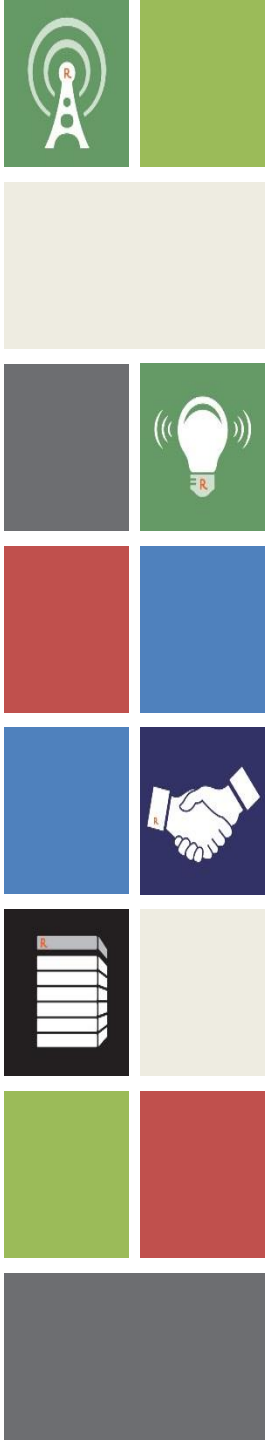
採智科技

Revolution R Enterprise

使用教學

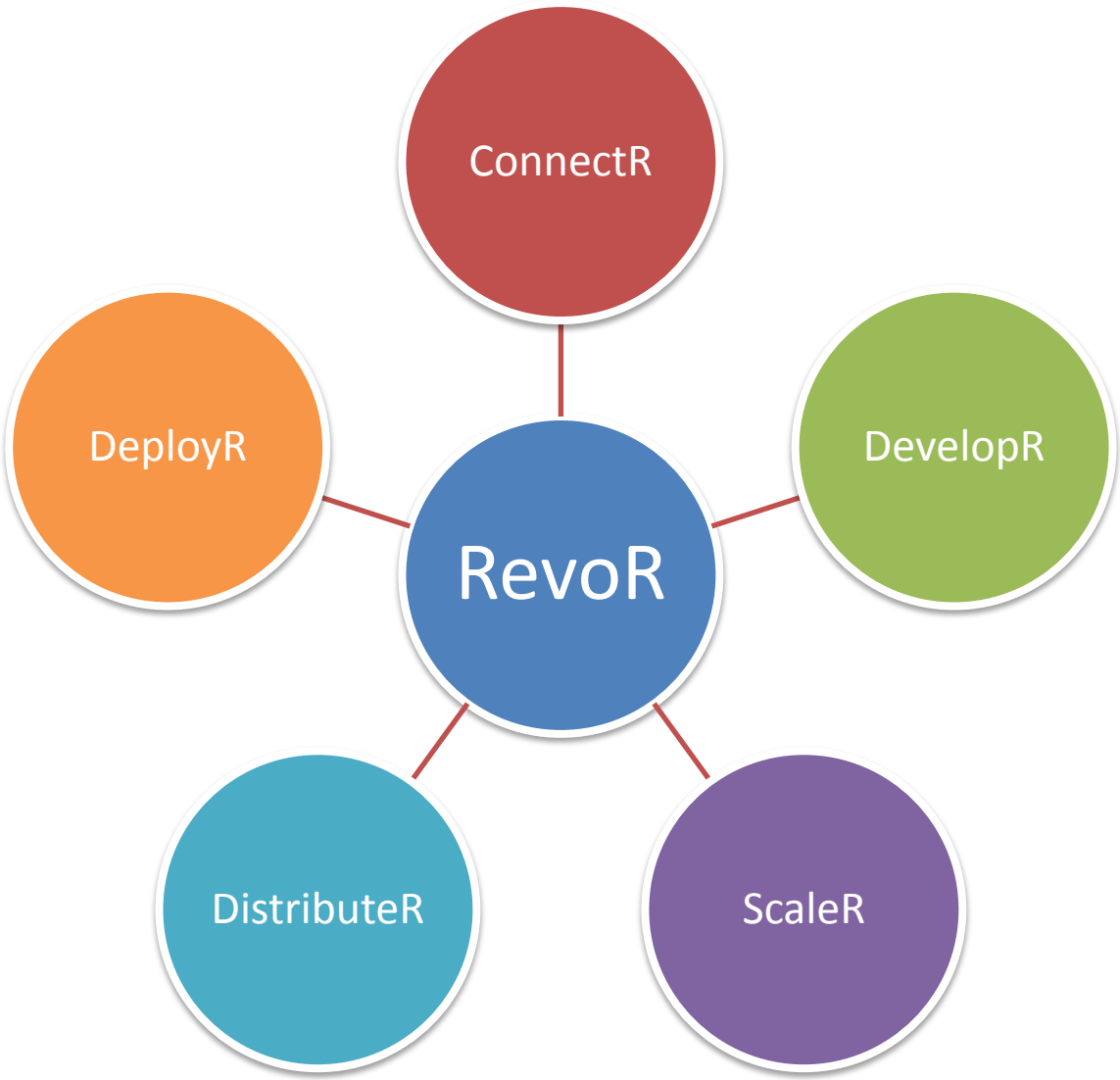
李秉鴻



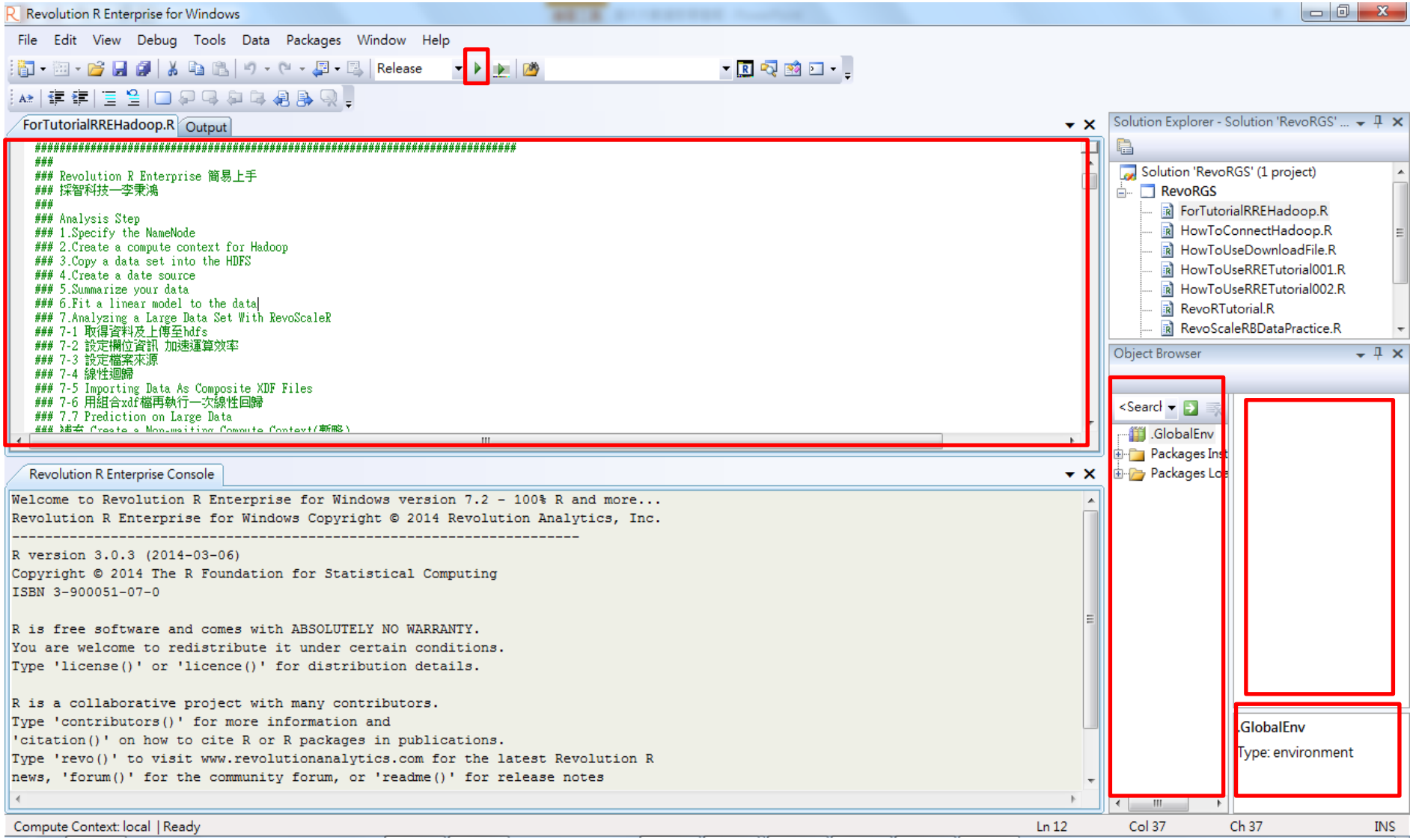
目錄

- RRE單機使用方式
- Linux簡易教學
- Hdfs指令延伸
- RRE操作Hdfs的函數
- RRE In Hadoop
- 附錄RRE安裝及Hadoop設定教學

Revolution R



簡易界面介紹


Revolution R Enterprise for Windows
File Edit View Debug Tools Data Packages Window Help
Release
ForTutorialRREHadoop.R
Output
Revolution R Enterprise 簡易上手
採智科技—李秉鴻
Analysis Step
1.Specify the NameNode
2.Create a compute context for Hadoop
3.Copy a data set into the HDFS
4.Create a date source
5.Summarize your data
