Data Quality, Analytics, and Privacy in Big Data

Xiaoni Zhang and Shang Xiang

Abstract. In today's world, companies not only compete on products or services but also on how they can analyze and mine data in order to gain insights for competitive advantages and long term growth. With the exponential growth of data, companies now face unprecedented challenges, however are also presented with numerous opportunities for competitive growth. Advancement in data capturing devices and the existence of multi-generation systems in organizations have increased the number of data sources. Typically, data generated from different devices may not be compatible with each other, which calls for data integration. Although, ETL market offers a wide variety of tools for data integration, it is still common for companies to use SQL to manually produce in-house ETL tools. There are technological and managerial challenges to deal with data integration. During data integration, data quality must be embedded in it.

Big data analytics delivers insights which can be used for effective business decisions. However, some of these insights may invade consumer privacy. With more and more data related to consumer behavior being collected and the advancement in big data analytics, privacy has become an increasing concern. Therefore, it is necessary to address issues related to privacy laws, consumer protections and best practices to safeguard privacy. In this chapter, we will discuss topics related to big data in the area of big data integration, big data quality, big data privacy, and big data analytics.

Keywords: big data, data quality, privacy, data analytics.

1 Introduction

Research conducted by the McKinsey Global Institute (2011) points to big data having the capability to create substantial value and commercial impact.

Xiaoni Zhang
Northern Kentucky University, USA
e-mail: zhangx@nku.edu

Shang Xiang
KPMG
e-mail: sxiang@kpmg.com

A.E. Hassanien et al.(eds.), *Big Data in Complex Systems*,
Studies in Big Data 9, DOI: 10.1007/978-3-319-11056-1_14

McKinsey found the potential of a 60 percent increase in retailers' operating margins, 0.7 percent increase in productivity in U.S. health care, all translating into a $300 billion value per year. In addition, there's the potential increase in demand for deep analytical talent positions, estimated between 140,000 and 190,000. Currently, many companies have or plan to implement big data solutions. Businesses changed their view on traditional view of assets to include big data in addition to cash, inventory, and fixed assets (Brands 2014). Not only businesses are exploring big data. Governments with advanced ICT infrastructure have invested in big data for national security, transparency, economic development, and operational efficiency (Gang-Hoon 2014).

In 2002, the Data Warehousing Institute reported the cost of poor data quality on US economy at $600 billion annually (TDWI 2002). Flawed data cost ten times as much to complete a simple work (Everett 2012). Data quality affects business decisions. When evaluating data quality, usually several dimensions are used: accuracy, relevancy, currency and completeness. As more and more big data becomes integrated, the demand for big data analytics increases. Unlike traditional data analysis in which data types are typically numeric, big data is complex in its data types. Data types include numbers, texts, images, voices, videos, etc. Thus big data analysis is performed on the combination of structured and unstructured data. Big data requires the ability to analyze diverse data sources and data types. As new technologies and techniques are emerging in this market, the need to explore the analytic issues and practices associated increases.

Big data is a buzzword and companies from all industries have been investing in big data technologies, techniques, applications and management. In today's world, companies compete not only on products or services but also on how they can analyze and mine data in order to gain insights on competitive advantages and long-term growth. With the exponential growth of data, companies now face unprecedented challenges, conversely are also presented with numerous opportunities. Advancement in data capturing devices and the existence of multi-generation systems in organizations have created many data sources. Although the data generated from different devices may not be compatible with each other, the need for data integration. The ETL market offers a wide variety of tools for data integration; it is still common for companies to use SQL to manually produce in-house ETL tools. There are technological, managerial challenges to deal with data integration. During data integration, data quality and appropriate tools must be used to ensure data quality. Data is the most valuable asset of any organization. Business decisions depend on high quality data. High quality data creates many opportunities for improvement from daily operations to strategic planning. Data quality management is a necessity for businesses and this management needs to be constantly updated, analyzed and perfected. New technologies, techniques, and methodologies on data quality management are constantly developed. Best practices and lessons learned are good sources to improve data quality management. However, data quality issue is not a new issue.

While acknowledging that Big Data has provided benefits to consumers in creating pricing transparency (Fulgoni 2013). Targeting of Internet advertising

based on data analysis is said to offer a means to maintaining or improving brand equity.

In this book chapter we present several interesting topics on big data, as detailed below:

- Section 2 - discuss data/information quality, market trends and data management
- Section 3 - focus on privacy and security issues in general and healthcare in specific
- Section 4 - overview on big data analytics and technologies
- Section 5 - discuss markets and prior publications on big data
- Section 6 - summarize what has been covered in this book chapter and provide future research directions and technology trends

2 Data/Information Quality and Data Integration

Data quality issues are prevalent and such issues attract attention of companies of all sizes from all industries. Data quality is critical for business operations, if there are any errors in business transactions, the consequences could be detrimental – ranging from lost sales, lost customers, lost competitions, failure to market a product. Flawed data could create chain effects affecting many business activities such as the delivery of products to specific locations, traditional marketing outreach to customers and prospects via mailings.

2.1 Definition

Data quality issues have been researched for many years and it has been used interchangeably with information quality. Petter et al. (2013) define information quality as the desirable characteristics of system outputs (content, reports and dashboards) and information quality is one of the key success factors for information systems. Setia et al. (2013) consider that information quality has these dimensions: completeness, accuracy, format and currency whereas Petter et al. (2013) note that information quality contains these dimensions: relevance, understandability, accuracy, conciseness, completeness, understandability, currency, timeliness, and usability.

Data quality is context specific. The data is of high quality if their meanings are well understood by different user groups. With the increasing variations of devices that data is generated from and coupled with different applications, data consistency could represent a problem. For example, customer names and contact information may be stored differently across applications, which in turn a person's age and birth date may conflict within different portions of a database.

Given the data quality dimensions, when managing data quality, it is important to note the quality at seven sources: entry quality, process quality, identification quality, integration quality, usage quality, aging quality, and organizational quality (McKnight 2009). The American SAP User Group reported that 93% of companies experienced data problems in their most recent projects (Woods 2009).

2.2 Overview of Markets

According to ResearchMoz (2013), global data quality tools market will grow at a compounding annual growth rate of 16.78 percent between 2012-2016. Improving productivity is one of the key factors that contribute to this market growth. The Global Data Quality Tools market has been witnessing the emerging SaaS-based data quality tools. Gartner estimates that this market reached $960 million in software revenue at the end of 2012. This translates to growth of 12.3% in constant-dollar terms over 2011 (a standout year in which this market grew by 17.5%). Gartner forecasts that the growth of data quality tools market will accelerate during the next few years and will become the fastest growing market. The market is expected to increase to 16% by 2017 and reach $2 billion in constant-dollar software revenue. The data quality tools in the market of unstructured data are expected to grow significantly in the future.

