🧾 Interim Report: Kara Data Product Project

Program: 10 Academy – Week 7
Participant: Yamlak Negash
Submission Date: July 13, 2025

✅ Executive Summary

The goal of this project is to build a fully automated, analytical data pipeline using Telegram data for Ethiopian medical businesses. The platform will enable Kara Solutions to answer key business questions such as frequently mentioned products, pricing trends, and image-based content prevalence using modern ELT practices.

To achieve this, a robust end-to-end pipeline is being constructed using:

- **Telethon** for Telegram scraping
- **PostgreSQL** for the data warehouse
- **dbt** for transformations
- **YOLOv8** for object detection on product images
- **FastAPI** for analytical API exposure
- **Dagster** for orchestration and scheduling

This report details progress made on **Task 0 (Setup)**, **Task 1 (Scraping)**, and **Task 2 (Transformation)**.

🧰 Task 0: Project Setup and Environment Management

A well-structured repository was initialized with the following:

- ✅ Git repository and `.gitignore`
- ✅ Docker environment for PostgreSQL and backend
- ✅ `requirements.txt` for Python dependencies
- ✅ `.env` file for API keys and credentials (excluded from Git)
- ✅ `python-dotenv` used for environment management

plaintext
CopyEdit

```
kara_data_product/
├── .env                # Contains API and DB credentials
├── Dockerfile
├── docker-compose.yml
├── requirements.txt
├── .gitignore
└── README.md
```

Docker services were tested to ensure PostgreSQL and scripts launch successfully.

🛰️ Task 1: Telegram Data Scraping and Collection

Tool Used: Telethon
Data Source: Ethiopian medical Telegram channels (e.g., Lobelia, Chemed)

Pipeline Steps Implemented:

- ✅ Connected to public Telegram channels
- ✅ Collected messages with metadata: `text`, `channel`, `timestamp`, and `media presence`

✅ Stored raw data as JSON in a timestamped, channel-partitioned structure:

bash

CopyEdit
`data/raw/lobelia4cosmetics/2025-07-12.json`

-

Logging was included to capture scraping status and error handling (e.g., rate limits, missing fields).

Sample schema:

json
CopyEdit

```
{
  "channel": "lobelia4cosmetics",
  "date": "2025-07-12T14:33:00",
  "message": "ጥሩ ንፁህ አበሳ ነው።",
  "media": {
    "has_image": true,
    "file_path": "path/to/image.jpg"
  }
}
```

🧱 Task 2: Data Modeling and dbt Transformation

Warehouse: PostgreSQL (via Docker)
Transformation Tool: dbt

✅ Setup Progress:

- Initialized `dbt` project and PostgreSQL adapter
- Created schemas: `raw`, `staging`, and `marts`

✅ Models Implemented:

📦 Staging Model

`stg_telegram_messages.sql`

- Extracts and normalizes fields like `text_length`, `channel`, `has_image`

📊 Marts Models

- `dim_channels.sql` – Metadata about channels
- `dim_dates.sql` – Date breakdown (weekday, hour, month)
- `fct_messages.sql` – Message facts with metrics

✅ Tests:

- Implemented `not_null` and `unique` tests on primary keys
- One custom test: Check if any message has `NULL text` in `stg_telegram_messages`

📌 Tools & Libraries Used

Tool	Purpose
Telethon	Telegram scraping
Python	Main scripting language
Docker	PostgreSQL + Python containerization
dbt	Transformation layer
pandas	Exploratory and feature engineering
psycopg2	PostgreSQL integration
dotenv	Credential management

📉 Early Insights

- Majority of channels post during afternoon hours (12PM–5PM)
- Image-based posts are frequent in cosmetics/pharmaceutical channels
- Message length varies by channel — some are heavy on visuals, others on descriptions

---

🔁 Next Steps

- Implement YOLOv8 enrichment to detect pills/creams in product images
- Expose final tables via FastAPI for analytical insights
- Automate pipeline using Dagster
- Conduct pipeline scheduling, testing, and dashboard reporting

---

📎 GitHub Repo Link

✅Interim Github Link