Caltech | Center for Technology & Management Education

Post Graduate Program in Cloud Computing

Powerd by simplilearn

Cloud

Cloud Computing

Caltech | Center for Technology & Management Education

PG CC - Microsoft Azure Architect Design: AZ:304

Powerd by simplilearn

Design Data Integration

Learning Objectives

By the end of this lesson, you will be able to:

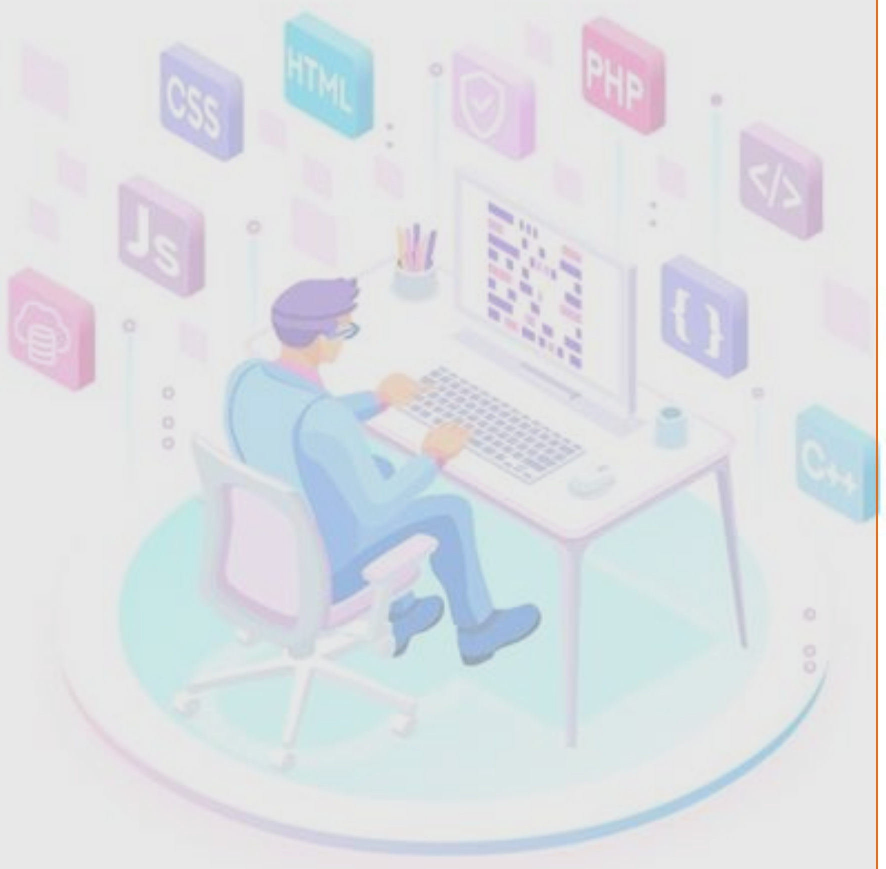
- Classify the modern data platform reference architecture
- Recommend a data flow to meet business requirements
- Classify the Azure synapse analytics architecture
- Recommend a solution for data integration

A Day in the Life of an Azure Architect

You are working as an Cloud Data Architect in ABC Inc. The company has decided to move its infrastructure and services to the cloud and wants to build a predictive analysis solution for their ecommerce application. Your company needs a production environment for their work.

Also, for the ecommerce application, company needs a data store where they can store the static images of the product.

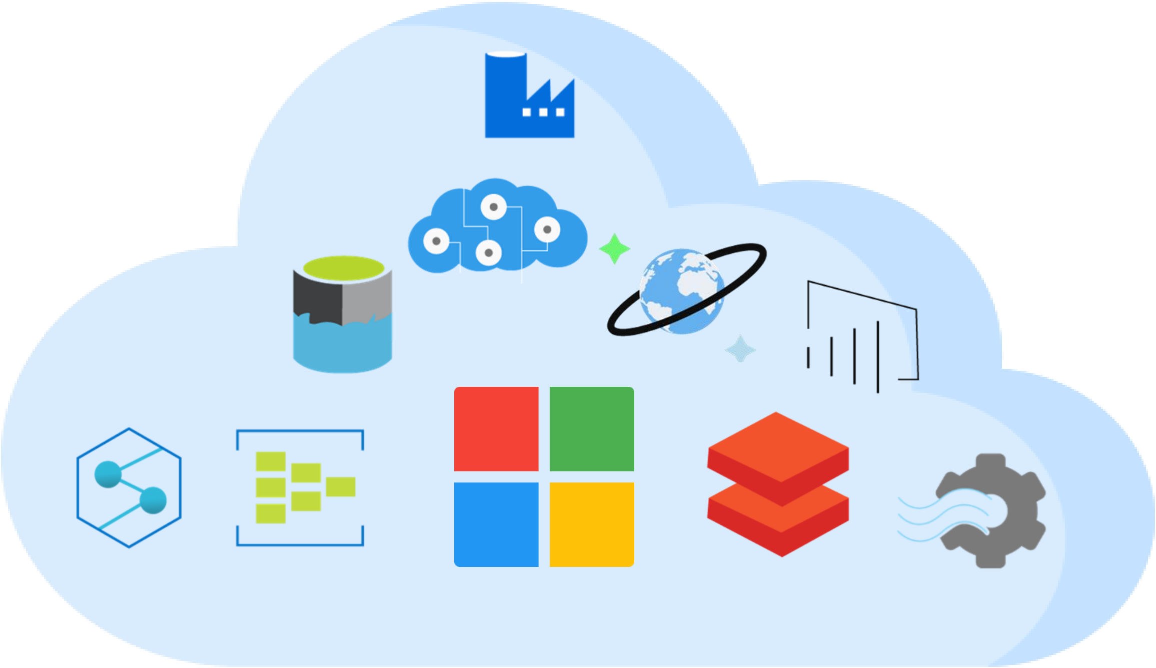
To achieve all of the above, along with some additional features, we would be learning a few concepts in this lesson that will help you find a solution for the above scenario.



