**Madiha Aimon Tappal**

[madihaaimon@gmail.com](mailto:madihaaimon@gmail.com)

**Data Engineering Batch – 1**

**Day – 16 Assignment**

**Azure Databricks**

**Cloud computing: -**

Cloud computing is a paradigm that involves delivering computing services, including storage, processing power, databases, networking, analytics, software, and intelligence, over the internet ("the cloud") to offer flexible resources and on-demand services. Instead of owning and maintaining physical servers or infrastructure, individuals and organizations can access and utilize computing resources as needed from cloud service providers.

Key characteristics of cloud computing include:

1. **On-Demand Self-Service:** Users can provision and manage computing resources as needed without requiring human intervention from the service provider.
2. **Broad Network Access:** Services are accessible over the internet through a variety of devices like laptops, smartphones, and tablets. Users can access cloud services from anywhere with an internet connection.
3. **Resource Pooling:** Cloud providers pool and allocate resources to multiple users, ensuring efficient utilization and cost-effectiveness. Resources are dynamically assigned based on demand.
4. **Rapid Elasticity:** Cloud resources can be scaled up or down quickly to meet changing workload demands. This enables users to pay only for the resources they consume.
5. **Measured Service:** Cloud computing resources are metered, and users are billed based on their usage. This pay-as-you-go model allows for cost optimization and efficient resource management.

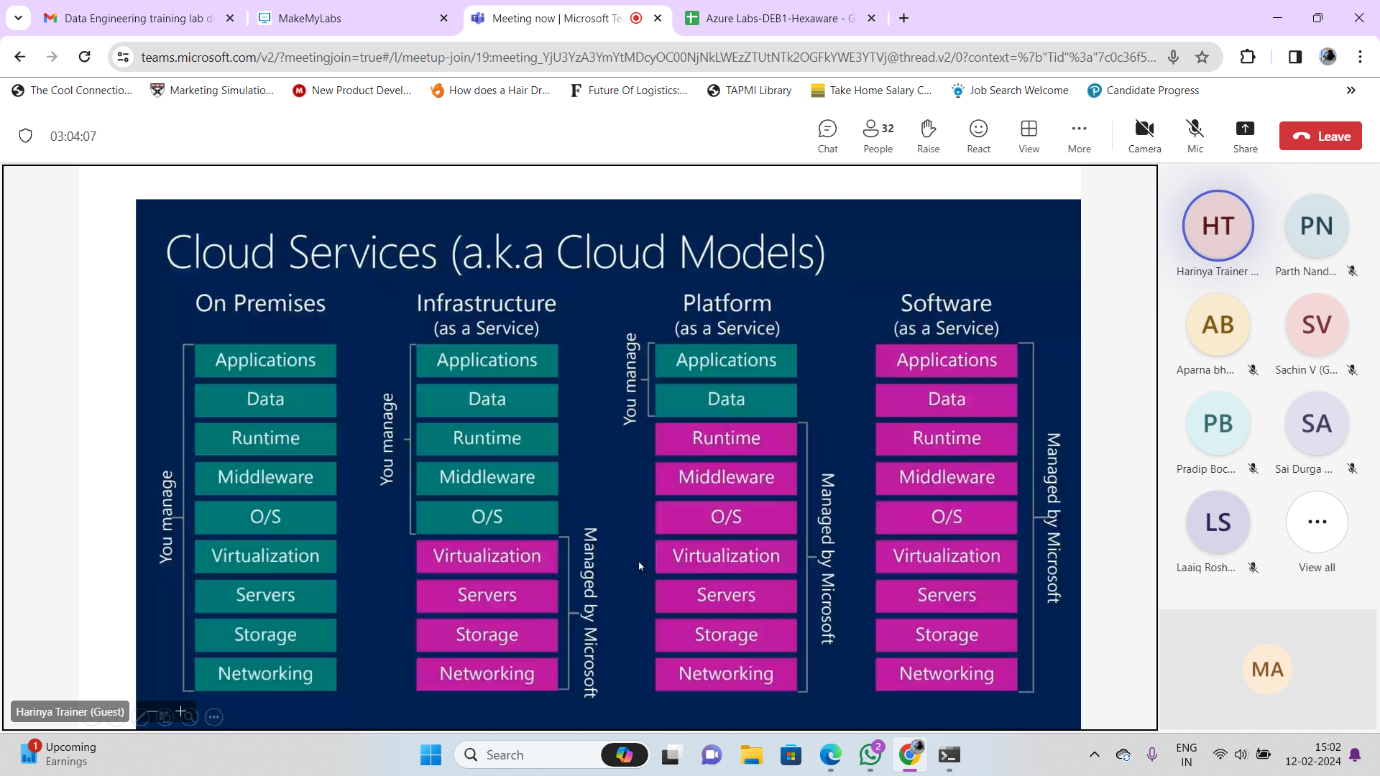
Cloud computing is generally categorized into three main service models and four deployment models:

**Service Models:**

* **Infrastructure as a Service (IaaS):** Provides virtualized computing resources over the internet. Users can rent virtual machines, storage, and networking components.
* **Platform as a Service (PaaS):** Offers a platform with tools and services for application development, simplifying the process for developers. Users can focus on building applications without worrying about the underlying infrastructure.
* **Software as a Service (SaaS):** Delivers software applications over the internet, eliminating the need for users to install, manage, and maintain the software locally.

**Deployment Models:**

* **Public Cloud:** Cloud resources and services are owned and operated by a third-party cloud service provider and are made available to the general public.
* **Private Cloud:** Cloud infrastructure is used exclusively by a single organization. It may be managed by the organization itself or by a third party.
* **Hybrid Cloud:** Combines elements of public and private clouds, allowing data and applications to be shared between them.
* **Multi-Cloud:** Involves using services from multiple cloud providers, providing redundancy, flexibility, and avoiding vendor lock-in.

****

**Microsoft Azure: -**

Microsoft Azure is a comprehensive cloud computing platform provided by Microsoft. It offers a wide range of services and resources, allowing businesses and individuals to build, deploy, and manage applications and services through a global network of data centers. Azure provides a scalable and flexible infrastructure that includes computing power, storage, databases, networking, analytics, machine learning, and more.

Key components and services of Microsoft Azure include:

1. **Compute Services:** Azure provides various options for computing, such as Virtual Machines (VMs), Azure App Service, Azure Functions (serverless computing), and more.
2. **Storage Services:** Azure offers different types of storage solutions, including Blob Storage for unstructured data, Table Storage for NoSQL data, Queue Storage for messaging between application components, and Azure Files for fully managed file shares.
3. **Database Services:** Azure provides a variety of managed database services, including Azure SQL Database, Azure Cosmos DB (a globally distributed, multi-model database service), Azure Database for MySQL, Azure Database for PostgreSQL, and more.
4. **Networking Services:** Azure offers a range of networking services, including Azure Virtual Network, Azure Load Balancer, Azure Application Gateway, VPN Gateway, and Azure DNS.
5. **Identity and Access Management:** Azure Active Directory (Azure AD) is a cloud-based identity and access management service that helps users sign in and access resources.
6. **AI and Machine Learning Services:** Azure includes services for artificial intelligence and machine learning, such as Azure Machine Learning, Azure Cognitive Services, and Azure Bot Services.
7. **Internet of Things (IoT):** Azure IoT services enable the development and management of IoT applications, with features for device management, data analytics, and integration with other Azure services.
8. **DevOps Services:** Azure DevOps provides a set of development and collaboration tools, including Azure Repos, Azure Pipelines, Azure Boards, and Azure Test Plans.
9. **Security and Compliance:** Azure offers a range of security services and compliance features, including Azure Security Center, Azure Policy, Azure Monitor, and more.
10. **Developer Tools and SDKs:** Azure supports a variety of programming languages, frameworks, and development tools, allowing developers to build applications using their preferred technologies.

