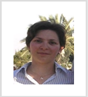
EMC IT Proven

EMC IT Proven

Smart Data Visualization: Helping Decision Makers Get the Picture

Posted on **December 12, 2014** by **EMC IT Proven**



By Dr. Lena Tenenboim-Chekina — Senior Data Scientist, EMC IT

Smart data visualization is proving to be an essential tool in maintaining increasingly complex Big Data systems in the cloud.

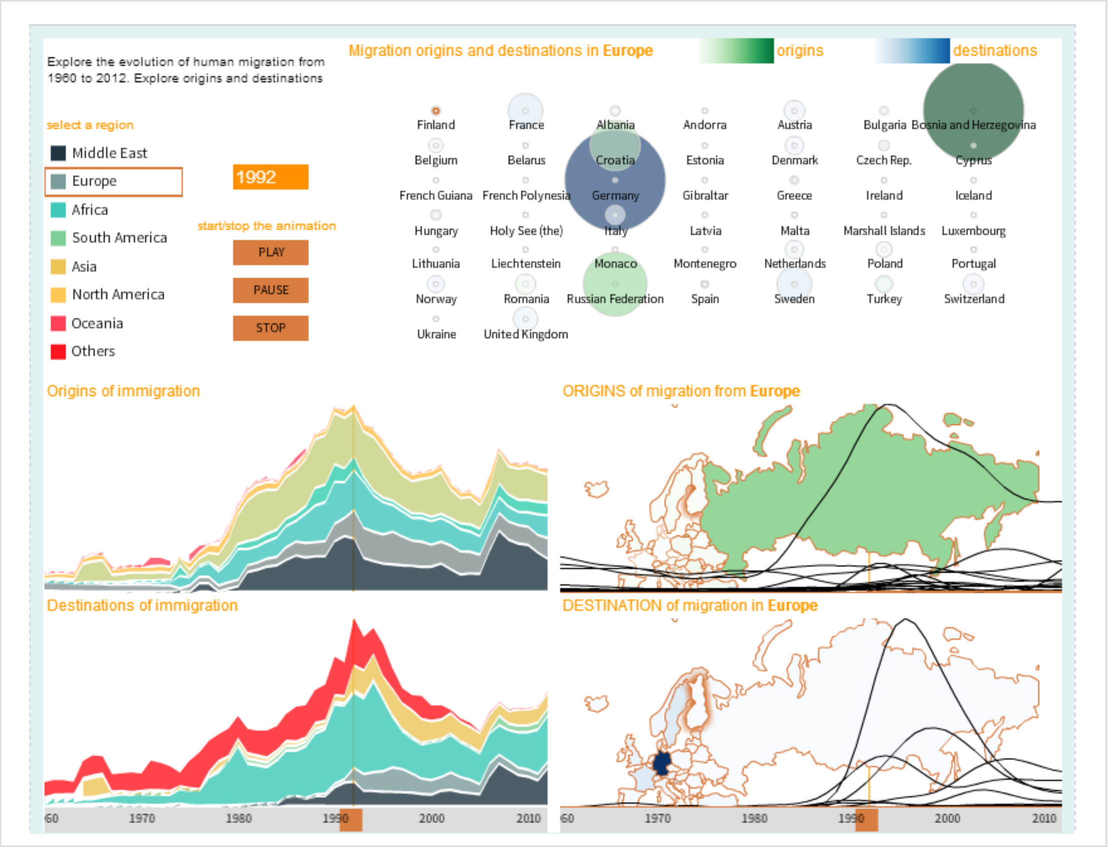
The adoption of Big Data tools and technology heavily relies on distributed scaled out computing. One of the main differences in this setting is that it includes systems that operate as a whole on top of several independent hosts. These hosts coordinate their actions with limited information and as a result maintenance complexity significantly increases. One way to overcome this challenge is smart data visualization, which helps the IT experts and management pinpoint the source of problems quickly.

The need for smart visualization is not unique to this problem. Representing complex data as a concise picture which tells decision-makers a story is a key part of any data analytics or data science project. Valuable results of a rigorous analysis may remain undiscovered due to a lack of a visualization clearly communicating the underlying information to the reader. The importance of data visualization is not a novelty. A number of visualization tools, as well as a general interest in data visualization topics, have exploded in popularity in recent years, as evidenced by the proliferation of literature available about infographics and visualization arcanum in both print and online media.

