

항공 데이터 분석

환경설정

파일 다운로드 및 업로드

```
$ sudo mkdir -p /usr/local/hadoop/data  
암 호 :  
$ sudo chown -R manager:hadoop /usr/local/hadoop/data/
```

<http://stat-computing.org/dataexpo/2009/the-data.html>

로컬 사이트:	원본사이트	원본사이트	리모트 사이트:
	원본사이트	원본사이트	/usr/local/hadoop/data
	util		hadoop
	분석자료		bin
	항공데이터		data
	2008.csv		etc
	빅데이터		hdfs
	정렬		include
	03_인공지능		jar
	04_IoT		lib
	05_nodejs		libexec
파일명	크기	파일 유형	파일명
..			..
2008.csv	689,413,3...	Microsoft Exce...	2008.csv

데이터 처리

데이터 분석

```
manager@master:~$ head --help
사 용 법 : head [<옵션>]... [<파일>]...
Print the first 10 lines of each FILE to standard output.
With more than one FILE, precede each with a header giving the file name.
```


```
cd /usr/local/hadoop/data/
```

```
manager@jin-VirtualBox:~$ head -3 2008.csv
Year,Month,DayofMonth,DayOfWeek,DepTime,CRSDepTime,
apsedTime,CRSElapsedTime,AirTime,ArrDelay,DepDelay,
ode,Diverted,CarrierDelay,WeatherDelay,NASDelay,S
2008,1,3,4,2003,1955,2211,2225,WN,335,N712SW,128,
2008,1,3,4,754,735,1002,1000,WN,3231,N772SW,128,1
```


```
manager@jin-VirtualBox:~$ sed -e '1d' 2008.csv > 2008_sub.csv
manager@jin-VirtualBox:~$ head -3 2008_sub.csv
2008,1,3,4,2003,1955,2211,2225,WN,335,N712SW,128,150,116,-14,8
2008,1,3,4,754,735,1002,1000,WN,3231,N772SW,128,145,113,2,19,I
2008,1,3,4,628,620,804,750,WN,448,N428WN,96,90,76,14,8,IND,BWI
```

Map의 Key와 value

프로젝트 / 패키지 만들기 / lib 등록

 New Java Project


Create a Java Project
Create a Java project in the workspace or in an external location.



Project name:

☒ Use default location

Location:

 New Java Package


Java Package
Create a new Java package.

Creates folders corresponding to packages.

Source folder:





Name:

☐ Create package-info.java

 Properties for airline

- > Resource
- Builders
- Coverage
- Java Build Path**
- > Java Code Style
- > Java Compiler
- > Java Editor
- Javadoc Location
- Project Facets
- Project Natures

Java Build Path

 Source  Projects  Libraries  Order and Export

JARs and class folders on the build path:

- Modulepath
 - hadoop-common-2.9.2.jar - F:\₩빅데이터 연습
 - hadoop-mapreduce-client-core-2.9.2.jar - F:\₩빅데이터 연습
 - JRE System Library [JavaSE-11]
- Classpath

파일 만들기

```
$ hadoop fs -mkdir /airdata
```

```
/usr/local/hadoop/data$ head 2008.csv > 2008_head.csv
```

```
manager@master:/usr/local/hadoop/data$ hadoop fs -put 2008_head.csv /airdata
```


코드 작성 & 분석

Value 자르기

2008_head.csv 분석

로컬 사이트: F:\빅데이터 연습\

- aws
- ETC
- Samsung
- System Volume Information
- Util
- 교재 PDF
- 까치
- 빅데이터 연습
- 사진

리모트 사이트: /usr/local/hadoop/data

- hadoop
 - bin
 - data
 - etc
 - hdfs
 - include
 - jar
 - lib
 - libexec

파일명	크기	파일 유형	최종 수정
..			
2008_head.csv	1,165	Microsoft Exce...	2019-07-18 오후 ...
2008_sub_keyTe...	289	Microsoft Exce...	2019-07-18 오후 ...
Big_data_Hadoo...	2,681,344	Microsoft Pow...	2019-07-16 오전 ...
CHANGES.txt	446,615	텍스트 문서	2019-07-01 오후 ...
FileZilla_3.43.0_...	8,916,472	응용 프로그램	2019-07-16 오전 ...
hadoop-comm...	3,906,902	ALZip JAR File	2019-07-16 오전 ...
hadoop-comm...	4,092,593	ALZip JAR File	2019-07-16 오후 ...
hadoop-mapred...	1,611,944	ALZip JAR File	2019-07-16 오전 ...
hadoop-mapred...	1,656,425	ALZip JAR File	2019-07-16 오후 ...
KeyValue.jar	2,882	ALZip JAR File	2019-07-18 오후 ...

파일명	크기
..	
2008.csv	689,413,...
2008_head.csv	1,165
2008_sub.csv	689,413,...
2008_sub_head.csv	961
2008_sub_keyTest.csv	289
KeyValue.jar	2,882
Quiz.csv	45

2008_head.csv 분석

FlightNum	TailNum	ActualElap
335	N712SW	128
3231	N772SW	128
448	N428WN	96
1746	N612SW	88
3920	N464WN	90
378	N726SW	101
509	N763SW	240
535	N428WN	233
11	N689SW	95

