# Data Discovery with BigSheets


© 2013 IBM Corporation



## Agenda

- **Overview of BigSheets**
  - What is BigSheets ?
  - It's uses ?
  - Common Scenario
- **BigSheets Processing**
  - Terminology
  - Architecture
  - Processing
- **Working with BigSheets**
  - Accessing BigSheets
  - Create & customize workbooks
  - Visualize results
- **Summary**


© 2013 IBM Corporation

# What is BigSheets?

- **Browser-based analytics tool for business users**

## Why BigSheets?

- Business users need a non-technical approach for analyzing Big Data
- Translating untapped data into actionable business insights is a common requirement
- Visualizing and drilling down into enterprise and Web data promotes new business intelligence

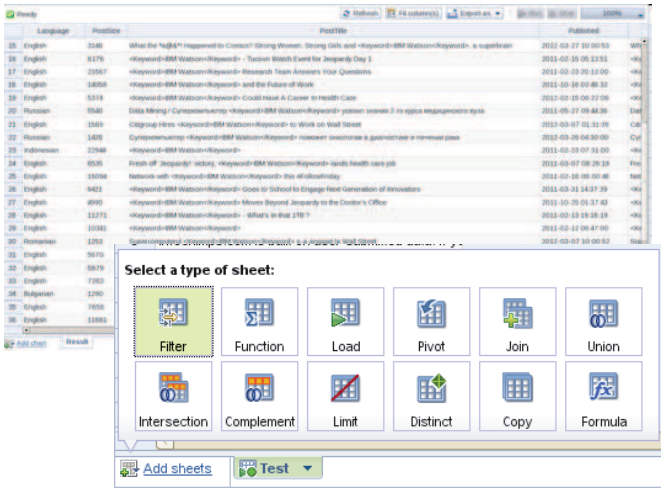
## How can BigSheets help?

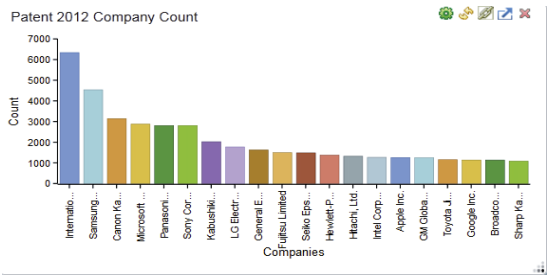
- Spreadsheet-like interface enables business users to gather and analyze data easily
- Built-in “readers” can work with data in several common formats (JSON arrays, CSV, TSV, Web crawler output, . . . )
- Users can combine and explore various types of data to identify “hidden” insights



# What you can 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)



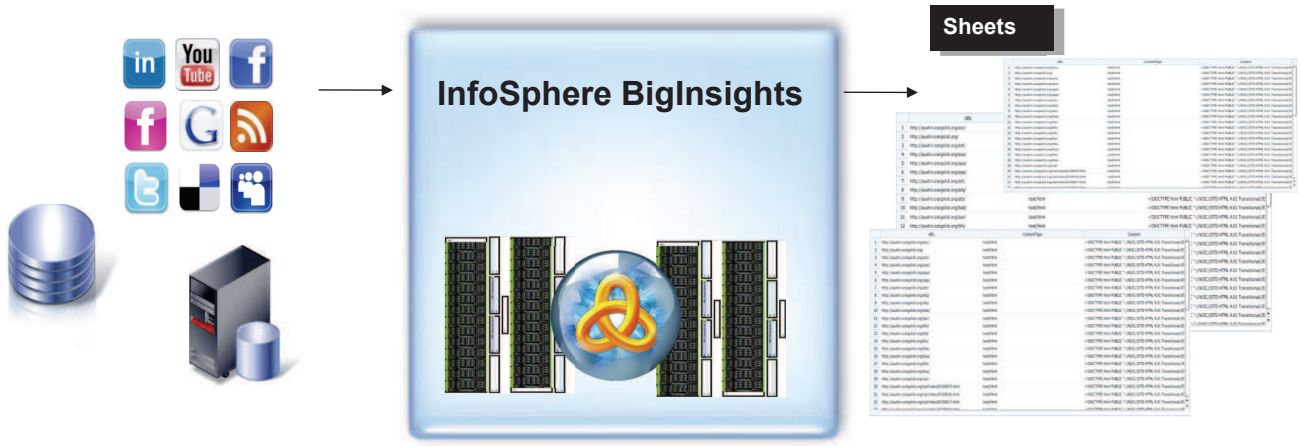




*No programming knowledge needed!*


© 2013 IBM Corporation

# BigSheets Scenario

- **Data Collection**
  - **WebCrawler app**
  - **DBMS import app**
  - **BoardReader app**
  - Flume
- **Data Storage**
  - Distributed file system
  - **Web based file browser and administration**
- **Data exploration**
  manipulation and analysis
  - **BigSheets**



