News Article Data Pipeline

Course Number: DSCI-6007

TEAM-11

Thomas Powell

Team Leader
Data Scientist

Simran Sattar

Data Analyst

Pavan Kumar Sunkara

Data Engineer

List of Contents

Introduction

The 24-hour news cycle presents challenges in tracking key trends due to the overwhelming volume of information. The News Article Data Pipeline addresses this by automating news extraction and analysis using Python for web scraping, AWS for storage and processing, and keyword retrieval. This system provides real-time insights, making it valuable for end-users like journalists and researchers.

Business Scenario Overview

Problem

- It is extremely difficult to keep track of current events with the 24-hours news cycle and multitude of platforms that publish news.
- It is impossible to deduce the personal importance or relevance of stories

Data Needed

- News Articles
  - Headlines, Media Outlet, Publishing Date

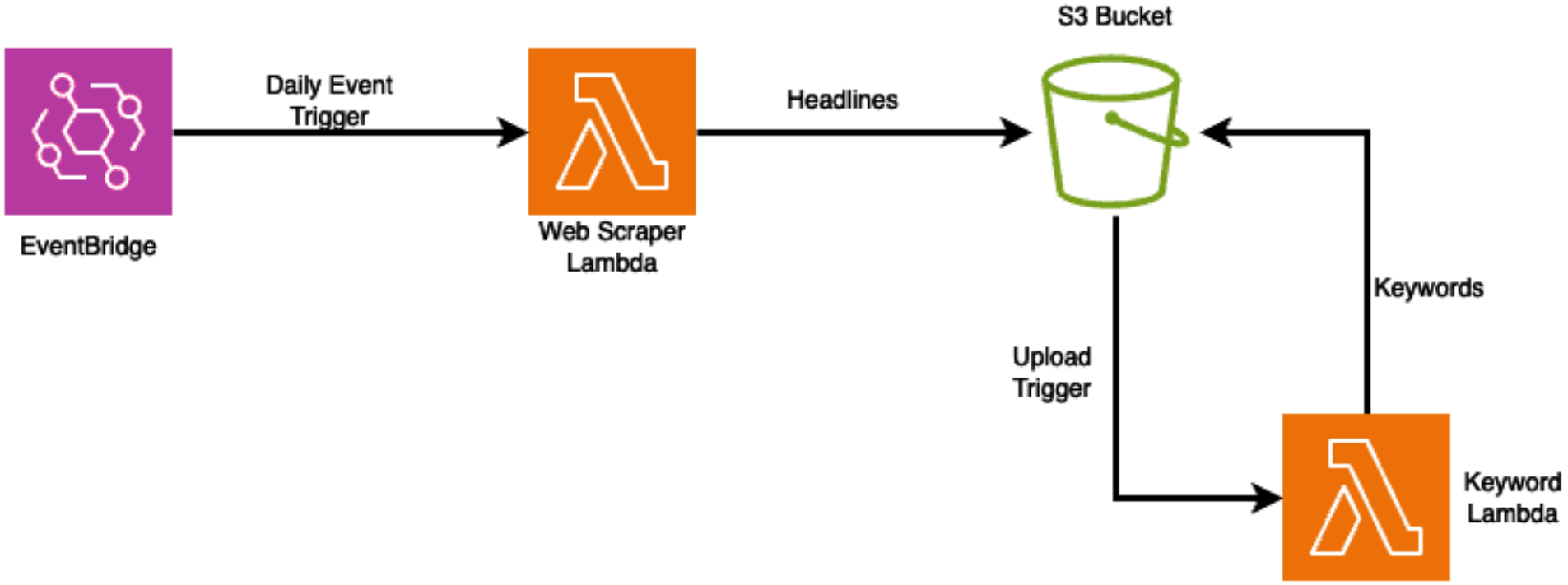
Data Obtainment

- Google News webpage

Solution Overview

- Business Understanding
  - Keep track of most popular topics, trends in daily news.
- Data Understanding
  - Scrape headlines from Google News articles.
- Data Preparation
  - Serverless Architecture (AWS Lambda), Web Scraping (BeautifulSoup), HTTP requests (Requests library), Data Storage (S3), Error Handling (AWS Cloudwatch).
- Evaluation
  - Topic Coherence, Article Consistency.
- Deployment
  - Automation with AWS Lambda, Storing the Data with AWS S3, Monitor Lambda execution logs in AWS CloudWatch.

Architecture Diagram of the Solution



Methodology

- **Business Understanding**: Identify news trends for decision-making.
- **Data Understanding**: Scraped Google News headlines (e.g., Notre Dame reopening, Syrian conflict).
- **Data Preparation**: Used Python for scraping; stored data in AWS S3.
- **Modeling**: Keyword obtainment and sorting (e.g., "Notre Dame," "Damascus").
- **Evaluation**: Focused on topic coherence and timeliness.
- **Deployment**: Automated with AWS Lambda, monitored via CloudWatch.

Data Engineering Pipeline

- **Data Ingestion**:

Tools: Python, BeautifulSoup, Requests.

- **Data Storage**:

Tools: AWS S3 for scalable and secure storage.

- **Data Processing**:

Tools: AWS Lambda for serverless execution, Python for transformations.

