



# DP-200T01: Implementing an Azure Data Solution



# Tissana Tanaklang

**Software and Solution Development Trainer**

Iverson Training Center Co., Ltd.

[tissana\\_t@hotmail.com](mailto:tissana_t@hotmail.com)

- Master of Science Program in Software Engineering  
King Mongkut's University of Technology Thonburi
- Bachelor of Science Program in Computer Science  
Naresuan University
- Microsoft Certified Solutions Associate (MCSA) - Web Application Development
- Microsoft Certified Azure Data Fundamentals
- Microsoft Certified Azure Fundamentals
- Microsoft Certified Trainer (MCT)

# Azure Learning Path

Level	Category	Code	Course	Role
Beginner (Fundamentals)	-	AZ-900	Microsoft Azure Fundamentals	IT Professional and Non-IT Professional (All)
	Data	DP-900	Microsoft Azure Data Fundamentals	Data Engineer, Database Administrator
	AI	AI-900	Microsoft Azure AI Fundamentals	AI Engineer, Data Scientist, Developer, Solutions Architect
Intermediate (Associate)	DevOps	AZ-104	Microsoft Azure Administrator	Administrator, DevOps Engineer
		AZ-204	Developing solutions for Microsoft Azure	Developer, DevOps Engineer
	Security	AZ-500	Microsoft Azure Security Technologies	Security Engineer
	Data	DP-300	Administering Relational Databases on Microsoft Azure	Database Administrator
		DP-200	Implementing an Azure Data Solution	Data Engineer
		DP-201	Designing an Azure Data Solution	
		DP-100	Designing and Implementing a Data Science Solution on Azure	Data Scientist
	AI	AI-100	Designing and Implementing an Azure AI Solution	AI Engineer
Advance (Expert)	DevOps	AZ-400	Designing and Implementing Microsoft DevOps solutions	DevOps Engineer
	Solutions Architect	AZ-303	Microsoft Azure Architect Technologies	Solutions Architect
		AZ-304	Microsoft Azure Architect Design	
Specialty	Data	DA-100	Analyzing Data with Power BI	Data Analyst
	-	AZ-220	Microsoft Azure IoT Developer	Developer



# Agenda

- About this course
- Course agenda
- Audience
- Prerequisites

# About this course

In this course, the students will implement various data platform technologies into solutions that are in line with business and technical requirements including on-premises, cloud, and hybrid data scenarios incorporating both relational and No-SQL data. They will also learn how to process data using a range of technologies and languages for both streaming and batch data.

The students will also explore how to implement data security including authentication, authorization, data policies and standards. They will also define and implement data solution monitoring for both the data storage and data processing activities. Finally, they will manage and troubleshoot Azure data solutions which includes the optimization and disaster recovery of big data, batch processing and streaming data solutions.

# Course Agenda

- Module 1
  - Azure for the Data Engineer
    - L01 - Explain the evolving world of data
    - L02 - Survey the services in the Azure Data Platform
    - L03 - Identify the tasks that are performed by a Data Engineer
    - L04 - Describe the use cases for the cloud in a case study
- Module 2
  - Working with Data Storage
    - L01 - Choose a data storage approach in Azure
    - L02 - Create an Azure Storage Account
    - L03 - Explain Azure Data Lake Storage
    - L04 - Upload data into Azure Data Lake

# Course Agenda (*continued #1*)

- Module 3
- Enabling Team Based Data Science with Azure Databricks
  - L01 - Explain Azure Databricks
  - L02 - Work with Azure Databricks
  - L03 - Read data with Azure Databricks
  - L04 - Perform transformations with Azure Databricks
- Module 4
- Building Globally Distributed Databases with Cosmos DB
  - L01 - Create an Azure Cosmos DB database built to scale
  - L02 - Insert and query data in your Azure Cosmos DB database
  - L03 - Build a .NET Core app for Azure Cosmos DB in Visual Studio Code
  - L04 - Distribute your data globally with Azure Cosmos DB

# Course Agenda (*continued #2*)

- Module 5
- Working with Relational Data Stores in the Cloud
  - L01 - Explain SQL Database
  - L02 - Explain SQL Data Warehouse
  - L03 - Provision and load data in Azure SQL Data Warehouse
  - L04 - Import data into Azure SQL Data Warehouse using PolyBase
- Module 6
- Performing Real-Time Analytics with Stream Analytics
  - L01 - Explain data streams and event processing
  - L02 - Data Ingestion with Event Hubs
  - L03 - Processing Data with Stream Analytics Jobs

# Course Agenda (*continued* #3)

- Module 7
  - Orchestrating Data Movement with Azure Data Factory
    - L01 - Explain how Azure Data Factory works
    - L02 - Create Linked Services and Datasets
    - L03 - Create Pipelines and Activities
    - L04 - Azure Data Factory pipeline execution and triggers
- Module 8
  - Securing Azure Data Platforms
    - L01 - Introduction to Security
    - L02 - Key Security Components
    - L03 - Securing Storage Accounts and Data Lake Storage
    - L04 - Security Data Stores

# Course Agenda (*continued #4*)

- Module 9
- Monitoring and Troubleshooting Data Storage and Processing
  - L01 - Explain the monitoring capabilities that are available
  - L02 - Troubleshoot common data storage issues
  - L03 - Troubleshoot common data processing issues
  - L04 - Manage disaster recovery

# Audience

## Primary audience

The audience for this course are data professionals, data architects, and business intelligence professionals who want to learn about the data platform technologies that exist on Microsoft Azure.

## 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.

# Prerequisites

In addition to their professional experience, students who take this training should have technical knowledge equivalent to the following courses:

[Azure fundamentals](#)



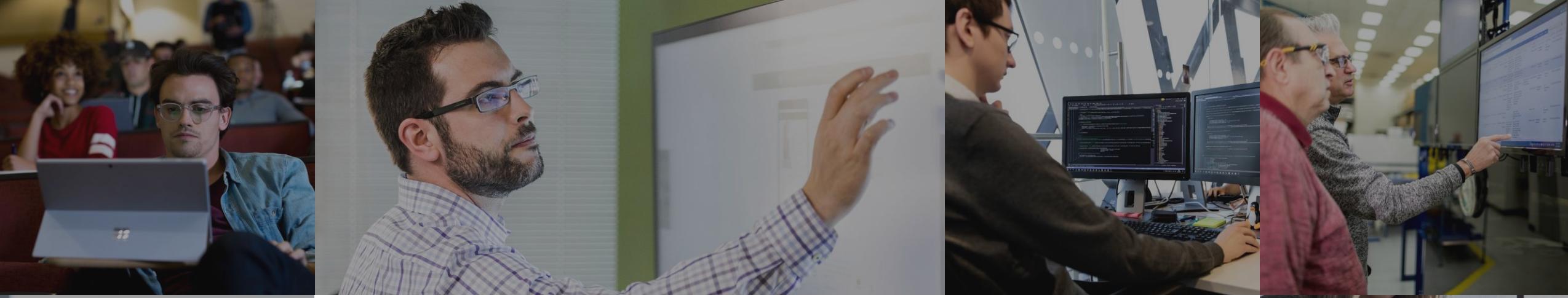
# Module 01:

## Azure for the Data Engineer



# Agenda

- L01 – Explain the evolving world of data
- L02 - Survey the services in the Azure Data Platform
- L03 - Identify the tasks that are performed by a Data Engineer
- L04 - Describe the use cases for the cloud in a case study



# Lesson 01

## The Evolving World of Data



# Lesson Objectives

- Data abundance
- Differences between on-premises and cloud data technologies
- How the role of the data professional is changing in organizations
- Identify use cases impacted by these changes

Data abundance

## Processes

Businesses are tasked to store, interpret, manage, transform, process, aggregate and report on data

## Consumers

There are a wider range of consumers using different types of devices to consume or generate data

## Variety

There's a wider variety of data types that need to be processed and stored

## Responsibilities

A data engineer's role is responsible for more data types and technologies

## Technologies

Microsoft Azure provides a wide set of tools and technologies

# On-premises versus cloud technologies



A large, semi-transparent background image of a city skyline, likely New York City, is visible against a blue sky with scattered white clouds.

Computing Environment



Licensing Model



Maintainability



Scalability

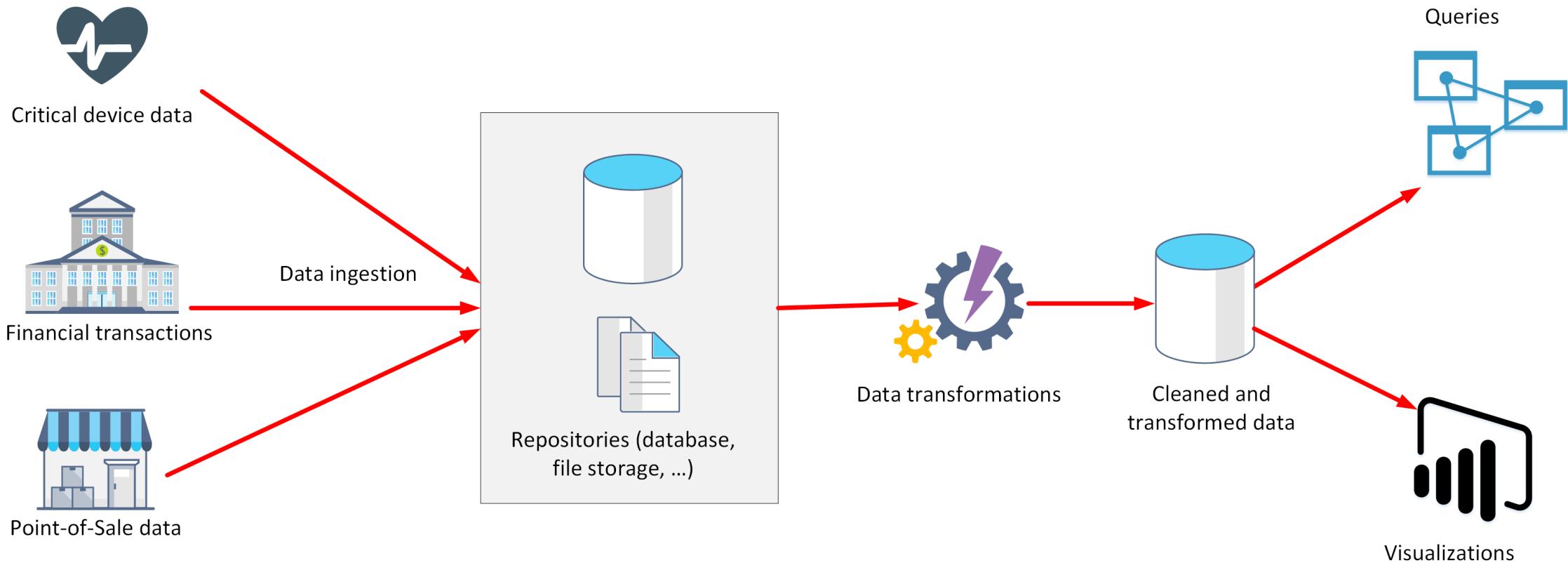


Availability



# Data engineering job responsibilities





# Use cases for the cloud

Here are some examples of industries making use of the cloud

Web retail

Using Azure Cosmos DB's multi-master replication model along with Microsoft's performance commitments, Data Engineers can implement a data architecture to support web and mobile applications that achieve less than a 10-ms response time anywhere in the world

Healthcare

Azure Databricks can be used to accelerate big data analytics and artificial intelligence (AI) solutions. Within the healthcare industry, it can be used to perform genome studies or pharmacy sales forecasting at petabyte scale

IoT scenarios

Hundreds of thousands of devices have been designed and sold to generate sensor data known as Internet of Things (IoT) devices. Using technologies like Azure IoT Hub, Data Engineers can easily design a data solution architecture that captures real-time data



# Lesson 02

## Survey the Services in the Azure Data Platform

# Lesson Objectives

- The differences between structured and unstructured data
- Azure Storage
- Azure Data Lake Storage
- Azure Databricks
- Azure Cosmos DB
- Azure SQL Database
- Azure SQL Data Warehouse
- Azure Stream Analytics
- Additional Azure Data Platform Services

	Schema	Data relationships	Examples
<b>Structured data</b>	Adheres to a schema, with the same data fields or properties.	Storable in relational database tables, with rows and columns.	Sensor data and financial data.
<b>Semi-structured data</b>	Has an ad hoc schema with less organized fields and properties.	Non-relational or NoSQL data, not storable in tables, rows and column.	Books, blogs, JSON, HTML documents.
<b>Unstructured data</b>	Has no designated schema or data structure.	Non-relational or blob data, with no restrictions on the kinds of data blobs contain.	PDFs, JPGs, videos.

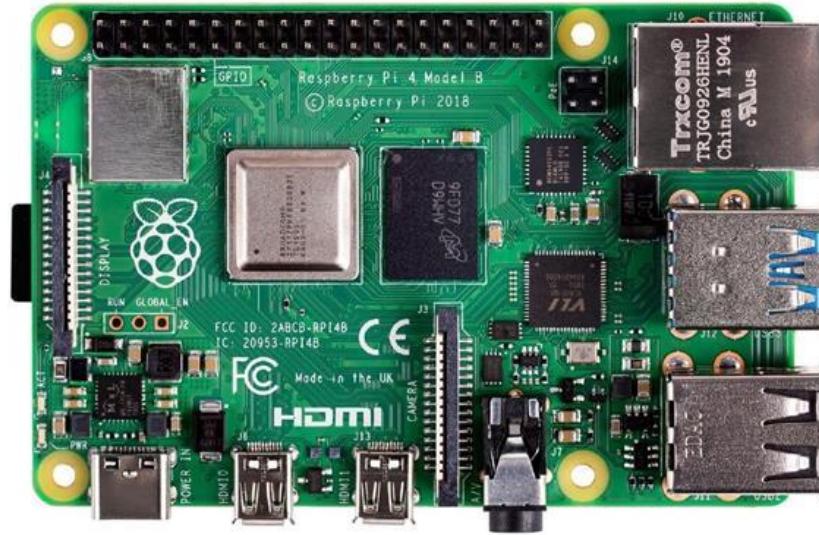
# Data Categories

# Structured versus unstructured data

There are three broad types of data and Microsoft Azure provides many data platform technologies to meet the needs of the wide varieties of data

Structured	Semi- Structured	Unstructured
<p>Structured data is data that adheres to a schema, so all of the data has the same fields or properties.</p> <p>Structured data can be stored in a database table with rows and columns</p>	<p>Semi-structured data doesn't fit neatly into tables, rows, and columns. Instead, semi-structured data uses _tags_ or _keys_ that organize and provide a hierarchy for the data</p>	<p>Unstructured data encompasses data that has no designated structure to it. Known as No-SQL., there are four types of No-SQL databases:</p> <ul style="list-style-type: none"><li>• Key Value Store</li><li>• Document Database</li><li>• Graph Databases</li><li>• Column Base</li></ul>

- IoT and telematics.
- Retail and marketing.
- Web and mobile applications.
- Gaming



## Non-Relational Database use case

```
{"latitude":37.8267,"longitude":-122.4233,"timezone":"America/Los_Angeles","currently":{"time":1598191217,"summary":"Partly Cloudy","icon":"partly-cloudy-day","nearestStormDistance":5,"nearestStormBearing":58,"precipIntensity":0,"precipProbability":0,"temperature":58.63,"apparentTemperature":58.63,"dewPoint":52.42,"humidity":0.8,"pressure":1011.8,"windSpeed":5.08,"windGust":7.73,"windBearing":210,"cloudCover":0.54,"uvIndex":0,"visibility":9.933,"ozone":291.2},"minutely":{"summary":"Partly cloudy for the hour. ","icon":"partly-cloudy-day","data":[{"time":1598191200,"precipIntensity":0,"precipProbability":0},{"time":1598191260,"precipIntensity":0,"precipProbability":0}, {"time":1598191320,"precipIntensity":0,"precipProbability":0}, {"time":1598191380,"precipIntensity":0,"precipProbability":0}, {"time":1598191440,"precipIntensity":0,"precipProbability":0}, {"time":1598191500,"precipIntensity":0,"precipProbability":0}, {"time":1598191560,"precipIntensity":0,"precipProbability":0}, {"time":1598191620,"precipIntensity":0,"precipProbability":0}, {"time":1598191680,"precipIntensity":0,"precipProbability":0}, {"time":1598191740,"precipIntensity":0,"precipProbability":0}, {"time":1598191800,"precipIntensity":0,"precipProbability":0}, {"time":1598191860,"precipIntensity":0,"precipProbability":0}, {"time":1598191920,"precipIntensity":0,"precipProbability":0}, {"time":1598191980,"precipIntensity":0,"precipProbability":0}, {"time":1598192040,"precipIntensity":0,"precipProbability":0}, {"time":1598192100,"precipIntensity":0,"precipProbability":0}, {"time":1598192160,"precipIntensity":0,"precipProbability":0}, {"time":1598192220,"precipIntensity":0,"precipProbability":0}, {"time":1598192280,"precipIntensity":0,"precipProbability":0}, {"time":1598192340,"precipIntensity":0,"precipProbability":0}, {"time":1598192400,"precipIntensity":0.0026,"precipIntensityError":0.0004,"precipProbability":0.01,"precipType":"rain"}, {"time":1598192460,"precipIntensity":0.0026,"precipIntensityError":0.0005,"precipProbability":0.01,"precipType":"rain"}, {"time":1598192520,"precipIntensity":0,"precipProbability":0}, {"time":1598192580,"precipIntensity":0.0026,"precipIntensityError":0.0005,"precipProbability":0.01,"precipType":"rain"}, {"time":1598192640,"precipIntensity":0.0026,"precipIntensityError":0.0005,"precipProbability":0.01,"precipType":"rain"}, {"time":1598192700,"precipIntensity":0.0027,"precipIntensityError":0.0005,"precipProbability":0.01,"precipType":"rain"}, {"time":1598192760,"precipIntensity":0.0027,"precipIntensityError":0.0005,"precipProbability":0.02,"precipType":"rain"}, {"time":1598192820,"precipIntensity":0.0026,"precipIntensityError":0.0005,"precipProbability":0.02,"precipType":"rain"}, {"time":1598192880,"precipIntensity":0,"precipProbability":0}], "hourly":{},"daily":{},"alerts":[]}
```

# Non-Relational Database use case

<https://api.darksky.net/>

## Open API :] สำหรับนักพัฒนา

แสดงค่าประจำวัน :

[//covid19.th-stat.com/api/open/today](https://covid19.th-stat.com/api/open/today)

ข้อมูลสรุปตามช่วงเวลา [เริ่มตั้งแต่วันที่ 01/01/20] :

[//covid19.th-stat.com/api/open/timeline](https://covid19.th-stat.com/api/open/timeline)

ข้อมูลแต่ละเคส :

[//covid19.th-stat.com/api/open/cases](https://covid19.th-stat.com/api/open/cases)

ข้อมูลสรุปจากเคส :

[//covid19.th-stat.com/api/open/cases/sum](https://covid19.th-stat.com/api/open/cases/sum)

แจ้งเตือนพื้นที่ตามคำประกาศ :

[//covid19.th-stat.com/api/open/area](https://covid19.th-stat.com/api/open/area)



กรมควบคุมโรค  
DEPARTMENT OF DISEASE CONTROL

# Non-Relational Database use case

- You might see the term *NoSQL* when reading about non-relational databases.
- NoSQL is a rather loose term that simply means non-relational.
- NoSQL (non-relational) databases generally fall into four categories:
  - key-value stores
  - document databases
  - column family databases
  - graph databases.

# What is NoSQL?

Key	Value
AAAAAA	1101001111010100110101111...
AABAB	1001100001011001101011110...
DFA766	000000000101010110101010...
FABCC4	1110110110101010100101101...

Opaque to  
data store

A key-value store is the simplest (and often quickest) type of NoSQL database for inserting and querying data.

# Key-Value Stores

Key	Document
1001	<pre>{   "CustomerID": 99,   "OrderItems": [     { "ProductID": 2010,       "Quantity": 2,       "Cost": 520     },     { "ProductID": 4365,       "Quantity": 1,       "Cost": 18     }   ],   "OrderDate": "04/01/2017" }</pre>
1002	<pre>{   "CustomerID": 220,   "OrderItems": [     { "ProductID": 1285,       "Quantity": 1,       "Cost": 120     }   ],   "OrderDate": "05/08/2017" }</pre>

A document database represents the opposite end of the NoSQL spectrum from a key-value store. In a document database, each document has a unique ID, but the fields in the documents are transparent to the database management system. Document databases typically store data in JSON format,

## Document Databases

RDBMS	MongoDB
Database	Database
Table	Collection
Tuple/Row	Document
column	Field
Table Join	Embedded Documents
Primary Key	Primary Key (Default key <code>_id</code> provided by mongoDB itself)

# Document Databases



Customer	
PK	CustomerID
FK1	Title FirstName LastName AddressID

Address	
PK	AddressID
	StreetAddress City State ZipCode



Customer Table

CustomerID	Title	FirstName	LastName	AddressID
1	Mr	Mark	Hanson	500
2	Ms	Lisa	Andrews	501
3	Mr	Walter	Harp	500

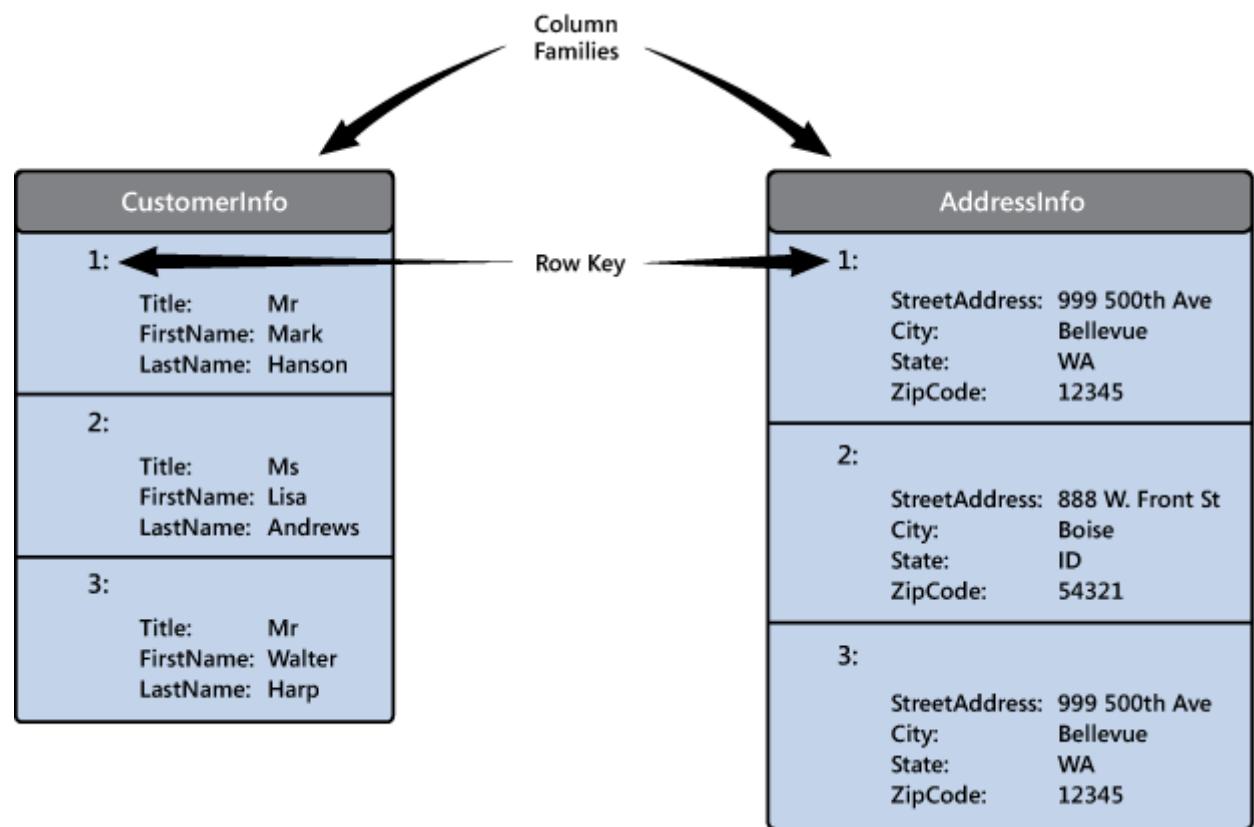
Address Table

AddressID	StreetAddress	City	State	ZipCode
500	999 500th Ave	Bellevue	WA	12345
501	888 W. Front St	Boise	ID	54321

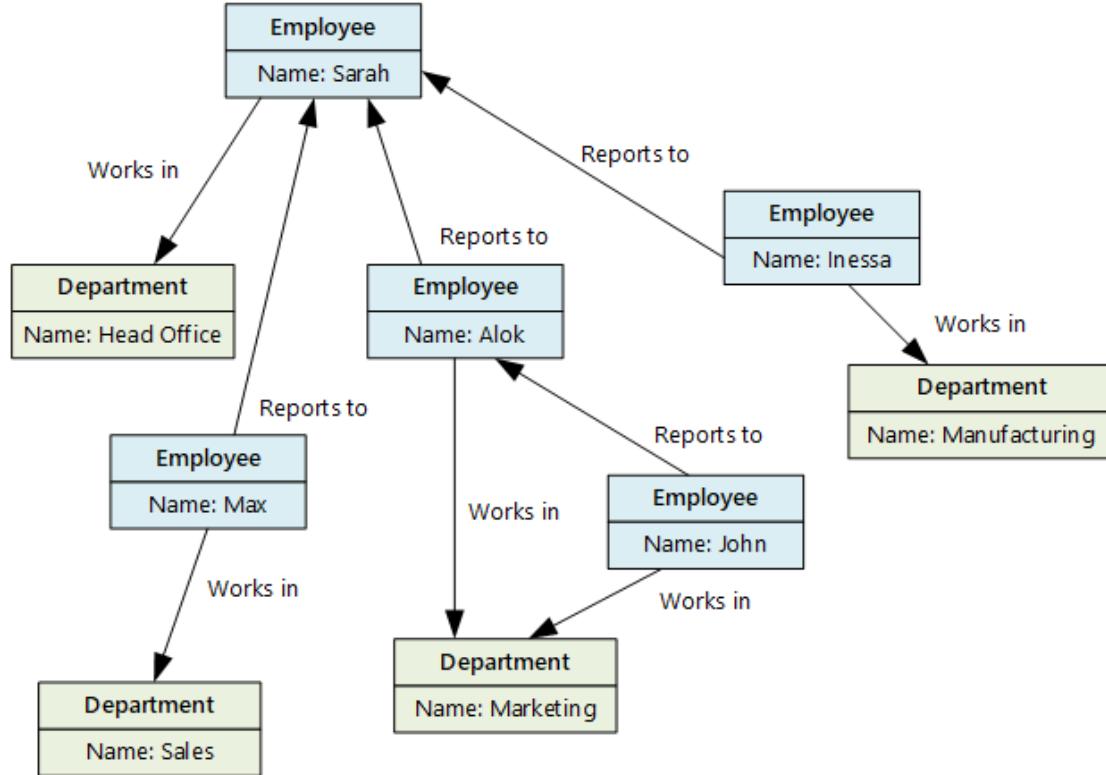
RDBMS is Row-based oriented

# Column Family Databases

Row Key	Column Families		
	CustomerInfo		AddressInfo
CustomerID	CustomerInfo:Title	CustomerInfo:FirstName	CustomerInfo:LastName
1	CustomerInfo:Title Mr	CustomerInfo:FirstName Mark	CustomerInfo:LastName Hanson
	AddressInfo:StreetAddress 999 500th Ave	AddressInfo:City Bellevue	AddressInfo:State WA
	AddressInfo:ZipCode 12345		
2	CustomerInfo:Title Ms	CustomerInfo:FirstName Lisa	CustomerInfo:LastName Andrews
	AddressInfo:StreetAddress 888 W. Front St	AddressInfo:City Boise	AddressInfo:State ID
	AddressInfo:ZipCode 54321		
3	CustomerInfo:Title Mr	CustomerInfo:FirstName Walter	CustomerInfo:LastName Harp
	AddressInfo:StreetAddress 999 500th Ave	AddressInfo:City Bellevue	AddressInfo:State WA
	AddressInfo:ZipCode 12345		



# Column Family Databases



Graph databases enable you to store entities, but the main focus is on the relationships that these entities have with each other.

A graph database stores two types of information: nodes that you can think of as instances of entities, and edges, which specify the relationships between nodes.

# Graph Databases

# What to use for Data



Storage Account



- When you need a **low cost, high throughput** data store.
- When you need to store **No-SQL** data.
- When you **do not need to query** the data directly. **No ad hoc query** support.
- Suits the storage of archive or **relatively static data**.
- Suits acting as a **HDInsight Hadoop** data store.

# What to use for Data



Data Lake Store



Module 02

- When you need a **low cost, high throughput** data store.
- **Unlimited storage** for **No-SQL** data
- When you **do not need to query** the data directly. **No ad hoc query** support.
- Suits the storage of archive or **relatively static data**.
- Suits acting as a **Databricks**, **HDInsight** and **IoT** data store.

# What to use for Data



Azure Databricks



- **Eases the deployment** of a Spark based cluster.
- Enables the **fastest processing** of Machine Learning solutions.
- **Enables collaboration** between data engineers and data scientists.
- Provides **tight enterprise security integration** with Azure Active Directory
- **Integration with other Azure Services** and **Power BI**.

