activity Name	Activity Type	Actions	Run Start	Duration	Status	Integration Runtime
Copy-copyfroms3	Copy	  	01/17/2018, 11:12:45 PM	00:04:00	 Succeeded	DefaultIntegrationRuntime (East US 2)

A photograph of two women in a meeting room. One woman is pointing at a whiteboard, and the other is looking at it. The image is dimmed to serve as a background for the text.

In Summary:

Ingesting Data with Azure Data Factory

- 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



Data Loading Best Practices

Luis Silva

Cloud Solution Architect, Data Platform

Agenda

 What is Azure Synapse Analytics

 Using Polybase to Load Data in a data warehouse

 Data Loading best practices

What is Azure Synapse Analytics?



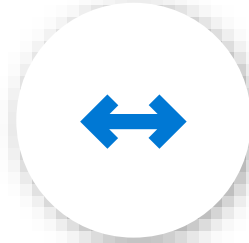
Azure Synapse Analytics

A **limitless** analytics service with **unmatched time to insight**, that delivers insights from all your data, **across data warehouses and big data** analytics systems, **with blazing speed**

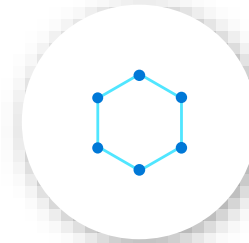
Azure Synapse Analytics



PaaS



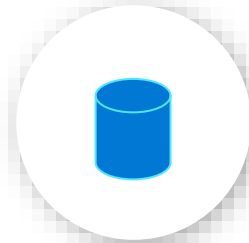
Elastic Scale



Big Data



Pause/Resume

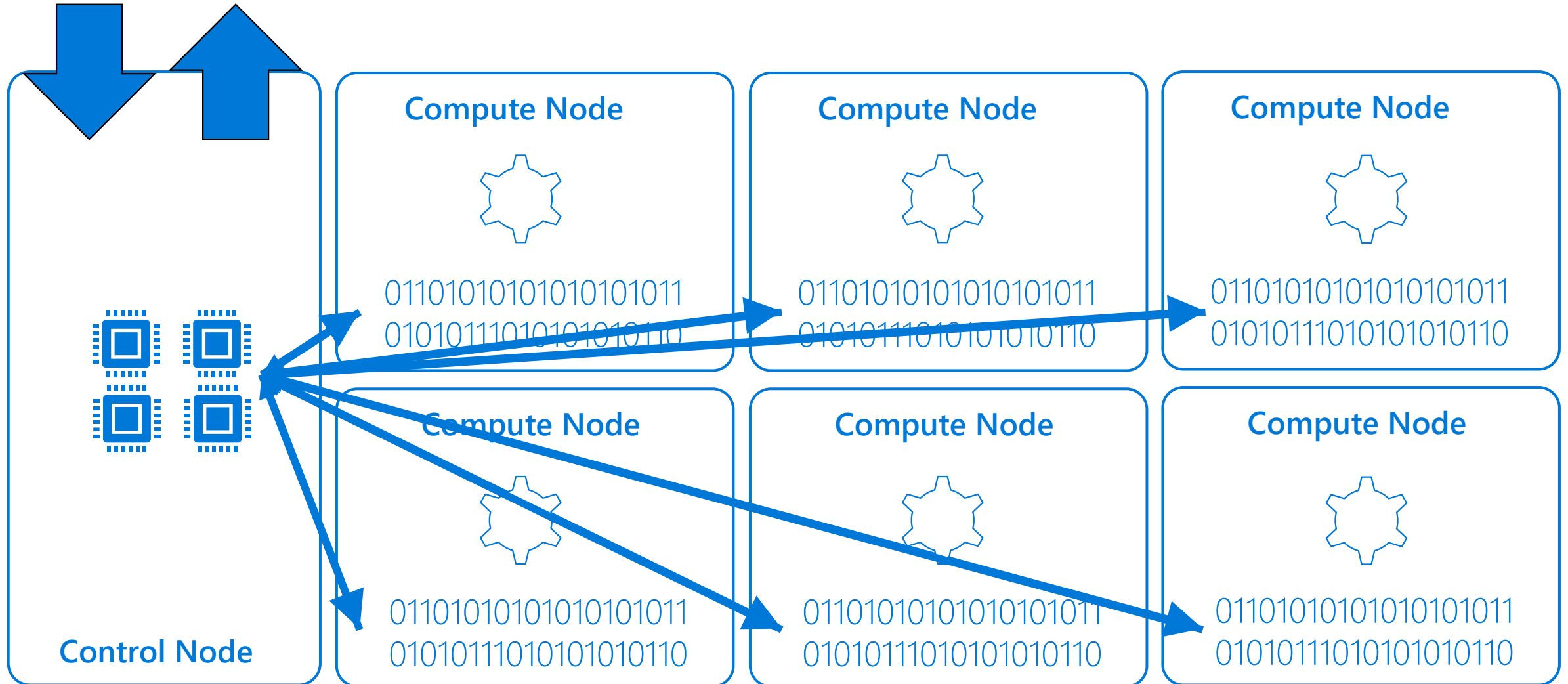


Separate
Storage/Compute



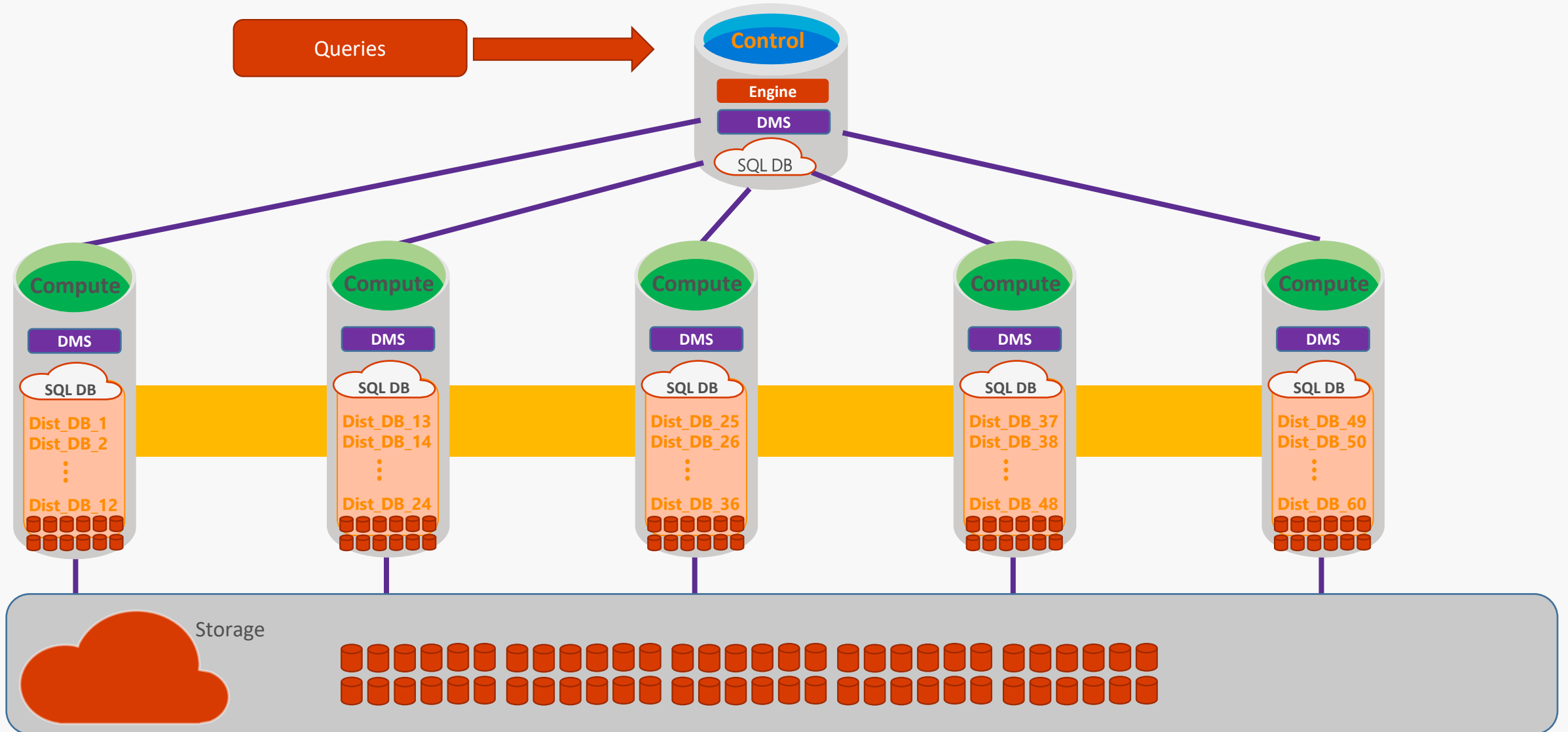
Workload
Management

Data Warehouse Architecture



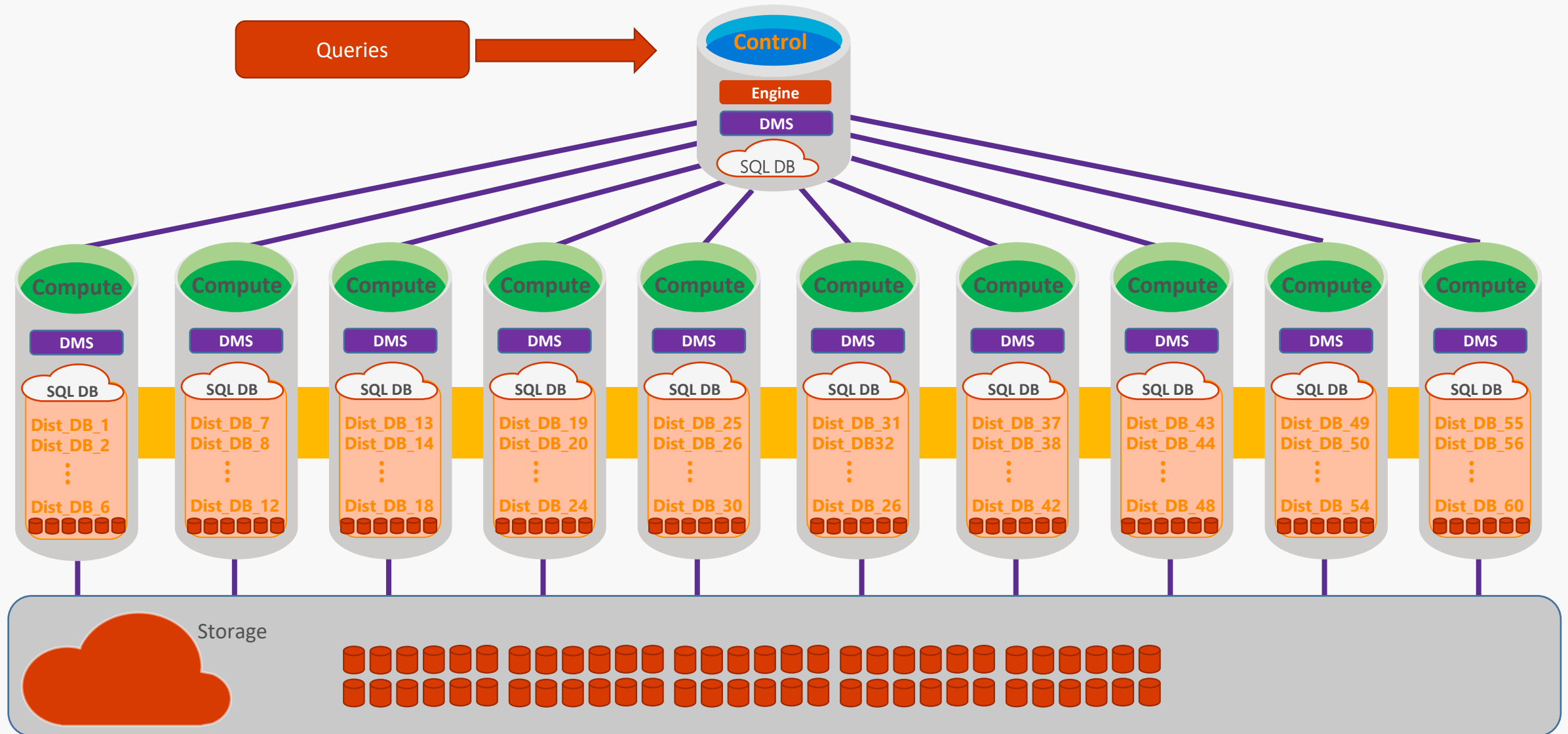
SQL Pool Scaling

DW2500c (5 compute nodes)

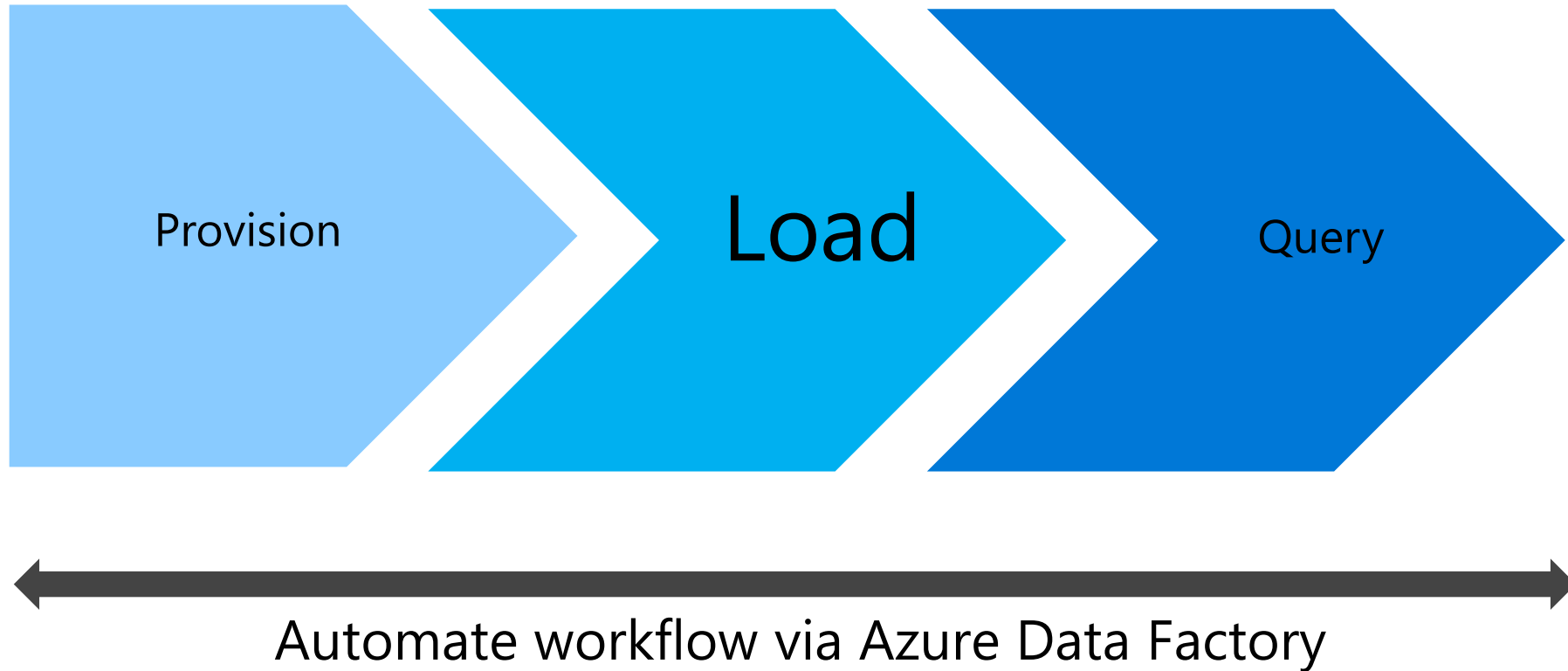


SQL Pool Scaling

DW5000c (10 compute nodes)



