Survey: Securing The Privacy In The World Of Big Data

Shilpa Batra, Pankaj Kumar, Arindam Batabyal
Amity University, ASET, U.P. Noida[1,2] , Gurgaon institute of technology & management[3]

Abstract

As we know that with the increase in expansion of internet and data sets with the passage of time, big data has taken birth. As of 2012, the size of data sets has grown tremendously due to accumulation of information from unambiguous sensing like internet search, finance, microphones, software logs etc. The capacity to store data has roughly doubled every 30months since 1980's. Big data is difficult to manage by traditional RDBMS and needs massive parallel servers running in tens and hundreds number. What matters is how an organisation manages and analyses its data sets. Firms like Sloan digital sky survey (SDSS) stores about 140TB of astronomical data; NASA stores 32PB of climatic information and simulation. Big data has served a critical role for United State President Obama's 2012 re-election campaign. Amazon.com handles about 7.8TB of data; Walmart handles 2.5PB of customer transactions and information and Facebook handles around 50 billion photos of user database. The data stored by these crucial organisations is highly confidential and critical. So, there arises the need of securing this amount of vast data as Big Data is distributed in nature. In this paper we will throw some light on the sources of attack on the databases and methods to prevent such attacks.

1. Introduction

Big data refers to the very large complex set of data with sizes beyond the commonly used software tools to capture, analyse, store, manage and process it within a specified time limit. The size of data has constantly moved from few terabytes to petabytes in a decade. This is the result of the improved technology which was followed in traditional DBMS and the new ability to handle large databases. It has been found in experiments that about 150million sensors deliver 40million times data per second and there are around 600million collisions per second. The collection of such an enormous size of data has been derived from sources like weblogs, internet documents, astrology, military surveillance, forecasting, medical records etc.

Due to the abundance of data set IBM has discovered that this big data can be categorised into 3 genres namely volume, variety and velocity abbreviated as 3V's.

The VOLUME factor deals with the enterprise's capacity to manage the ever-growing data size from some terabytes to petabytes of information.

The VELOCITY factor deals with the ability to process time sensitive information like catching frauds. For example, to analyse 200million daily call detail records in real time to predict the customer details faster.

The VARIETY factor deals with the data sets included in the big data. These can be structured as well as unstructured such as text, audio, video, log files etc.

But on the other side, according to SAS, there are 2 more dimensions while considering big data that are variability and complexity. The former determines the rate at which the data is being added or accessed .Daily, seasonal and event triggered peak data loads can be a big challenge to manage for social networking. The event triggered peak data deals with the huge volume of data coming from multiple sources. It is difficult to link, match,cleanse and transform data across systems. Therefore, it is essential to correlate relationships among databases to produce high quality information that is appropriate and up to date. The technologies required to deal with the big data's efficient management should comprise the features like cheap and abundant storage , high server processing capacity , affordable large memory like Hadoop, new storage technology , cloud computing and other flexible resource allocation.

Since big data is not dealt by the conventional databases, think about the environment that comprises big data. They use many notes for distributed data management and processing. There aremultiple copies of data across various nodes. This helps to let the environment provide parallel processing if one of the node fails. Since the data is too large, distributed and parallel, therefore, it becomes difficult to secure systems. The clusters are self-organising and allow multiple users to communicate and share data. Validating the user to access the information becomes a challenging task. The flexible nature of big data allows new nodes to be added to the existing cluster at any time and share data. Since we want security to be added to the big data; scalability,performance and self-organising issues are important factors.

Whether it is a monetary gain or revenge, hackers want to intercept an organisation's data. Databases are the pivotal point of attacks and cybercrime. Therefore, it needs to be guarded. To ensure the database security; It is important to determine, the types of attack,their sources to DBMS,and what they do once they get access.Once, it is determined, we

can guard our database with various fences of methodologies and technologies in order to prevent future attacks. The issues of securing gigabit networks, visualising the layered topology of network and improving the performance becomes a secondary issue. Big data brings the ease of expansion and easy accessibility but to deal with the security and thefts has become a critical issue of concern that will be put to light later in the paper.(WIKIPEDIA)

2. ATTACKERS APPROACHES TO DATABASES

Last year, it was found out that hackers gain crores of rupees through cybercrime. In that case, the hackers intercepted the debit cards of Fidelity National Information Services and accessed the company's database, which contained the information of the withdrawals and the limits cardholderswithin a 24-hour period. From this information, they either increase or eliminate their withdrawal limit. It was observed that the hackers duplicated(cloned)the cards and sent copies to fellow gang members in Greece, Russia,Spain, Sweden, Ukraine and the United Kingdom. When the prepaid amount on a particular card got low, the hackers would reload the fraudulent card remotely. All told, FNIS warned, as many as 7,170 prepaid accounts may have been put at risk, and three individual cardholders' personal information may have been disclosed.

