AZ-305T00A
Designing Microsoft Azure Infrastructure Solutions

Design a data integration solution

https://learn.microsoft.com/training/modules/design-data-integration/

Learning Objectives

- Design a data integration solution with Azure Data Factory
- Design a data integration solution with Azure Data Lake
- Design a data integration and analytics solution with Azure Databricks
- Design a data integration and analytics solution with Azure Synapse Analytics
- Design Azure Stream Analytics solution for Data Analysis
- Design a strategy for hot/warm/cold data path
- Case study
- Learning recap

AZ-305: Design Data Storage Solutions (20-25%)

Design Data Integration

- Recommend a solution for data integration
- Recommend a solution for data analysis

Design a data integration solution with Azure Data Factory

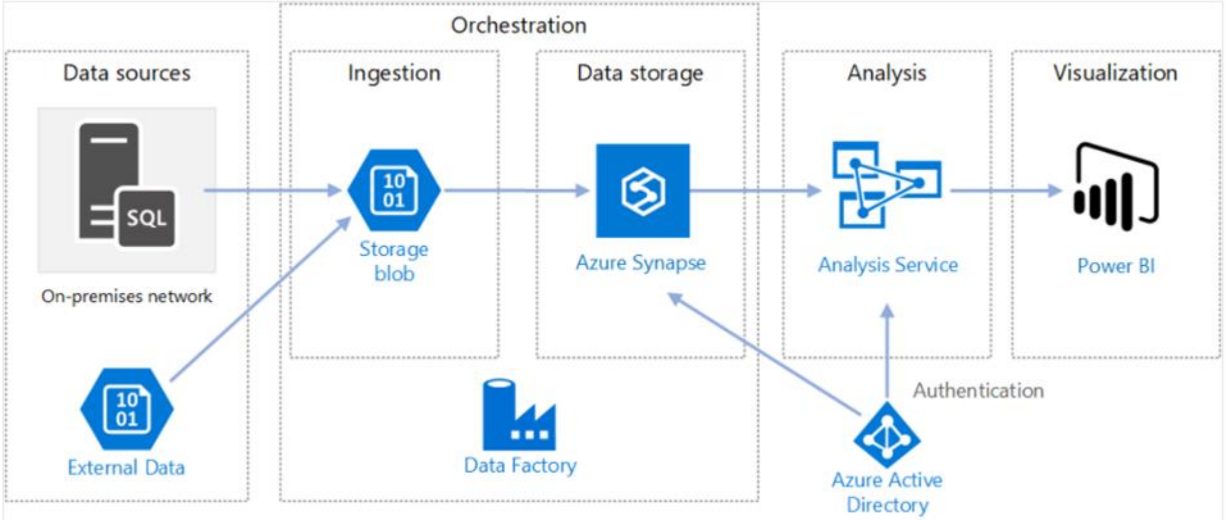
Data-driven workflows

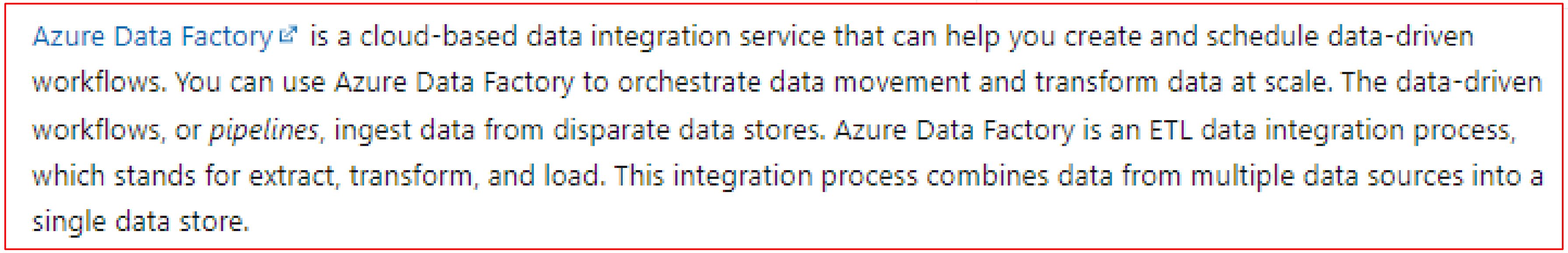
ETL (extract, transform, and load) and ELT (extract, load, and transform)

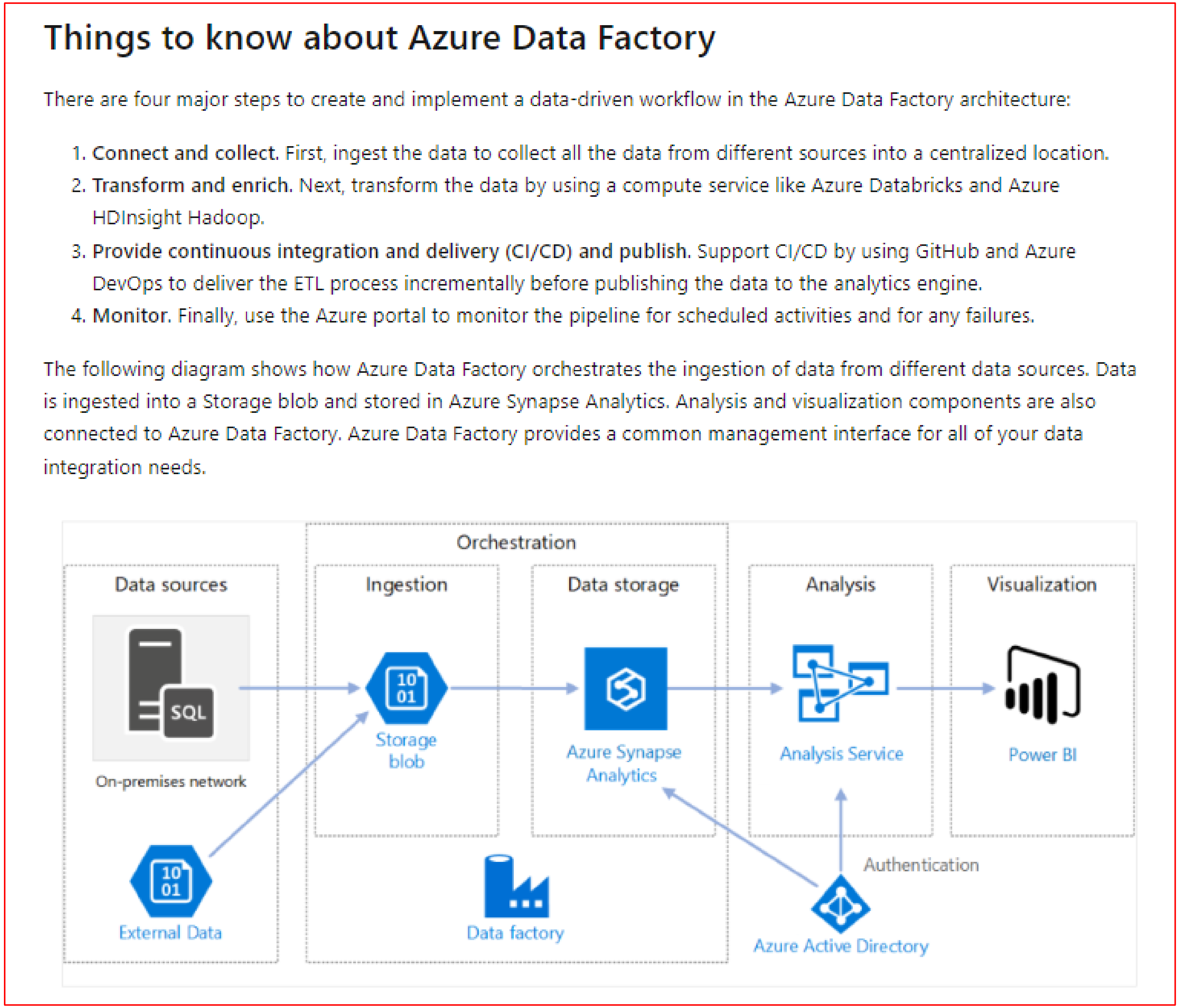
Azure Data Factory is a cloud-based ETL and data integration service that can help you create and schedule data-driven workflows (called pipelines) that can ingest data from disparate data stores.

You can use Azure Data Factory to:

1. Orchestrate data movement.
2. Transform data at scale.




Azure Data Factory is a cloud-based data integration service that can help you create and schedule data-driven workflows. You can use Azure Data Factory to orchestrate data movement and transform data at scale. The data-driven workflows, or *pipelines*, ingest data from disparate data stores. Azure Data Factory is an ETL data integration process, which stands for extract, transform, and load. This integration process combines data from multiple data sources into a single data store.

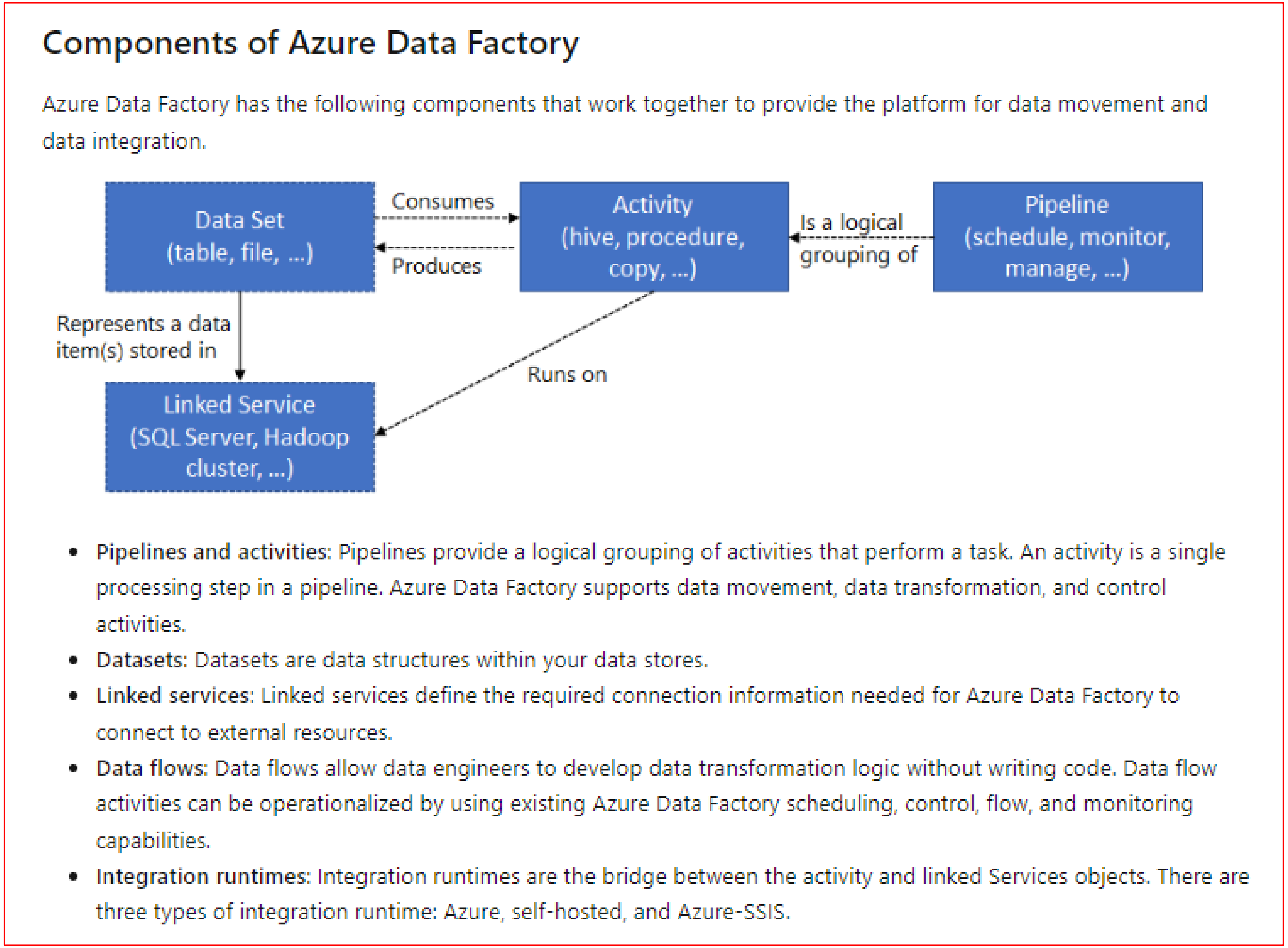


Things to know about Azure Data Factory

There are four major steps to create and implement a data-driven workflow in the Azure Data Factory architecture:

1. **Connect and collect.** First, ingest the data to collect all the data from different sources into a centralized location.
2. **Transform and enrich.** Next, transform the data by using a compute service like Azure Databricks and Azure HDInsight Hadoop.
3. **Provide continuous integration and delivery (CI/CD) and publish.** Support CI/CD by using GitHub and Azure DevOps to deliver the ETL process incrementally before publishing the data to the analytics engine.
4. **Monitor.** Finally, use the Azure portal to monitor the pipeline for scheduled activities and for any failures.

The following diagram shows how Azure Data Factory orchestrates the ingestion of data from different data sources. Data is ingested into a Storage blob and stored in Azure Synapse Analytics. Analysis and visualization components are also connected to Azure Data Factory. Azure Data Factory provides a common management interface for all of your data integration needs.

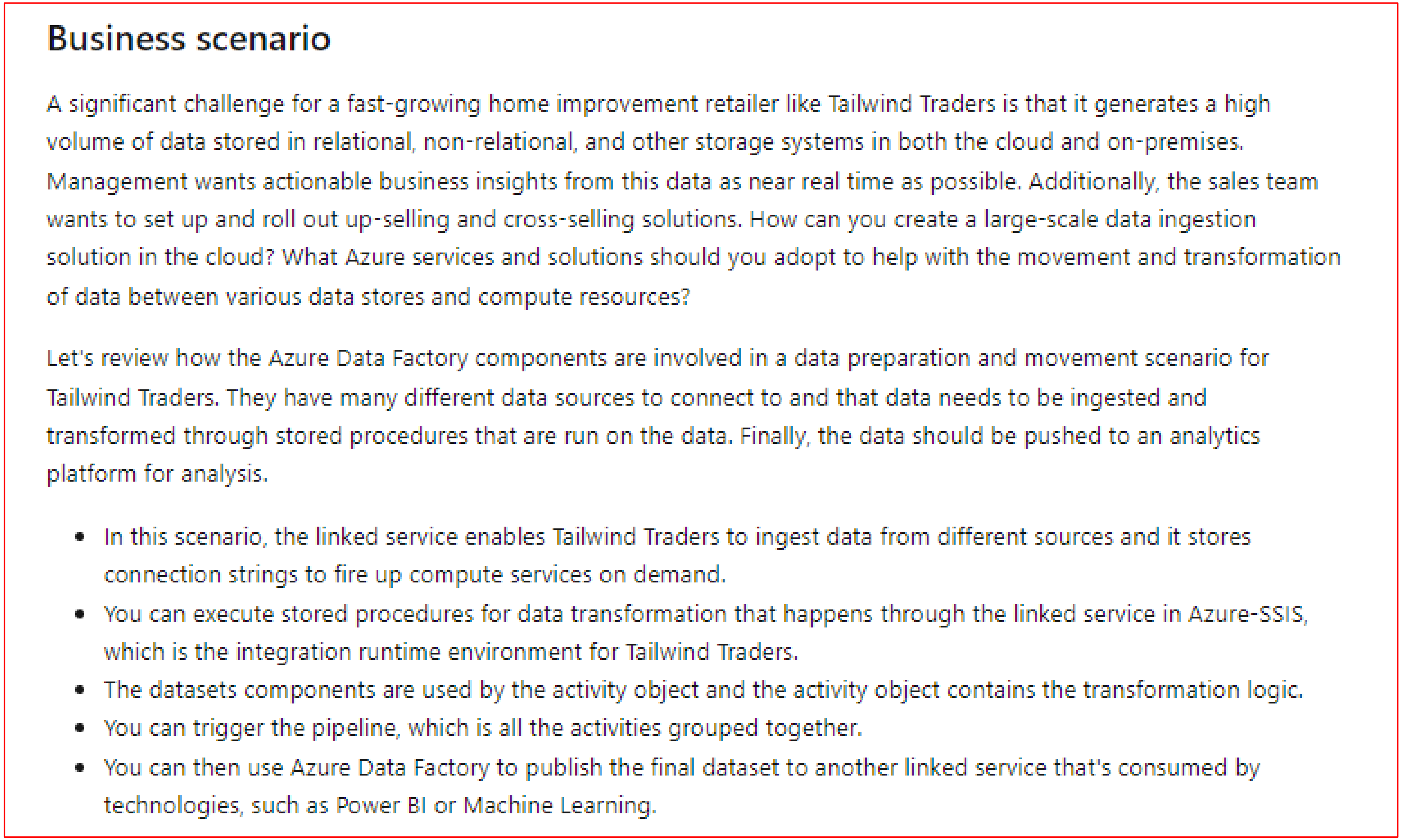


Components of Azure Data Factory

Azure Data Factory has the following components that work together to provide the platform for data movement and data integration.

Data Set
(table, file, ...)
Consumes
Produces
Activity
(hive, procedure, copy, ...)
Is a logical grouping of
Pipeline
(schedule, monitor, manage, ...)
Represents a data item(s) stored in
Runs on
Linked Service
(SQL Server, Hadoop cluster, ...)

- **Pipelines and activities:** Pipelines provide a logical grouping of activities that perform a task. An activity is a single processing step in a pipeline. Azure Data Factory supports data movement, data transformation, and control activities.
- **Datasets:** Datasets are data structures within your data stores.
- **Linked services:** Linked services define the required connection information needed for Azure Data Factory to connect to external resources.
- **Data flows:** Data flows allow data engineers to develop data transformation logic without writing code. Data flow activities can be operationalized by using existing Azure Data Factory scheduling, control, flow, and monitoring capabilities.
- **Integration runtimes:** Integration runtimes are the bridge between the activity and linked Services objects. There are three types of integration runtime: Azure, self-hosted, and Azure-SSIS.

