Map Reduce Programming

王耀聰 陳威宇 楊順發

jazz@nchc.org.tw

waue@nchc.org.tw

shunfa@nchc.org.tw

國家高速網路與計算中心(NCHC)

Outline

- 概念
- 程式基本框架及執行步驟方法
- 範例一:
  - Hadoop 的 Hello World => Word Count
  - 說明
  - 動手做
- 範例二:
  - 進階版=> Word Count 2
  - 說明
  - 動手做

程式基本框架

Program Prototype

Class MR{

Map 區

```
Class Mapper …{
}
```

Map 程式碼

Reduce 區

```
Class Reducer …{
}
```

Reduce 程式碼

設定區

```
main(){
Configuration conf = new Configuration();
Job job = new Job(conf, "the name");
job.setJarByClass(MR.class);
FileInputFormat.setInputPaths(job, new Path( " /user/hadoop/input " ));
FileOutputFormat.setOutputPath(job, new Path(" /user/hadoop/output " ));
job.setMapperClass(Mapper.class);
job.setReducerClass(Reducer.class);
job.waitForCompletion(true);
}}
```

其他的設定參數程式碼

Class Mapper

```
class MyMap extends MapReduceBase
implements Mapper < INPUT KEY , INPUT VALUE , OUTPUT KEY , OUTPUT VALUE >
{
// 全域變數區
public void map ( INPUT KEY key, INPUT VALUE value, Context context)
throws IOException, InterruptedException
          {
          // 區域變數與程式邏輯區
          context.write( NewKey, NewValue);
          }
}
```

程式基本框架

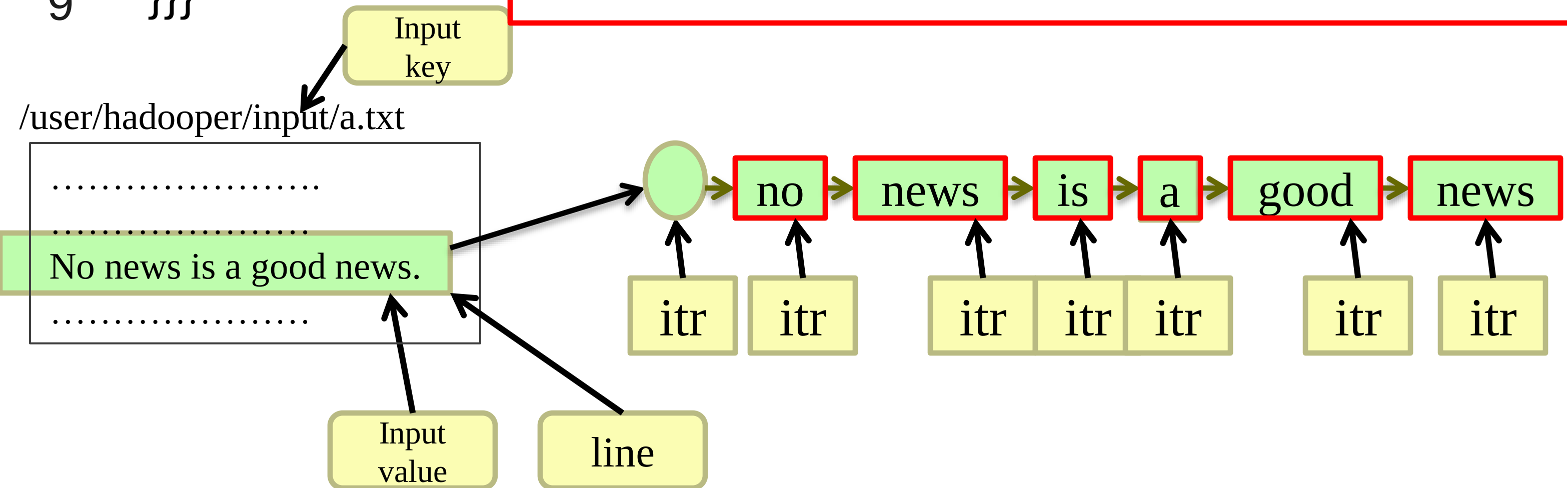
Class Reducer

```
class MyRed extends MapReduceBase
implements Reducer < INPUT KEY , INPUT VALUE , OUTPUT KEY , OUTPUT VALUE >
{
// 全域變數區
public void reduce ( INPUT KEY key, Iterator< INPUT VALUE > values,
          Context context) throws IOException, InterruptedException
          {
          // 區域變數與程式邏輯區
          output.collect( NewKey, NewValue);
          }
}
```