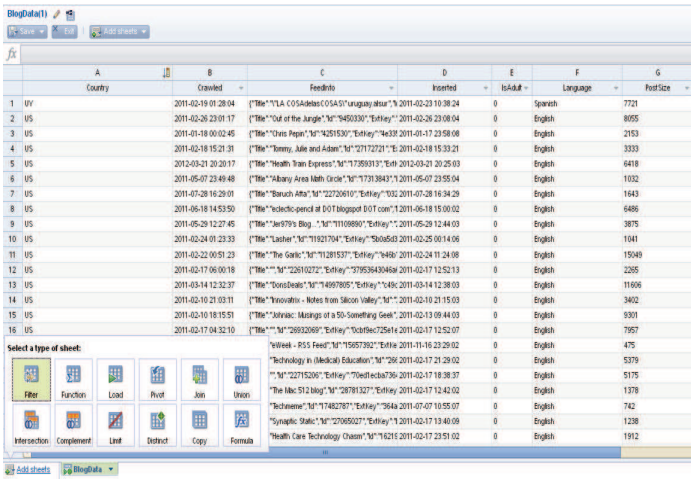


© 2013 IBM Corporation



# BigSheets Terminology - Workbook

- **What is a Workbook?**
  - Representation of data within the tool
  - Simple Spreadsheet-like structure
  - Created and managed by user
  - Name, Description, Tags
- **How to create a Workbook?**
  - Import new data
  - Build from other Workbooks
- **Master, Parent, Sibling, Child**
  - Related Workbooks that are associated with the current workbook
- **Other key terms:**
  - Formula
  - Function
  - Sheet

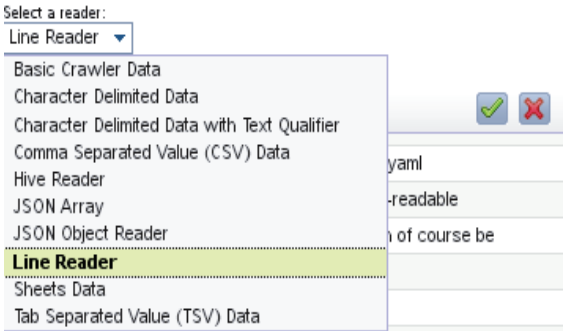



© 2013 IBM Corporation

# BigSheets Terminology - Readers

- **BigSheets Reader – Built-in "readers" know how to apply a "schema" to the data at run time**

- **Built-in Readers**
  - Basic Crawler Data
  - Character Delimited Data
  - Comma Separated Value (CSV) Data
  - Tab Separated Value (TSV) Data
  - Hive Reader
  - JSON Array and Object Reader
  - Line Reader
  - Sheets Data





© 2013 IBM Corporation



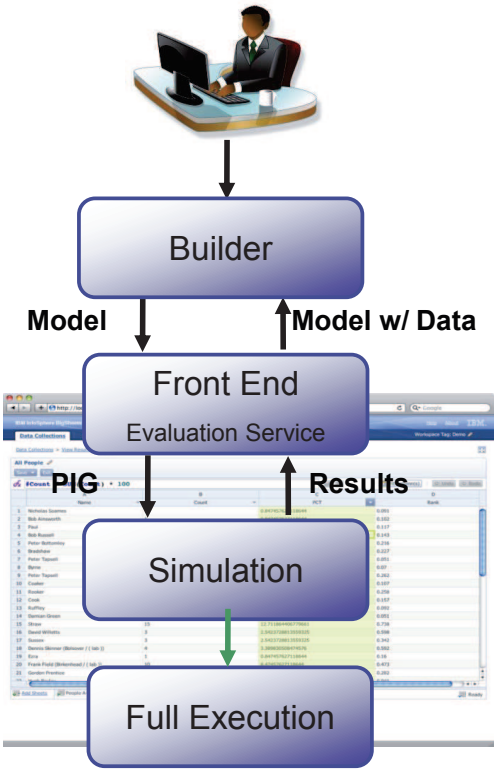
# BigSheets Processing

- 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 workbooks to compute results on the real data and explores the output





© 2013 IBM Corporation

# Working with BigSheets

- **Create workbook to model target data – Directly from an application or by using the Build new Workbook button.**
- **Customize workbook through graphical editor and built-in functions**
  - Filter data
  - Manipulate data
  - Join data from multiple workbooks
- **"Run" workbook: apply work to full data set**
- **Explore results in spreadsheet format and/or create diagrams**
- **Optionally, export your data**