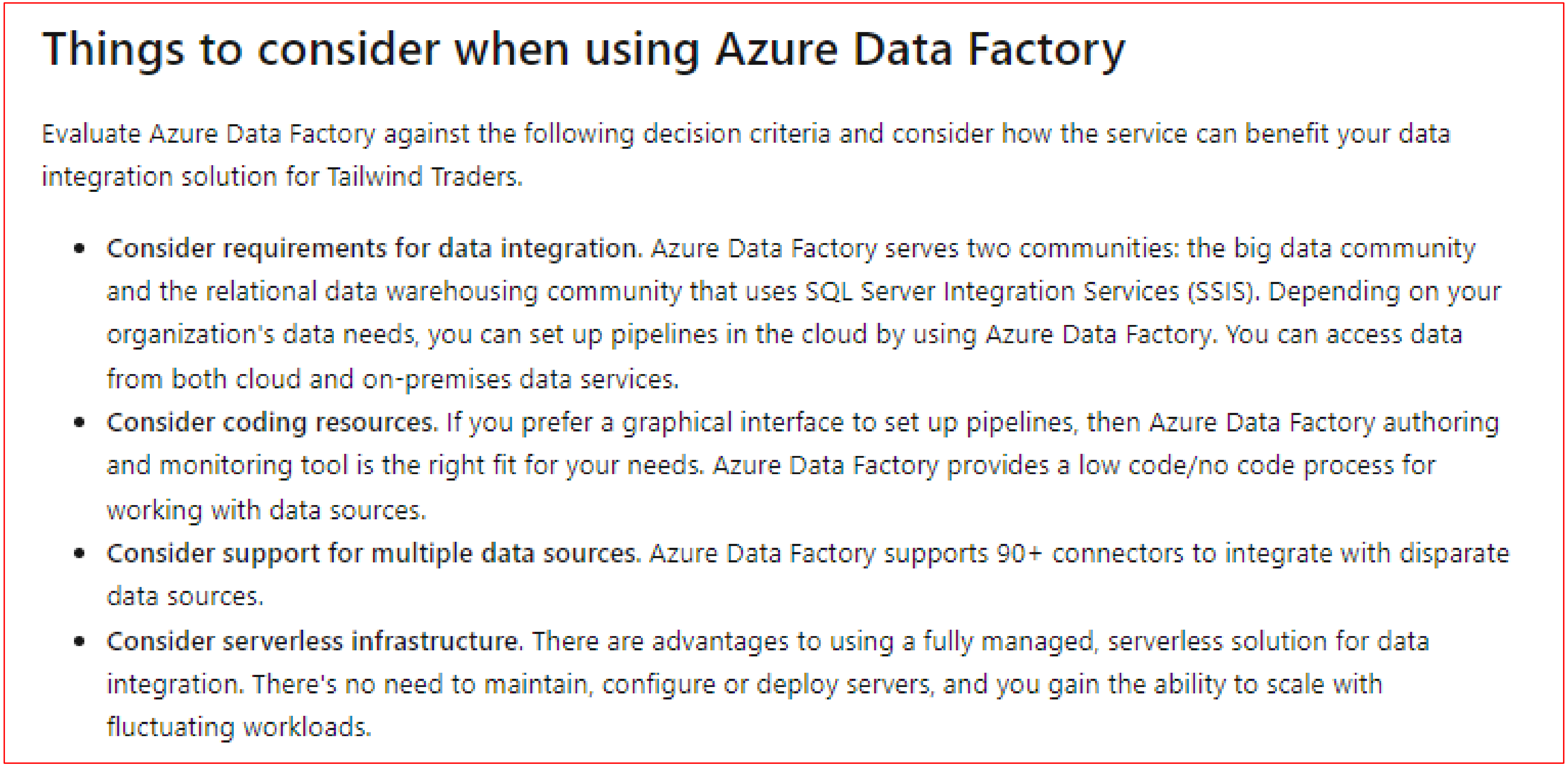


Business scenario

A significant challenge for a fast-growing home improvement retailer like Tailwind Traders is that it generates a high volume of data stored in relational, non-relational, and other storage systems in both the cloud and on-premises. Management wants actionable business insights from this data as near real time as possible. Additionally, the sales team wants to set up and roll out up-selling and cross-selling solutions. How can you create a large-scale data ingestion solution in the cloud? What Azure services and solutions should you adopt to help with the movement and transformation of data between various data stores and compute resources?

Let's review how the Azure Data Factory components are involved in a data preparation and movement scenario for Tailwind Traders. They have many different data sources to connect to and that data needs to be ingested and transformed through stored procedures that are run on the data. Finally, the data should be pushed to an analytics platform for analysis.

- In this scenario, the linked service enables Tailwind Traders to ingest data from different sources and it stores connection strings to fire up compute services on demand.
- You can execute stored procedures for data transformation that happens through the linked service in Azure-SSIS, which is the integration runtime environment for Tailwind Traders.
- The datasets components are used by the activity object and the activity object contains the transformation logic.
- You can trigger the pipeline, which is all the activities grouped together.
- You can then use Azure Data Factory to publish the final dataset to another linked service that's consumed by technologies, such as Power BI or Machine Learning.



Things to consider when using Azure Data Factory

Evaluate Azure Data Factory against the following decision criteria and consider how the service can benefit your data integration solution for Tailwind Traders.

- **Consider requirements for data integration.** Azure Data Factory serves two communities: the big data community and the relational data warehousing community that uses SQL Server Integration Services (SSIS). Depending on your organization's data needs, you can set up pipelines in the cloud by using Azure Data Factory. You can access data from both cloud and on-premises data services.
- **Consider coding resources.** If you prefer a graphical interface to set up pipelines, then Azure Data Factory authoring and monitoring tool is the right fit for your needs. Azure Data Factory provides a low code/no code process for working with data sources.
- **Consider support for multiple data sources.** Azure Data Factory supports 90+ connectors to integrate with disparate data sources.
- **Consider serverless infrastructure.** There are advantages to using a fully managed, serverless solution for data integration. There's no need to maintain, configure or deploy servers, and you gain the ability to scale with fluctuating workloads.

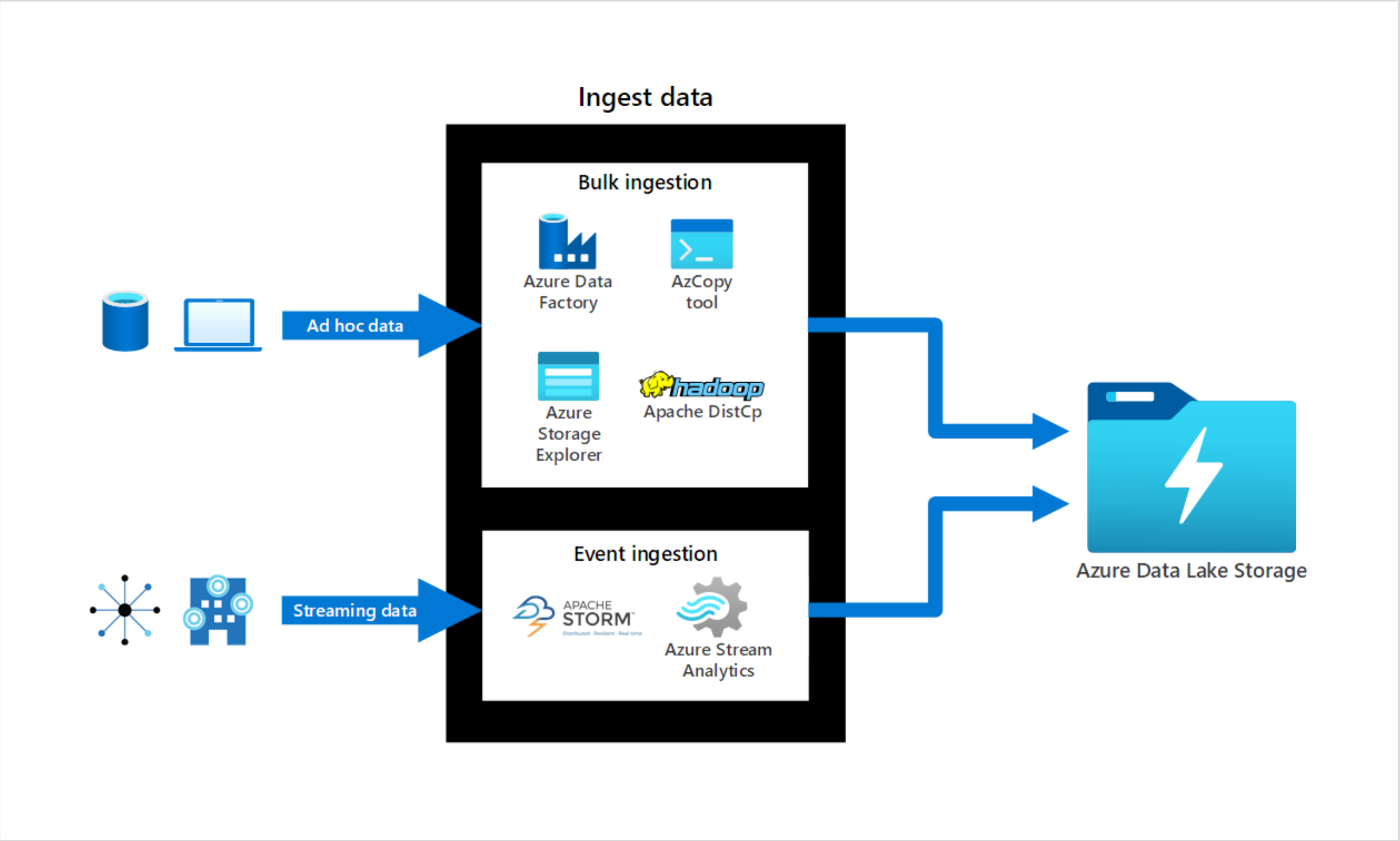
Design a data integration solution with Azure Data Lake

Azure Data Lake

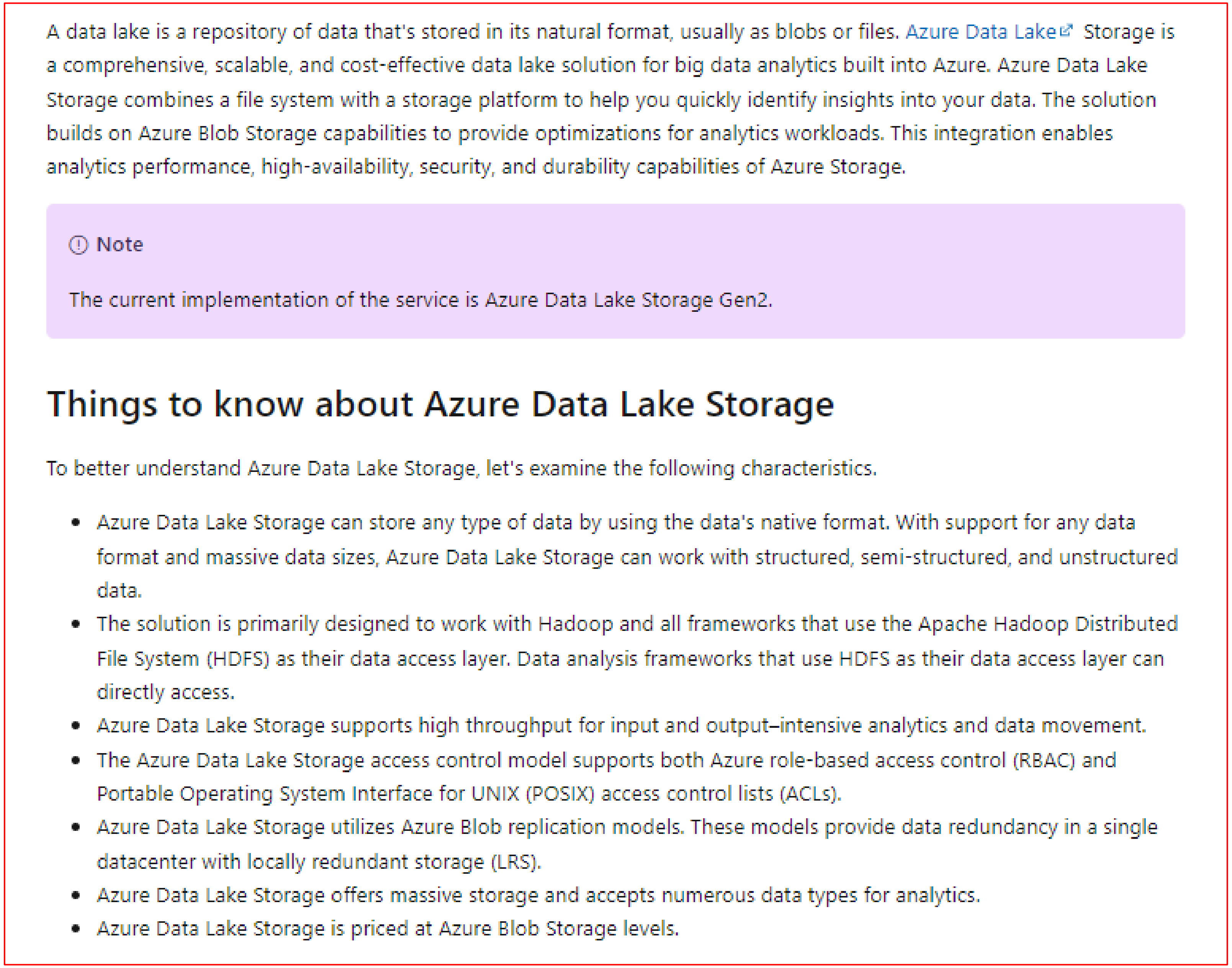
Azure Data Lake Storage is a comprehensive, scalable, and cost-effective data lake solution for big data analytics built into Azure.

Use Azure Data Lake when you need:

- a data repository on the cloud for managing large volumes of data
- To manage a diverse collection of data types such as JSON files, CSV, log files or other diverse formats
- Real-time data ingestion and storage



Design a data integration solution with Azure Data Lake



A data lake is a repository of data that's stored in its natural format, usually as blobs or files. Azure Data Lake Storage is a comprehensive, scalable, and cost-effective data lake solution for big data analytics built into Azure. Azure Data Lake Storage combines a file system with a storage platform to help you quickly identify insights into your data. The solution builds on Azure Blob Storage capabilities to provide optimizations for analytics workloads. This integration enables analytics performance, high-availability, security, and durability capabilities of Azure Storage.

ⓘ Note

The current implementation of the service is Azure Data Lake Storage Gen2.

Things to know about Azure Data Lake Storage

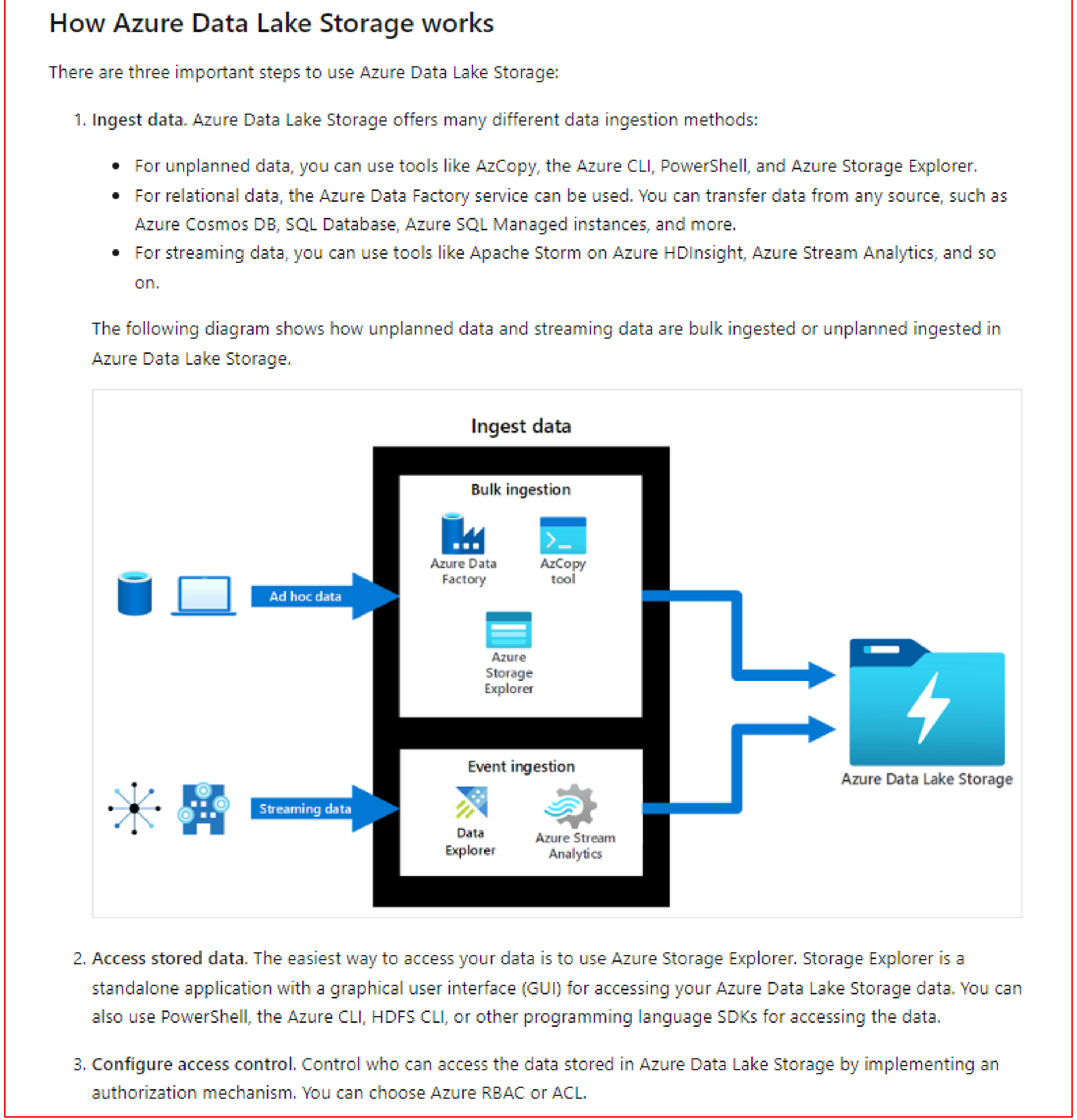
To better understand Azure Data Lake Storage, let's examine the following characteristics.

