THE JOEUN JEWELERY

빅데이터 개발환경을 사용한

쇼핑몰 상품 추천 시스템 구축

임지안
백종성
이지훈
김현강
지현규

CONTENTS

01 HYPOTHESIS & TARGET

02 UML DIAGRAM

03 ERD

04 WORK FLOW

05 SCHEDULE & WORK DIVISION

06 WEB SKILLS

07 SYSTEM SKILLS

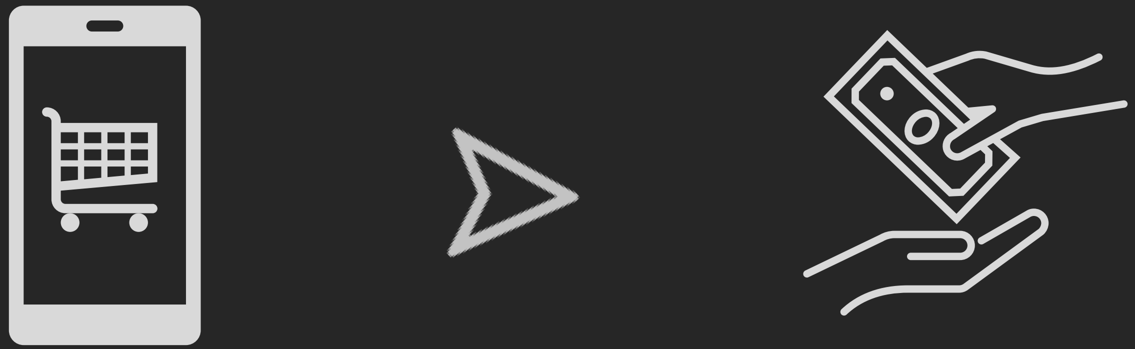
08 ANALYSIS

09 HYPOTHESIS

01.

HYPOTHESIS & TARGET

“ 추천된 상품을 구매할 확률이 높다 ”



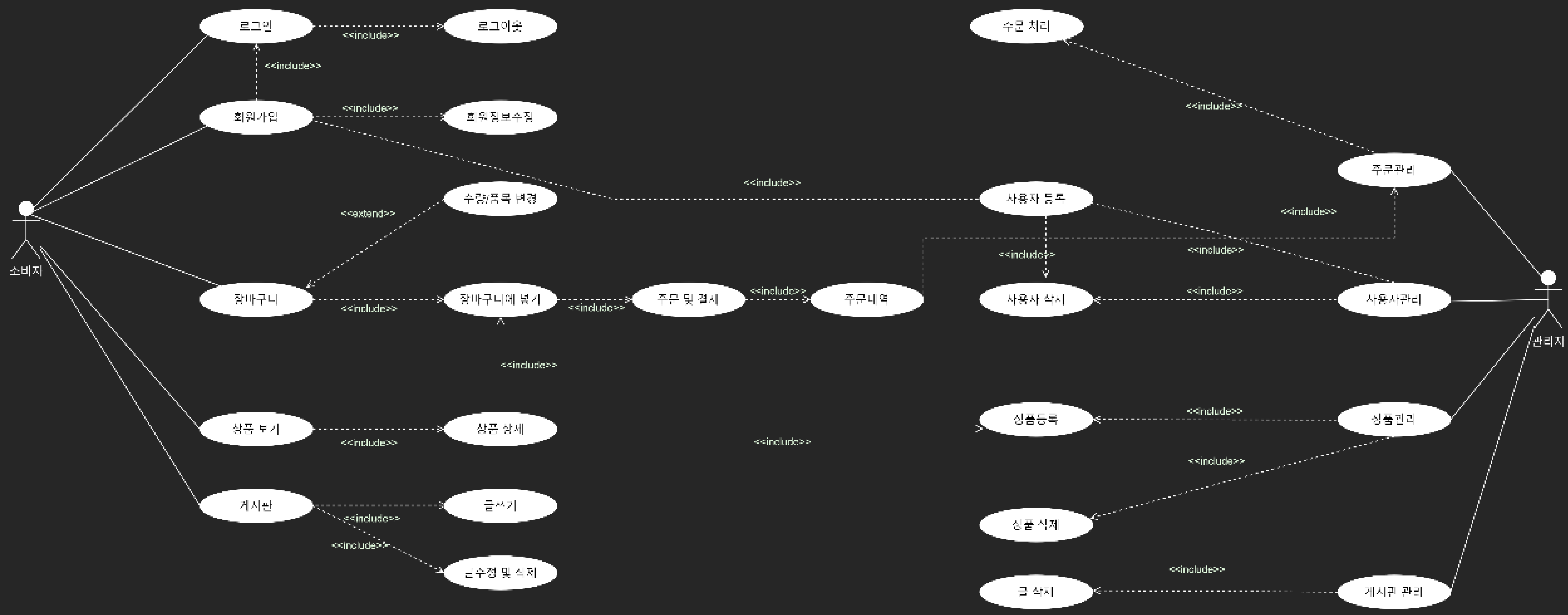
01.
HYPOTHESIS & TARGET

구매내역 데이터를 통한
협업필터링 모델 생성

회원 구매내역 기반
상품 추천

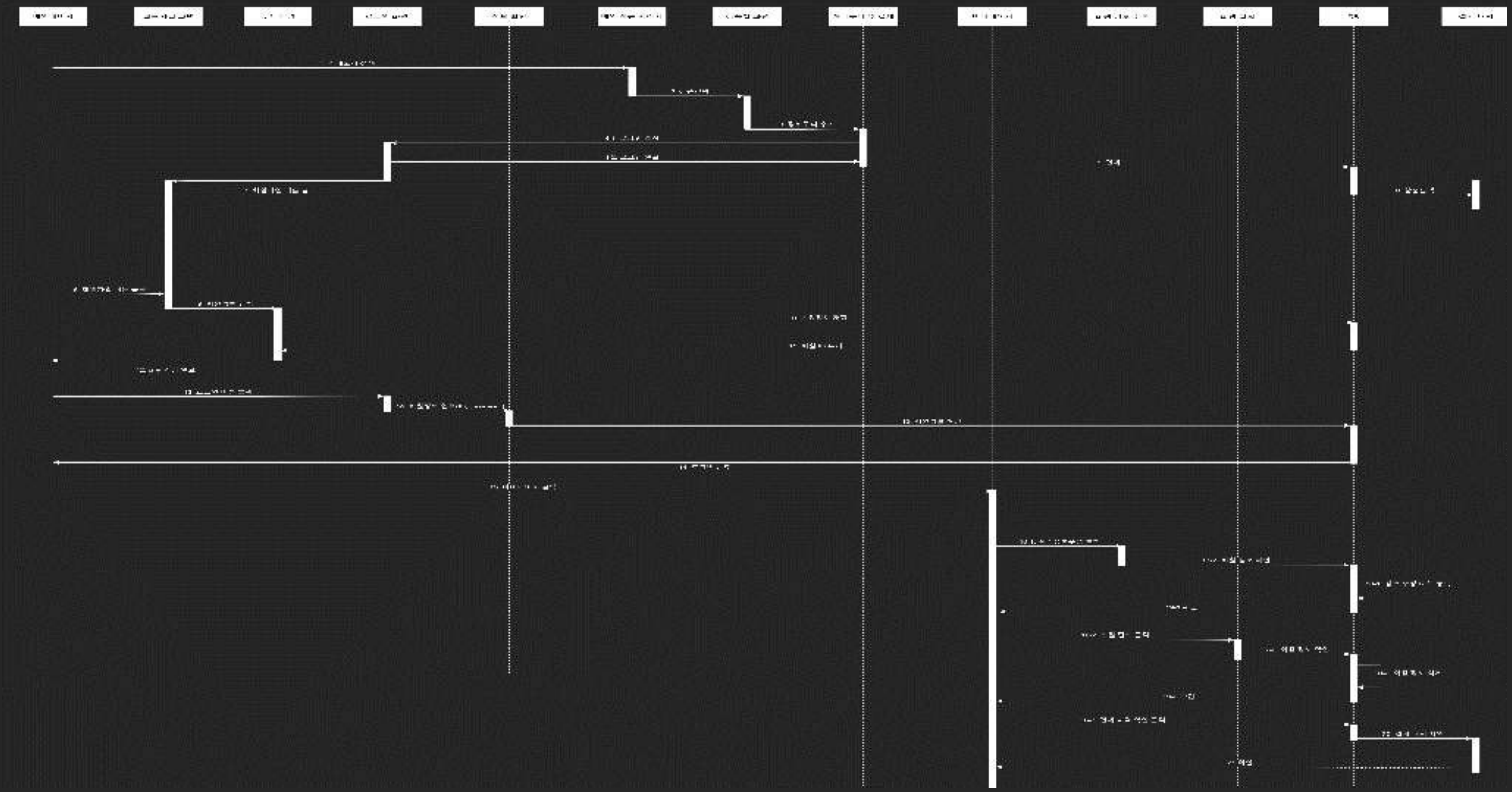
소비자의 취향에 따라
개인화된 추천 결과 제공

02-1.
USE CASE

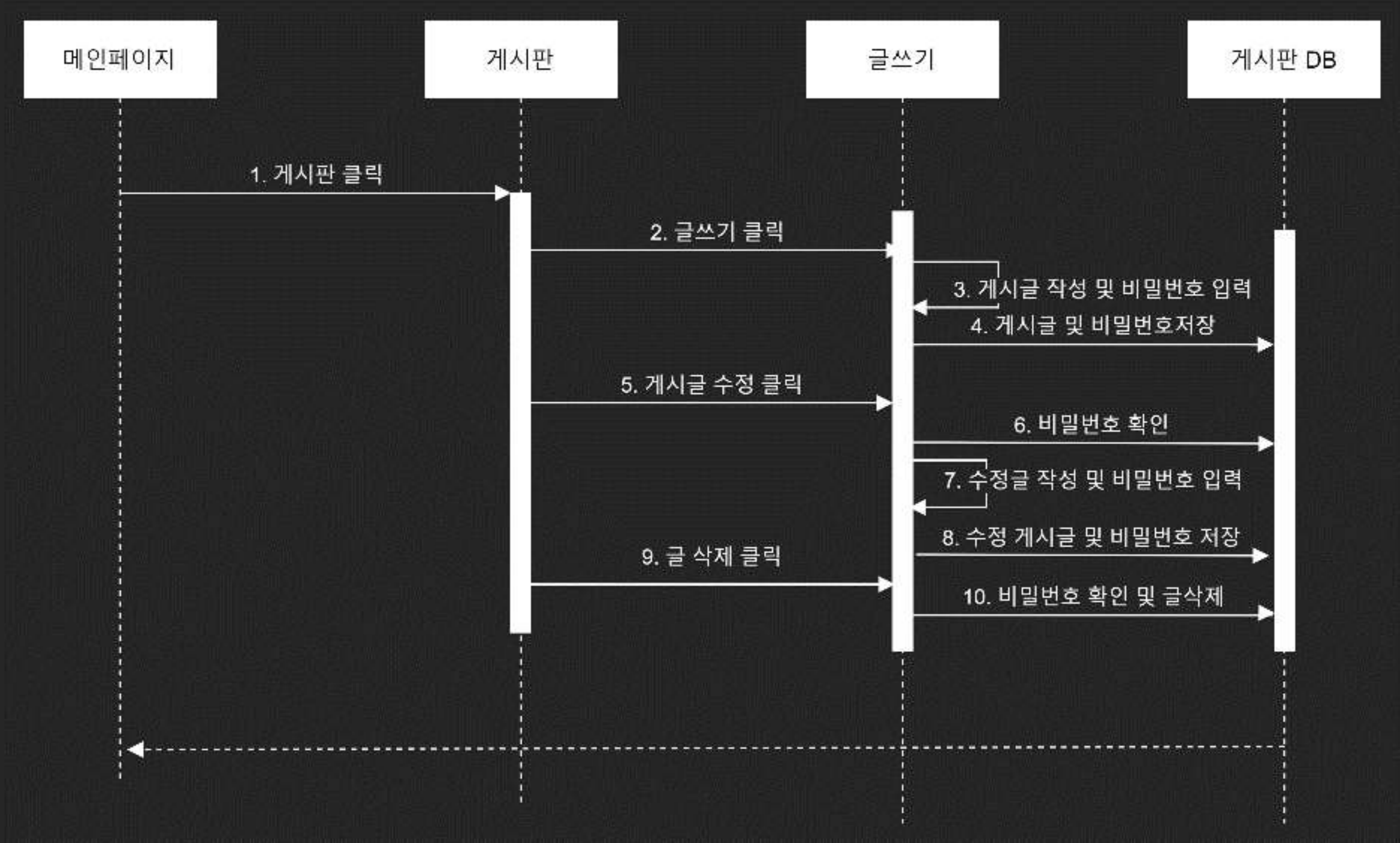

소비자
로그인
로그아웃
<<include>>
<<include>>
회원가입
회원정보수집
<<include>>
<<include>>
수량/품목 변경
<<extend>>
장바구니
<<include>>
장바구니에 넣기
<<include>>
주문 및 결제
<<include>>
주문내역
<<include>>
상품 보기
상품 상세
<<include>>
게시판
글쓰기
<<include>>
<<include>>
글수정 및 삭제
<<include>>
주문 처리
<<include>>
주문관리
사용자 등록
<<include>>
<<include>>
<<include>>
사용자 삭제
<<include>>
사용자관리
관리자
상품등록
<<include>>
상품관리
<<include>>
상품 삭제
<<include>>
글 삭제
게시판 관리

02-2.

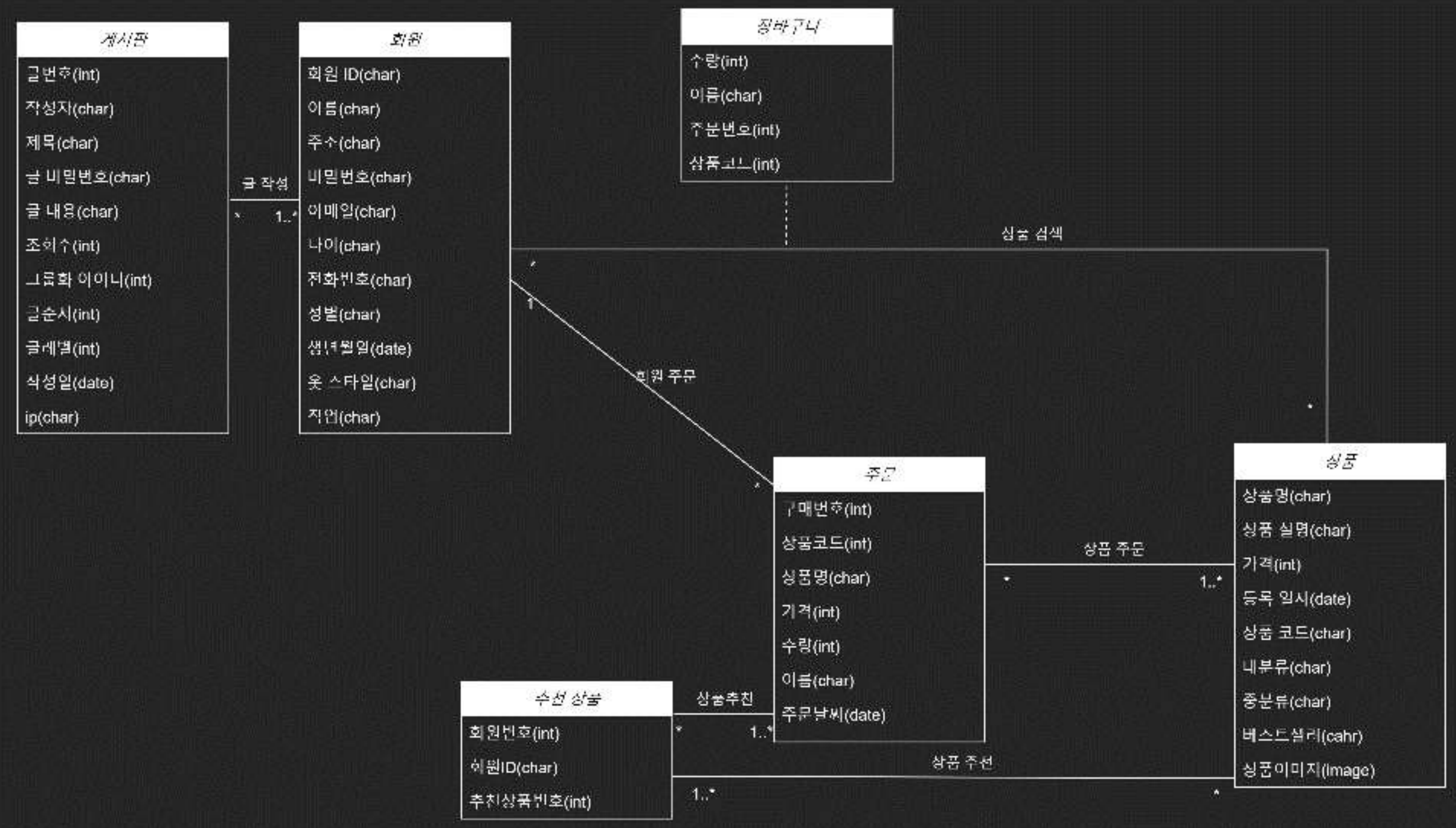
SEQUENCE DIAGRAM (Process)



02-3.
SEQUENCE DIAGRAM (Board)

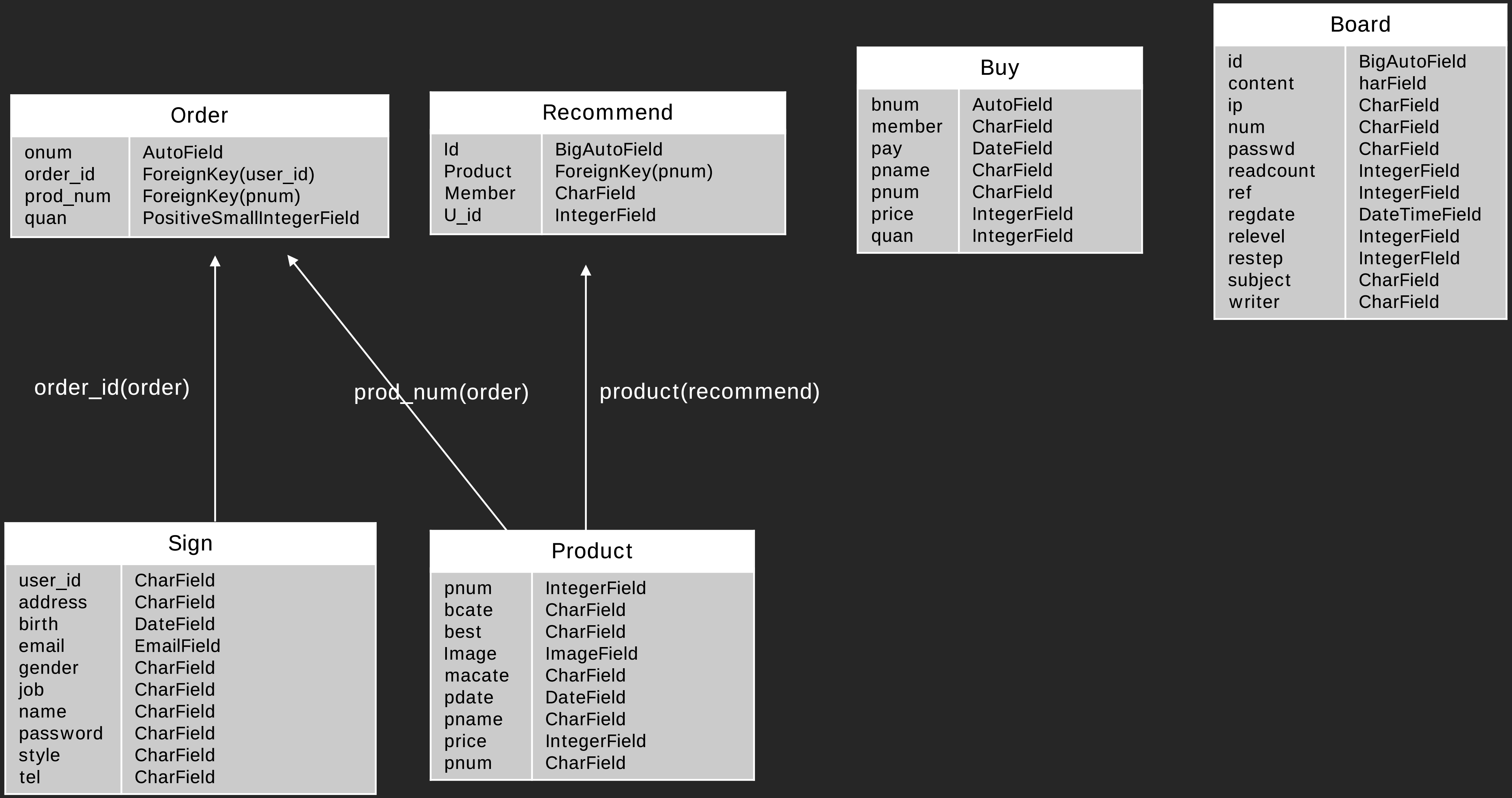

메인페이지
게시판
글쓰기
게시판 DB
1. 게시판 클릭
2. 글쓰기 클릭
3. 게시글 작성 및 비밀번호 입력
4. 게시글 및 비밀번호저장
5. 게시글 수정 클릭
6. 비밀번호 확인
7. 수정글 작성 및 비밀번호 입력
8. 수정 게시글 및 비밀번호 저장
9. 글 삭제 클릭
10. 비밀번호 확인 및 글삭제

02-4.
CLASS DIAGRAM



03.

