**Ansh Ranjan  
Azure Data  
  
Exercise 1 – Azure Storage Options for Data**

**TASK 1 and 3: Creating a new Azure Storage Account and Upload a dataset**

1. In Storage Accounts click on Create
2. Select subscription and resource group and give it a name and pick redundancy type   
   A screenshot of a computer

   AI-generated content may be incorrect.
3. Once created, to upload a file go to your storage account > Storage Browser > Blob Containers > Add container > Enter a name > Create
4. Then navigate to your container > Upload > Browser and select the file > click Upload  
   A screenshot of a computer

   AI-generated content may be incorrect.
5. Now your will see your dataset file in the container  
   A screenshot of a computer

   AI-generated content may be incorrect.

**TASK 2: Explore difference between Blob Storage, File Storage, Queue Storage, Table Storage**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Feature** | **Blob Storage** | **File Storage** | **Queue Storage** | **Table Storage** |
| **Data Type** | Unstructured | Structured (files) | Messages | Structured (key-value pairs) |
| **Structure** | Blobs in containers | Files in directories | Messages in queues | Entities in tables |
| **Access Protocol** | HTTP/HTTPS | SMB/NFS | HTTP/HTTPS | HTTP/HTTPS |
| **Use Cases** | Media, backups, analytics | File shares, migrations | Messaging, workflows | NoSQL database, IoT |
| **Scalability** | Highly scalable | Up to 100 TiB per share | Millions of messages | Petabytes of data |
| **Pricing** | Size and access tier | Provisioned capacity and tier | Operations and data transfer | Data stored and operations |

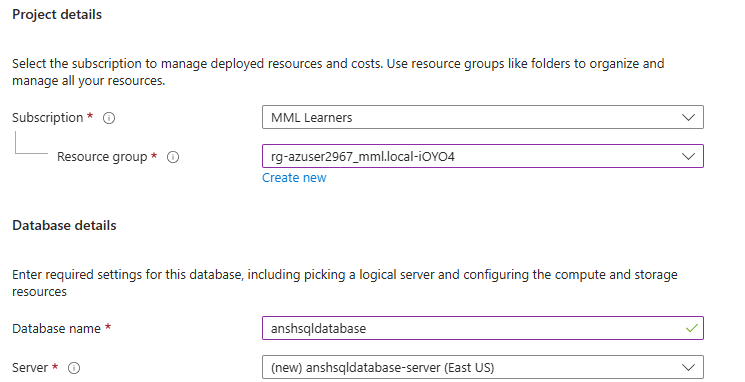
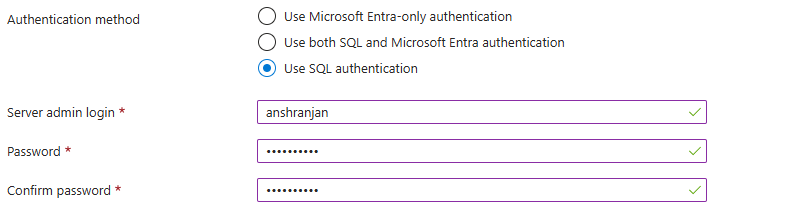
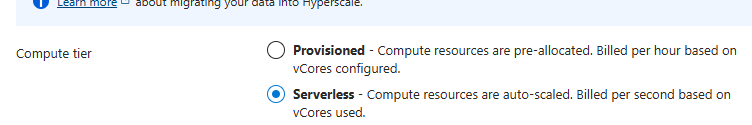
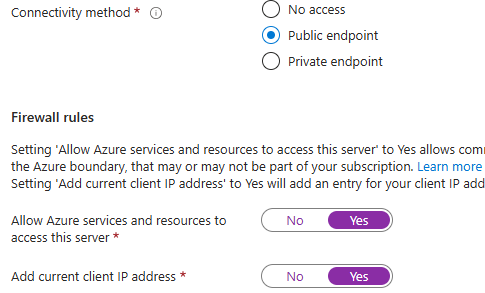
Each storage service in Azure is optimized for specific scenarios:

* Use **Blob Storage** for unstructured data like media files and backups.
* Use **File Storage** for shared file systems and legacy applications.
* Use **Queue Storage** for asynchronous messaging between components.
* Use **Table Storage** for structured NoSQL data with high scalability and performance.

