## **HW5 - spark**

**組員 : 0853408羅廷君、0853412吳宛儒、0853413楊雅雯、 0853418張巧欣**

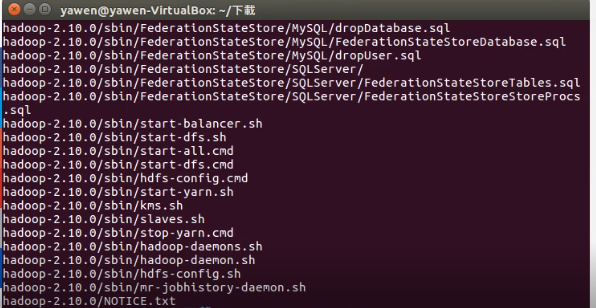
**I. 截圖與步驟描述**

**Virtual Box、Ubuntu安裝設定**

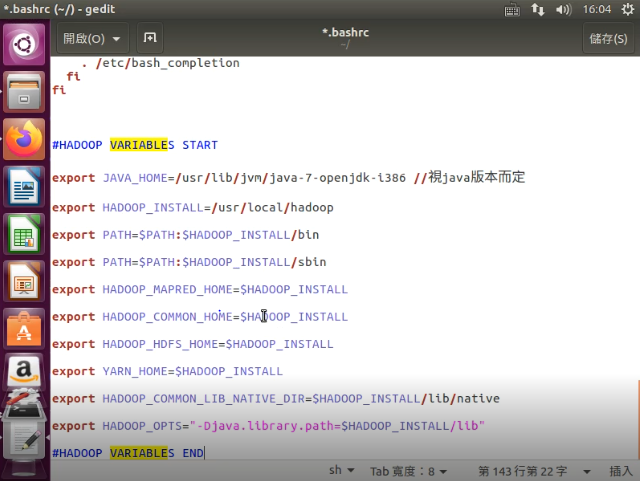
1. **安裝virtual box**
2. **virtual box 上裝 ubuntu**
3. **安裝JDK**

