

# Launching Your Career in Big Data

Sujee Maniyam

ElephantScale.com

[Sujee@elephantscale.com](mailto:Sujee@elephantscale.com)



# Who Invited This Guy?

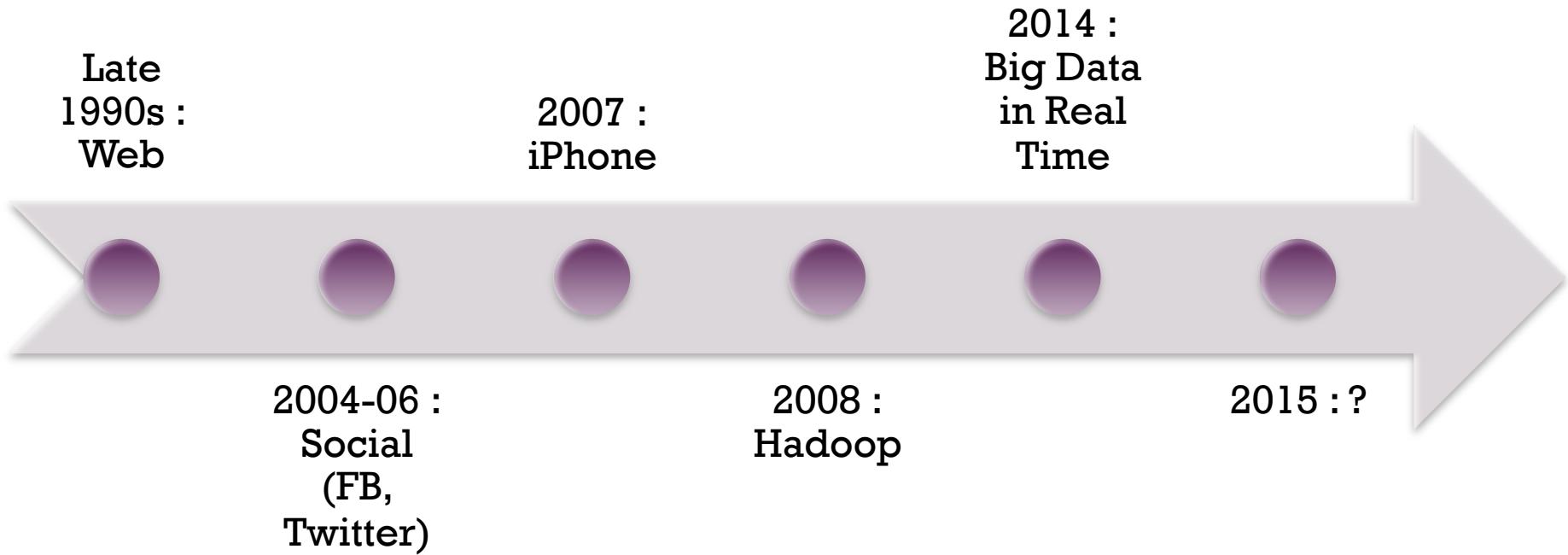
Hi, I am Sujee Maniyam 😊



- 15 years+ software development experience
- Consulting & Training in Big Data
- Author
  - “Hadoop illuminated” open source book
  - “HBase Design Patterns” coming soon
- Open Source contributor (including Hadoop)  
<http://github.com/sujee>
- Founder / Organizer of ‘**Big Data Guru**’ meetup  
<http://www.meetup.com/BigDataGurus/>
- <http://sujee.net/>



# Riding The Technology Wave





# Big Data Fad Or Real?

- It is very real !

money can't  
make you happy

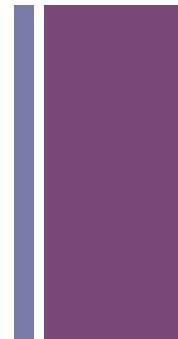


unless you roll  
around in it.





# Big Data Job Trend



## Hadoop Job Trends



This job trends graph shows the percentage of jobs we find that contain your search terms.