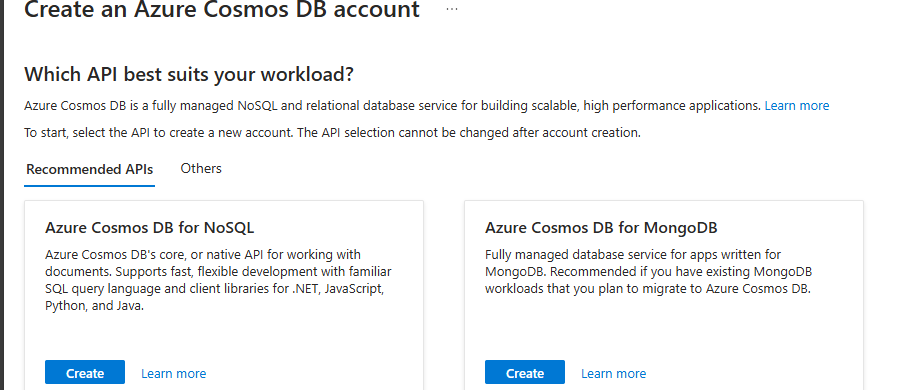
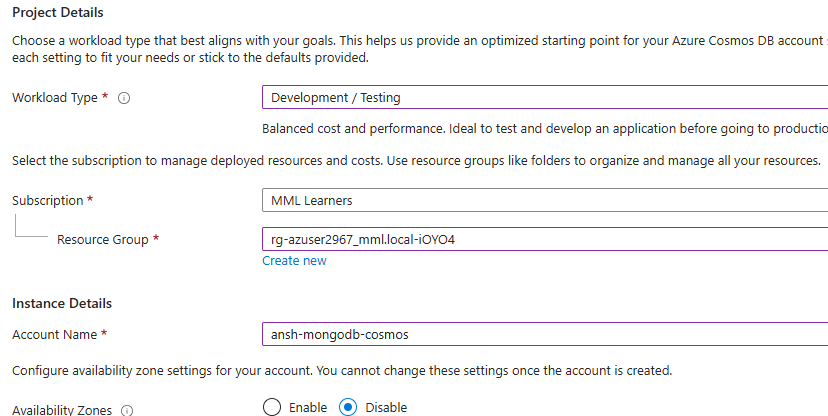
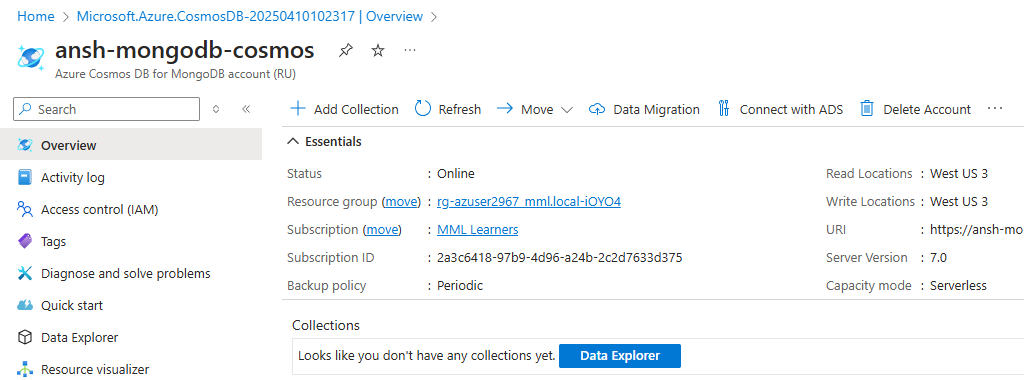
**EXERCISE 2 - Introduction to Azure Databases**

**TASK 1: Deploy a sample database in Azure Cosmos DB and Azure SQL Database**

* **SQL DATABASE**

1. Go to SQL Databases > Create > Enter details > Create a new database server if you do not have existing option.    
   
2. Under Network tab set Admin login and Password for your SQL database  
   
3. Picking serverless computer for cheaper computation   
   
4. Under Network settings, allow connection to database   
   

* **COSMOS DB MONGO DB**

1. Open Cosmos DB > Create > Cozmos DB for MongoDB   
   
2. Request Unit database account > Enter details > Review and Create   
   
3. Your Azure Cosmos DB API for Mongo DB will be created and running   
   

**TASK 2: Document key features and use cases for each**

**Azure Synapse Analytics**

**Key Features**:

* Unified platform for big data and data warehousing.
* Massively Parallel Processing (MPP) for large datasets.
* Integrated pipelines for ETL/ELT with Azure Data Factory.
* Synapse Studio for data exploration and analytics.
* Scalable, secure, and supports machine learning.

