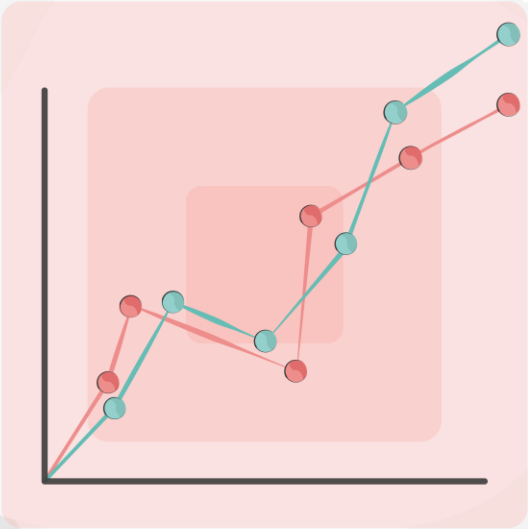
Level 5 Data Engineer
Module 5 Topic 2

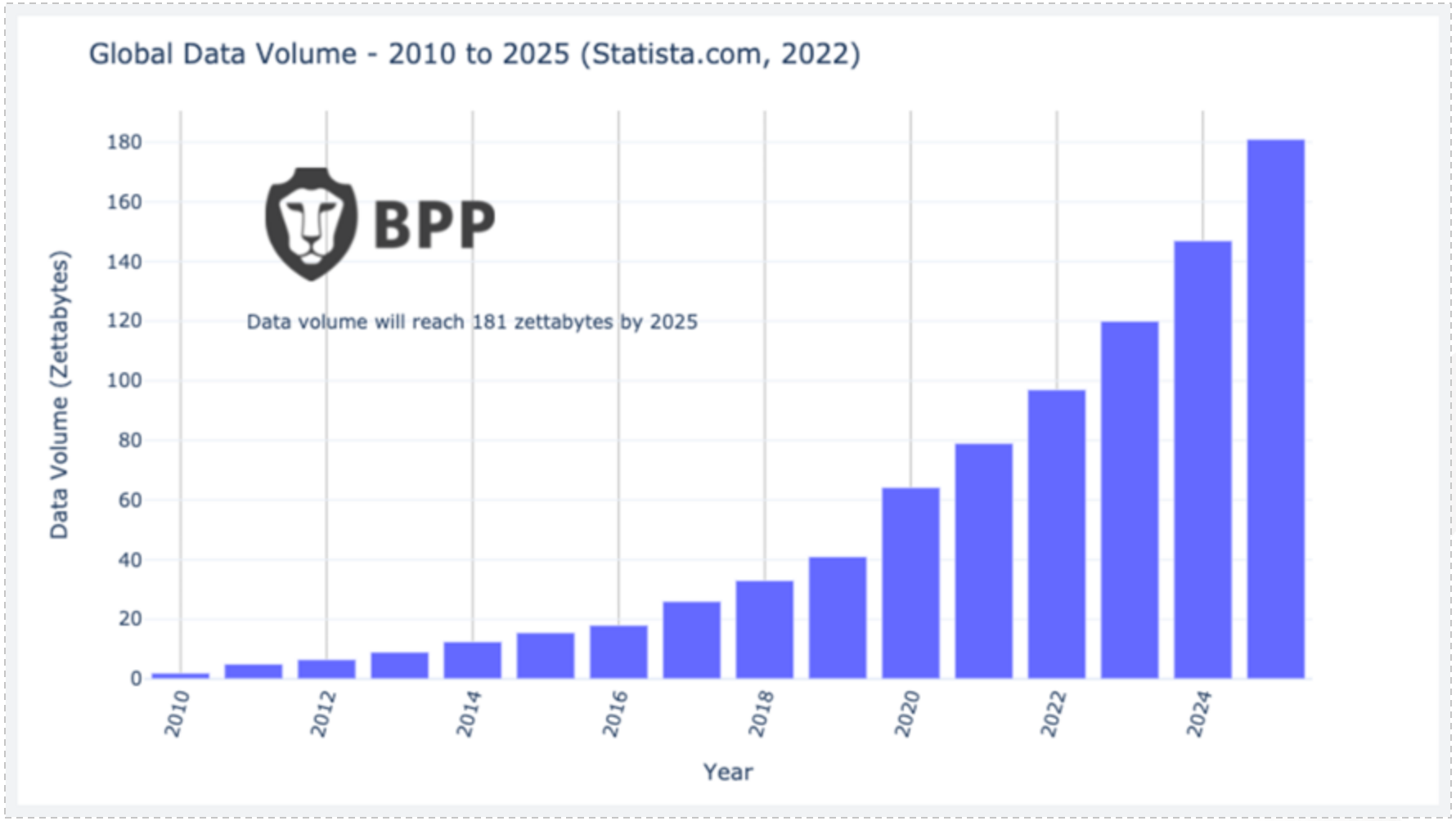
Data in the cloud



Welcome to today's webinar.