6.Fit a linear model to the data
7.Analyzing a Large Data Set With RevoScaleR
7-1 取得資料及上傳至hdfs
7-2 設定欄位資訊 加速運算效率
7-3 設定檔案來源
7-4 線性迴歸
7-5 Importing Data As Composite XDF Files
7-6 用組合xdf檔再執行一次線性回歸
7.7 Prediction on Large Data
Revolution R Enterprise Console
Welcome to Revolution R Enterprise for Windows version 7.2 - 100% R and more...
Revolution R Enterprise for Windows Copyright © 2014 Revolution Analytics, Inc.
R version 3.0.3 (2014-03-06)
Copyright © 2014 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
Type 'revo()' to visit www.revolutionanalytics.com for the latest Revolution R
news, 'forum()' for the community forum, or 'readme()' for release notes
Solution Explorer - Solution 'RevoRGS' ...
Solution 'RevoRGS' (1 project)
RevoRGS
ForTutorialRREHadoop.R
HowToConnectHadoop.R
HowToUseDownloadFile.R
HowToUseRRETutorial001.R
HowToUseRRETutorial002.R
RevoRTutorial.R
RevoScaleRBDataPractice.R
Object Browser
<Search
.GlobalEnv
Packages Inst
Packages Loa
.GlobalEnv
Type: environment
Compute Context: local | Ready
Ln 12
Col 37
Ch 37
INS

分析流程

資料匯入

資料探索

資料整理

資料分析

結果視覺化

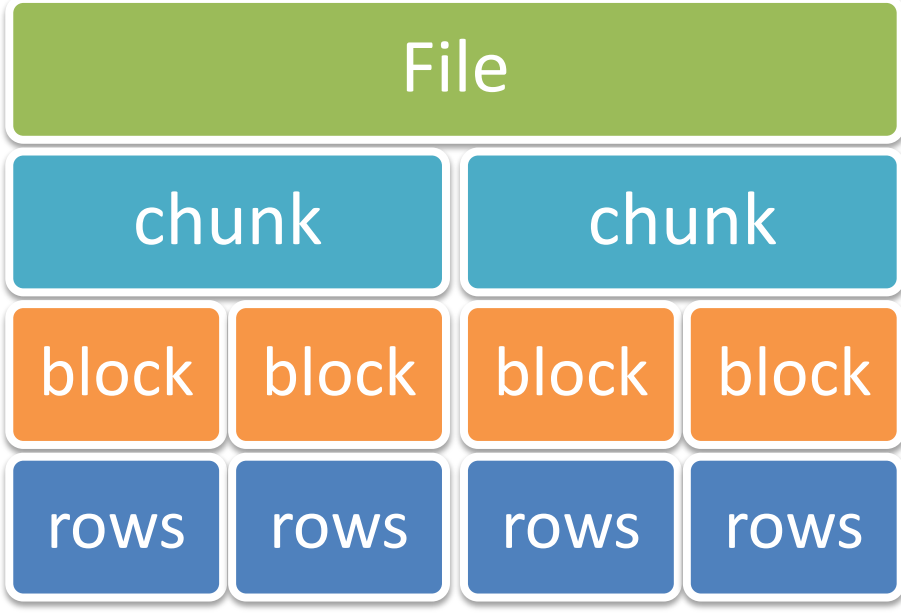
資料匯入-Xdf

單機使用可載入多種檔案, 但建議載入Xdf格式的檔案,可減少資料在硬碟與記憶體間的存取時間.

data.frame載入時即佔記憶體空間, 則Xdf載入時,僅記錄檔案位置,實際要做運算動作時才讀入.

