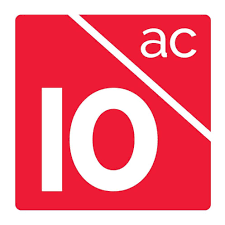
****

**Interim report:**

**10 Academy: Artificial Intelligence Mastery**

**Week 7 Challenge Document**

**Building a Data Warehouse to store data on Ethiopian medical business data scraped from telegram channels**

**By**

**Simone Asnake**

**Date: Feb 01/2025**

**Contents**

[List of Acronyms 3](#_Toc189345717)

[Executive summary 4](#_Toc189345718)

[Statement of the problem 4](#_Toc189345719)

[Scope of the projects 4](#_Toc189345720)

[Project objective 5](#_Toc189345721)

[General: 5](#_Toc189345722)

[Specific: 5](#_Toc189345723)

[Tools and packages/libraries 5](#_Toc189345724)

[Methodologies 5](#_Toc189345725)

[Findings and flow of the task 6](#_Toc189345726)

[Next action points 6](#_Toc189345727)

# List of Acronyms

KAIM: Kifya Artificial Intelligence Mastery

CDR: Challenge document report

# Executive summary

I have tried to focus on the following key components:

1. **Extract the medical business data from telegram app configuration and apply on the scrapper script to get the CSV data and images**
2. **Data Import and Libraries:** importing important libraries and packages analysis.
3. **Data Loading:** Three separate datasets (tg\_data1, tg\_data2, tg\_data3) are loaded from CSV files, representing Telegram data from different channels, each containing information like Channel Title, Channel Username, ID, Message, Date, and Media Path.
4. **Data Exploration:** The loaded datasets are displayed to provide an overview of the information contained in each dataset, showcasing details such as messages, dates, and media paths associated with the respective channels.
5. **Data Pre-processing:** A section on data pre-processing for text is mentioned, which involves text cleaning tasks like removing special characters and I got **198 and 402** missing values **from DoctorET** message and media path, and removing emoji’s.
6. **Tokenization and counting:** Another step involves tokenizing the text data into words for further processing, indicating a focus on text data analysis and natural language processing tasks.

# Statement of the problem

The process of creating a data warehouse for storing data on Ethiopian medical enterprises scraped from web or Telegram channels presents various issues that must be solved successfully, which are outlined below:

* Data Collection Challenges
* Data Quality and Integrity
* Object Detection Integration
* Analysis and Reporting Efficiency

# Scope of the projects

To address the difficulty of Ethiopian medical business data usages and extract essential insightful information, we will largely focus on the following scopes:

* Integrating data from many Telegram channels to provide full coverage of Ethiopian medical enterprises.
* Addressing scalability requirements by creating a data warehouse architecture capable of handling rising volumes of data while preserving performance and efficiency.
* Ensure data cleanliness, consistency, and integrity through effective ETL and ELT processes to allow for reliable analysis and decision-making.
* The scope includes integrating YOLO object detection skills to improve data analysis and extract important insights from data collected on Ethiopian medical businesses.
* Improving data analysis, querying, and reporting to identify trends and patterns for better decision-making in Ethiopia's medical industry sector.

# Project objective

# General:

The overall goal of the week 07 challenge document is to create a robust and scalable data warehouse for storing data on Ethiopian medical businesses collected from web sources or Telegram channels, as well as to integrate object detection using YOLO to improve data analysis capabilities.

# Specific:

* By using ETL (Extract, Transform, and Load) and ELT (Extract, Load, Transform) frameworks to maintain clean, consistent data in the warehouse for analysis.
* By using YOLO object detection to improve data analysis for Ethiopian medical businesses and gain useful insights.
* Improve querying and reporting efficiency to give timely and accurate intelligence to stakeholders.

# Tools and packages/libraries

* pandas
* numpy
* venv
* telethon
* dotenv
* pyspellchecker
* nltk
* nltk.tokenize
* word\_tokenize
* install dbt
* mysql-connector-python
* pymysql
* Jupyter notebook
* Xampp
* Vscode
* Chatgpt
* Deep seek
* Github

# Methodologies

 Data **Import and Libraries:**

* Import necessary libraries such as pandas, numpy, nltk.tokenize, etc..

 Data **extract and Loading:**

* Extract medical business data from telegram by sing telethon packages
* Load separate datasets from CSV files containing Telegram data from different channels.
* Explore the datasets to understand the structure, contents, and patterns present in the data.

 Data **Pre-processing:**

* Missing values
* Removing irrelevant columns
* Removing and cleaning emoji’s

 Text **Tokenization:**

* Implement tokenization techniques to break down the text data into words

 EDA

* Count the number of words
* Count the error spelling of the message files

# Findings

* By applying the above tools and methodologies I got a data inconstancy from Doctor ET scrapping file.
* There content not only focused only on medical business so they had have **content inconsistency** this lead to losing their stockholders but will see after the final analysis.

# Next action points

* Create data based table for scrapped file
* Connect the scrapped file to the database
* Implement YOLO model
* Reporting and documentation