**Ansh Ranjan  
Azure Data  
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