Firefox File Edit View History Bookmarks Tools Window Help Fri 1:12 PM

http://flinkf...ok/2BUU5KCBY

flinkforward-16-zeppelin.mybluemix.net/#/notebook/2BUU5KCBY

Search

Zeppelin Notebook Job Search your Notebooks anonymous

/ff2016/Intro Head default

FINISHED

Presentation Starting Soon...

Took 0 sec. Last updated by anonymous at September 12 2016, 6:56:13 AM. (outdated)

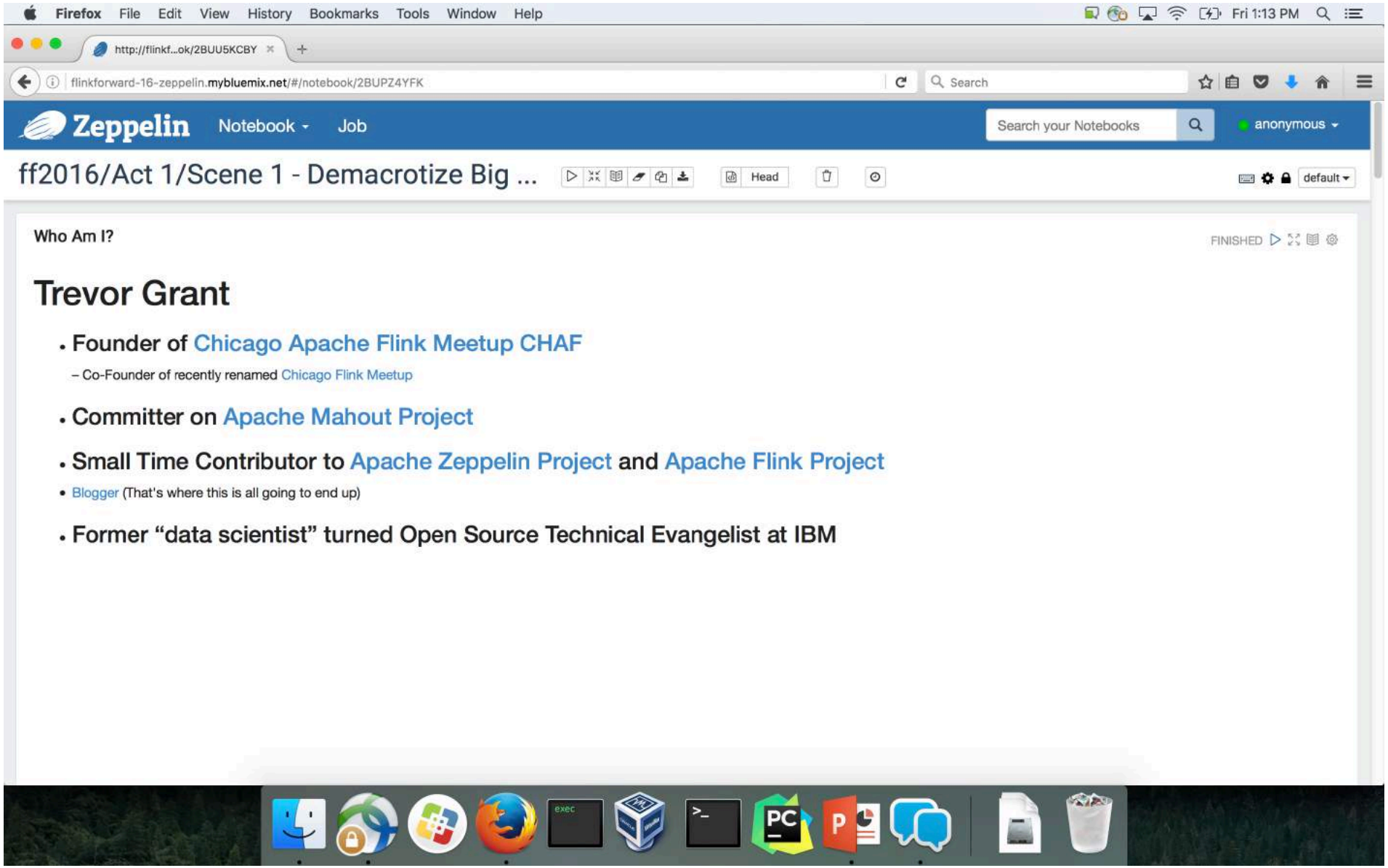
FINISHED

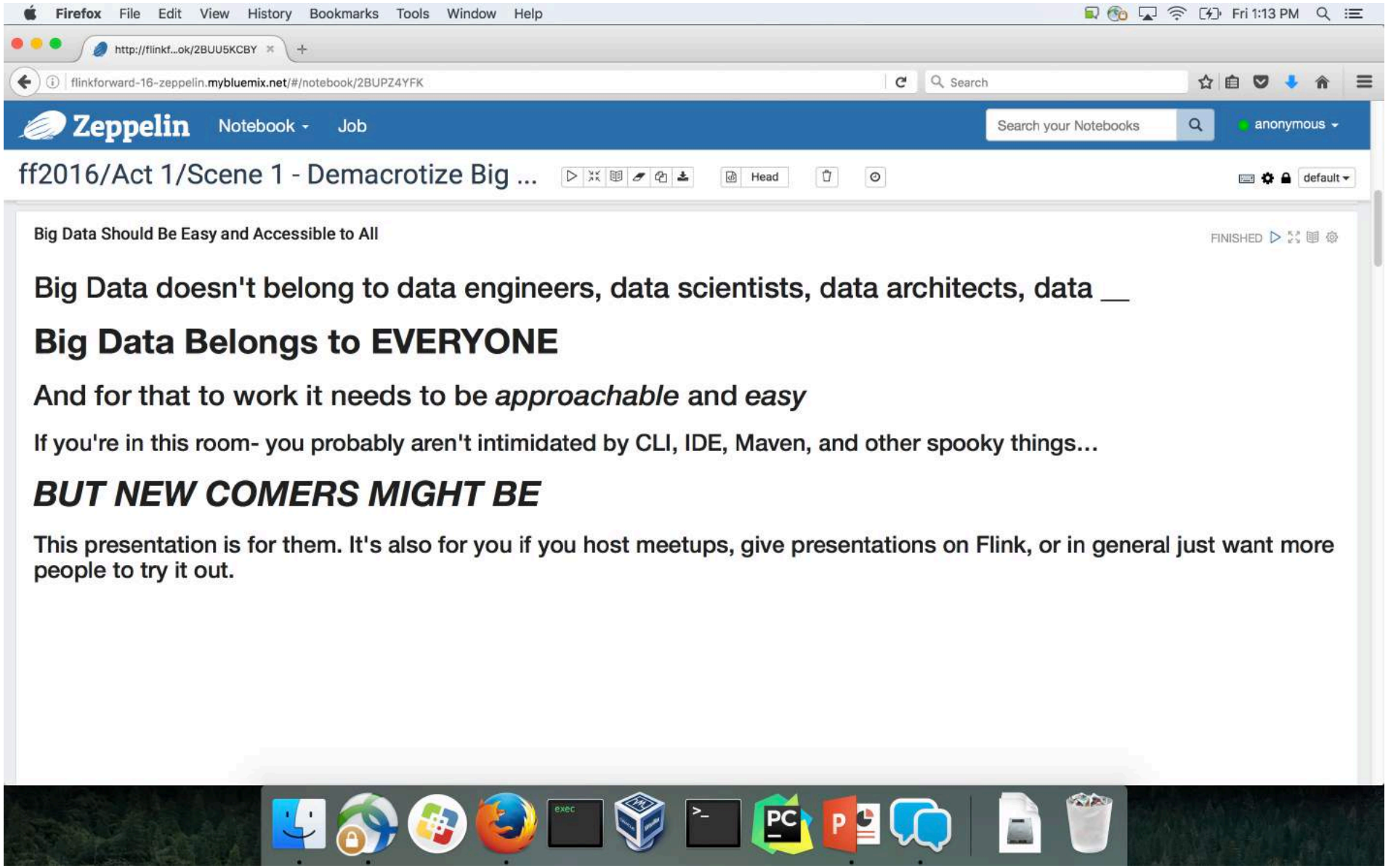
All Notebooks available on github at https://github.com/chi-apache-flink-meetup/flink_forward_2016

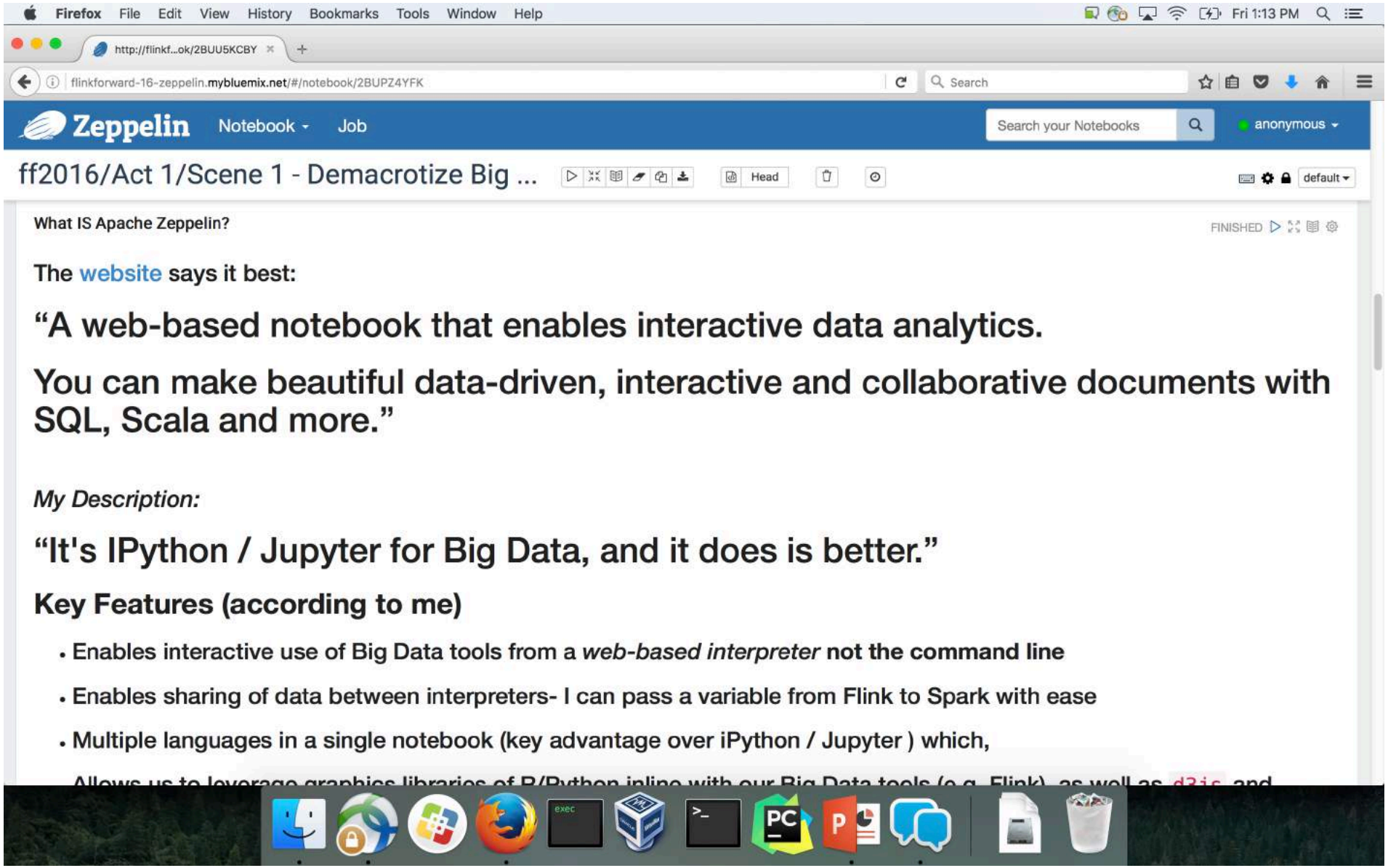
Took 0 sec. Last updated by anonymous at September 23 2016, 11:59:32 AM. (outdated)

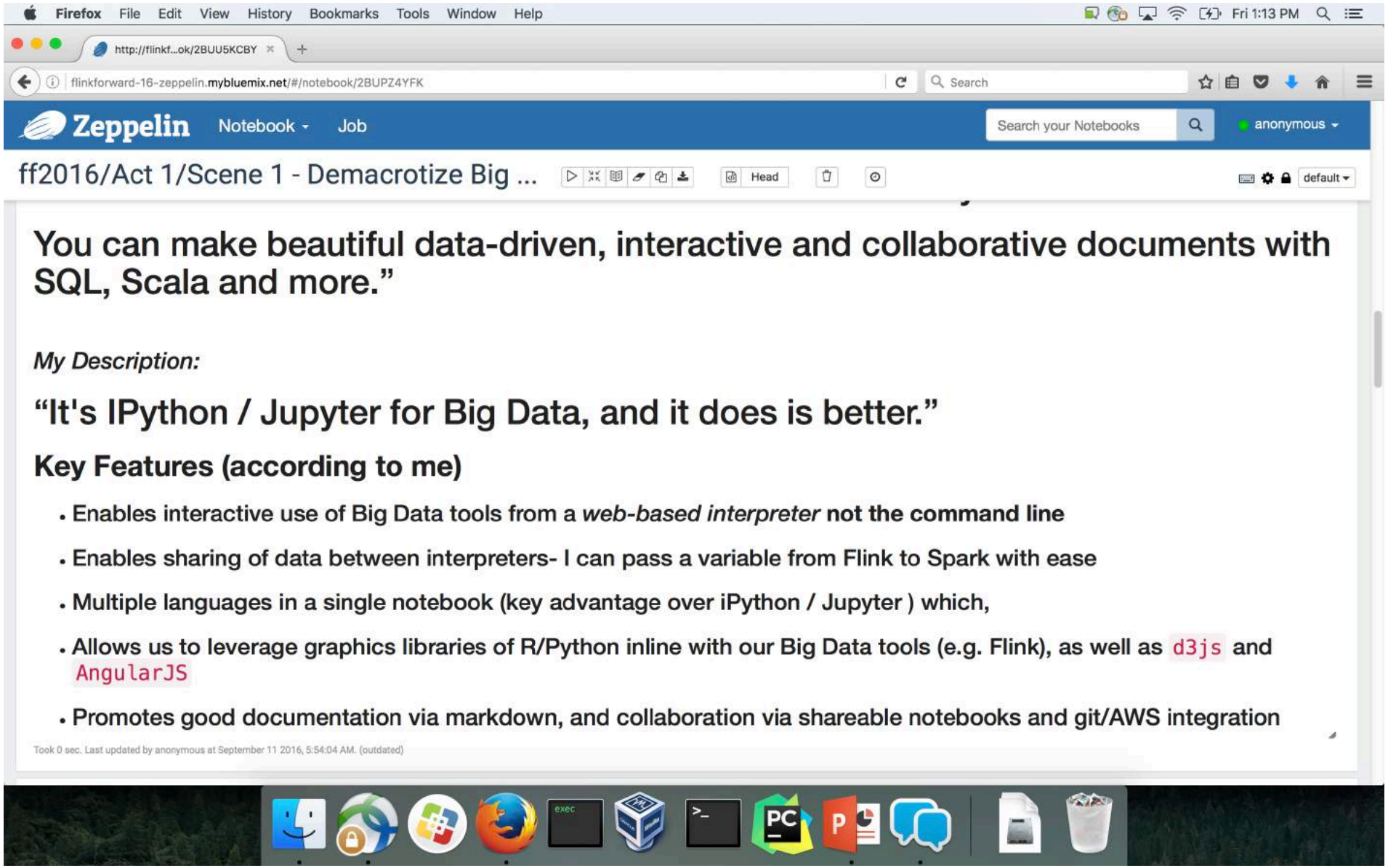
```
%md
```

READY


Firefox File Edit View History Bookmarks Tools Window Help
Fri 1:13 PM
http://flinkf...ok/2BUU5KCBY
flinkforward-16-zeppelin.mybluemix.net/#/notebook/2BUPZ4YFK
Search
Zeppelin
Notebook
Job
Search your Notebooks
anonymous
ff2016/Act 1/Scene 1 - Demacrotize Big ...
Head
default
Who Am I?
FINISHED
Trevor Grant
Founder of Chicago Apache Flink Meetup CHAF
– Co-Founder of recently renamed Chicago Flink Meetup
Committer on Apache Mahout Project
Small Time Contributor to Apache Zeppelin Project and Apache Flink Project
Blogger (That's where this is all going to end up)
Former “data scientist” turned Open Source Technical Evangelist at IBM


Firefox File Edit View History Bookmarks Tools Window Help
Fri 1:13 PM
http://flinkf...ok/2BUU5KCBY
flinkforward-16-zeppelin.mybluemix.net/#/notebook/2BUPZ4YFK
Search
Zeppelin
Notebook
Job
Search your Notebooks
anonymous
ff2016/Act 1/Scene 1 - Demacrotize Big ...
Head
default
Big Data Should Be Easy and Accessible to All
FINISHED
Big Data doesn't belong to data engineers, data scientists, data architects, data __
Big Data Belongs to EVERYONE
And for that to work it needs to be approachable and easy
If you're in this room- you probably aren't intimidated by CLI, IDE, Maven, and other spooky things...
BUT NEW COMERS MIGHT BE
This presentation is for them. It's also for you if you host meetups, give presentations on Flink, or in general just want more people to try it out.


