BigInsights Cloud Tutorial: Analytics for Hadoop on Bluemix

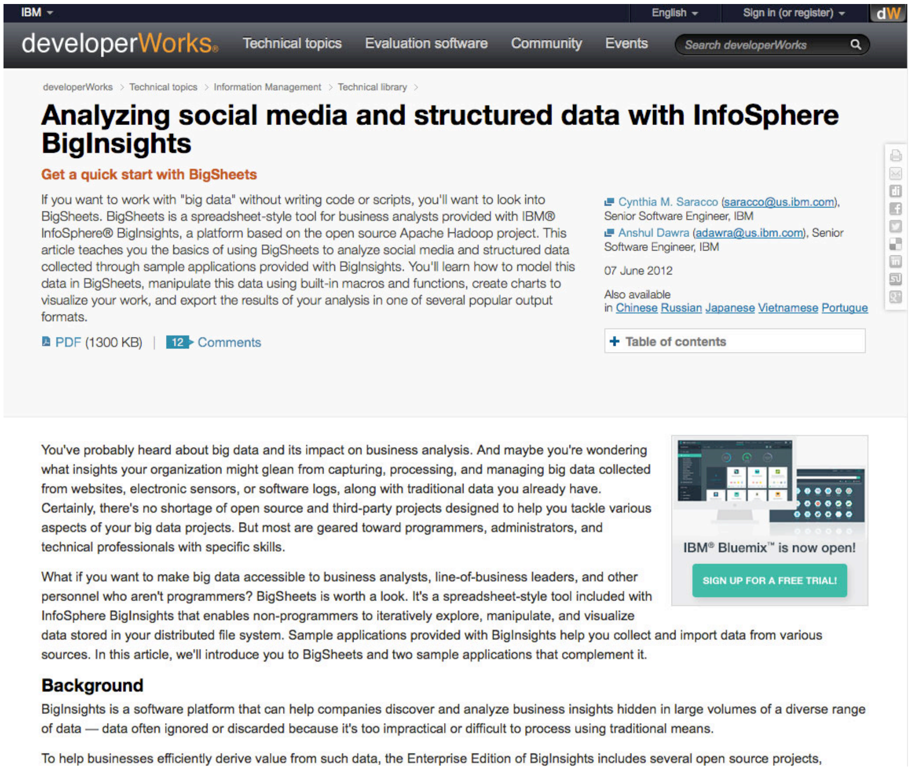
Tutorial 3: Loading data into BigInsights

Learn how to quickly load data into your InfoSphere® BigInsights™ Hadoop environment, the options available to you and how to quickly review the data at hand.

Business data is stored in various formats and sources. Before you import your data into the InfoSphere BigInsights distributed file system, you must determine what questions you want to answer through analysis, identify the data type of your sources, and use the tools and procedures that best fit your business need.

You can use InfoSphere BigInsights with your existing infrastructure or data warehouse to import data and content in its original formats, or you can import huge volumes of at-rest (static) data or incoming data in motion (continually updated data). After you import your data, you can explore the data separately or combine the data to complete exploration and analysis.

Many businesses might want to examine the popularity of a specific brand or service in social media. The data that is provided for this lesson is the result of a BoardReader application search for the instances of the phrase "IBM Watson™" on the Internet. This search is detailed in the developerWorks® article, Analyzing social media and structured data with InfoSphere BigInsights: Get a quick start with BigSheets.



Analyzing social media and structured data with InfoSphere BigInsights

Get a quick start with BigSheets

If you want to work with "big data" without writing code or scripts, you'll want to look into BigSheets. BigSheets is a spreadsheet-style tool for business analysts provided with IBM® InfoSphere® BigInsights, a platform based on the open source Apache Hadoop project. This article teaches you the basics of using BigSheets to analyze social media and structured data collected through sample applications provided with BigInsights. You'll learn how to model this data in BigSheets, manipulate this data using built-in macros and functions, create charts to visualize your work, and export the results of your analysis in one of several popular output formats.