Recognizing that data quality affects the performance of organizations, companies will focus their attention on data quality. With the developing cloud computing and big data markets, technologies and data quality market will be more complex. New services and vendors emerge, with two major groups. The market has data service providers and data management service providers. Data services providers offer data delivery, analysis, management, or governance-related services, whereas data management services providers operate in the area of finding, collecting, migrating, and integrating data.

With the popularity of Software as a Service (SaaS) and cloud based computing, Gartner (2013) reports that deployments reached 14% for SaaS and 6% for cloud-based tools. In the emerging data management area (big data and analytics), companies are likely to use third parties for its data management, with 69% of respondents reporting they currently use or will use a service provider (Green 2014).

Data quality initiatives for transactional, financial, location and product data are increasing and 78% of data quality projects address customer data quality (Lawson 2013). Overall, data governance drives many data quality initiatives. Data quality in the big data arena has not been taken off yet. Gartner report also shows that the current buyers do not consider data quality related issues. There are stand-alone data quality tools in the market that address the core functional requirements, such as: data profiling, data quality measurement, parsing and standardization, cleansing, matching, monitoring, and enrichment.

Table 1 shows the Gartner report on the magic quadrant for data quality tools between 2009-2013. Gartner classifies the major vendors into four categories: leaders, challengers, visionaries and niche players. In general the leaders control 50% share of the market. In the past five years, the vendors in the leaders category remain constant with such companies like Informatica, IBM, SAS, SAP and

Trillium. In the challenger category, Pitney Bowes is in this category consistently for the past five years, however Oracle joined this category in 2011. In the visionary category, Human Inference was in this category for four years, Talent and Atccama for three years, and Information Builders for the last two years. In the niche player category, five players: Red Point, Data Mentors, Datactics, Innovative Systems, Uniserv, are consistently ranked in this group for the past five years.

In terms of the market share large vendors like IBM, Informatica, Pitney Bowes, SAP and SAS takes 50% and all rest of the market is shared by mid-sized and small-sized companies. As shown in Table 1, in the past five years, the market players are stable with the exception that Oracle enters the data quality market in 2011.

Table 1 Magic Quadrant Data Quality Tools Market between 2009-2013

Year	**2013**	**2012**	**2011**	**2010**	**2009**
Leaders	Informatica	Informatica	Informatica	Informatica	Informatica
	IBM	IBM	IBM	IBM	IBM
	Triumlium	Triumlium	Triumlium	Triumlium	Triumlium
	SAP	SAP	SAP	SAP/ Business Objects	SAP
	SAS	SAS/ Dataflux	SAS/ Dataflux	Dataflux	Dataflux
Challengers	Oracle	Oracle	Oracle		
	Pitney Bowes	Pitney Bowes	Pitney Bowes	Pitney Bowes	Pitney Bowes
Visionaries	Neopost/ Human Inference	Human Inference		Human Inference	Human Inference
	Talend	Talend	Talend		
	Information Builders	Information Builders/ iWay		Datanomic	Datanomic
	Atccama	Atccama	Atccama		
	X88				
Niche Players	Red Point	Red Point (DataLever)	Human Inference		Netrics
	Data Mentors	Data Mentors	Data Mentors	Data Mentors	Data Mentors
	Datactics	Datactics	Datactics	Datactics	Datactics
	Innovative Systems	Innovative Systems	Innovative Systems	Innovative Systems	Innovative Systems
	Uniserv	Uniserv	Uniserv	Uniserv	Uniserv
			DataLever	DataLever	DataLever

2.3 Data/Information Quality Management

Poor data quality creates detrimental effects on businesses. Data flaws increase the costs for organization. In turn many functional areas are affected, including bids and proposals, research and development, human resources, and customer relationship management. Data management plays in many cost-saving initiatives. Good data management and data governance practices are helpful in ensuring data quality.

Ten years ago, Data Warehousing Institute reported that poor data quality costs six hundred billion dollars annually. Recently, English noted organizations spent 20-35% of operating revenue in recovery from process failure, information scrap and rework caused by poor data quality (2009). Poor data quality is costly for any organization. To resolve data quality issues, organizations invest in technologies to address data errors. Gartner's survey (2013) show more than two-thirds of organizations will increase their spending on data management services in the next year.

Total information quality management (TIQM) methodology developed by English (2003) is practical and useful in guiding and managing data quality. It follows the six sigma's define-measure-analyze-improve-control methodology. TIQM requires the establishment and deployment of roles, responsibilities, policies, and procedures concerning the acquisition, maintenance, dissemination, and disposition of data. Successful enforcement of TIQM depends on the partnership between business users and the technology group. The business users define business rules and meanings for data elements whereas the technology group builds architecture, databases, and applications to ensure the effective life span of data elements.

2.4 Big Data Quality

Master data management is crucial in data quality management. Master data management and data governance initiatives are critical programs organizations should have. In the big data era, with data volumes growing explosively, data strategy should focus on collecting the right and needed data. In the data quality tools market, there is a trend for data quality vendors and data integration vendors to converge. It is typical for companies with technologies of many generations made by various vendors to offer various methodologies.

Master data is the most important data for business. Commonly used master data includes customer, employee, products etc. All business transactions involve master data. Master data is used by many business applications and its definitions, standardization, and management is crucial in order to depend on the master data. Master data is used to generate reports for operational purposes, analyze trends, and identify anomalies for strategic purpose.

Data warehouses, data quality and data integration technologies work together to better safeguard data quality. Big data quality involves a variety of data types

with increased velocity. Assessing big data quality is more complicated than before with Hadoop or NoSQL. Technologies are the essential for managing big data; however new technologies are not mature and still need to evolve.

3 Data Privacy and Security

Privacy is of interest to each of us. However, as an individual, we do not know how to protect privacy. Technology advancements make "zero privacy" possible. Privacy means the right to be left alone. Information privacy or data protection laws prohibit the disclosure or misuse of information held on private individuals. Many countries in the world have enacted data privacy laws. Compared to Europe, the data privacy law is less regulated in the U.S.

The legal framework regulating the big data business model is built upon existing principles of intellectual property, confidentiality, contract and data protection law. Big data could potentially lead to big legal battles without proper security and privacy measures in place. Companies buy, sell and share data with other entities. It is important to know the sources of the data, and the extent to which the data can be re-used. When using big data, first and foremost companies must be certain that the data from various sources are anonymized.

Privacy invasions occur from time to time and the consequences of privacy invasions in healthcare are particularly salient. In this chapter, we focus on data privacy in healthcare due to the fact that the U.S. healthcare system is undergoing major changes and healthcare has more privacy regulations and requirements. Information privacy or data protection laws prohibit the disclosure or misuse of information held on private individuals. Laws, regulations, technologies, and hospital mergers all have played interwoven roles in privacy.

3.1 Healthcare Big Data

Most healthcare executives have high expectations for big data, but talent shortage and lack of resource are roadblocks to realization of big data's potential. Society of actuaries (2013) report that 87% of healthcare decision makers perceive big data as impactful in future; 84% find it difficult to find skilled people in optimizing big data; 45% plan to hire more skilled people in the next year. In terms of current value of big data, 45% states substantial benefits; 27% says that big data provides some benefits but not much; 22% of healthcare decision makers claims no benefits.

