IBM Training

## Demonstration 1

Analyzing Social Media and Structured Data

At the end of this demonstration, you should be able to:

- Create a Filter sheet
- Load data from a workbook into a second workbook
- Join data from two sheets
- Use the Pivot function on BigSheets data
- Utilize the visualization capabilities of BigSheets




*Demonstration 1: Analyzing Social Media and Structured Data*

# Demonstration 1: Analyzing Social Media and Structured Data

**Purpose:**
**In this demonstration, you will use BigSheets to analyze social media data and use various forms of BigSheets visualization to present the results. You will use all you have learned so far in this course to complete this demonstration. In the second portion of this demonstration, you will join the social media data with DBMS data to do further analysis and finally conclude with visualizing the data with BigSheets.**

User/Password: **biadmin/biadmin**
**Root/dalvm3**
Service Password: **ibm2blue**

## Task 1. Load the test data into HDFS.

The social media data used in this demonstration has been loaded on your local system. This data was created using the Boardreader application that comes with BigInsights. (Although to use this app, you must have a license.)
Also, data was exported from a database system info a *csv* format that you will use. You should have loaded the news-data.txt and blogs-data.txt from previous lab demonstrations already, but they may have been in a different directory. In any case, go ahead and upload them now in this section. Along with those two files, you will also be working with the a CSV file that was exported from a RDBMS to show you that not only can you work with big data, but also data from a RDBMS.

1. Open up a new terminal.
2. Do a listing of the /user/biadmin folder on the HDFS.

```
hdfs dfs -ls /user/biadmin
```

3. Make a note of which files already exist and use the following to upload each of the three files, as needed.

   If any of the files already exist, the command will fail.

   ```
   hdfs dfs -put
   /home/biadmin/labfiles/bigsheets/RDBMS_data.csv
   /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
   ```

4. If needed, do another listing of the **/user/biadmin** directory on hdfs to make sure you have all three files under **/user/biadmin**.

   ```
   hdfs dfs -ls /user/biadmin
   ```

## Task 2. Creating a BigSheets Workbook from the RDBMS data.

1. Open up the **BigSheets** page from BigInsights Home.
2. Click **New Workbook**.
3. Name the workbook, **Media Contacts**.
4. Select the ***RDBMS-data.csv*** file under **/user/biadmin**.
5. Change the **BigSheets** reader by selecting **Edit workbook reader**  .
6. Select the **Comma Separated Value (CSV) Data** reader, uncheck the **Headers included?** checkbox and then click **Set reader**  .
7. Click **Save workbook**  to create the workbook.

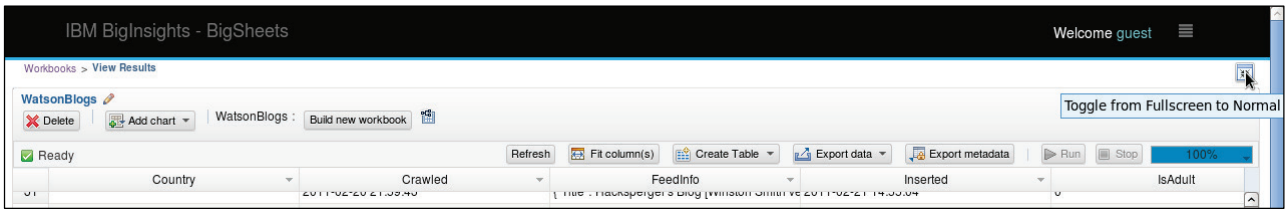
## Task 3. Creating BigSheets Workbooks from the Boardreader data.

Now you will create two more workbooks, one for each of the two files from the Boardreader application.