Recommend a Data Flow to Meet Business Requirements

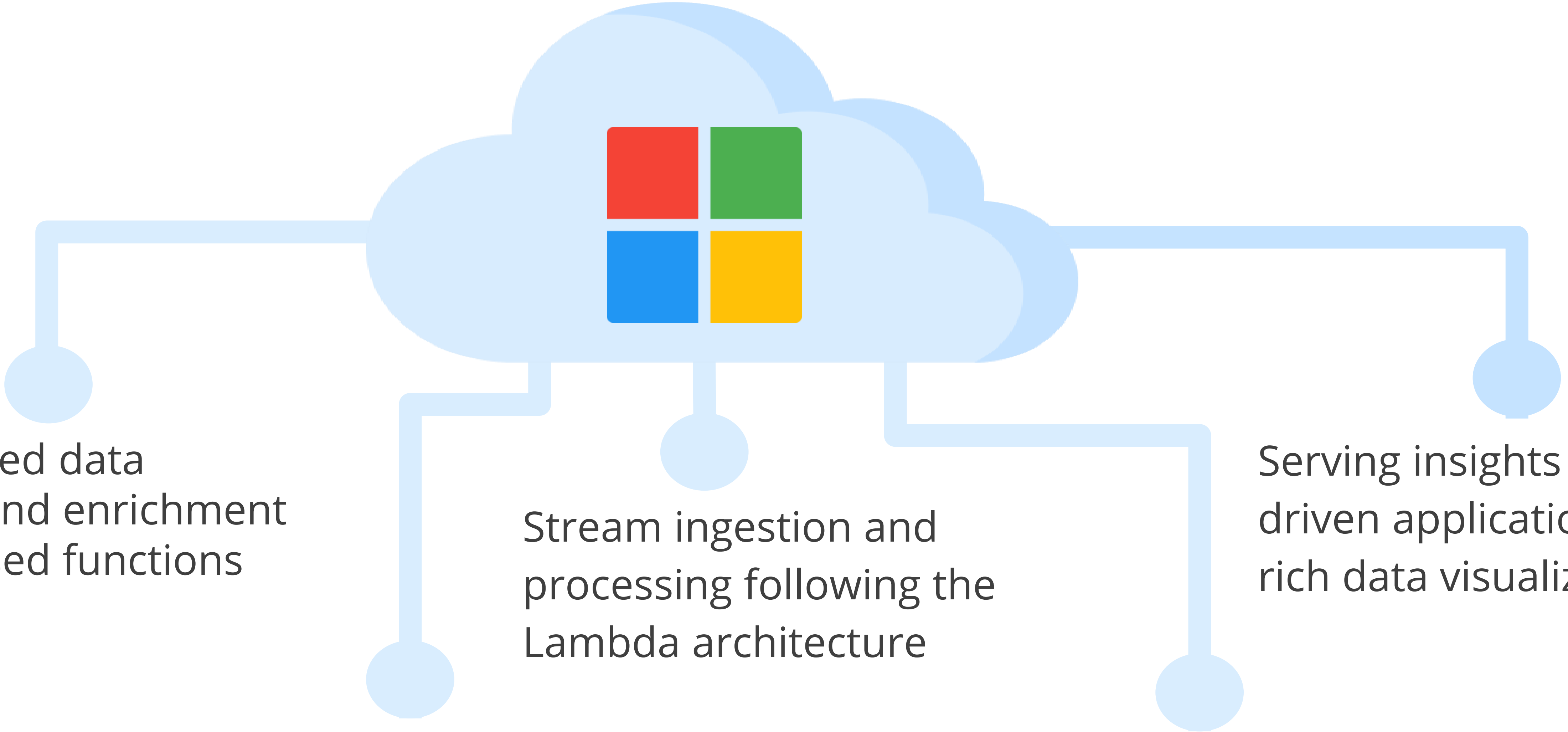
Azure Data Platform End-to-End

The Azure Data Services family is used to build a modern data platform that can handle the most common data challenges in an organization.



Azure Data Platform End-to-End

This solution architecture demonstrates how a single, unified data platform can be used to meet the most common requirements for:



Unstructured data ingestion and enrichment with AI-based functions

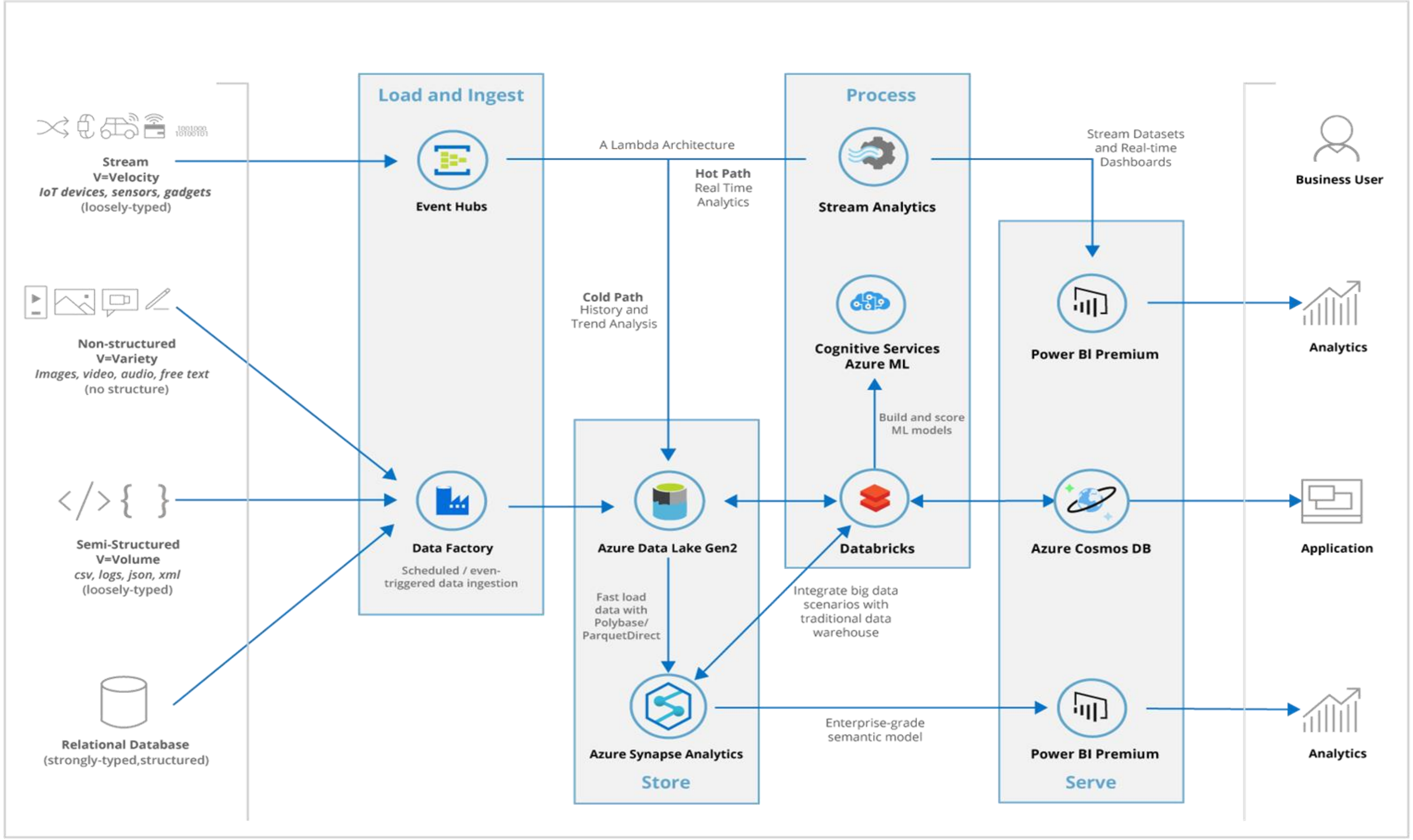
Traditional relational data pipelines

Stream ingestion and processing following the Lambda architecture

Big Data transformations

Serving insights for data-driven applications and rich data visualization

Modern Data Platform Reference Architecture



The essential components are:

