Course Guide

IBM BigInsights BigSheets v4.0

Course code DW644 ERC 1.0

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

IBM Training

August 2015

NOTICES

This information was developed for products and services offered in the USA.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service. IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not grant you any license to these patents. You can send license inquiries, in writing, to:

IBM Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
United States of America

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law: INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM websites are provided for convenience only and do not in any manner serve as an endorsement of those websites. The materials at those websites are not part of the materials for this IBM product and use of those websites is at your own risk.

IBM may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you. Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

TRADEMARKS

IBM, the IBM logo, ibm.com and BigInsights are trademarks or registered trademarks of International Business Machines Corp., registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the web at “Copyright and trademark information” at www.ibm.com/legal/copytrade.shtml.

Java and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.

Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.

© Copyright International Business Machines Corporation 2015.

This document may not be reproduced in whole or in part without the prior written permission of IBM.

US Government Users Restricted Rights - Use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM Corp.

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

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

Contents

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

© Copyright IBM Corp. 2012, 2015

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

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

© Copyright IBM Corp. 2012, 2015

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

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

© Copyright IBM Corp. 2012, 2015

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

Course overview

Preface overview

This course is designed to introduce the student to the capabilities of BigSheets. BigSheets is a component of IBM BigInsights through the Analyst and the Data Scientist module. It provides the analyst the ability to be able to visualize and analyze data stored on the HDFS using a spreadsheet type interface without any programming.

Intended audience

The course is designed for business analysts that does not want to deal with any coding to get insight on their data.

Topics covered

Topics covered in this course include:

IBM BigInsights BigSheets:

- Using BigSheets for data analysis
- Making data available to BigSheets
- Navigating a Workbook
- Discovering data with expressions, functions, and formulas
- Integrating with Big SQL
- Visualizing data with Big SQL

Course prerequisites

Participants should have:

- Students should be familiar with Hadoop and the Linux file system.
- Although not required, it would also be helpful for students to take the DW613 - IBM BigInsights Overview course to have a better understanding of how BigSheets fit into everything.
- Students can attend many free courses at www.bigdatauniversity.com to acquire the necessary requirements.

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

© Copyright IBM Corp. 2012, 2015

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

Document conventions

Conventions used in this guide follow Microsoft Windows application standards, where applicable. As well, the following conventions are observed:

- **Bold**: Bold style is used in demonstration and exercise step-by-step solutions to indicate a user interface element that is actively selected or text that must be typed by the participant.
- *Italic*: Used to reference book titles.
- CAPITALIZATION: All file names, table names, column names, and folder names appear in this guide exactly as they appear in the application.
  To keep capitalization consistent with this guide, type text exactly as shown.

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

© Copyright IBM Corp. 2012, 2015

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

Additional training resources

- Visit IBM Analytics Product Training and Certification on the IBM website for details on:
  - Instructor-led training in a classroom or online
  - Self-paced training that fits your needs and schedule
  - Comprehensive curricula and training paths that help you identify the courses that are right for you
  - IBM Analytics Certification program
  - Other resources that will enhance your success with IBM Analytics Software
- For the URL relevant to your training requirements outlined above, bookmark:
  - Information Management portfolio: http://www-01.ibm.com/software/data/education/
  - Predictive and BI/Performance Management/Risk portfolio: http://www-01.ibm.com/software/analytics/training-and-certification/

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

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM product help

Help type	When to use	Location
Task-oriented	You are working in the product and you need specific task-oriented help.	*IBM Product* - Help link
Books for Printing (.pdf)	You want to use search engines to find information. You can then print out selected pages, a section, or the whole book. Use Step-by-Step online books (.pdf) if you want to know how to complete a task but prefer to read about it in a book. The Step-by-Step online books contain the same information as the online help, but the method of presentation is different.	Start/Programs/*IBM Product*/Documentation
IBM on the Web	You want to access any of the following:	
	• IBM - Training and Certification	• http://www-01.ibm.com/software/analytics/training-and-certification/
	• Online support	• http://www-947.ibm.com/support/entry/portal/Overview/Software
	• IBM Web site	• http://www.ibm.com

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

© Copyright IBM Corp. 2012, 2015

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

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

© Copyright IBM Corp. 2012, 2015

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Using BigSheets for data analysis

IBM BigInsights v4.0

© Copyright IBM Corporation 2015
Course materials may not be reproduced in whole or in part without the written permission of IBM.

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

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

© Copyright IBM Corp. 2012, 2015

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Unit objectives

- Identify business and technical challenges in dealing with big data
- Describe how BigSheets can help with the business and technical challenges of big data
- Access BigSheets from the BigInsights Home console

© Copyright IBM Corporation 2015

Unit objectives

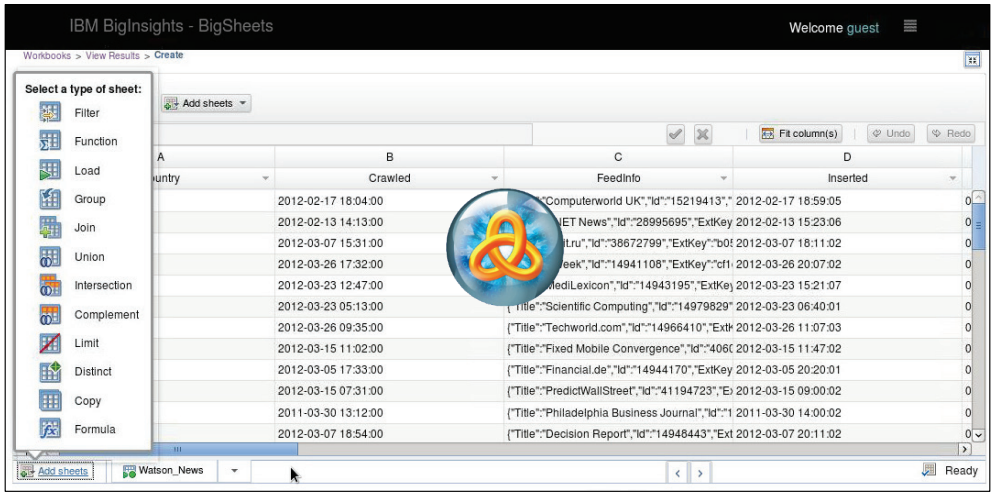
This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

What is BigSheets?

- Browser-based analytics tool for business users
- Spreadsheet like interface for analyzing big data
- A component of IBM BigInsights



Introduction to BigSheets

© Copyright IBM Corporation 2015

What is BigSheets?

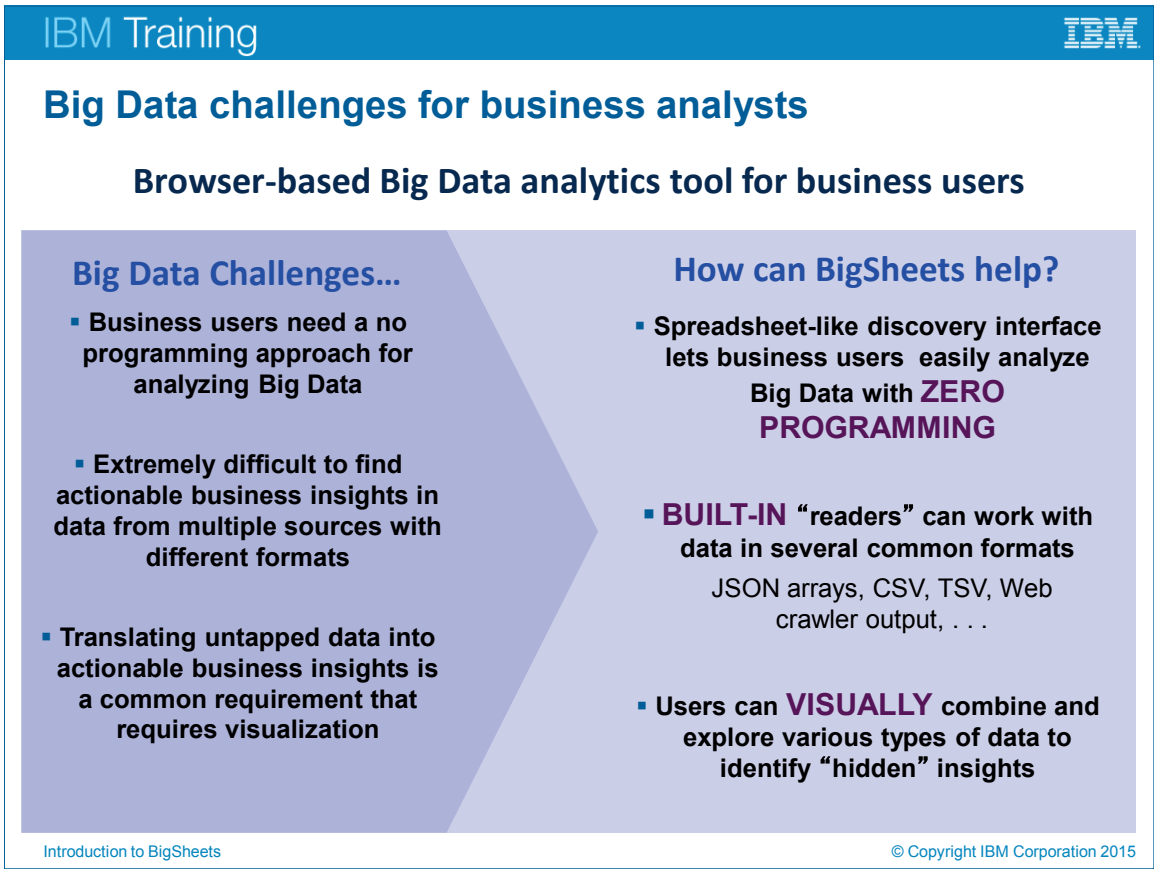
BigSheets is a browser-based analytic tool designed to work with Big Data. Unlike like many other Big Data tools, it is designed to support business users and non-technical professionals. To do so, it presents a familiar, spreadsheet-like interface that allows users to gather, filter, combine, explore, and visualize data from various sources.

IBM chose the spreadsheet as the model for organizing data because most users are already familiar with such software. If users want to represent the data in more complex ways,.

As an important part of IBM's Big Data strategy, BigSheets is a feature of IBM BigInsights

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

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.



Big Data challenges for business analysts

One of the Big Data challenges is "How do I get analysts to go out and analyze this data with zero programming". If you do not have such tooling, you create an unnatural dependency on development to code and build every piece of visualization and analysis. This is too expensive, inefficient, and time consuming. BigSheets gives you exactly this, with ZERO programming. Your analysts want to be able to visualize and analyze data in JSON, CVS and text file formats. They want a programming free crawler, and more, all of which is included in BigSheets. To the person using BigSheets, it looks like a spreadsheet, but under the covers, it generates PIG jobs to run on Hadoop.

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

© Copyright IBM Corp. 2012, 2015

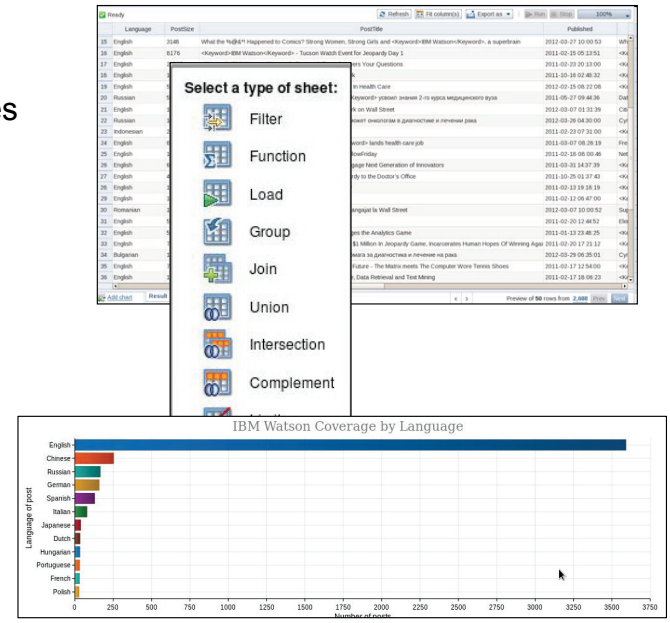
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

What can you do with BigSheets?

- Model “big data” collected from various sources in spreadsheet-like structures
- Filter and enrich content with built-in functions
- Combine data in different workbooks
- Visualize results through spreadsheets, charts
- Export data into common formats (if desired)

No programming knowledge needed!



Introduction to BigSheets

© Copyright IBM Corporation 2015

What can you do with BigSheets?

BigSheets is a browser-based visualization and analysis tool designed to help non-programmers work with Big Data. It ships with BigInsights Quick Start and Enterprise Editions.

With this tool, users model their big data in workbooks, or familiar spreadsheet-like tabular data structures. Once data is represented in a workbook, business analysts can filter and enrich its content using built-in functions and macros. Furthermore, analysts can combine data residing in different workbooks as well as generate charts and new "sheets" (workbooks) to visualize their data. They can even export data into a variety of common formats with a click of a button.

Here are some of the distinguishing characteristics of BigSheets:

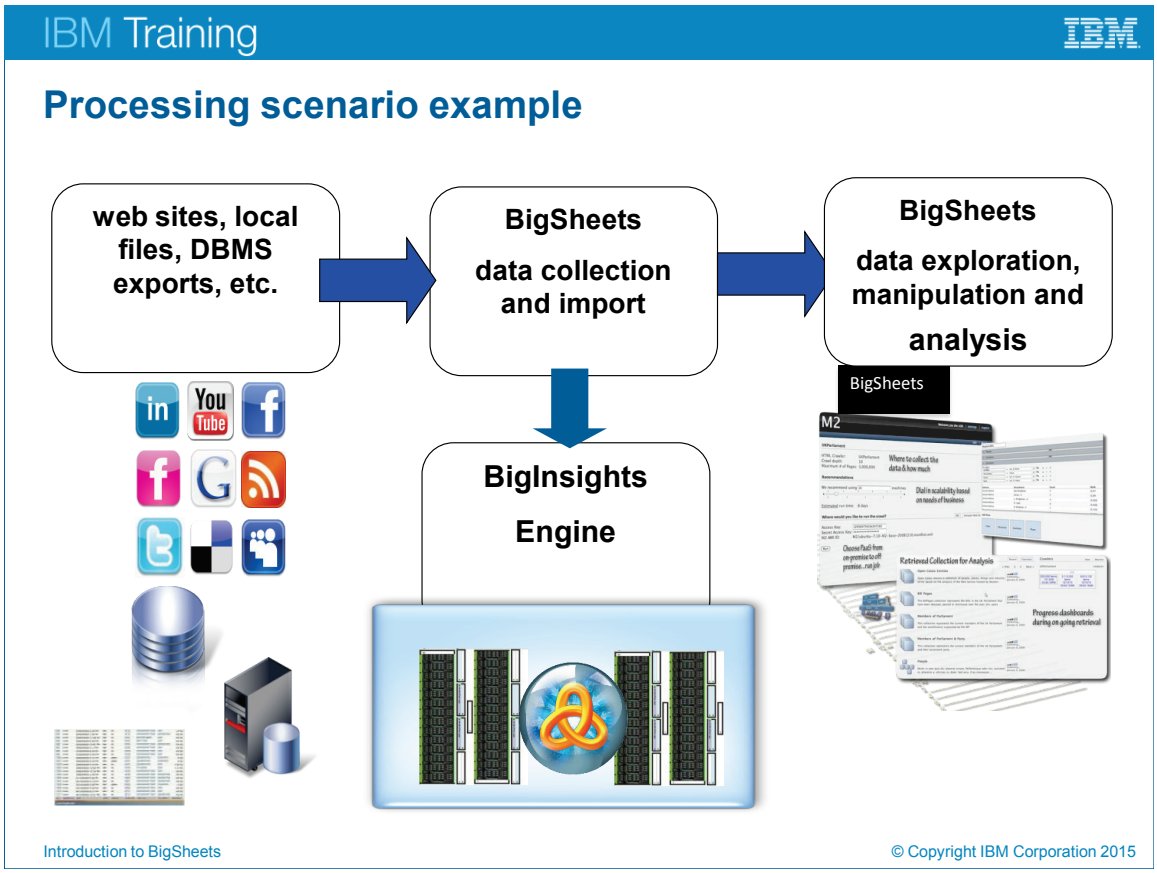
It presents a user interface developed specifically for business intelligence and non-technical business users to facilitate data gathering and analysis.

It can consume various kinds of data, such as CSV files produced by relational DBMSs and other applications or Web crawler data produced by a built-in application provided with BigInsights.

It can combine data sources from different sources, potentially enabling users to identify trends, opportunities, and risks "hidden" in the data.

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

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.



Processing scenario example

Before digging into the technical details of BigSheets, here is a walk through of a sample scenario that illustrates one way in which BigSheets can be used.

Firms can import data from web sites, local file systems, and other sources into BigSheets by using a simple graphical interface. Under the covers, BigSheets stores the data in BigInsights. Firms can then explore and manipulate the data using the BigSheets' simple spreadsheet interface and, if desired, generate charts to visualize specific results.

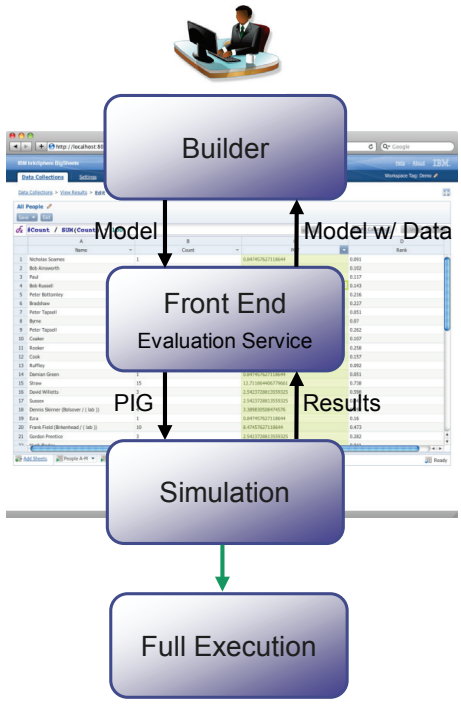
This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

BigSheets runtime processing

- User builds workbooks, filtering and transforming data as desired
- BigSheets evaluates and compiles front-end commands into executable work
- BigSheets executes work on a simulated environment of sample data
- User runs the workbook to compute results on the real data and explores the output



Introduction to BigSheets

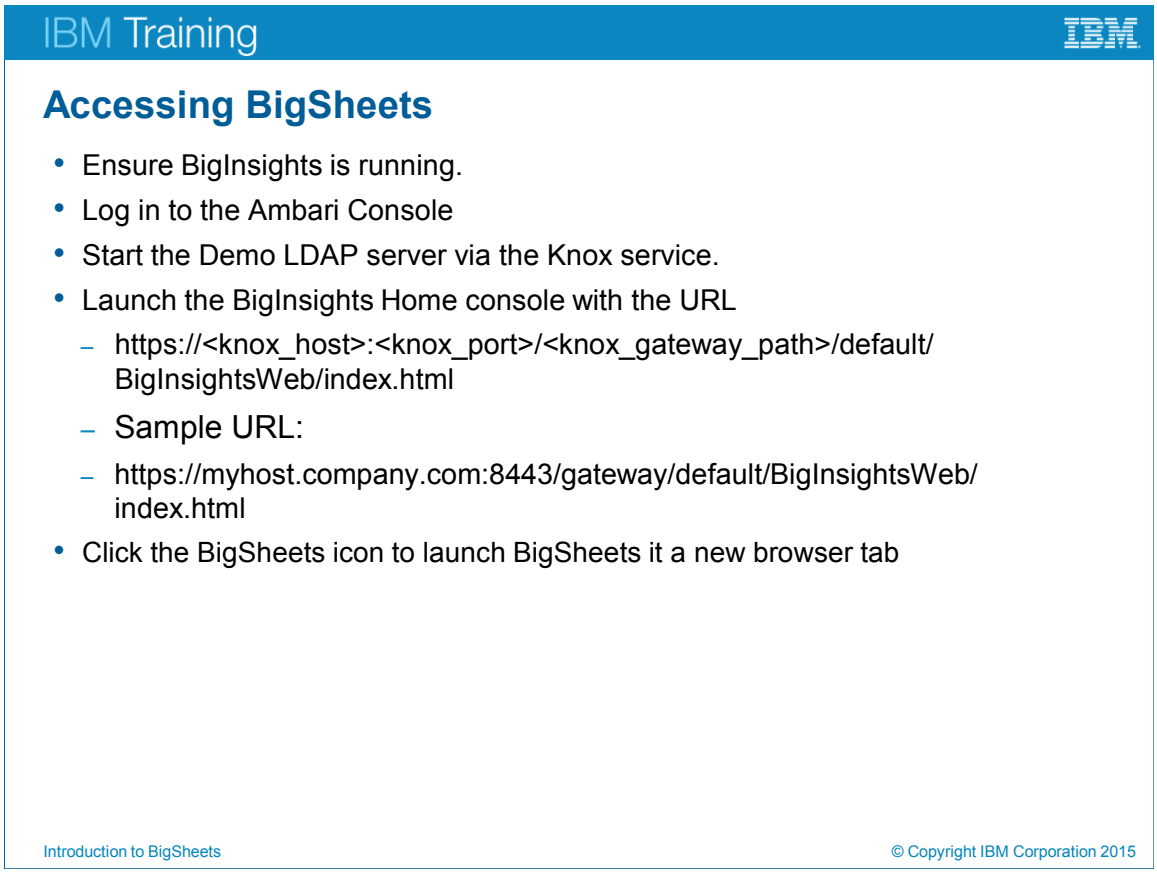
© Copyright IBM Corporation 2015

BigSheets runtime processing

The BigSheets graphical interface enables users to create workbooks (tabular data models) and filter or transform the data as desired. Under the covers, BigSheets generates PIG scripts as needed and executes the necessary work on a simulated environment consisting of sample data. This allows the user to iteratively explore various possibilities in an efficient manner. When satisfied, the user can "run" the workbook, which directs BigSheets to execute MapReduce jobs over the full set of data.

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

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.



Accessing BigSheets

To begin, make sure the BigInsights has been started. Log in to the Ambari console to start up the Demo LDAP server. This is required to start up the BigInsights Home where you will access BigSheets. The knox_host would be where BigInsights was installed. The default port is 8443. A sample url is provided. Once on the BigInsights Home page, click on BigSheets to launch it.

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

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Working with BigSheets

- Create a workbook to model target data
  - Directly from an application
  - From the BigInsights Console's BigSheets tab
  - From the BigInsights Console's Files tab
- Customize a workbook using the graphical editor and built-in functions
  - Filter data
  - Manipulate data
  - Join data from multiple workbooks
- "Run" the workbook
  - Applies work to full data set
- Explore results in spreadsheet format and/or create diagrams
- Optionally, export your data

Introduction to BigSheets

© Copyright IBM Corporation 2015

Working with BigSheets

Here are some typical steps that people take when working with the tool. Some of these step have been discussed.

Workbooks can be create from existing data residing in Hadoop or as a result of the output of a job running in Hadoop.

Additional workbooks can be created by applying functions to the data in an existing workbook.

A subset of the data is used during the workbook development process. This allows the analyst to view results in a timely manner. Later, the functions applied to a workbook can be "run" against the entire set of data.

Not only can the result be displayed in a spreadsheet format but they can also viewed in a chart format as well. The final results can be exported in order to be accessed by applications outside of BigSheets.

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

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Checkpoint

1. Where can you get BigSheets?
2. True or False? BigSheets executes work on a simulated environment of sample data.
3. List the four steps to work with BigSheets.

© Copyright IBM Corporation 2015

Checkpoint

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

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Checkpoint solutions

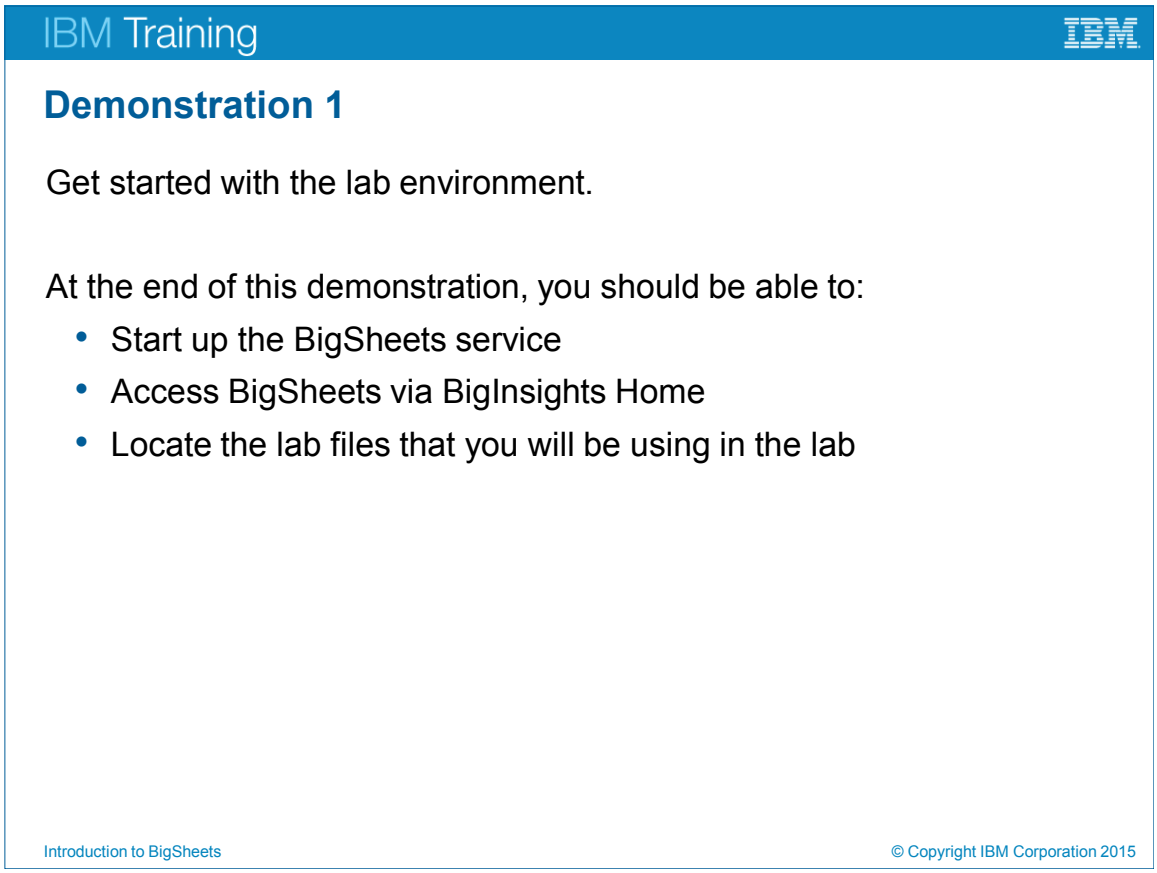
1. BigSheets is a component of **IBM BigInsights**
2. True, BigSheets first runs against a sample set of data. Once you are ready, you will need to run the work on the full data set.
3. Four steps to work with BigSheets:
   1. Create a workbook from the data
   2. Customize worksheet with formulas and built-in functions
   3. Run the workbook on the full data set
   4. Analyze results and / or create visualization diagrams

© Copyright IBM Corporation 2015

Checkpoint solutions

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

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.



Demonstration 1: Getting started with the lab environment

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

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

Demonstration 1: Getting started with the lab environment

Purpose:
You will start up the BigSheets service, access BigSheets via BigInsights home page and locate and upload lab files you will be using in the demonstration.

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

Task 1. Configure your image.

Important: Occasionally, when you suspend and resume the VM image, the network may assign a different IP address than the one you had configured. In these instances, the Ambari console and the services will not run. You will need to update /etc/hosts file with the newly assigned IP address to continue working with the image. No restart of the VM image is necessary - just give it a couple of minutes, at most. In some cases, you may need to restart the Ambari server, using *ambari-server restart* from the command line.

1. To open a new terminal, right-click the desktop, and then click **Open in Terminal**.
2. Type `ifconfig` to check for the current assigned IP address.
3. Take note of the IP address next to **inet**.
   Next, you need to edit the /etc/hosts file to map the hostname to the IP address.
4. To switch to root user, type `su -`.
5. When prompted for a password, type **dalvm3**.
6. To open the */etc/hosts* file, type **`gedit /etc/hosts`**.
7. Ensure that the contents of the file are similar to the following:
   ```
   10.0.0.118 ibmclass.localdomain ibmclass
   127.0.0.1 localhost.localdomain localhost
   ```
8. Update with the IP address of the first line from step 3.
9. Save and exit the file, and then close the terminal.

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

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

Task 2. Start the BigInsights components.

If all of the components have been started already, you may skip this task. Otherwise, follow the steps in this task to start up all of the components.

You will start up all the services via the Ambari console to ensure that everything is ready for the lab. You may stop what you don't need later, but for now, you will start everything.

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

   On the left side of the browser are the statuses of all the services. If it is currently yellow, wait for several minutes for it to become red before you start them up.
3. Once all the statuses are red, at the bottom of the left side, click **Actions** and then click **Start All** to start up the services.
4. In the **Confirmation** dialog, click **OK**.

   This will take a while to complete to complete.
5. When the services have started successfully, click **OK**.

Task 3. 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. To prepare for the next section of the lab, you will load in several sets of data into the HDFS.

1. To open a new terminal, right-click the desktop, and then click **Open in Terminal**.
2. To switch to the root user, type **`su -`**, and then type the password **`dalvm3`**.
3. Create the **biadmin** folder on the **hdfs** under **/user**:

   ```
   su - hdfs -c "hdfs dfs -mkdir -p /user/biadmin/"
   ```
4. Change the ownership of the folder to **biadmin**:

   ```
   su - hdfs -c "hdfs dfs -chown -R biadmin /user/biadmin"
   ```
5. To log out of the root user, type **`exit.`**
6. Do a listing of the **/user** directory to see that the **biadmin** directory has been created.

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

   On your local system, in the /home/biadmin folder, is a *labfiles* directory. In this directory are some of the data files that you will be using throughout this exercise.