- Azure Data Lake Storage can store any type of data by using the data's native format. With support for any data format and massive data sizes, Azure Data Lake Storage can work with structured, semi-structured, and unstructured data.
- The solution is primarily designed to work with Hadoop and all frameworks that use the Apache Hadoop Distributed File System (HDFS) as their data access layer. Data analysis frameworks that use HDFS as their data access layer can directly access.
- Azure Data Lake Storage supports high throughput for input and output–intensive analytics and data movement.
- The Azure Data Lake Storage access control model supports both Azure role-based access control (RBAC) and Portable Operating System Interface for UNIX (POSIX) access control lists (ACLs).
- Azure Data Lake Storage utilizes Azure Blob replication models. These models provide data redundancy in a single datacenter with locally redundant storage (LRS).
- Azure Data Lake Storage offers massive storage and accepts numerous data types for analytics.
- Azure Data Lake Storage is priced at Azure Blob Storage levels.

Compare Azure Data Lake to Azure Blob storage

Criteria	Azure Data Lake	Azure Blob Storage
Data type	Good for storing large volumes of text data	Good for storing unstructured non-text-based data such as photos, videos, backup etc.
Namespace support	Supports hierarchical namespaces	Supports flat namespaces
Hadoop compatibility	Optimized for big data, like Hadoop	Is not Hadoop compatible
Security	Access Control Lists (ACLs), shared keys, SAS and RBAC	Shared keys, SAS, and RBAC

Blob Storage feature support in Azure storage accounts | Microsoft Docs



How Azure Data Lake Storage works

There are three important steps to use Azure Data Lake Storage:

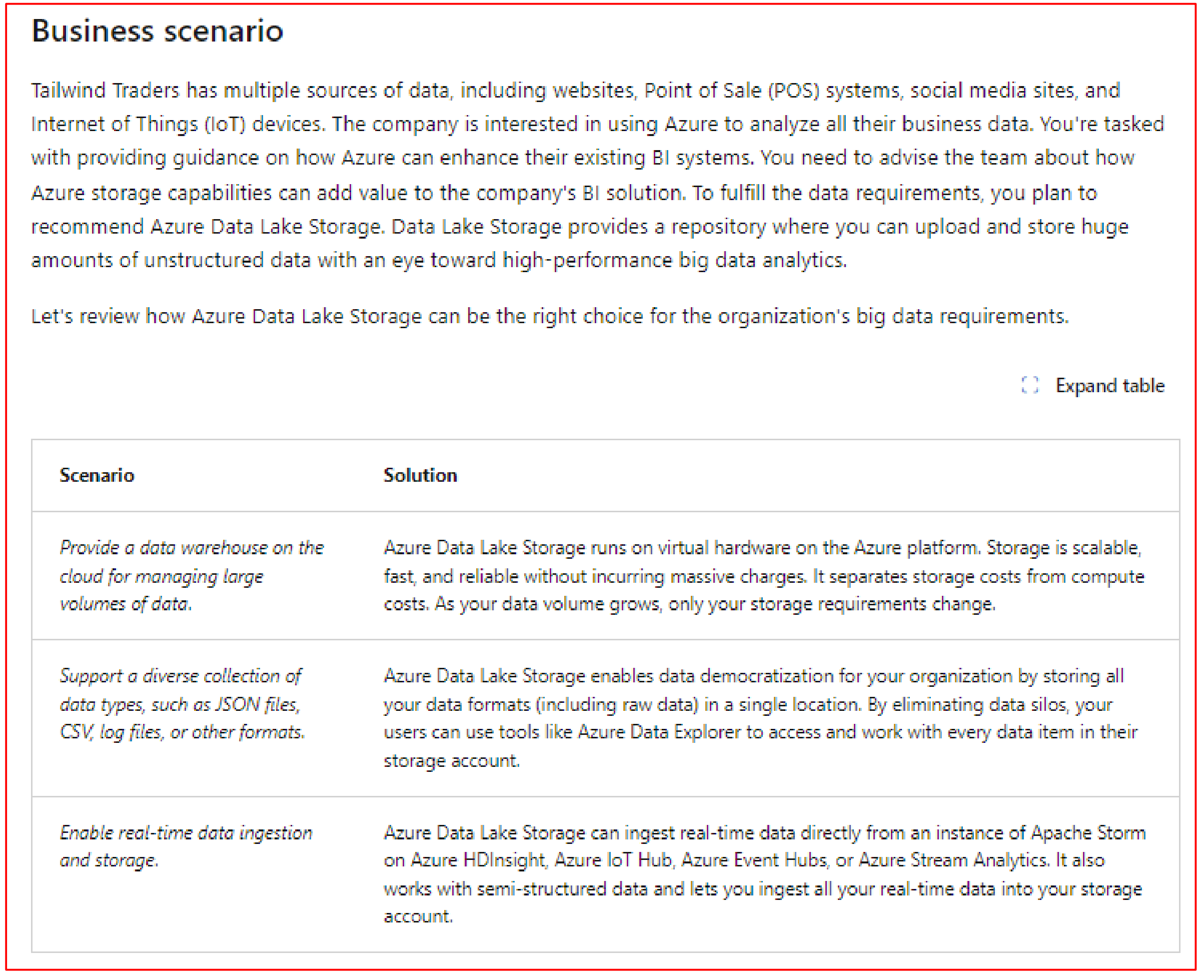
1. **Ingest data.** Azure Data Lake Storage offers many different data ingestion methods:

   - For unplanned data, you can use tools like AzCopy, the Azure CLI, PowerShell, and Azure Storage Explorer.
   - For relational data, the Azure Data Factory service can be used. You can transfer data from any source, such as Azure Cosmos DB, SQL Database, Azure SQL Managed instances, and more.
   - For streaming data, you can use tools like Apache Storm on Azure HDInsight, Azure Stream Analytics, and so on.

   The following diagram shows how unplanned data and streaming data are bulk ingested or unplanned ingested in Azure Data Lake Storage.

2. **Access stored data.** The easiest way to access your data is to use Azure Storage Explorer. Storage Explorer is a standalone application with a graphical user interface (GUI) for accessing your Azure Data Lake Storage data. You can also use PowerShell, the Azure CLI, HDFS CLI, or other programming language SDKs for accessing the data.

3. **Configure access control.** Control who can access the data stored in Azure Data Lake Storage by implementing an authorization mechanism. You can choose Azure RBAC or ACL.

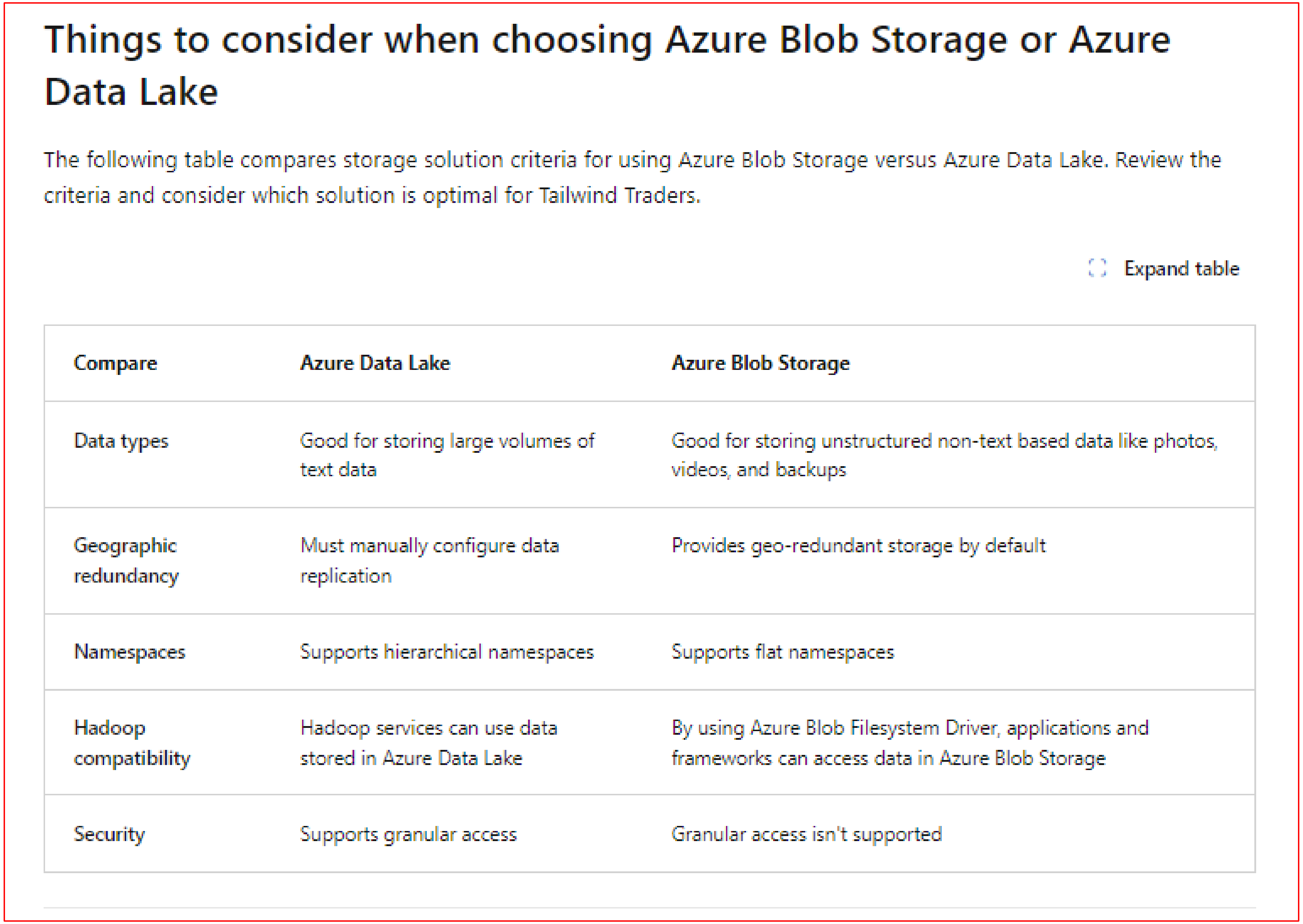


Business scenario

Tailwind Traders has multiple sources of data, including websites, Point of Sale (POS) systems, social media sites, and Internet of Things (IoT) devices. The company is interested in using Azure to analyze all their business data. You're tasked with providing guidance on how Azure can enhance their existing BI systems. You need to advise the team about how Azure storage capabilities can add value to the company's BI solution. To fulfill the data requirements, you plan to recommend Azure Data Lake Storage. Data Lake Storage provides a repository where you can upload and store huge amounts of unstructured data with an eye toward high-performance big data analytics.

Let's review how Azure Data Lake Storage can be the right choice for the organization's big data requirements.

Scenario	Solution
Provide a data warehouse on the cloud for managing large volumes of data.	Azure Data Lake Storage runs on virtual hardware on the Azure platform. Storage is scalable, fast, and reliable without incurring massive charges. It separates storage costs from compute costs. As your data volume grows, only your storage requirements change.
Support a diverse collection of data types, such as JSON files, CSV, log files, or other formats.	Azure Data Lake Storage enables data democratization for your organization by storing all your data formats (including raw data) in a single location. By eliminating data silos, your users can use tools like Azure Data Explorer to access and work with every data item in their storage account.
Enable real-time data ingestion and storage.	Azure Data Lake Storage can ingest real-time data directly from an instance of Apache Storm on Azure HDInsight, Azure IoT Hub, Azure Event Hubs, or Azure Stream Analytics. It also works with semi-structured data and lets you ingest all your real-time data into your storage account.



Things to consider when choosing Azure Blob Storage or Azure Data Lake

The following table compares storage solution criteria for using Azure Blob Storage versus Azure Data Lake. Review the criteria and consider which solution is optimal for Tailwind Traders.

Compare	Azure Data Lake	Azure Blob Storage
Data types	Good for storing large volumes of text data	Good for storing unstructured non-text based data like photos, videos, and backups
Geographic redundancy	Must manually configure data replication	Provides geo-redundant storage by default
Namespaces	Supports hierarchical namespaces	Supports flat namespaces
Hadoop compatibility	Hadoop services can use data stored in Azure Data Lake	By using Azure Blob Filesystem Driver, applications and frameworks can access data in Azure Blob Storage
Security	Supports granular access	Granular access isn't supported

Design a data integration and analytics solution with Azure Databricks

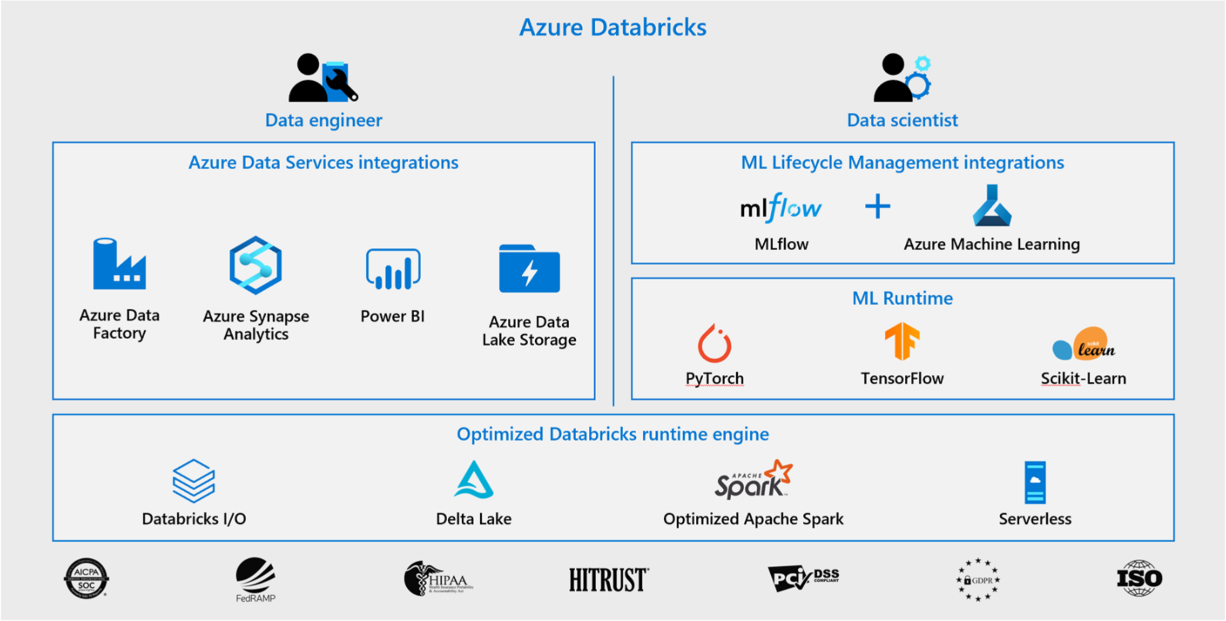
Azure Databricks

Azure Databricks is a fully managed, cloud-based Big Data and Machine Learning platform, which empowers developers to accelerate AI and innovation.

Provides data science and engineering teams with a single platform for Big Data processing and Machine Learning.

Offers three environments for developing data intensive applications:

- Databricks SQL
- Databricks Data Science & Engineering
- Databricks Machine Learning.



Azure Databricks

Data engineer

Data scientist

Azure Data Services integrations

Azure Data Factory

Azure Synapse Analytics

Power BI

Azure Data Lake Storage

ML Lifecycle Management integrations

mlflow + MLflow

Azure Machine Learning

ML Runtime

PyTorch

TensorFlow

Scikit-Learn

Optimized Databricks runtime engine

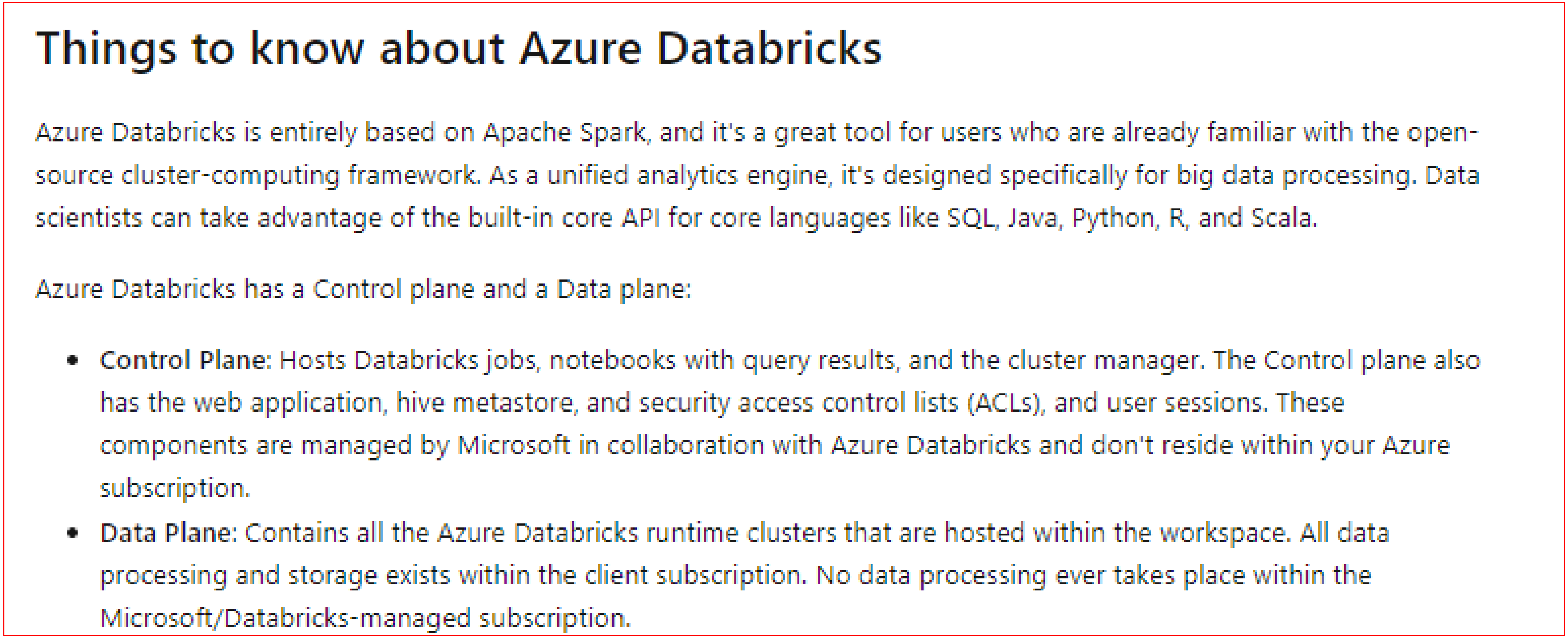
Databricks I/O

Delta Lake

Optimized Apache Spark

Serverless

Azure Databricks is a fully managed, cloud-based Big Data and Machine Learning platform, which empowers developers to accelerate AI and innovation. Azure Databricks provides data science and engineering teams with a single platform for big data processing and Machine Learning. The Azure Databricks managed Apache Spark platform makes it simple to run large-scale Spark workloads.