Firefox File Edit View History Bookmarks Tools Window Help
Fri 1:13 PM
http://flinkf...ok/2BUU5KCBY
flinkforward-16-zeppelin.mybluemix.net/#/notebook/2BUPZ4YFK
Search
Zeppelin
Notebook
Job
Search your Notebooks
anonymous
ff2016/Act 1/Scene 1 - Demacrotize Big ...
Head
default
What IS Apache Zeppelin?
FINISHED
The website says it best:
“A web-based notebook that enables interactive data analytics.
You can make beautiful data-driven, interactive and collaborative documents with SQL, Scala and more.”
My Description:
“It's IPython / Jupyter for Big Data, and it does is better.”
Key Features (according to me)
• Enables interactive use of Big Data tools from a web-based interpreter not the command line
• Enables sharing of data between interpreters- I can pass a variable from Flink to Spark with ease
• Multiple languages in a single notebook (key advantage over iPython / Jupyter) which,


Firefox File Edit View History Bookmarks Tools Window Help
Fri 1:13 PM
http://flinkf...ok/2BUU5KCBY
flinkforward-16-zeppelin.mybluemix.net/#/notebook/2BUPZ4YFK
Search
Zeppelin
Notebook
Job
Search your Notebooks
anonymous
ff2016/Act 1/Scene 1 - Demacrotize Big ...
Head
default
You can make beautiful data-driven, interactive and collaborative documents with SQL, Scala and more."
My Description:
"It's IPython / Jupyter for Big Data, and it does is better."
Key Features (according to me)
• Enables interactive use of Big Data tools from a web-based interpreter not the command line
• Enables sharing of data between interpreters- I can pass a variable from Flink to Spark with ease
• Multiple languages in a single notebook (key advantage over iPython / Jupyter) which,
• Allows us to leverage graphics libraries of R/Python inline with our Big Data tools (e.g. Flink), as well as d3js and AngularJS
• Promotes good documentation via markdown, and collaboration via shareable notebooks and git/AWS integration
Took 0 sec. Last updated by anonymous at September 11 2016, 5:54:04 AM. (outdated)

Firefox File Edit View History Bookmarks Tools Window Help Fri 1:13 PM

http://flinkf...ok/2BUU5KCBY

flinkforward-16-zeppelin.mybluemix.net/#/notebook/2BUPZ4YFK

Search

Zeppelin Notebook Job Search your Notebooks anonymous

ff2016/Act 1/Scene 1 - Demacrotize Big ... Head default

Get Zeppelin

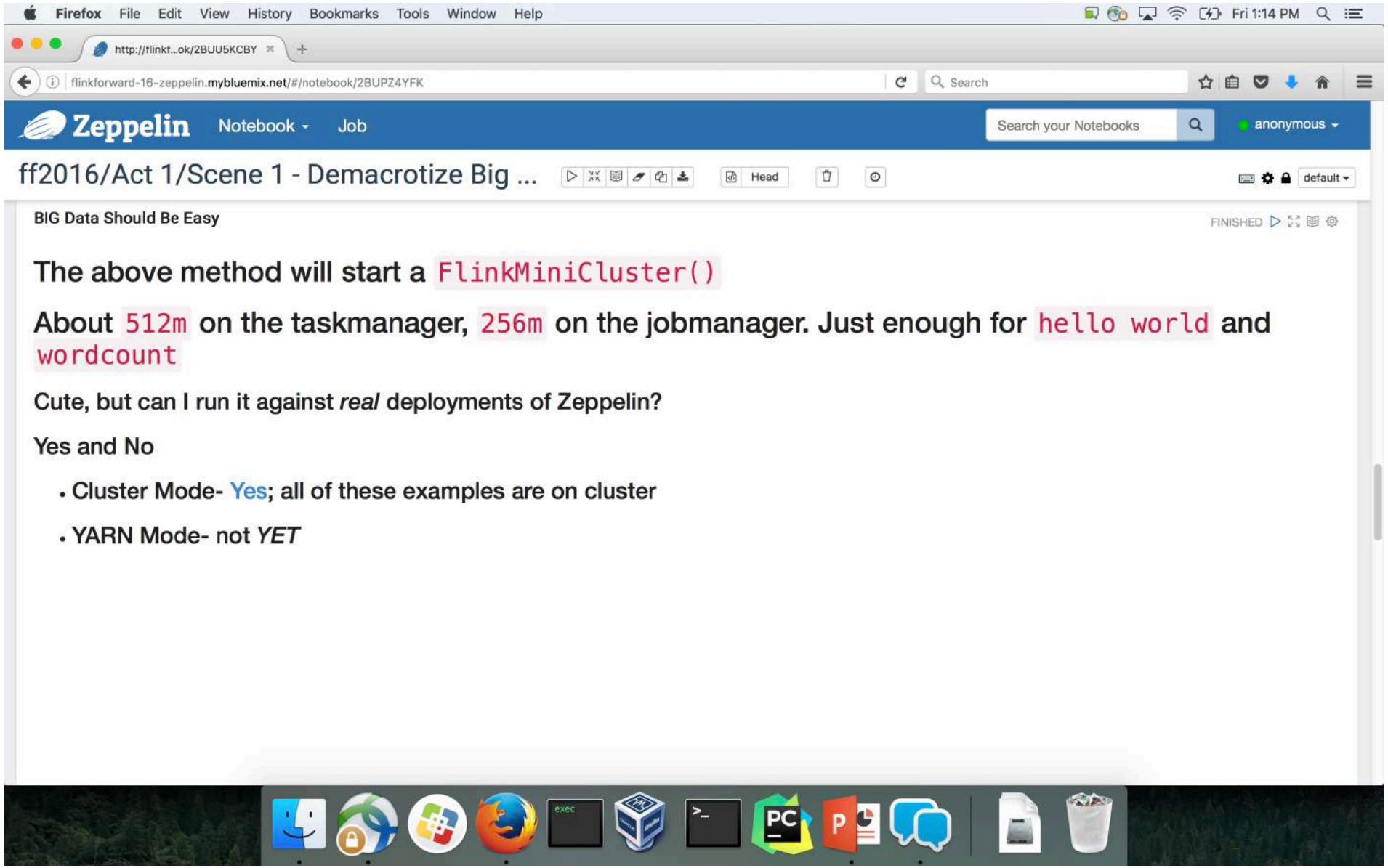
FINISHED

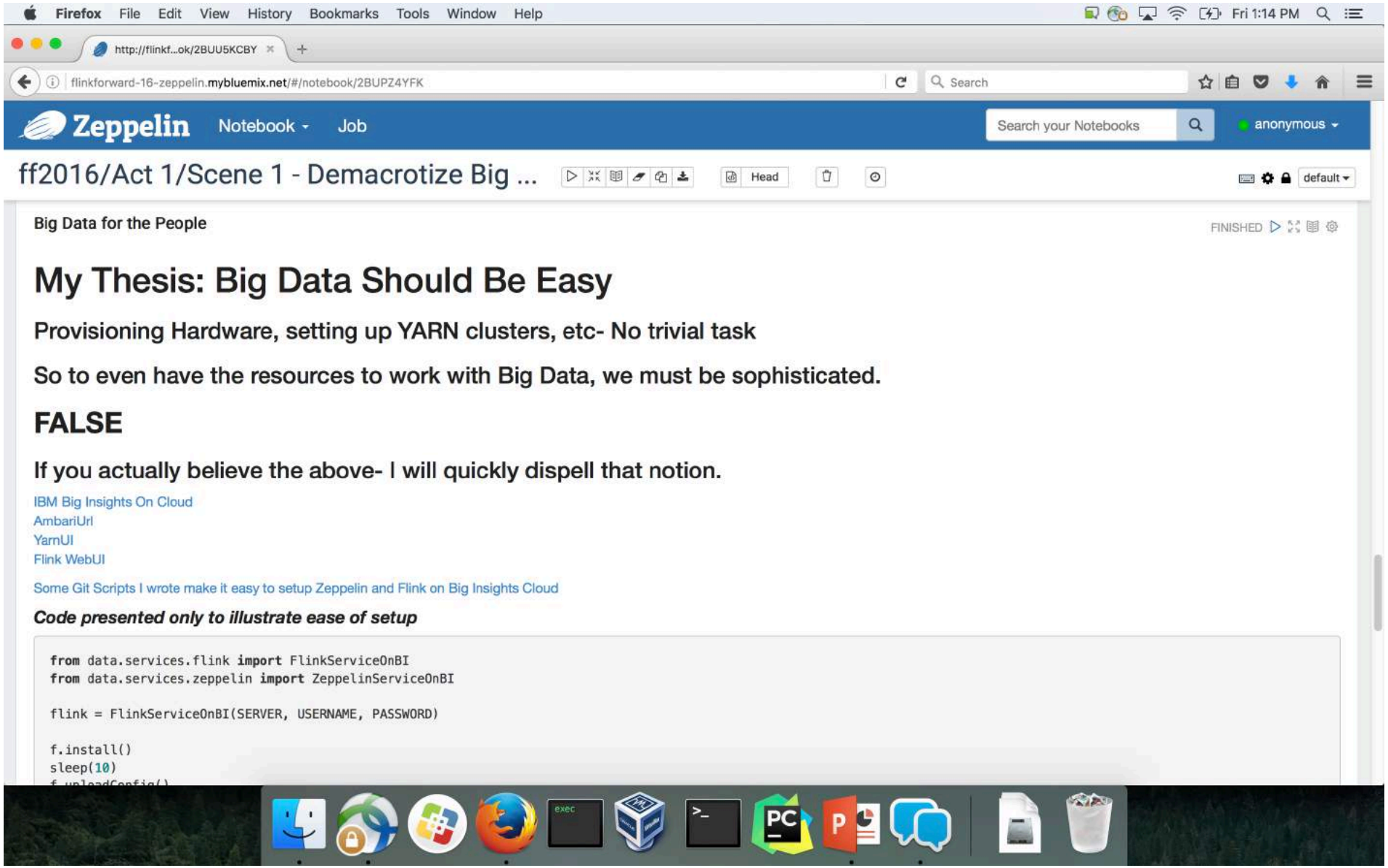
Zeppelin comes with a Flink!

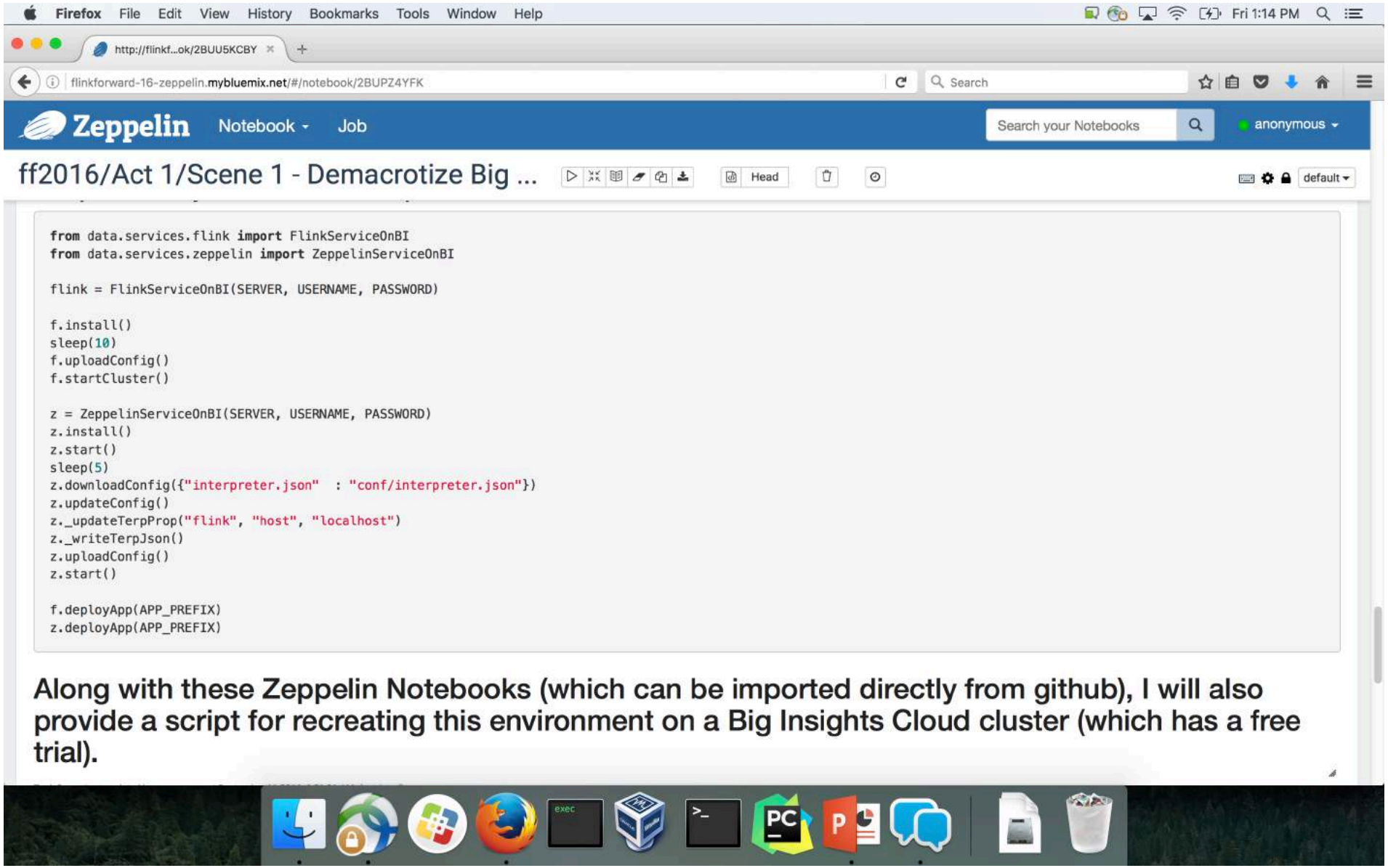
Install Zeppelin

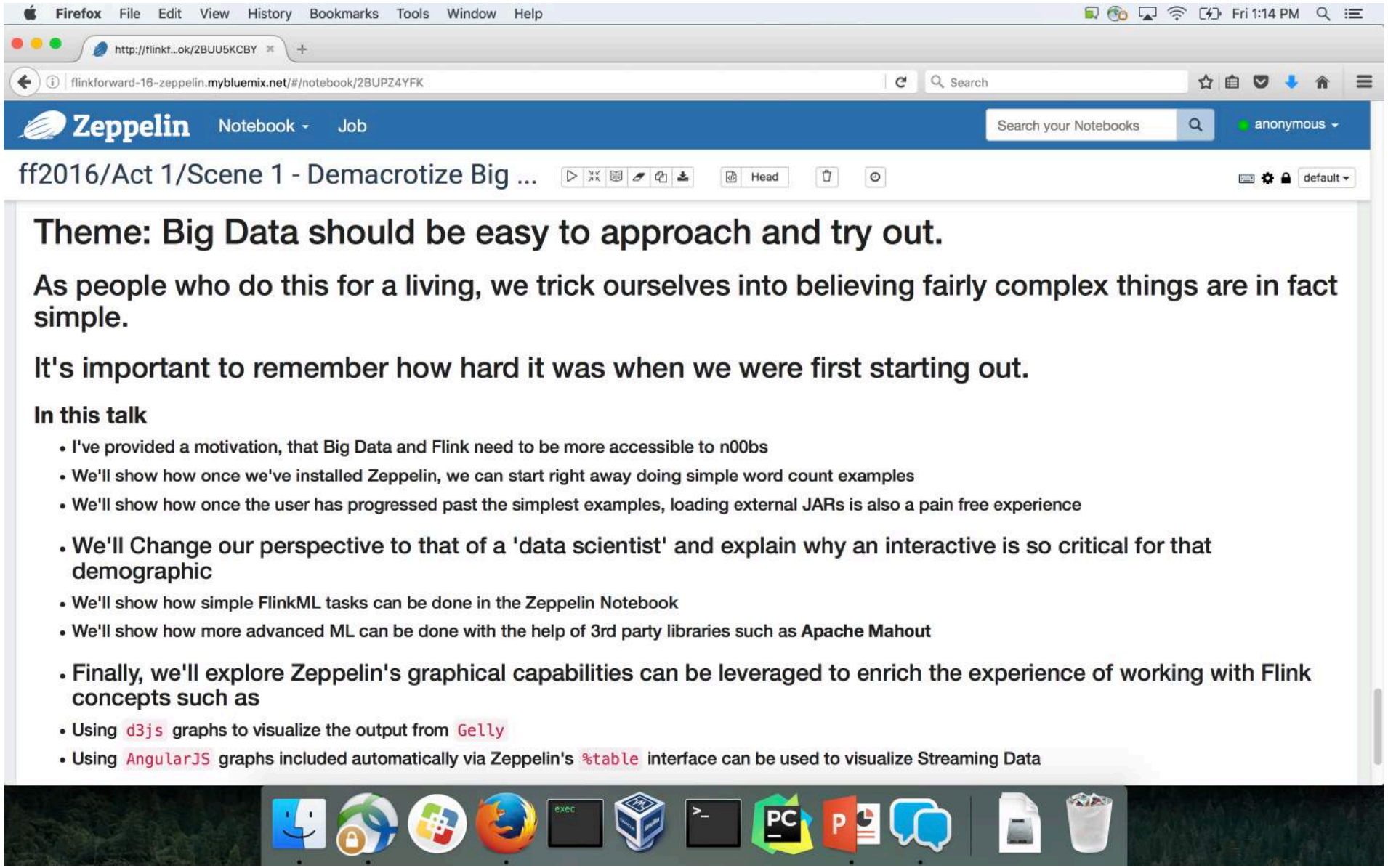
```
wget http://archive.apache.org/dist/zeppelin/zeppelin-0.6.0/zeppelin-0.6.0-bin-all.tgz
tar xzf zeppelin-0.6.0-bin-all.tgz
zeppelin-0.6.0-bin-all/bin/zeppelin-daemon.sh start
```

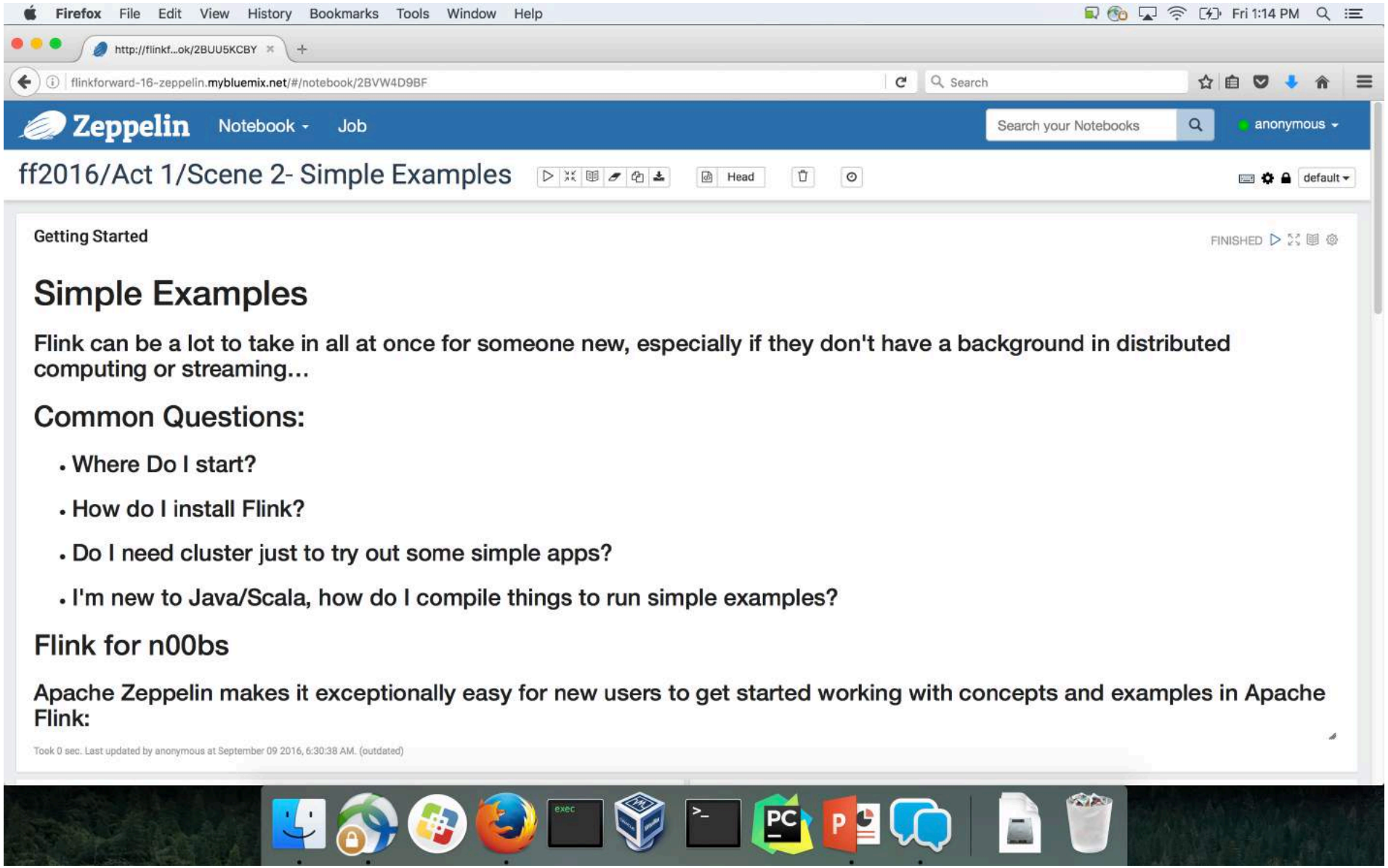
Open Zeppelin in your favorite web-browser at `http://localhost:8080` and type in a paragraph starting with `%flink`

