Data Technician

Name: Athiramol Padijarekudiyil Sukumaran

Course Date: 01 September - 04 September 2025

Table of contents

Day 1: Task 1

Please research and complete the below questions relating to key concepts of cloud.

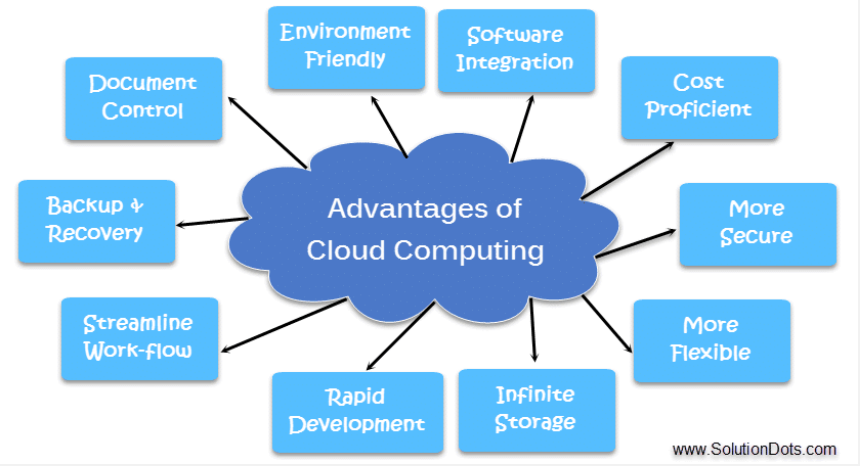
Be prepared to discuss the below in the group following this task.

What can cloud computing do for us in the real⦿world?

- Store and access data from anywhere.
- Run applications without needing powerful local computers.
- Process and analyze large amounts of data quickly.
- Host websites, apps, and services reliably.
- Provide backup and disaster recovery solutions.

Cloud computing can do a lot for us in the real world. Based on the examples from your image, here are some key things it enables:

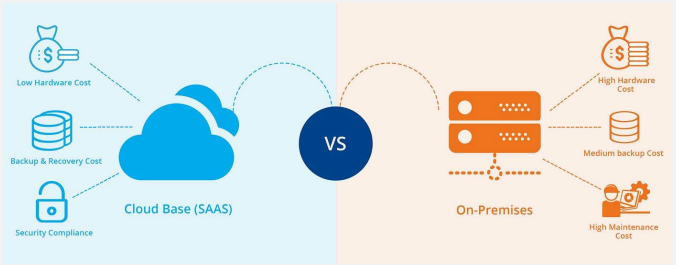
- **File Storage and Sharing** – Store files online and access them from anywhere (e.g., Google Drive, OneDrive, Dropbox).
- **Collaboration** – Work together in real time with teammates across the world (e.g., Microsoft 365, Google Workspace, Slack).
- **Customer Relationship Management (CRM)** – Manage customers, sales, and marketing efficiently (e.g., Salesforce, HubSpot).
- **Enterprise Resource Planning (ERP)** – Run and integrate business processes like finance, HR, and supply chain (e.g., SAP, Oracle Cloud ERP).
- **E⦿commerce** – Build and manage online stores without needing your own servers (e.g., Shopify, Magento, WooCommerce).
- **Streaming Services** – Deliver music, video, and entertainment instantly over the internet (e.g., Netflix, Spotify, Amazon Prime Video).

	• **Software as a Service (SaaS)** – Use software without installing it locally, paying only for what you need (e.g., Zoom, Adobe Creative Cloud, Trello). • **Development and Testing** – Developers can build, test, and deploy apps quickly on cloud platforms (e.g., GitHub, GitLab, Bitbucket). Cloud computing gives us **flexibility, cost savings, scalability, global access, and collaboration opportunities** across work, entertainment, and daily life.
How can it benefit a business?	• **Cost saving** – pay only for what you use, no big hardware costs. • **Scalability** – easily grow or shrink resources based on demand. • **Security** – built⦿in security, compliance, and backup features. • **Collaboration** – teams can work together from anywhere. • **Innovation** – access advanced AI, analytics, and automation tools. 

What's the alternative to cloud computing?

On⦿Premises Computing (using company⦿owned servers).

- The company buys and maintains physical servers.
- Full control, but expensive to maintain and less flexible.



What cloud providers can we use, what are their features and functions?

Amazon Web Services (AWS)

- Largest provider.
- Services: storage (S3), databases (RDS), compute (EC2), analytics (Athena, Redshift).

- Strength: scalability and wide range of services.

Microsoft Azure

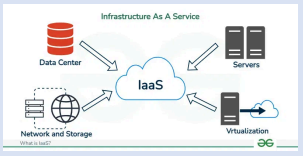
- Strong in enterprise/business integration.
- Services: Azure SQL Database, Azure Data Factory, Azure Machine Learning.
- Strength: works well with Microsoft tools (Excel, Power BI, Office 365).

Google Cloud Platform (GCP)

- Known for big data and AI.
- Services: BigQuery (analytics), Cloud Storage, AI/ML tools.
- Strength: strong in data science and machine learning.

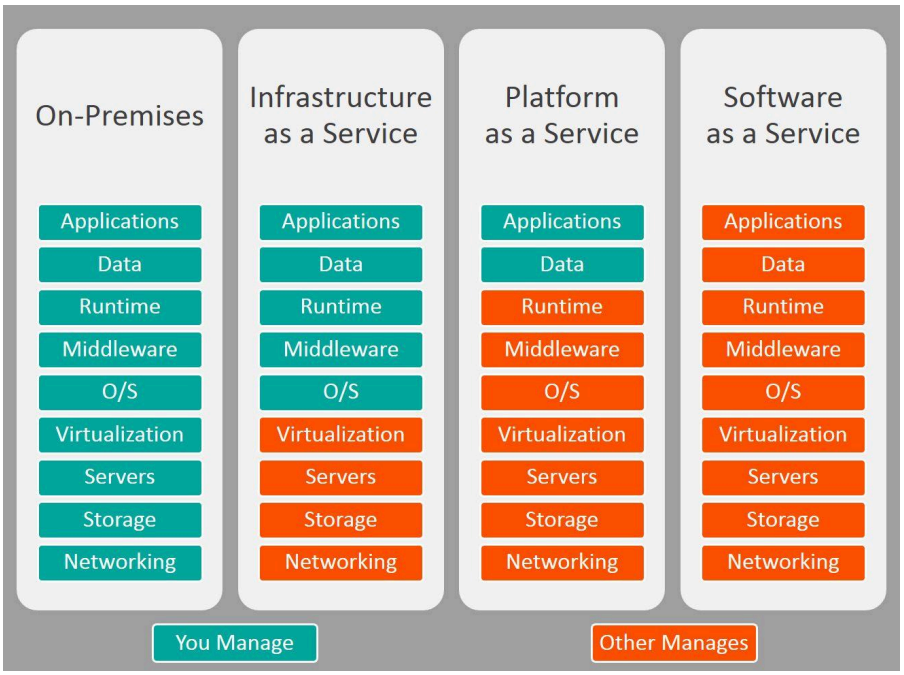
Day 1: Task 2

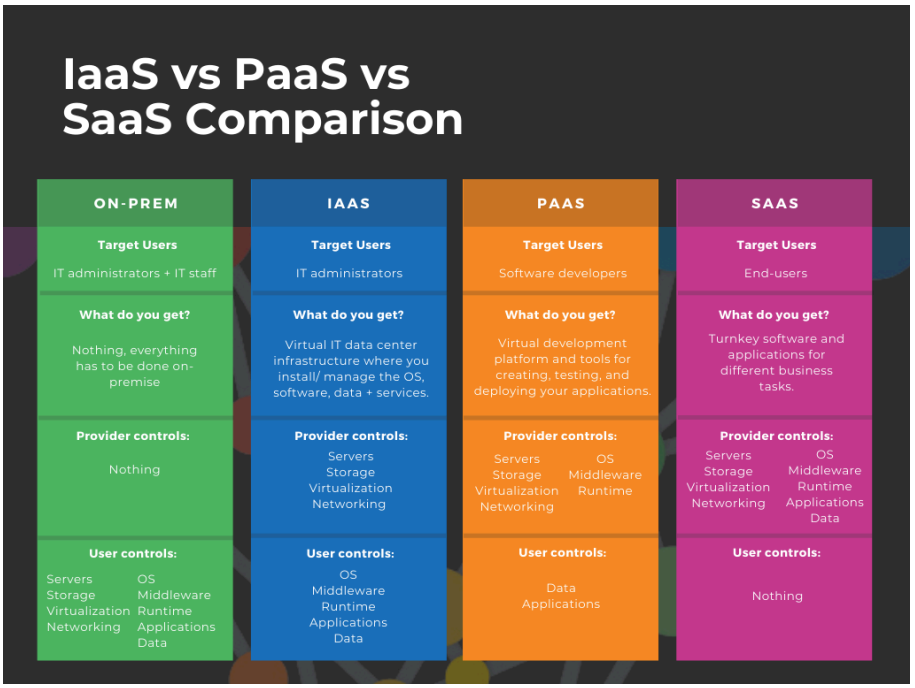
Please research the below cloud offerings, explain what they are and examples of use cases.

