**Madiha Aimon Tappal**

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

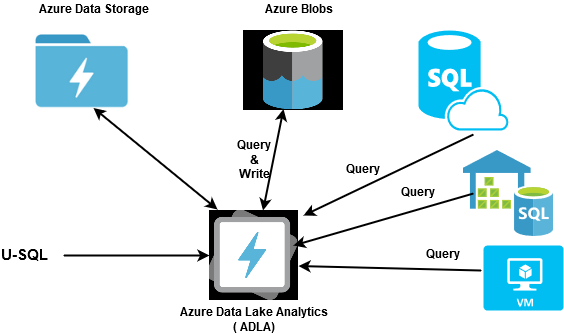
**Data Engineering Batch – 1**

**Day – 20 Assignment**

**Azure Databricks**

Azure Data Lake Storage (ADLS) is a cloud-based storage service provided by Microsoft Azure for big data analytics. It is designed to handle large amounts of data for analytics and machine learning workloads. Here are some key theoretical concepts related to Azure Data Lake Storage:

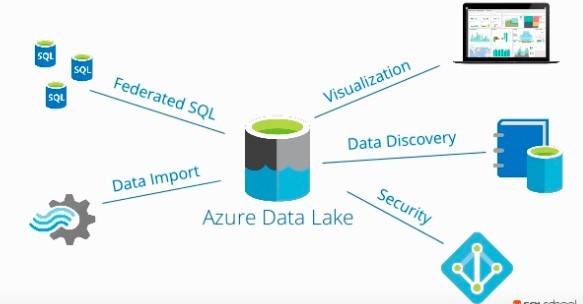
1. **Storage Accounts:**
   * Azure Data Lake Storage is associated with an Azure Storage Account, which is a general-purpose storage account that supports different Azure Storage services, including Blob storage, Azure Files, and Azure Queues.
2. **Hierarchical Namespace:**
   * ADLS Gen2 (second generation) introduces a hierarchical namespace, which allows you to organize and manage data in a hierarchical structure similar to a file system. This makes it more efficient for organizing and accessing large datasets.
3. **File System:**
   * Within an ADLS Gen2 account, you can create one or more file systems. Each file system acts as a container for directories and files. The hierarchical namespace provides a way to organize data within these file systems.
4. **Directories and Files:**
   * Data within ADLS is organized into directories and files. Directories are like folders and can contain both files and subdirectories. Files are individual units of data stored in the file system.
5. **Access Control:**
   * ADLS supports Azure Role-Based Access Control (RBAC), which allows you to control access to your data by defining roles and assigning those roles to users, groups, or applications. This helps in securing data at various levels.
6. **Data Lake Storage Gen1 vs. Gen2:**
   * ADLS Gen1 was the first version, and it did not have the hierarchical namespace, making it less efficient for certain types of data operations. ADLS Gen2 introduced the hierarchical namespace and is built on Azure Blob storage, combining the best features of both.
7. **Blob Storage Integration:**
   * ADLS Gen2 is built on top of Azure Blob storage, providing a unified storage solution. This integration allows you to use the capabilities of both ADLS Gen2 and Blob storage within the same storage account.
8. **Analytics and Processing:**
   * Azure Data Lake Storage is often used in conjunction with other Azure services like Azure Databricks, Azure HDInsight, and Azure Synapse Analytics for big data analytics, processing, and querying.
9. **Storage Tiers:**
   * ADLS Gen2 supports different storage tiers, including hot, cool, and archive tiers. This allows you to optimize costs based on the access patterns and the importance of the data.

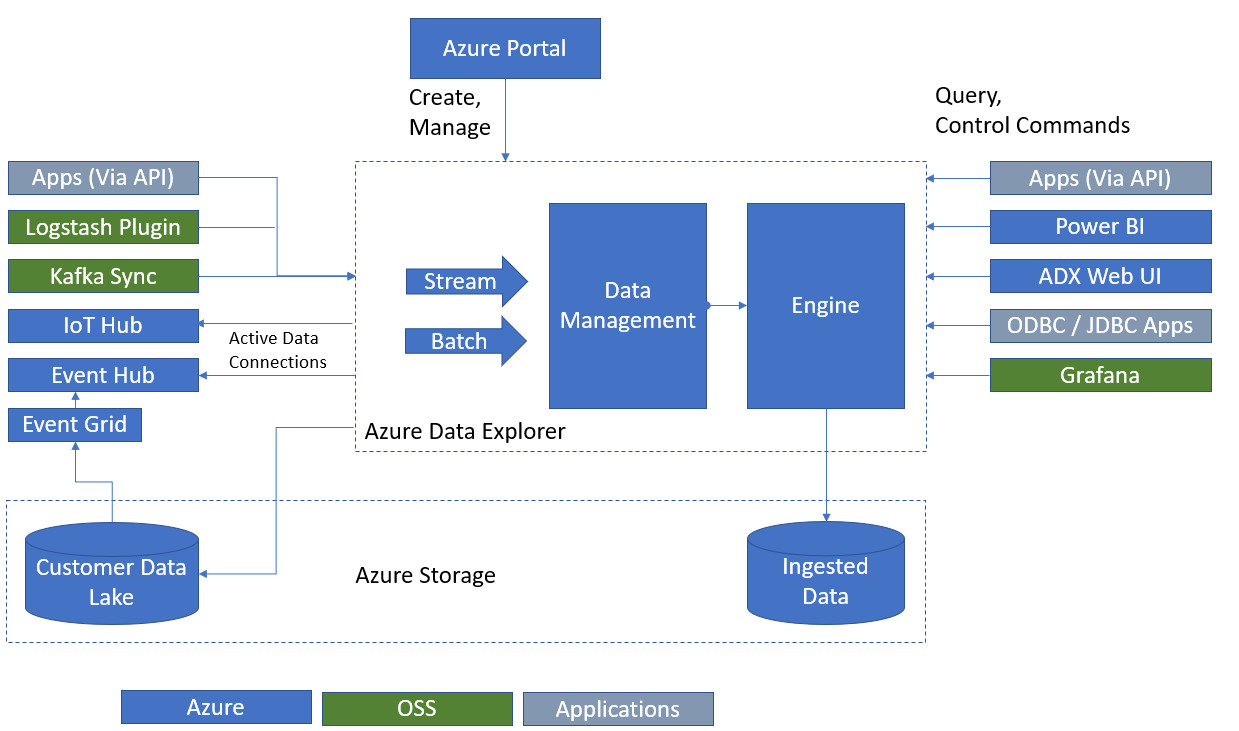


**Data lake gen 1 and gen 2 Difference: -**

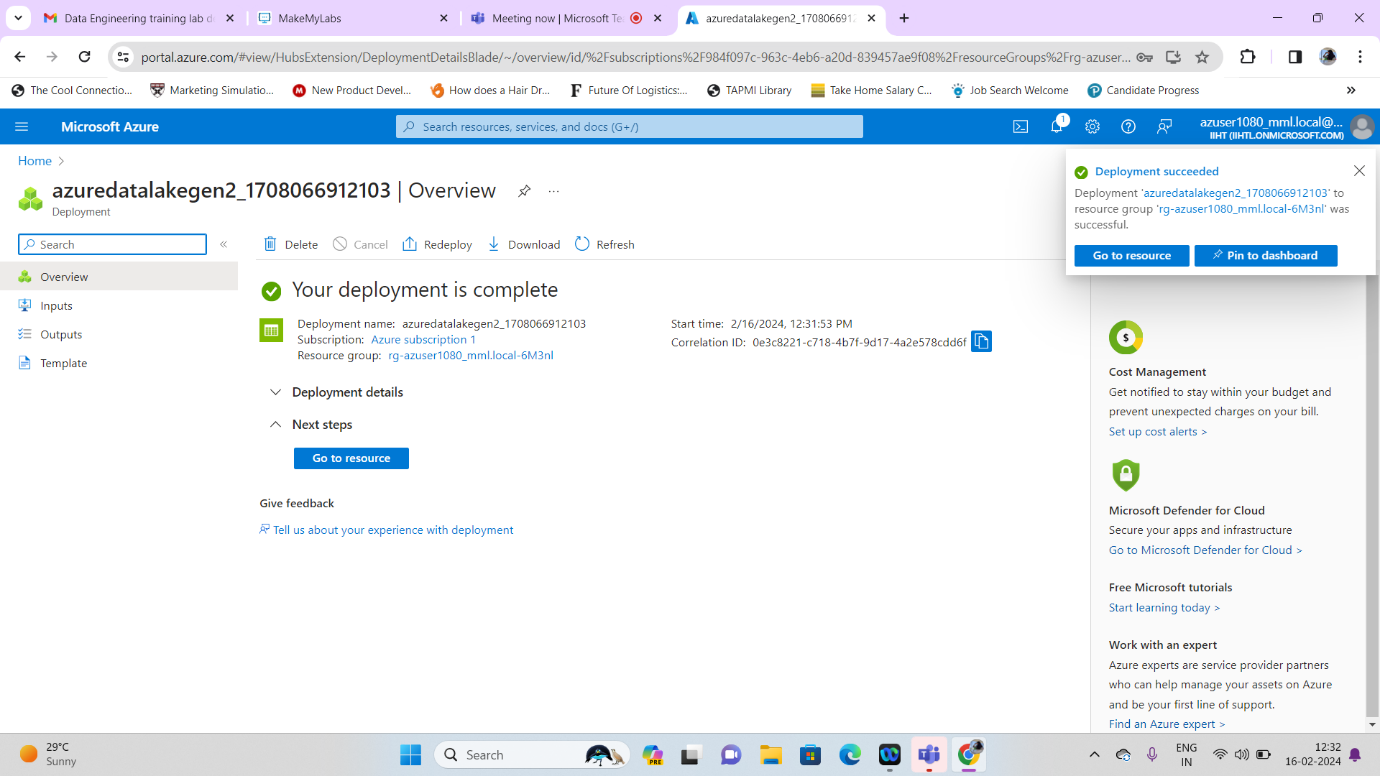
| **Feature** | **ADLS Gen1** | **ADLS Gen2** |
| --- | --- | --- |
| Hierarchical Namespace | Flat namespace | Hierarchical namespace |
| Integration with Blob Storage | Separate entities | Unified storage on top of Azure Blob storage |
| Compatibility | Less compatible with other Azure services | Compatible with Azure Blob storage and HDFS |
| Performance and Scalability | May have limitations | Improved performance and scalability |
| Access Control and Security | Limited access control options | Supports Azure RBAC for fine-grained control |
| Storage Tiers | No storage tiers | Supports hot, cool, and archive tiers |

