# Thoughts on Linux & Statistical Computing

...

Ryan Murphy
Spring 2016

# Why I'm giving this talk

1. Several experiences pushed me inevitably towards UNIX, serves as good anecdotal evidence
2. Recent switcher

## Experience pushing me towards Linux

1. Windows installations can be a huge pain in the rear

# Windows Intall: Pain in the Rear

- FreeSurfer and LaTeX
- R to C communication
- Spark

# Spark Windows Installation

stackoverflow

I found the easiest solution on Windows is to build from source.

You can pretty much follow this guide: http://spark.apache.org/docs/latest/building-spark.html

http://spark.apache.org/docs/latest/building-spark.html

You'll get and error for winutils.exe:

```
15/04/15 12:33:13 INFO MemoryStore: MemoryStore started with capacity 267.3 MB
15/04/15 12:33:20 WARN NativeCodeLoader: Unable to load native-hadoop library fo
r your platform... using builtin-java classes where applicable
15/04/15 12:33:21 ERROR Shell: Failed to locate the winutils binary in the hadoo
p binary path
java.io.IOException: Could not locate executable null\bin\winutils.exe in the Ha
doop binaries.
        at org.apache.hadoop.util.Shell.getQualifiedBinPath(Shell.java:318)
        at org.apache.hadoop.util.Shell.getWinUtilsPath(Shell.java:333)
        at org.apache.hadoop.util.Shell.<clinit>(Shell.java:326)
```

# Building Spark



Building Spark using Maven requires Maven 3.3.3 or newer and Java 7+. The Spark build can supply a suitable Maven binary; see below.

## Building with `build/mvn`

Spark now comes packaged with a self-contained Maven installation to ease building and deployment of Spark from source located under the

# Windows installation

```
>spark-shell --packages com.databricks:spark-csv_2.11:1.3.0
```


No, You Can't Has
Cheezburger
NOT YOURS

```
15/12/01 20:06:35 WARN General: Plugin (Bundle) "org.datanucleus.store.rdbms" is
 already registered. Ensure you dont have multiple JAR versions of the same plug
in in the classpath. The URL "file:/C:/spark/spark-1.5.2-bin-hadoop2.4/lib/datan
ucleus-rdbms-3.2.9.jar" is already registered, and you are trying to register an
 identical plugin located at URL "file:/C:/spark/spark-1.5.2-bin-hadoop2.4/bin/.
./lib/datanucleus-rdbms-3.2.9.jar."
15/12/01 20:06:35 WARN General: Plugin (Bundle) "org.datanucleus.api.jdo" is alr
eady registered. Ensure you dont have multiple JAR versions of the same plugin i
n the classpath. The URL "file:/C:/spark/spark-1.5.2-bin-hadoop2.4/bin/../lib/da
tanucleus-api-jdo-3.2.6.jar" is already registered, and you are trying to regist
er an identical plugin located at URL "file:/C:/spark/spark-1.5.2-bin-hadoop2.4/
lib/datanucleus-api-jdo-3.2.6.jar."
15/12/01 20:06:35 WARN General: Plugin (Bundle) "org.datanucleus" is already reg
istered. Ensure you dont have multiple JAR versions of the same plugin in the cl
asspath. The URL "file:/C:/spark/spark-1.5.2-bin-hadoop2.4/lib/datanucleus-core-
3.2.10.jar" is already registered, and you are trying to register an identical p
lugin located at URL "file:/C:/spark/spark-1.5.2-bin-hadoop2.4/bin/../lib/datanu
cleus-core-3.2.10.jar."
15/12/01 20:06:36 WARN Connection: BoneCP specified but not present in CLASSPATH
 (or one of dependencies)
15/12/01 20:06:36 WARN Connection: BoneCP specified but not present in CLASSPATH
 (or one of dependencies)
15/12/01 20:06:45 WARN ObjectStore: Version information not found in metastore.
hive.metastore.schema.verification is not enabled so recording the schema versio
n 1.2.0
15/12/01 20:06:45 WARN ObjectStore: Failed to get database default, returning No
SuchObjectException
15/12/01 20:06:47 WARN : Your hostname, RyanM-PC resolves to a loopback/non-reac
hable address: fe80:0:0:0:0:5efe:ac15:e752%net12, but we couldn't find any exter
nal IP address!
java.lang.RuntimeException: java.lang.RuntimeException: The root scratch dir: /t
mp/hive on HDFS should be writable. Current permissions are: rw-rw-rw-
        at org.apache.hadoop.hive.ql.session.SessionState.start(SessionState.jav
a:522)
        at org.apache.spark.sql.hive.client.ClientWrapper.<init>(ClientWrapper.s
cala:171)
```

There's some exception, but where in that mess does it originate?

