TAP Hands-On Workshop

Module 1: Performing Analytics on Your Data

Overview

This is module 1 of a two part workshop. This module uses the TAP Analytics Toolkit (TAP Analytics Toolkit) within the Trusted Analytics Platform (TAP) to execute a script that reads data from a TAP server and creates and outputs an amended dataset. The amended dataset becomes the input for Module 2: Visualizing Your Data with an App.

Estimated Length

This module will take approximately 1 hour to complete.

Setup

You will need to prepare your local environment/computer by installing software tools. The software you need includes:

- The software you need includes:
  - **Jupyter Notebook** (also known as Jupyter) - a tool for exploratory computation and data analysis that stores inputs/outputs in notebook documents. http://jupyter.readthedocs.io/en/latest/install.html
  - **Python (2.7)** - a widely used general-purpose, high-level programming language. https://www.python.org/downloads/

Objectives

After completing this module, you will have:

- Accessed Marketplace and TAP.
- Demonstrated the ability to create an instance of TAP Analytics Toolkit on your system.
- Used Python or Jupyter Notebook.
- Accessed the help functions for editing a script**.**
- Shown how to attach a TAP server to your project to input data.
- Executed an TAP Analytics Toolkit script that exports data.
- Demonstrated the ability to install multiple software tools needed to build and operate domain-specific applications.

Module Steps

1. Install all necessary developer tools listed in Setup.
2. Sign into Marketplace, view the Data Catalog functions and add a file into TAP.
3. Access Marketplace to create a new instance of TAP Analytics Toolkit on your system, view the running instance and copy the instance uri to use in the Netflow demo.
4. Netflow demo? Open the Netflow demo file in Python or Jupyter Notebook, review the help functions and edit the TAP Analytics Toolkit.server.uri field in the file.
5. Execute the Netflow demo to deploy the TAP Analytics Toolkit model. The demo first clears out old frames and graphs. Then it reads in the data, formats and prints the date/time data in columns, builds a graph for computing statistics, computes weighted and unweighted degree counts, computes a summary frame and downloads it to Pandas, computes and plots histograms, makes scatter plots, creates and trains an Support Vector Machine (SVM) model and downloads a scatter frame to Pandas.
6. Export the data to Hadoop Distributed File System (HDFS) as a comma separated value (CSV) file.

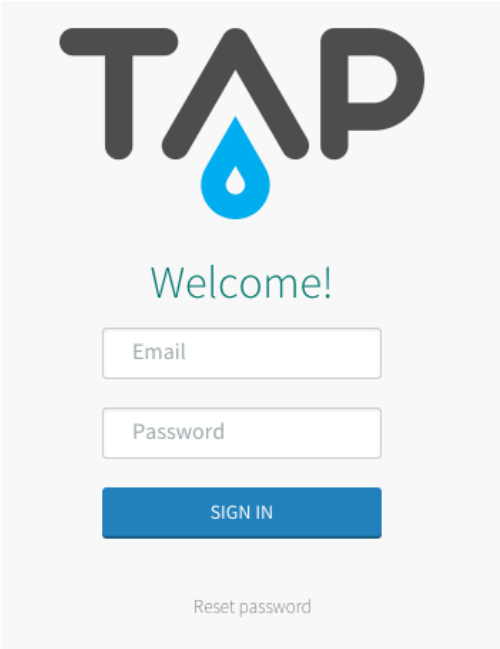
Important: The URLs and input commands provided in the workshop will change based on the instance used and the setup of your network. Check with the workshop instructor or consult with your development operations team (DevOps) to obtain the proper information.

Results

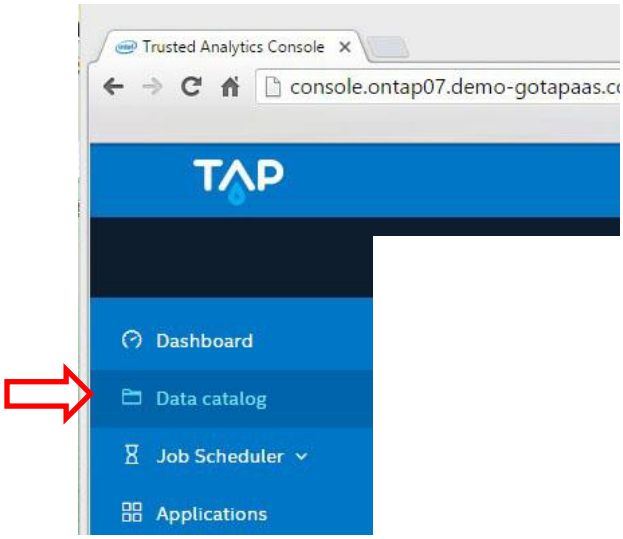
This section provides step-by-step instructions and their results.

1. **Install all necessary developer tools listed in Setup**
   1.1. No additional instructions are provided for this step.

2. **TAP Marketplace**
   2.1. Sign into TAP with your User ID and Password:

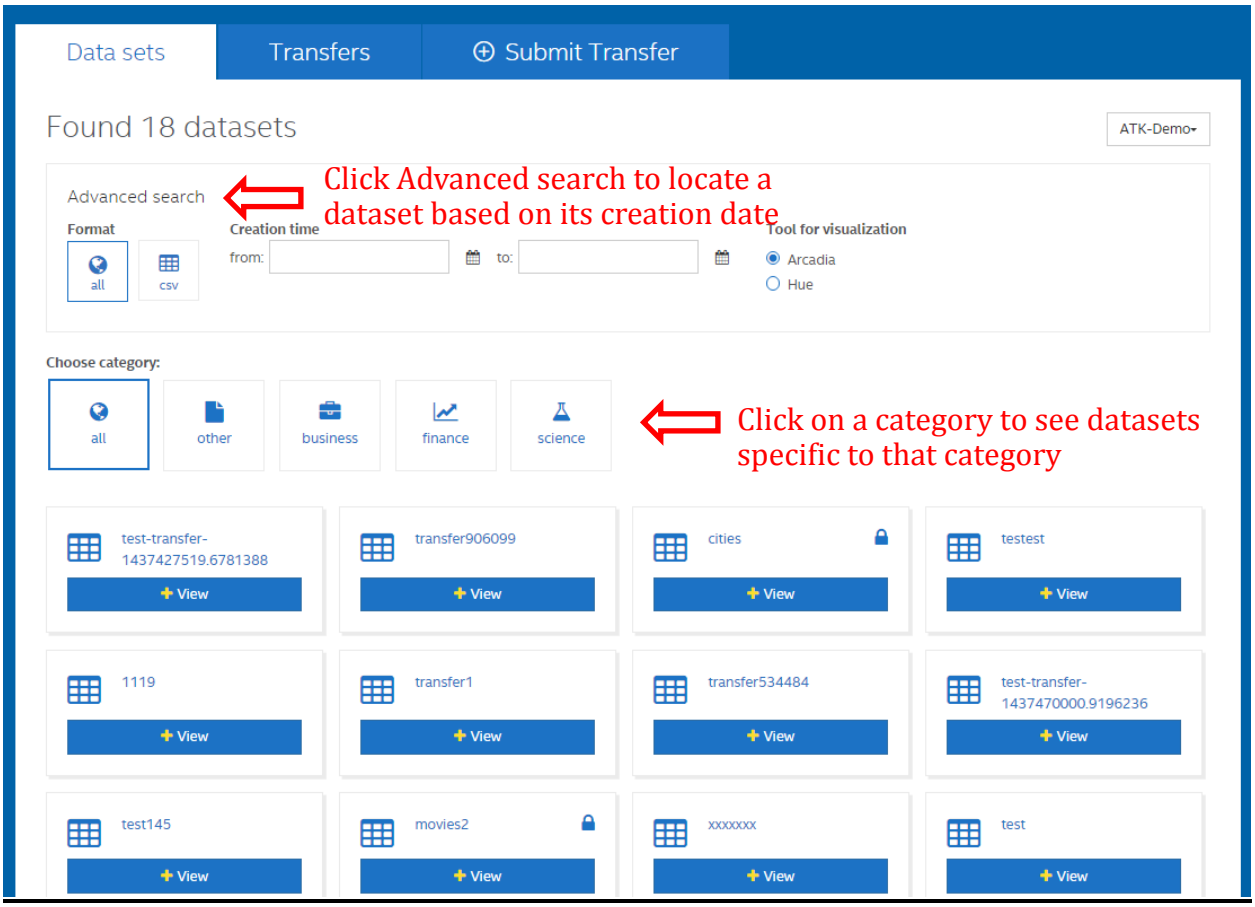


   2.2. View the Data Catalog functions:
   - Click on Data Catalog under Main navigation to access the three Data Catalog tabs – Data sets, Transfers and Submit Transfer.

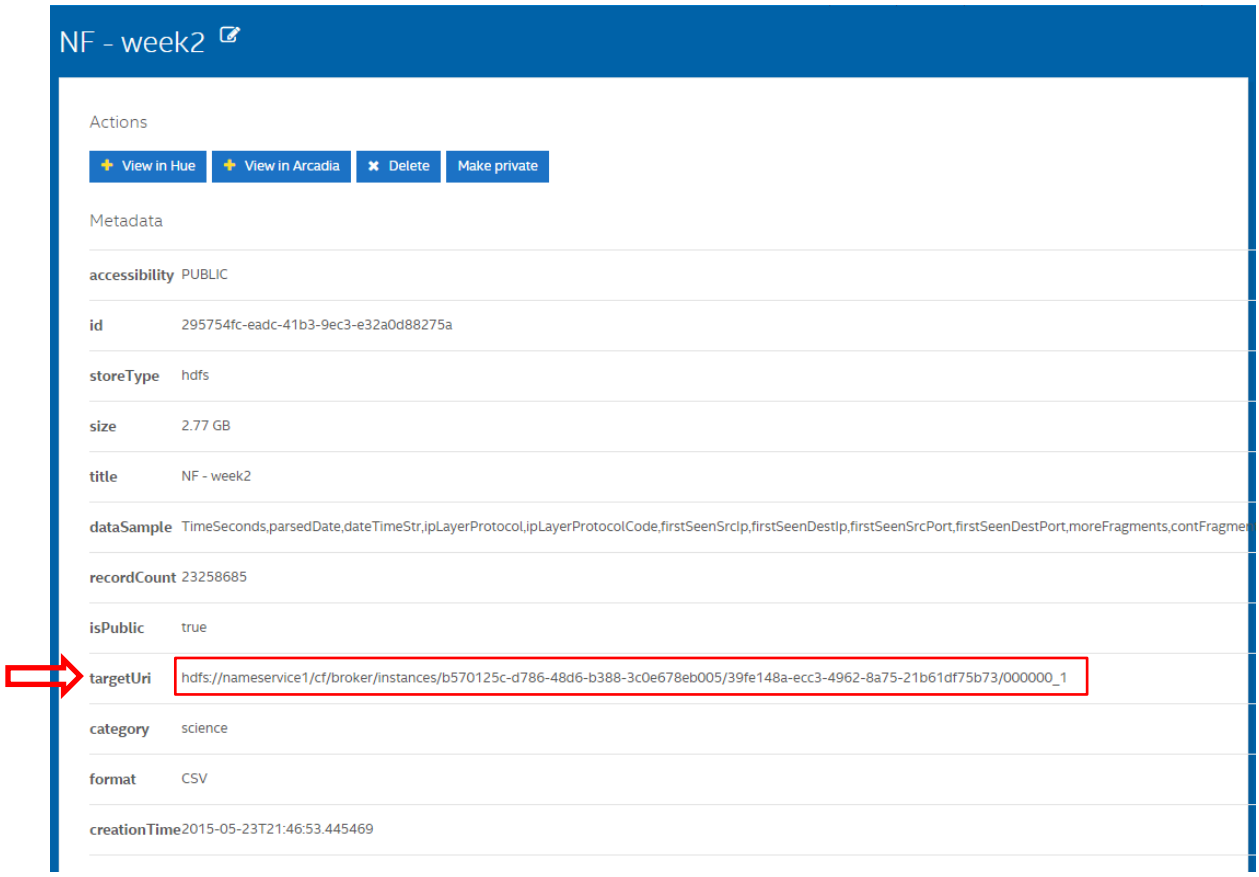


   - Click on the **Data sets** tab to see the datasets on TAP:

- If you have any data sets, you will see them listed on the first tab along with any data sets marked public. Use 'Advanced Search' or a category to limit the number of data sets you see.
- In this example, we are in the Demo organization in Marketplace. The Demo organization is only available in this TAP instance. The organization name may be different if you use your own instance.
- **Note**: Make sure you select the correct organization when you look for your dataset in Marketplace.

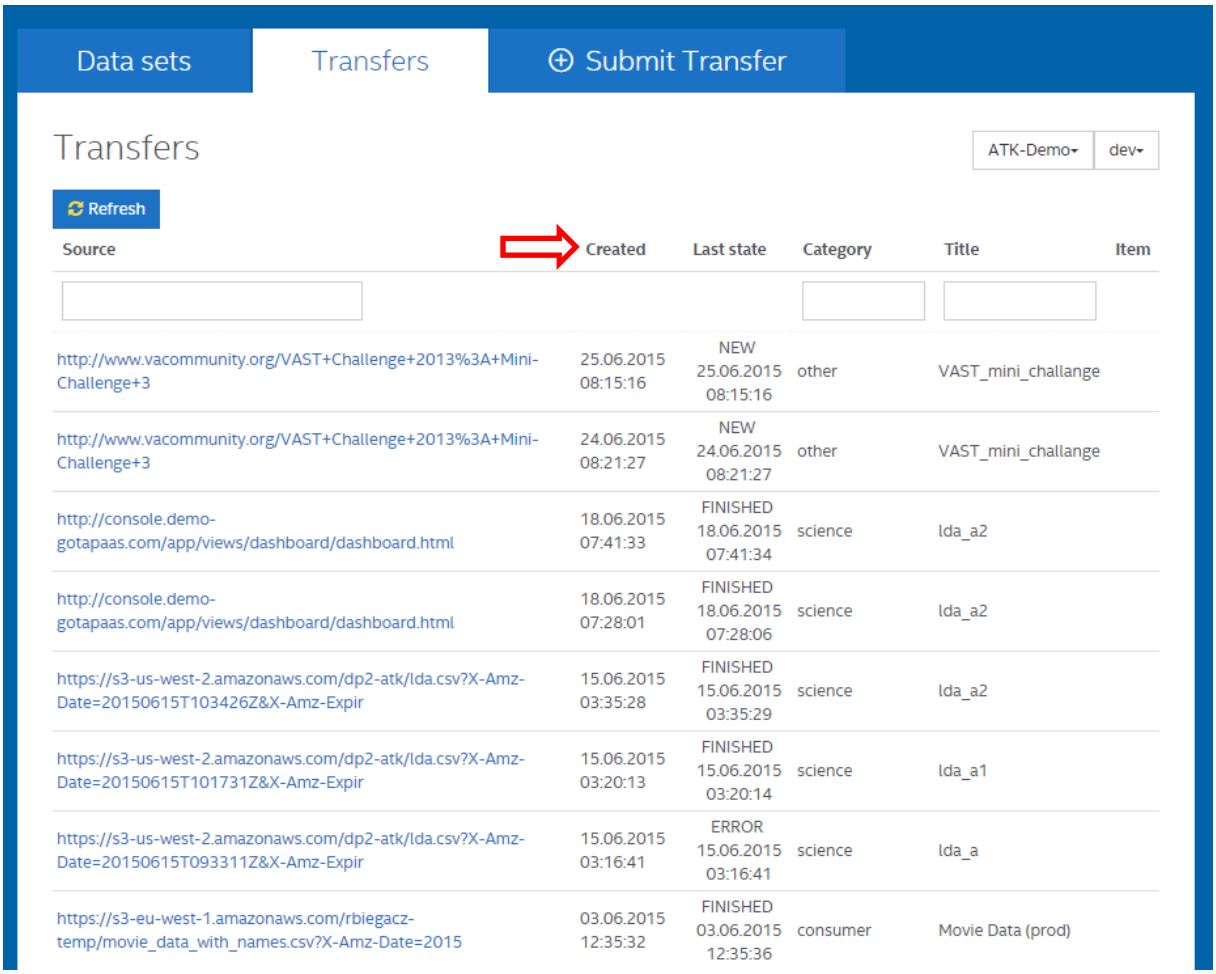


- Click the **Dataset Title** on any data set to display dataset information, such as:
  - Title, size, category …
  - **TargetUri** is a link to a dataset. The dataset link shown may not be the one needed in this exercise.
  - You can copy this link to use in other apps.



