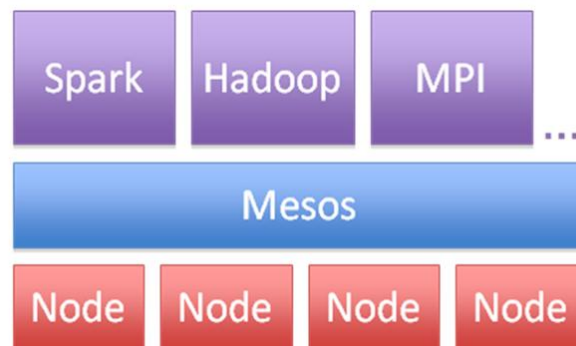
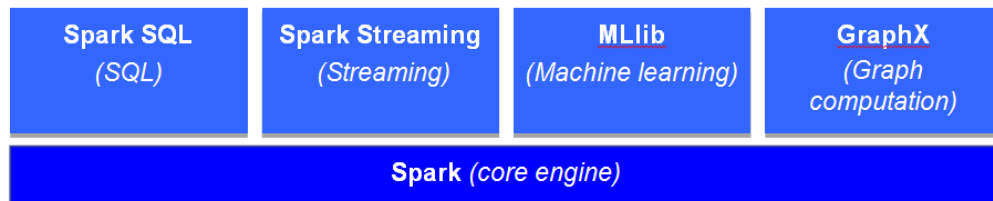
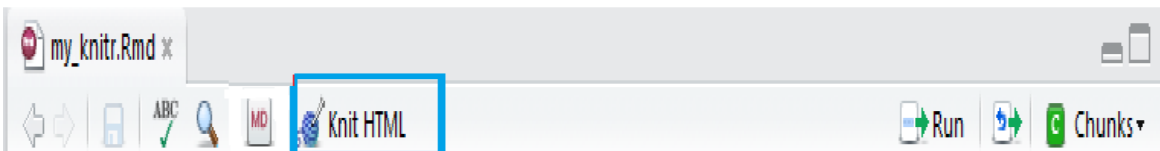
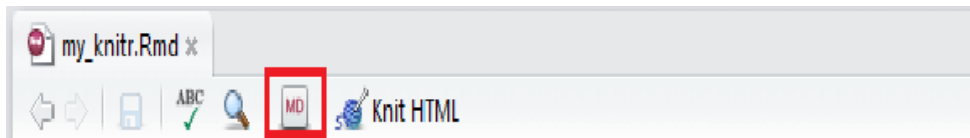
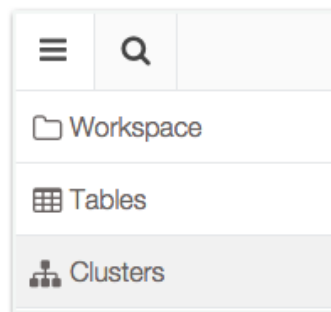
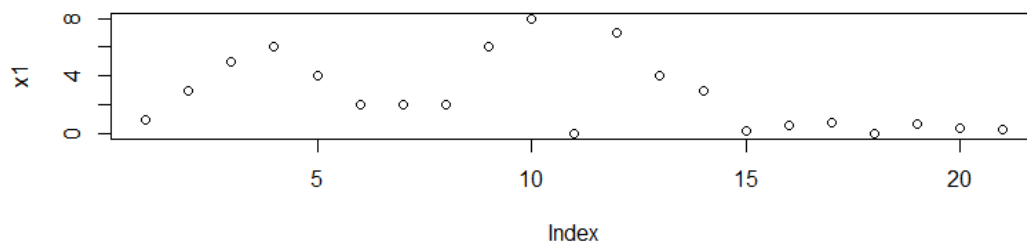


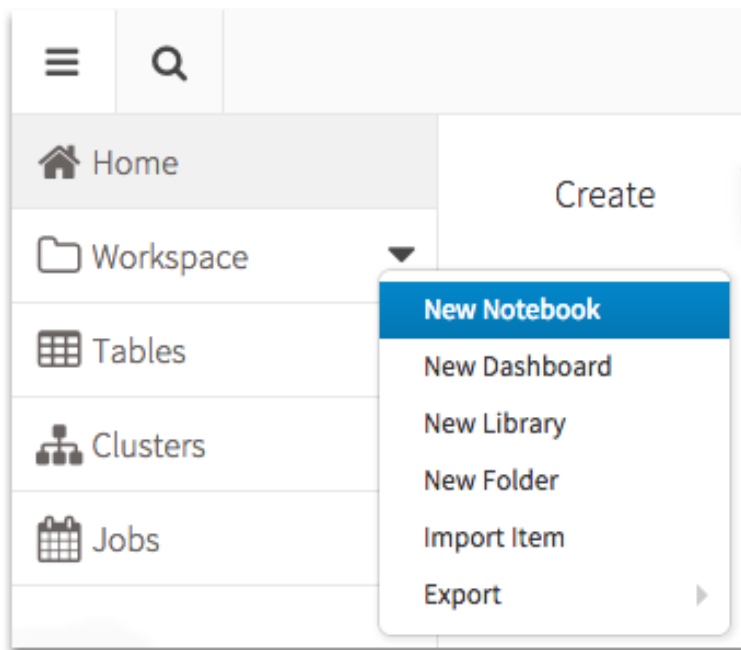
## Chapter 1: Spark for Machine Learning





## Chapter 2: Data Preparation for Spark ML





## Create Notebook

Name

Language 

Python

Scala

SQL

✓ R

Cluster

### Airlines Job

Task: [Set Notebook / Set JAR](#)

Cluster: 60 GB On-demand, Spark 1.4 (preview) [Edit](#)

Schedule: None [Edit](#)

Advanced ▶

#### Active runs

Run	Start Time
No active runs. <a href="#">Run Now</a>	

#### Completed runs

Latest successful run (refreshes automatically) • [View as a dashboard](#)

Previous 20

Run	Start Time
Previous 20	

### Select Notebook to Run

de

eld

hhd

hossein

ion

jeffpang

John

Joseph

kyle

lucian

matei

.D-moved...

.D-renam...

OT UsageLog Analysis

Airlines

Customer Analysis

Dashboards

etc

Investigatinos

Staging Pipeline

Tests

Wall display

airlines

Cancel

OK

## Chapter 3: A Holistic View on Spark

Q

Home

Workspace

Tables

Clusters

Jobs

Accounts

Q

### Table Import

Data Source

S3

AWS Key ID

S3

Secret Access Key

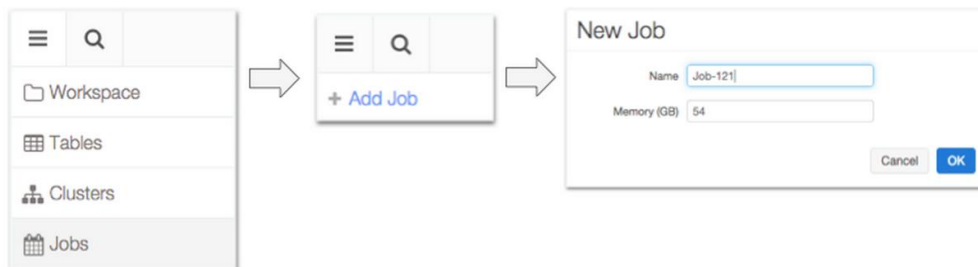
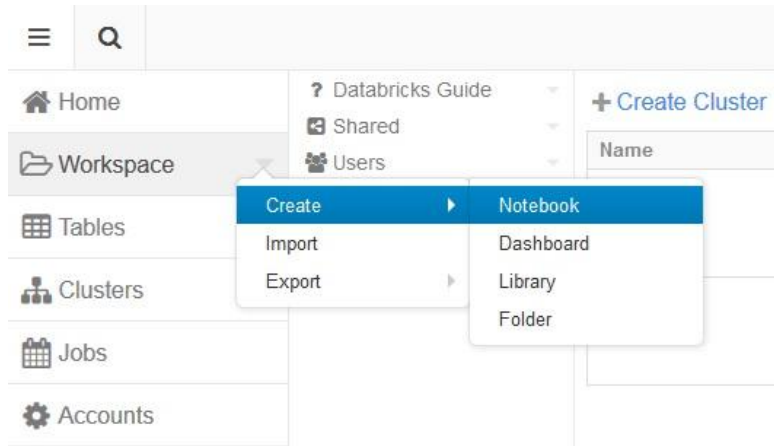
DBFS

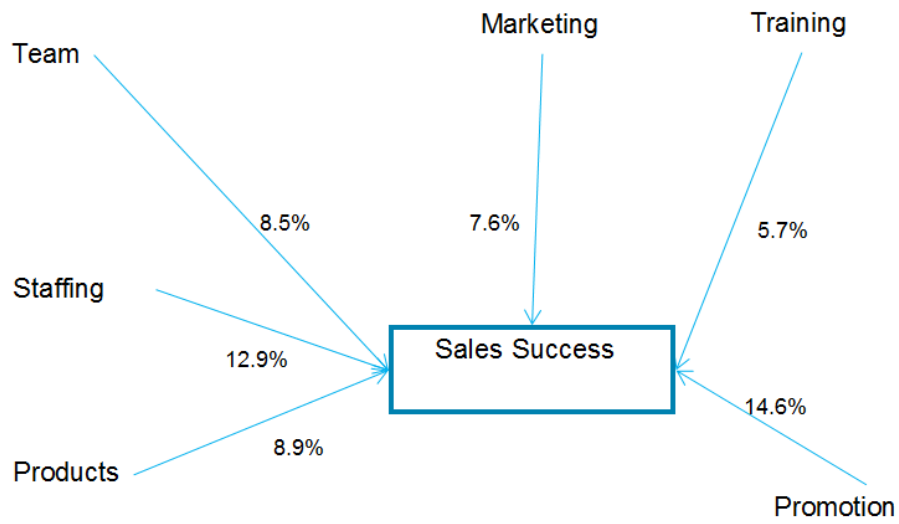
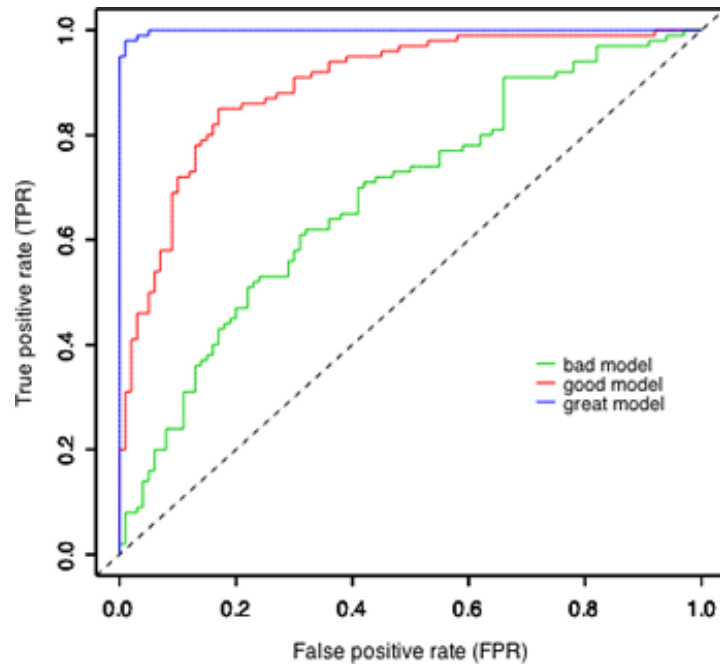
S3 Bucket Name