ERD



Order

onum	AutoField
order_id	ForeignKey(user_id)
prod_num	ForeignKey(pnum)
quan	PositiveSmallIntegerField

Recommend

Id	BigAutoField
Product	ForeignKey(pnum)
Member	CharField
U_id	IntegerField

Buy

bnum	AutoField
member	CharField
pay	DateField
pname	CharField
pnum	CharField
price	IntegerField
quan	IntegerField

Board

id	BigAutoField
content	harField
ip	CharField
num	CharField
passwd	CharField
readcount	IntegerField
ref	IntegerField
regdate	DateTimeField
relevel	IntegerField
restep	IntegerFIeld
subject	CharField
writer	CharField

order_id(order)

prod_num(order)

product(recommend)

Sign

user_id	CharField
address	CharField
birth	DateField
email	EmailField
gender	CharField
job	CharField
name	CharField
password	CharField
style	CharField
tel	CharField

Product

pnum	IntegerField
bcate	CharField
best	CharField
Image	ImageField
macate	CharField
pdate	DateField
pname	CharField
price	IntegerField
pnum	CharField

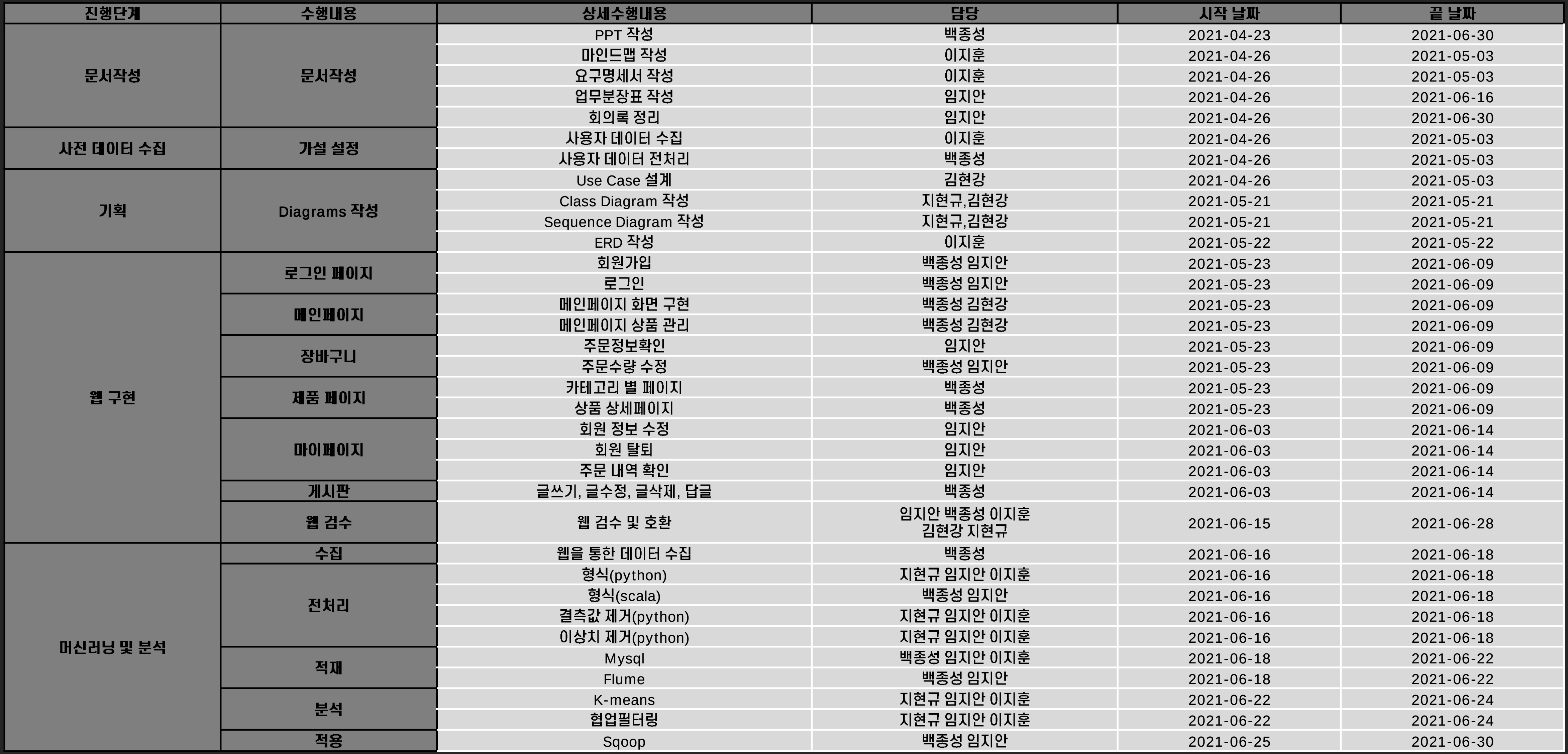
04.

WORK FLOW

수집	전처리	적재	분석	적용
Django Server Django 3 HTML CSS JavaScript Jquery Ajax	Linux Master Hadoop NameNode DataNode Spark		Linux Worker Hadoop DataNode	
	Python Scala	Flume MySQL SQOOP	Pyspark Crond	SQOOP MySQL

05.

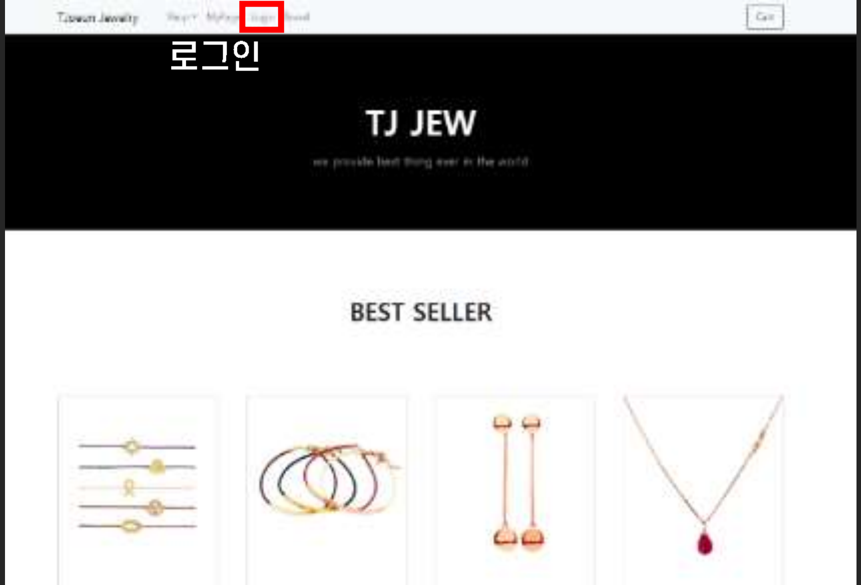
SCHEDULE & WORK DIVISION



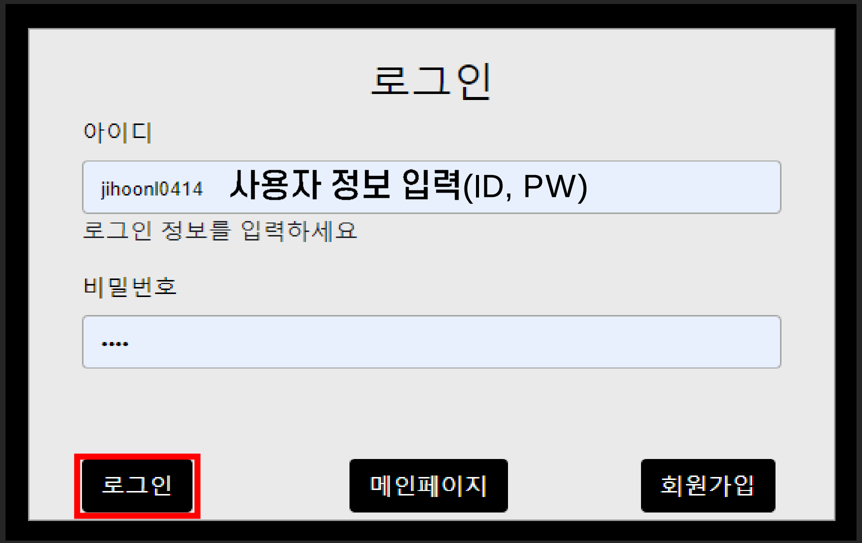
진행단계	수행내용	상세수행내용	담당	시작 날짜	끝 날짜
문서작성	문서작성	PPT 작성	백종성	2021-04-23	2021-06-30
		마인드맵 작성	이지훈	2021-04-26	2021-05-03
		요구명세서 작성	이지훈	2021-04-26	2021-05-03
		업무분장표 작성	임지안	2021-04-26	2021-06-16
		회의록 정리	임지안	2021-04-26	2021-06-30
사전 데이터 수집	가설 설정	사용자 데이터 수집	이지훈	2021-04-26	2021-05-03
		사용자 데이터 전처리	백종성	2021-04-26	2021-05-03
기획	Diagrams 작성	Use Case 설계	김현강	2021-04-26	2021-05-03
		Class Diagram 작성	지현규,김현강	2021-05-21	2021-05-21
		Sequence Diagram 작성	지현규,김현강	2021-05-21	2021-05-21
		ERD 작성	이지훈	2021-05-22	2021-05-22
웹 구현	로그인 페이지	회원가입	백종성 임지안	2021-05-23	2021-06-09
		로그인	백종성 임지안	2021-05-23	2021-06-09
	메인페이지	메인페이지 화면 구현	백종성 김현강	2021-05-23	2021-06-09
		메인페이지 상품 관리	백종성 김현강	2021-05-23	2021-06-09
	장바구니	주문정보확인	임지안	2021-05-23	2021-06-09
		주문수량 수정	백종성 임지안	2021-05-23	2021-06-09
	제품 페이지	카테고리 별 페이지	백종성	2021-05-23	2021-06-09
		상품 상세페이지	백종성	2021-05-23	2021-06-09
	마이페이지	회원 정보 수정	임지안	2021-06-03	2021-06-14
		회원 탈퇴	임지안	2021-06-03	2021-06-14
		주문 내역 확인	임지안	2021-06-03	2021-06-14
	게시판	글쓰기, 글수정, 글삭제, 답글	백종성	2021-06-03	2021-06-14
	웹 검수	웹 검수 및 호환	임지안 백종성 이지훈 김현강 지현규	2021-06-15	2021-06-28
머신러닝 및 분석	수집	웹을 통한 데이터 수집	백종성	2021-06-16	2021-06-18
	전처리	형식(python)	지현규 임지안 이지훈	2021-06-16	2021-06-18
		형식(scala)	백종성 임지안	2021-06-16	2021-06-18
		결측값 제거(python)	지현규 임지안 이지훈	2021-06-16	2021-06-18
		이상치 제거(python)	지현규 임지안 이지훈	2021-06-16	2021-06-18
	적재	Mysql	백종성 임지안 이지훈	2021-06-18	2021-06-22
		Flume	백종성 임지안	2021-06-18	2021-06-22
	분석	K-means	지현규 임지안 이지훈	2021-06-22	2021-06-24
		협업필터링	지현규 임지안 이지훈	2021-06-22	2021-06-24
	적용	Sqoop	백종성 임지안	2021-06-25	2021-06-30

06.

WEB SKILLS



로그인



로그인

아이디

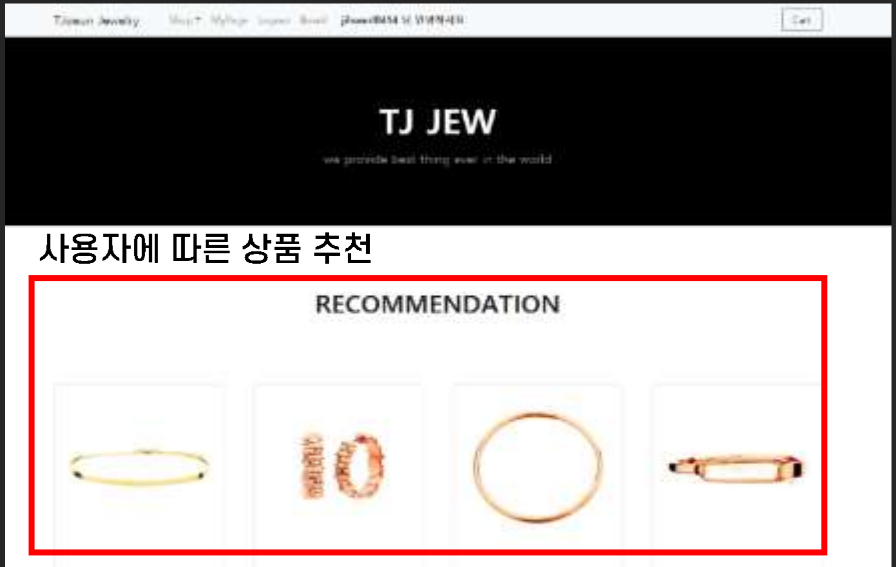
jihoonl0414

로그인 정보를 입력하세요

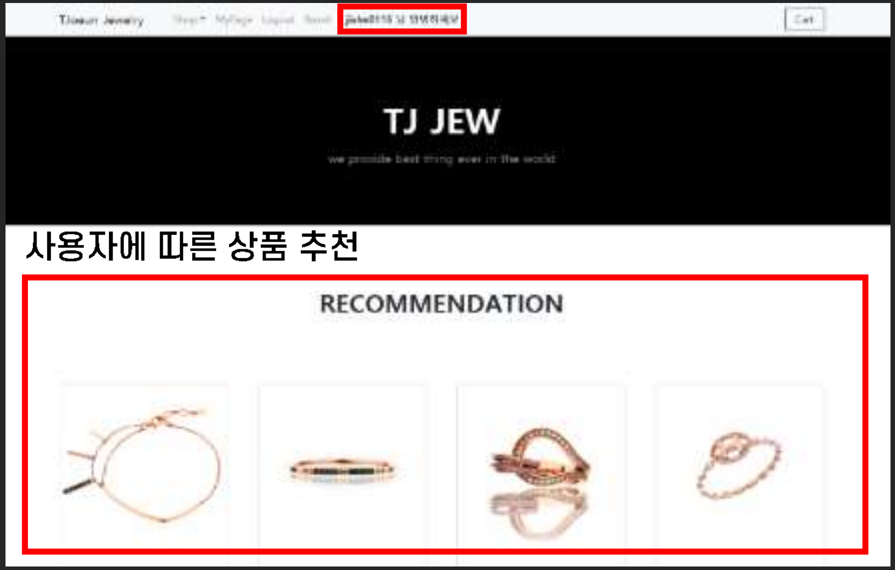
비밀번호

....

사용자 정보 입력(ID, PW)

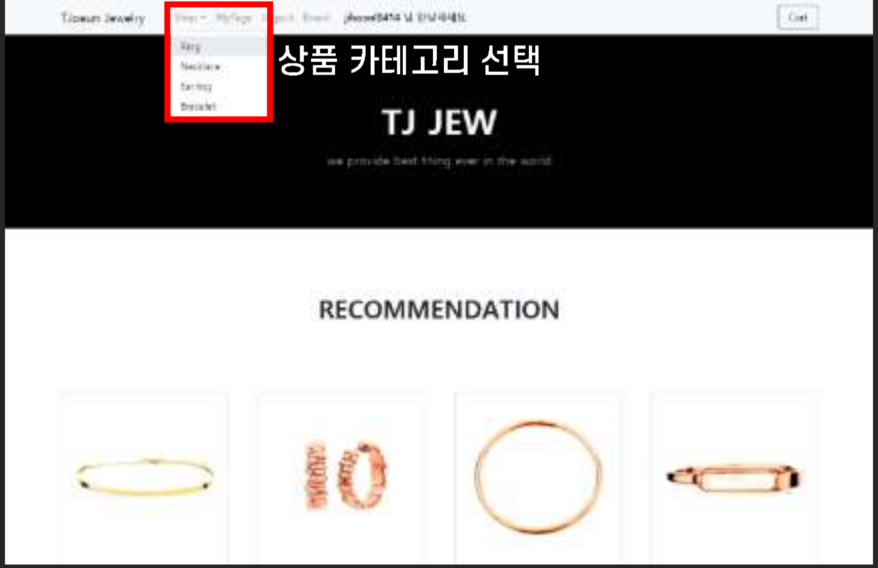


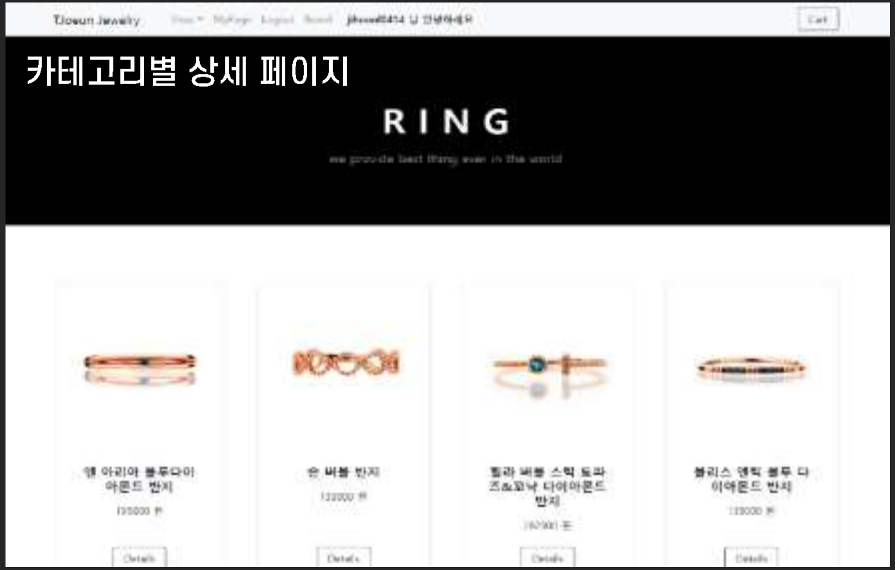
사용자에 따른 상품 추천



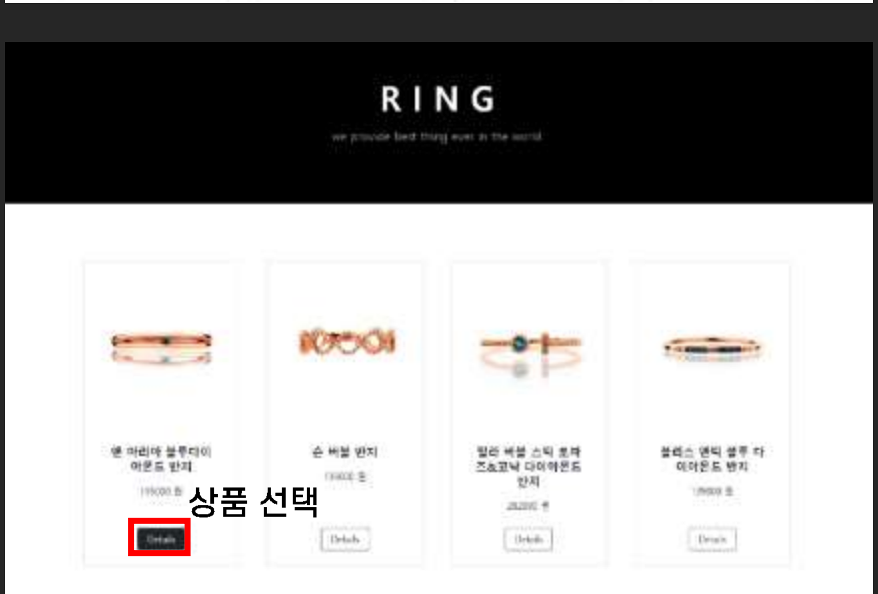
사용자에 따른 상품 추천

06.
WEB SKILLS

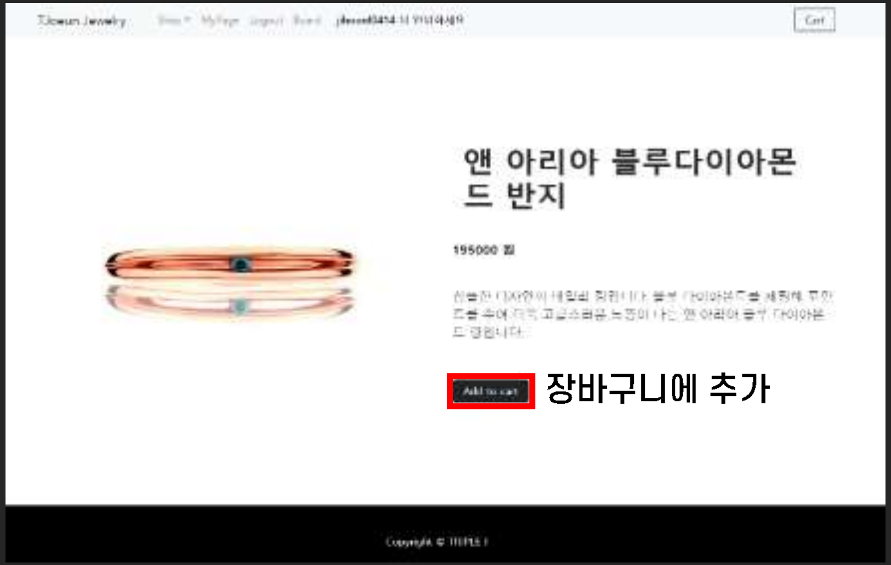

상품 카테고리 선택


카테고리별 상세 페이지

TJoeun Jewelry
Shop
MyPage
Logout
Board
jhseo0414 님 안녕하세요
Cart
RING
we provide best thing ever in the world
앤 아리아 블루다이아몬드 반지
195000 원
순 버블 반지
122000 원
릴라 버블 스틱 토파즈&코냑 다이아몬드 반지
162900 원
블리스 엔틱 블루 다이아몬드 반지
139000 원
Details


상품 선택

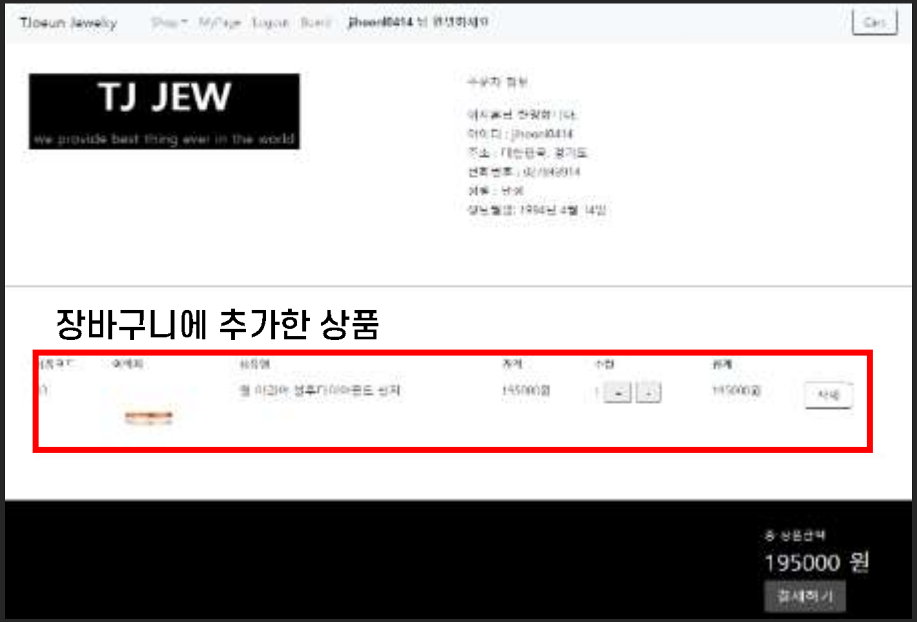
RING
we provide best thing ever in the world
앤 아리아 블루다이아몬드 반지
195000 원
순 버블 반지
122000 원
릴라 버블 스틱 토파즈&코냑 다이아몬드 반지
162900 원
블리스 엔틱 블루 다이아몬드 반지
139000 원
Details


