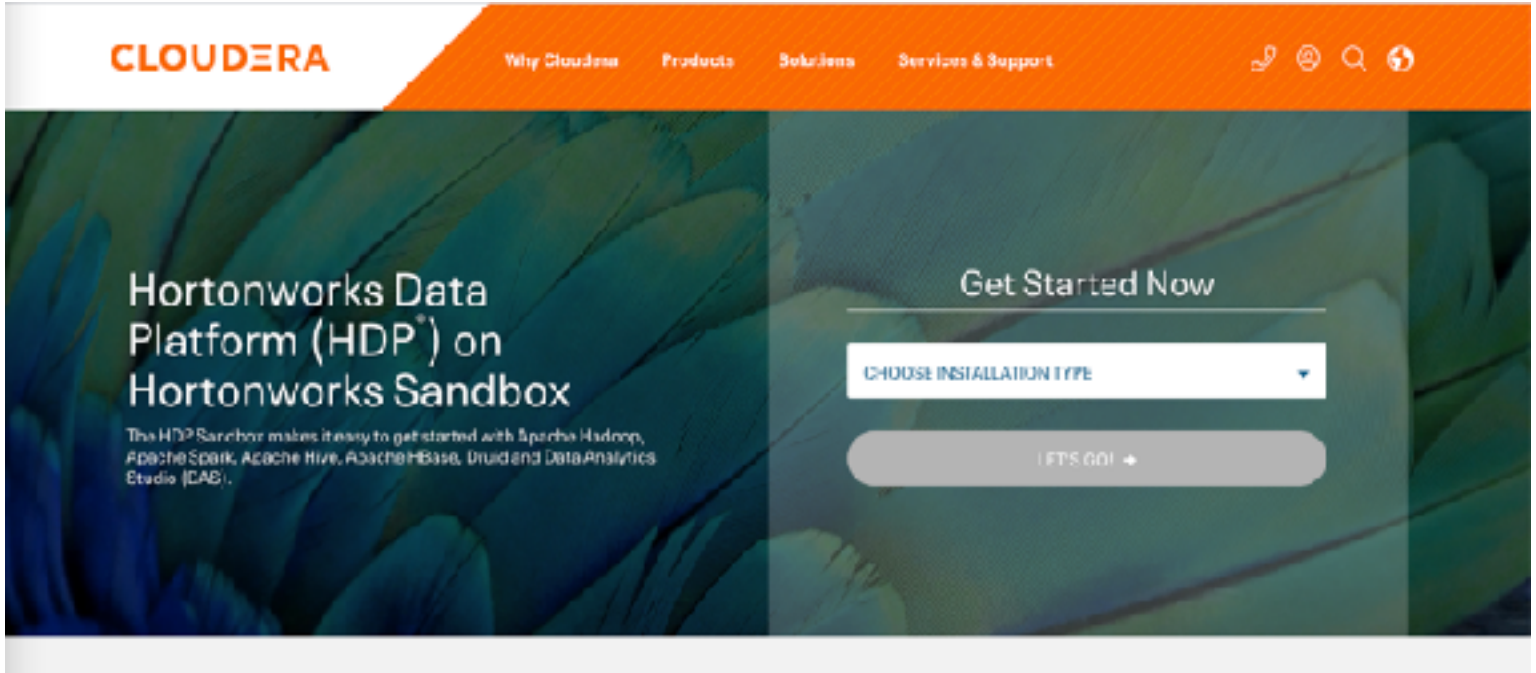
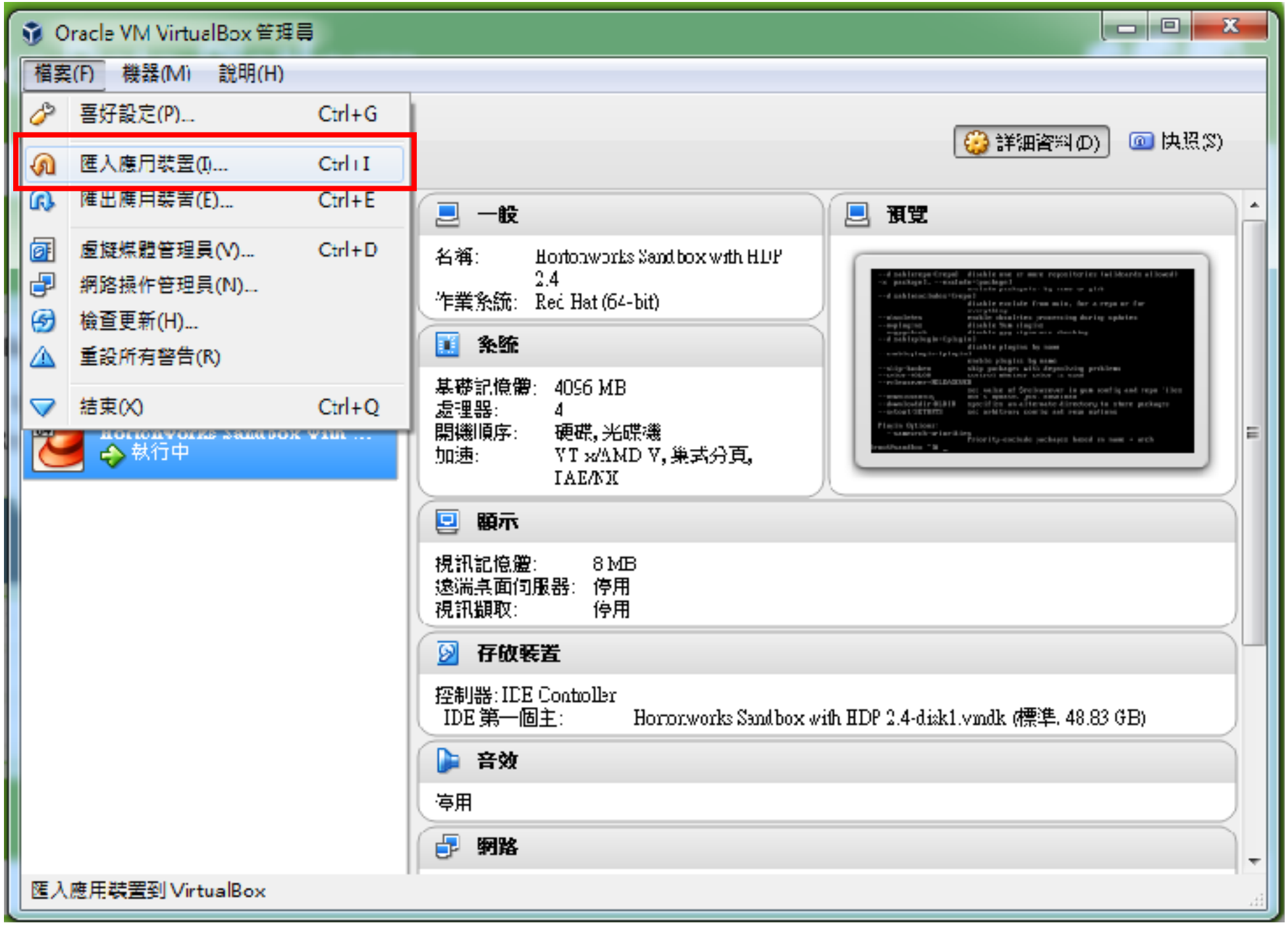
How to Install Hortontworks Sandbox and Build a MapReduce program

