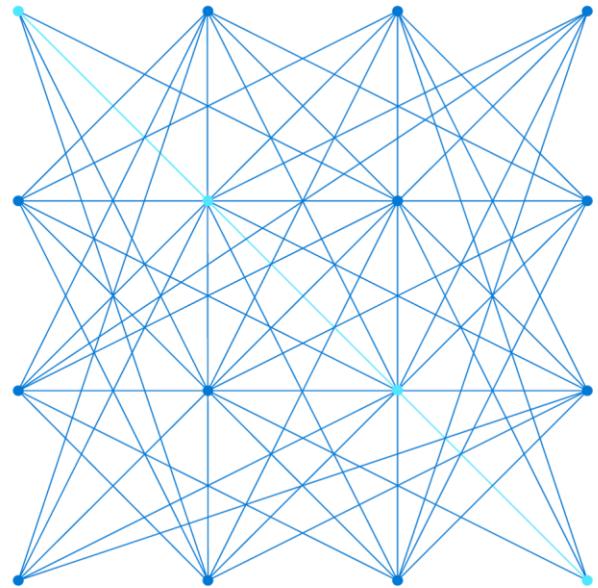




# Implementing an Azure Data Solution [DP-200]



1

## 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 Azure Fundamentals  
Microsoft Certified Azure Data Fundamentals  
Microsoft Certified Azure AI Fundamentals  
Microsoft Certified Solutions Associate (MCSA) - Web Application Development  
Microsoft Certified Trainer (MCT)

2

# Agenda



About this course



Audience



Course agenda



Prerequisites

3

## 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

4

## Course agenda

### Module 1

Azure for the Data Engineer

**Lesson 01** – Explain the evolving world of data

**Lesson 02** – Survey the services in the Azure Data Platform

**Lesson 03** – Identify the tasks that are performed by a Data Engineer

**Lesson 04** – Describe the use cases for the cloud in a case study

### Module 2

Working with Data Storage

**Lesson 01** – Choose a data storage approach in Azure

**Lesson 02** – Create an Azure Storage Account

**Lesson 03** – Explain Azure Data Lake Storage

**Lesson 04** – Upload data into Azure Data Lake

5

## Course agenda (*continued #1*)

### Module 3

Enabling team based Data Science with Azure Databricks

**Lesson 01** – Explain Azure Databricks

**Lesson 02** – Work with Azure Databricks

**Lesson 03** – Read data with Azure Databricks

**Lesson 04** – Perform transformations with Azure Databricks

### Module 4

Building globally distributed databases with Cosmos DB

**Lesson 01** – Create an Azure Cosmos DB database built to scale

**Lesson 02** – Insert and query data in your Azure Cosmos DB database

**Lesson 03** – Build a .NET Core app for Azure Cosmos DB in Visual Studio Code

**Lesson 04** – Distribute your data globally with Azure Cosmos DB

6

## Course agenda (*continued #2*)

### Module 5

Working with relational data stores in the cloud

**Lesson 01** – Explain SQL Database

**Lesson 02** – Explain SQL Data Warehouse

**Lesson 03** – Provision and load data in Azure SQL Data Warehouse

**Lesson 04** – Import data into Azure SQL Data Warehouse using PolyBase

### Module 6

Performing real-time analytics with Stream Analytics

**Lesson 01** – Explain data streams and event processing

**Lesson 02** – Data ingestion with Event Hubs

**Lesson 03** – Processing data with Stream Analytics jobs

7

## Course agenda (*continued #3*)

### Module 7

Orchestrating data movement with Azure Data Factory

**Lesson 01** – Explain how Azure Data Factory works

**Lesson 02** – Create linked services and datasets

**Lesson 03** – Create pipelines and activities

**Lesson 04** – Azure Data Factory pipeline execution and triggers

### Module 8

Securing Azure Data Platforms

**Lesson 01** – Introduction to security

**Lesson 02** – Key security components

**Lesson 03** – Securing storage accounts and Data Lake Storage

**Lesson 04** – Security data stores

**Lesson 05** – Securing streaming data

8

## Course agenda (*continued #4*)

### Module 9

Monitoring and troubleshooting Data Storage and processing

Lesson 01 – Explain the monitoring capabilities that are available

---

Lesson 02 – Troubleshoot common data storage issues

---

Lesson 03 – Troubleshoot common data processing issues

---

Lesson 04 – Manage disaster recovery

9

## 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

10



## Prerequisites

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

[Azure fundamentals](#)

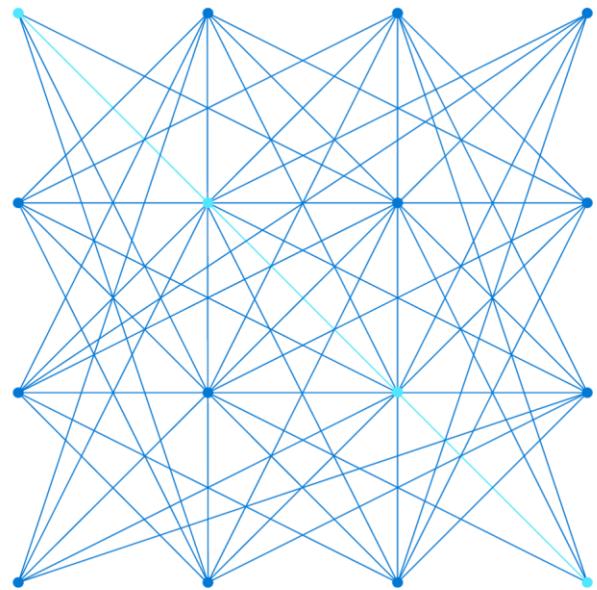
11

The slide features a dark blue background. In the top-left corner, the Microsoft Azure logo is displayed, consisting of four colored squares (blue, orange, yellow, green) followed by the word "Azure". At the bottom-left, there is a small line of fine print: "© Copyright Microsoft Corporation. All rights reserved." The rest of the slide is blank.

12



# Module 01: Azure for the Data Engineer



13

## Agenda



Lesson 01 – Explain the evolving world of data



Lesson 02 – Survey the services in the Azure Data Platform



Lesson 03 – Identify the tasks that are performed by a Data Engineer



Lesson 04 – Describe the use cases for the cloud in a case study

14

## Lesson 01: The evolving world of data



15

### 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

16

## Data abundance

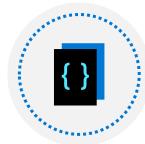
<b>Processes</b>	Businesses are tasked to store, interpret, manage, transform, process, aggregate and report on data
<b>Consumers</b>	There are a wider range of consumers using different types of devices to consume or generate data
<b>Variety</b>	There's a wider variety of data types that need to be processed and stored
<b>Responsibilities</b>	A data engineers role is responsible for more data types and technologies
<b>Technologies</b>	Microsoft Azure provides a wide set of tools and technologies

17

## On-premises versus cloud technologies



Computing Environment



Licensing Model



Maintainability



Scalability



Availability

18

## Data engineering job responsibilities



19

## Use cases for the cloud

Here are some examples of industries making use of the cloud

<b>Web retail</b>	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
<b>Healthcare</b>	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
<b>IoT scenarios</b>	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

20

## Lesson 02: Survey the services in the Azure Data Platform



21

### 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

22

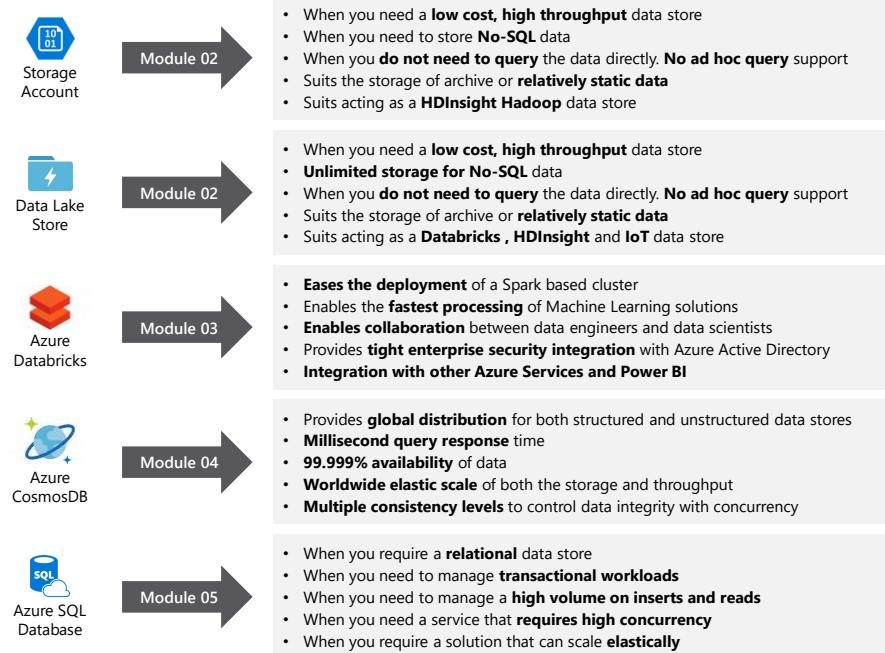
## Structured versus Unstructured data

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
Structured data is data that adheres to a schema, so all of the data has the same fields or properties. Structured data can be stored in a database table with rows and columns	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	Unstructured data encompasses data that has no designated structure to it. Known as No-SQL, there are four types of No-SQL databases: <ul style="list-style-type: none"> <li>• Key Value Store</li> <li>• Document Database</li> <li>• Graph Databases</li> <li>• Column Base</li> </ul>