**Azure Databricks: -**

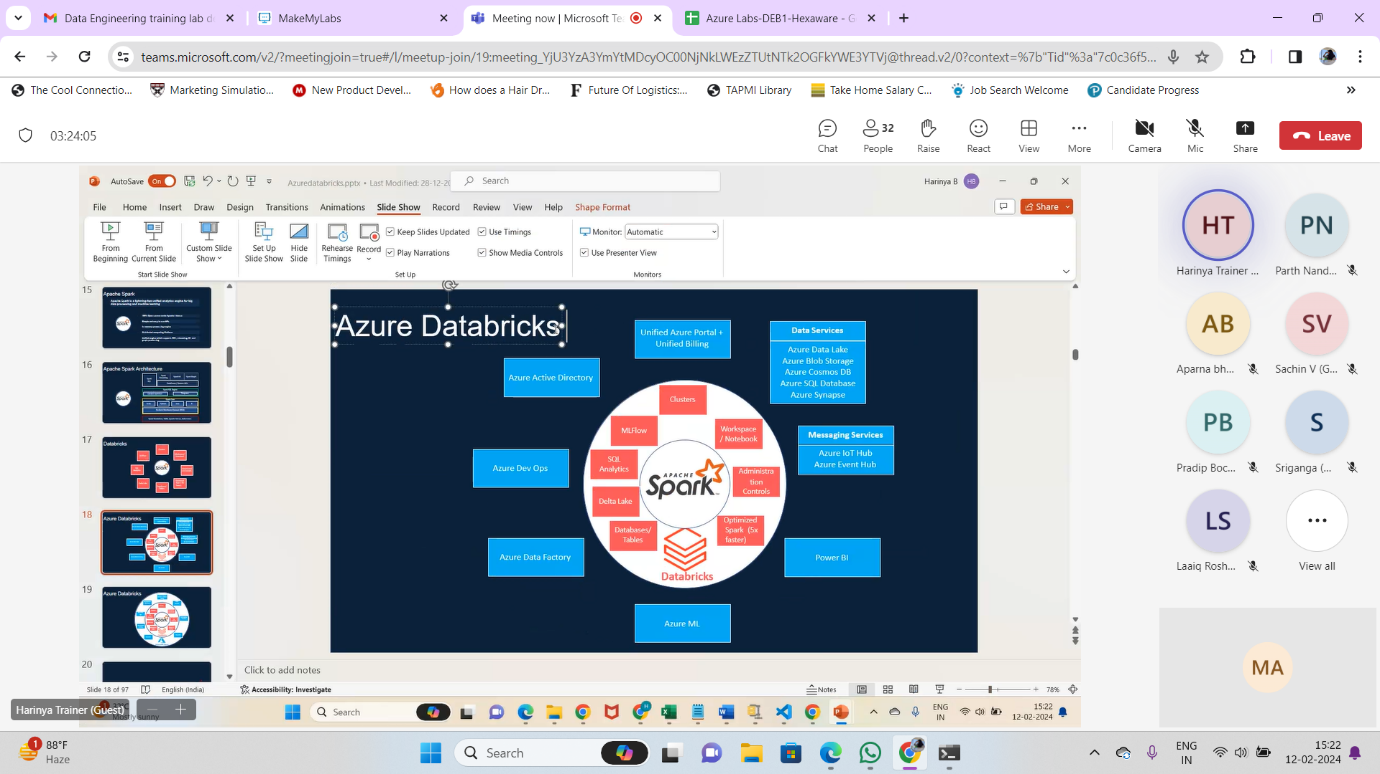
Azure Databricks is a cloud-based big data analytics platform that combines Apache Spark with Microsoft Azure cloud services. It is a collaborative environment for data science, machine learning, and big data analytics. Databricks provides a Unified Analytics Platform, offering a set of tools and services to process and analyze large volumes of data efficiently.

Key features of Azure Databricks include:

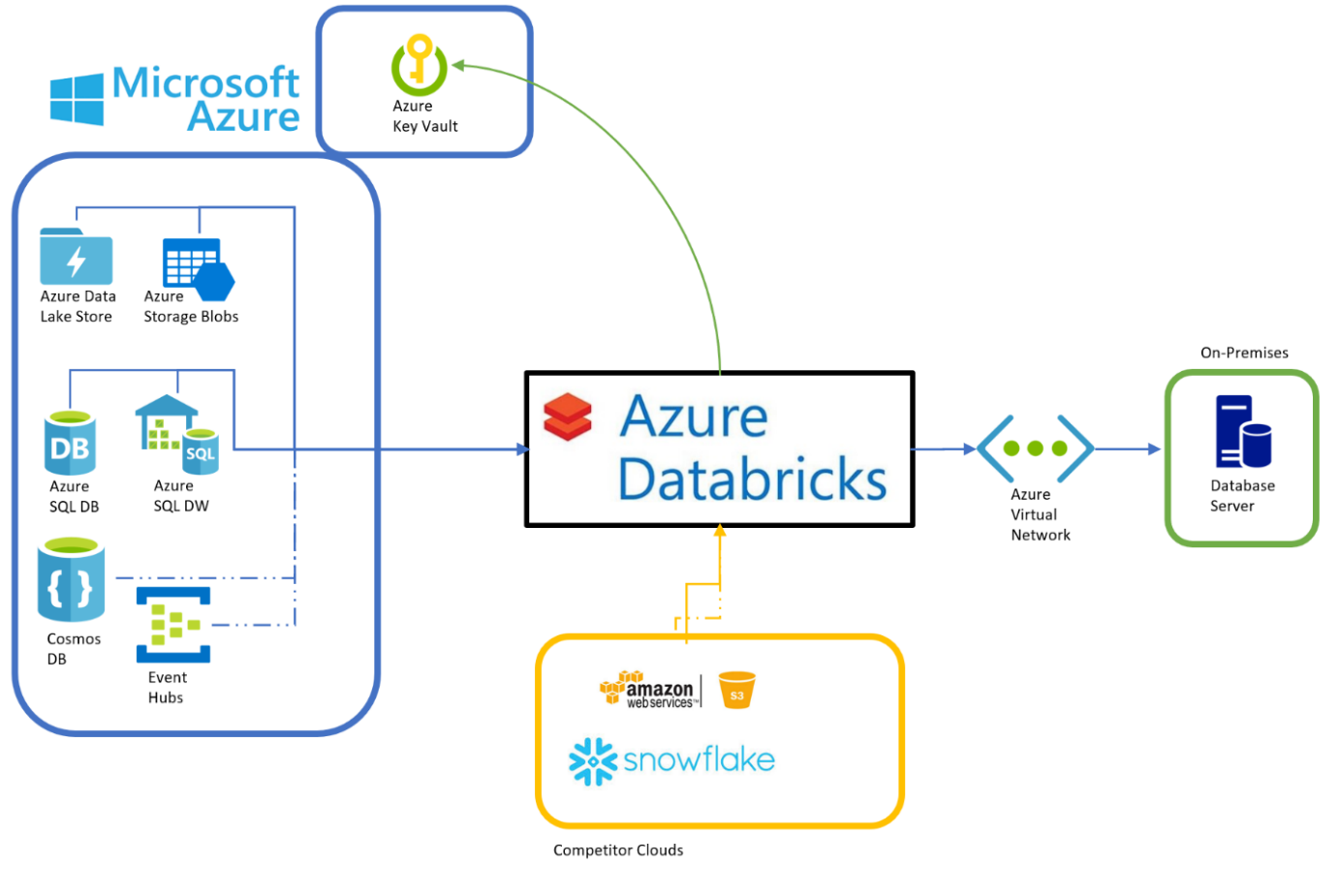
1. **Apache Spark Integration:** Databricks is built on Apache Spark, an open-source, distributed computing system. This allows users to leverage Spark's capabilities for data processing, machine learning, and graph processing.
2. **Collaborative Workspace:** Azure Databricks provides a collaborative workspace where data engineers, data scientists, and analysts can work together. It supports multiple programming languages such as Python, Scala, R, and SQL.
3. **Notebooks:** Users can create and share interactive notebooks, which contain code, visualizations, and narrative text. Notebooks are commonly used for data exploration, analysis, and building machine learning models.
4. **Libraries:** Databricks supports the installation of libraries and packages for different programming languages, enabling users to extend the platform's functionality with third-party tools and frameworks.
5. **Job Scheduling:** Users can schedule and automate data engineering and data science tasks using job scheduling features. This helps in managing and orchestrating complex workflows.
6. **Integrated Data Management:** Databricks integrates with Azure Data Lake Storage, Azure Blob Storage, and other Azure data services. This allows users to seamlessly access and manage their data stored in Azure.
7. **Machine Learning:** Azure Databricks includes MLlib, Spark's machine learning library, and supports integration with Azure Machine Learning. Users can build, train, and deploy machine learning models at scale.
8. **Security:** Databricks provides features for data security and compliance, including Azure Active Directory integration, role-based access control (RBAC), and encryption of data at rest and in transit.
9. **Streaming Analytics:** Databricks supports real-time data processing and analytics through its integration with Apache Spark Streaming, allowing users to analyze streaming data.
10. **AutoML (Automated Machine Learning):** The platform includes capabilities for automated machine learning, simplifying the process of model selection and hyperparameter tuning.

