

BigInsights Cloud Tutorial: Analytics for Hadoop on Bluemix

Step 4: Exploring data with BigSheets

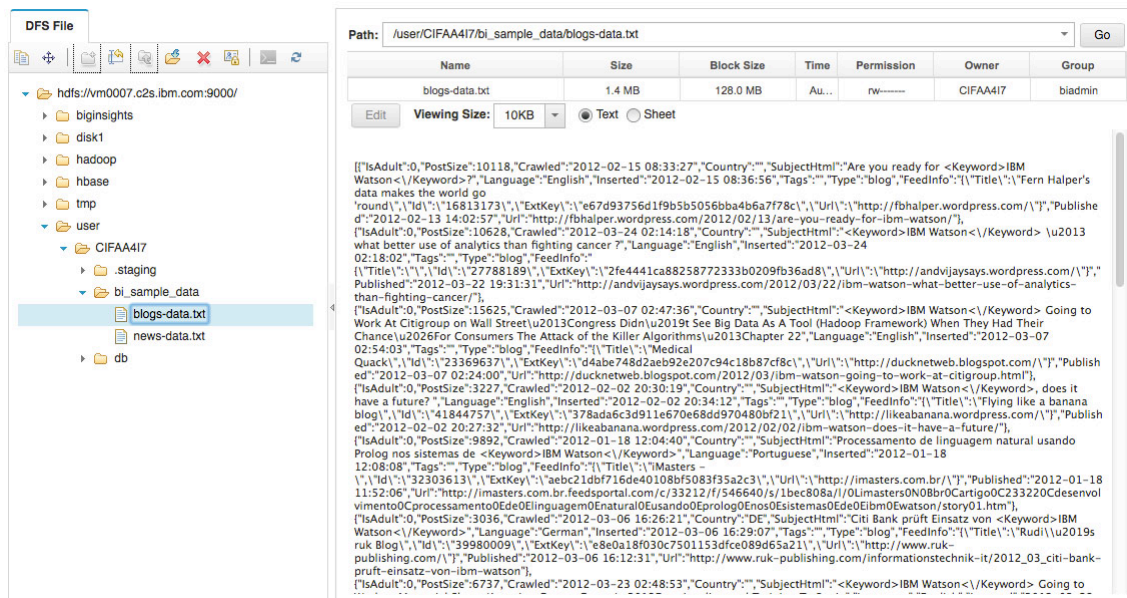
Learn how to immediately analyze and review big data using BigSheets: a browser based spreadsheet-like tool that can model, filter, combine, and chart data collected from multiple sources, and which ships with all versions of BigInsights.

In the past 3 brief tutorials, we have signed up for a Bluemix account, selected and started up our Analytics for Hadoop service, Explored the BigInsights web console and have uploaded data to BigInsights.

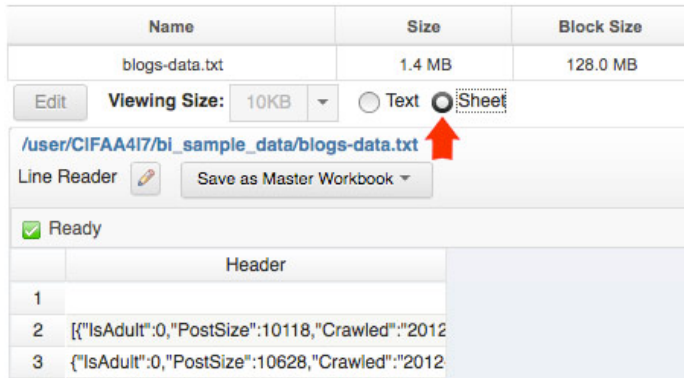
In this tutorial we will learn how to analyze and review big data using BigSheets: a browser based spreadsheet-like tool that can model, filter, combine, and chart data collected from multiple sources, and which ships with all versions of BigInsights.

Now let's get to the data.