# What to use for Data



Azure CosmosDB



- Provides **global distribution** for both structured and unstructured data stores.
- **Millisecond query response** time.
- **99.999% availability** of data.
- **Worldwide elastic scale** of both the storage and throughput
- **Multiple consistency levels** to control data integrity with concurrency

# What to use for Data



Azure SQL Database

- When you require a **relational** data store.
- When you need to manage **transactional workloads**
- When you need to manage a **high volume on inserts and reads**
- When you need a service that **requires high concurrency**
- When you require a solution that can scale **elastically**

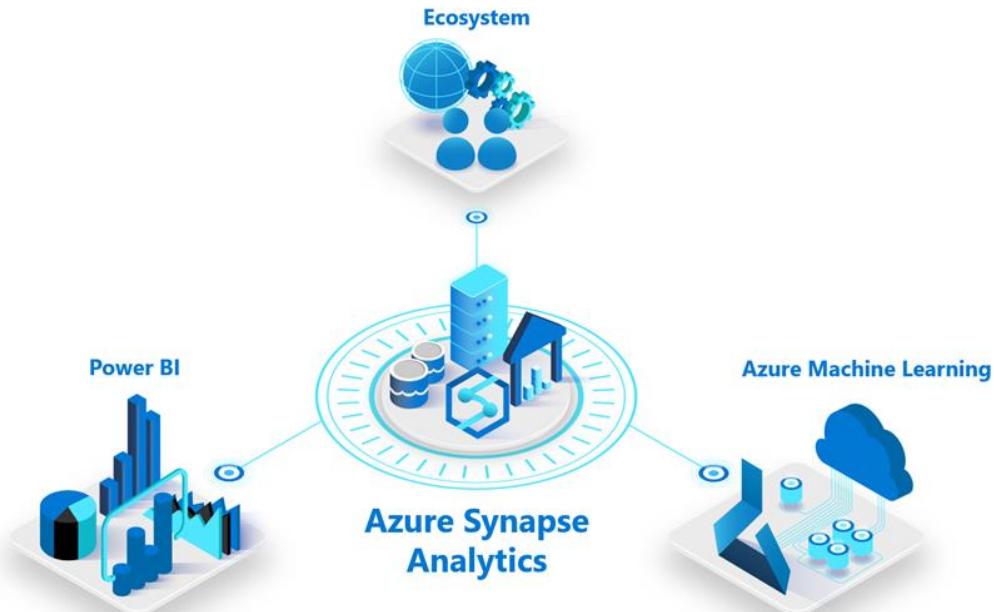
# What to use for Data



Azure Synapse  
Analytics

Module 05

- When you require an integrated **relational** and **big data** store.
- When you need to manage **data warehouse** and **analytical workloads**
- When you need **low cost storage**.
- When you require the ability to **pause and restart the compute**.
- When you require a solution that can scale **elastically**



# What to use for Data



Azure Stream Analytics



- When you require a **fully managed event processing engine**.
- When you require **temporal analysis of streaming data**.
- Support for analyzing **IoT streaming data**.
- Support for analyzing application data through **Event Hubs**.
- Ease of use with a **Stream Analytics Query Language**.

# What to use for Data



- When you want to **orchestrate the batch movement** of data.
- When you want to connect to **wide range of data platforms**.
- When you want to **transform or enrich** the data in movement.
- When you want to **integrate with SSIS packages**.
- Enables **verbose logging** of data processing activities.

# What to use for Data



Azure HDInsight



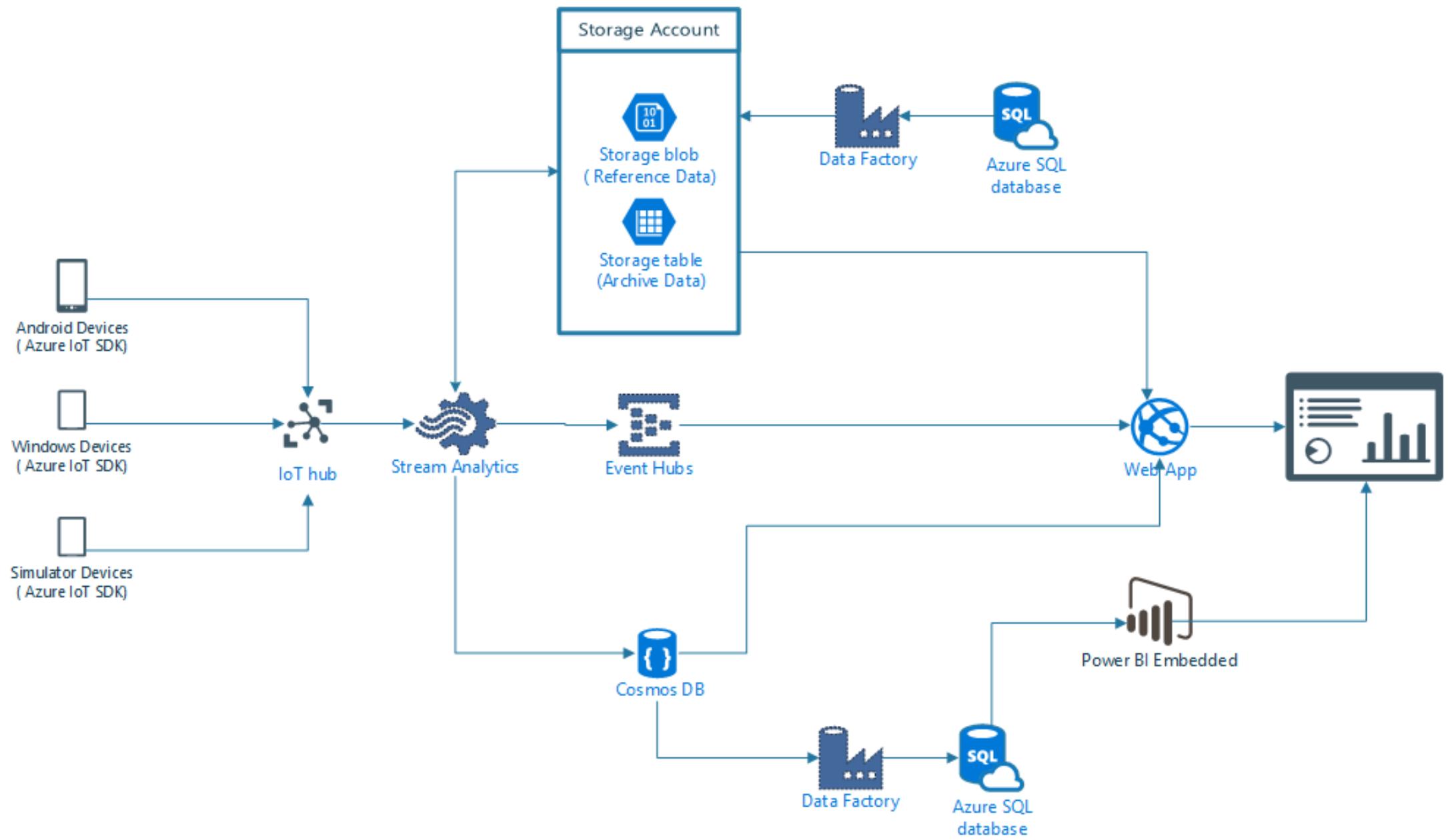
- When you need a **low cost, high throughput** data store.
- When you need to store **No-SQL** data.
- Provides a Hadoop **Platform as a Service** approach
- Suits acting as a **Hadoop, Hbase, Storm or Kafka** data store.
- **Eases the deployment and management** of clusters.



Azure Data Catalog



- When you require **documentation** of your data stores.
- When you require a **multi user** approach to documentation.
- When you need to **annotate data sources** with descriptive metadata.
- A **fully managed cloud service** whose users can discover the data sources.
- When you require **a solution that can help business users** understand their data.





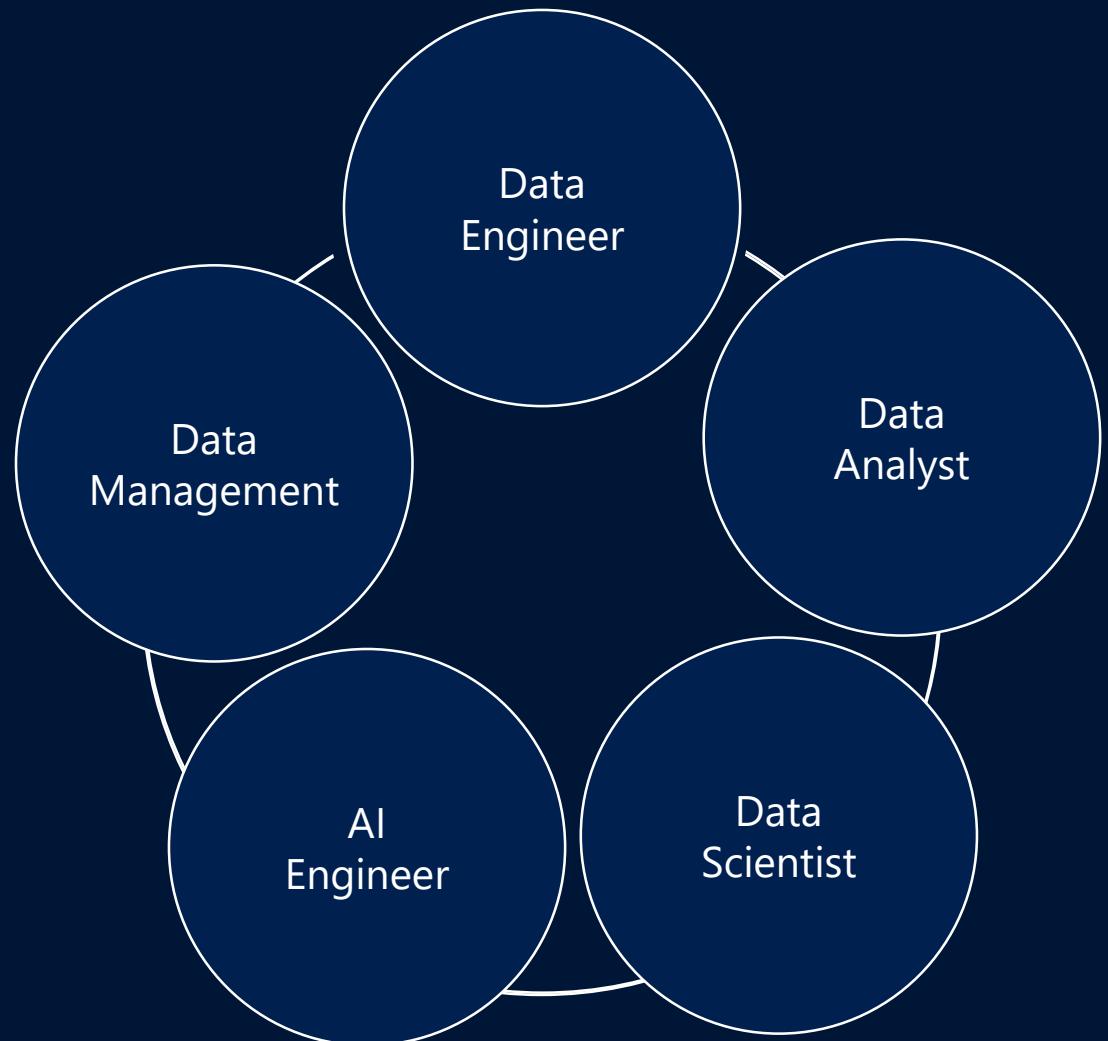
# Lesson 03

## Identify the Tasks Performed by a Data Engineer

# Lesson Objectives

- List the new roles of modern data projects
- Outline data engineering practices
- Explore the high-level process for architecting a data engineering project

# Roles in Data Projects



# Data Engineering Practices

The background of the slide is a blurred photograph of a professional workspace, likely a data center or control room. Several people are visible in the background, sitting at desks and working on computers. The desks are equipped with multiple monitors displaying various data and interface screens. The overall atmosphere is one of a busy, technical environment.

Provision

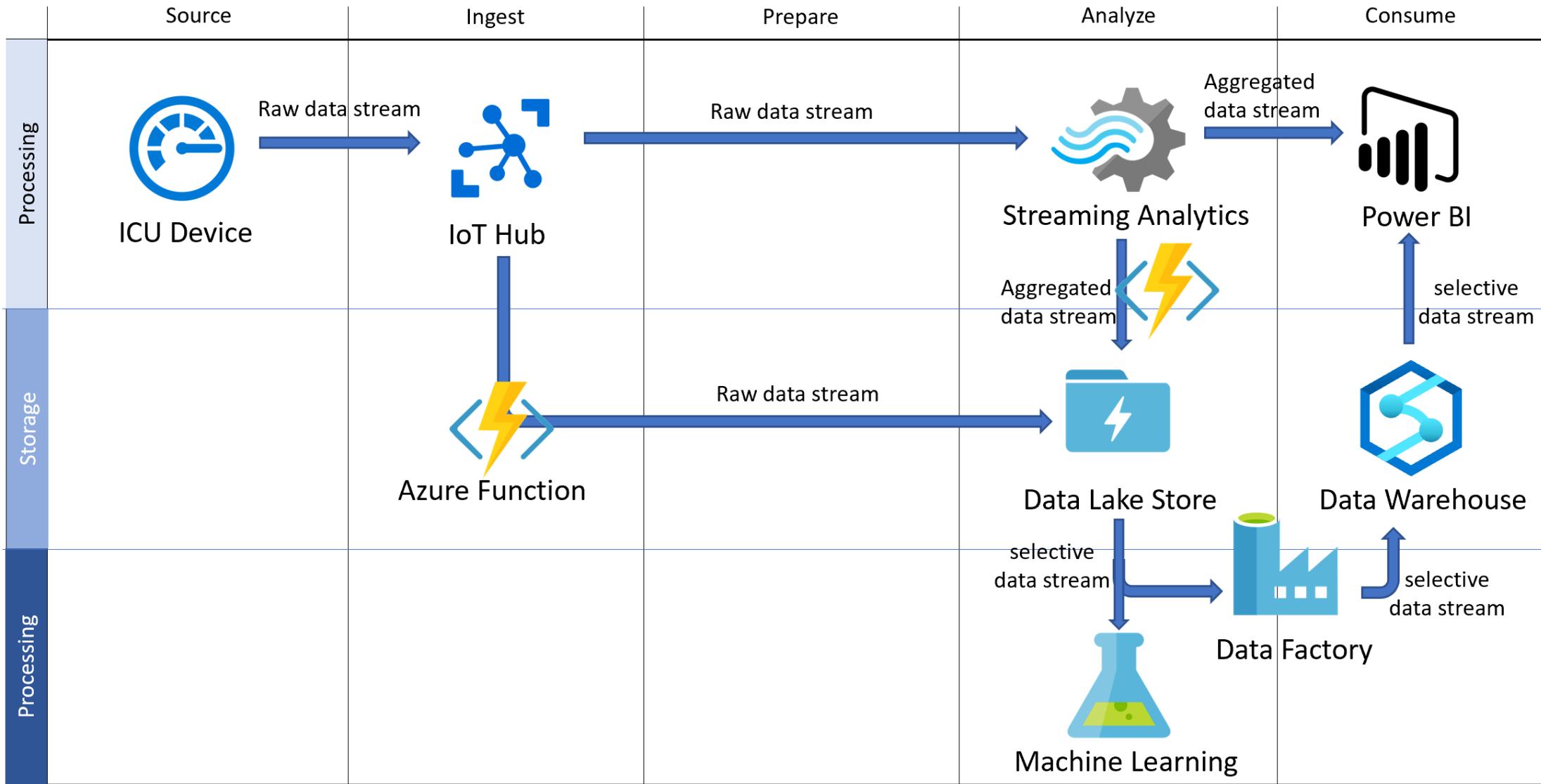
Process

Secure

Monitor

Disaster  
Recovery

# Architecting Projects – an example





# Lesson 04

## Course Case Study



# Lesson Objectives

- Read the course case study

# Course case study: AdventureWorks Cycles

## Read the case study

In this section of the course, the instructor will either:

- Allocate you 10 minutes to read through the case study.
- Or
- Spend 10 minutes walking through the case study with you as a group

**Note:**

This case study will be used in labs across the entire course. Each lab will drill down more into the detail of what is required as you perform each lab.



# Lab: Azure for the Data Engineer



# Module Summary ›

## In this module, you have learned about:

- The evolving world of data.
- The services in the Azure Data Platform.
- The tasks that are performed by a Data Engineer.
- A fictitious Case Study for use in labs.

## Next steps ›

After the course, consider visiting [[the Microsoft Customer Case Study site](#)]. Use the search bar to search by an industry such as healthcare or retail, or by a technology such as Azure Cosmos DB or Stream Analytics. Read through some of the customers stories.





# Module 02: Working with Data Storage



# Agenda

- L01 – Choose a data storage approach in Azure
- L02 - Create an Azure Storage Account
- L03 - Explain Azure Data Lake Storage
- L04 - Upload data into Azure Data Lake Store



The background of the slide features a collage of nine images. Top row: 1. Students in a classroom setting. 2. A man in a plaid shirt pointing at a whiteboard. 3. Two men working at a desk with multiple monitors. 4. Two men in a factory or industrial setting looking at a large screen. Middle row: 5. A woman looking down at her phone. 6. A woman writing on a clipboard. Bottom row: 7. A person wearing headphones playing a video game. 8. A person using a tablet to draw over architectural blueprints. 9. A woman working on a computer with a brain icon overlay.

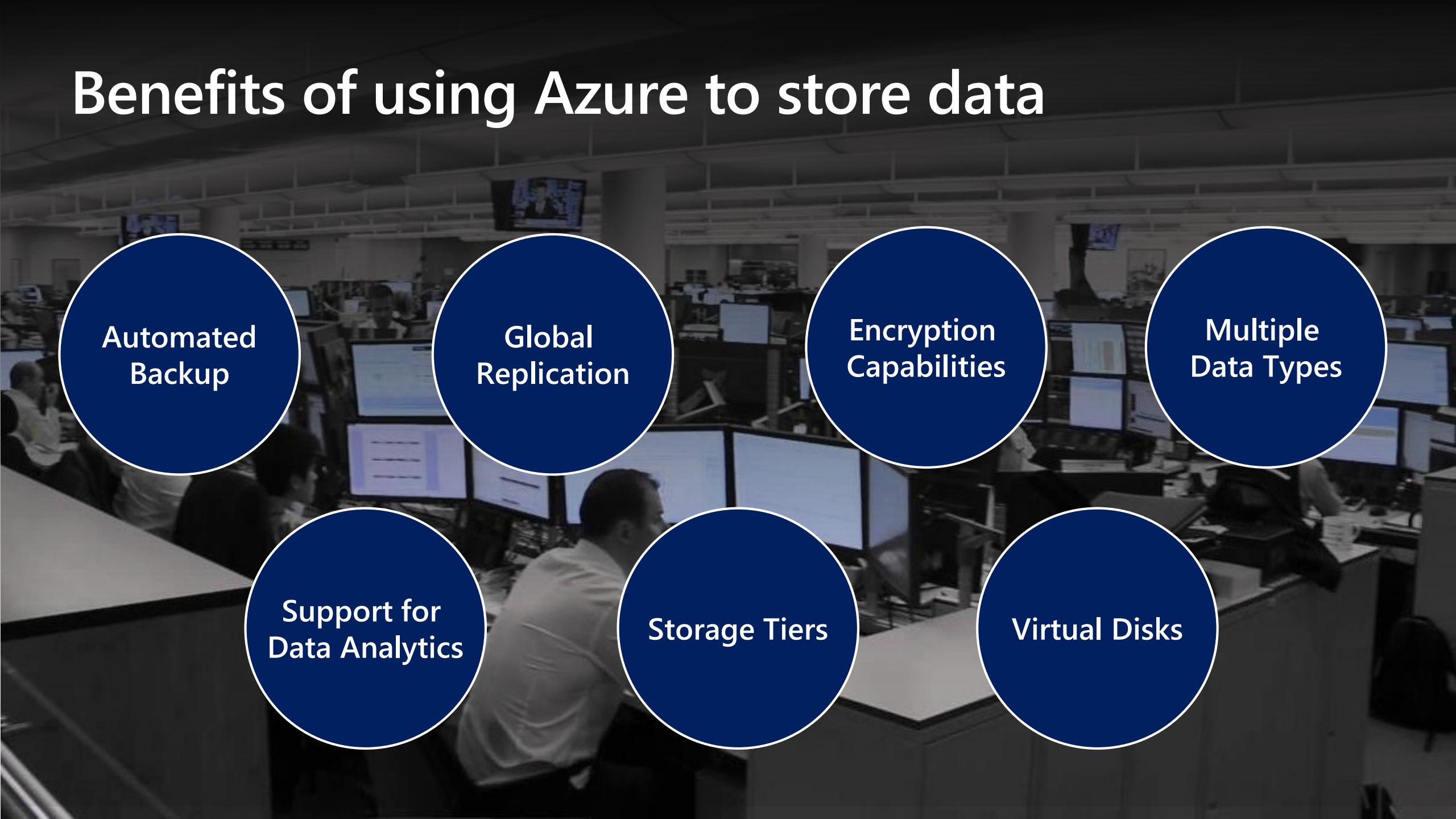
# Lesson 01

## Choose a Data Storage Approach in Azure

# Lesson Objectives

- The Benefits of using Azure to store data
- Compare Azure data storage with on-premises storage

# Benefits of using Azure to store data

The background of the slide is a grayscale photograph of a modern control room or data center. Several people are visible in the background, working at their respective stations. Each station is equipped with multiple computer monitors displaying various data visualizations and system logs. The room has a high ceiling with exposed structural elements and lighting fixtures.

Automated  
Backup

Global  
Replication

Encryption  
Capabilities

Multiple  
Data Types

Support for  
Data Analytics

Storage Tiers

Virtual Disks

# Comparing Azure to on-premises storage

The term "on-premises" refers to the storage and maintenance of data on local hardware and servers.

Cost effectiveness	Reliability	Storage types	Agility
On-premises storage requires up-front expenses. Azure data storage provides a pay-as-you-go pricing model	Azure data storage provides backup, load balancing, disaster recovery, and data replication to ensure safety and high availability. This capability requires significant investment with on-premises solutions	Azure data storage provides a variety of different storage options including distributed access and tiered storage	Azure data storage gives you the flexibility to create new services in minutes and allows you to change storage backends quickly



# Lesson 02

## Create an Azure Storage Account

# Lesson Objectives

- Describe storage accounts
- Determine the appropriate settings for each storage account
- Choose an account creation tool
- Create a storage account using the Azure portal

# Storage accounts

## What is a Storage Account

It is a container that groups a set of Azure Storage services. Only data services can be included in a storage account such as *Azure Blobs*, *Azure Files*, *Azure Queues*, and *Azure Tables*.

## How many do you need?

The number of storage accounts you need is typically determined by your data diversity, cost sensitivity, and tolerance for management overhead.

**The number of storage accounts you need is based on:**

### Data Diversity

Organizations often generate data that differs in where it is consumed and how sensitive it is.



### Cost Sensitivity

The settings you choose for the account do influence the cost of services, and the number of accounts you create



### Management Overhead

Each storage account requires some time and attention from an administrator to create and maintain.



# Storage Account settings

## Create storage account

Basics Networking Advanced Tags Review + create

Azure Storage is a Microsoft-managed service providing cloud storage that is highly available, secure, durable, scalable, and redundant. Azure Storage includes Azure Blobs (objects), Azure Data Lake Storage Gen2, Azure Files, Azure Queues, and Azure Tables. The cost of your storage account depends on the usage and the options you choose below.

[Learn more about Azure storage accounts](#)

### Project details

Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources.

Subscription \*

chtestao

Resource group \*

Select existing...

[Create new](#)

### Instance details

The default deployment model is Resource Manager, which supports the latest Azure features. You may choose to deploy using the classic deployment model instead. [Choose classic deployment model](#)

Storage account name \* ⓘ

Location \*

(US) South Central US

Performance ⓘ

Standard  Premium

Account kind ⓘ

StorageV2 (general purpose v2)

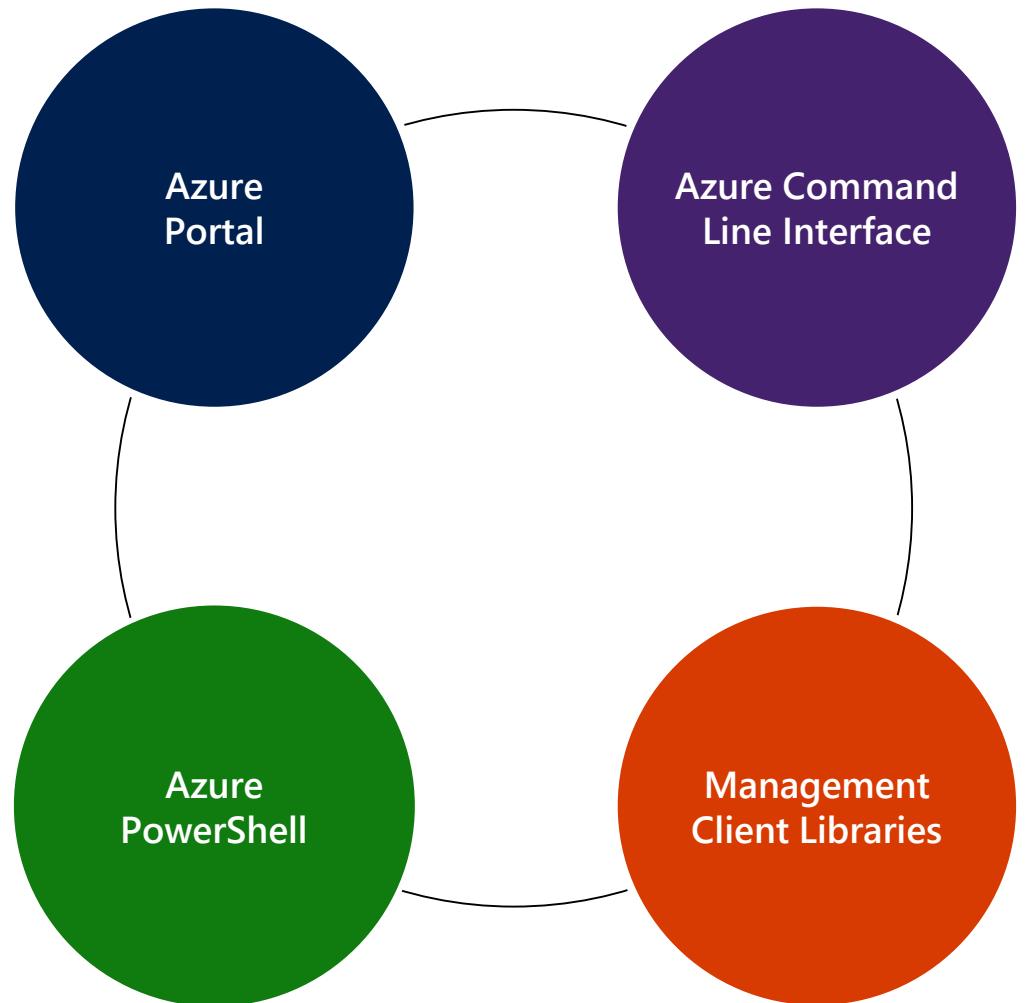
Replication ⓘ

Read-access geo-redundant storage (RA-GRS)

Access tier (default) ⓘ

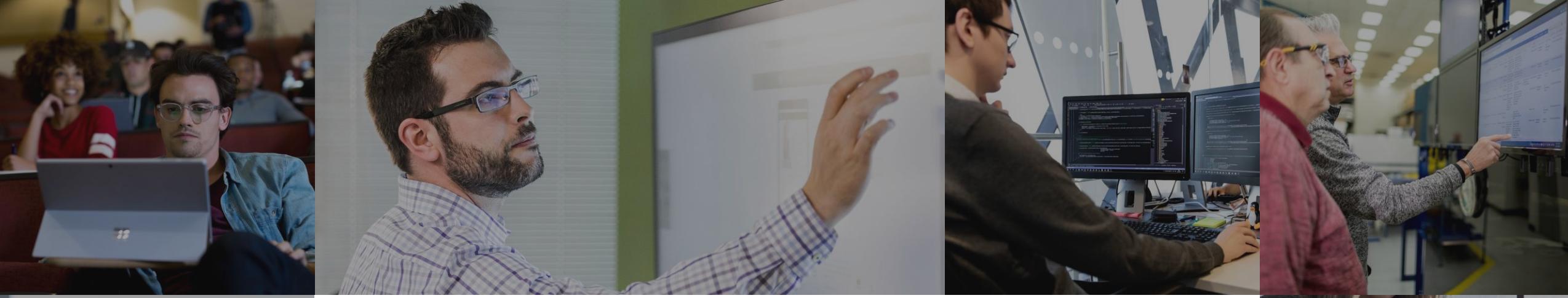
Cool  Hot

# Storage Account creation tool



# Create a Storage Account





# Lesson 03

## Azure Data Lake Storage

Start = 13.00



# Lesson Objectives

- Explain Azure Data Lake Storage
- Create an Azure Data Lake Store Gen 2 using the portal
- Compare Azure Blob Storage and Data Lake Store Gen 2
- Explore the stages for processing Big Data Using Azure Data Lake Store
- Describe the use cases for Data lake Storage

# Azure Data Lake Storage – Generation II



Hadoop  
Access



Security



Performance



Redundancy

# Create a Azure Data Lake Store (Gen II) using the Portal.

Home > New > Storage account > Create storage account

## Create storage account

Basics Networking Advanced Tags Review + create

### Security

Secure transfer required ⓘ  Disabled  Enabled

### Azure Files

Large file shares ⓘ  Disabled  Enabled

i The current combination of storage account kind, performance, replication and location does not support large file shares.

### Data protection

Blob soft delete ⓘ  Disabled  Enabled

i Data protection and hierarchical namespace cannot be enabled simultaneously.

