



大數據資料處理 – Hadoop MapReduce Programming

網路資料自動下載

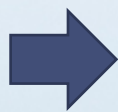
交通部高速公路闡道資料-”TDCS06A” 為例
(<http://tisvcloud.freeway.gov.tw/history/TDCS/M06A/>)

授課教授：王經篤 (Jing-Doo Wang)

Email: jdwang@asia.edu.tw

亞洲大學 資訊工程系 (CSIE, Asia University)

請進入左方網址



交通部高速公路閘道資料-"TDCS06A" 為例
[http://tisvcloud.freeway.gov.tw/history/
TDCS/M06A/](http://tisvcloud.freeway.gov.tw/history/TDCS/M06A/)

六、本資料庫即時資料下載位置：<http://tsvcloud.freeway.gov.tw>

七、本資料庫歷史資料下載位置：<http://tsvcloud.freeway.gov.tw/history>

roadlevel_info.xml.gz	路段靜態資訊
roadlevel_value.xml.gz	路段一分鐘動態資訊
roadlevel_value5.xml.gz	路段五分鐘動態資訊
roadlevel_threshold.xml.gz	服務水準門檻分級
vd_info.xml.gz	VD 靜態資訊
vd_value.xml.gz	VD 一分鐘動態資訊
vd_value5.xml.gz	VD 五分鐘動態資訊
cctv_info.xml.gz	CCTV 靜態資訊
cctv_value.xml.gz	CCTV 動態資訊
cms_info.xml.gz	CMS 靜態資訊
cms_value.xml.gz	CMS 動態資訊
avi_info.xml.gz	AVI 所在位置靜態資訊
avi_pair.xml.gz	AVI 配對靜態資訊
avi_value.xml.gz	AVI 動態資訊
TDCS_M03A	各類車種通行量統計
TDCS_M04A	站間各車種平均旅行時間
TDCS_M05A	站間各車種平均行駛車速
TDCS_M06A	各旅次路徑原始資料
TDCS_M07A	各類車種旅次平均長度
TDCS_M08A	各類車種旅次數
TDCS_PINGLIN	宜蘭地區eTag配對旅行時間



選取TDCS_MO6A

※如果想要選擇別的資料只需要在程式中的 Base Type變更名稱即可

進入後，會再裡面看到許多資料夾，資料夾記載國道闢道每天的資訊，可以依據自己的需求下載

Data Collection System,TDCS)使用手冊」。

(三) 下載資料內容為系統自動產出，系統無法預估系統斷線或設備故障等情況，故資料僅供參考，下載者皆不得以本局資料庫所提供之資訊主張任何權利。

(四) 下載資料內容或加值後對外提供時，應說明原始資料來源為本局。

(五) 使用者重複擷取資料週期不得小於四十秒，以免影響他人下載權力。若發現有違反情事，本局得逕行中斷其連線，使用者不得異議。

(六) 本資料庫資料不得進行任何違反中華民國法律之使用及應用。

三、使用者開始進行資料下載即視同同意本注意事項之相關規定。

四、交通部發布之「路側設施即時交通資訊發布標準格式」。

五、本局「國道高速公路電子收費交通資料蒐集支援系統(Traffic Data Collection System,TDCS)使用手冊」。

六、本資料庫即時資料下載位置：<http://tisvcloud.freeway.gov.tw>

七、本資料庫歷史資料下載位置：
<http://tisvcloud.freeway.gov.tw/history>

Name	Last modified	Size
 Parent Directory		-
 20170707/	08-Jul-2017 06:38	-
 20170706/	07-Jul-2017 06:37	-
 20170705/	06-Jul-2017 06:37	-
 20170704/	05-Jul-2017 06:36	-

滑鼠 - 按到手抽筋！

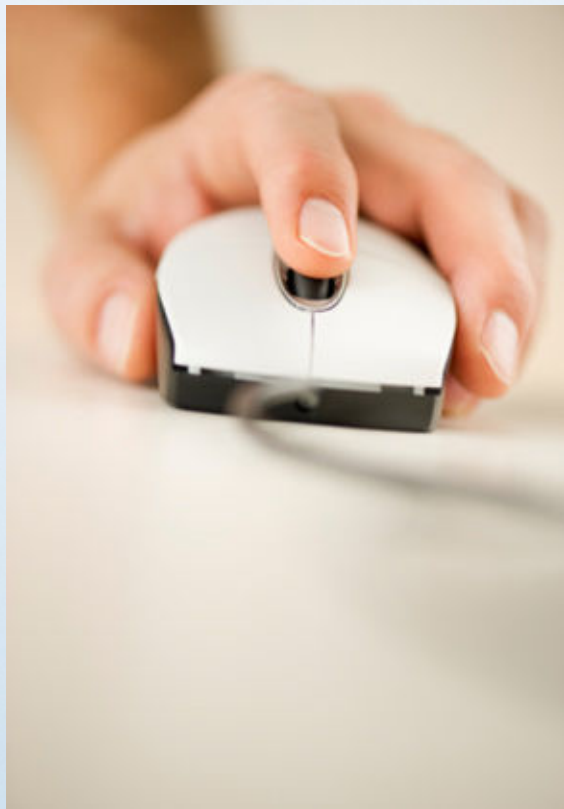


Photo From: <http://healthtalk.pixnet.net/album/photo/63323298>

網路機器人 Web Bot 、 Web Spider

交通部高速公路閘道資料-"TDCS06A" 為例

<http://tisvcloud.freeway.gov.tw/history/TDCS/M06A/>

自動下載範例程式（藉由金門大學陳鍾誠老師之程式改寫）：

dns2.asia.edu.tw/~jdwang/TeachingCourses/Hadoop/TDCS_WebURLDownload_jdwang_2017_10_20.zip



