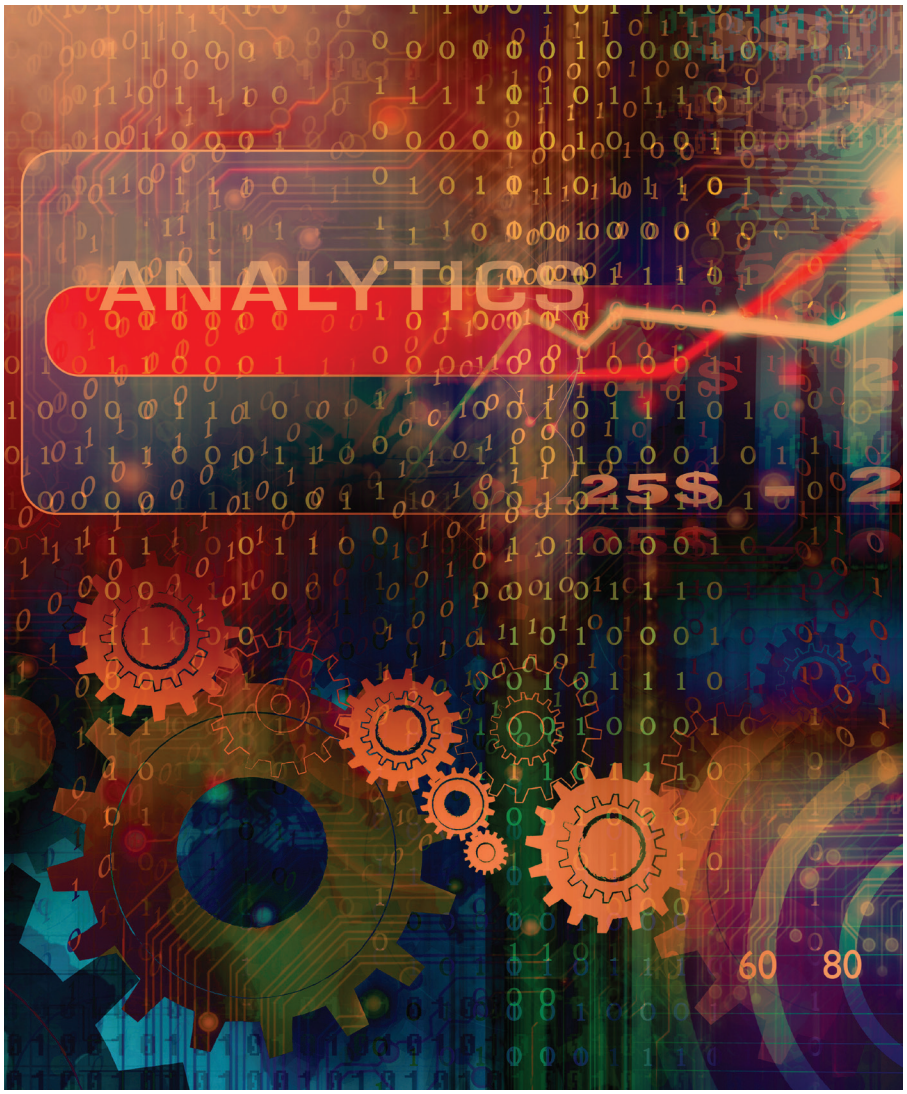
MIT TECHNOLOGY REVIEW CUSTOM
Produced in partnership with
ORACLE