# Contacted customer support December 1st...

From Purdue Customer Support <Central-Incident@purdue.ed

Subject **Thank you for contacting us about "SQL Context Error" ISSUE=591536 PROJ=17** 12/01/2015 09:47 PM

To murph213@purdue0.onmicrosoft.c

Tags Important

Here is my command:

```
spark-shell --packages com.databricks:spark-
csv_2.11:1.3.0
```

# Still trying to find a solution in January!

Subject **Customer Notice about SQL Context Error ISSUE=591536 PROJ=17** 01/11/2016 09:40 A

To murph213@purdue0.onmicrosoft.c

- Use the winutil.exe here: https://github.com/steveloughran/winutils /blob/master/hadoop-2.6.0/bin/winutils.exe, I am not sure if this can be the issue, but the size of winutil.exe from the above link is different from the one that is currently on your system.
- Put the winutil.exe in the same bin directory as other hadoop executables (i.e. hadoop, mapred, hdfs, yarn etc), then set the HADOOP_HOME to be that directory above bin. Hadoop needs to be able to find all its executables using HADOOP_HOME/bin, in your setup, the HADOOP_HOME/bin only contains the winutils.exe.
- Make sure to stop and start the daemons after you make any changes to the system.

If the error persists, I would suggest you to come to our coffee consultant to chat with your laptop. I will be at LavAzza at 2PM on Tuesday (https://www.rcac.purdue.edu/coffee/)

# Linux – did it last night

http://blog.prabeeshk.com/blog/2014/10/31/install-apache-spark-on-ubuntu-14-dot-04/

```
~ $  wget http://www.scala-lang.org/files/archive/scala-2.10.4.tgz
--2016-02-25 00:02:28--  http://www.scala-lang.org/files/archive/scala-2.10.4.
tgz
Resolving www.scala-lang.org (www.scala-lang.org)... 128.178.154.159
Connecting to www.scala-lang.org (www.scala-lang.org)|128.178.154.159|:80... c
onnected.
HTTP request sent, awaiting response... 200 OK
Length: 29937534 (29M) [application/x-gzip]
Saving to: 'scala-2.10.4.tgz'

 7% [=>                                   ] 2,114,633    719KB/s              ^
C
~ $
~ $
~ $
~ $  wget http://www.scala-lang.org/files/archive/scala-2.10.4.tgz
--2016-02-25 00:03:30--  http://www.scala-lang.org/files/archive/scala-2.10.4.
tgz
Resolving www.scala-lang.org (www.scala-lang.org)... 128.178.154.159
Connecting to www.scala-lang.org (www.scala-lang.org)|128.178.154.159|:80... c
onnected.
HTTP request sent, awaiting response... 200 OK
Length: 29937534 (29M) [application/x-gzip]
Saving to: 'scala-2.10.4.tgz.1'

100%[===================================>] 29,937,534  6.20MB/s   in 8.2s

2016-02-25 00:03:39 (3.48 MB/s) - 'scala-2.10.4.tgz.1' saved [29937534/2993753
4]
```

```
ows installed java version

a version "1.7.0_72"_ Java(TM) SE Runtime Environment (build 1.7.0_72-
)_ Java HotSpot(TM) 64-Bit Server VM (build 24.72-b04, mixed mode)
```

ext step is install Scala, follow the following instructions to set up Scala.
t download the Scala from here

y downloaded file to some location for example */urs/local/src*, untar the file
set path variable,

```
$ wget http://www.scala-lang.org/files/archive/scala-2.10.4.tgz
$ sudo mkdir /usr/local/src/scala
$ sudo tar xvf scala-2.10.4.tgz -C /usr/local/src/scala/
```

```
$ vi .bashrc
```