Azure Databricks is well-suited for organizations dealing with large-scale data analytics and data science workloads. It streamlines the process of working with big data and enables users to derive insights from their data in a collaborative and scalable environment. The platform leverages the cloud resources and services provided by Microsoft Azure, making it easy to integrate with other Azure services.

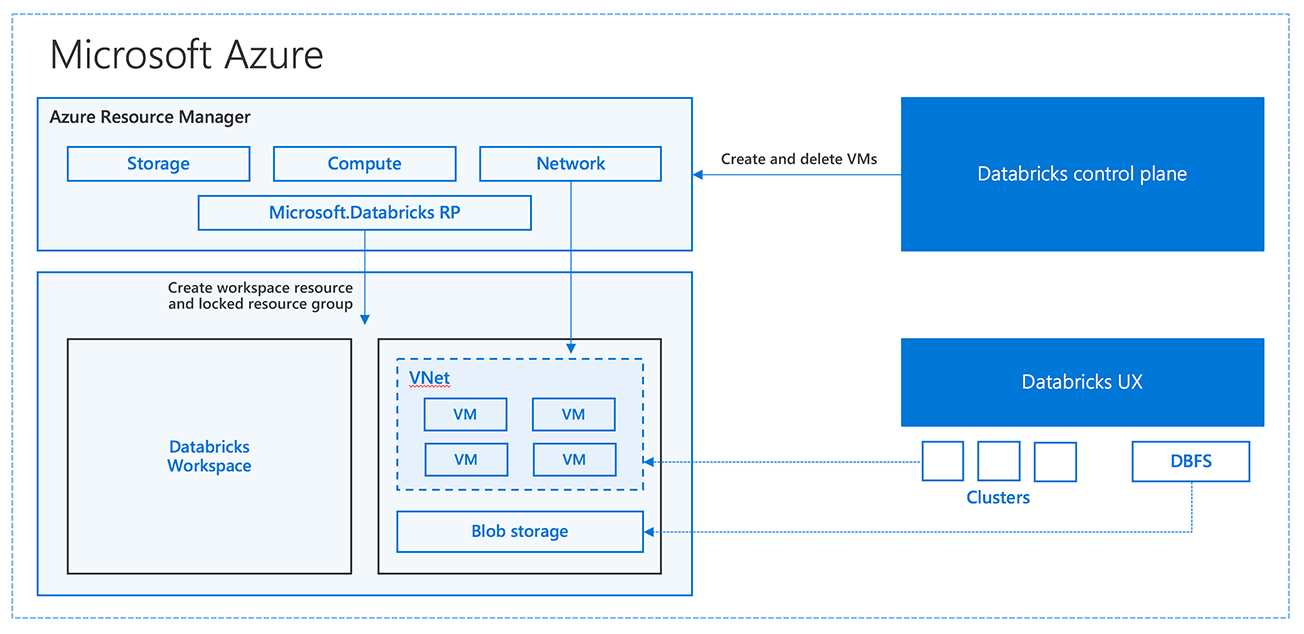
Top of Form

****

**Azure Databricks: -**



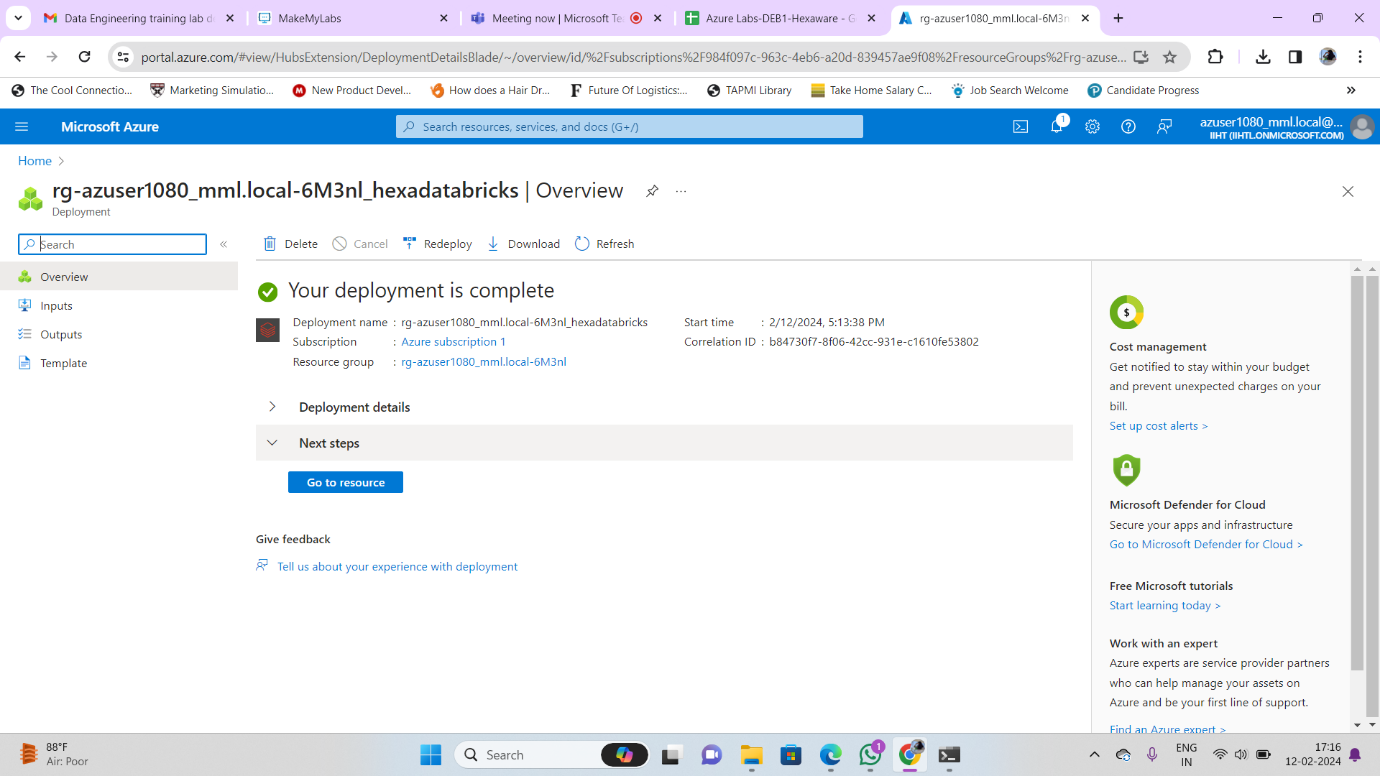
**Azure Databricks Architecture: -**

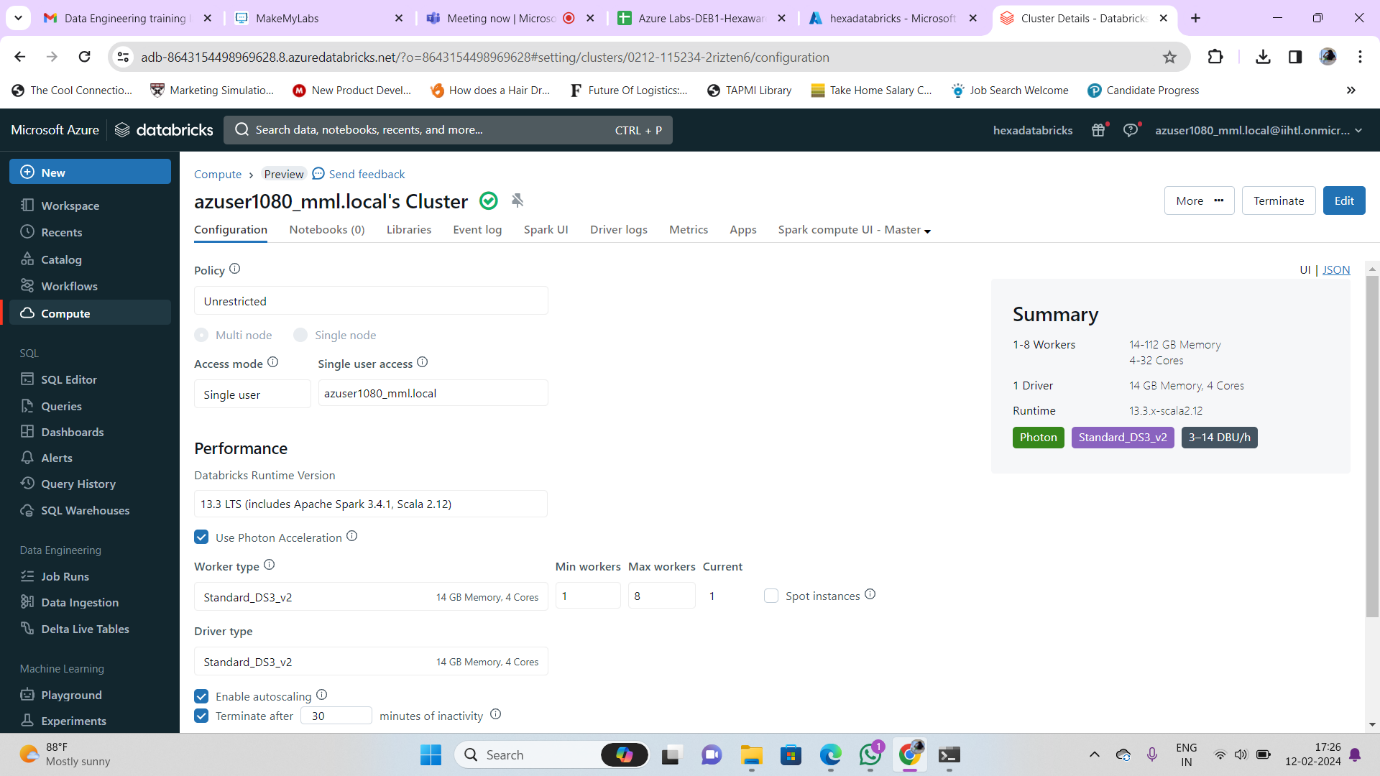


**Creating a cluster in Azure Databricks involves setting up a cluster of virtual machines that will run your Spark jobs and notebooks.**

the general steps to create a cluster in Azure Databricks: -

1. **Navigate to Azure Databricks Workspace:**
   * Go to the Azure portal (<https://portal.azure.com/>).
