

# Hadoop WordCount

빅데이터 시스템

- Scale-out으로 확장

Job의 병렬 처리로 성능 향상 –  
Tracker로 활동 여부 확인

HDFS(Hadoop Distributed File  
System): 데이터 수집을 분산하여  
저장.

MapReduce : HDFS에 저장된 데이  
터를 분산 처리하는 시스템.

1. map
2. reduce
3. jobConf : WordCount
4. jar파일로 export
5. hadoop시작
6. 분석할파일을 넣을 hdfs상의 폴더를 만든다.
7. 분석할파일을 6번에서 만든 폴더로 copy
8. hadoop 프로그램 시작 jar을 이용
9. Hive로 해당 파일 WordCount 분석
10. Windows Dbeaver와 Excel을 통해 전처리
11. Windows Python 에서 시각화

# Properties에서 Hadoop library 경로 확인하여 불러오기

The screenshot illustrates the steps to configure Hadoop libraries in an Eclipse IDE project named 'hadoop01'.

**Left Panel: Properties for hadoop01**

- Java Build Path** tab is selected.
- Libraries** sub-tab is active.
- Current libraries: `hadoop-common-2.9.1.jar`, `hadoop-mapreduce-client-core-2.9.1.jar`, and `JRE System Library [JavaSE-1.8]`.
- The **Add External JARs...** button is highlighted.

**Top Right Panel: JAR Selection**

This panel shows the file explorer for the `usr/local/hadoop/share/hadoop/mapreduce` directory.

이름	크기	종류	수정
lib			16 4월 2018
lib-examples			16 4월 2018
sources			16 4월 2018
hadoop-mapreduce-client-app-2.9.1.jar	573.2 kB	Archive	16 4월 2018
hadoop-mapreduce-client-common-2.9.1.jar	787.9 kB	Archive	16 4월 2018
<b>hadoop-mapreduce-client-core-2.9.1.jar</b>	<b>1.6 MB</b>	<b>Archive</b>	<b>16 4월 2018</b>
hadoop-mapreduce-client-hs-2.9.1.jar	200.6 kB	Archive	16 4월 2018
hadoop-mapreduce-client-hs-plugins-2.9.1.jar	32.8 kB	Archive	16 4월 2018
hadoop-mapreduce-client-jobclient-2.9.1.jar	71.4 kB	Archive	16 4월 2018
hadoop-mapreduce-client-jobclient-2.9.1-tests.jar	1.6 MB	Archive	16 4월 2018
hadoop-mapreduce-client-shuffle-2.9.1.jar	85.2 kB	Archive	16 4월 2018

**Bottom Right Panel: JAR Selection**

This panel shows the file explorer for the `usr/local/hadoop/share/hadoop/common` directory.

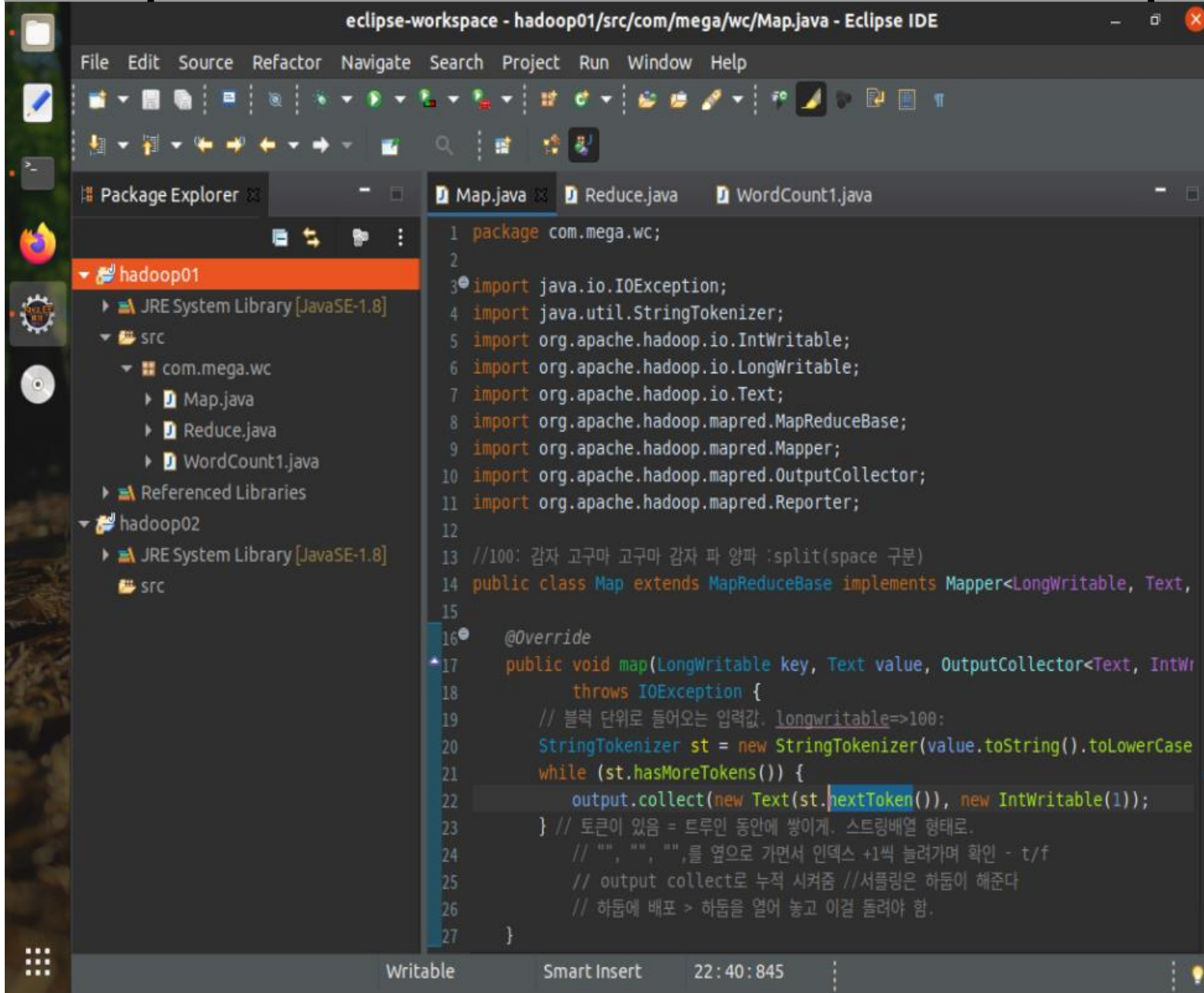
이름	크기	종류	수정
jdifff			16 4월 2018
lib			16 4월 2018
sources			16 4월 2018
templates			16 4월 2018
<b>hadoop-common-2.9.1.jar</b>	<b>3.9 MB</b>	<b>Archive</b>	<b>16 4월 2018</b>
hadoop-common-2.9.1-tests.jar	2.6 MB	Archive	16 4월 2018
hadoop-nfs-2.9.1.jar	188.7 kB	Archive	16 4월 2018

**Right Panel: Package Explorer**

- Project: `hadoop01`
- Source: `JRE System Library [JavaSE-1.8]`
- Source: `src`
- Package: `com.mega.wc`
- Referenced Libraries**

  - `hadoop-common-2.9.1.jar - /usr/local/hadoop/share/hadoop/common`
  - `hadoop-mapreduce-client-core-2.9.1.jar - /usr/local/hadoop/share/hadoop/mapreduce`

