INTEGRATE WEB APPS WITH APACHE

LINUX JOURNAL™

Since 1994: The Original Magazine of the Linux Community

WATCH:
ISSUE
OVERVIEW

MARCH 2017 | ISSUE 275
http://www.linuxjournal.com

BIG DATA, Hadoop and R

+ AUTOMATION TIPS for SysAdmins

MANIPULATING IMAGES with ImageMagick

A LOOK AT
Cutting the Cable Cord

Cloud-Scale Automation with Puppet

Author:
John S. Tonello

Sponsor:
Puppet

Why Innovative App Developers Love High-Speed OSDBMS

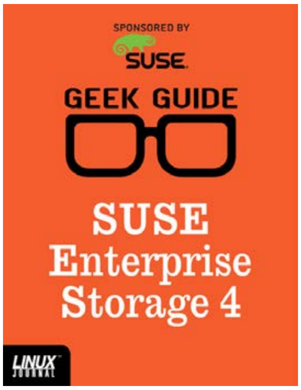
Author:
Ted Schmidt

Sponsor:
IBM

Tame the Docker Life Cycle with SUSE

Author:
John S. Tonello

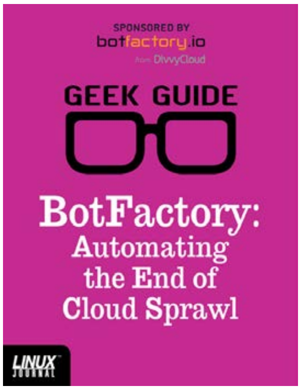
Sponsor:
SUSE



SUSE Enterprise Storage 4

Author:
Ted Schmidt

Sponsor:
SUSE



BotFactory: Automating the End of Cloud Sprawl

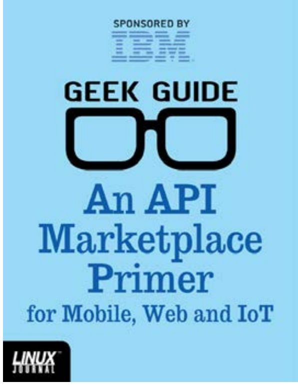
Author:
John S. Tonello

Sponsor:
BotFactory.io

Containers 101

Author:
Sol Lederman

Sponsor: Puppet



An API Marketplace Primer for Mobile, Web and IoT

Author:
Ted Schmidt

Sponsor:
IBM

Public Cloud Scalability for Enterprise Applications

Author:
Petros Koutoupis

Sponsor:
SUSE

CONTENTS

MARCH 2017
ISSUE 275

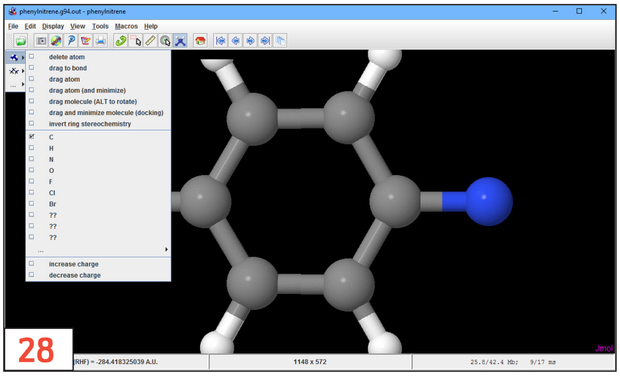
FEATURES

Cover Image © Can Stock Photo / kentoh

CONTENTS

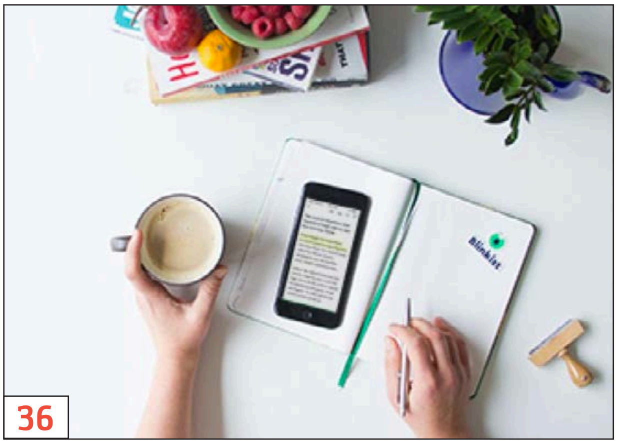
COLUMNS

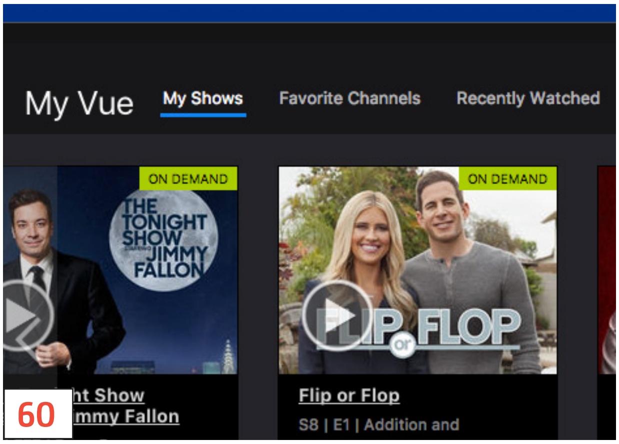
IN EVERY ISSUE



phenylnitrene.g94.out - phenylnitrene
File Edit Display View Tools Macros Help
delete atom
drag to bond
drag atom
drag atom (and minimize)
drag molecule (ALT to rotate)
drag and minimize molecule (docking)
invert ring stereochemistry
C
H
N
O
F
Cl
Br
??
??
??
...
increase charge
decrease charge
(RHF) = -284.418325039 A.U.
1148 x 572

28


36



60

ON THE COVER

- Integrate Web Apps with Apache, p. 96
- Big Data, Hadoop and R, p. 80
- A Look at Cutting the Cable Cord, p. 60
- Automation Tips for Sysadmins, p. 54
- Manipulating Images with ImageMagick, p. 48

LINUX JOURNAL (ISSN 1075-3583) is published monthly by Belltown Media, Inc., PO Box 980985, Houston, TX 77098 USA.
Subscription rate is $29.50/year. Subscriptions start with the next issue.

LINUX JOURNAL

Executive Editor	Jill Franklin jill@linuxjournal.com
Senior Editor	Doc Searls doc@linuxjournal.com
Associate Editor	Shawn Powers shawn@linuxjournal.com
Art Director	Garrick Antikajian garrick@linuxjournal.com
Products Editor	James Gray newproducts@linuxjournal.com
Editor Emeritus	Don Marti dmarti@linuxjournal.com
Technical Editor	Michael Baxter mab@cruzio.com
Senior Columnist	Reuven Lerner reuven@lerner.co.il
Security Editor	Mick Bauer mick@visi.com
Hack Editor	Kyle Rankin lj@greenfly.net
Virtual Editor	Bill Childers bill.childers@linuxjournal.com

Contributing Editors
Ibrahim Haddad • Robert Love • Zack Brown • Dave Phillips • Marco Fioretti • Ludovic Marcotte
Paul Barry • Paul McKenney • Dave Taylor • Dirk Elmendorf • Justin Ryan • Adam Monsen

President	Carlie Fairchild publisher@linuxjournal.com
Publisher	Mark Irgang mark@linuxjournal.com
Associate Publisher	John Grogan john@linuxjournal.com
Director of Digital Experience	Katherine Druckman webmistress@linuxjournal.com
Accountant	Candy Beauchamp acct@linuxjournal.com

***Linux Journal* is published by, and is a registered trade name of,**
Belltown Media, Inc.
PO Box 980985, Houston, TX 77098 USA

Editorial Advisory Panel
Nick Baronian
Kalyana Krishna Chadalavada
Brian Conner • Keir Davis
Michael Eager • Victor Gregorio
David A. Lane • Steve Marquez
Dave McAllister • Thomas Quinlan
Chris D. Stark • Patrick Swartz

Advertising
E-MAIL: ads@linuxjournal.com
URL: www.linuxjournal.com/advertising
PHONE: +1 713-344-1956 ext. 2

Subscriptions
E-MAIL: subs@linuxjournal.com
URL: www.linuxjournal.com/subscribe
MAIL: PO Box 980985, Houston, TX 77098 USA

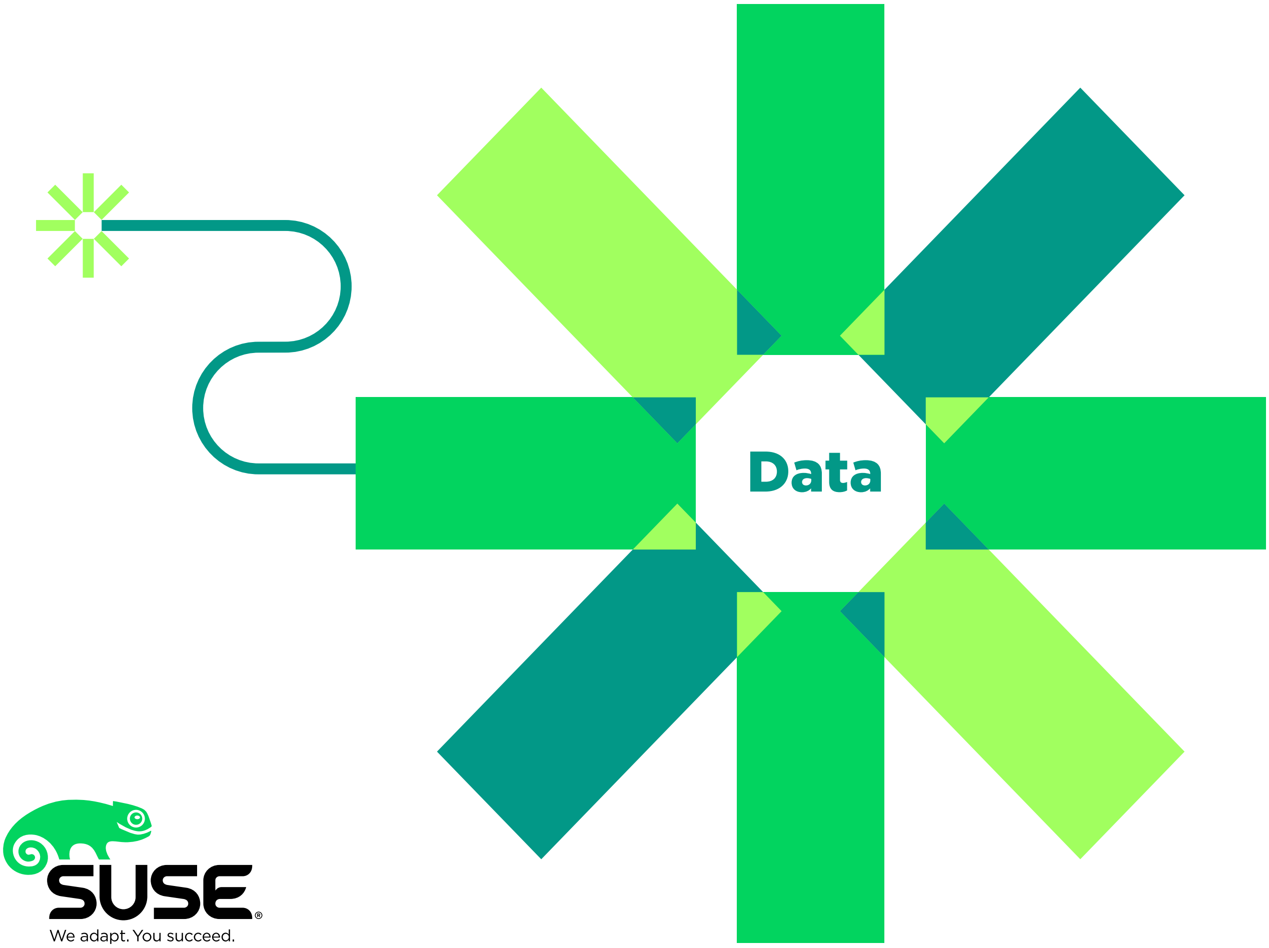
STORAGE R∃DEFINED:

You can~~not~~ keep up with data explosion.

Manage data expansion with SUSE Enterprise Storage.

SUSE Enterprise Storage, the leading open source storage solution, is highly scalable and resilient, enabling high-end functionality at a fraction of the cost.

suse.com/storage



SUSE
We adapt. You succeed.

Current_Issue.tar.gz

It’s a Bird! It’s a Plane! Nope, It’s My Server!

SHAWN POWERS

Shawn Powers is the Associate Editor for *Linux Journal*. He’s also the Gadget Guy for LinuxJournal.com, and he has an interesting collection of vintage Garfield coffee mugs. Don’t let his silly hairdo fool you, he’s a pretty ordinary guy and can be reached via email at shawn@linuxjournal.com. Or, swing by the #linuxjournal IRC channel on Freenode.net.

Like most fancy tech terms, “Cloud Computing” has lost its newness, and it’s now just a commodity we purchase. It’s often so much easier to provision virtual machines than it is to buy and host your own servers. Yes, there are concerns over privacy and security when your data is in the cloud. When you host in your own data center, however, there’s still the possibility of a rogue cleaning crew getting to your servers. (We’ve all seen the movies; it just takes a mop and a blue jumper to get you into the most secure data center.) Regardless of your stance on cloud computing, it’s here to stay. This month, we talk a bit about how to live in this bold new world.

Reuven M. Lerner starts things off with more information about machine learning. What if the software itself could suss out the important patterns and information from your data instead of feeding it presorted information for simple data mining? Reuven describes “unsupervised machine learning” this month, and it’s either awesome or terrifying, depending on how you feel about such things.

Dave Taylor follows with a look at one of my

▶

VIDEO: Shawn Powers runs through the latest issue.

favorite command-line application suites: ImageMagick. I use a few of the tools with my ongoing birdcam project. Dave explores all sorts of ways to manipulate images from the command line, and his article is only the tip of the iceberg!

Kyle Rankin delves into automation for sysadmins this month. Thanks to DevOps and the like, task automation is far more robust than it used to be. Kyle explains why we should automate, when we should automate and how we should go about doing it. Although it's tempting to replace your system administrator with a few DevOps tools, that human factor is invaluable.

I take a look at entertainment this month, specifically as it pertains to watching television. Many of you are "cord cutters", but for the few still holding on to their cable subscription (such as myself), it's important to understand the various options out there. Thankfully, we're closer to a complete television experience over the internet than ever before. This month, I look at some of the available options.

Rune Torbensen and Søren Top describe building a Linux cluster in the cloud for the purpose of data analysis with R. Building a huge cluster of machines in your own data center is impractical for most people, especially when that cluster isn't needed as a permanent fixture. Rune and Søren show how to take advantage of the cloud along with open-source tools like Hadoop and R to analyze logs.

Andy Carlson follows them with an article on writing custom Apache configurations in order to run specific applications on the web. Yes, it's the sort of thing you can do on your own hardware, but again, the cloud is a convenient way to treat computing like a commodity, so having methods that are flexible are vital in our modern IT world.

Love it or hate it, the cloud isn't going away any time soon. As someone who historically has had server racks full of last-generation servers in his basement (that's all I could afford), I can't express how happy the cloud makes me. Before you can safely take advantage of using someone else's servers for your data, it's important to understand not only how the cloud itself works, but also how your software integrates. And understanding open-source software is what we love most here at *Linux Journal*! ■

LETTERS

PREVIOUS
Current_Issue.tar.gz

NEXT
UpFront

Stop

Regarding Doc Searls' "Debugging Democracy" in the January 2017 issue: please stop printing childish personal insults in *Linux Journal*. When you refer to the President-Elect as "Internet troll Donald Trump", you are being personally insulting, childish by playing funny games with someone's name ("heh heh, he's got 'rump' in his name!"), and promoting politics of hate and fear. This has no place in a technical journal and no relevance to Linux or computing in general.

You would not refer to the losing candidate as a "shrew", nor would you play childish games with her name as a way of insulting her. In fact, you did refer to her, by her correct name, without personal insult.

I'm sorry, but the election process does not need "debugging" because your favorite lost. This has happened every four years since the ratification of the US Constitution—someone wins, someone loses. It's a direct and inevitable side effect of having one President instead of two. Every time, nearly half the voters are disappointed.

By picking sides and using personal insults to make political commentary after the election is over and done, you disappoint upwards of half of your subscribers. Doc Searls owes an apology to the readers, if not to Mr Trump himself.

—Mark Kramer

Doc Searls replies: *Mark is right. I do owe readers an apology. By calling*

Donald Trump a troll (take a look at the Wikipedia definition of internet troll at https://en.wikipedia.org/wiki/Internet_troll *and draw your own conclusions), I was being a troll as well. Even though trolling wasn't my intent, that has been the effect so far: every response to my January column, both here and on our website, has been as negative as Mark's, and for the same reasons.*

Opening with that remark also failed to support the main purpose of that column, which was to call for help in rescuing journalism—and real journals such as this one—from drowning in a sea of "content", way too much of which is crap routed by algorithms aimed by surveillance-gathered data into echo chambers of the like-minded. This has the effect of increasing enmity and blame toward those in echo chambers with opposing sympathies, which is worse than dangerous in democratic societies, because it tears apart the center spaces of basic agreement those societies require. You can see how this looks in The Wall Street Journal*'s Blue Feed, Red Feed site (*http://graphics.wsj.com/blue-feed-red-feed*), subtitled "See Liberal Facebook and Conservative Facebook, Side by Side".*

I am sure most of the systems driving us into hostile camps are built on Linux. (Isn't everything now?) So I don't think I'm off base calling for help here.

Shotcut Video Editor

I just read Shawn Powers' August 2016 *Linux Journal* article about his traveling gear (yes, I am that far behind) and besides being shocked that he uses a MacBook Air (just kidding, Apple hardware is good enough), I was also shocked

LINUX JOURNAL

At Your Service

SUBSCRIPTIONS: *Linux Journal* is available in a variety of digital formats, including PDF, .epub, .mobi and an online digital edition, as well as apps for iOS and Android devices. Renewing your subscription, changing your email address for issue delivery, paying your invoice, viewing your account details or other subscription inquiries can be done instantly online: **http://www.linuxjournal.com/subs**. Email us at subs@linuxjournal.com or reach us via postal mail at *Linux Journal*, PO Box 980985, Houston, TX 77098 USA. Please remember to include your complete name and address when contacting us.

ACCESSING THE DIGITAL ARCHIVE: Your monthly download notifications will have links to the various formats and to the digital archive. To access the digital archive at any time, log in at **http://www.linuxjournal.com/digital**.

LETTERS TO THE EDITOR: We welcome your letters and encourage you to submit them at **http://www.linuxjournal.com/contact** or mail them to *Linux Journal, PO Box 980985,* Houston, TX 77098 USA. Letters may be edited for space and clarity.

WRITING FOR US: We always are looking for contributed articles, tutorials and real-world stories for the magazine. An author's guide, a list of topics and due dates can be found online: **http://www.linuxjournal.com/author**.

FREE e-NEWSLETTERS: *Linux Journal* editors publish newsletters on both a weekly and monthly basis. Receive late-breaking news, technical tips and tricks, an inside look at upcoming issues and links to in-depth stories featured on **http://www.linuxjournal.com**. Subscribe for free today: **http://www.linuxjournal.com/enewsletters**.

ADVERTISING: *Linux Journal* is a great resource for readers and advertisers alike. Request a media kit, view our current editorial calendar and advertising due dates, or learn more about other advertising and marketing opportunities by visiting us on-line: **http://www.linuxjournal.com/advertising**. Contact us directly for further information: ads@linuxjournal.com or +1 713-344-1956 ext. 2.

that he uses Final Cut Pro for his video editing.

For the editing needs he described, there is no excuse for not using the excellent Shotcut video editor (https://shotcut.org). Besides the fact that it has more than enough features, it's open source and available for Linux, Windows and Mac OS.

I am just wondering what kind of excuse he will give for not adopting Shotcut for his video editing needs. I am not pressuring him with this, just playing and using this opportunity to let him know about Shotcut.

I'm not connected with the project; I just use, advocate and provide some help with my small knowledge in the Shotcut forum. See also the videos on YouTube for some help getting started.
—Luis Sismeiro

Shawn Powers replies: *It's easy to feel a bit defensive with questions like, "what's your excuse?", but tone is often easy to misinterpret in email, so I'm going to assume this was a friendly message. I don't think I need an excuse, because I don't think I've done anything wrong. But I'll answer the question of "why", because that's a fair one.*

It's possible that some of this was answered in my article, or in issues past, but really quickly: I have several jobs, and those jobs require me to have various computer systems. On a daily basis, I truly use Windows, OS X and Linux. I love open-source software, but I'm not a zealot. I use whatever tool I can use to get the job done. For Linux Journal, *it makes sense to use Linux software for video editing. However, all the times I've attempted to do that through the years, it's been inefficient at best and impossible at the worst.*

I've never tried Shotcut, but now that you've brought it to my attention, I'll be giving it a try. Heck, perhaps I'll write about it. The thing that's important to know though is that if it is a program that crashes or doesn't work well for me, I likely won't use it.

So after all that, thank you for bringing the project to my attention. I'll definitely check it out!

Fake News

I read Mr Searls' tantrum in the January 2017 issue with great amusement. It seems he is not a fan of Donald Trump, but rather than calling for the death of the Electoral College (the other fad *du jour*), he says "we need to hack the news back in a logical direction and away from the fact-free, misleading and emotion stirring ways that news is made today". In other words "Doc" is calling for global "fact-checking" on the internet—aka a globalized Wikipedia. And who, pray tell, will be trusted with that curation process? We don't need to look far for an answer, because others have suggested similar things before with regard to broadcast news: https://en.wikipedia.org/wiki/Fairness_Doctrine.

So the creative chaos of the Bazaar is good for Linux, but bad in the arena of politics and news? Interesting how when liberals lose elections their immediate instincts are to change the Constitutional process and call for clamping down on radio, TV and now the internet. The First Amendment applies only to whoever has the correct views. Wouldn't it be better if we had an educated electorate who could smell truth from fiction on their own? But how often is critical thinking taught today? Let us take this as a teachable moment.

Mr Searls goes on to show some pretty graphs and points out that "This kind of study does not show a mandate...." True, the election of Trump alone does not show a mandate, but what does this table show?

	2008(DEM/REP)	2016(DEM/REP)	
State Governors	28/22	18/31	from Wikipedia
State Senates	28/20	13/37	https://ballotpedia.org/Partisan_composition_of_state_senates
State House	32/16	18/31	https://ballotpedia.org/Partisan_composition_of_state_houses
US Senate	49/49	44/54	from Wikipedia
US House	233/198	194/241	https://ballotpedia.org/United_States_Congress_elections,_2016

It's almost as if a vast, right-wing conspiracy infiltrated every state/local/national news outlet, then hired an army of covert East-bloc operatives to create "fake news" on social media, all designed to sway the election up and down the ballot completely inverting the majority control at state and federal levels since Barack Obama has been elected.

Hillary would no doubt agree.

—Steve Langer

Doc Searls replies: *I meant "hack" in the broad sense it has been used here since* Linux Journal *began in 1994. If you want a specific definition (or a set of them), consult the Jargon File:* http://www.catb.org/jargon/html*.*

The Trump vs. Hillary contest was maximally interesting at the time I wrote the column, but it was also beside the main point I tried to make: there are dangerously dysfunctional ways our democracy now informs itself in the networked world. "Fake news" is currently the most obvious example, although I believe the real problems are deeper and more systematic than that. However one looks at it, some fixing is required.

It's about the how *of democracy, not about the* who*.*

Mars Lander Program

Regarding Dave Taylor's Mars Lander program in the September, October and November 2016 issues: it's a curious game in a few lines of shell script. I appreciated the final version. But `bc` is not so friendly of a tool, and it puts some unfair mistakes in the game. Here's a solution for `bc` uses:

```
speed=$( $bc <<< "scale=3; $speed + $gravity + $thrust" | sed -e
's/^\([-]*\)\./\10./')
...
altitude=$( $bc <<< "scale=3; $altitude + $speed" | sed -e
's/^\([-]*\)\./\10./')
```

Thank you, good work!

—José Nicolau

Dave Taylor replies: *Interesting tweak to the script. I'll have to send it over to my NASA friends to QA!*

kpcli—KeePass Command-Line Interface

Related to der.hans' January 2017 article on password managers ("Online Privacy and Security Using a Password Manager"), I thought that letting your audience know about kpcli might be useful: http://kpcli.sourceforge.net.

—Lester Hightower

Shawn Powers' Synology Review

Regarding Shawn Powers' "My Love Affair with Synology" in the January 2017 issue: many applications are difficult to install directly on a Synology because Synology does not use a standard repository format like RPM or APT, but it does support Docker! My tests with Docker have been great so far. I can build and test Docker images on my Debian server and then copy the finished images to the Synology for deployment.

In my case, I am trying to get PostGIS/Geoserver/Geonode going. I upgraded from 1 to 4GB of RAM to support it. I need to support a very small team of users sharing geospatial data, so we don't need a lot of compute horsepower, just a shared data store. End users will probably be using QGIS as a client, so most of the computation will take place on their laptops.

But the use case that I applied to justify buying the Synology was "Synology Cloud Station" (it might have come out after you wrote your review). It is great; files on the server are not kept locked up in a special container (like ownCloud), so you can seamlessly drop files into an ordinary folder on the Synology and have them replicate out to your Cloud Station clients.

This means LAN users can see files directly via Samba, AppleTalk or NFS and not have to copy them to their own hard drives, but I (working at home) also get access via synchronization. Working remotely, I can export a large format map to a PDF file, and it will be uploaded automatically via Cloud Station sync. Then my team can view the map from the Synology via LAN file share or web server (File Station) without syncing a copy to their own laptops.

I love ownCloud by the way, and it will run on a Synology (I tried it out), and I use it on my own Debian server, but Cloud Station fits our use cases better.
—Brian Wilson

Shawn Powers replies: *I'm fairly certain Cloud Station has been there for a while, but I have to admit I haven't tried it. (That will change!) My concerns with running things on the Synology directly are all horsepower-related. I love it for things like reverse proxying, web hosting and torrent management. My Plex Media Server, however, I put on a separate box because I fear the Synology wouldn't be able to manage the transcoding. I also share your frustration with the packages provided by Synology, but thankfully, there are some community-maintained programs that can be downloaded and installed.*

The Geo stuff you're doing sounds cool, by the way, and it sounds like a perfect use case since the CPU demands aren't too high. And thanks again for the tip about Cloud Station; I'll have to give it a try!

Curl Example in "Automated Slack Notifications" Doesn't Work

In the "Automatic Slack Notifications" piece in the December 2016 issue's UpFront section, the `curl` command to send a message to Slack doesn't work (at least on my MacBook). It gives the following error:

```
curl: (3) [globbing] unmatched close brace/bracket in column 8
```

In fact, the data block should be enclosed in single quotes, not double quotes.
—Khoa Le

Shawn Powers replies: *Double quotes and single quotes are often the bane of my programming. Half the time I get errors like you mention here, and the other half of the time I end up with output that looks like, "Thank you $NAME, your contribution to $THING was greatly appreciated!"*

In this case, the script worked for me on Linux, and once it worked, I didn't try it elsewhere. I could have, as I use OS X as well, but sadly, I didn't. Thanks for pointing out the issue. Hopefully everyone struggling will see this letter!

