

# BigSheets

*Hands-On Lab Guide*



## Table of Contents

<b>Analyzing social media and structured data using BigSheets.....</b>	<b>3</b>
1.1 Introduction .....	3
1.2 Collecting Social Media and Structured Data .....	3
1.3 Creating a BigSheets Workbook .....	6
1.3.1 <i>BigSheets from your RDBMS data</i> .....	6
1.3.2 <i>Creating a Workbook from Boardreader Data</i> .....	7
1.3.3 <i>Tailoring Your Workbooks</i> .....	10
1.3.4 <i>Exploring Your Workbook</i> .....	17
1.4 Digging Deeper: Filtering results and Extracting URL data.....	22
1.4.1 <i>Filtering Data</i> .....	22
1.4.2 <i>Extracting URL data</i> .....	23
1.5 Combining Social Media with RDBMS Data.....	29
1.5.1 <i>Running a built-in Function to extract data</i> .....	29
1.5.2 <i>Performing a Pivot or “Group by” Operation</i> .....	33
1.5.3 <i>Joining Social with Structured data</i> .....	36
1.6 Dashboard.....	39
1.7 Summary .....	42
1.8 Appendix.....	42
1.8.1 <i>Collecting Social Media Data using BoardReader</i> .....	42
1.8.2 <i>Working with RDBMS</i> .....	46

## Analyzing social media and structured data using BigSheets

### 1.1 Introduction

IBM's InfoSphere BigInsights 2.1 Enterprise Edition enables firms to store, process, and analyze large volumes of various types of data. In this exercise, you'll see how you can find and ingest Social Media Data from Boardreader and do some ad hoc data exploration with BigSheets.

#### What you will learn in this section:

After completing this hands-on lab, you'll be able to:

- Create a Big Sheets workbook in order to process some of the Boardreader input data that has been loaded into the HDFS for you.
- Connect BigInsights to DB2 in order to transfer data.
- Create Big Sheets visualizations to see some of the quantitative analytics.
- Allow 90 minutes to 2 hours to complete this lab.

Service	Username	Password
Linux	root	Password
BigInsights Web Console	biadmin	Password

### 1.2 Collecting Social Media and Structured Data

1. Select the **Start BigInsights** icon to start all services



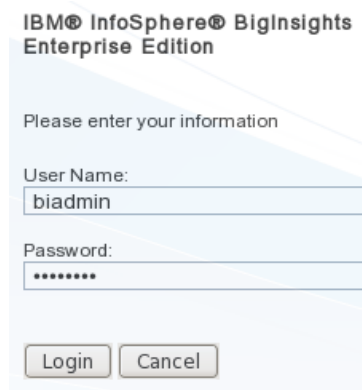
#### What is this program actually doing?

If BigInsights is not running, this will launch all services through the start-all.sh command in the \$BIGINSIGHTS\_HOME/bin directory. (This is typically /opt/ibm/biginsights/bin or /data/opt/ibm/biginsights/bin.)

2. Launch the BigInsights web console



- The BigInsights VMware image was configured with security enabled. So, you will be prompted to enter a user ID and password. Enter the user ID and password of biadmin/passw0rd as the userid/password. Click Login to log into the system.



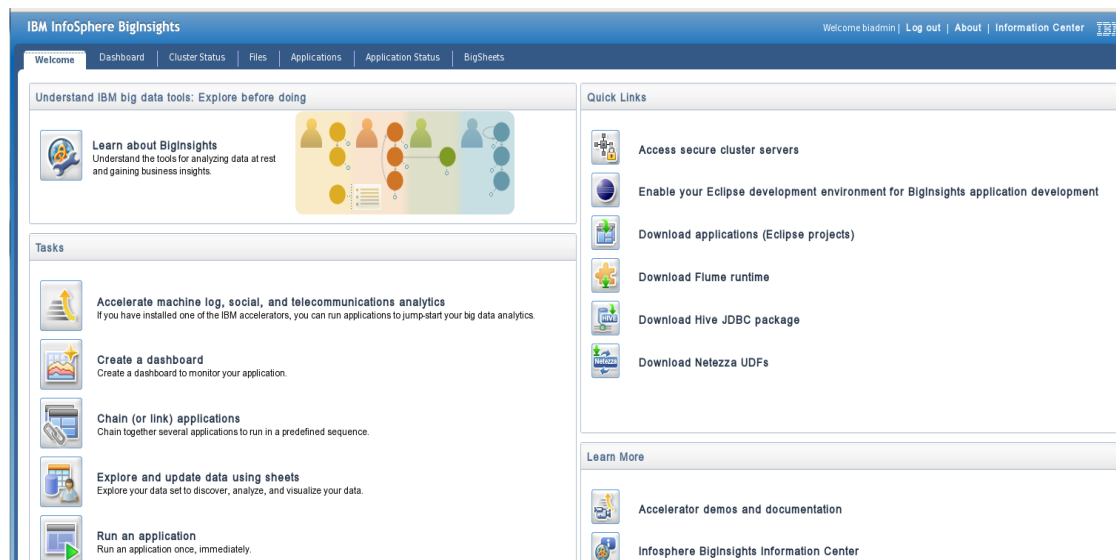
IBM InfoSphere BigInsights  
Enterprise Edition

Please enter your information

User Name:

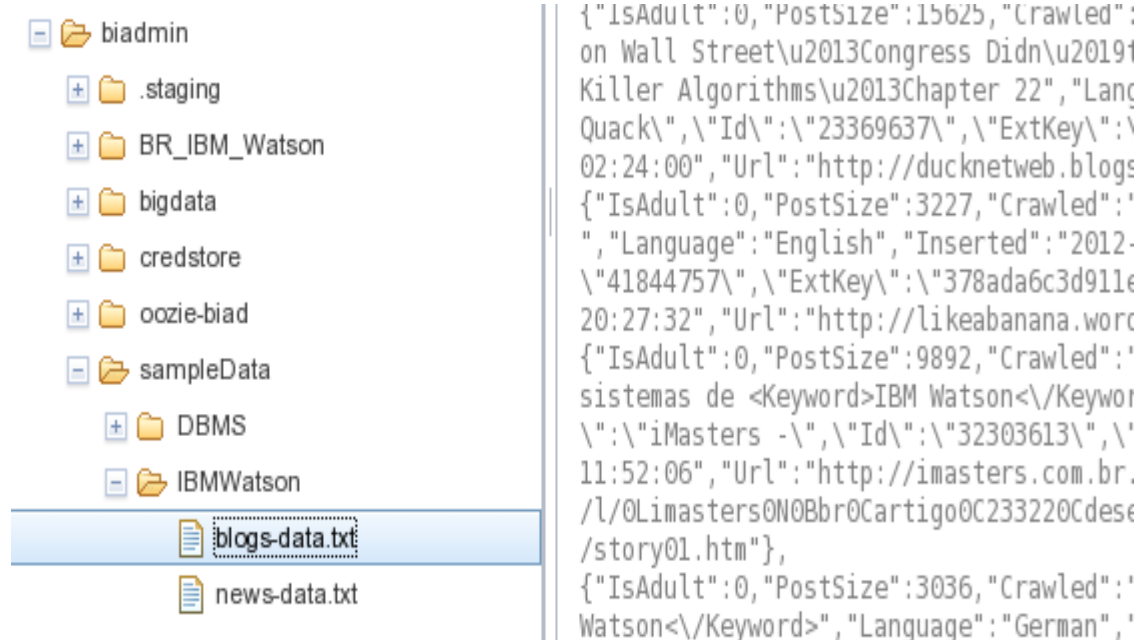
Password:

- Verify that your Web console appears similar to this:



The social media output that will be used for this lab has been preloaded into the HDFS. This data was retrieved using the Boardreader app. To learn how to use the Boardreader app to pull social media data, refer to the Appendix, Section 1.8.1.

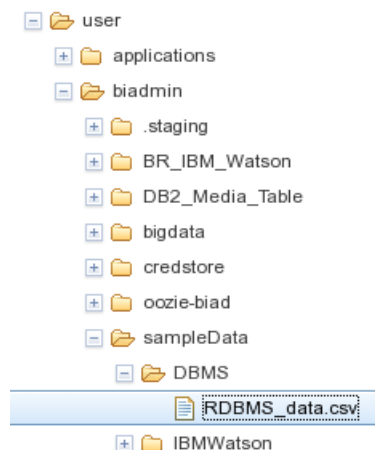
- To review this data, use the Files tab to navigate to the following folder (/user/biadmin/sampleData/IBMWatson) and select the blogs-data.txt file as shown below.



6. In a future section, we will display this text file within BigSheets in order to display, interact with, and visualize the data within BigInsights.

Many customers want to join data from their Big Data environment with data in an RDBMS. For this lab, the structured data has been preloaded into the HDFS, to learn how to import data from DB2 using a built-in app, refer to the Appendix, Section 1.8.2.

7. To review this data, use the Files tab to navigate to the following folder (/user/biadmin/sampleData/DBMS) and select the RDBMS\_data.csv file as shown below.



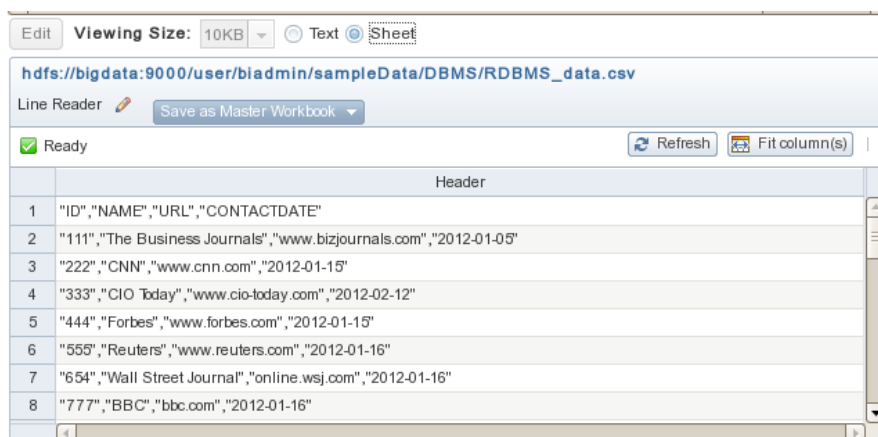
8. You should now see the data on the right-hand slide. We will work with this data in the next section.

### 1.3 Creating a BigSheets Workbook

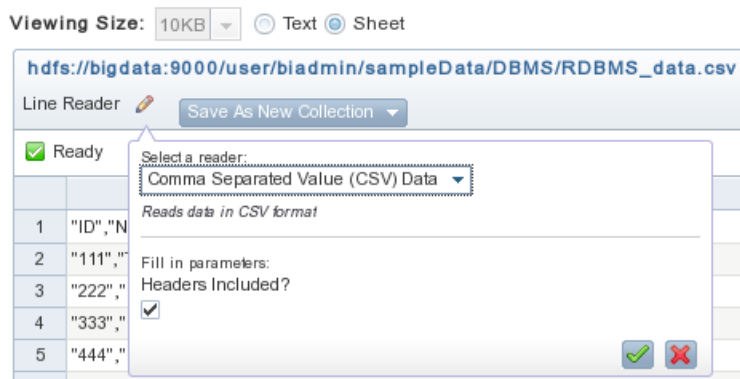
In this section, we will use the spread-sheet style interface known as BigSheets. BigSheets provides access to data in items known as “workbooks”.

#### 1.3.1 BigSheets from your RDBMS data

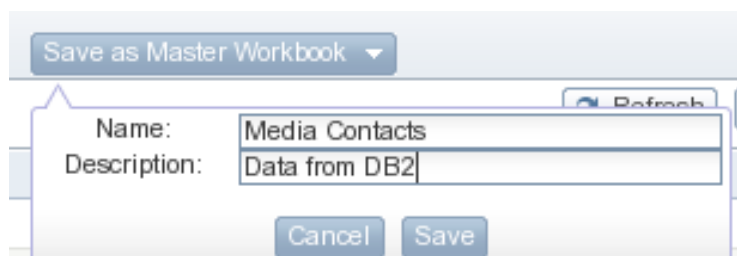
Starting from the file RDBMS\_data.csv shown in Section 1.2 Step 7, click on the Sheet radio button.



9. You will then click on the pencil icon to select the appropriate reader for this file. Select the reader: Comma Separated Value (CSV) Data from the pull-down menu.



10. Click the green check-mark to see the data represented with the CSV reader.
11. Click the “Save As Master Workbook” button, input a Name, for example “Media Contacts” and optionally a description. Click the Save button to save this data to the BigSheets tab.



12. The data should now be shown in a spreadsheet under the BigSheets tab. At this point, this data is now in HDFS and we will use it later to combine social media data with this data from DB2 to gain additional insights into our web-based marketing campaign results.

