Data Engineering: text-to-speech data collection with Kafka, Airflow, and Spark

Introduction

In short, data engineers are responsible for designing, developing, and maintaining the data platform, which includes the data infrastructure, data applications, data warehouse, and data pipelines. Lack and quality of data are two of the most important things to consider while building a model. So the purpose of this project is to build a data engineering pipeline that allows recording millions of Amharic and Swahili speakers reading digital texts in-app and web platforms.

Objective of the Project

The main objective of this project is to produce a tool that can be deployed to process posting and receiving text and audio files from and into a data lake, apply transformation in a distributed manner, and load it into a warehouse in a suitable format to train a speech-to-text model.

Data

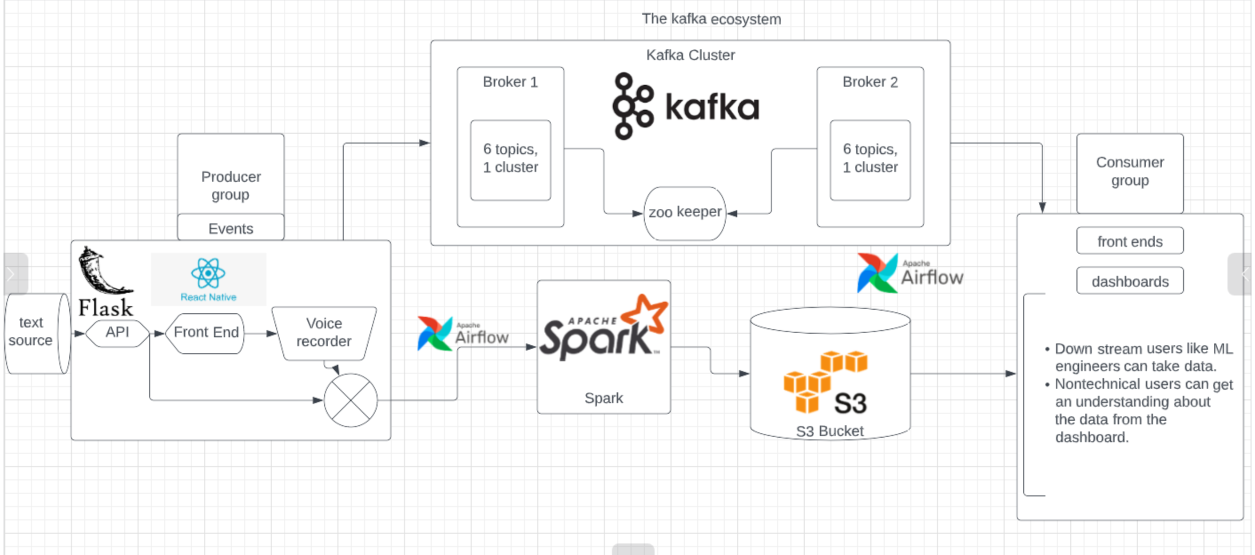
For this project, Amharic news text classification dataset with baseline performance dataset is used. You can find the dataset here.

Data understanding

From the dataset collected for each row there are 6 attributes which are:

- Date:- date it was published
- Views:- how many views it got
- Total data: 50706 articles
- Link :- from which link it was taken from
- Category:- what category is the news in
- Article: the content of the news
- Headlines — headlines of the news

Project Pipeline



Technologies and tools used to build the data capture pipeline

Frontend – is how we interact with users to upload an audio file or validate audio file submissions. The project frontend is developed using ReactJs. This ReactJs is an open-source, declarative, efficient, and flexible JavaScript library for building reusable UI components.

Backend - It is the infrastructure that supports the front end and is made up of parts of a piece of software regular users can't see. The backend is also called server side and is basically a website's brain**.** Django and python were used to develop the backend of this project. Django is a high-level Python web framework that encourages rapid development and clean, pragmatic design.

Apache Kafka:

Apache Kafka is a distributed data store optimized for ingesting and processing streaming data in real-time. Streaming data is data that is continuously generated by thousands of data sources, which typically send the data records in simultaneously. It combines messaging, storage, and stream processing.

Why would you use Kafka?

Kafka is used to build real-time streaming data pipelines and real-time streaming applications that adapt to the data streams. A data pipeline reliably processes and moves data from one system to another, and a streaming application is an application that consumes streams of data. So in our case, Kafka used for speech to text data collection (to store the unprocessed audio text in a fault-tolerant, durable way) because of its high throughput, high scalability, low latency, permanent storage and high availability.

Apache Airflow:

Apache Airflow is an open-source tool to programmatically author, schedule, and monitor workflows. Directed acyclic graphs (DAG) are used by Airflow to control workflow orchestration. In this project, it schedules spark jobs as well as Kafka cluster jobs at the same time.