- Relational database
- Semi-structured data sources
- Non-structured data sources
- Streaming

Architecture Use Cases

The architecture can also be used to:

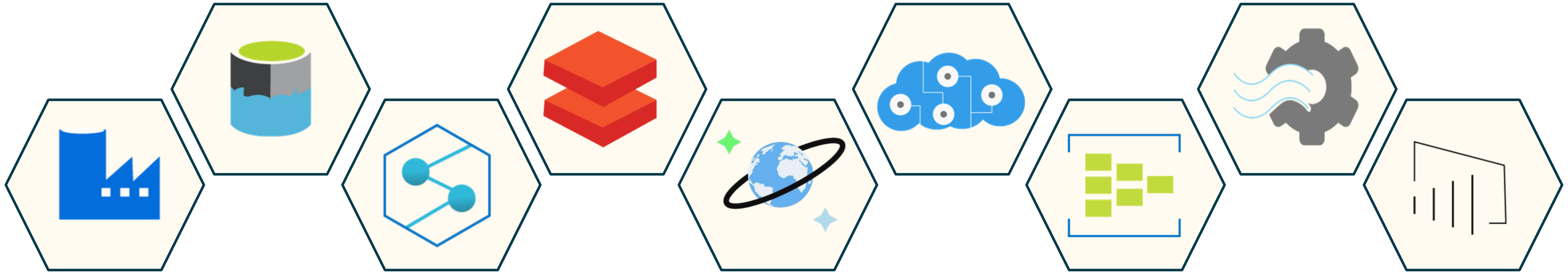
Provide an enterprise-wide data hub that includes a structured data warehouse and a semi-structured and unstructured data lake. This data hub becomes your data's single source of truth

Use Big Data processing techniques to bridge relational data sources with other unstructured datasets

Use semantic modeling and advanced visualization tools for better data analysis

Architecture Components

The following Azure services are used in the architecture::



Azure Data Factory

Azure Data Lake Gen2

Azure Synapse Analytics

Azure Databricks

Azure Cosmos DB

Azure Cognitive Services

Azure Event Hubs

Azure Stream Analytics

Microsoft Power BI

Recommend a Solution for Data Integration

Data Flows Using Azure Data Factory

Azure Data Factory uses Code-Free Extract, Transform, Load (ETL) Service

INGEST

- Multi-cloud on-premise hybrid copy data
- 90+ native connectors
- Serverless and auto-scale
- Use wizard for quick copy jobs

CONTROL FLOW

- Design code-free data pipelines
- Generate pipelines via SDK
- Utilize workflow constructs: loops, branches, conditional execution, etc.

DATA FLOW

- Code-free data transformations that execute in Spark
- Scale-out with Azure Integration Runtimes
- Generate data flows via SDK

Data Flows Using Azure Data Factory

Azure Data Factory uses Code-Free Extract, Transform, Load (ETL) Service

SCHEDULE

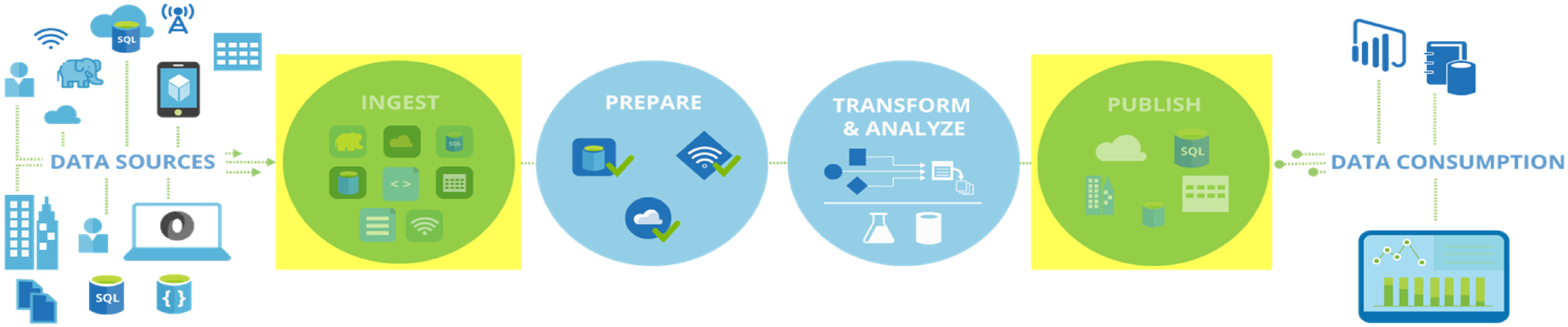
- Build and maintain operational schedules for your data pipelines
- Wall clock, event-based, tumbling windows, chained

MONITOR

- View active executions and pipeline history
- Detail activity and data flow executions
- Establish alerts and notifications

How Data Factory Works

You can use the *Copy* activity in Azure Data Factory to copy data between on-premises and cloud data storage.



How Data Factory Works

Azure Data Factory contains a series of interconnected systems for data engineers:

Connect and collect

Data Factory moves data from on-premises and cloud source data stores to a centralized data store in the cloud for analysis by using Copy activity in a data pipeline.

Transform and enrich

Data flows make it possible for data engineers to create data transformation graphs that run on Spark without having to know anything about Spark clusters or programming.