請先下載此連結

新建(N)	Alt+Shift+N
開啟檔案(..)	
關閉(C)	Ctrl+W
全部關閉(L)	Ctrl+Shift+W
儲存(S)	Ctrl+S
另存新檔(A)...	
全部儲存(E)	Ctrl+Shift+S
回復(T)	
移動(V)...	
重新命名(M)...	F2
重新整理(F)	F5
將行定界字元轉換成(V)	
列印(P)...	Ctrl+P
切換工作區(W)	
Restart	
匯入(I)...	
匯出(O)...	
內容(R)	Alt+Enter
1 presidentElection.java [PresidentEl...]	
2 TDCS_GIDSequence_MRE.java [TDCS_GID...]	
3 WebPageDownload_jdwang_1.java [WebP...]	
結束(X)	

```
package org.windoop;

import java.io.BufferedWriter;
import java.io.OutputStreamWriter;
import java.util.Random;
import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.*;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.lib.input.*;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;

public class presidentElection {

    public static void main(String[] args) throws Exception {
        String[] candidate = {"A","B","C"};
        String[] location = {"Taichung", "Taipei", "Kaohsiung"};
        Random r;
        boolean useResourceManager = true;
        //boolean useJobTracker = true; <-- for hadoop 1.0

        //Step 0. Set configuration
```

快速存取

Java Map/Reduce

概要

- org.windoop
 - presidentElection
 - main(String[]) : void

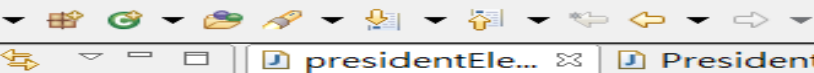
下載完之後，到
eclipse檔案點選匯
入

Problems Tasks @ Javadoc Map/Reduce Locations 主控台

<已終止> presidentElection [Java 應用程式] C:\Program Files\Java\jre1.8.0_121\bin\javaw.exe (2017年7月5日 上午9:40:17 直接存取)

```
Total committed heap usage (bytes)=735576064
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=9000000
File Output Format Counters
  Bytes Written=28

Job Done!
```



presidentEle... PresidentEle... PresidentEle... hdfs://local... TDCS_GIDSequ... WebPageDownl... »₂

```
package org.windowop;
```

```
import java.io.*;
import java.util.*;
import org.apache.hadoop.conf.*;
import org.apache.hadoop.fs.*;
import org.apache.hadoop.mapreduce.*;
import org.apache.hadoop.mapreduce.Mapper;
import org.apache.hadoop.mapreduce.Reducer;
```

```
public class PresidentElection {
```

```
    public static void main(String[] args) throws Exception {
        String input = "input.txt";
        String output = "output.txt";
        Random r = new Random();
        boolean success = true;
        // bootstrap
        // Step 1: Create input file
```



匯入

選取
自保存檔或目錄來建立新專案。

選取匯入來源(S) :

輸入過濾器文字

- 一般
 - 保存檔
 - 喜好設定
 - 匯入現有的專案至工作區**
 - 檔案系統
- 安裝
- 執行/除錯
- 團隊



點選匯入現有專案
至工作區

Problems Tasks



< 上一步(B)

下一步(N) >

完成(F)

