About Me

- A Random ML Engineer from DRCongo🇨🇩, But Lost in UK!
- By Day: Machine Learning Engineer, LBG London!
- By Night: NLP Engineer, I started learning NLP before ChatGPT. 👌🏾
- Almost 10 years of Programming Experience 💪🏾.
- 🐦‍⬛ On Twitter: esp_py.

Balobi Nini: A news summarizer on a budget!

by **Espoir Murhabazi Buzina**

> And when you get software right, something magical happens: You don't need hordes of programmers to keep it working. You don't need massive requirements documents and huge issue tracking systems. You don't need global cube farms and 24/7 programming.!
>
> - Rob Martin, Clean Architecture

Presentation Outline

Why Did I build this?

In this presentation, we will learn how I use :

- Traditional Machine Learning Methods,
- LLM

To build a fully fledge application on a budget without using any cloud provider by self hosting all the components.

I will also describe how much I spend so far to host the components.

All of this was done for learning Purpose.

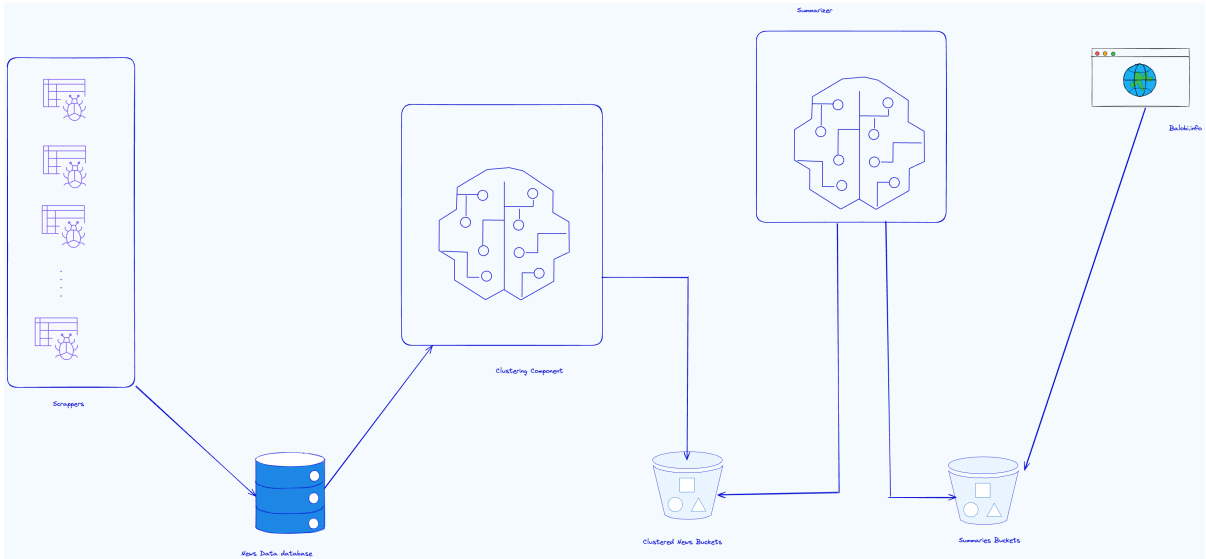
Balobi Nini

What are they talking about in Lingala, is a application that gives a daily summary of news happening in Congo.

How is that Done?

- Aggregates news from different news website from congo. News are in **French**!
- Group them in categories according to topics.
- For each topic it use a Language Model to generate news summaries.
- A front end in VueJs that display application.

The Application Architecture



Components

- Scraper
- Clustering
- Language Model
- Front end

Every component is enabled in a Docker image and run on a standalone docker machine! All the docker images were deployed on Docker hubs for free for now.

Hardware

Let talk about the hardware, mostly VPS product. I tried to avoid using cloud providers because the end goal will be to self host the application.