One-third of IT decision makers in healthcare say they have concerns about the potential of big data (35%); they consider big data as a double-edged sword with both opportunity and risk (34%); they also thinks that payers (55%) are more likely than providers (47%) to perceive big data as an opportunity. Only half of decision makers say their organization is very well-prepared to take full advantage of data growth (51%). Based on Society of Actuaries's survey, it seems that at health insurance companies are better prepared for big data than hospitals and health systems.

Decision makers at health insurance companies, hospitals and health systems set high expectations for big data. The benefits to be delivered by big data will include not only help set better financial decisions, in addition to population health and clinical outcomes. Yet, currently many of the expected benefits are not realized and healthcare is awaiting further developments in big data. As big data is such a new field, it is difficult to find people with the necessary big data experience and skill. In a few years, when big data technologies become increasingly mature and more people learn the skills, tangible benefits of implementing big data solutions will be realized.

3.2 Data Privacy in Healthcare

Data is generated at a compounding rate. With the increased healthcare data created at an astounding rate by advanced diagnostic equipment, PDAs, tablets, health monitoring devices etc. Data security and privacy issues have always been concerns in all industries. As consumers, we care about our identity and privacy. As patients, we want our medical records to be secured and the history of our medical records to be known only to those we deem important. Laws are important in restricting organizations and individuals' behavior and play vital roles in safeguarding privacy and security.

Health Insurance Portability and Accountability Act (HIPPA) was enacted in 1996. The Health Information Technology for Economic and Clinical Health (HITECH) Acts were signed into law in 2009. HIPPA and HITECH are designed to address the concerns associated with the electronic transmission of health information, the implementation of electronic information systems and the prevalence of cloud computing in healthcare. These new challenges have introduced unprecedented privacy and security challenges (Birk, 2013).

3.3 Data Security Overview

Experian reports "the healthcare industry, by far, will be the most susceptible to publicly disclosed and widely scrutinized data breaches in 2014" (Carr, 2014). In healthcare, social security numbers are stored and used as a unique identifier for many systems. Ponemon Institute survey reports that 94% of the respondents have at least one data breach in the last two years (2014). The 2010 HIMSS Analytics Report on Security of Patient Data shows that healthcare organizations are having major obstacles securing patient information, with efforts being largely reactive rather than proactive. 19% of the health organizations surveyed had a security breach as compared to 13% in 2008, of these, 84% were as a result of lost or stolen laptops, improper disposal of documents, stolen backup tapes, etc.; 87% of organizations note their data access, sharing, and security policies.

Today's hackers are more tech savvy and can infiltrate networks by exploiting vulnerabilities. IBM (2012) reports the commonly used avenues for breaches include exploiting default or easily guessed passwords, backdoor malware, use of

stolen credentials, exploiting backdoor or command and control channels, key loggers spyware, and SQL injection attacks.

Privacy and security risks can also come from connected medical devices and consumer health-monitoring gadgets. These devices play critical and important roles in remote patient monitoring and personal health management. Yet many of these devices do not have the basic security functions and do not provide privacy protections because the wireless data links used to transmit data and instructions are not encrypted (Carr 2013). Recently, many people are concerned with the HealthCare.gov website given the fact many entities can access the information on the website. The potential for misuse of healthcare information is great. In addition, patient portals present another source of risks. Patient Portals become a popular interface for patients to communicate with their healthcare providers. With patient portals, patients can leave a message for a doctor, make an appointment, request refills and enter health history data. Given so much personal health information available on the patient portal, security and privacy becomes a major concern.

It is important to note that data security and data privacy are different but interrelated issues. Data security prevents or grants access to data based upon authorization. Data privacy ensures that only those who have a valid need to view and/or utilize data can access data and that they work within the organization's policies. Privacy and security of patient health information must be at the core of any organization's policies. Policies and procedures should clearly define who has the right to access what part of health information and how data is protected, stored, and secured during transmission from one site to another. Security and privacy policies must be adhere to both federal and state laws.

3.4 Management and Policies

With the government mandating the application of electronic healthcare records systems, over half of the hospitals have implemented electronic medic records by 2012. As the data breach risk is quickly increasing, security and privacy takes the front stage. Data breaches are costly to any organizations. When such incidents happen, organizations face lawsuits, penalties and lost customers. Organizations must be vigilant about data security and privacy protocols. Organizations need to develop and implement various measures to strengthen its security and privacy. Organizations should set up a budget for security and privacy. Most of the health IT initiatives focus on meeting regulatory requirements, managing digital patient data, reducing costs, improving care, increasing clinical efficiency and collaboration. In the past security is not on top of the investment list, however, it should be a priority since the aftermath is far costlier. The average cost of encryption for a single device is $150. The average cost of an enterprise encryption system is between a quarter million to half a million dollars. Several vendors offer privacy-monitoring software to specifically meet the healthcare industry's needs. For high tech breaches, investment in encryption software is a necessity.

To manage security and privacy in the big data era, organizations should address not only external threats but also internal threats. In healthcare, traditionally many people do not think they are personally responsible for data management. In fact, technologies in healthcare are typically many years behind the financial industry and people in healthcare are less technology savvy and have limited knowledge of data security. Thus education and trainings are important in improving the understanding of privacy and security issues. Staff education is effective in dealing with low tech data breaches. Part of maintaining data security is educating end users in basic security measures and awareness of company policy. Educating employees regarding security dangers such as: clicking on email links, taping the password to the computer desk, or surfing unauthorized websites will help avert unintentional sabotage. Again, organizations need to install monitoring and assessment systems in place to ensure that employees are working within the boundaries of the organization's policies. According to Health Insurance Portability and Privacy Act providers may be responsible for their employees' data breaches in certain circumstances. With new threats coming in and continuous education on employees is the key to safeguard security.

Healthcare organizations develop policies to safeguard against security breaches and clearly define steps to deal with violations. Some of the measures use to protect patient data include security policy, data access monitoring, physical security, formal education among others. Data security should be a responsibility for the entire organization and not just specific departments. Developing this enterprise view on security will help reduce e so-called "low tech" security breaches and strengthen current policies and efforts. Therefore, organizations must utilize tactics that support such an initiative such as "rotating privacy and security audits to spot problems, including password on sticky notes and computers left unattended which displaying sensitive information; compliance with privacy and security rules; completion of education and training as metrics in performance evaluations; and daily hurdles to discuss patients on the floor who might spark curiosity – and inappropriate intrusion into their health information" (Birk, 2013).

It is important to have plans in action and be proactive. Reactive behaviors such as playing follow-up after the breach occurs are not effective. Healthcare data is ideal for the identity thief. Therefore, it is critical that the healthcare organizations setup risk management teams. The reality is that many organizations do not implement proper safety risk management when moving to web-based systems and the cloud. Organizations must always assess and reassess all systems and situations. It is crucial that proper steps are taken from the beginning to protect the organization against security and privacy breaches.