1. In your sample data directory (which was created in the last tutorial), select the blogs-data.txt file that you have uploaded – You will notice a raw snapshot view of this data in the right hand window

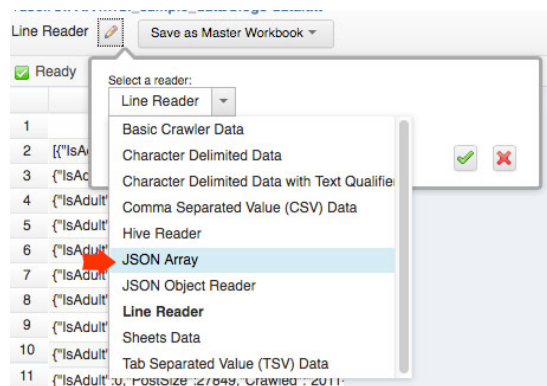


2. Above the raw snapshot of data - click the **Sheet** radio button. In the **Preview** area of the window, you see that the data is not displayed properly. It is formatted in a JSON Array structure (which stands for - JavaScript Object Notation)

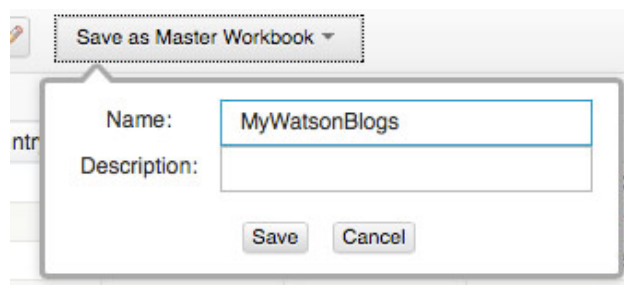






3. So we want to apply a reader to map the data to the spreadsheet format:

- Click the Edit icon ().
- Select **JSON Array** from the drop-down list, and click the green check mark (). You immediately see the data map to the columns and rows of the spreadsheet-like interface in the **Preview** area.



- Since the data columns exceed the viewing space, click **Fit column(s)**. The first eight columns display in the **Preview** area. **Note:** Depending on the size of your web browser window, you might need to scroll to see **Fit column(s)**.
- Click **Save as Master Workbook**.
- In the **Name** field, enter MyWatsonBlogs.



- f. In the **Description** field, enter Watson blog data from blogs-data.txt, then click **Save**.
4. Click the **Workbooks** link in the breadcrumb at the top of the window. You are moved to the BigSheets tab, and you see your new master workbook, MyWatsonBlogs.
5. Click **New Workbook**.
6. In the **Name** field, enter MyWatsonNews.
7. In the **Description** field, enter Watson news feed data from news-data.txt.
8. Under **File**, navigate to the /user/<USERID>/ directory and select the news-data.txt file. The right side of the window displays the file name and contents. This data is also in JSON Array format.
9. Click the Edit icon () , select **JSON Array** from the drop-down list, and click the green check mark () to apply the reader.
10. Since the data columns exceed the viewing space, click **Fit column(s)**. The first eight columns display in the **Preview** area.
11. Save the master workbook by clicking the green check mark () in the lower right corner of the screen. **Note:** Depending on the size of your web browser window, you might need to scroll to see the green check mark () . You are moved to the BigSheets tab, and you see your new workbook, MyWatsonNews.
12. View both new master workbooks by clicking the **Workbooks** link.

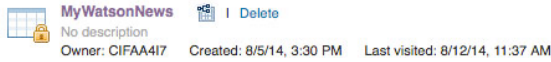
You are now ready to explore the data that you loaded.

Typically, before you analyze and explore data, you must tailor its format and content. Here you will create child workbooks from each master workbook and remove unwanted columns to refine the amount and type of your data.

In addition to protecting the original data, master workbooks set the data format (including the data types for the columns). Therefore, you must create child workbooks in which to modify your data. Child workbooks inherit their format and data from their master workbooks, but you can tailor their attributes to display only necessary data.

Steps

1. From the **BigSheets** tab of the InfoSphere® BigInsights™ Web Console, select the MyWatsonNews master workbook.



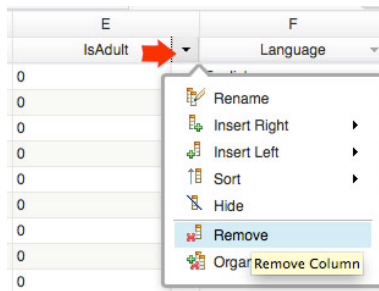
2. Click **Build new workbook**. A new workbook is created with the name: MyWatsonNews(1).