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

© Copyright IBM Corp. 2012, 2015

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

7. Navigate to **/home/biadmin/labfiles/bigsheets** to see the files.
8. Upload these files to **/user/biadmin/** on the HDFS

```
hdfs dfs -put
/home/biadmin/labfiles/bigsheets/employee_state.csv
/user/biadmin
hdfs dfs -put /home/biadmin/labfiles/bigsheets/product.csv
/user/biadmin
hdfs dfs -put /home/biadmin/labfiles/bigsheets/sales.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
hdfs dfs -put
/home/biadmin/labfiles/bigsheets/last_of_the_mohicans.txt
/user/biadmin
hdfs dfs -put
/home/biadmin/labfiles/bigsheets/RDBMS_data.csv
/user/biadmin
```

   You will work with these files in subsequent lab exercises. You just made sure everything was uploaded to the HDFS for use later.
9. Back in **Firefox**, and then if necessary, navigate to the **Ambari** login page, **http://ibmclass.localdomain:8080**.
10. Log in to the **Ambari** console as **admin/admin**.
11. Ensure that the **BigInsights - BigSheets** component is started.

    The BigInsights Home requires the LDAP server to be started.
12. Click the **Knox** component.
13. Under the **Service Actions** dropdown on the upper right, select **Start Demo LDAP**, and then click **OK** to close the confirmation window.
14. 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

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

© Copyright IBM Corp. 2012, 2015

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

If promoted 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:

15. Click **BigSheets**, to launch BigSheets.
16. You will continue with this in the next lab exercise.

Task 4. Troubleshooting (Optional).

1. On the BigInsights Home, if BigSheets shows as unavailable, restart the BigSheets service via Ambar.

Results:
You started up the BigSheets service, accessed BigSheets via BigInsights home page and located and uploaded lab files.

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

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Unit summary

- Identify business and technical challenges in dealing with big data
- Describe how BigSheets can help with the business and technical challenges of big data
- Access BigSheets from the BigInsights Home console

© Copyright IBM Corporation 2015

Unit summary

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

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Making data available to BigSheets

IBM BigInsights v4.0

© Copyright IBM Corporation 2015
Course materials may not be reproduced in whole or in part without the written permission of IBM.

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

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

© Copyright IBM Corp. 2012, 2015

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Unit objectives

- Define the term workbook when used in conjunction with BigSheets
- List the readers supplied with BigSheets
- Describe the different ways to get data into a master workbook
- Explain the concept of running a workbook
- Import and export a workbook

© Copyright IBM Corporation 2015

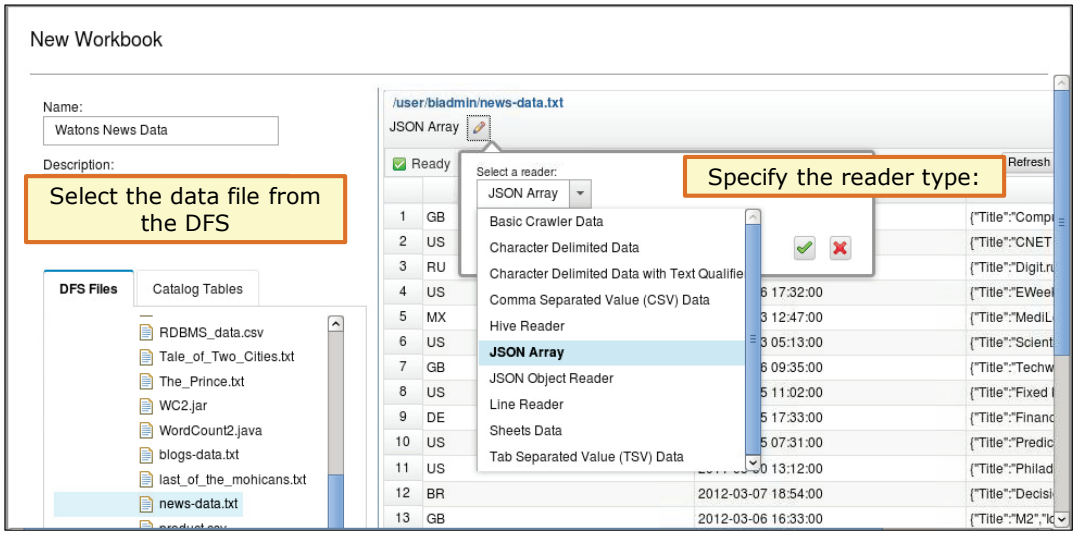
Unit objectives

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

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Create a master workbook from HDFS



Making data available to BigSheets

© Copyright IBM Corporation 2015

Create a master workbook from HDFS

For data to be accessible by BigSheets, it must be in a workbook. Create the workbook by clicking the New Workbook button. Then select the data from the DFS. By default the data in that file is displayed in a text format. Select the appropriate reader type to have the data placed in that format. At this point, the data can be saved into a workbook.

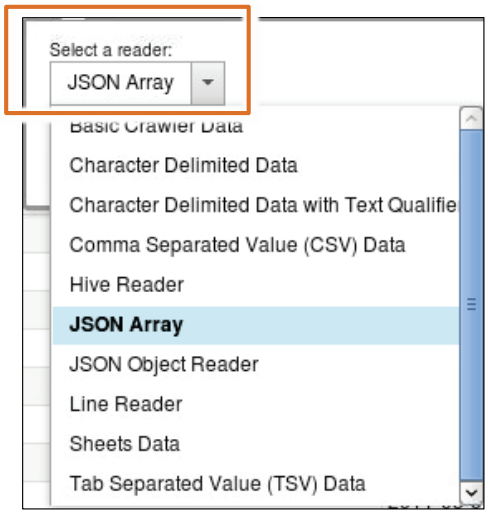
This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

BigSheets readers

- Select the reader based upon the data's format
  - Basic Crawler Data
  - Character Delimited Data
  - Comma Separated Value (CSV)
    - Can allow for a header record
  - Hive Reader
  - JavaScript Object Notation (JSON)
    - Both array and object reader
  - Tab Separated Value
    - Can allow for a header record
  - Sheets Data
  - Line (default)
- User-written plug-ins are supported



Making data available to BigSheets

© Copyright IBM Corporation 2015

BigSheets readers

In order for the data in a workbook to be parsed properly, the correct reader must be used. IBM supplies a number of different readers and user-written plug-ins are also supported. The supplied readers are:

Comma Separated Value - This also allows for a header record to be part of the file

Tab Separated Value - This also allows for a header record to be part of the file

Character Separated Value - This also allows for a header record to be part of the file. You must also specify the character that is used as the separator

JSON Array or Object Reader – This reader reads a JSON file that contains a JSON array of JSON objects.

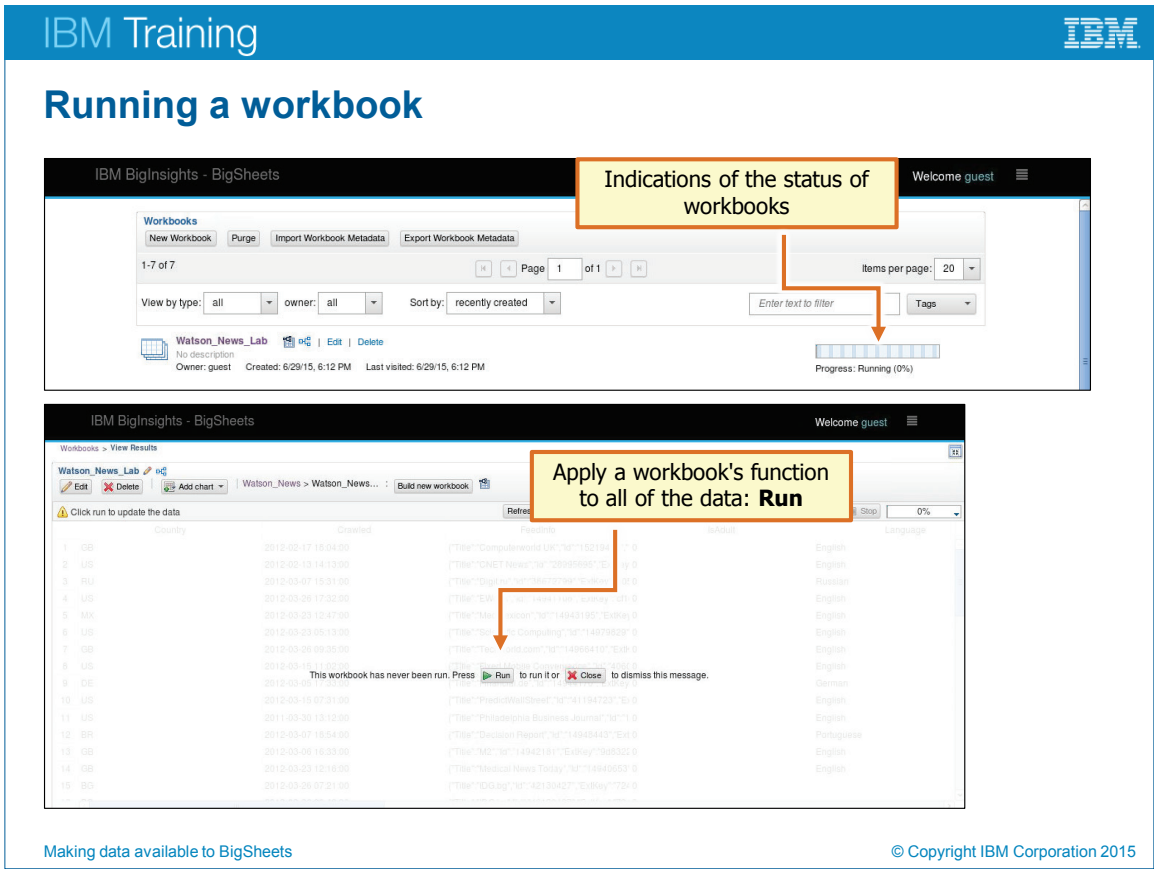
Basic Crawler - This is used to view the results from a web crawler

Hive – reads from the Hive default field separator output file

Line - This is the default reader

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

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.



Running a workbook

As stated before, BigSheets is designed to work on a subset of the data during the creation phase of a workbook. After the workbook has been test and is known to run properly on the subset of data, a *run* function can be executed so that the processes in the workbook are applied to the entire data set.

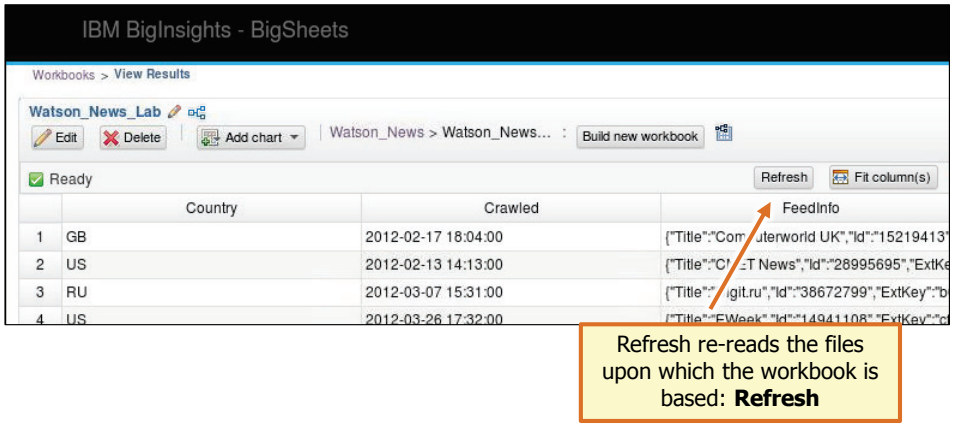
When viewing the list of workbooks, the status of each workbook is displayed in the form of a progress indicator which gives a visual indication as to whether a workbook needs to be run on the entire data set.

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

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Updating the data in a workbook



Making data available to BigSheets

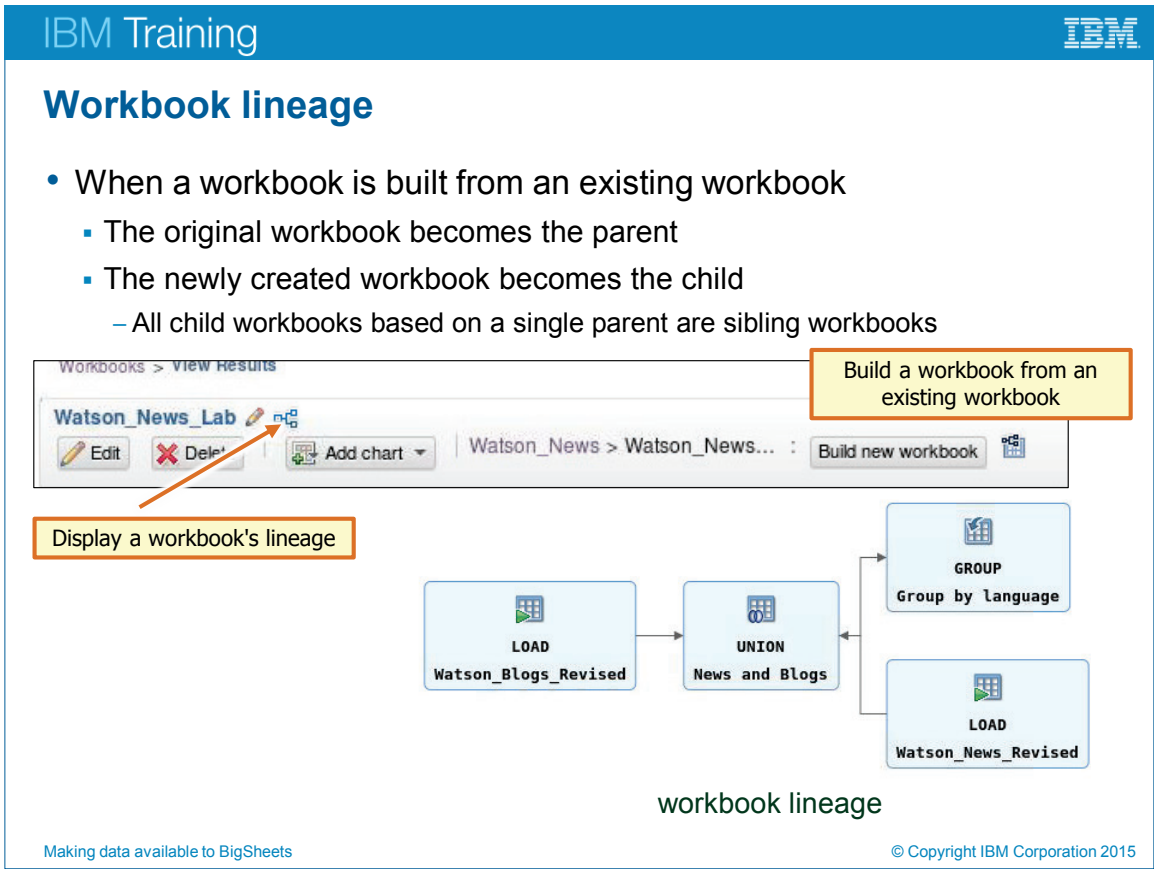
© Copyright IBM Corporation 2015

Updating the data in a workbook

Once data is imported into a workbook, that data is static. Changing the data in the file(s) upon which the workbook is based does not automatically change the data in the workbook. To update the workbook, open the workbook and click the *Refresh* pushbutton. This imports the updated data into the workbook.

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

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.



Workbook lineage

Additional workbooks can be built based upon existing workbooks. When a workbook is built from an existing workbook, a lineage is created. The original workbook becomes the parent and the newly created workbook becomes a child. All workbooks that have the same parent workbook are sibling workbooks.

Why have workbooks that are based on other workbooks? This allows for implementing incremental analysis. Small manipulations of the data are made so that the effect of each of those manipulations can be easily observed. This simplifies the analytics but also can create a documentation nightmare.

Assume that you have created ten workbooks that, in a step by step fashion, culminate in a workbook in which the data can be easily analyzed. Later someone looks at your final workbook and wonders how you manipulated the data to get the results. The lineage capability of BigSheets is a self-documenting mechanism that gives a person a way of seeing which workbooks were used in the creation of the final workbook and the order in which they were applied.

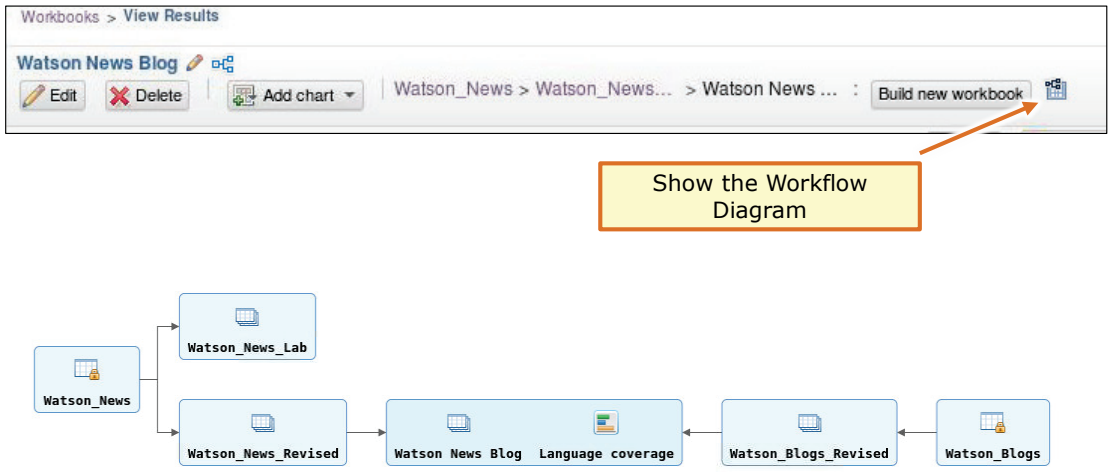
This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Workflow Diagram

- Shows the relationship between all the workbooks related to the current workbook



Making data available to BigSheets

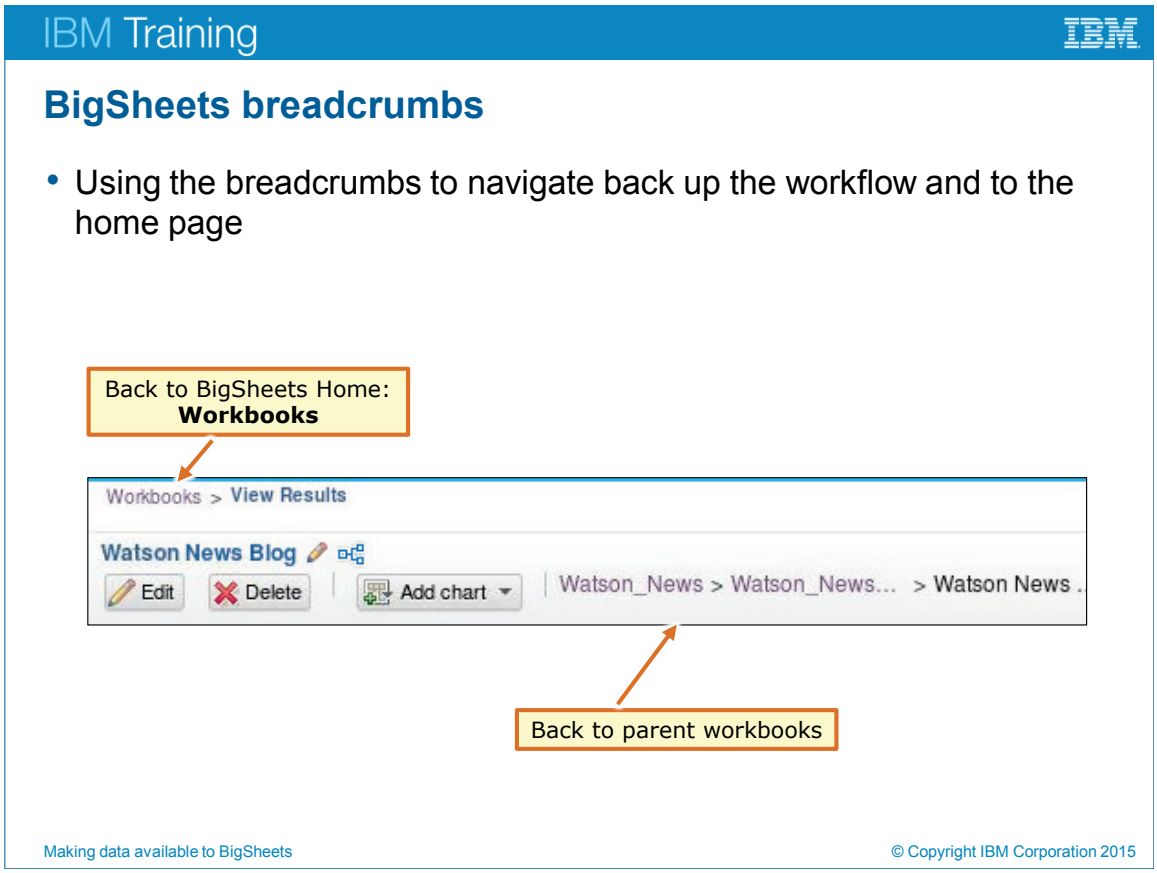
© Copyright IBM Corporation 2015

Workflow Diagram

The Workflow Diagram shows the relationship between the current workbook and all the workbooks that are related.

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

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.



BigSheets breadcrumbs

Use the breadcrumb links to go back to parent workbooks or to the BigSheets Home where you can create new workbooks or perform other workbook options, such as importing and exporting workbooks.

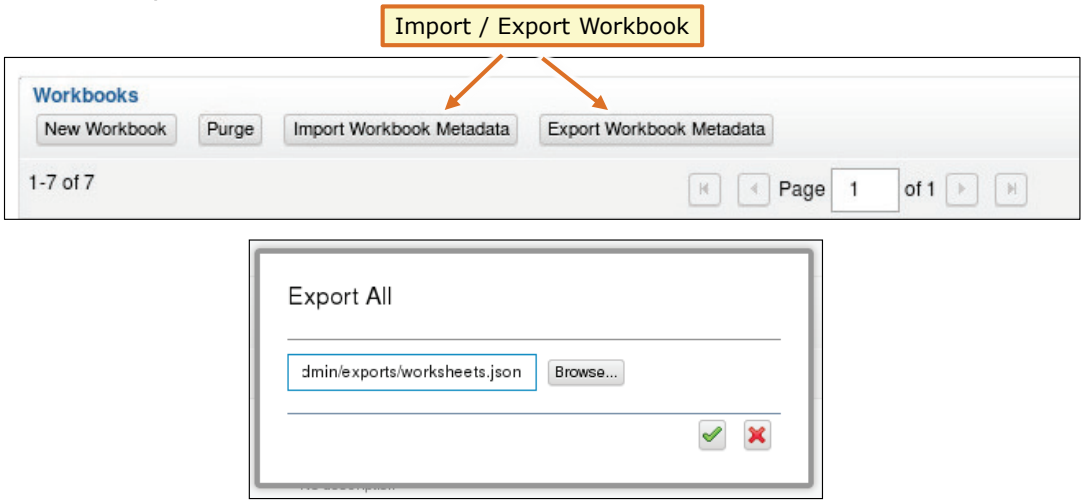
This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Import and export a workbook

- Import or export a workbook between clusters
  - Development cluster <> Test Cluster <> Production cluster



Making data available to BigSheets

© Copyright IBM Corporation 2015

Import and export a workbook

You can export and import workbooks between different clusters of your environment. Click on the Workbooks breadcrumb link and then select the Export or Import pushbuttons to perform the desired actions. Then select the appropriate worksheets or click the Export All pushbuttons.

To import a Workbook, if the data of the source exist, then the workbook's data will be generated. Otherwise, you will need to specify the source of where the data resides within the HDFS.

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

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Checkpoint

1. How do you create a workbook?
2. What is the purpose of BigSheets readers?
3. True or False? Data in a worksheet is automatically updated when the source data changes.
4. What allows you to find out if a workbook has a parent?

© Copyright IBM Corporation 2015

Checkpoint

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

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Checkpoint solutions

1. Click the New Workbook button and specify the data source from HDFS.
2. BigSheets readers allow a way to parse the data as it is loaded. CSV, TSV, JSON, Character-delimited, Hive, Line (default).
3. False. The data in the worksheet is static, you need to manually refresh the data. Open the workbook and click the Refresh pushbutton.
4. The **workbook lineage** or the **workflow diagram** shows a visual representation of the relationship between the workbooks.

© Copyright IBM Corporation 2015

Checkpoint solutions

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

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Demonstration 1

Import data into a workbook.

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

- Load data into various master workbooks

© Copyright IBM Corporation 2015

Demonstration 1: Importing data into a workbook

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

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

Demonstration 1: Importing data into a workbook

Purpose:
You will load data into various BigSheets master workbooks.

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

Task 1. Creating a workbook from a web crawler application.

In this task, you will be creating a workbook using the results of a Web Crawler application. The crawler was directed to extract information from a website that dealt with patents. Essentially the web crawler looked at a site that had a list of names. Each name is a hyperlink to a page that lists the patents for that person.

1. Launch **Firefox**, or open up a new tab in the existing browser and go to the following website:
   http://www.ibm.com/software/ebusiness/jstart/bigsheets/demo/Patents.html
   There you see the list of names.
2. Click on any name and it takes you to a page that lists all of the patents registered to that individual. This is to give you a frame of reference when doing this exercise.
3. You can close that tab.
4. Open a new terminal or use an existing one. To open, right-click anywhere on the desktop and click **Open in Terminal**.
5. Upload the results of the web crawler run into the HDFS.

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

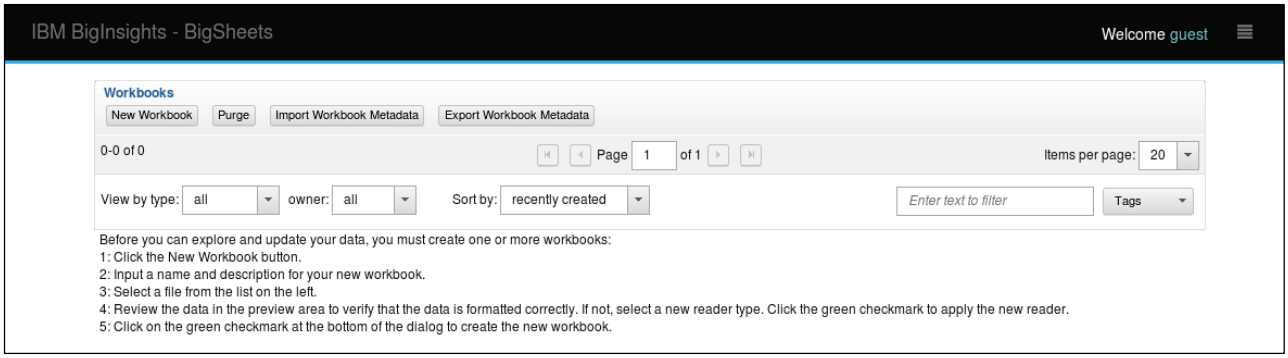
This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2012, 2015

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

6. Go to the **BigInsights Home** page in the Firefox browser and click on the **BigSheets** link (last step from the previous demonstration) to load the BigSheets landing page.

   You should be here now:



7. Click the **New Workbook** to bring up the dialog.
8. For the **Name** field, type **PatentCrawler**.
9. For the **Description** field, type **Results from Web Crawler**.
10. On the **DFS Files** tab, drill down to **/user/biadmin/** and then select **PatentCrawler_data.csv**.

    On the Preview pane, you will see that the default Line Reader was used. The file format is CSV, so you will select the CSV reader.

11. Click **Edit Workbook reader**  and then select the **Comma Separated Value (CSV) data.**

12. **Uncheck** the "Headers Included?" checkbox and then click **Apply settings**  to change the reader type.

13. Click **Save workbook**  to create the *PatentCrawler* workbook.
14. You will work with this workbook in a later lab exercise. For now, you are done with this task.

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

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

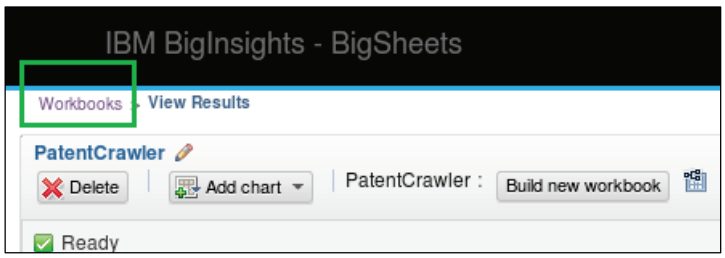
Task 2. Creating a workbook from the results of a WordCount application.

In this task, you will be creating a workbook from the results of a WordCount application that was ran against the file, last_of_the_mohicans.txt, which is located on your local system under */home/biadmin/labfiles/bigsheets.* The output file from the WordCount run is also located under the same directory -- *part-r-00000*. You will load this file to the HDFS in order to create a workbook from it.

1. Use an existing terminal, or open a new one.
2. Upload part-r-00000 to the HDFS.

   ```
   hdfs dfs -put /home/biadmin/labfiles/bigsheets/
   part-r-00000 /user/biadmin
   ```

3. Go back to the **BigSheets Home** page by clicking on the **Workbooks** breadcrumb.



4. Click the **New Workbook** to bring up the dialog.
5. In the *Name* field, type **Wordcount**.
6. In the *Description* field, type **Results from WordCount application**.
7. On the **DFS Files** tab, drill down to **/user/biadmin/** and select **part-r-00000**.

   On the Preview pane, you will see that the default Line Reader was used. This is sufficient for this file.

8. Click **Save workbook**  to create the Wordcount workbook.

   You will work with this workbook in a later lab exercise. For now, you are done with this task.

Task 3. Creating additional master workbooks.

In this task, you will create workbooks based off of the files you had uploaded to the HDFS in the first lab exercise. You will create three notebooks from the CSV files.

1. Go to the **BigSheets** Home page, and then click the **Workbooks** breadcrumb.
2. Click **New Workbook**.

   The New Workbook window appears.

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

© Copyright IBM Corp. 2012, 2015

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

3. In the **Name** field, type **Employees**.
4. In the **Description** field, type **Employees state data**.
5. On the **DFS Files** tab, navigate to your user directory (/user/biadmin/) and select the **employee_state.csv** file.

   The contents of the employee_state.csv are displayed in the right frame in a text format.

   The data has been read by the BigSheets line reader. Unfortunately, the data is in a comma-separated format, so you need to specify a different BigSheets reader.

6. Click **Edit workbook reader**  .
7. Select **Comma-Separated-Value (CSV) Data**, ensure the **Headers included?** checkbox is selected and then click **Set reader**  .
8. Create workbooks for the **product.csv** and **sales.csv** files.
9. Click **Save workbook**.

   Once you have the three workbooks created, you are done with this task. You will get to work with these workbooks in a later exercise.

Results:
You loaded data into various BigSheets master workbooks.