How Data Factory Works

CI/CD and publish

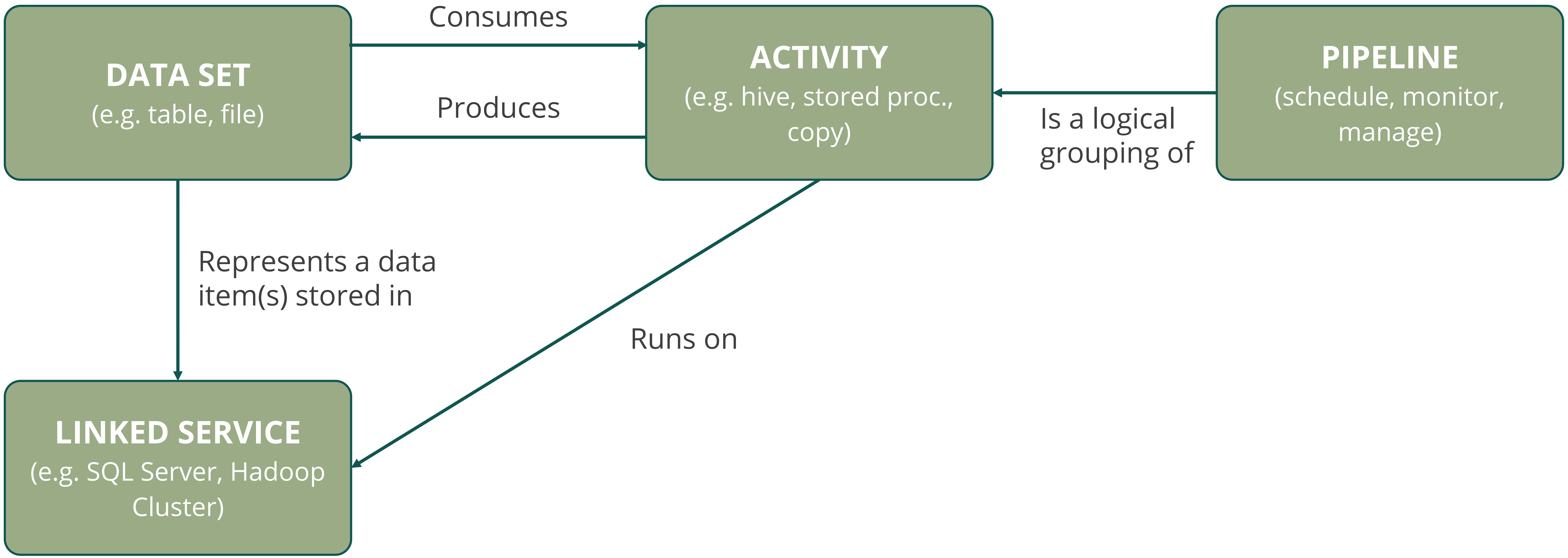
Full support for continuous integration and continuous delivery (CI/CD) of data pipelines utilizing Azure DevOps and GitHub enables incremental development and delivery of ETL procedures prior to publishing.

Monitor

Pipeline monitoring is built-in with Azure Monitor, API, PowerShell, Azure Monitor logs, and Azure portal health panels.

Data Factory: Process

Illustration of data factory process:



DATA SET
(e.g. table, file)

Consumes

Produces

ACTIVITY
(e.g. hive, stored proc., copy)

Is a logical grouping of

PIPELINE
(schedule, monitor, manage)

Represents a data item(s) stored in

Runs on

LINKED SERVICE
(e.g. SQL Server, Hadoop Cluster)

Data Factory: Key Concepts

The key concept of components of data factory process:

Dataset

Datasets are data structures within data stores that simply point to or refer to the data you want to use as inputs or outputs in your operations.

Activity

A processing step in a pipeline is represented by activity.
For instance, you could use a copy activity to copy data from one data store to another data store.

Data Factory: Key Concepts

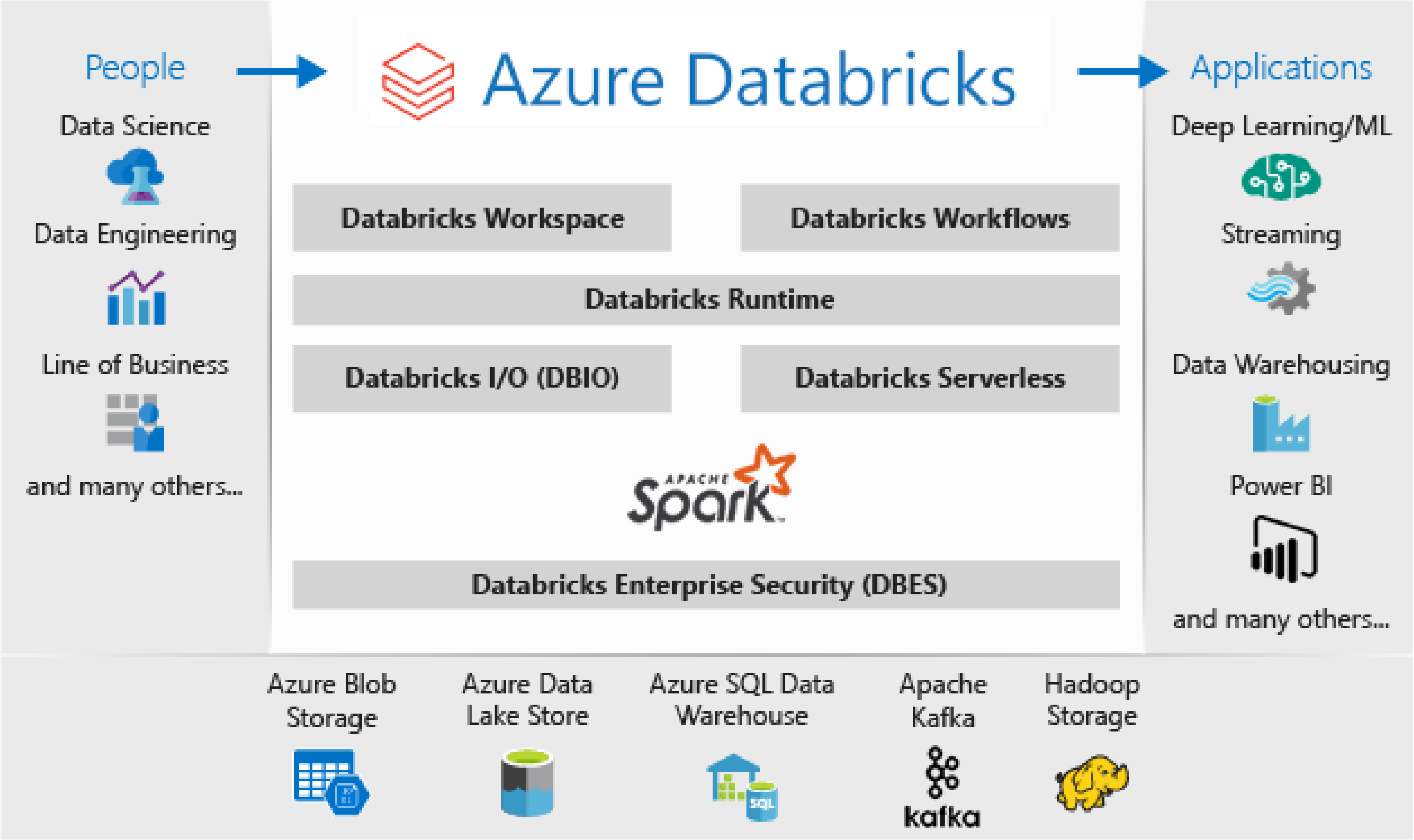
Pipeline

A pipeline is a logical grouping of activities that performs a unit of work. For instance, A pipeline can contain a group of activities that ingests data from an Azure blob, and then runs a Hive query on an HDInsight cluster to partition the data.