**Use Cases**:

* Data warehousing and big data analytics.
* Business intelligence with Power BI integration.
* Advanced analytics and IoT data processing.

**Azure SQL Database**

**Key Features**:

* Fully managed relational database service.
* High availability, scalability, and automated maintenance.
* Elastic pools for resource sharing.
* Advanced security and geo-replication.
* Seamless integration with other Azure services.

**Use Cases**:

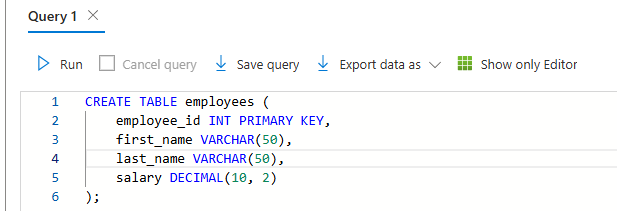
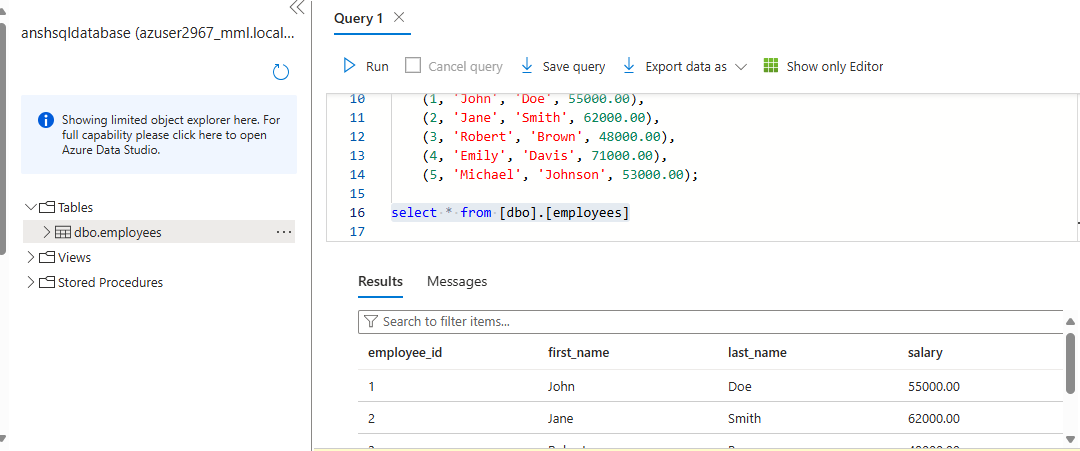
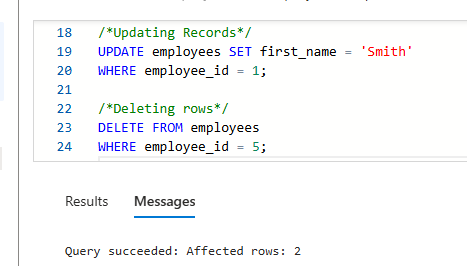
* Transactional workloads (OLTP).
* Backend for web/mobile apps and e-commerce systems.
* ERP/CRM databases and lightweight analytics.

