

# **Getting started with**

## **Text Analytics & BigSheets**

June, 2017

## *Contents*

<b>LAB 1</b>	<b>OVERVIEW .....</b>	<b>4</b>
1.1.	WHAT YOU'LL LEARN .....	4
<b>LAB 2</b>	<b>THE DATA.....</b>	<b>5</b>
2.1.	OUR TASK IN THE LAB .....	5
<b>LAB 3</b>	<b>LAUNCHING THE BIGINSIGHTS HOME AND TEXT ANALYTICS WEB.....</b>	<b>6</b>
3.1.	LAUNCHING THE BigINSIGHTS HOME AND BIG SQL WEB TOOLING .....	6
3.2.	CREATE A PROJECT .....	8
3.3.	THE LIBRARY AND PRE-BUILD EXTRACTORS .....	11
<b>LAB 4</b>	<b>BUILDING A SEQUENCE.....</b>	<b>14</b>
4.1.	CREATE THE FIRST SEQUENCE .....	14
4.2.	COMPLETING THE SEQUENCE.....	20
<b>LAB 5</b>	<b>ADDING A SECOND PATTERN.....</b>	<b>24</b>
<b>LAB 6</b>	<b>BIGSHEETS.....</b>	<b>28</b>

---

## Lab 1 Overview

This Hand-on lab will introduce you to the BigInsights Text Analytics Web Tooling for BigInsights and BigSheets v4.2.

There's a convenient stopping point when you've built a single sequence, but if you continue, you will sample all the features.

In this exercise, we extract information from log files written by Cisco routers. The use case is based on a paper Cisco published explaining how these files can be used to understand potentially malicious network activity. <http://www.cisco.com/web/about/security/intelligence/identify-incidents-via-syslog.html>

The data for the demo is already loaded into your image in a single file containing multiple log records.

[Allow **60 minutes** to complete.]

### 1.1. What you'll learn

After completing all exercises in this lab guide, you'll know how to

- How to use the Text Analytics web tooling.
- Learn about Rich library of pre-built extractors.
- Create your own extractor.
- How to use BigSheets.



You can find all the resources on GitHub:

<https://github.com/fjcanobailen/biginsights>



**About the screen captures, sample code, and environment configuration**

Screen captures in this lab depict examples and results that may vary from what you see when you complete the exercises. In addition, some code examples may need to be customized to match your environment.

---

## Lab 2 The data

The log records to be analyzed look like this:

```
Aug 24 2007 10:27:29: %ASA-6-106100: access-list OUTSIDE denied tcp outside/192.168.208.63(39675)-> inside/192.168.150.77(80) hit-cnt 1 first hit [0x22e8ac21, 0x0]
```

Key elements of each record are:

- Date and time - Aug 24 2007 10:27:29
- Code - %ASA-6-106100
- Access Control List - OUTSIDE
- IP address and port - 192.168.208.63(39675)

Codes are associated with events detected by the router. Access Control Lists identify groups of IP addresses, in our example they identify IP Addresses inside or outside the firewall.

Suspicious events are identified based on events caused by actions originating outside the firewall aimed at IP addresses inside the firewall. Log records for suspicious events will contain specific codes with an originating IP address outside the firewall and a target IP address inside the firewall.

In the example above we see an example where an %ASA-6-106100 code has been raised as a result of a request from outside the firewall targeted at an address inside the firewall.

Extracting this information can help the business understand what type and level of suspicious activity is taking place and where it's originating from. Attacks can include horizontal probes where a range of IP addresses are targeted and vertical probes that target a range of ports at one IP address.

### 2.1. Our task in the lab

We will perform a first level of analysis based on suspicious activity detected by the firewall. This analysis also delivers the originating IP addresses. With those in hand, the business can also determine if any probes from those addresses got through the firewall.

The demo uses a combination of: pre-built extractors; basic features using regex and dictionaries and patterns that reflect various message styles.

---

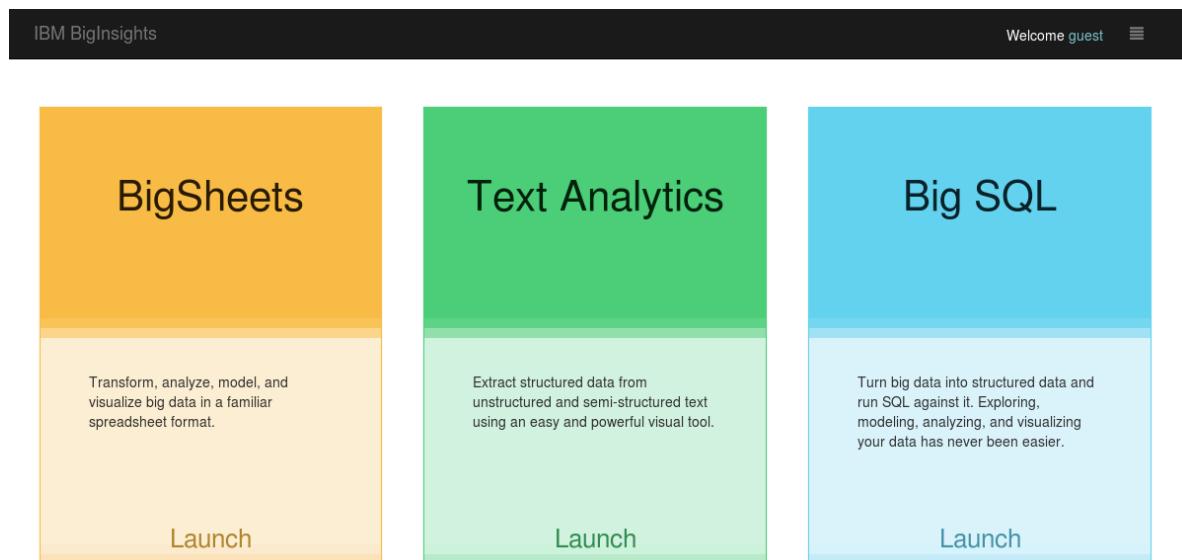
## Lab 3      Launching the BigInsights Home and Text Analytics Web

### 3.1. Launching the BigInsights Home and Big SQL web tooling

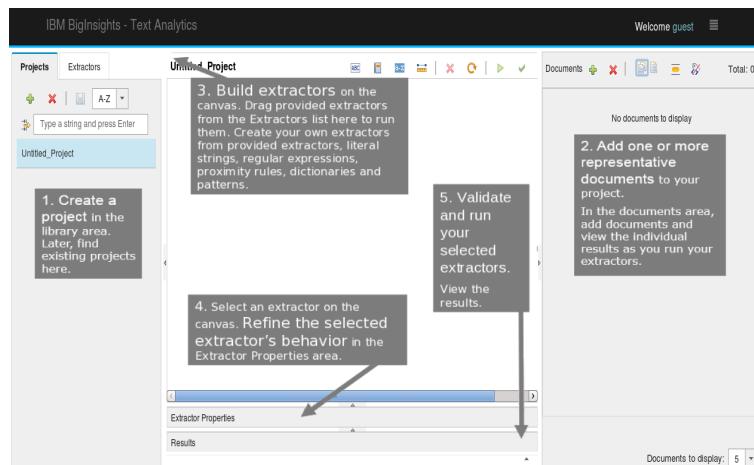
- 1. Launch BigInsights Home, providing the appropriate URL based on your installation's configuration. Assuming you installed BigInsights with Knox and accepted default installation values, the BigInsights Home URL is similar to the link shown below. **Substitute the location of the Knox gateway on your cluster for the italicized text in this example. Remember start Demo LDAP.**

<https://bigi01.localhost:8443/gateway/default/BigInsightsWeb/index.html>

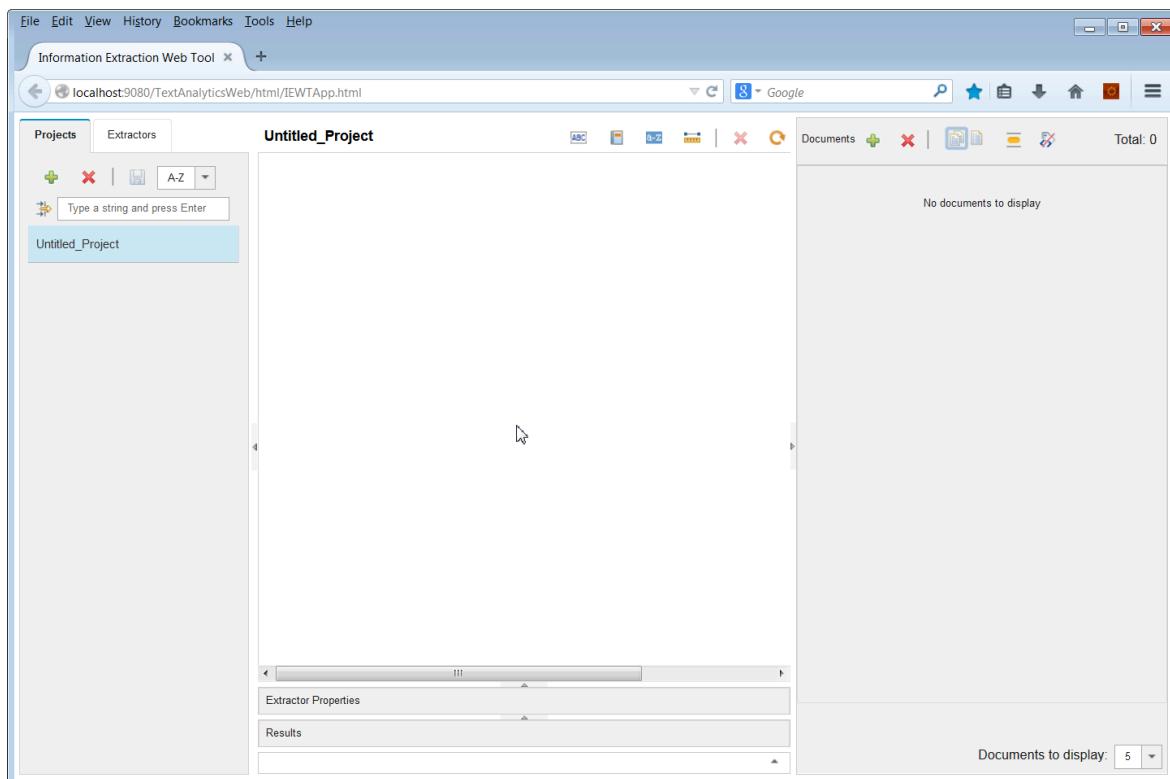
- 2. When prompted, enter a valid user ID and password for the Knox gateway. (Defaults are guest / guest-password).
- 3. Verify that BigInsights Home displays an item for Text Analytics. Depending on the size of your browser window and other BigInsights components installed on your cluster, you may need to scroll through the BigInsights Home page to locate the Text Analytics section.



