第一屆 全國大專校院 BigData 金象盃決賽題目

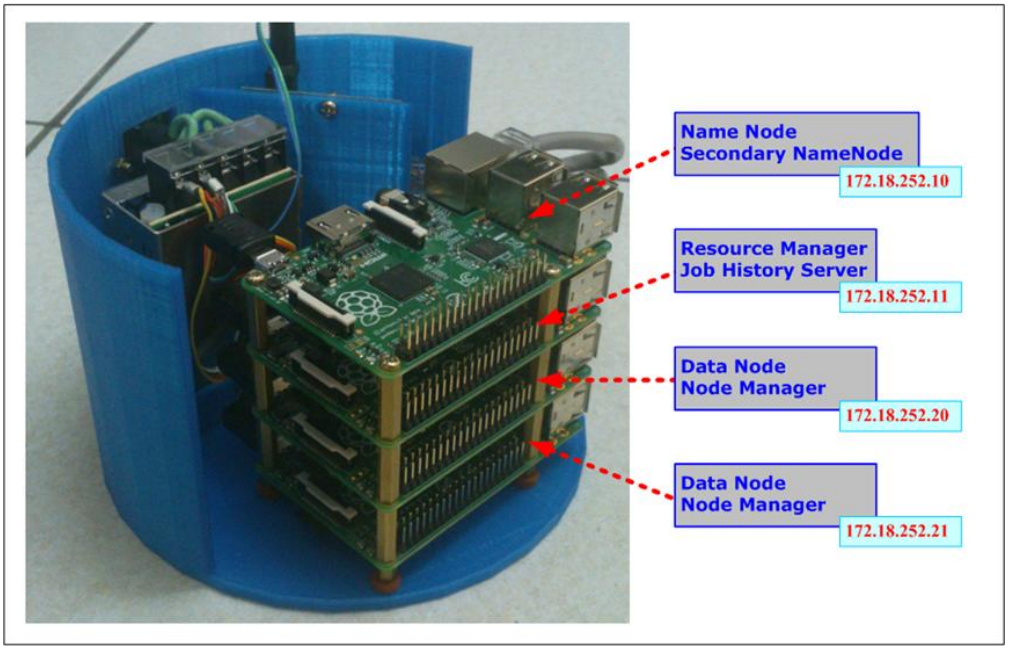
【決賽規則】

1. 決賽使用自造 Hadoop 小刀鋒(4 部樹莓派主機組成)，做為競賽主機，每部樹莓派主機，已預先安裝 RASPBIAN Debian Wheezy 作業系統，OpenJDK 7 及 Hadoop 2.5.2 套件(沒設定)。
2. 競賽方式，每隊只能派二名代表參加，決賽採用實機操作方式(4 部樹莓派主機)，各隊只能透過序列埠 (UART) 連接方式，各別登入 4 部樹莓派主機，從設定 IP 位址開始，逐一完成競賽題目。
3. 決賽採取現場制(競賽環境對外網路不連通)，請各隊依照比賽時間和地點參加競賽。
4. 報到時每隊隊長請洽考場人員領取『競賽資料袋』，內含：Hadoop 小刀鋒登入所需的帳號/密碼，決賽證明。主辦單位宣布競賽開始，方能統一拆開。
5. 排名會先依完成度為優先排序，再依時間做最後名次排序。

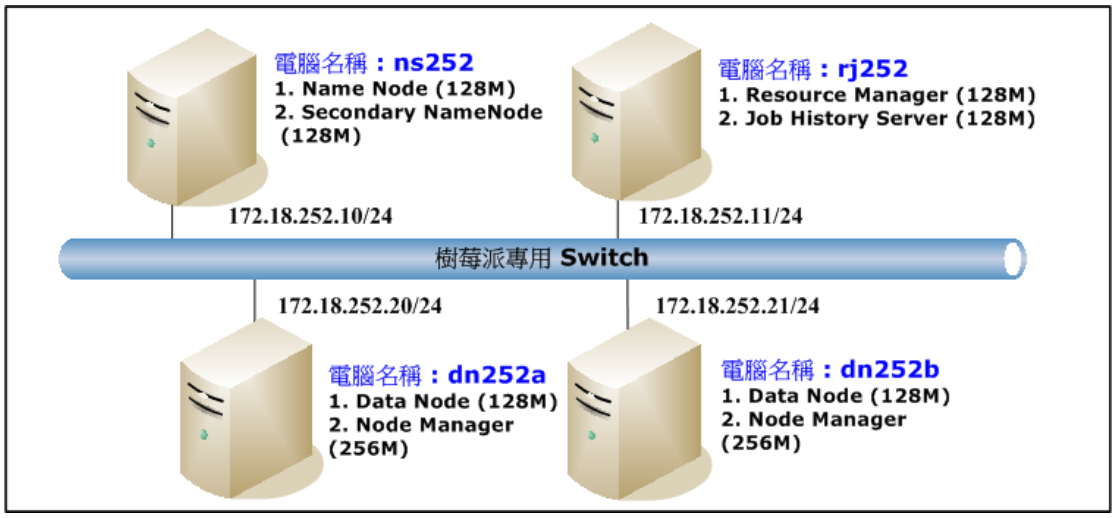
【決賽題目】

各隊使用主辦單位事先設定好的 終端機程式(putty) 登入 Hadoop 小刀鋒, 然後根據競賽題目的順序, 依序完成, 競賽最終目的是要完成 **自造 Hadoop 多點運算系統** 建置。