Cynthia M. Saracco (saracco@us.ibm.com), Senior Software Engineer, IBM
Anshul Dawra (adawra@us.ibm.com), Senior Software Engineer, IBM

07 June 2012

Also available in Chinese Russian Japanese Vietnamese Portugue

PDF (1300 KB) | 12 Comments

+ Table of contents

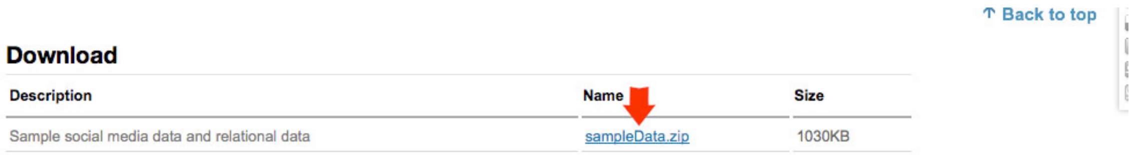
You've probably heard about big data and its impact on business analysis. And maybe you're wondering what insights your organization might glean from capturing, processing, and managing big data collected from websites, electronic sensors, or software logs, along with traditional data you already have. Certainly, there's no shortage of open source and third-party projects designed to help you tackle various aspects of your big data projects. But most are geared toward programmers, administrators, and technical professionals with specific skills.

What if you want to make big data accessible to business analysts, line-of-business leaders, and other personnel who aren't programmers? BigSheets is worth a look. It's a spreadsheet-style tool included with InfoSphere BigInsights that enables non-programmers to iteratively explore, manipulate, and visualize data stored in your distributed file system. Sample applications provided with BigInsights help you collect and import data from various sources. In this article, we'll introduce you to BigSheets and two sample applications that complement it.

Background

BigInsights is a software platform that can help companies discover and analyze business insights hidden in large volumes of a diverse range of data — data often ignored or discarded because it's too impractical or difficult to process using traditional means.

To help businesses efficiently derive value from such data, the Enterprise Edition of BigInsights includes several open source projects, You are going to need this data as referenced in the above article, so please download it before proceeding. (The download is at the bottom of the article, yet feel free to download directly, here)



↑ Back to top

Download

Description	Name	Size
Sample social media data and relational data	sampleData.zip	1030KB

Accept the terms and conditions and save the file article_sampleData to your local system. After you unzip the file, the article_sampleData folder should contain the files RDBMS_data.csv, blogs-data.txt, news-data.txt, and a README.txt file that details the data output.

Make sure to note where you save these files – we will be uploading them to the cloud shortly.

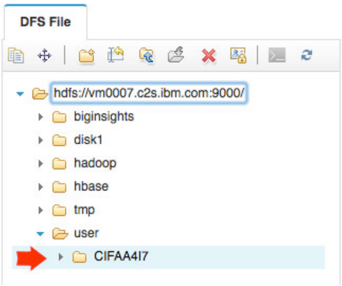
For this tutorial, and the related tutorial on BigSheets (Tutorial 4 for our Bluemix series), only the news and blog data that was returned by the search is used. The returned data was slightly modified to contain only a subset of the information that the BoardReader application collects from blogs and news feeds. The full-text/HTML content of posts, news items, and certain metadata, was removed to keep the size of each file manageable. For information on BoardReader, click here.

Now, let's look at the file system and create a folder where we will load the data.

Creating a Folder

1. Open the InfoSphere BigInsights Web Console.

2. From the **Files** tab, select the **HDFS** folder. (Note – the HDFS folder may already be open) – Within the **HDFS** folder navigate to the **USERS** folder and find your specific folder with your user id (note: this is the ID that is located on the IBM Analytics for Hadoop startup page). Navigate to this folder.



3. Create a directory to store this data in the distributed file system. Click the **Create**

Directory folder icon ( ) in the Files toolbar.

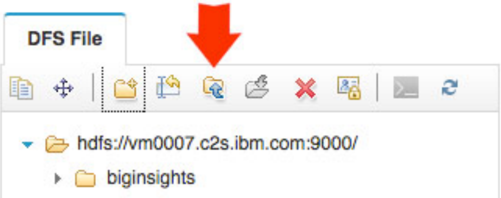
4. Name your directory. For this lesson, in the **HDFS** directory, create the directory bi_sample_data. For example the home directory is /user/<USERID>/.

You now have a directory to store all of your source data files and application results. Next, we will load the data.

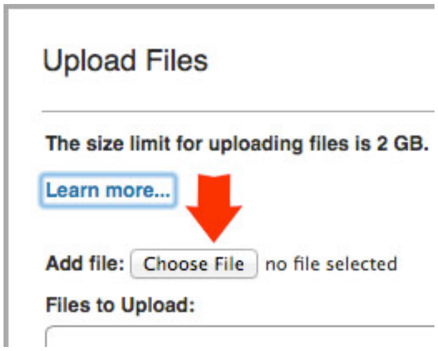
Loading the data:

Note: There are many ways to load data into BigInsights, but for this tutorial we are going to aim to keep it simple. We will be using the native Upload funciton in BigInsights, which is only recommended for smaller data sets (under 2GB). For larger data sets, you can leverage the Distributed File Copy application.