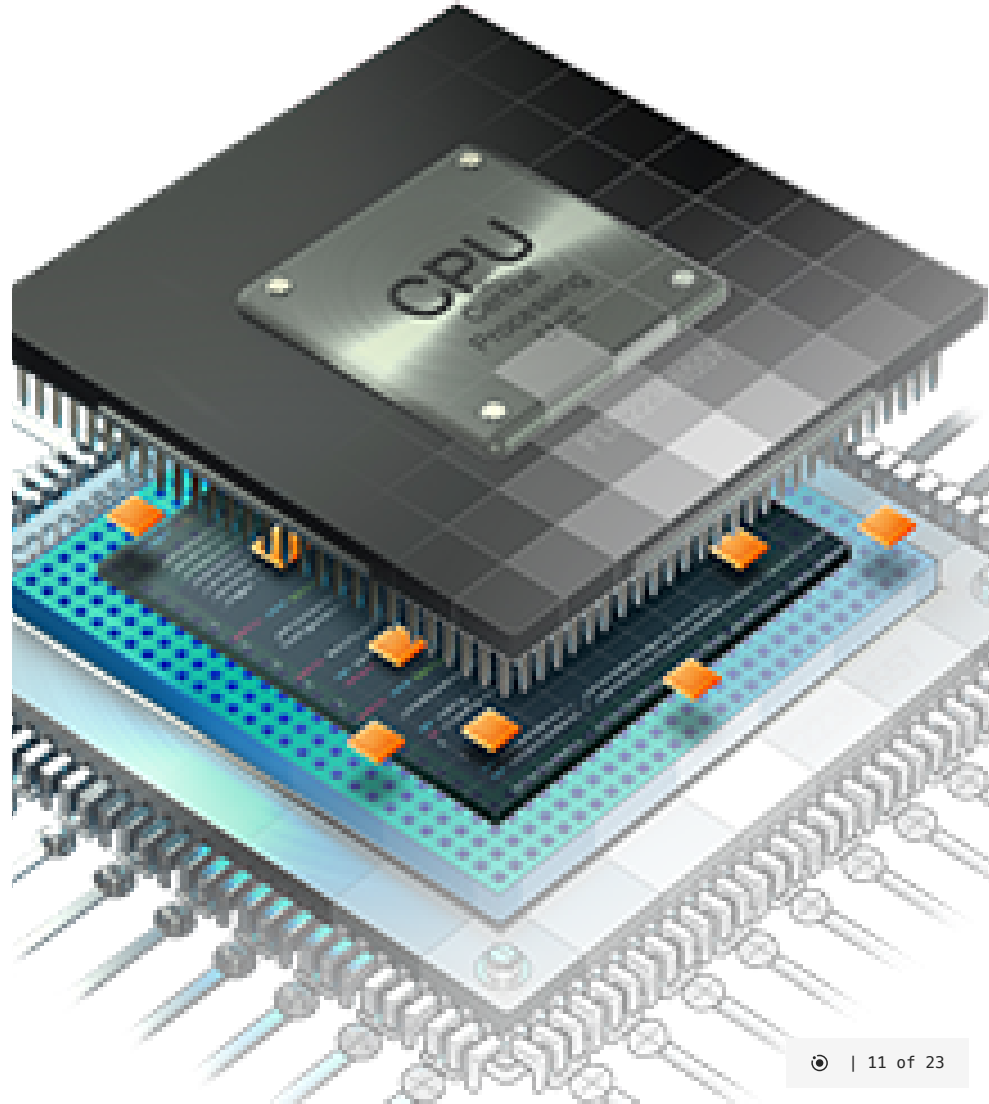
Relational Database

- Postgres Data: It is a custom database hosted on a VPS from Racknerd.
- Spec: 500 Mb RAM
- Software: Linux Alpine
- CPU: : Intel® Xeon® CPU E5-2690 v4 @ 2.60GHz
- Price 💸: 25 USD per year! 2.5 USD per month(Black Friday Deal)

First Linux VPS:

A linux VPS that run the scraper and The clustering pipeline

- RAM: 8Gb of RAM
- Software: Linux Ubuntu
- CPU: Intel Xeon Processor (Skylake, IBRS, no TSX), 2Ghz
- Price: 10 USD per month: 120 USD per year.
- Provider: Hetzner Cloud



Second Linux VPS:

A linux VPS that run the LLM and the generator app

- Spec: 16 Gb of RAM
- Software: Linux Ubuntu
- CPU: AMD 2.GHZ 16-Core 🤔 (It is a VPS)
- Price: 12.50 USD per month, 144 USD per year.
- Provider: Contabo

Object Storage

An S3 compatible API storage where I save blob and model data.

- Provider: BlackBlaze
- Cost: Free

Components

Now let us talk about the components.

Scrapper

- It is build with Python and Scrapy. We have more than 10 spiders, each spider is for a separate news website.
- Where does it Run: It runs on the First Linux VPS and save the data in the Postgres Database.
- Tools used: Scrappy, ScrappyD, Celery

Clustering Component

- For the clustering we use two main components, an embedding model and clustering algorithm.
- For the embedding I am using an embedding model called Stella which is based on Qwen2-1.5B instruct. It has 400 Millions parameter and have 1.5Gb and it quite good for French documents. It also uses sentence transformer to deploy the model. **Not langchain**
- For the modeling I used hierarchical clustering with plain Scipy and Numpy, **Not sklearn.** I have a blog on the clustering approach on my website.

sizes

- Model: 400M parameters without Quantization: 1.5Gb.
- Docker Image: 1.5Gb So far the largest one because of Sentence transformer and Pytorch.

Deployment Spec:

- Hardware use: First Linux VPS
- Database Postgres Database
- Object Storage.