### 1.3.2 Creating a Workbook from Boardreader Data

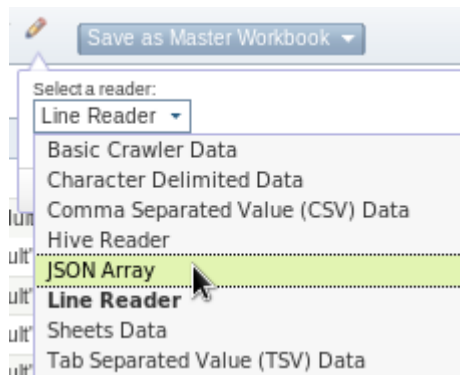
13. Return to the Files tab.
14. Navigate to the /user/biadmin/sampleData/IBMWatson/blogs-data.txt file. Make sure this file is selected.



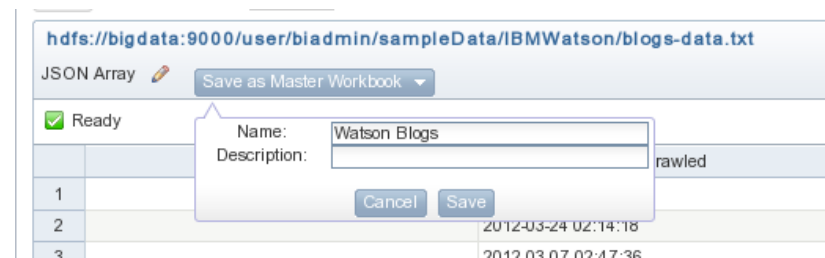
15. Again, click the Sheet radio button to view this data within a BigSheets interface.



16. The data that comes from the Boardreader application within BigInsights is formatted in a JSON Array structure. You will need to click on the pencil icon and select the JSON Array option for this file.



17. Save this as a Master Workbook named **“Watson Blogs”**. You can optionally provide a description. Click the Save button.

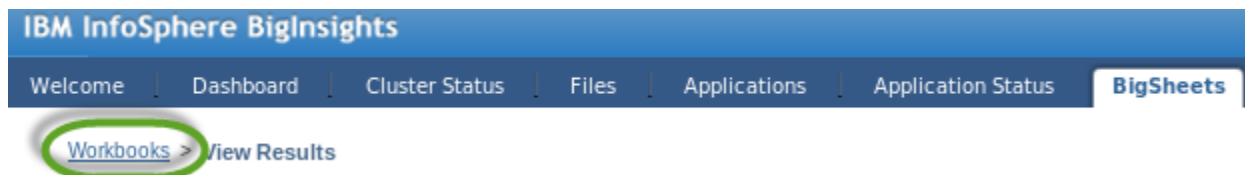


18. You will want to do the same thing for the news-data.txt file in the same folder. So, to do this, you will need to return to the Files tab, navigate to the file, and follow the same process. This time, you will provide the name **“Watson News”**.

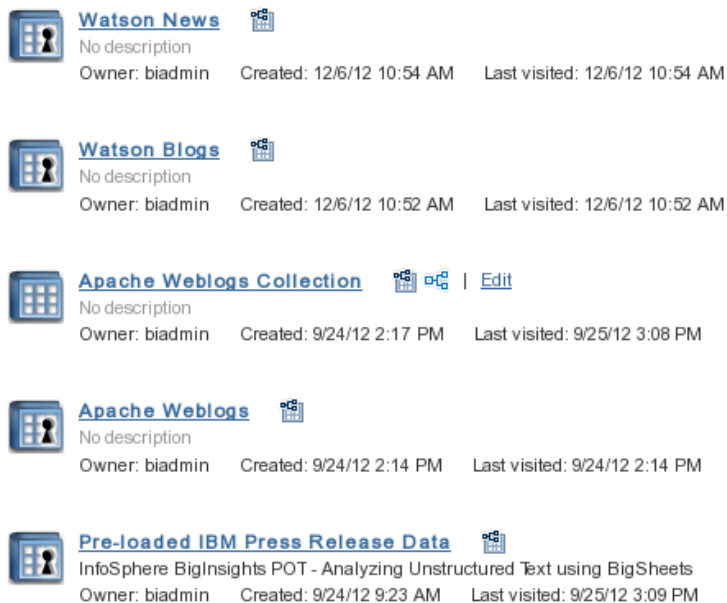




19. Click on the “Workbooks” link in the upper left-hand corner of the page.

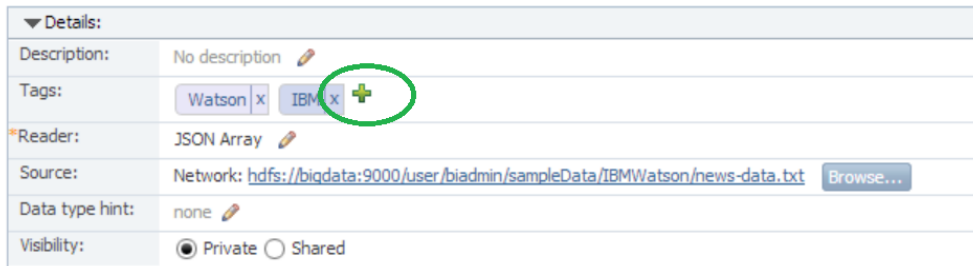


20. You should now see (2) or more workbooks on your system, depending on the configuration of your VM. The top (2) are the ones you created in the previous steps.

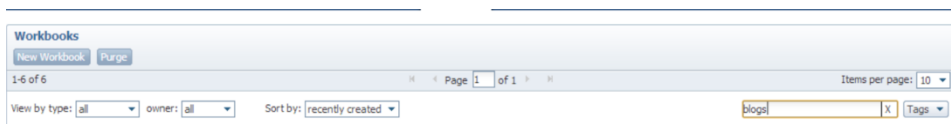


21. Adding tags allows us to quickly search and manage workbooks. Select the “Watson Blogs” workbook.

22. Under Workbook Details, Add the following tags “**Watson**” “**IBM**” “**Blogs**” by selecting the green + and adding each individually.



23. From the BigSheets tab, you can quickly filter workbooks and search for a specific tag. Enter the term “tag: Blogs” to see all workbooks that have the associated tag.

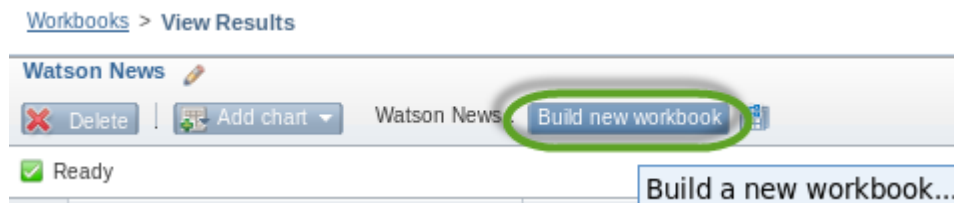


### 1.3.3 Tailoring Your Workbooks

In this section, we will first show you how to reduce the amount of data you are working with by removing unwanted columns from within your BigSheets Workbooks. We will also perform a merge or “union” of two data workbooks.

24. From the list of Workbooks (you launched in the previous steps), click on the link named “Watson News” to open this workbook in BigSheets.
25. This BigSheets Master Workbook is a “base” workbook and has a limited set of things you can edit. Therefore, in order to begin to manipulate the data contained within a workbook, we will want to create a dependent workbook.

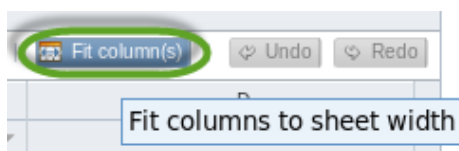
- Click the “Build new Workbook” button



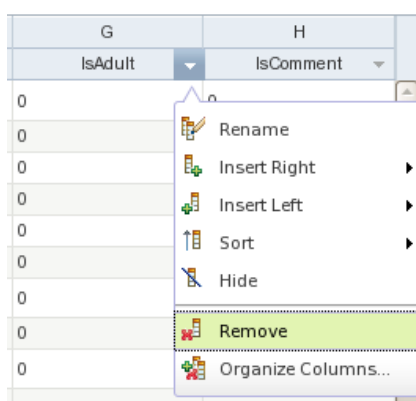
- When the new Workbook appears, you can change its default name (by clicking on the pencil icon next to the name) to the new name of “**Watson News Revised**”.



- Be sure to click the green check-mark button to actually change your new workbook's name.
- Click the Fit Columns button to more easily see columns A through H on your screen.



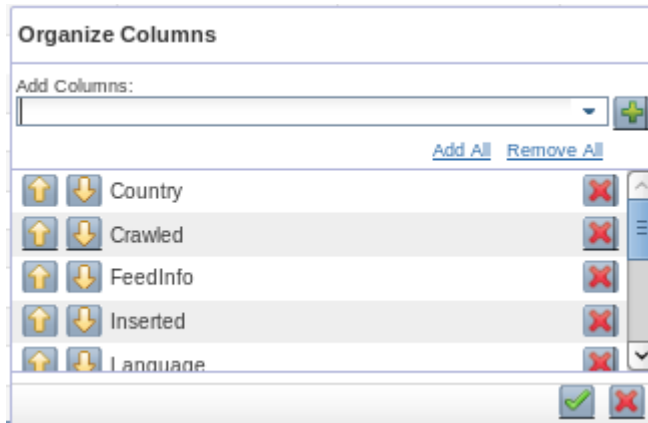
26. You have decided you would like to remove the column “IsAdult” from your workbook. This is currently column E. Click on the triangle next to the column name of “IsAdult” and select the “Remove” option to remove this from your new workbook.



#### Did I lose data?

Deleting a column does not remove data. Deleting a column in a workbook just removing the mapping to this column.

27. In this case, you want to keep only a few columns. In order, to more easily remove a larger number of columns (without having to do this same click-remove process), click the triangle again (from any column) and select the “Organize Columns...” option.

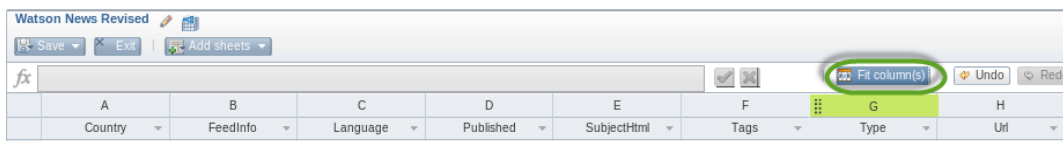


- Click the red X button next to each column title you want to remove.

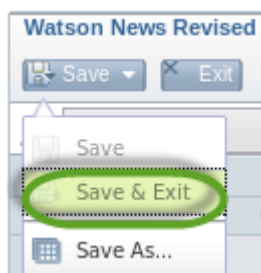
In this case, **KEEP** the following columns...

- Country
- FeedInfo
- Language
- Published
- SubjectHtml
- Tags
- Type
- Url

- Click the green check mark button when you are ready to remove the columns you have selected to remove.
28. Click on the Fit Columns button again to show columns A through H. You should see the following columns in your new workbook.



29. “Save and Exit.” The system will ask for an optional Description. Click Save to complete the Save.



30. After clicking Save, you will be shown two buttons (run and close). Click the Run button to run the workbook. You can monitor the progress of your request by watching the status bar indicator in the upper right-hand side of the page.

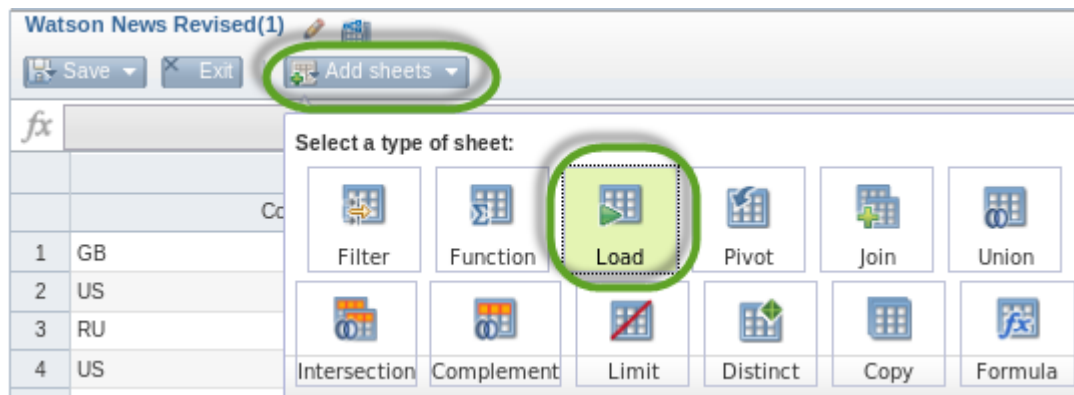


31. To reduce the unwanted columns in the “Watson Blogs” workbook, you will want to perform the same steps above in order to wind up with a new workbook called “Watson Blogs Revised”.

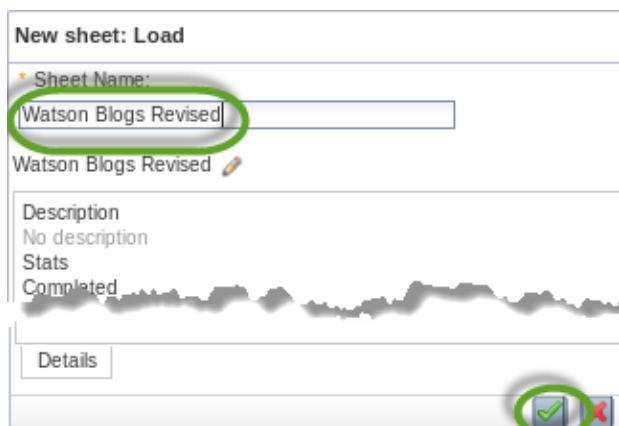
 A screenshot of a spreadsheet titled "Watson Blogs Revised". It shows a table with columns labeled A through H. The first row of data contains the following values: "P>Title", "CitySkv Wallnai Finnish", "2012/05/19 09:43:00", "First Academic Case Cn", "hlon", and "http://enhuokwai".
 

	A	B	C	D	E	F	G	H
	Country	FeedInfo	Language	Published	SubjectHtml	Tags	Type	Url
1	P>Title	CitySkv Wallnai Finnish		2012/05/19 09:43:00	First Academic Case Cn		hlon	http://enhuokwai

32. Now, since we have two workbooks with the exact same structure, we can perform a “union” of these two workbooks as the basis for exploring the coverage of IBM Watson across the sources that Boardreader provided.
33. To perform this action, make sure you are currently in the “Watson News Revised” workbook. Click the “Build New Workbook” button again.
34. In the top left-hand side or bottom left, you should see a link called “**Add sheets**”. This allows you to perform additional analysis on your data within the current workbook. Click the “Add sheets” link.



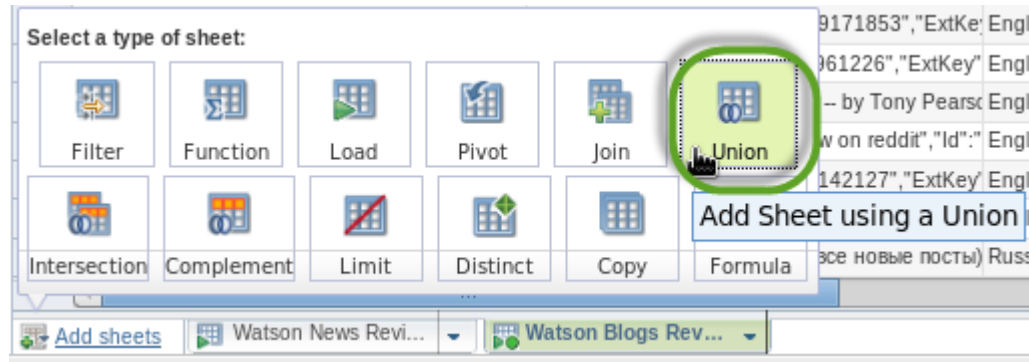
35. The Load option will allow you to load data into the current workbook from another workbook. Click the Load icon and select the “Watson Blogs Revised” workbook link.
36. The system will ask you for a \*Sheet Name and you should change Sheet1 to “Watson Blogs Revised” as the name of the new tab that will be created in your current workbook.



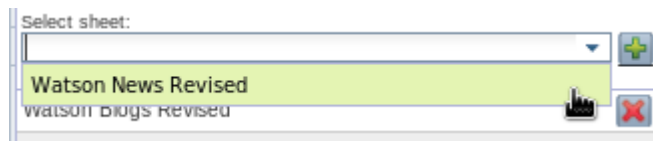
37. Click the green check-mark button at this time to load the new workbook into your current workbook.
38. You should now see two tabs at the bottom on your current workbook. If you move your mouse over the second one, a tooltip will show the action and the name you provided for this sheet / tab within you current workbook. (Giving your tabs meaningful names will help you and other that use your sheets an easy way to understand your data processing flow(s).)



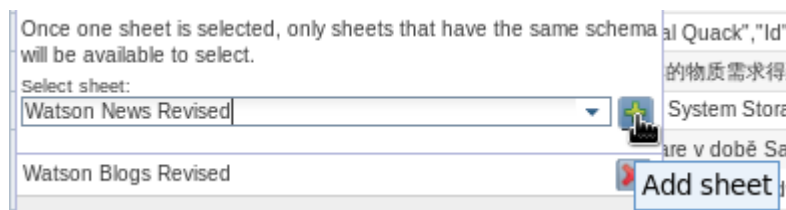
39. Now, we want to add another sheet in order to perform the Union function.



40. The Union function asked for the other “sheet” you would like to use. Select the triangle to expose the pull-down menu.



41. Select the item and then click the green plus-mark button.




42. Then provide a new \*Sheet Name. Before you click the green check-mark button to add this new tab/sheet to your workbook, make sure your options match the example below...


**New sheet: Union**

\* Sheet Name:

Once one sheet is selected, only sheets that have the same schema will be available to select.

Select sheet:

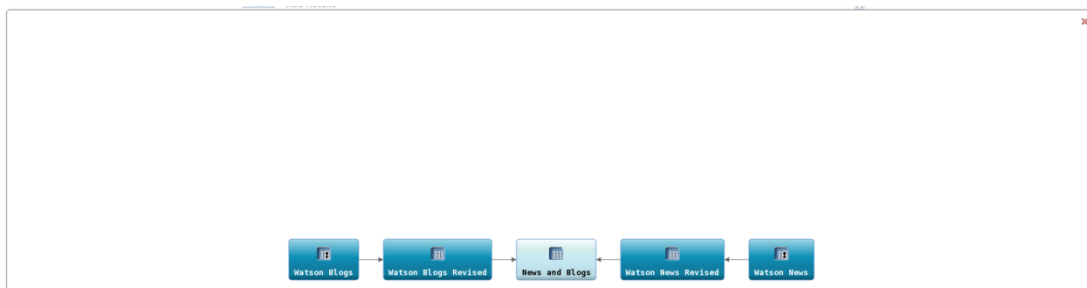
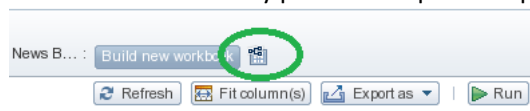
Watson Blogs Revised 

Watson News Revised 

43. Click the green check-mark button to add this tab to your workbook.



44. Save and Exit and then run this new workbook. When prompted for a Description, you can change the name of your new workbook from “Watson News Revised(1)” to “Watson News and Blogs”. Click the Save button. Then click the Run button to run the workbook.
45. Select the Workflow Diagram icon to see a mapping of the workbooks associated with the News and Blogs workbook. This can be done at any point to keep a clear picture of which workbooks you are extending to/from.



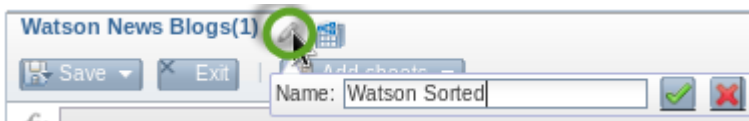


46. Close this frame and you can continue to the next section.

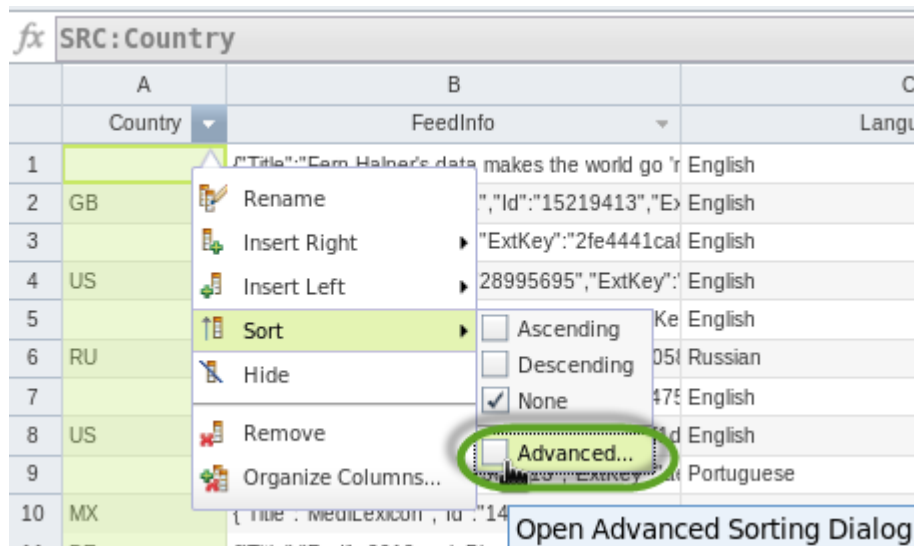
### 1.3.4 Exploring Your Workbook

In this section, we will explore the data in your new workbook. You will perform actions like sorting, charting, cleansing, and grouping to make analysis of the HDFS data easier via BigSheets.

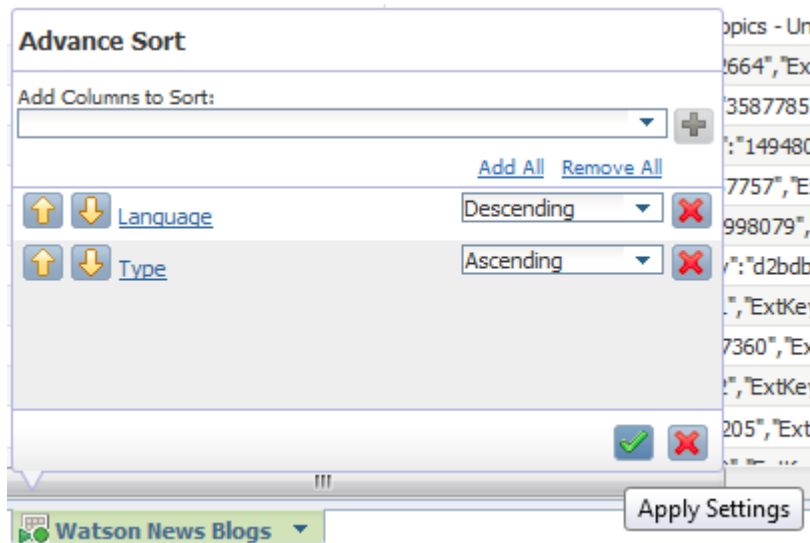
47. If you are not already in the workbook, open the “**Watson News and Blogs**” workbook.
48. In this case, we want to keep our initial workbook “as is” and produce another workbook that contains the records in sorted order. So, click the “**Build New Workbook**” button to do this.
49. As a way to keep track of what you are doing, I typically suggest that you rename your new workbook right away as this helps to remind you of what you are working on. So, provide a new name for your workbook, like “**Watson Sorted**”.



50. After looking at the data, you decide you would like to understand more about the language and the types of posts contained in your workbook.
51. Clicking on the triangle next to the column name of any column, you will want to click on the Sort -> Advanced option.



52. Then, you will click on the pull-down triangle to expose the list of columns under the “Add Columns to Sort” area. Click on the green plus-sign button to add the two columns you wish to sort on. Then, select the desired order for sorting each column. In this case, your “Advance Sort” should look like the following picture.



53. Click on the green check-mark button to continue and create the new tab/sheet with your desired sorting applied to it.
54. As with all new tabs/sheets, the system shows you a simulated result based on the rows of data BigSheets keeps in memory. You should be able to click on “Fit Columns” to review the contents of both the Language and Type columns to see that your “advanced sort” was applied to this simulated set of data.
55. Now, “save and exit” and then run your workbook. This will apply the sorting options to more than the first 2,000 rows the system operates on as a simulation. This will sort the entire, larger data in the workbook. So, you should see different results once your workbook has been run. For example, in the simulated data, only one Vietnamese row was showing. However, against the entire data set, you should see twenty (20) rows that are of the Vietnamese language. This is because more of the Vietnamese rows were in the data beyond the first 2,000 rows the system uses in memory for a simulated result before you click the run button. Review and confirm these results after the job reaches 100% and then you can move onto the next step.
56. To visualize this data better, you want to analyze the quantities of posts in the various languages in your current workbook. While still in the “Watson Sorted” workbook.
57. Click on the “Add chart” link in the lower left (this may take a minute to populate the chart types the first time it is run), click on chart, then pie, and fill out the following information to produce a pie chart of the languages used.