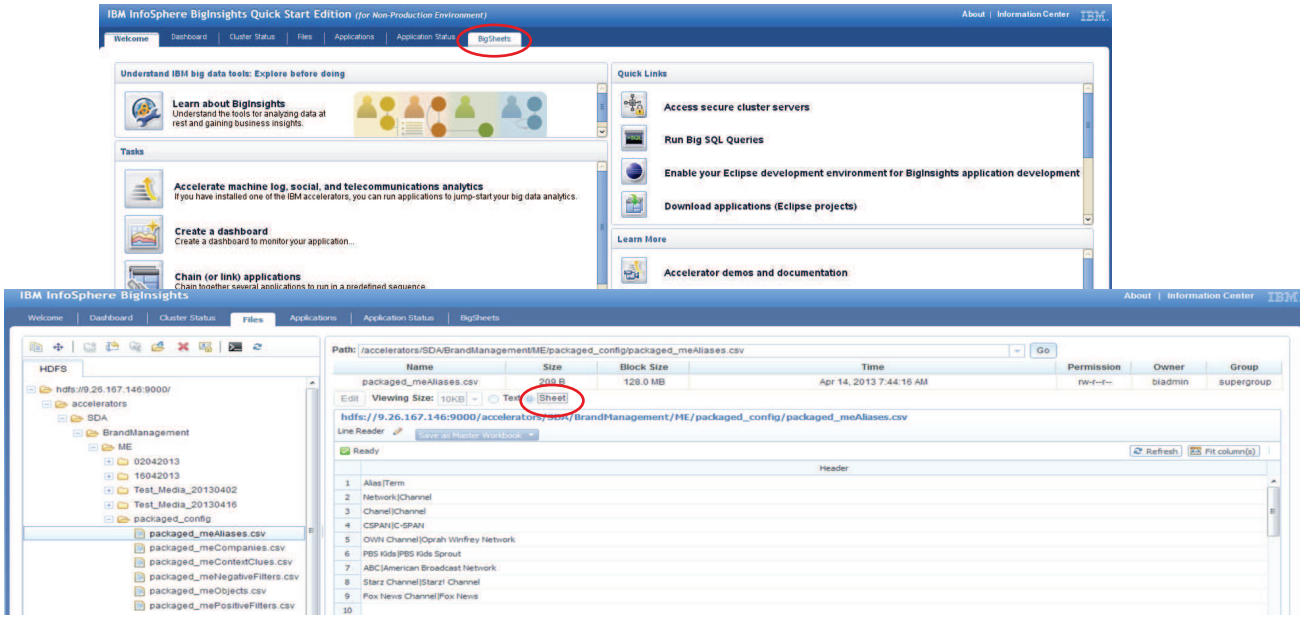


© 2013 IBM Corporation



# Accessing BigSheets

- **Ensure BigInsights Enterprise is running**
- **Launch the Web console with URL http://<host>:<port> or http://<host>:<port>/data/html/index.html**

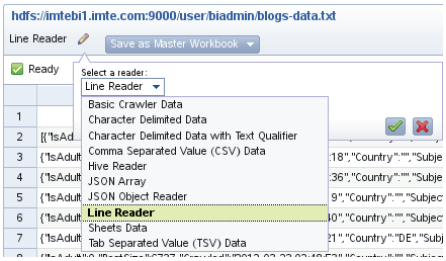


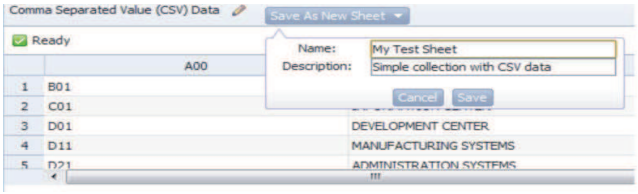


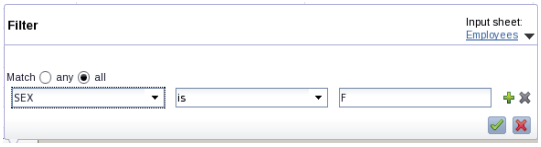
© 2013 IBM Corporation

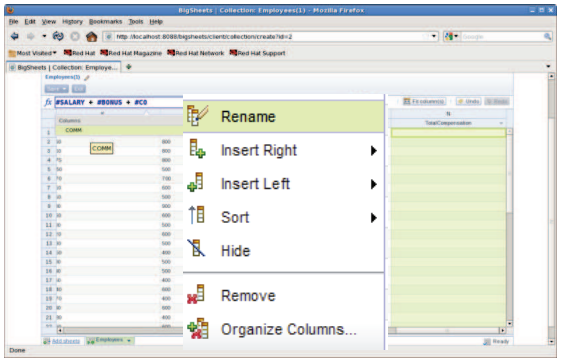
# Creating and Customizing a Workbook- Example

- **Locate data of interest**
- **Switch viewing preference to Sheets**
- **Specify a Sheet "reader"**
- **Save as a new Sheet**
- **Work with built-in "sheets" editor**
  - Add / delete columns
  - Filter data
  - Specify formulas to compute new values
  - Apply built-in or custom functions



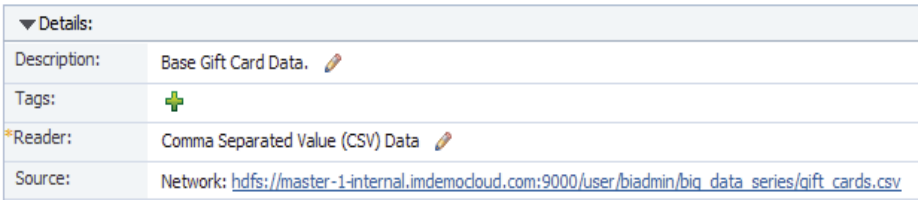


© 2013 IBM Corporation

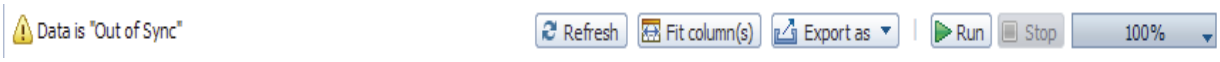


# BigSheets – Data Refresh Steps

- **Master (or base) workbooks can point to a file…**




- **…or a directory (as long as all files are the same format).**
- **To change the file contents of the Master workbook re-run the Master workbook.**
- **This will mark all dependent sheets as "Out of Sync"**