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

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Unit summary

- Define the term workbook when used in conjunction with BigSheets
- List the readers supplied with BigSheets
- Describe the different ways to get data into a master workbook
- Explain the concept of running a workbook
- Import and export a workbook

© Copyright IBM Corporation 2015

Unit summary

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

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

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

© Copyright IBM Corp. 2012, 2015

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Navigating a workbook

IBM BigInsights v4.0

© Copyright IBM Corporation 2015
Course materials may not be reproduced in whole or in part without the written permission of IBM.

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

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

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Unit objectives

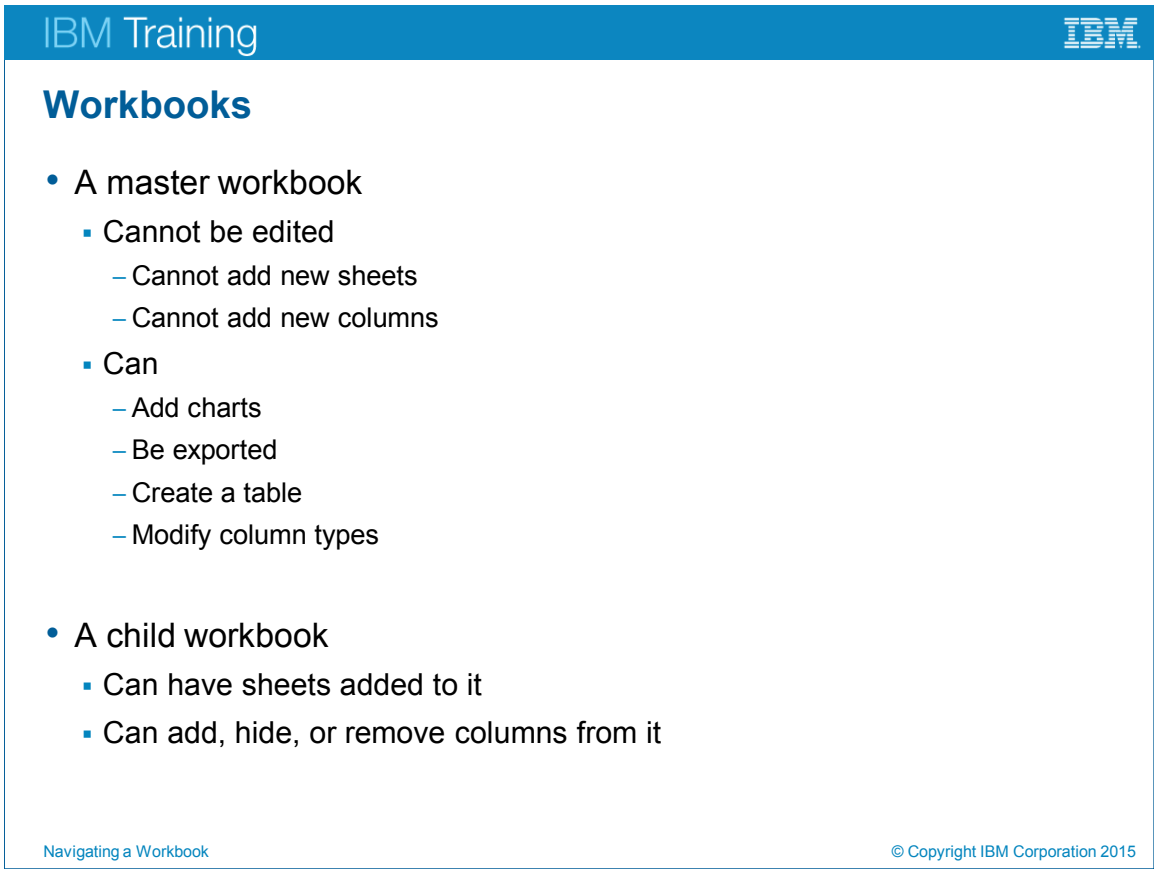
- Create a workbook from a master workbook
- List the different types of sheets that can be added to a workbook
- Describe what each type of sheet can do in a workbook
- Add a sheet to a workbook

© Copyright IBM Corporation 2015

Unit objectives

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

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.



Workbooks

When you create a master workbook by importing data from the Hadoop file system, that workbook is immutable. Columns cannot be modified and no new sheets can be added. Master workbook can add charts, have its metadata exported, create a table, and have its column types modified.

Child workbooks, created from an existing workbook can be changed. Sheets can be added to the new workbook and columns in any of the sheets can be modified. A child workbook takes on the characteristics of its parents. This means that any sheets defined in the parent are applied before the data is passed to a child workbook.

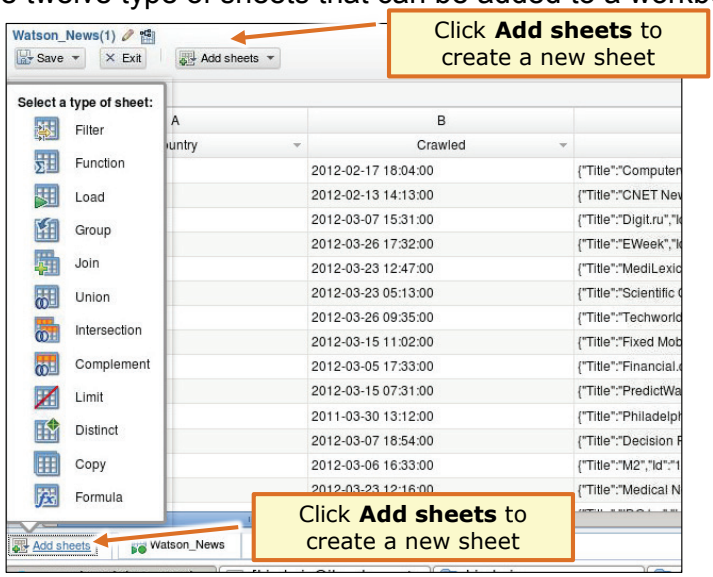
This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Adding sheets to a workbook

- There are twelve type of sheets that can be added to a workbook



Navigating a Workbook © Copyright IBM Corporation 2015

Adding sheets to a workbook

Normally, extracting data from unstructured text is a multi-step process. This is due to the complexity of the process. The best way to approach this task is to use "baby steps". It is for this reason that BigSheets gives you the ability to create child workbooks and each child workbook has access to twelve types of sheets. Each type of sheet is used for a specific purpose but taken together, they allow you to create a complex data extraction tool that does not require any programming on your part.

At both the top and the bottom of the BigSheets interface there is a mechanism that gives you access to the twelve types of sheets.

You create a sheet to apply some sort of process to the data. For example, one type of sheet allows you to filter data. A second allows you to join two sheets together. The order in which the sheets are defined in a workbook determine the order that these processes are applied to the data.

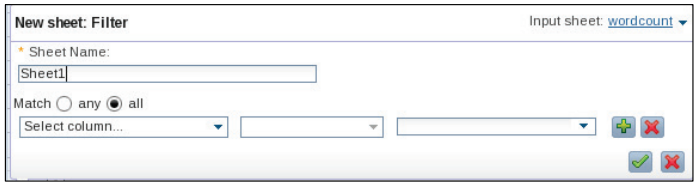
This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

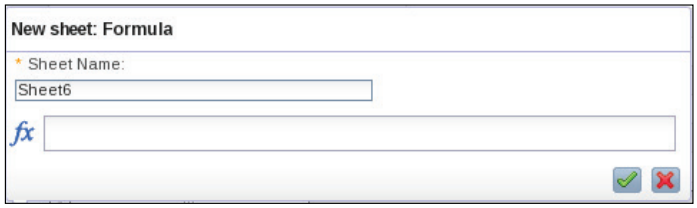
Workbook sheets (1 of 9)

- Filter
  - Removes data that does not match a specified criteria



Can specify multiple filtering criteria

- Formula
  - Allows complex formulae to be specified



Specify some formula

Navigating a Workbook

© Copyright IBM Corporation 2015

Workbook sheets

The *Filter* sheet allows you to code one or more conditions that causes data to be removed.

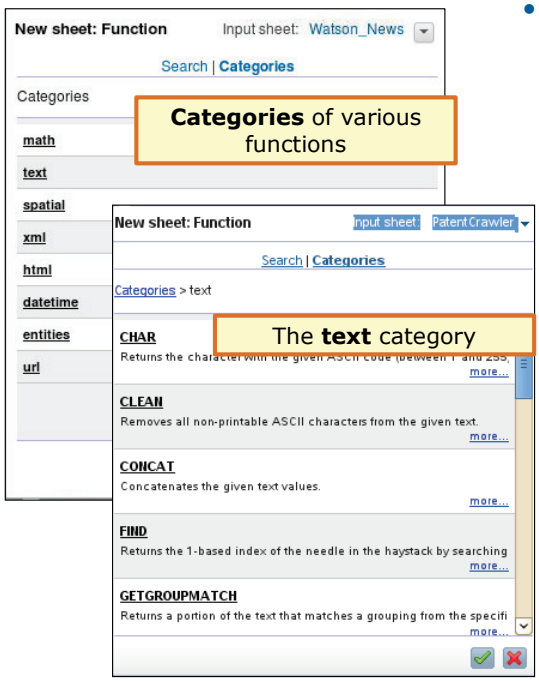
The *Formula* sheet gives you an entry field where you can code a specific formula.

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

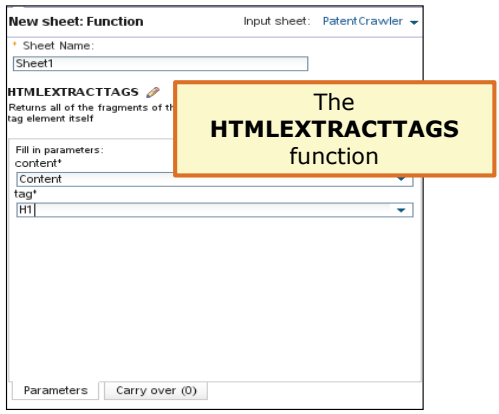
© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Workbook sheets (2 of 9)



- Function
  - Executes a function against each row in a sheet.
  - It is a function that takes rows of data as input and produces one or more rows of data as output



Navigating a Workbook

© Copyright IBM Corporation 2015

The *Function* sheet executes the specified function against each row in the sheet. Selecting the *Categories* hyperlink, presents a list of available function types. Selecting a function type presents a list of functions. Based upon the chosen function, you enter parameter values using the provided graphical interface.

There are a number of functions provided by BigInsights but you also have the option of defining additional functions.

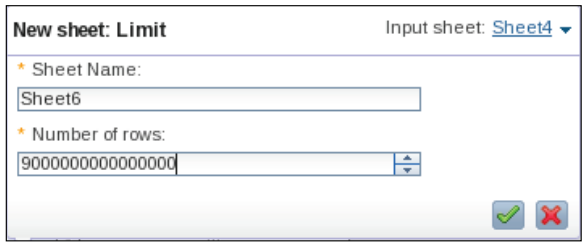
This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

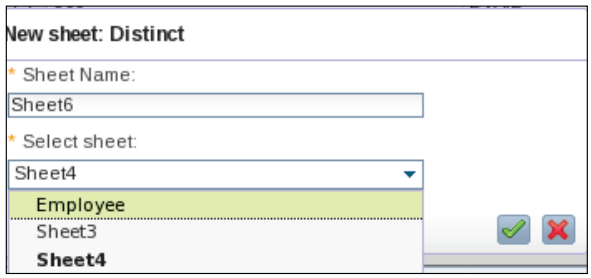
IBM Training

Workbook sheets (3 of 9)

- Limit
  - Limits the number of rows in a sheet



- Distinct
  - Eliminates duplicate rows in a sheet



Navigating a Workbook

© Copyright IBM Corporation 2015

The *Limit* sheet limits the number of rows in a sheet.

The *Distinct* sheet eliminates duplicate rows in a sheet.

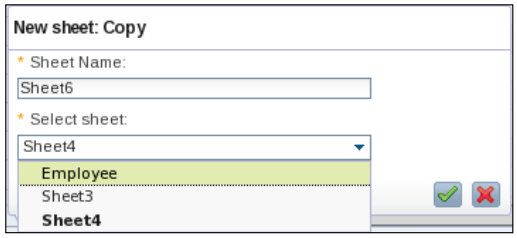
This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

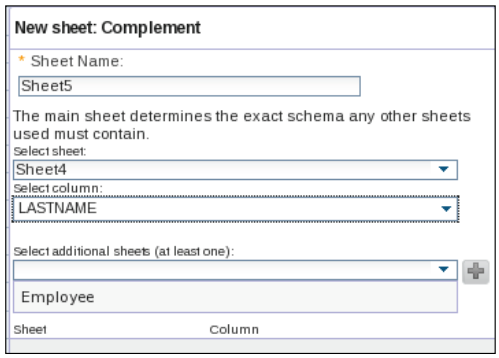
IBM Training

Workbook sheets (4 of 9)

- Copy
  - Copies a sheet, including the data and all the formulae used to create the data



- Complement
  - Creates a sheet from rows of two or more sheets where the first sheet contains values in the specified column that are not in the other sheets



Navigating a Workbook

© Copyright IBM Corporation 2015

The *Copy* sheet makes a copy of another sheet. This includes both the data and all formulas. You might want to do this because you want to try different techniques when working with a particular set of data.

The *Complement* sheet takes two or more sheets as input. A common column is chosen for all sheets. All rows in the first sheet that do not have the same values in the chosen column in the other sheets are retained. All sheets must have the exact same schema.

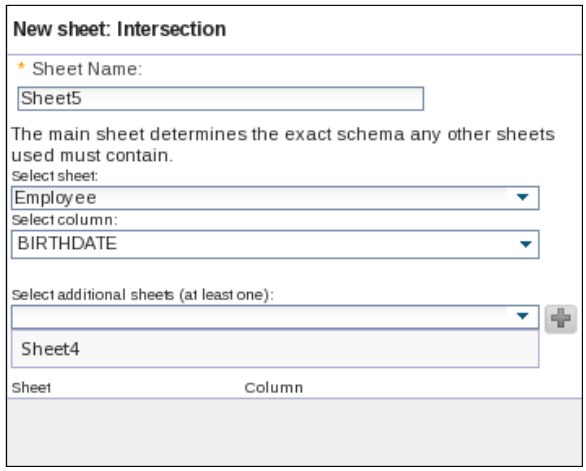
This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Workbook sheets (5 of 9)

- Intersection
  - Creates a sheet from two or more sheets consisting of rows where all sheets have the same values in the specified column



Navigating a Workbook

© Copyright IBM Corporation 2015

The *Intersection* sheet creates a new sheet from two or more existing sheets that contains those rows that have a common value in the chosen column. All sheets must have the same schema.

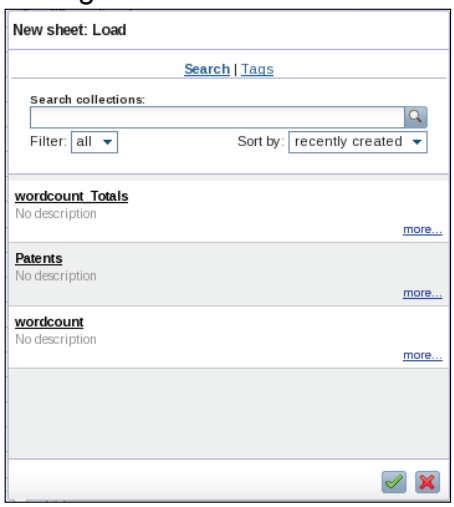
This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

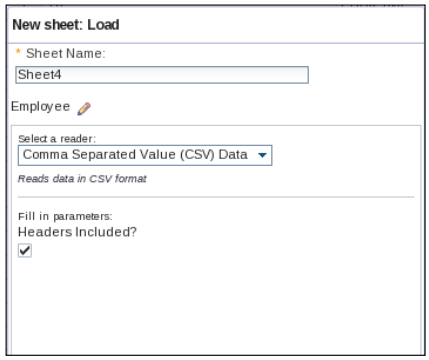
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Workbook sheets (6 of 9)

- Load
  - Brings in the data of another workbook as a sheet.
  - Useful in performing a combine, complement, intersection, or union operation using other sheets





Navigating a Workbook

© Copyright IBM Corporation 2015

The *Load* sheet brings in data from another workbook. It is useful in performing a combine, complement, intersection, or union operation using other sheets. The data, when imported into to the original workbook, required a particular reader in order for the data to be properly parsed. When the *Load* sheet is used, it follows somewhat the same process as importing the data. A reader needs to be specified to parse the data properly.

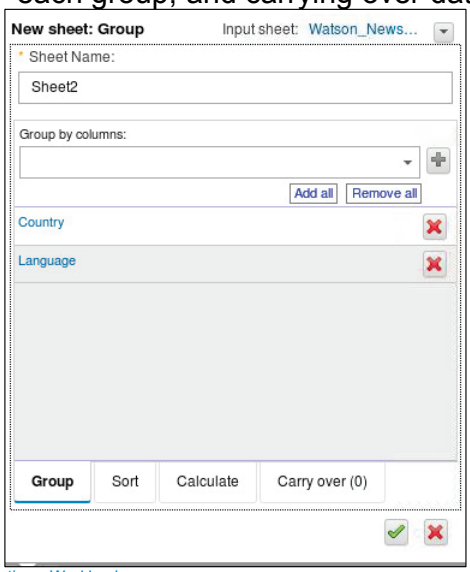
This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

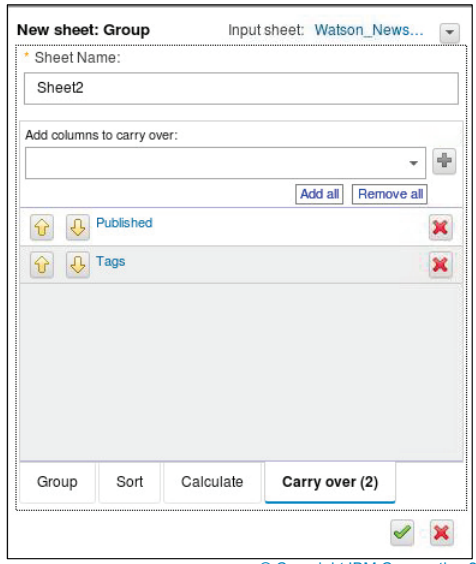
© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Workbook sheets (7 of 9)

- Group
  - Calculates values by grouping data in the workbook, applying functions to each group, and carrying over data





Navigating a Workbook

© Copyright IBM Corporation 2015

The *Group* sheet calculates values by grouping data in the workbook, possibly applying a function to each group, and specifying any columns that are to be carried over from the input sheet.

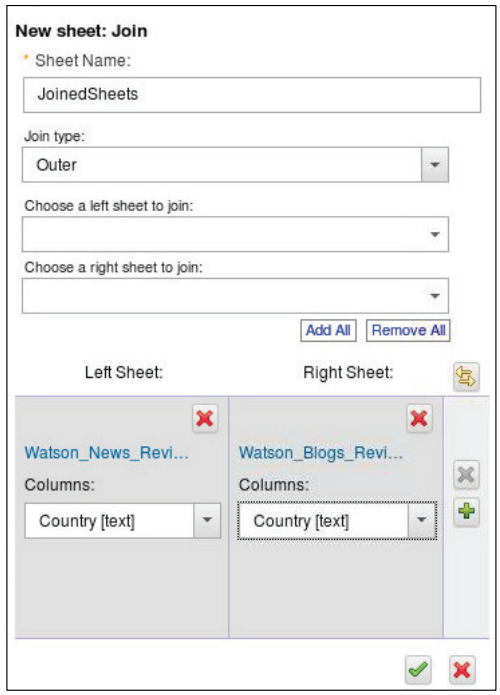
This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Workbook sheets (8 of 9)

- Join
  - Joins two or more sheets using a common column in all of the sheets



© Copyright IBM Corporation 2015

The *Combine* sheet joins two or more sheets using a common column in all of the sheets.

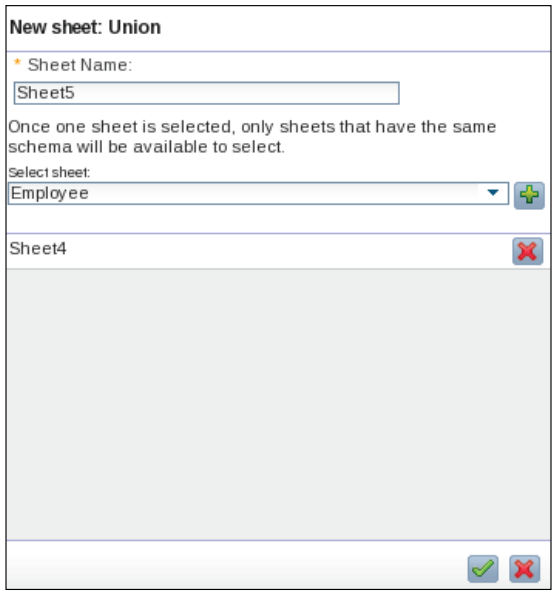
This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Workbook sheets (9 of 9)

- Union
  - Appends sheets



Navigating a Workbook

© Copyright IBM Corporation 2015

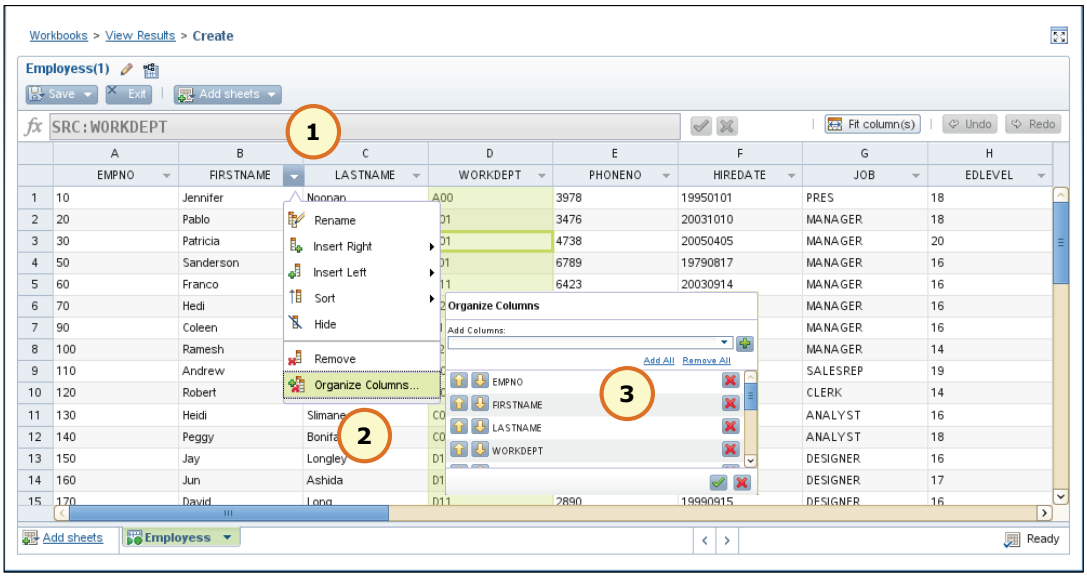
The *Union* sheet appends sheets into a single sheet. All sheets being unioned must have the same schema.

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

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Modifying a column



Navigating a Workbook

© Copyright IBM Corporation 2015

Modifying a column

Within a child workbook, columns can be added, removed, sorted, and hidden. In order to populate a newly inserted column, you specify a function. You can also change the name of a workbook by clicking on the pencil icon next to the name of the workbook.

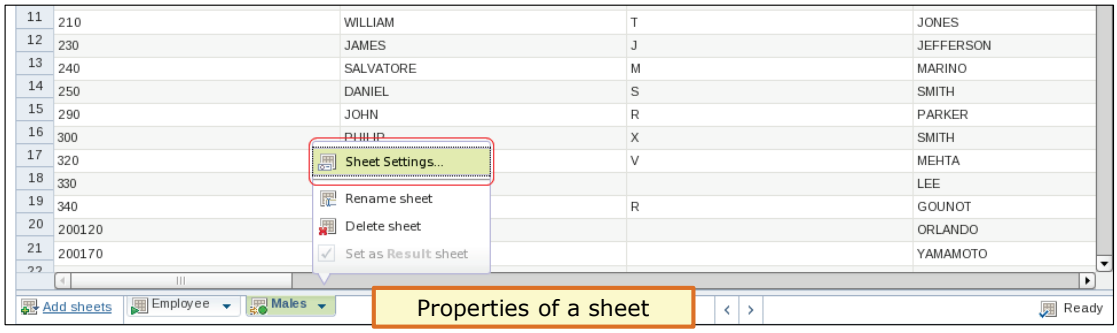
This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

The properties of a sheet

- In edit mode, click the sheet's tab, select *Sheet Settings*
  - Display the function used to create the sheet
  - Allows for that function to be modified
- Can rename the sheet
- Can delete the sheet



Navigating a Workbook

© Copyright IBM Corporation 2015

The properties of a sheet

Clicking on the triangle on a sheet's tab allows for modification of the sheet. The sheet can be renamed, deleted, or modified.

Selecting the *Sheet Settings* displays the function used to create that sheet and allows you to modify that function. For example, if the sheet is a filter, then selecting *Sheet Setting* displays the filtering criteria used to create the sheet and allows you to change that filtering criteria. If the function used to create the sheet was a load, then you would get the option to change the reader.

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

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Checkpoint

1. True or False? A master workbook can be edited to add new sheets or columns.
2. List at least three different sheets that you can create.
3. What information can you find when you view the properties of a sheet?

© Copyright IBM Corporation 2015

Checkpoint

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

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Checkpoint solutions

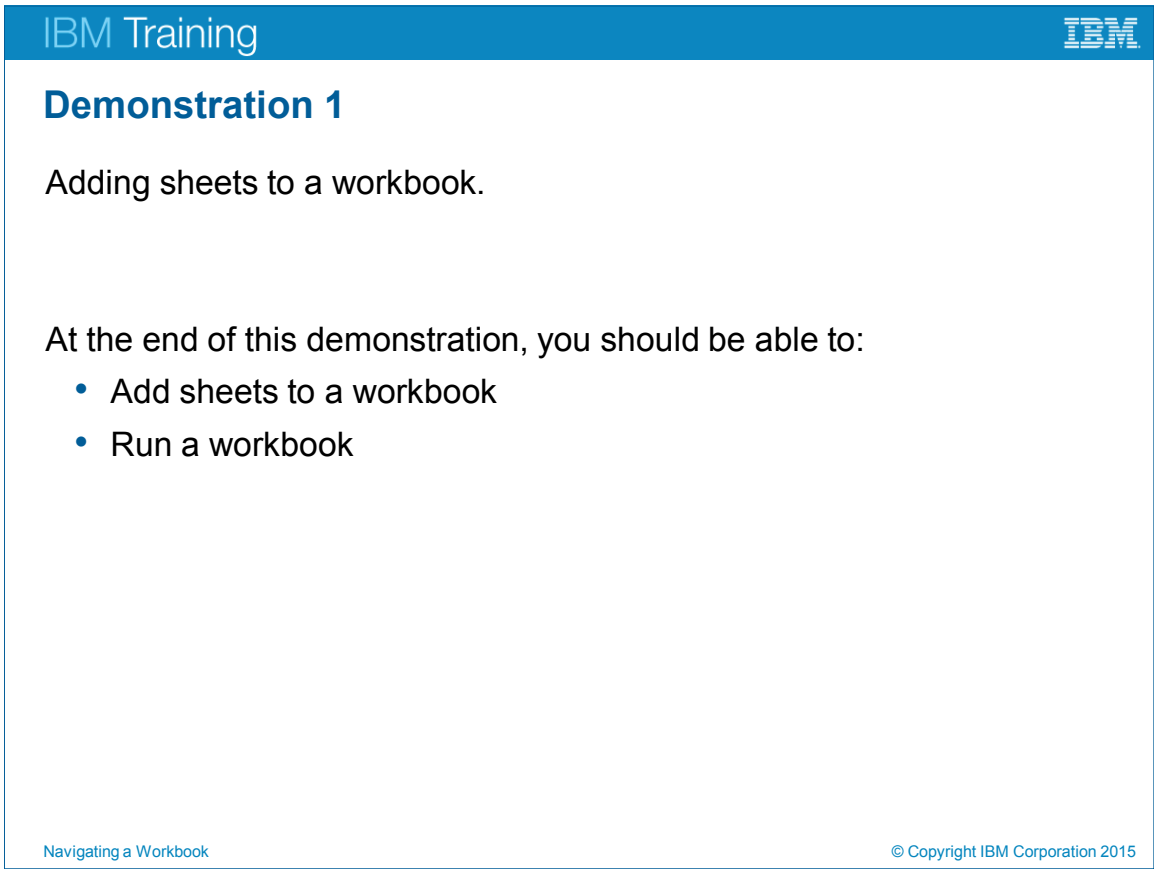
1. False. You cannot edit the master workbook to add any sheets or columns. You must first create a child workbook to do that.
2. The types of sheets available: filter, function, load, group, join, union, intersection, complement, limit, distinct, copy, formula
3. The properties of a sheet tell you formula or function that was used to created it. You can update that as well as delete or rename the sheet.

© Copyright IBM Corporation 2015

Checkpoint solutions

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

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.



Demonstration 1: Adding sheets to a workbook

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

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

Demonstration 1: Adding sheets to a workbook

Purpose:
You will create a new workbook based on the web crawler notebook and use sheets to extract data from web pages.

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

Task 1. Background.

In the previous demonstration you created a workbook based on the results of a web crawler. The crawler was directed to extract information from a website that dealt with patents. Essentially the web crawler looked at a site that had a list of names. Each name is a hyperlink to a page that lists the patents for that person.

1. If you wish, you can review the contents of the website that was crawled. Launch **Firefox**, or open up a new tab in the existing browser. If you wish to see it again, go to the following website:
   http://www.ibm.com/software/ebusiness/jstart/bigsheets/demo/Patents.html
2. There you see the list of names, Click on any name and it takes you to a page that lists all of the patents registered to that individual. This is to give you a frame of reference when doing this exercise.
3. You can close the newly opened tab.

Task 2. Create a Function sheet.

1. Go to the **BigSheets** Home page.
2. Open the **PatentCrawler** workbook.
   PatentCrawler is a master workbook. No modifications can be made to it. However, you can create a new workbook that is based on PatentCrawer and this new workbook can be modified.
3. Click **Build new workbook** 

   .
   Before you start working with sheets, you will rename each header into something more meaningful.
4. Rename the columns by clicking on the drop-down button next to the column header and then select **Rename**.

