Previews of TDWI course books offer an opportunity to see the quality of our material and help you to select the courses that best fit your needs. The previews cannot be printed.

TDWI strives to provide course books that are content-rich and that serve as useful reference documents after a class has ended.

This preview shows selected pages that are representative of the entire course book; pages are not consecutive. The page numbers shown at the bottom of each page indicate their actual position in the course book. All table-of-contents pages are included to illustrate all of the topics covered by the course.
This page intentionally left blank.
TDWI Data Governance Innovations

Adapting for Agile, Big Data, and Cloud
To learn:

- The data governance challenges and opportunities that arise from cloud services
- The risks, challenges, and opportunities of big data governance
- The apparent conflicts between data governance and agile and how to overcome them
- The roles, relationships, and complexities of metadata management for data governance
- Data governance challenges that arise from mobile devices and social media
- The importance of ethics as a data governance imperative
- New models, practices, and processes for modern data governance

TDWI takes pride in the educational soundness and technical accuracy of all of our courses. Please send us your comments—we'd like to hear from you. Address your feedback to:

info@tdwi.org

Publication Date: July 2014

© Copyright 2014 by TDWI. All rights reserved. No part of this document may be reproduced in any form, or by any means, without written permission from TDWI.
<table>
<thead>
<tr>
<th>Module 1</th>
<th>The Current State of Data Governance ..........</th>
<th>1-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Module 2</td>
<td>Agile Data Governance ................................</td>
<td>2-1</td>
</tr>
<tr>
<td>Module 3</td>
<td>Big Data Governance .....................................</td>
<td>3-1</td>
</tr>
<tr>
<td>Module 4</td>
<td>Cloud Data Governance ...................................</td>
<td>4-1</td>
</tr>
<tr>
<td>Module 5</td>
<td>Metadata and Governance .................................</td>
<td>5-1</td>
</tr>
<tr>
<td>Module 6</td>
<td>Ethics, Mobile, Social, and Governance ..........</td>
<td>6-1</td>
</tr>
<tr>
<td>Module 7</td>
<td>Contemporary Governance Techniques .............</td>
<td>7-1</td>
</tr>
<tr>
<td>Module 8</td>
<td>Summary and Conclusion .................................</td>
<td>8-1</td>
</tr>
<tr>
<td>Appendix A</td>
<td>Bibliography and References ..........................</td>
<td>A-1</td>
</tr>
</tbody>
</table>
Module 1
The Current State of Data Governance

<table>
<thead>
<tr>
<th>Topic</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Current Definitions</td>
<td>1-2</td>
</tr>
<tr>
<td>Current Focus</td>
<td>1-4</td>
</tr>
<tr>
<td>Current Organizations</td>
<td>1-8</td>
</tr>
<tr>
<td>Current Processes</td>
<td>1-12</td>
</tr>
<tr>
<td>Why Data Governance Innovations?</td>
<td>1-16</td>
</tr>
<tr>
<td>Areas of Innovation</td>
<td>1-28</td>
</tr>
</tbody>
</table>
Current Definitions
Assets, Management, and Policies

Data governance is an emerging, cross functional management program that treats data as an enterprise asset: A collection of corporate policies, standards, processes, people and technology essential to managing critical data to a set of goals.

Maria Villar & Theresa Kushner

Data governance is the organization and implementation of policies, procedures, structure, roles, and responsibilities which outline and enforce rules of engagement, decision rights, and accountabilities for the effective management of information assets.

John Ladley & Danette McGilvray
Current Definitions
Assets, Management, and Policies

MANAGING A VALUABLE ASSET

Maria Villar and Theresa Kushner define data governance as a “management program that treats data as an enterprise asset: a collection of corporate policies, standards, processes, people, and technology …”¹

John Ladley and Danette McGilvray define data governance as “policies, procedures, structure, roles and responsibilities which outline and enforce rules of engagement, decision rights, and accountabilities for the effective management of information assets.”²

Both definitions are informative and thought provoking. Considering them together provides a good sense of the core of data governance as an asset management practice with attention to data-related policies.

ADDITIONAL PERSPECTIVE

No single standard definition exists, but several other information management practitioners define data governance in a variety of ways that add depth to overall understanding of the subject:

Gwen Thomas defines it as “execution and enforcement of authority over the management of data and data-related processes.”³

Alex Berson and Larry Dubov define it as “a process focused on managing the quality, consistency, usability, security, and availability of information.”⁴

David Loshin describes data governance as “a program for defining information policies that relate to the constraints of the business …”⁵

ASSETS & VALUE MANAGEMENT

Jonathan Geiger says, “data governance recognizes that data is an important enterprise asset and applies the same rigor to managing this asset it does for any other asset.”⁶ When considering asset management for financial or property assets it is generally true that value is a key consideration. It follows, then, that data value is key in data governance.

² Source: Executing Data Quality Projects by McGilvray. Ladley is a well-known EIM consultant.
³ Source: Data Governance Defined by McGilvray (www.enterprisedatajournal.com/article/data-governance-defined.html).
⁴ Source: Master Data Management and Customer Data Integration for a Global Enterprise by Berson and Dubov.
⁵ Source: Master Data Management by Loshin.
⁶ Source: Data Governance Defined by Ladley (www.enterprisedatajournal.com/article/data-governance-defined.html). Geiger is Executive Vice President of Intelligent Solutions and a well-known speaker and consultant.
Current Focus
Decision Rights

Authority to make decisions that is strongly connected to responsibility for data assets and accountability for outcomes.
Current Focus

Decision Rights

**PEOPLE GOVERN**

Data governance is done by people. Many people in various roles interact to govern data – or more accurately to govern the behaviors of other people when working with data. Individuals in specific roles have designated responsibility, authority, and accountability.

Responsibility is the obligation incurred by an individual in a specific role to perform the duties of that role. The individual is obligated to take actions and produce results that affect the organization’s assets.

Authority is the power granted to an individual in a specific role to make decisions and direct others to follow those decisions.

Accountability is the individual liability created by the use of authority. It is a condition of being fully answerable for results and achievement of goals.

**DECISION RIGHTS IN GOVERNANCE**

Decision rights are the means by which authority is implemented in most data governance programs. They typically assign decision rights by identifying the individual with authority to make decisions for each combination of decision type (data naming standards, data sensitivity, regulatory interpretation, to name a few) and data domain (customer, product, employee, etc.)

**THE DOWNSIDE OF DECISION RIGHTS**

While it is important to designate decision makers, the decision rights approach to governance is a burden when decision making is a slow process that inhibits fast-moving projects.
Current Organizations

Organization Structure
Current Organizations

Organization Structure

**HIERARCHICAL**
Top-down data governance is based in authority, command, and control. Authority to govern is delegated by an executive-level sponsor such as a chief data officer. When security, compliance, and legal risk are among the drivers – when policy definition and enforcement is important – the top-down governance may be necessary. In this governance model, executive decisions flow downward through hierarchical organizations.

**THE DOWNSIDE OF HIERARCHY**
A typical hierarchy has three levels – data officer, data owners, and data stewards – before reaching the level of business stakeholders. Those who work with data on a daily basis often feel far removed from the decision processes that affect their work. This culture may amplify the sense of a data bureaucracy and inhibit buy-in and participation.

**FUNCTIONAL ORIENTATION**
One effect of hierarchy is functional focus – more attention to “what we do” than to “what we accomplish.” The thought process is something like: owners decide, stewards facilitate, and custodians manage. While each of those generalizations is more or less true, there is little cohesion between decision, facilitation, and management that is directly traceable to goals such as data quality, regulatory compliance, security and privacy, etc.
Current Processes

Policy Making

- a few essential policies
- growth of data & complexity
- exceptions, regulations & variation
- overload, uncertainty & confusion
Current Processes
Policy Making

CREATING DATA POLICIES

Policymaking is a central process of policy- and enforcement-focused data governance. The key to sustainable policy-focused governance is managing the number of policies that are in force. It is common to create a policy in response to an immediate circumstance, then to leave the policy in place forever. As the number of policies grows, the probability of conflict among them grows, the frequency of requests for exceptions to policy increases, and the level of uncertainty and confusion expands. Having too many policies achieves nearly the same result as having no policies, but is more time consuming. Ideally, to avoid policy overload, strive for a small number of goal-oriented policies that are bounded with a defined expiration or renewal date.
Why Data Governance Innovations?
Driving Forces
Why Data Governance Innovations?

Driving Forces

CONTINUOUS EVOLUTION

Data governance programs must continuously evolve because the world of data experiences frequent change. The environment in which we create, consume, and manage data experiences change from many sources – business, data, systems, stakeholders, expectations, regulations, processes, practices, etc. A sustainable data governance program must evolve with the changes.

ABCs OF DATA GOVERNANCE

There will always be something new and emerging in the field of data and information management. Today’s hot topics within data governance include Agile, Big Data, and Cloud. Agile BIB processes and practices, big data opportunities, and cloud technologies and deployments all have near-term governance implications.

BEYOND THE ABCs

Agile, big data, and cloud are only today’s governance pressures. Data governance programs need to continuously innovate to keep pace with changes in how we manage data. Data visualization has yet to become a governance focus, but you can be sure that it will do so in the near future. And the future holds many other sources of change:

- What about mobile?
- What about virtualization?
- Desktop analytics?
- Social media?
- And what is emerging that is unique to your industry?

STEPPING UP TO METADATA

Not all needs for governance innovation are related to recently emerged and future trends. Metadata management is a long-standing problem that remains unsolved in most organizations. Innovating and adapting a data governance program is a good opportunity to make progress in resolving the metadata muddle.
Areas of Innovation
From Current State to Future State
Areas of Innovation
From Current State to Future State

THE DATA GOVERNANCE FUTURE

Today’s data governance programs are directed primarily at enterprise data governance – managing data as an asset with little consideration of the broader ecosystem in which data is collected, managed, and used. Today’s pressures on data management include agile projects, big data, and cloud technologies as well as the long-standing problem of metadata management. Attention to project pressures, to external data and new data types, and to emerging and evolving technologies is essential for sustainable governance programs. Resolving the metadata muddle will certainly help to sustain governance. Static governance programs can be assured that they will falter and fail over time. Adaptive and evolving data governance has the potential to deliver real and meaningful data and business value.
Module 2
Agile Data Governance

<table>
<thead>
<tr>
<th>Topic</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agile Development</td>
<td>2-2</td>
</tr>
<tr>
<td>Agile Teams</td>
<td>2-6</td>
</tr>
<tr>
<td>Agile Projects</td>
<td>2-8</td>
</tr>
<tr>
<td>Agile Culture</td>
<td>2-14</td>
</tr>
<tr>
<td>Agile Governance</td>
<td>2-18</td>
</tr>
</tbody>
</table>
Agile Development
Agile Process and Values

Individuals and interactions over processes and tools
Working software over comprehensive documentation
Customer collaboration over contract negotiation
Responding to change over following a plan
Agile Development
Agile Processes and Values

WHAT IS AGILE?
If you look in the dictionary you’ll find that agile means “moving quickly and lightly” or “mentally quick.” Synonyms include nimble and spry. Yet the term has quickly taken on particular meaning in systems and software development.

AGILE SYSTEMS DEVELOPMENT
The Scrum approach to agile development is based on the principle of discovering requirements through active collaboration with business subject experts. Agile projects are short in duration and fast moving with each project delivering a small set of functionality and business capability. Several projects are typically required to build a robust system. Projects are executed iteratively with multiple iterations until a release-ready product is built. Iterations are constituted of daily planning and direction for a continuous build process. When a release-ready product is completed, the release activity begins including planning, documentation and formalization, production implementation, etc. Releases and iterations achieve cohesion and continuity by working within a strategic structure of shared vision and goals.

THE AGILE MANIFESTO
Several of the pioneers and leading practitioners of agile development (too many to list here) collectively authored the following Manifesto for Agile Software Development. Following these principles will make us quick and nimble in systems development whether called “agile” or not.

“We are uncovering better ways of developing software by doing it and helping others do it. Through this work we have come to value:

- Individuals and interactions over processes and tools
- Working software over comprehensive documentation
- Customer collaboration over contract negotiation
- Responding to change over following a plan

That is, while there is value in the items on the right, we value the items on the left more.” (source: http://agilemanifesto.org)

GOVERNANCE: FRIEND OR FOE?
It is a common belief that agile and governance are in conflict – that the processes of governance are too burdensome and become barriers to agile development. In reality, agile and governance can work together quite effectively but it requires a change in mindset. We need to think differently about agile teams and about how we govern – shifting the work of governance from external to part of the agile process, and focusing governance on value produced instead of processes followed.
Agile Teams

Two Levels

Agile Planning and Oversight

- Agile Planning and Oversight
- Strategy
- Release
- Iteration
- Daily
- Continuous

Legal, regulatory, and risk stakeholders

Financial and operational stakeholders

Sponsors, functional stakeholders, business SMEs

Agile Development

End-users, developers, team lead

Constraints, expectations, requirements
Agile Teams
Two Levels

**TEAM AND PROCESS ALIGNMENT**

The agile process has planning activities (strategy and release) and development activities (iteration and continuous building). It makes sense to organize agile teams to align with the agile process – organizing into a planning and oversight team and a separate development team.

**PLANNING AND OVERSIGHT**

The planning and oversight team is made up of sponsors, stakeholders, and business subject matter experts. This team has responsibility for vision, goals, funding, and project chartering. Their work is focused on strategy and release activities and they participate in iteration activities as a bridge from planning to development.

**DEVELOPMENT**

The development team is a collaboration of end users and developers working under the guidance of a team lead. This team is responsible to discover requirements, design solutions, and build systems. Their work is focused on continuous development, daily planning, and iteration as a bridge to planning and oversight activities.

**GOVERNANCE**

In this team structure, governance begins with the planning and development team. These stakeholders – functional, legal, regulatory, risk, financial, and operational – are positioned to understand governance needs and to express those needs as constraints, expectations, and requirements for the development team. The development team is responsible to include governance needs in the development process, and communicate with stakeholders to fully understand governance goals and constraints.
Agile Projects

Agile Projects and Governance Goals

requirements of legal, regulatory, and risk stakeholders

development team ... build in quality is an agile principle

requirements of legal, regulatory, and risk stakeholders

define strategy, plan at iteration, build to integrate

"good enough" in development, complete at release

requirements of legal, regulatory, risk & functional stakeholders

expectations of functional, operational & financial stakeholders ...

plus agile is value oriented by “building the right things”
Agile Projects

Agile Projects and Governance Goals

**POINTS OF FOCUS FOR GOVERNANCE**

Data governance has many different goal areas, each with a slightly different focus in the agile development process:

- Compliance, security, and privacy goals are driven by requirements of legal, regulatory, and risk stakeholders. Focus on these goals at strategy to express requirements and constraints and at release to confirm that the requirements are met.

- Quality goals are primarily a development team focus of continuous build, daily planning, and iteration activities. This aligns neatly with the agile principle to “build in quality.”

- Data integration goals have focal points throughout the agile process to strategically define integration goals, plan for integration as an iteration activity, and build to integrate as part of daily planning and continuous build activities.

- Metadata goals get a light touch in continuous build, collecting only the metadata that is necessary to support the build processes. The work of bringing metadata to enterprise standard is primarily a release activity.

- Data retention and disposal goals are driven by requirements of legal, regulatory, risk, and functional stakeholders. Focus on these goals at strategy to express requirements and constraints and at release to confirm that the requirements are met.

- Value and project risk goals are driven by expectations of functional, operational, and financial stakeholders as part of strategy. The agile development process also focuses on value and risk, minimizing risk with continuous end-user participation and achieving value by adapting to changing requirements to build the right things.
Agile Culture
Agile Culture and Governance Roles
Agile Culture

Agile Culture and Governance Roles

**PEOPLE AND COLLABORATION**

It is often said that agile development is more culture than process. Collaboration is fundamental to agile development, and inclusion is a key element of collaboration. From the perspective of agile data governance, this means including those who have governance roles in agile teams:

- Data owners as part of the agile planning and oversight team with participation in strategy and release activities.
- Data specialists and custodians as part of the agile development team participating in daily planning and continuous build activities.
- Data stewards in the critical role of connecting strategy with development by participating at each iteration cycle.
Agile Governance

Governering with Agility

- focus on value produced, not methodology and processes
- think “governance as a service” instead of “authority and control”
- strive for policy adoption over policy enforcement
- govern proactively ... introduce constraints as project requirements
- be quick and responsive, never a bottleneck
- no project delay from unresolved issues ... remediate first then resolve
- include data stewards on project teams
- write brief, concise, clear, and understandable policies
- agile data definition ... get definitions from user stories
- brief definitions ... what it is, what it isn’t, and an example or two
- evolve data models ... not “get it right the first time”
- jump start data models with data model patterns and universal models
- emphasize data model & database refactoring as core skills
- govern collaboratively and automate with collaboration tools
Agile Governance
Governing with Agility

GOVERNANCE TO ENABLE AGILE PROJECTS

Governing with agility requires a change in mindset from traditional governance structures and practices to those that enable and do not inhibit agile projects. Traditionally, governance is organized and operated as an external entity that exercises controls over projects and the results that they produce. Agile data governance positions governance as part of the projects (internal, not external) participating and collaborating in ways that help the project to succeed while simultaneously accomplishing the goals and purpose of data governance. Some of the practices that help to make the shift to agile governance include:

- Focus on value produced, not methodology and processes. This includes value to the project and enterprise value produced by meeting governance goals.
- Govern proactively by introducing constraints as requirements at the beginning of a project instead of seeking remedial action at the end. Attempting to introduce constraints in the midst of an active agile project is even more disruptive than end-point remediation.
- Strive for policy adoption over policy enforcement. Make it easy to comply with policies, communicate the reasons for and value that is created by the policies. Motivate people to adopt and little enforcement will be necessary.
- Write brief, concise, clear, and understandable policies. Use simple language that is not ambiguous or subject to interpretation.
- Include data stewards on project teams. They bring valuable knowledge and are generally great collaborators.
- Think “governance as a service” instead of “authority and control.”
- Practice agile data definition. Get definitions from user stories and write brief definitions that describe what it is, state what it isn’t, and provide an example or two.
- Be quick and responsive. Govern at the speed of agile.
- Don’t let unresolved issues become project delays. Think issue remediation first, with just enough to let the project continue. Then pursue issue resolution without disrupting projects.
- Evolve data models. Don’t expect them to be right the first time. Jump start with data model patterns, and adapt to change by developing data model and database refactoring as core skills.
- Govern collaboratively and automate with collaboration tools that enable working together without excessive meetings and bureaucracy.
Module 3
Big Data Governance

<table>
<thead>
<tr>
<th>Topic</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Big Data Basics</td>
<td>3-2</td>
</tr>
<tr>
<td>Big Data and Governance Goals</td>
<td>3-10</td>
</tr>
</tbody>
</table>
Big Data Basics

Big Data Defined

Data sets with sizes beyond ability of commonly-used software tools to capture, integrate, manage, and process within a reasonable amount of time.

Massive volumes of both structured and unstructured data that are so large that it’s difficult to process with traditional database and software techniques.

Data sets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze.

(The McKinsey Global Institute, 2011)
Big Data Basics

Big Data Defined

DESCRIPTION

The term “big data” has become popular to describe rapid growth in the volume, variety, and velocity of data that is now available in business – unstructured data, semi-structured data, social media data, location data, radio frequency data, and more. These types of data tend to yield data sets that are too large, complex, or unwieldy to work with traditional data management and analytics technologies.

BIG DATA TYPES AND TECHNOLOGY

Structured big data can be readily stored in tabular forms. It is typically used for analysis and is stored and manipulated using columnar databases and high speed analytic and data warehouse appliances based on multiple parallel processing (MPP) hardware architectures.

Unstructured and semi-structured big data includes formats such as text, social media content, multi-media content, and Web logs. The technologies to work with unstructured big data are emerging and evolving. Today’s most commonly used technologies include Hadoop, MapReduce, and NoSQL.

Hadoop is distributed file system designed to work with large data collections that combine structured with more complex types of data. MapReduce is a programming framework to write applications that work with Hadoop datasets.

NoSQL describes a class of database management systems whose common characteristics are that they are not based on a relational model and they do not use SQL as a programming language. NoSQL databases are optimized for distributed storage and retrieval of very large volumes of data, either structured or unstructured. NoSQL supports append and read functions. Insert, delete, update, and join are not typically supported.
Big Data and Governance Goals
Security and Privacy

- Identify sensitive big data
- Collect sensitivity metadata
- Evaluate privacy regulations by locale
- Identify data that crosses geographic boundaries
- Define internal big data security/privacy policies
- Monitor access to and use of big data
Big Data and Governance Goals

Security and Privacy

**BIG DATA VULNERABILITIES**

The origins of big data technologies led to some security and privacy exposures that are inherent in today’s big data solutions. Many of the solutions were not designed for the common uses that they see today. MapReduce, for example, was created by Google to store public links and make them ready for efficient searching. Security was not a concern because the data was public links. Today MapReduce and NoSQL solutions are used with medical records, financial records, credit and lending data, and much more that is privacy-sensitive.

**IDENTITY AND PRIVACY**

Much of the commonly used big data such as purchasing patterns and mobile device tracking removes personally identifying information (PII). Removing PII, however, is not sufficient to ensure privacy. Computer science research has demonstrated that it is possible, given enough data, to reconstitute identity with a high degree of accuracy through inference from patterns and relationships in the data.

**PRIVACY PRACTICES FOR BIG DATA**

Sunil Soares itemizes these best practices for privacy protection of big data.

- Identify sensitive big data
- Flag sensitive big data in the metadata repository
- Address privacy laws by country, state, and province
- Manage situations where personal data crosses international boundaries
- Monitor access to sensitive big data by privileged users

---

# Module 4

Cloud Data Governance

<table>
<thead>
<tr>
<th>Topic</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cloud Data Basics</td>
<td>4-2</td>
</tr>
<tr>
<td>Cloud Data and Governance Goals</td>
<td>4-6</td>
</tr>
<tr>
<td>Migrating to the Cloud</td>
<td>4-18</td>
</tr>
</tbody>
</table>
Cloud Data Basics
Data Options in the Cloud

private cloud
hybrid cloud
public cloud

DaaS, SaaS, PaaS, IaaS
Cloud Data Basics
Data Options in the Cloud

CLOUD SERVICES

Cloud computing is the widely used term for virtualization of computing services. Cloud technology fits into several categories:

<table>
<thead>
<tr>
<th>Service Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Software as a Service (SaaS)</td>
<td>A Web-based model for deployment of applications, SaaS allows on-demand use of applications without the need to license and install for every computer the software is used on.</td>
</tr>
<tr>
<td>Data as a Service (DaaS)</td>
<td>DaaS hosts data in the cloud. Data marketplace services make data available on a pay-per-use basis. The developer data hub provides Web services for developers to upload data, then build Web services to work with data. Store and synchronize provides services such as those of Dropbox and similar products.</td>
</tr>
<tr>
<td>Platform as a Service (PaaS)</td>
<td>PaaS is an approach to virtualizing the hardware, operating systems, applications frameworks, and technology stacks upon which applications are built and deployed.</td>
</tr>
<tr>
<td>Infrastructure as a Service (IaaS)</td>
<td>IaaS is an architecture of virtualized hardware and operating systems. On the surface it sounds a lot like another name for PaaS, but there are some subtle differences. PaaS provides a developer environment as well as operations environment but is limited to web applications. IaaS delivers only the operations environment but supports a broader range of applications.</td>
</tr>
</tbody>
</table>

PUBLIC, PRIVATE, AND HYBRID CLOUD

All variations of cloud services are characterized by multi-tenant architecture where many customers use a single instance of servers and software. Multi-tenancy is characteristic of public cloud services. Governance concerns such as security may be addressed with private and hybrid cloud options.

<table>
<thead>
<tr>
<th>Cloud Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Private Cloud</td>
<td>Private cloud describes a virtualized environment that is used exclusively by a single tenant. The private cloud may be implemented internally behind a corporate firewall, or it may be an isolated, single tenant environment hosted by a cloud service provider.</td>
</tr>
<tr>
<td>Hybrid Cloud</td>
<td>Hybrid cloud uses a combination of public and private options to gain efficiencies without compromising security and privacy.</td>
</tr>
</tbody>
</table>
Cloud Data and Governance Goals

Data Security and Privacy in the Cloud

- Access Control
- Storage Encryption
- Transport Encryption
- Firewalls
- Hardening of Servers
- Physical Security
Cloud Data and Governance Goals
Data Security and Privacy in the Cloud

**SECURITY AND PRIVACY**

Moving data to the cloud means less local and physical control. It is a common misconception that cloud-hosted data services raise the risk of data misuse and security breaches. Cloud services don’t necessarily increase the risks; they simply change them. When risks change, risk mitigation techniques must also change, with the service provider assuming some responsibilities. Minimize risk of data misuse, compromised privacy, and security breaches by seeking a provider who offers:

- Access control lists to define the permissions attached to the data objects.
- Storage encryption to protect against unauthorized access at the provider’s data center.
- Transport level encryption to protect data from intrusion when transmitting across networks.
- Firewalls that include Web application firewalls for protection from outside attacks launched against the provider’s servers.
- Hardening of the servers to protect against vulnerabilities in the operating system and software. Hardening is the process of securing a system by reducing its surface of vulnerability. A system has a larger vulnerability surface the more functions it fulfills; in principle a single-function system is more secure than a multipurpose one.
- Reducing available vectors of attack through removal of unnecessary software, unnecessary usernames or logins, and disabling or removing unnecessary services.
- Physically securing servers and networks to protect against unauthorized physical access to data.
Migrating to the Cloud
Data Governance Roles

- Role in Provider Selection
- Role in Service Level Agreements
- Policies Governing Cloud-Suitable Data
- Pre- and Post-Migration Data Audits
Migrating to the Cloud

Data Governance Roles

**PROVIDER SELECTION**

The processes of migrating data from on-premises to cloud-hosted raises data governance questions. What roles, responsibilities, and participation should data governance have in choosing cloud service providers and negotiating service agreements? This is a subjective question for which there is no definitive answer. The right answer depends on many variables including data governance maturity, organizational structure and culture, and working relationships among owners, stewards, and custodians of data. Though there is no single prescriptive answer, it is clear that cloud services agreements must not conflict with internal data governance policies, processes, responsibilities, and accountabilities.

**POLICIES**

It is unlikely that all data will migrate to the cloud; not all data is suited to cloud hosting. The data governance organization needs to develop policies about cloud-suited data and data that is explicitly prohibited from moving to cloud.

**AUDIT**

Do not assume that cloud migration is inherently error-free. The data governance organization should oversee the planning of appropriate pre-migration, post-migration, and ongoing data audits. The audit function may be expanded to include data profiling and data quality assessment.
Module 5
Metadata and Governance

<table>
<thead>
<tr>
<th>Topic</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metadata Basics</td>
<td>5-2</td>
</tr>
<tr>
<td>Metadata Types</td>
<td>5-12</td>
</tr>
<tr>
<td>Metadata Processes</td>
<td>5-22</td>
</tr>
<tr>
<td>Metadata Challenges</td>
<td>5-24</td>
</tr>
<tr>
<td>Metadata Governance</td>
<td>5-40</td>
</tr>
</tbody>
</table>
Metadata Basics
Metadata Defined

**Metadata:** Data about data.

**Metadata:** The data and information that is needed by an organization to effectively and efficiently manage its data and information resources.

<table>
<thead>
<tr>
<th>Dewey Decimal System</th>
</tr>
</thead>
<tbody>
<tr>
<td>000</td>
</tr>
<tr>
<td>100</td>
</tr>
<tr>
<td>200</td>
</tr>
<tr>
<td>300</td>
</tr>
<tr>
<td>400</td>
</tr>
<tr>
<td>500</td>
</tr>
<tr>
<td>600</td>
</tr>
<tr>
<td>700</td>
</tr>
<tr>
<td>800</td>
</tr>
<tr>
<td>900</td>
</tr>
</tbody>
</table>
Metadata Basics
Metadata Defined

The common definition of metadata is simply “data about data.” Although technically correct, it is too simplistic and not very informative. Expanding the definition helps to better understand what metadata is all about. Metadata encompasses all the data and information that is needed by an organization to effectively and efficiently manage its data and information resources. This includes: data to classify information (by topic for example), data to describe information (for example attributes about content and structure), data to guide users of information (to find and to understand), and data to control information (distribution, security, privacy, etc.)

Although rarely called metadata, the Dewey Decimal System is perhaps the oldest and certainly the most widely used metadata system in the world. Invented by Melvil Dewey in 1876, this system is used by libraries in more than 135 countries. The decimal system classifies books and other bibliographic items by content. The items are indexed in a card catalog (or today, the computerized version of a card catalog) where they can be searched by author, topic, and other attributes – then located on shelves where they are physically organized according to Dewey’s system. The books on shelves are information resources. The catalog and Dewey Decimal System are metadata. This fundamental concept of metadata as a key to managing information has been established for well over a century.
Metadata Types

The Scope of Metadata

- **Business Metadata**
  - Data about the business meaning, management, users, and uses of data.

- **Design Metadata**
  - Data about the design of systems & databases that collect, manage, & use data.

- **Implementation Metadata**
  - Data about the design of systems & databases that collect, manage, & use data.

- **Operations Metadata**
  - Data about schedules, events, & execution of systems that collect, manage, & use data.
Metadata Types
The Scope of Metadata

**KINDS OF METADATA**

Metadata can be classified and grouped in many ways. A common and simple classification uses two categories – business metadata and technical metadata. A more effective classification model to understand the broad scope of metadata subdivides into four categories: business metadata, design metadata, implementation metadata, and operations metadata. Each category is described in greater detail, and is placed into governance context on the following pages.
Metadata Processes
Metadata Management Activities

- Collect
- Document
- Structure
- Store
- Link
- Index
Metadata Processes

Metadata Management Activities

**Metadata Processes**

Metadata management involves several activities to gather and organize metadata and to make it easy to find and access:

- Metadata collection is the work of gathering needed metadata. This collection is usually not a stand-alone activity. It is a task done while performing other activities such as requirements gathering and system design. Various tools used for requirements analysis, systems design, development, and operations will each collect metadata.

- Metadata documentation records the metadata in forms and formats that are appropriate to the types of metadata collected. Standards, guidelines, and templates can improve consistency and quality of metadata documentation. Tools may also impose some standards for documenting metadata.

- Structuring metadata organizes it for ease of access and navigation with keys and relationships that describe the interrelationships among various metadata records. Metadata structures are typically pre-defined by the structures of one or more metadata repositories.

- Storage collects metadata records in databases or repositories. Many tools have proprietary metadata repositories, creating difficulties in achieving integrated metadata.

- Metadata linkage captures associations among metadata components to aid access, navigation, and even light integration. Linkage is typically performed in one of two ways: (1) by an association table that captures the relationships of metadata components stored in multiple repositories, or (2) by linking metadata components in multiple repositories through a shared logical and syntactic structure.

- Indexing metadata makes it searchable by associating metadata components with keywords or other search criteria.
Metadata Challenges
Metadata Management Issues

Determining Metadata Requirements
Collecting and Recording Metadata
Integrating Metadata
Maintaining and Updating Metadata
Providing Metadata Access
Metadata Challenges

Metadata Management Issues

**Metadata Requirements**

Understanding metadata requirements is fundamental to formulating a metadata strategy. Requirements need to consider (1) why metadata is needed, (2) how much and which metadata is needed, and (3) how much metadata integration is necessary.

The successful metadata program identifies essential metadata based upon business needs, selects a manageable initial scope (and cost) with clearly identified business benefits, and then incrementally rolls out metadata capabilities and achieves associated benefits.

**Metadata Collection**

Gathering all metadata is difficult. Identifying the sources and capture points is challenging enough. Once they are known, the metadata collection tasks may be labor-intensive. These tasks include the following:

- Collect and store technical metadata about warehousing data
- Build process metadata capture capabilities into ET
- Author business metadata about warehousing data
- Find and/or create business and technical metadata about sources

**Metadata Integration**

Multiple tools capture different and sometimes overlapping metadata. Integration involves identifying all the tools, applications, and processes that capture metadata, knowing what metadata they capture, and determining how (in a sustainable and cost effective manner) to realize integration across those components.

**Metadata Maintenance**

Keeping metadata complete and current is a continuous and resource intensive effort. The volume of metadata will increase as business users discover new ways to use information and new business questions are asked. Careful attention to standard processes and effective application of available tools will help keep the job manageable.

**Metadata Access**

Metadata access for the business is ideally integrated into the analytic tools being used. Metadata provides context, and is most effective in doing so when metadata access is seamless. Providing access for all metadata users – business and technical – is challenged when metadata storage is not integrated but an integrated access facility is desirable.
Metadata Governance

Governance for Metadata
# Metadata Governance

## Governance for Metadata

### Metadata Needs Governance

Data governance is more than just governing for business data. At best it also involves governing for metadata, which is as important as business data. It is, in fact, business data that is essential and needed to manage business information assets. The metadata scope of governance includes standards and practices, data naming and definition, metadata service levels, and metadata oversight.

### Standards and Practices

Standards and practices for metadata are many and varied. We’ve looked at several that are related to agile development, big data, and cloud data. In addition to more common standards and practices, these recent trends in data management push the boundaries and create needs for new standards and practices. The data governance challenge is in establishing, communicating, and fostering adoption of new standards and practices.

### Naming and Definition

As is true with any data, metadata is more useful and valuable when carefully named and meaningfully defined. Metadata naming and definition practices, however, are often more casual and lack the discipline that we apply to other business data. Data governance can change this reality by establishing standards for metadata names and definitions, assisting with naming, and facilitating definition activities.

### Service Levels

Service levels for metadata quality, completeness, availability, and usability are particularly important. Perhaps more than any other collection of data (finance, customer, HR, etc.), data about data affects every business function throughout an enterprise. And the impact and importance of metadata is growing as we expand the scope of data that we use beyond enterprise boundaries. Metadata governance can improve service levels with goal setting, measurement, and monitoring.

### Oversight

Data governance may also assume responsibility for oversight of metadata assets including repositories, registries, and portals. All other data collections in an enterprise have (or should have) designated data owners. Metadata, however, tends to be fragmented data where we assume that ownership is connected with that for the data it describes. For cohesive metadata management consider assigning ownership to a data governance council with data stewards as primary stakeholders.
Module 6

Ethics, Mobile, Social and Governance

<table>
<thead>
<tr>
<th>Topic</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ethics and Data Governance</td>
<td>6-2</td>
</tr>
<tr>
<td>Mobile Technology and Data Governance</td>
<td>6-10</td>
</tr>
<tr>
<td>Social Media and Data Governance</td>
<td>6-14</td>
</tr>
</tbody>
</table>
Ethics and Data Governance

Why Ethics Matters

How we collect data

How we use information

How we guide conduct

Good / Right

Others

Self

Bad / Wrong
Ethics and Data Governance

Why Ethics Matters

WHAT IS ETHICS
Ethics is the challenge of distinguishing right from wrong and good from bad. It is challenging because “right” and “good” are not always clear. Furthermore, right and wrong must account for all stakeholders. It is common to think of ethics as doing right for others, but when doing right for others is harmful to the self, ethics is not served. Nor is ethics served when doing right for oneself harms others. Ethics has two aspects – the tension of right vs. wrong and the tension of right vs. right.

ETHICS, BI, AND ANALYTICS
Business intelligence raises many new ethical questions about how we collect and use intelligence, and in how those choices guide and shape organizational and individual conduct. Every decision that we make and every action that we take shapes the perceptions of customers, employees, partners, competitors, and the public.

We are approaching an era where every BI program will need to actively manage ethics. More data, more kinds of data, and advanced analysis of data often conflict with concerns of data privacy, security, anonymity, and ownership. Ethics is especially important in the area of predictive analytics and predictions of individual behaviors, and the need is amplified with the abundance of personal data available from big data sources.
Mobile Technology and Data Governance

Governance Risks

Security ... Privacy ... Confidentiality ... Recoverability

Viruses ... Malware ... Hacking
Mobile Technology and Data Governance

Governance Risks

**MOBILE DATA SECURITY AND PRIVACY**

Data on mobile devices such as smartphones and tablets brings new data governance challenges. Sending data, receiving data, and storing data with mobile devices creates risks related to security, privacy, and confidentiality. Data stored only on a mobile device carries high risk of loss without the ability to recover that data. Beyond the risks of data on mobile devices, you also need to consider the risks of data in motion. Sending data across networks – especially open wireless networks and unsecured cellular networks – poses new security and privacy risks.

**MOBILE DATA MISUSE AND ABUSE**

Mobile devices also increase risk of data breaches because they are vulnerable to viruses, malware, and hacking. This is especially true when personal devices (“bring your own device”) that are not enterprise managed are used to store and transmit business data.

Mobile apps also create opportunity for data misuse and abuse. For internally developed apps for mobile workers as well as consumer apps that are generally available, data governance should consider the security, privacy, and ethical implications of data that is collected by mobile apps.
Social Media and Data Governance

Governance Risks

Security ... Privacy ... Confidentiality
Quality ... Integration ... Usability
Social Media and Data Governance

Governance Risks

**SOCIAL MEDIA, PRIVACY, AND CONFIDENTIALITY**
Social media brings risks for privacy and confidentiality similar to those of mobile. The security concerns are not as significant as with mobile because we typically don’t send, receive, or store data with social media (although there are exceptions). Privacy and confidentiality concerns are related more to attention to appropriate social media postings that do not compromise privacy or confidentiality. Although policies that address social media postings are not common in data governance programs, it is easy to make a case for creating them.

**SOCIAL MEDIA AS DATA SOURCES**
Social media data used as data sources raises questions about data quality, data integration, and usability. These topics were discussed in some depth earlier in the course and will only be briefly revisited here. As social media data becomes part of the information ecosystem, it is imperative that data governance address its unique risk factors.

**SOCIAL MEDIA ANALYTICS**
Inappropriate use is arguably the biggest risk of social media analytics. The world of social media is rich with data that identifies individuals and exposes many facts about them and their behaviors. The abundance of data creates opportunity for analysis that may cross ethical boundaries. Data governance can help to mitigate this risk, but must step up to the subject of ethics to do so.
Module 7

Contemporary Governance Techniques

<table>
<thead>
<tr>
<th>Topic</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Governance Zones</td>
<td>7-2</td>
</tr>
<tr>
<td>Flat Governance Organizations</td>
<td>7-4</td>
</tr>
<tr>
<td>BI for Governance</td>
<td>7-8</td>
</tr>
<tr>
<td>Technology for Governance</td>
<td>7-12</td>
</tr>
</tbody>
</table>
Data Governance Zones
Data Variety and Pragmatic Governance

One size fits all is old data governance. Data governance zones are today’s manifesto.

– Michele Goetz, Forrester Analyst
Data Governance Zones
Data Variety and Pragmatic Governance

BREAKING BIG CHALLENGES INTO LITTLE ONES

Michelle Goetz, a Forrester analyst, blogs that data governance is not “one size fits all” and suggests data governance zones as an appropriate way to fit governance solutions to specific needs. Governance zones may also be an effective way to take the very big challenge of data governance and subdivide it into a related collection of manageable challenges.

Data governance zones are somewhat analogous to the nation, state and province, and local subdivisions of geopolitical governance. Some of the ways that you might consider zoning data governance include:

- Master data vs. local data
- Enterprise vs. end user and spreadsheet
- Internal vs. externally hosted
- Traditional vs. big data
- Operational/transactional vs. decision-support and analytics
- Zoned by business function
- Zoned by compliance sensitivity
- Zoned by security sensitivity: personally identifiable, private, confidential, internal use, general use
- Zoned by affinity with highly integrated and shared clusters constituting zones

This sampling illustrates some of the possibilities in a new way to think about governance. Mix and match these and introduce zones not described here to get the best fit for your organization. Combine zoned governance with enterprise governance if that is what makes best sense. Consider the possibility of enterprise governance as a federation of zones. Think about the potential of lead data stewards designated by zone.

The concept of data governance zones is relatively new with only limited information. But it is a worthy idea that may provide the seeds of innovation for your data governance program.
Flat Governance Organizations
Simplify for Agility
Flat Governance Organizations
Simplify for Agility

RETHINKING THE GOVERNANCE ORGANIZATION

A simpler data governance organization is one of the keys to agile data governance. Can you create an organization with fewer layers of hierarchy, fewer committees, and minimal meetings? Do you minimize meetings by having fewer, shorter duration meetings, with fewer participants, or limited agendas?

Is it practical to simplify the governance organization and simultaneously maximize communication? Depending on culture, a simpler organization may be a path to more effective communication. Consider using mobile technologies and social media as communication methods – a means to engage many people while keeping the formal organization small and light.

These questions and thoughts are intended as stimuli that may seed innovation of your data governance organization.
BI for Governance
Applying BI Concepts for Data Governance

**BI and Business Performance Management**

Business Processes

- measure
- compare
- report
- analyze

Business Goals

adjust

**BI and Data Governance Management**

Data Governance Processes

- measure
- compare
- report
- analyze

Data Governance Goals

adjust
BI for Governance
Applying BI Concepts for Data Governance

**PERFORMANCE MANAGEMENT**

Data governance is a business function that is performed through business processes. The principles of business performance management – goal setting, measurement, monitoring, and adjustment – apply as readily to data governance as for any other business function. Data governance performance management *is* business performance management.
Technology for Governance
Opportunities and Applications
Technology for Governance
Opportunities and Applications

TOOLS FOR GOVERNANCE

Data governance, just like any other important business function, needs technology to be successful. The market offers a handful of software tool suites for data governance, but a tool suite isn’t always necessary. It may be most practical to select standalone tools for areas of greatest need. The common areas of tool support for data governance include:

- Policy management
- Workflow management
- Metadata management
- Sharing and collaboration
- Business rules management and repositories
- Data discovery and data profiling
- Data validation
- Data quality management