3.5 Big Security Data

Organizations of any size are exposed to the unprecedented number and variety of threats and risks to cyber security. Big Data will transform intelligence-driven models (Kar 2014). Research firm Gartner predicts that more than 25 percent of

global firms will adopt big data analytics for at least one security and fraud detection use by 2016.

The variety, volume and speed of security data have increased rapidly. As a result, analytic on big security data has become ever more important, especially as hacking algorithms and methods has become unexpectedly more sophisticated. The sheer volume of security related data makes it extremely difficult to identify a threat. However, big security data are useful in the safeguard of organizations security and the mining of security data making them proactive in defending their own networks and protecting organizations. McAfee (2014) security survey of IT decision makers show that 35% of organizations can detect data breaches within minutes of happening; 22% of organizations would need a day to identify a breach; 5% of the organizations would take up to a week. On average, it takes 10 hours for an organization to recognize a security breach. Given the serous security issues, big data analytics can play a critical role and allows organizations to access data, gain a complete view of business, perform effective security analysis, and detect advanced threats.

3.6 Security Products

Security experts have been predicting trends and challenges in the security activities. For example, Schwartz (2012) notes seven security trends for 2013 including: 1) mainstream cloud and mobile adoption seeks security;,2) businesses begin sandboxing smartphone apps; 3) cloud offers unprecedented attack strength; 4) post-flashback, cross-platform attack increase; 5) destructive malware targets critical infrastructure; 6) hackers target QR codes, TecTiles; 7) digital wallets become cybercrime targets. Hurst predicts ten security challenges for 2013 as follows: 1) state-sponsored espionage; 2) distributed denial of service (ddos) attacks; 3) cloud migration; 4) password management; 5) sabotage; 6) botnets; 7) insider threat; 8) mobility; 9) internet; 10) privacy laws.

With so many challenges identified and more to be discovered, security products can be effective in addressing challenges above. Table 2 shows Security Readers' Choice awards for top security products in 19 categories. In application security category, readers were asked to vote on Static and dynamic vulnerability scanners, and other source code analysis products and services used during development. In authentication category, readers voted on digital identity verification products, services, and management systems, including PKI, hardware and software tokens, smart cards, knowledge-based systems, digital certificates, biometrics, cell phone-based authentication. In Cloud security category, readers voted on Services and products designed to secure business use of cloud computing, including data encryption, identity and access management and network security.

In Data loss prevention category, readers voted on Network, client and combined data leakage prevention software and appliances for enterprise and midmarket deployments, as well as "DLP lite" email-only products. In email security category, readers voted on Antispam, antiphishing, email antivirus and

antimalware filtering, software and appliance products, as well as hosted "in-the-cloud" email security services. Includes email archiving and e-discovery products and services. In encryption category, readers voted on Hardware and software-based file and full-disk encryption, and network encryption products. In Endpoint security category, readers voted on business-grade desktop and server antimalware and endpoint protection suites that include antivirus and antispyware, using signature-, behavior- and anomaly-based detection, whitelisting, host-based intrusion prevention and client firewalls. In Enterprise firewalls category, readers voted on enterprise-caliber network firewall appliances and software, stateful packet filtering firewalls with advanced application layer and protocol filtering. In Identity and access management category, readers voted on User identity access privilege and authorization management, single sign-on, user identity provisioning, Web-based access control, federated identity, role-based access management, password management, compliance and reporting.

In Intrusion detection and prevention category, readers network-based intrusion detection and prevention appliances, using signature-, behavior-, anomaly- and rate-based technologies to identify denial-of service, malware- and hacker-attack traffic patterns. In Mobile data security category, readers voted on smartphone and tablet data protection products including antimalware, mobile access, platform-specific security (Android, iOS, Windows and BlackBerry), mobile device management, mobile application management and mobile application security. In Network access control category, readers voted on appliance, software and infrastructure user and device network access policy creation, compliance, enforcement (802.1X, client-based, DHCP) and remediation products. In Policy and risk management category, readers voted on risk assessment and modeling, and policy creation, monitoring and reporting products and services, IT governance, risk and compliance products, and configuration management. In remote access category, IPsec VPN, SSL VPN (stand-alone and as part of application acceleration and delivery systems) and combined systems and products, as well as other remote access products and services.

In SIEM category, readers voted on security information and event management software, appliances and managed services for SMB and enterprise security monitoring, compliance and reporting. In Unified threat management category, readers voted on UTM appliances that integrate firewall, VPN, gateway antivirus, URL Web filtering, antispam. In Vulnerability management category, readers voted on network vulnerability assessment scanners, vulnerability risk management, reporting, remediation and compliance, patch management and vulnerability lifecycle management. In Web application firewalls category, readers voted on standalone Web application firewalls and WAFs that are part of application acceleration and delivery systems. In web security category, readers voted on Software and hardware products, hosted Web services for inbound and outbound content filtering for malware activity detection/prevention, static and dynamic URL filtering and application control (IM, P2P).

Table 2 Security Readers' Choice Awards 2013

Security Category	Gold	Silver	Bronze
Application security	QualysGuard WAS, Qualys Inc.	Juniper Networks AppSecure, Juniper Networks Inc.	API Gateways, Layer7 Technologies Inc.
Authentication	SecurID, RSA, the security division of EMC Corp.	Symantec Managed PKI for SSL, Symantec Corp.	
	Symantec User Authentication Solutions, Symantec Corp.		
Cloud security	Juniper Networks vGW Virtual Gateway, Juniper Networks Inc.	Symantec Email Security.cloud, Symantec Corp.	Symantec O3, Symantec Corp.
Data loss prevention	Symantec Data Loss Prevention, Symantec Corp.	Websense Data Security Suite, Websense, Inc.	McAfee Total Protection for Data, McAfee, Inc.
Email security	Messaging Gateway powered by Brightmail, Symantec Corp.	Cisco Email Security Appliance (formerly IronPort), Cisco Systems	Google Message Security, Google
Encryption	Dell Data Protection - Encryption, Dell Inc.	Check Point Full Disk Encryption, Check Point Software Technologies Ltd	SecureData Enterprise, Voltage Security, Inc.
Endpoint security	Symantec Endpoint Protection 12, Symantec Corp.	Kaspersky Endpoint Security for Business, Kaspersky Lab	AVG AntiVirus Business Edition, AVG Technologies
Enterprise firewalls	McAfee Firewall Enterprise, McAfee, Inc.	Juniper Networks SRX Series Services Gateways for the Data Center, Juniper Networks, Inc.	Juniper Networks ISG Series Integrated Security Gateways, Juniper Networks, Inc.

