Putting Big Data to Work: Opportunities for Enterprises

By Brett Sheppard

Table of Contents

About Brett Sheppard

Brett Sheppard is an executive director at Zettaforce and previously served as a senior analyst at Gartner. His work focuses on big data analytics and collaboration.

About GigaOM Pro

GigaOM Pro gives you insider access to expert industry insights on emerging markets. Focused on delivering highly relevant and timely research to the people who need it most, our analysis, reports and original research come from the most respected voices in the industry. Whether you're beginning to learn about a new market or are an industry insider, GigaOM Pro addresses the need for relevant, illuminating insights into the industry's most dynamic markets.

Visit us at: http://pro.gigaom.com

Executive Summary

Business and IT leaders now face significant opportunities and challenges with big data. This is true across numerous industries, from health care and finance to travel booking sites and even sports.

In 2011, we're seeing extraordinary growth of big data in these areas along three dimensions: volume, complexity and speed. Additionally, data science is contributing significantly to operational efficiencies in these sectors as well as enabling more sales growth and even brand-new business models.

This report, published in conjunction with the GigaOM Structure Big Data 2011 conference, explores the rapidly evolving big data business and technology ecosystem. It examines big data in the context of several different industries: financial services, health care, sports, travel and media. We explore the different big data technologies — from Hadoop and in-database analytics to cloud-based collaboration tools — and their various benefits for enterprises. And we examine some of the existing challenges big data poses, and what enterprise IT leaders can do to overcome them.

Defining Big Data

The term "business intelligence" (BI) dates back to 1958, when IBM researcher Hans Peter Luhn coined the term in an *IBM Journal* article.[1] However, it took until the late 1980s and early 1990s for a group of industry thought leaders to popularize "business intelligence" as an umbrella term to cover software-enabled innovations in performance management, planning, reporting, querying, analytics, online analytical processing, integration with operational systems, predictive analytics and related areas.

[1] *H.P. Luhn, "A Business Intelligence System," IBM Journal, October1958.*

Much like the history surrounding the term "business intelligence," the current data marketplace lacks a definitive term that is widely accepted. Those that more or less overlap in meaning include "very large databases," "extremely large databases," "big data," "extreme data" and "total data."

Roger Magoulas and Ben Lorica at O'Reilly Media offer a good definition of big data: when the data size and performance requirements "become significant design and decision factors for implementing a data management and analysis system." In this definition, there's not an absolute size milestone between "data" and "big data."

For the purpose of this report, we can postpone any terminology debates to a later date and focus on managing and deriving business benefits from ever larger and more complex data sets that increasingly require real-time or near-real-time capabilities. Three characteristics — volume, complexity and speed — define the big data marketplace in 2011.

Volume

Bank of America, Dell, eBay and Wal-Mart Stores, along with multiple government agencies, are among the members of the "Petabyte Club," with enterprise data volumes in the petabytes (10^{15} bytes).

While IDC estimates that global data volumes are now in the zettabytes (10^{21} bytes), what constitutes big data is relative and varies by organization. For a large enterprise, data volumes may exceed multiple petabytes, while for a small or midsize enterprise, data volumes that grow into tens of terabytes may become problematic to manage and hence become "big data" in the context of that organization.

To start with a Fortune 500 enterprise example, at Apple, the Information Services and Technology department operates a Teradata enterprise data warehouse, along with Oracle databases. These provide reporting solutions for the company's cross-

functional business units, including marketing, sales, operations, support and finance. Extract transform load (ETL) and data integration tools from Informatica and other providers deliver access to multiple terabytes of data from SAP enterprise resource planning (ERP) software and other data sources.

Many small and midsize organizations, too, are grappling with data volume growth. For example, regional bank holding company Zions Bancorporation and its banking subsidiaries, which operate in 10 U.S. states, began with a single data mart that soon became overloaded. It migrated its financial data mart to an EMC Greenplum system and has since created several new data marts using EMC Greenplum.[2]

Complexity

In its most recent annual survey of corporate CEOs, IBM found that the majority of the more than 1,500 CEOs surveyed identified complexity as their organization's greatest challenge. According to IBM, complexity in this context means: "That new trends are interacting, creating levels of volatility, unpredictability, opportunity, and threat in forms never encountered before—be it upstart companies, market movements, the proliferation of data, or heightened expectations and scrutiny from customers."

Nearly 80 percent of CEOs believe their business environment is becoming more complex, yet few believe they are ready to handle that complexity. Those senior executives who sense more uncertainty on the horizon yet feel unprepared for it have fallen into what IBM terms the "complexity gap."[3]

As discussed by Professor of Computer Science Joseph M. Hellerstein at the University of California, Berkeley, we've entered the era of the "industrial revolution of data." The equivalents of "data factories" generate UPC barcode reads, RFID scans and GPS location data. Sensors are extending from heating, ventilation and air control (HVAC) and industrial plant monitoring to automotive sensors and motion detection in

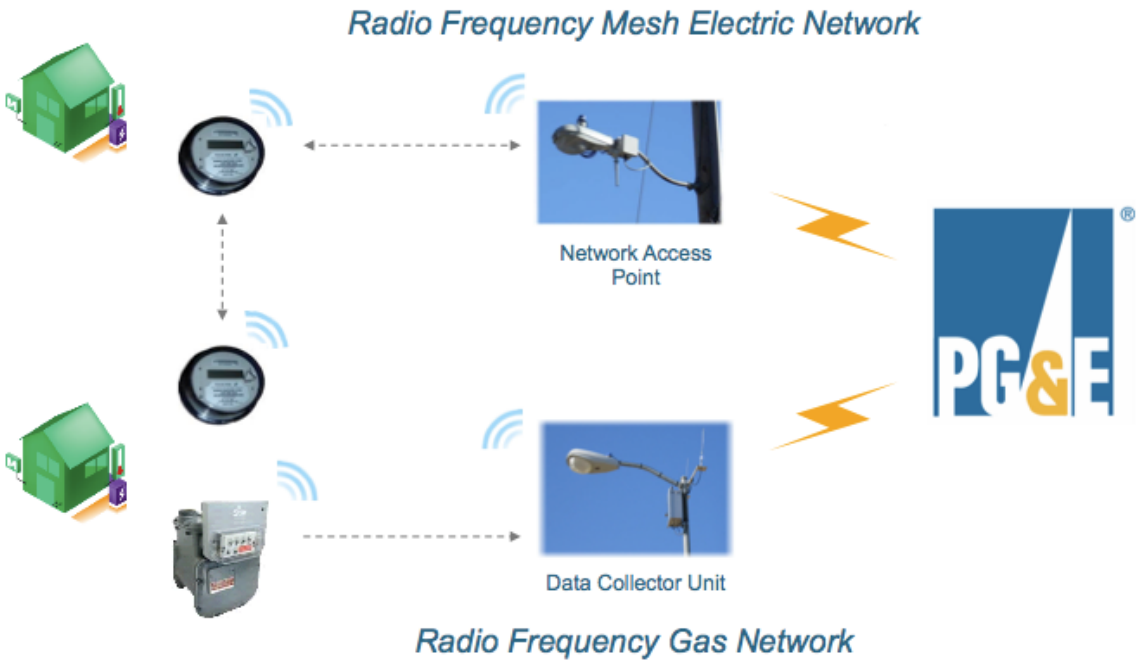
[2] Zions Bancorporation VP of Business Intelligence Clint Johnson, presentation at the 451 Group Enterprise Data Clouds forum, Feb. 24, 2010, in San Jose, Calif.
[3] IBM 2010 Global CEO Study "Capitalizing on Complexity," May 2010.

gaming. For this streaming or event processing data, individual packets may be quite modest in size but start to become "big data" when aggregated and analyzed over many days, months and years.

For example, the PG&E SmartMeter records electric and natural gas usage data from homes and businesses in 15-minute increments and communicates that to PG&E via a dual-architecture wireless sensor network. According to PG&E, the intermittent signals sent over the wireless network by each SmartMeter add up to an average of only 45 seconds a day. As PG&E SmartMeter data is aggregated for entire states and regions and is integrated into the analysis of the national energy grid, data volumes become much more substantial.

Furthermore, there is much complexity for PG&E in separating data for consumer and business accounts, providing rebates for customers who reduce their monthly usage and making energy grid decisions for individual communities and regions. All this happens while PG&E protects the confidentiality of personal nonpublic information such as social security, driver's license and credit card numbers that may be stored in account and billing records.

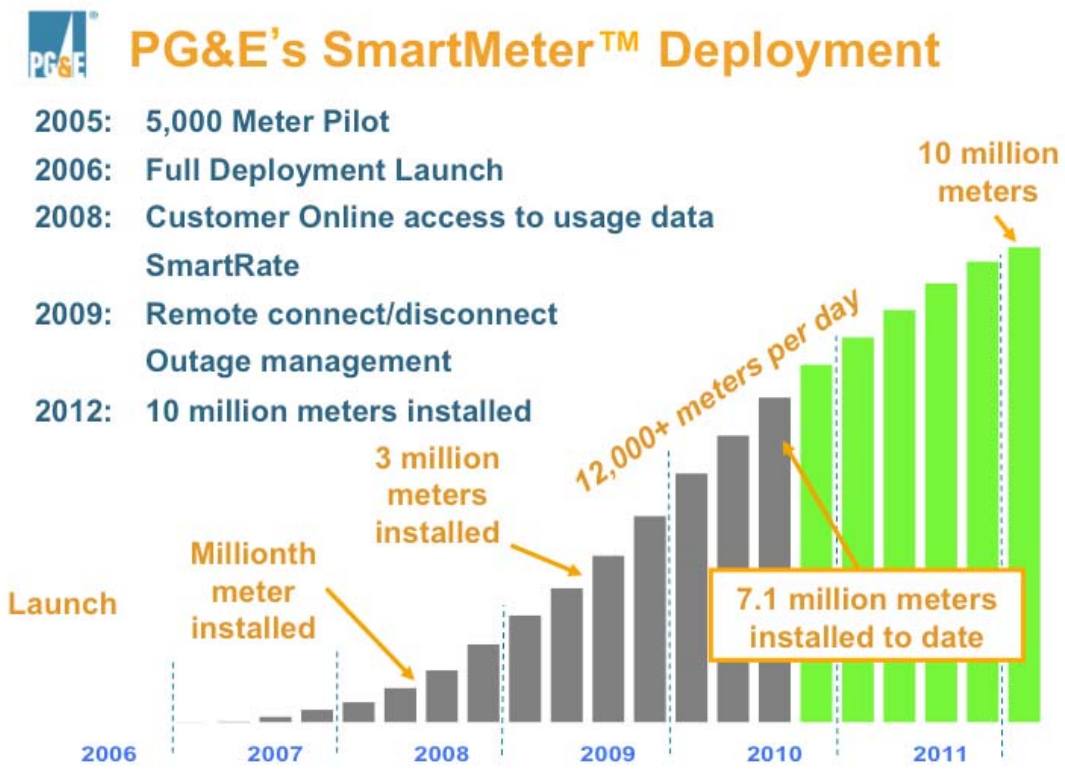
PG&E SmartMeter Wireless Network Architecture



Source: PG&E SmartMeter Program Overview, April 2009

Despite early stumbles in poor customer service and insufficient customer education for SmartMeter installations, PG&E is making progress with its large-scale SmartMeter deployment, as shown below. Forty-five seconds of communications a day for an individual SmartMeter sounds inconsequential, but multiply that by 10 million meters for multiple customer segments with both real-time and historical data and PG&E may start to face "big data" challenges.