add following in the end of the file

# Already working, for example, here's spark:

```
bin $ ./spark-shell --packages com.databricks:spark-csv_2.11:1.3.0
Ivy Default Cache set to: /home/murph213/.ivy2/cache
The jars for the packages stored in: /home/murph213/.ivy2/jars
:: loading settings :: url = jar:file:/usr/local/src/spark-1.6.0/assembl
y/target/scala-2.10/spark-assembly-1.6.0-hadoop2.2.0.jar!/org/apache/ivy
/core/settings/ivysettings.xml
com.databricks#spark-csv_2.11 added as a dependency
:: resolving dependencies :: org.apache.spark#spark-submit-parent;1.0
        confs: [default]
        found com.databricks#spark-csv_2.11;1.3.0 in central
        found org.apache.commons#commons-csv;1.1 in central
        found com.univocity#univocity-parsers;1.5.1 in central
downloading https://repo1.maven.org/maven2/com/databricks/spark-csv_2.11/1.3.0/s
park-csv_2.11-1.3.0.jar ...
        [SUCCESSFUL ] com.databricks#spark-csv_2.11;1.3.0!spark-csv_2.11.jar (56
ms)
downloading https://repo1.maven.org/maven2/org/apache/commons/commons-csv/1.1/co
mmons-csv-1.1.jar ...
        [SUCCESSFUL ] org.apache.commons#commons-csv;1.1!commons-csv.jar (32ms)
```

# Spark is working

```
scala> val input = sc.parallelize(List(1, 2, 3, 4))
input: org.apache.spark.rdd.RDD[Int] = ParallelCollectionRDD[0] at parallelize a

scala> val result = input.map(x => x * x)
result: org.apache.spark.rdd.RDD[Int] = MapPartitionsRDD[1] at map at <console>:

scala> result.collect()
res2: Array[Int] = Array(1, 4, 9, 16)

scala>
```

# Installation

- Really just a sequence of
  - wget
  - apt-get
  - update bash-rc

## Experience pushing me towards Linux

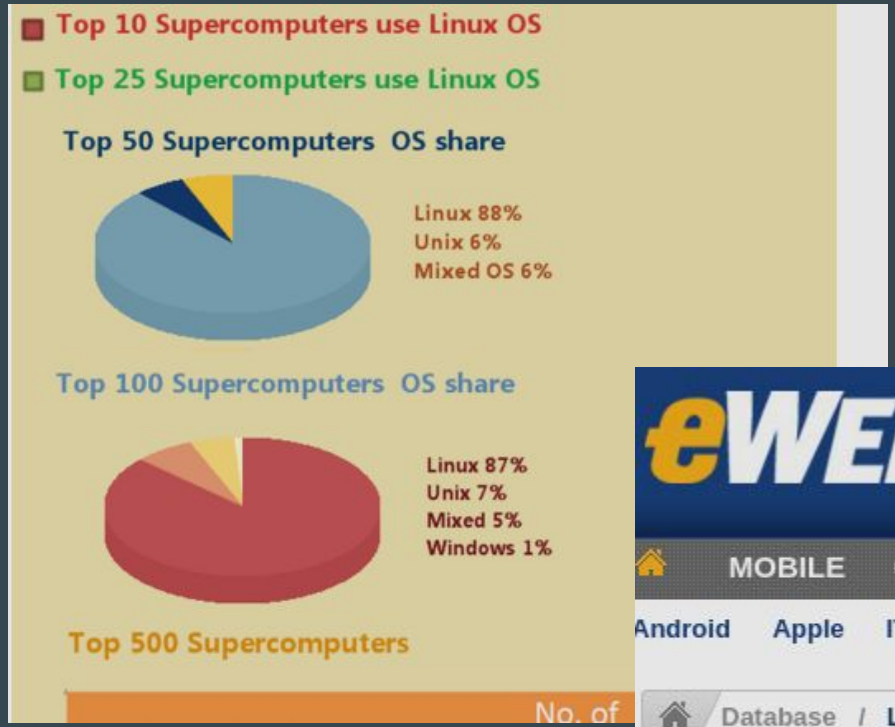
2. Prevalence of UNIX computing systems

# Prevalence of UNIX computing

- Purdue CS and STAT departments
- Impression of growing dominance

BOSTON -- Linux is now in the mainstream of enterprise adoption, according to analysts presenting new research here at the LinuxCon conference.

