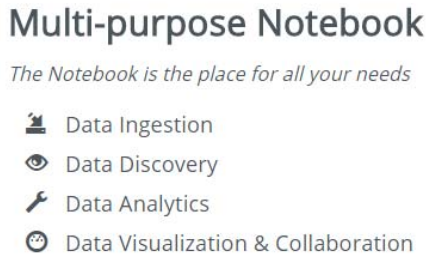
제플린 노트북

순천향대학교 컴퓨터공학과

이 상 정

아파치 제플린 (Apache Zeppelin) (1)

- 제플린 (Zeppelin)은 웹 기반의 노트북 (web-based notebook)
  - https://zeppelin.apache.org
  - 웹 상에서 코드를 작성-실행-결과확인-코드수정을 반복하면서 최종 원하는 결과를 생성하는 작업 환경
  - 데이터 수집, 발견, 분석, 시각화 & 협력작업 지원



아파치 제플린 (Apache Zeppelin) (2)

- 웹 상에서 파이썬, 스칼라 등을 플러그인하여 코드 작성
- 스파크 뿐만 아니라 Livy, Cassandra, Lens, SQL 등등의 다른 데이터 분석도구나 데이터베이스에 접근하여 질의 등의 확장 기능 지원

제플린 설치 – 다운로드

- 마스터 서버에 제플린 다운로드
  - http://mirror.navercorp.com/apache/zeppelin/zeppelin-0.8.2/zeppelin-0.8.2-bin-all.tgz
    - 최신 버전 0.8.2 (2020년 3월 기준) 설치

```
$ wget http://mirror.apache-kr.org/zeppelin/zeppelin-0.8.2/zeppelin-0.8.2-bin-all.tgz
```

- 압축 해제

```
$ tar -xvzf zeppelin-0.8.2-bin-all.tgz
```

제플린 설치 – 설정 (1)

- 기존 제플린 설정 파일들 복사 & 변경
  - 경로 : ~/zeppelin-0.8.2-bin-all/conf
  - zeppelin-site.xml.template
  - zeppelin-env.sh.template
  - zeppelin-env.cmd.template
  - shiro.ini.template

```
$ cd ~/zeppelin-0.8.2-bin-all/conf

$ cp zeppelin-site.xml.template zeppelin-site.xml
$ cp zeppelin-env.sh.template zeppelin-env.sh
$ cp zeppelin-env.cmd.template zeppelin-env.cmd
$ cp shiro.ini.template shiro.ini
```


제플린 설치 – 설정 (2)

- 제플린 계정 설정
  - ~/zeppelin-0.8.2-bin-all/conf/shiro.ini 파일
  - admin / password1, user1 / password2

```
[users]
admin = password1, admin
user1 = password2, role1, role2
user2 = password3, role3
user3 = password4, role2
```

- 제플린 서버 주소와 웹 포트 설정
  - ~/zeppelin-0.8.2-bin-all/conf/zeppelin-site.xml 설정 파일
  - addr, port 속성 변경

```
<property>
  <name>zeppelin.server.addr</name>
  <value>192.168.0.200</value>
  <description>Server address</description>
</property>
<property>
  <name>zeppelin.server.port</name>
  <value>17000</value>
  <description>Server port.</description>
</property>
```

제플린 설치 – 설정 (3)

❑ 제플린 환경 설정

- ~/zeppelin-0.8.2-bin-all/conf/zeppelin-env.sh 파일

```
export JAVA_HOME=/usr/lib/jvm/jdk1.8.0_201
export MASTER=spark://master:7077
export SPARK_HOME=/home/bigdata/spark-2.4.5-bin-hadoop2.7
export HADOOP_CONF_DIR=/home/bigdata/hadoop-2.7.7/etc/hadoop
```

❑ ~/.bash 환경 설정

- *배포한 가상머신 이미지에 설정 안되었으니 설정 필요*
- 추가 후 아래 명령을 통해 적용
  $ source ~/.bashrc

```
# Zeppelin Path
export ZEPPELIN_HOME=$HOME/zeppelin-0.8.2-bin-all
```