WRITE *LJ* A LETTER
We love hearing from our readers. Please send us your comments and feedback via http://www.linuxjournal.com/contact.

PHOTOS
Send your Linux-related photos to ljeditor@linuxjournal.com, and we'll publish the best ones here.

RETURN TO CONTENTS

LINUX JOURNAL
AN INDEPTH LOOK AT WI-FI TECHNOLOGY
DODGE BLOCKED NETWORKS with an RPi
+ Android Apps for Science
RASPBERRY PI
Amazon's Virtual Private Cloud and the New AWS CLI Tools
PERSONAL BOUNDARIES AND THE CLOUD
WATCH: ISSUE OVERVIEW

◀ PREVIOUS
Letters

NEXT
Editors' Choice ▶

diff -u

What's New in Kernel Development

Filesystem capabilities are supposed to be an improvement over simply running something as the root user. The idea is that you identify the specific special powers a program needs and then give it the ability to do only those special powers. Unfortunately, capabilities have become very complicated, with some individual capabilities being used to grant so many special powers that they might as well just be the root user after all.

In particular, kernel developers who create new powers don't always know of which capability that power should be a part, so any given capability can end up providing either too much or too little power to the program.

Michael Kerrisk recently began an effort to document some basic guidelines to help developers figure out which capability would best house any particular new power. For example, "Don't choose CAP_SYS_ADMIN if you can possibly avoid it!" Apparently **CAP_SYS_ADMIN** has become a huge dumping ground for powers of all sorts, falling prey to the might-as-well-be-root syndrome.

Unfortunately, **Casey Schaufler** pointed out some **POSIX** history that led to poor decisions being made early on, regarding how to organize filesystem capabilities. For example:

> Everyone involved was looking to use capabilities to meet B2 least privilege requirements in NSA security evaluations. Because those evaluations were of security policy, by far the easiest thing to do was to create a single capability for all the things that didn't show up in the security policy and declare that the people doing the evaluation didn't have to look over there.

Ultimately, my guess is that filesystem capabilities will have to be replaced by some kind of non-POSIX solution that's better thought out, but what that might look like remains an open question.

The **fbdev** drivers have been on the chopping block for quite a while now, as the **DRM** framework has been trying to replace them for years. But whenever anyone tries to get rid of those last straggling fbdev drivers, **Geert Uytterhoeven** or someone else always asks the same question: "Can DRM be used to create extremely simple display drivers?" And, the answer is always "Yes, absolutely! Not right now, but soon!" At which point Geert or someone else vetoes the expurgation, and the process begins again some months later.

But, this time it turned out, as **Daniel Vetter** noted, DRM has indeed advanced to the point of being able to produce those simple drivers right now. The community had heard the objections, and the community had answered. It took a little time, this time, for this to be made clear. Partly this was because Geert had gone through the rejection ritual so many times in the past that it had become an ingrained habit.

But once the truth finally became clear, Geert withdrew his objections, and now it looks as though the last few remaining fbdev drivers soon will be history—or at least, the remaining obstacles are no longer absolute deal-breakers.

This has been a long time coming, and the DRM folks have had to endure a lot of frustration in the process, so there was much ringing of bells when the path was cleared at last.

One thing that's no fun is when the **CPU** itself contains security holes. It's a real pain to discover that certain opcodes leak crucial information, because it essentially means that those opcodes never should be used. They're just wasted transistors, taking up space on the chip.

Paolo Bonzini recently wanted to disable several CPU instructions for **KVM**, such as the **SGDT** and **SIDT** opcodes, which he said could leak kernel addresses into userspace. Once leaked, those addresses could be used to defeat certain security measures, such as address layout randomization.

So, that's happening. Meanwhile, folks like **Liang Z Li** have offered to help lock down those issues.

Everyone wants to support **USB type C**, the new USB connector that works right side up and upside down. And, **Heikki Krogerus** of **Intel** recently posted some patches to support this. There was quite a bit of discussion and review of the patch, and enough problems arose that **Greg Kroah-Hartman** took a deeper look at the code. The patch turned out to have so many deep structural problems, Greg insisted that Heikki take the code back to Intel and have the engineers there give it a thorough going-over before Greg even would agree to look at it again.

So, that was harsh. Nobody likes to hear that their code is so bad that the upstream maintainer won't even look at future versions unless the downstream maintainers stage some sort of intervention. But, that's what happened.

Ultimately, USB type C support will be coming sooner rather than later. There's a lot of motivation to support it, given its popularity in the real world. I would imagine that the Intel engineers already are much closer to a proper patch.—Zack Brown

14th Annual

2017 HPC for Wall Street – Cloud & Data Centers Show & Conference

April 3, 2017 (Monday) **Roosevelt Hotel, NYC**

Madison Ave & 45th, next to Grand Central Station

Register Today.
Low-Cost Conference at $295.
Save $100.
Register for Free Show

The Next Big Game Changer.

New FinTech Technology Featured on April 3. All-Star Conference program for 2017.

Plan to attend the largest FinTech meeting of Cloud, Data Centers, HPC, Big Data, Networks, AI and Machine Learning, Trading, Low Latency for the Capital Markets.

- The Next-Generation FinTech Economy
- HPC Innovation for the FSI Market
- Addressing the Scalability Challenges of Deep Learning Model Training
- Optimizing New Data Technology and Developing HPC as a Service
- The AI Revolution Comes to Wall Street
- New Applications for High Performance Networking
- Enabling Financial Institutions to Build and Deploy Artificial Intelligence Applications using an Open Source Container Platform
- A New Approach to Data Storage and new Data Storage Architecture
- Hybrid Cloud, Private Cloud, Public Cloud
- Innovative Servers, Networks and Storage Applications

Plan to attend low-cost conference at $295. Save $100.

Qualified End Users may register as our guests, but must confirm their title and responsibility.

Register for the free show online at: www.flaggmgmt.com/linux

Partial List of 2017 Speakers.

Terry Roche, Principal, Head of FinTech Research, TABB Group

Donal Byrne CEO, Corvil

Lacee McGee Sr Prod Mgr, FSI Vertical Sols, HPE

Harvey Stein Head of Credit Risk, Modeling, Bloomberg

Dino Vitale TD Securities

Moiz Kohari (Invited) SVP Ch Tech Arch, State Street London

Anthony Golia FSI Chief Architect, Red Hat

David Rukshin CTO, WorldQuant (invited)

Asaf Wachtel Sr Dir, Business Dev, Mellanox Technologies

Roman Chwyl Head of Fin Svcs Google

Joseph George VP Solutions Strategy, SUSE

Andy Steinbach Sr Dir Global FSI Bus Dev, NVIDIA

Phil Filleul Segment Dir, Fin Svcs, Cray Inc

Felix Candelario Global Fin Svcs Sol Arch, Amazon Web Svcs (Invited)

Asif Alam Global Bus Dir Thomson Reuters

2017 Sponsors

DELL EMC · Hewlett Packard Enterprise · CRAY · redhat · SUSE · SUPERMICRO · Mellanox Technologies · NVIDIA · terasIC www.terasic.com · CAVIUM · LINUX NEW MEDIA The Pulse of Open Source · datanami · ENTERPRISETECH · HPCwire · LINUX JOURNAL · Integration developer news · information management · TABB GROUP · WSTA Wall Street Technology Association · AHA advanced hardware accelerators · BIOS IT · Chelsio Communications · Corvil · COMMVAULT · Mentor Graphics · CDW People Who Get It · Bright Computing · PENGUIN COMPUTING · DINI Group · InfoX · Myricom · RYFT · SOLARFLARE · SYMBOLIC IO · WD · SanDisk · Dolphin · FIBERBLAZE

Show Hours: Mon, Apr 3	8:00 - 4:00
Conference Hours:	8:30 - 4:50

Visit: www.flaggmgmt.com/linux

Show & Conference: Flagg Management Inc,
353 Lexington Avenue, New York 10016 (212) 286 0333 fax: (212) 286 0086
flaggmgmt@msn.com

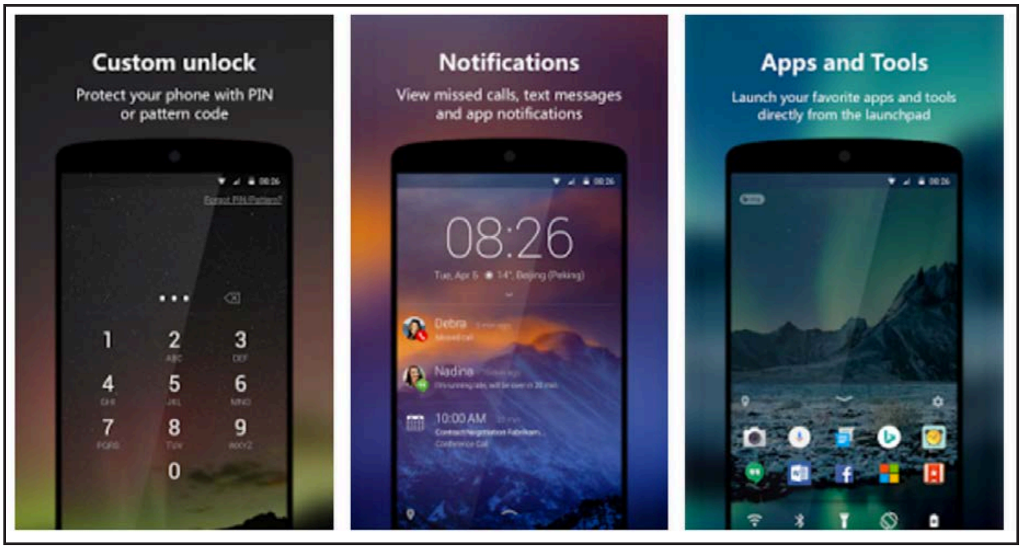
Natalia Vassilieva Sr Research Manager, Hewlett Packard Labs

Ed Turkel HPC Strategist, Dell EMC

Don Clegg VP Mktg Bus Dev Supermicro (Invited)

Pat McGinn BBA/IB CITP VP Prod Mktg CoolIT Systems

Android Candy: My World, in a Lock Screen


Custom unlock
Protect your phone with PIN or pattern code
Notifications
View missed calls, text messages and app notifications
08:26
Apps and Tools
Launch your favorite apps and tools directly from the launchpad

It feels weird to mention a Microsoft product in *Linux Journal*. But to be honest, there are some cool things coming out of the Microsoft Garage (http://www.microsoft.com/garage). One of those things is "Next Lock Screen", which is an Android app that brings interactive tools to the lock screen.

This concept isn't revolutionary, but with Next Lock Screen, it's done very well. It's possible to launch apps, interact with messages and get customized notifications all without unlocking your phone. Do you prefer to have your calendar events on your lock screen? Done. Want to control your music? Done. Again, nothing here is really new, it's just integrated and customizable in a way that takes a bunch of good ideas and repackages them into a slick lock screen. You also can get the Bing wallpaper on your lock screen, which is pretty cool, because honestly, the Bing photo of the day is almost always incredible.

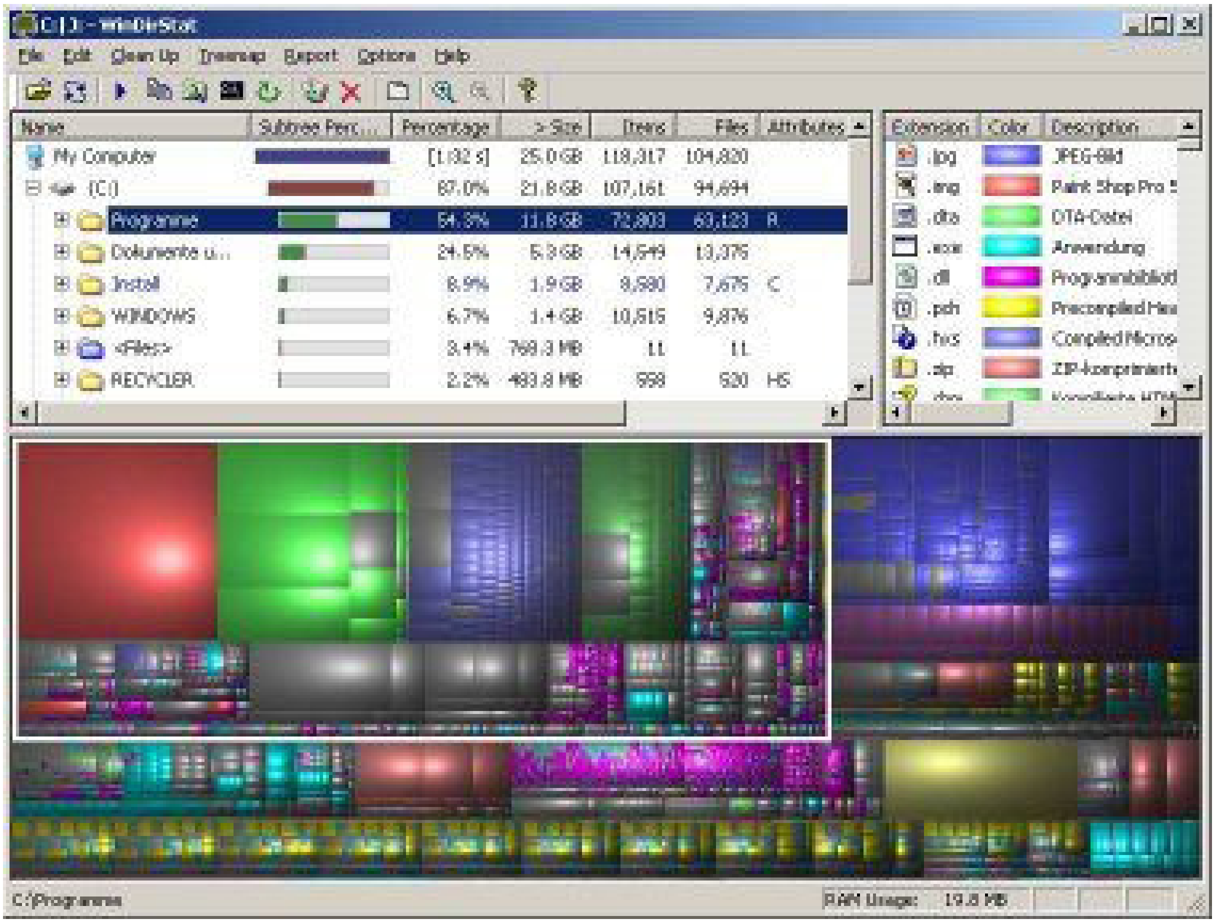
If you're not afraid to try an app developed by Microsoft, I urge you to check out Next Lock Screen. It makes a locked phone far more useful! (I should add that enabling interaction on your lock screen does make it far less secure, so be careful as to which features you enable.) Check it out at the Google Play Store.—Shawn Powers

Non-Linux FOSS: File Spelunking with WinDirStat

With Linux, it's fairy easy to find the large files on your system by doing something like this:

```
du -ahx / | sort -rh | head -20
```

Unfortunately, Windows users don't usually have equivalent tools.



Name	Subtree Perc...	Percentage	> Size	Items	Files	Attributes
My Computer		[1:32 s]	25.0 GB	118,317	104,820	
(C:)		87.0%	21.8 GB	107,161	94,694	
Programme		54.3%	11.8 GB	72,803	63,123	R
Dokumente u...		24.5%	5.3 GB	14,549	13,376	
Install		8.9%	1.9 GB	8,580	7,676	C
WINDOWS		6.7%	1.4 GB	10,515	9,876	
<Files>		3.4%	768.3 MB	11	11	
RECYCLER		2.2%	483.8 MB	598	520	HS

Extension	Color	Description
.jpg	🟦	JPEG-Bild
.img	🟥	Paint Shop Pro 5
.dta	🟩	DTA-Datei
.exe	🟦	Anwendung
.dll	🟪	Programmbibliot
.pch	🟨	Precompiled Hea
.hxs	🟦	Compiled Micros
.zip	🟥	ZIP-komprimierte
.chm	🟩	Kompilierte HTM

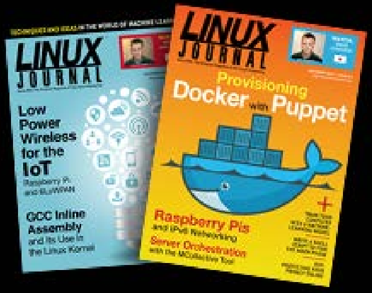
C:\Programme

RAM Usage: 19.8 MB

(Image via https://windirstat.net)

That's where something like WinDirStat comes into play. It's a file browser that uses incredible GUI elements to show you the files on your system with file size shown as rectangles. Big files are shown as big rectangles, and their file types are specified by color. It's a great visual way to sort your filesystem and get rid of (or at least find) extremely large files.

If you use Windows on a regular basis, but seem to have a shrinking hard drive, I urge you to download WinDirStat to get real-time statistics on your filesystem. It's open source and, of course, free to download at https://windirstat.net.—Shawn Powers



Is the Best Tea Ever Really Worth It?

I recently wrote about my "perfect tea"-steeping device. It's nothing more than a plastic mug-shaped device that has a sieve built in for straining tea leaves after the steeping process is complete. I use it every day. Really. And compared to the tech pieces I normally write in *Linux Journal*, that little UpFront blurb garnered me quite a few emails asking for details. It also got me a few messages explaining that brewing tea in a plastic cup was an unforgivable sin. One email, however, did make me think. I was asked about my Breville One Touch Tea Maker.

For Christmas (or a birthday?) a few years back, my incredible wife bought me a $250 tea maker. The One Touch treats making tea a bit like making coffee. You put the tea leaves into a basket, and the brewing process is automatic. The cool part is that it makes the water the perfect temperature and steeps the leaves for the exact right time depending on the type of tea you're making. And you know what? It makes the best tea I've ever had. Seriously. It makes tea noticeably better than any other method I've used. And yet, I rarely use it. Why?

It turns out that although the One Touch isn't exactly difficult to clean, it does take some effort. That's frustrating. The real gotcha, however, is that I don't drink the entire pot of tea quickly enough, and even though the One Touch keeps tea hot for an hour, I find myself heating cold tea in the microwave. It is possible to make a smaller batch of tea, but if you're going to go through the hassle of brewing a pot of tea, why wouldn't you make a full pot? So most of the time I use my One Touch pot to heat water, and then brew tea in the plastic steeper. It's crazy.

My Breville One Touch has been instrumental in my thinking when it comes to tech purchases, and I wish I'd learned its lesson sooner. My PlayStation 4 Pro? I actually find the PS3 is just as fun and a fraction of the price. My rackmount Xeon ESXi server? The few Raspberry Pi servers I have are actually more useful and flexible. Heck, even my fancy new F-150 pickup isn't half as fun as my 43-year-old Volkswagen Beetle.

So what did my $250 tea maker teach me? Marketing and popularity aren't what make things great. It's a lesson I should have learned years ago, because Linux is free, and yet it's the operating system that brings me the most joy! —Shawn Powers

THEY SAID IT

There is only one success—to be able to spend your life in your own way.

—Christopher Morley

When a thought is too weak to be expressed simply, it should be rejected.

—Marquis de Vauvenargues

Do something. If it doesn't work, do something else. No idea is too crazy.

—Jim Hightower

The most profound statements are often said in silence.

—Lynn Johnston

The most potent muse of all is our own inner child.

—Stephen Nachmanovitch

Jmol: Viewing Molecules with Java

Let's dig back into some chemistry software to see what kind of work you can do on your Linux machine. Specifically, let's look at Jmol, a Java application that is available as both a desktop application and a web-based applet (http://jmol.sourceforge.net).

You can use Jmol to help analyze the results you get from other software packages that actually calculate the chemical effects you are researching. It can read in dozens of different file formats, and you can use it to visualize everything from small molecules to huge macromolecules, like proteins. You also can visualize crystals and orbitals. You even can visualize animated events, such as chemical reactions and molecular vibrations.

Most Linux distributions should have Jmol available within their package management repositories. For example, you can install it on

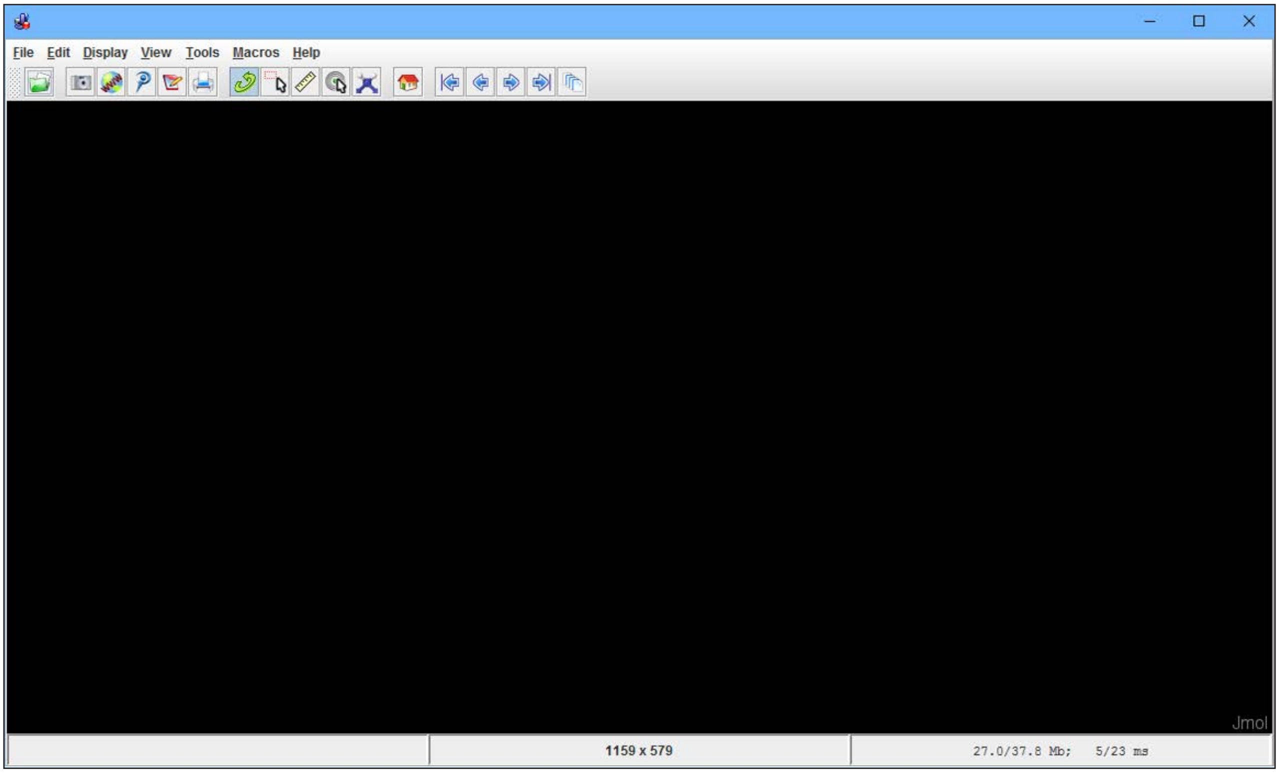


Figure 1. When you first start Jmol, you get a blank workspace ready for your work.

Debian-based distributions with this command:

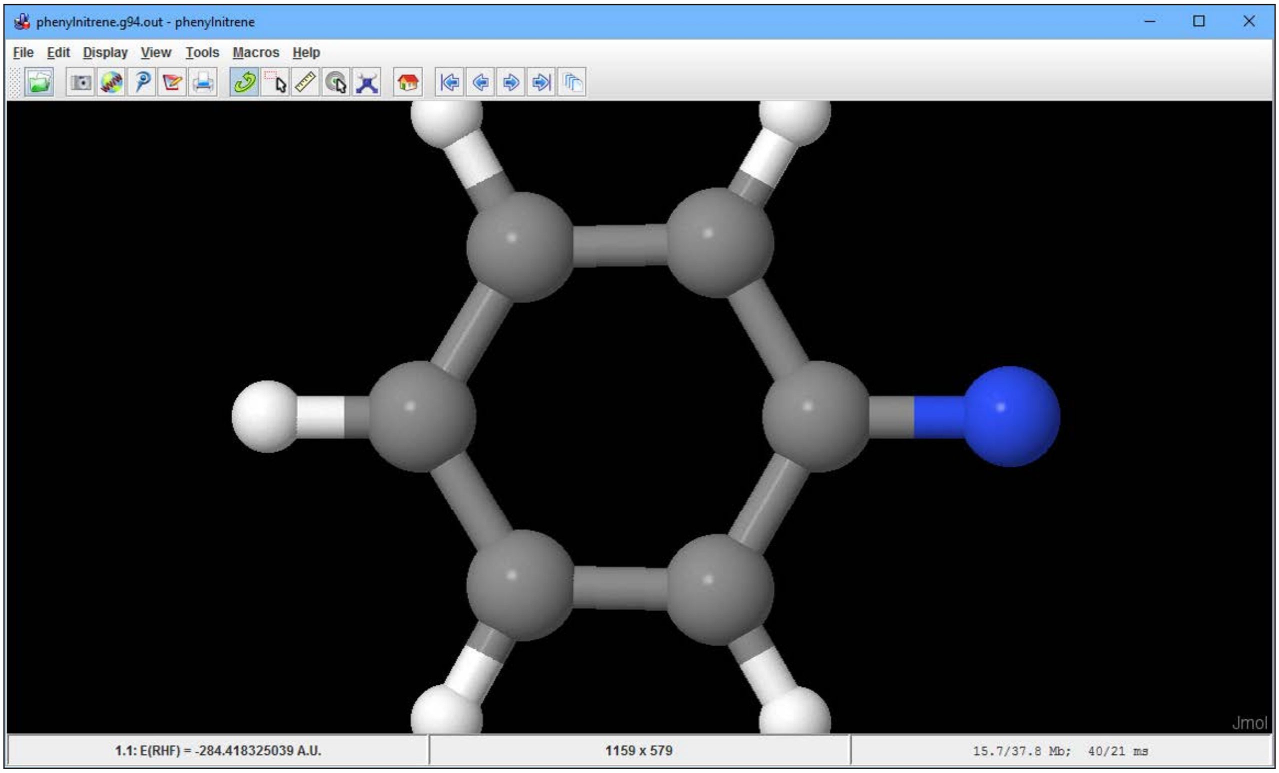
```
sudo apt-get install jmol
```

If you want to use the latest and greatest version, download it from the main project website. The download comes as a simple zip file containing everything you need to run Jmol. You also will need to install a Java virtual machine in order to run Jmol.

If you installed Jmol from the package manager, you probably will have a script available that will make running Jmol easier. If you install it from the binary zip file, you will need to run it manually by calling Java and using the JAR file as a command-line option.

