Data Warehous as a Service

conf.dr. Cristian KEVORCHIAN
Faculty of Mathematics and Infomatics
University of Bucharest

Data Warehouse

A data warehouse is a centralized repository of integrated data from one or more disparate sources

Data warehouses store current and historical data and are used for reporting and analysis of the data

To move data into a data warehouse, data is periodically extracted from various sources that contain important business information

Data warehouse architectures

Enterprise BI in Azure with Azure Synapse Analytics

Automated enterprise BI with Azure Synapse and Azure Data Factory

Solution lifecycle

The data warehouse can store historical data from multiple sources, representing a single source of truth

You can improve data quality by cleaning up data as it is imported into the data warehouse

Reporting tools don't compete with the transactional systems for query processing cycles

A data warehouse can consolidate data from different software

Data mining tools can find hidden patterns in the data using automatic methodologies

Data warehouses make it easier to provide secure access to authorized users, while restricting access to others

Data warehouses make it easier to create business intelligence solutions, such as OLAP cubes

Challenges

01

COMMITTING THE TIME REQUIRED TO PROPERLY MODEL YOUR BUSINESS CONCEPTS

02

PLANNING AND SETTING UP YOUR DATA ORCHESTRATION

03

MAINTAINING OR IMPROVING DATA QUALITY BY CLEANING THE DATA AS IT IS IMPORTED INTO THE WAREHOUSE

- Azure SQL Database
- SQL Server in a virtual machine
- Azure Synapse Analytics
- Apache Hive on HDInsight
- Interactive Query on HDInsight

Data warehousing in Azure

Data warehousing in Azure

- Azure SQL Data Warehouse Workload Patterns and Anti-Patterns
- Azure SQL Data Warehouse loading patterns and strategies
- Migrating data to Azure SQL Data Warehouse in practice
- Common ISV application patterns using Azure SQL Data Warehouse

- Managed service vs. Managing your servers infrastructure
- Working with extremely large data sets vs. highly complex, long-running queries
- For a large data set, use the data source structured or unstructured
- Separate your historical data from your current, operational data
- Integrate data from several sources, beyond your OLTP data store
- Multitenancy requirement
- Relational data store seems to be prefered
- Real-time reporting requirements

A Possible Feedback

In a big data architecture, there is often a need for an analytical data store that serves processed data in a structured format that can be queried using analytical tools

Analytical data stores that support querying of both hot-path and cold-path data are collectively referred to as the serving layer, or data serving storage

In the lambda architecture, the serving layer is subdivided into a speed serving layer, which stores data that has been processed incrementally, and a batch serving layer, which contains the batch-processed output

Architectural options for the analytical data store

- Key/value databases hold a single serialized object for each key value
- Document databases are key/value databases in which the values are documents
- Column-family databases are key/value data stores that structure data storage into collections of related columns called column families
- Graph databases store information as a collection of objects and relationships
- Telemetry and time series databases are an append-only collection of objects

Options when choosing an analytical data store

- Azure Synapse Analytics
- Azure Synapse Spark pools
- Azure Databricks
- Azure Data Explorer
- Azure SQL Database
- SQL Server in Azure VM
- HBase/Phoenix on HDInsight
- Hive LLAP on HDInsight
- Azure Analysis Services
- Azure Cosmos DB

Modern data warehouse

We present several solutions to modernize legacy data stores and explore big data tools and capabilities, without overextending current budgets and skillsets

These end-to-end Azure data warehousing solutions integrate easily with tools like Azure Machine Learning, Microsoft Power Platform, Microsoft Dynamics, and other Microsoft technologies

