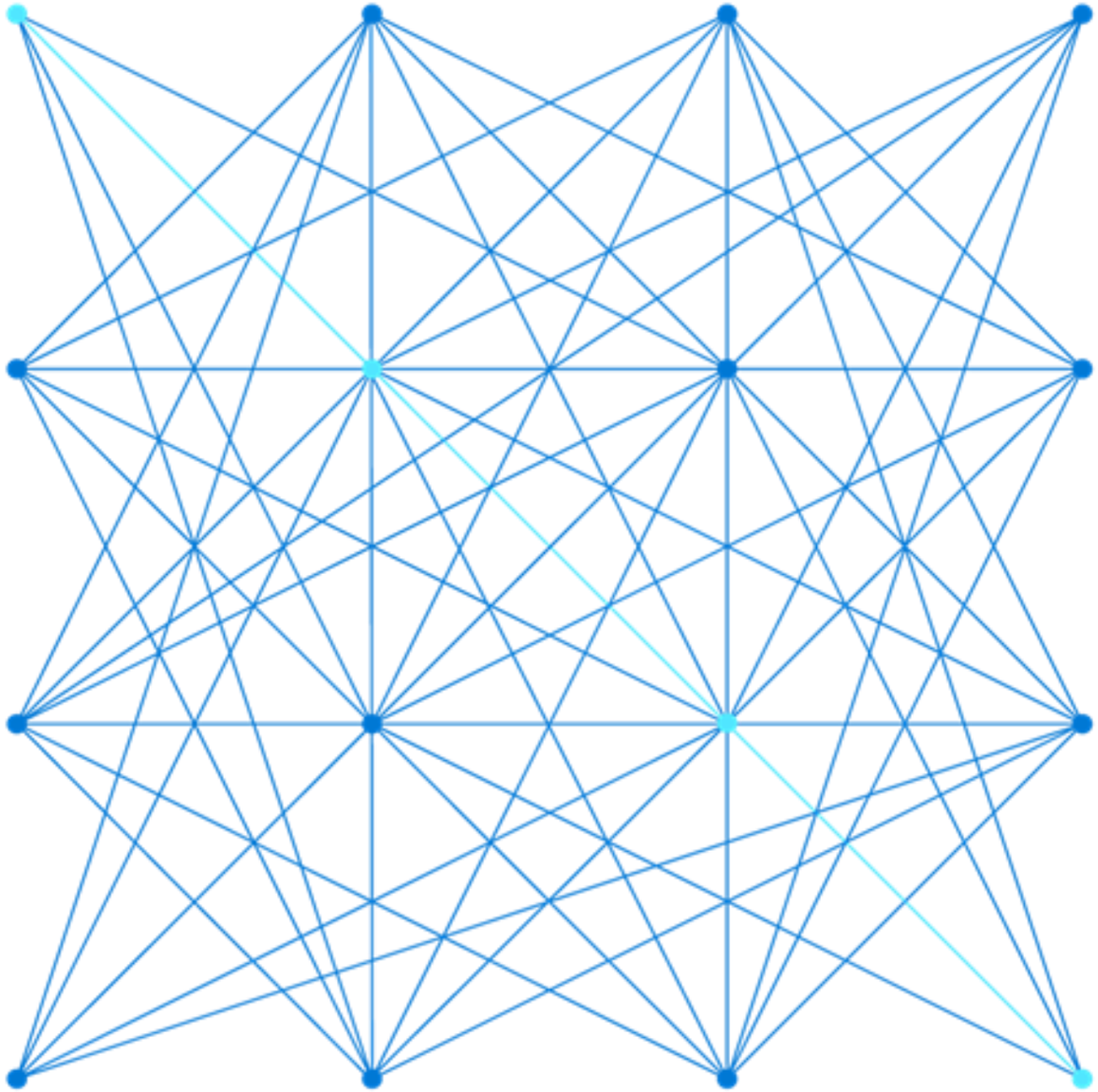
Microsoft Azure

DP-203T00: Data Engineering in Azure



- Data Scientist, Functional Analyst, formateur MCT indépendant
  PhD @ EPT: Data Warehousing, Social Media Analytics, AI, ML
  Research Msc @ ENIT, Information Processing Techniques
  Current certifications and expertise:
  PCEP – Certified Entry-Level Python Programmer
  PL-100 | PL-300 | PL-900 | AI-900 | AZ-900 | DP-300 | DP-900 | DP-100 | DP-203 | PSPO1 | PSM1
  IBM Watson Certified Application Developer V3

1. Quelles sont vos attentes de cette formation ? (Que souhaitez vous tirer de cette formation ?)
   - ✓ Découvrir surtout les concepts et les méthodes, Je suis moins intéressé par les outils.
   - ✓ Connaître globalement les concepts et les méthodes, je suis plus intéressé par les outils et les possibilités qu'ils offrent.
2. Quels sont vos objectifs à moyen et long terme en suivant cette formation ?

About this course

In this course, the student will learn about the data engineering patterns and practices as it pertains to working with batch and real-time analytical solutions using Azure data platform technologies. Students will begin by understanding the core compute and storage technologies that are used to build an analytical solution. They will then explore how to design an analytical serving layers and focus on data engineering considerations for working with source files.

The students will learn how to interactively explore data stored in files in a data lake. They will learn the various ingestion techniques that can be used to load data using the Apache Spark capability found in Azure Synapse Analytics or Azure Databricks, or how to ingest using Azure Data Factory or Azure Synapse pipelines. The students will also learn the various ways they can transform the data using the same technologies that is used to ingest data.

The student will spend time on the course learning how to monitor and analyze the performance of analytical system so that they can optimize the performance of data loads, or queries that are issued against the systems. They will understand the importance of implementing security to ensure that the data is protected at rest or in transit.

The student will then show how the data in an analytical system can be used to create dashboards or build predictive models in Azure Synapse Analytics.

