

Final Project

Graph mining on MapReduce and Spark

국민대학교 컴퓨터공학부

20143109

최지욱

빅데이터 최신 기술

프로젝트 개요

- 프로젝트 목표: 대용량 그래프의 군집계수(Clustering Coefficient) 분석을 위한 효율적인 분산 알고리즘을 설계, 구현 및 실험한다.
- 프로젝트는 다음과 같은 세부 과업으로 구성된다.
 - Task 1) 그래프의 중복 edge 및 self-loop 제거 (Hadoop)
 - Task 2) 각 node u 의 degree $d(u)$ 구하기 (Hadoop)
 - Task 3) 각 node u마다, u를 포함하는 삼각형 수 $t(u)$ 구하기 (Spark)
 - Task 4) 각 node u마다, 군집계수 $cc(u)$ 구하기 (Spark)

구현한 프로그램들을 구글 클라우드 플랫폼에서 실행

- LiveJournal Dataset 을 HDFS 에 업로드 하는 과정과 결과를 캡처하여 첨부
 - LiveJournal Dataset 다운로드 경로
<http://snap.stanford.edu/data/soc-LiveJournal1.html>
- 각 Task 의 프로그램을 실행하는 과정과 결과를 캡처하여 첨부
 - 과정에는 실행 명령어, 실행 중 및 실행 후의 Web UI 포함
 - 결과에는 결과파일 경로, 목록 및 결과파일 내용 일부 포함

soc-LiveJournal1.txt 파일에 edge list 정보를 제외하고는 삭제를 해주었다. <그림 1>처럼 문자들이 들어있어서 <그림 2>로 파일의 내용을 일부 수정하였다.

```
File Edit View Search Terminal Help
sexy-ji@sexy-ji:~/BigData$ head soc-LiveJournal1.txt
# Directed graph (each unordered pair of nodes is saved once): soc-LiveJournal1.
txt
# Directed LiveJournal friendship social network
# Nodes: 4847571 Edges: 68993773
# FromNodeId      ToNodeId
0      1
0      2
0      3
0      4
0      5
0      6
sexy-ji@sexy-ji:~/BigData$
```

<그림 1>

```
File Edit View Search Terminal Help
sexy-ji@sexy-ji:~/BigData/soc-LiveJournal$ head soc-LiveJournal1.txt
0      1
0      2
0      3
0      4
0      5
0      6
0      7
0      8
0      9
0      10
sexy-ji@sexy-ji:~/BigData/soc-LiveJournal$
```

<그림 2>

soc-LiveJournal1.txt 파일을 구글 클라우드 플랫폼(이하 GCP)에 업로드 하는 과정이다.

scp soc-LiveJournal1.txt kcoma2623@35.190.230.101:~/ 명령어를 사용하여 GCP 의 클러스터에 업로드 하였다.

```
File Edit View Search Terminal Help
sexy-ji@sexy-ji:~/BigData/soc-LiveJournal$ scp soc-LiveJournal1.txt    kcoma2623@35.190.230.101:~/
kcoma2623@35.190.230.101's password: [REDACTED]
```

<그림 3>

<그림 4>는 업로드하는 중간 과정이고 <그림 5>는 업로드를 완료한 그림이다.

```
File Edit View Search Terminal Help
sexy-ji@sexy-ji:~/BigData/soc-LiveJournal$ scp soc-LiveJournal1.txt    kcoma2623@35.190.230.101:~/
kcoma2623@35.190.230.101's password:
soc-LiveJournal1.txt                                13%  111MB  11.2MB/s   01:02 ETA[REDACTED]
```

<그림 4>

```
File Edit View Search Terminal Help
sexy-ji@sexy-ji:~/BigData/soc-LiveJournal$ scp soc-LiveJournal1.txt kkoma2623@35.190.230.101:~/
kkoma2623@35.190.230.101's password:
soc-LiveJournal1.txt                                              100%   815MB  11.4MB/s   01:11
sexy-ji@sexy-ji:~/BigData/soc-LiveJournal$
```

<그림 5>

클러스터의 ip 를 입력하고 비밀번호를 입력하여 클러스터에 ssh 접속하였다.

그리고 ls 명령어를 사용하여 soc-LiveJournal1.txt 파일이 업로드 된 것을 확인하였다.

```
File Edit View Search Terminal Help
sexy-ji@sexy-ji:~/BigData/soc-LiveJournal$ scp soc-LiveJournal1.txt kkoma2623@35.190.230.101:~/
kkoma2623@35.190.230.101's password:
soc-LiveJournal1.txt                                              100%   815MB  11.4MB/s   01:11
sexy-ji@sexy-ji:~/BigData/soc-LiveJournal$ ssh kkoma2623@35.190.230.101
kkoma2623@35.190.230.101's password:
Linux kmu-cluster-m 4.19.0-0.bpo.8-amd64 #1 SMP Debian 4.19.98-1~bpo9+1 (2020-03-09) x86_64

The programs included with the Debian GNU/Linux system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*copyright.

Debian GNU/Linux comes with ABSOLUTELY NO WARRANTY, to the extent
permitted by applicable law.
Last login: Fri Jun 26 04:32:06 2020 from 1.209.175.113
kkoma2623@kmu-cluster-m:~$ ls
soc-LiveJournal1.txt
kkoma2623@kmu-cluster-m:~$
```

<그림 6>

HDFS에 soc-LiveJournal1.txt 파일을 업로드하였다.

hdfs dfs -put soc-LiveJournal1.txt 명령어를 사용하여 업로드하였고, hdfs dfs -ls 명령어를 사용하여 HDFS에 업로드 한 것을 확인하였다.

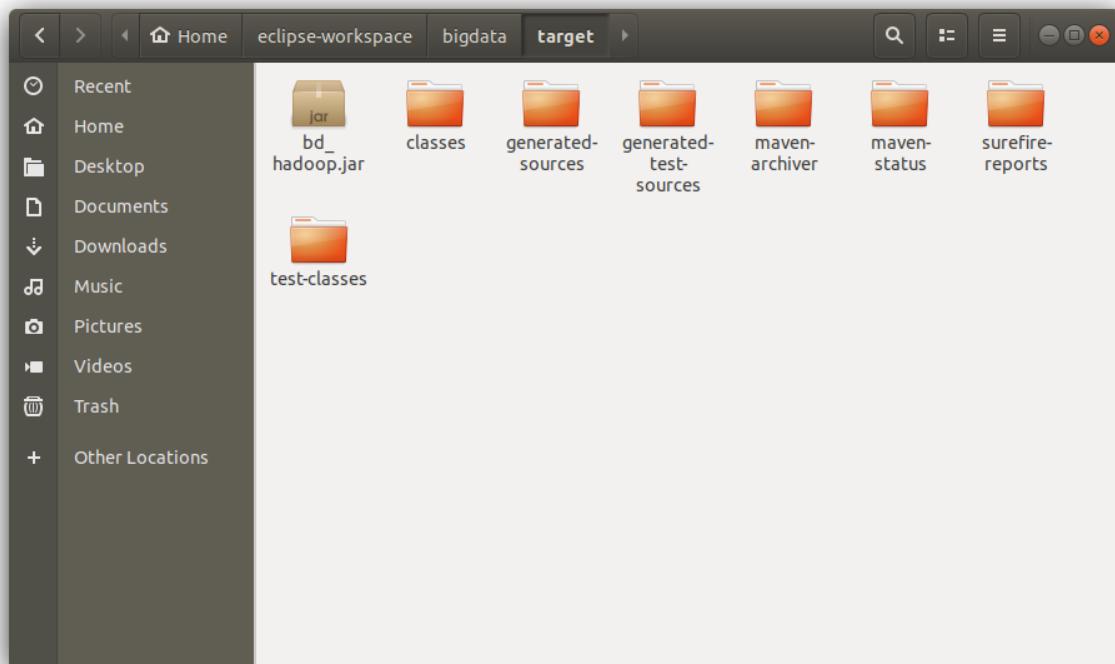
```
File Edit View Search Terminal Help
sexy-ji@sexy-ji:~/BigData/soc-LiveJournal$ scp soc-LiveJournal1.txt kkoma2623@35.190.230.101:~/kkoma2623@35.190.230.101's password:
soc-LiveJournal1.txt                                         100%   815MB  11.4MB/s   01:11
sexy-ji@sexy-ji:~/BigData/soc-LiveJournal$ ssh kkoma2623@35.190.230.101
kkoma2623@35.190.230.101's password:
Linux kmu-cluster-m 4.19.0-0.bpo.8-amd64 #1 SMP Debian 4.19.98-1~bpo9+1 (2020-03-09) x86_64

The programs included with the Debian GNU/Linux system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*copyright.

Debian GNU/Linux comes with ABSOLUTELY NO WARRANTY, to the extent
permitted by applicable law.
Last login: Fri Jun 26 04:32:06 2020 from 1.209.175.113
kkoma2623@kmu-cluster-m:~$ ls
soc-LiveJournal1.txt
kkoma2623@kmu-cluster-m:~$ hdfs dfs -put soc-LiveJournal1.txt
kkoma2623@kmu-cluster-m:~$ hdfs dfs -ls
Found 2 items
drwxr-xr-x  - kkoma2623 hadoop          0 2020-06-25 15:38 .sparkStaging
-rw-r--r--  2 kkoma2623 hadoop  854597439 2020-06-26 04:39 soc-LiveJournal1.txt
kkoma2623@kmu-cluster-m:~$
```

<그림 7>

Task1과 Task2를 구현한 프로그램이 있는 bd_hadoop.jar 파일을 생성하였다.



<그림 8>

bd_hadoop.jar 파일을 scp 명령어를 통해서 GCP에 업로드하였다.

```
File Edit View Search Terminal Help
sexy-ji@sexy-ji:~/eclipse-workspace/bigdata/target$ scp bd_hadoop.jar kkoma2623@35.190.230.101:~/
kkoma2623@35.190.230.101's password:
bd_hadoop.jar                                         100% 8031    207.9KB/s   00:00
sexy-ji@sexy-ji:~/eclipse-workspace/bigdata/target$
```

<그림 9>

클러스터에서 ls 명령어를 사용하여 업로드 된 것을 확인하였다.

```
File Edit View Search Terminal Help
sexy-ji@sexy-ji:~/BigData/soc-LiveJournal$ scp soc-LiveJournal1.txt kkoma2623@35.190.230.101:~/
kkoma2623@35.190.230.101's password:
soc-LiveJournal1.txt                                         100% 815MB  11.4MB/s   01:11
sexy-ji@sexy-ji:~/BigData/soc-LiveJournal$ ssh kkoma2623@35.190.230.101
kkoma2623@35.190.230.101's password:
Linux kmu-cluster-m 4.19.0-0.bpo.8-amd64 #1 SMP Debian 4.19.98-1~bpo9+1 (2020-03-09) x86_64

The programs included with the Debian GNU/Linux system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*/*copyright.

Debian GNU/Linux comes with ABSOLUTELY NO WARRANTY, to the extent
permitted by applicable law.
Last login: Fri Jun 26 04:32:06 2020 from 1.209.175.113
kkoma2623@kmu-cluster-m:~$ ls
soc-LiveJournal1.txt
kkoma2623@kmu-cluster-m:~$ hdfs dfs -put soc-LiveJournal1.txt
kkoma2623@kmu-cluster-m:~$ hdfs dfs -ls
Found 2 items
drwxr-xr-x  - kkoma2623 hadoop          0 2020-06-25 15:38 .sparkStaging
-rw-r--r--  2 kkoma2623 hadoop  854597439 2020-06-26 04:39 soc-LiveJournal1.txt
kkoma2623@kmu-cluster-m:~$ ls
bd_hadoop.jar  soc-LiveJournal1.txt
kkoma2623@kmu-cluster-m:~$
```

<그림 10>

bd_hadoop.jar 파일을 HDFS에 업로드하였다.

hdfs dfs -put bd_hadoop.jar 명령어를 통해 업로드하였고, hdfs dfs -ls 명령어를 통해 업로드된 것을 확인하였다.

```
File Edit View Search Terminal Help

The programs included with the Debian GNU/Linux system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*copyright.

Debian GNU/Linux comes with ABSOLUTELY NO WARRANTY, to the extent
permitted by applicable law.
Last login: Fri Jun 26 04:32:06 2020 from 1.209.175.113
kkoma2623@kmu-cluster-m:~$ ls
soc-LiveJournal1.txt
kkoma2623@kmu-cluster-m:~$ hdfs dfs -put soc-LiveJournal1.txt
kkoma2623@kmu-cluster-m:~$ hdfs dfs -ls
Found 2 items
drwxr-xr-x - kkoma2623 hadoop 0 2020-06-25 15:38 .sparkStaging
-rw-r--r-- 2 kkoma2623 hadoop 854597439 2020-06-26 04:39 soc-LiveJournal1.txt
kkoma2623@kmu-cluster-m:~$ ls
bd_hadoop.jar soc-LiveJournal1.txt
kkoma2623@kmu-cluster-m:~$ hdfs dfs -put bd_hadoop.jar
kkoma2623@kmu-cluster-m:~$ hdfs dfs -ls
Found 3 items
drwxr-xr-x - kkoma2623 hadoop 0 2020-06-25 15:38 .sparkStaging
-rw-r--r-- 2 kkoma2623 hadoop 8031 2020-06-26 04:43 bd_hadoop.jar
-rw-r--r-- 2 kkoma2623 hadoop 854597439 2020-06-26 04:39 soc-LiveJournal1.txt
kkoma2623@kmu-cluster-m:~$
```

<그림 11>

hadoop jar bd_hadoop.jar bigdata.Task1 -Dmapreduce.job.reduces=5 soc-LiveJournal1.txt task1Outputs 명령어를 통해 bigdata 패키지의 Task1을 실행하였다.

-Dmapreduce.job.reduces=5 옵션을 통해 리듀서의 개수를 5로 설정하였다.

args[0]에 입력파일 경로를, args[1]에 출력 파일 경로를 설정해주었다.