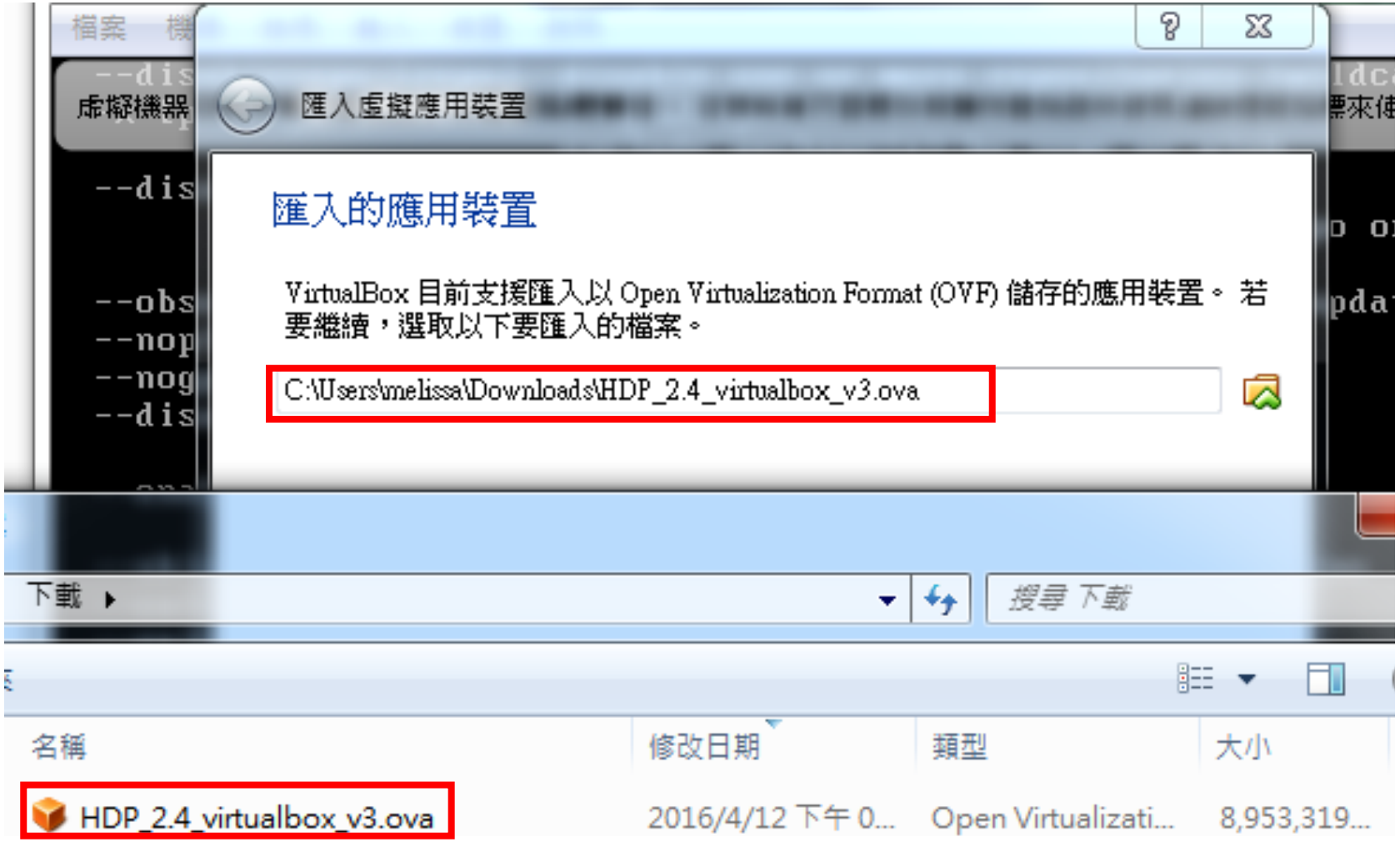
2021.09

Download Virtualbox and Sandbox

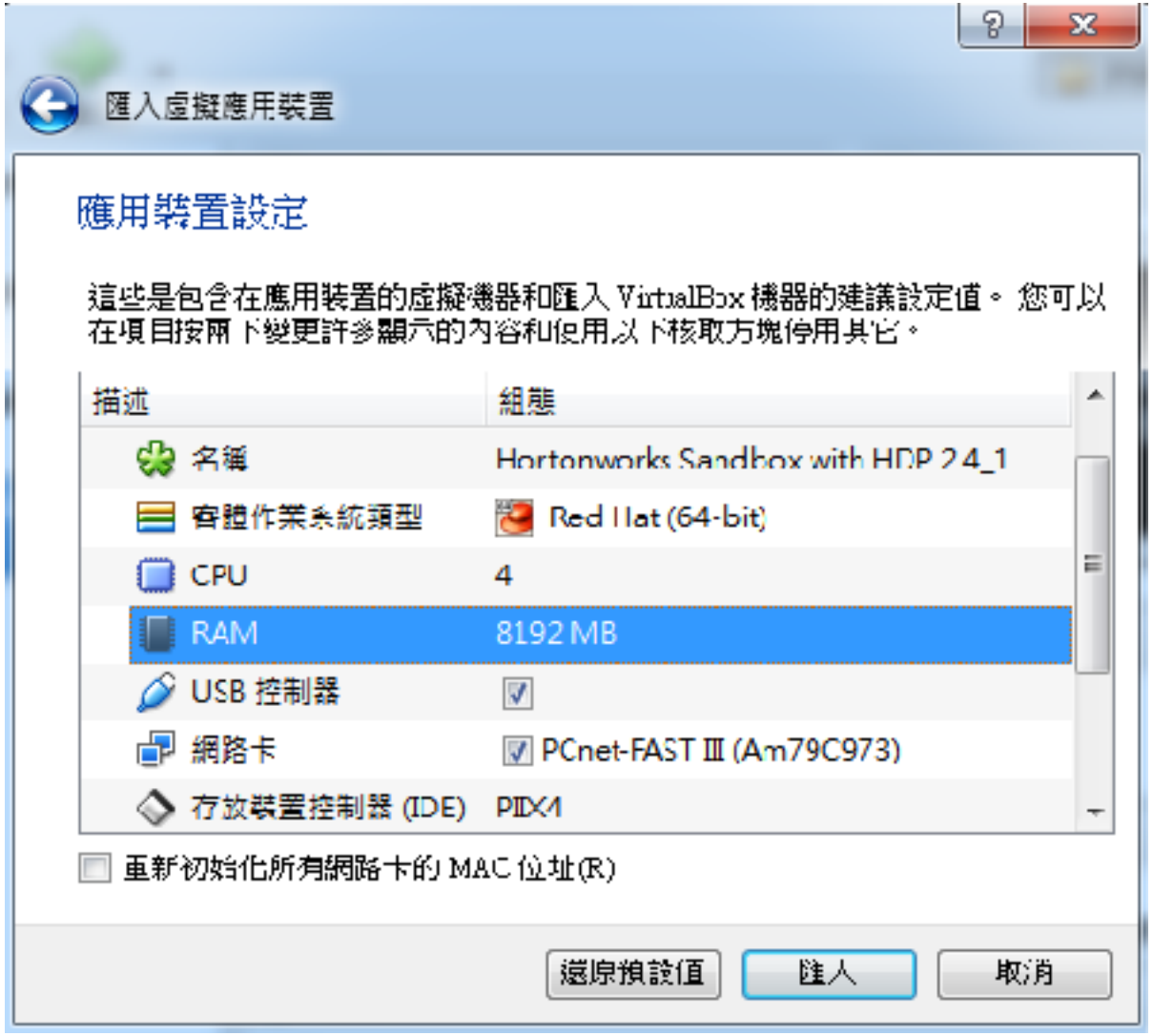
- Download Oracle VM Virtualbox
  - https://www.virtualbox.org/
- Download Hortonworks Sandbox (virtualbox)
  - https://www.cloudera.com/downloads/hortonworks-sandbox/hdp.html