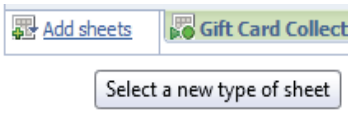



- **Manually re-run all sheets**
- **To change the source for a master workbook, the new source must have the same schema**

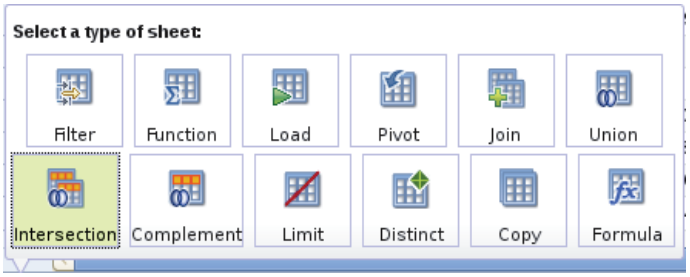

© 2013 IBM Corporation

# BigSheets – Subsequent Data Processing Options

- **Adding a "sheet" adds a tab's worth of processing**
- **Clicking the "Add sheets" link provides 12 options:**




- **Filter, Function, Load, Pivot, Join, Union, Intersection, Complement, Limit, Distinct, Copy, and Formula**




- **Details on the next slides…**

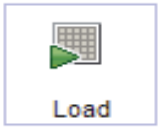

© 2013 IBM Corporation



# BigSheets – Filter, Macro, Load and Join


- **Filter** – Removes data that does not match some specified criteria


- **Function** – Executes a function against each row in a workbook. A function takes a row of data as input and produces one or more rows of data as output. (96, built-in functions)

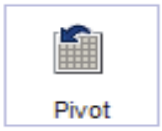



- **Load** – Brings the data of another workbook as a sheet. Load is useful if you want to perform a join or union operation.


- **Join**– Joins sheets, meaning that any data you want to combine needs to be a sheet in the spreadsheet.


© 2013 IBM Corporation

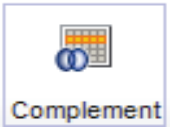
# BigSheets – Pivot, Union, Intersection, Complement




- **Pivot** – Calculates values by grouping the data in the workbook, applying functions to each group, and carrying over data.


- **Union** – Appends sheets, meaning that any data you want to perform a union operation on needs to be a sheet in the spreadsheet.


- **Intersection** – Calculates the intersection of two or more sheets on a certain column. All sheets must have the exact same schema.

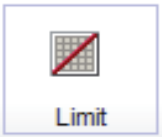



- **Complement** - Calculates the complement of two or more sheets on a certain column and returns all rows that contain a value in the specified column that only occurs in the first selected sheet. All sheets must have the exact same schema.

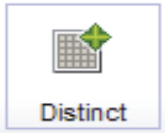

© 2013 IBM Corporation



# BigSheets – Limit, Distinct, Copy, and Formula




- **Limit** – Limits the number of rows in a sheet.




- **Distinct** – Eliminates duplicate rows in a sheet.

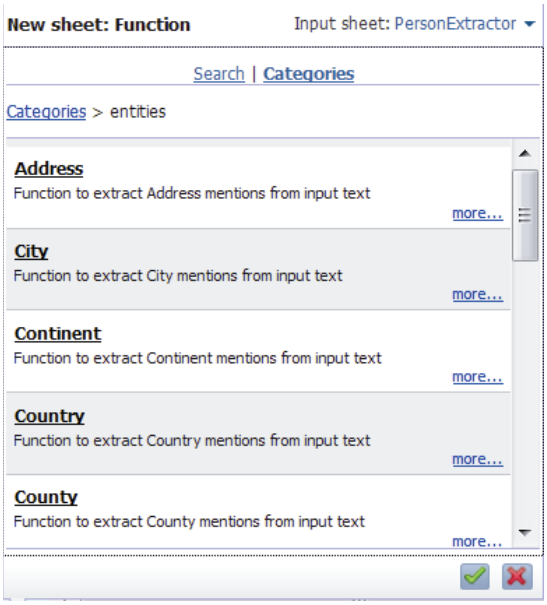

- **Copy** – Copies a sheet, including the data and all the formulas used to create the data. Copy is useful if you have sheets that were created in a similar way but you want to make minor adjustments to get different data.

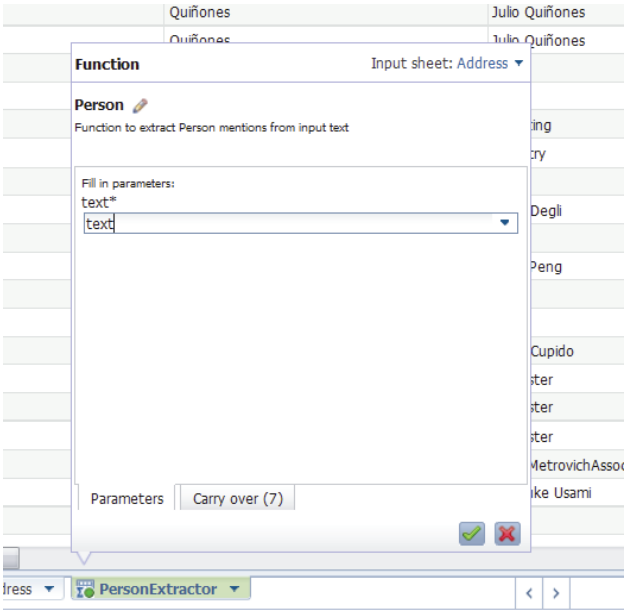

- **Formula** - Exposes a formula filed that lets you type in complex formulas.