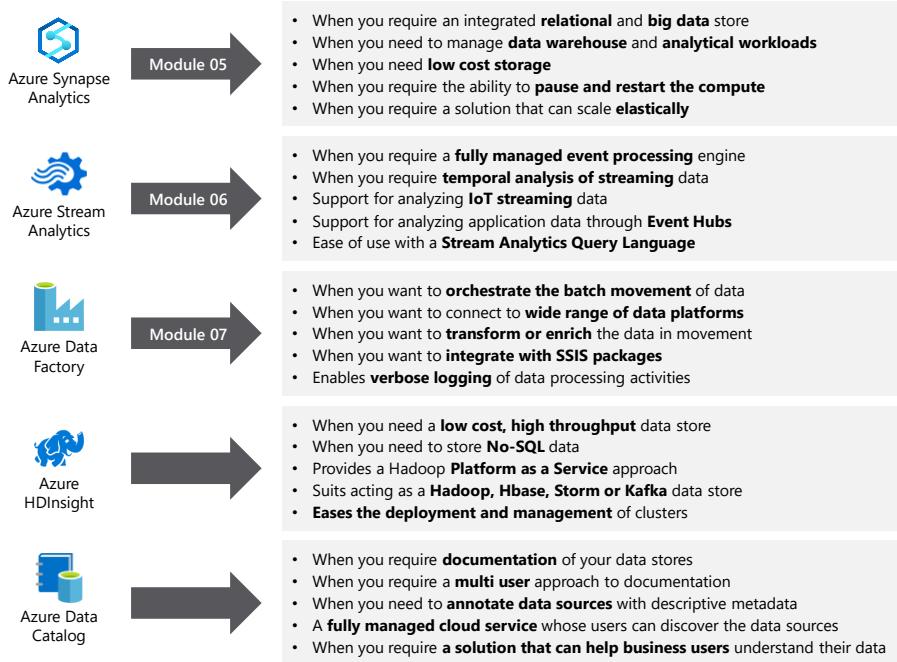
23

## What to use for Data



24

## What to use for Data



25

## Lesson 03: Identify the tasks performed by a Data Engineer



26

## Lesson objectives



List the new roles of modern data projects



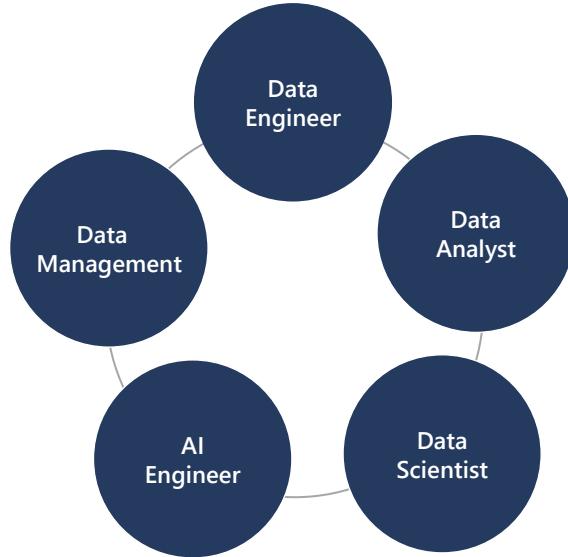
Outline data engineering practices



Explore the high-level process for architecting a data engineering project

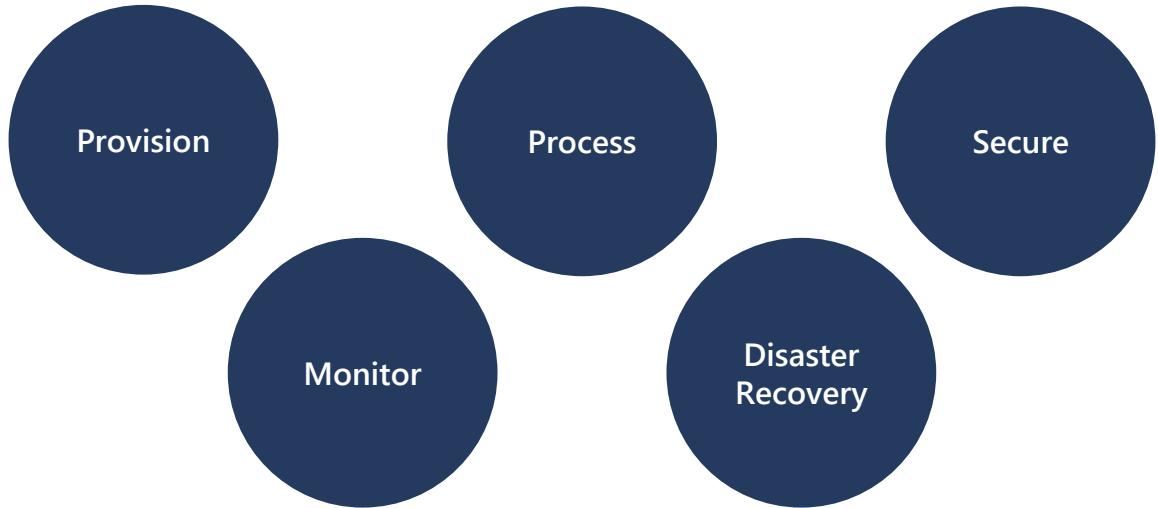
27

## Roles in data projects



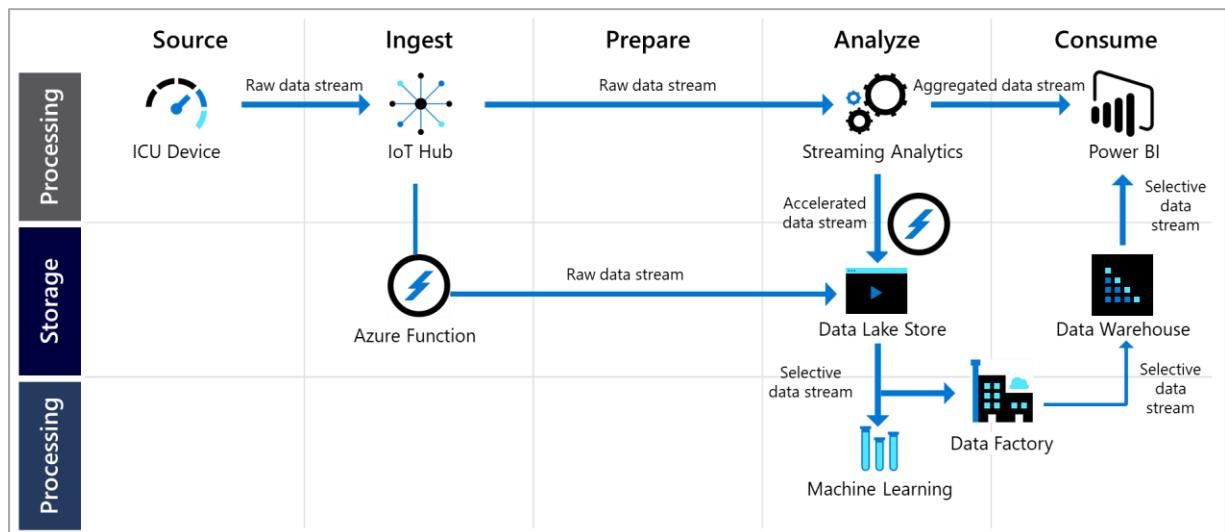
28

## Data engineering practices



29

## Architecting projects – An example

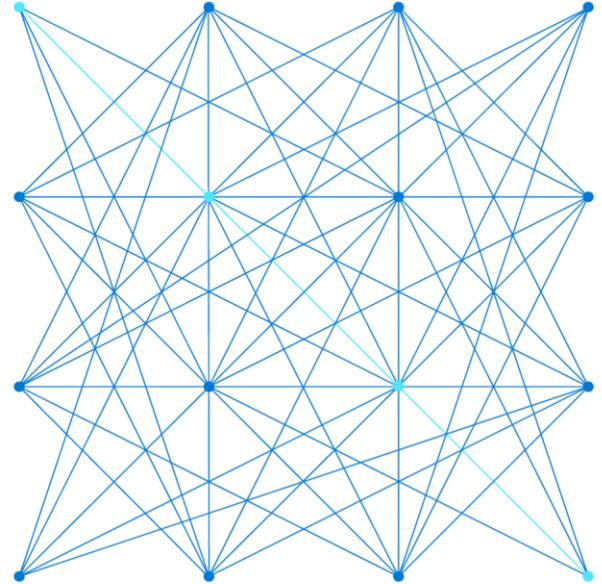


30

© Copyright Microsoft Corporation. All rights reserved.

31

## Module 02: Working with data storage



32

## Agenda



Lesson 01: Choose a data storage approach in Azure

---



Lesson 02: Create an Azure Storage Account

---



Lesson 03: Explain Azure Data Lake Storage

---



Lesson 04: Upload data into Azure Data Lake Store

33

**Lesson 01: Choose a data storage approach in Azure**

34

## Lesson objectives



The Benefits of using Azure to store data



Compare Azure data storage with on-premises storage

35

## Benefits of using Azure to store data



Automated backup



Support for data analytics



Global replication



Storage tiers



Encryption capabilities



Virtual disks



Multiple data types

36

## 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 back-ends quickly

37

## Lesson 02: Create Azure storage account



38

## 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

39

## 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.

40

## Storage account settings

Home > New > Storage account > Create storage account

**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 [View](#)

**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

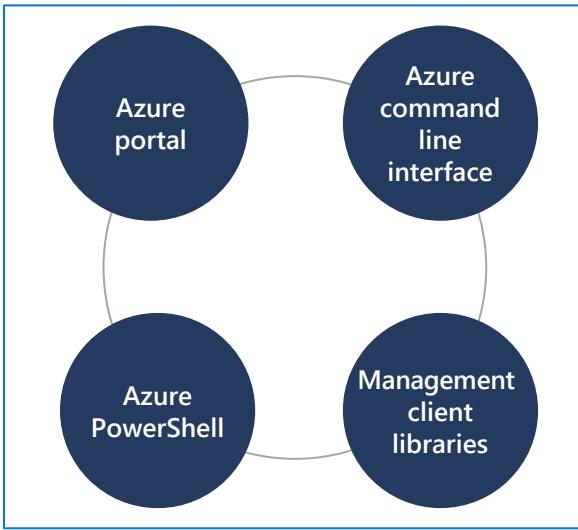
Account kind StorageV2 (general purpose v2)

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

Access tier (default) cool Hot

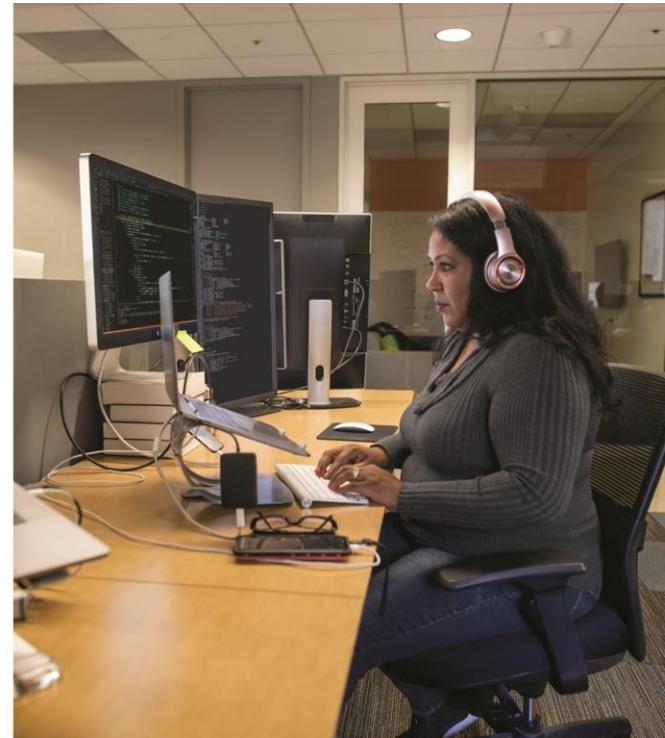
41

## Storage account creation tool



42

## Create a storage account



43

## Lesson 03: Azure Data Lake Storage



44

## 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



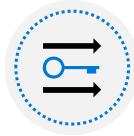
Explore the stages for processing Big Data using Azure Data Lake Store



Describe the use cases for Data lake Storage

45

## Azure Data Lake Storage – Generation II



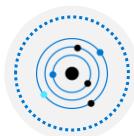
Hadoop access



Security



Performance



Redundancy

46

## 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

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

**Data protection**

Blob soft delete  Disabled  Enabled

Data protection and hierarchical namespace cannot be enabled simultaneously.

**Data Lake Storage Gen2**

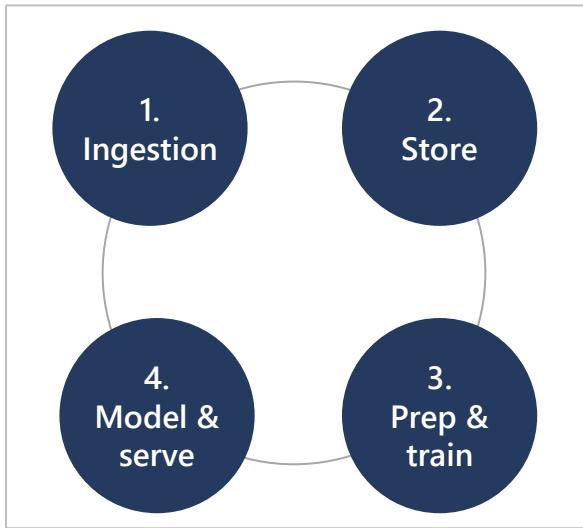
Hierarchical namespace  Disabled  Enabled

NFS v3  Disabled  Enabled

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

47

## Processing Big Data with Azure Data Lake Store



48

## 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

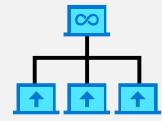
In this solution, Azure Data factory is transferring terabytes of web logs from a web server to the Data Lake on an hourly basis. This data is provided as features to the predictive model in Azure Databricks, which is then trained and scored. The result of the model is then distributed globally using 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

49

## Lesson 04: Upload data into Azure Data Lake Store



50

## 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

51

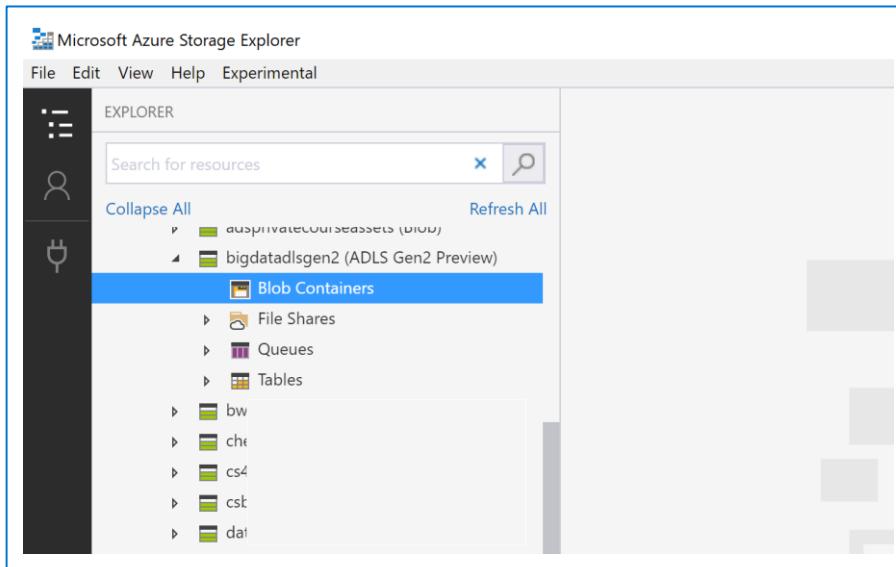
## Create a Azure Data Lake Store (Gen II) using PowerShell

```
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
```

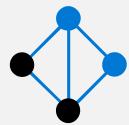
52

## Uploading data with Azure Storage Explorer



53

## Lab: Working with data storage



54

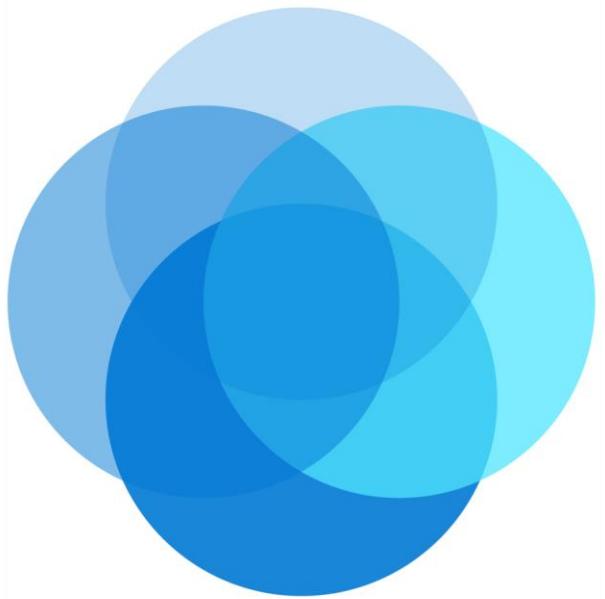


© Copyright Microsoft Corporation. All rights reserved.

