

# Cloudera Migration

1. [데이터 적재](#)
2. [Cluster to Cluster 데이터 이관](#)
3. [Cluster to S3 데이터 이관](#)
4. [비교](#)

## 1. 데이터 적재

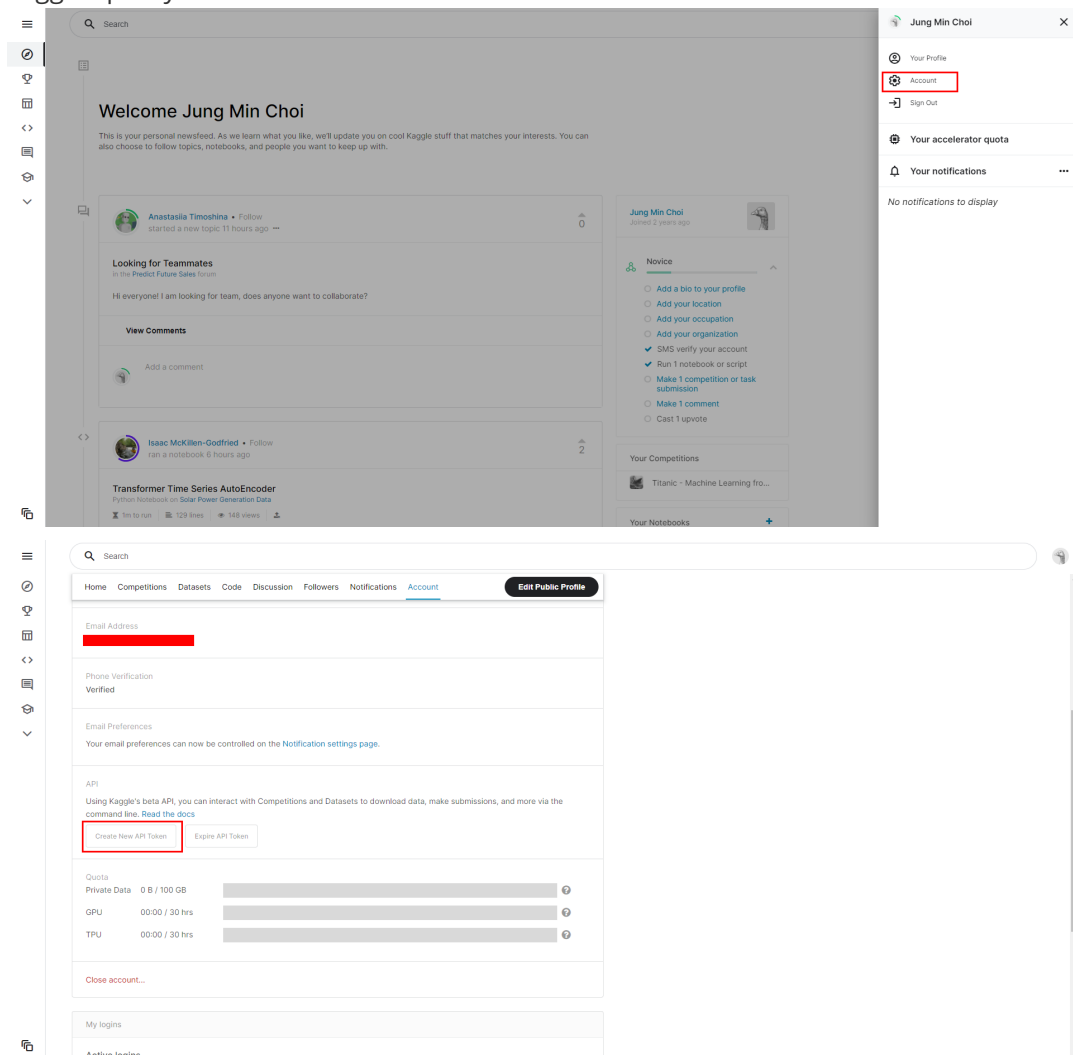
### 1. 데이터 선정 (Kaggle)

