Big Data

Nicholas Ross, PhD

Introduction

Biography

- **Education:**
  - Mathematics Undergraduate: UC Berkeley
  - Economics Masters: UC Davis
  - Accounting PhD: UCLA

Most Importantly!

Work Experience

- **Bates White**
  - Litigation Consulting
- **TinyCo**
  - Director of Analytics
- **Sega**
  - Director of Analytics
- **Currently:**
  - Professor of Accounting and Analytics

Things not covered in this talk

Big data is a Thing!

When accounting meets Big Data

September 8, 2015 By B

Accountants Dealing with Big Data and Other Big Issues

NEW YORK (JULY 22, 2014)

BY MICHAEL COHN

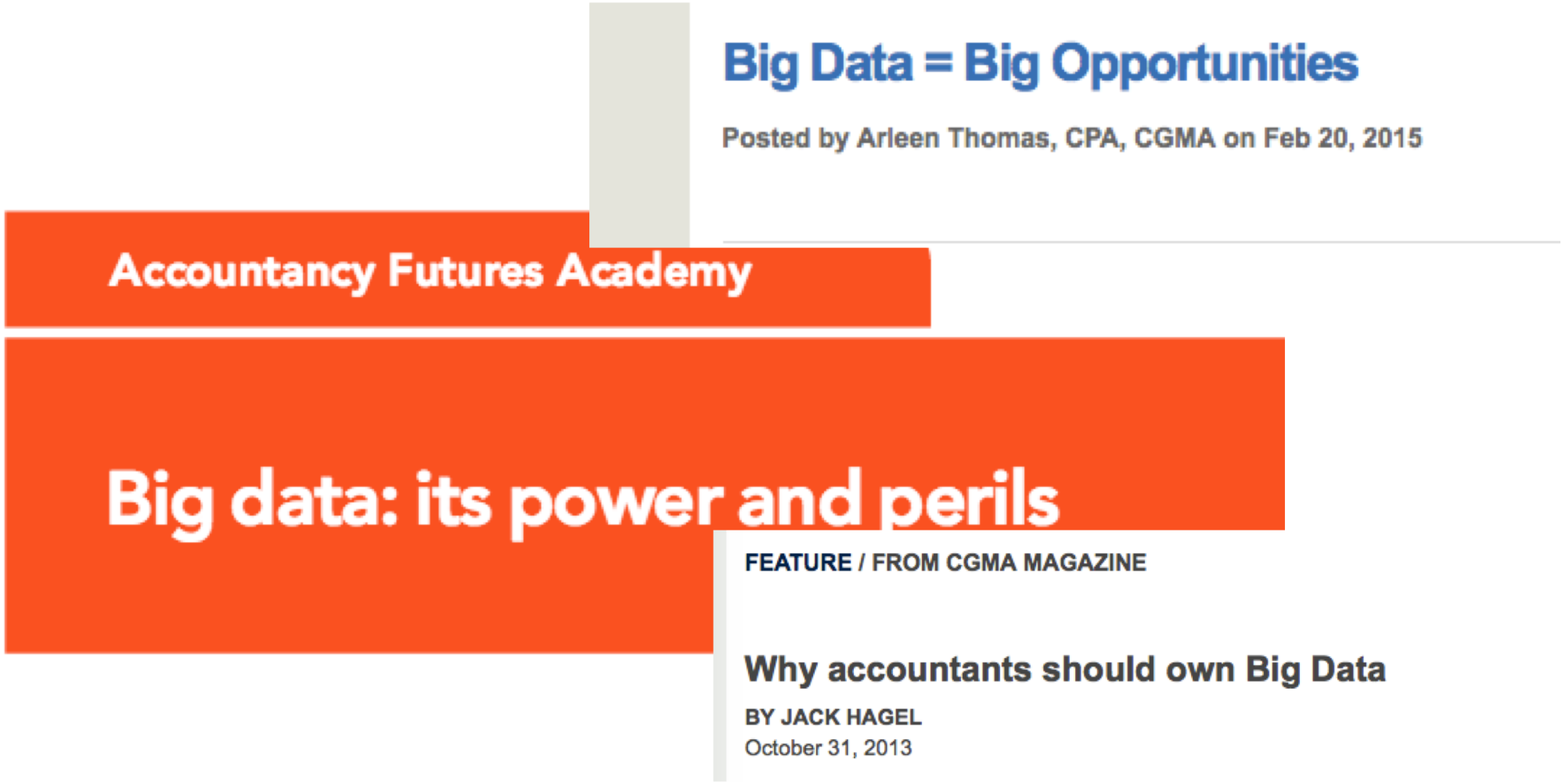
FEATURE / FROM CGMA MAGAZINE

Why accountants should own Big Data

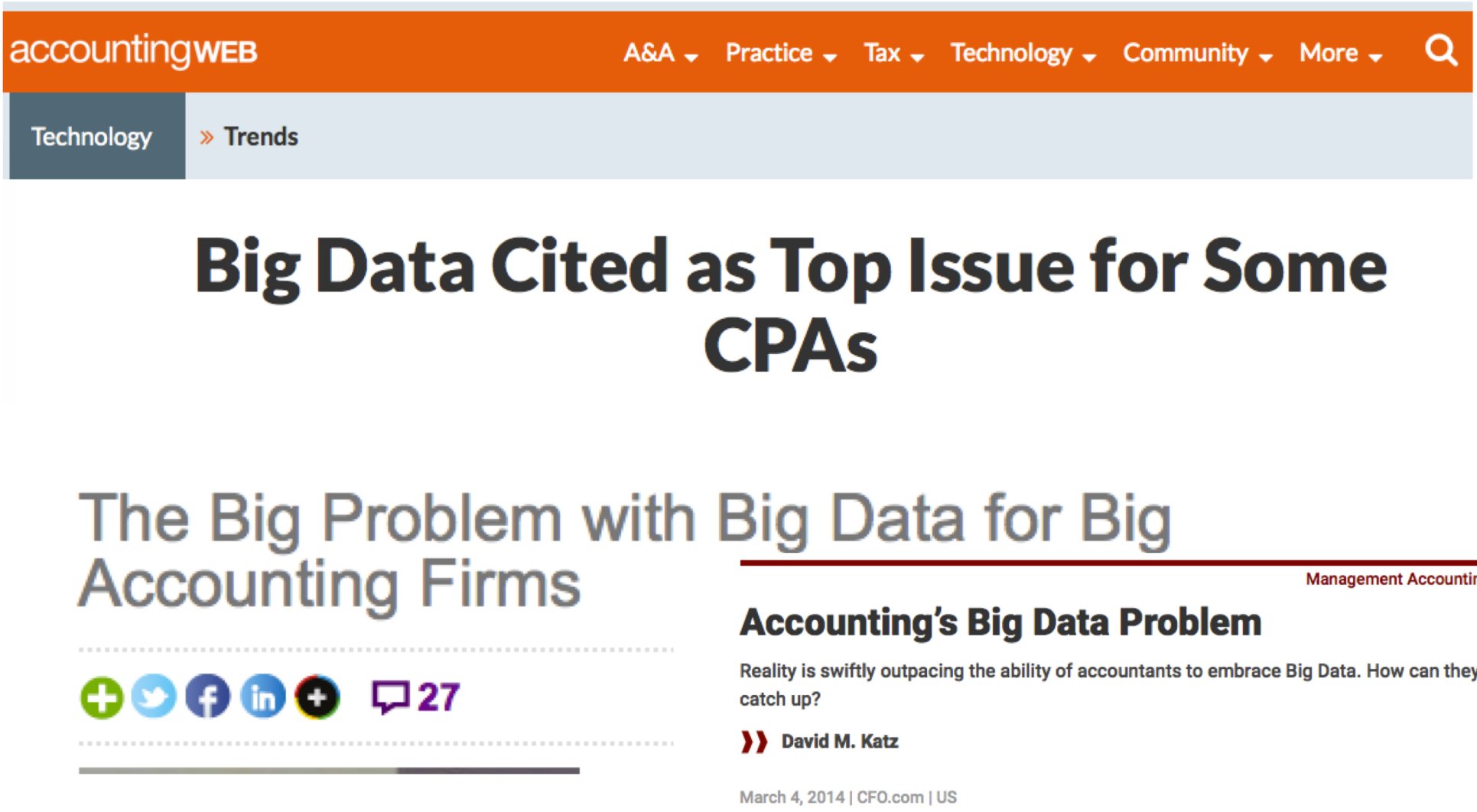
BY JACK HAGEL

October 31, 2013

Big data is Important!



Big data is Hard!



Learning Objectives

- How to evaluate data – which characteristics are important when evaluating a dataset against current capabilities
- Determine how to match workload against available tools
- Understand big data team structure, including what skills and roles are needed to hire to build a world-class team

Learning Objectives

- How do we usefully define big data?
