# Cloudera Migration

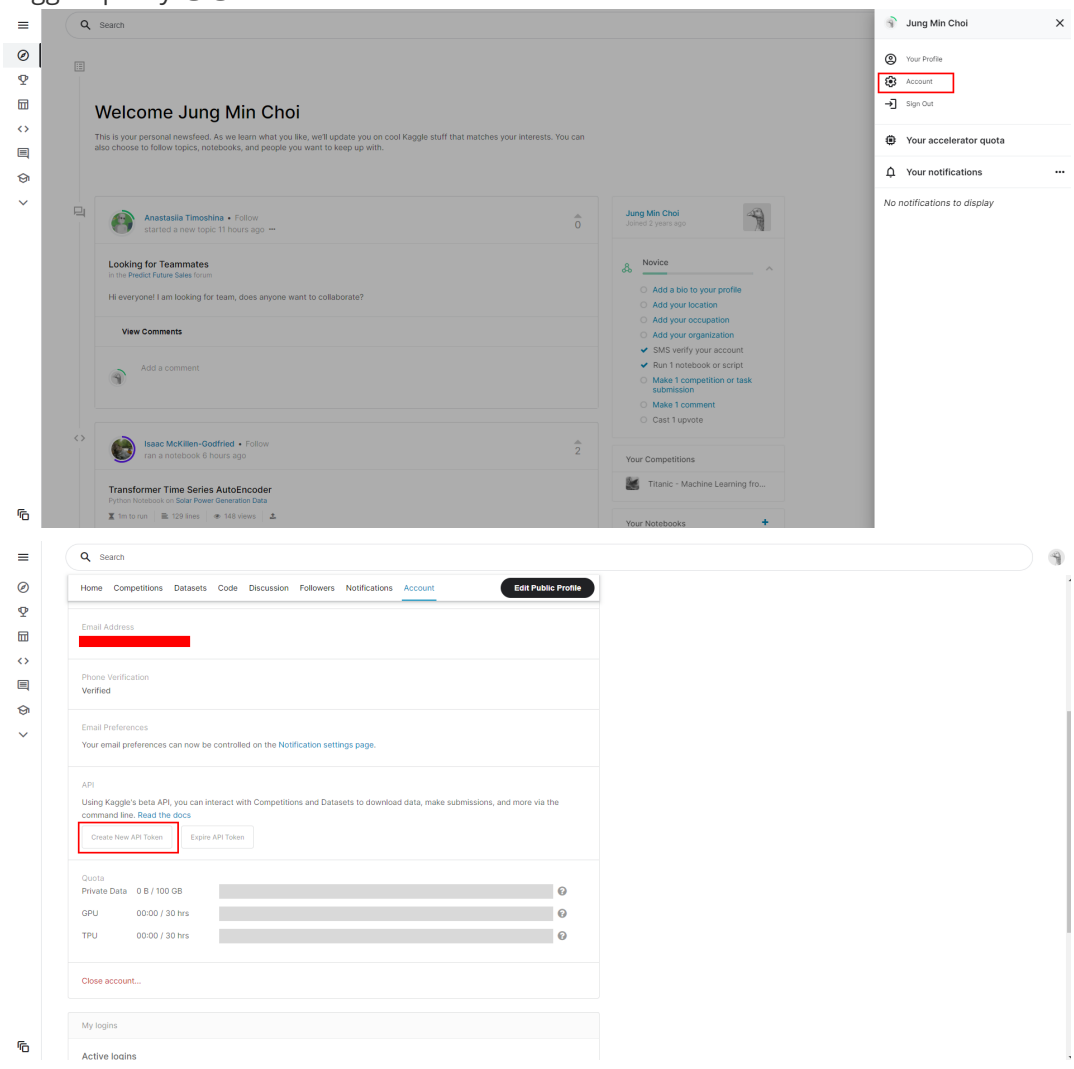


## 1. 데이터 적재

1. 데이터 선정 (Kaggle)
    1. allposts.csv: 43.09 GB
    2. deletedposts.csv : 5 GB
    3. chats_2021-04.csv : 11.17 GB
2. Kaggle 작업 환경 구성

