T-SA:

Twitter keyword Search API based Tweet Analysis
(트위터 키워드 검색 API기반 트윗 분석)

과 목 산학캡스톤디자인1(2019-1학기)
담 당 교 수 정현숙 교수님
팀 명 브이아이(VI)
발 표 자 서재익
발 표 일 자 2019.04.04

Project Github URL: https://github.com/SeokJune/BigData_VI_T-SA/

T-SA: Contents

Twitter Keyword Search API based Tweet Analysis

T-SA: Team Introduction

Twitter Keyword Search API based Tweet Analysis

Name	Lee SeokJune
Student ID	20165072
Cell Phone	010-4020-5717
E-mail	op2se1@gmail.com
Major Lang	Java
GitHub	https://github.com/SeokJune
Part	- MariaDB 환경 구축 및 관리 - Hadoop(Map)구현 - 문서 작성 및 수정

Name	Lee YunHyuck
Student ID	20165062
Cell Phone	010-4220-5134
E-mail	leeyh5134@naver.com
Major Lang	Python
GitHub	https://github.com/yunhyuck
Part	- Hadoop3 환경 구축 - Hadoop, DB 연동 구현 - Sqoop 환경 구축 - Hadoop(Reduce)구현

Name	Bae InGyu
Student ID	20165073
Cell Phone	010-4679-4968
E-mail	happykkk789@naver.com
Major Lang	Python
GitHub	https://github.com/BaeInGyu
Part	- Python, DB 연동 구현 - Visualization 구현

Name	Seo JaeIck
Student ID	20144773
Cell Phone	010-2460-7617
E-mail	nero8879@naver.com
Major Lang	Python
GitHub	https://github.com/nero8879
Part	- Twitter API 구현 - Visualization 구현

T-SA: Purpose of Development

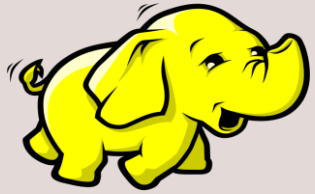
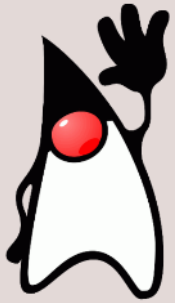
Twitter Keyword Search API based Tweet Analysis

대한민국 지역 및 특정 기간에 사용된 키워드 트렌드 분석

특정 인물의 트윗 스타일 분석

T-SA: Development Environment

Twitter Keyword Search API based Tweet Analysis

ubuntu	python	MariaDB	hadoop	OpenJDK	eclipse
					
18.04.2 LTS	3.6.8	10.1.38	3.2.0	1.8.0_191	2019-03(4.11)

T-SA: Development Environment

Twitter Keyword Search API based Tweet Analysis

To execute Python and Hadoop

Ubuntu Site:
- https://www.ubuntu.com/

T-SA: Development Environment

Twitter Keyword Search API based Tweet Analysis

Provides tweepy which is twitter API

Visualization of Data

Python Stie:
- https://www.python.org/

민형기, 파이썬으로 데이터 주무르기, 2017.12.29, 비제이퍼블릭

파이썬으로 데이터 주무르기 저자의 블로그 중 파이썬 목록
- https://pinkwink.kr/category/Software/Python

T-SA: Development Environment

Twitter Keyword Search API based Tweet Analysis

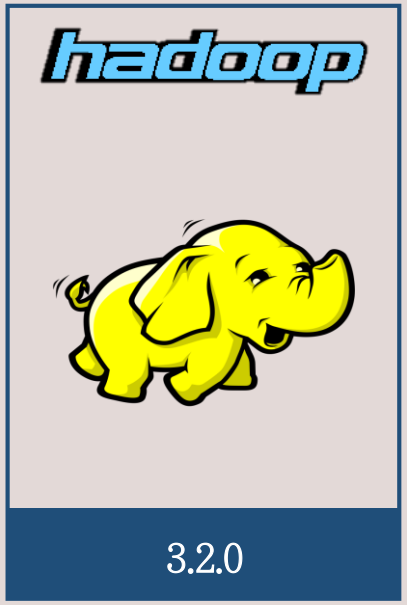
Open source RDBMS

Based on the same source as MySQL

MariaDB Stie:
- https://mariadb.com/kb/ko/mariadb

T-SA: Development Environment

Twitter Keyword Search API based Tweet Analysis



Distributed storge and Processing of big data

Pseudo-distributed

Hadoop Site:
- https://hadoop.apache.org/

정재화, 시작하세요! 하둡 프로그래밍 빅데이터 분석을 위한 하둡 기초부터 YARN까지 [개정2판], 2016.05.13, 위키북스

T-SA: Development Environment_W04

Twitter Keyword Search API based Tweet Analysis

Using Hadoop Ecosysytem

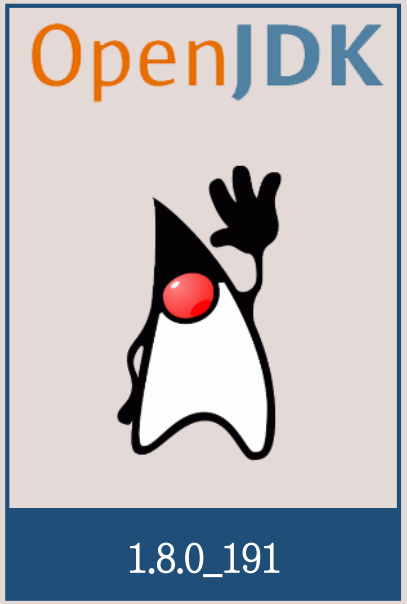
Sqoop
- RDBMS에서 데이터 수집
- 배치 처리 후 RDBMS에 데이터 저장