# Prevalence of Linux




# Making your life easier

…

# The makefile

```
all: gethttp git-commit

gethttp: gethttp.cpp openhttp.cpp SimpleHTMLParser.cpp
        g++ -o gethttp -g gethttp.cpp openhttp.cpp SimpleHTMLParser.cpp

git-commit:
        git add Makefile *.h *.cpp  >> .local.git.out
        git commit -a -m "Commit web-crawler" >> .local.git.out

clean:
        rm -f *.o gethttp core
```

# Regular Expressions

- Incredibly general way to search through text
- Real world example: finding one invalid record in a 100,000 line datafile
- Toy example ($FILES/1_Purdue/Seminars/regex/)

```
> dat <- read.table("datafile.dat")
> mean(dat$V1)
[1] NA
Warning message:
In mean.default(dat$V1) : argument is not numeric
NA
>
```

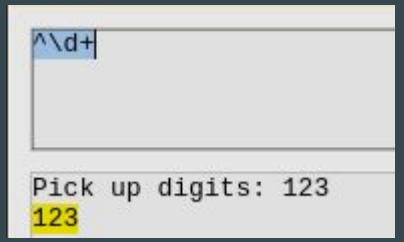
# Regular Expressions

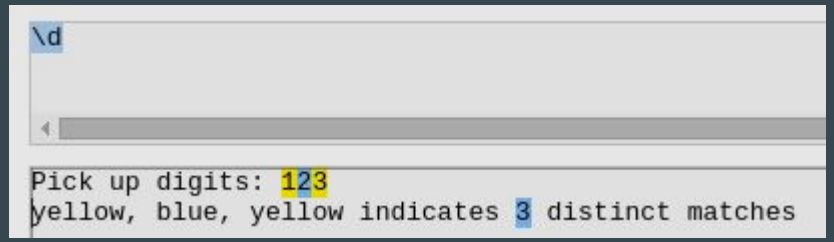
```
regex $ grep -Pn "^\D" datafile.dat
4:~ 5 0
```

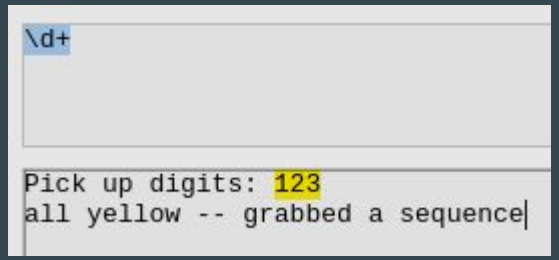
- P -- allows use of Patterns like \D
- n -- shows the line number
- grep = “global regular expression print”

# More on regular expressions

- Syntax: Expression -- Meaning
- \d -- Digits
- \w -- word characters
- \D -- non digit
- \W -- non word
- + -- Place after a symbol to grab a sequence (use in program to store text as a variable -- write your own web scrape?)
- ^ -- Look only at the start



# More on text files

- Big files can't be opened in notepad
  - (okay, okay, it's more about memory...)
- UNIX lets us view and manipulate text files without opening them
- File size - du
- File top and bottom - head or tail


```
-bash-4.1$ du -h  diabetic_data_cleaned.csv
12M     diabetic_data_cleaned.csv
```

# Piping

- Allows you to send the output of one command into another
- From a set of building blocks, your command line can perform a broad array of tasks.

```
regex $ cat datafile.dat
9 3 1
1 2 3
5 5 2
9 4 7
~ 5 0
regex $ sort datafile.dat | head -1
1 2 3
regex $ sort -d  datafile.dat | head -1
1 2 3
```

# Access to a UNIX system

- You too can have the power

- SSH isn't good enough
- Benefits of native support ("just push, no surprises")
- Macs
  - Expensive
  - Excessive hardware control
  - Deep level of customization
  - Linux makes you a first-class citizen in open source development

# Installation and Demo

- Linux install fest Monday, February 29th

1. Backup
2. Partition
3. Flash USB
4. Turn off “fast boot” and modify EFI settings
5. The Linux installer will set itself up in the open partition

# Thoughts

Pros

- Amazing for development
- Liberating
- Customizations
- Often like working with typical (proprietary) OS
- Computationally Efficient

Cons

- Learning Curve
- Office softwares are bad -- ie Word Processing just doesn’t compare
- Big troubles with Cisco Anyconnect