When taking into account monetary losses and bad publicity; they underscore the importance of strong database security policies. If data is a target, then the walls around that target must be as difficult to scale as possible. "Servers and databases have continued to be a crucial target for hackers"says John Harrison, senior manager at Symantec Security Response.

The open source SQL map tool can attack using five different SQL injectiontechniques or directly, if the user has DBMS credentials, an IP address, port and database name. It can enumerate users and password hashes, with in-line support to crack them with a dictionary-based attack, and supportsprivilege escalation through metasploit, get system command. But planning a good defence means understanding a good offense—in this case, what hackers do and what they want. From SQL injectionvulnerabilities to user provisioning to poor configurations, attackers will use whatever foothold they can find to circumvent security.

2.1 Entering through the side window[2][3]

It cannot be said that all the hackers want to steal your data; they might want to just splash it over internet. For instance, the hacking that occurred last year in the security of think tank Stratford Global Intelligence, in which hackers stole confidential information. The hackers were associated with 58% of the stolen records in the breaches. The firm's insider people like database administrator Steven Jingo Kim, who was sentenced for a year in prison in 2010 after admitting that he accessed the GEXA Energy corporate network after he was fired from the company, stole information and tampered with the database. It is difficult to predict what motivates the hackers—be it greed, anger or the desire to make a political statement—but what is certain is that they want your data. And to achieve what they want, they need to penetrate your database. This implies cracking the key—also known as a weak password—for the door that is guarding it or poking a hole in the door in order to tear it wide open.

However, it might also mean using a side window, so to speak, and many of today's hackers do so via Web applications vulnerable to SQL injectionSQL injection vulnerabilities remain a headache for Web app developers, security professionals and database administrators. The responses made SQL injection the most-cited attack vector on a list that included cross-sitescripting and privilege escalation.SQL injection attacks exploit non-validated user input to issue commands through an application to a back-end database. Finding the holes through which these attacks can be launched isn't all that difficult. One of the first things attackers like to do is to see how an applicationhandles errors. Another way to search for vulnerable sites is through Google hacking.Google hacking uses search engines to find security gaps by leveraging the mountains of data they index. An attacker might start by entering a search query—called a Google Dork—designed to locate results that could offer a clue about sites that might be vulnerable. There are a number of Google Dorks that can be useful for a hacker searching for SQL injection vulnerability to exploit.Examples wouldbe queries like allinurl:index.php?id= and allinurl:article.php?ID=.

From an enterprise perspective, dealing with the problem of Google Dorks is tricky. The ultimate solution is to secure the site susceptible to such an attack. Security managers can start to do so by making Google hacking part of the penetration tests.They launch to check the security of their companies. However, as any good attacker knows, Google hacking tells only part of the story. To determine if a site is truly vulnerable to SQL injection attacks, further testing is required. Fortunately, SQL injection vulnerabilities are relatively easy to prevent. One way is to use parameterized queries and stored procedures. According to the Open Web Application Security Project (OWASP), developers should consider using parameterized queries

(Preparedstatements) that use placeholders for parameters whose values are ultimately supplied at execution time. Though taking this approach can negatively impact performance, it can help block attacks.Using stored procedures can have the same effect, though dynamic SQL within stored procedures can still be vulnerable ifuser input is not properly sanitized. One final approach to tackle the SQL injection problemis to escape user input before putting it in a query. Professionals and developers can closethe gaps through which SQL injection attacks wend their way into corporate systems, at the end of the day, finding and exploiting SQL injection vulnerabilities is not that difficult—especially for skilled attackers.

There is no shortage of tools to help hackers. In fact, some of the tools available are actually intended for the kind of ethical hacking done as part of penetration tests. One of the more popular of these is SQLmap, an open source pen testing tool that features a wide range of functionality, including a SQL injection detection engine, data-fetching capabilities and the ability to enumerate databases. Other tools focused on SQL injection include SQLninja, Havij and Priamos.