제플린에서의 스파크 실행 에러 보정 – Netty

❑ 제플린(0.8.2)에서 스파크(2.4.5) 실행 시 발생하는 에러 보정

❑ Netty 에러 보정

- Netty는 자바 네트워크 응용 개발을 지원하는 비동기 이벤트 기반 네트워크 응용 프레임워크
- 스파크 버전의 Netty jar를 제플린에 복사
  - ~/spark-2.4.5-bin-hadoop2.7/jars/netty-all-4.1.42.Final.jar 를 ~/zeppelin-0.8.2-bin-all/lib에 복사
- 제플린의 기존 netty-all-4.0.23.Final.jars는 제거(백업)

```
$ cp $SPARK_HOME/jars/netty-all-4.1.42.Final.jar $ZEPPELIN_HOME/lib

$ mv $ZEPPELIN_HOME/lib/netty-all-4.0.23.Final.jar $ZEPPELIN_HOME/lib/netty-all-4.0.23.Final.jar-old
```

제플린에서의 스파크 실행 에러 보정 – Jackson

❑ Jackson 에러 보정

- Jackson은 JSON, XML 등 다양한 형식을 지원하는 자바 라이브러리
- 스파크 버전의 Jackon jar들을 제플린에 복사
  - ~/spark-2.4.5-bin-hadoop2.7/jars
    - jackson-annotations-2.6.7.jar
    - jackson-core-2.6.7.jar
    - jackson-databind-2.6.7.3.jar
  - 제플린의 기존의 Jackson jar들은 제거(백업)

```
$ cp $SPARK_HOME/jars/jackson-annotations-2.6.7.jar $ZEPPELIN_HOME/lib
$ cp $SPARK_HOME/jars/jackson-core-2.6.7.jar $ZEPPELIN_HOME/lib
$ cp $SPARK_HOME/jars/jackson-databind-2.6.7.1.jar $ZEPPELIN_HOME/lib

$ mv $ZEPPELIN_HOME/lib/jackson-annotations-2.8.0.jar $ZEPPELIN_HOME/lib/jackson-annotations-2.8.0.jar-old
$ mv $ZEPPELIN_HOME/lib/jackson-core-2.8.10.jar $ZEPPELIN_HOME/lib/jackson-core-2.8.10.jar-old
$ mv $ZEPPELIN_HOME/lib/jackson-databind-2.8.11.1.jar $ZEPPELIN_HOME/lib/jackson-databind-2.8.11.1.jar-old
```


제플린에서의 스파크 실행 에러 보정 – commons-lang

❑ commons-lang 에러 보정

- 스파크 버전 commons-lang3-3.5.jar를 제플린에 복사
- 제플린의 기존 commons-lang3-3.4.jar는 제거(백업)

```
$ cp $SPARK_HOME/jars/commons-lang3-3.5.jar $ZEPPELIN_HOME/lib

$ mv $ZEPPELIN_HOME/lib/commons-lang3-3.4.jar $ZEPPELIN_HOME/lib/commons-lang3-3.4.jar-old
```

제플린 서버 실행

- 실행
  - 위치 : ~/zeppelin-0.8.2-bin-all/bin
  - zeppelin-daemon.sh 실행
    - 제플린 서버 시작