Linked service

Linked service establishes a link to the data source, and a dataset describes the data's structure.
For instance, to connect to an Azure Storage account, an Azure Storage-linked service specifies a connection string.

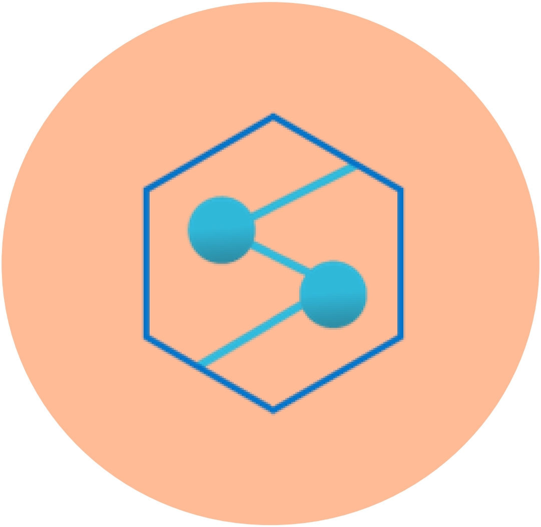
Integrate Data Factory and Databricks



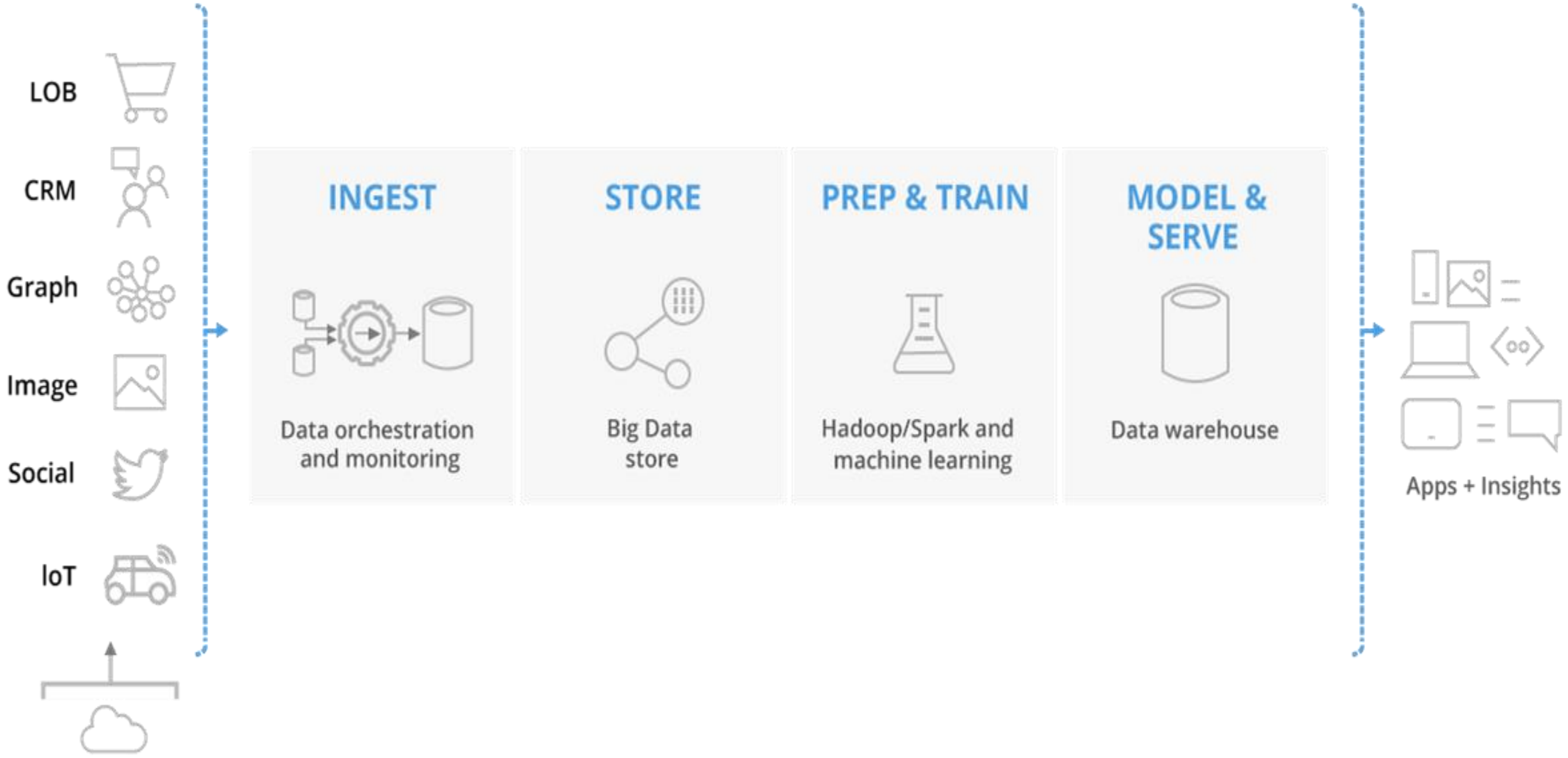
1. Create an Azure storage account
2. Create a Data Factory instance: Portal
3. Create a data workflow pipeline: Copy activity
4. Add a Databricks notebook to the pipeline
5. Analyze the data: train data

Azure Synapse Analytics

Azure Synapse is an analytics service that brings together enterprise data warehousing and Big Data analytics with a unified experience to ingest, prepare, manage, and serve data for immediate BI and machine learning needs.