2.2 Attacking the LOCK[2]

Hacking databases is not all about SQL injection.However,in addition to Web applications, attackers often target databases by compromising the network. Hacking a corporatenetwork can begin with a tool like Nmap, which enables attackers to scan for open ports and map targeted networks. These port scans can give hints as to what type of database an organization is using. For example,Microsoft SQL Server defaults to Port 1433, while the Oracle Database listener typicallyuses 1521.

Once an attacker has compromised the network, the hacker can turn to password crackers. The use of weak, blankand default passwords—not to mention the repetition of passwords—continues to be aproblem throughout the world of IT."Many passwords are shared among different databases. Just to give a short example, for an Oracle database, one has to select four different passwords for each instance: SYS,SYSTEM, OUTLN, SNMP," says Sean Roth, manager of product marketing for database security at McAfee. "If you have thousands of instances, this is a huge amount of passwords you need to manage. Most [people] either use the exact same password or some sort of aschema. The end result is that if hackers gain access to one database, they gain access to many more."As Roth points out, sharing passwords may be convenient, but it can also buy an attackeraccess to multiple databases once he or she has managed to crack or brute-force thepassword of one. For these reasons, organizations can save themselves some potentialheadaches by replacing any default passwords and ensuring that database administratorsand users are following best practices with regard to password complexity.

Alongthese same lines, organizations should consider using timeouts or lockouts after multiplefailed login attempts. Then there are the actual database vulnerabilities themselves. Like other software, databases occasionally have bugs that can be exploited by attackers. Database administratorsreceived a reminder about this recentlyafter security researcher. Joxean Koret publicized details of an Oracle vulnerability that affects the TNS Listener component responsiblefor routing connections from the client to the database server. If exploited, the flaw can enable attackers to intercept any connection between databases and clients without any user authentication. Vulnerability scanning products can help inthis regard by uncovering unpatched security holes in databases.

"Databases are critical components that make up part of core information systems ofmodern-day organizations," Shulman says "Changing the software of such a component isn't a process to be taken lightly. Any change to a production system component requires rigorous testing in a staging environment and a maintenance window for deploying to production. With such a critical component that usually affects a number of applications, and given that side effects could be unpredictable,I totally understand why organizations are lagging with respect to database patching." For that reason, it is important for organizations to have a patch management strategy that prioritizes at least critical database updates.

2.3 Abusing AccessPrivileges[2]

Network and security misconfigurations can be just as lethal to database security as unpatched software. Take, for example, the March breach of Medicaid and Child Health InsurancePlan benefit recipients in Utah. In that case, unencrypted personal information includingSocial Security numbers, in some cases—belonging to nearly 800,000 people was exposed due to configuration and policy mistakes. In its analysis of the breach, the state found that not only was the server not being protected by a firewall, but that factory issued default passwords were still being used.A similar situation occurred at the University of North Carolina at Charlotte, where configuration errors and incorrect access settings made during a system upgrade exposed data for nearly three months. In a separate situation, data belonging to the University's College of Engineering was discovered to have been exposed over the Internet for more than a decade. As these incidents show, misconfigurations in security tools can leave open pathways to data. Organizations should also be wary of database features that are not being used and are unnecessarily enabled. Assigning the proper privileges requires having an understanding of the business needs of a given user or user group. From there, you can map out what users need access to and what controls need to be inplace. In principle, least privilege means not giving anyone more access than needed to do thejob. The same is true for an application accessing the database. If, for example, an applicationis meant to provide read-only views of data, it should not have the ability to change information. Not taking a least privilege approach increases the amount of damage an attacker can do if he or she somehow manages to compromise an account and take on the access rights of victim.

Determining user privileges should be done at the beginning of the identity management life cycle. But while one might assume that special attention is always paid to managing identities, it is actually an area in which many organizations fall short. According to the IOUG survey, 54% of respondents said they do not monitor new account creation. An orphaned account is an account that belongs to a user who is no longer active, such as someone who was fired from the organization. Thirty-eight percent of respondents said they had no way of determining whether a current or former employee used an orphaned account to access information, and 15% said such activity had occurred once. The subjects of super-user accounts and separation of duties are closely related to theconcept of least privilege. In the IOUG survey, nearly half of the respondents confessed thatthey do not monitor all privileged user activities on production databases—a stat that is troubling both in terms of dealing with external hackers and in terms of the prospect of internal threats. The latter category has received more attention in recent years as vendors have pushed data loss prevention technology and companies have faced down compliance regulations such as the Payment Card Industry Data Security Standard (PCI DSS). Though insider attacks are responsible for only a small portion of data breaches— just 4% of those covered in Verizon's analysis last year—malicious insiders can take advantage of the mix of poor monitoring, excessive privileges and poor account auditing practices to make off with critical data.