# MapReduce 프로그래밍 - Map클래스 작성



```
1 package com.mega.wc;
2
3 import java.io.IOException;
4 import java.util.StringTokenizer;
5 import org.apache.hadoop.io.IntWritable;
6 import org.apache.hadoop.io.LongWritable;
7 import org.apache.hadoop.io.Text;
8 import org.apache.hadoop.mapred.MapReduceBase;
9 import org.apache.hadoop.mapred.Mapper;
10 import org.apache.hadoop.mapred.OutputCollector;
11 import org.apache.hadoop.mapred.Reporter;
12
13 //100: 감자 고구마 고구마 감자 파 양파 :split(space 구분)
14 public class Map extends MapReduceBase implements Mapper<LongWritable, Text,
15
16     @Override
17     public void map(LongWritable key, Text value, OutputCollector<Text, IntWr
18         throws IOException {
19         // 블록 단위로 들어오는 입력값. longwritable=>100:
20         StringTokenizer st = new StringTokenizer(value.toString().toLowerCase
21         while (st.hasMoreTokens()) {
22             output.collect(new Text(st.nextToken()), new IntWritable(1));
23         } // 토큰이 있음 = 트루인 동안에 쌓이게. 스트림배열 형태로.
24         // "", "", "", 를 옆으로 가면서 인덱스 +1씩 늘려가며 확인 - t/f
25         // output collect로 누적 시켜줌 //서플링은 하둡이 해준다
26         // 하둡에 배포 > 하둡을 열어 놓고 이걸 돌려야 함.
27     }
```

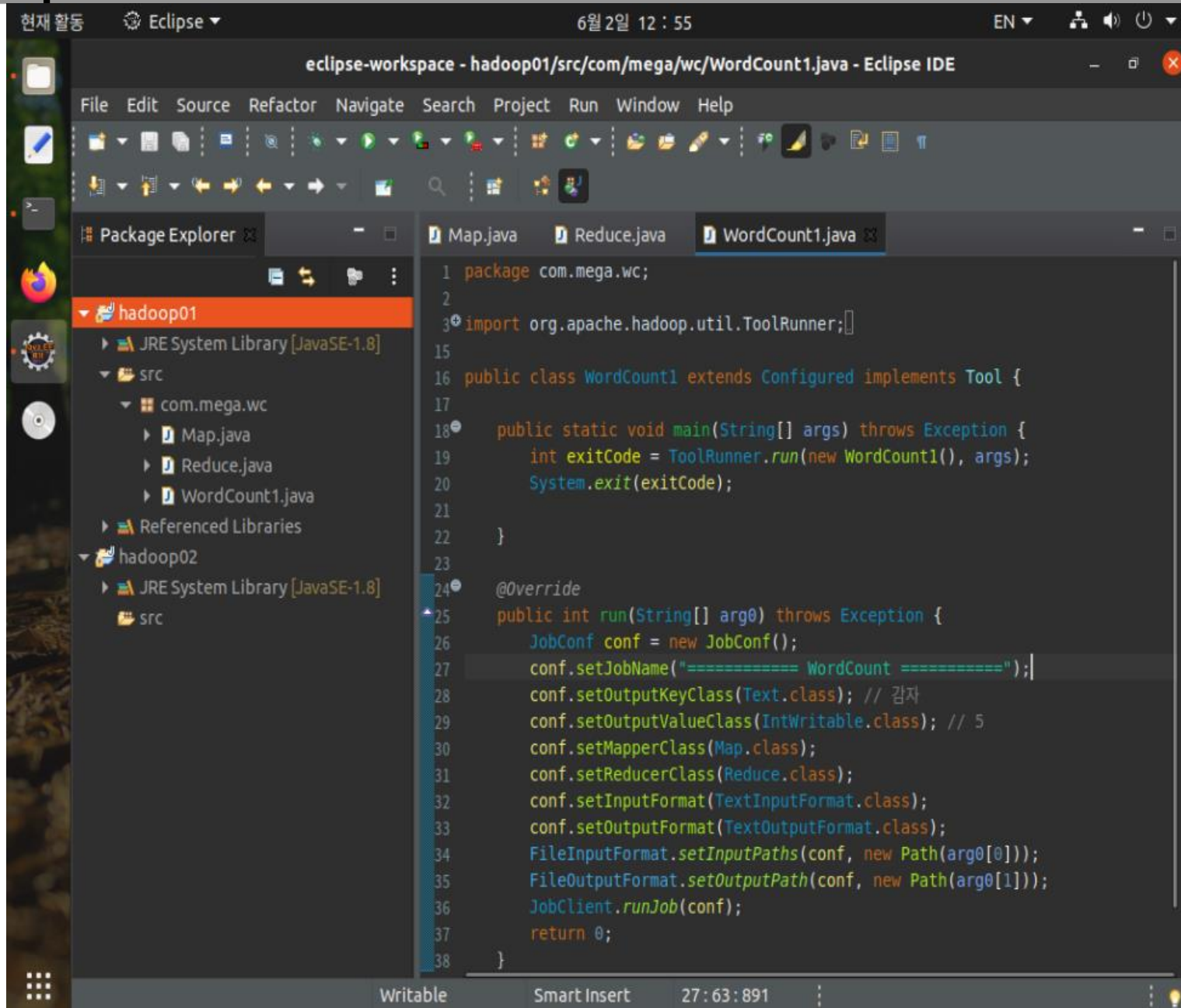
# MapReduce 프로그래밍 – Reduce 클래스 작성

The screenshot shows the Eclipse IDE interface. The top bar indicates the current date and time as 6월 2일 12:55. The title bar reads 'eclipse-workspace - hadoop01/src/com/mega/wc/Reduce.java - Eclipse IDE'. The menu bar includes File, Edit, Source, Refactor, Navigate, Search, Project, Run, Window, and Help. The Package Explorer on the left shows the project structure: hadoop01 (containing JRE System Library [JavaSE-1.8], src, com.mega.wc (containing Map.java, Reduce.java, WordCount1.java), and Referenced Libraries), and hadoop02 (containing JRE System Library [JavaSE-1.8] and src). The main editor displays the code for Reduce.java, which implements the Reducer interface. The code includes imports for java.io.IOException, java.util.Iterator, org.apache.hadoop.io.IntWritable, org.apache.hadoop.io.Text, org.apache.hadoop.mapred.MapReduceBase, org.apache.hadoop.mapred.OutputCollector, org.apache.hadoop.mapred.Reducer, and org.apache.hadoop.mapred.Reporter. The class Reduc extends MapReduceBase and implements Reducer<Text, IntWritable>. The reduce method is annotated with @Override and throws IOException. It contains comments in Korean explaining the logic: '서플링한 결과가 들어옴 : 결과를 반복적으로 꺼내오기 위해 Iterator.' and '//리듀스 하는 사람의 의도에 따라 가중치 두거나 평균 낼 수도 있음'. The method initializes a counter 'cnt' to 0, iterates over the 'values' using 'values.hasNext()', increments 'cnt' by 'values.next().get()', and finally collects the result using 'output.collect(key, new IntWritable(cnt))'.

```
1 package com.mega.wc;
2
3 import java.io.IOException;
4 import java.util.Iterator;
5 import org.apache.hadoop.io.IntWritable;
6 import org.apache.hadoop.io.Text;
7 import org.apache.hadoop.mapred.MapReduceBase;
8 import org.apache.hadoop.mapred.OutputCollector;
9 import org.apache.hadoop.mapred.Reducer;
10 import org.apache.hadoop.mapred.Reporter;
11
12 public class Reduc extends MapReduceBase implements Reducer<Text, IntWritable> {
13
14     @Override
15     public void reduce(Text key, Iterator<IntWritable> values, OutputCollector
16         throws IOException {
17         // 서플링한 결과가 들어옴 : 결과를 반복적으로 꺼내오기 위해 Iterator.
18         //리듀스 하는 사람의 의도에 따라 가중치 두거나 평균 낼 수도 있음
19         int cnt = 0;
20         while (values.hasNext()) {
21             cnt += values.next().get();
22         }
23         output.collect(key, new IntWritable(cnt));
24     }
25 }
26 }
27 }
```