입력 파일 경로는 soc-LiveJournal1.txt이고 출력 파일 경로는 task1Outputs이다.

```
File Edit View Search Terminal Help
kkoma2623@kmu-cluster-m:~$ hadoop jar bd_hadoop.jar bigdata.Task1 -Dmapreduce.job.reduces=5 soc-LiveJournal1.txt task1Outputs
20/06/26 04:54:12 INFO client.RMProxy: Connecting to ResourceManager at kmu-cluster-m/10.146.0.9:8032
20/06/26 04:54:12 INFO client.AHSProxy: Connecting to Application History server at kmu-cluster-m/10.146.0.9:10200
20/06/26 04:54:12 INFO input.FileInputFormat: Total input files to process : 1
20/06/26 04:54:12 INFO mapreduce.JobSubmitter: number of splits:7
20/06/26 04:54:13 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled
is deprecated. Instead, use yarn.system-metrics-publisher.enabled
20/06/26 04:54:13 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1593072483341_0014
20/06/26 04:54:13 INFO impl.YarnClientImpl: Submitted application application_1593072483341_0014
20/06/26 04:54:13 INFO mapreduce.Job: The url to track the job: http://kmu-cluster-m:8088/proxy/application_1593072483341_0014/
20/06/26 04:54:13 INFO mapreduce.Job: Running job: job_1593072483341_0014
```

<그림 12>

<그림 13>은 터미널에서 Task1 을 실행하는 과정을 본 그림이다.

```
File Edit View Search Terminal Help
20/06/26 04:54:12 INFO input.FileInputFormat: Total input files to process : 1
20/06/26 04:54:12 INFO mapreduce.JobSubmitter: number of splits:7
20/06/26 04:54:13 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled
20/06/26 04:54:13 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1593072483341_0014
20/06/26 04:54:13 INFO impl.YarnClientImpl: Submitted application application_1593072483341_0014
20/06/26 04:54:13 INFO mapreduce.Job: The url to track the job: http://kmu-cluster-m:8088/proxy/application_1593072483341_0014/
20/06/26 04:54:13 INFO mapreduce.Job: Running job: job_1593072483341_0014
20/06/26 04:54:20 INFO mapreduce.Job: Job job_1593072483341_0014 running in uber mode : false
20/06/26 04:54:20 INFO mapreduce.Job: map 0% reduce 0%
20/06/26 04:54:33 INFO mapreduce.Job: map 14% reduce 0%
20/06/26 04:54:42 INFO mapreduce.Job: map 25% reduce 0%
20/06/26 04:54:43 INFO mapreduce.Job: map 47% reduce 0%
20/06/26 04:54:48 INFO mapreduce.Job: map 51% reduce 0%
20/06/26 04:54:49 INFO mapreduce.Job: map 69% reduce 0%
20/06/26 04:54:54 INFO mapreduce.Job: map 70% reduce 0%
20/06/26 04:54:55 INFO mapreduce.Job: map 71% reduce 0%
20/06/26 04:55:14 INFO mapreduce.Job: map 76% reduce 0%
20/06/26 04:55:22 INFO mapreduce.Job: map 81% reduce 0%
20/06/26 04:55:23 INFO mapreduce.Job: map 91% reduce 0%
20/06/26 04:55:25 INFO mapreduce.Job: map 95% reduce 0%
20/06/26 04:55:26 INFO mapreduce.Job: map 100% reduce 0%
```

<그림 13>

<그림 14>, <그림 15>, <그림 16>는 web UI 를 통해 Task1 이 실행중인 것을 캡처한 그림이다.

The screenshot shows the Google Cloud DataProc Cluster details page for cluster 'bigdata-290408'. The main table displays various tasks and their execution status. Task1, with ID 'application_1593072483341_0014', is listed as 'RUNNING' with a progress of 88.9%. Other tasks listed include Task3, Task3, Task3, Task3, Task3, Task3, Spark shell, Spark shell, Spark shell, and Task1 again, all in a 'FINISHED' state with 0.0% progress. The table includes columns for ID, User, Name, Application Type, Queue, Application Priority, Start Time, Finish Time, State, Final Status, Running Containers, Allocated CPU VCores, Allocated Memory MB, Reserved CPU VCores, Reserved Memory MB, % of Queue, % of Cluster, Progress, Tracking UI, and Blacklisted Nodes.

<그림 14>

MapReduce Application application_1593072483341_0014

Active Jobs

Job ID	Name	State	Map Progress	Maps Total	Maps Completed	Reduce Progress	Reduces Total	Reduces Completed
job_1593072483341_0014	bd_hadoop.jar	RUNNING	[Progress Bar]	7	1	[Progress Bar]	5	0

Showing 1 to 1 of 1 entries

<그림 15>

MapReduce Job job_1593072483341_0014

Job Overview

Job Name:	bd_hadoop.jar
User Name:	ikoma2623
Queue Name:	default
State:	RUNNING
Uberized:	false
Started:	Fri Jun 26 04:54:18 UTC 2020
Elapsed:	50sec

ApplicationMaster

Attempt Number	Start Time	Node	Logs
1	Fri Jun 26 04:54:14 UTC 2020	kmu-cluster-w-0.asia-northeast1-a.c.bigdata-280408.internal:8042	/gateway/default/yarn/logs

Task Type

Map	Reduce	New	Running	Failed	Killed	Successful
0	5	8	0	0	1	0

<그림 16>

<그림 17>은 터미널에서 Task1 이 실행 된 후의 그림이다.

```
File Edit View Search Terminal Help
    Reduce shuffle bytes=554559050
    Reduce input records=55455884
    Reduce output records=36767837
    Spilled Records=120345478
    Shuffled Maps =35
    Failed Shuffles=0
    Merged Map outputs=35
    GC time elapsed (ms)=4714
    CPU time spent (ms)=322390
    Physical memory (bytes) snapshot=12116512768
    Virtual memory (bytes) snapshot=63408820224
    Total committed heap usage (bytes)=11914969088
    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=854622015
    File Output Format Counters
        Bytes Written=531608189
kkoma2623@kmu-cluster-m:~$
```

<그림 17>

hdfs dfs -ls 명령어를 통해 HDFS에 출력 파일 경로에 설정한 출력파일이 생성된 것을 확인하였다.

```
File Edit View Search Terminal Help
    Merged Map outputs=35
    GC time elapsed (ms)=4714
    CPU time spent (ms)=322390
    Physical memory (bytes) snapshot=12116512768
    Virtual memory (bytes) snapshot=63408820224
    Total committed heap usage (bytes)=11914969088
    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=854622015
    File Output Format Counters
        Bytes Written=531608189
kkoma2623@kmu-cluster-m:~$ hdfs dfs -ls
Found 4 items
drwxr-xr-x  - kkoma2623 hadoop          0 2020-06-25 15:38 .sparkStaging
-rw-r--r--  2 kkoma2623 hadoop     8031 2020-06-26 04:43 bd_hadoop.jar
-rw-r--r--  2 kkoma2623 hadoop  854597439 2020-06-26 04:39 soc-LiveJournal1.txt
drwxr-xr-x  - kkoma2623 hadoop          0 2020-06-26 04:55 task1outputs
kkoma2623@kmu-cluster-m:~$
```

<그림 18>

hdfs dfs -ls task1Outputs 명령어를 통해 task1Outputs 디렉터리에 생성된 결과 파일들을 확인하였다.

```
File Edit View Search Terminal Help
    CONNECTION=0
    IO_ERROR=0
    WRONG_LENGTH=0
    WRONG_MAP=0
    WRONG_REDUCE=0
    File Input Format Counters
        Bytes Read=854622015
    File Output Format Counters
        Bytes Written=531608189
kkoma2623@kmu-cluster-m:~$ hdfs dfs -ls
Found 4 items
drwxr-xr-x  - kkoma2623 hadoop      0 2020-06-25 15:38 .sparkStaging
-rw-r--r--  2 kkoma2623 hadoop     8031 2020-06-26 04:43 bd_hadoop.jar
-rw-r--r--  2 kkoma2623 hadoop  854597439 2020-06-26 04:39 soc-LiveJournal1.txt
drwxr-xr-x  - kkoma2623 hadoop      0 2020-06-26 04:55 task1Outputs
kkoma2623@kmu-cluster-m:~$ hdfs dfs -ls task1Outputs
Found 6 items
-rw-r--r--  2 kkoma2623 hadoop      0 2020-06-26 04:55 task1Outputs/_SUCCESS
-rw-r--r--  2 kkoma2623 hadoop  105746210 2020-06-26 04:55 task1Outputs/part-r-00000
-rw-r--r--  2 kkoma2623 hadoop  106476096 2020-06-26 04:55 task1Outputs/part-r-00001
-rw-r--r--  2 kkoma2623 hadoop  105959403 2020-06-26 04:55 task1Outputs/part-r-00002
-rw-r--r--  2 kkoma2623 hadoop  106038438 2020-06-26 04:55 task1Outputs/part-r-00003
-rw-r--r--  2 kkoma2623 hadoop  107388042 2020-06-26 04:55 task1Outputs/part-r-00004
kkoma2623@kmu-cluster-m:~$
```

<그림 19>

hdfs dfs -cat task1Outputs/part-r-00004 명령어를 통해 파일의 내용을 출력하겠다.

```
File Edit View Search Terminal Help
    CONNECTION=0
    IO_ERROR=0
    WRONG_LENGTH=0
    WRONG_MAP=0
    WRONG_REDUCE=0
    File Input Format Counters
        Bytes Read=854622015
    File Output Format Counters
        Bytes Written=531608189
kkoma2623@kmu-cluster-m:~$ hdfs dfs -ls
Found 4 items
drwxr-xr-x  - kkoma2623 hadoop      0 2020-06-25 15:38 .sparkStaging
-rw-r--r--  2 kkoma2623 hadoop     8031 2020-06-26 04:43 bd_hadoop.jar
-rw-r--r--  2 kkoma2623 hadoop  854597439 2020-06-26 04:39 soc-LiveJournal1.txt
drwxr-xr-x  - kkoma2623 hadoop      0 2020-06-26 04:55 task1Outputs
kkoma2623@kmu-cluster-m:~$ hdfs dfs -ls task1Outputs
Found 6 items
-rw-r--r--  2 kkoma2623 hadoop      0 2020-06-26 04:55 task1Outputs/_SUCCESS
-rw-r--r--  2 kkoma2623 hadoop  105746210 2020-06-26 04:55 task1Outputs/part-r-00000
-rw-r--r--  2 kkoma2623 hadoop  106476096 2020-06-26 04:55 task1Outputs/part-r-00001
-rw-r--r--  2 kkoma2623 hadoop  105959403 2020-06-26 04:55 task1Outputs/part-r-00002
-rw-r--r--  2 kkoma2623 hadoop  106038438 2020-06-26 04:55 task1Outputs/part-r-00003
-rw-r--r--  2 kkoma2623 hadoop  107388042 2020-06-26 04:55 task1Outputs/part-r-00004
kkoma2623@kmu-cluster-m:~$ hdfs dfs -cat task1Outputs/part-r-00004
```

<그림 20>

<그림 21>은 출력이 너무 길어서 keyboard interrupt를 통해서 출력을 멈춘 그림이다.

simple graph를 생성하는 것을 확인하였다.

```
File Edit View Search Terminal Help
4994 954804
4994 311743
4994 954803
4994 1722791
4994 321984
4994 642508
4994 741836
4994 22983
4994 2203112
4994 1605076
4994 190931
4994 1347014
4994 300496
4994 1526210
4994 1233399
4994 403936
4994 353772
4994 1893878
4994 168431
4994 168435
4994 2086382
4994 831992
49cat: Filesystem closed
kkoma2623@kmu-cluster-m:~$
```

<그림 21>

<그림 22>와 <그림 23>은 Task1 을 수행한 후에 web UI 를 통해 확인한 그림이다.

ID	User	Name	Application Type	Queue	Application Priority	StartTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU Vcores	Allocated Memory MB	Reserved CPU Vcores	Reserved Memory MB	% of Queue	% of Cluster	Progress	Tracking UI	Blacklisted Nodes
application_1593072483341_0014	kkoma2623	bd_hadoop.jar	MAPREDUCE	default	0	Fri Jun 26 13:54:13 +0900 2020	Fri Jun 26 13:55:59 +0900 2020	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	0.0	0.0	0.0	History	0	
application_1593072483341_0013	kkoma2623	bd_hadoop.jar	MAPREDUCE	default	0	Fri Jun 26 13:47:06 +0900 2020	Fri Jun 26 13:48:56 +0900 2020	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	0.0	0.0	0.0	History	0	
application_1593072483341_0012	kkoma2623	Task3	SPARK	default	0	Fri Jun 26 00:37:47 +0900 2020	Fri Jun 26 00:38:25 +0900 2020	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	0.0	0.0	0.0	History	0	
application_1593072483341_0011	kkoma2623	Task3	SPARK	default	0	Fri Jun 26 00:33:44 +0900 2020	Fri Jun 26 00:34:42 +0900 2020	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	0.0	0.0	0.0	History	0	
application_1593072483341_0010	kkoma2623	Task3	SPARK	default	0	Fri Jun 26 00:23:57 +0900 2020	Fri Jun 26 00:24:51 +0900 2020	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	0.0	0.0	0.0	History	0	
application_1593072483341_0009	kkoma2623	Task3	SPARK	default	0	Thu Jun 25 22:50:34 +0900 2020	Thu Jun 25 23:40:43 +0900 2020	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	0.0	0.0	0.0	History	0	
application_1593072483341_0008	kkoma2623	Spark shell	SPARK	default	0	Thu Jun 25 22:07:51 +0900 2020	Thu Jun 25 22:52:14 +0900 2020	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	0.0	0.0	0.0	History	0	
application_1593072483341_0007	kkoma2623	Spark shell	SPARK	default	0	Thu Jun 25 21:59:12 +0900 2020	Thu Jun 25 22:01:01 +0900 2020	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	0.0	0.0	0.0	History	0	

<그림 22>

The screenshot shows the Google Cloud DataProc Cluster details page for a cluster named 'kmu-cluster'. A specific job, 'MapReduce Job job_1593072483341_0014', is displayed as successful ('SUCCEEDED'). The job overview table includes columns for Job Name, User Name, Queue, State, Submitted, Started, Finished, and Elapsed times, along with diagnostic metrics like Average Map Time, Average Shuffle Time, Average Merge Time, and Average Reduce Time. Below this is a detailed ApplicationMaster table showing task counts (Total, Failed, Killed, Successful) for Map and Reduce tasks across different attempt numbers.

<그림 23>

hadoop jar bd_hadoop.jar bigdata.Task2 -Dmapreduce.job.reduces=5 task1Outputs/*
task2Outputs 명령어를 통해 bigdata 패키지의 Task2를 실행하였다.

