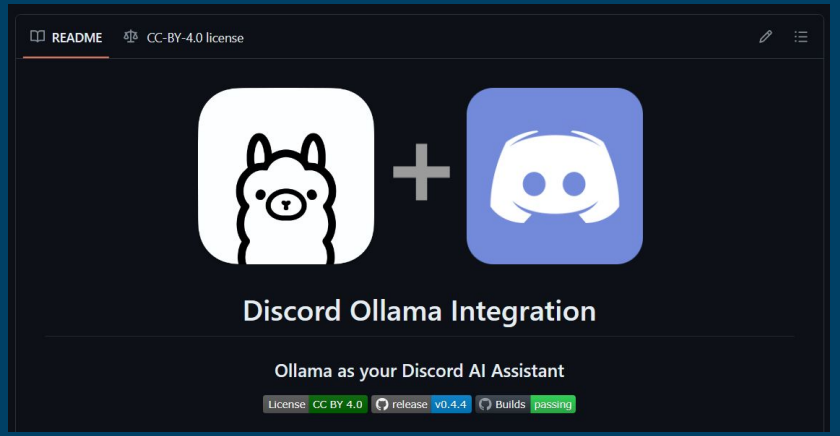
LLM Context Pipeline

CPSC 324 - Big Data Analytics
Jonathan Smoley

Goal

- **Discord-Ollama**: open-source LLM chat bot
- **Provide Purpose**: take a cool project and give it a clear direction to move in
- **Automation**: simplify data collection for model file context



Examples

Basic `Modelfile`

An example of a `Modelfile` creating a mario blueprint:

```
FROM llama3
# sets the temperature to 1 [higher is more creative, lower is more coherent]
PARAMETER temperature 1
# sets the context window size to 4096, this controls how many tokens the LLM can use as context to generate the next token
PARAMETER num_ctx 4096

# sets a custom system message to specify the behavior of the chat assistant
SYSTEM You are Mario from super mario bros, acting as an assistant.
```

To use this:

1. Save it as a file (e.g. `Modelfile`)
2. `ollama create choose-a-model-name -f <location of the file e.g. ./Modelfile>'`
3. `ollama run choose-a-model-name`
4. Start using the model!

Data Pipeline

Data Collection

- Web scrape www.gonzaga.edu
  - Alternative: University API
- Run on fixed interval
- Extract text and deeper links from accessible domains

Data Preparation

- Pull when current data is available
- Extract names, places, things, etc.
- Form contextual clues from information on Gonzaga University
- Standardize clues into a model file to be used in “training” an LLM

The Data Itself

During Web Scraping:

- Found in HTML documents available at www.gonzaga.edu
- HTML to Text, stored as separate instances found in documents

During Processing:

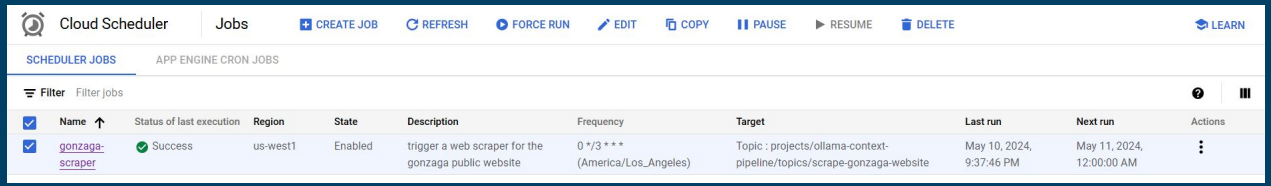
- Clue type, context found in, datetime found, etc.
- Fit for document storage (script of context clues)

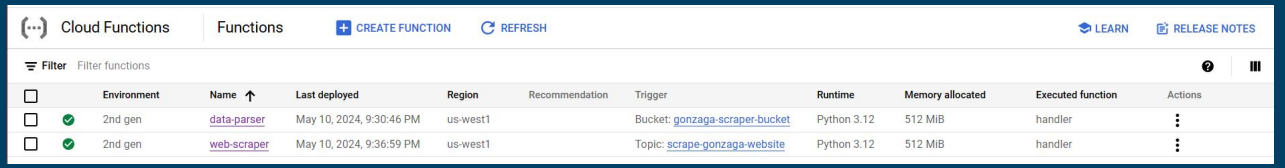
Challenges:

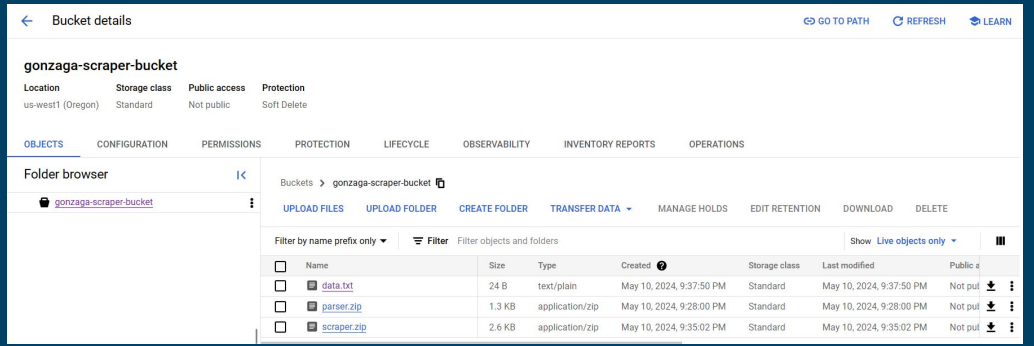
- Not easy to recognize which text is valuable to an LLM, at least with code alone
  - An addition of NLP analysis would come in handy here

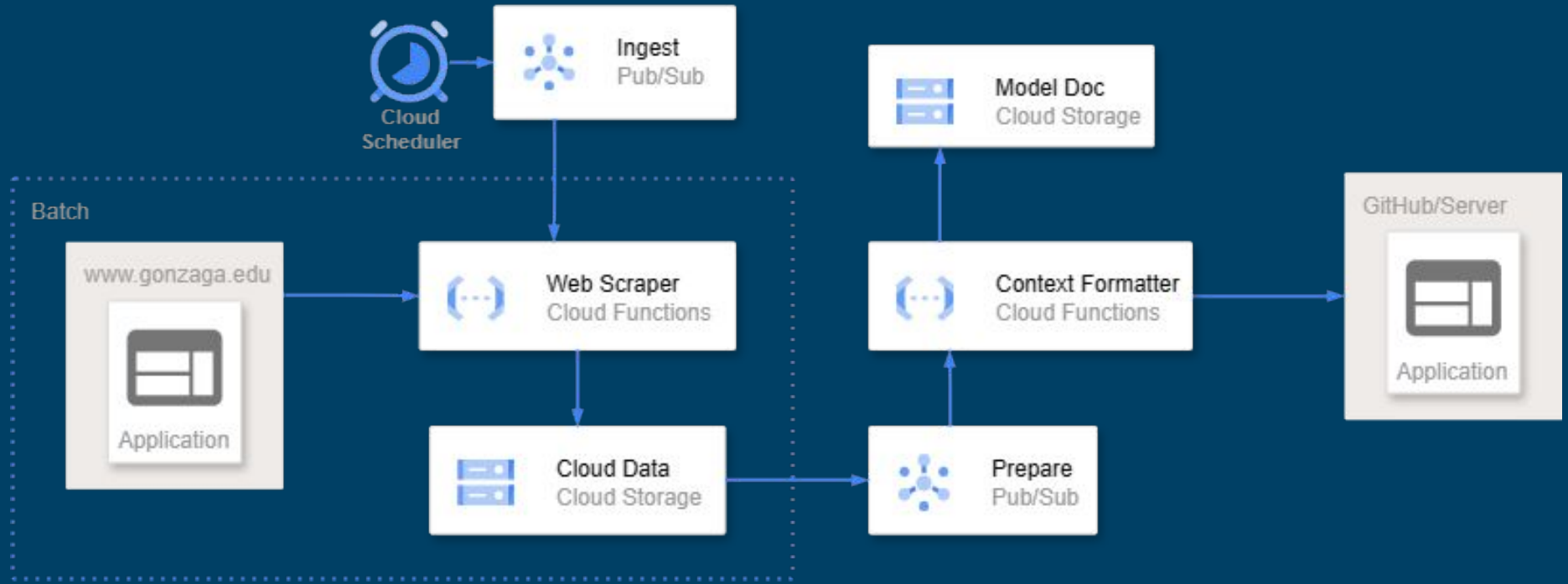
Services Involved

- Cloud Scheduler
- Pub/Sub
- Cloud Functions
- Cloud Storage








Cloud
Scheduler
Ingest
Pub/Sub
Model Doc
Cloud Storage
Batch
www.gonzaga.edu
Application
Web Scraper
Cloud Functions
Context Formatter
Cloud Functions
GitHub/Server
Application
Cloud Data
Cloud Storage
Prepare
Pub/Sub

Demo

- Services working together
- Existing scraper work
- Automation

Future Work

Short-Term

- Finalize web scraper
  - Assess typical document structure on www.gonzaga.edu
  - Choose elements to scan for
- Decide a final destination for results
  - Cloud Storage -> GitHub
  - Model File -> GitHub

Long-Term

- Build out parser
  - ML/NLP solution to look for keywords in text collected
- Set up model file (context clues)
  - Design sentence structure

Resources

- Ollama: https://ollama.com/
- Discord-Ollama: https://github.com/kevinthedang/discord-ollama