Data Warehouse Processes



Loading design goals

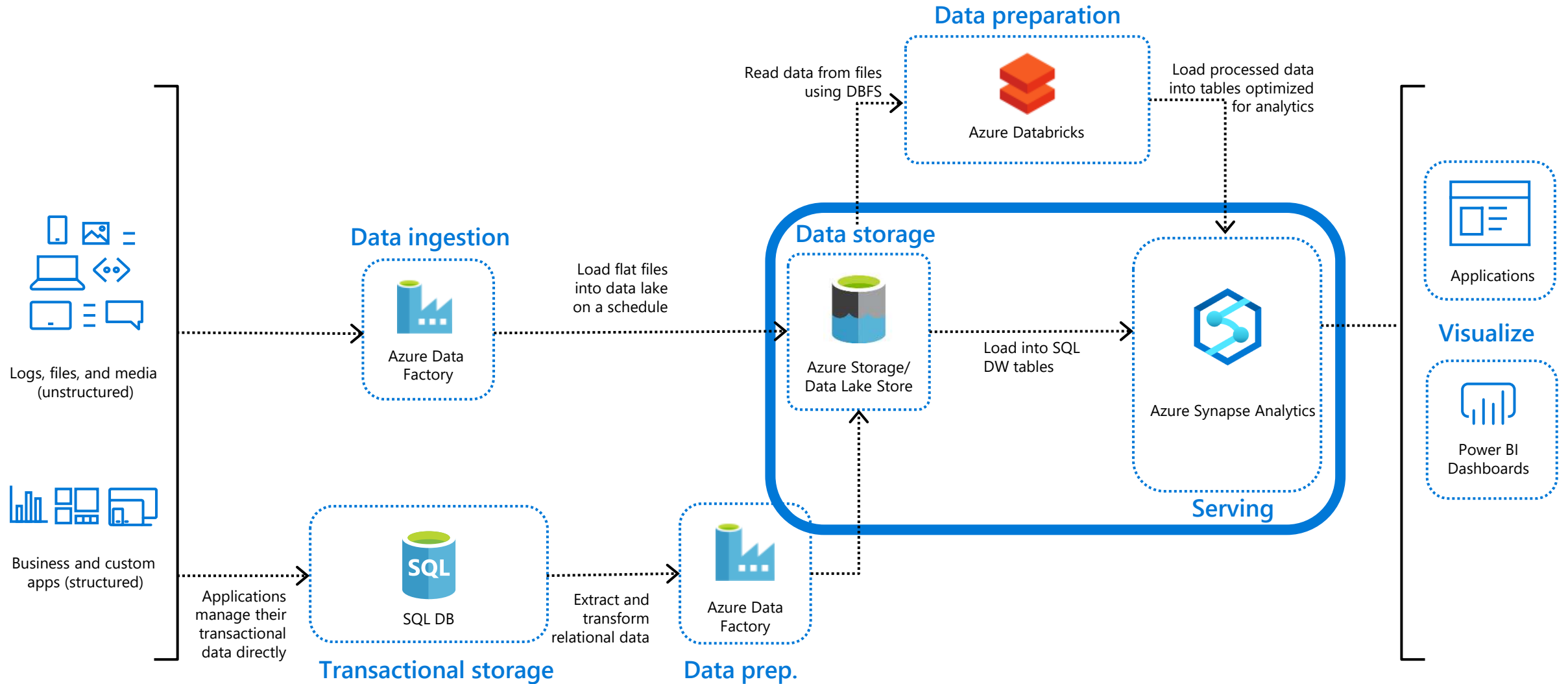
A technician wearing a blue cap and jacket is working on a server rack in a data center. The technician is pointing at a component on the rack. The background shows rows of server racks with cables.

Loading design goals

- Load data efficiently
- Load Data non-obtrusively, respecting concurrent queries and loads
- Reduce table fragmentation as much as possible
- Provide system recovery capabilities in the event of data load failure with minimal impact on concurrent queries
- Multiple methods of loading

