EECS 4415 Project 3 report

Boho Kim 217303033

## <System Architecture>

This project retrieve data from GitHub API using HTTP requests. Using Python, it transforms data and send it for Spark Streaming using TCP. In Apache Spark Cluster, analyse data that we need and extract using SQL context. Extracted data in data frame is sent for Webapp Service. Using Python, we visualize it so that people can see in webpage such as number, graph, string.

## <Service Interaction>

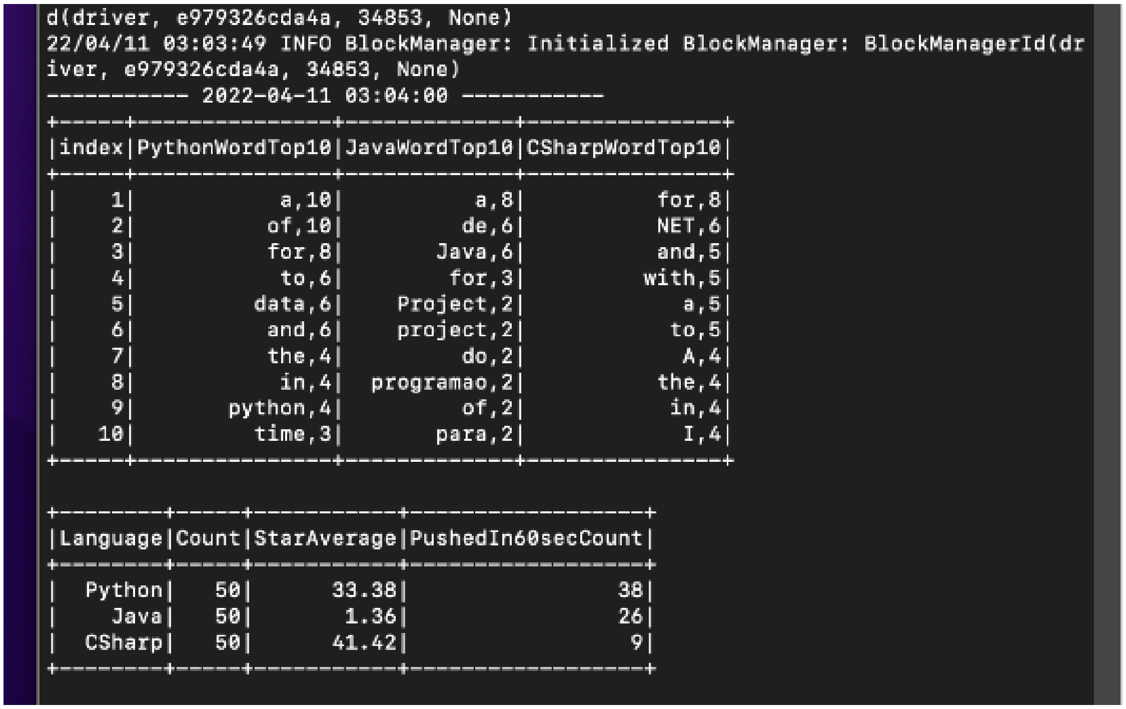
GitHub provides information about GitHub repositories using GitHub API. This service interacts with Data Source Service. Data Source Service requests data using HTTP Requests and GitHub API requires TOKEN for this information. Data Source Service can read the token from the environment variable. After getting information, it cleans data that we need. Using the same IP and Port that is used for TCP and Socket stream, Apache Spark receives data from Data source service. Then, it divides them into batches per time that is set in Python scripts. Then, scripts analyse them and transformed data is sent to to Web application for dashboard using requests. Then, web application transform data again to visualize it for web application user.

## <How the Spark Application processes the streaming data>

Spark Application processes the data in real time and does lazy evaluation. So, it collects Rdd and divide to batch but does not take an action until action is triggered. In each batch, they do map and reduce with data and, using Python Pyspark API, transforms Rdd to DataFrame that we can apply SQL context.