장바구니에 추가

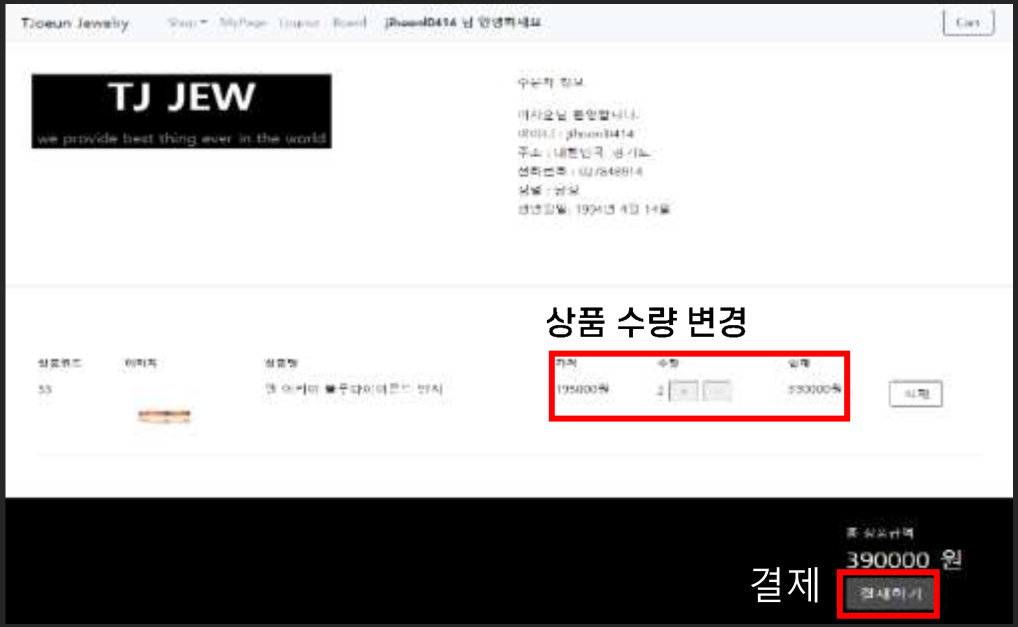
앤 아리아 블루다이아몬드 반지

195000 원

06.
WEB SKILLS

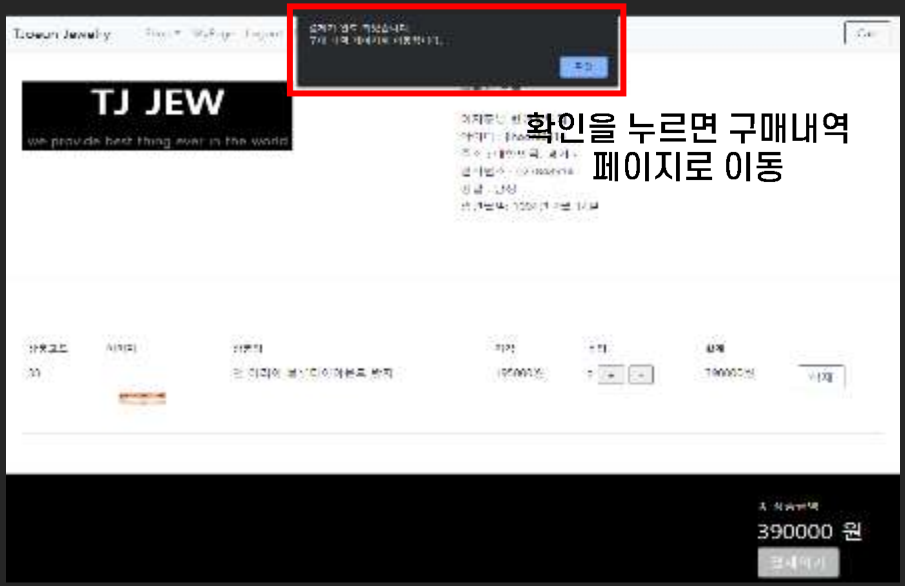


장바구니에 추가한 상품

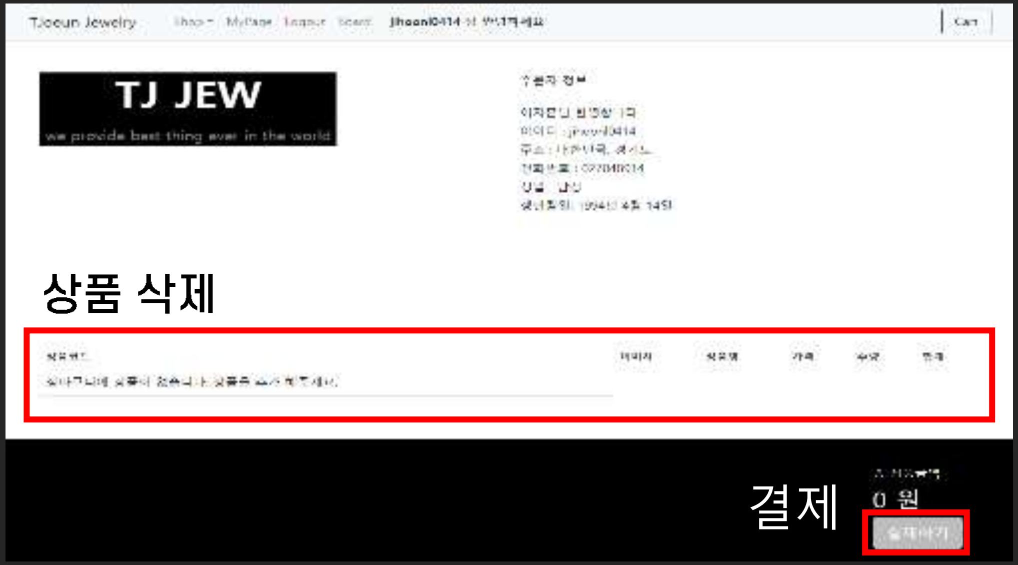


상품 수량 변경

결제



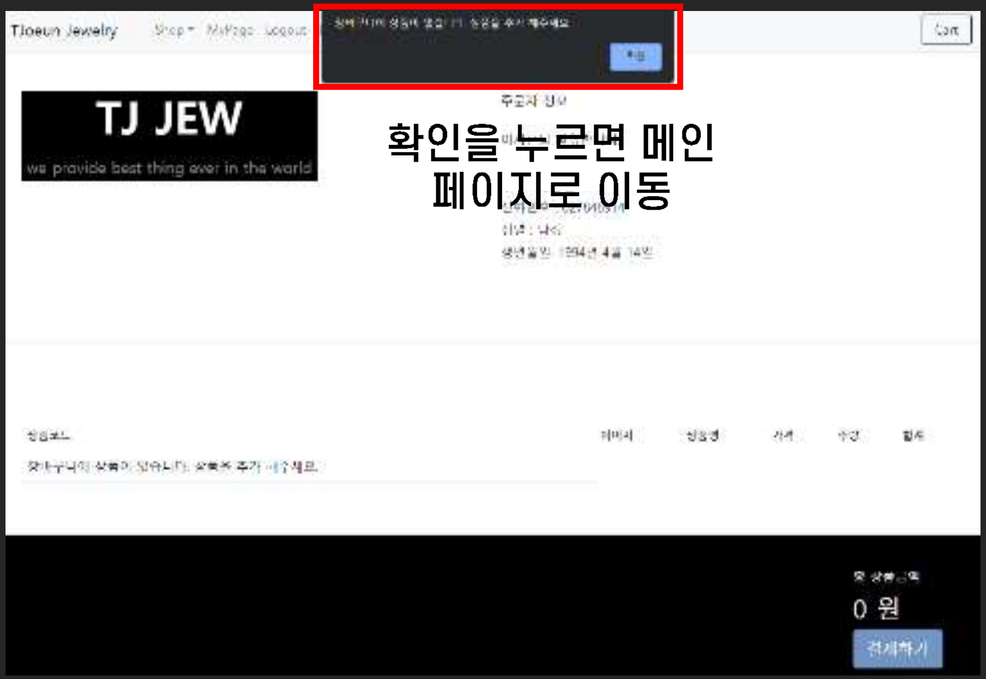
확인을 누르면 구매내역 페이지로 이동



상품 삭제

결제

06.
WEB SKILLS



Tjoeun Jewelry Shop MyPage Logout Cart

[illegible]

확인

TJ JEW

we provide best thing ever in the world

주문자 정보

성별 : 남자

생년월일 : 1994년 4월 14일

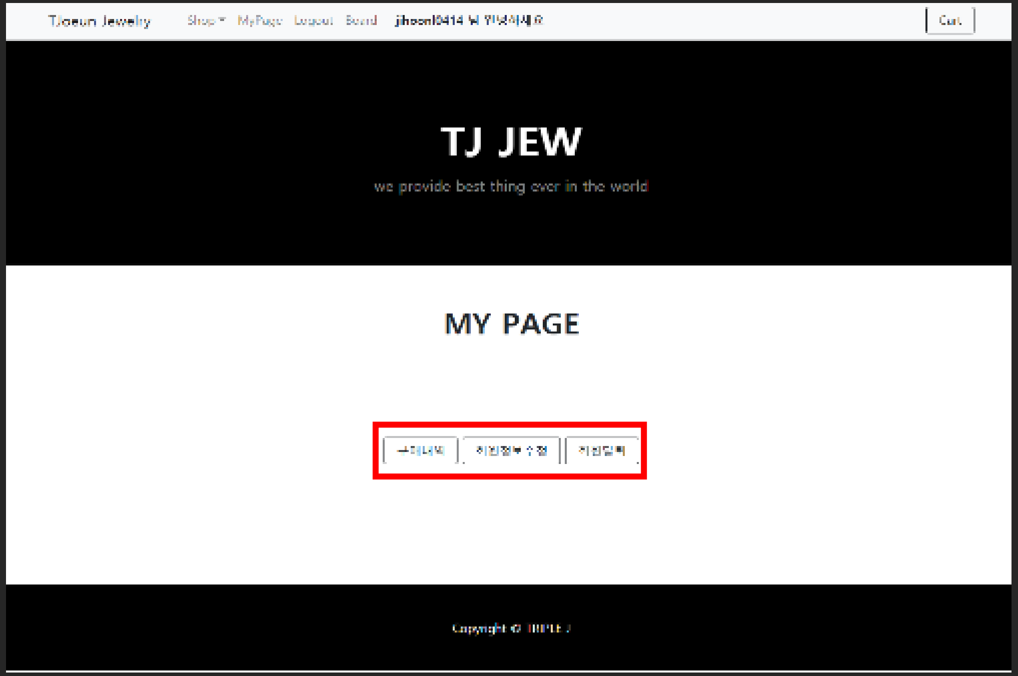
확인을 누르면 메인 페이지로 이동

상품코드	이미지	상품명	가격	수량	합계
장바구니에 상품이 없습니다. 상품을 추가해주세요.					

총 상품금액

0 원

결제하기



Tjoeun Jewelry Shop MyPage Logout Board jihoon0414 님 안녕하세요 Cart

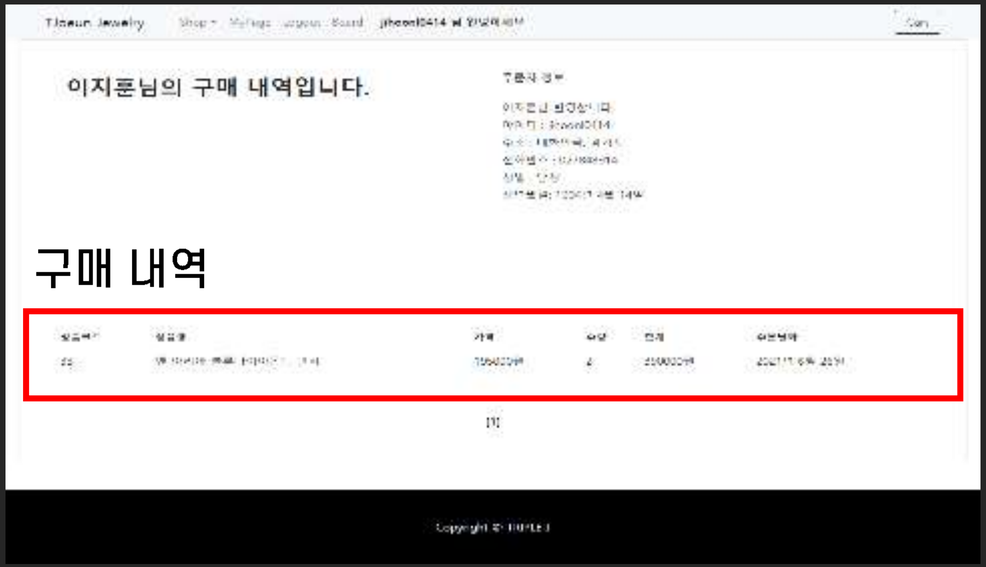
TJ JEW

we provide best thing ever in the world

MY PAGE

구매내역 회원정보수정 회원탈퇴

Copyright © TRIPLE J



Tjoeun Jewelry Shop MyPage Logout Board jihoon0414 님 안녕하세요 Cart

이지훈님의 구매 내역입니다.

주문자 정보

[illegible]

구매 내역

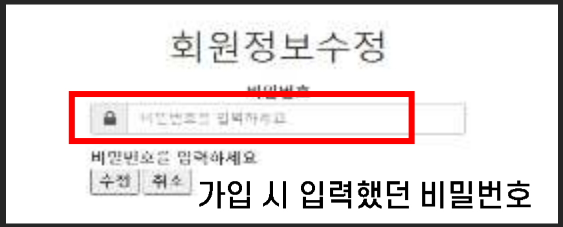
[illegible]

[1]

Copyright © TRIPLE J

[illegible]

기존에 있는 회원 정보
주소 : 경기도
옷 스타일 : 스트릿
직업 : 학생



회원정보수정

비밀번호

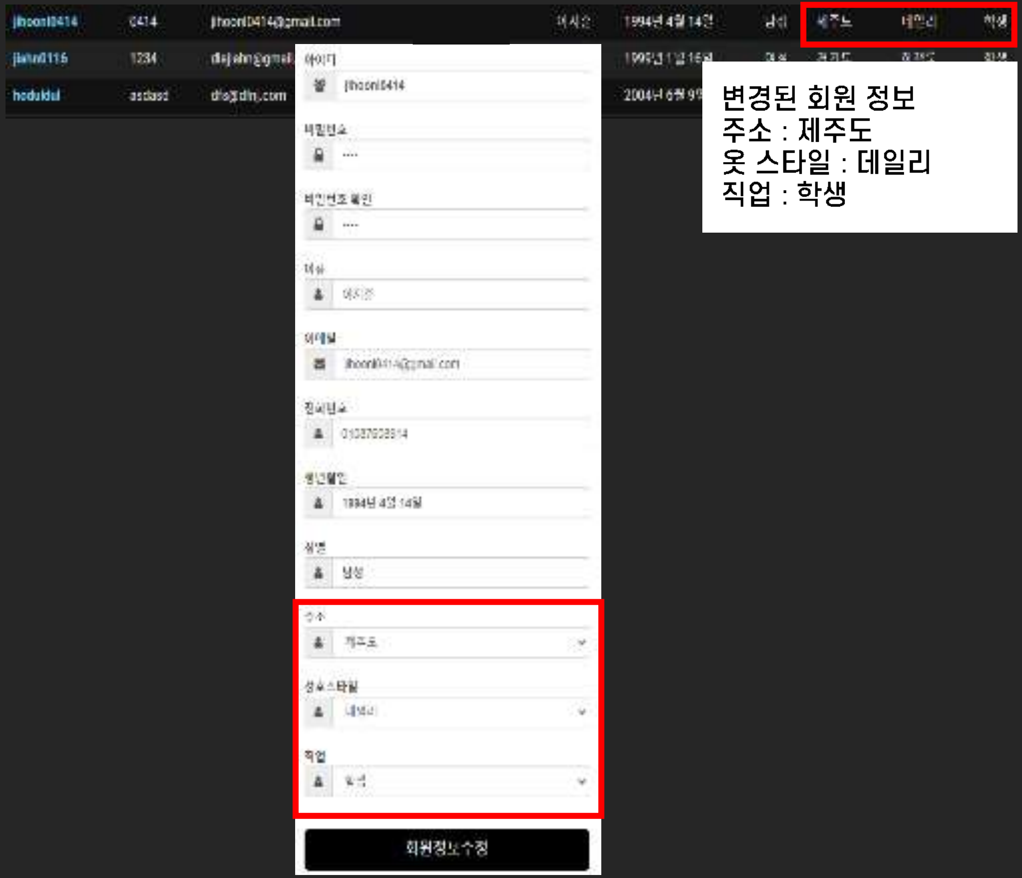
비밀번호를 입력하세요

비밀번호를 입력하세요

수정 취소

가입 시 입력했던 비밀번호

06.
WEB SKILLS



jhoon0414	0414	jhoon0414@gmail.com	이지훈	1994년 4월 14일	남성	제주도	데일리	학생
jiahn0115	1234	diejahn@gmail		1999년 1월 16일				
hodukdul	asdasd	dhs@dhn.com		2004년 6월 9				

아이디
jhoon0414

비밀번호
····

비밀번호 확인
····

이름
이지훈

이메일
jhoon0414@gmail.com

전화번호
01037028814

생년월일
1994년 4월 14일

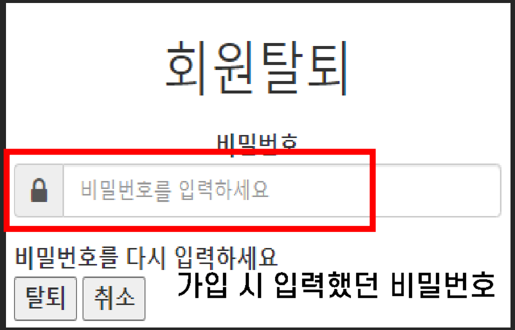
성별
남성

주소
제주도

선호스타일
데일리

직업
학생

변경된 회원 정보
주소 : 제주도
옷 스타일 : 데일리
직업 : 학생

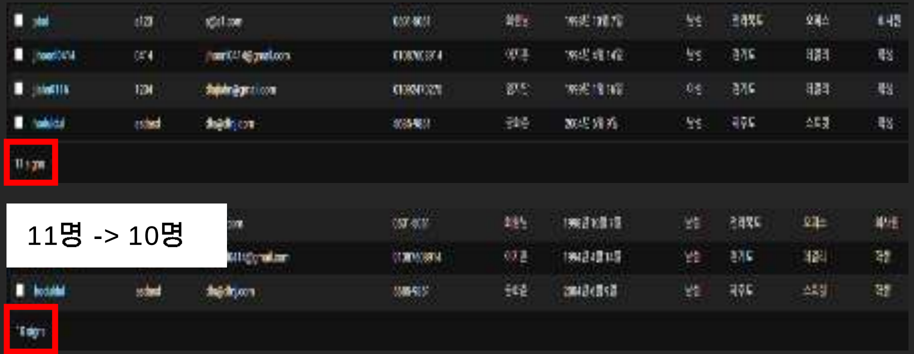


회원탈퇴

비밀번호
비밀번호를 입력하세요

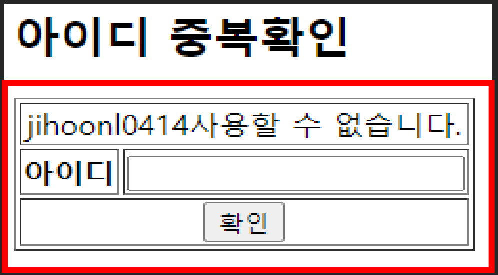
비밀번호를 다시 입력하세요

가입 시 입력했던 비밀번호



11명 -> 10명

06.
WEB SKILLS



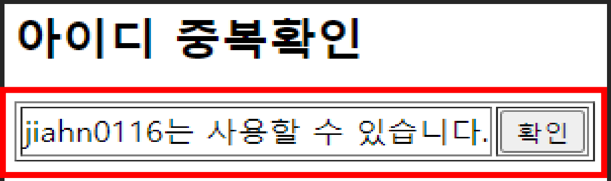
아이디 중복확인

jihoonl0414사용할 수 없습니다.

아이디

확인

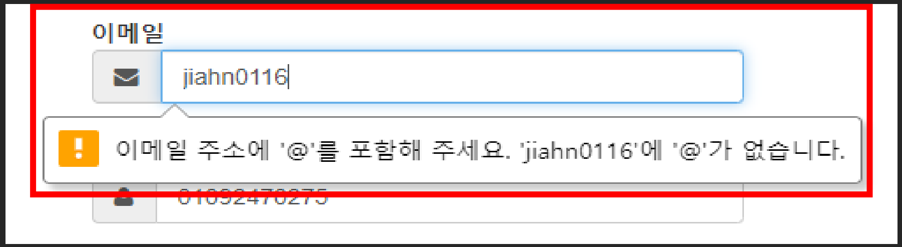
중복확인



아이디 중복확인

jiahn0116는 사용할 수 있습니다. 확인

사용할 ID 입력



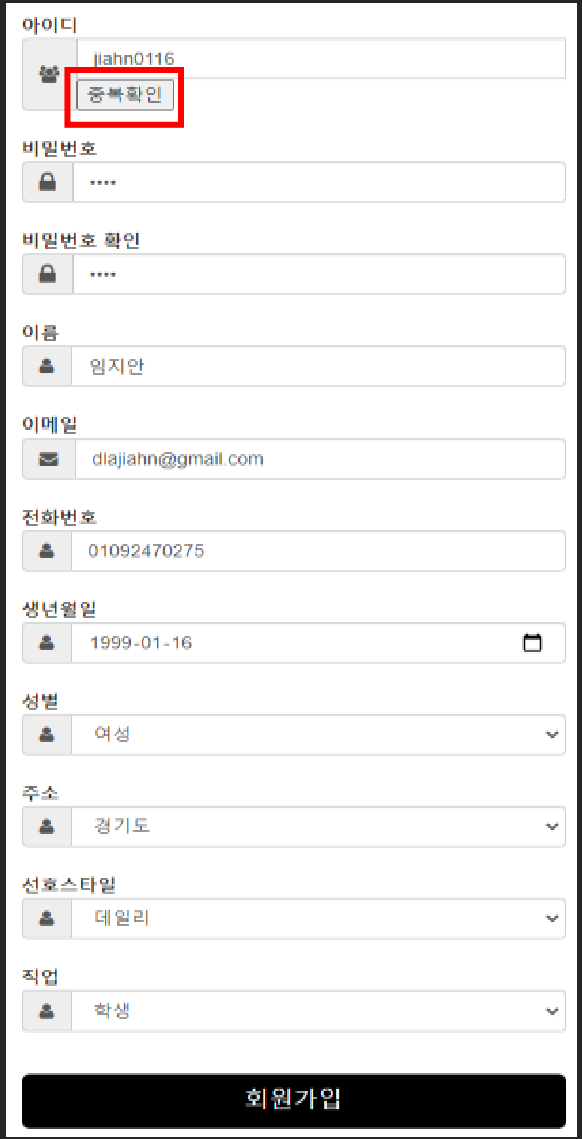
이메일

jiahn0116

! 이메일 주소에 '@'를 포함해 주세요. 'jiahn0116'에 '@'가 없습니다.