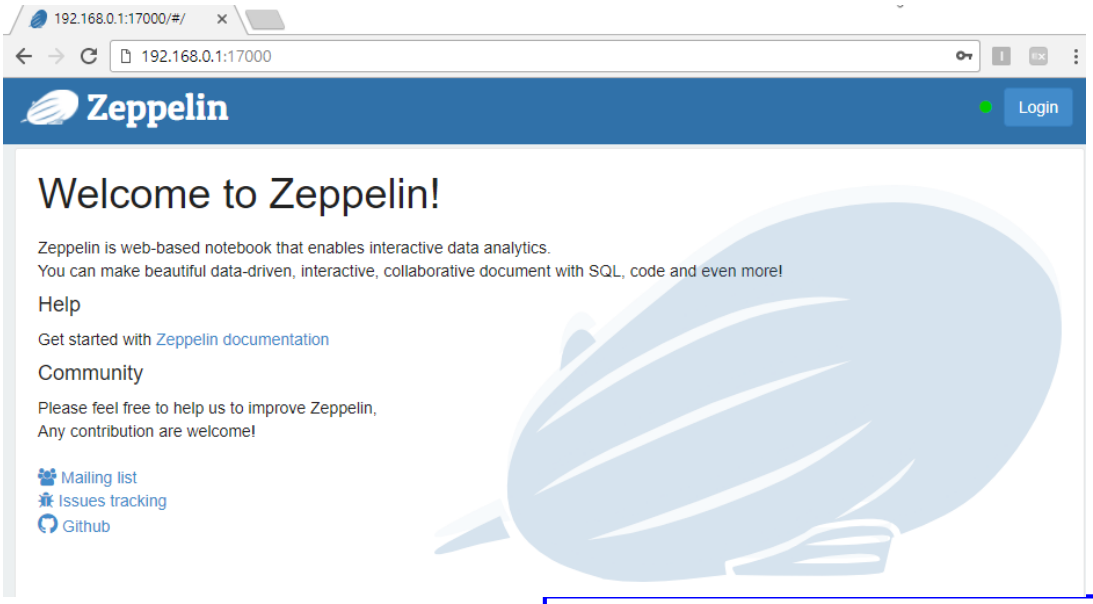
```
$ $ZEPPELIN_HOME/bin/zeppelin-daemon.sh start
```

  - 종료 시에는 stop, 재시작 시에는 restart

```
bigdata@master:~$ $ZEPPELIN_HOME/bin/zeppelin-daemon.sh start
Log dir doesn't exist, create /home/bigdata/zeppelin-0.8.0-bin-all/logs
Pid dir doesn't exist, create /home/bigdata/zeppelin-0.8.0-bin-all/run
Zeppelin start                                             [  OK  ]
bigdata@master:~$
bigdata@master:~$ jps
3168 NodeManager
2977 ResourceManager
2518 NameNode
4875 Master
2699 DataNode
10364 Jps
10335 ZeppelinServer
bigdata@master:~$
```

