data engineering — National School of Applied Sciences of Al Hoceima

03 jun 2024

project presentation

Patent Analysis of Sustainable Aviation Fuel Using a Big Data Solution

Support by

EL MEFTAHI Souhayla

El MANSOURI Samiha

Supervised by

Pr.Anas ELHADDADI

Plan

Introduction

project objectives

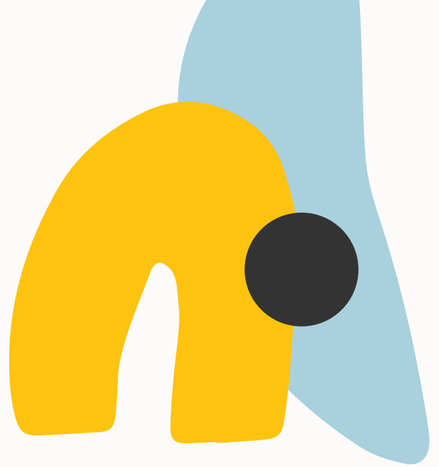
Architecture

Methodology

Demo

conclusion

Introduction



- Importance of patent analysis for intellectual property protection
- Rising number of patent applications globally
- Need for accessible, user-friendly patent analysis tools
- proprietary tools: PatSeer, Derwent Innovation, Orbit Intelligence
- Limitations: Expensive, complex, not open-source
- Public tools: Patent2Net, GooglePatents, PatentMiner
- Gap in accessible and easy-to-use tools