# Securing the Big Data Life Cycle

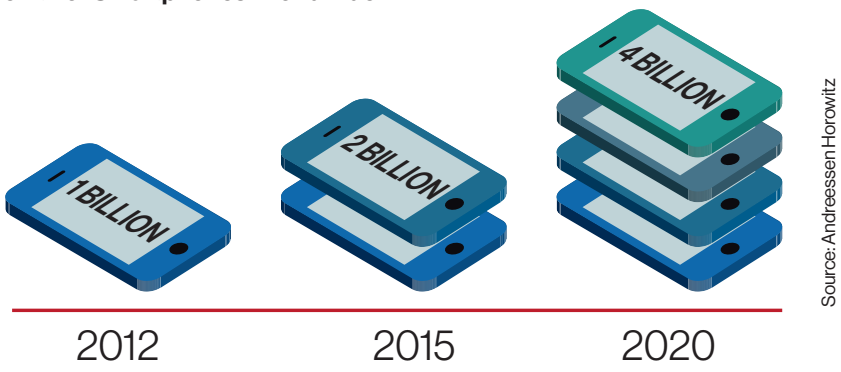



"Big data drives big benefits, from innovative businesses to new ways to treat diseases. The challenges to privacy arise because technologies collect so much data (e.g., from sensors in everything from phones to parking lots) and analyze them so efficiently (e.g., through data mining and other kinds of analytics) that it is possible to learn far more than most people had anticipated or can anticipate given continuing progress."

– U.S. President's Council of Advisors on Science and Technology(1)

The big data phenomenon is a direct consequence of the digitization and "datafication" of nearly every activity in personal, public, and commercial life. Consider, for instance, the growing impact of mobile phones. The global smartphone audience grew from 1 billion users in 2012 to 2 billion

**Growth of Smartphones Worldwide**




Source: Andreessen Horowitz

today, and is likely to double again, to 4 billion, by 2020, according to Benedict Evans, a partner with the venture capital firm Andreessen Horowitz.[2] That means that about a quarter of the world's population now uses smartphones at least occasionally, and data from all those calls and applications is being warehoused and mined.

Add to that the Internet of Things (IoT), the growing network of everyday objects equipped with sensors that can record, send, and receive data over the Internet without human intervention. Gartner Inc. estimates that the IoT currently includes 4.9 billion connected "things"—a 30 percent increase from 2014; analysts predict that the number will hit 25 billion by 2020.[3] Of course, all these devices will generate data.

## Big Data Requires Bigger Responsibility

Companies of all sizes and in virtually every industry are struggling to manage exploding amounts of data. But as both business and IT executives know all too well, managing big data involves far more than just dealing with storage and retrieval challenges—it requires addressing a variety of privacy and security issues as well.

In a talk at the Technology Policy Institute's 2013 Aspen Forum, Federal Trade Commission Chairwoman Edith Ramirez described some big data pitfalls to be avoided. Though many organizations use big data for collecting non-personal information, there are others that use it "in ways that implicate individual privacy," she noted, adding that the type of information collected may "reflect an individual's health concerns, browsing history, purchasing habits, social, religious and political preferences, financial data, and more."[4]

Ramirez described several potential pitfalls, including:

- Ubiquitous and indiscriminate data collection from a wide range of devices
- Unexpected uses of collected data, especially without customer consent
- Unintended data breach risks with larger consequences

As head of the governmental entity responsible for protecting U.S. consumers, Ramirez called for "big responsibility" with big data. "The larger the concentration of sensitive personal data, the more attractive a database is to criminals, both inside and outside a firm," she said. "The risk of consumer injury increases as the volume and sensitivity of the data grows."

Ramirez also called for stronger incentives to push companies to better safeguard sensitive information. "The FTC has urged Congress to give the agency civil penalty authority against companies that fail to maintain reasonable security," she said. "The advent of big data only bolsters the need for this legislation."

What this means for organizations is that if they fail to secure the life cycle of their big data environments, then they may face regulatory consequences, in addition to the significant brand damage that data breaches can cause.

A few years ago, an individual breach affected 1 million to 10 million records; today, in the age of mega-breaches, a single incident can involve 200 million records—or more.

## A Trio of Top Threat Vectors

More than 60 percent of 763 security practitioners surveyed reported successful cyberattacks on their midsize-to-large companies in the previous year, according to CyberEdge.[5] The Verizon 2015 Data Breach Investigations Report (DBIR) tallied nearly 80,000 security incidents, including 2,122 confirmed data breaches.[6] Security breaches affect organizations of all industries and sizes. A few years ago, an individual breach affected 1 million to 10 million records; today, in the age of mega-breaches, a single incident can involve 200 million records—or more.

