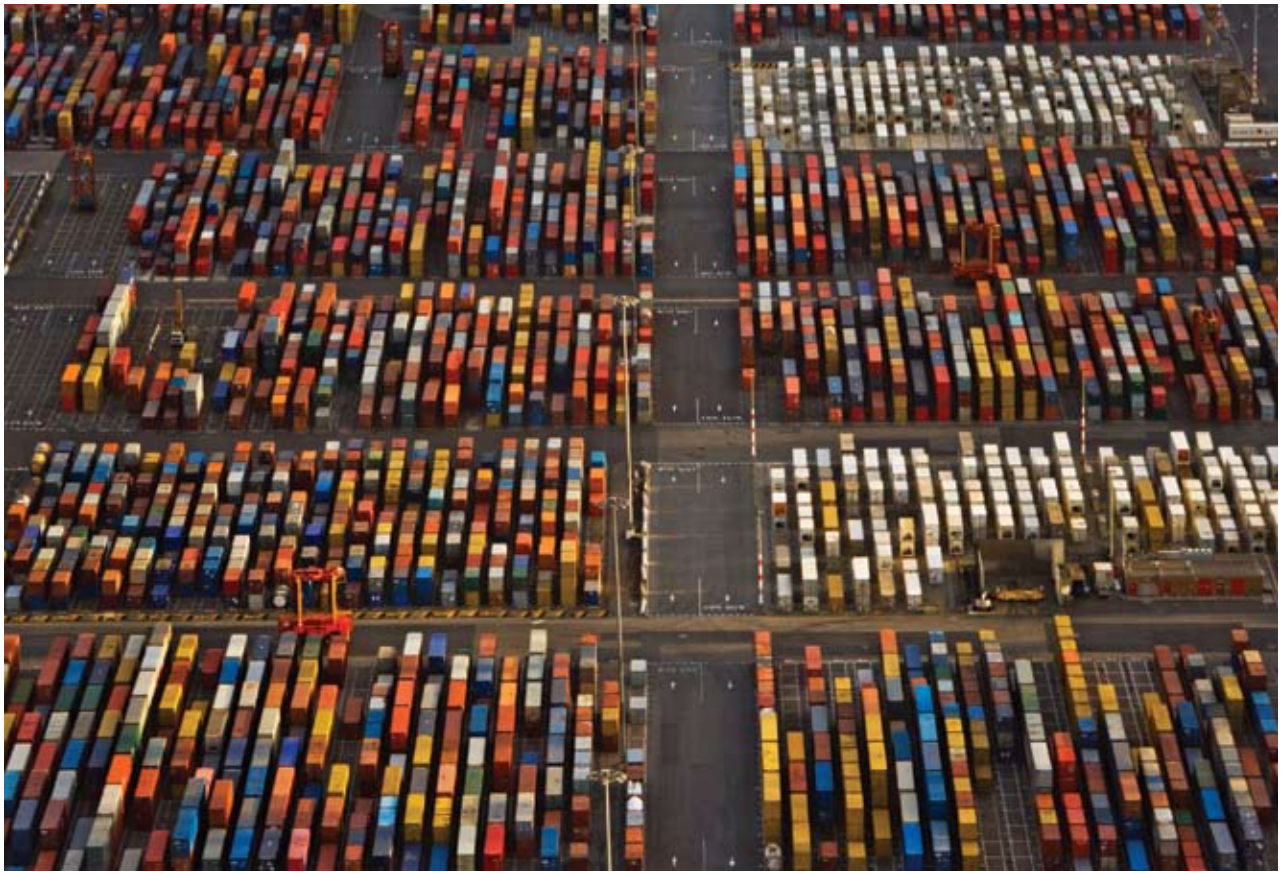
O'REILLY® radar

Release 2.0 11

Issue 2.0.11, February 2009
http://r2.oreilly.com



"Ultimately, big data is more about attitude than tools; data-driven organizations look at big data as a solution, not a problem."