55



## Module 03: Enabling team based Data Science with Azure Databricks



56

## Agenda



Lesson 01 – Describe Azure Databricks

---



Lesson 02 – Provision Azure Databricks and Workspaces

---



Lesson 03 – Read data using Azure Databricks

---



Lesson 04 – Perform transformations with Azure Databricks

57

## Lesson 01: Describe Azure Databricks



58

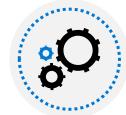
## Lesson objectives



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

59

## What is Azure Databricks



### Apache Spark-based analytics platform:

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



### Enterprise Security:

Utilizes the security capabilities of Azure



### Integration with other Cloud Services:

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

60

## 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?

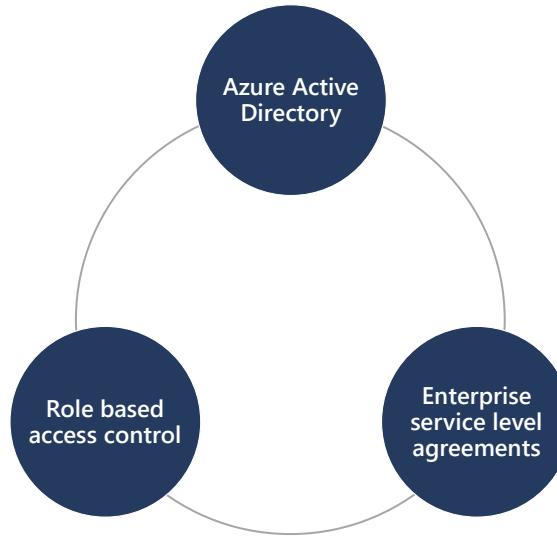
Azure Databricks is a wrapper around Apache Spark that simplifies the provisioning and configuration of a Spark cluster in a GUI interface

### Azure Databricks components:

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

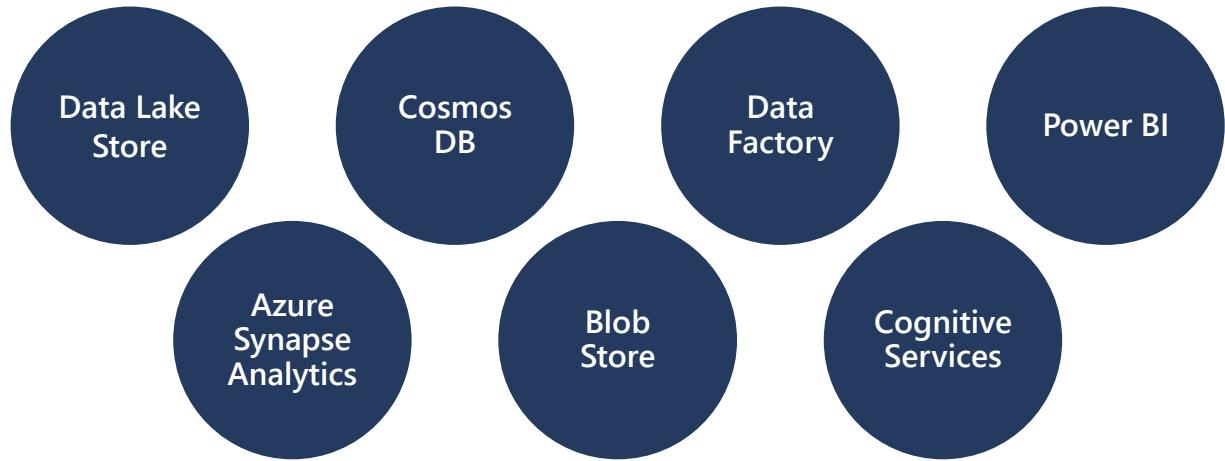
61

## Enterprise security



62

## Integration with cloud services



63

## Lesson 02: Provision Azure Databricks and Workspaces



64



## Lesson objectives

65

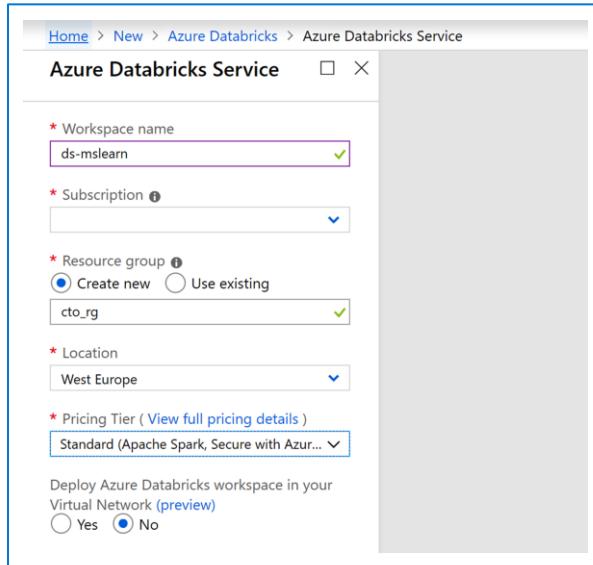


Create your own Azure Databricks workspace



Create a cluster and notebook in Azure Databricks

## Create an Azure Databricks Workspace



The screenshot shows the "Azure Databricks Service" creation page. The URL in the browser is [Home > New > Azure Databricks > Azure Databricks Service](#). The form fields are as follows:

- \* Workspace name: ds-mslearn
- \* Subscription: A dropdown menu showing a list of subscriptions.
- \* Resource group: Create new (radio button selected) cto\_rg
- \* Location: West Europe
- \* Pricing Tier: Standard (Apache Spark, Secure with Azur...)
- Deploy Azure Databricks workspace in your Virtual Network (preview): Yes (radio button selected)

66

## Create a Cluster and Notebook in Azure Databricks

The screenshot shows the Azure Databricks portal interface. At the top left is the Microsoft Azure logo and the 'Azure Databricks' icon. The top right features 'PORTAL', '@microsoft.com', a help icon, and a user profile icon. The main header says 'Azure Databricks'. Below the header are three main sections: 'Explore the Quickstart Tutorial' (with an icon of a document and a lightbulb), 'Import & Explore Data' (with an icon of a dashed box and a cloud), and 'Create a Blank Notebook' (with an icon of a document and a plus sign). A central dashed box contains the text 'Drop files or click to browse'. Below these sections are three tabs: 'Common Tasks', 'Recents', and 'Documentation'. The 'Common Tasks' tab is active, showing links for 'New Notebook', 'Upload Data', 'Create Table', 'New Cluster', 'New Job', 'New Mlflow Experiment' (marked as 'New'), 'Import Library', and 'Read Documentation'. The 'Recents' tab shows a placeholder 'Recent files appear here as you work.' The 'Documentation' tab lists 'Databricks Guide', 'Python, R, Scala, SQL', and 'Importing Data'.

67

## Lesson 03: Read data using Azure Databricks



68

## Lesson objectives



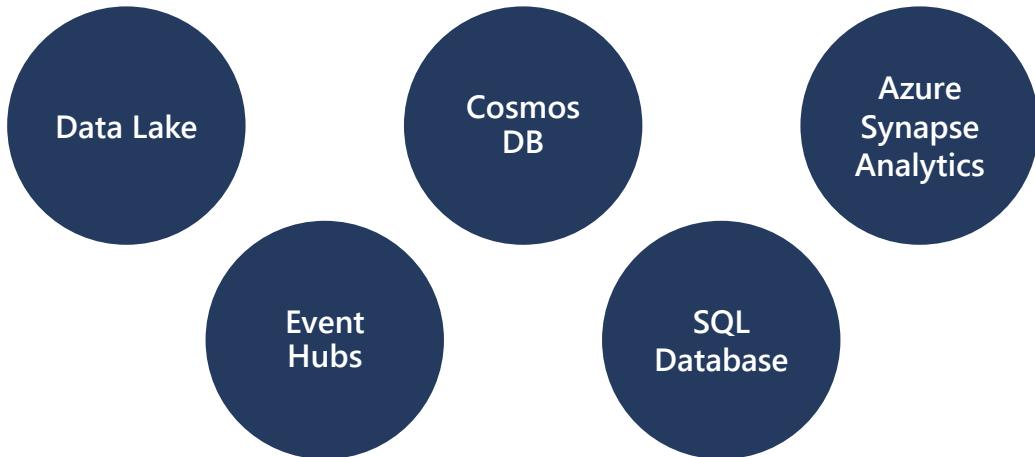
Use Azure Databricks to access data sources



Reading data in Azure Databricks

69

## Use Azure Databricks to access data sources



70

## 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)

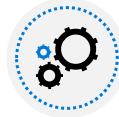
71

## Lesson 04: Perform transformations with Azure Databricks



72

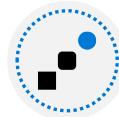
## Lesson objectives



Performing ETL to populate a data model



Perform basic transformations



Perform advanced transformations with user-defined functions

73

## 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
<b>Connect to many data stores:</b> Postgres SQL Server Cassandra Cosmos DB CSV, Parquet Many more..	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

74

## Basic transformation

-  Normalizing values
-  Missing/null data
-  De-duplication
-  Pivoting data frames

75

## 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

76

## Lab: Enabling team-based Data Science with Azure Databricks



77

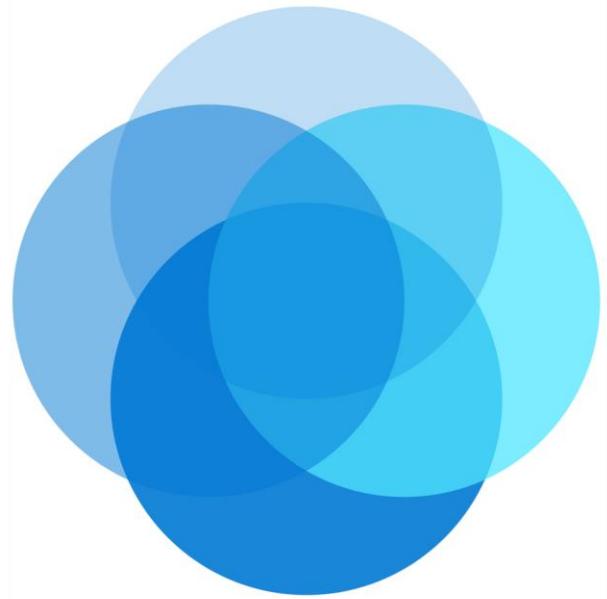
 Microsoft Azure

© Copyright Microsoft Corporation. All rights reserved.

78



# Module 04: Building globally distributed databases with Cosmos DB



79

## Agenda



Lesson 01: Create an Azure Cosmos DB database built to scale

---



Lesson 02: Insert and query data in your Azure Cosmos DB database

---



Lesson 03: Build a .NET Core app for Azure Cosmos DB in Visual Studio Code

---



Lesson 04: Distribute your data globally with Azure Cosmos DB

80

## Lesson 01: Create an Azure Cosmos DB database built to scale



81

### 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

82

## What is Azure Cosmos DB



Scalability



Performance



Availability



Programming model

83

## 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 \*  Resource Group \*  [Create new](#)

**Instance Details**

Account Name \*

API \*  Core (SQL)  Notebooks Notebooks with Apache Spark. None [Sign up for Apache Spark preview](#)

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

Location \*

Geo-Redundancy  Enable  Disable

Multi-region Writes  Enable  Disable

Up to 15% off multi-region writes is available to qualifying new discounts only. Accounts must be created between December 1, 2019 and February 29, 2020. Offer is limited to accounts with both account location 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.

84

## 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

85

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	$(500 * 10) + (500 * 48) = 29,000 \text{ RU/s}$

86

## 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

87

## Creating a Database and a Container in Cosmos DB

The screenshot shows the 'Add Container' dialog box. At the top, there's a note: 'Start at \$24/mo per database, multiple containers included' with a 'More details' link. The 'Database id' section has a radio button for 'Create new' (selected) and 'Use existing'. A text input field is provided for 'Type a new database id'. Below this is a checked checkbox for 'Provision database throughput'. Under 'Throughput (400 - 100,000 RU/s)', there are two options: 'Autopilot (preview)' (radio button) and 'Manual' (radio button selected), with a throughput value of '400' entered. A note below states 'Estimated spend (USD): \$0.032 hourly / \$0.77 daily (1 region, 400RU/s, \$0.00008/RU)'. The 'Container id' section has a text input field with 'e.g., Container1'. The 'Partition key' section has a text input field with 'e.g., /address/zipCode'. A checkbox 'My partition key is larger than 100 bytes' is unchecked. At the bottom, there's a 'Unique keys' section with a '+ Add unique key' button.

88

## Lesson 02: Insert and Query Data in your Azure Cosmos DB Database



89

### 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

90

## Create a product catalog documents in the Data Explorer

The screenshot shows the Azure Cosmos DB Data Explorer interface for the 'awcdbstudcto' account. The left sidebar includes links for Overview, Activity log, Access control (IAM), Tags, Diagnose and solve problems, Quick start, Notifications, and Data Explorer (which is selected). Under Settings, there are options for Replicate data globally, Default consistency, and Firewall and virtual networks. The main area displays a SQL API query: 'SELECT \* FROM c'. Below the query, a table titled 'Items' shows three rows of data with columns 'id' and '...'. To the right of the table is a JSON representation of the data:

```

1  [
2    {
3      "id": "1",
4      "productId": "33218896",
5      "category": "Women's Clothing",
6      "manufacturer": "Contoso Sport",
7      "description": "Quick dry crew neck t-shirt",
8      "price": "14.99",
9      "shipping": {
10        "weight": 1,
11        "dimensions": {
12          "width": 6,
13          "height": 8,
14          "depth": 1
15        }
16      },
17      "_rid": "P01tAMk-JP8BAAAAAAA=",
18      "_self": "dbs/P01tAA==/colls/P01tAMk-JP8=/docs"
19      "_etag": "\u22124100b36e-0000-0d00-0000-5d38cafe0"
20    }
21  ]
  
```

91

## Perform Azure Cosmos DB Queries

<h3>SELECT Query Basics</h3> <pre>SELECT &lt;select_list&gt; [FROM &lt;optional_from_specification&gt;] [WHERE &lt;optional_filter_condition&gt;] [ORDER BY &lt;optional_sort_specification&gt;] [JOIN &lt;optional_join_specification&gt;]</pre>	<h3>Examples</h3> <pre>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</pre>
---	---

92

## 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	User defined functions
<p>Stored procedures perform complex transactions on documents and properties.</p> <p>Stored procedures are written in JavaScript and are stored in a collection on Azure Cosmos DB</p>	<p>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</p>