The rise of data

Current volume at 100 Zettabytes



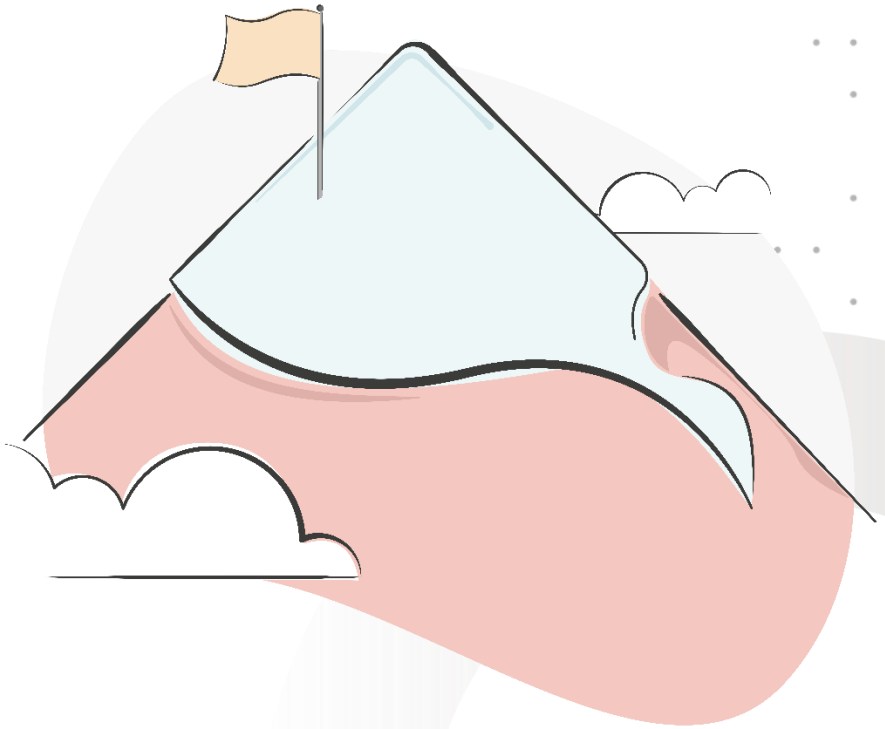
Session objectives

Building Careers Through Education

Everybody Matters. Stronger Together. Student, Learner and Client Centric. Trust and Respect. Embrace Change.

This webinar supports the following learning outcomes:

- Consider the available data sources internally and externally
- Critically evaluate limitations business may face (structured, semi-structured and unstructured data)
- Evaluate a data source with an appropriate cloud platform demonstrating your rationale and views
- Ethical the ethical issues with managing data sources
- Critique the usefulness of structures such as NoSQL databases, data warehouses and data lakes as sources of big data
- Demonstrate proficiency in a simple cloud storage tool (Azure Bob Storage)



BPP

Webinar Agenda

What we will cover in the webinar:

1. Internal and External Sources of Big Data
2. The Impact of Big Data Structure
3. Cloud Big Data Storage Platforms
4. Ethics of Big Data Storage
5. Azure Blob Storage – Demo

The drivers of Big Data

Traditional vs today…

Traditional → Structured data, Rows and Columns, Relational Databases, Data Warehouses

Data Today →

Google

YouTube

fitbit

And so much more!

Building Careers Through Education
Everybody Matters. Stronger Together. Student, Learner and Client Centric. Trust and Respect. Embrace Change.

BPP

Discussion

Internal and external sources of Big Data

1. What do you consider to be the main benefits of using an internal source for big data?

2. In contrast, what benefits of using an external source for big data can you identify?

3. Envisage a scenario where your organisation is considering making a internal big data set “open data”.

4. What benefits might be gained in doing so? Are there any risks involved?

Submit your responses to the chat!

The impact of Big Data structure

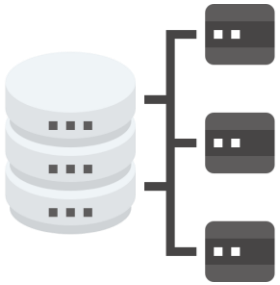
Point 1…

Structure?	Examples

Structured	`Name,Age,Species,Subspecies`
	`Nick,31,Human,`
	`Max,2,Cat,Persian`
	`Cecil,5,Dog,Yorkie-Chihuahua`

Structured Data

Defined Data Type, Format

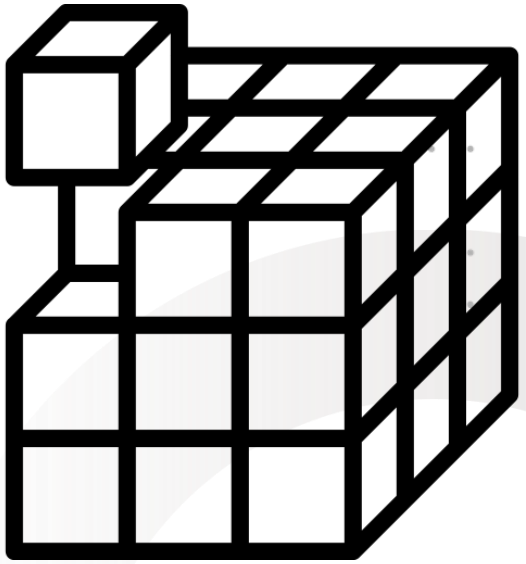


Transaction Data

Online Analytical Processing Data Cubes

Traditional RDBMS

CSV Files, Spreadsheets



The impact of Big Data structure

Point 2…

Structure?	Examples
Structured	Name,Age,Species,Subspecies Nick,31,Human, Max,2,Cat,Persian Cecil,5,Dog,Yorkie-Chihuahua
Semi-structured	[{“name”: “Nick”, “age”: 31, “species”: “Human”, “occupation”: “Lecturer”, “appearance”: “Smiling”, “wearing”: “Black jumper”}, {“name”: “Max”, “age”: 2, “species”: “Cat”, “subspecies”: “Persian”, “colour”: “Red”, “sat-on”: “Scratching post”}, {“name”: “Cecil”, “age”: 5, “species”: “Dog”, “appearance”: “Relaxed”, “subspecies”: [“Yorkie”, “Chihuahua”], “colour”: “Tan”, “sat-on”: “Dog beds”}]

Semi-structured Data

Building Careers Through Education

Everybody Matters. Stronger Together. Student, Learner and Client Centric. Trust and Respect. Embrace Change.