-Dmapreduce.job.reduces=5 옵션을 통해 리듀서의 개수를 5로 설정하였다.

args[0]에 입력파일 경로를, args[1]에 출력 파일 경로를 설정해주었다.

입력 파일 경로는 task1Outputs/*이고 출력 파일 경로는 task2Outputs이다.

```
File Edit View Search Terminal Help
4994    311743
4994    954803
4994    1722791
4994    321984
4994    642508
4994    741836
4994    22983
4994    2203112
4994    1605076
4994    190931
4994    1347014
4994    300496
4994    1526210
4994    1233399
4994    403936
4994    353772
4994    1893878
4994    168431
4994    168435
4994    2086382
4994    831992
49cat: Filesystem closed
kkoma2623@kmu-cluster-m:~$ hadoop jar bd_hadoop.jar bigdata.Task2 -Dmapreduce.job.reduces=5 task1Outputs/* task2Outputs
```

<그림 24>

<그림 25>는 Task2를 실행하는 그림이다.

```
File Edit View Search Terminal Help
4994 1233399
4994 403936
4994 353772
4994 1893878
4994 168431
4994 168435
4994 2086382
4994 831992
49cat: Filesystem closed
kkoma2623@kmu-cluster-m:~$ hadoop jar bd_hadoop.jar bigdata.Task2 -Dmapreduce.job.reduces=5 task1output
s/* task2outputs
20/06/26 05:15:26 INFO client.RMProxy: Connecting to ResourceManager at kmu-cluster-m/10.146.0.9:8032
20/06/26 05:15:26 INFO client.AHSProxy: Connecting to Application History server at kmu-cluster-m/10.146
.0.9:10200
20/06/26 05:15:27 INFO input.FileInputFormat: Total input files to process : 5
20/06/26 05:15:27 INFO mapreduce.JobSubmitter: number of splits:5
20/06/26 05:15:27 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled
is deprecated. Instead, use yarn.system-metrics-publisher.enabled
20/06/26 05:15:27 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1593072483341_0015
20/06/26 05:15:27 INFO impl.YarnClientImpl: Submitted application application_1593072483341_0015
20/06/26 05:15:28 INFO mapreduce.Job: The url to track the job: http://kmu-cluster-m:8088/proxy/applicat
ion_1593072483341_0015/
20/06/26 05:15:28 INFO mapreduce.Job: Running job: job_1593072483341_0015
```

<그림 25>

<그림 26>, <그림 27>, <그림 28>은 Task2를 실행중인 것을 Web UI를 통해서 보는 그림이다.

ID	User	Name	Type	Queue	Priority	Start Time	Finish Time	State	Final Status	Running Containers	Allocated CPU VCores	Allocated Memory MB	Reserved CPU VCores	Reserved Memory MB	% of Queue	% of Cluster	Progress	Tracking UI	Block UI	Application Master
application_1593072483341_0015	kkoma2623	bd_hadoop.jar	MAPREDUCE	default	0	Fri Jun 26 14:15:27 +0900 2020	N/A	RUNNING	UNDEFINED	6	6	24576	0	0	66.7	66.7	0	0	0	ApplicationMaster
application_1593072483341_0014	kkoma2623	bd_hadoop.jar	MAPREDUCE	default	0	Fri Jun 26 13:55:59 +0900 2020	N/A	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	0.0	0.0	0	0	0	History
application_1593072483341_0013	kkoma2623	bd_hadoop.jar	MAPREDUCE	default	0	Fri Jun 26 13:47:06 +0900 2020	N/A	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	0.0	0.0	0	0	0	History
application_1593072483341_0012	kkoma2623	Task3	SPARK	default	0	Fri Jun 26 00:37:47 +0900 2020	N/A	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	0.0	0.0	0	0	0	History
application_1593072483341_0011	kkoma2623	Task3	SPARK	default	0	Fri Jun 26 00:43:44 +0900 2020	N/A	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	0.0	0.0	0	0	0	History
application_1593072483341_0010	kkoma2623	Task3	SPARK	default	0	Fri Jun 26 00:23:57 +0900 2020	N/A	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	0.0	0.0	0	0	0	History
application_1593072483341_0009	kkoma2623	Task3	SPARK	default	0	Thu Jun 25 22:50:34 +0900 2020	N/A	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	0.0	0.0	0	0	0	History
application_1593072483341_0008	kkoma2623	Spark shell	SPARK	default	0	Thu Jun 25 22:07:51 +0900 2020	N/A	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	0.0	0.0	0	0	0	History

<그림 26>

The screenshot shows the Google Cloud DataProc Cluster details page for a cluster named 'kmu-cluster'. The main title is 'MapReduce Application application_1593072483341_0015'. On the left, there's a sidebar with navigation links: Cluster, Application, About jobs, and Tools. The 'Application' link is currently selected. Below the sidebar, the title 'Active Jobs' is displayed, followed by a table showing one active job entry:

Job ID	Name	State	Map Progress	Maps Total	Maps Completed	Reduce Progress	Reduces Total	Reduces Completed
job_1593072483341_0015	bd_hadoop.jar	RUNNING	[Progress Bar]	5	0	[Progress Bar]	5	0

At the bottom of the table, it says 'Showing 1 to 1 of 1 entries'.

<그림 27>

The screenshot shows the Google Cloud DataProc Cluster details page for a cluster named 'kmu-cluster'. The main title is 'MapReduce Job job_1593072483341_0015'. On the left, there's a sidebar with navigation links: Cluster, Application, Job, Overview, Counters, Configuration, Metrics, Reduce tasks, AM Logs, and Tools. The 'Job' link is currently selected. Below the sidebar, the title 'Job Overview' is displayed, followed by a table showing job details:

Job Name:	bd_hadoop.jar
User Name:	ikoma2623
Queue Name:	default
State:	RUNNING
Uberized:	false
Started:	Fri Jun 26 05:15:33 UTC 2020
Elapsed:	1mins, 0sec

Below the details table, there's a section titled 'ApplicationMaster' with a table showing task progress:

Attempt Number	Start Time	Node	Logs
1	Fri Jun 26 05:15:29 UTC 2020	kmu-cluster-w-1.asia-northeast1-a.c.bigdata-280408.internal:8042	/gateway/default/yarn/logs

Under the 'Logs' column, there are two tables: 'Map' and 'Reduce'.

Task Type	Progress	Total	Pending	Running	Complete
Map	[Progress Bar]	5	0	5	0
Reduce	[Progress Bar]	5	0	0	0

Attempt Type	New	Running	Failed	Killed	Successful
Maps	0	0	0	0	0
Reduces	5	0	0	0	0

<그림 28>

Task2를 실행 중인 것을 터미널을 통해 확인하는 그림이다.

```
File Edit View Search Terminal Help
20/06/26 05:15:27 INFO input.FileInputFormat: Total input files to process : 5
20/06/26 05:15:27 INFO mapreduce.JobSubmitter: number of splits:5
20/06/26 05:15:27 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled
is deprecated. Instead, use yarn.system-metrics-publisher.enabled
20/06/26 05:15:27 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1593072483341_0015
20/06/26 05:15:27 INFO impl.YarnClientImpl: Submitted application application_1593072483341_0015
20/06/26 05:15:28 INFO mapreduce.Job: The url to track the job: http://kmu-cluster-m:8088/proxy/application_1593072483341_0015/
20/06/26 05:15:28 INFO mapreduce.Job: Running job: job_1593072483341_0015
20/06/26 05:15:35 INFO mapreduce.Job: Job job_1593072483341_0015 running in uber mode : false
20/06/26 05:15:35 INFO mapreduce.Job: map 0% reduce 0%
20/06/26 05:15:55 INFO mapreduce.Job: map 18% reduce 0%
20/06/26 05:15:58 INFO mapreduce.Job: map 41% reduce 0%
20/06/26 05:16:01 INFO mapreduce.Job: map 42% reduce 0%
20/06/26 05:16:04 INFO mapreduce.Job: map 47% reduce 0%
20/06/26 05:16:19 INFO mapreduce.Job: map 51% reduce 0%
20/06/26 05:16:25 INFO mapreduce.Job: map 55% reduce 0%
20/06/26 05:16:37 INFO mapreduce.Job: map 57% reduce 0%
20/06/26 05:16:40 INFO mapreduce.Job: map 68% reduce 0%
20/06/26 05:16:43 INFO mapreduce.Job: map 74% reduce 0%
20/06/26 05:16:46 INFO mapreduce.Job: map 75% reduce 0%
20/06/26 05:16:48 INFO mapreduce.Job: map 79% reduce 0%
20/06/26 05:16:49 INFO mapreduce.Job: map 80% reduce 0%
```

<그림 29>

<그림 30>, <그림 31>은 Task2가 실행이 종료된 후 Web UI를 통해 본 그림이다.

ID	User	Name	Application Type	Queue	Application Priority	StartTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU Vcores	Allocated Memory MB	Reserved CPU Vcores	Reserved Memory MB	% of Queue	% of Cluster	Progress	Tracking UI	Blacklisted Nodes
application_1593072483341_0015	kkoma2623	bd_hadoop.jar	MAPREDUCE	default	0	Fri Jun 26 14:15:27 2020	Fri Jun 26 14:17:54 +0900 2020	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	0.0	0.0	History	0	
application_1593072483341_0014	kkoma2623	bd_hadoop.jar	MAPREDUCE	default	0	Fri Jun 26 13:54:13 2020	Fri Jun 26 13:55:59 +0900 2020	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	0.0	0.0	History	0	
application_1593072483341_0013	kkoma2623	bd_hadoop.jar	MAPREDUCE	default	0	Fri Jun 26 13:47:06 2020	Fri Jun 26 13:48:56 +0900 2020	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	0.0	0.0	History	0	
application_1593072483341_0012	kkoma2623	Task3	SPARK	default	0	Fri Jun 26 00:37:47 2020	Fri Jun 26 00:38:25 +0900 2020	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	0.0	0.0	History	0	
application_1593072483341_0011	kkoma2623	Task3	SPARK	default	0	Fri Jun 26 00:33:44 2020	Fri Jun 26 00:34:42 +0900 2020	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	0.0	0.0	History	0	
application_1593072483341_0010	kkoma2623	Task3	SPARK	default	0	Fri Jun 26 00:23:57 2020	Fri Jun 26 00:24:51 +0900 2020	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	0.0	0.0	History	0	
application_1593072483341_0009	kkoma2623	Task3	SPARK	default	0	Thu Jun 25 22:50:34 2020	Thu Jun 25 22:50:43 +0900 2020	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	0.0	0.0	History	0	
application_1593072483341_0008	kkoma2623	Spark shell	SPARK	default	0	Thu Jun 25 22:07:51 2020	Thu Jun 25 22:08:14 +0900 2020	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	0.0	0.0	History	0	

<그림 30>

The screenshot shows the Google Cloud DataProc Cluster details page for a cluster named 'kmu-cluster'. A specific job, 'MapReduce Job job_1593072483341_0015', is displayed as successful. The job overview table includes columns for Job Name, User Name, Queue, State, Submitted, Started, Finished, and Elapsed times. Diagnostics section shows average map, shuffle, merge, and reduce times. Below this is an ApplicationMaster table with columns for Attempt Number, Start Time, Node, and Logs. The table details task types (Map, Reduce) and attempt types (Maps, Reduces) with their respective counts and status (Failed, Killed, Successful).

<그림 31>

<그림 32>는 실행이 종료된 것을 터미널을 통해 확인한 그림이다.

```

File Edit View Search Terminal Help
Reduce shuffle bytes=735356890
Reduce input records=73535674
Reduce output records=4003121
Spilled Records=220607022
Shuffled Maps =25
Failed Shuffles=0
Merged Map outputs=25
GC time elapsed (ms)=4669
CPU time spent (ms)=410400
Physical memory (bytes) snapshot=8390676480
Virtual memory (bytes) snapshot=52830547968
Total committed heap usage (bytes)=8241283072
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=531608189
File Output Format Counters
Bytes Written=40490596
kkoma2623@kmu-cluster-m:~$ 

```

<그림 32>

hdfs dfs -ls 명령어를 통해 HDFS에 Task2에서 출력 경로로 설정해둔 곳에 파일이 생성된 것을 확인하였다.

```
File Edit View Search Terminal Help
    GC time elapsed (ms)=4669
    CPU time spent (ms)=410400
    Physical memory (bytes) snapshot=8390676480
    Virtual memory (bytes) snapshot=52830547968
    Total committed heap usage (bytes)=8241283072
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=531608189
File Output Format Counters
    Bytes Written=40490596
kkoma2623@kmu-cluster-m:~$ hdfs dfs -ls
Found 5 items
drwxr-xr-x  - kkoma2623 hadoop      0 2020-06-25 15:38 .sparkStaging
-rw-r--r--  2 kkoma2623 hadoop  8031 2020-06-26 04:43 bd_hadoop.jar
-rw-r--r--  2 kkoma2623 hadoop 854597439 2020-06-26 04:39 soc-LiveJournal1.txt
drwxr-xr-x  - kkoma2623 hadoop      0 2020-06-26 04:55 task1Outputs
drwxr-xr-x  - kkoma2623 hadoop      0 2020-06-26 05:17 task2Outputs
kkoma2623@kmu-cluster-m:~$
```

<그림 33>

hdfs dfs -ls task2Outputs 명령어를 통해 task2Outputs 디렉터리 안에 생성된 출력 파일들의 목록을 확인하였다.

```
File Edit View Search Terminal Help
    IO_ERROR=0
    WRONG_LENGTH=0
    WRONG_MAP=0
    WRONG_REDUCE=0
File Input Format Counters
    Bytes Read=531608189
File Output Format Counters
    Bytes Written=40490596
kkoma2623@kmu-cluster-m:~$ hdfs dfs -ls
Found 5 items
drwxr-xr-x  - kkoma2623 hadoop      0 2020-06-25 15:38 .sparkStaging
-rw-r--r--  2 kkoma2623 hadoop  8031 2020-06-26 04:43 bd_hadoop.jar
-rw-r--r--  2 kkoma2623 hadoop 854597439 2020-06-26 04:39 soc-LiveJournal1.txt
drwxr-xr-x  - kkoma2623 hadoop      0 2020-06-26 04:55 task1Outputs
drwxr-xr-x  - kkoma2623 hadoop      0 2020-06-26 05:17 task2Outputs
kkoma2623@kmu-cluster-m:~$ hdfs dfs -ls task2Outputs
Found 6 items
-rw-r--r--  2 kkoma2623 hadoop      0 2020-06-26 05:17 task2Outputs/_SUCCESS
-rw-r--r--  2 kkoma2623 hadoop  8100999 2020-06-26 05:17 task2Outputs/part-r-00000
-rw-r--r--  2 kkoma2623 hadoop  8098472 2020-06-26 05:17 task2Outputs/part-r-00001
-rw-r--r--  2 kkoma2623 hadoop  8094956 2020-06-26 05:17 task2Outputs/part-r-00002
-rw-r--r--  2 kkoma2623 hadoop  8098317 2020-06-26 05:17 task2Outputs/part-r-00003
-rw-r--r--  2 kkoma2623 hadoop  8097852 2020-06-26 05:17 task2Outputs/part-r-00004
kkoma2623@kmu-cluster-m:~$
```