Business and IT executives are learning through harsh experience that big data brings big security headaches. One big problem has been Hadoop, an open-source software framework for storing and processing big data in a distributed fashion. Hadoop wasn't built with security in mind; it was developed solely to address massive data storage and faster processing.

But despite its security weaknesses, Hadoop is being widely integrated with existing IT infrastructure.

Today's most common threats to big data environments in general, and Hadoop in particular, include:

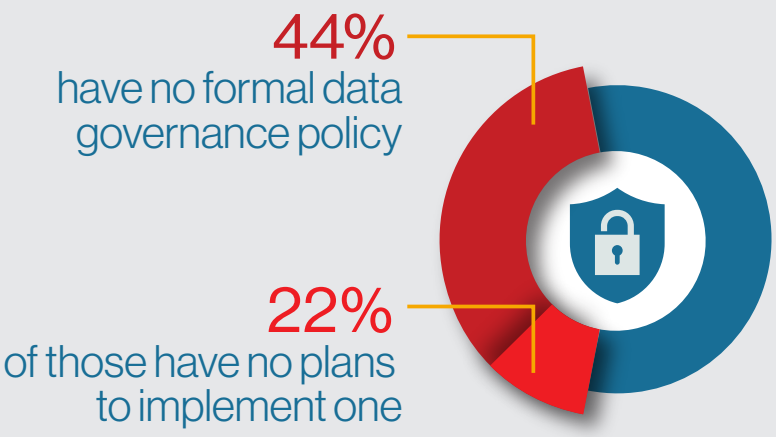
• **Unauthorized access.** Built under the principle of "data democratization"—so that all data is accessible by all users of the cluster—Hadoop has had challenges complying with certain rigorous compliance standards, such as the Health Insurance Portability and Accountability Act of 1996 (HIPAA) and the Payment Card Industry Data Security Standard (PCI DSS). That's due to the lack of access controls on data, including password controls, file and database authorization, and auditing.

• **Data provenance**. With open source Hadoop, it has been difficult to determine where a particular dataset originated and what data sources it was derived from. At a minimum, this means there is significant potential for garbage-in, garbage-out problems, but it also means that people might base their business decisions on analytics taken from suspect or compromised data. Users need to know the source of the data to ensure its validity for critical predictive activities.

Managing big data involves far more than just dealing with storage and retrieval challenges—it requires addressing a variety of privacy and security issues as well.

• **DIY Hadoop.** A build-your-own Hadoop cluster presents inherent risks, especially in shops where few engineers are capable of building and maintaining one. As a cluster grows from small project to advanced enterprise Hadoop, every period of growth—patching, tuning, verifying versions between Hadoop modules, OS libraries, utilities, user management, and so on—becomes more difficult and time-consuming. This leaves fewer engineers to attend to security holes, operational security, and stability, until a major disaster, such as a data breach, occurs.

**Percentage of Organizational Data Governance Policies**




Source: Rand Secure Archive Data Governance Survey

## Real-World Examples

A single successful data security breach comprising millions of records can result in multimillion-dollar losses and other serious consequences for the victimized organization. One high-profile example is the Target data breach of late 2013, in which cybercriminals stole account data for millions of Target's retail shoppers. At least 40 million credit cards were compromised, according to the retail chain's estimates, and the thieves also stole personal information, including names, addresses, email addresses, and phone numbers of as many as 110 million customers.

Target's stock dropped dramatically following the breach, and the company faced numerous lawsuits, including one from a group of financial institutions claiming tens of millions of dollars in damages. In the wake of the breach, the company's CEO and CIO, who had both worked at Target for decades, resigned.

Today, there are other ways, besides outright breaches, to "implicate individual privacy," in the sense meant by FTC chairwoman Ramirez. Groups or individuals can correlate datasets previously considered anonymous with other publicly available information to identify individuals in those datasets. That, in turn, can expose private information that the individuals disclosed when they thought their privacy was guaranteed.

