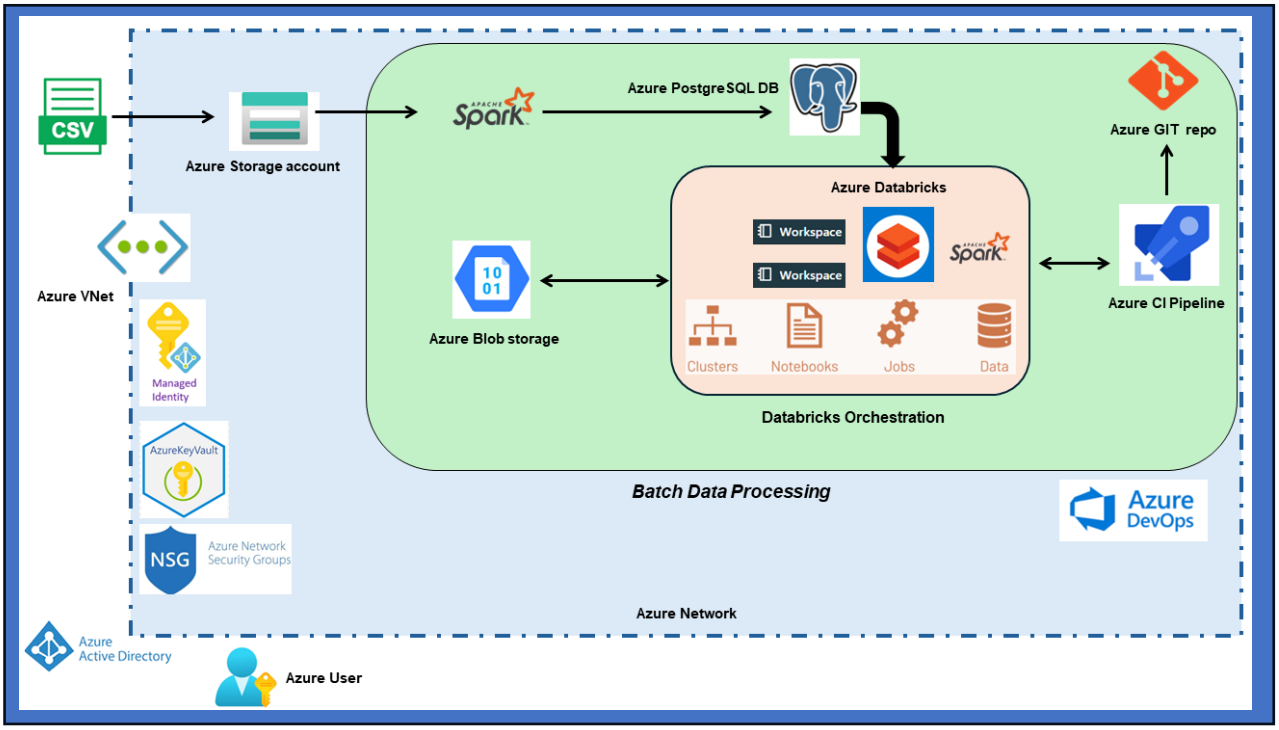
Project Aim:

The goal of this project is to use Azure services to design and build a reliable batch-processing data architecture that will be used as the backend of a machine learning application that processes enormous amounts of data. Creating a scalable, dependable, and maintainable system that can effectively process substantial amounts of data, store it, conduct the necessary preprocessing, and aggregate the data for use in machine learning applications is the main goal. The primary goal of this project is to construct the required data infrastructure by utilizing Azure DevOps, Azure Databricks, Azure Key Vault, Azure Blob Storage, Azure PostgreSQL, and Azure Network security. At the conceptual stage, we will focus on integrating canonical software components and frameworks and adapting common data engineering principles to create a state-of-the-art architecture for data processing. The design will prioritize batch scheduling of data processing tasks to coincide with the frontend application's quarterly execution cycle, which oversees producing new iterations of the machine learning model. Wilder, B. (2012).

Figure 1 : Batch data processing Architecture



Source: Own representation.

Data Source

The data set used for this project is from Kaggle and it is a random dataset generated using Python and is unrelated to any corporate entity. This dataset consists of more than a million records and includes timestamp which is the required criteria for our project.

Source link: https://www.kaggle.com/datasets/ksabishek/massive-bank-dataset-1-million-rows

Concepts and methodology

In this project, the focus is on batch-processing data architecture which will be built on the Azure platform. For this project below are the methods and technology that will be used.

Data ingestion:

Azure Databricks, which offers an unparalleled ETL (extract, transform, load) experience by fusing the power of Apache Spark with Delta Lake and proprietary tools, is used to induce the data to the storage. This can compose ETL logic using Python, Scala, and SQL, and then with a few clicks, orchestrate scheduled job deployment.

Processing and Storage:

For processing the data, Azure Databricks is used as it provides data processing scheduling and management, particularly ETL. Also managing security, governance, high availability, and disaster recovery. Data discovery, annotation, and exploration will be more convenient. Azure Databricks uses Spark core and RDD which is the underlying general execution engine for the Spark platform.

The Python API for Apache Spark is called PySpark. With Python, you can use it to process massive amounts of data in real-time within a distributed environment. Additionally, it offers a PySpark shell for interactive data analysis. PySpark allows anyone familiar with Python to process and analyse data of any size by fusing the ease of use and learnability of Python with the power of Apache Spark. Dudley, R. J. (2010).

In this project, two main data storage systems will be used, first Azure PostgreSQL which will store the processed Spark job data into SQL data structure and then this data which is fed to Azure Databrick will be converted to Parquet format data will be stored into Azure Blob storage for further analysis.

Data Governance, Secure, Scalability, Security, Protection and Maintainability

Azure provides a variety of services which include Governance, Secure, Scalability, Security, Protection, and maintenance of the data.

Protection: Tokens, passwords, certificates, API keys, and other secrets can be safely stored and strictly controlled with Azure Key Vault.

Maintainability: Azure offers four tiers of management: resources, resource groups, subscriptions, and management groups**.** Sharma, V., Nigam, V., & Sharma, A. K. (2020).

Scalability: Azure storage or the Azure clusters used in Databricks are auto-scalable based on the used requirements and processing.

Governance and Security: Network traffic between Azure resources in an Azure virtual network can be filtered using an Azure network security group. A network security group contains security

rules that allow or prohibit network traffic passing different types of Azure resources. Identity access management (IAM) provides governance as Cloud administrators configure and integrate coarse access control permissions for Unity Catalog.

References

Wilder, B. (2012). *Cloud architecture patterns: using Microsoft azure*. " O'Reilly Media, Inc.".

Collier, M., & Shahan, R. (2015). *Microsoft azure essentials-fundamentals of azure*. Microsoft Press.

Dudley, R. J. (2010). *Microsoft Azure: Enterprise Application Development*. Packt Publishing Ltd.

Galiveeti, S., Tawalbeh, L. A., Tawalbeh, M., & El-Latif, A. A. A. (2021). Cybersecurity analysis: Investigating the data integrity and privacy in AWS and Azure cloud platforms. In *Artificial intelligence and blockchain for future cybersecurity applications* (pp. 329-360). Cham: Springer International Publishing.

Sharma, V., Nigam, V., & Sharma, A. K. (2020). Cognitive analysis of deploying web applications on microsoft windows azure and amazon web services in global scenario. *Materials Today: Proceedings*, *11*.