Audience

Primary audience:

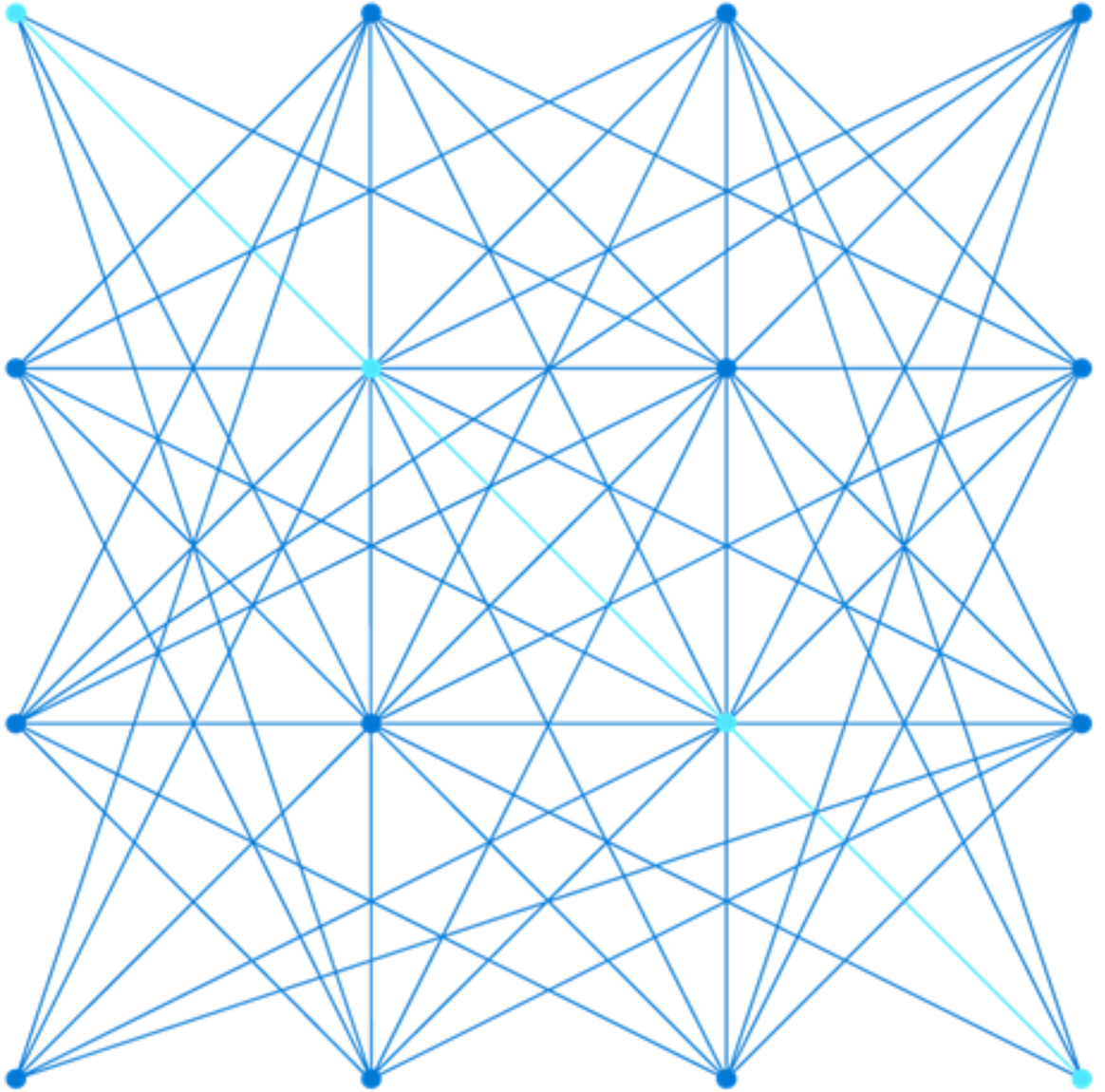
The audience for this course are data engineers, data professionals, data architects, and business intelligence professionals who want to learn about the data platform technologies that exist on Microsoft Azure that can be used to perform data engineering and storage for analytical solutions.

Secondary audience:

The secondary audience for this course are individuals who develop applications that deliver content from the data platform technologies that exist on Microsoft Azure.

Microsoft Azure

DP-203T00: Explore compute and storage options for data engineering workloads



Agenda

Lesson 01 – Introduction to Azure Synapse Analytics

Lesson 02 – Describe Azure Databricks

Lesson 03 – Introduction to Azure Data Lake storage

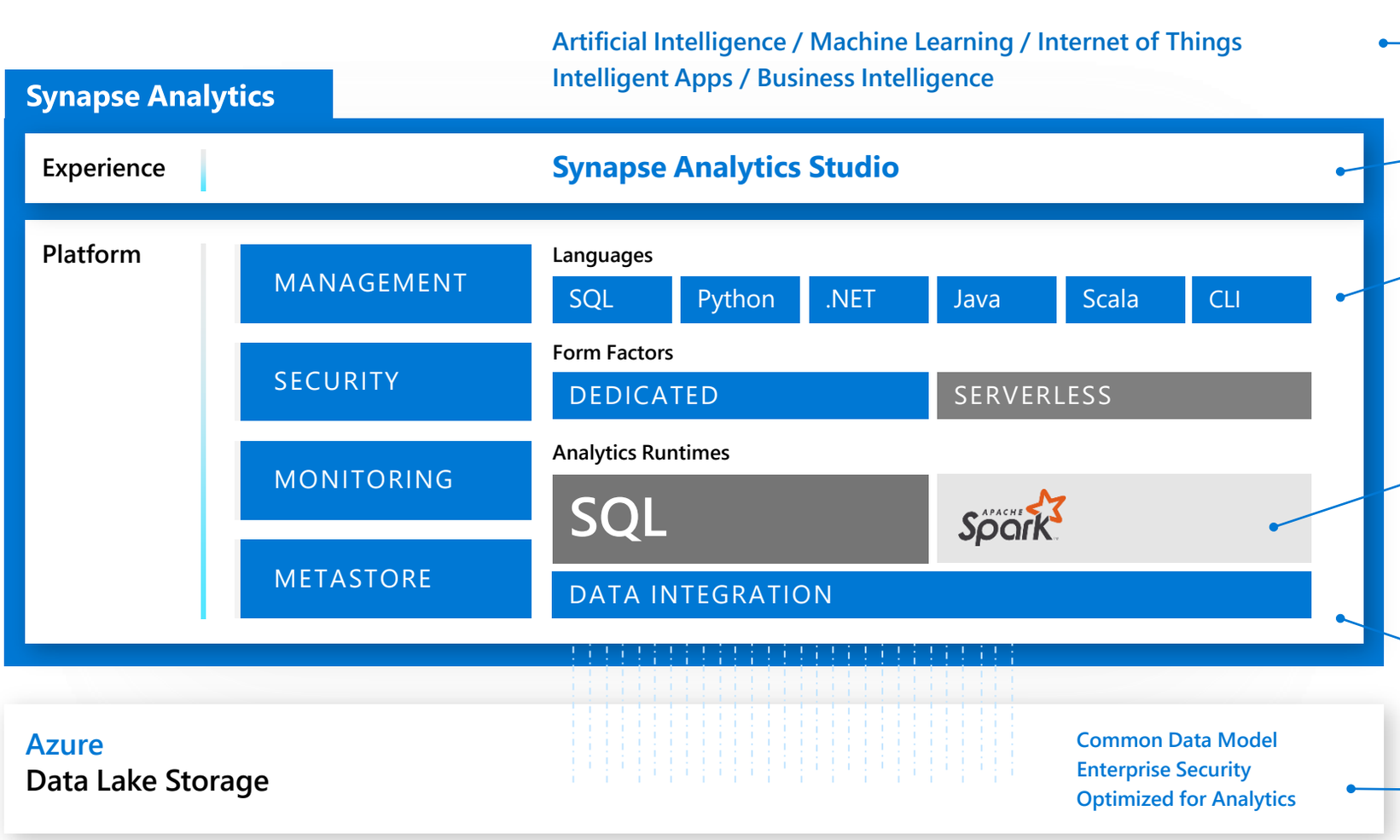
Lesson 04 – Describe Delta Lake architecture