The quintessential paper on big data risk, "Robust De-anonymization of Large Sparse Datasets,"[7] by two researchers from the University of Texas at Austin, illustrates the sensitivity of supposedly anonymous information. The authors were able to identify the publicly available and "anonymous" dataset of 500,000 Netflix subscribers by cross-referencing the data with the Internet Movie Database (IMDb). For some IMDb users who were also Netflix members, the researchers were able

to identify the users' complete Netflix movie rating history, information that can, in turn, be used to determine personal political, religious, or sexual preferences.

In a more recent instance of big data security concerns, the 2013 public release of a New York taxicab dataset was completely de-anonymized, according to Neustar, a real-time provider of cloud-based information services and data analytics. The data ultimately revealed cab drivers' annual incomes and, possibly more alarming, their passengers' weekly travel habits.[8]

These are just two scenarios that highlight the type of sensitive data being collected, organized, and used within big data environments in general and Hadoop in particular. As processing speed and techniques improve, identifying individuals will become even easier. Hadoop environments are increasingly likely to become the crown jewels targeted by both external cybercriminals and malicious insiders—if they're not effectively secured.

> Built under the principle of "data democratization"—so that all data is accessible by all users of the cluster—Hadoop is unable to comply with certain rigorous compliance standards, such as HIPAA and PCI DSS.

## Potential Pitfalls, Key Steps

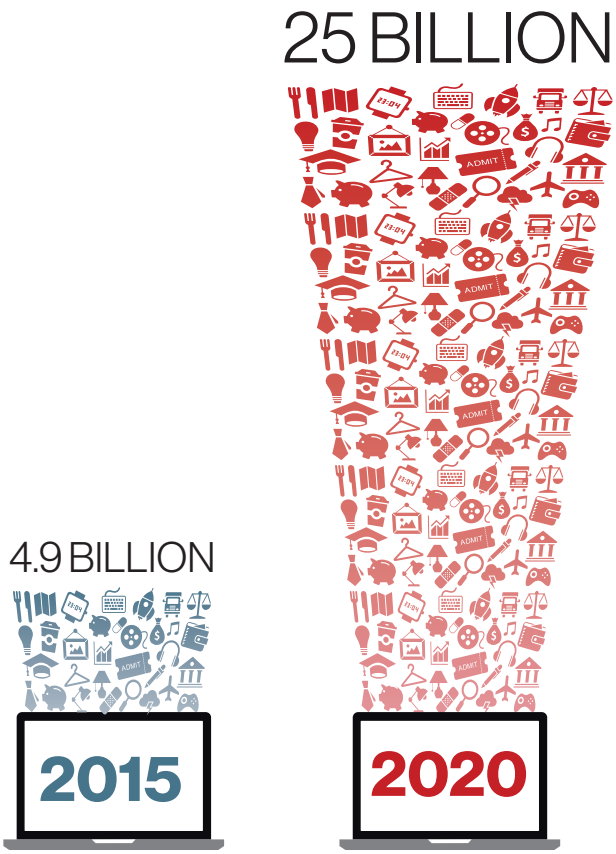
Meanwhile, as the number of big data environments grows rapidly, many companies make a key mistake in establishing them: They focus too much on the potential benefits of big data, such as gaining insights and generating revenues, and too little on making sure they're sufficiently securing that information. That lack of attention to security can, of course, leave the organization open to attacks from any number of unknown sources.