HDFS
- 분산 데이터 저장

Map/Reduce
- 분산 데이터 배치 처리

T-SA: Development Environment

Twitter Keyword Search API based Tweet Analysis

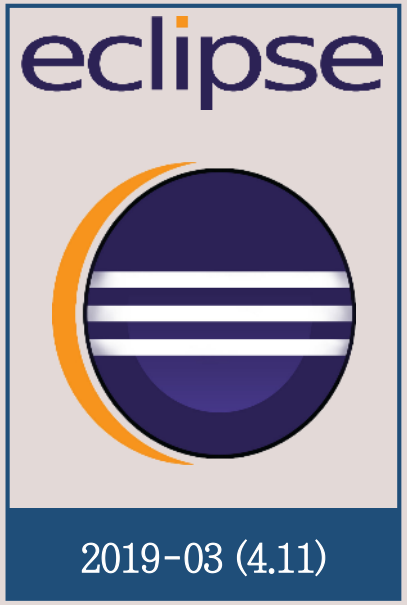


Used in Hadoop and Sqoop

OpenJDK Site:
- https://openjdk.java.net/

T-SA: Development Environment

Twitter Keyword Search API based Tweet Analysis



To Create Map/Reduce for Hadoop

Eclipse Site:
- https://www.eclipse.org/

T-SA: Development Environment

Twitter Keyword Search API based Tweet Analysis

Twitter API

5.0

To import data from Twitter

Twitter Developer Site:
- https://developer.twitter.com

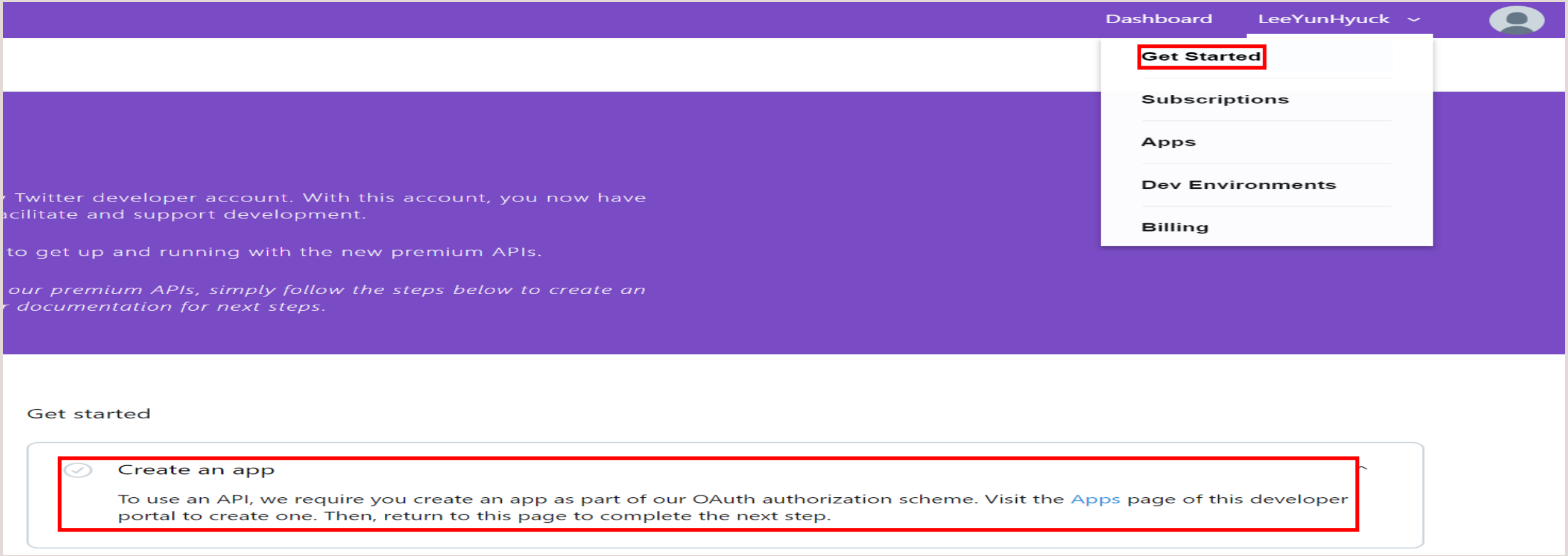
Tweepy Site:
- http://www.tweepy.org

Twitter Analysis Site:
- http://tweetrend.com/
- https://foller.me/

T-SA: Development Environment_W04

Twitter Keyword Search API based Tweet Analysis

Importing Twitter API Key



Get Started 를 누르면 아래와 같이 Twitter API를 사용하기 위한 인증키 발급을 받을 수 있는 목록을 받을 수 있다.
Create an app 을 제외한 나머지는 유료 이므로 무료로 사용하기 위한 인증키를 발급 받는다.

T-SA: Development Environment_W04

Twitter Keyword Search API based Tweet Analysis

Importing Twitter API Key

Tell us how this app will be used (required)

This field is only visible to Twitter employees. Help us understand how your app will be used. What will it enable you and your customers to do?

이 앱의 사용방법은 사용자의 키워드를 분석하여, 해당 키워드에 대해 분석을 통해 얻을 수 있는 정보들에 대해 시각화 하는 것에 목적이 있습니다.