1. Make sure that you are in your 'bi_sample_directory' folder
2. On the files toolbar, click the **Upload..** icon – A new window should open 'Upload Files'

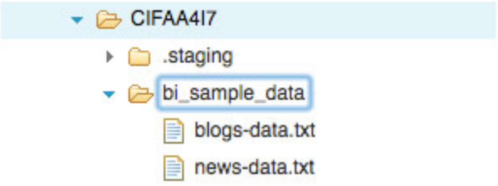


3. Click on the **Choose File** button in this new window.



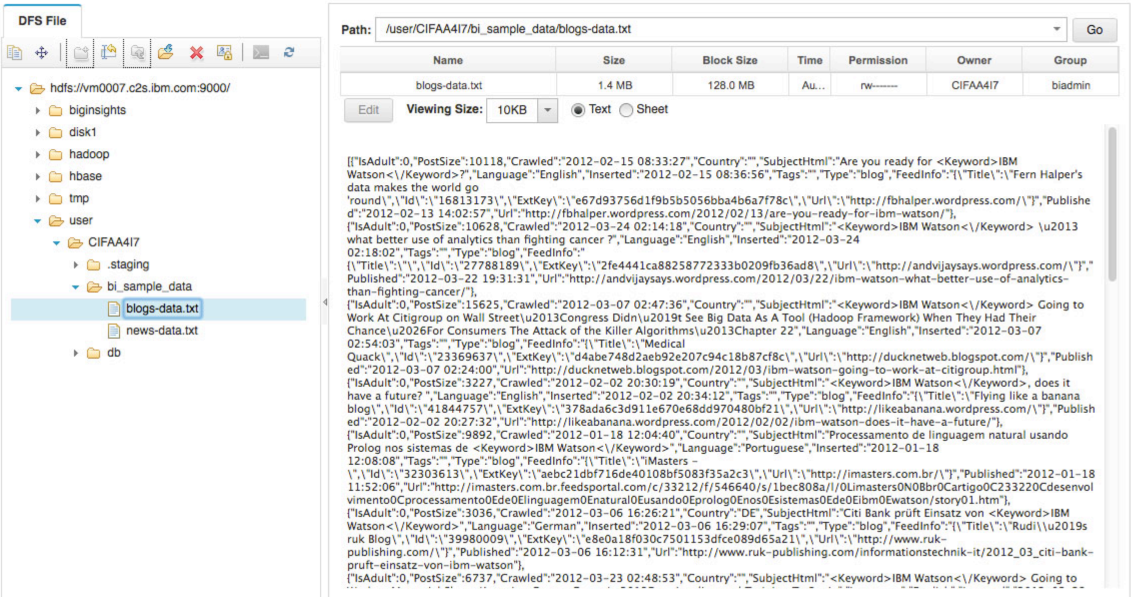
The size limit for uploading files is 2 GB.

4. Navigate your local file system to where you saved the blogs-data.txt & news-data.txt files earlier in the tutorial.
5. Choose blogs-data.txt and upload
6. Repeat the process and Choose news-data.txt
7. Both files should be in the **Upload Files** window. Click the Ok button.
8. Navigate back to your 'bi_sample_directory' folder to ensure that the files are in there. Note: you can hit refresh on the toolbar to refresh the view.



You have now loaded the data into BigInsights that we will be using for the 4th tutorial.

If you want to take a sneak peak at what you just loaded – click on either of the files that we just brought in – the window to the right will show you a quick sample of the semi structured data.



Path: /user/CIFAA4I7/bi_sample_data/blogs-data.txt

Name	Size	Block Size	Time	Permission	Owner	Group
blogs-data.txt	1.4 MB	128.0 MB	Au...	rw-------	CIFAA4I7	biadmin

