Perform text analytics with Bluemix analytics services

Lucas Silva
Software Engineer
Big Data University

Felipe Mosquetta
Business Analyst Intern
Big Data University

Marcos Rogério De Mello
Software Engineer
Big Data University

25 February 2015
(First published 21 August 2014)

This tutorial explains how to perform text analytics using the Analytics for Hadoop service and dashDB service (formerly known as the Analytics Warehouse service) on IBM Bluemix™. Most of the processes are performed on a client machine with Eclipse IDE and the BigInsights™ plugin installed. After extracting the desired text, we use R in dashDB to plot charts with the results.

Introduction

Sentiment analysis, or simply finding patterns in text, is one key tool many companies are using today to understand customers, monitor the reputations of their brands, or see what competitors are doing. Twitter and Facebook data are often used as input to perform such analyses. In this tutorial, we show you how you can easily gain valuable insight by performing such analyses. The 2014 Emmy Awards was used as the topic for our case scenario. Two lists with all the nominees were obtained from the USA Today web site. The first list includes the actors, and the second includes the series. These lists will be used to create our dictionaries to perform text analytics.

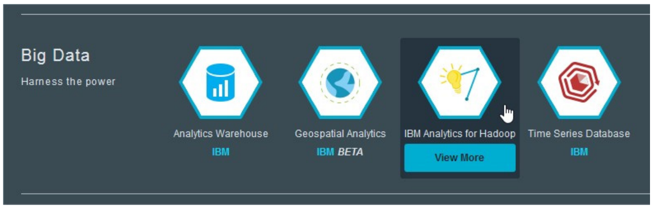
What you'll need for your application:

- Eclipse Classic 4.2.2
- An IBM Bluemix™ account

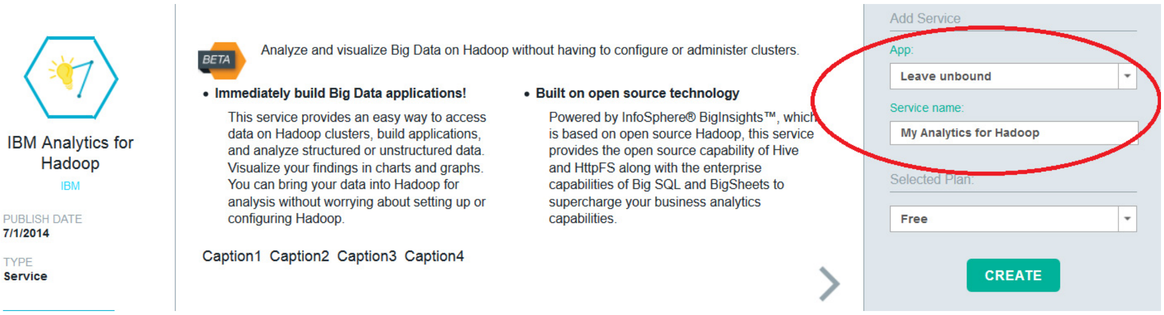
Setting up the Eclipse environment to perform text analytics

Assuming you already have Eclipse 4.2.2 installed on your computer, you need to add or install the InfoSphere® BigInsights Eclipse tools plugin:

1. Log in to your Bluemix account. On the menu, click **Catalog** and scroll down until you find the Big Data section. Then choose **IBM Analytics for Hadoop**.



2. In the open window, click on the combo box **App** and select **Leave unbound**. Feel free to give a name of your own preference; we used "My Analytics for Hadoop." Click **Create**.



Analyze and visualize Big Data on Hadoop without having to configure or administer clusters.

- **Immediately build Big Data applications!**

  This service provides an easy way to access data on Hadoop clusters, build applications, and analyze structured or unstructured data. Visualize your findings in charts and graphs. You can bring your data into Hadoop for analysis without worrying about setting up or configuring Hadoop.

- **Built on open source technology**

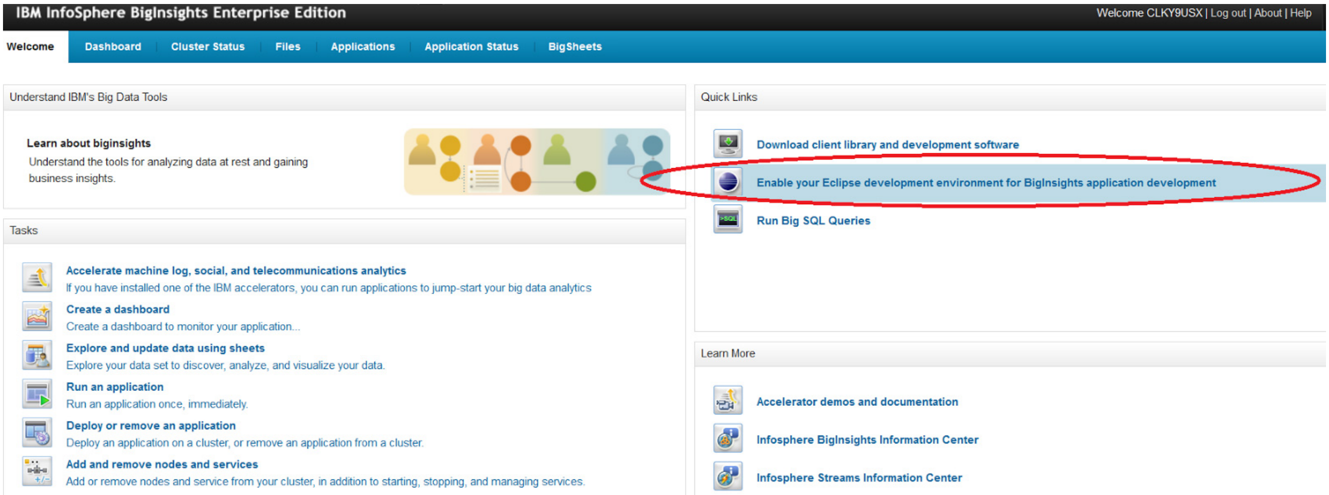
  Powered by InfoSphere® BigInsights™, which is based on open source Hadoop, this service provides the open source capability of Hive and HttpFS along with the enterprise capabilities of Big SQL and BigSheets to supercharge your business analytics capabilities.

App: Leave unbound

Service name: My Analytics for Hadoop

Selected Plan: Free

3. You will automatically be redirected to the service page. Click **Launch**.
4. You'll be sent to the InfoSphere BigInsights Enterprise Edition welcome page. In the Quick Links section, click **Enable your Eclipse development environment for BigInsights application development**.


IBM InfoSphere BigInsights Enterprise Edition
Welcome CLKY9USX | Log out | About | Help
Welcome
Dashboard
Cluster Status
Files
Applications
Application Status
BigSheets
Understand IBM's Big Data Tools
Quick Links
Tasks
Learn More

Learn about biginsights
Understand the tools for analyzing data at rest and gaining business insights.

Download client library and development software
Enable your Eclipse development environment for BigInsights application development
Run Big SQL Queries

Accelerate machine log, social, and telecommunications analytics
If you have installed one of the IBM accelerators, you can run applications to jump-start your big data analytics

Create a dashboard
Create a dashboard to monitor your application...

Explore and update data using sheets
Explore your data set to discover, analyze, and visualize your data.

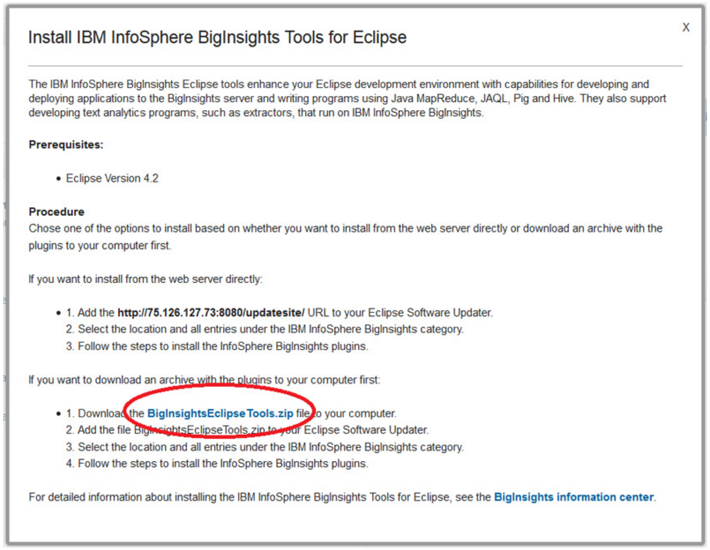
Run an application
Run an application once, immediately.