3. Rename the workbook by clicking the Edit icon (✎), entering MyWatsonNewsRevised, and clicking the green check mark (✅).



4. To see columns A through H within your web browser, click **Fit column(s)**.

5. For your analysis, you do not need the IsAdult column (column E). Remove it by clicking the down arrow in the column heading and selecting **Remove**.

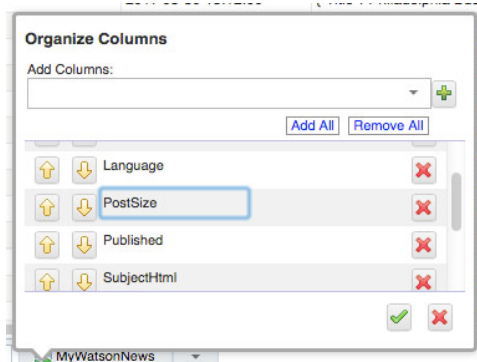


When you remove columns from a child workbook, you delete only the data from the child workbook. The master workbook on which this child workbook is based always contains the original data as it was loaded. If you decide later that you want the IsAdult data in your analysis, you can create another child workbook from the MyWatsonNews master workbook.

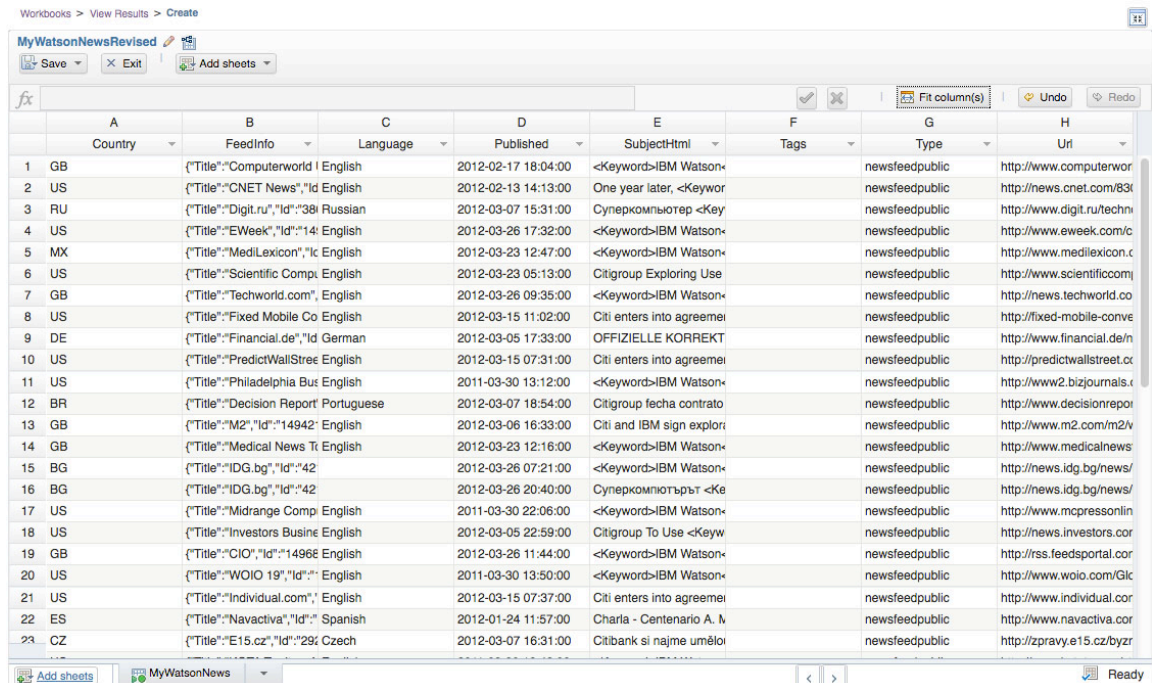
6. As you review the data in this MyWatsonNewsRevised child workbook, you decide that you do not need several other columns. You can use the same method, as in the previous step, to remove them one at a time or remove multiple columns at once:

- a. Click the down arrow in any column heading, and select **Organize Columns**.
- b. Click the red X (✖) next to the following columns to mark them for removal:
 - i. Crawled
 - ii. Inserted
 - iii. MoveoverUrl
 - iv. PostSize
- c. Click the green check mark (✅) to remove the columns. **Tip:** If you