PG&E SmartMeter Deployment Schedule



Source: PG&E SmartMeter Program All Hands Meeting presentation, December 6, 2010

Within enterprise and public sector IT networks, each network node and application generates log data. For enterprise IT, keeping track of log and sensor data is important to monitor security levels, validate regulatory compliance, address security risks and reduce operational costs. An organization facing a lawsuit or information request by a regulator may need to initiate an internal electronic discovery investigation that covers many years of communication and financial records across a myriad of file formats.

The Internet adds yet more data. Virtually every page view, ad impression and click worldwide generates log information that is aggregated, shared and analyzed by multiple players in the complex web-advertising ecosystem. For online advertising companies that use cookies or video beacons to track web activity, with every moment

and web user, more and more data is collected, filtered, organized, searched, visualized and archived. Advanced pattern recognition often involves looking not only at several weeks or months of sample data but at complex trends across years of complete data sets together with information from real-time or near-real-time event processing.

Speed

Big data has strained traditional data warehouse and business intelligence systems. Nightly batch loading is poorly suited for e-commerce, multimedia content delivery, ad targeting and other applications that occur in real time. This puts pressure on speeding up data loading at the same time that data volumes are skyrocketing. Data streaming, complex event processing (CEP) and related technologies — once mostly prevalent in financial services and government — are now emerging as requirements as part of enterprise data architectures in multiple industries.

Marketing and sales departments want a 360-degree lifetime view of the customer; they want a centralized view of the customer's retail, website, email and telephone interactions that can be pulled up immediately when a customer walks in the door or talks with a call center agent. Likewise, as more enterprises engage in social media to answer customer questions, correct misunderstandings or learn from customer feedback, responding in real time or in near real time to Facebook or Twitter updates becomes significantly more necessary.

At OmniTI, a web application and Internet architecture consulting and infrastructure provider, big data is not new. Several of the company's clients started managing terabytes of complex data 10 years ago. One new shift, though, over the past decade is that OmniTI's clients expect more immediacy in data results. For example, an online marketing firm may process billions of data points a day. What used to be a daily "rinse, wash and repeat" cycle of setting up campaigns, receiving back batch results

and altering plans, that happened over, say, 24 hours has now been compressed into a few seconds.[4]

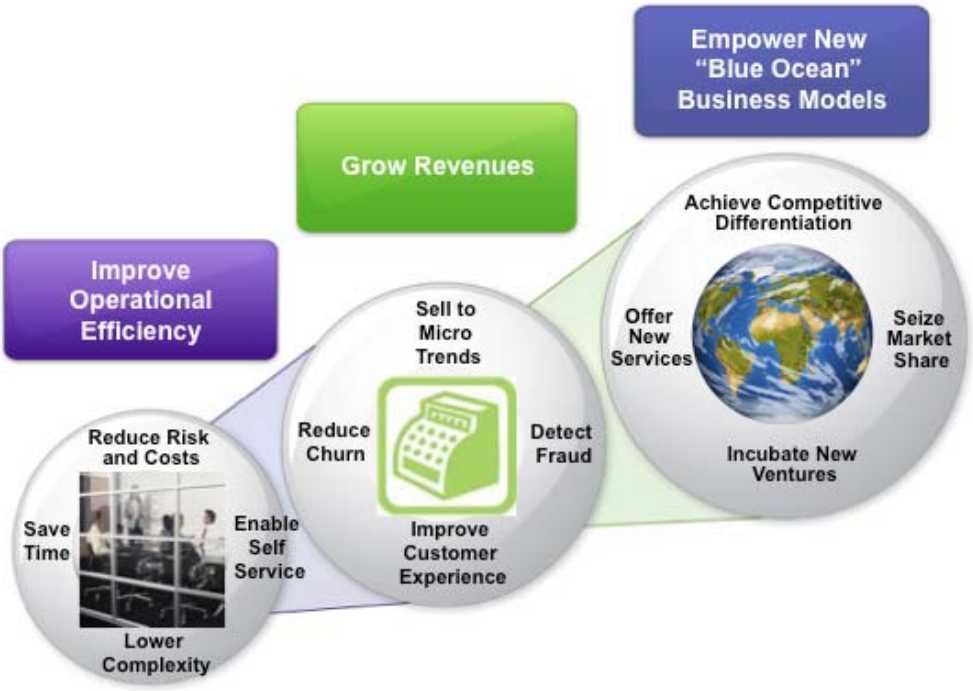
Likewise, sensor data can provide important immediate information in the event of an emergency. Making this data actionable, however, may require coordination with emergency response teams as well as automation and integration of operational processes. Following a Sept. 9, 2010, gas pipeline explosion in the community of San Bruno, Calif., PG&E staff required 89 minutes from the time of the explosion to close manual valves and stop the flow of gas. The blast killed eight people. In hearings before the National Transportation Safety Board, PG&E senior engineer Chih-hung Lee conceded that remote automated valves would have required substantially less time to close than manual valves.

PG&E also faced criticism when residents in San Bruno reported calling PG&E in the days and weeks prior to the explosion to complain of a gas leak. One of the potential public-safety applications of the PG&E SmartMeter infrastructure, when sensor data is integrated with automatic valve technology, will be to automatically shut down gas distribution in the event of a leak, versus waiting for customers to call a hotline phone number to notify the electric utility of a problem.[5] PG&E, could avoid this in the future and reap the benefits of big data's speed with SmartMeter readings every 15 minutes — thought this could require enhancements to older businesses processes and infrastructure.

As discussed in the following section, organizations that successfully adapt their enterprise architecture and processes to address these three attributes of big data — volume, complexity and speed — are improving operational efficiency, growing revenues and even empowering brand-new "blue ocean" business models.

[4] *Phone discussion with OmniTI CEO Theo Schlossnagle, March 14, 2011.*
[5] *San Francisco Examiner, "PG&E officials grilled by feds on San Bruno blast," March 1, 2011.*

Big Data Business Models



Empowerment of Business and Public-Sector Value

Large, complex data sets and requests for faster processing are becoming the norm in many enterprises and public sector organizations. Big data technologies are empowering employees, partners and customers as well as enabling better operational processes and new business models. Enterprises and public sector organizations are finding new ways to analyze, deliver and monetize information.

Improve Operational Efficiency

Self-service analytics enable customers to evaluate and pick from the multiple options available to them. For example, Amazon customers can log into their online account and make changes to open orders without phoning a call center or sending an email. And self-service BI can help an organization's customer-facing staff "to quickly drill

down on issues relevant to customer satisfaction and to identify the 'next best offer' in each customer interaction," says James Kobielus, a senior analyst at Forrester Research.[6] In other words, if a customer calls with a question or problem, analytics viewable by the customer call center staff can suggest a resolution to mention to the customer on the phone; this reduces call length and call wait times and maintains consistent customer service in the event of call center staff turnover.[7]

Grow Revenues

In addition to driving operational efficiencies, big data is helping organizations add revenue to their existing business models. In some cases, big data lets organizations preserve substantial revenue changes and challenges in their competitive environment.

Pay-television providers, for example, are beginning to customize TV ads to individual household demographics and TV viewing patterns. As cited by Streaming Media, eMarketer predicts that of the estimated $28 billion spent online in 2011, advertisers will devote about $2 billion of that total to video advertising. That's higher than the estimated spend for paid search or web banner ads.[8] Annual spending on addressable ads will reach $11.5 billion in the U.S. by 2015, according to projections from Bank of America Merrill Lynch. In a Comcast test run in Baltimore, homes receiving targeted ads changed the channel 32 percent less often than did households viewing non-targeted ads.[9] These test runs and future commercial systems create a lot of big data, as personalized TV ads and Internet-streamed video on demand are much more complex from a network perspective than showing the identical TV program and same TV ads on a fixed schedule set by traditional national and local television broadcasters.

Meanwhile, Disney has extended its existing architecture with a Hadoop cluster to improve information sharing and coordination among Disney's many departments

[6] James Kobielus, Forrester blog, "Predications and Plans for Business Analytics in 2011," Jan. 6, 2011.
[7] James Kobielus, Forrester blog, "Predications and Plans for Business Analytics in 2011," Jan. 6, 2011.
[8] Cited by Streaming Media, February / March 2011 issue, "Video Advertising 2011".
[9] The Wall Street Journal, "Targeted TV Ads Set for Takeoff," Dec. 20, 2010.

and affiliated businesses. Below is one of the architecture diagrams from an excellent case study prepared by Disney and PricewaterhouseCoopers. By pulling together diverse departmental data, most of which is stored separately by Disney's many business units and subsidiaries, data can now be analyzed for patterns across different but connected customer activities—such as attendance at a theme park, purchases from Disney stores and viewership of Disney's cable television programming.

One rationale for implementing the Hadoop cluster was the relatively low cost; Disney Principal Data Architect Matt Estes estimates that the Hadoop project cost between $300,000 and $500,000.[10] While a six-figure budget may sound extravagant for some young venture-backed start-ups, for a large enterprise with multiple business units that each generate hundreds of millions or billions of dollars of revenue, building a functional operational system for $500,000 or less is a comparative bargain. And the revenue benefits are substantially larger; even an increase of a few percentage points in additional sales through improved organization-wide analytics collaboration can in some cases deliver a disproportionately higher increase in business margins for existing sales channels. To quote from the Disney and PricewaterhouseCoopers case study: "Even in this early stage, [Disney EVP and CTO of Shared Services Bud Albers] is confident that the ability to ask more questions will lead to more insights that translate to both the bottom line and the top line. For example, Disney already is seeking to boost customer engagement and spending by making recommendations to customers based on pattern analysis of their online behavior."[11]

[10] Disney case study summarized from PricewaterhouseCoopers, Technology Forecast, Big Data issue, 2010.
[11] PricewaterhouseCoopers, Technology Forecast, Big Data issue, 2010.