01092470275

이메일 형식 체크



아이디

jiahn0116

중복확인

비밀번호

····

비밀번호 확인

····

이름

임지안

이메일

dlajiahn@gmail.com

전화번호

01092470275

생년월일

1999-01-16

성별

여성

주소

경기도

선호스타일

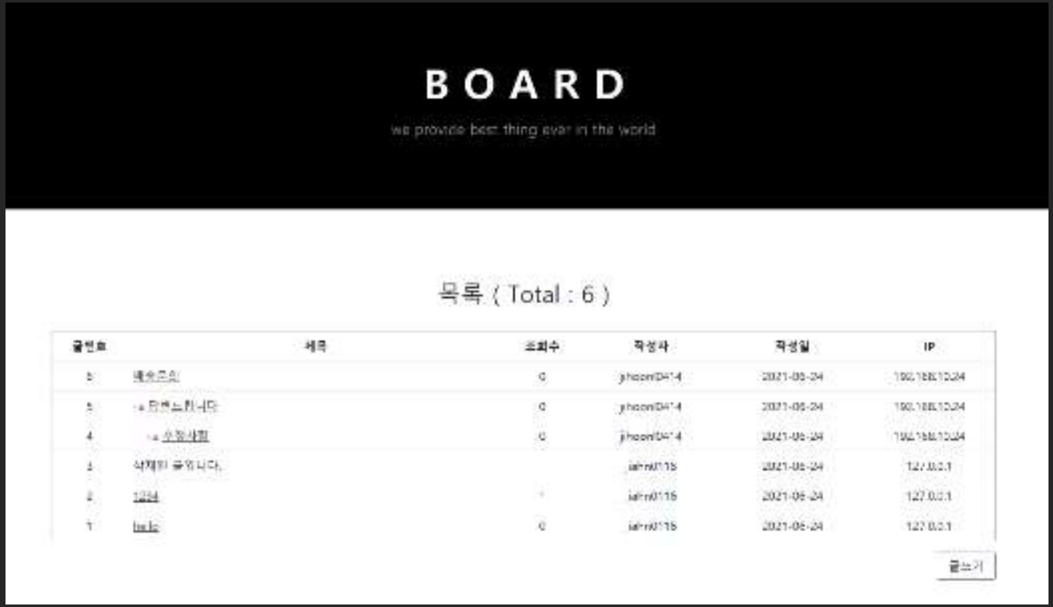
데일리

직업

학생

회원가입

06.
WEB SKILLS

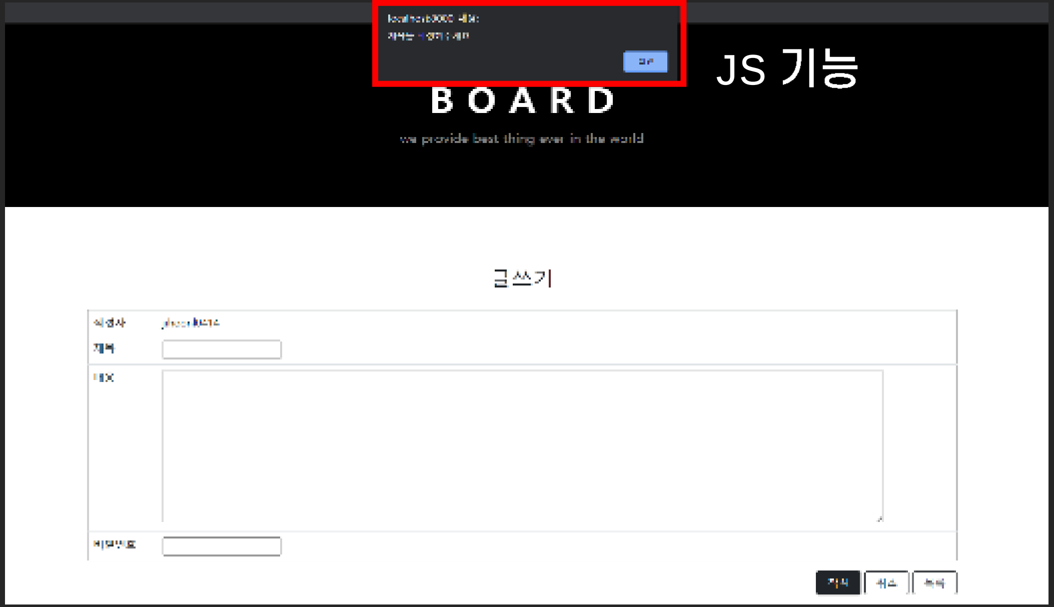


BOARD

we provide best thing ever in the world

목록 (Total : 6)

글번호	제목	조회수	작성자	작성일	IP
6	[illegible]	0	jhoon0414	2021-06-24	192.168.10.24
5	ㄴ 답변드립니다	0	jhoon0414	2021-06-24	192.168.10.24
4	ㄴ 수정사항	0	jhoon0414	2021-06-24	192.168.10.24
3	삭제된 글입니다.		jahn0115	2021-06-24	127.0.0.1
2	1234	[illegible]	jahn0115	2021-06-24	127.0.0.1
1	hello	0	jahn0115	2021-06-24	127.0.0.1



localhost:8000 내용:
[illegible]

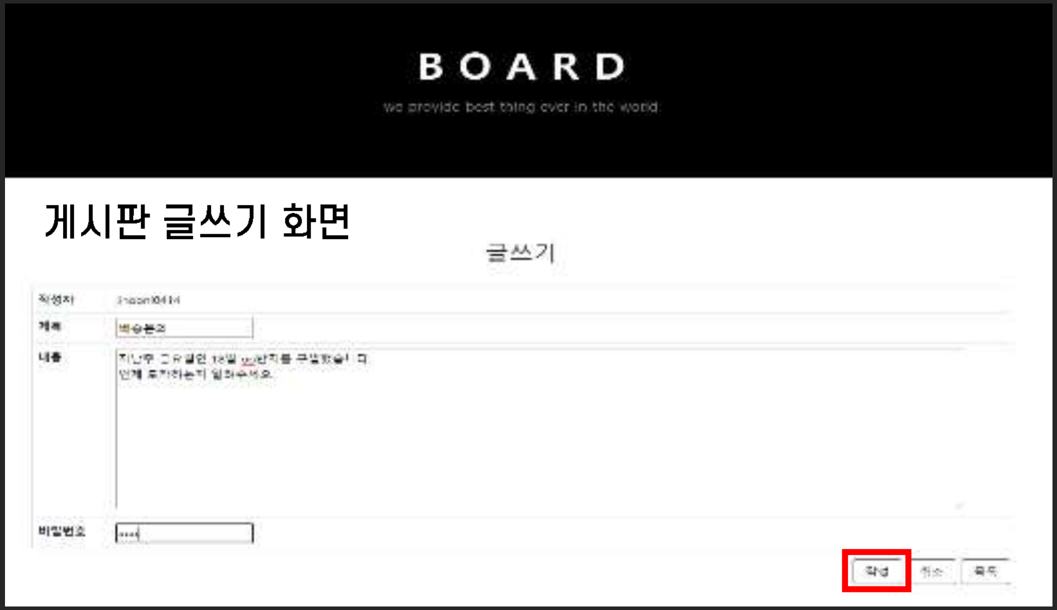
JS 기능

BOARD

we provide best thing ever in the world

글쓰기

작성자	jhoon0414
제목	
내용	
비밀번호	



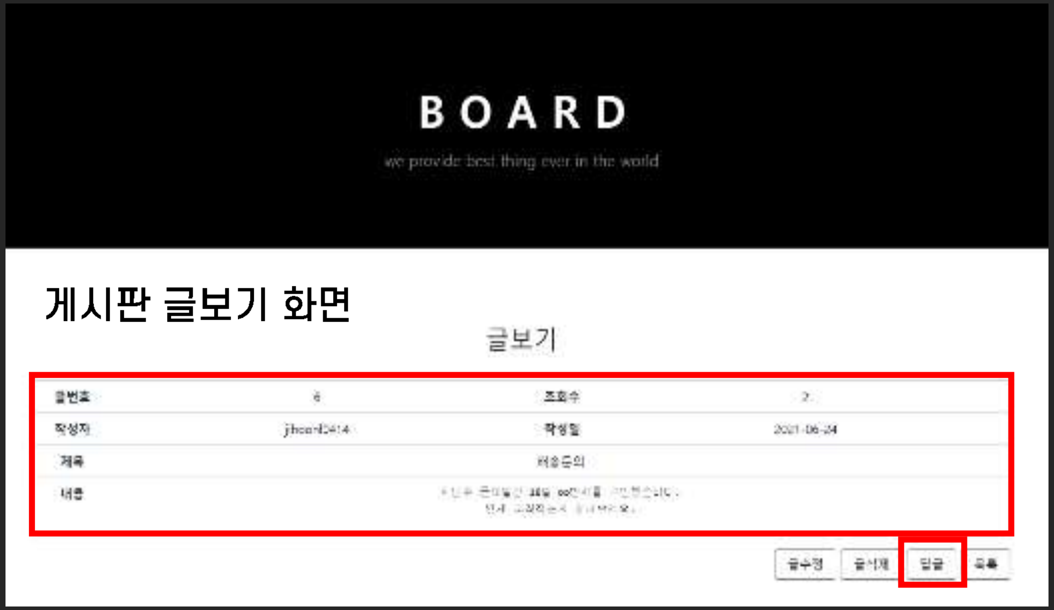
BOARD

we provide best thing ever in the world

게시판 글쓰기 화면

글쓰기

작성자	jhoon0414
제목	[illegible]
내용	[illegible]
비밀번호	****



BOARD

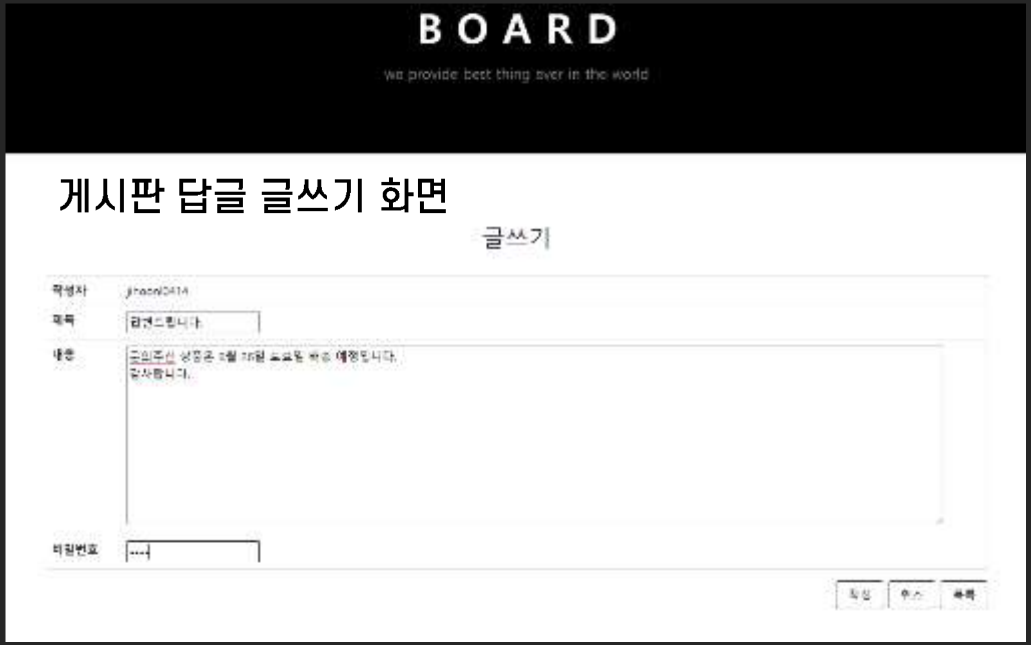
we provide best thing ever in the world

게시판 글보기 화면

글보기

글번호	6	조회수	2
작성자	jhoon0414	작성일	2021-06-24
제목	[illegible]		
내용	[illegible]		

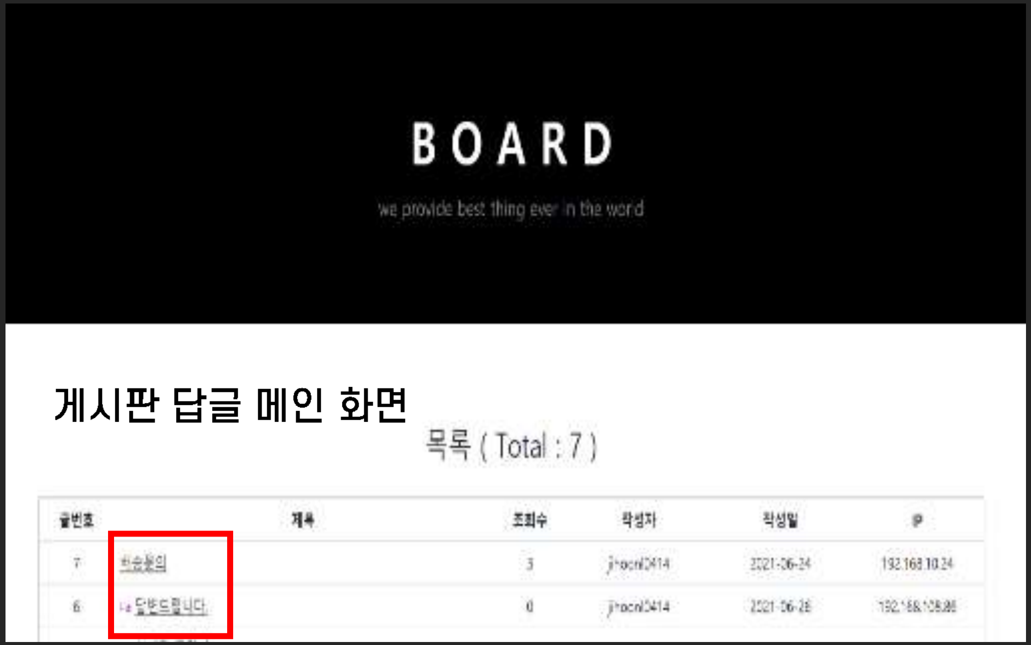
06.
WEB SKILLS



게시판 답글 글쓰기 화면

글쓰기

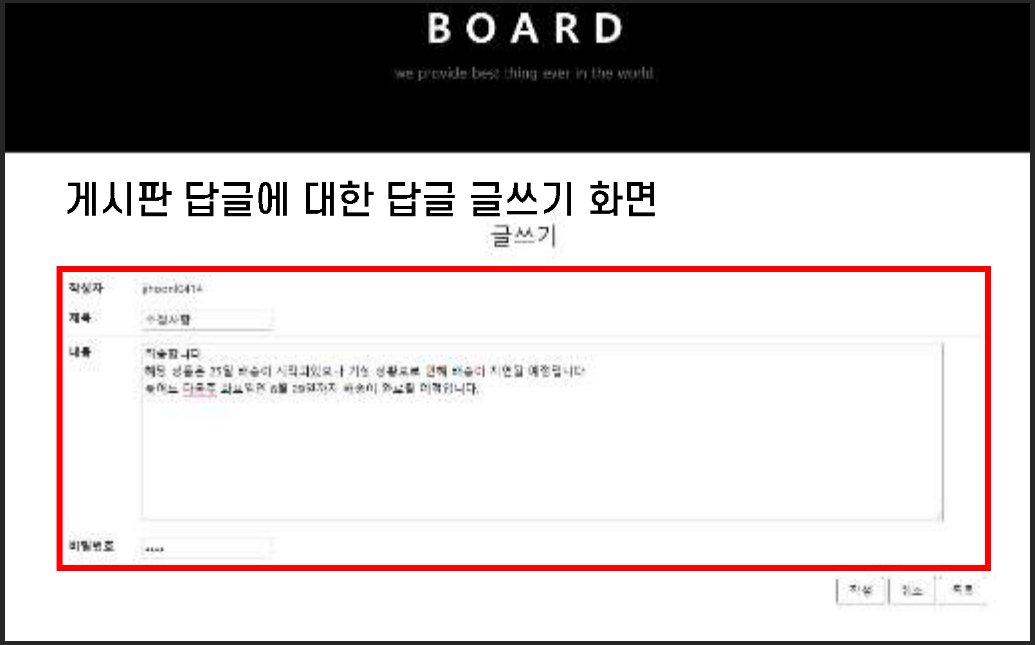
작성자	jihoonl0414
제목	답변드립니다.
내용	문의주신 상품은 6월 28일 오후에 배송 예정입니다. 감사합니다.
비밀번호	····



게시판 답글 메인 화면

목록 (Total : 7)

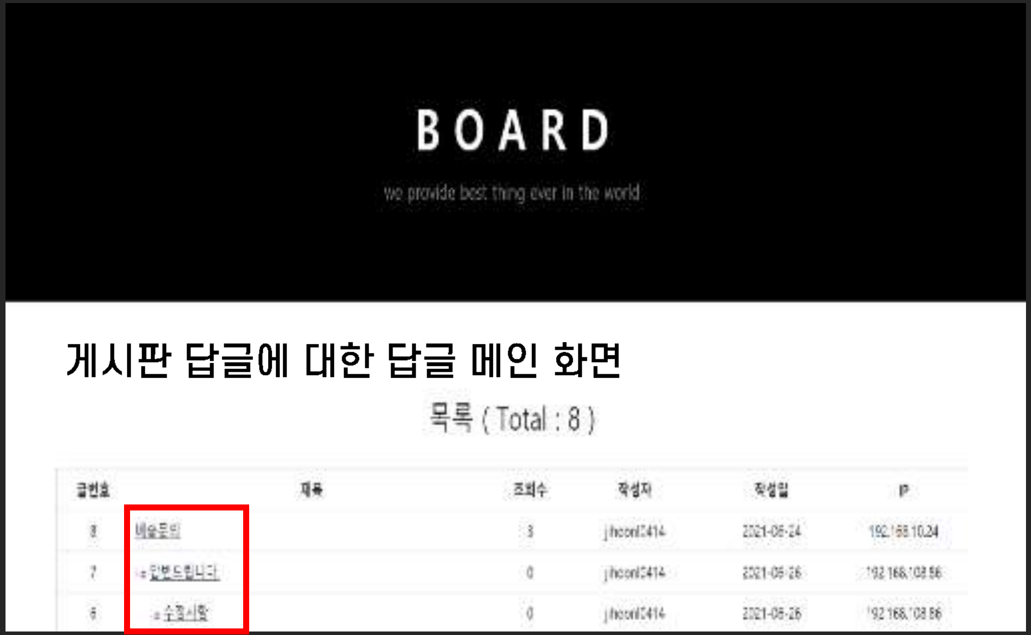
글번호	제목	조회수	작성자	작성일	IP
7	배송문의	3	jihoonl0414	2021-06-24	192.168.10.24
6	↳ 답변드립니다.	0	jihoonl0414	2021-06-28	192.168.108.86



게시판 답글에 대한 답글 글쓰기 화면

글쓰기

작성자	jihoonl0414
제목	수정사항
내용	죄송합니다 해당 상품은 25일 배송이 시작되었으나 기상 상황으로 인해 배송이 지연될 예정입니다 늦어도 다음주 화요일면 6월 29일까지 배송이 완료될 예정입니다.
비밀번호	····

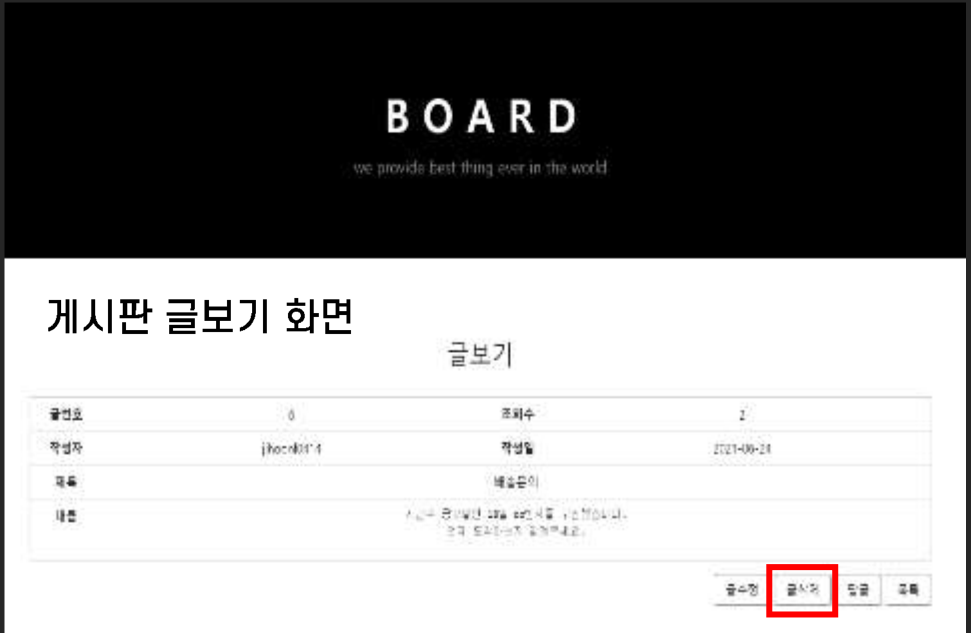


게시판 답글에 대한 답글 메인 화면

목록 (Total : 8)

글번호	제목	조회수	작성자	작성일	IP
8	배송문의	3	jihoonl0414	2021-06-24	192.168.10.24
7	↳ 답변드립니다	0	jihoonl0414	2021-06-26	192.168.108.86
6	↳ 수정사항	0	jihoonl0414	2021-06-26	192.168.108.86

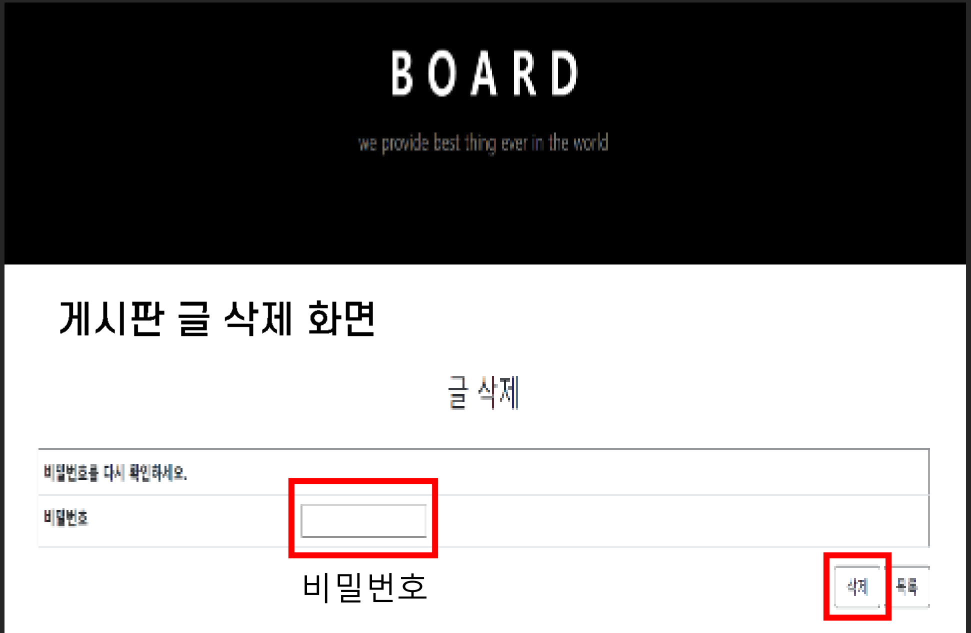
06.
WEB SKILLS



게시판 글보기 화면

글보기

글번호	6	조회수	2
작성자	jihoonl0414	작성일	2021-06-24
제목	배송문의		
내용	[illegible]		



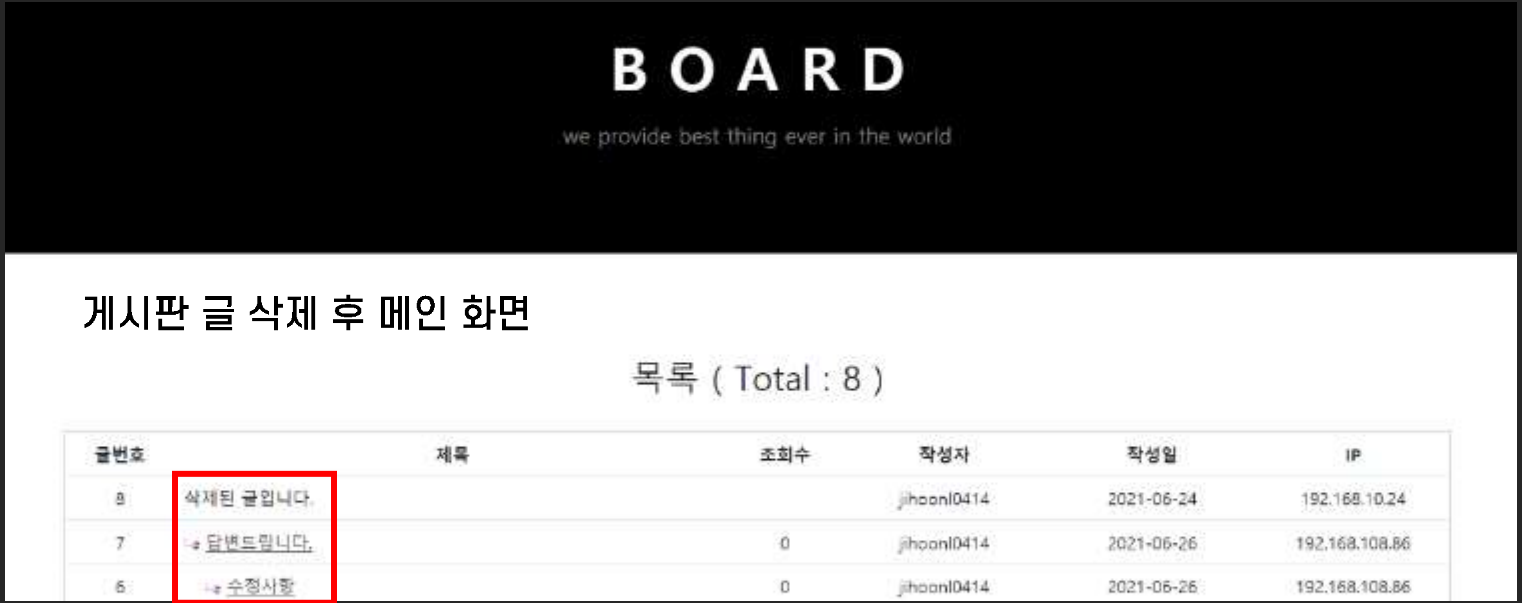
게시판 글 삭제 화면

글 삭제

비밀번호를 다시 확인하세요.