Deploy or remove an application
Deploy an application on a cluster, or remove an application from a cluster.

Add and remove nodes and services
Add or remove nodes and service from your cluster, in addition to starting, stopping, and managing services.

Accelerator demos and documentation
Infosphere BigInsights Information Center
Infosphere Streams Information Center

5. You have two options to install the InfoSphere BigInsights Eclipse tools: install from the web server or from a .zip file. We will use the second one here so download BigInsigthsEclipseTools.zip.



Install IBM InfoSphere BigInsights Tools for Eclipse

The IBM InfoSphere BigInsights Eclipse tools enhance your Eclipse development environment with capabilities for developing and deploying applications to the BigInsights server and writing programs using Java MapReduce, JAQL, Pig and Hive. They also support developing text analytics programs, such as extractors, that run on IBM InfoSphere BigInsights.

Prerequisites:

- Eclipse Version 4.2

Procedure
Chose one of the options to install based on whether you want to install from the web server directly or download an archive with the plugins to your computer first.

If you want to install from the web server directly:

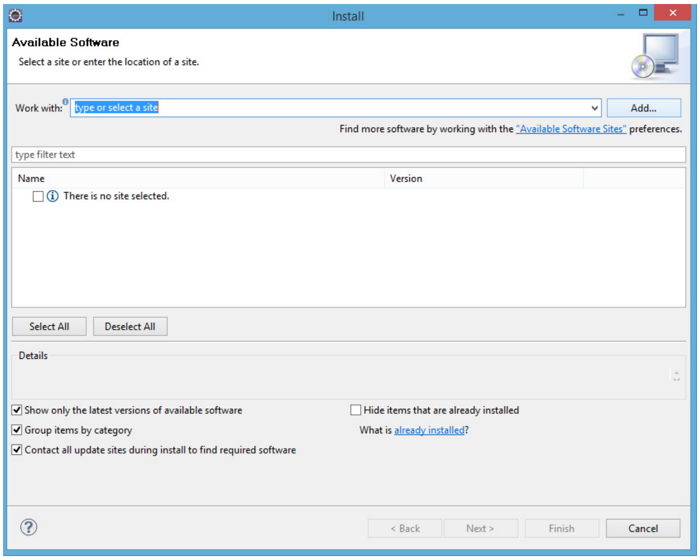
- 1. Add the **http://75.126.127.73:8080/updatesite/** URL to your Eclipse Software Updater.
  2. Select the location and all entries under the IBM InfoSphere BigInsights category.
  3. Follow the steps to install the InfoSphere BigInsights plugins.

If you want to download an archive with the plugins to your computer first:

- 1. Download the **BigInsightsEclipseTools.zip** file to your computer.
  2. Add the file BigInsightsEclipseTools.zip to your Eclipse Software Updater.
  3. Select the location and all entries under the IBM InfoSphere BigInsights category.
  4. Follow the steps to install the InfoSphere BigInsights plugins.

For detailed information about installing the IBM InfoSphere BigInsights Tools for Eclipse, see the **BigInsights information center**.

6. Now you need to install the plugin in the Eclipse IDE. Open Eclipse, select **Help > Install New Software**.



Install

Available Software
Select a site or enter the location of a site.

Work with: type or select a site

Find more software by working with the "Available Software Sites" preferences.

type filter text

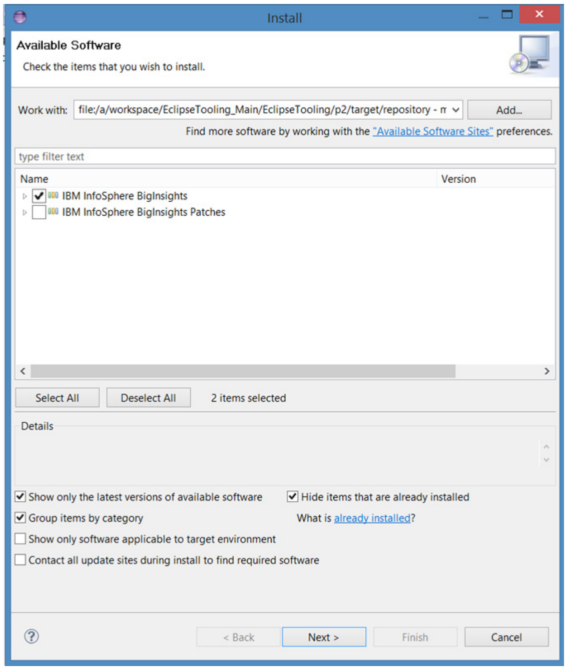
Name	Version
□ There is no site selected.	

Details

☑ Show only the latest versions of available software
☑ Group items by category
☑ Contact all update sites during install to find required software
□ Hide items that are already installed
What is already installed?

7. Click **Add** next to the **Work with** field, then select the BigInsightsEclipseTools.zip file (the file you downloaded) from your computer and click **OK**.

8. Check the InfoSphere BigInsights category and all of its features, then click **Next**. (Ensure that group items by category is checked so you see the category and its features).



Install

Available Software

Check the items that you wish to install.

Work with: file:/a/workspace/EclipseTooling_Main/EclipseTooling/p2/target/repository - n

Add...

Find more software by working with the "Available Software Sites" preferences.

type filter text

Name	Version
☑ IBM InfoSphere BigInsights	
□ IBM InfoSphere BigInsights Patches	

Select All | Deselect All | 2 items selected

Details

☑ Show only the latest versions of available software

☑ Hide items that are already installed

☑ Group items by category

What is already installed?

□ Show only software applicable to target environment

□ Contact all update sites during install to find required software

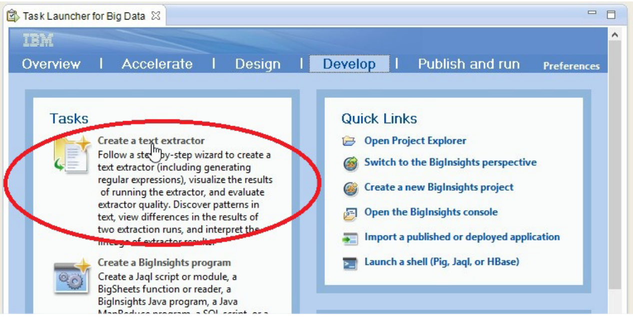
< Back | Next > | Finish | Cancel

9. Follow the installation instructions.
10. Click **OK** if a security warning message appears.
11. Restart Eclipse when prompted.

Creating a text analytics project

You need to create a new text analytics project on Eclipse IDE:

1. In the Task Launcher for Big Data, go to the Develop tab and choose **Create a text extractor** from the Tasks panel.



Task Launcher for Big Data

IBM

Overview | Accelerate | Design | Develop | Publish and run Preferences

Tasks

Create a text extractor

Follow a step-by-step wizard to create a text extractor (including generating regular expressions), visualize the results of running the extractor, and evaluate extractor quality. Discover patterns in text, view differences in the results of two extraction runs, and interpret the lineage of extractor results.

Create a BigInsights program

Create a Jaql script or module, a BigSheets function or reader, a BigInsights Java program, a Java MapReduce program, a SQL script, or a

Quick Links

Open Project Explorer

Switch to the BigInsights perspective

Create a new BigInsights project

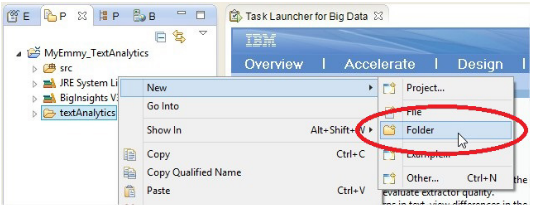
Open the BigInsights console

Import a published or deployed application

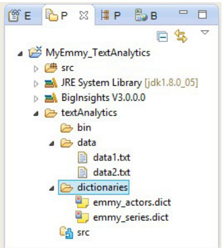
Launch a shell (Pig, Jaql, or HBase)

2. Name your project (suggestion: `MyEmmy_textAnalytics`) and click **Finish**.
3. Click **Yes** to switch to the InfoSphere BigInsights perspective when prompted.
4. Create a folder to put the data files. In the Project Explorer view, expand your project folder and find a folder called textAnalytics. Right click the textAnalytics folder

and select **New > Folder**. Name your folder (suggestion: `data`) and click **Finish**.