Must be 100 characters or longer

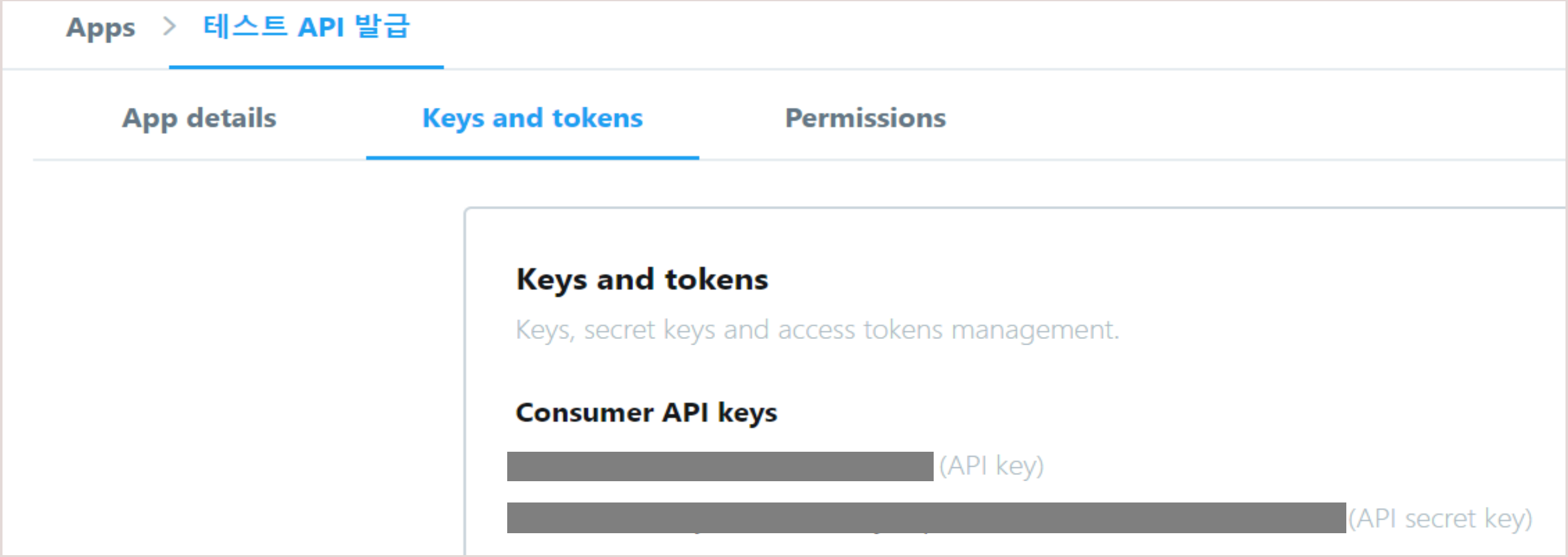
Minimum characters: **100**

앱의 이름, 앱의 설명, 사용하는 주소, 앱의 사용 방법에 대한 필수적인 요소를 작성합니다.

T-SA: Development Environment_W04

Twitter Keyword Search API based Tweet Analysis

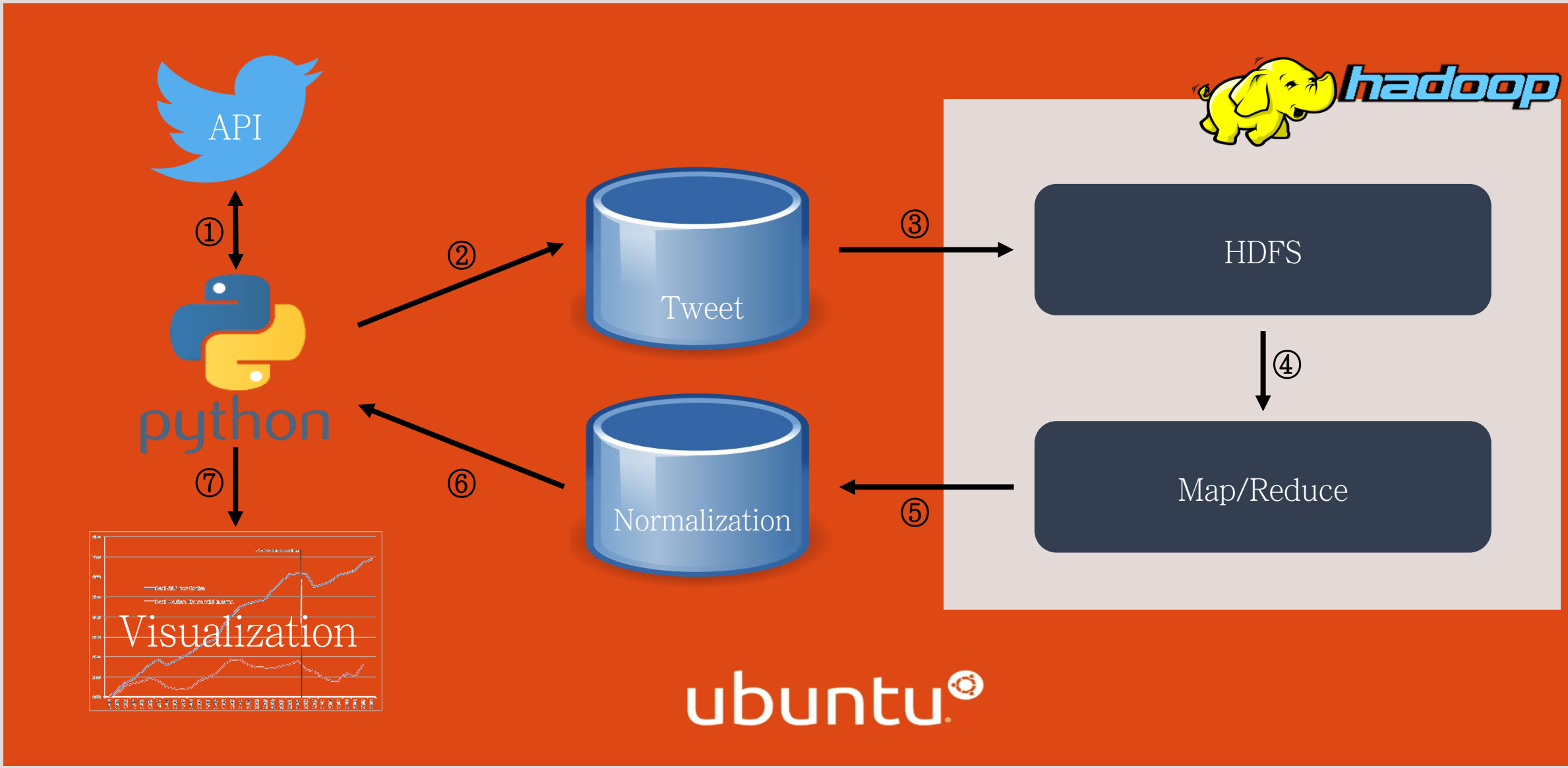
Importing Twitter API Key



개인 정보의 Apps를 통해 본인이 사전에 작성한 제목을 통해API 키가 발급이 된 것을 확인 할 수 있다.