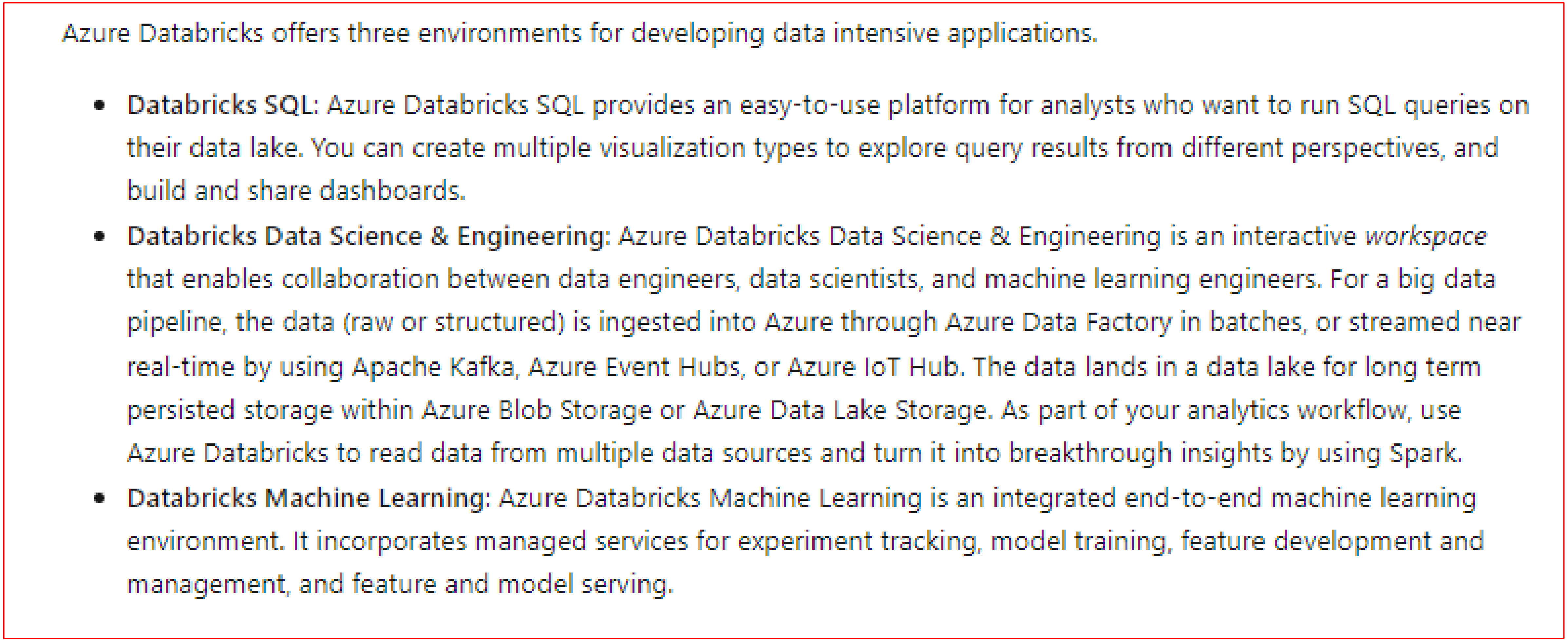


Things to know about Azure Databricks

Azure Databricks is entirely based on Apache Spark, and it's a great tool for users who are already familiar with the open-source cluster-computing framework. As a unified analytics engine, it's designed specifically for big data processing. Data scientists can take advantage of the built-in core API for core languages like SQL, Java, Python, R, and Scala.

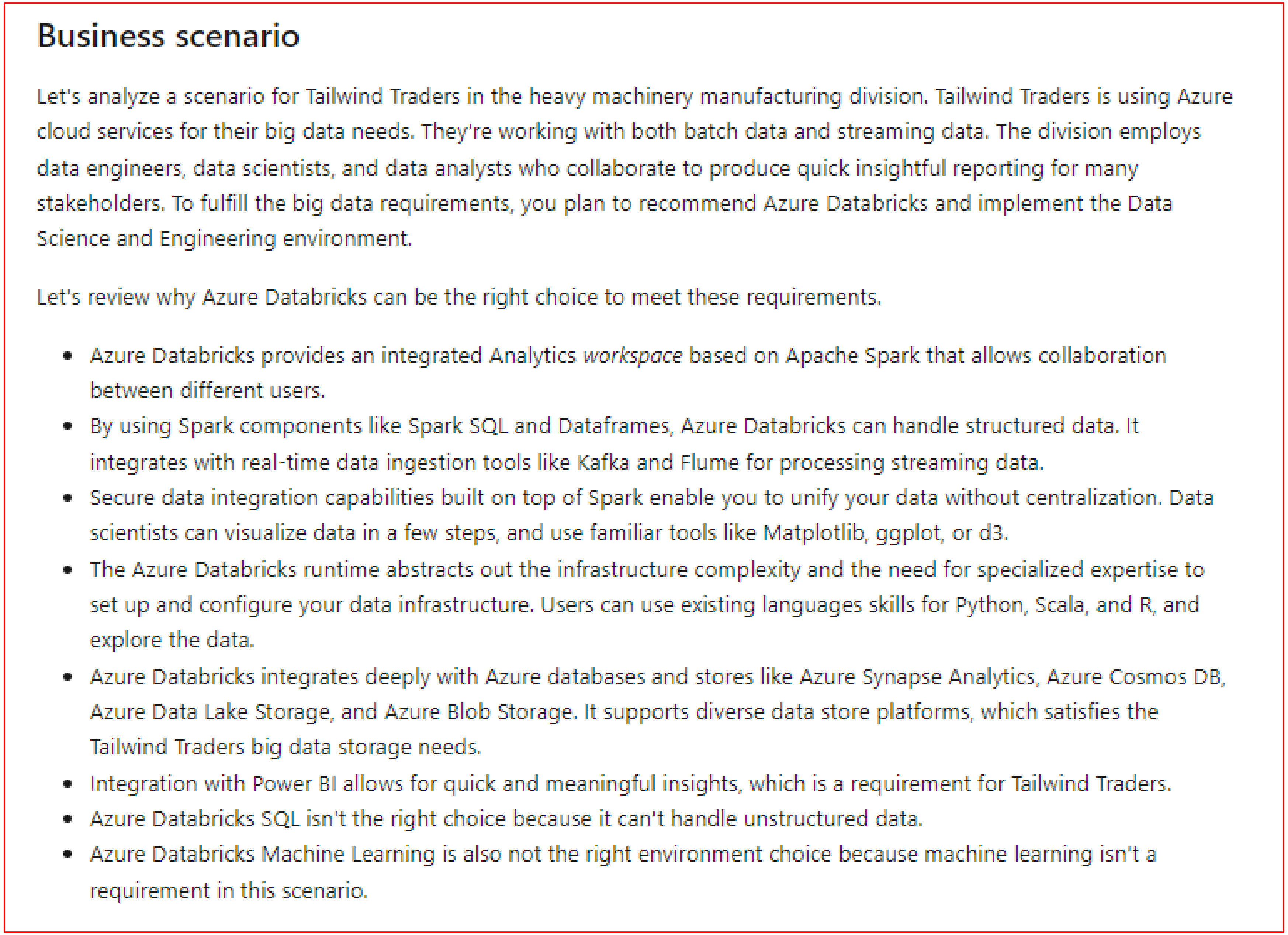
Azure Databricks has a Control plane and a Data plane:

- **Control Plane:** Hosts Databricks jobs, notebooks with query results, and the cluster manager. The Control plane also has the web application, hive metastore, and security access control lists (ACLs), and user sessions. These components are managed by Microsoft in collaboration with Azure Databricks and don't reside within your Azure subscription.
- **Data Plane:** Contains all the Azure Databricks runtime clusters that are hosted within the workspace. All data processing and storage exists within the client subscription. No data processing ever takes place within the Microsoft/Databricks-managed subscription.



Azure Databricks offers three environments for developing data intensive applications.

- **Databricks SQL**: Azure Databricks SQL provides an easy-to-use platform for analysts who want to run SQL queries on their data lake. You can create multiple visualization types to explore query results from different perspectives, and build and share dashboards.
- **Databricks Data Science & Engineering**: Azure Databricks Data Science & Engineering is an interactive *workspace* that enables collaboration between data engineers, data scientists, and machine learning engineers. For a big data pipeline, the data (raw or structured) is ingested into Azure through Azure Data Factory in batches, or streamed near real-time by using Apache Kafka, Azure Event Hubs, or Azure IoT Hub. The data lands in a data lake for long term persisted storage within Azure Blob Storage or Azure Data Lake Storage. As part of your analytics workflow, use Azure Databricks to read data from multiple data sources and turn it into breakthrough insights by using Spark.
- **Databricks Machine Learning**: Azure Databricks Machine Learning is an integrated end-to-end machine learning environment. It incorporates managed services for experiment tracking, model training, feature development and management, and feature and model serving.

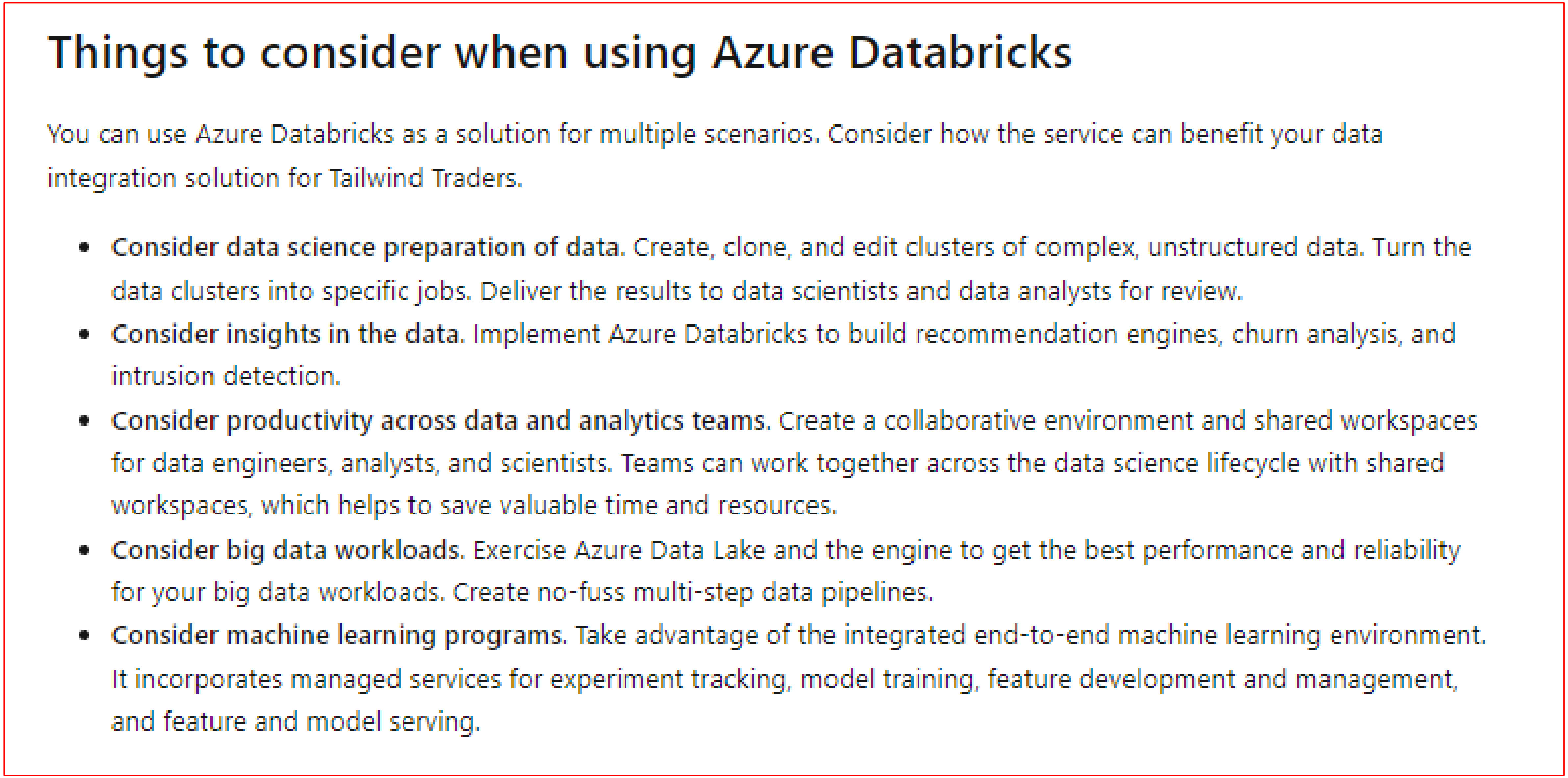


Business scenario

Let's analyze a scenario for Tailwind Traders in the heavy machinery manufacturing division. Tailwind Traders is using Azure cloud services for their big data needs. They're working with both batch data and streaming data. The division employs data engineers, data scientists, and data analysts who collaborate to produce quick insightful reporting for many stakeholders. To fulfill the big data requirements, you plan to recommend Azure Databricks and implement the Data Science and Engineering environment.

Let's review why Azure Databricks can be the right choice to meet these requirements.

- Azure Databricks provides an integrated Analytics *workspace* based on Apache Spark that allows collaboration between different users.
- By using Spark components like Spark SQL and Dataframes, Azure Databricks can handle structured data. It integrates with real-time data ingestion tools like Kafka and Flume for processing streaming data.
- Secure data integration capabilities built on top of Spark enable you to unify your data without centralization. Data scientists can visualize data in a few steps, and use familiar tools like Matplotlib, ggplot, or d3.
- The Azure Databricks runtime abstracts out the infrastructure complexity and the need for specialized expertise to set up and configure your data infrastructure. Users can use existing languages skills for Python, Scala, and R, and explore the data.
- Azure Databricks integrates deeply with Azure databases and stores like Azure Synapse Analytics, Azure Cosmos DB, Azure Data Lake Storage, and Azure Blob Storage. It supports diverse data store platforms, which satisfies the Tailwind Traders big data storage needs.
- Integration with Power BI allows for quick and meaningful insights, which is a requirement for Tailwind Traders.
- Azure Databricks SQL isn't the right choice because it can't handle unstructured data.
- Azure Databricks Machine Learning is also not the right environment choice because machine learning isn't a requirement in this scenario.



Things to consider when using Azure Databricks

You can use Azure Databricks as a solution for multiple scenarios. Consider how the service can benefit your data integration solution for Tailwind Traders.

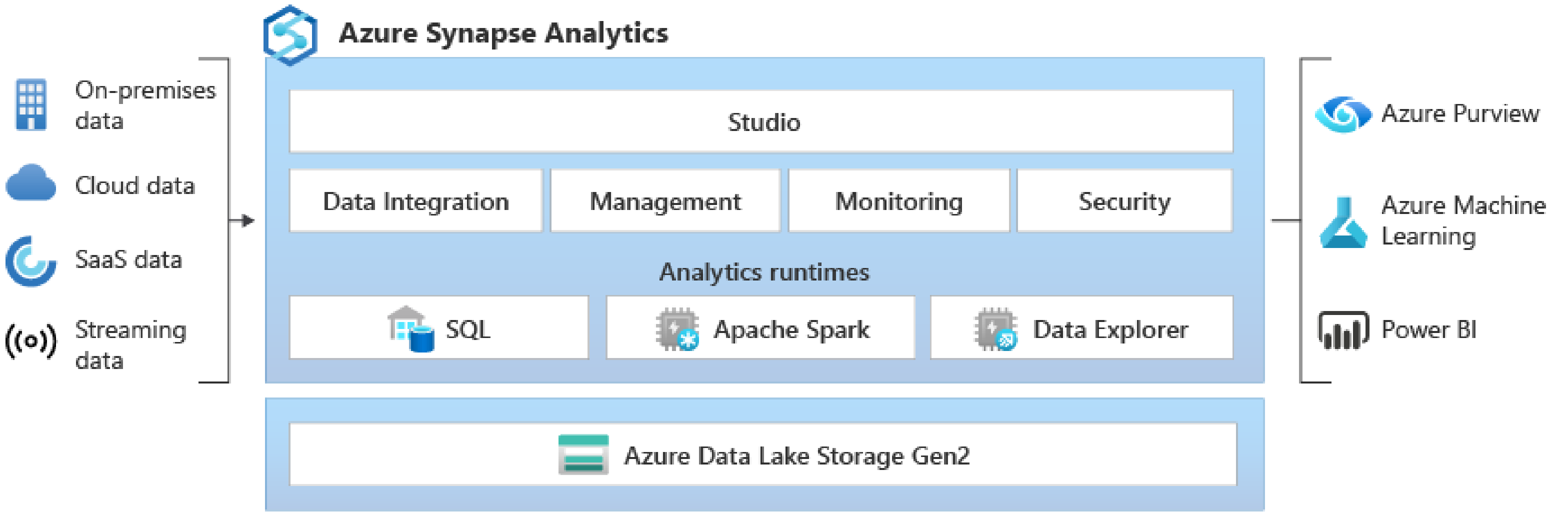
- **Consider data science preparation of data.** Create, clone, and edit clusters of complex, unstructured data. Turn the data clusters into specific jobs. Deliver the results to data scientists and data analysts for review.
- **Consider insights in the data.** Implement Azure Databricks to build recommendation engines, churn analysis, and intrusion detection.
- **Consider productivity across data and analytics teams.** Create a collaborative environment and shared workspaces for data engineers, analysts, and scientists. Teams can work together across the data science lifecycle with shared workspaces, which helps to save valuable time and resources.
- **Consider big data workloads.** Exercise Azure Data Lake and the engine to get the best performance and reliability for your big data workloads. Create no-fuss multi-step data pipelines.
- **Consider machine learning programs.** Take advantage of the integrated end-to-end machine learning environment. It incorporates managed services for experiment tracking, model training, feature development and management, and feature and model serving.

