Project Proposal

Real-Time Analysis of Public GitHub Activity

Roland Bernard, 19598, rolbernard@unibz.it

August 2025

1 Application Domain

This project's goal is to develop an application for monitoring public activity on GitHub. GitHub is a large software development platform hosting many open source projects. The application will capture and analyze a continuous stream of public events to provide insights into ongoing open source development trends. The events to be analyzed include the creation of new repositories, branches, and forks; pushes to public repositories; the opening and closing of issues or pull requests; and the starring of repositories. The application will offer a live dashboard that visualizes the data, by processing it in real-time.

2 Data Sources

The primary data source for this project will be the official GitHub Events API[1]. Specifically, the project will use the `api.github.com/events` endpoint, which provides the most recent events across the entire platform. Each event is delivered as a JSON object containing detailed information about the action performed. By periodically polling this endpoint, we can obtain a high-volume stream of events.

3 Technology and Architecture

The system will be built using a variation of the provided reference architecture, as can be seen in Fig. 1.

Data Ingestion A Java-based Kafka Producer will be developed to poll the GitHub Events API. This component will handle de-duplication of events before publishing them as JSON to a dedicated Kafka topic.

Stream Processing Apache Flink will be used for stream processing. It will consume the raw event stream from Kafka and execute a series of computations, to derive meaningful analytics.

Data Serving The processed data from Flink will be published to new Kafka topics. A Java-based server will consume these topics and expose the results via a WebSocket connection to the frontend.

Frontend The user interface will be a custom single-page application using HTML and JavaScript, served by the frontend server. It will leverage the ECharts library to create an interactive dashboard with dynamic charts that update in real-time as new data arrives.

[1] `https://docs.github.com/en/rest/activity/events`

4 Functionalities

The project will provide several features on the frontend dashboard:

Live Event Feed A filterable, real-time stream of raw events, allowing users to filter based on event type, username, or repository name.

Real-Time Event Counters Dynamic counters displaying the volume of different event types (e.g., commits, issues opened, forks, etc.) over sliding time windows, such as the last minute, 5 minutes, or hour.

Activity Leaderboards Continuously updated rankings of the most active repositories and users. This will be calculated based on the number of events over a sliding time window, such as the last 2 hours.

Trending Repository Detection Identification of repositories that are rapidly gaining popularity. This will be achieved by analyzing the rate of new stars. Flink will calculate the rate of new stars over time, allowing the system to flag repositories experiencing a significant increase in interest.

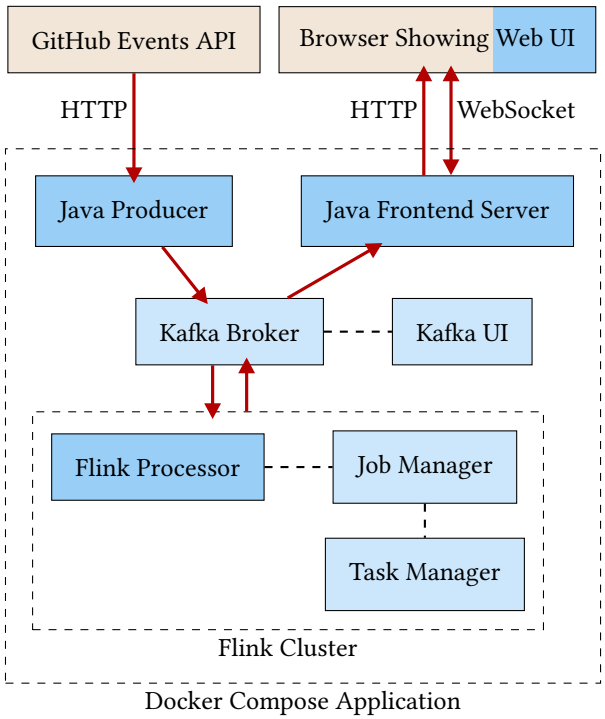

Figure 1: Proposed application architecture. Blue indicates parts of the architecture controlled by the project, while the brown indicates external components.