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
```

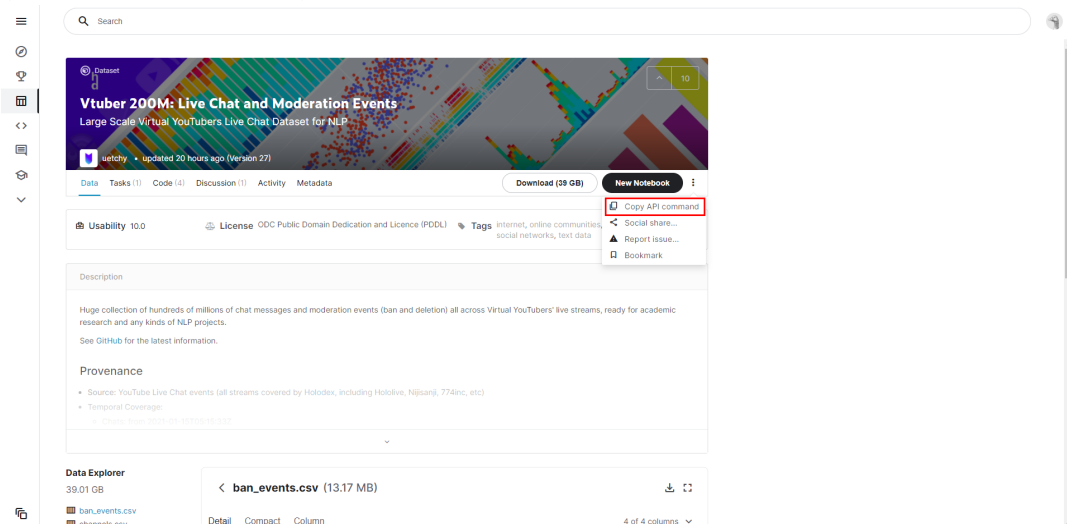
```
{KAGGLE API COMMAND} -p download_data # kaggle 명령어 실행해서 다운로드
```

```
mkdir -p datas # data_dir 폴더 생성
```

```
unzip download_data/{download zip file} -d datas/{data dir}
```

```
rm -rf download_data # download_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 소요)
```

```
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}
```

```
21/06/10 07:01:26 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
21/06/10 07:01:26 INFO impl.YarnClientImpl: Submitted application application_1623291459458_0005
21/06/10 07:01:26 INFO mapreduce.Job: The url to track the job: http://cdp-manager:8088/proxy/application_1623291459458_0005
21/06/10 07:01:26 INFO tools.DistCp: DistCp job-id: job_1623291459458_0005
21/06/10 07:01:26 INFO mapreduce.Job: Running job: job_1623291459458_0005
21/06/10 07:01:34 INFO mapreduce.Job: Job job_1623291459458_0005 running in uber mode : false
21/06/10 07:01:34 INFO mapreduce.Job: map 0% reduce 0%
21/06/10 07:01:51 INFO mapreduce.Job: map 100% reduce 0%
21/06/10 07:11:40 INFO mapreduce.Job: Job job_1623291459458_0005 completed successfully
21/06/10 07:11:40 INFO mapreduce.Job: Counters: 37
File System Counters
  FILE: Number of bytes read=0
  FILE: Number of bytes written=244678
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=4627111500
  HDFS: Number of bytes written=4627111124
  HDFS: Number of read operations=14
  HDFS: Number of large read operations=0
  HDFS: Number of write operations=5
  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)=603697
  Total time spent by all reduces in occupied slots (ms)=0
  Total time spent by all map tasks (ms)=603697
  Total vcore-milliseconds taken by all map tasks=603697
  Total megabyte-milliseconds taken by all map tasks=618185728
Map-Reduce Framework
  Map input records=1
  Map output records=0
  Input split bytes=115
  Spilled Records=0
  Failed Shuffles=0
  Merged Map outputs=0
  GC time elapsed (ms)=2580
  CPU time spent (ms)=184910
  Physical memory (bytes) snapshot=283209728
  Virtual memory (bytes) snapshot=2606091456
  Total committed heap usage (bytes)=124780544
  Peak Map Physical memory (bytes)=458637312
  Peak Map Virtual memory (bytes)=2606091456
File Input Format Counters
  Bytes Read=261
File Output Format Counters
  Bytes Written=0
DistCp Counters
  Bandwidth in Bbytes=7724763
  Bytes Copied=4627111124
  Bytes Expected=4627111124
  Files Copied=1
[ec2-user@ip-10-0-1-235 ~]$
```