Disney Use of a Hadoop Cluster for Data Aggregation from Multiple Departments and Affiliated Businesses

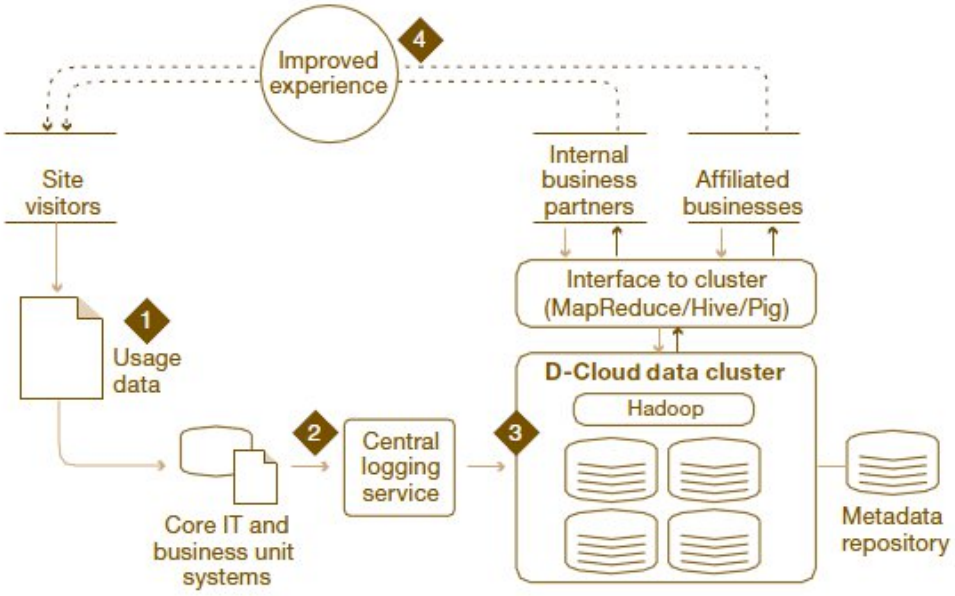


Figure 1: **Disney's Hadoop cluster and central logging service**

Disney's new D-Cloud data cluster can scale to handle (1) less-structured usage data through the establishment of (2) a central logging service, (3) a cost-effective Hadoop data analysis engine, and a commodity computer cluster. The result is (4) a more responsive and personalized user experience.

Source: Disney, 2010

Source: PricewaterhouseCoopers, Technology Forecast, Big Data issue, 2010

Empower "Blue Ocean" New Business Models

Beyond operational efficiencies and revenue enhancements, big data is enabling entirely new "Blue Ocean"[12] business models, to use a term coined by the authors W. Chan Kim and Renée Mauborgne in their book *Blue Ocean Strategy*. To quote the authors: "Blue ocean is an analogy to describe the wider, deeper potential of market space that is not yet explored."

[12] W. Chan Kim and Renée Mauborgne, *Blue Ocean Strategy*, 2005.

For example, Chaordix uses crowdsourced data to help clients in consumer-packaged goods, technology and other industries make product decisions, predict market reaction, enhance brand relevance and find problem-solving ideas. IBM uses Chaordix to power IBM's InfoGov Community, which enables big data practitioners in enterprises and public sector organizations to collaborate together and identify the best practices for data governance.

Peter Norvig, the director of research at Google, cites examples where Google stats for search terms such as "flu" and "cold relief" are helping to predict outbreaks of the flu in specific local communities days or weeks before the U.S. Centers for Disease Control and Prevention have enough data from hospitals and clinics to identify the same trends.[13] It's an example of how very large sets of immediate crowdsourced data can sometimes trump official but slower data sources.

Big Data by Industry

Gartner 2011 CIO Agenda Survey Results by Industry

Industry	2011 IT budget change		Percentage of responses		
	Weighted	Unweighted	Increasing	No change	Decreasing
Consumer, retail, media	+2.3%	+5.5%	47%	38%	15%
Education	+0.6%	+0.9%	29%	52%	19%
Energy and commodities	-2.9%	+3.7%	37%	40%	23%
Financial services	+1.0%	+3.9%	44%	42%	14%
Government	+1.0%	+1.2%	28%	51%	21%
Healthcare	+1.0%	+2.9%	38%	49%	13%
Manufacturing	+3.7%	+4.9%	51%	33%	16%
Professional services	+1.0%	+4.4%	42%	44%	16%
Telecom and technology	-0.1%	+4.1%	36%	47%	17%
Transportation and wholesale	+5.2%	+7.1%	46%	42%	12%
Utilities	-2.9%	+5.5%	44%	47%	9%
Global	**+1.0%**	**+3.9%**	**40%**	**44%**	**16%**

- Weighted IT budget change incorporates the size of the IT budget into the overall figure.
- Unweighted IT budgets are the average of each company regardless of budget size.

Source: webcast by Gartner, Dr. Mark McDonald, "Reimagining IT: The 2011 CIO Agenda," March 2011

[13] Google presentation at SDForum "Analytics Revolution" conference, held in Mountain View, Calif., in April 2010.

As organizations address unprecedented growth in big data volumes, complexity and speed, much of the work occurs within an industry context. There are commonalities in big data enterprise architectures that cross multiple industries, and organizations do seek to learn the best practices from their peers in other sectors of the economy. That said, for a variety of business, historical, legal, regulatory, technical and competitive reasons, there are some patterns that do distinguish how leading enterprises in their respective industries are managing and benefiting from big data.

Financial Services

Financial service companies across the globe mine and analyze big data to stay ahead of the competition, improve retail customer service, detect fraud, validate regulatory compliance and maximize operational efficiencies.

Regulatory oversight has increased following the subprime mortgage meltdown and its resulting impact on government's role as a "lender of last resort" in loans and bailouts. And financial service companies are preparing to comply with limitations on credit and debit card fees and other consumer protection measures while at the same time meeting the capital and risk-weighted asset measures that will come with the new Basel rules.

In credit card processing and global payments, business and regulatory requirements include: accepting local payment types, pricing in local currencies with dynamic currency conversion, fraud prevention and detection, protection of consumer nonpublic information and measures to counter drug money laundering and terrorist financing.

Fidelity National Information Services (FIS) uses big data analytics to detect credit card fraud. It sells credit card risk management and fraud detection services, backed by an analytics infrastructure that includes ParAccel's column-oriented Analytic Database. The nature of what FIS analysts do is highly ad hoc and interactive. They

frequently run complex queries correlating multiple activities in different data sets, to stay one step ahead of credit card thieves. As new methods of fraud are detected, they are encoded into their company's search algorithms and the operational systems that accept or decline a credit card transaction in real time. According to ParAccel: "With PADB, FIS can engage in two-way 'conversations' with [its] data to optimize detection for its customers, while minimizing impact on legitimate clients."[14]

Even the U.S. Securities and Exchange Commission (SEC) struggles to cope with data overload. Several days after the May 6, 2010, "flash crash," when the Dow Jones Industrial Average plummeted nearly 1,000 points in just minutes, SEC Chairwoman Mary L. Schapiro stated in testimony to the U.S. House Committee on Financial Services that "the technologies used for market oversight and surveillance have not kept pace with the technology and trading patterns of the rapidly evolving and expanding securities markets."

Enterprises in financial services and accounting are among those most impacted by the U.S. Sarbanes-Oxley Act of 2002 (SOX). SOX mandates that senior executives take individual responsibility for the accuracy and completeness of corporate financial reports. It also puts in place new reporting requirements for off-balance-sheet transactions, pro-forma figures and the stock trades of corporate officers. As a best practice, financial services and other organizations have used SOX compliance requirements to improve risk management, standardize IT architectures and drive operational efficiencies. Forrester analyst Chris McClean notes that SOX "has had a good impact on the way companies look at risk management and consider risk management in a lot of their business decisions."[15]

[14] *ParAccel Fraud Analytics case study and Brett Sheppard discussion with ParAccel CTO Barry Zane and CMO Tarun Loomba, Feb. 28, 2011.*
[15] *Quoted in Computerworld, June 29, 2010.*

Health Care

Large, complex data sets are becoming the norm in health care organizations. Some drivers of this data growth include the advent of electronic medical records, advances in medical imaging, genetic research and the use of huge databases in pharmaceutical studies. By applying data mining tools to data sets from a large number of patients, medical researchers are pinpointing causes of diseases and options for prevention, diagnosis and treatment.

The U.S. Health Insurance Portability and Accountability Act (HIPAA), first passed in 1996, requires health insurance portability for workers and their families following a change or loss of job. In addition, the law has important impacts for data management: It requires the establishment of national standards for electronic health care transactions and sets standards for the security and privacy of health data.

Electronic health care records are designed to include the patient's complete medical history, going back years and, in some cases, decades. These records may include a whole range of data in comprehensive or summary form, such as medical history, medications and allergies, immunization status, test results, radiology images, demographics and, of course, billing information.

Within appropriate group- and role-based security to maintain the privacy of patient records, leading health care organizations are enabling rapid access to a patient's medical records, including digital imagery such as x-rays and CAT scans, within a distributed network comprising large urban hospitals and remote medical clinics.