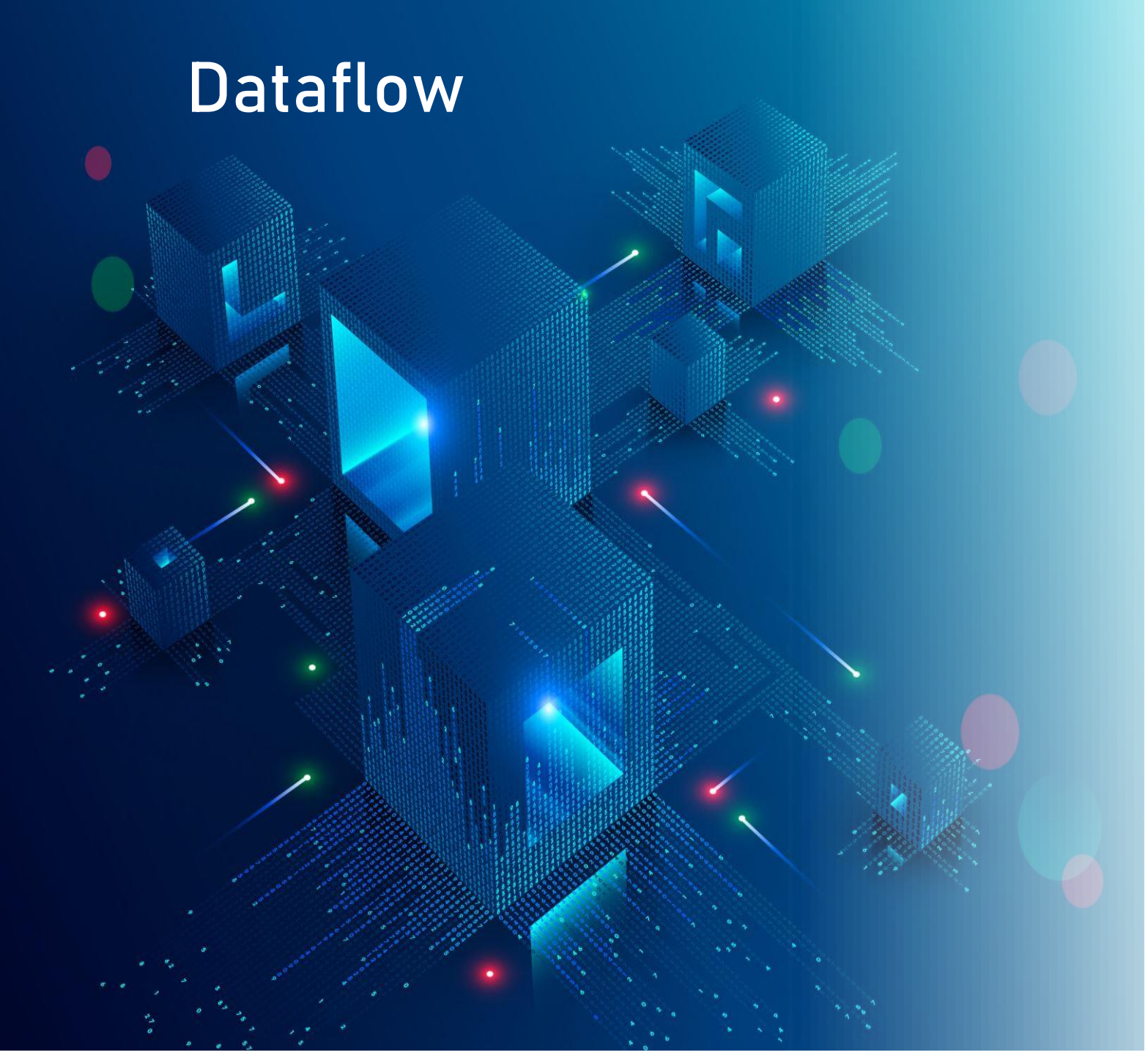
Architecture

- Unstructured data, like documents and graphics
- Semi-structured data, such as logs, CSVs, JSON, and XML files
- Structured relational data, including databases that use stored procedures for extract-transform-load/extract-load-transform activities

