Using BigSheets for Spreadsheet-like Analytics

Performing data analysis using BigSheets

Contents

Performing data analysis with BigSheets

In the last exercise, you loaded data into BigSheets. In this exercise you will be doing data analysis on social media data using BigSheets. To keep things simple, as this is only a lab exercise, you have been provided the files from the BoardReader application which collects information from blogs and forums. To use the BoardReader application to do the data collection yourself, you will need a license.

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

- Create a workbook from a master workbook
- Add sheets to a workbook
- Modify sheets in a workbook
- Create charts to visualize a workbook

Allow 30-45 minutes to complete this section of lab.

Throughout this lab you will be using the following account login information:

When to use:	Username	Password
Log in from the command-line to accept the licenses	root	password
Log in from the RHEL Desktop to access the BigInsights Desktop	virtuser	password
Log in from the Ambari console	admin	admin

2.1 Creating BigSheets workbook

Parent workbooks cannot have any sheets added to them, so you will be creating child workbooks to perform data analysis. The first two files that you upload into the system are acquired from the BoardReader application. The third file that you will use is from a sample RDBMS from which you will need to use for the analysis.

2.1.1 Creating the parent workbooks

__1. Make sure your BigInsights cluster has started. If not, start it by accessing the Ambari console.

__2. You will be using the **blogs-data.txt** and **news-data.txt** file as well as the **RDBMS_data.csv** file.

__3. Use the same process from lab one to create a workbook for news-data.txt, blogs-data.txt and RDBMS_data.csv.

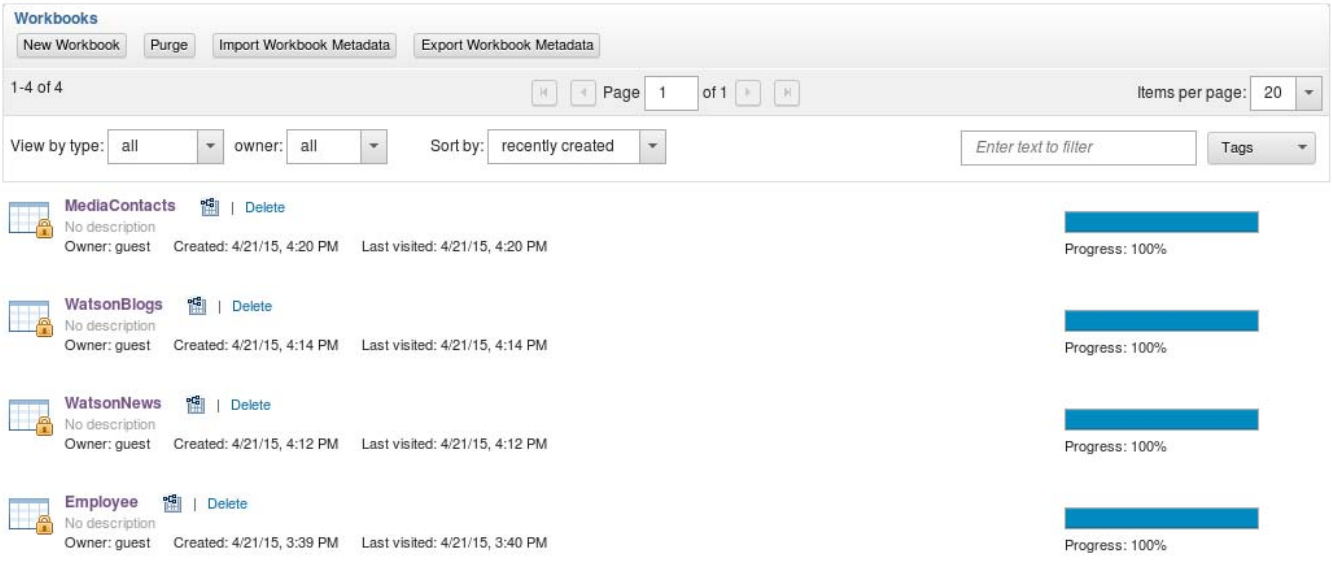
__4. For **news-data.txt**, use the JSON Array reader and give it the name of **WatsonNews.**