비밀번호

비밀번호

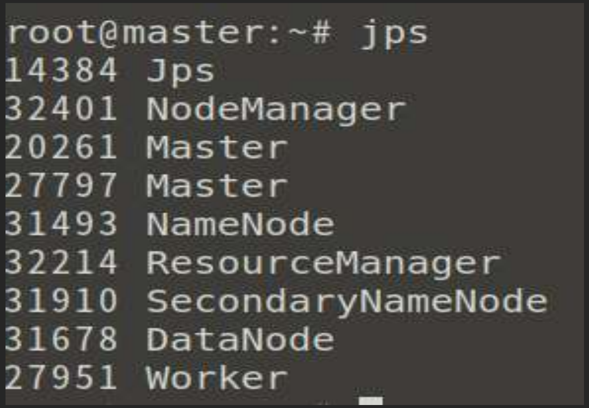


게시판 글 삭제 후 메인 화면

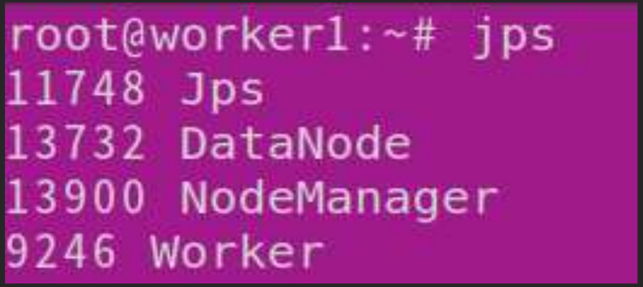
목록 (Total : 8)

글번호	제목	조회수	작성자	작성일	IP
8	삭제된 글입니다.		jihoonl0414	2021-06-24	192.168.10.24
7	↳ 답변드립니다.	0	jihoonl0414	2021-06-26	192.168.108.86
6	↳ 수정사항	0	jihoonl0414	2021-06-26	192.168.108.86

07.
SYSTEM SKILLS



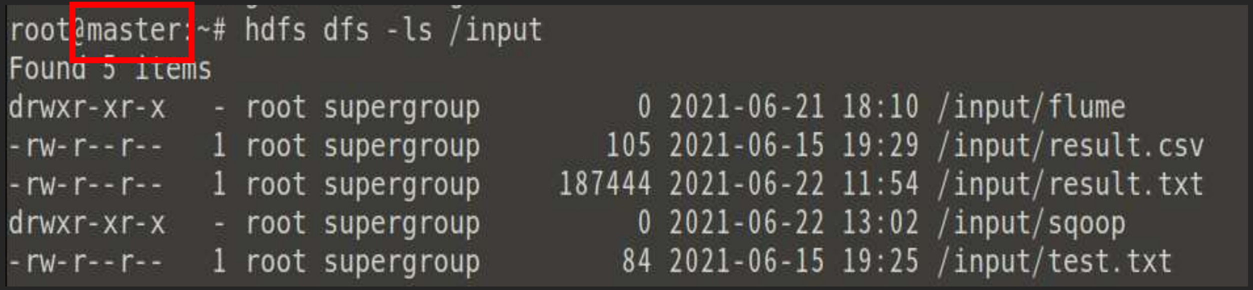
```
root@master:~# jps
14384 Jps
32401 NodeManager
20261 Master
27797 Master
31493 NameNode
32214 ResourceManager
31910 SecondaryNameNode
31678 DataNode
27951 Worker
```



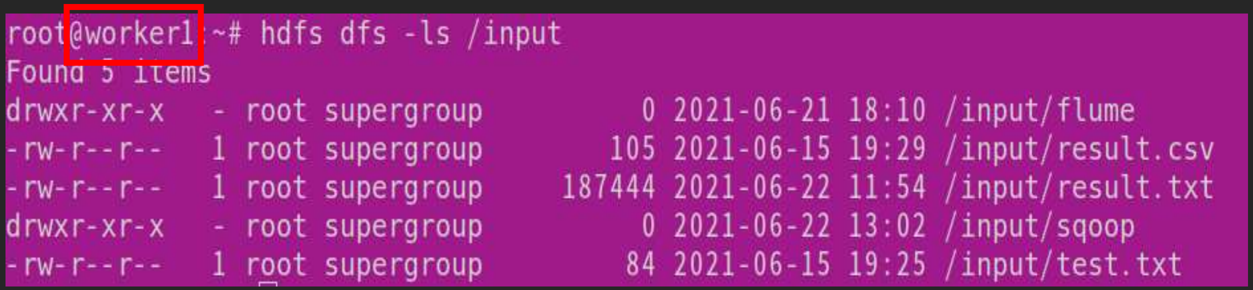
```
root@worker1:~# jps
11748 Jps
13732 DataNode
13900 NodeManager
9246 Worker
```

HDFS

DFS, YARN 실행



```
root@master:~# hdfs dfs -ls /input
Found 5 items
drwxr-xr-x   - root supergroup          0 2021-06-21 18:10 /input/flume
-rw-r--r--   1 root supergroup        105 2021-06-15 19:29 /input/result.csv
-rw-r--r--   1 root supergroup     187444 2021-06-22 11:54 /input/result.txt
drwxr-xr-x   - root supergroup          0 2021-06-22 13:02 /input/sqoop
-rw-r--r--   1 root supergroup         84 2021-06-15 19:25 /input/test.txt
```



```
root@worker1:~# hdfs dfs -ls /input
Found 5 items
drwxr-xr-x   - root supergroup          0 2021-06-21 18:10 /input/flume
-rw-r--r--   1 root supergroup        105 2021-06-15 19:29 /input/result.csv
-rw-r--r--   1 root supergroup     187444 2021-06-22 11:54 /input/result.txt
drwxr-xr-x   - root supergroup          0 2021-06-22 13:02 /input/sqoop
-rw-r--r--   1 root supergroup         84 2021-06-15 19:25 /input/test.txt
```

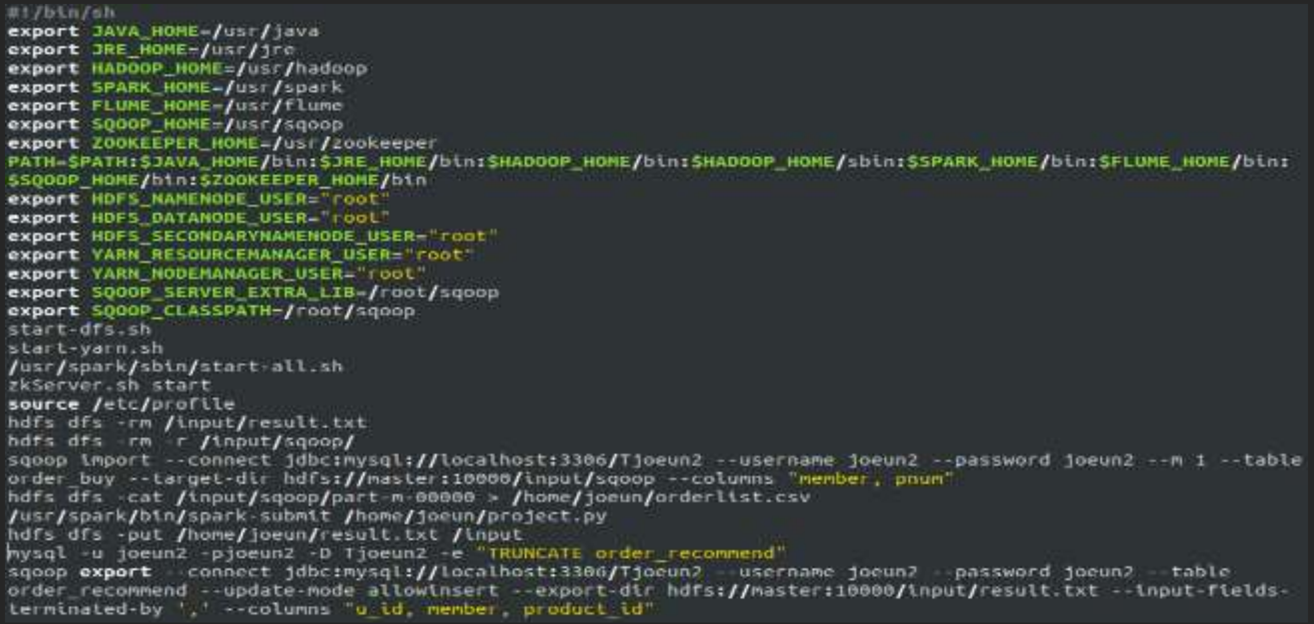
Master, Worker

HDFS

07.

SYSTEM SKILLS

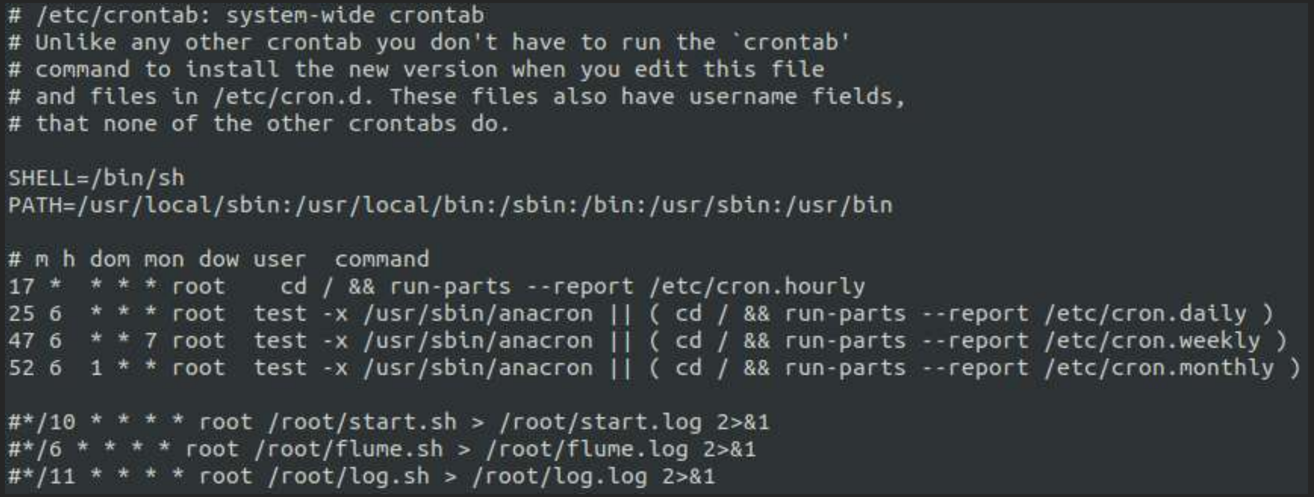
Crontab



```
#!/bin/sh
export JAVA_HOME=/usr/java
export JRE_HOME=/usr/jre
export HADOOP_HOME=/usr/hadoop
export SPARK_HOME=/usr/spark
export FLUME_HOME=/usr/flume
export SQOOP_HOME=/usr/sqoop
export ZOOKEEPER_HOME=/usr/zookeeper
PATH=$PATH:$JAVA_HOME/bin:$JRE_HOME/bin:$HADOOP_HOME/bin:$HADOOP_HOME/sbin:$SPARK_HOME/bin:$FLUME_HOME/bin:
$SQOOP_HOME/bin:$ZOOKEEPER_HOME/bin
export HDFS_NAMENODE_USER="root"
export HDFS_DATANODE_USER="root"
export HDFS_SECONDARYNAMENODE_USER="root"
export YARN_RESOURCEMANAGER_USER="root"
export YARN_NODEMANAGER_USER="root"
export SQOOP_SERVER_EXTRA_LIB=/root/sqoop
export SQOOP_CLASSPATH=/root/sqoop
start-dfs.sh
start-yarn.sh
/usr/spark/sbin/start-all.sh
zkServer.sh start
source /etc/profile
hdfs dfs -rm /input/result.txt
hdfs dfs -rm -r /input/sqoop/
sqoop import --connect jdbc:mysql://localhost:3306/Tjoeun2 --username joeun2 --password joeun2 --m 1 --table
order_buy --target-dir hdfs://master:10000/input/sqoop --columns "member, pnum"
hdfs dfs -cat /input/sqoop/part-m-00000 > /home/joeun/orderlist.csv
/usr/spark/bin/spark-submit /home/joeun/project.py
hdfs dfs -put /home/joeun/result.txt /input
mysql -u joeun2 -pjoeun2 -D Tjoeun2 -e "TRUNCATE order_recommend"
sqoop export --connect jdbc:mysql://localhost:3306/Tjoeun2 --username joeun2 --password joeun2 --table
order_recommend --update-mode allowinsert --export-dir hdfs://master:10000/input/result.txt --input-fields-
terminated-by ',' --columns "u_id, member, product_id"
```

Shell script

실행 프로세스 코딩



```
# /etc/crontab: system-wide crontab
# Unlike any other crontab you don't have to run the `crontab'
# command to install the new version when you edit this file
# and files in /etc/cron.d. These files also have username fields,
# that none of the other crontabs do.

SHELL=/bin/sh
PATH=/usr/local/sbin:/usr/local/bin:/sbin:/bin:/usr/sbin:/usr/bin

# m h dom mon dow user  command
17 *  * * * root    cd / && run-parts --report /etc/cron.hourly
25 6  * * * root  test -x /usr/sbin/anacron || ( cd / && run-parts --report /etc/cron.daily )
47 6  * * 7 root  test -x /usr/sbin/anacron || ( cd / && run-parts --report /etc/cron.weekly )
52 6  1 * * root  test -x /usr/sbin/anacron || ( cd / && run-parts --report /etc/cron.monthly )

#*/10 * * * * root /root/start.sh > /root/start.log 2>&1
#*/6 * * * * root /root/flume.sh > /root/flume.log 2>&1
#*/11 * * * * root /root/log.sh > /root/log.log 2>&1
```

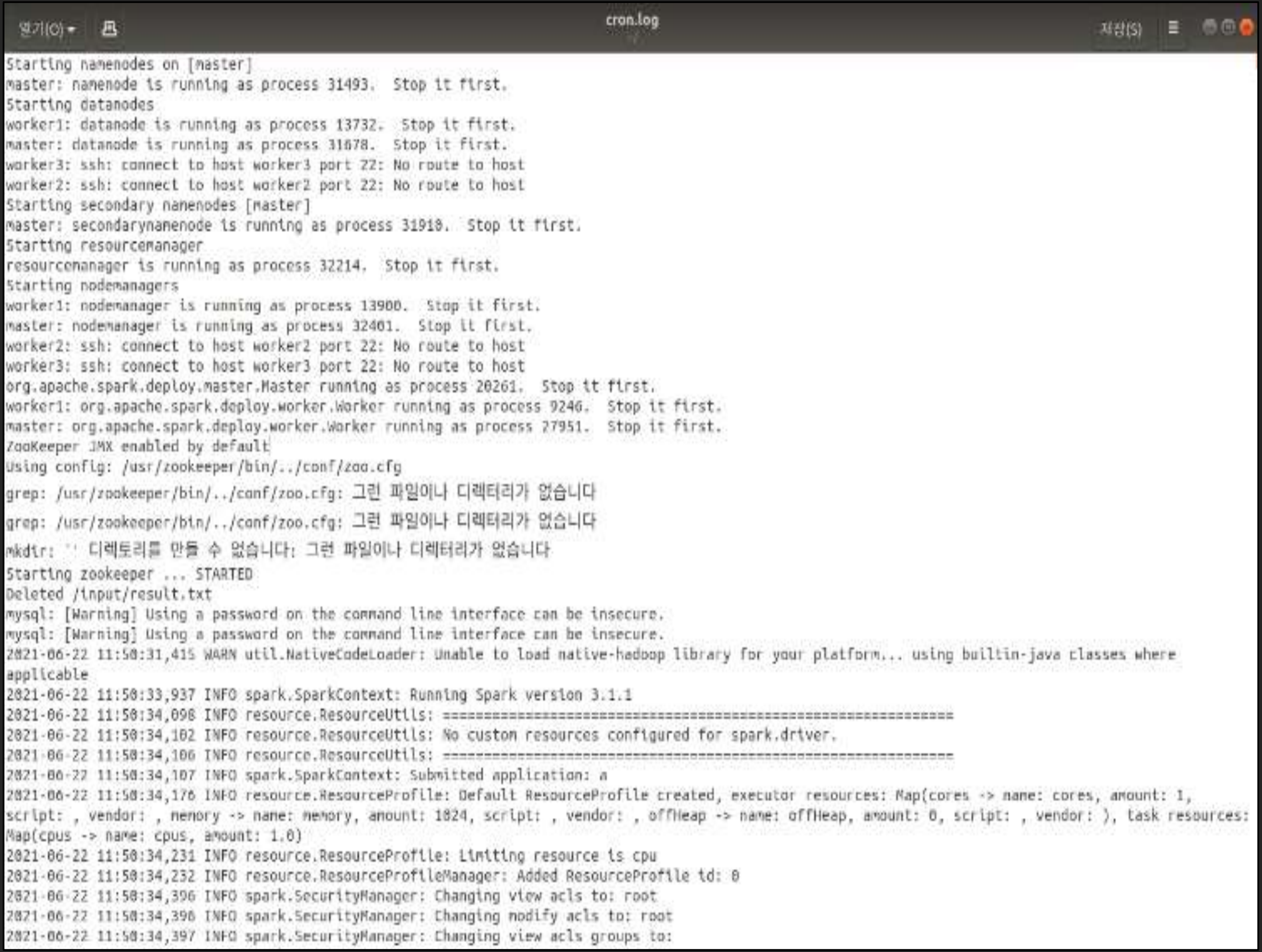
Crontab

.sh 실행시간 설정

07.

SYSTEM SKILLS

Crontab

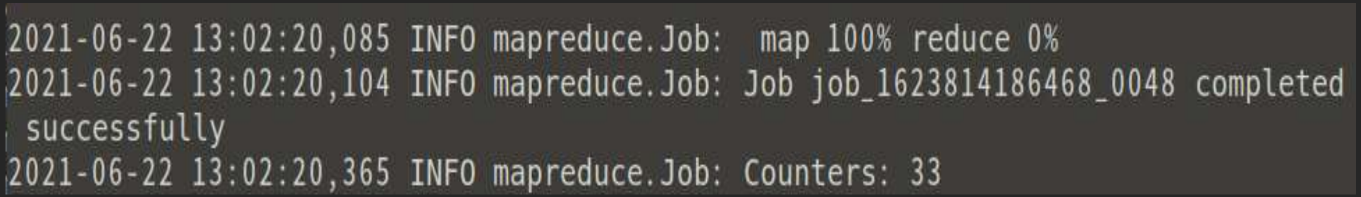


```
Starting namenodes on [master]
master: namenode is running as process 31493.  Stop it first.
Starting datanodes
worker1: datanode is running as process 13732.  Stop it first.
master: datanode is running as process 31678.  Stop it first.
worker3: ssh: connect to host worker3 port 22: No route to host
worker2: ssh: connect to host worker2 port 22: No route to host
Starting secondary namenodes [master]
master: secondarynamenode is running as process 31918.  Stop it first.
Starting resourcemanager
resourcemanager is running as process 32214.  Stop it first.
Starting nodemanagers
worker1: nodemanager is running as process 13900.  Stop it first.
master: nodemanager is running as process 32401.  Stop it first.
worker2: ssh: connect to host worker2 port 22: No route to host
worker3: ssh: connect to host worker3 port 22: No route to host
org.apache.spark.deploy.master.Master running as process 20261.  Stop it first.
worker1: org.apache.spark.deploy.worker.Worker running as process 9246.  Stop it first.
master: org.apache.spark.deploy.worker.Worker running as process 27951.  Stop it first.
ZooKeeper JMX enabled by default
Using config: /usr/zookeeper/bin/../conf/zoo.cfg
grep: /usr/zookeeper/bin/../conf/zoo.cfg: 그런 파일이나 디렉터리가 없습니다
grep: /usr/zookeeper/bin/../conf/zoo.cfg: 그런 파일이나 디렉터리가 없습니다
mkdir: '' 디렉토리를 만들 수 없습니다: 그런 파일이나 디렉터리가 없습니다
Starting zookeeper ... STARTED
Deleted /input/result.txt
mysql: [Warning] Using a password on the command line interface can be insecure.
mysql: [Warning] Using a password on the command line interface can be insecure.
2021-06-22 11:50:31,415 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where
applicable
2021-06-22 11:50:33,937 INFO spark.SparkContext: Running Spark version 3.1.1
2021-06-22 11:50:34,098 INFO resource.ResourceUtils: ==============================================================
2021-06-22 11:50:34,102 INFO resource.ResourceUtils: No custom resources configured for spark.driver.
2021-06-22 11:50:34,106 INFO resource.ResourceUtils: ==============================================================
2021-06-22 11:50:34,107 INFO spark.SparkContext: Submitted application: a
2021-06-22 11:50:34,176 INFO resource.ResourceProfile: Default ResourceProfile created, executor resources: Map(cores -> name: cores, amount: 1,
script: , vendor: , memory -> name: memory, amount: 1024, script: , vendor: , offHeap -> name: offHeap, amount: 0, script: , vendor: ), task resources:
Map(cpus -> name: cpus, amount: 1.0)
2021-06-22 11:50:34,231 INFO resource.ResourceProfile: Limiting resource is cpu
2021-06-22 11:50:34,232 INFO resource.ResourceProfileManager: Added ResourceProfile id: 0
2021-06-22 11:50:34,396 INFO spark.SecurityManager: Changing view acls to: root
2021-06-22 11:50:34,396 INFO spark.SecurityManager: Changing modify acls to: root
2021-06-22 11:50:34,397 INFO spark.SecurityManager: Changing view acls groups to:
```