Semi-Structured

Textual data files that enable parsing (XML data files)

```
<?xml version="1.0" encoding="UTF-8"?>
- <EmployeeData>
  - <employee id="34594">
      <firstName>Heather</firstName>
      <lastName>Banks</lastName>
      <hireDate>1/19/1998</hireDate>
      <deptCode>BB001</deptCode>
      <salary>72000</salary>
    </employee>
  - <employee id="34593">
      <firstName>Tina</firstName>
      <lastName>Young</lastName>
      <hireDate>4/1/2010</hireDate>
      <deptCode>BB001</deptCode>
      <salary>65000</salary>
    </employee>
</EmployeeData>
```

BPP

The impact of Big Data structure

Point 3…

Structure?	Examples

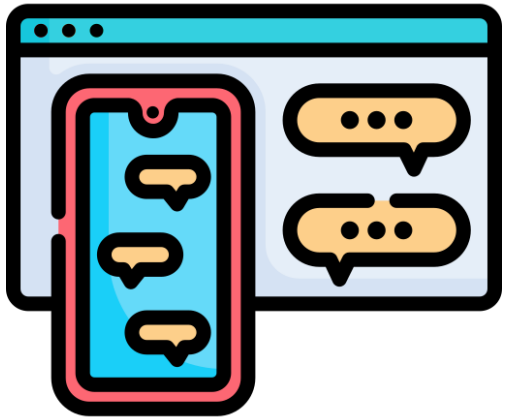
Structured	`Name,Age,Species,Subspecies`
	`Nick,31,Human,`
	`Max,2,Cat,Persian`
	`Cecil,5,Dog,Yorkie-Chihuahua`

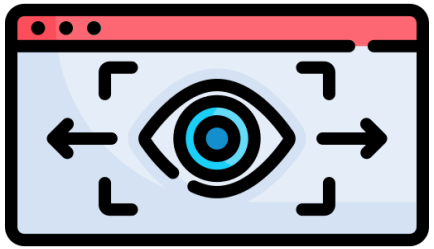
Semi-structured	`[`
	`  {“name”: “Nick”, “age”: 31, “species”: “Human”, “occupation”: “Lecturer”,`
	`  “appearance”: “Smiling”, “wearing”: “Black jumper”},`
	`  {“name”: “Max”, “age”: 2, “species”: “Cat”, “subspecies”: “Persian”,`
	`  “colour”: “Red”, “sat-on”: “Scratching post”},`
	`  {“name”: “Cecil”, “age”: 5, “species”: “Dog”, “appearance”: “Relaxed”,`
	`  “subspecies”: [“Yorkie”, “Chihuahua”], “colour”: “Tan”, “sat-on”: “Dog beds”}`
	`]`

Unstructured	  

Unstructured Data

No apparent or obvious structure
Difficult to parse or process





BPP

Discussion

The impact of Big Data structure

1. What examples of data that can be generated from image data can you think of?

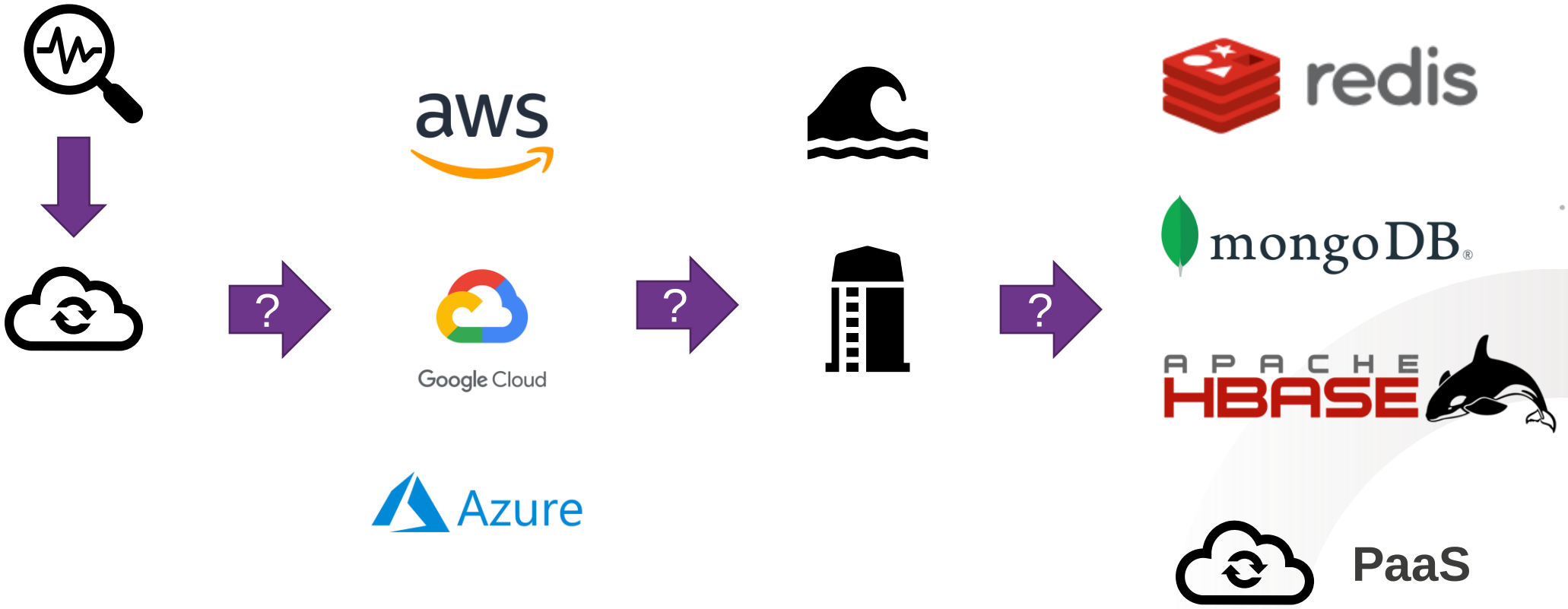
2. What techniques can be used when processing unstructured data?

3. In your experience, how do they compare to processing (semi-)structured data (in terms of ease of use, computational requirements, etc.)?

Submit your responses to the chat!

Cloud Big Data storage platforms

Decisions



Building Careers Through Education
Everybody Matters.
Stronger Together.
Student, Learner and Client Centric.
Trust and Respect.
Embrace Change.
BPP

Decision activity

Cloud Big Data storage platforms

- What factors would you consider to be the most significant when selecting an cloud big data storage platform?