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

© Copyright IBM Corp. 2012, 2015

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

5. Name the columns as follows:
6. Header1 = URL
7. Header2 = Type
8. Header3 = Content
9. Click **Add sheets** 

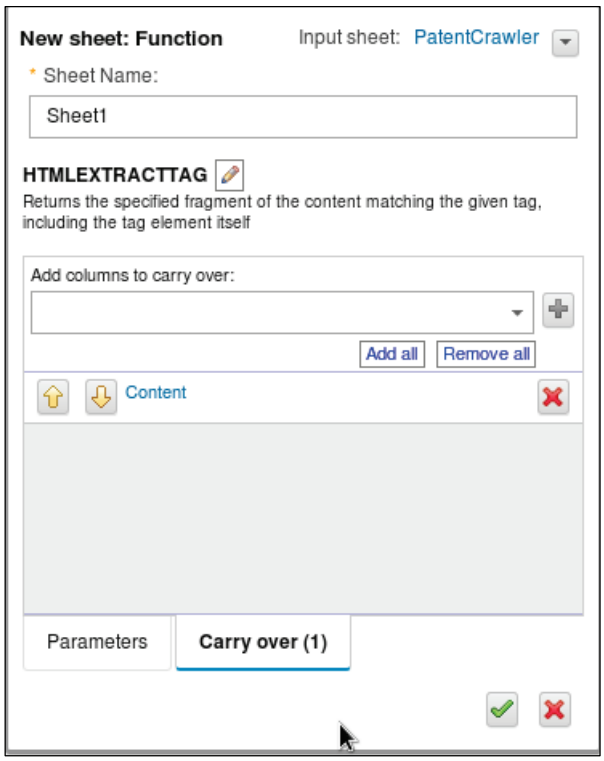
.
   You are to apply a function to some of the data.
10. Select **Function** from the list of sheet types.
    Since you are not aware of all of the functions at your disposal, it is best to get a list of them.
11. Click the **Categories** hyperlink.
    The ultimate goal is to get patent information for each person on the patents website. You want to apply some function that works with HTML tags.
12. Click the **html** hyperlink.
    For each individual's patent site, the person's name is associated with the HTML tag H1. The first thing that needs to be done is to get the content for each H1 tag.
13. Select the **HTMLEXTRACTTAG** hyperlink from the list of functions.
    You have the option of giving each new sheet a descriptive name, but you will use *Sheet1* as the name.
14. Click the **content** drop-down box.
    You are presented with a list of all columns in the sheet from which to make a selection.
15. Select **Content.**
    You want to get the names of the individuals. They are displayed using an HTML H1 tab.
16. In the *tag* field, type in **H1**.
17. In the **occurrence** field*,* enter a value of **1**.
18. At the bottom of the entry dialog (you may need to scroll down), click the **Carry over (0)** tab.
    You might not realize it, but you not only want the extracted information in this new sheet, but, for your purposes, you also want the Content column information as well.
19. From the **Add columns to carry over** drop-down box, select **Content.**

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

© Copyright IBM Corp. 2012, 2015

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

20. Click **Add column to carry over**  .
    This adds the Content column to the carry over list.



21. Click **Apply settings**.
    You now have a new sheet. But the heading name for column A needs some work.
22. Move the cursor to the **HTMLEXTRACTAG** column heading, click on the drop-down indicator and then select **Rename.**
23. Highlight the name and change it to **NameWithH1Tag.**
    The column name cannot have any spaces.
24. Click **Apply settings**.

Task 3. Further refine the collection.

You did not get the results that you really desired. There are HTML tags encapsulating the desired data. To keep progressing towards your goal, you must create another sheet. You want it based on *Sheet1.*

1. Ensure that the tab for *Sheet1* is selected and then click **Add sheets.**
2. You want to create another **Function** sheet.
3. Click **Categories** and then **html**.
   This time, you want to get the encapsulated values
4. Select the **HTMLTAGVALUE** function.

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

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

5. Keep the default sheet name, click the **elements** drop down box and then select **NameWithH1Tag.**
6. Click the **Carry over** tab and add content to the carry over list by clicking **Add column to carry over**.
7. Click **Apply settings**.
8. Change the name of column A to **PatentOwner.**

Task 4. The saga continues.

You will now get all of the patents for each individual.

1. Ensure **Sheet2** is selected and then click **Add** sheet.
2. Select a **Function** sheet, click **Categories** and then **html**.
3. Select **HTMLEXTRACTTAGS**. (with an S)
   HTMEXTRACTTAG lets you specify the occurrence. HTMLEXTRACTTAGS selects all.
4. Keep the default sheet name.
5. From the **Content** drop-down, select **Content**.
6. For **tag** type **H2**.
   This is the tag associated with the name of each patent for an individual.
7. Click the **Carry over** tab and then add **PatentOwner** to the list.
8. Click **Apply settings**.
   Now you have a row for each patent that is associated with each individual. If you notice, you also removed the blank name and the one called *Found.*
   Ultimately, your goal might be to count the number of patents for each individual. However, you are going to stop here. The presentation material has not covered the additional topics that are required to complete that goal. Later, once you have this additional information, you could come back here and code the additionally required sheets.
9. Rename your workbook.
10. At the top of the spreadsheet, click **Edit workbook**  .
11. Change the name to **Patent Extract** and then click **Save tag**  .
12. Click **Save**.
13. Select **Save & Exit**.

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

© Copyright IBM Corp. 2012, 2015

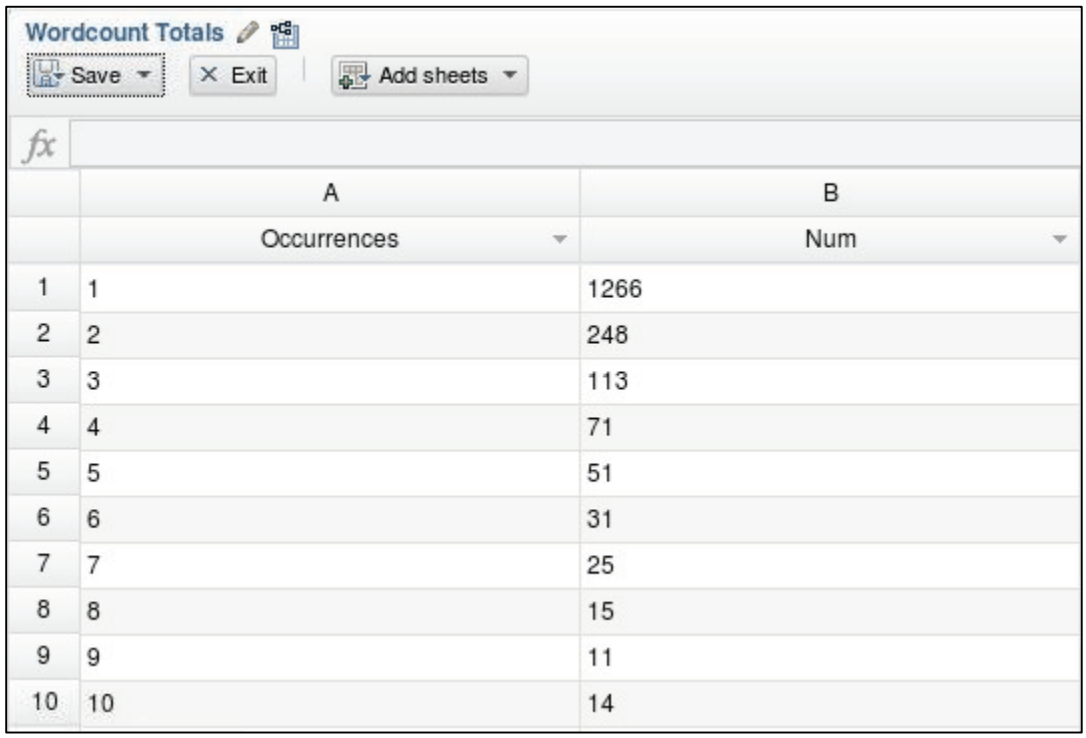
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

14. At this point you also have the option to rename the collection.
15. Click **Save.**

    Notice that this new workbook has not been run. You have only been working with a subset of data. In order for your work to be applied to all of the data, you must run the workbook.
16. Click either one of the **Run** buttons.

    There is a yellow triangle on the left side above the data that indicates that the workbook has not been run. (There is also a progress indicator on the right side.) When the processing completes, the triangle changes to a green checkmark. Also, the percentage complete on the right side goes to 100%.

    The results appear as follows:



	A	B
	Occurrences	Num
1	1	1266
2	2	248
3	3	113
4	4	71
5	5	51
6	6	31
7	7	25
8	8	15
9	9	11
10	10	14

Results:
You created a new workbook based on the web crawler notebook and used sheets to extract data from web pages.

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

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Unit summary

Create a workbook from a master workbook

List the different types of sheets that can be added to a workbook

Describe what each type of sheet can do in a workbook

Add a sheet to a workbook

© Copyright IBM Corporation 2015

Unit summary

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

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

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

© Copyright IBM Corp. 2012, 2015

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Discovering data with expressions, formulas, and functions

IBM BigInsights v4.0

© Copyright IBM Corporation 2015
Course materials may not be reproduced in whole or in part without the written permission of IBM.

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

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

© Copyright IBM Corp. 2012, 2015

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Unit objectives

- Describe the types of expressions supported by BigSheets
- Explain how to access
  - Values from other fields in the same sheet
  - Values from fields in other sheets
- List the different categories of functions supplied with BigSheets
- Describe the purpose of some of those functions
- List and describe some of the supported column data types

© Copyright IBM Corporation 2015

Unit objectives

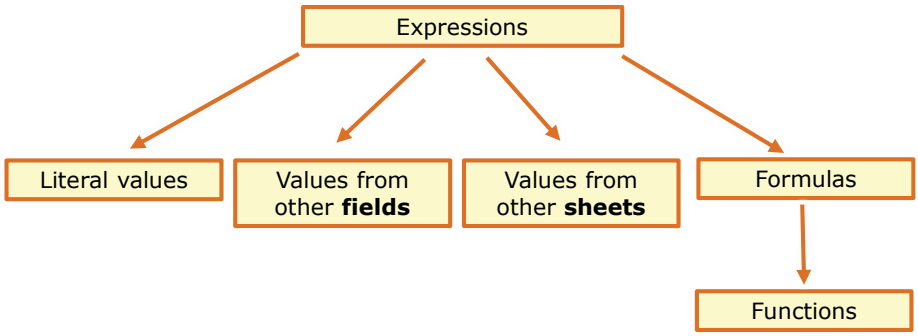
This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

BigSheets expressions

- Expressions allows you to refine, discover, and explore the data.
- Types of BigSheets expressions:
  - Literal values
  - Values from other fields
  - Values from other sheets
  - Formulas
    - Functions



Discovering data with expressions, formulas, and functions

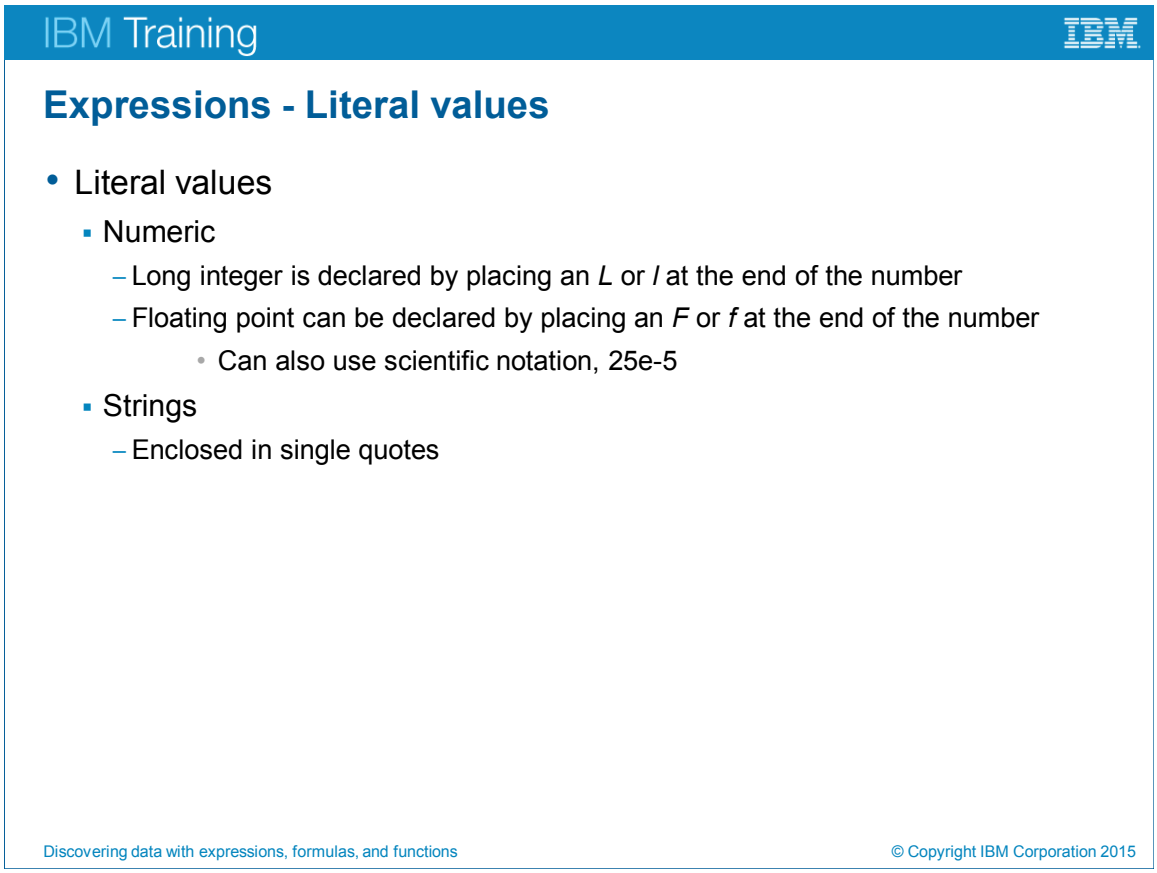
© Copyright IBM Corporation 2015

BigSheets Expressions

BigSheets expressions allow you to refine, discover and explore the data that you have imported. There are four types of expressions. Literal values are either numeric or string values. Values from other fields or columns can be used. Values from other sheets can also be used. Then, perhaps one of the biggest and customizable type of expression, formulas and functions can be used to analyze your data. Functions are basically predefined formulas that perform calculations on the data.

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

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.



Expressions - Literal values

BigSheets support both numeric and string expressions. Integer literals are any sequence of numbers. If you want to specify a long integer, then append an *L* (either an uppercase or lowercase). If you want to specify a floating-point number, then code the numeric string with a decimal point or append the number with an *F* (either uppercase or lowercase). You can also specify a floating-point value by using scientific notation, for example 198e-5.

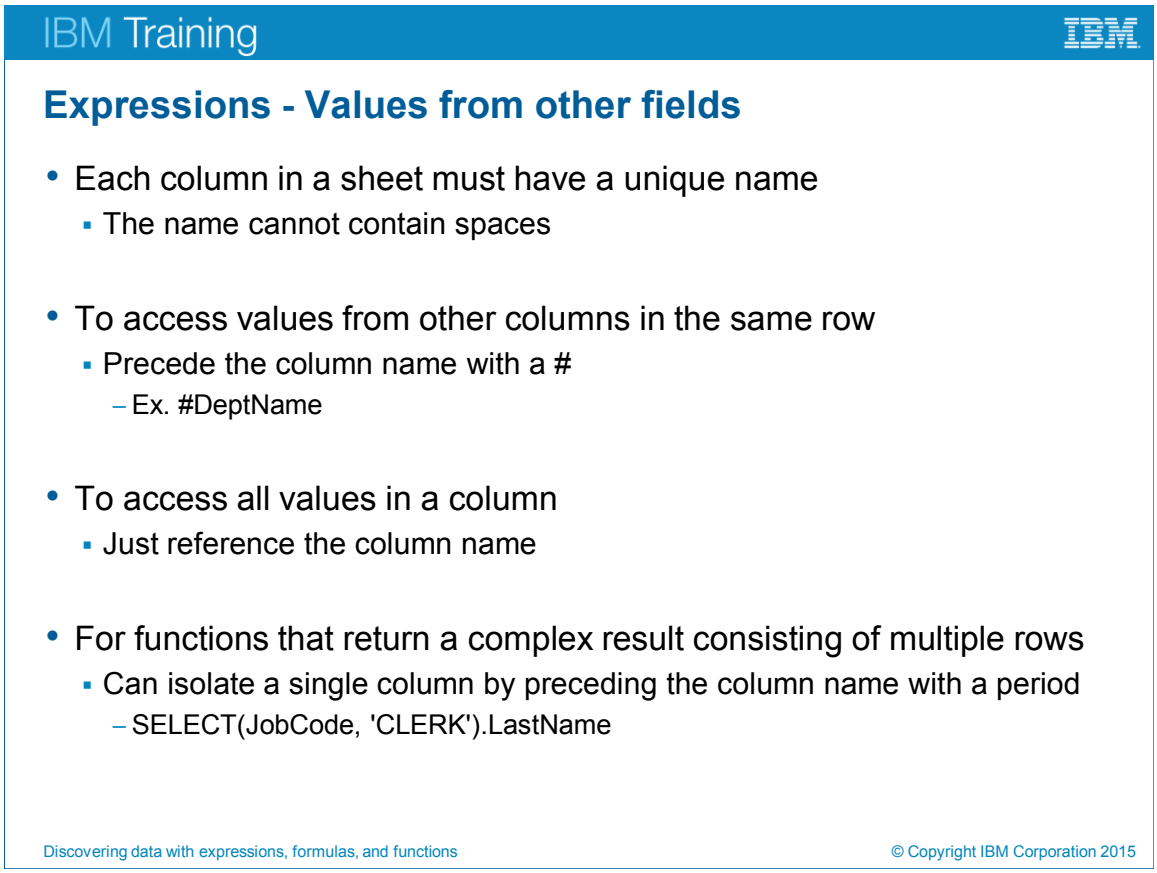
Strings are any set of characters that are enclosed in single quotes.

Important

Column names, worksheet names, and function names are all case-sensitive.

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

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.



Expressions - Values from other fields

Each column in a sheet must have a unique name. Because of that, values in other columns can be accessed using column names. To access the value from the same row for another column, then precede the column name with a #. Assume there is a column with a name of *DeptName*. To calculate the number of characters for each row value in the *DeptName* column, you code LEN(#DeptName).

To access all values in a column, then just reference the column name. Assume that the function assigned to a column, *NewColumn,* is COUNT(DeptName). This function counts the number of rows with values in the *DeptName* column and place the resulting value in each row of the *NewColumn* column.

Some functions return a complex result consisting of multiple rows. In that case, you can isolate a single column by separating it from the function by a period. The SELECT function is an example of a function that returns a complex result. So if you select all rows where there is a value of *CLERK* in the *JobCode* column and want to access the *LastName* column, you code SELECT(JobCode, 'CLERK').LastName.

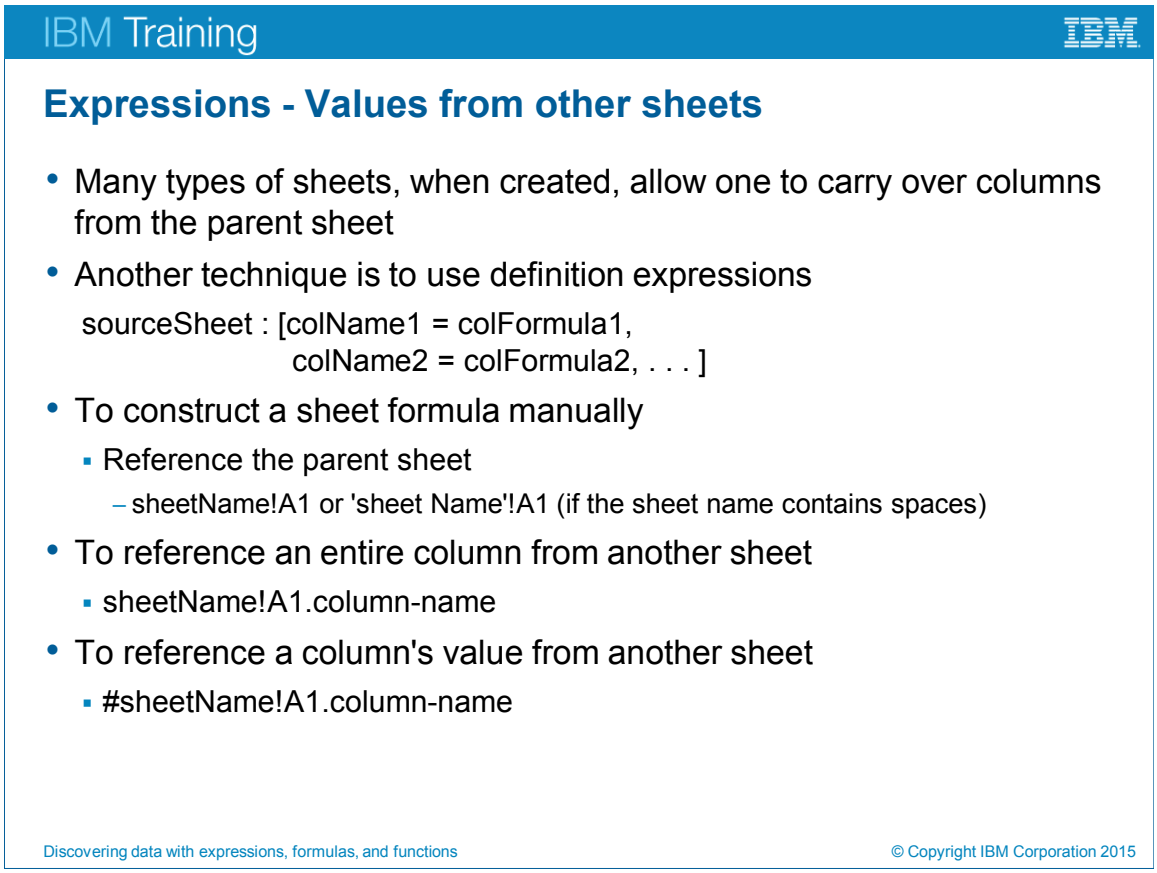
Note

In reality, you would need to code more than just the SELECT function. The SELECT function is coded within another function. Something along the lines of COUNT(SELECT(JobCode, 'CLERK').Lastname).

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

© Copyright IBM Corp. 2012, 2015

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.



Expressions - Values from other sheets

For a number of types of sheets, one is able to specify carry over columns. But if you have a need to do something more than just carry over columns, you might think about using definition expressions while utilizing a *Function* sheet. A definition expression specifies the name of the sheet from which the columns are to be accessed. It then allows you to code a list of new column names and the corresponding column formula to be applied. A definition expression has the following format:

```
sourceSheet : [ columnName1 = columnFormula1,
columnName2 = columnformula2, ...]
```

Example

```
Sheet1 : [DeptName = #Dept, DeptNameLth = LEN(#Dept)]
```

To reference an entire sheet manually in a *Function* sheet, you code *sheetName!A1.* (If the name of your sheet includes embedded spaces then you must include the sheet name in single quotes.

To reference an entire column from another sheet, use the following format:

sheetName!A1.column-name.

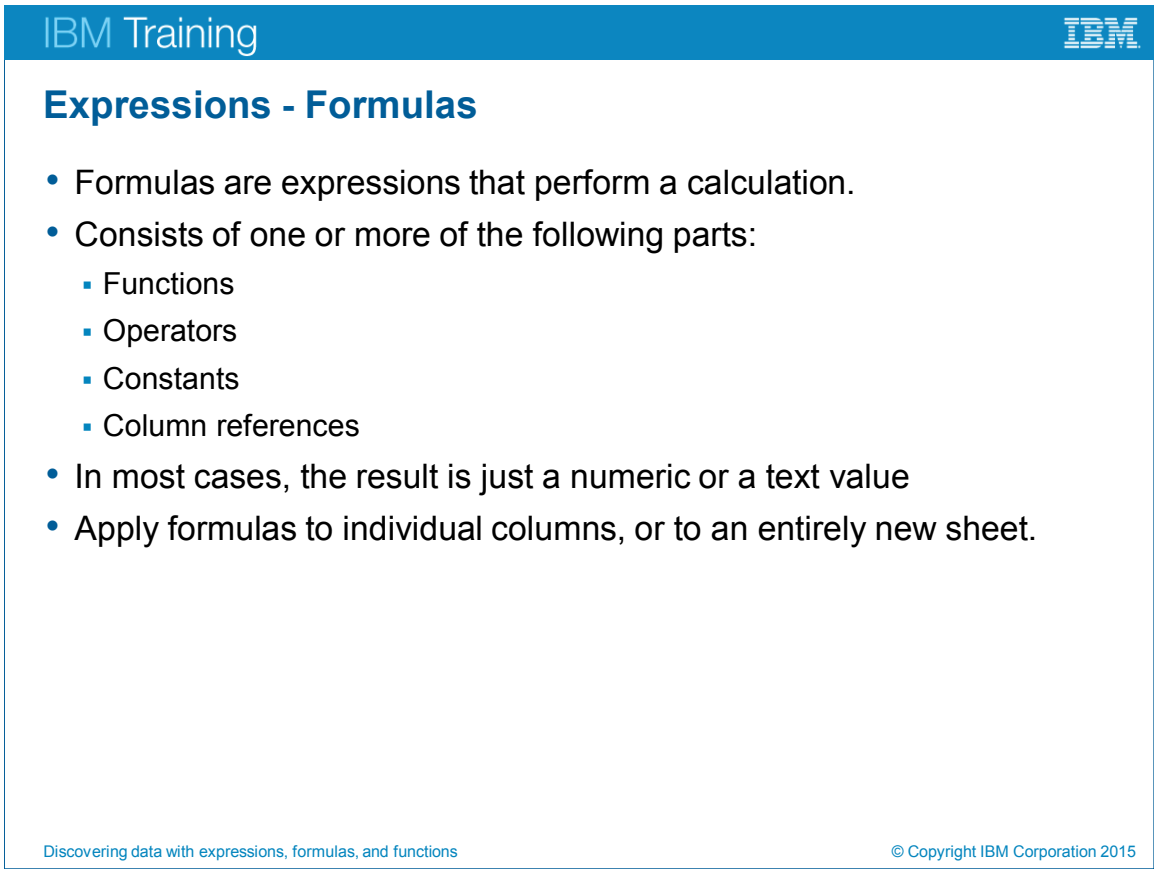
To reference the value of a column from another sheet, code:

#sheetName!A1.column-name.

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

© Copyright IBM Corp. 2012, 2015

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

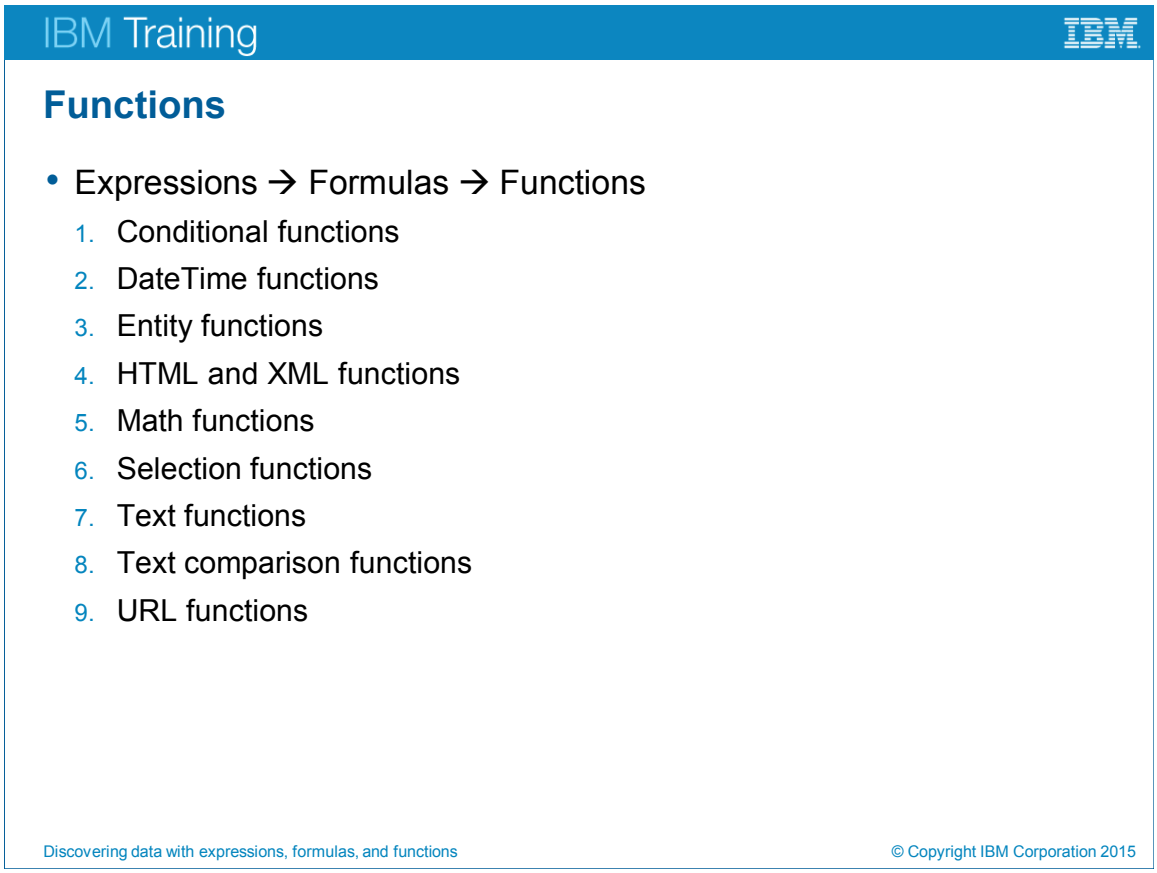


Expressions - Formulas

Formulas are a specific type of expression that performs a calculation and typically returns a single numeric or string value. Formulas consists of one or more of the following parts: function, operators, constants, and/or column references. You can apply formulas on specific columns, or create an entirely new sheet with the data that comes directly from a formula.

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

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.



Functions

There are nine categories of functions, which are essentially predefined formulas, supplied with BigSheets.