5. Now you can move your data to this folder. Select the folder you just created, copy your data files from your computer, and paste them in the Eclipse folder. You can also import the files with your data (**File > Import**) or drag and drop the file with your data from your local file system onto the data folder in Eclipse you just created.
6. Create a folder for the dictionaries. In the Project Explorer view, expand your project folder and select the **textAnalytics** folder. Right-click the textAnalytics folder and select **New > Folder**. Write `dictionaries` as the name of the folder (you can choose a different name, but this may result in divergences throughout this article) and click **Finish**.
7. Now you can move the dictionaries files to the folder that you just created. Two dictionaries contain the nominees of the International Emmy Awards: one with all the actors and actresses (emmy_actors.dict) and the other with all the series (emmy_series.dict). The next image shows the data, the dictionaries folders, and the files.

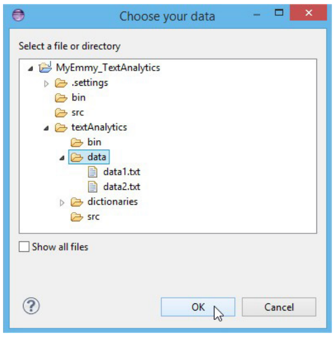


Now you are ready to start the text analytics program.

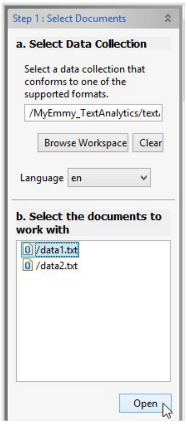
The text analytics workflow in Eclipse

In this section, we will go through the steps in the text analytics workflow in Eclipse to develop an extractor:

1. We need to specify the location of our data. In the Extraction Tasks view, go to **Step 1: Select Documents** and click **Browse Workspace**. Select the folder with your data (created in the previous section, called data) and click **OK**.



2. In the Language box, choose the language of your data. Select **en** for English. To use another language, you can check all the supported languages and their codes on the Supported languages webpage.
3. In Step 1b, select one of your data files and click **Open**. After this, you have completed Step 1 of the workflow.



a. Select Data Collection

Select a data collection that conforms to one of the supported formats.

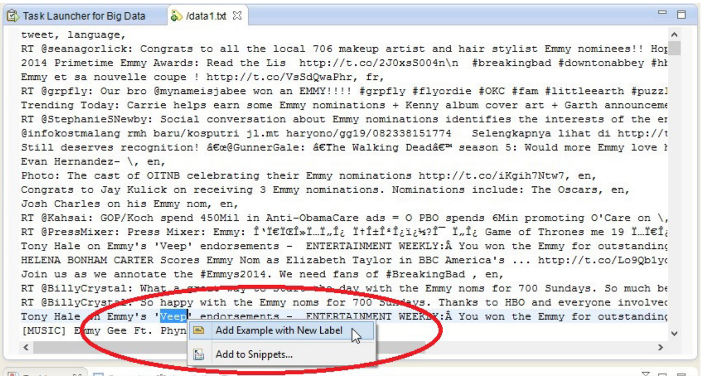
/MyEmmy_TextAnalytics/text

Language en

b. Select the documents to work with

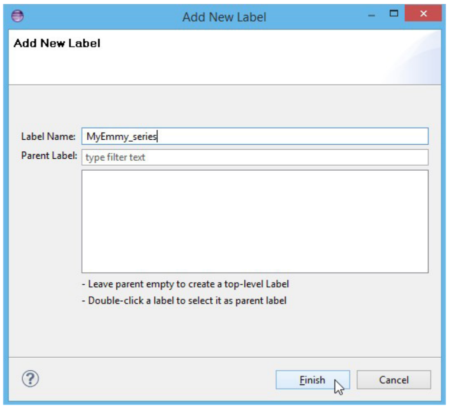
/data1.txt
/data2.txt

4. In Step 2, you will label examples in the data. This step is useful in helping you to develop the extractor. In the data1.txt file, select **Veep**, right-click, then click **Add example with New Label**.



```
tweet, language,
RT @seanagorlick: Congrats to all the local 706 makeup artist and hair stylist Emmy nominees!! Hop
2014 Primetime Emmy Awards: Read the Lis  http://t.co/2J0xsS004n\n  #breakingbad #downtonabbey #hb
Emmy et sa nouvelle coupe ! http://t.co/VsSdQwaPhr, fr,
RT @grpfly: Our bro @mynameisjabee won an EMMY!!!! #grpfly #flyordie #OKC #fam #littleearth #puzzl
Trending Today: Carrie helps earn some Emmy nominations + Kenny album cover art + Garth announceme
RT @StephanieSNewby: Social conversation about Emmy nominations identifies the interests of the en
@infokostmalang rmh baru/kosputri jl.mt haryono/gg19/082338151774   Selengkapnya lihat di http://t
Still deserves recognition! â€œ@GunnerGale: â€The Walking Deadâ€™ season 5: Would more Emmy love h
Evan Hernandez- \, en,
Photo: The cast of OITNB celebrating their Emmy nominations http://t.co/iKgih7Ntw7, en,
Congrats to Jay Kulick on receiving 3 Emmy nominations. Nominations include: The Oscars, en,
Josh Charles on his Emmy nom, en,
RT @Kahsai: GOP/Koch spend 450Mil in Anti-ObamaCare ads = O PBO spends 6Min promoting O'Care on \,
RT @PressMixer: Press Mixer: Emmy: Î'Ï€ÏŒÎ»Ï…Ï„Î¿ Ï†Î±Î¹Î¿ï¿½?Î¯ Ï„Î¿ Game of Thrones me 19 Ï…Ï€Î¿
Tony Hale on Emmy's 'Veep' endorsements -  ENTERTAINMENT WEEKLY:Â You won the Emmy for outstanding
HELENA BONHAM CARTER Scores Emmy Nom as Elizabeth Taylor in BBC America's ... http://t.co/Lo9Qb1yc
Join us as we annotate the #Emmys2014. We need fans of #BreakingBad , en,
RT @BillyCrystal: What a great way to start the day with the Emmy noms for 700 Sundays. So much be
RT @BillyCrystal: So happy with the Emmy noms for 700 Sundays. Thanks to HBO and everyone involved
Tony Hale on Emmy's 'Veep' endorsements -  ENTERTAINMENT WEEKLY:Â You won the Emmy for outstanding
[MUSIC] Emmy Gee Ft. Phyn
```

5. Name your label (suggestion: `MyEmmy_series`) in the Label Name field and leave the Parent label empty. Click **Finish**.



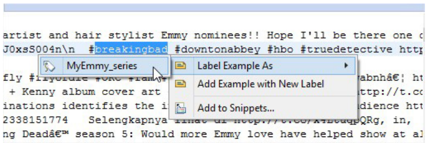
Add New Label

Label Name: MyEmmy_series

Parent Label: type filter text

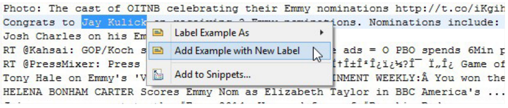
- Leave parent empty to create a top-level Label
- Double-click a label to select it as parent label

6. Select a different series (**Breaking Bad**, for example). Right-click, select **Label example as**, then click the name that you chose for your example in the previous step.



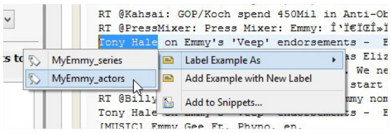
```
artist and hair stylist Emmy nominees!! Hope I'll be there one c
J0xsS004n\n  #breakingbad #downtonabbey #hbo #truedetective http
fly #                                            abnhâ€¦ ht
 + Kenny album cover art                         ttp://t.cc
inations identifies the i                        dience htt
2338151774   Selengkapnya                      QRg, in,
ng Deadâ€™ season 5: Would more Emmy love have helped show at al
```

7. Let's create another label for the actors and actresses. Select **Jay Kulick** in the data1.txt file, right-click, then click **Add example with New Label**.