<그림 34>

hdfs dfs -cat task2Outputs/part-r-00004 명령어를 통해 출력 파일의 일부를 확인하겠다.

```
File Edit View Search Terminal Help
    IO_ERROR=0
    WRONG_LENGTH=0
    WRONG_MAP=0
    WRONG_REDUCE=0
    File Input Format Counters
        Bytes Read=531608189
    File Output Format Counters
        Bytes Written=40490596
kkoma2623@kmu-cluster-m:~$ hdfs dfs -ls
Found 5 items
drwxr-xr-x  - kkoma2623 hadoop      0 2020-06-25 15:38 .sparkStaging
-rw-r--r--  2 kkoma2623 hadoop  8031 2020-06-26 04:43 bd_hadoop.jar
-rw-r--r--  2 kkoma2623 hadoop 854597439 2020-06-26 04:39 soc-LiveJournal1.txt
drwxr-xr-x  - kkoma2623 hadoop      0 2020-06-26 04:55 task1outputs
drwxr-xr-x  - kkoma2623 hadoop      0 2020-06-26 05:17 task2outputs
kkoma2623@kmu-cluster-m:~$ hdfs dfs -ls task2outputs
Found 6 items
-rw-r--r--  2 kkoma2623 hadoop      0 2020-06-26 05:17 task2outputs/_SUCCESS
-rw-r--r--  2 kkoma2623 hadoop  8100999 2020-06-26 05:17 task2outputs/part-r-00000
-rw-r--r--  2 kkoma2623 hadoop  8098472 2020-06-26 05:17 task2outputs/part-r-00001
-rw-r--r--  2 kkoma2623 hadoop  8094956 2020-06-26 05:17 task2outputs/part-r-00002
-rw-r--r--  2 kkoma2623 hadoop  8098317 2020-06-26 05:17 task2outputs/part-r-00003
-rw-r--r--  2 kkoma2623 hadoop  8097852 2020-06-26 05:17 task2outputs/part-r-00004
kkoma2623@kmu-cluster-m:~$ hdfs dfs -cat task2outputs/part-r-00004
```

<그림 35>

출력이 너무 길어서 keyboard interrupt를 사용하여 출력을 끊었다.

다음 그림은 Task2의 출력물 중 일부이다.

```
File Edit View Search Terminal Help
1529364 5
1529369 30
1529374 105
1529379 33
1529384 49
1529389 23
1529394 13
1529399 1
1529404 48
1529409 7
1529414 13
1529419 17
1529424 19
1529429 25
1529434 24
1529439 27
1529444 5
1529449 3
1529454 28
1529459 35
1529464 37
1529469 11
1529474 cat: Filesystem closed
kkoma2623@kmu-cluster-m:~$
```

<그림 36>

IntelliJ에서 sbt shell에서 package 명령어를 통해 Task3 와 Task4 를 수행할 수 있는 jar 파일을 생성하겠다.

The screenshot shows the IntelliJ IDEA interface with the following details:

- File Structure:** The left sidebar shows the project structure under 'Project'. It includes a 'src' folder containing 'main' and 'scala' subfolders. 'main' contains 'bigdata' with 'Task3.scala' and 'Task4.scala'. 'scala' contains 'bigdata' with 'countTriangle', 'Task3', and 'Task4'. There is also a 'test' folder and a 'target' folder containing 'scala-2.11'.
- Code Editor:** The main window displays the content of 'Task3.scala'. The code defines a Spark application 'Task3' that reads edges from a file, splits them into two sets based on their middle character, and then finds the union of these sets.
- Terminal:** At the bottom, the terminal window shows the output of the 'sbt shell bigdata' command. It lists various informational messages about the sbt build process, including loading global plugins, setting up the project, and defining the current project to 'bigdata'.

<그림 37>

<그림 38>은 jar 파일을 생성을 완료한 그림이다.

bigdata_2.11-0.1.jar 파일이 생성된 것을 확인하였다. 이를 bd_spark.jar로 리팩토링 하였다.

The screenshot shows the IntelliJ IDEA interface with the following details:

- Project Structure:** The left sidebar shows the project structure under 'Project'. It includes modules like 'bigdata', 'Task3', and 'Task4', and source files like 'Task3.scala' and 'Task4.scala'. A 'target' folder contains 'scala-2.11' which includes 'classes', 'update', and 'bigdata_2.11-0.1.jar'.
- Code Editor:** The main window displays 'Task3.scala' with the following code:

```
1 package bigdata
2
3 import org.apache.spark.{SparkConf, SparkContext}
4
5 import scala.collection.mutable.ListBuffer
6
7 object Task3 {
8   def main(args: Array[String]): Unit = {
9     val conf = new SparkConf().setAppName("Task3")
10    val sc = new SparkContext(conf)
11
12    val input = args(0)
13    val output = args(1)
14
15    val sortedEdges = sc.textFile(input).repartition(128) // 8 l
16
17    val edges1 = sortedEdges.map(x=>
18      val split = x.split("\\|")
19      (split(0).toInt, split(1).toInt)
20    ).sortByKey()
21
22    val edges2 = sortedEdges.map(x=>
23      val split = x.split("\\|")
24      (split(1).toInt, split(0).toInt)
25    ).sortByKey()
26
27    val edges = edges1.union(edges2)
28
29    val degrees = edges.map(x=>
30      (x._1, 1)
31    ).reduceByKey((a,b)=>
32      main(args: Array[String])  λ(x: Any)
```
- Terminal:** The bottom panel shows the sbt shell output:

```
sbtshell bigdata
[info] Loading settings for project bigdata from build.sbt ...
[info] set current project to bigdata (in build file:/home/seyy-i1/IdeaProjects/bigdata/)
[info] Defining Global / ideaPort
[info] The new value will be used by Compile / compile, Test / compile
[info] Reapplying settings...
[info] set current project to bigdata (in build file:/home/seyy-i1/IdeaProjects/bigdata/)
[info] set current project to bigdata (in build file:/home/seyy-i1/IdeaProjects/bigdata/)
[info] Compiling 1 Scala source to /home/seyy-i1/IdeaProjects/bigdata/target/scala-2.11/classes ...
[info] Done compiling.
[warn] Multiple main classes detected. Run 'show discoveredMainClasses' to see the list
[success] Total time: 5 s, completed Jun 26, 2028 2:25:14 PM
[]$sbt:bigdata:

```

<그림 38>

scp bd_spark.jar kkoma2623@35.190.230.101:~/ 명령어를 통해 bd_spark.jar 파일을 GCP에 업로드 하였다.

```
File Edit View Search Terminal Help
sexy-ji@sexy-ji:~/IdeaProjects/bigdata/target/scala-2.11$ scp bd_spark.jar kkoma2623@35.190.230.101:~
kkoma2623@35.190.230.101's password:
Permission denied, please try again.
kkoma2623@35.190.230.101's password:
bd_spark.jar                                         100%   29KB 381.5KB/s   00:00
sexy-ji@sexy-ji:~/IdeaProjects/bigdata/target/scala-2.11$
```

<그림 39>

클러스터에서 ls 명령어를 통해 bd_spark.jar 파일이 업로드 된 것을 확인하였다.

```
File Edit View Search Terminal Help
1529374 105
1529379 33
1529384 49
1529389 23
1529394 13
1529399 1
1529404 48
1529409 7
1529414 13
1529419 17
1529424 19
1529429 25
1529434 24
1529439 27
1529444 5
1529449 3
1529454 28
1529459 35
1529464 37
1529469 11
1529474 cat: Filesystem closed
kkoma2623@kmu-cluster-m:~$ ls
bd_hadoop.jar  bd_spark.jar  soc-LiveJournal1.txt
kkoma2623@kmu-cluster-m:~$
```

<그림 40>

hdfs dfs -put bd_spark.jar 명령어를 통해 bd_spark.jar 파일을 HDFS에 업로드 하였다.

hdfs dfs -ls 명령어를 통해 HDFS에 bd_spark.jar 파일이 업로드 된 것을 확인하였다.

```
File Edit View Search Terminal Help
1529419 17
1529424 19
1529429 25
1529434 24
1529439 27
1529444 5
1529449 3
1529454 28
1529459 35
1529464 37
1529469 11
1529474 cat: Filesystem closed
kkoma2623@kmu-cluster-m:~$ ls
bd_hadoop.jar bd_spark.jar soc-LiveJournal1.txt
kkoma2623@kmu-cluster-m:~$ hdfs dfs -put bd_spark.jar
kkoma2623@kmu-cluster-m:~$ hdfs dfs -ls
Found 6 items
drwxr-xr-x  - kkoma2623 hadoop      0 2020-06-25 15:38 .sparkStaging
-rw-r--r--  2 kkoma2623 hadoop    8031 2020-06-26 04:43 bd_hadoop.jar
-rw-r--r--  2 kkoma2623 hadoop   29497 2020-06-26 05:27 bd_spark.jar
-rw-r--r--  2 kkoma2623 hadoop  854597439 2020-06-26 04:39 soc-LiveJournal1.txt
drwxr-xr-x  - kkoma2623 hadoop      0 2020-06-26 04:55 task1Outputs
drwxr-xr-x  - kkoma2623 hadoop      0 2020-06-26 05:17 task2Outputs
kkoma2623@kmu-cluster-m:~$
```

<그림 41>

spark-submit --num-executors 12 --class bigdata.Task3 bd_spark.jar task1Outputs/*

task3Outputs 명령어를 통해 bigdata 패키지의 Task3를 실행하겠다.

--num-executors 12 옵션을 통해 executors를 12로 설정하였다.

args[0]에 입력파일 경로를, args[1]에 출력 파일 경로를 설정해주었다.

입력 파일 경로는 task1Outputs/*이고 출력 파일 경로는 task3Outputs이다.

```
File Edit View Search Terminal Help
1529424 19
1529429 25
1529434 24
1529439 27
1529444 5
1529449 3
1529454 28
1529459 35
1529464 37
1529469 11
1529474 cat: Filesystem closed
kkoma2623@kmu-cluster-m:~$ ls
bd_hadoop.jar bd_spark.jar soc-LiveJournal1.txt
kkoma2623@kmu-cluster-m:~$ hdfs dfs -put bd_spark.jar
kkoma2623@kmu-cluster-m:~$ hdfs dfs -ls
Found 6 items
drwxr-xr-x  - kkoma2623 hadoop      0 2020-06-25 15:38 .sparkStaging
-rw-r--r--  2 kkoma2623 hadoop    8031 2020-06-26 04:43 bd_hadoop.jar
-rw-r--r--  2 kkoma2623 hadoop   29497 2020-06-26 05:27 bd_spark.jar
-rw-r--r--  2 kkoma2623 hadoop  854597439 2020-06-26 04:39 soc-LiveJournal1.txt
drwxr-xr-x  - kkoma2623 hadoop      0 2020-06-26 04:55 task1Outputs
drwxr-xr-x  - kkoma2623 hadoop      0 2020-06-26 05:17 task2Outputs
kkoma2623@kmu-cluster-m:~$ spark-submit --num-executors 12 --class bigdata.Task3 bd_spark.jar task1Outputs/* task3Outputs
```

<그림 42>

bigdata 패키지의 Task3를 실행하는 화면이다.

```
File Edit View Search Terminal Help
-rw-r--r-- 2 kkoma2623 hadoop      8031 2020-06-26 04:43 bd_hadoop.jar
-rw-r--r-- 2 kkoma2623 hadoop      29497 2020-06-26 05:27 bd_spark.jar
-rw-r--r-- 2 kkoma2623 hadoop  854597439 2020-06-26 04:39 soc-LiveJournal1.txt
drwxr-xr-x - kkoma2623 hadoop       0 2020-06-26 04:55 task1outputs
drwxr-xr-x - kkoma2623 hadoop       0 2020-06-26 05:17 task2outputs
kkoma2623@kmu-cluster-m:~$ spark-submit --num-executors 12 --class bigdata.Task3 bd_spark.jar task1outputs/* task3outputs
20/06/26 05:29:51 INFO org.apache.spark.util.log: Logging initialized @2512ms
20/06/26 05:29:51 INFO org.apache.spark.project.jetty.server.Server: jetty-9.3.z-SNAPSHOT, build timestamp: unknown, git hash: unknown
20/06/26 05:29:51 INFO org.apache.spark.project.jetty.server.Server: Started @2588ms
20/06/26 05:29:51 INFO org.apache.spark.project.jetty.server.AbstractConnector: Started ServerConnector@5426b0d{HTTP/1.1,[http/1.1]}{0.0.0.0:4040}
20/06/26 05:29:51 WARN org.apache.spark.scheduler.FairSchedulableBuilder: Fair Scheduler configuration file not found so jobs will be scheduled in FIFO order. To use fair scheduling, configure pools in fairscheduler.xml or set spark.scheduler.allocation.file to a file that contains the configuration.
20/06/26 05:29:52 INFO org.apache.hadoop.yarn.client.RMProxy: Connecting to ResourceManager at kmu-cluster-m/10.146.0.9:8032
20/06/26 05:29:53 INFO org.apache.hadoop.yarn.client.AHSProxy: Connecting to Application History server at kmu-cluster-m/10.146.0.9:10200
20/06/26 05:29:55 INFO org.apache.hadoop.yarn.client.api.impl.YarnClientImpl: Submitted application application_1593072483341_0016
20/06/26 05:30:01 INFO org.apache.hadoop.mapred.FileInputFormat: Total input files to process : 5
```

<그림 43>

<그림 44>~<그림 50>은 Web UI에서 실행중인 것을 확인하는 그림이다.