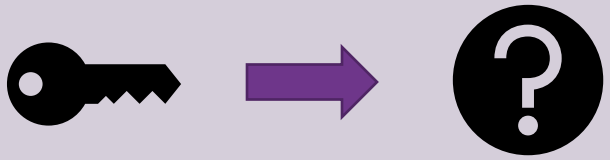
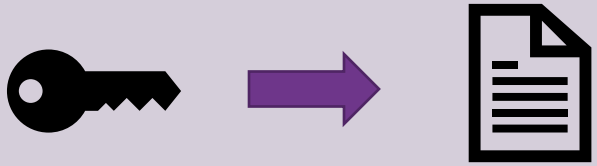
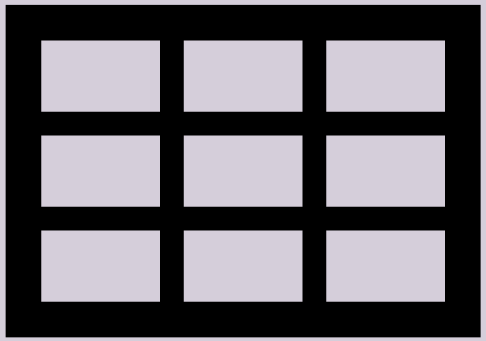
When storing data for a big data project, some organisations will select between having a data lake or a data warehouse, whilst others will use both types.

Identify a benefit and a drawback for each of these approaches.

Submit your responses to the chat!

Big Data storage platforms

NoSQL…

Key-Value	Document	Column-Orientated
		

Cloud Big Data storage platforms

NoSQL activity…

Aspect	Key-Value	Document	Column-Orientated
Data Structure			
Standard Queries			
Ease of Distribution (High-Medium-Low)			

Building Careers Through Education
Everybody Matters.
Stronger Together.
Student, Learner and Client Centric.
Trust and Respect.
Embrace Change.
BPP

Cloud Big Data storage platforms

NoSQL activity model answer…

Aspect	Key-Value	Document	Column-Orientated
Data Structure	Structured keys, unstructured values	Semi-structured documents	Structured values
Standard Queries	Keys, key prefixes	Keys, document terms, document structure	Columns, value comparisons
Ease of Distribution (High-Medium-Low)	High	Medium	Low

Building Careers Through Education
Everybody Matters.
Stronger Together.
Student, Learner and Client Centric.
Trust and Respect.
Embrace Change.
BPP

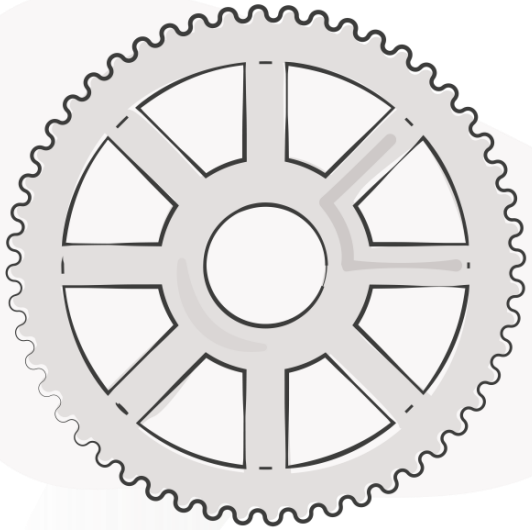
Walkthrough

Pricing NoSQL in the cloud

See the resource "L5DE M5T2 Azure Pricing Calculator".

Your tutor will guide you through this demo.

Objective: exploring data egress costs (focusing on NoSQL) and specifying data read and data write requirements



BPP

Activity

Ethics of Big Data Storage

1. Thinking back to a source of big data you previously identified, are there any privacy concerns that would be relevant when storing that data?

2. What requirements for data holders from relevant privacy laws are you aware of that would affect the storage of this data

3. How might they affect your choices when designing a big data storage solution?

Submit your responses to the chat!

Ethical Big Data storage

Relevant laws…

- Personal data (that identifies natural person) was regulated and protected in the UK under the **Data Protection Act (DPA) 1998** until late 2018

- This has been superseded by the **General Data Protection Regulation (GDPR)** and the updated **DPA 2028**

- **Everyone** is responsible for using personal data to follow strict rules called **‘data protection principles’**

Ethical Big Data storage

Compliance with DPA / GDPR…

- If you are processing any data that does, or could potentially, identify a person or groups of people, you need to **comply with the DPA 2018, and hence also GDPR**

- **Everyone** in your organization is responsible

  - As long as they may be exposed to personal data **(personal information)**

Ethical Big Data storage

Data protection jargon buster…

Processing is anything that is done with personal data:

collecting, recording, organising, structuring, storing, adapting, altering, erasing

A **data subject** is an identified or identifiable person.

A **controller**

determines why the personal data has to be processed.

A **processor**

processes data

on behalf of a controller.

Building Careers Through Education
Everybody Matters.
Stronger Together.
Student, Learner and Client Centric.
Trust and Respect.
Embrace Change.
BPP

Ethical Big Data storage

Enforcement of GDPR…

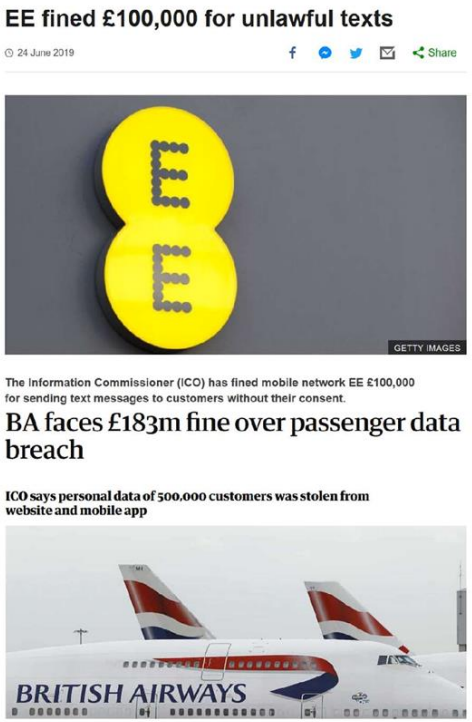
- The maximum fine under the GDPR is up to 4% of annual global turnover or $20 million – whichever is greter
- However, not all GDPR infringements lead to data protection fines. ICO (Information Commissioner's Office) can take a range of other actions, including:

➢ Issue **warnings** and reprimands

➢ Imposing a temporary or permanent **ban** on data processing

➢ Ordering the rectification, restriction or **erasure** of data, and;

