

# Data Discovery with BigSheets



Wednesday, August 07, 2013

© 2013 IBM Corporation

PureData Ecosystem

IBM

## 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**

# What is BigSheets?

- Browser-based analytics tool for business users

## Why BigSheets?

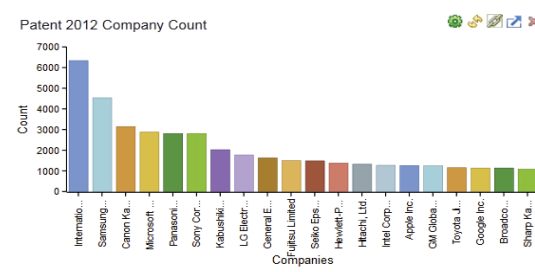
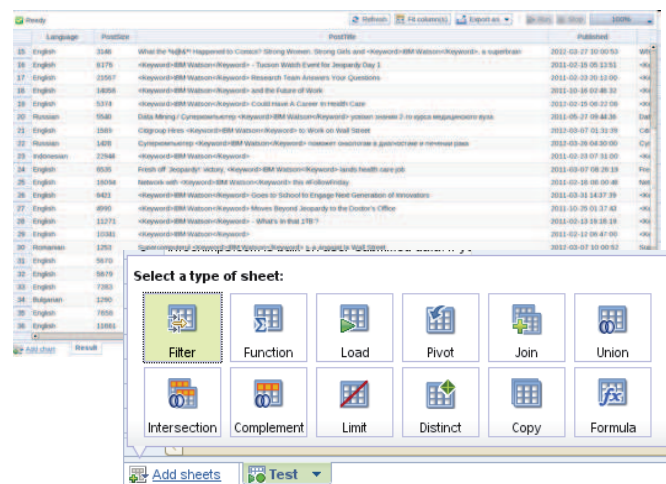
- Business users need a non-technical approach for analyzing Big Data
- Translating untapped data into actionable business insights is a common requirement
- Visualizing and drilling down into enterprise and Web data promotes new business intelligence

## 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

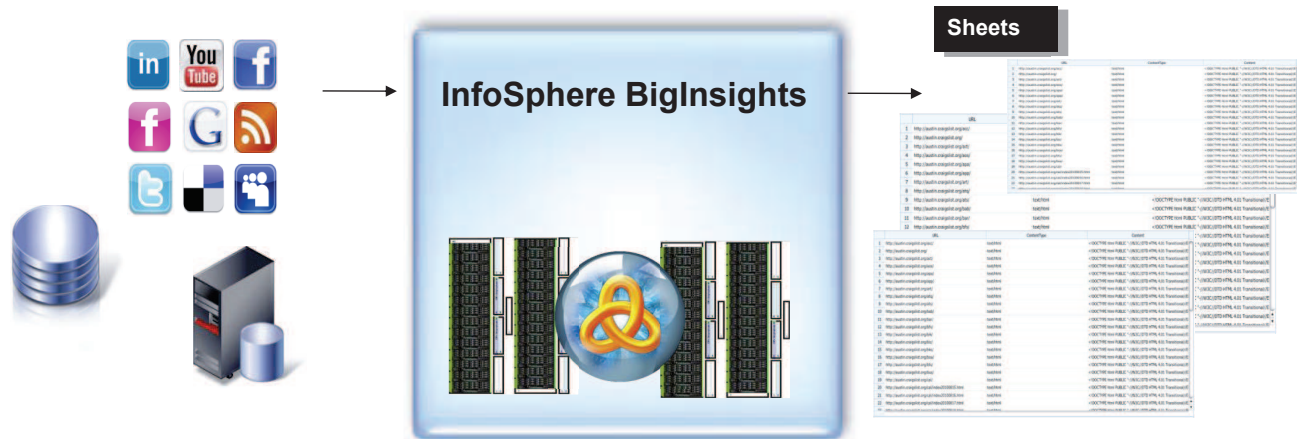
- Model “big data” collected from various sources in spreadsheet-like structures
- Filter and enrich content with built-in functions
- Combine data in different workbooks
- Visualize results through spreadsheets, charts
- Export data into common formats (if desired)