**New visualizer: Pie**

Chart Name:  
Language Coverage

Title:  
IBM Watson Coverage by Language

Value:  
Language

Count:  
Count occurrences of X axis values

Limit:  
12

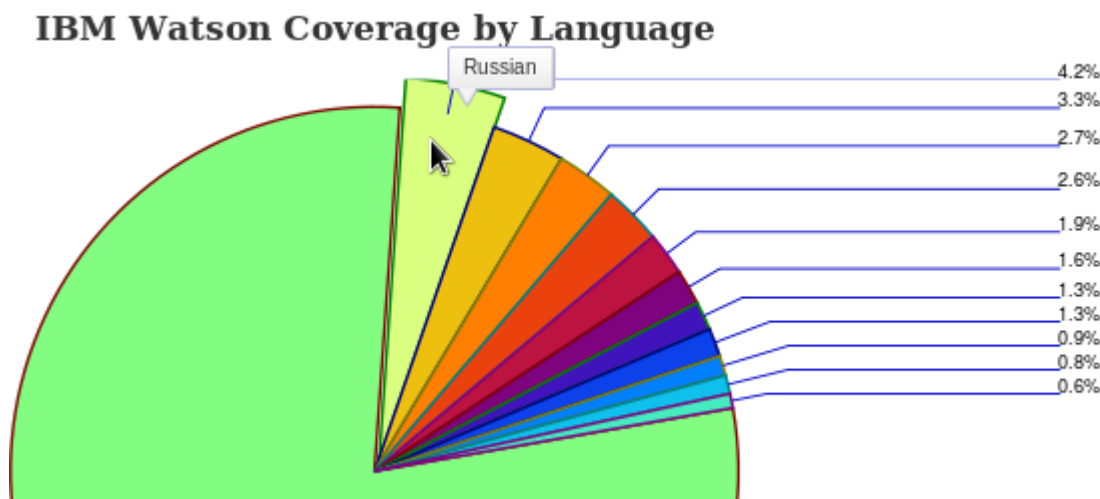
☐ ☐

58. Click on the green check-mark button to create the chart tab.

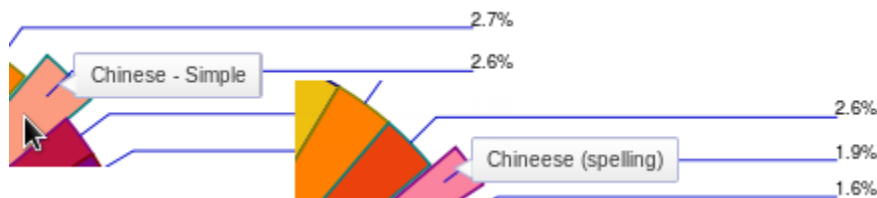
59. Just like working with tabular data, you will see a simulated visualization. Again, this is based on the rows in cache. (If you click on the Close button here, you can interact with the chart which is based on simulated data. You would then click the “run” button in the upper right.)



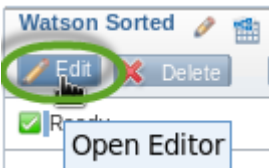
60. Click the run button to run the visualization against the entire data set.  
61. Once the chart has been run, you can interact with it to find out the second, most-popular language for posts regarding IBM Watson is Russian. Move your mouse over this item within your pie chart to see these results.



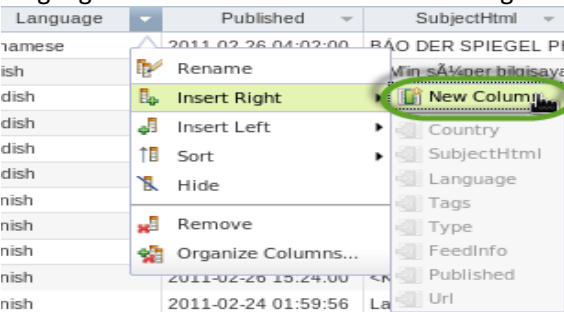
62. Additionally, you have decided you need to clean up some of the data within the workbook. Based on the fifth and sixth largest languages in the pie chart you just generated, you can see they are both variations on the Chinese language.



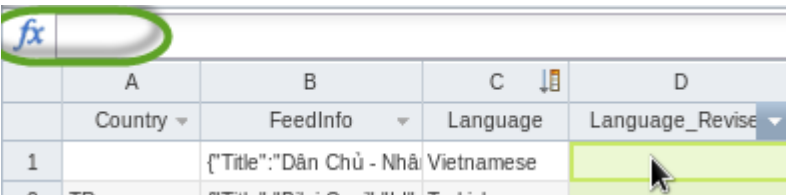
63. Let’s perform an action to clean up this data which will show all Chinese items in the same bucket. To do this, we will combine these two items into one item called Chinese.  
64. From the “Watson Sorted” workbook, you will want to click on the edit button in order to go back into edit mode for this workbook.



65. To make things easier, you can optionally click on the “Fit Columns” button to make your columns thinner and to see more data on the screen.
66. At this point, you have decided to add another column to your workbook. To do this, click on the triangle next to the Language column name. Select the Insert Right -> New Column option.



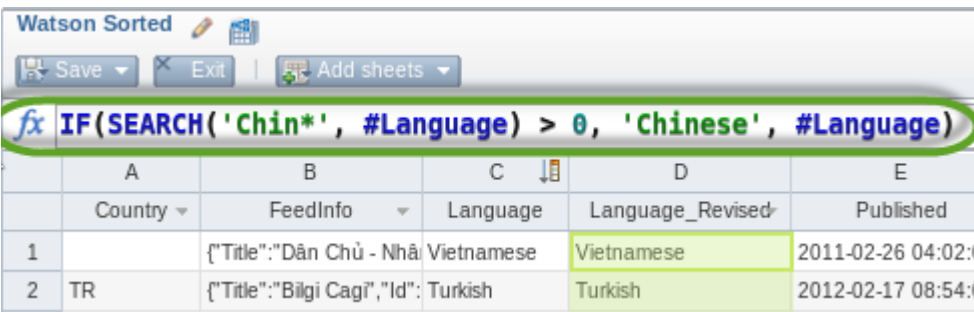
67. Then, you will provide a name for your new column, like “Language\_Revised” and then click the green check-mark button (or hit enter) to apply your new column name.
68. Your cursor is then moved to the fx (or function) area where you can provide the function to be used to generate the contents of your new column.



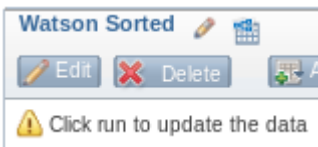
69. Enter the following formula as your function...  
IF(SEARCH('Chin\*', #Language) > 0, 'Chinese', #Language)

This formula looks at the Language column indicated by #Language. If the #Language column starts with 'Chin\*', then the new #Language\_Revised column will contain 'Chinese'. If it does not, the value of #Language is copied over to #Language\_Revised. (See the original article, URL at the top of this document, for additional explanation of this formula.)

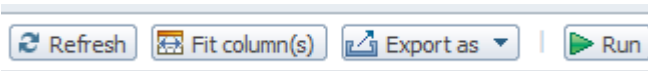
70. Clicking on the green check-mark button (or hitting enter) should produce content in your new column based on your new formula.



71. “Save and LanExit” and you should notice that you receive a message to “Click run to update the data”.

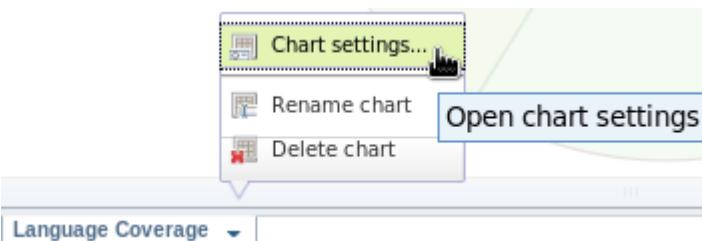


72. Click the run button in the upper right to run the workbook.



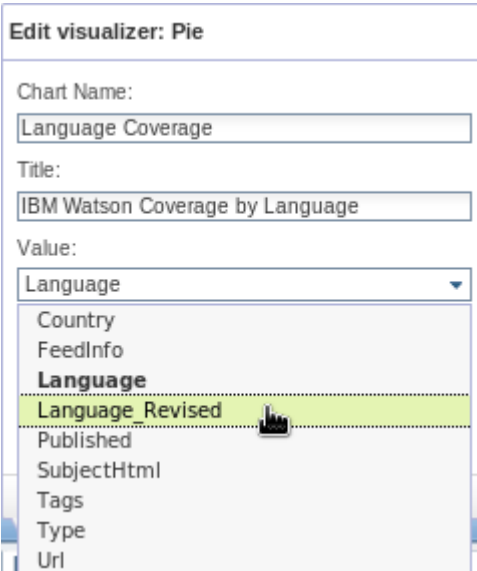
73. Now, click on the Language Coverage tab that contains your previously generated pie chart. This now has the status of “needs to be run”. Before we run it, we need to change one of the settings on the pie chart to use our newly generated column named Language\_Revised.

74. To change the settings, click on the triangle next to the Language Coverage tab.



75. Click to select the “Chart Settings” option.

76. Change the “Value:” item to be based on the new, Language\_Revised column.

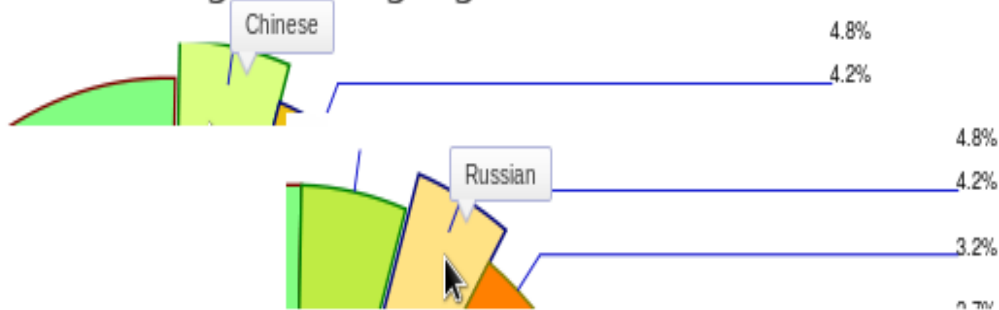


77. Click on the green check-mark button to apply your new settings.

78. Click on the run button to regenerate your pie chart.

79. Once your new pie chart has been generated, you should be able to see Chinese as a cleaned up, single item in your pie chart (compared to the two items you saw previously). With this cleansed data, Chinese is now the second largest and Russian is third.

## IBM Watson Coverage by Language



### 1.4 Digging Deeper: Filtering results and Extracting URL data

In this section, we'll derive a new workbook, based on the Watson Sorted workbook, that will only contain English-language records with URLs that end in .uk or a Country value of "GB" (for Great Britain). To do this, you will apply a filter and a function against the entire set of data in the workbook.

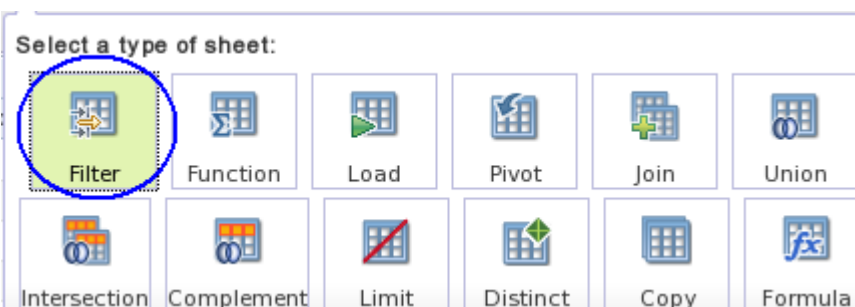
#### 1.4.1 Filtering Data

While viewing the "Watson Sorted" workbook, click on the "Build new workbook" button.

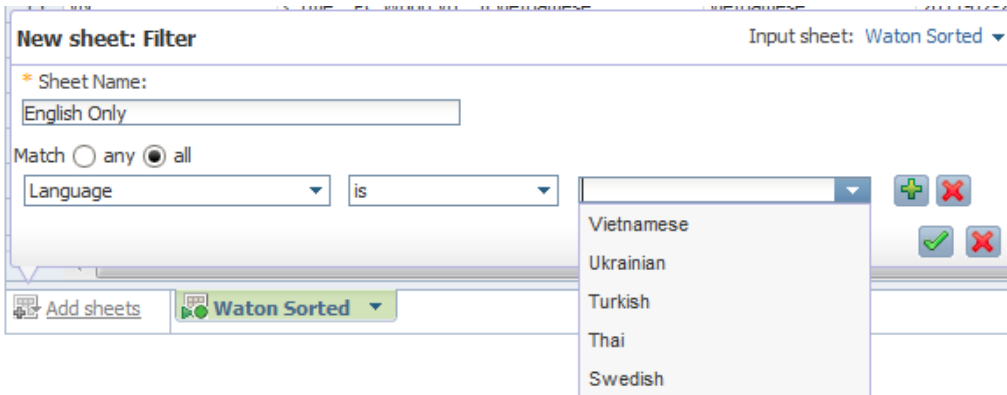
80. Again, as a way to make sure you are performing work on the proper workbook, you will want to rename your new workbook as a first step. You can use "Watson Sorted English UK" as a new name.



