**Name -Monika Gariya**

**Email-** [**monikagariya2023@gmail.com**](mailto:monikagariya2023@gmail.com)

**Data Engineering Batch 1**

**Date – 16-02-2024**

**TOPIC – AZURE DATA LAKE STORAGE (ADLS)**

**Azure Data Lake Storage**

Azure Data Lake Storage (ADLS) is a massively scalable and secure data lake for high-performance analytics workloads. It was formerly known and is sometimes still referred to as the Azure Data Lake Store.

Designed to eliminate data silos, Azure Data Lake Storage provides a single storage platform that organizations can use to integrate their data. It can help optimize costs with tiered storage and policy management. It also provides role-based access controls and single sign-on capabilities through Azure Active Directory.

Users can manage and access data within Azure Data Lake Storage using the Hadoop Distributed File System (HDFS). Therefore any tool that you’re already using that is based on HDFS will work with Azure Data Lake Storage.

## **When to use Azure Data Lake?**

* **Data warehousing:** Azure Data Lake supports any type of data, so you can use it to integrate all of your enterprise data in a single data warehouse.
* **Internet of Things (IoT) capabilities:** The Azure platform provides tools for processing streaming data in real time from multiple types of devices.
* **Support for hybrid cloud environments:** You can use the Azure HDInsight component to extend an existing on-premises big data infrastructure to the Azure cloud.
* **Enterprise features:** The environment is managed and supported by Microsoft and includes enterprise features for security, encryption, and governance. You can also extend your on-premises security solutions and controls to the Azure cloud environment.
* **Speed to deployment:** It’s pretty easy to get up and running quickly with the Azure Data Lake solution. All of the components are available through the portal and there are no servers to install and no infrastructure to manage.

## **Features of Azure Data Lake**

Here are some of the key features and benefits that make Azure Data Lake stand out in the world of big data:

* **Simplified data management:** With Azure Data Lake, you can bid farewell to the hassle of managing multiple data storage systems. It provides a single, unified platform for all your different data types.
* **Improved data accessibility:** Get your data quickly and easily with Azure Data Lake, making it a breeze to derive insights and make data-driven decisions.
* **Enhanced data security:** Rest easy knowing that Azure Data Lake's robust security features protect your sensitive data, ensuring compliance with industry regulations.
* **Cost-effective scalability:** As your data storage and processing needs grow, Azure Data Lake has your back with scalable capabilities without breaking the bank or dealing with on-premises infrastructure complexities.
* **Accelerated innovation:** Thanks to Azure Data Lake's support for advanced analytics, machine learning, and real-time processing, your organization can rapidly develop and deploy innovative data-driven applications and services.

## **Data Lake Storage Gen2**

Azure Data Lake Storage Gen2 refers to the current implementation of Azure's Data Lake Storage solution. The previous implementation, Azure Data Lake Storage Gen1 will be retired on February 29, 2024.

Unlike Data Lake Storage Gen1, Data Lake Storage Gen2 isn't a dedicated service or account type. Instead, it's implemented as a set of capabilities that you use with the Blob Storage service of your Azure Storage account. You can unlock these capabilities by enabling the hierarchical namespace setting.

**The stages for processing big data**

Data lakes have a fundamental role in a wide range of big data architectures. These architectures can involve the creation of:

* An enterprise data warehouse.
* Advanced analytics against big data.
* A real-time analytical solution.

There are four stages for processing big data solutions that are common to all architectures:

* **Ingest** - The ingestion phase identifies the technology and processes that are used to acquire the source data. This data can come from files, logs, and other types of unstructured data that must be put into the data lake. The technology that is used will vary depending on the frequency that the data is transferred. For example, for batch movement of data, pipelines in Azure Synapse Analytics or Azure Data Factory may be the most appropriate technology to use. For real-time ingestion of data, Apache Kafka for HDInsight or Stream Analytics may be an appropriate choice.
* **Store** - The store phase identifies where the ingested data should be placed. Azure Data Lake Storage Gen2 provides a secure and scalable storage solution that is compatible with commonly used big data processing technologies.
* **Prep and train** - The prep and train phase identify the technologies that are used to perform data preparation and model training and scoring for machine learning solutions. Common technologies that are used in this phase are Azure Synapse Analytics, Azure Databricks, Azure HDInsight, and Azure Machine Learning.
* **Model and serve** - Finally, the model and serve phase involves the technologies that will present the data to users. These technologies can include visualization tools such as Microsoft Power BI, or analytical data stores such as Azure Synapse Analytics. Often, a combination of multiple technologies will be used depending on the business requirements.

**PRACTICAL**