- 4. Click the Launch button in the Text Analytics box. (Your screen may appear somewhat different than the image below when you launch the tool for the first time.)



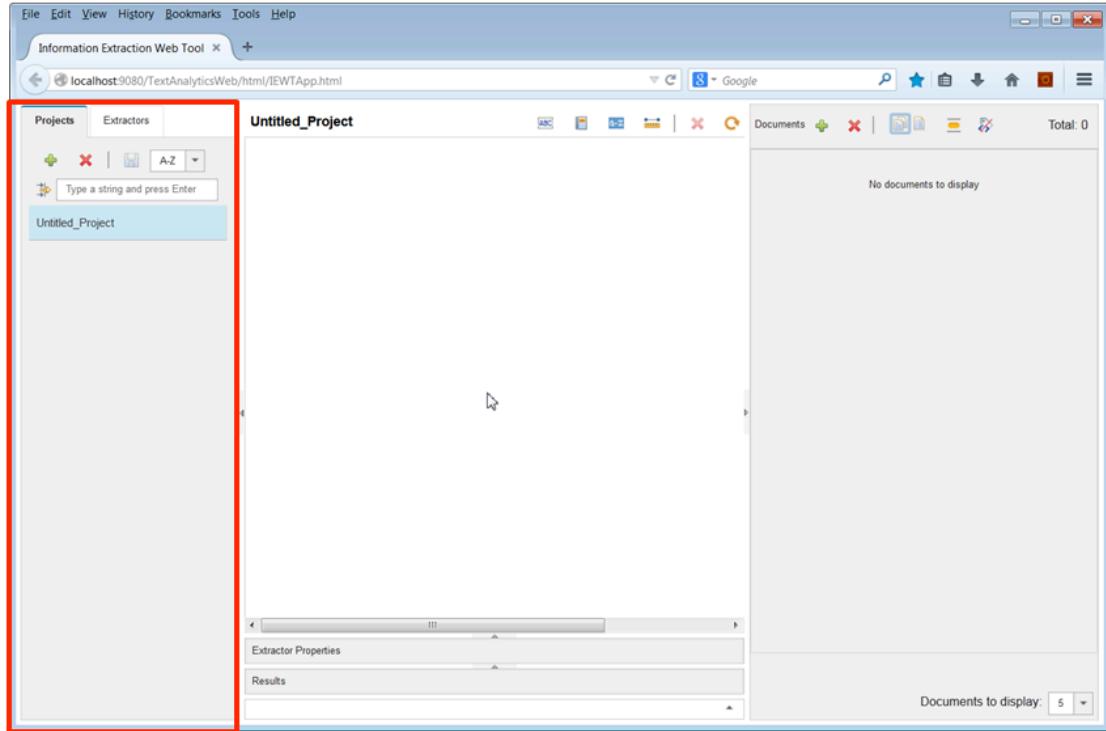
—5. Click in any place to remove the explanations.



There are 3 main panels on display in a new project.

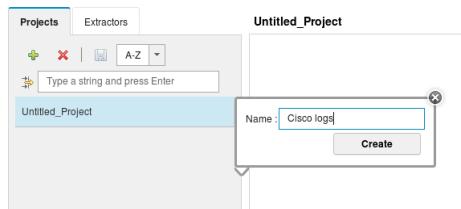
- On the left is the project panel where you can create, open, delete and search for projects.
- On the right is the documents panel where you load and delete documents that you want to extract information from
- In the middle is the working canvas where you will build your text analytics application

- In the middle is the working canvas where you will build your text analytics application

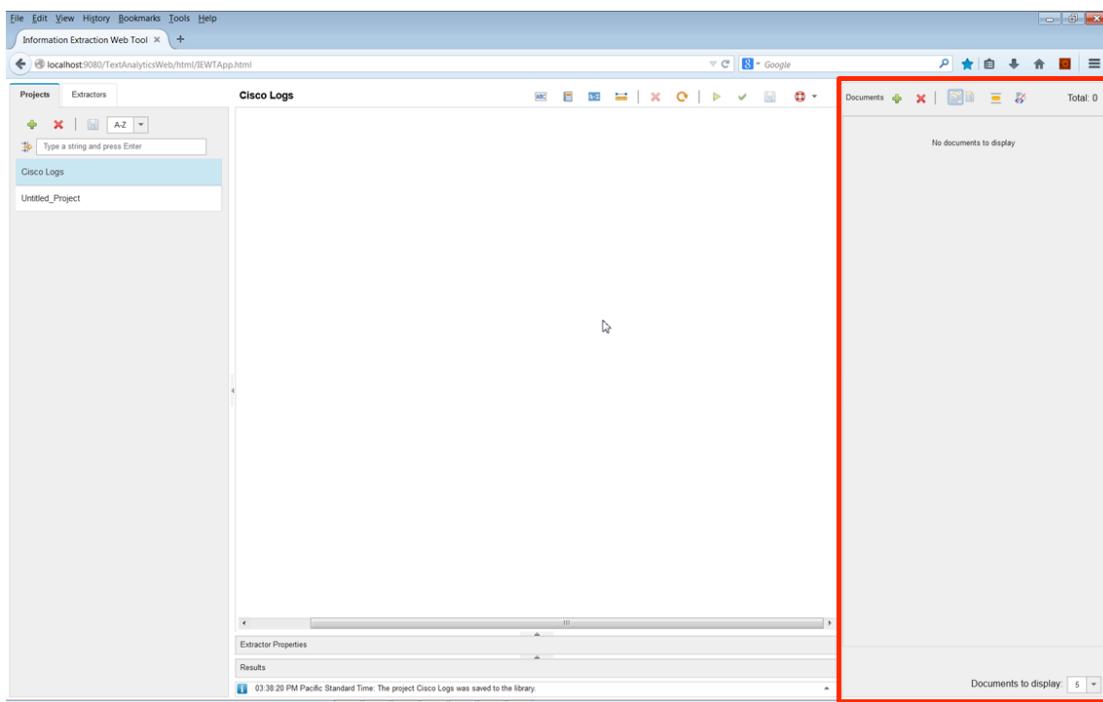


### 3.2. Create a Project

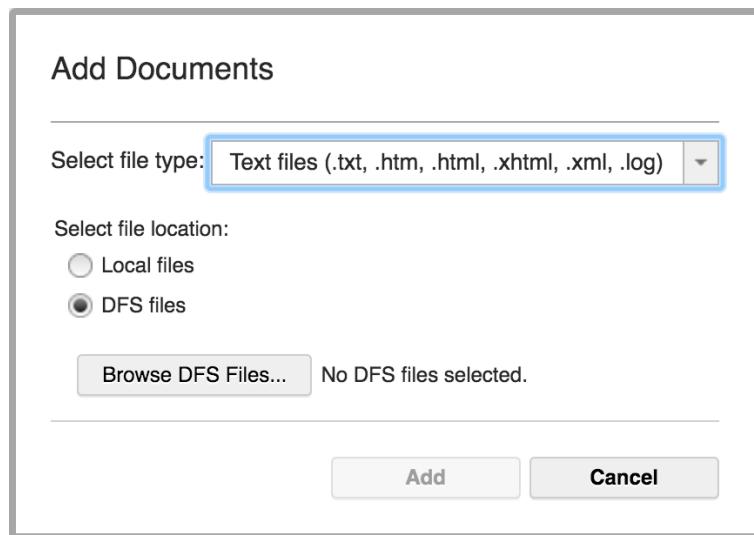
We will start with the Projects panel and create a new project by clicking on the + button and naming the project "Cisco logs" in the new project dialog



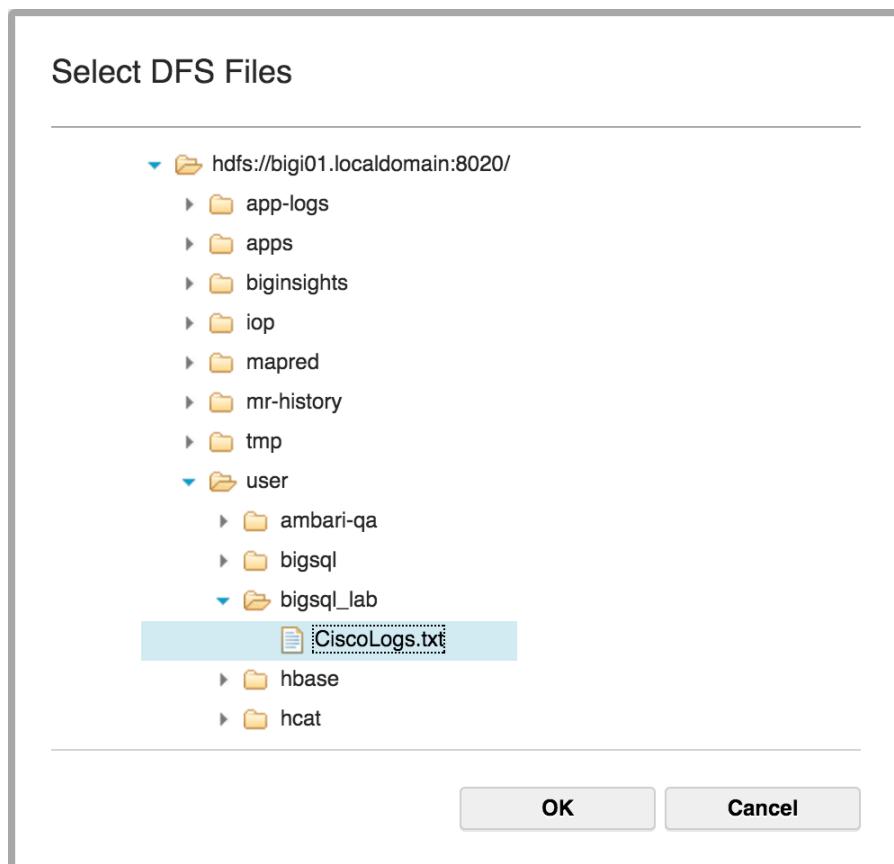
- Now we will move to the documents panel and load some data



Load a document set by clicking on the + button.



Note from the drop-down the various file types available and select text format. Leave the location as “DFS” and Browse to select the input file CiscoLogs.txt:



Click “Add” and you will see it appear in the document panel. You could use local files in your unix file system as well.

