Team - ;DROP TABLE

Team:
David B.A.De V. Velho
Rohit Ramesh Patekar

College - Vellore Institute of Technology, Vellore

Graduating Year - 2022

Meet the Team

DAVID B.A.DE V. VELHO

ROHIT RAMESH PATEKAR

About Ourselves

Our Participation

We've participated in various hacks like MLH IvyHacks, Eng Hack 2021, VIT Hack, Devspace, Devsoc, Women Techies, etc

Acolades and Awards

- SAMSUNG Prism research fellow
- TVS Defect detection pilot project
- Various certification courses

Our Projects

We've worked on various projects together -

(kindly view the next page)

Our Projects

To name a few

POSTBABY

Cross platform open source Postman alternative, written in C++ and OpenGL. Focused on Speed and portability

UNIFYPDF

Client side PDF Merger. No external Servers. Privacy oriented. All conversion done in browser.

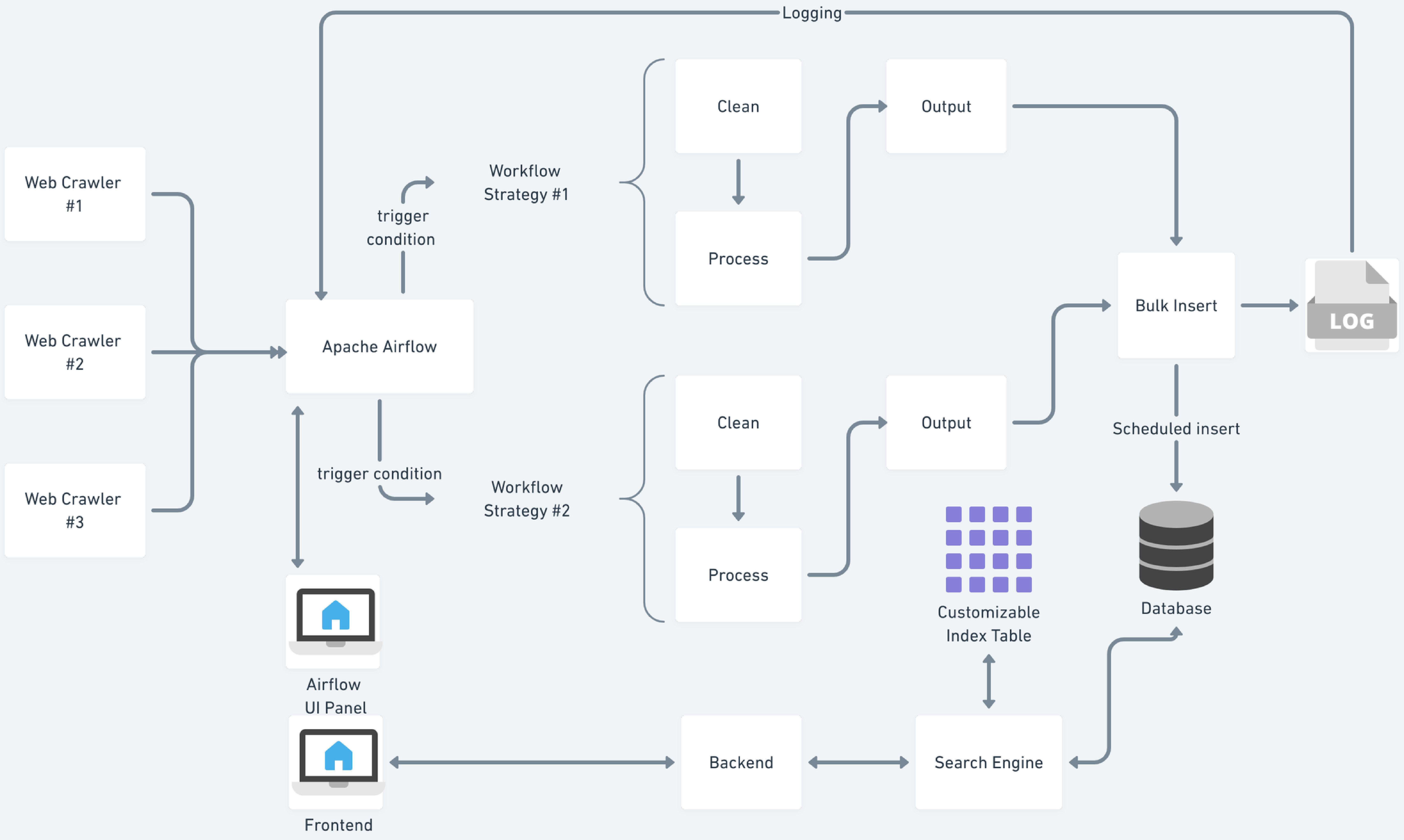
DEVCONNECTOR

A social media website for developers to interact. Made with ReactJs

Problem Statement

Website Search

OVERVIEW


Logging
Web Crawler #1
Web Crawler #2
Web Crawler #3
Apache Airflow
trigger condition
Workflow Strategy #1
trigger condition
Workflow Strategy #2
Clean
Process
Output
Clean
Process
Output
Bulk Insert
LOG
Scheduled insert
Database
Customizable Index Table
Airflow UI Panel
Frontend
Backend
Search Engine