```
[hadoop@ip-10-0-1-79 ~]$ sudo -u hdfs hadoop fs -ls /
Found 4 items
drwxr-xr-x - hdfs hadoop 0 2021-06-10 03:45 /apps
drwxrwxrwt - hdfs hadoop 0 2021-06-10 03:46 /tmp
drwxr-xr-x - hdfs hadoop 0 2021-06-10 03:45 /user
drwxr-xr-x - hdfs hadoop 0 2021-06-10 03:45 /var
[hadoop@ip-10-0-1-79 ~]$ sudo -u hdfs hadoop fs -ls /data
Found 1 items
-rw-r--r-- 3 hdfs hadoop 4627111124 2021-06-10 07:11 /data/allposts.csv
[hadoop@ip-10-0-1-79 ~]$
```

## 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}
```

```
21/06/10 07:42:29 INFO mapreduce.Job: Job job_1623291459458_0006 completed successfully
21/06/10 07:42:29 INFO mapreduce.Job: Counters: 43
File System Counters
  FILE: Number of bytes read=0
  FILE: Number of bytes written=490862
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=46271112023
  HDFS: Number of bytes written=0
  HDFS: Number of read operations=20
  HDFS: Number of large read operations=0
  HDFS: Number of write operations=4
  HDFS: Number of bytes read erasure-coded=0
  S3A: Number of bytes read=0
  S3A: Number of bytes written=4627111124
  S3A: Number of read operations=14
  S3A: Number of large read operations=0
  S3A: Number of write operations=2077
Job Counters
  Launched map tasks=2
  Other local map tasks=2
  Total time spent by all maps in occupied slots (ms)=662503
  Total time spent by all reduces in occupied slots (ms)=0
  Total time spent by all map tasks (ms)=662503
  Total vcore-milliseconds taken by all map tasks=662503
  Total megabyte-milliseconds taken by all map tasks=678493072
Map-Reduce Framework
  Map input records=2
  Map output records=0
  Input split bytes=228
  Spilled Records=0
  Failed Shuffles=0
  Merged Map outputs=0
  GC time elapsed (ms)=3841
  CPU time spent (ms)=807880
  Physical memory (bytes) snapshot=758522368
  Virtual memory (bytes) snapshot=579928832
  Total committed heap usage (bytes)=414187520
  Peak Map Physical memory (bytes)=461193216
  Peak Map Virtual memory (bytes)=2967109632
File Input Format Counters
  Bytes Read=671
File Output Format Counters
  Bytes Written=0
DistCp Counters
  Bandwidth in Bbytes=71076975
  Bytes Copied=4627111124
  Bytes Expected=4627111124
  Files Copied=1
  DIR_Copied=1
21/06/10 07:42:29 INFO impl.MetricsSystemImpl: Stopping s3a-file-system metrics system...
21/06/10 07:42:29 INFO impl.MetricsSystemImpl: s3a-file-system metrics system stopped.
21/06/10 07:42:29 INFO impl.MetricsSystemImpl: s3a-file-system metrics system shutdown complete.
[ec2-user@ip-10-0-1-235 ~]$
```

```
[hadoop@ip-10-0-1-79 ~]$ aws s3 ls s3://cjm-oregon/emr/data/cdp/
[hadoop@ip-10-0-1-79 ~]$ aws s3 ls s3://cjm-oregon/emr/data/cdp/
2021-06-10 07:42:02 4627111124 allposts.csv
[hadoop@ip-10-0-1-79 ~]$
```

Amazon S3 > cjm-oregon > emr/ > data/ > cdp/

cdp/ S3 URI 복사

객체 속성

객체 (1)

객체는 Amazon S3에 저장되어 있는 기본 엔티티입니다. [Amazon S3 인벤토리](#)를 사용하여 버킷에 있는 모든 객체의 목록을 얻을 수 있습니다. 다른 사용자가 객체에 액세스할 수 있게 하려면 명시적으로 권한을 부여해야 합니다. [자세히 알아보기](#)