*No programming knowledge needed!*

## 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



5

© 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

	Country	Crawled	FeedsIn	Inserted	IsAdult	Language	PostSize
1	US	2011-02-19 01:28:04	("Title": "LA COSA DELLA COSA SI", "Entry": "2011-02-23 10:38:24", "Country": "US", "Language": "English", "PostSize": 7721)	0	0	English	7721
2	US	2011-02-26 23:01:17	("Title": "Out of the Jungle", "Entry": "2011-02-26 23:08:04", "Country": "US", "Language": "English", "PostSize": 8055)	0	0	English	8055
3	US	2011-01-18 00:02:45	("Title": "Chris Pepper", "Entry": "2011-01-17 23:58:08", "Country": "US", "Language": "English", "PostSize": 2153)	0	0	English	2153
4	US	2011-02-18 15:21:31	("Title": "Tommy, Julie and Adam", "Entry": "2011-02-18 15:33:21", "Country": "US", "Language": "English", "PostSize": 3333)	0	0	English	3333
5	US	2012-05-21 20:20:17	("Title": "Health from Express", "Entry": "2012-05-21 20:25:03", "Country": "US", "Language": "English", "PostSize": 6418)	0	0	English	6418
6	US	2011-05-07 23:49:48	("Title": "Alamy Area Mink", "Entry": "2011-05-07 23:55:04", "Country": "US", "Language": "English", "PostSize": 1302)	0	0	English	1302
7	US	2011-01-28 16:20:01	("Title": "Baruch Area", "Entry": "2011-01-28 16:34:29", "Country": "US", "Language": "English", "PostSize": 1443)	0	0	English	1443
8	US	2011-06-18 14:53:50	("Title": "Science and", "Entry": "2011-06-18 15:00:02", "Country": "US", "Language": "English", "PostSize": 4486)	0	0	English	4486
9	US	2011-05-29 12:27:45	("Title": "Jas79's Blog", "Entry": "2011-05-29 12:44:03", "Country": "US", "Language": "English", "PostSize": 3875)	0	0	English	3875
10	US	2011-02-24 01:23:33	("Title": "Lester", "Entry": "2011-02-24 01:25:06", "Country": "US", "Language": "English", "PostSize": 1941)	0	0	English	1941
11	US	2011-02-22 09:51:23	("Title": "The Gals", "Entry": "2011-02-24 11:24:08", "Country": "US", "Language": "English", "PostSize": 15049)	0	0	English	15049
12	US	2011-02-17 06:00:18	("Title": "The Gals", "Entry": "2011-02-17 12:52:13", "Country": "US", "Language": "English", "PostSize": 1205)	0	0	English	1205
13	US	2011-03-14 12:32:37	("Title": "Gonsolo", "Entry": "2011-03-14 12:38:03", "Country": "US", "Language": "English", "PostSize": 11606)	0	0	English	11606
14	US	2011-02-11 21:03:11	("Title": "Kronos - Notes from Silicon Valley", "Entry": "2011-02-11 21:15:03", "Country": "US", "Language": "English", "PostSize": 3402)	0	0	English	3402
15	US	2011-02-10 18:15:51	("Title": "Johanna: Meetings of a So-Something Geek", "Entry": "2011-02-10 19:44:03", "Country": "US", "Language": "English", "PostSize": 9301)	0	0	English	9301
16	US	2011-02-17 04:32:10	("Title": "The Gals", "Entry": "2011-02-17 12:52:07", "Country": "US", "Language": "English", "PostSize": 7957)	0	0	English	7957