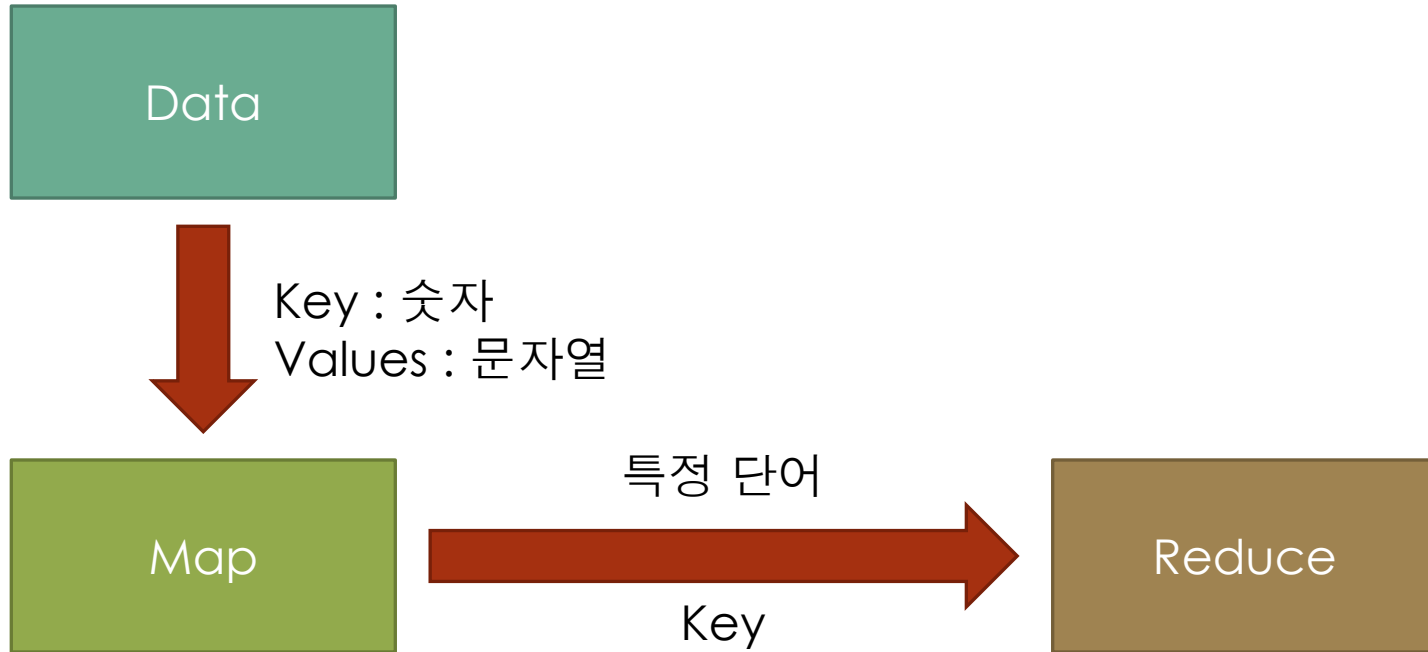
파일 수정

```
$ head -2 2008.csv > 2008_head.csv
```

```
$ hadoop fs -rm -r /airdata/2008_head.csv
```

```
$ hadoop fs -put 2008_head.csv /airdata
```

데이터 흐름



패키지 생성



New Java Package



Java Package

Create a new Java package.



Creates folders corresponding to packages.

Source folder:

Name:

☐ Create package-info.java

ValueMap

```
package com.jin.values;

import java.io.IOException;

import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapred.MapReduceBase;
import org.apache.hadoop.mapred.Mapper;
import org.apache.hadoop.mapred.OutputCollector;
import org.apache.hadoop.mapred.Reporter;

public class ValueMap extends MapReduceBase implements
Mapper<LongWritable, Text, Text, IntWritable>{
    @Override
    public void map(LongWritable key, Text value,
        OutputCollector<Text, IntWritable> output, Reporter reporter)
        throws IOException {
        String [] airData = value.toString().split(",");
        output.collect(new Text(airData[10]), new IntWritable((int)key.get()));
    }
}
```


ValueReduce

```
package com.jin.values;

import java.io.IOException;
import java.util.Iterator;

import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapred.MapReduceBase;
import org.apache.hadoop.mapred.OutputCollector;
import org.apache.hadoop.mapred.Reducer;
import org.apache.hadoop.mapred.Reporter;

public class ValueReduce extends MapReduceBase implements
Reducer<Text, IntWritable, Text, IntWritable>{
    @Override
    public void reduce(Text key, Iterator<IntWritable> values,
        OutputCollector<Text, IntWritable> output, Reporter reporter)
        throws IOException {
        output.collect(key, new IntWritable(values.next().get()));
    }
}
```

ValueMain

```
public class ValueMain extends Configured implements Tool {
    public static void main(String[] args) throws Exception {
        int exitCode = ToolRunner.run(new ValueMain(), args);
        System.exit(exitCode);
    }
    @Override
    public int run(String[] arg0) throws Exception {
        JobConf conf = new JobConf(ValueMain.class);

        conf.setJobName("value Check");

        conf.setOutputKeyClass(Text.class);
        conf.setOutputValueClass(IntWritable.class);

        conf.setMapperClass(ValueMap.class);
        conf.setReducerClass(ValueReduce.class);

        conf.setInputFormat(TextInputFormat.class);
        conf.setOutputFormat(TextOutputFormat.class);

        FileInputFormat.setInputPaths(conf, new Path(arg0[0]));
        FileOutputFormat.setOutputPath(conf, new Path(arg0[1]));

        JobClient.runJob(conf);
        return 0;
    }
}
```

Select the resources to export:

- > ☒ airline
- > ☐ WordCount

- ☐ .classpath
- ☐ .project

- ☒ Export generated class files and resources
- ☐ Export all output folders for checked projects
- ☐ Export Java source files and resources
- ☐ Export refactorings for checked projects. [Select refactorings...](#)

Select the export destination:

JAR file: F:\빅데이터 연습\ValueTest.jar

Options:

리모트 사이트: /usr/local/hadoop/jar

- hadoop
 - ? bin
 - data
 - ? etc
 - ? hdfs
 - ? include
 - jar
 - ? lib
 - libexec

파일명	크기	파일 유형
..		
CHANGES.txt	446,615	텍스트 문서
KeyValue.jar	2,882	ALZip JAR ...
TestKey.jar	3,074	ALZip JAR ...
ValueTest.jar	3,059	ALZip JAR ...
WordCount.jar	3,211	ALZip JAR ...