Table 2 (*continued*)

Identity and access management	Oracle Identity and Access Management Suite Plus, Oracle Corp.	RSA Identity Protection and Verification Suite, RSA, the security division of EMC	CA IdentityMinder, CA Technologies Inc.
Intrusion detection and prevention	Juniper Networks IDP Series Intrusion and Prevention Appliances, Juniper Networks Inc.	Fortinet FortiGate, Fortinet Inc.	Check Point IPS Software Blade, Check Point Software Technologies Ltd.
Mobile data security	McAfee Enterprise Mobility Management, McAfee Inc.	AirWatch MDM, AirWatch LLC	Check Point Mobile Access Software Blade, Check Point Software Technologies Ltd.
Network access control	Unified Access Control, Juniper Networks Inc.	McAfee Network Access Control, McAfee Inc.	Cisco NAC Appliance, Cisco Systems
Policy and risk management	IBM Tivoli Compliance Insight Manager, IBM Corp.	VMware vCenter Configuration Manager, VMware Inc.	McAfee ePolicy Orchestrator, McAfee Inc.
Remote access	Check Point Remote Access VPN Software Blade, Check Point Software Technologies LTD.	Juniper Networks SA Series SLL VPN Appliances, Juniper Networks	Netgear ProSafe VPN Firewall, Netgear
SIEM	Splunk Enterprise, Splunk Inc.	HP ArcSight Enterprise Security Manager (ESM), Hewlett-Packard Co.	McAfee Security Information and Event Manager, McAfee, Inc.
Unified threat management	Dell SonicWall, Dell Corp.	Check Point Unified Threat Management, Check Point Software Technologies LTD.	FortiGate, Fortinet, Inc.

Table 2 (*continued*)

Vulnerability management	Shavlik Protect, LANDesk Software	QualysGuard Vulnerability Management, Qualys, Inc.	Nessus Vulnerability Scanner, Tenable Network Security
Web application firewalls	Citrix NetScaler AppFirewall, Citrix Systems Inc.	FortiWeb-400C, Fortinet, Inc.	F5 Networks BIG-IP Application Security Manager, F5 Networks
Web security	Websense Web Security Gateway, Websense Inc.	Blue Coat Systems ProxySG appliances, Blue Coat Systems, Inc. 90-100 words	Symantec Web Security.cloud, Symantec Corp.

Source: http://searchsecurity.techtarget.com/essentialguide/Security-Readers-Choice-Awards-2013

4 Big Data Analytics

4.1 Overview

Big data analytics applies advanced analytics to very large data sets. According to a 2009 TDWI survey, 38% of organizations surveyed reported practicing advanced analytics, whereas 85% said they would be practicing it within three years. Forrester (2013) reports that 70% of IT decision makers consider big data a top priority now or in a year. In addition, a majority of companies that Forrester surveyed estimate that they are only analyzing 12% of the data they have. Big Data is changing the way of products, solutions, and services are being marketed. McKinsey & Company (2012) reported that big data and improved analytics can improve sales by $200 billion.

Big data analytics deal with complex data types. Querying such complex data types could be challenging and analysis consumes large amount of resources: storage, memory, CPU. Query performance could be affected. Building a solid infrastructure to support fast data through output and improving query performance are critical in big data analytics. Currently, analytics are used in reporting, dashboard, performance analytics, web analytics, and process, predictive, location analytics, advanced visualization, text analytics and streaming analytics.

4.2 Technologies

New technology – Hadoop holds the promise big data analytics. Hadoop and related products make data capturing, storage and analysis in a cost effective way. When investing in big data technologies, scalability is the key. Organizations must think ahead and be future-oriented. The volume of data is exploding and a variety of devices (mobile phones, sensors, websites) produce different data types (web data, image files, video and audio files). The new innovative devices coming to market will continue to change the future of the big data market. Thus, big data infrastructure must be able to accommodate future data growth needs.

Hadoop is a distributed file system that handles massive volumes of file-based unstructured data. It becomes the de facto standard for big data technologies. It is an open source software project administered by the Apache Software Foundation. Because Hadoop is essentially a distributed file system, it lacks some functionality of database management systems. Furthermore, a set of related software technologies (Pig, MapReduce, Hive, HBase) work together to become the Hadoop family of products. Both the Apache Software Foundation and several software vendors offer Hadoop family products. With the high hopes for Hadoop in dealing with big data, more and more vendors make an effort in integrating their products with Hadoop.

Pig is a platform for analyzing large data sets that consists of a high-level language for expressing data analysis programs and it can be used to create mapper and reducer to work on large data sets. Pig consists of two components: PigLatin and the runtime environment. MapReduce is a software framework for creating applications to handle vast amounts of data (multi-terabyte data-sets) in-parallel on large clusters (thousands of nodes) of commodity hardware. Hive is the Apache data warehouse software performing querying and managing large datasets located in distributed storage. Hive is not a standard SQL but resemble SQL. HBase is a non-relational database and can host very large tables (billions of rows and millions of columns) for random, realtime read/write access. HBase was modeled after Google's Bigtable and has similar functions as Bigtable.

The commonly referred Hadoop essentially contains a set of related technologies in the big data environment and each of these technologies has their own unique advantages in handling and processing large data sets. Together, Hadoop, Pig, MapReduce, Hive, Hbase etc. leverage computing resources in order to efficiently process big data. Hadoop is designed for multi-structured data types whereas a data warehouse handles structured data.

According to TWDI survey (2012), most organizations are planning to integrate Hadoop into their existing architecture. TDWI also predicts that Hadoop technologies will complement the well-established products and practices for business intelligence (BI), data warehousing (DW), data integration (DI), and analytics. However, Hadoop is a new technology, its security and administrative tools need improvements. In addition it is hard to find Hadoop users and technical professionals. As time passes by, there will be more Hadoop users. 10% of organizations surveyed have a Hadoop implementation in production today (Russon 2013)

For example, an organization can store aggregated web log data in their relational database, while keeping the complete datasets at the most granular level in Hadoop. This allows them to run new queries against the full historical data at any time to find new insights, which can be a true game-changer as organizations aggressively look for new insights and offerings to differentiate from the competition. The popular Hadoop products include MapReduce, HDFS, Java, Hive, HBase, and Pig. Mahout, Zookeeper, and HCatalog are taking off.

Enabling big data analytics is the leading benefit of Hadoop, whereas a lack of Hadoop skills is the leading barrier. BI/DW aside, a few respondents also anticipate using Hadoop as a live archive (23%) or as a platform for content management (35%).