With 621 hospital beds and 60,000 in-patient hospitalizations a year, Harvard Medical School's teaching hospital, Beth Israel Deaconess Medical Center, created a private cloud of archival storage and started 2010 with 175 terabytes of storage. The

center developed a cloud-based community image–sharing solution that enables immediate search and does not require local storage to exchange images among their affiliated clinicians.[16]

Kaiser Permanente

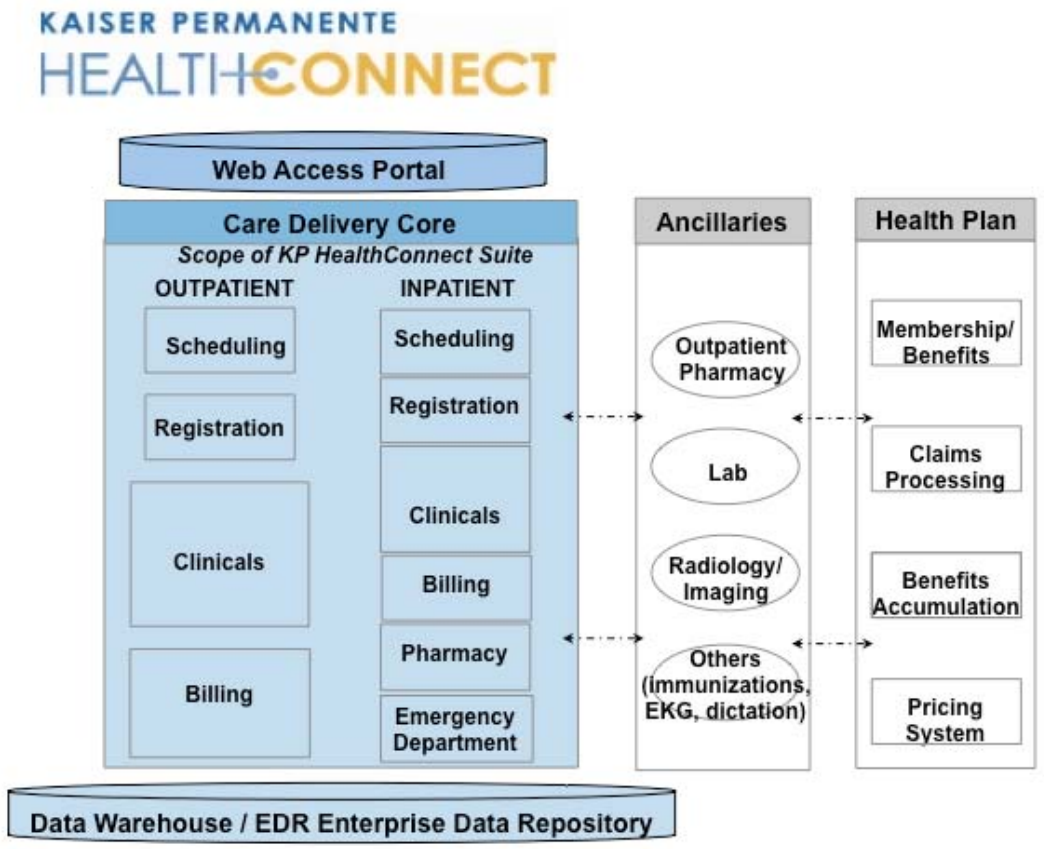
For Kaiser Permanente and its more than 8 million members, big data is about improving the quality of care and reducing costs. Kaiser standardized an electronic health record (EHR) system for all of its 36 hospitals and more than 400 medical offices. Kaiser's data architecture includes:

- Electronic health care records software from Epic Systems
- SAP BusinessObjects and Crystal Enterprise reporting
- SOA-based application and service development
- Data warehouses and marts including Oracle 9i/10g, SQL Server and Teradata
- Informatica PowerCenter for data integration
- Data center outsourcing services from IBM

Kaiser's big data is multidimensional: Inpatient, outpatient, pharmacy, finance, cost management and other groups at Kaiser use decision-support software to improve the quality of care and reduce costs. These departments need to analyze many factors at the same time: treatment; demographics, such as age and sex; lab test results; prescriptions; diagnosis; medical plan; and payment records. Integrating all this disparate information together, Kaiser's decision-support software helps doctors and nurses understand the patient's complete history and choose the best course of care.

[16] John D. Halima, "Life as a Healthcare CIO" blog, March 2010.

Kaiser Permanente HealthConnect Multidimensional Data



Source: presentation by Robert Crane, Institute for Health Policy, Kaiser Permanente

The Kaiser Permanente Center for Health Research maintains a virtual data warehouse with a series of standardized files. Content areas and data elements that are commonly required for research studies are available. Data dictionaries, meanwhile, are created for each of the content areas, specifying the format of each of the elements, such as variable name, variable label, extended definition, code values and value labels.

The results today of the Kaiser electronic health care implementation come after many years of investments, perseverance and, at times, missteps and restarts. While technology has been part of the challenge, people processes across multiple hospitals and clinics were a major factor in the ultimate success of the EHR implementation

organization-wide. As noted in a case study of lessons learned: "Good communication among many different parties and a supportive team spirit from local interested parties are necessary to facilitate the building and maintenance of a high quality research data structure."

Sports

During the 2011 National Football League (NFL) playoff TV broadcasts—amid commercials with recording artist Eminem and auto racing driver Danica Patrick—there was an ad with an IBM researcher talking about data analytics. While at first glance NFL TV broadcasts may seem an unusual forum for a discussion of data analytics, information management and analysis play an important role in professional sports. As discussed at the MIT Sloan Sports Analytics Conference held earlier this month, teams prioritize draft choices based on sophisticated criteria, scrutinize game video to look for opponent weaknesses and reference statistics to guide play-calling. Chicago-based EXACT Sports, for example, makes sports science its core business: It provides behavioral assessments, sports psychology tools and other analytics-enabled aids for player development.

In world football, A.C. Milan works with Milan Lab and Microsoft business intelligence software for sports injury prevention and treatment. Milan Lab founder Jean-Pierre Meersseman and his team perform sophisticated tests and analysis to determine physical wholeness, including equilibrium, endurance and coordination. For example, in a "dynajump" drill, they test a player's jump to measure the angles of the knees with electromyography hooked up to the muscles. By combining the dynajump test results with neural networking, they can predict an injury rate with an estimated 70 percent accuracy—significantly higher than previous evaluation rates. Milan Lab tests the players each fortnight, and cause-and-effect data is gathered each day. According to A.C. Milan, in the first full season of working with Milan Lab, days lost to player injuries declined by two-thirds. Milan Lab uses Microsoft BI software to store and analyze the player data.[17]

[17] *Mail Online, "A.C. Milan's ageing starts," March 9, 2011.*

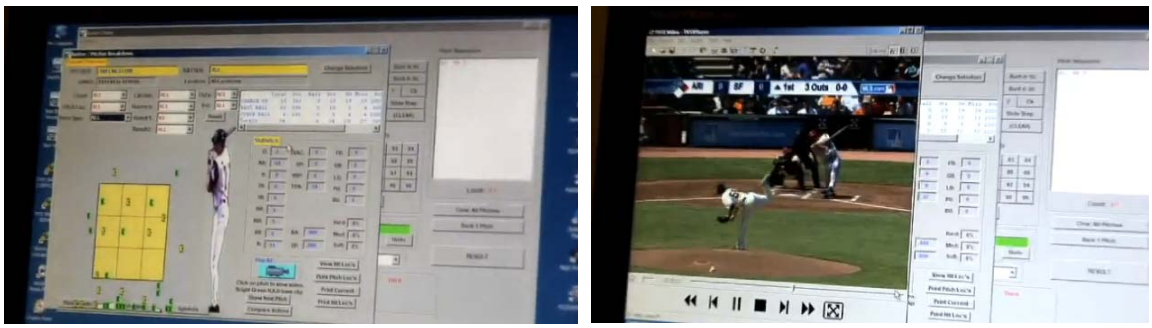
TeamRankings.com, based in San Francisco, develops data-driven sports predictions, rankings and analysis tools. They use math and technology to help sports fans excel in prediction-based contests and markets, from March Madness brackets to NFL survivor pools to sports betting. They deliver content directly via their website and via distribution partners such as ESPN Insider.

San Francisco Giants

Video analytics has become a major behind-the-scenes component of team and individual preparation for each major league baseball (MLB) game for all the clubs, including the 2010 World Series–winning San Francisco Giants.

SF Giants Senior Vice President and Chief Information Officer (CIO) Bill Schlough values the contributions made by advanced analytics for his team on the field and in the box office. In a preview video for his upcoming keynote talk at the March 25, 2011, SDForum "Analytics: The Next Wave" conference, at Stanford, Schlough highlighted the roles that advanced analytics play for two key areas for the Giants' franchise: (1) video coaching and scouting tools and (2) dynamic pricing for game tickets.

Video Coaching System Employed by SF Giants Players and Coaches



Source: screen shots from NetApp promotional video with the San Francisco Giants, March 15, 2010

From a player's perspective, video coaching tools are one of the most visible and often-used big data applications. The SF Giants clubhouse, at AT&T Park, includes multiple stations where players can sit before the game or during a game and access footage from various camera angles. Hitters can compare and contrast different at bats against

the same pitcher or in a time sequence over several days or weeks. Capabilities include side-by-side views to compare successful at bats with poor ones. And pitchers can plan their strategy for each hitter.

This video coaching system uses NetApp systems for storage. It already exceeds 10 terabytes of storage; growth is planned to 100 terabytes or more to meet future requirements. The SF Giants were the first MLB team to install an all-digital video system. In previous decades, teams used VCR tapes that involved manual tape changes and slow fast-forwarding and rewinding. Now every MLB team has an all-digital video system for both hitting and pitching evaluations.[18]

With annual revenues in excess of $200 million, the SF Giants staff a large organization off the field, although the team's revenues represent about half of the close to $400 million earned annually by the New York Yankees and the Boston Red Sox. Marketing partnerships form part of the equation in the choice of technology vendors, which keeps the SF Giants' capital technology investment budget between $5 million and $15 million a year, depending on which new IT projects the team takes on that year.[19]

For box office sales and online ticket sales at the team website, the SF Giants have employed dynamic pricing, using software for Qcue, a company based in Austin, Texas. This dynamic pricing for game tickets is not unlike the variable pricing used for years in the airline industry (although thankfully without fees for bringing carry-on bags to baseball games). Dynamic pricing uses advanced analytics to adjust prices on the fly. For years, teams may have charged more for games against a division rival, based on an annual schedule.

With dynamic pricing, if a player goes on a potential record-breaking hitting streak or the weather is particularly nice outside or two teams are neck and neck in the standings, pricing for the available tickets to remaining games is adjusted accordingly.

[18] *Phone discussion with NetApp Office of the CTO Strategic Planning Team head and "NetApp Cloud Czar" Val Bercovici, March 9, 2011.*
[19] *Ben Thompson, "A Whole Different Ballgame," Business Management magazine, July 4, 2010.*

Qcue helps identify opportunities for markups for prime seating or popular games, as well as in some cases price discounts in order to fill less desirable seats in the upper reaches of the stadium or distant outfield bleacher seats. With dynamic pricing, the cost of the ticket directly reflects the value based on demand.

Travel

Expedia, Kayak, Orbitz, Priceline, Sabre Holdings/Travelocity, TripAdvisor and others compete aggressively for travel bookings, cross-sell with rental car reservations and unique website visitors.

While there are competitive differences in the companies—Expedia, including its Hotels.com subsidiary, does particularly well with hotel room bookings, while Orbitz earns the majority of its revenue from airline bookings—there is substantial overlap, too, among these Tier 1 online travel leaders. These overlaps in business model among the travel website companies increase the importance of big data infrastructure and analytics as the online travel websites compete for consumer and business travel wallet-share. For example, as highlighted in one of the case studies below, while Orbitz is not a hotel specialist, it uses a Hadoop infrastructure to deliver hotel recommendations for online hotel bookings on the Orbitz travel website.