Hadoop 小刀鋒配置圖



自造 Hadoop 多點運算系統架構圖



請務必依照上圖中, 正確設定每個 Hadoop 服務 (NameNode, Resource Manager,...) 的記憶體大小。

【評分方式】

所有 **評分命令** 必須在指定的樹莓派主機中執行

1. 在 ns252 主機, 執行以下命令,確認 HDFS 分散檔案系統是否正常運作 (25%)

$ hdfs dfsadmin –report

```
Configured Capacity: 122331324416 (113.93 GB)
Present Capacity: 98226593792 (91.48 GB)
DFS Remaining: 98226544640 (91.48 GB)
DFS Used: 49152 (48 KB)
DFS Used%: 0.00%
Under replicated blocks: 0
Blocks with corrupt replicas: 0
Missing blocks: 0

-------------------------------------------------
Live datanodes (2):
```

2. 在 ns252 主機, 上載 103 年大專校院名錄檔 ulist.txt (已存在 ns252) 至 HDFS 分散檔案系統 (15%)

$ hdfs dfs -put ulist.txt

3. 在 rj252 主機, 執行以下命令, 確認 YARN 分散運算系統是否正常運作 (25%)

$ yarn node –list -all

```
14/11/27 17:18:51 INFO client.RMProxy: Connecting to ResourceManager at rj252/172.18.252.11:8032
Total Nodes:2
         Node-Id             Node-State Node-Http-Address       Number-of-Running-Containers
    dn252a:51034                RUNNING       dn252a:8042                                  0
    dn252b:49459                RUNNING       dn252b:8042                                  0
```

4. 在 rj252 主機, 執行 MapReduce 程式 (15%)

$ hadoop jar /opt/hadoop-2.5.2/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.5.2.jar pi 1 1

::

```
        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=236
        File Output Format Counters
                Bytes Written=97
Job Finished in 209.835 seconds
Estimated value of Pi is 4.00000000000000000000
```

5. 在 rj252 主機, 執行 ulist.pig 程序檔 (已存在 rj252 主機), 計算全國各縣市的大學總數 (20%)

$ pig ulist.pig

檢視分析結果

$ hdfs dfs -cat sr/part-r-00000

[01]新北市 19
[02]宜蘭縣 4
[03]桃園縣 12
[04]新竹縣 2
[05]苗栗縣 4
[06]臺中市 6
[07]彰化縣 5
[08]南投縣 2
[09]雲林縣 3
[10]嘉義縣 4
[11]臺南市 11
[12]高雄市 10
[13]屏東縣 5
[14]臺東縣 2
[15]花蓮縣 5
[16]澎湖縣 1

[17]基隆市 3
[18]新竹市 6
[19]臺中市 11
[20]嘉義市 3
[21]臺南市 5
[32]臺北市 1
[33]臺北市 5
[34]臺北市 2
[35]臺北市 2
[38]臺北市 3
[39]臺北市 1
[40]臺北市 3
[41]臺北市 4
[42]臺北市 5
[52]高雄市 1
[54]高雄市 2
[55]高雄市 4
[58]高雄市 1
[61]高雄市 1
[71]金門縣 1

HDFS 設定檔

1. core-site.xml

```
<configuration>
  <property>
      <name>fs.default.name</name>
      <value> </value>
  </property>
</configuration>
```

2. hdfs-site.xml

```
<configuration>
  <property>
    <name>dfs.replication</name>
    <value>      </value>
  </property>
  <property>
    <name>dfs.namenode.name.dir</name>
    <value>       </value>
  </property>
  <property>
    <name>dfs.datanode.data.dir</name>
    <value>       </value>
  </property>
  <property>
    <name>fs.permissions.umask-mode</name>
    <value>       </value>
  </property>
</configuration>
```

3. hadoop-env.sh

```
                    ::
# The java implementation to use.
export JAVA_HOME=
                    ::
The maximum amount of heap to use, in MB. Default is 1000.
export HADOOP_HEAPSIZE=
                    ::
export HADOOP_CLIENT_OPTS=
                    ::
export HADOOP_LOG_DIR=
```

YARN 設定檔

1. mapred-site.xml

```
<?xml version="1.0"?>
<configuration>
    <property>
        <name>mapreduce.framework.name</name>
        <value>        </value>
    </property>
</configuration>
```

2. yarn-site.xml

```
<?xml version="1.0"?>
<configuration>
  <property>
      <name>yarn.nodemanager.aux-services</name>
      <value>          </value>
  </property>
  <property>
      <name>yarn.nodemanager.aux-services.mapreduce.shuffle.class</name>
      <value>        </value>
  </property>
  <property>
    <name>yarn.resourcemanager.webapp.address</name>
    <value>       </value>
  </property>
  <property>
    <name>yarn.resourcemanager.hostname</name>
    <value>       </value>
  </property>
  <property>
    <name>yarn.nodemanager.local-dirs</name>
    <value>     </value>
  </property>
  <property>
    <name>mapreduce.jobhistory.address</name>
    <value>       </value>
  </property>
</configuration>
```

3. mapred-env.sh

::
export HADOOP_JOB_HISTORYSERVER_HEAPSIZE=
::
JobHistory log file path, 會自動產生目錄
export HADOOP_MAPRED_LOG_DIR=

4. yarn-env.sh

::
JAVA_HEAP_MAX=
::
YARN_HEAPSIZE=
::
default log directory & file
export YARN_LOG_DIR= **(加入這一行)**
if ["$YARN_LOG_DIR" = ""]; then
  YARN_LOG_DIR="$HADOOP_YARN_HOME/logs"
fi