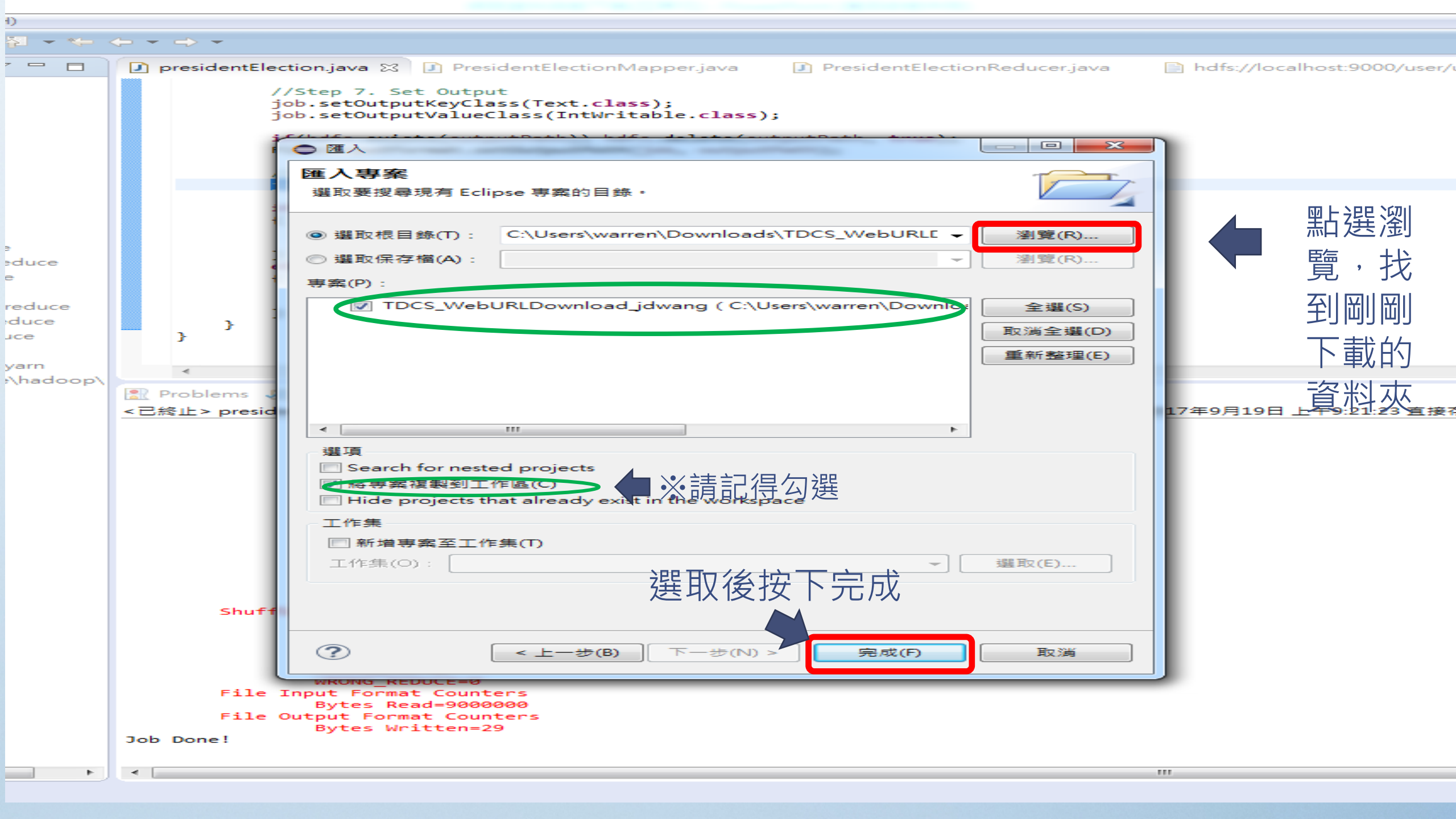
取消

<已終止> presidentElec

上午9:40:17 直接存取)

```
Total committed heap usage (bytes)=735576064
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=9000000
File Output Format Counters
Bytes Written=28
```

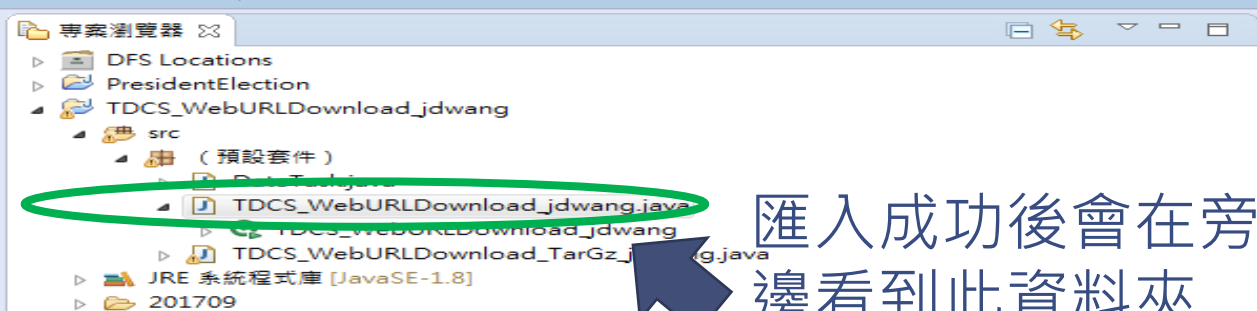
Job Done!



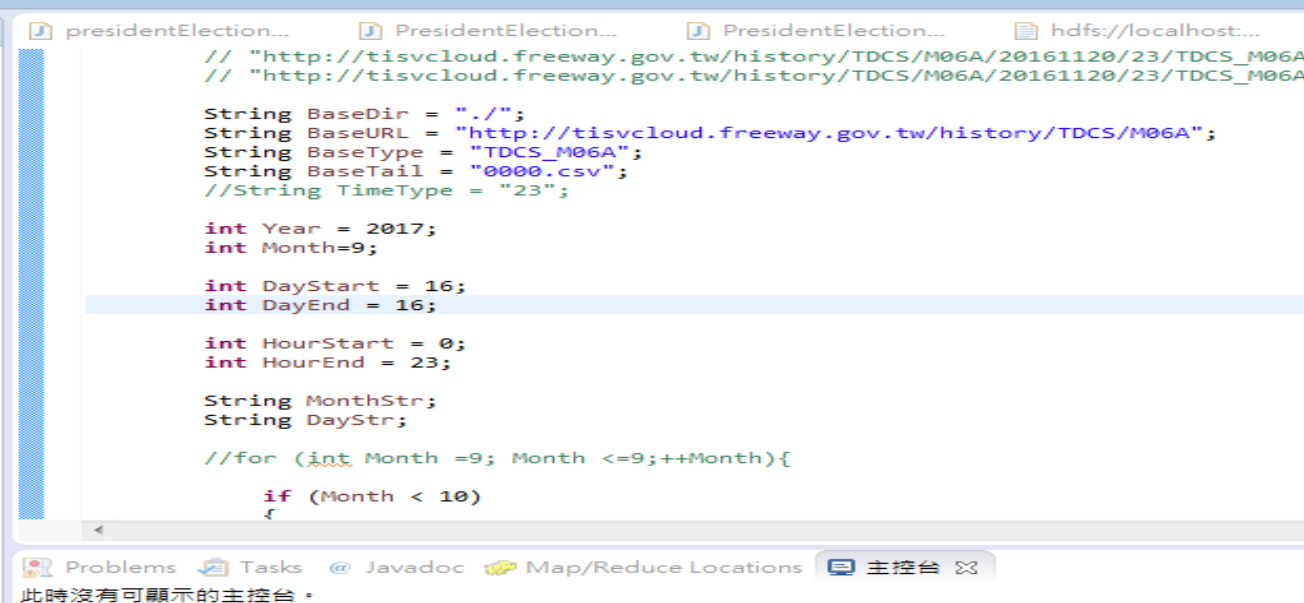
點選瀏覽，找到剛剛下載的資料夾

※請記得勾選

選取後按下完成



匯入成功後會在旁邊看到此資料夾



(五) 使用者重複擷取資料週期**不得小於四十秒**，以免影響他人下載權力。

① tsvcloud.freeway.gov.tw/history/TDCS/M06A/

交通資料庫

交通部臺灣區國道高速公路局「交通資料庫」下載注意事項

104年6月發布

一、訂定目的

交通部臺灣區國道高速公路局(以下簡稱本局)「交通資料庫」(以下簡稱本資料庫)為彙整國道即時及歷史交通資訊，提供用路人更多資訊取得管道，交通資訊內容免費對外開放使用。為規範下載及使用之方式，特訂定本注意事項。