Executive customers of the Data Science-as-a-Service (DSaaS) team can't review every detail in the data they use. In order to make data-driven decisions and draw conclusions, what they need is a distilled version of the data. This is where smart visualization can be of the highest importance. It can allow readers understand "what is going on" in the data in just a few moments instead of having to undertake an annoying, time consuming analysis.

Nowadays, advanced data visualizations go beyond graphs and charts to help in the process of making crucial business decisions. Several visualization formats are available: static, zooming, clickable, animated, video, or interactive. The choice between these depends on the overall objectives of the visualization. While static is the simplest and the most common form of visualization, the interactive options are becoming more popular because they give users some control over the displayed information.

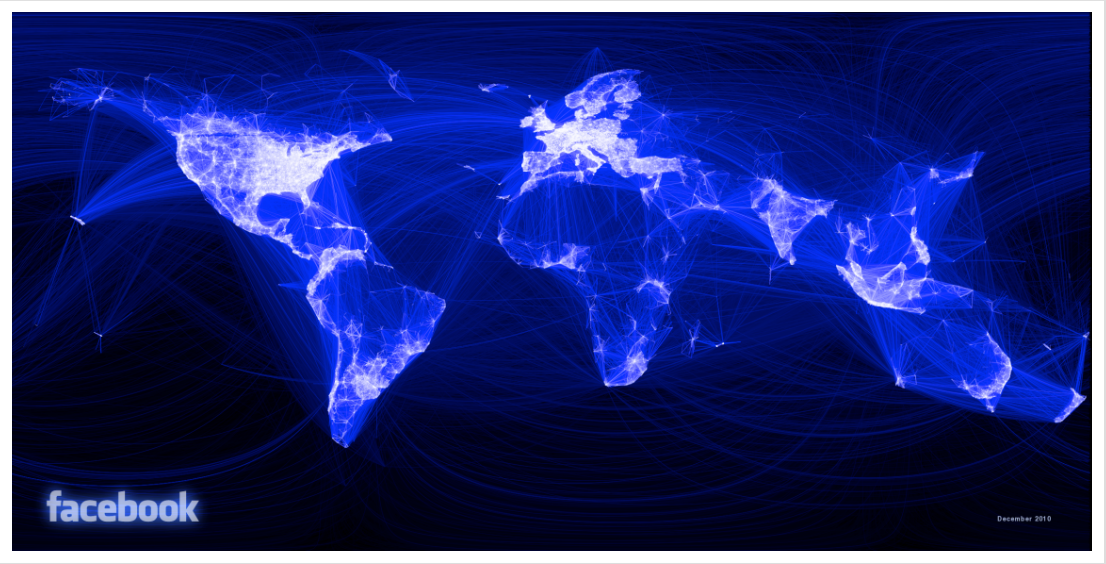
Many interesting and good visualizations can be found on the Visualizing.org portal (**http://www.visualizing.org/**) which can provide you with ideas for new visualization forms. One such example is an interactive dashboard providing global and local trends in human migration during the last 50 years:



Explore the evolution of human migration from 1960 to 2012. Explore origins and destinations

In just a few seconds this dashboard reveals some interesting patterns, such as the migration “boom” of the 1980s, the consequent growth of migration worldwide and how armed conflicts produce big human movements in specific periods.

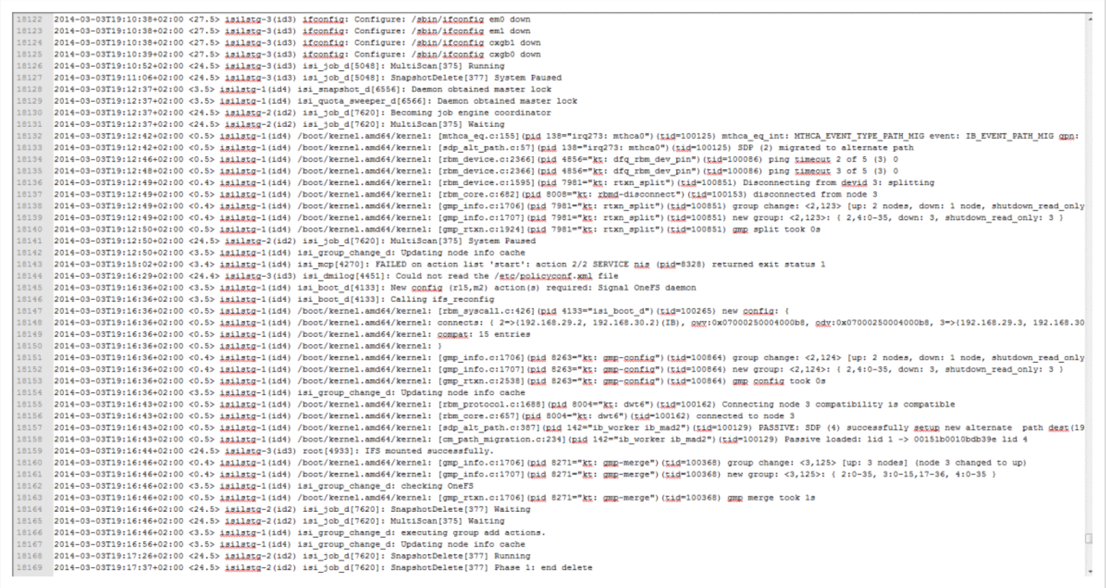
Another interesting visualization is Facebook’s friendship locality map where the lines represent real human relationships:



Besides the fact that continents and certain international borders are visible on this map, a reader can observe the level of Facebook popularity and connectivity between people in different areas of the word. Also, the areas of low or lack of activity become immediately apparent and can be considered for more powerful marketing efforts. This is a nice example of a visualization highlighting specific relations within the data. While in the case of Facebook these are relations between social network users, in other cases it can be relations between devices, various types of events or services.

The above visualizations are examples of an efficient way to communicate data. Successful data visualizations are very space efficient and display all the data within a single field of view. This allows a reader to see the entire picture with minimal eye movement and without scrolling or flipping between the pages. In this post we will show how a very simple data analysis along with a creative visualization can help to asses a server's status in a few seconds and save several days of work.

As a customer-facing organization, an important task for field support specialist at EMC is Root Cause Analysis (RCA) of a problem in a customer's installation. RCA usually requires looking at records and manually correlating events from multiple log files. Although log files are commonly used for various support purposes, they are often very large and take a textual form that is difficult to follow. For example, below is a screen capture of part of a log file:



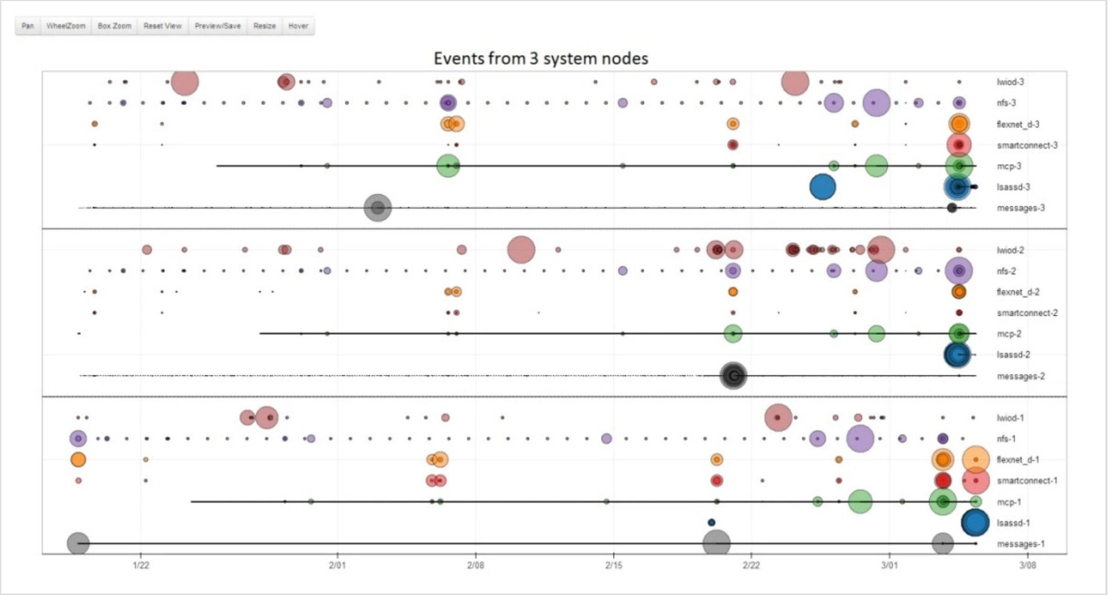
```
2014-03-03T19:10:38+02:00 <27.5> isilstg-3(id3) ifconfig: Configure: /sbin/ifconfig em0 down
2014-03-03T19:10:38+02:00 <27.5> isilstg-3(id3) ifconfig: Configure: /sbin/ifconfig em1 down
2014-03-03T19:10:38+02:00 <27.5> isilstg-3(id3) ifconfig: Configure: /sbin/ifconfig cxgb1 down
2014-03-03T19:10:39+02:00 <27.5> isilstg-3(id3) ifconfig: Configure: /sbin/ifconfig cxgb0 down
2014-03-03T19:10:52+02:00 <24.5> isilstg-3(id3) isi_job_d[5048]: MultiScan[375] Running
2014-03-03T19:11:06+02:00 <24.5> isilstg-3(id3) isi_job_d[5048]: SnapshotDelete[377] System Paused
2014-03-03T19:12:37+02:00 <3.5> isilstg-1(id4) isi_snapshot_d[6556]: Daemon obtained master lock
2014-03-03T19:12:37+02:00 <3.5> isilstg-1(id4) isi_quota_sweeper_d[6566]: Daemon obtained master lock
2014-03-03T19:12:37+02:00 <24.5> isilstg-2(id2) isi_job_d[7620]: Becoming job engine coordinator
2014-03-03T19:12:37+02:00 <24.5> isilstg-2(id2) isi_job_d[7620]: MultiScan[375] Waiting