Note - Clicking the 'View' button will prompt the data set to be published in Hue through a pop up.

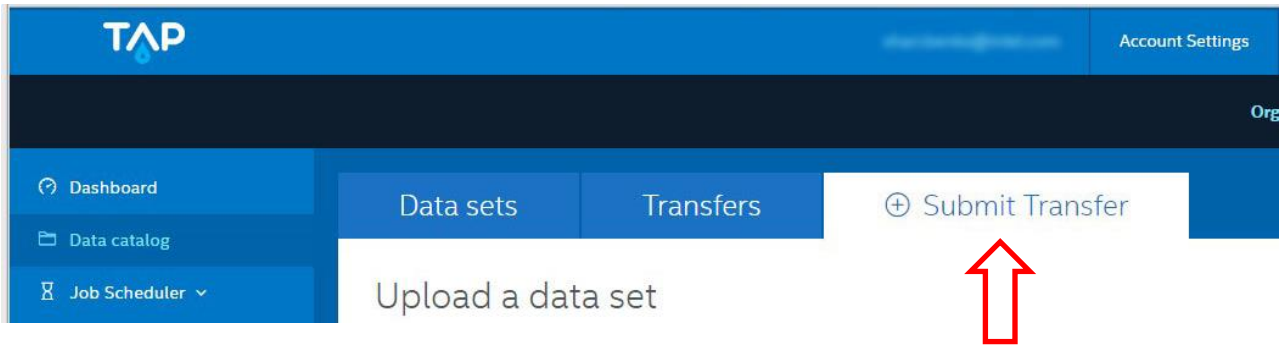
- Click on the **Data Catalog** tab.
- Click on the **Transfers** tab to see what files have been uploaded to TAP and to get information on when a file was added to the system.



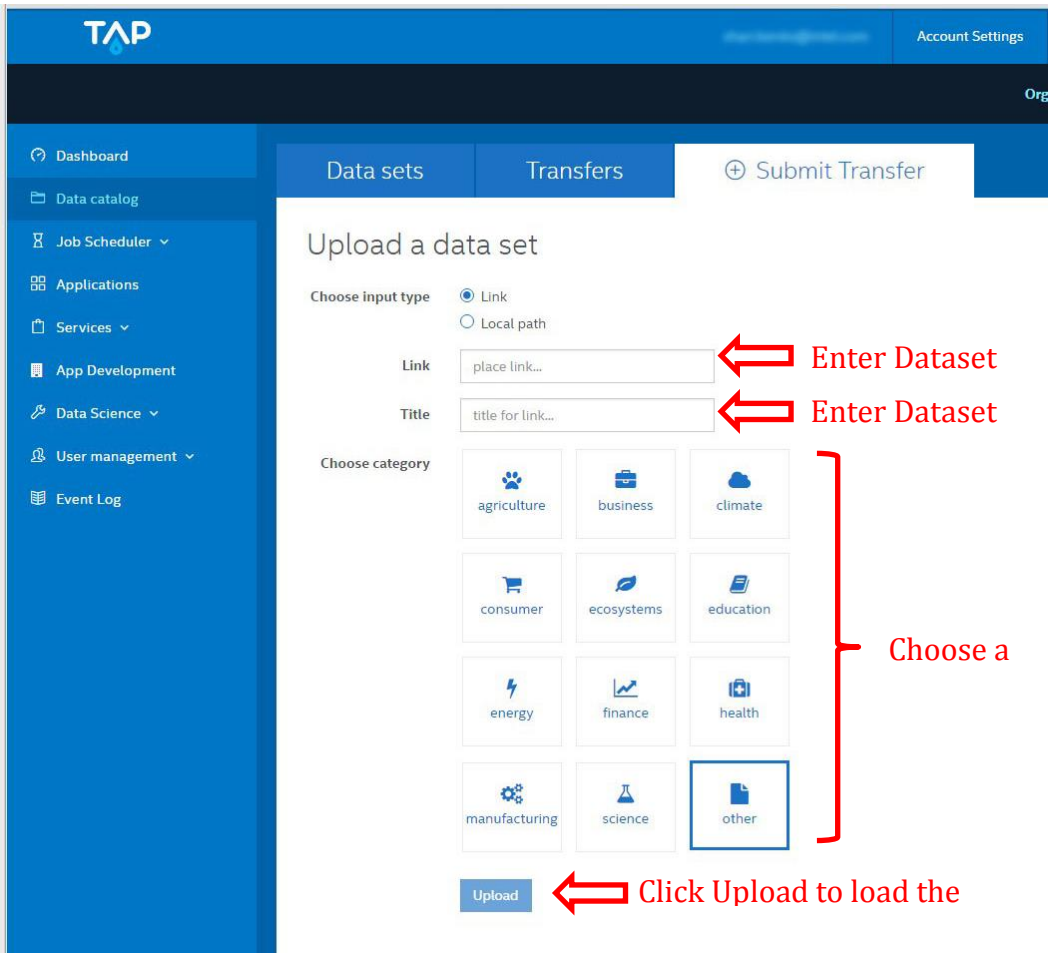
Source	Created	Last state	Category	Title	Item
http://www.vacommunity.org/VAST+Challenge+2013%3A+Mini-Challenge+3	25.06.2015 08:15:16	NEW 25.06.2015 08:15:16	other	VAST_mini_challange	
http://www.vacommunity.org/VAST+Challenge+2013%3A+Mini-Challenge+3	24.06.2015 08:21:27	NEW 24.06.2015 08:21:27	other	VAST_mini_challange	
http://console.demo-gotapaas.com/app/views/dashboard/dashboard.html	18.06.2015 07:41:33	FINISHED 18.06.2015 07:41:34	science	lda_a2	
http://console.demo-gotapaas.com/app/views/dashboard/dashboard.html	18.06.2015 07:28:01	FINISHED 18.06.2015 07:28:06	science	lda_a2	
https://s3-us-west-2.amazonaws.com/dp2-atk/lda.csv?X-Amz-Date=20150615T103426Z&X-Amz-Expir	15.06.2015 03:35:28	FINISHED 15.06.2015 03:35:29	science	lda_a2	
https://s3-us-west-2.amazonaws.com/dp2-atk/lda.csv?X-Amz-Date=20150615T101731Z&X-Amz-Expir	15.06.2015 03:20:13	FINISHED 15.06.2015 03:20:14	science	lda_a1	
https://s3-us-west-2.amazonaws.com/dp2-atk/lda.csv?X-Amz-Date=20150615T093311Z&X-Amz-Expir	15.06.2015 03:16:41	ERROR 15.06.2015 03:16:41	science	lda_a	
https://s3-eu-west-1.amazonaws.com/rbiegacz-temp/movie_data_with_names.csv?X-Amz-Date=2015	03.06.2015 12:35:32	FINISHED 03.06.2015 12:35:36	consumer	Movie Data (prod)	

2.3. To add a file into TAP:

- Click the **Submit Transfer** tab.



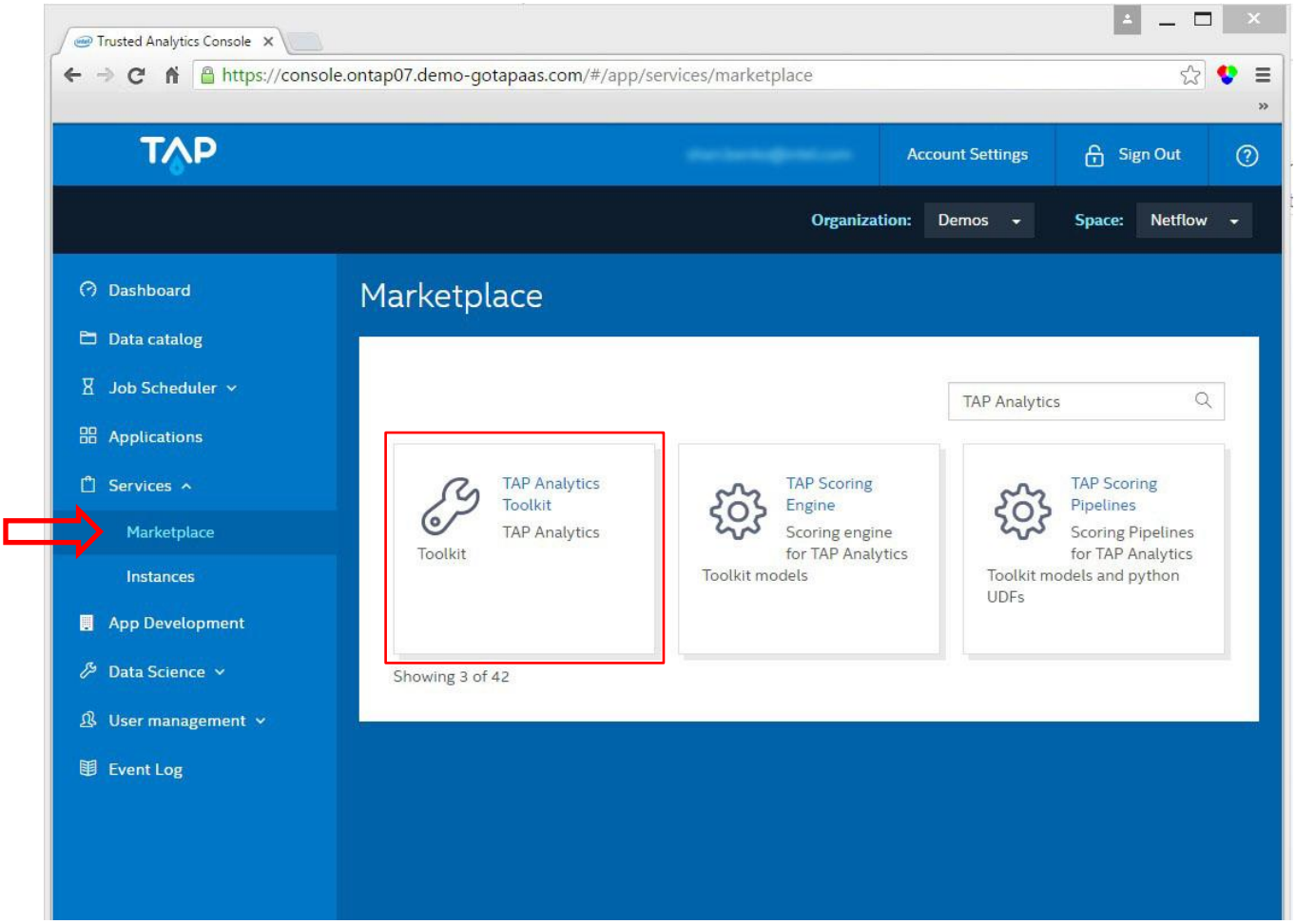
- Complete the Submit Transfer fields:
  - **Category** – Click on a category, based on your needs.
  - **Link field** – Enter the URL of the dataset you want to upload.
  - **Title field** – Enter a file name for your dataset.
- Click on Upload after you have entered the necessary information.



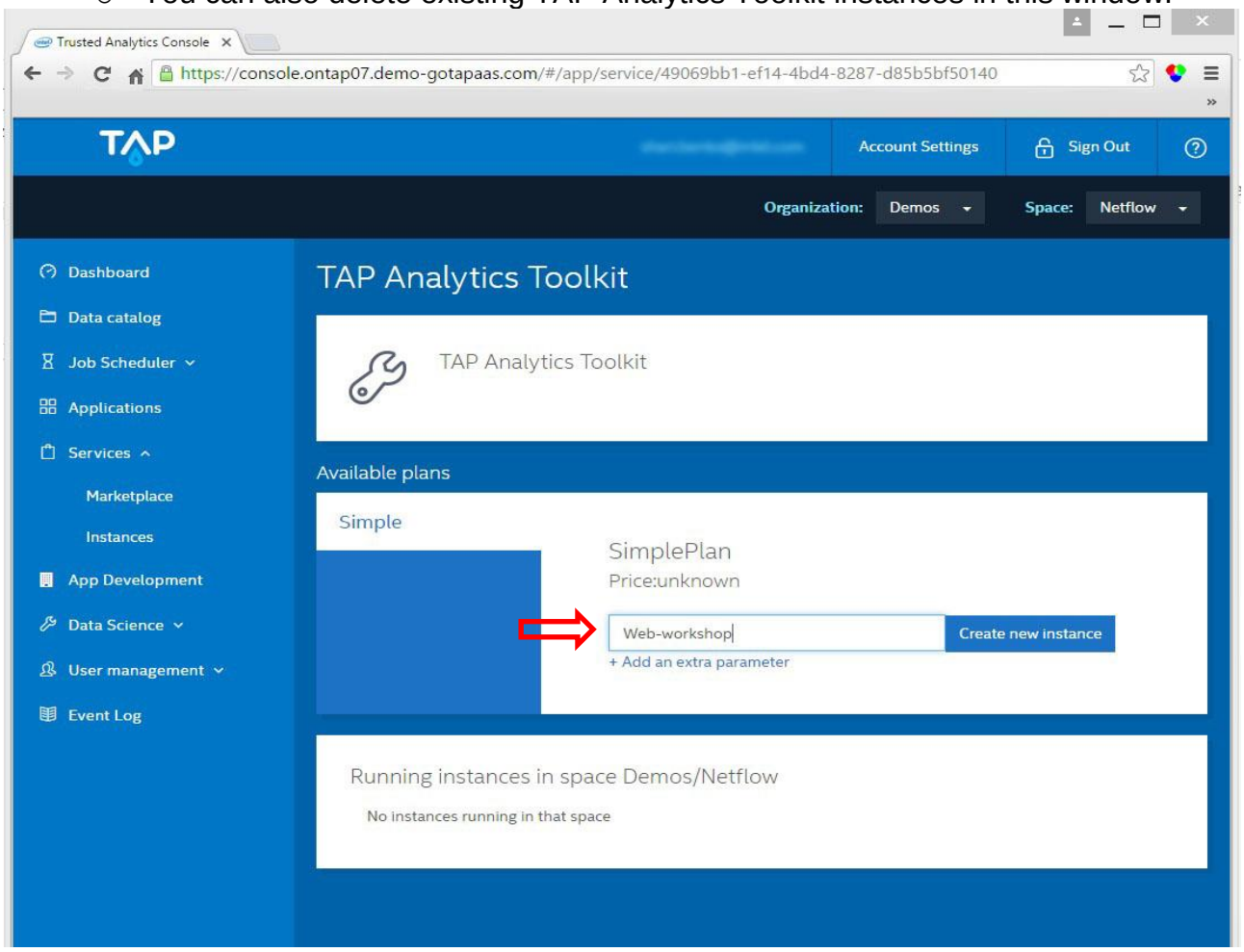
3. TAP Marketplace

3.1. Create a new instance of TAP Analytics Toolkit in TAP:

- Click on **Services > Marketplace** under Main navigation.
- Search for then click on TAP Analytics Toolkit.