3. Steps to secure an organisation's data sources[1]

As the database attacks have increased, the number of SQL injection attacks has jumped by more than two thirds SQL injection attacks have been around for a long time, but according to the 2012 IBM Security Systems X-Force Report, they are still the most common type of attack on the Internet. Because of thesereasons, various steps were taken to impose regulations on cybercrimes. Besides,the hackers are becoming more efficient as they are able to build more sophisticated to tools to crack the wall. The rise of social media, cloud computing, mobility and big data are making threats harder to identify. Thus spies are able to pass confidential information from inside the organisation to an external person. Organizations need to adopt a more proactive and systematic approach to securing sensitive data and addressing compliance requirements amid the digital information explosion. This approach must be able to deal with the dispersed networks. We need to automate the discovery process in order to retrieve accurate and reliable data than manual analysis. In addition to exposing confidential information, SQL injection vulnerabilities allow attackers to embed other attacks inside the database that can then be used against visitors to the website.

Step 1: Identify the discovery rules and implement policies

Organizations are responsible to protect data even if the data is not revealed but only showing the relationships among databases. One needs to identify the sensitive assets of the databases and instances of discovery works by examining data values across multiple sources and determining their complex rules. The following figure shows how the discovery of security policies and classification is done.

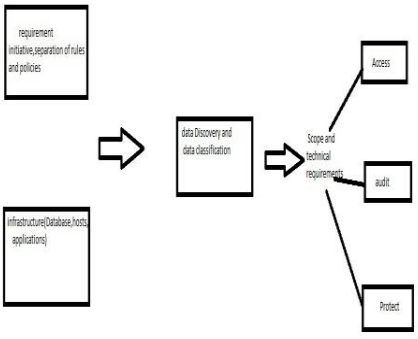

Fig 1: Use discovery solutions as a foundation for all security activities.

Step2: Finding vulnerabilities and loopholes

Organizations need to assess the configuration of databases,warehouses and Hadoop-based systems to ensure they don't havesecurity holes this comprises of verifying the way the DBMS is being installed and the security permissions granted to the sensitive tables. Alongside,organisations need to check the database versions with known vulnerabilities. Don't build your own checklists. There are several benchmarks from organizations such as they provide tests to check for common vulnerabilities including missing patches, weakpasswords, and misconfigured privileges and default accounts.

Step 3:Tightening the database

The result of a vulnerability assessment is often a set of specific recommendations to eliminate as much securityrisks as possible. Implementing these recommendations, such as setting a baseline for system configuration settings and locking user access to data, is the first step to hardening the database, warehouse or Hadoop-based system. Other elements of hardening involve removing all functions and options not in use. IT security professionals must be vigilant in establishing security policies, access controls and data usage policies to ensure that both security and business requirements of the organization are being met.

Step 4:Alert on change and audit

After discovering data sources, classifying the sensitive datatypes and hardening configurations, the organisations need to keep a check that data sources do not deviate from their secure configuration.Change tools and immediately alert that any updation has been made to the system policies. You must audit changes to data and the database executable. A database installation has numerous executable, and each one can be used by an attacker to compromise an environment. An attacker can replace one executable file with a version, that in addition to doing the regular work, also stores user names and passwords in a readable file. Changing audits would detect changes to the executable file and any other file created by the attacker.[1]

Step 5:Monitoring activities

Real-time monitoring of database, data warehouse orHadoop-based system activity is the key to limiting security risks.Activity monitoring across systems collects information fromdifferent sources for advanced analytics. Organizations can then create policies based on this security intelligence, such asalerting, masking or even halting malicious activity. Purposebuilt activity monitoring solutions offer a level of granular inspection of databases and repositories not found in any other tool. Use activity monitoring to provide immediatedetection of intrusions, misuse and unusual access patterns (which are characteristic of SQL injection attacks) unauthorized changes to financial data, elevation of account privileges, configuration changes executed via SQLcommands, and other malicious events. In addition to sending alerts, mature activity monitoring solutions can help prevent a disaster by taking corrective actions in real time, such as transaction blocking, user quarantine or data masking. Activity monitoring also helps with vulnerability assessment because it goes beyond traditional static assessments to include dynamic assessments of "behavioural vulnerabilities," such as multiple users sharing privileged credentials or an excessive number of

failed database logins.The given figure shows the procedure of monitoring activities.[1]