6

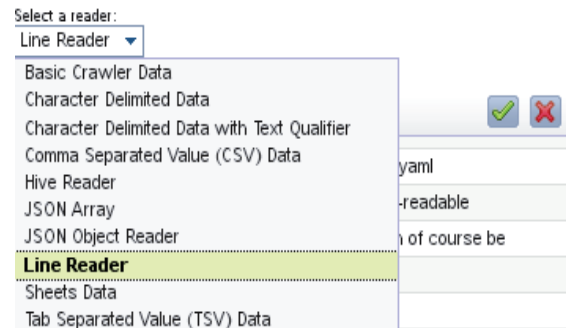
© 2013 IBM Corporation

## 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

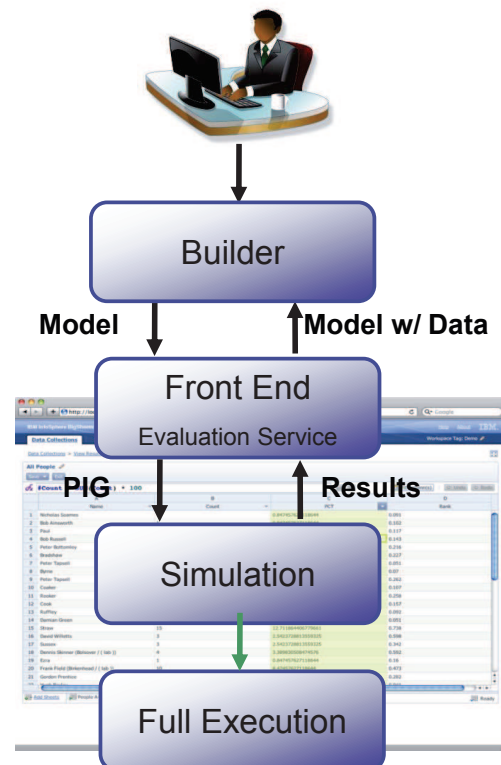


7

© 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



## 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

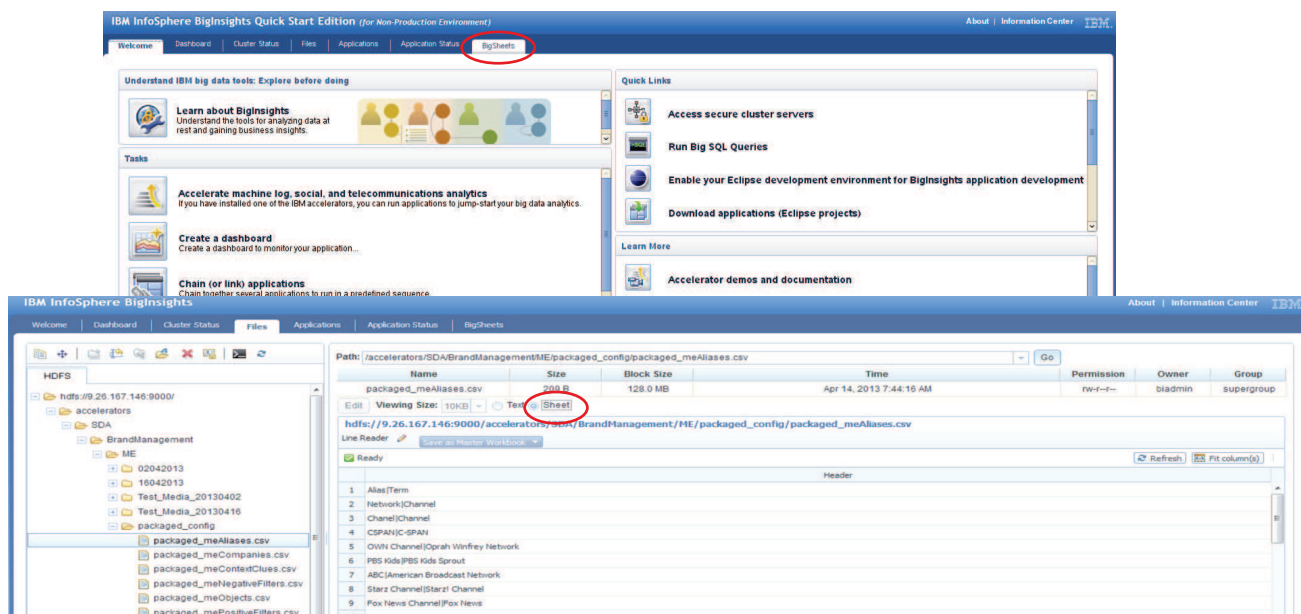


9

© 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>

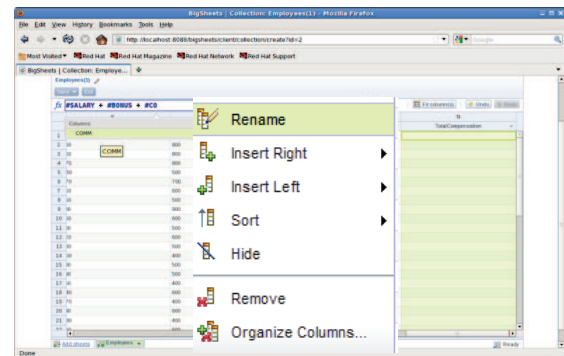
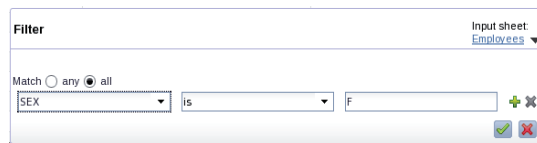
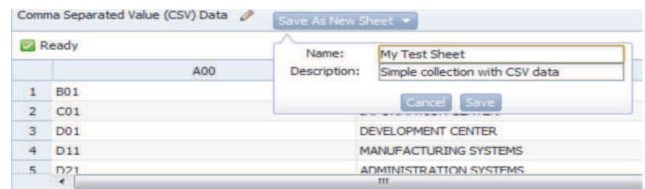
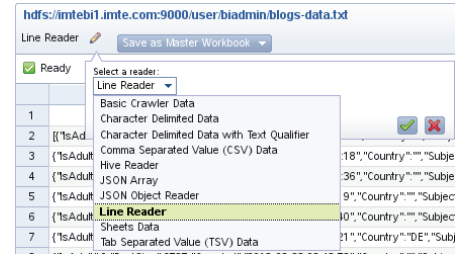


10

© 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



11

© 2013 IBM Corporation

## BigSheets – Data Refresh Steps

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

▼ Details:	
Description:	Base Gift Card Data.
Tags:	
Reader:	Comma Separated Value (CSV) Data
Source:	Network: <a href="https://master-1-internal.imdemocloud.com:9000/user/biadmin/big_data_series/gift_cards.csv">https://master-1-internal.imdemocloud.com:9000/user/biadmin/big_data_series/gift_cards.csv</a>

- ...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”

Data is "Out of Sync"



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

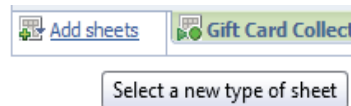
12

© 2013 IBM Corporation

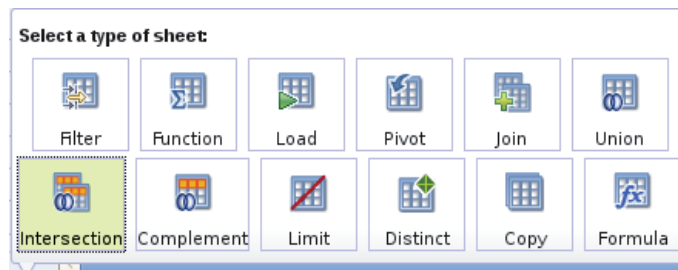


## BigSheets – Subsequent Data Processing Options

- 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

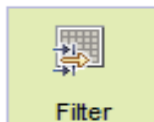


- Details on the next slides...