Azure Synapse Components

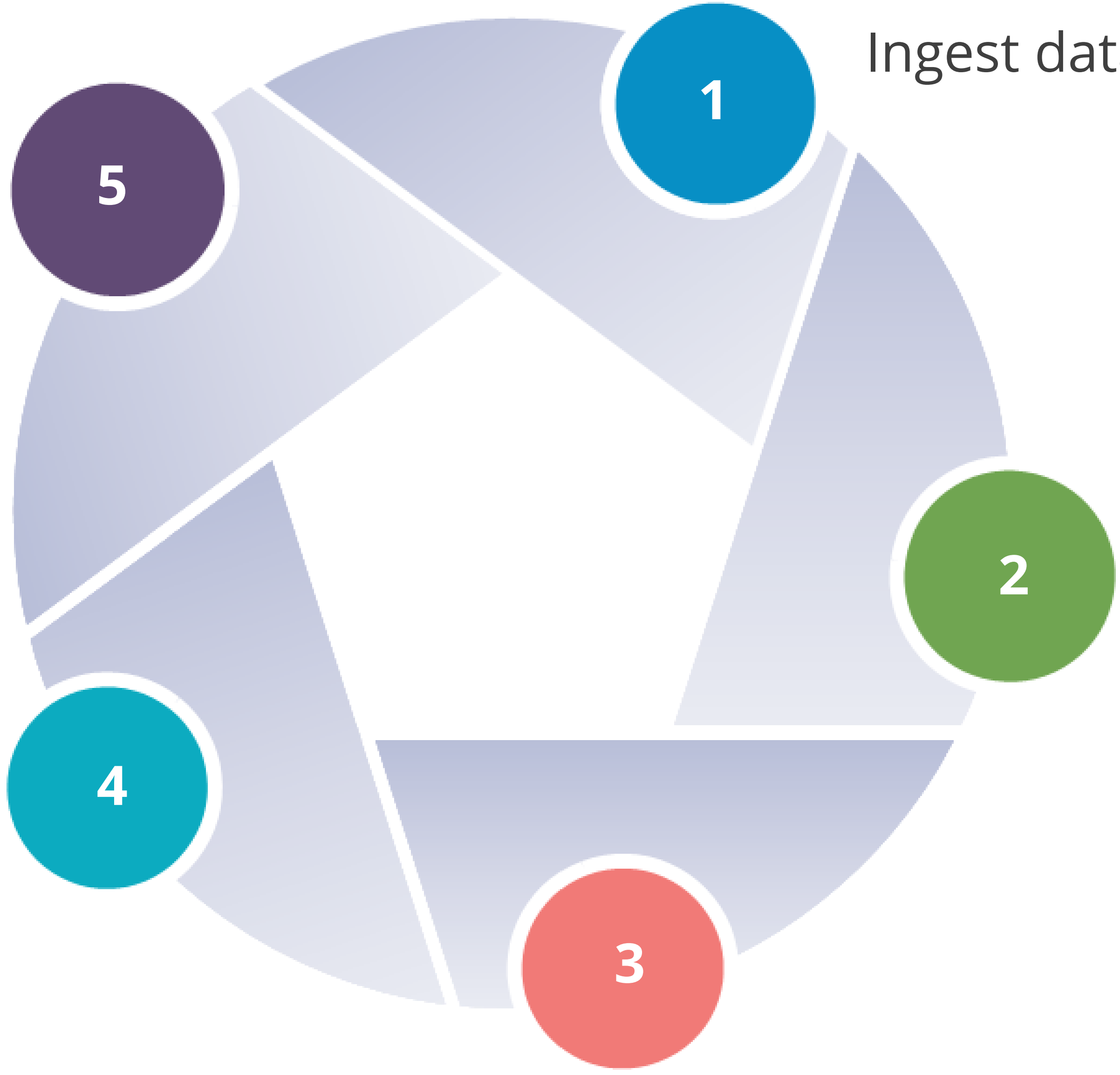


Azure Synapse has four components:

- Synapse SQL
- Spark
- Synapse Pipelines
- Studio

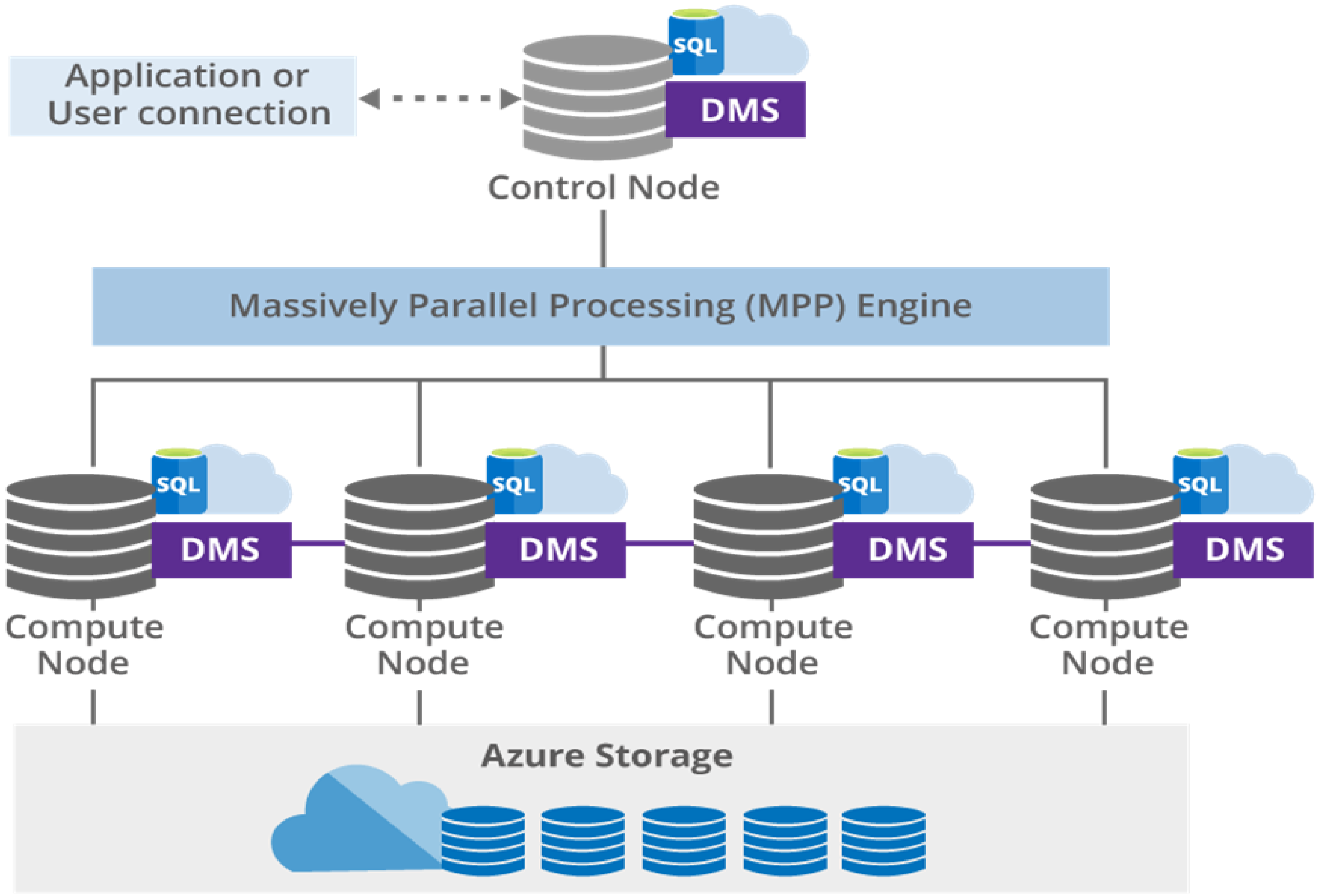
Data Flow

The process of data flow in Azure:



1. Ingest data
2. Prepare data (Hadoop, Spark, and machine learning)
3. PolyBase to query Big Data stores
4. T-SQL queries to populate pool tables
5. Stores data in relational tables with columnar storage

Azure Synapse Analytics Architecture



Azure Storage Sharding Patterns

- Hash
- Round Robin
- Replicate

Control node

- The Massively Parallel Processing (MPP) engine
- T-SQL query

Compute nodes

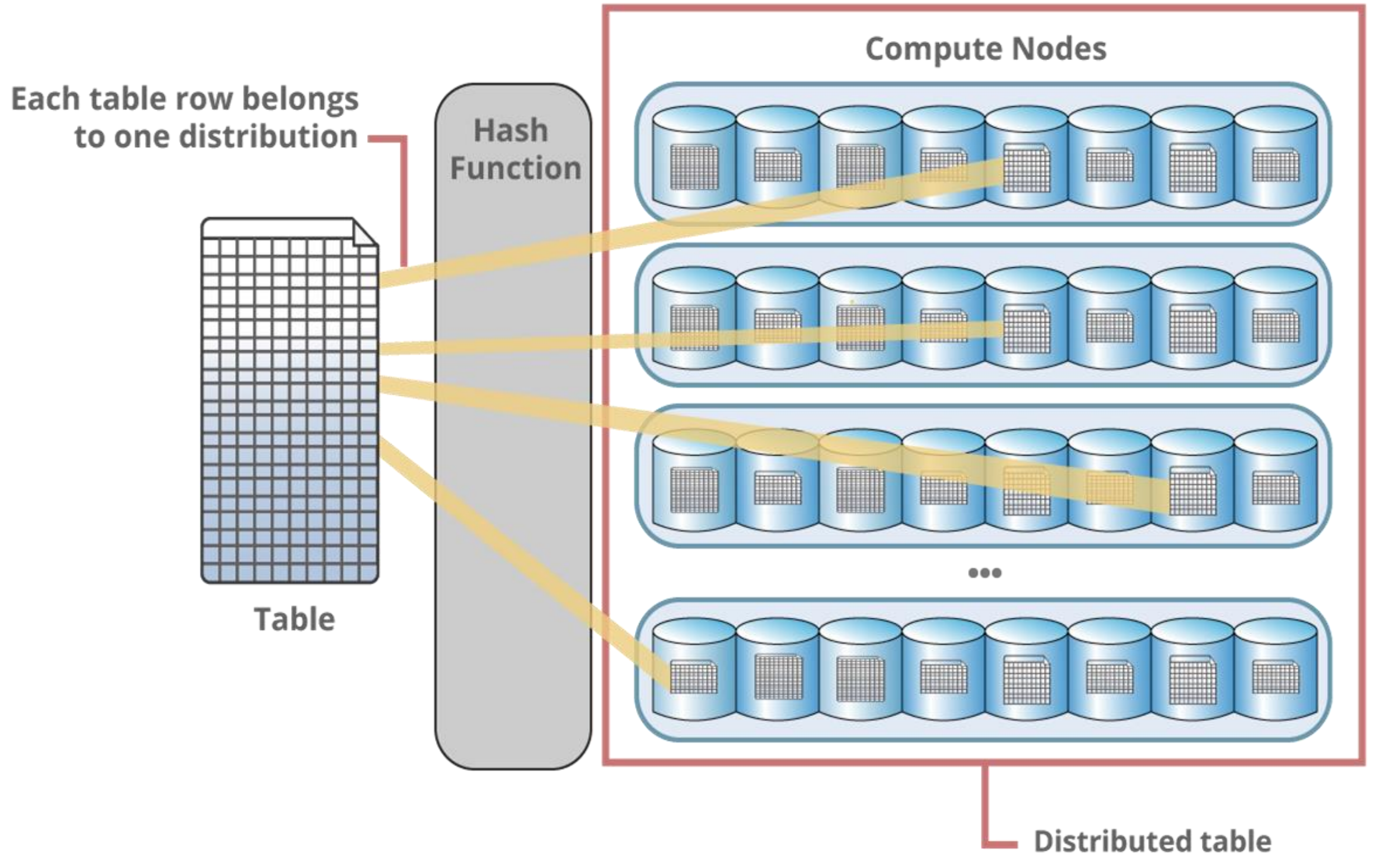
- Distributed processing

Data Movement Service

- Transport between notes

Distributions

A distribution is the basic unit of storage and processing for parallel queries that run on distributed data.