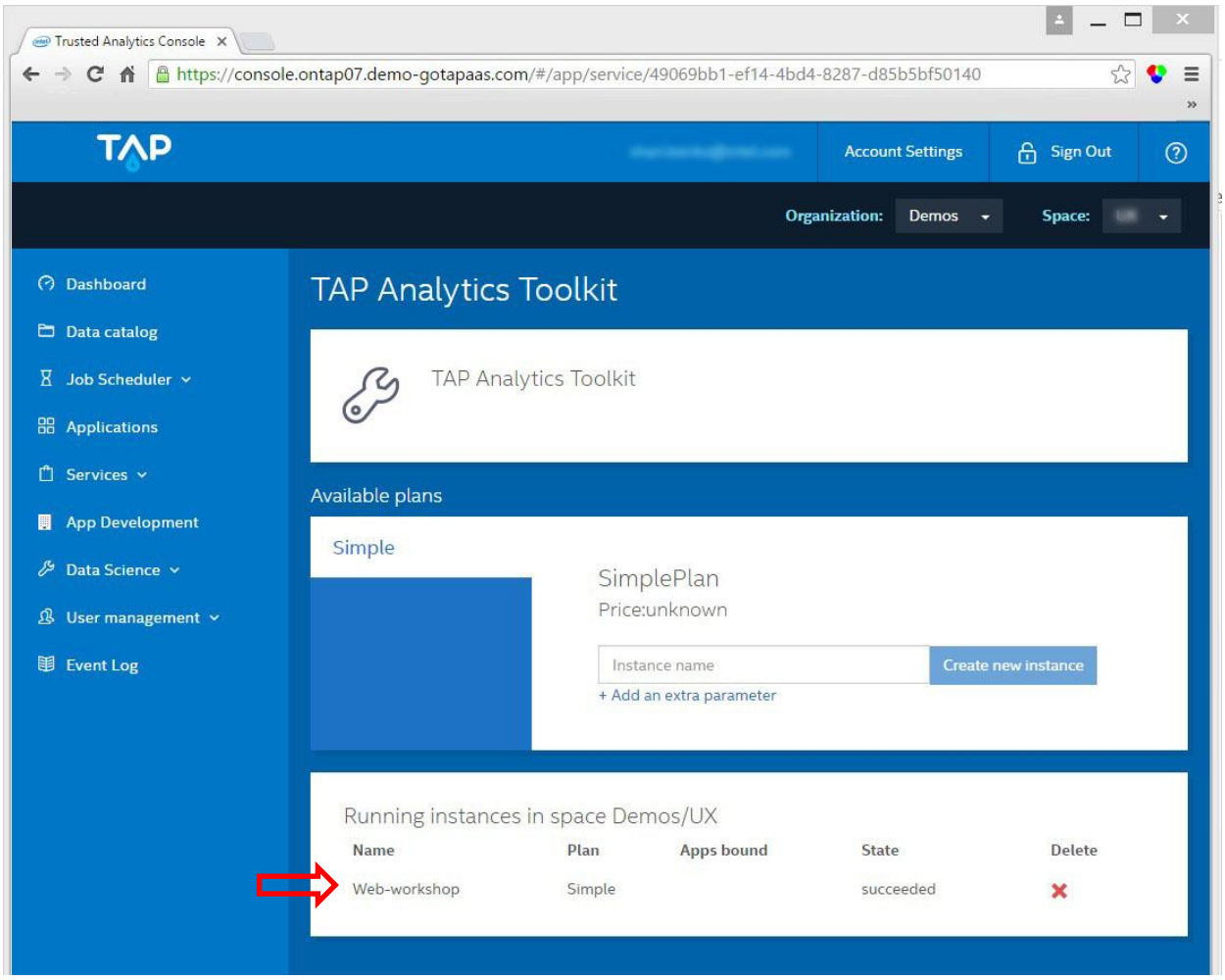


- In the TAP Analytics Toolkit window, enter a name for your instance and click Create new instance. In this example, we created an instance called Web-Workshop.
- You can also delete existing TAP Analytics Toolkit instances in this window.



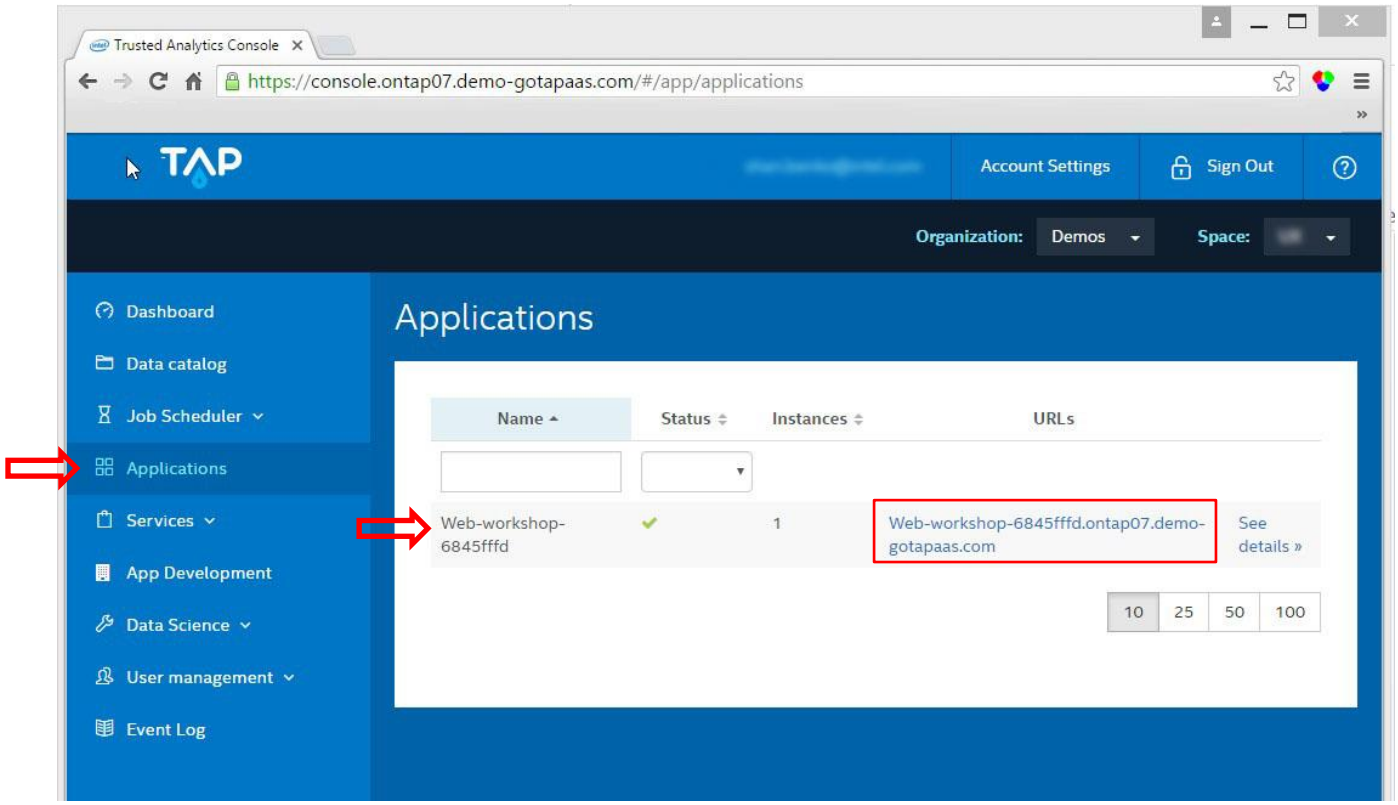
3.2. View the running instance:

- Refresh your window to see the instance you created.



3.3. Copy the instance URL to use in the Netflow demo:

- Click Applications under Main navigation to view the status of the new instance in the Applications window.
- Copy the URL of the instance. You will need the URL to run the Netflow demo in the next part of the workshop.
  - The red box in the Status column indicates the instance is still initiating and not yet available. The red box is replaced by a green check when the instance is running.
  - Click on the instance's URL to see if the instance is up and running.



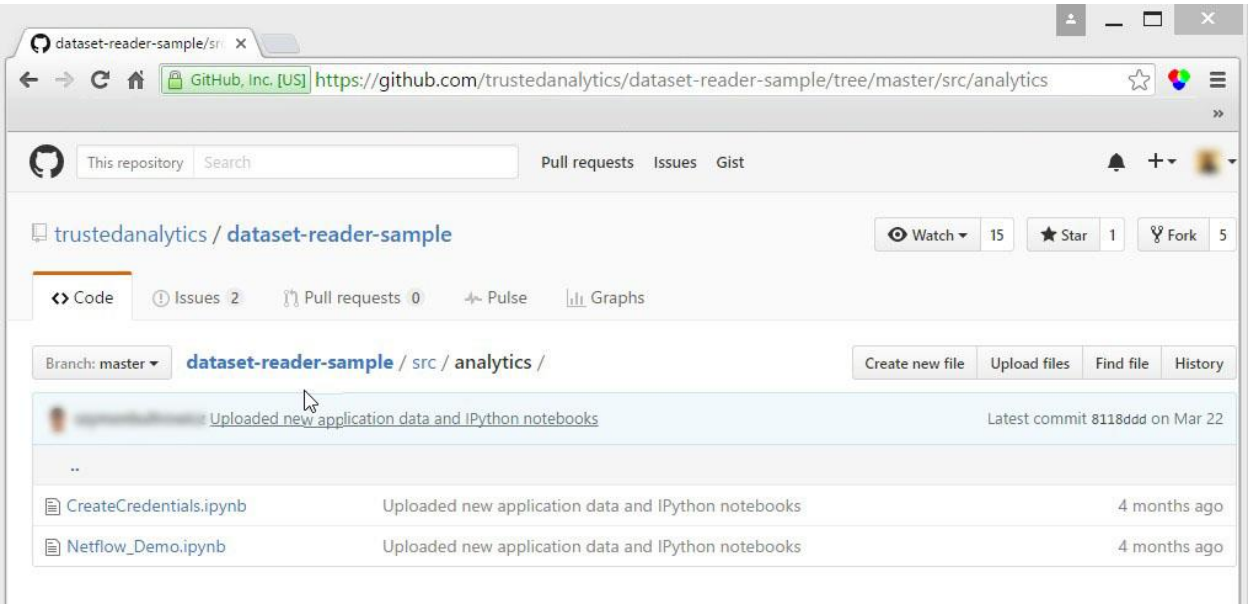
4. Netflow demo

4.1. What is the Netflow demo?

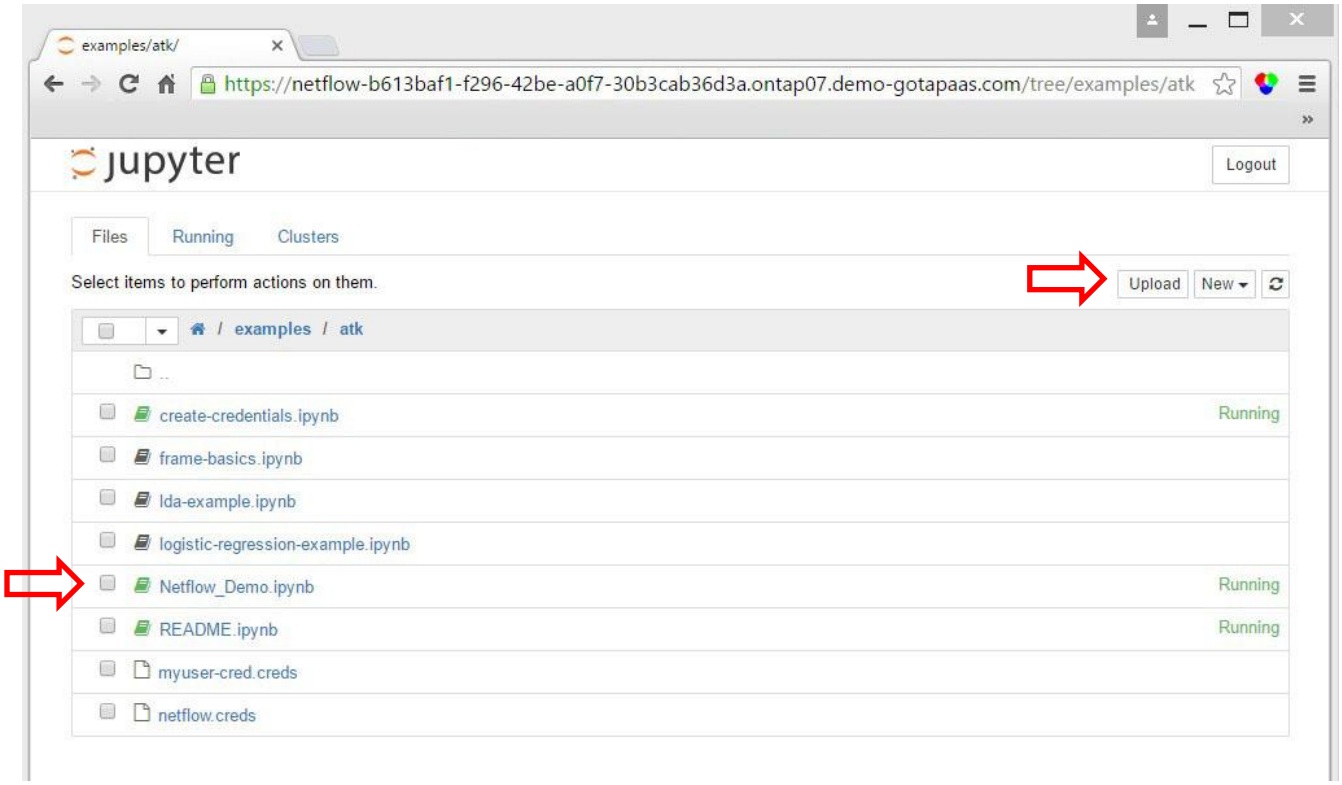
- The Netflow dataset contains information from web servers. The dataset contains information on tracked network traffic, which is incoming and outgoing internet traffic.
- The demo reads and trains the dataset. The demo then establishes a normal traffic pattern by tying the network edges and nodes together.
- It builds a graph showing all the connections based on the nodes' IP addresses.
- Finally, it produces a number of charts and graphs to show the network traffic flow and display outliers, which are outside the normal traffic pattern.