****

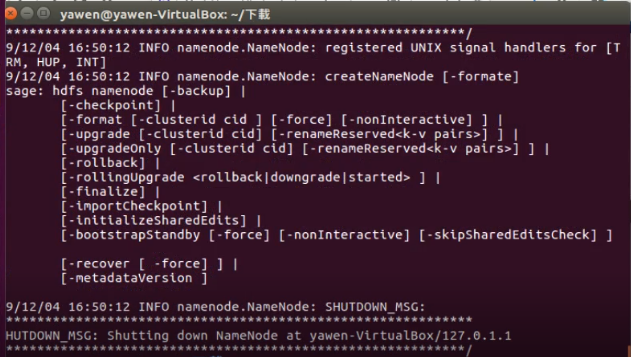
1. **設定SSH無密碼登入**
2. **下載安裝Hadoop**

****

1. **Hadoop各項設定**

****

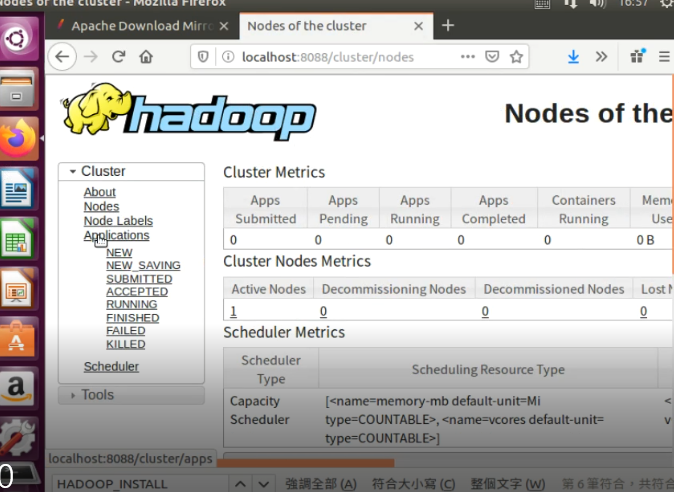
1. **建立與格式化HDFS目錄**

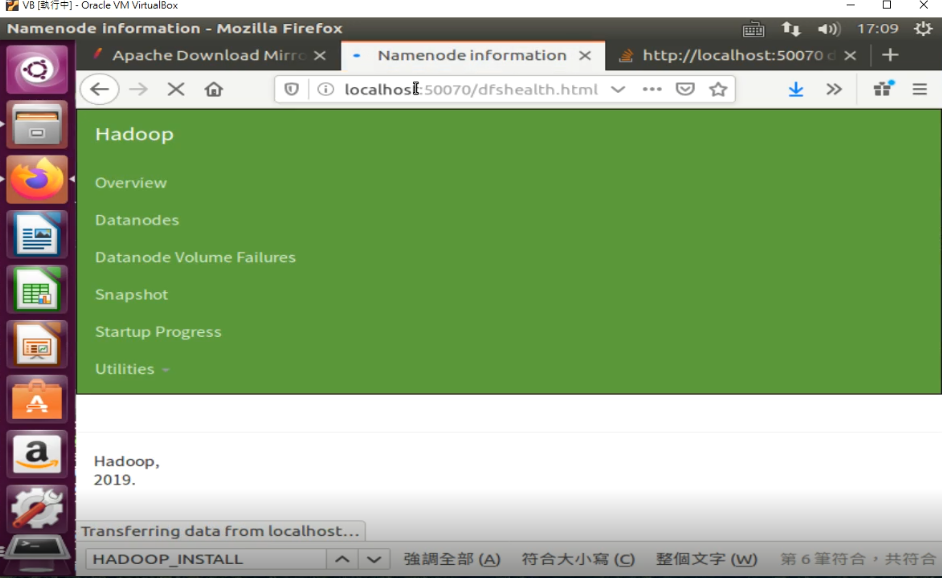
****

1. **開啟Hadoop ResourceManager Web介面**
2. **start Hadoop**

****

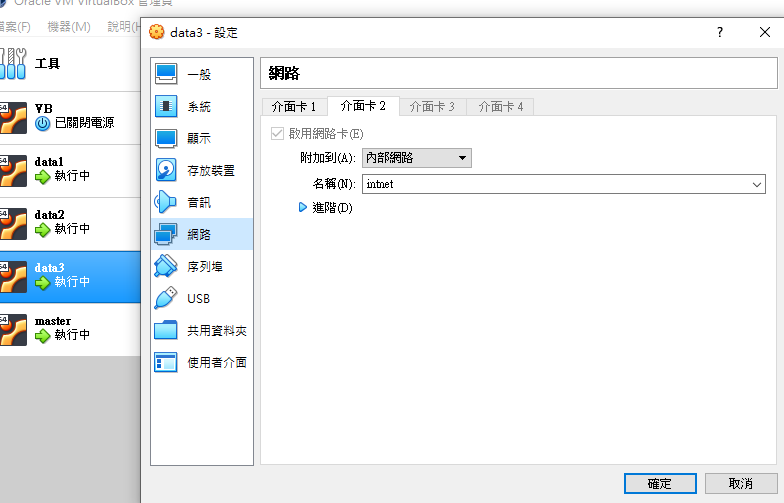
1. **NameNode HDFS Web介面**

****

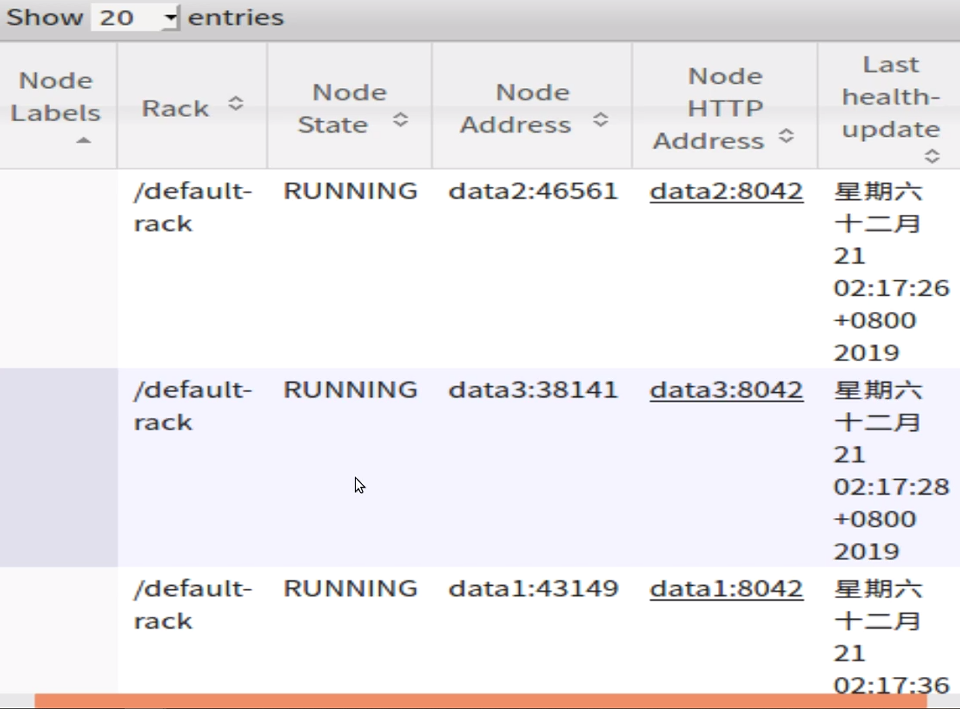
****

**Hadoop Multi Node Cluster安裝**

1. **複製 single node cluster到data1**