File>chunk>Block>rows

一次讀一個chunk



設定RevoScaleR範例路徑

Revolution因應一些需求設計此函數,可透過此函數去找出範例檔或執行流程等.

```
sampleDataDir <- rxGetOption("sampleDataDir")
```

取得此資料夾下的AirlineDemoSmall.csv的路徑

```
inputFile<-file.path(sampleDataDir,"AirlineDemoSmall.csv")
```

資料匯入-rxImport

- rxImport(inData,outFile=null,...)
  - inData為欲輸入的檔案位置
  - outFile為輸出的xdf檔案位置
    - outFile若不輸入檔案位置, 則此函數會回傳一個data.frame
    - 若有寫位置,則會在該位置生成xdf檔, 回傳一個路徑, 不直接將檔案存在R的記憶體中, 分析時才讀取
    - 更多參數, 詳見附件網址.

資料匯入-rxImport

```
airDS<-rxImport(
            inData=inputFile,
            outFile="ADS.xdf",
            missingValueString="M",
            stringsAsFactors=T,
            overwrite=T
            )
```

資料匯入-RxXdfData

若xdf檔案已經存在, 可直接使用RxXdfData函數直接讀取Xdf檔案

RxXdfData()

或直接以csv等檔案做運算

RxTextData()

輔助資料

rxImport

http://www.rdocumentation.org/packages/RevoScaleR/functions/rxImport

RxXdfData

http://www.rdocumentation.org/packages/RevoScaleR/functions/RxXdfData

RxTextData

http://www.rdocumentation.org/packages/RevoScaleR/functions/RxTextData

實作時間

- http://packages.revolutionanalytics.com/datasets/mortDefault.zip

- 下載此資料，並匯入RRE，並且在匯入時轉存成xdf檔

分析流程

資料匯入

資料探索

資料整理

資料分析

結果視覺化

資料探索

過去

nrow(airDS)
ncol(airDS)
head(airDS)

資料探索

現在

rxGetInfo(airDS,getVarInfo=T,numRows=6)

得知變數資訊, 知道變數的類型

若為類別變數, 則知道有哪些類別

資料探索

調閱敘述性統計, 得知資料長相

```
adsSummary<-rxSummary(
        ~ArrDelay+CRSDepTime+DayOfWeek,
        data=airDS
        )
adsSummary
```

資料探索

```
rxSummary(~DayOfWeek,data=airDS)

rxSummary(~ArrDelay+DayOfWeek,data=airDS)

rxSummary(~ArrDelay:DayOfWeek,data=airDS)

rxSummary(ArrDelay~DayOfWeek,data=airDS)
```

透過rxSummary可以得知

- 變數的均值, 標準差, 最大值, 最小值
- 若變數為類別變數的話, 可知道每一個類別有多少筆資料

以視覺化方式了解資料長相

```
rxHistogram(~DayOfWeek, data=airDS)
```

```
rxHistogram(~CRSDepTime|DayOfWeek,data=airDS)
```

輔助資料

rxGetInfo

http://www.rdocumentation.org/packages/RevoScaleR/functions/rxGetInfoXdf

rxSummary

http://www.rdocumentation.org/packages/RevoScaleR/functions/rxSummary

rxHistogram

http://www.rdocumentation.org/packages/RevoScaleR/functions/rxHistogram

分析流程

資料匯入

資料探索

資料整理

資料分析

結果視覺化

資料整理(新增變數)

```
airExtraDS <- rxDataStep(
        inData = airDS, outFile="temp.xdf",
        transforms=list(
                Late = ArrDelay > 15,
                DepHour = as.integer(CRSDepTime),
                Night = DepHour >= 20 | DepHour <= 5,
                test = ArrDelay+10
                ),
        overwrite=T
        )

rxGetInfo(airExtraDS, getVarInfo=TRUE, numRows=5)
```

資料整理(刪除變數)

```
airExtraDS<-rxDataStep(
        inData= airExtraDS,
        outFile="ADS2.xdf",
        varsToDrop=c("test"),
        overwrite=T
        )

rxGetInfo(airExtraDS, getVarInfo=TRUE, numRows=5)
```