Roger Magoulas and Ben Lorica, from *Big Data: Technologies and Techniques for Large Scale Data*, page 32

Release 2.0
Issue 11, February 2009
ISSN 1935-9446

Published six times a year by
O'Reilly Media, Inc.,
1005 Gravenstein Highway North,
Sebastopol, CA 95472
http://r2.oreilly.com

This newsletter covers the world of information technology and the Internet—and the business and societal issues they raise.

subscription information
Release 2.0
PO Box 17046
North Hollywood, CA 91615-9588
http://dispatchservice.oreilly.com

customer service
1.800.889.8969
1.707.827.7019
r2@oreilly.com

Contents

Roger Magoulas is the Director of Research at O'Reilly. **Ben Lorica** is a Senior Analyst in O'Reilly's Research group.

Big Data:
Technologies and Techniques for Large-Scale Data

Preface: Stories from the Field

You take a leave of absence from an organization known for handling big data to work on the data analysis systems for the Obama campaign. You're faced with one big server and five terabytes of messy voter registration data from multiple sources in multiple formats. You're tasked with optimizing "get out the vote" efforts by finding out who has already voted and removing those names from the call-bank lists used by canvassers—real-time, on election day. With only a few weeks to build the system, you assemble a small team of people comfortable with several aspects of big data management, i.e., the size and state of the data, analytics, and serving the data to many users and many devices.

In the end, while there are problems on election day, you are able to clean 1.6 million voters from the call lists the campaign distributes to canvassers that afternoon, making those lists 25% shorter, on average. Your knowledge and experience with big data management makes a complex task manageable in a tight timeframe with a small team. And, you spare 1.6 million supporters an unnecessary phone call.

Alternatively, you work for a large social networking website and you're tasked with creating an analysis infrastructure that serves a unique group of users. The data is already large and growing fast. There's no real business plan and the data is interesting beyond just its business insights—there's fodder for real sociology research in your social graph. There's no time and no guidance, but everyone thinks the data is important.

You determine that the Hadoop implementation of MapReduce provides the scaling, performance, and flexibility you need. There's no requirement to predefine a schema, so you can just throw data into the Hadoop platform. Your developers and analysts can build and use the various access points, with the help of a few tutorials. Over time, the most effective and repeatable analysis

patterns become clear and you start to develop access tools that support different classes of users, e.g., programming language APIs for developers and SQL-like access for analysts. Your now data-driven organization has the infrastructure to capture all the data generated by your website, run experiments and one-off analyses, and identify opportunities to build more formal, repeatable data analysis interfaces when needed.

We heard these stories from folks doing leading-edge big data implementations, and their experiences aren't isolated or unusual. More and more organizations are facing the challenges and opportunities of working with big data, and they're transforming themselves in the process. We've seen that the organizations that best handle their big data challenges can gain competitive advantage and improve their product and service offerings.

For those organizations facing new challenges and new opportunities regarding big data, we present a roadmap of choices and trade-offs for large-scale data management. There's a lot to make sense of and many competing perspectives. The high-level descriptions and guidance regarding what to consider can inform a deeper dive into making decisions about your big data environment.

Introduction to Big Data

Big Data: when the size and performance requirements for data management become significant design and decision factors for implementing a data management and analysis system. For some organizations, facing hundreds of gigabytes of data for the first time may trigger a need to reconsider data management options. For others, it may take tens or hundreds of terabytes before data size becomes a significant consideration.

We're at the front edge of a data deluge, brought on by new, pervasive data sources. A few years ago, a retailer analyzing thousands of T-shirt sales to discern customer behavior thought it was dealing with big data. Today, social networking companies examine hundreds of millions of personal interactions to identify social trends and relationships, and energy utilities plow through petabytes of sensor data to understand use trends and demand projections. Given the scale of today's data sets, traditional approaches to data acquisition, management, and analysis don't always measure up.

Over the past year, we've noticed a number of faint signals, from the people we talk to and the data we research, that making sense of large-scale data

stores is increasingly interesting and important. What we find most notable is the broad array of organizations, from large enterprises and government agencies down to startups, that are tackling big data—the size of the organization no longer directly correlates with the size of the data challenges. We're also seeing the role of data becoming more central to business strategy. Companies like Google (where analytics is at the heart of how they manage ad revenue), Facebook, (which is attempting to harness the power of data on its social graph of users to develop its business plan), or Twitter (focusing on analyzing its micro-messaging data as the basis for a business model) are examples of companies organized around data insight. Vendors and various open source communities are responding with a new set of tools and techniques to handle this emerging focus on data. We see new big data challenges, growing interest in the topic, and an increasingly diverse set of tools available to address these challenges.

