Cloud Computing

Caltech | Center for Technology & Management Education

Designing Infrastructure Solutions on Azure

Powerd by simplilearn

Design Data Integration

A Day in the Life of an Azure Architect

You are working as an architect in an organization that has decided to develop SaaS apps. The company assigns built-in roles to users, groups, service principals, and managed identities. When assigning a role to a user, consider what actions the role can perform and what the scope of those operations is.

- You can load data into the storage, then transform or compute with Spark.
- You can load the result set into the app storage that is the backend for our SaaS app.

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

Learning Objectives

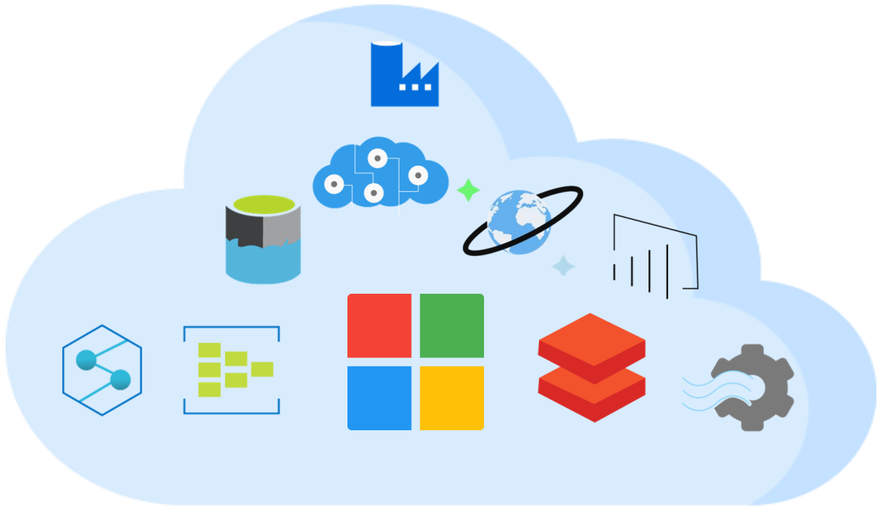
By the end of this lesson, you will be able to:

- 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

Recommend a Data Flow to Meet Business Requirements

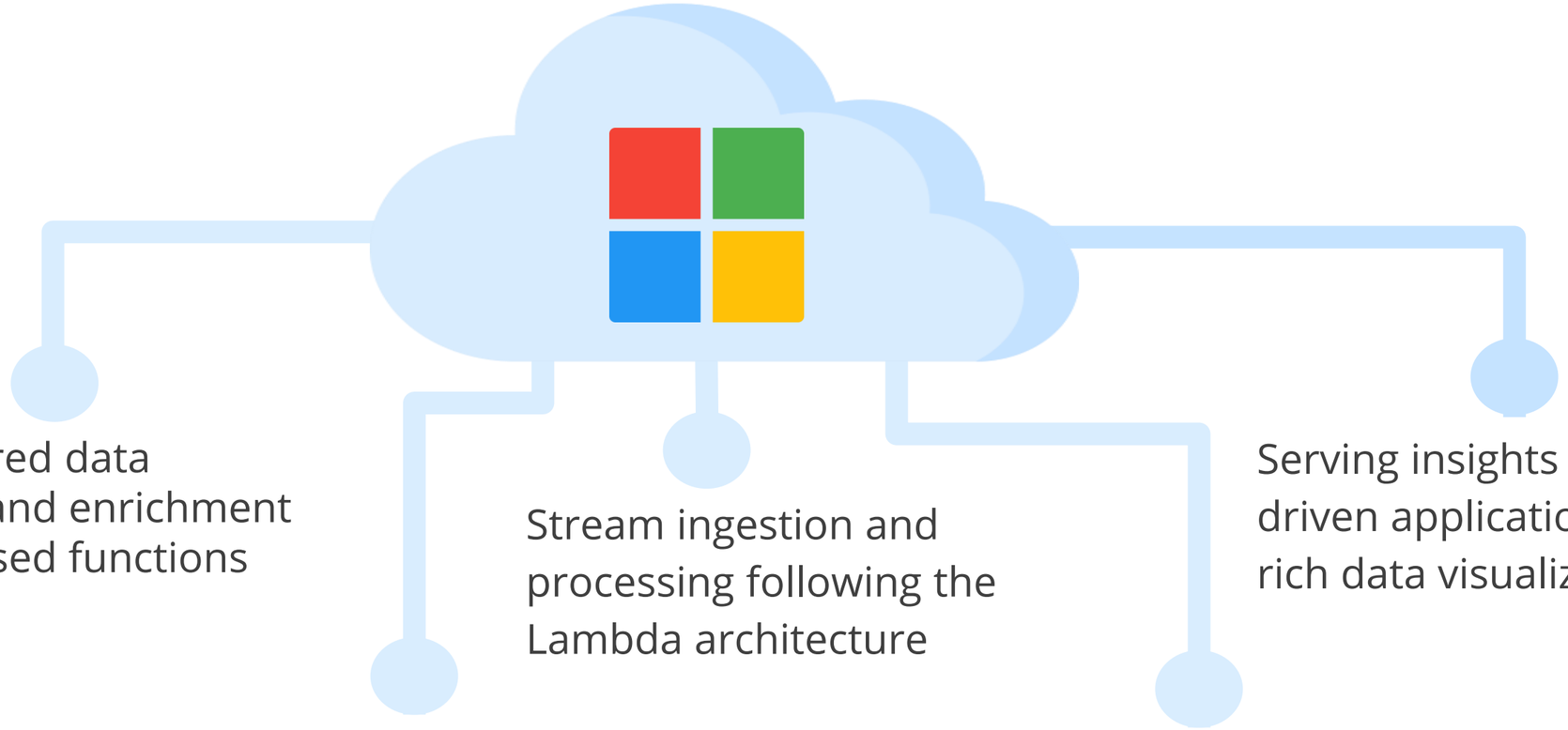
Azure Data Platform End-to-End

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

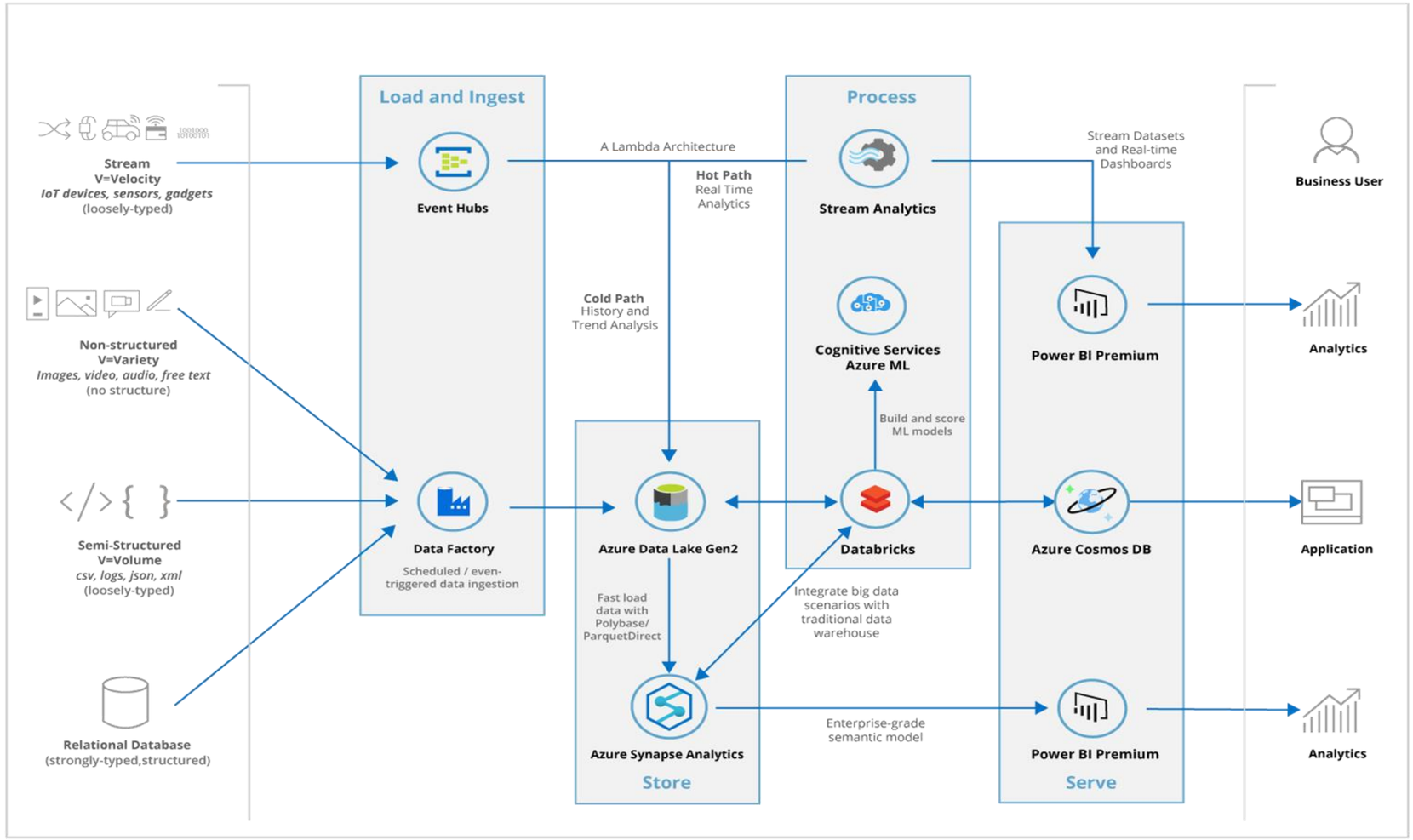
Stream ingestion and processing following the Lambda architecture

Serving insights for data-driven applications and rich data visualization

Traditional relational data pipelines

Big Data transformations

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:

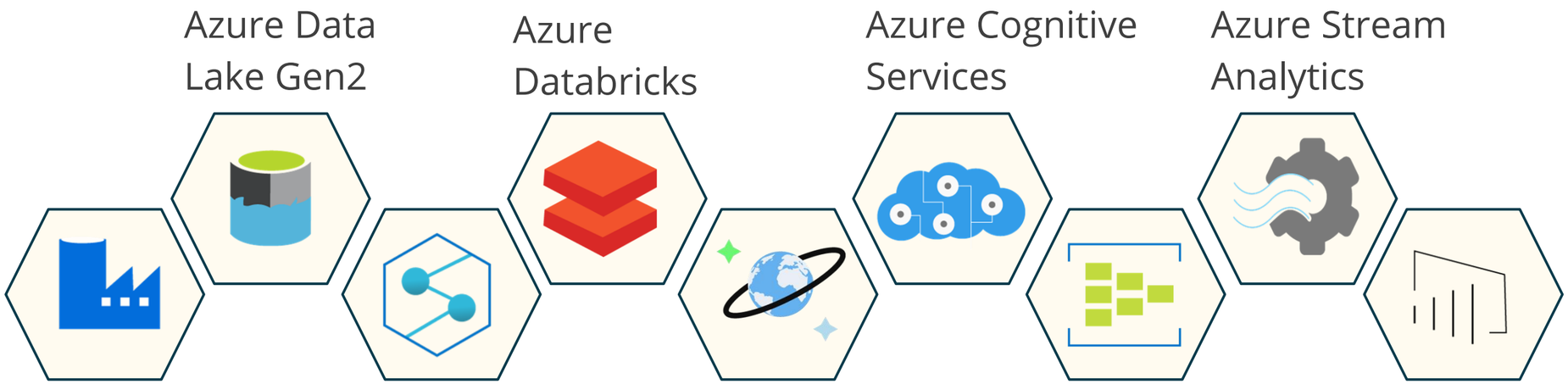
Provide an enterprise-wide data hub that includes a structured data warehouse and a semi-structured and unstructured data lake. This data hub becomes the data's single source of truth

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:



Recommend a Solution for Data Integration

Data Flows Using Azure Data Factory

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

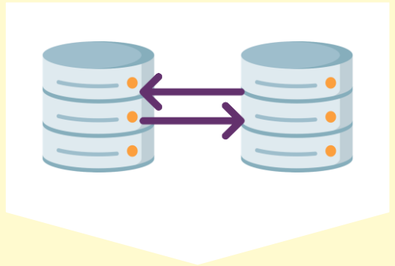
Ingest

- 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

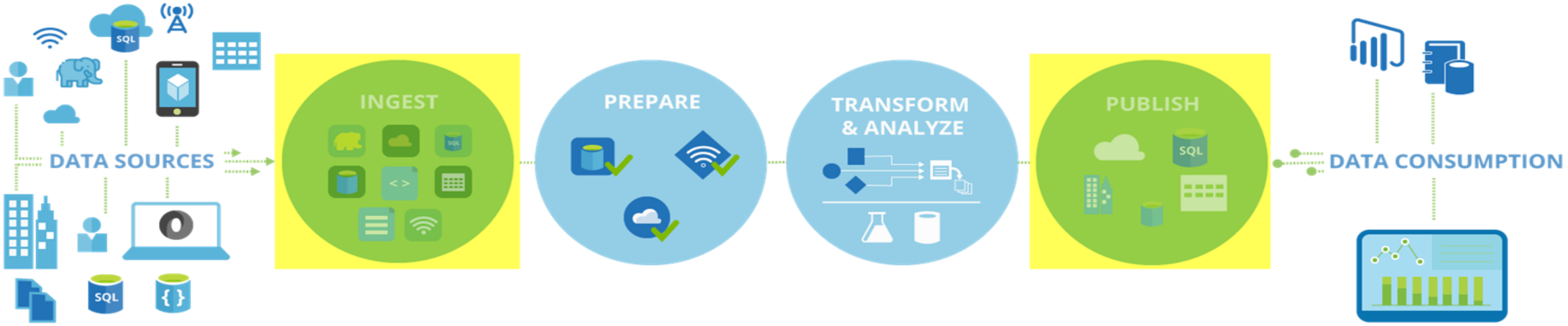
- Build and maintain operational schedules for the data pipelines
- Set clock and maintain event-based, tumbling chained windows