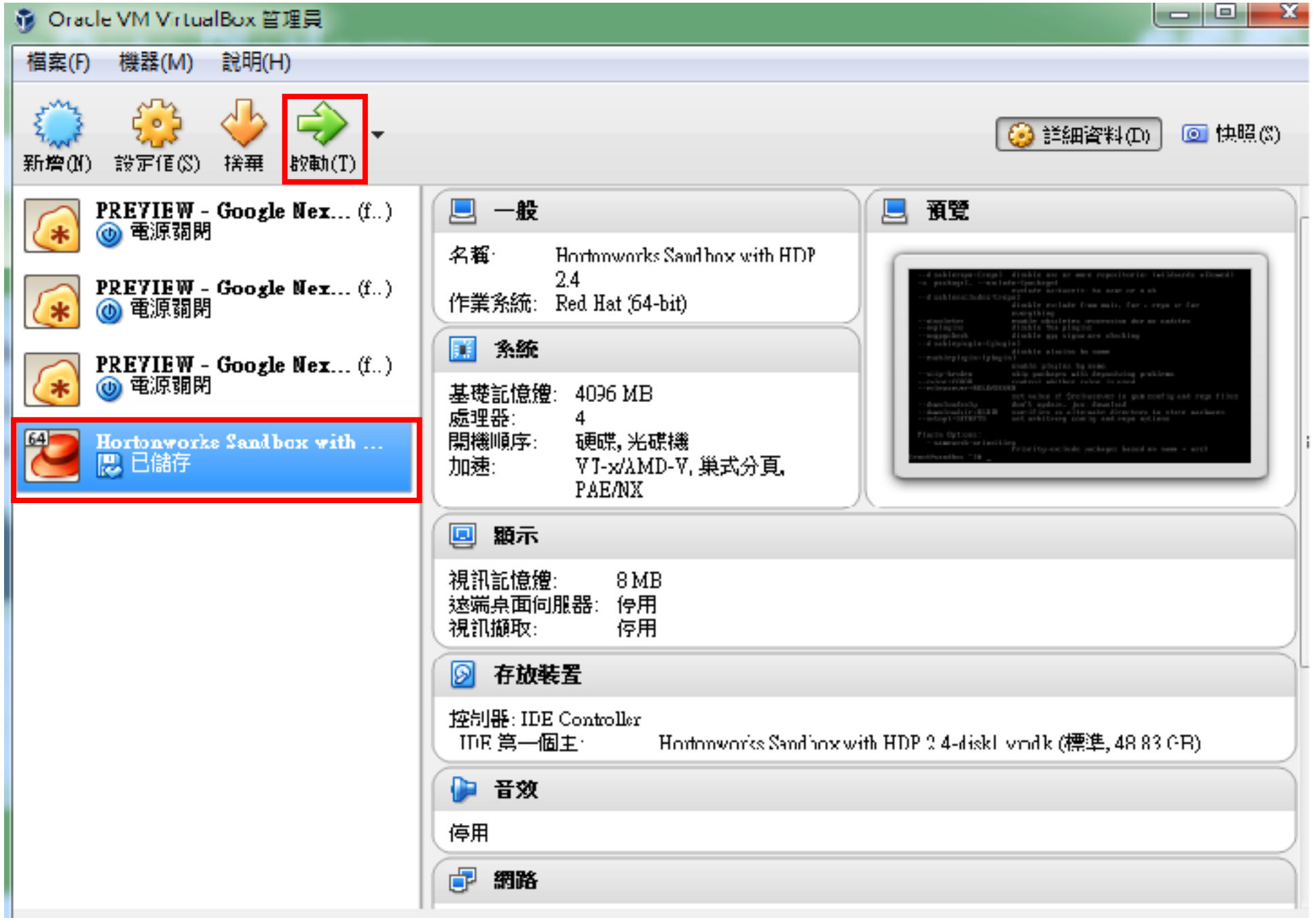

Oracle VM VirtualBox 管理員
檔案(F)
機器(M)
說明(H)
喜好設定(P)... Ctrl+G
匯入應用裝置(I)... Ctrl+I
匯出應用裝置(E)... Ctrl+E
虛擬媒體管理員(V)... Ctrl+D
網路操作管理員(N)...
檢查更新(H)...
重設所有警告(R)
結束(X) Ctrl+Q
執行中
詳細資料(D)
快照(S)
一般
名稱: Hortonworks Sandbox with HDP 2.4
作業系統: Red Hat (64-bit)
預覽
系統
基礎記憶體: 4096 MB
處理器: 4
開機順序: 硬碟, 光碟機
加速: VT-x/AMD-V, 巢式分頁, PAE/NX
顯示
視訊記憶體: 8 MB
遠端桌面伺服器: 停用
視訊擷取: 停用
存放裝置
控制器: IDE Controller
IDE 第一個主: Hortonworks Sandbox with HDP 2.4-disk1.vmdk (標準, 48.83 GB)
音效
停用
網路
匯入應用裝置到 VirtualBox

