**Phase 1: Setup & Configuration**

1. **Creating a Azure Blob Storage**

**Azure Blob Storage** is a cloud service for storing **massive amounts of unstructured data** like text, images, videos, backups, and logs—it's great for storing files you want to access from anywhere over the internet.

A screenshot of a computer

AI-generated content may be incorrect.

Created Resource group called “CIS-660\_Final\_project” . Contains all the resources required for this Project.

Azure Blob Storage name : **puneethdiabetesstorage**

A screenshot of a computer

AI-generated content may be incorrect.

Enabling Hierarchical Namespace so that Azure Data Lake Storage Gen2 gets Enabled

A close-up of a white background

AI-generated content may be incorrect.

Azure Blob Storage named **puneethdiabetesstorage** deployed under resource **CIS\_660\_Final\_project**

A screenshot of a computer

AI-generated content may be incorrect.

Following **Medallion architecture** where

**Bronze**: Stores raw, unprocessed data as it comes in.

**Silver**: Contains cleaned and structured data for analysis.

**Gold**: Holds final, aggregated data ready for reporting or ML.

A screenshot of a computer

AI-generated content may be incorrect.

Created three containers (bronze, gold silver ) inside ADLS-Gen2

A screenshot of a computer

AI-generated content may be incorrect.

1. **Creating Azure Data Factory for Data orchestration**

**Azure Data Factory (ADF)** is a cloud-based data integration service that lets you **build, schedule, and manage data pipelines** to move and transform data from various sources to storage or analytics systems—**without writing much code**.

A screenshot of a computer

AI-generated content may be incorrect.

Creating Data Factory named **diabetes-etl-adf** under the same resource group called **CIS-660\_Final\_Project**

A screenshot of a computer

AI-generated content may be incorrect.

Enabling Configure Git later

A white background with black text

AI-generated content may be incorrect.

Azure Data Factory got Deployed under resource CIS-660\_Final\_project

A screenshot of a computer

AI-generated content may be incorrect.

1. **Creating Azure Data Bricks for Transformations**

**Azure Databricks** is a fast, scalable analytics platform that combines **Apache Spark** with **Azure’s cloud services**, allowing you to run big data and machine learning workloads easily using **Python, SQL, R, or Scala** in collaborative notebooks.

A screenshot of a computer

AI-generated content may be incorrect.

Databricks workspace named **diabetes-etl-db1** is created along with default storage named **diabetes\_managed\_group** .

**A screenshot of a computer

AI-generated content may be incorrect.**

**A screenshot of a computer

AI-generated content may be incorrect.**

1. **Creating Initial ADF Pipelines**

Before Ingesting pipelines we need to provide access permissions to azure data factory on blob storage (ADLS Gen2)

Hence providing Storage Blob Data Contributor access through managed identity from storage

**A screenshot of a computer

AI-generated content may be incorrect.**

**A screenshot of a computer

AI-generated content may be incorrect.**

The data Factory diabetes-etl-adf now has contributor access on storage where it can read , write and delete files from storage

**Ingesting raw data from Kaggle**

**A diagram of a system

AI-generated content may be incorrect.**

**Image Reference : CloudBox Academy**

1. **Creating Linked service for Github and Data Lake**

**A screenshot of a computer

AI-generated content may be incorrect.**

**A screenshot of a computer

AI-generated content may be incorrect.**

Linked my github account with ADF

**A screenshot of a computer

AI-generated content may be incorrect.**

Similarly created a linked service for Azure data lake storage as well

1. **Creating source and sink Datasets**

**A screenshot of a computer

AI-generated content may be incorrect.**

Naming the dataset created to perform copy activity from http and also linking the data with created linked service

**Source Dataset (Files from Github):**

**Name:** ds\_diabetes\_data\_github\_http

**A screenshot of a computer

AI-generated content may be incorrect.**

Since the file is Zipped setting the compression type to gzip and column delimiter as Comma

**A screenshot of a computer

AI-generated content may be incorrect.**

**Sink Data Set (Files to store into):**

**Dataset Name :** ds\_puneethdiabetesdtorage\_raw (format :dataset\_storage-name\_container\_name)

**A screenshot of a computer

AI-generated content may be incorrect.**

1. **PipeLine and Copy activity**

Creating a pipeline to copy files from github to bronze(raw) container in the storage using **copy activity**

**A screenshot of a computer

AI-generated content may be incorrect.**

**A screenshot of a computer

AI-generated content may be incorrect.**

Specified **Source** as **ds\_diabetes\_data\_github\_http** and **sink** as **ds\_puneethdiabetesdtorage\_raw**

Pipeline named **pl\_ingest\_github\_raw** succeeded the first copy activity

**A screenshot of a computer

AI-generated content may be incorrect.**

The file from GitHub has been stored into bronze container

A screenshot of a computer

AI-generated content may be incorrect.