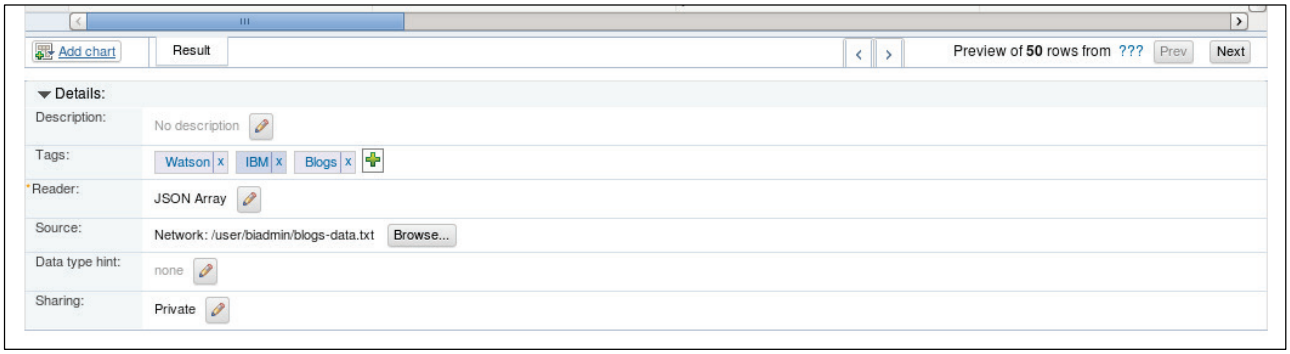
1. To go back to the **BigSheets** home page, click the **Workbooks** breadcrumb.
2. Click **New Workbook**.
3. Name the new workbook **WatsonBlogs**.
4. Select **blogs-data.txt** under **/user/biadmin**.

5. Click **Edit workbook reader**, select the **JSON Array** reader and then click **Set reader**  .
6. Now with the data formatted properly, scroll down (if you have to) and click **Save workbook**.

   Tag your sheet. This allows you to quickly search and manage your workbooks.
7. Scroll to the bottom of the workbook to view the workbook details. If you do not see the detail information, click **Toggle from Normal to Fullscreen** which is located above the workbook data in the upper right of the *BigSheets* page.




8. In the **Tags** section, click **Add new tag**  .
9. In the Tag value box type **Watson**, and then click **Save tag**  .
10. Repeat by adding **IBM** and **Blogs** as tag values.

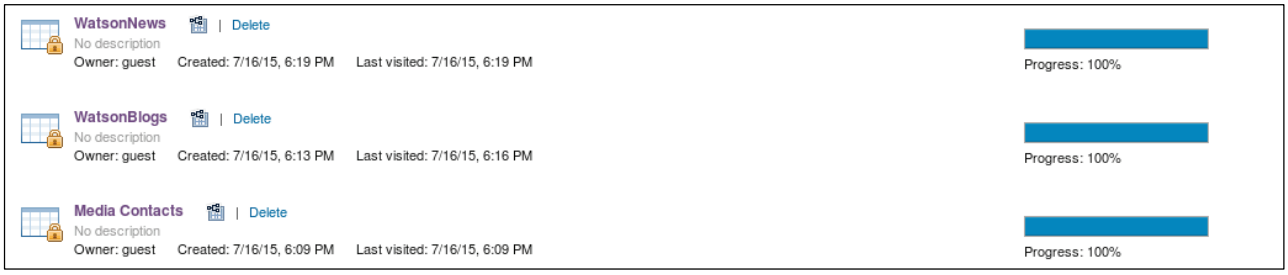


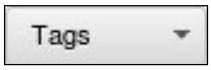

You will now create a second workbook for the news-data.txt file.

11. Create a new workbook called **WatsonNews**.
12. Add **Watson**, **IBM**, and **News** as tags to this new workbook.

13. Click the **Workbooks** breadcrumb tab to get a list of all workbooks.

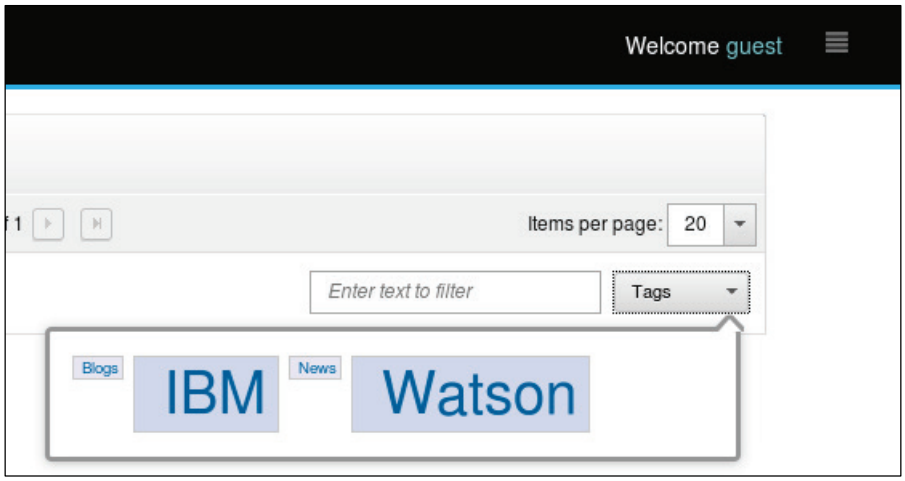
    You should see the three workbooks you created in this demonstration as well as others from previous demonstrations.