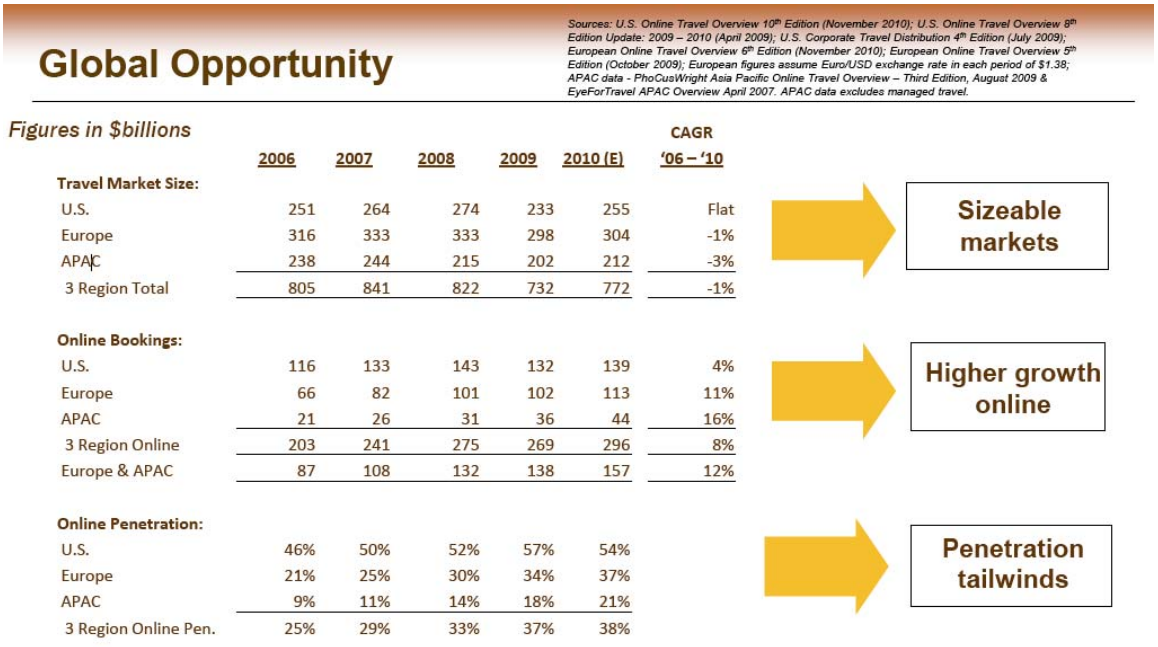
In terms of unique monthly website visitors, Expedia remains the market leader, but the gap has narrowed since mid-2009 (see below). Expedia's market weight is likewise reflected in its revenues; Expedia earned $3.3 billion in 2010, backed by total Expedia bookings that year of $26.0 billion.

Comparison of Travel Website Unique Visitors

Source: Compete.com, 2010

According to data compiled by Expedia (shown below), online travel bookings in 2010 totaled $296 billion, out of an estimated total travel (hotel plus airfare plus rental car) spend last year of $772 billion, giving them a 38-percent share of the online bookings.

Growth in Online Travel Bookings out of the Total Travel Bookings Marketplace



Global Opportunity

Sources: U.S. Online Travel Overview 10th Edition (November 2010); U.S. Online Travel Overview 8th Edition Update: 2009 – 2010 (April 2009); U.S. Corporate Travel Distribution 4th Edition (July 2009); European Online Travel Overview 6th Edition (November 2010); European Online Travel Overview 5th Edition (October 2009); European figures assume Euro/USD exchange rate in each period of $1.38; APAC data - PhoCusWright Asia Pacific Online Travel Overview – Third Edition, August 2009 & EyeForTravel APAC Overview April 2007. APAC data excludes managed travel.

Figures in $billions

	2006	2007	2008	2009	2010 (E)	CAGR '06 – '10
Travel Market Size:						
U.S.	251	264	274	233	255	Flat
Europe	316	333	333	298	304	-1%
APAC	238	244	215	202	212	-3%
3 Region Total	805	841	822	732	772	-1%
Online Bookings:						
U.S.	116	133	143	132	139	4%
Europe	66	82	101	102	113	11%
APAC	21	26	31	36	44	16%
3 Region Online	203	241	275	269	296	8%
Europe & APAC	87	108	132	138	157	12%
Online Penetration:						
U.S.	46%	50%	52%	57%	54%	
Europe	21%	25%	30%	34%	37%	
APAC	9%	11%	14%	18%	21%	
3 Region Online Pen.	25%	29%	33%	37%	38%	

Sizeable markets

Higher growth online

Penetration tailwinds

Source: Expedia.com, Q4 2010 company overview

Airfare distribution is built on schedules published through OAG Aviation, fares published through ATPCO and availability data housed on airline passenger reservation systems. According to an informative article published last year by Norm Rose at Travel Tech Consulting, what ITA Software has in fact built is a "next generation airline passenger reservation system" through the real-time, cloud-computing-based linking of these three data sets. If the Google acquisition of ITA Software is approved, Google/ITA may compete in online travel meta search.

The Google acquisition of ITA Software remains pending as of March 2011, as the U.S. Department of Justice reviews antitrust implications. Multiple travel website companies that are otherwise competitors—including Expedia, Kayak and Travelocity—have joined one another under a lobbying and marketing collation called FairSearch.org to oppose the acquisition, on the grounds that they would lose business due to potential integration between Internet search and ITA's infrastructure and services. In addition to the incumbent provider GDS, there is also another start-up in that space, Vayant, but it has not to date attracted the sizeable customer base that ITA

Software has.

Microsoft Bing is not an online travel website—at least not yet. But, as pointed out in the USA Today Travel section earlier this month, Bing does store historical airfares and provides tools for forecasting fares for upcoming travel, to help consumers determine prices and the best times to purchase airline tickets. This provides part of the infrastructure that Microsoft Bing could use to enter the online travel industry. However, part of the Bing infrastructure for travel is based on ITA Software, which helps explain Microsoft's participation in the FairSearch.org coalition. Microsoft would prefer not to be dependent on a competitor for travel search results such as airline fares and hotel room availability, which are displayed via the ITA Software system on Microsoft Bing search results.

ITA Software

A group of computer scientists from MIT founded ITA in 1996. ITA delivers air travel pricing and inventor management capabilities for several dozen airlines and online travel websites. ITA Software offers a Matrix Airfare Search capability that Kayak, Orbitz and others customize through an XML interface with their choice of configuration, company logo and layout; they embed this matrix in their websites to help consumers pick the best flight schedule and travel price for their needs. ITA developed an alternative to GDS for a scalable high-volume online system for airfare pricing, shopping and availability management. The ITA Dynamic Availability Pricing System (DAPS) processes more than 1 million queries per second.

The online travel industry has quite complex requirements. Capacity is limited: There are a fixed number of commercial flights and seats on each flight a day. When a consumer selects a flight and seat, the airline management inventory must be updated immediately, to avoid another consumer making a duplicate purchase. Likewise, if that consumer ends up changing or canceling their flight travel plans, their chosen seat must then be shown as available as quickly as possible, to avoid a potential revenue loss for that airline.

Example of an ITA Travel Search Application for Mobile Phones

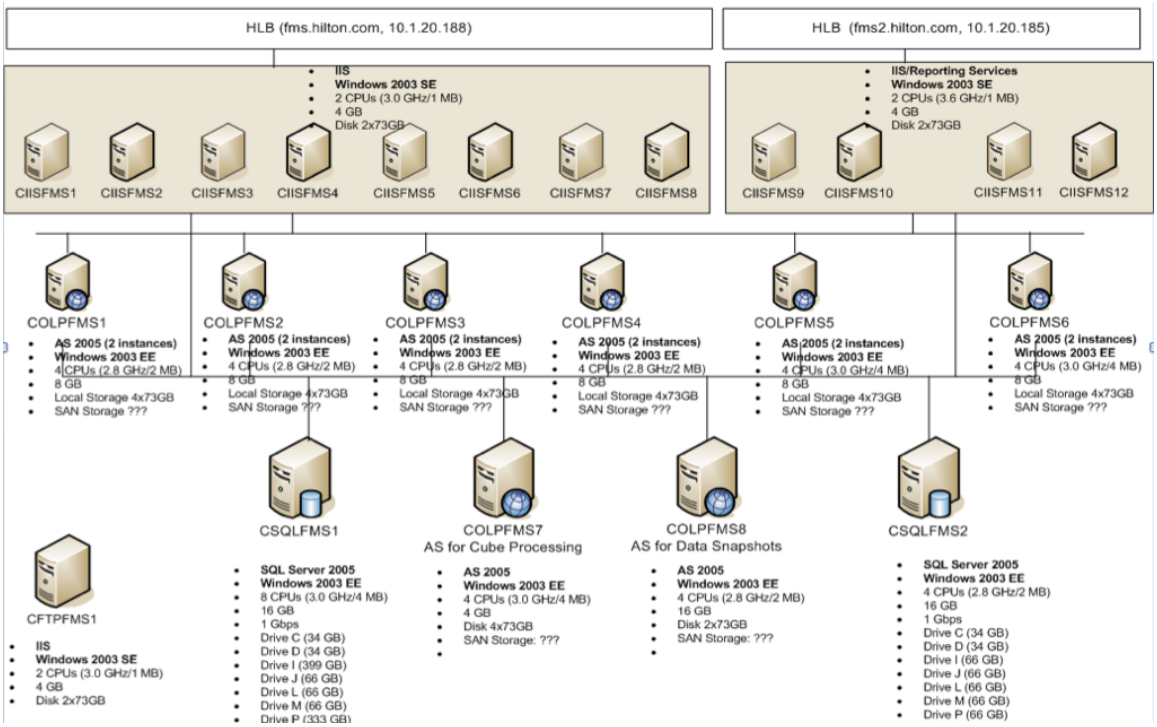
ITA Software has adapted its technology and expertise in high-volume, concurrent management of complex data sets to offer services for data acquisition, integration, cleansing, analyzing and publishing, in a service portfolio called Needlebase.

Hilton Hotels

Hilton Hotels uses a room forecasting system based on Microsoft SQL Server products, including SQL, AS, IS and RS, together with Hyperion Essbase for implementation and the development of data extracts from source systems, load rules, advanced calculation scripting, application tuning and report writing. Hilton's Brands & Commercial Services Center, IT and Operations Support Groups are based in Memphis, Tenn.