JDBC

File

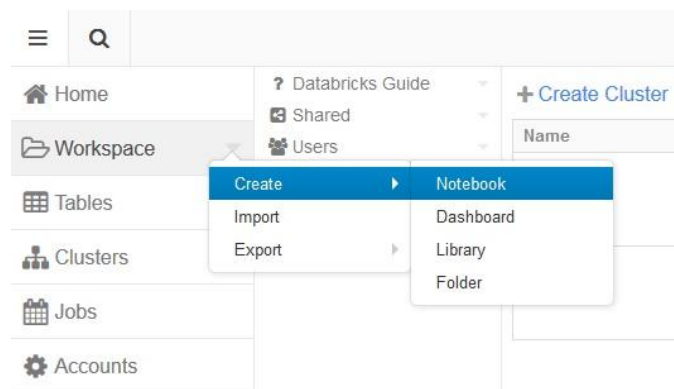
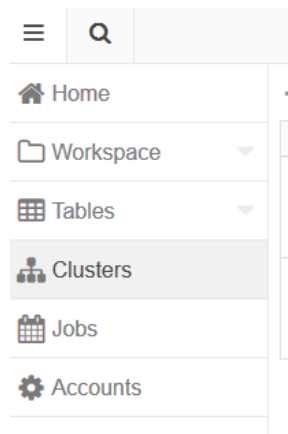
Browse Bucket

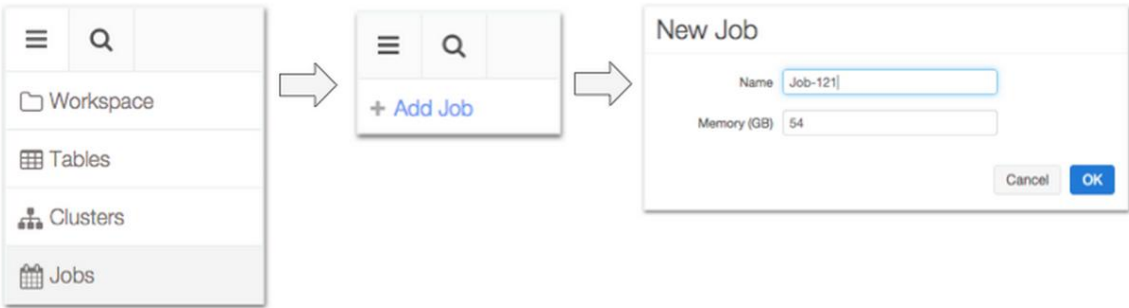






## Chapter 4: Fraud Detection on Spark






## Chapter 5: Risk Scoring on Spark



The banner features a blue header with the Big Data University logo and name. Below this is a blue background with a white network diagram. A white box in the center contains the 'Data Scientist Workbench' logo and tagline. The bottom section is white and displays logos for R, Python, Scala, and Spark. A URL is provided at the bottom, and a footer contains navigation icons, copyright information, and a Twitter handle.

**BIG DATA UNIVERSITY**

 **Data Scientist Workbench**  
Prepare data. Analyze data. Get answers.

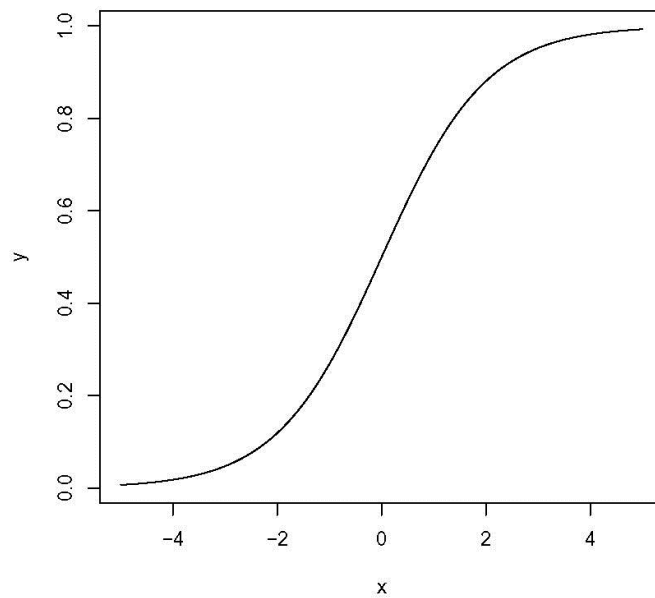
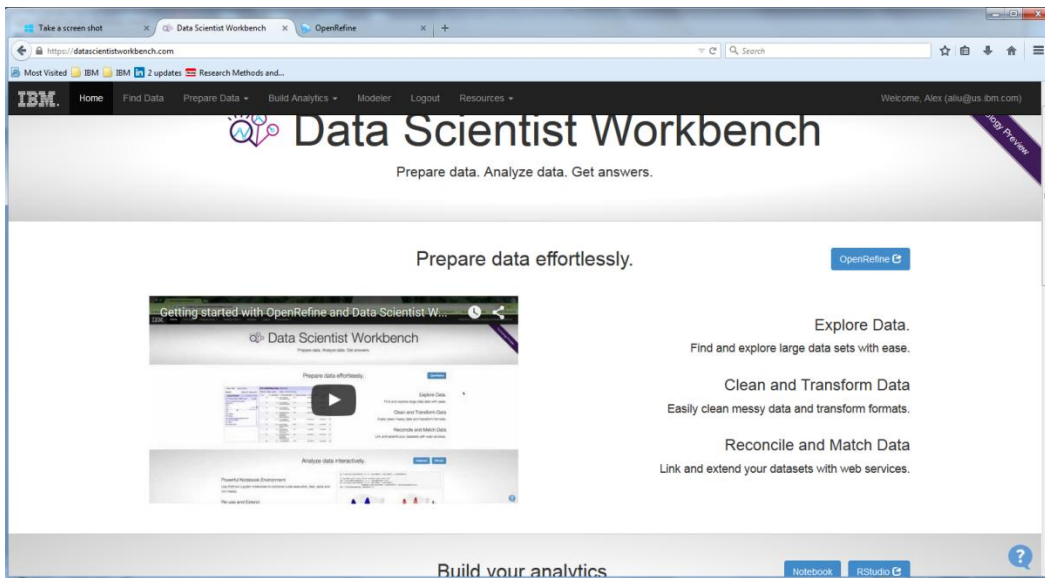
  **python**<sup>TM</sup>

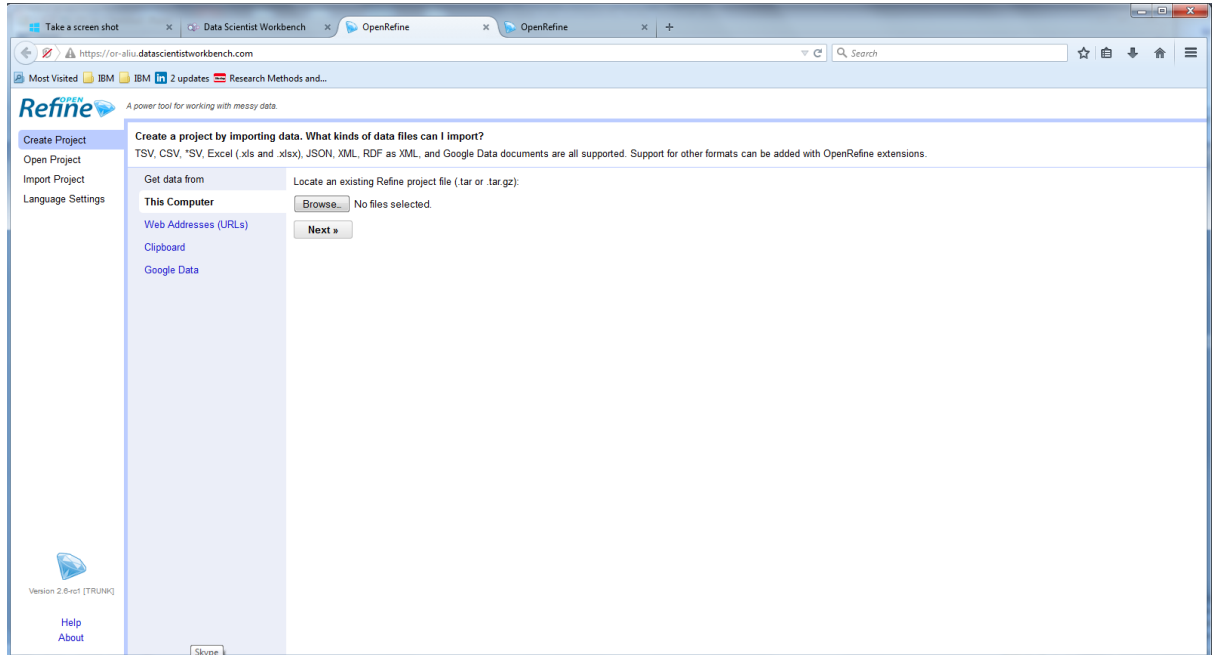
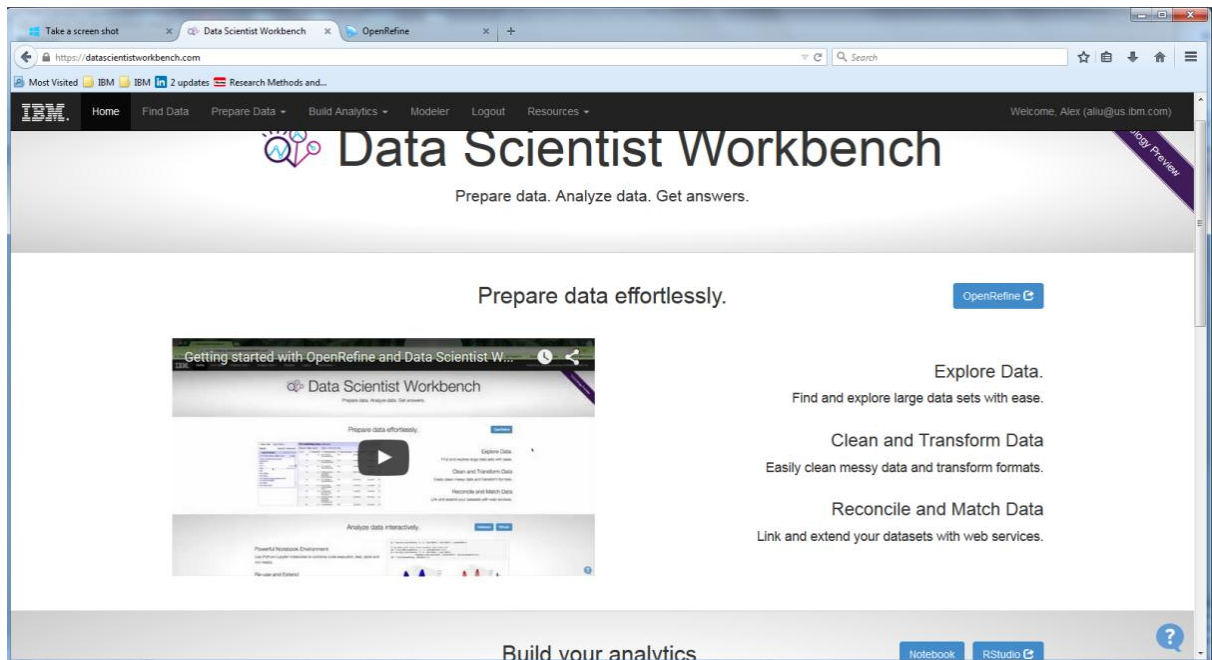
 **Scala**

**Spark** 

<https://datascientistworkbench.com>

← ↻ ⓘ 2015 BigDataUniversity.com  @bigdatau





Take a screen shot

Data Scientist Workbench

OpenRefine

OpenRefine

https://datascientistworkbench.com

Most Visited IBM IBM 2 updates Research Methods and...

IBM Home Find Data Prepare Data Build Analytics Modeler Logout Resources

Welcome, Alex (alex@us.ibm.com)

# Build your analytics.

NotebookRStudio

## Notebooks

Use notebooks to combine code execution, text, plots and rich media. Use preinstalled Python, Scala and R libraries. Install others as needed.