© 2013 IBM Corporation

# Text Analytics Integration

- **Provides 10+ build-in functions to extract names, addresses, organizations, email, and phone numbers.**



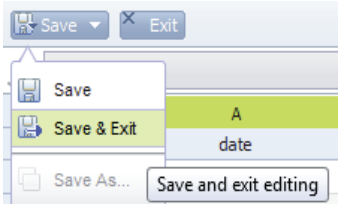





© 2013 IBM Corporation



# BigSheets – "Preview" and "Run" Details

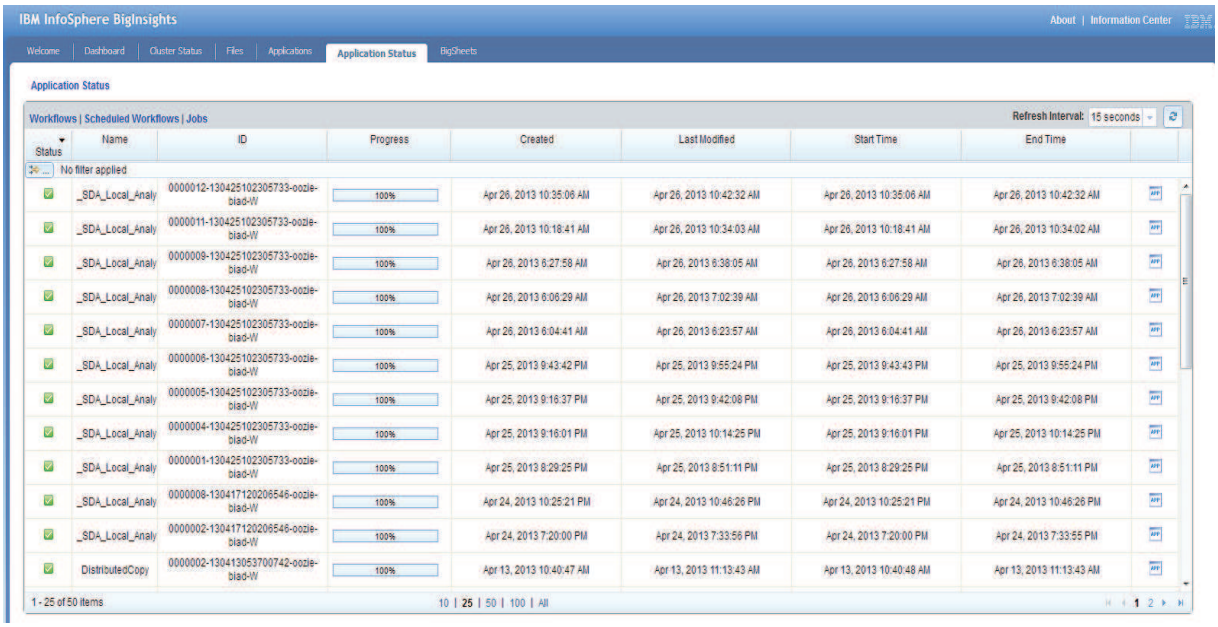



- **Save – allows you to save your changes.**
- **Exit – must exit "edit" mode to see your results.**
  - Once you Exit, the system "runs" a simulation on the first 2,000 rows of the data set and displays the first 50 rows.
- **This "preview" lets you see a simulated example only.**
- **Click "run" to see the results against the entire data set within the Workbook.**


© 2013 IBM Corporation

# Inspecting Runtime Statistics

- **Monitor the progress of the jobs**
- **Drill down to statistics relevant to the job**