# MapReduce 프로그래밍 – WordCount 클래스 작성



The screenshot shows the Eclipse IDE interface. The Package Explorer on the left displays the project structure: hadoop01 (containing JRE System Library [JavaSE-1.8] and src) and hadoop02 (containing JRE System Library [JavaSE-1.8] and src). The src folder of hadoop01 contains the com.mega.wc package, which includes Map.java, Reduce.java, and WordCount1.java. The main editor displays the code for WordCount1.java. The code defines a public class WordCount1 that extends Configured and implements Tool. It includes a main method that calls ToolRunner.run and a run method that configures the job and runs it.

```
1 package com.mega.wc;
2
3 import org.apache.hadoop.util.ToolRunner;
4
15
16 public class WordCount1 extends Configured implements Tool {
17
18     public static void main(String[] args) throws Exception {
19         int exitCode = ToolRunner.run(new WordCount1(), args);
20         System.exit(exitCode);
21     }
22
23
24     @Override
25     public int run(String[] arg0) throws Exception {
26         JobConf conf = new JobConf();
27         conf.setJobName("===== WordCount =====");
28         conf.setOutputKeyClass(Text.class); // 감자
29         conf.setOutputValueClass(IntWritable.class); // 5
30         conf.setMapperClass(Map.class);
31         conf.setReducerClass(Reduce.class);
32         conf.setInputFormat(TextInputFormat.class);
33         conf.setOutputFormat(TextOutputFormat.class);
34         FileInputFormat.setInputPaths(conf, new Path(arg0[0]));
35         FileOutputFormat.setOutputPath(conf, new Path(arg0[1]));
36         JobClient.runJob(conf);
37         return 0;
38     }
39 }
```

- JobConf  
: WordCount

# Java파일 재설정, jar로 export

The image shows the Eclipse IDE interface with the Package Explorer on the left. The project 'hadoop03' is selected, and the 'src' folder contains 'WordCount3.java'. The 'Export' dialog is open, showing the 'JAR file' option selected under the 'Java' category. The 'JAR File Specification' dialog is also open, showing the 'hadoop03' project selected for export. The 'Export generated class files and resources' checkbox is checked. The 'JAR file' field shows the path '/usr/local/hadoop/jar/wordcount3.jar'. The 'Options' section has 'Compress the contents of the JAR file' checked.

The terminal window on the right shows the Hadoop startup process. It displays the command 'start-all.sh' and the output of the script, which includes the startup of the NameNode, SecondaryNameNode, and DataNode. The output shows the Hadoop configuration and the startup of the NameNode, SecondaryNameNode, and DataNode.

```
hadoop@hadoop: /usr/local/hadoop
파일(F) 편집(E) 보기(V) 검색(S) 터미널(T) 도움말(H)
hdfs/namenode/current/fsimage.ckpt_00000000000000000000 of size 322 bytes saved in 0
seconds .
21/06/04 10:36:27 INFO namenode.NNStorageRetentionManager: Going to retain 1 images
with txid >= 0
21/06/04 10:36:27 INFO namenode.NameNode: SHUTDOWN_MSG:
/*****
SHUTDOWN_MSG: Shutting down NameNode at hadoop/127.0.1.1
*****/
hadoop@hadoop:/usr/local/hadoop$ start-all.sh
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
21/06/04 10:36:35 WARN util.NativeCodeLoader: Unable to load native-hadoop library f
or your platform... using builtin-java classes where applicable
Starting namenodes on [localhost]
hadoop@localhost's password:
localhost: starting namenode, logging to /usr/local/hadoop/logs/hadoop-hadoop-nameno
de-hadoop.out
hadoop@localhost's password:
localhost: starting datanode, logging to /usr/local/hadoop/logs/hadoop-hadoop-datano
de-hadoop.out
Starting secondary namenodes [localhost]
hadoop@localhost's password:
localhost: starting secondarynamenode, logging to /usr/local/hadoop/logs/hadoop-hado
op-secondarynamenode-hadoop.out
21/06/04 10:37:04 WARN util.NativeCodeLoader: Unable to load native-hadoop library f
or your platform... using builtin-java classes where applicable
starting yarn daemons
starting resourcemanager, logging to /usr/local/hadoop/logs/yarn-hadoop-resourcemana
ger-hadoop.out
hadoop@localhost's password:
localhost: starting nodemanager, logging to /usr/local/hadoop/logs/yarn-hadoop-nodem
anager-hadoop.out
hadoop@hadoop:/usr/local/hadoop$ jps
11666 ResourceManager
12052 Jps
2757 org.eclipse.equinox.launcher_1.5.700.v20200207-2156.jar
11510 SecondaryNameNode
11848 NodeManager
11083 NameNode
11278 DataNode
hadoop@hadoop:/usr/local/hadoop$
```



# hadoop 시작&확인, hdfs 폴더 생성

```
ubuntu@ubuntu-VirtualBox: /usr/local/hadoop
21/06/02 12:49:31 INFO util.GSet: 0.029999999329447746% max memory 966.7 MB = 297.0 KB
21/06/02 12:49:31 INFO util.GSet: capacity = 2^15 = 32768 entries
21/06/02 12:49:31 INFO namenode.FSImage: Allocated new BlockPoolId: BP-404838968-127.0.1.1-1622605771088
21/06/02 12:49:31 INFO common.Storage: Storage directory /usr/local/hadoop/hdfs/namenode has been successfully formatted.
21/06/02 12:49:31 INFO namenode.FSImageFormatProtobuf: Saving image file /usr/local/hadoop/hdfs/namenode/current/fsimage.ckpt_000000000000000000 using no compression
21/06/02 12:49:31 INFO namenode.FSImageFormatProtobuf: Image file /usr/local/hadoop/hdfs/namenode/current/fsimage.ckpt_000000000000000000 of size 323 bytes saved in 0 seconds .
21/06/02 12:49:31 INFO namenode.NNStorageRetentionManager: Going to retain 1 images with txid >= 0
21/06/02 12:49:31 INFO namenode.NameNode: SHUTDOWN_MSG:
/*****
SHUTDOWN_MSG: Shutting down NameNode at ubuntu-VirtualBox/127.0.1.1
*****/
ubuntu@ubuntu-VirtualBox: /usr/local/hadoop$ start-all.sh
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
21/06/02 12:56:37 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Starting namenodes on [localhost]
ubuntu@localhost's password:
localhost: starting namenode, logging to /usr/local/hadoop/logs/hadoop-ubuntu-namenode-ubuntu-VirtualBox.out
ubuntu@localhost's password:
localhost: starting datanode, logging to /usr/local/hadoop/logs/hadoop-ubuntu-datanode-ubuntu-VirtualBox.out
Starting secondary namenodes [localhost]
ubuntu@localhost's password:
localhost: starting secondarynamenode, logging to /usr/local/hadoop/logs/hadoop-ubuntu-secondarynamenode-ubuntu-VirtualBox.out
21/06/02 12:57:04 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
starting yarn daemons
starting resourcemanager, logging to /usr/local/hadoop/logs/yarn-ubuntu-resourcemanager-ubuntu-VirtualBox.out
ubuntu@localhost's password:
localhost: starting nodemanager, logging to /usr/local/hadoop/logs/yarn-ubuntu-nodemanager-ubuntu-VirtualBox.out
ubuntu@ubuntu-VirtualBox: /usr/local/hadoop$
```

하둡 전체 시작

```
ubuntu@ubuntu-VirtualBox: /usr/local/hadoop$ jps
9872 NameNode
10035 DataNode
10804 Jps
10230 SecondaryNameNode
10359 ResourceManager
10668 NodeManager
4285 org.eclipse.equinox.launcher_1.6.100.v20201223-0822.jar
ubuntu@ubuntu-VirtualBox: /usr/local/hadoop$
```

하둡의 셸 커맨드(Hadoop fs -명령어)를 이용하여 HDFS에 접근한다.

```
ubuntu@ubuntu-VirtualBox: /usr/local/hadoop$ ls
LICENSE.txt  bin          eclipse-workspace  hadoop-2.9.1.tar.gz  jar      logs  snap  템플릿
NOTICE.txt   eclipse      etc                hdfs                  lib      sbin  tmp
README.txt   eclipse-installer  hadoop-2.9.1      include               libexec  share  공개
ubuntu@ubuntu-VirtualBox: /usr/local/hadoop$ hadoop fs -mkdir -p /wordcount/input2
21/06/02 12:58:31 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
ubuntu@ubuntu-VirtualBox: /usr/local/hadoop$ hadoop fs -copyFromLocal /usr/local/hadoop/README.txt /wordcount/input2
21/06/02 13:00:57 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
```



