Ugur Can Avcu
215814239
Project – 3 Report

Project-3 Report

The project of the system architecture consists of main two components which are data-source.py and spark_app.py. The data source script makes three GitHub requests based on language using the token to get 50 pages of repos for each request. Then it basically loads to data into JSON and filters the data to get the needed key, value pairs for the functions that will be performed on spark_app.py. Finally, sends the data as JSON over TCP, port 9999.

Spark_app.py is where all required calculations are supposed to be performed. It ingests the data from TCP and data source being 9999. Batch interval is being adjusted, and MapReduce functions are being applied. After applying functions and calculating analytics, it sends the result to the dashboard over HTTP.

This Flask web app provides a very simple dashboard that illustrates the real-time live analytics gathered from Github data sent by the spark app. The web app is listening on port 5000.

All apps are designed to be run in Docker containers. The streaming folder is mounted for all containers using the volumes attribute in docker-compose.yaml.

Ugur Can Avcu
215814239
Project – 3 Report

How to run the app

To run the data streaming pipeline, we need to run two commands:

$ docker-compose up

$ docker exec streaming_spark_1 spark-submit /streaming/spark_app.py