실행

```
manager@master:/usr/local/hadoop/data$ cd ..  
manager@master:/usr/local/hadoop$ cd jar/
```

```
manager@master:/usr/local/hadoop/jar$ hadoop jar  
ValueTest.jar com.jin.values.ValueMain /airdata/2  
008 head.csv /airdata/outputValue
```

```
$ hadoop fs -ls /airdata/outputValue
```

```
$ hadoop fs -cat /airdata/outputValue/part-00000
```

```
N428WN    974  
N464WN    684  
N612SW    589  
N689SW    1075  
N712SW    300  
N726SW    778  
N763SW    877  
N772SW    399  
TailNum  0
```

결과 분석

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Year	Month	DayofMor	DayOfWee	DepTime	CRSDepTi	ArrTime	CRSArrTim	UniqueCa	FlightNum	TailNum	ActualElap	CRSElapse
2	2008	1	3	4	2003	1955	2211	2225	WN	335	N712SW	128	150
3	2008	1	3	4	754	735	1002	1000	WN	3231	N772SW	128	145
4	2008	1	3	4	628	620	804	750	WN	448	N428WN	96	90
5	2008	1	3	4	926	930	1054	1100	WN	1746	N612SW	88	90
6	2008	1	3	4	1829	1755	1959	1925	WN	3920	N464WN	90	90
7	2008	1	3	4	1940	1915	2121	2110	WN	378	N726SW	101	115
8	2008	1	3	4	1937	1830	2037	1940	WN	509	N763SW	240	250
9	2008	1	3	4	1039	1040	1132	1150	WN	535	N428WN	233	250
10	2008	1	3	4	617	615	652	650	WN	11	N689SW	95	95

```
ture release
N428WN    974
N464WN    684
N612SW    589
N689SW    1075
N712SW    300
N726SW    778
N763SW    877
N772SW    399
TailNum 0
```

Quiz

- 아래와 같은 내용에서 4번째 값을 기준으로 결과가 출력될 수 있도록 코드를 수정하고 실행 결과를 분석하시오

1	2	3	4
0	0	0	a
0	0	0	b
0	0	0	c
0	0	0	d

```
4      0
a      9
b     18
c     27
d     36
```

Quiz

- ▶ 2005년부터 2008년까지의 모든 데이터를 하둡에 저장하고 년도별 총 운항 횟수를 구하시오.

2005	7140596
2006	7141922
2007	7453215
2008	7009728

MapReduce 동작 이해

컴파일

hadoop

jar

xxx.jar

packageName+className

args...

Main Method

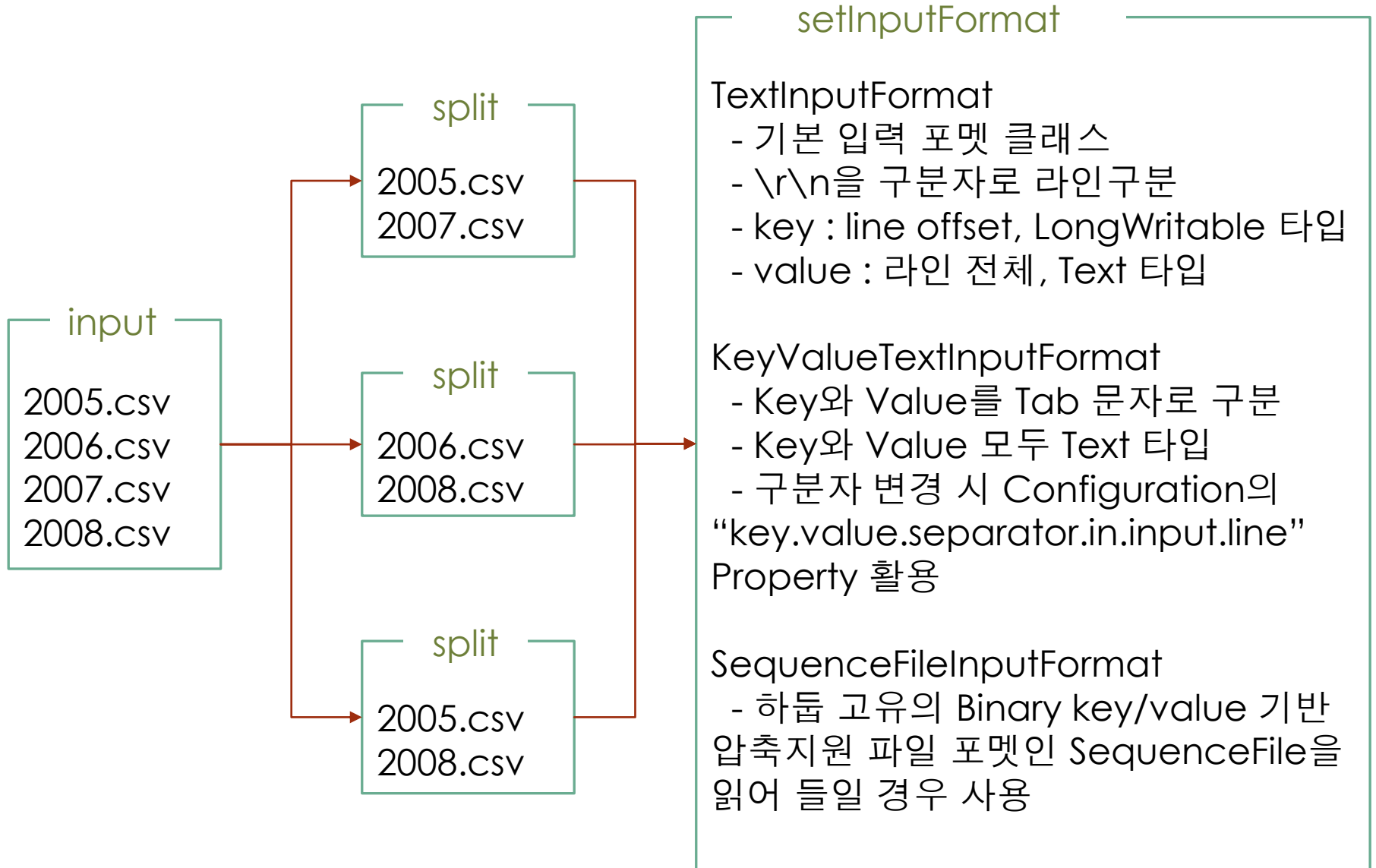
```
ToolRunner.run(  
    new MainClassName(),  
    args  
)
```