二、下載資訊使用範圍

(一) 本資料庫資料來源為「高快速公路整體路網交通管理系統」與「高速公路電子收費系統」，資料提供免費下載使用。

(二) 本系統資料格式係依據交通部發布之「路側設施即時交通資訊發布標準格式」及本局「國道高速公路電子收費交通資料蒐集支援系統(Traffic Data Collection System,TDCS)使用手冊」。

(三) 下載資料內容為系統自動產出，系統無法預估系統斷線或設備故障等情況，故資料僅供參考，下載者皆不得以本局資料庫所提供之資訊主張任何權利。

(四) 下載資料內容或加值後對外提供時，應說明原始資料來源為本局。

(五) 使用者重複擷取資料週期不得小於四十秒，以免影響他人下載權力。若發現有違反情事，本局得逕行中斷其連線，使用者不得異議。

(六) 本資料庫資料不得進行任何違反中華民國法律之使用及應用。

三、使用者開始進行資料下載即視同同意本注意事項之相關規定。

四、交通部發布之「路側設施即時交通資訊發布標準格式」。

五、本局「國道高速公路電子收費交通資料蒐集支援系統(Traffic Data

開啟此程式碼

```
// "http://tisvcloud.freeway.gov.tw/history/TDCS/M06A/20161120/23/TDCS_M06A_20161120_230000.csv"  
// "http://tisvcloud.freeway.gov.tw/history/TDCS/M06A/20161120/23/TDCS_M06A_20161120_000000.csv"
```

```
String BaseDir = "./";  
String BaseURL = "http://tisvcloud.freeway.gov.tw/history/TDCS/M06A";  
String BaseType = "TDCS_M06A";  
String BaseTail = "0000.csv";  
//String TimeType = "23";
```

```
int Year = 2017;  
int Month=9;
```

```
int DayStart = 16;  
int DayEnd = 16;
```

```
int HourStart = 0;  
int HourEnd = 23;
```

```
String MonthStr;  
String DayStr;
```

```
//for (int Month =9; Month <=9;++Month){
```

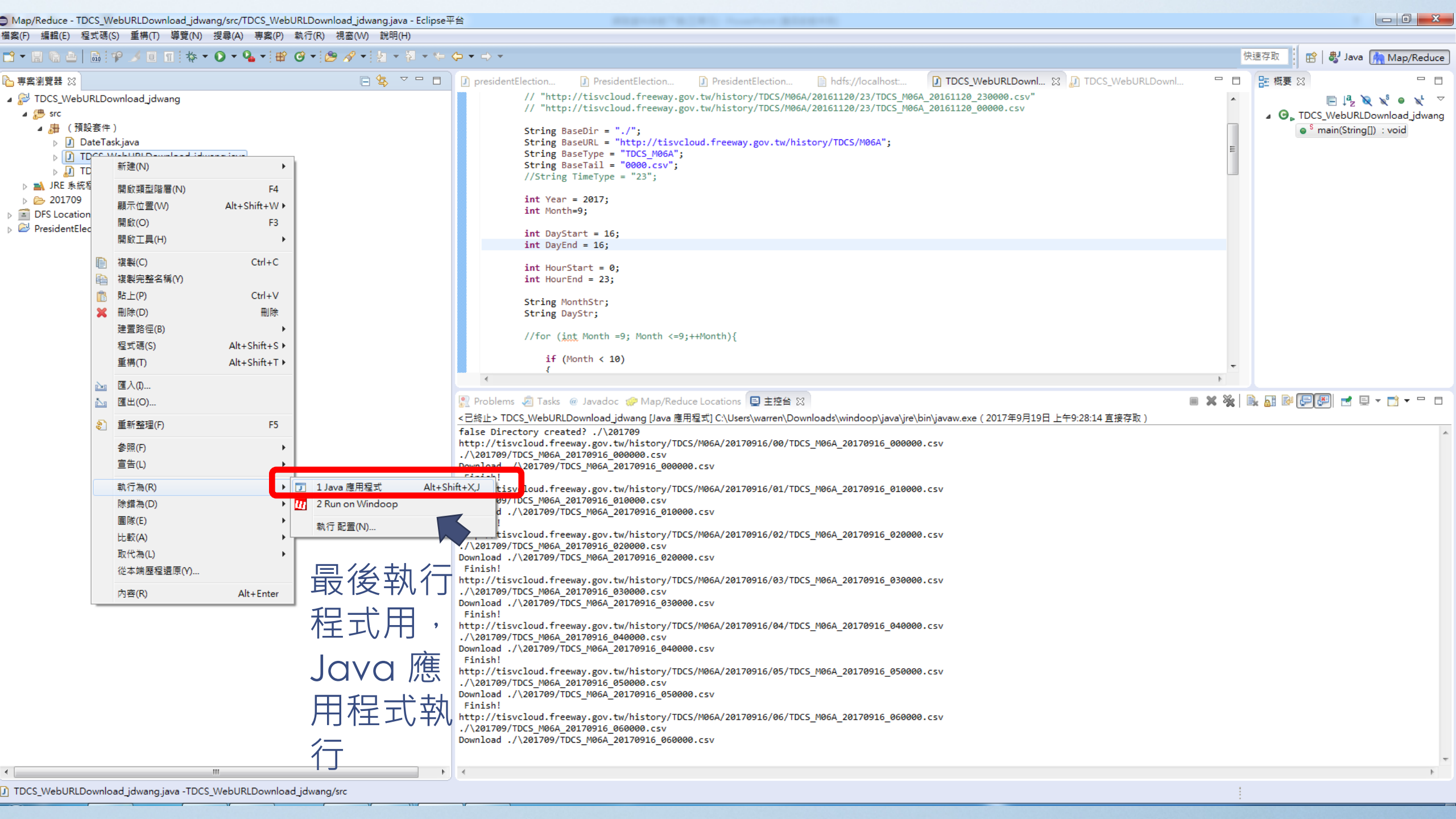
```
if (Month < 10)
```

```
{
```

依照資料夾年月日，
改變程式年月日的日期，
就可以選取高公
局資料並下載

<已終止> TDCS_WebURLDownload_jdwang [Java 應用程式] C:\Users\warren\Downloads\windoop\java\jre\bin\javaw.exe (2017年9月19日 上午9:28:14 直接存取)