▶ [Email to a friend](#)

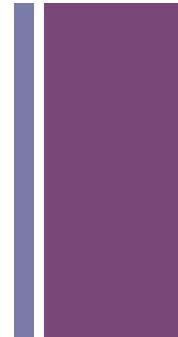
▶ [Post on your blog/website](#)

### Top Job Trends

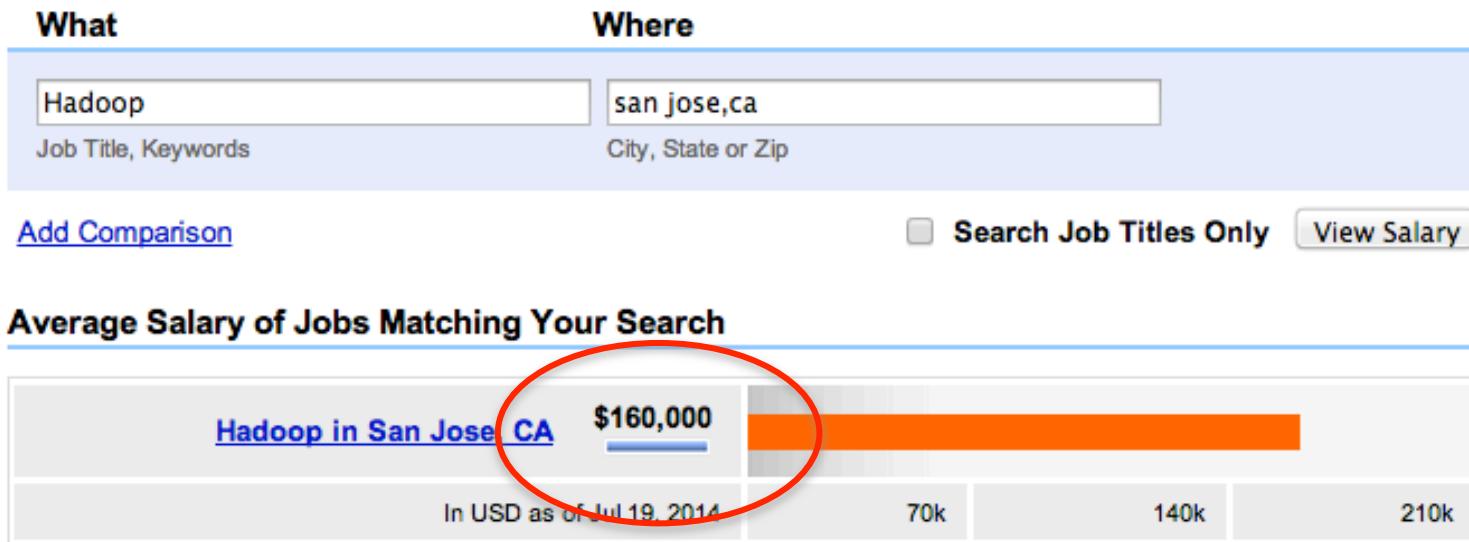
1. [HTML5](#)
2. [MongoDB](#)
3. [iOS](#)
4. [Android](#)
5. [Mobile app](#)
6. [Puppet](#)
7. **Hadoop**
8. [jQuery](#)
9. [PaaS](#)
10. [Social Media](#)



... and \$\$\$



## Hadoop Salary in San Jose, CA



+

Audience For This Talk...

# Developers



+

# This Doesn't Work....

- Quit Job on Friday
- Attend Big Data Bootcamp over the Weekend
- Start as a Big Data Developer on Monday

- ☺ sorry !