# MapReduce 프레임워크에서 WordCount jar 전송, job 실행

```
ubuntu@ubuntu-VirtualBox: /usr/local/hadoop$ hadoop jar jar/WordCount1.jar com.mega.wc.WordCount1 /wordcount/input2 wordcount/output2
```

```
21/06/02 13:03:18 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
21/06/02 13:03:19 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
21/06/02 13:03:19 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
21/06/02 13:03:19 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with processName=JobTracker, sessionId=
```

```
ubuntu@ubuntu-VirtualBox: /usr/local/hadoop$
```

```
21/06/02 13:03:21 INFO mapreduce.Job: Job job_local351488205_0001 running on uber mode : false
21/06/02 13:03:21 INFO mapreduce.Job: map 0% reduce 0%
21/06/02 13:03:22 INFO mapred.LocalJobRunner:
21/06/02 13:03:22 INFO mapred.MapTask: Starting flush of map output
21/06/02 13:03:22 INFO mapred.MapTask: Spilling map output
21/06/02 13:03:22 INFO mapred.MapTask: bufstart = 0; bufend = 2055; bufvoid = 104857600
21/06/02 13:03:22 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26213684(104854736); length = 713/6553600
21/06/02 13:03:22 INFO mapred.MapTask: Finished spill 0
21/06/02 13:03:22 INFO mapred.Task: Task attempt_local351488205_0001_m_000000_0 is done. And is in the process of committing
21/06/02 13:03:22 INFO mapred.LocalJobRunner: hdfs://localhost:62350/wordcount/input2/README.txt:0+1366
21/06/02 13:03:22 INFO mapred.Task: Task 'attempt_local351488205_0001_m_000000_0' done.
21/06/02 13:03:22 INFO mapred.LocalJobRunner: Finishing task: attempt_local351488205_0001_m_000000_0
21/06/02 13:03:22 INFO mapred.LocalJobRunner: map task executor complete.
21/06/02 13:03:22 INFO mapred.LocalJobRunner: Waiting for reduce tasks
21/06/02 13:03:22 INFO mapred.LocalJobRunner: Starting task: attempt_local351488205_0001_r_000000_0
21/06/02 13:03:22 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
21/06/02 13:03:22 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
21/06/02 13:03:22 INFO mapred.Task: Using ResourceCalculatorProcessTree : [ ]
21/06/02 13:03:22 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task.reduce.Shuffle@13397f15
21/06/02 13:03:22 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=363285696, maxSingleShuffleLimit=98821424, mergeThreshold=239768576, ioSortFactor=10, memToMemMergeOutputsThreshold=10
21/06/02 13:03:22 INFO reduce.EventFetcher: attempt_local351488205_0001_r_000000_0 Thread started: EventFetcher for fetching Map Completion Events
21/06/02 13:03:22 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local351488205_0001_m_000000_0 decomp: 2415 len: 2419 to MEMORY
21/06/02 13:03:22 INFO reduce.InMemoryMapOutput: Read 2415 bytes from map-output for attempt_local351488205_0001_m_000000_0
21/06/02 13:03:22 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 2415, inMemoryMapOutputs.size() -> 1, commitMemory -> 0, usedMemory -> 2415
21/06/02 13:03:22 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning
21/06/02 13:03:22 INFO mapred.LocalJobRunner: 1 / 1 copied.
21/06/02 13:03:22 INFO reduce.MergeManagerImpl: finalMerge called with 1 in-memory map-outputs and 0 on-disk map-outputs
21/06/02 13:03:22 INFO mapred.Merger: Merging 1 sorted segments
21/06/02 13:03:22 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 240
```

```
ubuntu@ubuntu-VirtualBox: /usr/local/hadoop$
```

```
21/06/02 13:03:22 INFO mapred.Merger: Merging 1 sorted segments
21/06/02 13:03:22 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 2406 bytes
21/06/02 13:03:22 INFO mapred.LocalJobRunner: 1 / 1 copied.
21/06/02 13:03:22 INFO mapred.Task: Task attempt_local351488205_0001_r_000000_0 is done. And is in the process of committing
21/06/02 13:03:22 INFO mapred.LocalJobRunner: 1 / 1 copied.
21/06/02 13:03:22 INFO mapred.Task: Task attempt_local351488205_0001_r_000000_0 is allowed to commit now
21/06/02 13:03:22 INFO output.FileOutputCommitter: Saved output of task 'attempt_local351488205_0001_r_000000_0' to hdfs://localhost:62350/user/ubuntu/wordcount/output2/_temporary/0/task_local351488205_0001_r_000000
21/06/02 13:03:22 INFO mapred.LocalJobRunner: reduce > reduce
21/06/02 13:03:22 INFO mapred.Task: Task 'attempt_local351488205_0001_r_000000_0' done.
21/06/02 13:03:22 INFO mapred.LocalJobRunner: Finishing task: attempt_local351488205_0001_r_000000_0
21/06/02 13:03:22 INFO mapred.LocalJobRunner: reduce task executor complete.
21/06/02 13:03:22 INFO mapreduce.Job: map 100% reduce 100%
21/06/02 13:03:23 INFO mapreduce.Job: Job job_local351488205_0001 completed successfully
21/06/02 13:03:24 INFO mapreduce.Job: Counters: 35

File System Counters
  FILE: Number of bytes read=5202
  FILE: Number of bytes written=939935
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=2732
  HDFS: Number of bytes written=1239
  HDFS: Number of read operations=13
  HDFS: Number of large read operations=0
  HDFS: Number of write operations=4

Map-Reduce Framework
  Map input records=31
  Map output records=179
  Map output bytes=2055
  Map output materialized bytes=2419
  Input split bytes=102
  Combine input records=0
  Combine output records=0
  Reduce input groups=123
  Reduce shuffle bytes=2419
  Reduce input records=179
  Reduce output records=123
  Spilled Records=358
  Shuffled Maps =1
  Failed Shuffles=0
  Merged Map outputs=1
  GC time elapsed (ms)=57
  Total committed heap usage (bytes)=457318400

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=1366

File Output Format Counters
  Bytes Written=1239
```

```
ubuntu@ubuntu-VirtualBox: /usr/local/hadoop$
```

```
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=2732
HDFS: Number of bytes written=1239
HDFS: Number of read operations=13
HDFS: Number of large read operations=0
HDFS: Number of write operations=4

Map-Reduce Framework
  Map input records=31
  Map output records=179
  Map output bytes=2055
  Map output materialized bytes=2419
  Input split bytes=102
  Combine input records=0
  Combine output records=0
  Reduce input groups=123
  Reduce shuffle bytes=2419
  Reduce input records=179
  Reduce output records=123
  Spilled Records=358
  Shuffled Maps =1
  Failed Shuffles=0
  Merged Map outputs=1
  GC time elapsed (ms)=57
  Total committed heap usage (bytes)=457318400

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=1366

File Output Format Counters
  Bytes Written=1239
ubuntu@ubuntu-VirtualBox: /usr/local/hadoop$
```



# WordCount의 결과 확인(cat 명령어 사용)

```
ubuntu@ubuntu-VirtualBox: /usr/local/hadoop
ing builtin-java classes where applicable
cat: '/wordcount/output2/part-00000': No such file or directory
ubuntu@ubuntu-VirtualBox: /usr/local/hadoop$ hadoop fs -cat wordcount/output2/part-00000
21/06/02 13:08:40 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... us
ing builtin-java classes where applicable
(bis), 1
(eccn) 1
(see 1
(tsu) 1
5d002.c.1, 1
740.13) 1
<http://www.wassenaar.org/> 1
about 1
administration 1
algorithms. 1
and 6
and/or 1
another 1
any 1
apache 1
as 1
asymmetric 1
at: 2
before 1
bis 1
both 1
bureau 1
by 1
check 1
classified 1
code 1
code. 1
commerce, 1
commodity 1
concerning 1
control 1
core 1
country 1
country's 1
please 2
policies 1
possession, 2
project 1
provides 1
re-export 2
regulations 1
regulations, 1
reside 1
restrictions 1
section 1
security 2
see 2
software 4
software, 2
software. 2
software: 1
source 1
ssl 1
technology 1
the 12
this 4
to 2
u.s. 1
under 1
unrestricted 1
use, 2
uses 1
using 2
visit 1
website 1
which 2
wiki, 1
with 1
written 1
you 1
your 1
ubuntu@ubuntu-VirtualBox: /usr/loc
```