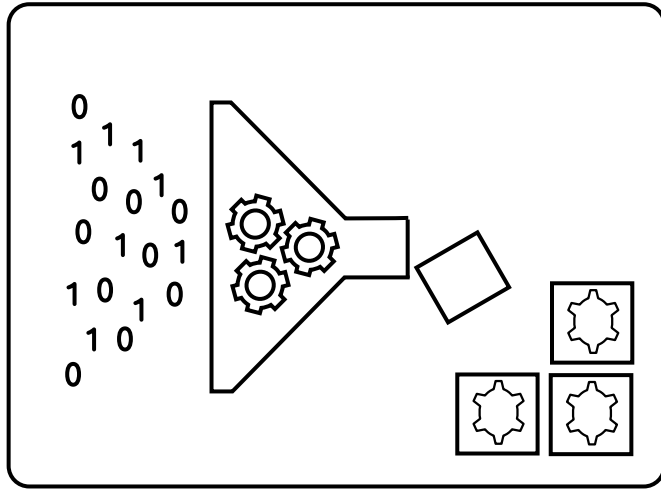
Data warehousing loading in Azure

Loading data into a data warehouse in Azure Synapse Analytics



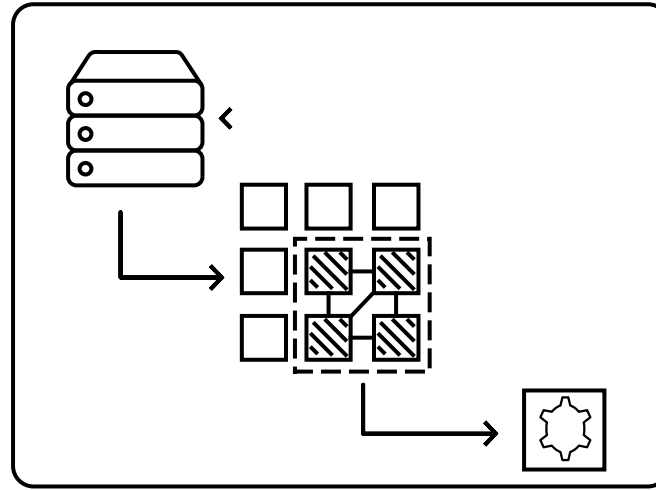
Loading Methods

BCP



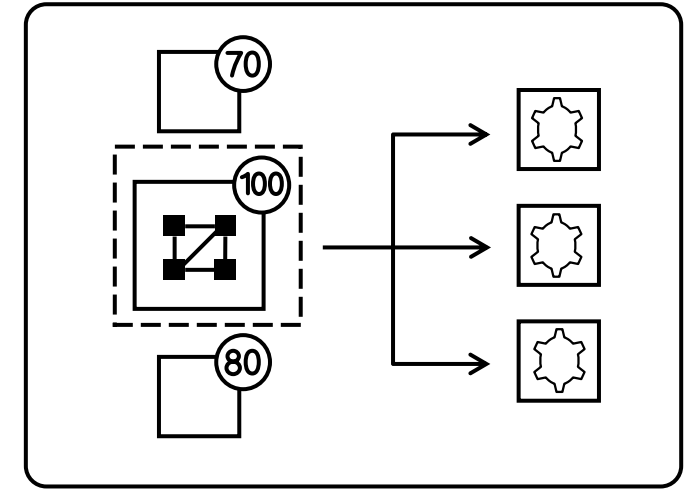
File based

SSIS



Heterogenous

PolyBase



File based

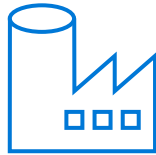
PolyBase benefits

Best practices for loading large amount of data



Leverages MPP architecture

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



Azure Data Factory support

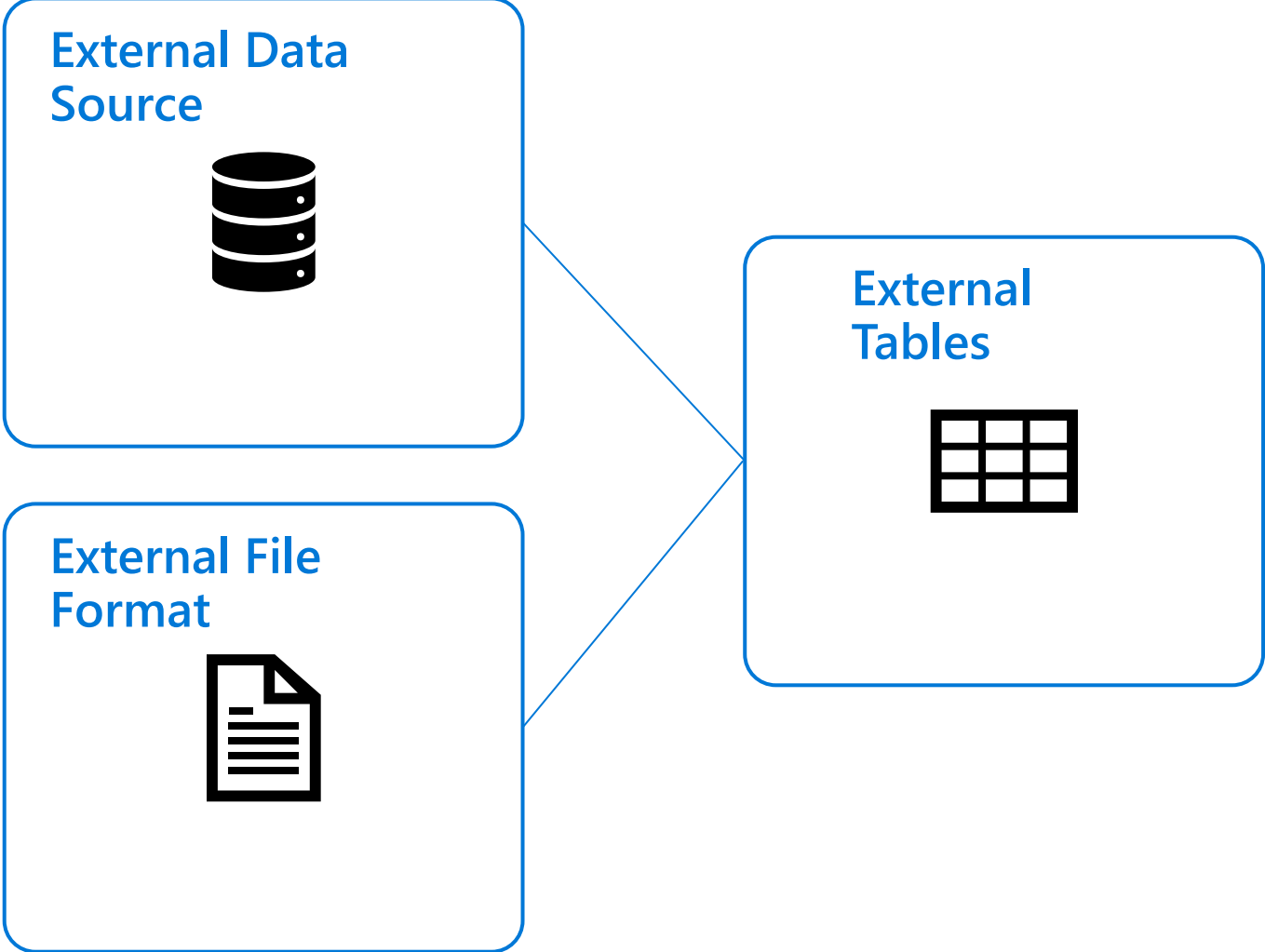
Azure Data Factory also supports PolyBase loads and can achieve similar performance to running PolyBase manually



Variety of file formats

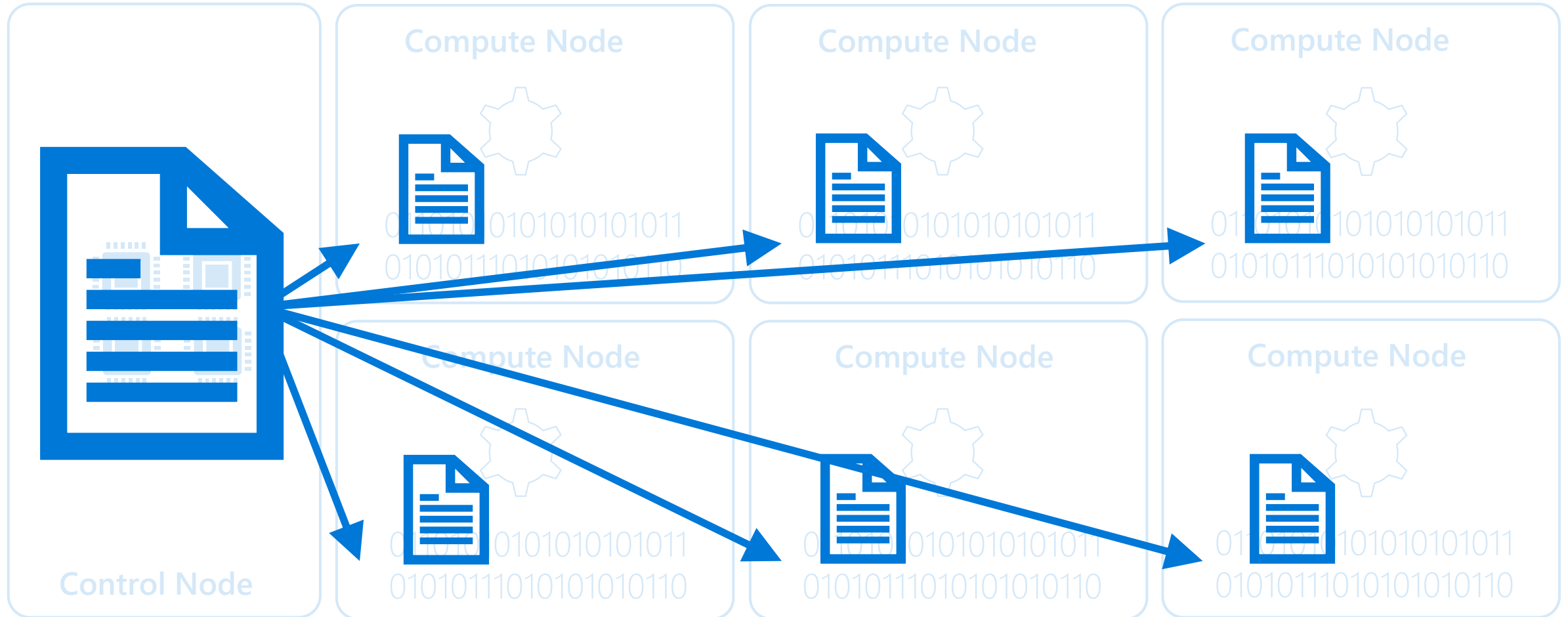
PolyBase supports a variety of file formats including RC, ORC and Gzip files.