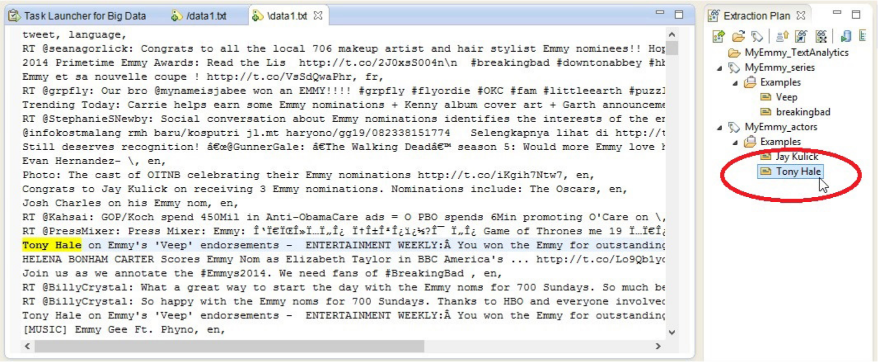
```
Photo: The cast of OITNB celebrating their Emmy nominations http://t.co/iKgih
Congrats to Jay Kulick                          ions. Nominations include:
Josh Charles on his Em
RT @Kahsai: GOP/Koch s                          e ads = 0 PBO spends 6Min p
RT @PressMixer: Press                           Î†Î±Î²Î¿ï¿½?Î¯ Ï„Î¿ Game of
Tony Hale on Emmy's 'V                          INMENT WEEKLY:Â You won the
HELENA BONHAM CARTER Scores Emmy Nom as Elizabeth Taylor in BBC America's ...
```

8. Write a name for your label (suggestion: `MyEmmy_actors`) in the Label Name field and leave the Parent label empty again. Click **Finish**.
9. Select a different actor's name (**Tony Hale**, for example), right-click, select **Label example as**, then click the name you chose for your example in the previous step.



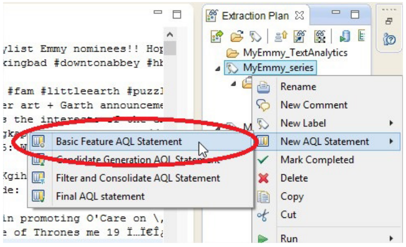
```
RT @Kahsai: GOP/Koch spend 450Mil in Anti-Ob
RT @PressMixer: Press Mixer: Emmy: Î'Ï€ÏŒÎ»Ï
Tony Hale on Emmy's 'Veep' endorsements -  E
                                 as Eliz
                                 . We ne
                                 start
RT @Billy                         mmy nom
Tony Hale                        ts -  E
[MUSIC] Emmy Gee Ft. Phyno, en,
```

What we did in Step 2 was to select examples of actors and series present in the data1.txt file and assign them their specific labels. So we selected Veep and Breaking Bad as examples of series, and Jay Kulick and Tony Hale as examples of actors. We won't create more labels because we just need to find actors and series in our data. You can continue labeling the text, adding more examples of actors and series. At the end, look in the Extraction Plan to view the examples that you added. Double-click on one of them to highlight its location in the text.

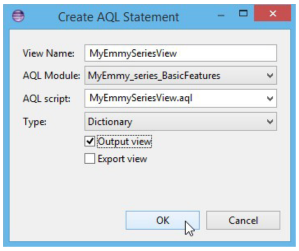


```
tweet, language,
RT @seanagorlick: Congrats to all the local 706 makeup artist and hair stylist Emmy nominees!! Hop
2014 Primetime Emmy Awards: Read the Lis  http://t.co/2J0xsS004n\n  #breakingbad #downtonabbey #hk
Emmy et sa nouvelle coupe ! http://t.co/VsSdQwaPhr, fr,
RT @grpfly: Our bro @mynameisjabee won an EMMY!!!! #grpfly #flyordie #OKC #fam #littleearth #puzzl
Trending Today: Carrie helps earn some Emmy nominations + Kenny album cover art + Garth announceme
RT @StephanieSNewby: Social conversation about Emmy nominations identifies the interests of the en
@infokostmalang rmh baru/kosputri jl.mt haryono/gg19/082338151774   Selengkapnya lihat di http://t
Still deserves recognition! â€œ@GunnerGale: â€The Walking Deadâ€™ season 5: Would more Emmy love h
Evan Hernandez- \, en,
Photo: The cast of OITNB celebrating their Emmy nominations http://t.co/iKgih7Ntw7, en,
Congrats to Jay Kulick on receiving 3 Emmy nominations. Nominations include: The Oscars, en,
Josh Charles on his Emmy nom, en,
RT @Kahsai: GOP/Koch spend 450Mil in Anti-ObamaCare ads = 0 PBO spends 6Min promoting O'Care on \,
RT @PressMixer: Press Mixer: Emmy: Î'Ï€ÏŒÎ»Ï…Î„Î¿ Ï†Î±Î²Î¿ï¿½?Î¯ Ï„Î¿ Game of Thrones me 19 Ï…Ï€Î¿
Tony Hale on Emmy's 'Veep' endorsements -  ENTERTAINMENT WEEKLY:Â You won the Emmy for outstanding
HELENA BONHAM CARTER Scores Emmy Nom as Elizabeth Taylor in BBC America's ... http://t.co/Lo9Qb1yc
Join us as we annotate the #Emmys2014. We need fans of #BreakingBad , en,
RT @BillyCrystal: What a great way to start the day with the Emmy noms for 700 Sundays. So much be
RT @BillyCrystal: So happy with the Emmy noms for 700 Sundays. Thanks to HBO and everyone involved
Tony Hale on Emmy's 'Veep' endorsements -  ENTERTAINMENT WEEKLY:Â You won the Emmy for outstanding
[MUSIC] Emmy Gee Ft. Phyno, en,
```

10. In Step 3 of the text analytics workflow, you will create an extractor. In the Extraction Plan view, right-click on the name of your first label (**MyEmmy_series** if you gave the same name as suggested), select **New AQL Statement > Basic Feature AQL Statement**.



11. Give a name to your view (suggestion: `MyEmmy_seriesView`) in the View Name field and leave the AQL Module field with its default value. Then specify a name for your script (suggestion: `MyEmmy_seriesView.aql`) in the AQL script field. In the Type field, you have to select the type of your extract statement. To learn more, see The Extract Statement reference. Here, we will use dictionary as the type of our extractors, so we have to choose this option from the drop-down menu. Finally, select the **Output view** checkbox and click **OK**.

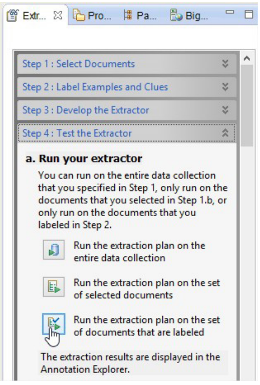


View Name: MyEmmySeriesView
AQL Module: MyEmmy_series_BasicFeatures
AQL script: MyEmmySeriesView.aql
Type: Dictionary
☑ Output view
□ Export view

12. After that, an AQL file opens in the editor with some errors. We just have to fix them. We are going to start with the following code: `create dictionary MyEmmySeriesViewDict from file '<path to your dictionary here>' with language as 'en';`.
13. Replace `<path to your dictionary here>` for the path with the dictionaries with all the names of the series: `../../dictionaries/emmy_series.dict`. We need the `../` because the folder with the dictionaries is under the textAnalytics folder, and the current directory is the Emmy_series_BasicFeatures module. If you don't want to put it, just move the dictionary you want to use to your AQL file level. You can do this in the Project Explorer view. You can also create dictionaries from the examples labeled in the previous steps or add them to existing dictionaries. To learn more about this, read Step 10.b of Lesson 3: Writing and testing AQL.
14. Now we are going to work with this part of the code: `create view MyEmmySeriesView as extract dictionary MyEmmySeriesViewDict on R.<input column> as match from <input view> R;`.
15. Replace `<input column>` for text and `<input view>` for document and save your file. Then you should have code similar to the one below and with no errors.

```
module MyEmmy_series_BasicFeatures; create dictionary MyEmmySeriesViewDict
from file '../../dictionaries/emmy_series.dict' with language as 'en';
create view MyEmmySeriesView as extract dictionary
MyEmmySeriesViewDict on R.text as match from Document R; output view
MyEmmySeriesView;
```

16. We are going to test our extractor. There are three ways to run it: run the extraction plan on the entire data collection, on the set of selected documents, or on the set documents labeled. Select the third option: **Run the extraction plan on the set of documents that are labeled**.