The screenshot shows a browser window with the URL <https://zpnfpbpizdrgrhg.com/gateway/default/sparkhistory?showIncomplete=true>. The page title is "bigData-290408 > kmu-cluster". It displays the "Spark 2.3.4 History Server" interface. The application details are as follows:

App ID	App Name	Started	Spark User	Last Updated	Event Log
application_1593072483341_0016	Task3	2020-06-26 14:29:50	kkoma2623	2020-06-26 14:30:51	Download

Below the table, it says "Showing 1 to 1 of 1 entries" and has a link "Back to completed applications".

<그림 44>

Screenshot of the Google Cloud DataProc Cluster details - Spark application UI. The cluster is named 'bigdata-280408' and is associated with 'kmu-cluster'. The UI shows the following information:

- Spark Jobs** (1 active):

Job Id	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
2	sortByKey at Task3.scala:33 sortByKey at Task3.scala:33	2020/06/26 05:31:13	33 s	0/5	141/725 (6 running)
- Completed Jobs** (2):

Job Id	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
1	sortByKey at Task3.scala:25 sortByKey at Task3.scala:25	2020/06/26 05:30:50	22 s	1/1 (1 skipped)	120/120 (5 skipped)
0	sortByKey at Task3.scala:20 sortByKey at Task3.scala:20	2020/06/26 05:30:01	49 s	2/2	125/125

<그림 45>

Screenshot of the Google Cloud DataProc Cluster details - Spark application UI. The cluster is named 'bigdata-280408' and is associated with 'kmu-cluster'. The UI shows the following information:

- Details for Job** (Status: RUNNING):

Details for Job										
Status	RUNNING									
Active Stages	1									
Pending Stages	3									
Completed Stages	1									
Event Timeline DAG Visualization										
Active Stages (1)										
Stage Id	Pool Name	Description	Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Read	Shuffle Write	
5	default	map at Task3.scala:22	+details 2020/06/26 05:31:13	59 s	88/120 (2 running)			345.7 MB	205.8 MB	
- Pending Stages (3)**:

Stage Id	Pool Name	Description	Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Read	Shuffle Write
8	default	sortByKey at Task3.scala:33	+details Unknown	Unknown	0/240				
7	default	map at Task3.scala:29	+details Unknown	Unknown	0/240				
4	default	repartition at Task3.scala:15	+details Unknown	Unknown	0/5				
- Completed Stages (1)**:

Stage Id	Pool Name	Description	Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Read	Shuffle Write
6	default	map at Task3.scala:17	+details 2020/06/26 05:31:13	37 s	120/120			471.5 MB	283.6 MB

<그림 46>

https://rzpnfpbpizdrghmzpk2tjseu-dot-asia-northeast1.dataproc.googleusercontent.com/gateway/default/sparkhistory/history/application_1593072483341_0016/jobs?user.name=anonymous

Sign out

bigdata-280408 > kmu-cluster

Spark 2.3.4 Jobs Stages Storage Environment Executors Task3 application UI

Spark Jobs (2)

User: kkoma023
Total Uptime:
Scheduling Mode: FAIR
Active Jobs: 1
Completed Jobs: 3

Event Timeline

Active Jobs (1)

Job Id	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
3	sortByKey at Task3.scala:56 sortByKey at Task3.scala:56	2020/06/26 05:33:37	5.8 min	4/9	856/1565 (6 running)

Completed Jobs (3)

Job Id	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
2	sortByKey at Task3.scala:33 sortByKey at Task3.scala:33	2020/06/26 05:31:13	2.4 min	4/4 (1 skipped)	720/720 (5 skipped)
1	sortByKey at Task3.scala:25 sortByKey at Task3.scala:25	2020/06/26 05:30:50	22 s	1/1 (1 skipped)	120/120 (5 skipped)
0	sortByKey at Task3.scala:20 sortByKey at Task3.scala:20	2020/06/26 05:30:01	49 s	2/2	125/125

<그림 47>

유동 시간이 오래 걸리는 Job 의 Event Timeline 을 캡처해보았다.

https://rzpnfpbpizdrghmzpk2tjseu-dot-asia-northeast1.dataproc.googleusercontent.com/gateway/default/sparkhistory/history/application_1593072483341_0016/jobs?user.name=anonymous

Sign out

bigdata-280408 > kmu-cluster

Spark 2.3.4 Jobs Stages Storage Environment Executors Task3 application UI

Spark Jobs (2)

User: kkoma023
Total Uptime:
Scheduling Mode: FAIR
Active Jobs: 1
Completed Jobs: 4

Event Timeline

Enable zooming

Executors

- Added
- Removed

Jobs

- Succeeded
- Failed
- Running

Executor driver added
Executor 5 added
Executor 4 added
Executor 3 added
Executor 2 added
Executor 1 added

sort
sortByKey at Task3.scala:56 (job 3)
sortByKey at Task3.scala:25
sortByKey at Task3.scala:20
sortByKey at Task3.scala:68 (job 4)

05.30 05.31 05.32 05.33 05.34 05.35 05.36 05.37 05.38 05.39 05.40 05.41 05.42 05.43 05.44 05.45 05.46 05.47 05.48 05.49 05.50 05.51 05.52 05.53 05.54 05.55
Fri 26 June

Active Jobs (1)

Job Id	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
4	sortByKey at Task3.scala:68 sortByKey at Task3.scala:68	2020/06/26 05:49:23	5.6 min	0/12	110/2165 (6 running)

Completed Jobs (4)

Job Id	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
3	sortByKey at Task3.scala:56 sortByKey at Task3.scala:56	2020/06/26 05:33:37	16 min	5/5 (4 skipped)	1080/1080 (485 skipped)
2	sortByKey at Task3.scala:33 sortByKey at Task3.scala:33	2020/06/26 05:31:13	2.4 min	4/4 (1 skipped)	720/720 (5 skipped)
1	sortByKey at Task3.scala:25 sortByKey at Task3.scala:25	2020/06/26 05:30:50	22 s	1/1 (1 skipped)	120/120 (5 skipped)
0	sortByKey at Task3.scala:20 sortByKey at Task3.scala:20	2020/06/26 05:30:01	49 s	2/2	125/125

<그림 48>

실행중인 Task3 의 stages 를 캡처해보았다.

Active Stages: 2
Pending Stages: 10
Completed Stages: 12
Skipped Stages: 6

Active Stages (2)

Stage Id	Pool Name	Description	Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Read	Shuffle Write
27	default	map at Task3.scala:58	+details 2020/06/26 05:49:23	7.2 min	44/120	96.7 MB	84.9 MB		
26	default	flatMap at Task3.scala:45	+details 2020/06/26 05:49:23	7.2 min	66/240 (6 running)	58.5 MB	1148.9 MB		

Pending Stages (10)

Stage Id	Pool Name	Description	Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Read	Shuffle Write
29	default	sortByKey at Task3.scala:68	+details Unknown	Unknown	0/240				
28	default	flatMap at Task3.scala:61	+details Unknown	Unknown	0/240				
25	default	map at Task3.scala:37	+details Unknown	Unknown	0/240				
24	default	map at Task3.scala:35	+details Unknown	Unknown	0/240				
23	default	filter at Task3.scala:33	+details Unknown	Unknown	0/240				
22	default	sortByKey at Task3.scala:20	+details Unknown	Unknown	0/120				
21	default	map at Task3.scala:29	+details Unknown	Unknown	0/240				
20	default	map at Task3.scala:22	+details Unknown	Unknown	0/120				
19	default	map at Task3.scala:17	+details Unknown	Unknown	0/120				
18	default	repartition at Task3.scala:15	+details Unknown	Unknown	0/5				

Completed Stages (12)

Stage Id	Pool Name	Description	Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Read	Shuffle Write
17	default	sortByKey at Task3.scala:56	+details 2020/06/26 05:37:34	12 min	240/240	213.0 MB			
16	default	map at Task3.scala:37	+details 2020/06/26 05:35:56	1.6 min	240/240	363.9 MB	213.0 MB		
15	default	map at Task3.scala:35	+details 2020/06/26 05:34:12	1.7 min	240/240	260.6 MB	344.4 MB		
14	default	filter at Task3.scala:33	+details 2020/06/26 05:33:37	13 s	240/240	32.4 MB	19.5 MB		
13	default	sortByKey at Task3.scala:20	+details 2020/06/26 05:33:37	35 s	120/120	283.6 MB	241.2 MB		
8	default	sortByKey at Task3.scala:33	+details 2020/06/26 05:33:25	12 s	240/240	32.4 MB			
7	default	map at Task3.scala:29	+details 2020/06/26 05:32:21	1.1 min	240/240	564.2 MB	32.4 MB		
6	default	map at Task3.scala:17	+details 2020/06/26 05:31:13	37 s	120/120	471.5 MB	283.6 MB		
5	default	map at Task3.scala:22	+details 2020/06/26 05:31:13	1.1 min	120/120	471.5 MB	280.6 MB		
3	default	sortByKey at Task3.scala:25	+details 2020/06/26 05:30:50	22 s	120/120	471.5 MB			
1	default	sortByKey at Task3.scala:20	+details 2020/06/26 05:30:22	28 s	120/120	471.5 MB			
0	default	repartition at Task3.scala:15	+details 2020/06/26 05:30:01	19 s	5/5	507.0 MB			471.5 MB

<그림 49>

실행중인 Task3 의 executors 를 캡처해보았다.

Executors

Summary

RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write	Blacklisted
Active(6)	0.0 B / 16.6 GB	0.0 B	5	6	0	2155	2161	1.6 h (5.4 min)	531.6 MB	4 GB	3.3 GB	0
Dead(0)	0.0 B / 0.0 B	0.0 B	0	0	0	0	0	0 ms (0 ms)	0.0 B	0.0 B	0.0 B	0
Total(6)	0.0 B / 16.6 GB	0.0 B	5	6	0	2155	2161	1.6 h (5.4 min)	531.6 MB	4 GB	3.3 GB	0

Executors

Executor ID	Address	Status	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write	Logs
driver	kmu-cluster-m-asia-northeast1-a.c.bigdata-280408.internal:37933	Active	0	0.0 B / 2 GB	0.0 B	0	0	0	0	0	0 ms (0 ms)	0.0 B	0.0 B	0.0 B	
1	kmu-cluster-w-0-asia-northeast1-a.c.bigdata-280408.internal:45557	Active	0	0.0 B / 2.9 GB	0.0 B	1	1	0	170	171	17 min (1.0 min)	106 MB	381.1 MB	511.3 MB	stdout stderr
2	kmu-cluster-w-1-asia-northeast1-a.c.bigdata-280408.internal:45703	Active	0	0.0 B / 2.9 GB	0.0 B	1	1	0	792	793	23 min (1.3 min)	105.7 MB	1.3 GB	762.6 MB	stdout stderr
3	kmu-cluster-w-2-asia-northeast1-a.c.bigdata-280408.internal:36965	Active	0	0.0 B / 2.9 GB	0.0 B	1	1	0	277	278	17 min (54 s)	106.5 MB	662.3 MB	754.2 MB	stdout stderr
4	kmu-cluster-w-0-asia-northeast1-a.c.bigdata-280408.internal:45809	Active	0	0.0 B / 2.9 GB	0.0 B	1	1	0	162	163	18 min (1.1 min)	106 MB	362.8 MB	512.8 MB	stdout stderr
5	kmu-cluster-w-1-asia-northeast1-a.c.bigdata-280408.internal:39411	Active	0	0.0 B / 2.9 GB	0.0 B	1	2	0	754	756	23 min (1.1 min)	107.4 MB	1.3 GB	730.6 MB	stdout stderr

<그림 50>

<그림 51> ~ <그림 55>는 Task3 가 종료되고 WebUI 를 통해 확인한 그림이다.

App ID	App Name	Started	Completed	Duration	Spark User	Last Updated	Event Log
application_1593072483341_0016	Task3	2020-06-26 14:29:50	2020-06-26 15:19:21	50 min	kkoma2623	2020-06-26 15:19:21	<button>Download</button>
application_1593072483341_0012	Task3	2020-06-26 00:37:42	2020-06-26 00:38:25	43 s	kkoma2623	2020-06-26 00:38:25	<button>Download</button>
application_1593072483341_0011	Task3	2020-06-26 00:33:39	2020-06-26 00:34:42	1.0 min	kkoma2623	2020-06-26 00:34:42	<button>Download</button>
application_1593072483341_0010	Task3	2020-06-26 00:23:52	2020-06-26 00:24:51	59 s	kkoma2623	2020-06-26 00:24:51	<button>Download</button>
application_1593072483341_0009	Task3	2020-06-25 22:50:29	2020-06-25 23:40:43	50 min	kkoma2623	2020-06-25 23:40:43	<button>Download</button>
application_1593072483341_0008	Spark shell	2020-06-25 22:07:47	2020-06-25 22:52:14	44 min	kkoma2623	2020-06-25 22:52:14	<button>Download</button>
application_1593072483341_0007	Spark shell	2020-06-25 21:59:09	2020-06-25 22:01:01	1.9 min	kkoma2623	2020-06-25 22:01:01	<button>Download</button>
application_1593072483341_0001	Spark shell	2020-06-25 17:35:28	2020-06-25 21:23:58	3.8 h	kkoma2623	2020-06-25 21:23:58	<button>Download</button>
application_1593072483341_0004	Task3	2020-06-25 20:07:41	2020-06-25 20:56:15	49 min	kkoma2623	2020-06-25 20:56:15	<button>Download</button>

Showing 1 to 9 of 9 entries
Show incomplete applications

<그림 51>

Job Id	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
6	runJob at SparkHadoopWriter.scala:78 runJob at SparkHadoopWriter.scala:78	2020/06/26 06:18:38	43 s	2/2 (14 skipped)	480/480 (2405 skipped)
5	sortByKey at Task3.scala:77 sortByKey at Task3.scala:77	2020/06/26 06:16:59	1.6 min	4/4 (11 skipped)	720/720 (1925 skipped)
4	sortByKey at Task3.scala:68 sortByKey at Task3.scala:68	2020/06/26 05:49:23	28 min	4/4 (8 skipped)	840/840 (1325 skipped)
3	sortByKey at Task3.scala:56 sortByKey at Task3.scala:56	2020/06/26 05:33:37	16 min	5/5 (4 skipped)	1080/1080 (485 skipped)
2	sortByKey at Task3.scala:33 sortByKey at Task3.scala:33	2020/06/26 05:31:13	2.4 min	4/4 (1 skipped)	720/720 (5 skipped)
1	sortByKey at Task3.scala:25 sortByKey at Task3.scala:25	2020/06/26 05:30:50	22 s	1/1 (1 skipped)	120/120 (5 skipped)
0	sortByKey at Task3.scala:20 sortByKey at Task3.scala:20	2020/06/26 05:30:01	49 s	2/2	125/125