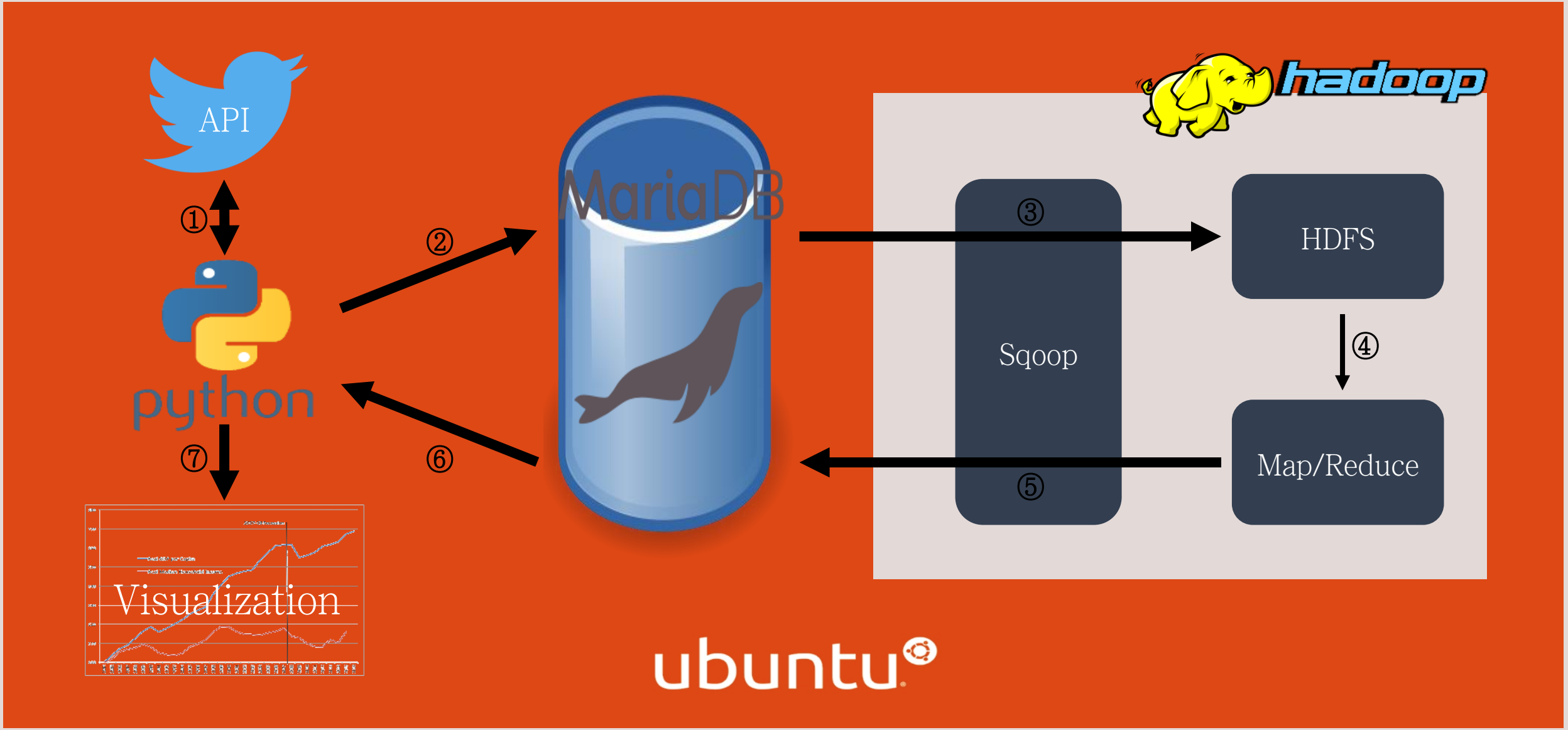
T-SA: Program Flowchart

Twitter Keyword Search API based Tweet Analysis



T-SA: Program Flowchart_W04

Twitter Keyword Search API based Tweet Analysis



T-SA: Program Flowchart(1)_W04

Twitter Keyword Search API based Tweet Analysis

Python에서 제공하는 TwitterAPI인 Tweepy를 이용한 데이터 크롤링

Search API를 이용한 키워드 관련 데이터 검색

Timeline API를 이용한 유저 관련 데이터 검색

T-SA: Program Flowchart(2)_W04

Twitter Keyword Search API based Tweet Analysis

크롤링된 데이터를 MariaDB에 저장

T-SA: Program Flowchart(3)_W04

Twitter Keyword Search API based Tweet Analysis

MariaDB에 저장된 데이터를 Sqoop을 이용해 HDFS에 분산 저장 처리

T-SA: Program Flowchart(4)_W04

Twitter Keyword Search API based Tweet Analysis

Map/Reduce를 이용해 분산 데이터 배치 처리

T-SA: Program Flowchart(5)_W04

Twitter Keyword Search API based Tweet Analysis

분산 배치 처리된 데이터를 Sqoop을 이용해 MariaDB로 저장

T-SA: Program Flowchart(6)_W04

Twitter Keyword Search API based Tweet Analysis