4.2. Open the Netflow demo file:

- Open the Netflow demo file using a Jupyter Notebook instance on TAP using the following steps:
  - Using the left side navigation, click 'Data Science'. Three additional options appear, click 'Jupyter'.
  - Specify an instance name for your new Jupyter instance and then click 'Create new Jupyter instance.' A notification will appear saying your new instance is created but may take a while. Refresh the page and you should see a new Jupyter instance on the page.
  - If you are using the Jupyter Notebook instance for the first time, you'll need to create a new set of credentials to connect to TAP Analytics Toolkit. You can use the file provided here: https://github.com/trustedanalytics/dataset-reader-sample/tree/master/src/analytics.
  - Download the Netflow demo script file (Netflow_demo.ipynb) located in same repository https://github.com/trustedanalytics/dataset-reader-sample/tree/master/src/analytics.

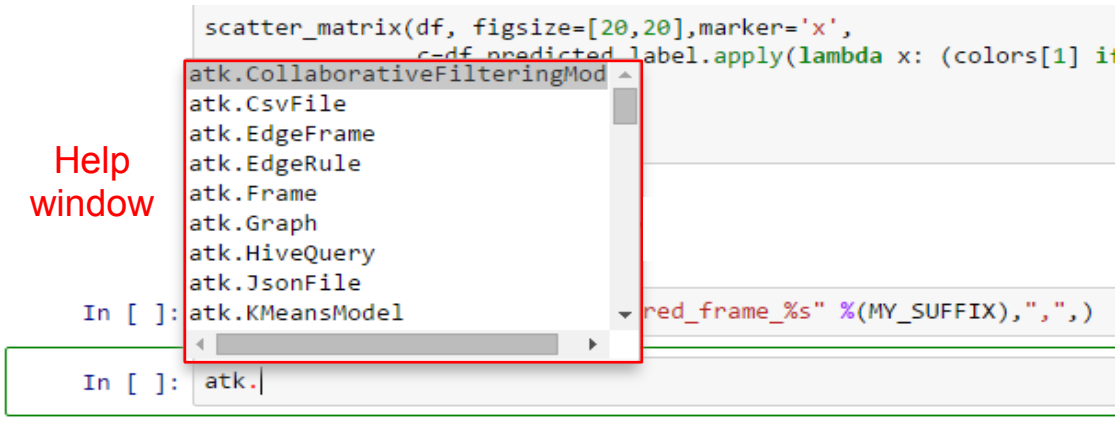


- Upload the demo script file using the upload button in the upper right of your Jupyter Instance, then click on the file to open the demo.
- You can find the training dataset located on the Trusted Analytics Community site: https://community.trustedanalytics.org/docs/DOC-1058

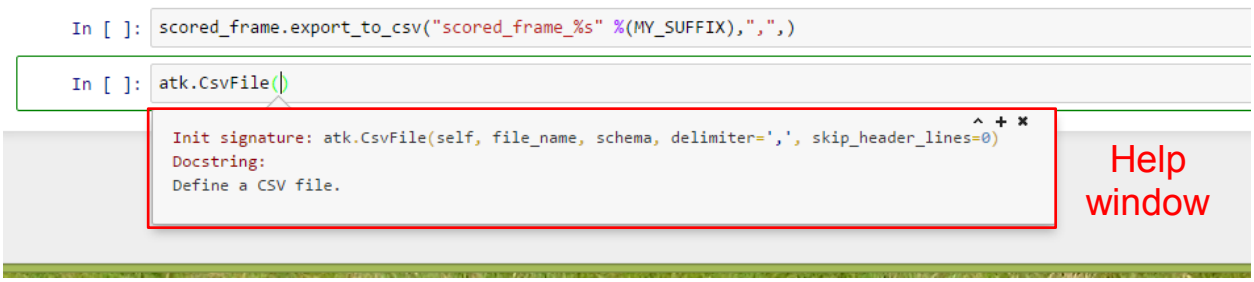


4.3. Review the help functions:

- If you need help knowing what commands are available:
  - Type the command into the In [ ] field; for example, "**TAP Analytics Toolkit.**"
  - Then press the **tab key** to see the available command options.

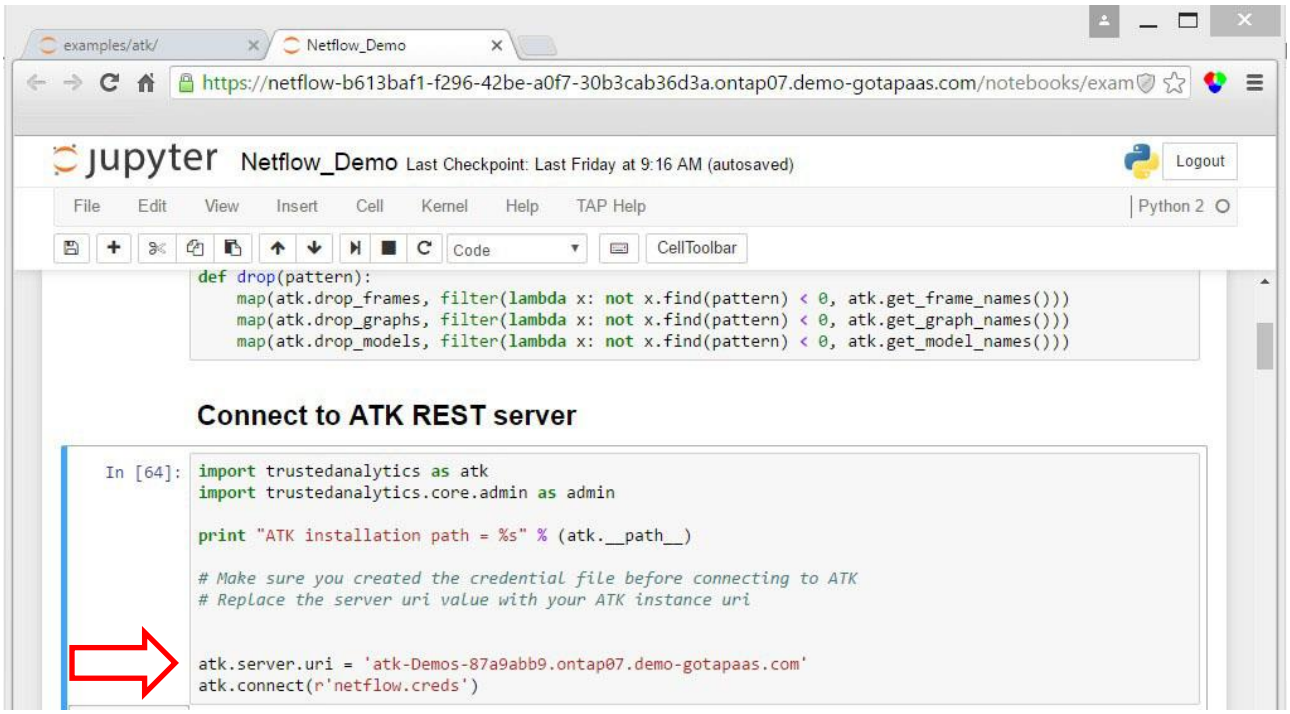


- If you need help knowing what the default parameters are for a command:
  - Type the command into the In field. For example, **"TAP Analytics Toolkit.CsvFile( )".**
  - Then press the **shift** and **tab keys** to see the default parameters for that command.



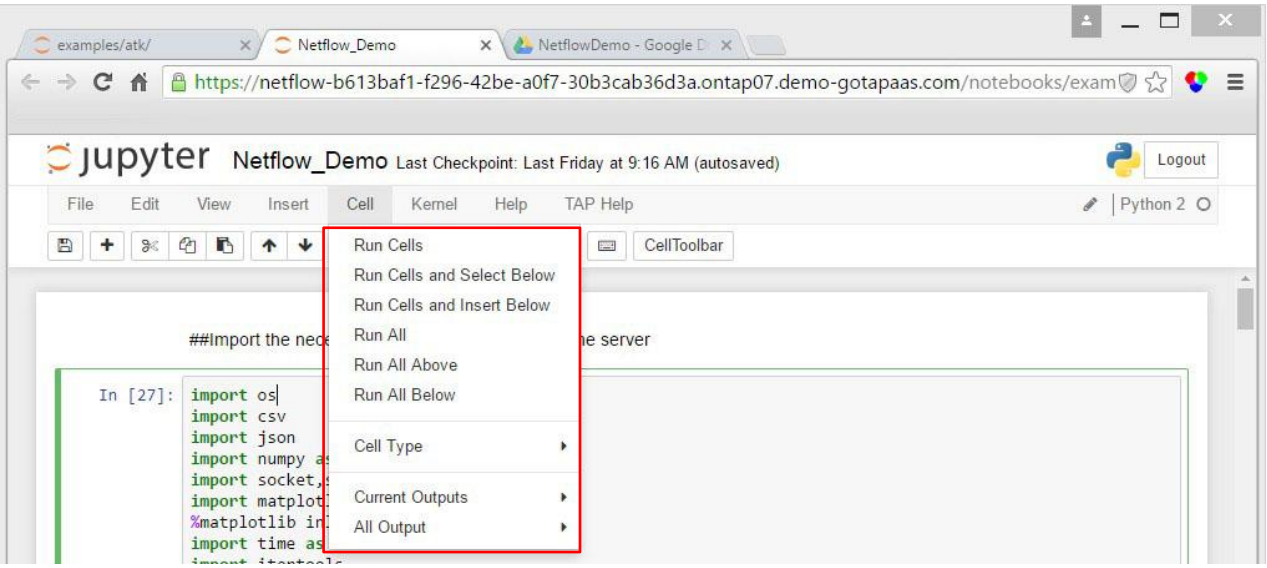
4.4. Connect to ATK REST Server

- Change the atk.server.uri in the demo script to access the TAP Analytics Toolkit instance you created in TAP.
- Copy the URL for your Analytics Toolkit instance you created in Step 3.3.
- Paste the URL of the instance into the atk.server.uri = field of the Netflow demo script.
- **Important**: You must remove the “http://” at the beginning of the URL and the “/” at the end, before you enter the uri information.
- You should already have created a credentials file, when you installed Python or Jupyter.



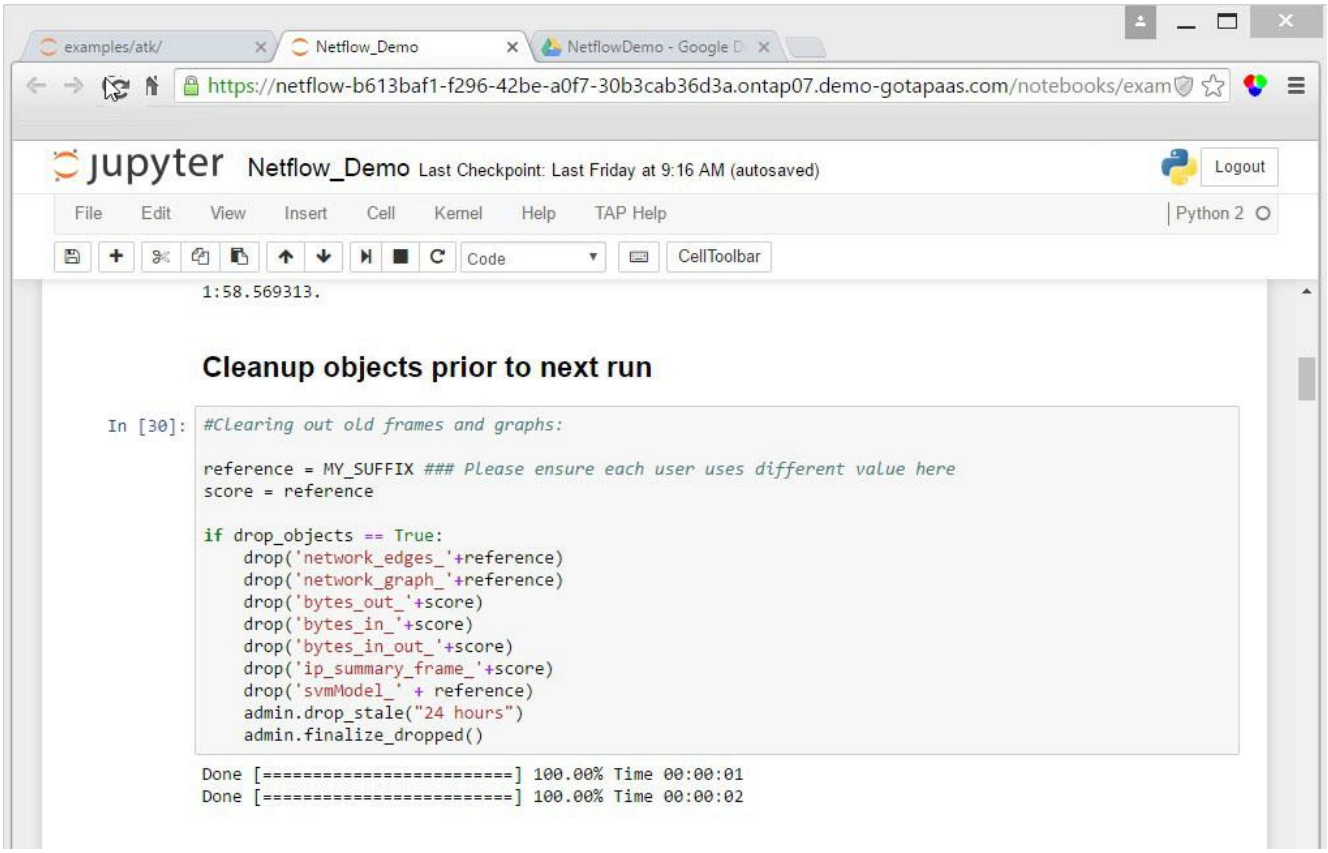
5. Deploy the TAP Analytics Toolkit Model

- Execute the Netflow demo, which deploys the TAP Analytics Toolkit.
- To begin, click Cell and select Run All. This runs all the scripts in the demo.
- If you need to rerun a single task, choose Run to execute a specific script in the demo.
- Or, if you need to run a group of scripts, choose Run All Below.