run Method

```
FileInputFormat.setInputPaths(conf, new Path(arg0[0]));  
FileOutputFormat.setOutputPath(conf, new Path(arg0[1]));
```

```
hadoop jar airline.jar  
com.jin.Ex01.MainClass  
/airdata/*.csv  
/airdata/output
```

conf.setInputFormat(TextInputFormat.class)



Mapper <k1, v1, k2, v2>

Reducer <k2, v2, k3, v3>

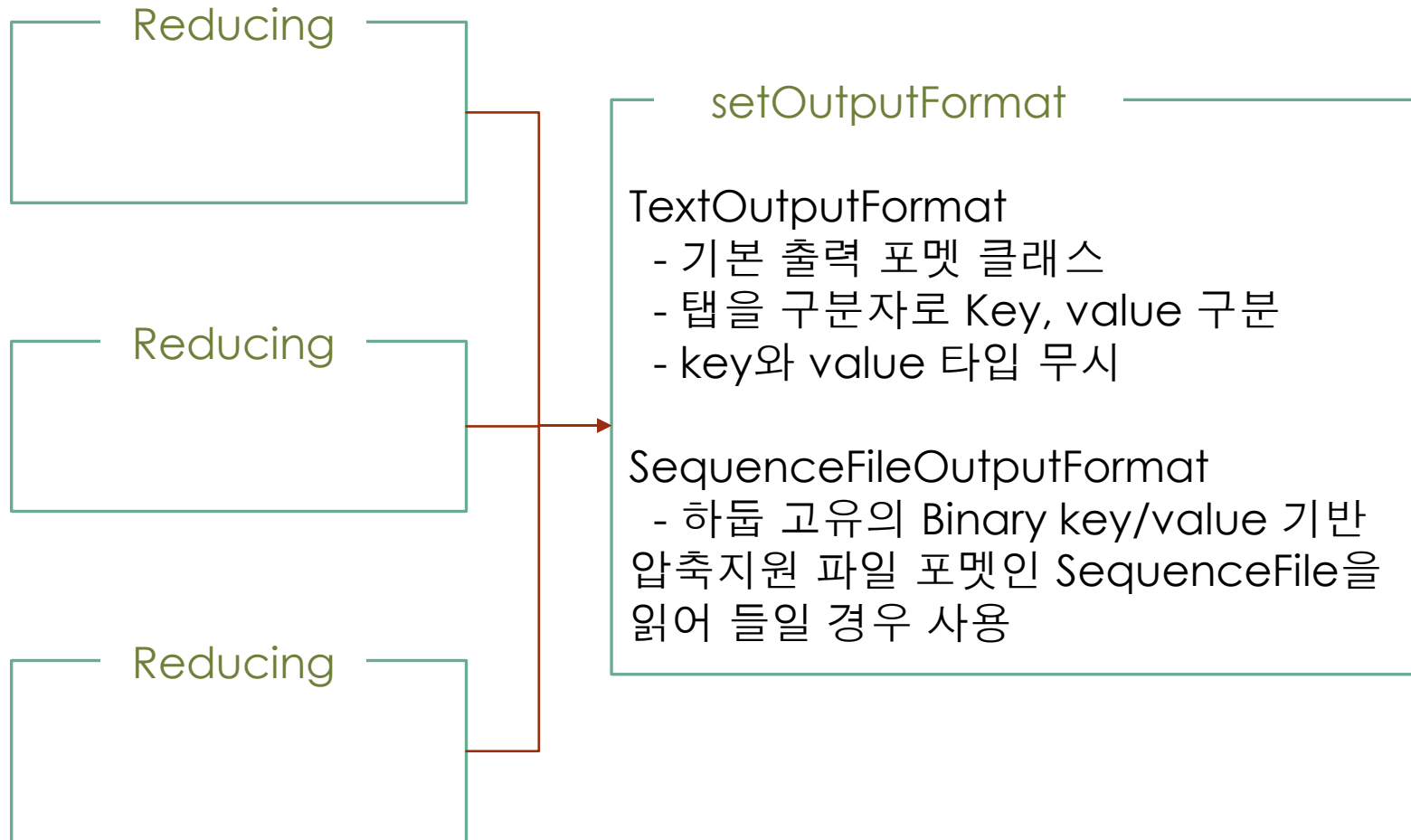
setMapOutputKeyClass(k2)

setMapOutputValueClass(v2)

setOutputKeyClass(k3)

setOutputValueClass(k3)

conf.setOutputFormat(TextInputFormat.class)



Quiz

- ▶ 매달 운항횟수를 확인해서 가장 비행을 많이 한 달과 적게 한 달을 구하고 이를 통해 이벤트를 실시하려 한다.
- ▶ 매달 운항횟수를 확인하는 프로그램을 작성하시오

```
1      2403535
10     2390627
11     2280756
12     2336198
2      2211419
3      2478056
4      2392617
5      2455623
6      2445455
7      2525696
8      2525194
9      2300285
```