When you first start Jmol, you'll see a blank screen, ready for input. Across the top is a series of icons allowing for easy access to the key functions available within Jmol. If you already have data files to analyze, you can use them. Otherwise, you may need some sample files in order to play with the functionality available.

The binary distribution doesn't include any sample files in order to save



1.1: E(RHF) = -284.418325039 A.U.
1159 x 579
15.7/37.8 Mb; 40/21 ms

Figure 2. The basic display you get when you load a molecule is a ball and stick display.

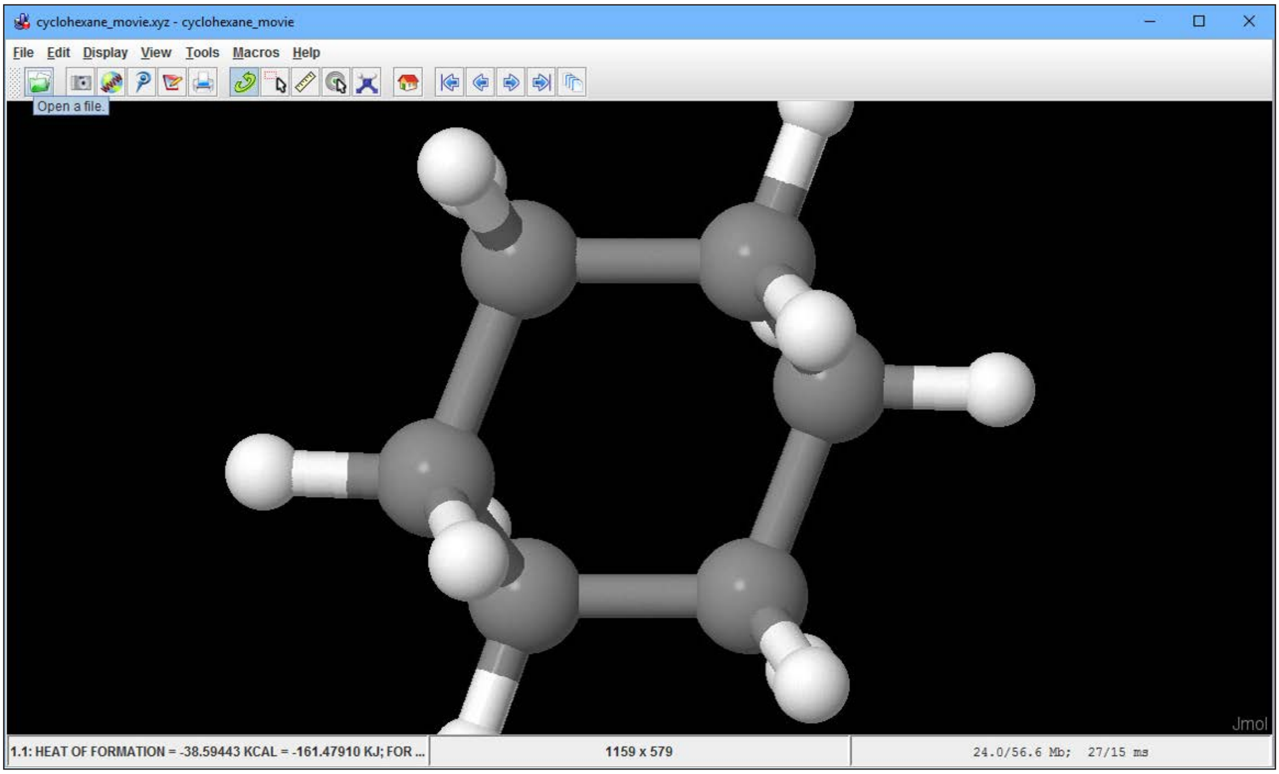
on download bandwidth; however, several sample data files are available from the main website. You can get the entire set by downloading a snapshot of the source files. In the examples for the rest of this article, I'm using several of the sample data files available from the source snapshot download.

The simplest example is just to load a data file and see what it looks like. Figure 2 shows what you get when you load the sample file Jmol-datafiles/gaussian/phenylnitrine.g94.out.

The data display is an interactive one. Using your mouse, you can click and drag the molecule to rotate it around to see all of the details.

The Display menu item provides a number of options to play with the molecule. The Atom menu item allows you to change how much of the van der Waals force field to show. The Bond menu item shows how thick to make the bonds between atoms. With these two options, you can tailor the display so that the appropriate amount of detail is shown. The Label menu item allows you to add either symbols, names or atomic numbers to the atoms within the molecule.

Near the bottom of the Display menu, there is a check box for whether



```
1.1: HEAT OF FORMATION = -38.59443 KCAL = -161.47910 KJ; FOR ...    1159 x 579    24.0/56.6 Mb;  27/15 ms
```

Figure 3. When you load an animation, it starts with a static image of your molecule.

hydrogen atoms are displayed within the display of the molecule.

While I'm talking about how to affect the display of the molecule, I should mention that the View menu item provides a number of presets on how to line up the molecule. So, with a single click, you can view the molecule along any of its axes.

Jmol also can display animations of events, along with static images. The animations subdirectory contains several examples that you can play with. When you load it up, you start with a static image of the molecule as before.

Within the icon bar at the top of the window, there are a series of buttons at the far right-hand side that allow you to step through the frames of the animation frame by frame. If you want to see the full animation, there's set of options under the Tools→Animate menu item. Here, you can go through the animation once, or you can put it on a loop.

You can even use a mode called Palindrome that goes forward through the animation and then backward. That way, you need to calculate only one half of the motion, yet you still are able to visualize the entire range of the motion.

Several more analysis tools are available. Clicking the Tools→Spectra→JSpecView menu item pops up a new window. Under the File menu item, you'll find options to add extra files, or do H1 or C13

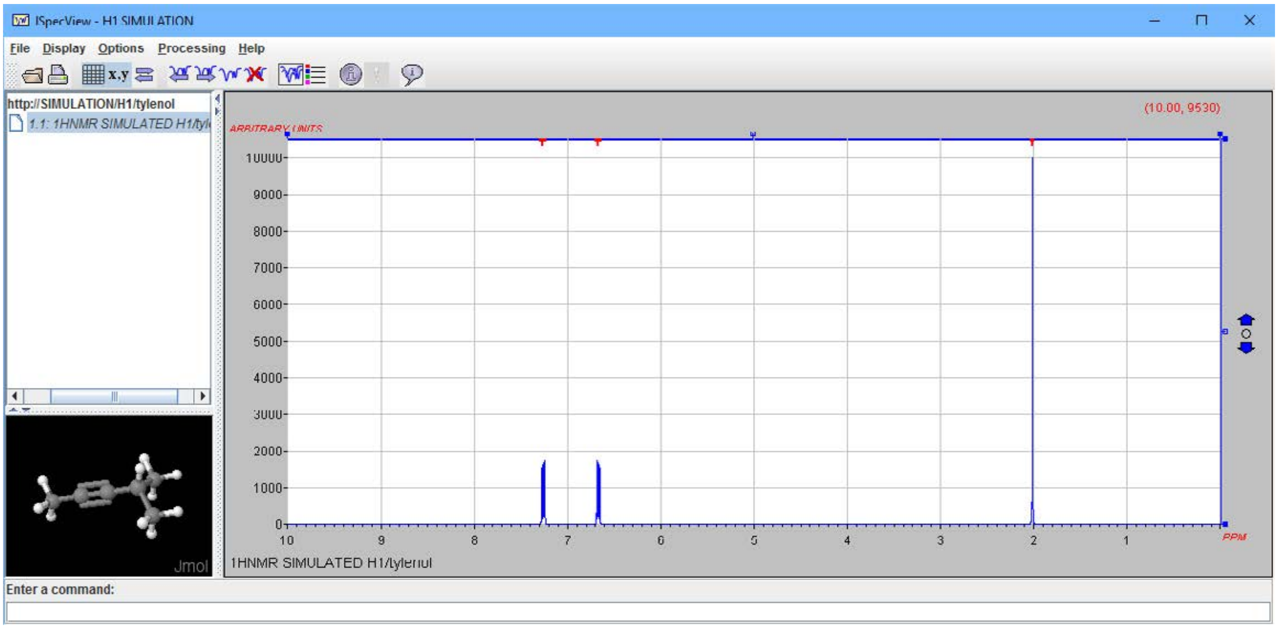


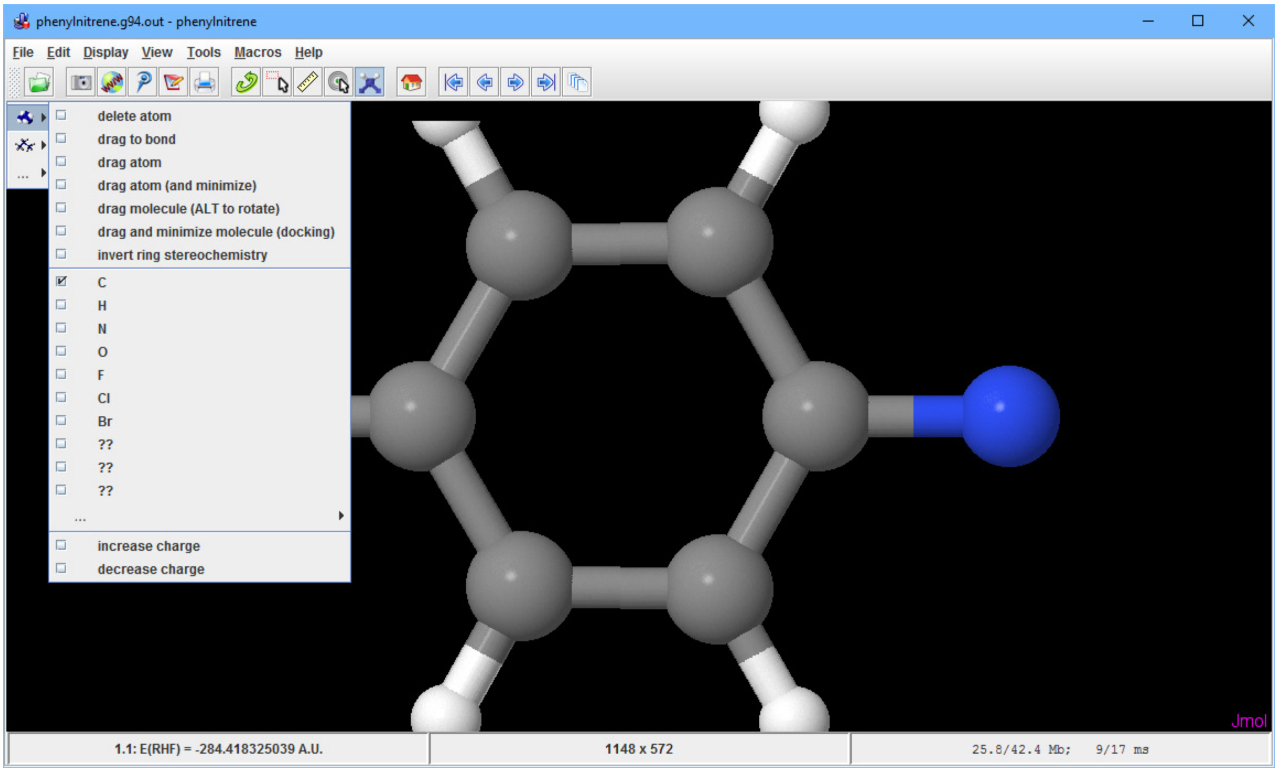
Figure 4. JSpecView is an extra tool available for looking at the spectra of molecules.

simulations. You can select Tools→Measurements to measure the distance between atoms within your molecule, and you can set the units used for those measurements with the Tools→Distance Units menu item. You actually can edit the molecule after it is loaded too.

If you click the icon button with the hover-over description "Open the model kit", you'll get a small set of drop-down items on the top left side of the display window. It allows you to delete atoms, move bonds around or even change the atom species at specific locations.

If you have some type of analysis that you need to repeat several times, Jmol supports the addition of macros. Macros are just simple text files that contain a set of Jmol instructions. If you save them in the ~/.jmol/macros directory, Jmol will pick them up and provide them within the Macros menu item.

The language for the macros is the same used for Jmol's scripting capabilities. This scripting language is based on RasMol, with some minor changes. There is a full language reference available at http://chemapps.stolaf.edu/jmol/docs.


phenylnitrene.g94.out - phenylnitrene
File
Edit
Display
View
Tools
Macros
Help
☐ delete atom
☐ drag to bond
☐ drag atom
☐ drag atom (and minimize)
☐ drag molecule (ALT to rotate)
☐ drag and minimize molecule (docking)
☐ invert ring stereochemistry
☑ C
☐ H
☐ N
☐ O
☐ F
☐ Cl
☐ Br
☐ ??
☐ ??
☐ ??
...
☐ increase charge
☐ decrease charge
Jmol

1.1: E(RHF) = -284.418325039 A.U.

1148 x 572

25.8/42.4 Mb; 9/17 ms

Figure 5. Jmol also lets you edit molecules.

You also can use scripts interactively by clicking the File→Script Editor menu item. This pops up a new window where you can write your script, check its syntax and then run it within Jmol. This provides a huge amount of power, allowing you to get the exact type of analysis you need.

Once you've done your analysis, several output options are available. The File→Export menu item gives you four options. You can select Export Image to save a static image in one of several image file formats.

Because Jmol also operates as a Java applet, you can select Export to Web Page to generate a web page that you then can use within your own website to share your research results.

If you want a higher-resolution image of a molecule, you can select Render in POV-Ray to use the POV-Ray external program to render a high quality 3D image.

The last export option is Write State, which saves the current workspace so that you can reload it later and continue your analysis. There also is an extra output option under Tools→Gaussian that pops up another window. Here you can set several options for a

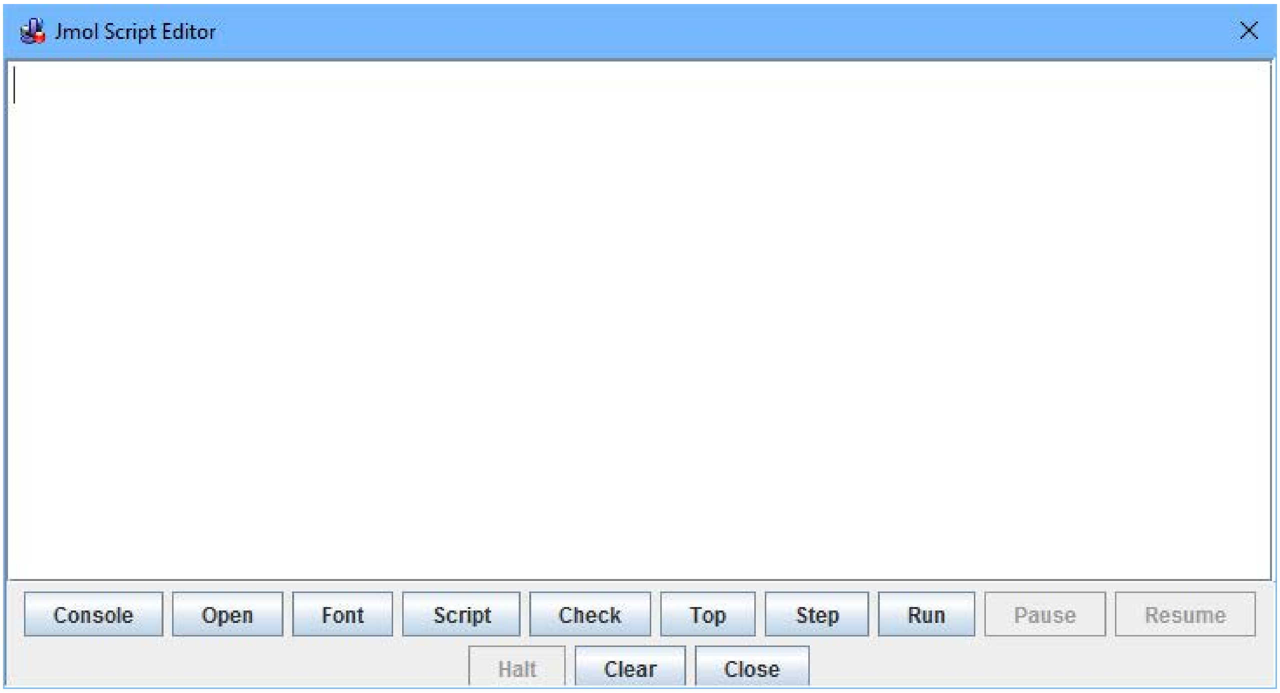
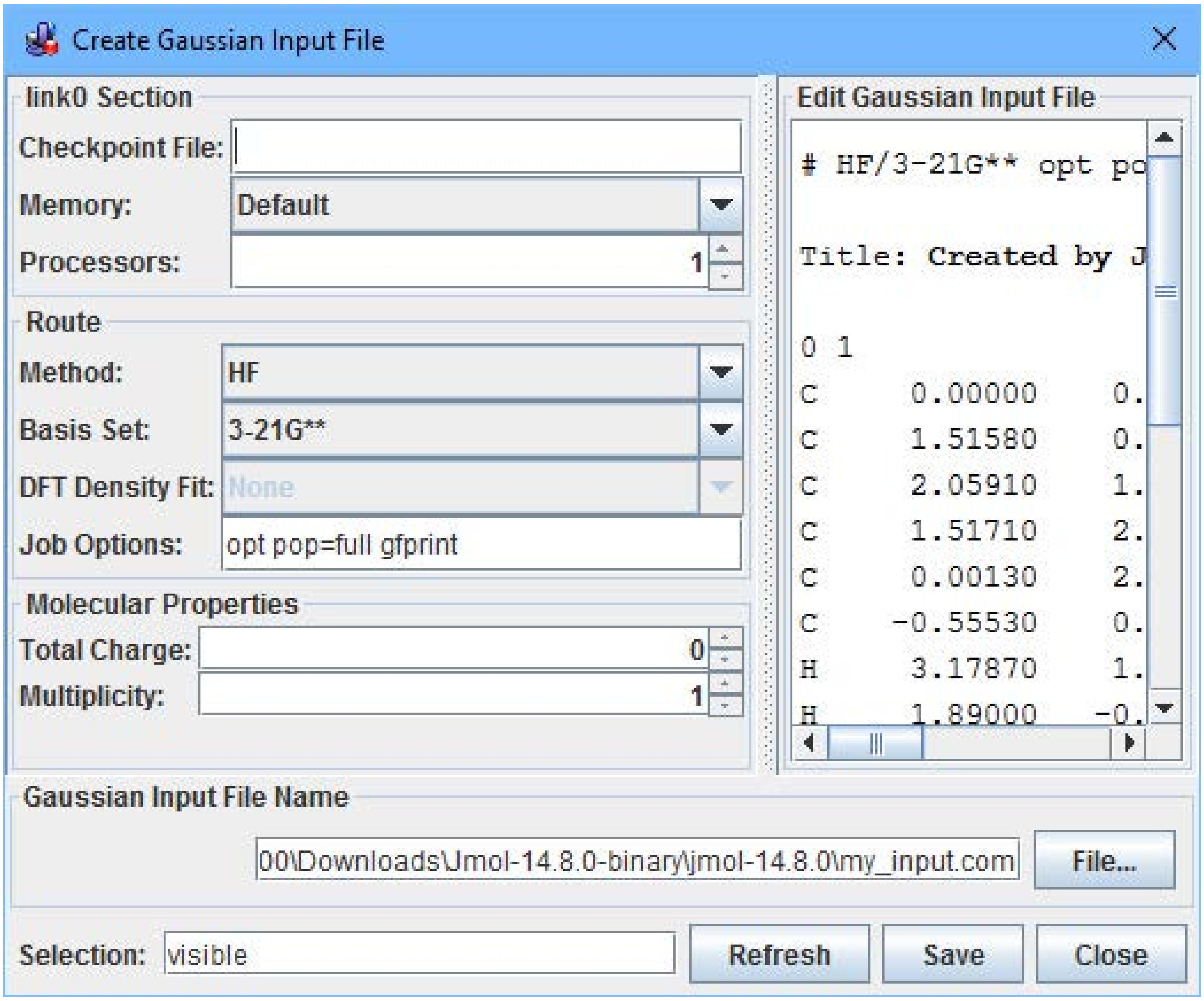


Figure 6. Jmol provides a full scripting language to help automate your analysis steps.



link0 Section

Checkpoint File:

Memory: Default

Processors: 1

Route

Method: HF

Basis Set: 3-21G**

DFT Density Fit: None

Job Options: opt pop=full gfprint

Molecular Properties

Total Charge: 0

Multiplicity: 1

```
# HF/3-21G** opt po

Title: Created by J

0 1
C      0.00000    0.
C      1.51580    0.
C      2.05910    1.
C      1.51710    2.
C      0.00130    2.
C     -0.55530    0.
H      3.17870    1.
H      1.89000   -0.
```

Gaussian Input File Name

00\Downloads\Jmol-14.8.0-binary\jmol-14.8.0\my_input.com

Selection: visible

Figure 7. You can use Jmol to generate Gaussian input files based on your analysis.

Gaussian input file that you can then use to run further simulations of your molecule.

With these tools, you easily can share your research results with others and build on the work you are doing.—Joey Bernard

RETURN TO CONTENTS

BALTIMORE
DRUPALCON 2017
APRIL 24-28

PREVIOUS
UpFront

NEXT
Reuven M. Lerner's
At the Forge

Read a Book in the Blink of an Eye!

LINUX JOURNAL EDITORS' CHOICE ★

I love reading. Sadly, the 24 hours I get per day seems to be inadequate for the tasks I need to accomplish. That might change as my teenagers turn into college kids and then begin to start families of their own. For now, however, between drama class and basketball practice, it seems like it takes about 30 hours to accomplish a 24-hour day. Needless to say,

I don't read as many books as I'd like.

Normally I take advantage of commute time to listen to audiobooks. That actually works quite well, and I'm able to read 30–40 books a year. Most of those books are fiction, but still, I'm grateful for audiobooks. Not long ago, I discovered a different sort of audiobook. I honestly have mixed feelings about the concept, but imagine if Cliff Notes and Audiobooks had a baby. That baby might be called "Blinkist".

Blinkist is a company that condenses books into very short summaries. They are available via text (for Kindle and the like), but for me, the professionally narrated audio versions are really what work best. Rather than reading an audiobook over the course of a week, I can "read" a book on the way to the grocery store. I'm shocked to admit, the summaries of entire books are surprisingly useful. For many books, the summary from Blinkist is enough. For some, the "blinks" make me want to read the entire book. That means although it's not a 100% replacement for reading, it adds value (and knowledge) to my life.

There is a three-day free trial that allows you to read as many books as you like. I urge you to give it a try. After the three days, you can either default to the free account, which allows you to listen (or read) one pre-chosen free book a day, or opt for a paid subscription. For text-only "blinks", it's $50/year. For unlimited text and audio "blinks", it's $80/year. Thankfully, three days is a enough time to figure out if it's something you find worth buying.

Thanks to its cool way of fitting more information into our over-busy lives and its handy mobile app for "blinking" on the go, Blinkist gets the Editors' Choice award this month. If nothing else, check out the free trial at http://www.blinkist.com. You can read a surprising number of books in three free days!—Shawn Powers

RETURN TO CONTENTS

Unsupervised Learning

In this article, Reuven moves from supervised learning to unsupervised learning, where you ask the computer to tell you something interesting about the data.

◀ PREVIOUS
Editors' Choice

NEXT
Dave Taylor's
Work the Shell ▶

REUVEN M. LERNER

Reuven M. Lerner, a longtime Web developer, offers training and consulting services in Python, Git, PostgreSQL and data science. He has written two programming ebooks (*Practice Makes Python* and *Practice Makes Regexp*) and publishes a free weekly newsletter for programmers, at http://lerner.co.il/newsletter. Reuven tweets at @reuvenmlerner and lives in Modi'in, Israel, with his wife and three children.

IN MY LAST FEW ARTICLES, I've looked into machine learning and how you can build a model that describes the world in some way. All of the examples I looked at were of "supervised learning", meaning that you loaded data that already had been categorized or classified in some way, and then created a model that "learned" the ways the inputs mapped to the outputs. With a good model, you then were able to predict the output for a new set of inputs.

Supervised learning is a very useful technique and is quite widespread. But, there is another set of techniques in machine learning known as *unsupervised learning*. These techniques, broadly speaking, ask the computer to find the hidden structure in the data—in other words, to "learn" what the meaning of the data is, what relationships it contains, which features are of importance, and which data records should be considered to be outliers or anomalies.

Unsupervised learning also can be used for what's known as "dimensionality reduction", in which the model functions as a preprocessing step, reducing the number of features in order to simplify the inputs that you'll hand to another model.

In other words, in supervised learning, you teach the computer about your data and hope that it understands the relationships and categorization well enough to categorize data it hasn't seen before successfully.

In unsupervised learning, by contrast, you're asking the computer to tell you something interesting about the data.

This month, I take an initial look at the world of unsupervised learning. Can a computer categorize data as well as a human? How can you use Python's scikit-learn to create such models?

Unsupervised Learning

There's a children's card game called *Set* that is a useful way to think about machine learning. Each card in the game contains a picture. The picture contains one, two or three shapes. There are several different shapes, and each shape has a color and a fill pattern. In the game, players are supposed to identify three-card groups of cards using any one of those properties. Thus, you could create a group based on the color green, in which all cards are green in color (but contain different numbers of shapes, shapes and fill patterns). You could create a group based on the number of shapes, in which every card has two shapes, but those shapes can be of any color, any shape and any fill pattern.

The idea behind the game is that players can create a variety of different groups and should take advantage of this in order to win the game.

I often think of unsupervised learning as asking the computer to play a game of *Set*. You give the computer a data set and ask it to divide that large bunch of data into separate categories. The model may choose any feature, or set of features, and that might (or might not) be a feature that humans would consider to be important. But, it will find those connections, or at least try to do so.

One of the most common machine-learning models for beginners is the "iris" dataset, containing 150 flowers, 50 from each of three types of irises. Several months ago, I showed how you could create a supervised model to identify irises. In other words, you could create and train a model that would categorize irises accurately based on their petal and sepal sizes.

Can unsupervised learning achieve the same goal? That is, can you create a model that will divide the flowers into three different groups, doing the same job (or close to it) that humans have done?

Another way of asking this question is whether the way in which biologists distinguish between varieties of flowers is supported by the underlying measurement data.

Let's load the iris data and then start to create an unsupervised model. Assuming that I'm working within the Jupyter notebook, I can execute the following:

```
%pylab inline
import pandas as pd
from pandas import DataFrame, Series

from sklearn.datasets import load_iris
iris = load_iris()

df = DataFrame(iris.data, columns=iris.feature_names)
df['response'] = iris.target
```

In other words, I've created a Pandas data frame containing five columns—the four features and also the response (that is, the classification). You won't be passing the classification to the model (although that might improve the model's ability to classify the flowers), but it's convenient to keep everything together in this way.