#### Data Lake Storage Gen2

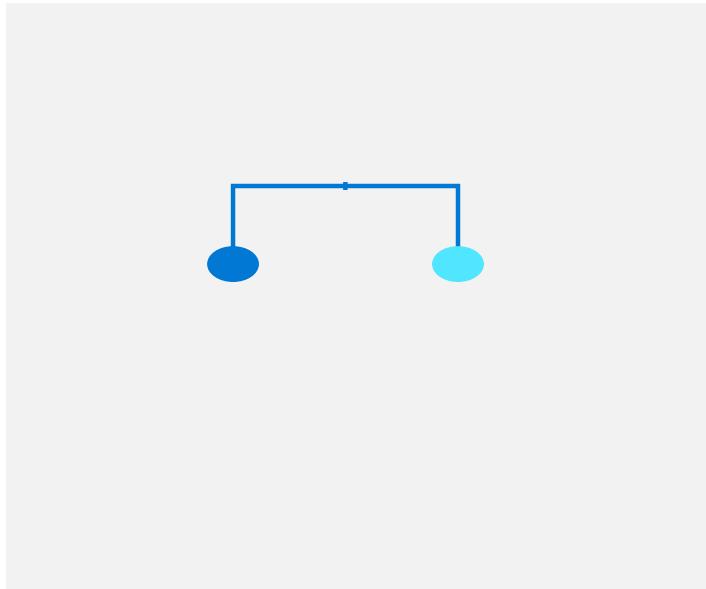
Hierarchical namespace ⓘ  Disabled  Enabled

### NFS v3 ⓘ

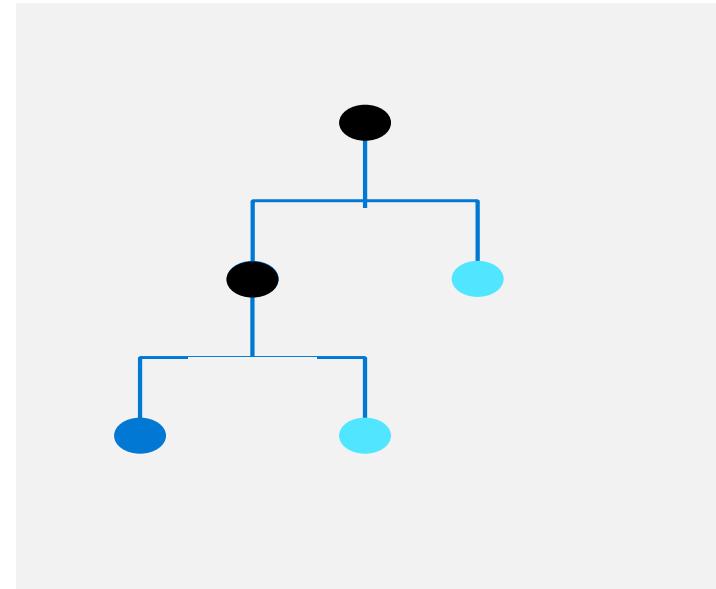
Signup is currently required to utilize the the NFS v3 feature on a per-subscription basis. [Signup for NFS v3](#)

# Compare Azure Blob Storage and Data Lake Store Gen 2

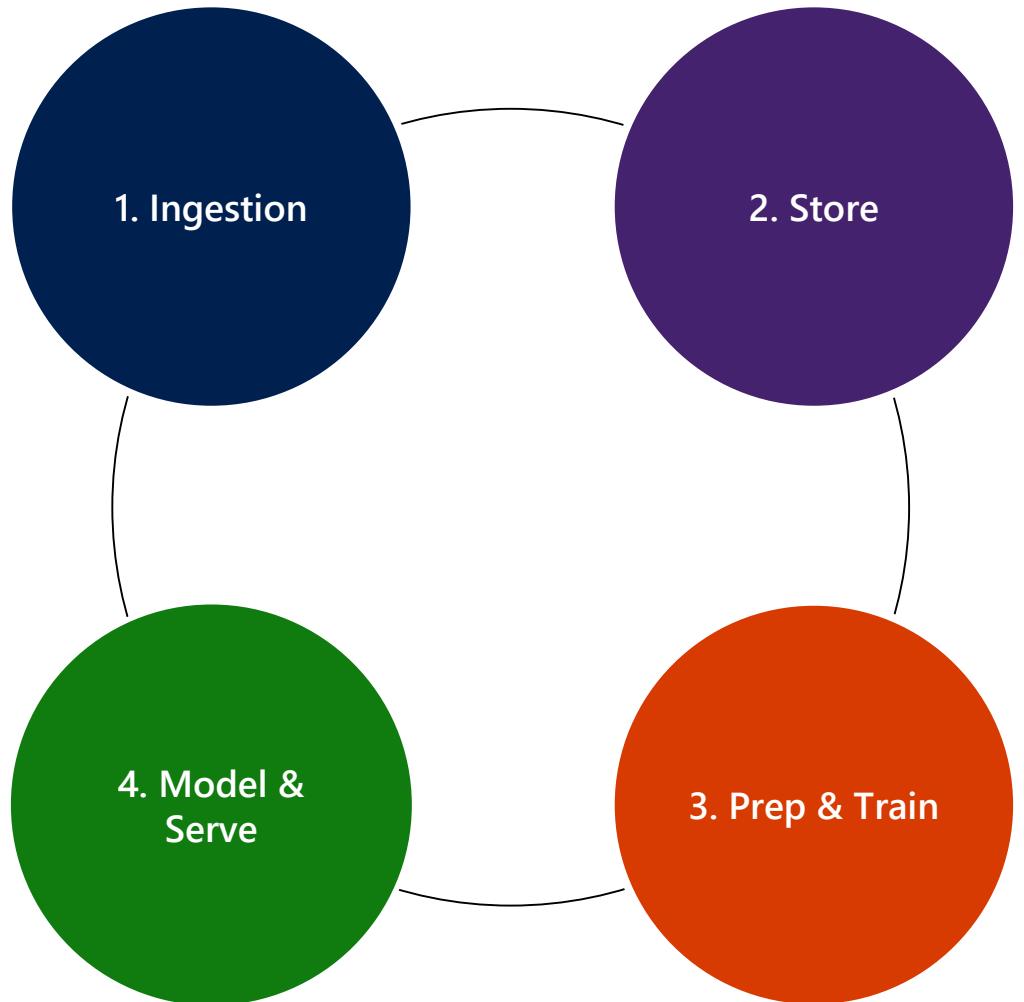
Azure Blob  
Flat Namespace



Data Lake (Gen II)  
Hierarchical Namespace



# Processing Big Data with Azure Data Lake Store



# 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

## Upload Data into Azure Data Lake Store

# Lesson Objectives

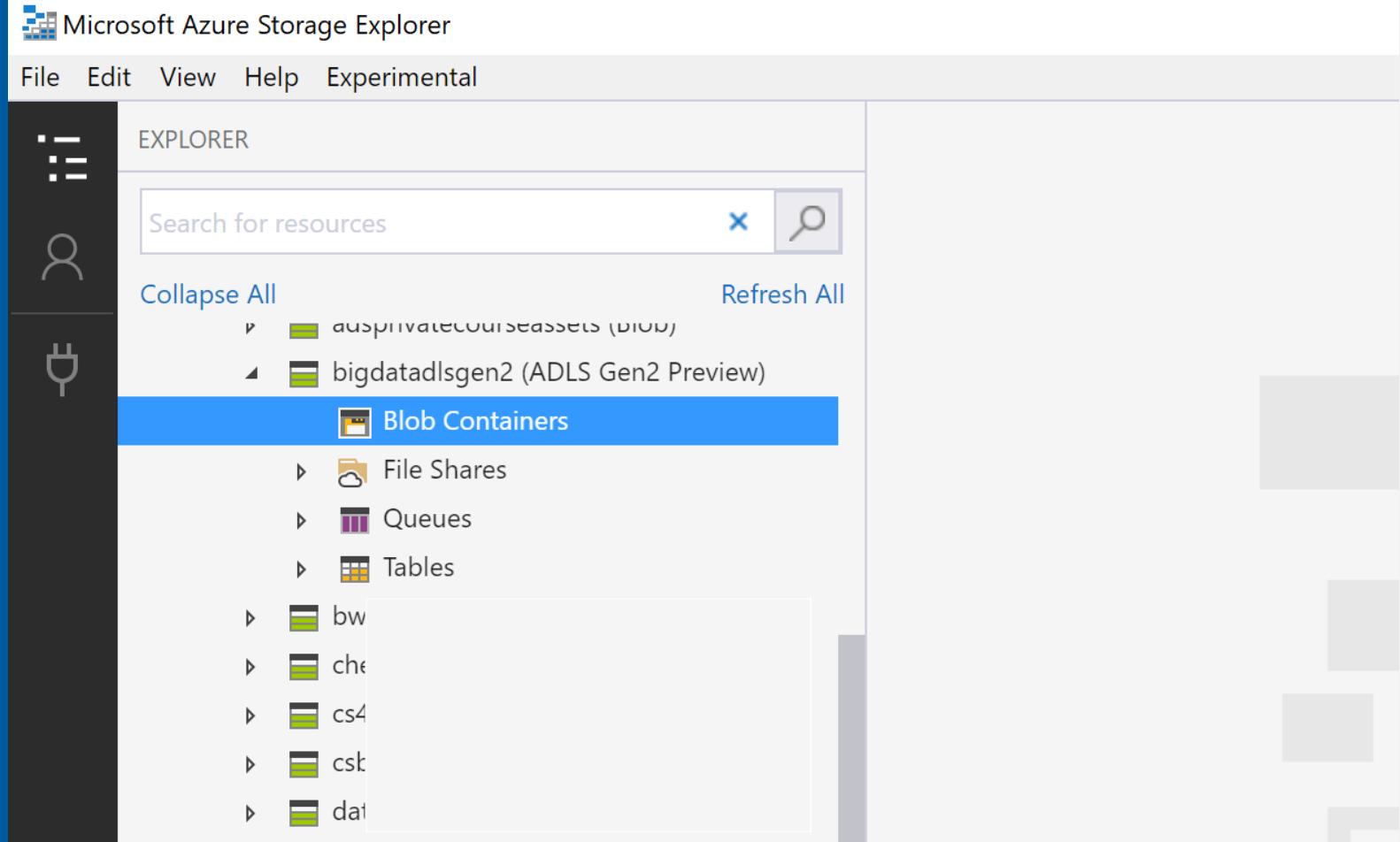
- Create an Azure Data Lake Gen2 Store using PowerShell
- Upload data into the Data Lake Storage Gen2 using Azure Storage Explorer
- Copy data from an Azure Data Lake Store Gen1 to an Azure Data Lake Store Gen2

Create a Azure  
Data Lake Store (Gen II)  
using PowerShell.

```
Windows PowerShell
Copyright (C) Microsoft Corporation. All rights reserved.

PS C:\Users> $location = "westus2"
>>
>> New-AzStorageAccount -ResourceGroupName $resourceGroup
>> -Name "storagequickstart"
>> -Location $location
>> -SkuName Standard_LRS
>> -Kind StorageV2
>> -EnableHierarchicalNamespace $True
```

# Uploading data with Azure Storage Explorer

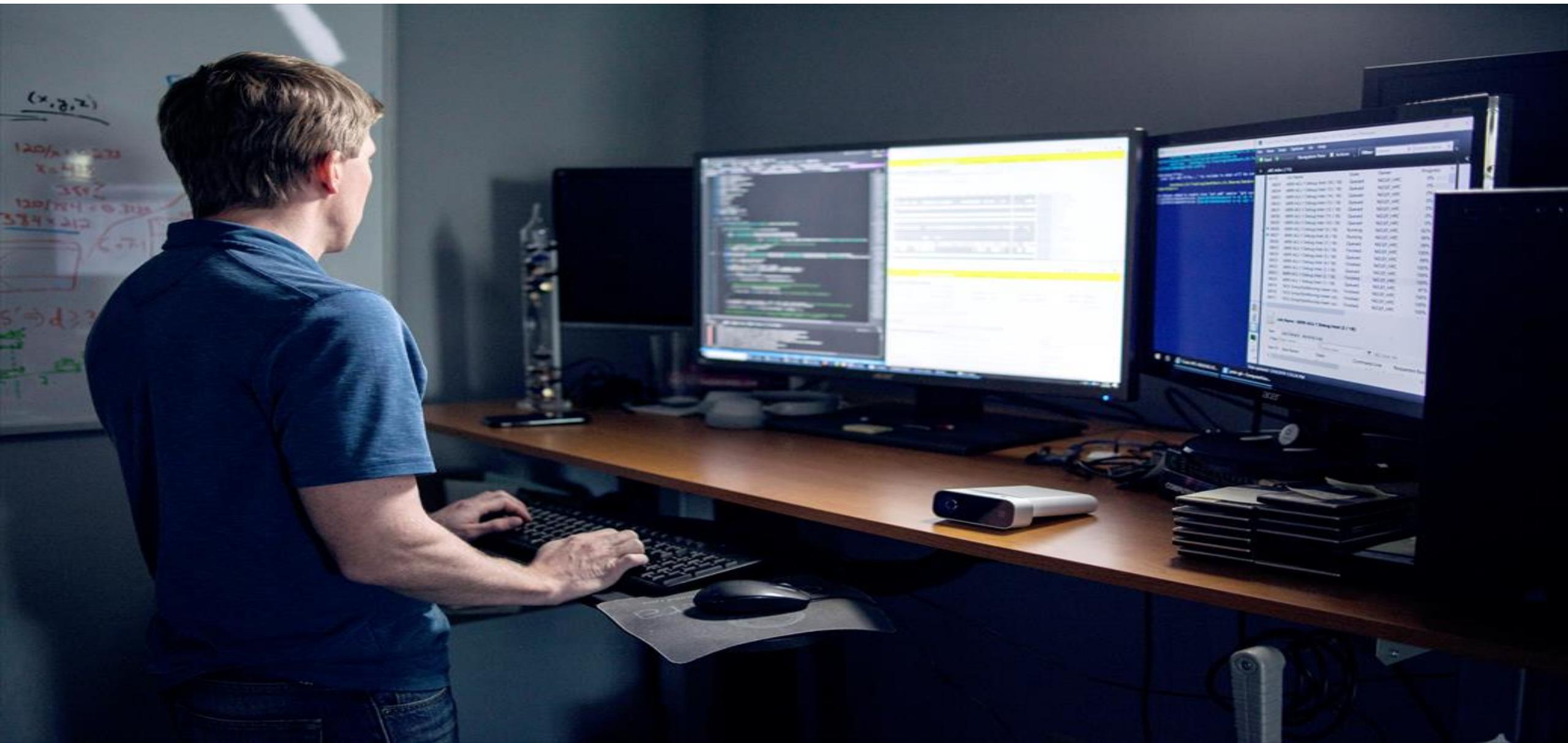


Copy data from an Azure Data Lake Store Gen1 to an Azure Data Lake Store Gen2

The screenshot shows the Microsoft Azure Data Factory interface. At the top, the navigation bar includes 'Microsoft Azure' and 'Data Factory' with a sub-path 'awadf'. A purple banner at the top right says 'Help us improve. [Click here](#) to tell us how we are doing.' Below the banner, the title 'Azure Data Factory' is displayed, followed by the large text 'Let's get started'. There are four main buttons arranged horizontally: 'Create pipeline' (represented by a blue pipe icon), 'Create pipeline from template' (represented by a flowchart icon), 'Copy Data' (represented by two blue cylinders with yellow stars), and 'Configure Integration' (represented by a blue cylinder with a green plus sign). On the far left, there is a vertical sidebar with three icons: a blue folder, a blue pencil, and a red circular icon.

- Create pipeline**
- Create pipeline from template**
- Copy Data**
- Configure Integration**

# Lab: Working with Data Storage





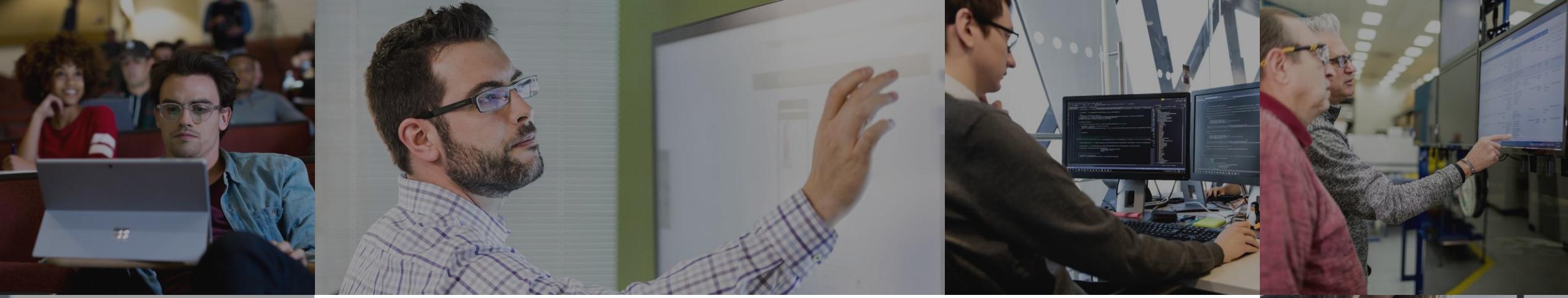
# Module 03:

## Enabling Team Based Data Science with Azure Databricks



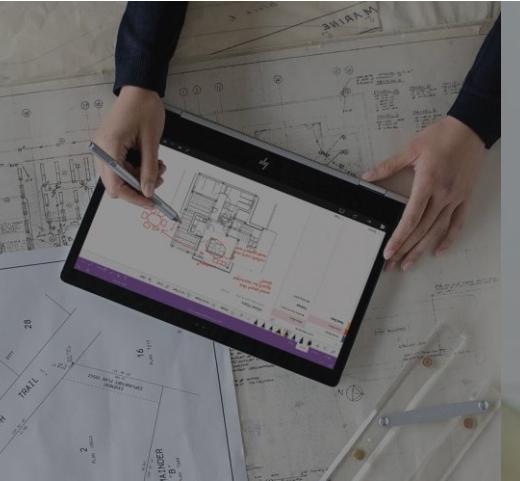
# Agenda

- L01 – Describe Azure Databricks
- L02 - Provision Azure Databricks and Workspaces
- L03 - Read data using Azure Databricks
- L04 - Perform transformations with Azure Databricks



# Lesson 01

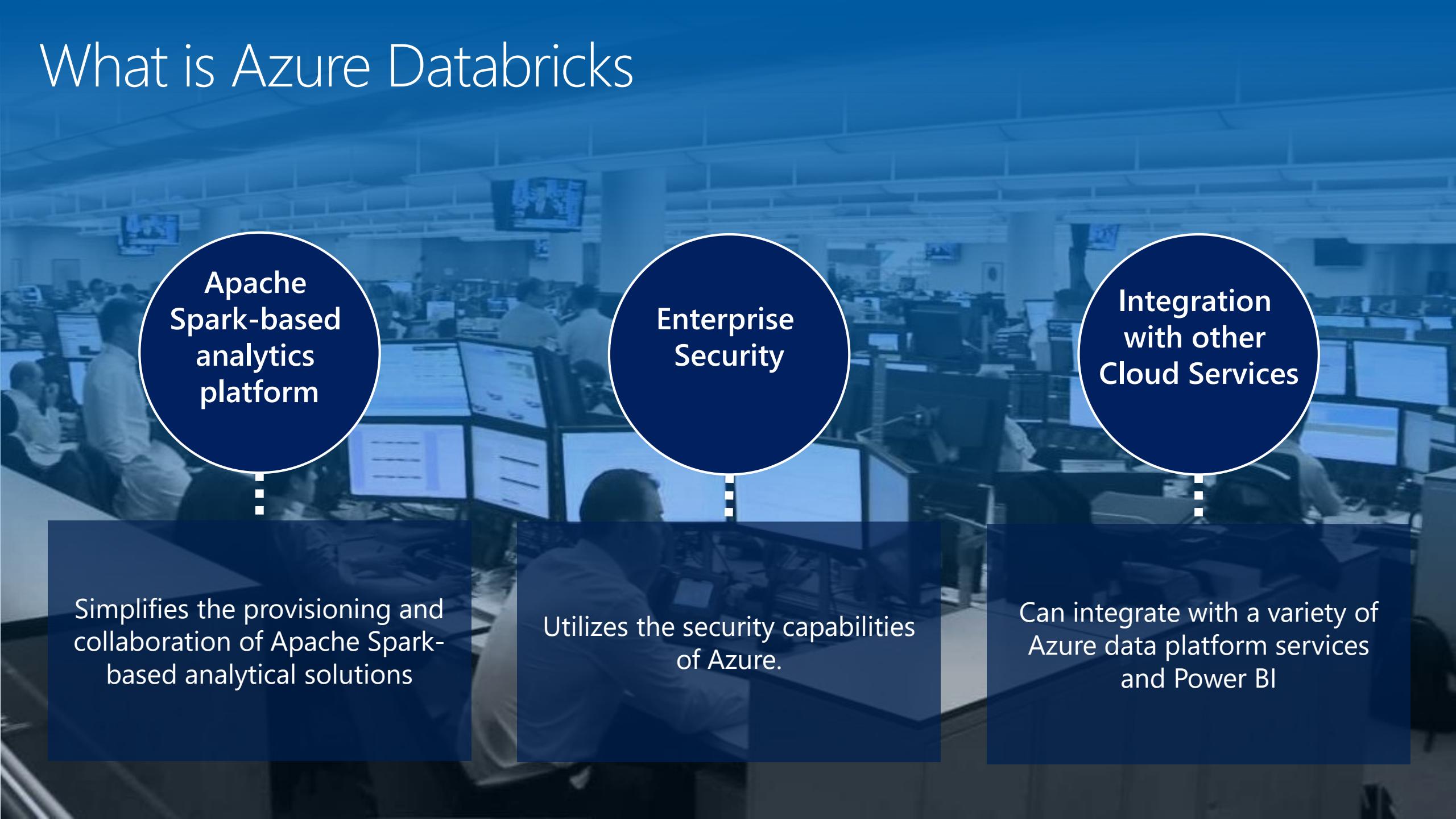
## Describe Azure Databricks



# Lesson Objectives

- What is Azure Databricks
- What are Spark based analytics platform
- How Azure Databricks integrates with enterprise security
- How Azure Databricks integrates with other cloud services

# What is Azure Databricks



A circular callout in the upper left corner contains the text "Apache Spark-based analytics platform".

Apache  
Spark-based  
analytics  
platform

A circular callout in the upper middle contains the text "Enterprise Security".

Enterprise  
Security

A circular callout in the upper right contains the text "Integration with other Cloud Services".

Integration  
with other  
Cloud Services

Simplifies the provisioning and collaboration of Apache Spark-based analytical solutions

Utilizes the security capabilities of Azure.

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

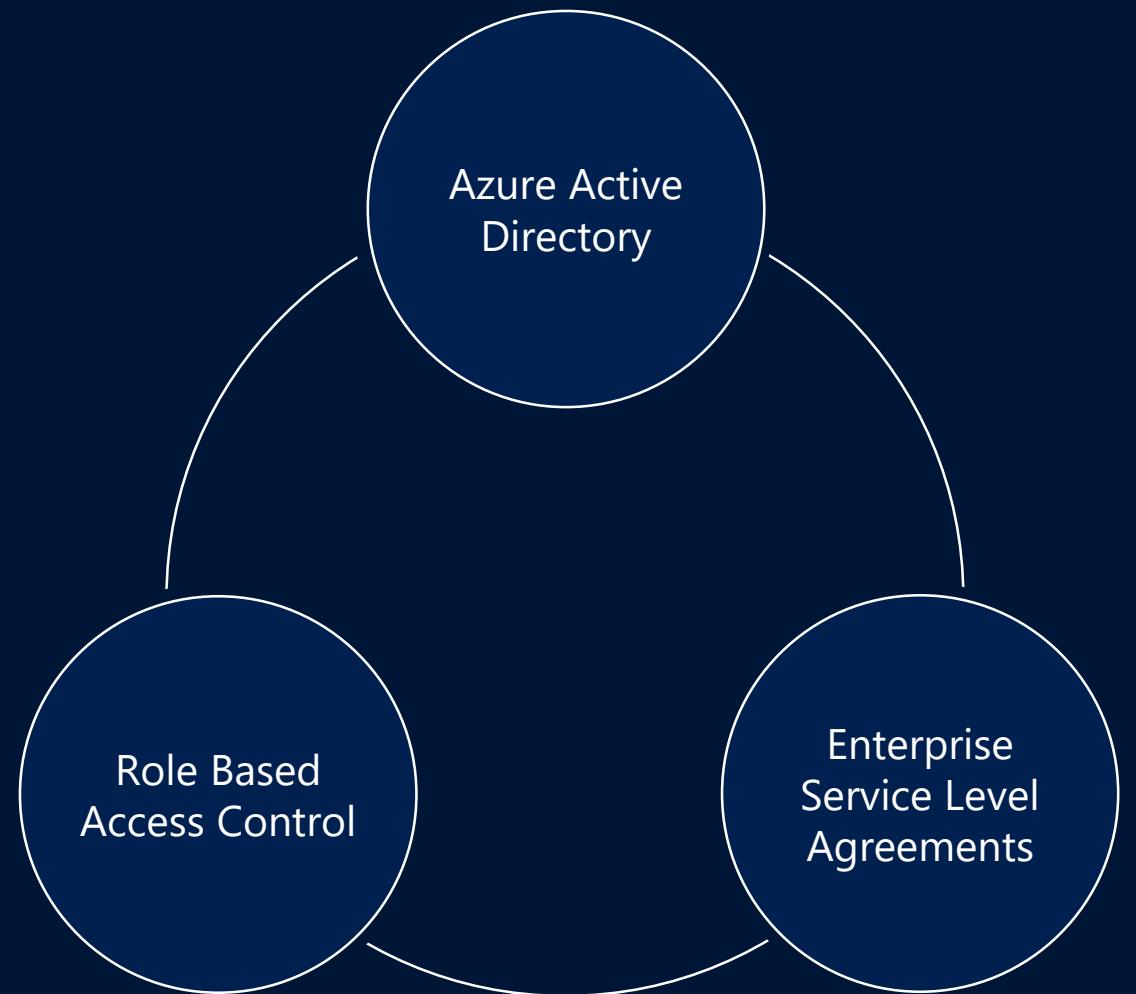
## Why use Azure Databricks

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

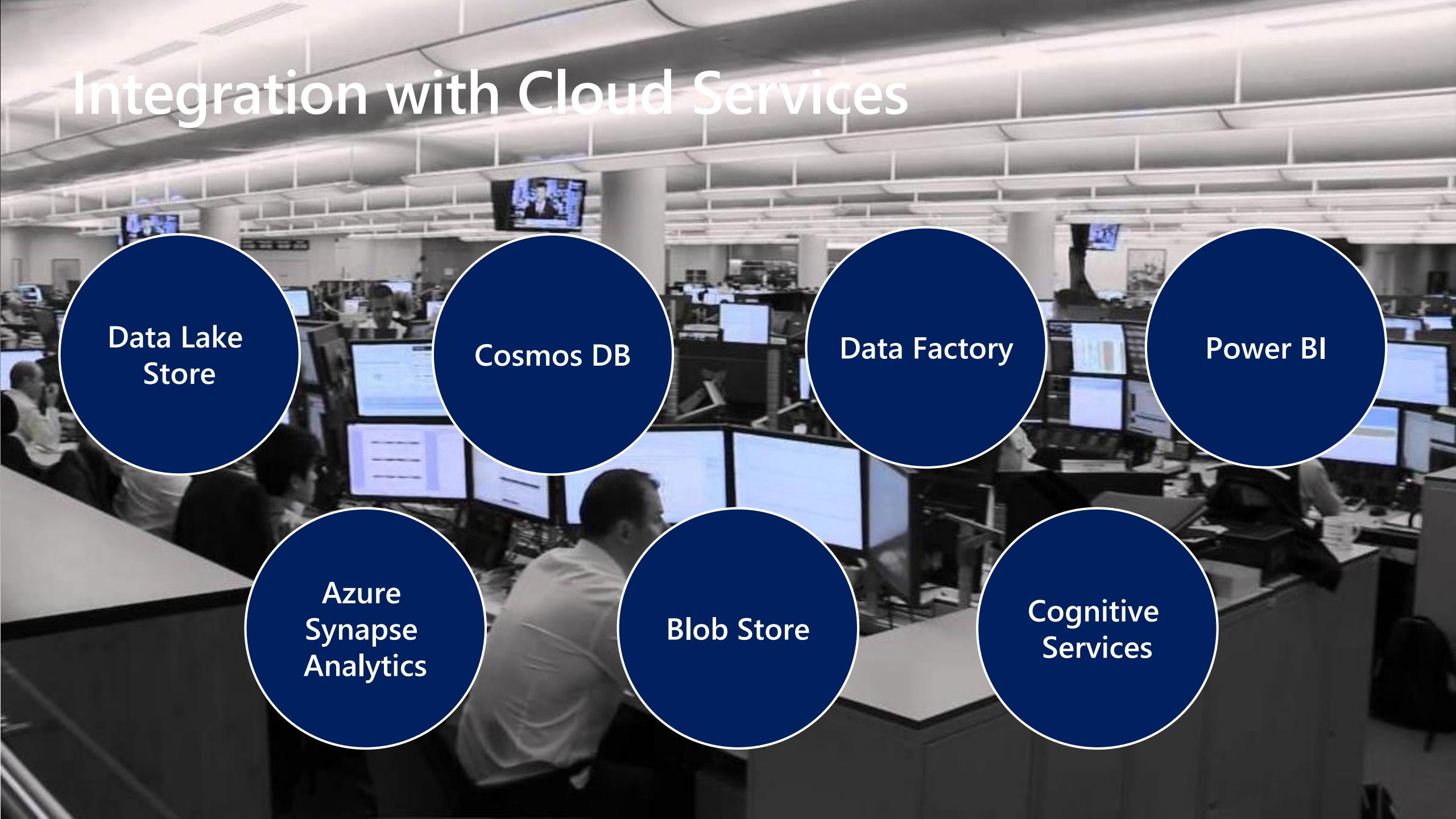
## Azure Databricks components:

- Spark SQL and DataFrames
- Streaming
- Mlib
- GraphX
- Spark Core API

# Enterprise Security



# Integration with Cloud Services



Data Lake Store

Cosmos DB

Data Factory

Power BI

Azure  
Synapse  
Analytics

Blob Store

Cognitive  
Services



# Lesson 02

## Provision Azure Databricks and Workspaces

# Lesson Objectives

- Create your own Azure Databricks workspace
- Create a cluster and notebook in Azure Databricks

# Create an Azure Databricks Workspace.

Home > New > Azure Databricks > Azure Databricks Service

## Azure Databricks Service

\* Workspace name  
ds-mslearn

\* Subscription

\* Resource group   
 Create new  Use existing  
cto\_rg

\* Location  
West Europe

\* Pricing Tier ( [View full pricing details](#) )  
Standard (Apache Spark, Secure with Azur...