Lesson 05 – Work with data streams by using Azure Stream Analytics

Lesson 01: Introduction to Azure Synapse Analytics

Azure Synapse Analytics

Limitless analytics service with unmatched time to insight



Designed for analytics **workloads at any scale**

SaaS **developer experiences** for code free and code first

Multiple **languages** suited to different analytics workloads

Integrated analytics runtimes available as a dedicated and serverless offering

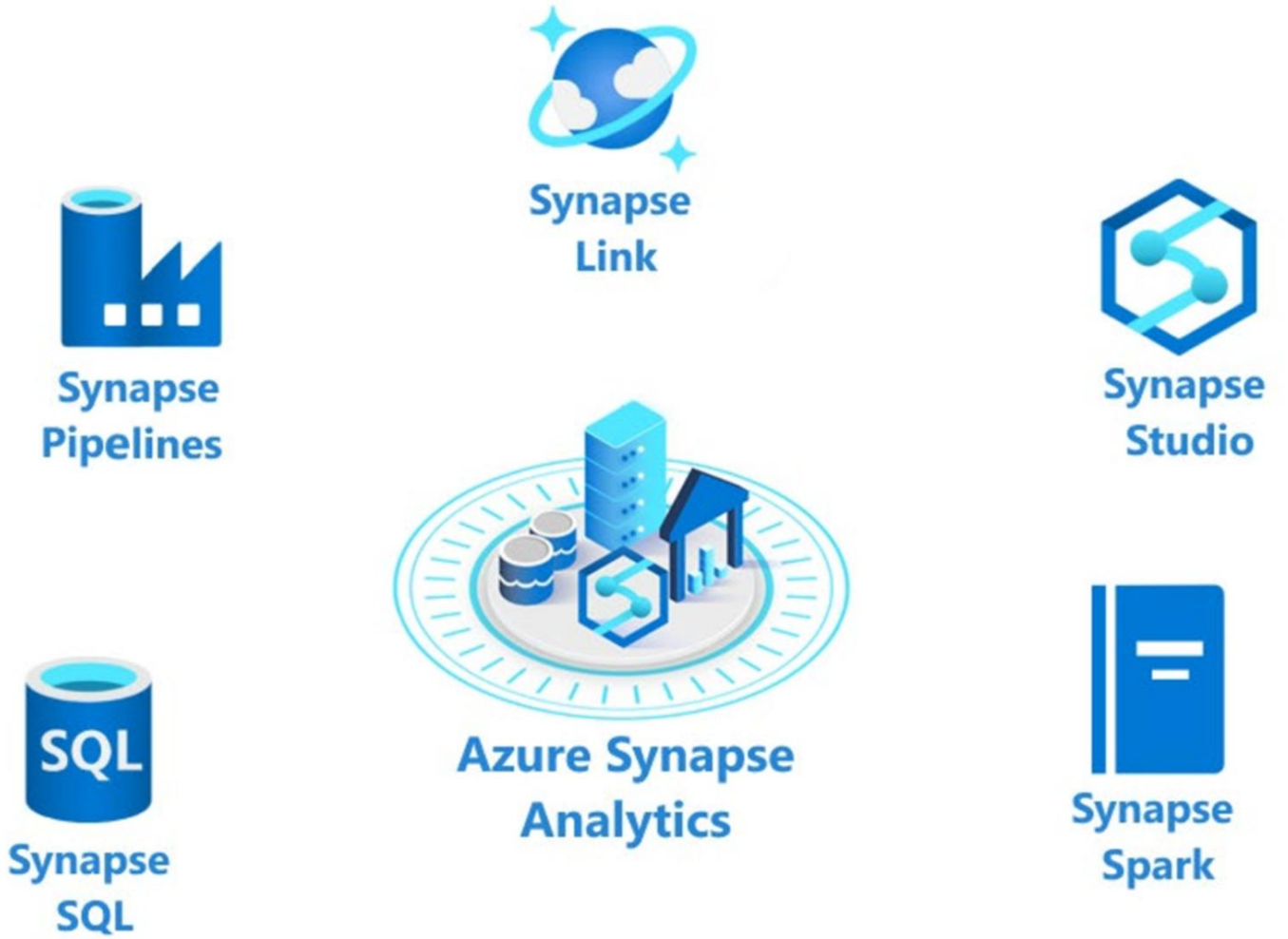
SQL pools offering T-SQL for batch, streaming and interactive processing