    참고 자료

    1. kaggle api key 생성





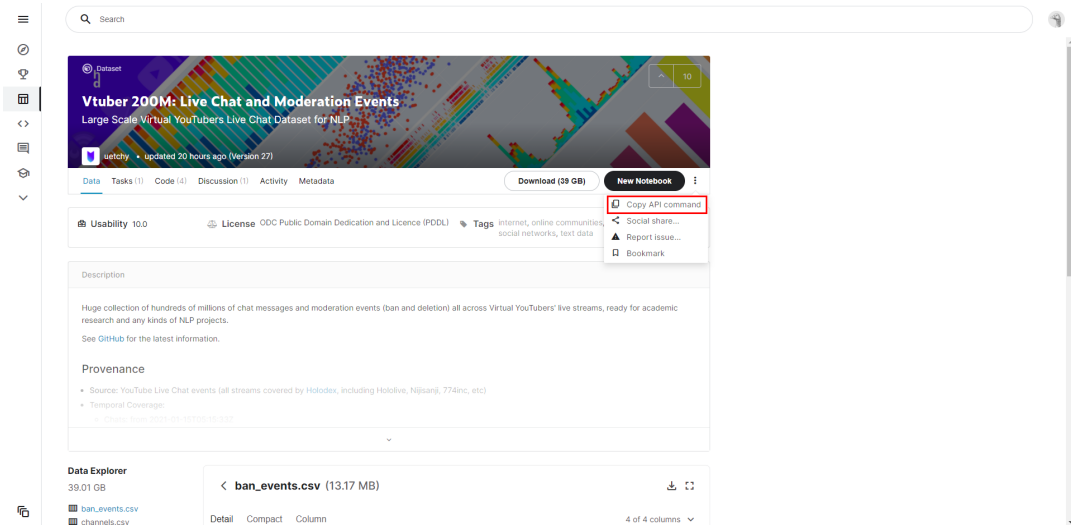


    `kaggle.json` 파일이 로컬에 download 됩니다.

    2. sftp로 kaggle.json 파일 server에 upload
    3. library 설치 및 데이터 download

```
pip3 install --user kaggle

# sftp 로 kaggle.json 파일 ~/에 upload

mkdir -p ~/.kaggle # 유저의 홈디렉토리에 .kaggle 폴더 생성
cp kaggle.json ~/.kaggle/kaggle.json # 현재 폴더의 kaggle.json 파일을 복사
chmod 600 ~/.kaggle/kaggle.json # kaggle.json을 오너만 읽기, 쓰기 권한 할당

export PATH=$PATH:/home/ec2-user/.local/bin # kaggle 명령어를 실행어를 어디서나 실행하기 위해 Path 설정
# 아래 명령어는 위에서 Kaggle Dataset API 복사 된 것을 붙이기 하세요
{KAGGLE API COMMAND}  -p download_data # kaggle 명령어 실행해서 다운로드

mkdir -p datas # data_dir 폴더 생성
unzip  download_data/{download zip file} -d datas/{data dir}

rm -rf download_data # downlaod_dir 폴더 제거
```

{KAGGLE API COMMAND} :