Step 1 : Select Documents
Step 2 : Label Examples and Clues
Step 3 : Develop the Extractor
Step 4 : Test the Extractor

a. Run your extractor

You can run on the entire data collection that you specified in Step 1, only run on the documents that you selected in Step 1.b, or only run on the documents that you labeled in Step 2.

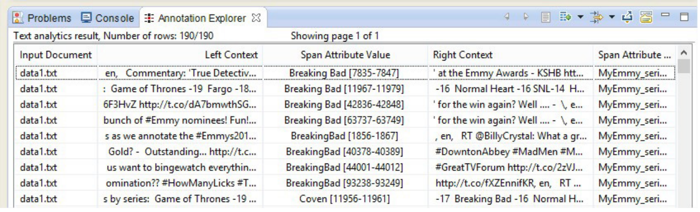
Run the extraction plan on the entire data collection

Run the extraction plan on the set of selected documents

Run the extraction plan on the set of documents that are labeled

The extraction results are displayed in the Annotation Explorer.

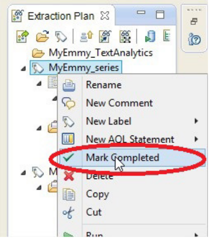
17. The results are shown in the Annotation Explorer view — specifically, in the Span Attribute Value column.



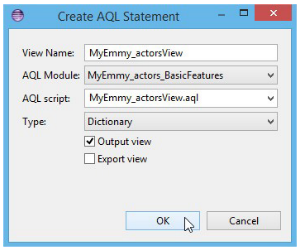
Text analytics result, Number of rows: 190/190 Showing page 1 of 1

Input Document	Left Context	Span Attribute Value	Right Context	Span Attribute ...
data1.txt	en, Commentary: 'True Detectiv...	Breaking Bad [7835-7847]	' at the Emmy Awards - KSHB htt...	MyEmmy_seri...
data1.txt	: Game of Thrones -19 Fargo -18...	Breaking Bad [11967-11979]	-16 Normal Heart -16 SNL-14 H...	MyEmmy_seri...
data1.txt	6F3HvZ http://t.co/dA7bmwthSG...	Breaking Bad [42836-42848]	' for the win again? Well .... - \, e...	MyEmmy_seri...
data1.txt	bunch of #Emmy nominees! Fun!...	Breaking Bad [63737-63749]	' for the win again? Well .... - \, e...	MyEmmy_seri...
data1.txt	s as we annotate the #Emmys201...	BreakingBad [1856-1867]	, en, RT @BillyCrystal: What a gr...	MyEmmy_seri...
data1.txt	Gold? - Outstanding... http://t.c...	BreakingBad [40378-40389]	#DowntonAbbey #MadMen #M...	MyEmmy_seri...
data1.txt	us want to bingewatch everythin...	BreakingBad [44001-44012]	#GreatTVForum http://t.co/2zVJ...	MyEmmy_seri...
data1.txt	omination?? #HowManyLicks #T...	BreakingBad [93238-93249]	http://t.co/fXZEnnifKR, en, RT ...	MyEmmy_seri...
data1.txt	s by series: Game of Thrones -19 ...	Coven [11956-11961]	-17 Breaking Bad -16 Normal H...	MyEmmy_seri...

18. Now we can check this label as completed. Right-click on the name of the label and click **Mark Completed**.



19. Now, we need to repeat the same process to the MyEmmy_actors label. Do from Step 10 to Step 18 of this section for each label you have (in our case, we just have the MyEmmy_actors label left). The difference here is that you can create your view in the same AQL file of the previous one or create another aql file for your view. If you choose to include additional views in the same module, at the end, you will get just one tam file; otherwise, you will get one for each module. In this article, we will create the MyEmmy_actorsView in another module, which is the first option below.
    - To create in another AQL file, in the Create AQL Statement window (obtained in Step 11), write a name for the view in the View Name field (suggestion: MyEmmy_actorsView), select MyEmmy_actors_BasicFeatures as the AQL Module, write a name for your AQL file, choose dictionary from the drop-down menu and, finally, check the Output view checkbox. Click **OK**.

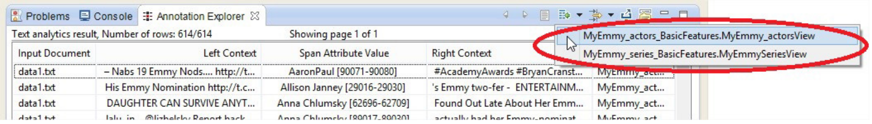
      

      View Name: MyEmmy_actorsView
      AQL Module: MyEmmy_actors_BasicFeatures
      AQL script: MyEmmy_actorsView.aql
      Type: Dictionary
      ☑ Output view
      □ Export view

    - To create in the same AQL file of the previous view, in the Create AQL Statement window (obtained in Step 11), select the same AQL module of the first label (MyEmmy_series_BasicFeatures) and the same AQL script (MyEmmy_seriesView.aql). Leave the other fields as shown in the preceding image.
20. After creating your view and fixing the errors in the template (steps 12-15), you should have the following code.

```
module MyEmmy_actors_BasicFeatures; create dictionary MyEmmy_actorsViewDict
                      from file '../../dictionaries/emmy_actors.dict' with language as 'en';
                      create view MyEmmy_actorsView as extract dictionary
                      MyEmmy_actorsViewDict on R.text as match from Document R; output view
                      MyEmmy_actorsView;
```

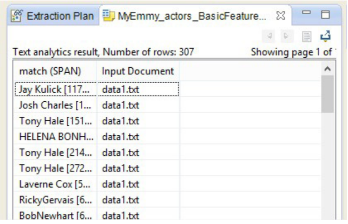
21. After running the extractor again (Step 16), you can select one of your views to see the results of each view separately. In this step, you can select the option to run your extraction plan on the entire data collection, unlike what was done in Step 16. Doing that, you will get results from all of your data files, not only from the labeled file.


Problems
Console
Annotation Explorer
MyEmmy_actors_BasicFeatures.MyEmmy_actorsView
MyEmmy_series_BasicFeatures.MyEmmySeriesView

Text analytics result, Number of rows: 614/614   Showing page 1 of 1

Input Document	Left Context	Span Attribute Value	Right Context	
data1.txt	– Nabs 19 Emmy Nods.... http://t...	AaronPaul [90071-90080]	#AcademyAwards #BryanCranst...	MyEmmy_act...
data1.txt	His Emmy Nomination http://t.c...	Allison Janney [29016-29030]	's Emmy two-fer - ENTERTAINM...	MyEmmy_act...
data1.txt	DAUGHTER CAN SURVIVE ANYT...	Anna Chlumsky [62696-62709]	Found Out Late About Her Emm...	MyEmmy_act...
data1.txt	Jalu_in... @lizbelsky Report back	Anna Chlumsky [89017-89030]	actually had her Emmy-nominat	MyEmmy_act

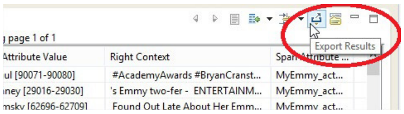
22. The results are shown in a different window beside the Extraction Plan view.