# Road Map For Launching Your Big Data Career

- (1) Learn
- (2) Network
- (3) Be Known
- (4) get hired

I don't always analyze data



But when I do, I prefer a lot of it

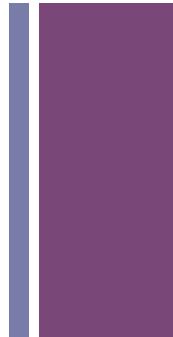


# Big Data / Hadoop Jobs

	Dev Ops / Admin	Developer	Data Scientist
What you do?	Keep things running	Build applications	Make sense of data
Skill sets	<ul style="list-style-type: none"><li>- Linux admin</li><li>- Scripting (python, shell)</li><li>- Puppet / chef</li><li>- Mad trouble shooting skills</li></ul>	<ul style="list-style-type: none"><li>- Java, Python, Ruby...</li><li>- Distributed systems</li><li>- Linux</li><li>- Hadoop / NoSQL concepts</li></ul>	<ul style="list-style-type: none"><li>- R, python</li><li>- Statistics, math</li><li>- Domain knowledge</li><li>(insurance, banking)</li><li>- Big Data tools</li></ul>
Best fit for	Admins	Developers	Analysts



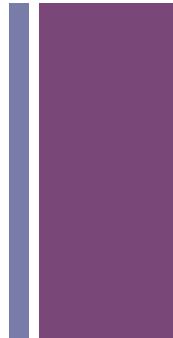
# Step (1) Learning



- Preferred Option : learn on your current job ☺
  - Take a training class
  - Do a Project
  
- If That is not possible (dead end job / employer)
  - Learn on your own



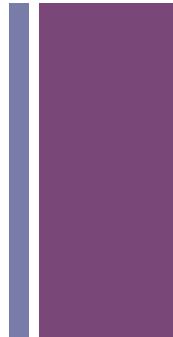
# Learning 1 : Learn



- Books
  - So many,
  - Start with 'Hadoop' by Tom White
  - Also checkout our free, open source book : 'hadoop illuminated' ☺
- Tutorials (Cloudera / HortonWorks)
- Blogs (Major vendors)
- Webinars
  - Free, watch at your own time
  - Signup at Cloudera / HortonWorks / DataStax
- Meetups
  - Plenty to choose from
  - My meetup : Big Data Gurus in San Jose ☺



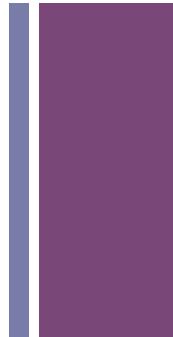
# Learning 2 : Practice



- Hands-on a must !!!!
- Get a Hadoop VM / Sandbox from a vendor
  - Easiest way to get Hadoop running
  - Free !
  - Every thing pre-installed and configured... ready to go!
- Use Hadoop version 2.x
  - Cloudera 5.x series
  - HortonWorks 2.x series



# Learning 2 : Practice

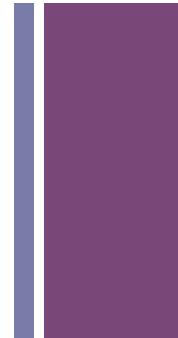


- Most VMs come with some tutorials pre-installed.. Do them
- We maintain an open-source Hadoop labs on github
  - <https://github.com/elephantscale/HI-labs>
  - 50+ labs on HDFS / MapReduce / Pig / Hive
- Where to get data?
  - [http://hadoopilluminated.com/hadoop\\_illuminated/  
Public\\_Bigdata\\_Sets.html](http://hadoopilluminated.com/hadoop_illuminated/Public_Bigdata_Sets.html)
  - Amazon hosts some big data sets



# Learning 2 : Practice ++

## How to stand-out



- Need more than 'hello world'
- Practice running Hadoop as a cluster
  - Use cloud providers like Amazon, Rackspace
- Cost ?



# of instances	5	1 hr	6 hrs
instance	price		
m1.medium	\$0.09	\$0.45	\$2.70
m1.large	\$0.18	\$0.90	\$5.40



# Challenges In Self Learning

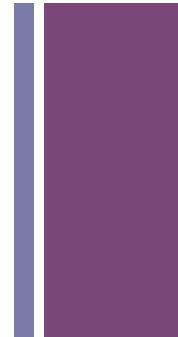
- Takes up a lot of personal time!

- Can loose motivation
  - Pair-study with some-one
  - Motivate / teach each other



+

Very quickly....



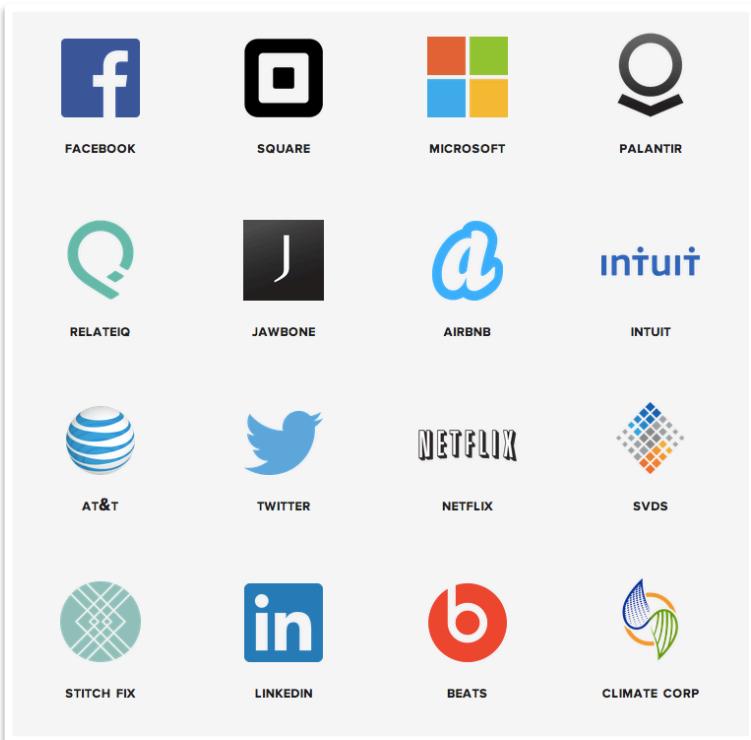


# Insight Data Engineering Fellowship (I am mentor!)

- 6 weeks, full time, professional fellowship
- **Completely free** for students ! (yes, really)
- Mentored by top industry experts (Nathan Marz – creator of Storm, Databricks– creators of Spark ..etc)
- Learn cool Data Engineering tools
- Build an awesome project
  - Motif finder at large scale
  - Inspect / visualize BitCoin transactions in real-time
- Demo to top companies  
(Netflix, Yelp, Facebook, Square)
- Get hired !



# INSIGHT



## Insight Data Engineering Fellows Program

[InsightDataEngineering.com](http://InsightDataEngineering.com)

- 2 sessions done in 2014
- Next session : Jan 2015
- Application deadline : **Oct 27**



# Big Data Skill Chasm



DAVE GRANLUND © [www.davegranlund.com](http://www.davegranlund.com)



# Do I Need a Certification?

- Certifications are good
  - if you have no practical experience
  - Consultant
- Cloudera / Horton Works
- You don't need to take their courses
- Just take the certification exams
  - Reasonably easy with preparation
  - Very affordable (few hundred \$)



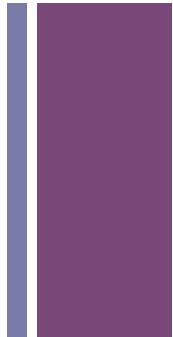
# Do I Need To Take A Pay Cut?

- Don't hesitate to take a paycut for your first Big Data job
- You will gain valuable practical experience
- And the next job you will make lot more



# +

# Step (2) Networking



- First get your OWN business card ☺
- Meetups
  - tons of meetup in this area
  - My meetup : <http://www.meetup.com/BigDataGurus/> ☺
- Conferences
  - Can be expensive (Strata \$3k)
    - Beg some one for a 'visitor pass' ☺
  - Cheap conferences (HbaseConf \$400, Hadoop Summit : \$500)
    - Money well spent... great connections!