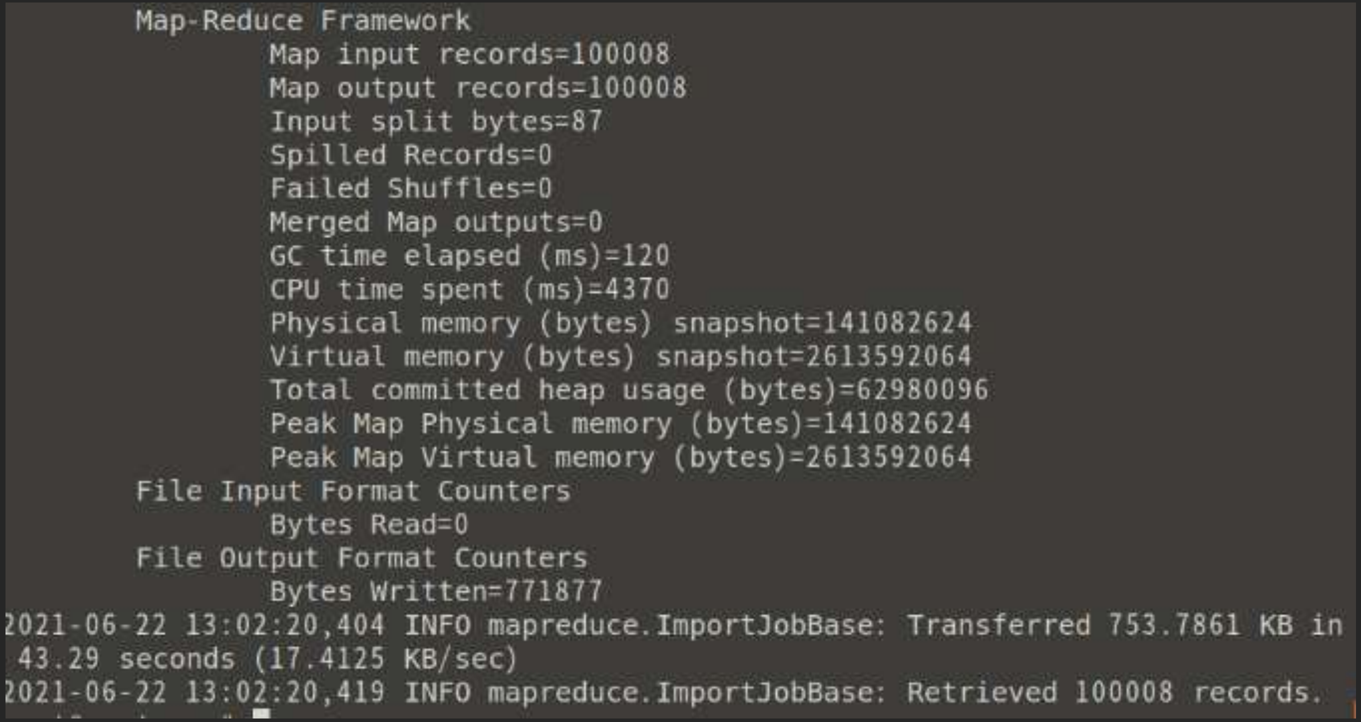
Cron.log

실행 로그 저장

07.
SYSTEM SKILLS



```
2021-06-22 13:02:20,085 INFO mapreduce.Job:  map 100% reduce 0%
2021-06-22 13:02:20,104 INFO mapreduce.Job: Job job_1623814186468_0048 completed
 successfully
2021-06-22 13:02:20,365 INFO mapreduce.Job: Counters: 33
```



```
        Map-Reduce Framework
                Map input records=100008
                Map output records=100008
                Input split bytes=87
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=120
                CPU time spent (ms)=4370
                Physical memory (bytes) snapshot=141082624
                Virtual memory (bytes) snapshot=2613592064
                Total committed heap usage (bytes)=62980096
                Peak Map Physical memory (bytes)=141082624
                Peak Map Virtual memory (bytes)=2613592064
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=771877
2021-06-22 13:02:20,404 INFO mapreduce.ImportJobBase: Transferred 753.7861 KB in
 43.29 seconds (17.4125 KB/sec)
2021-06-22 13:02:20,419 INFO mapreduce.ImportJobBase: Retrieved 100008 records.
```

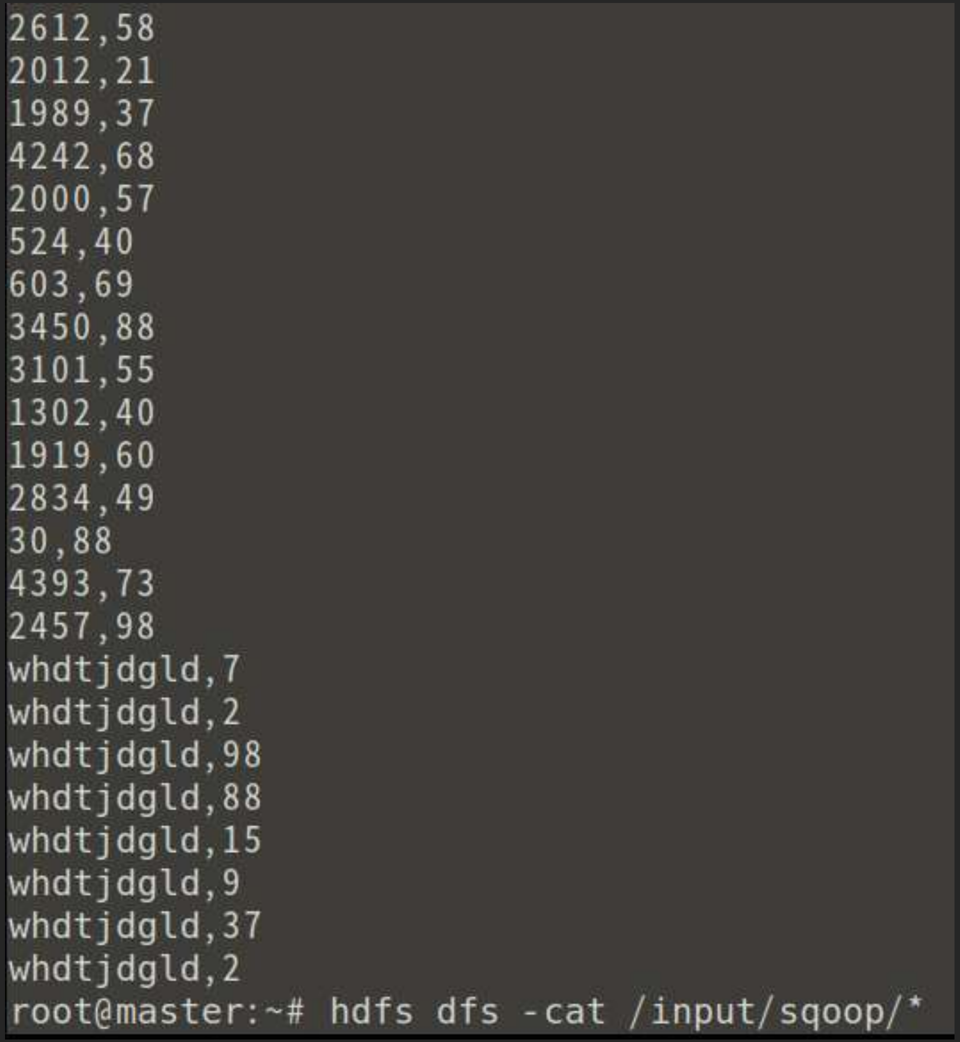
Map reduce

Mysql -> HDFS

Import

완료

07.
SYSTEM SKILLS

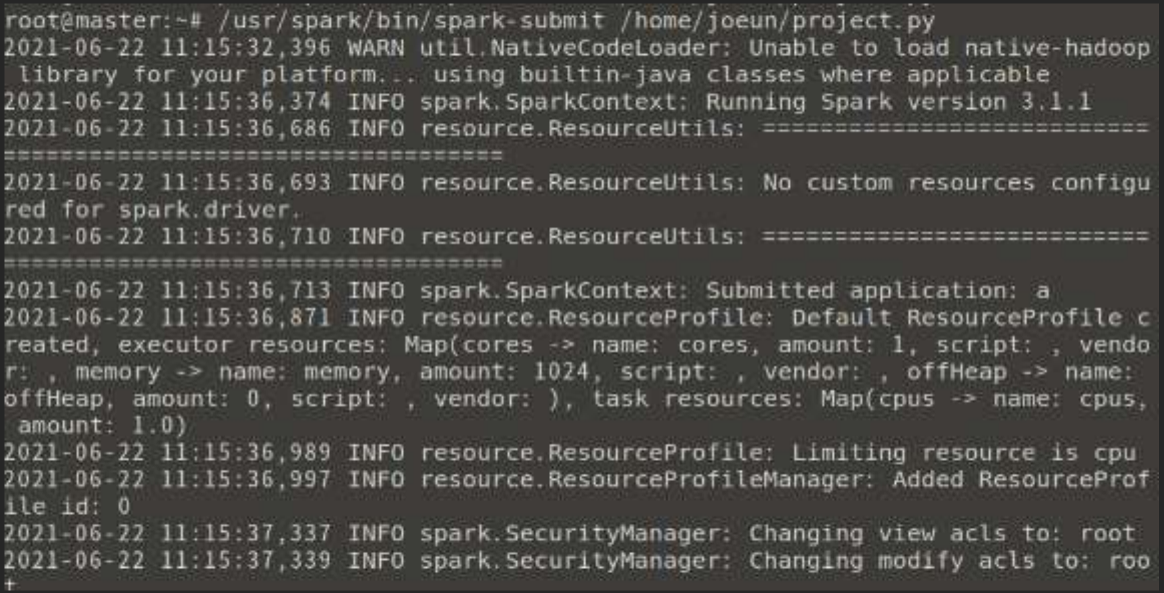


```
2612,58
2012,21
1989,37
4242,68
2000,57
524,40
603,69
3450,88
3101,55
1302,40
1919,60
2834,49
30,88
4393,73
2457,98
whdtjdgld,7
whdtjdgld,2
whdtjdgld,98
whdtjdgld,88
whdtjdgld,15
whdtjdgld,9
whdtjdgld,37
whdtjdgld,2
root@master:~# hdfs dfs -cat /input/sqoop/*
```

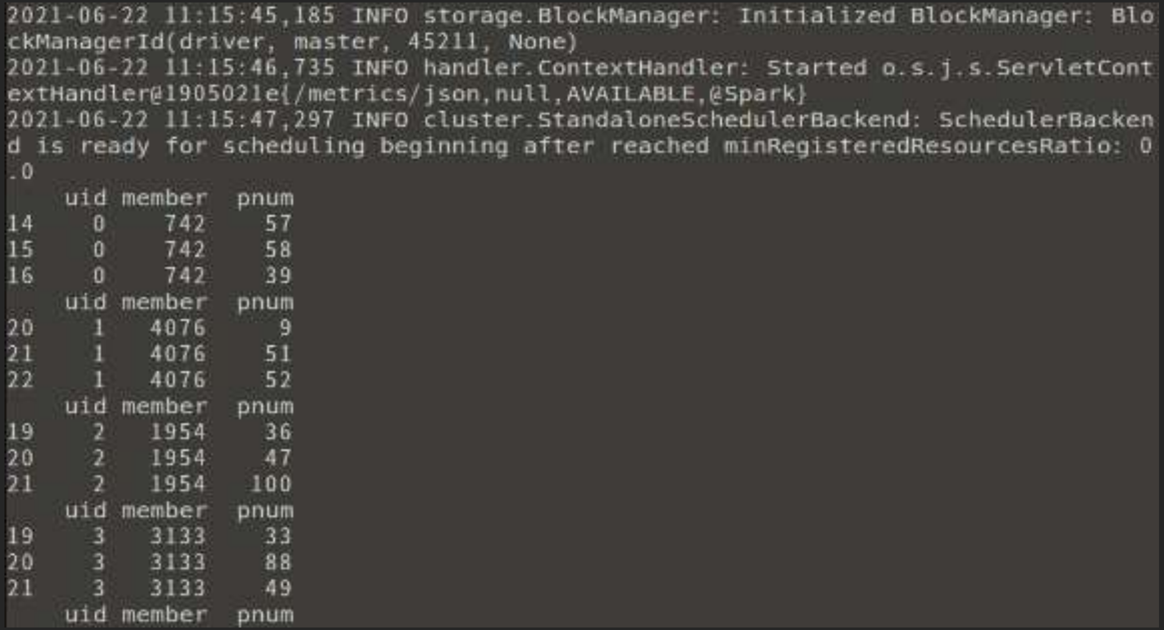
Import 결과

HDFS 저장

07.
SYSTEM SKILLS



```
root@master:~# /usr/spark/bin/spark-submit /home/joeun/project.py
2021-06-22 11:15:32,396 WARN util.NativeCodeLoader: Unable to load native-hadoop
 library for your platform... using builtin-java classes where applicable
2021-06-22 11:15:36,374 INFO spark.SparkContext: Running Spark version 3.1.1
2021-06-22 11:15:36,686 INFO resource.ResourceUtils: ============================
====================================
2021-06-22 11:15:36,693 INFO resource.ResourceUtils: No custom resources configu
red for spark.driver.
2021-06-22 11:15:36,710 INFO resource.ResourceUtils: ============================
====================================
2021-06-22 11:15:36,713 INFO spark.SparkContext: Submitted application: a
2021-06-22 11:15:36,871 INFO resource.ResourceProfile: Default ResourceProfile c
reated, executor resources: Map(cores -> name: cores, amount: 1, script: , vendo
r: , memory -> name: memory, amount: 1024, script: , vendor: , offHeap -> name:
offHeap, amount: 0, script: , vendor: ), task resources: Map(cpus -> name: cpus,
 amount: 1.0)
2021-06-22 11:15:36,989 INFO resource.ResourceProfile: Limiting resource is cpu
2021-06-22 11:15:36,997 INFO resource.ResourceProfileManager: Added ResourceProf
ile id: 0
2021-06-22 11:15:37,337 INFO spark.SecurityManager: Changing view acls to: root
2021-06-22 11:15:37,339 INFO spark.SecurityManager: Changing modify acls to: roo
t
```



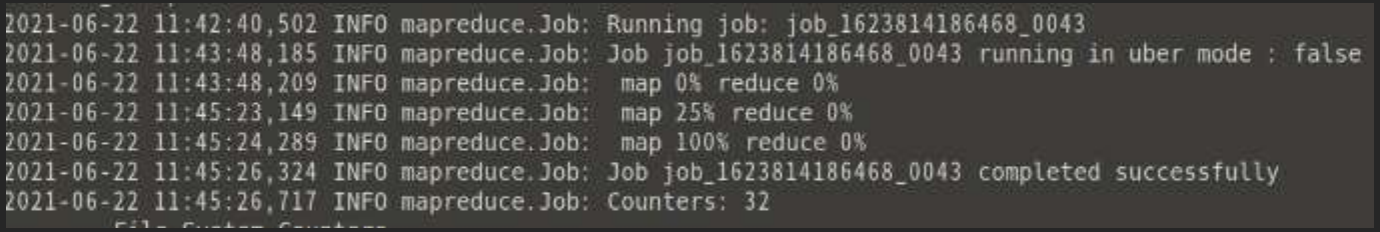
```
2021-06-22 11:15:45,185 INFO storage.BlockManager: Initialized BlockManager: Blo
ckManagerId(driver, master, 45211, None)
2021-06-22 11:15:46,735 INFO handler.ContextHandler: Started o.s.j.s.ServletCont
extHandler@1905021e{/metrics/json,null,AVAILABLE,@Spark}
2021-06-22 11:15:47,297 INFO cluster.StandaloneSchedulerBackend: SchedulerBacken
d is ready for scheduling beginning after reached minRegisteredResourcesRatio: 0
.0
    uid  member  pnum
14    0     742    57
15    0     742    58
16    0     742    39
    uid  member  pnum
20    1    4076     9
21    1    4076    51
22    1    4076    52
    uid  member  pnum
19    2    1954    36
20    2    1954    47
21    2    1954   100
    uid  member  pnum
19    3    3133    33
20    3    3133    88
21    3    3133    49
    uid  member  pnum
```

분산 분석

.py Script 실행

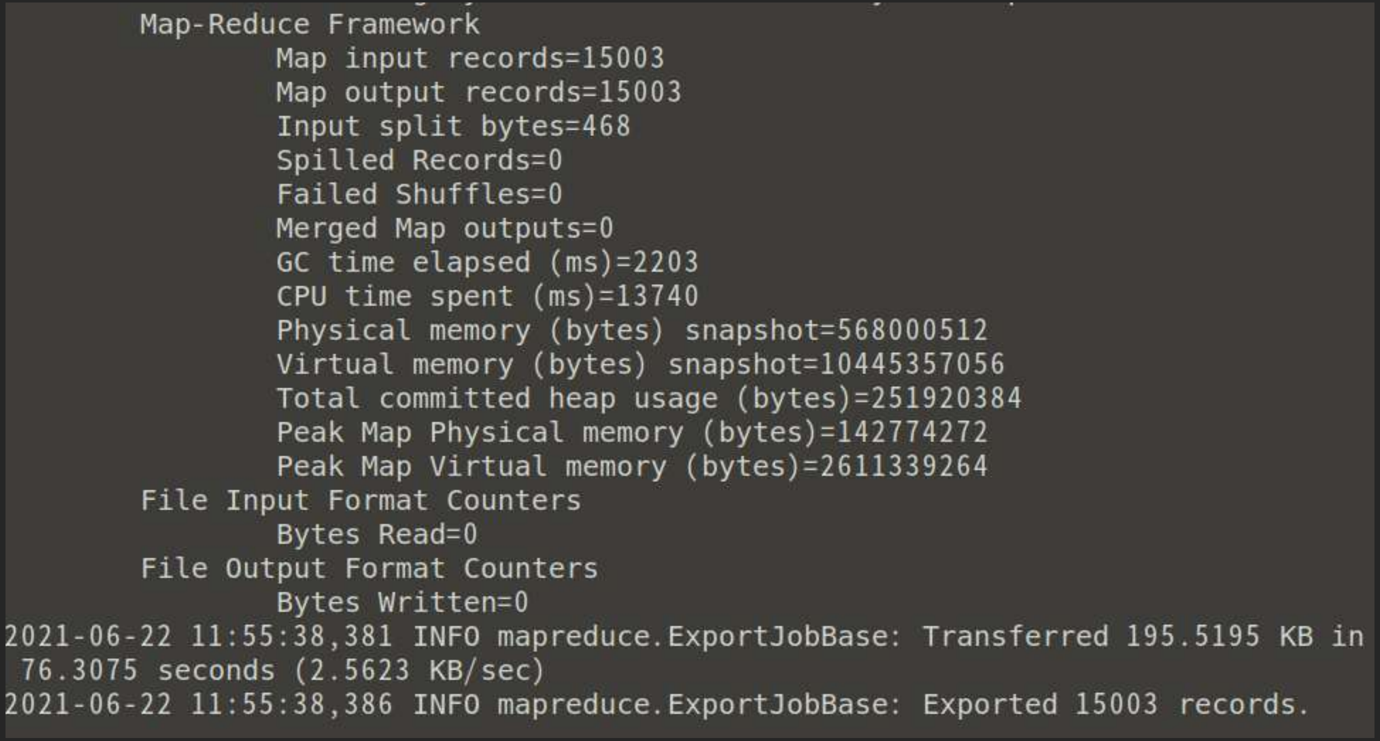
분석 진행

07.
SYSTEM SKILLS



```
2021-06-22 11:42:40,502 INFO mapreduce.Job: Running job: job_1623814186468_0043
2021-06-22 11:43:48,185 INFO mapreduce.Job: Job job_1623814186468_0043 running in uber mode : false
2021-06-22 11:43:48,209 INFO mapreduce.Job:  map 0% reduce 0%
2021-06-22 11:45:23,149 INFO mapreduce.Job:  map 25% reduce 0%
2021-06-22 11:45:24,289 INFO mapreduce.Job:  map 100% reduce 0%
2021-06-22 11:45:26,324 INFO mapreduce.Job: Job job_1623814186468_0043 completed successfully
2021-06-22 11:45:26,717 INFO mapreduce.Job: Counters: 32
        File System Counters
```

Map reduce



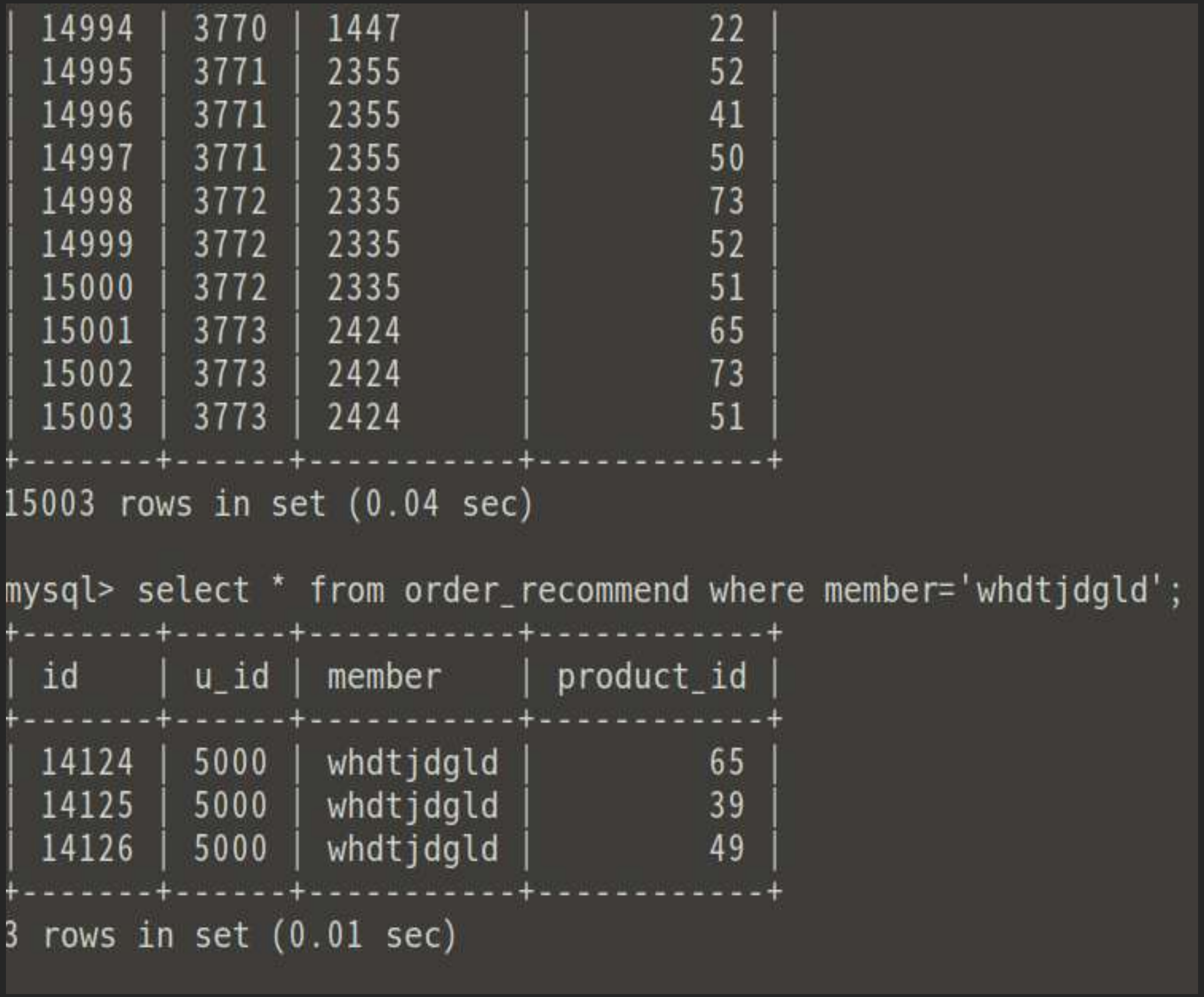
```
        Map-Reduce Framework
                Map input records=15003
                Map output records=15003
                Input split bytes=468
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=2203
                CPU time spent (ms)=13740
                Physical memory (bytes) snapshot=568000512
                Virtual memory (bytes) snapshot=10445357056
                Total committed heap usage (bytes)=251920384
                Peak Map Physical memory (bytes)=142774272
                Peak Map Virtual memory (bytes)=2611339264
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=0
2021-06-22 11:55:38,381 INFO mapreduce.ExportJobBase: Transferred 195.5195 KB in
 76.3075 seconds (2.5623 KB/sec)
2021-06-22 11:55:38,386 INFO mapreduce.ExportJobBase: Exported 15003 records.
```