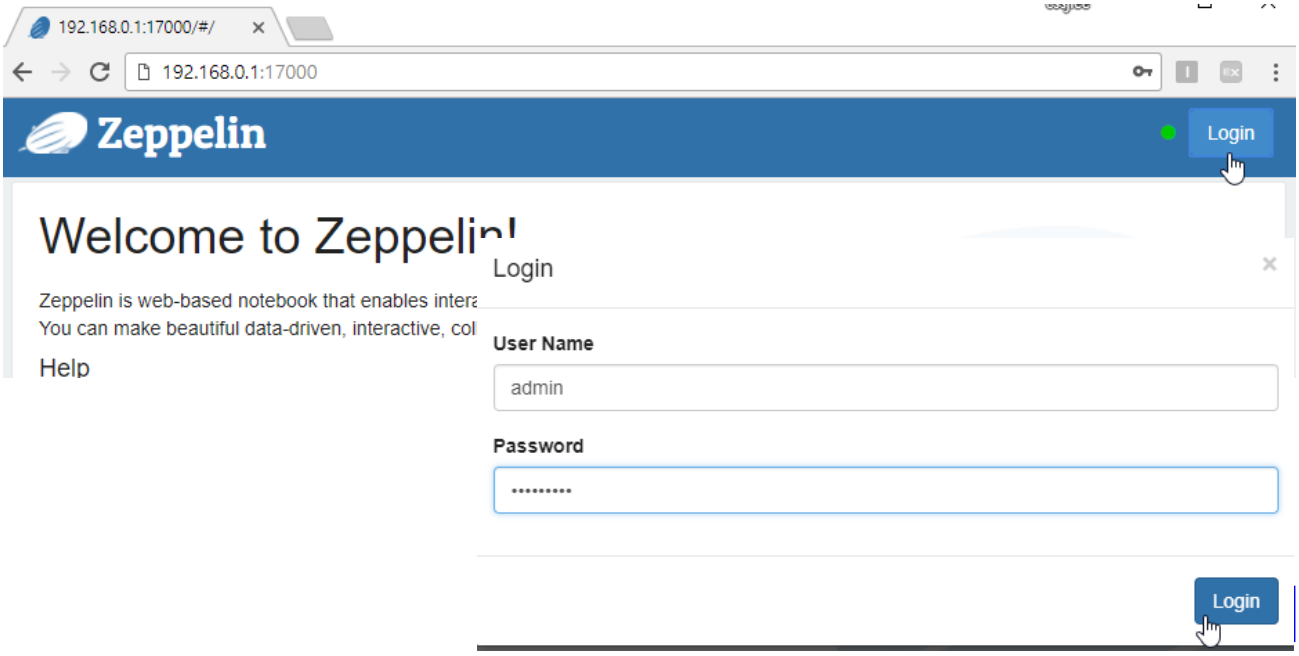

제플린 서버 접속

- zeppelin-site.xml 에서 설정한 포트 17000 으로 접속
  - http://192.168.0.1:17000



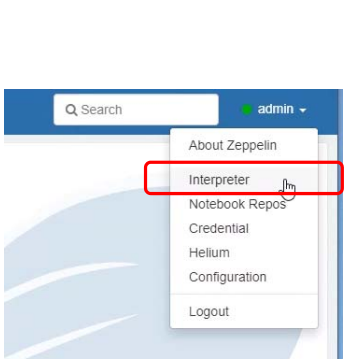
제플린 로그인

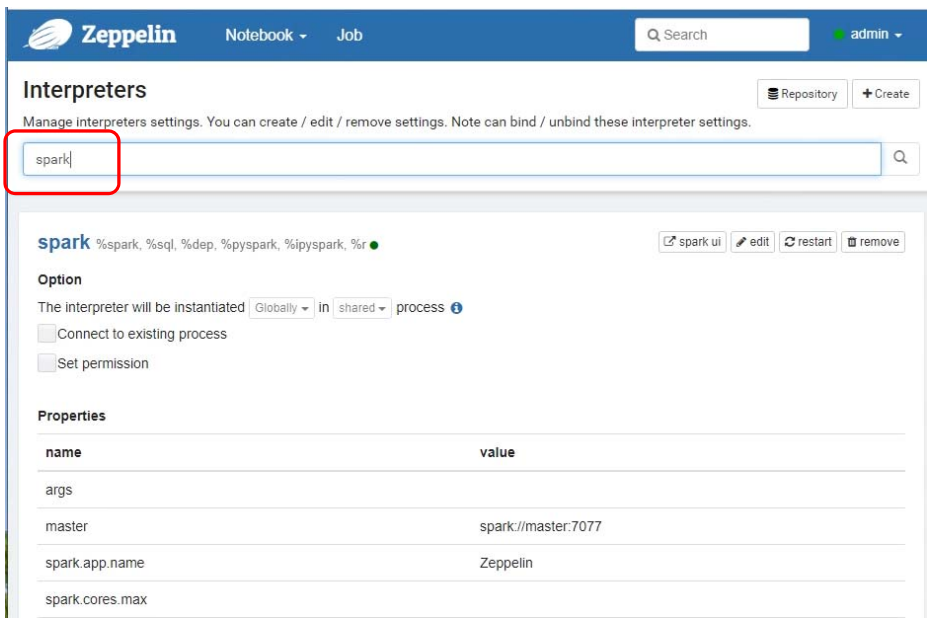
- 계정 설정(shiro.ini) 에서 설정한 사용자로 접속
  - admin / password1, user1 / password2