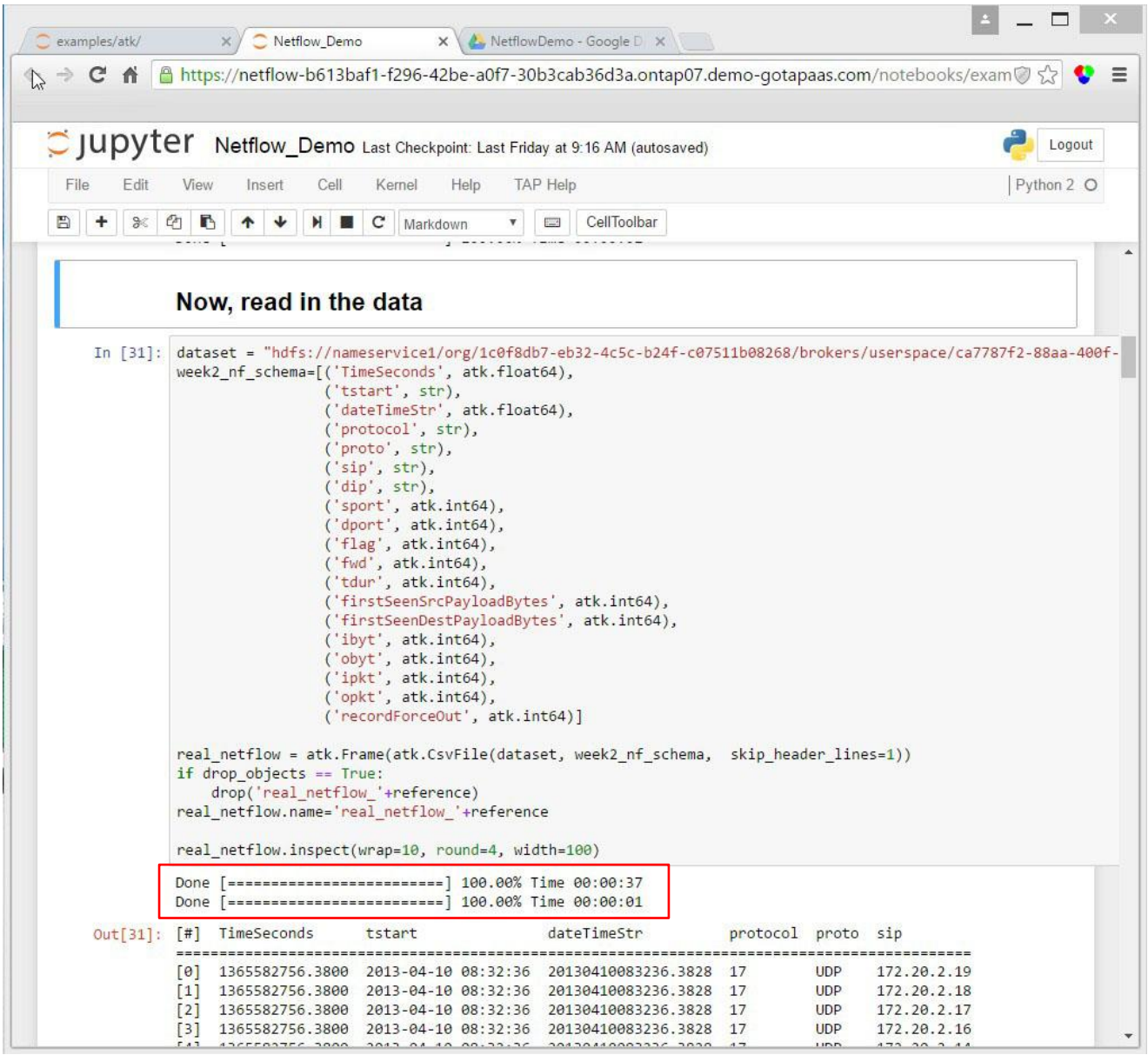
5.1. Clear out the old frames and graphs:

- The first step in the script cleans up the objects to ensure the code can be rerun multiple times.
- This script resets all the counters to zero.



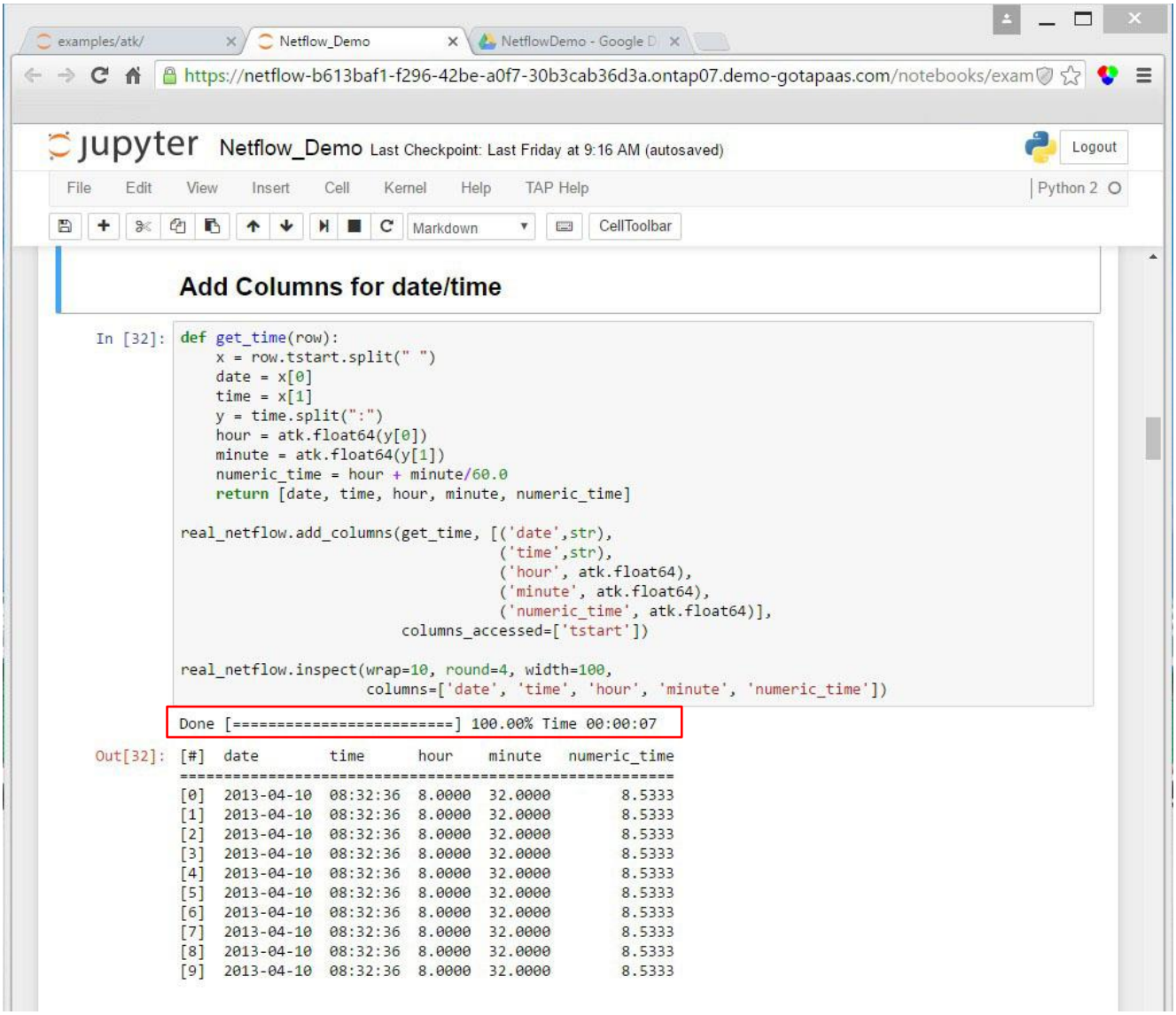
5.2. Now, read in the data:

- This step reads in the data file that was loaded onto the TAP Analytics Toolkit server.
- The Netflow dataset includes: date and time information, node IP addresses, protocols, bytes in and out, packets in and out …
- Confirm that the script completed 100% of the tasks and review the log for error messages. You may see several progress bars when multiple tasks run in the same section.

