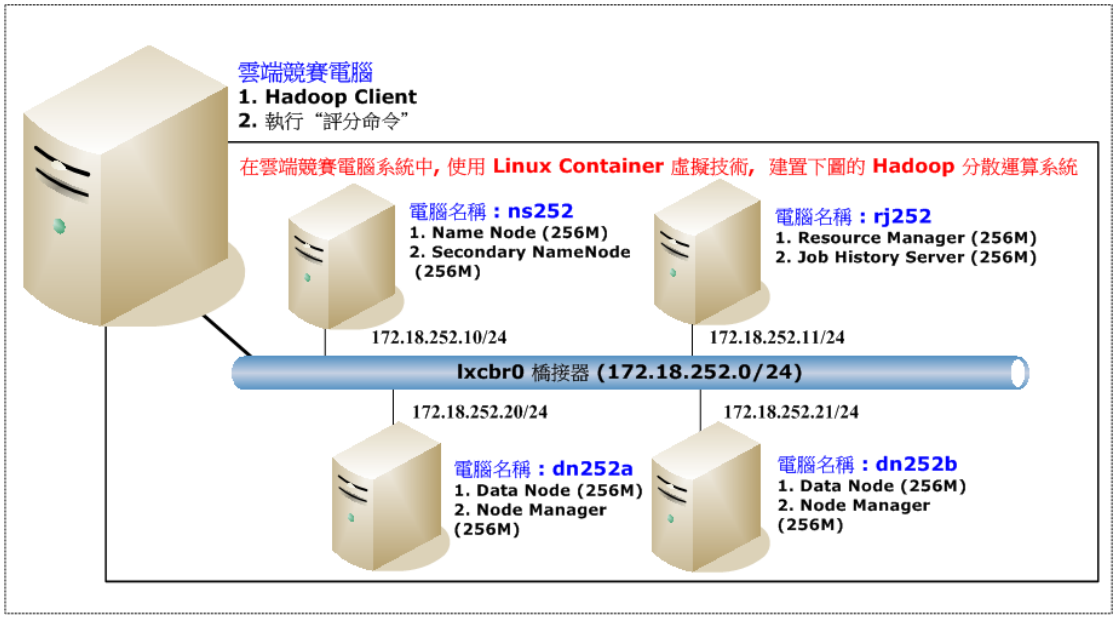
第一屆 全國大專校院 BigData 金象盃初賽題目

登入 **Big Data 競賽電腦** 的帳號是 **bigred**, 密碼是 **2Password!**, 在 **Big Data 競賽電腦** 的系統內,已存在一台名字為 **bigred** 的 Linux Container (LXC) 虛擬電腦, 登入帳號及密碼均為 **ubuntu**. 此部 LXC 虛擬電腦(bigred), 主要做為模板運算電腦, 各隊必須使用此模板運算電腦, 來建置其它 Hadoop 分散運算電腦。

在各隊 **Big Data 競賽電腦** 系統中的 /opt 目錄, 已存在以下檔案

1. hadoop-2.5.2.tar.gz
2. jdk-7u71-linux-i586.tar.gz
3. pig-0.14.0.tar.gz
4. ulist.txt
5. ulist.pig
6. hadoopxml.txt (HDFS, YARN 設定範本檔)

請利用上述的模板運算電腦及 /opt 目錄中的檔案, 完成建置下圖中的多點 Hadoop 分散運算系統.



1. **ns252** : 啟動 Name Node 及 Secondary NameNode 這二個服務,
   IP 位址為 172.18.252.10
2. **rj252** : 啟動 Resource Manager 及 Job History Server 這二個服務,
   IP 位址為 172.18.252.11
3. **dn252a** : 啟動 Data Node 及 Node Manager 這二個服務,
   IP 位址為 172.18.252.20
4. **dn252b**: 啟動 Data Node 及 Node Manager 這二個服務,
   IP 位址為 172.18.252.21

【評分方式】

所有 **評分命令** 均在各隊的 **Big Data 競賽電腦** 系統中執行, 不會在任何一部 LXC 虛擬電腦系統 (ns252,rj252,dn252a,dn252b) 中執行

1. 確認使用 LXC 虛擬技術完成 Hadoop 分散運算系統, 所需的四部運算電腦建置 (20%)

$ sudo lxc-ls –f

```
NAME    STATE    IPV4           IPV6  AUTOSTART
------------------------------------------------
bigred  STOPPED  -              -     NO
dn252a  RUNNING  172.18.252.20  -     NO
dn252b  RUNNING  172.18.252.21  -     NO
ns252   RUNNING  172.18.252.10  -     NO
rj252   RUNNING  172.18.252.11  -     NO
```

2. 確認 HDFS 分散檔案系統正常運作 (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):
```

3. 確認 YARN 分散運算系統正常運作 (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. 執行 MapReduce 程式 (10%)

$ 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. 執行 Pig 資料分析 (20%)

上載 103 年大專校院名錄檔 (ulist.txt) 至 HDFS 分散檔案系統

$ hdfs dfs -put ulist.txt

編輯 ulist.pig 程序檔

$ nano ulist.pig

```
school = LOAD 'ulist.txt' AS (sno:int,name:chararray,city:chararray);
frec = FILTER school by sno is not null OR city != '縣市名稱';
sdiv = GROUP frec BY city;
counts = foreach sdiv generate group,COUNT(frec);
store counts into 'sr';
```

執行 ulist.pig 程序檔

$ 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
```