Design a data integration and analytics solution with Azure Synapse Analytics

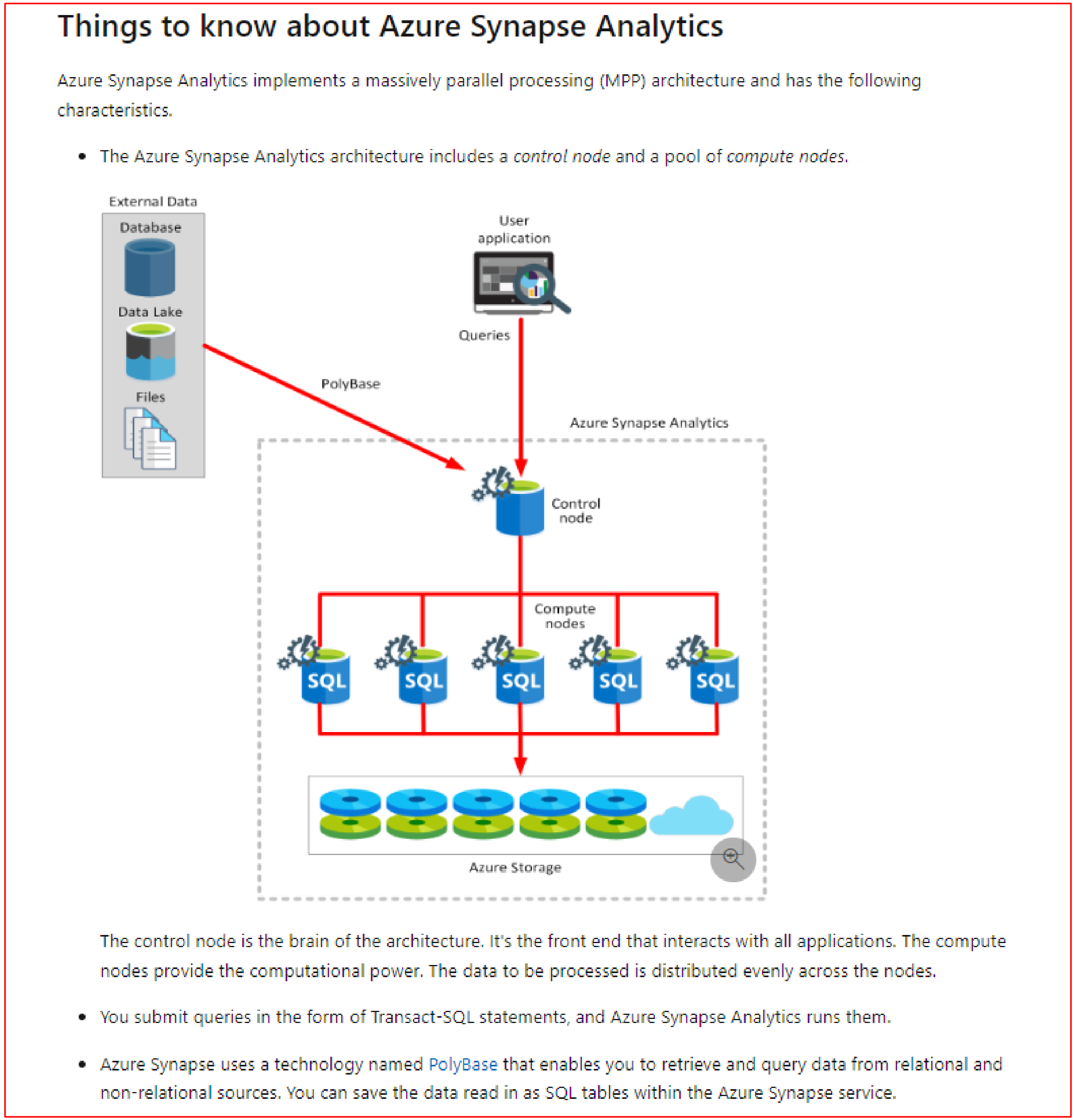
Azure Synapse Analytics

Azure Synapse Analytics is an integrated analytics platform that brings together data integration, enterprise data warehousing, big data analytics and visualization into a single service. Azure Synapse Analytics is an evolution of Azure SQL Data Warehouse.

- Modern data warehousing
- Advanced analytics
- Data exploration and discovery
- Real time analytics
- Data integration
- Integrated analytics
- Machine Learning



Azure Synapse Analytics combines features of big data analytics, enterprise data storage, and data integration. The service lets you run queries on serverless data or data at scale. Azure Synapse supports data ingestion, exploration, transformation, and management, and supports analysis for all your BI and machine learning needs.



Things to know about Azure Synapse Analytics

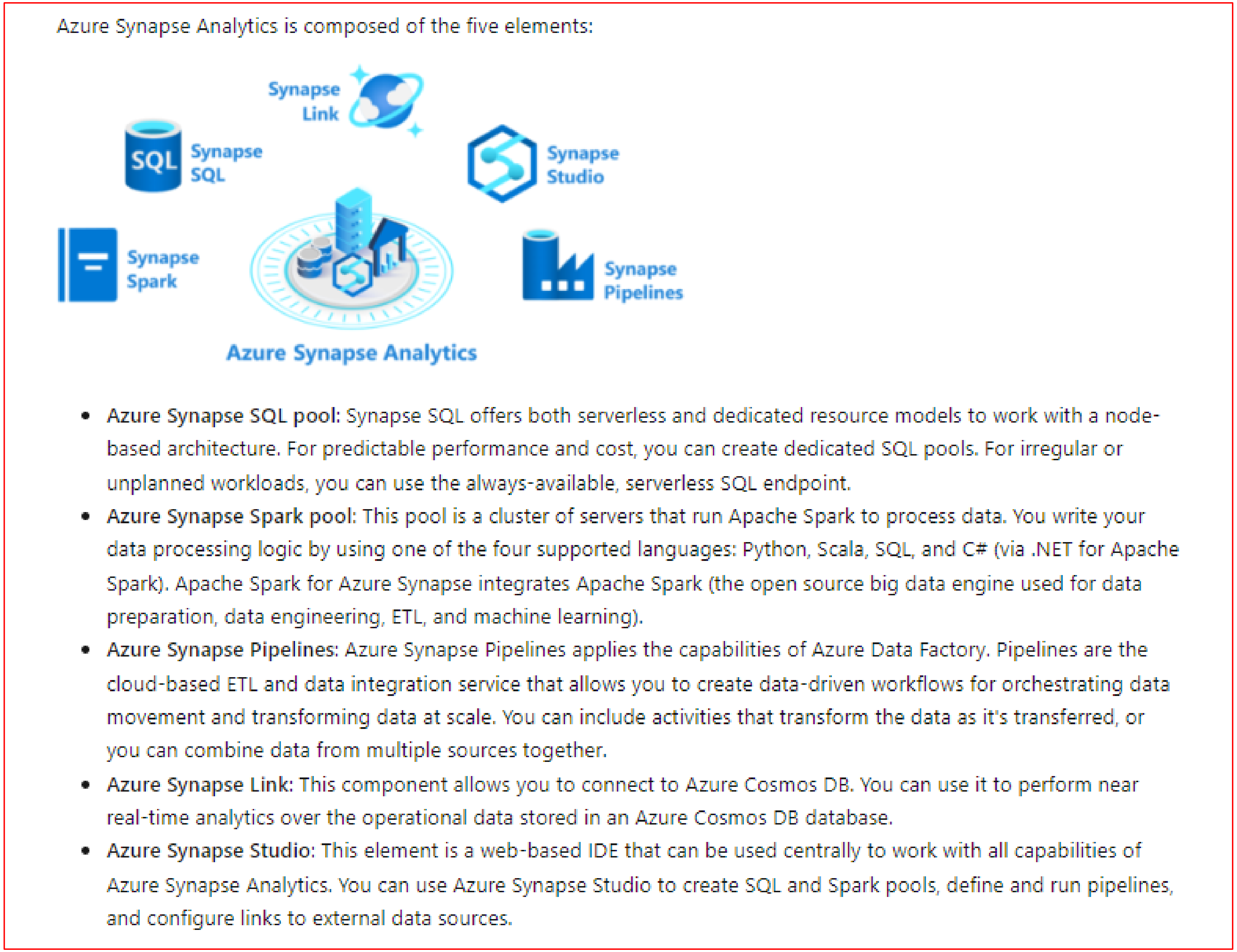
Azure Synapse Analytics implements a massively parallel processing (MPP) architecture and has the following characteristics.

- The Azure Synapse Analytics architecture includes a *control node* and a pool of *compute nodes*.

The control node is the brain of the architecture. It's the front end that interacts with all applications. The compute nodes provide the computational power. The data to be processed is distributed evenly across the nodes.

- You submit queries in the form of Transact-SQL statements, and Azure Synapse Analytics runs them.
- Azure Synapse uses a technology named PolyBase that enables you to retrieve and query data from relational and non-relational sources. You can save the data read in as SQL tables within the Azure Synapse service.

Components of Azure Synapse Analytics



Azure Synapse Analytics is composed of the five elements:

- **Azure Synapse SQL pool:** Synapse SQL offers both serverless and dedicated resource models to work with a node-based architecture. For predictable performance and cost, you can create dedicated SQL pools. For irregular or unplanned workloads, you can use the always-available, serverless SQL endpoint.
- **Azure Synapse Spark pool:** This pool is a cluster of servers that run Apache Spark to process data. You write your data processing logic by using one of the four supported languages: Python, Scala, SQL, and C# (via .NET for Apache Spark). Apache Spark for Azure Synapse integrates Apache Spark (the open source big data engine used for data preparation, data engineering, ETL, and machine learning).
- **Azure Synapse Pipelines:** Azure Synapse Pipelines applies the capabilities of Azure Data Factory. Pipelines are the cloud-based ETL and data integration service that allows you to create data-driven workflows for orchestrating data movement and transforming data at scale. You can include activities that transform the data as it's transferred, or you can combine data from multiple sources together.
- **Azure Synapse Link:** This component allows you to connect to Azure Cosmos DB. You can use it to perform near real-time analytics over the operational data stored in an Azure Cosmos DB database.
- **Azure Synapse Studio:** This element is a web-based IDE that can be used centrally to work with all capabilities of Azure Synapse Analytics. You can use Azure Synapse Studio to create SQL and Spark pools, define and run pipelines, and configure links to external data sources.

What kind of analytics can you do with Azure Synapse Analytics?

Descriptive analytics - "What is happening?"

Azure Synapse Analytics leverages the dedicated SQL pool capability that enables you to create a persisted data warehouse to perform this type of analysis.

Diagnostic analytics - "Why is it happening?"

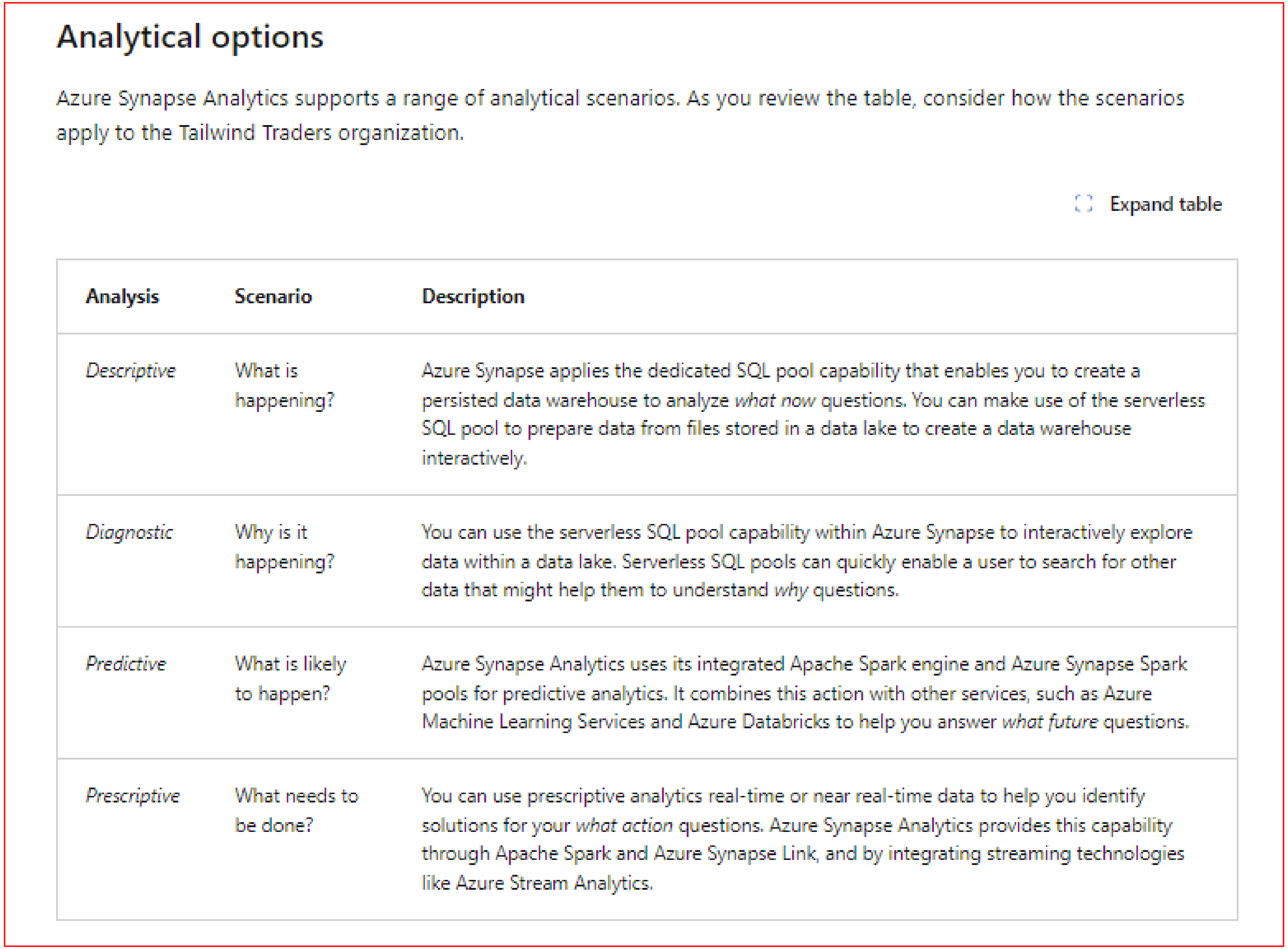
You can use the serverless SQL pool capability within Azure Synapse Analytics that enables you to interactively explore data within a data lake.

Predictive analytics - "What is likely to happen?"

Azure Synapse Analytics uses its integrated Apache Spark engine and Azure Synapse Spark pools for predictive analytics with other services such as Azure Machine Learning Services, or Azure Databricks.

Prescriptive analytics - "What needs to be done?"

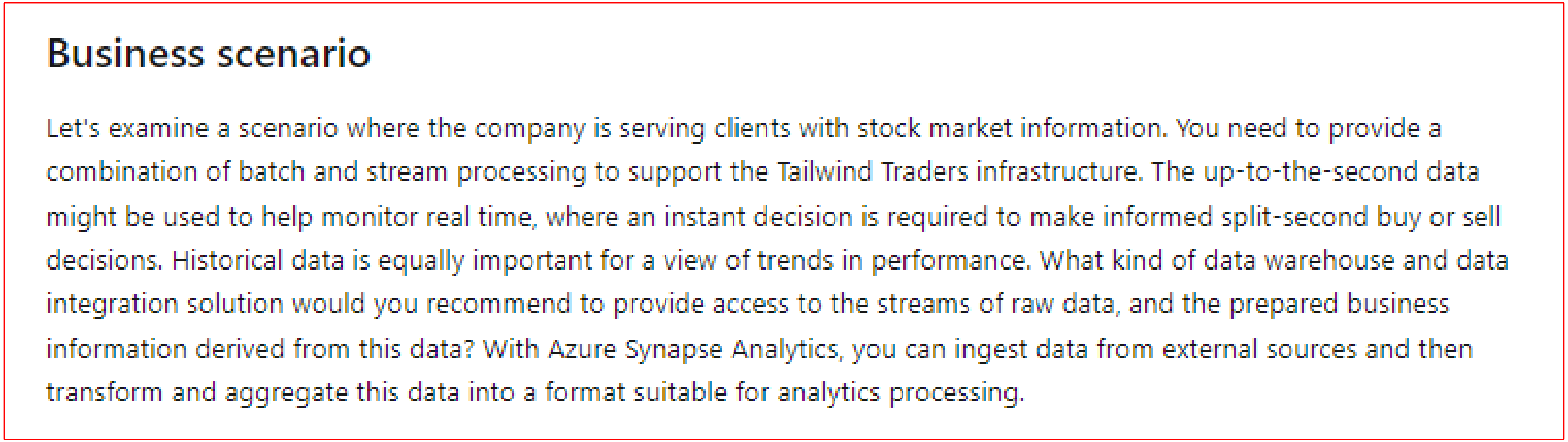
This type of analytics looks at executing actions based on real-time or near real-time analysis of data, using predictive analytics.



