10 academy

November 2022

Text-to-speech data collection

Data Engineering Project

@thamar-niyomukiza

Outline

01

Business Objectives

02

EDA

03

Implement
ation

04

Conclusion and Future plan

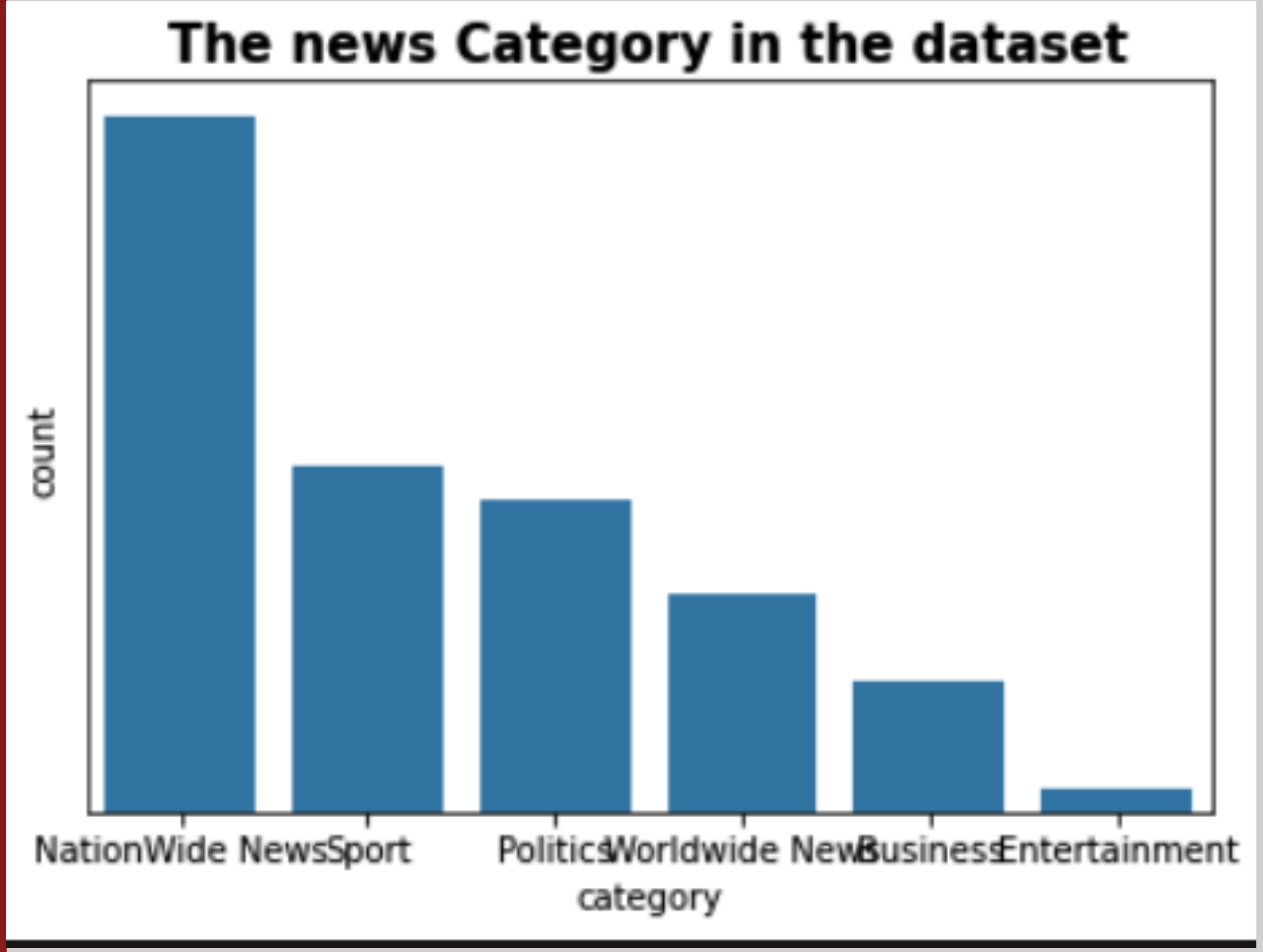