➢ Suspending data **transfers** to third countries



Ethical Big Data storage

Recent ICO enforcements…

- Marriot International fined £90 million
- British Airways fined £183 million
    - For ‘insufficient technical and organisational measures to ensure information security’
- Estate Agents Life Residential fined £80,000 for leaving 19,000 customer’s personal data exposed for almost two years

Ethical Big Data storage

Personal data…

- **GDPR applies to personal data**

- **Personal data** is information that relates to an identified or identifiable individual

- The GDPR provides a non-exhaustive list of **identifies**, including:

  - **Name;**
  - **Identification number;**
  - **Location data, and;**
  - **An online identifier (IP or cookie data)**
  - **Etc…**

Ethical Big Data storage

Non-personal, personal, sensitive…

Non-personal data	Personal data	Sensitive personal data
Address without a name	Individual's name and address	Political views
Corporate accounts summary	Personal email address	Ethnic origin
Company name and website address	A web cookie	Sexual orientation
A receipt with date, time, Items and last 4 digits of credit card		
A generic email address such as enquiries@lsbf.uk	Name and last 4 digits of credit card	Religious belief

Ethical Big Data storage

Personal data in context…

House price – A. used to determine the level of house prices in a district

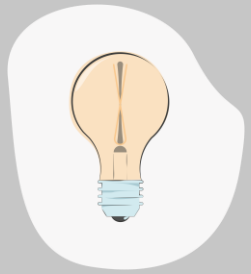
House price – B. used to determine the level of taxes an occupier may be paying

Discussion activity

True or false

Is the following statement true or false?

- *‘anonymised data is outside of data protection law?’*



Wrong!

If it is feasible to (re)-identify a natural person, data protection principles apply – you must assess the risk.

Anonymised personal data can often be identified to a natural person by use of additional information – so should be considered personal information (on a identifiable natural person).

Submit your responses to the chat!

Ethical Big Data storage

GDPR: The 7 data protection principles…

1. **Lawfulness, fairness and transparency**
2. **Purpose limitation**
3. **Data minimisation**
4. **Accuracy**
5. **Storage limitation**
6. **Integrity and confidentiality**
7. **Accountability**

Ethical Big Data storage

Data protection principles and implementation…

Data protection principles	Implementation
Lawfulness, fairness and transparency	e.g. data asset registers, explainable AI
Purpose limitation	e.g. mechanism to attach purpose to processing activity
Data minimisation	e.g. masking, pseudonymising, rounding, reducing size / attributes of dataset, differential privacy
Accuracy	e.g. mechanisms for cross-checking attributes
Storage limitation	e.g. attach data retention periods to processing activities
Integrity and confidentiality	e.g. encryption, differential privacy, access controls
Accountability	e.g. transparent pipelines with logs

Building Careers Through Education
Everybody Matters.
Stronger Together.
Student, Learner and Client Centric.
Trust and Respect.
Embrace Change.
BPP

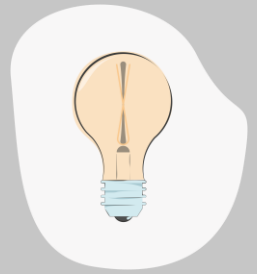
Ethical Big Data storage

Art. 25 GDPR Data protection by design and default…

- …implement appropriate technical and organisational measures
  - Such as pseudonymisation

- Which are designed to implement data-protection principles
  - Such as data minimization

- and to integrate the necessary safeguards into the processing in order to **protect the right of data subjects**



Data protection by design and by default requires you to assess risks and implement suitable mitigation for those risks BEFORE ANY PROCESSING HAS TAKEN PLACE!

Ethical Big Data storage

Applying the 7 data protection principles

This will enable your data subjects to exercise their:
Right to be informed
Right of access
Right of rectification
Right to erasue
Right to restrict processing
Right to data portability
Right to object

Rights related to opting out of automated decision-making including profiling.

Lawful basis

All processing needs to have an appropriate legal basis, either…

A. **Consent –** the individual agreed that you process their data for a stated purpose
B. **Contract –** the processing is necessary for a contract your have with the individual
C. **Legal obligation**
D. **Vital interests –** the processing is necessary to protect someone's life
E. **Public task –** the processing is necessary for you to perform a task in the public interest for your official functions, and the task or function has a clear basis in law
F. **Legitimate interests –** the processing is necessary for your legitimate interests or the legitimate interests of a third party, unless there is a good reason to protect the individual's personal data which overrides those legitimate interests

Data protection impact assessments

Applying the 7 data protection principles

Potential high-risk processing require extra measures (DPIA) because they deal with **sensitive personal information**, such as:

- Evaluation or scoring
- Automated decision-making with legal or significant effect
- Systematic monitoring of someone
- Sensitive data or data of a highly personal nature
- Data processed on a large scale
- Matching or combining vulnerable data subjects
- Data concerning vulnerable data subjects
- Innovative use or applying new technological or organisational solutions
- Preventing data subjects from exercising a right or using a service or contract

Data is power

What do we mean by this statement?

Having more information means you have more control

- Better at predicting and reacting to different outcomes

Those with information about us can have control over us

- Targeted marketing is the best example

Data ethics means that we need to pay special attention to how companies and governments use data about people (and other living beings).