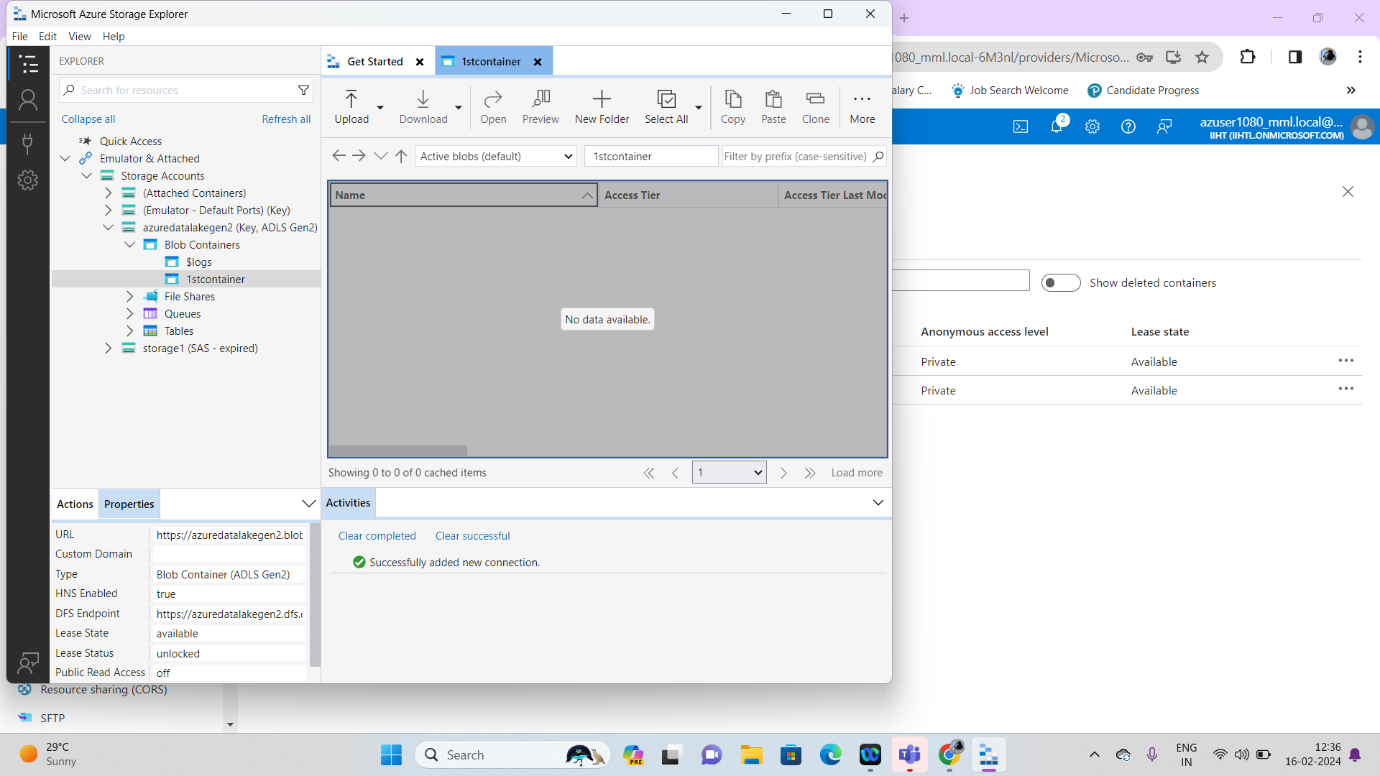
**Certainly, here's a concise table summarizing the key differences between Azure Data Lake Storage Gen1 and Gen2:**