Quiz

- ▶ 년도별로 구분된 데이터를 이용하여 월별 실적을 분석해 보자

2005년 10월	592712
2005년 11월	566138
2005년 12월	572343
2005년 1월	594924
2005년 2월	545332
2005년 3월	617540
2005년 4월	594492
2005년 5월	614802
2005년 6월	609195
2005년 7월	627961
2005년 8월	630904
2005년 9월	574253
2006년 10월	611718
2006년 11월	586197

분석

CSV 파일 생성

```
public class Map extends MapReduceBase implements
Mapper<LongWritable, Text, Text, IntWritable>{

    @Override

    public void map(LongWritable key, Text value,
        OutputCollector<Text, IntWritable> output, Reporter reporter)
        throws IOException {

        AirlineParser ap = new AirlineParser(value);

        output.collect(new Text(ap.getYear()+"."+ap.getMonth()+"."), new
IntWritable(1));

    }

}
```


파일 가져오기

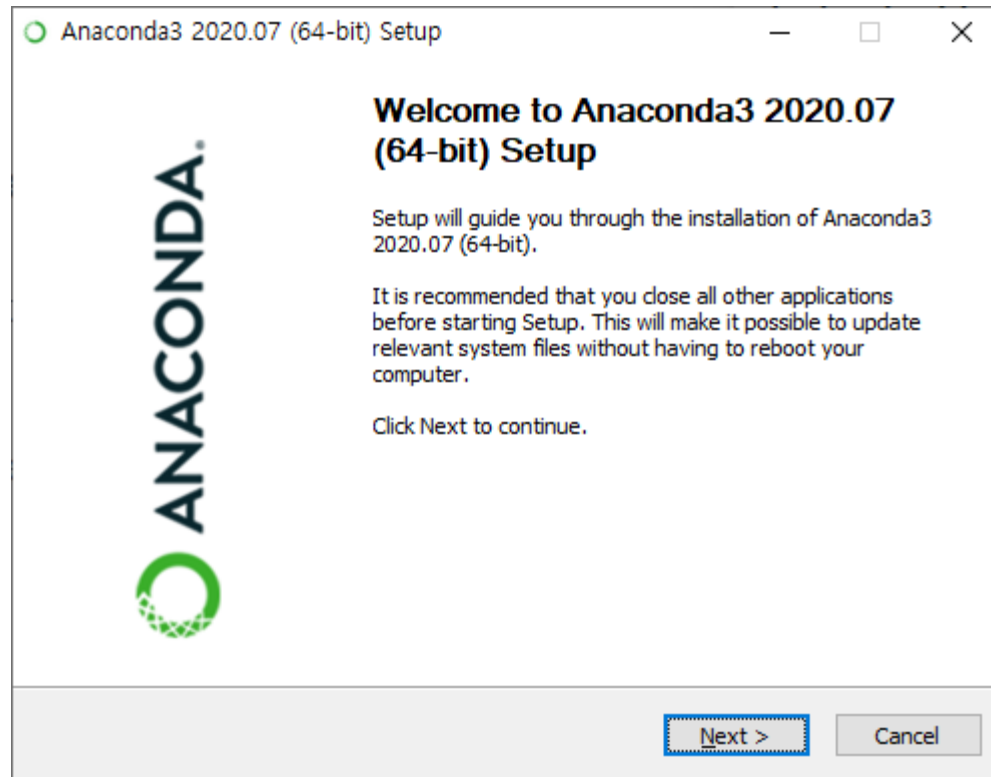
```
hadoop jar /usr/local/hadoop/jar/month.jar com.jin.Airline /airdata  
/output1
```

```
hadoop fs -ls /output1
```

```
hadoop fs -get /output1/part-00000 monthData.csv
```

아나콘다 설치

➤ <https://www.anaconda.com/products/individual>



환경설정

The image shows the Anaconda Navigator application interface. On the left, a sidebar contains a menu with '최근에 추가한 앱' (Recently added apps) and a list of applications including Spyder, Anaconda Navigator, and Anaconda Prompt. A red circle with the number '1' is placed over the Anaconda Navigator icon. The main window displays the 'Environments' tab, which shows a list of environments including 'base (root)'. A red circle with the number '2' is placed over the 'Environments' tab. A 'Create new environment' dialog box is open in the foreground. It contains fields for 'Name' (set to 'jin'), 'Location' (set to 'C:\Users\3-19\anaconda3\envs\jin'), and 'Packages' (with 'Python 3.8' selected and 'R' unselected). A red circle with the number '4' is placed over the 'Python 3.8' package selection. At the bottom of the dialog, there are 'Cancel' and 'Create' buttons. A red circle with the number '3' is placed over the 'Create' button. The bottom of the main window shows a toolbar with 'Create', 'Clone', 'Import', and 'Remove' buttons. A red circle with the number '3' is placed over the 'Create' button in the toolbar.

환경 설정

The screenshot displays the Anaconda Navigator application window. The interface includes a sidebar on the left with navigation options: Home, Environments, Learning, and Community. The main panel shows a list of environments, with 'base (root)' and 'jin' visible. A search bar and a filter dropdown (set to 'Installed') are at the top of the environment list. A red circle '1' highlights the 'jin' environment. A red circle '2' highlights the 'Open Terminal' button next to it. A red circle '3' highlights the 'Environments' sidebar item. A red circle '4' highlights the 'Open Terminal' button in the context menu. Below the environment list, a terminal window is open, showing the command prompt '(jin) C:\Users\W3-19>'. Another terminal window is shown in the foreground, displaying the command '(jin) C:\Users\W3-19>pip install matplotlib' and the output of the installation process, including 'Collecting matplotlib', 'Downloading matplotlib-3.3.2-cp38-cp38-win_amd64.whl | 8.5 MB', and 'Collecting python-dateutil>=2.1'.

