

Exercise 1

Working with BigSheets

Exercise 1: Working with BigSheets

Exercise 1: Working with BigSheets

Purpose:

You will create a BigSheets workbook and derive a chart from it to visualize your data.

Estimated time: **30 minutes**
 User/Password: **biadmin/biadmin**
root/dalvm3
 Services Password: **ibm2blue**

Important: Before doing this exercise, ensure that your access and services are configured and running. Check that:

- /etc/hosts displays your environment's IP address
- in the Ambari console, ensure that all BigInsights services are running

If you are unsure of the steps, please refer to Unit 1, Exercise 1 to ensure that your environment is ready to proceed. You should review the steps in Task 1 (Configure your image) and Task 2 (Start the BigInsights components).

Task 1. Loading data into BigSheets.

BigSheets allows you to analyze the data residing on the HDFS. You can create master workbooks, apply various sheets types to refine and filter the data, and then create charts to visualize the data. This task will walk you through the start to the end from creating a workbook to visualizing the data with charts. More functions and features will be covered in the BigSheets specific module.

You will load in two set of data into the HDFS.

1. To open a new terminal, right-click **biadmin's Home**, and then click **Open in Terminal**.
2. Navigate to **/home/biadmin/labfiles/bigsheets** to see the files.
3. Upload **blogs-data.txt** and **news-data.txt** to **/user/biadmin/**.

```
hdfs dfs -put /home/biadmin/labfiles/bigsheets/blogs-data.txt /user/biadmin
hdfs dfs -put /home/biadmin/labfiles/bigsheets/news-data.txt /user/biadmin
```

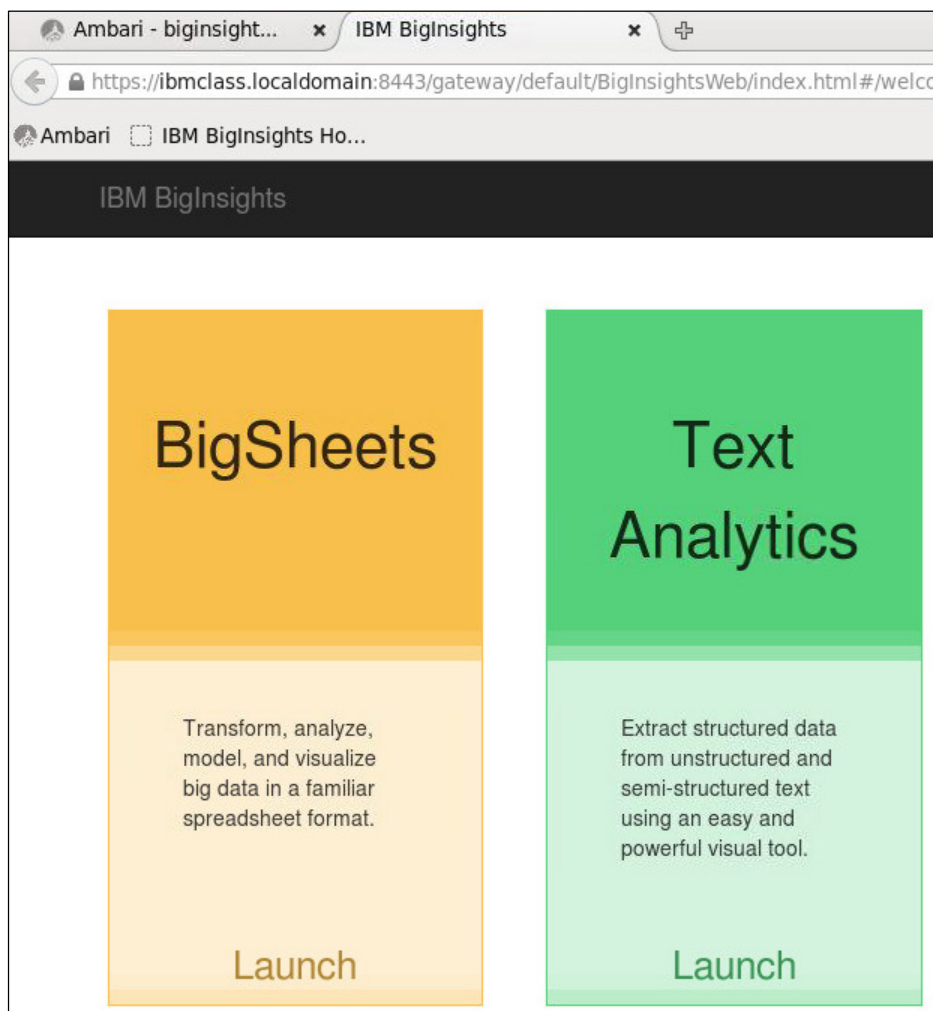
Once the files are inside of the HDFS, you are ready to create the BigSheets workbook.