Creating a Model

Once you've loaded the data, it's time to create a model. You're looking to do what's known as "clustering", which means that the computer will divide the data set into categories or clusters.

So, now what? In supervised learning, you would create a new model from a classifier and then train it using scikit-learn's "fit" method. You then could give the trained model one or more data points and ask it to categorize those based on the model.

In unsupervised learning, it's a bit trickier—after all, you're asking the computer to do the categorization. If you don't have any pre-labeled categories, it's going to be hard to know whether your model is useful, accurate or both.

But before getting into the evaluation, let's build a model. Sklearn comes with a number of classifiers that handle clustering. One popular classifier is known as "K-means". In K-means clustering, the idea is that the model puts each data point inside the cluster whose mean is the closest. Thus, if there are three clusters, each cluster will contain points that are calculated to be closest. The "inertia" is a measurement of how coherent the groups are—that is, how closely associated with one another the elements that have been grouped together fit.

I should note that because K-means uses distances to calculate how to compose a group, you likely will want all of your features to be on the same scale. In the case of the flowers, all are within the same order of magnitude. But, you can imagine that if three measurements are on a scale of 1–10 and a fourth is on a scale of 1–1 million, the calculations might not work out as well. For this reason, it can be a good idea to use a scaler—several of which come with sklearn—to put all of your data onto the same scale. Such scaling is often important when creating models; it helps the calculations to identify two or more items as being close by.

So, using Python's scikit-learn, you can say:

```
from sklearn.cluster import KMeans
k = KMeans(n_clusters=3)
```

The above code indicates that you're going to use the K-means algorithm. You create a new model, indicating when you do so that you want three groups.

Now, right away you might be asking yourself how to know that there will be three categories—and the cop-out answer is that you guess. You can try different values for `n_clusters` and evaluate the model to see how well it does. But in many cases, you'll have to experiment a bit.

Let's now run K-means on the data. The X (that is, input matrix) is going to be the data frame, minus the "response" column. You can create that as follows:

```
X = df.drop('response', axis=1)
```

With supervised learning, the "fit" method is the process in which you teach the model to make associations between the input matrix X and the output vector y. In unsupervised learning, you're asking the model itself to make such divisions and to create an output vector. You do this with "fit":

```
k.fit(X)
```

Evaluating the Model

The first question you'll ask the model is: "How did it divide up the flowers?" You know that the irises should be divided into three different groups, each with 50 flowers. How did K-means do?

You can ask the model itself using a variety of attributes. These attributes often end with an underscore (_), indicating that they may continue to change over time, as the model is trained more.

And indeed, this is an important point to make. When you invoke the "fit" method, you are teaching the model from scratch. However, there are times when you have so much data, you cannot reasonably teach the model all at once. For such cases, you might want to try an algorithm that supports the "partial_fit" method, which allows you to grab inputs a little bit at a time, teaching the model iteratively. However, not all algorithms support partial_fit; a large number of data points might force your hand and reduce the number of algorithms from which you can choose.

For this example, and in the case of K-means, you cannot teach the model incrementally. Let's ask the model for its measure of inertia:

```
k.inertia_
```

(Again, notice the trailing underscore.) The value that I get is 0.78.9. The inertia value isn't on a scale; the general sense is that the lower the inertia score, the better, with zero being the best.

What if I were to divide the flowers into only two groups, or four groups? Using scikit-learn, I can do that pretty quickly and determine whether the computer thinks the manual classification (into three groups) was a good choice:

```
output = [ ]
for i in range(2,20):
    model = KMeans(n_clusters=i)
    model.fit(X)
    output.append((i, model.inertia_))
kmeans = DataFrame(output, columns=['i', 'inertia'])
```

Now, it might seem ridiculous to group 150 flowers into up to 19 different groups! And indeed, the lowest inertia value that I get is when I set `n_clusters=19`, with the inertia rising as the number of groups goes down.

Perhaps this means that every flower is unique and cannot be categorized? Perhaps. But it seems more likely that our data isn't appropriate for K-means. Maybe it's the wrong shape. Maybe its values aren't varied enough. And indeed, when you look at the way in which the flowers were clustered for `n_clusters=3`, you see that the clustering was quite different from what people came up with. I can turn the automatically labeled flowers into a Pandas Series, and then count how many of each flower was found:

```
Series(k.labels_).value_counts()
```

I get:

```
2    62
1    50
0    38
```

Well, it could be worse—but it also could be much better. Perhaps you can and should try another algorithm and see if it's better able to group the flowers together.

I should note that this now falls under the category of "semi-supervised learning"—that is, trying to see whether an unsupervised technique can achieve the same results, or at least similar results, to a previously used supervised technique.

In such a case, you can evaluate your model using not just statistical tests, but also one of the techniques I described in my previous articles on supervised learning, namely train-test-split. You use unsupervised learning on a portion of the input data and then predict on the remaining part. Comparing the model's outputs with the expected outputs for that subset can help you evaluate and tune your model.

A Different Algorithm

But in this case, let's try using a different model to achieve a different result, simply to see how easily sklearn allows you to try different models. One common choice in unsupervised learning is Gaussian Mixture, known in previous versions of scikit-learn as GMM. Let's use it:

```
from sklearn.mixture import GaussianMixture
model = GaussianMixture(n_components=3)
model.fit(X)
```

Now, let's have the model predict with the data used to train it, which will return a NumPy array with the categories:

```
model.predict(X)
```

How did that do? Let's pop this data into a Pandas Series object and then count the values:

```
Series(model.predict(X)).value_counts()
```

And sure enough, the results:

```
2    55
1    50
0    45
```

This is still imperfect—assuming that the human classification counts as "perfect", but it's clearly better than the attempts with K-means. And because this is semi-supervised learning here, in which you have some of the original scores, you can use some of sklearn's metrics to find how good (or bad) the model is:

```
from sklearn import metrics
labels_true = iris.target
labels_pred = model.predict(X)
```

Let's find out how well it did:

```
metrics.homogeneity_score(labels_true, labels_pred)
0.89832636726027748

metrics.completeness_score(labels_true, labels_pred)
0.90106489086402064
```

Hey, pretty good! Not perfect (that is, 1.0), but not bad at all. And if you compare this against the K-means model:

```
labels_pred = k.labels_
metrics.homogeneity_score(labels_true, labels_pred)
0.75148540219883375

metrics.completeness_score(labels_true, labels_pred)
0.76498615144898152
```

In other words, my intuition was right. The GaussianMixture model was better at clustering the flowers than the K-means model.

Conclusion

In many ways, unsupervised learning is the true magic and potential in the machine-learning world. By using computers to identify patterns and groups in your data, more quickly and accurately than you could do yourself, you can start to identify and predict all sorts of things. As with supervised learning though, unsupervised learning requires that you try a variety of models, compare them against one another and understand that each model has its own advantages, disadvantages and biases.

The world of data science in general, and machine learning in particular, continues to grow at an extremely rapid rate, with new ideas, techniques and tutorials available all of the time. The Resources section here describes several places where you can learn more and start your journey in this set of concepts and technologies.■

RESOURCES

I used Python (http://python.org) and the many parts of the SciPy stack (NumPy, SciPy, Pandas, Matplotlib and scikit-learn) in this article. All are available from PyPI (http://PyPI.python.org) or from https://www.scipy.org.

I recommend a number of resources for people interested in data science and machine learning.

One long-standing weekly email list is "KDNuggets" at http://www.kdnuggets.com. You also should consider the "Data Science Weekly" newsletter (https://www.datascienceweekly.org) and "This Week in Data" (https://datarepublicblog.com/category/this-week-in-data), describing the latest data sets available to the public.

I am a big fan of podcasts, and I particularly love "Partially Derivative". Other good ones are "Data Stories" and "Linear Digressions". I listen to all three on a regular basis and learn from them all.

If you're looking to get into data science and machine learning, I recommend Kevin Markham's Data School (http://dataschool.org) and Jason Brownlie's "Machine Learning Mastery" (http://machinelearningmastery.com), where he sells a number of short and dense, but high-quality ebooks on these subjects.

Send comments or feedback via
http://www.linuxjournal.com/contact
or to ljeditor@linuxjournal.com.

RETURN TO CONTENTS

The **Free Software Foundation**, GNU Project and
MIT's SIPB invite you to

LIBRE PLANET 2017

"Roots of Freedom"

A conference for everyone who loves free software

- + Hackers
- + Activists
- + Beginners
- + Users
- + Writers
- + Artists

March 25 & 26 at MIT. Students and FSF members attend gratis.

libreplanet.org/conference

Image Manipulation with ImageMagick

DAVE TAYLOR

Dave Taylor has been hacking shell scripts on UNIX and Linux systems for a really long time. He's the author of *Learning Unix for Mac OS X* and *Wicked Cool Shell Scripts*. You can find him on Twitter as @DaveTaylor, or reach him through his tech Q&A site: http://www.AskDaveTaylor.com.

Dave switches gears this month and begins delving into the more functional topic of image manipulation.

◀ PREVIOUS
Reuven M. Lerner's
At the Forge

NEXT ▶
Kyle Rankin's
Hack and /

IN MY LAST ARTICLE, I had some fun looking at the children's game of rock, paper, scissors, writing a simple simulator and finding out that some strategies are better than others. Yes, I used "strategy" and "rock, paper, scissors" in the same sentence!

So for this article, I thought it would be interesting to delve into something more functional and pragmatic: image manipulation. Ordinary shell scripts don't tend to do much with images because you can't display anything from the command line.

But let's be honest here. The chance that you're running Linux or a similar command-line interface raw on a computer terminal is pretty darn low. More likely, you've got a terminal window open on your X11 system or, like I often have, you're running a command-line interface app within a modern OS like Mac OS X. And this means, yes, you do have the ability to display graphics, just not within the terminal app itself.

Get Yourself a Copy of ImageMagick

The first step is to download and install a copy of the ImageMagick suite of graphics-related commands. You already might have it installed if you're lucky: Just type `convert -version`, and if you have it installed, you'll see something similar to this:

```
$ convert -version
Version: ImageMagick 6.9.6-6 Q16 x86_64 2016-12-31
 ➥http://www.imagemagick.org
Copyright: Copyright (C) 1999-2016 ImageMagick Studio LLC
License: http://www.imagemagick.org/script/license.php
Features: Cipher DPC Modules
Delegates (built-in): bzlib djvu fftw fontconfig freetype gslib
 ➥jbig jng jp2 jpeg lcms ltdl lzma openexr png ps tiff
 ➥webp x xml zlib
```

If you don't have it installed, it can be quite a task to get it all up and running. Everything lives at http://www.imagemagick.org, which is where you want to get started.

On a Linux system, you can use the package manager of choice for your distro. You can grab a compressed tar image from the site, or you can use `rpm`, like this:

```
rpm -Uvh ImageMagick-7.0.4-1.x86_64.rpm
```

Of course, there's a bit more to it, but that'll get you started.

On a Mac, you'll want to start by installing MacPorts (http://www.macports.org), which you can't do until you install Xcode (free from Apple, get it through the App Store). Once you've installed Xcode and MacPorts, you can install ImageMagick, and you're good to go.

You know you're good to go when the test command `convert -version` returns something meaningful. As always, when you install new software, you'll want to log out and log in again for the PATH changes and shell command-line hash to include all the newest programs.

Converting Image Formats

One of the most useful tasks ImageMagick can help you with is converting image file formats. It's a remarkably well built suite of programs and can read or write more than 200 different formats. Don't believe me? Try this command:

```
convert -list format | more
```

Among the most common formats that you'll actually encounter in your day-to-day computer usage are the following:

- BMP: MS Windows bitmapped image.
- GIF: Graphics Interchange Format.
- JPG: JPEG image format.
- PNG: Progressive Network Graphic format.
- TIFF: Tagged Image File Format.

ImageMagick knows oodles of other formats too, including all the major video formats (MKV, MP4, AVI, MOV). It also can convert things like EPSF (Encapsulated Postscript) and even PDF (Portable Document Format), which can be useful in specific instances.

Armed with that knowledge, conversion between image file formats is really ridiculously simple. Let's say you want to convert an image from

JPEG to PNG. It's as simple as:

```
convert image.jpeg image.png
```

Since the ImageMagick utilities are glob-aware (that is, you can use wild cards and specify multiple filenames), you also can convert a group of GIF images to JPG with the `convert` command or, more easily, with its cousin `mogrify`:

```
mogrify -format jpg *.gif
```

Let's give it a whirl with a folder that contains a half-dozen GIF images, using `ls` to show the folder contents before and after the mogrification (is that a word?):

```
$ ls -s
total 272
  8 add-to-google-reader.gif  24 blogger-1.gif
  8 dave.gif                    8 add-to-newsgator.gif
 24 blogger-2.gif            176 manga.gif
 16 aw-logo.gif                8 blogger-3.gif
$ mogrify -format jpg *gif
$ ls -s
total 752
  8 add-to-google-reader.gif  24 blogger-1.gif
  8 dave.gif                    8 add-to-google-reader.jpg
112 blogger-1.jpg              8 dave.jpg
  8 add-to-newsgator.gif           24 blogger-2.gif
176 manga.gif                   8 add-to-newsgator.jpg
128 blogger-2.jpg            168 manga.jpg
 16 aw-logo.gif                8 blogger-3.gif
 24 aw-logo.jpg               24 blogger-3.jpg
```

Simple enough. Use `convert` for individual images and `mogrify` for bulk conversions. It'd be an easy script to differentiate between these two cases and invoke the correct command with the correct arguments too. I'll leave that up to you!

Checking Image Sizes

Another useful feature of the ImageMagick suite is to be able to identify the dimensions of a graphic image. The latest version of the `file` command can offer this information on some systems:

```
$ file manga*
manga.gif: GIF image data, version 89a, 358 x 313
manga.jpg: JPEG image data, JFIF standard 1.01,
 ➥aspect ratio, density 1x1, segment length 16,
 ➥baseline, precision 8, 358x313, frames 3
manga.png: PNG image data, 358 x 313, 8-bit/color RGB,
 ➥non-interlaced
```

But on most Linux systems, one or more of these would exclude the actual dimensions. Further, look closely at the above output, and you'll see it's quite inconsistent, making it difficult to parse out the dimensions if you don't encode specific rules for each format—which is, uh, lame.

Instead, you can glean image size with the `identify` command, as shown here:

```
manga.gif GIF 358x313 358x313+0+0 8-bit sRGB 256c 88.5KB
 ➥0.000u 0:00.000
manga.jpg JPEG 358x313 358x313+0+0 8-bit sRGB 85.4KB 0.000u
 ➥0:00.000
manga.png PNG 358x313 358x313+0+0 8-bit sRGB 266KB 0.000u
 ➥0:00.000
```

That's better. It's consistently the third parameter, which means that a simple script can strip out everything but the image dimensions:

```
$ for image in manga*; do   identify $image | cut -f1,3 -d\  ; done
manga.gif 358x313
manga.jpg 358x313
manga.png 358x313
```

Easy enough, and notice that the `cut` command is invoked both with a space as the default field delimiter and specifying that you want field 1 and 3 but none of the others.

And Next Month...

Okay, ImageMagick is complicated. In fact, I didn't really get much into scripting this month. But, come back next month; I'll explain how to turn all this knowledge of `convert`, `mogrify` and `identify` into some pretty sick scripts. See you then!■

Send comments or feedback via http://www.linuxjournal.com/contact or to ljeditor@linuxjournal.com.

RETURN TO CONTENTS

Sysadmin 101: Automation

Approach automation in the right way, and you might find you've automated yourself out of a job.

KYLE RANKIN

Kyle Rankin is a Sr. Systems Administrator in the San Francisco Bay Area and the author of a number of books, including *The Official Ubuntu Server Book*, *Knoppix Hacks* and *Ubuntu Hacks*. He is currently the president of the North Bay Linux Users' Group.

PREVIOUS
Dave Taylor's
Work the Shell

NEXT
Shawn Powers'
The Open-Source
Classroom

THIS IS THE SECOND IN A SERIES OF ARTICLES ON SYSTEMS ADMINISTRATOR FUNDAMENTALS. These days, DevOps has made even the job title "systems administrator" seem a bit archaic, much like the "systems analyst" title it replaced. These DevOps positions are rather different from sysadmin jobs in the past. They have a much larger emphasis on software development far beyond basic shell scripting, and as a result, they often are filled by people with software development backgrounds without much prior sysadmin experience. In the past, a sysadmin would enter the role at a junior level and be mentored by a senior sysadmin on the team, but in many cases currently, companies go quite a while with cloud outsourcing before their first DevOps hire. As a result, the DevOps engineer might be thrust into the role at a junior level with no mentor around apart from search engines and Stack Overflow posts.

In this series, I'm going to expound on some of the lessons I've learned through the years that might be obvious to longtime sysadmins but may be news to someone just coming into this position.

In the first article in this series, I talked about how to approach alerting and on-call rotations as a sysadmin. In this article, I discuss how to automate yourself out of your job. There is a quote that you see from time to time in sysadmin circles that goes something along the lines of "Be careful or I will replace you with a tiny shell script." Good system administrators hate performing mundane tasks and constantly seek to apply that saying to themselves. That said, there are many different approaches to automation, and not all of them result in a time-savings. Here, I discuss my experience with automation and describe what, when, why and how you should (and shouldn't) automate.

Why You Should Automate

There are a number of different reasons why you should take steps to automate your work as a sysadmin:

1) It frees up time spent doing mundane tasks to focus on more important work. With all of the automation that's already built in to servers these days, it's easy to take for granted just how many mundane tasks sysadmin have had to perform in the past. Logs weren't always rotated automatically; backups usually were home-grown affairs that often were triggered manually. Even now, there still are system administrators who install every single server by hand, log in to a machine manually and install or update software, and configure server configuration files on the host by hand.

Let's take server OS installation as an example—a modern interactive server OS installation may take anywhere from 15 minutes to an hour of sysadmin time to walk through and answer questions. These are the kinds of actions that don't really require a sysadmin's expertise once you've made the initial decisions about how you want a server to be set up. By automating these mundane tasks, you can get back to the more difficult work that does require your expertise.

2) Automation reduces mistakes in routine tasks. The thing about performing the same task over and over by hand is that it is easy to make mistakes, and if it's something you do every day, eventually you even may stop paying attention to whether your task succeeded. Also, the way that you may perform a certain task might be a little bit different from how a different administrator on the team does it. By automating a task, the team can agree on the ideal way to perform it and know that when you run your automation script, it is performed the same way every single time with no skipped steps or commands run in the wrong order.

3) Automation allows everyone on the team to be productive. With automation, you can take even a complex process and reduce it down to a command. That command then becomes something that anyone on the team can run, whereas the complex process may have required more senior members of the team. For instance, if you take production software deployment as an example, often there can be a complex arrangement of triggering load balancer and monitoring maintenance modes, software versions to check, mirrors to sync up, and services to restart and test. Even though these individual steps may be mundane, combined, they become pretty complicated and could overwhelm a junior member of the team—especially when production uptime hangs in the balance. By automating that process, senior administrators can put all of their expertise into creating the right process that performs the right checks, and they can go on vacation knowing that anyone else on the team now can perform the task the right way.

4) Automation reduces documentation workload. Often instead of automating a task, a sysadmin team will spend time documenting a process. There is still an important place for documentation, and in the next section, I discuss when that makes sense and when it doesn't. The fact is though, if you take take an entire process and put it into a single automated task, you no longer need a full wiki page of documentation (that inevitably will become out of date), because you've reduced it down to "run this command". Because the process is now automated, you also know the process is kept up to date; otherwise, the script wouldn't work.

What You Should Automate

Not everything is appropriate for automation, and even things that may be good candidates for automation may not be good candidates today (the next section covers when you should automate). Following are a few different types of tasks that make good candidates for automation.

1) Routine tasks. In general, tasks that you perform frequently (at least monthly) are good candidates for automation. The more frequent the task, in theory, the more time-savings you would get from automating it. Tasks that you perform only once a year may not be worth the effort to build automation around, and instead, those are the kinds of tasks that benefit from good documentation.

2) Repeatable tasks. If you could document a process as a series of commands, and then copy and paste them one by one in a terminal and the task would be complete, that's a repeatable task that may be a good candidate for automation. On the other hand, one-off tasks that have custom inputs or are something you may never have to do again aren't worth the time and effort to automate.

3) Complex tasks. The more complex a task, the more opportunities you have for mistakes if you do it manually. If a task has multiple steps, in particular steps that require you to take the output from one step and use it as input for another, or steps that use commands with a complex string of arguments are all great candidates for automation.

4) Time-consuming tasks. The longer the tasks take to complete (especially if there are periods of running a command, waiting for it to complete, and then doing something with that command's output), the better a candidate it is for automation. OS installation and configuration is a great example of this, as when you install an OS, there are periods when you enter installation settings and periods when you wait for the installation to complete. All of that waiting is wasted time. By automating long-running tasks, you can go do some other work and come back to the automation (or better, have it alert you) to see if it is complete.

When You Should Automate

My coworkers know that I enjoy automating myself out of my job, and sometimes in the past they have been surprised to learn that I haven't automated a task that by all measures is a prime candidate for automation. My answer is usually "Oh I plan to, I'm just not ready yet." The fact is that even if you have a task that is a great candidate for automation, it may not necessarily be the right time to automate it.

When I need to perform a new task that's a series of mundane, manual steps, I like to force myself to perform it step by step at least a few times "in the wild" before I start automating it. I find I usually need to perform a task a few times to understand where automation makes the most sense, what areas of the task may require extra attention, and what sorts of variables I might encounter for the task. Otherwise, if I just charge ahead and write a script, I may find yourself rewriting it from scratch a few weeks later because I discover the process needs to be adapted to a new variation of the task. If I'm not quite sure about parts of a process, I may automate only the parts I am sure of first and get those right. Later on when the rest of the process starts to gel in my mind, I then go back and incorporate it into the automation I've already completed.

I also avoid automating tasks if I'm not sure I can do so securely. For instance, a number of organizations are big fans of using ChatOps (automating tasks using bots inside a chatroom) for automation. Although I know that many bots can authenticate tasks before they perform them, I still worry about the potential for abuse with a service that's usually shared across the whole company, not to mention the fact that production changes are being triggered by a host outside the production environment. With my current threat model, I have to maintain strict separation between development and production environments, so having a bot accessible to anyone in the company, or having a Jenkins continuous integration server in the development environment performing my production tasks, just doesn't work. In many cases, I have fully automated tasks up to the point that it still requires an administrator with the proper access to go to the production environment (thereby proving that they are authorized to be there) before they push "the button".

How You Should Automate

Since the whole goal of automation is to save time, I don't like to waste time refactoring my automation. If I don't feel like I understand a process and its variables well enough to automate it, I wait until I do or automate only the parts I feel good about. In general, I'm a big fan of building a foundation of finished work that I then build upon. I like to start with automating tasks that will give me the biggest time-savings or encourage the most consistency and then build off them.

I like doing the hard work up front so that it's easier down the road, and that is why I am a big fan of configuration management to automate server configuration. Once something like that is in place, rolling out changes to configuration becomes trivial, and creating new servers that match existing ones should be easy. These big tasks may take time up front, but they provide huge cost savings from then on, so I try to automate them first.

I also favor automation tasks that can be used in multiple ways down the road. For instance, I think all administrators these days should have a simple, automated way to query their environment for whether a package is installed and on what hosts, and then be able to update that package easily on the hosts that have it. Some administrators refer to this as part of orchestration, a subject I covered a few months back in a series on MCollective.

Package updates are something that sysadmins do constantly both for in-house software that changes frequently and system software that needs security updates. If a security update is a burden, many sysadmin won't bother. Having automation in place to make package updates easy means administrators save time on a task they have to perform frequently. Sysadmins then can use that automated package update process both for security patches, in-house software deployments and other tasks where package updates are just one component of many.

As you write your automation, be careful to check that your tasks succeeded, and if not, alert the sysadmin to the problem. That means shell scripts should check for exit codes, and error logs should be forwarded somewhere that gets the administrator's attention. It's all too easy to automate something and forget about it, but then check back weeks later and discover it stopped working!

In general, approach automation as a way to free up your brain, time and expertise toward tasks that actually need them. For me, I find that means time spent improving automation and otherwise dealing with exceptions—things that fall outside the normal day. If you keep it up, you eventually will find that when there are no crises or new projects, the day-to-day work should be automated to the point that your task is just to keep an eye on your well-oiled machine to make sure everything's running. That is when you know you have replaced yourself with a shell script.■

Send comments or feedback via http://www.linuxjournal.com/contact or to ljeditor@linuxjournal.com.

RETURN TO CONTENTS

The Post-TV Age?

I have lots of streaming packages, but I just can't seem to cut the cord!

PREVIOUS
Kyle Rankin's
Hack and /

NEXT
New Products

SHAWN
POWERS

Shawn Powers is the Associate Editor for *Linux Journal*. He's also the Gadget Guy for LinuxJournal.com, and he has an interesting collection of vintage Garfield coffee mugs. Don't let his silly hairdo fool you, he's a pretty ordinary guy and can be reached via email at shawn@linuxjournal.com. Or, swing by the #linuxjournal IRC channel on Freenode.net.

THE MOST BASIC CABLE PACKAGE FROM CHARTER (SPECTRUM?) COSTS ME MORE THAN $70 PER MONTH, and that's without any equipment other than a single cable card. It's very clear why people have been cutting the cord with cable TV companies. But, what options exist? Do the alternatives actually cost less? Are the alternatives as good? I've been trying to figure that out for a few months now, and the results? It depends.

The idea of cord cutting isn't new. For years, people have been severing their ties with cable companies in order to save money. The ever-persistent question is this: how do the options compare?

Real Time or On Demand?

When replacing cable TV, there are two main types of media in question. Services like Netflix, Amazon Prime and Hulu are great, but they don't provide live television. In fact, depending on the show and service, you might need to wait until the next day or even the end of a season before your desired shows are available. You usually get the advantage of no commercials, but the waiting often is unbearable if you're into television shows that end with cliffhangers.

It is interesting though, now that Netflix and Amazon have been so successful with their streaming services, they're beginning to get their own exclusive shows. This means that not only are the shows not delayed, but they're also actually not available at all via cable TV! Admittedly that phenomenon is fairly new (only the last few years), but it makes the case for streaming far stronger. Why pay $70 per month and still not get to watch *Jessica Jones*?

Also, many individual stations are starting to offer their own streaming options, so the days of paying for cable so you can see a particular HBO show are over. Broadcast networks are starting to offer streaming options too, so if you're just looking for the ability to watch particular television shows, even paying for multiple online accounts is cheaper than paying for cable—usually.