Cloud Offerings	Explain what it is	When / how might you use this service in the real⦿world?
IaaS (Infrastructure as a service)	⦿ A service model where the cloud provider delivers core infrastructure: virtual machines, networking, storage, and servers over the internet. ⦿ Users are responsible for installing and managing operating systems, middleware, and applications. ⦿ Offers scalability and flexibility without investing in physical hardware. 	⦿ **Startups & SMEs**: Quickly set up servers to launch websites or apps without buying costly hardware. ⦿ **IT departments**: Create test & development environments that can be spun up and deleted easily. ⦿ **Disaster Recovery**: Backup and replicate servers across regions for high availability. ⦿ **Big Data processing**: Spin up powerful servers temporarily for analytics workloads. **Examples** ⦿ **Microsoft Azure Virtual Machines**: Rent Windows/Linux servers. ⦿ **Amazon EC2 (Elastic Compute Cloud)**: Scalable virtual servers for apps. ⦿ **Google Compute Engine**: Custom VM instances on Google Cloud.
PaaS (Platform as a service)	⦿ A service model where the provider supplies a complete development & deployment platform, including infrastructure, OS, runtime,	⦿ **App Development**: Build and deploy web/mobile apps quickly without worrying about servers. ⦿ **API/Microservices**: Deploy small services that scale automatically with demand. ⦿ **Collaboration**: Multiple developers work on the same project in a managed environment.

	databases, and developer tools. ⦿ Developers focus only on writing code, while the provider handles scaling, updates, and security patches. ⦿ Speeds up development cycles and reduces management overhead. 	⦿ **Startups**: Reduce time⦿to⦿market by using pre⦿built environments. **Examples** ⦿ **Microsoft Azure App Service**: Host web apps and APIs easily. ⦿ **Google App Engine**: Auto⦿scaling apps without managing servers. ⦿ **AWS Elastic Beanstalk**: Deploy apps in Java, .NET, Python, PHP, Node.js with one click. ⦿ **Heroku**: Popular PaaS for startups and rapid prototyping.
SaaS (Software as a service)	⦿ A service model where fully functional software is delivered over the internet on a subscription basis. ⦿ Accessible via web browser or app, with no installation or maintenance required. ⦿ Updates, security, and infrastructure are managed by the provider.	⦿ **Business Communication**: Use tools like Slack or Zoom for team collaboration. ⦿ **Office Productivity**: MS Office 365, Google Workspace for word processing, spreadsheets, and email. ⦿ **Customer Relationship Management (CRM)**: Salesforce for managing leads and sales pipeline. ⦿ **Storage & File Sharing**: Dropbox, Google Drive for collaboration and backup. ⦿ **Everyday users**: Streaming services like Netflix (entertainment SaaS). **Examples** ⦿ **Zoom**: Cloud⦿based video conferencing. ⦿ **Microsoft 365 / Google Workspace**: Office productivity and collaboration tools.

	⦿ Highly scalable and cost⦿efficient. 	⦿ **Salesforce**: CRM solution. ⦿ **Dropbox**: Cloud storage and sharing. ⦿ **Netflix / Spotify**: SaaS entertainment platforms.

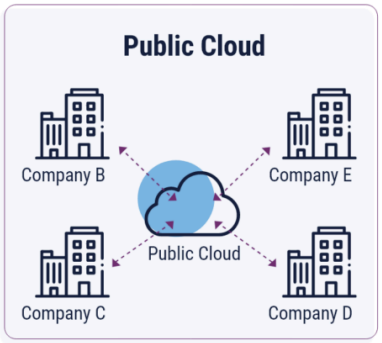




Day 1: Task 3

Please research the below terms and explain what they are, when they would be appropriate and a real-world example of where it could be implemented (i.e. what type of organisation).

Public Cloud



Definition

- A cloud environment owned and managed by third-party providers (e.g., AWS, Microsoft Azure, Google Cloud).
- Delivered over the internet and shared among multiple customers ("multi-tenant" model).

Characteristics

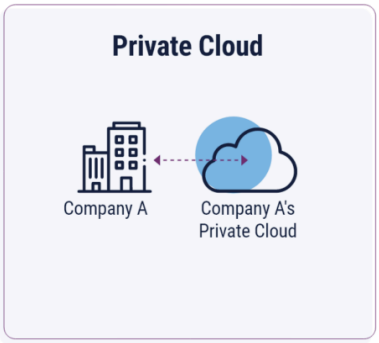
- Pay-as-you-go pricing.
- On-demand scalability and elasticity.
- Minimal management responsibility for the customer.
- Lower upfront cost compared to private setups.

Benefits

- Cost-effective for startups and SMEs.
- Global accessibility and high availability.
- Fast deployment without the need for physical infrastructure.

Limitations

- Less control over infrastructure.
- Potential security and compliance concerns.

	• Vendor lock-in risk. **When appropriate** • For businesses that need to scale quickly, handle variable workloads, or avoid heavy infrastructure investment. • Suitable for SaaS products, e-commerce platforms, and web applications. **Real-world example** • **Netflix** runs much of its global streaming services on AWS Public Cloud to handle massive amounts of traffic and deliver content worldwide.
Private Cloud	 **Definition** • A cloud environment dedicated solely to one organization. • Can be hosted on-premises (internal data center) or by a third-party vendor. **Characteristics** • Greater control and customization of resources. • Single-tenant model → only one organization uses the infrastructure. • Can integrate with existing on-premises IT setups. **Benefits** • Stronger security and compliance.

- Predictable performance since resources are not shared.
- Easier to enforce governance policies.

Limitations

- Higher upfront and ongoing costs.
- Requires more in-house expertise to manage.
- Less scalability compared to public cloud.

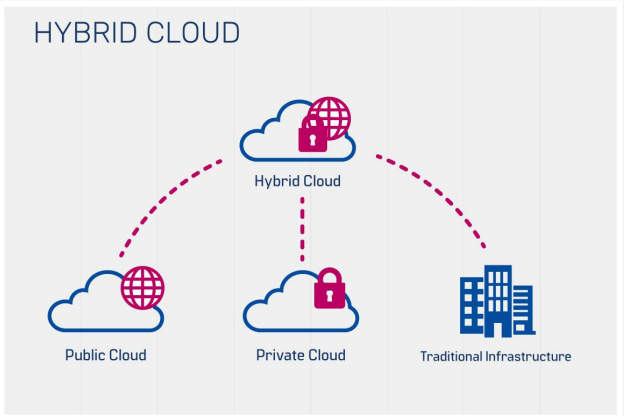
When appropriate

- For industries handling highly sensitive data (e.g., healthcare, finance, government).
- Organizations with strict compliance requirements (GDPR, HIPAA, PCI DSS).

Real-world example

- **JP Morgan Chase (Banking)** uses private cloud to process and store sensitive financial transactions while meeting strict regulatory standards.

Hybrid Cloud



Definition

- A mix of public and private clouds where data and applications can move between the two.
- Enables organizations to leverage the benefits of both models.