__5. For **blogs-data.txt**, use the JSON Array reader and name it **WatsonBlogs**.

__6. For **RDBMS-data.csv**, use the CSV reader, no headers, and name it **MediaContacts**

2.1.2 Modifying the workbook

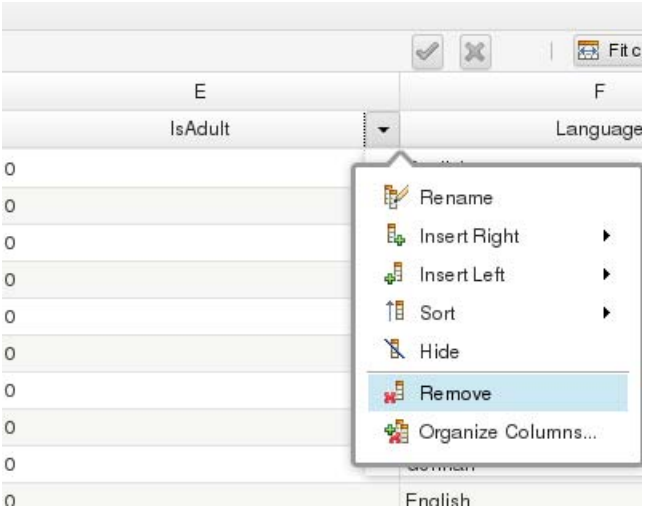
__1. All three parent workbooks have been created. Now you will need to create child workbooks before you can start doing any work to them. Go to the BigSheets home page, where all the workbooks are located.



__2. Click on the **WatsonNews** workbook**.**

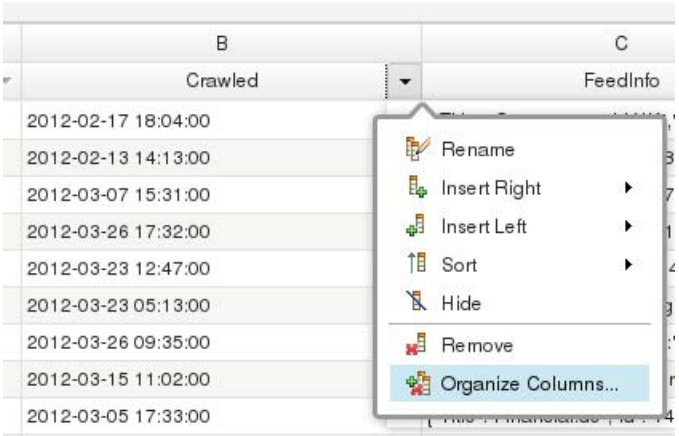
__3. Click the **Build new workbook** pushbutton.

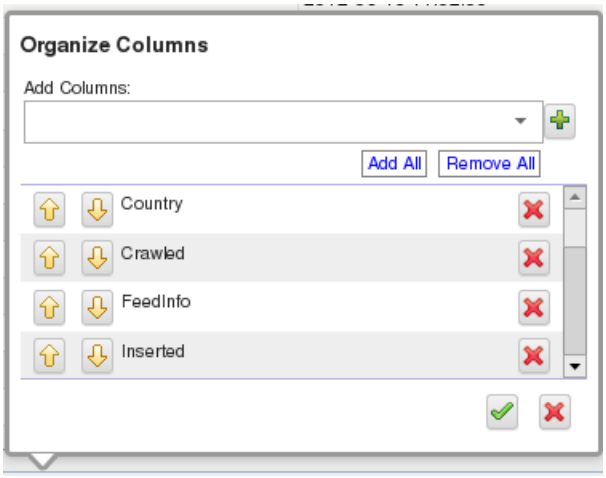
__4. Now that the child workbook has been created, we want to remove some columns to simplify the workbook. Remember, if we build additional child workbooks from this one, all the columns will be inherited so it is best to simplify early. You will want to remove the *isAdult* column. Click the drop down arrow next on the isAdult column and select **Remove:**



__5. You will want to remove all the columns from the workbook EXCEPT the following columns. Keep these columns: **Country, FeedInfo, Language, Published, SubjectHtml, Tags, Type, and Url**

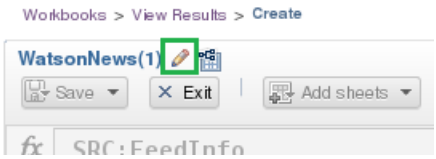
An alternative way to remove columns is by selecting **Organize Columns…** from the column options.