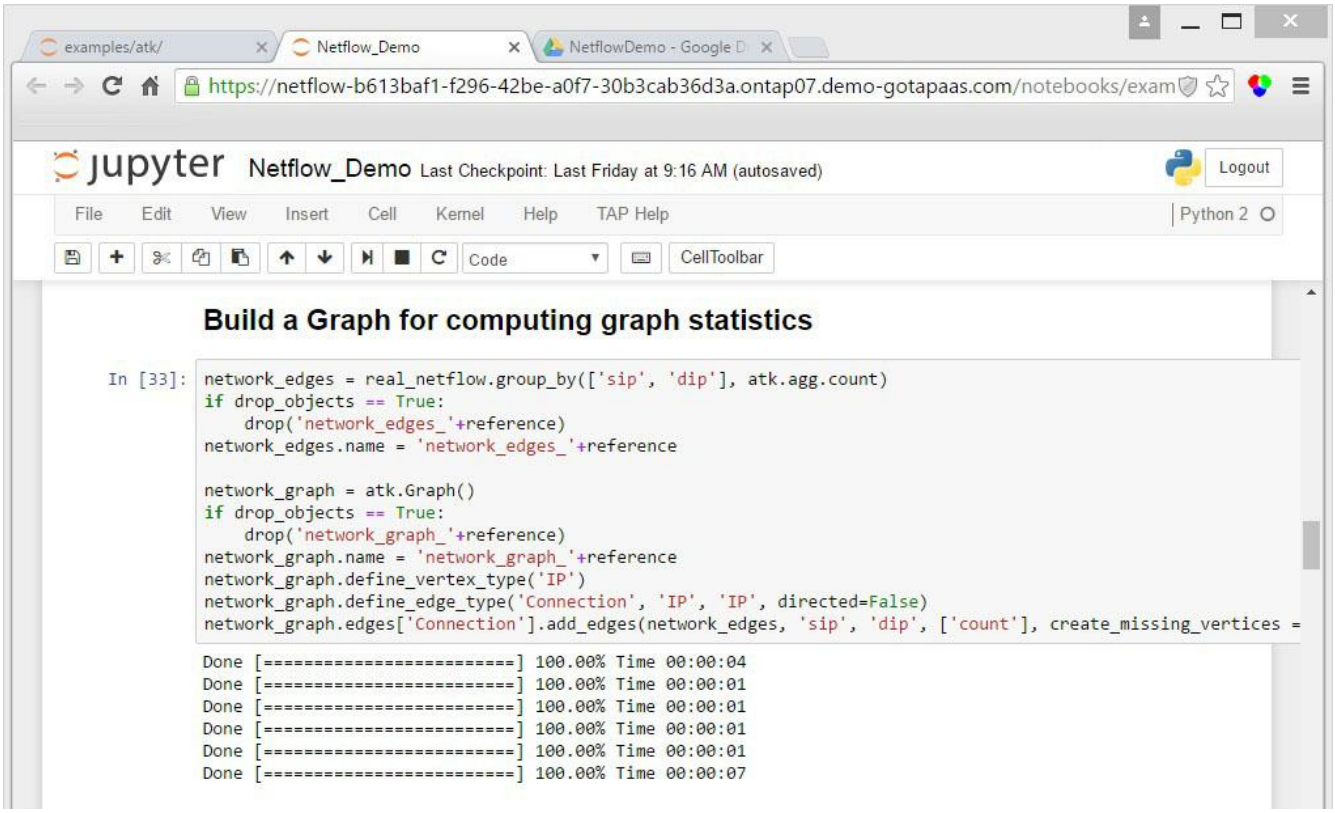


5.3. Add Columns for date/time

- This script formats the data into columns for date, time …

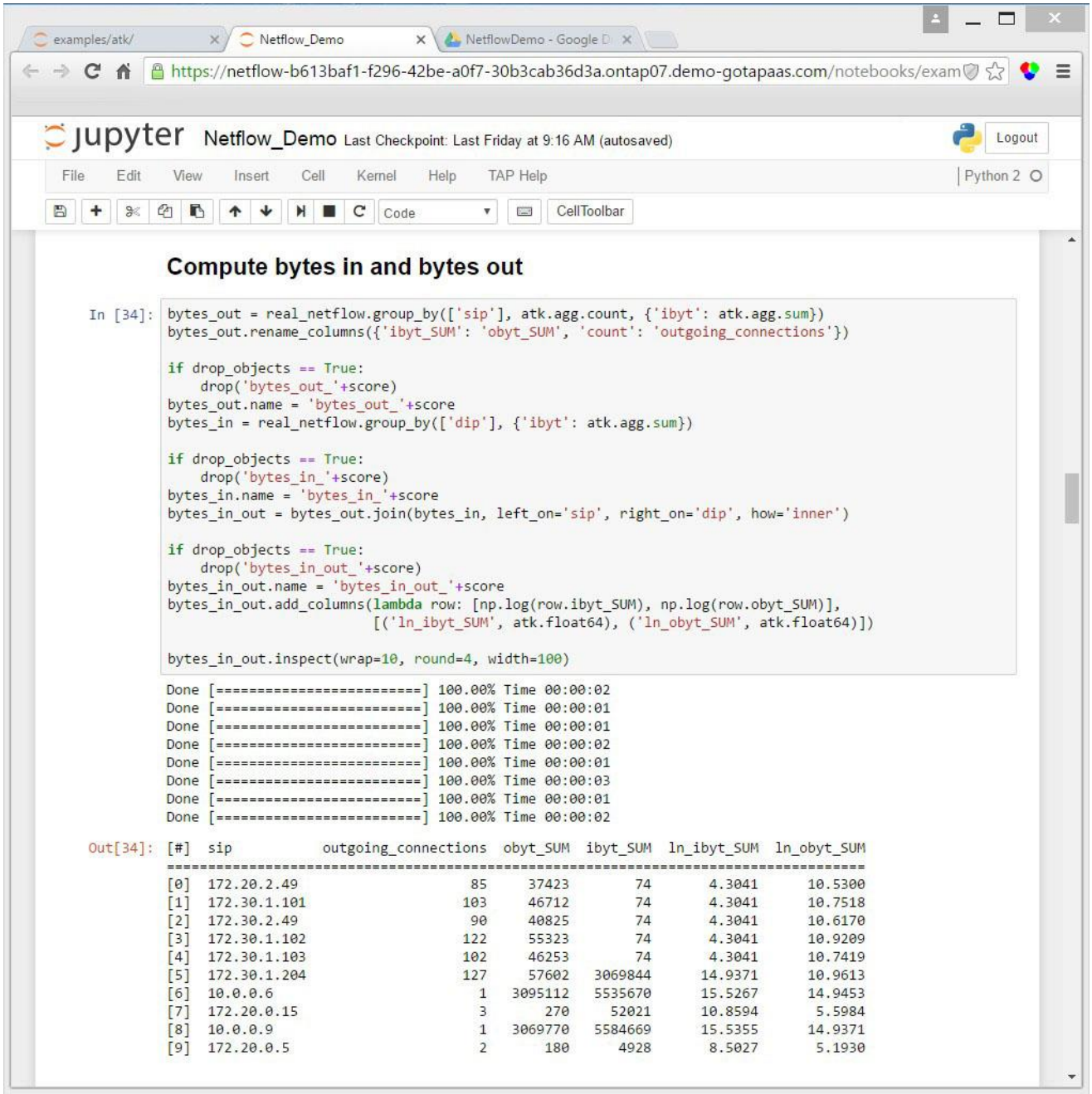


5.4. Build a Graph for computing graph statistics



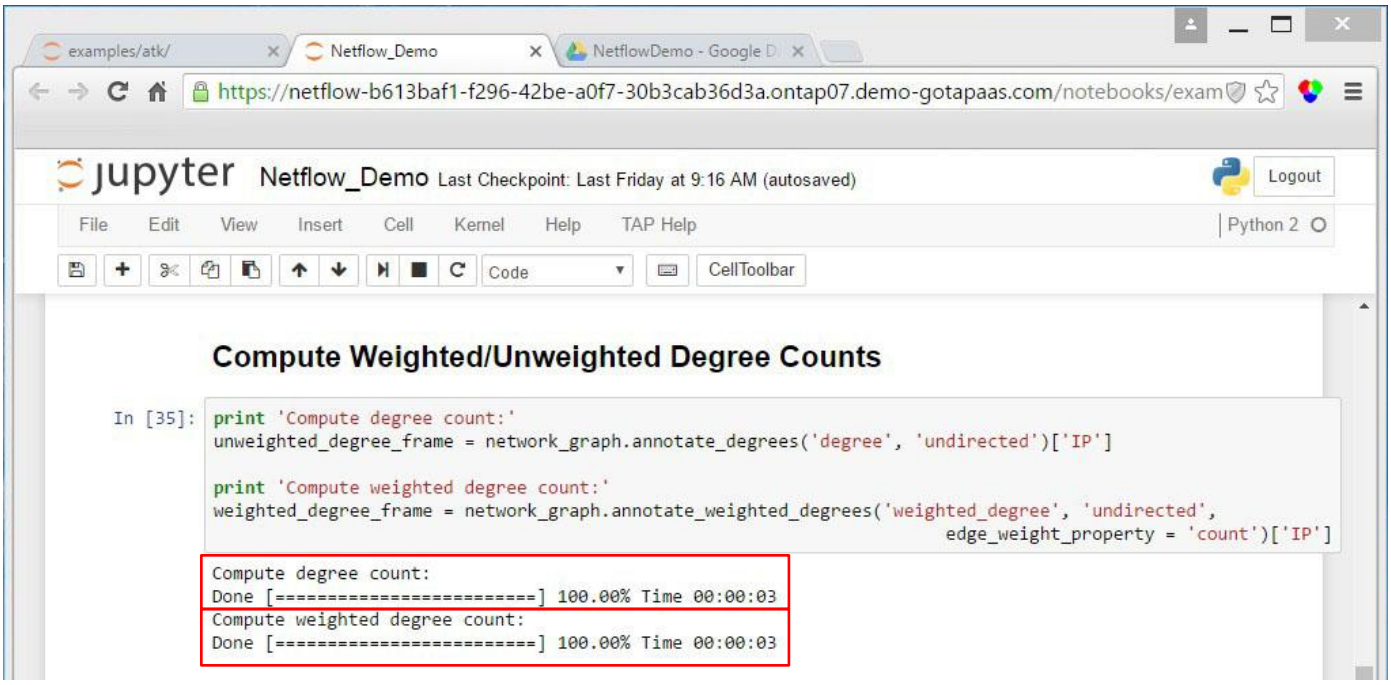
5.5. Computes bytes in and bytes out:

- This script computes the number of bytes in and bytes out for each node and creates total I/O counts for the nodes.



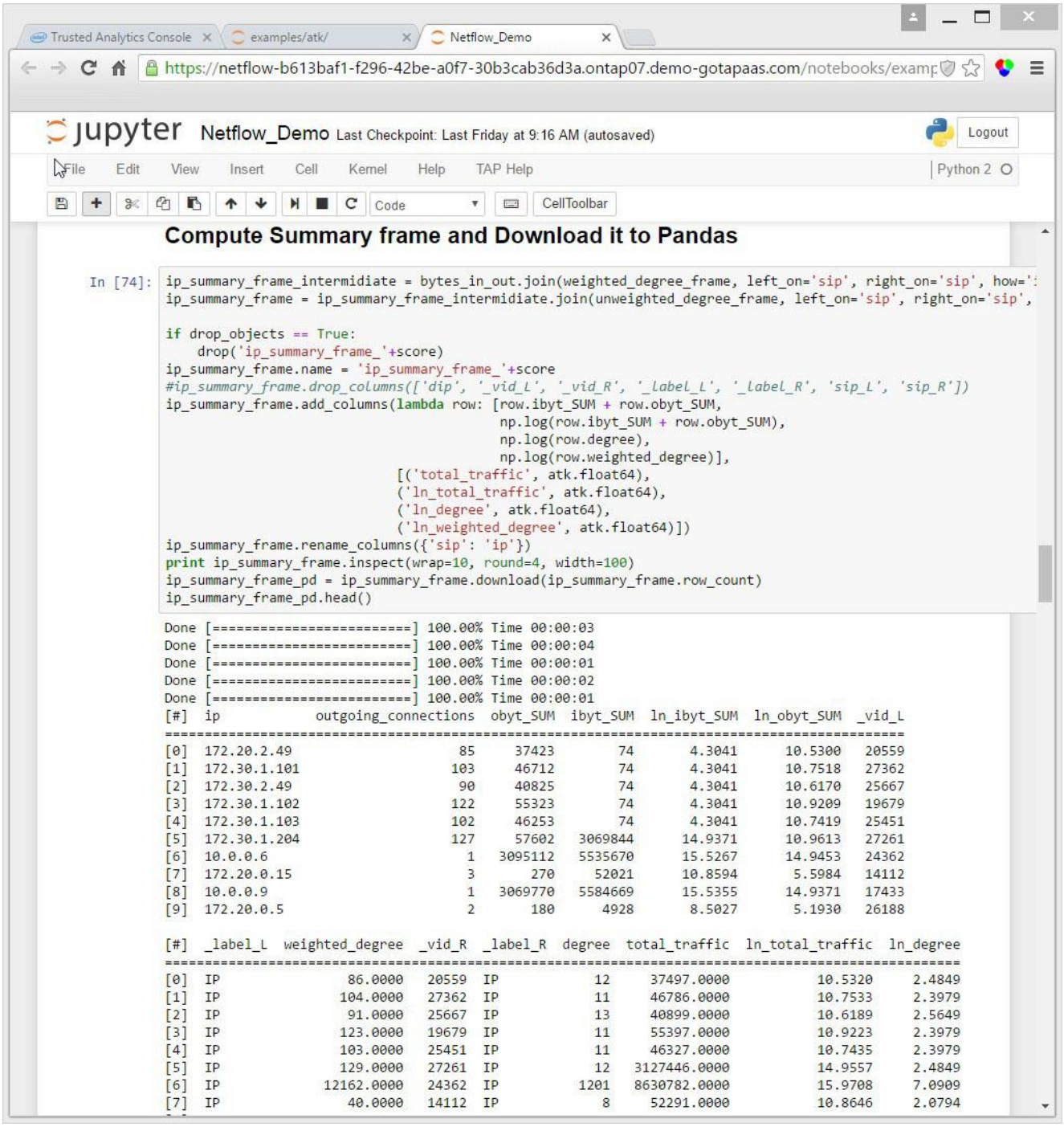
5.6. Compute weighted and unweighted degree counts:

- The degree of a node is the number of relations (edges) it has.
- Relations (edges) are borders between 2 IP nodes. If a node has a high degree, it means it has a lot of nodes as neighbors.
- The weighted degree of a node is like the degree. It's based on the number of edges for a node but is influenced by the weight of each edge.
- In addition to calculating the degree counts, this script identifies the 2 tasks it runs using print statements, e.g., print 'Compute degree count:' and print 'Compute weighted degree count:'. Notice the printout located just above each of the Tasks %.



5.7. Compute summary frame and download to Pandas:

- A data frame is a table, or two-dimensional array-like structure, in which each column contains measurements on one variable and each row contains one case. A summary frame is the sum of each row or column.
- This script prints the summary frame results in a tabular format.
- Additionally, this script downloads the data to your local machine. This allows you to execute commands and scripts against the sample data set.

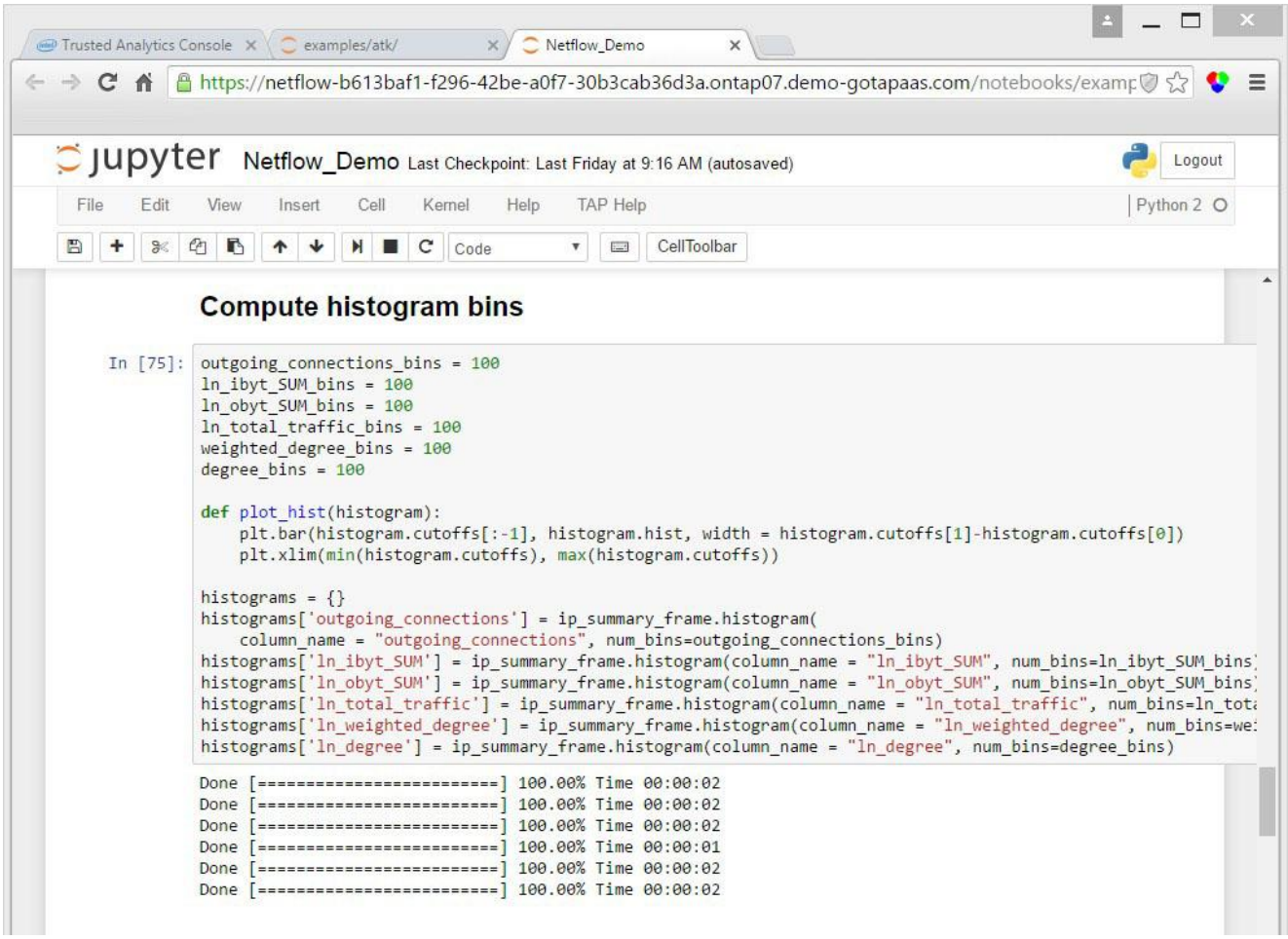


- This is only a sample of the summary frame table.

	ip	outgoing_connections	obyt_SUM	ibyt_SUM	ln_ibyt_SUM	ln_obyt_SUM	_vid_L	_label_L	weigh
0	172.20.2.49	85	37423	74	4.304065	10.530041	20559	IP	86.0
1	172.30.1.101	103	46712	74	4.304065	10.751756	27362	IP	104.0
2	172.30.2.49	90	40825	74	4.304065	10.617050	25667	IP	91.0
3	172.30.1.102	122	55323	74	4.304065	10.920944	19679	IP	123.0
4	172.30.1.103	102	46253	74	4.304065	10.741882	25451	IP	103.0

5.8. Compute histogram bins:

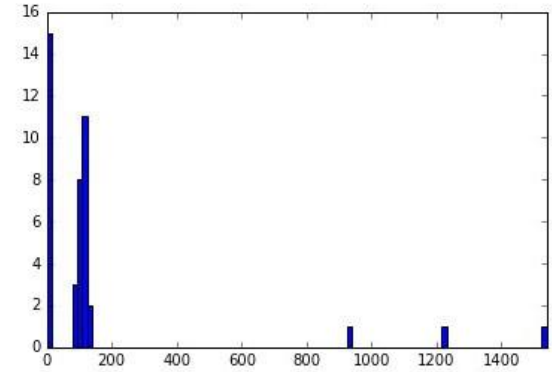
- A histogram is a visual interpretation of numerical data indicating the number of data points that lie within a range of values, called a class or a bin. The frequency of the data in each class is depicted by the use of a bar.
- This script separates the data into six groups—outgoing connections, in byte sum, out byte sum, total traffic, weighted degree and degree.
- It displays the results in 6 histograms, seen on the next page.