The new summarizer

An LLM: **Qwen-Instruct 1.5B with 8bits quantization**. Best for it size! I have used **llama.cpp** to host the model and build the summarizer pipeline using requests. The results are saved in a bucket.

Sizes:

- Model: 1.5 Gb 8 bits quantization
- Llama.cpp image: 139Mb for the docker image, Generator: 172 Mb

Deployment Spec:

- Hardware Used: Second Linux VPS
- Cloud Storage

Performance:

```
| model                          |       size |     params
| ------------------------------ | ---------: | ---------:
| qwen2 1.5B Q8_0                |   1.53 GiB |     1.54 B
```

Prompt:

```
"""
Give a title and a short summary 2 to 3 sentences in french
{{content}}
 Describes it in a style of a french new paper reporters.

Don't summarize each document separately, the content in al


The answer should have the format:

Titre:
Résumé:

The title and summary should be in french not in English.
"""
```

The front end

The front end is a Vue JS application that run on cloudflare, for free .

It read the summarizer and display them on the front end.

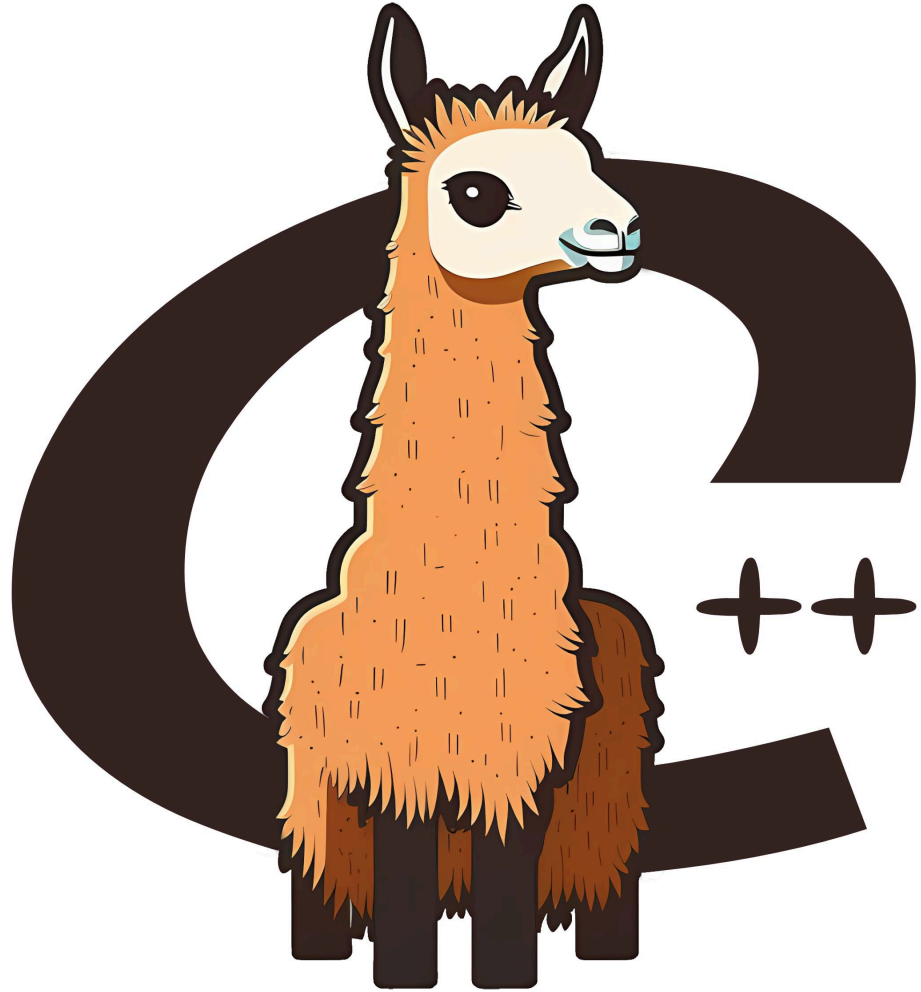
Deployment

CloudFlare: For free, they handle traffic and CDN for me. Domain Name: 17 USD for two year.

The app: Balobi.info

Next Steps

- Get Rid of the One VPS and Keep everything on one VPS
- Use a lightweight version of Kubernetes and Install Flyte to ochestrate the platform
- Use a smaller model Qwen 0.5 and Finetune it on the Synthetic data I have generated so far.
- A lot of work on the user interaction to generate buisness values from the application



Summary

A full functioning Application that cost: **255 USD per year.**

> When software is done right, it requires a fraction of the human resources to create and maintain. Changes are simple and rapid. Defects are few and far between. Effort is minimized, and functionality and flexibility are maximized.
>
> - Rob Martin, Clean Architecture

My Pledge 🙏🏾: Support Goma