Click the **red X** icon in the popup dialog to remove the columns that you want to delete. Once you are done, click the **green checkmark.**

__6. Before you save the workbook, rename it to **WatsonNewsRevised.** Click the Pencil icon next to the *WatsonNews(1)* text to bring up the dialog to rename the workbook.



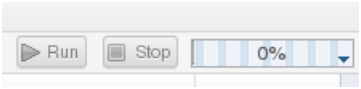
__7. Click the **Save** pushbutton and select the **Save & Exit** option.

__8. Notice that you could have also updated the workbook name in the dialog that popped up. In our case, we have already made the update, so click on the **Save** pushbutton.

Remember that a workbook shows only a sample of your entire data set. Once you save and exit the workbook after the changes have been made, you will need to run your workbook so that the changes are applied to the entire workbook. Behind the scenes, BigSheets runs Pig scripts that initiate MapReduce jobs on your collection. The time it takes for the job to complete depends on the volume of your collection.

__9. Click the **Run** pushbutton to update the data.

__10. Once you click Run, you will see a progress indicator:



__11. When it completes, you will be able to view the entire workbook with the changes you made

__12. Ultimately, you will want to consolidate the news and the blogs data for analysis, follow the same approach for creating the **WatsonBlogsRevised** workbook**.**

__13. Open the **WatsonBlogs** workbook.

__14. Click the **Build new workbook** pushbutton.

__15. Keep these **Country, FeedInfo, Language, Published, SubjectHtml, Tags, Type, and Url** and remove all other columns like you did for the WatsonNews.

__16. Click **Save & Exit** and specify the name of **WatsonBlogsRevised** (if you had not already done so).

__17. Run the workbook.

2.2 Consolidating two sheets for analysis

In this section, you will be merging the two workbooks: WatsonNews and WatsonBlogs with a union operation as a basis for exploring the data. To do so, both workbooks must have the same structure (or schema). In the last section, you modified the two workbooks to have the same columns so both workbooks are ready to be merged.

Before you can do a union operation, both sheets must be in the same workbook. You will open the WatsonNewsRevised and bring in the WatsonBlogsRevised sheet using the load operation.

__18. Open the **WatsonNewsRevised** workbook.

__19. Click the **Build new workbook** pushbutton.

__20. Click the **Add sheets** pushbutton.

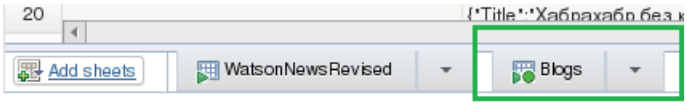
__21. Select the **Load** sheet.

__22. Give the new sheet the name: **Blogs.**

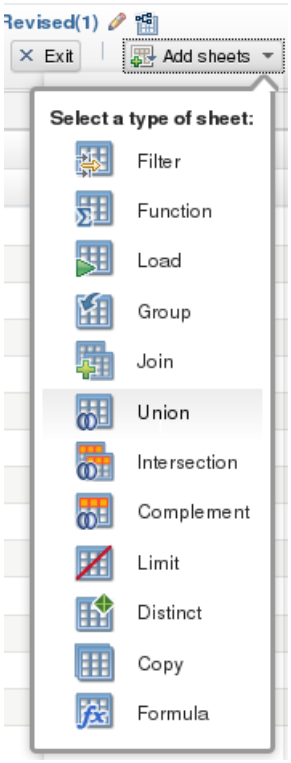
__23. Select the **WatsonBlogsRevised** workbook.

__24. Click the **green checkmark** to run the load operation.

__25. Once the operation completes, at the bottom left of the window, you will see that a new tab that show the Blog sheet that was just loaded:

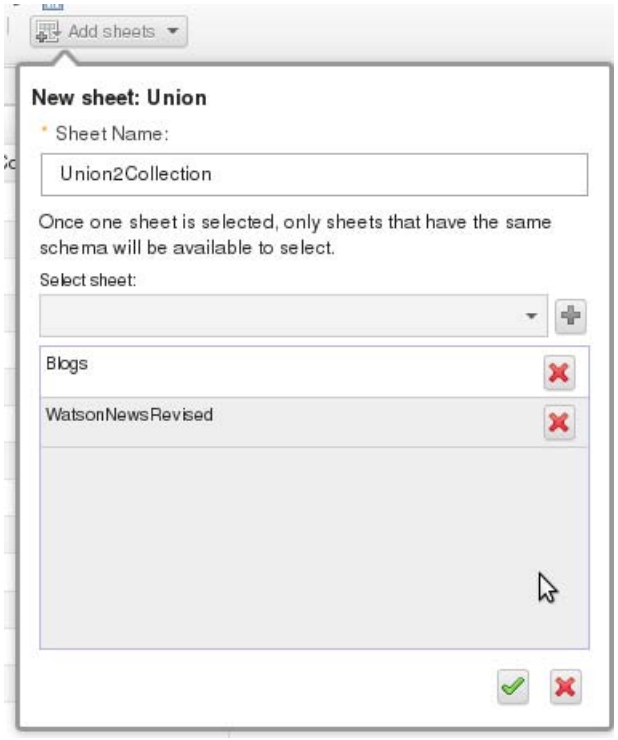


__26. Now you are ready to create a union of these two sheets. Click **Add sheets** and select the **Union** operation.



__27. Name the sheet: **Union2Collection**

__28. From the **Select sheet** dropdown, add the **Blogs** and the **WatsonNewsRevised** sheet to be used for the *Union* operation.



__29. Click the **green checkmark** to run the operation.

__30. You will see a new tab at the bottom when the operation completes.

__31. Save & Exit the sheet. Name the sheet: **WatsonNewsBlogs**.

__32. Run the workbook.

2.3 Data analysis using BigSheets

You are now ready for to do some analysis on the data. You have loaded data into the BigSheets environment, created workbooks from those data, remove the unnecessary columns and combined those two sheets into one.

2.3.1 Sorting the workbook

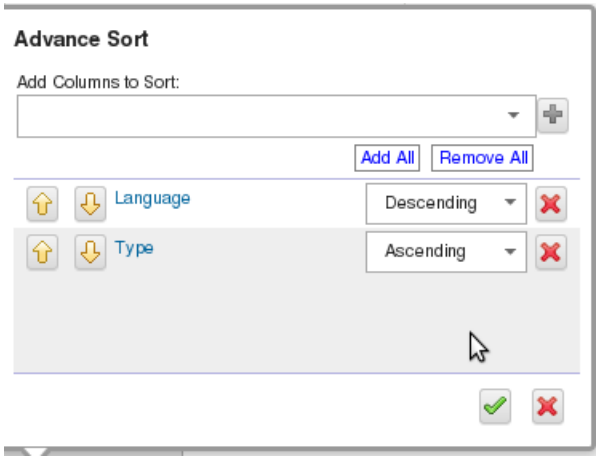
Looking at the news and blog entries, you see that there is a language column. Sort by the language column to see the Watson coverage around the world.

__33. Open the **WatsonNewsBlogs** workbook.

__34. Create a new child workbook from it.

__35. Click the column options and select the **Sort → Advanced Sort** option.

__36. Add the columns Language and Type to be sorted. Sort the Language column as *descending* and the Type column as *ascending.* Make sure the Language column is the primary search column (first column in the list).



__37. Click the **green check mark** to run the sort operation.

__38. Save and run the workbook as **WatsonSorted.**

__39. When the run completes, you will see more languages in the workbook.

2.3.2 Visualizing the workbook

Now you will create a graph to visualize your results. To set up the workbook, you need to filter out any possible null from it; otherwise the Chart function will throw an error. Create a new workbook.