- What are the Critical Success Factors ("CSFs") for successfully working with big data?
- What your team composition says about your current big data abilities?

My strong opinions

- **Investing in big data technology is rarely a long-term win**
  - Costs
  - Change / Lock-in
- **Big data isn't your core competency**
  - Big data is a tool
  - If you are a gold miner you need a good shovel, but you don't need to own a shovel factory
- **Analysis isn't the hard part**

Overview

- ~~Introduction~~
- CSFs of big data
- Defining big data
- Big data team roles
- Matching engagements to teams
- What is the current “standard” big data technology stack
- Hiring

Big Data Critical Success Factors:

- Reproducible Processes
  - Do my tools generate the same answers repeatedly?
  - Who or what guarantees that my analysis generates the same answers each time?
- Ease of Iteration
  - As my understanding changes, how does that work its way back into my analysis and numbers?
  - What is the analysis “round-trip” time?

Critical Success Factors (cont.)

- **Transparency**
  - What does my verification process look like?
  - How do I check a junior employee's work product?
- **Ease of the deep dive**
  - Can I quickly pull an Excel-sized cut of the data?
- **Cost**
  - Are my costs higher than the value of the engagement?

Overview

- ~~Introduction~~
- ~~CSFs of big data~~
- Defining big data
- Big data team roles
- Matching engagements to teams
- What is the current “standard” big data technology stack
- Hiring

What is *Big* Data?

- **Data that doesn't work in your current environment**
  - Specific to your own operations
- **Articulate:**
  - What am I capable of?
  - What is my next engagement?
- **Use the 3 V's:**
  - Volume
  - Velocity
  - Variety

3 V's of Data

Volume: How much data do I have?

- Easiest dimension to understand:
  - "I have a 50 Megabyte excel file"
  - "The database has 200 million rows"
- Most of the time described in bytes, or, for some applications, row or object counts is used

3 V's of data

Velocity: At what rate is data being created?

- How much *new* data is being created?
- Example: every month, your client sends you 2 MB of that month's inventory data
- Many, if not most, engagements will have a velocity of zero
- If the velocity is greater than zero, need to have an **Extract, Transform and Loading ("ETL") plan.**
  - How frequently does data load?
  - When does it load?
  - What checks do I have in place to verify data integrity?