제플린 스파크 설정 (1)

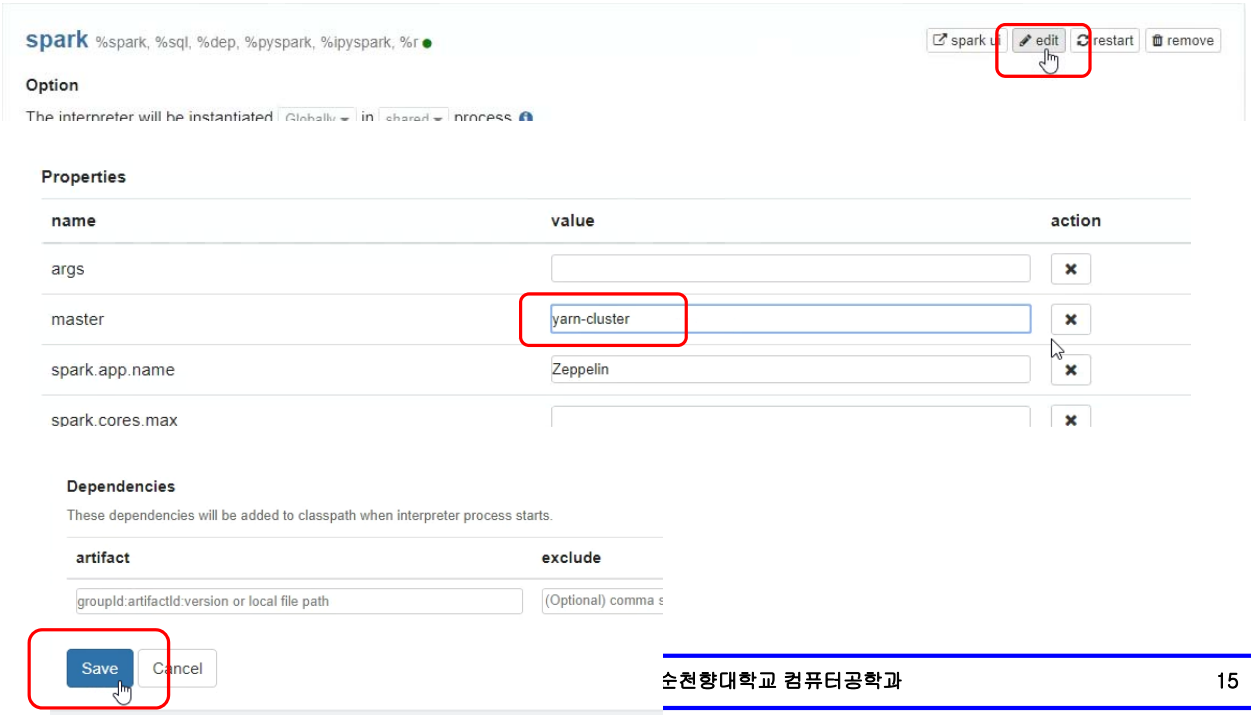
- 스파크 인터프리터 설정 확인
  - Interpretes에서 spark 검색하여 스파크 설정 확인
    - admin 계정