__40. Within the WatsonSorted workbook, build a new child workbook.

__41. Remove the **tag** column.

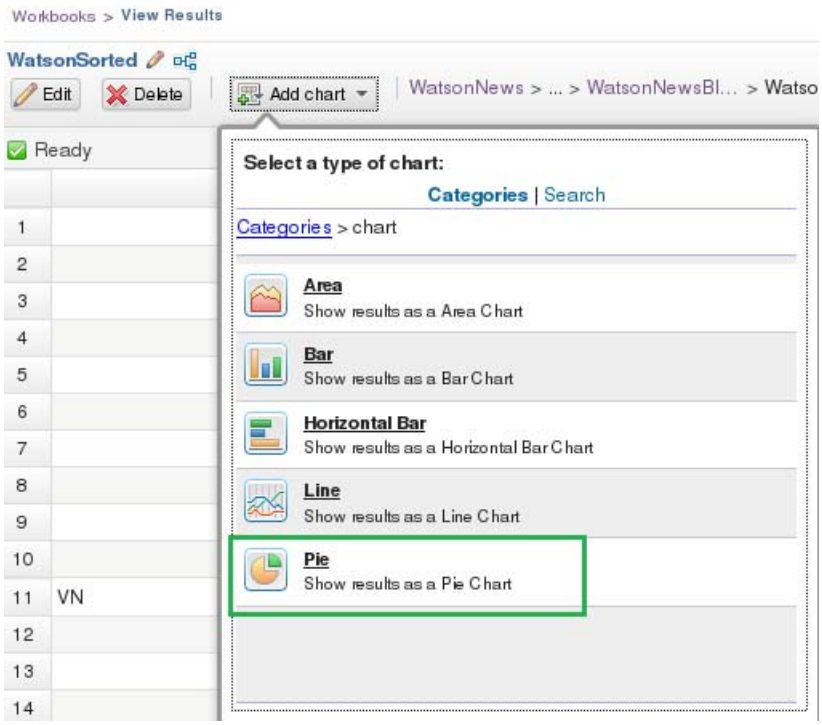
__42. Add a filter sheet. Name it: **CountryIsNotEmpty**

__43. Save the workbook as **WatsonSortedFiltered**

__44. Save and exit.

__45. Run the workbook.

__46. With the *WatsonSortedFiltered* workbook still open, click the **Add chart** pushbutton and select **Chart → Pie**



__47. Provide the following values for the *Pie* chart:

Chart Name: Language coverage

Title: Watson coverage by language

Value: Language

Count: Count occurrences of X axis values

Sort by: Count

Occurrences Order: Descending

Limit: 12

Template: Soda Cap

__48. Click the **green checkmark** to create the chart.

__49. You will need to run the chart against the full set of the data, so click **Run.**

__50. Once the run completes, you will see that English has the biggest slice of pie. What is the second most appeared language? If you move the mouse over the second biggest piece of the pie, you'll see that it is Russia.

__51. Move the mouse pointer over the fourth and sixth biggest slice and you will see that they're both Chinese. Chinese (Simplified) and Chinese (Spelling). This shows one of the common situations involving data from multiple sources where you may need to do additional refactoring of the data in order to get what you need. In this case, you have multiple entries that you need to treat as identical. For the purpose of our exercise, you will stop here, but feel free to play around with the different sheets and functions to see other types of operations you can perform on the data.

2.3.3 Data analysis with structured data

In this section, you will see another common situation where you may need to combine your social media data with your internal structured data for analysis.

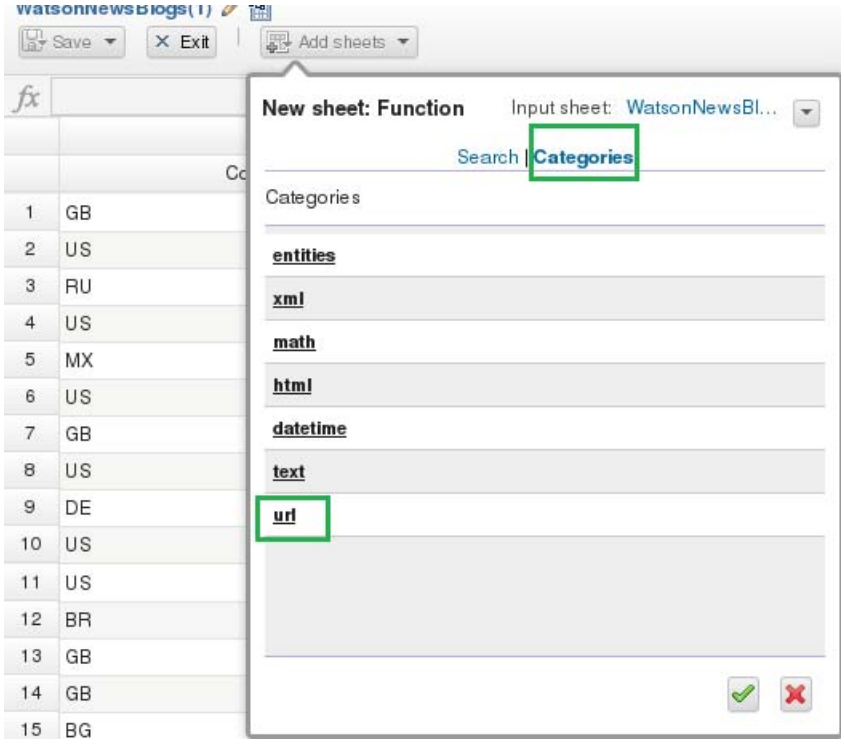
__52. Open the **WatsonNewsBlogs** workbook.

