MLB Game Score Automation Pipeline-demo

Project Overview

This project presents a robust, automated Python pipeline designed to scrape daily MLB (Major League Baseball) game scores and matchups directly from Baseball-Reference.com. The collected data is then seamlessly uploaded to a designated Google Sheet, providing a real-time, shareable, and structured dataset for sports analytics, trend tracking, and historical review.

This pipeline exemplifies advanced data mining capabilities, including targeted web scraping, data structuring, and integration with cloud services, making it an ideal solution for clients requiring up-to-date sports data for various analytical purposes.

Key Features & Capabilities

- • Automated Daily Game Score Collection
- • Targeted Web Scraping using requests and BeautifulSoup4
- • Google Sheets API Integration
- • Real-time Data Sharing via sharable links
- • Structured Output (Team Matchups and Scores)
- • Comprehensive Logging of All Pipeline Activities
- • Easy Configuration via config.json

Pipeline Output & Demonstrative Results

The pipeline runs through `run_scraper.py`, producing console logs and populating a real-time Google Sheet.

Example Log Snippet (Truncated):

```
- Successfully fetched yearly schedule for 2024
- Collected: Seattle Mariners (11) @ Los Angeles Angels (0)
- Uploaded 29 scores/matchups to 'Scores'
- Spreadsheet shared as 'reader'
- Scores Spreadsheet URL:
https://docs.google.com/spreadsheets/d/1CuZDCK_uCwLiqjdrWjh9U8gWNREx-7edYaaPZRyvKW8/edit#gid=0
```

Example Output in Google Sheets

Sheet: 'Scores'
Columns: Date | Away_Team | Away_Score | Home_Team | Home_Score

Sample Rows:

July 11, 2024 | Seattle Mariners | 11 | Los Angeles Angels | 0

July 11, 2024 | Atlanta Braves | 0 | Arizona D'Backs | 1

July 10, 2024 | Texas Rangers | 2 | Los Angeles Angels | 7

Live Sheet: https://docs.google.com/spreadsheets/d/1CuZDCK_uCwLiqjdrWjh9U8gWNREx-7edYaaPZRyvKW8/edit#gid=195384354

Technology Stack

- • Python 3.x
- • requests, BeautifulSoup4, lxml for scraping
- • gspread, oauth2client for Google Sheets API
- • python-dateutil, pytz for time handling
- • json for config loading
- • logging for pipeline diagnostics

Quick Start Guide

1. Clone the Repository:

```
git clone https://github.com/YourUsername/MLB-Game-Score-Automation.git

cd MLB-Game-Score-Automation
```

2. Install Requirements:

```
pip install -r requirements.txt
```

3. Add `credentials.json` to the root directory (DO NOT commit this to Git).

4. Modify `config.json` as needed.

5. Run the script:

```
python run_scraper.py
```

Summary

This project clearly demonstrates end-to-end automation of sports data scraping and cloud-based delivery. It is highly relevant to any role that requires real-time or historical sports analytics, and it aligns closely with the requirements of multi-sport data tracking initiatives like the one outlined in your Upwork posting.