- Click on the document and select the single document button in the toolbar to switch to single document view so you can see more of the file

The screenshot shows the 'Documents' tab in the top navigation bar. A single document titled 'CiscoLogs.txt' is selected. The content pane displays several log entries:

```

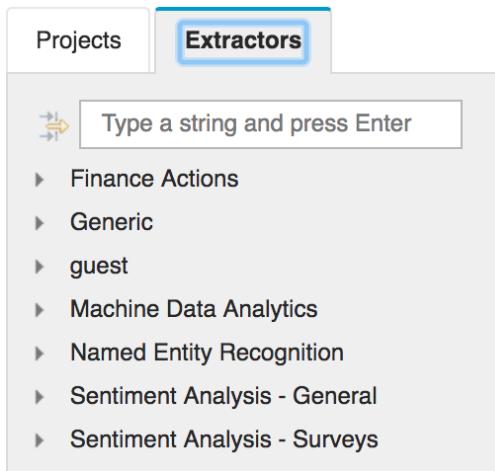
Aug 24 2007 10:27:29: %ASA-6-106100: access-list
OUTSIDE denied tcp outside/192.168.208.63(39675)->
inside/192.168.150.77(80) hit-cnt 1 first hit [0x22e8ac21,
0x0]
Aug 24 2007 10:27:31: %ASA-6-106100: access-list
OUTSIDE denied tcp outside/192.168.208.63(39676) ->
inside/192.168.150.77(80) hit-cnt 1 first hit [0x22e8ac21,
0x0]
Aug 24 2007 10:27:22: %ASA-4-400014: IDS:2004 ICMP
echo request from 192.168.208.63:39676 to
192.168.150.70(80) on interface outside
Aug 24 2007 10:27:22: %ASA-6-302020: Built ICMP
connection for faddr 192.168.208.63/15343 gaddr
192.168.150.70/0 laddr 192.168.150.70/0
Aug 24 2007 10:27:22: %ASA-6-106015: Deny TCP (no
connection) from 192.168.208.63/49827 to

```

Now we are ready to build our extractor.

### 3.3. The Library and Pre-Build extractors

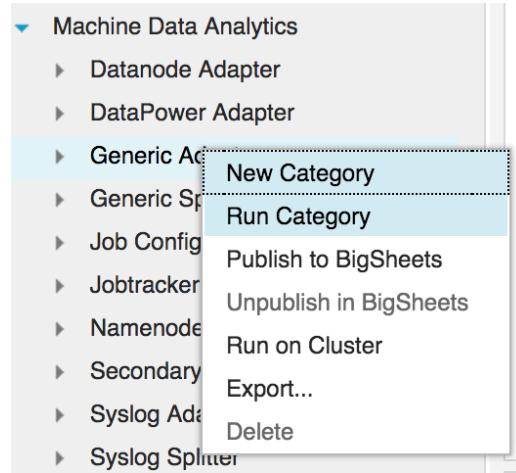
- Click on the Extractors tab in the left panel to bring the extractor library into view



An important feature of the new Web Tooling is an easy way to reuse extractors from the extensive library of pre-built extractors.

- Expand the Named Entity Recognition category to show the rich set of extractors.  
IBM will continue to add to this library and you can add your own extractors, as you will demonstrate later.

- Close the Named Entity Recognition category and open the Machine Data Analytics category. Since we're analyzing log files, there will be useful extractors in this category. Since we don't know what will match our Cisco log records, an easy way to get started is to run the category. Right click on the Generic Adapter category and select 'Run Category'



The Category will run and the extractors that match against the sample document will be automatically added to the canvas.

Document	span (Span)	text (String)	field_type (String)
CiscoLogs.txt	Aug 24 2007	Aug 24 2007	Date
CiscoLogs.txt	Aug 24 2007	Aug 24 2007	Date
CiscoLogs.txt	Aug 24 2007	Aug 24 2007	Date
CiscoLogs.txt	Aug 24 2007	Aug 24 2007	Date
CiscoLogs.txt	Aug 24 2007	Aug 24 2007	Date

Note how the default extractors have already matched parts of the document, with colour-coding to the dates and IP addresses.

- Minimize the properties panel (underneath the canvas) and click in the grid in the results panel to highlight one of the extracted spans. Note that clicking in the grid highlights that result in the document. Click on the tabs in the results panel to show the results from each of the extractors.

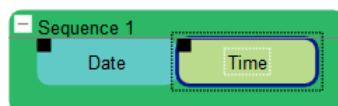
## Lab 4 Building a Sequence

### 4.1. Create the first Sequence

To match elements of log records and extract the IP addresses associated with suspicious activity we will build a sequence to match records like the first one in the document that include: an ASA-6 code and the OUTSIDE ACL:

```
Aug 24 2007 10:27:29: %ASA-6-106100: access-list OUTSIDE denied tcp  
outside/192.168.208.63(39675)-> inside/192.168.150.77(80) hit-cnt 1 first hit [0x22e8ac21, 0x0]
```

- Start the sequence by dragging the Date extractor further down toward the middle left of the canvas, then drag the Time extractor and dock it on the right of Date



- We can create new extractors for literals, dictionaries, regular expressions and token gaps. Looking at the first record in the document panel, notice that the next element in the log record is a colon. Match this by clicking on the literal button in the toolbar

On the new literal that was added to the canvas, type a colon, hit enter and then drag it and dock it to the right of Time



- Right click on sequence 1 and choose run selected from the context menu. You could also select the sequence and then hit the run button in the toolbar.

CiscoLogs.txt

```
Aug 24 2007 10:27:29 %ASA-6-106100 access-list  
OUTSIDE denied tcp outside/192.168.208.63(39675)->  
inside/192.168.150.77(80) hit-cnt 1 first hit [0x22e8ac21,  
0x0]  
Aug 24 2007 10:27:31 %ASA-6-106100 access-list  
OUTSIDE denied tcp outside/192.168.208.63(39676) ->  
inside/192.168.150.77(80) hit-cnt 1 first hit [0x22e8ac21,  
0x0]  
Aug 24 2007 10:27:22 %ASA-4-400014 IDS 2004 ICMP  
echo request from 192.168.208.63/15346 to  
192.168.150.70(80) on interface outside  
Aug 24 2007 10:27:22 %ASA-6-302020 Built ICMP  
connection for faddr 192.168.208.63/15343 gaddr  
192.168.150.70/0 laddr 192.168.150.70/0  
Aug 24 2007 10:27:22 %ASA-6-106015 Deny TCP (no  
connection) from 192.168.208.63/49827 to  
192.168.150.70/80 flags ACK on interface outside  
Aug 24 2007 10:27:22 %ASA-6-302020 Built ICMP  
connection for faddr 192.168.208.63/15343 gaddr  
192.168.150.70/0 laddr 192.168.150.70/0  
Aug 24 2007 10:27:22 %ASA-6-302015 Built inbound UDP  
connection 732748 for outside/192.168.208.63/49804 to  
inside/192.168.150.70/53
```

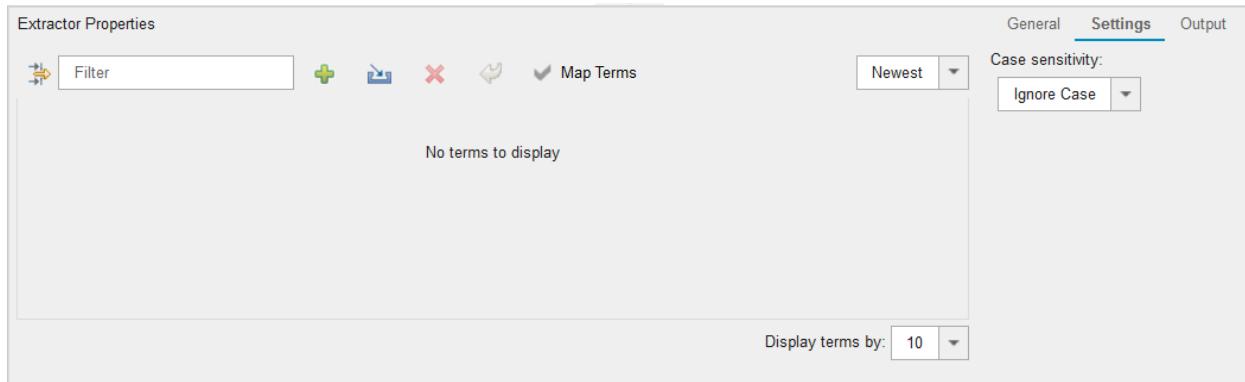
The highlighting now shows the matches, you can see the rapid iterative way of working that is a feature of this interface.

- The next element in the log record is the code %ASA-6-106100. We will use a dictionary to match this and any other codes that are caused by suspicious events.



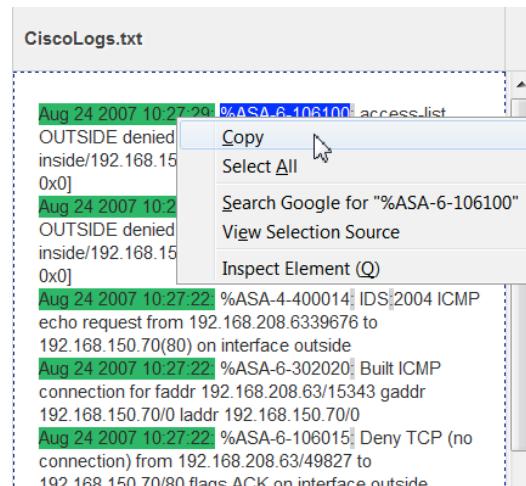
Click on the dictionary button in the toolbar, type Codes to name the new extractor and hit enter.

- Expand the extractor properties panel below the canvas.

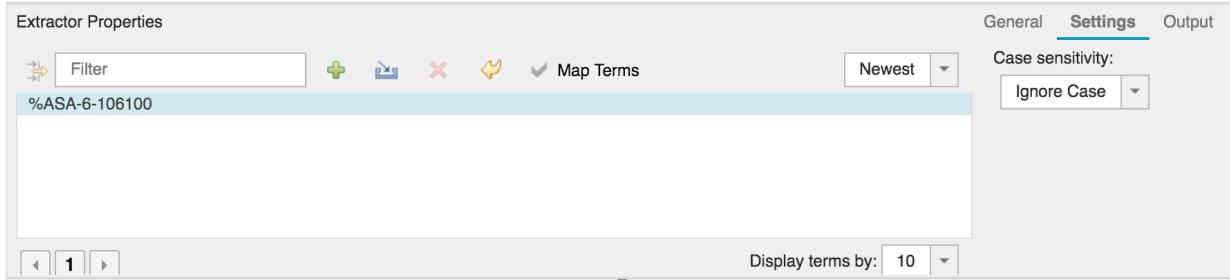


Look at the General tab where we could add more detail about this extractor (note that each of the pre-built extractors has extensive descriptive metadata). We will use the Output tab later when we need to define the output of our extractor.