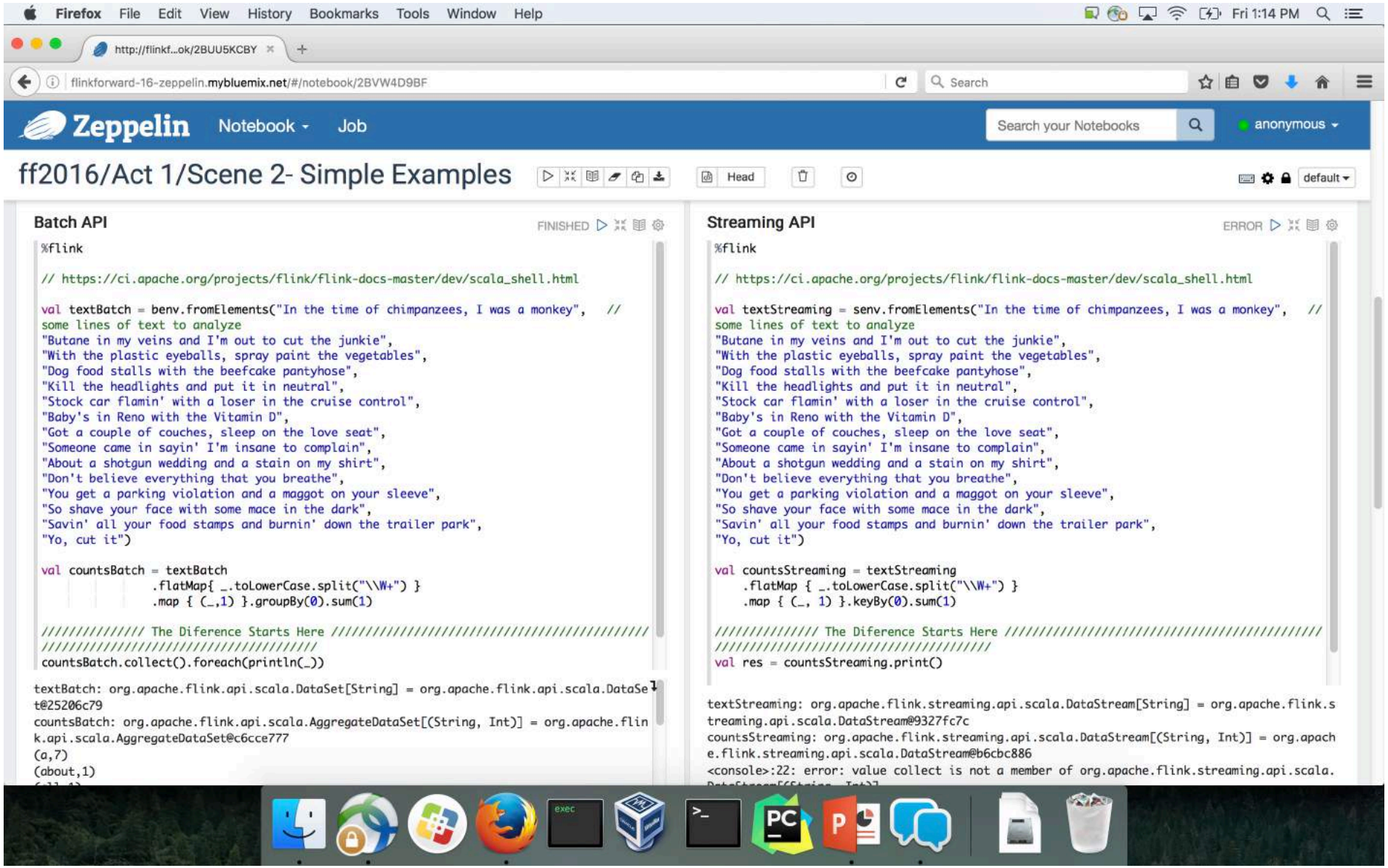

Firefox File Edit View History Bookmarks Tools Window Help
Fri 1:14 PM
http://flinkf...ok/2BUU5KCBY
flinkforward-16-zeppelin.mybluemix.net/#/notebook/2BUPZ4YFK
Search
Zeppelin
Notebook
Job
Search your Notebooks
anonymous
ff2016/Act 1/Scene 1 - Demacrotize Big ...
Head
default
BIG Data Should Be Easy
FINISHED
The above method will start a FlinkMiniCluster()
About 512m on the taskmanager, 256m on the jobmanager. Just enough for hello world and wordcount
Cute, but can I run it against real deployments of Zeppelin?
Yes and No
Cluster Mode- Yes; all of these examples are on cluster
YARN Mode- not YET

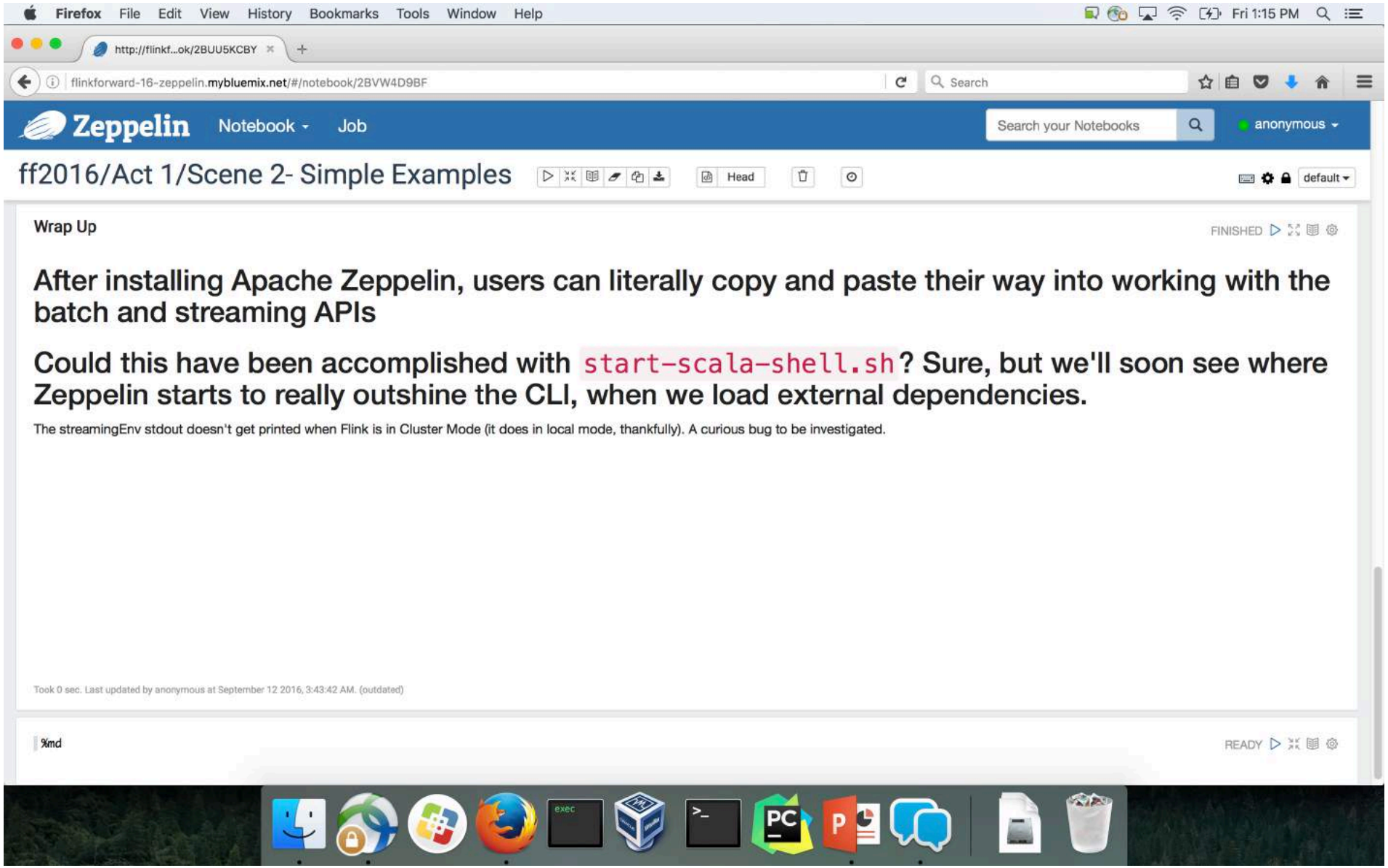

Firefox File Edit View History Bookmarks Tools Window Help
Fri 1:14 PM
http://flinkf...ok/2BUU5KCBY
flinkforward-16-zeppelin.mybluemix.net/#/notebook/2BUPZ4YFK
Search
Zeppelin
Notebook
Job
Search your Notebooks
anonymous
ff2016/Act 1/Scene 1 - Demacrotize Big ...
Head
default
Big Data for the People
FINISHED
My Thesis: Big Data Should Be Easy
Provisioning Hardware, setting up YARN clusters, etc- No trivial task
So to even have the resources to work with Big Data, we must be sophisticated.
FALSE
If you actually believe the above- I will quickly dispell that notion.
IBM Big Insights On Cloud
AmbariUrl
YarnUI
Flink WebUI
Some Git Scripts I wrote make it easy to setup Zeppelin and Flink on Big Insights Cloud
Code presented only to illustrate ease of setup
from data.services.flink import FlinkServiceOnBI
from data.services.zeppelin import ZeppelinServiceOnBI
flink = FlinkServiceOnBI(SERVER, USERNAME, PASSWORD)
f.install()
sleep(10)

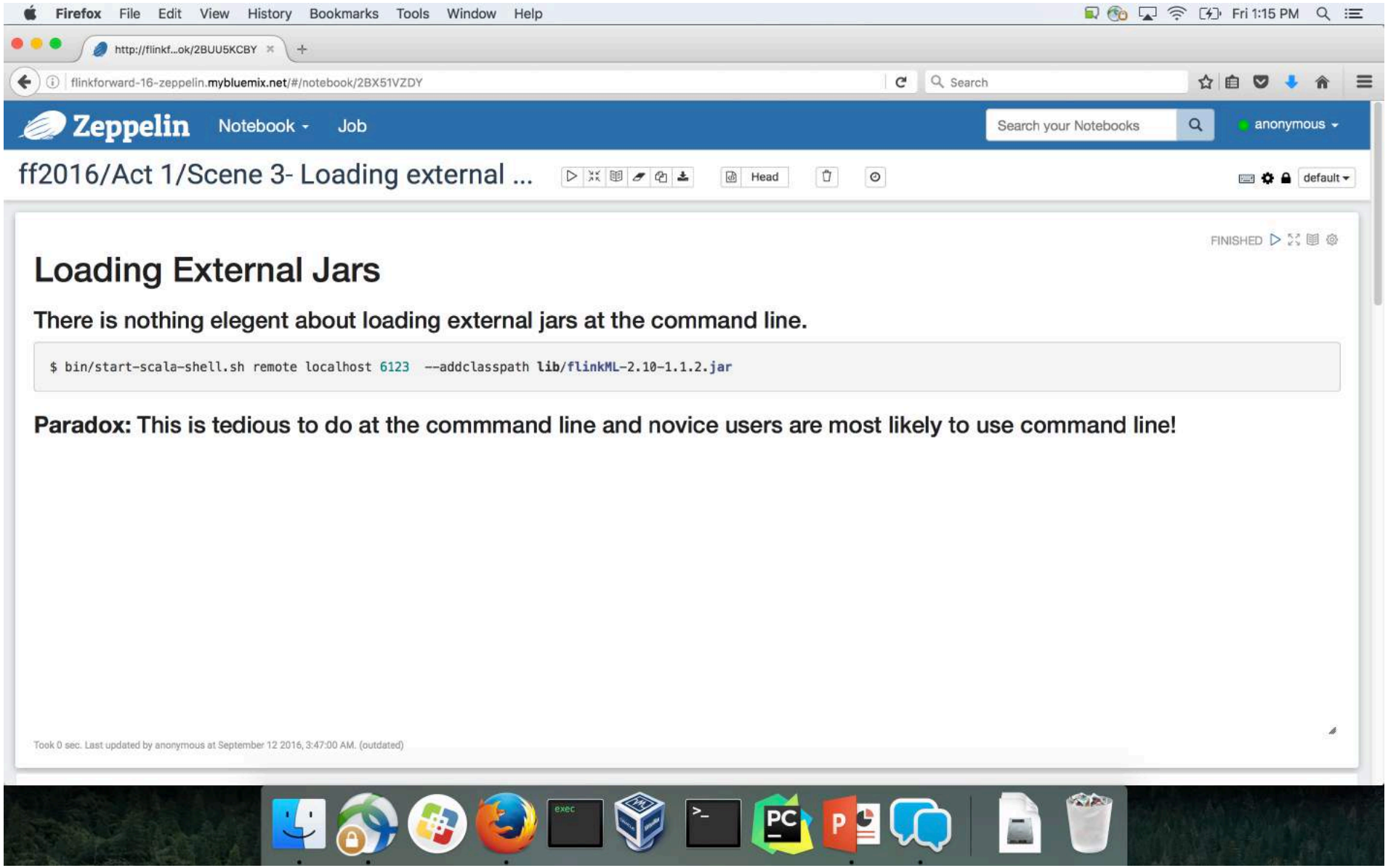

Firefox File Edit View History Bookmarks Tools Window Help
Fri 1:14 PM
http://flinkf...ok/2BUU5KCBY
flinkforward-16-zeppelin.mybluemix.net/#/notebook/2BUPZ4YFK
Search
Zeppelin
Notebook
Job
Search your Notebooks
anonymous
ff2016/Act 1/Scene 1 - Demacrotize Big ...
Head
default
from data.services.flink import FlinkServiceOnBI
from data.services.zeppelin import ZeppelinServiceOnBI
flink = FlinkServiceOnBI(SERVER, USERNAME, PASSWORD)
f.install()
sleep(10)
f.uploadConfig()
f.startCluster()
z = ZeppelinServiceOnBI(SERVER, USERNAME, PASSWORD)
z.install()
z.start()
sleep(5)
z.downloadConfig({"interpreter.json" : "conf/interpreter.json"})
z.updateConfig()
z._updateTerpProp("flink", "host", "localhost")
z._writeTerpJson()
z.uploadConfig()
z.start()
f.deployApp(APP_PREFIX)
z.deployApp(APP_PREFIX)
Along with these Zeppelin Notebooks (which can be imported directly from github), I will also provide a script for recreating this environment on a Big Insights Cloud cluster (which has a free trial).

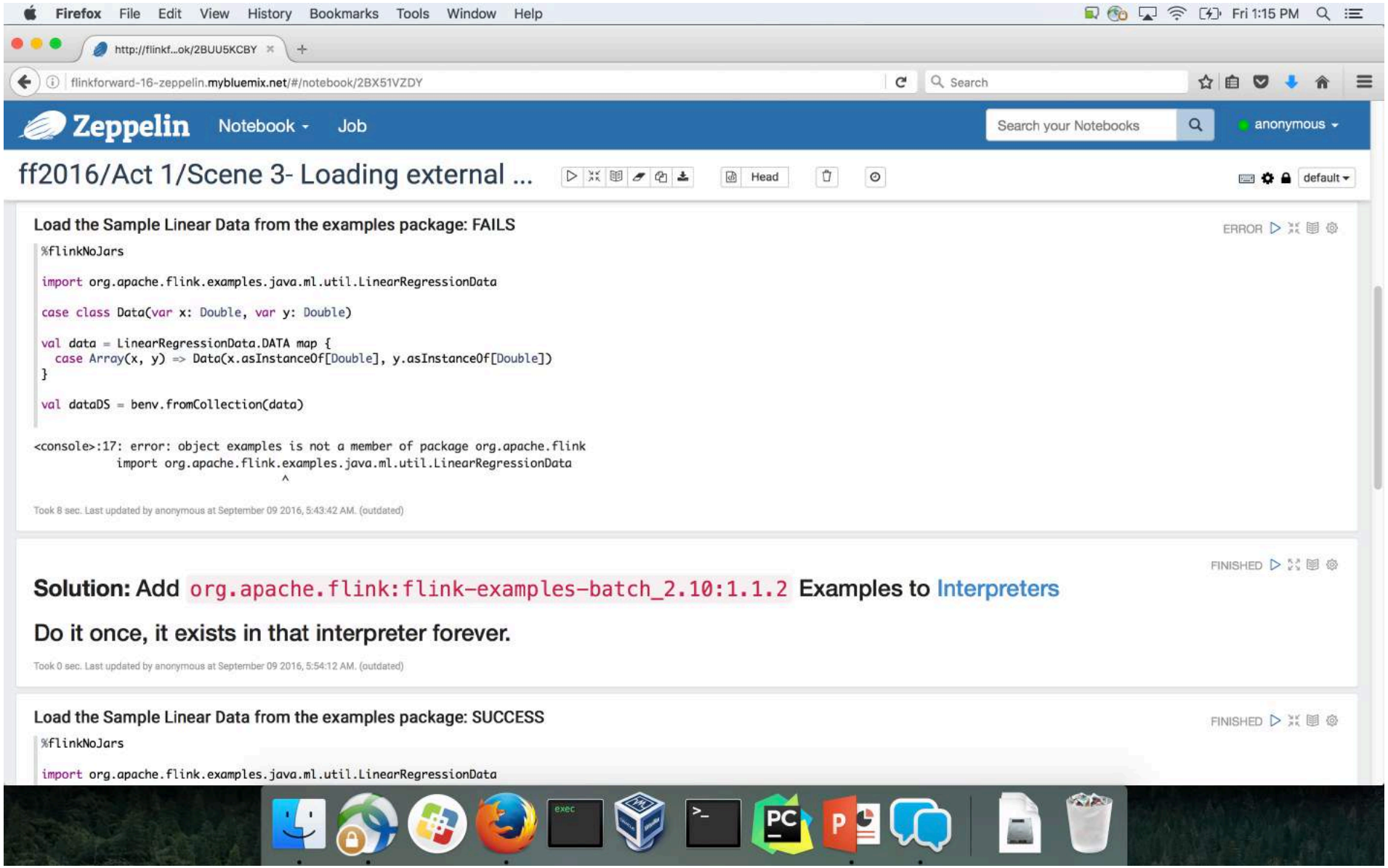

Firefox File Edit View History Bookmarks Tools Window Help
Fri 1:14 PM
http://flinkf...ok/2BUU5KCBY
flinkforward-16-zeppelin.mybluemix.net/#/notebook/2BUPZ4YFK
Search
Zeppelin
Notebook
Job
Search your Notebooks
anonymous
ff2016/Act 1/Scene 1 - Demacrotize Big ...
Head
default
Theme: Big Data should be easy to approach and try out.
As people who do this for a living, we trick ourselves into believing fairly complex things are in fact simple.
It's important to remember how hard it was when we were first starting out.
In this talk
I've provided a motivation, that Big Data and Flink need to be more accessible to n00bs
We'll show how once we've installed Zeppelin, we can start right away doing simple word count examples
We'll show how once the user has progressed past the simplest examples, loading external JARs is also a pain free experience
We'll Change our perspective to that of a 'data scientist' and explain why an interactive is so critical for that demographic
We'll show how simple FlinkML tasks can be done in the Zeppelin Notebook
We'll show how more advanced ML can be done with the help of 3rd party libraries such as Apache Mahout
Finally, we'll explore Zeppelin's graphical capabilities can be leveraged to enrich the experience of working with Flink concepts such as
Using d3js graphs to visualize the output from Gelly
Using AngularJS graphs included automatically via Zeppelin's %table interface can be used to visualize Streaming Data

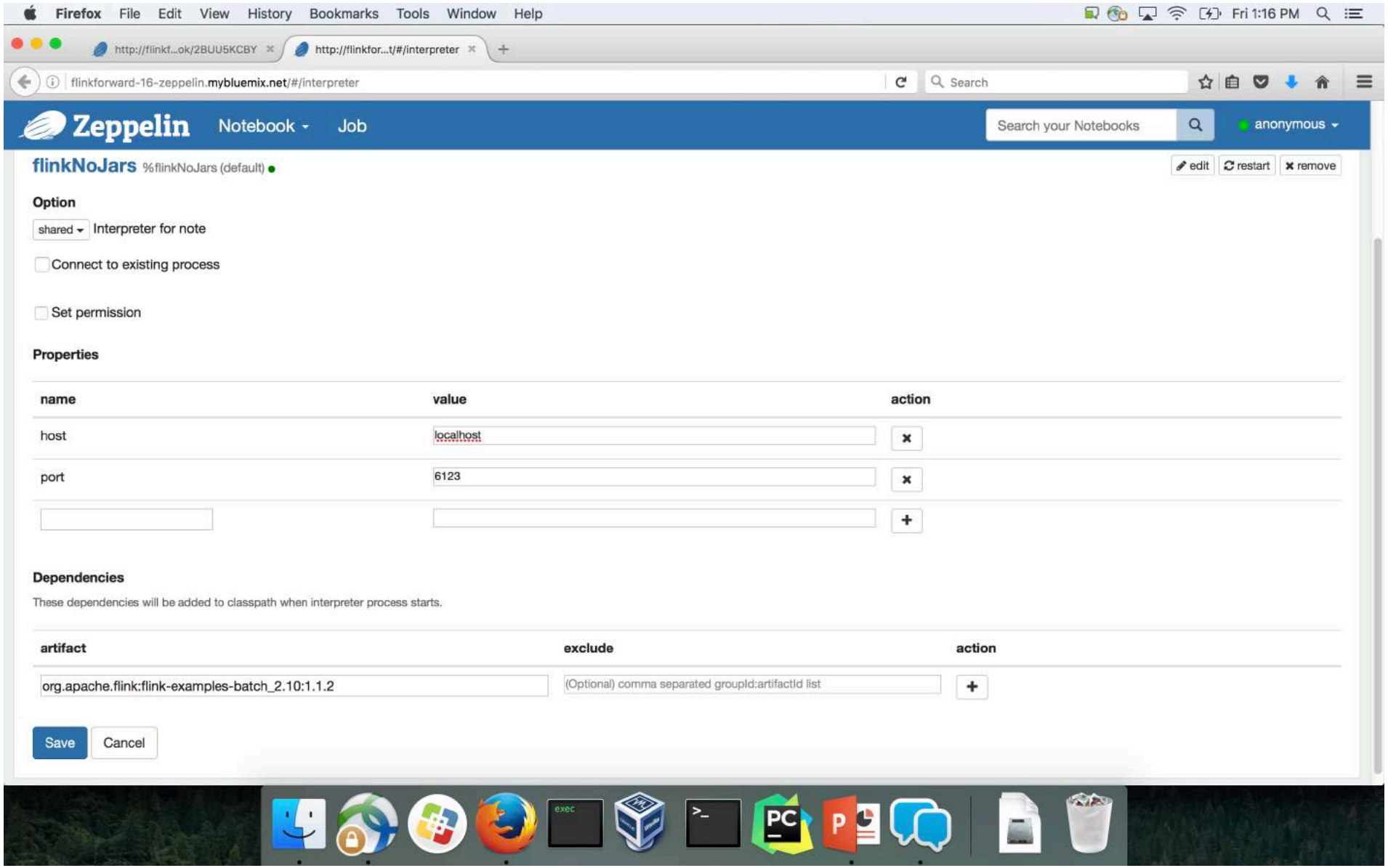

Firefox File Edit View History Bookmarks Tools Window Help
Fri 1:14 PM
http://flinkf...ok/2BUU5KCBY
flinkforward-16-zeppelin.mybluemix.net/#/notebook/2BVW4D9BF
Search
Zeppelin
Notebook
Job
Search your Notebooks
anonymous
ff2016/Act 1/Scene 2- Simple Examples
Head
default
Getting Started
FINISHED
Simple Examples
Flink can be a lot to take in all at once for someone new, especially if they don't have a background in distributed computing or streaming...
Common Questions:
Where Do I start?
How do I install Flink?
Do I need cluster just to try out some simple apps?
I'm new to Java/Scala, how do I compile things to run simple examples?
Flink for n00bs
Apache Zeppelin makes it exceptionally easy for new users to get started working with concepts and examples in Apache Flink:
Took 0 sec. Last updated by anonymous at September 09 2016, 6:30:38 AM. (outdated)

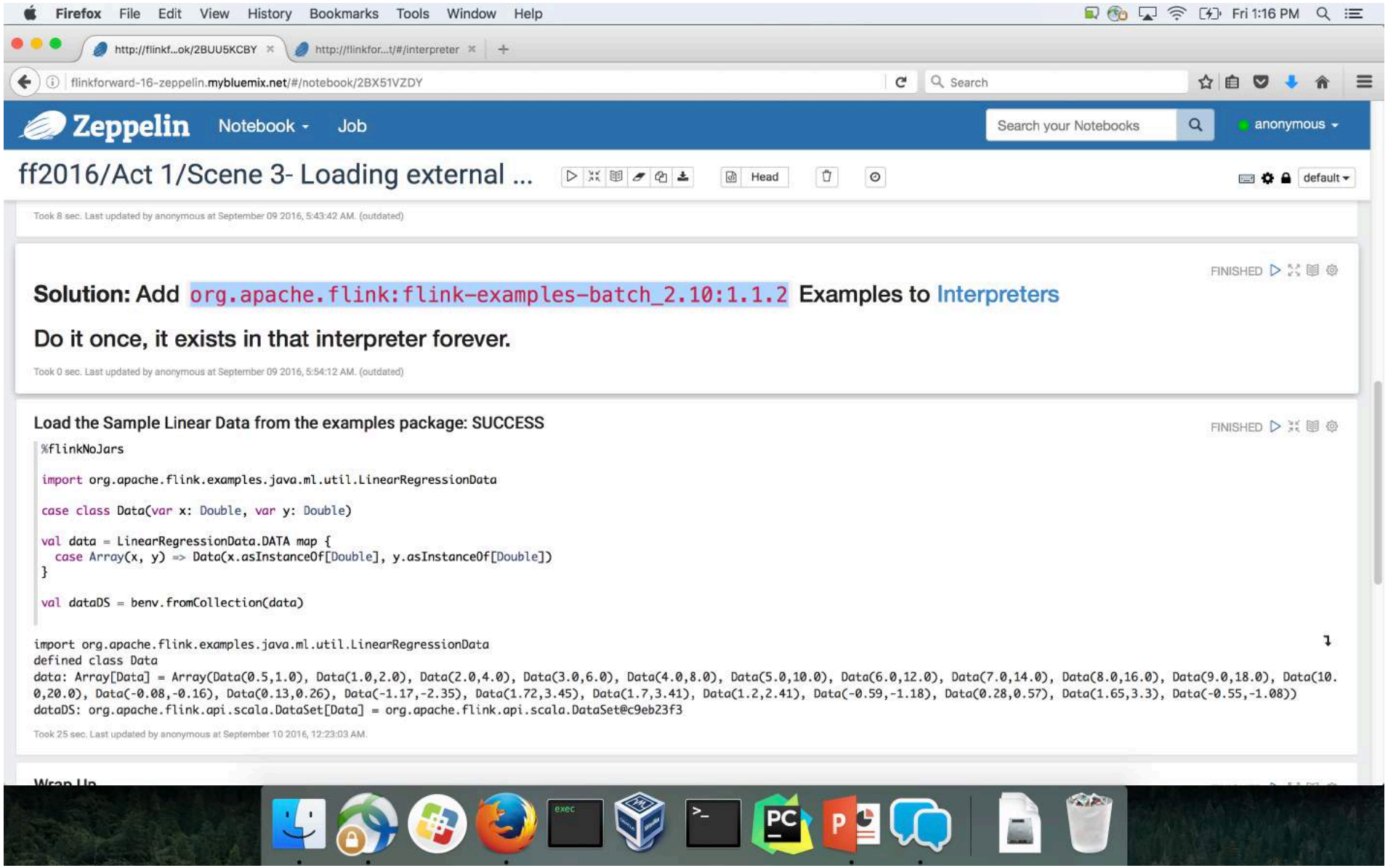

Firefox File Edit View History Bookmarks Tools Window Help
Fri 1:14 PM
http://flinkf...ok/2BUU5KCBY
flinkforward-16-zeppelin.mybluemix.net/#/notebook/2BVW4D9BF
Search
Zeppelin
Notebook
Job
Search your Notebooks
anonymous
ff2016/Act 1/Scene 2- Simple Examples
Head
default
Batch API
FINISHED
%flink
// https://ci.apache.org/projects/flink/flink-docs-master/dev/scala_shell.html
val textBatch = benv.fromElements("In the time of chimpanzees, I was a monkey", //
some lines of text to analyze
"Butane in my veins and I'm out to cut the junkie",
"With the plastic eyeballs, spray paint the vegetables",
"Dog food stalls with the beefcake pantyhose",
"Kill the headlights and put it in neutral",
"Stock car flamin' with a loser in the cruise control",
"Baby's in Reno with the Vitamin D",
"Got a couple of couches, sleep on the love seat",
"Someone came in sayin' I'm insane to complain",
"About a shotgun wedding and a stain on my shirt",
"Don't believe everything that you breathe",
"You get a parking violation and a maggot on your sleeve",
"So shave your face with some mace in the dark",
"Savin' all your food stamps and burnin' down the trailer park",
"Yo, cut it")
val countsBatch = textBatch
.flatMap{ _.toLowerCase.split("\\W+") }
.map { (_,1) }.groupBy(0).sum(1)
/////////////// The Diference Starts Here //
//////////////////////////////////////
countsBatch.collect().foreach(println(_))
textBatch: org.apache.flink.api.scala.DataSet[String] = org.apache.flink.api.scala.DataSet@25206c79
countsBatch: org.apache.flink.api.scala.AggregateDataSet[(String, Int)] = org.apache.flink.api.scala.AggregateDataSet@c6cce777
(a,7)
(about,1)
Streaming API
ERROR
%flink
// https://ci.apache.org/projects/flink/flink-docs-master/dev/scala_shell.html
val textStreaming = senv.fromElements("In the time of chimpanzees, I was a monkey", //
some lines of text to analyze
"Butane in my veins and I'm out to cut the junkie",
"With the plastic eyeballs, spray paint the vegetables",
"Dog food stalls with the beefcake pantyhose",
"Kill the headlights and put it in neutral",
"Stock car flamin' with a loser in the cruise control",
"Baby's in Reno with the Vitamin D",
"Got a couple of couches, sleep on the love seat",
"Someone came in sayin' I'm insane to complain",
"About a shotgun wedding and a stain on my shirt",
"Don't believe everything that you breathe",
"You get a parking violation and a maggot on your sleeve",
"So shave your face with some mace in the dark",
"Savin' all your food stamps and burnin' down the trailer park",
"Yo, cut it")
val countsStreaming = textStreaming
.flatMap { _.toLowerCase.split("\\W+") }
.map { (_, 1) }.keyBy(0).sum(1)
/////////////// The Diference Starts Here //
//////////////////////////////////////
val res = countsStreaming.print()
textStreaming: org.apache.flink.streaming.api.scala.DataStream[String] = org.apache.flink.streaming.api.scala.DataStream@9327fc7c
countsStreaming: org.apache.flink.streaming.api.scala.DataStream[(String, Int)] = org.apache.flink.streaming.api.scala.DataStream@b6cbc886
<console>:22: error: value collect is not a member of org.apache.flink.streaming.api.scala.

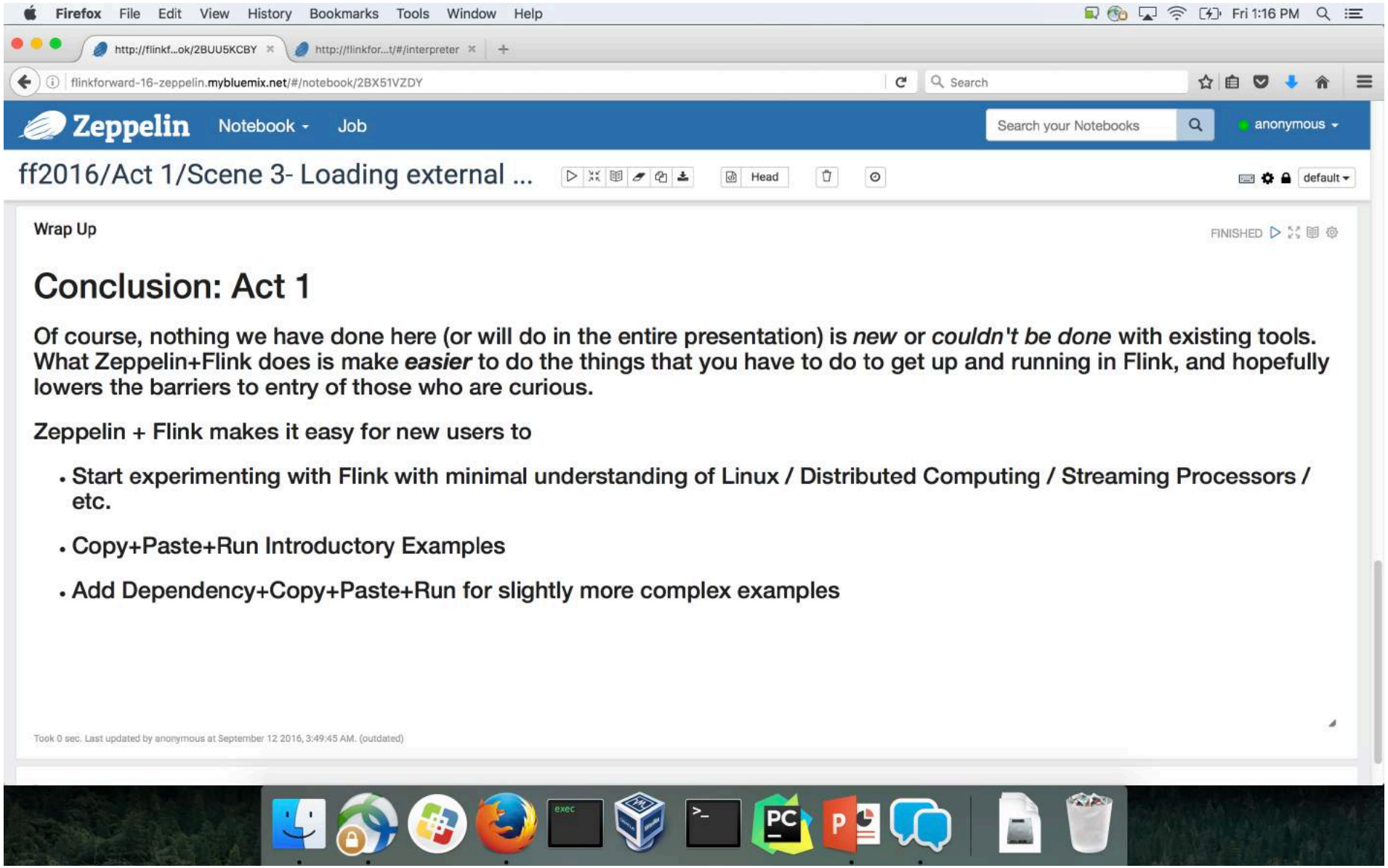

Firefox File Edit View History Bookmarks Tools Window Help
Fri 1:15 PM
http://flinkf...ok/2BUU5KCBY
flinkforward-16-zeppelin.mybluemix.net/#/notebook/2BVW4D9BF
Search
Zeppelin
Notebook
Job
Search your Notebooks
anonymous
ff2016/Act 1/Scene 2- Simple Examples
Head
default
Wrap Up
FINISHED
After installing Apache Zeppelin, users can literally copy and paste their way into working with the batch and streaming APIs
Could this have been accomplished with start-scala-shell.sh? Sure, but we'll soon see where Zeppelin starts to really outshine the CLI, when we load external dependencies.
The streamingEnv stdout doesn't get printed when Flink is in Cluster Mode (it does in local mode, thankfully). A curious bug to be investigated.
Took 0 sec. Last updated by anonymous at September 12 2016, 3:43:42 AM. (outdated)
%md
READY

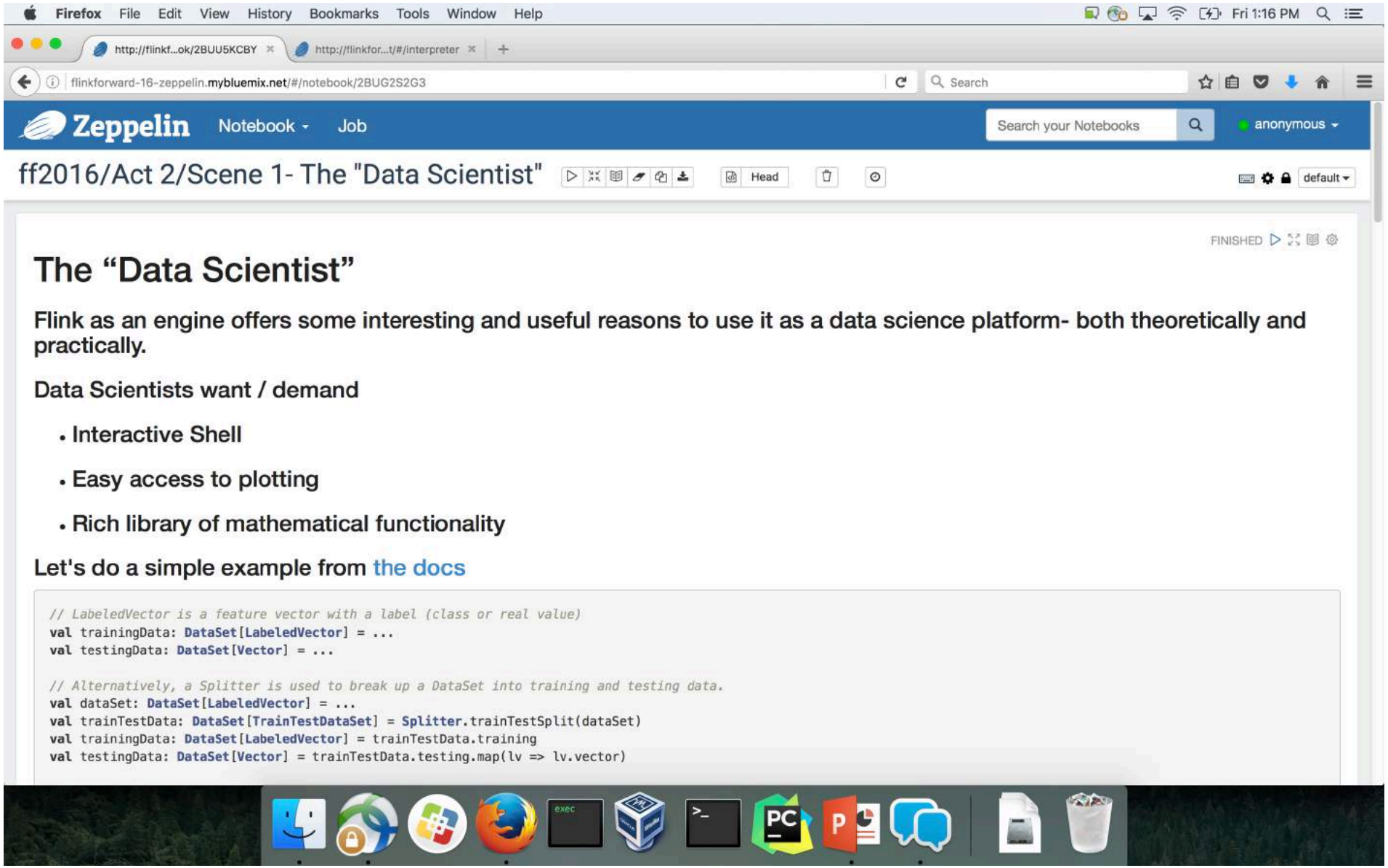

Firefox File Edit View History Bookmarks Tools Window Help
Fri 1:15 PM
http://flinkf...ok/2BUU5KCBY
flinkforward-16-zeppelin.mybluemix.net/#/notebook/2BX51VZDY
Search
Zeppelin
Notebook
Job
Search your Notebooks
anonymous
ff2016/Act 1/Scene 3- Loading external ...
Head
default
FINISHED
Loading External Jars
There is nothing elegent about loading external jars at the command line.
$ bin/start-scala-shell.sh remote localhost 6123 --addclasspath lib/flinkML-2.10-1.1.2.jar
Paradox: This is tedious to do at the commmand line and novice users are most likely to use command line!
Took 0 sec. Last updated by anonymous at September 12 2016, 3:47:00 AM. (outdated)

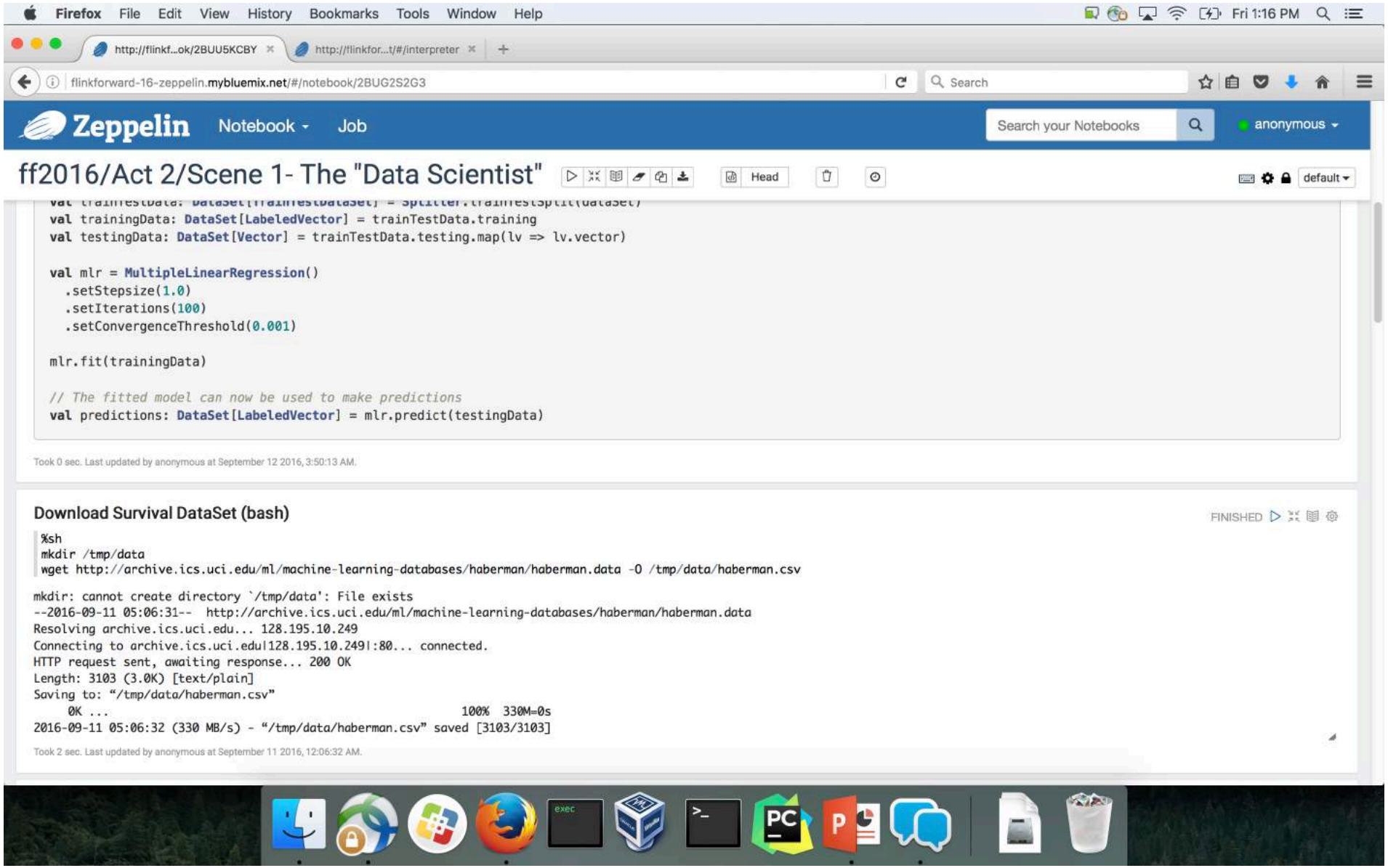

Firefox File Edit View History Bookmarks Tools Window Help
Fri 1:15 PM
http://flinkf...ok/2BUU5KCBY
flinkforward-16-zeppelin.mybluemix.net/#/notebook/2BX51VZDY
Search
Zeppelin
Notebook
Job
Search your Notebooks
anonymous
ff2016/Act 1/Scene 3- Loading external ...
Head
default
Load the Sample Linear Data from the examples package: FAILS
ERROR
%flinkNoJars
import org.apache.flink.examples.java.ml.util.LinearRegressionData
case class Data(var x: Double, var y: Double)
val data = LinearRegressionData.DATA map {
case Array(x, y) => Data(x.asInstanceOf[Double], y.asInstanceOf[Double])
}
val dataDS = benv.fromCollection(data)
<console>:17: error: object examples is not a member of package org.apache.flink
import org.apache.flink.examples.java.ml.util.LinearRegressionData
^
Took 8 sec. Last updated by anonymous at September 09 2016, 5:43:42 AM. (outdated)
FINISHED
Solution: Add org.apache.flink:flink-examples-batch_2.10:1.1.2 Examples to Interpreters
Do it once, it exists in that interpreter forever.
Took 0 sec. Last updated by anonymous at September 09 2016, 5:54:12 AM. (outdated)
Load the Sample Linear Data from the examples package: SUCCESS
FINISHED
%flinkNoJars
import org.apache.flink.examples.java.ml.util.LinearRegressionData

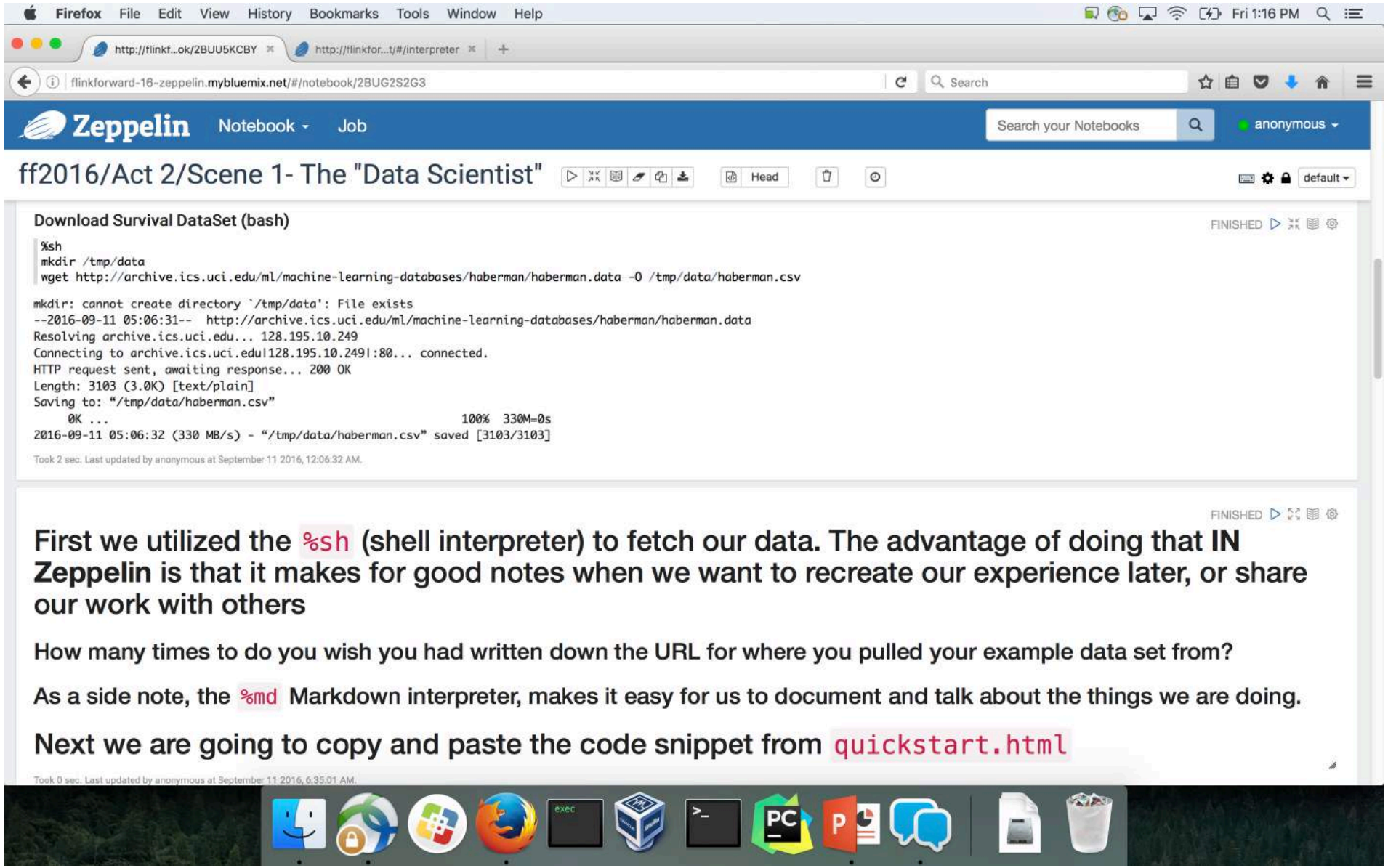

Firefox File Edit View History Bookmarks Tools Window Help
Fri 1:16 PM
http://flinkf...ok/2BUU5KCBY
http://flinkfor...t/#/interpreter
flinkforward-16-zeppelin.mybluemix.net/#/interpreter
Search
Zeppelin
Notebook
Job
Search your Notebooks
anonymous
flinkNoJars %flinkNoJars (default)
edit
restart
remove
Option
shared
Interpreter for note
Connect to existing process
Set permission
Properties
name
value
action
host
localhost
port
6123
Dependencies
These dependencies will be added to classpath when interpreter process starts.
artifact
exclude
action
org.apache.flink:flink-examples-batch_2.10:1.1.2
(Optional) comma separated groupId:artifactId list
Save
Cancel

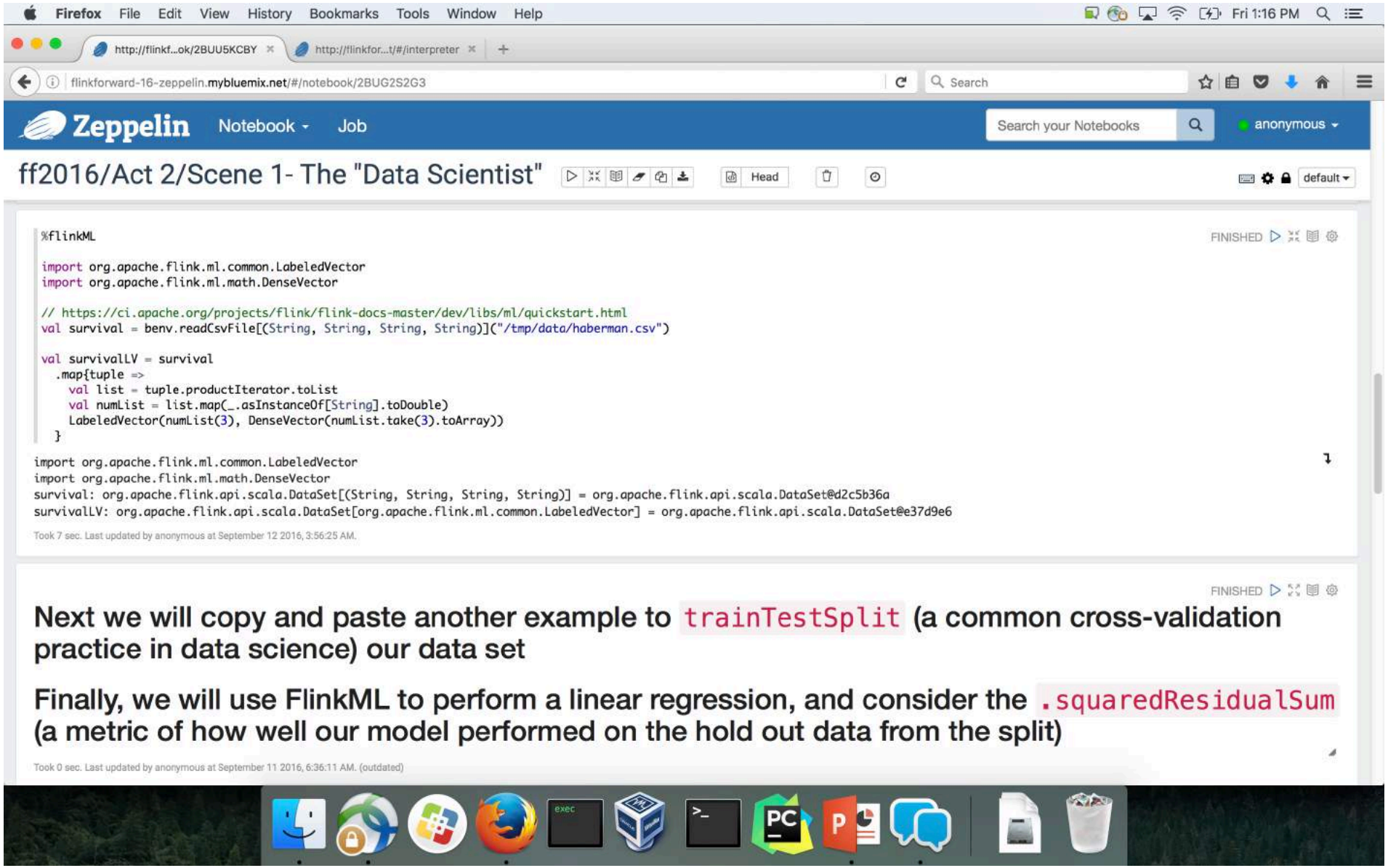

Firefox File Edit View History Bookmarks Tools Window Help
Fri 1:16 PM
http://flinkf...ok/2BUU5KCBY
http://flinkfor...t/#/interpreter
flinkforward-16-zeppelin.mybluemix.net/#/notebook/2BX51VZDY
Search
Zeppelin
Notebook
Job
Search your Notebooks
anonymous
ff2016/Act 1/Scene 3- Loading external ...
Head
default
Took 8 sec. Last updated by anonymous at September 09 2016, 5:43:42 AM. (outdated)
FINISHED
Solution: Add org.apache.flink:flink-examples-batch_2.10:1.1.2 Examples to Interpreters
Do it once, it exists in that interpreter forever.
Took 0 sec. Last updated by anonymous at September 09 2016, 5:54:12 AM. (outdated)
Load the Sample Linear Data from the examples package: SUCCESS
FINISHED
%flinkNoJars
import org.apache.flink.examples.java.ml.util.LinearRegressionData
case class Data(var x: Double, var y: Double)
val data = LinearRegressionData.DATA map {
case Array(x, y) => Data(x.asInstanceOf[Double], y.asInstanceOf[Double])
}
val dataDS = benv.fromCollection(data)
import org.apache.flink.examples.java.ml.util.LinearRegressionData
defined class Data
data: Array[Data] = Array(Data(0.5,1.0), Data(1.0,2.0), Data(2.0,4.0), Data(3.0,6.0), Data(4.0,8.0), Data(5.0,10.0), Data(6.0,12.0), Data(7.0,14.0), Data(8.0,16.0), Data(9.0,18.0), Data(10.0,20.0), Data(-0.08,-0.16), Data(0.13,0.26), Data(-1.17,-2.35), Data(1.72,3.45), Data(1.7,3.41), Data(1.2,2.41), Data(-0.59,-1.18), Data(0.28,0.57), Data(1.65,3.3), Data(-0.55,-1.08))
dataDS: org.apache.flink.api.scala.DataSet[Data] = org.apache.flink.api.scala.DataSet@c9eb23f3
Took 25 sec. Last updated by anonymous at September 10 2016, 12:23:03 AM.
Wrap Up

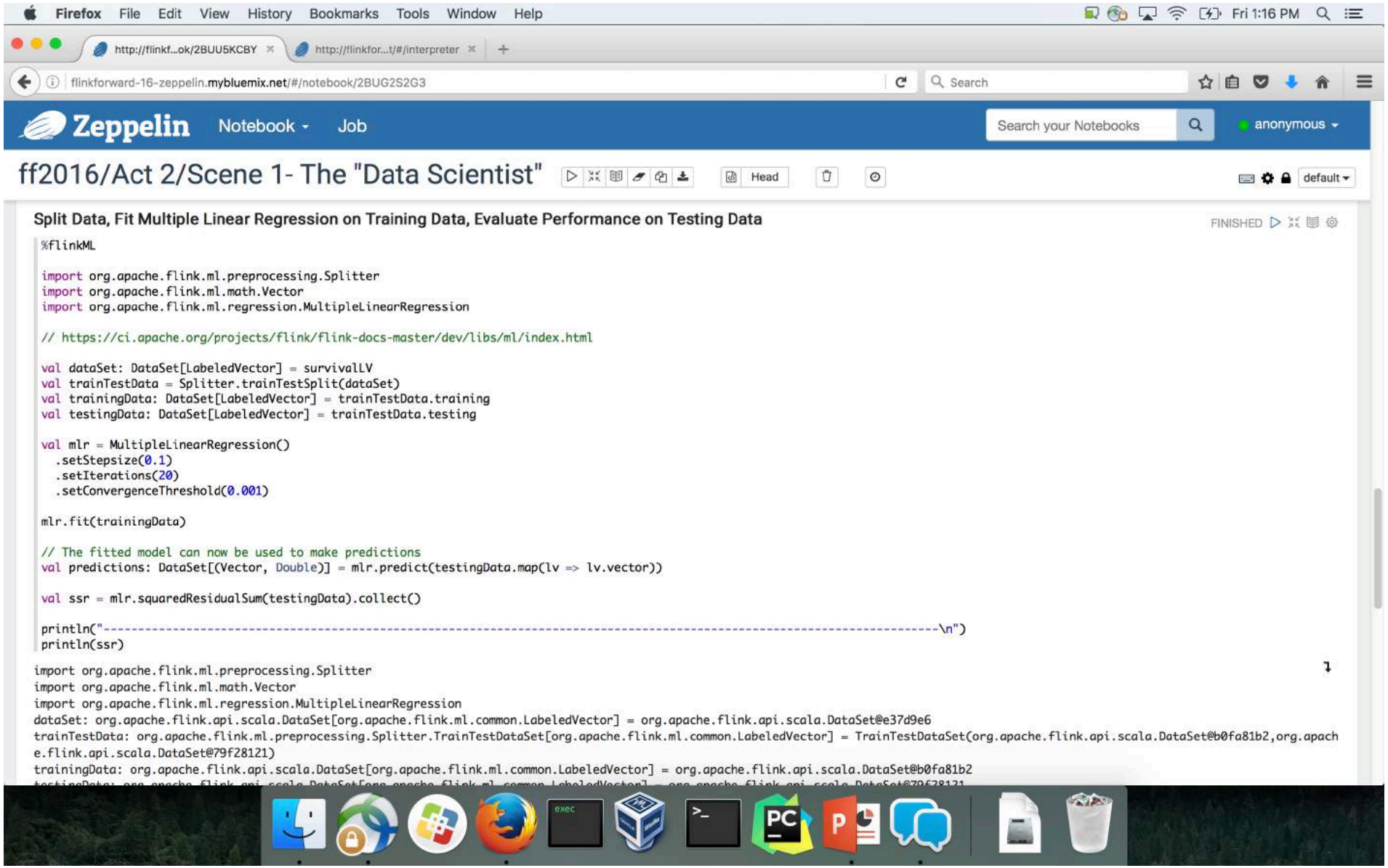

Firefox File Edit View History Bookmarks Tools Window Help
Fri 1:16 PM
http://flinkf...ok/2BUU5KCBY
http://flinkfor...t/#/interpreter
flinkforward-16-zeppelin.mybluemix.net/#/notebook/2BX51VZDY
Search
Zeppelin
Notebook
Job
Search your Notebooks
anonymous
ff2016/Act 1/Scene 3- Loading external ...
Head
default
Wrap Up
FINISHED
Conclusion: Act 1
Of course, nothing we have done here (or will do in the entire presentation) is new or couldn't be done with existing tools. What Zeppelin+Flink does is make easier to do the things that you have to do to get up and running in Flink, and hopefully lowers the barriers to entry of those who are curious.
Zeppelin + Flink makes it easy for new users to
• Start experimenting with Flink with minimal understanding of Linux / Distributed Computing / Streaming Processors / etc.
• Copy+Paste+Run Introductory Examples
• Add Dependency+Copy+Paste+Run for slightly more complex examples
Took 0 sec. Last updated by anonymous at September 12 2016, 3:49:45 AM. (outdated)

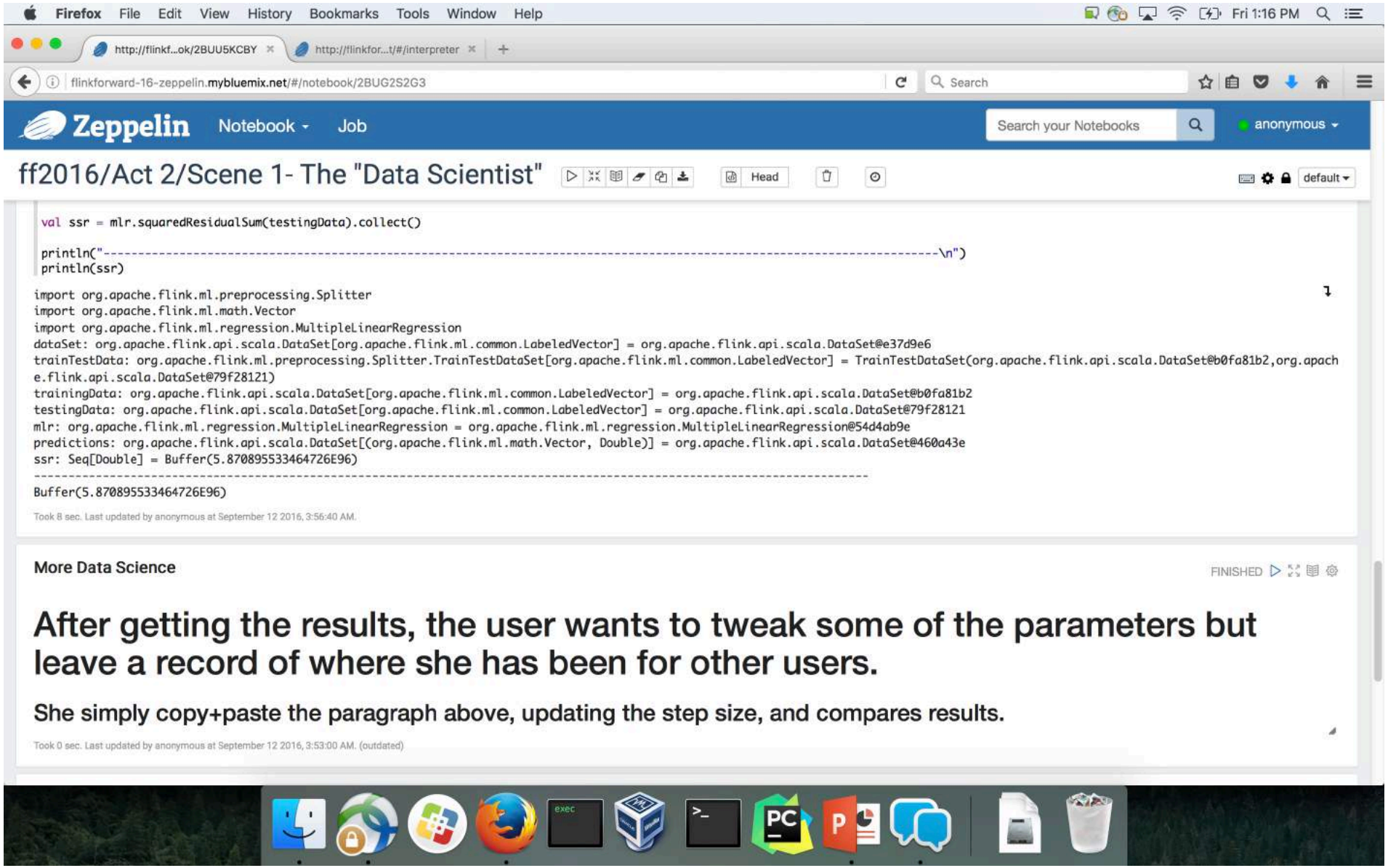

Firefox File Edit View History Bookmarks Tools Window Help
Fri 1:16 PM
http://flinkf...ok/2BUU5KCBY
http://flinkfor...t/#/interpreter
flinkforward-16-zeppelin.mybluemix.net/#/notebook/2BUG2S2G3
Search
Zeppelin
Notebook
Job
Search your Notebooks
anonymous
ff2016/Act 2/Scene 1- The "Data Scientist"
Head
default
FINISHED
The “Data Scientist”
Flink as an engine offers some interesting and useful reasons to use it as a data science platform- both theoretically and practically.
Data Scientists want / demand
Interactive Shell
Easy access to plotting
Rich library of mathematical functionality
Let's do a simple example from the docs
// LabeledVector is a feature vector with a label (class or real value)
val trainingData: DataSet[LabeledVector] = ...
val testingData: DataSet[Vector] = ...
// Alternatively, a Splitter is used to break up a DataSet into training and testing data.
val dataSet: DataSet[LabeledVector] = ...
val trainTestData: DataSet[TrainTestDataSet] = Splitter.trainTestSplit(dataSet)
val trainingData: DataSet[LabeledVector] = trainTestData.training
val testingData: DataSet[Vector] = trainTestData.testing.map(lv => lv.vector)


Firefox File Edit View History Bookmarks Tools Window Help
Fri 1:16 PM
http://flinkf...ok/2BUU5KCBY
http://flinkfor...t/#/interpreter
flinkforward-16-zeppelin.mybluemix.net/#/notebook/2BUG2S2G3
Search
Zeppelin
Notebook
Job
Search your Notebooks
anonymous
ff2016/Act 2/Scene 1- The "Data Scientist"
Head
default
val trainingData: DataSet[LabeledVector] = trainTestData.training
val testingData: DataSet[Vector] = trainTestData.testing.map(lv => lv.vector)
val mlr = MultipleLinearRegression()
.setStepsize(1.0)
.setIterations(100)
.setConvergenceThreshold(0.001)
mlr.fit(trainingData)
// The fitted model can now be used to make predictions
val predictions: DataSet[LabeledVector] = mlr.predict(testingData)
Took 0 sec. Last updated by anonymous at September 12 2016, 3:50:13 AM.
Download Survival DataSet (bash)
FINISHED
%sh
mkdir /tmp/data
wget http://archive.ics.uci.edu/ml/machine-learning-databases/haberman/haberman.data -O /tmp/data/haberman.csv
mkdir: cannot create directory `/tmp/data': File exists
--2016-09-11 05:06:31-- http://archive.ics.uci.edu/ml/machine-learning-databases/haberman/haberman.data
Resolving archive.ics.uci.edu... 128.195.10.249
Connecting to archive.ics.uci.edu|128.195.10.249|:80... connected.
HTTP request sent, awaiting response... 200 OK
Length: 3103 (3.0K) [text/plain]
Saving to: “/tmp/data/haberman.csv”
0K ... 100% 330M=0s
2016-09-11 05:06:32 (330 MB/s) - “/tmp/data/haberman.csv” saved [3103/3103]
Took 2 sec. Last updated by anonymous at September 11 2016, 12:06:32 AM.

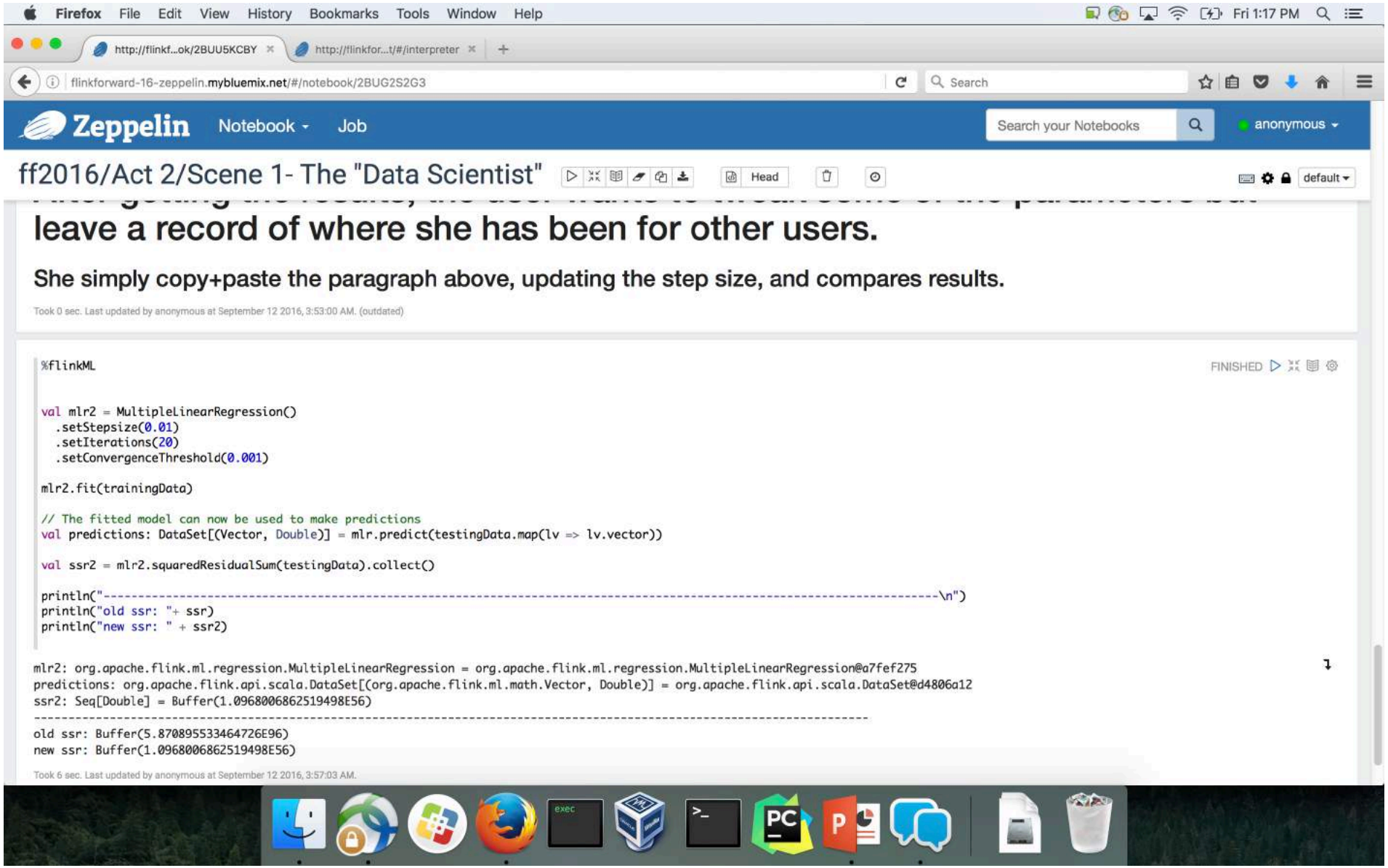

Firefox File Edit View History Bookmarks Tools Window Help
Fri 1:16 PM
http://flinkf...ok/2BUU5KCBY
http://flinkfor...t/#/interpreter
flinkforward-16-zeppelin.mybluemix.net/#/notebook/2BUG2S2G3
Search
Zeppelin
Notebook
Job
Search your Notebooks
anonymous
ff2016/Act 2/Scene 1- The "Data Scientist"
Head
default
Download Survival DataSet (bash)
FINISHED
%sh
mkdir /tmp/data
wget http://archive.ics.uci.edu/ml/machine-learning-databases/haberman/haberman.data -O /tmp/data/haberman.csv
mkdir: cannot create directory `/tmp/data': File exists
--2016-09-11 05:06:31-- http://archive.ics.uci.edu/ml/machine-learning-databases/haberman/haberman.data
Resolving archive.ics.uci.edu... 128.195.10.249
Connecting to archive.ics.uci.edu|128.195.10.249|:80... connected.
HTTP request sent, awaiting response... 200 OK
Length: 3103 (3.0K) [text/plain]
Saving to: "/tmp/data/haberman.csv"
0K ... 100% 330M=0s
2016-09-11 05:06:32 (330 MB/s) - "/tmp/data/haberman.csv" saved [3103/3103]
Took 2 sec. Last updated by anonymous at September 11 2016, 12:06:32 AM.
FINISHED
First we utilized the %sh (shell interpreter) to fetch our data. The advantage of doing that IN Zeppelin is that it makes for good notes when we want to recreate our experience later, or share our work with others
How many times to do you wish you had written down the URL for where you pulled your example data set from?
As a side note, the %md Markdown interpreter, makes it easy for us to document and talk about the things we are doing.
Next we are going to copy and paste the code snippet from quickstart.html
Took 0 sec. Last updated by anonymous at September 11 2016, 6:35:01 AM.

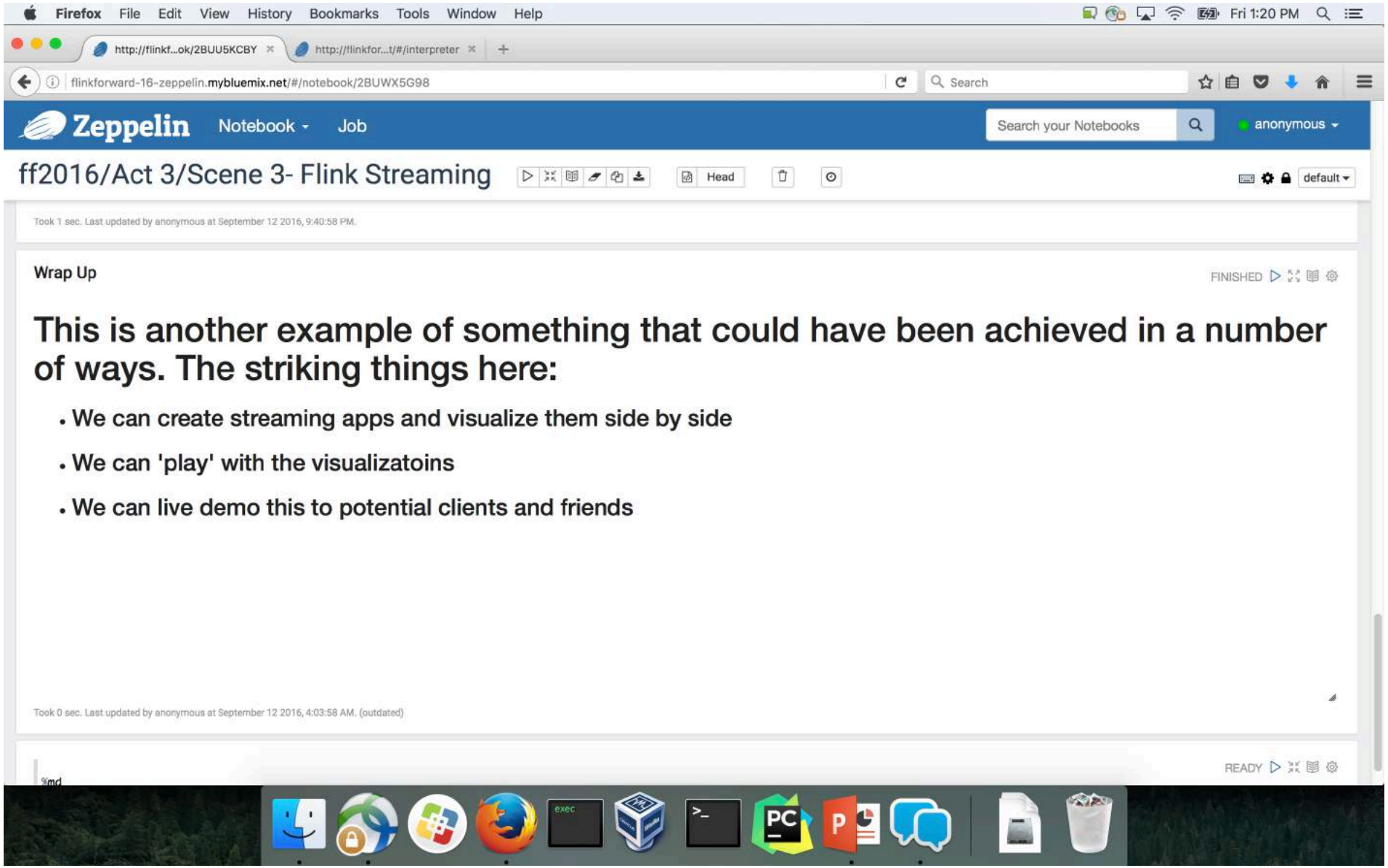

Firefox File Edit View History Bookmarks Tools Window Help
Fri 1:16 PM
http://flinkf...ok/2BUU5KCBY
http://flinkfor...t/#/interpreter
flinkforward-16-zeppelin.mybluemix.net/#/notebook/2BUG2S2G3
Search
Zeppelin
Notebook
Job
Search your Notebooks
anonymous
ff2016/Act 2/Scene 1- The "Data Scientist"
Head
default
%flinkML
FINISHED
import org.apache.flink.ml.common.LabeledVector
import org.apache.flink.ml.math.DenseVector
// https://ci.apache.org/projects/flink/flink-docs-master/dev/libs/ml/quickstart.html
val survival = benv.readCsvFile[(String, String, String, String)]("/tmp/data/haberman.csv")
val survivalLV = survival
.map{tuple =>
val list = tuple.productIterator.toList
val numList = list.map(_.asInstanceOf[String].toDouble)
LabeledVector(numList(3), DenseVector(numList.take(3).toArray))
}
import org.apache.flink.ml.common.LabeledVector
import org.apache.flink.ml.math.DenseVector
survival: org.apache.flink.api.scala.DataSet[(String, String, String, String)] = org.apache.flink.api.scala.DataSet@d2c5b36a
survivalLV: org.apache.flink.api.scala.DataSet[org.apache.flink.ml.common.LabeledVector] = org.apache.flink.api.scala.DataSet@e37d9e6
Took 7 sec. Last updated by anonymous at September 12 2016, 3:56:25 AM.
FINISHED
Next we will copy and paste another example to trainTestSplit (a common cross-validation practice in data science) our data set
Finally, we will use FlinkML to perform a linear regression, and consider the .squaredResidualSum (a metric of how well our model performed on the hold out data from the split)
Took 0 sec. Last updated by anonymous at September 11 2016, 6:36:11 AM. (outdated)

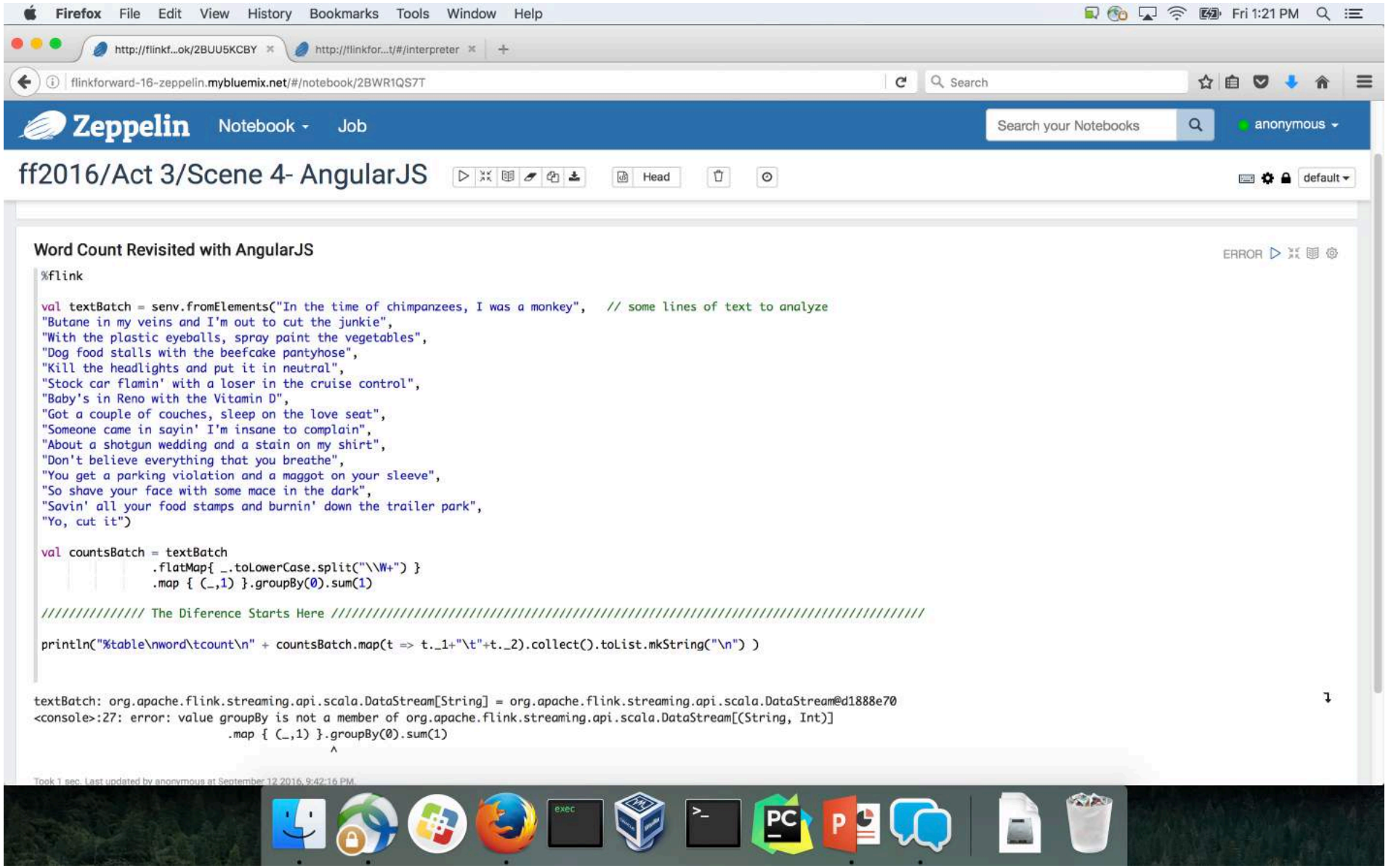

Firefox File Edit View History Bookmarks Tools Window Help
Fri 1:16 PM
http://flinkf...ok/2BUU5KCBY
http://flinkfor...t/#/interpreter
flinkforward-16-zeppelin.mybluemix.net/#/notebook/2BUG2S2G3
Search
Zeppelin
Notebook
Job
Search your Notebooks
anonymous
ff2016/Act 2/Scene 1- The "Data Scientist"
Head
default
Split Data, Fit Multiple Linear Regression on Training Data, Evaluate Performance on Testing Data
FINISHED
%flinkML
import org.apache.flink.ml.preprocessing.Splitter
import org.apache.flink.ml.math.Vector
import org.apache.flink.ml.regression.MultipleLinearRegression
// https://ci.apache.org/projects/flink/flink-docs-master/dev/libs/ml/index.html
val dataSet: DataSet[LabeledVector] = survivalLV
val trainTestData = Splitter.trainTestSplit(dataSet)
val trainingData: DataSet[LabeledVector] = trainTestData.training
val testingData: DataSet[LabeledVector] = trainTestData.testing
val mlr = MultipleLinearRegression()
.setStepsize(0.1)
.setIterations(20)
.setConvergenceThreshold(0.001)
mlr.fit(trainingData)
// The fitted model can now be used to make predictions
val predictions: DataSet[(Vector, Double)] = mlr.predict(testingData.map(lv => lv.vector))
val ssr = mlr.squaredResidualSum(testingData).collect()
println("----------\n")
println(ssr)
import org.apache.flink.ml.preprocessing.Splitter
import org.apache.flink.ml.math.Vector
import org.apache.flink.ml.regression.MultipleLinearRegression
dataSet: org.apache.flink.api.scala.DataSet[org.apache.flink.ml.common.LabeledVector] = org.apache.flink.api.scala.DataSet@e37d9e6
trainTestData: org.apache.flink.ml.preprocessing.Splitter.TrainTestDataSet[org.apache.flink.ml.common.LabeledVector] = TrainTestDataSet(org.apache.flink.api.scala.DataSet@b0fa81b2,org.apache.flink.api.scala.DataSet@79f28121)
trainingData: org.apache.flink.api.scala.DataSet[org.apache.flink.ml.common.LabeledVector] = org.apache.flink.api.scala.DataSet@b0fa81b2

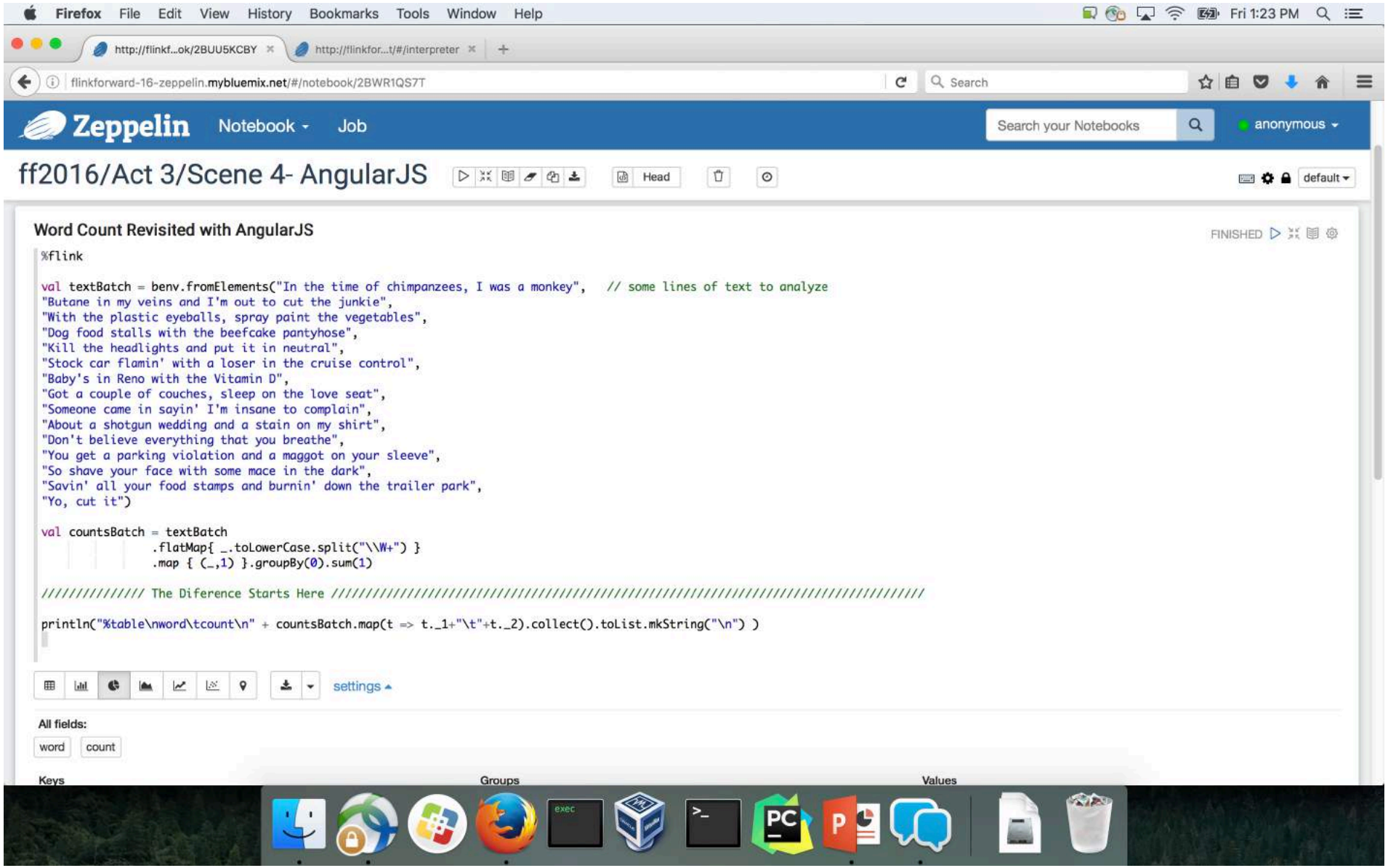

Firefox File Edit View History Bookmarks Tools Window Help
Fri 1:16 PM
http://flinkf...ok/2BUU5KCBY
http://flinkfor...t/#/interpreter
flinkforward-16-zeppelin.mybluemix.net/#/notebook/2BUG2S2G3
Search
Zeppelin
Notebook
Job
Search your Notebooks
anonymous
ff2016/Act 2/Scene 1- The "Data Scientist"
Head
default
val ssr = mlr.squaredResidualSum(testingData).collect()
println("----------\n")
println(ssr)
import org.apache.flink.ml.preprocessing.Splitter
import org.apache.flink.ml.math.Vector
import org.apache.flink.ml.regression.MultipleLinearRegression
dataSet: org.apache.flink.api.scala.DataSet[org.apache.flink.ml.common.LabeledVector] = org.apache.flink.api.scala.DataSet@e37d9e6
trainTestData: org.apache.flink.ml.preprocessing.Splitter.TrainTestDataSet[org.apache.flink.ml.common.LabeledVector] = TrainTestDataSet(org.apache.flink.api.scala.DataSet@b0fa81b2,org.apache.flink.api.scala.DataSet@79f28121)
trainingData: org.apache.flink.api.scala.DataSet[org.apache.flink.ml.common.LabeledVector] = org.apache.flink.api.scala.DataSet@b0fa81b2
testingData: org.apache.flink.api.scala.DataSet[org.apache.flink.ml.common.LabeledVector] = org.apache.flink.api.scala.DataSet@79f28121
mlr: org.apache.flink.ml.regression.MultipleLinearRegression = org.apache.flink.ml.regression.MultipleLinearRegression@54d4ab9e
predictions: org.apache.flink.api.scala.DataSet[(org.apache.flink.ml.math.Vector, Double)] = org.apache.flink.api.scala.DataSet@460a43e
ssr: Seq[Double] = Buffer(5.870895533464726E96)

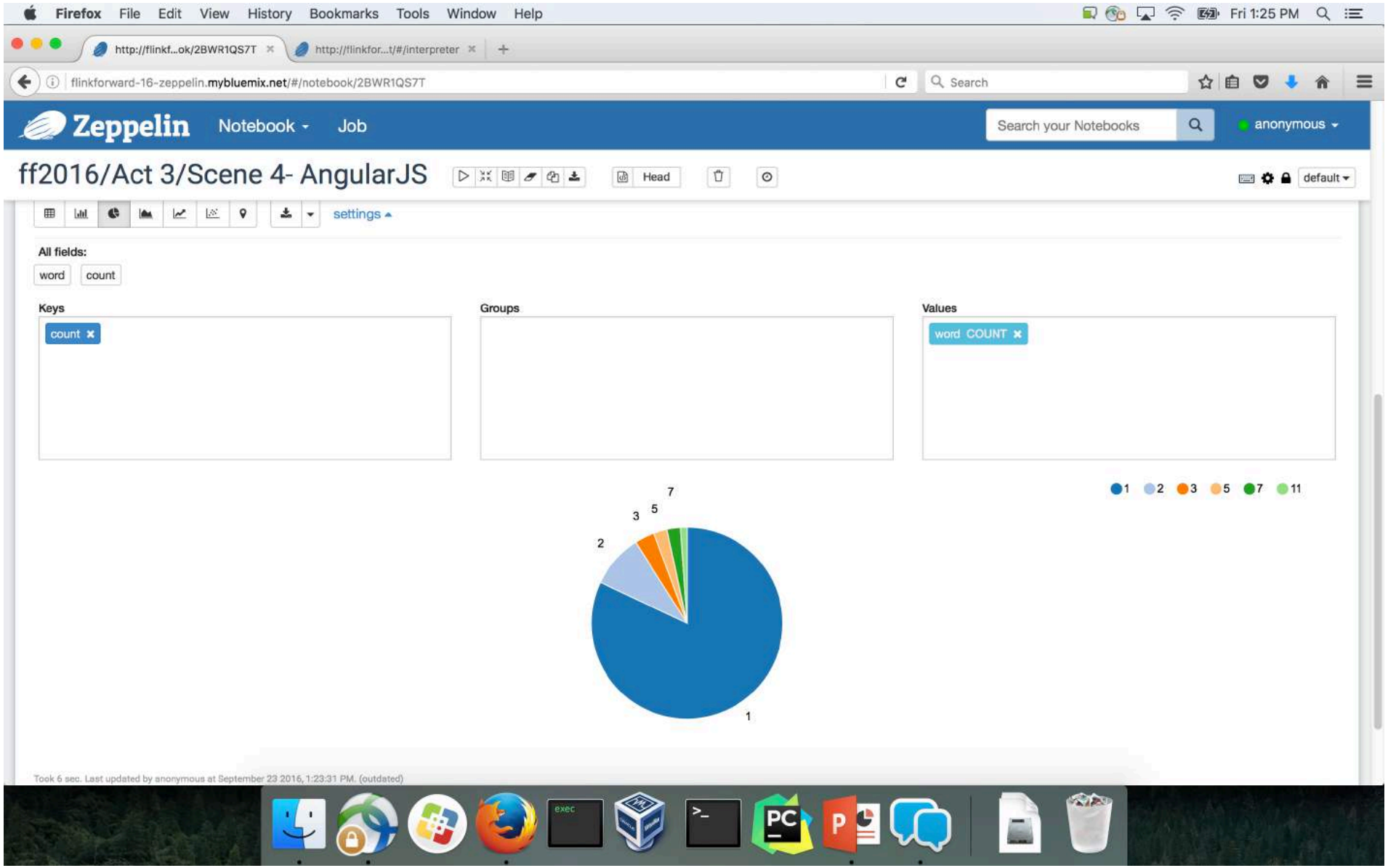
Buffer(5.870895533464726E96)
Took 8 sec. Last updated by anonymous at September 12 2016, 3:56:40 AM.
More Data Science
FINISHED
After getting the results, the user wants to tweak some of the parameters but leave a record of where she has been for other users.
She simply copy+paste the paragraph above, updating the step size, and compares results.
Took 0 sec. Last updated by anonymous at September 12 2016, 3:53:00 AM. (outdated)

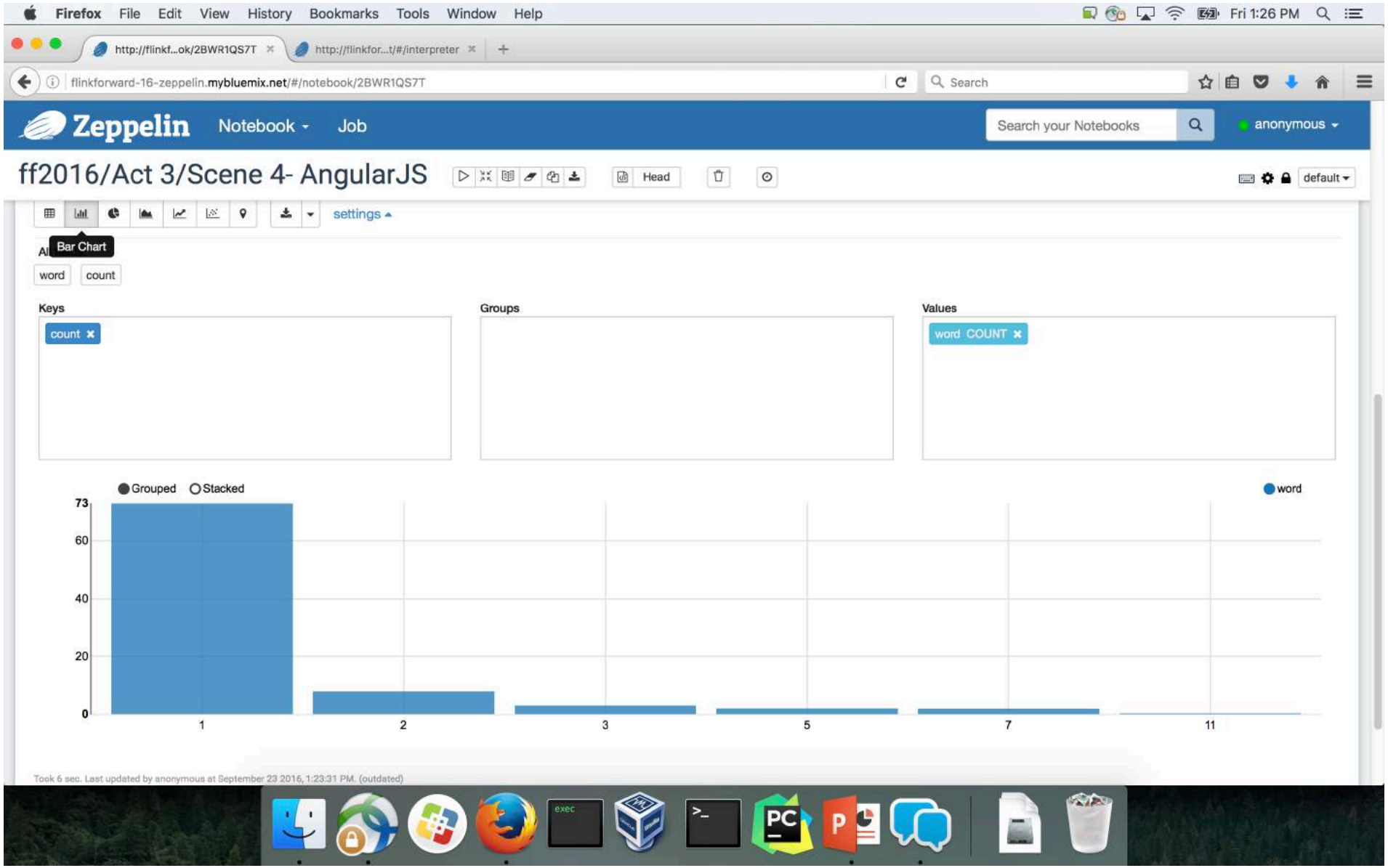

Firefox File Edit View History Bookmarks Tools Window Help
Fri 1:17 PM
http://flinkf...ok/2BUU5KCBY
http://flinkfor...t/#/interpreter
flinkforward-16-zeppelin.mybluemix.net/#/notebook/2BUG2S2G3
Search
Zeppelin
Notebook
Job
Search your Notebooks
anonymous
ff2016/Act 2/Scene 1- The "Data Scientist"
Head
default
leave a record of where she has been for other users.
She simply copy+paste the paragraph above, updating the step size, and compares results.
Took 0 sec. Last updated by anonymous at September 12 2016, 3:53:00 AM. (outdated)
%flinkML
FINISHED
val mlr2 = MultipleLinearRegression()
.setStepsize(0.01)
.setIterations(20)
.setConvergenceThreshold(0.001)
mlr2.fit(trainingData)
// The fitted model can now be used to make predictions
val predictions: DataSet[(Vector, Double)] = mlr.predict(testingData.map(lv => lv.vector))
val ssr2 = mlr2.squaredResidualSum(testingData).collect()
println("----------\n")
println("old ssr: "+ ssr)
println("new ssr: " + ssr2)
mlr2: org.apache.flink.ml.regression.MultipleLinearRegression = org.apache.flink.ml.regression.MultipleLinearRegression@a7fef275
predictions: org.apache.flink.api.scala.DataSet[(org.apache.flink.ml.math.Vector, Double)] = org.apache.flink.api.scala.DataSet@d4806a12
ssr2: Seq[Double] = Buffer(1.0968006862519498E56)
old ssr: Buffer(5.870895533464726E96)
new ssr: Buffer(1.0968006862519498E56)
Took 6 sec. Last updated by anonymous at September 12 2016, 3:57:03 AM.

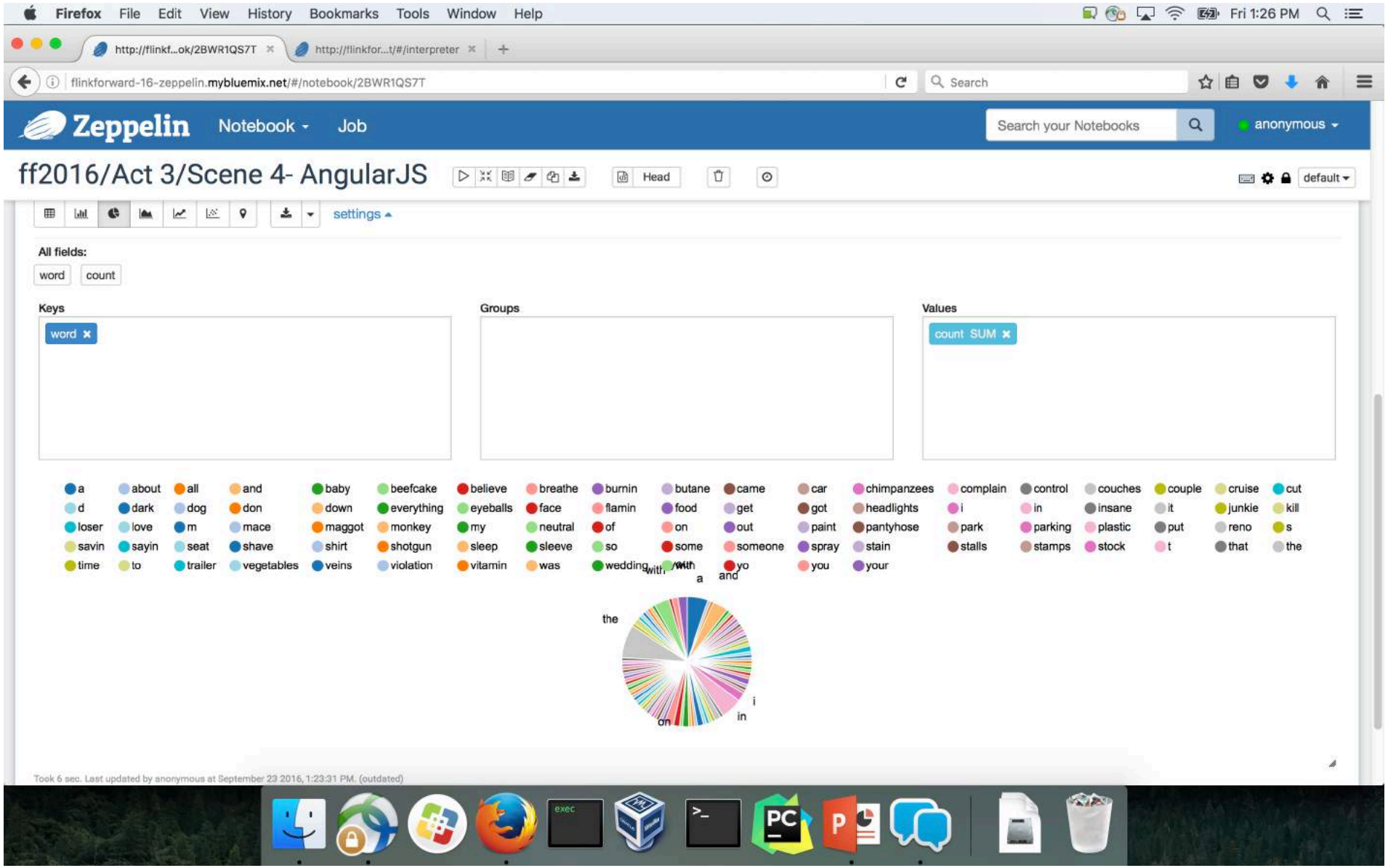

Firefox File Edit View History Bookmarks Tools Window Help
Fri 1:20 PM
http://flinkf...ok/2BUU5KCBY
http://flinkfor...t/#/interpreter
flinkforward-16-zeppelin.mybluemix.net/#/notebook/2BUWX5G98
Search
Zeppelin
Notebook
Job
Search your Notebooks
anonymous
ff2016/Act 3/Scene 3- Flink Streaming
Head
default
Took 1 sec. Last updated by anonymous at September 12 2016, 9:40:58 PM.
Wrap Up
FINISHED
This is another example of something that could have been achieved in a number of ways. The striking things here:
• We can create streaming apps and visualize them side by side
• We can 'play' with the visualizatoins
• We can live demo this to potential clients and friends
Took 0 sec. Last updated by anonymous at September 12 2016, 4:03:58 AM. (outdated)
READY
%md

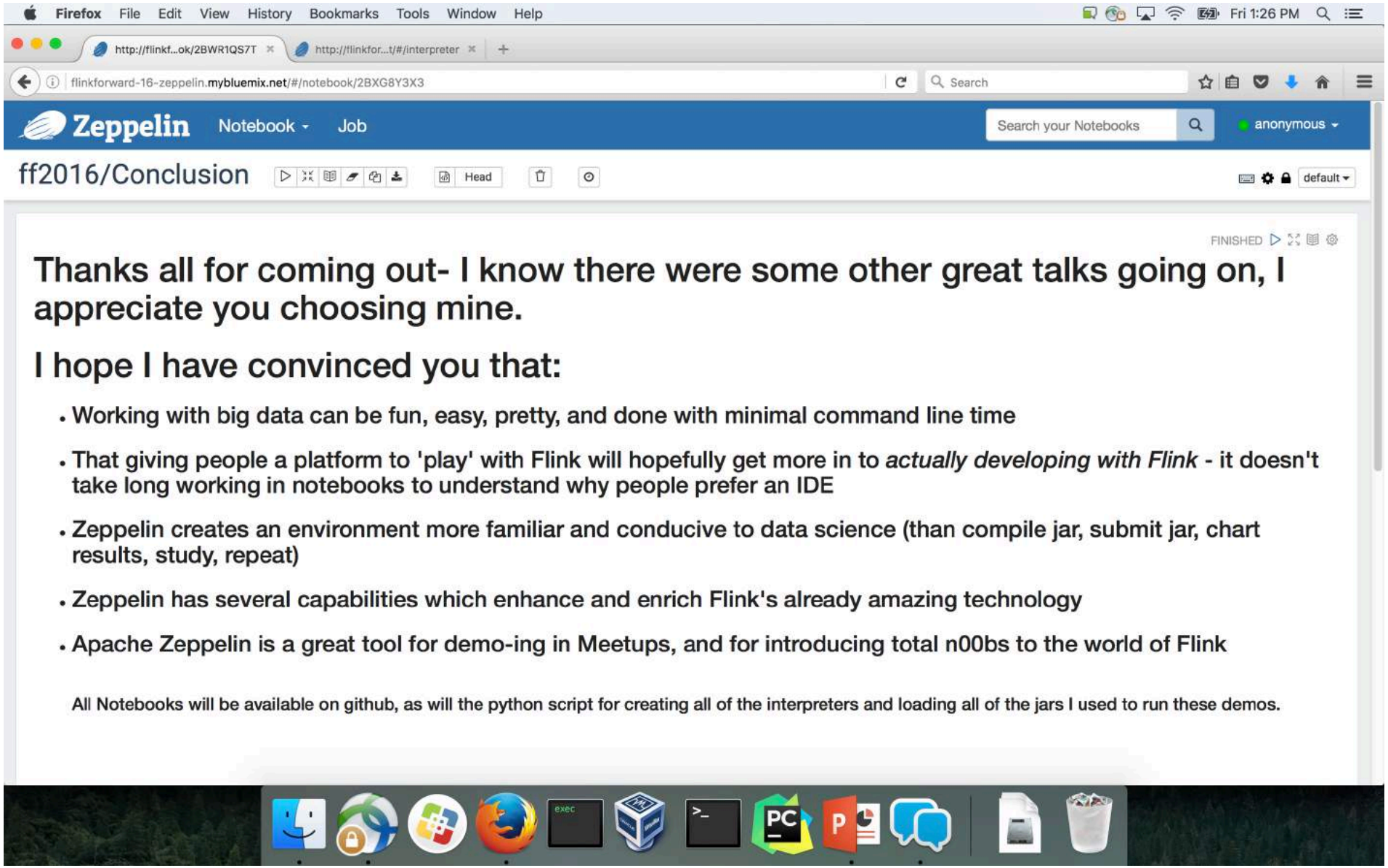

Firefox File Edit View History Bookmarks Tools Window Help
Fri 1:21 PM
http://flinkf...ok/2BUU5KCBY
http://flinkfor...t/#/interpreter
flinkforward-16-zeppelin.mybluemix.net/#/notebook/2BWR1QS7T
Search
Zeppelin
Notebook
Job
Search your Notebooks
anonymous
ff2016/Act 3/Scene 4- AngularJS
Head
default
Word Count Revisited with AngularJS
ERROR
%flink
val textBatch = senv.fromElements("In the time of chimpanzees, I was a monkey", // some lines of text to analyze
"Butane in my veins and I'm out to cut the junkie",
"With the plastic eyeballs, spray paint the vegetables",
"Dog food stalls with the beefcake pantyhose",
"Kill the headlights and put it in neutral",
"Stock car flamin' with a loser in the cruise control",
"Baby's in Reno with the Vitamin D",
"Got a couple of couches, sleep on the love seat",
"Someone came in sayin' I'm insane to complain",
"About a shotgun wedding and a stain on my shirt",
"Don't believe everything that you breathe",
"You get a parking violation and a maggot on your sleeve",
"So shave your face with some mace in the dark",
"Savin' all your food stamps and burnin' down the trailer park",
"Yo, cut it")
val countsBatch = textBatch
.flatMap{ _.toLowerCase.split("\\W+") }
.map { (_,1) }.groupBy(0).sum(1)
/////////////// The Diference Starts Here //
println("%table\nword\tcount\n" + countsBatch.map(t => t._1+"\t"+t._2).collect().toList.mkString("\n"))
textBatch: org.apache.flink.streaming.api.scala.DataStream[String] = org.apache.flink.streaming.api.scala.DataStream@d1888e70
<console>:27: error: value groupBy is not a member of org.apache.flink.streaming.api.scala.DataStream[(String, Int)]
.map { (_,1) }.groupBy(0).sum(1)
^
Took 1 sec. Last updated by anonymous at September 12 2016, 9:42:16 PM.


Firefox File Edit View History Bookmarks Tools Window Help
Fri 1:23 PM
http://flinkf...ok/2BUU5KCBY
http://flinkfor...t/#/interpreter
flinkforward-16-zeppelin.mybluemix.net/#/notebook/2BWR1QS7T
Search
Zeppelin
Notebook
Job
Search your Notebooks
anonymous
ff2016/Act 3/Scene 4- AngularJS
Head
default
Word Count Revisited with AngularJS
FINISHED
%flink
val textBatch = benv.fromElements("In the time of chimpanzees, I was a monkey", // some lines of text to analyze
"Butane in my veins and I'm out to cut the junkie",
"With the plastic eyeballs, spray paint the vegetables",
"Dog food stalls with the beefcake pantyhose",
"Kill the headlights and put it in neutral",
"Stock car flamin' with a loser in the cruise control",
"Baby's in Reno with the Vitamin D",
"Got a couple of couches, sleep on the love seat",
"Someone came in sayin' I'm insane to complain",
"About a shotgun wedding and a stain on my shirt",
"Don't believe everything that you breathe",
"You get a parking violation and a maggot on your sleeve",
"So shave your face with some mace in the dark",
"Savin' all your food stamps and burnin' down the trailer park",
"Yo, cut it")
val countsBatch = textBatch
.flatMap{ _.toLowerCase.split("\\W+") }
.map { (_,1) }.groupBy(0).sum(1)
/////////////// The Diference Starts Here //
println("%table\nword\tcount\n" + countsBatch.map(t => t._1+"\t"+t._2).collect().toList.mkString("\n"))
settings
All fields:
word
count
Keys
Groups
Values


Firefox File Edit View History Bookmarks Tools Window Help
Fri 1:25 PM
http://flinkf...ok/2BWR1QS7T
http://flinkfor...t/#/interpreter
flinkforward-16-zeppelin.mybluemix.net/#/notebook/2BWR1QS7T
Search
Zeppelin
Notebook
Job
Search your Notebooks
anonymous
ff2016/Act 3/Scene 4- AngularJS
Head
default
settings
All fields:
word
count
Keys
count
Groups
Values
word COUNT
1 2 3 5 7 11
7
5
3
2
1
Took 6 sec. Last updated by anonymous at September 23 2016, 1:23:31 PM. (outdated)


Firefox File Edit View History Bookmarks Tools Window Help
Fri 1:26 PM
http://flinkf...ok/2BWR1QS7T
http://flinkfor...t/#/interpreter
flinkforward-16-zeppelin.mybluemix.net/#/notebook/2BWR1QS7T
Search
Zeppelin
Notebook
Job
Search your Notebooks
anonymous
ff2016/Act 3/Scene 4- AngularJS
Head
default
settings
Bar Chart
word
count
Keys
count
Groups
Values
word COUNT
Grouped
Stacked
word
73
60
40
20
0
1
2
3
5
7
11
Took 6 sec. Last updated by anonymous at September 23 2016, 1:23:31 PM. (outdated)


Firefox File Edit View History Bookmarks Tools Window Help
Fri 1:26 PM
http://flinkf...ok/2BWR1QS7T
http://flinkfor...t/#/interpreter
flinkforward-16-zeppelin.mybluemix.net/#/notebook/2BWR1QS7T
Search
Zeppelin
Notebook
Job
Search your Notebooks
anonymous
ff2016/Act 3/Scene 4- AngularJS
Head
default
settings
All fields:
word
count
Keys
word
Groups
Values
count SUM
a about all and baby beefcake believe breathe burnin butane came car chimpanzees complain control couches couple cruise cut
d dark dog don down everything eyeballs face flamin food get got headlights i in insane it junkie kill
loser love m mace maggot monkey my neutral of on out paint pantyhose park parking plastic put reno s
savin sayin seat shave shirt shotgun sleep sleeve so some someone spray stain stalls stamps stock t that the
time to trailer vegetables veins violation vitamin was wedding with yo you your
the
a
and
i
in
on
Took 6 sec. Last updated by anonymous at September 23 2016, 1:23:31 PM. (outdated)


Firefox File Edit View History Bookmarks Tools Window Help
Fri 1:26 PM
http://flinkf...ok/2BWR1QS7T
http://flinkfor...t/#/interpreter
flinkforward-16-zeppelin.mybluemix.net/#/notebook/2BXG8Y3X3
Search
Zeppelin
Notebook
Job
Search your Notebooks
anonymous
ff2016/Conclusion
Head
default
FINISHED
Thanks all for coming out- I know there were some other great talks going on, I appreciate you choosing mine.
I hope I have convinced you that:
• Working with big data can be fun, easy, pretty, and done with minimal command line time
• That giving people a platform to 'play' with Flink will hopefully get more in to actually developing with Flink - it doesn't take long working in notebooks to understand why people prefer an IDE
• Zeppelin creates an environment more familiar and conducive to data science (than compile jar, submit jar, chart results, study, repeat)
• Zeppelin has several capabilities which enhance and enrich Flink's already amazing technology
• Apache Zeppelin is a great tool for demo-ing in Meetups, and for introducing total n00bs to the world of Flink
All Notebooks will be available on github, as will the python script for creating all of the interpreters and loading all of the jars I used to run these demos.

Firefox File Edit View History Bookmarks Tools Window Help Fri 1:26 PM

http://flinkf...ok/2BWR1QS7T | http://flinkfor...t/#/interpreter

flinkforward-16-zeppelin.mybluemix.net/#/notebook/2BXG8Y3X3

Search

Zeppelin Notebook Job Search your Notebooks anonymous

ff2016/Conclusion Head default

Q & A

"And she's climbing the stairway to heaven…"

Took 0 sec. Last updated by anonymous at September 12 2016, 7:59:39 AM. (outdated)

%sh

READY