4. Launch **Firefox**, and then if necessary, navigate to the **Ambari** login page, **http://ibmclass.localdomain:8080**.
5. Log in to the **Ambari** console as **admin/admin**.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

6. Ensure that the **BigInsights - BigSheets** component is started.
The BigInsights Home requires the LDAP server to be started.
7. Click the **Knox** component.
8. Under the **Service Actions** dropdown on the upper right, select **Start Demo LDAP**, and then click **OK** to close the confirmation window.
9. In Firefox, open a new tab, and navigate to the **Web UI (BigInsights Home)** page with the following URL.
<https://ibmclass.localdomain:8443/gateway/default/BigInsightsWeb/index.html>
If prompted for a login, use guest / guest-password. It should be saved in the Firefox browser so you can click OK to continue with the login.

The results appear as follows:



10. Click **BigSheets**, to launch BigSheets.
In the next few steps, you will create two parent workbooks. One for the news-data.txt and one for news-blogs.txt residing on the HDFS.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

11. Click the **New Workbook** button.
12. On the New Workbook window, under **Name**, type **News Data**.
You can leave the description field blank.
13. On the **DFS Files** tab, navigate to **/user/biadmin/** and select the **news-data.txt** file.



The data will be previewed on the pane to the right.

By default, it is using the Line Reader to parse the data. You will want to select the JSON Array reader so that the data is parsed correct.



The results appear as follows:

The screenshot shows the 'New Workbook' window. On the left, the 'DFS Files' tab is active, showing a file tree with 'user' expanded and 'biadmin' selected. The file 'news-data.txt' is highlighted. On the right, the 'Line Reader' is selected, and the data is previewed in a table format. The table has a 'Header' row and 13 data rows. The data is JSON-formatted text, such as '{"PostSize":6597,"Crawled":"2012-02-17 18:0'.