13

© 2013 IBM Corporation

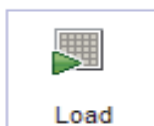
## BigSheets – Filter, Macro, Load and Join



- **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)

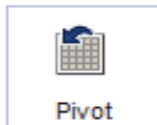


- **Load** – Brings the data of another workbook as a sheet. Load is useful if you want to perform a join or union operation.



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

## BigSheets – Pivot, Union, Intersection, Complement



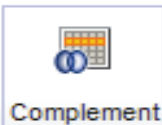
- **Pivot** – Calculates values by grouping the data in the workbook, applying functions to each group, and carrying over data.



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

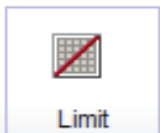


- **Intersection** – Calculates the intersection of two or more sheets on a certain column. All sheets must have the exact same schema.

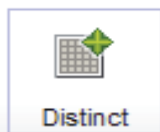


- **Complement** - Calculates the complement of two or more sheets on a certain column and returns all rows that contain a value in the specified column that only occurs in the first selected sheet. All sheets must have the exact same schema.

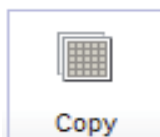
## BigSheets – Limit, Distinct, Copy, and Formula



- **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.

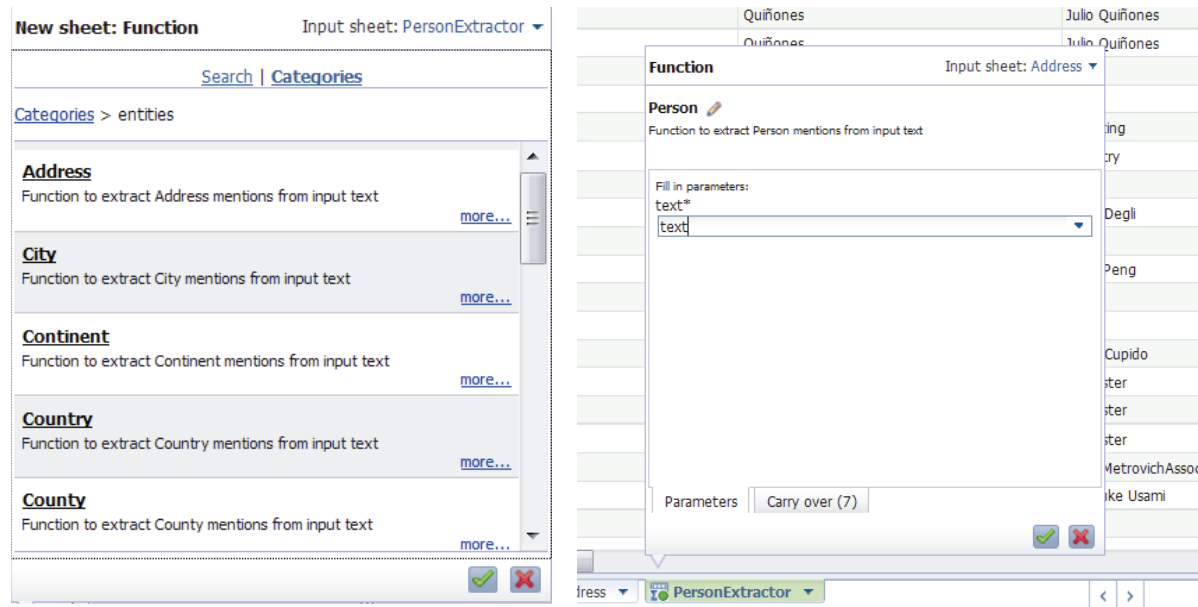


- **Formula** - Exposes a formula field that lets you type in complex formulas.



## Text Analytics Integration

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

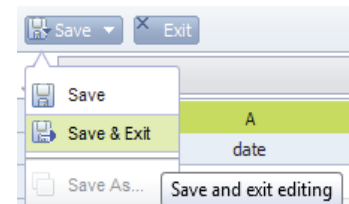


17

© 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.