Characteristics

- Flexibility in resource usage (burst to public cloud during demand spikes).
- Supports disaster recovery and backup strategies.
- Balanced cost vs. control approach.

Benefits

- Optimizes cost efficiency while maintaining data security.
- Greater agility compared to pure private cloud.
- Useful for gradual migration from on-prem to cloud.

Limitations

- Complex to manage and integrate.
- Potential security risks when data moves between environments.
- Requires robust networking and orchestration.

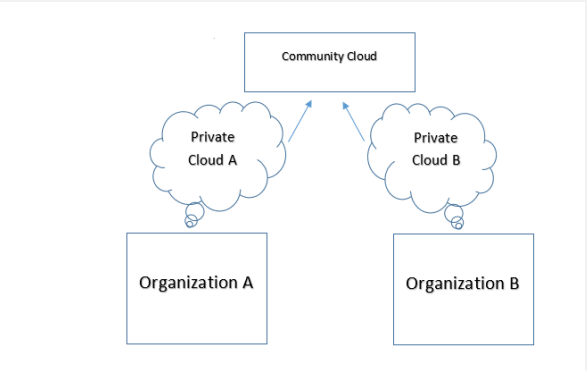
When appropriate

- For organizations needing to protect sensitive data but also benefit from scalable public cloud workloads.
- Ideal for companies with fluctuating workloads.

Real-world example

- **Walmart** uses hybrid cloud — keeping sensitive customer/financial data in private cloud while using public cloud for big data analytics and retail operations.

Community Cloud



Definition

- A shared cloud model where multiple organizations with common needs share the infrastructure.
- Can be managed internally or by a third-party vendor.

Characteristics

- Shared by organizations with similar regulations, missions, or security needs.
- Cost is distributed among members.
- Often industry-specific.

Benefits

- Cost-effective compared to private cloud (since shared).
- Collaborative platform for research, compliance, or government projects.
- Higher security than public cloud, as access is restricted.

Limitations

- Limited scalability compared to public cloud.
- Shared governance can create conflicts.
- Customization is limited since it must meet the needs of all members.

When appropriate

- For government bodies, universities, or industry sectors that need to share resources/data securely.
- Ideal for joint research projects, healthcare providers, or financial regulators.

Real-world example

- **European Union (EU) institutions** use community clouds to share data and applications securely among member states for governance and compliance.

Cloud Model	What it is	Benefits	Limitations	When to Use	Real-World Example
Public Cloud	Shared cloud owned by third-party (AWS, Azure, GCP)	Scalable, cost-effective, global access	Less control, compliance issues	Startups, SaaS apps, e-commerce	Netflix streaming on AWS
Private Cloud	Dedicated infrastructure for one org	Security, control, compliance	Expensive, less scalable	Finance, healthcare, gov	JP Morgan banking systems
Hybrid Cloud	Combination of public & private	Balance of cost, flexibility, disaster recovery	Complex to manage, security risks	Retail, enterprises with variable workloads	Walmart retail ops
Community Cloud	Shared by organizations with common needs	Shared cost, secure collaboration	Limited scalability, governance challenges	Government, universities, healthcare	EU institutions data sharing

Day 2: Task 1

Describe, with examples, the **three** major areas that the Computer Misuse Act deals with.

Area	Description	Example
Unauthorised Access to Computer Material	This is often called "hacking." It covers situations where a person **accesses a computer or network without permission**, even if no damage is done. The law recognises that simply breaching someone's system or data privacy is harmful. The offence occurs regardless of whether the hacker intends to steal, damage, or misuse the data.	A student guesses the password of a school computer system to access exam results without permission. Even if they only view the results and do not change anything, this is illegal.
Unauthorised Access with Intent to Commit or Facilitate a Crime	This is more serious than simple unauthorised access. Here, the hacker **accesses a computer without permission with the purpose of committing a further criminal offence**. This could include fraud, theft, or blackmail. The intent to commit a crime makes the act more severely punishable under the law.	Someone hacks into a company's payroll system intending to transfer money to their own bank account. The unauthorised access combined with the intent to steal is covered by this area.
Unauthorised Modification of Computer Material	This area criminalises **modifying, deleting, or adding data to a computer without authorisation**. It includes spreading viruses, malware, ransomware, or deleting important files. Even if no financial gain is intended, causing disruption or damage is illegal.	A disgruntled employee installs ransomware on the company network, encrypting files and demanding payment to restore access. Another example is deleting school records to disrupt administration.

The computer misuse act 1990 is an act where an individual can be criminalised because of computer related offense. Describe three extra powers that the Police and Justice Act 2006 (Computer Misuse) has added.

Description
1. Criminalising the Making, Supplying, or Obtaining of "Hacker Tools" This new offence (Section 3A of the CMA) specifically targets the tools used to commit computer crimes, such as malware, viruses, and hacking software. It makes it a criminal offense to: • **Make or adapt** any article (including programs or data) with the intent that it be used for a CMA offence. • **Supply or offer to supply** any article with the belief that it is likely to be used for a CMA offence. • **Obtain** any article with a view to its being supplied for use in a CMA offence. **Example:** A person who develops and sells a malicious program designed to bypass security systems for others to use in hacking would be committing an offence under this section.
2. Extending Unauthorised Acts to Include Recklessness The Act updated the offence under Section 3 of the CMA, which deals with actions that impair a computer's operation. The original law required an "intent" to impair, but the 2006 Act expanded this to also include **recklessness**. This means a person can be prosecuted if they perform an unauthorised act and are simply reckless as to whether it will impair the computer's operation. **Example:** Someone who launches a Denial of Service (DoS) attack, even if their primary intent isn't to cause damage, can be prosecuted if they are reckless as to whether the attack would hinder the operation of the targeted computer system.
3. Increasing Sentencing Powers The Police and Justice Act 2006 significantly increased the maximum prison sentences for some of the key CMA offences. This was done to reflect the growing seriousness and financial impact of computer-related crimes. For instance, the maximum sentence for an offence under the new Section 3 (Unauthorised Acts) was increased from five years to ten years. **Example:** Before the 2006 Act, a person found guilty of a serious hacking and data impairment offence might have received a five-year sentence, but after the changes, they could be sentenced for up to ten years.

Look at the below website to answer the questions:
https://www.gov.uk/personal-data-my-employer-can-keep-about-me

Write down three items of data which a company can store about an employee.
• Name, address, and date of birth
• Education and qualifications
• Emergency contact details

Give three more examples of data that an employer can only store if they first get the employee's permission.
• Race and ethnicity
• Religious or philosophical beliefs
• Trade union membership

Conduct further research to answer the below questions.

Question	Answer
Provide one example of: Copyright infringement	**Anthropic Settles Lawsuit Over Pirated Books Used to Train AI** In August 2025, artificial intelligence company Anthropic settled a class action lawsuit filed by U.S. authors who alleged that Anthropic used pirated books to train its AI assistant, Claude. The authors claimed that Anthropic had downloaded up to seven million pirated books, potentially exposing the company to billions in damages. The settlement marked the first major resolution among ongoing AI-related copyright disputes and highlighted the legal complexities surrounding the use of copyrighted material in AI training.
Provide one example of: Plagiarism	**Influencer Plagiarizes Personal Story on TikTok** In August 2025, journalist Esme Hewitt accused an influencer of plagiarizing her deeply personal story about being adopted during China's one-child policy. The influencer shared a TikTok video repeating entire sections of Hewitt's article verbatim without crediting her. Despite Hewitt's call-out, the influencer's husband dismissed the

	claims and accused her of stealing. The incident sparked a public debate about plagiarism on social media platforms and the need for proper attribution.
What are two consequences of copyright infringement and software piracy?	• **Legal Penalties:** Individuals found guilty of copyright infringement or software piracy may face substantial fines and imprisonment. For example, in the United States, civil penalties can reach up to $150,000 per infringement, and criminal fines can be as high as $250,000, with potential prison sentences of up to five years. • **Reputational Damage:** Being involved in copyright infringement or software piracy can severely damage an individual's or company's reputation. Public exposure of such activities can lead to loss of trust among customers, partners, and the general public, impacting future business opportunities and personal credibility.
Give three possible consequences for individuals when using pirated software	• **Exposure to Malware and Viruses:** Pirated software often lacks proper security updates and may contain malicious code, leading to potential data breaches and system compromises. • **Legal Repercussions:** Using pirated software constitutes a violation of copyright laws and can result in legal actions, including fines and imprisonment. • **Lack of Support and Updates:** Users of pirated software do not have access to official customer support or software updates, which can lead to unresolved technical issues and security vulnerabilities.