┌ Note ————————————

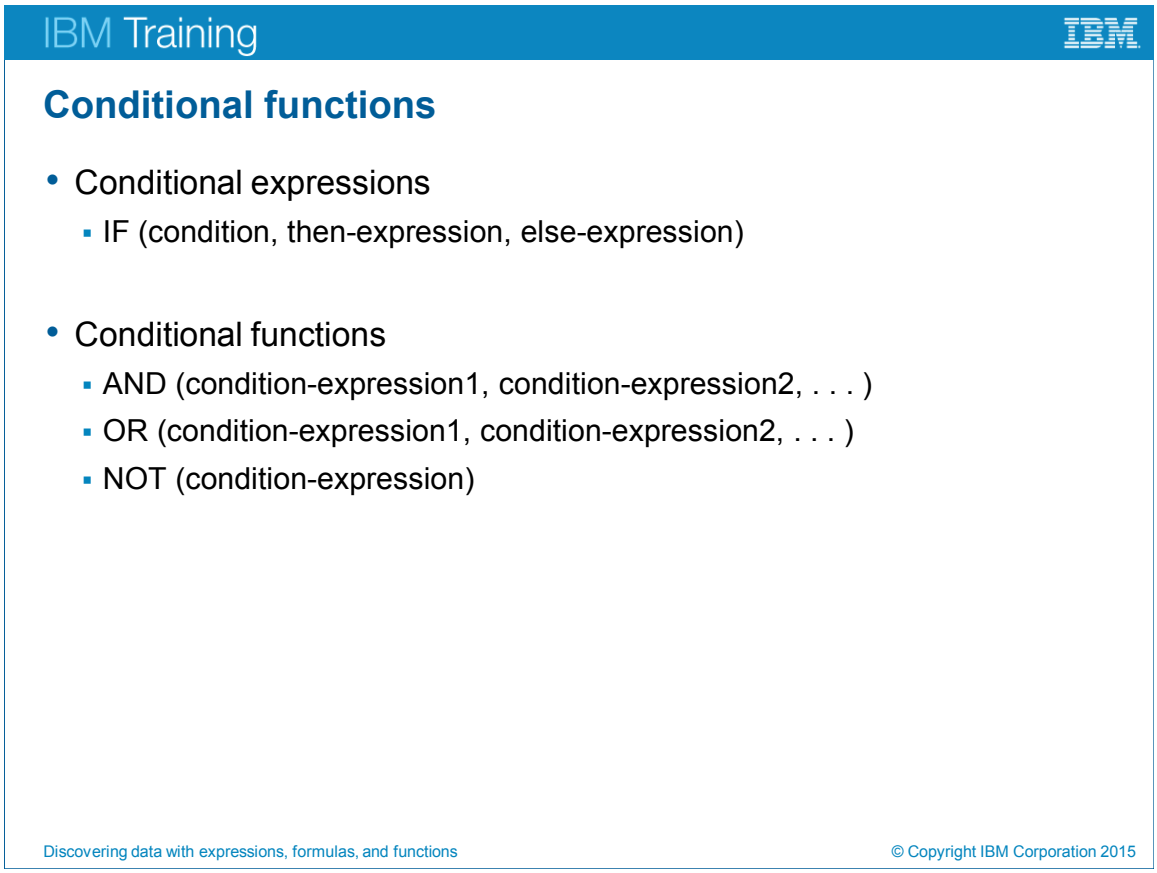
It is also possible for you to write your own functional plug-ins.

└————————————————

There are too many functions here, to cover each one individually. The following few visuals are just to give you an idea of some of the supplied capabilities. Refer to the IBM Knowledge Center to get detailed information about each function.

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

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.



Conditional functions

To apply conditional logic to a set of functions, use the *IF* function. This is of the format:

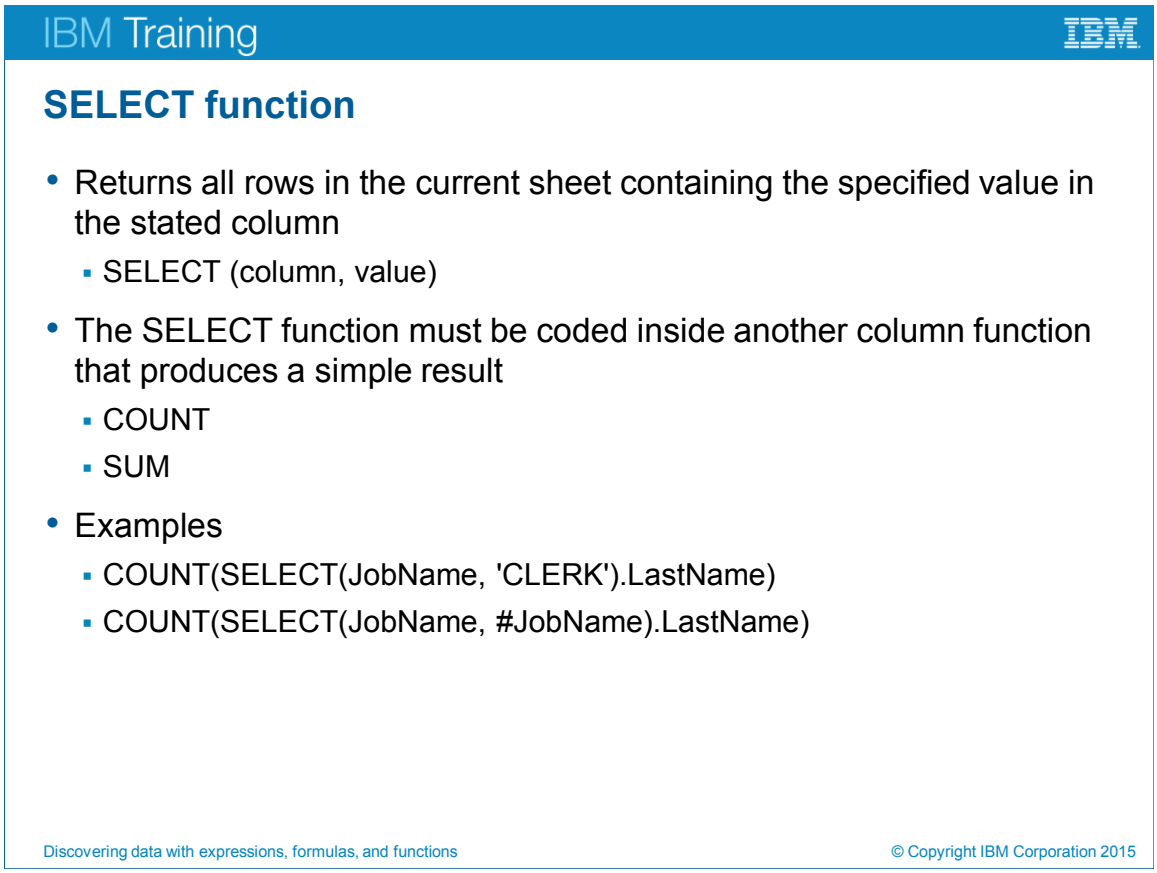
IF (condition, then-expression, else-expression)

If the condition evaluates to true, then the *then-expression* is executed. If the condition evaluates to false, then the *else-condition* is executed. The *then-expression* and the *else-expression* are any formula expressions that produce a string or numeric value.

Boolean logic is supported in an *IF* condition through the use of conditional functions. The *AND* and *OR* condition functions require at least two condition expressions and can have an unlimited number. A condition expression can be another condition function.

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

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.



SELECT function

The SELECT function returns all rows in the current sheet containing the specified value in the specified column. For example,

SELECT(GENDER, 'F')

selects all rows where the *GENDER* column is equal to an *F.*

The SELECT function returns a complex result. Essentially, this means that the returned result is not a single string or numeric value but rather the values from multiple rows. For this reason the SELECT function must be coded inside another column function that does return a simple result. Coding the SELECT function within the COUNT function allows you to get the total number of rows returned. Coding the SELECT function within the SUM function allows you to total the values in the returned column for all of the rows that met the select criteria.

If you wanted to count the number of employees that held the position of CLERK, then you might code

COUNT(SELECT(JobName, 'CLERK').LastName)

The SELECT function returns the *LastName* column for all rows where there is a value of CLERK in the *JobName* column. The COUNT function then counts the number of rows.

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

© Copyright IBM Corp. 2012, 2015

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

If you coded

COUNT(SELECT(JobName, #JobName).LastName)

then you get a much different result. #JobName refers to the value in the *JobName* column for each row. So

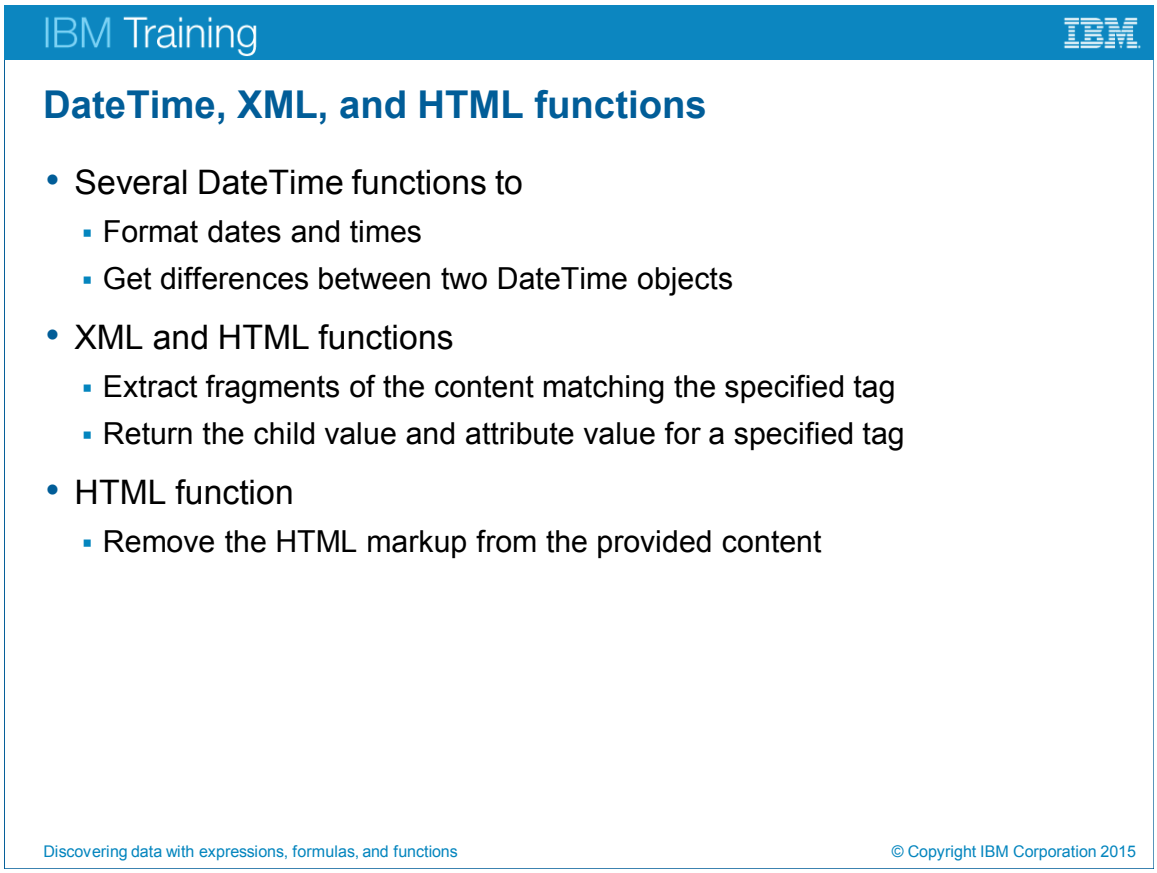
SELECT(JobName, #JobName).LastName

takes the value in the *JobName* column for the first row and then returns all rows where the value in the *JobName* column equals the value in the first row. When the COUNT function is applied, the number of rows return by the *SELECT* are counted and that value is returned. The *SELECT* next looks at the *JobName* value in the second row and goes through the process again. This is repeated for each row in the sheet.

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

© Copyright IBM Corp. 2012, 2015

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.



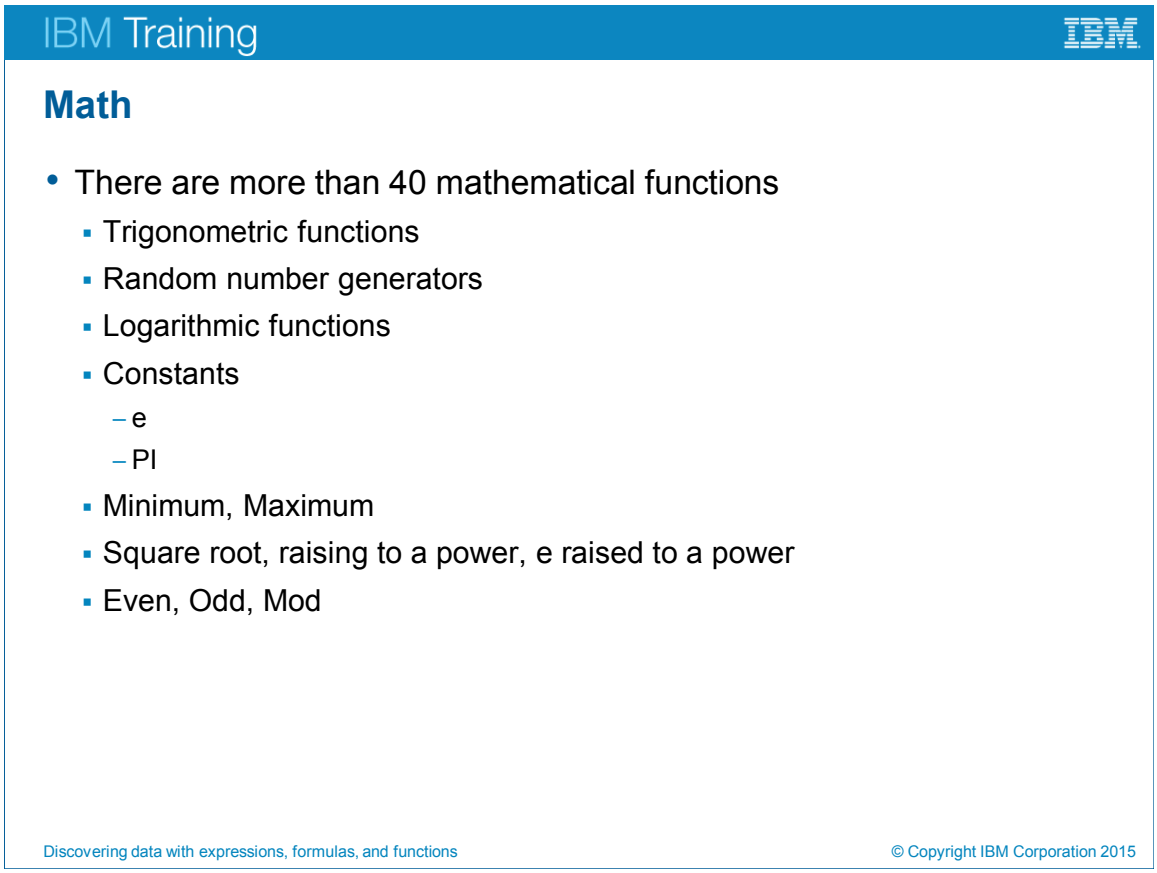
DateTime, XML, and HTML functions

There are several date functions that return a formatted date. There is one that gives you the difference between two dates.

For XML and HTML data, there are functions that extracts fragments of the content matching the specified tag. Also you can get access to the child value and attribute value for a specified tag.

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

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

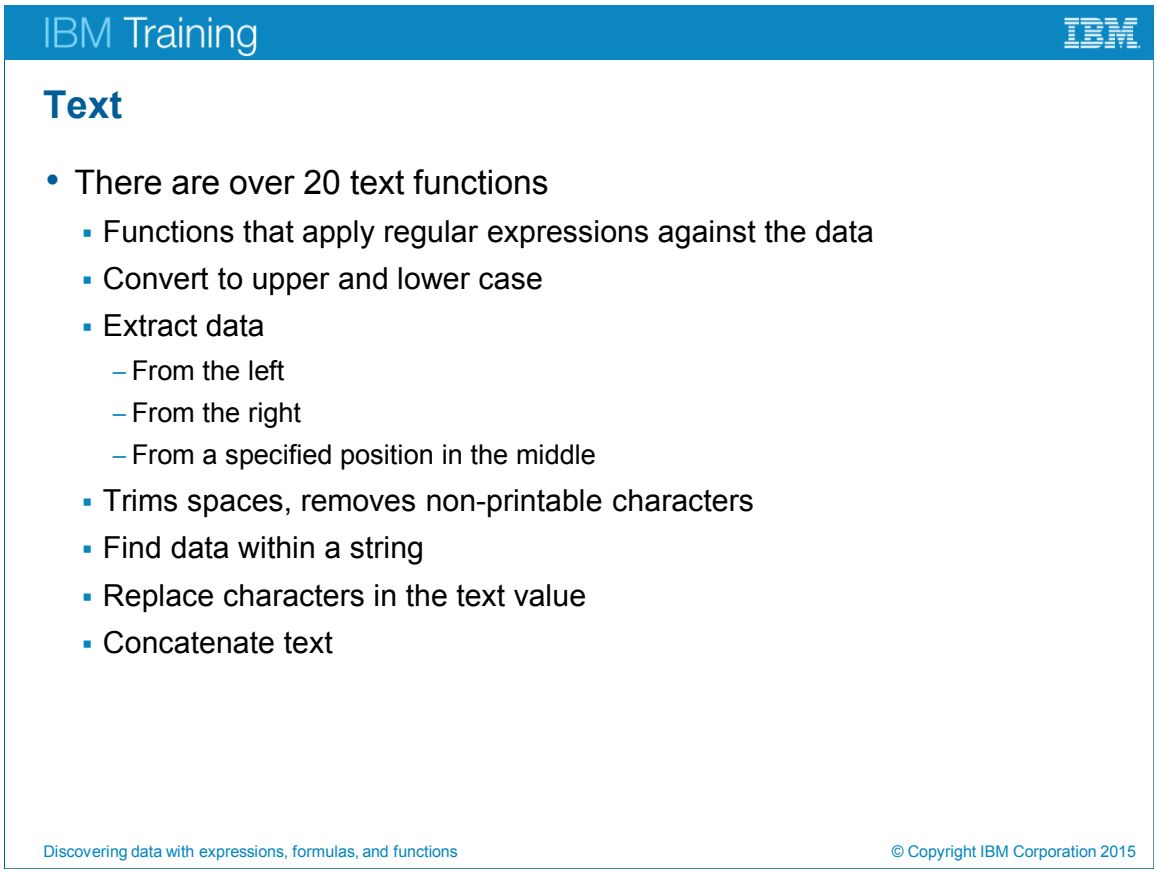


Math

There are 48 mathematical functions supplied. There are a number of trigonometrical functions, random number generators, both between 0 and 1 and between two supplied values. There are logarithmic functions. You have access to constants, like PI and e. You can find the minimum or maximum values of a column, find the square root of a value, raise a value to some power or raise the constant e to some power. You can round a number up to the nearest odd or even integer, take the modulo of a number, find the absolute value, truncate a number to an integer and test to see if a string is a valid number.

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

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

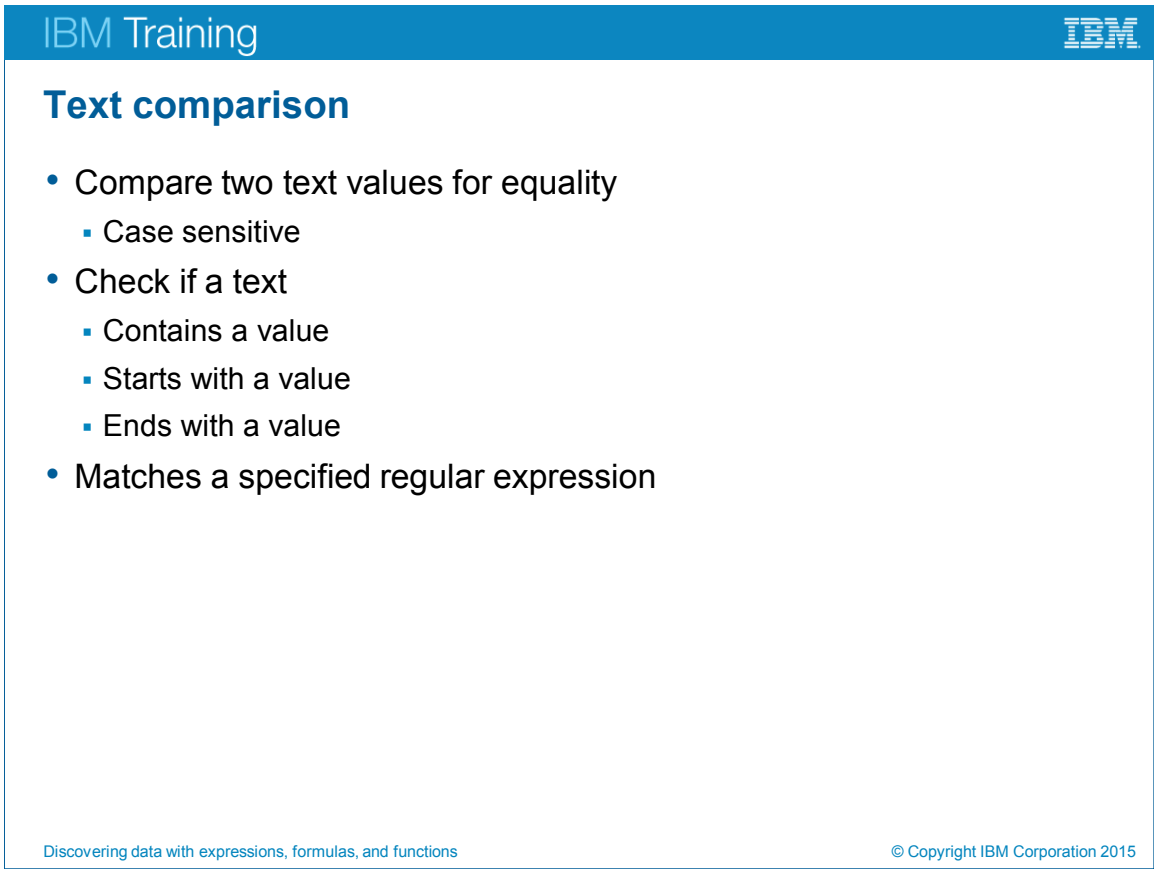


Text

There are 28 text functions. There are a number that work with regular expressions in some way. You can convert the text to upper or lower case, concatenate text values, and extract data from various areas of the text field. You can search for some occurrence of a string within the text and replace characters within the text. You can get the length of your text string and you can count the number of individual words within the provided text.

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

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.



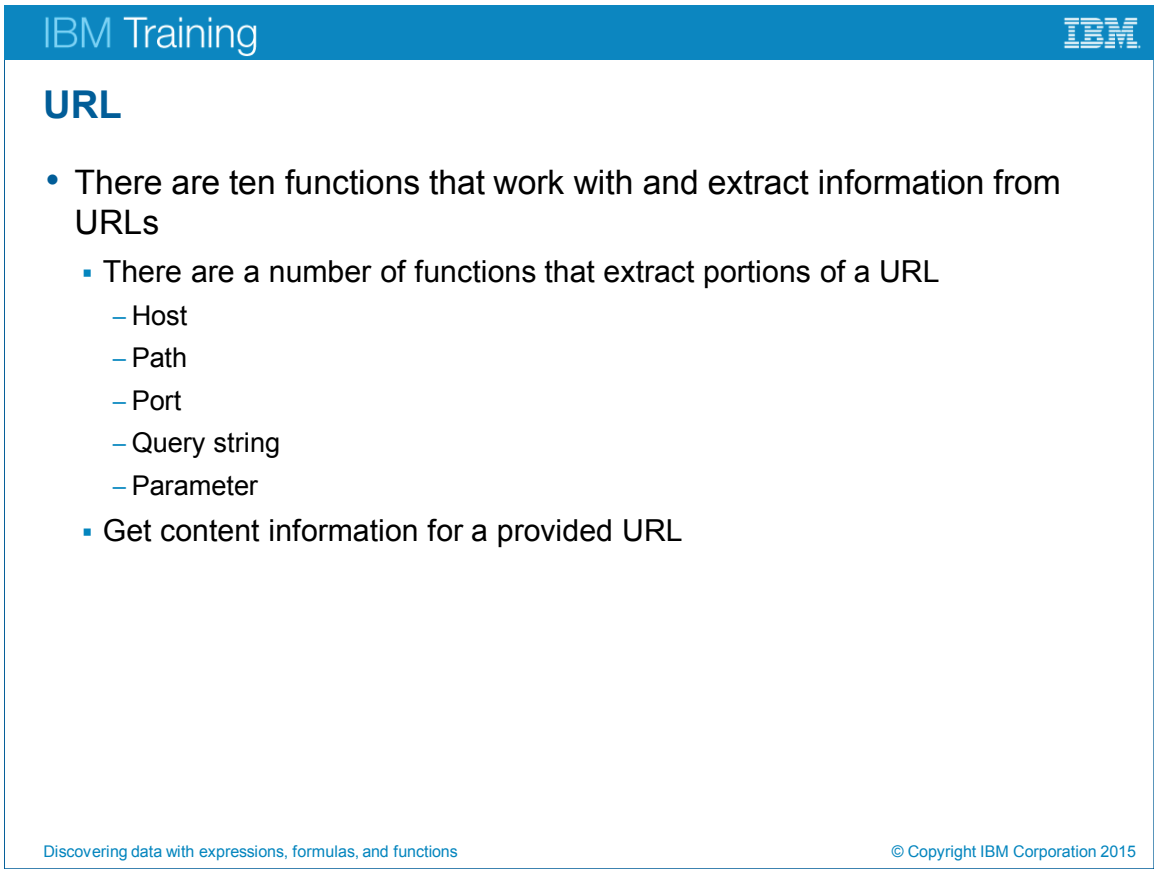
Text comparison

There are a few text comparison functions that return values of true and false. Most likely these functions would be used in the IF (...) function.

You can compare if two values are equal. This is based upon case sensitivity. You can check if a text value contains some value, starts with a value or ends with a value. And you can test if a text value matches a supplied regular expression.

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

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.



URL

There are ten functions that work with and extract information from URLs. Most of these functions provide portions of the URL, like the host, port, path, query string, and parameter. Also there are functions that return content information, assuming that a connection can be established.

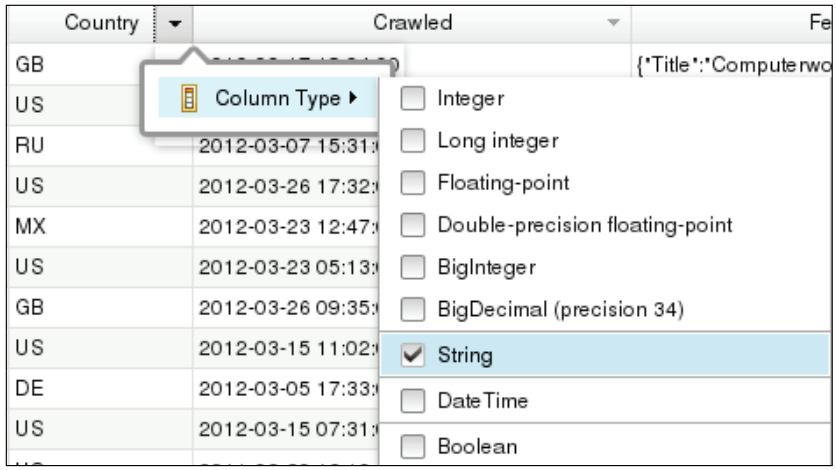
This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

BigSheets data types

- BigSheets support common data types that you can use with the columns of your worksheets.
- Specify this on the master workbook:
  - BigDecimal
  - BigInteger
  - Boolean
  - DateTime
  - Double
  - Float
  - Integer
  - Long
  - String



Discovering data with expressions, formulas, and functions

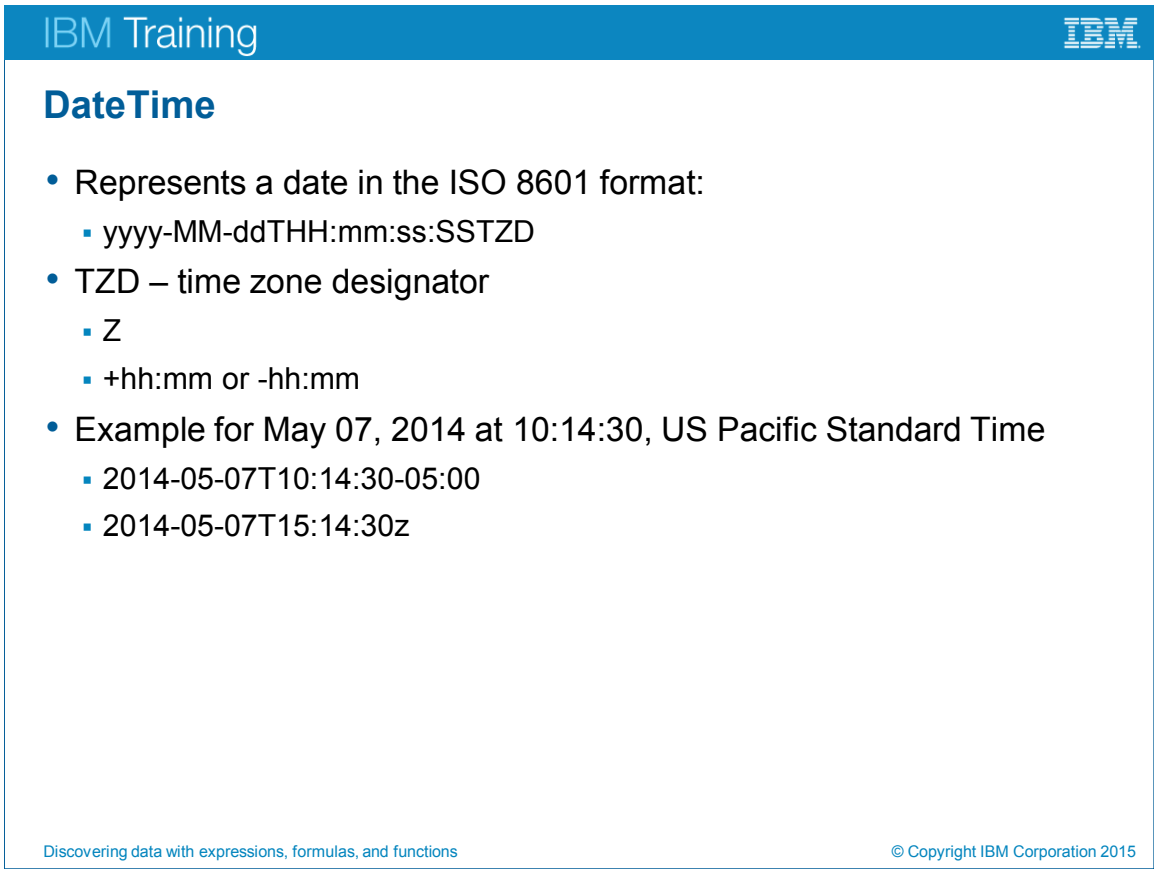
© Copyright IBM Corporation 2015

BigSheets data types

BigSheets workbooks supports common data types. There are too many here to for us to cover everything individually and some of these are common enough where we do not need to cover them. The following visuals will give you an idea of some of the data types available. For the others, you can refer to the IBM Knowledge Center for more information.