篩選資料

```
myData <- rxDataStep(
        inData = airDS,
        rowSelection = ArrDelay > 240 & ArrDelay <= 300,
        varsToKeep = c("ArrDelay", "DayOfWeek")
        )

rxHistogram(~ArrDelay, data = myData)
```

資料整理(篩選資料列)

從資料列最開始到最末列, 每十列抽一筆

```
myData<-rxDataStep(
            inData=airDS,
             rowSelection = (seq(from=.rxStartRow,length.out=.rxNumRows)%%10==0),
            )
```

資料切割

使用rxSplit

切割檔案做Train, Validation, Test

或切多份,每一份都輪流當Test

輔助資料

rxDataStep

http://www.rdocumentation.org/packages/RevoScaleR/functions/rxDataStep

rxSplit

http://www.rdocumentation.org/packages/RevoScaleR/functions/rxSplitXdf

分析流程

資料匯入

資料探索

資料整理

資料分析

結果視覺化

資料分析

目前RRE提供的HPA(High-Performance-Analysis)

Regression

Classify

Cluster

Regression

- rxLinMod
  - 線性迴歸
- rxLogit
  - LogisticRegression
- rxGlm
  - 廣義線性模型
- rxPredict
  - 預測

建立模型

```
logitObj <- rxLogit(
        Late~DepHour + Night,
        data = airExtraDS
        )

summary(logitObj)
```

預測

```
predictDS <- rxPredict(
        modelObject = logitObj,
        data = airExtraDS,
        outData = airExtraDS
        )
```

預測

rxPredict函數的參數modelObj目前僅接受rxLinMod, rxGlm,rxLogit三種函數回傳的Object

輔助資料

rxLinMod
http://www.rdocumentation.org/packages/RevoScaleR/functions/rxLinMod

rxGlm
http://www.rdocumentation.org/packages/RevoScaleR/functions/rxGLM

rxLogit
http://www.rdocumentation.org/packages/RevoScaleR/functions/rxLogit

rxPredict
http://www.rdocumentation.org/packages/RevoScaleR/functions/rxPredict

分析流程

資料匯入

資料探索

資料整理

資料分析

結果視覺化

視覺化

可用rx或開源R套件視覺化

用開源R套件視覺化，有些套件不支援xdf，須轉存成data.frame，要注意記憶體可能無法裝載。

視覺化

- http://rgm.ogalab.net/RGM/R_image_list?page=1761&init=true

ggplot2

rCharts

互動式網頁呈現Shiny

小結

- rxImport, RxTextData等函數設定資料來源
- rxDataStep(資料整理)
- 資料分析(rxGlm, rxLogit等)
- 結果視覺化

小結

我記不住那麼多function，怎麼辦？

打開RRE→打開script→點擊滑鼠右鍵
→選擇Insert Snippet

Revolution R Enterprise+Hadoop

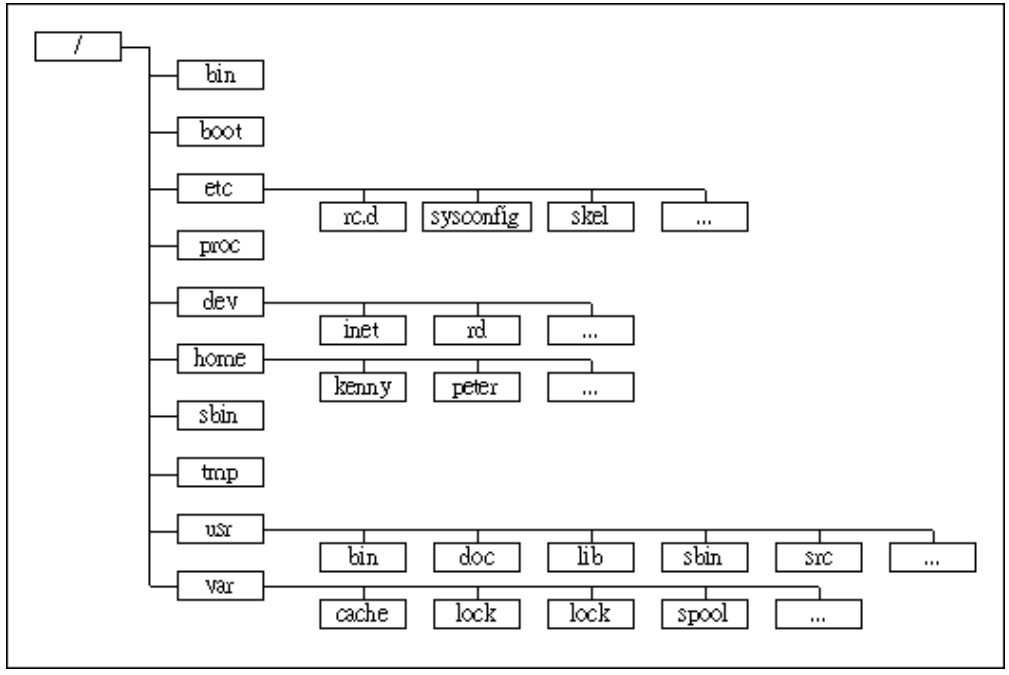
- 1.Linux
- 2.Hdfs
- 3.RRE 操控Hdfs的function
- 4.用RRE做InHadoop的運算