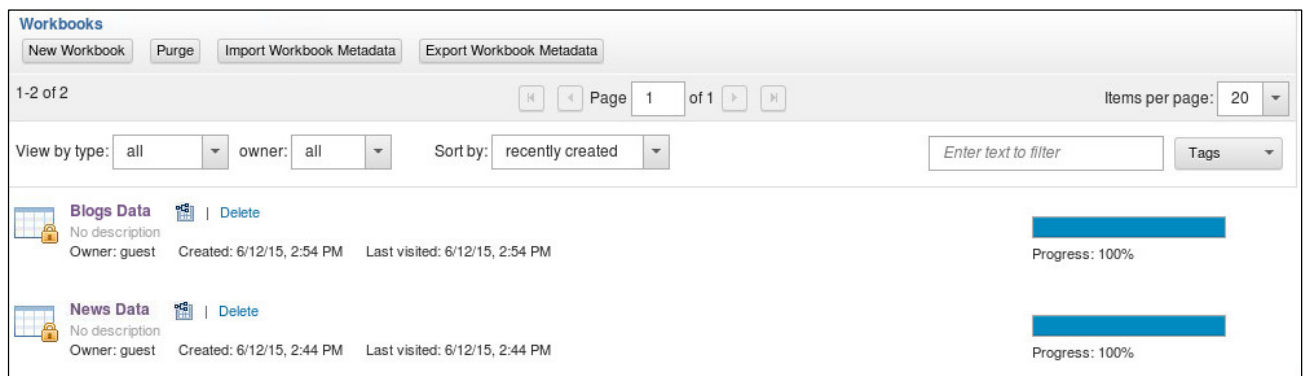
	Header
1	
2	["PostSize":6597,"Crawled":"2012-02-17 18:0
3	["PostSize":3739,"Crawled":"2012-02-13 14:1
4	["PostSize":2431,"Crawled":"2012-03-07 15:3
5	["PostSize":3982,"Crawled":"2012-03-26 17:3
6	["PostSize":4433,"Crawled":"2012-03-23 12:4
7	["PostSize":2820,"Crawled":"2012-03-23 05:1
8	["PostSize":2900,"Crawled":"2012-03-26 09:3
9	["PostSize":2745,"Crawled":"2012-03-15 11:0
10	["PostSize":2651,"Crawled":"2012-03-05 17:3
11	["PostSize":2730,"Crawled":"2012-03-15 07:3
12	["PostSize":5917,"Crawled":"2011-03-30 13:1
13	["PostSize":4881,"Crawled":"2012-03-07 18:5

14. Beside **Line Reader**, click **Edit workbook reader** .
 15. In the **Select a reader** list, select **JSON Array**, and then click **Set reader** .
- You can see now that the data is properly parsed.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

16. At the bottom of the New Workbook window, click **Save workbook** . Using the same steps, you will create the Blog Data workbook.
17. Click the **Workbooks** link (breadcrumb) to go back to the BigSheets home page.
18. Click the **New Workbook** button.
19. On the New Workbook window, under **Name**, type **Blogs Data**. You can leave the description field blank.
20. On the **DFS Files** tab, navigate to **/user/biadmin/** and select the **blogs-data.txt** file.
21. Specify the **JSON Array** reader.
22. At the bottom of the New Workbook window, click **Save workbook** .
23. Click the **Workbooks** link to return to the BigSheets home page.






The results appear as follows:






Task 2. Creating and editing child workbooks.

1. Click **News Data** to open the workbook. You will create a child workbook.
2. Beside **News Data**, click the **Build new workbook** button. You will now remove unnecessary columns.
3. In the **IsAdult** column header, expand the dropdown menu, and then click **Remove**.
4. In any column header, expand the dropdown menu, and then click **Organize Columns**.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

5. Click **Remove item**  beside each of the following columns to delete them.
 - **Crawled**
 - **Inserted**
 - **MoreoverUrl**
 - **PostSize**
 - **URL**
6. Click **Apply settings**  to confirm your actions.
7. Beside **News Data(1)**, click **Edit workbook name** , beside **Name**, type **NewsDataRevised**, and then click **Save Tag** .
8. Expand the **Save** dropdown, and then click **Save and Exit**.
9. Note that you could also name the workbook in here. Since we had already named it, click the **Save** button.
 Part of BigSheets is the feature to view what you intend to do on the subset of the data. In order for the changes to take effect on the full dataset, you must run the workbook. When you save and exit from the workbook, you will be prompted to Run the workbook.
10. Click **Run** to run the workbook on the full set of data.
11. Click **Workbooks**.
 You will use the steps above to revise the Blogs Data workbook.
12. Click **Blogs Data**.
13. Beside **Blogs Data**, click the **Build new workbook** button.
 You will now remove unnecessary columns.
14. In the **IsAdult** column header, expand the dropdown menu, and then click **Remove**.
15. In any column header, expand the dropdown menu, and then click **Organize Columns**.
16. Click **Remove item**  beside each of the following columns to delete them.
 - **Crawled**
 - **Inserted**
 - **Url**
 - **PostSize**