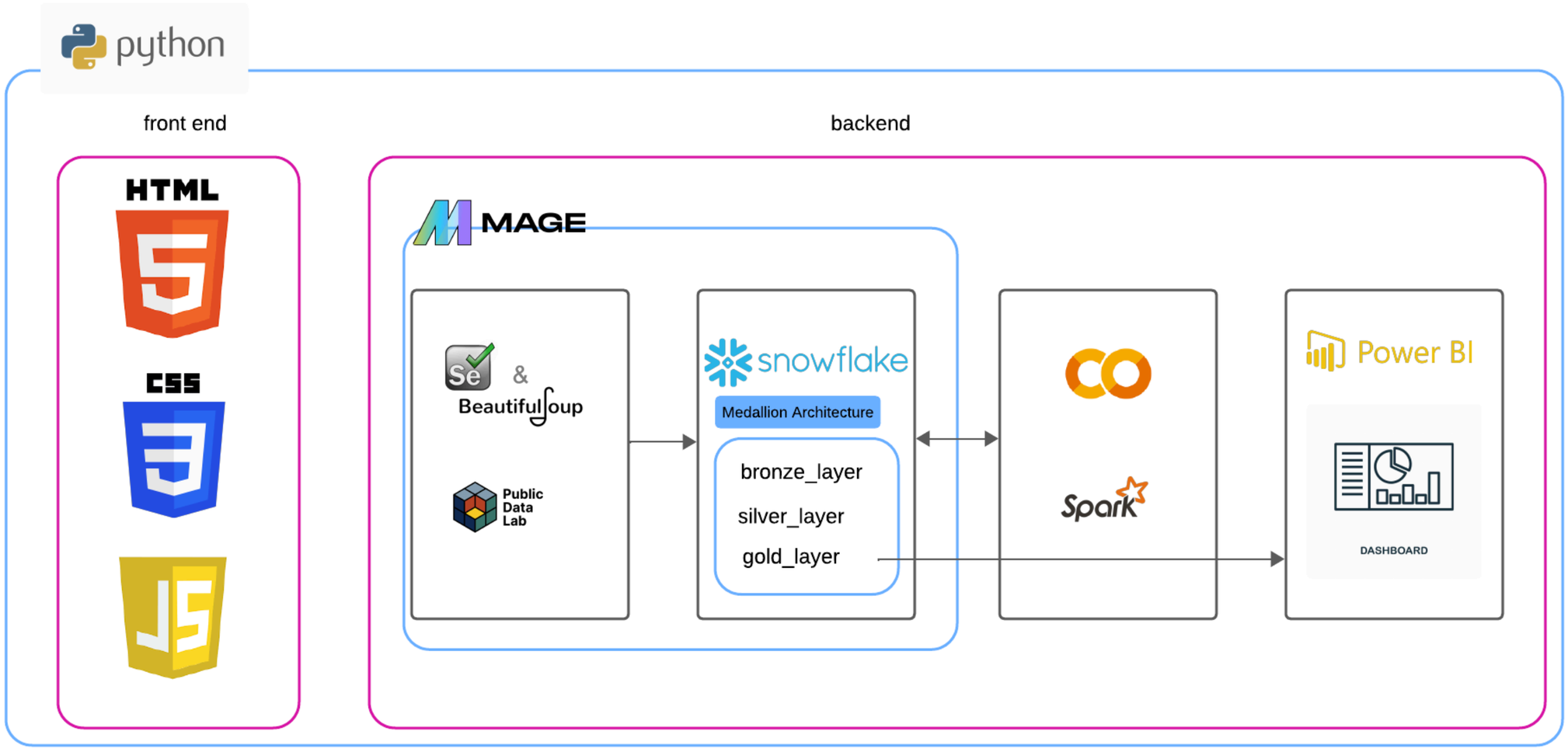
Project Objectives

- Develop an open-source patent analysis tool
- Focus on sustainable aviation fuel patents
- Provide descriptive analyses, thematic axes, and citation networks

“Ensah_Aviation Tool Overview

We created Ensah_Aviation, an open-source tool implemented in Python. It leverages data from multiple sources including USPTO, Google Patent, and Espace Net, providing a comprehensive analysis of patent data.

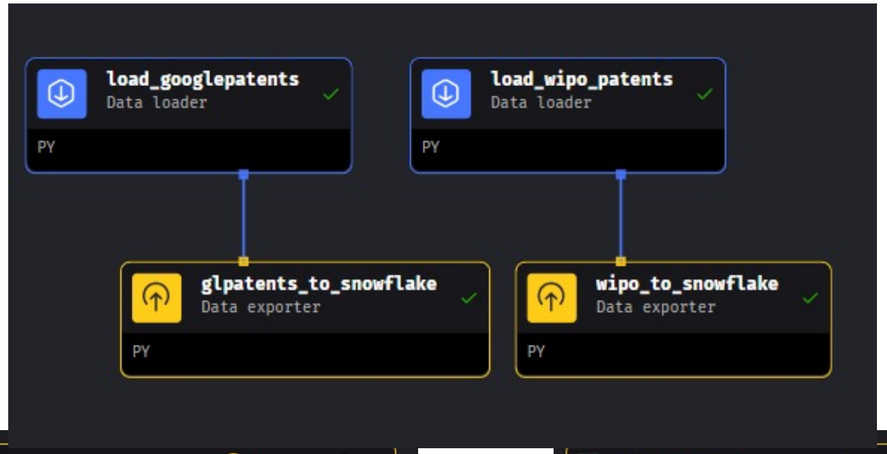
Methodology



data collection

- Sources: USPTO, Google Patent Lens, UPTO Net Space,wipo
- Methods: Web scraping using BeautifulSoup and Selenium.,using public data set
- Automation: using selenium for automized process of importing public data set

Tools: Python ,selenuim, Mage.ai for data flow automation



PY DATA EXPORTER glpatents_to_snowflake 1 parent

Positional arguments for decorated function:

```
@data_exporter
def export_data(data):
    data → load_googlepatents
```

```python
from mage_ai.settings.repo import get_repo_path
from mage_ai.io.config import ConfigFileLoader
from mage_ai.io.snowflake import Snowflake
from pandas import DataFrame
from os import path

if 'data_exporter' not in globals():
    from mage_ai.data_preparation.decorators import data_exporter


@data_exporter
def export_data_to_snowflake(df: DataFrame, **kwargs) → None:
    """
    Template for exporting data to a Snowflake warehouse.
    Specify your configuration settings in 'io_config.yaml'.

    Docs: https://docs.mage.ai/design/data-loading#snowflake
    """
    table_name = 'raw_wipo'
    database = 'Bronze_Layer'
    schema = 'RAW'
    config_path = path.join(get_repo_path(), 'io_config.yaml')
    config_profile = 'default'

    with Snowflake.with_config(ConfigFileLoader(config_path, config_profile)) as loader:
        loader.export(
            df,
            table_name,
            database,
            schema,
            if_exists='replace',  # Specify resolution policy if table already exists
        )
```

PY DATA EXPORTER glpatents_to_snowflake 1 parent

Positional arguments for decorated function:

```
@data_exporter
def export_data(data):
    data → load_googlepatents
```

```python
from mage_ai.settings.repo import get_repo_path
from mage_ai.io.config import ConfigFileLoader
from mage_ai.io.snowflake import Snowflake
from pandas import DataFrame
from os import path

if 'data_exporter' not in globals():
    from mage_ai.data_preparation.decorators import data_exporter


@data_exporter
def export_data_to_snowflake(df: DataFrame, **kwargs) → None:
    """
    Template for exporting data to a Snowflake warehouse.
    Specify your configuration settings in 'io_config.yaml'.

    Docs: https://docs.mage.ai/design/data-loading#snowflake
    """
    table_name = 'raw_wipo'
    database = 'Bronze_Layer'
    schema = 'RAW'
    config_path = path.join(get_repo_path(), 'io_config.yaml')
    config_profile = 'default'

    with Snowflake.with_config(ConfigFileLoader(config_path, config_profile)) as loader:
        loader.export(
            df,
            table_name,
            database,
            schema,
            if_exists='replace',  # Specify resolution policy if table already exists
        )
```

data storage

Medallion Architecture

Bronze Layer

- Raw, unprocessed data
- Stores raw data in its original form.