Linux檔案結構

/為根目錄

/home -> 根目錄下的home資料夾

/home/kenny ->根目錄下的home資料夾下的kenny資料夾



Linux

目前所在位置為/home，想要到kenny資料夾

絕對路徑

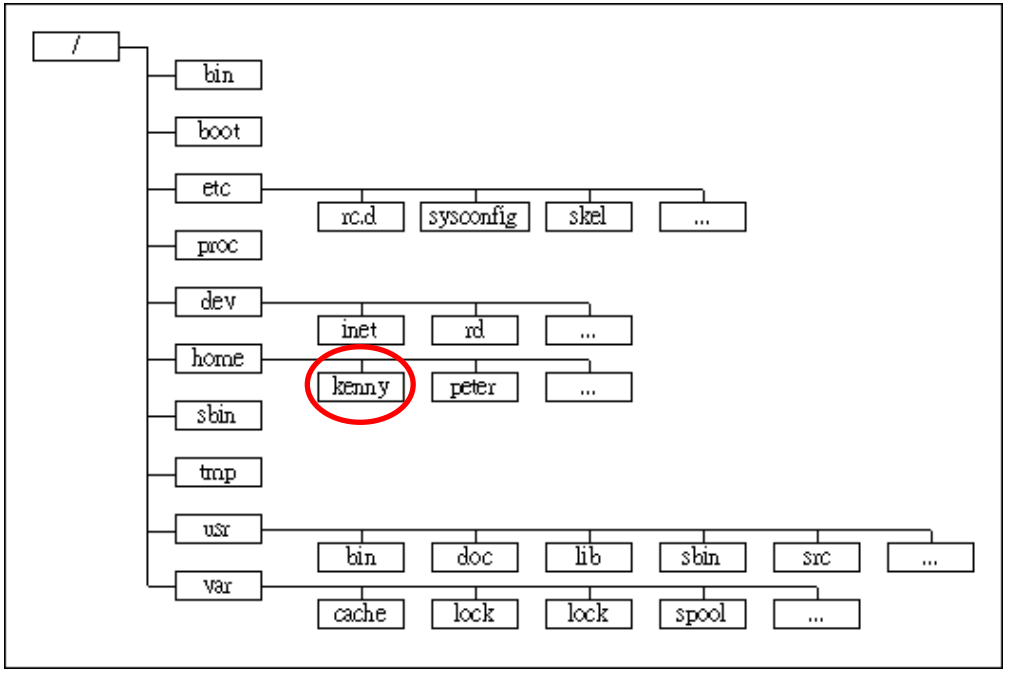
cd /home/kenny

相對路徑

cd ./kenny

. 意指當前資料夾

..意指從當前資料夾
往上跳一層



Linux基本指令

ls:檢視目錄與檔案

~$ls [欲查詢的位置,預設為當前目錄]

ls –l 調閱檔案詳細資訊

cd :切換目錄

~$cd /home/cloudera/Downloads

mkdir:建立資料夾

Downloads$mkdir ./LBH

Downloads$mkdir –p ./LBH/LBJ/Wade

Linux基本指令

先切換至 LBH資料夾　Downloads$cd LBH

touch:新增檔案

LBH$touch LBHvsWade

LBH$touch LBHvsLBJ

rm:刪除檔案

LBH$rm LBHvsWade

LBH$cd ..