Text analytics result, Number of rows: 307   Showing page 1 of

match (SPAN)	Input Document
Jay Kulick [117...	data1.txt
Josh Charles [1...	data1.txt
Tony Hale [151...	data1.txt
HELENA BONH...	data1.txt
Tony Hale [214...	data1.txt
Tony Hale [272...	data1.txt
Laverne Cox [5...	data1.txt
RickyGervais [6...	data1.txt
BobNewhart [6...	data1.txt

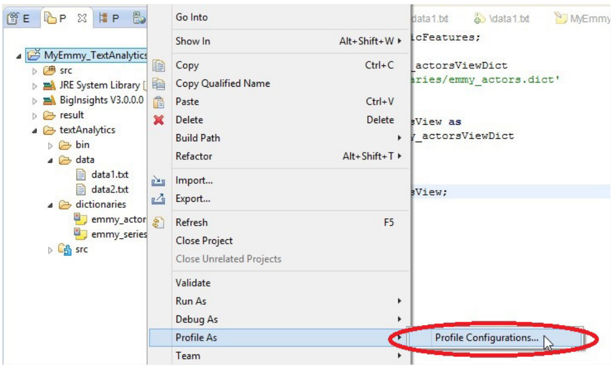
23. To finish Step 4 of the text analytics workflow, mark the MyEmmy_actors label as Completed (Step 18 of this section). Note: In this tutorial, we just use the Basic Feature AQL Statement, but you can also use the other ones. You can learn about them in Guidelines for writing AQL.
24. You can also export your data as HTML and CSV files after running the extraction plan on the entire data collection. In the Annotation Explorer view, click the **Export Results** icon.



g page 1 of 1

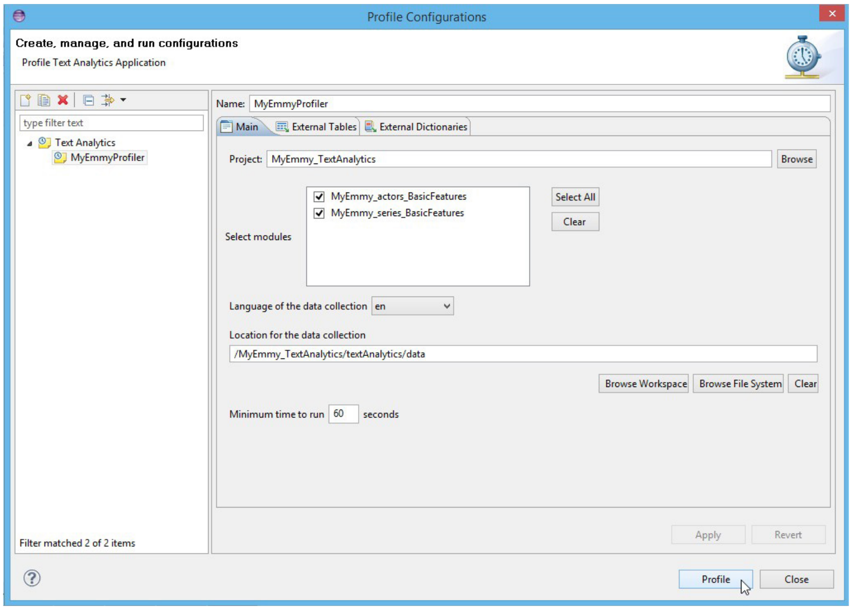
Attribute Value	Right Context	Span Attribute ...
ul [90071-90080]	#AcademyAwards #BryanCranst...	MyEmmy_act...
ney [29016-29030]	's Emmy two-fer - ENTERTAINM...	MyEmmy_act...
msky [62696-62709]	Found Out Late About Her Emm...	MyEmmy_act...

25. You now have to specify a path where the files will be saved. Click **Browse File System** and select the directory in which to save the results, then click **Finish**.
26. In Step 5 of the text analytics workflow, we can investigate the performance of our extractor. Go to the Project Explorer view, right-click on the name of your project, then go to Profile As and select **Profile Configurations**.



27. Double-click on **Text Analytics**, name your profile in the name field, select your modules, and specify the location for the data collection. You can also change the minimum time to the profiler runs. Then click **Profile**. After the profiler runs, you will get the performance runtime

results of your extractor. To learn more about the AQL Profiler, see AQL Logic optimization.



Name: MyEmmyProfiler

Project: MyEmmy_TextAnalytics

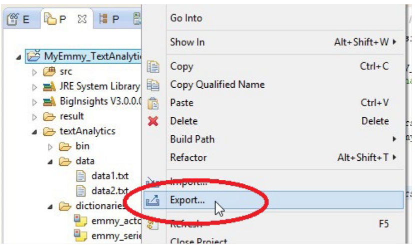
Select modules
☑ MyEmmy_actors_BasicFeatures
☑ MyEmmy_series_BasicFeatures

Language of the data collection: en

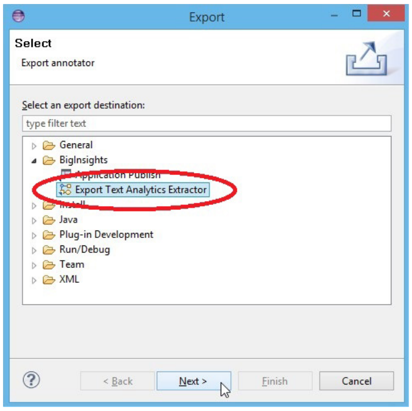
Location for the data collection: /MyEmmy_TextAnalytics/textAnalytics/data

Minimum time to run 60 seconds

28. The last step in the workflow is exporting the extractor. In the Project Explorer view, right-click on the name of your project, then go to Export.

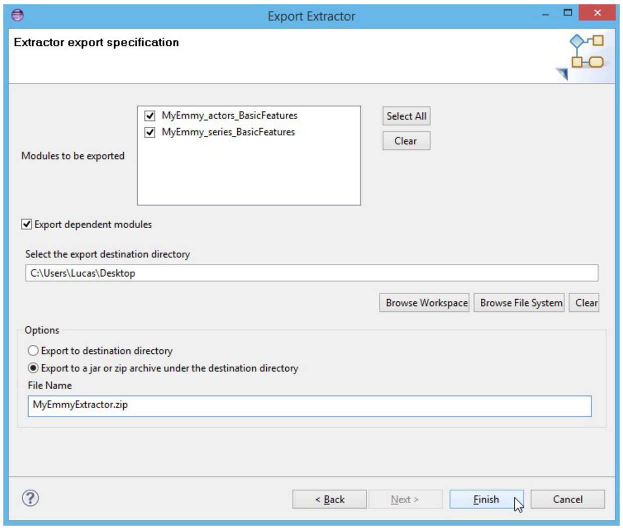


29. Expand the BigInsights folder, select **Export Text Analytics Extractor** and click **Next**.



30. Select the module to export and the **Export dependent modules** checkbox. Click **Browse File System** and specify the directory to save your extractor. In Options, select **Export to**

a jar or zip archive under the destination directory and write a name in the file name field.



Export Extractor

Extractor export specification

Modules to be exported
☑ MyEmmy_actors_BasicFeatures
☑ MyEmmy_series_BasicFeatures

☑ Export dependent modules

Select the export destination directory
C:\Users\Lucas\Desktop

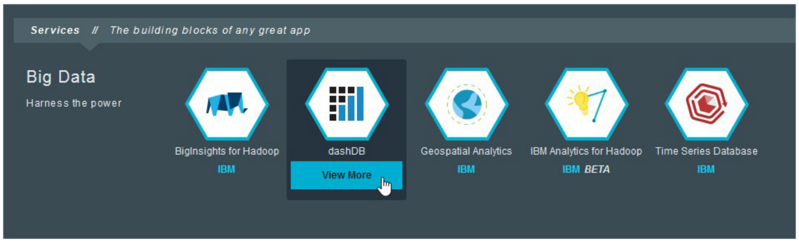
Options
○ Export to destination directory
● Export to a jar or zip archive under the destination directory
File Name
MyEmmyExtractor.zip

You should receive a confirmation message indicating that the process was successfully done. The steps in the text analytics workflow in Eclipse are complete. Additionally, you can upload your exported extractor to the HDFS on a Hadoop cluster to run it using much more data as input. To learn more, see the BigInsights text analytics documentation .

