Rocky Team Internship Presentation 2019

•••

By Doug Talbert

Mentored by Greg Dungca

Purpose

- Looking for ways to improve team productivity
- Create a tool to extract data from Github and generate insights to enhance productivity

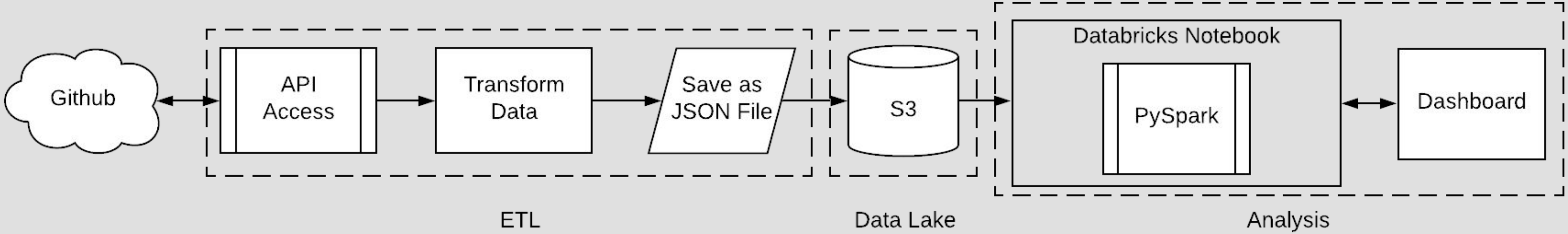
Methods

- Divide and conquer
  - ETL (Extract, transform, load) Python script
  - Analytics with PySpark in Databricks notebook

Insights

- Generate different charts that show the following for a given repository:
  - Open pull requests
  - Weekly Activity
  - Histogram of time until pull request is closed
  - Contributing users
  - Proportion of pull requests that get merged

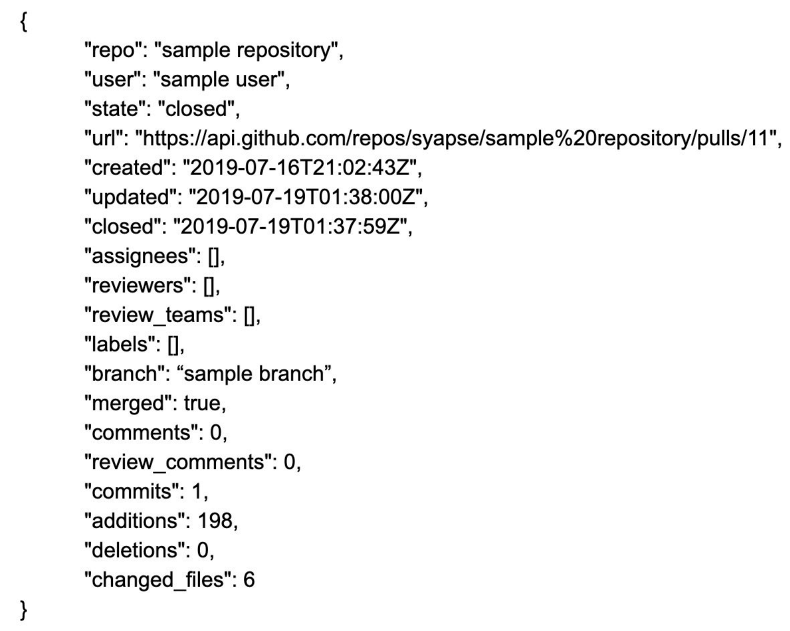
Design


Github
API Access
Transform Data
Save as JSON File
S3
Databricks Notebook
PySpark
Dashboard
ETL
Data Lake
Analysis