Downloads$rm –rf ./LBH

rm –rf是強制刪除指定資料夾，包含子目錄與檔案(請小心使用)

Linux基本指令

linux 檔案有權限管理

owner group others

rwx rwx rwx

r:可讀 w:可寫 x:可執行

Linux基本指令

改變檔案屬性與權限三指令

chmod

chmod -R 777 dirname/filename

777 意指全部人可以讀寫執行

chgrp

chgrp groupname [-R] dirname/filename

chown

chown username [-R] dirname/filename

Linux基本練習

- 請回到使用者cloudera 資料夾下
- 建立test資料夾
- 並在test資料夾內建立兩個file, 名為file1 file2
- 刪除file1,
- 然後將test資料夾權限改為全部人皆可讀寫執行
- 刪除test資料夾
- 建立一個file, 名為test

- 建議每一個步驟做完,都用ls –l 觀察變化 更能加深印象

解答

- $cd /home/cloudera
- $mkdir test
- $touch ./test/file1 ./test/file2
- $rm ./test/file1
- $chmod 777 ./test
- $rm –rf ./test

參考網站與可學習資源

- http://linux.vbird.org/linux_basic/redhat6.1/linux_06command.php
- http://linux.vbird.org/linux_basic/0210filepermission.php

趁中午練習

- http://packages.revolutionanalytics.com/datasets/mortDefault.zip

- 下載此資料集後，練習把資料讀入RRE，觀察資料，對資料作整理，分析資料。

HDFS指令

- hadoop fs –ls 欲查詢位置

  $hadoop fs –ls /

  查詢HDFS的根目錄下有哪些目錄與檔案

- hadoop fs –put 將本地端檔案上傳至HDFS

  $hadoop fs –put 本地端檔案 hdfs位置

  $hadoop fs –put ./test /user/cloudera/

HDFS指令

- hadoop fs –get 從HDFS將檔案取回至本地端

  $hadoop fs –get /user/cloudera/test /home/cloudera/Downloads/

- hadoop fs –cat 欲觀看檔案的路徑

  $hadoop fs –cat /user/cloudera/test

  看根目錄下的user目錄下的cloudera目錄下的test檔案

HDFS指令

- hadoop fs –rm 刪除 HDFS上的檔案

  $hadoop fs –rm /user/cloudera/test

- hadoop fs –mkdir 建立目錄

  $hadoop fs –mkdir /user/cloudera/testDir

  建立一個在根目錄下的user目錄下的cloudera目錄下的目錄,名為testDir

HDFS指令

- hadoop fs –chmod 欲更改的檔案權限

  hadoop fs –chmod 777 /user/cloudera/testDir

  網頁介面 檢視hdfs

  localhost:50070

  選擇右上角Utilities ,再選擇Browse the file system

更多HDFS

- http://www.ewdna.com/2013/04/Hadoop-HDFS-Comics.html
- http://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-common/FileSystemShell.html
- http://my.oschina.net/crxy/blog/348868
- http://my.oschina.net/u/2009649/blog/351694