## RStudio

Use your favorite IDE to build, debug and test your R code.

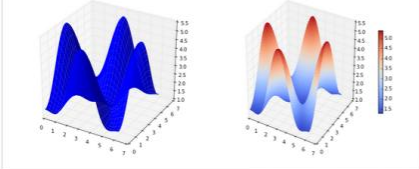
## Analytics at Scale

Submit your analytic jobs to large-scale Hadoop, Spark, BigR, BigSQL, and dashDB clusters.

## Collaborate and Share

Build on what others have done and easily share your analysis.

```
p = mx.plot_surface(x, y, z, rstride=1, cstride=1, linewidth=0)
# surface plot with color grading and color bar
ax = fig.add_subplot(1, 2, 2, projection='3d')
p = mx.plot_surface(x, y, z, rstride=1, cstride=1,
                    cmap=plt.cm.coolwarm, linewidth=0, antialiased=False)
cb = fig.colorbar(p, shrink=0.5)
```



Get answers quickly.

?

Take a screen shot Data Scientist Workbench OpenRefine OpenRefine

https://datascientistworkbench.com

Most Visited IBM IBM 2 updates Research Methods and...

IBM Home Find Data Prepare Data Build Analytics Modeler Logout Resources

Welcome, Alex (aliu@us.ibm.com)

Notebook RStudio

## Build your analytics.

Notebook RStudio

### Notebooks

Use notebooks to combine code execution, text, plots and rich media. Use preinstalled Python, Scala and R libraries. Install others as needed.

### RStudio

Use your favorite IDE to build, debug and test your R code.

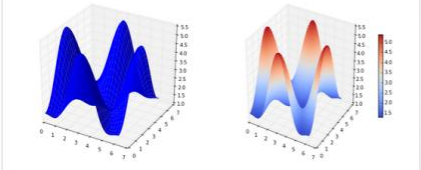
### Analytics at Scale

Submit your analytic jobs to large-scale Hadoop, Spark, BigR, BigSQL, and dashDB clusters.

### Collaborate and Share

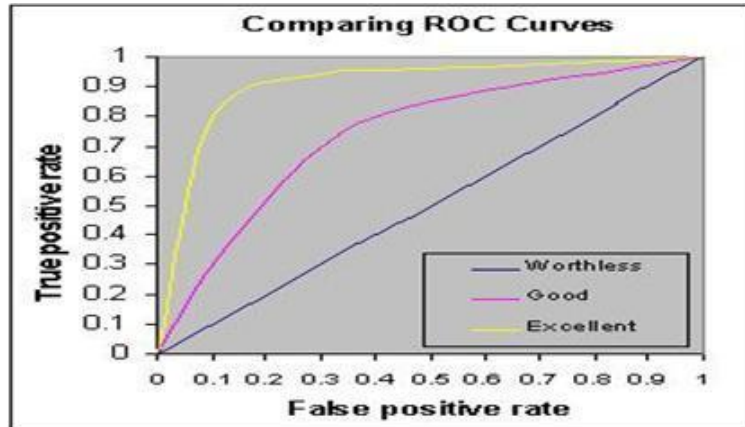
Build on what others have done and easily share your analysis.

```
p = aa.plot_surface(x, y, z, rstride=1, cstride=1, linewidth=0)
# surface plot with color grading and color bar
ax = fig.add_subplot(1, 2, 1, projection='3d')
p = aa.plot_surface(x, y, z, rstride=1, cstride=1,
                    cmap=cm.coolwarm, linewidth=0, antialiased=False)
cb = fig.colorbar(p, shrink=0.5)
```



Get answers quickly.

https://datascientistworkbench.com/workbench



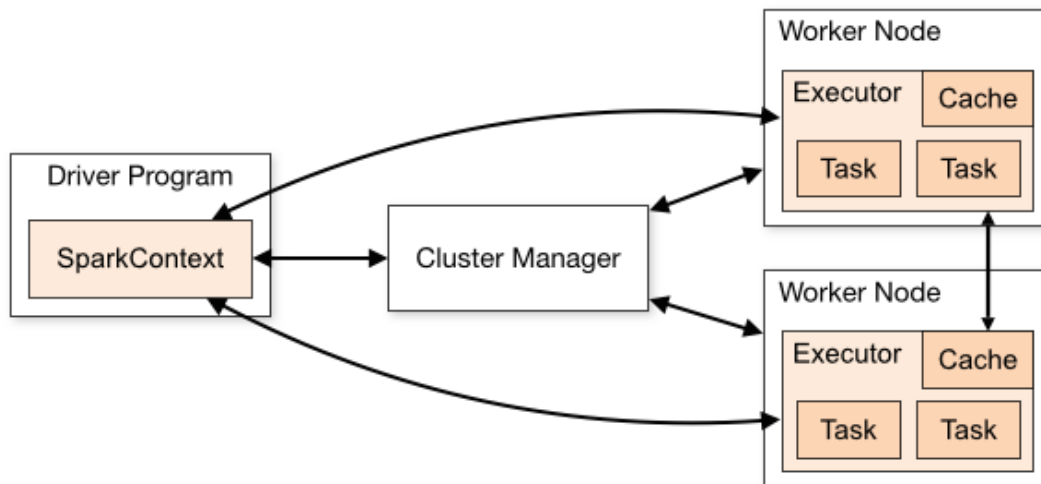
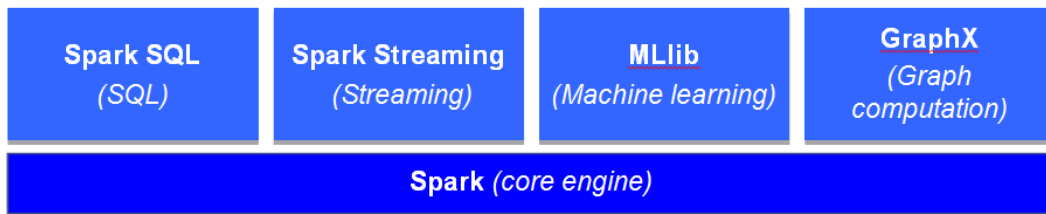
$$\ln(\frac{P}{1-P}) = a + bX$$

$$\frac{P}{1-P} = e^{a+bX}$$

$$P = \frac{e^{a+bX}}{1+e^{a+bX}}$$



## Chapter 6: Churn Prediction on Spark

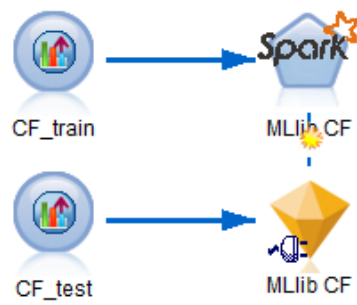
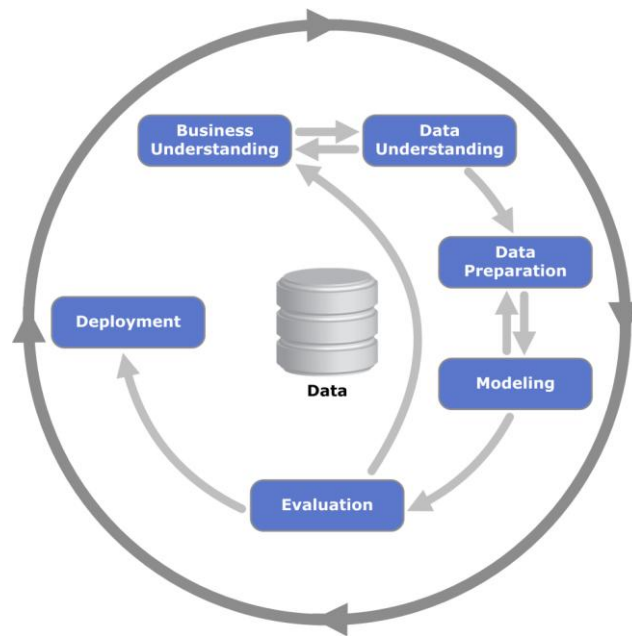


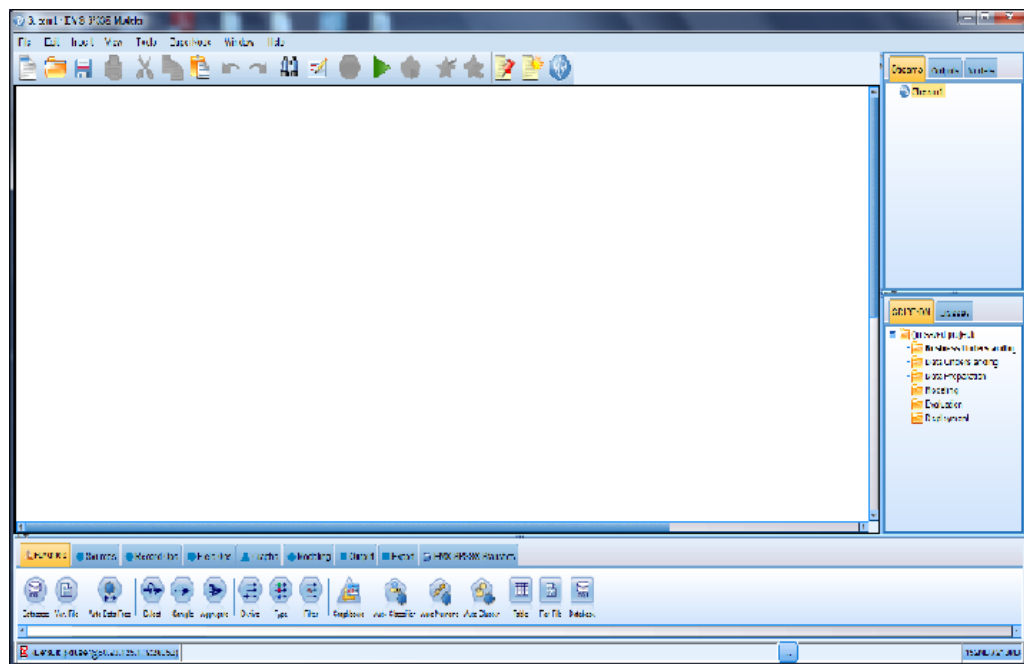
$$\ln\left(\frac{P}{1-P}\right) = a + bX$$