Viewing Size: 10KB ● Text ○ Sheet

```
[{"IsAdult":0,"PostSize":10118,"Crawled":"2012-02-15 08:33:27","Country":"","SubjectHtml":"Are you ready for <Keyword>IBM
Watson<\/Keyword>?","Language":"English","Inserted":"2012-02-15 08:36:56","Tags":"","Type":"blog","FeedInfo":"{\"Title\":\"Fern Halper's
data makes the world go
'round\",\"Id\":\"16813173\",\"ExtKey\":\"e67d93756d1f9b5b5056bba4b6a7f78c\",\"Url\":\"http://fbhalper.wordpress.com/\"}","Publishe
d":"2012-02-13 14:02:57","Url":"http://fbhalper.wordpress.com/2012/02/13/are-you-ready-for-ibm-watson/"},
{"IsAdult":0,"PostSize":10628,"Crawled":"2012-03-24 02:14:18","Country":"","SubjectHtml":"<Keyword>IBM Watson<\/Keyword> \u2013
what better use of analytics than fighting cancer ?","Language":"English","Inserted":"2012-03-24
02:18:02","Tags":"","Type":"blog","FeedInfo":"
{\"Title\":\"\",\"Id\":\"27788189\",\"ExtKey\":\"2fe4441ca88258772333b0209fb36ad8\",\"Url\":\"http://andvijaysays.wordpress.com/\"}","
Published":"2012-03-22 19:31:31","Url":"http://andvijaysays.wordpress.com/2012/03/22/ibm-watson-what-better-use-of-analytics-
than-fighting-cancer/"},
{"IsAdult":0,"PostSize":15625,"Crawled":"2012-03-07 02:47:36","Country":"","SubjectHtml":"<Keyword>IBM Watson<\/Keyword> Going to
Work At Citigroup on Wall Street\u2013Congress Didn\u2019t See Big Data As A Tool (Hadoop Framework) When They Had Their
Chance\u2026For Consumers The Attack of the Killer Algorithms\u2013Chapter 22","Language":"English","Inserted":"2012-03-07
02:54:03","Tags":"","Type":"blog","FeedInfo":"{\"Title\":\"Medical
Quack\",\"Id\":\"23369637\",\"ExtKey\":\"d4abe748d2aeb92e207c94c18b87cf8c\",\"Url\":\"http://ducknetweb.blogspot.com/\"}","Publish
ed":"2012-03-07 02:24:00","Url":"http://ducknetweb.blogspot.com/2012/03/ibm-watson-going-to-work-at-citigroup.html"},
{"IsAdult":0,"PostSize":3227,"Crawled":"2012-02-02 20:30:19","Country":"","SubjectHtml":"<Keyword>IBM Watson<\/Keyword>, does it
have a future? ","Language":"English","Inserted":"2012-02-02 20:34:12","Tags":"","Type":"blog","FeedInfo":"{\"Title\":\"Flying like a banana
blog\",\"Id\":\"41844757\",\"ExtKey\":\"378ada6c3d911e670e68dd970480bf21\",\"Url\":\"http://likeabanana.wordpress.com/\"}","Publish
ed":"2012-02-02 20:27:32","Url":"http://likeabanana.wordpress.com/2012/02/02/ibm-watson-does-it-have-a-future/"},
{"IsAdult":0,"PostSize":9892,"Crawled":"2012-01-18 12:04:40","Country":"","SubjectHtml":"Processamento de linguagem natural usando
Prolog nos sistemas de <Keyword>IBM Watson<\/Keyword>","Language":"Portuguese","Inserted":"2012-01-18
12:08:08","Tags":"","Type":"blog","FeedInfo":"{\"Title\":\"iMasters -
\",\"Id\":\"32303613\",\"ExtKey\":\"aebc21dbf716de40108bf5083f35a2c3\",\"Url\":\"http://imasters.com.br/\"}","Published":"2012-01-18
11:52:06","Url":"http://imasters.com.br.feedsportal.com/c/33212/f/546640/s/1bec808a/l/0Limasters0N0Bbr0Cartigo0C233220Cdesenvol
vimento0Cprocessamento0Ede0Elinguagem0Enatural0Eusando0Eprolog0Enos0Esistemas0Ede0Eibm0Ewatson/story01.htm"},
{"IsAdult":0,"PostSize":3036,"Crawled":"2012-03-06 16:26:21","Country":"DE","SubjectHtml":"Citi Bank prüft Einsatz von <Keyword>IBM
Watson<\/Keyword>","Language":"German","Inserted":"2012-03-06 16:29:07","Tags":"","Type":"blog","FeedInfo":"{\"Title\":\"Rudi\\u2019s
ruk Blog\",\"Id\":\"39980009\",\"ExtKey\":\"e8e0a18f030c7501153dfce089d65a21\",\"Url\":\"http://www.ruk-
publishing.com/\"}","Published":"2012-03-06 16:12:31","Url":"http://www.ruk-publishing.com/informationstechnik-it/2012_03_citi-bank-
pruft-einsatz-von-ibm-watson"},
{"IsAdult":0,"PostSize":6737,"Crawled":"2012-03-23 02:48:53","Country":"","SubjectHtml":"<Keyword>IBM Watson<\/Keyword> Going to
```

We will tame and make sense of this data in the next tutorial, Exploring Data with BigSheets.