**Comparison of Use Cases**

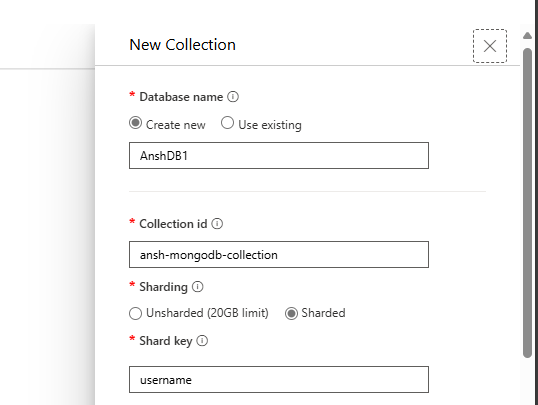
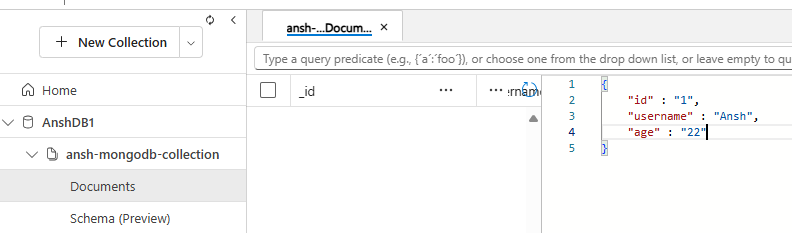
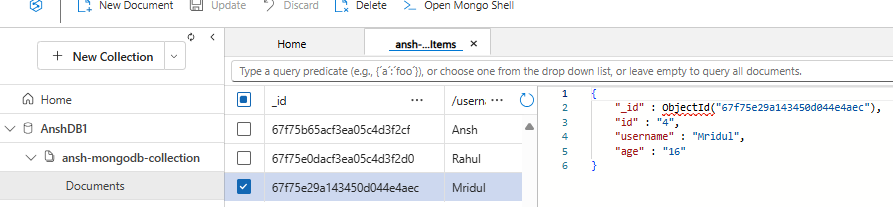
|  |  |  |
| --- | --- | --- |
| **Feature/Use Case** | **Azure Synapse Analytics** | **Azure SQL Database** |
| **Primary Focus** | Analytical workloads (OLAP) | Transactional workloads (OLTP) |
| **Data Volume** | Petabytes of data | Gigabytes to terabytes of data |
| **Scalability** | Massively parallel processing | Elastic scaling for transactional data |
| **Integration** | Big data tools, Power BI, Data Lake | Web apps, mobile apps, and business apps |
| **Machine Learning** | Advanced analytics and AI workloads | Limited to lightweight analytics |
| **Use Case Examples** | Data warehousing, predictive analytics | E-commerce systems, ERP/CRM databases |

**TASK 3: Perform basic CRUD operations**

* **SQL DATABASE**

1. Go to your DB > Query Editor > Login with admin ID and password
2. You will be presented with query page. Write a query to create a table in your database   
   
3. Insert records into your database table and read them   
   
4. Updating and Deleting records   
   

* **COSMOS DB MONGO DB**

1. Go to Data Explorer > Create Database > New Collection > Select Database or create new, give collection id, shard key   
   
2. Once your collection is created > click New Document and enter data in document > Save   
   
3. You can add more documents now that your mongo db is up and running   
   

**EXERCISE 3 - Data Security and Compliance in Azure**

**1. Research and Document Azure's Data Encryption Methods**

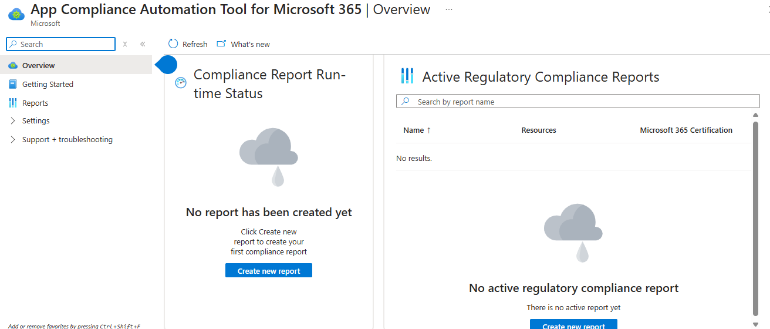
* **What to do**:
* Research how Azure secures data through encryption. Include details on encryption at rest, encryption in transit, and customer-managed keys.
* Cover services like Azure Storage encryption, SQL Database encryption (TDE), and Azure Key Vault.
* Summarize your findings in **1 page** and save it as a **PDF document**.
* **How to do it**:
* Use Azure documentation (<https://learn.microsoft.com/en-us/azure/>) for research.
* Tools like Word or Google Docs can help you write and export your document as a PDF.

**2. Enable Encryption for a Sample Azure Storage Account**

* **What to do**:
* Create or use an existing Azure Storage Account.
* Enable encryption for the storage account (Azure encrypts all storage accounts at rest by default).
* Optionally, configure **customer-managed keys** for encryption using Azure Key Vault.
* **How to do it**:
* Log in to the Azure Portal ([https://portal.azure.com](https://portal.azure.com/)).
* Navigate to your storage account.
* Under **Settings**, go to **Encryption** and verify or enable encryption.
* If using customer-managed keys, integrate Azure Key Vault.