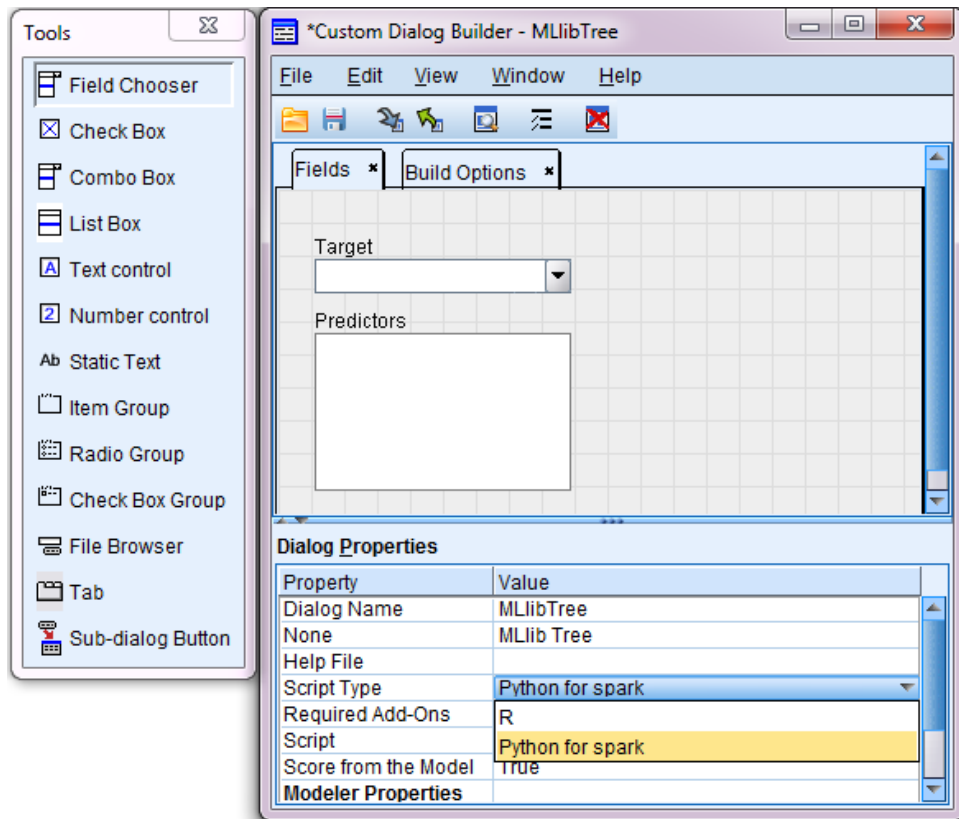
$$\frac{P}{1-P} = e^{a+bX}$$

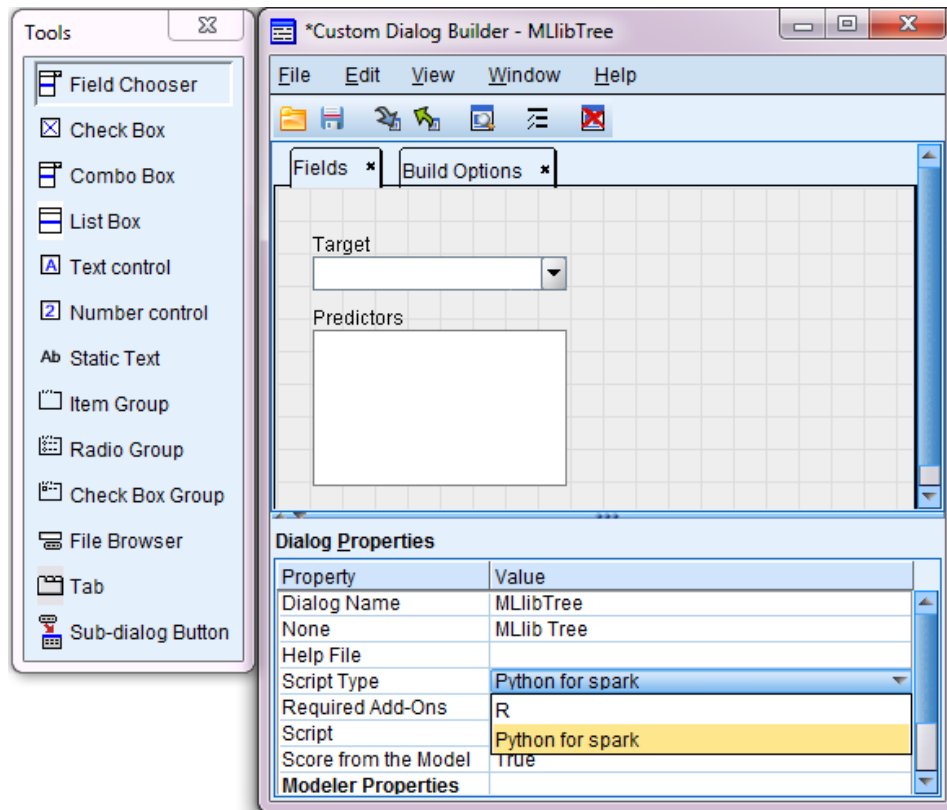
$$P = \frac{e^{a+bX}}{1+e^{a+bX}}$$

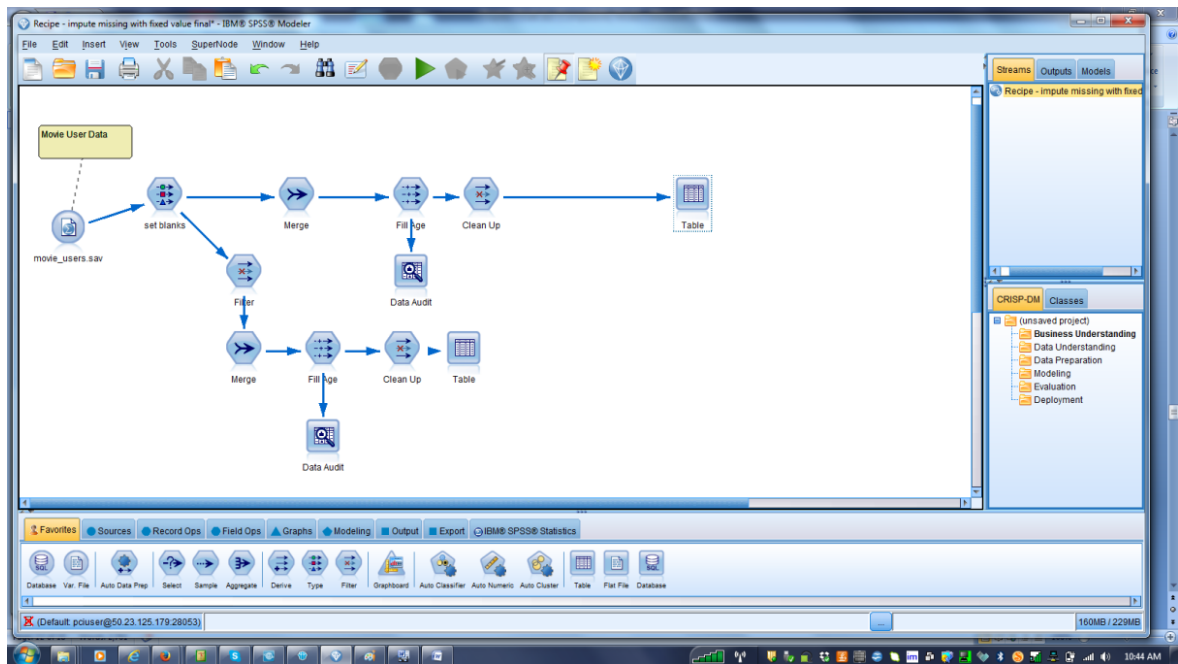
## Chapter 7: Recommendations on Spark











Server Login

Select, add or edit the server to which you want to connect in the table. By default, the connection marked as Default is used at start-up.

Default	Server Name	Description	Port
<input type="checkbox"/>	Local Server		
<input checked="" type="checkbox"/>	50.20.125.300		28053

Add...

Edit...

Delete

Search...

Default data path:

\$MODELERSERVER/data

...

☒ Set Credentials

User ID:

rec-user

Password:

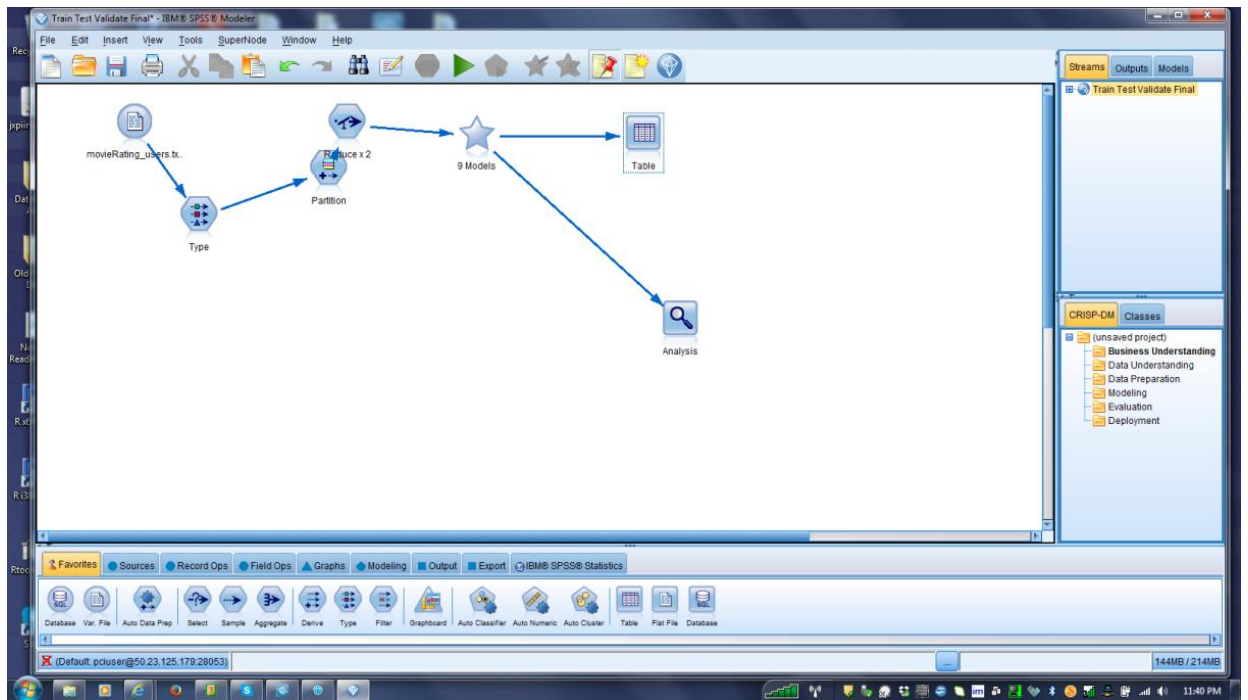
Domain:

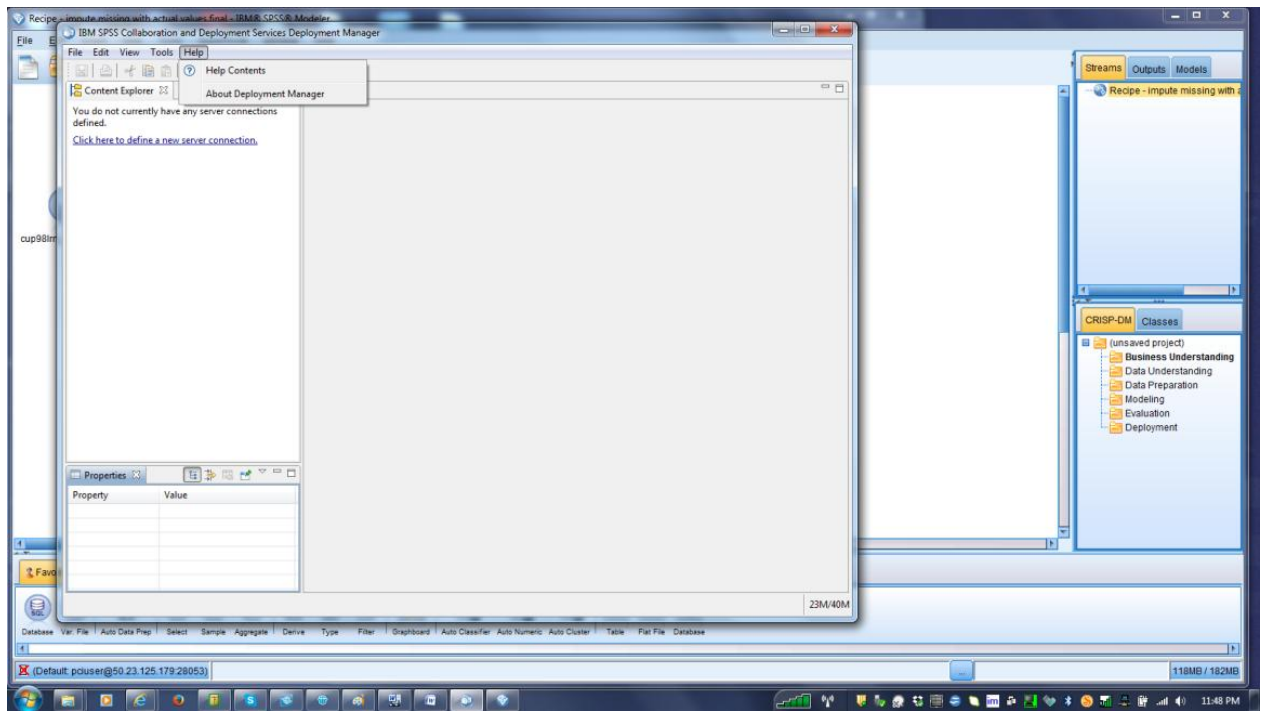
OK

Cancel

Help







## Chapter 8: Learning Analytics on Spark



The screenshot shows the Zeppelin web interface. At the top, there is a blue header bar with the Zeppelin logo, "Notebook", "Interpreter", and a "Connected" status indicator. Below the header, the main content area has a "Welcome to Zeppelin!" message, followed by a brief description of Zeppelin as a web-based notebook. To the left, there is a "Notebook" section with a "Create new note" button and a list of existing notebooks. To the right, there is a "Help" section with links to documentation and a "Community" section with a message and links to a mailing list, issues tracking, and GitHub. A large, stylized blue blimp illustration is on the right side of the page.