- How is this information shared and with whom
- What is this information used for

Case study

Facial recognition…

Picture: Hong Kong protesters tear down a facial recognition lamp post (Getty images)

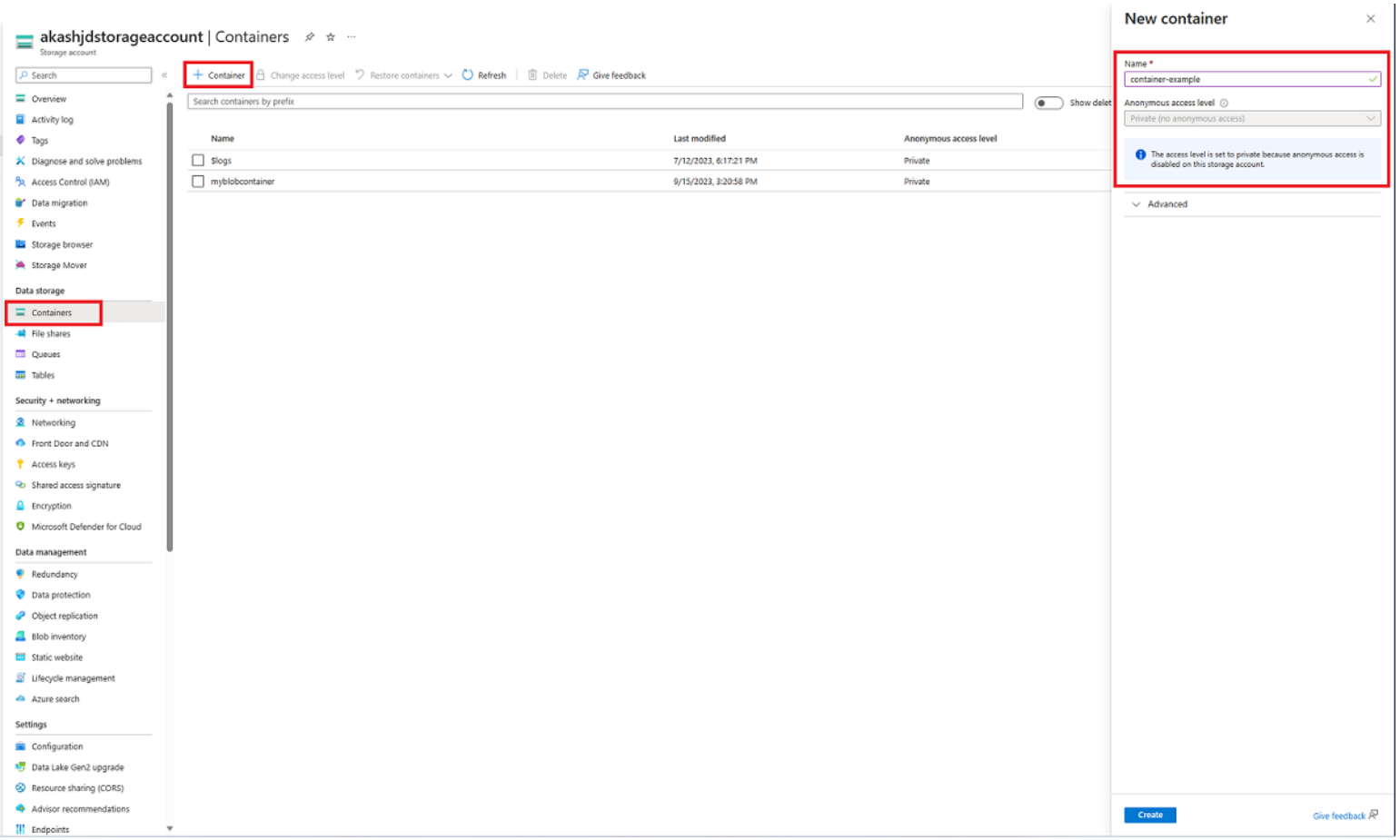
Data anonymisation techniques

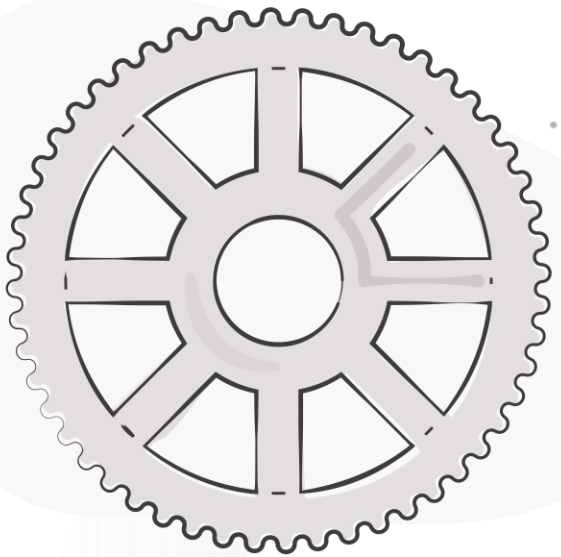
There are eight in total…

1. Attribute suppression
2. Record suppression
3. Character masking
4. Pseudonymisation
5. Generalisation
6. Permutation
7. Data perturbation
8. Data aggregation