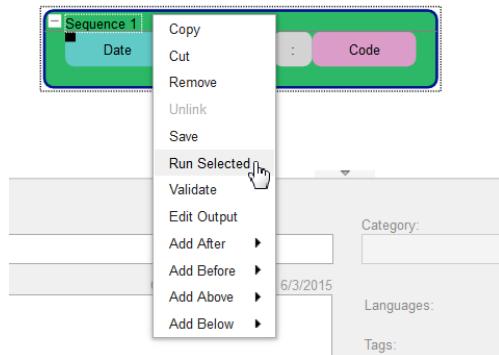
Click back to the Settings tab. We can enter dictionary terms directly or load them from a file. Select %ASA-6-106100 in the document panel and copy the text to the clipboard.



In the Properties panel select the + button, paste the text from the clipboard into the entry field and then hit enter.



- Drag the Code extractor and dock it to the right of the colon in the sequence, then highlight the sequence and either hit the run button on the toolbar or right click and choose run selected from the context menu.



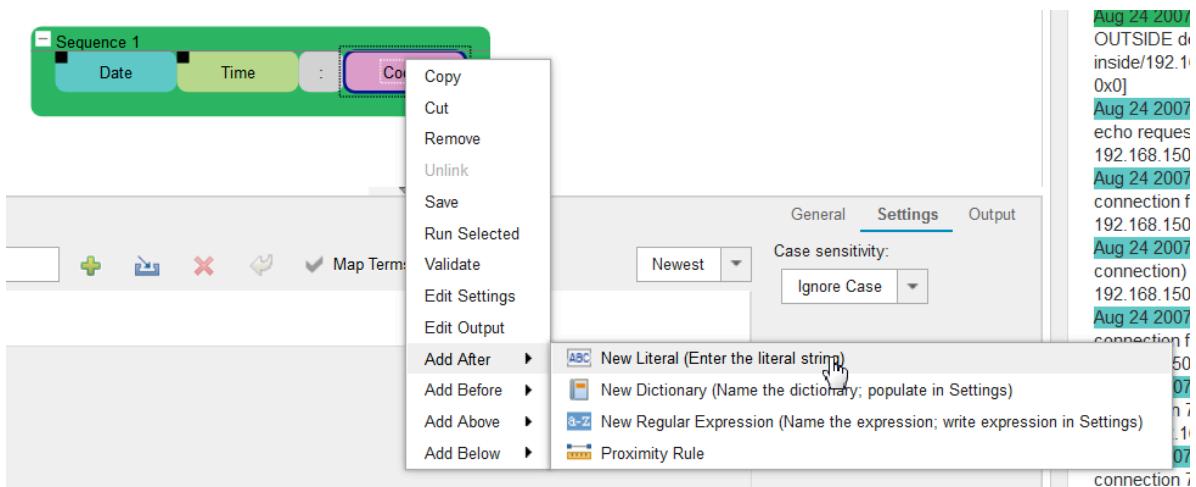
- Now we are only matching the records that contain the ASA-6-106100 code we just added to the dictionary:

```
CiscoLogs.txt

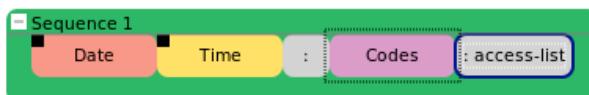
Aug 24 2007 10:27:29: %ASA-6-106100: access-list
OUTSIDE denied tcp outside/192.168.208.63(39675)->
inside/192.168.150.77(80) hit-cnt 1 first hit [0x22e8ac21,
0x0]
Aug 24 2007 10:27:31: %ASA-6-106100: access-list
OUTSIDE denied tcp outside/192.168.208.63(39676) ->
inside/192.168.150.77(80) hit-cnt 1 first hit [0x22e8ac21,
0x0]
Aug 24 2007 10:27:22: %ASA-4-400014: IDS|2004 ICMP
echo request from 192.168.208.63/1539676 to
192.168.150.70(80) on interface outside
Aug 24 2007 10:27:22: %ASA-6-302020: Built ICMP
connection for faddr 192.168.208.63/15343 gaddr
192.168.150.70/0 laddr 192.168.150.70/0
Aug 24 2007 10:27:22: %ASA-6-106015: Deny TCP (no
connection) from 192.168.208.63/49827 to
192.168.150.70/80 flags ACK on interface outside
Aug 24 2007 10:27:22: %ASA-6-302020: Built ICMP
connection for faddr 192.168.208.63/15343 gaddr
192.168.150.70/0 laddr 192.168.150.70/0
```

- Scroll down in the document view to show additional matches and click on the page 2 button at the bottom of the panel to show paging. Switch back to page 1.
- The next part of the sequence includes some text ': access-list' and a reference to the ACL - OUTSIDE. We will use a literal and a dictionary to match these elements.

Select and copy the text ': access-list' from the first record in the document panel. Right click on the Code extractor that's already in the sequence and select Add After->New Literal from the context menu.

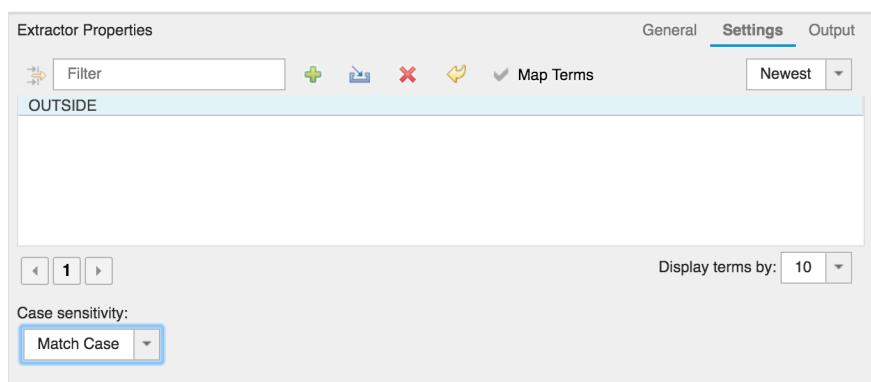


Paste the text you just copied and hit enter.



- Right click on the new literal you just added and from the context menu select Add After->New Dictionary. Type ACL as the name and hit enter. In the Properties panel type OUTSIDE in the entry field and hit enter.

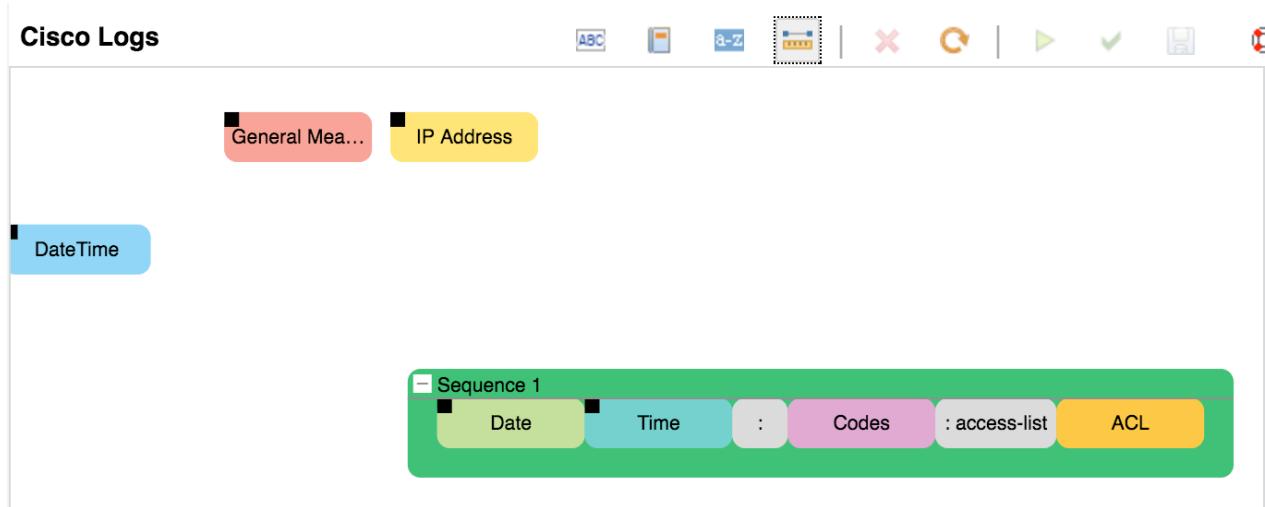
Change the Case sensitivity setting on the right to Match Case.



Select the full sequence on the canvas, run again and check the document panel to validate that you are matching the message through to the ACL OUTSIDE.

- The next element we're interested in is the TCPIP address. Between the ACL and the TCPIP address there's some text that we don't need to match exactly - 'denied tcp outside' and then a slash '/'. We can use a token gap (proximity rule) to match the text, a literal to match the slash and the pre-built to extract TCPIP address.

Click on the proximity rule button in the toolbar.



Type 1-5 (no spaces) on the extractor input box and hit enter. Dock the proximity rule to the right of ACL in the sequence. The proximity rule will match between 1 and 5 tokens. A token is either a consecutive sequence of 1 or more characters or a punctuation mark. The text in our log message 'denied tcp outside' contains 3 tokens. We could have specified 3-3, matching precisely this text, but allowing 1 - 5 tokens gives us some flexibility.

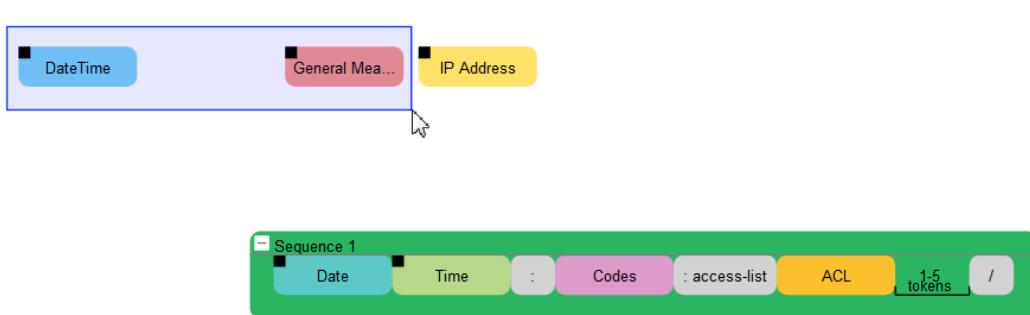
- Add a new literal, type a slash, hit enter and if you didn't use the 'Add After' method dock it to the right of the proximity rule.