(5)과정을 통해 저장된 데이터를 Python에 로드

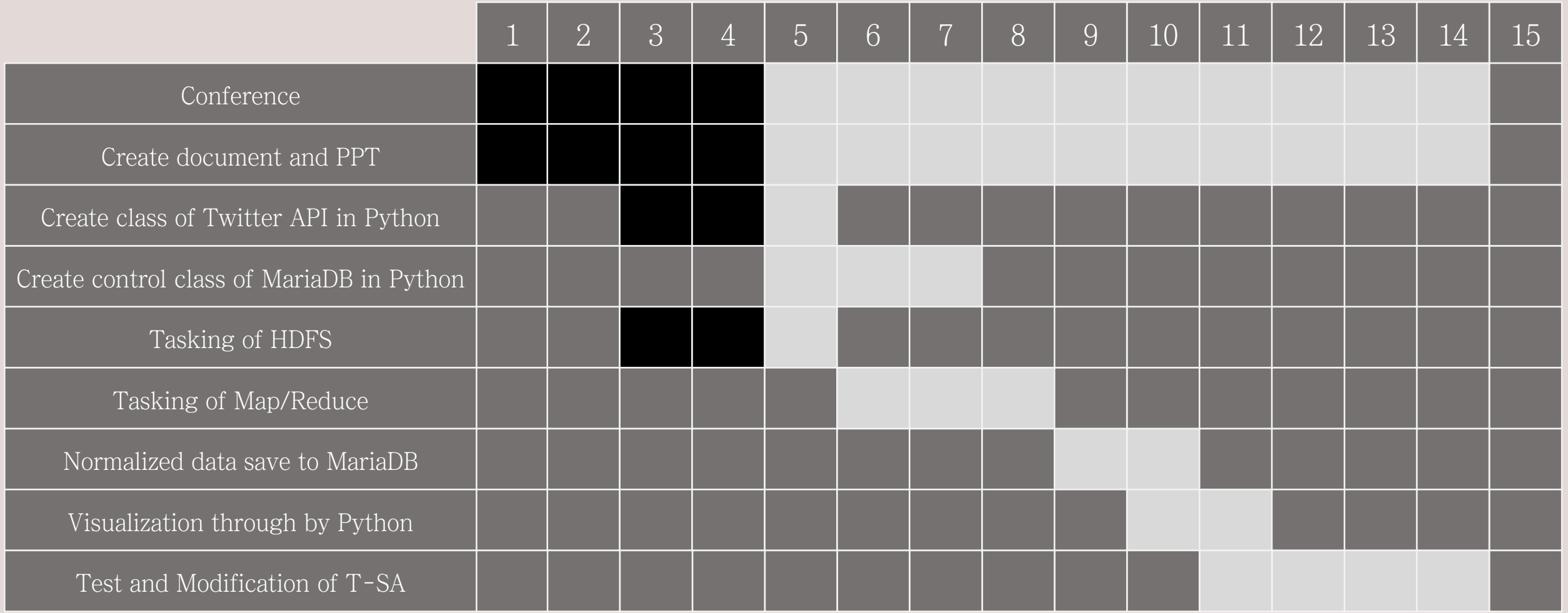
T-SA: Program Flowchart(7)_W04

Twitter Keyword Search API based Tweet Analysis

Python에서 제공되는 시각화 라이브러리를 이용한 데이터 시각화

T-SA: Development Schedule

Twitter Keyword Search API based Tweet Analysis



T-SA: Weekly Progress_W03

Twitter Keyword Search API based Tweet Analysis

Lee SeokJune

문서 작성 및 수정

HIVE 관련 자료 수집

HIVE-RDB 연동 자료 수집

Lee YunHyuck

문서 작성 및 수정

MariaDB 자료 수집

HIVE-RDB 연동 자료 수집

Bae InGyu

Twitter API 자료 수집, 정리 및 간단한 예제 테스트

발표 준비

Seo JaeIck

Twitter API 자료 수집, 정리 및 간단한 예제 테스트

T-SA: Weekly Progress_W04

Twitter Keyword Search API based Tweet Analysis

Lee SeokJune

문서 작성 및 수정

MariaDB 구축

Lee YunHyuck

Sqoop 구축

DB, Hadoop 연동

Bae InGyu

Python, MariaDB 연결 및
DML(select, delete) 구현

Seo JaeIck

발표 준비

T-SA: Weekly Progress_W04

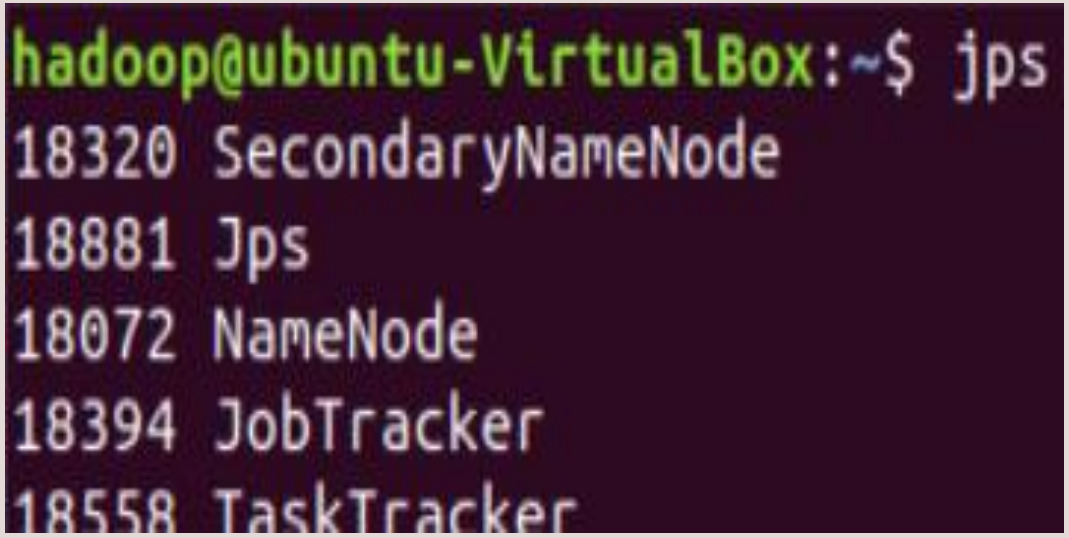
Twitter Keyword Search API based Tweet Analysis