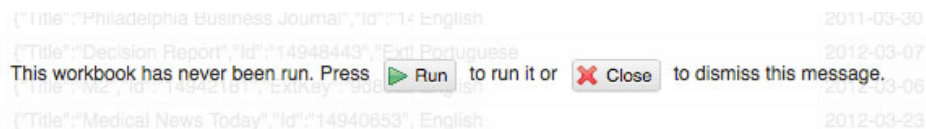
accidentally remove more columns than you intend, you can undo your last action by clicking **Undo**.



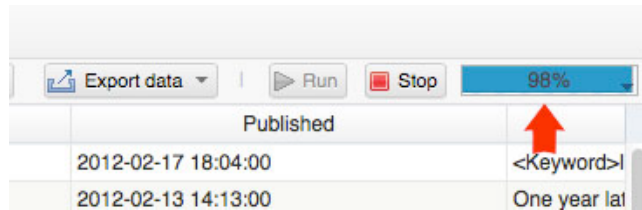
7. Click **Fit column(s)** to resize the remaining columns.





8. Save and exit the workbook by clicking **Save** and selecting **Save & Exit**. If you are prompted with a Save workbook window, you can save the workbook with or without entering a description.
9. You are prompted with the message "This workbook has never been run." Press Run to run it or Close to dismiss this message. Click **Run**.



You see a progress indicator in the upper right corner of the window.



Until now, you have been working with a subset of the Watson and internal IBM data. BigSheets keeps only a limited number of rows in memory. The lower right corner displays a message that indicates you are seeing only a simulated sample of 50 rows of data. When you run the data, you apply all changes that you made since the last time you saved the workbook to the full data set. The progress bar monitors the progress of the job. Behind the scenes, Pig scripts initiate MapReduce jobs. The runtime performance depends upon the volume of data that is associated with your data collection and the system resources that are available.

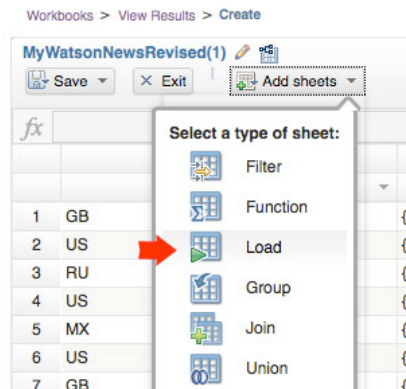
10. Now, create a child workbook from the MyWatsonBlogs master workbook, and remove the columns that are not needed for your analysis:
 - a. To return to the page that displays all your workbooks, click the **Workbooks** link.
 - b. Select the MyWatsonBlogs master workbook, and click **Build new workbook**. A new workbook is created with the name: MyWatsonBlogs(1).
 - c. Rename the new child workbook by clicking the Edit icon (), typing MyWatsonBlogsRevised, and clicking the green check mark ().
 - d. Use the **Organize Columns** function to remove the following columns:
 - i. Crawled
 - ii. Inserted
 - iii. IsAdult
 - iv. PostSize
 - e. Remember to select the green check mark in the Organize Columns window. Now, the MyWatsonNewsRevised and MyWatsonBlogsRevised workbooks contain the same columns. To merge workbooks, each workbook must contain the same data types and columns, or *schema*.
 - f. Save and exit the workbook.
 - g. When prompted, click **Run** to apply the changes that you made to the child workbook.

Because both new child workbooks have the same schema, you can merge them into a new workbook, where you can explore and analyze your data.

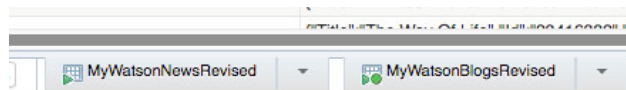
To merge the data, we will create a new workbook from an existing workbook, then load the data from the second workbook into the new workbook.