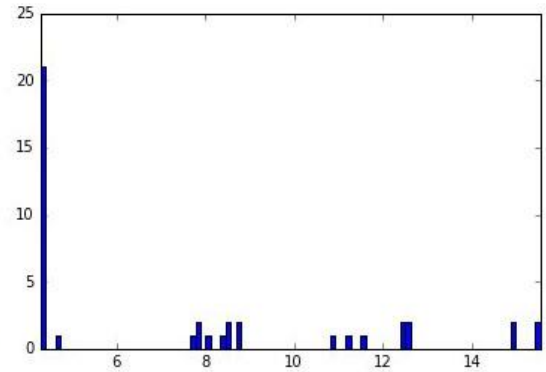
- Six histograms:

Plot histograms

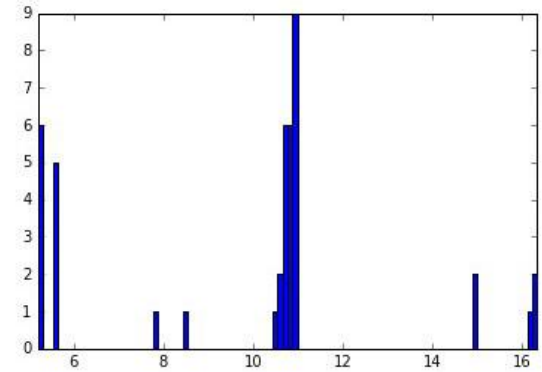
```
plot_hist(histograms['outgoing_connections'])
```



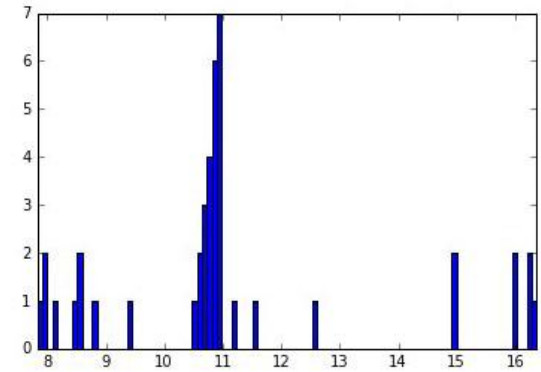
```
plot_hist(histograms['ln_ibyt_SUM'])
```



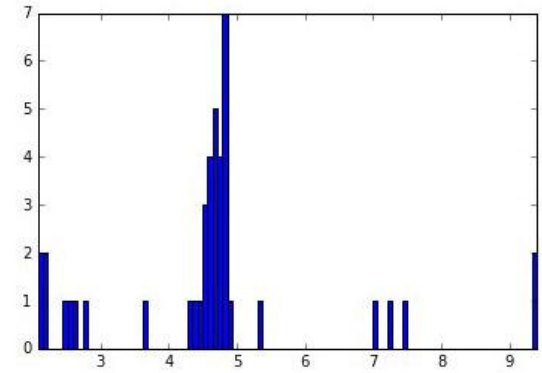
```
plot_hist(histograms['ln_obyt_SUM'])
```



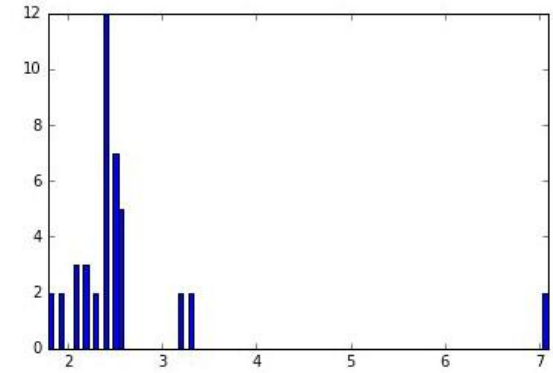
```
plot_hist(histograms['ln_total_traffic'])
```



```
plot_hist(histograms['ln_weighted_degree'])
```

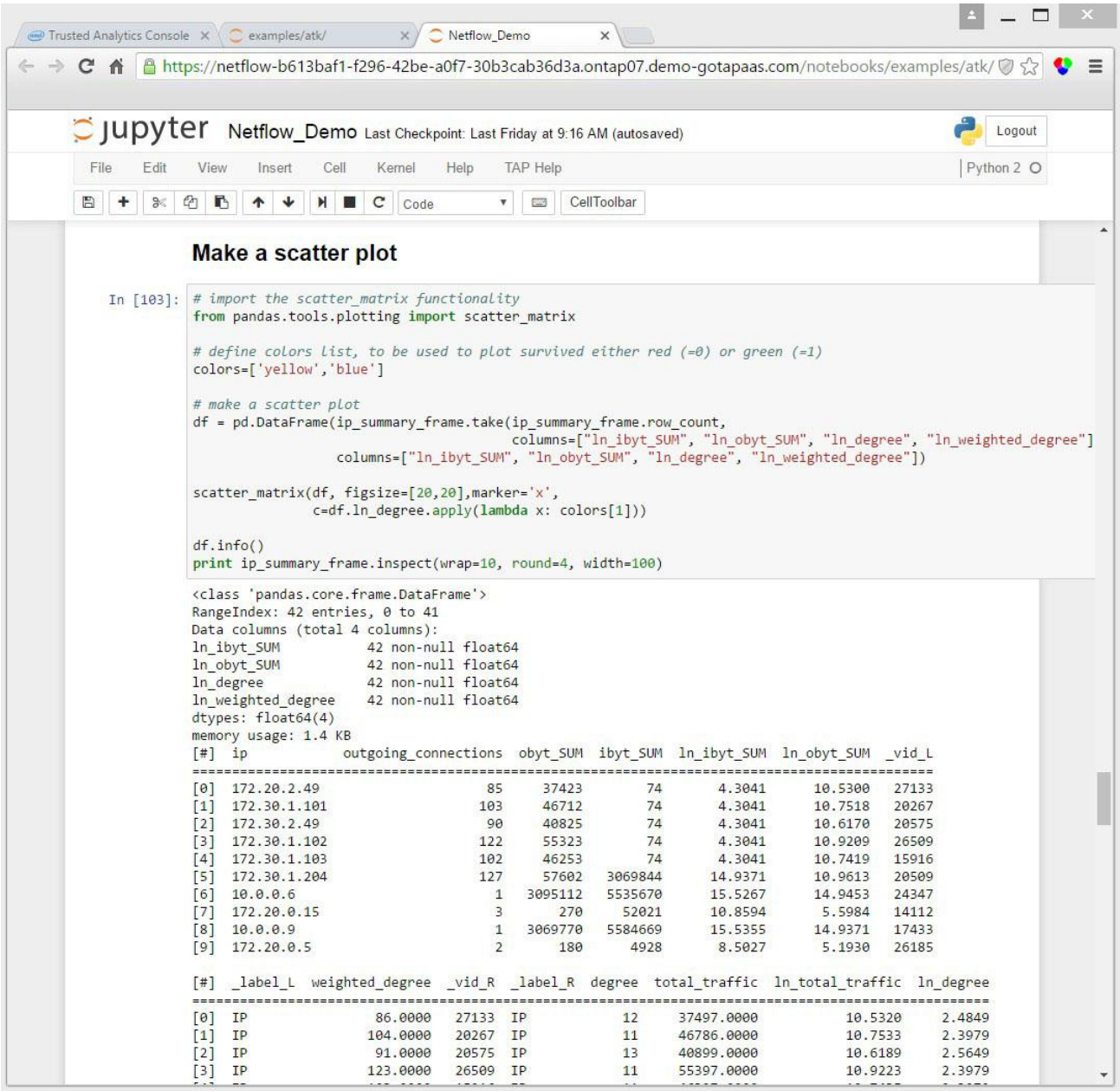


```
plot_hist(histograms['ln_degree'])
```

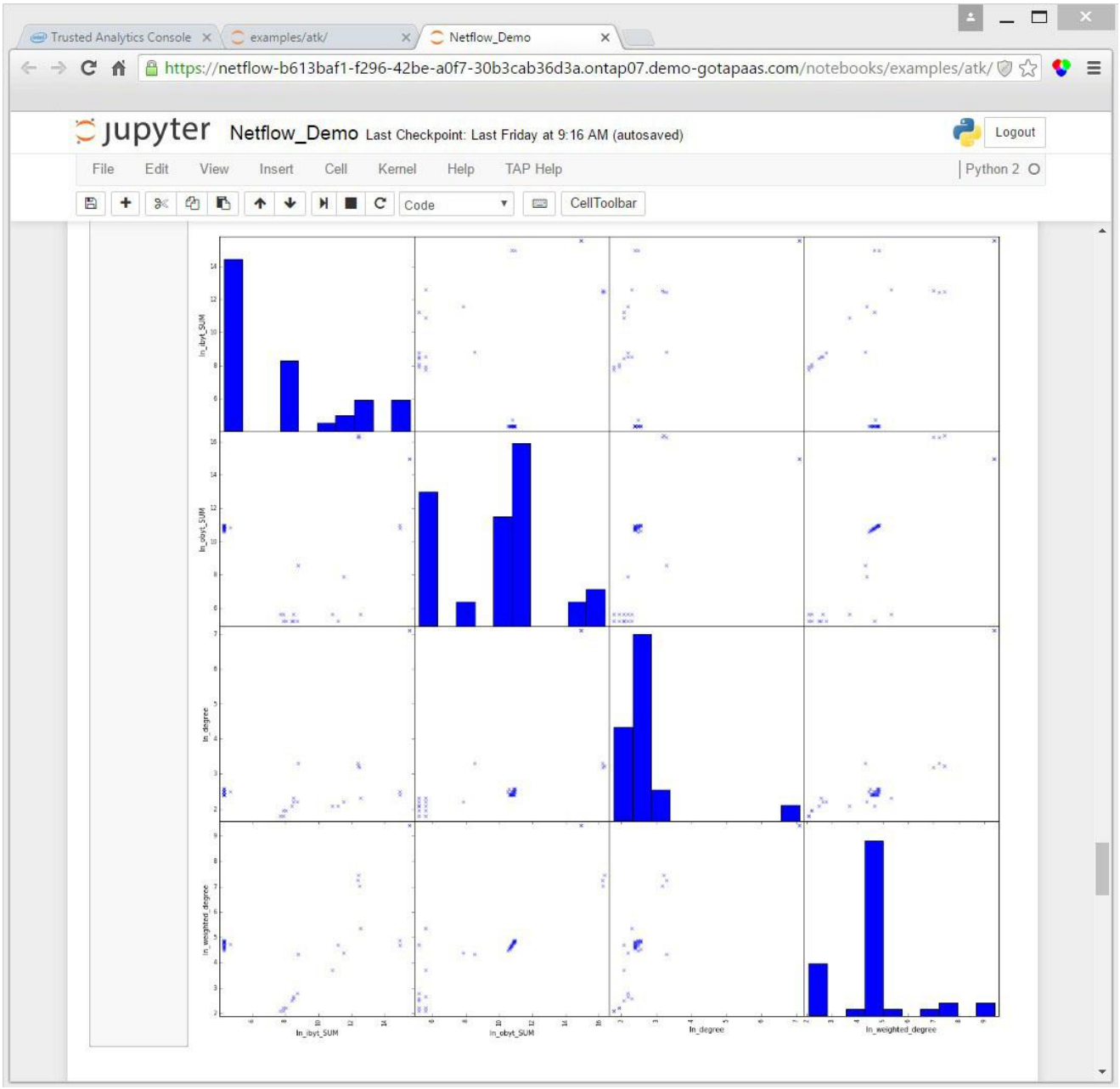


5.9. Make a Scatter Plot

- A scatter (XY) plot has points that show the relationship between 2 sets of data. The relationship between two variables is called a correlation.
- This script uses 4 groups of data—in byte sum, out byte sum, degree, and weighted degree.
- It creates a four-by-four scatter-plot matrix.



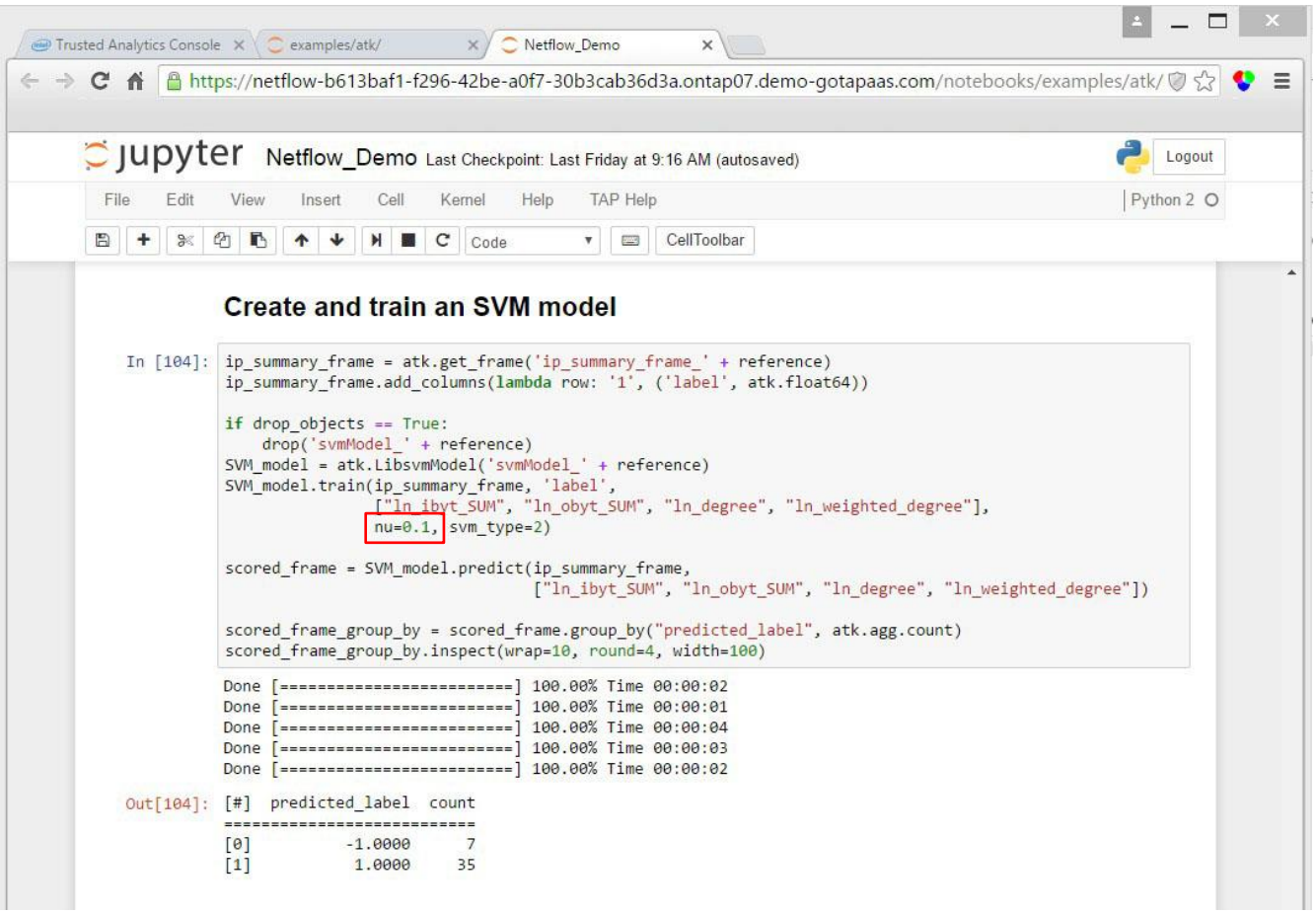
- Scatter-plot matrix:



- The scatter-plot matrix is replicated in Module 2: Visualizing Your Data with an App.