Spark pools for big data processing with Python, Scala, R and .NET

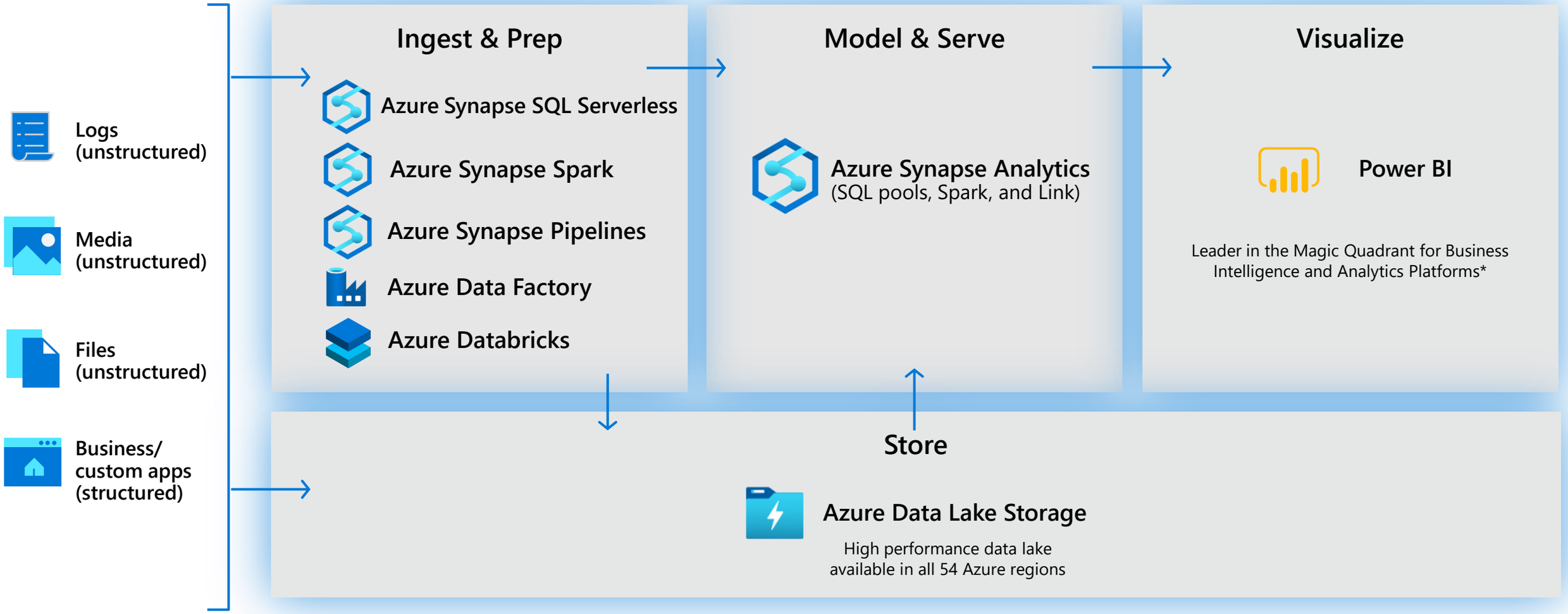
Integrated **platform services** for, management, security, monitoring, and metastore

Data **lake integrated** and Common Data Model aware

Introduction to Azure Synapse Analytics



Modern data warehousing pattern with Azure Synapse Analytics



Lesson 02: Describe Azure Databricks

What is Azure Databricks

Apache Spark-based analytics platform:

Simplifies the provisioning and collaboration of Apache Spark-based analytical solutions dealing with batch and streaming data

Comprehensive Spark library support:

Support includes SQL, DataFrames, MLlib, Hyperspace and MSSparkUtil

Enterprise Security:

Utilizes the security capabilities of Azure

Integration with other Cloud Services:

Can integrate with a variety of Azure data platform services and Power BI

What is Apache Spark

Apache Spark emerged to provide a parallel processing framework that supports in-memory processing to boost the performance of big-data analytical applications on massive volumes of data

Interactive Data Analysis:
Used by business analysts or data engineers to analyze and prepare data

Streaming Analytics:
Ingest data from technologies such as Kafka and Flume to ingest data in real-time

Machine Learning:
Contains a number of libraries that enables a Data Scientist to perform Machine Learning

Why use Azure Databricks?

Azure Databricks is a wrapper around Apache Spark that simplifies the provisioning and configuration of a Spark cluster in a GUI interface

Azure Databricks components:

Spark SQL and DataFrames

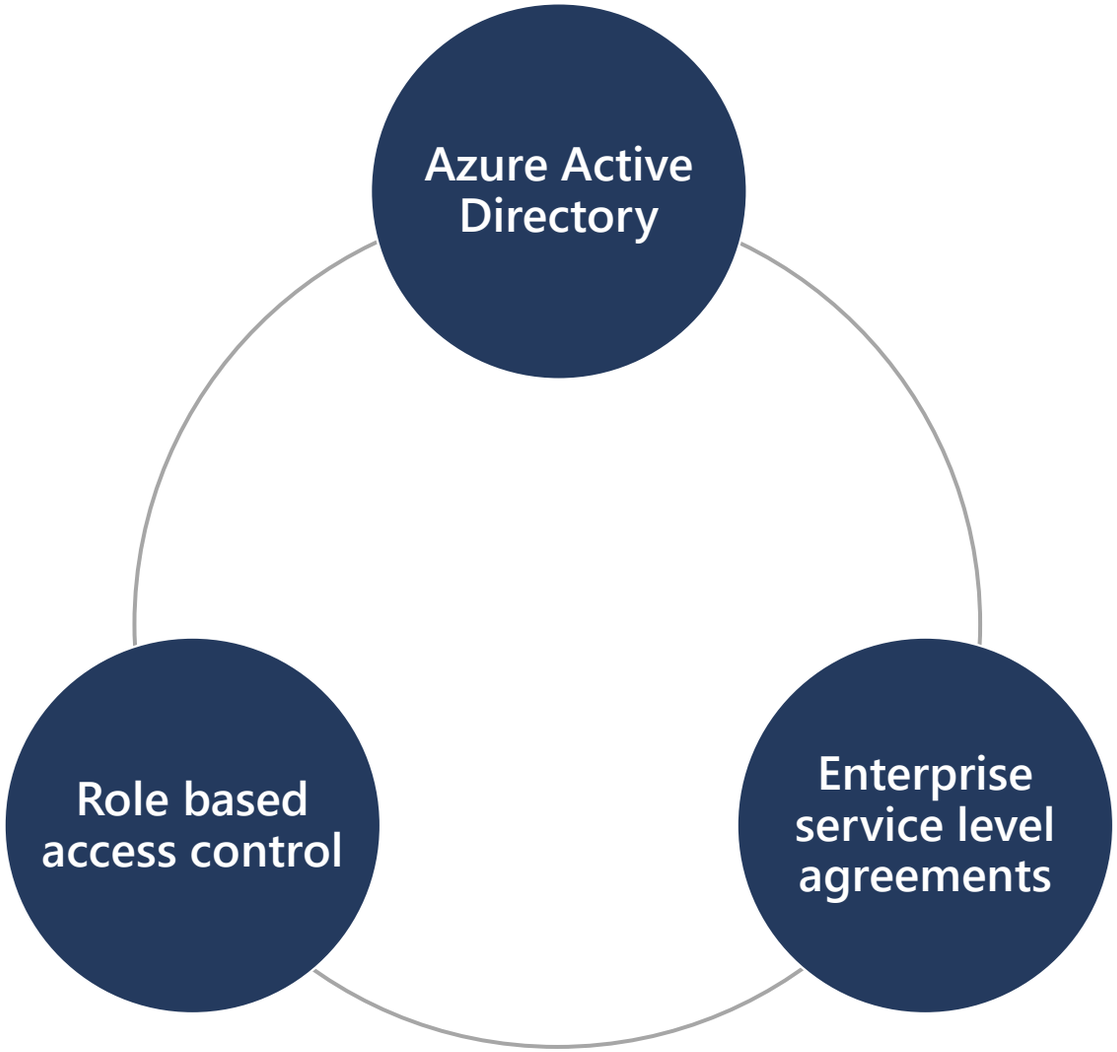
Streaming

Mlib

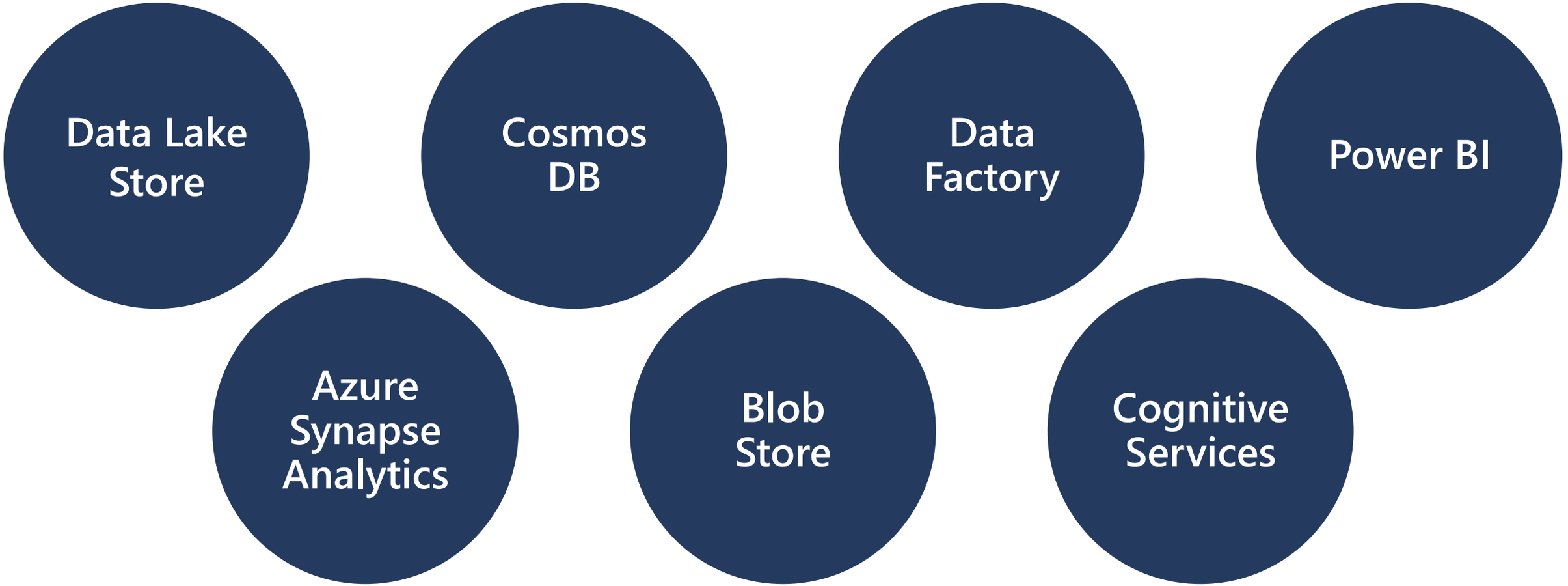
GraphX

Spark Core API

Enterprise security



Integration with cloud services

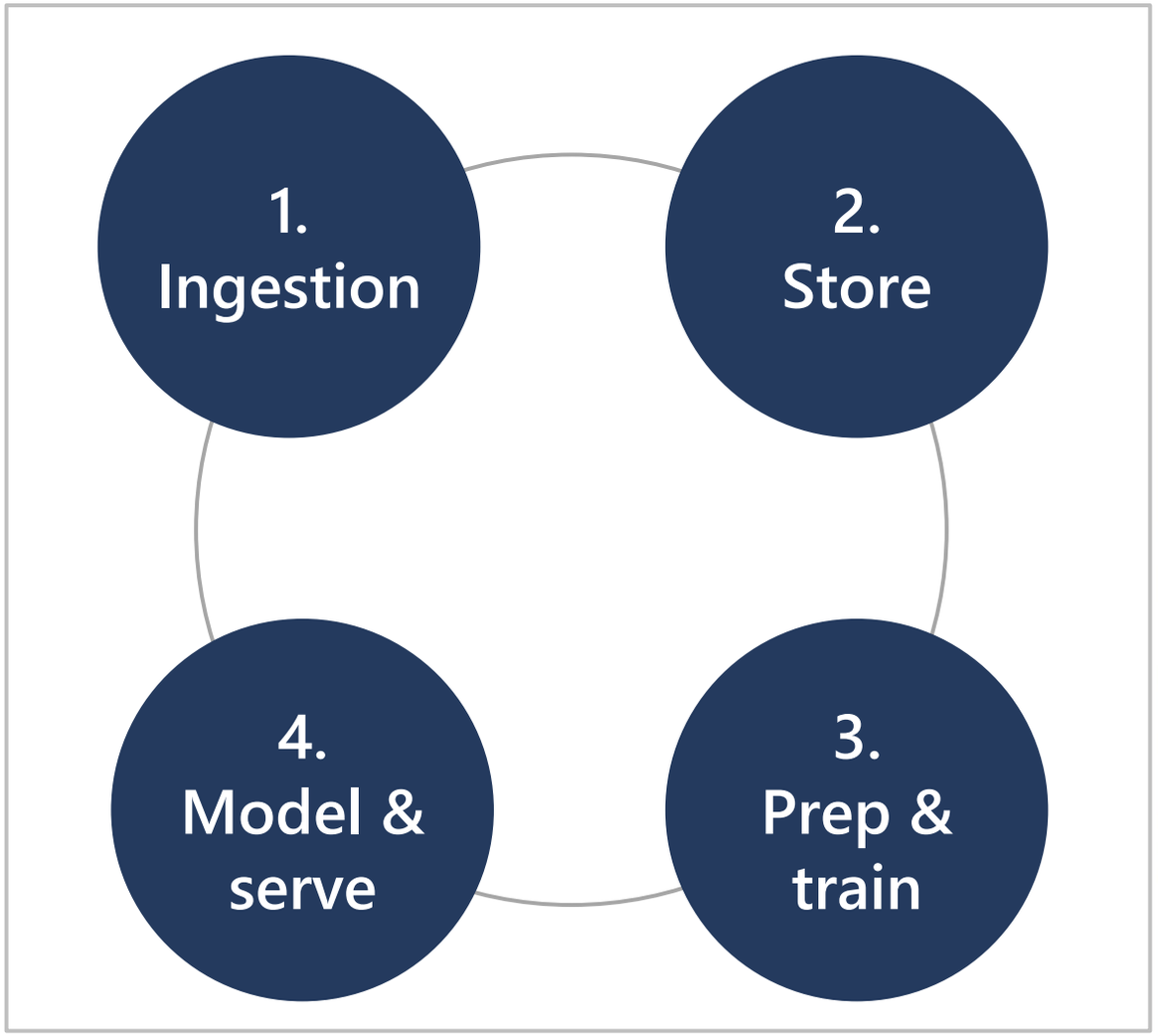


Spark: what to use when and where

	Apache Spark	HDInsight	Azure Databricks	Synapse Spark
WHAT	Is an Open Source memory optimized system for managing big data workloads	Microsoft implementation of Open Source Spark managed within the realms of Azure	A managed Spark as a Service solution	Embedded Spark capability within Azure Synapse Analytics
WHEN	When you want to benefits of spark for big data processing and/or data science work without the Service Level Agreements of a provider	When you want to benefits of OSS spark with the Service Level Agreement of a provider	Provides end to end data engineering and data science solution and management platform	Enables organizations