Other evolving circumstances also contribute to a wide range of security-related risks, hurdles, and potential pitfalls associated with big data. As the Cloud Security Alliance, an industry group, notes: "Large-scale cloud infrastructures, diversity of data sources and formats, the streaming nature of data acquisition, and high-volume inter-cloud migration all create unique security vulnerabilities."[9]

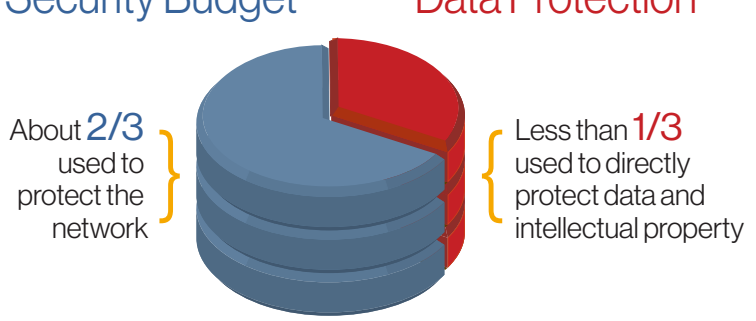
Two additional complicating factors include:

1. **Outdated approaches.** Previous perimeter-based approaches to security are no longer sufficient. A CSO Market

**Growth of the Internet of Things**




Source: Gartner Inc.




Source: CSO Market Pulse Survey

Pulse survey found that "two-thirds of security budgets are used to protect the network, with less than a third used to directly protect the data and intellectual property that reside inside the organization."[10] Unfortunately for most organizations, breaches of servers occurred more frequently than network breaches. In fact, less than 1 percent of breaches were detected using network perimeter security controls such as switches, firewalls and routers, according to the 2014 Verizon DBIR.

2. **Insufficient governance.** Forty-four percent of organizations have no formal data governance policy, and 22 percent of these firms have no plans to implement one, according to the 2013 Rand Secure Archive Data Governance Survey.[11] Big data increases companies' data ingestion by many orders of magnitude, adding to the complexity. Without overall management of the availability, usability, integrity, and security of big data employed in an enterprise, organizations will find it tough to address the stronger privacy and regulatory mandates called for by the FTC and the European Union.

Ultimately, securing the big data life cycle requires that organizations take steps in four key areas, according to the Cloud Security Alliance's Big Data Working Group:

To derive real business value from big data, you need the right tools to capture and organize a wide variety of data types from different sources, the ability to analyze each type within the context of your enterprise data—and do it all securely.

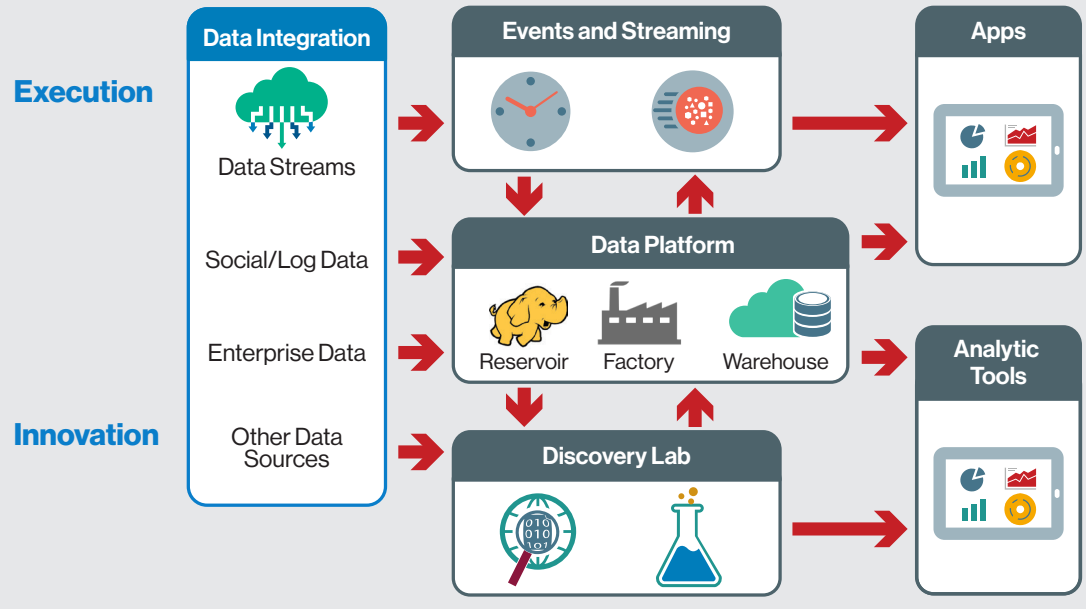
1. **Infrastructure security.** Secure computations in distributed programming frameworks as well as in nonrelational data stores.