Overview

01

Web Crawlers scrape websites

Scraping performed on user defined strategy. Data submitted to Apache Airflow

02

Workflow triggers

Apache Airflow triggers workflows depending on the selected user strategy and input data, which is then cleaned

03

Bulk Insert

We bulk insert the data during low load and enable logging and retry on failure

Overview

04

Search Engine

The Search Engine queries the DB for documents and caches them in an internal cache using a user specified cache policy

05

New addition triggers bots

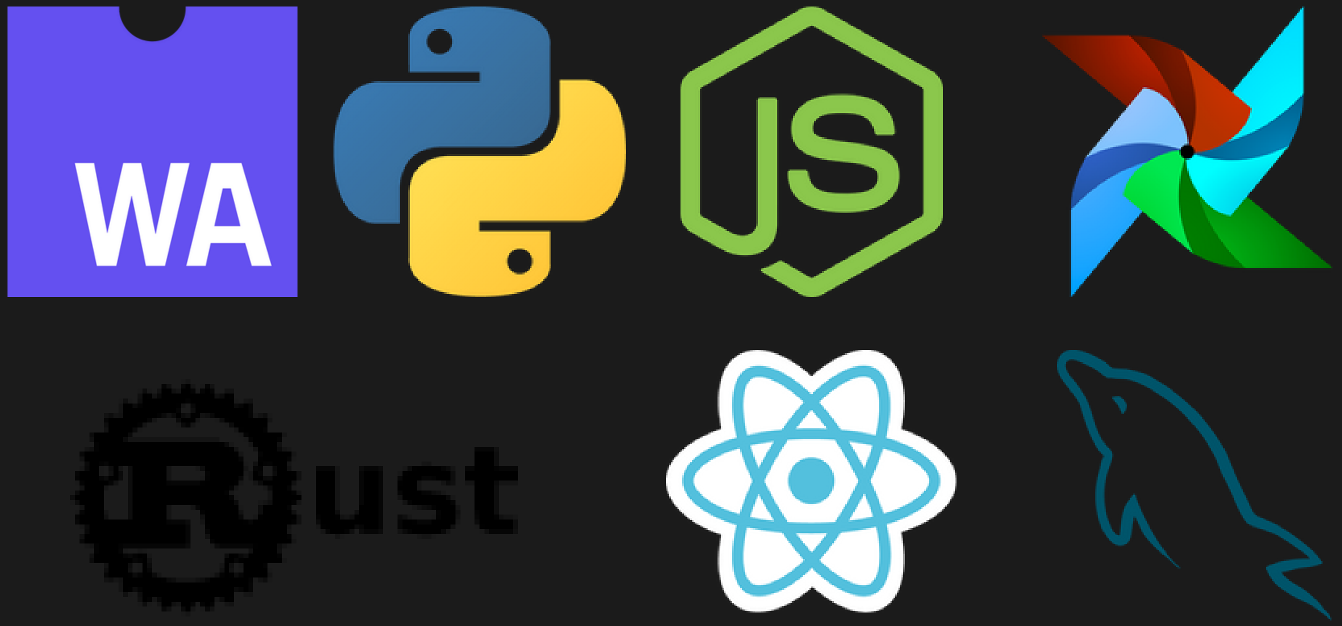
When a new website is added, the web crawlers are triggered which in turn trigger a workflow

06

Visual Representation

The workflow is represented as a DAG (Directed Acyclic Graph). This enables us to customize the pipeline to our needs

Tech
Stack


WA
JS
Rust

Discover Templates
Browse by category
Featured collections
Facebook Posts

Backend

- Rust
- Python
- NodeJS

Frontend

- React
- SASS

Cloud Providers

- AWS
- Alibaba
- GCP

Database

- Cloud hosted RDBMS

Others

- Apache Airflow
- Scikit Learn

Unique Selling Point

Completely autonomous

No manual intervention required when indexing new websites

Insanely Fast

Built in Rust, the search engine provides answers in < 50ms along with customizable indexes

Plug n Play system

The proposed system is independent of the URL. Simply provide the necessary secrets and start the application. The system will take care of the rest.

Unique Selling Point

Microservice approach

The components of the system are designed to be deployed as micro-services

Self Contained

All dependencies are self contained. The system is designed to work out of the box. No extensive setup required

Dynamic Strategy Switching

Our proposed system will have multiple 'strategies' to choose from to provide better results and can dynamically switch between them

Unique Selling Point

Extensive Logging and failovers

Every event is logged and can be viewed visually. Retry on failure is enabled so that data is never lost

Visual Representation

We can view the workflows as visual graphs with nodes and edges to represent the data and the processing done on it. Airflow maintains it's state as a DAG (Directed Acyclic Graph)

Workflow Scheduling and distribution

We can schedule when a job gets triggered and customize the various execution paths that a workflow can take

Future Enhancements

WASM Integration

The search engine, being written in Rust, can be compiled into WASM and be used directly in the frontend. This would reduce latency and increase performance