18

© 2013 IBM Corporation

## Inspecting Runtime Statistics

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

IBM InfoSphere BigInsights

Application Status

Status	Name	ID	Progress	Created	Last Modified	Start Time	End Time
✓	_SDA_Local_Analy	0000012-130425102305733-oodie-biaa-W	100%	Apr 26, 2013 10:35:06 AM	Apr 26, 2013 10:42:32 AM	Apr 26, 2013 10:35:06 AM	Apr 26, 2013 10:42:32 AM
✓	_SDA_Local_Analy	0000011-130425102305733-oodie-biaa-W	100%	Apr 26, 2013 10:18:41 AM	Apr 26, 2013 10:34:03 AM	Apr 26, 2013 10:18:41 AM	Apr 26, 2013 10:34:02 AM
✓	_SDA_Local_Analy	0000008-130425102305733-oodie-biaa-W	100%	Apr 26, 2013 8:27:58 AM	Apr 26, 2013 8:38:05 AM	Apr 26, 2013 8:27:58 AM	Apr 26, 2013 8:38:05 AM
✓	_SDA_Local_Analy	0000008-130425102305733-oodie-biaa-W	100%	Apr 26, 2013 8:06:29 AM	Apr 26, 2013 7:02:39 AM	Apr 26, 2013 8:06:29 AM	Apr 26, 2013 7:02:39 AM
✓	_SDA_Local_Analy	0000007-130425102305733-oodie-biaa-W	100%	Apr 26, 2013 6:04:41 AM	Apr 26, 2013 6:23:57 AM	Apr 26, 2013 6:04:41 AM	Apr 26, 2013 6:23:57 AM
✓	_SDA_Local_Analy	0000006-130425102305733-oodie-biaa-W	100%	Apr 25, 2013 9:43:42 PM	Apr 25, 2013 9:55:24 PM	Apr 25, 2013 9:43:43 PM	Apr 25, 2013 9:55:24 PM
✓	_SDA_Local_Analy	0000005-130425102305733-oodie-biaa-W	100%	Apr 25, 2013 9:16:37 PM	Apr 25, 2013 9:42:08 PM	Apr 25, 2013 9:16:37 PM	Apr 25, 2013 9:42:08 PM
✓	_SDA_Local_Analy	0000004-130425102305733-oodie-biaa-W	100%	Apr 25, 2013 9:16:01 PM	Apr 25, 2013 10:14:25 PM	Apr 25, 2013 9:16:01 PM	Apr 25, 2013 10:14:25 PM
✓	_SDA_Local_Analy	0000001-130425102305733-oodie-biaa-W	100%	Apr 25, 2013 8:29:25 PM	Apr 25, 2013 8:51:11 PM	Apr 25, 2013 8:29:25 PM	Apr 25, 2013 8:51:11 PM
✓	_SDA_Local_Analy	0000008-130417120206546-oodie-biaa-W	100%	Apr 24, 2013 10:25:21 PM	Apr 24, 2013 10:46:26 PM	Apr 24, 2013 10:25:21 PM	Apr 24, 2013 10:46:26 PM
✓	_SDA_Local_Analy	0000002-130417120206546-oodie-biaa-W	100%	Apr 24, 2013 7:20:00 PM	Apr 24, 2013 7:33:58 PM	Apr 24, 2013 7:20:00 PM	Apr 24, 2013 7:33:58 PM
✓	DistributedCopy	0000002-130413053700742-oodie-biaa-W	100%	Apr 13, 2013 10:40:47 AM	Apr 13, 2013 11:13:43 AM	Apr 13, 2013 10:40:48 AM	Apr 13, 2013 11:13:43 AM

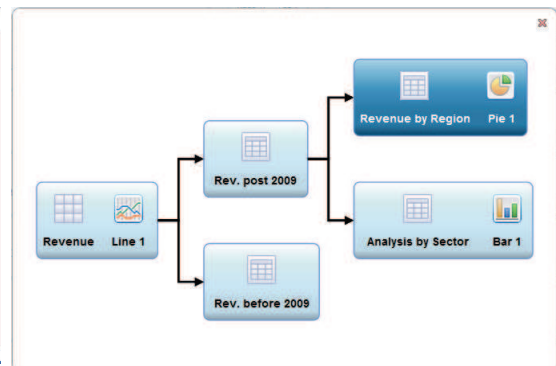
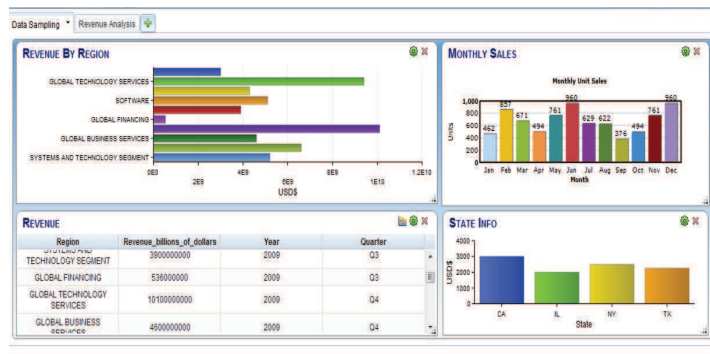
1 - 25 of 50 items