93

## 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'))"
]
```

94

## Lesson 03: Build a .NET Core App for Azure Cosmos DB in VS Code



95

### 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



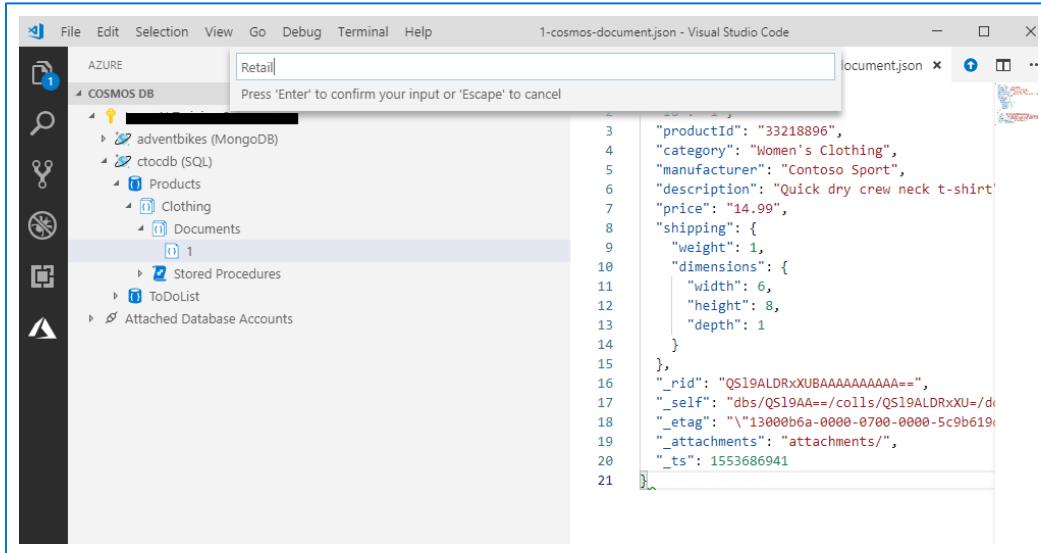
Use the Terminal in Visual Studio Code to quickly create a console application



Add Azure Cosmos DB functionality with the help of the Azure Cosmos DB extension for Visual Studio Code

96

## Creating Azure Cosmos DB in Visual Studio Code



The screenshot shows the Visual Studio Code interface with the Azure extension installed. On the left, there's a sidebar with icons for file operations like copy, paste, search, and refresh. The main area has a tree view under 'AZURE' labeled 'COSMOS DB'. It lists several databases: 'adventbikes (MongoDB)', 'ctcldb (SQL)', 'Products', 'Clothing', 'Documents', and 'Stored Procedures'. Under 'Documents', there's a node labeled '1'. To the right of the tree view is a code editor window titled '1-cosmos-document.json - Visual Studio Code'. The JSON document contains the following data:

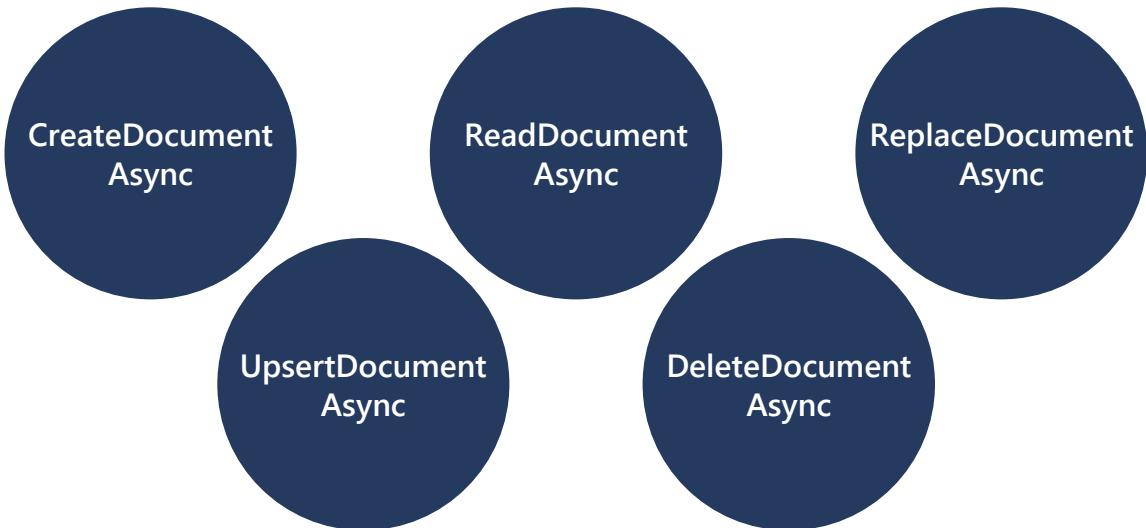
```

3 "productId": "33218896",
4 "category": "Women's Clothing",
5 "manufacturer": "Contoso Sport",
6 "description": "Quick dry crew neck t-shirt",
7 "price": "14.99",
8 "shipping": {
9   "weight": 1,
10  "dimensions": {
11    "width": 6,
12    "height": 8,
13    "depth": 1
14  }
15 },
16 "_rid": "QSl9ALDRxxUBAAAAAA==",
17 "_self": "dbs/QSl9AA==/colls/QSl9ALDRxxU=/docs/1",
18 "_etag": "\"13000b6a-0000-0700-0000-5c9b619",
19 "_attachments": "attachments/",
20 "_ts": 1553686941
21

```

97

## Working with documents programmatically



98

## Querying document programmatically

```
{
    // Set some common query options
    FeedOptions queryOptions = new FeedOptions { MaxItemCount = -1, EnableCrossPartitionQuery = true };

    // Here we find nelapin via their LastName
    IQueryable<User> userQuery = this.client.CreateDocumentQuery<User>(
        UriFactory.CreateDocumentCollectionUri(databaseName, collectionName), queryOptions)
        .Where(u => u.LastName == "Pindakova");

    // The query is executed synchronously here, but can also be executed asynchronously via the IDocumentQuery<T> interface
    Console.WriteLine("Running LINQ query...");
    foreach (User in userQuery)
    {
        Console.WriteLine("\tRead {0}", user);
    }

    // Now execute the same query via direct SQL
    IQueryable<User> userQueryInSql = this.client.CreateDocumentQuery<User>(
        UriFactory.CreateDocumentCollectionUri(databaseName, collectionName),
        "SELECT * FROM User WHERE User.lastName = 'Pindakova'", queryOptions );

    Console.WriteLine("Running direct SQL query...");
    foreach (User in userQueryInSql)
    {
        Console.WriteLine("\tRead {0}", user);
    }

    Console.WriteLine("Press any key to continue ...");
    Console.ReadKey();
}
```

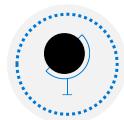
99

Lesson 04: Distribute your data globally with Azure Cosmos DB



100

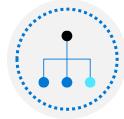
## Lesson objectives



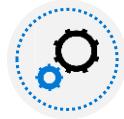
Learn about the benefits of writing and replicating data to multiple regions around the world



Cosmos DB multi-master replication



Cosmos DB failover management



Change the consistency setting for your database

101

## Benefits of writing and replicating data to multiple regions

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

**Replicate data globally**  
ctoedb

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	✓	✓	trash
UK South	✓	✓	trash
Japan West	✓	✓	trash
South Africa North	✓	✓	trash

102

## Cosmos DB multi-master replication



103

## 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

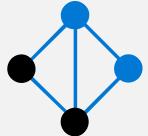
Azure Cosmos DB accounts with a read region in one of the affected regions are automatically disconnected from their write region and marked offline

### 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

104

## Lab: Building globally distributed databases with Cosmos DB



105

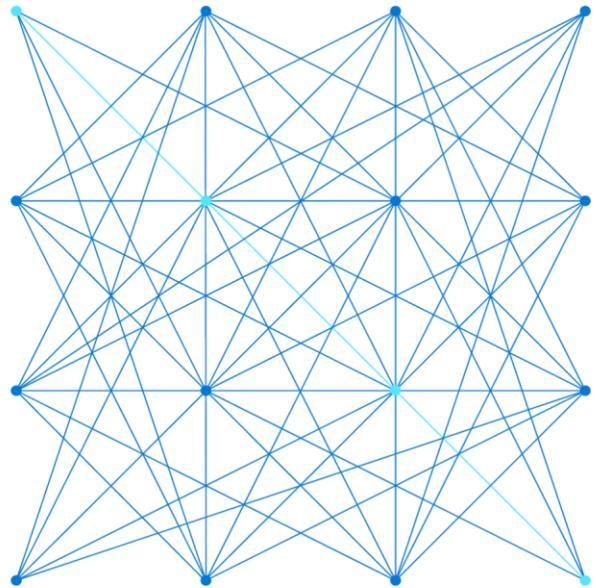
 Microsoft Azure

© Copyright Microsoft Corporation. All rights reserved.

106



# Module 05: Working with relational data stores in the cloud



107

## Agenda



Lesson 01: Work with Azure SQL database

---



Lesson 02: Work with Azure Synapse Analytics

---



Lesson 03: Provision and query data in Azure Synapse Analytics

---



Lesson 04: Import data into Azure Synapse Analytics using PolyBase

108

## Lesson 01: Azure SQL database



109

### 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

110

## Why Azure SQL Database is a good choice



Convenience



Cost



Scale



Security

111

## 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	vCores	SQL elastic pools	SQL managed instances
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	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 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	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

112

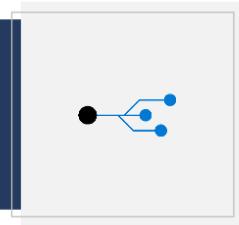
## Create an Azure SQL database

This screenshot shows the 'Create SQL Database' wizard on the 'Basics' tab. It includes fields for Subscription (chtestao), Resource group (Select existing...), Database name (Enter database name), Server (Select a server), and Compute + storage (Please select a server first). A note at the top says: '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'.

This screenshot shows the 'Create SQL Database' wizard on the 'Additional settings' tab. It highlights the 'Backup' section, which includes tabs for 'None', 'Backup' (selected), and 'Sample'. A note says: 'Start with a blank database, restore from a backup or select sample data to populate your new database.' Below are sections for 'Data source' (Use existing data or Backup), 'Database Collation' (myserver (West Europe)), and 'Collation' (Note: Database collation defines the rules that sort and compare data, and cannot be changed after database creation. The default database collation is SQL\_Latin1\_General\_CI\_AS). Step numbers 1 through 4 are overlaid on the interface.

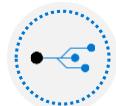
113

## Lesson 02: Azure Synapse Analytics



114

## Lesson objectives



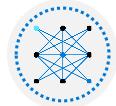
Explain Azure Synapse Analytics



Explain Azure Synapse Analytics features



Types of solution workloads



Explain massively parallel processing concepts



Compare table geometries

115

## 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

116

## 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: <ul style="list-style-type: none"> <li>• Workload Groups</li> <li>• Workload Classification</li> <li>• Workload Importance</li> </ul>
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

117

## 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

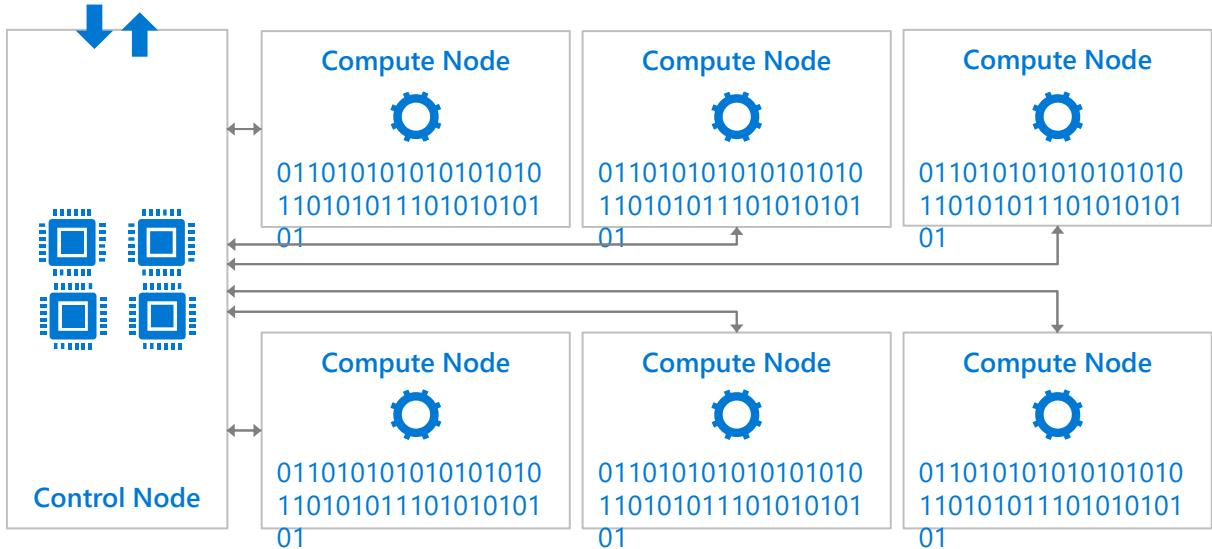
We're trying to predict when our customers churn

### Real-time analytics

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

118

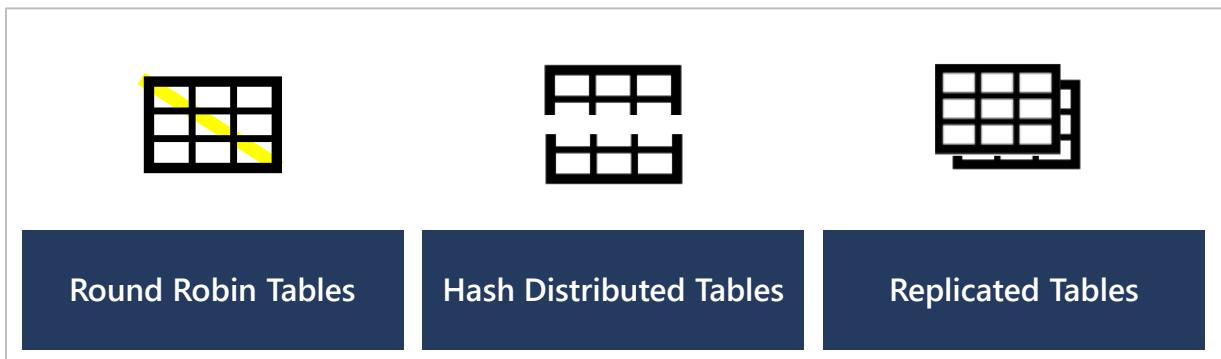
## Massively Parallel Processing (MPP) concepts



119

## Table geometries

### Table distribution



120

## Lesson 03: Creating and querying an Azure Synapse Analytics



121

### 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

122

## Create an Azure Synapse Analytics

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 \*  [Select performance level](#)

123

## 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
```