範例
程式一

Word Count Sample (1)

```
class MapClass extends MapReduceBase implements
Mapper<LongWritable, Text, Text, IntWritable> {
        private final static IntWritable one = new IntWritable(1);
        private Text word = new Text();
        public void map( LongWritable key, Text value,
          Context context) throws IOException {
                String line = ((Text) value).toString();
                StringTokenizer itr = new StringTokenizer(line);
                while (itr.hasMoreTokens()) {
                        word.set(itr.nextToken());
                        context.write(word, one);
}}}
```

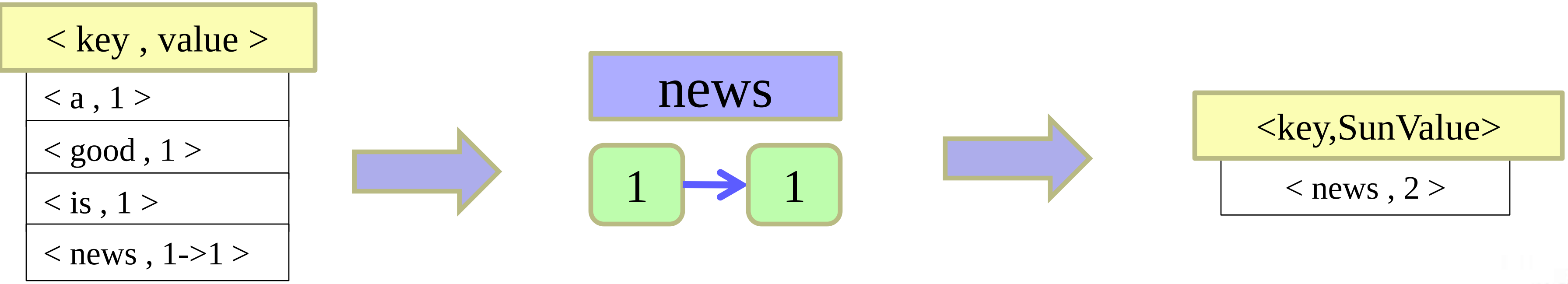


<word,one>
< no , 1 >
< news , 1 >
< is , 1 >
< a, 1 >
< good , 1 >
< news, 1 >

範例
程式一

Word Count Sample (2)

```
class ReduceClass extends MapReduceBase implements Reducer< Text,
IntWritable, Text, IntWritable> {
        IntWritable SumValue = new IntWritable();
        public void reduce( Text key, Iterator<IntWritable> values,
        Context context)
        throws IOException {
                int sum = 0;
                while (values.hasNext())
                        sum += values.next().get();
                SumValue.set(sum);
                context.write(key, SumValue);
}}
```



< key , value >

< a , 1 >

< good , 1 >

< is , 1 >

< news , 1->1 >

news

1 → 1

<key,SunValue>

< news , 2 >

範例
程式一

Word Count Sample (3)

```
Class WordCount{
.. main(){
        Configuration conf = new Configuration();
        Job job = new Job(conf, "Word Count");
        job.setJarByClass(WordCount.class);
        job.setMapperClass(Mapper.class);
        job.setReducerClass(Reducer.class);
        job.setOutputKeyClass(Text.class);
        job.setOutputValueClass(IntWritable.class);
        FileInputFormat.addInputPath(job, new Path(args[0]));
        FileOutputFormat.setOutputPath(job, new Path(args[1]));
        job.waitForCompletion(true);
}}
```

編譯與執行

1. 編譯
   - javac △ -classpath △ hadoop-*-core.jar △ -d △ MyJava △ MyCode.java
2. 封裝
   - jar △ -cvf △ MyJar.jar △ -C △ MyJava △ .
3. 執行
   - bin/hadoop △ jar △ MyJar.jar △ MyCode △ HDFS_Input/ △ HDFS_Output/