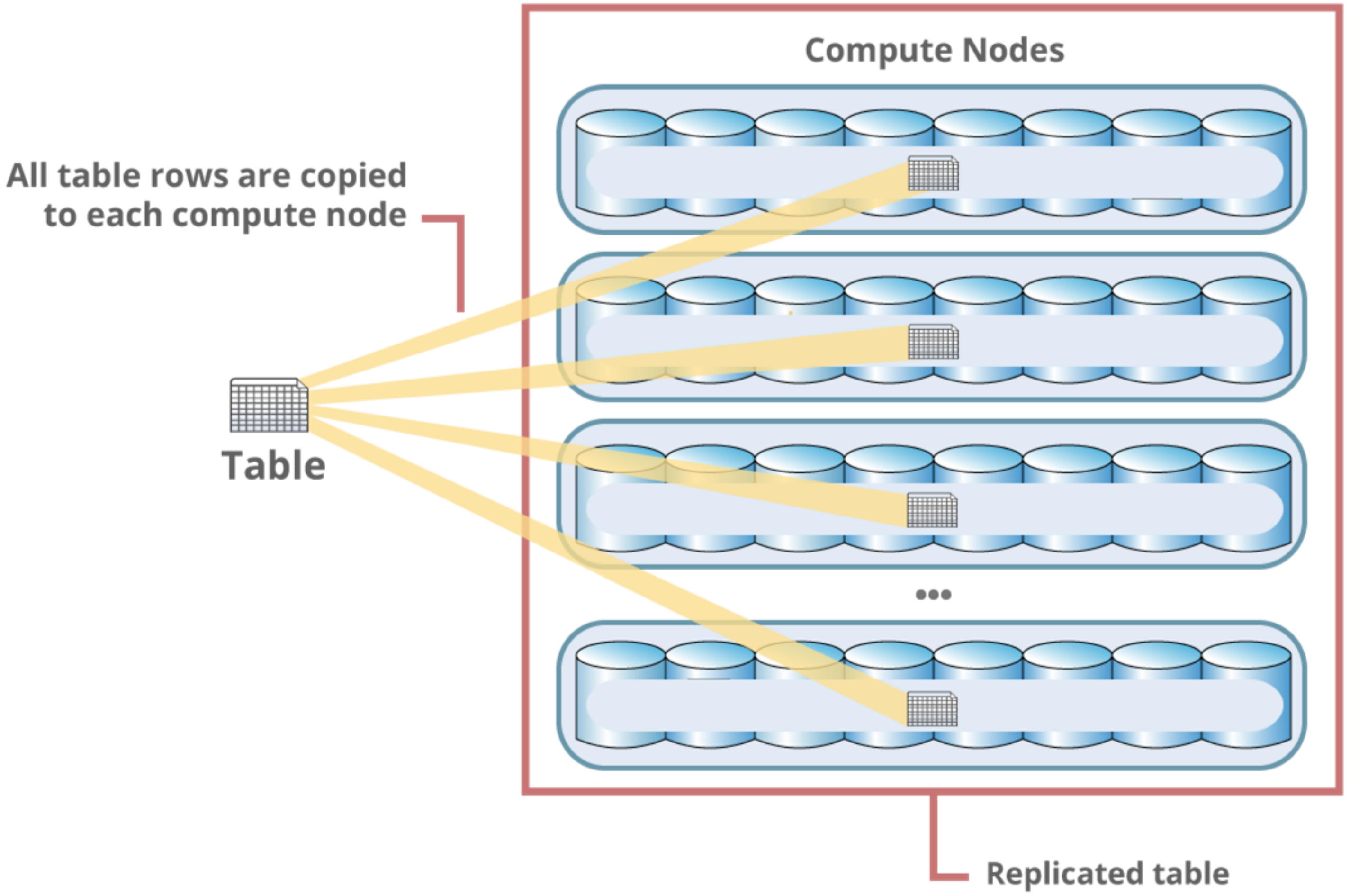
Hash-Distributed Tables

- Each row belongs to one distribution
- A deterministic hash algorithm assigns each row to one distribution
- The number of table rows per distribution varies

Round-Robin Distributed Tables

- A round-robin distributed table distributes data evenly across the table
- A distribution is first chosen at random

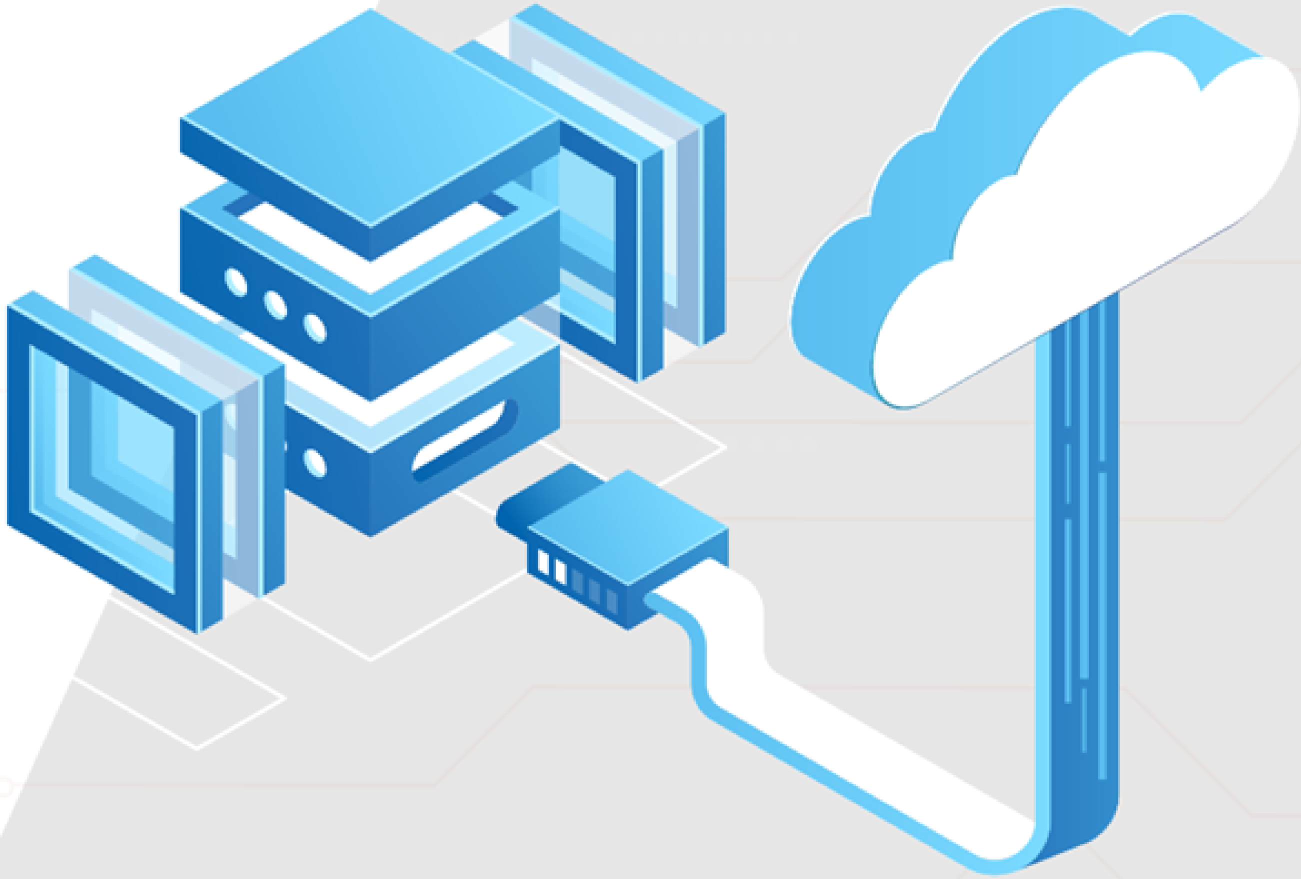
Replicated Tables



- A replicated table caches a full copy of the table on each compute node.
- It removes the need to transfer data among compute nodes.
- It is best utilized with small tables.

Key Takeaways

- The Azure Data Services family is used to build a modern data platform that can handle the most common data challenges in an organization.
- In the modern data platform reference architecture, the most important components are Relational, Semi-structured, Non-structure, and Streaming.
- Azure Synapse is an analytics service that brings together enterprise data warehousing and Big Data analytics.
- A distribution is the basic unit of storage and processing for parallel queries that run on distributed data.



Thank you