GitHub Repository Data Pipeline

COSC 448 - Directed Studies

Abijeet Dhillon

Objective

01

EXTRACT

Collect data from software repositories

02

TRANSFORM

Structure and enrich collected data into json artifacts

03

LOAD

Index and store structured data using Elasticsearch

High-Level Architecture

Developer

Writes code, opens issues, pushes commits, creates PRs, etc.

GitHub Repo

Stores commits, issues, PRs, metadata, etc.

Retrieval

Collect repo metadata, issues, PRs, commits, contributors, git blame data, and more

Normalized JSON Output

Outputted json files (issues.json, commits.json, etc.)

Index

Flatten & transform JSON for search, apply mappings, and bulk index

Elasticsearch Database

Full-text database which is our backend for the LLM to query

Data Source and GitHub APIs

Primary Data Source:

- GitHub Repositories

GitHub REST v3 API:

- Used to extract majority of our data
- Each endpoint returns a predefined set of data

GitHub GraphQL v4 API:

- Used to extract git blame data
- Allows you to specify data fields in the query

Retrieval Component

config.py

- Centralizes settings
- Holds REPOS list

http_client.py

- API requests
- Token rotation

collectors.py

- Fetch & cache data
- Write JSON output

linkers.py

- Extract references between issues, PRs, repos

runner.py

- Runs retrieval component only

From GitHub to JSON

Retrieval Component

- Call APIs to fetch data
- Handle errors & limits
- Add token rotation

Why JSON Normalization

- Ensure stable schema
- Easier to link data
- Easier to index & query

Normalize into JSON

- Convert raw GitHub responses into clean, consistent JSON files
- Structure fields uniformly across all repos so schemas reusable

Indexing Component

config.py

- Load ES connection vars

schema.py

- ES index mappings
- Map JSON to index

client.py

- ES API wrapper to help with upload

indexer.py

- Scan JSON folder
- Inject repo name
- validates indices
- Bulk-indexes into ES

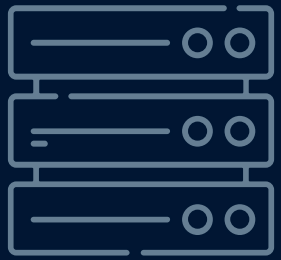
runner.py

- Runs indexing component only

From JSON to Elasticsearch

Indexing Component

- Read JSON artifacts
- Flatten nested data



Why Elasticsearch

- Ensure stable schema
- Easier to link data
- Easier to index & query

Indexing into Elasticsearch

- Create indices with explicit mappings (e.g. issues, PRs, etc.)
- Bulk index documents in batches to ensure quick upload

Pipeline Workflow

Configure Settings

Retrieval fetches data and writes it into JSON files. Indexing creates ES indices, applies mappings, and bulk-loads JSON files into database.

Query & Analyze

Data & Metrics From The Data

Metrics: popularity trends, repository age, release cadence

Repository-Level Data

Metrics: time-to-close, reopen rate, labes, bug vs feature ratio

Issues & PRs

Metrics: author frequency, hot files, commit message content

Commits & Contributors

Metrics: repo coupling, dependency hubs

Cross-Repo References & Blame Summaries

Outcomes

- End-to-end working pipeline for an ETL workflow
- Unified & queryable datasets across multiple repositories in Elasticsearch database
- LLM-ready schema design
- Ability to view deeper insights not visible in GitHub's UI easily

Challenges & Reflection

Challenges

- GitHub Rate Limits - rotate multiple tokens & backoff
- Data Quality - spot check fields using a small repository
- Time To Pull Data - unsolved, large repos (e.g. prettier/prettier) are slow to fetch issues/commits/blame

What I'd Do Differently

- Define the schema first
- Introduce GraphQL earlier
- Manage time better

Demo: Searching for Unassigned “Bug” Issues

ES|QL Query:

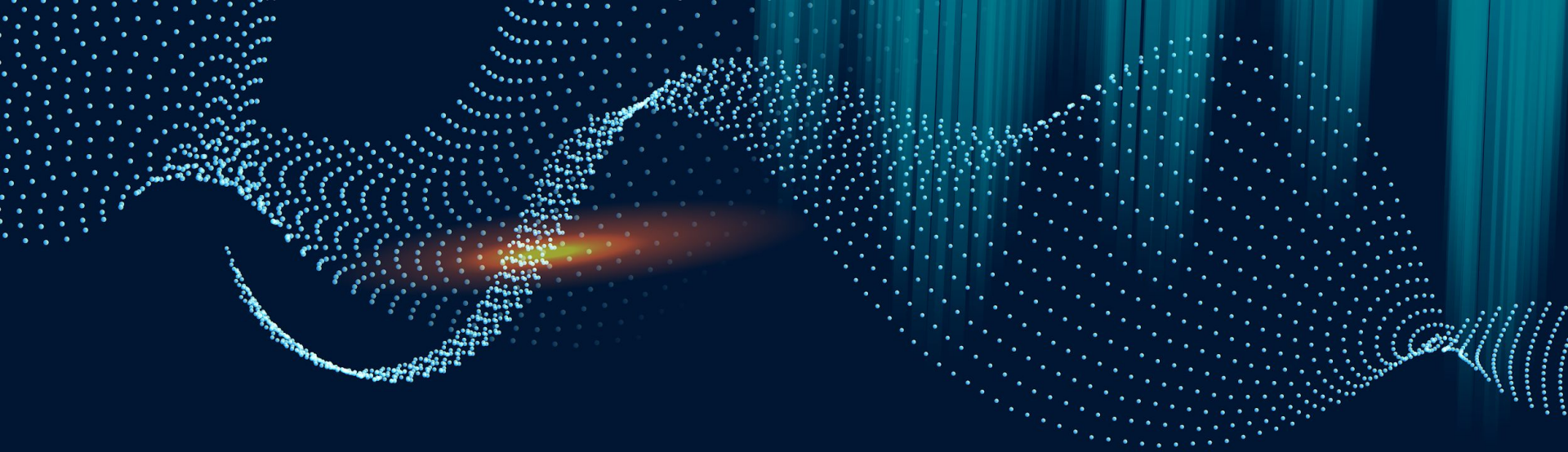
```
FROM issues

WHERE state == “open” AND
labels.name.keyword == “bug” AND
comments >= 5 AND assignee.login
IS NULL

KEEP repo_name, number, title,
comments, created_at, labels.name

SORT created_at ASC
```

6 results

Columns 6 | Sort fields

repo_name	number	title	comments	created_at	labels.name
standard/standard	786	Standard allows certain unnecessary semicolons	8	Feb 13, 2017 @ 20:56:03.000	bug
standard/standard	1,282	Still impossible to use Vue plugin	6	May 24, 2019 @ 11:02:30.000	bug
standard/standard	1,328	Global standard doesn't find global babel-eslint since …	5	Jul 12, 2019 @ 04:02:05.000	bug
micromatch/micromatch	212	filepaths with dots in them do not match with brace ranges	6	Mar 10, 2021 @ 05:03:29.000	bug
laravel-mix/laravel-mix	3,021	webpack-dev-server - 4.0.0-beta.3 breaks secure WSS websocket …	7	Jun 21, 2021 @ 06:40:46.000	bug
laravel-mix/laravel-mix	3,239	URL rewriting not working as expected with hot-reload (HMR)	5	Mar 30, 2022 @ 07:14:14.000	bug

ES query showing open ‘bug’ issues with heavy discussion but no assigned owner



Thank You!