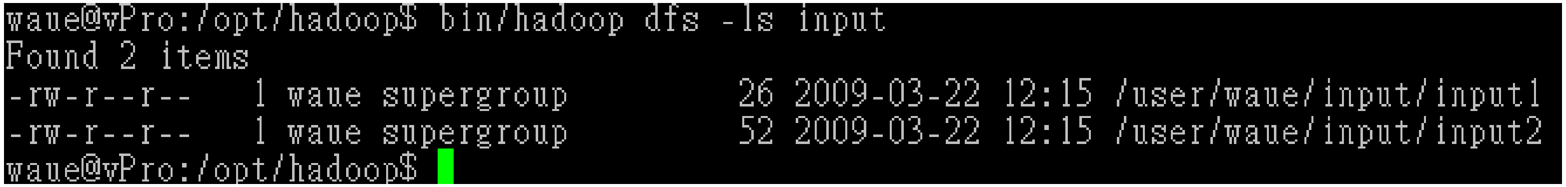
- 所在的執行目錄為Hadoop_Home
- ./MyJava = 編譯後程式碼目錄
- Myjar.jar = 封裝後的編譯檔

- 先放些文件檔到HDFS上的input目錄
- ./input; ./ouput = hdfs的輸入、輸出目錄

範例一
動手做

WordCount1 練習 (I)

1. cd $HADOOP_HOME
2. bin/hadoop dfs -mkdir input
3. echo "I like NCHC Cloud Course." > inputwc/input1
4. echo "I like nchc Cloud Course, and we enjoy this crouse." > inputwc/input2
5. bin/hadoop dfs -put inputwc inputwc
6. bin/hadoop dfs -ls input



```
waue@vPro:/opt/hadoop$ bin/hadoop dfs -ls input
Found 2 items
-rw-r--r--   1 waue supergroup         26 2009-03-22 12:15 /user/waue/input/input1
-rw-r--r--   1 waue supergroup         52 2009-03-22 12:15 /user/waue/input/input2
waue@vPro:/opt/hadoop$
```

範例一
動手做

WordCount1 練習 (II)

1. 編輯WordCount.java
   http://trac.nchc.org.tw/cloud/attachment/wiki/jazz/Hadoop_Lab6/WordCount.java?format=raw
2. mkdir MyJava
3. javac -classpath hadoop-*-core.jar -d MyJava WordCount.java
4. jar -cvf wordcount.jar -C MyJava .
5. bin/hadoop jar wordcount.jar WordCount input/ output/