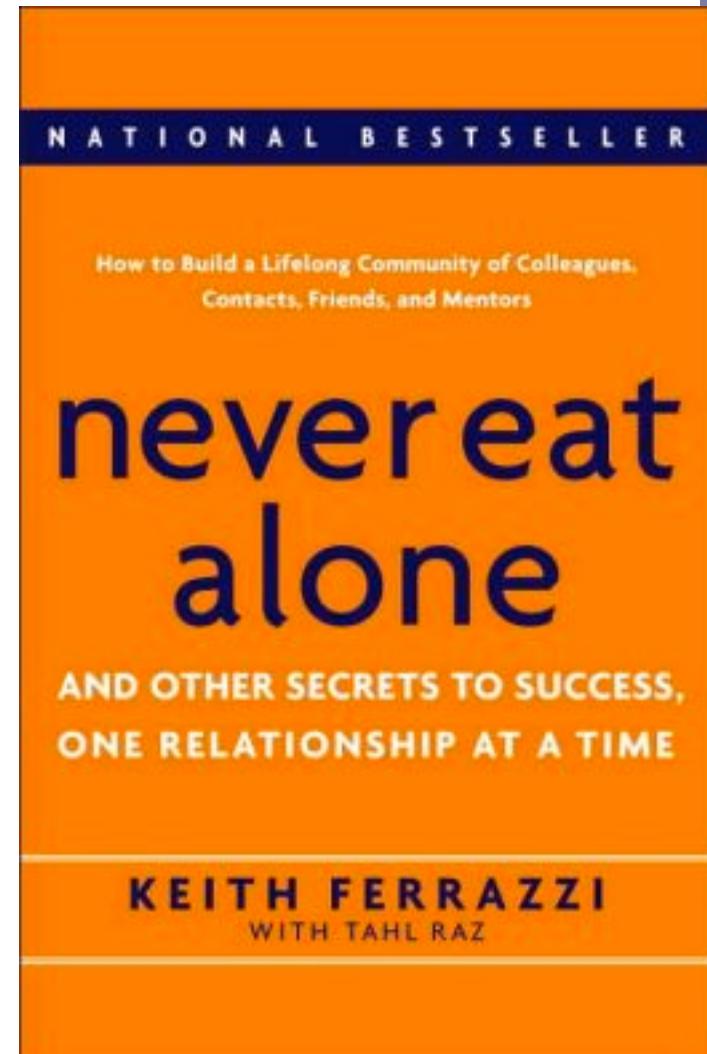
# Networking : How To Find Decision Makers?





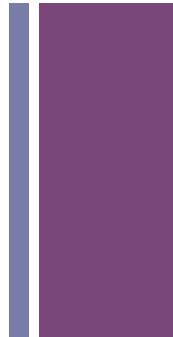
# Networking : Uber Networking Tips

- Read this book →
- Soft Networking
  - Become the connector, do intros
  - People will REMEMBER you!
- Volunteering
  - Help to run a meetup / event
  - You will get to know people you otherwise wouldn't meet  
(Board of directors ...etc)