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

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

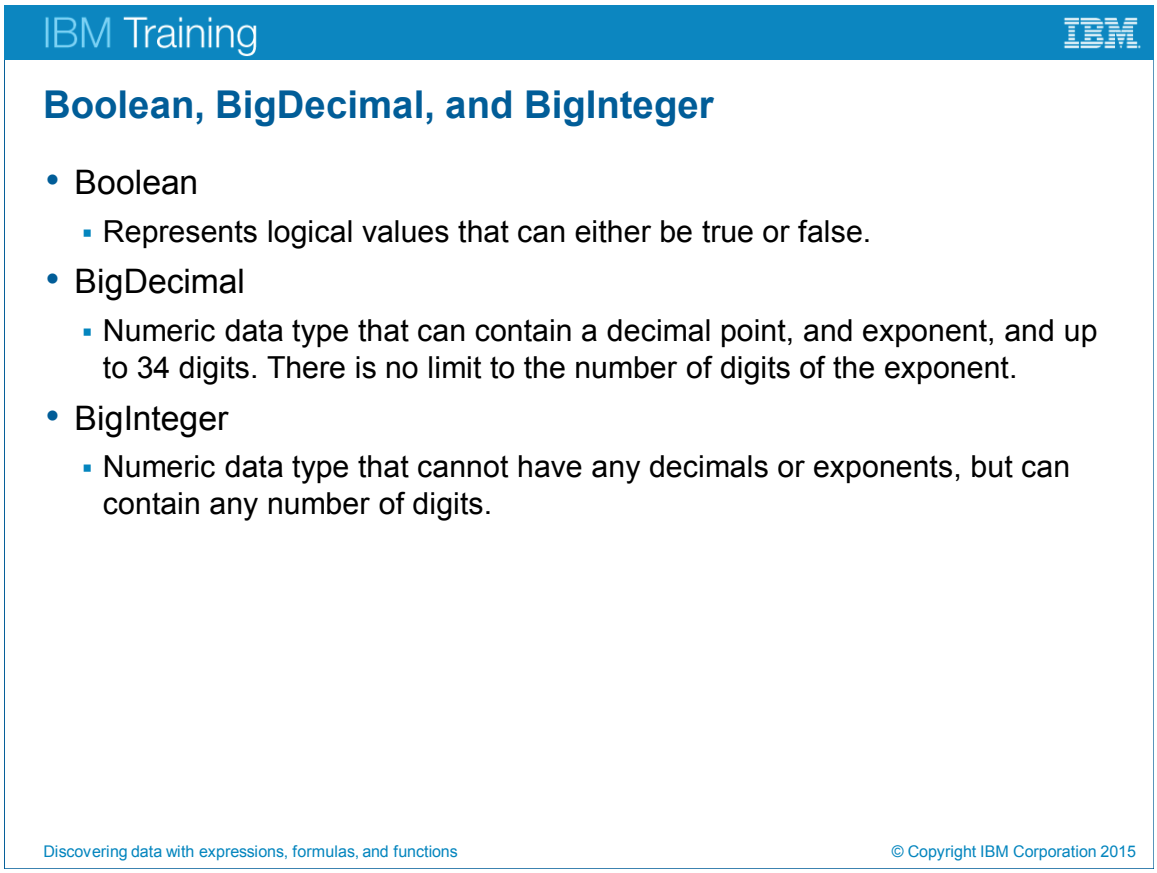


DateTime

The DateTime function is useful for when you need to extract information in your worksheet based on a date range. The format is ISO 8601 where you need to specify the year, month, date, followed by time. Then there is the time zone designator that you specify to indicate the time zone. The example shown here illustrates what a particular date would look like in the ISO 8601 format.

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

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

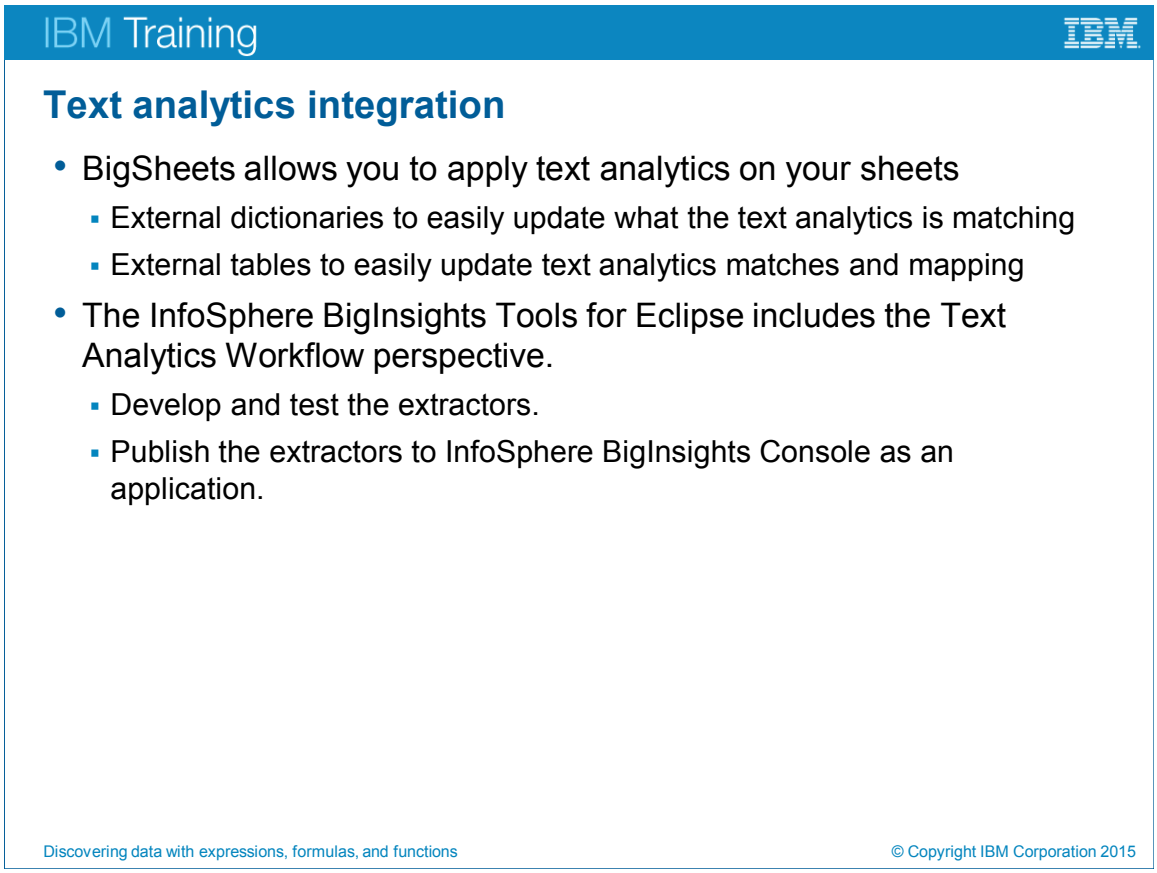


Boolean, BigDecimal, and BigInteger

Boolean represents logical values that can either be true or false. BigIntegers are integers values that cannot have any decimal points or exponents but can contain any number of digits. BigDecimals can contain a decimal point, an exponent, an up to 34 digits. There is no limit on the number of digits of the exponent.

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

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.



Text analytics integration

BigSheets allow for text analytics on your sheets. You can use external dictionaries to let you easily update what the text analytics is matching. You can also use external tables as well.

There is the Text Analytics Workflow perspective within the InfoSphere BigInsights Tools for Eclipse. Use that to create your extractors and then you can deploy it as an application to the BigInsights Console.

The details of this text analytics integration is outside the scope of this course, but I wanted you to be aware of this feature and what is available.

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

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Checkpoint

1. List the four types of BigSheets expressions.
2. How do you refer to the value from a different column in the same row?
3. True or False? The SELECT function must be coded inside another function to get any meaningful data.

© Copyright IBM Corporation 2015

Checkpoint

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

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Checkpoint solutions

1. Four types of expressions:
   1. literal values
   2. values from another field
   3. values from another sheet
   4. formulas
2. To refer to the value from a different column of the same row, use the # symbol.
   - For example. #DeptName
3. True. The SELECT function must be coded inside another function that returns a simple value such as COUNT or SUM.
   - For example: COUNT(SELECT(JobName, 'CLERK').LastName)

© Copyright IBM Corporation 2015

Checkpoint solutions

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

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Demonstration 1

Working with Functions

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

- Use some of the functions learned to work with and extract data in a sheet

© Copyright IBM Corporation 2015

Demonstration 1: Working with Functions

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

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

Demonstration 1: Working with Functions

Purpose:
You will work with the workbook that was created from the WordCount program. You will add new columns with specific functions to the sheet.

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

Task 1. Background.

The WordCount application generated a record for each unique character string found in the specified document. It then totaled the number of occurrences of each unique character string. Your goal is to create a sheet that has the number of occurrences for all character strings that occur the same number of times.

1. In **BigSheets** select the **Wordcount** workbook.

   Displayed for each row is a character string, followed by a tab character, followed by a number. The number represents the occurrences of that character string in the *last_of_the_mohicans.txt* document.

   Your goal is to count the number of character strings that occur the same number of times.

   In the previous exercise, you used Function sheets to apply functions to your data. This time, you are going to add new columns to a sheet and code the needed functions directly.

Task 2. Create a new workbook.

1. From within the **Wordcount** workbook, click **Build new workbook**.
2. Click **Add Sheets** and then select **Formula**.

   Keep the default sheet name.

   You are to create a new sheet and specify a formula that references the *Header* column from the Wordcount workbook. The result is to have a new column called Occurrences. Then, use the *GETGROUPMATCH()* function to populate this new column. Here is the format of that function as shown in the BigInsights Knowledge Center. *GETGROUPMATCH (text,regex,group,number*)

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

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

You are provided the regular expression to use. You know that you want the second group. But what about the text parameter? You want this function to be applied to each row for the Header column. To indicate that, you specify **#Header**. As in:

GETGROUPMATCH(#Header,'(.+\t)([0-9]+)',2)

You are to reference the Wordcount sheet and, as stated, this new column is to have the name of **Occurrences**. The final parameter, 2, says to extract the second group from the regular expression.

3. Code the following in the **fx** field.

```
Wordcount!A1 : [Occurrences =
GETGROUPMATCH(#Header,'(.+\t)([0-9]+)',2) ]
```

4. Click **Apply settings**.

Task 3. Count the occurrences.

Add another column to your sheet.

1. Click the drop-down for **Occurrences** and then select **Insert Right->New Column**.
2. Type **Num** for the name.

   You want to take the value for each row in the Occurrences column and select all rows that have that same value. Then count the number of those rows returned.
3. In the **fx** field for the **Num** column, type the following formula:

```
COUNT(SELECT(Occurrences,#Occurrences).Occurrences)
```

4. Click **Save formula**  .

   You can see that there were 1266 unique character strings that only occurred once in the document. But you do not want to see that 1266 times. You need to remove duplicate rows.
5. Make sure Sheet1 is selected, add a new sheet and then select **Distinct.**

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

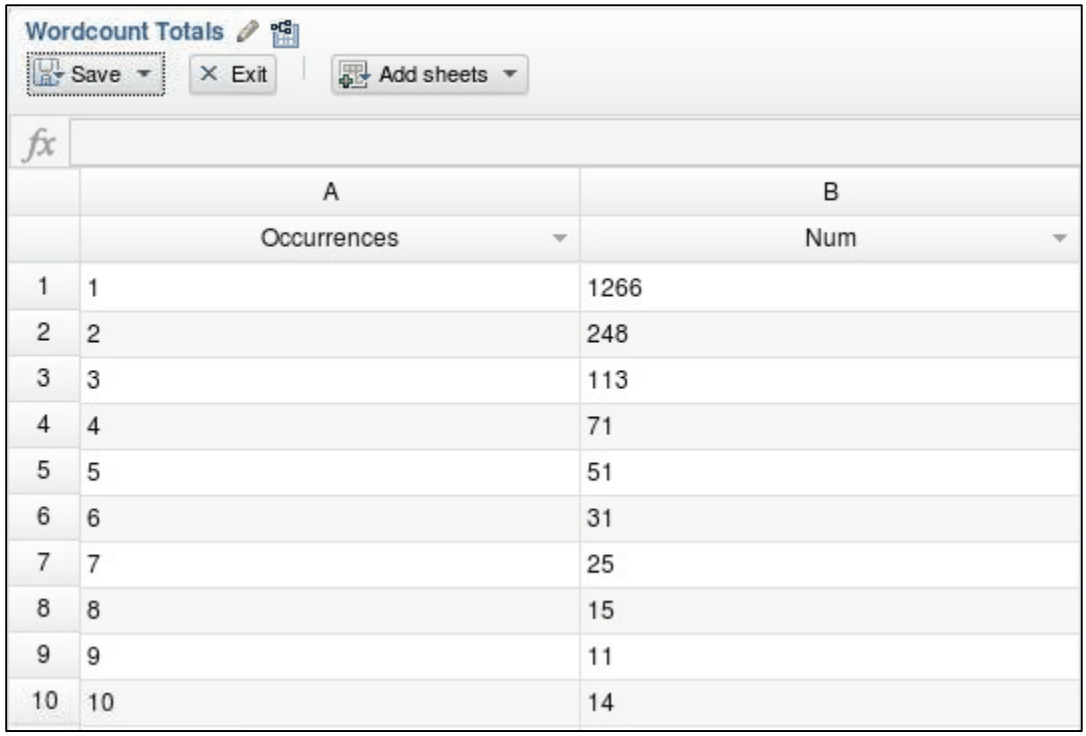
© Copyright IBM Corp. 2012, 2015

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

6. Keep the default name for the new sheet and click **Apply settings**.

   Now you have your results. There are 1266 unique character strings that occurred once. Fourteen unique character strings that occurred 10 times, etc.

   The results appear as follows:



	A	B
	Occurrences	Num
1	1	1266
2	2	248
3	3	113
4	4	71
5	5	51
6	6	31
7	7	25
8	8	15
9	9	11
10	10	14

7. Click **Save**, **Save and Exit**, and name the workbook **WordCount Totals**.
8. Click **Save**.

Results:
Using the workbook that was created from the WordCount program, you added new columns with specific functions to the sheet.

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

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Unit summary

- Describe the types of expressions supported by BigSheets
- Explain how to access
  - Values from other fields in the same sheet
  - Values from fields in other sheets
- List the different categories of functions supplied with BigSheets
- Describe the purpose of some of those functions
- List and describe some of the supported column data types

© Copyright IBM Corporation 2015

Unit summary

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

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Integrating with Big SQL

IBM BigInsights v4.0

© Copyright IBM Corporation 2015
Course materials may not be reproduced in whole or in part without the written permission of IBM.

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

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

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Unit objectives

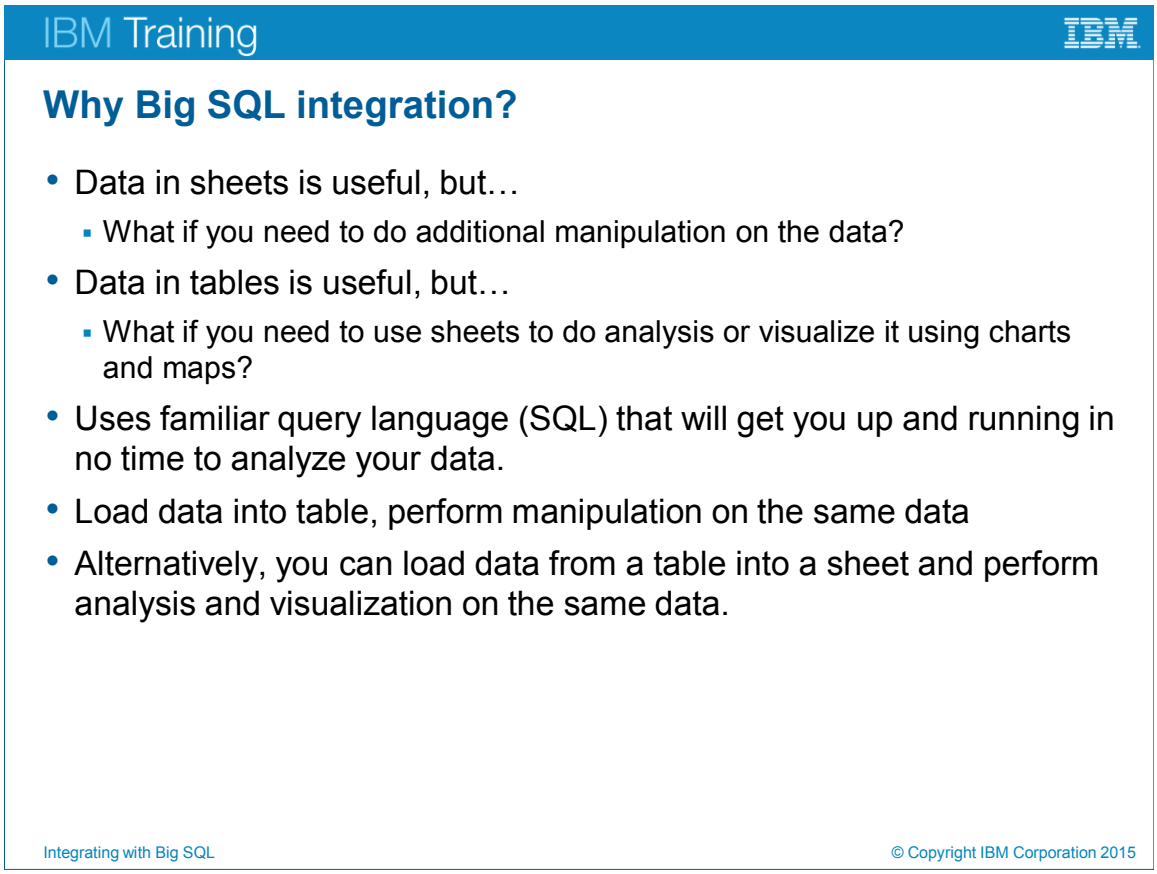
- Manage the data lifecycle through Big SQL
  - Create new tables using the same data as the sheet
  - Create new sheets using the same data as the table

© Copyright IBM Corporation 2015

Unit objectives

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

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.



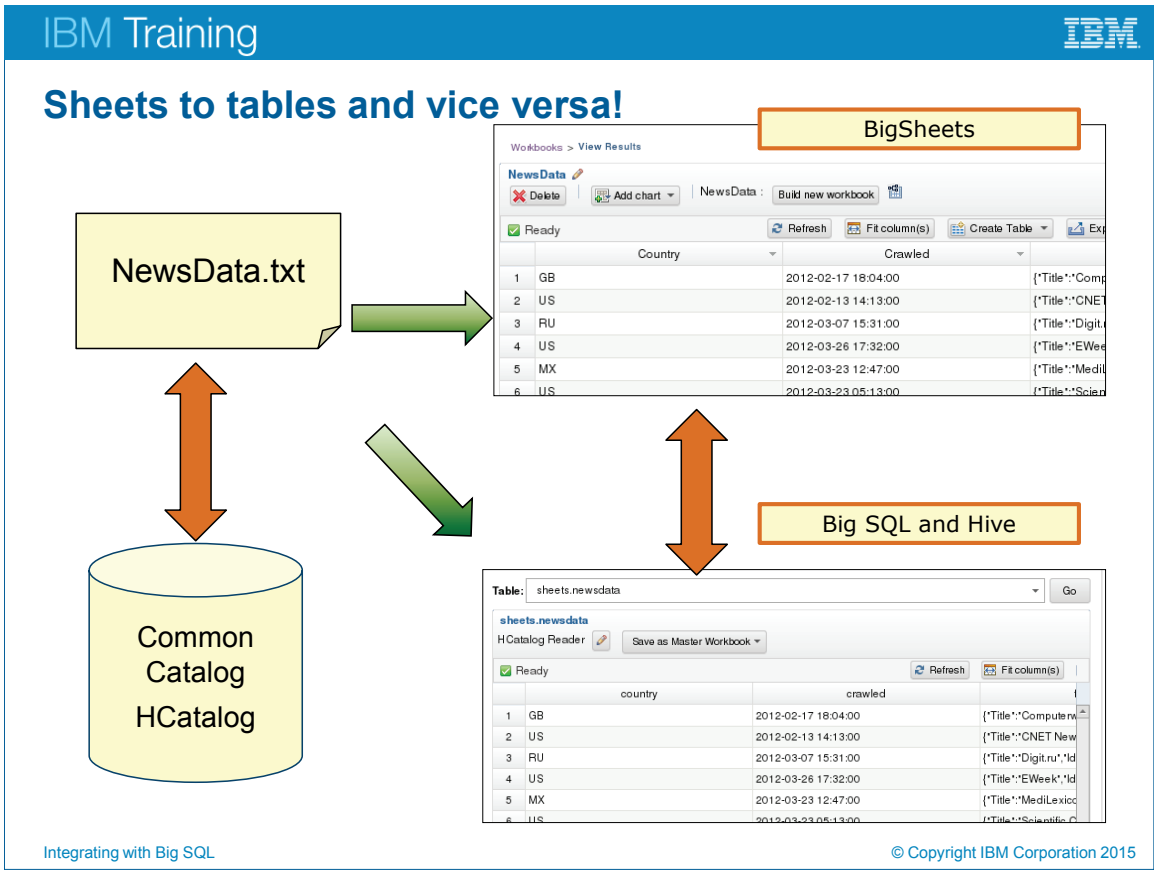
Why Big SQL integration?

BigSheets provides a simple and quick way for you to integrate with Big SQL tables. You may be thinking, why would I want to do that? Well, the answer is simple, it depends on the type of calculation and analysis that you need to do. BigSheets provides spreadsheet-like capabilities, but if you have additional need for query languages, then Big SQL offers that ability. Big SQL queries uses the common SQL, which will enable you to work with your data without the need to learn anything more. The scope of this unit only covers the integration aspect of Big SQL and BigSheets. To learn more about the usage of Big SQL, refer to the BigInsights Knowledge Center or find more information in the Big SQL course.

Likewise, if your data currently resides in tables within Big SQL and you need the analytics and visualization tools that BigSheets has to offer, then you can quickly load that data into a sheet.

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

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.



Sheets to tables and vice versa!

Here you see that the initial data in the worksheet came from a text file. The same text file is used when you choose to create a table from your sheet. You can think of the two as different views of the same data. Both share a common catalog within BigInsights.

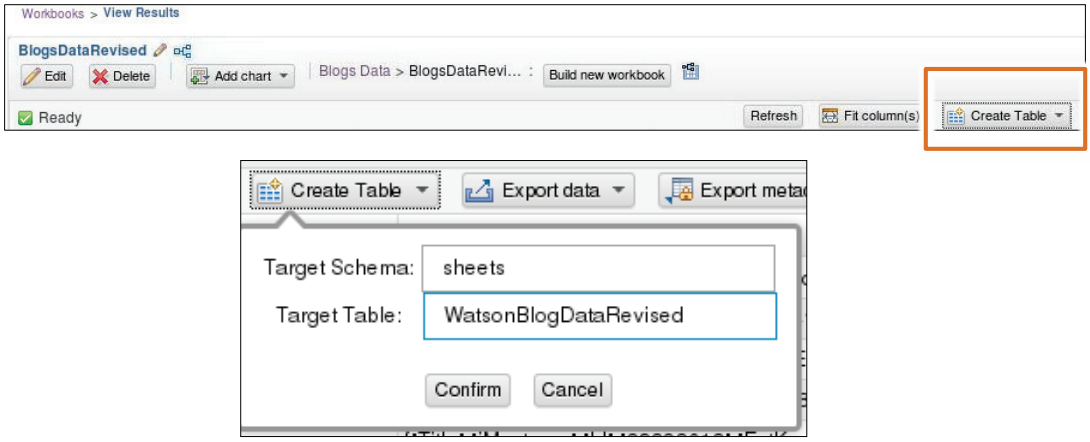
This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Creating a new table using the same data as the sheet

- Create a table from within the **BigSheets workbook.**
  - The table will be created in the common catalog
  - One table per sheet
- **Export the Data** and create a SQL query to create the table.



Integrating with Big SQL

© Copyright IBM Corporation 2015

Creating a new table using the same data as the sheet

You can create a table using the same data as a sheet in your workbook. The table will be created in the common catalog. When the sheet updates, the table also gets the new data since they are sharing the same data. Each sheet can only have a single table. Once a table has been created for a particular sheet, the option to *Create Table*, becomes the option to *Delete Table,* which you can use to delete the table from the catalog.

Alternatively, you can create a table by exporting the workbook data, and creating a query to insert that data. This allows for more flexibility as you can use the data from your workbook and insert it into other database systems.

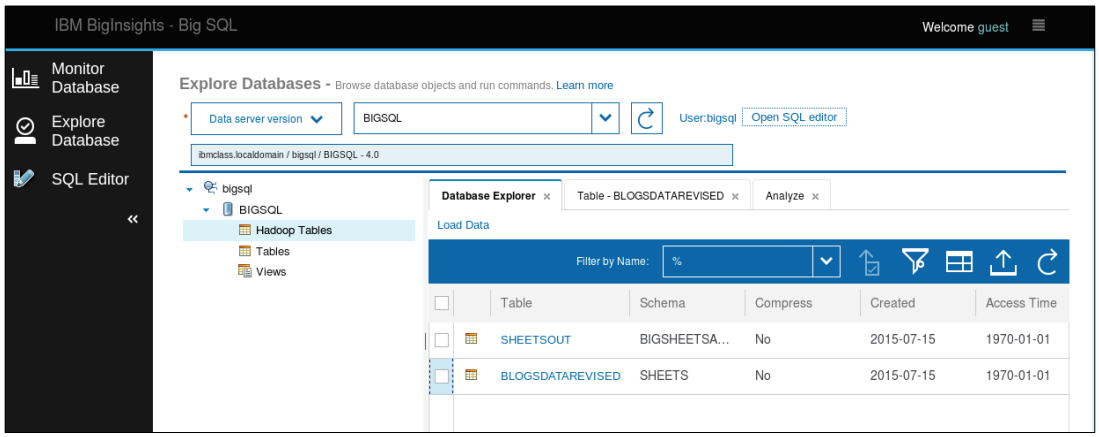
This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

View the data in the BigInsights Home, Big SQL page

- **Explore Database** and select **Hadoop Tables**
- Run Big SQL query on the data
  - Using the **SQL Editor**
  - Using the JSqsh shell



Integrating with Big SQL

© Copyright IBM Corporation 2015

View the data in the BigInsights Home, Big SQL page

Once a table has been created, you can view the data and schema on the **Explore Database** page.

BigInsights provides two interfaces for you to run Big SQL queries against the data. You can work with the Big SQL tables using the SQL Editor on this page from BigInsights Home, or use the JSqsh shell.

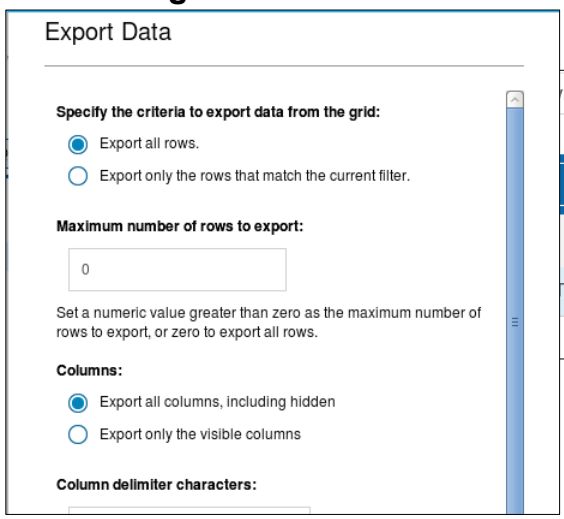
This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

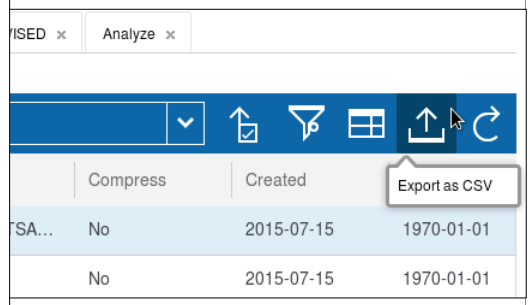
© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Creating a new sheet using the same data as the table

- Use data from existing tables to create a sheet.
- **Export** the table as CSV
- Create a **BigSheets** workbook from that CSV





Integrating with Big SQL

© Copyright IBM Corporation 2015

Creating a new sheet using the same data as the table

Alternatively, you can also create a sheet using data that resides in a table. You first export the table as CSV and then create a BigSheets workbook from that data.

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

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Checkpoint

1. How do you create a Big SQL table from a workbook?
2. How do you create a workbook from a Big SQL table?
3. True or False? Both the sheet and the table share the same data source.

© Copyright IBM Corporation 2015

Checkpoint

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

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Checkpoint solutions

1. To create a Big SQL table from a sheet:
   1. Click the **Create Table** pushbutton and specify the table schema and name. Alternatively, you can also export out as a TSV format and create a SQL query to create the table based on that.
2. To create a sheet from a Big SQL table:
   1. Export the table's data as CSV and create a workbook from that file.
3. True. Both the sheet and table share the same data.

© Copyright IBM Corporation 2015

Checkpoint solutions

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

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Demonstration 1

Integrating with Big SQL

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

- Use Big SQL tables with BigSheets

© Copyright IBM Corporation 2015

Demonstration 1: Integrating with Big SQL

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

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

Demonstration 1: Integrating with Big SQL

Purpose:
The integration with Big SQL allows you to perform additional analytical queries against the data. It extends upon the capabilities of BigSheets allowing you to use common SQL queries to get insight from the data. You will use Big SQL tables with BigSheets.

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

Task 1. Loading the test data into the HDFS.

You are going to use the **blogs-data.txt** file for this demonstration. Do the following steps to load the file into the HDFS. The data in the blogs-data.txt file comes from blogs that reference the term IBM Watson.