Listed below are some laws which we have covered today:

1. Computer Misuse Act 1990
2. Police and Justice Act 2006 (Computer Misuse)
3. Copyright, Designs and Patents Act 1988
4. Copyright (Computer Programs) Regulations 1992
5. The Health and Safety (Display Screen Equipment) Regulations 1992
6. Data Protection Act 2018
7. Consumer Rights Act 2015

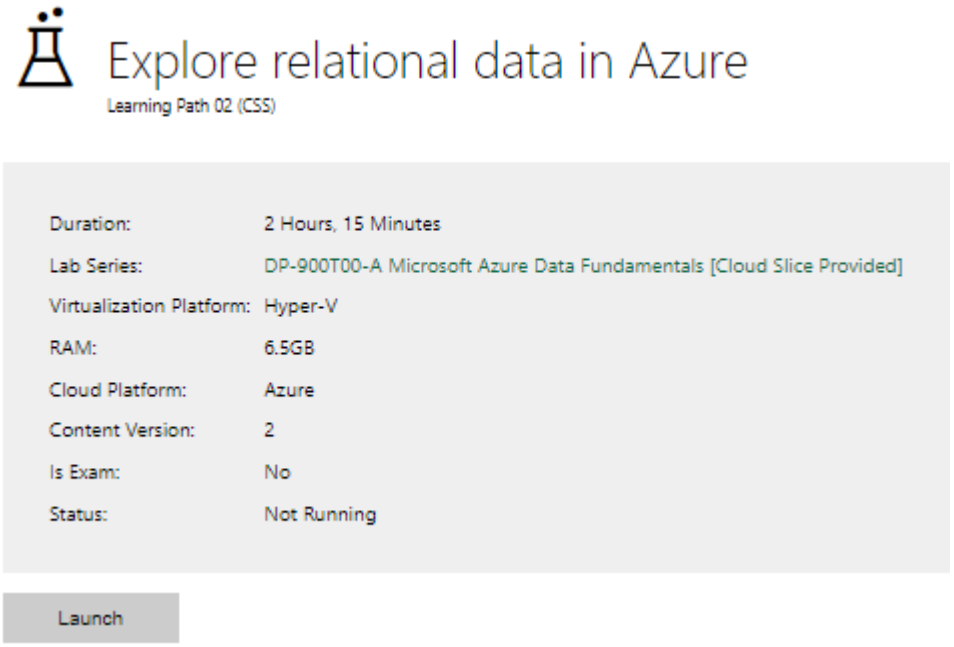
- Insert a number in the first column of each row to match each of the statements with one of the above Acts.

- One of the statements is incorrect and not illegal. For this statement, write 'Not illegal'.

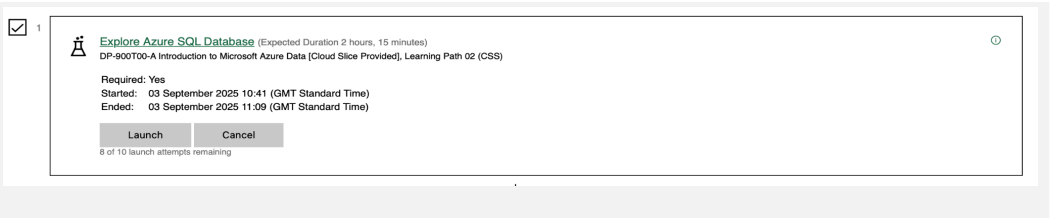
Act number	Clause
	With some exceptions, it is illegal to use unlicensed software
	Any product, digital or otherwise, must be fit for the purpose it is supplied for
	Unauthorised modification of computer material is illegal
	It is illegal to create or use a hacking tool for penetration testing
	Personal data may only be used for specified, explicit purposes
	Employers must provide their computer users with adequate health and safety training for any workstation they work at
	It is illegal to distribute hacking tools for criminal purposes
	It is illegal to distribute an illicit recording
	Personal data may not be kept longer than necessary
	Gaining unauthorised access to a computer system is illegal
	Employers must ensure that employees take regular and adequate breaks from looking at their screens
	It is illegal to prevent or hinder access (e.g. by a denial⦿of⦿service attack) to any program or data held in any computer
	Personal data must be accurate and where necessary kept up to date

Day 3: Task 1

Please complete the below lab (3) *'Explore relational data in Azure'* and paste evidence of the completed lab in the box provided.