2. **Data privacy.** Secure the data itself using a privacy-preserving approach for data mining and analytics. Also, protect sensitive data through the use of cryptographically enforced data-centric security and granular access control.

3. **Data management.** Manage the enormous volume of data using scalable, distributed solutions to secure data stores and enable efficient audits and data provenance.

4. **Integrity and reactive security.** Use endpoint validation and filtering to check the integrity of streaming data, and real-time security monitoring and analytics to help prevent and address security problems.[9]

**Today's Big Data Architecture**




Source: Oracle

Sensitive and non-sensitive data flow into big data environments from multiple sources, including online applications, ERP systems, social media, and enterprise systems. Requirements include engines for actionable events, a data reservoir (Hadoop) that is integrated with a traditional data warehouse, and business analytics that provide actionable insights and information.

## The Solution: Defense in Depth

No question about it: In the age of big data, organizations need to adopt a data-centric approach to security. Specifically, they need to employ three key types of security controls:

**Preventive:** Securing the data itself with controls such as encryption of data at rest and in motion, redaction of data in applications, and use of identity and access management.

**Detective:** Looking for anomalous behavior by, for instance, auditing operating systems, Hadoop services, data activity, and monitoring systems throughout the big data environment, and providing compliance reports or alerts about potential problems.

**Administrative:** Implementing tools that enable the processes and procedures for security, such as sensitive data discovery, privileged user analysis, configuration management, and encryption key management capabilities.

A comprehensive data security approach ensures that the right people, internal or external, get access to the appropriate data and information at the right time and place, within the right channel. Defense-in-depth security prevents and safeguards against malicious attacks and protects organizational information assets by securing and encrypting data while it is in motion and at rest. It also enables organizations to separate roles and responsibilities and protect sensitive data without compromising privileged user access. Furthermore, it extends monitoring, auditing, and compliance reporting across traditional data management to big data systems.

No question about it: In the age of big data, organizations need to adopt a data-centric approach to security.

Organizations are now in need of big data environments that include enterprise-grade authentication and authorization (Kerberos or LDAP and Apache Sentry project), and auditing that can be automatically set up on installation, greatly simplifying the process of hardening Hadoop.

## The Case for a Big Data Technology Platform

Businesses are finding that big data works best in an environment that combines Hadoop, NoSQL, and relational databases. To realize a robust and successful big data strategy, it's important to determine how to integrate these technologies under a big data technology platform. Such a platform is where the company governs all of its data and makes it securely available to the rest of the organization for use and analysis. The platform also includes the critical systems currently used to run the business.

Securing the big data life cycle requires the following security controls:

- Authentication and authorization of users, applications, and databases
- Privileged user access and administration
- Encryption of data at rest and in motion
- Data redaction and masking for both production and non-production environments

- Separation of responsibilities and roles
- Implementing least privilege
- Transport security
- API security
- Monitoring, auditing, alerting, and reporting

In addition, organizations must be able to address regulatory compliance and extend existing governance policies across their big data platforms.

## Oracle for Enterprise Big Data

Analyzing new and diverse digital data streams can uncover new sources of economic value, provide fresh insights into customer behavior, and identify market trends early in the game. Unfortunately, big data solutions that provide these benefits often create high-risk environments due to the concentration of personally identifiable information, personal health information, and other sensitive and regulated data. To derive real business value from big data, organizations need the right solution to capture and organize a wide variety of data types from different sources, the ability to analyze each type within the context of their enterprise data—and do it all securely.

With Oracle, organizations can take advantage of the future of big data while preserving the value of existing investments in technology and skills. Oracle removes the barriers between Hadoop, NoSQL, and relational databases in the cloud and on-premises to support a wider array of analytics, applications, and algorithms.