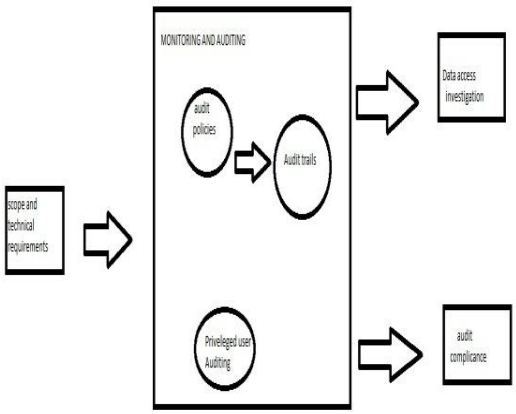

Fig: 2 Use case for database activity monitoring and auditing.

Step 6:Auditing and compliance reporting

Secure, non-reputable audit trails must be generated andmaintained for any data source activity that impacts securityposture, data integrity or viewing sensitive data. In addition tobeing a key compliance requirement, having granular audittrails is also important for forensic investigations. However, these approaches are often found to be lacking because of their complexity and high operational costs due to manual efforts. Other disadvantages include high performance overhead, lack of separation of duties (DBAs caneasily tamper with the contents of database logs, therebyaffecting non-repudiation) and the need to purchase and manage large amounts of storage capacity to handle massive amounts of unfiltered transaction information .Fortunately, next-generation activity monitoring solutions are available to provide granular, DBMS-independent auditing with minimal impact on performance, while reducing operational costs via automation, centralized cross-DBMS policies and audit repositories, filtering, and compression. Without the ability to quickly provide independent access security reporting, organizations will face significant costs and possible audit failure even with the best security controls. A data security solution should centralize reporting acrossdatabases, data warehouses, file shares and Hadoop-based systems with a customizable workflow automation solution to generate compliance reports on a scheduled basis. [1]

Step 7:Authentication and access control

Not all data and not all users are created equally. Regulatory mandates and security requirements are requiring organizations to adopt strong, multifactor authentication methods to protect against unauthorized and unidentified access. Organizations must authenticate users, ensure full accountability per user and manage privileges to limit access to data. Organizations should enforce these privileges, even for the most privileged database users. Periodic review of entitlement reports (also called User Right Attestation reports) as part of a formal audit process will result in better enterprise data security. [1]

Step 8: Data transformation

A key component of an information governance strategy, data transformation helps organizations address the challenges of information volume and variety with solutions for data protection and privacy—regardless of the type of data, the location or the usage. Use encryption, masking or redaction to render sensitive data unusable, so that an attacker cannot gain unauthorized access to data from outside the data repository or inadvertently reveal sensitive data. Data transformation techniques should protect data in transit, so that an attacker cannot eavesdrop at the networking layer (and gain access to the data when it is sent to the databaseclient), as well as protect data at rest, so that an attacker cannot extract the data (even with access to the media files). Employing the correct data transformation technique will ensure that both structured and unstructured data is protected, while still allowing the needed business data to be shared. Organizations must validate the flow of trusted information by applying the appropriate business rules and privacy procedures to manage data and continue to demonstrate and prove compliance to third-party auditors on an ongoing basis.[1]

4. CONCLUSION

Over the last decade, big data has become an important issue among organisations. It has become a path way to improve efficiencies, eradicate waste, increase productivity and make reliable, data driven decisions. Vast amount of data is being stored by organisations, so the main a motto is how do we store and effectively analyse this data. There is an issue concerning all organisations; that is finding the most vital items of intelligence in real-time, without browsing through huge amountsof data.

Many new big data tools are being developed to aidfirms to analyseresults drawn from theenormous information sources. These new technologies, such as predictive analytics and cloud computing, are helping big data analytics possible.