Components of PolyBase

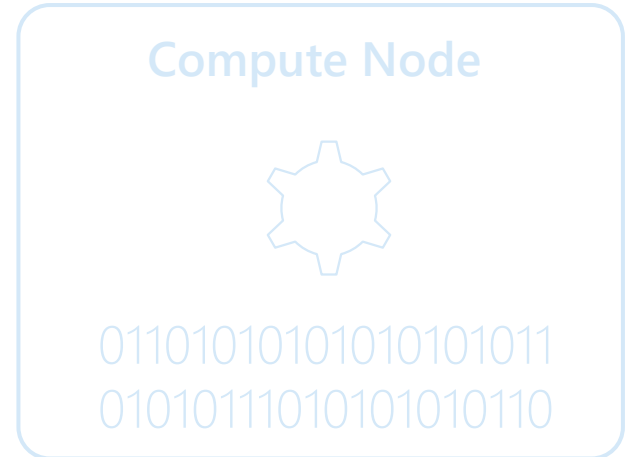
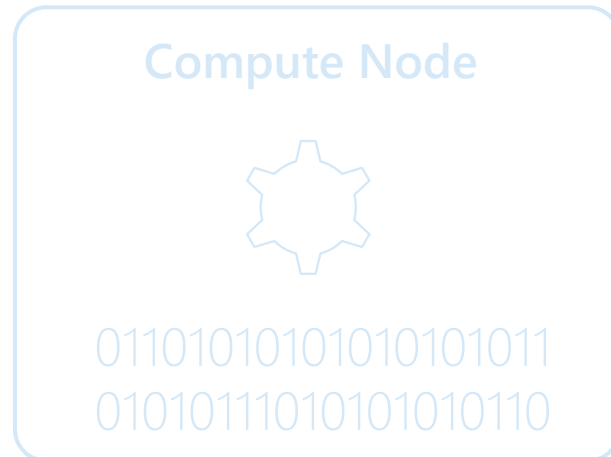
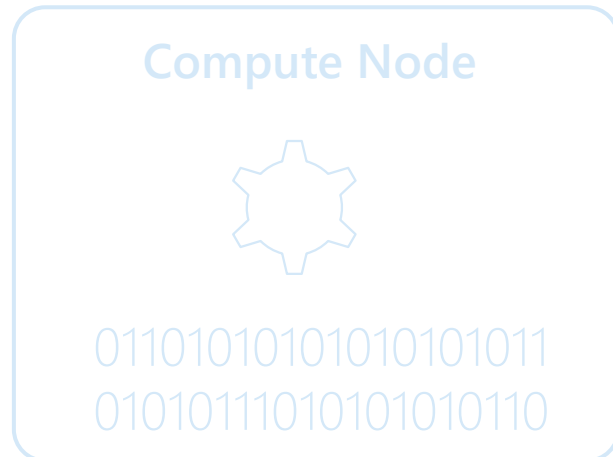
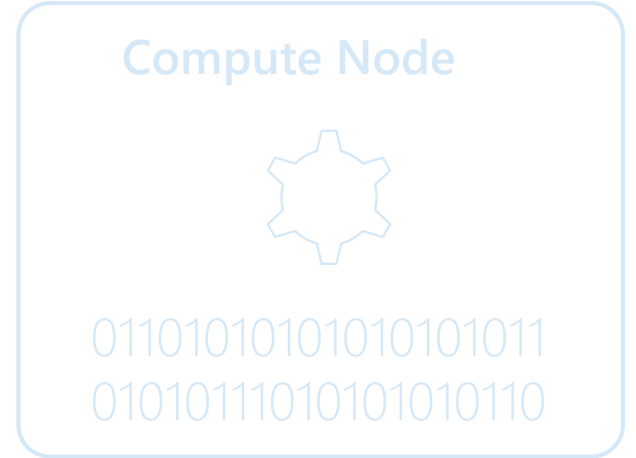
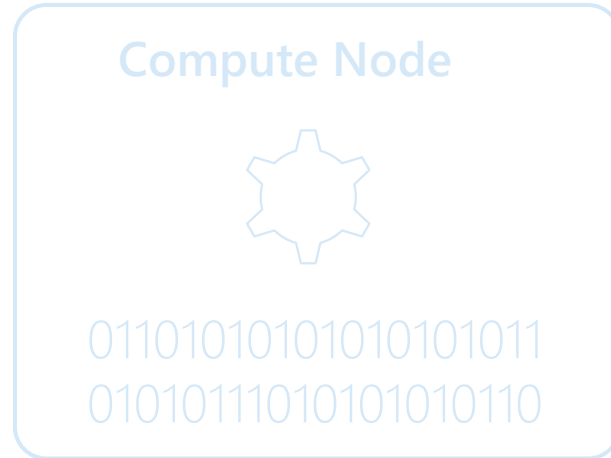
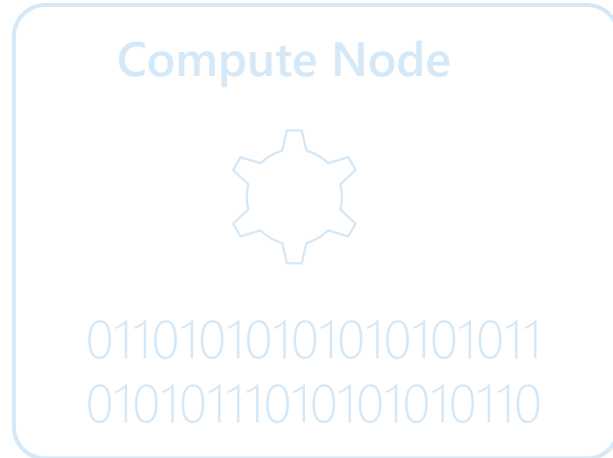
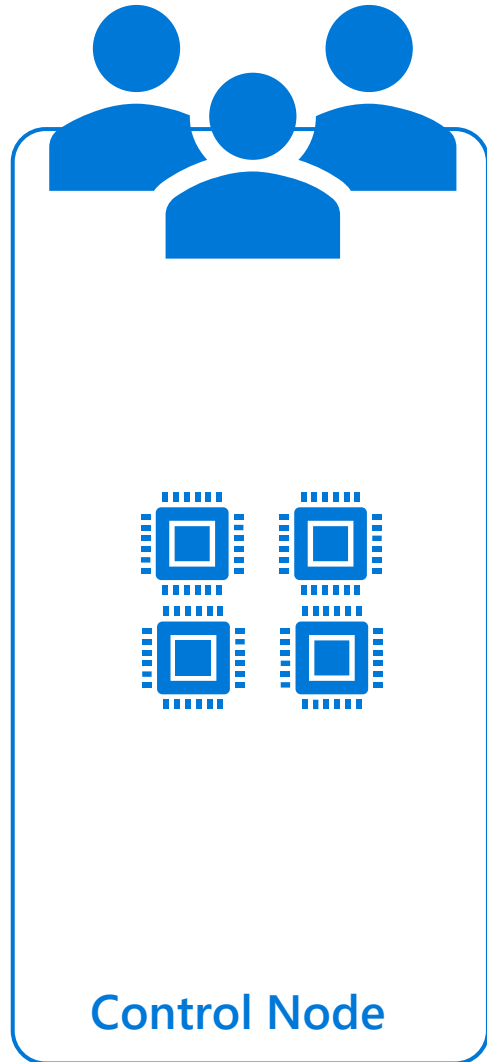