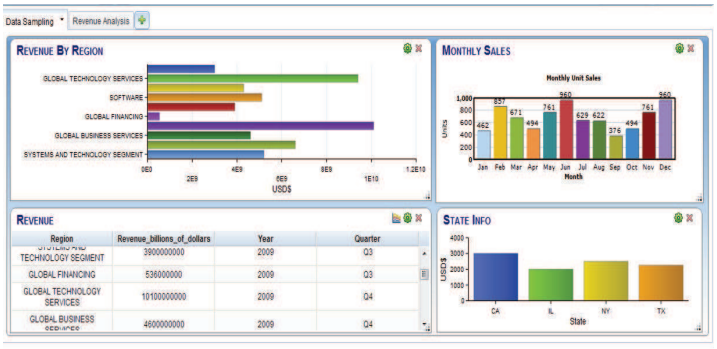


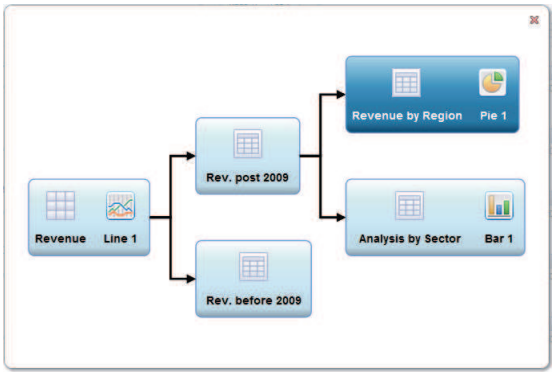
© 2013 IBM Corporation



# Visualizing Results

- **Centralized dashboard to visualize analytic results**
  - BigSheets workbooks
  - Analytic application results
  - Monitoring metrics
- **Application Linking**
  - Compose new applications from existing applications and BigSheets
  - Invoke analytics applications from web console, integration with BigSheets
- **BigSheets usability enhancements**
  - View BigSheets data flows between and across data sets to quickly navigate and relate analysis and charts
  - Inner outer joins, enhanced filters for BigSheets columns, column data-type mapping for workbooks and application of analytics to BigSheets columns
- **New application type support**
  - BigSheets macro and BigSheets reader