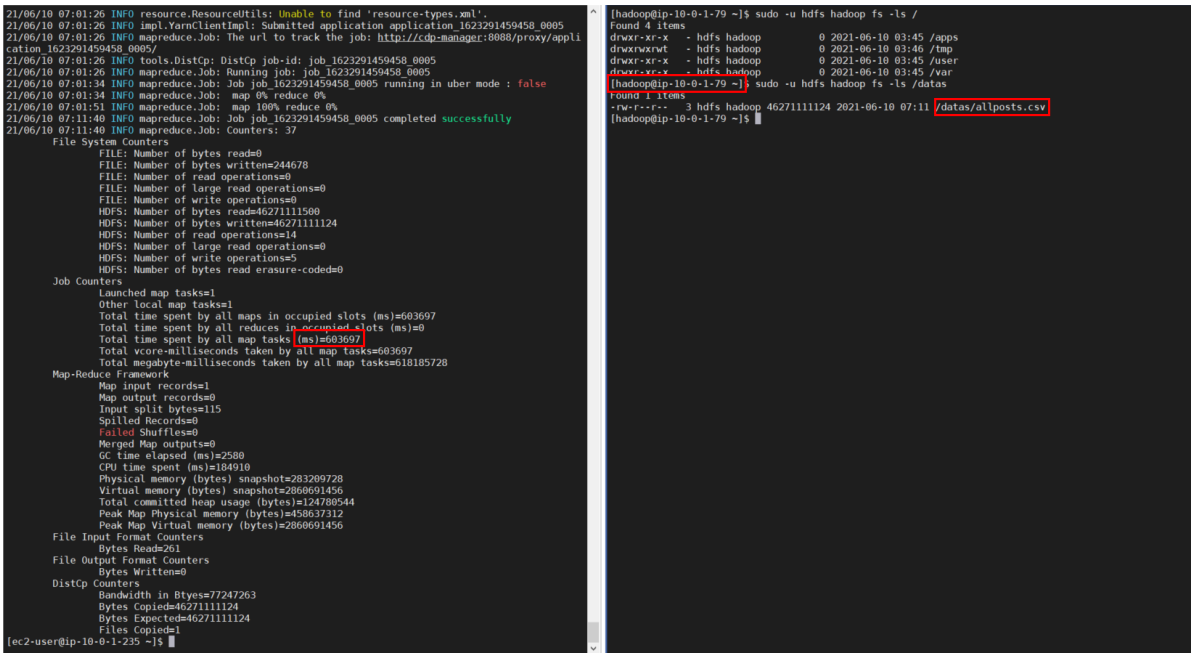

3. 데이터 적재

```
# hdfs 폴더 생성
sudo -u hdfs hadoop fs -mkdir {target_dir}

# hdfs 권한 변경
sudo -u hdfs hadoop fs -chmod 777 {target_dir}

# hdfs 파일 내역 확인
sudo -u hdfs hadoop fs -ls /

# hdfs put 명령어로 데이터 넣기  (40G 기준 약 15분 소모)
sudo -u hdfs hadoop fs -put {source_dir} {target_dir}

# hdfs file size 확인
sudo -u hdfs hadoop fs -du -h /
```