Activity

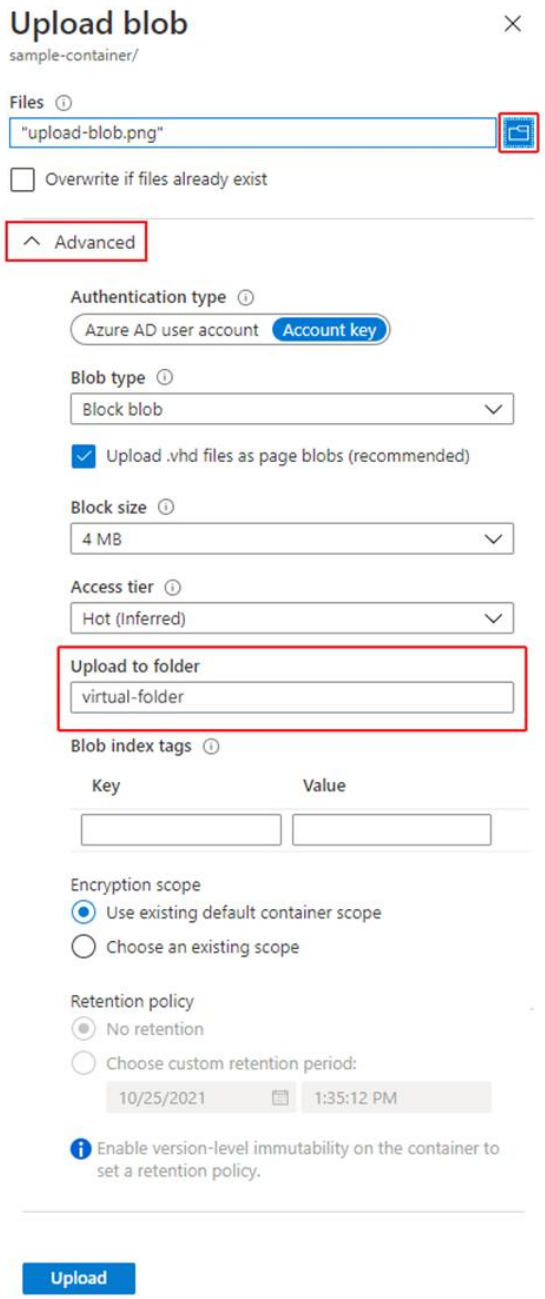
Azure data storage

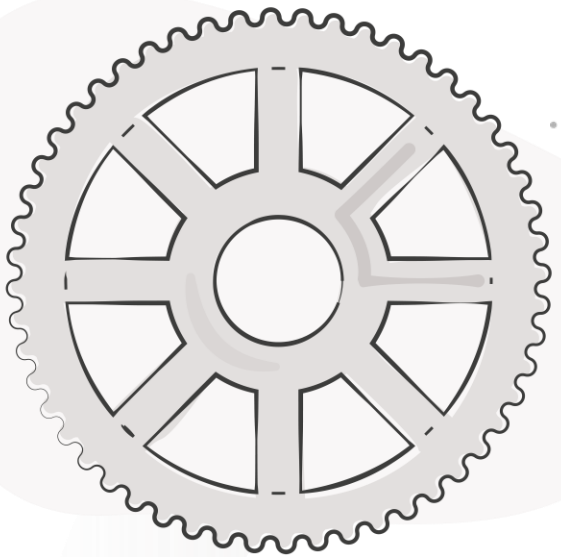




Activity

Azure data storage

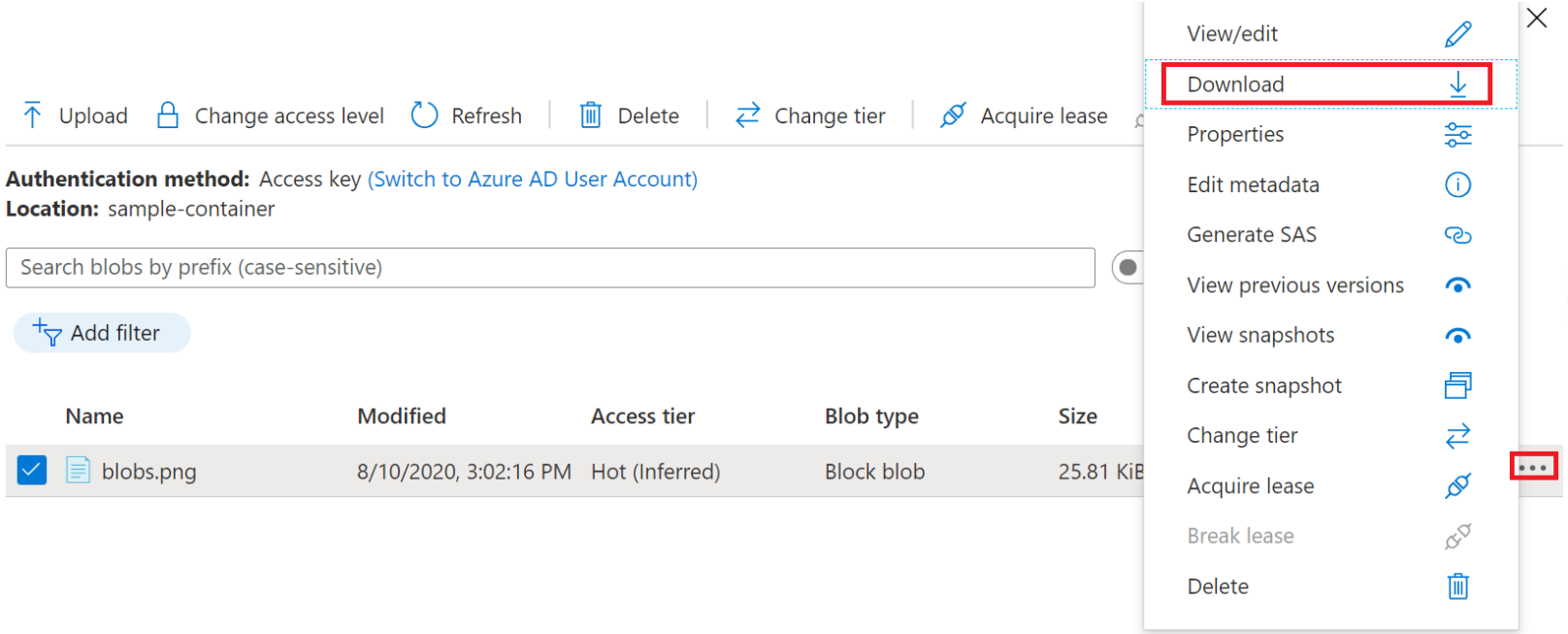


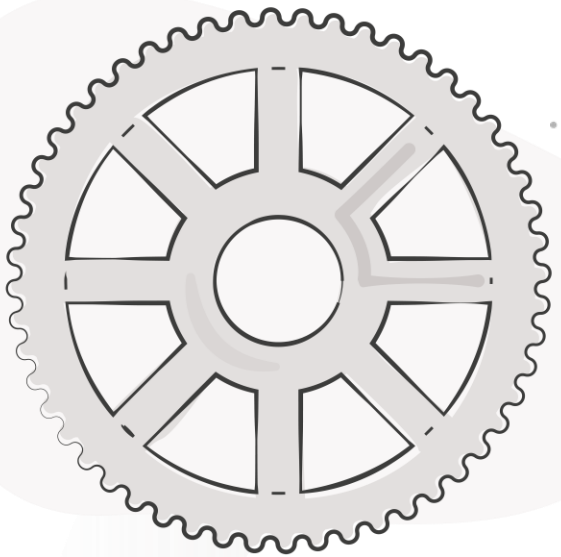


Building Careers Through Education
Everybody Matters.
Stronger Together.
Student, Learner and Client Centric.
Trust and Respect.
Embrace Change.
BPP