   * In the left sidebar, click on "Resource groups" and select the resource group where your Databricks workspace is located.
2. **Open Azure Databricks Workspace:**
   * In the resource group, find and click on your Databricks workspace.
3. **Access Databricks Workspace:**
   * Once inside your Databricks workspace, click on the "Launch Workspace" button. This will open the Databricks workspace in a new tab.
4. **Navigate to Clusters:**
   * In the Databricks workspace, click on the "Clusters" icon in the left sidebar.
5. **Create a New Cluster:**
   * Click on the "Create Cluster" button.
6. **Configure Cluster Settings:**
   * Provide a name for your cluster.
   * Choose the Databricks Runtime version (the version of Apache Spark that Databricks is built on).
   * Select the type of worker nodes you want in your cluster, specifying the number of worker nodes and the VM size.
   * Optionally, configure advanced options, such as auto-scaling, cluster tags, and initialization scripts.
7. **Configure Advanced Options (Optional):**
   * You can configure advanced options, such as auto-scaling, custom libraries, initialization scripts, etc., based on your specific requirements.
8. **Create Cluster:**
   * Click the "Create Cluster" button to provision the cluster.
9. **Monitor Cluster Creation:**
   * You will be redirected to the Clusters page where you can monitor the progress of your cluster creation. It may take a few minutes for the cluster to be fully provisioned.
10. **Access the Cluster:**
    * Once the cluster is created, you can access it by going to the Clusters page and clicking on the cluster name. From there, you can attach notebooks, run Spark jobs, and manage the cluster settings.
11. **Terminate Cluster (Optional):**
    * To save costs, remember to terminate the cluster when you're done using it. You can do this by going to the Clusters page, selecting the cluster, and clicking the "Terminate" button.

****

****