+

## Step (3) -- Be Known



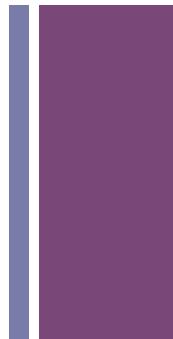
**It is not what you know**

**It is who you know**

**It is WHO knows YOU !**



# How to Be Known (aka How To Become an Expert!)



## ■ Open Source

- Huge boost to your resume

## ■ Write quality blogs, articles

- Lot of magazines wants contributors

## ■ Write a Book

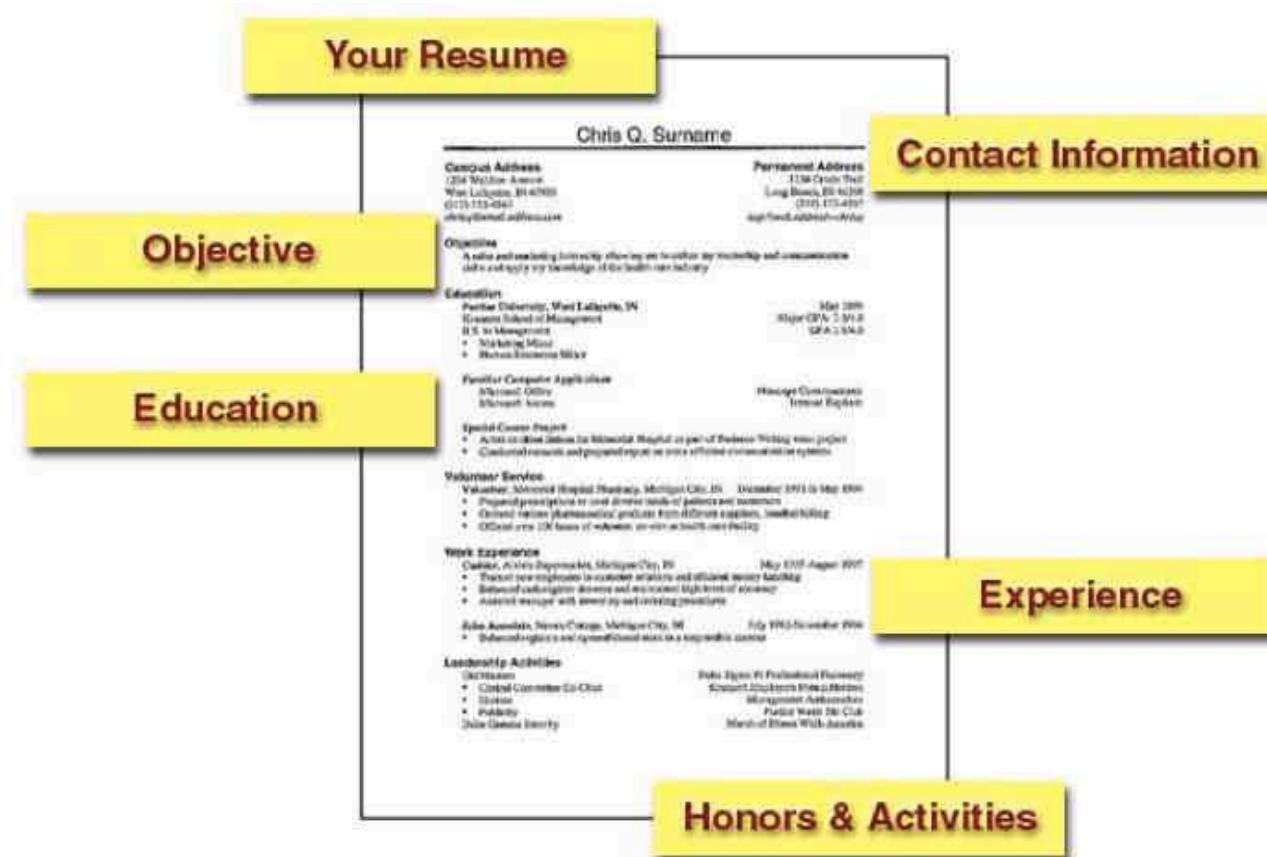
- We did it on our own – without a publisher

- ‘Hadoop Illuminated’ : <http://hadoopilluminated.com/>

## ■ Speak at meetups / conferences

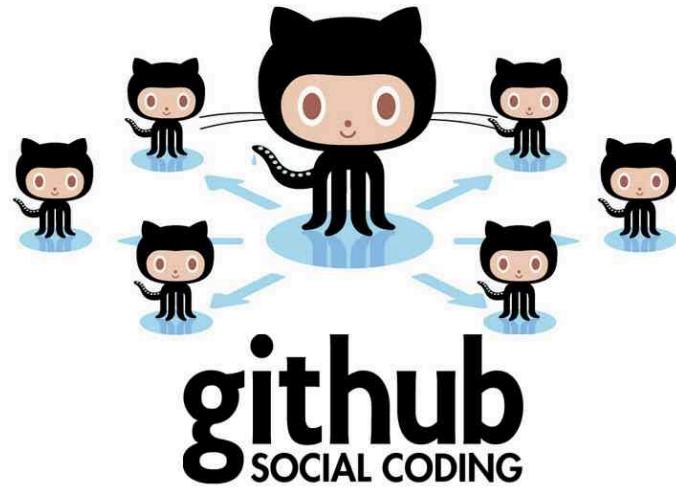
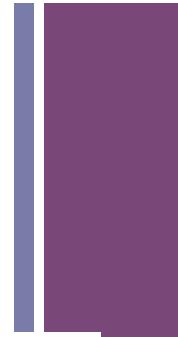