Activity

Azure data storage





Activity

Azure data storage

Dashboard > storagesamples >

sample-container
Container

Search (Ctrl+/)

- Overview
- Access Control (IAM)

Settings

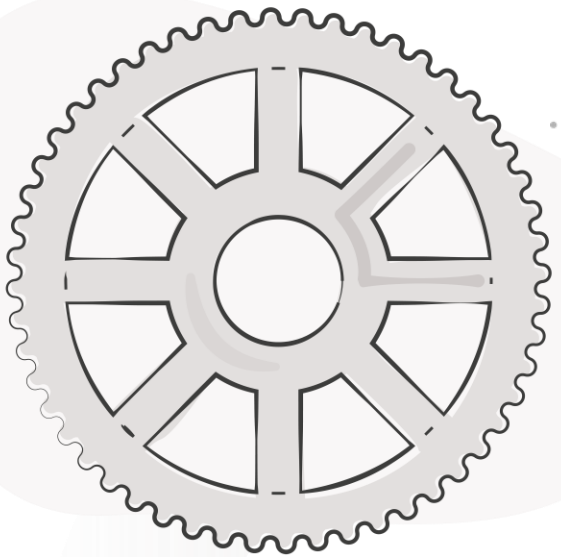
- Access policy
- Properties
- Metadata
- Editor (preview)

Upload | Change access level | Refresh | Delete | Change tier | Acquire lease | Break lease | ...

Authentication method: Azure AD User Account (Switch to Access key)
Location: sample-container

Search blobs by prefix (case-sensitive) | Show deleted blobs

Name	Modified	Access tier	Blob type	Size
level1				
blob1.txt	11/27/2019, 1:53:15 P...	Hot (Inferred)	Block blob	48 B
blob2.txt	3/11/2019, 10:18:04 P...	Hot (Inferred)	Block blob	20 B
blob3.txt	2/25/2019, 11:59:30 ...	Hot (Inferred)	Block blob	9 B
blob4.txt	2/25/2019, 12:01:16 P...	Hot (Inferred)	Block blob	9 B
blob5.txt	2/25/2019, 12:01:39 P...	Hot (Inferred)	Block blob	9 B
blob6.txt	2/25/2019, 12:06:56 P...	Hot (Inferred)	Block blob	9 B
blob7.txt	2/25/2019, 12:06:56 P...	Hot (Inferred)	Block blob	9 B
blob8.txt	2/25/2019, 12:06:55 P...	Hot (Inferred)	Block blob	9 B
blob9.txt	2/25/2019, 12:06:55 P...	Hot (Inferred)	Block blob	9 B
logfile.txt	9/19/2019, 4:06:49 PM	Hot (Inferred)	Block blob	13 B



Building Careers Through Education
Everybody Matters. Stronger Together. Student, Learner and Client Centric. Trust and Respect. Embrace Change.
BPP

Activity

Finishing the exercise...

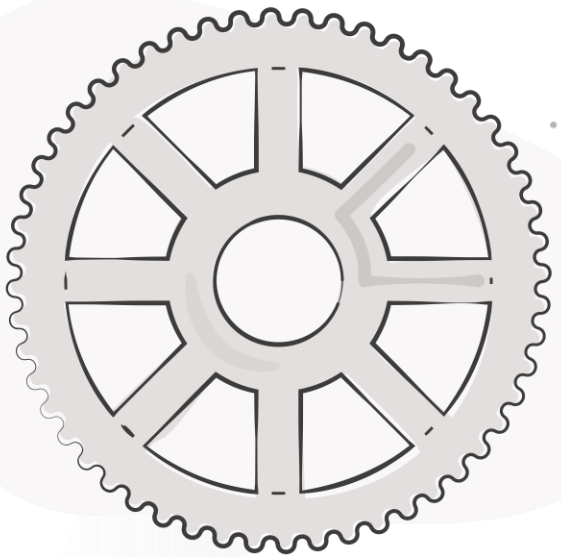
To delete the container:

1. In the Azure portal, navigate to the list of containers in your storage account.

2. Select the container to delete.

3. Select the **More** button (**...**), and select **Delete**.

4. Confirm that you want to delete the container.



Key Learning Summary

The key takeaways from this session are as follows:

- The transformative power of cloud computing illustrated by Netflix.
- Understanding and selecting appropriate compute models.
- Importance of calculating ROI for informed decision-making.
- Leveraging hyperscalers and edge computing for advanced solutions.
- Navigating cloud services and implementing robust IAM practices.

BPP

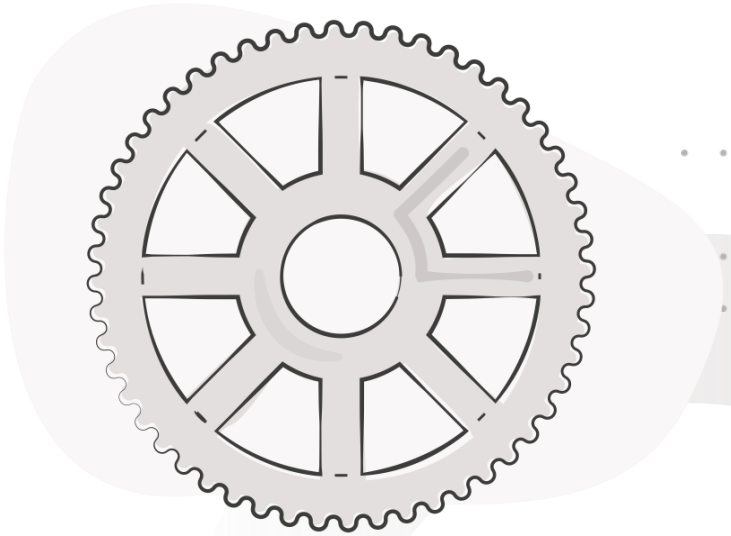
Post-webinar tasks

Apply…

- **Task 1:** Reflect on the article

https://blog.consoleconnect.com/the-truth-about-cloud-data-egress-fees

- **Task 2:** Write a reflection on comparing data access control methods in the cloud (see e-learning for details)



Thank you

Do you have any questions, comments, or feedback?