2014-03-03T19:12:42+02:00 <0.5> isilstg-1(id4) /boot/kernel.amd64/kernel: [mthca_eq.c:155](pid 138="irq273: mthca0")(tid=100125) mthca_eq_int: MTHCA_EVENT_TYPE_PATH_MIG event: IB_EVENT_PATH_MIG gpn:
2014-03-03T19:12:42+02:00 <0.5> isilstg-1(id4) /boot/kernel.amd64/kernel: [sdp_alt_path.c:57](pid 138="irq273: mthca0")(tid=100125) SDP (2) migrated to alternate path
2014-03-03T19:12:46+02:00 <0.5> isilstg-1(id4) /boot/kernel.amd64/kernel: [rbm_device.c:2366](pid 4856="kt: dfq_rbm_dev_pin")(tid=100086) ping timeout 2 of 5 (3) 0
2014-03-03T19:12:48+02:00 <0.5> isilstg-1(id4) /boot/kernel.amd64/kernel: [rbm_device.c:2366](pid 4856="kt: dfq_rbm_dev_pin")(tid=100086) ping timeout 3 of 5 (3) 0
2014-03-03T19:12:49+02:00 <0.4> isilstg-1(id4) /boot/kernel.amd64/kernel: [rbm_device.c:1595](pid 7981="kt: rtxn_split")(tid=100851) Disconnecting from devid 3: splitting
2014-03-03T19:12:49+02:00 <0.5> isilstg-1(id4) /boot/kernel.amd64/kernel: [rbm_core.c:682](pid 8008="kt: rbmd-disconnect")(tid=100153) disconnected from node 3
2014-03-03T19:12:49+02:00 <0.4> isilstg-1(id4) /boot/kernel.amd64/kernel: [gmp_info.c:1706](pid 7981="kt: rtxn_split")(tid=100851) group change: <2,123> [up: 2 nodes, down: 1 node, shutdown_read_only
2014-03-03T19:12:49+02:00 <0.4> isilstg-1(id4) /boot/kernel.amd64/kernel: [gmp_info.c:1707](pid 7981="kt: rtxn_split")(tid=100851) new group: <2,123>: { 2,4:0-35, down: 3, shutdown_read_only: 3 }
2014-03-03T19:12:50+02:00 <0.5> isilstg-1(id4) /boot/kernel.amd64/kernel: [gmp_rtxn.c:1924](pid 7981="kt: rtxn_split")(tid=100851) gmp split took 0s
2014-03-03T19:12:50+02:00 <24.5> isilstg-2(id2) isi_job_d[7620]: MultiScan[375] System Paused
2014-03-03T19:12:50+02:00 <3.5> isilstg-1(id4) isi_group_change_d: Updating node info cache
2014-03-03T19:15:02+02:00 <3.4> isilstg-1(id4) isi_mcp[4270]: FAILED on action list 'start': action 2/2 SERVICE nis (pid=8328) returned exit status 1
2014-03-03T19:16:29+02:00 <24.4> isilstg-3(id3) isi_dmilog[4451]: Could not read the /etc/policyconf.xml file
2014-03-03T19:16:36+02:00 <3.5> isilstg-1(id4) isi_boot_d[4133]: New config (r15,m2) action(s) required: Signal OneFS daemon
2014-03-03T19:16:36+02:00 <3.5> isilstg-1(id4) isi_boot_d[4133]: Calling ifs_reconfig
2014-03-03T19:16:36+02:00 <0.5> isilstg-1(id4) /boot/kernel.amd64/kernel: [rbm_syscall.c:426](pid 4133="isi_boot_d")(tid=100265) new config: {
2014-03-03T19:16:36+02:00 <0.5> isilstg-1(id4) /boot/kernel.amd64/kernel: connects: { 2=>(192.168.29.2, 192.168.30.2)(IB), gwy:0x07000250004000b8, gdv:0x07000250004000b8, 3=>(192.168.29.3, 192.168.30
2014-03-03T19:16:36+02:00 <0.5> isilstg-1(id4) /boot/kernel.amd64/kernel: compat: 15 entries
2014-03-03T19:16:36+02:00 <0.5> isilstg-1(id4) /boot/kernel.amd64/kernel: }