Monitor

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

How Data Factory Works

The *Copy* activity in Azure Data Factory copies data between on-premises and cloud data storage.



How Data Factory Works

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

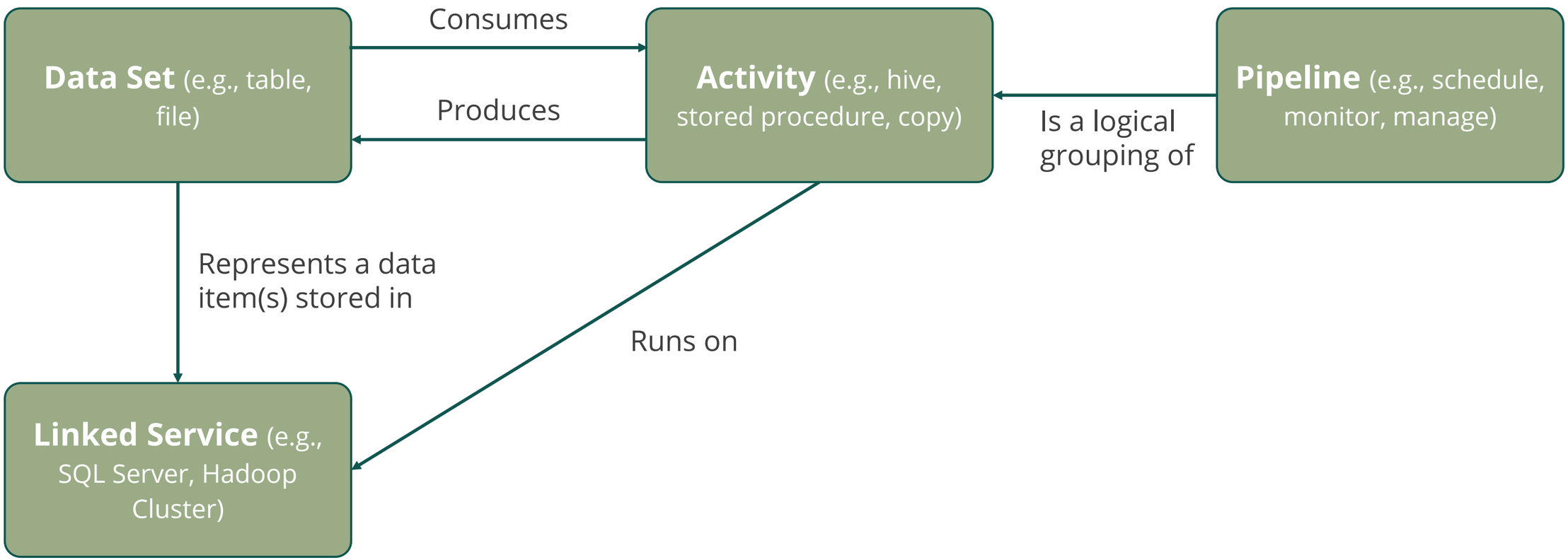
Connect and collect	Data Factory moves data from on-premises and cloud source data stores to a centralized data store in the cloud for analysis by using Copy activity in a data pipeline.
Transform and enrich	Data flows make it possible for data engineers to create data transformation graphs that run 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 Factory: Key Concepts

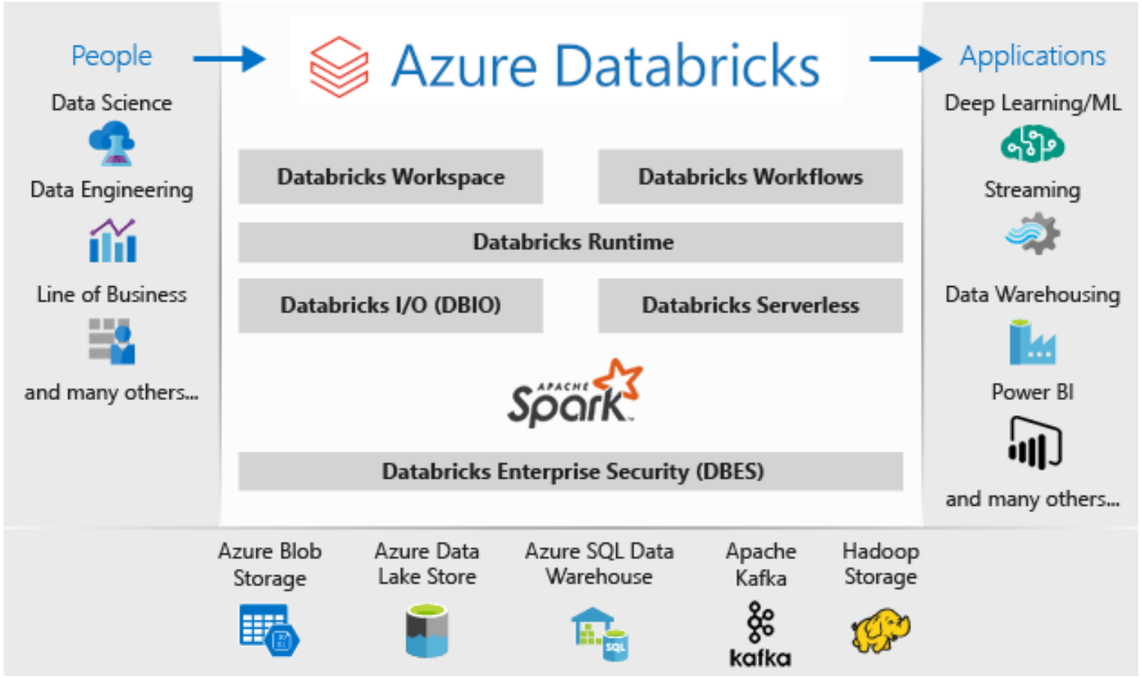
The key concept of components of data factory process:

Dataset	Datasets are data structures within data stores that simply refer to the data that users want to use as inputs or outputs in their operations.
Activity	A processing step in a pipeline is represented by activity. For instance, a user 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 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, an Azure Storage-linked service specifies a connection string to connect to an Azure Storage account.