Silver Layer

- Cleaned and preprocessed data
- Removes duplicates and errors
- Prepares data for deeper analysis.

Gold Layer

- Transformed and aggregated data ready for analysis.
- Aggregates data to meet end-user needs.
- Optimizes data for queries and business analysis.

Tools: Snowflake for data storage, Mage.ai for data flow automation

data transformation

Preprocessing

- remove duplicates

- cloumns selections
- add country from id for google patents data set
- handling some missing values
- merge 4 tables into one by chosen shared columns

Computations

- Adding calculated columns (publication_to_grant-duration, filing_to_priority_duration_days,filling-to-grant-duration).

- Using Support Vector Machine (SVM) for classifying technology sectors.
- Creating CSV files for fact tables and importing them into Snowflake.

Tools: Snowflake,Spark ,SVM,Python,Google colab

BRONZE_LAYER	Local	ACCOUNTADMIN
GOLD_LAYER	Local	ACCOUNTADMIN
SILVER_LAYER	Local	ACCOUNTADMIN

BRONZE_LAYER / RAW

Create

Schema ACCOUNTADMIN 2 days ago

Schema Details | Tables

4 Tables

Search | All Tables

NAME	TYPE	CLASSIFICATION	OWNER	RO...	BY...	CREATED
RAW_ESPACENET	Table	—	ACCOUNTADMIN	120	2.5...	2 days a...
RAW_GOOGLEPAT...	Table	—	ACCOUNTADMIN	120	2.5...	2 days a...
RAW_USPTO	Table	—	ACCOUNTADMIN	120	2.5...	2 days a...
RAW_WIPO	Table	—	ACCOUNTADMIN	120	2.5...	2 days a...

SILVER_LAYER / PUBLIC

Create

Schema ACCOUNTADMIN 2 days ago

Schema Details | Tables

1 Table

Search | All Tables

NAME	TYPE	CLASSIFICATION	OWNER	RO...	BY...	CREATED
PATENTS_DA...	Table	—	ACCOUNTADMIN	120	2.5...	2 days a...

BIG_DATA / PATENT

Create

Schema ACCOUNTADMIN 3 weeks ago

Schema Details | Tables

9 Tables

Search | All Tables

NAME	TYPE	CLASSIFICATION	OWNER	RO...	BY...	CREATED
ASSIGNEE	Table	—	ACCOUNTADMIN	0	0.0B	3 weeks...
COUNTRY	Table	—	ACCOUNTADMIN	0	0.0B	3 weeks...
FACTPATENT	Table	—	ACCOUNTADMIN	0	0.0B	3 weeks...
FILING_DATE	Table	—	ACCOUNTADMIN	0	0.0B	3 weeks...
GRANT_DATE	Table	—	ACCOUNTADMIN	0	0.0B	3 weeks...
INVENTOR	Table	—	ACCOUNTADMIN	0	0.0B	3 weeks...
PATENT	Table	—	ACCOUNTADMIN	0	0.0B	3 weeks...
PRIORITY_DATE	Table	—	ACCOUNTADMIN	0	0.0B	3 weeks...
PUBLICATION_D...	Table	—	ACCOUNTADMIN	0	0.0B	3 weeks...

data visualization

Page 1: Overview
Metrics Visualized:
Total and average patent durations (publication, grant, and filing)
General distribution of patents by source

page 2: Detailed Analysis
Metrics Visualized:
Patent distribution by sector (e.g., Automotive, Biotechnology)
Top authors and assignees

Page 3: In-depth Analysis
Metrics Visualized:
Detailed patent durations (filing to grant, creation to publication)
Patent processing efficiency

Tools: Snowflake ,Power BI

demo

References

bibliographics

- · Georgiou, K.; Mittas, N.; Ampatzoglou, A.; Chatzigeorgiou, A.; Angelis, L. What is being Patented in Software Engineering? Empirical Evidence from USPTO. IEEE Softw. 2023, 1–7. [Google Scholar] [CrossRef]
- · Albino, V.; Ardito, L.; Dangelico, R.M.; Petruzzelli, A.M. Understanding the development trends of low-carbon energy technologies: A patent analysis. Appl. Energy 2014, 135, 836–854. [Google Scholar] [CrossRef]

thanks !