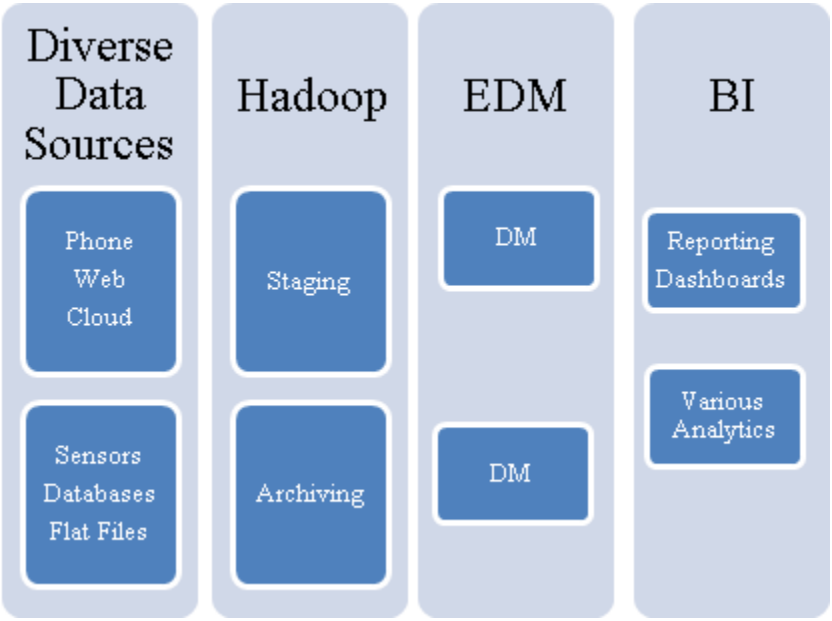

Fig. 1 Big Data Architecture here

4.3 Business Decision Making

Big data analytics operate on large data sets captured by diverse devices from sensors, devices, third parties, web applications, and social. Detailed and granular level data of business operations can be analyzed and new sights can be generated. Advanced analytics techniques such as: predictive analytics, data mining, statistics, and natural language processing can be applied to unstructured and structured data types consequently create fact-based decisions. Results derived from big data analytics may help in the areas of customer segmentation, fraud detection, risk analysis, and tracking evolving customer behaviors. Business decisions that use

deep and advanced data analytics provide benefits on operational, tactical and strategic levels. Furthermore, there are many advantages of implementing big data analytics, which result in optimization, performance improvement, costs reduction of the organization as a whole.

While long-term expectations are high, most of healthcare executives say they have yet to see substantial benefits from big data. The Society of Actuaries (SOA 2013) recent report found that 66% of leaders are enthusiastic about the potential of big data, while more than 87% said that data analytics will have an important impact on the business of healthcare in the future. Another half of payers saying there is substantial business benefit (53%) by implementing big data analytics.

Healthcare has collected enormous amount of data across many disparate systems. In order to reduce costs and improve care outcomes, analytics must be used. In the past, structured data have been used in the administrative and financial area of healthcare, with more and more text data captured. In terms of patient care, text data provide more valuable insights in understanding cause and outcome. Textual and predictive analytics tools can reveal hidden patterns.

5 Discussion

Though many organizations have and are eager to jump on the big data wagon, implementing big data solutions is not easy. It is quite common that large organizations have technologies, tools, and architectures in different generations with diverse, complex information compliance and security policies. Business Leaders are not as confident about data as the volume, variety and velocity of information increases. Data quality issue still persists. If business leaders do not think data they have is trustworthy, they will not make decisions based on analytics generated on data.

Data quality, privacy and security are the main issues that should be tackled by any organization. Processes and procedures are the root causes for data quality issues and these cannot be corrected by technologies. A disciplined methodology should be followed through in data quality management.

Data quality in the big data area is more complex. However proven data management methodologies still apply to the big data environment. To be effective in big data management, data quality tools, data integration tools, data stewardship, policies and procedures must be in place. In addition, top management support in big data quality initiative is also a necessity. Furthermore, to ensure success in big data implementations, it is important to focus on security intelligence solutions.

5.1 Market Demand for Big Data Talents

McKinsey Global Institute (2011) reports that there will be a shortage of 140,000 to 190,000 people with deep analytical skills and 1.5 million managers and analysts who have the analytic skills of big data for effective decision-making. In addition, most organizations are not prepared to address both the technical and management challenges posed by big data.

In healthcare, the shortage of big data skills is also a concern. The overall shortfalls of resources including staffing, budget, and infrastructure are the biggest barriers to the adoption of big data analytics in their organization. Payers and providers have difficulty obtaining the funding for recruiting staff skilled to gain the full benefits of big data. Most providers and payers have difficulty finding staff that can consolidate complex datasets and glean actionable information from them. In addition, it is hard for payers and providers to hire the right talent who can identify the business opportunities big data provides. It would be good idea for healthcare to find talents in other industry.

Furthermore, we need to educate professionals in the importance of big data security. It has become obvious that technology alone will not solve security issues. People are backbone in in protecting cyberspace. Though security education has been addressed in the past, threats and attacks have been increasingly insidious and harmful. In order to be more effective, we need to hold mandatory security training session for all employees on all aspects: workflow, process, security policies, software, antivirus software etc. For universities, we need to offer full-blown curriculum addressing security to better prepare students for cyber defense.

5.2 Big Data Solutions Implementations

When it comes to big data solution implementations, Forrester (2013) finds six challenges as 1) integrating big data solutions in a complex, heterogeneous data management environment; 2) technical implementation skills with employees; 3) meeting business demand for big data analytics; 4) understanding business values of big data solutions; 5) infrastructure budget constraints; 6) difficulty in finding and hiring people with needed skills. In addition, Forrester (2013) also recommends seven qualities for production-ready big data solutions: 1) manageability, 2) availability, 3) performance, 4) scalability, 5) adaptability, 6) security, 7) cost.

5.3 Analysis of Big Data Publications

We analyzed 220 white papers on big data published between 2010-2014 using SAS text miner. We examined the concept links of the top four terms (data, big, business, big data) based on frequency. Data is linked to analytics, big, base, design, big data, analyze, type, software, and analysis. Analytics links to unstructured, predictive, unstructured data, business intelligence, visualization, predictive analytics and sensor. Big data is linked to Hadoop, network, open, storage, industry, variety, amount, source, enterprise, support, access, information, cluster, and exist. Big, is linked to development, industry, big data, software, environment, manage, strategy and challenge. Business, performance, thing, decision, practice, result, opportunity, insight, strategy, and analytics.

Table 3 shows the frequencies of terms in descending order in the 220 white papers we selected. As shown TDWI publishes most papers related in big data. The combination of words tdwi, user, organization and analytic are commonly

mentioned in 10 documents. Next frequently occurred word group is pillar, four, governance, Hadoop and node and the combination of these words occur in 16 documents.