**Zeppelin** Notebook Interpreter Connected

### Welcome to Zeppelin!

Zeppelin is web-based notebook that enables interactive data analytics.  
You can make beautiful data-driven, interactive, collaborative document with SQL, code and even more!

#### Notebook

Create new note

- Air Pollution & CA Weather (Henry)
- DashDB Example
- Heating Complaints in Bronx
- Heating Problems in Manhattan
- Modeling Heating Problem in New York City
- Modeling Heating Problems in New York City - Part 2
- Rinterpreter
- Step By Step Predictive Modeling
- Stream Example
- Streaming Twitter
- test
- vtest1
- XML example
- Zeppelin Tutorial

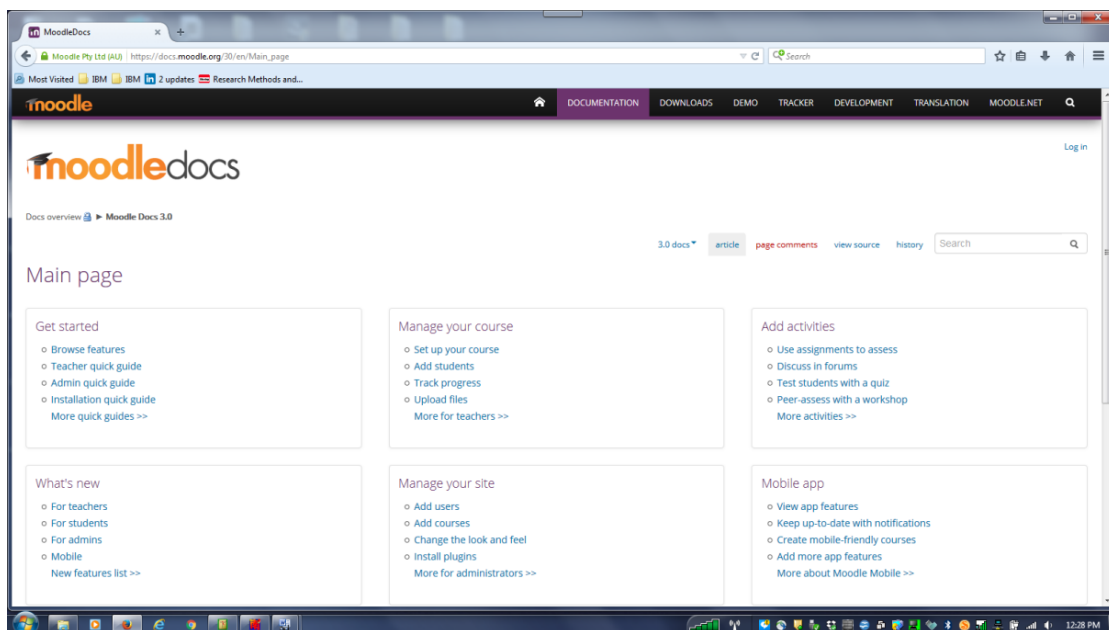
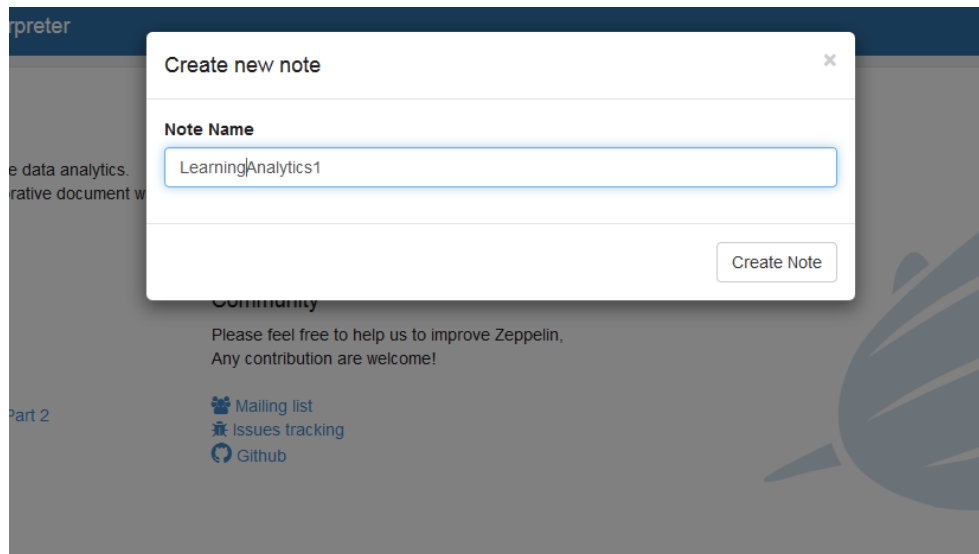
#### Help

Get started with Zeppelin documentation

#### Community

Please feel free to help us to improve Zeppelin.  
Any contribution are welcome!

- Mailing list
- Issues tracking
- GitHub



Notebook
Interpreter
Connected

Learning Analy
▶
🔍
📄
🗑️
👤
⚙️
🔗

🔍
⚙️
default

```

// Train a RandomForest model.

val treeStrategy = Strategy.defaultStrategy("Classification")
val numTrees = 300
val featureSubsetStrategy = "auto" // Let the algorithm choose.
val model = RandomForest.trainClassifier(trainingData,
  treeStrategy, numTrees, featureSubsetStrategy, seed = 12345)

import org.apache.commons.io.IOUtils
import java.net.URL
import java.nio.charset.Charset
bankText: org.apache.spark.rdd.RDD[String] = ParallelCollectionRDD[0] at parallelize at <console>:31
defined class Bank
bank: org.apache.spark.sql.DataFrame = [age: int, job: string, marital: string, education: string, balance: int]
Took 20 seconds. (outdated)

```

Notebook
Interpreter
Connected

Learning Analy
▶
🔍
📄
🗑️
👤
⚙️
🔗

🔍
⚙️
default

Took 20 seconds. (outdated)

```

%sql
select age, count(1) value
from bank
where age < 30
group by age
order by age

```

📄
📊
📈
📉
📊
📈
📉
📊
📈
📉
settings

● Grouped
○ Stacked
value

Took 4 seconds.

FINISHED ▶ 🔍 📄 🗑️ 👤 ⚙️ 🔗

```

%sql
select age, count(1) value
from bank
where age < ${maxAge:30}
group by age
order by age

```

maxAge

📄
📊
📈
📉
📊
📈
📉
📊
📈
📉
settings

● Grouped
○ Stacked
value

Took 1 seconds.

FINISHED ▶ 🔍 📄 🗑️ 👤 ⚙️ 🔗

```

%sql
select age, count(1) value
from bank
where marital=${marital:single,single|divorced|married}
group by age
order by age


```

marital

📄
📊
📈
📉
📊
📈
📉
📊
📈
📉
settings

● Grouped
○ Stacked
value





Took 1 seconds.

 **Zeppelin**

Notebook - Interpreter

Connected

Learning Analyt

default

```
// Train a RandomForest model.

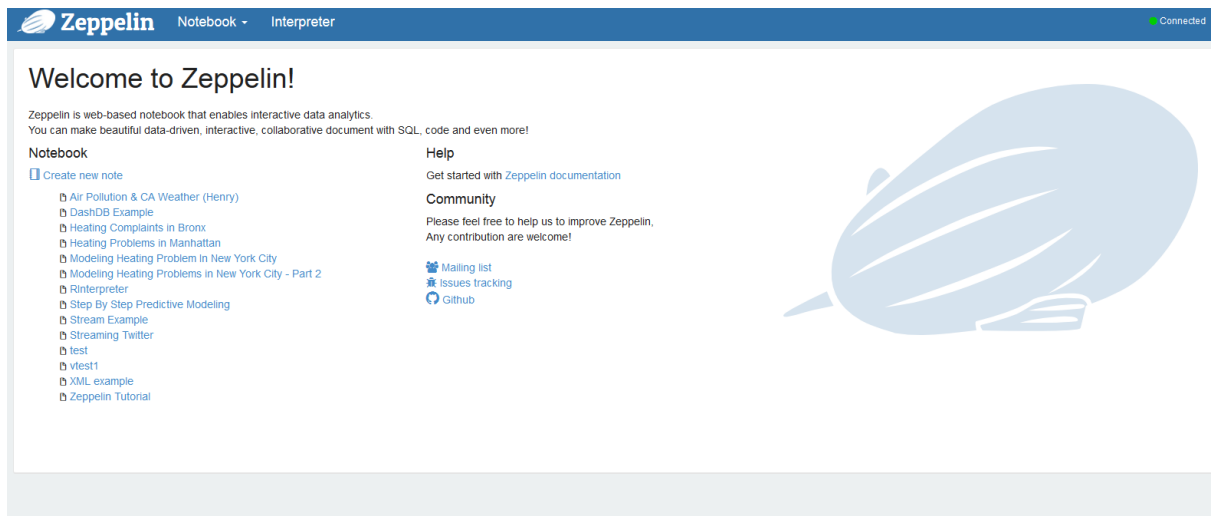
val treeStrategy = Strategy.defaultStrategy("Classification")
val numTrees = 300
val featureSubsetStrategy = "auto" // Let the algorithm choose.
val model = RandomForest.trainClassifier(trainingData,
  treeStrategy, numTrees, featureSubsetStrategy, seed = 12345)

import org.apache.commons.io.IOUtils
import java.net.URL
import java.nio.charset.Charset
bankText: org.apache.spark.rdd.RDD[String] = ParallelCollectionRDD[0] at parallelize at <console>:31
defined class Bank
bank: org.apache.spark.sql.DataFrame = [age: int, job: string, marital: string, education: string, balance: int]

Took 20 seconds (outdated)
```

$$f(z) = \frac{1}{1 + e^{-z}}$$

## Chapter 9: City Analytics on Spark



The image shows the Zeppelin Notebook web interface. At the top, there's a blue header with the Zeppelin logo, "Notebook" and "Interpreter" tabs, and a "Connected" status indicator. The main content area has a "Welcome to Zeppelin!" heading, followed by a brief description of Zeppelin as a web-based notebook. Below this, there are three sections: "Notebook" with a "Create new note" button and a list of example notebooks; "Help" with a link to documentation; and "Community" with links to a mailing list, issues tracking, and GitHub. A large, stylized blue blimp illustration is on the right side of the page.