14. Click **Tags** 

.

    A cloud list of tags gets displayed.




15. Click **News**.

    Only those workbooks with that tag are displayed.

16. In the **filter** field, (to the left of the *Tags* button) type **tag:Watson**, and then press **Enter**.

    This is another way of filtering on a tag.

## Task 4. Tailoring a workbook.

1. Click the **WatsonNews** workbook.

   Create a child workbook based on this master workbook.

2. Click **Build new workbook**.

3. Change the workbook name by clicking **Edit workbook name**  , and then change the name to **WatsonNewsRevised**.

4. To view more of the columns, click **Fit column(s)**.

   You are not going to need the IsAdult column.

5. Click the drop-down arrow for the **isAdult** column and then select **Remove**.
   The data was not actually deleted. The mapping to the column was removed.
   You need to remove a number of columns.
6. Click the drop-down arrow for any column and then select **Organize Columns**.
   Clicking a *red X* removes that column.
7. Click the **red X** to remove the following columns.
   a. Crawled
   b. Inserted
   c. MoreoverUrl
   d. PostSize
8. Click **Apply settings**.
9. Click **Save**, select **Save & Exit** and then click **Save**.
10. Run your workbook.
11. In the **Watson Blogs** workbook, follow the same steps as above to remove the following columns.
    e. Crawled
    f. Inserted
    g. isAdult
    h. PostSize
12. Save this new workbook as **WatsonBlogsRevised**.
13. Run the **WatsonBlogsRevised** workbook.

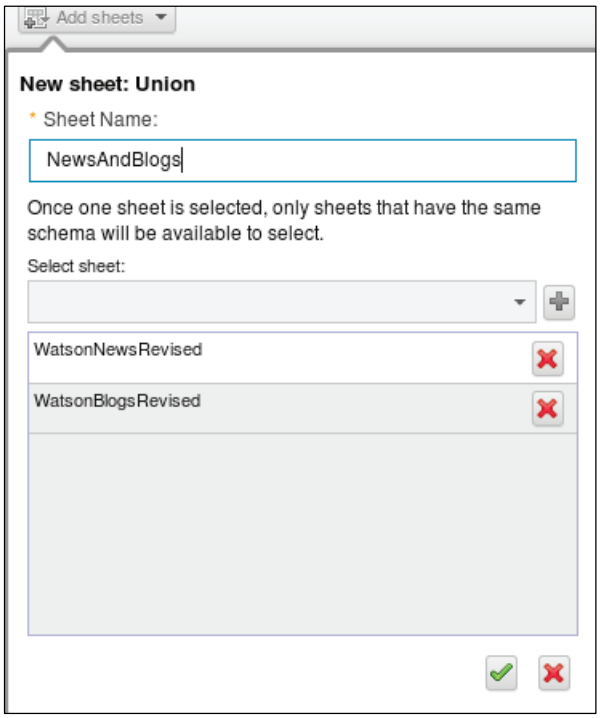
## Task 5. Union the two workbooks.

Because both workbooks have the same structure now, you can union them. This becomes the basis for exploring the coverage of IBM Watson across the sources that the Boardreader provided.

1. Click the **Workbooks** breadcrumb tab and select the **WatsonNewsRevised** workbook.
2. Click **Build new workbook**.
3. Click **Add sheets**.
4. Click **Load** and then click **WatsonBlogsRevised**.
5. Change the sheet name to **WatsonBlogsRevised** and then click **Apply settings**.
   Now the data from both revised workbooks is accessible in order to add the data into a single sheet.