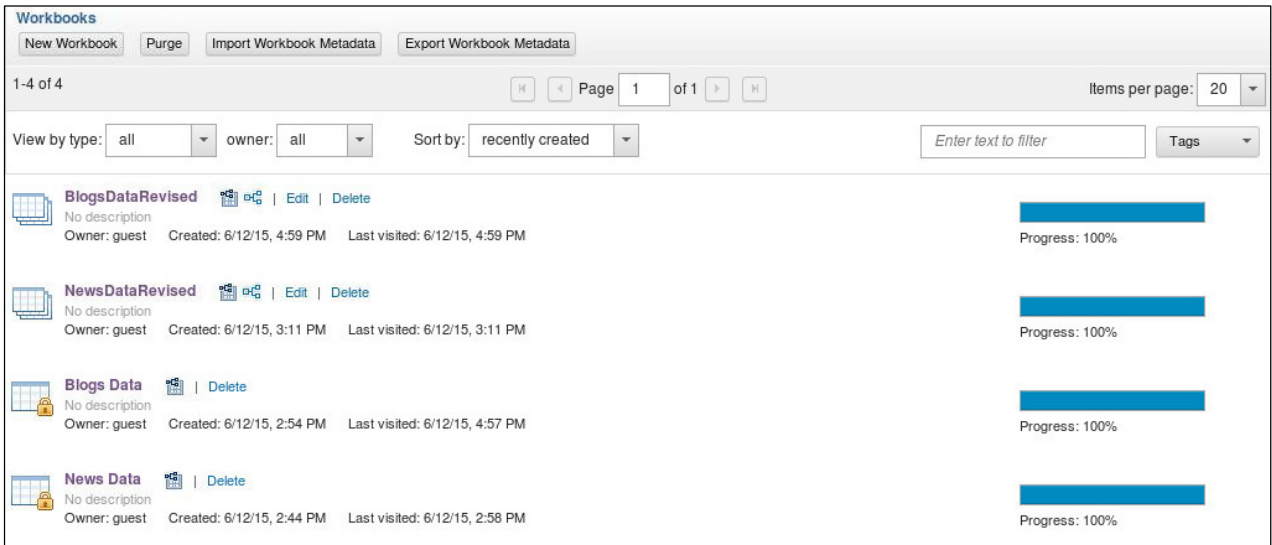
This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

17. Click **Apply settings**  to confirm your actions.
18. Beside **Blogs Data(1)**, click **Edit workbook name** , beside **Name**, type **BlogsDataRevised**, and then click **Save Tag** .
19. Expand the **Save** dropdown, and then click **Save and Exit**.
20. Note that you could also name the workbook in here. Since we had already named it, click the **Save** button.

Part of BigSheets is the feature to view what you intend to do on the subset of the data. In order for the changes to take effect on the full dataset, you must run the workbook. When you save and exit from the workbook, you will be prompted to Run the workbook.

21. Click **Run** to run the workbook on the full set of data.

The results appear as follows:



The screenshot displays the 'Workbooks' section of the IBM BigInsights interface. At the top, there are buttons for 'New Workbook', 'Purge', 'Import Workbook Metadata', and 'Export Workbook Metadata'. Below these, a pagination bar shows '1-4 of 4' items, 'Page 1 of 1', and 'Items per page: 20'. A filter section includes 'View by type: all', 'owner: all', 'Sort by: recently created', a search box 'Enter text to filter', and a 'Tags' dropdown. The main content area lists four workbooks, each with a grid icon, name, description, owner, creation/last visited timestamps, and a progress bar.

Workbook Name	Description	Owner	Created	Last Visited	Progress
BlogsDataRevised	No description	guest	6/12/15, 4:59 PM	6/12/15, 4:59 PM	100%
NewsDataRevised	No description	guest	6/12/15, 3:11 PM	6/12/15, 3:11 PM	100%
Blogs Data	No description	guest	6/12/15, 2:54 PM	6/12/15, 2:57 PM	100%
News Data	No description	guest	6/12/15, 2:44 PM	6/12/15, 2:58 PM	100%

Leave The BigInsights - BigSheets window open for the next task.