Table 3 Term/Topic Frequency

Document Cutoff	Term Cutoff	Topic	Number of Terms	# Docs
0.554	0.038	tdwi,+tdwi,+user organization,+analytic,+bi	713	10
0.539	0.036	+pillar,four,governance,hadoop,+node	709	16
0.569	0.039	+integrator,+five,+data integration,+integration,heterogeneous	672	14
0.553	0.042	intelligence unit,+organisation,limited,+unit,rst	648	11
0.608	0.043	hr,hr,+bi,+customer,+workforce	643	26
0.637	0.045	hadoop,+hadoop,+home,+big,+big data	630	22
0.26	0.03	+mobile,digital,+agency,+big data,+sector	607	30
0.32	0.031	+quality,+review,+propose,+theory,+opinion	603	20
0.443	0.034	+analytical,+analytics,+bi,+predictive,+big	580	34
0.487	0.037	+patient,+healthcare,clinical,+readmissions,+health	575	12
0.496	0.038	,+myth,+quality,integrity,+data quality	566	13
0.429	0.036	+builder,iway,+quality,+information builders,ltd	558	10
0.397	0.035	governance,home,+home,+li,+data governance	557	22
0.498	0.035	+federation,hadoop,+lasr,hdfs,+cache	556	20
0.486	0.038	infosphere,+biginsights,+puredata,+puredata system,hadoop	527	26
0.449	0.034	+inmemory,+cube,cognos,+dynamic,+cache	520	17

Table 3 *(continued)*

0.494	0.037	hadoop,forrester,+forrester wave,edw,+wave	499	10
0.227	0.028	+sap,hp,hp,+tb,+in-memory	496	19
0.393	0.033	+chart,+visual,+visualization,+plot, +visualize	490	14
0.33	0.031	+backup,+encryption,+protection,+cloud,data protection	490	9
0.158	0.026	+chapter,+springer,+book,+quo,+style	488	13
0.468	0.041	+security,+threat,+attack,+malware, +breach	456	15
0.415	0.038	+title,+reference,+style,+pt,+head	404	9
0.431	0.037	s p o n s,o f,pa g e,pa,e d b y home	359	19
0.283	0.029	+bi,+rs,+benchmark,+dan,+ar	357	10

Table 4 shows terms appeared in documents in descending order. The word group (analytics, analytic, BI,big, predictive) appears in 34 documents; the word group (mobile,digital,+agency,+big data,+sector) occur in 30 documents; the word group (hr,hr,+bi,+customer,+workforce) appears in 26 documents. The occurrence of the words show the importance of the concepts.

Table 4 Document Frequencies

Document Cutoff	Term Cutoff	Topic	Number of Terms	# Docs
0.443	0.034	+analytical,+analytics,+bi,+predictive, +big	580	34
0.26	0.03	+mobile,digital,+agency,+big data,+sector	607	30
0.608	0.043	hr,hr,+bi,+customer,+workforce	643	26
0.486	0.038	infosphere,+biginsights,+puredata,+puredata system,hadoop	527	26
0.637	0.045	hadoop,+hadoop,+home,+big,+big data	630	22
0.397	0.035	governance,home,+home,+li,+data governance	557	22
0.32	0.031	+quality,+review,+propose,+theory,+opinion	603	20

Table 4 *(continued)*

0.498	0.035	+federation,hadoop,+lasr,hdfs,+cache	556	20
0.227	0.028	+sap,hp,hp,+tb,+in-memory	496	19
0.431	0.037	s p o n s,o f,pa g e,pa,e d b y home	359	19
0.449	0.034	+in-memo-ry,+cube,cognos,+dynamic,+cache	520	17
0.539	0.036	+pillar,four,governance,hadoop,+node	709	16
0.468	0.041	+security,+threat,+attack,+malware,+breach	456	15
0.569	0.039	+integrator,+five,+data integra-tion,+integration,heterogeneous	672	14
0.393	0.033	+chart,+visual,+visualization,+plot,+visualize	490	14
0.496	0.038	,+myth,+quality,integrity,+data quality	566	13
0.158	0.026	+chapter,+springer,+book,+quo,+style	488	13
0.487	0.037	+patient,+healthcare,clinical,+readmissions,+health	575	12
0.553	0.042	intelligence unit,+organisation,limited,+unit,rst	648	11
0.554	0.038	tdwi,+tdwi,+user organiza-tion,+analytic,+bi	713	10
0.429	0.036	+builder,iway,+quality,+information builders,ltd	558	10
0.494	0.037	hadoop,forrester,+forrester wave,edw,+wave	499	10
0.283	0.029	+bi,+rs,+benchmark,+dan,+ar	357	10
0.33	0.031	+backup,+encryption,+protection,+cloud,data protection	490	9
0.415	0.038	+title,+reference,+style,+pt,+head	404	9

5.4 Big Data Security

Cyber security is of critical importance to any company and individual. Cyber-attacks have grown increasingly sophisticated, stealthy, and dangerous. According to the Verizon 2013 Data Breach Investigations Report, 66% of breaches are not

discovered for months. Security intelligence is a proactive strategy in advance to manage security. Though companies have many security systems, typically these security systems do not talk to each other. Integrating security data from many disparate systems is essential in security management. More security data, transaction data, unstructured data can help identify vulnerability and filter out noise. The main sources of vulnerability can be classified into two categories: internal and external. In general for internal vulnerability, human errors are the weakness links in the security model. Security experts should be able to analyze if vulnerability will be raised because of errors.

With the increasing level of sophistication of threat compounding the fact that disparate security systems cannot talk to each other, it is infeasible to obtain a consolidated view about security situation. Some attacks occur quietly and attack slowly. In reality, it's not a matter of if attacks come in; it is a matter of when attacks come in and how long it takes for the security team to figure out. Security teams should start with normal behavior, normal network traffic, and normal user activities. Using the normal behaviors as security baseline will help find anomaly.

It is critical that organizations build a model to address Advanced Persistent (APTs). APTs are hard to identify with disparate security systems. Insider threats come from entities who have the motivation and ability to harm enterprises. APT is a set of stealthy and continuous hacking processes which usually intend to harm organizations, nations for business or political motives. APT processes can be very long and slow. This type of malware hides itself on the system over a long period of time.

The Ponemon Institute reports that 67% of organizations state their current security activities are insufficient to stop a targeted attack. Targeted attacks are difficult to predict, diagnose and defend. A custom attack requires a custom defense. There are companies specializing security technologies that detect and analyze APTs and targeted attacks.

APTs are dangerous and its hackers are well organized with an intent to steal valuable intellectual property, such as confidential project descriptions, contracts, and patent information. Grimes (2012) suggest five signs to watch for APTs: 1) increasing in elevated log-ons late at night, 2) finding widespread backdoor Trojans, 3) unexpected information flows, 4) discovering unexpected data bundles, 5) detecting pass-the-hash hacking tools. In addition, Frank and Watson (2013) recommend the importance of detecting account abuse by insiders and APTs, pinpointing data exfiltration by APTs, alerting new program execution.

Privacy of patient information must be a key priority when implementing the new analytics systems. In addition, the information should be secure, especially when being transported via the web and/or the cloud from one place to another. HIPAA requires that this data be private and secure. As web-based systems and cloud computing gains ever increasing popularity, these types of systems bring their unique characteristics, which expose to new security breaches. Organizations should implement proper safety risk management before moving to web-based

systems and the cloud. Organizations need to continuously assess and reassess all systems and situations for possible data breaches. It is crucial that proper steps are taken from the very beginning to protect the organization against security and privacy breaches.