Select the sequence, run it and check the results.



## 4.2. Completing the Sequence

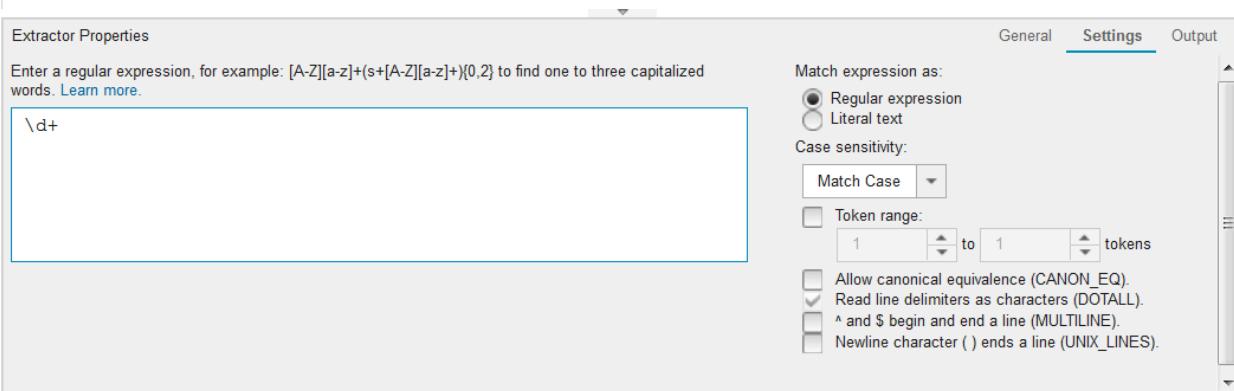
- Before we complete the sequence, tidy up a little by selecting and deleting the two unused pre-builtts that we don't need.



- Drag the IP address extractor and dock it to the right of the slash.



- Right click on the IP Address extractor, select Add After->New Literal, enter '(' character and hit enter.
- We will use a regular expression to match the port number. Right click on the literal you just created, select Add After-> New Regular Expression. Enter Port as the name, hit enter and in the Properties panel enter the regular expression \d+



- Complete the sequence by adding a literal for the ')' character.
- In the general properties for the sequence, change the name to 'Suspect IP Addresses'. Run the extractor and check the results. You will see full matches against the first two log records and in the grid you can see various elements being returned by the extractor

Results												
ACL (13)	Codes (8)	Date (43)	IP Address (94)	Port (893)	Suspect IP Addresses (8)	Time (47)						
Document	Suspect IP Address (Span)	span (Span)	span_1 (Span)	Literal_1 (Span)	Codes (Span)	Literal_2 (Span)	ACL (Span)	Literal_3 (Span)	span_2 (Span)	Literal_4 (Span)	Port (Span)	Literal_5 (Span)
CiscoLog: Aug 24 2007 10:27:29: %ASA-6-1 access-list OUTSIDE	Aug 24 2007	10:27:29	:	%ASA-6-1 : access-lis	OUTSIDE /	192.168.2 (	39675 )					

- We want the extractor to return just the IP address and port. We control this from the Output tab of the Properties. Select the Output tab and then click on the + to the left of the output column names and select Hide All Columns.

Extractor Properties

Select an extractor or structure and format your output into columns. [Learn more](#).

Output													
General	Settings	Output											
<input checked="" type="checkbox"/> Suspect IP Addresses <input checked="" type="checkbox"/> span <input checked="" type="checkbox"/> span_1 <input checked="" type="checkbox"/> Literal_1 <input checked="" type="checkbox"/> Codes <input checked="" type="checkbox"/> Literal_2 <input checked="" type="checkbox"/> ACL <input checked="" type="checkbox"/> Literal_3 <input checked="" type="checkbox"/> span_2 <input checked="" type="checkbox"/> Literal_4 <input checked="" type="checkbox"/> Port <input checked="" type="checkbox"/> Literal_5	<span>span</span> <span>Span</span>	<span>span_1</span> <span>Span</span>	<span>Literal_1</span> <span>Span</span>	<span>Codes</span> <span>Span</span>	<span>Literal_2</span> <span>Span</span>								
	<input type="checkbox"/> Manage overlapping matches Output column: Suspect IP Addresses Method: Contained Within												
Result													
43	IP Address (94)	Port (893)	Suspect IP Addresses (8)	Time (47)									
			span_1 (Span)	Literal_1 (Span)	Codes (Span)	Literal_2 (Span)	ACL (Span)	Literal_3 (Span)	span_2 (Span)	Literal_4 (Span)	Port (Span)	Literal_5 (Span)	
CiscoLog: Aug 24 2007 10:27:29: %ASA-6-1 access-list OUTSIDE	Aug 24 2007	10:27:29	:	%ASA-6-1 : access-lis	OUTSIDE /	192.168.2 (	39675 )						

Do Show All Columns Hide All Columns New Column Delete Column

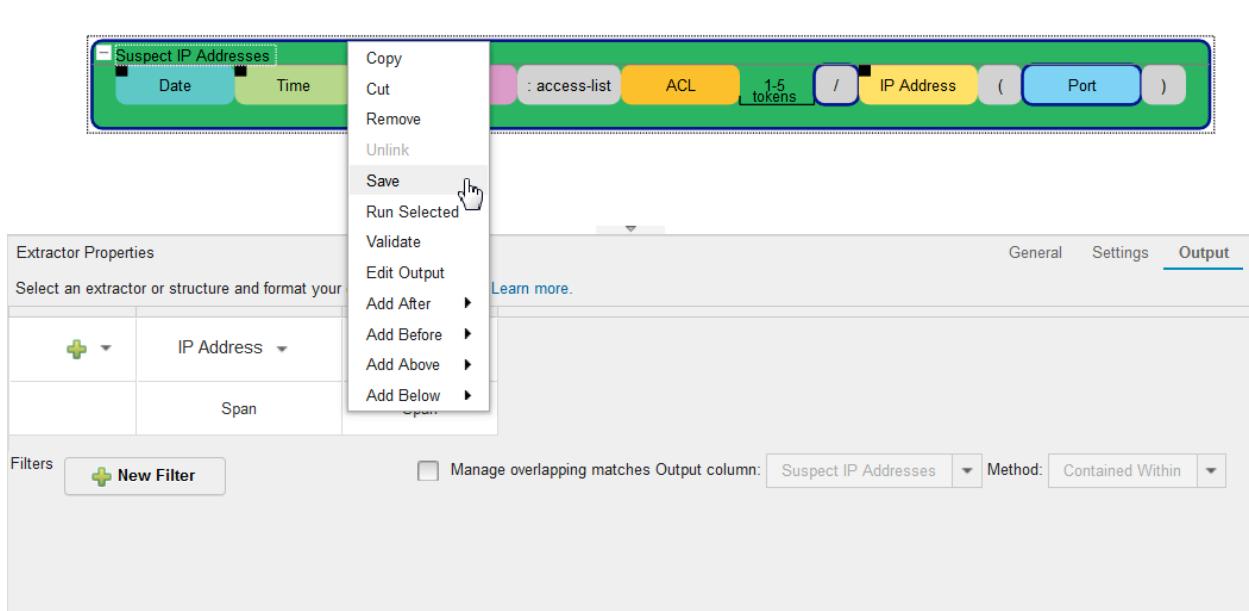
- Select span\_2 and Port - note that you should uncheck the first entry - Suspect IP Addresses since it was left as the default output by the hide all columns operation.
- In the output column names grid, select the menu on the span\_2 column and choose Rename.

The screenshot shows the 'Extractor Properties' window with the 'Output' tab selected. A context menu is open over the 'span\_2' column header, with 'Rename' highlighted. Other options in the menu include 'Edit Default Value', 'Choose Column', 'Convert To String', 'Convert To Lowercase String', 'Trim', 'New Column From Single Column...', and 'New Column From Two Columns...'. Below the menu, there are filter settings: 'Mapping matches Output column: Suspect IP Addresses' and 'Method: Contained Within'.

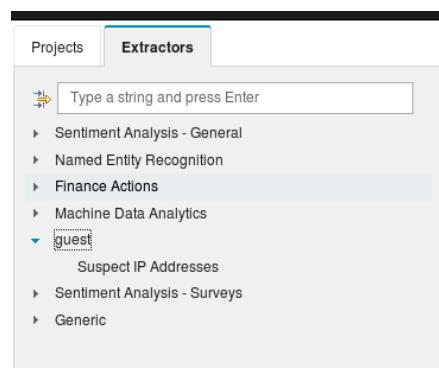
Type 'IP Address' and hit enter. Run the extractor and show the new results.

Results						
ACL (13)	Codes (8)	Date (43)	IP Address (94)	Port (893)	Suspect IP Addresses (8)	Time (47)
Document				IP Address (Span)		Port (Span)
<i>CiscoLogs.txt</i>				192.168.208.63		39675
<i>CiscoLogs.txt</i>				192.168.208.63		39676
<i>CiscoLogs.txt</i>				192.168.208.63		39675
<i>CiscoLogs.txt</i>				192.168.208.63		39676

- Finally, since this is a useful extractor we are likely to use again, we will save it to the library. Right click on the extractor and select Save from the context menu.



Select your userid (guest) and click OK in the dialog. You will now see your extractor in the library:



This is a good place to stop if you are running short of time. The only features you haven't shown are copy/paste with linked extractors and unions.

## Lab 5 Adding a second pattern

The data contains another log record for denied connections that we will add a second pattern to detect. Look down to the fifth record:

Aug 24 2007 10:27:22: %ASA-6-106015: Deny TCP (no connection) from 192.168.208.63/49827 to 192.168.150.70/80 flags ACK on interface outside

We can use another sequence pattern to detect these records and then union together both patterns to deliver a complete result set.

- Start by renaming the current extractor Suspect IP Addresses 1 in the General page of the Properties panel. Dismiss the warning message which tells you that you are making changes that will not be reflected in the copy of the extractor in the library.
- Copy the extractor and right click on the canvas. There are two paste options:
  - Paste will paste a linked copy of the extractor. Changes made to one will be automatically reflected in all linked copies, keeping them all consistent.
  - Paste as New Copy will paste a new, independent copy. New copy is just for the sequence (the internal extractor continue been a linked extractor).

We need an independent copy, so select Paste as New Copy and then rename it to Suspect IP Addresses 2.