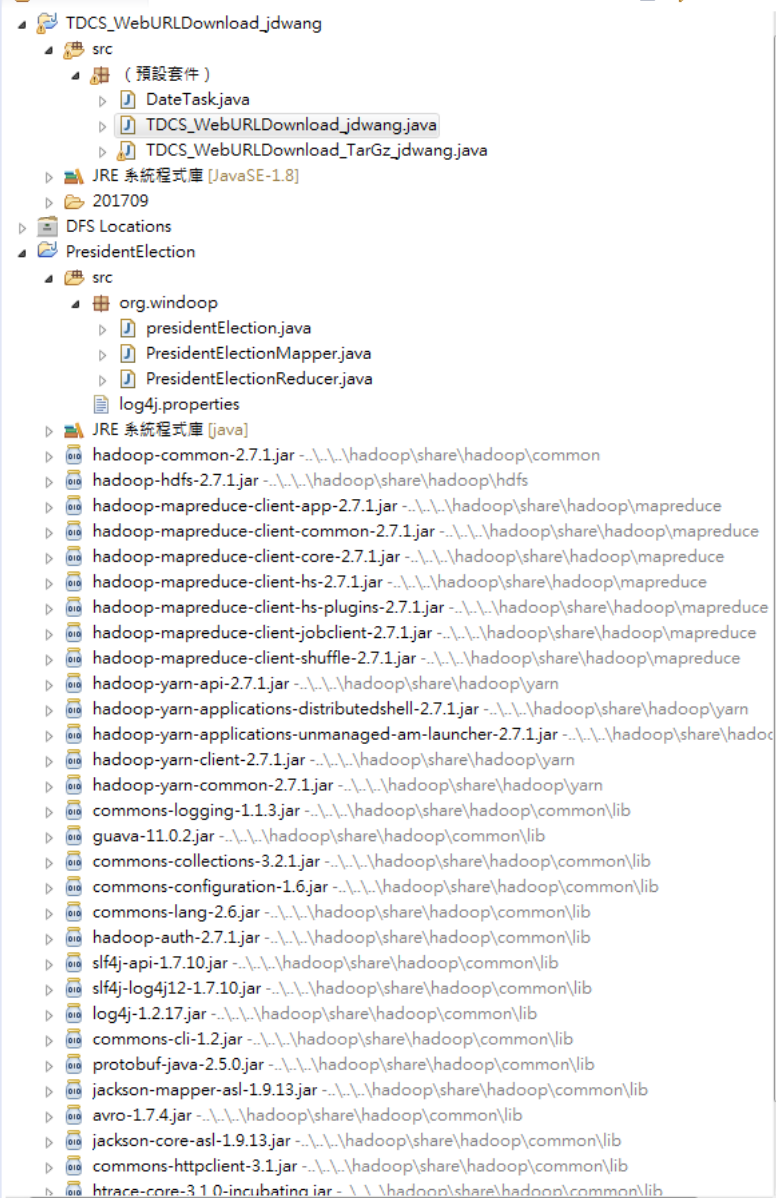
```
false Directory created? ./\201709  
http://tisvcloud.freeway.gov.tw/history/TDCS/M06A/20170916/00/TDCS_M06A_20170916_000000.csv  
./\201709/TDCS_M06A_20170916_000000.csv  
Download ./\201709/TDCS_M06A_20170916_000000.csv  
Finish!  
http://tisvcloud.freeway.gov.tw/history/TDCS/M06A/20170916/01/TDCS_M06A_20170916_010000.csv  
./\201709/TDCS_M06A_20170916_010000.csv  
Download ./\201709/TDCS_M06A_20170916_010000.csv  
Finish!  
http://tisvcloud.freeway.gov.tw/history/TDCS/M06A/20170916/02/TDCS_M06A_20170916_020000.csv  
./\201709/TDCS_M06A_20170916_020000.csv  
Download ./\201709/TDCS_M06A_20170916_020000.csv  
Finish!  
http://tisvcloud.freeway.gov.tw/history/TDCS/M06A/20170916/03/TDCS_M06A_20170916_030000.csv  
./\201709/TDCS_M06A_20170916_030000.csv  
Download ./\201709/TDCS_M06A_20170916_030000.csv  
Finish!  
http://tisvcloud.freeway.gov.tw/history/TDCS/M06A/20170916/04/TDCS_M06A_20170916_040000.csv  
./\201709/TDCS_M06A_20170916_040000.csv  
Download ./\201709/TDCS_M06A_20170916_040000.csv  
Finish!  
http://tisvcloud.freeway.gov.tw/history/TDCS/M06A/20170916/05/TDCS_M06A_20170916_050000.csv  
./\201709/TDCS_M06A_20170916_050000.csv  
Download ./\201709/TDCS_M06A_20170916_050000.csv  
Finish!  
http://tisvcloud.freeway.gov.tw/history/TDCS/M06A/20170916/06/TDCS_M06A_20170916_060000.csv  
./\201709/TDCS_M06A_20170916_060000.csv  
Download ./\201709/TDCS_M06A_20170916_060000.csv
```