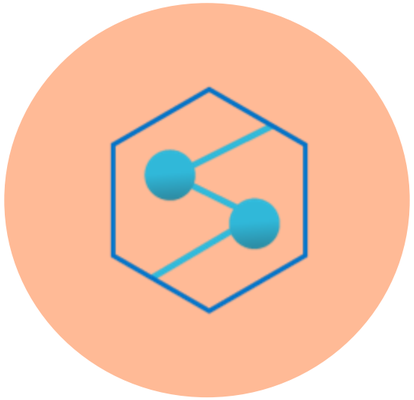
Integrate Data Factory and Databricks



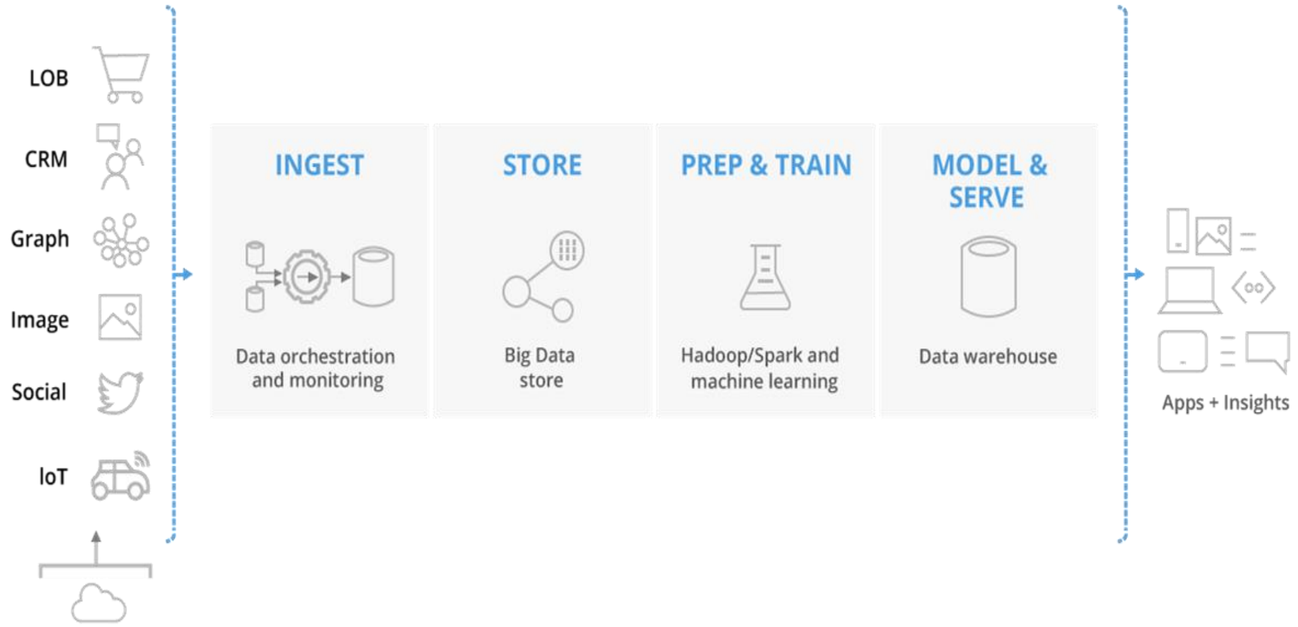
1. Create an Azure storage account
2. Create a Data Factory instance: Portal
3. Create a data workflow pipeline: Copy activity
4. Add a Databricks notebook to the pipeline
5. Analyze the data: train data

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

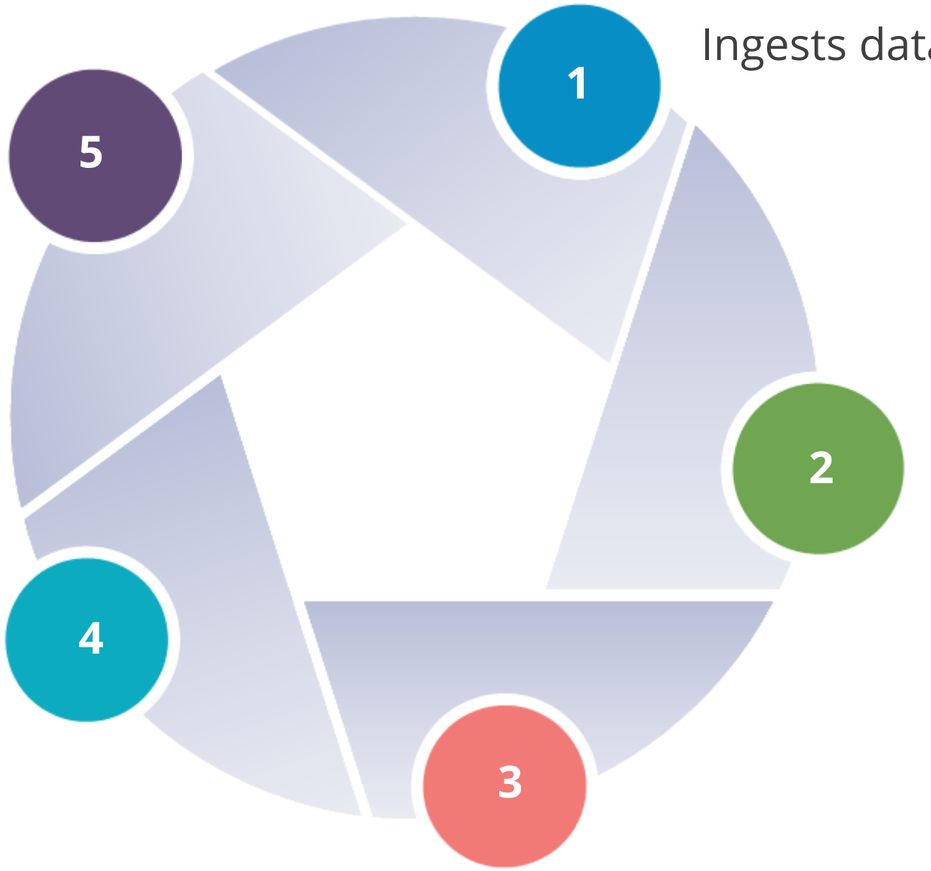


Azure Synapse has four components:

- Synapse SQL
- Spark
- Synapse Pipelines
- Studio

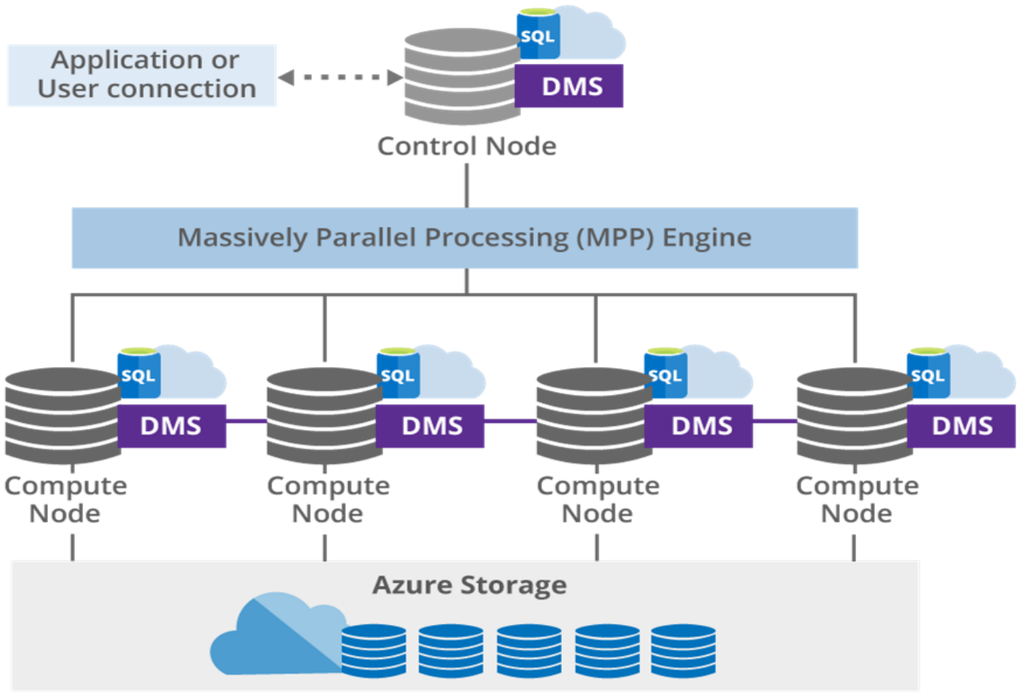
Data Flow

The process of data flow in Azure:



1. Ingests data
2. Prepares 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

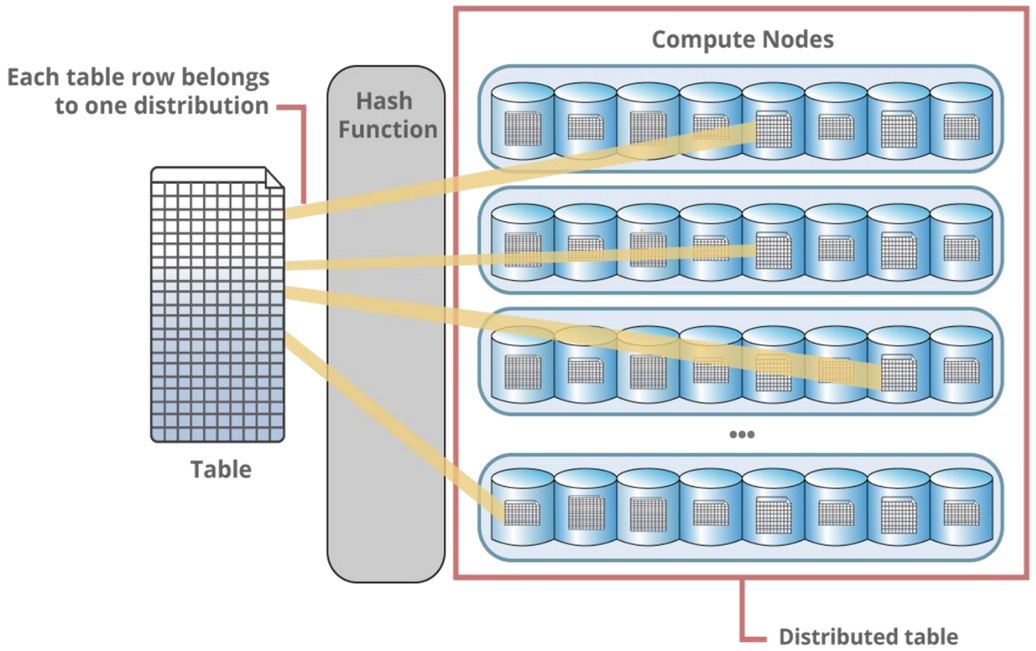
- Distributed processing

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.