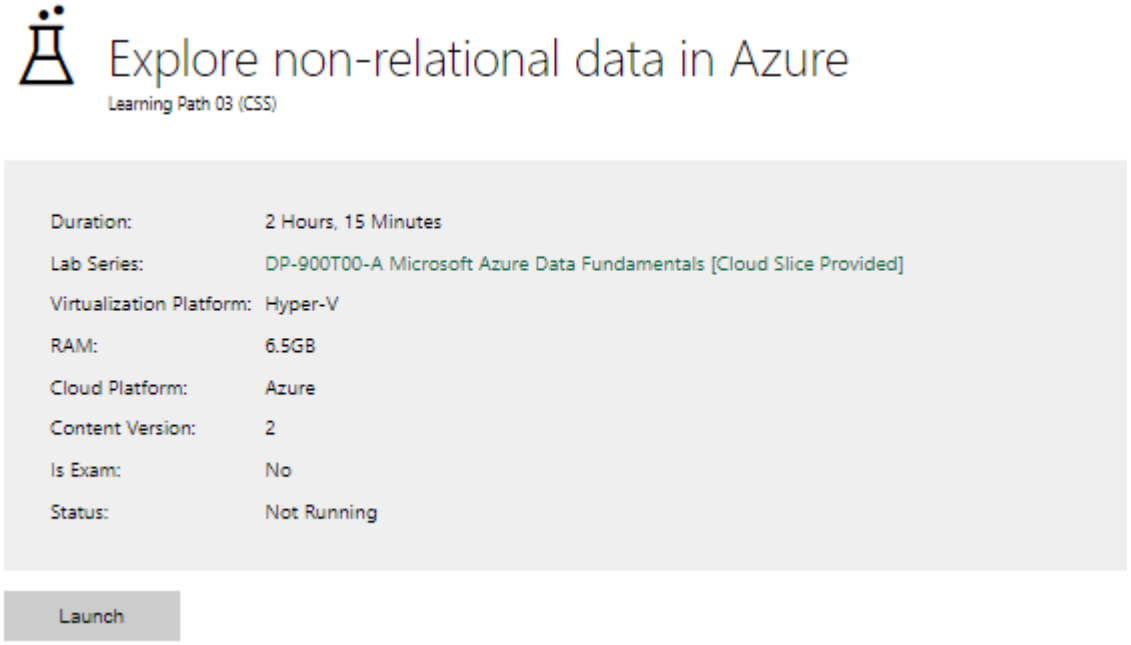


Completed lab



Day 3: Task 2

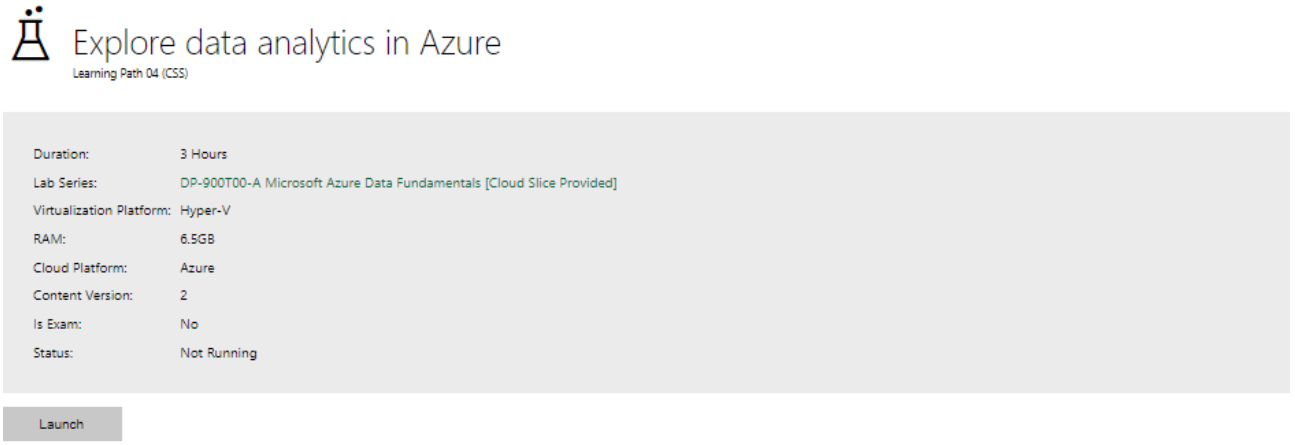
Please complete the below lab (4) *'Explore non⦿relational data in Azure'* and paste evidence of the completed lab in the box provided.



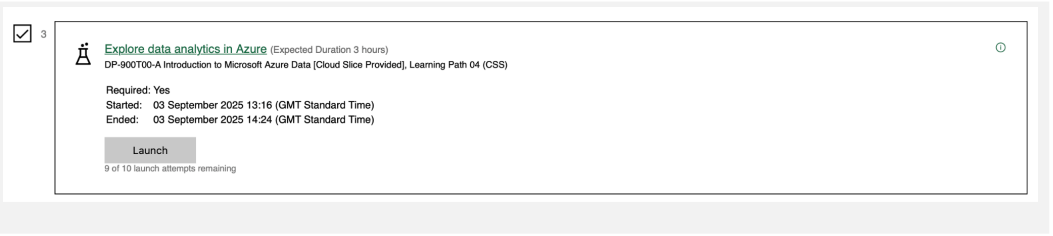
Completed lab

Day 3: Task 3

Please complete the below lab (5) 'Explore data analytics in Azure' and paste evidence of the completed lab in the box provided.



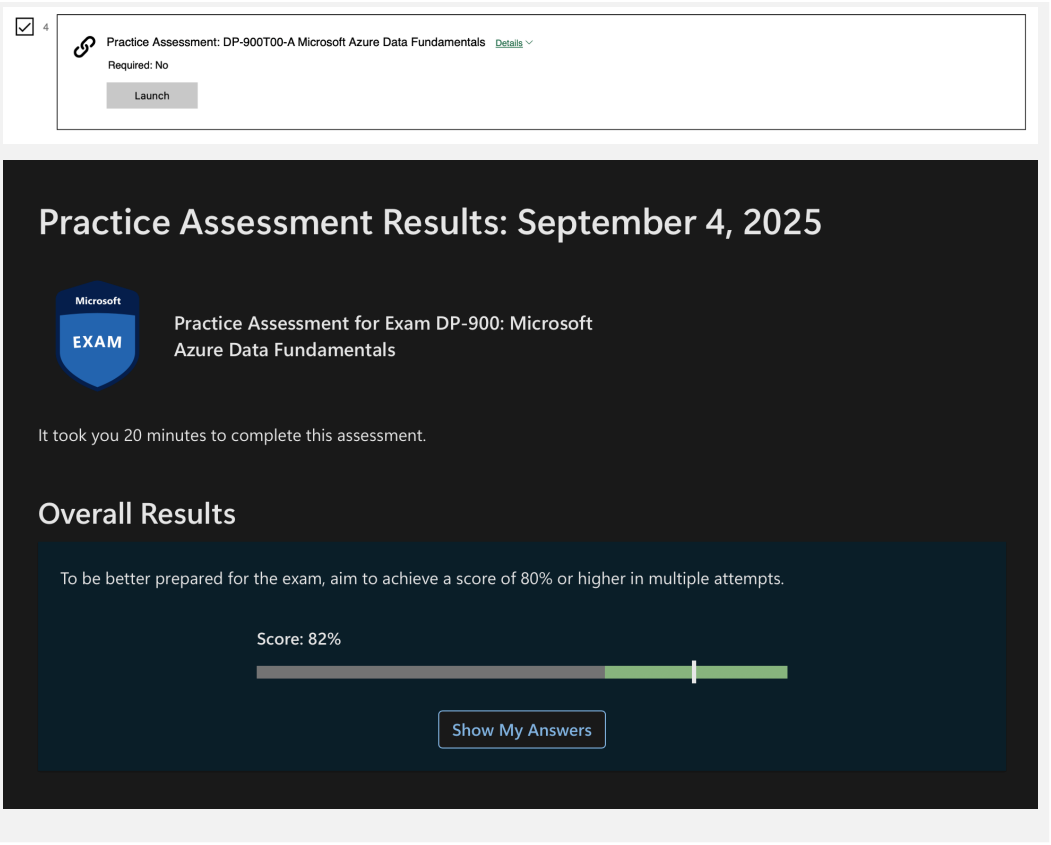
Completed lab



Day 4: Task 1

In your teams, complete the Azure DP-900 practice exam and paste your result below – this is an open book and please research and discuss your answers as a team.

Result



Day 4: Task 2

1. Scenario Background

"Paws & Whiskers" is a growing pet shop that aims to improve its business by analysing sales, customer information, and inventory data. Currently, the data is collected manually or stored in spreadsheets. Management is interested in transitioning to Microsoft Azure to streamline data storage, analysis, and reporting, enabling them to make data⦿driven decisions.

2. Data Laws and Regulations

Identify and explain the data laws and regulations relevant to handling customer data within the proposal. Ensure you cover the following points:

- **GDPR Compliance**: Highlight the importance of adhering to the General Data Protection Regulation (GDPR), particularly as it relates to storing and processing customer information.
- **Data Protection Act (DPA) 2018**: Outline how the DPA 2018 may affect the way "Paws & Whiskers" collects and stores data, ensuring compliance with UK laws on data privacy.

- **Other Industry Standards**: Research any additional data protection standards or regulations that may apply to pet shop data, particularly if they involve sensitive or payment information.

3. Azure Service Recommendations

Recommend Microsoft Azure services that would suit the company's data analysis needs and explain why these services are suitable. Your recommendations should include:

- **Data Storage**: Identify suitable storage options, such as **Azure Blob Storage** or **Azure SQL Database**, and discuss the benefits of each for storing large datasets, including inventory, sales transactions, and customer details.
- **Data Analysis Tools**: Recommend tools such as **Azure Machine Learning** for customer behaviour analysis or **Azure Synapse Analytics** for analysing sales trends.
- **Data Integration and Automation**: Explain how services like **Azure Data Factory** could automate data collection and integration processes, improving efficiency.

4. Data Types and Data Modelling

Define the types of data "Paws & Whiskers" will need to work with and describe your approach to data modelling:

- **Data Categories**: Identify key data types, such as customer demographics, transaction history, pet inventory, and product categories.
- **Data Modelling Approach**: Outline how you would structure this data using a relational model or a data warehouse approach, considering factors like tables, entities, relationships, and primary keys.

5. Data Storage Formats and Structures in Azure

Discuss how you would store data within Azure and the formats you would recommend:

- **Data Formats**: Specify recommended formats (e.g., CSV for raw data imports, JSON for structured data, Parquet for analytics) and explain why these formats are suitable for specific data types.
- **Data Security and Encryption**: Include recommendations for securing data using Azure's built-in encryption features and access controls to ensure compliance with data privacy regulations.

6. Additional Considerations

Provide any other considerations that might enhance data handling and efficiency in Azure, such as:

- **Backup and Disaster Recovery**: Outline a backup plan using **Azure Backup** or **Azure Site Recovery** to safeguard against data loss.
- **Data Visualisation**: Discuss potential use of **Power BI** within Azure for creating dashboards that provide management with real-time insights into sales and customer trends.
- **Future Scalability**: Comment on how Azure services can scale as the business grows, accommodating larger datasets and more complex analyses.