Dataflow

- Azure Synapse Analytics pipelines ingest the legacy data warehouses into Cloud Solution
  - The pipelines orchestrate the flow of migrated or partially refactored legacy databases and SSIS packages into Azure SQL Database
  - The pipelines can also pass unstructured, semi-structured, and structured data into Azure Data Lake Storage for centralized storage and analysis with other sources
- Microsoft Dynamics data sources can be used to build centralized BI dashboards on augmented datasets using Synapse Serverless analysis tools
- Real-time data from streaming sources can also enter the system via Azure Event Hubs
- The data can also enter the centralized Data Lake for further analysis, storage, and reporting

Dataflow



- Serverless analysis tools are available in the Azure Synapse Analytics workspace
- Serverless pools are ideal for
  - Ad hoc data science explorations in T-SQL format
  - Early prototyping for data warehouse entities
  - Defining views that consumers can use, for example in Power BI, for scenarios that can tolerate performance lag

Components

- Azure Synapse Analytics is an analytics service that combines data integration, enterprise data warehousing, and big data analytics
  - An Azure Synapse Workspace promotes collaboration between data engineers, data scientists, data analysts, and business intelligence professionals
  - Azure Synapse pipelines orchestrate and ingest data into SQL Database and Data Lake Storage Gen
  - Azure Synapse serverless SQL pools analyze unstructured and semi-structured data in Data Lake Storage Gen2 on demand
  - Azure Synapse serverless Apache Spark pools do code-first explorations in Data Lake Storage Gen2 with Spark languages like Spark SQL, pySpark, and Scala
- Azure SQL Database is an intelligent, scalable, relational database service built for the cloud