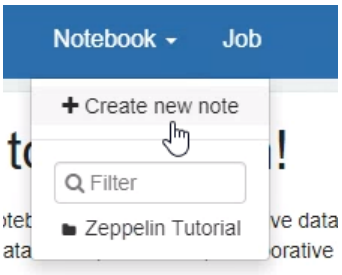


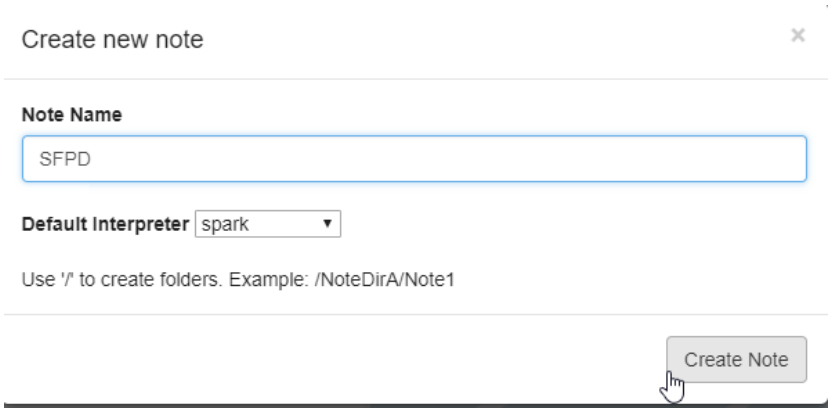
제플린 스파크 설정 (2)