# 분석한 파일을 지정 폴더로 copy

```
ubuntu@ubuntu-VirtualBox: /usr/local/hadoop/output2
written 1
you 1
your 1
ubuntu@ubuntu-VirtualBox:/usr/local/hadoop$ hadoop fs -get wordcount/output2
21/06/02 13:09:48 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
ubuntu@ubuntu-VirtualBox:/usr/local/hadoop$ ls
LICENSE.txt  bin          eclipse-workspace  hadoop-2.9.1.tar.gz  jar  logs  share  공개
NOTICE.txt   eclipse      etc                hdfs                 lib  output2  snap  템플릿
README.txt   eclipse-installer  hadoop-2.9.1      include              libexec  sbin    tmp
ubuntu@ubuntu-VirtualBox:/usr/local/hadoop$ cd output2
ubuntu@ubuntu-VirtualBox:/usr/local/hadoop/output2$ ls
_SUCCESS part-00000
ubuntu@ubuntu-VirtualBox:/usr/local/hadoop/output2$ cat part-00000
(bis), 1
(eccn) 1
(see 1
(tsu) 1
5d002.c.1, 1
740.13) 1
<http://www.wassenaar.org/> 1
about 1
administration 1
algorithms. 1
and 6
and/or 1
another 1
any 1
apache 1
as 1
asymmetric 1
at: 2
before 1
bis 1
both 1
bureau 1
by 1
check 1
```

```
please 2
policies 1
possession, 2
project 1
provides 1
re-export 2
regulations 1
regulations, 1
reside 1
restrictions 1
section 1
security 2
see 2
software 4
software, 2
software. 2
software: 1
source 1
ssl 1
technology 1
the 12
this 4
to 2
u.s. 1
under 1
unrestricted 1
use, 2
uses 1
using 2
visit 1
website 1
which 2
wiki, 1
with 1
written 1
you 1
your 1
ubuntu@ubuntu-VirtualBox:/usr/local
```

# 분석결과 보고서 확인

열기(O) ▼

🔍

열기(O) ▼

🔍

part-00000  
/usr/local/hadoop/output2

1 (bis), 1

2 (eccn) 1

3 (see 1

4 (tsu) 1

5 5d002.c.1, 1

6 740.13) 1

7 <http://www.wassenaar.org/> 1

8 about 1

9 administration 1

10 algorithms. 1

11 and 6

12 and/or 1

13 another 1

14 any 1

15 apache 1

16 as 1

17 asymmetric 1

18 at: 2

19 before 1

20 bis 1

21 both 1

22 bureau 1

23 by 1

24 check 1

25 classified 1

26 code 1

27 code. 1

28 commerce, 1

29 commodity 1

30 concerning 1

31 control 1

32 core 1

33 country 1

34 country's 1

35 country, 1

36 cryptographic 3

37 currently 1

1 (bis), 1

2 (eccn) 1

3 (see 1

4 (tsu) 1

5 5d002.c.1, 1

6 740.13) 1

7 <http://www.wassenaar.org/> 1

8 about 1

9 administration 1

10 algorithms. 1

11 and 6

12 and/or 1

13 another 1

14 any 1

15 apache 1

16 as 1

17 asymmetric 1

18 at: 2

19 before 1

20 bis 1

21 both 1

22 bureau 1

23 by 1

24 check 1

25 classified 1

26 code 1

27 code. 1

28 commerce, 1

29 commodity 1

30 concerning 1

31 control 1

32 core 1

33 country 1

34 country's 1

35 country, 1

36 cryptographic 3

37 currently 1

문서 통계  
part-00000

	문서	선택한 부분
줄	123	0
단어	265	0
글자(공백 포함)	1238	0
글자(공백 없음)	993	0
바이트	1238	0

일반 텍스트 ▼

# 성공시 동일 과정으로 목표 파일 시도

열기(O) ▼

🔍

vacc.txt  
-/

저장(S)

☰

⌵

⌵

⌵

⌵

국가기간뉴스 통신사 연합뉴스

[영상] 美제공 안센 백신 내일 새벽 한국에...백악관 "한국 상황 특별"

송고시간2021-06-04 11:45

<https://youtu.be/P-0xjp6b62U>

(서울=연합뉴스) 미국 백악관은 3일(현지시간) 신종 코로나바이러스 감염증(코로나19) 백신 100만 회분이 이날 저녁 한국으로 떠난다고 밝혔습니다.

백악관은 미국 정부가 해외에 공유하겠다고 앞서 밝힌 코로나19 백신 8천만 회분 중 2천500만 회분에 대한 세부 지원계획을 발표하면서 이같이 밝혔는데요.

제프 자이언츠 백악관 코로나19 조정관은 브리핑에서 "조 바이든 대통령이 한국에 제공을 약속한 100만 회분의 안센 백신이 캘리포니아로 2천 마일을 이동, 항공기에 실려 오늘 저녁 한국으로 떠날 것"이라고 말했습니다.

코로나19 예방접종대응추진단도 앞서 101만회분의 안센 백신을 실은 군 수송기가 한국시간으로 5일 오전 1시께 성남 서울공항에 도착할 예정이라고 밝혔습니다. 이를 위해 공군 다목적 공중급유수송기인 'KC-330'이 지난 2일 김해기지에서 이륙해 미국 현지로 이동했습니다.

특히 이날 브리핑에서 한국에 대한 백신 지원에 대해 제이크 설리번 백악관 국가안보보좌관이 직접 나서 주한미군 보호에 방점을 두며 특별한 상황이라고 설명했습니다.

설리번 보좌관은 또 "미국은 백신을 받는 어떤 나라에도 어떤 것도 요청하지 않는다. 양보를 얻어내려 하지 않으며 갈취하지 않는다. (조건 부과는) 백신을 제공하는 다른 나라들이 하는 방식"이라고 말했는데 이는 중국과 러시아 등을 겨냥한 발언으로 해석됐습니다.

영상으로 보시죠.

<제작 : 황윤정·서정인>

<영상 : 연합뉴스TV·로이터>

[영상] 美제공 안센 백신 내일 새벽 한국에...백악관 "한국 상황 특별"