Task 3. Combining workbooks.

In this task, you will be merging the two workbooks: NewsDataRevised and BlogsDataRevised 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 task, 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 NewsDataRevised and bring in the BlogsDataRevised sheet using the load operation.

1. Open the **NewsDataRevised** workbook.
2. Click the **Build new workbook** button.
3. Expand the **Add sheets** dropdown, and then click the **Load** sheet.
4. Under **Sheet Name**, type **BlogsDataRevised**.
5. Click the **BlogsDataRevised** workbook.

6. Click **Apply Settings**  to run the load operation.

Once the operation completes, at the bottom left of the window, you will notice a new tab showing the Blog sheet that was just loaded.

Now you are ready to create a union of these two sheets.

7. Click **Add sheets** and select the **Union** operation.
8. Name the sheet: **Union2Collection**.
9. From the **Select sheet** dropdown, add the **BlogsDataRevised** and the **NewsDataRevised** sheet to be used for the Union operation, and then click **Apply Settings**.

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

The results appear as follows:



10. Click **Save**, click **Save & Exit**, in the **Name** box, type **NewsAndBlogsData**, and then click **Save**.

11. Run the workbook.

The results appear as follows:

The screenshot displays the 'Workbooks' section of the IBM BigInsights interface. At the top, there are buttons for 'New Workbook', 'Purge', 'Import Workbook Metadata', and 'Export Workbook Metadata'. Below these, a pagination bar shows '1-5 of 5' items, 'Page 1 of 1', and 'Items per page: 20'. A filter section allows users to 'View by type' (all), 'owner' (all), and 'Sort by' (recently created). A search bar with the placeholder 'Enter text to filter' and a 'Tags' dropdown are also present. The main content area lists five workbooks, each with a grid icon, name, description, owner, creation and last visit timestamps, and a progress bar showing 100% completion.

Workbook Name	Description	Owner	Created	Last Visited	Progress
NewsAndBlogsData	No description	guest	6/15/15, 9:35 AM	6/15/15, 9:35 AM	100%
BlogsDataRevised	No description	guest	6/12/15, 4:59 PM	6/14/15, 11:46 PM	100%
NewsDataRevised	No description	guest	6/12/15, 3:11 PM	6/15/15, 9:28 AM	100%
Blogs Data	No description	guest	6/12/15, 2:54 PM	6/12/15, 4:57 PM	100%
News Data	No description	guest	6/12/15, 2:44 PM	6/12/15, 2:58 PM	100%

Task 4. Sorting and creating charts.

1. Open the **NewsAndBlogsData** workbook.
2. To create a new child workbook, click **Build new workbook**.
3. From any column options menu, point to **Sort**, and then click **Advanced**.
4. In the **Add Columns to Sort** list, add the columns **Language** and **Type** to be sorted.

Hint: Click  to add to the list.

5. Beside **Language**, select **Descending**, beside **Type** select **Ascending**, and then use the arrows to ensure that Language is the primary sort column.

The results appear as follows:

The screenshot shows the 'Advance Sort' dialog box. It has a section 'Add Columns to Sort:' with a dropdown menu showing 'Country' and a green plus icon to add more. Below this are 'Add All' and 'Remove All' buttons. The main area lists the selected columns: 'Language' and 'Type'. For 'Language', the sort order is 'Descending' with a red 'X' button to remove it. For 'Type', the sort order is 'Ascending' with a red 'X' button to remove it. At the bottom right, there is a green checkmark and a red 'X' button.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

