Data Warehouse for Ethiopian Medical Businesses

# Prepared by: Abrham

## Introduction

Kara Solutions, a leading data science company, has embarked on a project to build a robust and scalable data warehouse   
 that stores data on Ethiopian medical businesses scraped from web and Telegram channels. This data warehouse will enable   
 comprehensive data analysis, pattern detection, and object detection using advanced machine learning algorithms such as YOLO.  
 This report outlines the steps and considerations for building the data warehouse, along with the progress made on each   
 task to ensure the success of the project.

## 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.

## Task 3: Object Detection using YOLO

YOLO (You Only Look Once) is integrated to detect objects in images scraped from Telegram channels.   
 The collected images are used to train YOLO models, allowing for advanced analysis such as identifying medical products and equipment.

Steps for Object Detection:

1. Collect images from Telegram channels like Lobelia4Cosmetics.  
 2. Train YOLO to detect relevant objects in the images.  
 3. Integrate the object detection results with the data warehouse for further analysis.

## Task 4: Data Warehouse Design and Implementation

The data warehouse is designed to store the cleaned and transformed data in a structured format that supports querying and   
 reporting. The schema is designed to accommodate various types of data such as channel messages, images, and metadata.   
 A scalable solution such as PostgreSQL or BigQuery is considered to ensure that the system can handle large datasets   
 efficiently.

## Task 5: Data Integration and Enrichment

Enriching the data is essential to provide more value to the analysis. This step involves integrating additional data sources   
 to enhance the scraped data. This can include metadata on businesses, geographic information, and other relevant datasets   
 to support more insightful analysis.

## Key Deliverables

1. A functional data scraping pipeline that pulls data from Telegram channels.  
 2. Cleaned and transformed data ready for analysis.  
 3. YOLO-powered object detection for identifying medical products in images.  
 4. A scalable data warehouse that stores all data in a structured format.  
 5. Enriched data with additional context for better decision-making.