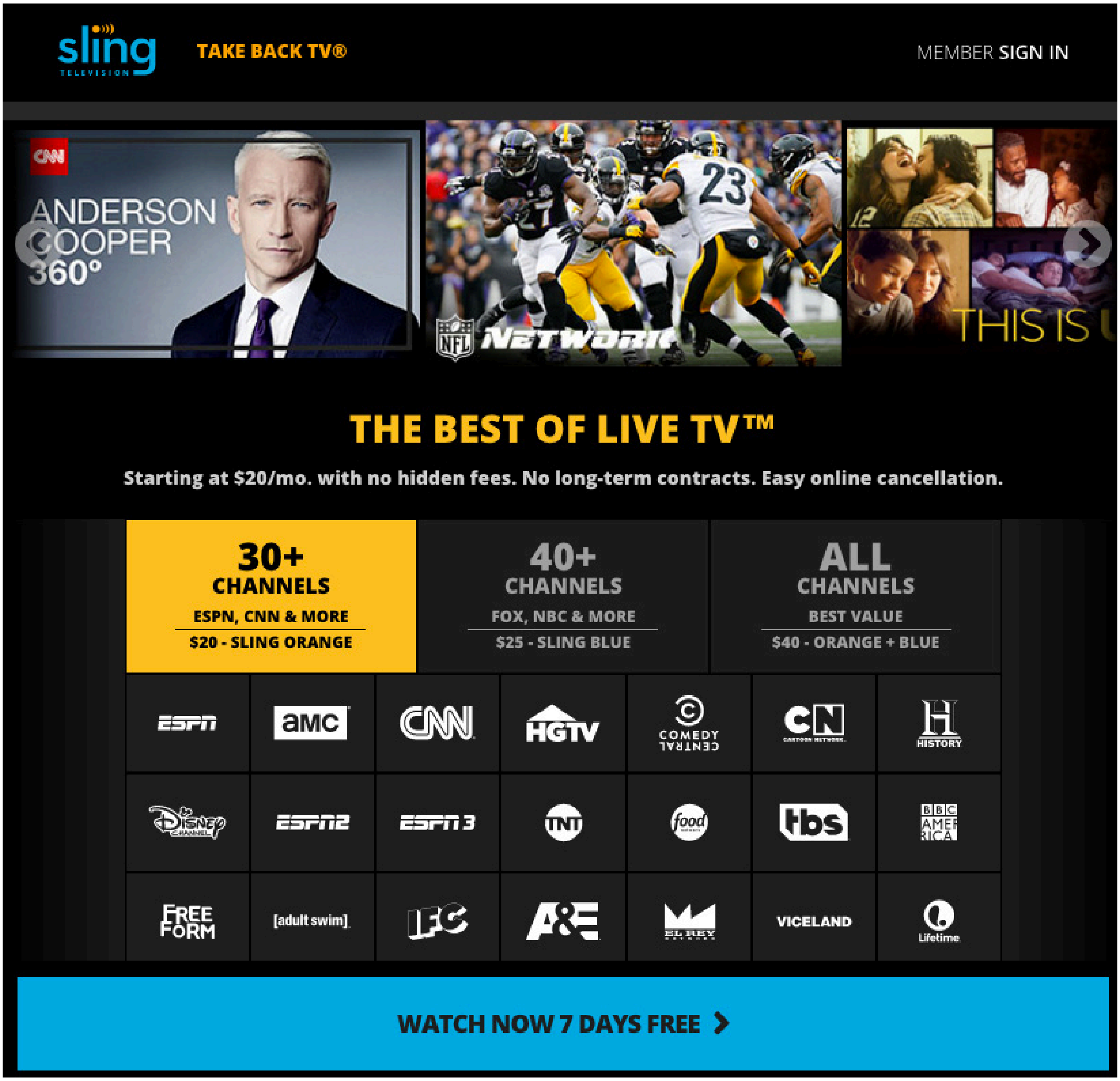
All Those Cable Channels...

Some of the biggest hurdles for cord-cutters are cable-only channels. I have a relative who watches only shows on the History Channel. And my mother-in-law couldn't live without watching movies on the Hallmark Channel. And everyone I know in real life is addicted to HGTV and its tiny house program. Those channels aren't big enough to support a full streaming platform (or are owned by actual cable companies, so they won't offer a non-cable alternative). So what's a cord-cutter to do?

Until recently, not much. Now, however, there are three really good options for streaming cable television stations, and one is almost really good. Those options aren't exactly cheap, and they're mostly US-only, but they're far less expensive than cable TV. The three options each have their quirks, but any of them are worth looking into if you have reliable internet speeds that aren't dependent on cable TV bundling. Currently, the three main options are Sling TV, PlayStation Vue and DirecTV Now.

Streaming Cable: Sling TV

Sling TV has been around the longest and is owned by Blockbuster (yes, *that* Blockbuster!), who in turn is a subsidiary of Dish Network. It has a large lineup of cable stations and several tiers of options that include packages like premium cable channels. Depending on promotions and where you live, the packages range from $20–$40 per month. If you live in a big city, you also might get local broadcast stations (NBC, ABC, CBS,



THE BEST OF LIVE TV™

Starting at $20/mo. with no hidden fees. No long-term contracts. Easy online cancellation.

30+ CHANNELS
ESPN, CNN & MORE
$20 - SLING ORANGE

40+ CHANNELS
FOX, NBC & MORE
$25 - SLING BLUE

ALL CHANNELS
BEST VALUE
$40 - ORANGE + BLUE

Figure 1. Sling TV has been around a long time, but the lack of DVR and video glitches make it less than stellar in my experience.

PBS, FOX), but for most of the country, you get those channels only "on demand", which means recordings of popular shows the next day.

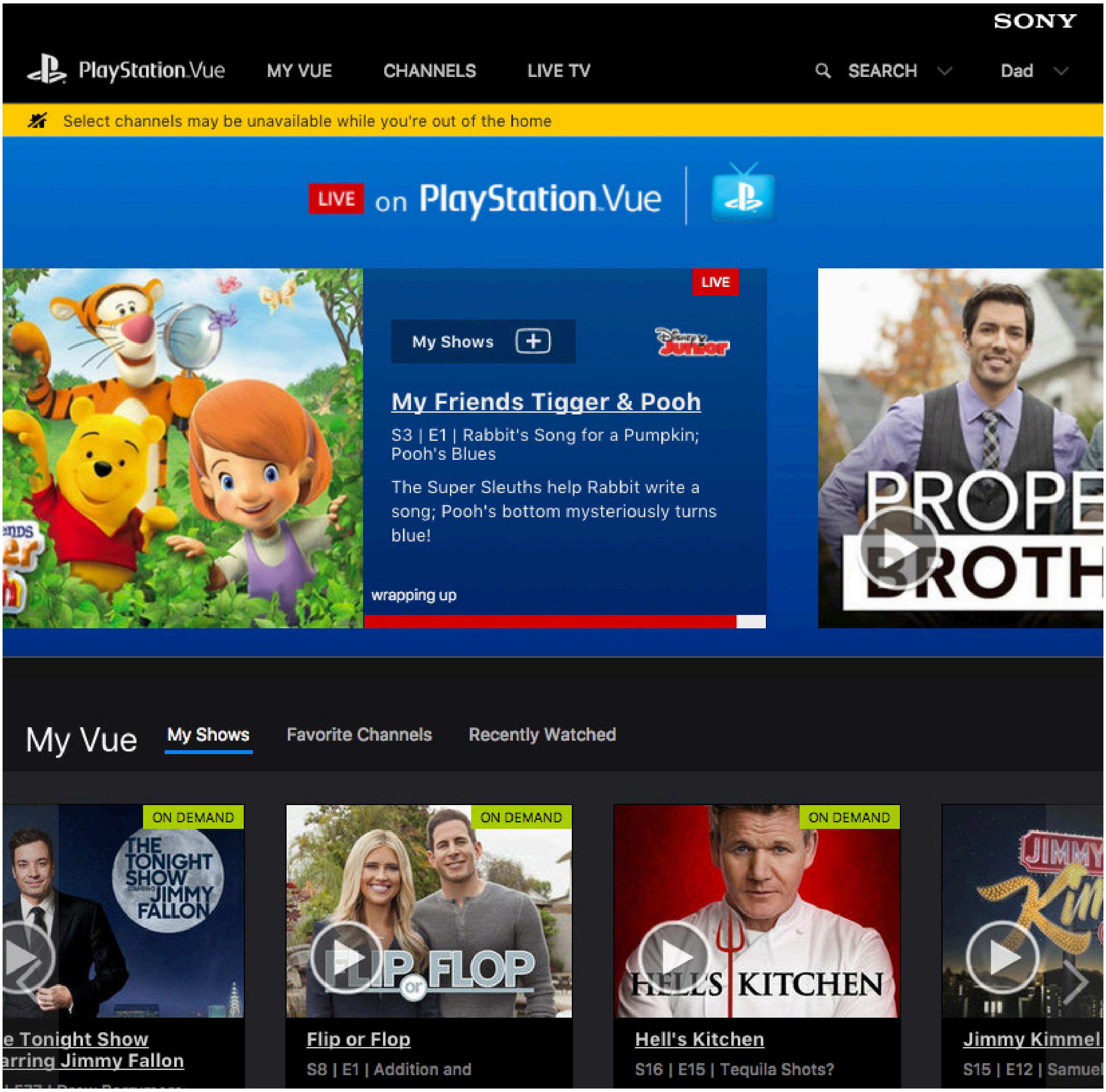
The technology details of Sling TV are a little confusing. If you subscribe to the lowest tier, you can stream only one channel per account at a time. That means if you are watching TV in your living room, you can't watch something else on your phone. If you subscribe to a higher-priced tier, you can have up to three streams at once. Also, although the streams usually are good quality, my anecdotal experience shows that there are a few more artifacts and glitches with Sling TV than with the other options, but nothing that makes it a showstopper. (I get glitches with my cable television too, so nothing is perfect.)

There's a free trial with Sling TV, so it's worth checking out. Just be sure to cancel it before your credit card auto-renews at the end of the trial, unless you decide to keep it. Also, because it's been around for a long time, Sling TV has apps on multiple platforms. Xbox users can install Sling TV, along with Android TV and Roku users. Like most streaming services, Roku does a great job of staying vendor-neutral, which means it usually can provide services regardless of who is providing them.

PlayStation Vue

PlayStation Vue is a bit more of a surprise, since Sony PlayStation is synonymous with gaming rather than television. Its offerings are impressive, however. The lineups are similar to Sling TV, but the breakdowns are a little different. The lowest-price service is around $30 per month, with other tiers available that add more channels. Sony gives you a price break if you're not in one of the cities that has local channels available, so for me in rural Michigan, it's cheaper than if I lived in Chicago. (That means I don't get local channels though, which is frustrating.)

Although the slightly higher price seems frustrating, the technology included might make up for it. Not only can you stream to five devices simultaneously, but it also provides "Cloud DVR", which automatically stores recorded content for you. All you need to do is mark a program as a favorite, and all episodes are saved for 30 days. It's not possible to schedule a timed event, but the DVR feature is extremely nice, and it provides a far better experience than the live-only Sling TV.



Select channels may be unavailable while you're out of the home

My Friends Tigger & Pooh

S3 | E1 | Rabbit's Song for a Pumpkin; Pooh's Blues

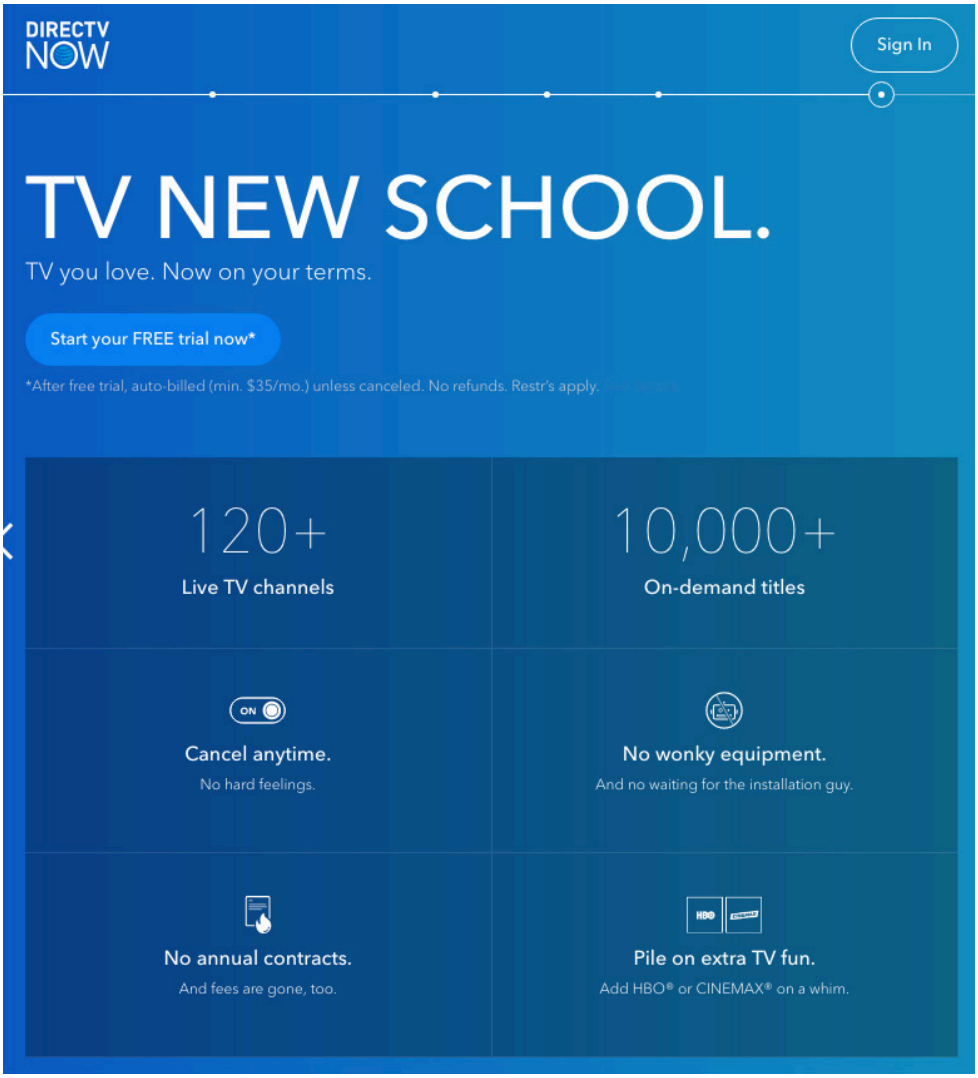
The Super Sleuths help Rabbit write a song; Pooh's bottom mysteriously turns blue!

Figure 2. PlayStation Vue is remarkable, until it's not. The video quality is amazing, and the DVR is superb. The geolocation frustrations along with PS4 console problems make it difficult to love.

The video quality with PlayStation Vue is shockingly good. Whether watching on a mobile device, a Roku or a PlayStation system, the video is far more reliable in my anecdotal trials. The five streams means people can watch TV in multiple rooms, and since Vue allows for individual profiles, different family members can have their own DVR'd shows. The only really big issue I've had with PlayStation Vue is that it's not possible to watch streams from the same account on two different PlayStation 4 consoles. I have a console in my office and a PlayStation Pro in the living room, and it's not possible to watch Vue on both devices. That is particularly frustrating, because watching on multiple Roku units works fine, but not on the actual Sony hardware! There's also some frustration with geolocation. Sony often thinks I'm not home, so it limits what I can watch. I would understand if my IP address changed, but I have a static IP address and I'm always connecting from home! (See the notice in Figure 2.)

DirecTV Now

DirecTV Now is the new kid on the block when it comes to cable TV streaming. The packages are similar to the other services I mentioned,



TV NEW SCHOOL.

TV you love. Now on your terms.

*After free trial, auto-billed (min. $35/mo.) unless canceled. No refunds. Restr's apply.

120+ Live TV channels

10,000+ On-demand titles

Cancel anytime. No hard feelings.

No wonky equipment. And no waiting for the installation guy.

No annual contracts. And fees are gone, too.

Pile on extra TV fun. Add HBO® or CINEMAX® on a whim.

Figure 3. DirecTV Now is the new kid on the block. The $35/month is a trial cost and likely will increase before this article is published.

with some initial low-priced options available to entice users away. (Note: with all these services being contract-free, the potential for moving in order to save a few bucks is very legitimate!) DirecTV Now has similar limitations regarding live broadcast stations (that is, at the time of this writing there aren't any available), but DirecTV Now has the additional limitation that even on-demand content from CBS isn't available. The kerfuffle that DirecTV and CBS have been having extends to the streaming service as well.

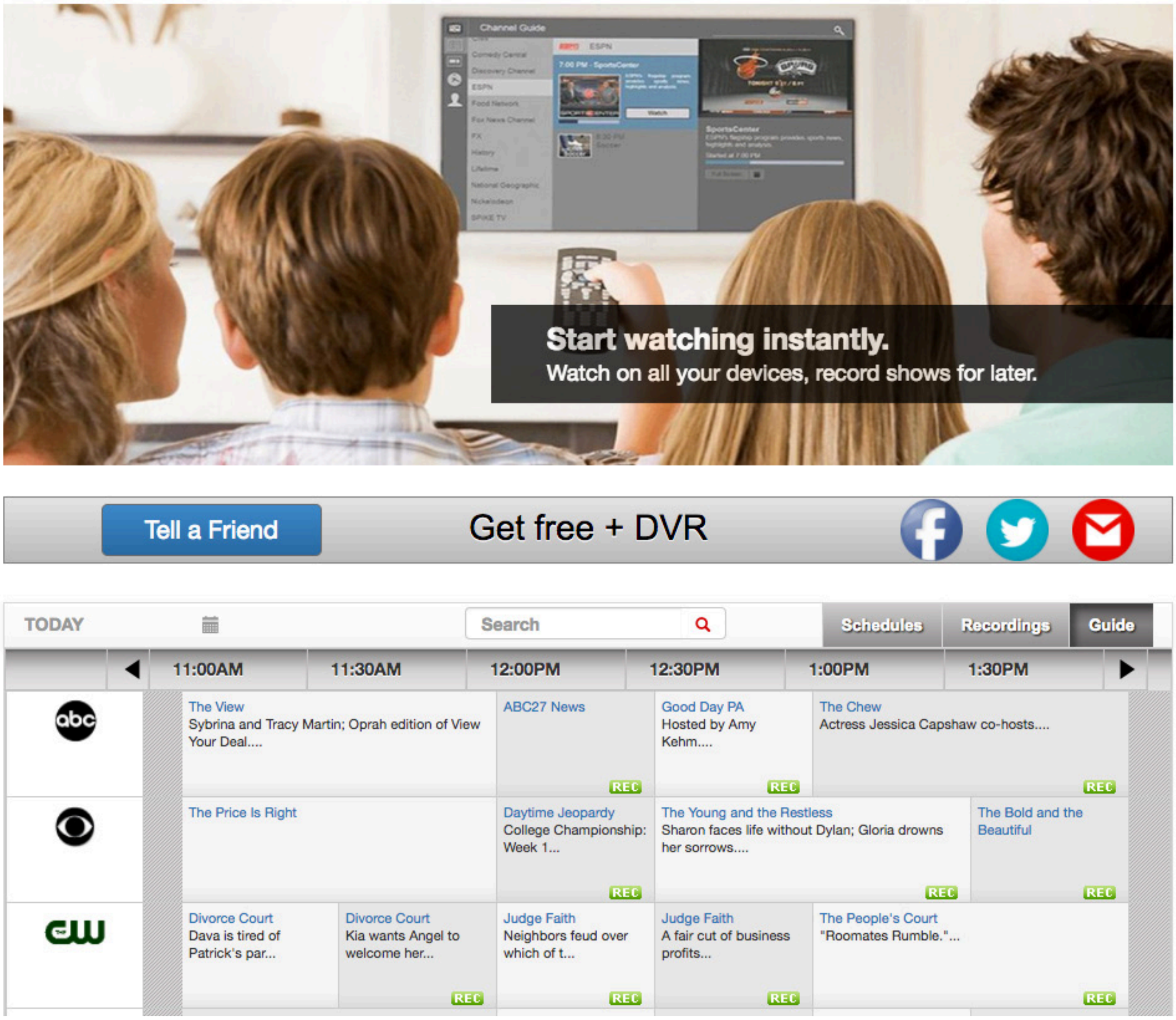
I haven't personally used the DirecTV Now service, because none of my devices currently are supported. (Apple TV is its main device, and you can get one free if you pre-pay for three months of service.) I have friends who've used it though, and they say the quality is very good. Like Sling TV, however, it doesn't currently have any DVR capability.

Since DirecTV Now is new, it's not fair to criticize its lack of hardware support yet. Roku streaming is slated for Q1 2017, and it's possible other non-competitors will get apps as well. As is usually the case, Roku likely will be one of the premiere ways to watch streaming cable TV service, because its compatibility will allow for service-hopping without hardware reinvestment.

USTVnow, the Sort of Option

USTVnow is a service designed for US citizens living outside the US and, therefore, unable to get US television. It's a streaming service that provides *live* network channels (ABC, CBS, NBC, PBS, FOX) for free, and for a monthly fee, it adds a few cable channels (28 total) and HD streaming as well. There is also some DVR service included with the premium packages. Any time I've tried to use USTVnow inside the US, it's worked perfectly, so there aren't any apparent geographical restrictions. Honestly, on paper, it's the best thing going.

Unfortunately, the times I've used USTVnow, I've had lots of glitches. Usually it's during busy times (Super Bowl party, for instance) that the service glitches, but since those are the times I want it to work the most, it's been a frustrating service. The pricing is competitive, however, especially since the SD free tier is really free and provides live broadcast stations. As with most services, Roku seems to be the best way, apart from a browser, to consume USTVnow.


Start watching instantly.
Watch on all your devices, record shows for later.
Tell a Friend
Get free + DVR
TODAY
Search
Schedules
Recordings
Guide

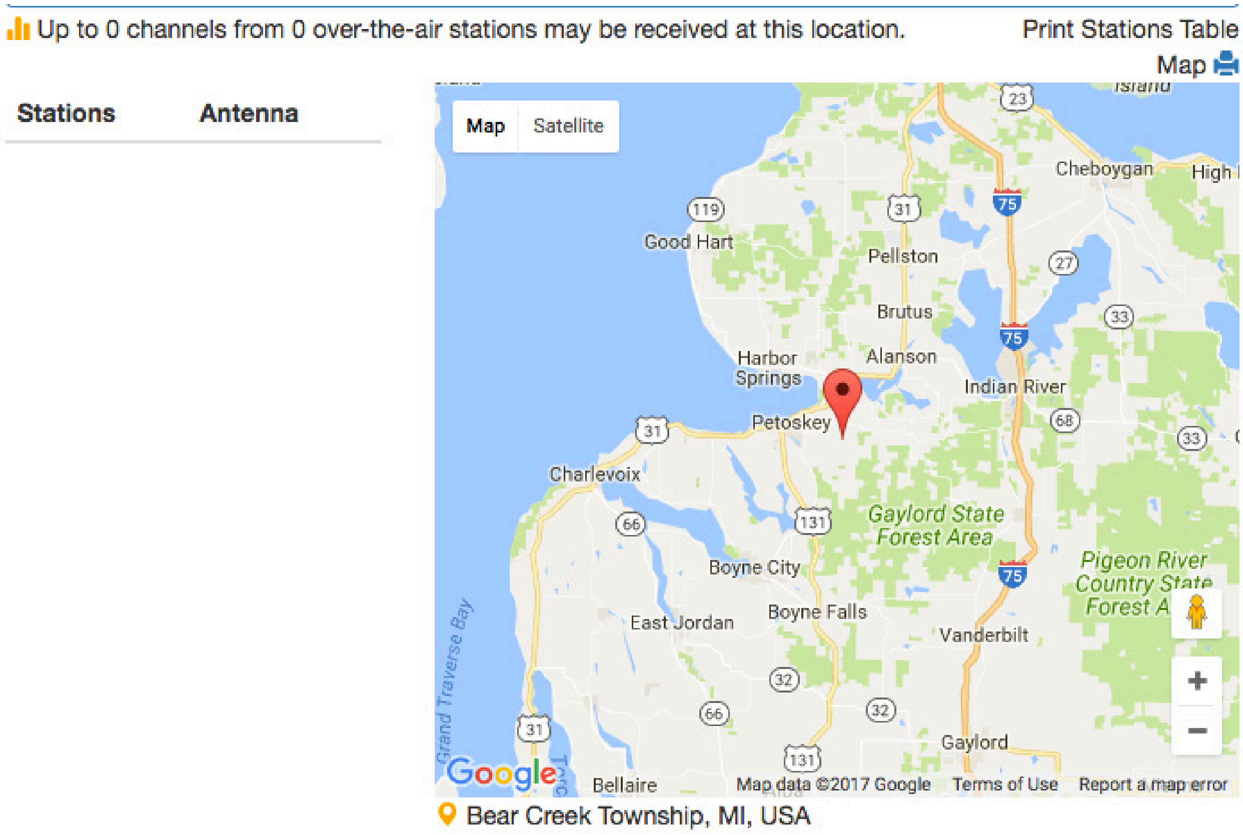
	11:00AM	11:30AM	12:00PM	12:30PM	1:00PM	1:30PM
abc	The View Sybrina and Tracy Martin; Oprah edition of View Your Deal....		ABC27 News REC	Good Day PA Hosted by Amy Kehm.... REC	The Chew Actress Jessica Capshaw co-hosts.... REC	
	The Price Is Right		Daytime Jeopardy College Championship: Week 1... REC	The Young and the Restless Sharon faces life without Dylan; Gloria drowns her sorrows.... REC		The Bold and the Beautiful REC
cw	Divorce Court Dava is tired of Patrick's par...	Divorce Court Kia wants Angel to welcome her... REC	Judge Faith Neighbors feud over which of t... REC	Judge Faith A fair cut of business profits... REC	The People's Court "Roomates Rumble."... REC	

Figure 4. I want to love USTVnow, and perhaps now that there is a paid service, the reliability will improve. I just hope it's able to keep providing live broadcast channels in the US.

I want to love USTVnow. I have no idea how it's able to provide service in the US when the other options struggle to provide broadcast stations. Hopefully, it's not a loophole that will be closed, because for some folks, it's the only way to get broadcast channels at all, even if they *do* live in the US.

Rabbit Ears

Yes, obviously using an antenna is a great way to get local television. In



Up to 0 channels from 0 over-the-air stations may be received at this location.

Bear Creek Township, MI, USA

Figure 5. "Up to 0 channels" is a sad thing to see; I hope your location is better.

fact, you can head over to http://antennaweb.org and see what channels are available in your area and what sort of antenna you'll need. The site even will tell you what direction to point your antenna for the best signal. If you're just looking for some old-fashioned television, an antenna is often a good option. Plus, apart from the hardware, it's totally free.

The problem is, even though I live in a (small) city, I get exactly zero channels from my location. That is due to geography, because I live on the side of a hill, but nonetheless, I can't get any channels using even a rooftop antenna. Even if you can, however, it's worth considering whether that sort of system is acceptable for you. I don't want to switch my input source on the television every time I want to watch TV. And TiVo has spoiled me; I want to pause live TV. It's possible to get something like an HD Homerun device from Silicon Dust and convert your antenna signal into a digital stream, but integrating that into your entertainment system is often challenging. Plus, I had so much frustration with my HD Homerun setup in our last house that I opted to just buy a cable TV subscription.