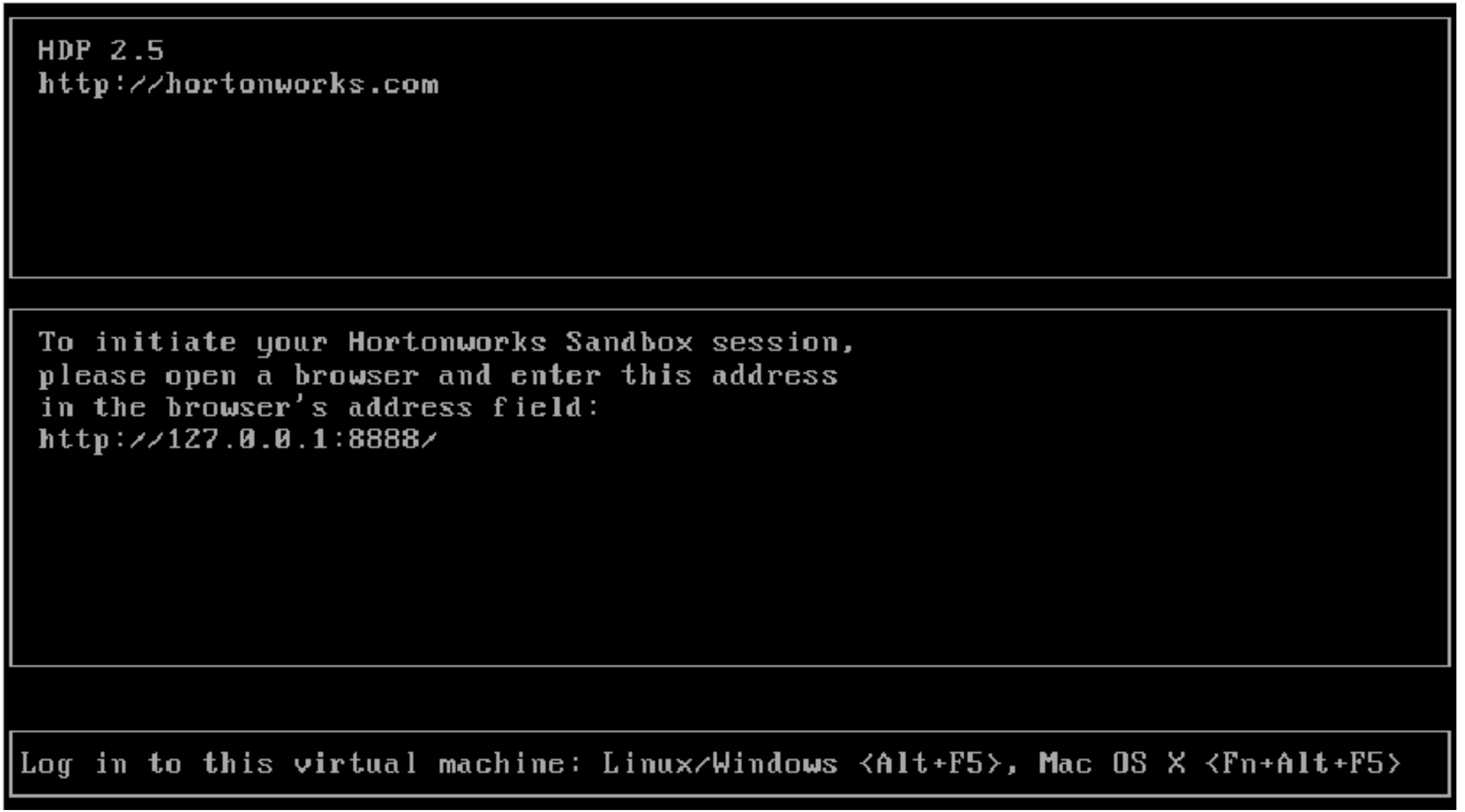

檔案
虛擬機器
匯入虛擬應用裝置
匯入的應用裝置
VirtualBox 目前支援匯入以 Open Virtualization Format (OVF) 儲存的應用裝置。 若要繼續，選取以下要匯入的檔案。
C:\Users\melissa\Downloads\HDP_2.4_virtualbox_v3.ova
下載
搜尋 下載
名稱
修改日期
類型
大小
HDP_2.4_virtualbox_v3.ova
2016/4/12 下午 0...
Open Virtualizati...
8,953,319...

PLEASE check your computer's RAM size and make sure your computer can handle with the setting of RAM.



Ex: My computer's RAM size is 8G, so I set **4096 MB** here.