ETL Input and Output

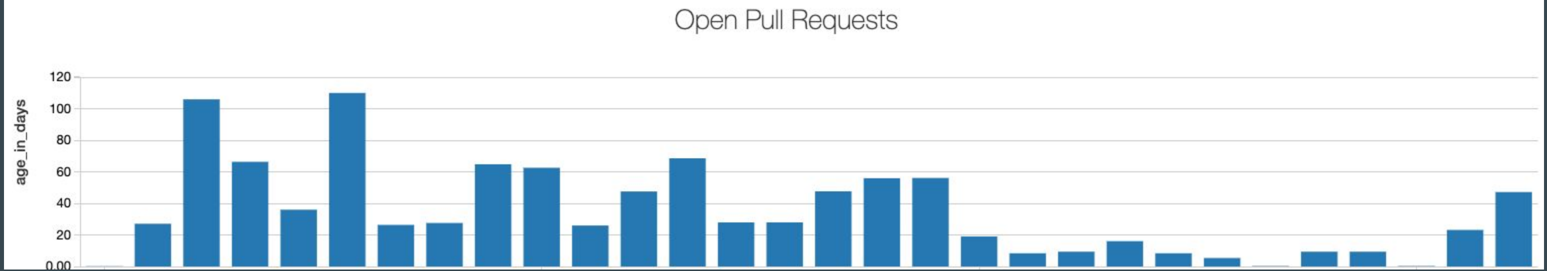
API Return (Single PR)

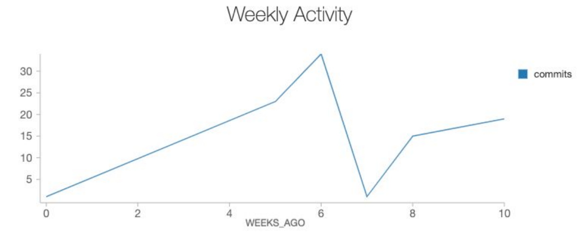
Transformed PR

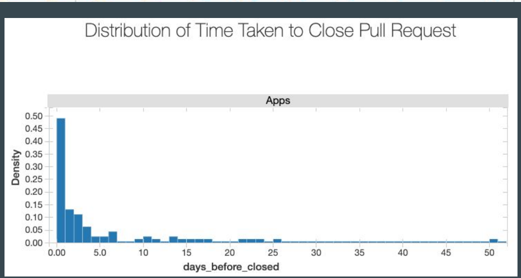


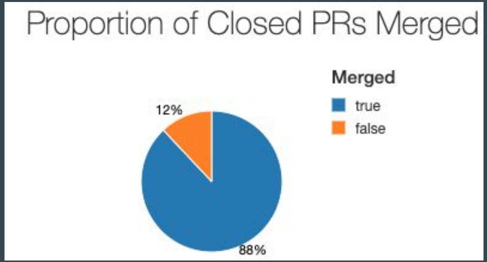
```
{
        "repo": "sample repository",
        "user": "sample user",
        "state": "closed",
        "url": "https://api.github.com/repos/syapse/sample%20repository/pulls/11",
        "created": "2019-07-16T21:02:43Z",
        "updated": "2019-07-19T01:38:00Z",
        "closed": "2019-07-19T01:37:59Z",
        "assignees": [],
        "reviewers": [],
        "review_teams": [],
        "labels": [],
        "branch": “sample branch”,
        "merged": true,
        "comments": 0,
        "review_comments": 0,
        "commits": 1,
        "additions": 198,
        "deletions": 0,
        "changed_files": 6
}
```

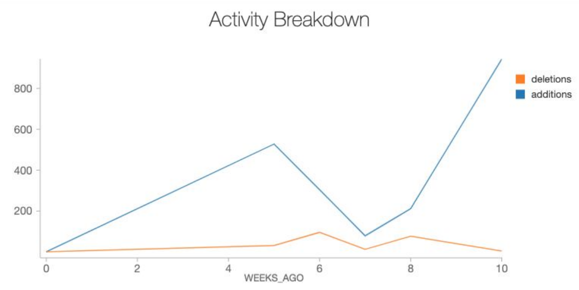
Analysis

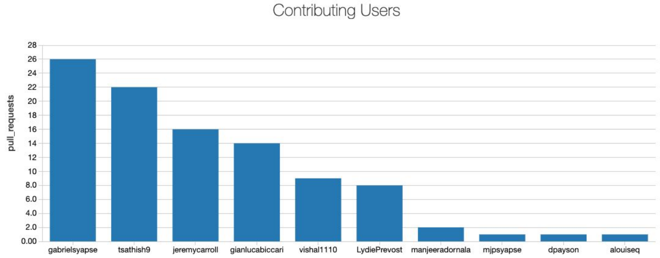

Open Pull Requests


Weekly Activity


Distribution of Time Taken to Close Pull Request


Proportion of Closed PRs Merged


Activity Breakdown


Contributing Users

Questions?