Anaconda Navigator

File Help

ANACONDA NAVIGATOR

Home

Environments

Learning

Community

Search Environments

base (root)

jin

Open Terminal

Open with Python

Open with IPython

Open with Jupyter Notebook

(jin) C:\WINDOWS\system32\cmd.exe

(jin) C:\Users\W3-19>

(jin) C:\Users\W3-19>pip install matplotlib

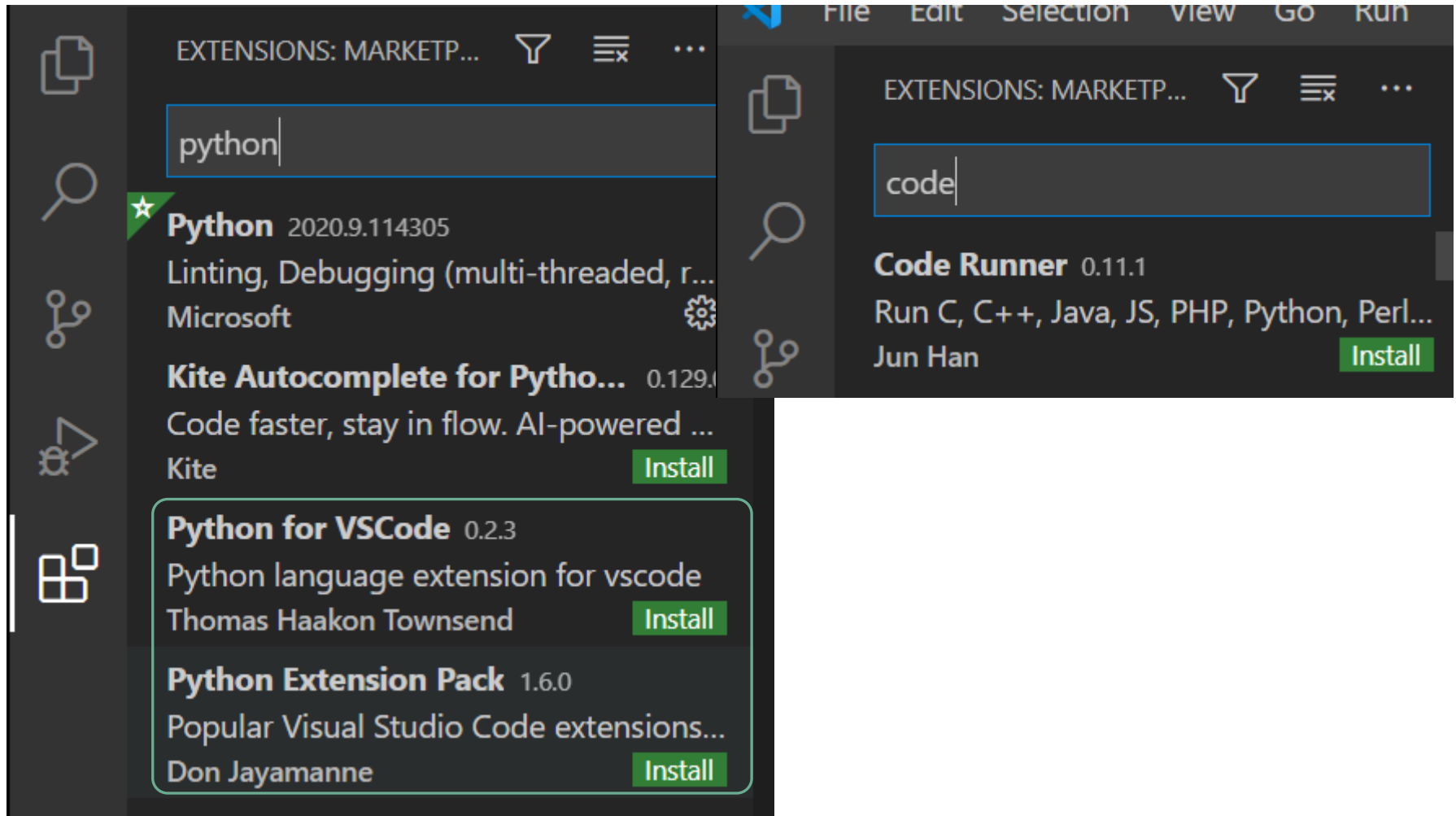
Collecting matplotlib

Downloading matplotlib-3.3.2-cp38-cp38-win_amd64.whl | 8.5 MB

Collecting python-dateutil>=2.1

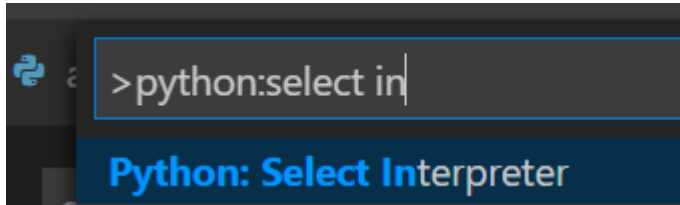
Using cached python_dateutil-2.8.1-py2.py3-no

파이썬 연동



아나콘다 연결

➡ ctrl + shift + p



Enter interpreter path...