19

© 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

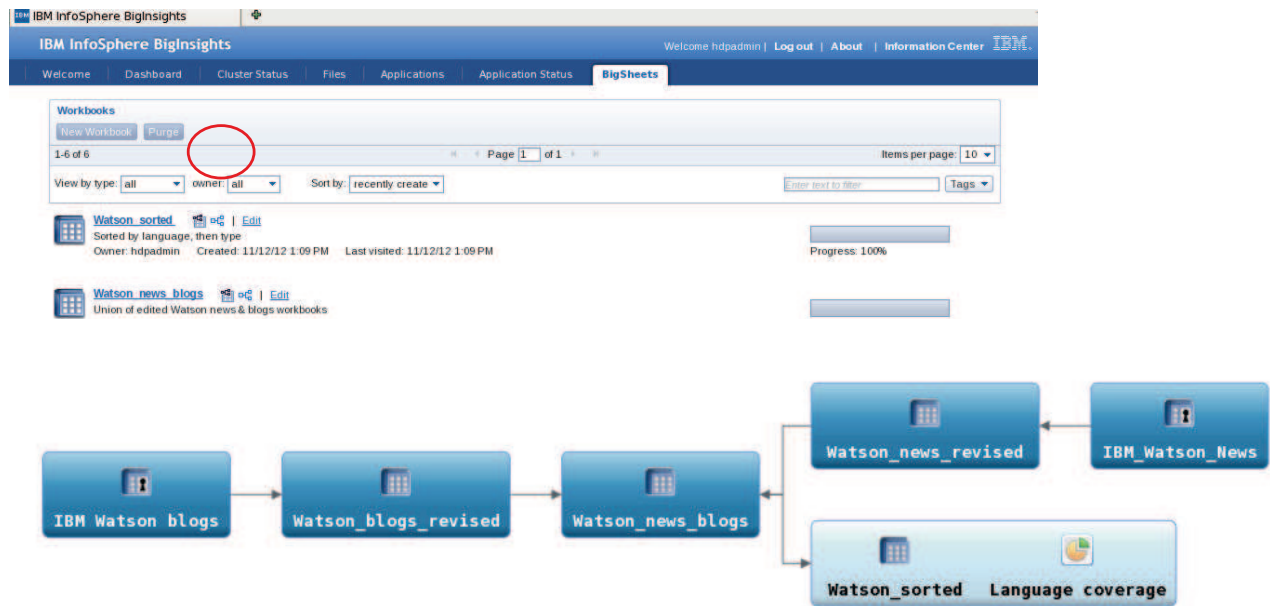


20

© 2013 IBM Corporation

## Displaying the workflow diagram

### ■ Preview of BigSheets and workflow

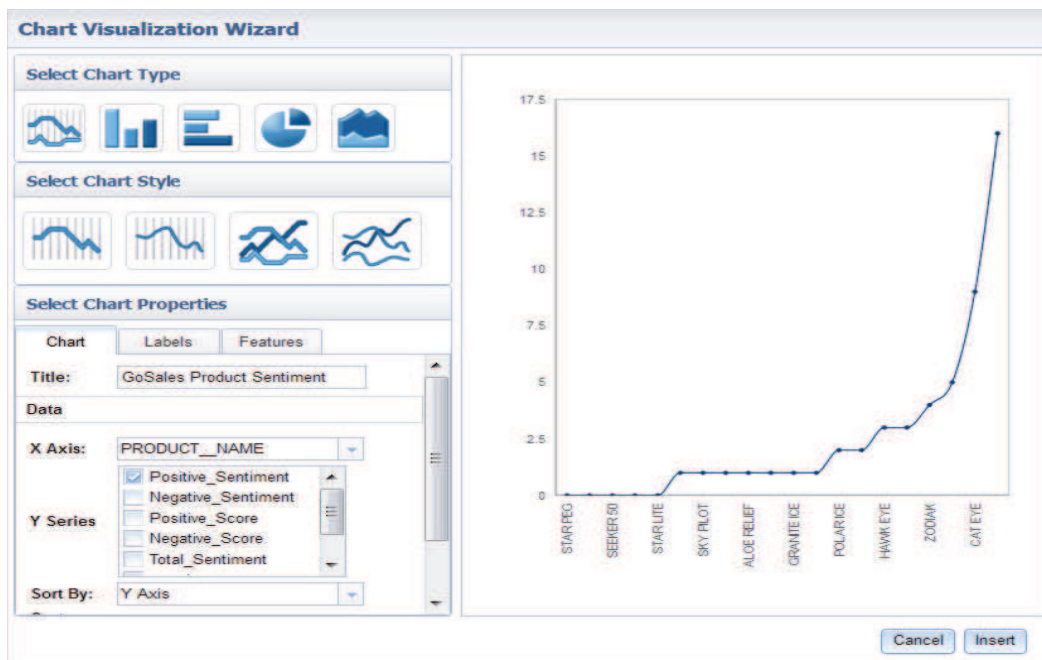


21

© 2013 IBM Corporation

## BigSheets Dashboard Enhancements

### ■ Instant preview of new charts in Dashboard

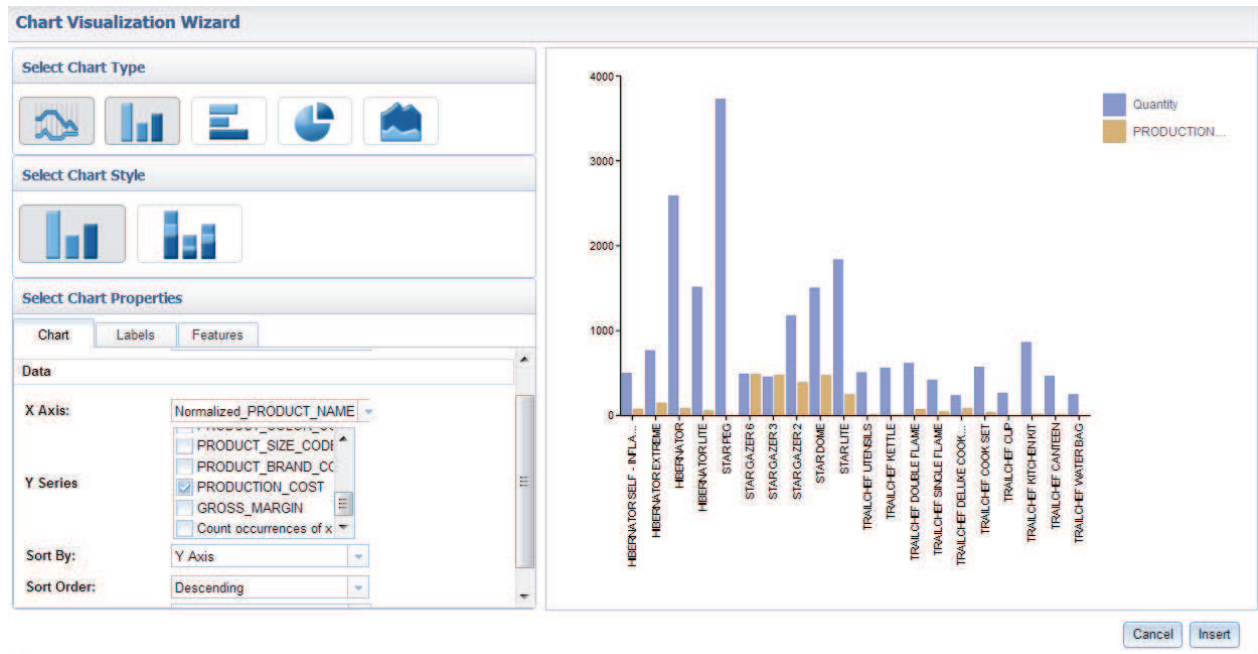