Oracle VM VirtualBox 管理員
檔案(F)
機器(M)
說明(H)
新增(N)
設定值(S)
捨棄
啟動(T)
詳細資料(D)
快照(S)
PREVIEW - Google Nex... (f..)
電源關閉
PREVIEW - Google Nex... (f..)
電源關閉
PREVIEW - Google Nex... (f..)
電源關閉
Hortonworks Sandbox with ...
已儲存
一般
名稱: Hortonworks Sandbox with HDP 2.4
作業系統: Red Hat (64-bit)
系統
基礎記憶體: 4096 MB
處理器: 4
開機順序: 硬碟, 光碟機
加速: VT-x/AMD-V, 巢式分頁, PAE/NX
預覽
顯示
視訊記憶體: 8 MB
遠端桌面伺服器: 停用
視訊擷取: 停用
存放裝置
控制器: IDE Controller
IDE 第一個主: Hortonworks Sandbox with HDP 2.4-disk1.vmdk (標準, 48.83 GB)
音效
停用
網路


HDP 2.5
http://hortonworks.com
To initiate your Hortonworks Sandbox session,
please open a browser and enter this address
in the browser's address field:
http://127.0.0.1:8888/
Log in to this virtual machine: Linux/Windows <Alt+F5>, Mac OS X <Fn+Alt+F5>

Connect using SSH

You can use Terminal or MobaXterm(Windows) to connect to the server.

- ssh root@127.0.0.1 -p 2222
- Password: hadoop
- Change the password when you first time login

Install Maven(1/2)

wget http://mirror.cc.columbia.edu/pub/software/apache/maven/maven-3/3.0.5/binaries/apache-maven-3.0.5-bin.tar.gz

tar xzf apache-maven-3.0.5-bin.tar.gz -C /usr/local

cd /usr/local

ln -s apache-maven-3.0.5 maven

sudo vi /etc/profile.d/maven.sh

加入

```
export M2_HOME=/usr/local/maven
export PATH=${M2_HOME}/bin:${PATH}
```

Install Maven(2/2)

- Finally, log out and log in again to activate the above environment variables.
- To verify successful installation of maven, check the version of maven:

```
mvn -version
```

```
[root@sandbox ~]# mvn -version
Apache Maven 3.0.5 (r01de14724cdef164cd33c7c8c2fe155faf9602da; 2013-02-19 13:51:28+0000
Maven home: /usr/local/maven
Java version: 1.7.0_95, vendor: Oracle Corporation
Java home: /usr/lib/jvm/java-1.7.0-openjdk-1.7.0.95.x86_64/jre
Default locale: en_US, platform encoding: UTF-8
OS name: "linux", version: "2.6.32-573.18.1.el6.x86_64", arch: "amd64", family: "unix"
```

Start a new MapReduce

- https://azure.microsoft.com/en-us/documentation/articles/hdinsight-develop-deploy-java-mapreduce-linux/
- (root/.m2 裡面建立一個檔案settings.xml，然後貼上下圖的程式)

```
<settings>
    <mirrors>
        <mirror>
          <id>controlhttps</id>
          <mirrorOf>central</mirrorOf>
          <name>Maven central https</name>
          <url>http://insecure.repo1.maven.org/maven2/</url>
        </mirror>
      </mirrors>
      </settings>
```

- FROM **建立 Maven 專案**
- TO **建置應用程式**
- **編輯檔案時如果遇到notepad command not found，請改用sudo vi 檔案名稱編輯**

Copy your input file to hdfs

- Make a new file on hdfs

```
hadoop fs -mkdir -p /user/root/data/
```

- Then copy your input file to hdfs

```
hadoop fs -copyFromLocal inputfile.txt /user/root/data/
```

- See more hadoop commands in the last page

Run the MapReduce

```
cd wordcountjava/target
```

```
yarn jar wordcountjava-1.0-SNAPSHOT.jar
org.apache.hadoop.examples.WordCount /user/
root/data/inputfile.txt output/out1
```

- So you can find your output in **/user/root/** on hdfs

```
hadoop fs -cat /user/root/output/out1/*
```

Hadoop commands

- Make a directory name “remote”

```
hadoop fs -mkdir /remote
```

- List files recursive in directory “remote”
(-ls: list, -ls -R: list recursive)

```
hadoop fs -ls -R /remote
```

- Print every file in directory “remote”

```
hadoop fs -cat /remote/*
```

- Remove directory “remote” recursive (-rm: remove file)

```
hadoop fs -rm -R /remote
```

- List all command options

```
hadoop fs -help
```