일반 텍스트 ▼

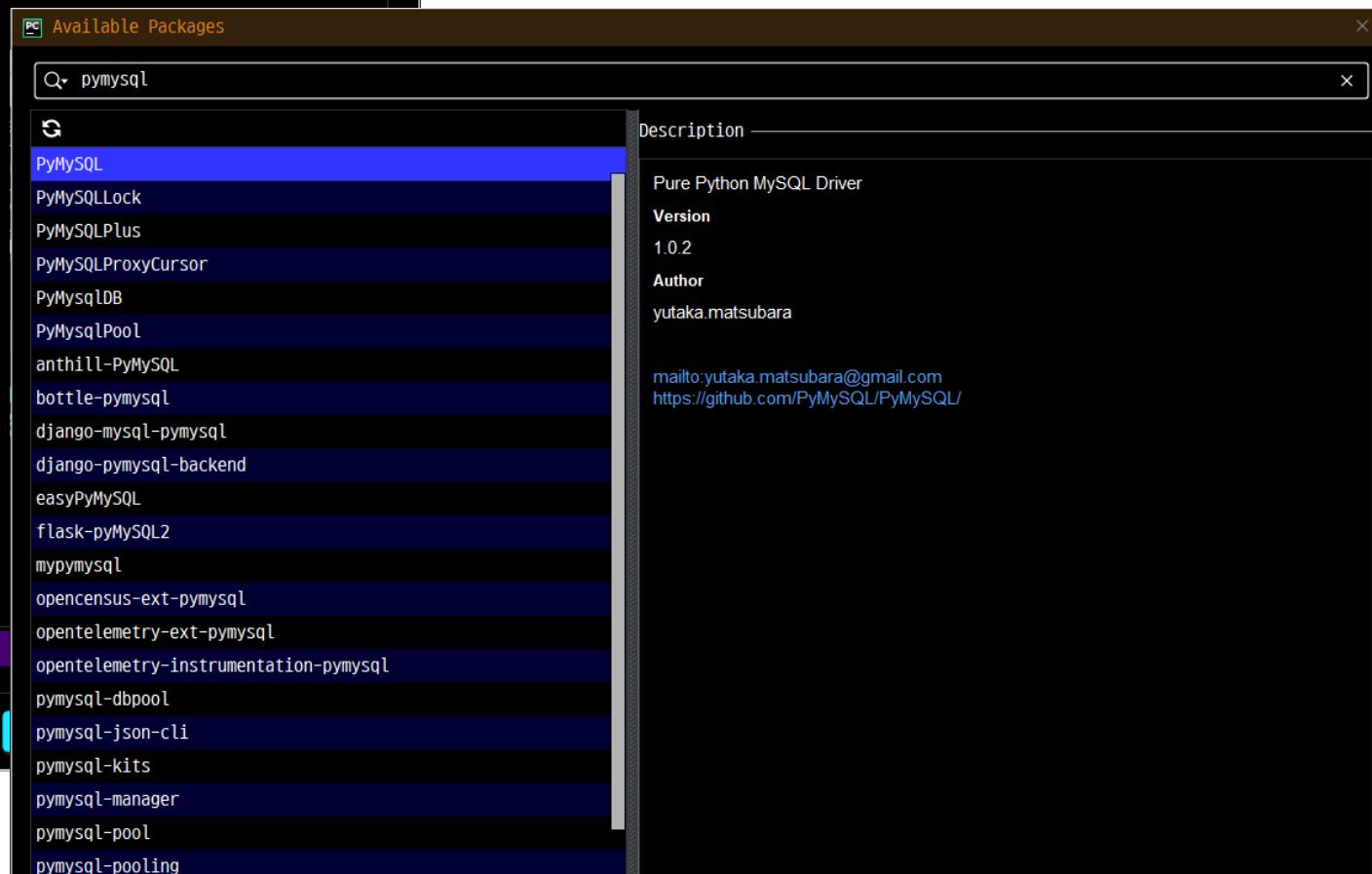
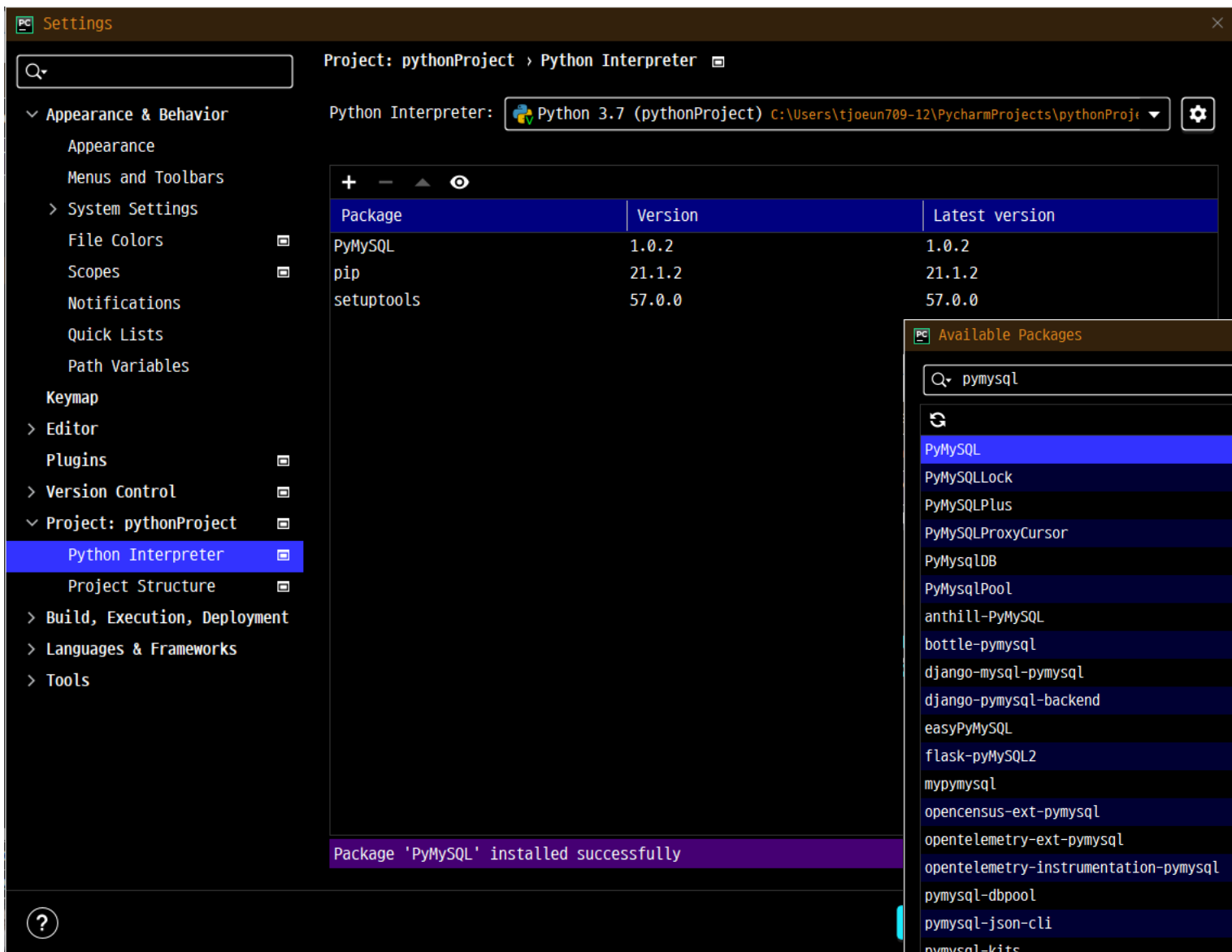
탭 너비: 8 ▼

276행, 1열 ▼

삽입



# Windows, Pymysql 패키지 설치



# 목표 파일 불러들이기

```
hadoop@hadoop:/usr/local/hadoop$ hdfs dfs -copyFromLocal /usr/local/hadoop/vacc.txt /input
21/06/04 13:13:54 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your
platform... using builtin-java classes where applicable

hadoop@hadoop:/usr/local/hadoop$ hdfs dfs -ls /input
21/06/04 13:15:11 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your
platform... using builtin-java classes where applicable
Found 1 items
-rw-r--r-- 1 hadoop supergroup 26467 2021-06-04 13:13 /input/vacc.txt

hadoop@hadoop:/usr/local/hadoop$ hadoop jar /usr/local/hadoop/share/hadoop/mapreduce/hadoop-
mapreduce-examples-2.9.1.jar wordcount /input/vacc.txt ~
21/06/04 14:38:01 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your
platform... using builtin-java classes where applicable
21/06/04 14:38:02 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs
.metrics.session-id
21/06/04 14:38:02 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker,
sessionId=
21/06/04 14:38:03 INFO input.FileInputFormat: Total input files to process : 1
21/06/04 14:38:03 INFO mapreduce.JobSubmitter: number of splits:1
21/06/04 14:38:04 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local155650229
8_0001
21/06/04 14:38:04 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
21/06/04 14:38:04 INFO mapreduce.Job: Running job: job_local1556502298_0001
21/06/04 14:38:04 INFO mapred.LocalJobRunner: OutputCommitter set in config null
21/06/04 14:38:04 INFO output.FileOutputCommitter: File Output Committer Algorithm version i
s 1
21/06/04 14:38:04 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _tempora
ry folders under output directory:false, ignore cleanup failures: false
21/06/04 14:38:04 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapreduce
.lib.output.FileOutputCommitter
21/06/04 14:38:04 INFO mapred.LocalJobRunner: Waiting for map tasks
21/06/04 14:38:04 INFO mapred.LocalJobRunner: Starting task: attempt_local1556502298_0001_m_
000000 0
```