6 Conclusion

Big data has generated a lot of interest not only in industry but also in academia. In this book chapter we have discussed data quality, data integration, privacy, security and analytics. We believe that these topics continue to be important in the next decade. New technologies, techniques, policies are to be developed to keep up with big data development. In addition, ethical and legal issues regarding big data present numerous challenges and opportunities for researchers.

According to IDC, about 22% of digital information is suitable for analysis, a lot of data reside in data silos (Lev-ram 2014). Data integration efforts continue to sustain and the market for technological vendors are becoming more promising. Data centers will continue to grow. In the future big data will become bigger data (Lev-ram 2014). More and more companies will invest in their data centers and expect to derive values from these.

Data quality is of interest to both practitioners and academia. To ensure data quality, technologies, people, policies and procedures working together to produce desired effect. Data quality must be ensured into order for data analysis to be reliable. The commonly used dimensions of data quality will be cherished across organizations: data accuracy, relevancy, timeliness, trustworthy. As data is treated as an asset of an organization, the value of data quality should be emphasized across the different hierarchies.

Only 5% of digitized data is currently being analyzed (Lev-ram 2014). Given such a small percentage， big data analytics has many areas waiting to be explored. Some suggest technologies, new statistical techniques are useful for big data analysis and others state the future of big data analytics should focus on effect size and variance explained (George 2013) rather than p values. In addition, big data visualization is another key research directions in the future as it makes big data meaningful by transcribing massive amount of information into images that are easy for people to understand.

Finally, in the education arena, a few universities across the country started data science as a new major. The purpose of such a discipline is to train students in big data technologies and big data analytics to meet the emerging market demand. In fact, with the new job title coming to the market (e.g. chief financial technology officer), CFOs are required to understand technology and big data. Brand (2014) states that “The future of the accountancy profession lies at the intersection of finance, technology and information”. As an educator, we need to keep up with technological developments and collaborate with our colleagues to provide cross disciplinary skills for students.

References

Birk, S.: Protecting patient medical data. Healthcare Executive 28(5), 20–28 (2013)

Brands, K.: Big Data and Business Intelligence for Management Accountants. Strategic Finance 96(6), 64–65 (2014)

Brand, H.: Big data: adapt or die (2014), https://www.accountancylive.com/big-data-adapt-or-die (accessed June 15, 2014)

Carr, D.F.: Hackers outsmart pacemakers, fitbits: worried yet? InformationWeek (2013), http://www.informationweek.com/healthcare/security-and-privacy/hackers-outsmart-pacemakers-fitbits-worried-yet/d/d-id/1113000?image_number=3 (accessed June 14, 2014)

English, L.P.: Information quality applied: best practices for improving business information, Processes and Systems. Wiley (2009)

Forrester, Is your big data solution production-ready? (2013), http://www.itworld.com/data-center/417766/your-big-data-solution-production-ready (accessed June 15, 2014)

Fulgoni, G.: Big data: friend or foe of digital advertising? Five ways marketers should use digital big data to their advantage. Journal of Advertising Research 53(4), 372–376 (2013)

Gartner report, Magic quadrant for data quality tools (2013), http://www.gartner.com/technology/reprints.do?id=1-1LE6U4H&ct=131008&st=sg (accessed February 26, 2014)

Kim, G.-H., Trimi, S., Chung, J.-H.: Big-data applications in the government sector. Communications of the ACM 57(3), 78–85 (2014)

George, G., Haas, M.R., Pentland, A.: Big data and management. Academy of Management Journal 57(2), 321–326 (2014)

Research Moz, Global data quality tools market is expected to reach a CAGR of 16.78% in 2016 (2013), http://www.prweb.com/releases/2013/11/prweb11352256.htm (accessed February 25, 2014)

Green, C.: Organizations will rapidly ramp up their data services in 2014 (2014), http://blogs.forrester.com/charles_green/14-02-06-organizations_will_rapidly_ramp_up_data_services_spend_in_2014 (accessed February 25, 2014)

Grimes, R.: 5 signs you've been hit with an advanced persistent threat (2012), http://www.infoworld.com/d/security/5-signs-youve-been-hit-advanced-persistent-threat-20494 (accessed March 24, 2014)

Hurst, S.: Top 10 security challenges for 2013. SC Magazine (2013)

IBM Corporation. Three guiding principles to improve data security and compliance: A holistic approach to data protection for a complex threat landscape (2012)

Kar, S.: Gartner report: big data will revolutionalize cybersecurity in the next two years. CloudTimes (2014)

Lawson, L.: Eight questions to ask before investing in data quality tools (2014), http://www.itbusinessedge.com/blogs/integration/eight-questions-to-ask-before-investing-in-data-quality-tools.html (accessed February 26, 2014)

McAffee, Needle in a datastack: the rise of big security data (2013), http://www.mcafee.com/us/about/news/2013/q2/20130617-01.aspx (accessed January 15, 2014)

Lev-ram, M.: What's the next big thing in big data? Bigger data. Fortune 169(8), 233–238 (2014)

Mcknight, W.: Seven sources of poor data quality. Information Management 19(2), 32–33 (2009)

McKinsey Global Institute, Big data: next frontier for innovation, competition, and productivity (2011)

McMillan, M., Cerrato, P.: Healthcare data breaches cost more than you think. InformationWeek Reports (2014)

Nunan, D., Di Domenico, M.: Market research and the ethics of big data. International Journal of Market Research 55(4), 2–13 (2013)

Petter, S., DeLone, W., McLean, E.R.: Information systems success: the quest for the independent variables. Journal of Management Information Systems 29(4), 7–62 (2013)

Russom, P.: Integrating hadoop into business intelligence and datawarehousing. TWDI Research (2013), http://www.cloudera.com/content/dam/cloudera/Resources/PDF/TDWI%20Best%20Practices%20report%20-%20Hadoop%20foro%20BI%20and%20DW%20-%20April%202013.pdf (accessed February 15, 2014)

Schwartz, M.J.: 7 Top Information security trends for 2013. InformationWeek (2012), http://www.darkreading.com/risk-management/7-top-information-security-trends-for-2013/d/d-id/1107955? (accessed January 23, 2014)

Setia, P., Venkatesh, V., Joglekar, S.: Leveraging digital technologies: how informa-tion quality leads to localized capabilities and customer service performance. MIS Quarterly 37(2), 565-A4 (2013)

Smith, R.F., Watson, B.: 3 Big data security analytics techniques you can apply now to catch advanced persistent threats. HP Enterprise Security (2013)

Society of Actuaries, Healthcare decision makers perspectives on big data (2013)

TDWI's Data Quality Report, http://tdwi.org/research/2002/02/tdwis-data-quality-report.aspx (accessed March 2, 2014)

Verizon. 2013 Data Breach Investigations Report, http://www.verizonenterprise.com/resources/reports/rp_data-breach-investigations-report-2013_en_xg.pdf (accessed March 2, 2014)

Woods, D.: Why data quality matters (2009), http://www.forbes.com/2009/08/31/software-engineers-enterprise-technology-cio-network-data.html (accessed February 25, 2014)