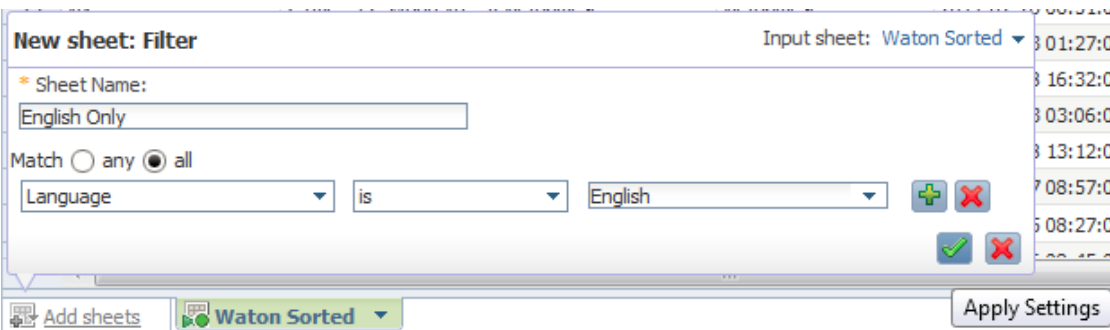
81. In the lower left, you should see the "Add sheets" link. We now want to add a Filter sheet based on a certain set of criteria.
82. Click Add sheets and then Filter.



83. In this case, you can name your new sheet, something like "English only" and then choose the column of Language to filter it by "is English". If you perform the pull down on the third option, you see only five (5) languages shown and English is not one of them. This only means that the system did not find the "English" value in the subset of data in the cache.



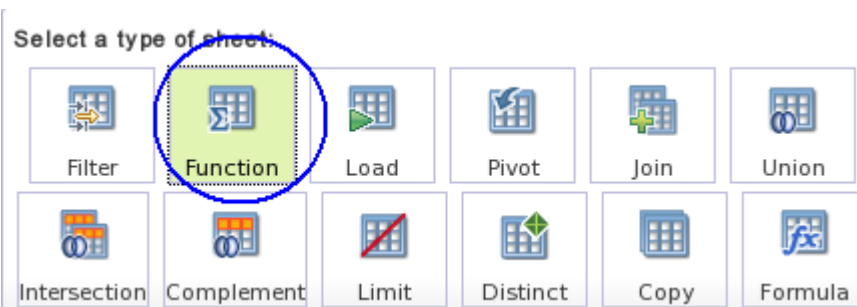
84. To provide the right value, you can actually type “English” (case sensitive) into the language area (as the value does not have to exist to perform the filter). So, type “English” into this field.



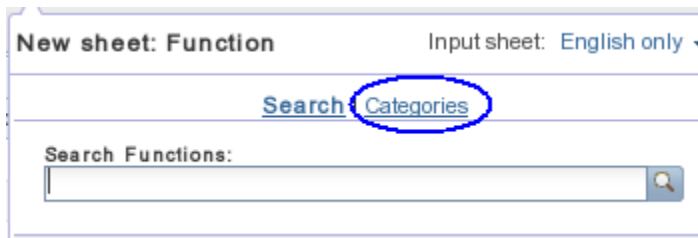
85. Click the green check-mark button to apply these settings to this new tab/sheet.  
86. You will see the rows in cache that match this new Filter setting.

#### 1.4.2 Extracting URL data

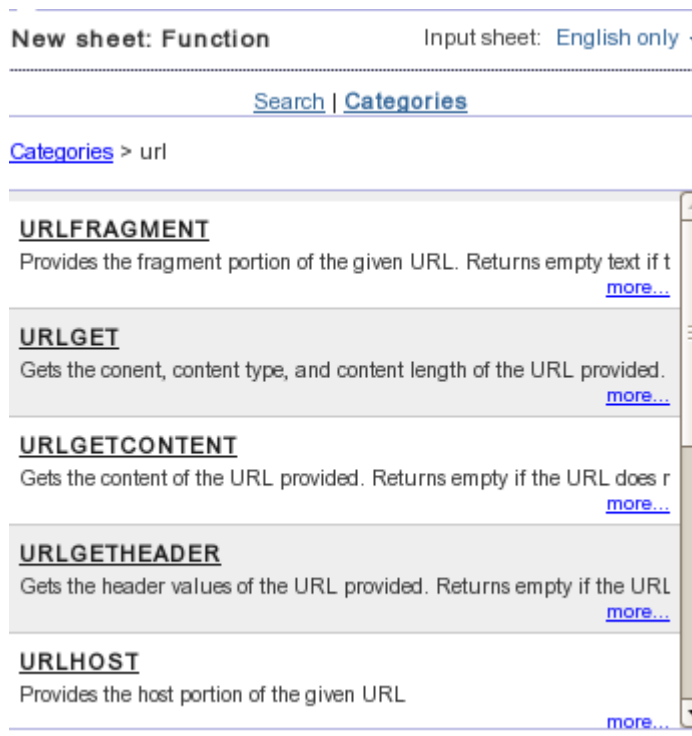
87. To build upon this new subset of data, we want to now perform an extraction to pull the Host portion of a URL into a new column.  
88. To do this, we need to add another sheet based on running a built-in function across the entire data workbook. Name it URLHOST.



89. By default, you are not shown any functions by name. Since the system currently ships with 96 functions, the system stays efficient and assumes you want to search for a particular function by name (if you know the name of the function) or you can look for functions by category.



90. Click on the Categories link and then click on the URL area to see the built-in URL.





91. Click on the URLHOST function. Fill out and select the information that the function requires.

**New sheet: Function** Input sheet: [URLHOST](#)

\* Sheet Name:

**URLHOST** 

Provides the host portion of the given URL

Fill in parameters:

url\*

Parameters Carry over (0)

92. Before clicking on the green check-mark button, there are additional items that need to be specified on the “Carry over (0)” tab. Click on the Carry over tab and click the “add all” link to “carry over” all of the columns from the “English Only” tab into your new URLHOST tab.

93. Your function specification should look like this now.

**New sheet: Function** Input sheet: URLHOST

---

\* Sheet Name:

URLHOST

**URLHOST**

Provides the host portion of the given URL

Add columns to carry over:

[Add all](#) [Remove all](#)

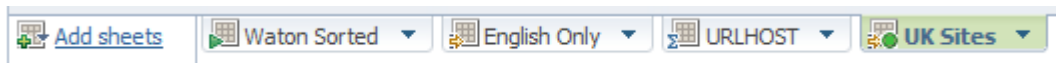
	Country	
	FeedInfo	
	Language	
	Language Revised	
	Published	
	SubjectHtml	
	Tags	

ParametersCarry over (9)

94. Click on the green check-mark button to “Apply Settings” and create this new tab.
95. One additional filter to apply would be a way to look at only URLs that end in “.uk” along with posts from the Country of Great Britain (code GB). So, add one more tab/sheet that filters on the following... (You may have to use the green plus-sign button to add the second condition. Also, be sure to use “Match any” to imply an “or” condition with your filter settings here. Lastly, be sure to manually “type in” any entries needed to create the appropriate conditions for your filter.)

96. If you click the green check-mark and you do not see any rows shown, you potentially forgot to change your “Match” radio button from “all” to “any”. So, just be sure to check this before you continue.

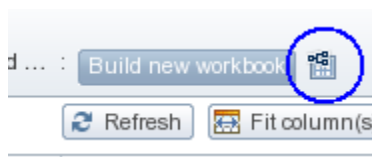
97. As one final double check, your workbook should look like this.

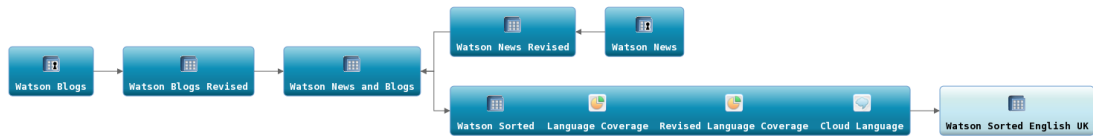


98. Once you run this, BigSheets will automatically run the processing within each tab from left to right. This will result in English only sites that end in “uk”.

99. You can now “save and exit” and then run the “Watson Sorted English UK” workbook.

100. View the Workflow Diagram





101. To quickly visualize the domain (URL Host Name) ending in “uk” with the most “English language” posts about Watson, you can create a Tag Cloud “chart”.
102. Click on the “Add chart” link.
103. Then click “Cloud” and then “Tag Cloud”.
104. Use the following settings.

#### New chart: Tag Cloud

Chart Name:	<input type="text" value="Top Sites"/>
Title:	<input type="text" value="Top 10 Sites with IBM Watson Coverage"/>
Tags:	<input type="text" value="URLHOST"/>
Count:	<input type="text" value="Count occurrences of value field"/>
Occurrence Order:	<input type="text" value="Descending"/>
Limit:	<input type="text" value="10"/>
<input type="button" value="✓"/> <input type="button" value="✗"/>	

105. Click the green check-mark button.
106. Click “Run” to run the visualization.
107. You can now see “itbriefing.net” has 16 occurrences (which ties [www.computerworlduk.com](http://www.computerworlduk.com)). You can discover they both have 16 by mousing-over each of the entries.

---

Top 10 Sites with IBM Watson Coverage

itbriefing.net [rss.feedsportal.com](http://rss.feedsportal.com)  
[www.computerweekly.com](http://www.computerweekly.com)

[www.computerworlduk.com](http://www.computerworlduk.com)  
[www.supercomputingonline.com](http://www.supercomputingonline.com) [www.techeye.net](http://www.techeye.net) [www.telecareaware.com](http://www.telecareaware.com) [www.thespoof.com](http://www.thespoof.com)  
[www.vadvert.co.uk](http://www.vadvert.co.uk) [www.zdnet.co.uk](http://www.zdnet.co.uk)

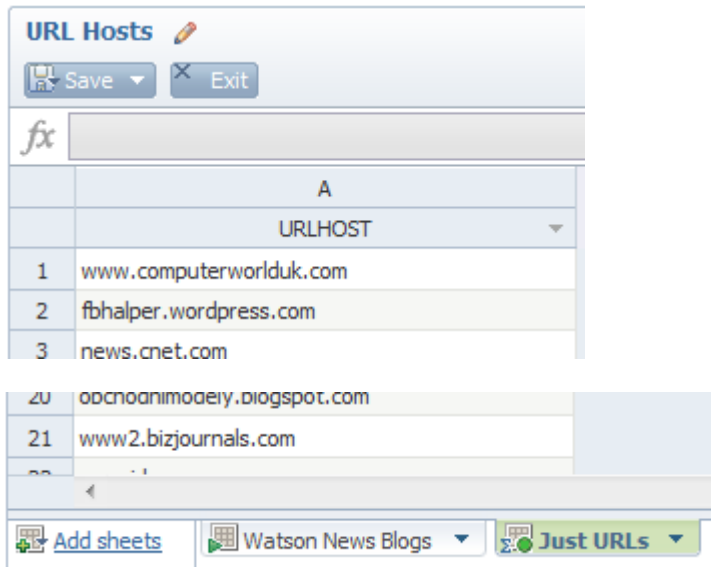
## 1.5 Combining Social Media with RDBMS Data

In this section, we will use the workbook of data we pulled from DB2 (earlier in this lab) and merge it with the data we collected from Boardreader. From an analysis standpoint, we want to look at the top 12 worldwide sites for coverage of IBM Watson. We also have, albeit fictitious, data about IBM’s public relations team efforts. In this case, we want to understand the coverage by sites where IBM’s media team performed outreach activities. This is to show how you might combine internal data from a relational database with external data from Social Media websites (from Boardreader).

### 1.5.1 Running a built-in Function to extract data

In this example, you'll use BigSheets to determine the number of distinct news and blog sites with coverage of IBM Watson. For this lab, you will need to start with the Data Workbook named “Watson News and Blogs”. Click on the “Watson News and Blogs” link under the BigSheets tab to open this workbook.

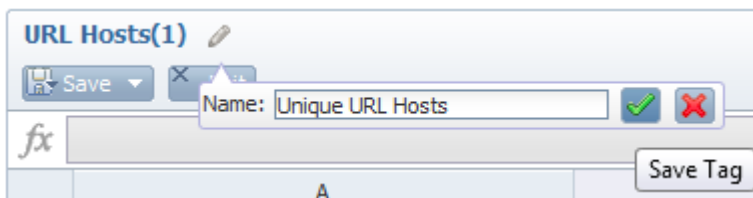
108. Borrowing a similar technique from the last lab...
109. Build a new workbook.
110. Give the name of “URL Hosts”.
111. Add a sheet that performs the URLHOST function against the Url column.
112. Click the green check-mark button.
113. As a quick check, you should be here now...