・所在的執行目錄為Hadoop_Home（因為hadoop-*-core.jar ）

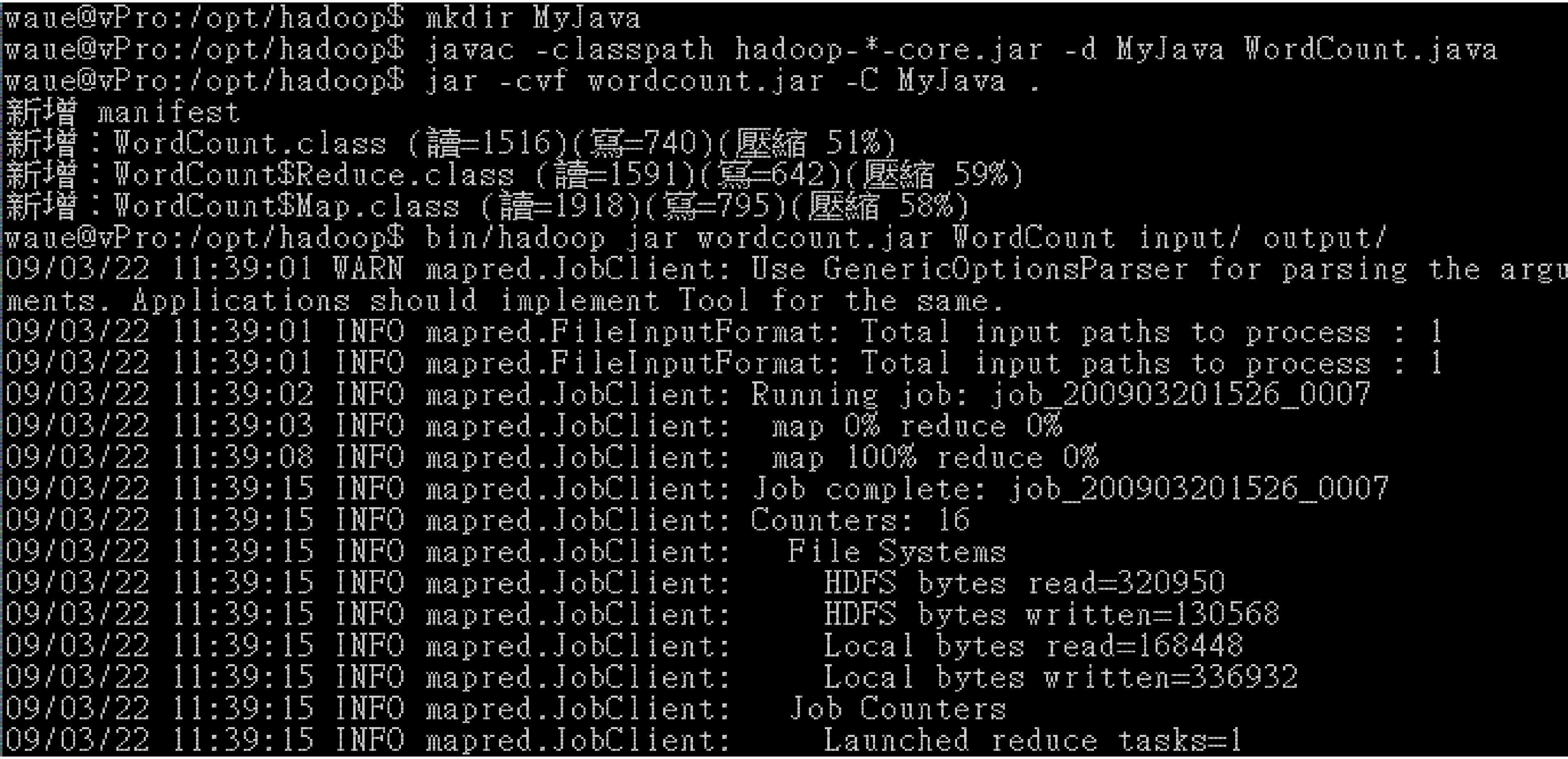
・javac編譯時需要classpath, 但hadoop jar時不用

・wordcount.jar = 封裝後的編譯檔，但執行時需告知class name

・Hadoop進行運算時，只有 input 檔要放到hdfs上，以便hadoop分析運算；執行檔（wordcount.jar）不需上傳，也不需每個node都放，程式的載入交由java處理