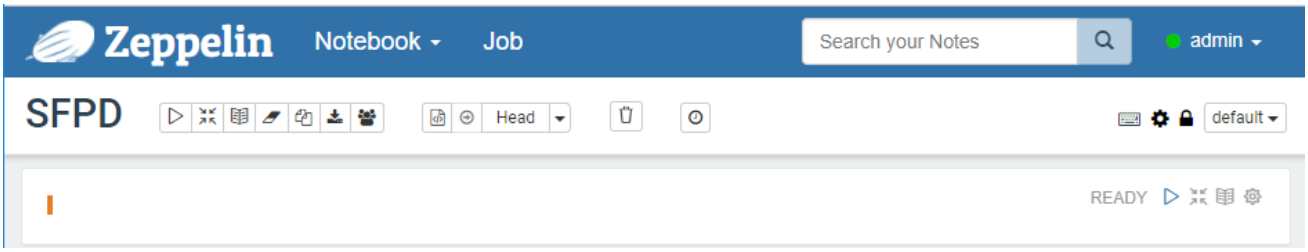
❑ master 속성을 yarn-cluster로 설정



SFPD 노트북 생성

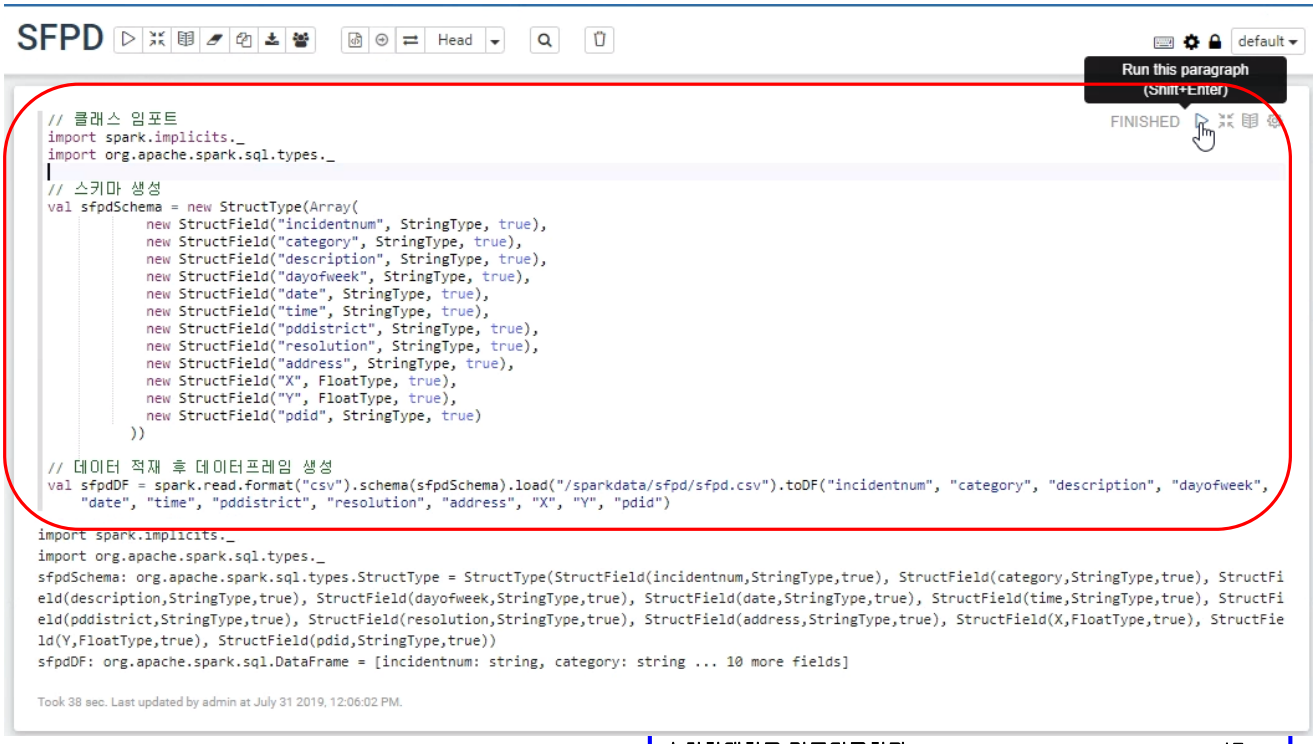






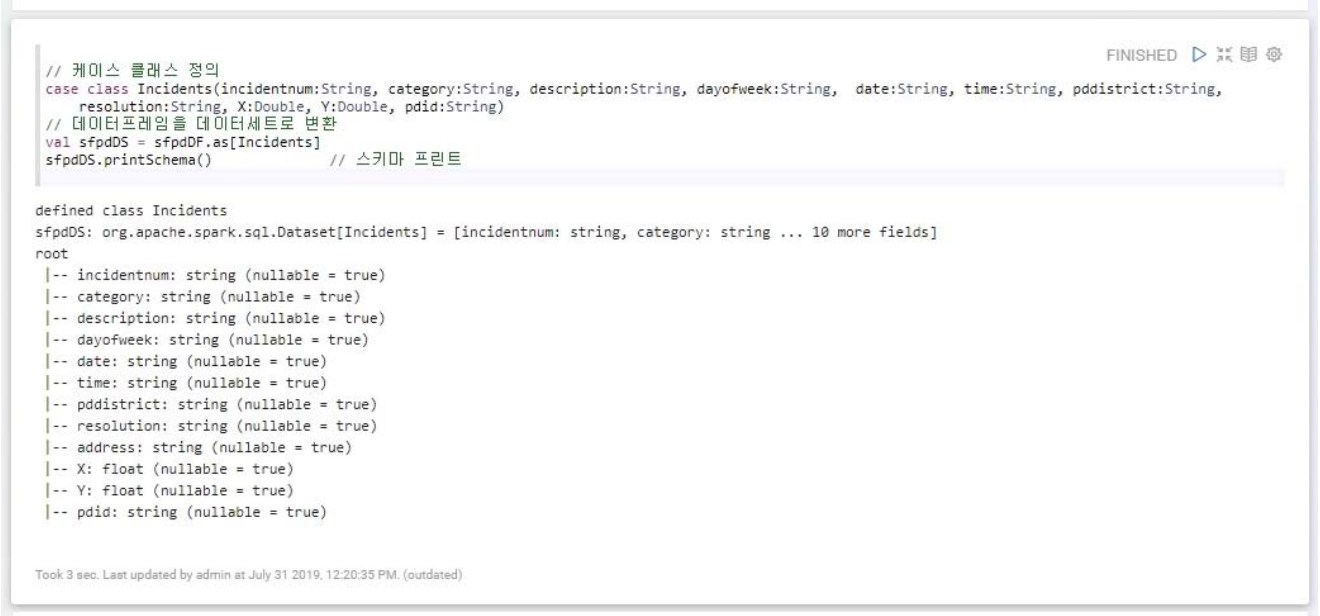
제플린 노트북

SFPD 데이터프레임 생성 실행 (1)