22

© 2013 IBM Corporation

## BigSheets Chart Types Enhancements

- New chart types including multi-series charts

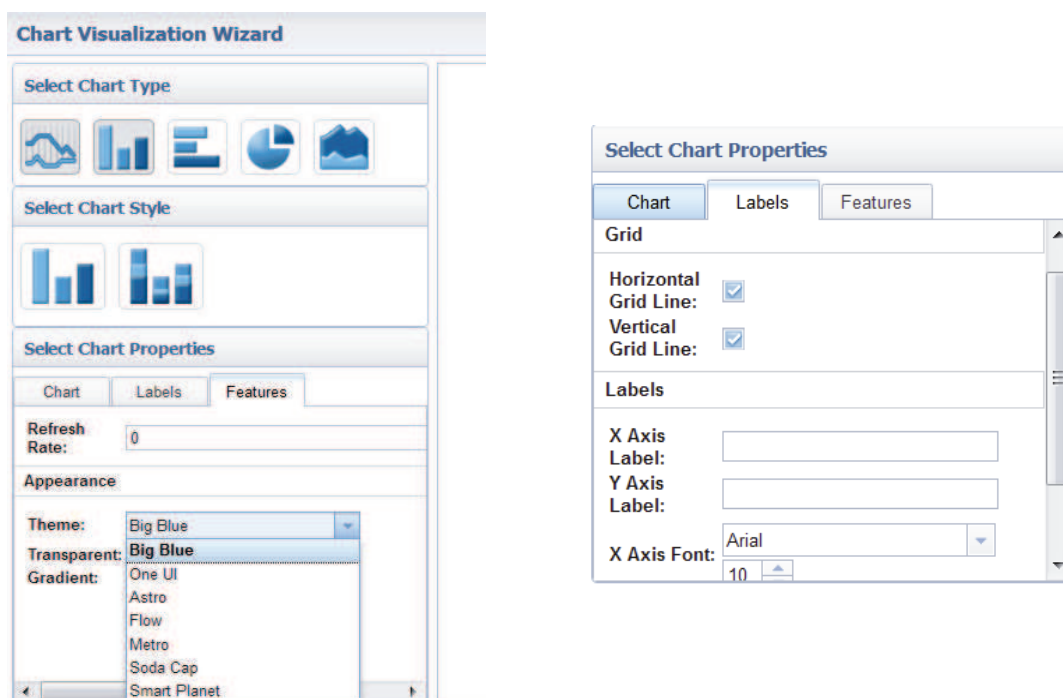


23

© 2013 IBM Corporation

## BigSheets Chart Types Enhancements – Continued

- Themes and enhanced customization of charts

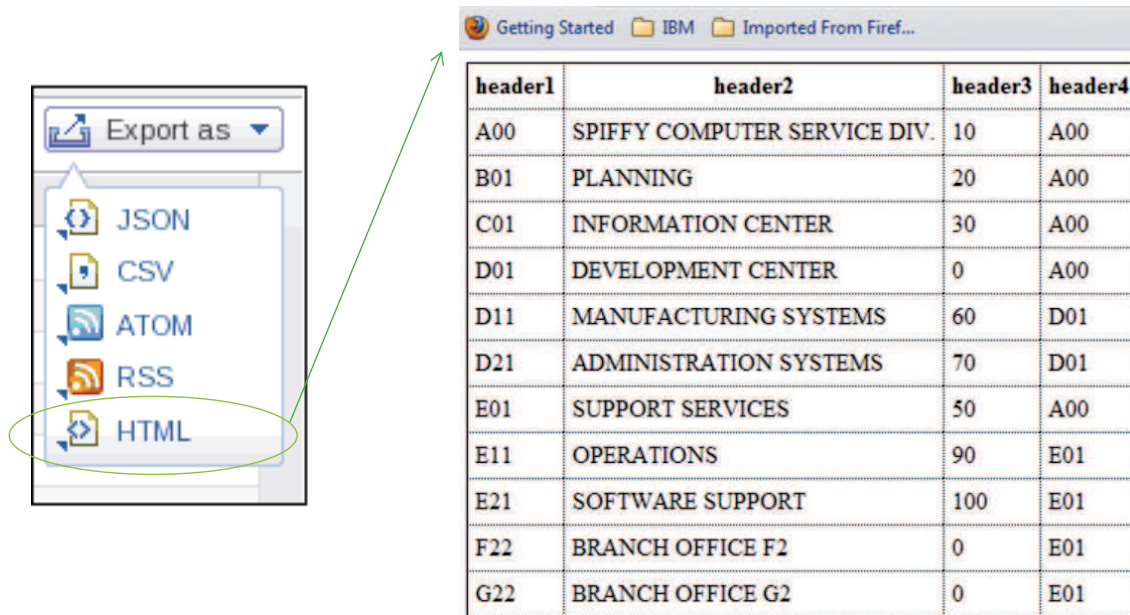


24

© 2013 IBM Corporation

## Exporting data

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