Submission Guidelines:

1. **Structure**: Ensure your report is well-organised, with sections for each task (e.g., Data Laws, Azure Services, Data Types, etc.).
2. **Formatting**: Include headings, bullet points where appropriate, and any visuals or diagrams that support your explanations.
3. **References**: Cite any resources or regulations referenced in the report.
4. **Length**: Aim for 1500-2000 words.

Data Management and Analysis in Azure – Paws & Whiskers

Course Notes

It is recommended to take notes from the course, use the space below to do so, or use the revision guide shared with the class:

- Day 1

What is Cloud Computing?

- On-demand delivery of IT resources over the internet with **pay-as-you-go pricing**.
- Eliminates the need to buy, own, and maintain physical servers and data centers.
- Examples:
  - **File Storage**: Google Drive, OneDrive, Dropbox
  - **Collaboration**: MS 365, Google Workspace
  - **CRM**: Salesforce, HubSpot
  - **ERP**: SAP, Oracle Cloud
  - **SaaS Apps**: Zoom, Trello, Adobe Creative Cloud

Cloud in Data

- **Scalability** – expand resources easily for large datasets.
- **Collaboration** – real-time teamwork from different locations.
- **Data Storage** – secure handling of structured & unstructured data.
- **Processing Power** – high-performance computing.
- **Advanced Analytics/AI** – pre-built ML & AI services.
- **Real-time Analytics** – up-to-the-minute insights.

- **Security & Compliance** – built-in protocols & certifications.
- **Backup & Recovery** – automated protection against data loss.
- **Integration** – seamless with BI tools like Power BI, Tableau.

Benefits of Cloud

- Agility
- Elasticity (scaling up & down)
- Cost savings
- Global deployment
- High availability

Cloud Service Models

1. **IaaS** (Infrastructure as a Service)
    - Virtual machines, storage, networking.
    - Flexible, highest control.
2. **PaaS** (Platform as a Service)
    - Pre-built environments with OS, databases.
    - Developers focus on building apps.
3. **SaaS** (Software as a Service)
    - Fully managed applications (e.g., Gmail, Office 365).
    - No infrastructure management.

Cloud Deployment Models

- **Public Cloud** – Infrastructure owned by providers (Azure, AWS, GCP). Multi-tenant, scalable, pay-as-you-go.
- **Private Cloud** – Dedicated to one organization. More control but higher cost.
- **Hybrid Cloud** – Mix of public & private, combining flexibility with control.

Comparison Factors:

- Ownership of infrastructure
- Scalability (unlimited in public cloud, limited in private)
- Maintenance (done by provider in public, by organization in private)
- Accessibility (public = open to anyone, private = restricted)
- Costs (public = operational expense, private = capital + operational).

Azure Costing

- Pricing depends on **usage and services**.
- Use Azure Calculator to estimate expenses.
- Public cloud can be more cost-effective for scaling but may get expensive

with long-term heavy use.

- Day 2

Cybersecurity & Legislation

- **Cyber Crime (UK 2024 stats):**
  - 45% of businesses reported cyber-attacks.
  - Average cost of breach: **£4.88 million**.
  - Ransomware attacks doubled.
  - Phishing attacks are increasing.
- **Key Security Terms:**
  - Anti-Virus, DoS/DDoS, Malware, Phishing, Keyloggers, Ransomware, VPN.
- **How to Protect as a Technician:**
  - Strong access controls (MFA).
  - Regular backups & recovery tests.
  - Encrypt data (in transit & at rest).
  - Monitor & log usage.
  - Secure data disposal.
  - Follow GDPR/Data Protection Act.
  - Staff training on best practices.

Current Legislation

- **Computer Misuse Act (1990):**
  - Illegal: unauthorised access, modification, hacking.
- **Police & Justice Act (2006):**
  - Strengthened CMA 1990.
  - Made hacking tools illegal.
  - DoS & DDoS made criminal offences (up to 10 years prison).
- **Consumer Rights Act (2015):**
  - Products must be of **satisfactory quality**, **fit for purpose**, and **as described**.
- **Data Protection Act (1998 & 2018) / GDPR:**
  - Ethical use of personal data.
  - Security of personal data.
  - GDPR (2018) replaced outdated laws to adapt to technology.

Microsoft Azure Overview

- **What is Azure?**

    - Microsoft's **cloud computing platform** for creating, deploying, and managing applications.
- **Azure Services Categories:**
    - Compute, Networking, Storage, Databases, AI/ML, Analytics, Security, IoT, DevOps, Governance.

Azure Storage Types

- **Blobs:** Scalable object storage (text, binary, big data, analytics). Types: Block, Append, Page.
- **Files:** Managed file shares (supports containers, cloud & on-premises).
- **Queues:** Message storage & processing between components.
- **Tables:** NoSQL storage for structured but non-relational data.
- **Disks:** Managed block storage for VMs.

Microsoft Cloud Services for Data

- **Data Stores:**
    - Azure SQL (relational DB).
    - Azure Database for MySQL, PostgreSQL, MariaDB.
    - Cosmos DB (global-scale NoSQL).
    - Azure Storage (files, blobs, tables, data lake).
- **Data Engineering & Analytics:**
    - Azure Data Factory (pipelines).
    - Synapse Analytics (end-to-end analytics).
    - Databricks (Apache Spark).
    - HDInsight (open-source big data).
    - Stream Analytics (real-time IoT data).
    - Data Explorer (real-time log/telemetry).
    - Purview (data governance).
    - Power BI (analytics & visualization).

Azure Resource Manager (ARM) Terms

- **Resource:** Individual service (VM, DB, etc.).
- **Resource Group:** Container for related resources.
- **Resource Provider:** Service offering (e.g., Microsoft.Storage).

- Day 3

Azure Storage Recap

- **Blob Storage:**
  - Stores binary large objects (BLOBs).
  - Types:
    - **Block blobs** – up to 4.7 TB, best for large files.
    - **Page blobs** – up to 8 TB, used for VM disks.
    - **Append blobs** – optimized for log data, up to 195 GB.
  - Tiers: **Hot** (fast, costly), **Cool** (cheaper, slower), **Archive** (lowest cost, highest latency).
- **Data Lake Gen2:**
  - Built on Blob storage with hierarchical namespace.
  - Combines Blob + Data Lake Gen1 for analytics.
  - Supports file/directory-level access control.
- **Azure Files:**
  - Cloud-based file shares.
  - Supports **SMB** & **NFS (premium)**.
  - Redundant & encrypted.
- **Azure Table Storage:**
  - Key-Value NoSQL store.
  - Rows identified by **PartitionKey** + **RowKey**.
  - Schema-less, scalable storage for structured data.

Azure Cosmos DB

- **Multi-model NoSQL database** with global scalability.
- Features:
  - Fast reads/writes.
  - Multi-region replication.
  - Multiple APIs supported:
    - Cosmos DB for **NoSQL**
    - **MongoDB** API
    - **PostgreSQL** API
    - **Table** API (key-value)
    - **Apache Cassandra** API
    - **Gremlin** API (graph database).

Large-Scale Data Analytics

- **Elements:**
  - Data ingestion & processing (ETL/ELT, batch + real-time).
  - Analytical data stores (data lakes, warehouses).
  - Analytical models (semantic models, cubes).
  - Visualization (reports, dashboards).
- **Processing Engines:**
  - **Relational DBs** (SQL).

    - **Apache Spark** (distributed, multi-language).
- **Architectures:**
    - **Data Warehouse** – SQL-based, optimized schemas (star/snowflake).
    - **Data Lakehouse** – combines data lakes + relational access.

Azure Data Analytics Services

- **PaaS Analytics:**
    - **Azure Synapse Analytics** – unified SQL, Spark, pipelines.
    - **Azure Databricks** – scalable Spark + ML workflows.
    - **Azure HDInsight** – Apache frameworks (Hadoop, Kafka, HBase, Spark).
- **SaaS Analytics (Microsoft Fabric):**
    - Unified platform with **OneLake** storage.
    - Includes Data Factory, Synapse, Power BI, Data Activator.