Deploy Azure Databricks workspace in your Virtual Network (preview)  
 Yes  No

# Create a Cluster and Notebook in Azure Databricks

Microsoft Azure PORTAL @microsoft.com

Azure Databricks Home Workspace Recents Data Clusters Jobs Search

## Azure Databricks

 Explore the Quickstart Tutorial

Spin up a cluster, run queries on preloaded data, and display results in 5 minutes.

 Import & Explore Data

Quickly import data, preview its schema, create a table, and query it in a notebook.

 Create a Blank Notebook

Create a notebook to start querying, visualizing, and modeling your data.

**Common Tasks**

-  New Notebook
-  Upload Data
-  Create Table
-  New Cluster
-  New Job
-  New MLflow Experiment New
-  Import Library
-  Read Documentation

**Recents**

Recent files appear here as you work.

**Documentation**

-  Databricks Guide
-  Python, R, Scala, SQL
-  Importing Data



# Lesson 03

## Read Data

### using Azure Databricks

# Lesson Objectives

- Use Azure Databricks to access data sources
- Reading Data in Azure Databricks

Use Azure Databricks to access data sources.

Data Lake

Cosmos DB

Event Hubs

SQL Database

Azure  
Synapse  
Analytics

# Reading Data in Azure Databricks.

SQL	DataFrame (Python)
SELECT col_1 FROM myTable	df.select(col("col_1"))
DESCRIBE myTable	df.printSchema()
SELECT * FROM myTable WHERE col_1 > 0	df.filter(col("col_1") > 0)
..GROUP BY col_2	..groupBy(col("col_2"))
..ORDER BY col_2	..orderBy(col("col_2"))
..WHERE year(col_3) > 1990	..filter(year(col("col_3")) > 1990)
SELECT * FROM myTable LIMIT 10	df.limit(10)
display(myTable) (text format)	df.show()
display(myTable) (html format)	display(df)



# Lesson 04

## Perform Transformations with Azure Databricks

# Lesson Objectives

- Performing ETL to populate a data model
- Perform basic transformations
- Perform advanced transformations with user-defined functions

# Performing ETL to populate a data model

The goal of transformation in Extract Transform Load (ETL) is to transform raw data to populate a data model.

Extraction	Data Validation	Transformation	Corrupt Record Handling	Loading Data
Connect to many data stores: <ul style="list-style-type: none"><li>• Postgres</li><li>• SQL Server</li><li>• Cassandra</li><li>• Cosmos DB</li><li>• CSV, Parquet</li><li>• Many more..</li></ul>	Validate that the data is what you expect.	Applying structure and schema to your data to transform it into the desired format.	Built-in functions of Databricks allow you to handle corrupt data such as missing and incomplete information.	Highly effective design pattern involves loading structured data back to DBFS as a parquet file.

Basic transformation

Normalizing Values

Missing/Null data

De-duplication

Pivoting Data frames

# Advanced Transformations

Advanced data transformation using custom and advanced user-defined functions, managing complex tables and loading data into multiple databases simultaneously.

## User-defined functions

This fulfils scenarios when you need to define logic specific to your use case and when you need to encapsulate that solution for reuse. UDFs provide custom, generalizable code that you can apply to ETL workloads when Spark's built-in functions won't suffice.

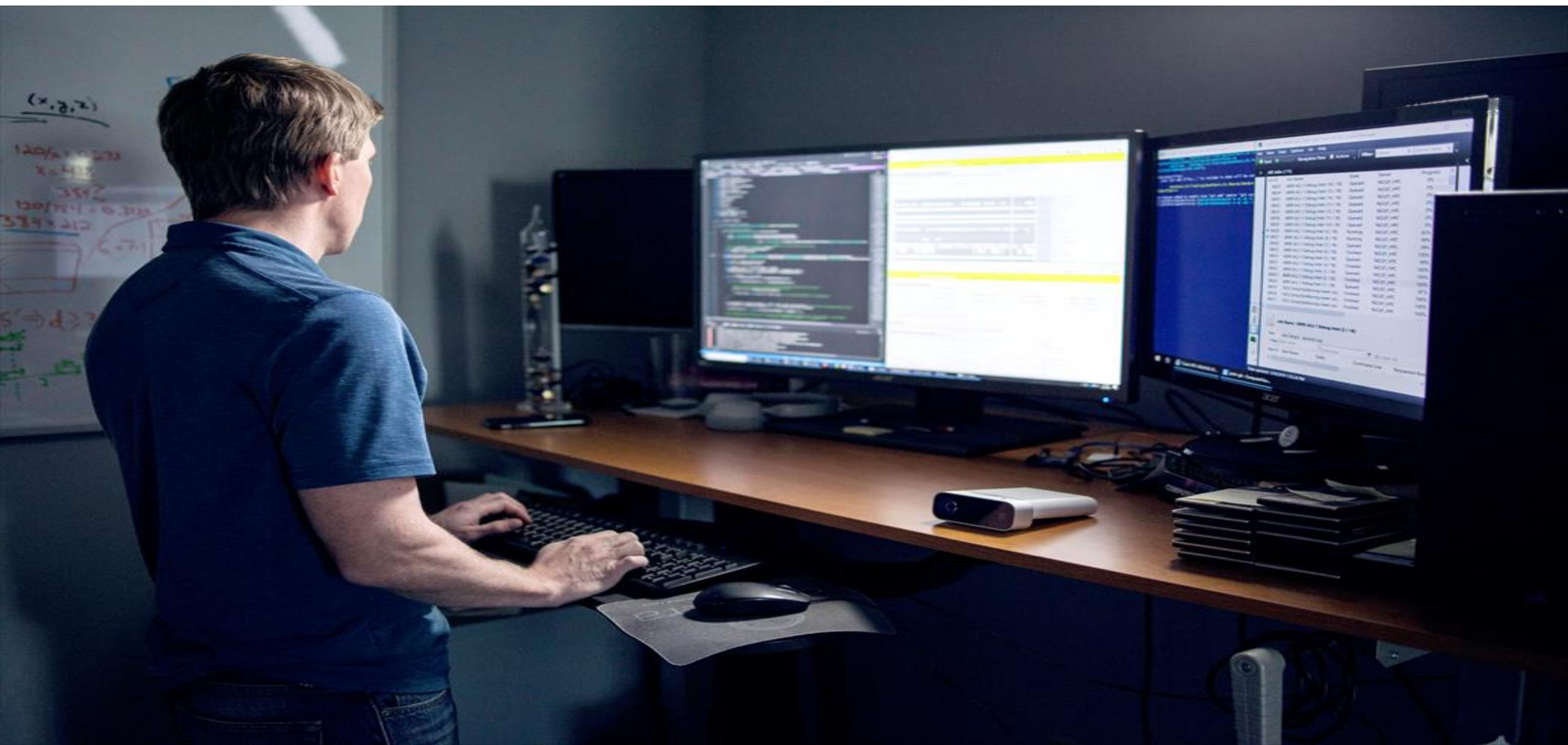
## Joins and lookup tables

A standard (or shuffle) join moves all the data on the cluster for each table to a given node on the cluster. This is an expensive operation. Broadcast joins remedy this situation when one DataFrame is sufficiently small enough to duplicate on each node of the cluster, avoiding the cost of shuffling a bigger DataFrame.

## Multiple databases

Loading transformed data to multiple target databases can be a time-consuming activity. Partitions and slots are options to get optimum performance from database connections. A partition refers to the distribution of data while a slot refers to the distribution of computation.

# Lab: Enabling Team Based Data Science with Azure Databricks





# Module 04:

## Building Globally Distributed Databases with Cosmos DB



# Agenda

- L01 – Create an Azure Cosmos DB database built to scale
- L02 - Insert and query data in your Azure Cosmos DB database
- L03 - Build a .NET Core app for Azure Cosmos DB in Visual Studio Code
- L04 - Distribute your data globally with Azure Cosmos DB



# Lesson 01

## Create an Azure Cosmos DB Database Built to Scale



# Lesson Objectives

- What is Cosmos DB
- Create an Azure Cosmos DB account
- What is a Request Unit
- Choose a partition key
- Create a database and container for NoSQL data in Azure Cosmos DB

# What is Azure Cosmos DB



Scalability



Performance



Availability



Programming  
Models

# Create an Azure Cosmos DB account.

Home > New > Create Azure Cosmos DB Account

## Create Azure Cosmos DB Account

Basics Networking Tags Review + create

Azure Cosmos DB is a globally distributed, multi-model, fully managed database service. [Try it for free](#), for 30 days with unlimited renewals. Go to production starting at \$24/month per database, multiple containers included. [Learn more](#)

**Project Details**

Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources.

Subscription \* chtestao

Resource Group \* Select existing... [Create new](#)

**Instance Details**

Account Name \* Enter account name

API \* Core (SQL)

Apache Spark [None](#) [Notebooks](#) [Notebooks with Apache Spark](#) [Sign up for Apache Spark preview](#)

Location \* (US) West US

Geo-Redundancy [Enable](#) [Disable](#)

Multi-region Writes [Enable](#) [Disable](#)

\*Up to 33% off multi-region writes is available to qualifying new accounts only. Accounts must be created between December 1, 2019 and February 29, 2020. Offer limited to accounts with both account locations and geo-redundancy, and applies only to multi-region writes in those same regions. Both Geo-Redundancy and Multi-region Writes must be enabled in account settings. Actual discount will vary based on number of qualifying regions selected.

# What are Request Units

Throughput is important to ensure you can handle the volume of transactions you need.

Database throughput

Database throughput is the number of reads and writes that your database can perform in a single second

What is a Request Unit

Azure Cosmos DB measures throughput using something called a request unit (RU). Request unit usage is measured per second, so the unit of measure is request units per second (RU/s). You must reserve the number of RU/s you want Azure Cosmos DB to provision in advance

Exceeding throughput limits

If you don't reserve enough request units, and you attempt to read or write more data than your provisioned throughput allows, your request will be rate-limited

# Request unit basics

Item size	Reads/second	Writes/second	Request units
1 KB	500	100	$(500 * 1) + (100 * 5) = 1,000 \text{ RU/s}$
1 KB	500	500	$(500 * 1) + (500 * 5) = 3,000 \text{ RU/s}$
4 KB	500	100	$(500 * 1.3) + (100 * 7) = 1,350 \text{ RU/s}$
4 KB	500	500	$(500 * 1.3) + (500 * 7) = 4,150 \text{ RU/s}$
64 KB	500	100	$(500 * 10) + (100 * 48) = 9,800 \text{ RU/s}$
64 KB	500	500	$(500 * 10) + (500 * 48) = 29,000 \text{ RU/s}$

# Choosing a Partition Key

## Why have a Partition Strategy

Having a partition strategy ensures that when your database needs to grow, it can do so easily and continue to perform efficient queries and transactions

## What is a Partition Key

A partition key is the value by which Azure organizes your data into logical divisions.

## Best Practice.

### Range of Values

The more values your partition key has, the more scalability you have

### Review Queries

To determine the best partition key for a read-heavy workload, review the top three to five queries you plan on using

### Transactional Workloads

For write-heavy workloads, you'll need to understand the transactional needs of your workload

# Creating a Database and a Container in Cosmos DB

Add Container X

**Start at \$24/mo per database, multiple containers included**  
[More details](#)

**\* Database id** ⓘ  
 Create new  Use existing  
Type a new database id

Provision database throughput ⓘ

**\* Throughput (400 - 100,000 RU/s)** ⓘ  
 Autopilot (preview)  Manual  
400

Estimated spend (USD): **\$0.032 hourly / \$0.77 daily** (1 region, 400RU/s, \$0.00008/RU)

**\* Container id** ⓘ  
e.g., Container1

**\* Partition key** ⓘ  
e.g., /address/zipCode

My partition key is larger than 100 bytes

Unique keys ⓘ

+ Add unique key



# Lesson 02

## Insert and Query Data in your Azure Cosmos DB Database

# Lesson Objectives

- Create a product catalog document in the Data Explorer
  - Add data
- Perform Azure Cosmos DB queries
  - Query types
  - Run queries
- Running complex operations on your data
- Working with graph data

# Create a product catalog documents in the Data Explorer.

awcdbstudcto - Data Explorer  
Azure Cosmos DB account

Search (Ctrl+ /)

Overview

Activity log

Access control (IAM)

Tags

Diagnose and solve problems

Quick start

Notifications

Data Explorer

Settings

Replicate data globally

Default consistency

Firewall and virtual networks

SQL API

Products

Scale

Clothing

Items

Settings

Stored Procedures

User Defined Functions

Triggers

ToDoList

Items

Items

SELECT \* FROM c

Edit Filter

	id	...
1	332...	
2	332...	
3	332...	
4		
5		
6		
7		
8		
9		
10		
11		
12		
13		
14		
15		
16		
17		
18		

```
1 "id": "1",  
2 "productId": "332...  
3 "category": "W...  
4 "manufacturer":  
5 "description":  
6 "price": "14.99"  
7 "shipping": {  
8 "weight": 1  
9 "dimensions":  
10 "width":  
11 "height":  
12 "depth":  
13 }  
14 },  
15 "_rid": "Po1tAM...  
16 "_self": "dbs/P...  
17 "_etag": "\\"410...  
18 "
```

# Perform Azure Cosmos DB Queries.

## SELECT Query Basics

```
SELECT <select_list>
[FROM <optional_from_specification>]
[WHERE <optional_filter_condition>]
[ORDER BY <optional_sort_specification>]
[JOIN <optional_join_specification>]
```

### Examples

```
SELECT *
FROM Products p WHERE p.id ="1"
```

```
SELECT p.id, p.manufacturer, p.description
FROM Products p WHERE p.id ="1"
```

```
SELECT p.price, p.description, p.productId
FROM Products p ORDER BY p.price ASC
```

```
SELECT p.productId
FROM Products p JOIN p.shipping
```

# Running complex operations on data

Multiple documents in your database frequently need to be updated at the same time. The way to perform these transactions in Azure Cosmos DB is by using stored procedures and user-defined functions (UDFs)

## Stored Procedures

Stored procedures perform complex transactions on documents and properties. Stored procedures are written in JavaScript and are stored in a collection on Azure Cosmos DB.

## User Defined Functions

User Defined Functions are used to extend the Azure Cosmos DB SQL query language grammar and implement custom business logic, such as calculations on properties and documents.

# Working with Graph Data

```
from gremlin_python.driver import client,  
serializer  
import sys, traceback  
  
CLEANUP_GRAPH = "g.V().drop()  
  
INSERT_NATIONAL_PARK_VERTICES = [  
    "g.addV('Park').property('id',  
'p1').property('name',  
'Yosemite').property('Feature', 'El Capitan')",  
    "g.addV('Park').property('id',  
'p2').property('name', 'Joshua  
Tree').property('Feature', 'Yucca Brevifolia')",  
    "g.addV('State').property('id',  
's1').property('name',  
'California').property('Location', 'USA')",  
    "g.addV('Ecosystem').property('id',  
'e1').property('name', 'Alpine')",  
    "g.addV('Ecosystem').property('id',  
'e2').property('name', 'Desert')",  
    "g.addV('Ecosystem').property('id',  
'e3').property('name', 'High Altitude')"  
]  
  
INSERT_NATIONAL_PARK_EDGES = [  
    "g.V('p1').addE('is in').to(g.V('s1'))",  
    "g.V('p2').addE('is in').to(g.V('s1'))",  
    "g.V('p1').addE('has ecosystem  
of').to(g.V('e1'))",  
    "g.V('p2').addE('has ecosystem  
of').to(g.V('e2'))",  
    "g.V('p1').addE('has ecosystem  
of').to(g.V('e3'))",  
    "g.V('p2').addE('has ecosystem  
of').to(g.V('e3'))"  
]
```



# Lesson 03

## Build a .NET Core App for Azure Cosmos DB in VS Code

# Lesson Objectives

- Create an Azure Cosmos DB account, database, and container in Visual Studio Code using the Azure Cosmos DB extension
- Create an application to store and query data in Azure Cosmos DB
- Use the Terminal in Visual Studio Code to quickly create a console application
- Add Azure Cosmos DB functionality with the help of the Azure Cosmos DB extension for Visual Studio Code

# Creating Azure Cosmos DB in Visual Studio Code

The screenshot shows the Visual Studio Code interface with the Azure extension installed. The left sidebar features icons for File Explorer, Search, Task Manager, and Terminal. The main area has tabs for '1-cosmos-document.json' and 'document.json'. A search bar at the top says 'Retail'. The left pane displays the Azure Cosmos DB explorer, showing databases like 'adventbikes' (MongoDB) and 'ctocdb' (SQL), and collections like 'Products' under 'ctocdb'. A specific document named '1' is selected in the 'Documents' collection. The right pane shows the JSON content of this document:

```
1   "id": "1",
2   "productId": "33218896",
3   "category": "Women's Clothing",
4   "manufacturer": "Contoso Sport",
5   "description": "Quick dry crew neck t-shirt",
6   "price": "14.99",
7   "shipping": {
8     "weight": 1,
9     "dimensions": {
10       "width": 6,
11       "height": 8,
12       "depth": 1
13     }
14   },
15   "_rid": "QSl9ALDRxXUBAAAAAA==",
16   "_self": "dbs/QSl9AA==/colls/QSl9ALDRxXU=/docs/1",
17   "_etag": "\"13000b6a-0000-0700-0000-5c9b619c0000\"",
18   "_attachments": "attachments/",
19   "_ts": 1553686941
20 }
```

# Working with documents programmatically

CreateDocument  
Async

ReadDocument  
Async

ReplaceDocument  
Async

UpsertDocument  
Async

DeleteDocument  
Async



The collage consists of nine square images arranged in a grid. Top row: 1. Students in a classroom setting. 2. A man in a plaid shirt pointing at a whiteboard. 3. Two men working at a desk with multiple monitors displaying code. 4. Two men in a factory or industrial setting looking at a large screen. Middle row: 5. A woman looking down at her smartphone. 6. A woman in a green sweater writing on a clipboard. Bottom row: 7. A person wearing headphones playing a video game on a laptop. 8. A person using a tablet to draw on a wall covered in technical diagrams. 9. A woman working on a computer with a monitor showing data visualizations.

# Lesson 04

## Distribute your data globally with Azure Cosmos DB



# Lesson Objectives

- Learn about the benefits of writing and replicating data to multiple regions around the world
- Cosmos DB multi-master replication
- Cosmos DB failover management
- Change the consistency setting for your database

# Benefits of writing and replicating data to multiple regions

Home > Resource groups > cto\_rg > ctocdb > Replicate data globally

## Replicate data globally

ctocdb

Save Discard Manual Failover Automatic Failover

Click on a location to add or remove regions from your Azure Cosmos DB account.

\* Each region is billable based on the throughput and storage for the account. [Learn more](#)



Configure regions

Configure the regions available for reads and writes. [+ Add region](#)

REGIONS	READS ENABLED	WRITES ENABLED	
West US	✓	✓	
UK South	✓	✓	
Japan West	✓	✓	
South Africa North	✓	✓	

# Cosmos DB multi-master replication



# Cosmos DB failover management

Automated fail-over is a feature that comes into play when there's a disaster or other event that takes one of your read or write regions offline, and it redirects requests from the offline region to the next most prioritized region.

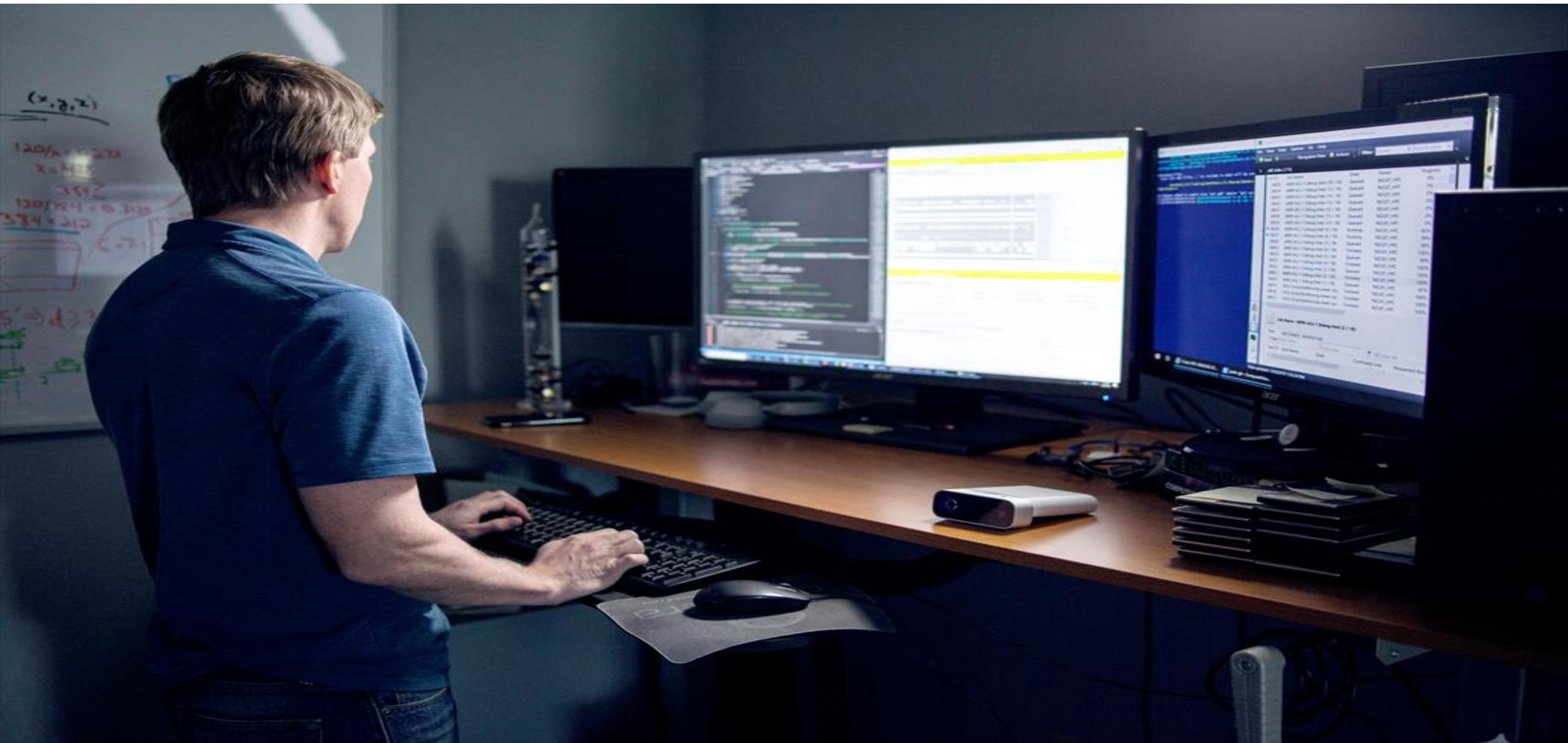
## Read region outage

Azure Cosmos DB accounts with a read region in one of the affected regions are automatically disconnected from their write region and marked offline

## Write region outage

If the affected region is the current write region and automatic fail-over is enabled, then the region is automatically marked as offline. Then, an alternative region is promoted as the write region

# Lab: Building Globally Distributed Databases with Cosmos DB



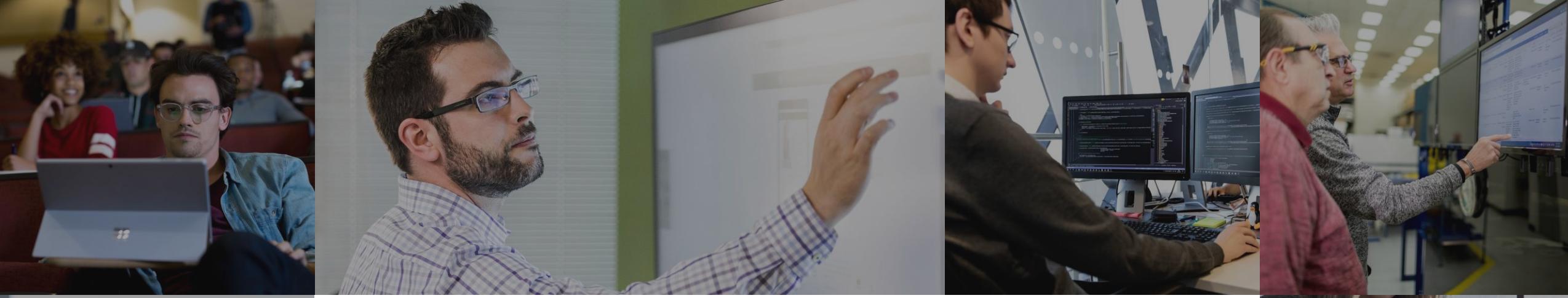
# Module 05: Working with Relational Data Stores in the Cloud

Start : 14.50 Uhr.



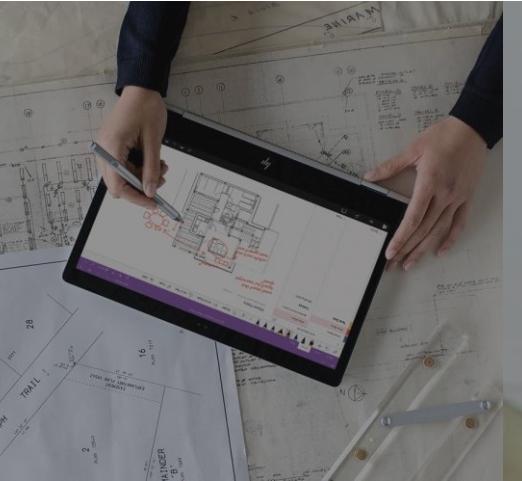
# Agenda

- L01 - Work with Azure SQL Database
- L02 - Work with Azure Synapse Analytics
- L03 - Provision and query data in Azure Synapse Analytics
- L04 - Import data into Azure Synapse Analytics using PolyBase



# Lesson 01

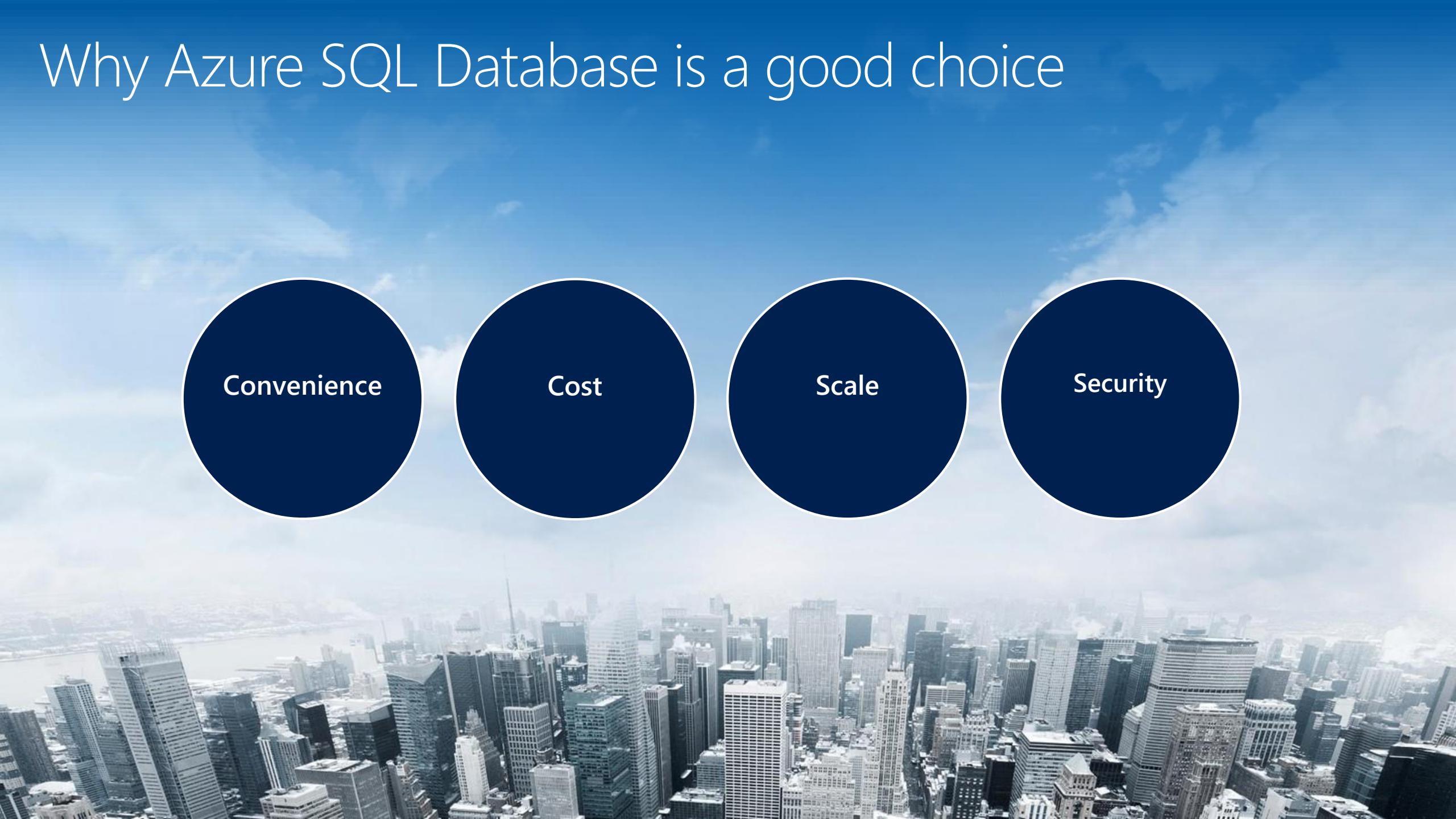
## Azure SQL Database



# Lesson Objectives

- Why Azure SQL Database is a good choice for running your relational database
- What configuration and pricing options are available for your Azure SQL database
- How to create an Azure SQL database from the portal
- How to use Azure Cloud Shell to connect to your Azure SQL database, add a table, and work with data

# Why Azure SQL Database is a good choice

The background of the slide features a wide-angle aerial photograph of a dense urban skyline, likely New York City, viewed from a high vantage point. The city is filled with numerous skyscrapers of varying heights, with some prominent buildings like the One World Trade Center clearly visible. The sky above is a vibrant blue with scattered white, wispy clouds.

Convenience

Cost

Scale

Security

# Azure SQL Database configuration options

When you create your first Azure SQL database, you also create an *Azure SQL logical server*. Think of a logical server as an administrative container for your databases.

