

Cardiovascular Disease Prediction Project Flow

I. GitHub

Purpose: To use GitHub as a data source.

1. Create a GitHub public repository.
2. Upload the dataset file to the repository.
3. Get the "raw" URL of the dataset file.

II. Azure Cloud

1. Resource Group

Purpose: Simplify resource management by grouping all related resources.

1. Create a resource group (e.g., big_data_project) and attach every Azure service/resource to this group.

2. Storage Account

Purpose: To store data securely in Azure cloud.

1. Create a storage account (e.g., cardiodisease), enabling the hierarchical storage option.
2. Create a container (e.g., dataset) to hold your data.
3. Create a directory inside the container (e.g., cardio_data) to store the dataset.

3. Data Factory (for data ingestion)

Purpose: Ingest and transfer data from GitHub to Azure storage.

1. Create a Data Factory resource (e.g., dataset-ingestion).
2. Go to "Author" → Create pipeline → Select "Move and Transform" → Drag "Copy data".
3. Name the pipeline (e.g., data_ingestion_pipeline).
4. Add source to the pipeline: Select new source → HTTP → Choose file format (CSV).
5. Add sink to the pipeline: Choose Azure Data Lake Storage Gen2 → Select file format (CSV).
6. Validate and debug the pipeline to ensure it's working correctly.

4. App Registration (to give Databricks access)

Purpose: Allow Databricks to access your Azure Storage account.

1. Search for App Registration in Azure and name it (e.g., App_cardio_disease).
2. Copy the "Application (client) ID" and "Directory (tenant) ID".
3. Go to "Certificates & Secrets" → Create new secret → Copy the value.

5. Databricks (for machine learning)

Purpose: To run machine learning models using Spark MLlib in a distributed environment.

1. Create a Databricks resource (e.g., databricks_cardiodisease).
2. Select 'Compute' option in left hand panel of the databricks and then create new cluster and give it name.
3. Select create new notebook and connect it to our cluster.
4. Run the first block of the code from notebook, replace the 'Application (client) ID', 'Secret Value', and 'Directory (tenant) ID'
5. Give mount name to load your file at location in databricks

Flow Diagram