Streaming & Real-Time Analytics

- **Batch vs Stream Processing:**
    - Batch = periodic processing.
    - Stream = real-time event handling.
- **Azure Stream Analytics:**
    - Ingests from **Event Hubs, IoT Hub, Blob Storage**.
    - Processes with continuous queries.
    - Outputs to **Blob, SQL, Synapse, Power BI, Functions**.
- **Azure Data Explorer (KQL):**
    - Optimized for telemetry & time-series analysis.
    - Supports log queries.
- **Fabric Real-Time Analytics:**
    - Capture event streams.
    - Store in Lakehouse/KQL DB.
    - Real-time visualizations via SQL/KQL.

- Lab - Explore Azure SQL Database

Lab Objective

Provision an Azure SQL Database with the **AdventureWorks sample data**, then connect and query it using the **Query Editor** in the Azure portal.

Step 1: Provision Azure SQL Database

1. In the **Azure portal**, select **+ Create a resource** → **Azure SQL** → **SQL databases** → **Create**.
2. Fill in:
    - **Subscription:** Your lab subscription
    - **Resource group:** `RG1`
    - **Database name:** `AdventureWorks`
    - **Server:** Create new → `sqlserver54328897` (location: `UK South`)
        - Auth: SQL authentication
        - Username: *your name*
        - Password: *complex password (remember this!)*
    - **Elastic pool?** → No
    - **Workload environment:** Development
    - **Backup storage:** Locally-redundant
3. **Networking tab** → Public endpoint → Allow Azure services + current client IP.
4. **Security tab** → Microsoft Defender for SQL → Not now.
5. **Additional settings** → Use existing data: **Sample**.
6. Select **Review + Create** → **Create**.
7. Wait for deployment, then go to the resource.

Step 2: Connect with Query Editor

1. From the SQL Database resource → left menu → **Query editor (preview)**.
2. Sign in with the **SQL admin login + password** you set earlier.
3. If blocked by IP → click **Allowlist IP** in the error message.

Step 3: Explore Data with SQL Queries

Run a simple select:

```
SELECT * FROM SalesLT.Product;
```

1. → Shows all columns from the Product table.

Run a filtered select:

```
SELECT ProductID, Name, ListPrice, ProductCategoryID
FROM SalesLT.Product;
```

2. → Returns only key product details.

Run a JOIN query:

```
SELECT
    p.ProductID,
    p.Name AS ProductName,
```


```
    c.Name AS Category,
    p.ListPrice
FROM SalesLT.Product AS p
INNER JOIN SalesLT.ProductCategory AS c
    ON p.ProductCategoryID = c.ProductCategoryID;
```

3. → Combines product info with category names.

- Lab- Explore Non- Relational data in Azure

Lab Objective-1

Provision an **Azure Storage account** and explore its **non-relational storage options**:

- Blob storage
- Data Lake Storage Gen2
- Azure Files
- Azure Tables

Step 1: Provision an Azure Storage Account

1. In **Azure portal** → **+ Create a resource** → **Storage account** → **Create**.
2. Fill in:
    - **Subscription:** your lab subscription
    - **Resource group:** create new (e.g., `RG1`)
    - **Storage account name:** unique lowercase name (e.g., `storage12345`)
    - **Region:** same region as lab resources
    - **Performance:** Standard
    - **Redundancy:** Locally-redundant (LRS)
3. **Advanced tab:** leave hierarchical namespace unchecked.
4. **Data protection tab:** deselect all soft delete options.
5. Select **Review + Create** → **Create**.

Step 2: Explore Blob Storage

1. In the storage account → **Containers** → **+ Add container**.

- Name: `data`
- Access level: **Private**

2. Download product1.json.
3. Go to **Storage browser** → **Blob containers** → **data** → **+ Add Directory** → name `products`.
   - Note: Folders are **virtual** until a blob exists inside.
4. Upload `product1.json` into a **folder named** **`product_data`**.
   - Confirms virtual path creation.

Step 3: Explore Data Lake Storage Gen2

1. Download product2.json.
2. In storage account → **Settings** → **Data Lake Gen2 upgrade** → enable hierarchical namespace.
   - This makes folders **real** and manageable.
3. Navigate back to **data** → **product_data**.
   - The old `product1.json` is still there.
4. Upload `product2.json`.
5. Now folder-level options (rename, permissions) are available.

Step 4: Explore Azure Files

1. In storage account → **File shares** → **+ File share**.
   - Name: `files`
   - Tier: Transaction optimized
   - Backup: Disabled
2. Open the share → select **Connect**.
   - View OS-specific mount scripts (Windows, Linux, macOS).

Step 5: Explore Azure Tables

1. In storage account → **Tables** → **+ Table**.
   - Name: `products`
2. In **Storage browser** → **Tables** → **products** → **+ Add entity**.
   - Entity 1:
     - PartitionKey = 1
     - RowKey = 1
     - Name = Widget
     - Price = 2.99
   - Entity 2:
     - PartitionKey = 1
     - RowKey = 2
     - Name = Kniknak

        - Price = 1.99
        - Discontinued = true
3. Verify rows are added → Timestamp column auto-created.

Step 6: Cleanup

- Delete the **resource group** to remove all resources and avoid charges.

Key Takeaways

- **Blob Storage:** stores unstructured data (virtual folders only).
- **Data Lake Gen2:** hierarchical folders + ACLs, best for analytics.
- **Azure Files:** SMB/NFS file shares, good for lift-and-shift.
- **Azure Tables:** NoSQL key-value store, schema-less and cheap.

Lab 2: Explore Azure Cosmos DB (NoSQL)

Goal: Create a Cosmos DB account, add JSON items, and query them.

Step 1: Create a Cosmos DB Account

1. In the **Azure portal** → **+ Create a resource** → search **Azure Cosmos DB** → select **Azure Cosmos DB for NoSQL** → **Create**.
2. Fill in:
    - **Workload type:** Learning
    - **Subscription:** Concierge Subscription (sandbox)
    - **Resource group:** use the existing sandbox RG (e.g., `learn-xxxx`)
    - **Account name:** unique (e.g., `cosmosdb12345`)
    - **Availability Zones:** Disable
    - **Location:** recommended closest region
    - **Capacity mode:** Provisioned throughput
    - **Free-tier discount:** Apply (if available)
    - **Limit throughput:** leave unselected
3. **Review + Create** → **Create**.
    - Wait for deployment, then go to the resource.

Step 2: Create Sample Database

1. In your new **Cosmos DB account** → **left panel** → **Data Explorer**.
2. Select **Launch quick start**.
3. A **SampleDB** (database) and **SampleContainer** (container) are created automatically with sample items.

Step 3: View & Add Items

1. Expand **SampleDB** → **SampleContainer** → **Items**.
    - Browse the JSON items already there.
2. To add your own:
    - Click **New Item**.

Paste this JSON:

```
{
  "name": "Road Helmet,45",
  "id": "123456789",
  "categoryID": "123456789",
  "SKU": "AB-1234-56",
  "description": "The product called \"Road Helmet,45\" ",
  "price": 48.74
}
```

    -
    - Select **Save**.
    - Cosmos DB will automatically add system fields like `_rid`, `_ts`, `_etag`.

Step 4: Query the Database

1. In **Data Explorer** → **New SQL Query**.

Default query:

```
SELECT * FROM c
```

2. → Runs and shows **all items**.

Modify and run:

```
SELECT *
FROM c
WHERE CONTAINS(c.name, "Helmet")
```

3. → Returns only items with `"Helmet"` in the name.

Step 5: Cleanup

- When done, **delete the resource group** (sandbox will also auto-expire).

Key Learnings

- Cosmos DB stores **non-relational JSON** documents.
- No rigid schema → you can add new fields anytime.
- Querying uses **SQL-like syntax** (`SELECT`, `WHERE`, `CONTAINS`).
- System fields (`_rid`, `_etag`, `_ts`) are automatically added for tracking & performance.

- Lab- Explore data analytics in Azure

Lab Objective-1

Practice data ingestion, querying, and analytics in **Microsoft Fabric Lakehouse** using the NYC Taxi dataset.