- **Data Consumption**:

Tools: Analytical dashboards (future integration).

- **Model Deployment**:

Environment: Serverless deployment using AWS Lambda.

- **Data Visualization**:

Comprehensive visualizations include topic distributions and word clouds.

Operationalization

Added Value for End Users:

- Provides real-time insights into news trends.
- Scalable for processing vast amounts of data efficiently.
- Usability for journalists, researchers, and other end-users to understand prevalent topics.

Results

Key Topics Identified:

Notre Dame Reopening:

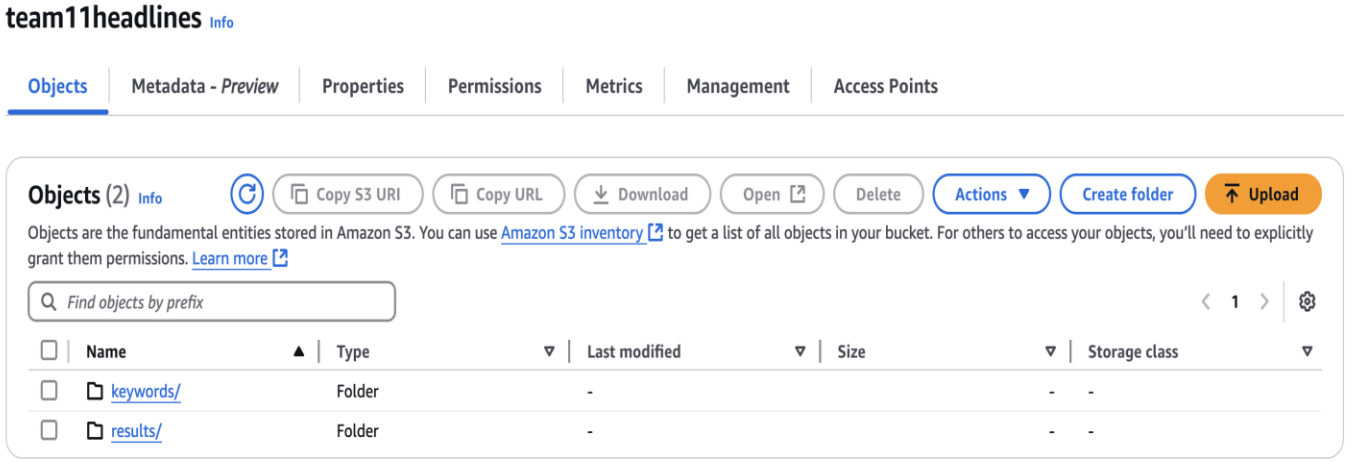
- Global headlines on reopening event.
- Keywords: Notre Dame, Paris, cathedral, reopening.

Syrian Conflict:

- Escalating tensions in Damascus.
- Keywords: Damascus, rebels, Assad.

Visualizations:

- Word clouds and topic distributions.



Results

result_20241207_201634.txt

result_20241207_201634.txt

```
1  Trump attends Notre Dame Cathedral reopening in Paris
2  World leaders gather for reopening of Notre-Dame Cathedral in Paris
3  Live updates: Notre Dame bells ring for the first time since fire
4  An archbishop's knock formally restores Notre Dame to life as winds howl and heads of state look on
5  Syrian rebels begin to encircle Damascus amid denials Assad has fled
6  Syrian Forces Withdraw From Damascus Suburbs, Monitors Say: Live Updates
7  Syrian rebels threaten Damascus, Assad from north and south
8  Syrian rebels battle for Homs and advance on Damascus, Assad's rule at stake
```

keyword_analysis_20241207_211059.txt

keyword_analysis_20241207_211059.txt

```
1  notre: 4
2  dame: 4
3  syrian: 4
4  damascus: 4
5  rebels: 3
6  assad: 3
7  cathedral: 2
8  reopening: 2
9  paris: 2
10 live: 2
```

Lessons learned

Challenges Overcome:
- Handling real-time data extraction errors.
- Configuring AWS services for scalability.

Helpful Resources:
- Python libraries (BeautifulSoup, Requests).
- AWS Documentation for Lambda and S3.

New Skills:
- Implementation of EventBridge and Cloudwatch

Next Steps:
- Expanding data sources beyond Google News.
- Incorporating sentiment analysis.

Summary

The **News Article Data Pipeline** automates news extraction and analysis using Python and AWS Lambda and S3.

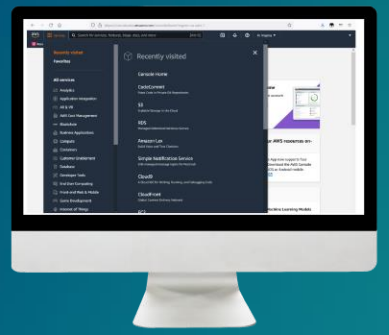
It provides real-time insights into key trends, helping users navigate the 24-hour news cycle effectively with a scalable and reliable solution.

Conclusion

This project showcases the integration of cloud computing and data science to analyze news trends efficiently.

Future improvements include expanding data sources and adding sentiment analysis for deeper insights.

Demo



Github Link: https://github.com/tpowell48/News-Article-Data-Pipeline

Thank you!☻