<그림 52>

Google Chrome https://rznfpbpibzdrgh... New Tab https://rznfpbpibzdrghmfpk2tj5eu-dot-asia-northeast1.dataproc.googleusercontent.com/gateway/default/sparkhistory/history/application_1593072483341_0016/stages Sign out

bigdata-280408 > kmu-cluster Task3 application UI

Jobs Stages Storage Environment Executors

Stages for All Jobs

Completed Stages: 22 Skipped Stages: 39

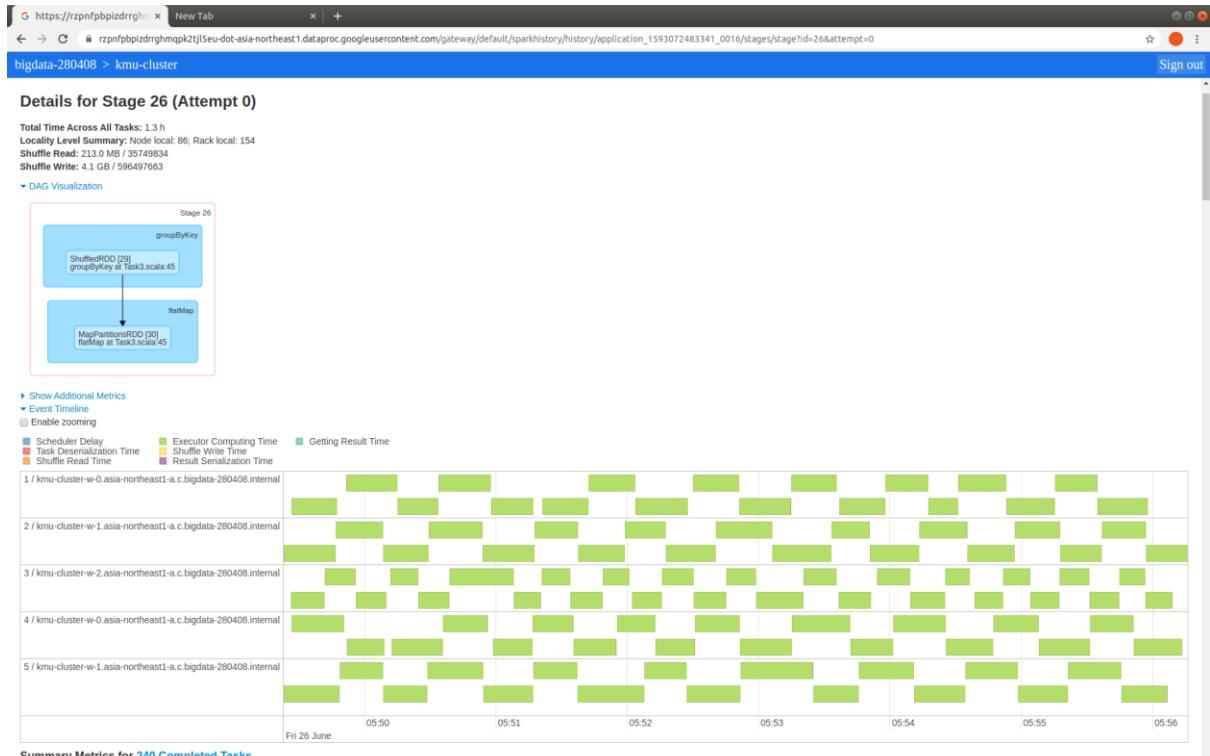
Completed Stages (22)

Stage Id	Pool Name	Description	Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Read	Shuffle Write
60	default	runJob at SparkHadoopWriter.scala:78	+details 2020/06/26 06:19:06	16 s	240/240	25.8 MB	14.2 MB		
59	default	map at Task3.scala:75	+details 2020/06/26 06:18:38	27 s	240/240	40.4 MB	14.2 MB		
44	default	sortByKey at Task3.scala:77	+details 2020/06/26 06:18:12	26 s	240/240	40.4 MB	99.6 MB	22.2 MB	
43	default	reduceByKey at Task3.scala:73	+details 2020/06/26 06:17:58	14 s	120/120	99.6 MB	22.2 MB		
42	default	flatMap at Task3.scala:71	+details 2020/06/26 06:16:59	59 s	120/120	283.6 MB	99.6 MB		
41	default	reduceByKey at Task3.scala:66	+details 2020/06/26 06:16:59	19 s	240/240	93.8 MB	18.3 MB		
29	default	sortByKey at Task3.scala:68	+details 2020/06/26 06:16:45	14 s	240/240	93.8 MB			
28	default	flatMap at Task3.scala:61	+details 2020/06/26 06:05:29	11 min	240/240	4.3 GB	93.8 MB		
27	default	map at Task3.scala:58	+details 2020/06/26 05:49:23	16 min	120/120	283.6 MB	231.6 MB		
26	default	flatMap at Task3.scala:45	+details 2020/06/26 05:49:23	16 min	240/240	213.0 MB	4.1 GB		
17	default	sortByKey at Task3.scala:56	+details 2020/06/26 05:37:34	12 min	240/240	213.0 MB			
16	default	map at Task3.scala:37	+details 2020/06/26 05:35:56	1.6 min	240/240	363.9 MB	213.0 MB		
15	default	map at Task3.scala:35	+details 2020/06/26 05:34:12	1.7 min	240/240	260.6 MB	344.4 MB		
14	default	filter at Task3.scala:33	+details 2020/06/26 05:33:37	13 s	240/240	32.4 MB	19.5 MB		
13	default	sortByKey at Task3.scala:20	+details 2020/06/26 05:33:37	35 s	120/120	283.6 MB	241.2 MB		
8	default	sortByKey at Task3.scala:33	+details 2020/06/26 05:33:25	12 s	240/240	32.4 MB			
7	default	map at Task3.scala:29	+details 2020/06/26 05:32:21	1.1 min	240/240	564.2 MB	32.4 MB		
6	default	map at Task3.scala:17	+details 2020/06/26 05:31:13	37 s	120/120	471.5 MB	283.6 MB		
5	default	map at Task3.scala:22	+details 2020/06/26 05:31:13	1.1 min	120/120	471.5 MB	280.6 MB		
3	default	sortByKey at Task3.scala:25	+details 2020/06/26 05:30:50	22 s	120/120	471.5 MB			
1	default	sortByKey at Task3.scala:20	+details 2020/06/26 05:30:22	28 s	120/120	471.5 MB			
0	default	repartition at Task3.scala:15	+details 2020/06/26 05:30:01	19 s	5/5	507.0 MB		471.5 MB	

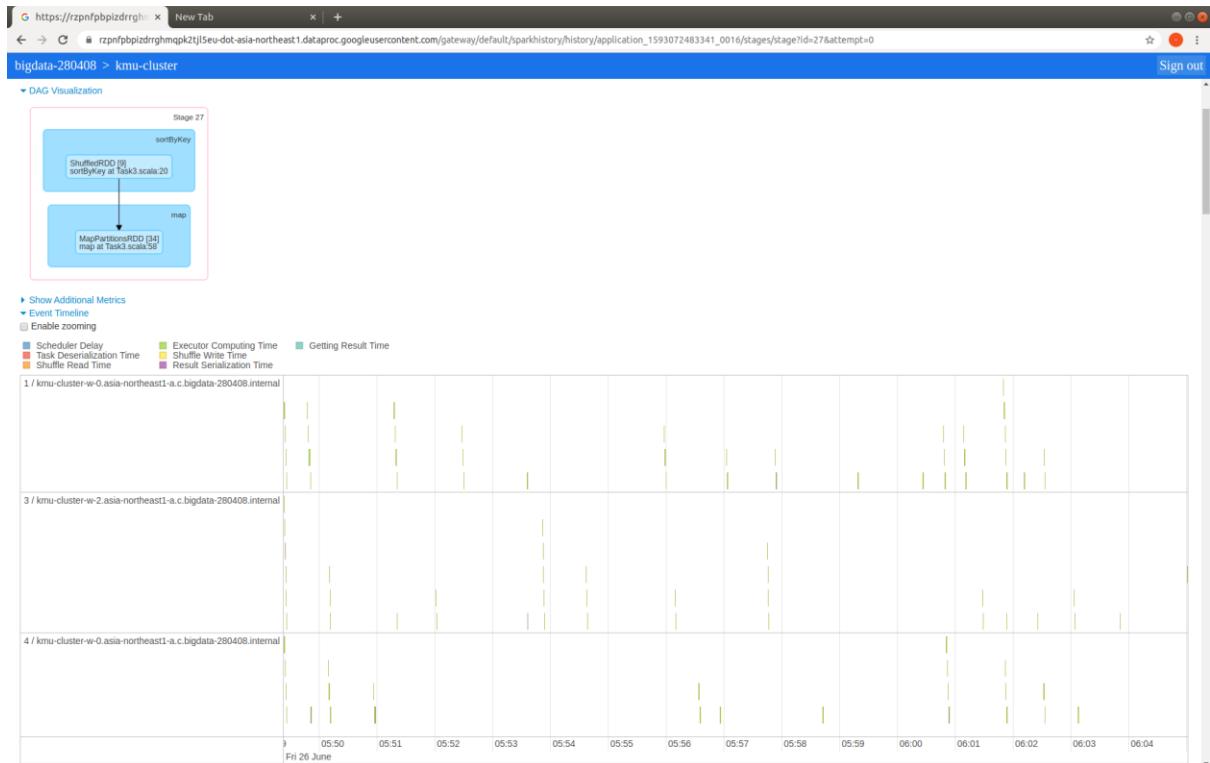
Skipped Stages (39)

Stage Id	Pool Name	Description	Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Read	Shuffle Write
58	default	reduceByKey at Task3.scala:66	+details Unknown	Unknown	0/240				
57	default	reduceByKey at Task3.scala:73	+details Unknown	Unknown	0/120				
56	default	flatMap at Task3.scala:71	+details Unknown	Unknown	0/120				
55	default	flatMap at Task3.scala:61	+details Unknown	Unknown	0/240				

<그림 53>



<그림 54>



<그림 55>

<그림 56>은 Task3 가 종료된 것을 터미널을 통해 확인한 그림이다.

```

File Edit View Search Terminal Help
-rw-r--r-- 2 kkoma2623 hadoop 854597439 2020-06-26 04:39 soc-LiveJournal1.txt
drwxr-xr-x - kkoma2623 hadoop 0 2020-06-26 04:55 task10outputs
drwxr-xr-x - kkoma2623 hadoop 0 2020-06-26 05:17 task20outputs
kkoma2623@kmu-cluster-m:~$ spark-submit --num-executors 12 --class bigdata.Task3 bd_spark.jar task10outputs/* task30outputs
20/06/26 05:29:51 INFO org.spark_project.jetty.util.log: Logging initialized @2512ms
20/06/26 05:29:51 INFO org.spark_project.jetty.server.Server: jetty-9.3.z-SNAPSHOT, build timestamp: unknown, git hash: unknown
20/06/26 05:29:51 INFO org.spark_project.jetty.server.Server: Started @2588ms
20/06/26 05:29:51 INFO org.spark_project.jetty.server.AbstractConnector: Started ServerConnector@5426b0d6{HTTP/1.1,[http/1.1]}{0.0.0.0:4040}
20/06/26 05:29:51 WARN org.apache.spark.scheduler.FairSchedulableBuilder: Fair Scheduler configuration file not found so jobs will be scheduled in FIFO order. To use fair scheduling, configure pools in fairscheduler.xml or set spark.scheduler.allocation.file to a file that contains the configuration.
20/06/26 05:29:52 INFO org.apache.hadoop.yarn.client.RMProxy: Connecting to ResourceManager at kmu-cluster-m/10.146.0.9:8032
20/06/26 05:29:53 INFO org.apache.hadoop.yarn.client.AHSProxy: Connecting to Application History server at kmu-cluster-m/10.146.0.9:10200
20/06/26 05:29:55 INFO org.apache.hadoop.yarn.client.api.impl.YarnClientImpl: Submitted application application_1593072483341_0016
20/06/26 05:30:01 INFO org.apache.hadoop.mapred.FileInputFormat: Total input files to process : 5
20/06/26 06:19:21 INFO org.spark_project.jetty.server.AbstractConnector: Stopped Spark@5426b0d6{HTTP/1.1,[http/1.1]}{0.0.0.0:4040}
kkoma2623@kmu-cluster-m:~$ 

```

<그림 56>

hdfs dfs -ls 명령어를 통해 출력 경로로 설정해준 곳에 출력 파일이 생성된 것을 확인하였다.

hdfs dfs -ls task3Outputs 명령어를 통해 task3Outputs 디렉터리에 출력 파일들이 생성된 것을 확인해보았다.

```
File Edit View Search Terminal Help
kkoma2623@kmu-cluster-m:~$ hdfs dfs -ls
Found 7 items
drwxr-xr-x 2 kkoma2623 hadoop 0 2020-06-26 06:19 .sparkStaging
-rw-r--r-- 2 kkoma2623 hadoop 8031 2020-06-26 04:43 bd_hadoop.jar
-rw-r--r-- 2 kkoma2623 hadoop 29497 2020-06-26 05:27 bd_spark.jar
-rw-r--r-- 2 kkoma2623 hadoop 854597439 2020-06-26 04:39 soc-LiveJournal1.txt
drwxr-xr-x 2 kkoma2623 hadoop 0 2020-06-26 04:55 task1outputs
drwxr-xr-x 2 kkoma2623 hadoop 0 2020-06-26 05:17 task2outputs
drwxr-xr-x 2 kkoma2623 hadoop 0 2020-06-26 06:19 task3outputs
kkoma2623@kmu-cluster-m:~$ hdfs dfs -ls task3outputs
Found 241 items
-rw-r--r-- 2 kkoma2623 hadoop 0 2020-06-26 06:19 task3outputs/_SUCCESS
-rw-r--r-- 2 kkoma2623 hadoop 92994 2020-06-26 06:19 task3outputs/part-00000
-rw-r--r-- 2 kkoma2623 hadoop 95125 2020-06-26 06:19 task3outputs/part-00001
-rw-r--r-- 2 kkoma2623 hadoop 103451 2020-06-26 06:19 task3outputs/part-00002
-rw-r--r-- 2 kkoma2623 hadoop 101313 2020-06-26 06:19 task3outputs/part-00003
-rw-r--r-- 2 kkoma2623 hadoop 99283 2020-06-26 06:19 task3outputs/part-00004
-rw-r--r-- 2 kkoma2623 hadoop 109561 2020-06-26 06:19 task3outputs/part-00005
-rw-r--r-- 2 kkoma2623 hadoop 97421 2020-06-26 06:19 task3outputs/part-00006
-rw-r--r-- 2 kkoma2623 hadoop 100612 2020-06-26 06:19 task3outputs/part-00007
-rw-r--r-- 2 kkoma2623 hadoop 102194 2020-06-26 06:19 task3outputs/part-00008
-rw-r--r-- 2 kkoma2623 hadoop 122543 2020-06-26 06:19 task3outputs/part-00009
-rw-r--r-- 2 kkoma2623 hadoop 112342 2020-06-26 06:19 task3outputs/part-00010
-rw-r--r-- 2 kkoma2623 hadoop 109817 2020-06-26 06:19 task3outputs/part-00011
```