Sqoop

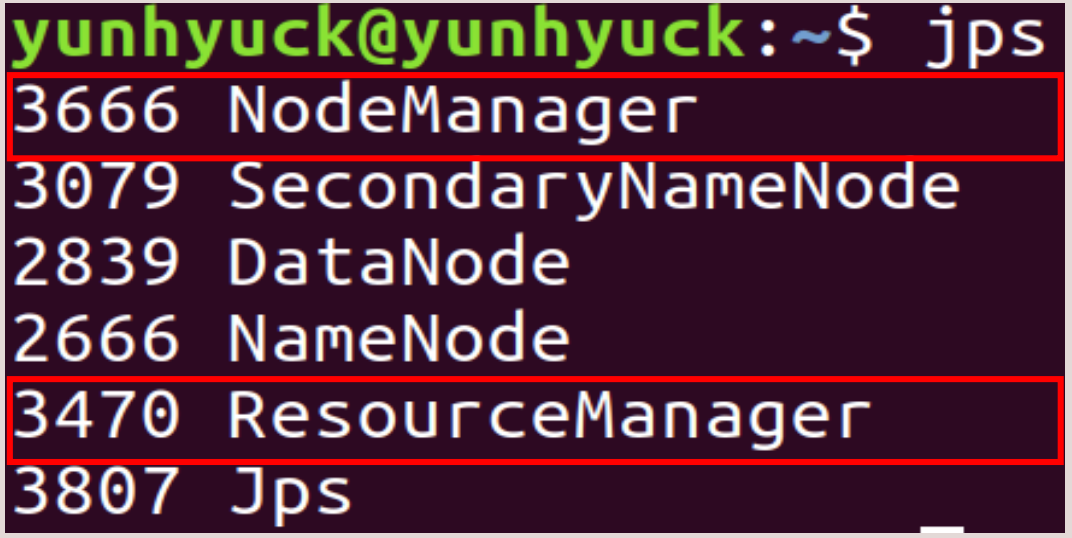
T-SA: Weekly Progress_W04

Twitter Keyword Search API based Tweet Analysis

Hadoop 1.2.1



Hadoop 3.2.0



JobTracker

- 클러스터 전체의 리소스 관리, 잡 스케쥴링 및 모니터링 기능 -> ResourceManager

TaskTracker

- slave node에서 map reduce작업을 수행 -> NodeManager

T-SA: Weekly Progress_W04

Twitter Keyword Search API based Tweet Analysis

Sqoop 실행

```
yunhyuck@yunhyuck:~$ sqoop
Warning: /home/yunhyuck/sqoop1/sqoop-1.4.7/../hbase does not exist! HBase imports will fail.
Please set $HBASE_HOME to the root of your HBase installation.
Warning: /home/yunhyuck/sqoop1/sqoop-1.4.7/../hcatalog does not exist! HCatalog jobs will fail.
Please set $HCAT_HOME to the root of your HCatalog installation.
Warning: /home/yunhyuck/sqoop1/sqoop-1.4.7/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
Warning: /home/yunhyuck/sqoop1/sqoop-1.4.7/../zookeeper does not exist! Accumulo imports will fail.
Please set $ZOOKEEPER_HOME to the root of your Zookeeper installation.
/home/yunhyuck/hadoop/libexec/hadoop-functions.sh: 줄 2364: HADOOP_ORG.APACHE.SQOOP.SQOOP_USER: bad substitution
/home/yunhyuck/hadoop/libexec/hadoop-functions.sh: 줄 2459: HADOOP_ORG.APACHE.SQOOP.SQOOP_OPTS: bad substitution
Try 'sqoop help' for usage.
```

T-SA: Weekly Progress_W04

Twitter Keyword Search API based Tweet Analysis

Sqoop Job 실행

sqoop import --connect "jdbc:mysql://localhost/yunhyuck" --username yunhyuck -P --table “test" -m 1

```
2019-04-03 17:31:10,195 INFO mapreduce.Job: Running job: job_1554280042468_0001
2019-04-03 17:31:22,362 INFO mapreduce.Job: Job job_1554280042468_0001 running in uber mode : false
2019-04-03 17:31:22,363 INFO mapreduce.Job:  map 0% reduce 0%
2019-04-03 17:31:32,455 INFO mapreduce.Job:  map 100% reduce 0%
2019-04-03 17:31:33,474 INFO mapreduce.Job: Job job_1554280042468_0001 completed successfully
2019-04-03 17:31:33,553 INFO mapreduce.Job: Counters: 33
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=232265
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=87
                HDFS: Number of bytes written=2
                HDFS: Number of read operations=6
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Launched map tasks=1
                Other local map tasks=1
                Total time spent by all maps in occupied slots (ms)=7400
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=7400
                Total vcore-milliseconds taken by all map tasks=7400
                Total megabyte-milliseconds taken by all map tasks=7577600
```

```
        Map-Reduce Framework
                Map input records=1
                Map output records=1
                Input split bytes=87
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=57
                CPU time spent (ms)=1060
                Physical memory (bytes) snapshot=194822144
                Virtual memory (bytes) snapshot=2671808512
                Total committed heap usage (bytes)=173015040
                Peak Map Physical memory (bytes)=194822144
                Peak Map Virtual memory (bytes)=2671808512
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=2
2019-04-03 17:31:33,558 INFO mapreduce.ImportJobBase: Transferred 2 bytes in 29.6853 seconds (0.0674 bytes/sec)
2019-04-03 17:31:33,561 INFO mapreduce.ImportJobBase: Retrieved 1 records.
```

T-SA: Weekly Progress_W04

Twitter Keyword Search API based Tweet Analysis

Sqoop Job 결과 확인

http://localhost:9870/

Browse Directory

/user/yunhyuck/test4 Go!