Analytical options

Azure Synapse Analytics supports a range of analytical scenarios. As you review the table, consider how the scenarios apply to the Tailwind Traders organization.

Analysis	Scenario	Description
Descriptive	What is happening?	Azure Synapse applies the dedicated SQL pool capability that enables you to create a persisted data warehouse to analyze *what now* questions. You can make use of the serverless SQL pool to prepare data from files stored in a data lake to create a data warehouse interactively.
Diagnostic	Why is it happening?	You can use the serverless SQL pool capability within Azure Synapse to interactively explore data within a data lake. Serverless SQL pools can quickly enable a user to search for other data that might help them to understand *why* questions.
Predictive	What is likely to happen?	Azure Synapse Analytics uses its integrated Apache Spark engine and Azure Synapse Spark pools for predictive analytics. It combines this action with other services, such as Azure Machine Learning Services and Azure Databricks to help you answer *what future* questions.
Prescriptive	What needs to be done?	You can use prescriptive analytics real-time or near real-time data to help you identify solutions for your *what action* questions. Azure Synapse Analytics provides this capability through Apache Spark and Azure Synapse Link, and by integrating streaming technologies like Azure Stream Analytics.

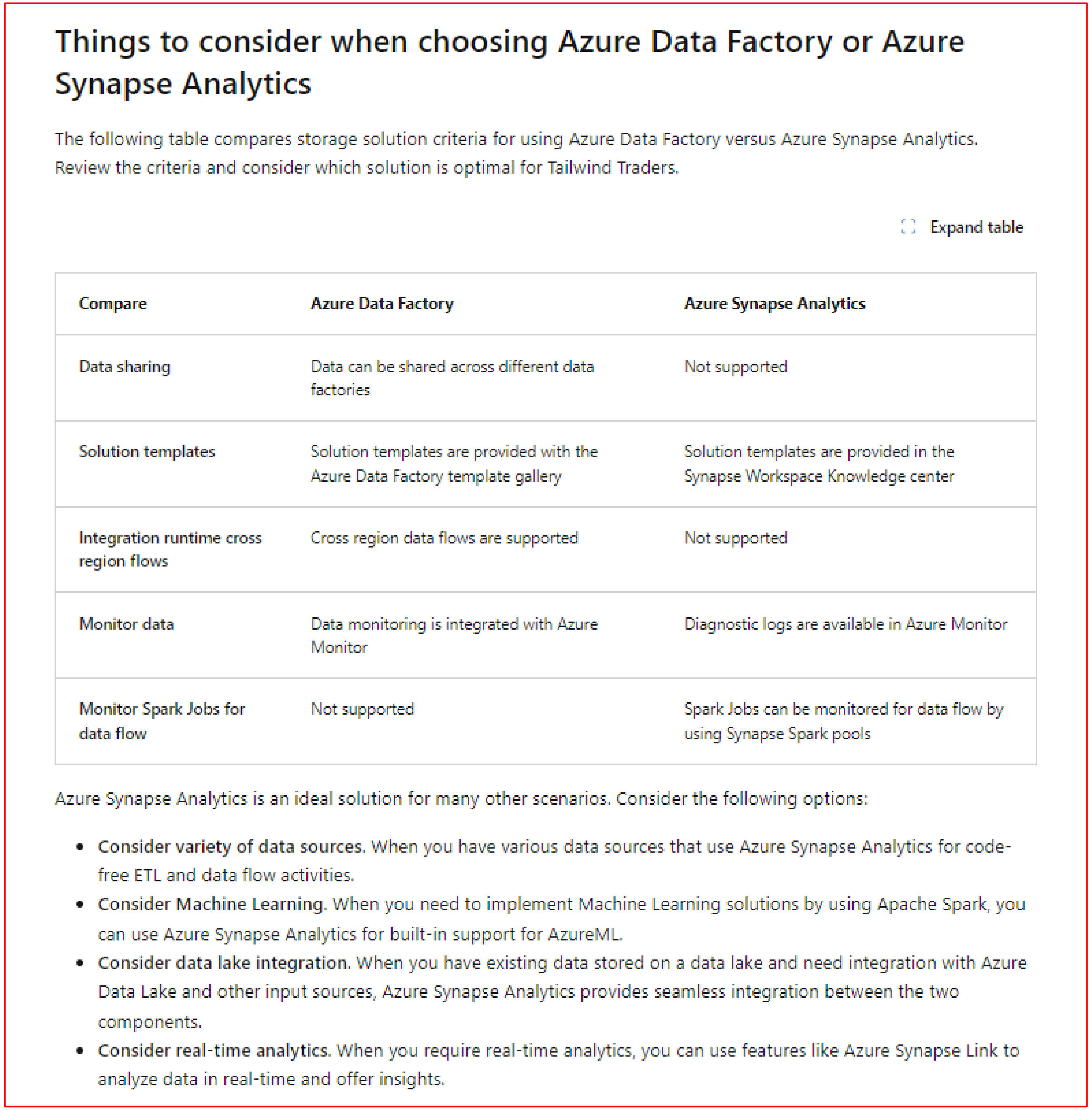


Business scenario

Let's examine a scenario where the company is serving clients with stock market information. You need to provide a combination of batch and stream processing to support the Tailwind Traders infrastructure. The up-to-the-second data might be used to help monitor real time, where an instant decision is required to make informed split-second buy or sell decisions. Historical data is equally important for a view of trends in performance. What kind of data warehouse and data integration solution would you recommend to provide access to the streams of raw data, and the prepared business information derived from this data? With Azure Synapse Analytics, you can ingest data from external sources and then transform and aggregate this data into a format suitable for analytics processing.

Compare Azure Data Factory to Azure Synapse Analytics

Criteria	Azure Data Factory	Azure Synapse Analytics
Integration runtime sharing	Can be shared across different data factories	No sharing
Solution templates	Provided with Azure Data Factory template gallery	Provided with Synapse Workspace Knowledge center
Integration Runtime cross region support	Support Cross region data flows	Does not support cross region data flows
Monitoring of Spark Jobs for Data Flow	Not supported	Supported by the Synapse Spark pools



Things to consider when choosing Azure Data Factory or Azure Synapse Analytics

The following table compares storage solution criteria for using Azure Data Factory versus Azure Synapse Analytics. Review the criteria and consider which solution is optimal for Tailwind Traders.

Compare	Azure Data Factory	Azure Synapse Analytics
Data sharing	Data can be shared across different data factories	Not supported
Solution templates	Solution templates are provided with the Azure Data Factory template gallery	Solution templates are provided in the Synapse Workspace Knowledge center
Integration runtime cross region flows	Cross region data flows are supported	Not supported
Monitor data	Data monitoring is integrated with Azure Monitor	Diagnostic logs are available in Azure Monitor
Monitor Spark Jobs for data flow	Not supported	Spark Jobs can be monitored for data flow by using Synapse Spark pools

Azure Synapse Analytics is an ideal solution for many other scenarios. Consider the following options:

- **Consider variety of data sources.** When you have various data sources that use Azure Synapse Analytics for code-free ETL and data flow activities.
- **Consider Machine Learning.** When you need to implement Machine Learning solutions by using Apache Spark, you can use Azure Synapse Analytics for built-in support for AzureML.
- **Consider data lake integration.** When you have existing data stored on a data lake and need integration with Azure Data Lake and other input sources, Azure Synapse Analytics provides seamless integration between the two components.
- **Consider real-time analytics.** When you require real-time analytics, you can use features like Azure Synapse Link to analyze data in real-time and offer insights.

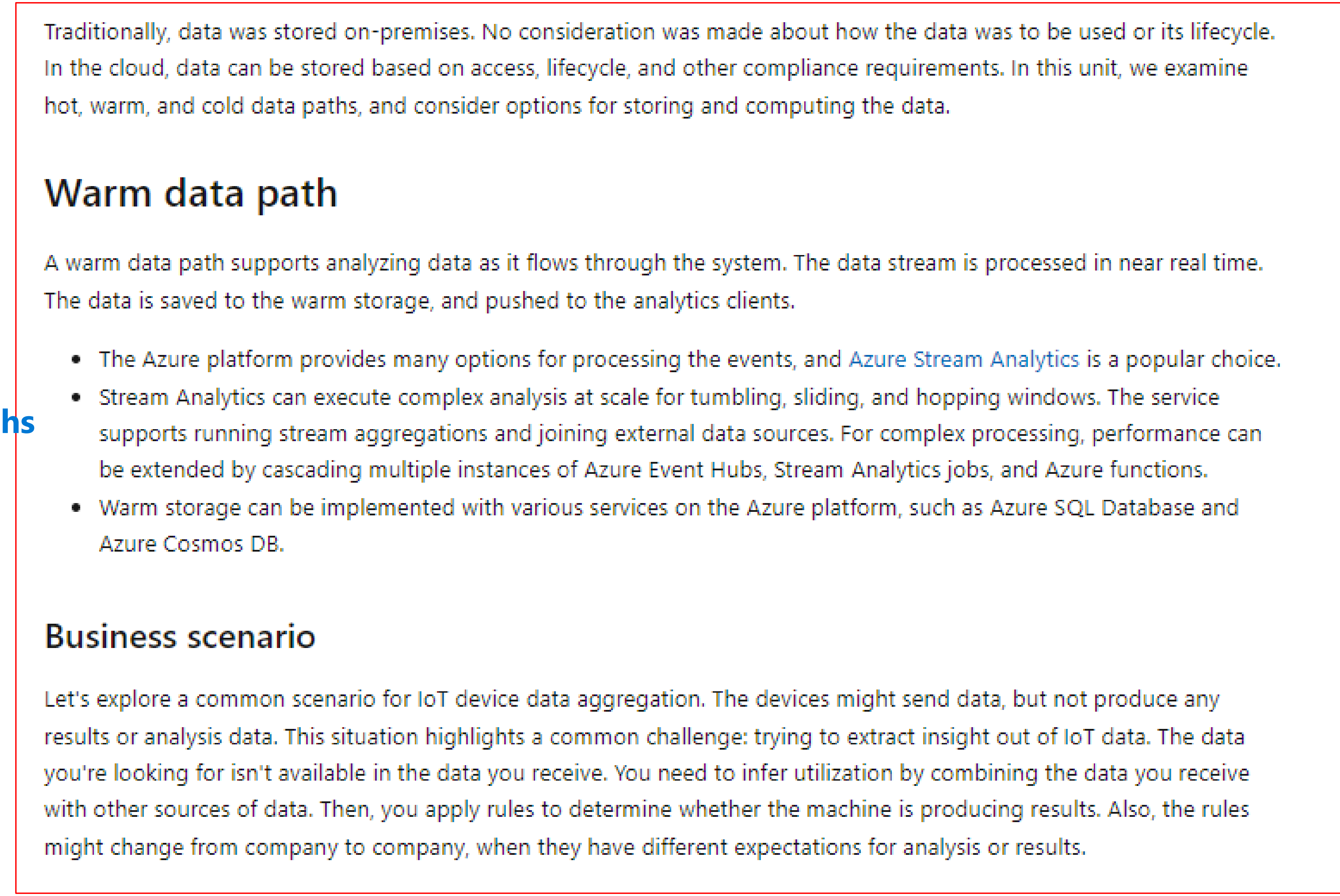
Compare Synapse to Databricks

Azure Synapse Analytics and Azure Databricks offer different capabilities which may be combined if required

Capabilities	Databricks	Synapse
Machine Learning	• Optimized runtimes with support for TensorFlow, PyTorch, and Keras. • GPU support	• Built-in support Azure ML • Train models using SparkML, MLlib and other open-source libraries • GPU-accelerated pools
Feature Set	• Optimized Apache Spark environment	• Distributed T-SQL system • Spark environment • Data Integration • Unified experience with Synapse Studio
Reporting	• Azure Databricks connection available in PowerBI	• PowerBI available directly from Synapse Studio

Design a strategy for hot/warm/cold data path

Design strategies for hot, warm, and cold data paths



Traditionally, data was stored on-premises. No consideration was made about how the data was to be used or its lifecycle. In the cloud, data can be stored based on access, lifecycle, and other compliance requirements. In this unit, we examine hot, warm, and cold data paths, and consider options for storing and computing the data.

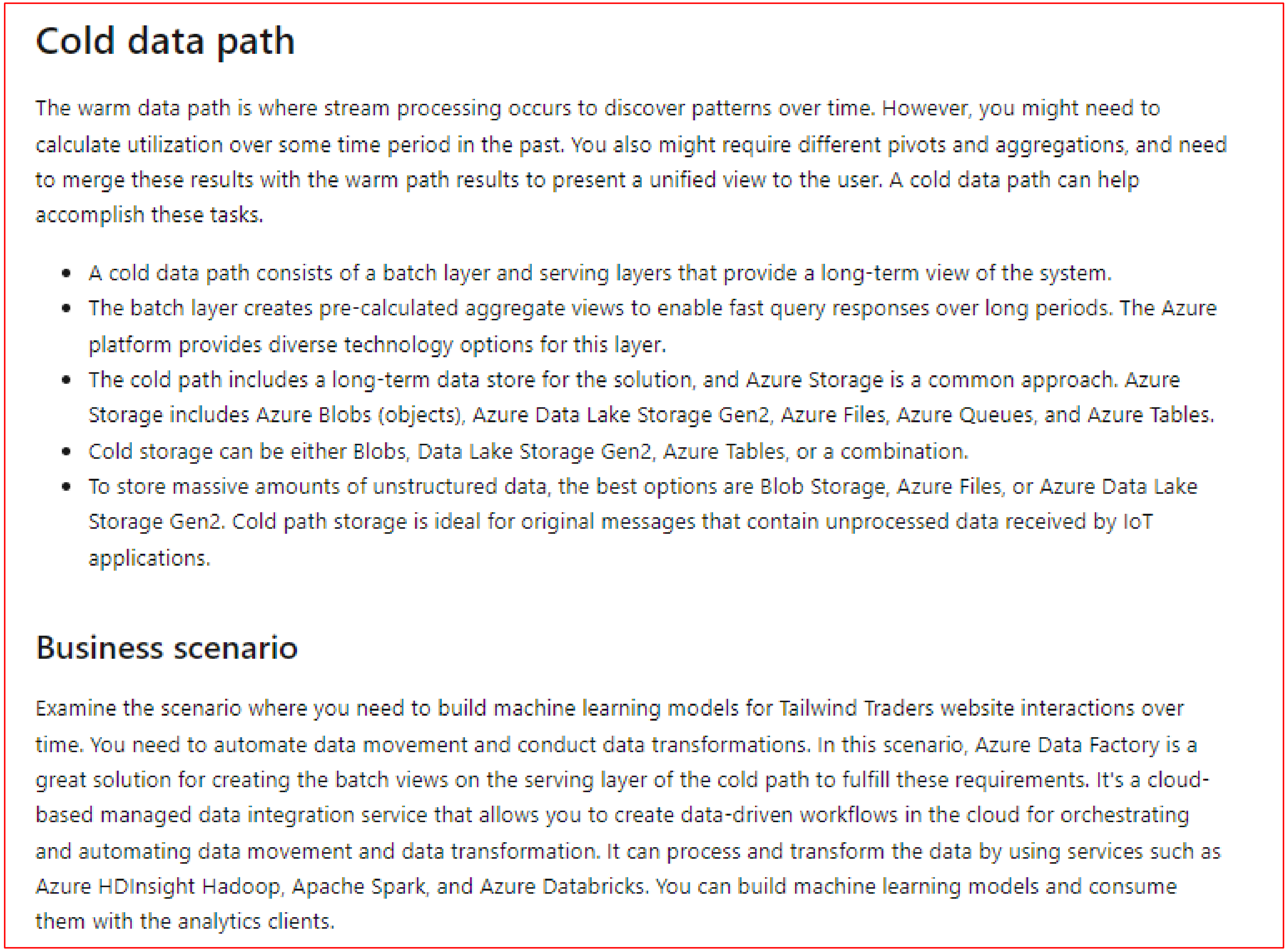
Warm data path

A warm data path supports analyzing data as it flows through the system. The data stream is processed in near real time. The data is saved to the warm storage, and pushed to the analytics clients.

- The Azure platform provides many options for processing the events, and Azure Stream Analytics is a popular choice.
- Stream Analytics can execute complex analysis at scale for tumbling, sliding, and hopping windows. The service supports running stream aggregations and joining external data sources. For complex processing, performance can be extended by cascading multiple instances of Azure Event Hubs, Stream Analytics jobs, and Azure functions.
- Warm storage can be implemented with various services on the Azure platform, such as Azure SQL Database and Azure Cosmos DB.

Business scenario

Let's explore a common scenario for IoT device data aggregation. The devices might send data, but not produce any results or analysis data. This situation highlights a common challenge: trying to extract insight out of IoT data. The data you're looking for isn't available in the data you receive. You need to infer utilization by combining the data you receive with other sources of data. Then, you apply rules to determine whether the machine is producing results. Also, the rules might change from company to company, when they have different expectations for analysis or results.