Steps

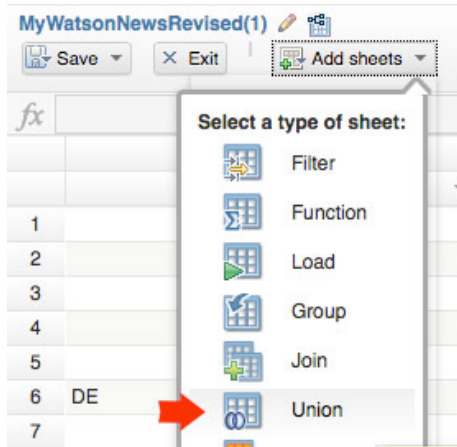
13. Go back to the workbooks link (breadcrumb), and select the MyWatsonNewsRevised workbook.
14. Click **Build new workbook**. The name of the workbook is MyWatsonNewsRevised(1), indicating that it is a child workbook of MyWatsonNewsRevised. You change the name of this workbook later when you save and exit the workbook.
15. Click **Add sheets**, and select **Load**.



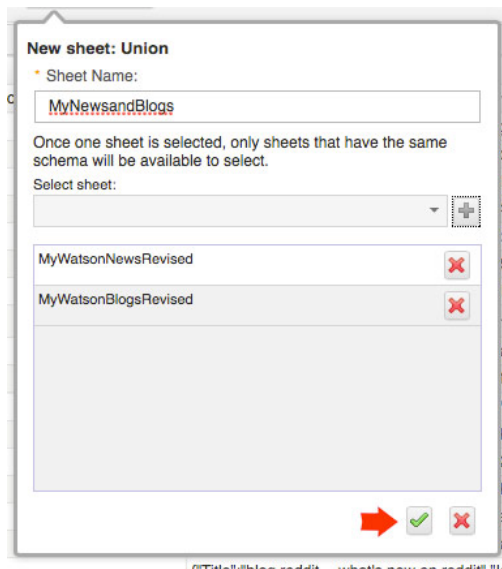
16. In the Load window, select the MyWatsonBlogsRevised workbook link from the list of existing workbooks.
17. In the **Sheet Name** field, enter MyWatsonBlogsRevised, and click the green check mark (✓). In the Load window, you see details of the columns and the first few rows of data in that workbook. At the bottom of your workbook, you see two tabs, **MyWatsonNewsRevised** and **MyWatsonBlogsRevised**.



18. Click **Add sheets**, and select **Union**.



19. In the **Sheet Name** field of the New Sheets: Union dialog, enter MyNewsandBlogs to indicate that this sheet contains the merged data.
20. From the **Select sheet** drop-down list, select the MyWatsonNewsRevised sheet, click the green plus sign (+) to add the sheet (you see the sheet move to the bottom of the dialog), then do the same for MyWatsonBlogsRevised – adding it to the union, ...and then click the green check mark (✓). Your workbook now displays the new tab, MyNewsandBlogs, at the bottom of your screen.



21. Click **Save**. When prompted for a name and description, enter Watson News Blogs in the **Name** field and Combined news and blogs data in the **Description** text box, and click **Save**.




You have now successfully combined the blog and news data into one workbook, where you can analyze and explore the data as a whole.

Next, you will learn how to create columns by grouping similar information. You want to discover how many news articles and blog posts are written in each

language. You accomplish this goal by using the Group sheet and its functions to combine, calculate, and sort the language data.

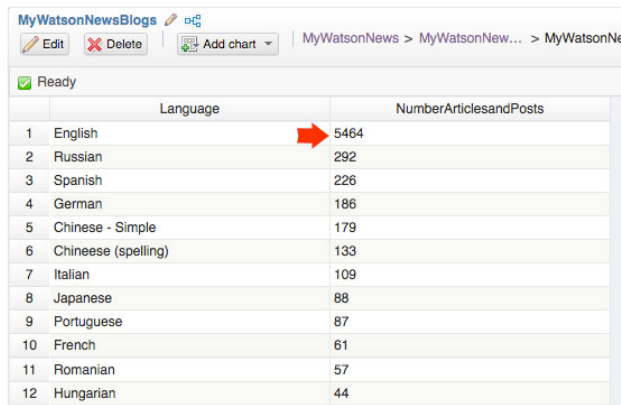
First, use the Calculate function to count the number of articles and posts by language. Then, sort the column by language to display the most popular languages first.