without existing Spark implementations to fire up a Spark cluster to meet data engineering needs without the overheads of the other Spark platforms listed
WHO	Open Source Professionals	Open Source Professionals wanting SLA's and Microsoft Data Platform experts	Data Engineers and Data Scientists working on big data projects every day	Data Engineers, Data Scientists, Data Platform experts and Data Analysts
WHY	To overcome the limitations of SMP systems imposed on big data workloads	To take advantage of the OSS Big Data Analytics platform with SLA's in place to ensure business continuity	It provides the ability to create and manage an end to end big data/data science project using one platform	It provides the ability to scale efficiently with spark clusters within a one stop shop Data Warehousing platform of Synapse.

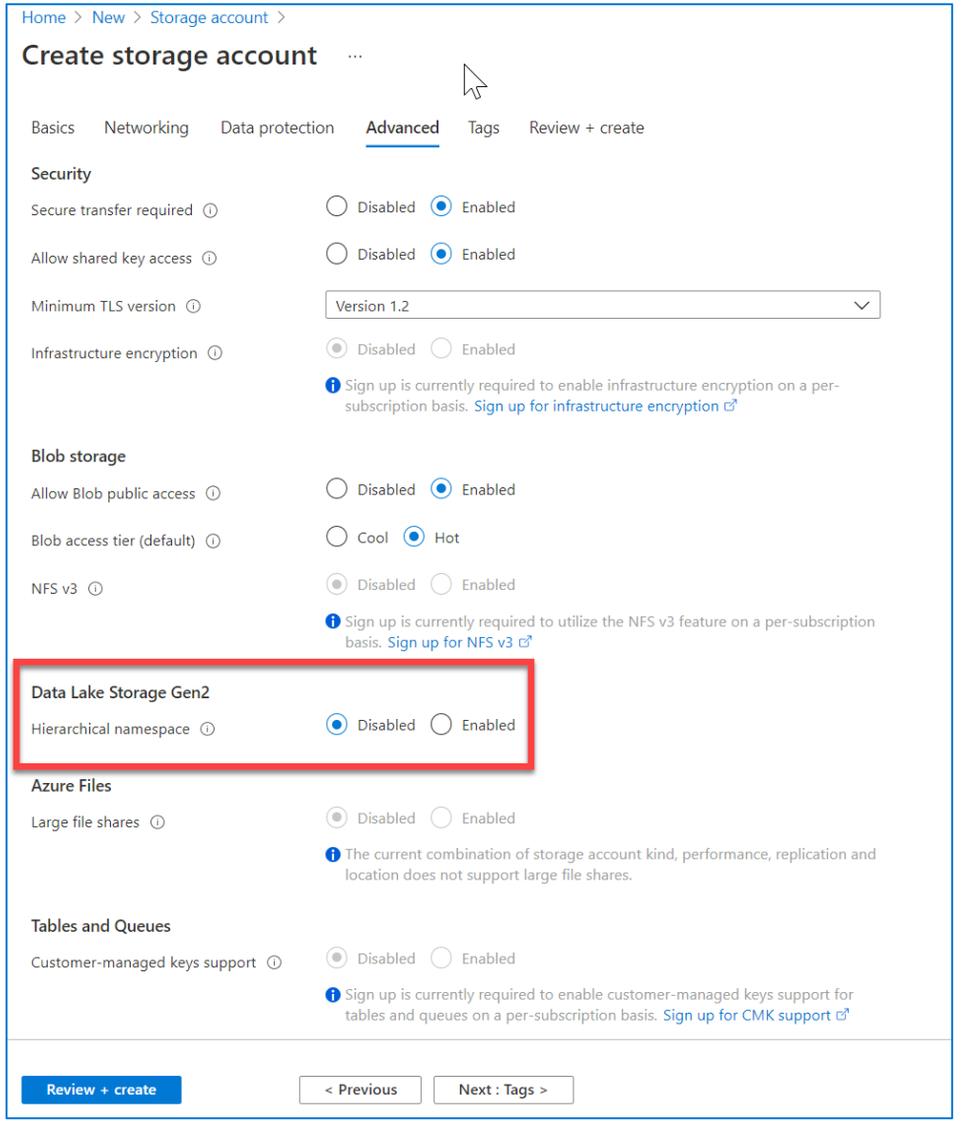
Lesson 03: Introduction to Azure Data Lake storage