Oracle's big data solutions help organizations quickly discover and predict real-world patterns by using insights about customer demand, or current events, before the situation changes. Oracle helps to simplify access to all data so that applications can query the complete big data environment. Organizations can secure and govern any critical data even when it's mixed with third-party data from customers and partners, or with data from mobile applications and connected devices.

Organizations can achieve all the benefits that big data has to offer while providing a comprehensive, inside-out security approach that ensures that the right people, internally and external, receive access to the appropriate data at the right time and place.

> "The larger the concentration of sensitive personal data, the more attractive a database is to criminals, both inside and outside a firm. The risk of consumer injury increases as the volume and sensitivity of the data grows."
>
> – Edith Ramirez, Chairwoman, U.S. Federal Trade Commission.

The Oracle big data solutions are integrated together to detect and prevent while securing sensitive data assets. By following the best practices of "least privilege" and "separation of duties," Oracle enables organizations to separate roles and responsibilities and protect sensitive data without compromising access for privileged users. Furthermore, Oracle's solutions provide monitoring, auditing, and compliance reporting across big data systems and traditional data-management systems. All those factors reflect one key imperative: It's important not to start from scratch, but to instead extend existing security policies, processes, and tools to new data—and new data management technologies.

## Bringing It All Together

Today's treasure trove of big data has created unprecedented opportunities all across the business landscape. There's no question that companies that use big data effectively can expect to gain significant competitive advantage. But despite big data's many benefits, it's generated new worries as well—and nowhere more so than in the areas of security and privacy. Most significantly, organizations are exposing their sensitive information to increased risk as they integrate open-source Hadoop into their IT environments. For that reason, companies serious about using big data effectively need to make sure they're doing so securely, protecting their valuable information and securing private data so that it stays private.

For more information, visit www.oracle.com/bigdata.

## REFERENCES


1. "Report to the President, Big Data and Privacy: A Technological Perspective," President's Council of Advisors on Science and Technology, 2014

2. "The Truly Personal Computer," The Economist, 2015.

3."Gartner says 4.9 Billion Connected 'Things' Will Be in Use in 2015," Gartner Inc., 2014

4. "The Privacy Challenges of Big Data: A View from the Lifeguard's Chair," speech by FTC Chairwoman Edith Ramirez, 2013

5. "2014 Cyberthreat Defense Report," CyberEdge Group, 2014

6. "2015 Data Breach Investigations Report," Verizon, 2015

7. "Robust De-anonymization of Large Sparse Datasets," by Arvind Narayanan and Vitaly Shmatikal, University of Texas, 2008

8. "Riding With the Stars: Passenger Privacy in the NYC Taxicab Dataset," Neustar Inc., 2014

9. "Big Data Taxonomy," Big Data Working Group, Cloud Security Alliance, 2014

10. "An Inside-Out Approach to Enterprise Security," Oracle/CSO Custom Solutions Group white paper, 2013

11. "Rand Secure Archive Releases North American Survey Results on Data Governance," Rand Secure Archive, a division of Rand Worldwide, 2013


## ADDITIONAL RESOURCES


"Big Data," Forrester Research

Trending Topics: Big Data," Gartner Inc.

"Expanded Top Ten Big Data Security and Privacy Challenges," Big Data Working Group, Cloud Security Alliance, 2013



**MIT TECHNOLOGY REVIEW CUSTOM**

Produced in partnership with

ORACLE®


### About MIT Technology Review Custom

Built on over 100 years of excellence in technology journalism, MIT Technology Review Custom is the arm of global media company MIT Technology Review that's responsible for creation and distribution of custom content. Our expert staff develops meaningful and relevant content from concept to completion, distributing it to users when and where they want it, in digital, print, online, or in-person experiences. The turn-key solutions offer everything from writing and editing expertise to promotional support, and are customized to fit clients' content marketing goals—positioning them as thought leaders aligned with the authority on technology that matters.