They use scale-out AS and RS running on IBM xSeries and IBM Blade Center servers, with load-balanced analysis services reader machines. Shown below is Hilton's load-balanced SQL server production architecture, current as of 2005. Each RS server supports 40 to 50 concurrent users and delivers complex queries to many clients based on large data sets.

Hilton Forecasting Management System (FMS) Production Architecture



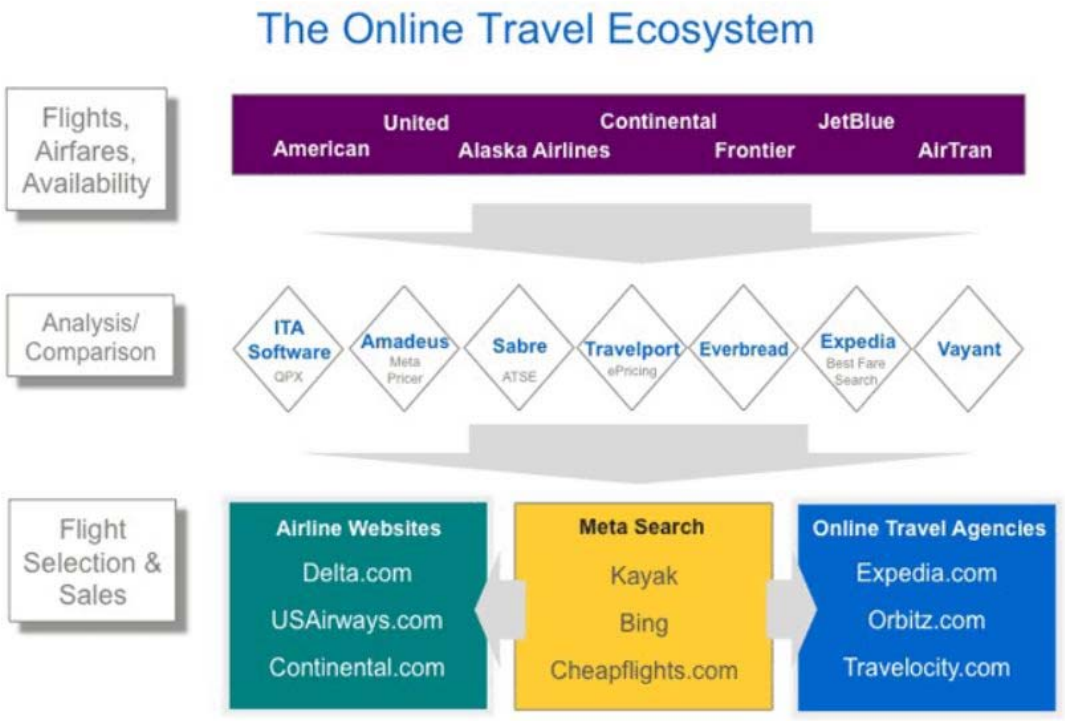
Source: Microsoft VLDB SQL Server Products Customer Advisory Team, presentation at Denver SQL Users Group

Using this SQL Server architecture, Hilton hotel-level budgeting and forecasting system (HLBFS) analysts provide daily and weekly EBITDA reporting for hundreds of hotels in the Hilton corporate family. These include Hilton, Hilton Garden Inn, Hilton Suites, Waldorf Astoria Collection, Embassy Suites, Doubletree Hotels, Doubletree Guest Suites, Doubletree Club, Hampton, Hampton Inn & Suites, Homewood Suites and independent hotels.

These levels of the online travel industry ecosystem occur in real time in parallel, with changes in one level impacting the others, similar to the fictional multilayered structure of dreaming portrayed in the *Inception* movie release last year. In the following visual, Google outlines its view of this online travel ecosystem. Given ITA Software's impressive customer base, this specific chart from Google–produced as part of Google's campaign for regulatory approval to acquire ITA–may understate ITA

Software's market share at that level of the online travel ecosystem. That said, the rest of the diagram paints an interesting picture of the multilayered, interdependent ecosystem for online travel.

View of the Online Travel Ecosystem from Google's Perspective



Source: Google presentation on proposed acquisition of ITA Software, July 2010

Web 2.0: Consumer, Retail and Media

The "consumerization of IT" continues to be a major trend in 2011. Organizations are seeking to implement collaboration software for employee and partner use that is modeled after social networking sites such as Facebook, LinkedIn and Twitter. After years of gradual enterprise adoption, video conferencing is starting to take off as consumers bring video webcam-enabled devices and knowledge of their use into the workplace. Online marketplaces, eBay, for instance, have inspired entrepreneurs to launch a new generation of industry-tailored B2B marketplaces. Even for an "old-

school" retailer such as Macy's, online sales are growing as a percent of total revenue, and innovations in big data management are enabling fine-tuning sales approaches. And advertising related to online video accounts for a rising percent of total advertising spending. Streaming online video is upending the traditional broadcasting industry and providing mobile content for smartphones and mobile PCs.

Netflix

Netflix uses computing capacity on Amazon Web Services, specifically AWS SimpleDB and S3, for encoding video for three-screen distribution: PC, mobile and TV. For Netflix, AWS provides better availability and scalability than would building out and managing more data centers. (As of late 2008, Netflix had a single data center.) Instead of becoming a data center manager, Netflix chose to focus on its core competency to deliver movies, TV shows and other high-quality video content.[20]

SimpleDB and S3 provide the following:

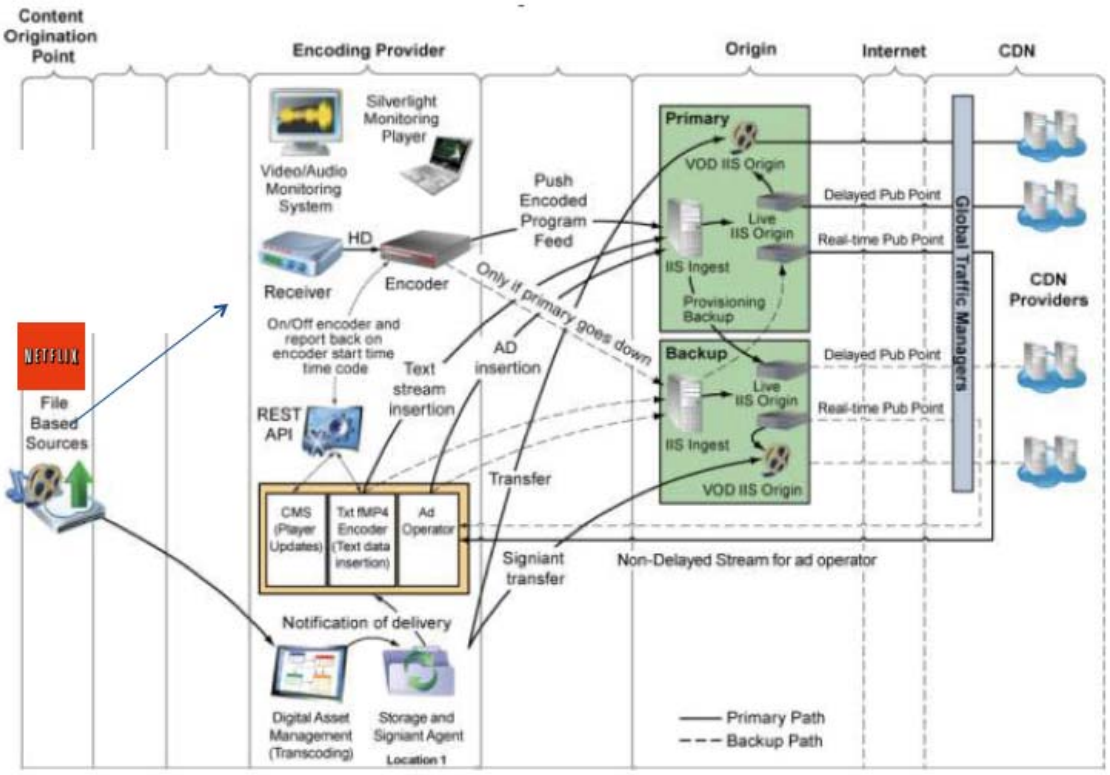
- Disaster recovery
- Managed fail-over and fail-back
- Cross-zone distribution
- Persistence
- Eventual consistency
- Support for the SimpleDB subset of SQL

According to Netflix's Sid Anand, "SimpleDB's support for an SQL-like language appealed to our application developers as they had mostly worked within the confines of SQL."

For video streaming, Netflix uses Microsoft adaptive streaming together with content delivery network (CDN) capacity and services from Level 3, Akamai and other CDNs.

[20] *Presentation and white paper by Sid Anand at Netflix, October 2010 and March 2011.*

Netflix Internet Content Delivery Using Microsoft Streaming together with CDN Partners



Adapted from Microsoft, Senior Product Manager Chris Knowlton, Streaming Media West 2010, November 2-3, 2010

Big Data Technologies

On the technology side, big data is contributing to and benefiting from innovations across enterprise IT architectures and multiple layers of the technology stack. The following is a partial list of examples:

- **Services:** infrastructure as a service; platform as a service; public, private and hybrid clouds; software as a service; implementation and consulting services; and data science advanced analytics services
- **Applications:** analytics; BI software; collaboration software; data quality tools;

social media; spreadsheets; and visualization tools

- **Middleware:** data integration tools; data management tools; data modeling tools; extract load transform (ELT); and service-oriented architecture (SOA)
- **Infrastructure:** application and network accelerators; complex event processing (CEP); data warehouses; data marts; distributed data stores; Gigabit Ethernet networking; memcached; operational data stores; and server, storage and desktop virtualization
- **Data sources:** call data records; EHRs; e-commerce; security event management; sensor networks; and web analytics
- **Databases:** columnar databases; document-oriented databases; Hadoop/MapReduce; and NoSQL derivatives from Amazon Dynamo and Google Bigtable, among others

At each level in this big data technology stack, there are important developments and innovations under way this year, as profiled below.

Data as a Service

A crop of innovative data as a service companies have emerged; they offer access to public and/or private data sets, along with data integration, reporting and other functions. Companies include:

1. **DataMarket** enables users to search, visualize, compare and download data from multiple providers, including Eurostat, Gapminder, the UN and World Bank. The company offers a combination of free and premium data sources.
2. **Factual** offers a data mashup service where users can combine data sets on any subdget. It provides toolsets to help the community build and maintain a trusted source of structured data.
3. **Google**'s Google Public Data is a project in Google Labs. It's value proposition is to make it easier to explore, visualize and communicate data. Google also acquired Freebase, to provide "a Wiki for structured data." Freebase offers a

Creative Commons-licensed collection of openly available structured data and a platform for accessing and manipulating that data via an API.