🔄 S3 URI 복사 📄 URL 복사 📄 다운로드 🔗 열기 🗑️ 삭제 📄 작업 📄 폴더 만들기 📄 업로드

🔍 접두사로 객체 찾기

<input type="checkbox"/>	이름	유형	마지막 수정	크기	스토리지 클래스
<input type="checkbox"/>	allposts.csv	csv	2021. 6. 10. pm 4:42:02 PM KST	43.1GB	Standard

## # S3 to EMR 이관

s3-dist-cp --src=s3a://{S3\_Bucket}/{path} --dest=hdfs://{target\_dir}

```
Launched map tasks=1
Launched reduce tasks=5
Back-local map tasks=1
Total time spent by all maps in occupied slots (ms)=203424
Total time spent by all reduces in occupied slots (ms)=168308544
Total time spent by all map tasks (ms)=2110
Total time spent by all reduce tasks (ms)=876607
Total vcore-milliseconds taken by all map tasks=2110
Total vcore-milliseconds taken by all reduce tasks=876607
Total megabyte-milliseconds taken by all map tasks=6599568
Total megabyte-milliseconds taken by all reduce tasks=5385873408

Map-Reduce Framework
  Map input records=1
  Map output records=1
  Map output bytes=114
  Map output materialized bytes=150
  Input split bytes=165
  Combine input records=0
  Combine output records=0
  Reduce input groups=1
  Reduce shuffle bytes=158
  Reduce input records=1
  Reduce output records=0
  Spilled Records=2
  Shuffled Maps =5
  Failed Shuffles=0
  Merged Map outputs=5
  GC time elapsed (ms)=3652
  CPU time spent (ms)=423720
  Physical memory (bytes) snapshot=2020375488
  Virtual memory (bytes) snapshot=39859728384
  Total committed heap usage (bytes)=1514143744
  Peak Map Physical memory (bytes)=468959232
  Peak Map Virtual memory (bytes)=429455616
  Peak Reduce Physical memory (bytes)=641798144
  Peak Reduce Virtual memory (bytes)=7127773184

Shuffle Errors
  BAD_ID=0
  CONNECTION=0
  IO_ERROR=0
  WRONG_LENGTH=0
  WRONG_MAP=0
  WRONG_REDUCE=0

File Input Format Counters
  Bytes Read=209
File Output Format Counters
  Bytes Written=0
2021-06-10 08:25:31.136 INFO s3distcp.S3DistCp: Try to recursively delete hdfs://tmp/380a869a-ad19-4de8-bda9-8cafeb54912
2021-06-10 08:25:31.171 INFO impl.MetricsSystemImpl: Stopping s3a-file-system metrics system...
2021-06-10 08:25:31.171 INFO impl.MetricsSystemImpl: s3a-file-system metrics system stopped.
2021-06-10 08:25:31.171 INFO impl.MetricsSystemImpl: s3a-file-system metrics system shutdown complete.
[hadoop@ip-10-0-1-79 ~]$ sudo -u hdfs hadoop fs -ls /s3_data
Found 1 items
-rw-r--r-- 1 hadoop hadoop 4627111124 2021-06-10 08:25 /s3_data/allposts.csv
[hadoop@ip-10-0-1-79 ~]$
```

## 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	<pre>sudo -u hdfs hadoop distcp -Dfs.s3a.access.key={Access_Key} -Dfs.s3a.secret.key={Secret_Key} s3a://{Bucket_Name}/{Path} {Target_Dir}</pre>	<pre>s3-dist-cp -- src=s3a://{Bucket_Name}/{Path} --dest=hdfs://{Target_Dir}</pre>
수행시간 (ms)	908182	907408