Loading best practices

Manage your files



Reduce concurrent access



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

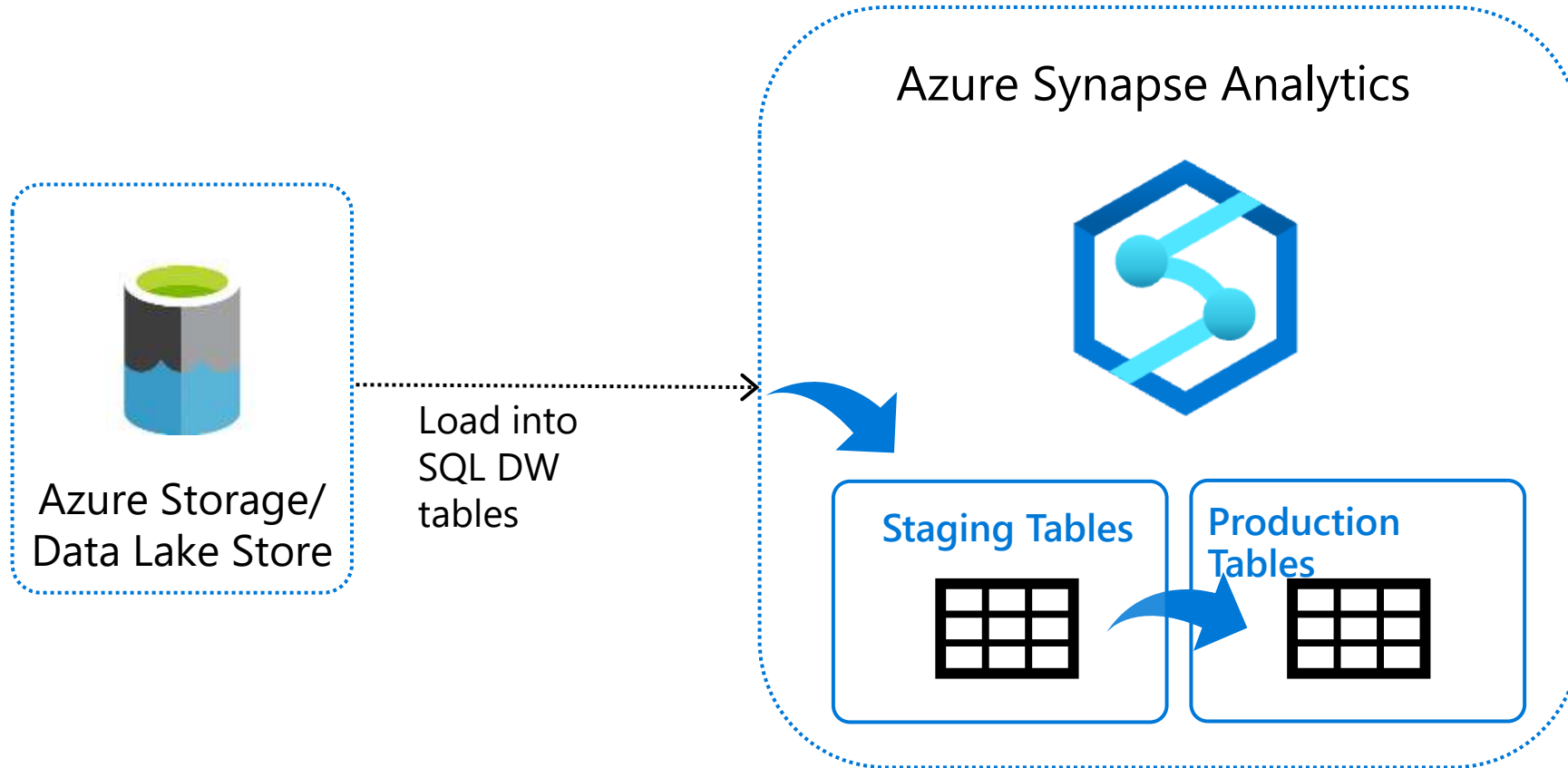
Compute Node



01101010101010101011
01010111010101010110

Optimize your loads

Staging data, a 2 step process



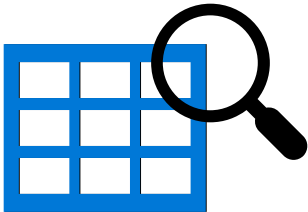
Create statistics after loading

Improve the query performance for users

Azure Synapse Analytics



Production Tables



Demo:
Loading data into
Azure Synapse Analytics Data Warehouse



Optimizing data warehousing query performance

Luis Silva

Cloud Solution Architect, Data Platform

Agenda

 What is Azure Synapse Analytics

 Maximizing performance

 Query performance tuning

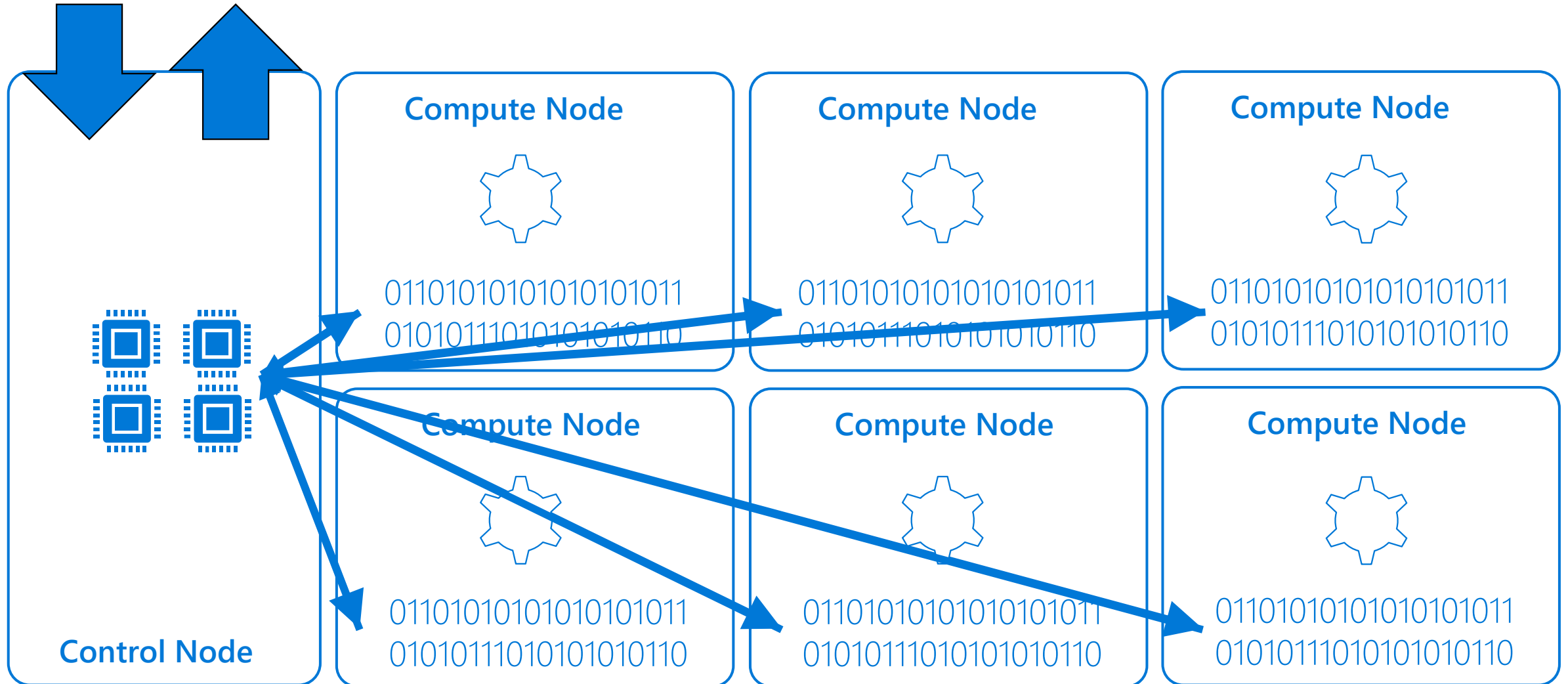
What is Azure Synapse Analytics?



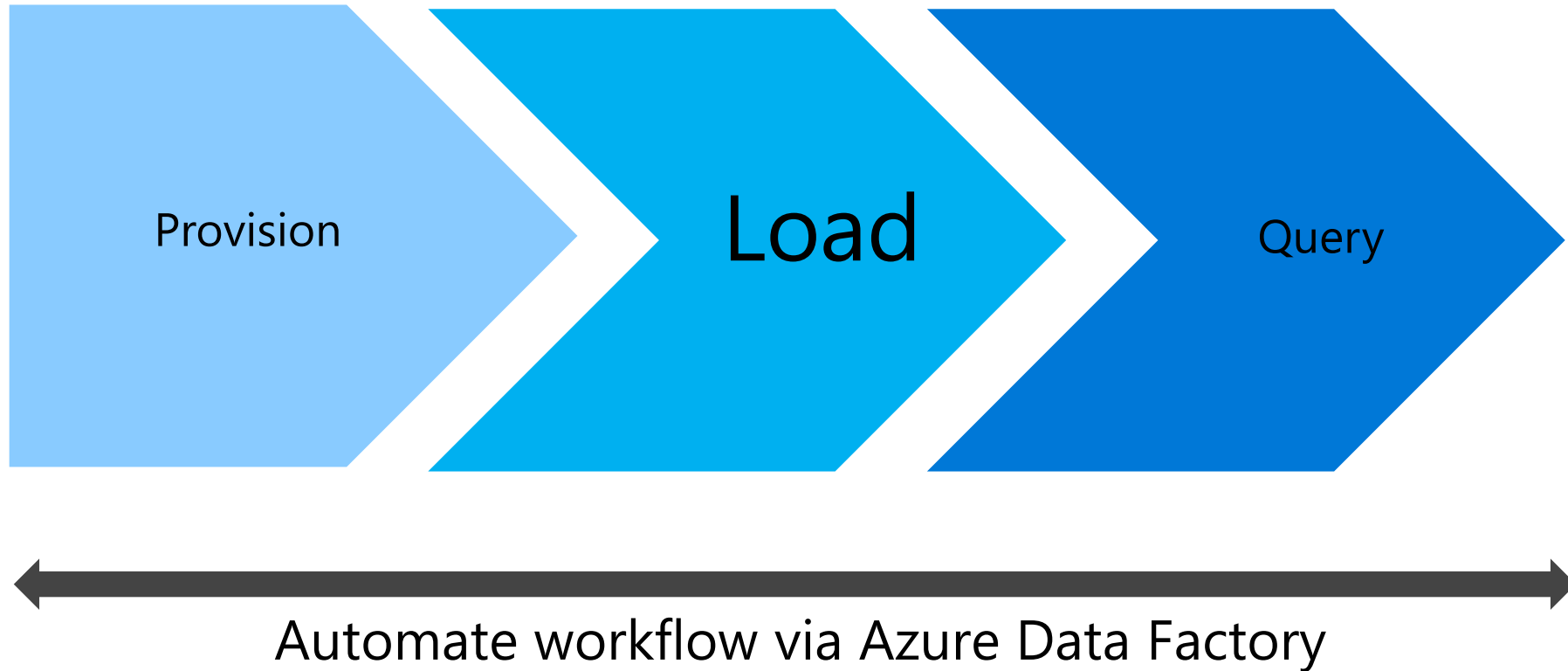
Azure Synapse Analytics

A **limitless** analytics service with **unmatched time to insight**, that delivers insights from all your data, **across data warehouses and big data** analytics systems, **with blazing speed**