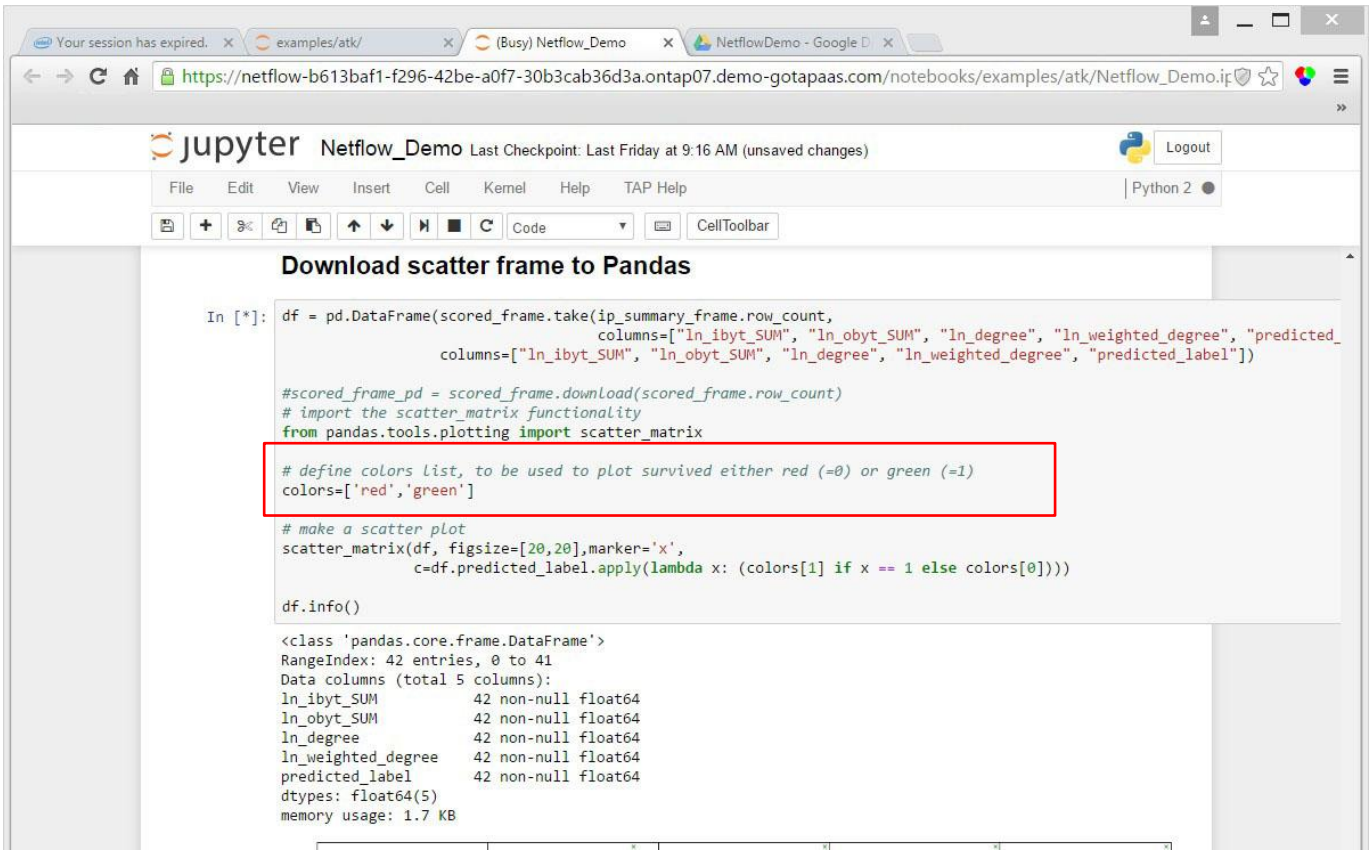
5.10. Create and train an SVM model:

- The script calculates the normal distribution of network traffic and the outliers. It detects anomalies on the network, such as DOS attach.
- The **nu** value determines the precision of our calculation. The value is the reciprocal of the dispersive power of a medium, known as constringency.
- After completing the remaining workshop steps, we encourage you to change the nu value in this script to see how it affects the data.
- Changing the nu value:
  - You **must** rerun step **5.7 "Compute summary frame and download to Pandas"** each time you change the nu value.
  - Then run this step and the last step in the workshop: **5.11 "Download the scatter frame to Pandas."**
  - Change the nu value to 0.2 or higher and notice how the graphs change.
  - Locate the upper limit of the nu value.



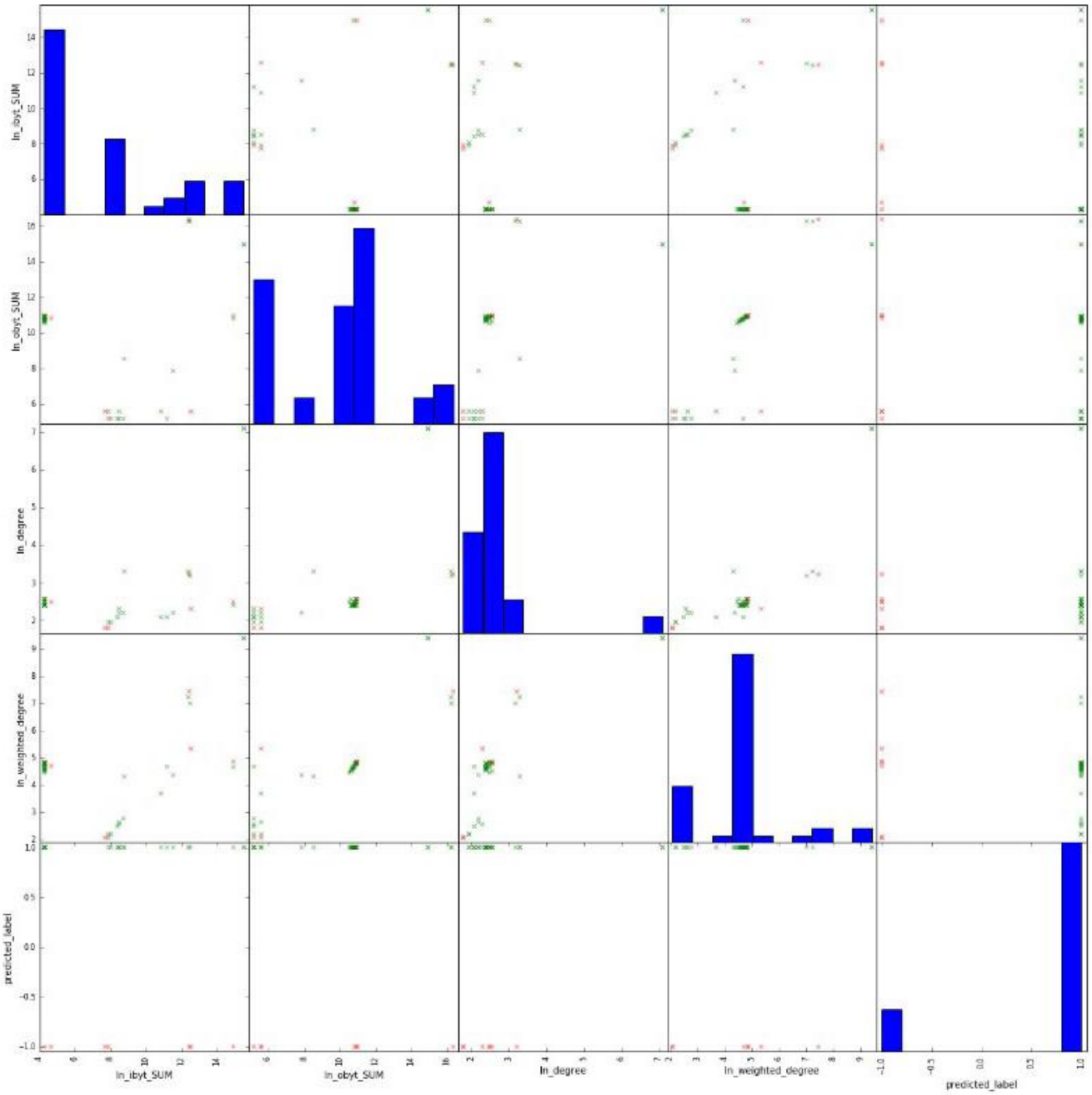
5.11. Download the scatter frame to Pandas:

- This script creates plot graphs from the results calculated in the “Create and train an SVM model” script in the previous step.
- The areas in green show the normal distribution. The points in red show the outliers.
- The points that are away from the “green cloud” in each frame should be investigated, as they fall outside the normal traffic patterns.
- To change the graph’s colors, modify the # *define colors list* in the script.



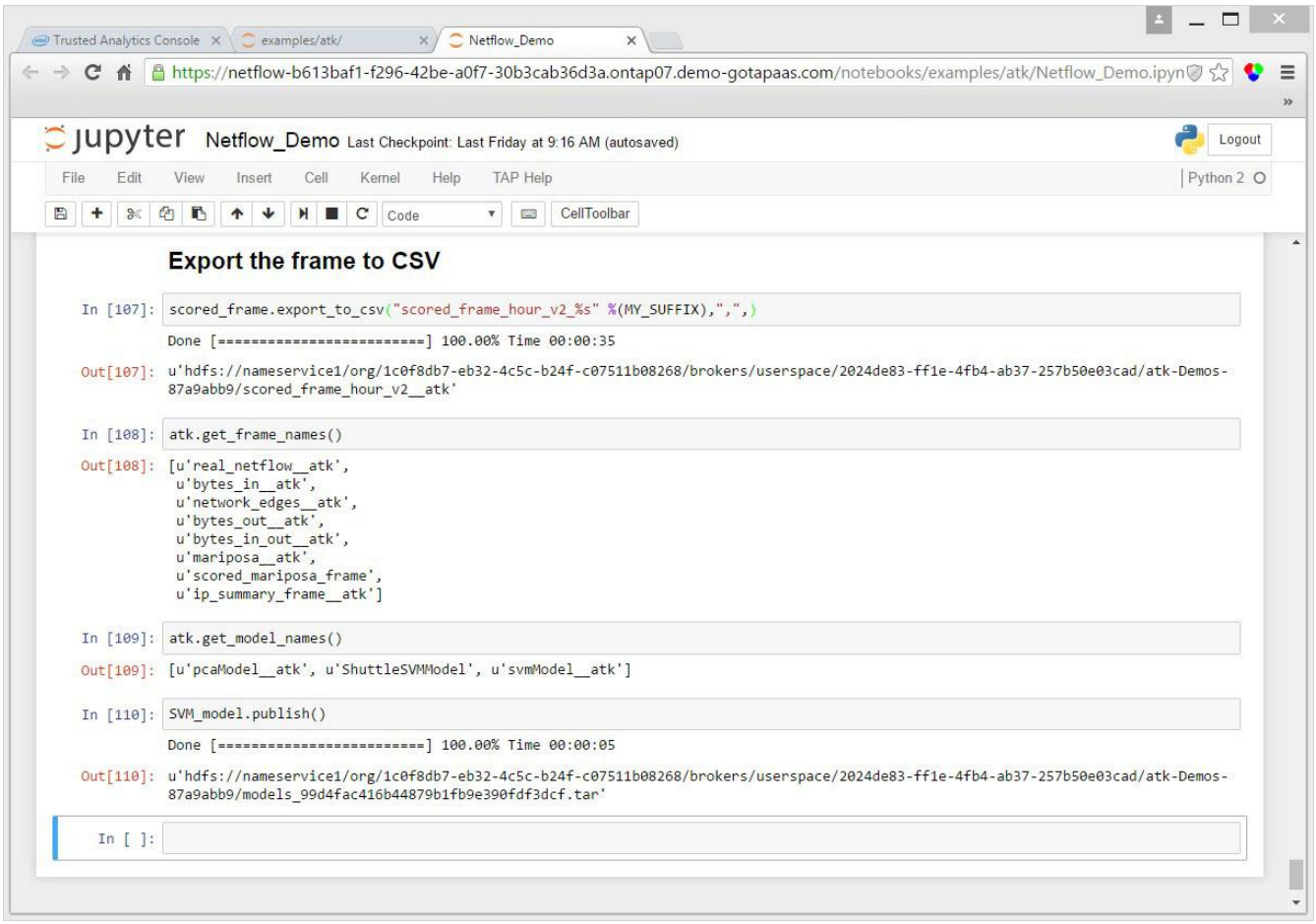
- See the scatter-plot graphs on the next page.

- Scatter-plot graphs:

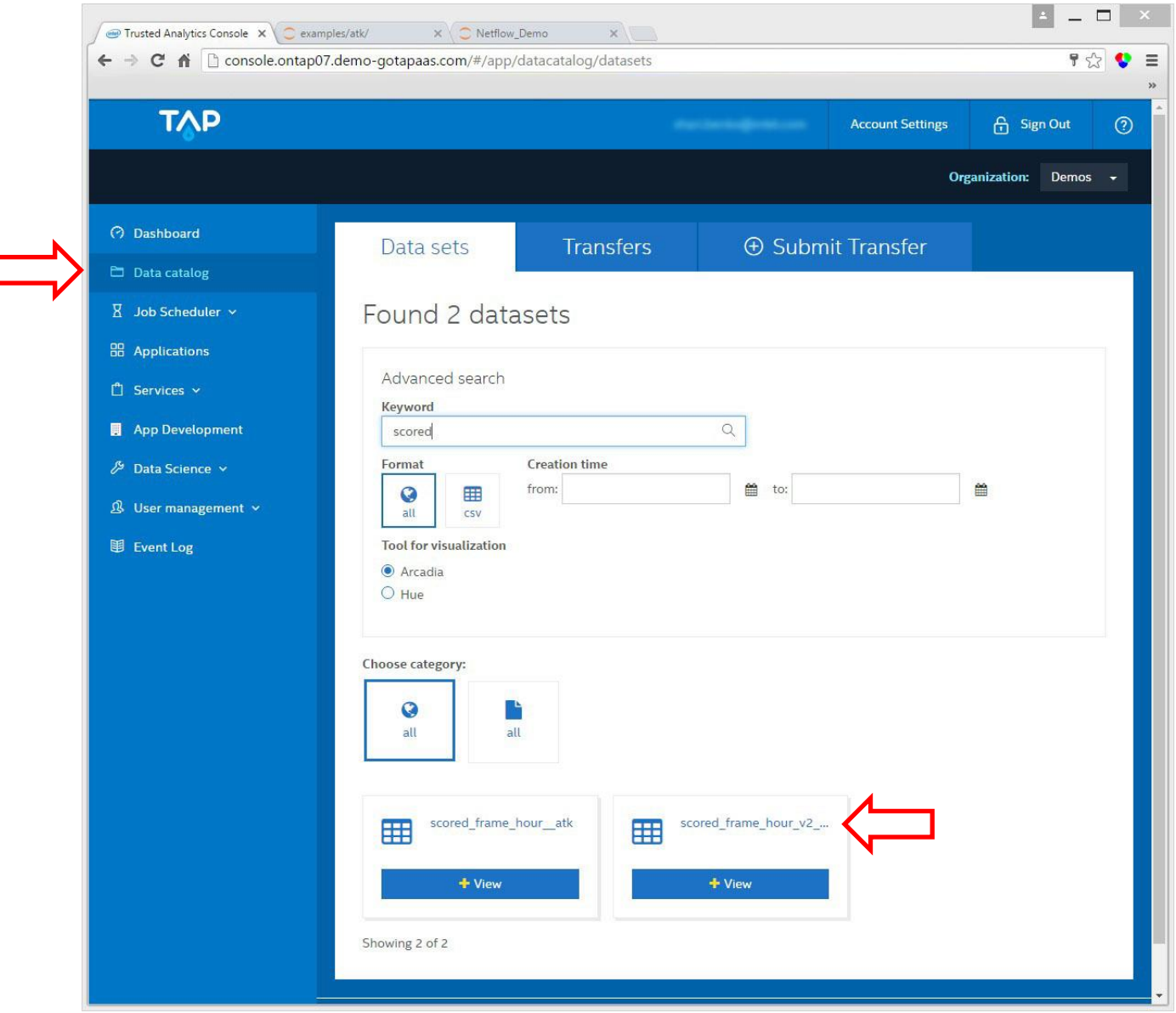


6. Export the data to HDFS as a CSV file

- This script exports the data to CSV files, e.g., Microsoft Excel.



- The CSV file will not be available to view in your TAP environment located in the Data Catalog tab.



CONGRATULATIONS!

You have successfully completed the workshop.