最後執行
程式用，
Java 應
用程式執
行



專案瀏覽器



```
public class TDCS_WebURLDownload_jdwang
{
    public static void main(String[] args) throws Exception
    {
        // "http://tisvcloud.freeway.gov.tw/history/TDCS/M06A/20161120/23/TDCS_M06A_20161120_230000.csv"
        // "http://tisvcloud.freeway.gov.tw/history/TDCS/M06A/20161120/23/TDCS_M06A_20161120_00000.csv"

        String BaseDir = ".";
        String BaseURL = "http://tisvcloud.freeway.gov.tw/history/TDCS/M06A";
        String BaseType = "TDCS_M06A";
        String BaseTail = "0000.csv";
        //String TimeType = "23";

        int Year = 2017;
        int Month=9;

        int DayStart = 16;
        int DayEnd = 16;

        int HourStart = 0;
        int HourEnd = 23;

        String MonthStr;
        String DayStr;
```

快速存取

Java Map/Reduce

概要

TDCS_WebURLDownload_jdwang

main(String[]) : void

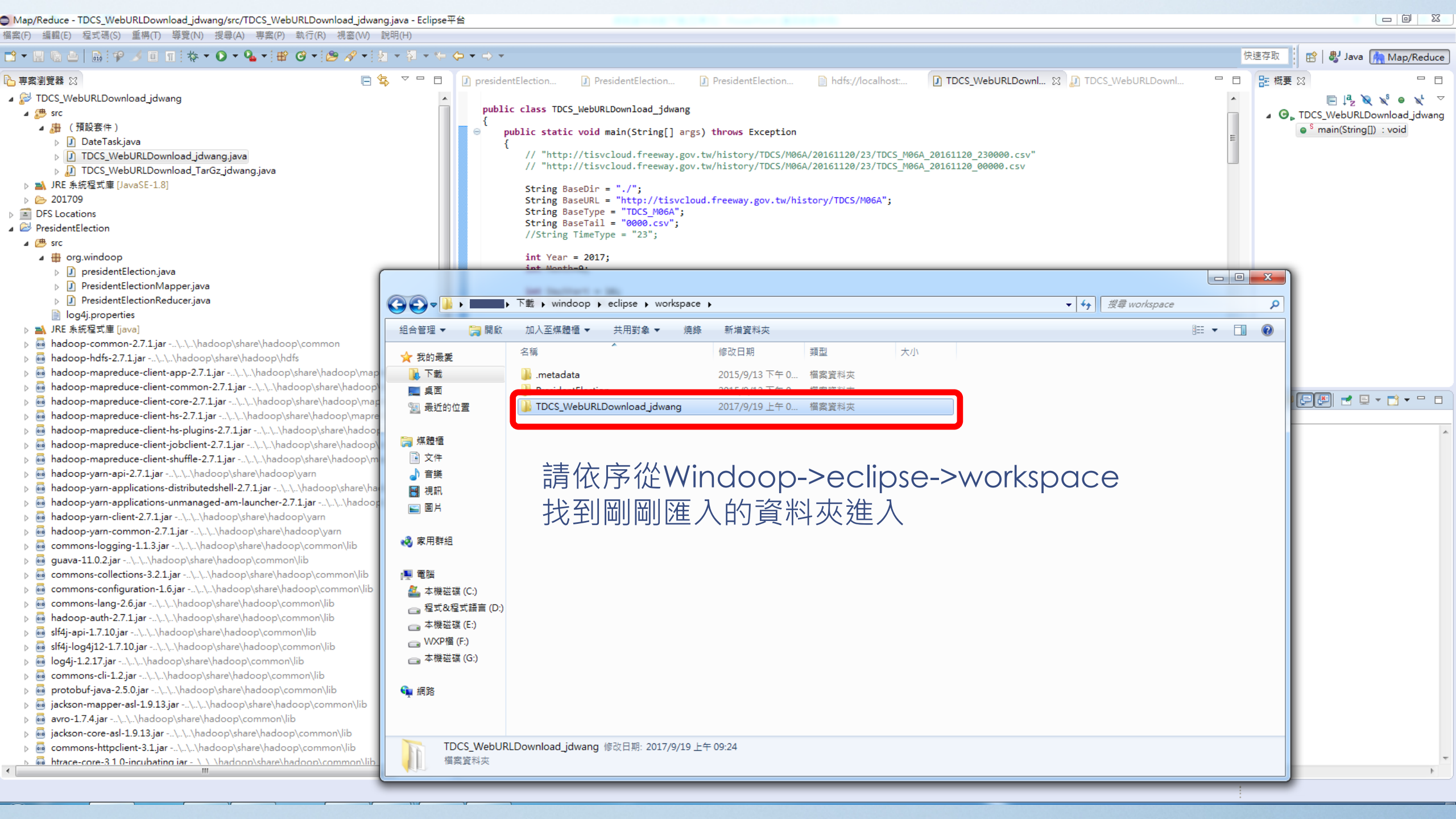
Problems Tasks Javadoc Map/Reduce Locations 主控台

TDCS_WebURLDownload_jdwang [Java 應用程式 C:\Users\warren\Downloads\windoop\java\ire\bin\javaw.exe (2017年9月19日 上午9:28:14 直接存取)]

```
false Directory created? ./\201709
http://tisvcloud.freeway.gov.tw/history/TDCS/M06A/20170916/00/TDCS_M06A_20170916_000000.csv
./\201709/TDCS_M06A_20170916_000000.csv
Download ./\201709/TDCS_M06A_20170916_000000.csv
Finish!
http://tisvcloud.freeway.gov.tw/history/TDCS/M06A/20170916/01/TDCS_M06A_20170916_010000.csv
./\201709/TDCS_M06A_20170916_010000.csv
Download ./\201709/TDCS_M06A_20170916_010000.csv
Finish!
http://tisvcloud.freeway.gov.tw/history/TDCS/M06A/20170916/02/TDCS_M06A_20170916_020000.csv
./\201709/TDCS_M06A_20170916_020000.csv
Download ./\201709/TDCS_M06A_20170916_020000.csv
Finish!
http://tisvcloud.freeway.gov.tw/history/TDCS/M06A/20170916/03/TDCS_M06A_20170916_030000.csv
./\201709/TDCS_M06A_20170916_030000.csv
Download ./\201709/TDCS_M06A_20170916_030000.csv
Finish!
http://tisvcloud.freeway.gov.tw/history/TDCS/M06A/20170916/04/TDCS_M06A_20170916_040000.csv
./\201709/TDCS_M06A_20170916_040000.csv
Download ./\201709/TDCS_M06A_20170916_040000.csv
```