# 2. Cluster to Cluster 데이터 이관

40GB 기준 약 10분 소요

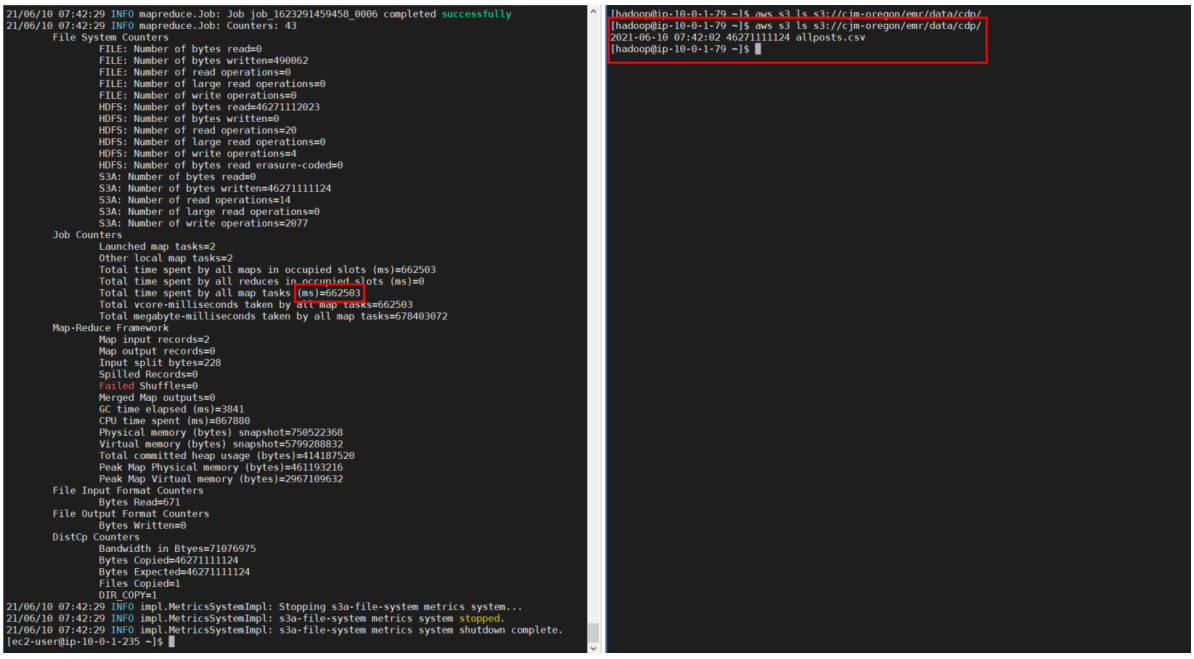
```
# Node Storage 확인
sudo -u hdfs hadoop dfsadmin -report

# 데이터 이관
sudo -u hdfs hadoop distcp -skipcrccheck -update {source_dir}
hdfs://{EMR_Master_Private_Ip}:8020{target_dir}
```