4. **Infochimps** sources data from public and for-fee data sets, and allows customers to query it or obtain the data in a variety of formats, including mailing lists or spreadsheets. Its aim is to become a search engine for data sets.
5. Microsoft envisions its **Windows Azure Marketplace DataMarket** as a data publishing service, with integrations to PowerPivot and Excel.
6. **Socrata**'s OpenData appeals to local governments and smaller enterprises that want to obtain a hosted solution for social data discovery.
7. **Timetric** offers services to publish, manage and enable social analysis and visualization of large collections of time-series data.
8. **WolframAlpha** offers, as an extension of its Mathematica technical computing software suite, the ability to undertake dynamic computations based on a built-in analytics library.

Middleware

Software tools can aid the aggregation and analysis of huge data volumes, such as IT logs and security event data. Splunk helps enterprises and public sector organizations analyze live-streaming IT data along with terabytes or more of historical IT data. Within the Hadoop open-source ecosystem, Chukwa can help for large-scale log collection and analysis.

The HP acquisition of ArcSight leads the Gartner Magic Quadrant for Security Information and Event Management (SIEM); SIEM can reduce management overhead of monitoring multiple point devices, as well as identify patterns that may only be evident when comparing data from multiple security, network or application layers. For example, comparing alerts from both anti-virus software and virtual private network nodes might identify a coordinated hacker attack that is less noticeable when viewing logs from only one network element.

Infrastructure

Flash memory can help speed mission-critical applications. Fusion-io provides flash-based, solid-state storage application accelerators to HP, IBM, Dell and others. For example Fusion-io customer Answers.com achieved an 8x improvement in disaster recovery backup times. Another customer of Fusion-io's flash-based accelerators, messaging security provider Cloudmark, improved database replication performance by five times.[21]

Memcached servers can alleviate performance bottlenecks as a component within tiered web architectures, where they are deployed on web or application servers or independently alongside a traditional database layer. As traffic demands increase, web applications can scale out horizontally by increasing the number of web servers and caching servers, alleviating database load and improving web application performance.

Databases

Apache HBase, a NoSQL database modeled from Google's Bigtable system, is part of the Hadoop ecosystem and offers specific advantages for row-level access. Guy Harrison has a written a good case study, "Real World NoSQL: HBase at Trend Micro."

CouchDB is a top-level Apache Software Foundation open source project. It has been running in production environments for three years, and reflects a document-oriented, schema-free database model. In other words, instead of sorting data into tables or columns, CouchDB houses data as documents. CouchDB uses a JavaScript-based view model to aggregate structures and produce reports. It has some similarities to Lotus Notes. For an intro to CouchDB, Joe Lennon, a software developer at Core International, has a good IBM developerWorks article.

Software projects inspired by Amazon Dynamo and Google Bigtable — including Cassandra (open-sourced by Facebook and now an Apache project, with commercial

[21] Fusion-io presentation at Accel Partners, New Data Stack conference, July 2010.

services from DataStax and others), Riak and Project Voldemort — can enable structured storage systems to scale horizontally, preserve states if machines fail and scale data writes.

Revolution Analytics, headquartered in Palo Alto, Calif., offers advanced analytics software using open-source R statistics software. Its software supports XDF files, a new binary big data file format with an interface to the R language that provides high-speed access to rows, blocks and/or columns of data.

In addition to these open-source tools, there are a variety of proprietary software applications that offer alternatives to traditional relational databases. For example, MarkLogic offers a purpose-built database for unstructured information for organizations in publishing, government, financial services and other industries. MarkLogic customers range from publishers Simon & Schuster and Pearson Education to aviation's Boeing and the U.S. Federal Aviation Administration.

Open-Source Options: Hadoop and More

Hadoop traces its origins to a landmark paper on MapReduce by Google in December 2004, which inspired Doug Cutting at Nutch to team with colleagues at Yahoo. Cutting (now at Cloudera) named Hadoop after his young son's stuffed elephant. MapReduce takes an application and divides it into multiple fragments of work, each of which can be executed on any node in the cluster, and Hadoop Distributed File System (HDFS) stores data on nodes in the cluster with the goal of providing greater bandwidth across the cluster.

Multiple organizations now package distributions of Hadoop together with related software tools. Hadoop distributions include Apache Hadoop, Amazon Elastic MapReduce (EMR), Cloudera, and IBM.

Cloudera has developed a training package that helps enterprises that are not specialists in HDFS and MapReduce to benefit from Hadoop-driven analytics. Cloudera partner Karmasphere offers front-end client software that enables developers and analysts to develop, debug and deploy Hadoop jobs to private, public or hybrid Hadoop clusters. Another Cloudera partner, Datameer, enables business users to employ a familiar spreadsheet interface to access and analyze big data volumes stored in Hadoop.

Despite recurring "SQL vs. NoSQL" technical debates, many organizations will continue to use a combination of relational databases, Hadoop, column stores such as ParAccel or Vertica and other data architecture approaches. The focus is "the right tool for the job," versus the big data equivalent of trying to use a hammer to install screws when what you really need is a screwdriver.

In-database Analytics

There is growing progress with in-memory systems that can enable faster responses for analytic queries and operational business intelligence. IBM SPSS, KXEN, SAP and SAS, among others, offer advanced analytics programs and tools that can run in-database. By pushing complex analytics computations closer to the data, processing times are faster, and organizations require less server and network infrastructure to move data between storage and memory. To accomplish this, an increasing number of enterprises and public sector organizations are executing predictive analysis, data mining, and other computation-intense big data applications within their data warehouses.

Depending on the platform, business analysts and statisticians who may not be experts in writing complex SQL can work in R, S-Plus, Predictive Modeling Markup Language (PMML), Eclipse-based integrated development environments (IDEs), Hadoop, service-oriented architectures (SOA), or other interfaces / programming languages to execute in-database analytics.

Database analytics are becoming "table stakes" in 2011. In-database advanced analytics are extending beyond existing adopters — such as marketing departments at finance companies — to other industries such as retail, through the work of IBM Netezza partners that develop industry-specific solution architectures based on the IBM Netezza platform.[22]

SAS and Teradata, meanwhile, are showing signs of progress with their in-database analytics collaboration. With this integration, SAS users avoid the costs and time delay of using a Teradata data warehouse and loading the extracted data into a separate system for analysis. SAS embeds some core SAS procedures into the Teradata database, and integrates a library within the Teradata database to process SAS formats. While Teradata CRM competes with SAS, and the overlap between the two companies is not great, it seems, overall, to be an effective partnership.

SAS and Teradata's joint customer, Cabela's, received a NCDM Gold Award for database marketing. Using in-database analytics, business analysts at Cabela's can better understand customer buying patterns in various locations, including brick-and-morter stores, catalogs and online. Analysts can develop marketing models faster, spend less time building the data and run queries faster, according to a SAS and Teradata case study.

SAS competitor SPSS offers an extensive range of in-database analytics capabilities: predictive analytics and data mining, for complex structured and unstructured data; support for a variety of data warehouse and DBMS platforms; and increasing integration with new parent company IBM's Cognos, InfoSphere and DB2 software.

KXEN is also a popular predictive analytics and data mining tool among business analysts and consultants. One of its partners is Teradata; KXEN taps the Teradata database to identify business operation drivers and build enterprise models. Data

[22] *Phone discussion with IBM Netezza Phil Francisco and Michele Chambers, April 2010.*

mining staff can then use the Teradata Warehouse Miner to build and deploy production data models based on the KXEN results.

Organizations using Oracle can integrate with MapReduce through Parallel Pipelined Table Functions; Oracle senior principal product manager Jean-Pierre Dijcks has a good blog post that explains how.

Following a single-vendor integrated stack strategy, Oracle's primary in-database value proposition focuses on: in-database mining in the Oracle Database kernel, together with data-preparation tools within the Oracle DBMS; the ability to analyze large volumes of text and other unstructured information; and support for statistical algorithms and variable selection techniques.

In addition to SAP Sybase support for MapReduce, Sybase has ported Fuzzy Logix DB Lytix statistical and predictive analytics algorithms library on Sybase IQ using an in-database analytics API (application programming interface). DB Lytix offers the ability to perform advanced in-database analytics through simple SELECT and EXECUTE statements.

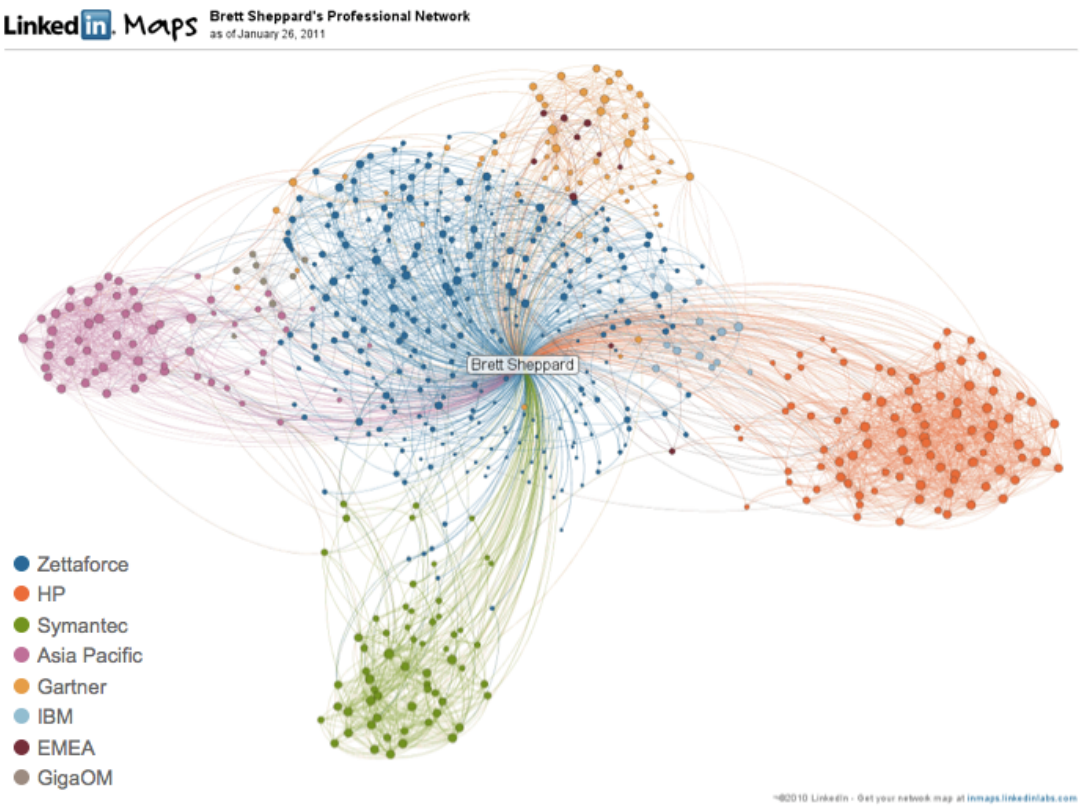
IBM Netezza works with Fuzzy Logix and other analytics libraries to provide quantitative models and web-based solutions for Internet marketing optimization, behavioral segmentation, predictive analytics, inventory optimization and other solution applications.