<그림 57>

```
File Edit View Search Terminal Help
-rw-r--r-- 2 kkoma2623 hadoop 124268 2020-06-26 06:19 task3outputs/part-00217
-rw-r--r-- 2 kkoma2623 hadoop 102091 2020-06-26 06:19 task3outputs/part-00218
-rw-r--r-- 2 kkoma2623 hadoop 109814 2020-06-26 06:19 task3outputs/part-00219
-rw-r--r-- 2 kkoma2623 hadoop 108048 2020-06-26 06:19 task3outputs/part-00220
-rw-r--r-- 2 kkoma2623 hadoop 123095 2020-06-26 06:19 task3outputs/part-00221
-rw-r--r-- 2 kkoma2623 hadoop 96623 2020-06-26 06:19 task3outputs/part-00222
-rw-r--r-- 2 kkoma2623 hadoop 118105 2020-06-26 06:19 task3outputs/part-00223
-rw-r--r-- 2 kkoma2623 hadoop 110051 2020-06-26 06:19 task3outputs/part-00224
-rw-r--r-- 2 kkoma2623 hadoop 112449 2020-06-26 06:19 task3outputs/part-00225
-rw-r--r-- 2 kkoma2623 hadoop 105263 2020-06-26 06:19 task3outputs/part-00226
-rw-r--r-- 2 kkoma2623 hadoop 114207 2020-06-26 06:19 task3outputs/part-00227
-rw-r--r-- 2 kkoma2623 hadoop 111059 2020-06-26 06:19 task3outputs/part-00228
-rw-r--r-- 2 kkoma2623 hadoop 101467 2020-06-26 06:19 task3outputs/part-00229
-rw-r--r-- 2 kkoma2623 hadoop 101650 2020-06-26 06:19 task3outputs/part-00230
-rw-r--r-- 2 kkoma2623 hadoop 112978 2020-06-26 06:19 task3outputs/part-00231
-rw-r--r-- 2 kkoma2623 hadoop 114960 2020-06-26 06:19 task3outputs/part-00232
-rw-r--r-- 2 kkoma2623 hadoop 105642 2020-06-26 06:19 task3outputs/part-00233
-rw-r--r-- 2 kkoma2623 hadoop 106357 2020-06-26 06:19 task3outputs/part-00234
-rw-r--r-- 2 kkoma2623 hadoop 124673 2020-06-26 06:19 task3outputs/part-00235
-rw-r--r-- 2 kkoma2623 hadoop 109498 2020-06-26 06:19 task3outputs/part-00236
-rw-r--r-- 2 kkoma2623 hadoop 99635 2020-06-26 06:19 task3outputs/part-00237
-rw-r--r-- 2 kkoma2623 hadoop 106954 2020-06-26 06:19 task3outputs/part-00238
-rw-r--r-- 2 kkoma2623 hadoop 118564 2020-06-26 06:19 task3outputs/part-00239
kkoma2623@kmu-cluster-m:~$
```

<그림 58>

hdfs dfs -cat task3Outputs/part-00239 명령어를 통해 출력 파일의 일부를 확인해보겠다.

```
File Edit View Search Terminal Help
-rw-r--r-- 2 kkoma2623 hadoop 124268 2020-06-26 06:19 task3Outputs/part-00217
-rw-r--r-- 2 kkoma2623 hadoop 102091 2020-06-26 06:19 task3Outputs/part-00218
-rw-r--r-- 2 kkoma2623 hadoop 109814 2020-06-26 06:19 task3Outputs/part-00219
-rw-r--r-- 2 kkoma2623 hadoop 108048 2020-06-26 06:19 task3Outputs/part-00220
-rw-r--r-- 2 kkoma2623 hadoop 123095 2020-06-26 06:19 task3Outputs/part-00221
-rw-r--r-- 2 kkoma2623 hadoop 96623 2020-06-26 06:19 task3Outputs/part-00222
-rw-r--r-- 2 kkoma2623 hadoop 118105 2020-06-26 06:19 task3Outputs/part-00223
-rw-r--r-- 2 kkoma2623 hadoop 110051 2020-06-26 06:19 task3Outputs/part-00224
-rw-r--r-- 2 kkoma2623 hadoop 112449 2020-06-26 06:19 task3Outputs/part-00225
-rw-r--r-- 2 kkoma2623 hadoop 105263 2020-06-26 06:19 task3Outputs/part-00226
-rw-r--r-- 2 kkoma2623 hadoop 114207 2020-06-26 06:19 task3Outputs/part-00227
-rw-r--r-- 2 kkoma2623 hadoop 111059 2020-06-26 06:19 task3Outputs/part-00228
-rw-r--r-- 2 kkoma2623 hadoop 101467 2020-06-26 06:19 task3Outputs/part-00229
-rw-r--r-- 2 kkoma2623 hadoop 101650 2020-06-26 06:19 task3Outputs/part-00230
-rw-r--r-- 2 kkoma2623 hadoop 112978 2020-06-26 06:19 task3Outputs/part-00231
-rw-r--r-- 2 kkoma2623 hadoop 114960 2020-06-26 06:19 task3Outputs/part-00232
-rw-r--r-- 2 kkoma2623 hadoop 105642 2020-06-26 06:19 task3Outputs/part-00233
-rw-r--r-- 2 kkoma2623 hadoop 106357 2020-06-26 06:19 task3Outputs/part-00234
-rw-r--r-- 2 kkoma2623 hadoop 124673 2020-06-26 06:19 task3Outputs/part-00235
-rw-r--r-- 2 kkoma2623 hadoop 109498 2020-06-26 06:19 task3Outputs/part-00236
-rw-r--r-- 2 kkoma2623 hadoop 99635 2020-06-26 06:19 task3Outputs/part-00237
-rw-r--r-- 2 kkoma2623 hadoop 106954 2020-06-26 06:19 task3Outputs/part-00238
-rw-r--r-- 2 kkoma2623 hadoop 118564 2020-06-26 06:19 task3Outputs/part-00239
kkoma2623@kmu-cluster-m:~$ hdfs dfs -cat task3Outputs/part-00239
```

<그림 59>

빠르게 출력이 되어서 따로 keyboard interrupt를 주지 않았다.

<그림 60>은 출력 화면이다.

```
File Edit View Search Terminal Help
4845946 4
4846047 47
4846057 15
4846264 8
4846366 349
4846367 432
4846368 398
4846483 3
4846582 1
4846748 19
4846754 55
4846882 1
4847014 1
4847054 247
4847090 24
4847185 5
4847231 3
4847233 1
4847235 1
4847261 1
4847392 1
4847399 3
4847442 17
kkoma2623@kmu-cluster-m:~$
```

<그림 60>

spark-submit --num-executors 12 --class bigdata.Task4 bd_spark.jar task2Outputs/*
task3Outputs/* task4Outputs 명령어를 통해 bigdata 패키지의 Task4를 실행하겠다.

--num-executors 12 옵션을 통해 executors를 12로 설정하였다.

args[0]와 args[1]에 입력파일 경로를, args[2]에 출력 파일 경로를 설정해주었다.

입력 파일 경로는 task2Outputs/* 와 task3Outputs/* 이고 출력 파일 경로는
task4Outputs이다.

```
File Edit View Search Terminal Help
4847261 1
4847392 1
4847399 3
4847442 17
kkoma2623@kmu-cluster-m:~$ spark-submit --num-executors 12 --class bigdata.Task4 bd_spark.jar task2Outputs/* task3Outputs/* task4Outputs
20/06/26 07:03:07 INFO org.spark_project.jetty.util.log: Logging initialized @2448ms
20/06/26 07:03:07 INFO org.spark_project.jetty.server.Server: jetty-9.3.z-SNAPSHOT, build timestamp: unknown, git hash: unknown
20/06/26 07:03:07 INFO org.spark_project.jetty.server.Server: Started @2535ms
20/06/26 07:03:07 INFO org.spark_project.jetty.server.AbstractConnector: Started ServerConnector@1e8f130{HTTP/1.1,[http/1.1]}{0.0.0.0:4040}
20/06/26 07:03:07 WARN org.apache.spark.scheduler.FairSchedulableBuilder: Fair Scheduler configuration file not found so jobs will be scheduled in FIFO order. To use fair scheduling, configure pools in fairscheduler.xml or set spark.scheduler.allocation.file to a file that contains the configuration.
20/06/26 07:03:09 INFO org.apache.hadoop.yarn.client.RMProxy: Connecting to ResourceManager at kmu-cluster-m/10.146.0.9:8032
20/06/26 07:03:09 INFO org.apache.hadoop.yarn.client.AHSProxy: Connecting to Application History server at kmu-cluster-m/10.146.0.9:10200
20/06/26 07:03:12 INFO org.apache.hadoop.yarn.client.api.impl.YarnClientImpl: Submitted application application_1593072483341_0017
20/06/26 07:03:18 INFO org.apache.hadoop.mapred.FileInputFormat: Total input files to process : 5
20/06/26 07:03:18 INFO org.apache.hadoop.mapred.FileInputFormat: Total input files to process : 240
```

<그림 61>

Task4를 실행중인 것을 WebUI를 통해 확인하였다.

The screenshot shows the Apache Spark History Server web interface. The URL is https://rzpnfpbpizdrgh:8483/applications. The page displays the following information:

App ID	App Name	Started	Spark User	Last Updated	Event Log
application_1593072483341_0017	Task4	2020-06-26 16:03:06	kkoma2623	2020-06-26 16:03:45	Download

Below the table, there is a message: "Showing 1 to 1 of 1 entries" and a link: "Back to completed applications".

<그림 62>

너무 빠르게 끝나서 실행중인 화면은 한 장 밖에 캡처하지 못했다.

다음은 실행이 종료된 후 캡처한 WebUI 그림들이다.

The screenshot shows the Apache Spark History Server UI at <https://rzpnfpbpizdrgh:8080/sparkhistory/>. The title bar says "bigdata-280408 > kmu-cluster". The page displays a table of completed applications:

App ID	App Name	Started	Completed	Duration	Spark User	Last Updated	Event Log
application_1593072483341_0017	Task4	2020-06-26 16:03:06	2020-06-26 16:04:07	1.0 min	kkoma2623	2020-06-26 16:04:07	Download
application_1593072483341_0016	Task3	2020-06-26 14:29:50	2020-06-26 15:19:21	50 min	kkoma2623	2020-06-26 15:19:21	Download
application_1593072483341_0012	Task3	2020-06-26 00:37:42	2020-06-26 00:38:25	43 s	kkoma2623	2020-06-26 00:38:25	Download
application_1593072483341_0011	Task3	2020-06-26 00:33:39	2020-06-26 00:34:42	1.0 min	kkoma2623	2020-06-26 00:34:42	Download
application_1593072483341_0010	Task3	2020-06-26 00:23:52	2020-06-26 00:24:51	59 s	kkoma2623	2020-06-26 00:24:51	Download
application_1593072483341_0009	Task3	2020-06-25 22:50:29	2020-06-25 23:40:43	50 min	kkoma2623	2020-06-25 23:40:43	Download
application_1593072483341_0008	Spark shell	2020-06-25 22:07:47	2020-06-25 22:52:14	44 min	kkoma2623	2020-06-25 22:52:14	Download
application_1593072483341_0007	Spark shell	2020-06-25 21:59:09	2020-06-25 22:01:01	1.9 min	kkoma2623	2020-06-25 22:01:01	Download
application_1593072483341_0001	Spark shell	2020-06-25 17:35:28	2020-06-25 21:23:58	3.8 h	kkoma2623	2020-06-25 21:23:58	Download
application_1593072483341_0004	Task3	2020-06-25 20:07:41	2020-06-25 20:56:15	49 min	kkoma2623	2020-06-25 20:56:15	Download

Showing 1 to 10 of 10 entries
Show incomplete applications

<그림 63>

The screenshot shows the Apache Spark Job UI at https://rzpnfpbpizdrgh:8080/gateway/default/spark/history/application_1593072483341_0017/jobs/. The title bar says "bigdata-280408 > kmu-cluster". The page displays a table of completed jobs:

Job Id	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
0	runJob at SparkHadoopWriter.scala:78 runJob at SparkHadoopWriter.scala:78	2020/06/26 07:03:19	48 s	5/5	605/605