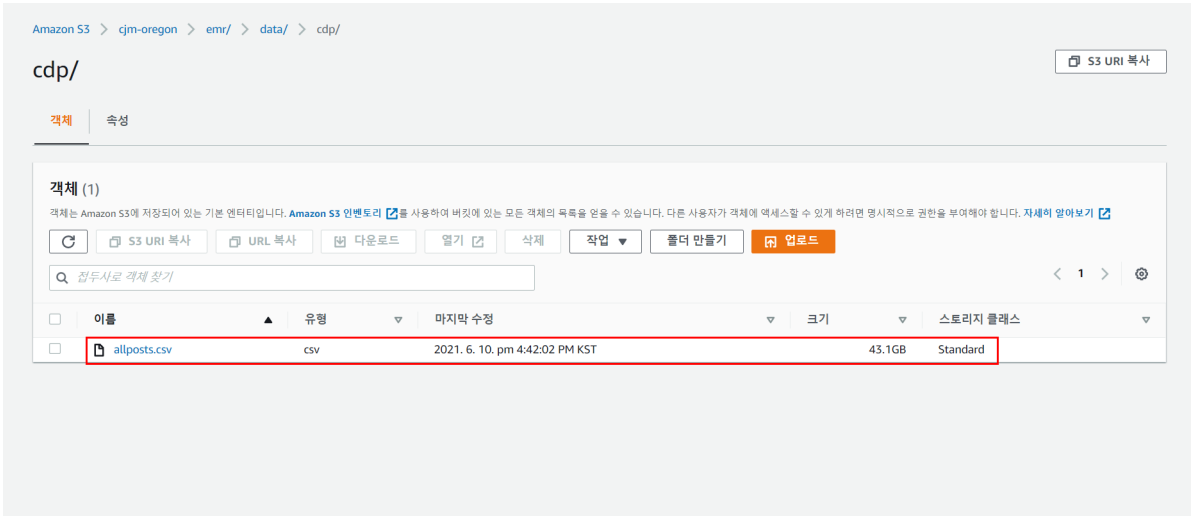



# 3. Cluster to S3 데이터 이관

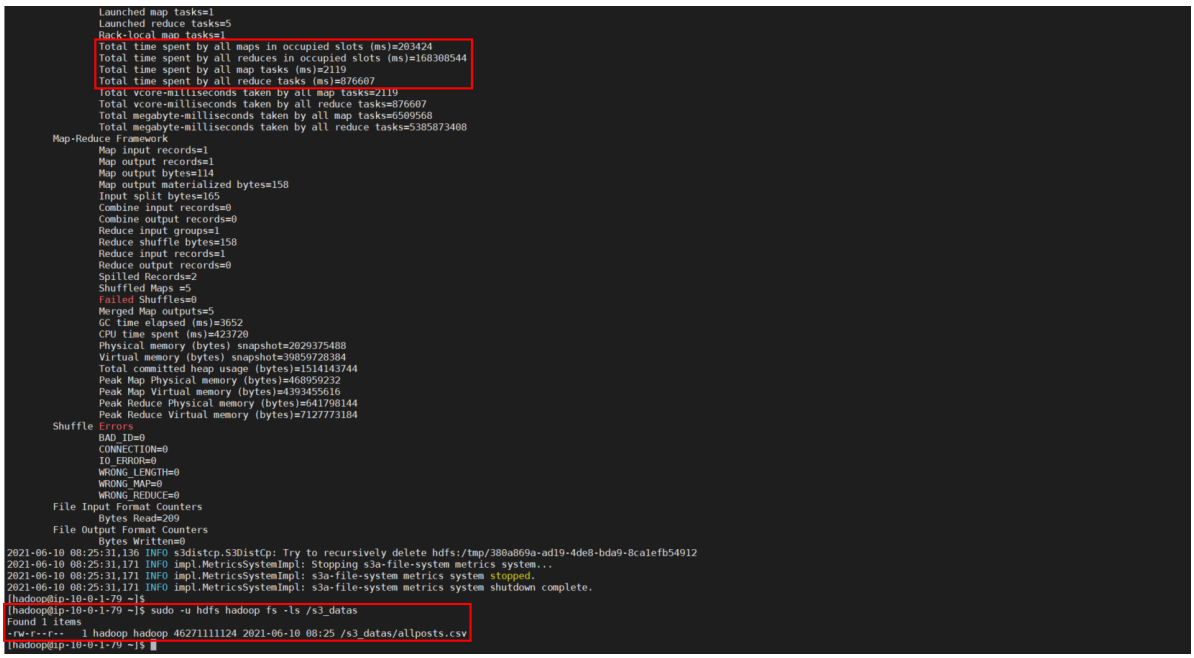
40GB 기준 약 11분 소요

```
# CDP to S3 이관
sudo -u hdfs hadoop distcp -Dfs.s3a.access.key={ACCESS_KEY} -Dfs.s3a.secret.key=
{SECRET_KEY} {source_dir} s3a://{S3_Bucket}/{path}
```






```
# S3 to EMR 이관
s3-dist-cp --src=s3a://{S3_Bucket}/{path} --dest=hdfs://{target_dir}
```




# 4. 비교

비교 방법 : S3 의 파일(약 40GB) distcp로 이동

1. 설정 값

| 유형 | Cloudera | EMR |
| --- | --- | --- |
| Type | t2.large | m5.xlarge |
| vCPU | 2 | 4 |
| 메모리 | 8 | 16 |
| 스토리지 | EBS 전용 | EBS 전용 |
| 명령어 | sudo -u hdfs hadoop distcp -Dfs.s3a.access.key={Access_Key} -Dfs.s3a.secret.key={Secret_Key} s3a://{Bucket_Name}/{Path} {Target_Dir} | s3-dist-cp --src=s3a://{Bucket_Name}/{Path} --dest=hdfs://{Target_Dir} |
| 수행시간 (ms) | 908182 | 907408 |

2. 결과
   약 15분 정도로 비슷한 결과값 도출