2014-03-03T19:16:36+02:00 <0.4> isilstg-1(id4) /boot/kernel.amd64/kernel: [gmp_info.c:1706](pid 8263="kt: gmp-config")(tid=100864) group change: <2,124> [up: 2 nodes, down: 1 node, shutdown_read_only
2014-03-03T19:16:36+02:00 <0.4> isilstg-1(id4) /boot/kernel.amd64/kernel: [gmp_info.c:1707](pid 8263="kt: gmp-config")(tid=100864) new group: <2,124>: { 2,4:0-35, down: 3, shutdown_read_only: 3 }
2014-03-03T19:16:36+02:00 <0.5> isilstg-1(id4) /boot/kernel.amd64/kernel: [gmp_rtxn.c:2538](pid 8263="kt: gmp-config")(tid=100864) gmp config took 0s
2014-03-03T19:16:36+02:00 <3.5> isilstg-1(id4) isi_group_change_d: Updating node info cache
2014-03-03T19:16:43+02:00 <0.5> isilstg-1(id4) /boot/kernel.amd64/kernel: [rbm_protocol.c:1688](pid 8004="kt: dwt6")(tid=100162) Connecting node 3 compatibility is compatible
2014-03-03T19:16:43+02:00 <0.5> isilstg-1(id4) /boot/kernel.amd64/kernel: [rbm_core.c:657](pid 8004="kt: dwt6")(tid=100162) connected to node 3
2014-03-03T19:16:43+02:00 <0.5> isilstg-1(id4) /boot/kernel.amd64/kernel: [sdp_alt_path.c:387](pid 142="ib_worker ib_mad2")(tid=100129) PASSIVE: SDP (4) successfully setup new alternate  path dest(19
2014-03-03T19:16:43+02:00 <0.5> isilstg-1(id4) /boot/kernel.amd64/kernel: [cm_path_migration.c:234](pid 142="ib_worker ib_mad2")(tid=100129) Passive loaded: lid 1 -> 00151b0010bdb39e lid 4
2014-03-03T19:16:44+02:00 <24.5> isilstg-3(id3) root[4933]: IFS mounted successfully.
2014-03-03T19:16:46+02:00 <0.4> isilstg-1(id4) /boot/kernel.amd64/kernel: [gmp_info.c:1706](pid 8271="kt: gmp-merge")(tid=100368) group change: <3,125> [up: 3 nodes] (node 3 changed to up)
2014-03-03T19:16:46+02:00 <0.4> isilstg-1(id4) /boot/kernel.amd64/kernel: [gmp_info.c:1707](pid 8271="kt: gmp-merge")(tid=100368) new group: <3,125>: { 2:0-35, 3:0-15,17-36, 4:0-35 }
2014-03-03T19:16:46+02:00 <3.5> isilstg-1(id4) isi_group_change_d: checking OneFS
2014-03-03T19:16:46+02:00 <0.5> isilstg-1(id4) /boot/kernel.amd64/kernel: [gmp_rtxn.c:1706](pid 8271="kt: gmp-merge")(tid=100368) gmp merge took 1s
2014-03-03T19:16:46+02:00 <24.5> isilstg-2(id2) isi_job_d[7620]: SnapshotDelete[377] Waiting
2014-03-03T19:16:46+02:00 <24.5> isilstg-2(id2) isi_job_d[7620]: MultiScan[375] Waiting
2014-03-03T19:16:46+02:00 <3.5> isilstg-1(id4) isi_group_change_d: executing group add actions.
2014-03-03T19:16:56+02:00 <3.5> isilstg-1(id4) isi_group_change_d: Updating node info cache
2014-03-03T19:17:26+02:00 <24.5> isilstg-2(id2) isi_job_d[7620]: SnapshotDelete[377] Running
2014-03-03T19:17:37+02:00 <24.5> isilstg-2(id2) isi_job_d[7620]: SnapshotDelete[377] Phase 1: end delete
```