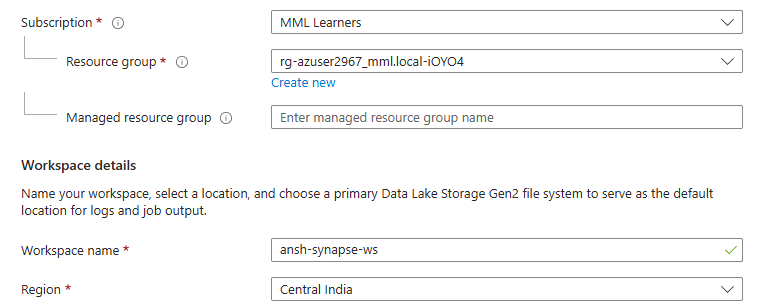
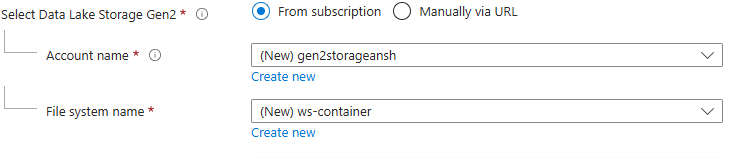
**3. Explore Azure Compliance Manager for Data Regulation**

* **What to do**:
* Use Azure Compliance Manager to explore how Azure helps meet regulatory requirements (e.g., GDPR, HIPAA).
* Review the compliance score and understand how to improve it.
* **How to do it**:
* Log in to the Azure Portal.
* Search for **Compliance Manager** in the portal.
* Explore the dashboard, review assessments, and understand the regulatory frameworks Azure supports.

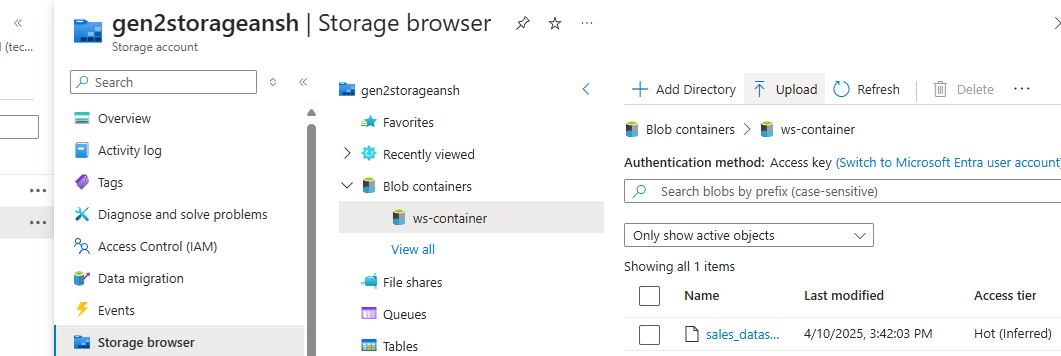


**EXERCISE 4: Azure Synapse Analytics**

**TASK 1: Deploy a sample Azure Synapse Analytics workspace**

1. Go to Azure Synapse Analytics > Create > Enter Details   
   
2. Select or create a Dala Lake Gen 2 storage    
   
3. Once again set a sql server admin login and password
4. Click on Review and Create

**TASK 2: Load a sample dataset and perform basic queries**

1. Upload a dataset in your container created   


1. Go to Data side tab > browse to your container > right click on ingested data > Select top 100 rows  
   A screenshot of a computer

   AI-generated content may be incorrect.
2. This will write a sql script for you to display first 100 rows of your ingested dataset.  
   A screenshot of a computer code

   AI-generated content may be incorrect.
3. You can execute the query using the serverless sql pool provided while creating the workspace.
4. However due to firewall limitations we are unable to use either serverless sql pools of dedicated sql pools in this azure account.

**TASK 3: Document Azure Synapse Key Benefits and Use cases**

**Azure Synapse Key Benefits:**

1. **Unified Analytics Platform**: Combines big data and data warehousing into a single platform, allowing for streamlined data analysis.
2. **Scalability**: Offers on-demand scalability, enabling you to handle large datasets and workloads without managing infrastructure.
3. **Integrated AI and Machine Learning**: Built-in integration with Azure Machine Learning and cognitive services to run advanced analytics and AI models.
4. **Real-time Analytics**: Supports real-time data streaming and analytics, providing insights with minimal delay.
5. **End-to-End Security**: Features robust security with encryption, firewalls, threat protection, and compliance with industry standards.
6. **Optimized Query Performance**: Leverages in-memory processing, parallel query execution, and caching for faster query performance.
7. **Serverless and Provisioned Models**: Offers both serverless querying and provisioned resources, allowing cost flexibility.
8. **Easy Integration with Other Azure Services**: Seamlessly integrates with tools like Power BI, Azure Data Factory, and Azure Databricks for enhanced data processing and visualization.