表示執行成功請
移至資料夾檢查
是否匯入資料







專案瀏覽器

- TDCS_WebURLDownload_jdwang
 - src
 - (預設套件)
 - DateTask.java
 - TDCS_WebURLDownload_jdwang.java
 - TDCS_WebURLDownload_TarGz_jdwang.java

JRE 系統程式庫 [JavaSE-1.8]

201709

DFS Locations

PresidentElection

src

org.windoop

presidentElection.java

PresidentElectionMapper.java

PresidentElectionReducer.java

log4j.properties

JRE 系統程式庫 [java]

hadoop-common-2.7.1.jar ...\..\..\hadoop\share\hadoop\common

hadoop-hdfs-2.7.1.jar ...\..\..\hadoop\share\hadoop\hdfs

hadoop-mapreduce-client-app-2.7.1.jar ...\..\..\hadoop\share\hadoop\map

hadoop-mapreduce-client-common-2.7.1.jar ...\..\..\hadoop\share\hadoop

hadoop-mapreduce-client-core-2.7.1.jar ...\..\..\hadoop\share\hadoop\map

hadoop-mapreduce-client-hs-2.7.1.jar ...\..\..\hadoop\share\hadoop\map

hadoop-mapreduce-client-hs-plugins-2.7.1.jar ...\..\..\hadoop\share\hadoop

hadoop-mapreduce-client-jobclient-2.7.1.jar ...\..\..\hadoop\share\hadoop

hadoop-mapreduce-client-shuffle-2.7.1.jar ...\..\..\hadoop\share\hadoop\m

hadoop-yarn-api-2.7.1.jar ...\..\..\hadoop\share\hadoop\yarn

hadoop-yarn-applications-distributedshell-2.7.1.jar ...\..\..\hadoop\share\ha

hadoop-yarn-applications-unmanaged-am-launcher-2.7.1.jar ...\..\..\hadoop

hadoop-yarn-client-2.7.1.jar ...\..\..\hadoop\share\hadoop\yarn

hadoop-yarn-common-2.7.1.jar ...\..\..\hadoop\share\hadoop\yarn

commons-logging-1.1.3.jar ...\..\..\hadoop\share\hadoop\common\lib

guava-11.0.2.jar ...\..\..\hadoop\share\hadoop\common\lib

commons-collections-3.2.1.jar ...\..\..\hadoop\share\hadoop\common\lib

commons-configuration-1.6.jar ...\..\..\hadoop\share\hadoop\common\lib

commons-lang-2.6.jar ...\..\..\hadoop\share\hadoop\common\lib

hadoop-auth-2.7.1.jar ...\..\..\hadoop\share\hadoop\common\lib

slf4j-api-1.7.10.jar ...\..\..\hadoop\share\hadoop\common\lib

slf4j-log4j12-1.7.10.jar ...\..\..\hadoop\share\hadoop\common\lib

log4j-1.2.17.jar ...\..\..\hadoop\share\hadoop\common\lib

commons-cli-1.2.jar ...\..\..\hadoop\share\hadoop\common\lib

protobuf-java-2.5.0.jar ...\..\..\hadoop\share\hadoop\common\lib

jackson-mapper-asl-1.9.13.jar ...\..\..\hadoop\share\hadoop\common\lib

avro-1.7.4.jar ...\..\..\hadoop\share\hadoop\common\lib

jackson-core-asl-1.9.13.jar ...\..\..\hadoop\share\hadoop\common\lib

commons-httpclient-3.1.jar ...\..\..\hadoop\share\hadoop\common\lib

htrace-core-3.1.0-incubating.jar ...\..\..\hadoop\share\hadoop\common\lib

public class TDCS_WebURLDownload_jdwang

{

public static void main(String[] args) throws Exception

{

// "http://tisvcloud.freeway.gov.tw/history/TDCS/M06A/20161120/23/TDCS_M06A_20161120_230000.csv"

// "http://tisvcloud.freeway.gov.tw/history/TDCS/M06A/20161120/23/TDCS_M06A_20161120_000000.csv"

String BaseDir = ".";

String BaseURL = "http://tisvcloud.freeway.gov.tw/history/TDCS/M06A";