- Looking at the fifth log record, notice that the first change we need to make is to add the code %ASA-6-106015 to the Codes dictionary. Select the dictionary in the new sequence and notice

that it's also selected in the first sequence. The extractors within the sequence are linked. Select **unlinked** in the second one (right click over extractor) and renamed with Codes 2

Copy the code %ASA-6-106015 from the fifth log record and on the Settings page of the Properties panel, click on the '+' button, paste the new code into the entry field and hit enter. Dismiss the warning message (in case that you get it) - note development are planning to give the user an option to suppress these messages after the first one.

- To modify the literal, change the literal text to ': Deny TCP'.
- Unlink and remove the ACL extractor from the sequence - right click and select remove from the context menu.
- Leave the proximity gap, drag the literal slash to the other side of IP Address and remove the literals for left and right parenthesis.



- Select and run the sequence. It may be difficult to see the results clearly in the documents view. To make it clearer, collapse the extractor by selecting the '-' to the left of the name and run it again. Now you will only see matches for the full sequence, not the contained elements.

Cisco Logs

Documents Total: 1

CiscoLogs.txt

```

Aug 24 2007 10:27:29: %ASA-6-106100:
access-list OUTSIDE denied tcp
outside/192.168.208.63(39675)->
inside/192.168.150.77(80) hit-cnt 1 first hit
[0x2de8ac21, 0x0]
Aug 24 2007 10:27:31: %ASA-6-106100:
access-list OUTSIDE denied tcp
outside/192.168.208.63(39675)->
inside/192.168.150.77(80) hit-cnt 1 first hit
[0x2de8ac21, 0x0]
Aug 24 2007 10:27:22: %ASA-4-100014:
IDS:2004 ICMP echo request from
192.168.208.63(99676) to 192.168.150.70(80)
on interface outside
Aug 24 2007 10:27:22: %ASA-6-302020: Built
ICMP connection for faddr
192.168.208.63/15343 gaddr 192.168.150.70/0
laddr 192.168.150.70/0
Aug 24 2007 10:27:22: %ASA-6-106015: Deny
TCP (no connection) from
192.168.208.63(49827) to 192.168.150.70/80
flags ACK on interface outside
Aug 24 2007 10:27:22: %ASA-6-302020: Built
ICMP connection

```

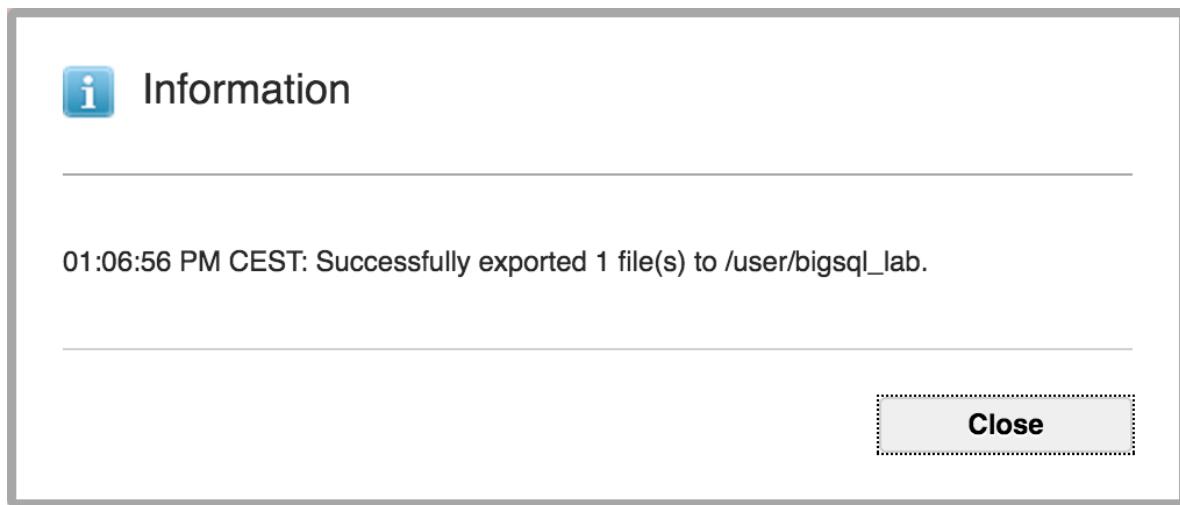
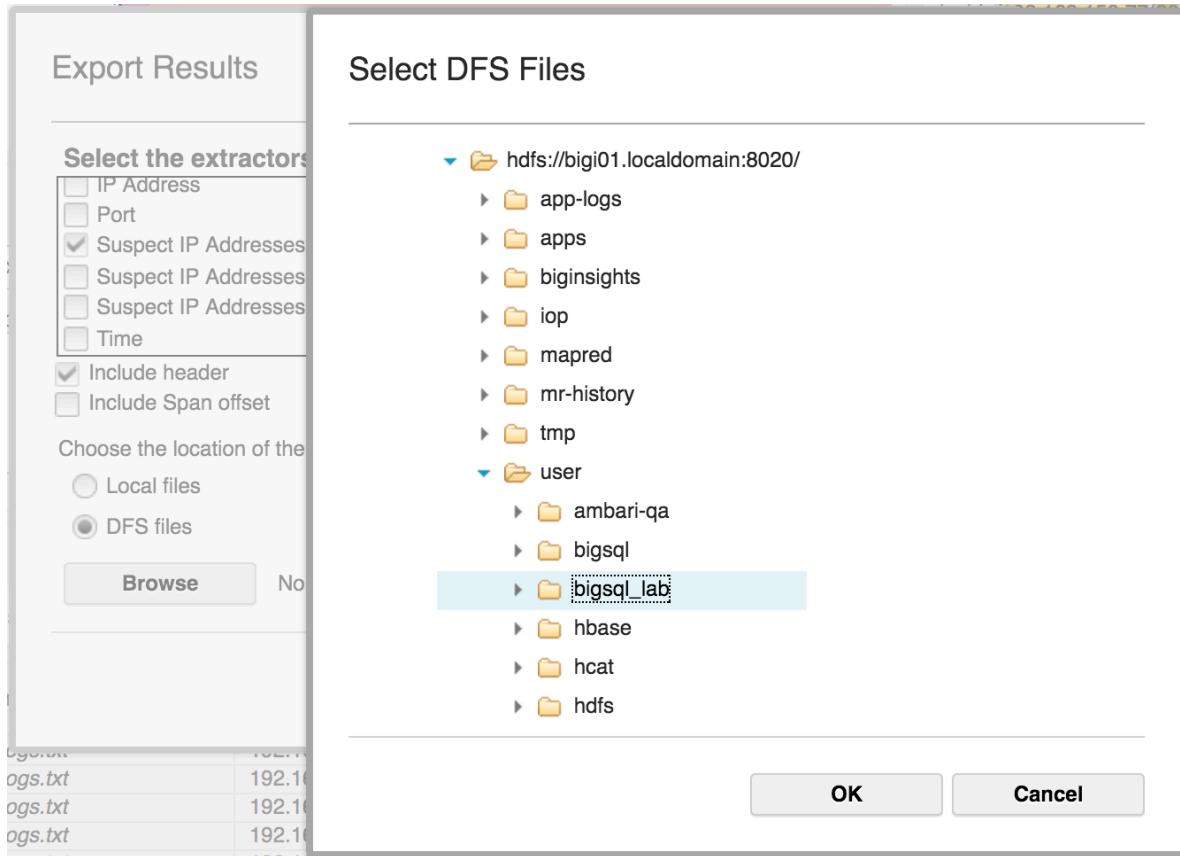
- Notice that the output of the new extractor includes a more fields. Go to the Output page of the Properties, select the + to the left of the output columns grad and uncheck the all except IP and Port
- To union together the two extractors, drag the new one and dock it under the first one:



- Run the union and check the results. You should see that all the suspicious activity in this log file is coming from one IP Address.
- Rename Union1 to Suspect IP Addresses and save it.
- Press in Suspect IP Addresses tab, deselect all, select Suspect IP Addresses, uncheck Include Span offset and save result in DFS path /user/bigsql\_lab:

The screenshot shows the Apache Nifi Results interface. The 'Suspect IP Addresses' tab is selected. The table has three columns: Document, IP Address (Span), and Port (Span). The data is as follows:

Document	IP Address (Span)	Port (Span)
CiscoLogs.txt	192.168.208.63	39675
CiscoLogs.txt	192.168.208.63	39676
CiscoLogs.txt	192.168.208.63	49827
CiscoLogs.txt	192.168.208.63	39675
CiscoLoas.txt	192.168.208.63	39676



## Lab 6 BigSheets

To help business analysts and those without a programming background analyze big data, IBM provides a spreadsheet-style tool called BigSheets. In this lab, you'll learn how you can explore big data through this tool without writing scripts or MapReduce applications.

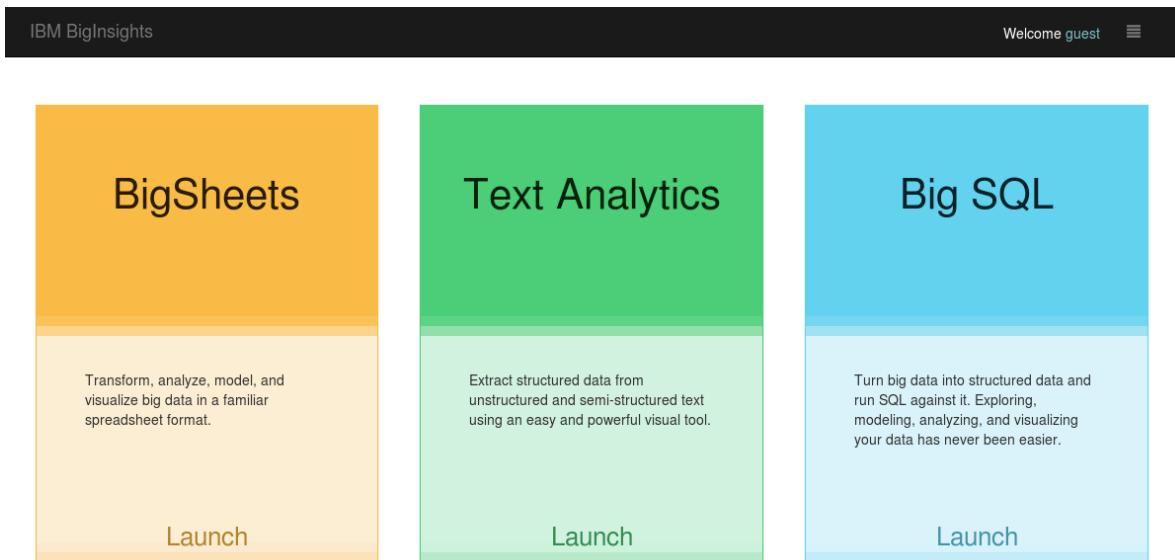
BigSheets is a browser-based analytic tool that you use to break large amounts of unstructured data into consumable, situation-specific business contexts.

