Cloud Computing 09/16/2025

Y Combinator Job Market Analysis Pipeline

By Team

Unattached and Unhinged

Team member 1: Varad Paradkar (801418318)
Team member 2: Mitali Yadav (801453849)
Team member 3: Sudeepta Bal(801455628)
Team member 4: Ruthwik Dovala(801431661)

Project Overview and Features

- A scalable pipeline for continuous job data collection, processing, and analytics - integrating Python scrapers, Cloudflare Workers, and Spark analytics.
- Three main modules - **Scraping Engine**, **Cloudflare Worker Backend**, and **Spark Analytics Pipeline** - handle data extraction, API management, and insights generation.
- Automated daily/weekly scraping, real-time analytics, R2/D1 data storage integration, modular configuration, and expansion-ready design with monitoring through GitHub Actions.

Progress so far

1. The project has successfully built a multi-layer automated job scraping pipeline.
2. Daily scraping workflow (GitHub Actions) is operational – collecting job data from Y Combinator’s job boards.
3. Enhanced YC Job Scraper implemented in Python using Requests, and environment-based configuration.
4. Cloudflare Worker backend is live – handles API endpoints for job submission, retrieval, analytics, and export to R2 storage.
5. The Worker integrates with D1 database for storing scraped job data and R2 for exporting datasets.
6. Weekly Analytics workflow is configured using Spark for advanced insights on job trends, top companies, and remote work patterns.
7. Simplified analytics (simple_analytics.py) added for GitHub Actions compatibility to test analytics.
8. Database schema and indexes for “jobs”, “scraping_jobs”, and “data_exports” tables have been created.
9. Setup.py automates environment initialization, dependency installation, and .env configuration.

Current Challenges

- ❖ Spark S3A CloudFlare R2 Read Fails

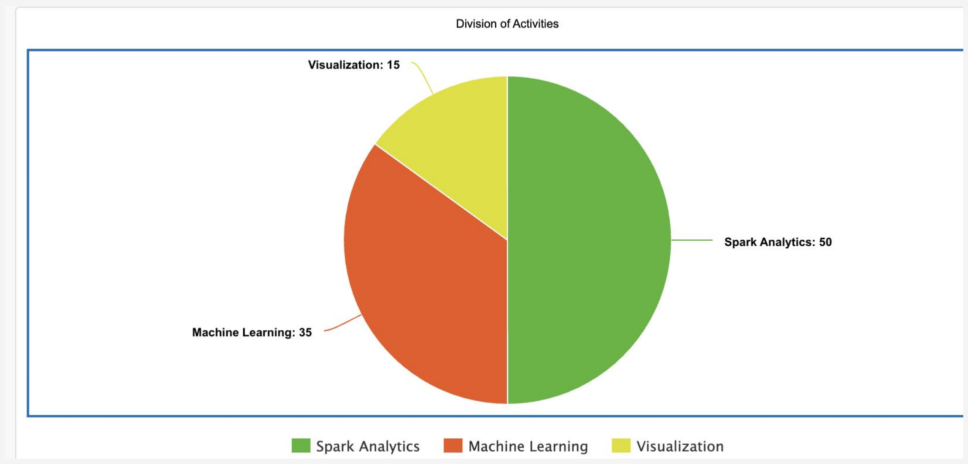
On trying to read a **JSON file** from CloudFlare R2 with PySpark, when running final_daily_analytics.py, the error pops up: **NumberFormatException: For input string: “60s”**

On setting all S3A timeouts as integers (e.g., “60000”), used the right Hadoop AWS packages, and checked all the configs, but the error still persists. Maybe, a hidden config is still set to “60s” but have to dig deeper!

Next Week activities

System Integration & Automation:

- ❖ Complete integration of scrapers with Cloudflare Worker APIs and automated Spark analytics deployment in the cloud for daily and weekly processing.



Thank You!