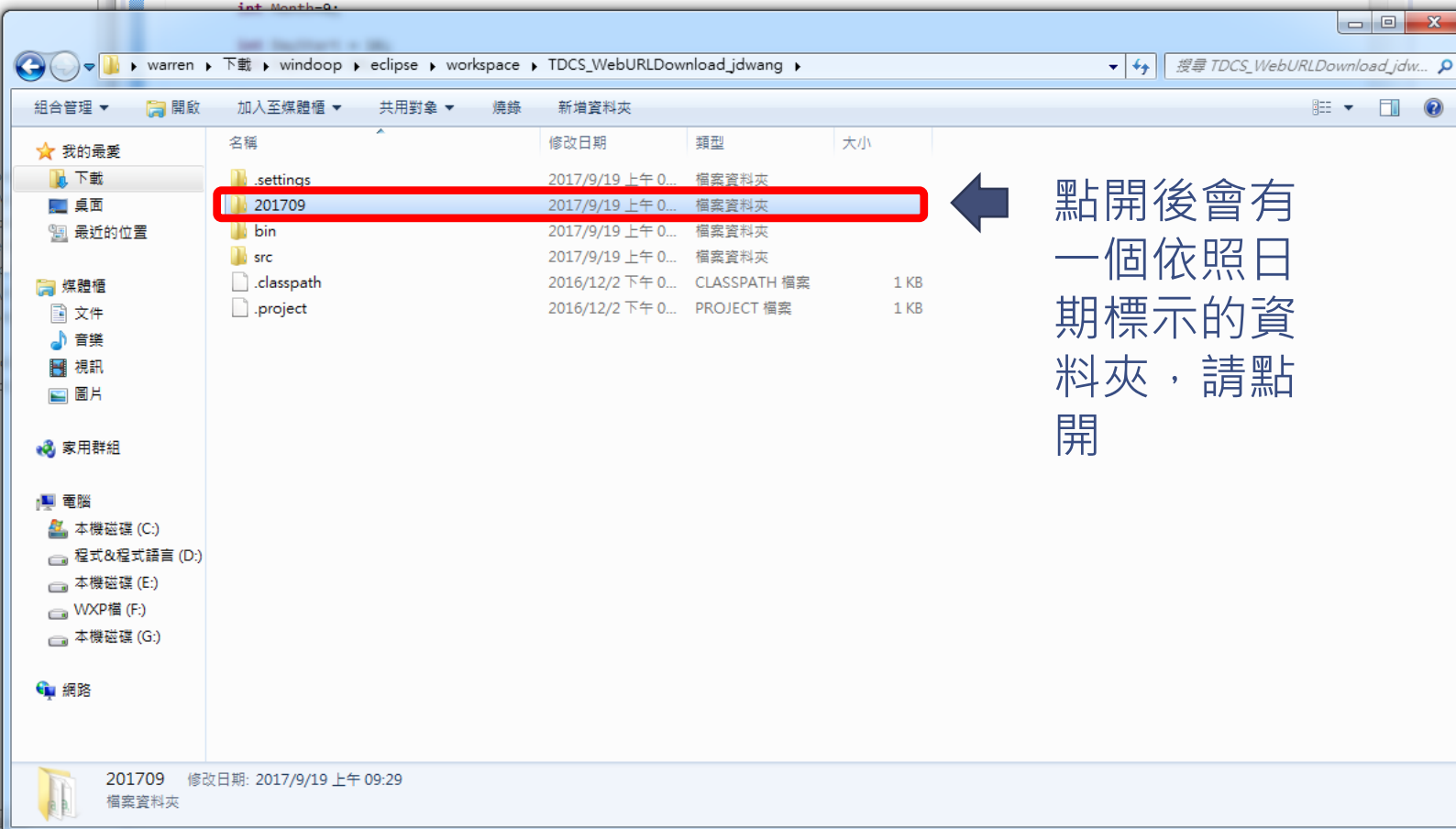
String BaseType = "TDCS_M06A";

String BaseTail = "0000.csv";

//String TimeType = "23";

int Year = 2017;

int Month = 9;



點開後會有一個依照日期標示的資料夾，請點開

TDCS_WebURLDownload_jdwang

- src
 - (預設套件)
 - DateTask.java
 - TDCS_WebURLDownload_jdwang.java
 - TDCS_WebURLDownload_TarGz_jdwang.java

JRE 系統程式庫 [JavaSE-1.8]

201709

DFS Locations

PresidentElection

- src
 - org.windoop
 - presidentElection.java
 - PresidentElectionMapper.java
 - PresidentElectionReducer.java
 - log4j.properties

JRE 系統程式庫 [java]

- hadoop-common-2.7.1.jar
- hadoop-hdfs-2.7.1.jar
- hadoop-mapreduce-client-app-2.7.1.jar
- hadoop-mapreduce-client-common-2.7.1.jar
- hadoop-mapreduce-client-core-2.7.1.jar
- hadoop-mapreduce-client-hs-2.7.1.jar
- hadoop-mapreduce-client-hs-plugins-2.7.1.jar
- hadoop-mapreduce-client-jobclient-2.7.1.jar
- hadoop-mapreduce-client-shuffle-2.7.1.jar
- hadoop-yarn-api-2.7.1.jar
- hadoop-yarn-applications-distributedshell-2.7.1.jar
- hadoop-yarn-applications-unmanaged-am-launcher-2.7.1.jar
- hadoop-yarn-client-2.7.1.jar
- hadoop-yarn-common-2.7.1.jar
- commons-logging-1.1.3.jar
- guava-11.0.2.jar
- commons-collections-3.2.1.jar
- commons-configuration-1.6.jar
- commons-lang-2.6.jar
- hadoop-auth-2.7.1.jar
- slf4j-api-1.7.10.jar
- slf4j-log4j12-1.7.10.jar
- log4j-1.2.17.jar
- commons-cli-1.2.jar
- protobuf-java-2.5.0.jar
- jackson-mapper-asl-1.9.13.jar
- avro-1.7.4.jar
- jackson-core-asl-1.9.13.jar
- commons-httpclient-3.1.jar
- htrace-core-3.1.0-incubating.jar

```
public class TDCS_WebURLDownload_jdwang
{
    public static void main(String[] args) throws Exception
    {
        // "http://tisvcloud.freeway.gov.tw/history/TDCS/M06A/20161120/23/TDCS_M06A_20161120_230000.csv"
        // "http://tisvcloud.freeway.gov.tw/history/TDCS/M06A/20161120/23/TDCS_M06A_20161120_000000.csv"

        String BaseDir = ".";
        String BaseURL = "http://tisvcloud.freeway.gov.tw/history/TDCS/M06A";
        String BaseType = "TDCS_M06A";
        String BaseTail = "0000.csv";
        //String TimeType = "23";

        int Year = 2017;
        int Month=0;
```

下載完成的資料會出現在剛剛匯入的資料夾中