Use BigSheets to create master workbooks from data files in your distributed file system.

After you collect data, you load data into a master workbook. Then, you format and explore the data by building sheets, which resemble spreadsheets, in workbooks that are based on the master workbook. You can combine columns from different workbooks, run formulas, and filter data. These manipulations form the basis of your analysis. You can also combine data with functions that are designed for BigInsights® text analytics to drill further into information and derive content out of raw data. These deep insights help you to filter and manipulate data from sheets even further.

Following the previous stored results, we are going to load the data on BigSheets and we are going to create a column bar chart.

- Login on BigSheets from BigInsights Home:



- BigSheets Home:

IBM BigInsights - BigSheets

Welcome guest

**Workbooks**

New Workbook Purge Import Workbook Metadata Export Workbook Metadata Manage Plugins

0-0 of 0 Page 1 of 1 Items per page: 20

View by type: all owner: all Sort by: recently created Enter text to filter Tags

Before you can explore and update your data, you must create one or more workbooks:  
 1: Click the New Workbook button.  
 2: Input a name and description for your new workbook.  
 3: Select a file from the list on the left.  
 4: Review the data in the preview area to verify that the data is formatted correctly. If not, select a new reader type. Click the green checkmark to apply the new reader.  
 5: Click on the green checkmark at the bottom of the dialog to create the new workbook.

- Click on New Workbook and fill up name and description and choose DFS File on /user/bigsql\_lab/Suspect\_IP\_Addresses.csv

**New Workbook**

Name: Suspect IP Addresses  
Description: Suspect IP Addresses

**DFS Files** Catalog Tables

/user/bigsql\_lab/Suspect\_IP\_Addresses.csv

Line Reader

Ready

	Header
1	Document,IP Address,Port
2	CiscoLogs.txt,192.168.208.63,39675
3	CiscoLogs.txt,192.168.208.63,39676
4	CiscoLogs.txt,192.168.208.63,39675
5	CiscoLogs.txt,192.168.208.63,39676
6	CiscoLogs.txt,192.168.208.63,51609
7	CiscoLogs.txt,192.168.208.63,51610
8	CiscoLogs.txt,192.168.208.63,52978
9	CiscoLogs.txt,192.168.208.63,52981
10	CiscoLogs.txt,192.168.208.63,49827
11	
12	
13	
14	
15	
16	
17	
18	
19	
...	

Refresh Fit column(s)

- As you can see the reader is not the correct one, select CSV reader:

/user/bigsql\_lab/Suspect\_IP\_Addresses.csv

Line Reader

Ready

Select a reader: Comma Separated Value (CSV) Data

Reads data in CSV format

Fill in parameters:  
Headers Included?

/user/bigsql\_lab/Suspect\_IP\_Addresses.csv  
Comma Separated Value (CSV) Data

Ready

	Document	IP_Address	Port
1	CiscoLogs.txt	192.168.208.63	39675
2	CiscoLogs.txt	192.168.208.63	39676
3	CiscoLogs.txt	192.168.208.63	39675
4	CiscoLogs.txt	192.168.208.63	39676
5	CiscoLogs.txt	192.168.208.63	51609
6	CiscoLogs.txt	192.168.208.63	51610
7	CiscoLogs.txt	192.168.208.63	52978
8	CiscoLogs.txt	192.168.208.63	52981
9	CiscoLogs.txt	192.168.208.63	49827
10			
11			
12			
13			
14			
15			
16			
17			
18			
19			
~~			



- Save the workbook
- Review how BigSheets can extract ColumnType

Suspect IP Addresses

Ready

	Document	IP_Address	Port
1	CiscoLogs.txt	192.168.208.63	39675
2	CiscoLogs.txt	192.168.208.63	39676
3	CiscoLogs.txt	192.168.208.63	39675
4	CiscoLogs.txt	192.168.208.63	39676
5	CiscoLogs.txt	192.168.208.63	51609
6	CiscoLogs.txt	192.168.208.63	51610
7	CiscoLogs.txt	192.168.208.63	52978
8	CiscoLogs.txt	192.168.208.63	52981
9	CiscoLogs.txt	192.168.208.63	49827
10			
11			

ColumnType

Integer  
Long integer  
Floating-point  
Double-precision floating-point  
BigInteger  
BigDecimal (precision 38)  
String  
DateTime  
Boolean

- Set Port column as String type

Port

39675  
39676  
39675  
39676  
51609  
51610  
52978  
52981  
49827

ColumnType

Integer  
Long integer  
Floating-point  
Double-precision floating-point  
BigInteger  
BigDecimal (precision 38)  
**String**  
DateTime  
Boolean

- Review Workbook details at bottom of the page

▼ Details:

Description:	Suspect IP Addresses 
Tags:	
Reader:	Comma Separated Value (CSV) Data 
Source:	Network: /user/bigsql_lab/Suspect_IP_Addresses.csv 
Data type hint:	none 
Sharing:	Private 

- Click on Build new workbook

Workbooks > View Results

**Suspect IP Addresses** 

 Delete |  Add chart | Suspect IP Add... :  

Ready  Refresh  Fit column(s)  Create Table

	Document	IP_Address	Port
1	CiscoLogs.txt	192.168.208.63	39675
2	CiscoLogs.txt	192.168.208.63	39676
3	CiscoLogs.txt	192.168.208.63	39675
4	CiscoLogs.txt	192.168.208.63	39676
5	CiscoLogs.txt	192.168.208.63	51609
6	CiscoLogs.txt	192.168.208.63	51610
7	CiscoLogs.txt	192.168.208.63	52978
8	CiscoLogs.txt	192.168.208.63	52981
9	CiscoLogs.txt	192.168.208.63	49827

- Edit name

**Suspect IP Addresses(1)**

	A	B	
	Document	IP_Address	
1	CiscoLogs.txt	192.168.208.63	39675
2	CiscoLogs.txt	192.168.208.63	39676
3	CiscoLogs.txt	192.168.208.63	39675
4	CiscoLogs.txt	192.168.208.63	39676
5	CiscoLogs.txt	192.168.208.63	51609
6	CiscoLogs.txt	192.168.208.63	51610
7	CiscoLogs.txt	192.168.208.63	52978
8	CiscoLogs.txt	192.168.208.63	52981
9	CiscoLogs.txt	192.168.208.63	49827
10			

- Click on Document column menu and Remove it

**Review Suspect IP Addresses**

	A	B	C
	Document	IP_Address	Port
1	CiscoLogs.txt	192.168.208.63	39675
2	CiscoLogs.txt	192.168.208.63	39676
3	CiscoLogs.txt	192.168.208.63	39675
4	CiscoLogs.txt	192.168.208.63	39676
5	CiscoLogs.txt	192.168.208.63	51609
6	CiscoLogs.txt	192.168.208.63	51610
7	CiscoLogs.txt	192.168.208.63	52978
8	CiscoLogs.txt	192.168.208.63	52981
9	CiscoLogs.txt	192.168.208.63	49827
10			

- Click on Add sheet > Function

Workbooks > View Results > Create

Review Suspect IP Addresses

Save Exit Add sheets

fx

1	192.168.208
2	192.168.208
3	192.168.208
4	192.168.208
5	192.168.208
6	192.168.208
7	192.168.208
8	192.168.208
9	192.168.208
10	
11	
12	
13	
14	
15	
16	

Select a type of sheet:

- Filter
- Function
- Load
- Group
- Join
- Union
- Intersection
- Complement
- Limit
- Distinct
- Copy
- Formula

- Search concat function

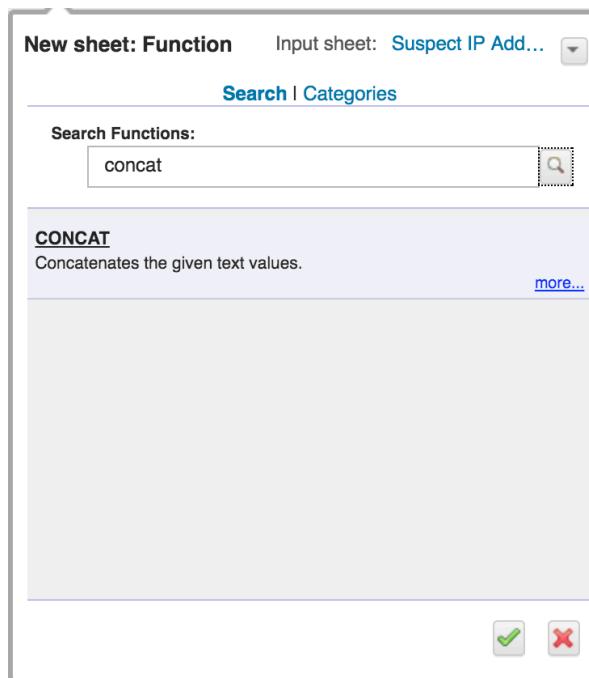
New sheet: Function Input sheet: Suspect IP Add...