**Zeppelin** Notebook Interpreter Connected

### Welcome to Zeppelin!

Zeppelin is web-based notebook that enables interactive data analytics.  
You can make beautiful data-driven, interactive, collaborative document with SQL, code and even more!

**Notebook**

Create new note

- Air Pollution & CA Weather (Henry)
- DashDB Example
- Heating Complaints in Bronx
- Heating Problems in Manhattan
- Modeling Heating Problem in New York City
- Modeling Heating Problems in New York City - Part 2
- Rinterpreter
- Step By Step Predictive Modeling
- Stream Example
- Streaming Twitter
- test
- viest1
- XML example
- Zeppelin Tutorial

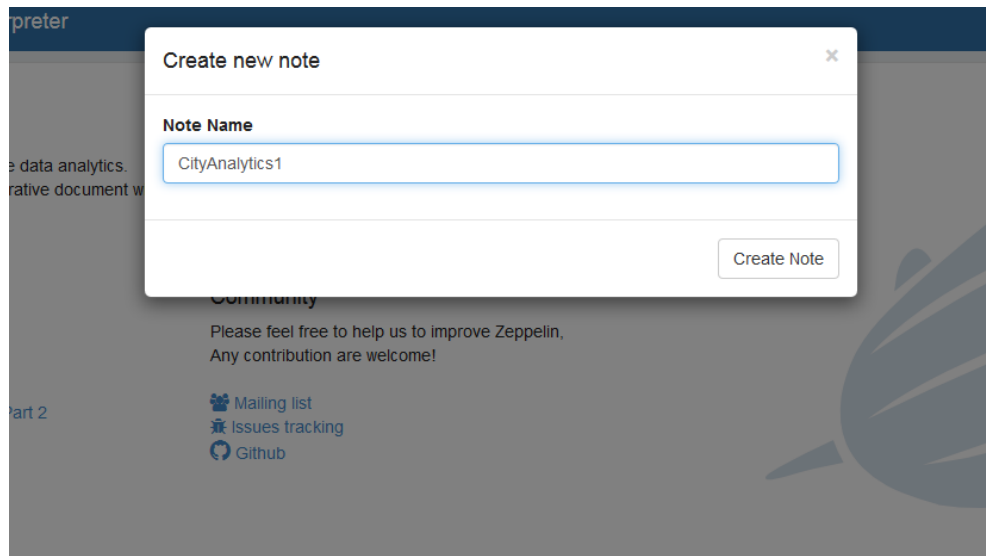
**Help**

Get started with [Zeppelin documentation](#)

**Community**

Please feel free to help us to improve Zeppelin,  
Any contribution are welcome!

- [Mailing list](#)
- [Issues tracking](#)
- [Github](#)



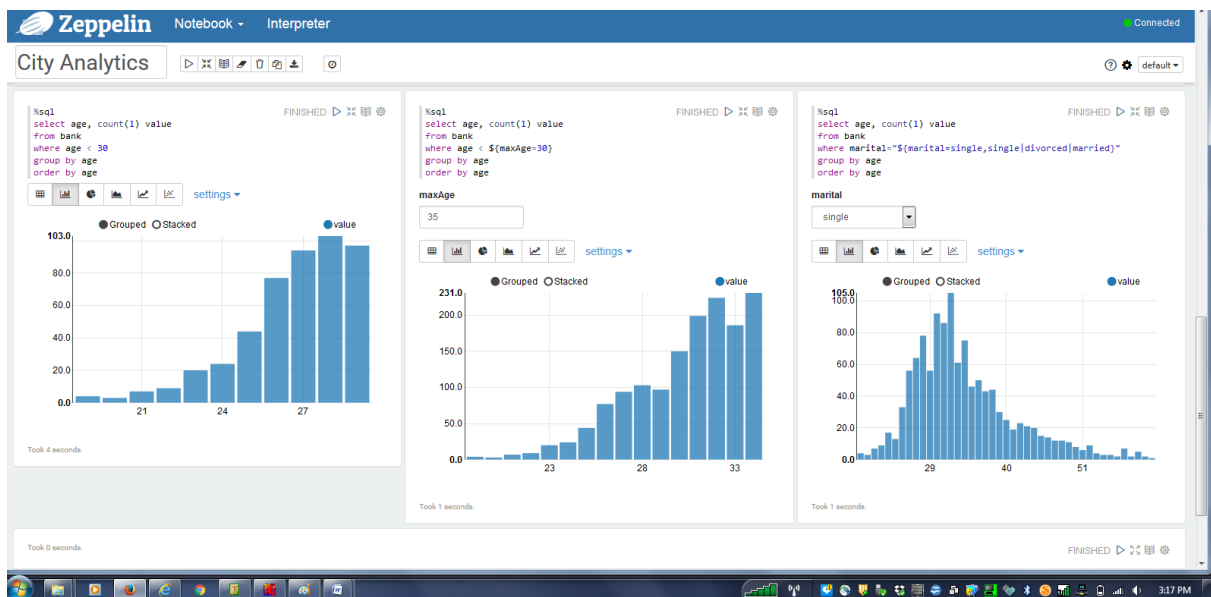
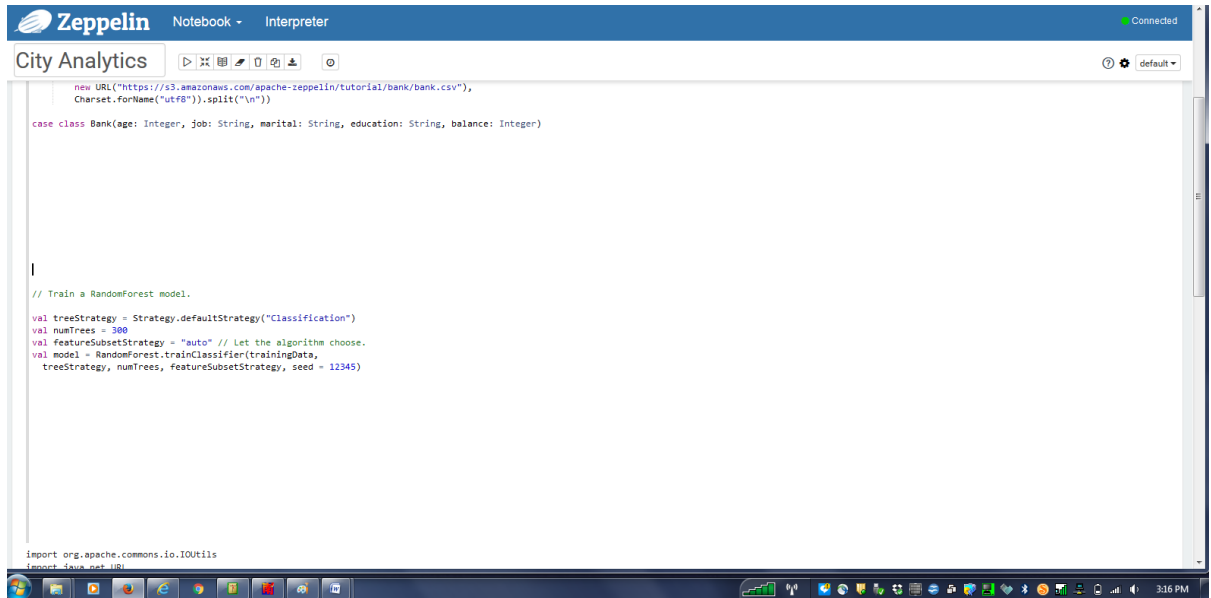
The image shows a "Create new note" dialog box overlaid on the Zeppelin Notebook interface. The dialog has a title bar with a close button. Inside, there's a "Note Name" label and a text input field containing "CityAnalytics1". At the bottom right of the dialog is a "Create Note" button.