Enter path or find an existing interpreter

Python 3.8.3 64-bit ('base': conda)

~\anaconda3\python.exe

Python 3.8.5 64-bit ('jin': conda)

~\anaconda3\envs\jin\python.exe

Python 3.9.0 64-bit

~\AppData\Local\Programs\Python\Python39\python.exe

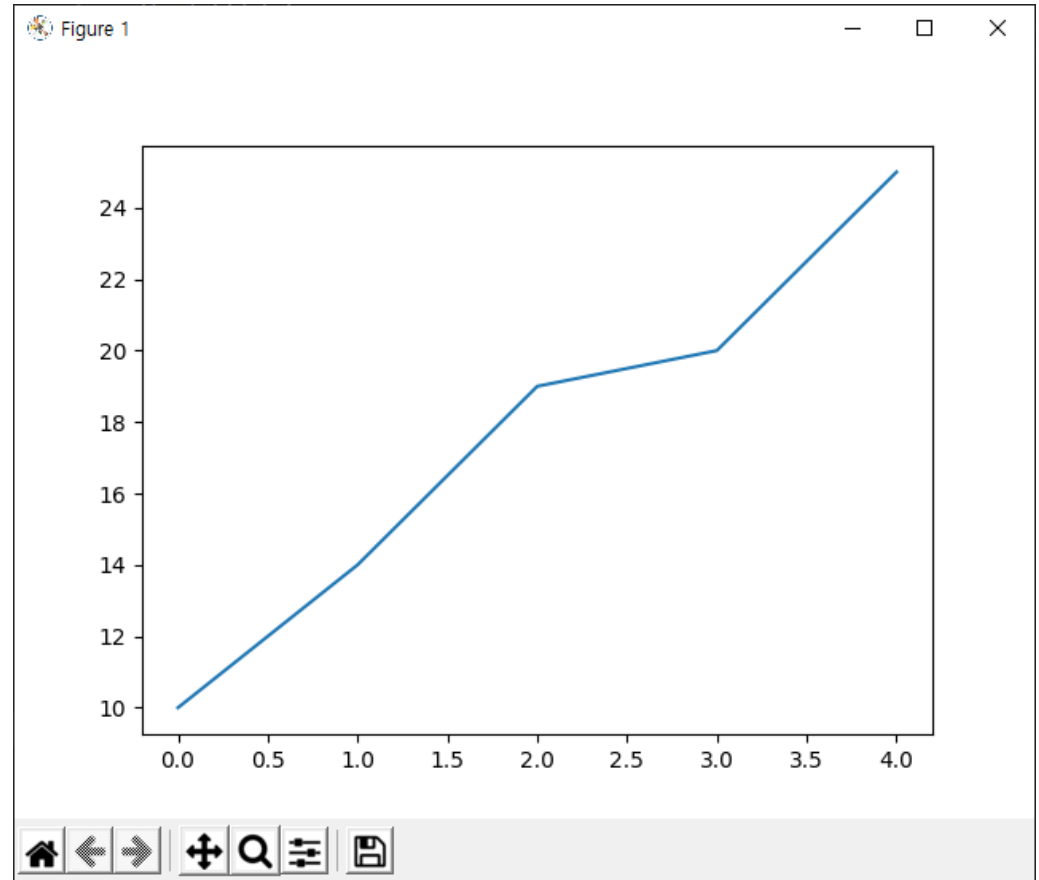
동작 확인

```
import matplotlib.pyplot as plt
```

```
data=[10, 14, 19, 20, 25]
```

```
plt.plot(data)
```