__53. Click the **Build a new workbook** pushbutton

__54. Click the **Add sheets** icon and select the **Function** sheet.

__55. Click the **Categories** hyperlink.

__56. Click the **url** hyperlink.



__57. Name the sheet: **URLHOST**

__58. For the parameters field, select the **Url** parameter from the dropdown menu.

__59. At the bottom, click the **Carry over** tab.

__60. Click the **Add all** hyperlink to add all of the columns to carry-over.

__61. Click the **green checkmark** to add this new sheet to your workbook.

__62. Next, add the MediaContacts workbook into your existing workbook as a sheet using the Load. Click the **Add sheets.**

__63. Select the **Load** sheet.

__64. Sheet name is: **Contacts**

__65. Select the **MediaContacts** workbook.

__66. Click the **green checkmark**.

__67. Rename the four columns to: **ID, Title, URL and LastContact,** in that order.

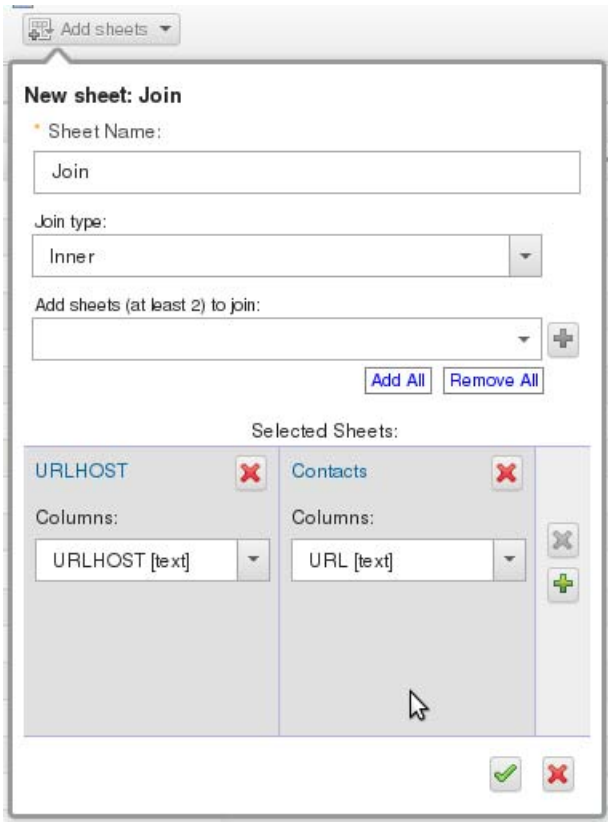
__68. Now you are going to add another sheet that combines the URLHOST and the Contacts sheets based on the values of URLHOST and URL columns. Click **Add sheets** and select **Join.**

__69. Name the sheet: **Join.**

__70. The *Join type* is **Inner**

__71. Add the two sheets from the dropdown menu, **URLHOST** and **Contacts**.

__72. Specify the columns **URLHOST** and **URL**, respectively.



__73. Click the **green checkmark**.

__74. Reorganize the columns to something like this (if you want): URLHOST, NAME, Published, LastContact, FeedInfo, Country, Language, SubjectHtml, Tags, Type

__75. Save the sheet as **WatsonNewsBlogCombined.**

__76. **Run** the sheet.

__77. To visualize the workbook, create a **Horizontal Bar** chart.

Summary

Having completed this exercise, you should now be able create workbooks and sheets that will perform some type of operation on the workbook. You combined two sets of data from which you had modified to have the same structure. By doing so, you have one set of data from which you can do the analysis. You also combined structural data from a RDBMS to the social media data. You also used visualization techniques offered by BigSheets to see your data as charts and graphs.

NOTES

NOTES