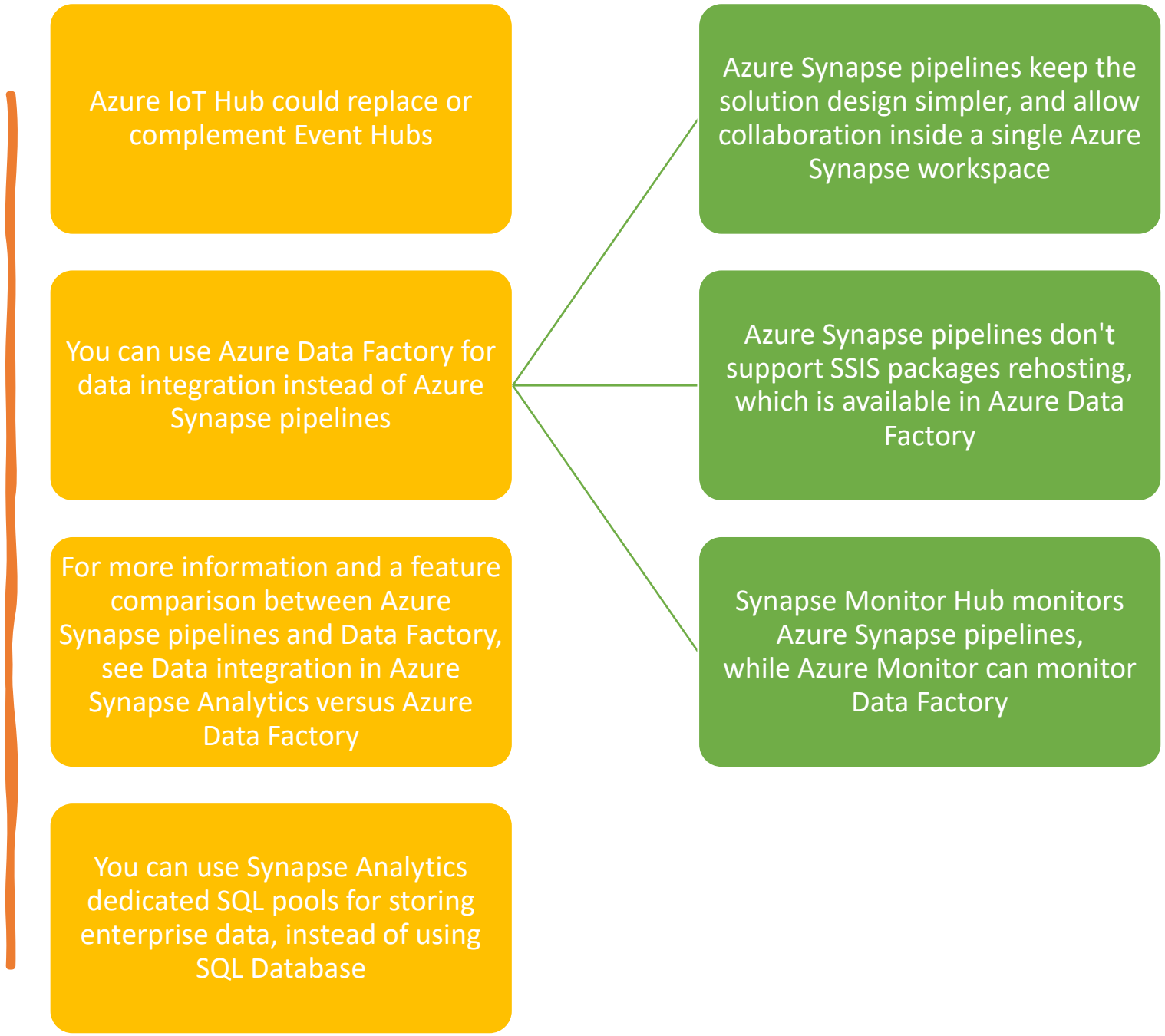
Components

Azure Event Hubs is a real-time data streaming platform and event ingestion service

Azure Stream Analytics is a real-time, serverless analytics service for streaming data

Azure Machine Learning is a toolset for data science model development and lifecycle management

Alternatives



Scenario details

Small and medium businesses face a choice when modernizing their on-premises data warehouses for the cloud

They can adopt big data tools for future extensibility, or keep traditional, SQL-based solutions for cost efficiency, ease of maintenance, and smooth transition

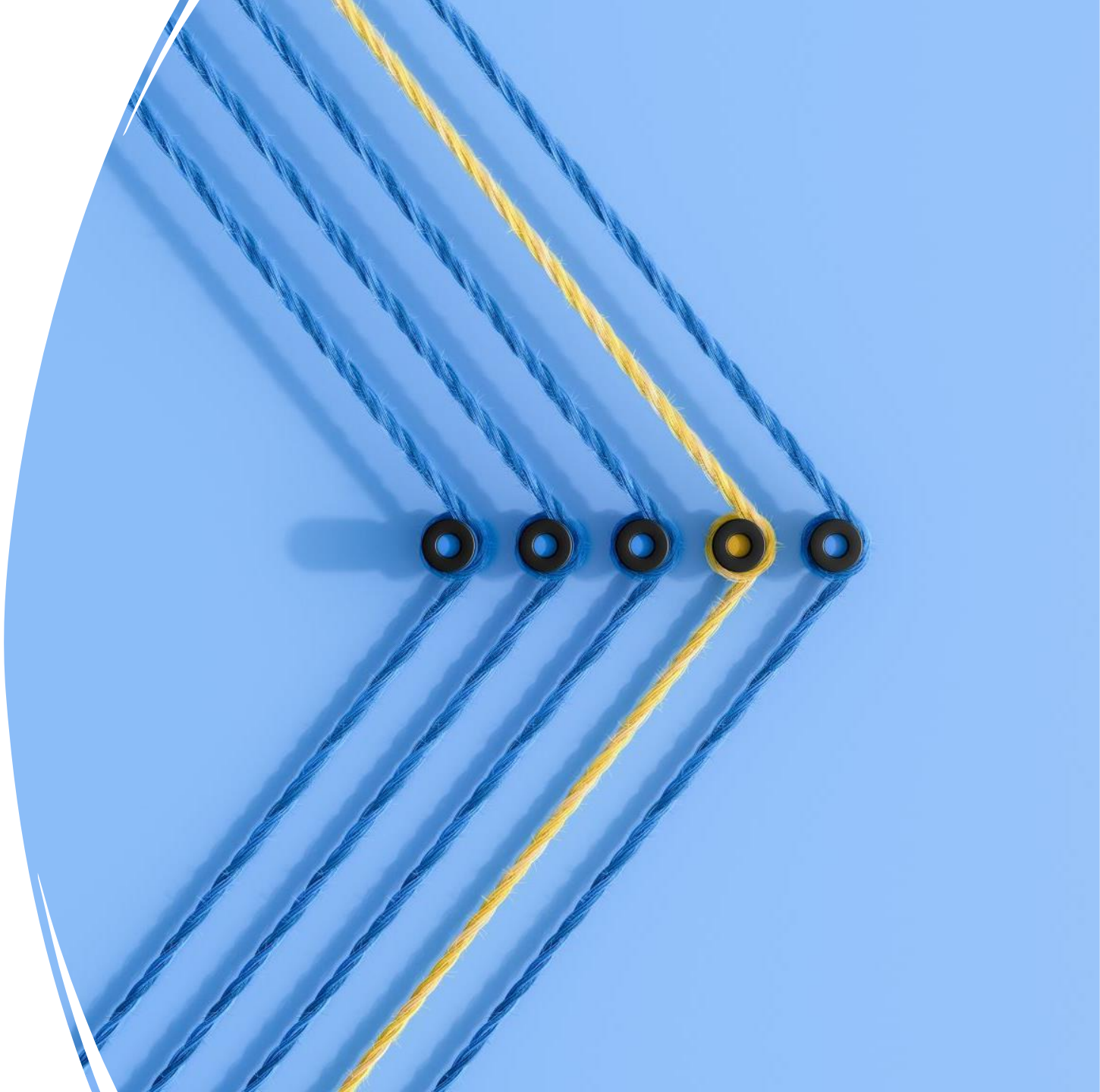
These end-to-end Azure data warehousing solutions integrate easily with Azure and Microsoft services and tools like Azure Machine Learning, Microsoft Power Platform, and Microsoft Dynamics