Objective

Design and build a robust, large-scale, fault-tolerant DE pipeline that

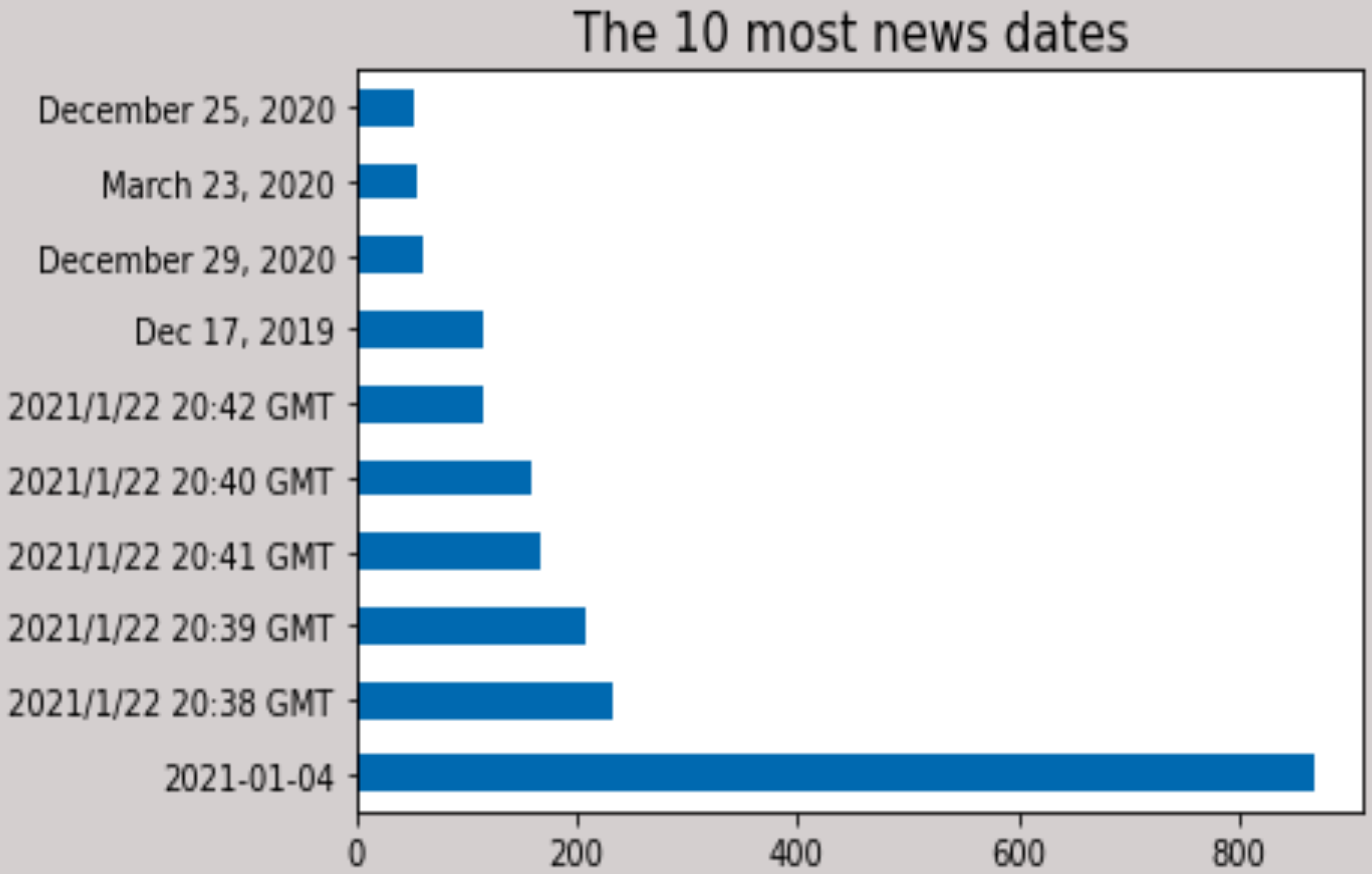
  - allows post text and record millions of Amharic and Swahili speakers reading digital texts in apps and web platform.

- Apply transformation in a distributed manner, and load it into a warehouse in a suitable format to train a speech-to-text model

Exploratory Data Analysis



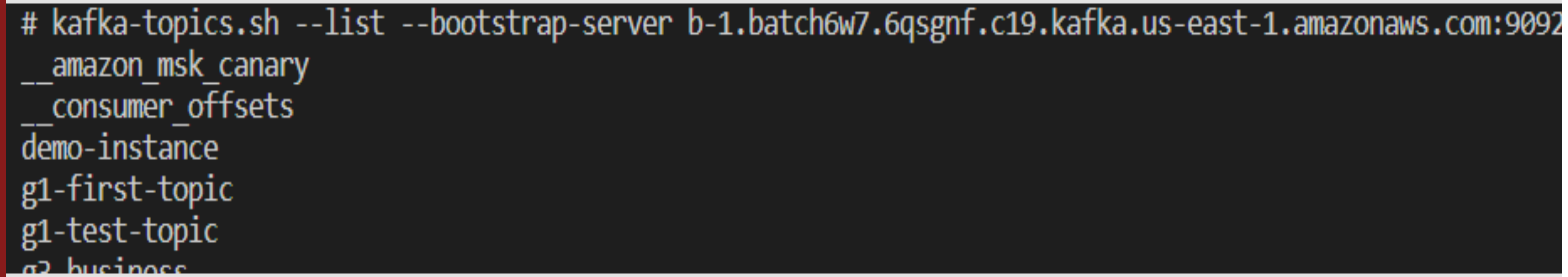
The news Category in the dataset



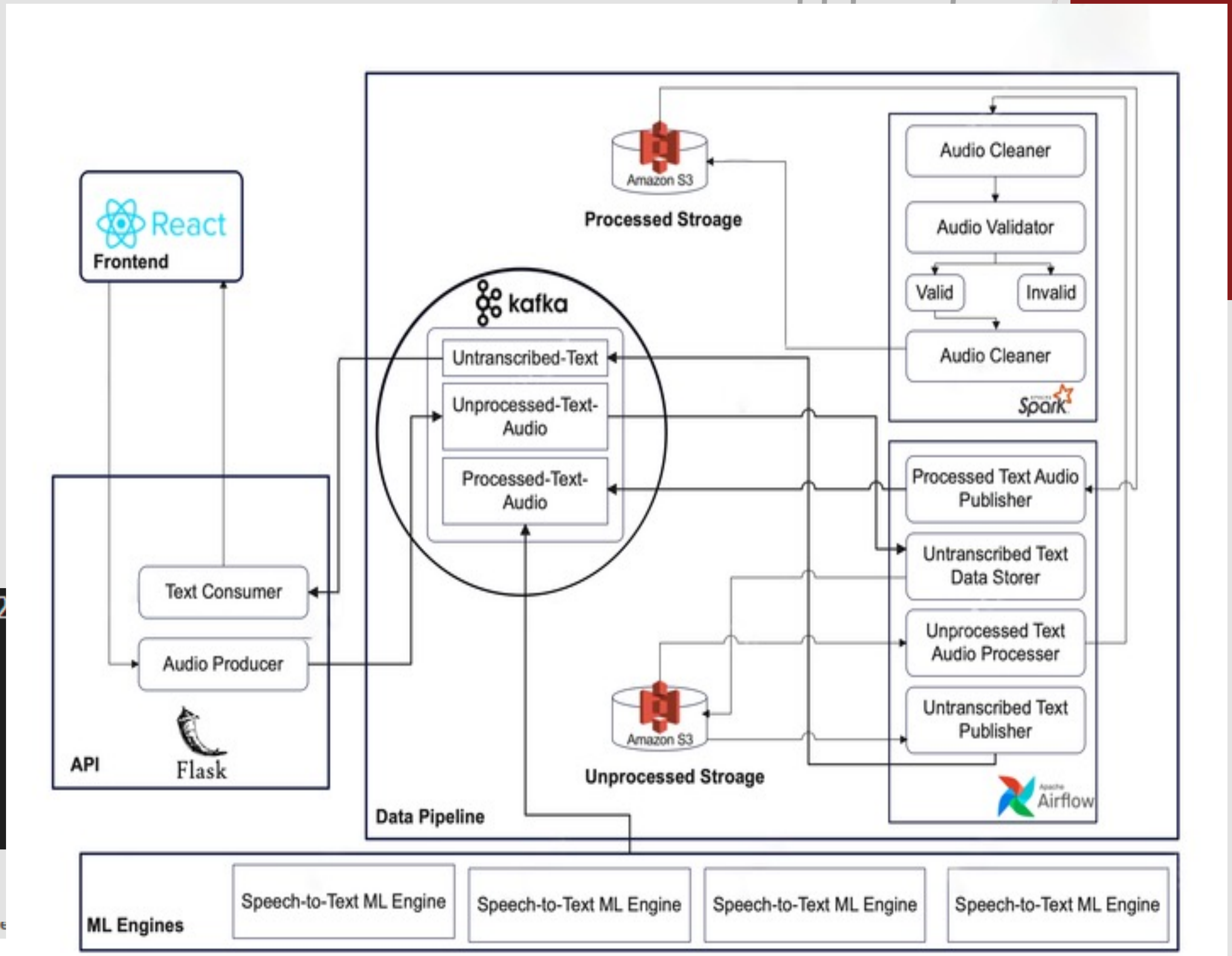
The 10 most news dates

Back end

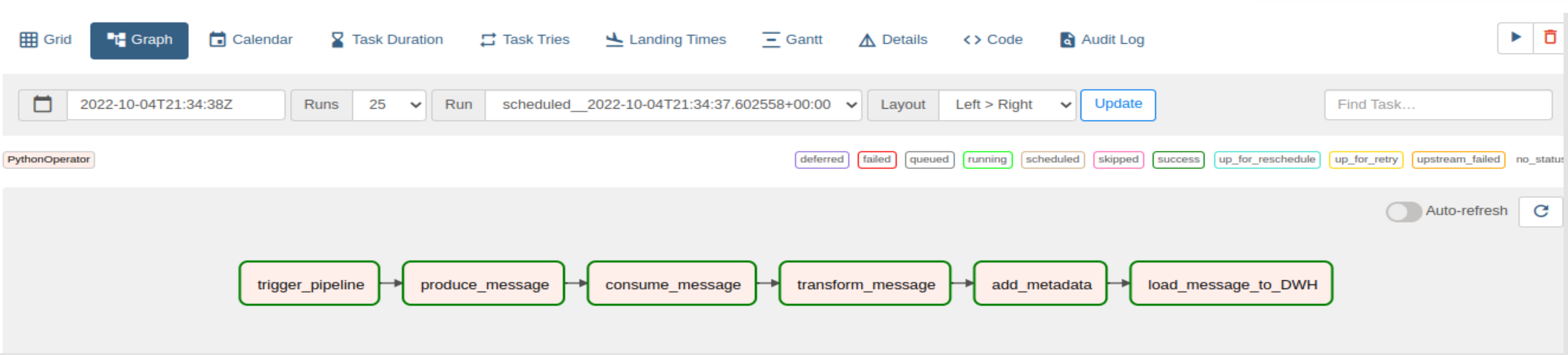
- Backend accepts the recorded audio and sends that to the Kafka cluster.
- Airflow schedules the Kafka consumers to consume the latest messages
- then saves to the unprocessed storage to aws s3



```
# kafka-topics.sh --list --bootstrap-server b-1.batch6w7.6qsgnf.c19.kafka.us-east-1.amazonaws.com:9092
__amazon_msk_canary
__consumer_offsets
demo-instance
g1-first-topic
g1-test-topic
g2-business
```

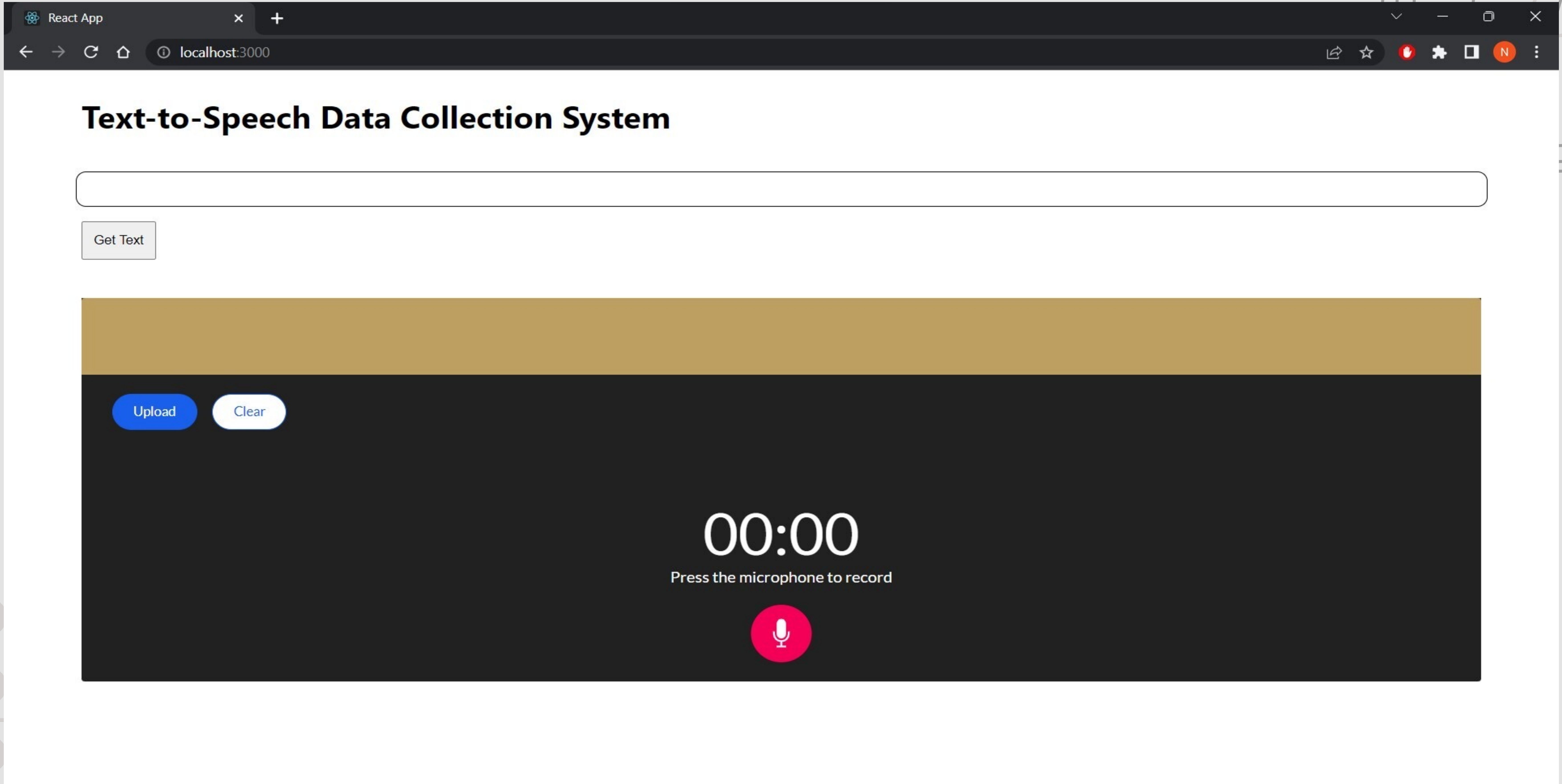


DAG: KAFKA_CLUSTERS_data_pipeline A data Extraction and loading pipeline for week 6 of 10 academy project success Sche



2022-10-04T21:34:38Z Runs 25 Run scheduled__2022-10-04T21:34:37.602558+00:00 Layout Left > Right Update Find Task...