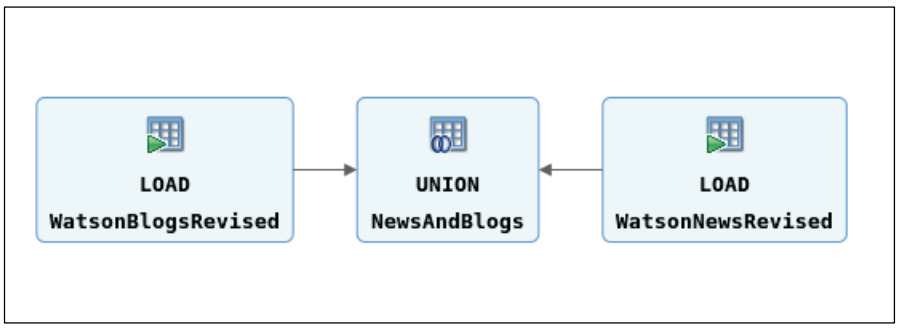
6. Click **Add sheets** and then select **Union**.
7. Change the Sheet Name to **NewsAndBlogs**.
8. In the **Select Sheet** drop down, select **WatsonNewsRevised**, and then click **Add sheet**.
9. Select the **WatsonBlogsRevised**, and then select **Apply settings**.

   If you forgot to change the name of the sheet, you can click the drop-down on the sheet's tab and choose to rename it.




10. Save the workbook as **Watson News and Blogs** and then exit the workbook.
11. **Run** the workbook.
12. Click **Workbook Diagram**  .

    The Watson News and Blogs workbook was created by loading two workbooks and then doing a union.



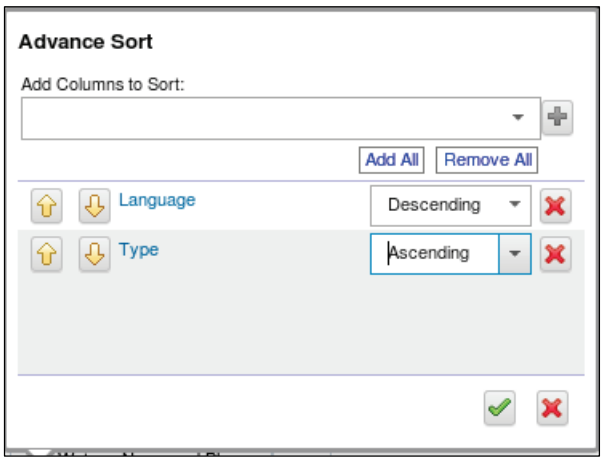

13. Close this window.

14. Click **Workflow Diagram**  which is to the right of Build new workbook.
    This shows the workbooks that were used to create the current workbook. Close this window.


## Task 6. Exploring the Workbook.

You should still be in the Watson News and Blogs workbook and you do not want to modify this workbook.

1. Click **Build new workbook**.
2. Click **Edit workbook name** and then change the workbook name to **WatsonSorted**.
   You will now take a closer look at the languages and types of posts in the data.
3. Click the drop-down menu for any column.
4. Select **Sort->Advanced**.
5. Click the **Add Columns to Sort** drop down box, select **Language** and then click **Add column sort**  .
6. Choose to sort the values in the **Language** column in **Descending** sequence.
7. Click the **Add Columns to Sort** drop down box, select **Type** and then click **Add column sort**.
   Keep the default of Ascending sequence.




8. Click **Apply settings**.

9. Click **Fit column(s)** so that you can see both the Language and the Type columns.

   The sort that was performed is only running on a subset of the data. When you save and run the workbook, the sort gets applied to all of the data so you might see some differences. For example, the subset of data has only a few records where the Language is Vietnamese. This changes when all of the data is used.

10. **Save**, **exit** and **run** your workbook.

## Task 7. Visualize your data.

You should be in the **WatsonSorted** workbook. Assume that you are interested in seeing the number of posts associated with each language.

1. Click **Add chart**.
2. Select the **chart** hyperlink and choose **Pie**.
3. Add the following in the Pie chart info.
   a. Chart Name - Language Coverage
   b. Title - IBM Watson Coverage by Language
   c. Value - Language
   d. Count - Count occurrences of X axis values
   e. Sort by: Value
   f. Limit - 12
4. Click **Apply settings**.
5. Click **Run** and wait for all of the data to be processed.
6. Move the cursor over the various segments to see that **Chinese - Simple** is next to a segment for **Chinese (Spelling)**.

   After reviewing you want to have all of the Chinese posts in a single segment.
7. Click **Edit**.

   To do the combination trick, you need to add a new column. Move the cursor over the *Language* column. Then, click the drop-down that is displayed.
8. Select **Insert Right->New Column**.

9. Name this new column **Language_Revised**.

   There cannot be any spaces in column names.

   After saving the column name, the cursor was moved to the **fx** field. The idea is that you are going to provide a function that is to be used to populate this new column.