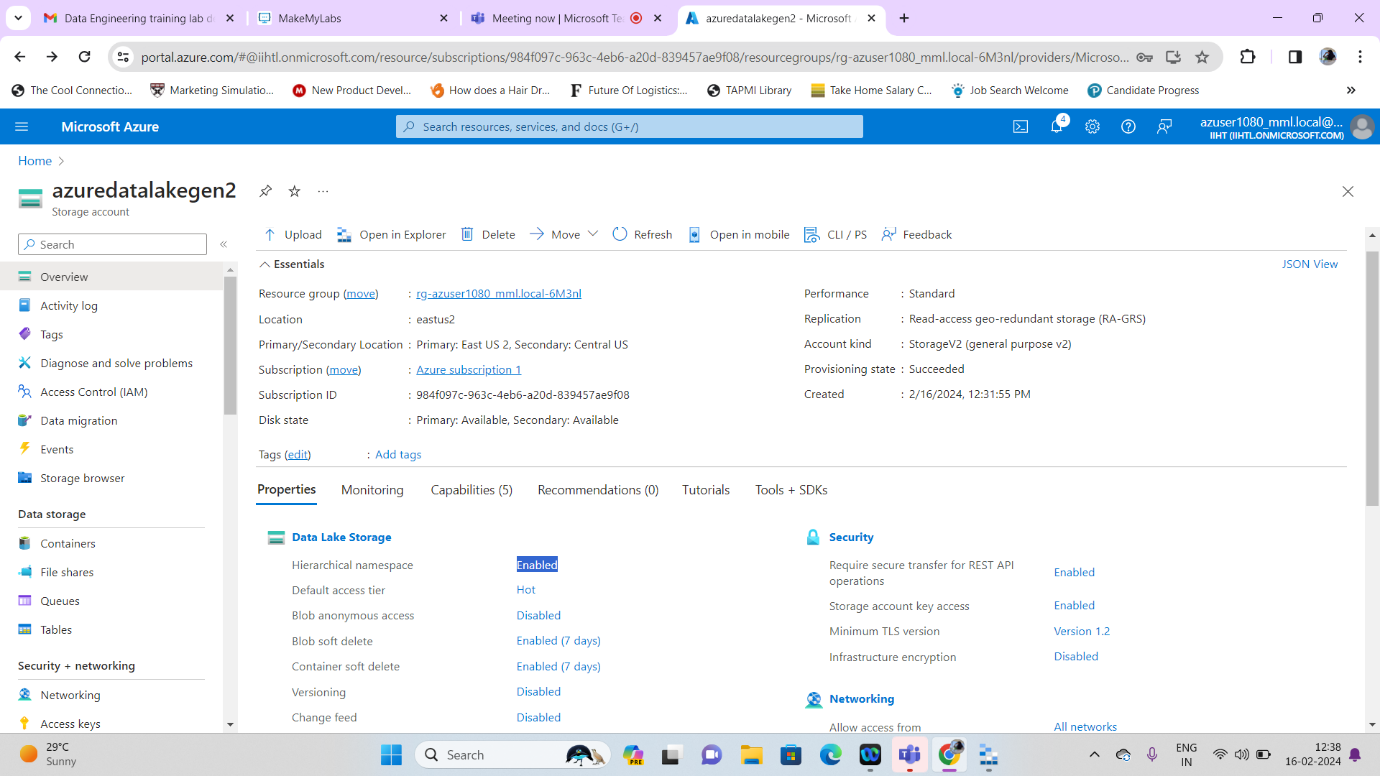


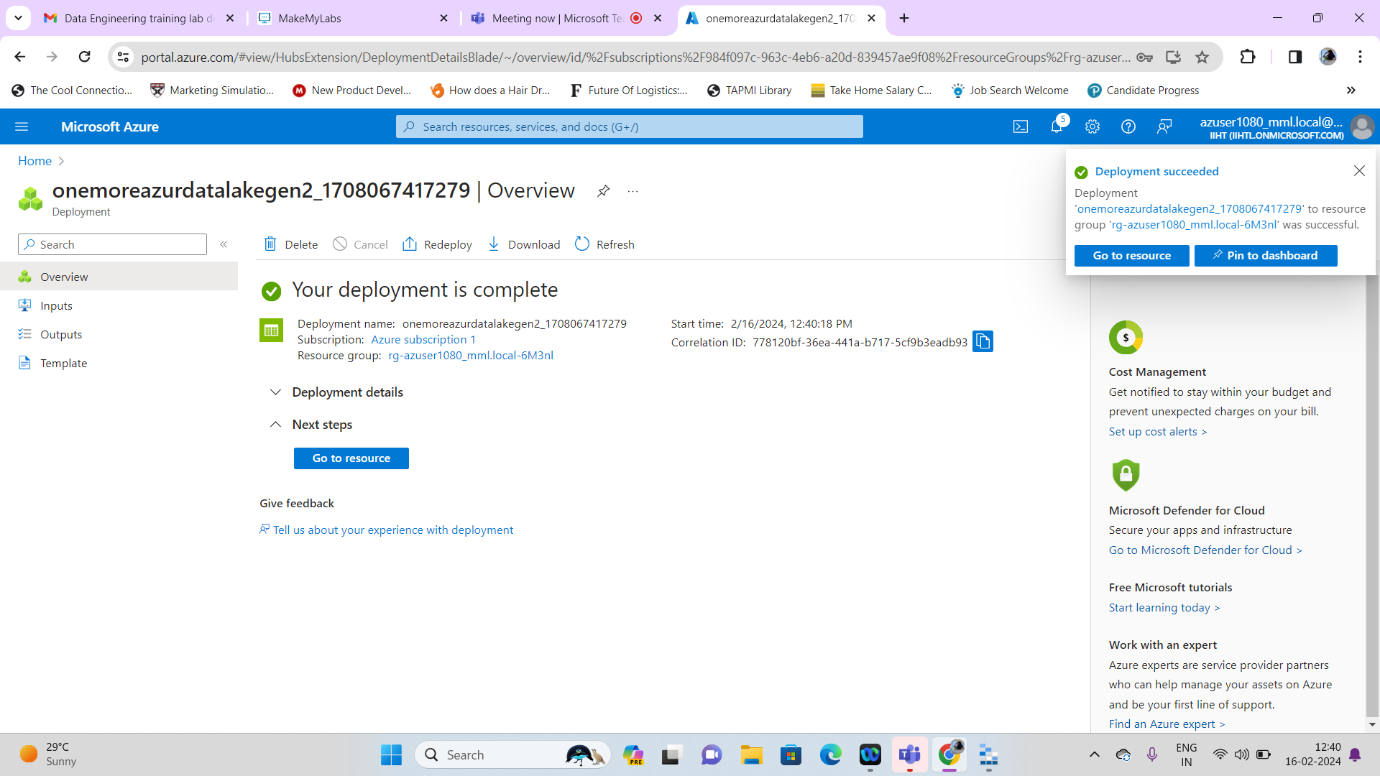


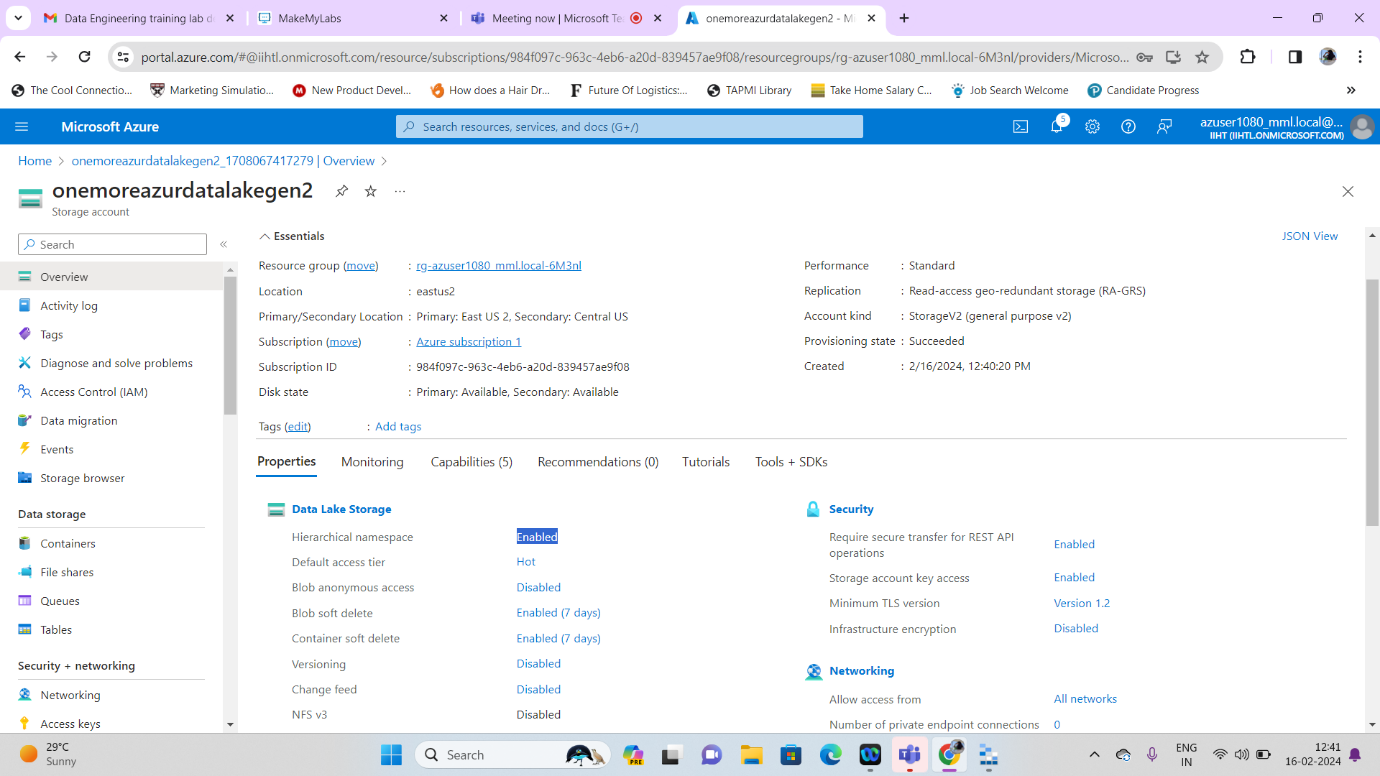
**Hands on: -**

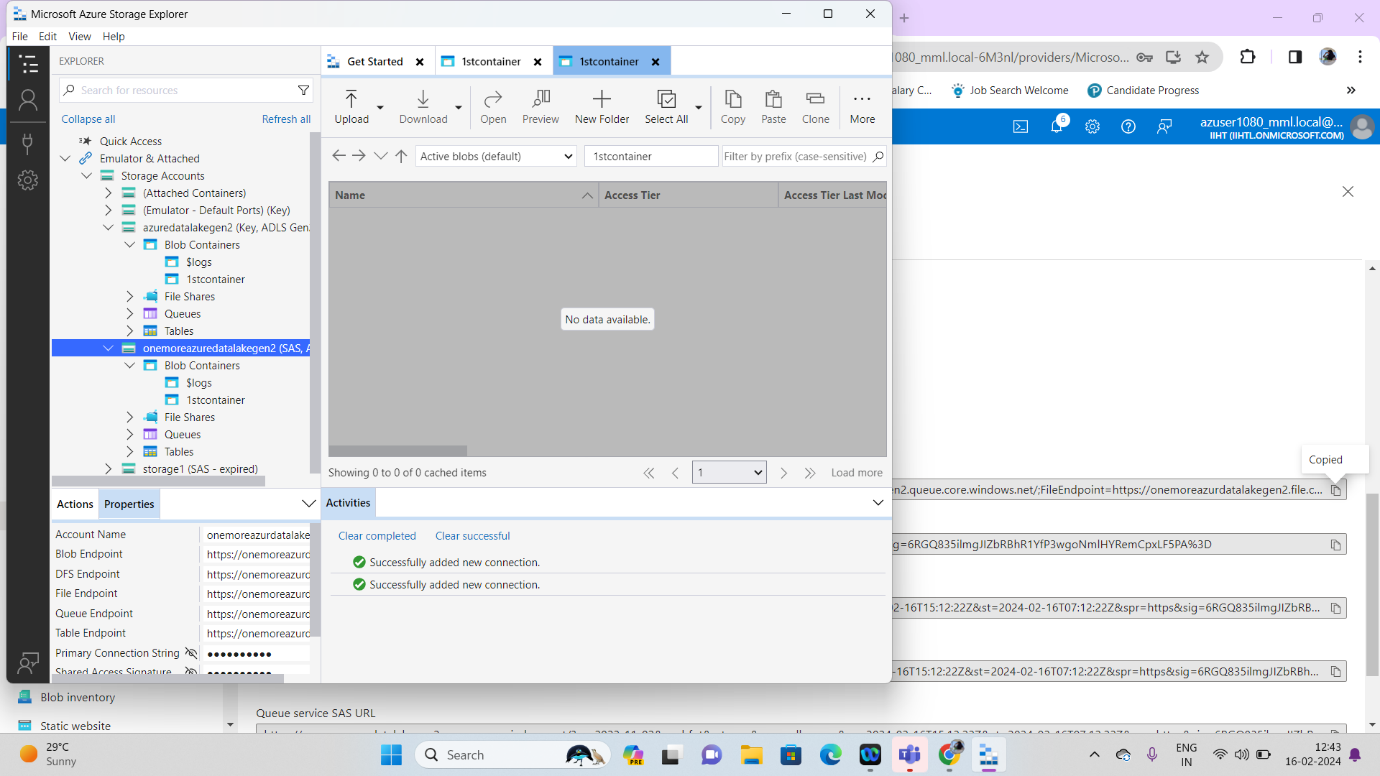
****

****

****

****

****

****

**Notes: -**