## DTUs

DTU stands for Database Transaction Unit and is a combined measure of compute, storage, and IO resources. Think of the DTU model as a simple, preconfigured purchase option

## vCores

vCore gives you greater control over what compute and storage resources you create and pay for. vCore model enables you to configure resources independently

## SQL elastic pools

SQL elastic pools relate to eDTUs. They enable you to buy a set of compute and storage resources that are shared among all the databases in the pool. Each database can use the resources they need

## SQL Managed Instances

The SQL managed instance creates a database with near 100% compatibility with the latest SQL Server on-premises Enterprise Edition database engine, useful for SQL Server customers who would like to migrate on-premises servers instance in a "lift and shift" manner

## Create SQL Database

Microsoft

[Basics](#) • [Networking](#) [Additional settings](#) [Tags](#) [Review + create](#)

Create a SQL database with your preferred configurations. Complete the Basics tab then go to Review + Create to provision with smart defaults, or visit each tab to customize. [Learn more](#)

## Project details

Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources.

Subscription \* ⓘ

Resource group \* ⓘ

 [Create new](#)

## Database details

Enter required settings for this database, including picking a logical server and configuring the compute and storage resources

Database name \*

Server \* ⓘ

 [Create new](#) The value must not be empty.

Want to use SQL elastic pool? \* ⓘ

 Yes  No

Compute + storage \* ⓘ

Please select a server first.

[Configure database](#)

# Create an Azure SQL Database.

Use Azure Cloud Shell to connect to your Azure SQL database

```
sqlcmd -S tcp:contoso-1.database.windows.net,1433  
-d Logistics -U martina -P "password1234$" -N -l 30
```

```
CREATE TABLE Drivers (DriverID int, LastName  
varchar(255), FirstName varchar(255), OriginCity  
varchar(255)); GO
```

```
SELECT name FROM sys.tables; GO
```

```
INSERT INTO Drivers (DriverID, LastName,  
FirstName, OriginCity) VALUES (123, 'Zirne', 'Laura',  
'Springfield'); GO
```



# Lesson 02

## Azure Synapse Analytics

# Lesson Objectives

- Explain Azure Synapse Analytics
- Explain Azure Synapse Analytics features
- Types of solution workloads
- Explain Massively Parallel Processing concepts
- Compare table geometries

# Azure Synapse Analytics

## What is Azure Synapse Analytics

A unified environment by combining the enterprise data warehouse of SQL, the Big Data analytics capabilities of Spark, and data integration technologies to ease the movement of data between both, and from external data sources.

## Data warehouse capabilities.

### SQL Analytics

A centralized data warehouse store that provides a relational analytics and decision support services across the whole enterprise

### SQL Pools

CPU, memory, and IO are bundled into units of compute scale called SQL, determined by Data Warehousing Units (DWU)

### Future features

Will include a Spark engine, a data integration and Azure Synapse Analytics Studio

# Azure Synapse Analytics features

## Workload Management

This capability is used to prioritize the query workloads that take place on the server using Workload Management. This involves three components:

- Workload Groups
- Workload Classification
- Workload Importance

## Result-Set Cache

Result-set caching can be used to improve the performance of the queries that retrieve these results. When result-set caching is enabled, the results of the query are cached in the SQL pool storage.

## Materialized Views

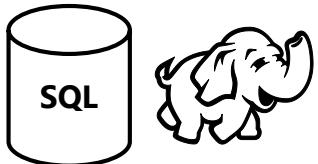
A materialized view pre-computes, stores, and maintains its data like a table. They are automatically updated when data in underlying tables are changed.

## SSDT CI/CD support

Database project support in SQL Server Data Tools (SSDT) allows teams of developers to collaborate over a version-controlled Azure Synapse Analytics, and track, deploy and test schema changes

# Types of solution workloads

The modern data warehouse extends the scope of the data warehouse to serve Big Data that's prepared with techniques beyond relational ETL



## Modern data warehousing

---

"We want to integrate all our data—including Big Data—with our data warehouse"



## Advanced analytics

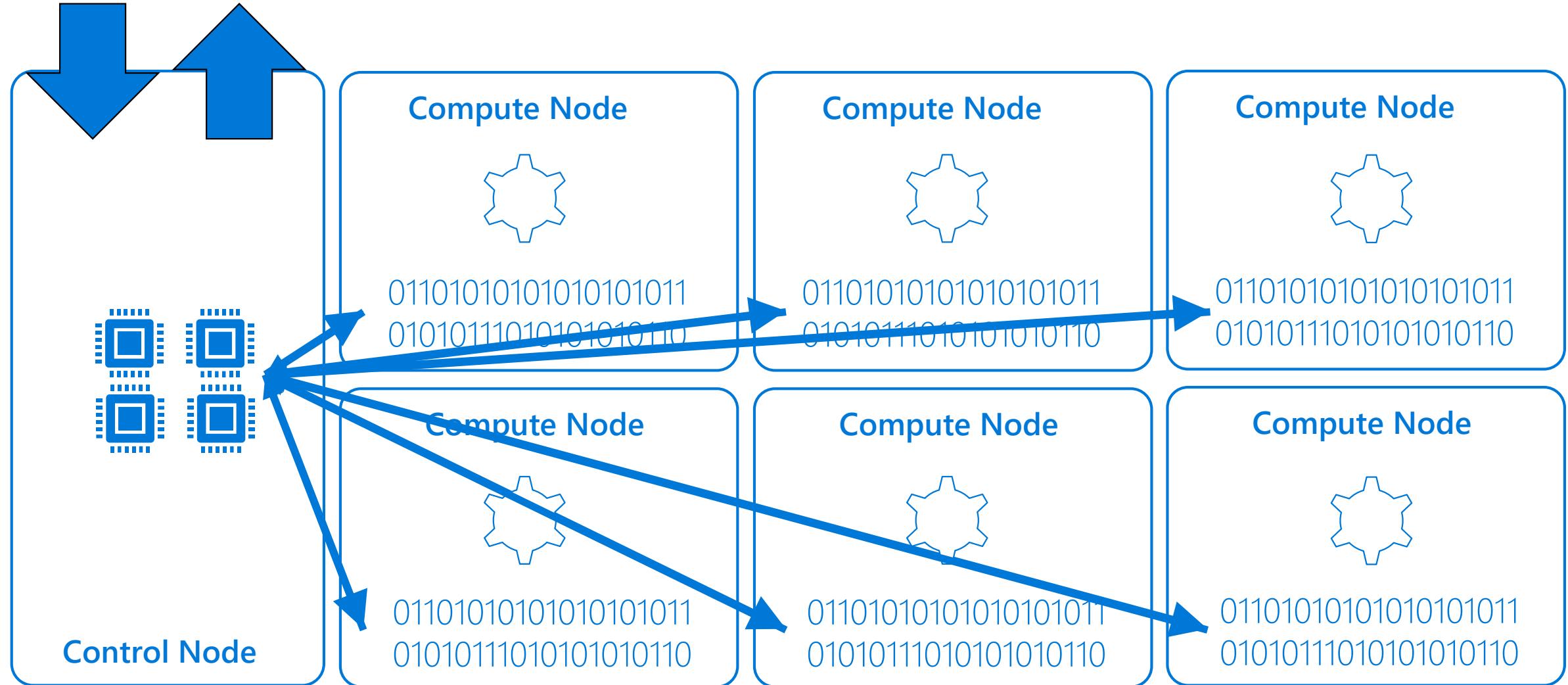
"We're trying to predict when our customers churn"



## Real-time analytics

"We're trying to get insights from our devices in real-time"

# Massively Parallel Processing (MPP) concepts





# Lesson 03

## Creating and Querying an Azure Synapse Analytics

# Lesson Objectives

- Create an Azure Synapse Analytics sample database
- Query the sample database with the SELECT statement and its clauses
- Use the queries in different client applications such as SQL Server Management Studio, and PowerBI

# Create an Azure Synapse Analytics

Home > New > Azure Synapse Analytics (formerly SQL DW) > SQL Data Warehouse

## SQL Data Warehouse

Welcome to Azure Synapse Analytics (formerly known as Azure SQL Data Warehouse). [Learn more](#)

**Basics** • Additional settings \* Tags Review + create

Create a SQL data warehouse with your preferred configurations. Complete the Basics tab then go to Review + Create to provision with smart defaults, or visit each tab to customize. [Learn more](#)

### Project details

Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources.

Subscription \*  Resource group \*   
[Create new](#)

### Data warehouse details

Enter required settings for this data warehouse, including picking a logical server and configuring the performance level.

Data warehouse name \*

Server \*   
[Create new](#)

✖ The value must not be empty.

Performance level \*

# Perform Azure Synapse Analytics Queries.

## **SELECT Query Basics**

```
SELECT <select_list>
[FROM <optional_from_specification>]
[WHERE <optional_filter_condition>]
[ORDER BY <optional_sort_specification>]
[JOIN <optional_join_specification>]
```

### **Examples**

```
SELECT *
FROM Products p WHERE p.id ="1"
```

```
SELECT p.id, p.manufacturer, p.description
FROM Products p WHERE p.id ="1"
```

```
SELECT p.price, p.description, p.productId
FROM Products p ORDER BY p.price ASC
```

```
SELECT p.productId
FROM Products p JOIN p.shipping
```

Perform  
Azure Synapse Analytics  
Queries.

## Create Table as Select (CTAS)

Used in parallel data loads

```
CREATE TABLE
[ database_name . [ schema_name ] . |
schema_name. ] table_name
    [ ( { column_name } [ ,...n ] ) ]
WITH ( DISTRIBUTION =
    { HASH( distribution_column_name )
        REPLICATE | ROUND_ROBIN }
    [ , <CTAS_table_option> [ ,...n ] ]
)
AS <select_statement> [ ; ]
```

### Example

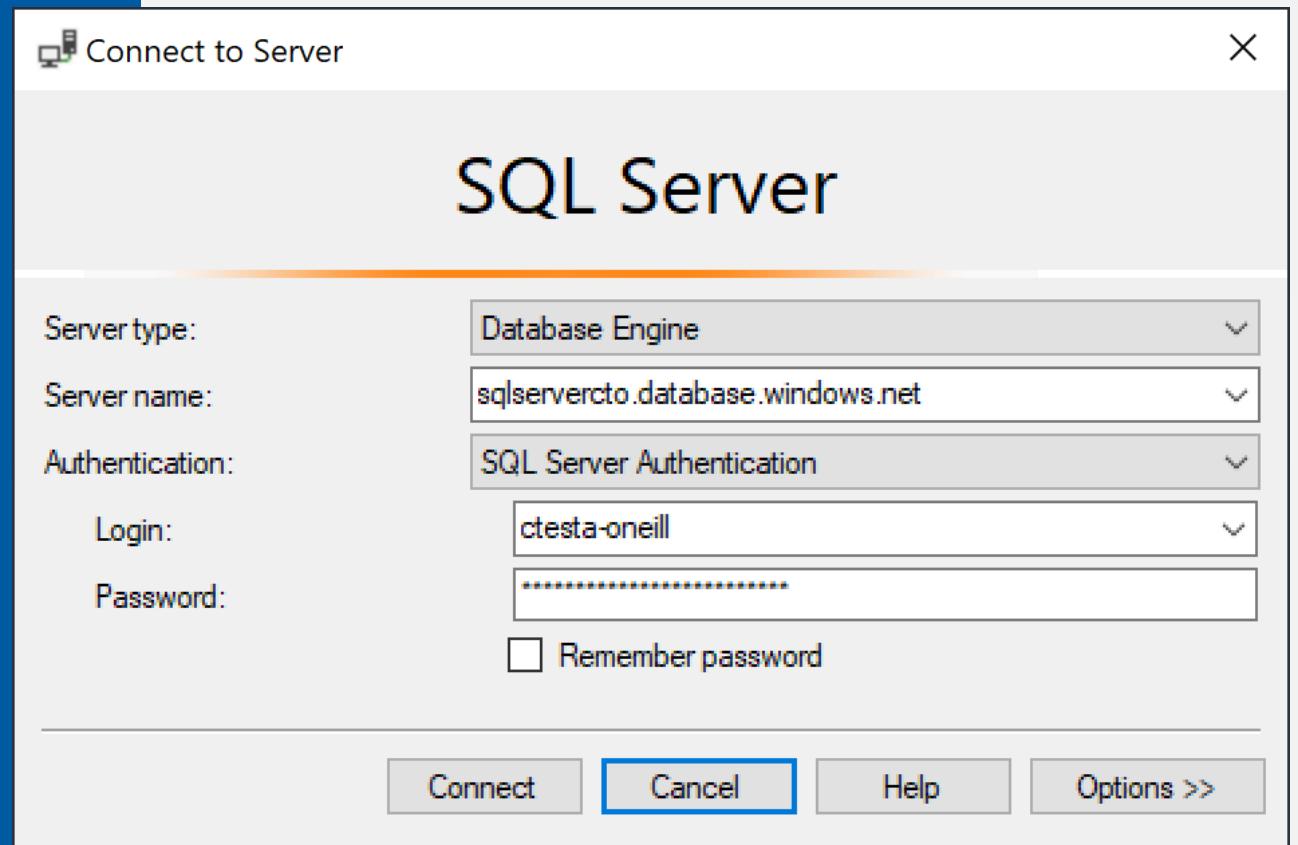
```
CREATE TABLE FactInternetSales_Copy
```

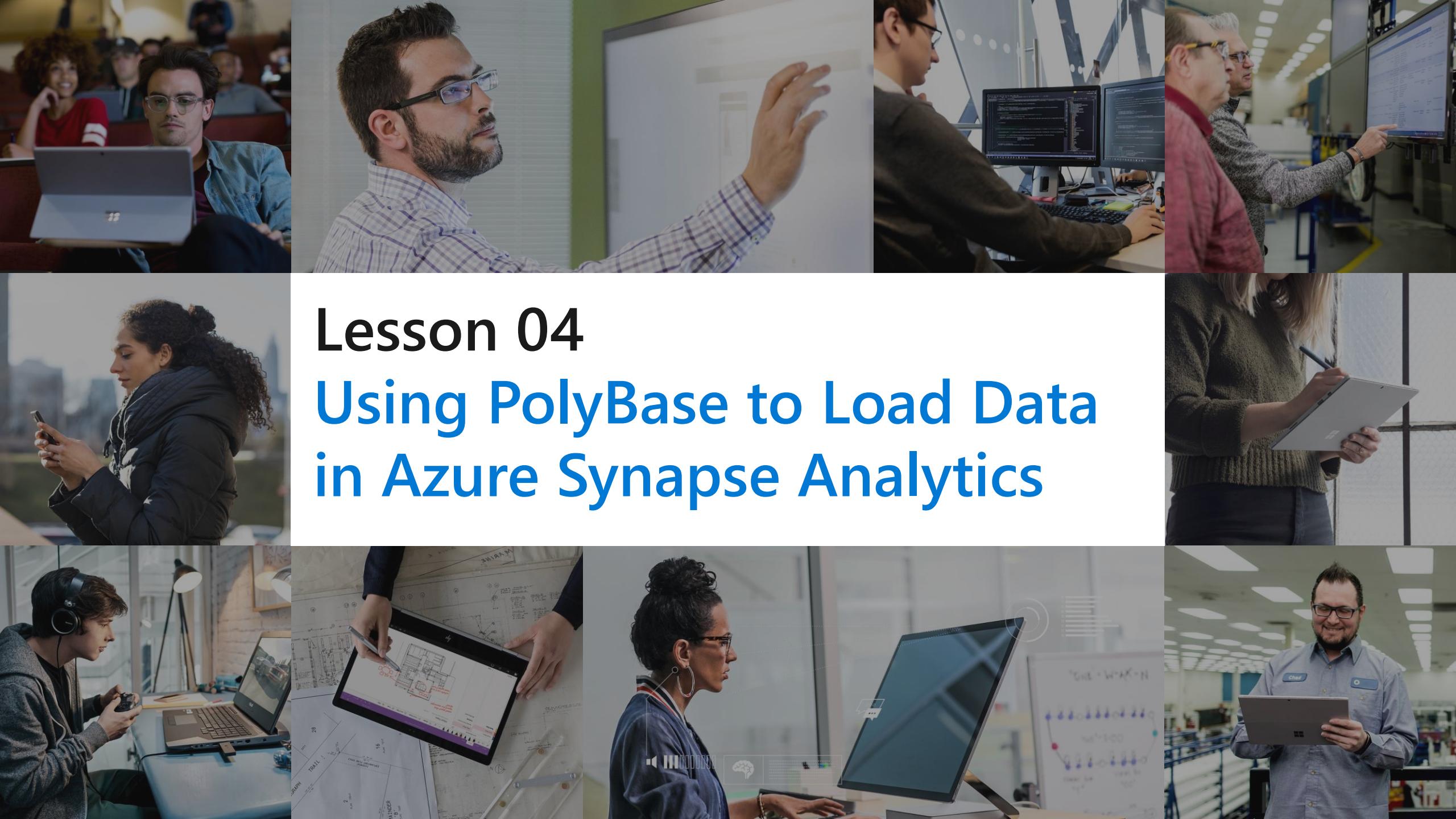
```
WITH
```

```
(DISTRIBUTION = HASH(SalesOrderNumber))
```

```
AS SELECT * FROM FactInternetSales
```

Querying with different client applications.





# Lesson 04

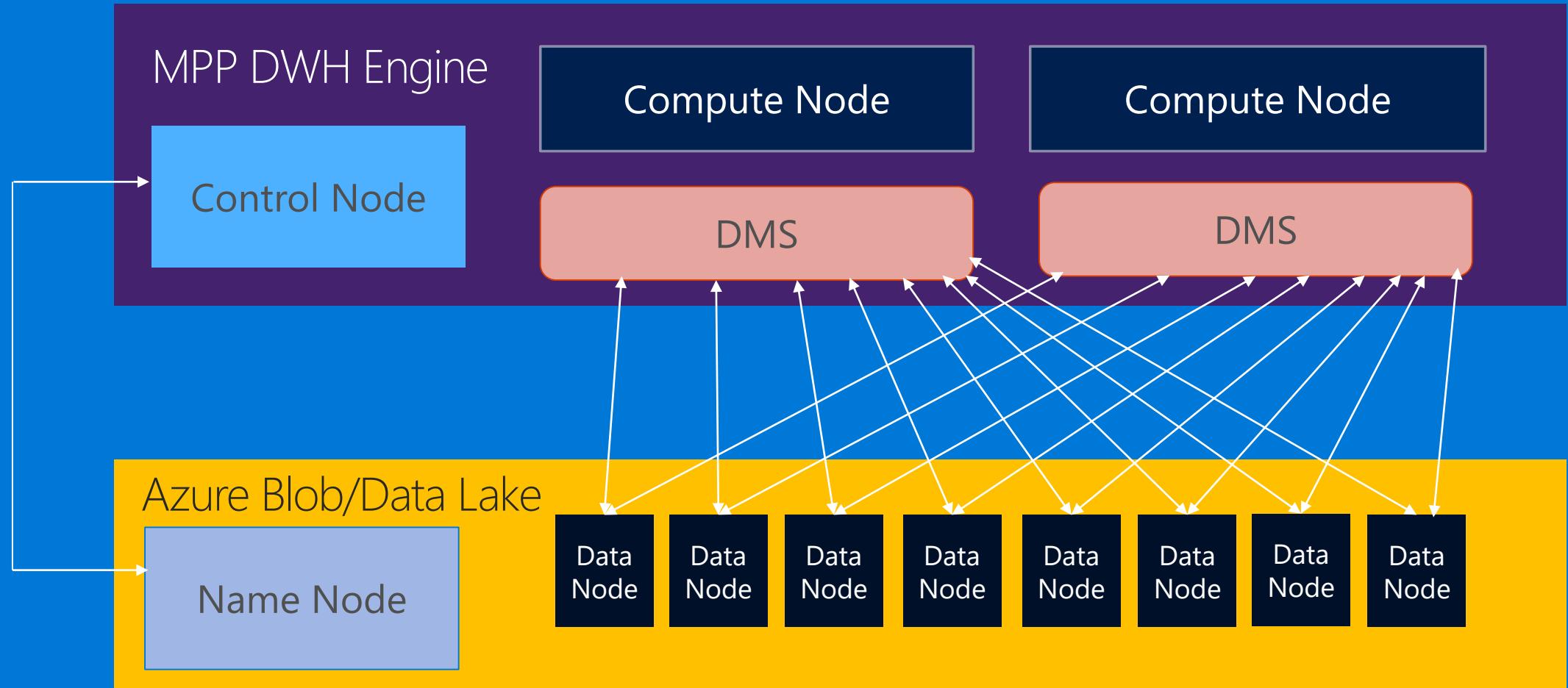
## Using PolyBase to Load Data in Azure Synapse Analytics

# Lesson Objectives

- Explore how PolyBase works
- Upload text data to Azure Blob store
- Collect the security keys for Azure Blob store
- Create an Azure Synapse Analytics
- Import data from Blob Storage to the Data Warehouse

# How PolyBase works

## The MPP engine's integration method with PolyBase



## Create storage account

[Basics](#) [Advanced](#) [Tags](#) [Review + create](#)

Azure Storage is a Microsoft-managed service providing cloud storage that is highly available, secure, durable, scalable, and redundant. Azure Storage includes Azure Blobs (objects), Azure Data Lake Storage Gen2, Azure Files, Azure Queues, and Azure Tables. The cost of your storage account depends on the usage and the options you choose below. [Learn more](#)

## PROJECT DETAILS

Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources.

\* Subscription

\* Resource group

[Create new](#)

## INSTANCE DETAILS

The default deployment model is Resource Manager, which supports the latest Azure features. You may choose to deploy using the classic deployment model instead. [Choose classic deployment model](#)

\* Storage account name [i](#)

\* Location

 Standard  PremiumPerformance [i](#)Account kind [i](#)Replication [i](#)Access tier (default) [i](#) Cool  Hot[Review + create](#)[Previous](#)[Next : Advanced >](#)

# Collect the Storage keys

toazureblob - Access keys

keys

Use access keys to authenticate your applications when making requests to this Azure storage account. Store your access keys securely - for example, using Azure K Vault - and don't share them. We recommend regenerating your access keys regularly. You are provided two access keys so that you can maintain connections using one key while regenerating the other.

When you regenerate your access keys, you must update any Azure resources and applications that access this storage account to use the new keys. This action will interrupt access to disks from your virtual machines. [Learn more](#)

Storage account name  
ctoazureblob

**key1** 

Key  
eU7...Cg==

Connection string  
Def...9YrQ...

**key2** 

Key  
NWD...VpUgB5w==

Connection string  
Def...Ns6...



# Create an Azure Synapse Analytics

Home > New > Azure Synapse Analytics (formerly SQL DW) > SQL Data Warehouse

## SQL Data Warehouse

Welcome to Azure Synapse Analytics (formerly known as Azure SQL Data Warehouse). [Learn more](#)

**Basics** • Additional settings \* Tags Review + create

Create a SQL data warehouse with your preferred configurations. Complete the Basics tab then go to Review + Create to provision with smart defaults, or visit each tab to customize. [Learn more](#)

### Project details

Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources.

Subscription \*  Resource group \*  [Create new](#)

### Data warehouse details

Enter required settings for this data warehouse, including picking a logical server and configuring the performance level.

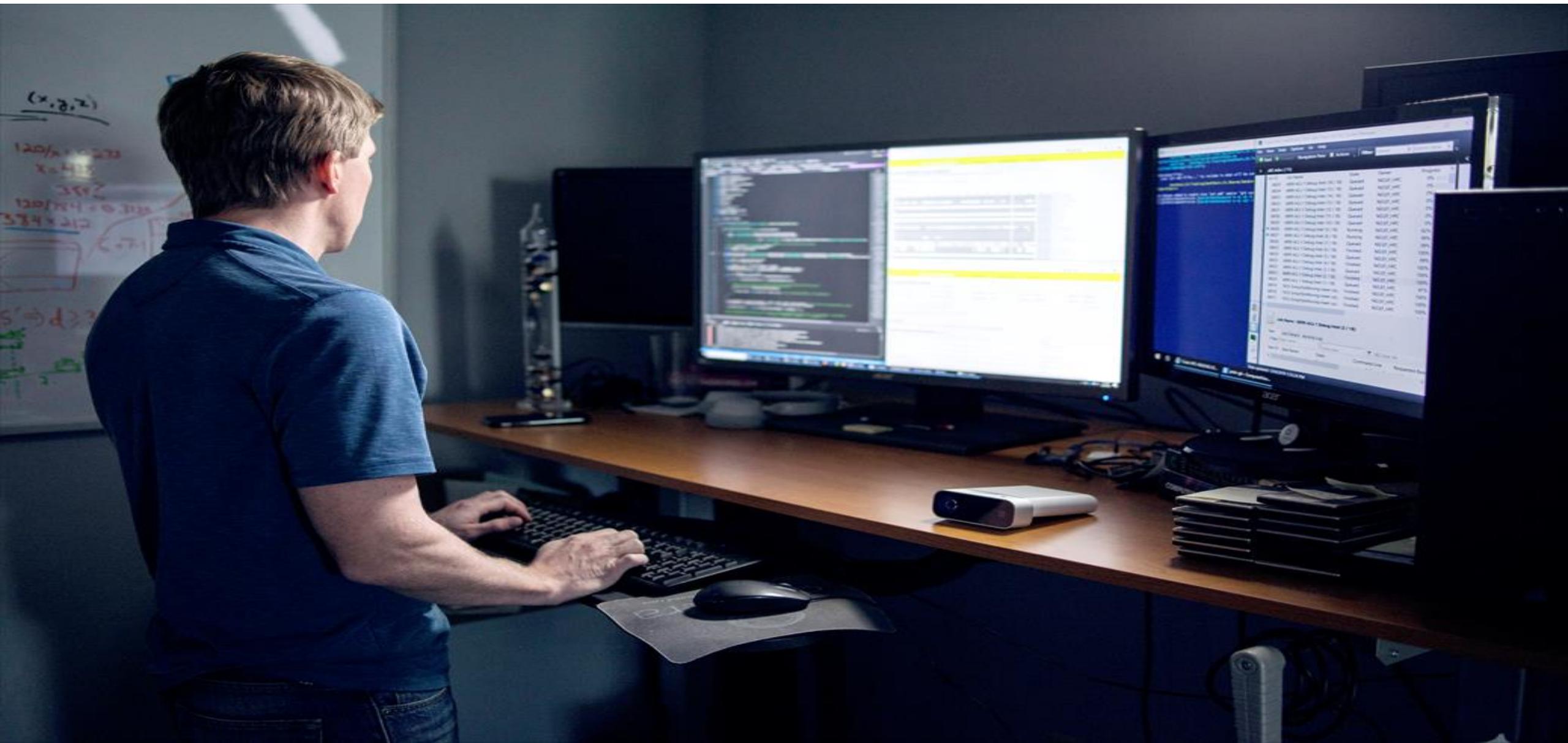
Data warehouse name \*

Server \*  [Create new](#)

✖ The value must not be empty.

Performance level \*

# Lab: Working with Relational Data Stores in the Cloud



# Module 06:

## Performing Real-Time Analytics with Stream Analytics



# Agenda

- L01 - Data streams and event processing
- L02 - Data ingestion with Event Hubs
- L03 - Processing data with Stream Analytics Jobs



The background of the slide features a collage of nine images. Top row: 1. Students in a classroom setting. 2. A man in a plaid shirt interacting with a large whiteboard. 3. Two men working at a desk with multiple monitors displaying code. 4. Two men in a factory or industrial setting looking at a large screen. Middle row: 5. A woman looking down at her phone. 6. A woman writing on a clipboard. Bottom row: 7. A person wearing headphones playing a video game. 8. A person using a tablet to draw on a wall with engineering plans. 9. A woman working on a computer with a brain icon overlay. 10. A man in a factory holding a tablet.

# Lesson 01

## Data Streams and Event Processing

# Lesson Objectives

- Explain data streams
- Explain event processing
- Learn about processing events with 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

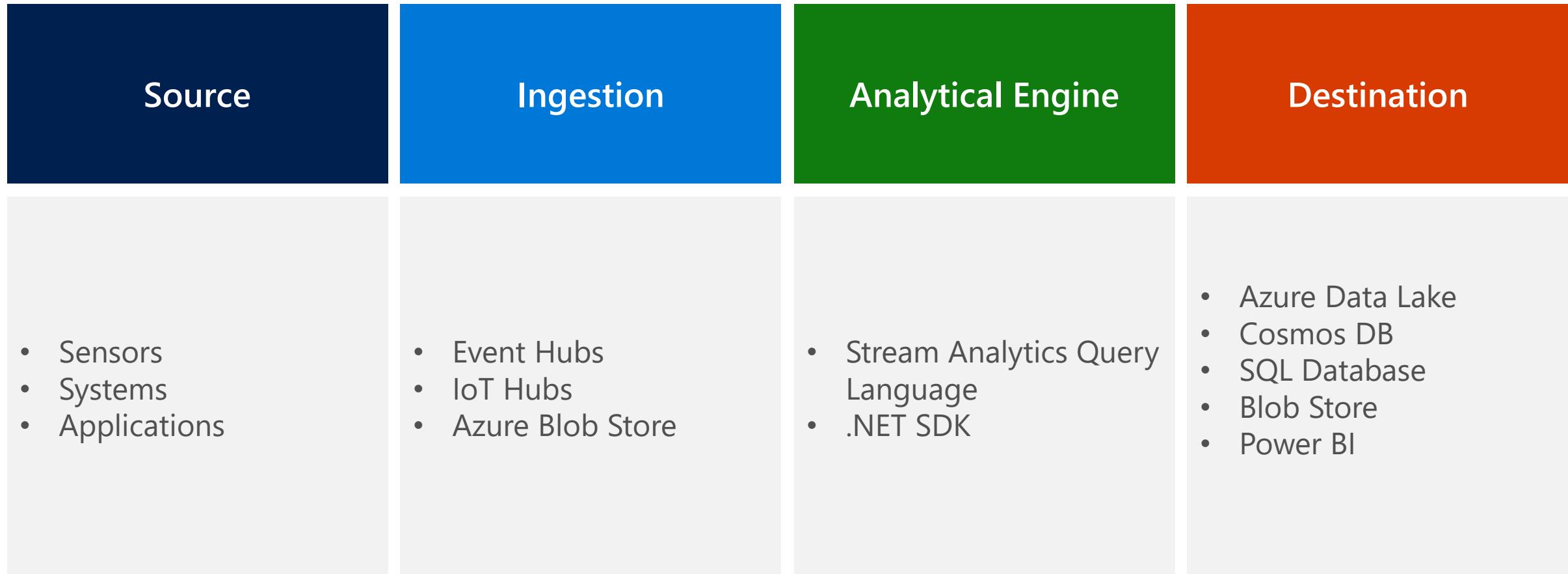
An engine to consume event data streams and deriving insights from them. Depending on the problem space, event processors either process one incoming event at a time (such as a heart rate monitor) or process multiple events at a time (such as a highway toll lane sensor)

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.