In this demonstration you are going to turn that text data into a BigSheets workbook, and then use the functions in BigSheets to format the data into something that is easier to understand.

To examine the blogs data in the blogs-data.txt file, you will create a workbook and use that data for a new Big SQL table. This demonstration introduces a way of creating tables from data that you analyze by using BigSheets and a TSV reader format and a JSON Array format.

1. Open up a terminal and enter in this command:

```
hdfs dfs -mkdir /user/biadmin/Watson
```

1. Use this command to upload the blogs-data.txt

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

2. Do another listing to confirm that the file has been loaded:

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

3. Change the permissions on the Watson directory:

```
hdfs dfs -chmod 777 /user/biadmin/Watson
```

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

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

Task 2. Start up the Big SQL service.

1. Inside the Ambari console, ensure **BigInsights - Big SQL** has started. If not, start it up.

   Big SQL requires that the monitoring utility package is started as well.

2. Open up a terminal to start the monitoring utility package.
3. Switch to the root user, by typing `su -` and then type the password `dalvm3`.
4. Change directory to the following path:

   ```
   cd /usr/ibmpacks/bigsql/4.0/dsm/1.1/ibm-datasrvrmgr/bin/
   ```

   You will run the dsmKnoxSetup script as the root user.

5. Run the script `/dsmKnoxSetup.sh -knoxHost <knox-host>`

   where <knox-host> is the host where the Knox gateway is running. In our case, it would be *ibmclass.localdomain*

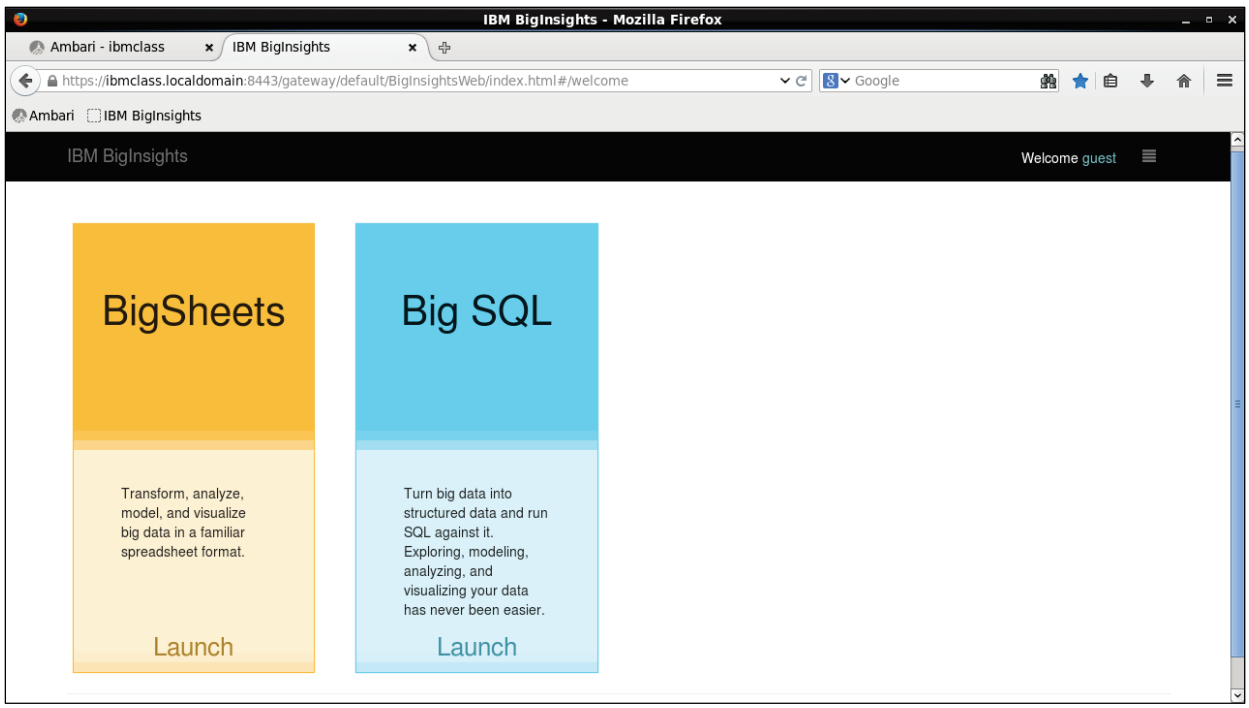
   ```
   ./dsmKnoxSetup.sh -knoxHost ibmclass.localdomain
   ```

2. When prompted to continue running with the above value, select **1**.

   Remember, when you restart the Knox server, you will have to run the dsmKnoxSetup script again.

   You will now be able to access Big SQL via the **BigInsights Home** page.

6. Check to see that **Big SQL** is available.



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

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

Task 3. Create a BigSheets Workbook.

1. Click the **BigSheets** link from the **BigInsights** Home.

   This takes you to the list of all workbooks.

2. Click **New Workbook** 

.
3. In the **Name** field type **BlogsData.**
4. Drill down to **/user/biadmin**, expand the **Watson** directory and the select the **blogs-data.txt** file.
5. Click **Edit workbook**  .
6. From the **Select a reader** drop-down, select **JSON Array** reader.
7. Click **Set Reader**  .
8. Now with the data formatted properly, scroll down (if you have to) and click **Save workbook**  .
9. You do not need all the columns in your Big SQL table, so you will remove some now. First you need to create a child workbook.
10. Click **Build new workbook**.
11. Rename the workbook by clicking **Edit workbook name** and then type **BlogsDataRevised**.
12. Remove multiple columns by following these steps:
    a. Click the down arrow in any column heading and select **Organize Columns**
    b. Click the **X** next to the following columns to mark them for removal
        i. Crawled
        ii. Inserted
        iii. IsAdult
        iv. PostSize
    c. Click **Apply Settings** to remove the marked columns
13. Click **Save** to save the workbook.
14. Click **Exit**, and then click **Run** to run the workbook.

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

© Copyright IBM Corp. 2012, 2015

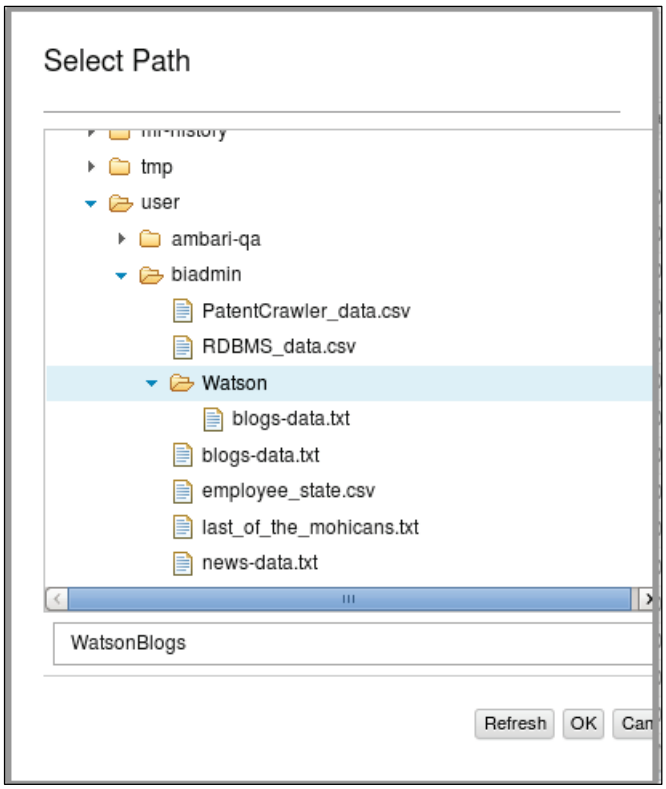
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

Task 4. Exporting the BigSheets blog data workbook to a TSV file.

You can export your BigSheets workbook to a file. Then, use that file to analyze the data in Big SQL. This capability allows you to not only use BigSheets data with Big SQL, but also to any number of systems. You can export the data into a number of different formats.

When run has reached 100%, proceed with the following steps.

1. In the menu bar of the **BlogsDataRevised** workbook, click **Export data**.
2. In the drop-down window, select **TSV** in the **Format Type** field.
3. In the **Export to** radio buttons, select **File** as the export target.
4. Click **Browse** to select a destination directory in the DFS.
5. Select your path as **/user/biadmin/Watson**.
6. Type **WatsonBlogs** as the name of the file.



7. Click **OK**.
8. Ensure that the **Include Headers** check box is not checked and then click **OK**.
9. Click **OK** to close the message dialog.

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

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

10. Make a note of the column names and the type of data from the BigSheets workbook that you want to define in Big SQL. You exported these columns from BigSheets:

    Country - contains a two-letter country identifier.

    FeedInfo - contains information from web feeds, with varying lengths.

    Language - contains the string that identifies the language of the feed.

    Published - contains a date and time stamp.

    SubjectHtml - contains a subject that is of varying length.

    Tags - contains a string of varying length that provides categories.

    Type - contains the source of the web feed, whether a news blog or a public feed.

    URL - contains the web address of the feed, with varying length.

Task 5. Creating a Big SQL script that creates Big SQL tables from the exported TSV file.

In this section, you create an SQL script to create Big SQL queries based on the BigSheets blogs data workbook.

1. In the Linux command line, create a SQL script named **NewsBlogs.sql**, type or paste the following code:

   cat > /home/biadmin/labfiles/bigsheets/NewsBlogs.sql

```
CREATE SCHEMA IF NOT EXISTS BigSheetsAnalysis;
USE BigSheetsAnalysis;

CREATE HADOOP TABLE BigSheetsAnalysis.sheetsOut
 (country VARCHAR(2), FeedInfo VARCHAR(300),
  language VARCHAR(25), published VARCHAR(25),
  subject VARCHAR(300), tags VARCHAR(100),
  type VARCHAR(20), url VARCHAR(100))
  ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t';

LOAD HADOOP USING FILE URL
 '/user/biadmin/Watson/WatsonBlogs.tsv'
  with SOURCE PROPERTIES ('field.delimiter'='\t')
INTO TABLE BigSheetsAnalysis.sheetsOut OVERWRITE;

SELECT * FROM BigSheetsAnalysis.sheetsOut;
```

3. Click **CTRL-D** to save and exit out of the file.

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

© Copyright IBM Corp. 2012, 2015

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

4. View the contents of the file to ensure you created it properly by typing the following:

   ```
   cat /home/biadmin/labfiles/bigsheets/NewsBlogs.sql
   ```

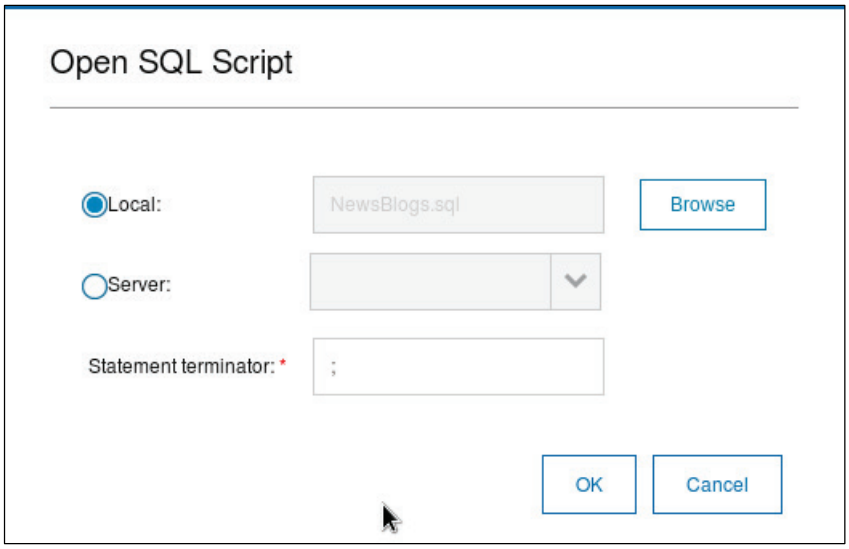
5. Go to the BigInsights - Home page by using the browser bookmark or type in this URL:

   https://ibmclass.localdomain:8443/gateway/default/BigInsightsWeb/index.html

6. Click the **Big SQL** link to open up the Big SQL UI.

   Note: If Big SQL is unavailable, you need to run the dsmKnoxSetup script as indicated in Task 2 of this lab exercise.

7. Click the **SQL Editor** link from the left side.
8. Click the **Open** link.
9. With Local selected, click the **Browse** button and navigate to **/home/biadmin/labfiles/bigsheets/NewsBlogs.sql**.



10. Click **OK** to open the script.

    The script you created earlier should now be inside the editor.

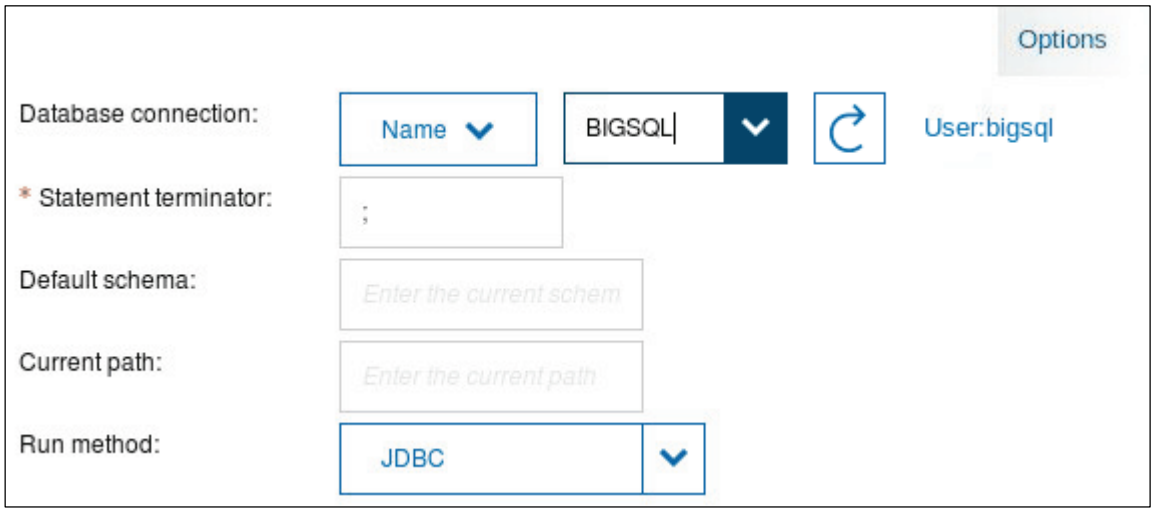
11. Click the **Options** link (far right of the window).

    This will expand the pane with more options.

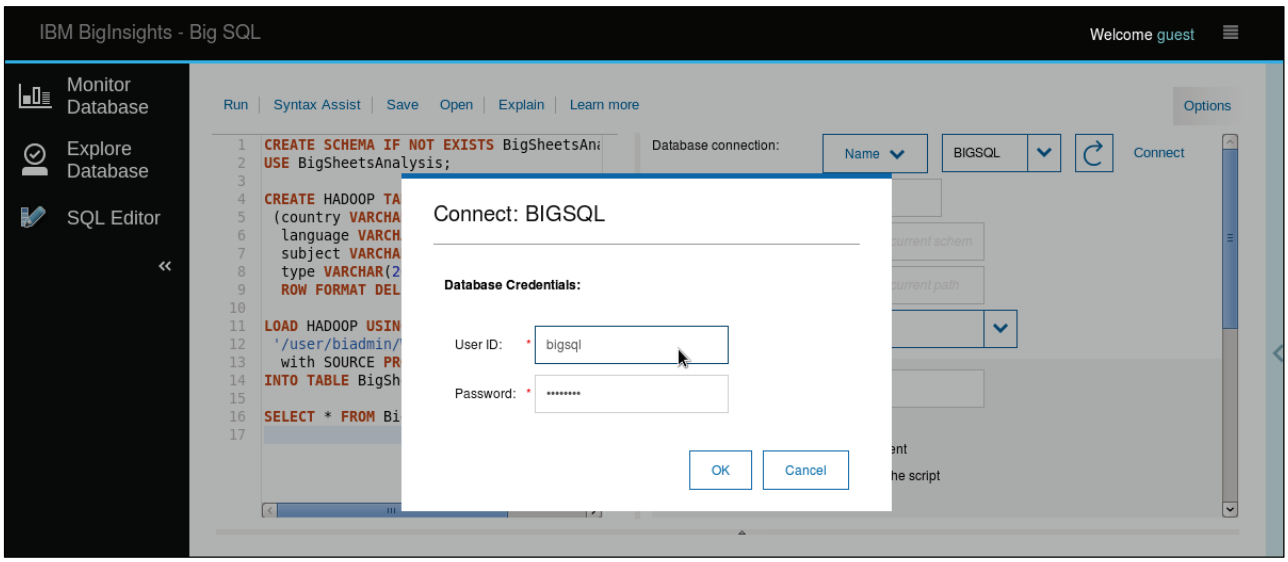
This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

12. Under **Database connection**, select **Name**, and then select **BIGSQL** from the dropdown.



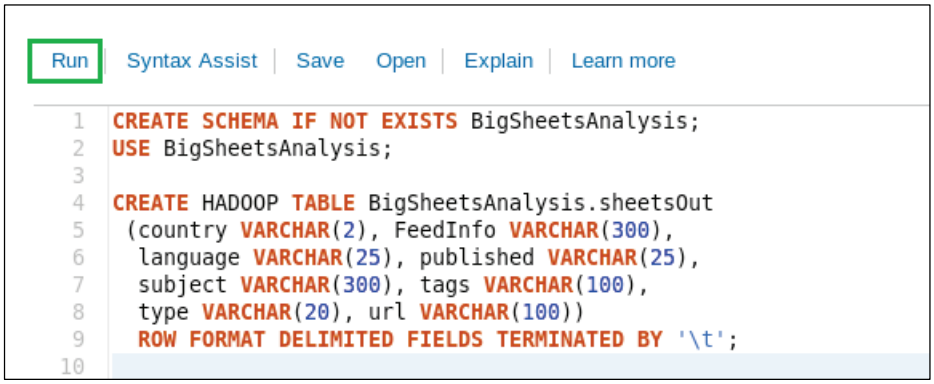
13. Click **Connect**.
14. Provide the login credentials **bigsql/ibm2blue**, and then click **OK**.



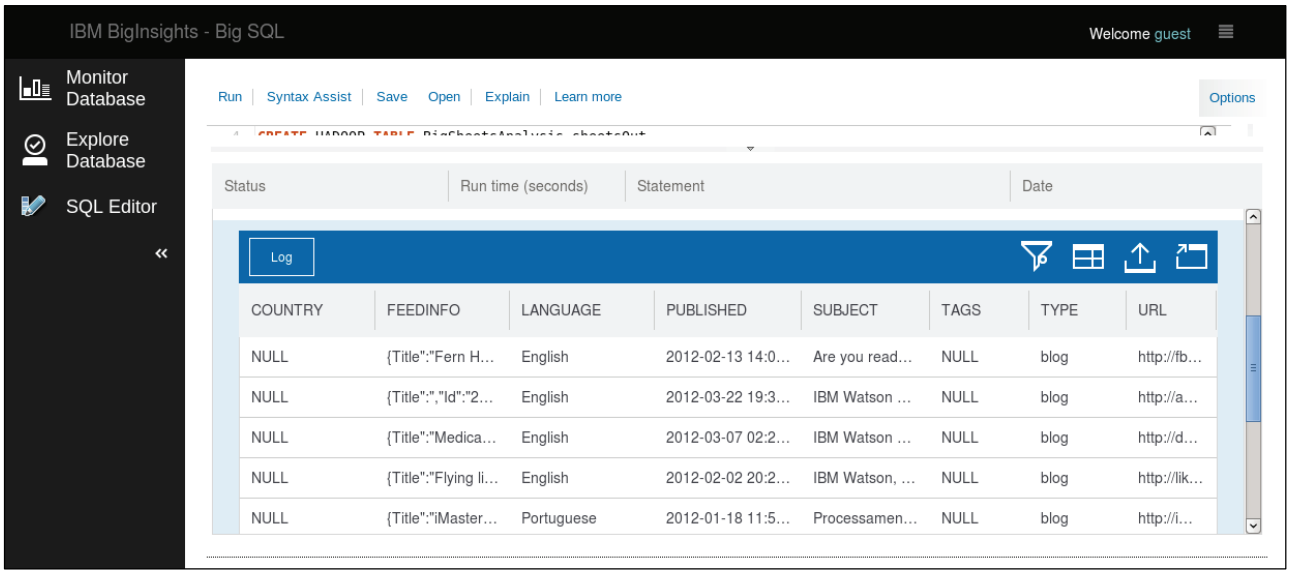
This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

15. Click the **Run** link to run the query.



The results are shown below:



Now that the results of the workbook are imported into Big SQL, you can perform analytical queries. That is beyond the scope of this lab. Refer to the Big SQL course for more information.

Task 6. Creating a Big SQL table using built-in integration.

You saw in the previous section how to export BigSheets workbook as a TSV file. In fact, you can export out in a number of different file formats to be used with any number of database systems, such as Big SQL. However, there is an added bonus if you choose to use Big SQL. I'm sure you have heard of the "easy" button, and that's exactly what you have here.

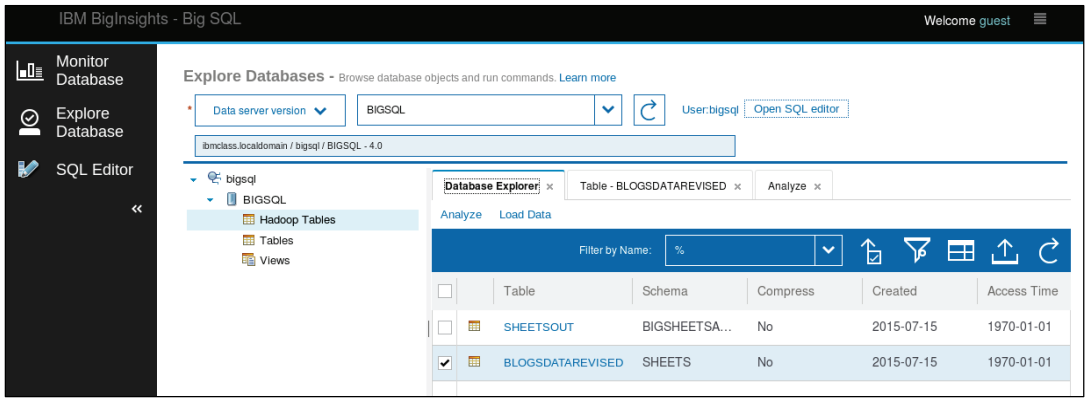
1. In the **BlogsDataRevised** workbook, click **Create Table** and keep the default schema and table name (*sheets.BlogsDataRevised*).
2. Click **Confirm.**

   Notice that the button changed its label to Delete Table. This means that you can only have one Big SQL table for a workbook.

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

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

3. Go to the **Big SQL** page and click the **Explore Database** link.
4. Connect to the **BIGSQL** database using your login credentials.
5. Navigate through the Hadoop Tables view to find your newly created table (from the click of a button).



Again, now you can work with this table as if it were any database table. You can run queries against it to find out more insight from the data.

Task 7. Troubleshooting (Optional).

Big SQL has some limitations running on a single node cluster (actually, anything less than a 3 node cluster) and for good reasons too. You do not ever want to have a single node cluster for your production environment.

In the training world, simpler is better. Your lab environment should have been configured with this fix to allow to create Big SQL tables, but if it wasn't, you can run this yourself:

1. First, open up a new terminal.
2. Switch to the **bigsql** user.
3. Run this command to connect to the bigsql database:

```
db2 connect to bigsql
Database Connection Information
Database server        = DB2/LINUXX8664 10.6.3
SQL authorization ID   = BIGSQL
Local database alias   = BIGSQL
```

4. Run this command for the fix:

```
db2set DB2_DYNAMIC_PMAP=INCLUDE_HEAD_NODE
```

5. Restart the **Big SQL** service from **Ambari**.

   This will allow you to create Big SQL tables (if you were not able to before).

Purpose:
You used Big SQL tables with BigSheets.

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

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Unit summary

- Manage the data lifecycle through Big SQL
  - Create new tables using the same data as the sheet
  - Create new sheets using the same data as the table

© Copyright IBM Corporation 2015

Unit summary

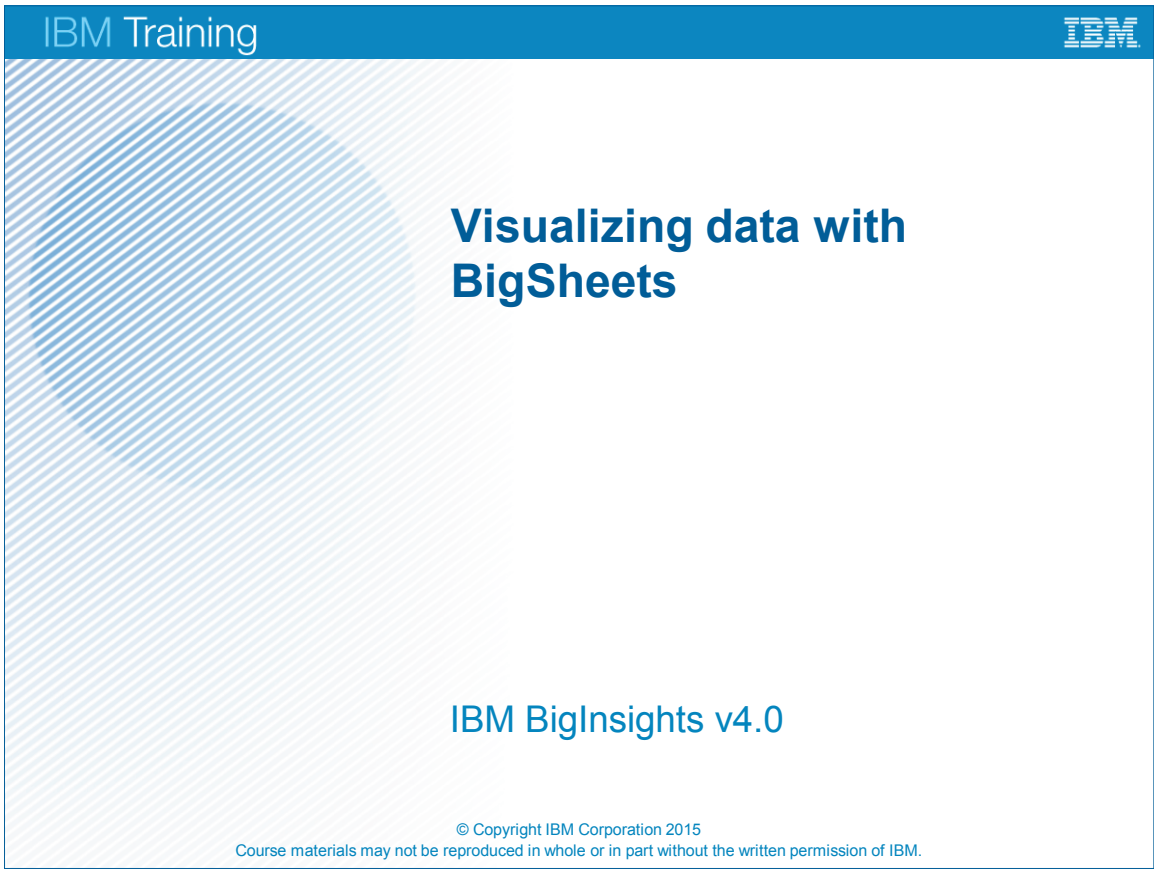
This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

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

© Copyright IBM Corp. 2012, 2015

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.



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

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

© Copyright IBM Corp. 2012, 2015

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Unit objectives

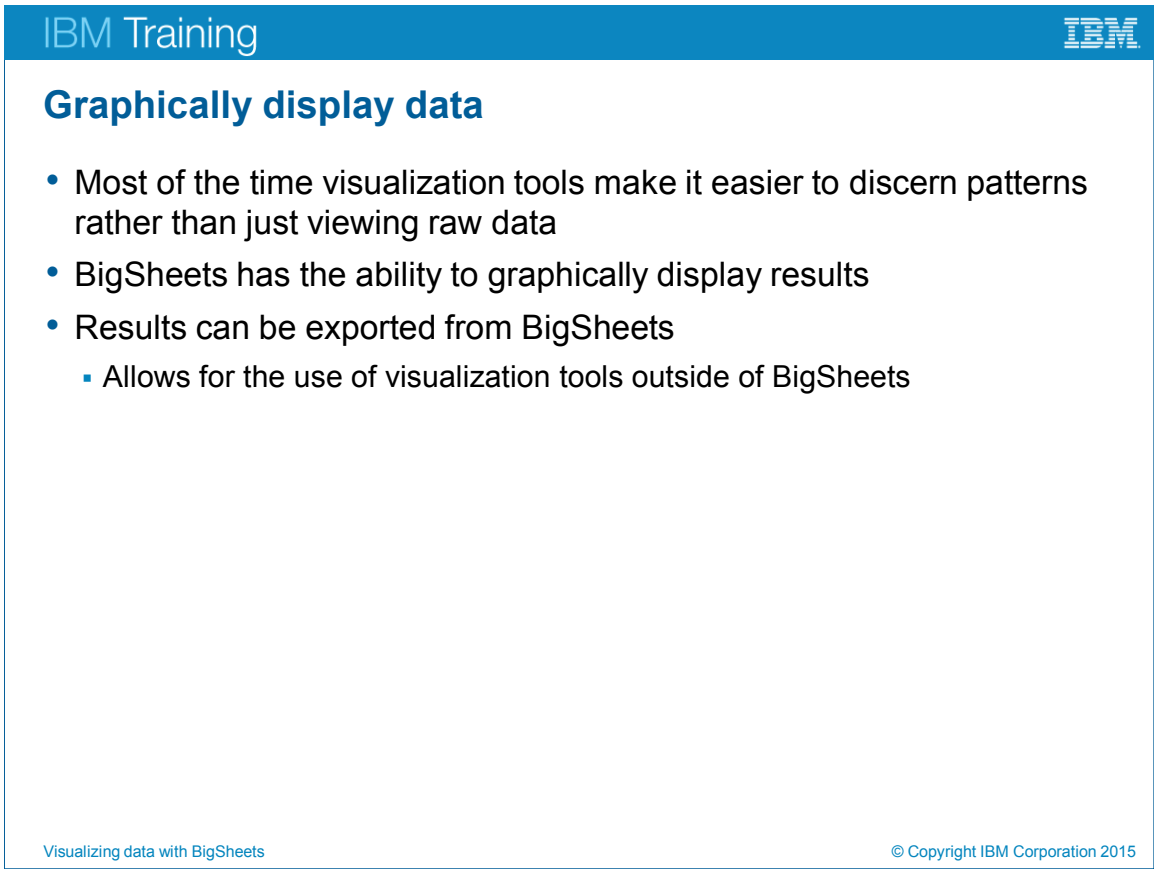
- Describe the visualization techniques supported by BigSheets
- Explain how to export data from a collection

© Copyright IBM Corporation 2015

Unit objectives

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

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.