3 V's of data

- **Variety: How many different forms does my data take?**
  - What are the number of different data sets and sources in this engagement?
  - Each unique dataset requires taking time to evaluate and understand, as well as to clean and merge
  - The same data from different sources should be considered different data

Variety: Structured vs. Unstructured

- **Structured data:**
  - "Normal" data that fits into rows and column easily
- **Unstructured data:**
  - The opposite: data without an easy organization structure
  - Examples: emails, voicemails, medical records, books and images
- **Unstructured data is generally more costly to understand and analyze**
- **This presentation focuses on structured data**

Recap

- **We now can define big data is (3 V's) and what we want (CSF's):**
  - What is your team's capability?
  - What types of engagements can you handle?
- **What should our team look like if we want to expand our capabilities?**

Overview

- ~~Introduction~~
- ~~CSFs of big data~~
- ~~Defining big data~~
- Big data team roles
- Matching engagements to teams
- What is the current “standard” big data technology stack
- Hiring

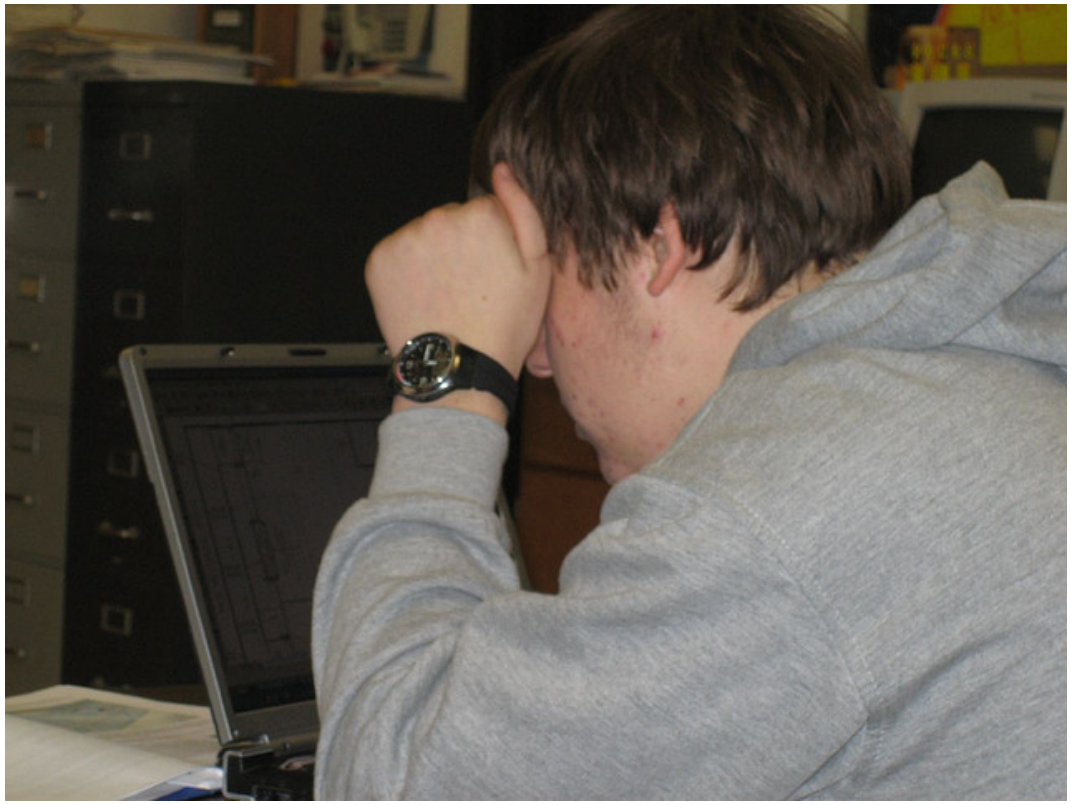
Team roles

Roles: Junior / Senior Data Analyst

- **Kevin and Rebecca:**
  - $35-$100K
  - R, Stata and Excel
- **Rebecca**
  - 2 years' experience
- **Kevin**
  - Graduated 2 months ago
- **Hiring CSF:**
  - Business applied to data

Roles: Data Engineer

- **Travis**
  - $100 - $200K
  - Python, Java and AWS
  - 5+ years' experience
- **Hiring CSF:**
  - Can solve own problems
  - Pass a code test
  - Cloud computing
  - Always looks bored