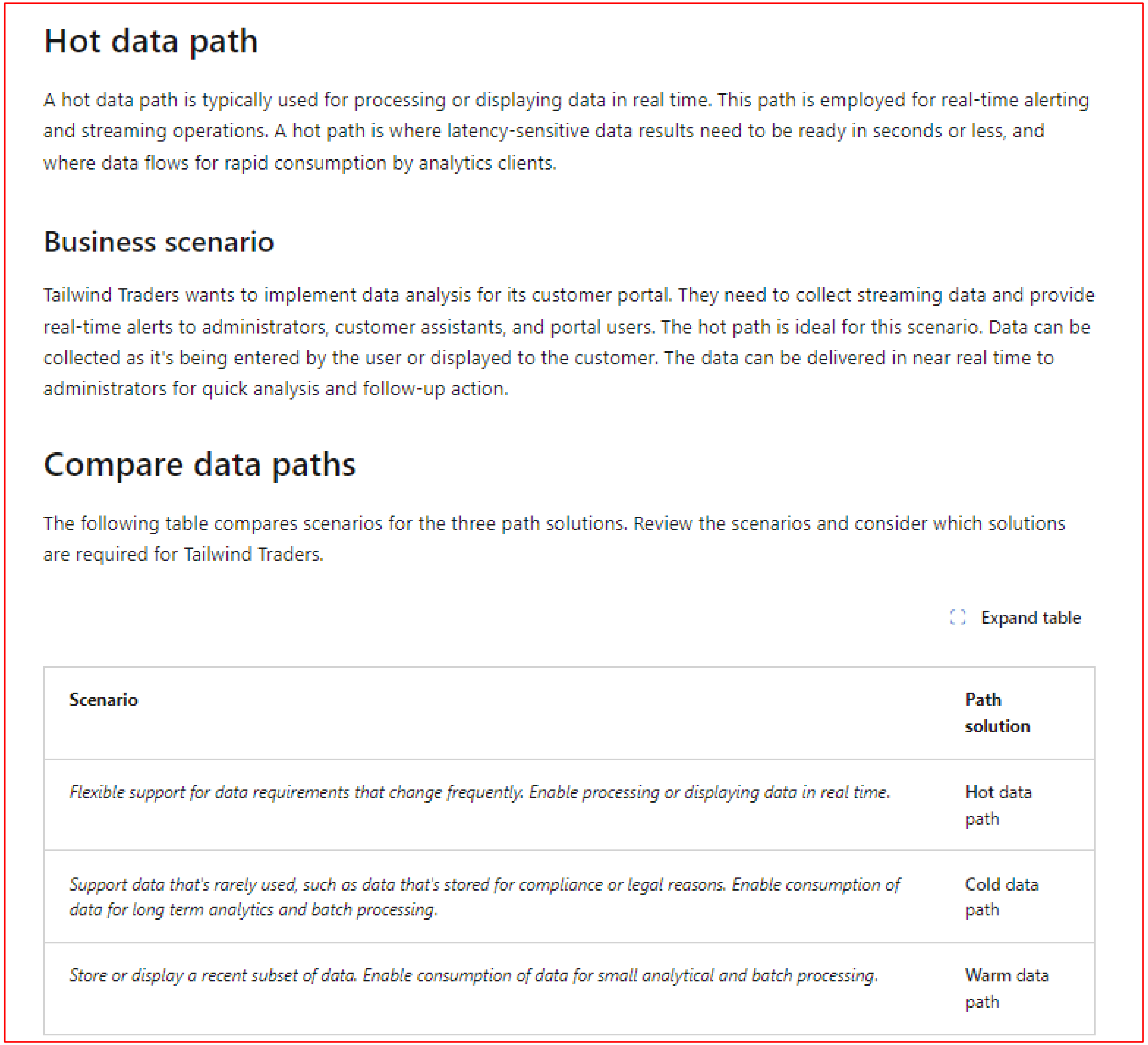
Cold data path

The warm data path is where stream processing occurs to discover patterns over time. However, you might need to calculate utilization over some time period in the past. You also might require different pivots and aggregations, and need to merge these results with the warm path results to present a unified view to the user. A cold data path can help accomplish these tasks.

- A cold data path consists of a batch layer and serving layers that provide a long-term view of the system.
- The batch layer creates pre-calculated aggregate views to enable fast query responses over long periods. The Azure platform provides diverse technology options for this layer.
- The cold path includes a long-term data store for the solution, and Azure Storage is a common approach. Azure Storage includes Azure Blobs (objects), Azure Data Lake Storage Gen2, Azure Files, Azure Queues, and Azure Tables.
- Cold storage can be either Blobs, Data Lake Storage Gen2, Azure Tables, or a combination.
- To store massive amounts of unstructured data, the best options are Blob Storage, Azure Files, or Azure Data Lake Storage Gen2. Cold path storage is ideal for original messages that contain unprocessed data received by IoT applications.

Business scenario

Examine the scenario where you need to build machine learning models for Tailwind Traders website interactions over time. You need to automate data movement and conduct data transformations. In this scenario, Azure Data Factory is a great solution for creating the batch views on the serving layer of the cold path to fulfill these requirements. It's a cloud-based managed data integration service that allows you to create data-driven workflows in the cloud for orchestrating and automating data movement and data transformation. It can process and transform the data by using services such as Azure HDInsight Hadoop, Apache Spark, and Azure Databricks. You can build machine learning models and consume them with the analytics clients.



Hot data path

A hot data path is typically used for processing or displaying data in real time. This path is employed for real-time alerting and streaming operations. A hot path is where latency-sensitive data results need to be ready in seconds or less, and where data flows for rapid consumption by analytics clients.

Business scenario

Tailwind Traders wants to implement data analysis for its customer portal. They need to collect streaming data and provide real-time alerts to administrators, customer assistants, and portal users. The hot path is ideal for this scenario. Data can be collected as it's being entered by the user or displayed to the customer. The data can be delivered in near real time to administrators for quick analysis and follow-up action.

Compare data paths

The following table compares scenarios for the three path solutions. Review the scenarios and consider which solutions are required for Tailwind Traders.

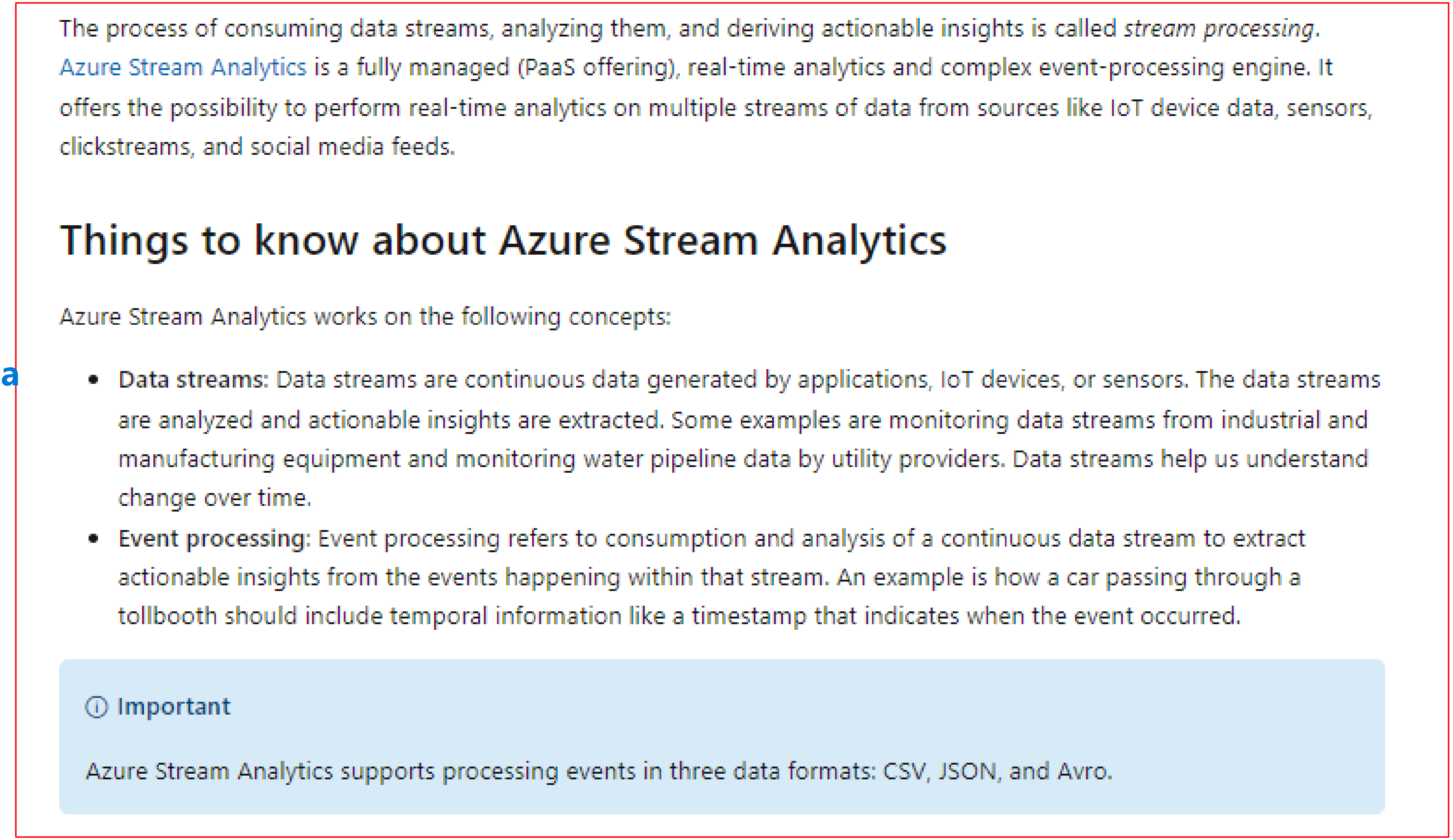
Scenario	**Path solution**
Flexible support for data requirements that change frequently. Enable processing or displaying data in real time.	Hot data path
Support data that's rarely used, such as data that's stored for compliance or legal reasons. Enable consumption of data for long term analytics and batch processing.	Cold data path
Store or display a recent subset of data. Enable consumption of data for small analytical and batch processing.	Warm data path

When to use Hot/Warm/Cold data path

Path	Requirement
Hot data path	• When data requirements are known to change frequently • When processing or displaying data in real time
Warm data path	• When you need to store or display a recent subset of data • Used for data that is consumed for small analytical and batch processing
Cold data path	• When data is rarely used. The data might be stored for compliance or legal reasons • Used for data that is consumed for long term analytics and batch processing

Design Azure Stream Analytics solution for Data Analysis

Design an Azure Stream Analytics solution for data analysis



The process of consuming data streams, analyzing them, and deriving actionable insights is called *stream processing*. Azure Stream Analytics is a fully managed (PaaS offering), real-time analytics and complex event-processing engine. It offers the possibility to perform real-time analytics on multiple streams of data from sources like IoT device data, sensors, clickstreams, and social media feeds.

Things to know about Azure Stream Analytics

Azure Stream Analytics works on the following concepts:

- **Data streams:** Data streams are continuous data generated by applications, IoT devices, or sensors. The data streams are analyzed and actionable insights are extracted. Some examples are monitoring data streams from industrial and manufacturing equipment and monitoring water pipeline data by utility providers. Data streams help us understand change over time.
- **Event processing**: Event processing refers to consumption and analysis of a continuous data stream to extract actionable insights from the events happening within that stream. An example is how a car passing through a tollbooth should include temporal information like a timestamp that indicates when the event occurred.

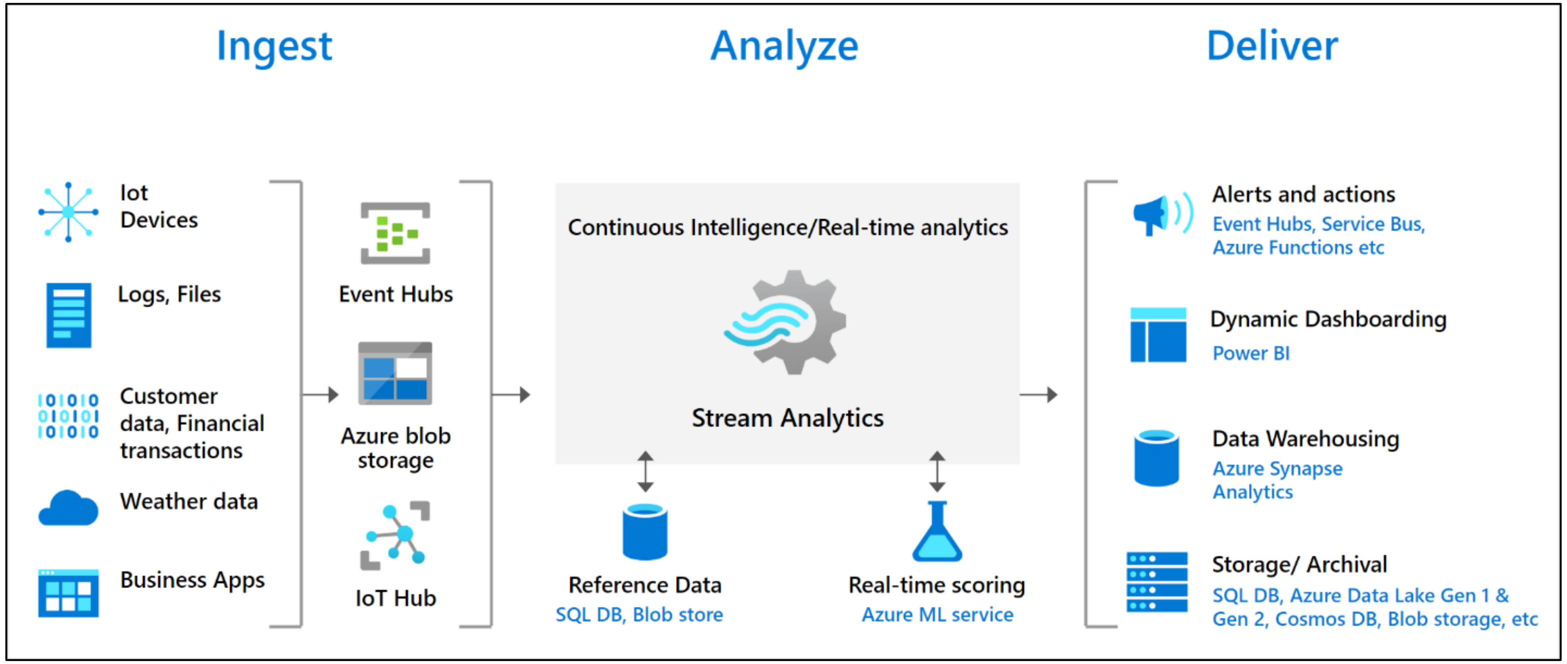
ⓘ **Important**

Azure Stream Analytics supports processing events in three data formats: CSV, JSON, and Avro.

Azure Stream Analytics

Azure Stream Analytics is a real-time analytics and complex event-processing engine that is designed to analyze and process high volumes of fast streaming data from multiple sources simultaneously.

- Analyze real-time telemetry streams from IoT devices
- Web logs/clickstream analytics
- Geospatial analytics for fleet management and driverless vehicles
- Remote monitoring and predictive maintenance of high value assets
- Real-time analytics on point-of-sale data for inventory control and anomaly detection



Ingest

Iot Devices

Logs, Files

Customer data, Financial transactions

Weather data

Business Apps

Event Hubs

Azure blob storage

IoT Hub

Analyze

Continuous Intelligence/Real-time analytics

Stream Analytics

Reference Data
SQL DB, Blob store

Real-time scoring
Azure ML service

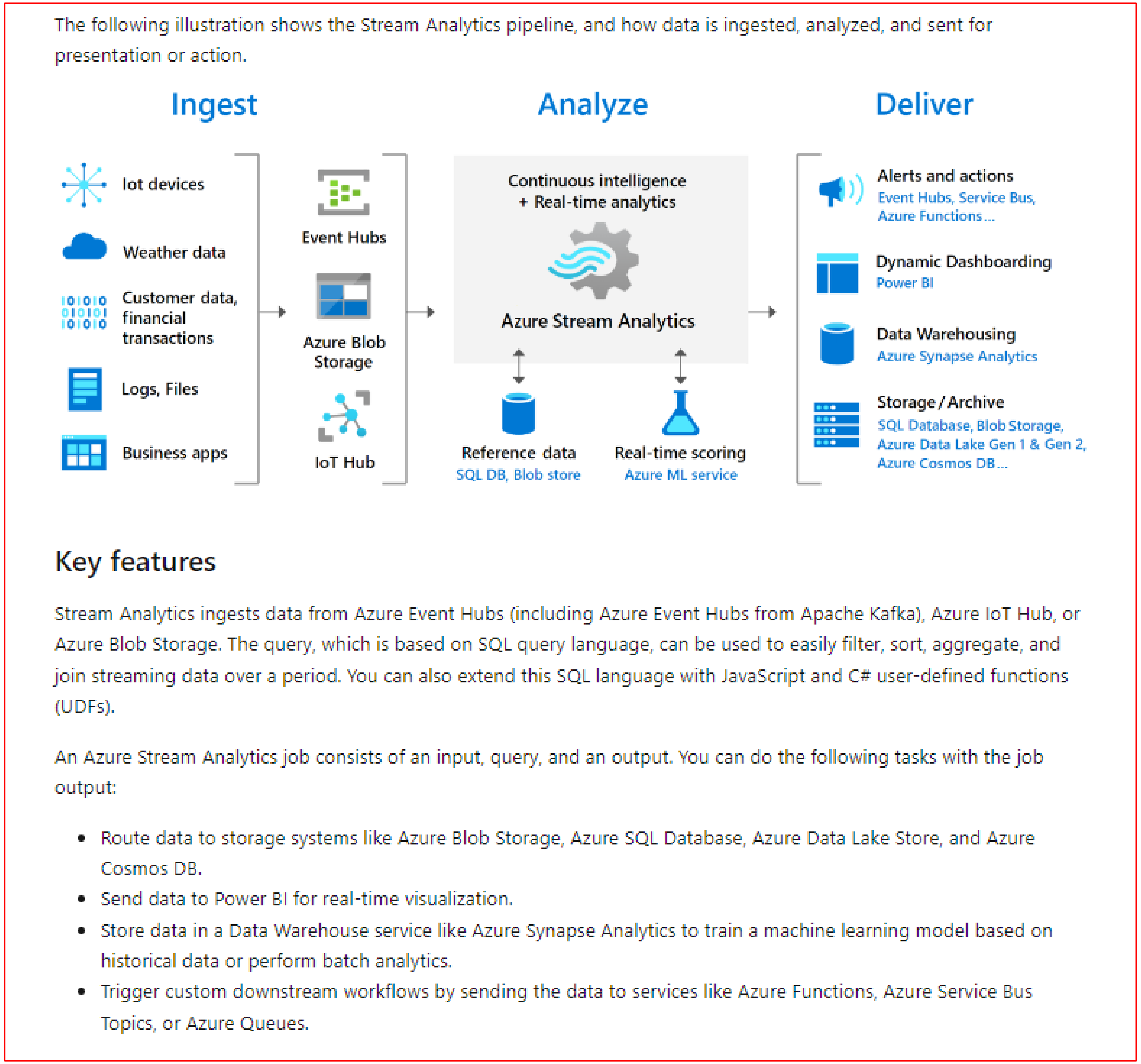
Deliver

Alerts and actions
Event Hubs, Service Bus, Azure Functions etc

Dynamic Dashboarding
Power BI

Data Warehousing
Azure Synapse Analytics

Storage/ Archival
SQL DB, Azure Data Lake Gen 1 & Gen 2, Cosmos DB, Blob storage, etc



The following illustration shows the Stream Analytics pipeline, and how data is ingested, analyzed, and sent for presentation or action.