Big data is a big topic. To help make sense of all that big data entails, we divide the topic into three broad activities:

- Data acquisition
- Data management
- Analysis and insight

Data acquisition—whether from data-collecting sensors, increasingly computerized systems, web content, telematics, social networks, or ubiquitous computing—leads to the need to store and manage more data, data that can become valuable with access and iterative, repeatable analysis. This puts data management at the center of big data—scaling to acquire more data and providing fast, convenient access and sophisticated analysis to all that data. In this report we focus on data management as a critical link in the big data story. We'll investigate different approaches to handling large-scale data, describe the technology, identify key trade-offs, and address resource requirements.

Why Big Data Matters

We believe that organizations need to embrace and understand data to make better sense of the world. (We believe it so much that O'Reilly co-sponsored an "unconference" covering Collective Intelligence topics, e.g., data mining and analytics around human behavior.) Big data matters because:

- The world is increasingly awash in sensors that create more data—both explicit sensors like point-of-sales scanners and RFID tags, and implicit sensors like cell phones with GPSs and search activity.

Key Takeaway

Building and making sense of massive databases is the core competency of the information age. Being better at data is why Google beat Yahoo! and Microsoft in search and one reason why Barack Obama beat John McCain. A bad economy accelerates the importance of big data—companies without big data competencies will be left behind.

- Harnessing both explicit and implicit human contribution leads to far more profound and powerful insights than traditional data analysis alone, e.g.:
  - Google can detect regional flu outbreaks seven to ten days faster than the Centers for Disease Control and Prevention by monitoring increased search term activity for phrases associated with flu systems, [M. Helft, "Google Uses Searches to Track Flu's Spread," *New York Times,* 11/11/2008]
  - MIT researchers were able to predict location and social interactions by analyzing patterns in geo/spatial/proximity data collected from students using GPS-enabled cell phones for a semester, [N. Eagle and A. Pentland, "Reality mining: sensing complex social systems," *Personal and Ubiquitous Computing,* Vol 10, #4, 255-268]
  - IMMI captures media rating data by giving participants special cell phones that monitor ambient noise and identify where and what media (e.g., TV, radio, music, video games) a person is watching, listening to, or playing, [J. Pontin, "Are Those Commercials Working? Just Listen." *New York Times,* 9/9/2007]
- Competitive advantage comes from capturing data more quickly, and building systems to respond automatically to that data.
- The practice of sensing, processing, and responding (based on pre-built models of what matters, "the database of expectations," so to speak) is arguably the hallmark of living things. We're now starting to build computers that work the same way. And we're building enterprises around this new kind of sense-and-respond computing infrastructure.
- As our aggregate behavior is measured and monitored, it becomes feedback that improves the overall intelligence of the system, a phenomenon Tim O'Reilly refers to as harnessing collective intelligence.
- With more data becoming publicly available, from the Web, from public data sharing sites like Infochimps, Swivel, and IBM's Many Eyes, from increasingly transparent government sources, from science organizations, from data analysis contests (e.g., Netflix), and so on, there are more opportunities for mashing data together and open sourcing analysis. Bringing disparate data sources together can provide context and deeper insights than what's available from the data in any one organization.
- Experimentation and models drive the analysis culture.
  - At Google, the search quality team has the authority and mandate to fine-tune search rankings and results. To boost search quality and relevancy, they focus on tweaking the algorithms, not analyzing the data.

- Models improve as more data becomes available, e.g., Google's automatic language translation tools keep getting better over time as they absorb more data*.
- Models and algorithms become the focus, not data management.

Big data repositories provide the opportunity, via analysis, for insights that can help you understand and guide your organization's activities and behaviors. You can improve results by combining more data from more sources with more sophisticated analysis and models.