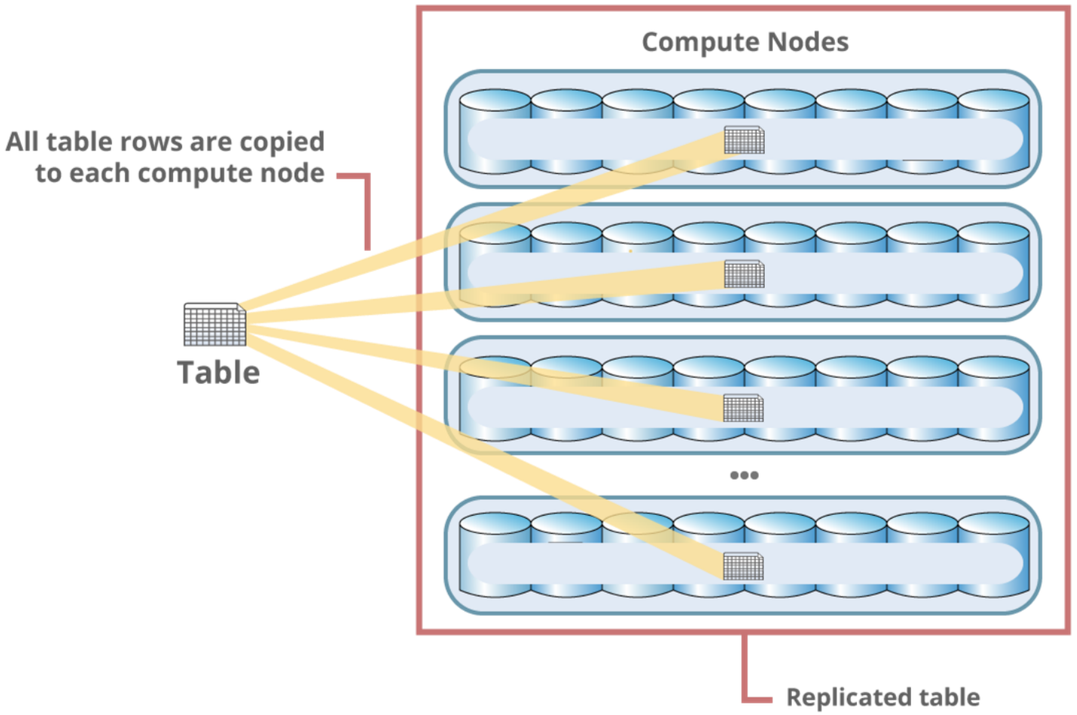
Hash-Distributed Tables

- 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

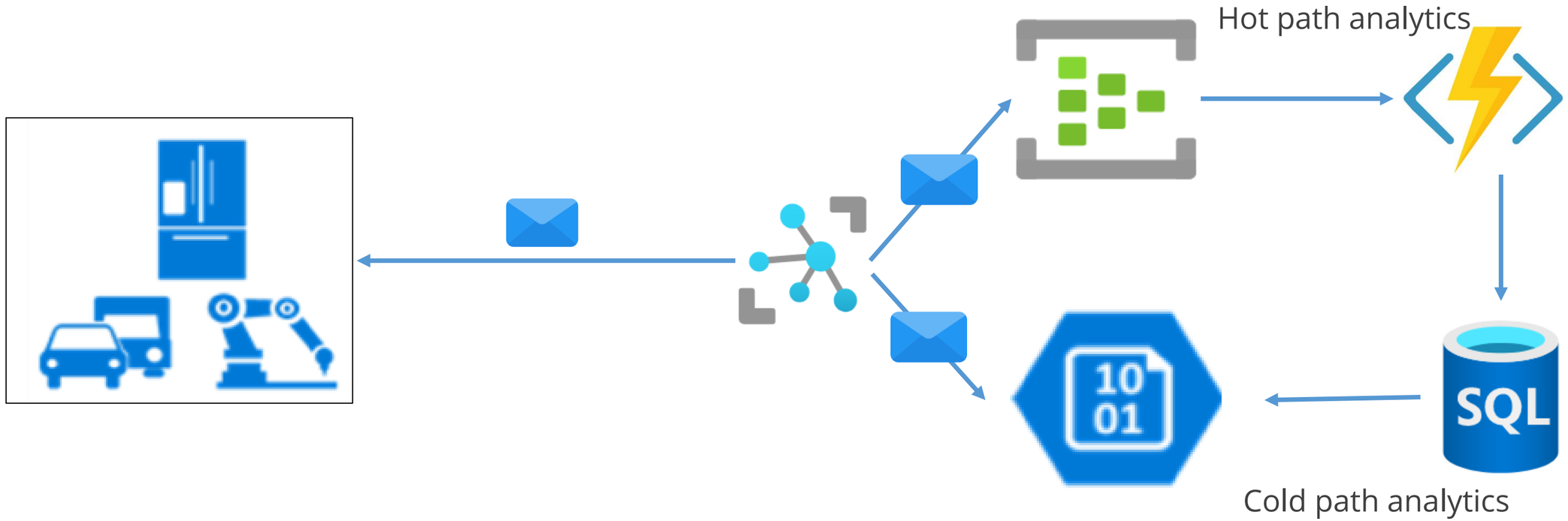


- A replicated table caches a full copy of the table on each compute node.
- It removes the need to transfer data among compute nodes.
- It is best utilized with small tables.

Design a Strategy for Hot, Warm, and Cold Data Path

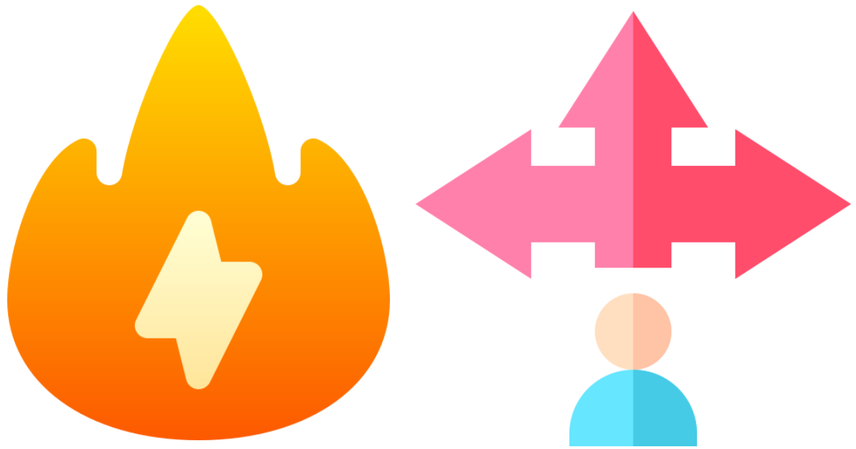
Working with Azure IoT Data

There are some patterns in how IoT use cases consume, analyze, and store data. The following diagram depicts an IoT pipeline in which IoT devices deliver data to Azure IoT Hub and then to Azure SQL Database:



Hot Path

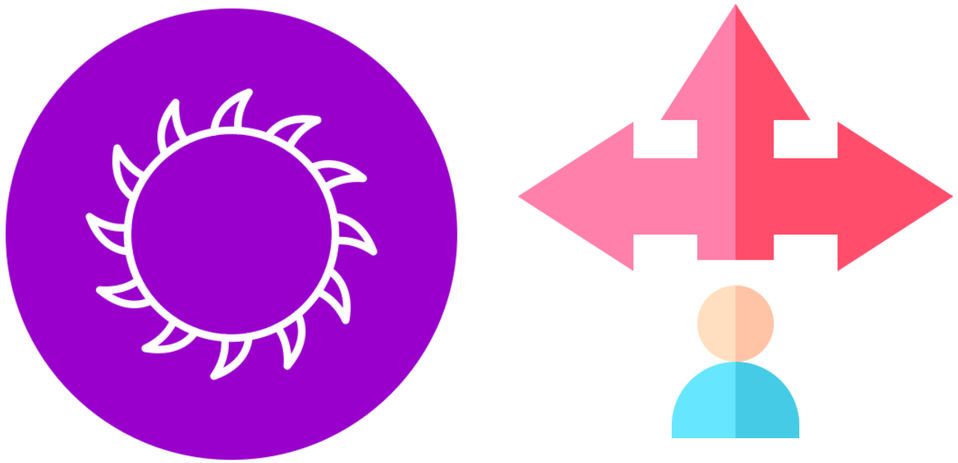
The hot path performs real-time data processing or display.