Key outcomes:

- Understand Fabric Lakehouse concepts.
- Ingest data using pipelines.
- Query data with SQL.
- Manage resources responsibly.

Step 1: Create a Workspace

1. Go to Microsoft Fabric and sign in with your account.
2. Click **Workspaces** · in the left menu.
3. Select **Create Workspace** → give it a unique name.
4. Choose a capacity with **Fabric enabled** to allow data engineering tasks.
5. The workspace opens empty.

Tip: Keep labs isolated by using a dedicated workspace.

Step 2: Create a Lakehouse

1. In your workspace, select **Create** → **Lakehouse** (under Data Engineering).
2. Name your lakehouse uniquely.
3. After creation, explore:
    - **Tables** → managed Delta tables, SQL-ready.
    - **Files** → raw or staged data files in OneLake.

Tip: Use **Files** for raw data; **Tables** for curated datasets.

Step 3: Ingest Data Using Pipelines

1. In the lakehouse, select **Get data** → **New data pipeline** → name it `Ingest Data`.
2. **Copy Data wizard:**
    - **Source:** Sample data → NYC Taxi - Green dataset.
    - **Destination:**
        - Root folder: Tables
        - Load to: New table
        - Table name: `taxi_rides`
        - Column mappings: leave default
        - Partitioning: off
3. Select **Start data transfer immediately** → **Save + Run**.
4. Monitor the pipeline in the **Output pane**. Wait until it succeeds (dataset: 75M+ rows, ~2.5 GB).

Tip: Refresh **Tables** node in Lakehouse explorer to see `taxi_rides`.

Step 4: Query Data in SQL

1. Switch to **SQL analytics endpoint** in the lakehouse.
2. Click **New SQL query**.
3. Example query:

```
SELECT DATENAME(dw, lpepPickupDatetime) AS Day,
       AVG(tripDistance) AS AvgDistance
FROM taxi_rides
GROUP BY DATENAME(dw, lpepPickupDatetime);
```

4. Run to see **average trip distance by day of the week**.

Tip: Lakehouse tables are SQL-friendly; you can run queries without moving data elsewhere.

Lab Objective-2

Practice real-time data ingestion, querying, and analytics using **Microsoft Fabric Real-Time Intelligence** with a taxi dataset.

Key outcomes:

- Create a workspace with Fabric capacity.
- Ingest real-time streaming data using an eventstream.
- Store and query data in an eventhouse (KQL database).
- Analyze real-time data and visualize trends.
- Clean up resources efficiently.

Step 1: Create a Workspace

1. Go to Microsoft Fabric and sign in.
2. Click **Workspaces** · in the left menu → **Create Workspace**.
3. Give it a unique name and select **Fabric-enabled capacity** (Trial, Premium, or Fabric).
4. The workspace opens empty.

Tip: Keep lab resources isolated by using a separate workspace.

Step 2: Create an Eventstream

1. Select **Real-Time hub** in the left menu.
2. Under **Data sources**, connect to the **Yellow taxi sample**.
3. Name the source `taxi` and the eventstream `taxi-data`. The default stream will be `taxi-data-stream`.
4. Wait for creation and select **Open eventstream** to see it on the design canvas.

Tip: Eventstreams ingest real-time data and can include transformations before sending data to destinations.

Step 3: Create an Eventhouse

1. In the left menu, select **Create** → **Eventhouse** (under Real-Time Intelligence).
2. Name the eventhouse uniquely.
3. The eventhouse contains a **KQL database** where you can create tables.
4. Select the database → **Get data** → **Eventstream** → **Existing eventstream**.
5. Create a new table named `taxi` and connect it to the `taxi-data` eventstream as `taxi-table`.
6. Finish the configuration and verify that the eventstream shows the

destination table.

Tip: The eventhouse provides durable storage and allows querying with Kusto Query Language (KQL).

Step 4: Query the Captured Data

1. In the KQL database, open the **queryset**.
2. Quick sample query to validate data:

```
taxi
| take 100
```

3. Query the number of taxi pickups per hour:

```
taxi
| summarize PickupCount = count() by
bin(todatetime(tpep_pickup_datetime), 1h)
```

4. Re-run queries periodically to see real-time data updates.

Tip: `bin(..., 1h)` groups events into hourly intervals for trend analysis.

Lab Objective-3

Learn how to:

- Import and model data from multiple sources in Power BI Desktop.
- Create relationships, hierarchies, and categories for deeper analysis.
- Build and format interactive reports with tables, charts, and maps.
- Explore data insights using drill-down and cross-highlighting.

Step 1: Install Power BI Desktop

1. Download from Power BI Desktop installer.
2. Run the setup wizard and complete the installation.
3. Open Power BI Desktop.

Step 2: Import Data

Use **Get Data** → **Web** three times with these sample CSV files:

- **Customers** → `https://github.com/MicrosoftLearning/DP-900T00A-Azure-Data-Fundamentals/raw/master/power-bi/customers.csv`
- **Products** → `https://github.com/MicrosoftLearning/DP-900T00A-Azure-Data-Fundamentals/raw/master/power-bi/products.csv`
- **Orders** → `https://github.com/MicrosoftLearning/DP-900T00A-Azure-Data-Fundamentals/raw/master/power-bi/orders.csv`

Now I have 3 related tables (Customers, Products, Orders).

Step 3: Refine the Data Model

1. Go to the **Model tab** to view the tables.
2. In **Orders**: Set *Revenue* → **Format = Currency**.
3. In **Products**:
    - Right-click *Category* → **Create hierarchy**.
    - Add *ProductName* to hierarchy → rename it **Categorized Product**.
4. In **Customers**: Set *City* column → **Data Category = City**.

Relationships, hierarchies, and categories prepare your data for better visuals.

Step 4: Create a Report

1. **Enable map visuals**: File → Options → Security → check *Use Map and Filled Map visuals*.
2. In **Report view**:
    - Add a **Text Box** → title: *Sales Report* (Bold, size 32).

Visuals to Build:

- **Column Chart (Revenue by Category)**
    - Add *Categorized Product* + *Revenue*.
    - Turn on **Drill-down** → click a category to see *Product* level.
- **Pie Chart (Quantity by Category)**
    - Add *Quantity* (Orders) + *Category* (Products).
- **Map (Revenue by City)**
    - Add *City* (Customers) + *Revenue* (Orders).
    - Interact with map → selecting a city cross-highlights other visuals.

Step 5: Save & Share

1. Save report as a **.pbix** file.
2. (Optional) Publish to **Power BI Service** → collaborate, refresh data, and share with others.

- Day 4

Data Stores

Data Warehouse

- Relational database for analytics (not transactions).
- Uses **fact tables** (numeric metrics/events) + **dimension tables** (time, location, products).
- Optimized for queries, aggregations, and reporting.

Data Lake

- Stores all data types (structured, semi-structured, unstructured).
- **Schema-on-read**: structure is applied only when data is queried.
- Handles diverse formats (text, images, videos) using tools like **Hadoop & Spark**.

Hybrid Approach

- Combines strengths of data warehouses & data lakes.

Azure Services (Analytical Stores)

- **IoT Hub**: Connects and manages IoT devices (supports HTTPS, AMQP, MQTT).
- **Azure Synapse Analytics**: Data warehouse + big data analytics.
- **Azure Cosmos DB**: Global, scalable NoSQL database.
- **Azure Stream Analytics**: Real-time stream processing.

Data Modelling

- **Purpose**: Organize data into tables, define relationships, and support analysis.
- **Fact Tables**: Store measures (e.g., sales revenue, quantity).
- **Dimension Tables**: Store descriptive attributes (e.g., customer names, product categories).

- **Schemas**:
  - **Star schema**: Simple, fact table linked directly to dimensions.
  - **Snowflake schema**: More complex, with dimensions linked to sub-dimensions.

We have included a range of additional links to further resources and information that you may find useful, these can be found within your revision guide.

END OF WORKBOOK

Please check through your work thoroughly before submitting and update the table of contents if required.

Please send your completed work booklet to your trainer.