範例一
動手做

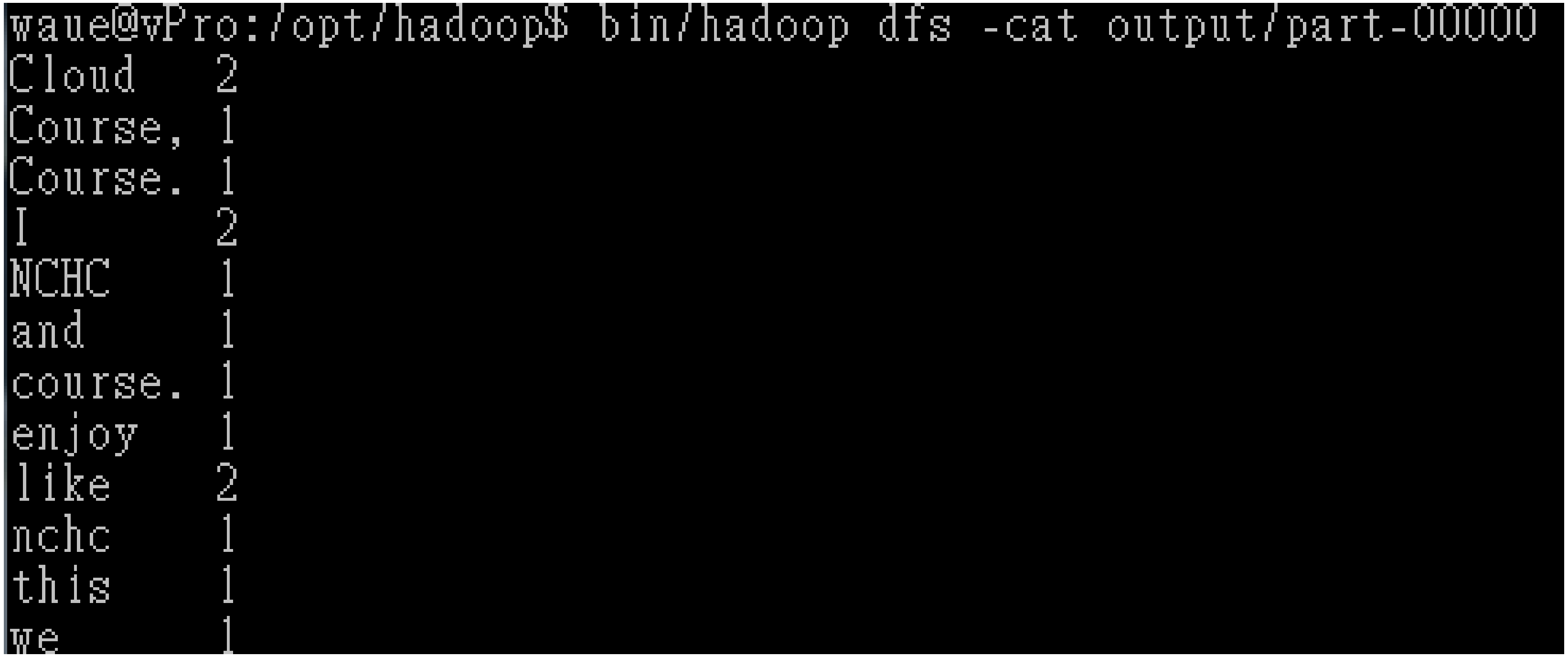
WordCount1 練習(III)



```
waue@vPro:/opt/hadoop$ mkdir MyJava
waue@vPro:/opt/hadoop$ javac -classpath hadoop-*-core.jar -d MyJava WordCount.java
waue@vPro:/opt/hadoop$ jar -cvf wordcount.jar -C MyJava .
新增 manifest
新增：WordCount.class (讀=1516)(寫=740)(壓縮 51%)
新增：WordCount$Reduce.class (讀=1591)(寫=642)(壓縮 59%)
新增：WordCount$Map.class (讀=1918)(寫=795)(壓縮 58%)
waue@vPro:/opt/hadoop$ bin/hadoop jar wordcount.jar WordCount input/ output/
09/03/22 11:39:01 WARN mapred.JobClient: Use GenericOptionsParser for parsing the argu
ments. Applications should implement Tool for the same.
09/03/22 11:39:01 INFO mapred.FileInputFormat: Total input paths to process : 1
09/03/22 11:39:01 INFO mapred.FileInputFormat: Total input paths to process : 1
09/03/22 11:39:02 INFO mapred.JobClient: Running job: job_200903201526_0007
09/03/22 11:39:03 INFO mapred.JobClient:  map 0% reduce 0%
09/03/22 11:39:08 INFO mapred.JobClient:  map 100% reduce 0%
09/03/22 11:39:15 INFO mapred.JobClient: Job complete: job_200903201526_0007
09/03/22 11:39:15 INFO mapred.JobClient: Counters: 16
09/03/22 11:39:15 INFO mapred.JobClient:   File Systems
09/03/22 11:39:15 INFO mapred.JobClient:     HDFS bytes read=320950
09/03/22 11:39:15 INFO mapred.JobClient:     HDFS bytes written=130568
09/03/22 11:39:15 INFO mapred.JobClient:     Local bytes read=168448
09/03/22 11:39:15 INFO mapred.JobClient:     Local bytes written=336932
09/03/22 11:39:15 INFO mapred.JobClient:   Job Counters
09/03/22 11:39:15 INFO mapred.JobClient:     Launched reduce tasks=1
```

範例一
動手做

WordCount1 練習(IV)



```
waue@vPro:/opt/hadoop$ bin/hadoop dfs -cat output/part-00000
Cloud   2
Course, 1
Course. 1
I       2
NCHC    1
and     1
course. 1
enjoy   1
like    2
nchc    1
this    1
we      1
```