# Old Resume



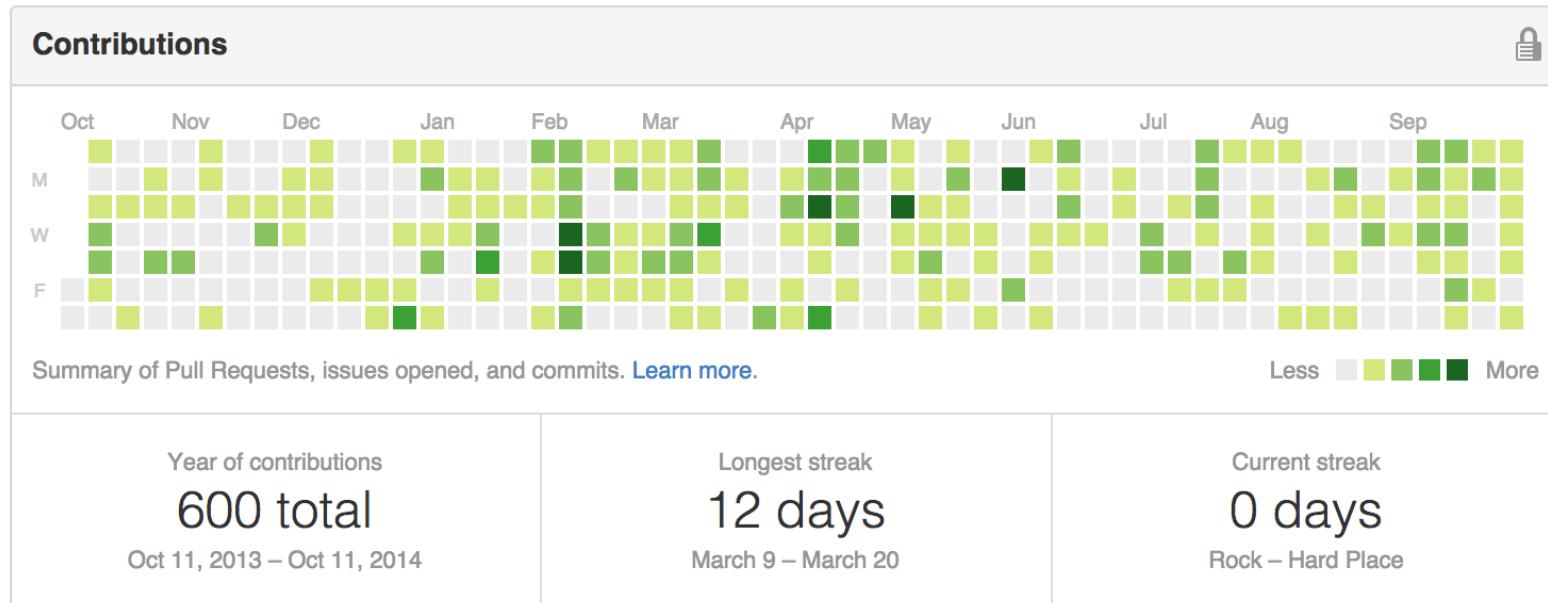
+

# Your New Resume





# Github activity log (employers check these !)





# Open Source Contributions

- Open Source involvement is a ‘hot skill’
- Just saying ‘I used TomCat’ isn’t enough ☺
- Open source tells me two things about you:
  - You are **passionate** about technology  
(not just b/c this gig pays well)
  - You dedicate your personal time → **initiative**
- Best option : Get Paid to work on open source ☺  
cloudera, linkedin, netflix....
- If not, you can still do meaningful contributions..