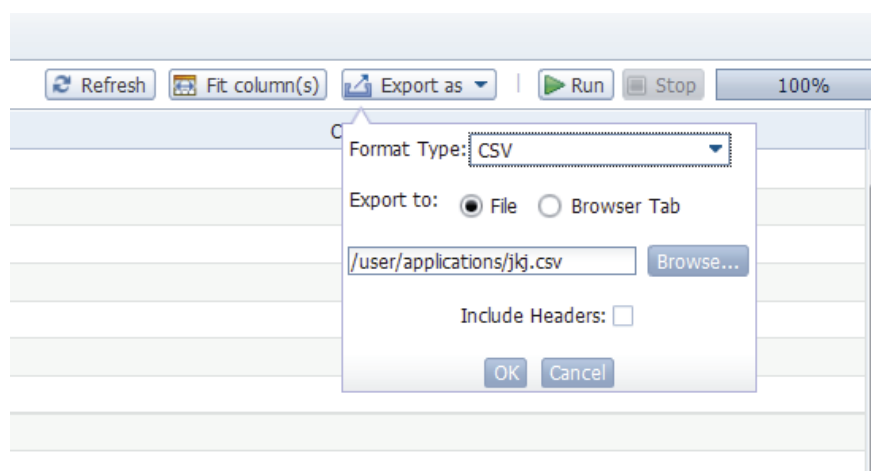


25

© 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



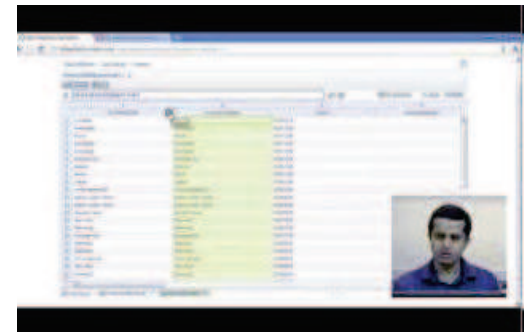
26

© 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>



## 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





Questions?

E-mail: [impe.biginsights@ca.ibm.com](mailto:impe.biginsights@ca.ibm.com)

Subject: Big data bootcamp