Potential use cases

- Migrating a traditional, on-premises relational data warehouse that's smaller than 1 TB and extensively uses SQL Server Integration Services packages to orchestrate stored procedures
- Meshing existing Dynamics or Power Platform Dataverse data with batched and real-time Azure Data Lake sources
- Using innovative techniques to interact with centralized Data Lake Storage Gen2 data
- Setting up eCommerce companies to adopt a data warehouse to optimize their operations
- Greenfield deployment of data warehouses that are estimated to be > 1 TB within one year
- Migrating on-premises data warehouses that are > 1 TB or projected to grow to that size within a year

Considerations

- These considerations implement the pillars of the Azure Well-Architected Framework, which is a set of guiding tenets that can be used to improve the quality of a workload



Availability

SQL Database is a PaaS service that can meet your high availability and disaster recovery requirements

Be sure to pick the SKU that meets your requirements

For guidance, see High availability for Azure SQL Database

Cost optimization

SQL Database bases costs on the selected Compute and Service tiers, and the number of vCores and Database Transaction Units

Data Lake Storage Gen2 pricing depends on the amount of data you store and how often you use the data

Azure Synapse pipelines base costs on the number of data pipeline activities, integration runtime hours, data flow cluster size, and execution and operation charges

Azure Synapse Spark pool bases pricing on node size, number of instances, and uptime

Azure Synapse serverless SQL pool bases pricing on TBs of data processed

Event Hubs bills based on tier, throughput units provisioned, and ingress traffic received

Stream Analytics bases costs on the number of provisioned streaming units

MDX example

- SELECT NON EMPTY { [Measures].[Internet Total Sales] } ON COLUMNS, NON EMPTY { ([Geography].[City].[City].ALLMEMBERS) } DIMENSION PROPERTIES MEMBER_CAPTION, MEMBER_UNIQUE_NAME ON ROWS FROM [Model] CELL PROPERTIES VALUE, BACK_COLOR, FORE_COLOR, FORMATTED_VALUE, FORMAT_STRING, FONT_NAME, FONT_SIZE, FONT_FLAGS