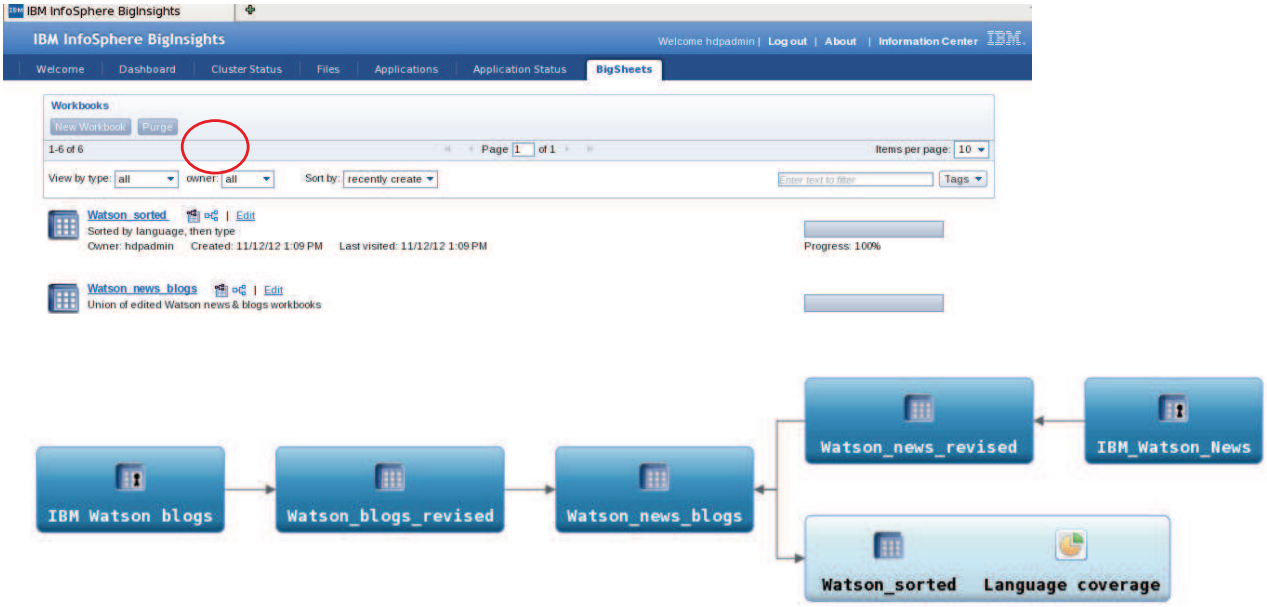





© 2013 IBM Corporation

# Displaying the workflow diagram

- **Preview of BigSheets and workflow**



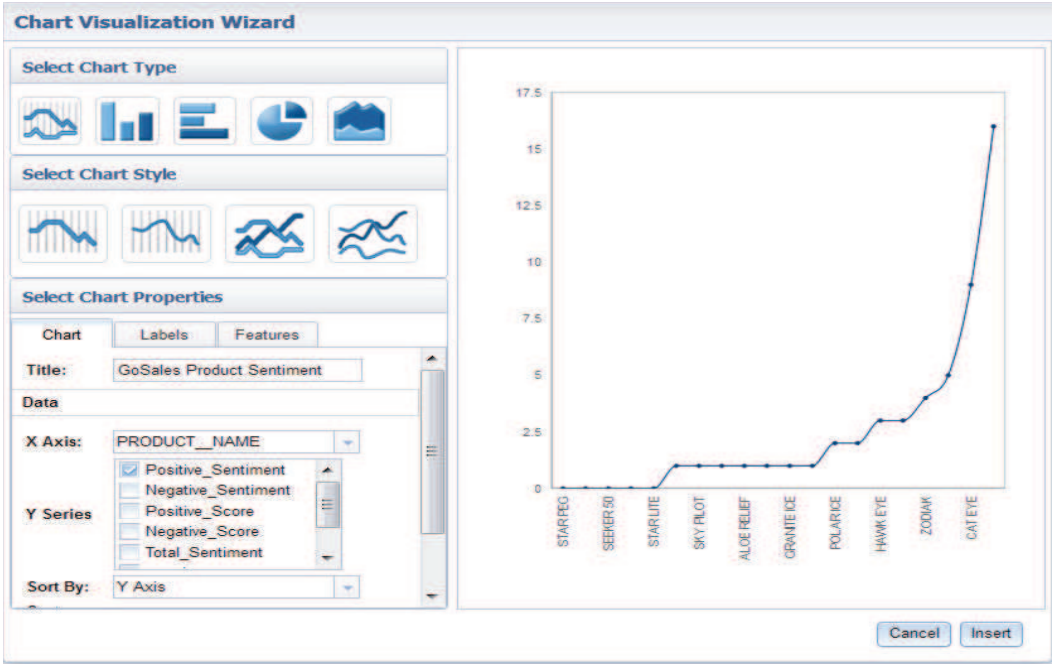


© 2013 IBM Corporation



# BigSheets Dashboard Enhancements

- **Instant preview of new charts in Dashboard**

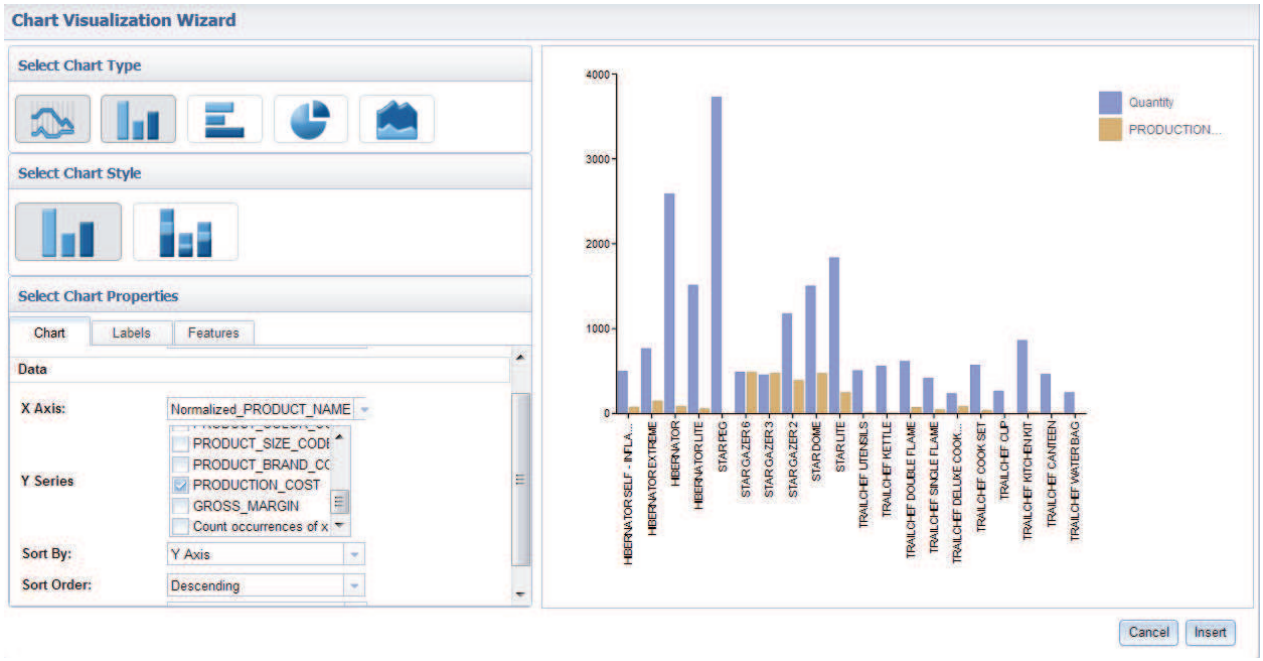




© 2013 IBM Corporation

# BigSheets Chart Types Enhancements

- **New chart types including multi-series charts**



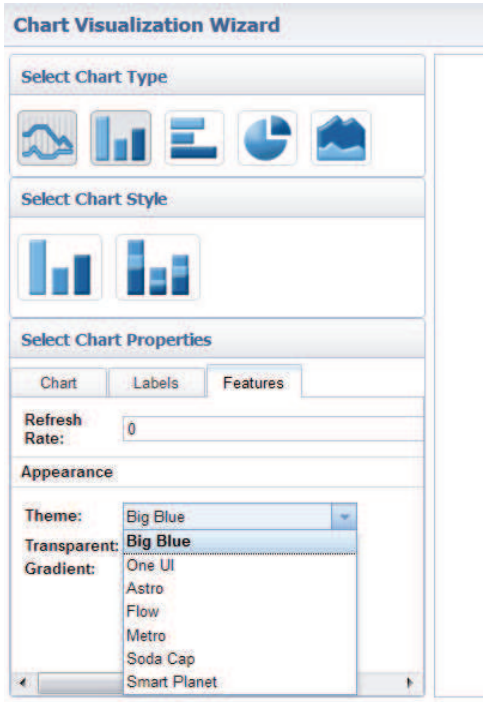


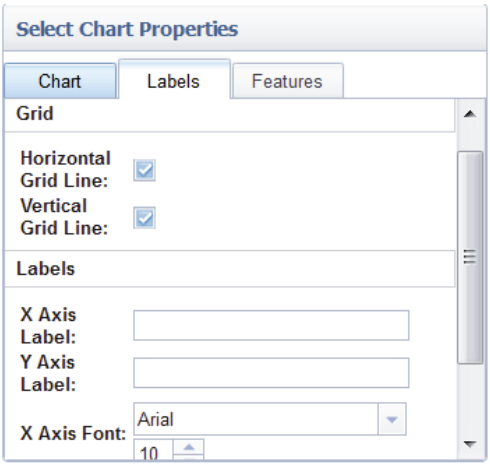
© 2013 IBM Corporation



# BigSheets Chart Types Enhancements – Continued

- **Themes and enhanced customization of charts**



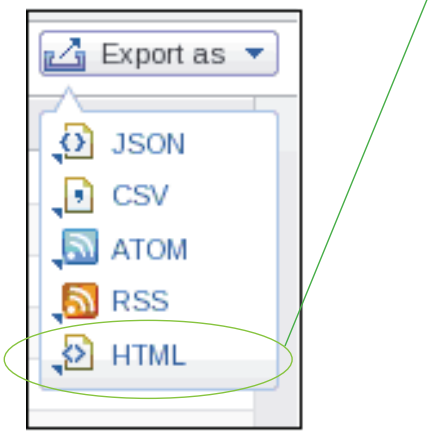





© 2013 IBM Corporation

# Exporting data

- **Useful for sharing with downstream applications**
- **Several common formats supported**




Getting Started | IBM | Imported From Firef...

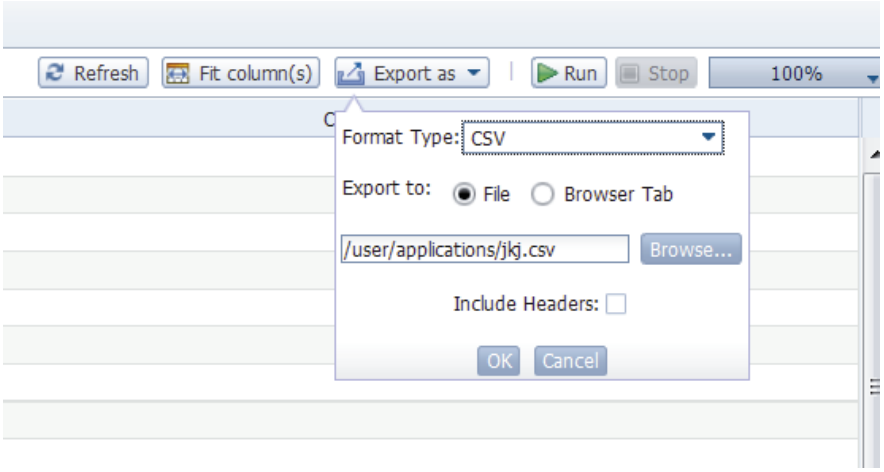
| header1 | header2 | header3 | header4 |
|---|---|---|---|
| A00 | SPIFFY COMPUTER SERVICE DIV. | 10 | A00 |
| B01 | PLANNING | 20 | A00 |
| C01 | INFORMATION CENTER | 30 | A00 |
| D01 | DEVELOPMENT CENTER | 0 | A00 |
| D11 | MANUFACTURING SYSTEMS | 60 | D01 |