Steps

22. Still in the MyWatsonNewsBlogs workbook - Click **Add Sheets**, and select **Group**.
23. In the New Sheet: Group window, complete the required information:
 - a. In the **Sheet name** field, enter Group by language.
 - b. From the **Group by columns** drop-down list, select **Language**, and click the green plus sign () to add the column. The **Language** column name displays in the bottom of the dialog.
 - c. At the bottom of the window, click the **Calculate** tab.
 - d. In the **Create columns based on groups** text box, enter NumberArticlesandPosts, and click the green plus sign ()
 - e. From the **NumberArticlesandPosts** drop-down list, select **COUNT**.
 - f. From the **Column** drop-down list, select **Language**, then click the green check mark ()
24. On the Group by language sheet, you see two columns, **Language** and **NumberArticlesandPosts**. The **Language** column displays all the languages from the News and Blogs sheet. The **NumberArticlesandPosts** column counts the number of posts in each language.
25. To see the most common languages for posts about IBM Watson™, sort the Pivot sheet by the number of posts. Click the drop-down arrow to the right of the **NumberArticlesandPosts** column, select **Sort**, and click **Descending**. You see that English is the most popular language with 3169 posts, followed by Russian, Spanish, and Chinese - Simple. But notice that Chinese (spelling) and Chinese - Traditional are also near the top of the list. In other versions of this tutorial you will be able to merge these all into one – but for the sake of brevity we will not cover.
26. Click **Save & Exit** to save and close the workbook.

Click **Run** to save, sort, and process the entire data set for the workbook. You see a progress indicator in the upper right corner of the window. After you run the

workbook, you see different results for the number of English posts in the **NumberArticlesandPosts** column, 5464.



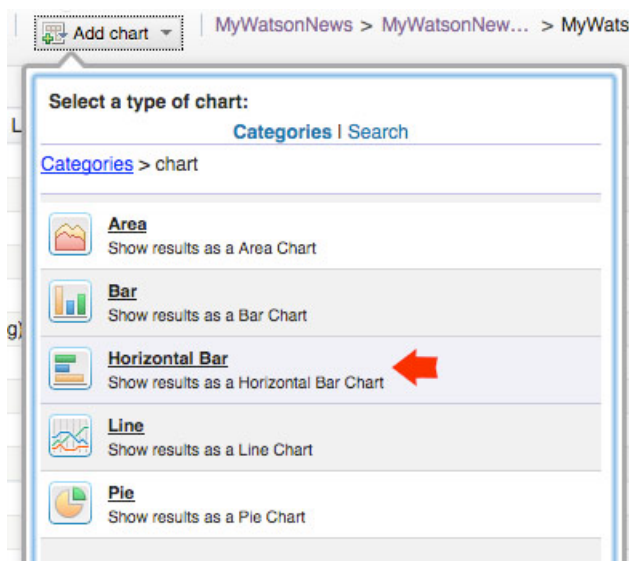
| | Language | NumberArticlesandPosts |
|----|--------------------|------------------------|
| 1 | English | 5464 |
| 2 | Russian | 292 |
| 3 | Spanish | 226 |
| 4 | German | 186 |
| 5 | Chinese - Simple | 179 |
| 6 | Chinese (spelling) | 133 |
| 7 | Italian | 109 |
| 8 | Japanese | 88 |
| 9 | Portuguese | 87 |
| 10 | French | 61 |
| 11 | Romanian | 57 |
| 12 | Hungarian | 44 |

For the last part of this tutorial we will take a very quick look at the visualization capabilities of BigInsights.

BigSheets provides various charts and maps. A *chart* plots data points in a grid, such as a typical pie or bar chart. A *cloud* shows the importance of values by displaying the size of the words relative to their importance. A *map* contains charts that represent geographic data, such as a heat map that shows the concentration of data points geographically.

Procedure

1. Open the MyWatsonNewsBlogs workbook, click **Add chart**, and then select **chart** > **Horizontal Bar**. It might take a few minutes to populate the categories of charts the first time.