Graphically display data

For many, just looking a column (or columns) of values does not facilitate discerning patterns in the data. Pictures tend to bring patterns into focus. Because of this, BigSheets has the ability of display data graphically. There is also the capability to export the results of a collection into a variety of data formats. This allows for the use of additional visualization tools not supplied by BigSheets.

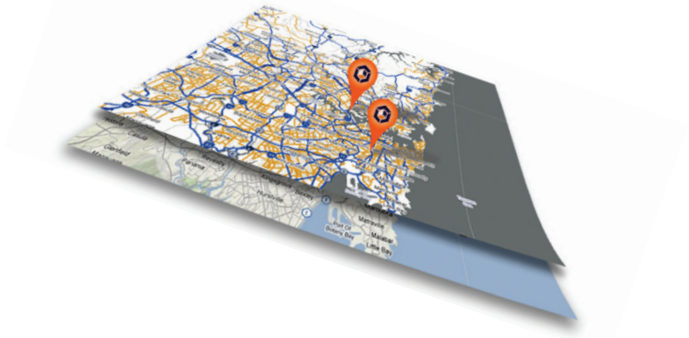
This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

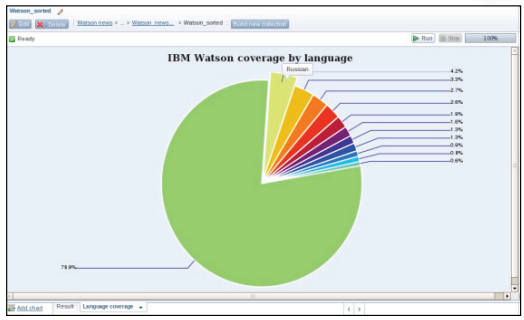
© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Chart types: categories

- Map
  - Geo Map
  - Heat Map
  - Map
- Cloud
  - Bubble Cloud
  - Text Cloud
- Chart
  - Area
  - Bar
  - Horizontal Bar
  - Line
  - Pie





Visualizing data with BigSheets

© Copyright IBM Corporation 2015

Chart types: categories

There are three types of chart categories to visualize your workbook data.

1) Map
2) Cloud
3) Chart

Geospatial analytics adds another meaningful dimension to analysis. This is created using the Geo Map. For example, it enables companies to analyze customer movement through a space, help policemen to see patterns of crime locations, and help for municipalities to understand where people most often request taxicabs or other services.

Heat Map - Charts the locations and concentrations of data points, and then shows the magnitude of those points in relation to other points overlaid on either a map of the world or a map of the United States.

Map - Counts and charts data points by name, and then shows those data values overlaid on either a map of the world or a map of the United States.

Map - Counts and charts data points by name, and then shows those data values overlaid on either a map of the world or a map of the United States.

Bubble Cloud - Shows the value, the relative size of the bubble reflect its value.

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

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

Tag Cloud - Shows the value; the relative size of a word reflects its value.

Area - Shows a trend in data over time by connecting a series of points that represent individual measurements.

Pie - Shows proportionate relationships; the relative size of the "slice" reflects the proportion of the data.

Bar - Shows proportionate relationships; the relative length of a rectangular bar reflects the proportion of the data. The bars are displayed vertically.

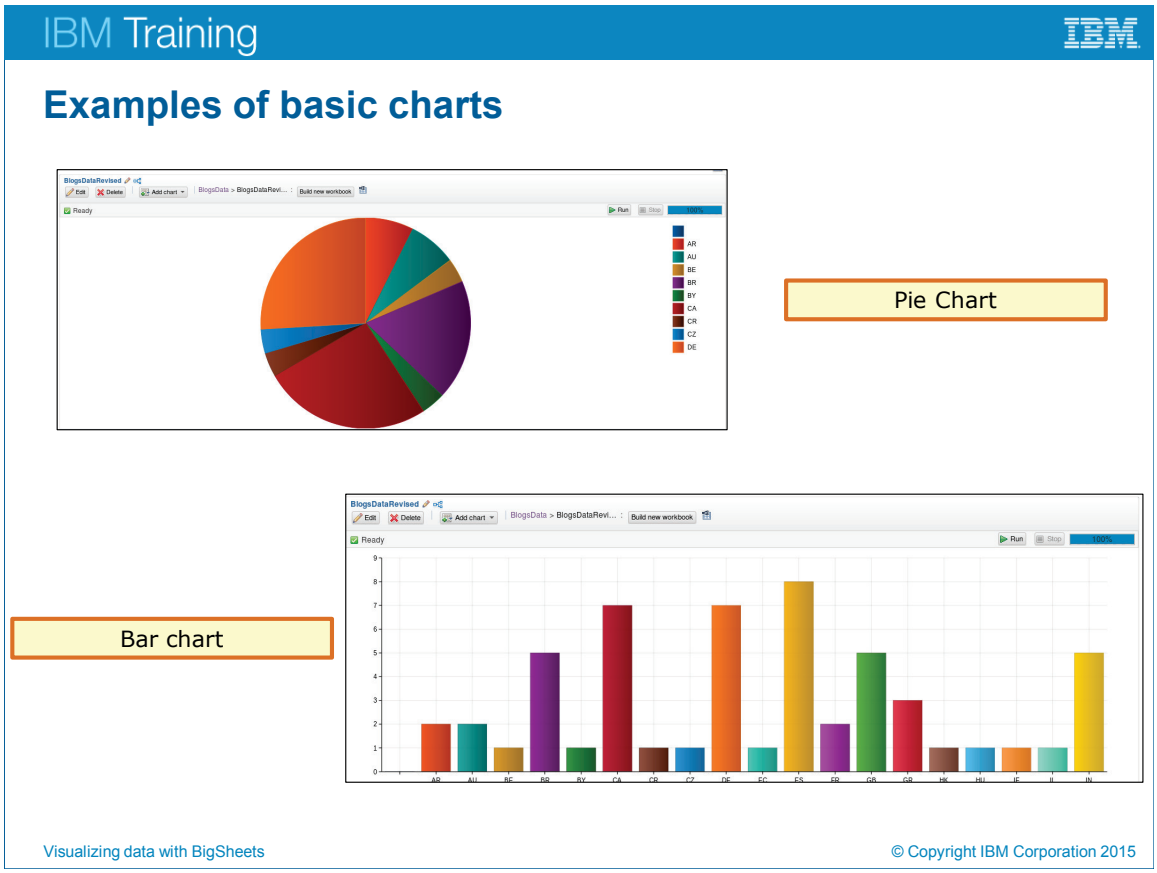
Line - Shows a trend in data over time by connecting a series of points that represent individual measurements.

Horizontal Bar - Shows proportionate relationships; the relative length of a rectangular bar reflects the proportion of the data. The bars are displayed horizontally.

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

© Copyright IBM Corp. 2012, 2015

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

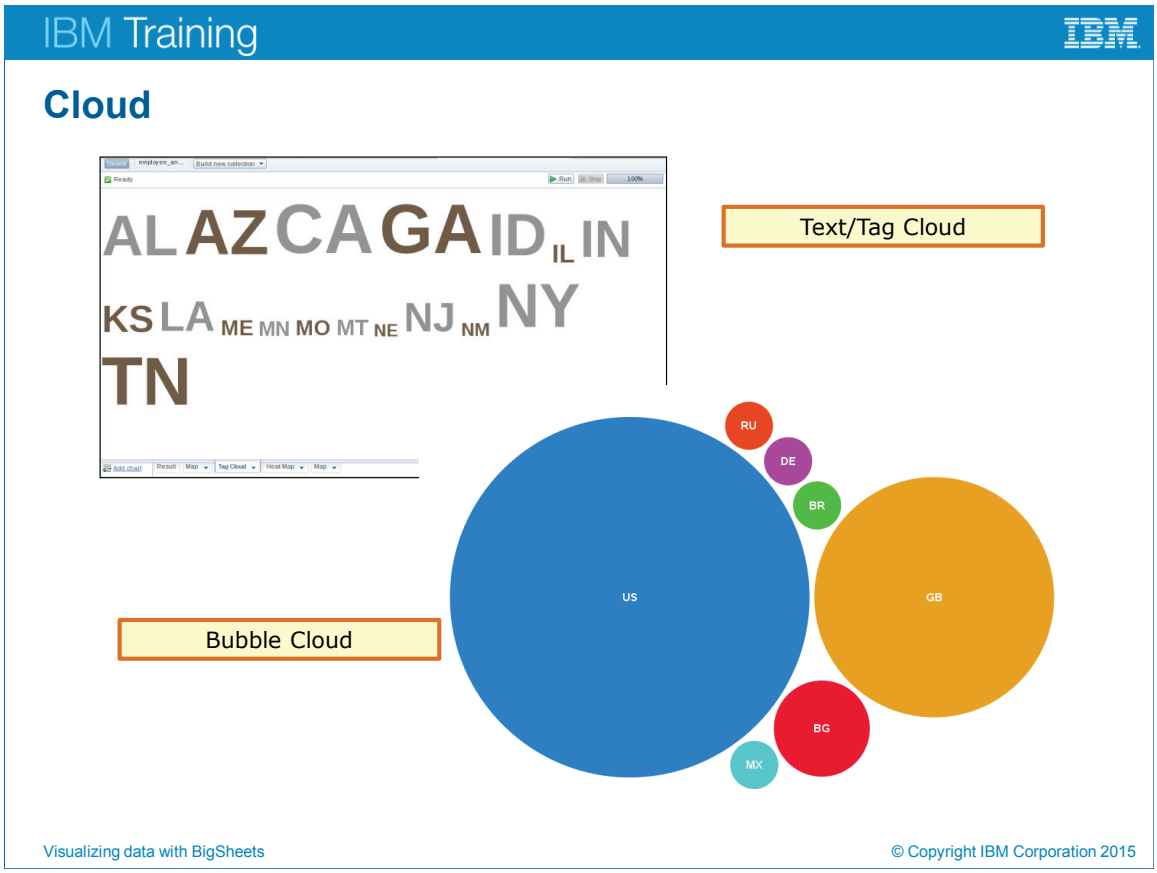


Examples of basic charts

Here are some examples of basic chart types.

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

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.



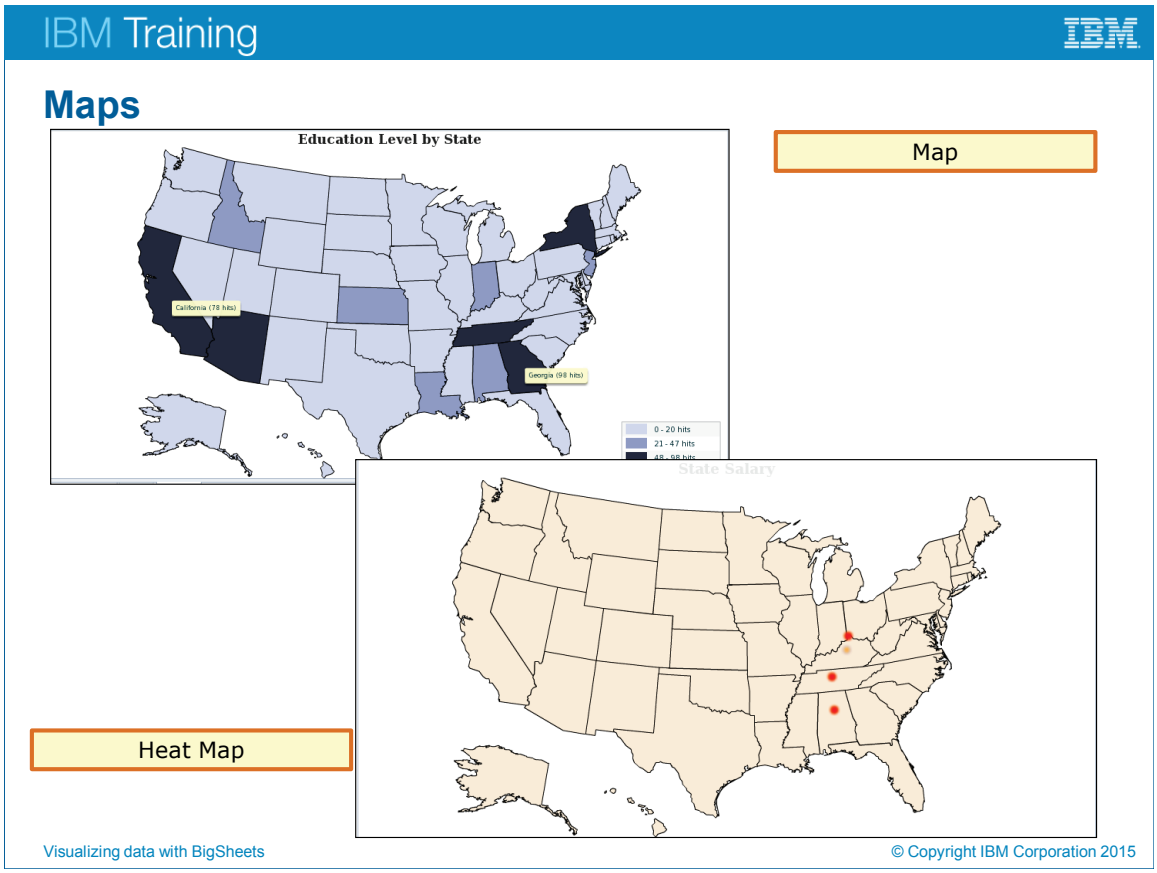
Cloud

This is an example of a Tag Cloud chart. The difference in font sizes has particular mean, that is, the larger the font the more numerous the values. In this case the graph is dealing with occurrences of records that have the same state value. States like Arizona, California, Georgia, and Tennessee have a greater number of references in the data when compared to Maine, Montana, New Mexico, and Nebraska.

The example of the Bubble Cloud shows the magnitude of the value. The size of the bubble is proportional to the magnitude of the data within the worksheet.

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

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.



Maps

The final way to visualize data is through maps. There are two types. The *map type* counts and charts data points by name, and then shows those data values overlaid on either a map of the world or a map of the United States.

The second type is a *heat map*. This type charts the locations and concentrations of data points, and then shows the magnitude of those points in relation to other points overlaid on either a map of the world or a map of the United States.

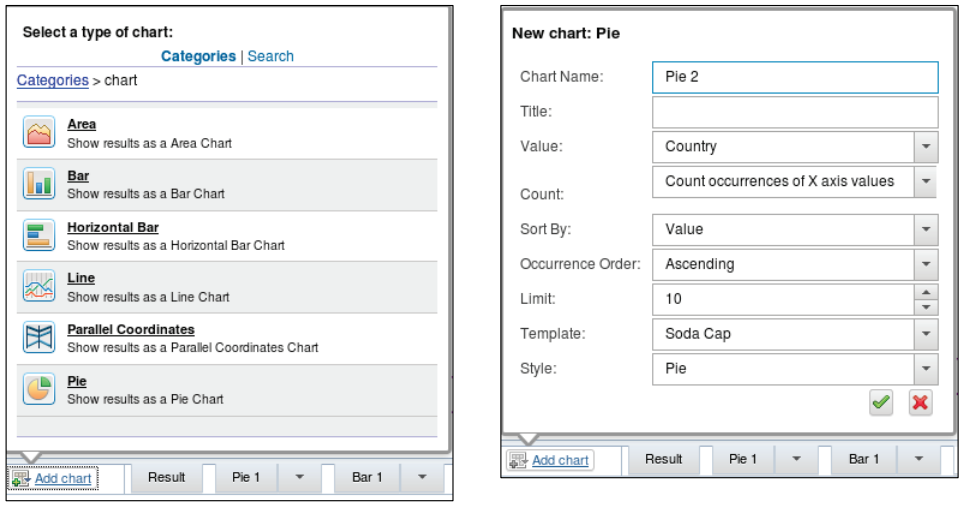
Heat maps require 3 primary values: valid latitude and longitude as well as a value that you want to represent at those points. Latitude and longitude columns that you select from the collection must be represented in decimal point expressions. Longitude values west of the prime meridian are designated as negative values. Likewise latitude values south of the equator are designated as negative values.

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

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Creating a chart



Visualizing data with BigSheets

© Copyright IBM Corporation 2015

Creating a chart

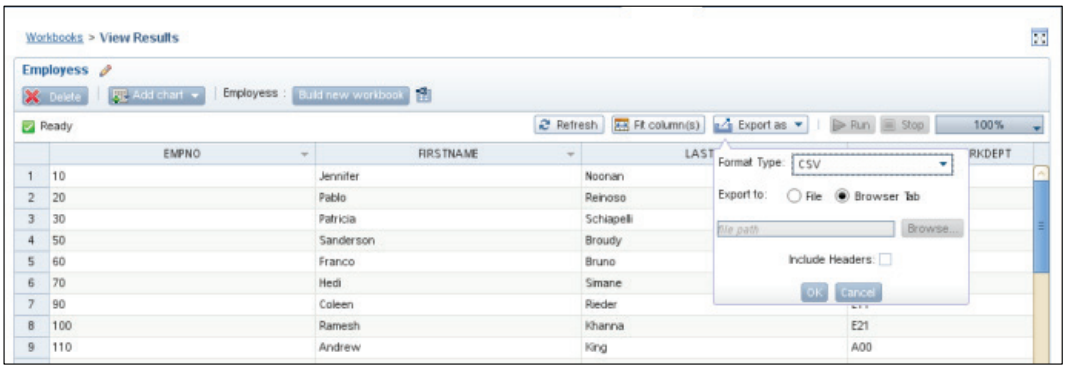
To create a chart, select the *Add chart* hyperlink for a workbook. Even master workbooks can have charts. Choose the type of chart and then specify the appropriate input parameters.

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

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Exporting data



- Select Export as
  - Choose the data format
- Specify whether to export to a file or a browser window
- Select whether to include headers

Visualizing data with BigSheets

© Copyright IBM Corporation 2015

Exporting data

Data in a BigSheets workbook can be exported into a variety of formats. The export process allows you to choose as the destination either a file or a web browser. If a browser is chosen, a browser window is opened and by default, 500 rows are exported. You can change how many rows to export to a browser by specifying a new number for &amount=*500* in the URL and refreshing the page. You can then copy the contents of the browser window to another location to save it.

You saw in the previous unit that you can use this method to export your data and import it into Big SQL. A different type of database system can also be used.

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

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Checkpoint

1. List three categories of charts supported by BigSheets.
2. List two types of clouds.
3. True or False? Master workbook can have charts.

© Copyright IBM Corporation 2015

Checkpoint

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

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Checkpoint solutions

1. The types of charts:
   1. Map
   2. Cloud
   3. Chart (basic)
2. Bubble cloud and tag cloud.
3. True. Master workbook can have charts.

© Copyright IBM Corporation 2015

Checkpoint solutions

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

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

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

© Copyright IBM Corporation 2015

Demonstration 1: Analyzing Social Media and Structured Data

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

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

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
```

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

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

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**.

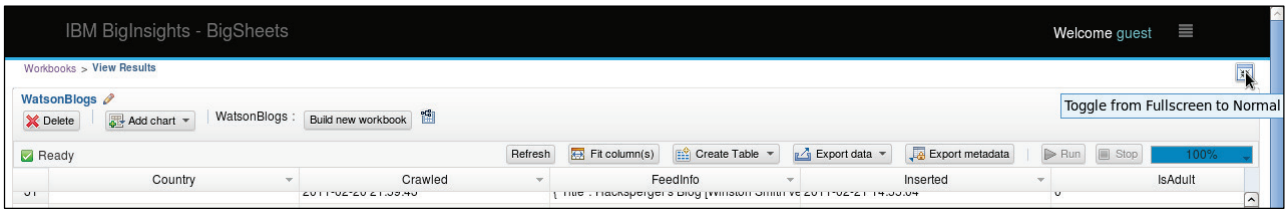
This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2012, 2015

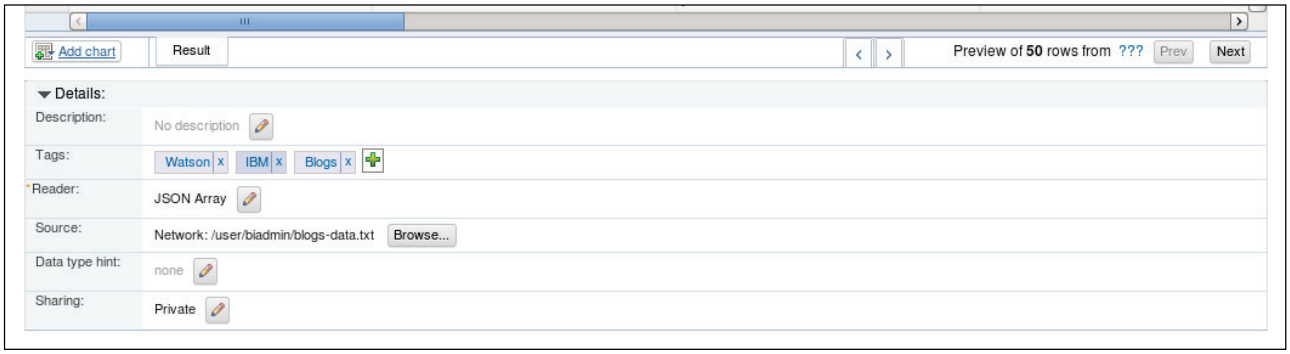
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

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.

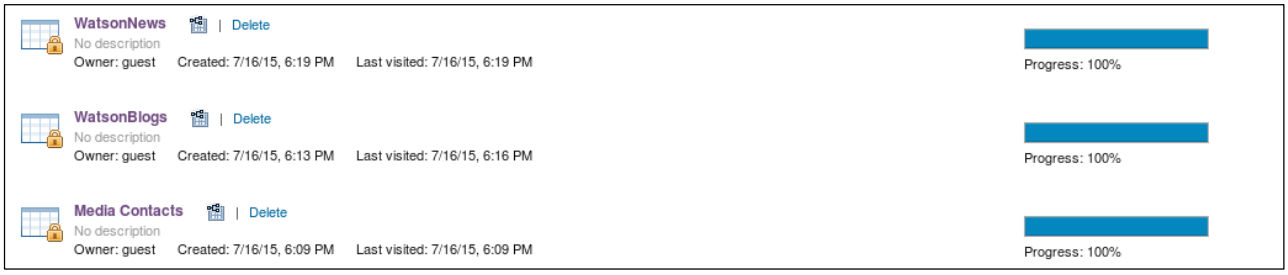
11. Create a new workbook called **WatsonNews**.
12. Add **Watson**, **IBM**, and **News** as tags to this new workbook.

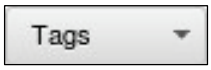
This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

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

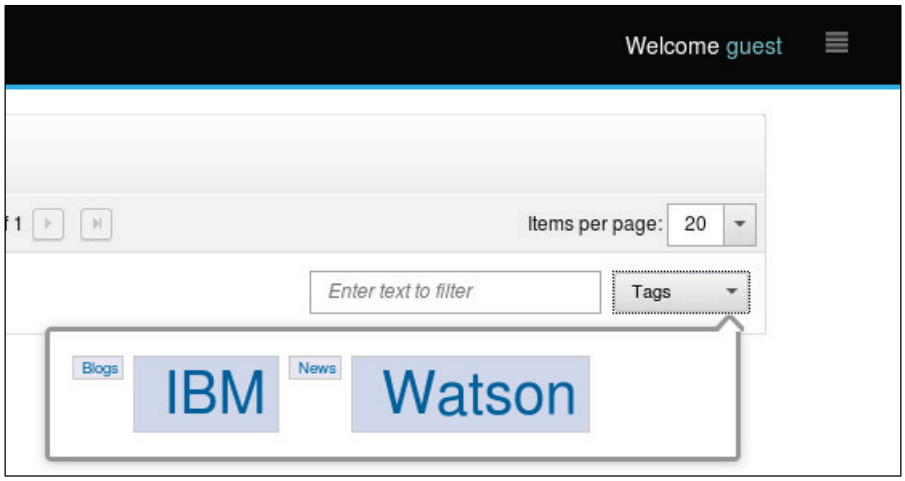
    You should see the three workbooks you created in this demonstration as well as others from previous demonstrations.



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.

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

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

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.

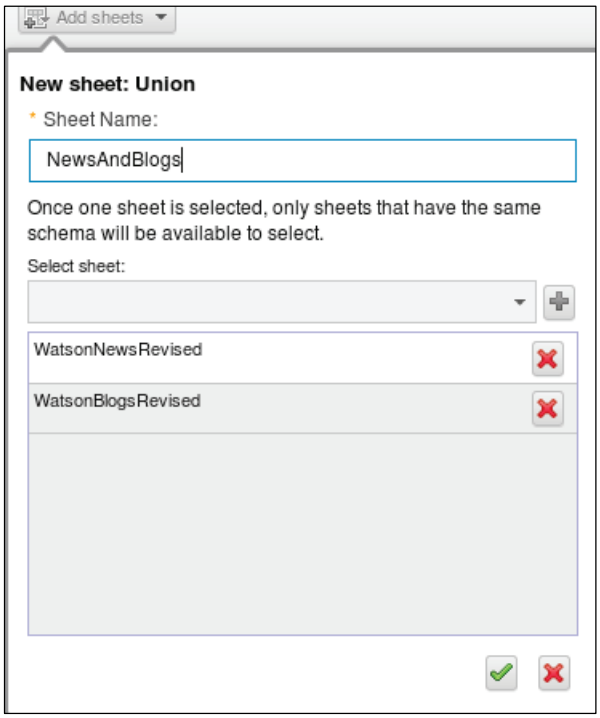
1. Click the **Workbooks** breadcrumb tab and select the **WatsonNewsRevised** workbook.
2. Click **Build new workbook**.
3. Click **Add sheets**.
4. Click **Load** and then click **WatsonBlogsRevised**.
5. Change the sheet name to **WatsonBlogsRevised** and then click **Apply settings**.
   Now the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

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

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

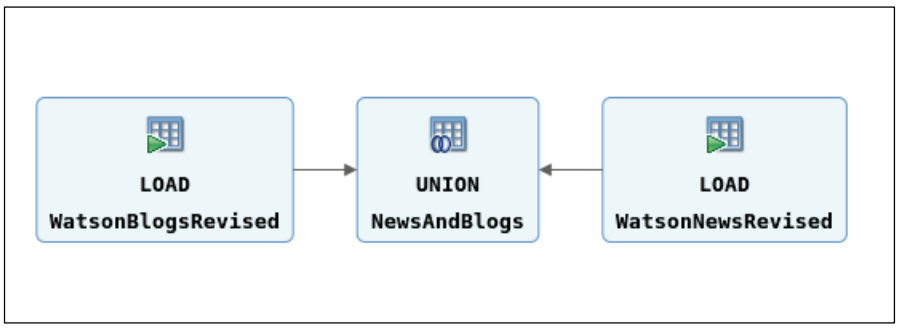
6. Click **Add sheets** and then select **Union**.
7. Change the Sheet Name to **NewsAndBlogs**.
8. In the **Select Sheet** drop down, select **WatsonNewsRevised**, and then click **Add sheet**.
9. Select the **WatsonBlogsRevised**, and then select **Apply settings**.

   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.

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

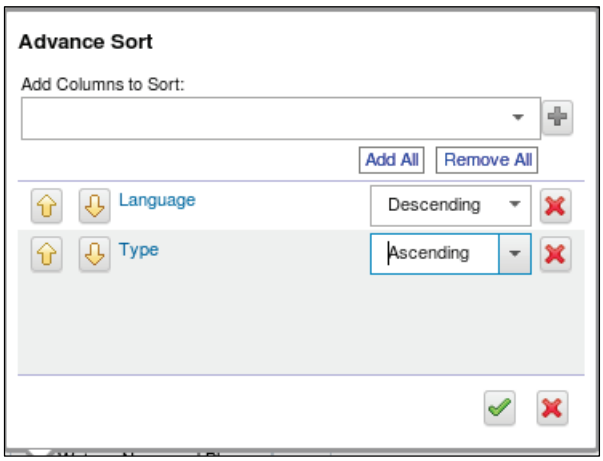
© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

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**.

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

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

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**.

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

© Copyright IBM Corp. 2012, 2015

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

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**.

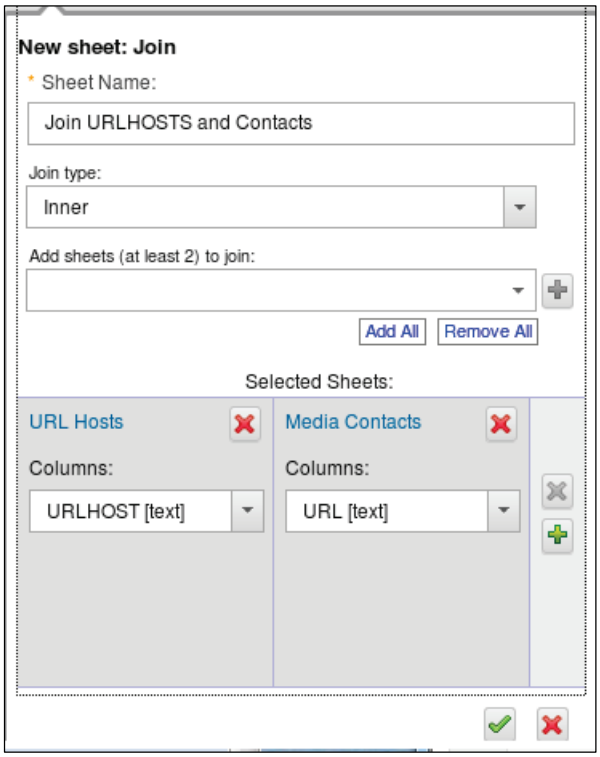
This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© Copyright IBM Corp. 2012, 2015

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

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**.

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

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

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.

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

© Copyright IBM Corp. 2012, 2015

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

IBM Training

Unit summary

- Describe the visualization techniques supported by BigSheets
- Explain how to export data from a collection

© Copyright IBM Corporation 2015

Unit summary

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

© Copyright IBM Corp. 2012, 2015
Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

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

IBM Training

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

© Copyright IBM Corporation 2015. All Rights Reserved.