제플린 노트북

SFPD 데이터프레임 생성 실행 (2)



SFPD 데이터프레임 생성 실행 (3)

실습 시작/종료 셀 (1)

- 마스터에 하둡/스파크/제플린 실습 시작/종료 셀
  - 시작: ~/start-lab.sh

```
# start Hadoop
start-all.sh
# start Spark
$SPARK_HOME/sbin/start-all.sh
# start Zeppelin
$ZEPPELIN_HOME/bin/zeppelin-daemon.sh start
```

  - 종료: ~/stop-lab.sh

```
# stop Zeppelin
$ZEPPELIN_HOME/bin/zeppelin-daemon.sh stop
# stop Spark
$SPARK_HOME/sbin/stop-all.sh
# stop Hadoop
stop-all.sh
```

실습 시작/종료 셸 (2)

- 셸 작성 후 실행 권한 부여
  - $ chmod +x start-lab.sh
  - $ chmod +x stop-lab.sh

❑ 실습 시작

- $ ~/start-lab.sh

```
bigdata@master:~$ ~/start-lab.sh
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
Starting namenodes on [master]
master: starting namenode, logging to /home/bigdata/hadoop-2.7.7/logs/hadoop-bigdata-namenode-master.out
slave1: starting datanode, logging to /home/bigdata/hadoop-2.7.7/logs/hadoop-bigdata-datanode-slave1.out
master: starting datanode, logging to /home/bigdata/hadoop-2.7.7/logs/hadoop-bigdata-datanode-master.out
Starting secondary namenodes [slave1]
slave1: starting secondarynamenode, logging to /home/bigdata/hadoop-2.7.7/logs/hadoop-bigdata-secondarynamenode-slave
1.out
starting yarn daemons
starting resourcemanager, logging to /home/bigdata/hadoop-2.7.7/logs/yarn-bigdata-resourcemanager-master.out
slave1: starting nodemanager, logging to /home/bigdata/hadoop-2.7.7/logs/yarn-bigdata-nodemanager-slave1.out
master: starting nodemanager, logging to /home/bigdata/hadoop-2.7.7/logs/yarn-bigdata-nodemanager-master.out
starting org.apache.spark.deploy.master.Master, logging to /home/bigdata/spark-2.3.3-bin-hadoop2.7/logs/spark-bigdata
-org.apache.spark.deploy.master.Master-1-master.out
slave1: starting org.apache.spark.deploy.worker.Worker, logging to /home/bigdata/spark-2.3.3-bin-hadoop2.7/logs/spark
-bigdata-org.apache.spark.deploy.worker.Worker-1-slave1.out
Zeppelin start                                             [  OK  ]
/home/bigdata/start-lab.sh: line 7: /home/bigdata: Is a directory
bigdata@master:~$
```


실습 시작/종료 셸 (3)

❑ 실습 종료

- $ ~/stop-lab.sh

```
bigdata@master:~$ ~/stop-lab.sh
Zeppelin stop                                              [  OK  ]
slave1: stopping org.apache.spark.deploy.worker.Worker
stopping org.apache.spark.deploy.master.Master
This script is Deprecated. Instead use stop-dfs.sh and stop-yarn.sh
Stopping namenodes on [master]
master: stopping namenode
slave1: stopping datanode
master: stopping datanode
Stopping secondary namenodes [slave1]
slave1: stopping secondarynamenode
stopping yarn daemons
stopping resourcemanager
master: stopping nodemanager
slave1: stopping nodemanager
no proxyserver to stop
bigdata@master:~$
bigdata@master:~$
```

실습 과제

- 강의 시간의 제플린 실습 내용을 정리하여 제출

텀 프로젝트 과제 – 제플린

- 텀 프로젝트 데이터의 스파크 데이터세트 생성을 제플린에서 실행

참고 자료

- 아파치 제플린
  - https://zeppelin.apache.org