2. **virtual box介面卡設定**
3. **設定data1伺服器**
4. **複製data1伺服器至data2、data3、master**

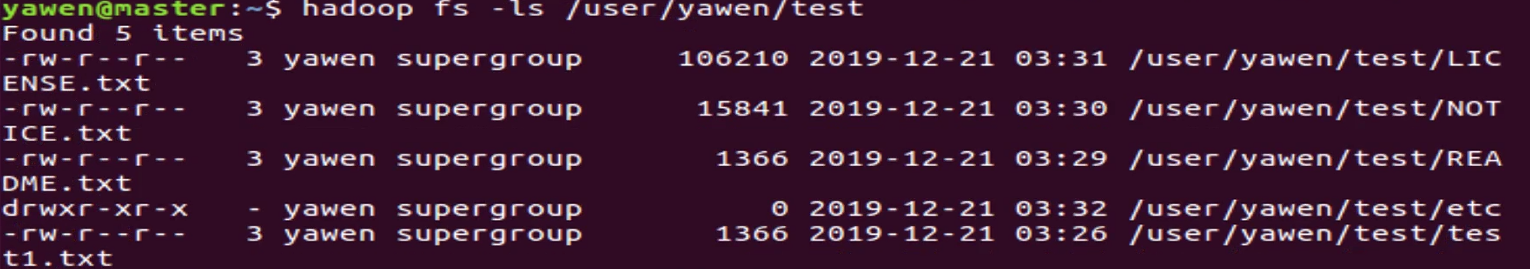
****

1. **設定data2、data3、master伺服器**
2. **建立與格式化NameNode HDFS目錄**
3. **開啟Hadoop ResourceManager Web**

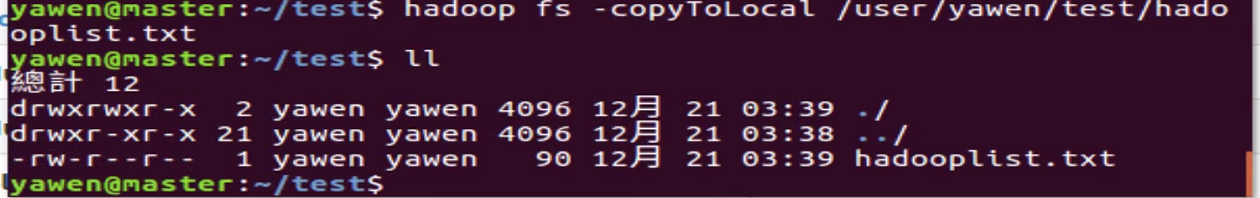
****

**Hadoop HDFS命令介紹**

1. **啟動Hadoop Multi-Node Cluster**
2. **建立與查看HDFS目錄**

****

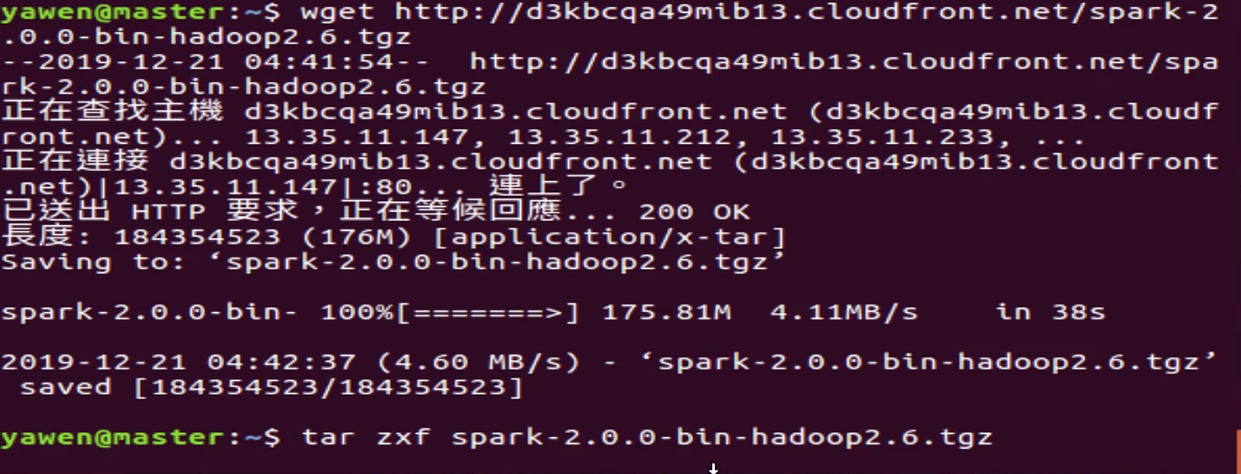
1. **從本機複製檔案到HDFS**
2. **將HDFS上的檔案複製到本機**