HDFS -> Mysql

Export

완료

07.
SYSTEM SKILLS



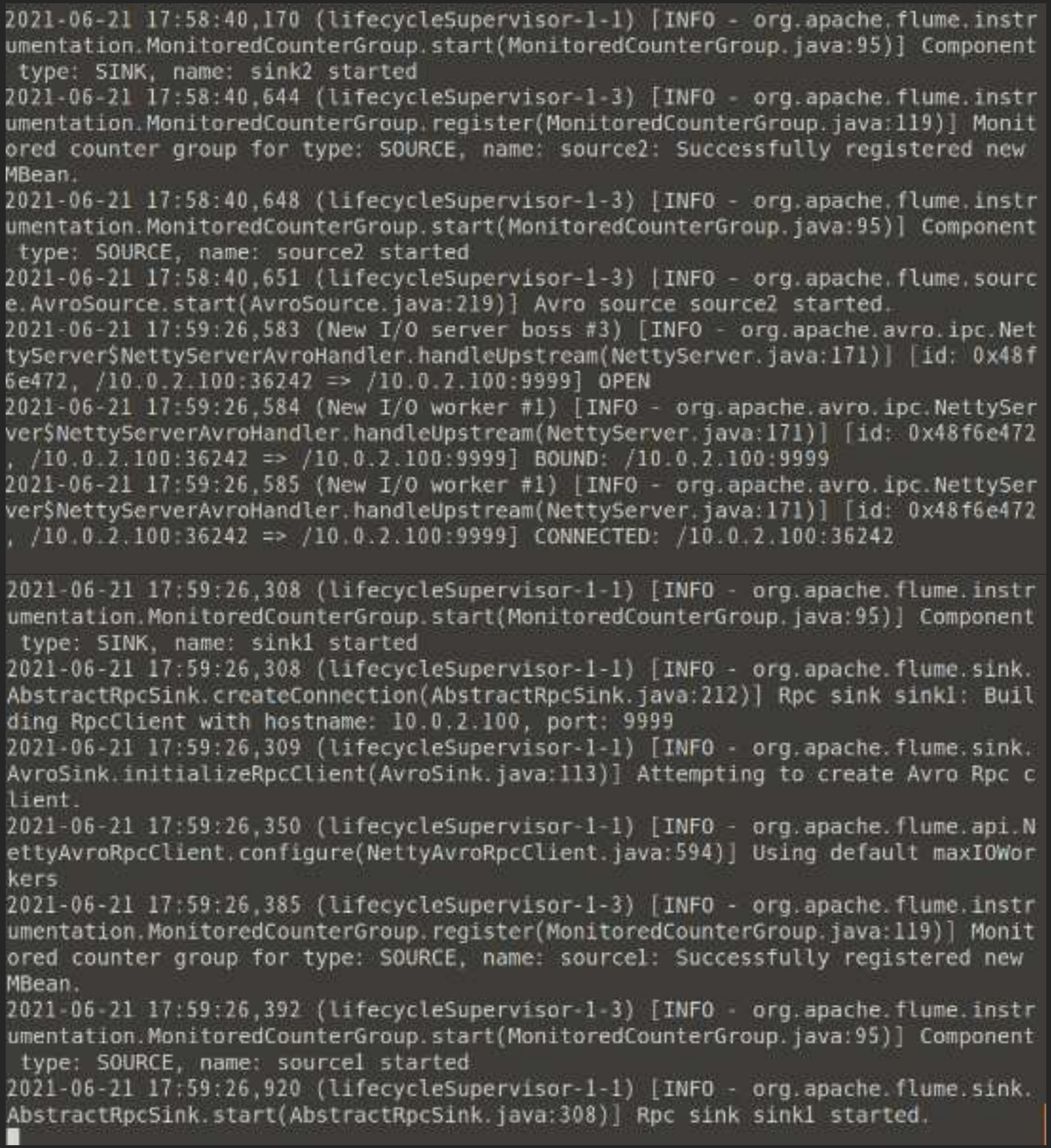
```
| 14994 | 3770 | 1447      |         22 |
| 14995 | 3771 | 2355      |         52 |
| 14996 | 3771 | 2355      |         41 |
| 14997 | 3771 | 2355      |         50 |
| 14998 | 3772 | 2335      |         73 |
| 14999 | 3772 | 2335      |         52 |
| 15000 | 3772 | 2335      |         51 |
| 15001 | 3773 | 2424      |         65 |
| 15002 | 3773 | 2424      |         73 |
| 15003 | 3773 | 2424      |         51 |
+-------+------+-----------+------------+
15003 rows in set (0.04 sec)

mysql> select * from order_recommend where member='whdtjdgld';
+-------+------+-----------+------------+
| id    | u_id | member    | product_id |
+-------+------+-----------+------------+
| 14124 | 5000 | whdtjdgld |         65 |
| 14125 | 5000 | whdtjdgld |         39 |
| 14126 | 5000 | whdtjdgld |         49 |
+-------+------+-----------+------------+
3 rows in set (0.01 sec)
```

Export 결과

Mysql

Table에 저장

07.
SYSTEM SKILLS



```
2021-06-21 17:58:40,170 (lifecycleSupervisor-1-1) [INFO - org.apache.flume.instr
umentation.MonitoredCounterGroup.start(MonitoredCounterGroup.java:95)] Component
 type: SINK, name: sink2 started
2021-06-21 17:58:40,644 (lifecycleSupervisor-1-3) [INFO - org.apache.flume.instr
umentation.MonitoredCounterGroup.register(MonitoredCounterGroup.java:119)] Monit
ored counter group for type: SOURCE, name: source2: Successfully registered new
MBean.
2021-06-21 17:58:40,648 (lifecycleSupervisor-1-3) [INFO - org.apache.flume.instr
umentation.MonitoredCounterGroup.start(MonitoredCounterGroup.java:95)] Component
 type: SOURCE, name: source2 started
2021-06-21 17:58:40,651 (lifecycleSupervisor-1-3) [INFO - org.apache.flume.sourc
e.AvroSource.start(AvroSource.java:219)] Avro source source2 started.
2021-06-21 17:59:26,583 (New I/O server boss #3) [INFO - org.apache.avro.ipc.Net
tyServer$NettyServerAvroHandler.handleUpstream(NettyServer.java:171)] [id: 0x48f
6e472, /10.0.2.100:36242 => /10.0.2.100:9999] OPEN
2021-06-21 17:59:26,584 (New I/O worker #1) [INFO - org.apache.avro.ipc.NettySer
ver$NettyServerAvroHandler.handleUpstream(NettyServer.java:171)] [id: 0x48f6e472
, /10.0.2.100:36242 => /10.0.2.100:9999] BOUND: /10.0.2.100:9999
2021-06-21 17:59:26,585 (New I/O worker #1) [INFO - org.apache.avro.ipc.NettySer
ver$NettyServerAvroHandler.handleUpstream(NettyServer.java:171)] [id: 0x48f6e472
, /10.0.2.100:36242 => /10.0.2.100:9999] CONNECTED: /10.0.2.100:36242
```

```
2021-06-21 17:59:26,308 (lifecycleSupervisor-1-1) [INFO - org.apache.flume.instr
umentation.MonitoredCounterGroup.start(MonitoredCounterGroup.java:95)] Component
 type: SINK, name: sink1 started
2021-06-21 17:59:26,308 (lifecycleSupervisor-1-1) [INFO - org.apache.flume.sink.
AbstractRpcSink.createConnection(AbstractRpcSink.java:212)] Rpc sink sink1: Buil
ding RpcClient with hostname: 10.0.2.100, port: 9999
2021-06-21 17:59:26,309 (lifecycleSupervisor-1-1) [INFO - org.apache.flume.sink.
AvroSink.initializeRpcClient(AvroSink.java:113)] Attempting to create Avro Rpc c
lient.
2021-06-21 17:59:26,350 (lifecycleSupervisor-1-1) [INFO - org.apache.flume.api.N
ettyAvroRpcClient.configure(NettyAvroRpcClient.java:594)] Using default maxIOWor
kers
2021-06-21 17:59:26,385 (lifecycleSupervisor-1-3) [INFO - org.apache.flume.instr
umentation.MonitoredCounterGroup.register(MonitoredCounterGroup.java:119)] Monit
ored counter group for type: SOURCE, name: source1: Successfully registered new
MBean.
2021-06-21 17:59:26,392 (lifecycleSupervisor-1-3) [INFO - org.apache.flume.instr
umentation.MonitoredCounterGroup.start(MonitoredCounterGroup.java:95)] Component
 type: SOURCE, name: source1 started
2021-06-21 17:59:26,920 (lifecycleSupervisor-1-1) [INFO - org.apache.flume.sink.
AbstractRpcSink.start(AbstractRpcSink.java:308)] Rpc sink sink1 started.
```

Master Flume

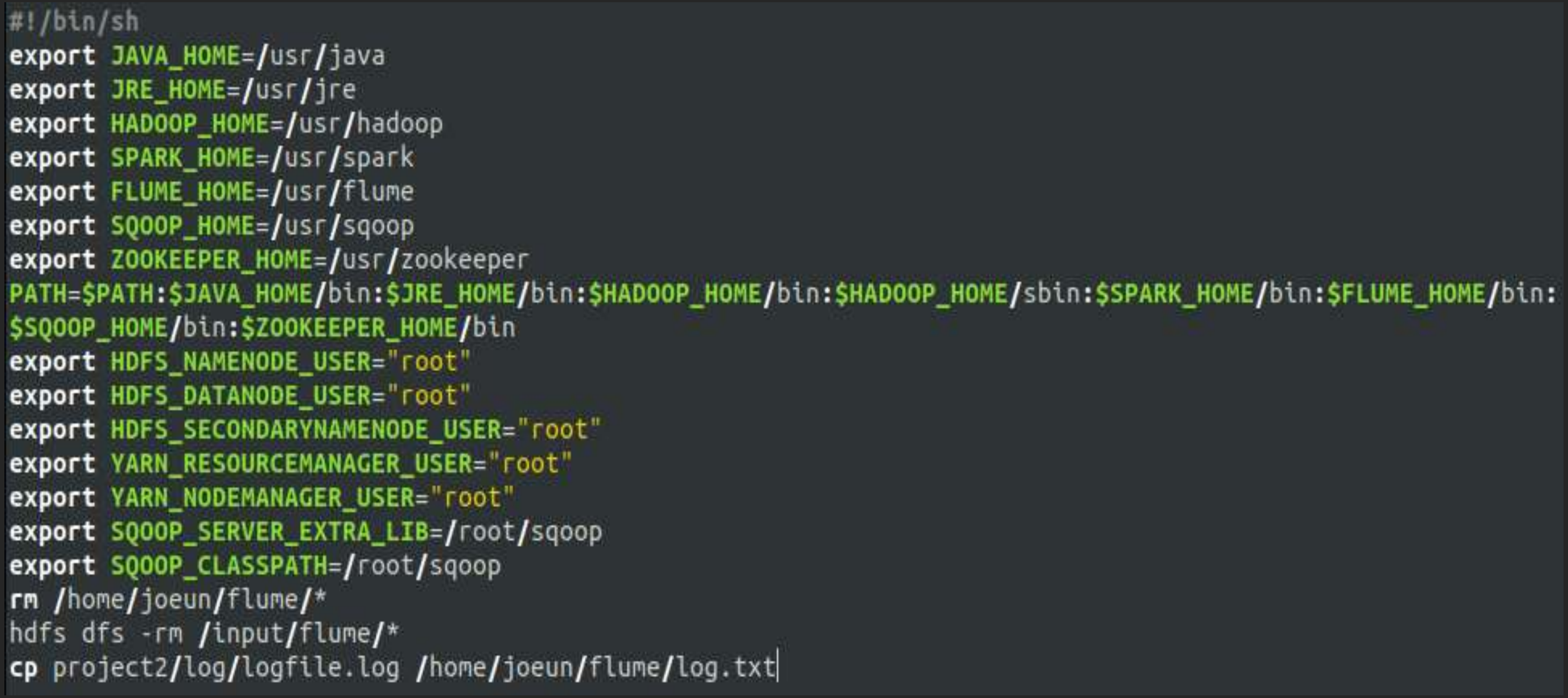
서버 실행

Master Flume

클라이언트 실행

07.

SYSTEM SKILLS

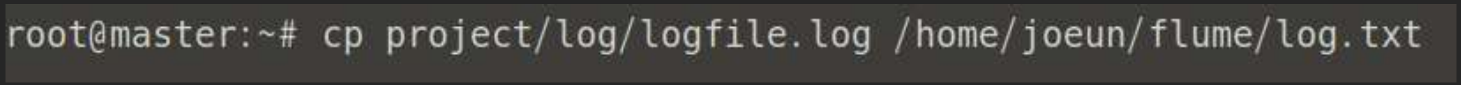


```
#!/bin/sh
export JAVA_HOME=/usr/java
export JRE_HOME=/usr/jre
export HADOOP_HOME=/usr/hadoop
export SPARK_HOME=/usr/spark
export FLUME_HOME=/usr/flume
export SQOOP_HOME=/usr/sqoop
export ZOOKEEPER_HOME=/usr/zookeeper
PATH=$PATH:$JAVA_HOME/bin:$JRE_HOME/bin:$HADOOP_HOME/bin:$HADOOP_HOME/sbin:$SPARK_HOME/bin:$FLUME_HOME/bin:
$SQOOP_HOME/bin:$ZOOKEEPER_HOME/bin
export HDFS_NAMENODE_USER="root"
export HDFS_DATANODE_USER="root"
export HDFS_SECONDARYNAMENODE_USER="root"
export YARN_RESOURCEMANAGER_USER="root"
export YARN_NODEMANAGER_USER="root"
export SQOOP_SERVER_EXTRA_LIB=/root/sqoop
export SQOOP_CLASSPATH=/root/sqoop
rm /home/joeun/flume/*
hdfs dfs -rm /input/flume/*
cp project2/log/logfile.log /home/joeun/flume/log.txt
```

flume.sh

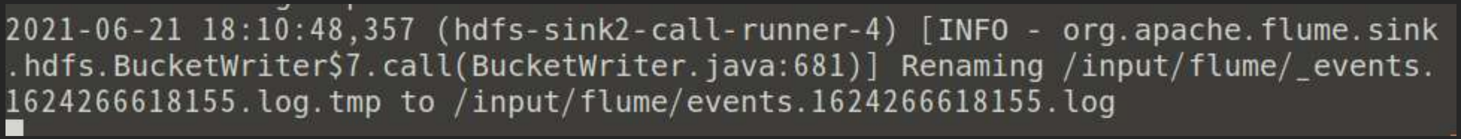
07.

SYSTEM SKILLS



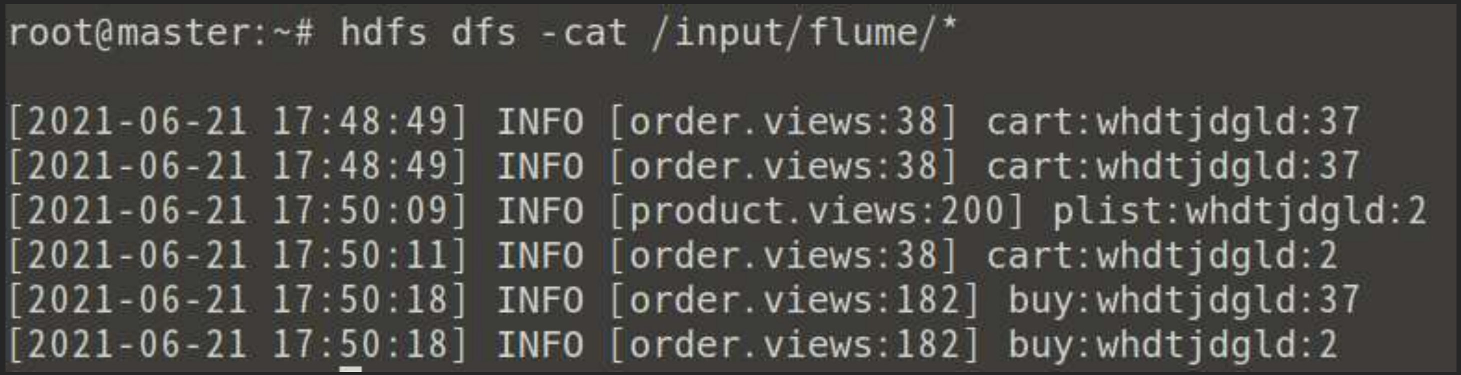
```
root@master:~# cp project/log/logfile.log /home/joeun/flume/log.txt
```

로그파일 전달



```
2021-06-21 18:10:48,357 (hdfs-sink2-call-runner-4) [INFO - org.apache.flume.sink
.hdfs.BucketWriter$7.call(BucketWriter.java:681)] Renaming /input/flume/_events.
1624266618155.log.tmp to /input/flume/events.1624266618155.log
```

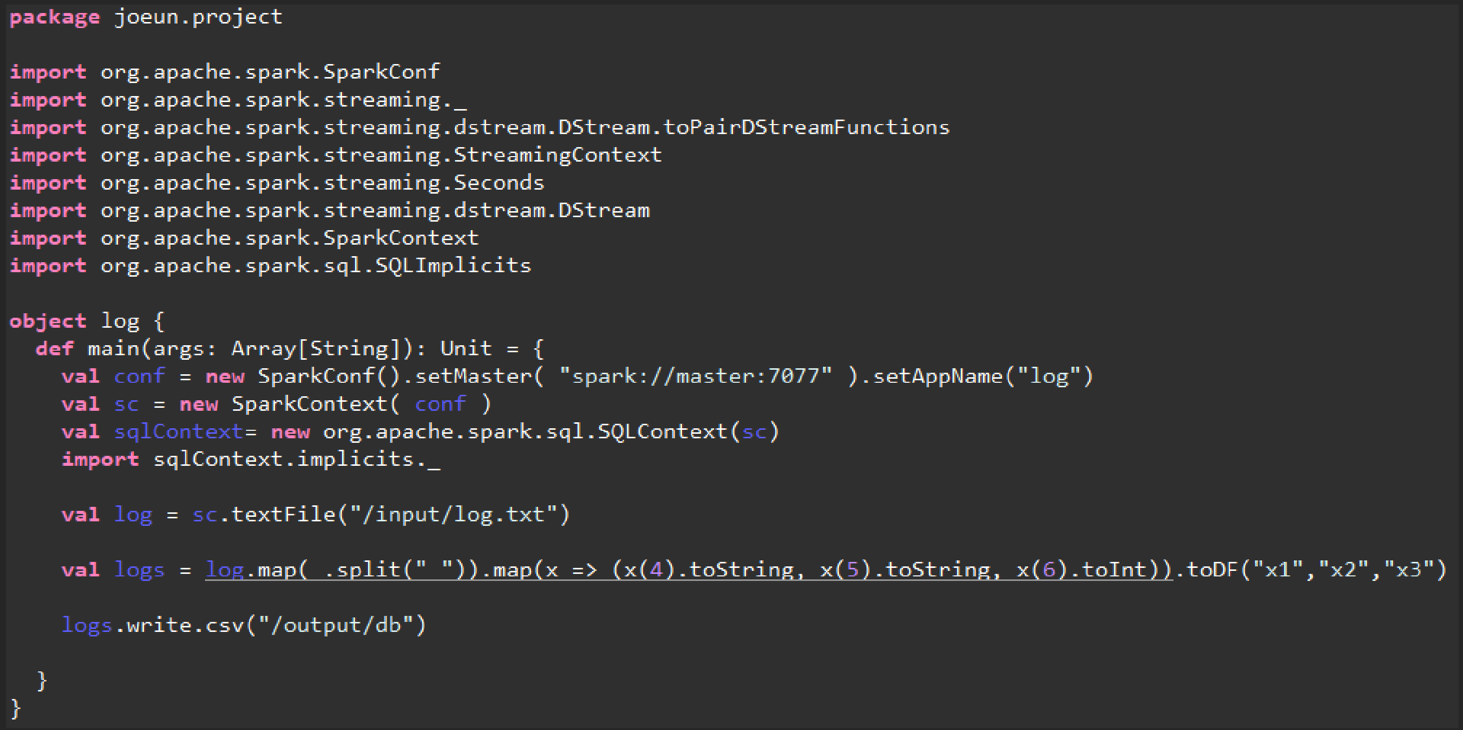
HDFS에 저장



```
root@master:~# hdfs dfs -cat /input/flume/*

[2021-06-21 17:48:49] INFO [order.views:38] cart:whdtjdgld:37
[2021-06-21 17:48:49] INFO [order.views:38] cart:whdtjdgld:37
[2021-06-21 17:50:09] INFO [product.views:200] plist:whdtjdgld:2
[2021-06-21 17:50:11] INFO [order.views:38] cart:whdtjdgld:2
[2021-06-21 17:50:18] INFO [order.views:182] buy:whdtjdgld:37
[2021-06-21 17:50:18] INFO [order.views:182] buy:whdtjdgld:2
```

저장된 파일내용

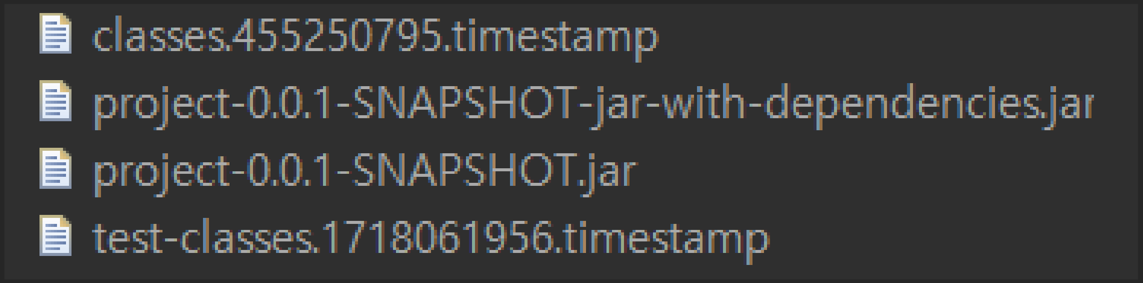
07.
SYSTEM SKILLS



```scala
package joeun.project

import org.apache.spark.SparkConf
import org.apache.spark.streaming._
import org.apache.spark.streaming.dstream.DStream.toPairDStreamFunctions
import org.apache.spark.streaming.StreamingContext
import org.apache.spark.streaming.Seconds
import org.apache.spark.streaming.dstream.DStream
import org.apache.spark.SparkContext
import org.apache.spark.sql.SQLImplicits

object log {
  def main(args: Array[String]): Unit = {
    val conf = new SparkConf().setMaster( "spark://master:7077" ).setAppName("log")
    val sc = new SparkContext( conf )
    val sqlContext= new org.apache.spark.sql.SQLContext(sc)
    import sqlContext.implicits._

    val log = sc.textFile("/input/log.txt")

    val logs = log.map(_.split(" ")).map(x => (x(4).toString, x(5).toString, x(6).toInt)).toDF("x1","x2","x3")

    logs.write.csv("/output/db")

  }
}
```

Log.txt 전처리

Scala 코딩



```
classes.455250795.timestamp
project-0.0.1-SNAPSHOT-jar-with-dependencies.jar
project-0.0.1-SNAPSHOT.jar
test-classes.1718061956.timestamp
```