   You want to look at the Language value for each row. If that value begins with Chin, then you want the value in the Language_Revised column for that row to be Chinese. Otherwise, you want the value to be what is in the *Language* column.
10. Type the following in the ***fx*** field.

    ```
    IF(SEARCH('Chin*', #Language) > 0, 'Chinese', #Language)
    ```
11. Click **Save formula**  .
12. **Save**, **exit** and **run** the workbook.
13. Click the drop-down menu for the **Language Coverage** tab at the bottom to modify the chart settings.
14. Select **Chart Settings**.
15. Change the **Value** field to **Language_Revised** and then click **Apply settings**.
16. Click the **Language Coverage** tab to bring up the modified chart.
17. Click **Run**.

    Now you can see that the Chinese segment is the second largest.

## Task 8. Joining Social with Structured Data.

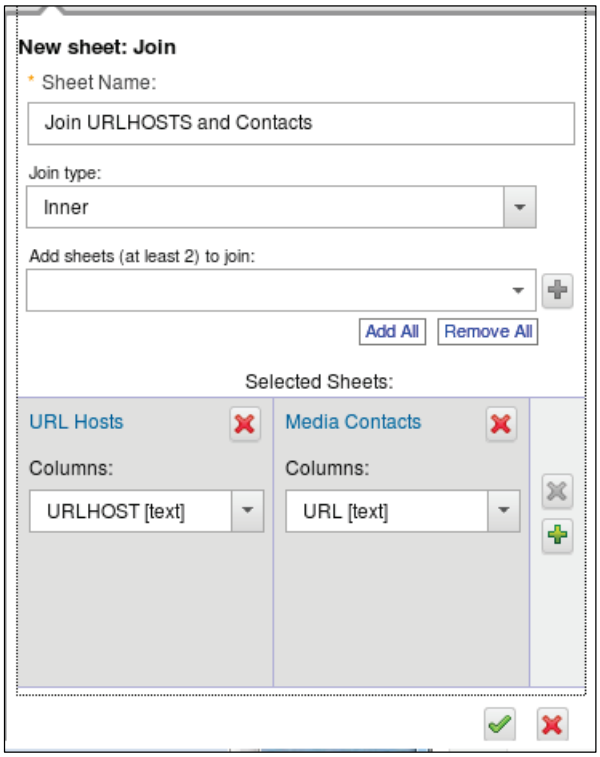
Last but not least, you will begin to work with the RDBMS data, pulled into a BigSheets workbook at the beginning of this exercise. As you might remember, you pulled data into a workbook and named it *Media Contacts*. Now, join this structured data with the Social Media data. By joining these two workbooks, you can explore how corporate media outreach efforts correlate to coverage by third-party websites.

1. In order to start with a workbook that has all of the items in it that you need, open the **Watson News And Blogs workbook**.
2. Build a new workbook and name it **Watson Media Analytics**.

   Again, you need the URLHOST column added to your new workbook.
3. Ensure the **URLHosts** column is showing values. If not, add **URL** from **sheet settings**.
4. Add a sheet that runs the **URLHOST** function and **carries over** all of the columns.
5. Name the sheet **URL Hosts**.

6. Add a sheet that **Loads the Media Contacts workbook** into your new, **Watson Media Analytics** workbook.
7. Name this sheet **Media Contacts**.
8. To make the last column of the **Media Contacts** more clear, rename it **Last_Contact**.

   Move the cursor over the *header4* column and click the drop-down. Choose to rename the column.
9. Change the name of the **header3** column to **URL.**
10. Join the data, add a **new sheet** and then select **Join**.
11. Name the sheet **Join URLHOSTS and Contacts**.
12. From the **Join Type** drop-down menu, select **inner** join.
13. In the **Add sheets** drop-down, select **URL Hosts** and then click **Add sheet**.
14. Add the **Media Contacts** sheet and then click **Add sheet**.
15. For the **URL Hosts** sheet, select the **URLHOST** column and for the **Media Contacts** sheet, select the **URL** column.




16. Click **Apply settings**.

17. As an additional way to make your results look more intuitive, you can reorganize the order of the columns by using the Organize Columns option or by dragging and dropping the column. Do that by a left-click-mouse-grab on the letter above the column name. Also, another option is selecting Fit Columns.
18. **Save**, **exit**, and **run** the workbook.

    You have now joined different data sources together in a BigSheets workbook.

**Results:**
**You joined social media data with DBMS data, and then analyzed the data with BigSheets.**