Create new note

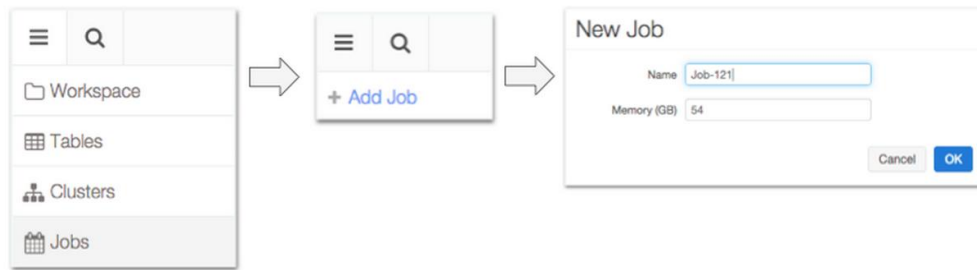
Note Name

CityAnalytics1

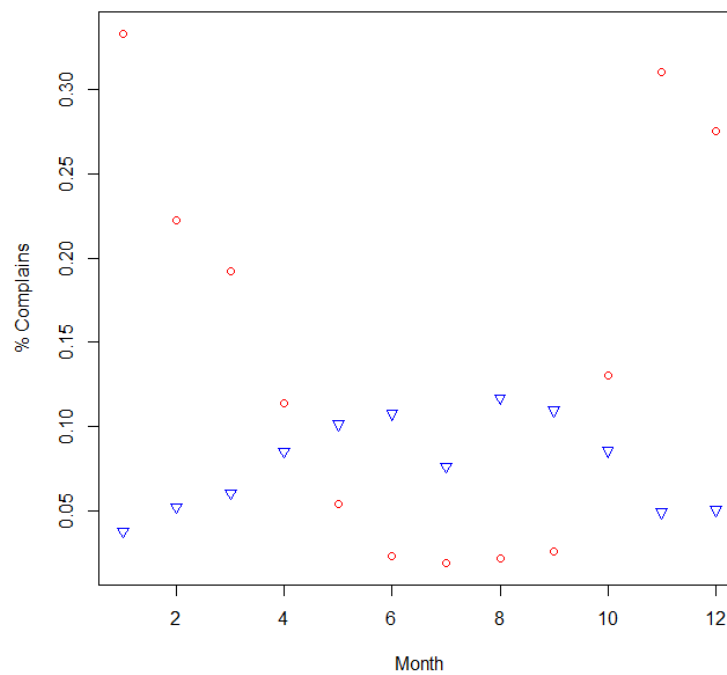
Create Note

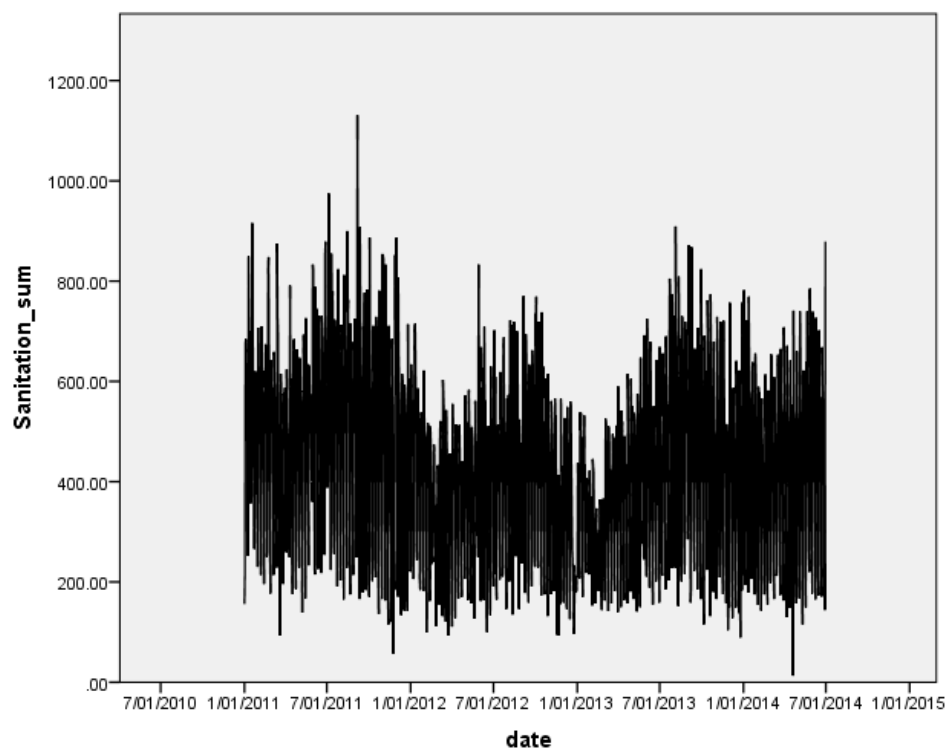


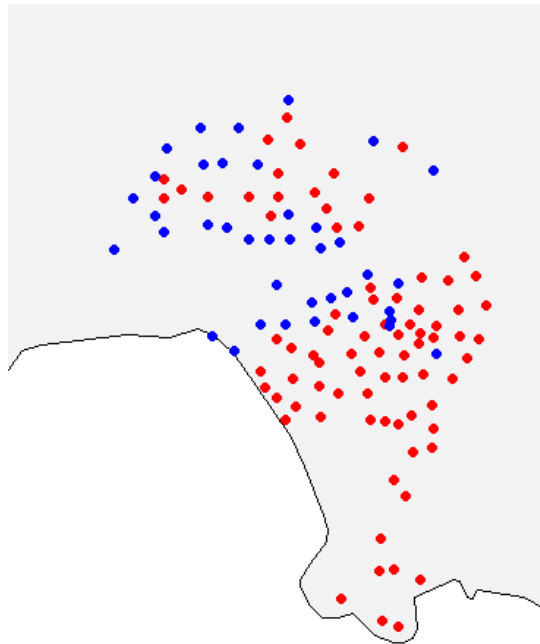
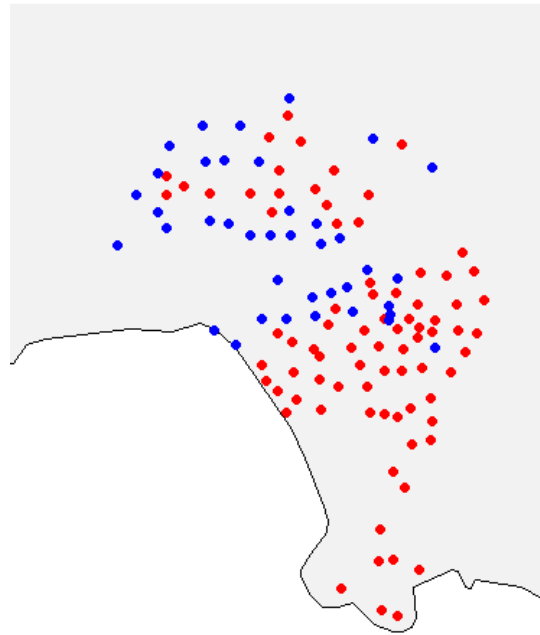




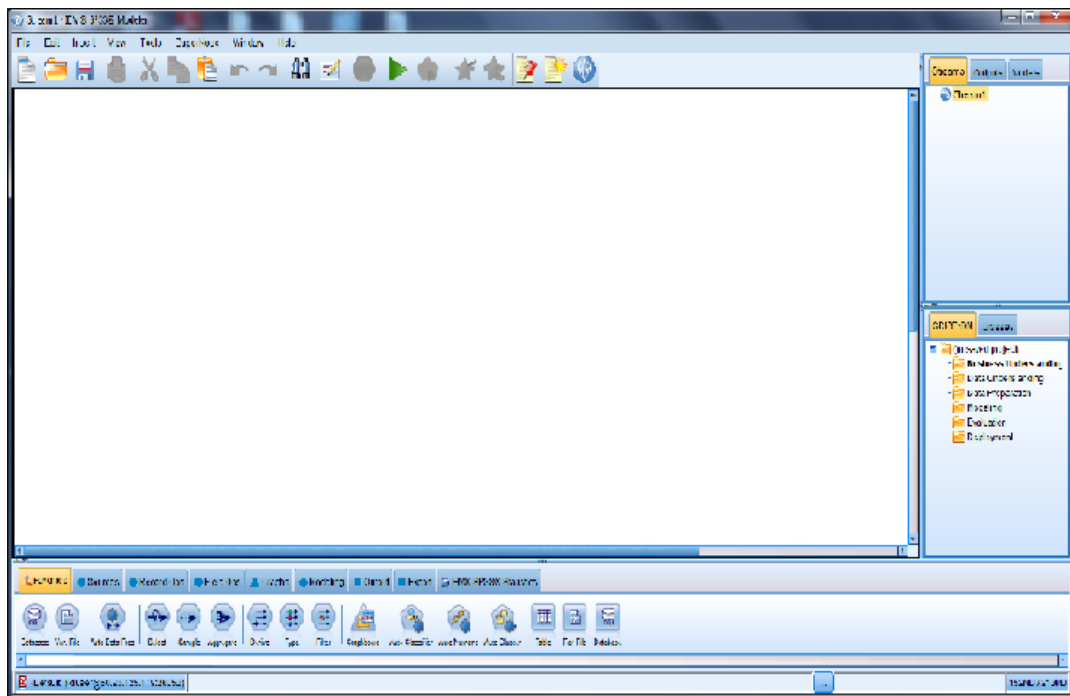
**% Heating and Noise Complains by Month**

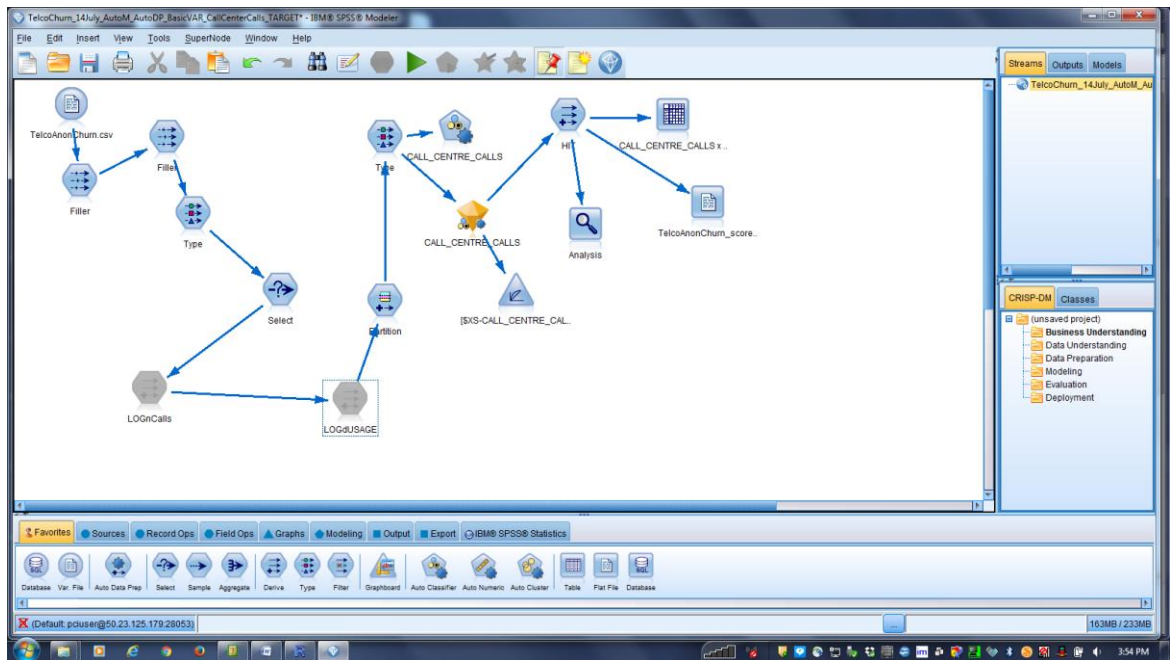






## Chapter 10: Learning Telco Data on Spark





Server Login

Select, add or edit the server to which you want to connect in the table. By default, the connection marked as Default is used at start-up.

Default	Server Name	Description	Port
<input type="checkbox"/>	Local Server		
<input checked="" type="checkbox"/>	50.20.125.300		28053

Add...

Edit...

Delete

Search...

Default data path: \$MODELERSERVER/data

☒ Set Credentials

User ID: rec-user

Password:

Domain:

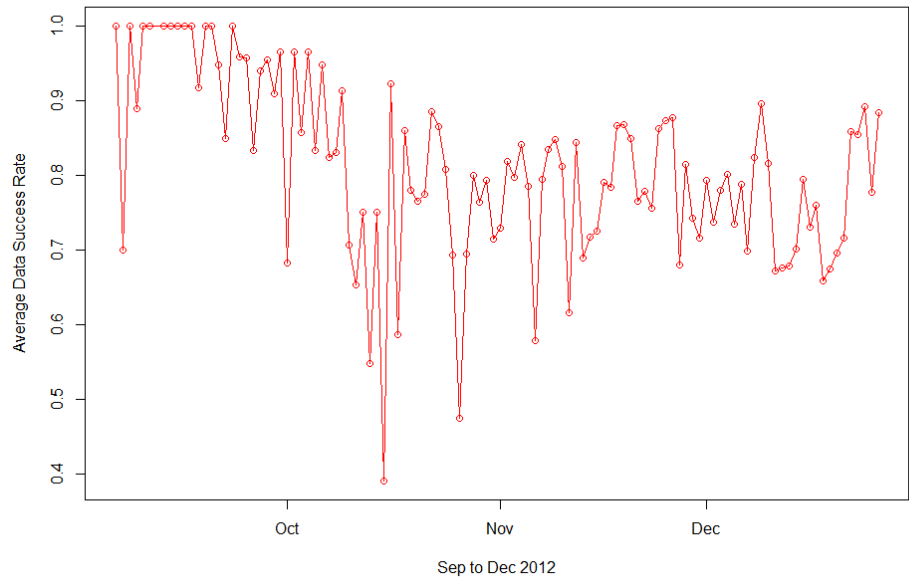
OK

Cancel

Help

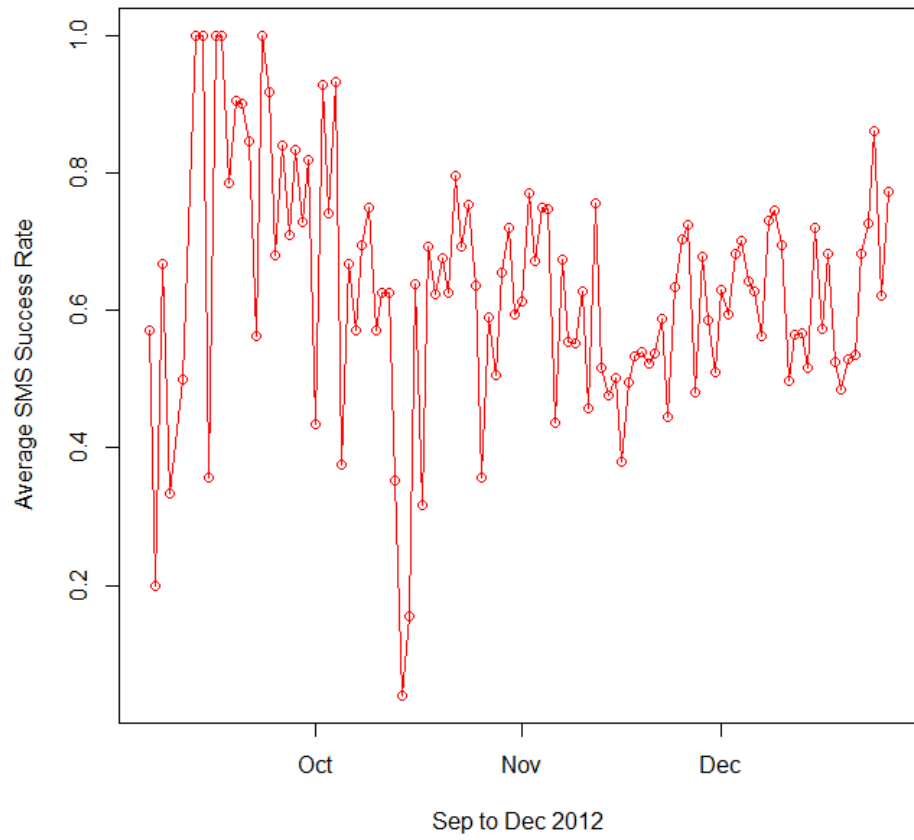


Data Success Rate in 2012

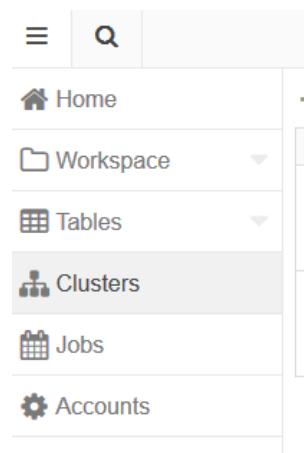
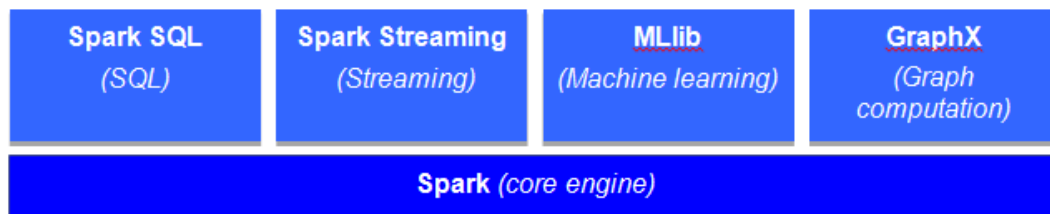




**SMS Success Rate in 2012**



## Chapter 11: Modeling Open Data on Spark



☰

🔍

## Table Import

Data Source

AWS Key ID

Secret Access Key

S3 Bucket Name

S3

S3

DBFS

JDBC

File

Browse Bucket



 BIG DATA UNIVERSITY



# Data Scientist Workbench

Prepare data. Analyze data. Get answers.

  python™

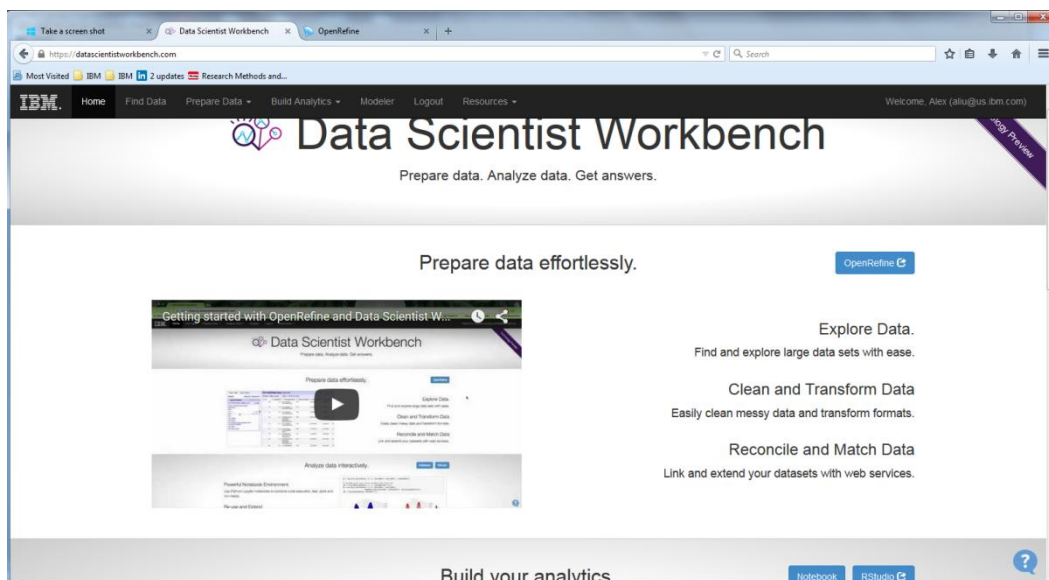
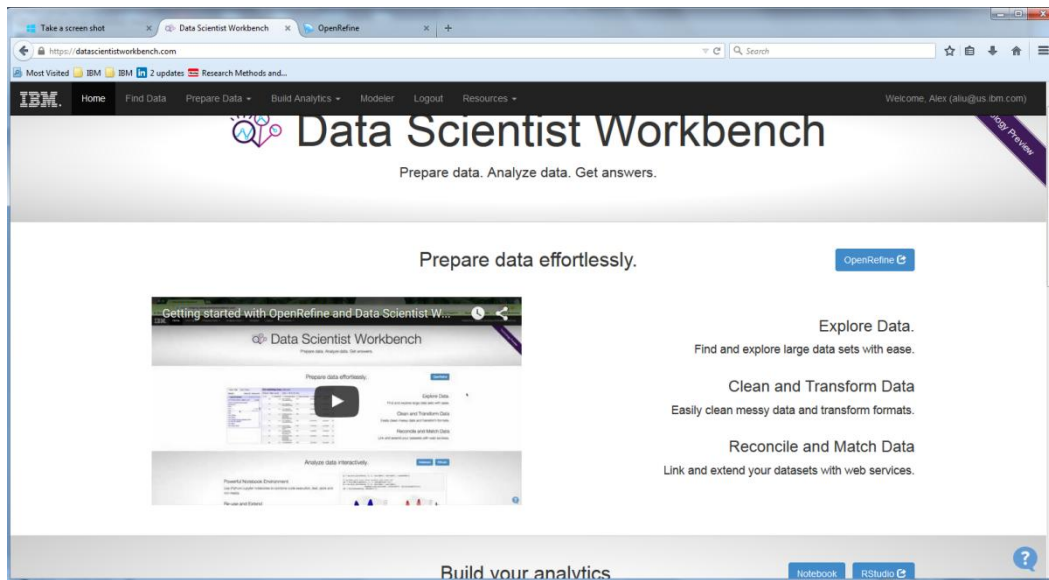
 

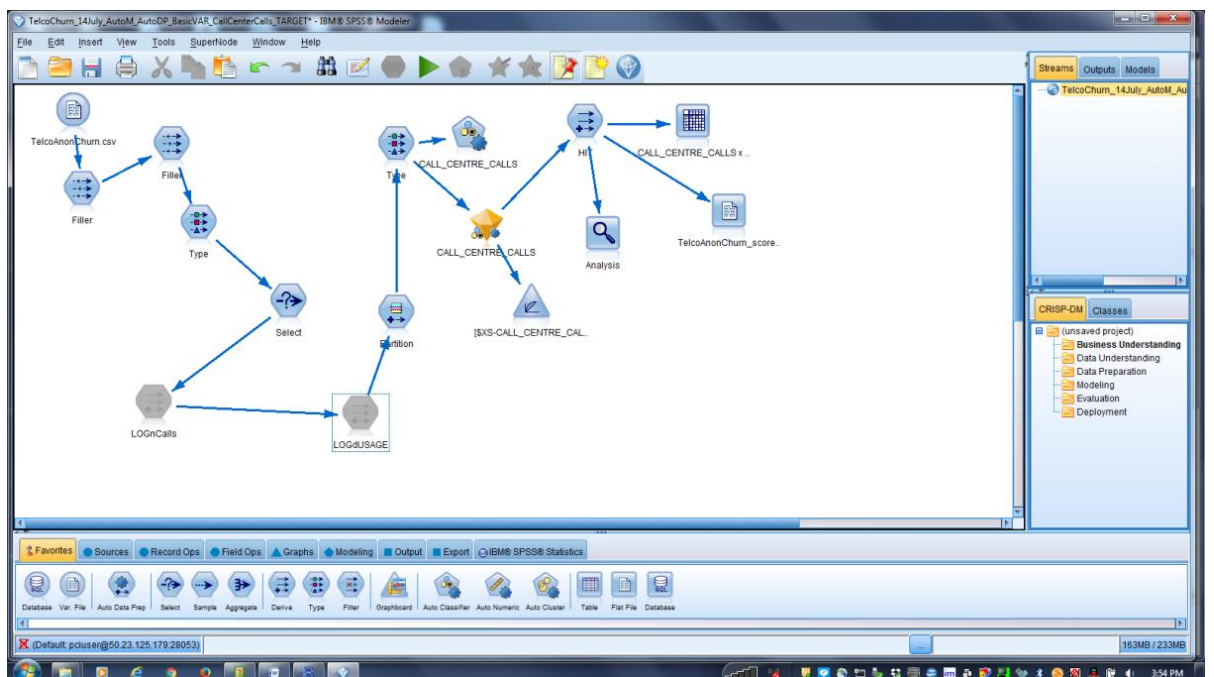
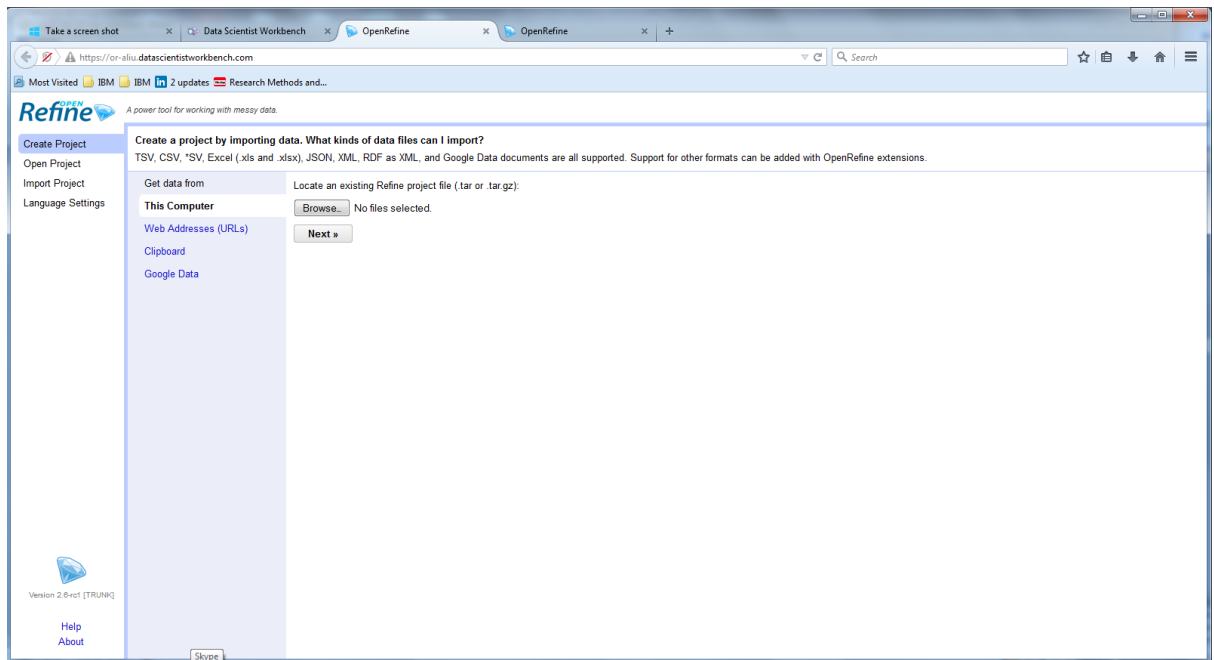
<https://datascientistworkbench.com>

🏠 📄 🔄

© 2015 BigDataUniversity.com

🐦 @bigdatau





Cluster Dendrogram