The background of the slide features a collage of nine images. Top row: 1. Students in a classroom setting. 2. A man in a plaid shirt interacting with a large whiteboard. 3. Two men working at a desk with multiple monitors displaying code. 4. Two men in a factory or industrial setting looking at a large screen. Middle row: 5. A woman looking down at her phone. 6. A woman writing on a clipboard. Bottom row: 7. A person wearing headphones playing a video game. 8. A person using a tablet to draw on a wall map. 9. A woman working at a desk with a computer monitor displaying data visualizations.

# Lesson 02

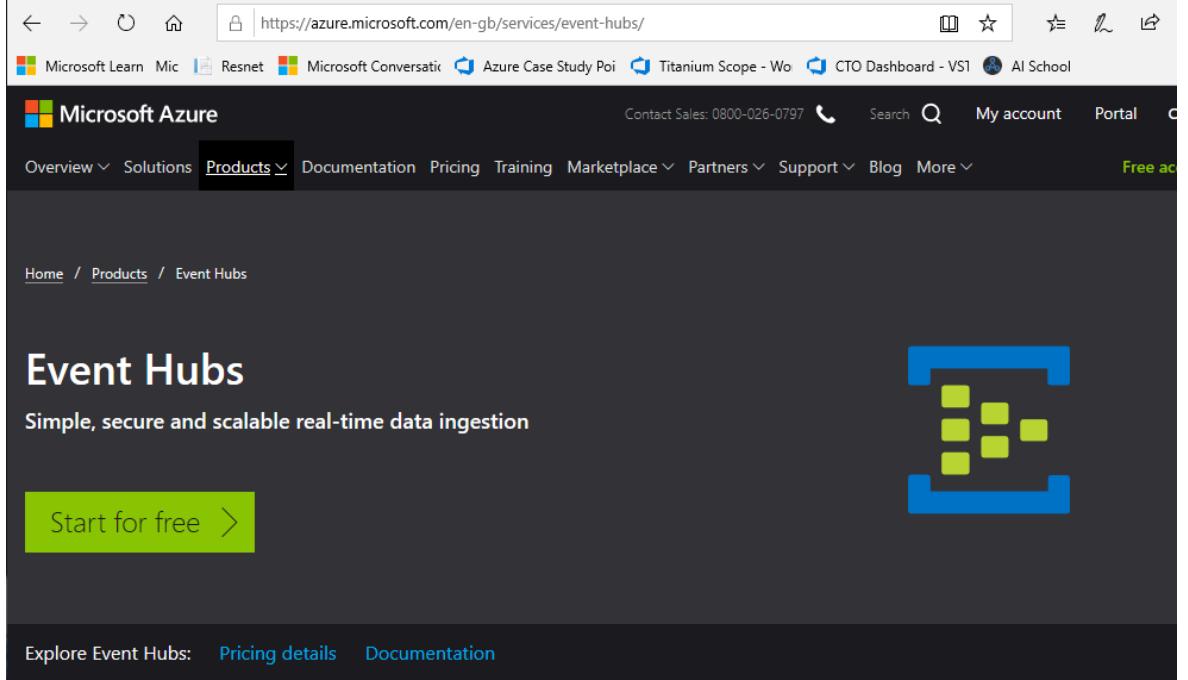
## Data Ingestion with Event Hubs

# Lesson Objectives

- Describe Azure Event Hubs
- Create an Event Hub
- Evaluate the performance of an Event Hub
- Configure applications to use an Event Hub

# Azure Event Hubs

*"Azure Event Hubs is a highly scalable publish-subscribe service that can ingest millions of events per second and stream them into multiple applications."*



The screenshot shows the Microsoft Azure website at https://azure.microsoft.com/en-gb/services/event-hubs/. The top navigation bar includes links for Microsoft Learn, Resnet, Microsoft Conversatric, Azure Case Study Poi, Titanium Scope - Wo, CTO Dashboard - VS1, AI School, Contact Sales, Search, My account, Portal, and Free acc. The main content area features a large title "Event Hubs" with the subtitle "Simple, secure and scalable real-time data ingestion". A green "Start for free >" button is prominent. Below this, there's a section titled "Explore Event Hubs" with links to "Pricing details" and "Documentation". To the right, there's a large blue and yellow graphic of a stylized "T" shape made of squares. At the bottom, there's a paragraph about Event Hubs being a fully managed, real-time data ingestion service, and another paragraph about integrating with other Azure services like Apache Kafka.

Event Hubs

Simple, secure and scalable real-time data ingestion

Start for free >

Explore Event Hubs: [Pricing details](#) [Documentation](#)

Event Hubs is a fully managed, real-time data ingestion service that's simple, trusted and scalable. Stream millions of events per second from any source to build dynamic data pipelines and immediately respond to business challenges. Keep processing data during emergencies using the [geo-disaster recovery](#) and geo-replication features.

Integrate seamlessly with other Azure services to unlock valuable insights. Allow existing Apache Kafka clients and applications to talk to Event Hubs without any code changes – you get a managed Kafka experience without having to manage your own clusters. Experience real-time data ingestion and microbatching on the same stream.

[Link to video >](#)

# Create an Event Hub

Microsoft Azure Search resources, services, and docs (G+/) ...

All services > myehubrg > New > Event Hubs > Create Namespace

## Create Namespace

Event Hubs

[Basics](#) [Features](#) [Tags](#) [Review + create](#)

**PROJECT DETAILS**

Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources.

Subscription \*

Resource group \*  [Create new](#)

**INSTANCE DETAILS**

Enter required settings for this namespace, including a price tier and configuring the number of throughput units.

Namespace name \*  .servicebus.windows.net

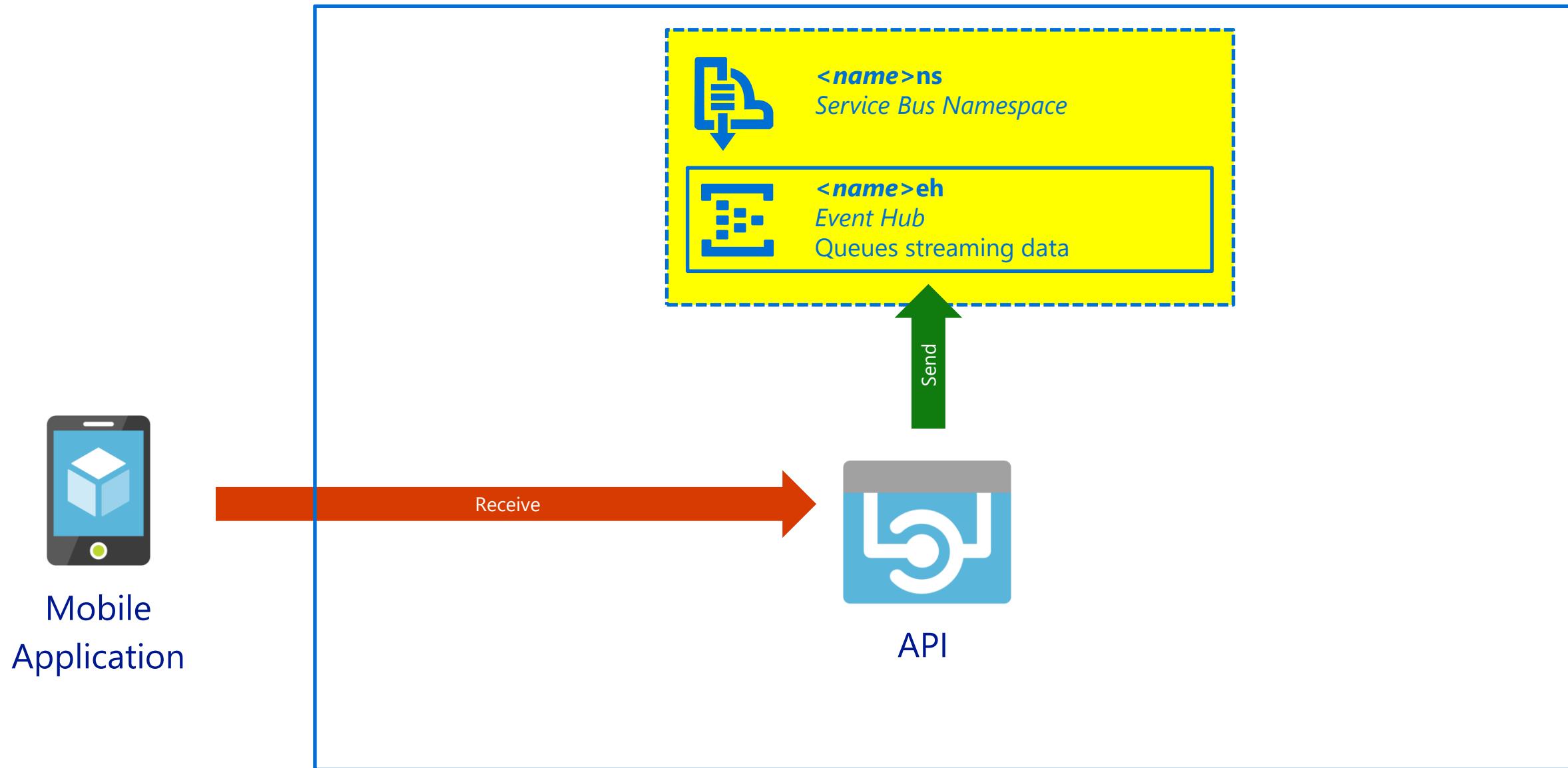
Location \*

Pricing tier ([View full pricing details](#)) \*

Throughput Units \*

[Review + create](#) [< Previous](#) [Next: Features >](#)

# Configure Applications to use Event Hubs





The background of the slide features a collage of nine images. Top row: 1. Students in a classroom setting. 2. A man in a plaid shirt interacting with a large whiteboard. 3. Two men working at a desk with multiple computer monitors. 4. Two men in a factory or industrial setting looking at a large screen. Middle row: 5. A woman looking down at her smartphone. 6. A woman in a green sweater writing on a clipboard. Bottom row: 7. A person wearing headphones playing a video game. 8. A person using a tablet to draw on a wall with engineering plans. 9. A woman working on a computer with a brain scan graphic overlaid.

# Lesson 03

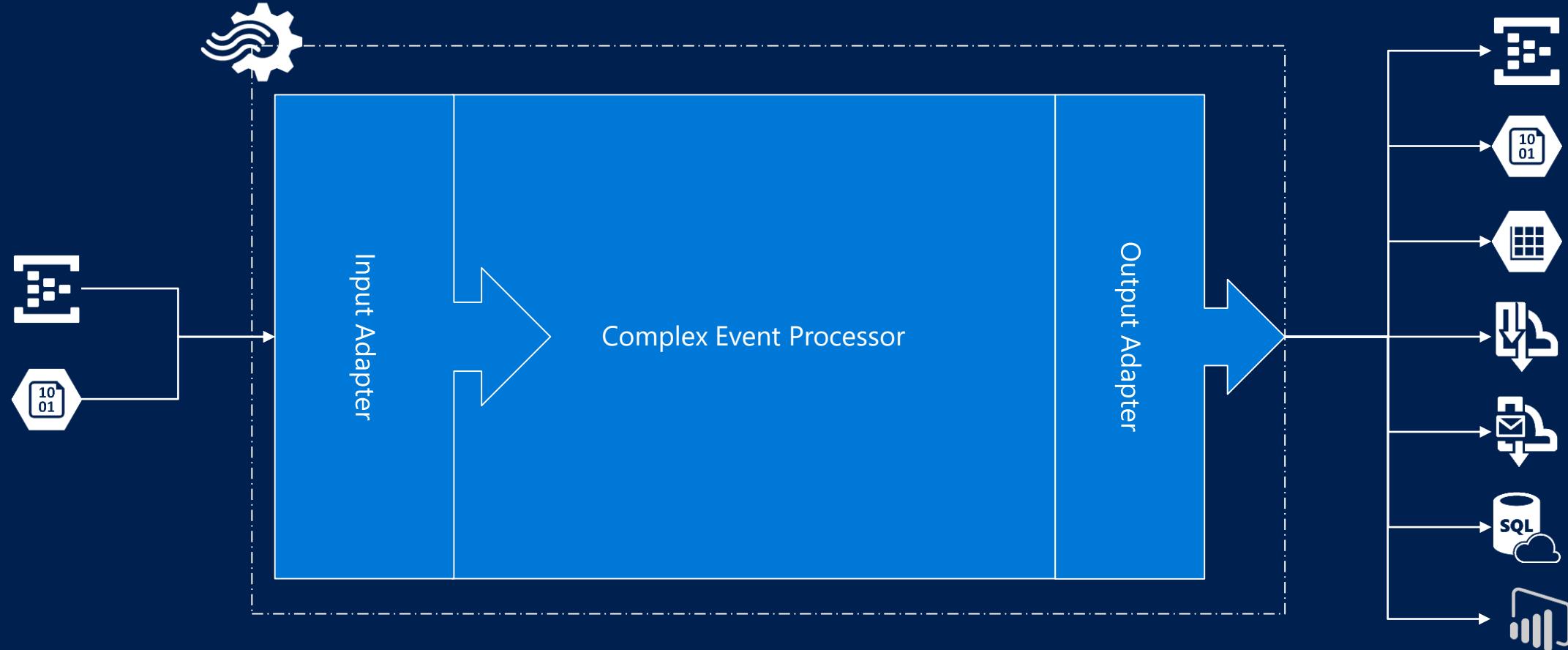
## Processing Data with Stream Analytics Jobs

# Lesson Objectives

- Explore the Streaming Analytics workflow
- Create a Stream Analytics Job
- Configure a Stream Analytics job input
- Configure a Stream Analytics job output
- Write a transformation query
- Start a Stream Analytics job

# Azure Stream Analytics Workflow

*Complex Event Processing of Stream Data in Azure*



# Create Stream Analytics Service

- Job name
- Subscription
- Resource group
- Location

Home > New > Stream Analytics job > New Stream Analytics job

## New Stream Analytics job

\* Job name  
cto-asa-job1

\* Subscription

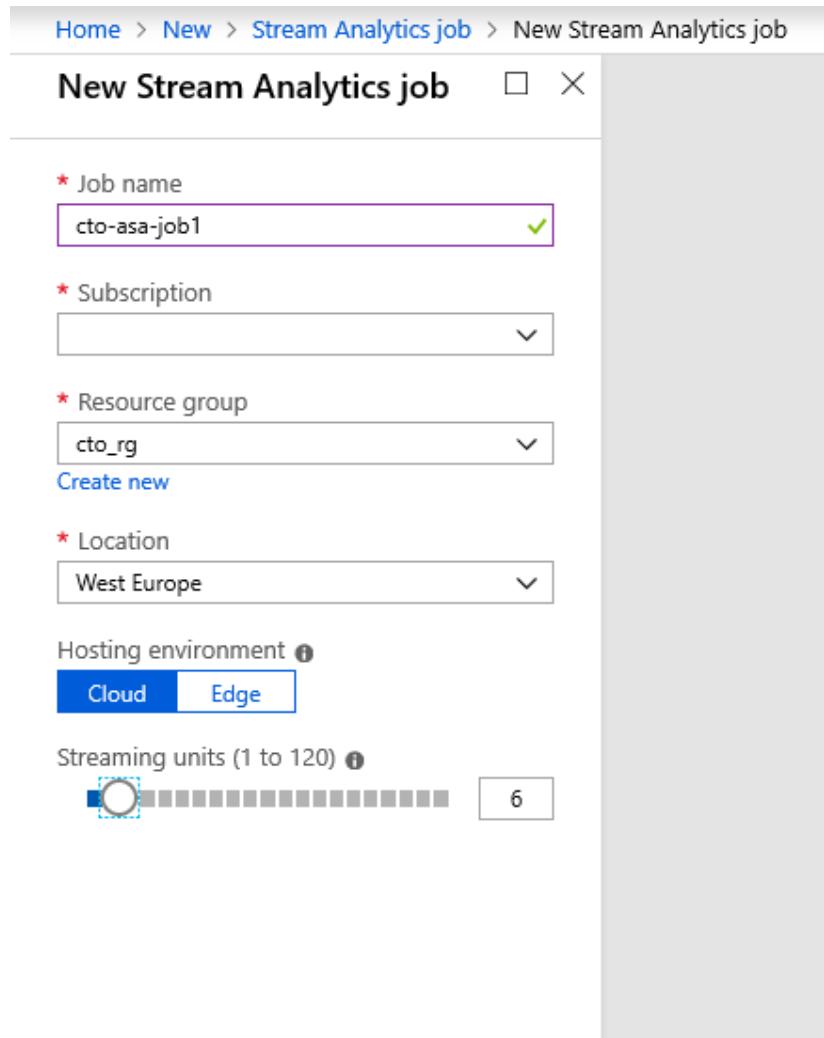
\* Resource group  
cto\_rg

Create new

\* Location  
West Europe

Hosting environment ⓘ  
Cloud Edge

Streaming units (1 to 120) ⓘ  
6



# Create a Stream Analytics Job Input.

**Event Hub**

New input

\* Input alias  
cto-asa-input01 

Provide Event Hub settings manually  
 Select Event Hub from your subscriptions

Subscription  
LearnAI Training Subscription 

\* Event Hub namespace   
cto-eh-ns 

\* Event Hub name   
 Create new  Use existing  
cto-name-eh 

\* Event Hub policy name   
RootManageSharedAccessKey 

Event Hub policy key  
\*\*\*\*\* 

Event Hub consumer group   


\* Event serialization format   
JSON 

Encoding   
UTF-8 

Event compression type   
None 

# Create a Stream Analytics Job Output.

Home > Resource groups > cto\_rg > cto-asa-job1 > Outputs

## Outputs

+ Add

- Event Hub
- SQL Database
- Blob storage
- Table storage
- Service Bus topic
- Service Bus queue
- Cosmos DB
- Power BI
- Data Lake Storage Gen1

SINK

**Blob storage**

\* Output alias  
cto-asa-output01 ✓

Provide Blob storage settings manually  
 Select Blob storage from your subscriptions

Subscription  
LearnAI Training Subscription

\* Storage account ✓  
ctoazureblob

\* Storage account key  
\*\*\*\*\*

\* Container  
 Create new  Use existing

\* Container name  
socialmedia ✓

Path pattern ✓

Date format  
YYYY/MM/DD

Time format  
HH

\* Event serialization format ✓  
JSON

Encoding ✓  
UTF-8

# Write a transformation query

The screenshot shows the Azure Stream Analytics job interface for 'cto-asa-job1'. The left sidebar contains navigation links: Overview (selected), Activity log, Access control (IAM), Tags, Diagnose and solve problems, Settings, Locks, Job topology, Inputs, Functions, Query, and Outputs. The main area displays the job's details: Resource group (change) : cto\_rg, Status : Created, Location : West Europe, Subscription (change) : LearnAI Training Subscription, and Subscription ID : 5be49961-ea44-42ec-8021-b728be90d58c. Below this, the 'Inputs' section shows one input named 'cto-asa-input01' with a red box around it. The 'Outputs' section shows one output named 'cto-asa-output01' with a red box around it. A large red box highlights the 'Query' section on the right, which contains the following T-SQL code:

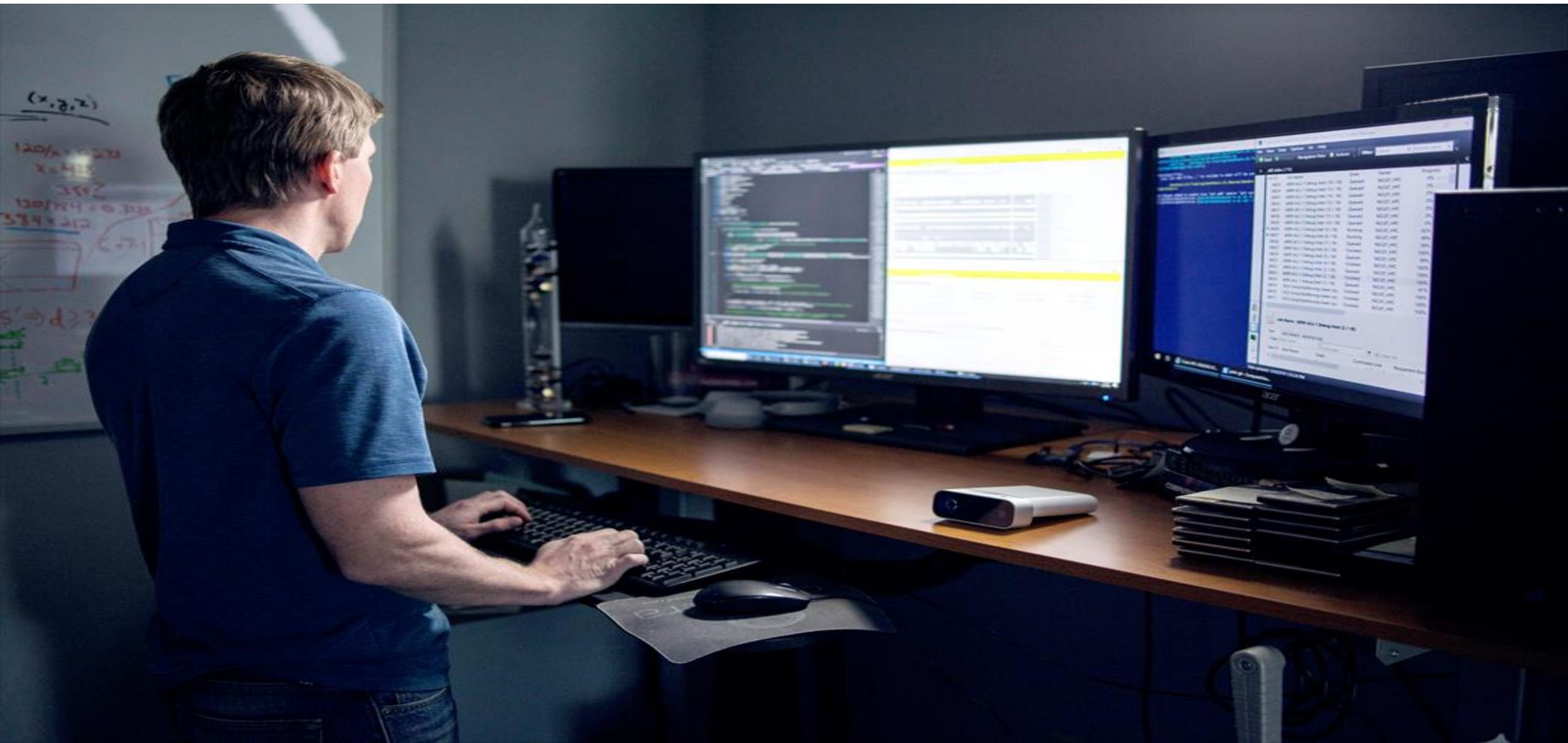
```
1 SELECT *
2   INTO [cto-asa-output01]
3     FROM [cto-asa-input01]
```

# Start a Stream Analytics Job

The screenshot shows the Azure Stream Analytics job configuration page for 'cto-asa-job1'. The job is currently in a 'Created' state. The 'Inputs' section lists one input named 'cto-asa-input01'. The 'Outputs' section lists one output named 'cto-asa-output01'. The 'Query' section displays the following T-SQL code:

```
1 SELECT *  
2 INTO [cto-asa-output01]  
3 FROM [cto-asa-input01]
```

# Lab: Performing Real-Time Analytics with Stream Analytics





# Module 07:

## Orchestrating Data Movement with

## Azure Data Factory



# Agenda

- L01 - Introduction to Azure Data Factory
- L02 - Understand Azure Data Factory components
- L03 - Integrate Azure Data Factory with Databricks



# Lesson 01

## Introduction to

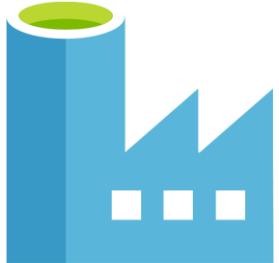
## Azure Data Factory



# Lesson Objectives

- What is Azure Data Factory
- The Data Factory process
- Azure Data Factory components
- Azure Data Factory security

# What is Azure Data Factory



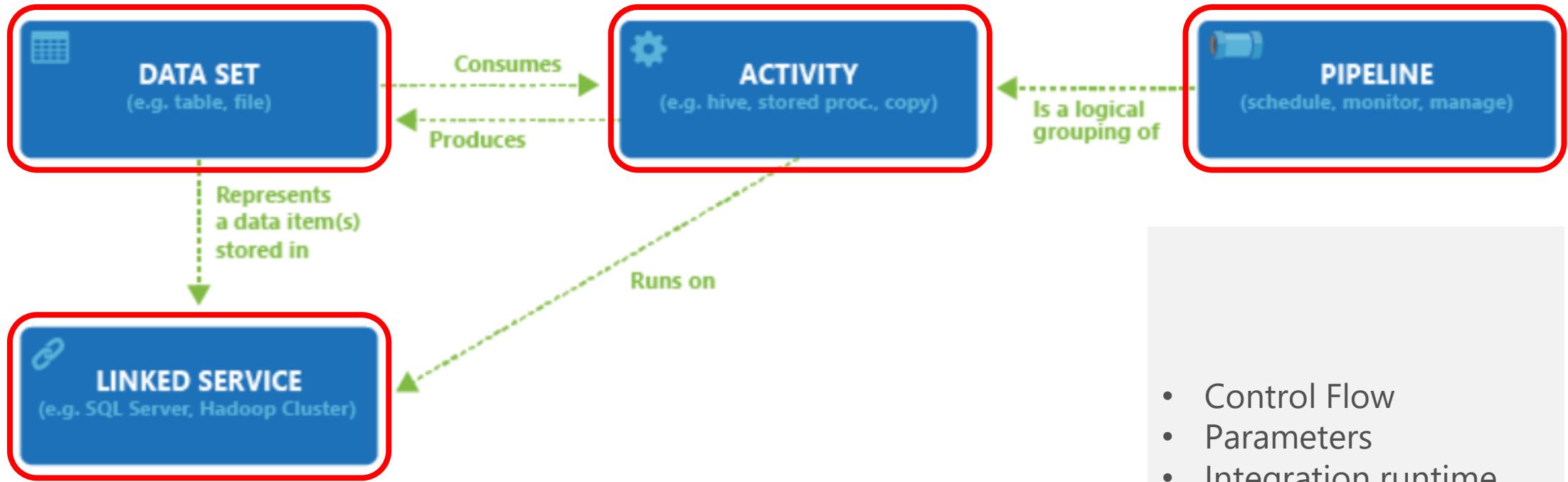
Creates, orchestrates, and automates the movement, transformation and/or analysis of data through in the cloud.



# The Data Factory Process



# Azure Data Factory Components



- Control Flow
- Parameters
- Integration runtime

# Azure Data Factory Security

## **Data Factory Contributor Role**

1. Create, edit, and delete data factories and child resources including datasets, linked services, pipelines, triggers, and integration runtimes.
2. Deploy Resource Manager templates. Resource Manager deployment is the deployment method used by Data Factory in the Azure portal.
3. Manage App Insights alerts for a data factory.
4. At the resource group level or above, lets users deploy Resource Manager template.
5. Create support tickets.