RevolutionR+Hadoop

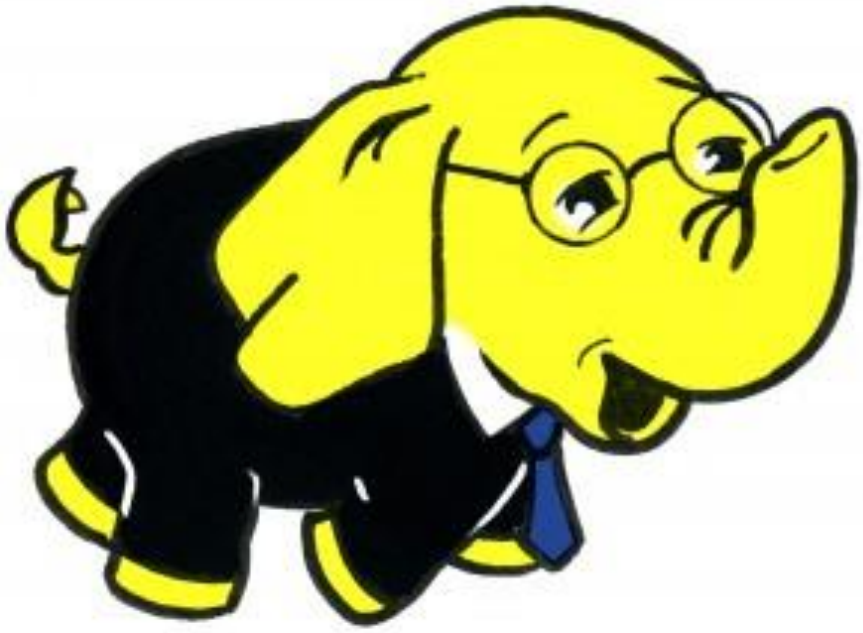


Illustration by Michael Capozzola

RevolutionR+Hadoop

- rxHadoopCommand

  可直接輸入指令操控hadoop, 函數內放要執行的語句 不用加hadoop

  $hadoop fs –ls /user/cloudera/testDir

  rxHadoopCommand("fs –ls /user/cloudera/testDir")

- rxHadoopCopyFromLocal
- rxHadoopCopyToLocal
- rxHadoopListFiles
- rxHadoopRemove
- rxHadoopCopy
- rxHadoopMove
- rxHadoopMakeDir
- rxHadoopRemoveDir

RRE In Hadoop

- 設定ComputeContext
- 設定檔案位置及類型(資料匯入)
- 資料探索
- 資料整理
- 資料分析
- 輸出結果

設定ComputeContext

	RxLocalSeq	RxHadoopMR
DelimitedText (RxTextData)	x	x
Fixed-FormatText (RxTextData)	x	
.xdf data files (RxXdfData)	x	x
SAS data files (RxSasData)	x	
SPSS data files (RxSpssData)	x	
ODBC data (RxOdbcData)	x	
Teradata database (RxTeradata)	x	

設定ComputeContext

- RRE預設使用RxLocalSeq,若要使用Hadoop去計算, 則須建立一個RxHadoopMR.

- 建立RxHaoopMR,有三種方式,此處示範兩種.

在Hadoop Cluster Node上執行RRE

- executeOnNode<-RxHadoopMR()
- rxSetComputeContext(executeOnNode)

遠端電腦連線至Node

要設定成免密碼連線

要先把你家的鎖拿過去給他家,請他用你的鎖做個你能進的門

用自己的key開他為你設的門, 進去他家

現場跟我一起做。

遠端電腦連至Node

```
# 欲連線的node之使用者帳號
mySshUsername<-"cloudera"

# 存在putty內的session名稱
mySshHostname<-"CDH4.7Cluster"
```

遠端電腦連至Node

```
# 在node本地端那邊供RRE存取的資料夾路徑，請勿更動此行
myShareDir <- paste("/var/RevoShare",mySshUsername,sep="/")
```

```
# 在hdfs供RRE存取的資料夾路徑，請勿更動此行
myHdfsShareDir <-paste("/user/RevoShare",mySshUsername,sep="/")
```

遠端電腦連至Node

```
# sshClientDir 請設定你putty的儲存路徑

myHadoopCluster <- RxHadoopMR(
        hdfsShareDir=myShareDir,
        shareDir=myShareDir,
        sshUsername=mySshUsername,
        sshHostname=mySshHostname,
        sshClientDir="D:\\putty"
        )
```

遠端電腦連線至Node

```
rxSetComputeContext(myHadoopCluster)
```

匯入資料

用RxHadoopMR為ComputeContext時

謹記只能接收兩種資料源

RxTextData

RxXdfData

匯入資料

單機所使用的xdf檔與Hdfs上的xdf檔不相同, 請勿直接把單機使用的xdf丟至hdfs上

請把原始檔案放在hdfs上, 再調整ComputeContext為RxHadoopMR(), 再進行rxImport()

匯入資料

在HDFS上的Xdf會以一個大目錄存在,裡面包含兩個小資料夾,

一個資料夾放副檔名為xdfm的檔案 ,
contains the metadata for all of the .xdfd files

一個資料夾放副檔名為xdfd的檔案
Data is split into individual ‘.xdfd’ files such that each file remains within a single HDFS block

後續流程

請參閱HowToUseRxForHadoop.R

學習資料

RRE線上教學與實作

https://www.datacamp.com/courses/introduction-to-revolution-r-enterprise-for-big-data-analytics

RevoScaleR包所擁有的函數

http://www.rdocumentation.org/packages/RevoScaleR