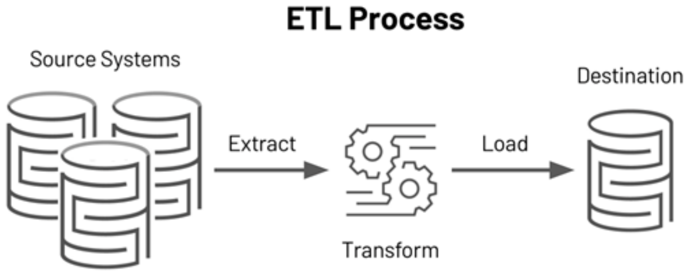
Apache Spark:

Apache Spark is a multi-language engine for executing data engineering, data science, and machine learning on single-node machines or clusters. Apache Spark is the fast, flexible, and developer-friendly leading platform for large-scale SQL, batch processing, and stream processing. So in our case, we use spark with airflow scheduler to load, process and transform streams of audio data.

To develop this project we will apply the ETL (Extract, Transform, and Load) approach. ETL is a data integration process that combines data from multiple data sources into a single, consistent data store that is loaded into a data warehouse or other target system.

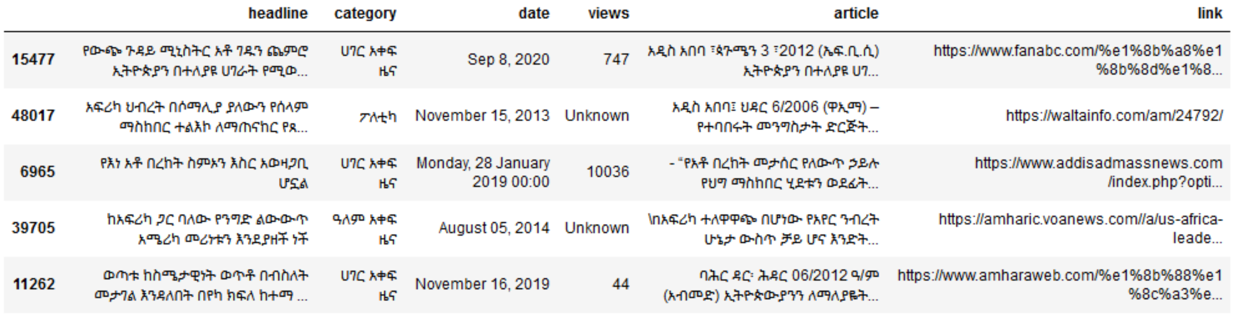
How ETL works?

During data extraction, raw data is copied or exported from source locations to a staging area. Data management teams can extract data from a variety of data sources, which can be structured or unstructured. In the staging area, the raw data undergoes data processing. Here in transformation, the data is transformed and consolidated for its intended analytical use case. During the load phase, the transformed data is moved from the staging area into a target data warehouse. Typically, this involves an initial loading of all data, followed by periodic loading of incremental data changes and, less often, full refreshes to erase and replace data in the warehouse.