Processing Big Data with Azure Data Lake Store

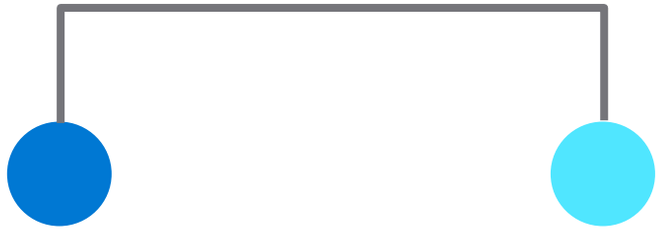
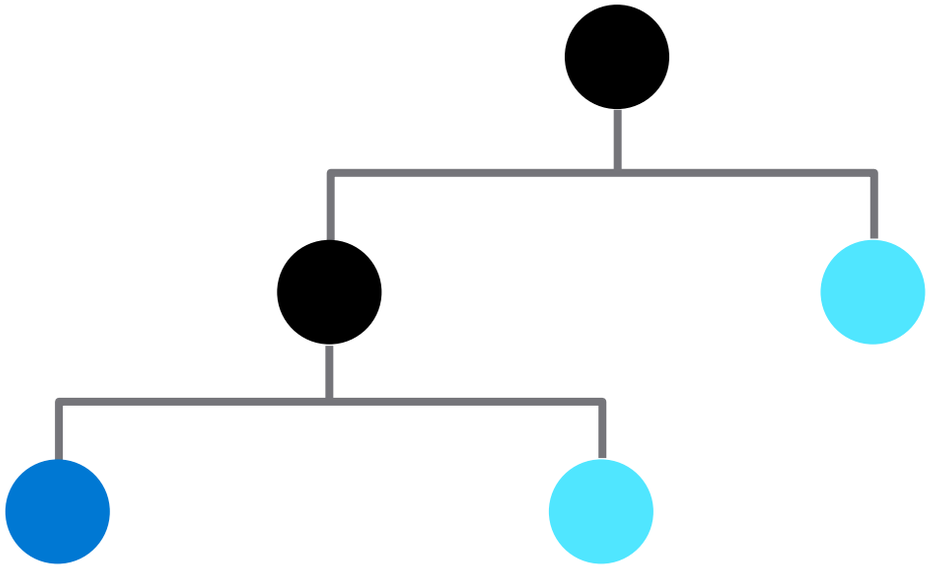


Introduction to Azure Data Lake storage

- Hadoop compatible
- Security
- Performance
- Redundancy



Compare Azure Blob Storage and Data Lake Store Gen 2

Azure Blob Flat namespace	Data Lake (Gen II) Hierarchical namespace
	

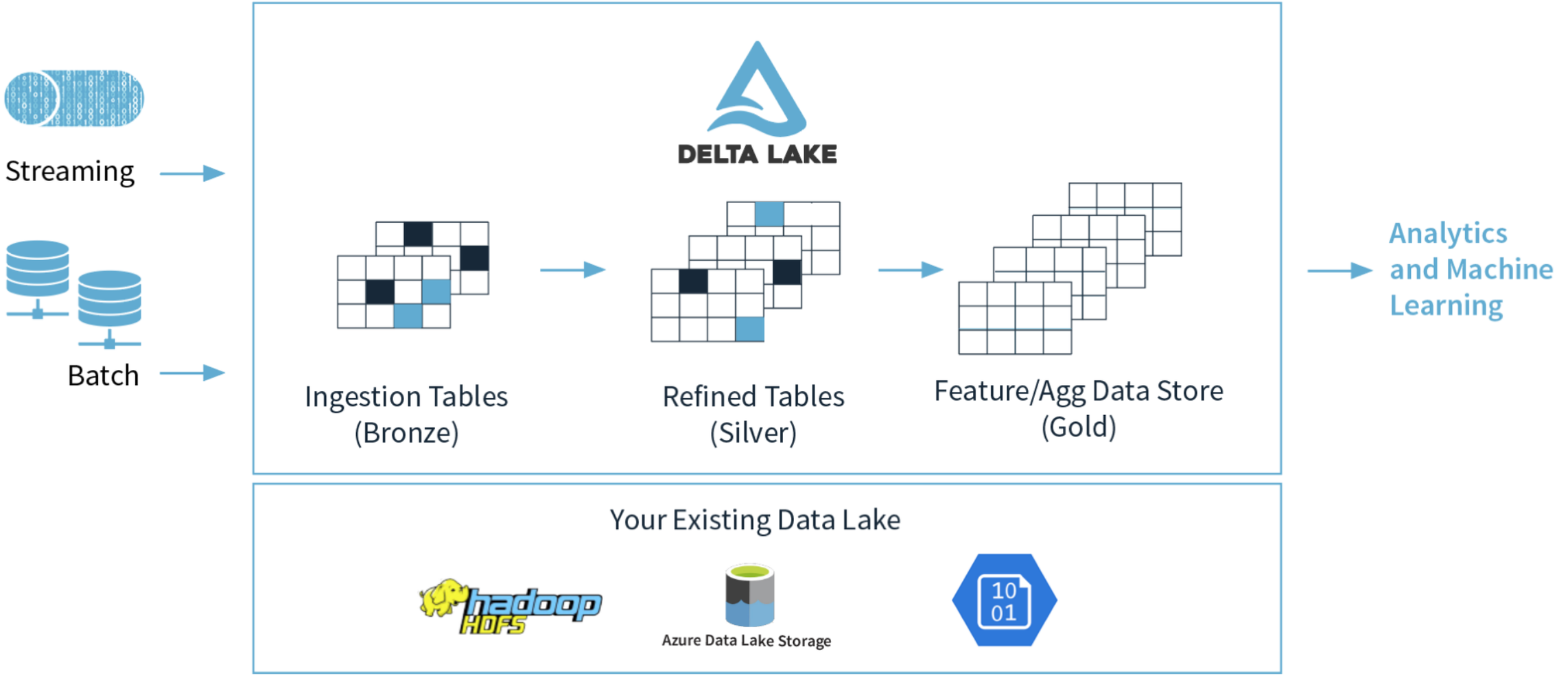
Big Data use cases

Let’s examine three use cases for leveraging an Azure Data Lake Store

Modern data warehouse	This architecture sees Azure Data Lake Storage at the heart of the solution for a modern data warehouse. Using Azure Data Factory to ingest data into the Data Lake from a business application, and predictive models built in Azure Databricks, using Azure Synapse Analytics as a serving layer
Advanced analytics	In this solution, Azure Data factory is transferring terabytes of web logs from a web server to the Data Lake on an hourly basis. This data is provided as features to the predictive model in Azure Databricks, which is then trained and scored. The result of the model is then distributed globally using Azure Cosmos DB, that an application uses
Real time analytics	In this architecture, there are two ingestion streams. Azure Data Factory is used to ingest the summary files that are generated when the HGV engine is turned off. Apache Kafka provides the real-time ingestion engine for the telemetry data. Both data streams are stored in Data Lake store for use in the future