2. In the New chart: Horizontal Bar window, enter or select the following values:
- In the **Chart Name** field, enter Language Coverage. The chart name is the name that displays on the tab at the bottom of the worksheet.
 - In the **Title** field, enter IBM Watson Coverage by Language. The title of the chart displays at the top of the chart.
 - From the **X Axis** drop-down list, select **NumArticlesandPosts**.
 - In the **X Axis Label**, enter Number of posts.
 - From the **Y Axis** drop-down list, select Language.
 - In the **Y Axis Label**, enter Language of post.
 - From the **Sort By** drop-down list, select **X Axis**. You want to sort by the number of posts.
 - From the **Occurrence Order** drop-down list, select **Descending**. You want to see the language with the highest number of posts first.
 - In the **Limit** field, enter 12. You want to see only the top 12 languages by the number of posts.

New chart: Horizontal Bar

Chart Name: Language Coverage

Title: IBM Watson Coverage by Language

X Axis: NumberArticlesandPosts

X Axis Label: Number of Posts

Y Axis: Language

Y Axis Label: Language of Post

Sort By: X Axis

Occurrence Order: Descending

Limit: 12

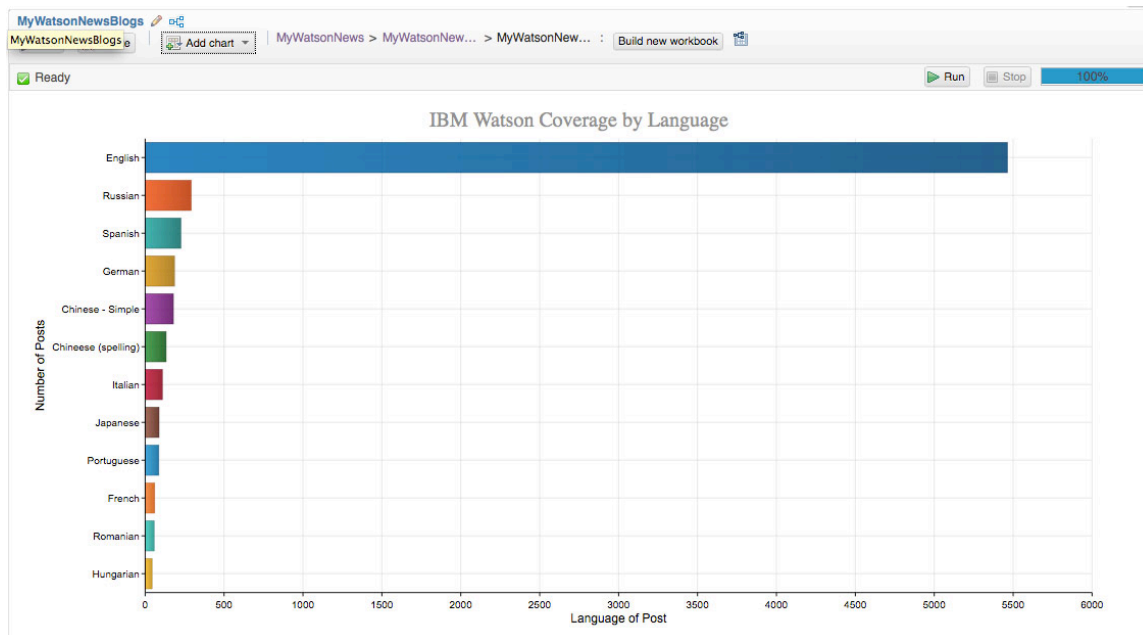
Template: Soda Cap

Style: Regular HBar

- j. Click the green check mark () to preview the chart with sample data.

Click **Run** to generate the chart from the full set of workbook data. Even though you see the preview chart immediately, the actual chart is not displayed until you see 100% on the progress bar. It might take some time to generate the chart from the full set of data. Use the progress bar to monitor the status of the completed chart. After the bar chart is generated, you can see that Russian is the second most popular

language for posts. You also see that the fifth and sixth most popular languages are variations on the Chinese language.



This is a very limited tutorial that gets you hands on quick with Hadoop in the Cloud – but lacks the rich content that our longer, more robust labs offer. If you have any interest in learning more – I would highly recommend that you check out the latest BigInsights labs in the [Hadoop Developer Community](#) (..Including text analytics, BigSQL and other unique technologies that we have built on top of apache open source Hadoop,... and download our FREE non-production BigInsights QuickStart Edition – all of the power of our enterprise version, free to use in non production environments.