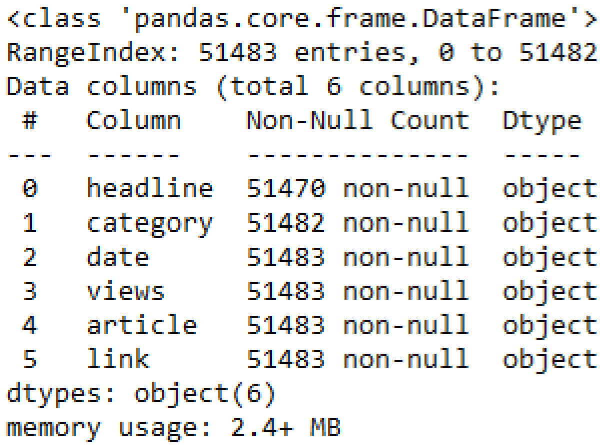
Data exploratory Analysis

Overview of the dataset



	headline	category	date	views	article	link
15477	የውጭ ጉዳይ ሚኒስትር አቶ ገዱን ጨምሮ ኢትዮጵያን በተለያዩ ሀገራት የሚወ...	ሀገር አቀፍ ዜና	Sep 8, 2020	747	አዲስ አበባ ፣ጳጉሜን 3 ፣2012 (ኤፍ.ቢ.ሲ) ኢትዮጵያን በተለያዩ ሀገ...	https://www.fanabc.com/%e1%8b%a8%e1%8b%8d%e1%8...
48017	አፍሪካ ህብረት በሶማሊያ ያለውን የሰላም ማስከበር ተልእኮ ለማጠናከር የጸ...	ፖለቲካ	November 15, 2013	Unknown	አዲስ አበባ፤ ህዳር 6/2006 (ዋኢማ) – የተባበሩት መንግስታት ድርጅት...	https://waltainfo.com/am/24792/
6965	የእነ አቶ በረከት ስምኦን እስር አወዛጋቢ ሆኗል	ሀገር አቀፍ ዜና	Monday, 28 January 2019 00:00	10036	- "የአቶ በረከት መታሰር የለውጥ ኃይሉ የህግ ማስከበር ሂደቱን ወደፊት...	https://www.addisadmassnews.com/index.php?opti...
39705	ከአፍሪካ ጋር ባለው የንግድ ልውውጥ አሜሪካ መሪነቱን እንደያዘች ነች	ዓለም አቀፍ ዜና	August 05, 2014	Unknown	\nአፍሪካ ተለዋዋጭ በሆነው የአየር ንብረት ሁኔታ ውስጥ ቻይ ሆና እንድት...	https://amharic.voanews.com//a/us-africa-leade...
11262	ውጣቱ ከስሜታዊነት ወጥቶ በብስለት መታገል እንዳለበት በየካ ክፍለ ከተማ ...	ሀገር አቀፍ ዜና	November 16, 2019	44	ባሕር ዳር፡ ሕዳር 06/2012 ዓ/ም (አብመድ) ኢትዮጵያውያንን ለማለያዬት...	https://www.amharaweb.com/%e1%8b%88%e1%8c%a3%e...

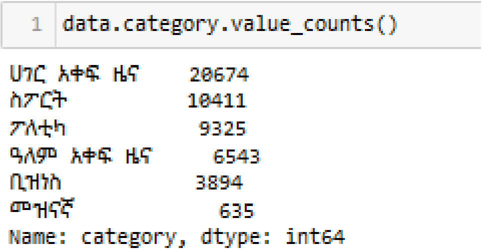
Columns and their Data Types



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 51483 entries, 0 to 51482
Data columns (total 6 columns):
 #   Column    Non-Null Count  Dtype 
---  ------    --------------  ----- 
 0   headline  51470 non-null  object
 1   category  51482 non-null  object
 2   date      51483 non-null  object
 3   views     51483 non-null  object
 4   article   51483 non-null  object
 5   link      51483 non-null  object
dtypes: object(6)
memory usage: 2.4+ MB
```

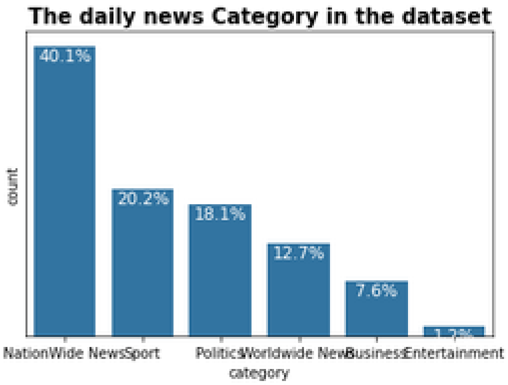
During data pre-processing phase, we discovered that there are no missing values and that all of the fields contain only string object type data.

The number of category in the dataset



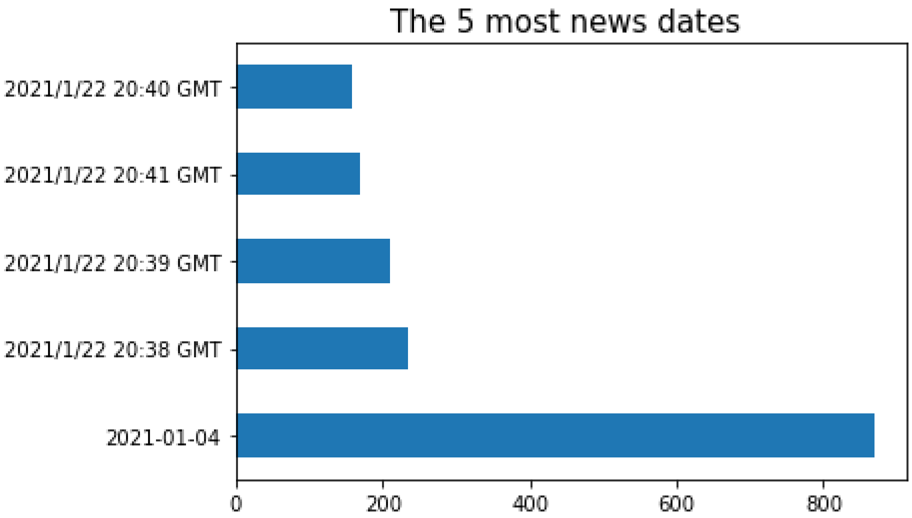
```
data.category.value_counts()
```