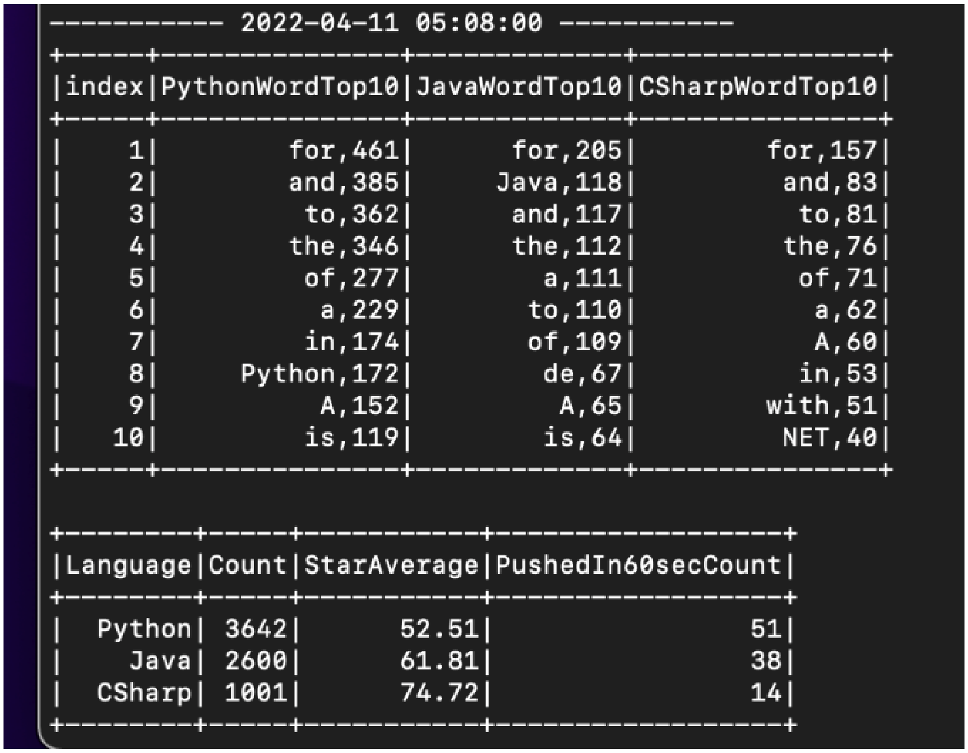
In this project, it takes data and divide to batch per 60 seconds. Using filter, organize data for each different category for our purpose and map and reduce. And, it transform them to Pyspark data frame that can be processed with SQL context. It makes 2 data frame that one is for repository count, average of stars, pushed time in 60 seconds and one is for top 10 words that are the most frequently used.

<2-hour running result>



```
d(driver, e979326cda4a, 34853, None)
22/04/11 03:03:49 INFO BlockManager: Initialized BlockManager: BlockManagerId(dr
iver, e979326cda4a, 34853, None)
----------- 2022-04-11 03:04:00 -----------
+-----+---------------+-------------+---------------+
|index|PythonWordTop10|JavaWordTop10|CSharpWordTop10|
+-----+---------------+-------------+---------------+
|    1|           a,10|          a,8|          for,8|
|    2|          of,10|         de,6|          NET,6|
|    3|          for,8|       Java,6|          and,5|
|    4|           to,6|        for,3|         with,5|
|    5|         data,6|    Project,2|            a,5|
|    6|          and,6|    project,2|           to,5|
|    7|          the,4|         do,2|            A,4|
|    8|           in,4|  programao,2|          the,4|
|    9|       python,4|         of,2|           in,4|
|   10|         time,3|       para,2|            I,4|
+-----+---------------+-------------+---------------+

+--------+-----+-----------+------------------+
|Language|Count|StarAverage|PushedIn60secCount|
+--------+-----+-----------+------------------+
|  Python|   50|      33.38|                38|
|    Java|   50|       1.36|                26|
|  CSharp|   50|      41.42|                 9|
+--------+-----+-----------+------------------+
```



```
----------- 2022-04-11 05:08:00 -----------
+-----+---------------+-------------+---------------+
|index|PythonWordTop10|JavaWordTop10|CSharpWordTop10|
+-----+---------------+-------------+---------------+
|    1|        for,461|      for,205|        for,157|
|    2|        and,385|     Java,118|         and,83|
|    3|         to,362|      and,117|          to,81|
|    4|        the,346|      the,112|         the,76|
|    5|         of,277|        a,111|          of,71|
|    6|          a,229|       to,110|           a,62|
|    7|         in,174|       of,109|           A,60|
|    8|     Python,172|        de,67|          in,53|
|    9|          A,152|         A,65|        with,51|
|   10|         is,119|        is,64|         NET,40|
+-----+---------------+-------------+---------------+

+--------+-----+-----------+------------------+
|Language|Count|StarAverage|PushedIn60secCount|
+--------+-----+-----------+------------------+
|  Python| 3642|      52.51|                51|
|    Java| 2600|      61.81|                38|
|  CSharp| 1001|      74.72|                14|
+--------+-----+-----------+------------------+
```