**Azure Synapse Use Cases:**

1. **Data Warehousing**: Store, manage, and analyze large datasets with high-performance query capabilities.
2. **Real-Time Analytics**: Process and analyze real-time streaming data for quick decision-making.
3. **Big Data Processing**: Work with massive amounts of unstructured and structured data using Spark and other big data tools.
4. **Business Intelligence**: Integrate with Power BI for advanced reporting, dashboards, and visual analytics.
5. **Advanced Analytics and AI**: Run predictive analytics and machine learning models directly within the platform.
6. **ETL/ELT Pipelines**: Build data pipelines using Azure Data Factory to move, transform, and load data for further analysis.
7. **Data Lakes**: Store raw, unstructured data in Azure Data Lake and process it using Synapse's integrated tools.
8. **Cost-Effective Storage**: Archive historical data with minimal cost by utilizing the platform's tiered storage model.

**EXERCISE 5: DataBricks for Data Engineering**

**TASK 1: Deploy a DataBricks workspace**

1. **Sign in to the Azure Portal**:
   * Go to [Azure Portal](https://portal.azure.com/) and log in with your credentials.
2. **Create a New Resource**:
   * Click **Create a resource** on the left-hand menu.
   * Search for **Azure Databricks** in the search bar.
   * Select **Azure Databricks** from the results and click **Create**.

A screenshot of a computer

Description automatically generated

1. **Configure the Workspace**:
   * **Resource Group**: Choose an existing resource group or create a new one.
   * **Workspace Name**: Provide a unique name for your Databricks workspace.
   * **Region**: Select the Azure region closest to your users or data source for better performance.
   * **Pricing Tier**: Select a pricing tier based on your requirements (Standard, Premium, or Trial).

A screenshot of a computer

Description automatically generated

1. **Networking (Optional)**:
   * Configure network settings if required, such as deploying the workspace in a Virtual Network (VNet).
2. **Review + Create**:
   * Click **Review + Create** to validate your configuration.
   * Once validation is successful, click **Create** to deploy the workspace.
3. **Wait for Deployment**:
   * The deployment process might take a few minutes. Monitor the progress in the **Notifications** section of the portal.

A white background with black text

Description automatically generated

1. **Access the Workspace**:
   * Once deployment is complete, go to the **Resource** to access the Databricks workspace.
   * Click the **Launch Workspace** button to open the Azure Databricks environment.

**TASK 2: Process a big sample dataset**

* 1. Go to your DataBricks Workspace > Create > New Notebook and initialize your spark session. You will need a cluster to run queries  
     A screen shot of a computer

     Description automatically generated
  2. Extraction  
     A screenshot of a computer

     Description automatically generatedA screenshot of a computer

     Description automatically generated

1. Since no column names are provided, we need to add them manually  
   A screen shot of a computer code

   Description automatically generated
2. Getting Schema of the dataframe  
   A screen shot of a computer program

   Description automatically generated
3. Repeating same steps for other dataframe and creating Views  
   A screenshot of a computer program

   Description automatically generated
4. Deriving Analytics from the data  
   Gender wise user breakdown  
   A screen shot of a computer

   Description automatically generated  
     
   **Give the top 5 movies which are reviewed maximum number of times**A screen shot of a computer

   Description automatically generatedList the top 10 movies which received highest number of 5 star ratings  
   A screen shot of a computer

   Description automatically generated

**TASK 3: DataBricks Key Features and use cases**

**Key Features of Databricks**

* Unified platform for **data engineering, analytics, and ML**
* Built on **Apache Spark** for fast, distributed processing
* Supports **Delta Lake** with ACID transactions and schema enforcement
* Collaborative **notebooks** with multi-language support (Python, SQL, etc.)
* Built-in **visualizations** and **MLflow** for ML lifecycle
* **Auto-scaling clusters**, CI/CD, and cloud integration (Azure, AWS, GCP)

**Use Cases**

* **ETL & Data Pipelines**: Ingest, transform, and clean large datasets
* **Data Lakehouse**: Unified storage and analytics using Delta Lake
* **Machine Learning**: Build, train, and deploy ML models
* **Real-Time Analytics**: Process streaming data
* **Business Intelligence**: Connect with Power BI/Tableau for reporting