The power of big data needs to be tempered with the responsibility of protecting privacy and civil liberties, preventing sensitive data from getting hacked or inappropriately shared, and treating people generating the data fairly. The insights gained from big data can be used to improve products and customer service, but they can also be used in ways that creep out customers and make them feel uncomfortable or watched. The industry doesn't have all the answers, e.g., academic research shows it's difficult to create truly anonymous data. There are techniques, such as aggregating data beyond the level of an individual, that can protect privacy while still allowing insightful analysis. Although it's beyond the scope of this report to address privacy issues in detail, you'll need to consider them as you work with big data.

Key Takeaway

Big data is driving new approaches: MPP, MapReduce, column-oriented data are all becoming essential parts of the database toolkit.

The relational model is no longer the only database that matters. For many problems, MapReduce-style processing is superior. What's more, it's easier for many programmers to understand and implement than SQL.

Parallelism is the answer to big data challenges: it lets you "divide and conquer," and it's built to scale.

*** FOOTNOTE:**

"How Google translates without understanding; Most of the right words, in mostly the right order," by Bill Softky, *The Register,* May 15, 2007. http://www.theregister.co.uk/2007/05/15/google_translation/

Calendar

A selection of significant public events over the next few months.

February 9–11

TOC: Tools of Change for Publishing (New York, NY)
http://en.oreilly.com/toc2009/public/content/home
Reinventing an industry for our networked, customer-driven world.

February 23–24

FOWA: Future of Web Apps (Miami, FL)
http://events.carsonified.com/fowa/2009/Miami
The title says it: future of the browser; web business models; online community; plus a Bar Camp. Across the pond, the Dublin version is March 6.

March 1–2

DEMO (Palm Desert, CA)
http://www.demo.com
Seventy 6–minute presentations from companies hoping to be the Next Big Thing.

March 4–7

DrupalCon (Washington, DC)
http://dc2009.drupalcon.org
The official unconference for developers and users of open source CMS Drupal.

March 9–12

ETech (San Jose, CA)
http://en.oreilly.com/et2009
Because innovation's more important than ever in "interesting" times.

March 13–17

SXSW Interactive (Austin, TX)
http://sxsw.com/interactive
Digital creatives head south.

March 18–20

Mix 09 (Las Vegas, NV)
http://2009.visitmix.com
The March Marathon continues for webheads, with Microsoft's web design and development conference.

March 31–April 3

Web 2.0 Expo (San Francisco, CA)
http://en.oreilly.com/where2009/
Our prescient theme (picked last summer): The Power of Less

April 15–17

The Next Web (Amsterdam, NL)
http://2009.thenextweb.com
Even more web goodness, this time in fabulous Amsterdam.

May 19–21

Where 2.0 (San Jose, CA)
http://www.web2expo.com/webexsf2009
Knowing "where" opens up a world of possibilities, from the mundane (finding pizza) to the world-changing (cultural preservation via indigenous maps).

June 9–11

Found: The Search Acquisition and Architecture Conference
(Burlingame, CA)
http://en.oreilly.com/found
A new O'Reilly conference for web developers, designers, SEO specialists, marketing strategies, and online entrepreneurs.

Release 2.0 Subscription Form

Complete this form and join the other industry executives who rely on *Release 2.0* to stay ahead of the headlines. You can also subscribe online at *http://r2.oreilly.com*.

Your annual *Release 2.0* subscription costs $495 per year, and includes both the print and electronic versions of six every-other-month issues, access to the complete *Release 1.0* and *Release 2.0* archives, a discount on attendance to O'Reilly conferences, and full access to our website, *http://r2.oreilly.com*.

name

title company

address

city state zip country

telephone fax

email (personal email required for electronic access)

url

☐ My colleagues should read *Release 2.0*, too!

Send me information about multiple copy subscriptions and electronic site licenses.

☐ check enclosed ☐ charge my (check one) ☐ american express ☐ mastercard ☐ visa

card number expiration date cvs code

name and billing address ☐ same as above ☐ see below

name

address

city state zip country

signature

Please fax this form to 1.818.487.4501 or mail it to Release 2.0, P.O. Box 17046, North Hollywood CA 91615-9588

Payment must be included with this form. Your satisfaction is guaranteed or your money back.

If you have any questions, please call us at 1.800.889.8969 or 1.707.827.7019 or email us at *r2@oreilly.com*.