# Lesson 02

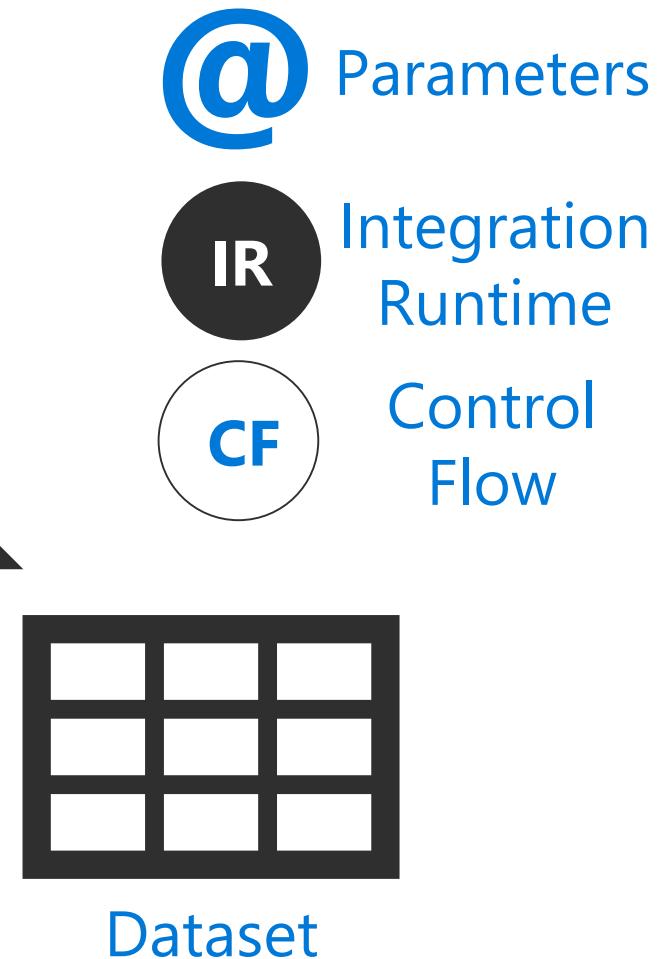
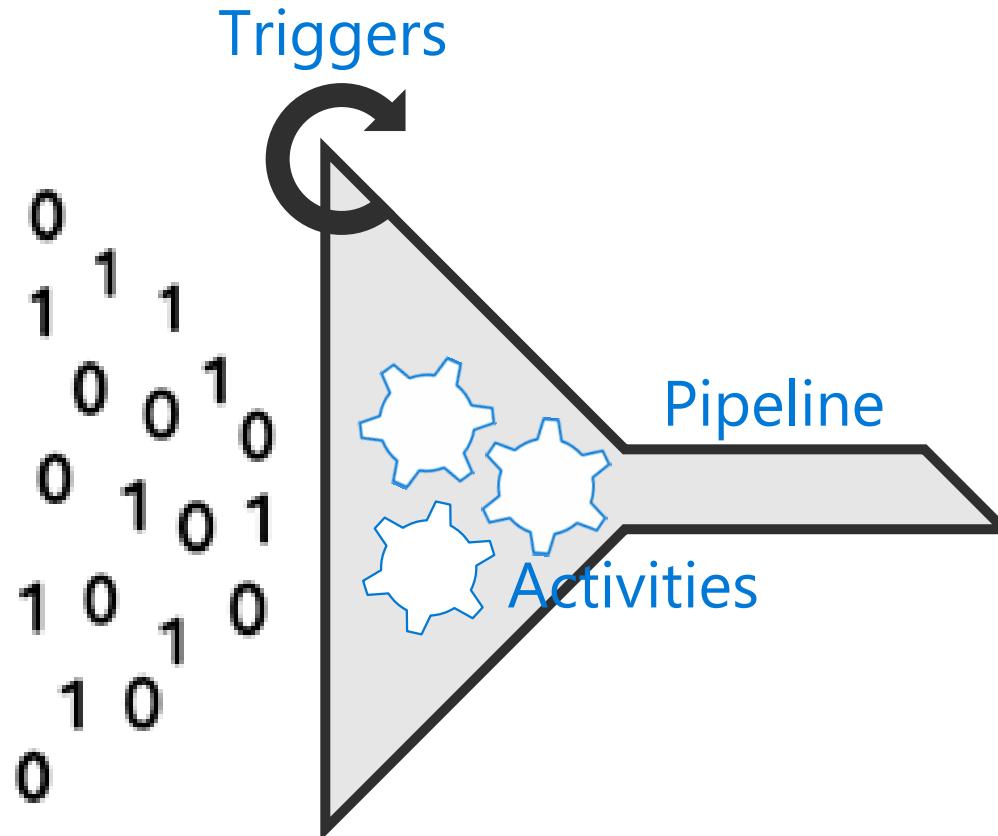
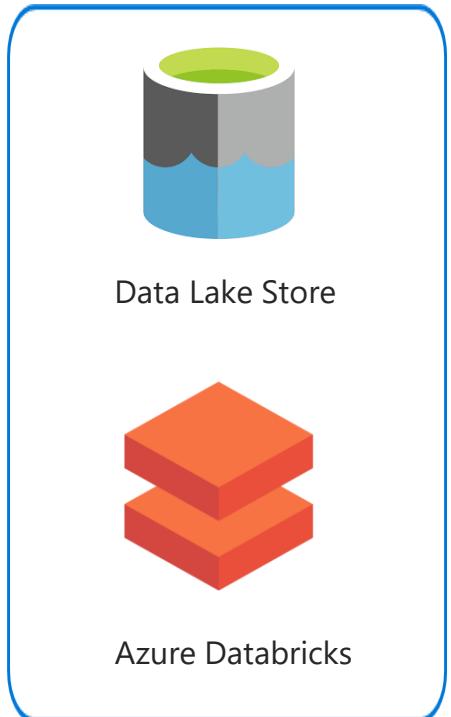
## Azure Data Factory Components

# Lesson Objectives

- Linked Services
- Datasets
- Data Factory activities
- Pipelines
- Pipeline example

# Azure Data Factory components

## Linked Service



# Data Factory Activities

Activities within Azure Data Factory defines the actions that will be performed on the data and there are three categories including:

## Data movement activities

Data movement activities simply move data from one data store to another. A common example of this is in using the Copy Activity.

## Data transformation activities

Data transformation activities use compute resource to change or enhance data through transformation, or it can call a compute resource to perform an analysis of the data.

## Control Activities

Control flow orchestrate pipeline activities that includes chaining activities in a sequence, branching, defining parameters at the pipeline level, and passing arguments while invoking the pipeline on-demand or from a trigger

# Pipelines

**Pipeline** is a grouping of logically related **activities**.

**Pipeline** can be **scheduled** so the activities within it get **executed**.

**Pipeline** can be **managed** and **monitored**.



The collage consists of nine images arranged in a grid. Top row: 1. Students in a classroom setting. 2. A man in a plaid shirt interacting with a large whiteboard. 3. Two men working at a desk with multiple monitors displaying code. 4. Two men in a factory or industrial setting looking at a large screen. Middle row: 5. A woman looking down at her phone. 6. A woman in a green sweater writing on a clipboard. Bottom row: 7. A person wearing headphones playing a video game. 8. A person using a tablet to draw on a wall map. 9. A woman working on a computer with a brain scan graphic overlaid.

# Lesson 03

## Ingesting and Transforming data

# Lesson Objectives

- How to setup Azure Data Factory
- Ingest data using the Copy Activity
- Transforming data with the Mapping Data Flow

# Create Azure Data Factory

## New data factory

Name \*

Version ⓘ

V2

Subscription \*

chtestao

Resource Group \*

Select existing...  
Create new

Location \* ⓘ

South Central US

Enable GIT ⓘ



GIT URL \*

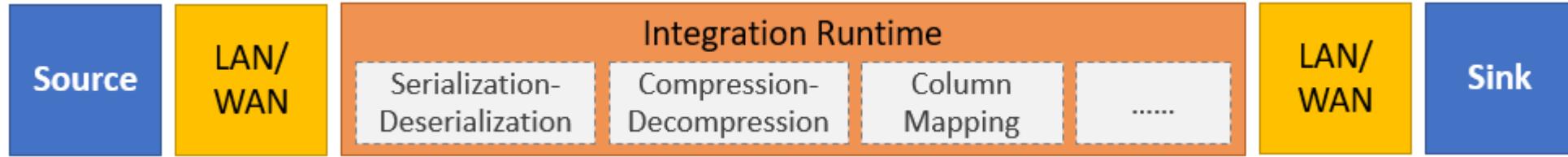
Repo name \*

Branch Name \*

Root folder \*

Create

# Ingesting data with the Copy Activity

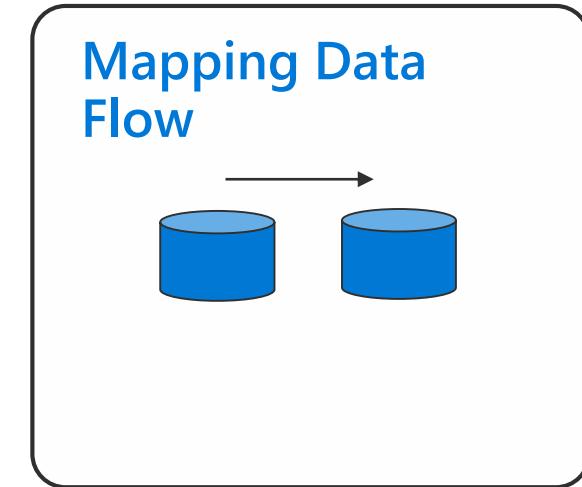


- Reads data from a source data store.
- Performs serialization/deserialization, compression/decompression, column mapping, and so on. It performs these operations based on the configuration of the input dataset, output dataset, and Copy activity.
- Writes data to the sink/destination data store

# Transforming data with the Mapping Data Flow

## Code free data transformation at scale

- Perform data cleansing, transformation, aggregations, etc.
- Enables you to build resilient data flows in a code free environment
- Enable you to focus on building business logic and data transformation
- Underlying infrastructure is provisioned automatically with cloud scale via Spark execution





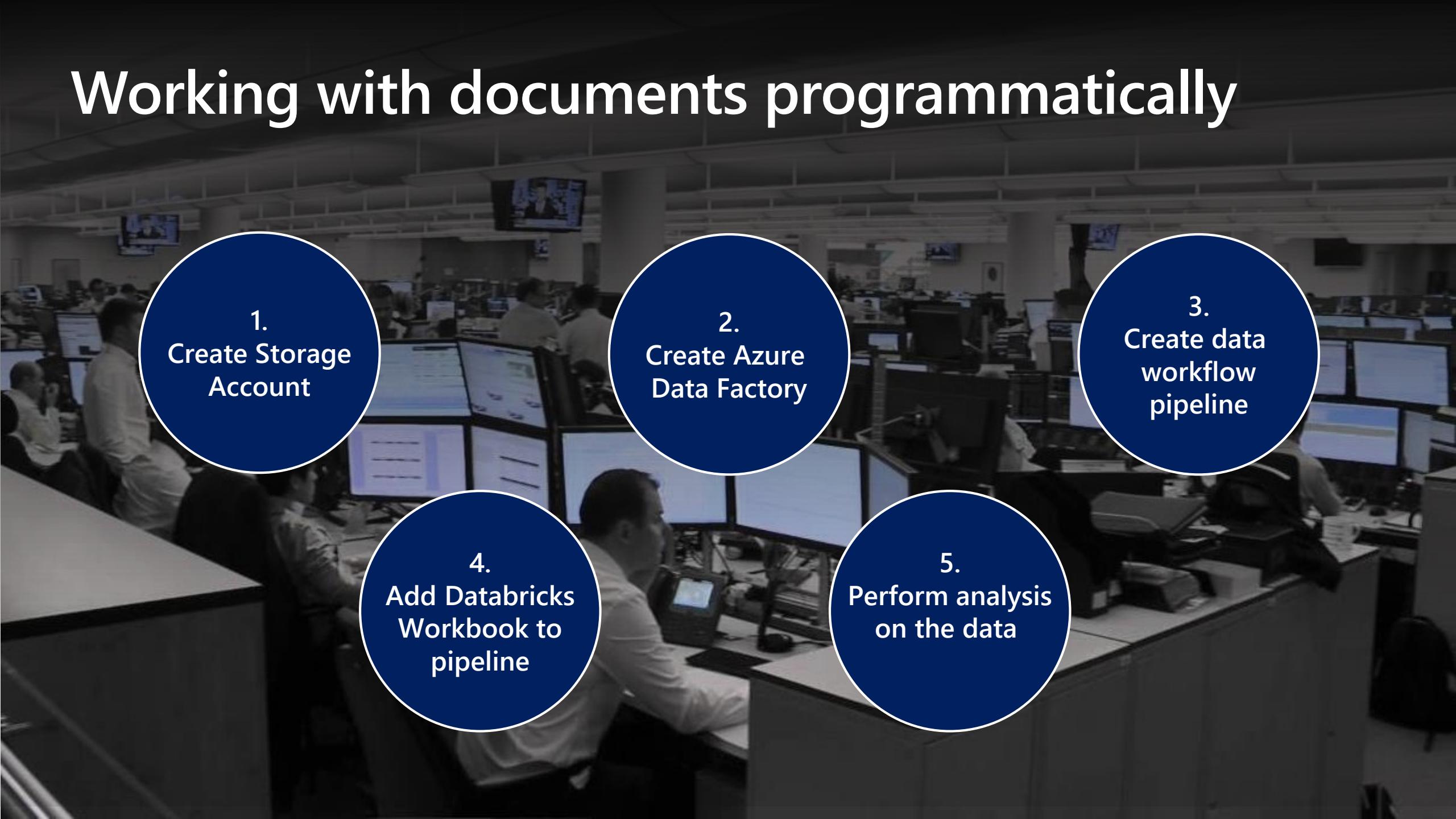
# Lesson 04

## Integrate Azure Data Factory with Azure Databricks

# Lesson Objectives

- Use Azure Data Factory (ADF) to ingest data and create an ADF pipeline.
- Create Azure Storage account and the Azure Data Factory instance
- Use ADF to orchestrate data transformations using a Databricks Notebook activity.

# Working with documents programmatically

A black and white photograph of a control room or monitoring center. Several people are seated at workstations, each equipped with multiple computer monitors displaying various data visualizations and information. The room has a high ceiling with exposed infrastructure and several small screens or cameras mounted on the wall.

1.  
Create Storage  
Account

2.  
Create Azure  
Data Factory

3.  
Create data  
workflow  
pipeline

4.  
Add Databricks  
Workbook to  
pipeline

5.  
Perform analysis  
on the data

# Create Azure Storage account and the Azure Data Factory Instance

Home > New > Storage account > Create storage account

### Create storage account

[Basics](#) [Advanced](#) [Tags](#) [Review + create](#)

Azure Storage is a Microsoft-managed service providing cloud storage that is highly available, secure, durable, scalable, and redundant. Azure Storage includes Azure Blobs (objects), Azure Data Lake Storage Gen2, Azure Files, Azure Queues, and Azure Tables. The cost of your storage account depends on the usage and the options you choose below. [Learn more](#)

**PROJECT DETAILS**

Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources.

\* Subscription:

\* Resource group:  [Create new](#)

**INSTANCE DETAILS**

The default deployment model is Resource Manager, which supports the latest Azure features. You may choose to deploy using the classic deployment model instead. [Choose classic deployment model](#)

\* Storage account name:

\* Location: West Europe

Performance:  Standard  Premium

Account kind: StorageV2 (general purpose v2)

Replication: Read-access geo-redundant storage (RA-GRS)

Access tier (default):  Cool  Hot

[Review + create](#) [Previous](#) [Next : Advanced >](#)

Home > New > Data Factory > New data factory

### New data factory

Name \*:

Version \*: V2

Subscription \*: chtestao

Resource Group \*:  [Select existing...](#) [Create new](#)

Location \*: South Central US

Enable GIT:

GIT URL \*:

Repo name \*:

Branch Name \*:

Root folder \*:

[Create](#)

# Use ADF to orchestrate data transformations using a Databricks Notebook activity

Microsoft Azure

03-Data-Transformation (Python)

Detached File View: Code Permissions Run All Clear Schedule

Cmd 1

## Data Transformation via Azure Data Factory

As you saw at the end of the previous lesson, different cities use different field names and values to indicate crimes, dates, etc. within their crime data.

For example:

- Some cities use the value "HOMICIDE", "CRIMINAL HOMICIDE" or "MURDER".
- In the New York data, the column is named `offenseDescription` while in the Boston data, the column is named `OFFENSE_CODE_GROUP`.
- In the New York data, the date of the event is in the `reportDate`, while in the Boston data, there is a single column named `MONTH`.

In the case of New York and Boston, here are the unique characteristics of each data set:

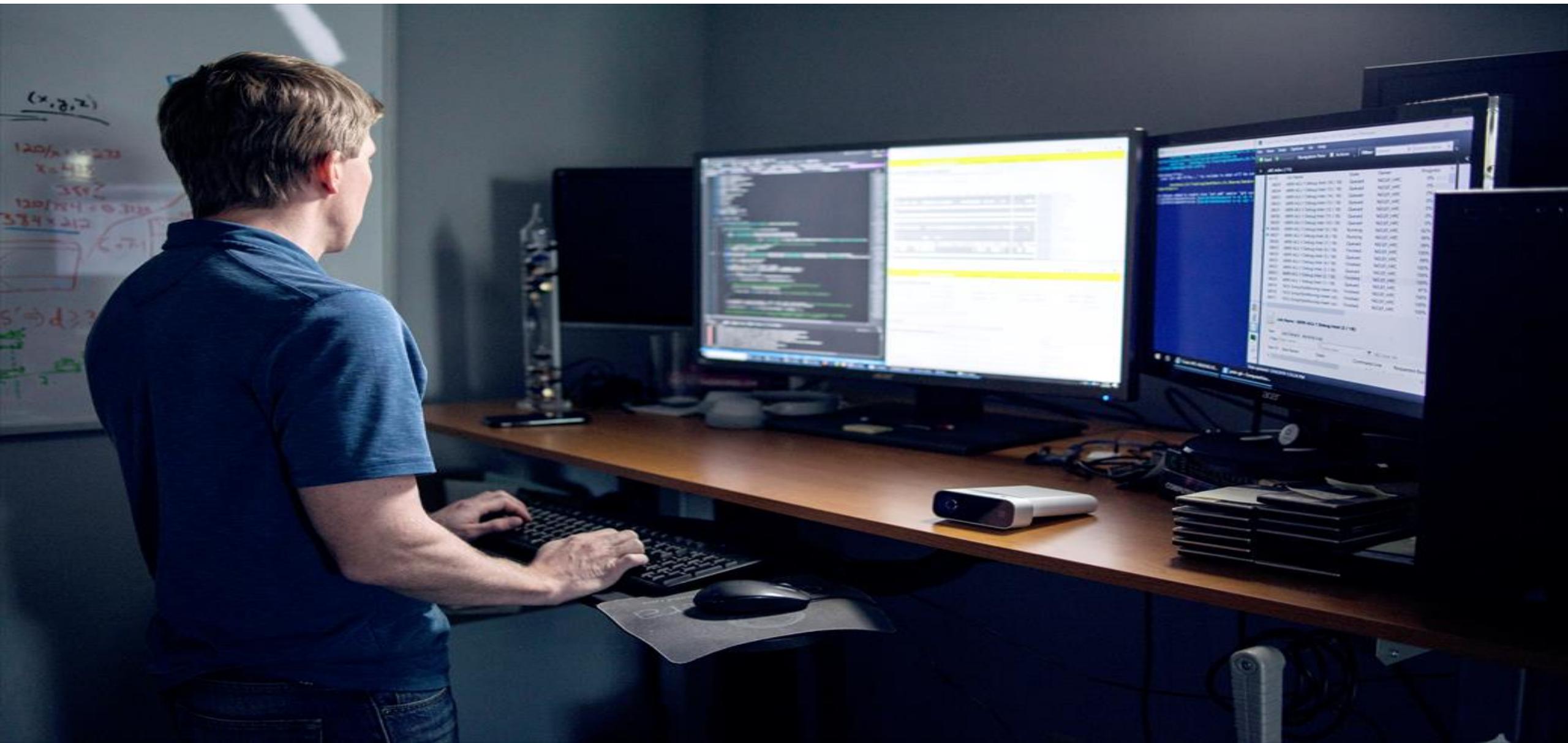
	Offense-Column	Offense-Value	Reported-Column	Reported-Data Type
New York	<code>offenseDescription</code>	starts with "murder" or "homicide"	<code>reportDate</code>	<code>timestamp</code>
Boston	<code>OFFENSE_CODE_GROUP</code>	"Homicide"	<code>MONTH</code>	<code>integer</code>

In this notebook, we will use an ADF Databricks Notebooks activity to perform transformations on and extract homicide statistics from the crime data being processed.

In this lesson you:

1. Create Databricks Access Token.
2. Add Databricks Notebook activity to pipeline.
3. Connect Copy Activities to Notebook Activity.
4. Publish the updated pipeline.
5. Trigger and Monitor the pipeline run.
6. Verify transformations of data by looking at the generated table in Databricks.
7. Perform a simple aggregation of the data.

# Lab: Orchestrating Data Movement with Azure Data Factory





# Module 08: Securing Azure Data Platforms

13.30 ₪.



# Agenda

- L01 - An introduction to security
- L02 - Key security components
- L03- Securing Storage Accounts and Data Lake Storage
- L04 - Securing data stores



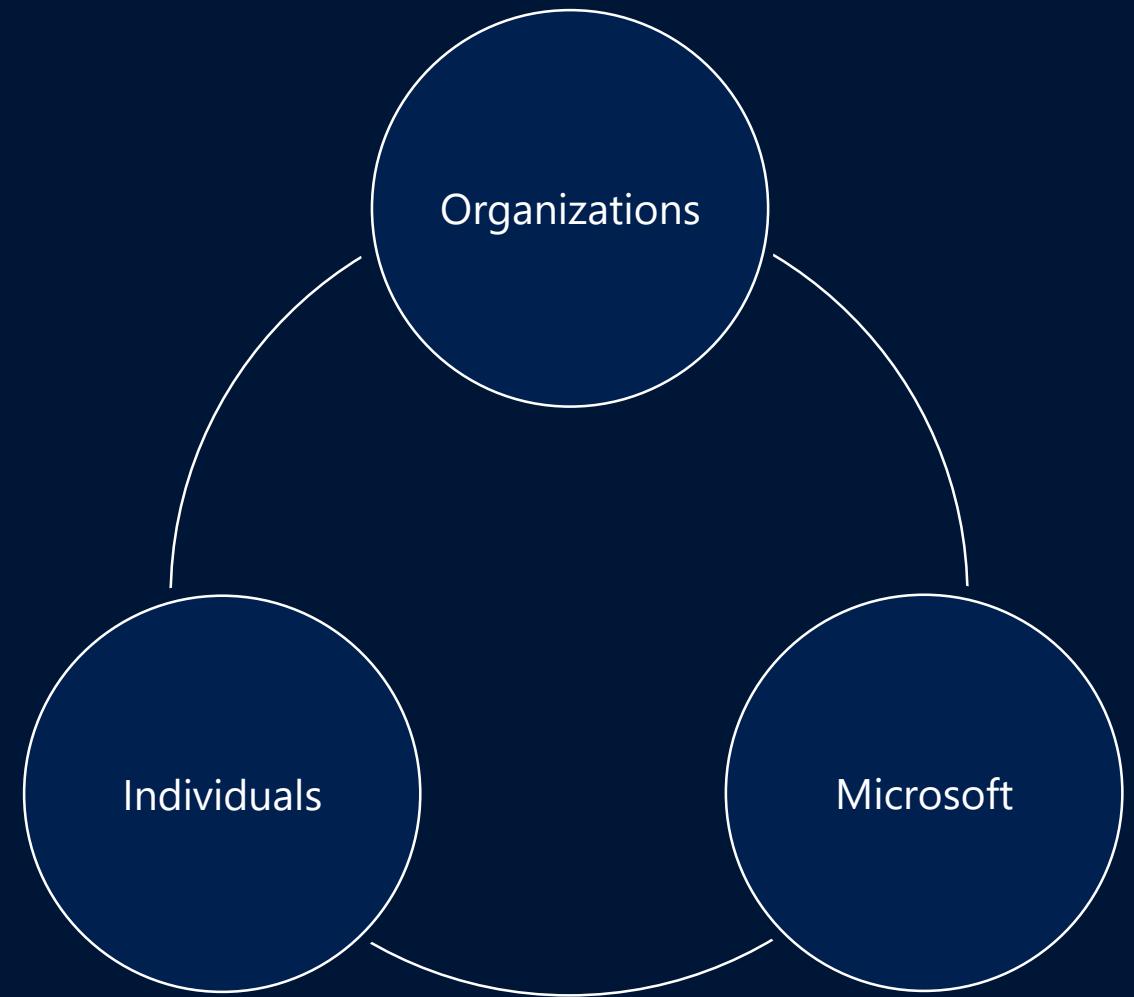
# Lesson 01

## An Introduction to Security

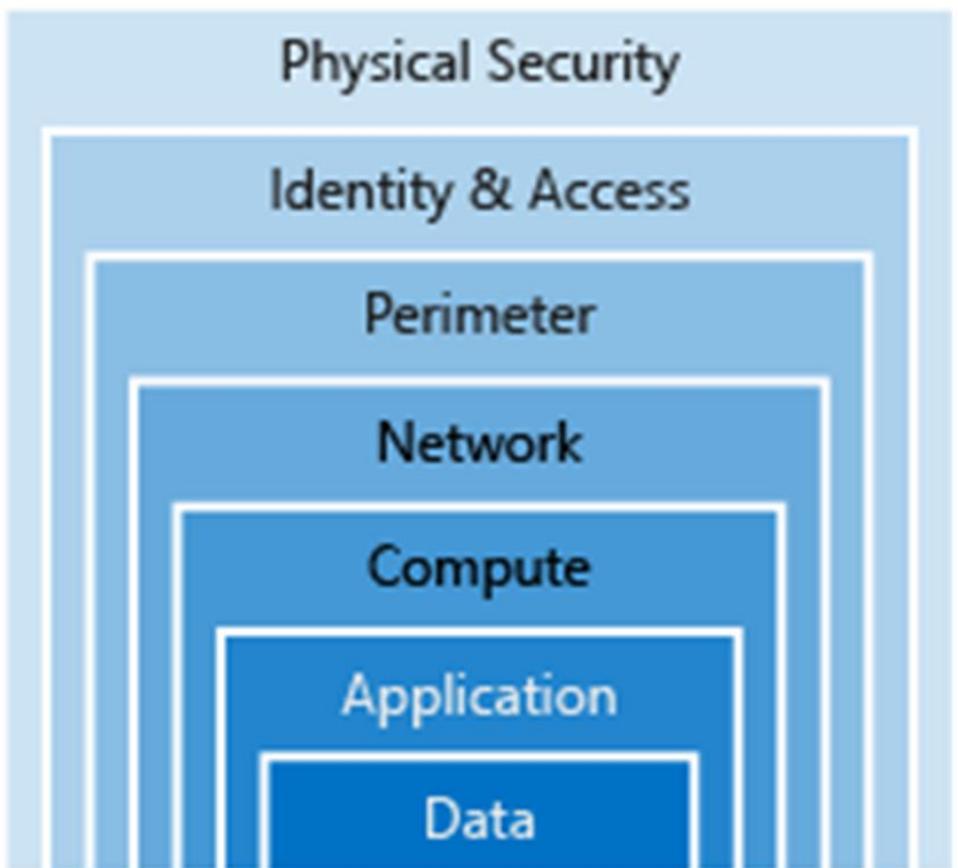
# Lesson Objectives

- Shared Security Responsibility
- A layered approach to security
- The Azure Security Center
- Azure Government

# Shared Security Responsibility



# A Layered Approach to Security.



# Azure Security Center

The screenshot shows the Microsoft Azure homepage with a dark theme. At the top, there's a navigation bar with links for Overview, Solutions, Products (which is highlighted in a black box), Documentation, Pricing, Training, Marketplace, and Partners. Below the navigation, the main heading "Azure Security Center" is displayed in large white text. A sub-headline "Gain unmatched hybrid security management and threat protection" follows. There are two prominent buttons: a blue button labeled "Turn on Security Center >" and a white button labeled "Not yet subscribed to Azure? Start free >". At the bottom of the main content area, there's a horizontal navigation bar with links for Pricing, Documentation, Updates, and Training. Below this, a section titled "Turn on protection you need" contains a paragraph about Microsoft's security measures.

Microsoft Azure

Overview Solutions **Products** Documentation Pricing Training Marketplace Partners

## Azure Security Center

Gain unmatched hybrid security management and threat protection

[Turn on Security Center >](#)

Not yet subscribed to Azure? [Start free >](#)

Pricing > Documentation > Updates > Training >

**Turn on protection you need**

Microsoft uses a wide variety of physical, infrastructure, and additional actions you need to take to help safeguard your security posture and protect against threats.

## Use for incident response

You can use Security Center during the detection, assessment, and diagnosis of security at various stages.

## Use to enhance security.

Reduce the chances of a significant security event by configuring a security policy, and then implementing the recommendations provided by Azure Security Center.

# Azure Government

The background of the slide features a wide-angle, aerial photograph of a dense urban skyline, likely New York City, viewed from a high vantage point. The buildings are mostly skyscrapers, and the sky above them is a clear, pale blue with wispy white clouds.

**Modernize  
Government  
Services**

**Provide a  
platform of  
agility**

**Advanced  
Government  
Mission**

**Physically  
separate from  
Azure**



# Lesson 02

## Key Security Components

# Lesson Objectives

- Network security
- Identity and access management
- Encryption capabilities built into Azure
- Azure Threat Protection

# Network Security

Securing your network from attacks and unauthorized access is an important part of any architecture.

## Internet Protection

Assess the resources that are internet-facing, and to only allow inbound and outbound communication where necessary. Make sure you identify all resources that are allowing inbound network traffic of any type.

## Firewalls

To provide inbound protection at the perimeter, there are several choices:

- Azure Firewall
- Azure Application Gateway
- Azure Storage Firewall

## DDoS Protection

The Azure DDoS Protection service protects your Azure applications by scrubbing traffic at the Azure network edge before it can impact your service's availability.

## Network Security Groups

Network Security Groups allow you to filter network traffic to and from Azure resources in an Azure virtual network. An NSG can contain multiple inbound and outbound security rules.

# Identity and Access

## Authentication

This is the process of establishing the identity of a person or service looking to access a resource. Azure Active Directory is a cloud-based identity service that provides this capability.

## Authorization

This is the process of establishing what level of access an authenticated person or service has. It specifies what data they're allowed to access and what they can do with it. Azure Active Directory also provides this capability.

## Azure Active Directory Features.

### Single Sign-On

Enables users to remember only one ID and one password to access multiple applications.

### Apps & Device Management

You can manage your cloud and on-premises apps and devices and the access to your organization's resources.

### Identity Services

Manage Business-to-business (B2B) identity services and Business-to-Customer (B2C) identity services.

# Encryption

## Encryption at rest

Data at rest is the data that has been stored on a physical medium. This could be data stored on the disk of a server, data stored in a database, or data stored in a storage account.

## Encryption in transit

Data in transit is the data actively moving from one location to another, such as across the internet or through a private network. Secure transfer can be handled by several different layers.

## Encryption on Azure.

### Raw Encryption

Enables the encryption of:

- Azure Storage
- V.M. Disks
- Disk Encryption

### Database Encryption

Enables the encryption of databases using:

- Transparent Data Encryption

### Encrypting Secrets

Azure Key Vault is a centralized cloud service for storing your application secrets.

# Azure Threat Protection

Azure Advanced Threat Protection | contoso-corp | Timeline PREVIEW

4:04 PM Today

Honeytoken activity Updated

The following activities were performed by [Bob Minion](#):

- Logged in to [2 computers](#) via [Contoso-DC](#).
- Authenticated from [2 computers](#) using Kerberos when accessing [5 resources](#) against [Contoso-DC](#).
- Authenticated from [ITARGOET-T4705](#) using NTLM against corporate resources via [Contoso-DC](#).

Started at 3:08 PM Jan 22, 2018

3:23 PM Jan 22, 2018

Remote execution attempt detected

The following remote execution attempts were performed on [Contoso-DC](#) from [ALICE-DESKTOP](#):

- Attempted remote execution of one or more WMI methods by [AdminUser](#).