So OTA (over the air) channels are worth checking out, and for some people, they are more than enough. For me, however, even if I could get a good signal, I want more.

What about the Parade?!

If you live in a big city and can get local channels via services like Sling TV or PlayStation Vue, things like watching the Thanksgiving Day parade are possible. For me, the only way I can watch live events is with the not-always-reliable USTVnow. And even those channels aren't local, so I can't ever watch the local news. My big issue used to be how to watch the Olympics, but thankfully, many streaming options are available now. Still, watching the Thanksgiving Day parade is something I've done my whole life, and without some way to see local channels, all the cable channels in the world don't help.

Because I have thousands of dollars invested in my TiVo infrastructure (lifetime Roamio subscription and four TiVo Minis), I'm still paying for the lowest tier of cable TV. I find that we almost never switch over to the TiVo, however, so in the next few months, I might bite the bullet and cancel cable TV altogether. For most folks, services like Sling TV, PlayStation Vue and DirecTV Now provide more than enough service at a fraction of the cost. So if you live in a big city or can live without live broadcast channels, I urge you to give them a try. Each is available with a free trial, and if you sort through the various pros and cons, coming up with a satisfactory service is pretty easy.

If you've cut the cord (many of you I've spoken with already have done so), I'd love to hear about your specific solution. Do you just switch sources on your TV and use rabbit ears? Do you strictly Netflix and chill? Have you sold your television and reverted to books? (I'm often tempted.) Please let me know. I'd love to follow up with some alternatives for folks like myself who are still struggling to cut the cord.■

Send comments or feedback via http://www.linuxjournal.com/contact or to ljeditor@linuxjournal.com.

RETURN TO CONTENTS

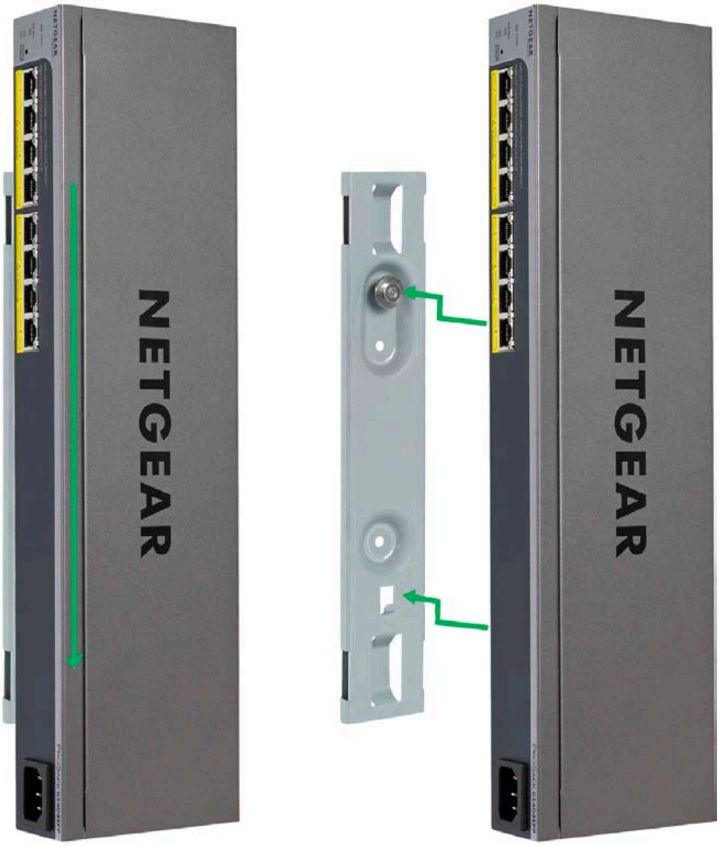
PREVIOUS
Shawn Powers'
The Open-Source
Classroom

NEXT
Feature: Big Data
Demonstrator

HOSTING™

HOSTING Monitoring Insights

An important need for today's CIOs is gaining greater granular visibility into hybrid cloud and on-premises environments to maximize the business value of their IT assets. To make progress in this arena, two natural allies—managed cloud services provider HOSTING and hybrid IT monitoring specialist ScienceLogic—teamed up to develop HOSTING Monitoring Insights, an innovative hybrid cloud monitoring solution that delivers "the industry's first holistic and comprehensive view of hybrid cloud environments". A view into today's situation for many organizations reveals the use of multiple platforms to monitor devices across disparate cloud and on-premises environments. The core customer benefit from Monitoring Insights is capacity to manage the health of critical business processes proactively across the whole environment and significantly reduce the quantity of monitoring tools required. Monitoring Insights is available both directly from HOSTING or through a HOSTING partner, and at various service levels based on customer need. http://hosting.com



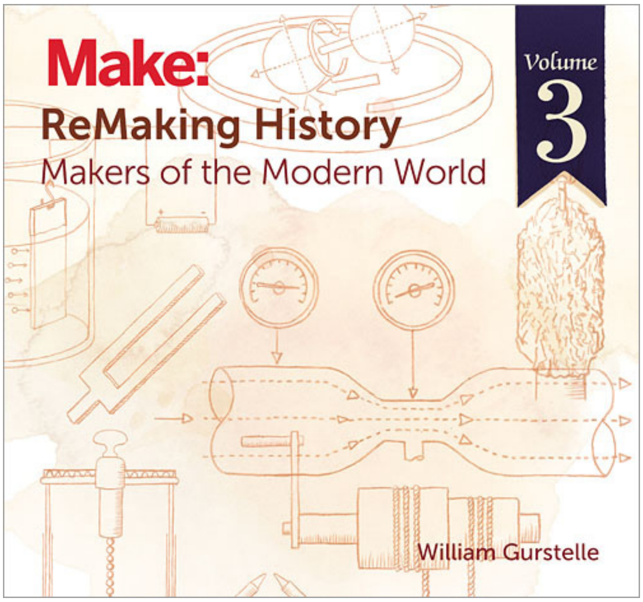
NETGEAR, Inc.'s GSS108EPP and GS408EPP Switches

Two new NETGEAR, Inc., Power-Over-Ethernet (PoE) switches feature a novel "Virtually Anywhere" mounting system that delivers modern high-power PoE+ to devices that others cannot. The mounting system on the two models—the ProSAFE 8-port Gigabit Ethernet Web Managed PoE+ Click Switch (GSS108EPP) and the ProSAFE Easy-Mount 8-port Gigabit Ethernet PoE+ Web Managed Switch (GS408EPP)—offers ultimate flexibility in placement so that PoE+ ports are available exactly where needed to power WAPs, VoIP phones, IP surveillance cameras and IoT devices. In any orientation, they can be mounted on a wall, strapped to a pole or tucked under a desk or tabletop. In addition, the GS408EPP's unique design allows for two switches to be mounted in a single 1U rack slot, saving valuable rack space while allowing for future expansion. Both switches also provide configurable, advanced Layer 2 network features. http://netgear.com

Minifree Ltd.'s GNU+Linux Computers

Minifree Ltd.—doing business as "Ministry of Freedom"—exists mainly for reasons Linuxers will like: to make it easier for people to get computers that respect their freedom and privacy, and to provide funding for a meaningful project, called Libreboot. Minifree describes Libreboot as a free (libre) and open-source BIOS/UEFI replacement that offers faster boot speeds, better security and many advanced features compared to most proprietary boot firmware. Minifree recently announced availability of three computers: the Libreboot C201 laptop, the Libreboot D16 Desktop and Libreboot D16 Server. All come with the Libreboot firmware and Debian GNU+Linux operating system preinstalled and are free of unwanted bloatware, DRM, spyware or restrictions on computer usage rights. The Libreboot C201 laptop is a configurable, lightweight and portable laptop ideal for anyone needing a small, lightweight computer for travel, work or general entertainment purposes. The Libreboot D16 Desktop is a configurable, high-end, business-grade, secure owner-controlled workstation free of backdoors implanted by the NSA and other agencies. Finally, the Libreboot D16 Server is a configurable, high-end, business-grade, secure owner-controlled server, also free of the aforementioned backdoors. Minifree ships its machines worldwide from the United Kingdom.

http://minifree.org

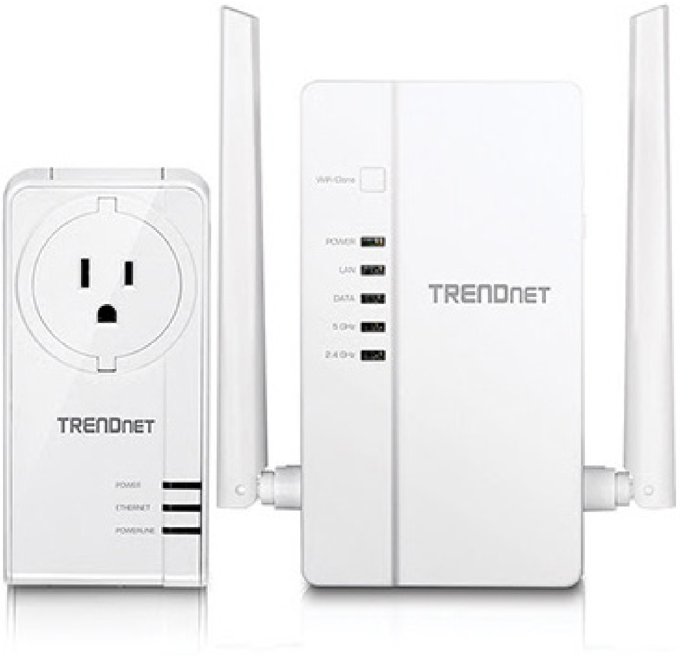


William Gurstelle's *ReMaking History*, Volume 3 (Maker Media, Inc.)

In William Gurstelle's *ReMaking History* series from Maker Media, Inc., readers get exponentially closer to the inventors who shaped our modern world compared to other histories of technology. This is because Gurstelle doesn't merely tell the stories of remarkable inventors from the past, he gets into their fascinating minds by illustrating how to make one's own version of the inventor's handiwork. The new Volume 3 of *ReMaking History* bearing the subtitle *Makers of the Modern World* explores the early modern era and builds on the earlier two volumes covering pre-modern history to the Industrial Age. In this volume, seven inventors and their technologies—destined to fill basements and garages everywhere—include Alessandro Volta and electroplating; Humphrey Davy and the first electric light; George Cayley and the aeronautical glider; the Lumiere Brothers and the movie projector; Rudolf Diesel and the automobile engine; Hans Goldschmidt and the thermite reaction; August Möbius and the Möbius Strip; and Louis Poinsot and loads, moments and torques.
http://oreilly.com

William Rothwell and Nick Garner's *Certified Ethical Hacker* (CEH) Complete Video Course (Pearson IT Certification)

Watch William Rothwell and Nick Garner's new *Certified Ethical Hacker* (CEH) Complete Video Course and learn everything you need to know to ace the CEH exam in less than 11 hours. Divided into five modules and containing a complete overview of the topics in the EC-Council Blueprint, Rothwell and Garner's intermediate-level video-training course helps viewers master the essentials needed to pass the exam. The course commences with a general overview of security essentials, followed by an exploration of system, network and web services security, and a dive in to wireless and internet security. To test one's chops, the course offers quizzes, exercises and two full practice exams. By providing the breadth of coverage necessary to learn the full security concepts behind the CEH exam, this video course helps prepare viewers for a career as a security professional.
http://informit.com



TRENDnet®

TRENDnet's WiFi Everywhere Powerline 1200 AV2 Access Point, Model TPL-430AP

TRENDnet recently released an innovative new solution that creates dead-spot-free home Wi-Fi by leveraging a home's existing electrical system. TRENDnet's WiFi Everywhere Powerline 1200 AV2 Access Point adapter, model TPL-430AP, uses existing electrical lines to create a wired home network and doubles as an access point to deliver a wireless signal from nearly any power outlet. The adapter supports Powerline 1200 networking with a built-in dual-band wireless AC1200 access point and features MIMO with Beamforming technology to enhance performance and range. WiFi Clone support duplicates one's existing wireless network settings, reducing Wi-Fi setup time from minutes to just seconds. Powerline range functions up to 984 feet (300m) over existing electrical lines.
http://trendnet.com

SUSE Linux Enterprise High Availability Extension

Historically, data replication has been available only piecemeal through proprietary vendors. In a quest to remediate history, SUSE and partner LINBIT announced a solution that promises to change the economics of data replication. The two companies' collaborative effort is the headliner in the updated SUSE Linux Enterprise High Availability Extension, which now includes LINBIT's integrated geo-clustering technology. Providing a new capability to replicate data across unlimited distances, LINBIT has enhanced SUSE's high availability solution that is built on open-source software and runs on commodity hardware. The LINBIT solution guards against failures or disasters by providing policy-driven mechanisms for customer applications, and data, to continue operations in another geographically dispersed data center. SUSE Linux Enterprise High Availability Extension is an integrated suite of open-source clustering technologies—from LINBIT and others—that enables customers to eliminate single points of failure, thus helping to maintain business continuity, enable compliance, protect data integrity, maintain isolation for multiple tenants and reduce unplanned downtime for mission-critical workloads. http://suse.com and http://linbit.com

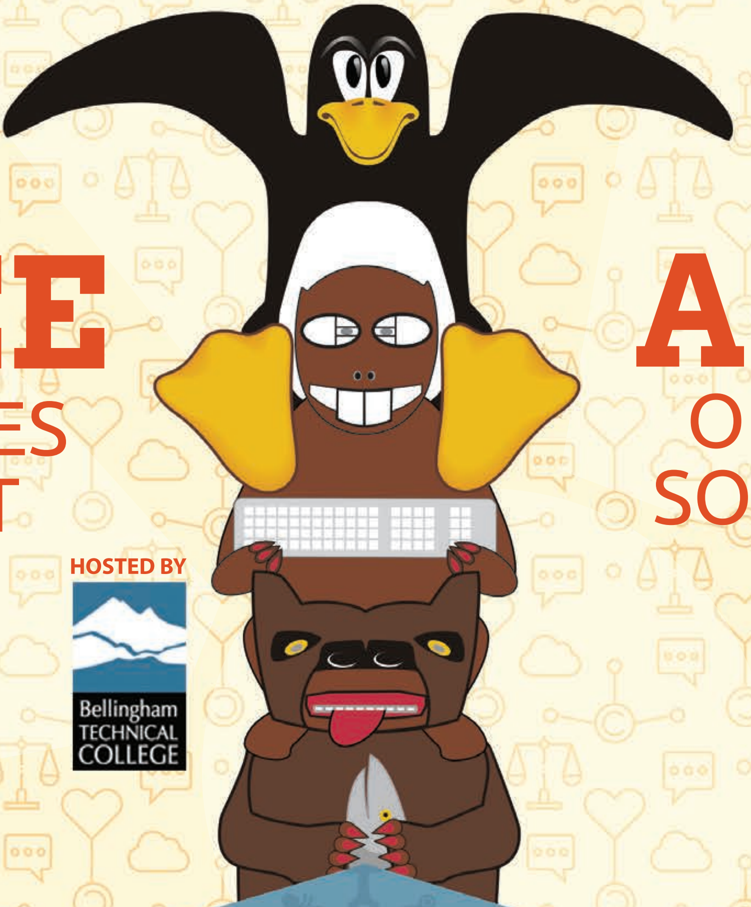
EXHIBIT HALL A

Briggs & Stratton 8,000 Watt Elite Series Portable Generator with StatStation Wireless

Although *Linux Journal* readers might not equate Milwaukee with tech, a new Briggs & Stratton product portends the bright future of smartened "legacy" devices from the industrial heartland. The Wisconsin-based maker of engines and industrial products recently announced a smarter—and "techier"—way to produce on-demand power in the form of the new Briggs & Stratton 8,000 Watt Elite Series Portable Generator with StatStation Wireless, featuring Bluetooth technology. The accompanying StatStation app for Android and iOS, with support from Bluetooth connectivity, provides valuable remote visibility into key metrics, such as fuel level and remaining runtime, runtime meter, percent of available Watt consumption, maintenance reminders, dealer locator, reference guides and how-to videos.
http://briggsandstratton.com

Please send information about releases of Linux-related products to newproducts@linuxjournal.com or New Products c/o Linux Journal, PO Box 980985, Houston, TX 77098. Submissions are edited for length and content.

RETURN TO CONTENTS



BIG DATA DEMONSTRATOR

USING HADOOP TO BUILD A LINUX CLUSTER FOR LOG DATA ANALYSIS USING R

THIS ARTICLE WALKS THROUGH THE STEPS TO CREATE A HADOOP LINUX CLUSTER IN THE CLOUD AND OUTLINES HOW TO ANALYZE DEVICE LOG DATA VIA AN EXAMPLE IN THE R PROGRAMING LANGUAGE.

RUNE TORBENSEN and SØREN TOP

◀ PREVIOUS
New Products

NEXT
Feature: Integrating Web Applications with Apache ▶

This article describes why device log data analysis is useful and briefly introduces the involved technologies and how they fit together. Linux is the basis for the "Infrastructure as a Service" standard that makes the proposed solution portable between cloud providers. Furthermore, we describe the steps you need to go through to create a Hadoop cluster based on Linux in an Amazon cloud. The steps involve bash/install scripts placed in a GitHub repository that allows the automatic installation of all the necessary components and configuration.

Big Data technology and the Internet of Things (IoT) are a strong combination. The IoT is a great source of information and Big Data technology allows for analysis of vast amounts of data. Possible applications are prediction, anomaly detection and device improvement/development. The latter is the case we have been working on in order to investigate why devices break. We need Big Data technology, because classical single-server approaches were unable to process the large amounts of data fast enough for an efficient analysis work cycle.

In this article, we use an example of data analysis of device log data to illustrate how to use a demonstrator setup to find unknown correlations between parameters in log data. However, in this article, we will not go into the device details, but use an abstracted device model approach.

DEMONSTRATOR OVERVIEW

The demonstrator setup consists of selected technologies (Figure 1), developed scripts and installation instructions. This allows for the reproduction of the setup.

The so-called cloud is a dynamic market for computer resources. Prices are decreasing over time,

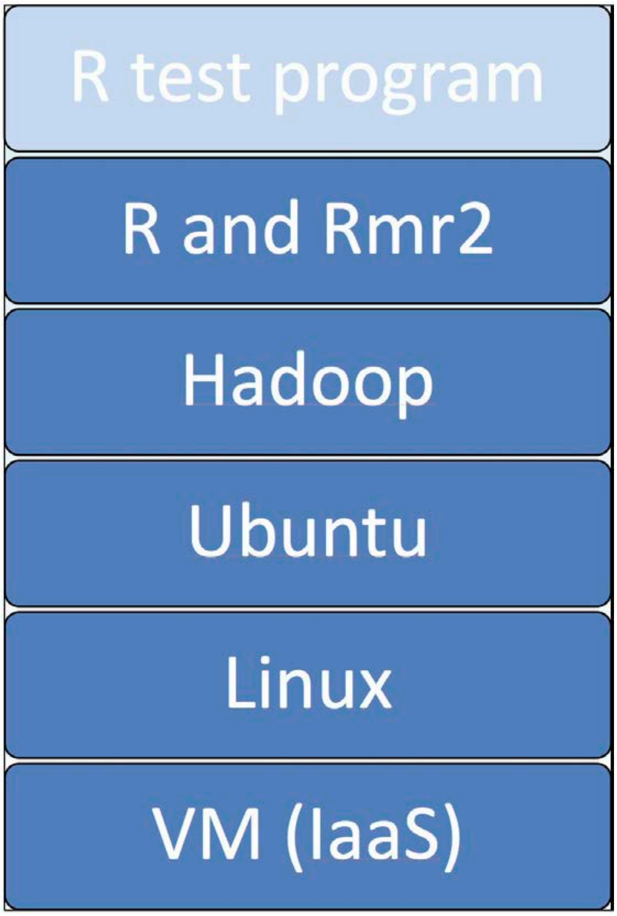


Figure 1. Selected Technologies Layer Model

but it is far from free, and a credit card is required to get a cloud account. The cloud is important in Big Data analysis, because we require large computer resources only from case to case, and a standing in-house computer cluster is in many situations too expensive, especially for smaller organizations to begin learning Big Data analysis.

When we want to analyze data, we select a cloud provider and create a cluster, do the analysis, and then afterward, we destroy the cluster. This way, we pay only for the computer resources used during the data analysis.

We have selected Infrastructure as a Service (IaaS) as the cloud technology, because it allows for portability of the demonstrator between different cloud providers.

All the cloud providers that we know of offer virtual Ubuntu Linux machines. Ubuntu is a well known Linux distribution, which is why we developed the demonstrator (installation script) for Ubuntu.

We chose to work on Amazon Web Services (AWS), since it's a well established, stable business with well defined and documented interfaces, and it offers a free tier that is very convenient for development work.

Hadoop is a kind of overlay operating system for a cloud cluster of Linux computers. It handles all the resources in the cluster and allows programs to be executed in a distributed manner. Hadoop is written in Java and is rather memory-consuming (when compared to smaller jobs and testing). Hadoop has been the de facto standard for Big Data processing for the past ten years or so. Hadoop consists of a number of components—see Figure 2.

At the bottom is the Hadoop Distributed File System (HDFS) that allows for high data throughput. It handles the I/O bottleneck problem when analyzing vast amounts of data. The data is spread out over the cluster, and the idea is that data is processed where it is stored.

YARN (Yet-Another-Resource-Negotiator) is the central component that allocates resources to Hadoop jobs. Keep in mind that this is no trivial task, because a Hadoop cluster may contain 1000+ nodes. YARN has built-in logic to handle node and job failure in a graceful way. Nodes may disappear and reappear on the cluster network, but jobs must be taken over by other nodes in the meantime.

Map-Reduce is the component that handles the parallelization of

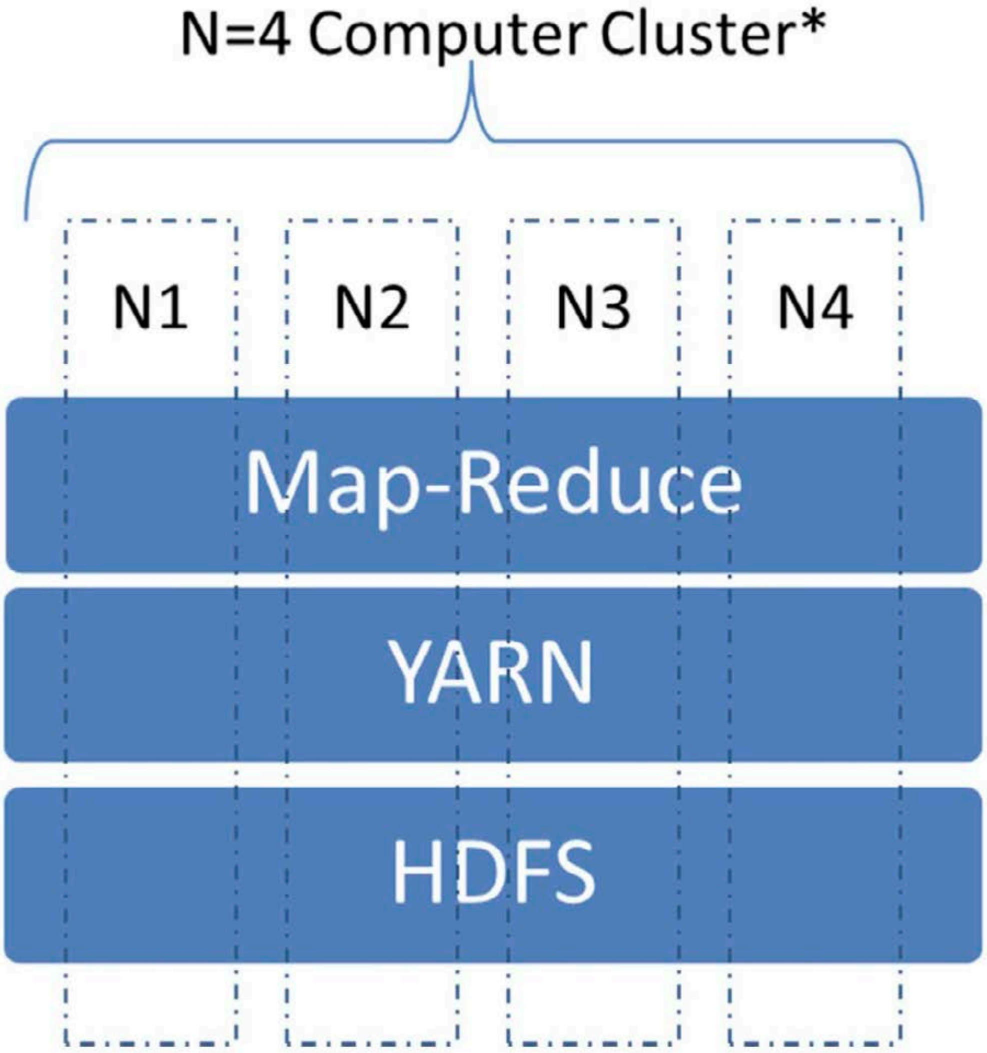


Figure 2. The Main Components of Hadoop

analysis tasks. Hard disk and network issues are abstracted away from the developer in order to allow the developer to concentrate on developing the analysis (program). The Hadoop system handles these issues automatically. Be aware that the Map-Reduce framework enforces parallel programming by constraining the programing model, and that can be difficult to get used to. There is a Map function that is responsible for importing data and converting to the internal data format (key-value)—for example, the key is the device ID and the value is a list of temperature data points. There is also a reduce function, one per slave node, that processes data with a certain key. This means, in our case, that all data for one device will be processed by the same slave node.

The Hadoop system will feed the map function with data records. It may be line by line or file by file. The system will distribute the load by dividing the files among the slave nodes for "map" processing. There is no direct filesystem access. The only way to output results is to emit a key-value pair or a list of key-value pairs. There is no shared memory between map instances; each map-node is doing the work on its own—hence allowing for decoupled, parallel processing of data. The developer has no control over which node will process what data. The key-values emitted by map instances are sorted by the system according to the key. Other than this, you cannot assume any ordering of key-value-pairs.

In order to speed up execution, the map function may be used to filter out records that are not relevant for the analysis. The Reducer function will receive a list of key-value pairs with a certain key for processing. When done, the function emits key-value pairs. The Hadoop system combines all the key-value pairs from the reducers into the output.

On top of the Hadoop Linux cluster, we have chosen R as the data analysis software. R is a generic math tool that provides a fast interactive process, which is fundamental for data analysis. R is a high-level programing language with many extension packages. This stems from the fact that R is open source and has a large community. Among its packages is data mining. R has a command line that allows an interactive process and fits well with the UNIX environment (scripting).