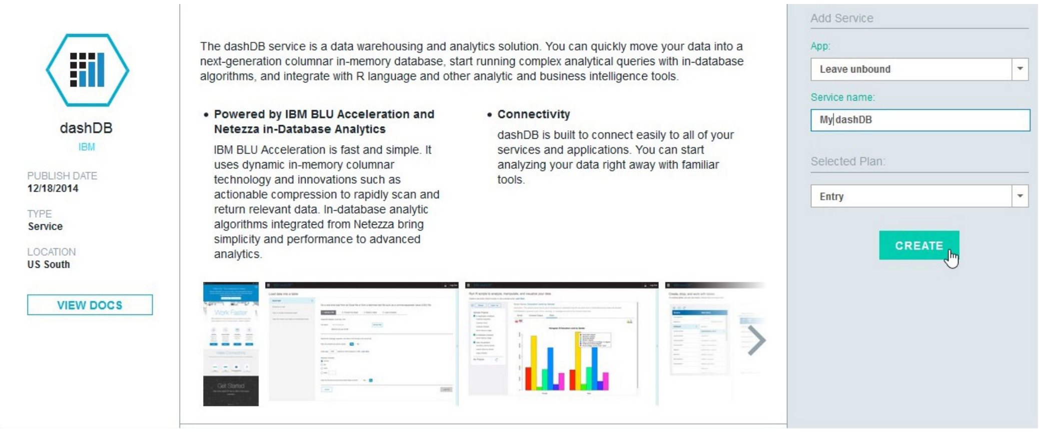
Analysis of results using R

We will use R included in the dashDB service in Bluemix to plot charts about the matches extracted in our data. Follow these steps to work with dashDB:

1. From the Bluemix dashboard, click **Catalog** in the menu, scroll down to the Big Data section, and click **dashDB**.



2. Change the name to `My dashDB` and click **CREATE**.



dashDB
IBM

PUBLISH DATE
12/18/2014

TYPE
Service

LOCATION
US South

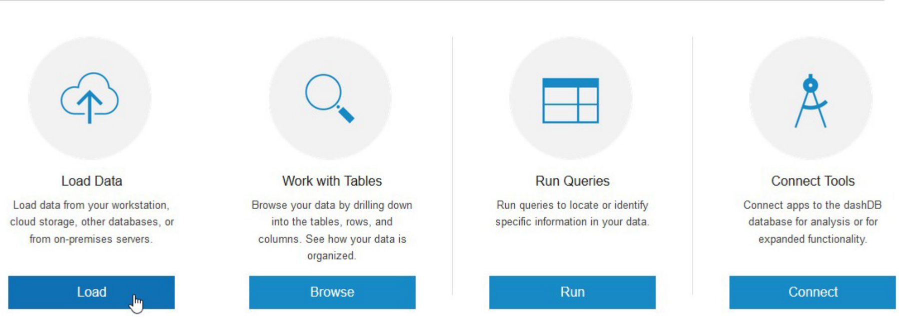
VIEW DOCS

The dashDB service is a data warehousing and analytics solution. You can quickly move your data into a next-generation columnar in-memory database, start running complex analytical queries with in-database algorithms, and integrate with R language and other analytic and business intelligence tools.

- **Powered by IBM BLU Acceleration and Netezza in-Database Analytics**
  IBM BLU Acceleration is fast and simple. It uses dynamic in-memory columnar technology and innovations such as actionable compression to rapidly scan and return relevant data. In-database analytic algorithms integrated from Netezza bring simplicity and performance to advanced analytics.
- **Connectivity**
  dashDB is built to connect easily to all of your services and applications. You can start analyzing your data right away with familiar tools.

Add Service
App: Leave unbound
Service name: My dashDB
Selected Plan: Entry
CREATE

3. The dashDB service is created.
4. Click **LAUNCH**.
5. The dashDB service is running. As input, use the two CSV files obtained in The text analytics workflow in Eclipse. One file contains the list of actors and the other file contains the list of series. Load these datasets to dashDB. In **Load Data**, click **Load**.



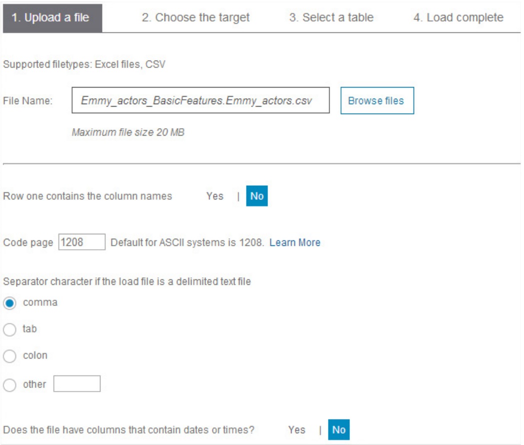
Load Data
Load data from your workstation, cloud storage, other databases, or from on-premises servers.
Load

Work with Tables
Browse your data by drilling down into the tables, rows, and columns. See how your data is organized.
Browse

Run Queries
Run queries to locate or identify specific information in your data.
Run

Connect Tools
Connect apps to the dashDB database for analysis or for expanded functionality.
Connect

6. Locate the input CSV files on your computer. The files do *not* include a header row. We show you how to do this for one file. Repeat the process for the other one.



1. Upload a file 2. Choose the target 3. Select a table 4. Load complete

Supported filetypes: Excel files, CSV

File Name: Emmy_actors_BasicFeatures.Emmy_actors.csv Browse files

Maximum file size 20 MB

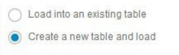
Row one contains the column names Yes | No

Code page 1208 Default for ASCII systems is 1208. Learn More

Separator character if the load file is a delimited text file
● comma
○ tab
○ colon
○ other

Does the file have columns that contain dates or times? Yes | No

7. Scroll down and click **Load File**.
8. Click **Next**. Select **Create a new table and load**.



○ Load into an existing table
● Create a new table and load

9. Name the only column in the dataset as **MyColumn**, the table as **Actors**, and click **Finish**. The name of your table cannot have spaces. You will need this name soon.

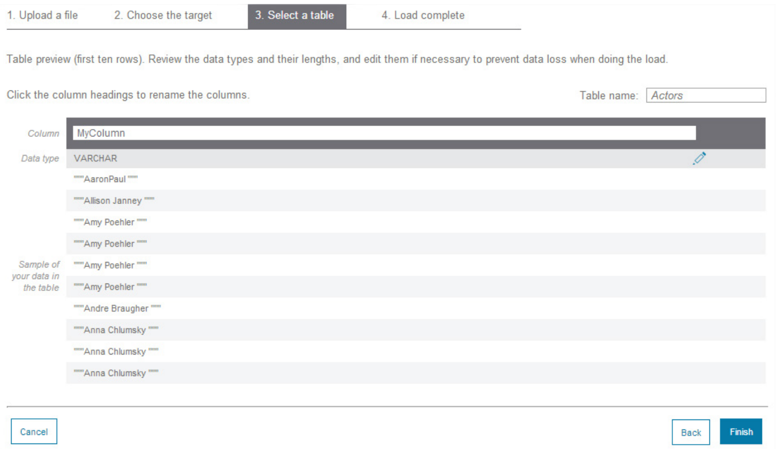


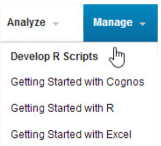
Table preview (first ten rows). Review the data types and their lengths, and edit them if necessary to prevent data loss when doing the load.

Click the column headings to rename the columns.

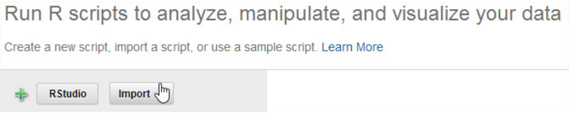
Table name: Actors

Column	MyColumn
Data type	VARCHAR
Sample of your data in the table	"""AaronPaul """
	"""Allison Janney """
	"""Amy Poehler """
	"""Amy Poehler """
	"""Amy Poehler """
	"""Amy Poehler """
	"""Andre Braugher """
	"""Anna Chlumsky """
	"""Anna Chlumsky """
	"""Anna Chlumsky """

10. With the data loaded to dashDB, let's perform some analysis using R. On the menu, choose **Analyze > Develop R Scripts**.



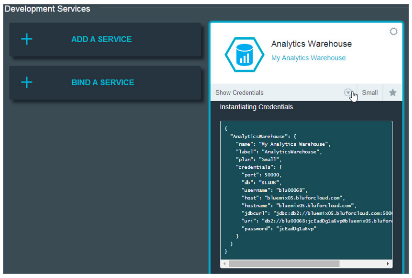
11. Now that you know how to load data, repeat the process to get the dataset with the series' name.
12. Upload the R Script. Click **Import** and locate the file in your computer.