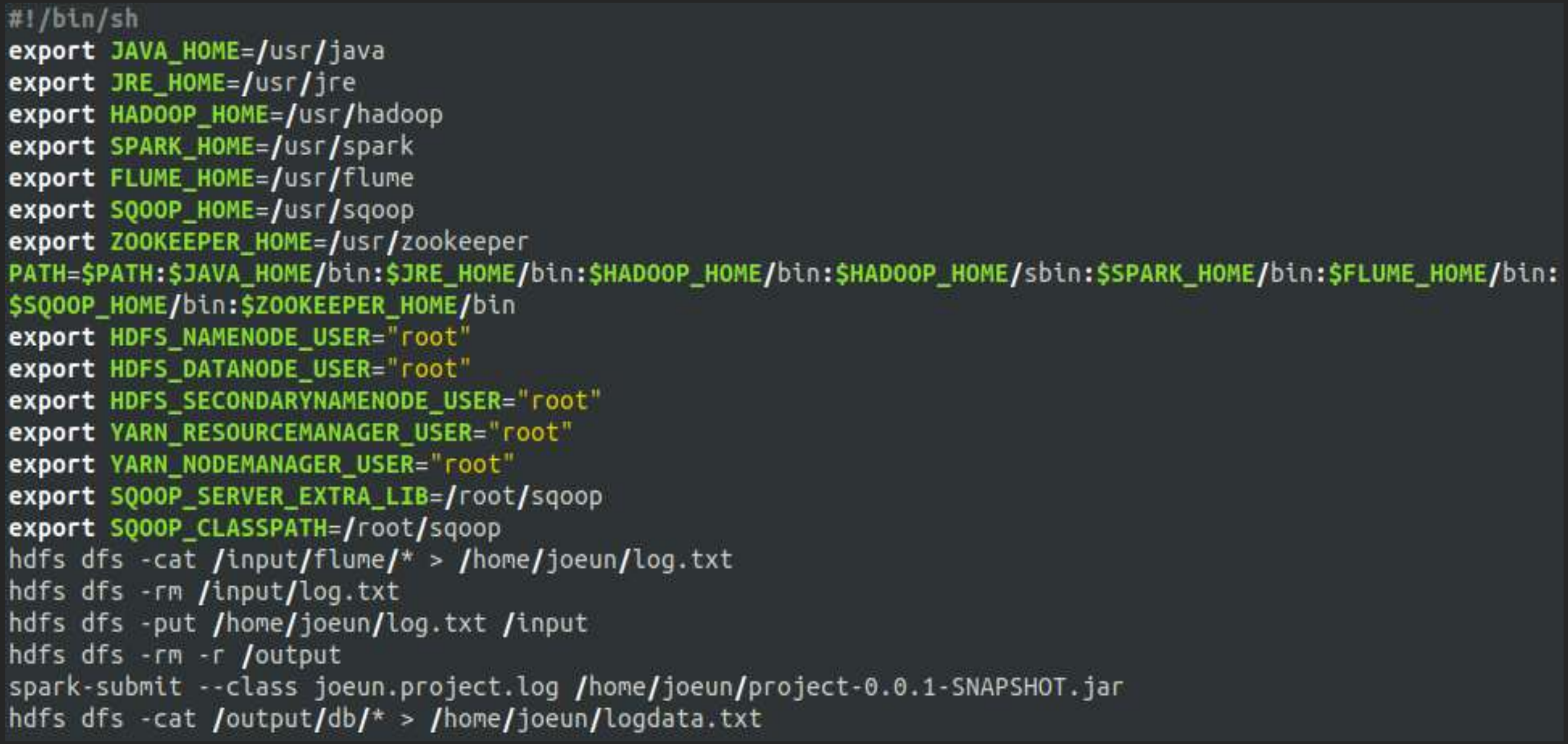
Maven

배포용 .jar파일 생성

07.
SYSTEM SKILLS

Scala

Maven



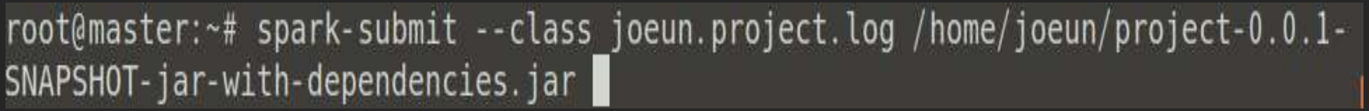
```
#!/bin/sh
export JAVA_HOME=/usr/java
export JRE_HOME=/usr/jre
export HADOOP_HOME=/usr/hadoop
export SPARK_HOME=/usr/spark
export FLUME_HOME=/usr/flume
export SQOOP_HOME=/usr/sqoop
export ZOOKEEPER_HOME=/usr/zookeeper
PATH=$PATH:$JAVA_HOME/bin:$JRE_HOME/bin:$HADOOP_HOME/bin:$HADOOP_HOME/sbin:$SPARK_HOME/bin:$FLUME_HOME/bin:
$SQOOP_HOME/bin:$ZOOKEEPER_HOME/bin
export HDFS_NAMENODE_USER="root"
export HDFS_DATANODE_USER="root"
export HDFS_SECONDARYNAMENODE_USER="root"
export YARN_RESOURCEMANAGER_USER="root"
export YARN_NODEMANAGER_USER="root"
export SQOOP_SERVER_EXTRA_LIB=/root/sqoop
export SQOOP_CLASSPATH=/root/sqoop
hdfs dfs -cat /input/flume/* > /home/joeun/log.txt
hdfs dfs -rm /input/log.txt
hdfs dfs -put /home/joeun/log.txt /input
hdfs dfs -rm -r /output
spark-submit --class joeun.project.log /home/joeun/project-0.0.1-SNAPSHOT.jar
hdfs dfs -cat /output/db/* > /home/joeun/logdata.txt
```

log.sh

07.
SYSTEM SKILLS

Scala

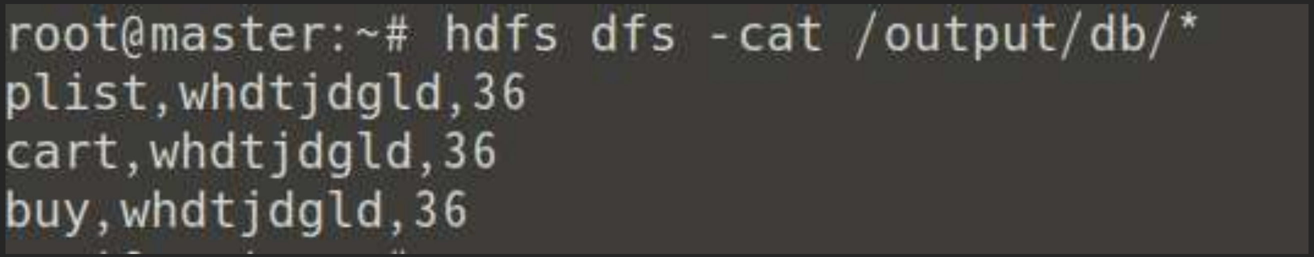
Maven



```
root@master:~# spark-submit --class joeun.project.log /home/joeun/project-0.0.1-
SNAPSHOT-jar-with-dependencies.jar
```

.jar 파일

Script 실행

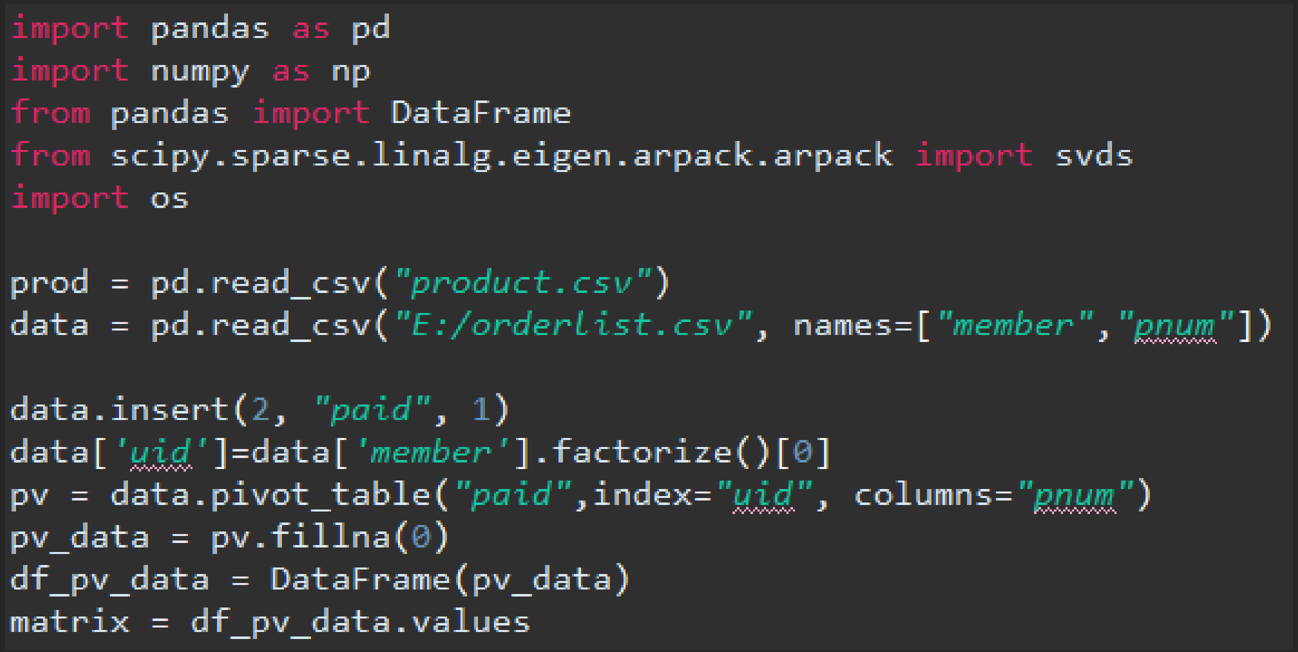


```
root@master:~# hdfs dfs -cat /output/db/*
plist,whdtjdgld,36
cart,whdtjdgld,36
buy,whdtjdgld,36
```

Log.txt 전처리

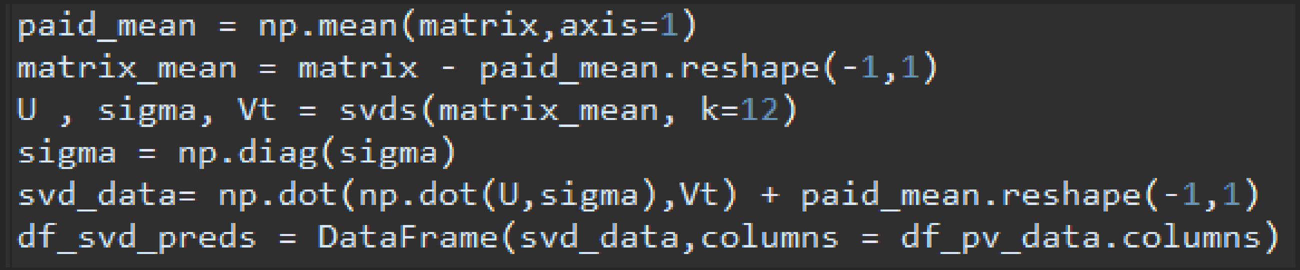
결과

08.
ANALYSIS



```python
import pandas as pd
import numpy as np
from pandas import DataFrame
from scipy.sparse.linalg.eigen.arpack.arpack import svds
import os

prod = pd.read_csv("product.csv")
data = pd.read_csv("E:/orderlist.csv", names=["member","pnum"])

data.insert(2, "paid", 1)
data['uid']=data['member'].factorize()[0]
pv = data.pivot_table("paid",index="uid", columns="pnum")
pv_data = pv.fillna(0)
df_pv_data = DataFrame(pv_data)
matrix = df_pv_data.values
```

사용자 구매 데이터 생성

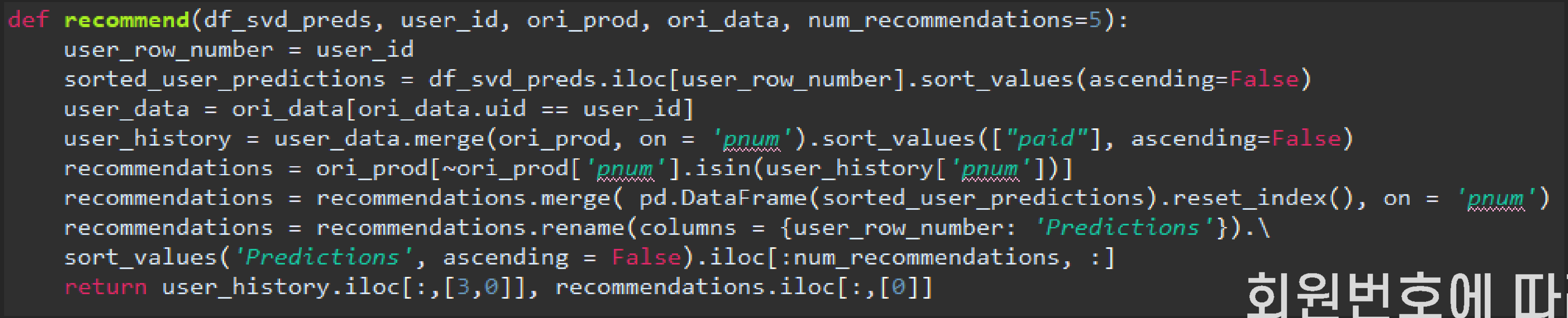


```python
paid_mean = np.mean(matrix,axis=1)
matrix_mean = matrix - paid_mean.reshape(-1,1)
U , sigma, Vt = svds(matrix_mean, k=12)
sigma = np.diag(sigma)
svd_data= np.dot(np.dot(U,sigma),Vt) + paid_mean.reshape(-1,1)
df_svd_preds = DataFrame(svd_data,columns = df_pv_data.columns)
```

행렬분해를 이용한 구매 예측
데이터 생성

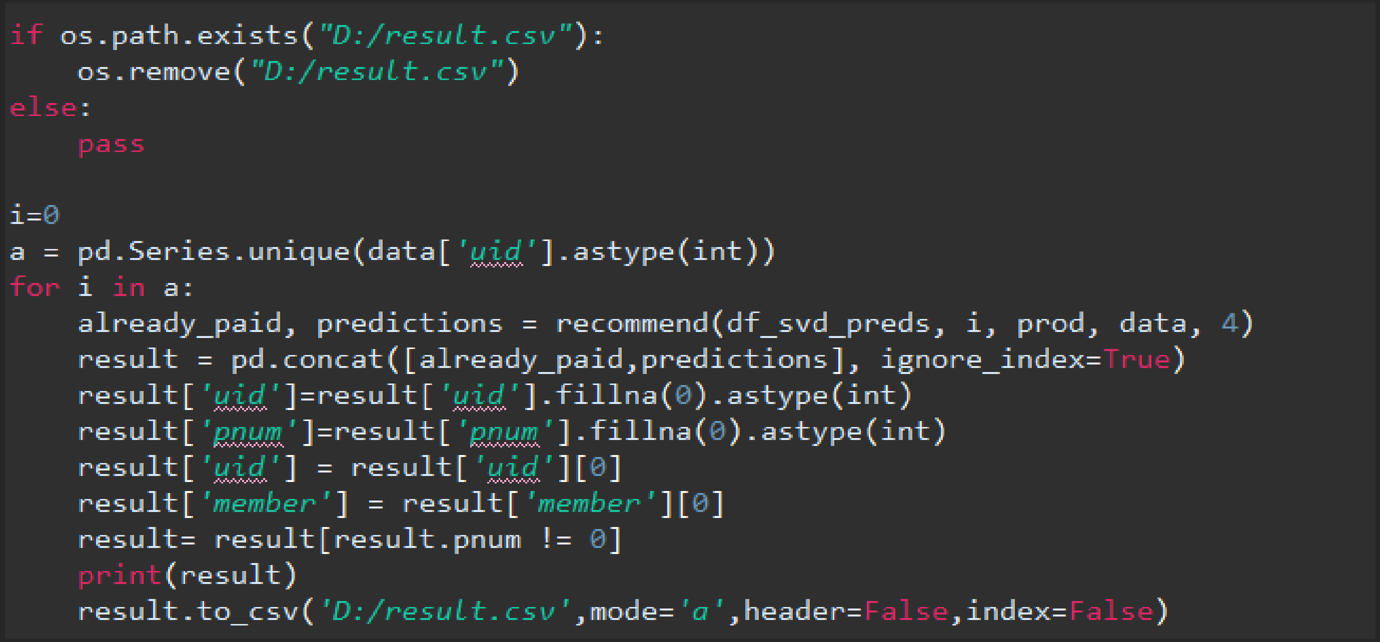
08.

ANALYSIS



```
def recommend(df_svd_preds, user_id, ori_prod, ori_data, num_recommendations=5):
    user_row_number = user_id
    sorted_user_predictions = df_svd_preds.iloc[user_row_number].sort_values(ascending=False)
    user_data = ori_data[ori_data.uid == user_id]
    user_history = user_data.merge(ori_prod, on = 'pnum').sort_values(["paid"], ascending=False)
    recommendations = ori_prod[~ori_prod['pnum'].isin(user_history['pnum'])]
    recommendations = recommendations.merge( pd.DataFrame(sorted_user_predictions).reset_index(), on = 'pnum')
    recommendations = recommendations.rename(columns = {user_row_number: 'Predictions'}).\
    sort_values('Predictions', ascending = False).iloc[:num_recommendations, :]
    return user_history.iloc[:,[3,0]], recommendations.iloc[:,[0]]
```

회원번호에 따라서 추천 상품을 출력하기 위한 함수

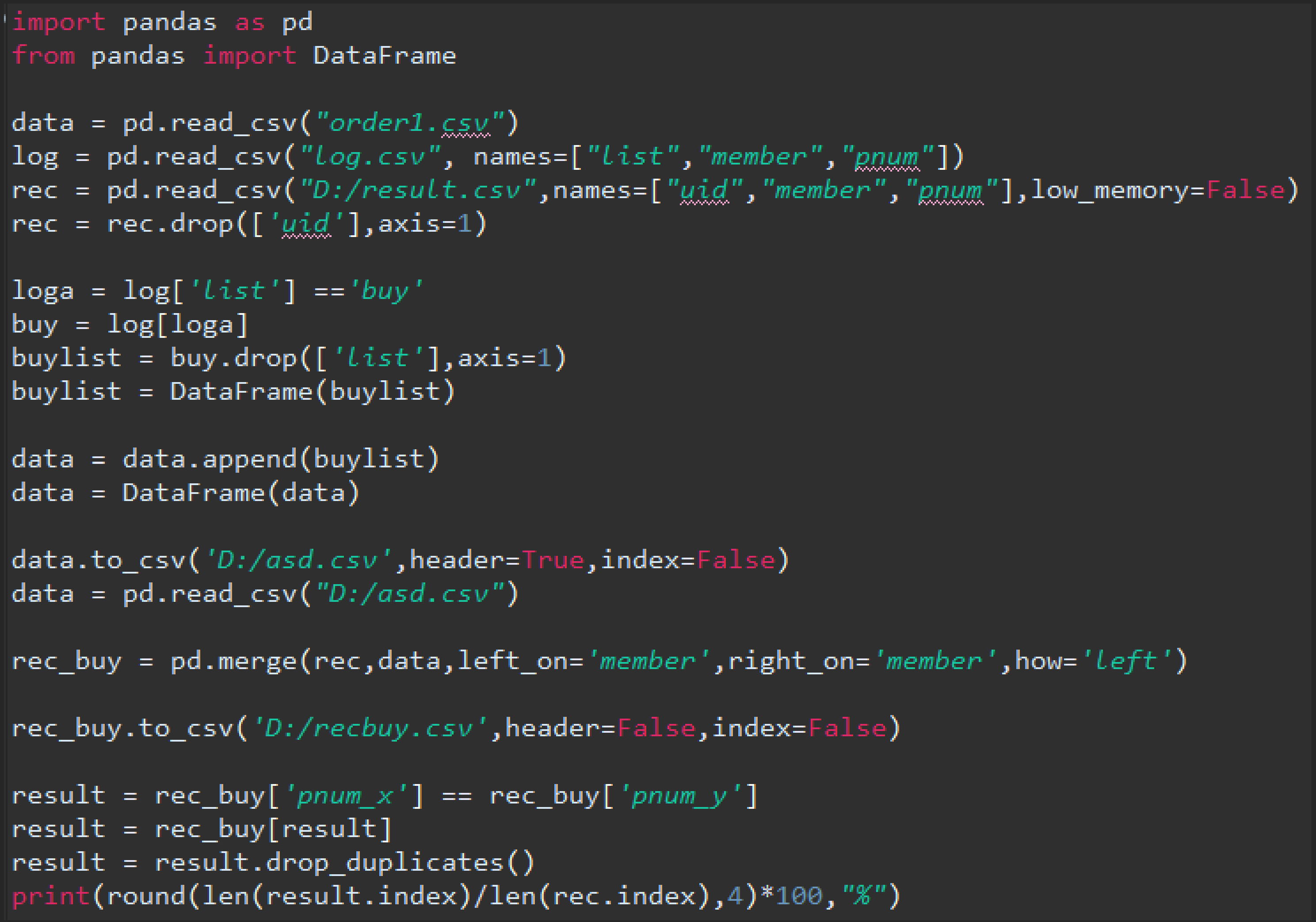


```
if os.path.exists("D:/result.csv"):
    os.remove("D:/result.csv")
else:
    pass

i=0
a = pd.Series.unique(data['uid'].astype(int))
for i in a:
    already_paid, predictions = recommend(df_svd_preds, i, prod, data, 4)
    result = pd.concat([already_paid,predictions], ignore_index=True)
    result['uid']=result['uid'].fillna(0).astype(int)
    result['pnum']=result['pnum'].fillna(0).astype(int)
    result['uid'] = result['uid'][0]
    result['member'] = result['member'][0]
    result= result[result.pnum != 0]
    print(result)
    result.to_csv('D:/result.csv',mode='a',header=False,index=False)
```

회원 전체의 추천을 만드는 반복문

09.

HYPOTHESIS



```
import pandas as pd
from pandas import DataFrame

data = pd.read_csv("order1.csv")
log = pd.read_csv("log.csv", names=["list","member","pnum"])
rec = pd.read_csv("D:/result.csv",names=["uid","member","pnum"],low_memory=False)
rec = rec.drop(['uid'],axis=1)

loga = log['list'] =='buy'
buy = log[loga]
buylist = buy.drop(['list'],axis=1)
buylist = DataFrame(buylist)

data = data.append(buylist)
data = DataFrame(data)

data.to_csv('D:/asd.csv',header=True,index=False)
data = pd.read_csv("D:/asd.csv")

rec_buy = pd.merge(rec,data,left_on='member',right_on='member',how='left')

rec_buy.to_csv('D:/recbuy.csv',header=False,index=False)

result = rec_buy['pnum_x'] == rec_buy['pnum_y']
result = rec_buy[result]
result = result.drop_duplicates()
print(round(len(result.index)/len(rec.index),4)*100,"%")
```

상품 추천 시 구매율

51%

REFERENCES

뉴스 핌
https://www.newspim.com/news/view/20201213000103

월곡 주얼리 산업 진흥재단
https://w-jewel.or.kr/wjrc

WJRC 행사 [2019 한국 주얼리 산업 전략 포럼]
https://blog.naver.com/wjrc1858/221726932299

지인의 익명의 쇼핑몰 판매 데이터

THANK YOU

Q&A