**Data Warehouse for Ethiopian Medical Businesses**

Prepared by: Abrham

Date: October 11, 2024

## Task 1: Data Scraping and Collection Pipeline

The first task involves collecting data from relevant public Telegram channels.   
 The channels scraped include DoctorsET, Lobelia4Cosmetics, Yetenaweg, and other medical-related channels.  
 Telethon, a Python package, is used to interact with the Telegram API and pull data such as messages and images for further   
 analysis. The scraping process also incorporates a monitoring and logging system to ensure that data collection happens   
 seamlessly and without error.

Steps for Data Scraping:

1. Install Telethon using pip and create a script to extract messages.  
 2. Collect data from channels such as DoctorsET, Lobelia4Cosmetics.  
 3. Log the process to capture errors and monitor progress.  
 4. Store the scraped data temporarily in a local database or file.

## Task 2: Data Cleaning and Transformation

After scraping, the raw data needs to be cleaned and standardized to ensure accuracy.   
 This involves removing duplicates, handling missing values, and validating the data.   
 Python libraries like Pandas are used for this task. After cleaning, DBT (Data Build Tool) is used to define transformation   
 models to shape the data in a format that is suitable for the warehouse.

Steps for Data Cleaning:

1. Remove duplicates using Pandas.  
 2. Handle missing values and standardize formats.  
 3. Set up DBT for data transformation and define models using SQL files.  
 4. Run the DBT models to load the transformed data into the data warehouse.