Run R scripts to analyze, manipulate, and visualize your data

Create a new script, import a script, or use a sample script. Learn More

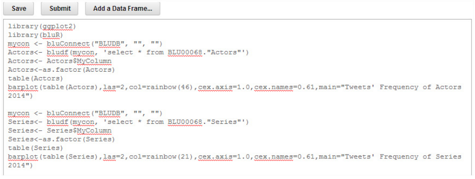
13. The script is loaded and shown on the screen. Edit the script so the connection string uses the appropriate user ID and password. This information can be obtained from the Bluemix VCAP services. In the DashDB menu, click **Set Up > Connect Applications**. In the Connection settings section, the Bluemix VCAP services information includes the user ID. Copy this user ID.



Instantiating Credentials

```
{
  "AnalyticsWarehouse": {
    "name": "My Analytics Warehouse",
    "label": "AnalyticsWarehouse",
    "plan": "Small",
    "credentials": {
      "port": 50000,
      "db": "BLUDB",
      "username": "blu00068",
      "host": "bluemix05.bluforcloud.com",
      "hostname": "bluemix05.bluforcloud.com",
      "jdbcurl": "jdbc:db2://bluemix05.bluforcloud.com:500
      "uri": "db2://blu00068:jcEadDg1a6vp@bluemix05.blufor
      "password": "jcEadDg1a6vp"
    }
  }
}
```

14. Go back to the R script and paste the username in the appropriate location (on the fourth line in the following screen capture. (In this case, the user ID is dash101859). Between quotes, put the name of the table.



```
library(ggplot2)
library(bluR)
mycon <- bluConnect("BLUDB", "", "")
Actors<- bludf(mycon, 'select * from BLU00068."Actors"')
Actors<- Actors$MyColumn
Actors<-as.factor(Actors)
table(Actors)
barplot(table(Actors),las=2,col=rainbow(46),cex.axis=1.0,cex.names=0.61,main="Tweets' Frequency of Actors
2014")

mycon <- bluConnect("BLUDB", "", "")
Series<- bludf(mycon, 'select * from BLU00068."Series"')
Series<- Series$MyColumn
Series<-as.factor(Series)
table(Series)
barplot(table(Series),las=2,col=rainbow(21),cex.axis=1.0,cex.names=0.61,main="Tweets' Frequency of Series
2014")
```

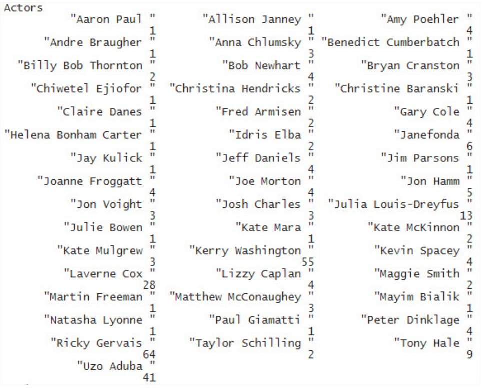
For the actors information, the fourth line of code reads:

```
Actors<- bludf(mycon, 'select * from dash101859, "Actors"')
```

For the series information, the eleventh line of code reads:

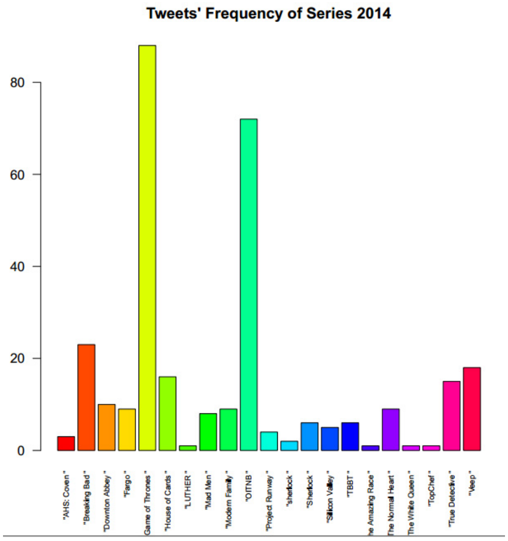
```
Series<- bludf(mycon, 'select * from dash101859, "Series"')
```

15. Click **Save** and then **Submit**. The Console Output shows a list of actors:



```
Actors
          "Aaron Paul "      "Allison Janney "          "Amy Poehler "
                      1                      1                       4
      "Andre Braugher "       "Anna Chlumsky " "Benedict Cumberbatch "
                      1                      3                       1
  "Billy Bob Thornton "         "Bob Newhart "       "Bryan Cranston "
                      2                      4                       3
    "Chiwetel Ejiofor " "Christina Hendricks "   "Christine Baranski "
                      1                      2                       1
        "Claire Danes "        "Fred Armisen "            "Gary Cole "
                      1                      2                       4
"Helena Bonham Carter "          "Idris Elba "            "Janefonda "
                      1                      2                       6
          "Jay Kulick "        "Jeff Daniels "          "Jim Parsons "
                      1                      4                       1
     "Joanne Froggatt "          "Joe Morton "             "Jon Hamm "
                      4                      4                       5
          "Jon Voight "        "Josh Charles "  "Julia Louis-Dreyfus "
                      3                      3                      13
         "Julie Bowen "           "Kate Mara "        "Kate McKinnon "
                      1                      1                       2
        "Kate Mulgrew "    "Kerry Washington "         "Kevin Spacey "
                      3                     55                       4
         "Laverne Cox "        "Lizzy Caplan "         "Maggie Smith "
                     28                      4                       2
      "Martin Freeman " "Matthew McConaughey "         "Mayim Bialik "
                      1                      3                       1
      "Natasha Lyonne "       "Paul Giamatti "       "Peter Dinklage "
                      1                      1                       4
       "Ricky Gervais "    "Taylor Schilling "            "Tony Hale "
                     64                      2                       9
           "Uzo Aduba "
                     41
```

16. On the Plots tab is an icon for a PDF file that contains a plot of the desired information. Click the icon to download the PDF file. This file includes a plot of the Emmy series data and Emmy actors data. The following screen capture was cropped to show only the frequency of Tweets for the series nominations in 2014.



Our goal was to know whom people were cheering for in Emmy Awards 2014. So, according to this graph, we can easily observe three main crowds: *Game of Thrones*, *Orange Is the New Black*, and *Breaking Bad*. This graph shows that over 80 tweets were about *Game of Thrones*. We see, still by this graph, that people have expectations of the awards. For example, they do not care much about reality TV shows, such as *Top Chef* and *The Amazing Race*. It is good to remind you that this data was collected during the first four days after the official nomination.

Conclusion

In this tutorial, we demonstrated how to create a text analytics program and some analytics based on a simple example of the Emmy Awards. You can use what you have learned here and apply in more complex examples based on your needs.

Downloads

Description	Name	Size
Code file	codigo.txt	1KB
CSV files	csvfiles.zip	2KB
Data files	dataFiles.zip	63KB
Dictionaries	dictionaries.zip	2KB

Resources

Learn

- Learn more about text analytics in the Knowledge Center.
- Learn more about the InfoSphere BigInsights text analytics.
- Learn more about the supported languages by annotators and dictionaries.
- Learn more about the extract statement.
- Learn more about writing and testing AQL.
- Learn more about the guidelines for writing AQL.
- Learn more about the AQL Profiler.
- Get Eclipse Classic 4.2.2 .
- Visit the developerWorks Information Management zone to find more resources for DB2 developers and administrators.
- Stay current with developerWorks technical events and webcasts focused on a variety of IBM products and IT industry topics.
- Follow developerWorks on Twitter.

Get products and technologies

- Evaluate software in the way that suits you best: Download a product trial, try a product online, or use a product in a cloud environment.

Discuss

- Get involved in the developerWorks community to connect with other developerWorks users while exploring the developer-driven blogs, forums, groups, and wikis.

About the authors

Lucas Silva

Lucas Lopes da Silva is a software engineer at Big Data University and fourth-year computer science student currently on a scholarship program at the University of Toronto.

Felipe Mosquetta

Felipe Henrique Mosquetta Oliveira is a business analyst intern at BDU and fourth-year statistics student on scholarship at the University of Toronto.

Marcos Rogério De Mello

Marcos Rogério de Mello is a third-year computer science student, an international student on scholarship by the Science Without Borders Program at the University of Toronto, and an intern at Big Data University.