| D21 | ADMINISTRATION SYSTEMS | 70 | D01 |
| E01 | SUPPORT SERVICES | 50 | A00 |
| E11 | OPERATIONS | 90 | E01 |
| E21 | SOFTWARE SUPPORT | 100 | E01 |
| F22 | BRANCH OFFICE F2 | 0 | E01 |
| G22 | BRANCH OFFICE G2 | 0 | E01 |


© 2013 IBM Corporation



# Exporting data – Continued

- **New JSON Object Reader**
  - Support reading JAQL and pig output
- **GPFS support in BigSheets.**
- **Export discovery to distributed filesystem**

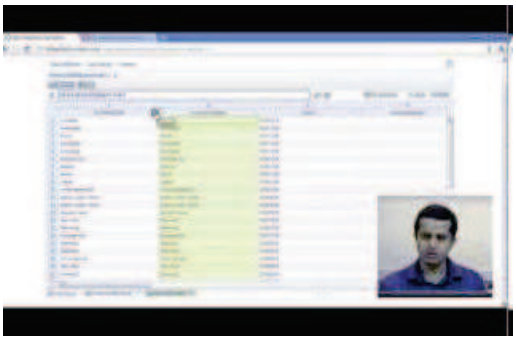




© 2013 IBM Corporation

# On-demand videos

- **Available on YouTube's IBM Big Data Channel at http://www.youtube.com/user/ibmbigdata**
- **"Analyzing Social Media for IBM Watson"**
- **"Big Data Patent Analysis with BigSheets"**
- **"Big Data for Business Users"**
- **"BigSheets in Action"**
- **See the full list of videos at http://tinyurl.com/biginsights**




© 2013 IBM Corporation



# Summary

- BigSheets – Spreadsheet-style tool
- Allows LOB to perform ad-hoc analysis on unstructured and structured content
- Gather – Analyze – Visualize
- Leverages Hadoop and Map Reduce technologies


© 2013 IBM Corporation

# Questions?

E-mail: impe.biginsights@ca.ibm.com
Subject: Big data bootcamp


© 2013 IBM Corporation