With the increase in sharing of information and data, it is important to ensure data security and privacy protection. Individuals' personal information like health, location, and online activity is scrutinised, which enhances issues about profiling, discrimination, and loss of control. The variousde-identification methods used by organizations such as encryption, key-coding enable analysis to proceed while increasing privacy concerns. Over the years, computer scientists have found that anonymised data can be re-linked to specific individuals.The conclusion for government and businesses can be sharply clear, if the de-identification becomes a key component of numerous business models, mostly

used for health data, online behavioural advertising, and cloud computing.

The principles such as data minimization and purpose limitation and individual control over information are based on privacy and data protection laws. Still it is an ambiguity, whether minimizing datacollection is always a pragmatic approach to privacy in the age of big data or not. The rules governing the privacy and data protection must be balanced with values such as public health, national security, environmental protection, and economic efficiency. A consistentstructure would be premised on a risk matrix, considering the value of different uses of information against therisks to individual autonomy and privacy. Where the advantages of potential data use clearly outweighs privacy threats, the justification of processing should be considered even if individuals reject. For example, web analytics—the measurement, collection, analysis, and reporting of internet data for purposes of understanding and optimizing web usage—creates rich value by ensuring that products and services can be improved to better serve consumers. Privacy threats are reduced, if analytics are put into effect, interactswith statistical data, in de-identified form.

The Policymakers should address the responsibility of consent in the privacy framework.Presently, many processing activities are based on individual consent. Still individuals are not-placed at correct position to make responsible decisions regarding their personal data given, although, well-documented cognitive prejudices, and on the other hand the increasing complexity of the information ecosystem. A privacy policy serves more as liability disclaimers for businesses rather than a positive declaration of privacy for consumers.

Simultaneously, action problems may produce a less substantial equilibrium where individuals fail to opt into societally beneficial data processing in the hope of free riding on the goodwill of their peers. Consider, for example, internet browser crash reports, which very few users opt into, not so much because of real privacy concerns but rather due to a (misplaced) belief that others will do so instead. This process is observed in other references where the difference between opt-in and opt-out process is not clear. In countries where organ donation is opt-in, donation rates tend to be minimal as compared to the rates in countries that are culturally similar but have an opt-out regime. Finally, a consent-based regulatory model tends to be less developed, since individuals' hopes are based on existing perceptions. Let's say, if Facebook had not launched its News Feed system in 2006 and had waited for users to opt-in,Facebook might not have been so beneficial to the users as we know it today. The users became used to the change after the information started flowing

The data regulators have increasingly denounced the era of big data as the growing ubiquity of data collection and the increasingly robust uses of information enabled by powerful processors and unlimited storage has been observed. Researchers, businesses, and entrepreneurs continueinnovation that may rely on the collection of large data sets. We want to initiate the development of a model where the advantages of data for businesses and researchers are balanced against individual privacy rights. Such a model would help us find whether processing can be justified based on legitimate business interest or only subject to individual consent.

5. REFERENCES

[1] Understanding the holistic database security, "IMW14277-USEN-01IBM Corporation Software Group Route 100 Somers, NY 10589 produced in the United States of America October 2012"

[2] Brian Prince, "How attackers find and exploit database vulnerabilities", Report ID: S4920512 InformationWeek, Reproduction Prohibited June 2012.

[3]Jeffrey Wheatman, "Database activities you should be monitoring," 14-March-2012ID: G00231531, GARTNER.

[4]Adrian Lane, "A guide to practical database monitoring", Report ID: S6281212 InformationWeek, Reproduction Prohibited December 2012.

[5] John Kindervag, Stephanie Balaouras, Brian W. Hill, and Kelley Mak, "Control and protect sensitive info in the era of big data". © 2012 Forrester Research, Inc. Reproduction Prohibited July 12, 2012.

[6] Gali Halevi, MLS, PhD Dr. Henk Moed , "The Evolution of Big Data as a Research and Scientific Topic Overview of the Literature" Research Trends Issue 30 September 2012.

[7] James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, Angela Hung Byers, "Big data: The next frontier for innovation, competition, and productivity" June 2011 © McKinsey & Company 2011

[8] Michael Cooper & Peter Mell, "NIST Information Technology Laboratory Computer Security Division, presentation".

[9] Big Data Strategy — Issues Paper, © Commonwealth of Australia 2013. Australian govt. Department of finance and deregulation

[10] Neil MacDonald, "Information Security Is Becoming a Big Data Analytics Problem" 8/22/12 Gartner