```
ሀገር አቀፍ ዜና     20674
ስፖርት            10411
ፖለቲካ             9325
ዓለም አቀፍ ዜና      6543
ቢዝነስ             3894
መዝናኛ              635
Name: category, dtype: int64
```



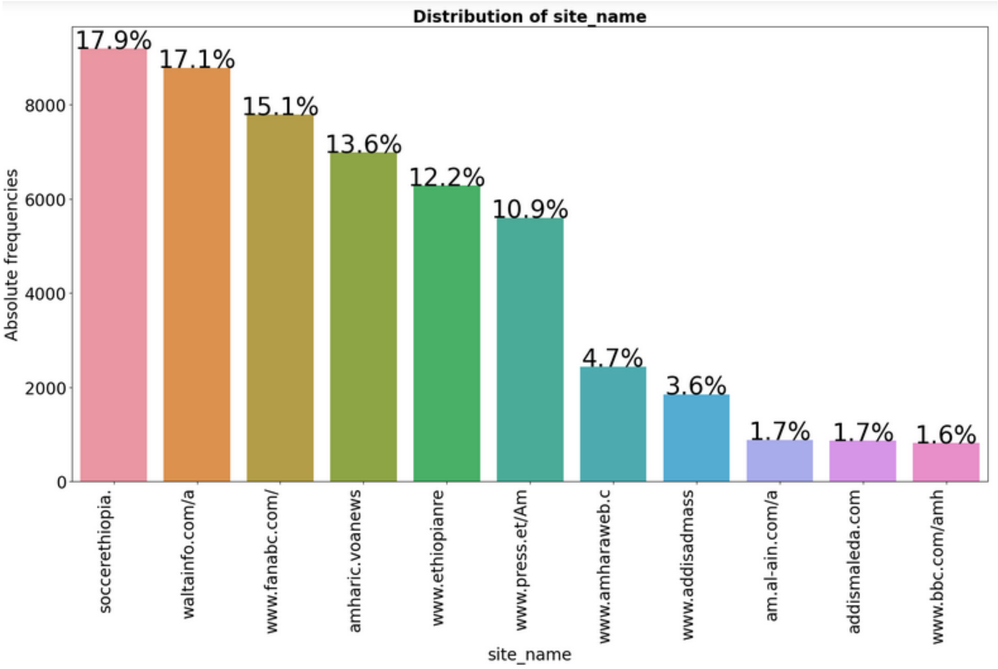
From this plot, we observe that the most daily news are nation wide news(ሀገር አቀፍ ዜና) and sport news (ስፖርት) is on the second rank.

Distribution of dates in terms of broadcasted news.



The above insight depicts that the five most news date. So from this we can conclude that, most of the dataset's news is broadcasted on 2021-01-04.

Distribution of site names



This insight shows that, soccer Ethiopia is the frequently posted site and waltainfo.com/ is on the second level.

What was implemented?

As we can observe from the above project pipeline, we started our implementation with the development of the front-end part of the system and with configuration of the require tools and packages. The frontend was a react based user interface that the users would interact with. Generally, in this project we implement a tool that allows user to read text/sentences, record the audio of those text/sentences and store it on s3 bucket for further speech-to-text model training.

For more detail understanding you can recap the tools we used to develop this project and the project pipeline.

Lessons learned

From this project we have been able to create and maintain an Apache Kafka cluster, work with Apache Airflow and Apache Spark, apply structured streaming to process streaming data, building data pipelines and orchestration workflows. Specifically:

- Amazon Web Service AWS
- What, why and how to work with Apache Kafka, Apache Airflow and Spark.
- ReactJS programing languages(able to improve our skill).
- Data collection approaches

Future plans

- Having more data helps machine learning model to increase its' learning capability. So to develop an accurate machine learning model we have planned to collect more data.
- Applying other approaches on the project pipeline to compare their quality and performance to choose the efficient one.
- Adding more functionality to make it complete: due to time constraints and problems of device configuration (unable to access the instances), we could not finish everything we wanted to do on this project.
- Improve the frontend to be user-friendly and attractive for real world use

Reference

https://github.com/IsraelAbebe/An-Amharic-News-Text-classification-Dataset
https://kafka.apache.org/documentation/#configuration
https://towardsdatascience.com/why-apache-airflow-is-a-great-choice-for-managing-data-pipelines-48effcce3e41
https://medium.com/vantageai/keeping-your-ml-model-in-shape-with-kafka-airflow-and-mlflow-143d20024ba6
https://sparkbyexamples.com/spark/spark-streaming-with-kafka/
https://airflow.apache.org/docs/apache-airflow/stable/index.html#

Github repo to the dapp

https://github.com/Choquet-Bruhat/Text-to-speech-data-collection-with-Kafka-Airflow-and-Spark

CONTRIBUTERS

- Kibatu Woldemariam
- Michael Getachew
- Josias Ounsinli
- Genet Shanko.
- Amanuel Zewdu
- Tegisty Hailay