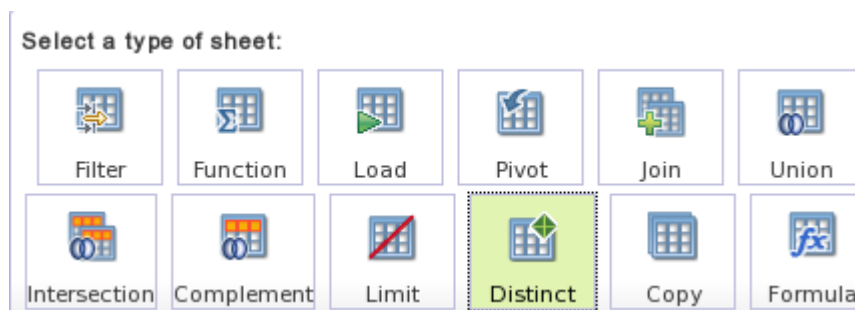
114. “Save and Exit” and run “URL Hosts” as you will need to “build new workbook” from it in the next step.

115. In order to quickly determine the number of unique, URLHOST entries in your new “URL Hosts” workbook, you can start with “URL Hosts”, build a new workbook, and add a new sheet to perform the distinct routine.

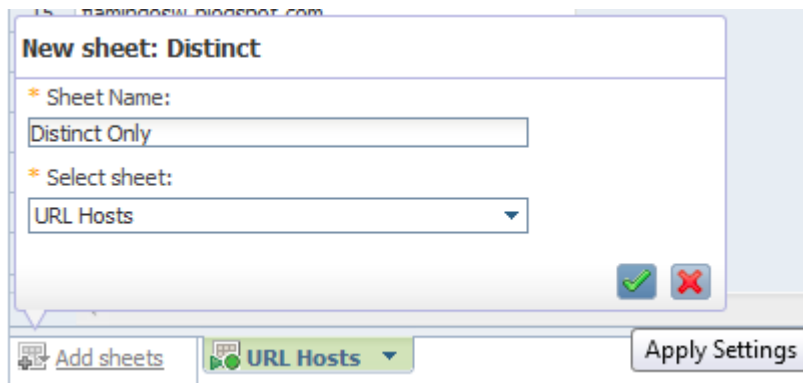
116. Build a new workbook.



117. Add the Distinct sheet.



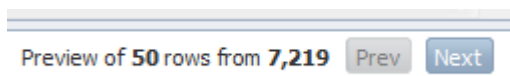
118. Select the proper sheet to perform the activity against.



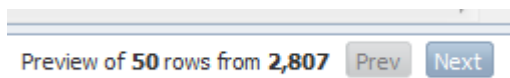
119. Click the green check-mark button.

120. “Save and Exit” and run the your new, “Unique URL Hosts” workbook.

121. Back in the “Watson News Blog” workbook, you can see that it contains 7,219 rows.



122. Once the “Distinct” processing is applied to the URLHOST column **within the “Unique URL Hosts” workbook**, you see only 2,807 distinct hosts. If you see “???” instead of 2,807, select the Next button.



123. Now, you might want to know more about these 2,807 sites. Let’s assume you want to know how many posts came from each site (or at least the top 12 sites by posting volume).

124. Navigate to the “URL Hosts” workbook and add a column chart to see the top 12 sites and then Run.

#### New chart: Column

Chart Name:

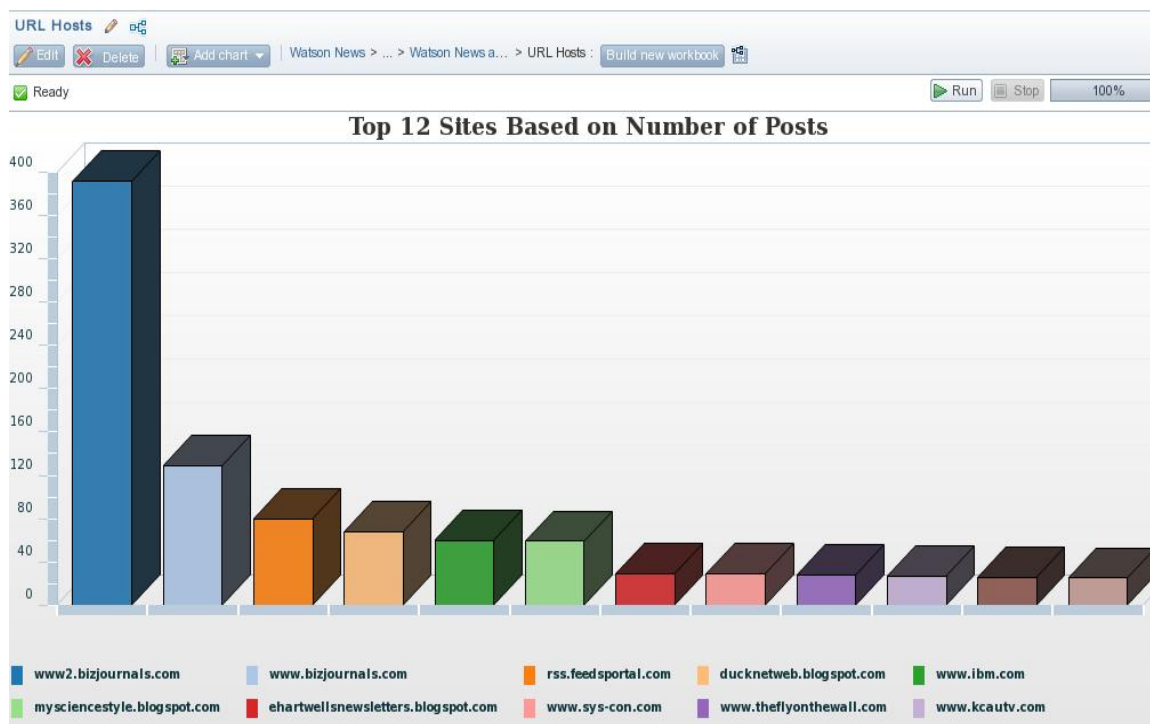
Title:

X Axis:

Y Axis:

Limit:

125. In this case, the top three sites are not even “ibm.com” as you might have expected. The top two sites are some variation of “bizjournals.com”. (So, just like the “Chinese” language cleanup you performed earlier, you could do that again here.)





### 1.5.2 Performing a Pivot or “Group by” Operation

Identifying the Top 12 sites, in the last section, might make you curious about the number of posts for each URL host site and other details in addition to just the total (as shown in the visualization above). This next example uses an easy approach to obtaining the total as well as individual URL details that an analyst might need to obtain through these analytics.

126. To perform these tasks, you will need to start by opening the “Watson News and Blogs” workbook.

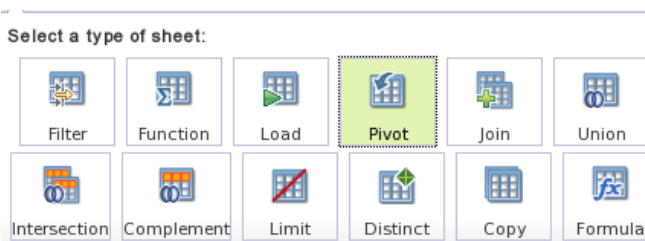
127. Build a new workbook and name it “Watson URL Details”.



128. Add a sheet to perform the URLHOST function and carry over all columns.

129. Now, you will want to add one additional sheet in order to perform a Pivot (or “group by”) operation.

130. Add a Pivot sheet.



131. Add URLHOST in the Group by Column and hit the green + button

New sheet: Pivot      Input sheet: URLHOST

\* Sheet Name:  
Pivot URLHOST

Group by columns:  
Add all   Remove all

URLHOST

Group   Calculate   Carry over (0)

132. Then, click on the “Calculate” tab in order to add another column to your new tab/sheet based on a particular calculation. Type in the name of the column you wish to add.

New sheet: Pivot      Input sheet: URL Host Added

\* Sheet Name:  
Pivot URLHOST

Create columns based on groups:  
Count\_URLHOST

Add column to sort

133. Then click the green plus-sign button and select the following items for your new column.

New sheet: Pivot      Input sheet: URL Host Added

\* Sheet Name:  
Pivot URLHOST

Create columns based on groups:  
Count\_URLHOST

Fill in parameters:  
Column: URLHOST

134. Then, we will add one more calculation before leaving this tab. In this case, we want a new column to contain a merge of all of the rows in our dataset that match the URLHOST that we are performing the Pivot (or “group by”) action on. Type in the new name of this new column we are going to add and click the green plus-sign button to add it.

Create columns based on groups:

Merge\_URL

Count\_URLHOST = COUNT

Fill in parameters:

Column: URLHOST

Merge\_URL = MERGE

Fill in parameters:

Column: Url

Separator: ,

Add column to sort

135. Then, configure the settings for this new column to use the MERGE operator on the Url column with “,” as a separator (that’s a comma and a space only, as you will not enter the double quotes... see below).

Column: URLHOST

Merge\_URL = MERGE

Fill in parameters:

Column: Url

Separator: ,

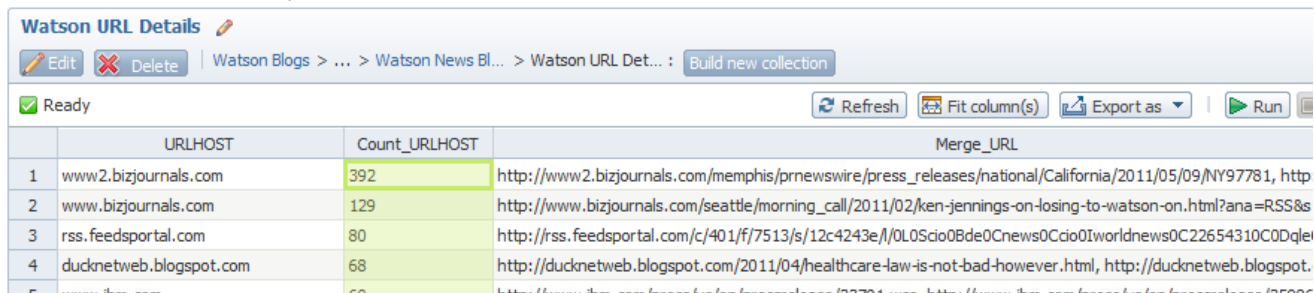
136. Click the green check-mark to build your new tab/sheet and produce the desired three (3) columns.

	A	B	C
	URLHOST	Count_URLHOST	Merge_URL
1	127.tumblr.com	1	http://127.tumblr.com/post/3362717688
2	1ubuntu.blogspot.com	1	http://1ubuntu.blogspot.com/2011/01/ibm-watson
3	2scottmontgomery.blogspot.com	1	http://2scottmontgomery.blogspot.com/2011/01/f

137. To make it easier to see the largest number of posts at the top, sort the Count\_URLHOST descending.

B	C
Count_URLHOST	Merge_URL
55	http://www.bizjournals.com/albany/prnews
50	com/albany/prnews
26	s://www.ibm.com/d
19	n/blog/2011/01/ibm
19	re.blogspot.com/2
10	
10	
9	
9	

138. “Save and Exit” and then run this workbook. You can now see the details of the URLs for each of the Host Names. Here is an example...

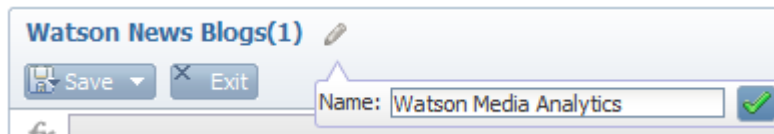


	URLHOST	Count_URLHOST	Merge_URL
1	www2.bizjournals.com	392	http://www2.bizjournals.com/memphis/prnewswire/press_releases/national/California/2011/05/09/NY97781, http
2	www.bizjournals.com	129	http://www.bizjournals.com/seattle/morning_call/2011/02/ken-jennings-on-losing-to-watson-on.html?ana=RSS&s
3	rss.feedsportal.com	80	http://rss.feedsportal.com/c/401/f/7513/s/12c4243e//0L0Scio0Bde0Cnews0Ccio0Iworldnews0C22654310C0Dgle
4	ducknetweb.blogspot.com	68	http://ducknetweb.blogspot.com/2011/04/healthcare-law-is-not-bad-however.html, http://ducknetweb.blogspot.