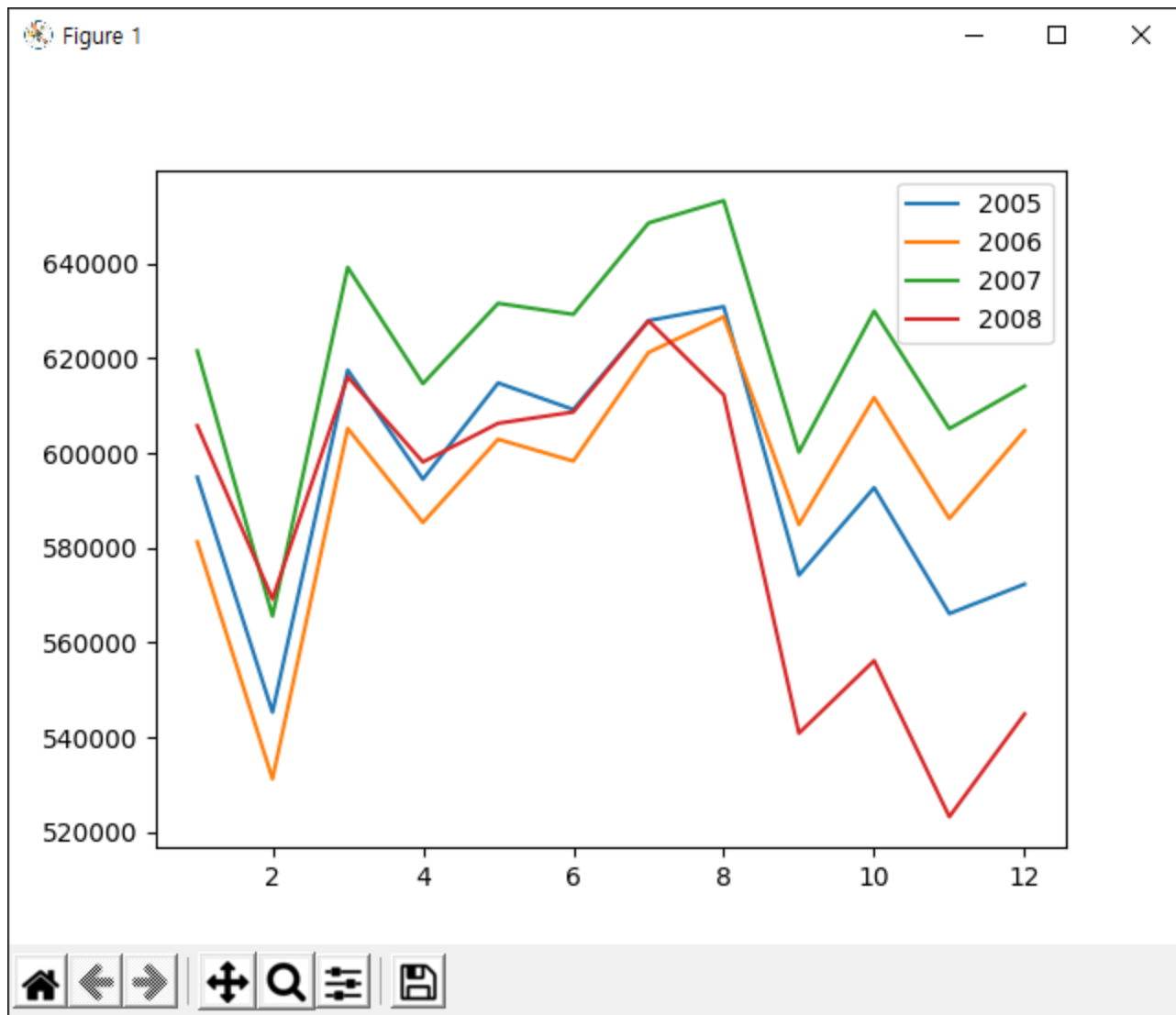
```
plt.show()
```



차트

```
import pandas as pd
import matplotlib.pyplot as plt

#데이터 추출
s1=pd.read_csv("monthData.csv", names=['year', 'month', 'data'])
#데이터 정렬
data1 = s1.set_index(['year','month']).sort_index()
#멀티인덱스 해제
data2=data1.reset_index(inplace=False)
#시각화 데이터 입력
for i in range(4):
    data3 = data2.loc[0+(i*12):11+(i*12)]
    plt.plot(data3['month'], data3['data'])
#범례 작성
plt.legend(['2005', '2006', '2007', '2008'])
#시각화
plt.show()
```

Quiz

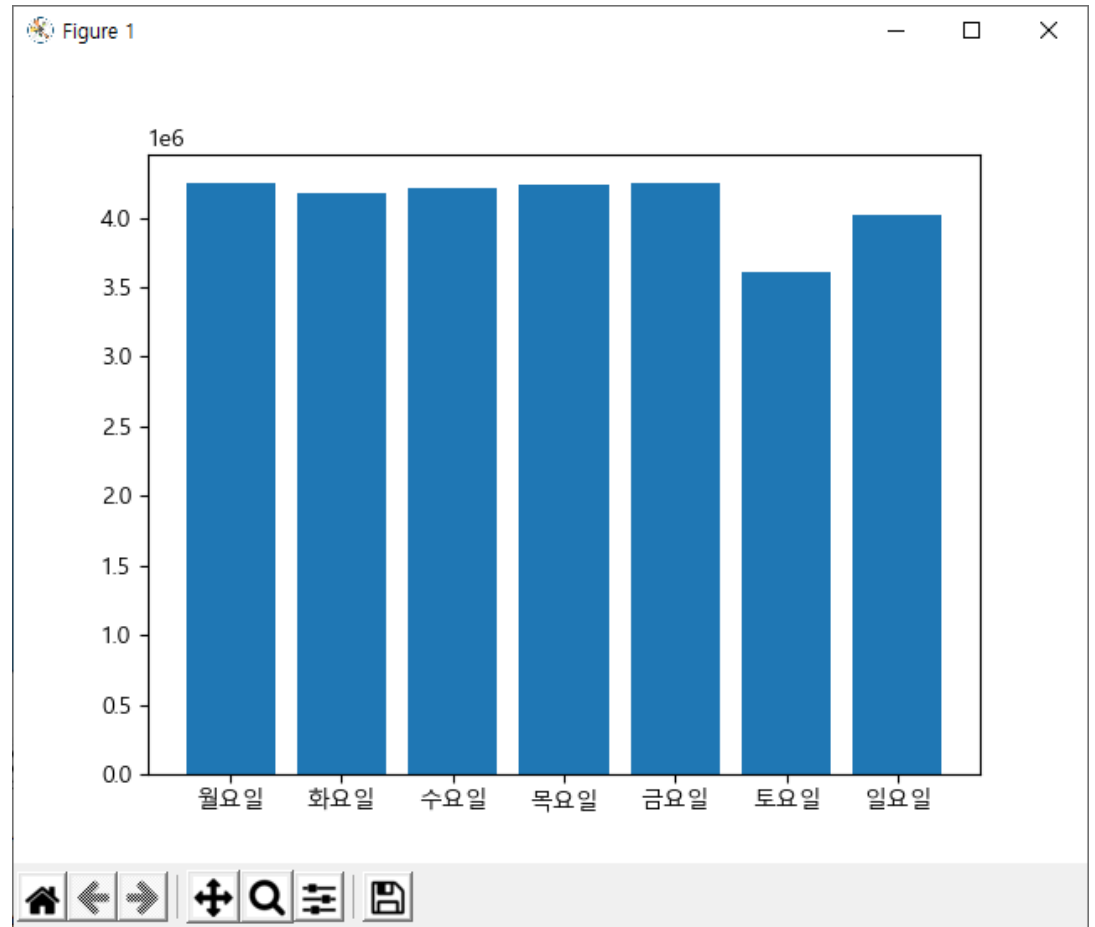
Quiz

▣ 아래의 코드가 1월부터 나올 수 있도록 코딩하시오.

```
2005년 10월      592712
2005년 11월      566138
2005년 12월      572343
2005년 1월       594924
2005년 2월       545332
2005년 3월       617540
2005년 4월       594492
2005년 5월       614802
2005년 6월       609195
2005년 7월       627961
2005년 8월       630904
2005년 9월       574253
2006년 10월      611718
2006년 11월      586197
```

Quiz

월	요	일	4245697
화	요	일	4178222
수	요	일	4214683
목	요	일	4230660
금	요	일	4244446
토	요	일	3609847
일	요	일	4021906



Quiz

▶ 월별 결항 횟수를 구하십시오

2007년 08월	12295
2007년 09월	6507
2007년 10월	7327
2007년 11월	6279
2007년 12월	21493
2008년 01월	17308
2008년 02월	20596
2008년 03월	16183
2008년 04월	10355
2008년 05월	6229
2008년 06월	10931
2008년 07월	10598