Search | Categories

Search Functions: concat

✓ ✘

- Click on Concat function



- Set name and select the right column and write ':' as delimiter and press 

**New sheet: Function**    Input sheet: Suspect IP Add...

\* Sheet Name:

**CONCAT**  Concatenates the given text values.

Fill in parameters:

text1*	<input type="text" value="IP_Address"/>
text2*	<input type="text" value=":"/>
text3	<input type="text" value="Port"/>

Parameters Carry over (0)

- Click on Add sheet > Group

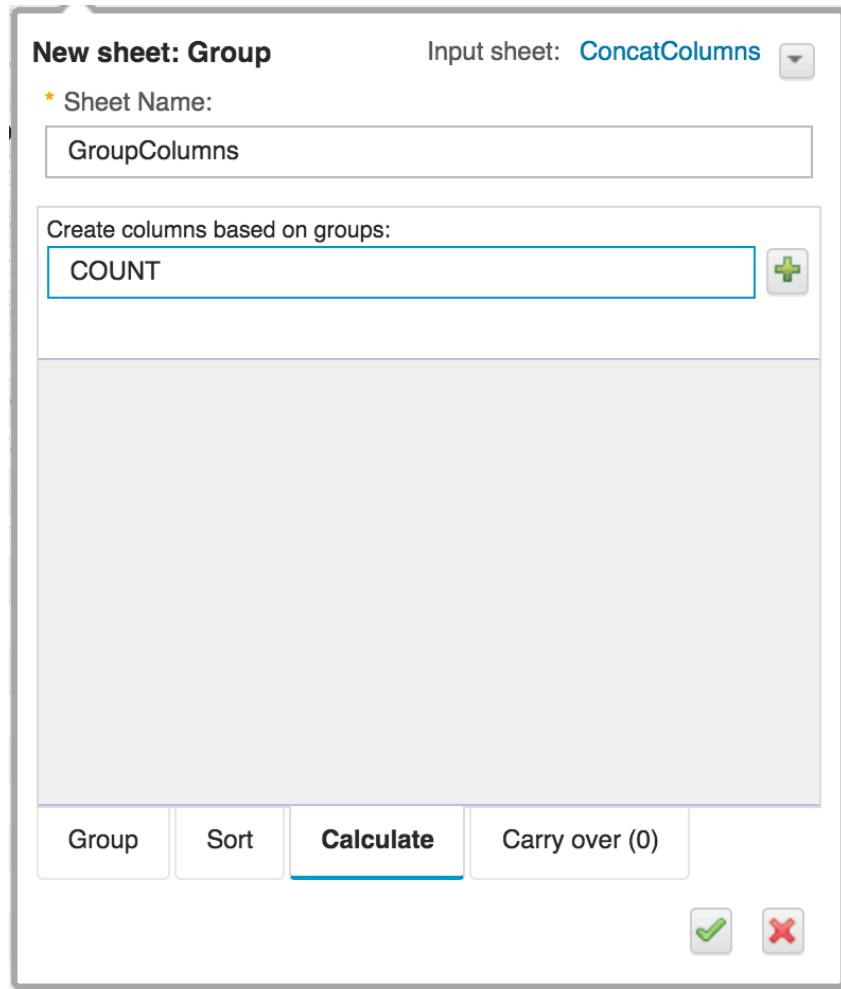
The screenshot shows a workspace titled "Review Suspect IP Addresses". A context menu is open at the top right, with the "Add sheets" option selected. A sub-menu titled "Select a type of sheet:" lists various sheet types with corresponding icons:

- Filter
- Function
- Load
- Group** (selected)
- Join
- Union
- Intersection
- Complement
- Limit
- Distinct
- Copy
- Formula

- Set name and click on Add all

The screenshot shows the "New sheet: Group" configuration dialog. The "Sheet Name" field is set to "GroupColumns". The "Group by columns:" section is empty. The "CONCAT" section contains a single row labeled "CONCAT". At the bottom, there are tabs for "Group", "Sort", "Calculate", and "Carry over (0)", with "Calculate" being the active tab. There are also "Check" and "Cancel" buttons at the bottom right.

- Click on Calculate tab and set COUNT column name a press '+'



- Select COUNT operation and select CONCAT column

New sheet: Group      Input sheet: **ConcatColumns**

\* Sheet Name:  
GroupColumns

Create columns based on groups:

COUNT = COUNT

Fill in parameters:  
Column: CONCAT

Group    Sort    **Calculate**    Carry over (0)

- Press 

- Review the results

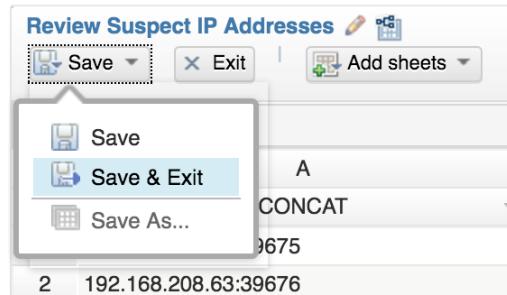
Workbooks > View Results > **Create**

**Review Suspect IP Addresses**  

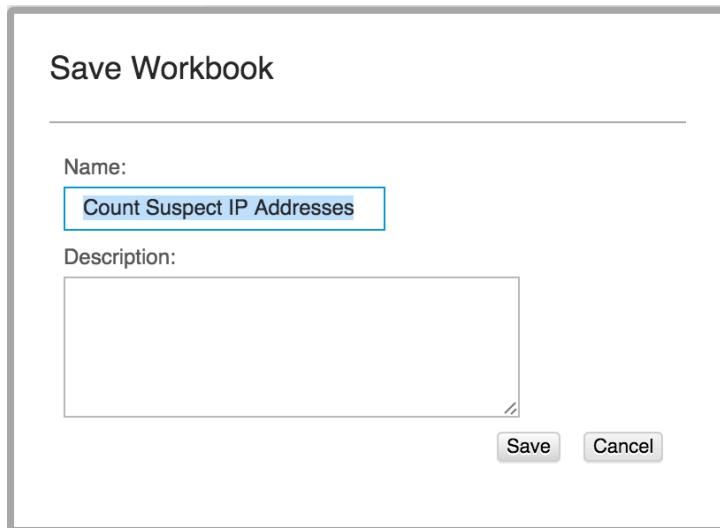
Save  Exit  Add sheets 

	A	B
	CONCAT	COUNT
1	192.168.208.63:39675	2
2	192.168.208.63:39676	2
3	192.168.208.63:49827	1
4	192.168.208.63:51609	1
5	192.168.208.63:51610	1
6	192.168.208.63:52978	1
7	192.168.208.63:52981	1

- Save & Exit



- Set name



- Press Run (this will run your flow over the data)

Workbooks > View Results

**Count Suspect IP Addresses**

Click run to update the data

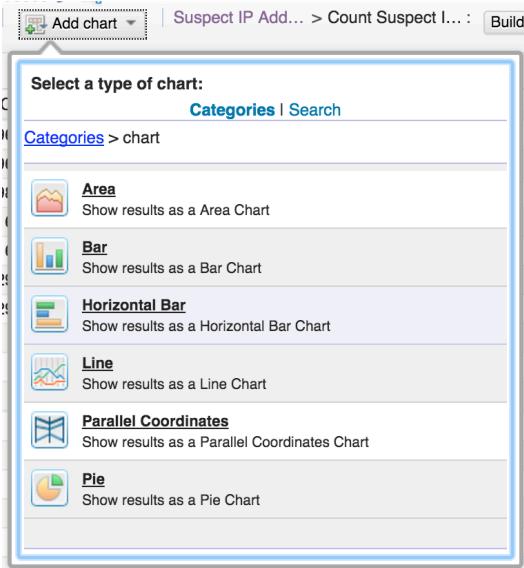
CONCAT	COUNT
1 192.168.208.63:39675	2
2 192.168.208.63:39676	2
3 192.168.208.63:49827	1
4 192.168.208.63:51609	1
5 192.168.208.63:51610	1
6 192.168.208.63:52978	1
7 192.168.208.63:52981	1
8	
9	
10	
11	
12	

This workbook has never been run. Press **Run** to run it or **Close** to dismiss this message.

- If you press over the progress bar, you will see the statistics. You can press on More link to review the process



- Press Add chart and select Horizontal Bar



- Set the values and press 

**Edit chart: Horizontal Bar**

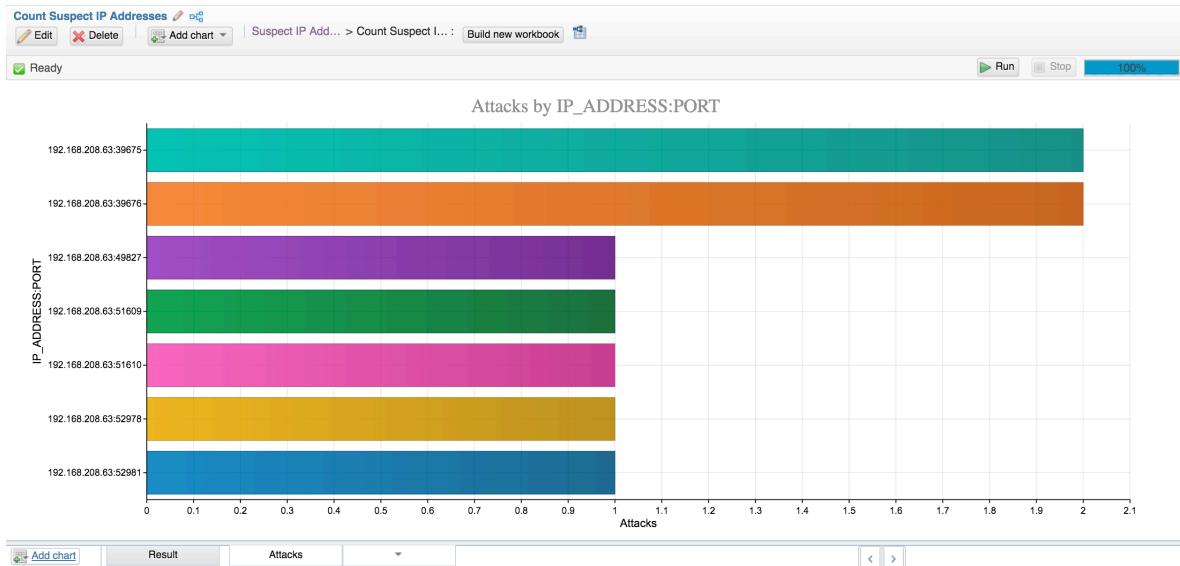
Chart Name:	Attacks
Title:	Attacks by IP_ADDRESS:PORT
X Axis:	COUNT
X Axis Label:	Attacks
Y Axis:	CONCAT
Y Axis Label:	IP_ADDRESS:PORT
Sort By:	Y Axis
Occurrence Order:	Ascending
Limit:	20
Template:	Smart Planet
Style:	Stacked

- Press Run



- Review results



- Click on Workbooks link

**Count Suspect IP Addresses**

Icon	Name	Description	Owner	Created	Last visited	Progress
Document icon	Count Suspect IP Addresses	No description	Owner: guest	Created: 02/06/2017 18:19	Last visited: 02/06/2017 18:44	Progress: 100%
Document icon	Suspect IP Addresses	Analyse suspect IP addresses	Owner: guest	Created: 02/06/2017 15:14	Last visited: 02/06/2017 18:14	Progress: 100%

- Check that you have stored your workbooks

Icon	Name	Description	Owner	Created	Last visited	Progress
Document icon	Count Suspect IP Addresses	No description	Owner: guest	Created: 02/06/2017 18:19	Last visited: 02/06/2017 18:44	Progress: 100%
Document icon	Suspect IP Addresses	Analyse suspect IP addresses	Owner: guest	Created: 02/06/2017 15:14	Last visited: 02/06/2017 18:14	Progress: 100%

- If you press on you can review the hierarchy of the workbooks and review you developed work