### 1.5.3 Joining Social with Structured data

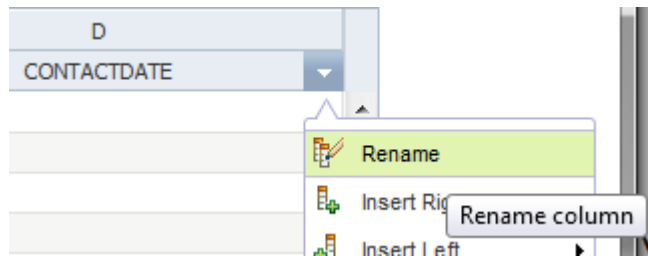
Last but not least, we will begin to work with the RDBMS data we pulled into a BigSheets workbook towards the beginning of this lab. As you might remember, we pulled data into a workbook and named it “Media Contacts”. We want to join this structured data with our Social Media data. By joining these two workbooks, you'll be able to explore how corporate media outreach efforts correlate to coverage by third-party websites.

139. In order to start with a workbook that has all of the items in it we need, we will again start with the “Watson News and Blogs” workbook. Open this workbook.
140. Build a New Workbook and call it “Watson Media Analytics”.

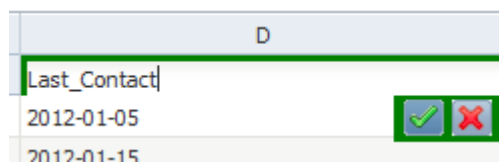


141. Again, we need the URLHOST column added to our new workbook. So, add a sheet that runs the URLHOST function and carries over all of the columns.
142. Add a sheet that Loads the “Media Contacts” workbook into your new, Watson Media Analytics workbook.
143. To make the data contents more clear, rename the CONTACTDATE column to Last\_Contact.

144. Click the triangle next to CONTACTDATE and select the rename option.



145. Change the column name to Last\_Contact.

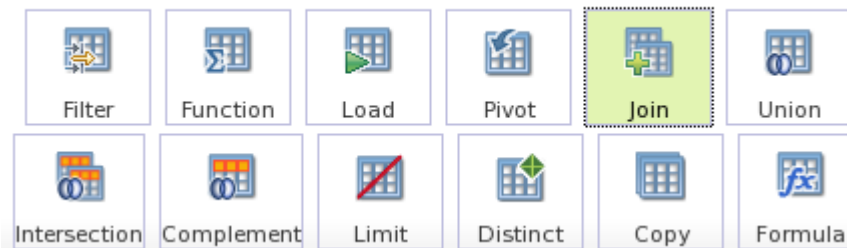


146. Click the green check-mark button (or hit enter) to change the column name.

147. Now, we will need to perform the “Join” of the proper sheets to build our final combination of social data with our data from the RDBMS.

148. Add a sheet that performs a Join.

Select a type of sheet:




149. Click to set the appropriate settings for this “Join” action (by using the pull down and the green plus-sign button).

**New sheet: Join**

\* Sheet Name:  
Join URLHOSTS & Contacts

Join type:  
Inner



Add sheets (at least 2) to join:  



[Add All](#) [Remove All](#)

Selected Sheets:

Watson Media Analy...	Media Contacts
Columns: URLHOST [text]	Columns: URL [text]

150. Select the appropriate Join column. (Since this is an “inner join”, only the rows that match will wind up in the results tab after this workbook gets run.)
151. Click the green check-mark button.
152. At this point, all of the columns from the URLHOST tab/sheet and now all four (4) columns of the Media Contacts workbook have been merged based on the join column from both workbooks. You should be able to see a simulated data result by scrolling to the right.
153. To make it easier to inspect the results, you can delete the ID column from your current sheet.
154. As an additional way to make your results look more intuitive, you can reorganize the order of the columns in the following order (by using the “Organize Columns” option or by dragging and dropping the column by a left-click-mouse-grab on the letter above the column name). Hint: Don’t forget about the “Fit Columns” button.

URLHOST  
NAME  
Published  
Last\_Contact  
FeedInfo  
Country  
Language  
SubjectHtml  
Tags  
Type  
Url

155. “Save and Exit” and then run this workbook.

156. The results contain the posts from each of the “Media Contacts” database. In this case, 169 posts matched up with data from the RDBMS. (Remember, this is just an example.)

Preview of 50 rows from 169

157. Generate a column chart of the Top 7 targeted sites and their posting quantity.

**Edit chart: Column**

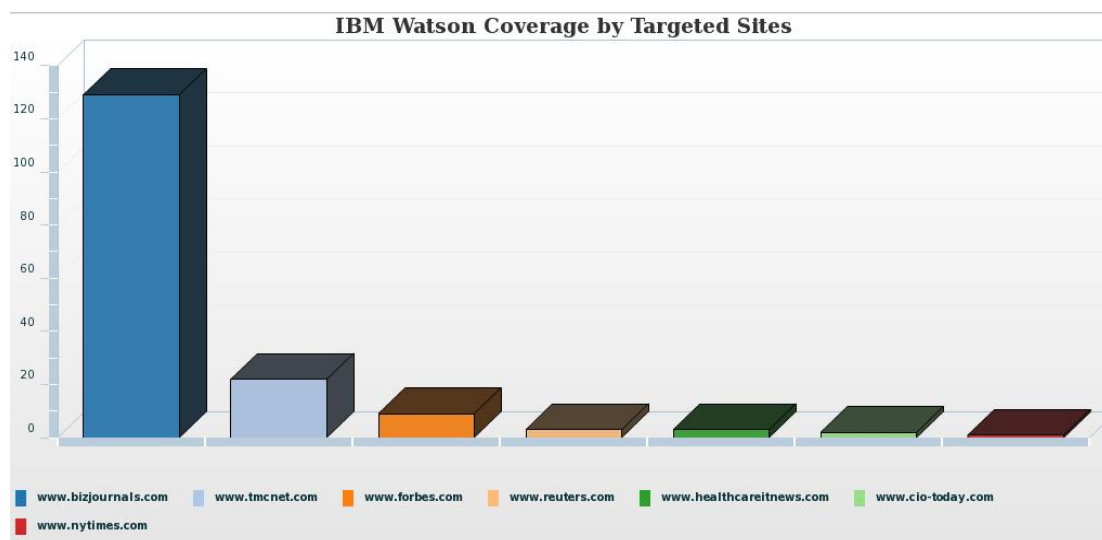
Chart Name:

Title:

X Axis:

Y Axis:

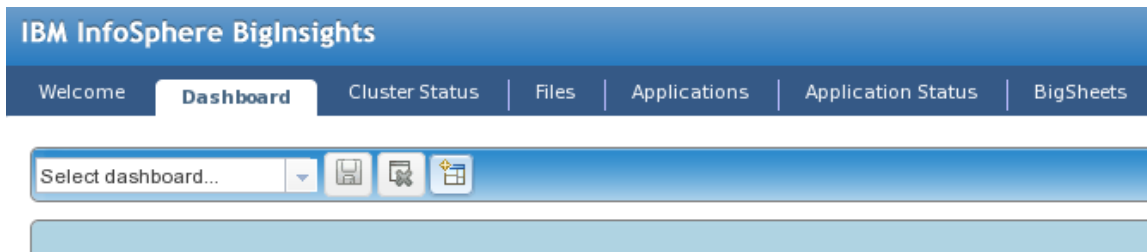
Limit:



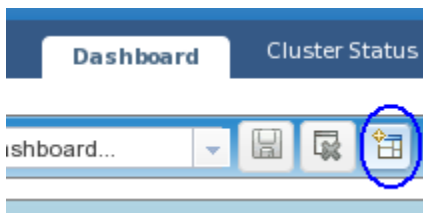
## 1.6 Dashboard

This section introduces you to the quickly and easily creating and managing custom dashboards. A custom dashboard allows you to gain total visibility over a set of data, a system, or analysis on a set of data depending on the types of widgets being managed by the dashboard. We will cover a simple dashboard with 3 widgets showcasing charts and data from other parts of this lab.

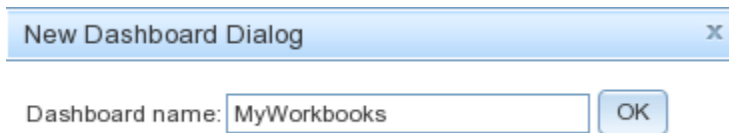
From the BigInsights web console, navigate to the Dashboard tab.



158. Add a new dashboard by clicking the “New Dashboard” icon



159. Name the new dashboard “MyWorkbooks”



160. Add two widgets: Watson\_sorted workbook and its associated pie chart (based on Language values) by selecting the Add Widget icon.



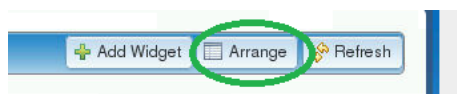


161. If Widgets do not come up, follow these instructions:
- In the biadmin console, type stop.sh derby console
  - Remove the following index folder

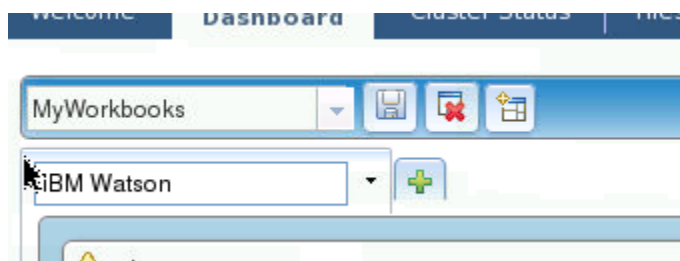
```
[biadmin@bigdata logs]$ rm -rf /opt/ibm/biginsights/sheets/work_data/index/
```

- In the biadmin console, type start.sh derby console and try again

162. "Arrange" the 2 widgets on the dashboard so they display nicely.



163. Rename the Dashboard tab to "IBM Watson" by clicking on the "New Tab" text.




164. From the Watson\_sorted workbook widget, click the "Open on BigSheets" icon in the upper right corner. Note that a new tab opens in your browser to display the target workbook in BigSheets.



165. Return to the MyWorkbooks dashboard and click on the "Create a chart" icon in the upper right corner of the Watson\_sorted workbook widget. Select "Bar" as the chart type. Select "Language" as the target column, change the Occurrence Order to "Descending" and retain all other default values. Click OK.

166. When the new chart appears, click "Arrange" to display all 3 widgets clearly on your dashboard.



**Why isn't my chart displaying properly?**  
Note that the widget for your new bar chart contains an icon with a yellow triangle at upper left. This indicates that sample data is displayed that the new chart needs to be run. Click the refresh button for this widget or, alternately, the refresh button for the entire dashboard. When the operation completes, inspect the results

167. Delete 1 widget from the new dashboard.

168. Optionally, delete the entire dashboard.

## 1.7 Summary

In this Lab, you have seen how you can work BigInsights to perform analytics on Social Media data along with structured data from an RDBMS.

This lab explored how BigInsights enables business analysts to work with big data without writing code or scripts. In particular, it introduced two sample applications for gathering social media and RDBMS data and explained how analysts can model, manipulate, analyze, combine, and visualize this data using BigSheets, a spreadsheet-style tool designed for business analysts. To keep things simple, this article explored a subset of BigSheets operators and functions, concentrating on those most relevant to our sample application scenario involving coverage of IBM Watson, a research project that uses Apache Hadoop to perform complex analytics to answer questions presented in a natural language.

## 1.8 Appendix

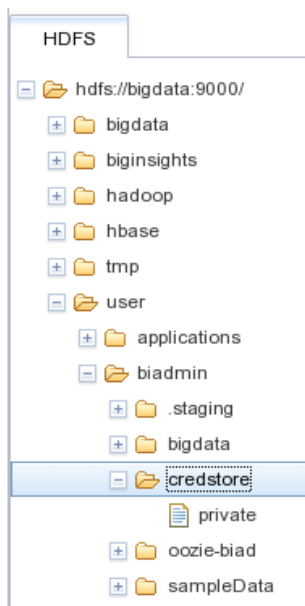
### 1.8.1 Collecting Social Media Data using BoardReader

If you already have a Boardreader license key, you can place it into a text file of the correct format, within the right folder, with the proper file access authority settings. This entire step is optional as the output of this step has already been included in the VMware image you are running. If you are running this lab on the IM Demo Cloud, a "shared" Boardreader license key will be provided on that environment for you. Please use this key sparingly as it is a shared pool of monthly requests across all users of the IM Demo Cloud system.