Visualization

With all of the information available today on the public Internet and within internal corporate sites, it's easy to feel overwhelmed. Visualization tools are important to help business stakeholders overcome a feeling of data overload, share big data information internally and with partners, and recognize actionable insights. Visualization helps ideas to be "sticky" and thus more memorable and effective.

Visualizations help business users identify patterns and take actionable steps. For example, LinkedIn Maps (below) enables users to map professional networks and understand relationships among connections. Your map is color-coded to represent different affiliations or groups from your professional career, such as your previous employer, college classmates or industries you've worked in. When you click on a contact within a circle you'll see their profile pop up on the right, as well as lines highlighting how they're connected to your connections.

Visualizing Social Media Networks: LinkedIn Maps



Source: LinkedIn

The groundbreaking graphical diagrams presented by Edward Tufte are no longer just pretty pictures in design books. They are helping to inspire modern business visualizations with real benefits. While executive dashboards displaying real-time key performance metrics do offer tremendous value in aligning and managing organizations, forming actionable insights from mounds of data can require more sophisticated and clever data visualization that is easily understood and actionable by larger groups of employees, customers and partners.

Collaboration: Friending Your Big Data

As an industry, we're only at the tip of the iceberg of enabling extraordinarily deep analysis of huge, complex data volumes to be easily shared across an extended organization and its partners.

Chatter, by Salesforce.com, is an example of one of many new software tools designed to help facilitate data-rich collaboration. Salesforce.com Chairman and CEO Marc Benioff expressed frustration with Facebook's ease of use compared to his perception of enterprise tools like Microsoft SharePoint or IBM Lotus Notes. With Chatter as a built-in extension to Salesforce.com sales and service clouds, salespeople, customer call center staff, marketers and other stakeholders can collaborate using Facebook-like interfaces that include employee updates, profiles and groups.

Big data collaboration tools like Chatter are delivering real-life business benefits. For example, using a collaboration tool extension to its EMC Greenplum databases, Zions Bancorporation can now more quickly and effectively answered SEC regulatory requests by enabling multiple departments to coordinate together. By logging into the collaboration tool, adding their documentation and notes and seeing the work of their colleagues, each department can add value, much like a virtual assembly line.

At Gartner BI Summit 2010, Gartner analyst John Hagerty advised BI advocates to give up trying to wean business users off Excel, and instead accept that the program is here to stay as one option for data analysis and information exchange. Hagerty advises IT departments to follow a rapid-iteration model to create and update reports, and allow business users to decide how to deploy data, whether in spreadsheets, BI software interfaces, dashboards, SharePoint or other collaboration tools.

IBM BigSheets, from the IBM Software Group's Emerging Technologies division, organizes information in a large spreadsheet, where users can analyze it using the sort of tools and macros found in desktop spreadsheet software. BigSheets is an extension of the mashup paradigm that integrates large sets of unstructured data, enriches that

data using semantic logic structure tools, such as LanguageWare or OpenCalais, and lets you explore and visualize enriched data in tools such as IBM Many Eyes.[23] Both IBM BigSheets and Datameer provide enablement for business analysts or other users who are not Hadoop specialists to import data into Hadoop, run queries and report results with familiar-to-use spreadsheet, charting and graphing tools.

Collaboration tools can raise important privacy questions, for both consumers and businesses, that are important to address in enabling appropriate role- and group-based access to big data. In January 2011, for example, Facebook ended a brief excursion into enabling external APIs to access users' emails and telephone numbers.

Before enabling Excel users or other user groups to access data warehouse files, organizations should evaluate their policies and enforcement for privacy and governance. For example, in addition to other security controls to prevent misuse of patient medical records, Stanford University requires users who export non-public data into Excel to sign and certify reports produced with that data, and to retain the data in the same format as it existed in the original SAP BusinessObjects, Oracle BI or Oracle Hyperion databases.

Big Data Limitations

While it's exciting to see significant industry advances in managing big data volumes, complexity and speed, it's also worthwhile to recognize at least a few of big data's limitations. As every enterprise early adopter knows, there will be technical hiccups in implementing new technologies. What starts off sounding like an inexpensive, easy-to-use, MacGyver-style "put together with duck tape but works surprisingly well" option can easily turn into a MacGruber-style comic catastrophe.

Enterprises that rely on the multitude of features on relational database management system (RDBMS) may find some emerging technologies to be bare bones. For

[23] *Interview with IBM JStart at O'Reilly Structure,*

example, at eBay, there are more restrictive policies for the number of named and concurrent users for the company's Hadoop clusters compared with its Teradata enterprise data warehouse and related Singularity semi-relational data warehouse, which shares use of Teradata SQL user interfaces. One reason for the limitations named and concurrent users on Hadoop is that a poorly written SQL query may run indefinitely on Hadoop until there is a system crash or a manual intervention by a system administration. In contrast, in eBay's Teradata system, workload management tools will stop a poorly written query that's taking up too much capacity as well as enable more-sophisticated workload balancing among applications and systems.[24]

Eventually workload tools will become more mature within the Hadoop ecosystem. But those initiatives take time, compared to the decades that Teradata and its competitors in relational databases have spent perfecting workload management in large-scale enterprise and public sector customer environments.

More fundamentally, businesses processes often adapt more slowly than technology. In many large organizations, and in some smaller ones, too, data is still silo'ed. In some cases, this reflects a conglomerate business model, where over time companies may acquire or spin out different lines of business. For a conglomerate, instead of adopting the consolidated enterprise data warehouse approach popularized by Bill Inmon, the organization may be better off with the distributed data mart approach associated with Ralph Kimball.

In other cases, data remains silo'ed because of conflicting needs among different departments or lines of business. And there are regulatory requirements and corporate governance practices, too, that can complicate organization-wide collaboration on big data.

Even when data is integrated organization-wide with appropriate role-based and group-based access rights, enterprises face adventures in data quality, as data

[24] Discussion with eBay Senior Director of Architecture and Operations Oliver Ratzesberger, on-site at eBay HQ campus technology building, March 15, 2011.

definitions and related aspects of data consistency will vary by business unit and by geography. Agreement on definitions and business rules for master data management can be a slow, painful process for IT and business leads.

To solve these organizational challenges, big data requires individuals who combine deep-down technical expertise in statistics with wide-ranging business skills to collaborate across organizations. It's this combination of deep technical and wide business skills that enables successful data scientists to access corporate silo'ed data, summarize and catalog data, understand natural language processing and present compelling visualizations that drive business actions.

While insights gleaned from big data can improve decision making, they do not rule out the vagaries of human behavior, even if we are nearing what Ray Kurzweil eloquently describes as a technology "singularity." All too often, the late David Sandler's observation remains true: People make decisions emotionally and then justify them with data.[25] In other words, data is only worthwhile if decision makers understand and use it. As Google's Hal Varian notes, while technology provides new tools for data science, "the complimentary scarce factor is the ability to understand that data and extract value from it."[26] For example, data on NFL team profitability and TV revenues can help inform discussions between the NFL and the players' union, but only the two sides' business representatives can recognize common ground and negotiate a new collective bargaining agreement to avert a work stoppage.

Overcoming these technical and organizational obstacles can require significant time and resources. In his latest book, *Profiles in Performance*, Howard Dresner, who helped popularize the term "business intelligence" during his tenure at Gartner, discusses enablers for a performance-directed business culture. In one of his case studies, the BI team at Cleveland Clinic persevered as outliers within the IT department for several years, until new leadership brought a different vision and a greater effort at organization cohesion, which included a BI-enabled performance-

[25] David H. Sandler, *You Can't Teach a Kid to Ride a Bike at a Seminar*, Bay Head Publishing, 1995.
[26] Hal Varian, quoted in The McKinsey Quarterly, January 2009.

driven culture as a key initiative. Highlighting that BI is a journey and not a quick fix; it took Dresner's profiled companies an average of eight years to create a performance-directed culture.

Key Takeaways

- The three attributes of big data — volume, complexity and speed — are increasingly becoming important design factors in enterprise architectures for both the government and private sectors.

- Once primarily associated with financial services and government, complex event processing (CEP) and related technologies are extending to industries as diverse as online travel and Web 2.0 e-commerce, which seek to derive a competitive advantage by maximizing the real-time or near-real-time velocity of bid data processing.

- Using cloud-computing technologies, organizations are experimenting with distributed data stores, cloud-compute capacity for data analytics, hosted data integration and even operational databases in the cloud, adapting public cloud technologies for use in private and hybrid clouds.

- Initially inspired and funded by several global Internet and social media companies, Hadoop/MapReduce has moved past testing and development to become a viable extension or, in some cases, an alternative to an RDBMS for managing and analyzing huge data sets, albeit without yet the range of workload management and other features that have been built over decades for RDBMSs.

- Visualization and collaboration tools, and their adoption, are important to enable the benefits of analytics to extend beyond a small group of data scientists to a broader group of business colleagues and authorized external partners.

Further Reading

Big Data 2011 Preview

Using cloud-computing technologies, organizations are experimenting with distributed data stores, cloud compute capacity for data analytics, hosted data integration and even operational databases in the cloud. Though the space is not without its obstacles, including plenty of privacy concerns, there are numerous sales-growth opportunities and new business models finally surfacing in 2011.

Health Care's Climb to the Cloud

From patient records to biomedical research to insurance claims, data and the ability to manage large amounts of it is a major concern for hospitals, insurers and researchers. Cloud computing offers each of these players a potentially more cost-effective alternative to traditional data storage and management.

Big Data, ARM and Legal Troubles Transformed Infrastructure in Q4

Some might call this past quarter in the infrastructure space transformative. The rise of ARM-based processing suggests the days of x86 dominance might be coming to an end, while the Amazon Web Services-WikiLeaks controversy cast new light on the legal aspects of cloud computing. Big data got bigger, meanwhile, as the Hadoop ecosystem expanded, and amid all these cutting-edge technologies, two archaic topics — Novell and Java — proved they aren't going anywhere soon.

Want More Information?

Contact Brett Sheppard the author of this report,
or any of the other experts at GigaOM Pro.

Discuss this report online.

Suggest a research topic.