Show 25 entries Search:

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	yunhyuck	supergroup	0 B	Apr 03 17:31	1	128 MB	_SUCCESS
-rw-r--r--	yunhyuck	supergroup	2 B	Apr 03 17:31	1	128 MB	part-m-00000

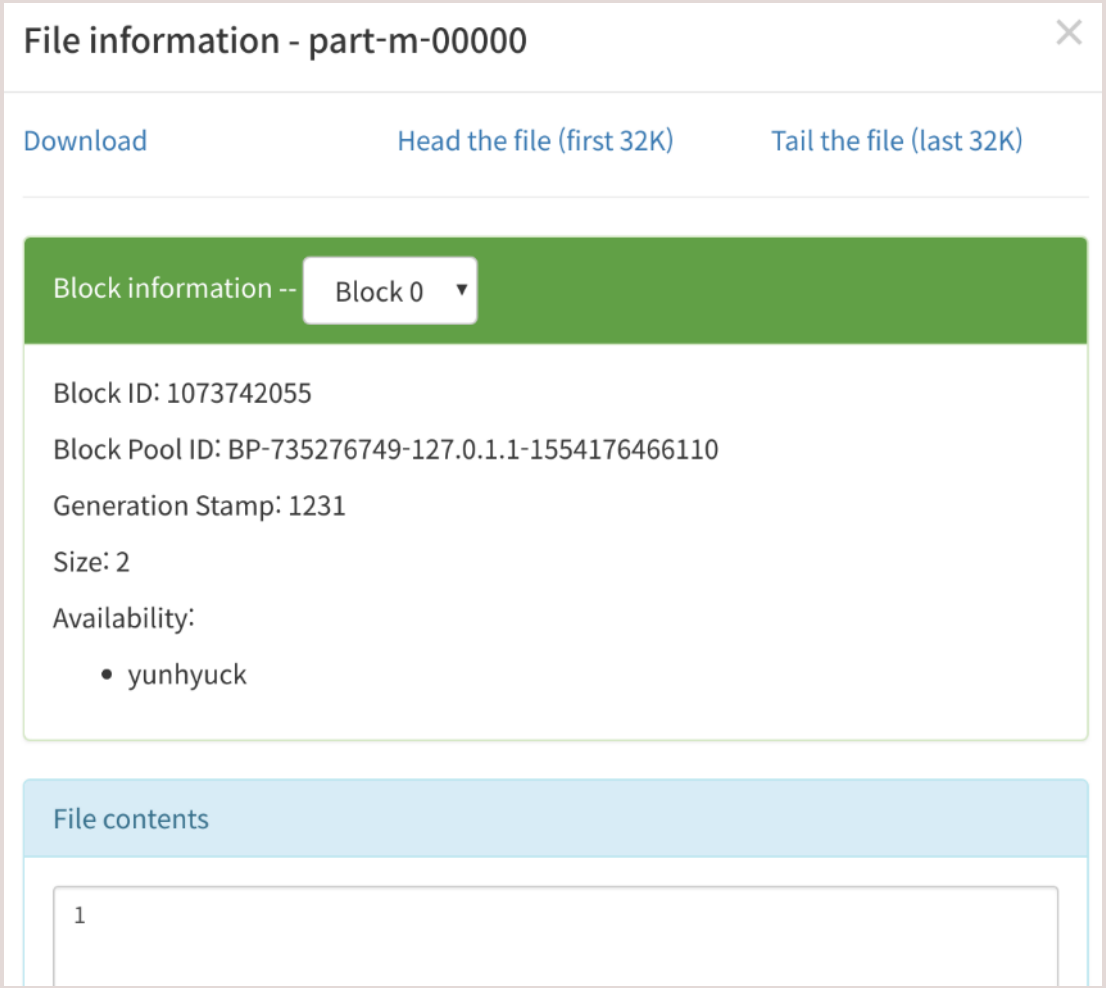
Showing 1 to 2 of 2 entries

Previous 1 Next

T-SA: Weekly Progress_W04

Twitter Keyword Search API based Tweet Analysis

Sqoop Job 상세 결과 확인



T-SA: Weekly Progress_W04

Twitter Keyword Search API based Tweet Analysis

Python(DB)

T-SA: Weekly Progress_W04

Twitter Keyword Search API based Tweet Analysis

DBModule.selectDB

```
# 테이블의 데이터 모두조회
def selectDB (table) :
        # MariaDB연결 및 Cursor 생성
        conn = dbModule.dbconn()
        curs = conn.cursor()

        # 테이블 조회
        sql = "select * from "+table+";"
        curs.execute(sql)


        # 테이블 데이터출력
        i = 0
        rows = curs.fetchall()
        result = rows

        # Select종료시 MariaDB연결종료
        conn.close()

        return print("조회완료\n", result)
```

```
vi@vi-X510UQR:~/다운로드$ python3 dbMain.py
데이터 모두 조회할 테이블명 입력: Student
조회완료
 ((1, 'Kim', '010-0000-0001', 'Computer', 'A'), (2, 'Lee', '010-0000-0002', 'Computer', 'A+'), (3, 'Pack', '010-0000-0003', 'Computer', 'C+')
, (4, 'Oh', '010-0000-0004', 'Computer', 'B'), (5, 'Song', '010-0000-0005', 'Computer', 'F'), (6, 'Yoon', '010-0000-0006', 'Computer', 'D'))
```

T-SA: Weekly Progress_W04

Twitter Keyword Search API based Tweet Analysis

DBModule.deleteDB

```
# 테이블의 데이터 모두삭제
def deleteDB (table) :
        # MariaDB연결 및 Cursor 생성
        conn = dbModule.dbconn()
        curs = conn.cursor()

        # Data삭제
        sql = "delete from "+table+";"
        curs.execute(sql)
        conn.commit()

        # Connection 닫기
        conn.close()

        return print("삭제완료")
```