However, R is classic single computer software; therefore, the R package

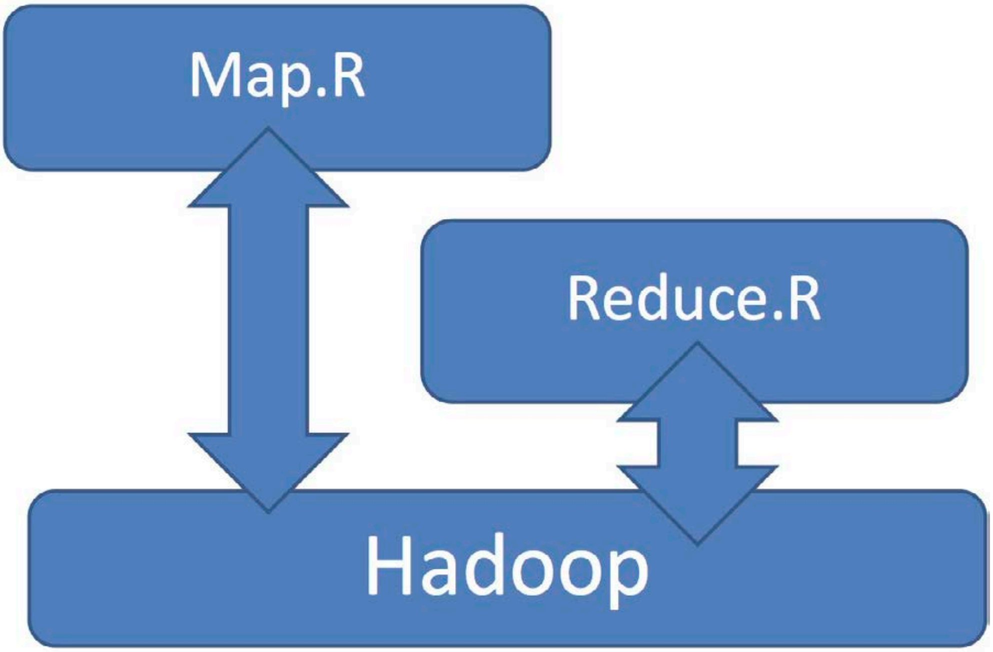


Figure 3. Map-Reduce Functions in R

Rmr2 is needed to allow R programs to run on a Hadoop Linux cluster. Rmr2 is mainly a wrapper of Hadoop into the R environment, and the map-reduce programming model applies to it as shown in Figure 3.

You now can use the entire R language and extensions to write the map/reduce functions. To use Rmr, you must follow these steps: 1) Store input data in HDFS. 2) Write the functions: Map and Reduce. 3) From R, call the map-reduce framework and point to the map and reduce functions. 4) Get output data from HDFS.

USING THE DEMONSTRATOR

Before diving into the steps of creating a Hadoop Linux cluster, we want to describe the work flow of the data analysis that we propose (Figure 4).

First, you have to collect and store data in order to have anything to analyze. It has to be stored on a server somewhere on the internet. Next

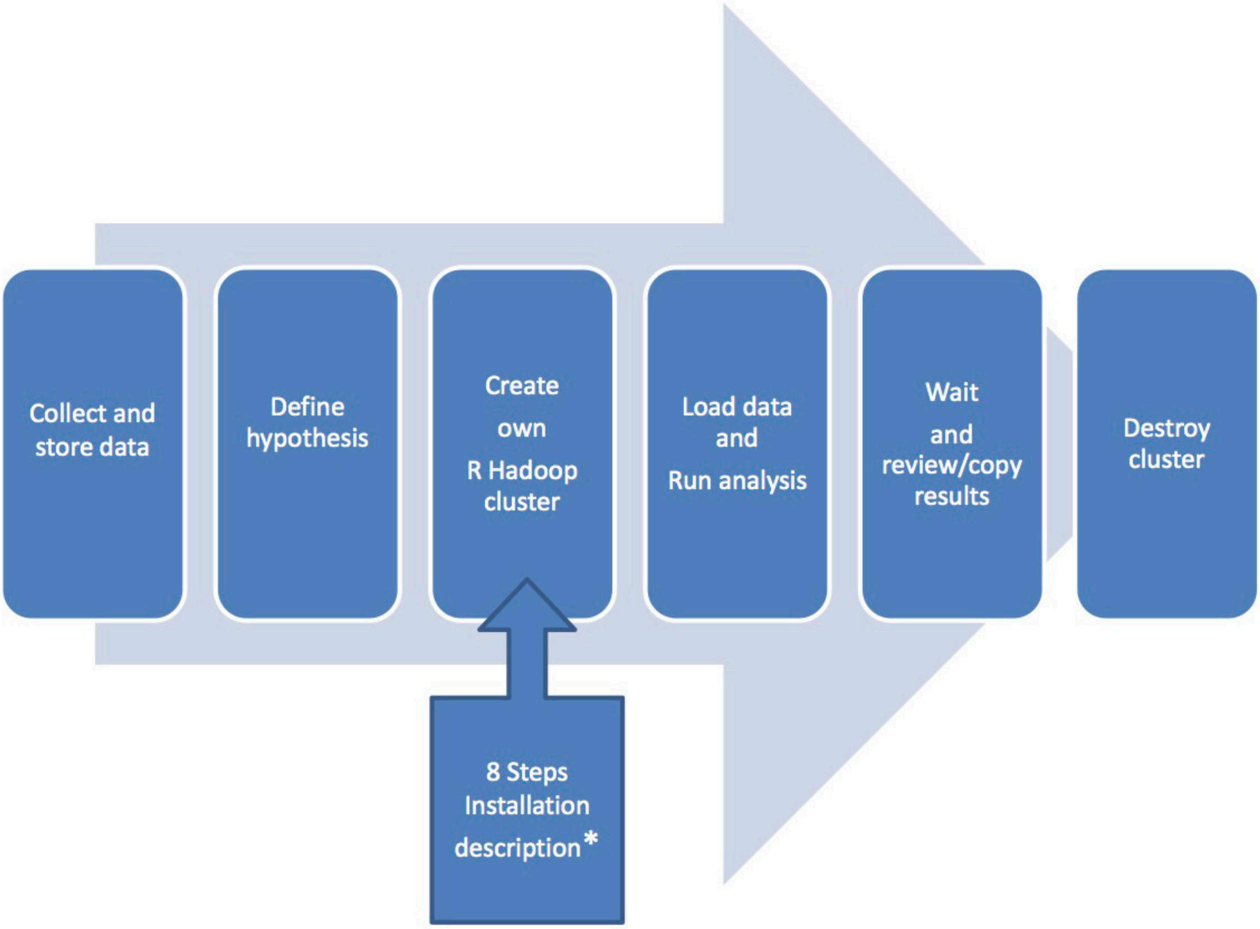

Collect and store data
Define hypothesis
Create own R Hadoop cluster
Load data and Run analysis
Wait and review/copy results
Destroy cluster
8 Steps Installation description*

Figure 4. Work Flow with the Demonstrator (*Requires Cloud Account)

you should determine what and how you want to analyze. In other words, you define your hypothesis and write an analysis program in R. Then create your own R Hadoop cluster. We'll describe the details in the next section. When your R Hadoop Linux cluster is ready, you load your data from the external server into the HDFS and run the analysis program using the R command prompt.

When the results are ready, you should review them and copy the result data to a storage server. If the results are not satisfactory, you should change the analysis program and run it again on the cluster. Finally, when done, you should destroy the cluster, since keeping disk and CPU allocation will cost too much if you are not using it. However, it may be sensible to keep the master (template) image, if you want to do more analysis in the future.

CREATE YOUR OWN HADOOP LINUX CLUSTER

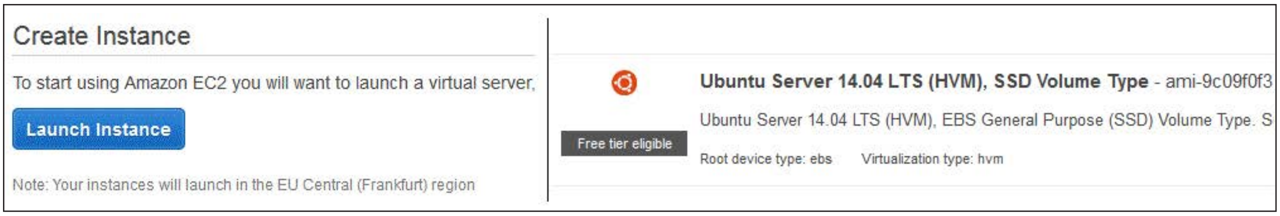
In this section, we provide one-line commands that can be copy/pasted into a Linux SSH console. We assume familiarity with AWS virtual machines; there are many tutorials and videos online.

First, install an Ubuntu server that will serve as Master and template for all the slaves:

1) Launch an Ubuntu server 14.04 LTS via Amazon WS web interface (Figure 5).

2) Log in using SSH to your server, and enter the following commands. These will download the installation script and run it:

```
> wget https://raw.githubusercontent.com/Rustor/EE-DIGI/
➥master/install-big-tools-demo.sh
> bash install-big-tools-demo.sh
```



Create Instance

To start using Amazon EC2 you will want to launch a virtual server,

Note: Your instances will launch in the EU Central (Frankfurt) region

Ubuntu Server 14.04 LTS (HVM), SSD Volume Type - ami-9c09f0f3

Ubuntu Server 14.04 LTS (HVM), EBS General Purpose (SSD) Volume Type. S

Root device type: ebs   Virtualization type: hvm

Figure 5. On the left is a screen dump of the Launch button, and on the right is the Ubuntu Server used.

The script will install SSH, Hadoop, R and Rmr2 on the Ubuntu server. Hadoop is installed in the "ubuntu" user's home directory, and the Hadoop data files (HDFS) will be placed in /tmp. In addition, it will copy the Hadoop configuration data (basic configuration) from the EEdigi GitHub repo to the Hadoop installation on the server. You can find further code comments in the script. A variation point is, for instance, that you can comment out the R part of the script before running it and install other analysis software, such as Python.

Next, here are the steps to create the template.

1) Shut down the server with:

```
> sudo init 0
```

2) Make a snapshot of the master in the AWS web interface (Figure 6).

3) Start the master again in the web interface, and log in via SSH.

Then, start forming the cluster based on the master template. All the slaves will know the master's internet address (the hostname is defined as "fe1") and public key, because you have generated the slaves from the master image. However, the slaves are unknown to the master, and therefore, you now run the following script on the master. The script opens a server that accepts slaves into a list of slaves. The slaves-list is the single point that defines the Hadoop Linux cluster.

4) Run auto-config.sh on the master to accept new slaves:

```
> bash auto-config.sh
```

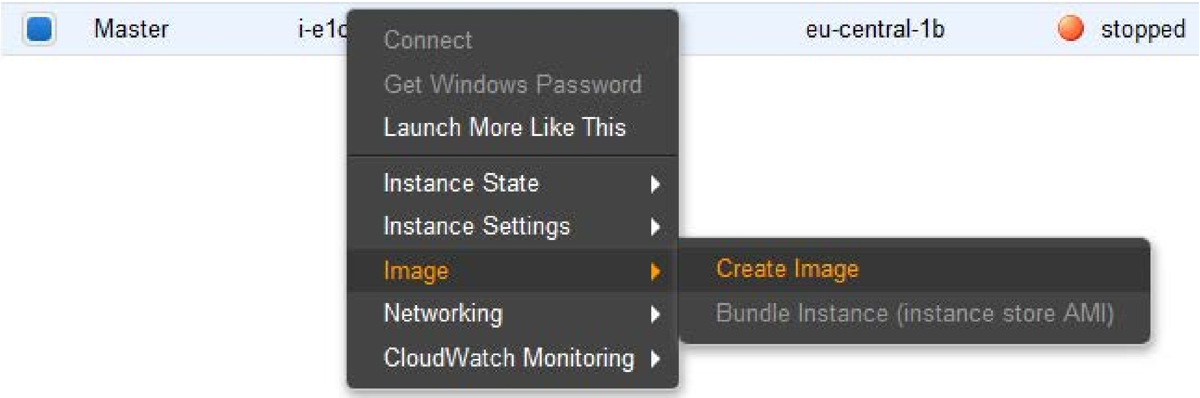
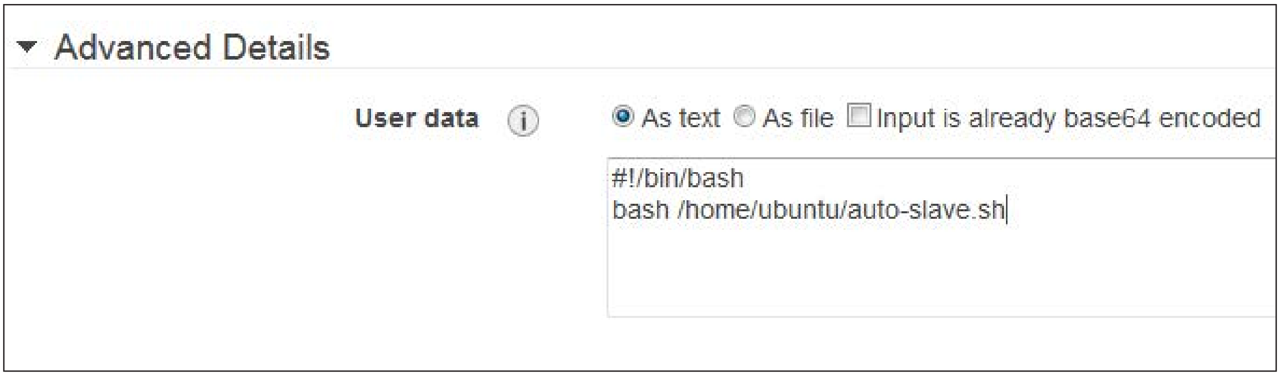


Figure 6. Screen Dump of AWS Create Image Menu



```
#!/bin/bash
bash /home/ubuntu/auto-slave.sh
```

Figure 7. Screen Dump of the AWS Field for User Data

5) Based on the master template, launch two (variation point) instances in the AWS web interface (Launch Step1: My AMIs) and insert these two lines in the AWS interface (user data), as shown in Figure 7:

```
#!/bin/bash
bash /home/ubuntu/auto-slave.sh
```

6) Wait some (ten) minutes for all the slaves to come online. If you want, on the master, in another console window, check the online slaves on the master during the process:

```
> cat cluster-config.file
```

7) Break auto-config.sh (press Ctrl-C to break) to stop accepting more slaves.

8) Append the cluster.config.file to /etc/hosts, because Hadoop requires DNS names for slaves:

```
> sudo cp -n /etc/hosts /etc/hosts.org
> sudo -- sh -c -e  "cat cluster-config.file >> /etc/hosts"
```

9) Update the Hadoop slaves list with the DNS names from the cluster-config.file, but remove the internet address information:

```
> cat cluster-config.file | cut -f 4 -d " " >
 ➥/home/ubuntu/hadoop/etc/hadoop/slaves
```

10) Add the slaves to ssh/known.hosts to avoid SSH warnings:

```
> ssh-keyscan  -H  -f  /home/ubuntu/hadoop/etc/hadoop/slaves  >>
 ➥~/.ssh/known_hosts
```

11) Copy /etc/hosts to all slaves to allow them to communicate using DNS names:

```
> wget https://raw.githubusercontent.com/Rustor/EE-DIGI/
➥master/update-all-slaves.sh
> bash update-all-slaves.sh
```

In this section, you have now, on AWS, created a Linux Hadoop cluster based on a single master installation. This installation has all the necessary software tools and configuration so that you can dive in to using and testing the cluster—see the next section.

DATA ANALYSIS METHOD

We wrote an R program to find correlations between data logs. To develop this program, we needed test data. Therefore, we have generated log files from random data. Each log file contains a number of variable data or signals. The R program (demonstrator) requires a format where each data record has a device ID and is stored on one line with time-date:

```
Time-date, device_id, var1, var2, var3, var4
```

However, we have multiplied one of the signals in one of the log files with a sinus curve. Without knowing which log, we have designed the R program to use a correlation function to find it. Correlation (cor) is a mathematical function that given two signals will output a number between 1 and –1. Zero (0) means that the two signals are unrelated or not detectable by the algorithm. A value above 0.75 or below –0.75 is considered a significant correlation.

An R program using the cor function was able to find the data log that was multiplied with a sinus curve. It is difficult to tell which one

YOU CAN SCALE UP THE EXAMPLE, IF YOU ARE WILLING TO PAY FOR MACHINES WITH MORE RAM.

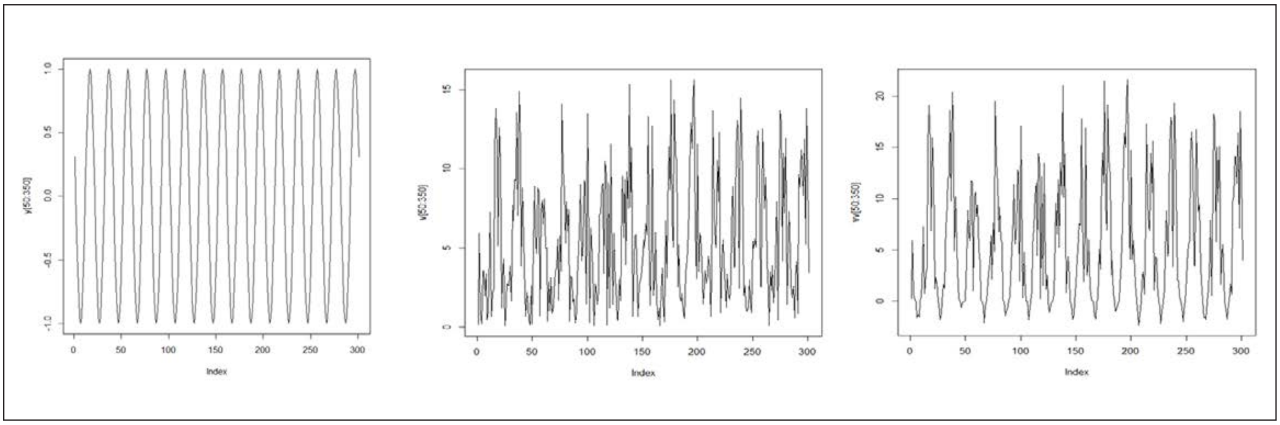


Figure 8. Illustrations of a couple device log files. On the left is the sinus signal. On the right of this are two log file signals: V (middle) and VV (right).

of the signals matches just by inspection. Correlation of the signals in Figure 8 result in the following output:

```
cor(v,y)
[1] 0.5779192

cor(vv,y)
[1] 0.7557141
```

As you can see, the signal VV has a correlation of 0.7557141 with y. We have found the log/signal that was multiplied with the sinus curve.

TESTING THE CLUSTER BY ANALYZING DATA

Based upon what we have presented in the previous section, we made a small example (25MB) with ten "spreadsheet files", each with 4,000 lines and 9 columns of data. We have scaled down the test that we present in this article in order to be able to run in the Amazon free tier on three computers: one master and two slaves. You can scale up the example, if you are willing to pay for machines with more RAM. The R analysis code does not require changes.

Enter the Hadoop directory, format the HDFS and start the whole system (Hadoop dæmons on the different machines). The following will initialize all nodes:

```
> cd hadoop
> bin/hdfs namenode  -format
> sbin/start-all.sh
```

Then you have to generate the test data. You should download and run the script found in the EEdigi repo on GitHub:

```
> wget https://raw.githubusercontent.com/Rustor/EE-DIGI/
➥master/genEEdigi-test-data.R
> Rscript genEEdigi-test-data.R
```

Next, put the test data into HDFS, and for that, you make a directory structure (as required by the analysis program):

```
> bin/hadoop  fs  -mkdir   /rhadoop/eedigi/xdata
> bin/hadoop  fs  -put  -f  TESTdata*.csv  /rhadoop/eedigi/xdata
```

Now the cluster should be running with data loaded. You need to download the EEdigi analysis program and run it after you first have set up the Rmr2 environment in an R session. The important parts of the R code are shown in Listing 1.

Get the EEdigi data analysis program:

```
> wget https://raw.githubusercontent.com/Rustor/EE-DIGI/
➥master/EEdigitest.R
```

Run the program in the R command prompt:

```
> R
> source("etc/hadoop/hset.R")
> source("EEdigitest.R")
> q()
```

Listing 1. The "mapper" takes the input files and selects key/value before emitting. "Reducer" analyses data with the "cor" function (correlation) and emits the result as value. Before printing, the R analysis program selects the data with high (more than 0.75) significance.

```
Mapper = function(key, val.df) {
        val.df = subset(val.df, id != 'id'  )
        output.key = data.frame(id=val.df$id, stringsAsFactors=F)
        output.val = val.df[,c('t1', 't2', 't3', 't4')]
        return( keyval(output.key, output.val))}

...

reducer = function(key, val.df) {
        output.key = key
   output.val = c(cor(as.numeric(val.df$t2),as.numeric(val.df$t4)),
                  cor(as.numeric(val.df$t1),as.numeric(val.df$t2)))
        return( keyval(output.key, output.val))}

...

print( results.df[which(results.df$val> 0.75),])
```

CLOSE DOWN THE CLUSTER

Use the following instructions if you want to close down the cluster:

```
> sbin/stop-all.sh
```

1) Shut down all machines in the AWS web interface.

2) Terminate the slaves (keep the master instance and image) in the interface.

CONCLUSION

After reading this article, you should be able to set up a Hadoop Linux cluster in a short amount of time. We have also provided a way for you to test the cluster using R. The open-source tool chain fits nicely together, and you'll be able to learn about Big Data analysis at no cost.

With the current cloud price structures, our recommendation is to use a cloud cluster when you have a small budget (for computing), and your need for it is transient. If at some point you are using many machines constantly for an entire month, you should consider building a local computer cluster. While doing the project, we learned that buying 50 strong computers as hardware would have cost a similar amount as renting 50 strong computers for one month.

When using a free tier, remember that the cloud will cost you money over time. It is the cloud providers' business to charge for using their computing resources. Cloud providers have extensive price information, and they will charge you for every use of a virtual server—for example, storage of hard disk images and machine templates, data out of the data center (region). Even when you use the free tier, read your bill carefully for unexpected costs due to a tiny mistake on your side—for example, using more storage than included in the free tier. The best you can do is to make a small budget for cloud usage in order to prepare you and your organization's mindset. Then you should run trials based on your typical analysis tasks to determine what type of machines and which provider is most suited.

Do not use the intuitively impressive-sounding Software as a Service (SaaS), but stick to generic virtual machines, as we used in this article. Otherwise, you most likely will be investing in learning a proprietary SaaS interface and will lose the ability to switch cloud providers without high switching costs. Managing Linux machines (servers) for short periods should not be a problem in terms of operational costs. In our experience, in-depth knowledge about Linux and computer hardware (performance) is highly relevant when using cloud computer resources. After all, you are given complete control of a real CPU and RAM, if only for short period.

ACKNOWLEDGEMENTS

This work was supported by The Southern Denmark Growth Forum and the regional EU project "IoTstyring", which is a project about energy efficiency in embedded control systems using Big Data and IoT technologies.■

Rune Torbensen is Postdoc at University of Southern Denmark, SDU Mechatronics, Sønderborg, Denmark and has an IoT PhD with a focus on wireless embedded communication from Aalborg University. He has used embedded Linux in most of his experiments during the past ten years. Recently, he became interested in Big Data technology due to its strong relation with IoT.

Søren Top, associate Professor PhD, is a lecturer at University of Southern Denmark, SDU Mechatronics, Sønderborg, Denmark. He has taught operating systems and embedded systems for decades, and he uses Linux for both topics on a daily basis.

RESOURCES

Hadoop: http://hadoop.apache.org

Hadoop: The Definitive Guide by Tom White.

Rmr2: https://github.com/RevolutionAnalytics/RHadoop/wiki

Guide to creating virtual machines with the AWS interface:
https://www.youtube.com/watch?v=Ix5IDuyamuY

Send comments or feedback via
http://www.linuxjournal.com/contact
or to ljeditor@linuxjournal.com.

RETURN TO CONTENTS

The Best SharePoint 2016 and Office 365 Training!

SPTechCon
The SharePoint Technology Conference

April 2-5, 2017 • AUSTIN, TEXAS

SPTechCon offers classes and tutorials for IT professionals, business decision makers, information workers, developers and software and information architects. Each presenter at SPTechCon is a true SharePoint expert, with many drawn from Microsoft's tech teams or holding Microsoft MVP status.

Whether you're looking to upgrade to a more current version, making a move to the cloud, or simply need answers to those daunting problems you've been unable to overcome, SPTechCon is the place for you! Come join us!

A BZ Media Event

- **Choose from more than 80 classes and panel sessions**
- **Improve your skills and broaden your knowledge of Microsoft's collaboration and productivity software**
- **Learn about SharePoint 2016, the latest on-premises server release from Microsoft**
- **Tips and tricks for working with SharePoint 2013 and 2010, and Office 365**
- **Practical information you can put to use on the job right away!**
- **The most knowledgeable instructors working in SharePoint today**

www.sptechcon.com

Integrating Web Applications **with Apache**

Learn how to write your own custom Apache configurations to make your applications work the way you want.

ANDY CARLSON

◀ PREVIOUS
Feature:
Big Data Demonstrator

NEXT
Doc Searls' EOF ▶

When you deploy a web application, how do end users access it? Often web applications are set behind a gateway device through which end users can access it. One of the popular products to act as an application gateway on Linux is the Apache Web Server. Although it can function as a normal web server, it also has the ability to connect through it to other web servers.

In this article, I discuss what it takes to integrate a web application into Apache. This includes integrating the HTTP protocol functionality, customizing content to render properly and reusing pieces of configuration. Once you understand those basic bits of functionality, you'll have the tools you need to maximize your web applications' usability. So, let's get started!

Crash Course in RegEx

A mechanism that I use throughout this article that might need a brief introduction is Regular Expressions (or regex). Regex is used to define a text pattern to search for within a URL or to find and replace text within content, such as HTML or JavaScript. The text processing command `sed` uses regex to do searches and substitutions.

For each example below there will be three parts: input, regex pattern and output. The pattern will be applied to the input text and determine the value of the output text.

Example 1:

```
Input:
  Name: Frank Sinatra
  Genre: Jazz
  Name: 2Pac
  Genre: Rap
  Name: Reel Big Fish
  Genre: Ska

Regex pattern: "^Name: "

Output:
  Name: Frank Sinatra
  Name: 2Pac
  Name: Reel Big Fish
```

This example searches the input text for text that matches the pattern `"^Name: "`. This pattern says, "Look for the text 'Name: ' at the beginning of each line." Since there are two lines that begin with that text, only those two lines are returned. While "^" represents the beginning of a line, "$" represents the end of a line. So if you were to apply the pattern "a$", two lines would be returned (Frank Sinatra and Ska). Let's expand on that example and use the input from Example 1 with a new pattern.

Example 2:

```
Regex pattern: "^Name: [0-9]"

Output:
  Name: 2Pac
```

As you can see, I've taken the original regex pattern and added `[0-9]` to the end. This will search for a single character that can be any number from 0 to 9, which is why "2Pac" was the only line returned. You also can specify a range with alphabetic characters (`[a-z]` or `[A-Z]`).