Data Warehouse Architecture

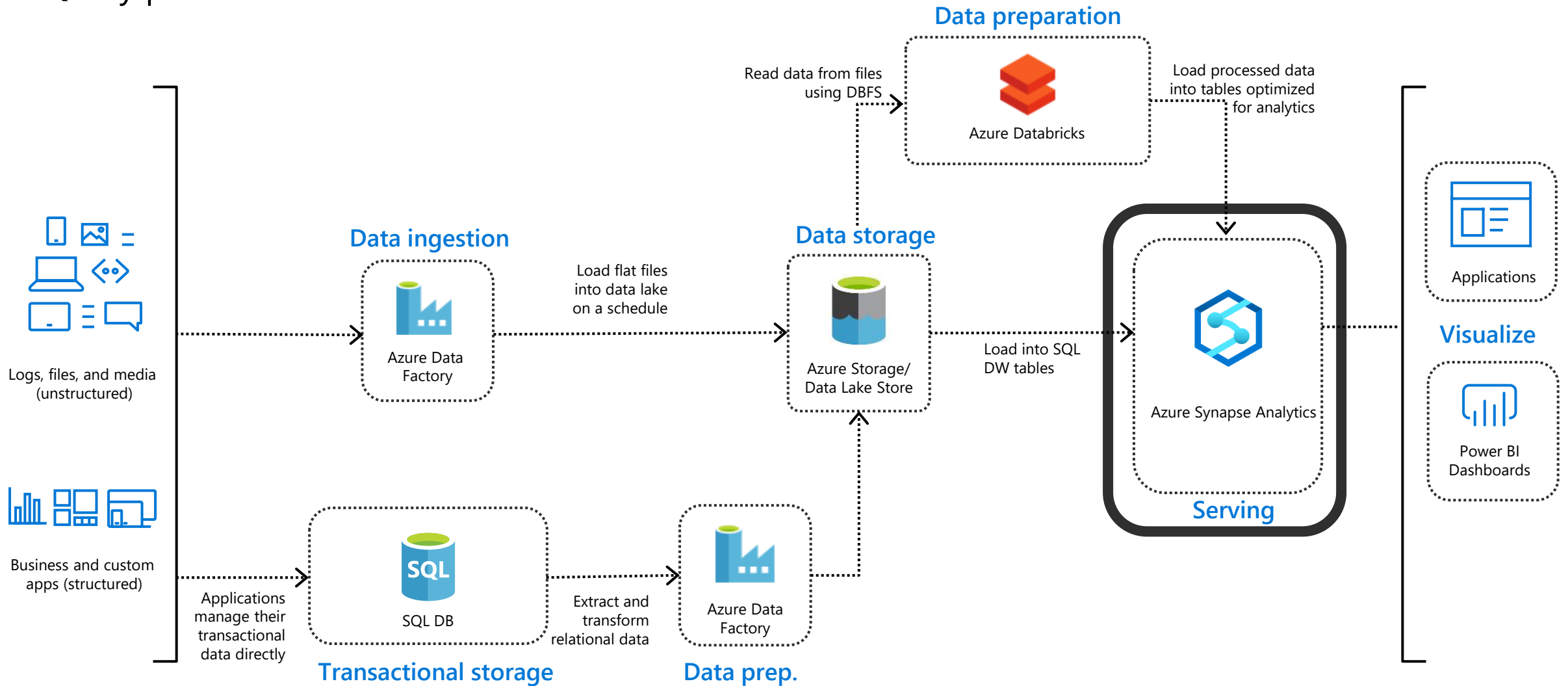


Data Warehouse Processes



Data warehouse performance in Azure Synapse Analytics

Query performance

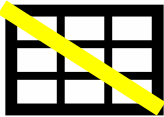


Maximizing Performance


Maximizing Query Performance

Table distribution

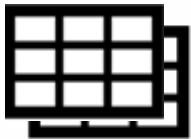
Round Robin
Tables



Hash Distributed
Tables

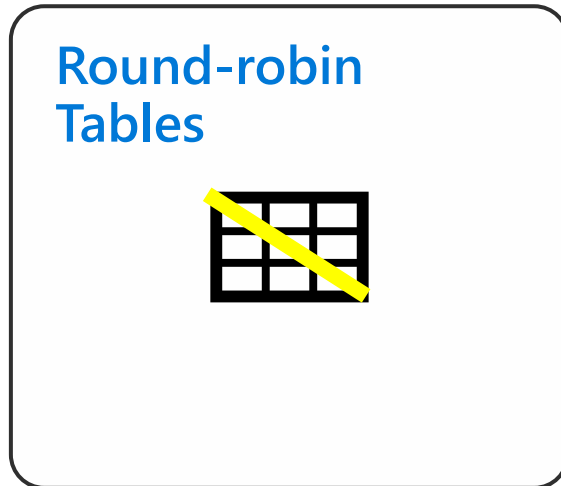


Replicated
Tables



Maximizing Query Performance

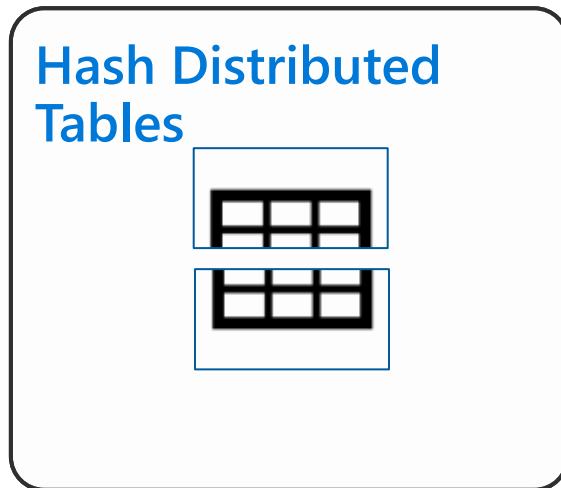
Round-robin distribution



- 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

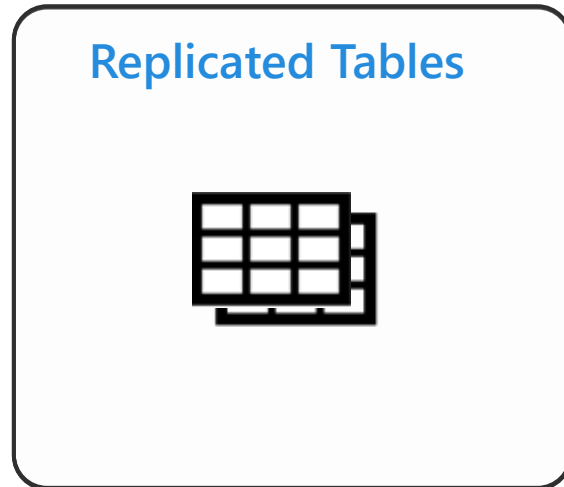
Hash distribution



- 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
- > You can convert existing round-robin tables to replicated tables to take advantage of the feature using a CTAS statement

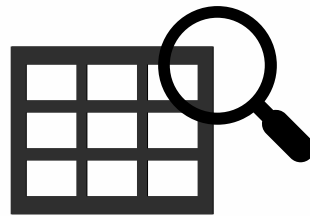
Create statistics after loading

Improve the query performance for users

Azure Synapse Analytics



Production Tables



Demo:

Creating distributed tables

Query Performance Tuning



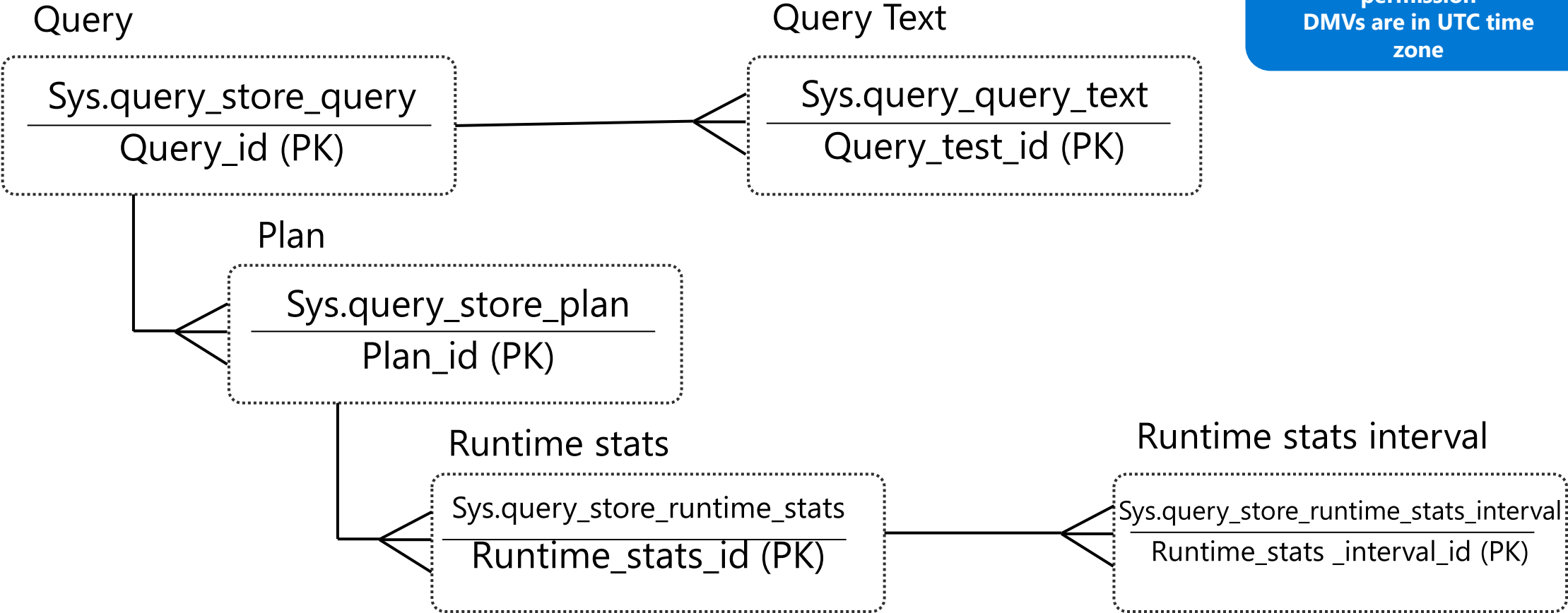
Query Data Store

- Overcomes the 10,000-row limit of DMV's output
- Pinpoint and fix queries with plan regression
 - View queries which produce multiple plans
 - 7-day retention period
 - Full query text
- A/B Testing with your Azure Synapse Analytics (SQL DW)
- Identify, improve and tune ad hoc queries
 - Top hitting queries for performance tuning

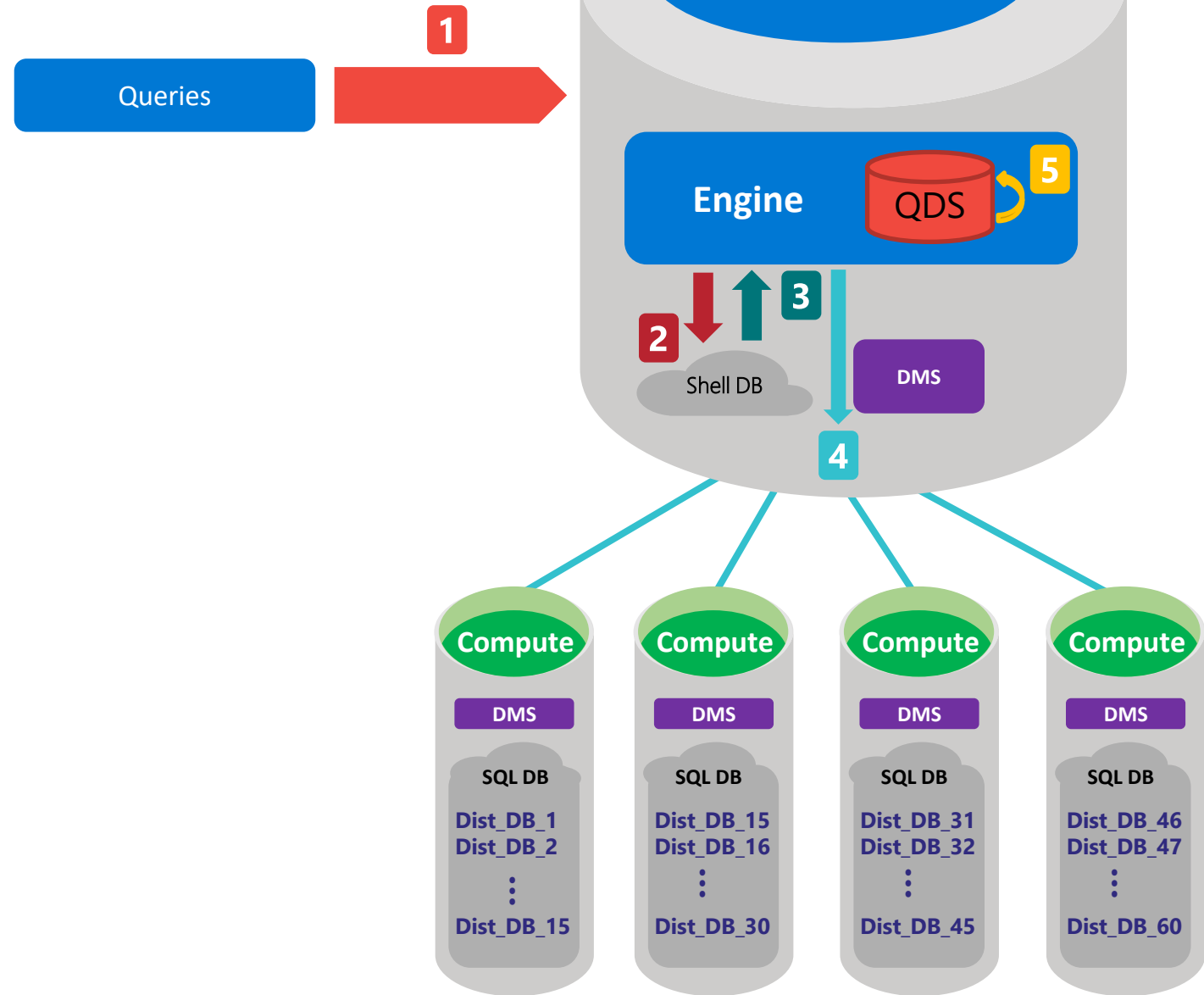
Query Data Store

Dynamic Management Views

VIEW DATABASE STATE
permission
DMVs are in UTC time
zone



Query execution with Query Data Store



- > Flush to disc every 15 minutes seconds
- > 10GB is the max storage size
- > Retention period is 7 days
- > Maximum plans per query is 200

Azure Synapse Analytics recommendations

Azure Synapse Analytics



Telemetry



Recommendation generation (every 24 hours)

Data skew + Replicate tables

Stats

Tempdb

Adaptive Cache

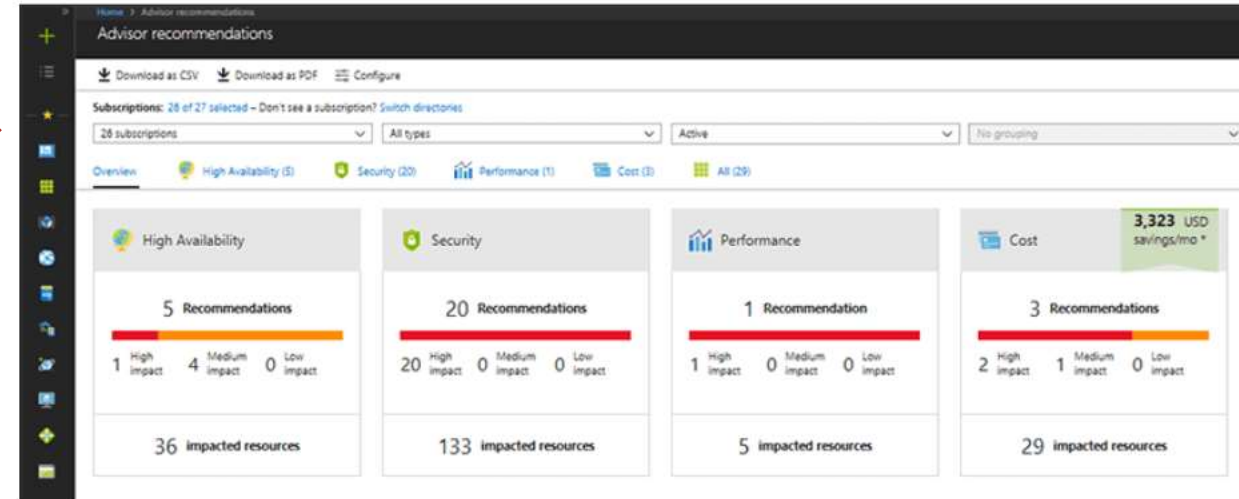
Recommendation API



Azure Advisor Recommendation

You have free Azure Advisor recommendations!

Azure Advisor is a free offering that analyzes your Azure usage and provides recommendations on how you can save money, improve performance, be more secure, and improve reliability of the solutions you already have running in Azure. [Learn more](#)



View my free recommendations



In Summary:

Query Performance

- Select the proper table distribution
- Detect data skew
 - Use Query Data store
 - Consider changing key columns
 - Only as fast as your slowest distribution
- Provision additional adaptive cache capacity
- Reduce tempdb contention
- Create and update statistics

Demo: Query Performance Tuning



Thank you!