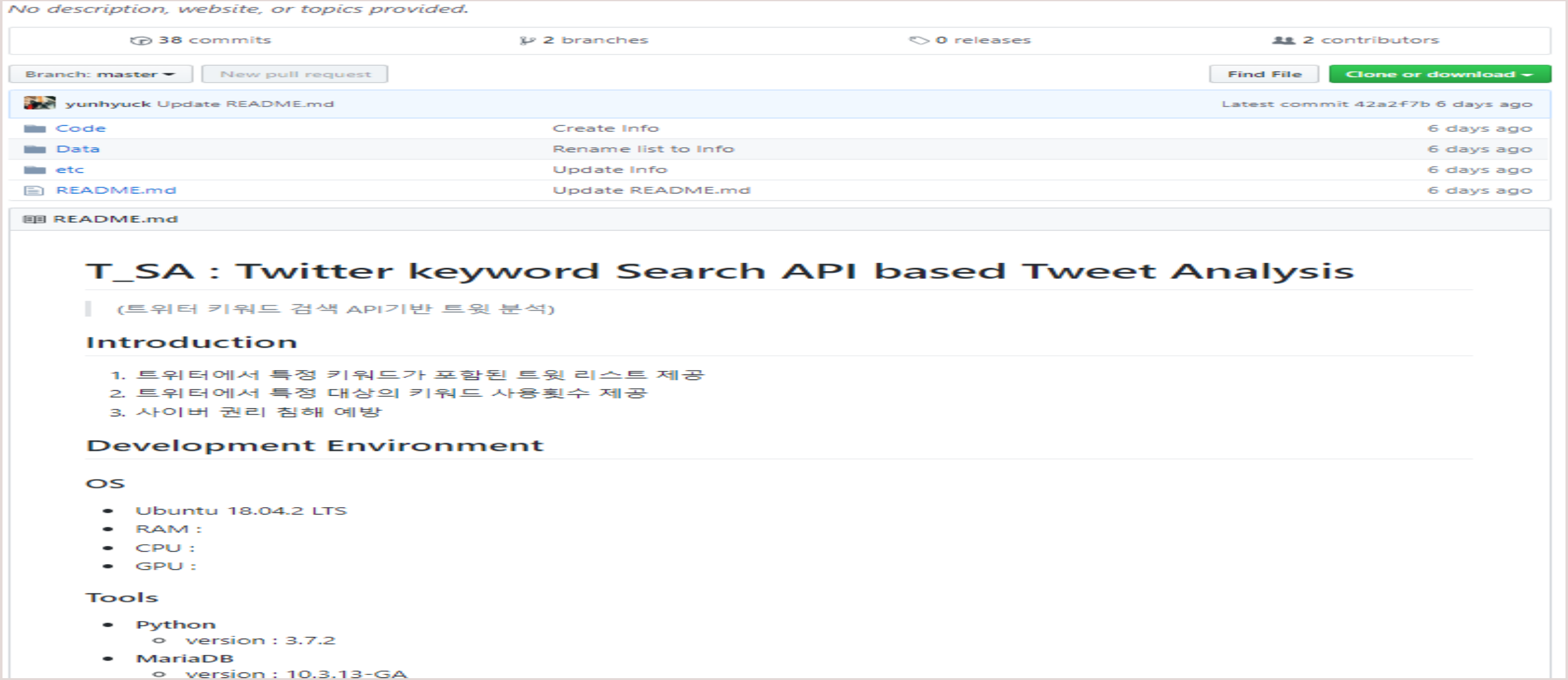
```
데이터 전부 삭제할 테이블명 입력: Student
삭제완료
```

```
MariaDB [mysql]> select * from Student;
Empty set (0.00 sec)
```

T-SA: Github

Twitter Keyword Search API based Tweet Analysis

Project Github URL: https://github.com/SeokJune/BigData_VI_T-SA/



No description, website, or topics provided.

38 commits | 2 branches | 0 releases | 2 contributors

Branch: master | New pull request | Find File | Clone or download

yunhyuck Update README.md — Latest commit 42a2f7b 6 days ago

Code	Create Info	6 days ago
Data	Rename list to Info	6 days ago
etc	Update Info	6 days ago
README.md	Update README.md	6 days ago

README.md

T_SA : Twitter keyword Search API based Tweet Analysis

(트위터 키워드 검색 API기반 트윗 분석)

Introduction

1. 트위터에서 특정 키워드가 포함된 트윗 리스트 제공
2. 트위터에서 특정 대상의 키워드 사용횟수 제공
3. 사이버 권리 침해 예방

Development Environment

OS

- Ubuntu 18.04.2 LTS
- RAM :
- CPU :
- GPU :

Tools

- Python
  - version : 3.7.2
- MariaDB
  - version : 10.3.13-GA

T-SA: Github_W04

Twitter Keyword Search API based Tweet Analysis

Project Github URL: https://github.com/SeokJune/BigData_VI_T-SA/

SeokJune / BigData_VI_T-SA

Watch 0 | Unstar 3 | Fork 0

Code | Issues 0 | Pull requests 0 | Projects 0 | Wiki | Insights | Settings

- Pulse
- Contributors
- Community
- Traffic
- Commits
- Code frequency
- Dependency graph
- Alerts
- Network
- Forks

March 3, 2019 – April 3, 2019

Period: 1 month

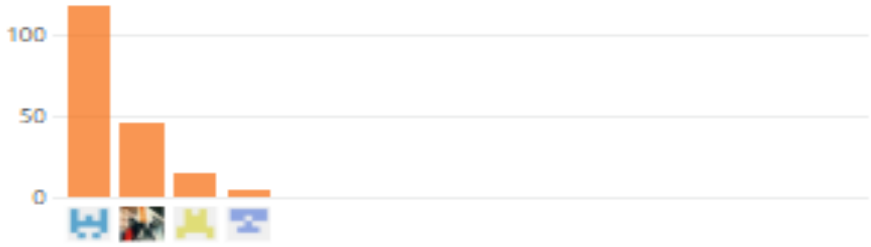
Overview

0 Active Pull Requests

0 Active Issues

0 Merged Pull Requests	0 Proposed Pull Requests	0 Closed Issues	0 New Issues

Excluding merges, **4 authors** have pushed **182 commits** to master and **183 commits** to all branches. On master, **72 files** have changed and there have been 2,964 **additions** and 0 **deletions**.



Q & A

Thank you.