There are tens of thousands of entries in a log file and hundreds of such files aggregated for each distinct machine or system. In scaled out systems, where a single device is comprised of several execution nodes, an additional layer of complexity is added. In this setting, manually digging through the heaps of logs becomes a difficult and time-consuming task that may require several days to complete. Although a few tools for analyzing log files exist, none are suitable for the specific RCA purpose which each of the field support specialists at EMC faces on a daily basis.

Below we show how a quick 30-minute python script can help reduce the time and effort invested by a specialist in analyzing multiple log files from several days to a few minutes or seconds. As an example, let us consider the analysis of log files from one of EMC's systems. Recently, in one of our engagements, we experimented with a set of log files originating from a storage system with multiple nodes.

The purpose was to provide an easy-to-use support system for the correlation of multiple events over time so that a field support engineer can quickly identify events of interests that may have triggered a given problem. Note that it is different from the server's health status monitoring where continuous and detailed analysis of numerous parameters is needed. Oppositely, a summary view over a long time period and relations between multiple events are of the main interest here.

After some data cleaning and processing, we can derive a list of the events occurring in the system across time (see the detailed explanation of the preprocessing at the bottom of this post) and use this data to power an explorative visualization of events. The events timeline was provided as an HTML file with several interactive capabilities (e.g. zoom, resize, hover, etc.) as shown in the screenshot below.



The chart represents a count of events across almost two months period of time taken from a set of predefined log files residing in each node of the system. Larger circles represent a higher count value (circle sizes are normalized with respect to each log file). Our visualization detects a system-wide ("global") event, clearly visible in the right hand side of the figure, towards the end of the monitoring period. The field support specialist can use the visualization tool to zoom in on areas of interest for further analysis within a few seconds or get additional information about events by hovering over the circles.

As mentioned above, this is just an example on how to extend the basic capabilities log analysis tools have. The general idea is to leverage an existing distributed file system such as HDFS, and build a monitoring tool which processes and analyzes the data from log files using MapReduce in parallel. The results can then be published as an easy-to-use, web-based, interface that can drill down into the system, examining its overall health. Among other things, we envision features such as a visual timeline of events, log content analysis, and quick data access via zooming in on events of interest. For applications where real-time logs processing might be valuable, other higher performance tools such as GemFire can be leveraged.

For interested readers, we provide Python code sample which can be used for the basic log files parsing. As one may note from the code snippets the most interesting and challenging here was the visualization part: to fit hundreds thousands of events and a two-month timeline with five minute intervals on a single screen so that it can be easily and clearly interpreted by a human; while the actual parsing and analytical calculations happen within a few python code rows only.

```
+ expand source
```

For events timeline visualization we have used Bokeh – an interactive web plotting library for Python.

```
+ expand source
```

To learn more about EMC IT Data Science efforts, read previous blogs from our data scientists:

- ***Data Science as a Service: Driving Agility and Innovation to the ITaaS Model***
- ***Text Analytics: Easy Classification for Routing Service Requests***
- ***The Price is Right: Predicting Cost of Support Contracts for Complex Products***

SHARE THIS:

Share 0 Tweet Share G+

Like

One blogger likes this.

RELATED

From Neuroscience to Data Science
In "Analytics"

Predictive Analytics for IT Operations: Continuing the Journey
In "Analytics"

The Business Data Lake from a Data Scientist Perspective
In "Big Data"

This entry was posted in **Analytics**, **Big Data**, **Cloud**, **Data Science**, **Virtualization** and tagged **Analytics**, **Big Data**, **Cloud**, **data science** by **EMC IT Proven**. Bookmark the **permalink**

[https://emcvirtualizationjourney.wordpress.com/2014/12/12/smart-data-visualization-helping-decision-makers-get-the-picture/] .

About EMC IT Proven

Come a long for the ride with EMC IT and experience our in-house virtualization journey. As we encounter challenges and develop new IT best practices. The series will continue until EMC IT has developed their own Enterprise Private Cloud.
View all posts by EMC IT Proven →

2 THOUGHTS ON "SMART DATA VISUALIZATION: HELPING DECISION MAKERS GET THE PICTURE"

AnalyticExec
on **December 14, 2014 at 12:53 am** said:

Reblogged this on The Analytic Executive.

Pingback: From Neuroscience to Data Science | EMC IT Proven