Hadoop MapReduce

Workshop

R05922087 張逸寧
R05922114 林明璟
NTU CSIE

工作站環境

```
ssh <STUDENT_ID>@140.112.31.199
```

因網路安全因素, 僅限台大 內部 IP Address 可登入。

```
$ hadoop version

$ java -version
```

- Hadoop 2.7.2
- OpenJDK 1.7.0_95, 64-bit Server VM

Complie Java Program

Download WordCount.tar

```
wget --no-check-certificate http://judgegirl.csie.org/downloads/hadoop/WordCount.tar
tar xvf WordCount.tar
cd WordCount && make
```

After compiling

```
$ tree
    .
    ├── bin
    │   ├── WordCount.class
    │   ├── WordCount$Map.class
    │   └── WordCount$Reduce.class
    ├── Makefile
    ├── src
    │   └── WordCount.java
    ├── WordCount.jar
    └── WordCount.tar
```

Deploy Hadoop Job

```
Usage: hadoop jar <jar> [mainClass] args...
$ hadoop jar WordCount.jar WordCount $(INPUT) $(OUTPUT)
```

How to use HDFS ?

```
$ hdfs dfs
Usage: hdfs dfs [generic options]
       [-appendToFile <localsrc> ... <dst>]
       [-cat [-ignoreCrc] <src> ...]
       [-checksum <src> ...]
       [-chgrp [-R] GROUP PATH...]
       [-chmod [-R] <MODE[,MODE]... | OCTALMODE> PATH...]
       [-chown [-R] [OWNER][:[GROUP]] PATH...]
       [-copyFromLocal [-f] [-p] [-l] <localsrc> ... <dst>]
       [-copyToLocal [-p] [-ignoreCrc] [-crc] <src> ... <localdst>]
       [-count [-q] [-h] <path> ...]
       [-cp [-f] [-p | -p[topax]] <src> ... <dst>]
       ...
```

HDFS - User Path

- 假設使用者 <USER> 操作 hdfs dfs 相關指令
- 預設家目錄 /user/<USER>
- 查閱家目錄下的檔案操作

```
$ hdfs dfs -ls
$ hdfs dfs -ls /user/<USER>
```

HDFS - Example

運行前, 移除先前的輸出結果 (預設行為不可覆蓋檔案)

```
$ hdfs dfs -rm -r -f wordcnt-output
```

將測試資料放入 HDFS

```
$ hdfs dfs -copyFromLocal local-wordcnt-input wordcnt-input
```

運行後, 將運行結果抓回本機 查閱

```
$ hdfs dfs -copyToLocal wordcnt-output local-wordcnt-output
```

Website Monitor

HDFS

```
140.112.31.199:50070
```

YARN

```
140.112.31.199:8088
```

Workshop 1

WordCound

Problem Discription

- 計算文檔中的單詞
- 使用預設的 LineReader, 以行為單位分散處理有多少單詞
- 輸出格式按照 <STRING:TOKEN> <INTEGER:COUNT>

Word Count

Sample Input

```
see the word count, character count, line
count, and paragraph count. Learn how to insert
the word count, count words as you type, and
more
```

Sample Output

```
Learn   1
and     2
as      1
character       1
count   1
count,  4
count.  1
how     1
insert  1
line    1
more    1
paragraph       1
see     1
the     2
to      1
type,   1
word    2
words   1
you     1
```

Experiment

下載 WordCount.tar (http://judgegirl.csie.org/downloads/hadoop/WordCount.tar)

- [1 pt] 統計目前放在 HDFS 上的資料, 路徑位置為 /user/hadoop/wordcnt-input/*
- [1 pt] 生一組自己的輸入測資, 手動上傳到 HDFS 後, 運行一次 Word Count 得到預期的正確結果

請保留 **操作指令** 和 **輸入/輸出結果**

在 Workshop 時, 由助教檢查運行結果

Warning

程式寫壞，卡在佇列中。

```
# Find your JobId
$ hadoop job -list
# For example, <JobId>=job_1464687596301_0089
$ hadoop job -kill <JobId>
```

請勿隨意砍掉他人的程序，同學自行協調。

Workshop 2

Average

Problem Discription

給定數筆學生成績 (以行為單位), 請找出每個學生的平均分數

- 輸入保證每一行為單一筆學生資料, 其每行格式為 `<STRING:NAME> <INTEGER:SCORE>`
- 輸出時, 按照 `<STRING:NAME> <FLOAT:AVERAGE>` 輸出

約束

- `<STRING:NAME>` 只包含英文大小寫字母
- `<INTEGER:SCORE>` 為 0 到 100 (含) 之間的整數
- 保證每個學生的分數總和不超過 `32-bit integer`

Average

Sample Input

```
Amy 35
Bob 60
Amy 70
andy 88
david 100
andy 25
Amy 70
```

Sample Output

```
Amy 58.33
Bob 60.00
andy 56.50
david 100.00
```

Experiment

下載 Average-TODO.tar (http://judgegirl.csie.org/downloads/hadoop/Average-TODO.tar)

- [1 pt] Task I: 完成附檔所缺少的 Mapper/Reducer 函數, 並成功運行 /user/hadoop/avg-input/avg-input.large
- [2 pt] Task II: 完成附檔所缺少的 Combiner 函數, 運行後回報效能改善情況

請保留 **操作指令** 和 **輸入/輸出結果**

在 Workshop 時, 由助教檢查運行結果

參考算法

```
// Mapper
Amy 35        -> (Amy, (35, 1))
Bob 60        -> (Bob, (60, 1))
Amy 70        -> (Amy, (70, 1))
andy 88       -> (andy, (88, 1))
----- data partitioning -----
andy 25       -> (andy, (25, 1))
Amy 70        -> (Amy, (70, 1))
```

```
// Combiner
(Amy, [(35, 1), (70, 1)]) -> (Amy, (105, 2))
(Bob, (60, 1))            -> (Bob, (60, 1))
(andy, (88, 1))           -> (andy, (88, 1))
----- data partitioning -----
(andy, (25, 1))           -> (andy, (25, 1))
(Amy, (70, 1))            -> (Amy, (70, 1))
```

```
// Reducer
(Amy, [(105, 2), (70, 1)])  -> (Amy, 58.33)
(Bob, (60, 1))              -> (Bob, 60)
----- data partitioning -----
(andy, [(25, 1), (88, 1)])  -> (andy, 56.5)
```

Note

- 請優先以 `avg-input.small` 測試正確性
- 上述程序只是參考程序, 無強制使用 `<Text, IntPair>`
- 可使用自行設計的 `Mapper/Reducer/Combiner`

Reference

- Apache Hadoop 2.7.2 Document
- Michael G. Noll, Writing an Hadoop MapReduce Program in Python
- Morris God Hadoop Something