169. Create your private/<key or properties> file per the instructions found here:  
[http://pic.dhe.ibm.com/infocenter/bigins/v1r4/index.jsp?topic=%2Fcom.ibm.swg.im.infosphere.biginsights.dev.doc%2Fdoc%2Fc\\_sample\\_apps\\_boardreader.html](http://pic.dhe.ibm.com/infocenter/bigins/v1r4/index.jsp?topic=%2Fcom.ibm.swg.im.infosphere.biginsights.dev.doc%2Fdoc%2Fc_sample_apps_boardreader.html)  
with additional instructions regarding the credential store process found here:  
<http://pic.dhe.ibm.com/infocenter/bigins/v1r4/index.jsp?topic=%2Fcom.ibm.swg.im.infosphere.biginsights.admin.d>

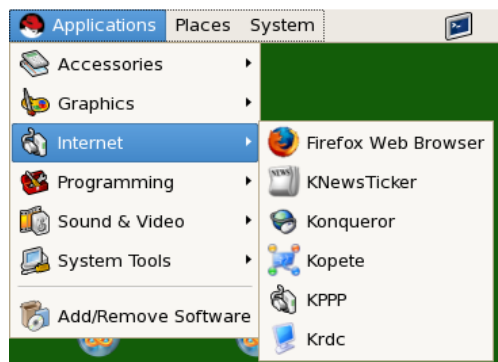
[oc%2Fdoc%2Fcredentials\\_store.html&resultof%3D%2522%2563%2572%2565%2564%2573%2574%256f%2572%2565%2522%2520%2522%2563%2572%2565%2564%2573%2574%256f%2572%2522%2520](oc%2Fdoc%2Fcredentials_store.html&resultof%3D%2522%2563%2572%2565%2564%2573%2574%256f%2572%2565%2522%2520%2522%2563%2572%2565%2564%2573%2574%256f%2572%2522%2520)

170. If you want to execute the Boardreader application, you will need to complete this step. The Files tab under BigInsights should contain the following folders and files as a result of this step.



171. Lastly, to run this step, you will need to make sure your VMware image is able to communicate and pull data from the Internet.

172. Launch a web browser (Firefox) from within your VMware image.



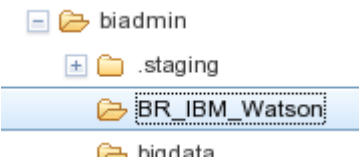
173. Navigate to a public URL to ensure your system can pull data from the Internet.



174. At this point, we will want to create an empty folder for your Boardreader output to go into. Navigate to the Files tab.



175. Expand the directory tree structure and create a folder under /user/biadmin called BR\_IBM\_Watson. (If you have completed the Lab Console Lab, you should be familiar with the process and the result should appear as shown below.)



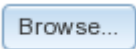
176. Navigate to the Applications tab.

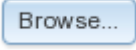


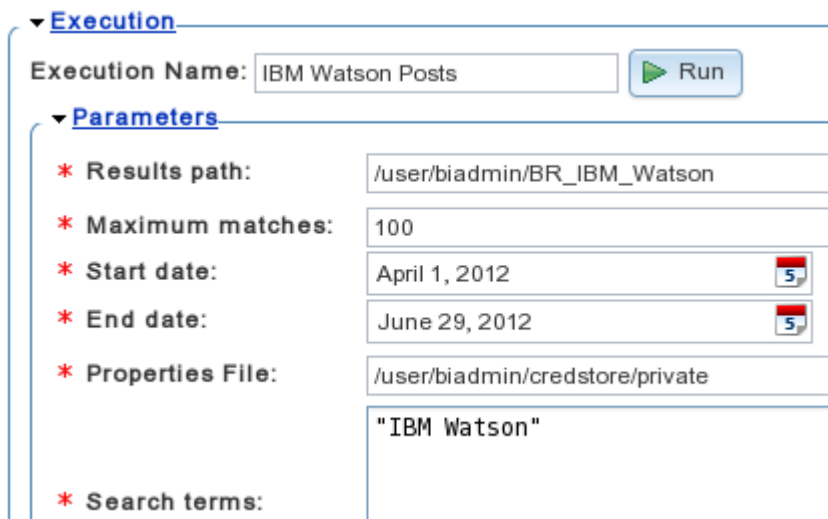
177. Click on the Boardreader Application icon on the left-hand side.




178. On the right-hand side, you will need to do the following: Enter an Execution Name. For example, “IBM Watson Posts” but do NOT click run yet.

179. Select the right path for your Results path: by clicking the  button.


180. Leave the Max matches set to 100 (as every Boardreader key typically has a monthly limit and you don't want to be the one to use it all up).
181. Set the start date to something at least a month back (so that you will get all 100 results you are about to request).
182. Select the right path for your Properties File: by click the  button.
183. In the Search terms box, enter "IBM Watson" as the search term you want to pull from Social Media. To ensure you pull the proper string, be sure to add the double-quotes when entering this search term.
184. Your entries should look very similar to the following...



**Execution**

Execution Name:  



**Parameters**

- \* Results path:
- \* Maximum matches:
- \* Start date:  
- \* End date:  
- \* Properties File:
- \* Search terms:

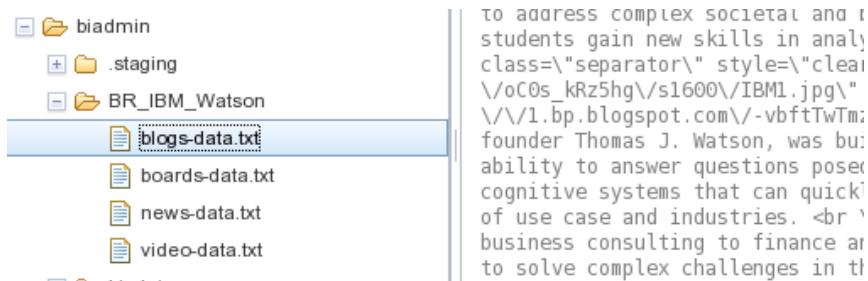
185. Click the Run button to make the request to Boardreader.
186. If everything was entered correctly (and your credstore/private file was correct), you should see the following in the bottom half of the window...

Application History		
Status	Execution Name	Progress
 ... No filter applied		
	IBM Watson Posts	<div>100%</div>

187. You can now click on the magnification glass icon under the output column to easily navigate to your application's (job's) results.

Elapsed Time (sec)	Output	Details
26		

188. Expanding the BR\_IBM\_Watson folder and clicking on the blogs-data.txt file name should show you the contents on the right-hand side, in text format by default.

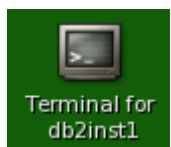


### 1.8.2 Working with RDBMS

Part of the lab will require access to DB2. Double click the following icon:



189. A terminal window should launch and then disappear.  
190. To quickly check if DB2 is running, find and double-click on the "Terminal for db2inst1" icon.



191. In the Terminal window, run a "db2 connect to sample" command and if you see the following, DB2 is running.

```
[db2inst1@bigdata ~]$ db2 connect to sample

Database Connection Information

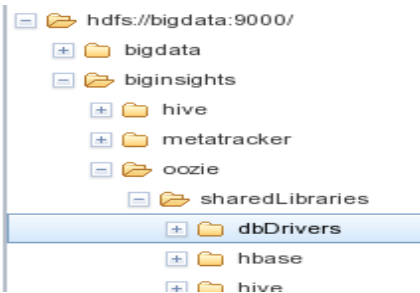
Database server      = DB2/LINUX8664 9.7.4
SQL authorization ID = DB2INST1
Local database alias = SAMPLE

[db2inst1@bigdata ~]$
```

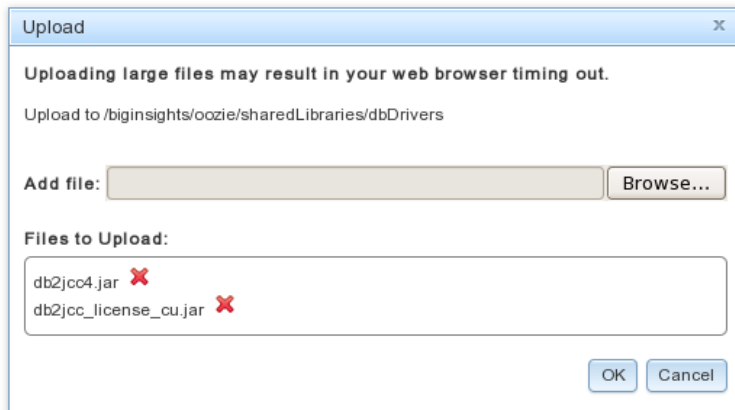
192. You can now close this Terminal window.

193. If you want to actually do the pull from DB2 on your image, we will need to configure your image in order to add the DB2 drivers to BigInsights. Depending on the version of the VMware image and/or the IM Demo Cloud image, these files may have already been put into place by the creator of the image. If they are already there, that's fine. You may want to read through the following steps to better understand what it takes to prepare BigInsights to connect to DB2 if you performed a "clean install" yourself.

194. Using the Files tab, you will need to create the following, dbDrivers folder



195. Then, you will need to upload two files (db2jcc4.jar and db2jcc\_license\_cu.jar) into this folder from your /opt/ibm/db2/V9.7/java folder...



196. Clicking on OK will complete the file upload of these two, JDBC driver files for BigInsights.

197. Last item to prep is your credstore/private/<your properties> file. If one of these does not yet exist on your system, you will need to add the following four (4) lines to your hdfs://bigdata:9000/user/biadmin/credstore/private/db2\_credentials.properties file.

On the IM Demo Cloud, your DB2 connection information is already provided in the /user/<your userid>/credstore/private/mykeys file. You can easily add more files to this folder in order to safely maintain connections to multiple data sources.

The text version before it is loaded into HDFS via the credstore loading process described within the BigInsights Information Center...

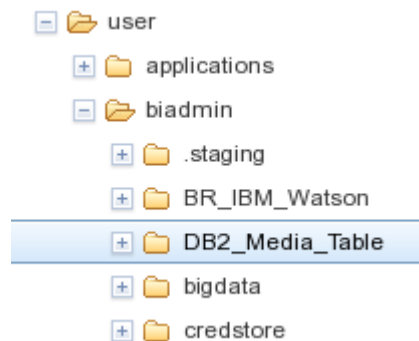
```
dbdriver=com.ibm.db2.jcc.DB2Driver
dburl=jdbc:db2://localhost:50000/SAMPLE
dbuser=db2inst1
dbpwd=[encode]passw0rd
```

If you view the file in HDFS after it has been loaded (and since it is in the “private” folder, the password gets encoded as indicated above) as well as certain characters get a backslash (\) in front of them to also be escaped...

```
#BigInsights Credential Store file
#Tue Jul 03 03:53:42 EDT 2012
dburl=jdbc\:db2\\:localhost\:50000/SAMPLE
dbdriver=com.ibm.db2.jcc.DB2Driver
dbpwd=[base64]Lz4sLChvLTs\=
dbuser=db2inst1
```

These are specific to the VMware image you are currently using. Obviously, if you needed to connect to another database, you would need to change the input parameters to match and optionally create another credstore/private/<dbms> file for use with that database connection.

198. Before we can connect and pull any data to BigInsights, we need to create an output directory. So, just like in the last lab, create a folder (under the Files tab) for your Database Import “app” to place the results.



199. Now, go to the Applications tab, click on the Database Import app icon on the left, and fill out the right hand-side to look like the following (before you click the Run button)...



▼ **Execution**

Execution Name:

▼ **Parameters**

\* Properties file :

\* SQL statement:

\* output format:

\* Output directory:

\* CSV delimiter:

\* Include Column Headers: ☒

If you are on the IM Demo Cloud, you will use the mykeys file and  
SELECT \* FROM SARACCO.MEDIA instead of the items shown above.

200. If everything looks fine, you can click the Run button.
201. Once the run has completed, you can click the magnifying glass under the output column, expand the DB2\_Media\_Table folder, and then click to view the contents of the output.txt file.



© Copyright IBM Corporation 2013  
All Rights Reserved.

IBM Canada  
8200 Warden Avenue  
Markham, ON  
L6G 1C7  
Canada

IBM, the IBM logo, ibm.com and Tivoli are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries, or both. If these and other IBM trademarked terms are marked on their first occurrence in this information with a trademark symbol (® or ™), these symbols indicate U.S. registered or common law trademarks owned by IBM at the time this information was published. Such trademarks may also be registered or common law trademarks in other countries. A current list of IBM trademarks is available on the Web at "Copyright and trademark information" at [ibm.com/legal/copytrade.shtml](http://ibm.com/legal/copytrade.shtml)

Other company, product and service names may be trademarks or service marks of others.

References in this publication to IBM products and services do not imply that IBM intends to make them available in all countries in which IBM operates.

No part of this document may be reproduced or transmitted in any form without written permission from IBM Corporation.

Product data has been reviewed for accuracy as of the date of initial

publication. Product data is subject to change without notice. Any statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.

THE INFORMATION PROVIDED IN THIS DOCUMENT IS DISTRIBUTED "AS IS" WITHOUT ANY WARRANTY, EITHER EXPRESS OR IMPLIED. IBM EXPRESSLY DISCLAIMS ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE OR NON-INFRINGEMENT.

IBM products are warranted according to the terms and conditions of the agreements (e.g. IBM Customer Agreement, Statement of Limited Warranty, International Program License Agreement, etc.) under which they are provided.