Along with pattern selection, you also can do substitution with regex. There are two formats for regex substitutions: s|pattern|replace|modifier or s/pattern/replace/modifier. In Apache, I find it easier to use the pipe-style substitution. Example 3 uses the same input with a new pattern.

Example 3:

```
Regex pattern: "s|^(.*)Frank(.*)$|\1Dwezil\2|g"

Output:
  Name: Dwezil Sinatra
  Genre: Jazz
  Name: 2Pac
  Genre: Rap
  Name: Reel Big Fish
  Genre: Ska
  Name: Dwezil Zappa
  Genre: Unknown
```

This pattern has a lot to dissect. One of the great features of regex is the ability to match any character. The dot operator will match any one character. The asterisk operator will match 0 or more of whatever character or operator preceded it. Putting these two operators together matches 0 or more of any character. Enclosing this in parentheses allows the matched text to be represented in the replace portion of the pattern with a variable. In this case, `\1` represents the first block of text within parentheses and `\2` represents the second. The only characters that are explicitly being matched are "Frank". As such, the lines containing "Frank" will be replaced with everything up to "Frank" (represented by `\1`), "Dwezil", and everything following "Frank" (represented by `\2`). As you can see, the entirety of the text input was sent to the output although modified by the pattern.

Protocol Integration

When it is decided that an application would benefit from Apache integration, there is a high likelihood that it will reside on a separate server from Apache. To integrate applications being accessed via HTTP fully, any or all of these modules may be used: `mod_rewrite`, `mod_proxy`, `mod_ssl` and `mod_headers`. Each of these modules allows you to customize the way communication between the end user and web servers occurs from modifying HTTP header data to managing proxy connections to other servers.

First, let's look at `mod_rewrite`. There are a number of directives within the `mod_rewrite` module, but I cover only a handful here: `RewriteEngine`, `RewriteCond` and `RewriteRule`. The `RewriteEngine` directive simply enables URL rewriting and is invoked as follows:

```
RewriteEngine on
```

`RewriteRule` allows the server to respond to an HTTP request to a specific URL by, among other things, returning an HTTP redirect (code 301 or 302), which will redirect the end user to a specified URL or send a proxied request to a back-end server. Here's an example of issuing an HTTP redirect:

```
RewriteRule /google http://www.google.com [R=301]
```

In this example, when the URL of /google is accessed, the server will respond with an HTTP 301 that will redirect the user to http://www.google.com. This example will work only if the request URL is exactly equal to "/google". If the need is to redirect on any URL starting with "/google", you would define a conditional redirect using `RewriteCond` as follows:

```
RewriteCond "%{REQUEST_URI}" "/google.*$"
RewriteRule "^.*$" http://www.google.com [R=301]
```

The `RewriteCond` directive has two parts: a string value to check and a substring to search for. In this example, you are looking in the `REQUEST_URI` HTTP session variable for anything beginning with "/google". If that condition is met, the `RewriteRule` on the following line is executed. Because you are determining the value of the target URL in the `RewriteCond`, the value of the target URL in the `RewriteRule` is defined as `"^.*$"`.

The examples given here are all user-facing events like a 301 redirect. The `RewriteRule` directive also can be used to proxy requests to a server. This is done behind the scenes unlike an HTTP redirect, so the request is forwarded without the users' knowledge. A proxied request may be configured like the example below:

```
RewriteRule "/home/(.*)$" http://back-end01.test:8080/$1 [P]
```

The above illustrates an example of a virtual root directory. When the user accesses anything underneath /home (note the ".*" expression), the request is sent to back-end01.test on port 8080 with the location set to the URL path beneath /home. For example, if the user tries to access /home/test/image.jpg, the request is sent to back-end01.test:8080 with a location of /test/image.jpg. A proxied `RewriteRule` also may be used in conjunction with `RewriteCond` for further customization. Note that this statement proxies only the HTTP request. Proxying of HTTP responses will require `mod_proxy`.

Another option for proxying HTTP connections through Apache is `mod_proxy`, which provides `ProxyPass`, `ProxyPassReverse` and `ProxyPassMatch` among many other directives that provide more robust proxying options. I focus primarily on these three directives here. As mentioned previously, RewriteRule provides proxying of HTTP requests. Let's compare the example already given for proxying with RewriteRule and an example for ProxyPass:

```
ProxyPass /home http://back-end01.test:8080/
```

This ProxyPass statement provides roughly the same level of functionality as the RewriteRule statement with a more simplistic command. When a request comes in for any URL beginning with "/home", the request header will be rewritten so that the request will be received properly by http://back-end01.test:8080/. Consider the following first lines of an HTTP request:

```
From user to server:    GET /home/test/image.jpg HTTP/1.1
From server to back-end:    GET /test/image.jpg HTTP/1.1
```

The first line of the header contains the method (GET in this case) and the URL being requested. When the server receives the request from the client, it strips off "/home", as specified in the ProxyPass directive and forwards the request to the back-end server. If you want to proxy response packets as well as request packets, the following ProxyPassReverse statement can be paired with the previous ProxyPass statement:

```
ProxyPassReverse /home http://back-end01.test:8080/
```

The syntax is exactly the same as ProxyPass, adding to the simplicity of the mod_proxy configuration. This will take any HTTP response matching an HTTP request for /home and forward the response back to the original client. If you need to add some programmatic proxying (similar to RewriteCond), you can use the ProxyPassMatch. When implementing a forward/reverse proxy configuration, ProxyPassMatch can replace ProxyPass. Here's an example:

```
ProxyPassMatch "^/home/([a-z0-9]*/docs)" http://docserver01.test:8080/$1
ProxyPassReverse /home http://docserver01.test:8080/
```

Along with rewriting URLs, it may be necessary to rewrite HTTP request or response header fields.

This example suggests that within the /home folder, there are many sub-folders (let's say user names) and within each of those exists a folder named "docs". The USERNAME/docs URL exists on docserver01.test:8080 in the root of the web server, as denoted by the $1 in the server URL. The `ProxyPassReverse` will function in the same manner as it did in the previous example.

Securing websites with SSL in Apache is accomplished with `mod_ssl`. Although I won't discuss configuring SSL from the ground up, a few directives relate to proxied SSL connections: `SSLProxyCheckPeerExpire`, `SSLProxyCheckPeerName` and `SSLProxyCheckPeerCN`. It is a common practice to use self-signed certificates on back-end servers (provided a valid cert is in place on the user-facing server), and these directives address common issues that can arise when using self-signed certs. Any of these directives can have one of two arguments provided: "on" or "off". If set to "off", `SSLProxyCheckPeerExpire` will skip checking the expiration date on the SSL cert used on a back-end server. To avoid checking a certificate's common name or alternate names against the server name used to access a back end, set `SSLProxyCheckPeerName` to "off". In older versions of Apache, you might be able to use `SSLProxyCheckPeerCN` (set to "off") instead of `SSLProxyCheckPeerName`.

Along with rewriting URLs, it may be necessary to rewrite HTTP request or response header fields. In Apache, this is done with `mod_headers`. There are only two directives within this module: `Header` and `RequestHeader`. These directives are used to modify response and request header fields, respectively. Many actions can be used with either of these directives, but here, let's look at the `set` and `edit` actions—for example:

```
Header set ReceiveTime "%t"
```

This example will add and replace any existing header in an HTTP response named ReceiveTime and give it the value of the UNIX timestamp when the request was received by the server (represented by "%t").

If you need to replace the value of a header that comes from a back-end server, you would use the edit action. Consider the following example:

```
Header edit Location "^http://back-end01.test:8080/(.*)$"
➥"http://public.test/$1"
```

This example will replace the Location attribute in an HTTP response, which will exist in a 301/302 redirect. If it finds http://back-end01.test:8080 at the beginning of the Location header, it replaces that part with "http://public.test" (the user-facing URL).

Content Integration

Once a remote application is integrated with an Apache server, from a protocol standpoint, it may be necessary to integrate content. This will generally manifest itself as URLs coded into HTML or JavaScript that are specific to a back-end server and not to a user-facing server. The basic necessity is to be able to search and replace bits of HTML or JavaScript content, so that it can render and perform correctly when accessed through an Apache proxy. The module that accomplishes this is mod_substitute and specifically the Substitute directive. Substitute allows a simple regex substitute to be performed on the payload data of an HTTP response.

Something to consider before attempting to replace text is to account for whether the back-end web server compresses data before sending it over the network. If it does, your Substitute statements might not work, as it will be searching for ASCII text within binary compressed data. To account for this, you can instruct Apache to decompress the data, manipulate the response and then re-compress it. This is done using the SetOutputFilter directive, which is part of Apache core functionality. Here's how it works:

```
SetOutputFilter INFLATE;SUBSTITUTE;DEFLATE
```

Reading the arguments from left to right, this tells Apache to `INFLATE` (decompress) the data from the back-end server, perform the substitute and `DEFLATE` (compress) the data before returning it to the end user.

The `Substitute` statement uses a regex substitute expression. As I mentioned previously, I found it easier to use the pipe-style substitute expression in Apache. To recap, the syntax is s|search|replace|options. Two common options that I tend to use: "i", which denotes a case-insensitive search, and "n", to allow the search and replace values to be processed as regex. Here's a common use example:

```
Substitute "s|(href="http)(://)back-end01.test:8080|$1s$2public.test|in"
```

For this example, let's assume that the user-facing site (public.test) runs HTTPS, and the back-end server (back-end01.test) runs HTTP on port 8080. This would be a solution if the back-end web server returned hyperlinks that were specific to itself as opposed to the user-facing site. In the search portion of the regex substitute, this splits out two groups of text in parentheses: `(href=\"http)` and `(://)`. These are blocks of text that you want preserved in the replace section of the regex. In the replace, you are inserting an "s" after http and replacing the hostname/port with the user-facing site name. After processing, the resulting string will be `href="https://public.test`. This will update hyperlinks that use "href" attributes (<a> and <link>). For <img> and <script> tags, you could use this same Substitute statement and replace "href" with "src". Another consideration would be to account for double or single quotes delimiting attribute values (`href='` vs. `href="`).

Another application of `Substitute` is to extend the functionality of a page without manipulating the original source code. Consider the following example:

```
Substitute "s|(<body.*>)|\1<div style=\"font-size:14pt;
➥font-weight:bold;background-color:#ff0000;color:
➥#ffffff;display:block;text-align:center;\">This site
 ➥will be down for 24 hours beginning at 8 pm tonight</div>|in"
```

If a website needs to be taken off-line for maintenance, this is an easy way to alert the user population of the outage without modifying the application itself.

If a website needs to be taken off-line for maintenance, this is an easy way to alert the user population of the outage without modifying the application itself. This example simply inserts a red bar along the top of the page (right after the <body> tag), which displays information about the outage. Depending on how your page is rendered, you might need to choose another tag to act as your starting point instead of <body>.

Streamlining Future Integrations

All of the topics presented here can be configured and maintained relatively easily if you have only a few statements. In the real world, there typically will be many sites that use a similar configuration and having to define the functionality for each site can be time-consuming and can lead to mistakes. Luckily, Apache provides a mechanism to repeat functionality throughout your configuration through the use of `mod_macro`. The `<Macro>` directive within an Apache config functions very much like a function or subroutine. Once a macro is defined, it can be referenced as many times as is necessary, leaving you with one place within your config to maintain your detailed functionality. Here's an example macro:

```
<Macro RedirectSecure $host $path>
        RewriteCond "%{REQUEST_URI}" "^$path"
        RewriteRule "^/(.*)$" "https://$host/$1"
</Macro>
```

When called, this macro will define a `RewriteCond` and `RewriteRule` that, if they access a URL starting with the value of the $path argument, will redirect the user to http://$host/$1, where $host is the hostname specified as a macro argument and $1 is the entire URL path. The following syntax would be used to call this macro:

```
Use RedirectSecure public.test /users
```

Something to consider is the location within the Apache config from which a macro is called. A `RewriteRule`, for example, cannot be called outside a `<VirtualHost>` block. As such, if the macro is called outside a `<VirtualHost>` block, Apache will throw an error and not start. Here's another example:

```
<Macro ReplaceContentURL $backendurl $publicurl>
        Substitute "s|(href=\")$backendurl|$1$publicurl|in"
        Substitute "s|(src=\")$backendurl|$1$publicurl|in"
</Macro>
```

This macro expands on the replacing of URLs that I covered previously. This will search for tag attributes of "href" and "src" and replace the hyperlinks of the back-end server with that of the user-facing server. Here's an example of how this might be called:

```
Use ReplaceContentURL http://back-end01.test:8080 https://public.test
```

This will search for http://back-end01.test:8080, beginning with either `href="` or `src="` and replace the URL with https://public.test. Macros can be used for any piece of Apache configuration. They can be used to do small tasks as shown here as well as whole site configurations. Although macros are pretty simple, they make the difference between a large amount of difficult-to-maintain configuration files and a simplified reusable configuration.

At this point, you have some basic knowledge of integrating HTTP, customizing content and reproducing configuration within Apache. Although many directives and modules weren't covered here, this will be a great starting point and can help you get started with accessing your applications through Apache. ■

Andy Carlson has worked in IT for the past 13 years doing networking and server administration. He is thankful to have chosen a career that he loves, grows in and learns from. He and his amazing wife have three daughters and a son, and they currently reside in Cincinnati, Ohio. He enjoys playing the guitar and spending time with family and friends.

RESOURCES

The following are some articles I've found useful along with some example Apache configs I've written.

Apache Module Reference (2.2): httpd.apache.org/docs/2.2/mod

Apache Module Reference (2.4): httpd.apache.org/docs/2.4/mod

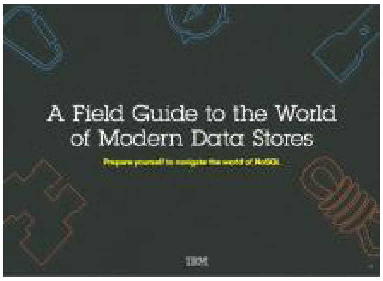
Git Instaweb Reverse Proxy: https://gist.github.com/bng44270/cff67619db3e3f915957

Monit Reverse Proxy: https://gist.github.com/bng44270/287277ea1975b9a3e3526d5a5bcb017c

Adobe Experience Manager Apache Config: https://github.com/bng44270/aem-dispatcher-config

Send comments or feedback via http://www.linuxjournal.com/contact or to ljeditor@linuxjournal.com.

RETURN TO CONTENTS



A Field Guide to the World of Modern Data Stores

Prepare yourself to navigate the world of NoSQL

IBM

A Field Guide to the World of Modern Data Stores

There are many types of databases and data analysis tools to choose from when building your application. Should you use a relational database? How about a key-value store? Maybe a document database? Is a graph database the right fit? What about polyglot persistence and the need for advanced analytics?

If you feel a bit overwhelmed, don't worry. This guide lays out the various database options and analytic solutions available to meet your app's unique needs.

You'll see how data can move across databases and development languages, so you can work in your favorite environment without the friction and productivity loss of the past.

Sponsor: IBM

> **https://geekguide.linuxjournal.com/content/field-guide-world-modern-data-stores**

Why NoSQL?

IBM

Why NoSQL? Your database options in the new non-relational world

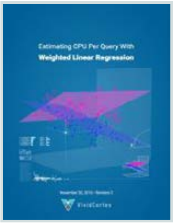
The continual increase in web, mobile and IoT applications, alongside emerging trends shifting online consumer behavior and new classes of data, is causing developers to reevaluate how their data is stored and managed. Today's applications require a database that is capable of providing a scalable, flexible solution to efficiently and safely manage the massive flow of data to and from a global user base.

Developers and IT alike are finding it difficult, and sometimes even impossible, to quickly incorporate all of this data into the relational model while dynamically scaling to maintain the performance levels users demand. This is causing many to look at NoSQL databases for the flexibility they offer, and is a big reason why the global NoSQL market is forecasted to nearly double and reach USD3.4 billion in 2020.

Sponsor: IBM

> **https://geekguide.linuxjournal.com/content/why-nosql-your-database-options-new-non-relational-world**

Estimating CPU Per Query With Weighted Linear Regression



Estimating CPU Per Query With Weighted Linear Regression

VividCortex

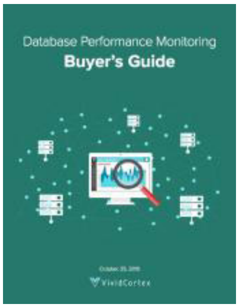
Your database server is suddenly using a lot of CPU resources. Quick, what caused it? This is a familiar question for engineers of all persuasions. And it's often impossible to answer.

There are good reasons why it's hard to figure out what consumes resources like CPU, IO, and memory in a complex piece of software such as a database. The first problem is that most database server software doesn't offer any way to measure or inspect that type of performance data. The database server isn't observable. This problem arises in turn from the complexity of the database server software and the way it does its work, which actually precludes measuring resource consumption accurately!

Author: Baron Schwartz

Sponsor: VividCortex

> **https://geekguide.linuxjournal.com/content/estimating-cpu-query-weighted-linear-regression**



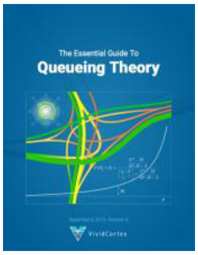
Database Performance Monitoring Buyer's Guide

More and more companies have begun to recognize database performance management as a vital need. Despite its widespread importance, good database performance management requires specialized expertise with custom approaches--yet all too often, organizations rely on one-size-fits-all solutions that theoretically "check the box" but in practice do little or nothing to help them find or prevent database-related outages and performance problems.

This buyer's guide is designed to help you understand what database management really requires, so your investments in a solution provide the greatest possible ultimate value.

Sponsor: VividCortex

> **https://geekguide.linuxjournal.com/content/database-performance-monitoring-buyer%E2%80%99s-guide**



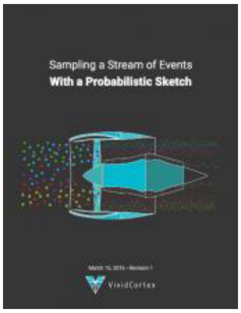
The Essential Guide To Queueing Theory

Whether you're an entrepreneur, engineer, or manager, learning about queueing theory is a great way to be more effective. Queueing theory is fundamental to getting good return on your efforts. That's because the results your systems and teams produce are heavily influenced by how much waiting takes place, and waiting is waste. Minimizing this waste is extremely important. It's one of the biggest levers you will find for improving the cost and performance of your teams and systems.

Author: Baron Schwartz

Sponsor: VividCortex

> **https://geekguide.linuxjournal.com/content/essential-guide-queueing-theory**



Sampling a Stream of Events With a Probabilistic Sketch

Stream processing is a hot topic today. As modern Big Data processing systems have evolved, stream processing has become recognized as a first-class citizen in the toolbox. That's because when you take away the how of Big Data and look at the underlying goals and end results, deriving real-time insights from huge, high-velocity, high-variety streams of data is a fundamental, core use case. This explains the explosive popularity of systems such as Apache Kafka, Apache Spark, Apache Samza, Apache Storm, and Apache Apex—to name just a few!

Author: Baron Schwartz

Sponsor: VividCortex

> **https://geekguide.linuxjournal.com/content/sampling-stream-events-probabilistic-sketch**

The Problem with "Content"

Real journalism is getting programmatically corrupted and harder to find. Fortunately, there's a fix.

DOC SEARLS

Doc Searls is Senior Editor of *Linux Journal*. He is also a fellow with the Berkman Center for Internet and Society at Harvard University and the Center for Information Technology and Society at UC Santa Barbara.

PREVIOUS
Feature: Integrating Web Applications with Apache

Back in the early '00s, John Perry Barlow (https://en.wikipedia.org/wiki/John_Perry_Barlow) said "I didn't start hearing about 'content' until the container business felt threatened." *Linux Journal* was one of those containers—so was every other magazine, newspaper and broadcast station. Today, those containers are bobbing around in an ocean of "content" on the internet. Worse, the stuff inside the containers, which we used to call "editorial", is now a breed of "content" too.

In the old days, editorial lived on one side of a "Chinese wall" between itself and the publishing side of a newspaper or magazine. The same went for the programming and advertising sides of a commercial broadcast station or network. The wall was transparent, meaning it was possible for a writer, a photographer, a newscaster or a performing artist to see what funded the operation, but the ethical thing was to ignore what happened on the other side of that wall. Which was easy to do, because everything on the other side of that wall was somebody else's job.

Today that wall has been destroyed by the imperatives of "content production", which is the new job of journalists and everybody else devoted to "generating content" in maximum volumes, all the better to attract "programmatic" advertising.

You can see the wreckage of one such wall in a January 2017 *The New York Times* story titled "In New Jersey, Only a Few Media Watchdogs Are Left" (https://www.nytimes.com/2017/01/03/nyregion/in-new-jersey-only-a-few-media-watchdogs-are-left.html?_r=2), by David Chen. In it he writes, "The *Star-Ledger*, which almost halved its newsroom eight years ago, has mutated into a digital media company requiring most reporters to reach an ever-increasing quota of page views as part of their compensation."

As I explained in my January 2016 article "What We Can Do with Ad Blocking's Leverage" (http://www.linuxjournal.com/content/what-we-can-do-ad-blockings-leverage), the advertising we're talking about here isn't the old Madison Avenue kind that lived on the other side of journalism's Chinese wall. It's a new all-digital kind called *adtech*. While adtech is *called* advertising and *looks* like advertising, it is actually a breed of direct marketing (https://en.wikipedia.org/wiki/Direct_marketing), a cousin of spam descended from junk mail.

Like junk mail, adtech is data-driven, wants to get personal, finds success in tiny-percentage responses and excuses massive negative externalities. Those include wanton and unwelcome surveillance, annoying the crap out of people and filling the world with crap—including fake news and fraudulent advertising.

Here's one way to tell the difference between real advertising and adtech, using the *Star-Ledger* as an example:

- Real advertising wants to be in the *Star-Ledger* because it values the paper's journalism and readership.

- Adtech wants to push ads at readers anywhere it can find them, based on gathered intelligence, algorithms and whatever else shows up in live auction markets for eyeballs.

In the old advertising-supported publishing world, *journalism* was what mattered most. In the new adtech-supported publishing world, *content* is what matters most.

Real advertisers in the old publishing world were flattered to be in the *Star-Ledger*. Adtech-oriented advertisers in the new publishing world just want to "go digital", whatever it takes. And there are thousands of intermediaries (http://cdn.chiefmartec.com/wp-content/uploads/2016/03/marketing_technology_landscape_2016.jpg) to help with that.

As I wrote in "Separating advertising's wheat and chaff" (https://medium.com/@dsearls/separating-advertisings-wheat-and-chaff-47858adfcb20#.pgujy36im), it is because of that orientation, and those intermediaries, that "Madison Avenue fell asleep, direct response marketing ate its brain, and it woke up as an alien replica of itself."

That's also why, to operate in publishing's new body-snatched economy, journalists are incentivized to meet that "ever-increasing quota of page views". When the incentives are volume-based. what happens to quality?

It is essential to note that adtech, by design, doesn't care about journalism at all. That's because adtech values a maximized sum of content in the world, regardless of how good that content is or where it comes from. The more content, the more places ads can be run.

It is also ridiculously easy to make adtech money with content, especially since there is nothing about content as a substance that requires facts to back it up. This is why, according to *Buzzfeed* (https://www.buzzfeed.com/craigsilverman/how-macedonia-became-a-global-hub-for-pro-trump-misinfo?utm_term=.gcAwkw26PA#.gs5dxdRJEz), teenagers in one town in Macedonia made as much as $3,000 a day by generating fake news (such as "Pope endorses Trump") during the 2016 US presidential election.

In my January 2017 EOF "Debugging Democracy" (http://www.linuxjournal.com/content/debugging-democracy), I made the mistake of opening with negative remarks about the winner of that election, which distracted readers from my main point, which was that journalism is corrupted, marginalized and suffering in a world where a business based on surveillance stokes people's prejudices and drives them into mutually hostile echo chambers (http://graphics.wsj.com/blue-feed-red-feed), damaging every democracy that depends on having at least some common ground on which agreement, or at least compromise, can be found. And I thought the topic was a good one for *Linux Journal* because readers such as ours are in a good position to help fix it.

I also think publishing needs to re-brain Madison Avenue and its employers who are drunk on digital (https://medium.com/@dsearls/tv-viewers-to-madison-avenue-please-quit-driving-drunk-on-adtech-6657728ab597#.klgyhav24) and demand real advertising by real advertisers who want to sponsor real journalism.

To support that effort, I will now cross our own Chinese wall here at *Linux Journal* and thank the advertisers listed each month below my column. Those advertisers are brands in the best sense of the word. I hope we attract more like them to *Linux Journal* with this simple fact: even though countless $billions (or perhaps $trillions by now) have been spent on adtech, not one brand has been built by it.

Bonus link: Don Marti's "Targeting failure: legit sites lose, intermediaries win": (http://zgp.org/targeted-advertising-considered-harmful/#targeting-failure-legit-sites-lose-intermediarieswin).■

Send comments or feedback via
http://www.linuxjournal.com/contact
or to ljeditor@linuxjournal.com.

RETURN TO CONTENTS

ADVERTISER INDEX

Thank you as always for supporting our advertisers by buying their products!

ADVERTISER	URL	PAGE #
DrupalCon Baltimore	https://events.drupal.org/baltimore2017	35
Drupalize.me	http://drupalize.me	114
HPC Wallstreet	http://www.flaggmgmt.com/linux	21
Libre Planet 2017	http://libreplanet.org/conference	47
LinuxFest Northwest	http://linuxfestnorthwest.org	79
Peer 1 Hosting	http://go.peer1.com/linux	115
SCALE 15x	http://www.socallinuxexpo.org	77
Silicon Mechanics	http://www.siliconmechanics.com	23
SPTechCon	http://www.sptechcon.com	95
SUSE	http://suse.com/storage	7

ATTENTION ADVERTISERS

The *Linux Journal* brand's following has grown to a monthly readership nearly one million strong. Encompassing the magazine, Web site, newsletters and much more, *Linux Journal* offers the ideal content environment to help you reach your marketing objectives. For more information, please visit http://www.linuxjournal.com/advertising

```
 *
 * Usage: $(obj).insertAtCursor(text);
 *   obj = a textarea or textfield
 *   text = a string to insert
 */
$.fn.extend({
  insertAtCaret: function(myValue){
    return $(this).each(function(){
      // If target element is hidden, don't do it.
      if ($(this).is(':hidden') || $(this).parents(':hidde
        return;
      }
      if (document.selection) {
        this.fo
        sel = d
        sel.tex
        this.fo
      }
      else if (
        var sta
        var end
        var scr
        this.va
        this.fo
        // High
```

● macromaker.module
× macromaker.js
× jquery.insertAtCaret.js
drupalize.me
home
pricing & signup
Learn Drupal & jQuery
The Super Convenient Way
OVER 60 VIDEOS & 87 HOURS OF DRUPAL TRA
Search
Views
Configuring Views relationships
Introduction to Advanced Theming for Drupal
Site building with Drupal Using Views
Overview of Views