- Used for real-time alerting and streaming activities
- Uses an Azure Function App, Signal R, and a web app hosted for real-time alerts and data streaming
- Enables the ability to stream data through a WebSocket based connection through Azure Signal R and the web

Warm Path

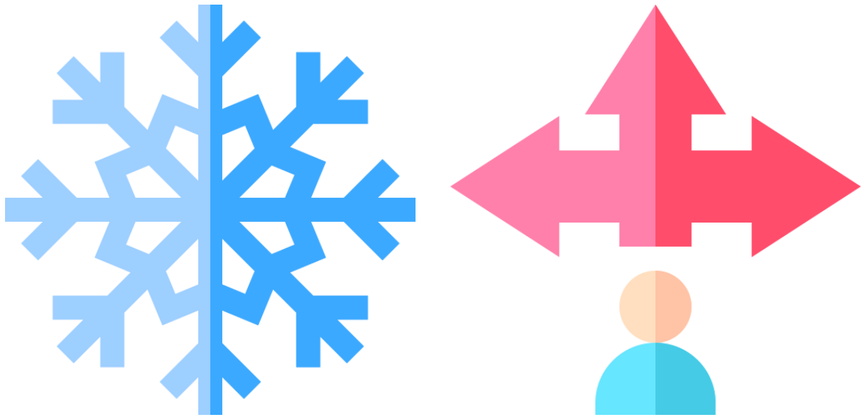
In warm path, small analytic and batch processing procedures are done on this data to save or show only the most current portion of the data.



A web app based on an Azure app service is utilized here since it can query and show the latest 24 hours of temperature data from Azure data explorer per device.

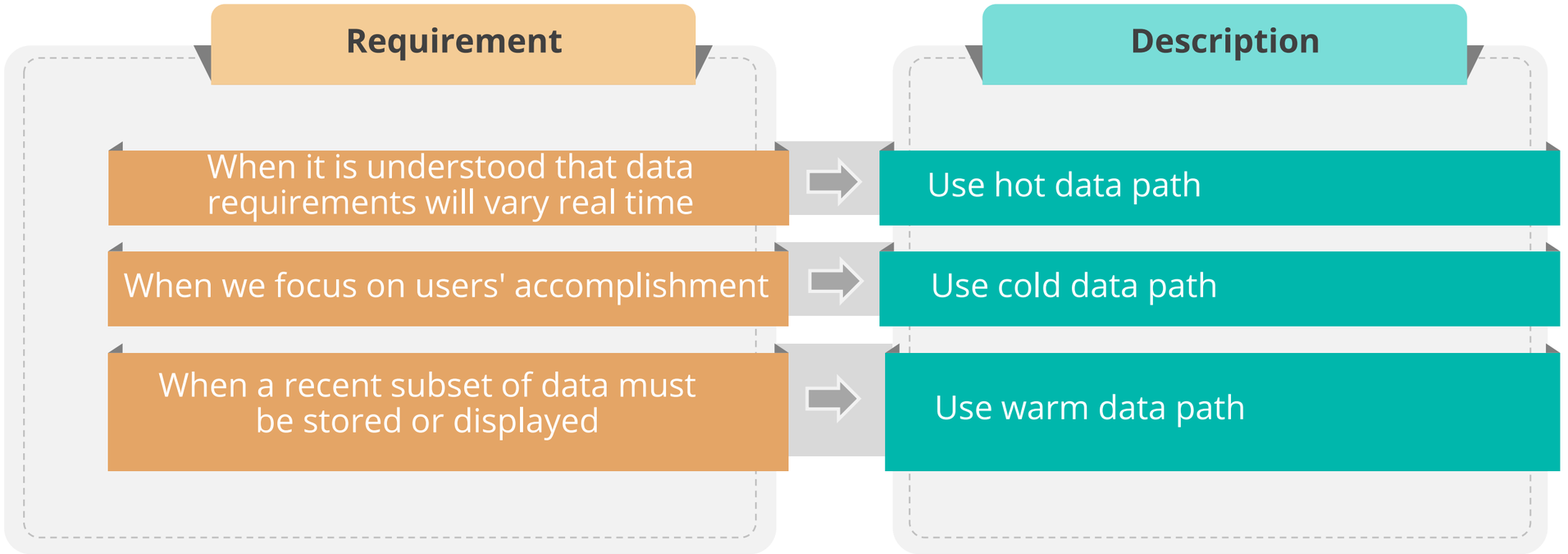
Cold Path

A cold path is used for long-term data storage. This data is subjected to time-consuming analytics and batch processing.



Azure data explorer is utilized here because it efficiently saves data for long periods, presently with a default of 100 years, and is an easy-to-use analytic engine built on top of the Kusto Query Language (KQL)

Scenarios to Use Hot, Warm, And Cold Data Path



Design Azure Stream Analytics Solution for Data Analysis

Azure Stream Analytics

Microsoft's Azure Stream Analytics is a fully managed, serverless real-time analytics engine.

Scenarios to Use Azure Stream Analytics

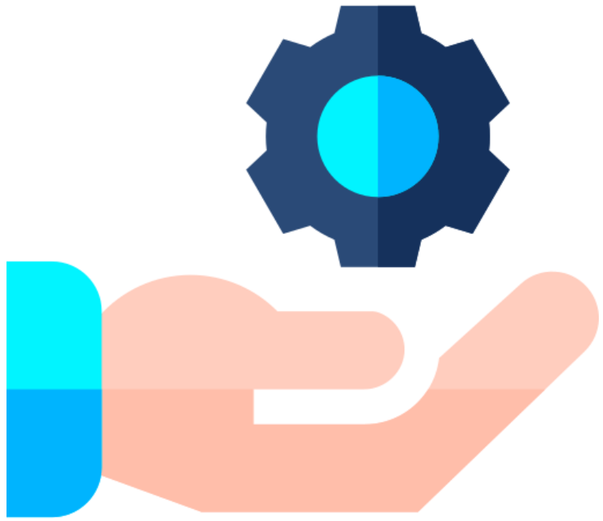
Azure stream analytics is made up of the following use cases:

- Dashboarding in real time using Power BI (monitoring purposes)
- Storing streaming data for later accessed by other cloud services
- Transforming and analyzing data in real-time trigger processes
- Sending alerts
- Making judgments in real time
- Working in machine learning limited application in more complex analytics
- Operating in Blob storage accounts

Azure Stream Analytics Example

- Send data to services like Azure Functions, Service Bus Topics, or Queues to trigger communications or custom actions further down the line
- Send data to a Power BI dashboard for real-time dashboarding
- Store data in other Azure storage services (for example, Azure Data Lake, Azure Synapse Analytics, and so on) to train a machine learning model or execute batch analytics based on historical data

Key Features of Azure Stream Analytics



- Ease of use
- Ease to adopt
- Fully managed
- Reliability
- Security
- Performance

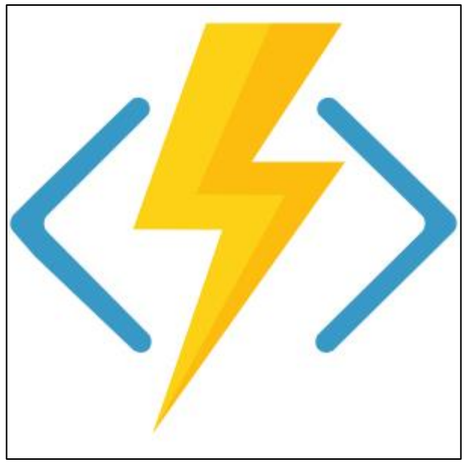
Limitations of Azure Stream Analytics

- It only works with SQL.
- The input data must be in AVRO, JSON, or CSV format.
- Blob storage can only be used to store static data.
- Only Azure services can be integrated.
- Users cannot benefit from dynamic reference data join functionality.
- There is no such thing as automated scaling (scale job in Azure Portal).