Lesson 04: Describe Delta Lake architecture

Describe a Delta Lake architecture



Lesson 05: Work with data streams by using Azure Stream Analytics

What are data streams

Data streams:

In the context of analytics, data streams are event data generated by sensors or other sources that can be analyzed by another technology

Data stream processing approach:

There are two approaches. Reference data is streaming data that can be collected over time and persisted in storage as static data. In contrast, streaming data have relatively low storage requirements. And run computations in sliding windows

Data streams are used to:

Analyze data:
Continuously analyze data to detect issues and understand or respond to them

Understand systems:
Understand component or system behavior under various conditions to fuel further enhancements of said system

Trigger actions:
Trigger specific actions when certain thresholds are identified

Event processing

The process of consuming data streams, analyzing them, and deriving actionable insights out of them is called Event Processing and has three distinct components:

Event producer	Examples include sensors or processes that generate data continuously such as a heart rate monitor or a highway toll lane sensor
Event processor	An engine to consume event data streams and deriving insights from them. Depending on the problem space, event processors either process one incoming event at a time (such as a heart rate monitor) or process multiple events at a time (such as a highway toll lane sensor)
Event consumer	An application which consumes the data and takes specific action based on the insights. Examples of event consumers include alert generation, dashboards, or even sending data to another event processing engine

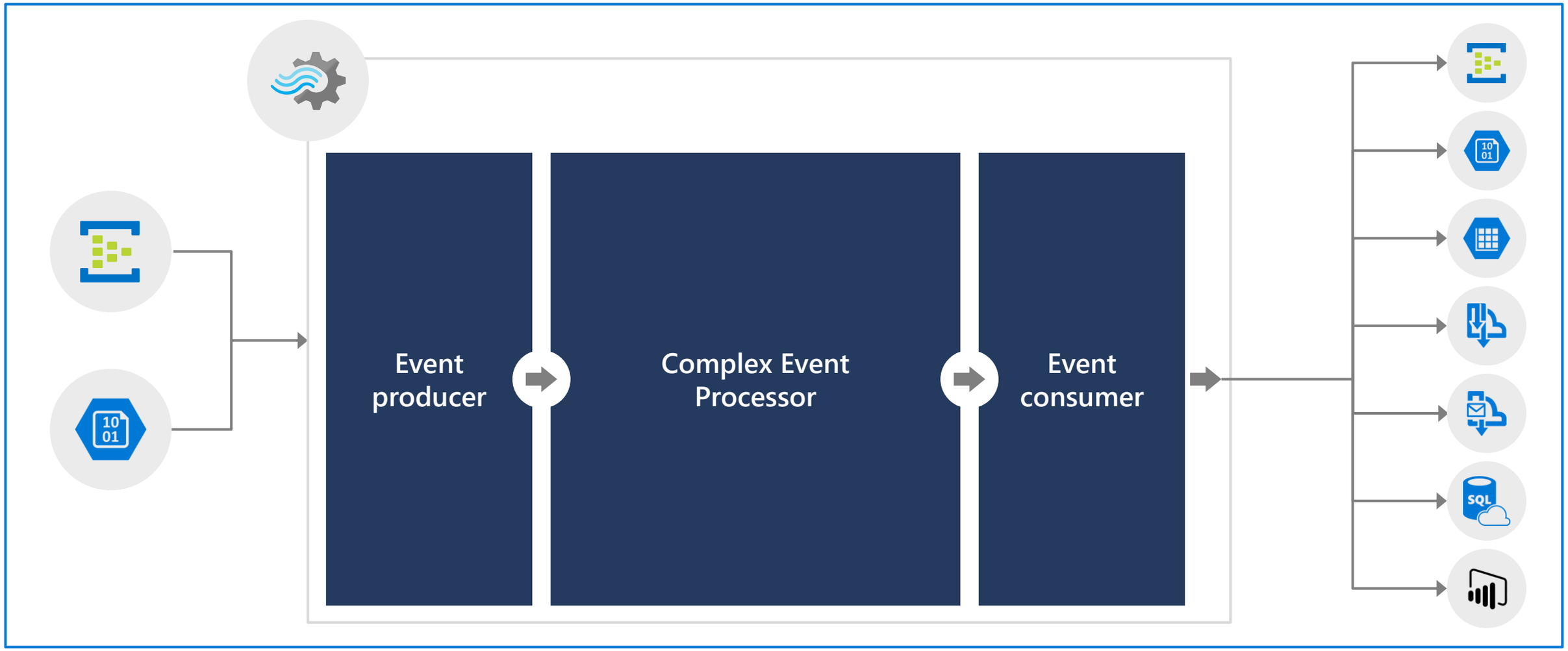
Processing events with Azure Stream Analytics

Microsoft Azure Stream Analytics is an event processing engine. It enables the consumption and analysis of high volumes of streaming data in real time

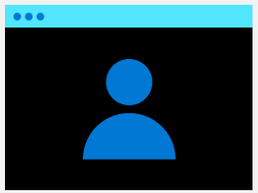
Source	Ingestion	Analytical engine	Destination
Sensors	Event Hubs	Stream Analytics Query Language	Azure Data Lake
Systems	IoT Hubs	.NET SDK	Cosmos DB
Applications	Azure Blob Store		SQL Database
			Blob Store
			Power BI

Work with data streams by using Azure Stream Analytics

Complex event processing of Stream Data in Azure



Lab: Explore compute and storage options for data engineering workloads



Lab overview

This lab teaches ways to structure the data lake, and to optimize the files for exploration, streaming, and batch workloads. The student will learn how to organize the data lake into levels of data refinement as they transform files through batch and stream processing. The students will also experience working with Apache Spark in Azure Synapse Analytics. They will learn how to create indexes on their datasets, such as CSV, JSON, and Parquet files, and use them for potential query and workload acceleration using Spark libraries including Hyperspace and MSSParkUtils.

Lab objectives

After completing this lab, you will be able to:

Work with a Delta Lake architecture

Working with Apache Spark in Azure Synapse Analytics