3:06 PM Jan 22, 2018

Suspicious service creation

[AdminUser](#) created [10 services](#) in order to execute potentially malicious commands on [Contoso-DC](#).

3:03 PM Jan 22, 2018

Brute force attack using LDAP simple bind

200 password guess attempts were made on [2 accounts](#) from [ALICE-DESKTOP](#). 2 account passwords were successfully guessed.

2:59 PM Jan 22, 2018

Reconnaissance using account enumeration

Suspicious account enumeration activity using Kerberos protocol, originating from [ALICE-DESKTOP](#), was detected. The attacker performed a total of 101 guess attempts for account names. 2 guess attempts matched existing account names in Active Directory.

12:38 PM Jan 21, 2018

Malicious replication of directory services

Malicious replication requests were attempted by [Alice Liddle](#), from [ALICE-DESKTOP](#) against [Contoso-DC](#).

11:59 AM Jan 21, 2018

Reconnaissance using DNS

Suspicious DNS activity was observed, originating from [ALICE-DESKTOP](#) (which is not a DNS server) against [Contoso-DC](#).

This screenshot shows the Azure Advanced Threat Protection Timeline interface. It displays a chronological list of security events for the organization 'contoso-corp'. The events are categorized into several types of threats, such as Honeytoken activity, Remote execution attempt, Suspicious service creation, Brute force attack, Reconnaissance using account enumeration, Malicious replication of directory services, and Reconnaissance using DNS. Each event entry includes a timestamp, a brief description, and a link to 'OPEN' for further details. The interface has a clean, modern design with a dark header and light-colored cards for each event. The Microsoft logo is visible in the top right corner.



# Lesson 03

## Securing Storage Accounts and Data Lake Storage

# Lesson Objectives

- Storage Account security features
- Explore the authentication options available to access data
  - Storage Account Key
  - Shared Access Signature
- Control network access to the data
- Managing encryption
- Azure Data Lake Storage Gen II security features

# Storage Account Security Features

A blurred background image of a control room or monitoring center. Several people are seated at desks, each equipped with multiple computer monitors displaying various data. The room is filled with rows of similar workstations, suggesting a high-tech surveillance or operational environment.

Encryption at Rest

Encryption in Transit

Role Based Access Control

Auditing Access

# Storage Account Keys

Home > [Resource groups](#) > [cto\\_rg](#) > [ctoazureblob - Access keys](#)

## ctoazureblob - Access keys

Storage account

 Overview

 Activity log

 Access control (IAM)

 Tags

 Diagnose and solve problems

 Events

 Storage Explorer (preview)

### Settings

 Access keys

 Geo-replication

 CORS

 Configuration

 Encryption

 Shared access signature

Use access keys to authenticate your applications when making requests to this Azure storage account. Store your access keys securely - for example, using Azure K Vault - and don't share them. We recommend regenerating your access keys regularly. You are provided two access keys so that you can maintain connections using one key while regenerating the other.

When you regenerate your access keys, you must update any Azure resources and applications that access this storage account to use the new keys. This action will interrupt access to disks from your virtual machines. [Learn more](#)

Storage account name

ctoazureblob

**key1** 

Key

eU7 [REDACTED] ICg==

Connection string

Def [REDACTED] 9YrQ...

**key2** 

Key

NW0 [REDACTED] V/pUgB5w==

Connection string

Def [REDACTED] Ns6...

# Shared Access Signatures

Home > Resource groups > cto\_rg > ctoazureblob - Shared access signature

**ctoazureblob - Shared access signature**  
Storage account

Search (Ctrl+/  
)

- Overview
- Activity log
- Access control (IAM)
- Tags
- Diagnose and solve problems
- Events
- Storage Explorer (preview)

**Settings**

- Access keys
- Geo-replication
- CORS
- Configuration
- Encryption
- Shared access signature**
- Firewalls and virtual networks
- Advanced Threat Protection
- Static website
- Properties
- Locks
- Export template

Blob service

A shared access signature (SAS) is a URI that grants restricted access rights to Azure Storage resources. You can provide a shared access signature to clients who should not be trusted with your storage account key but whom you wish to delegate access to certain storage account resources. By distributing a shared access signature URI to these clients, you grant them access to a resource for a specified period of time.

An account-level SAS can delegate access to multiple storage services (i.e. blob, file, queue, table). Note that stored access policies are currently not supported for an account-level SAS.

[Learn more](#)

Allowed services ⓘ  
 Blob  File  Queue  Table

Allowed resource types ⓘ  
 Service  Container  Object

Allowed permissions ⓘ  
 Read  Write  Delete  List  Add  Create  Update  Process

Start and expiry date/time ⓘ  
Start  
2019-03-29  11:59:33

End  
2019-03-29  19:59:33   
(UTC+00:00) --- Current Time Zone ---

Allowed IP addresses ⓘ  
for example, 168.1.5.65 or 168.1.5.65-168.1.5.70

Allowed protocols ⓘ  
 HTTPS only  HTTPS and HTTP

Signing key ⓘ  
key1

**Generate SAS and connection string**

# Control network access to data

## Firewalls and virtual networks

Save Discard Refresh

**Info** Firewall settings allowing access to storage services will remain in effect for up to a minute after saving updated settings restricting access.

Allow access from  
 All networks  Selected networks

Configure network security for your storage accounts. [Learn more](#).

Virtual networks  
Secure your storage account with virtual networks. [+ Add existing virtual network](#) [+ Add new virtual network](#)

VIRTUAL NETWORK	SUBNET	ADDRESS RANGE	ENDPOINT STATUS	RESOURCE GROUP	SUBSCRIPTION
No network selected.					

Firewall  
Add IP ranges to allow access from the internet or your on-premises networks. [Learn more](#).

Add your client IP address ('86.184.235.180') [?](#)

**ADDRESS RANGE**

Exceptions

Allow trusted Microsoft services to access this storage account [?](#)

Allow read access to storage logging from any network

Allow read access to storage metrics from any network

# Managing Encryption

Databases stores information that is sensitive, such as physical addresses, email addresses, and phone numbers. The following can be used to protect this data:

## Transport Layer Security (TLS)

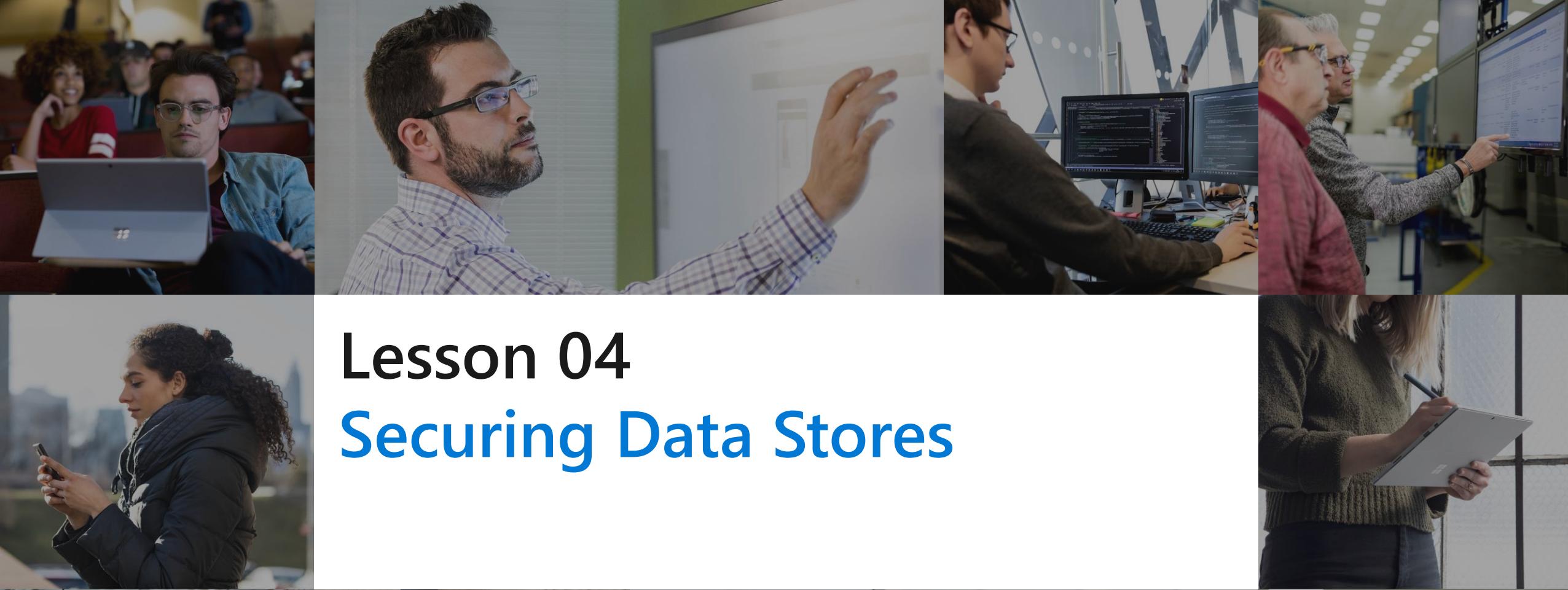
Azure SQL Database and Data Warehouse enforces Transport Layer Security (TLS) encryption at all times for all connections, which ensures all data is encrypted "in transit" between the database and the client.

## Transparent data encryption

Both Azure Data Warehouse and SQL Database protects your data at rest using transparent data encryption (TDE). TDE performs real-time encryption and decryption of the database, associated backups, and transaction log files at rest without requiring changes to the application.

## Application encryption

Data in transit is a method to prevent man-in-the-middle attacks. To encrypt data in transit, specify **Encrypt=true** in the connection string in your client applications



# Lesson 04

## Securing Data Stores

# Lesson Objectives

- Control network access to your data stores using firewall rules
- Control user access to your data stores using authentication and authorization
- Dynamic Data Masking
- Audit and monitor your Azure SQL Database for access violations

# Control network access to your data stores using firewall rules

There are a number of ways you can control access to your Azure SQL Database or Data Warehouse over the network.

## Server-level firewall rules

These rules enable clients to access your **entire Azure SQL server**, that is, all the databases within the same logical server.

## Database level firewall rules

These rules allow access to an individual database on a logical server and are stored in the database itself. For database-level rules, only **IP address rules** can be configured.

# Control user access to your data stores using authentication and authorization

## Authentication

SQL Database and Azure Synapse Analytics supports two types of authentication: SQL authentication and Azure Active Directory authentication.

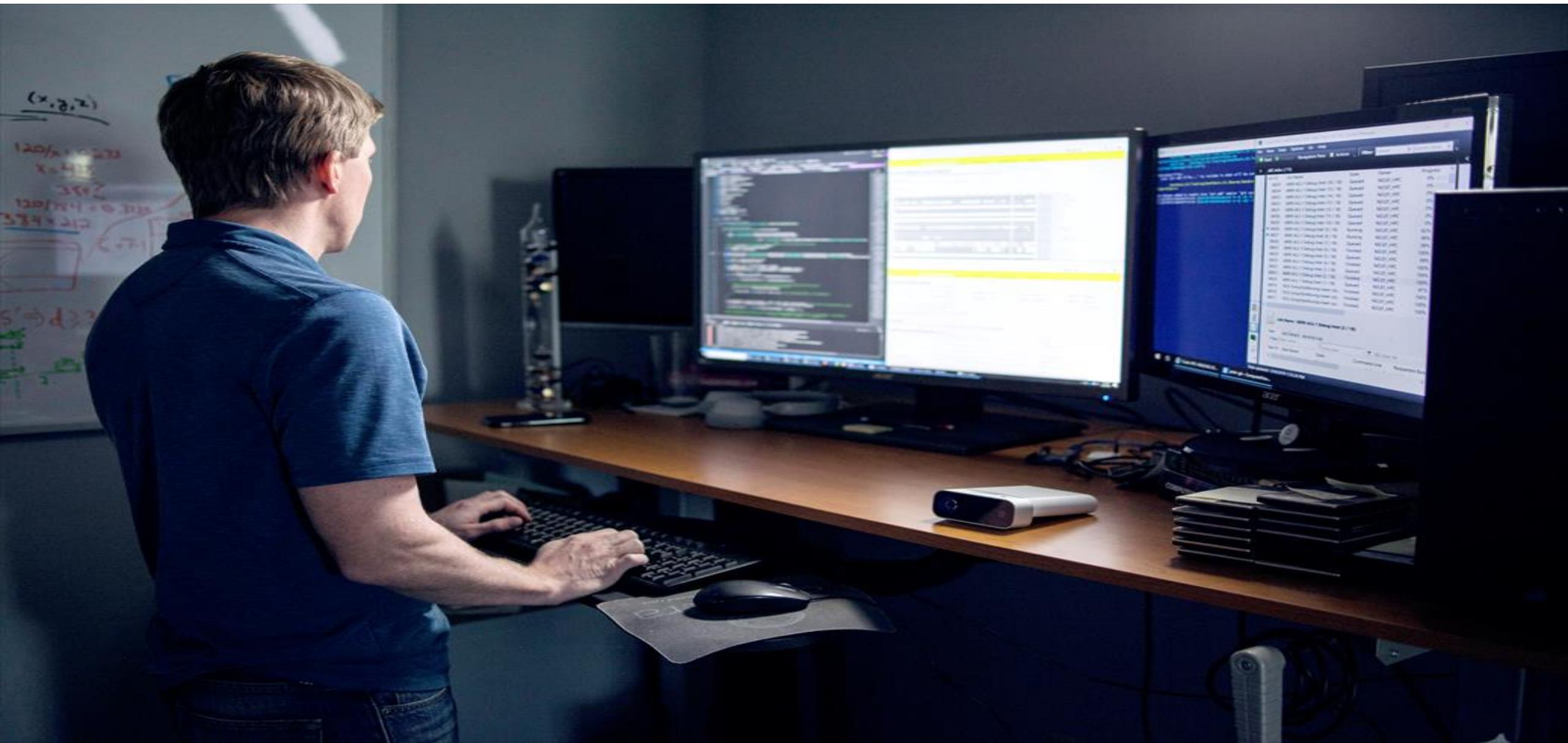
## Authorization

Authorization is controlled by permissions granted directly to the user account and/or database role memberships. A database role is used to group permissions together to ease administration

## Auditing and Monitoring



# Lab: Securing Azure Data Platforms





# Module 09:

## Monitoring and Troubleshooting

## Data Storage and Processing



# Agenda

L01 - General Azure monitoring capabilities

L02 - Troubleshoot common data storage issues

L03 - Troubleshoot common data processing issues

L04 - Manage disaster recovery



# Lesson 01

## General Azure Monitoring Capabilities



# Lesson Objectives

Azure Monitor

Monitoring the network

Diagnose and Solve Problems

# Azure Monitor

Azure Monitor provides a holistic monitoring approach by collecting, analyzing, and acting on telemetry from both cloud and on-premises environments

## Metric Data

Provides quantifiable information about a system over time that enables you to observe the behavior of a system.

## Log Data

Logs can be queried and even analyzed using Azure Monitor logs. In addition, this information is typically presented in the overview page of an Azure Resource in the Azure portal.

## Alerts

Alerts notify you of critical conditions and potentially take corrective automated actions based on triggers from metrics or logs

# Monitoring the network

Azure Monitor logs within Azure monitor has the capability to monitor and measure network activity.

## Network Performance Monitor

Network Performance Monitor measures the performance and reachability of the networks that you have configured.

## Application Gateway Analytics

Application Gateway Analytics contains rich, out-of-the box views you can get insights into key scenarios, including:

- Monitor client and server errors.
- Check requests per hour

# Diagnose and Solve Issues

The screenshot shows the 'ctocdb - Diagnose and solve problems' blade in the Azure portal. The left sidebar lists navigation options: Overview, Activity log, Access control (IAM), Tags, Diagnose and solve problems (selected), Quick start, Notifications, Data Explorer, Settings, Replicate data globally, Default consistency, Firewall and virtual networks, and CORS. The main content area includes a search bar, a 'RESOURCE HEALTH' section showing 'Available' status, a 'RECENT ACTIVITY' section for the past 24 hours, and a 'SOLUTIONS TO COMMON PROBLEMS' section with items like 'My database is slow', 'My request unit (RU) charging is unclear', 'I need more storage/throughput', 'My queries are slow', 'MongoDB API Support', and 'Import MongoDB data into CosmosDB'.

Home > ctocdb - Diagnose and solve problems

## ctocdb - Diagnose and solve problems

Azure Cosmos DB account

Overview

Activity log

Access control (IAM)

Tags

Diagnose and solve problems

Quick start

Notifications

Data Explorer

Settings

Replicate data globally

Default consistency

Firewall and virtual networks

CORS

### RESOURCE HEALTH

Available

There aren't any known problems affecting this Cosmos DB database account [More details](#)

### RECENT ACTIVITY

Activity for the past 24 hours

[Quick Insights](#) | [See all activity](#)

### SOLUTIONS TO COMMON PROBLEMS

- My database is slow
- My request unit (RU) charging is unclear
- I need more storage/throughput
- My queries are slow
- MongoDB API Support
- Import MongoDB data into CosmosDB



The collage consists of nine square images arranged in a grid. Top row: 1. Students in a classroom setting, one using a laptop. 2. A man in a plaid shirt and glasses pointing at a whiteboard. 3. Two men working at a desk with multiple computer monitors displaying code. 4. Two men in a factory or industrial setting looking at a large screen. 5. A woman in a green sweater writing on a clipboard. Bottom row: 1. A woman with curly hair looking at her phone. 2. A man wearing headphones and playing a video game on a laptop. 3. A person using a tablet to draw on a wall covered in technical blueprints. 4. A woman in profile working on a computer with a brain icon overlay. 5. A man in a blue shirt holding a tablet in a factory setting.

# Lesson 02

## Troubleshoot Common Data Storage Issues



# Lesson Objectives

Connectivity issues

Performance issues

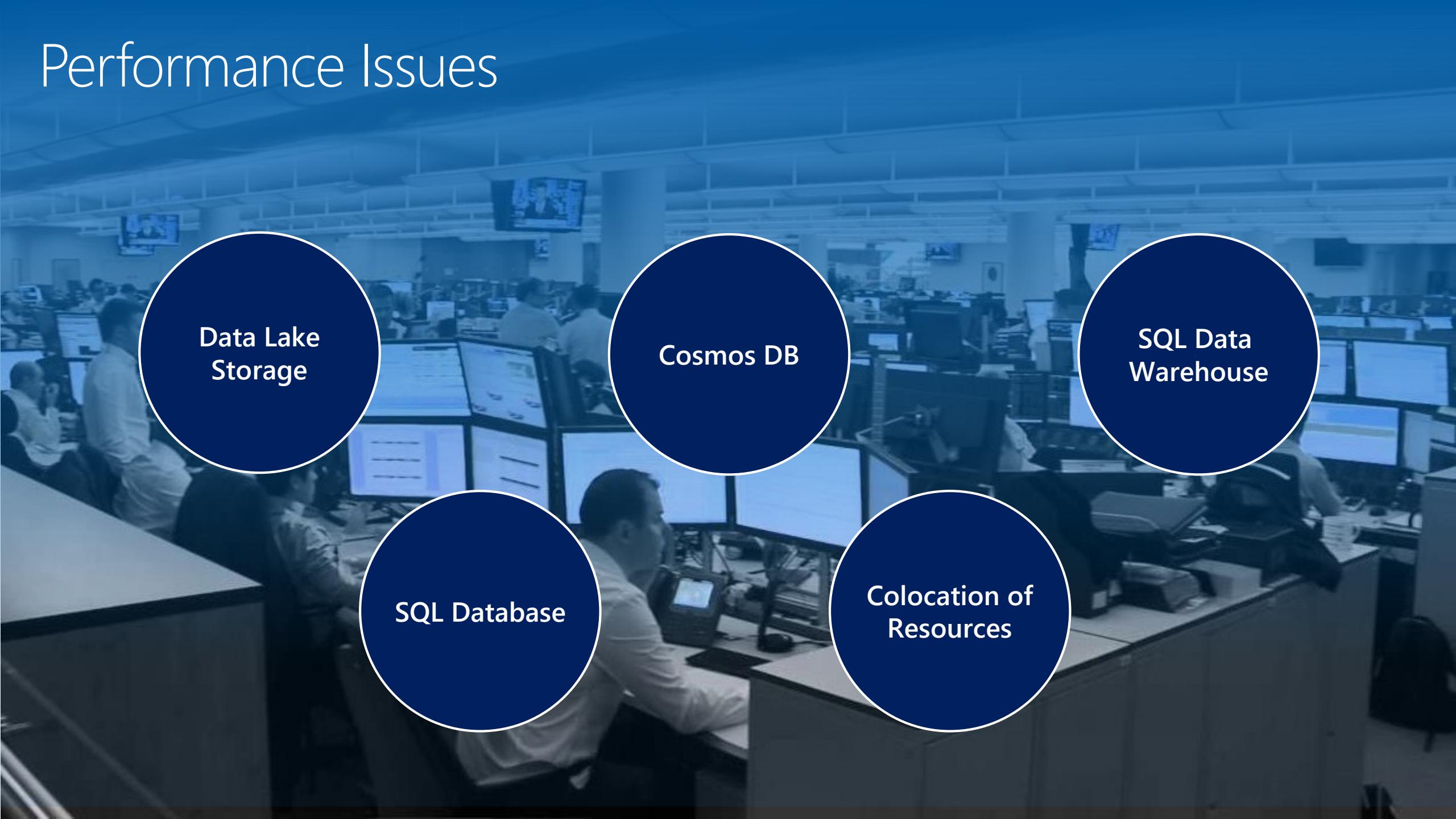
Storage issues

# Connectivity Issues

There are a range of issues that can impact connectivity issues, including:

Unable to connect to the data platform	Authentication Failures	Cosmos DB Mongo DB API errors	SQL Database Failover
<ul style="list-style-type: none"><li>The first area that you should check is the firewall configuration.</li><li>Test the connection by accessing it from a location external to your network.</li><li>Check maintenance schedules</li></ul>	<ul style="list-style-type: none"><li>The first check is to ensure that the user name and password is correct.</li><li>Check the storage account keys and ensure that they match in the connection string.</li></ul>	<ul style="list-style-type: none"><li>Mongo client drivers establishes more than one connection.</li><li>On the server side, connections which are idle for more than 30 minutes are automatically closed down.</li><li>Check for timeouts</li></ul>	Should you receive an "unable to connect" message (error code 40613) in the Azure SQL Database, this scenario commonly occurs when a database has been moved because of deployment, failover, or load balancing.

# Performance Issues

The background of the slide shows a blurred image of a control room or monitoring center. Several people are visible at workstations, each equipped with multiple computer monitors displaying various data. The room has a high ceiling with exposed infrastructure and multiple rows of workstations.

Data Lake Storage

Cosmos DB

SQL Data Warehouse

SQL Database

Colocation of Resources



# Lesson 03

## Troubleshoot Common Data Processing Issues



# Lesson Objectives

Troubleshoot streaming data

Troubleshoot batch data loads

Troubleshoot Azure Data Factory

# Troubleshoot streaming data

When using Stream Analytics, a Job encapsulates the Stream Analytic work and is made up of three components:

## Job input

The job input contains a **Test Connection** button to validate that there is connectivity with the input. However, most errors associated with a job input is due to the malformed input data that is being ingested.

## Job query

A common issue associated with Stream Analytics query is the fact that the output produced is not expected. In this scenario it is best to check the query itself to ensure that there is no mistakes on the code there.

## Job output

As with the job input, there is a \*\*Test Connection\*\* button to validate that there is connectivity with the output, should there be no data appearing. You can also use the \*\*Monitor\*\* tab in Stream Analytics to troubleshoot issues.

# Troubleshoot batch data loads

When trying to resolve data load issues, it is first pragmatic to make the holistic checks on Azure, as well as the network checks and diagnose and solve issue check. After that, then check:

## Azure Blob and Data Lake Store

Notwithstanding network errors; occasionally, you can get timeout or throttling errors that can be a symptom of the availability of the storage accounts.

## SQL Data Warehouse

- Make sure you are always leveraging PolyBase.
- Ensure CTAS statements are used to load data
- Break data down into multiple text files.
- Consider DWU usage

## Cosmos DB

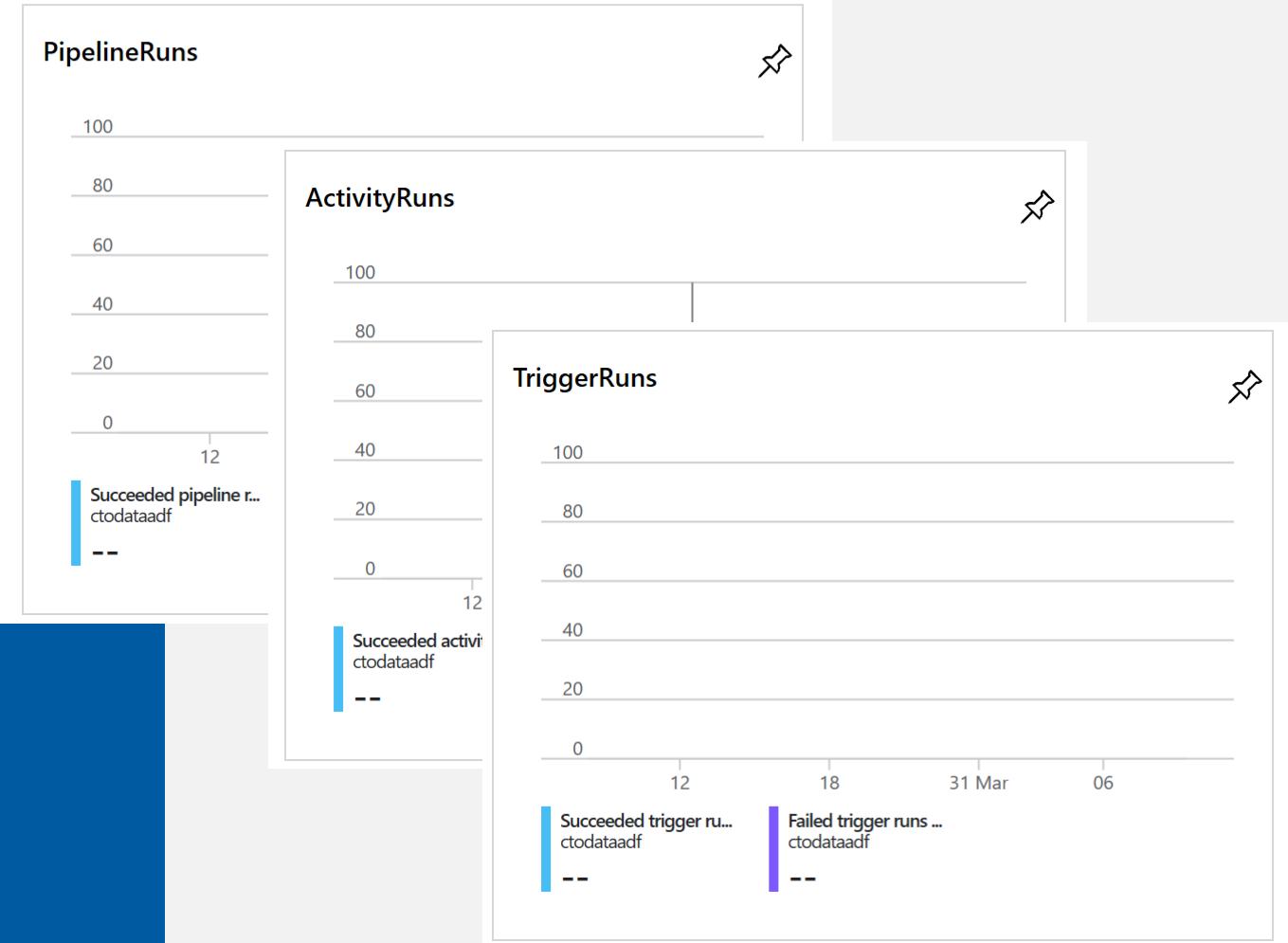
- Check that you have provisioned enough RU's
- Review partitions and partitioning keys
- Check for client connection string settings

## SQL Database

- Check that you have provisioned enough DTU's
- Review whether the database would benefit from elastic pools
- A wide range of tools can be used to troubleshoot SQL Database

# Troubleshoot Azure Data Factory

## Monitoring





# Lesson 04

## Managing Disaster Recovery

# Lesson Objectives

Data redundancy

Disaster recovery

# Data redundancy

Data redundancy is the process of storing data in multiple locations to ensure that it is highly available.

## Azure Blob and Data Lake Store

- Locally redundant storage (LRS)
- Zone-redundant storage (ZRS)
- Geo-redundant storage (GRS)
- Read-access geo-redundant storage (RA-GRS)

## SQL Data Warehouse

SQL Data Warehouse performs a **geo-backup** once per day to a paired data center. The RPO for a geo-restore is 24 hours.

## Cosmos DB

Azure Cosmos DB is a globally distributed database service. You can configure your databases to be globally distributed and available in any of the Azure regions.

## SQL Database

- Check that you have provisioned enough DTU's
- Review whether the database would benefit from elastic pools
- A wide range of tools can be used to troubleshoot SQL Database

# Disaster Recovery

There should be processes that are involved in backing up or providing failover for databases in an Azure data platform technology. Depending on circumstances, there are numerous approaches that can be adopted.

Azure Blob and Data Lake Store	SQL Data Warehouse	Cosmos DB	SQL Database
<p>Supports account failover for geo-redundant storage accounts.</p> <p>You can initiate the failover process for your storage account if the primary endpoint becomes unavailable.</p>	<p>SQL Data Warehouse performs a <b>geo-backup</b> once per day to a paired data center.</p> <p>Data warehouse snapshot feature that enables you to create a restore point to create a copy of the warehouse to a previous state.</p>	<p>Takes a backup of your database every <b>4 hours</b> and at any point of time</p> <p>Only the latest 2 backups are stored.</p>	<p>Creates database backups that are kept between 7 and 35 days</p> <p>Uses Azure read-access geo-redundant storage (RA-GRS) to ensure that they preserved even if data center is unavailable.</p>

# Lab: Monitoring and Troubleshooting Data Storage and Processing