124

## Perform Azure Synapse Analytics queries

### Create Table as Select (CTAS)

Used in parallel data loads

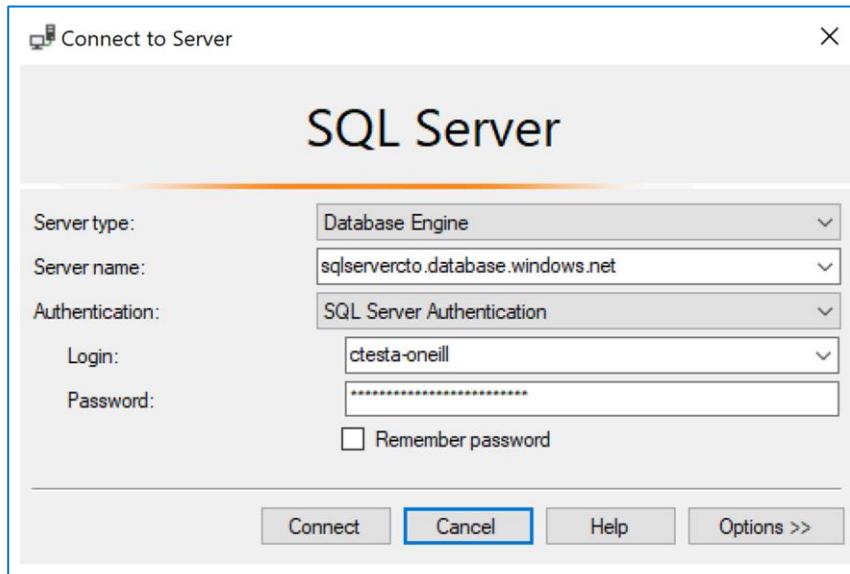
```
CREATE TABLE
[ database_name . [ schema_name ] . | schema_name. ] table_name
    [ ( { column_name } [ ,...n ] ) ]
WITH ( DISTRIBUTION =
    { HASH( distribution_column_name )
        [ , <CTAS_table_option> [ ,...n ] ]
    }
) AS <select_statement> [ ; ]
```

### Example

```
CREATE TABLE FactInternetSales_Copy
WITH
(DISTRIBUTION = HASH(SalesOrderNumber))
AS SELECT * FROM FactInternetSales
```

125

## Querying with different client applications



126

## Lesson 04: Using PolyBase to Load Data in Azure Synapse Analytics



127

### 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

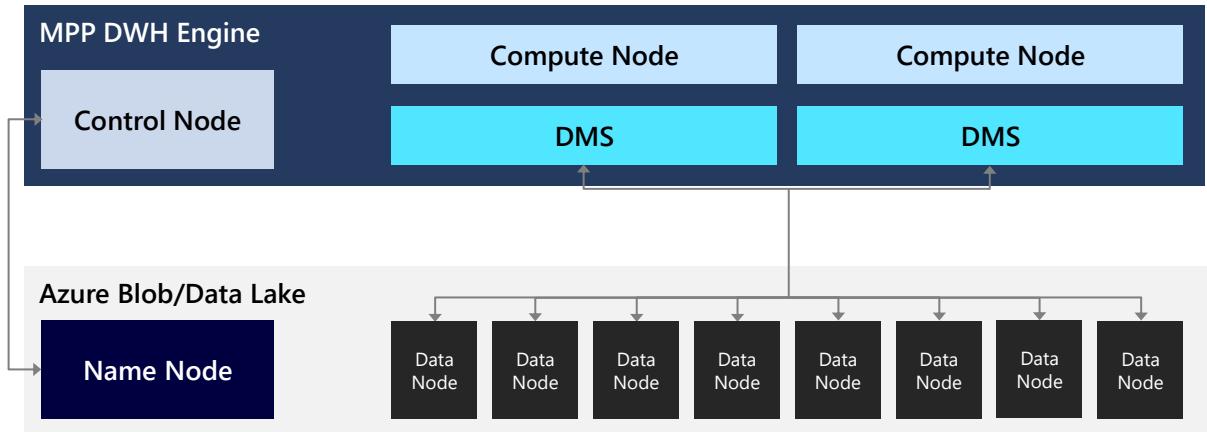


Import data from Blob Storage to the Data Warehouse

128

## How PolyBase works

### The MPP engine's integration method with PolyBase



129

## Upload text data to Azure Blob store

This screenshot shows the 'Create storage account' wizard in the Azure portal, specifically the 'Basics' step.

**Basics Tab:**

- Subscription:** A dropdown menu showing available subscriptions.
- Resource group:** A dropdown menu or input field for creating a new resource group, with a 'Create new' link.

**INSTANCE DETAILS:**

- Storage account name:** An input field containing 'teststorage123'.
- Location:** A dropdown menu set to 'West Europe'.
- Performance:** Radio buttons for 'Standard' (selected) and 'Premium'.
- Account kind:** A dropdown menu set to 'StorageV2 (general purpose v2)'.
- Replication:** A dropdown menu set to 'Read-access geo-redundant storage (RA-GRS)'.
- Access tier (default):** Radio buttons for 'Cool' and 'Hot' (selected).

**Buttons at the bottom:**

- 'Review + create'
- 'Previous'
- 'Next : Advanced >' (highlighted in blue)

130

## Collect the storage keys

The screenshot shows the 'Access keys' section for an Azure storage account named 'toazureblob'. It displays two sets of access keys: 'key1' and 'key2'. Each key includes a 'Key' value (redacted here) and a 'Connection string' value (also redacted). A note at the top states: '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.' Below this note, it says: '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.' A link 'Learn more' is provided.

131

## Create an Azure Synapse Analytics

The screenshot shows the 'Create a SQL Data Warehouse' wizard in the Azure portal. The 'Basics' tab is selected. The 'Subscription' dropdown is set to 'chtestao'. The 'Resource group' dropdown has 'Select existing...' highlighted. Under 'Data warehouse details', the 'Data warehouse name' field is empty and highlighted in red. The 'Server' dropdown is also empty and highlighted in red, with an error message: 'The value must not be empty.' The 'Performance level' dropdown is empty and highlighted in green, with an error message: 'Please select a server first.' A note at the bottom left says: 'Enter required settings for this data warehouse, including picking a logical server and configuring the performance level.'

132

## Lab: Working with relational data stores in the cloud



133

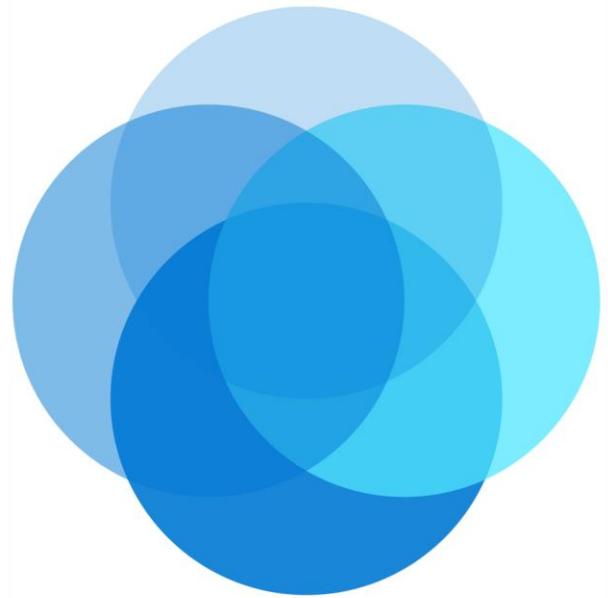
 Microsoft Azure

© Copyright Microsoft Corporation. All rights reserved.

134

# Module 06:

## Performing real-time analytics with Stream Analytics



135

### Agenda



Lesson 01 – Data streams and event processing

---



Lesson 02 – Data ingestion with Event Hubs

---



Lesson 03 – Processing data with Stream Analytics Jobs

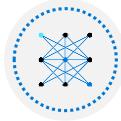
136

## Lesson 01: Data streams and event processing

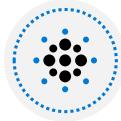


137

### Lesson objectives



Explain data streams



Explain event processing



Learn about processing events with Azure Stream Analytics

138

## What are data streams

<p><b>Data streams:</b></p> <p>In the context of analytics, data streams are event data generated by sensors or other sources that can be analyzed by another technology</p>	<p><b>Data streams are used to:</b></p> <table border="0"> <tr> <td>Analyze data: Continuously analyze data to detect issues and understand or respond to them</td><td>Understand systems: Understand component or system behavior under various conditions to fuel further enhancements of said system</td><td>Trigger actions: Trigger specific actions when certain thresholds are identified</td></tr> </table>	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
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		
<p><b>Data stream processing approach:</b></p> <p>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</p>				

139

## 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:

<p><b>Event producer</b></p>	<p>Examples include sensors or processes that generate data continuously such as a heart rate monitor or a highway toll lane sensor</p>
<p><b>Event processor</b></p>	<p>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)</p>
<p><b>Event consumer</b></p>	<p>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</p>

140

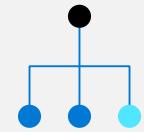
## Processing events with Azure Stream Analytics

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

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

141

## Lesson 02: Data ingestion with Event Hubs



142

## Lesson objectives

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

143

## 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 Event Hubs landing page. At the top, there's a green 'Start for free' button. Below it, a heading says 'Event Hubs' and describes it as a 'Simple, secure and scalable real-time data ingestion' service. A detailed description follows: '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.' Below this, another section discusses integration: '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.' At the bottom, there's a 'Link to video >' button.

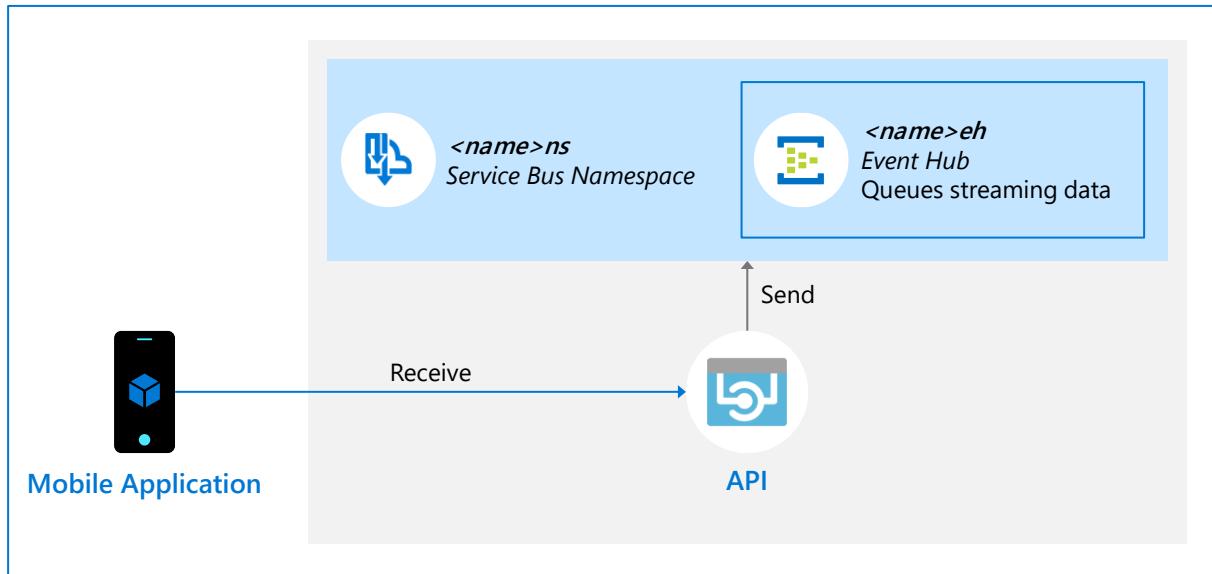
144

## Create an Event Hub

Create an event hub namespace	Create an event hub
<ol style="list-style-type: none"> <li>In the <a href="#">Azure portal</a>, select NEW, type Event Hubs, and then select Event Hubs from the resulting search. Then select Create</li> <li>Provide a name for the event hub, and then create a resource group. Specify <code>xx-name-eh</code> and <code>xx-name-rg</code> respectively, XX- represent your initials to ensure uniqueness of the Event Hub name and Resource Group name</li> <li>Click the checkbox to Pin to the dashboard, then select the Create button</li> </ol>	<ol style="list-style-type: none"> <li>After the deployment is complete, click the <code>xx-name-eh</code> event hub on the dashboard</li> <li>Then, under Entities, select Event Hubs</li> <li>To create the event hub, select the + Event Hub button. Provide the name <code>socialstudy-eh</code>, and then select Create</li> <li>To grant access to the event hub, we need to create a shared access policy. Select the <code>socialstudy-eh</code> event hub when it appears, and then, under Settings, select Shared access policies</li> <li>Under Shared access policies, create a policy with MANAGE permissions by selecting + Add. Give the policy the name of <code>xx-name-eh-sap</code>, check MANAGE, and then select Create</li> <li>Select your new policy after it has been created, and then select the copy button for the CONNECTION STRING – PRIMARY KEY entity</li> <li>Paste the CONNECTION STRING – PRIMARY KEY entity into Notepad, this is needed later in the exercise</li> <li>Leave all windows open</li> </ol>