Streaming Analytics Competitors

Steaming Analytics has few competitors, namely:

Apache kafka streaming



Azure functions

Components Of Azure Stream Analytics Job

Three components define the Azure stream analytics task: streaming data input sources, transforming data using a SQL-like query, and transforming data using a SQL-like query.

Working of Azure Stream Analytics

Workflows

- An Azure Stream Analytics task is made up of three parts: an input, a query, and an output.
- It retrieves data from Azure event hubs, Azure IoT hubs, and Azure blob storage.
- The query is written in SQL and may be used to easily filter, sort, aggregate, and merge streaming data.
- Users can reference a workflow within another workflow.

High Level Approach by Data Analytics

Data can be ingested to Azure Streaming Platform from IoT devices, log files, Azure Blob, financial data such as credit card transactions.

Users can use a Stream Analytics query to perform analytics on the data, such as sorting, filtering, aggregations, and many more.

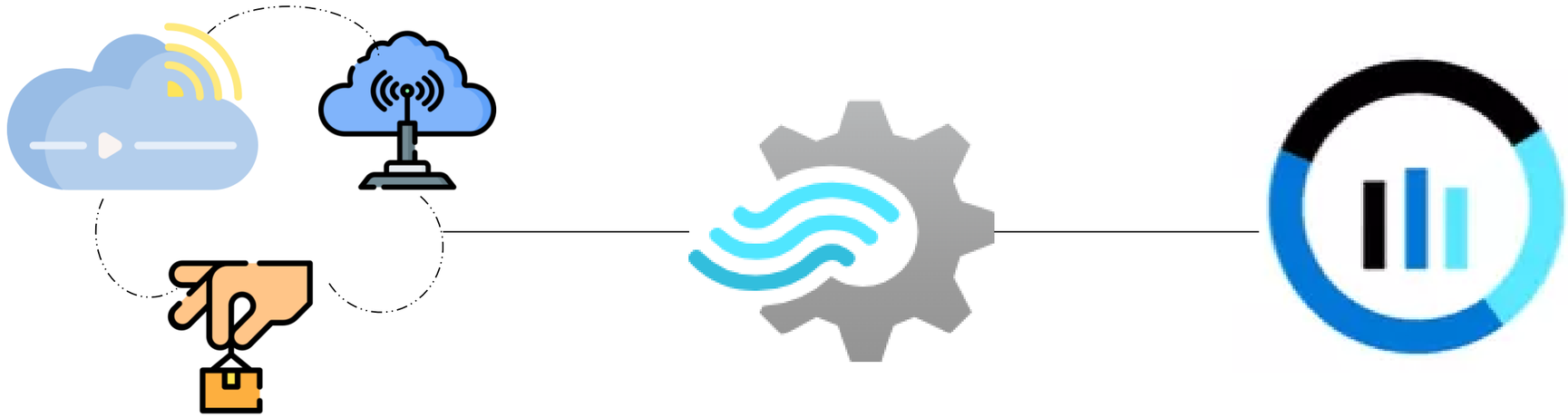
Downstream applications such as a data warehouse and reporting tools can use the data for additional analysis and issuing alerts and notifications.

Azure Stream Analytics Pricing Structure

Parameters	Standard	Dedicated
Resource type	Stream analytics job	Stream analytics cluster
Streaming unit	$0.11 per hour with a minimum of 1 SU	$0.11 per hour with a minimum of 36 SU*
Visual network support	No	Yes
C# user-defined functions	Limited to West Central US, North Europe, East US, West US, East US 2, and West Europe	All regions

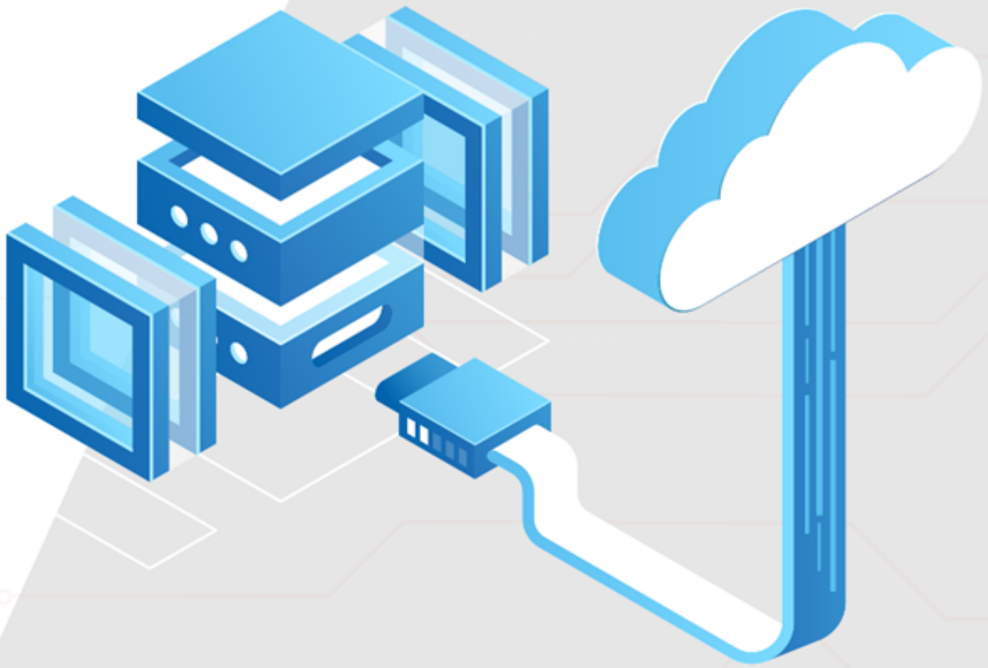
Azure Stream Analytics on IoT Edge

The Azure IoT Edge is a fully managed service based on the Azure IoT Hub, which enables businesses to install cloud service workloads and other third-party services on IoT Edge devices.



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, Relational, Semi-structured, Non-structure, and Streaming are the most important components
- 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