Front end



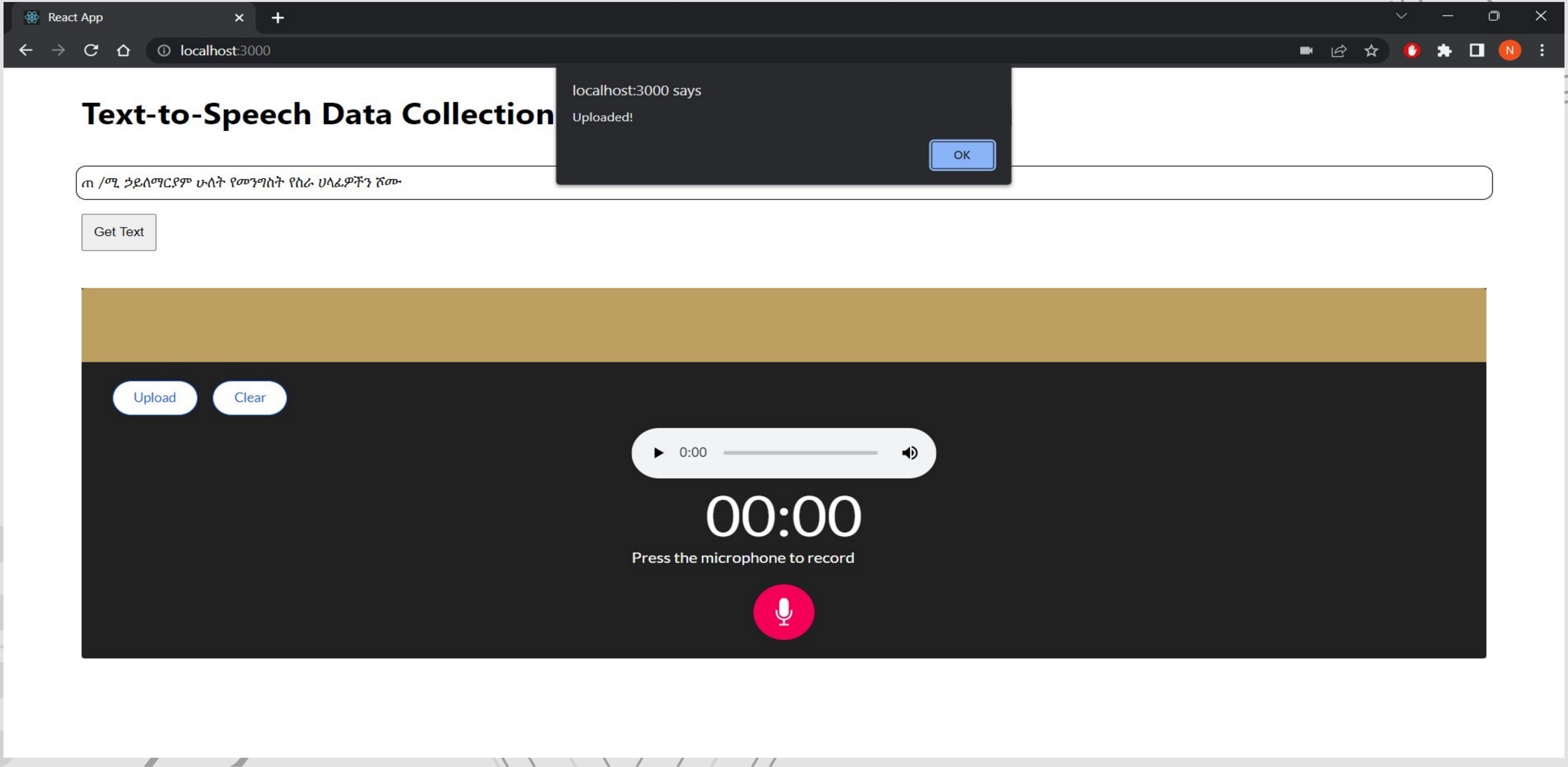
Text-to-Speech Data Collection System

00:00

Press the microphone to record

Front end

The Frontend asks the user to insert the recording for their given news article. As we can see here the user will record and upload the audio.



Text-to-Speech Data Collection

localhost:3000 says

Uploaded!

ጠ /ሚ ኃይለማርያም ሁለት የመንግስት የስራ ሀላፊዎችን ሾሙ

00:00

Press the microphone to record

Future Plan

- Schedule tasks to send the unprocessed data to Kafka for further transformation

- Then save the processed audio and the text in a separate file with same file names

November 2022

THANKYOU

@Thamar-niyo