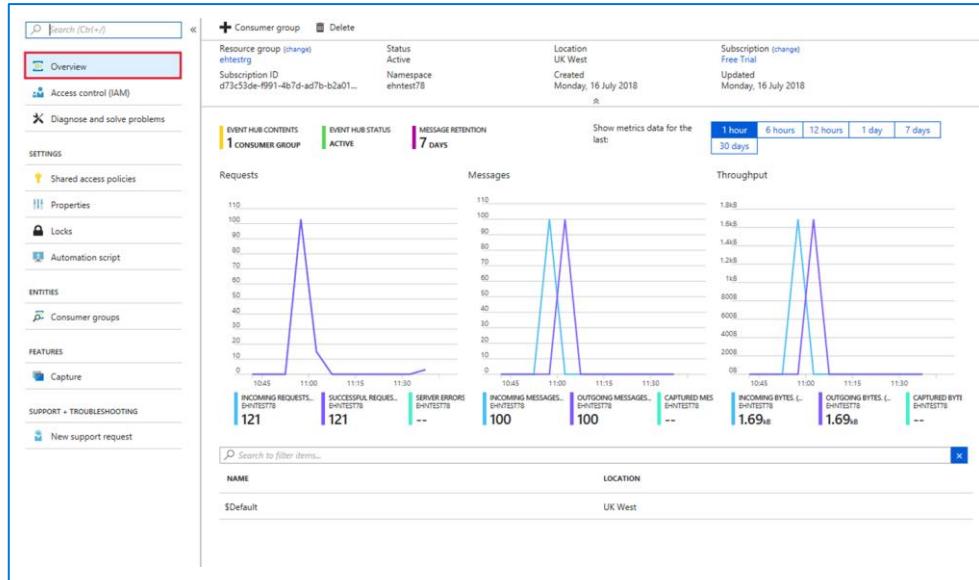
145

## Configure applications to use Event Hubs



146

## Evaluating the performance of Event Hubs



147

## Lesson 03: Processing data with Stream Analytics Jobs



148

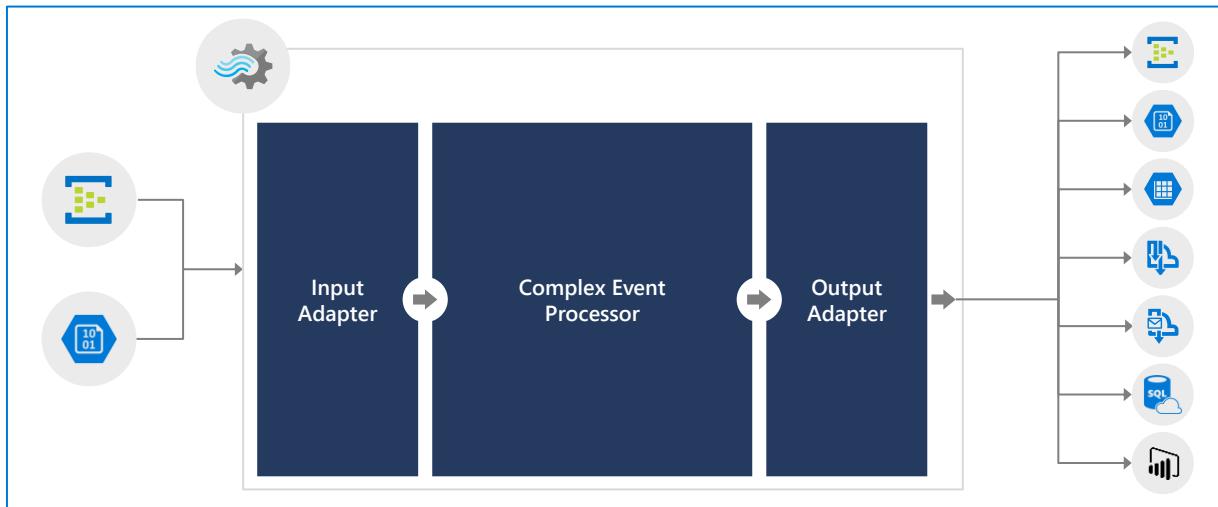
## Lesson objectives

-  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

149

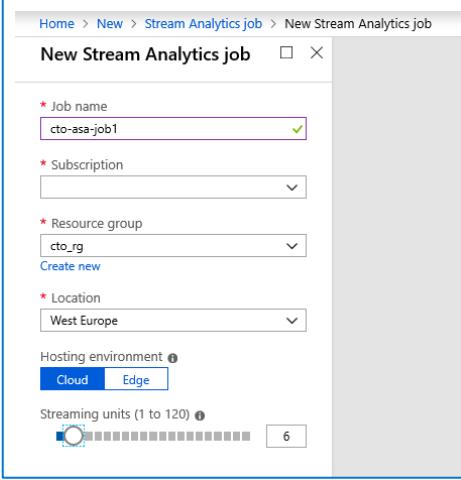
## Azure Stream Analytics workflow

Complex event processing of Stream Data in Azure



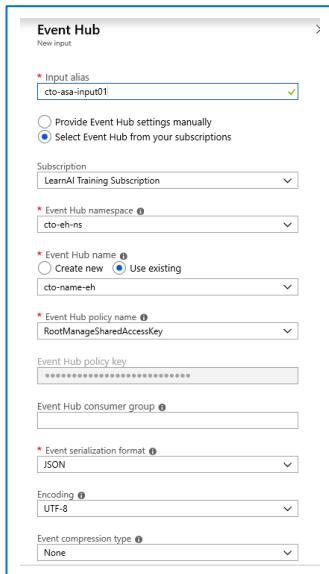
150

## Create Stream Analytics service

Job name	Subscription	Resource group	Location
			

151

## 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

152

## Create a Stream Analytics Job output

Outputs

**SINK**

**Blob storage**

**Subscription:** LearnAI Training Subscription

**Storage account:** ctoazureblob

**Container:** socialmedia

**Date format:** YYYY/MM/DD

**Time format:** HH

**Event serialization format:** JSON

**Encoding:** UTF-8

153

## Write a transformation query

cto-asa-job1

Resource group (change) : cto\_rg

Status : Created

Location : West Europe

Subscription (change) : LearnAI Training Subscription

Subscription ID : 5be49961-ea44-42ec-b728be90d58c

Inputs

1  
cto-asa-input01

Outputs

1  
cto-asa-output01

Query

```

1 SELECT
2 *
3 INTO
4 [cto-asa-output01]
5 FROM
6 [cto-asa-input01]

```

154

## Start a Stream Analytics Job

cto-asa-job1  
Stream Analytics job

Search (Ctrl+)

Start Stop Delete

Resource group (change) : cto\_rg  
Status : Created  
Location : West Europe  
Subscription (change) : LearnAI Training Subscription  
Subscription ID : 5be49961-ea44-42ec-8021-b728be90d58c

Send feedback  
Created  
Started  
Output water  
Hosting envir

Overview Activity log Access control (IAM) Tags Diagnose and solve problems

Inputs  
1 cto-asa-input01

Outputs  
1 cto-asa-output01

Query

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

155

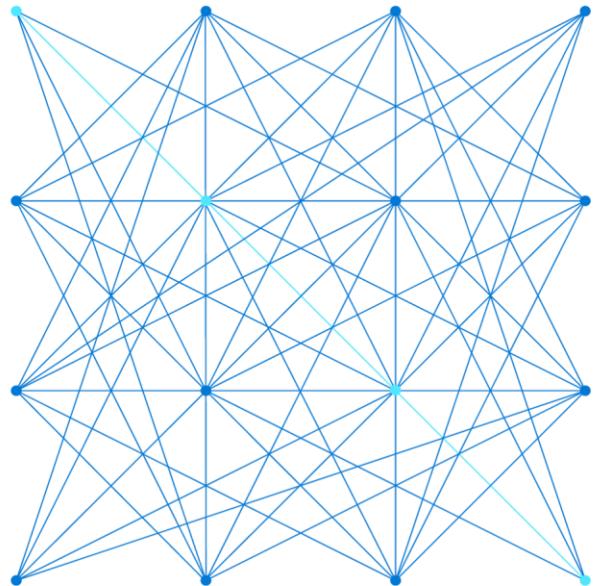
Lab: Performing real-time analytics with Stream Analytics



156

## Module 07: Orchestrating data movement with Azure Data Factory

Start : 09.15



## Agenda



Lesson 01 – Introduction to Azure Data Factory

---



Lesson 02 – Understand Azure Data Factory components

---



Lesson 03 – Integrate Azure Data Factory with Databricks

159

Lesson 01: Introduction to Azure Data Factory



160

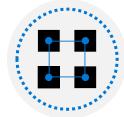
## Lesson objectives



What is Azure Data Factory



The Data Factory process



Azure Data Factory components



Azure Data Factory security

161

## What is Azure Data Factory

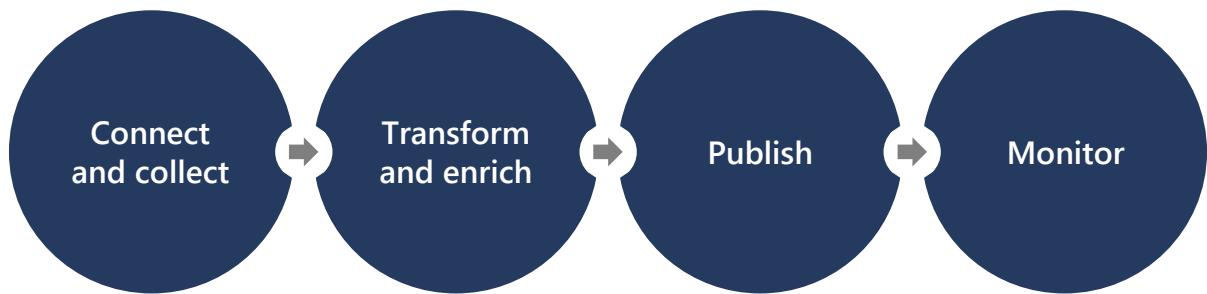


Creates, orchestrates, and automates the movement, transformation and/or analysis of data through in the cloud



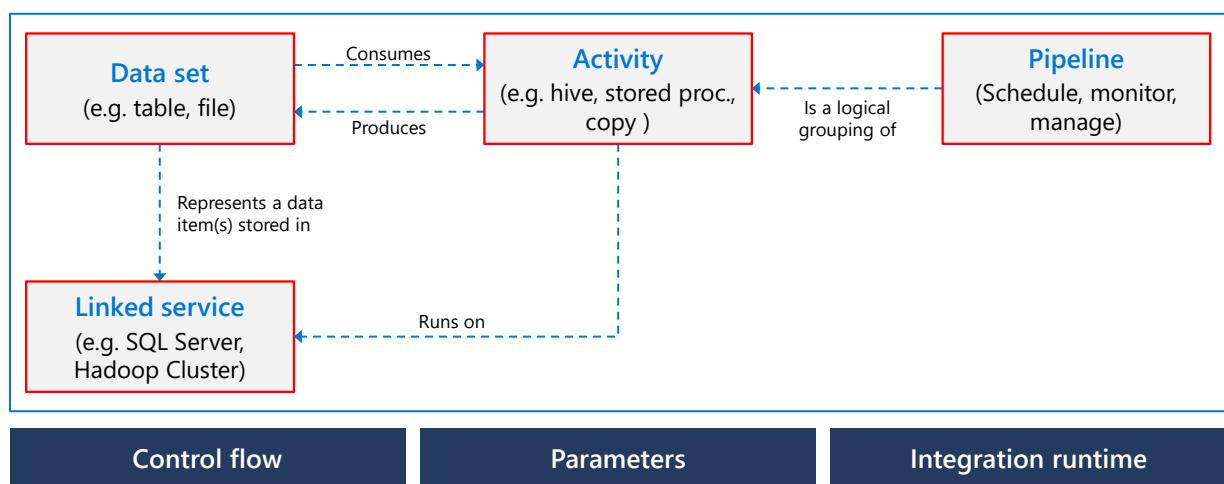
162

## The Data Factory process



163

## Azure Data Factory components



164

## 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

165

## Lesson 02: Azure Data Factory components



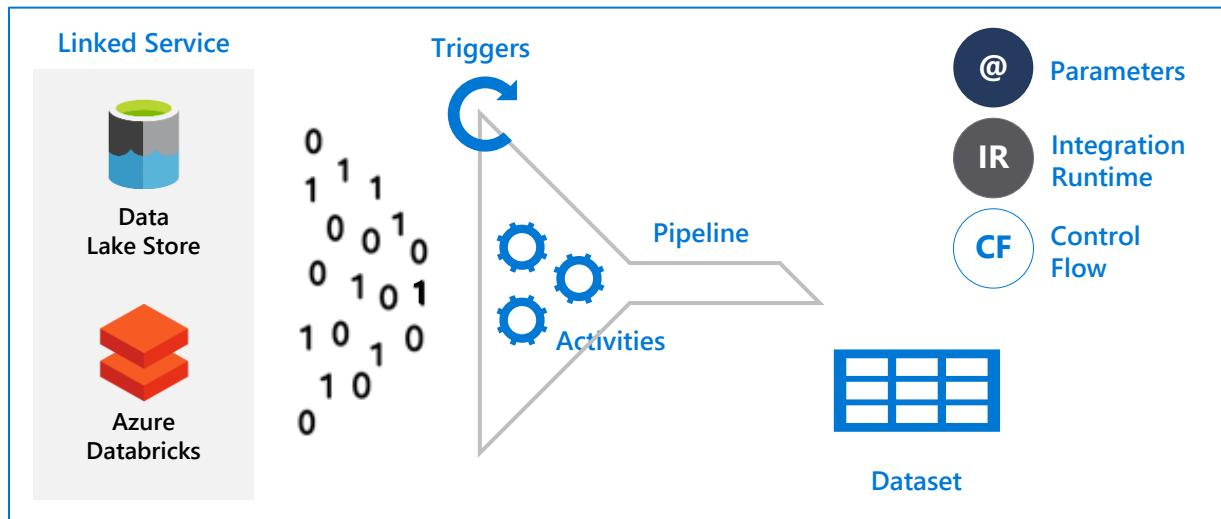
166

## Lesson objectives



167

## Azure Data Factory components



168

## Data factory activities

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

<b>Data movement activities</b>	Data movement activities simply move data from one data store to another. A common example of this is in using the Copy Activity
<b>Data transformation activities</b>	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
<b>Control Activities</b>	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