****

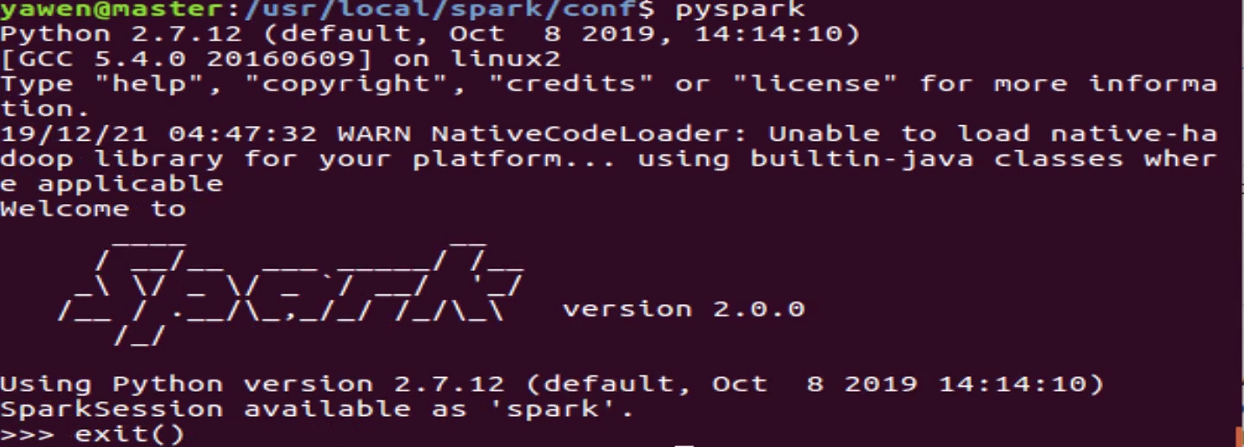
1. **複製與刪除HDFS檔案**

**Python Spark介紹與安裝**

1. **安裝Scala介紹與安裝**
2. **安裝spark**

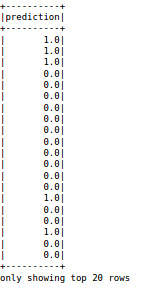
****

1. **啟動pyspark互動介面**

****

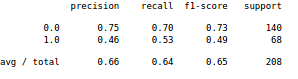
**II.作業目標對應之結果**

1. **Dataset : character-deaths.csv**
2. **預測結果(前20筆):**

****

1. **accuracy、precision、recall**

****

****

**III.討論**

做hadoop時，因為版本已經更新了，和老師所推薦使用的課本不一樣(課本為Ubuntu 14.04版)，所以**花了很多時間**在安裝設定部分，尤其要輸入的指令、要更改的設定檔很多，手動輸入容易打錯，再加上Hadoop Multi Node Cluster 這邊，打指令經常有問題，datanode或namenode會不出現，導致live node沒有出現、數量不正確，後來重做了很多次才成功。

預測結果的準確度(accuracy)大約64%，precision和recall也大約在六成左右。在target的部分有許多缺失值，我們以補零的方式填補，而train data裡面較不會有這樣的問題，缺失值很可能導致資訊不夠準確，因此預測準確度沒辦法達到很高，也許需要在調整參數或是換其他模型才能再增加精準度。可能特徵是角色在書本出現的時間與是否為貴族和年齡等特徵，可以分析的特徵其實並不是很多，也因此導致預測不準確。而且冰與火之歌是在爭權奪位的奇幻故事，女性角色比較不回上戰場，女生的存活率會比較高，貴族會去打仗，所以死亡率較高。