<그림 64>

https://rznfpbpibzdrgh... > kmu-cluster

Sign out

Jobs Stages Storage Environment Executors Task4 application UI

Details for Job 0

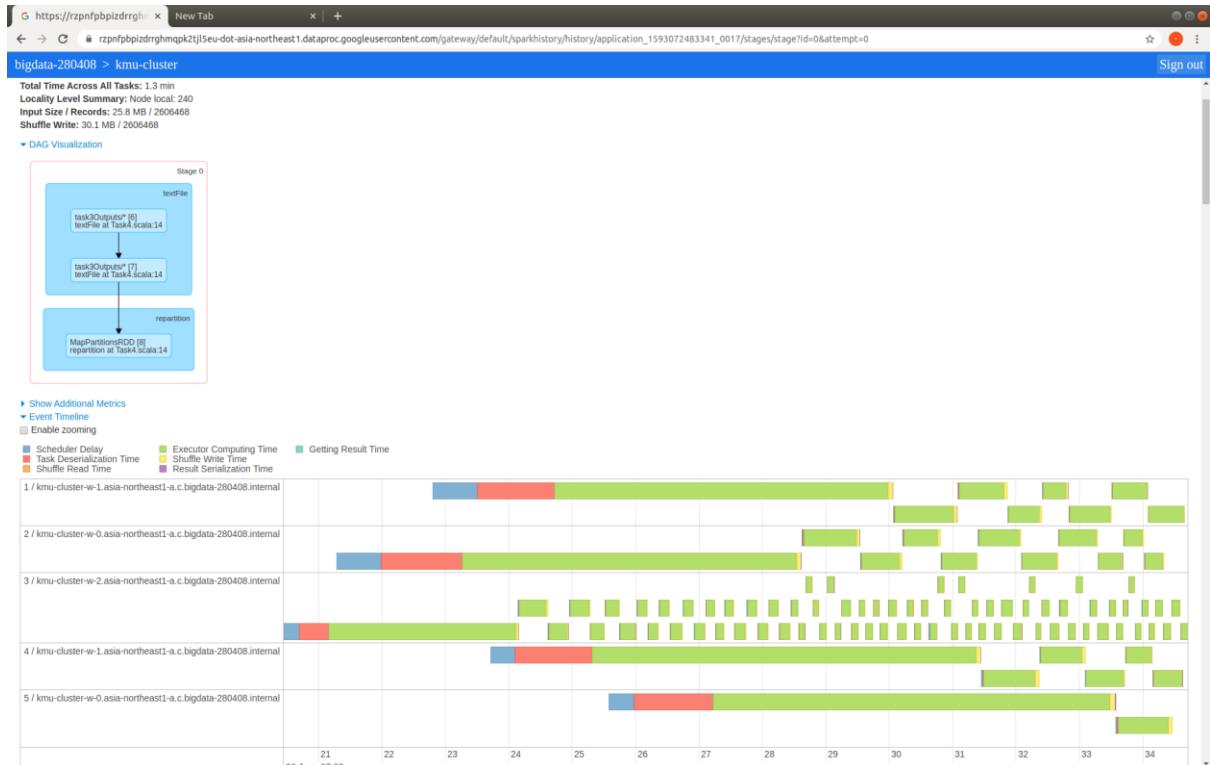
Status: SUCCEEDED
Completed Stages: 5

- Event Timeline
- DAG Visualization

Completed Stages (5)

Stage Id	Pool Name	Description	Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Read	Shuffle Write
4	default	runJob at SparkHadoopWriter.scala:78	+details 2020/06/26 07:03:54	12 s	120/120	42.0 MB	68.9 MB		
3	default	map at Task4.scala:16	+details 2020/06/26 07:03:43	8 s	120/120		39.9 MB		40.2 MB
2	default	repartition at Task4.scala:13	+details 2020/06/26 07:03:19	2 s	5/5	38.6 MB			39.9 MB
1	default	map at Task4.scala:20	+details 2020/06/26 07:03:41	12 s	120/120		30.1 MB		28.7 MB
0	default	repartition at Task4.scala:14	+details 2020/06/26 07:03:19	21 s	240/240	25.8 MB			30.1 MB

<그림 65>



<그림 66>

Task4 가 종료된 것을 터미널을 통해 확인하였다.

```
File Edit View Search Terminal Help
4847399 3
4847442 17
kkoma2623@kmu-cluster-m:~$ spark-submit --num-executors 12 --class bigdata.Task4 bd_spark.jar task20outputs/* task30outputs/* task40outputs
20/06/26 07:03:07 INFO org.spark_project.jetty.util.log: Logging initialized @2448ms
20/06/26 07:03:07 INFO org.spark_project.jetty.server.Server: jetty-9.3.z-SNAPSHOT, build timestamp: unknown, git hash: unknown
20/06/26 07:03:07 INFO org.spark_project.jetty.server.Server: Started @2535ms
20/06/26 07:03:07 INFO org.spark_project.jetty.server.AbstractConnector: Started ServerConnector@1e8f1300{HTTP/1.1,[http/1.1]}{0.0.0.0:4040}
20/06/26 07:03:07 WARN org.apache.spark.scheduler.FairSchedulableBuilder: Fair Scheduler configuration file not found so jobs will be scheduled in FIFO order. To use fair scheduling, configure pools in fairscheduler.xml or set spark.scheduler.allocation.file to a file that contains the configuration.
20/06/26 07:03:09 INFO org.apache.hadoop.yarn.client.RMProxy: Connecting to ResourceManager at kmu-cluster-m/10.146.0.9:8032
20/06/26 07:03:09 INFO org.apache.hadoop.yarn.client.AHSProxy: Connecting to Application History server at kmu-cluster-m/10.146.0.9:10200
20/06/26 07:03:12 INFO org.apache.hadoop.yarn.client.api.impl.YarnClientImpl: Submitted application application_1593072483341_0017
20/06/26 07:03:18 INFO org.apache.hadoop.mapred.FileInputFormat: Total input files to process : 5
20/06/26 07:03:18 INFO org.apache.hadoop.mapred.FileInputFormat: Total input files to process : 240
20/06/26 07:04:07 INFO org.spark_project.jetty.server.AbstractConnector: Stopped Spark@1e8f1300{HTTP/1.1,[http/1.1]}{0.0.0.0:4040}
kkoma2623@kmu-cluster-m:~$
```

<그림 67>

hdfs dfs -ls 명령어를 통해 HDFS 에 출력 경로로 설정해둔 곳에 출력물이 생성된 것을 확인하였다.

```
File Edit View Search Terminal Help
20/06/26 07:03:07 WARN org.apache.spark.scheduler.FairSchedulableBuilder: Fair Scheduler configuration file not found so jobs will be scheduled in FIFO order. To use fair scheduling, configure pools in fairscheduler.xml or set spark.scheduler.allocation.file to a file that contains the configuration.
20/06/26 07:03:09 INFO org.apache.hadoop.yarn.client.RMProxy: Connecting to ResourceManager at kmu-cluster-m/10.146.0.9:8032
20/06/26 07:03:09 INFO org.apache.hadoop.yarn.client.AHSProxy: Connecting to Application History server at kmu-cluster-m/10.146.0.9:10200
20/06/26 07:03:12 INFO org.apache.hadoop.yarn.client.api.impl.YarnClientImpl: Submitted application application_1593072483341_0017
20/06/26 07:03:18 INFO org.apache.hadoop.mapred.FileInputFormat: Total input files to process : 5
20/06/26 07:03:18 INFO org.apache.hadoop.mapred.FileInputFormat: Total input files to process : 240
20/06/26 07:04:07 INFO org.spark_project.jetty.server.AbstractConnector: Stopped Spark@1e8f1300{HTTP/1.1,[http/1.1]}{0.0.0.0:4040}
kkoma2623@kmu-cluster-m:~$ hdfs dfs -ls
Found 8 items
drwxr-xr-x  - kkoma2623 hadoop      0 2020-06-26 07:04 .sparkStaging
-rw-r--r--  2 kkoma2623 hadoop  8031 2020-06-26 04:43 bd_hadoop.jar
-rw-r--r--  2 kkoma2623 hadoop  29497 2020-06-26 05:27 bd_spark.jar
-rw-r--r--  2 kkoma2623 hadoop 854597439 2020-06-26 04:39 soc-LiveJournal1.txt
drwxr-xr-x  - kkoma2623 hadoop      0 2020-06-26 04:55 task10outputs
drwxr-xr-x  - kkoma2623 hadoop      0 2020-06-26 05:17 task20outputs
drwxr-xr-x  - kkoma2623 hadoop      0 2020-06-26 06:19 task30outputs
drwxr-xr-x  - kkoma2623 hadoop      0 2020-06-26 07:04 task40outputs
kkoma2623@kmu-cluster-m:~$
```

<그림 68>

hdfs dfs -ls task4Outputs 명령어를 통해 task4Outputs 디렉터리 안에 생성된 파일들을 확인하였다.

```
File Edit View Search Terminal Help
kkoma2623@kmu-cluster-m:~$ hdfs dfs -ls task4outputs
Found 121 items
-rw-r--r-- 2 kkoma2623 hadoop 0 2020-06-26 07:04 task4outputs/_SUCCESS
-rw-r--r-- 2 kkoma2623 hadoop 250180 2020-06-26 07:03 task4outputs/part-00000
-rw-r--r-- 2 kkoma2623 hadoop 255062 2020-06-26 07:03 task4outputs/part-00001
-rw-r--r-- 2 kkoma2623 hadoop 263255 2020-06-26 07:03 task4outputs/part-00002
-rw-r--r-- 2 kkoma2623 hadoop 276200 2020-06-26 07:03 task4outputs/part-00003
-rw-r--r-- 2 kkoma2623 hadoop 295444 2020-06-26 07:03 task4outputs/part-00004
-rw-r--r-- 2 kkoma2623 hadoop 317981 2020-06-26 07:03 task4outputs/part-00005
-rw-r--r-- 2 kkoma2623 hadoop 340162 2020-06-26 07:03 task4outputs/part-00006
-rw-r--r-- 2 kkoma2623 hadoop 364987 2020-06-26 07:03 task4outputs/part-00007
-rw-r--r-- 2 kkoma2623 hadoop 389062 2020-06-26 07:03 task4outputs/part-00008
-rw-r--r-- 2 kkoma2623 hadoop 414294 2020-06-26 07:03 task4outputs/part-00009
-rw-r--r-- 2 kkoma2623 hadoop 430932 2020-06-26 07:03 task4outputs/part-00010
-rw-r--r-- 2 kkoma2623 hadoop 454135 2020-06-26 07:03 task4outputs/part-00011
-rw-r--r-- 2 kkoma2623 hadoop 467936 2020-06-26 07:03 task4outputs/part-00012
-rw-r--r-- 2 kkoma2623 hadoop 476632 2020-06-26 07:03 task4outputs/part-00013
-rw-r--r-- 2 kkoma2623 hadoop 479917 2020-06-26 07:03 task4outputs/part-00014
-rw-r--r-- 2 kkoma2623 hadoop 479048 2020-06-26 07:03 task4outputs/part-00015
-rw-r--r-- 2 kkoma2623 hadoop 477882 2020-06-26 07:03 task4outputs/part-00016
-rw-r--r-- 2 kkoma2623 hadoop 467654 2020-06-26 07:03 task4outputs/part-00017
-rw-r--r-- 2 kkoma2623 hadoop 452827 2020-06-26 07:03 task4outputs/part-00018
-rw-r--r-- 2 kkoma2623 hadoop 436770 2020-06-26 07:03 task4outputs/part-00019
-rw-r--r-- 2 kkoma2623 hadoop 418311 2020-06-26 07:03 task4outputs/part-00020
```

<그림 69>

hdfs dfs -cat task4Outputs/part-00119 명령어를 통해 생성된 파일의 일부를 출력해보겠다.

```
File Edit View Search Terminal Help
-rw-r--r-- 2 kkoma2623 hadoop 366691 2020-06-26 07:04 task4outputs/part-00097
-rw-r--r-- 2 kkoma2623 hadoop 391463 2020-06-26 07:04 task4outputs/part-00098
-rw-r--r-- 2 kkoma2623 hadoop 412455 2020-06-26 07:04 task4outputs/part-00099
-rw-r--r-- 2 kkoma2623 hadoop 436719 2020-06-26 07:04 task4outputs/part-00100
-rw-r--r-- 2 kkoma2623 hadoop 454333 2020-06-26 07:04 task4outputs/part-00101
-rw-r--r-- 2 kkoma2623 hadoop 467401 2020-06-26 07:04 task4outputs/part-00102
-rw-r--r-- 2 kkoma2623 hadoop 477678 2020-06-26 07:04 task4outputs/part-00103
-rw-r--r-- 2 kkoma2623 hadoop 481295 2020-06-26 07:04 task4outputs/part-00104
-rw-r--r-- 2 kkoma2623 hadoop 483020 2020-06-26 07:04 task4outputs/part-00105
-rw-r--r-- 2 kkoma2623 hadoop 475454 2020-06-26 07:04 task4outputs/part-00106
-rw-r--r-- 2 kkoma2623 hadoop 468157 2020-06-26 07:04 task4outputs/part-00107
-rw-r--r-- 2 kkoma2623 hadoop 454504 2020-06-26 07:04 task4outputs/part-00108
-rw-r--r-- 2 kkoma2623 hadoop 436942 2020-06-26 07:04 task4outputs/part-00109
-rw-r--r-- 2 kkoma2623 hadoop 416001 2020-06-26 07:04 task4outputs/part-00110
-rw-r--r-- 2 kkoma2623 hadoop 392413 2020-06-26 07:04 task4outputs/part-00111
-rw-r--r-- 2 kkoma2623 hadoop 370410 2020-06-26 07:04 task4outputs/part-00112
-rw-r--r-- 2 kkoma2623 hadoop 348643 2020-06-26 07:04 task4outputs/part-00113
-rw-r--r-- 2 kkoma2623 hadoop 325910 2020-06-26 07:04 task4outputs/part-00114
-rw-r--r-- 2 kkoma2623 hadoop 302232 2020-06-26 07:04 task4outputs/part-00115
-rw-r--r-- 2 kkoma2623 hadoop 284832 2020-06-26 07:04 task4outputs/part-00116
-rw-r--r-- 2 kkoma2623 hadoop 270725 2020-06-26 07:04 task4outputs/part-00117
-rw-r--r-- 2 kkoma2623 hadoop 258821 2020-06-26 07:04 task4outputs/part-00118
-rw-r--r-- 2 kkoma2623 hadoop 252666 2020-06-26 07:04 task4outputs/part-00119
kkoma2623@kmu-cluster-m:~$ hdfs dfs -cat task4outputs/part-00119
```

<그림 70>

다음과 같이 출력이 되는 것을 확인할 수 있다.

```
File Edit View Search Terminal Help
19667  0.14285715
2103328 0.2
3105640 0.6666667
1751122 0.12270531
1152969 0.16666667
1383576 0.23809524
1327776 0.16507937
1637844 0.33333334
53984   0.064313725
1068459 0.6666667
799556  0.019607844
2653111 1.0
3330271 0.42857143
3320173 0.33333334
1369716 0.16666667
1274676 0.32666665
2022247 0.8
2018521 0.115384616
1175739 0.33333334
1702252 0.1904762
3041515 0.18947369
1356846 0.33333334
57746   0.32019705
kkoma2623@kmu-cluster-m:~$
```

<그림 71>