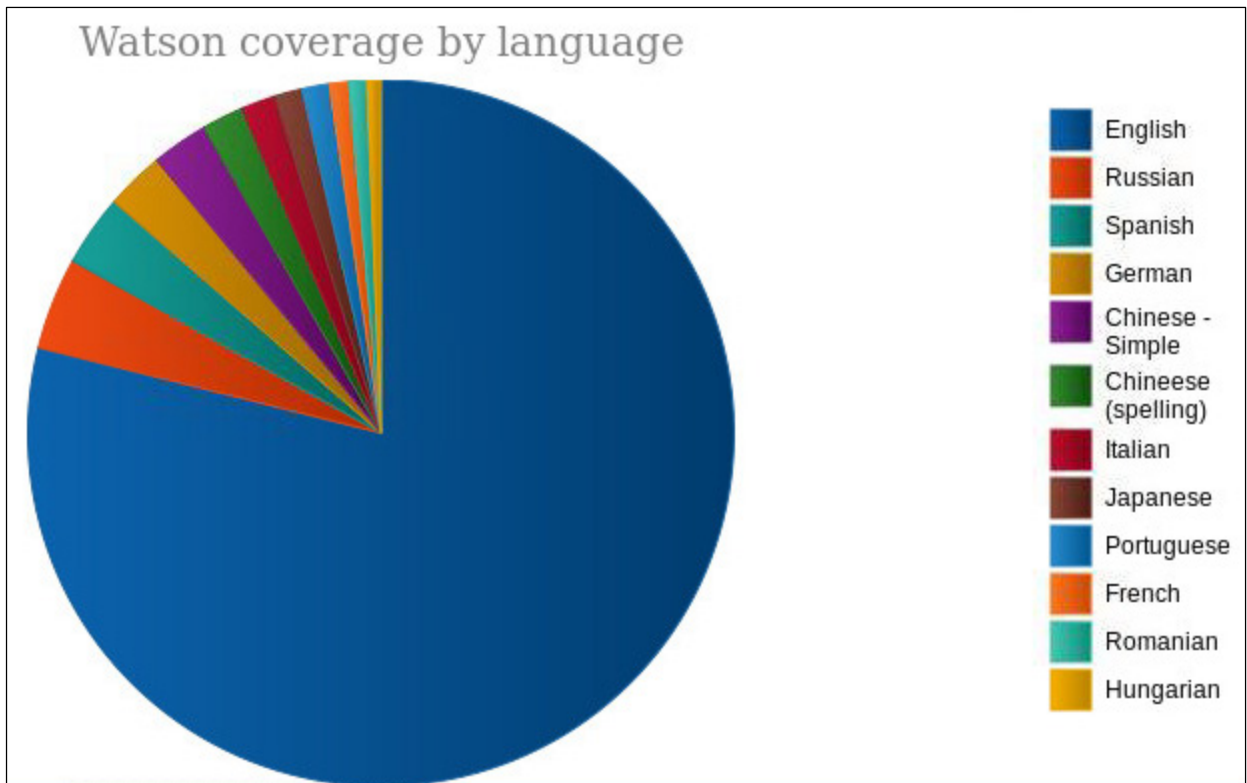
6. Click **Apply Settings** to run the sort operation.
7. Save and run the workbook as **NewsAndBlogsDataSorted**.
When the run completes, you will see more languages in the workbook.
Now you will create a graph to visualize your results.
8. Within the **NewsAndBlogsDataSorted** workbook, expand the **Add chart** dropdown, click **Chart**, and then click **Pie**
9. 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**
 - Style: **Pie**
10. Click **Apply Settings** to create the chart, and then click **Run**.
Once the run completes, you will see that English has the largest slice of pie.
What is the second most appeared language?

11. Point to the second largest piece of the pie.

Russia has the second largest piece of the pie.

Move the mouse pointer over the fifth and sixth largest 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.

The results appear as follows:



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 if you have some time, you may play around with the different sheets and functions to see other types of operations you can perform on the data.

12. Close all open windows.

Results:

You have created a BigSheets workbook and a chart from it to visualize your data.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE