Data governance literature review

Nine H

Abstract

User privacy and the mishandling of user data has led to a crisis of data governance, where the historical wisdom on big data has been to hoard it and mine it intensely, new legislation threatens to disrupt internet technologies, while simultaneously industry figures, courts, and governments battle to shape the industry, technology, and future of the internet and web services.

Data Governance

Introduction

In this paper I will be making a thorough literature review of Data Governance, as it relates to the enterprise and with a focus on long term strategies, management of large data sets, and novel applications for data. Over the past five years, big data has become a reality, in that time technologies, strategies, and techniques have had time to fail, succeed, and become battle tested.

Section 1: Data Governance strategies and the impact of big data

According to Thom and Chronje while many organisations are seeking to collect and use big data, effective data governance is still elusive. The emergence of big data technologies has had the consequence that organisations of all sizes have begun capturing data as part of their day to day operations this has caused an exponential increase data production worldwide. Their paper defines the role of data governance as:

- Managing and mitigating risks posed by data storage and (mis)use.
- Management of people, processes, and technology to utilise data.
- Developing data standards, policies, and oversight.
- Managing oversight, ownership, and custody of data

And the mechanisms of governance as:

- Strategies
- Goals
- Policy
- Plans
- Standards

We can define the role, but this begs the question, how do we manage data? What strategies? What’s the best practice? Is there a framework we can apply? What are people doing in research and enterprise? Based on the literature these are difficult questions to answer.

Thompson et al (2015) examined the information assets of the WA public sector and from two case studies involving the state firearms registry, and emergency services came up with a number of negative examples of data governance and their causes. The report is a damning look at the failures of these departments but at a high level the issues were:

- Failing to maintain records accurately.
- Failing to validate recorded data.
- Failing to account for fraud and forgery.
- Allowing poor data integrity to hide critical information and impede departmental duties.
- Requiring manual processing of records for basic operations, putting an excessive burden on staff to process records.
- Failing to identify or record illegal access.
- Failure to differentiate users on the system or provide granular authentication/access privileges.

It’s important to note that the systems in question were conventional databases for which all of these issues have ready solutions. These were not technical failures, they were governance failures that were not mitigated by the backups mandated by common IT security policy (these were usually found to be functioning normally but backing up garbage data). The authors made three core recommendations.

1. Keep things simple
   Documentation, systems, and procedures should be as simple and understandable as possible for all users of the system. Complex procedures should be automated and validated by the system rather than a user.
2. Make compliance easier for end users than noncompliance
   Users that can override a system that they do not understand, malfunctions, or presents irrelevant errors that they can ignore and continue their workday most likely will.
3. Reinforce positive behaviour
   Positive reinforcement for good behaviour can remind employees that their actions do count and can encourage a culture that takes security and compliance seriously.

However, for many organisations simply implementing any form of data governance would be a good start.

Section 2: Enterprise applications of data

Big data enables a number of technologies and applications that are extremely data intensive. Chen (2014) identifies the opportunities afforded by big data to existing organisations:

- Improving efficiency
  Data can be used to identify waste within organisations and quantify the impact of small inefficiencies over a large scale organisation.
- Strategic planning
  Data can be used to inform and evaluate organisational strategies objectively.
- Improving service levels
  Finding the driving factors that contribute to missed SLAs, delays, bottlenecks.
- Identifying opportunities, developing products and services
  Data can help to identify gaps in products and services, and help organisations identify products and services to pivot towards or expand.
- Improving customer satisfaction
  Big data analytics can be applied to social data, community feedback to improve customer satisfaction.
- Identifying new markets
  Big data search can help find gaps in markets and opportunities for businesses to capitalise on.
- Faster turnaround to market
  Data can be used in product research and development to decrease the lead time on products and integrate lessons learned from successes and failures into future products much faster.
- Complying with regulations
  Big data processing can aggregate vast numbers of records and serve to automate compliance and reporting.

In addition to improving existing organisations a number of novel technologies have been developed in conjunction with and supported by big data datasets and technologies.

- Social network analysis
  The unfathomable amount of data generated by online social networks mandates big data analysis technologies in order to create actionable insights from large unstructured datasets.
- Machine learning
  Self organising computer networks that create models and novel solutions to navigation, search, and prediction problems consume vast datasets as their training data.
- Distributed compute
  Large datasets can exceed the processing capabilities of a single machine. Big data has necessitated the development of technologies to batch out the storage and processing of data over a network of machines working in tandem.
- Visualisation
  Huge collections require summarisation to become usable information, executive dashboards, infographics, and interactive models fed in real time allow big data to be understood and acted upon more rapidly.

Chen et al provided some examples of fields of study and problem sets that have been significantly improved by the application of big data.

- Navigation
- Social networks
- Finance
- Biomedicine
- Astronomy
- Intelligent transport
- IoT

Organisations of all sizes will increasingly need to not only manage essential data but develop knowledge and strategies around big data in order to compete and navigate the market and their fields.

Section 3: Big data technologies

There are a number of technologies for working with big data, I will be referring to the Journal of King Saud University's excellent review of technologies based on the apache Hadoop platform.

Big data processing has separated itself out into a layered model. Different and competing technologies can be deployed at different layers in a big data processing architecture to form a complete stack. From top to bottom these layers are:

- Management
  This is the highest layer and where system administrators perform management of systems and datasets.
- Analytics
  This layer is where applications that facilitate real time and passive monitoring of data flow, events, and performance indicators reside.
- Access
  This layer manages access to data and the big data infrastructure. This layer is where security and validation of applications and users takes place.
- Query
  This layer comprises of technologies that handle user queries, particularly scheduling, distribution, and batching of them. This is the intermediate layer that correlates and consolidates data that’s processed in the lower layers to be sent back up the stack.
- Processing
  This layer is for technologies that operate on chunks of data.
- Storage layer
  This layer manages storage and caching of data.

There are now several distributions of the hadoop platform for users to select from. As these distributions continue to mature and improve it would be advisable to select and incorporate one into a data governance policy according to the usual criteria for evaluating open source software rather than trying to construct one’s own software stack.

Section 4: Data integrity management

Big Data has a number of challenges Oussos et. al. define these as:

- Management
  Collecting, integrating, and storing big data. Evaluating hardware and software. Managing the scale of big data and optimising organisational capabilities against costs.
- Cleaning
  The removal of incorrect or misleading datapoints, the act of correcting bias in a dataset, and the act of shaping data to isolate useful representative samples.
- Aggregation
  It can be difficult to integrate heterogeneous independent datasets or integrate open source external datasets.
- Imbalanced systems capabilities
  Differing report frequency or accuracy from, for example, different IoT sensors sampling the same information may introduce a bias to the generated dataset.
- Imbalanced data
  Datasets that contain an unknown or unaccounted for bias can invalidate research and insights derived from them.

Each of these challenges are multidimensional and complex. A typical big data dataset is diverse and heterogeneous, meaning that it contains multiple data points measured at different frequencies from different devices and platforms. Even datasets collected by a large organisation such as a social media company are heterogeneous with respect to time by the action of models changing and the capability of the organisation to capture and store data increasing over time.

This is an example of an imbalance within data, but not the only one. Capturing large amounts of data in specific areas can introduce a bias to the data, as can the uneven distribution of or physical and configuration differences between collection devices. Having more data is not a panacea for bad data and without taking care to counter these imbalances one can expect normal garbage-in garbage-out operation like any other system. Organisations that wish to use data within their company need to have strategies and processes in place to review data collection, and review stored data to identify gaps and shortcomings in their datasets.

Section 5: Privacy and security concerns of data governance

In section one I introduced Trom and Cronje, in their review of governance they identified privacy, quality, and security as the key risks surrounding data governance. This has been compounded by IoT and smart medical sensors collecting extremely personal data that can potentially be used to de-anonymise data subjects back into individuals. This begs the question, how should an organisation mitigate these risks?

Bertino and Ferrari (2017) state that the security considerations of big data are much the same as any computer system: Confidentiality, Integrity, and Availability (CIA) but with the addition of Privacy. The mechanisms with which to address security concerns are:

- Identity validation and authorisation
  Management of keys, tokens, passwords, 2FA, or other technologies that validate identity.
- Authorisations management
  Managing identity, sessions, roles, responsibilities, and access. As part of management configuration and authorisation changes must be attributed and recorded.
- Platform controls
  As cloud data platforms[1] *become mainstream the role of data governance will be shared with them and they will assume some of the responsibility*[2] *to manage the use and abuse of their platforms, infrastructure, and applications by individuals and partners.*
- Cryptography
  The use of plaintext protocols and storage should be avoided. This dovetails into identity validation and authorisation. One must also manage cryptographic standards, migration, and monitoring of changing standards.
- Application security
  Ensuring the application is fit for purpose and well engineered so as to prevent data breaches.
- Network security
  Ensuring the network that handles data is correctly firewalled, domains are separated, topology is not visible to outside observers, and servers are free of defects and vulnerabilities.

They also identify a number of areas for further research:

- Access control of merging and integration
  Some datasets are harmless in isolation but may be joined to create harmful breaches of privacy.
- Proof of purpose
  As data platforms emerge it may become necessary to validate database queries against the stated aims of researchers to prevent abuse, this is a non-trivial problem.
- Personal and population privacy
  Personal privacy is intuitive, population privacy is the idea that information specific to a given population, that may be a cause of discrimination ought not be readily available.[3]

[1] Bluemix, BigQuery, etc.

[2] Platform neutrality for OSNs is being debated in public discourse, but this is an issue that's yet to be addressed for platforms and datasets outside of online publication.

[3] This idea is culturally and politically contentious and there are differing opinions on the rights of special interest groups over the rights of individuals.

- Integration of user preferences and privacy policy
  The EU recently passed user privacy regulations that govern online platforms. Mozilla Firefox can attach a header to all requests that indicates that the user does not want to be tracked by online services and networks.
- Relationship based access control
  Current privacy models on online social networks (OSNs) for user data are not granular enough to meet all the needs of users, access decisions need to be made taking into account interpersonal relationships.
- Risk modelling
  Methods, models, and frameworks to anticipate the impact a given data set could have in different contexts: open, closed, secure, leaked, etc.
- Privacy based data lifecycle models
  The conventional wisdom in big data is that all data is value and should be kept indefinitely. However public policy initiatives like the EU's right to be forgotten outline a need for guidelines regarding timely use of data and data disposal.
- Data ownership
  Legal rulings, torts, and statutes that clarify the legal owners and custodians of personal, private, and public data on individual data subjects.

Section 5: Data governance, new legal obligations

NB: this section deals with laws that are just coming in and may be subject to change. I'm am not a lawyer and this does not constitute legal advice. This section does not include citations of the same standard as the rest of this paper and for this reason was excluded from assessment. All information was sourced from DLAPiper's website.

One of the consequences of poor data governance, and lack of ethical guidelines has been the creation of laws that legislate limits around data collection, privacy, ethics, security, and the sharing of datasets. These laws were created in response to a number of controversies reported in the media, most notably the cambridge analytica scandal in which a partisan political lobbying group was given unprecedented access to a private user dataset from Facebook.

The two main international laws to examine are:

- EUGDPR - EU General Data Protection Regulation
  The EUGDPR implements rules around tracking, data collection, data storage, procedures to have user data expunged, and many non-eu tech companies are choosing to comply with these laws universally in order to serve european customers without handling the additional complexity of managing a separate EU website.
- CCPR - California Consumer Privacy Act
  This has yet to be finalised, but when completed will be the landmark US data protection legislation, the first one designed under the US constitution, and the one that will determine the compliance burden placed on the Silicon Valley tech giants: Facebook, Amazon, Apple, Netflix, Google etc.

The relevant Australian federal laws are:

- The Federal Privacy Act (1988)
  States that public sector entities and private sector entities with an annual turnover exceeding $3million are subject to the Privacy Commission and are also subject to civil penalties for privacy violations.
- Assistance and Access Act (2018)
  Government and law enforcement agencies gain the ability to issue various notices to communications providers to force them to provide data, persistent backdoors, and additional product features to intelligence agencies. It's unclear whether this is a legal obligation to organisations based outside of Australia but we can expect many to simply roll over when they see an intelligence letterhead.
- Consumer Data Right (TBD)
  This is a draft (not implemented) bill that would amend the Competition and Consumer Act to obligate companies to provide a mechanism for individuals and organisations to collect all data stored about them by a third party.

In addition each state has its own data protection legislation, these are:

- Information Privacy Act 2014 (Australian Capital Territory)
- Information Act 2002 (Northern Territory)
- Privacy and Personal Information Protection Act 1998 (New South Wales)
- Information Privacy Act 2009 (Queensland)
- Personal Information Protection Act 2004 (Tasmania)
- Privacy and Data Protection Act 2014 (Victoria)

Conclusion

In conclusion, effective data governance, while currently underutilised and under attained is far from a black art. In lieu of a common framework we can still enumerate the roles, responsibilities, and goals of effective data governance, identify the need to define the requirements for hardware, software, and compliance, but most importantly and fundamentally the need to take ownership and responsibility for data governance policy within organisations. Moving on from failures, controversies, and crises caused by poor governance in public and private organisations the coming decade will be the one where innovators and lawmakers will determine the future of data.

References

Trom, L., & Cronje, J. (2019). *Analysis of Data Governance Implications on Big Data. Perspectives on Asian Tourism*, 645–654.

Oussous, A. Benjelloun, F.Z., Lahcen, A., Belfkh, S. (2018). *Big Data Technologies: A survey. Journal of King Saud University - Computer and Information Sciences*. 431-448.

Bertino, E., & Ferrari, E. (2017). *Big Data Security and Privacy. A Comprehensive Guide Through the Italian Database Research Over the Last 25 Years,* 425–439.

Philip Chen, C. L., & Zhang, C.-Y. (2014). *Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. Information Sciences,* 275, 314–347.

Thompson, N., Ravindran, R., Nicosia, S. (2015). *Government data does not mean data governance: Lessons learned from a public sector application audit. Government Information Quarterly*. Vol 32, Issue 3, 316-322.