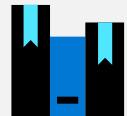
169

## Pipelines



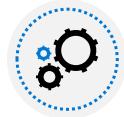
170

## Lesson 03: Ingesting and transforming data



171

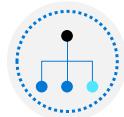
### Lesson objectives



How to setup Azure Data Factory



Ingest data using the Copy Activity



Transforming data with the Mapping Data Flow

172

## Create Azure Data Factory

Home > New > Data Factory > New data factory

New data factory

Name \*

Version

Subscription \*

Resource Group \*

Location \*

Enable GIT

GIT URL \*

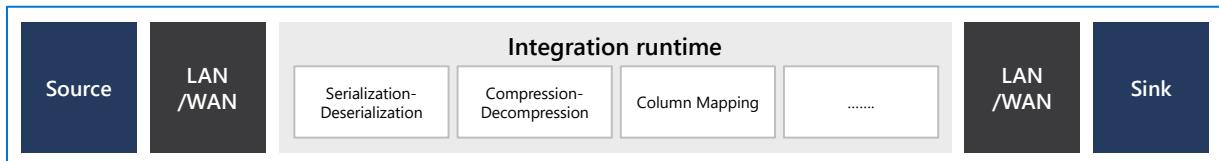
Repo name \*

Branch Name \*

Root folder \*

173

## 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

174

## 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

### Mapping Data Flow



175

## Lesson 04: Integrate Azure Data Factory with Azure Databricks



176

## Lesson objectives



Use Azure Data Factory (ADF) to ingest data and create an ADF pipeline



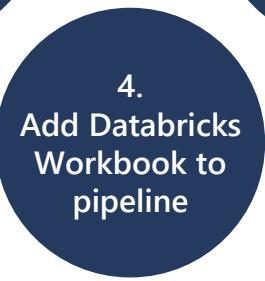
Create Azure Storage account and the Azure Data Factory instance



Use ADF to orchestrate data transformations using a Databricks Notebook activity

177

## Working with documents programmatically



178

## Create Azure Storage account and the Azure Data Factory instance

The screenshot shows two parallel creation processes in the Azure portal:

- Left Panel (Storage Account Creation):**
  - Basics Tab:** Shows the subscription as 'cthestao' and the resource group as 'Create new'.
  - Project Details:** Selects 'Resource Manager' deployment model.
  - Instance Details:** Sets the storage account name to 'l', location to 'West Europe', and performance tier to 'Standard'.
  - Review + Create:** Step 1 of 3.
- Right Panel (Data Factory Creation):**
  - New Data Factory:** Sets the name to 'l', version to 'V2', and subscription to 'cthestao'.
  - Resource Group:** Set to 'Create new'.
  - Location:** Set to 'South Central US'.
  - Enable GIT:** Checked.
  - GIT URL:** Left empty.
  - Repo Name:** Left empty.
  - Branch Name:** Left empty.
  - Root Folder:** Left empty.
  - Create:** Step 1 of 1.

179

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

The screenshot shows a Databricks notebook titled "03-Data-Transformation (python)" with the following content:

### 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.

In this lesson you:

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

180

## Lab: Orchestrating data movement with Azure Data Factory



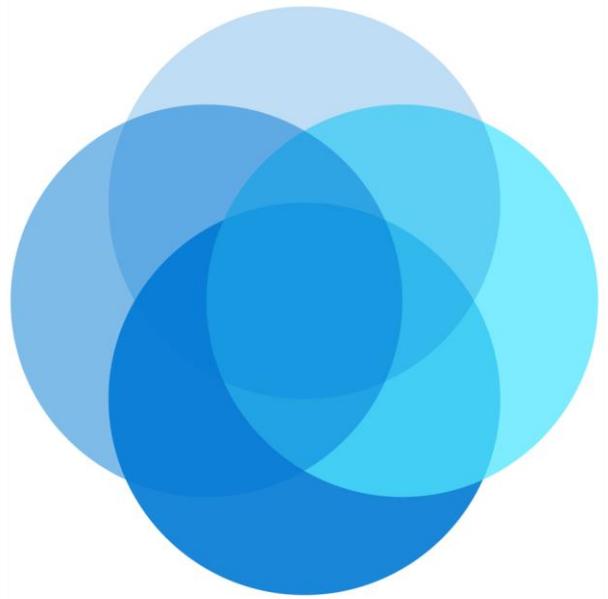
181

 Microsoft Azure

© Copyright Microsoft Corporation. All rights reserved.

182

# Module 08: Securing Azure Data Platforms



183

## Agenda



Lesson 01 – An introduction to security

---



Lesson 02 – Key security components

---



Lesson 03 – Securing storage accounts and Data Lake Storage

---



Lesson 04 – Securing data stores

---



Lesson 05 – Securing streaming data

184

## Lesson 01: An Introduction to security



185

### Lesson objectives



Shared security responsibility



A layered approach to security



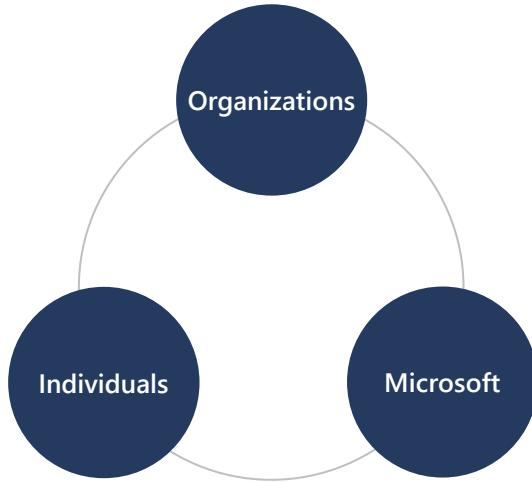
The Azure security center



Azure Government

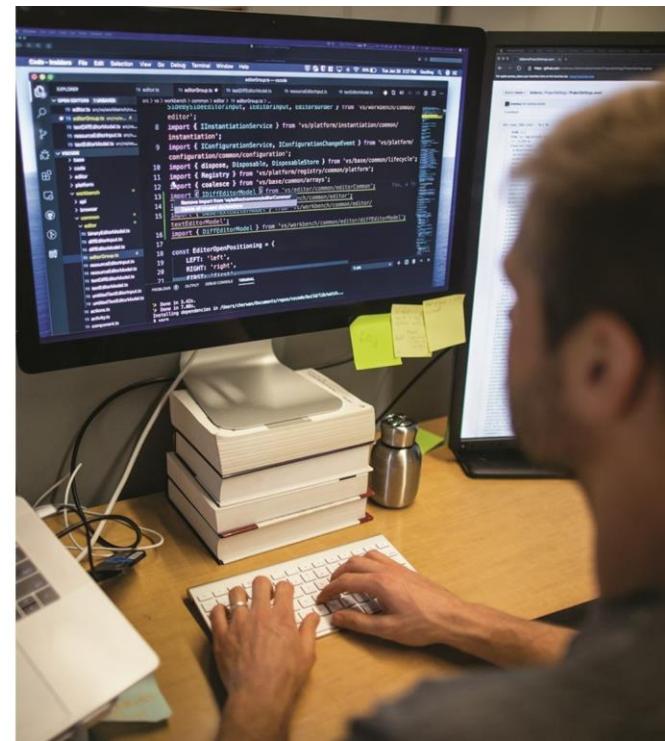
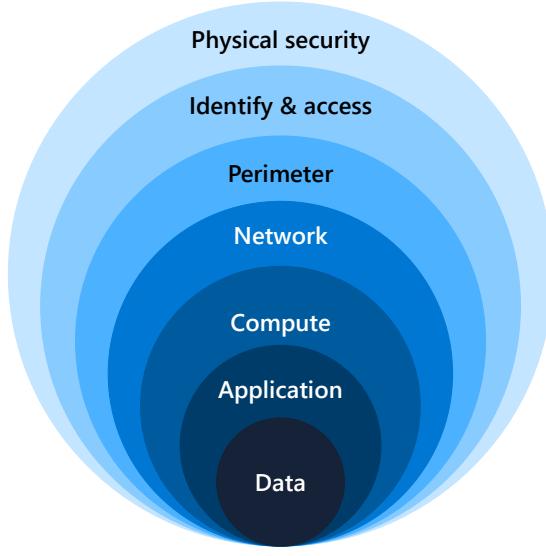
186

## Shared security responsibility



187

# A layered approach to security



188

## Azure security center

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:

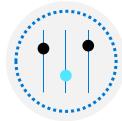
Reduce the chances of a significant security event by configuring a security policy, and then implementing the recommendations provided by Azure Security Center

189

## Azure Government



**Modernize  
Government  
services**



**Provide a  
platform of agility**



**Advanced  
Government  
mission**



**Physically separate  
from Azure**

190

## Lesson 02: Key security components



191

### Lesson objectives



Network security



Identity and access management



Encryption capabilities built into Azure



Azure threat protection

192

## Network security

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

Internet protection	Firewalls	DDoS protection	Network security groups
<p>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</p>	<p>To provide inbound protection at the perimeter, there are several choices:</p> <ul style="list-style-type: none"> <li>• Azure Firewall</li> <li>• Azure Application Gateway</li> <li>• Azure Storage Firewall</li> </ul>	<p>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</p>	<p>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</p>

193

## Identity and access

<p><b>Authentication</b></p> <p>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</p>	<p><b>Azure Active Directory features</b></p> <table border="1"> <tbody> <tr> <td> <p><b>Single sign-on</b></p> <p>Enables users to remember only one ID and one password to access multiple applications</p> </td><td> <p><b>Apps &amp; device management</b></p> <p>You can manage your cloud and on-premises apps and devices and the access to your organization's resources</p> </td><td> <p><b>Identity services</b></p> <p>Manage Business-to-business (B2B) identity services and Business-to-Customer (B2C) identity services</p> </td></tr> </tbody> </table>		<p><b>Single sign-on</b></p> <p>Enables users to remember only one ID and one password to access multiple applications</p>	<p><b>Apps &amp; device management</b></p> <p>You can manage your cloud and on-premises apps and devices and the access to your organization's resources</p>	<p><b>Identity services</b></p> <p>Manage Business-to-business (B2B) identity services and Business-to-Customer (B2C) identity services</p>
<p><b>Single sign-on</b></p> <p>Enables users to remember only one ID and one password to access multiple applications</p>	<p><b>Apps &amp; device management</b></p> <p>You can manage your cloud and on-premises apps and devices and the access to your organization's resources</p>	<p><b>Identity services</b></p> <p>Manage Business-to-business (B2B) identity services and Business-to-Customer (B2C) identity services</p>			
<p><b>Authorization</b></p> <p>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</p>					

194

## 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

195

## Azure threat protection

The screenshot shows the Azure Advanced Threat Protection Timeline interface for the contoso-corp tenant. The timeline lists several threat events:

- 4:04 PM Today**: Honeypot activity. The following activities were performed by Bob Minion:
  - Logged in to 2 computers via Contoso-DC.
  - Authenticated from 2 computers using Kerberos when accessing 1 resources against Contoso-DC.
  - Authenticated from PARISDET-FA100 using NTLM against corporate resources via Contoso-DC.
- 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.
- 2:09 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.
- 1:38 PM Jan 21, 2018**: Malicious replication of directory services. Malicious replication requests were attempted by Alice Lübel, from ALICE-DESKTOP against Contoso-DC.
- 1:09 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.

196

## Lesson 03: Securing storage accounts and Data Lake Storage



197

### 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

198

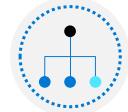
## Storage account security features



Encryption  
at rest



Encryption  
in transit



Role based  
access control



Auditing  
access

199

## Storage account keys

Home > Resource groups > cto\_rg > ctoazureblob - Access keys

**ctoazureblob - Access keys**

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

Use access keys to authenticate your applications when making requests to this Azure storage account. Store your access keys securely - for example, using Azure Key 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: eU[REDACTED]Cg==

Connection string: DefaultEndpointsProtocol=https;AccountName=ctoazureblob;AccountKey=eU[REDACTED]Cg==;EndpointSuffix=core.windows.net

**key2**

Key: NW[REDACTED]VpUgB5w==

Connection string: DefaultEndpointsProtocol=https;AccountName=ctoazureblob;AccountKey=NW[REDACTED]VpUgB5w==;EndpointSuffix=core.windows.net

200

## Shared access signatures

The screenshot shows the 'Shared access signature' blade for the 'ctoazureblob' storage account. It includes sections for Allowed services (Blob, File, Queue, Table), Allowed resource types (Service, Container, Object), and Allowed permissions (Read, Write, Delete, List, Add, Create, Update, Process). It also shows the Start and expiry date/time, Allowed IP addresses (e.g., 168.1.5.65 or 168.1.5.65-168.1.5.70), Allowed protocols (HTTPS only selected), and a Signing key dropdown set to 'key1'. A 'Generate SAS and connection string' button is at the bottom.

201

## Control network access to data

The screenshot shows the 'Firewalls and virtual networks' blade. It has sections for Firewall settings (warning about immediate effect), Allow access from (Selected networks selected), Virtual networks (Add existing virtual network, Add new virtual network), and Firewall rules (No network selected). It also includes ADDRESS RANGE and Exceptions sections (Allow trusted Microsoft services checked).