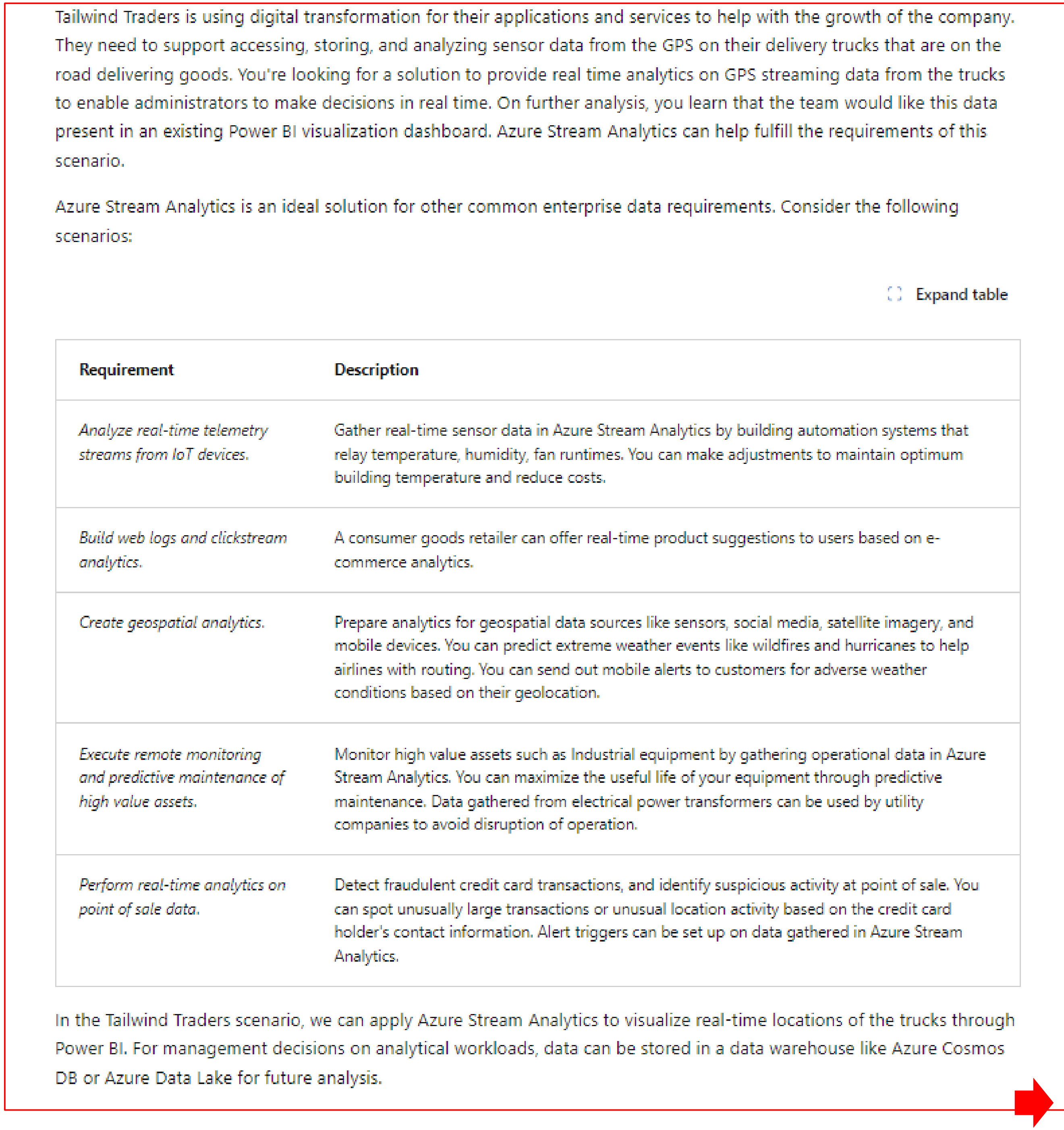
Key features

Stream Analytics ingests data from Azure Event Hubs (including Azure Event Hubs from Apache Kafka), Azure IoT Hub, or Azure Blob Storage. The query, which is based on SQL query language, can be used to easily filter, sort, aggregate, and join streaming data over a period. You can also extend this SQL language with JavaScript and C# user-defined functions (UDFs).

An Azure Stream Analytics job consists of an input, query, and an output. You can do the following tasks with the job output:

- Route data to storage systems like Azure Blob Storage, Azure SQL Database, Azure Data Lake Store, and Azure Cosmos DB.
- Send data to Power BI for real-time visualization.
- Store data in a Data Warehouse service like Azure Synapse Analytics to train a machine learning model based on historical data or perform batch analytics.
- Trigger custom downstream workflows by sending the data to services like Azure Functions, Azure Service Bus Topics, or Azure Queues.

Business scenario

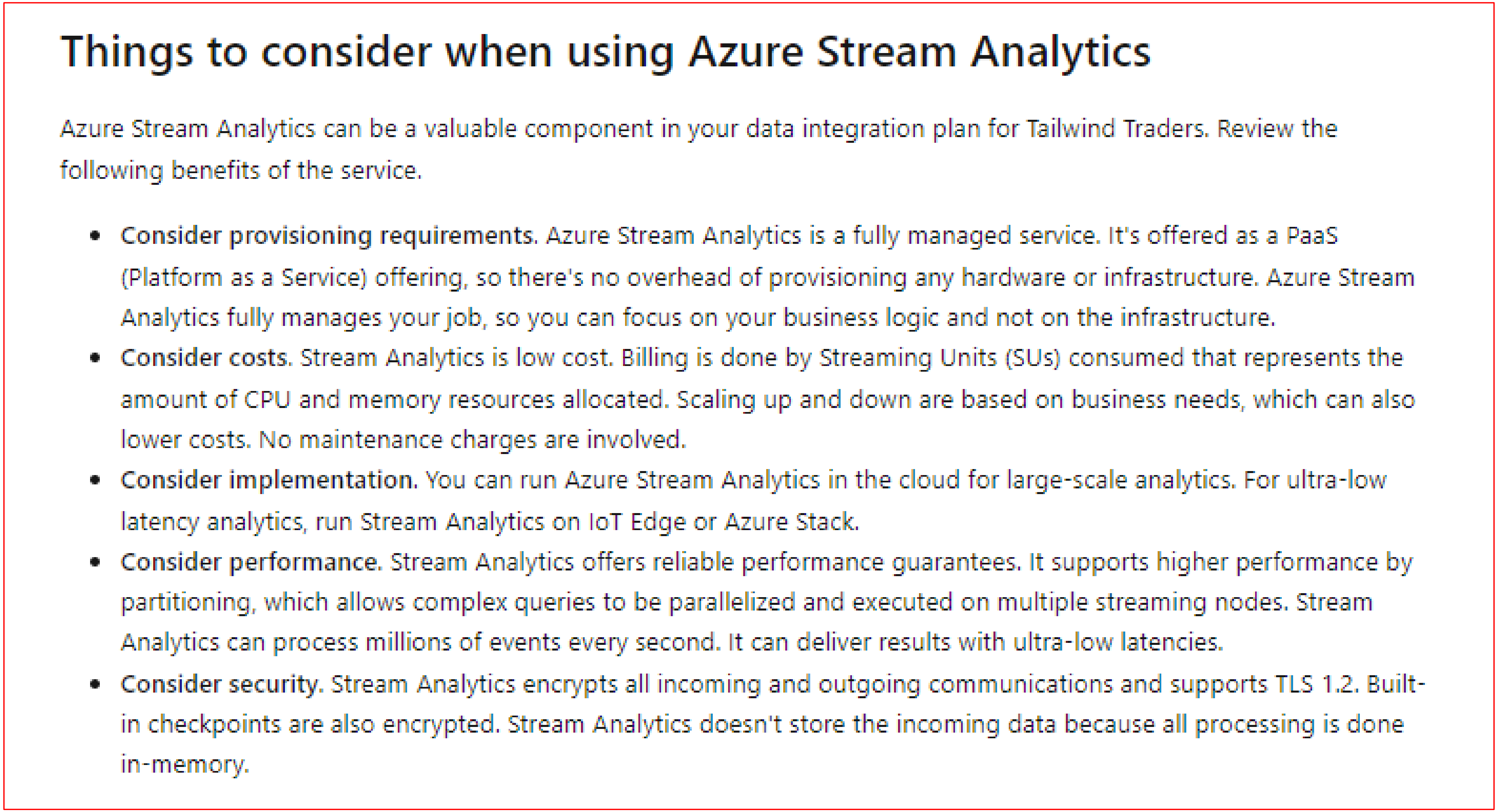


Tailwind Traders is using digital transformation for their applications and services to help with the growth of the company. They need to support accessing, storing, and analyzing sensor data from the GPS on their delivery trucks that are on the road delivering goods. You're looking for a solution to provide real time analytics on GPS streaming data from the trucks to enable administrators to make decisions in real time. On further analysis, you learn that the team would like this data present in an existing Power BI visualization dashboard. Azure Stream Analytics can help fulfill the requirements of this scenario.

Azure Stream Analytics is an ideal solution for other common enterprise data requirements. Consider the following scenarios:

Requirement	Description
Analyze real-time telemetry streams from IoT devices.	Gather real-time sensor data in Azure Stream Analytics by building automation systems that relay temperature, humidity, fan runtimes. You can make adjustments to maintain optimum building temperature and reduce costs.
Build web logs and clickstream analytics.	A consumer goods retailer can offer real-time product suggestions to users based on e-commerce analytics.
Create geospatial analytics.	Prepare analytics for geospatial data sources like sensors, social media, satellite imagery, and mobile devices. You can predict extreme weather events like wildfires and hurricanes to help airlines with routing. You can send out mobile alerts to customers for adverse weather conditions based on their geolocation.
Execute remote monitoring and predictive maintenance of high value assets.	Monitor high value assets such as Industrial equipment by gathering operational data in Azure Stream Analytics. You can maximize the useful life of your equipment through predictive maintenance. Data gathered from electrical power transformers can be used by utility companies to avoid disruption of operation.
Perform real-time analytics on point of sale data.	Detect fraudulent credit card transactions, and identify suspicious activity at point of sale. You can spot unusually large transactions or unusual location activity based on the credit card holder's contact information. Alert triggers can be set up on data gathered in Azure Stream Analytics.

In the Tailwind Traders scenario, we can apply Azure Stream Analytics to visualize real-time locations of the trucks through Power BI. For management decisions on analytical workloads, data can be stored in a data warehouse like Azure Cosmos DB or Azure Data Lake for future analysis.



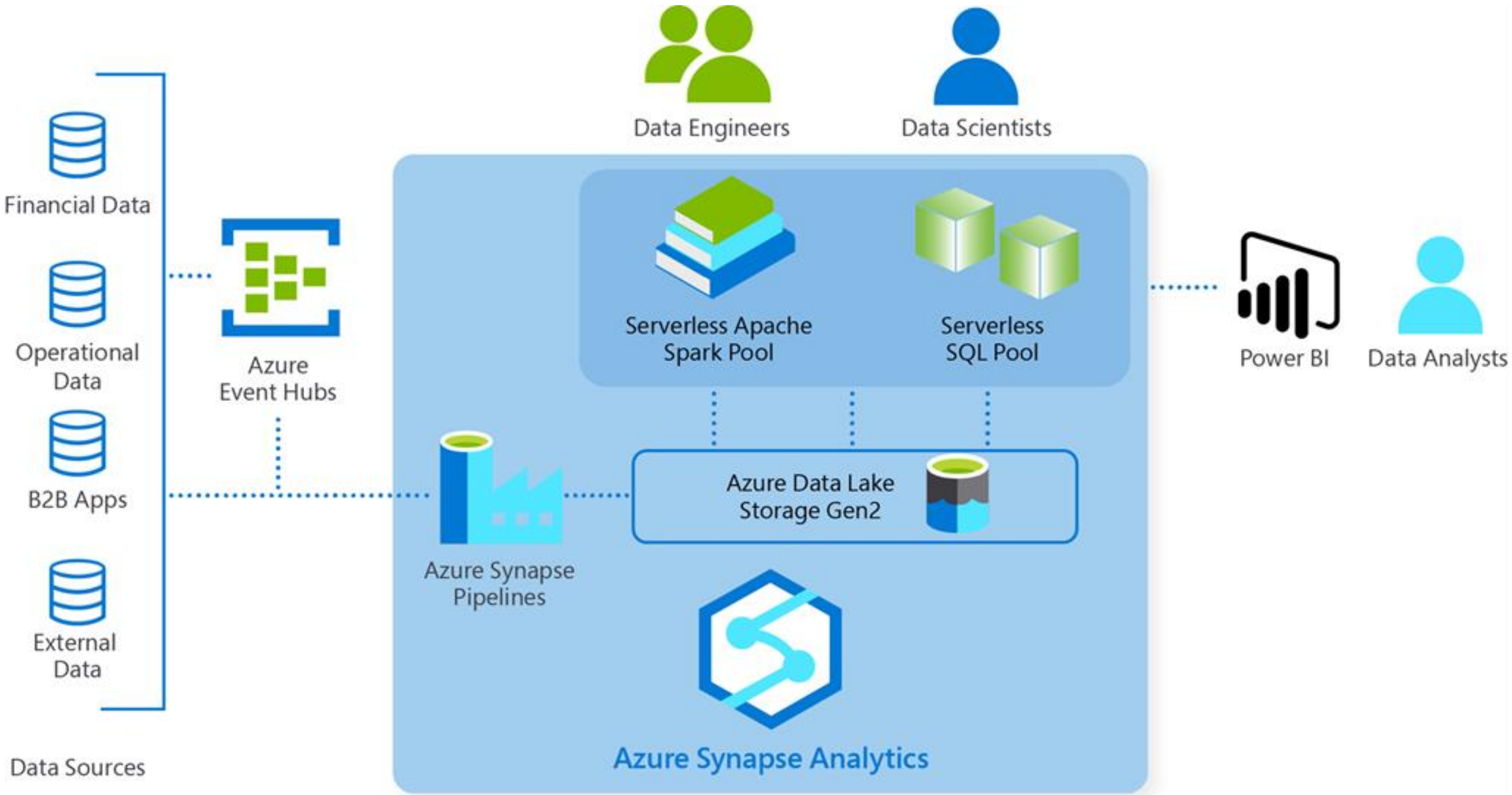
Things to consider when using Azure Stream Analytics

Azure Stream Analytics can be a valuable component in your data integration plan for Tailwind Traders. Review the following benefits of the service.

- **Consider provisioning requirements.** Azure Stream Analytics is a fully managed service. It's offered as a PaaS (Platform as a Service) offering, so there's no overhead of provisioning any hardware or infrastructure. Azure Stream Analytics fully manages your job, so you can focus on your business logic and not on the infrastructure.
- **Consider costs.** Stream Analytics is low cost. Billing is done by Streaming Units (SUs) consumed that represents the amount of CPU and memory resources allocated. Scaling up and down are based on business needs, which can also lower costs. No maintenance charges are involved.
- **Consider implementation.** You can run Azure Stream Analytics in the cloud for large-scale analytics. For ultra-low latency analytics, run Stream Analytics on IoT Edge or Azure Stack.
- **Consider performance.** Stream Analytics offers reliable performance guarantees. It supports higher performance by partitioning, which allows complex queries to be parallelized and executed on multiple streaming nodes. Stream Analytics can process millions of events every second. It can deliver results with ultra-low latencies.
- **Consider security.** Stream Analytics encrypts all incoming and outgoing communications and supports TLS 1.2. Built-in checkpoints are also encrypted. Stream Analytics doesn't store the incoming data because all processing is done in-memory.

Case study and review

Use Case: Just-in-time inventory



Review this article: https://azure.microsoft.com/blog/4-common-analytics-scenarios-to-build-business-agility/

Aggreko is a global leader in the supply of temporary power generation, temperature control systems, and energy services, providing backup energy and power supply whenever and wherever their customers need it. Aggreko uses Azure Synapse to increase operational efficiency with the just-in-time supply of their specialist equipment.

Aggreko’s data ingestion pipeline was set up to run every eight hours because it took four hours to run the ingestion (batch) jobs. Moreover, the data warehouse had to be rebuilt every day due to storage limitations. This meant that there was a lag of 8-24 hours between when the data arrived and when it was available for data analytics pipelines:

By adopting Azure Synapse, Aggreko was able to significantly improve its time-to-insight by reducing ingestion complexities and improving speed. **Ingestion time was reduced from four hours to less than five minutes**. This in turn meant that for Aggreko, data is now available for analytics pipelines in near real-time (less than five minutes’ lag). The team also estimated that they have saved 30-40 percent of their time—this time was spent solving technology problems in their legacy systems. By adopting Azure Synapse, data is now available for instant exploration, which means that the Aggreko team has more time to focus on solving business problems.

> "*Azure Synapse gives us a single environment to explore and query the data without moving it. So at a spectrum of the volume of data, we can achieve exponentially faster insights, by querying directly over the lake before outputting insight to Power BI*." —Elizabeth Hollinger, Director of Data Insights at Aggreko

As mentioned before, this use case is based on a real-world scenario where Aggreko adopted Azure Synapse as their analytics platform. To learn more about this customer story, you can watch this interview with Aggreko’s Director of Data Insights.

2 other examples: 4 common analytics scenarios to build business agility | Azure Blog and Updates | Microsoft Azure