Roles: IT / DBA

- **Dell, our IT guy**
  - 50K+
- **Server configuration**
- **Basic database maintenance**
- **CSF:**
  - Limited database capabilities

Overview

- ~~Introduction~~
- ~~CSFs of big data~~
- ~~Defining big data~~
- ~~Big data team roles~~
- Matching engagements to teams
- What is the current “standard” big data technology stack
- Hiring

Matching team to engagements

0 V Problem: Just Kevin and Rebecca

- **Can do everything when all 3 V's are small:**
  - Less than a few GB of data
  - No more than a few different datasets
  - Data is static – it's velocity is essentially zero
- **Excel, Stata or R and some bootstrapping around how analysis is shared**
- **Biggest Risk: Reproducibility**
- **If any V increases, chaos ensues**

1 V Problem: Kevin, Rebecca and Dell

Problem with 1 V…

- Volume: Maybe more data (A few TB)
- Or Velocity: Maybe data loads occurring once or twice a day, but not significantly large data
- Or Variety: Maybe dataset variety increasing significantly

Now:

- Kevin and Rebecca learn **SQL**
- Dell sets up a beefy server at the office
- Dell shows Kevin and Rebecca some scripting tools

Biggest Risk: Ease of Iteration

- What is Dell's primary responsibility?

>1 V Problem: Kevin, Rebecca and Travis

- **>1 V issue:**
  - Velocity: Large amount (> 500 MBs) loaded daily
  - Volume: Hundreds of TBs
  - Variety: Significant variety of sources
- **Sorry Dell!**
  - Cloud services (such as AWS) now cost effective
- **Now:**
  - Kevin and Rebecca use SQL
  - Travis writes scripts to help Kevin and Rebecca when they hit their limits
- **Biggest Risk: Cost**

Key take-away: Analysis vs. storage

As V problems increase:

- The distinction between working on data “storage” vs. data “analysis” increase
- Rebecca and Kevin can do analysis with some storage, but will get stuck and need help:
  - For small V problems, Dell can assist
  - For large V problems, call Travis

Cost Example

- **15 TB of data**
- **Year-long engagement**
  - Travis could spend time on another engagement
  - Doesn't include benefits, etc.
- **Using AWS**
  - About $4K month, SQL style server

Cost	Yearly Total
Travis	$150,000
10 Med. Servers @ 3TB	$50,000
Rebecca and Kevin (75K Per)	$150,000
Total	**$350,000**

Overview

- ~~Introduction~~
- ~~CSFs of big data~~
- ~~Defining big data~~
- ~~Big data team roles~~
- ~~Matching engagements to teams~~
- What is the current “standard” big data technology stack
- Hiring

Standard Tooling

SQL

- Established technology
- Easy for analysts to learn
- **Requires configuration work in proportion to data size!**

The “Current” Standard Stack:

- Python scripts that load data
- Amazon Redshift (SQL Server)

Cost and Tooling caveat

- **Use technology properly:**
  - Lose leverage of Kevin and Rebecca: Turn them into data monkeys, not analysts
  - Costs will be even higher if Kevin and Rebecca aren't utilized efficiently
- **Tendency to overinvest in tooling**

Overview

- ~~Introduction~~
- ~~CSFs of big data~~
- ~~Defining big data~~
- ~~Big data team roles~~
- ~~Matching engagements to teams~~
- ~~What is the current "standard" big data technology stack~~
- Hiring

On hiring: Kevin, Rebecca and Travis

Hiring in data is hard

- **Common data employee types:**
  - Data Analysts
  - Data Scientists
  - Data Engineers
  - IT
- **Data scientists are costly and often unnecessary:**
  - Skilled at implementing “new” things – not following process
- **Turnover is high**
- **Cross-discipline hires are even more difficult**
  - Accounting

Thanks!

- **Feel free to contact me on LinkedIn or via email with any questions!**
- **Linkedin:**
  - https://www.linkedin.com/in/drdata
- **Email:**
  - nickross510@gmail.com