202

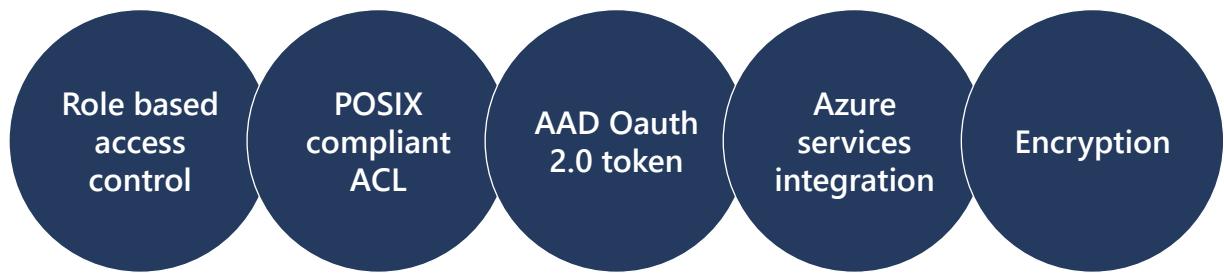
## Managing encryption

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

<b>Transport Layer Security (TLS)</b>	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
<b>Transparent data encryption</b>	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
<b>Application encryption</b>	Data in transit is a method to prevent man-in-the-middle attacks. To encrypt data in transit, specify <a href="#">Encrypt=true</a> in the connection string in your client applications

203

## Azure Data Lake Storage Gen2 security features



204

## Lesson 04: Securing data stores



205

### 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

206

## 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

207

## 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

208

## Dynamic data masking

Masking rules

MASK NAME	MASK FUNCTION
You haven't created any masking rules.	
SQL users excluded from masking (administrators are always excluded) <small>i</small>	
<input checked="" type="checkbox"/> SQL users excluded from masking (administrators are always excluded)	

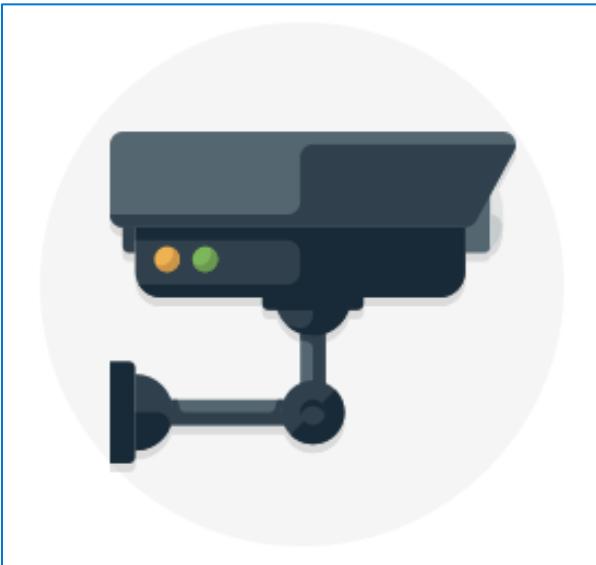
Recommended fields to mask

SCHEMA	TABLE	COLUMN	
SalesLT	Address	AddressID	<button>Add mask</button>
SalesLT	Address	AddressLine1	<button>Add mask</button>
SalesLT	Address	AddressLine2	<button>Add mask</button>
SalesLT	Customer	FirstName	<button>Add mask</button>
SalesLT	Customer	LastName	<button>Add mask</button>

[Load more](#)

209

## Auditing and monitoring



210

## Lesson 05: Securing streaming data



211

### Lesson objectives



Understand Stream Analytics security



Understand Event Hub security

212

## Stream Analytics security

### Data in transit

Azure Stream Analytics encrypts all incoming and outgoing communications and supports Transport Layer Security v 1.2

### Data at rest

Stream Analytics doesn't store the incoming data since all processing is done in-memory. Therefore, consider setting security for services such as Event Hubs or Internet of Things Hubs, or for data stores such as Cosmos DB

213

## Event Hub security

### Authentication

Authentication makes use of Shared Access Signatures and Event Publishers to ensure that only applications or devices with valid credentials are only allowed to send data to an Event Hub. Each client is assigned a token

### Token management

Once the tokens have been created, each client is provisioned with its own unique token. If a token is stolen by an attacker, the attacker can impersonate the client whose token has been stolen. Adding a client to a blocked recipients list renders that client unusable

214

## Lab: Securing Azure Data Platforms



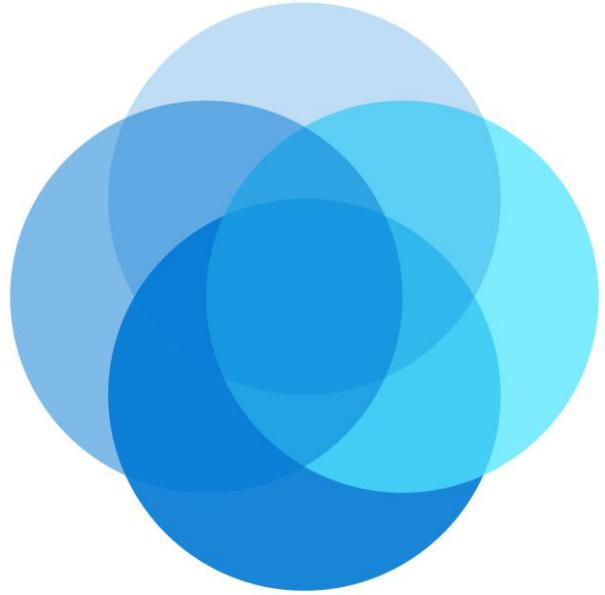
215

 Microsoft Azure

© Copyright Microsoft Corporation. All rights reserved.

216

# Module 09: Monitoring and troubleshooting data storage and processing



217

## Agenda



Lesson 01: General Azure monitoring capabilities

---



Lesson 02: Troubleshoot common data storage issues

---



Lesson 03: Troubleshoot common data processing issues

---



Lesson 04: Manage disaster recovery

218

## Lesson 01: General Azure monitoring capabilities

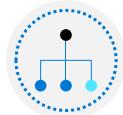


219

## Lesson objectives



Azure Monitor



Monitoring the network



Diagnose and solve problems

220

## 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

221

## Monitoring the network

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

Network performance monitor	Application gateway analytics
Network Performance Monitor measures the performance and reachability of the networks that you have configured	<p>Application Gateway Analytics contains rich, out-of-the box views you can get insights into key scenarios, including:</p> <ul style="list-style-type: none"> <li>• Monitor client and server errors</li> <li>• Check requests per hour</li> </ul>

222

## Diagnose and solve issues

The screenshot shows the 'ctocdb - Diagnose and solve problems' blade in the Azure portal. The left sidebar contains navigation links: 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 area is titled 'RESOURCE HEALTH' and shows a green 'Available' status with the message: 'There aren't any known problems affecting this Cosmos DB database account'. Below this is the 'RECENT ACTIVITY' section, which is currently empty. The 'SOLUTIONS TO COMMON PROBLEMS' section lists several items with dropdown arrows:

- 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

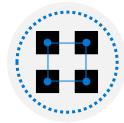
223

## Lesson 02: Troubleshoot common data storage issues



224

## Lesson objectives



Connectivity issues



Performance issues



Storage issues

225

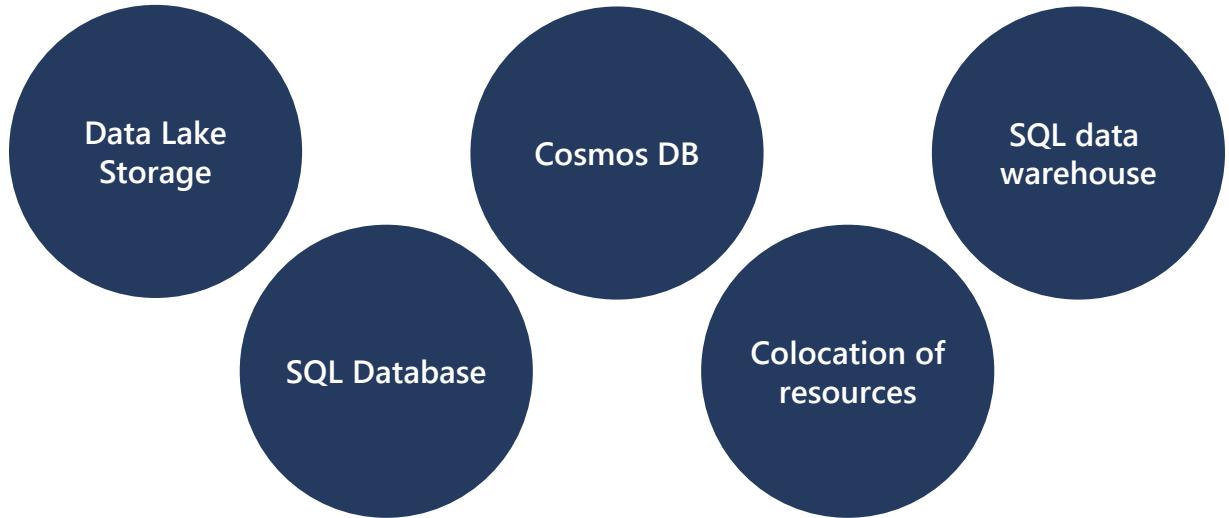
## Connectivity issues

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

Unable to connect to the data platform	Authentication failures	Cosmos DB MongoDB API errors	SQL database failover
<p>The first area that you should check is the firewall configuration</p> <p>Test the connection by accessing it from a location external to your network</p> <p>Check maintenance schedules</p>	<p>The first check is to ensure that the username and password is correct</p> <p>Check the storage account keys and ensure that they match in the connection string</p>	<p>Mongo client drivers establishes more than one connection</p> <p>On the server side, connections which are idle for more than 30 minutes are automatically closed down</p> <p>Check for timeouts</p>	<p>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</p>

226

## Performance issues



227

## Storage issues

### Consistency:

Consider the consistency levels of the following data stores that can impact data consistency:

- Cosmos DB
- SQL Data Warehouse
- SQL Database

### Corruption:

Data corruption can occur on any of the data platforms for a variety of reasons. You should have an appropriate disaster recovery strategy

228

## Lesson 03: Troubleshoot common data processing issues



229

### Lesson objectives



Troubleshoot streaming data



Troubleshoot batch data loads



Troubleshoot Azure Data Factory

230

## Troubleshoot streaming data

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

<b>Job input</b>	The job input contains a <b>Test Connection</b> 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
<b>Job query</b>	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
<b>Job output</b>	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

231

## Troubleshoot batch data loads

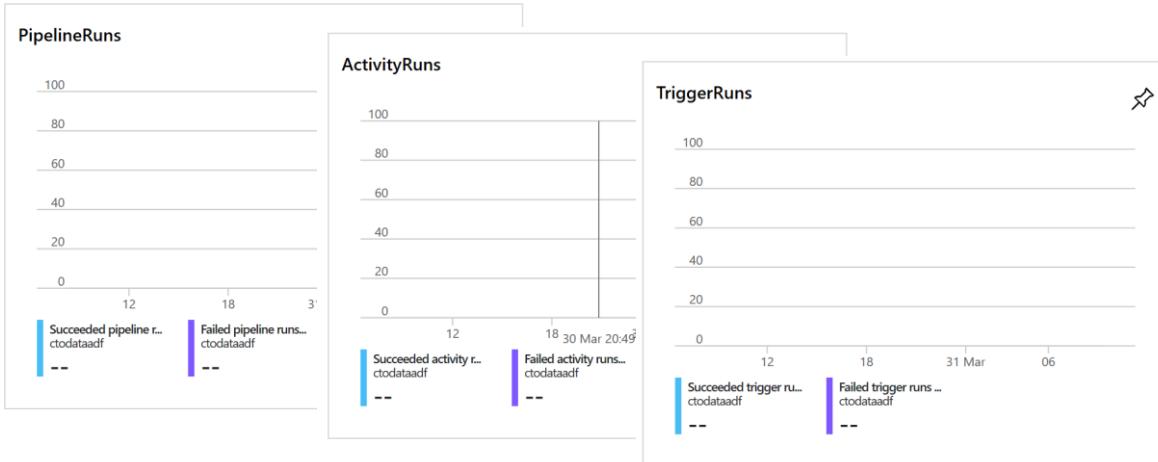
When trying to resolve data load issues, it is first pragmatic to make the holistic checks on Azure, as well as the network checks and diagnose and solve issue check. After that, then check:

Azure Blob and Data Lake Store	SQL Data Warehouse	Cosmos DB	SQL Database
Notwithstanding network errors; occasionally, you can get timeout or throttling errors that can be a symptom of the availability of the storage accounts	<ul style="list-style-type: none"> <li>Make sure you are always leveraging PolyBase</li> <li>Ensure CTAS statements are used to load data</li> <li>Break data down into multiple text files</li> <li>Consider DWU usage</li> </ul>	<ul style="list-style-type: none"> <li>Check that you have provisioned enough RU's</li> <li>Review partitions and partitioning keys</li> <li>Check for client connection string settings</li> </ul>	<ul style="list-style-type: none"> <li>Check that you have provisioned enough DTU's</li> <li>Review whether the database would benefit from elastic pools</li> <li>A wide range of tools can be used to troubleshoot SQL Database</li> </ul>

232

## Troubleshoot Azure Data Factory

### Monitoring



233

## Lesson 04: Managing disaster recovery



234

## Lesson objectives



Data redundancy



Disaster recovery

235

## 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	SQL Data Warehouse	Cosmos DB	SQL Database
Locally redundant storage (LRS) Zone-redundant storage (ZRS) Geo-redundant storage (GRS) Read-access geo-redundant storage (RA-GRS)	SQL Data Warehouse performs a <a href="#">geo-backup</a> once per day to a paired data center. The RPO for a geo-restore is 24 hours	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	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

236

## 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>

237

Lab: Monitoring and troubleshooting data storage and processing



238



© Copyright Microsoft Corporation. All rights reserved.