# How to Contribute To Open Source

- Step 1) Get a GitHub account (free)
- Step 2) Look for open source projects to contribute
  - Hadoop, cassandra, Spark
  - Start fixing bugs ('low hanging fruits')
- Step 3) Starting your own project
  - Has to be unique (not another word count example using Spark!)
  - Solve some thing you know about
    - E.g. : Mark Kerzner – eDiscovery & Hadoop



# Personal Story : Contributing to HBase

- [code] Improve benchmarking tool
  - Started as a hobby project
  - Submitted patch
  - Lots help from senior developers
  - <https://issues.apache.org/jira/browse/HBASE-4440>
- [documentation] improve patch submission process
  - You don't need to be a code-ninja to contribute !!
  - Documentation is badly needed in open source projects
  - <https://issues.apache.org/jira/browse/HBASE-5577>



# JIRA : HBASE-4440



HBase / HBASE-4440

## add an option to presplit table to PerformanceEvaluation

▼ Sujee Maniyam added a comment - 30/Nov/11 23:56

patch attached

▼ ramkrishna.s.vasudevan added a comment - 01/Dec/11 02:06

@Sujee

Wait till there not regions in RIT instead of sleep like how we have done in TestMasterFailover

```
log("Waiting for no more RIT");
ZKAssign.blockUntilNoRIT(zkw);
```

▼ Nicolas Spiegelberg added a comment - 06/Dec/11 02:35

thanks Sujee!

▼ Hudson added a comment - 06/Dec/11 05:53

Integrated in HBase-TRUNK-security #23 (See <https://builds.apache.org/job/HBase-TRUNK-security/23/>)  
**HBASE-4440** add an option to presplit table to PerformanceEvaluation



# Tips On Submissions

- **Make it easy for committers**
  - Don't create extra work for them !



HBase / HBASE-5577

improve 'patch submission' section in HBase book

▼ [Sujee Maniyam](#) added a comment - 28/Mar/12 00:17

hmm, looks like build system is trying to apply this 'patch' against trunk.  
should I have pasted the doc-changes, in the comment section?

▼ [Jonathan Hsieh](#) added a comment - 28/Mar/12 03:57

Sujee – ideally you'd modify the docs source code – I believe the files is `src/docbkx/developer.xml` – and submit a patch. 😊

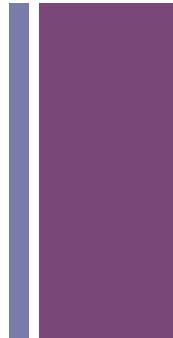


# Speaking at Meetups / Conferences

- Start with meetups
- Have a unique topic
  - “solving XXX using Spark” ..etc
- Having a popular open source project will help ☺ ☺
- Submit proposals to many conferences
  - You may not get into Strata first ☺
- Pay attention to ‘local’ conferences
  - SVCodeCamp, Dataweek in SF ...etc
- Big Data conference calendar  
[http://elephantsscale.com/bigdata\\_conferences](http://elephantsscale.com/bigdata_conferences)



# Acing The Interview



- Interviewer : So , have you used Hadoop at your work? What kind of practical experience you have?
- **If no, then usually interview ends here**
- You :  
Ahem, I haven't had a chance to use Hadoop at work...  
But let me tell you about the open source project I am working on...  
\* walk to whiteboard, start drawing, explain ...\*  
\* gets hired ! \*

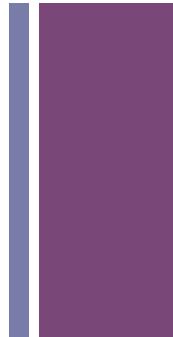


## Step (4) Get Hired





Thanks !



# Sujee Maniyam

[sujee@elephantscale.com](mailto:sujee@elephantscale.com)

<http://elephantscale.com>

Expert consulting & training in Big Data