# WordCount 기능 작동 확인

```
hadoop@hadoop: /usr/local/hadoop
파일(F) 편집(E) 보기(V) 검색(S) 터미널(T) 도움말(H)
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=52934
HDFS: Number of bytes written=21183
HDFS: Number of read operations=13
HDFS: Number of large read operations=0
HDFS: Number of write operations=4
Map-Reduce Framework
  Map input records=280
  Map output records=2565
  Map output bytes=36589
  Map output materialized bytes=27694
  Input split bytes=102
  Combine input records=2565
  Combine output records=1629
  Reduce input groups=1629
  Reduce shuffle bytes=27694
  Reduce input records=1629
  Reduce output records=1629
  Spilled Records=3258
  Shuffled Maps =1
  Failed Shuffles=0
  Merged Map outputs=1
  GC time elapsed (ms)=108
  Total committed heap usage (bytes)=428875776
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=26467
File Output Format Counters
  Bytes Written=21183
hadoop@hadoop: /usr/local/hadoop$ hdfs dfs -ls ~
21/06/04 14:38:16 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your
platform... using builtin-java classes where applicable
Found 2 items
-rw-r--r-- 1 hadoop supergroup 0 2021-06-04 14:38 /home/hadoop/_SUCCESS
-rw-r--r-- 1 hadoop supergroup 21183 2021-06-04 14:38 /home/hadoop/part-r-00000
hadoop@hadoop: /usr/local/hadoop$
```

```
hadoop@hadoop: /usr/local/hadoop$ hdfs dfs -cat ~/part-r-00000
21/06/04 14:39:49 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your
platform... using builtin-java classes where applicable
"(현재) 1
"2월에 1
"7월 1
"7월부터 1
"8월안에 1
"미국은 1
"백신 1
"수도권이 1
"예약자가 1
"위중증 1
"인구 1
"접종 1
"제일 1
"조 1
"치명률 1
"코로나19 2
"하반기 1
"한국 2
& 1
'KC-330'이 1
'교차 1
'녹색 1
'백신 1
'병상 2
'스포츠라이트' 1
'웃돈의 1
'이규연의 2
(C) 2
(JTBC 1
(monnie@kbs.co.kr) 1
(사망자 1
(서울=연합뉴스) 2
(아스트라제네카, 1
(조건 1
(코로나19 1
(환자 1
- 2
/ 1
0.0004%에 1
21-06-04 1
화이자 9
화이자, 2
화이나 1
화자로 2
화자를 1
화자와 2
확보한 1
확산 1
확산과 1
확산세가 1
확인됐다. 1
확진자 2
확진자가 1
확진자는 1
확충해야 1
환자 1
환자가 3
환자의 1
활성화하는 1
황윤정·서정인> 1
화복의 1
화복할 2
화복했나. 1
회분 1
회분, 2
회분밖 1
회분밖에 2
회분으로 1
회분을 6
회분이 6
회분입니다. 1
회의에서 1
효과 1
효과가 3
효과를 2
효과" 1
효율적으로 1
효율화 1
효율화' 1
환술면서 1
힘으로 1
"미국 1
```



# Windows에서 파일 정리(tab으로 구분>csv로)

텍스트 마법사 - 3단계 중 2단계

데이터의 구분 기호를 설정합니다. 미리 보기 상자에서 적용된 텍스트를 볼 수 있습니다.

구분 기호

☒ 탭(T)    ☐ 연속된 구분 기호를 하나로 처리(R)

☐ 세미콜론(M)    텍스트 한정자(Q): {없음} ▼

☐ 쉼표(C)

☐ 공백(S)

☐ 기타(O):

데이터 미리 보기(P)

3종미	1
3주로	1
3주로,	1
3차	3
3천	1

< >

취소 < 뒤로(B) 다음(N) > 마침(F)

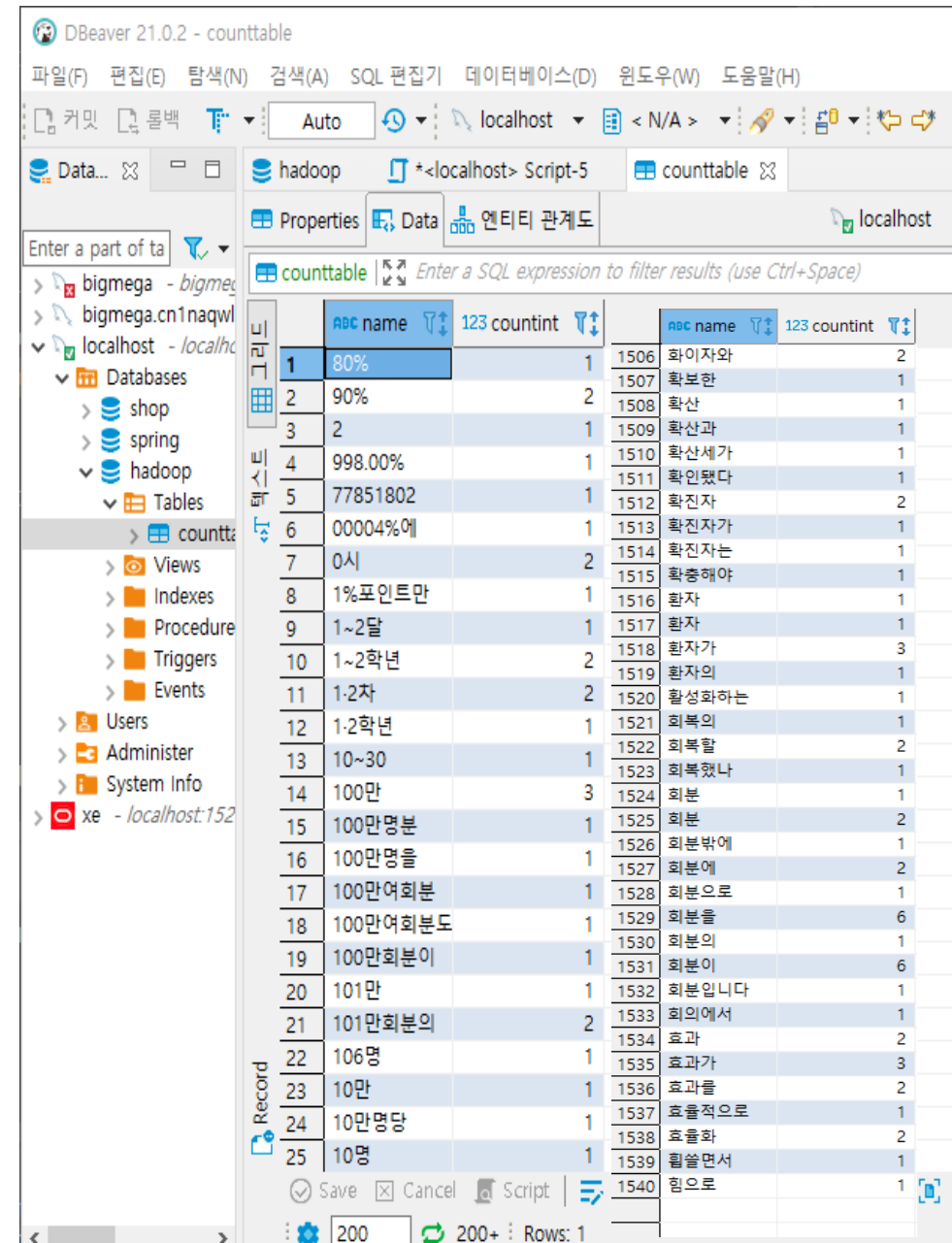
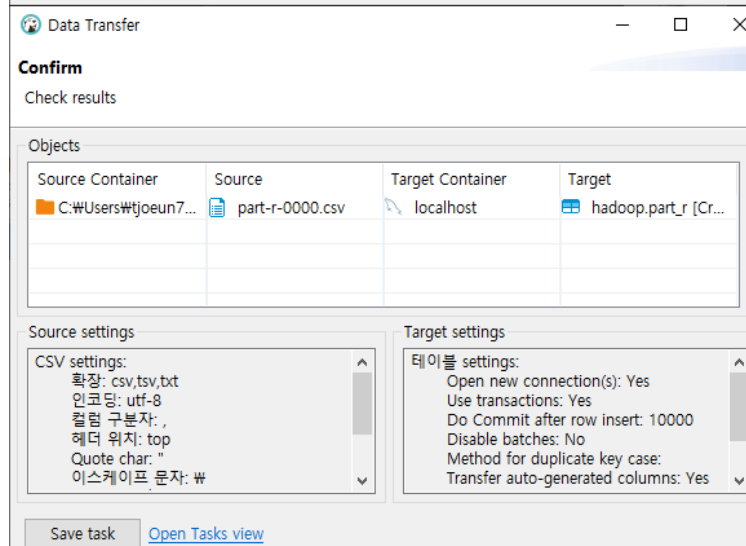
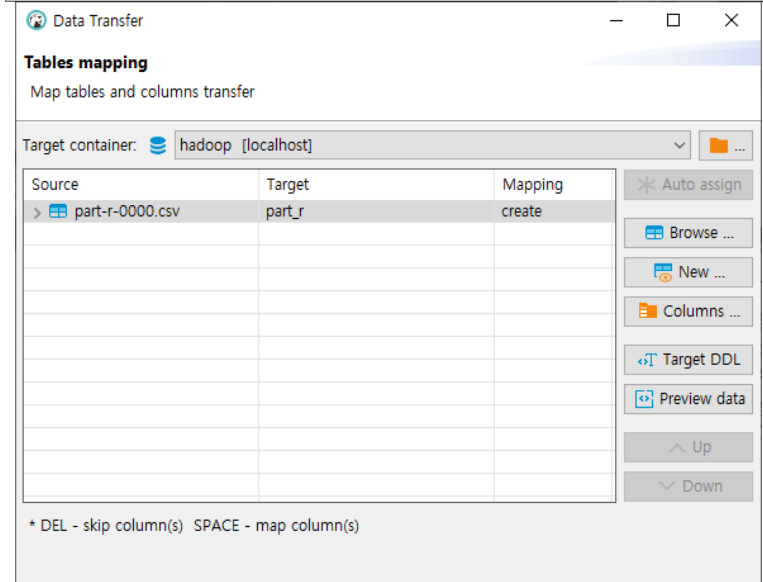
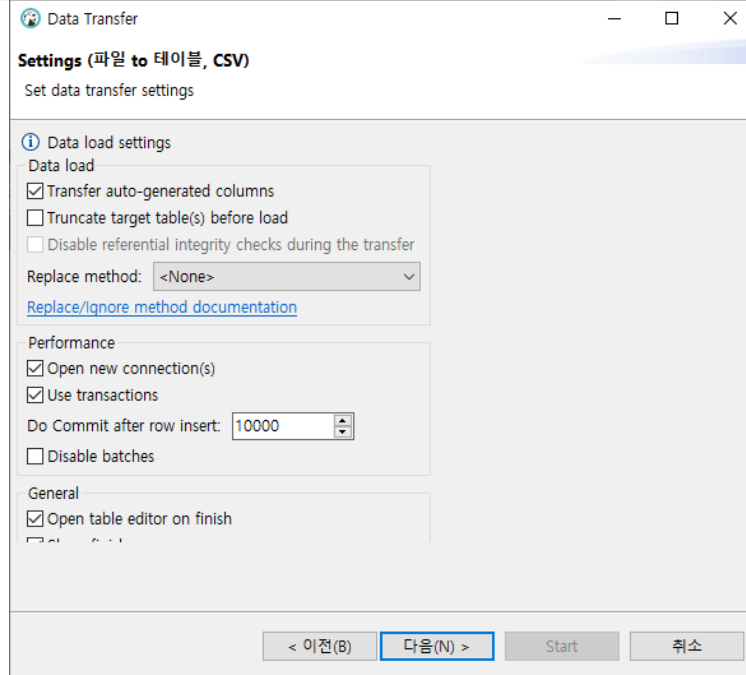
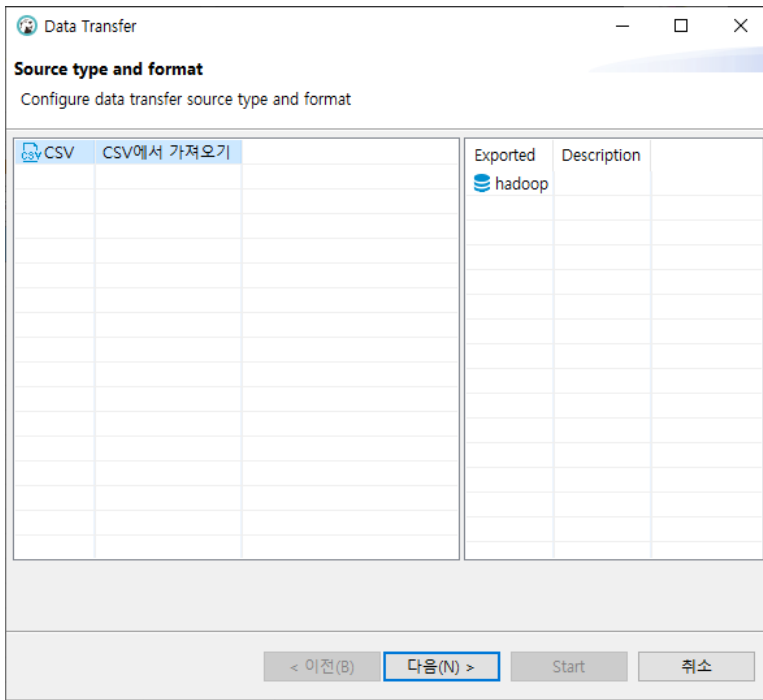
part-r-0000 - Windows 메모장