## 2. 결과

약 15분 정도로 비슷한 결과값 도출

<pre>21/06/15 02:36:17 INFO mapreduce.Job: Job job_162372273859_0001 completed successfully 21/06/15 02:36:17 INFO mapreduce.Job: Counters: 43 File System Counters   FILE: Number of bytes read=0   FILE: Number of bytes written=49068   FILE: Number of read operations=0   FILE: Number of large read operations=0   FILE: Number of write operations=0   HDFS: Number of bytes read=871   HDFS: Number of bytes written=4627111124   HDFS: Number of read operations=22   HDFS: Number of large read operations=0   HDFS: Number of write operations=0   HDFS: Number of bytes read &lt;resource&gt;=0   S3A: Number of bytes read=4627111124   S3A: Number of bytes written=0   S3A: Number of read operations=3   S3A: Number of large read operations=0   S3A: Number of write operations=0 Job Counters   Launched map tasks=2   Other local map tasks=2   Total time spent by all maps in occupied slots (ms)=908182   Total time spent by all reduces in occupied slots (ms)=0   Total time spent by all map tasks (ms)=908182   Total vcore-milliseconds taken by all map tasks=908182   Total megabyte-milliseconds taken by all map tasks=92978368 Map-Reduce Framework   Map input records=2   Map output records=0   Input split bytes=226   Spilled Records=0   Failed Shuffles=0   Merged Map outputs=0   GC time elapsed (ms)=5801   CPU time spent (ms)=600670   Physical memory (bytes) snapshot=32880640   Virtual memory (bytes) snapshot=566525128   Total committed heap usage (bytes)=671088640   Peak Map Physical memory (bytes)=569026384   Peak Map Virtual memory (bytes)=2840809472 File Input Format Counters   Bytes Read=645 File Output Format Counters   Bytes Written=0 DistCp Counters   Bandwidth in Bytes=51815354   Bytes Copied=4627111124   Bytes Expected=4627111124   Files Copied=1   DIR Copied=1 21/06/15 02:36:17 INFO impl.MetricsSystemImpl: Stopping s3a-file-system metrics system... 21/06/15 02:36:17 INFO impl.MetricsSystemImpl: s3a-file-system metrics system stopped. 21/06/15 02:36:17 INFO impl.MetricsSystemImpl: s3a-file-system metrics system shutdown complete. [ec2-user@ip-10-0-1-235 ~]\$</pre>	<pre> S3A: Number of read operations=1 S3A: Number of large read operations=0 S3A: Number of write operations=0 Job Counters   Launched map tasks=1   Launched reduce tasks=5   Data-local map tasks=1   Total time spent by all maps in occupied slots (ms)=437856   Total time spent by all map tasks (ms)=4591   Total time spent by all reduce tasks (ms)=907408   Total vcore-milliseconds taken by all map tasks=4561   Total vcore-milliseconds taken by all reduce tasks=907408   Total megabyte-milliseconds taken by all map tasks=14011392   Total megabyte-milliseconds taken by all reduce tasks=5575114752 Map-Reduce Framework   Map input records=1   Map output records=1   Map output bytes=114   Map output materialized bytes=158   Input split bytes=166   Combine input records=0   Combine output records=0   Reduce input groups=1   Reduce shuffle bytes=158   Reduce input records=1   Reduce output records=0   Spilled Records=2   Shuffled Maps=5   Failed Shuffles=0   Merged Map outputs=5   GC time elapsed (ms)=3707   CPU time spent (ms)=429840   Physical memory (bytes) snapshot=1874915328   Virtual memory (bytes) snapshot=39863812096   Total committed heap usage (bytes)=1488453632   Peak Map Physical memory (bytes)=426950656   Peak Map Virtual memory (bytes)=1394057728   Peak Reduce Physical memory (bytes)=550714880   Peak Reduce Virtual memory (bytes)=7115907072 Shuffle Errors   BAD_ID=0   CONNECTION=0   IO_ERROR=0   WRONG_LENGTH=0   WRONG_MAP=0   WRONG_REDUCE=0 File Input Format Counters   Bytes Read=209 File Output Format Counters   Bytes Written=0 2021-06-15 02:36:18.035 INFO s3distcp.S3DistCp: Try to recursively delete hdfs:/tmp/2e4a89d7-6355-49 7d-ad23-955f15e38cbe 2021-06-15 02:36:18.048 INFO impl.MetricsSystemImpl: Stopping s3a-file-system metrics system... 2021-06-15 02:36:18.048 INFO impl.MetricsSystemImpl: s3a-file-system metrics system stopped.</pre>
--	---