파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)

"(현재)	1	
"2월에	1	
"7월	1	
"7월부터	1	
"8월안에	1	
"미국은	1	
"백신	1	
"수도권이	1	
"예약자가	1	
"위중증	1	
"인구	1	
"접종	1	
"제일	1	
"조	1	
"치명률	1	
"코로나19	2	
"하반기	1	
"한국	2	
&	1	
'KC-330'이	1	
'교차	1	
'녹색	1	
'백신	1	
'병상	2	
'스포츠라이트'	1	
'옷돈의	1	
'이규연의	2	
(C)	2	

Ln 1, Col 1    100%    Unix (LF)    UTF-8

# DBeaver로 WordCount 결과 파일 확인



# 해당 sql문을 활용하여 python에서 시각화

localhost> Script-5

검색(A) SQL 편집기 데이터베이스(D) 윈도우(W) 도움말(H)

Auto localhost < N/A >

hadoop \*localhost> Script-5 counttable

```
select count(*) from hadoop.counttable where countint >=10
select name from hadoop.counttable where countint >=10
select max(countint) from hadoop.counttable
select * from hadoop.counttable where countint >2 order by countint desc
```

counttable 1

select name from hadoop.counttable

ABC name
등
미국
백신
백신을
수
아스트라제네카
이날
있다
접종
접종을
코로나19

Save Cancel Script

200 11 Rows: 1

KST ko 쓰기 가능 스마트 삽입 3:55:116

hadoopProject1 - findcount.py

```
import pymysql
import matplotlib.pyplot as plt
from matplotlib import font_manager, rc
rcParams['axes.unicode_minus'] = False
font_path = "/Library/Fonts/AppleGothic.ttf"
font_path = "C:/Windows/Fonts/malgun.ttf"
font_name = font_manager.FontProperties(fname=f_path).get_name()
('font', family=font_name)

def select_all():
    conn = pymysql.connect(
        user='root',
        passwd='1234',
        host='localhost',
        port=3306,
        db='hadoop',
        charset='utf8'
    )
    # print(conn.db)
    # cursor = conn.cursor(pymysql.cursors.DictCursor) #딕셔너리로 가져옴
    cursor = conn.cursor() # 튜플로 가져옴
    sql = "select * from hadoop.counttable where countint >10 order by countint desc"
    cursor.execute(sql, )
    conn.commit()
    result = cursor.fetchall()
    print(result)
    name = list()
    countint = list() # 튜플 처음 항목을 이름 리스트에 모아주고, countint를 한 리스트에
    for x, y in result:
        print(x, y)
        name.append(x)
        countint.append(y)
```

Python Packages Python Console

Installed packages: 'matplotlib' (today 오후 1:00)

findcount

```
C:\Users\tjoeun709-12\AppData\Local\Programs\Python\Python37\python.exe C:/Users/tjoeun709-12/PycharmProjects/hadoopProject1/findcount.py
 (('백신', 61), ('코로나19', 28), ('백신을', 28), ('수', 20), ('접종', 19), ('있다', 18), ('접종을', 17), ('미국', 13), ('이날', 11))
 Process finished with exit code 0
```

TODO Problems Terminal Python Packages Python Console

Installed packages: 'matplotlib' (today 오후 1:00)

