People and Processes

Linus Rundberg Streuli

Table of contents

1 Introduction

- The tools used when building a data governance strategy are important, but not enough by themselves.
- The people involved, and the process of implementing data governance are just as important.

2 People

- A well functioning data governance program is dependent on a complex interplay of roles and responsibilities.
- Many organizations struggle to match the roles and responsibilities of the data governance framework, due to
  - lack of employee skill set, or
  - lack of headcount.
- Many people working in IT and data have different responsibilities that may or may not line up with their actual role or job title, wearing different *user hats* during their workday.

2.1 Data User Categories

- Data governance roles and responsibilities can be broadly divided into three categories:
  - **Governors**: specify the governance policy
  - **Data stewards/Approvers**: implement the governance policy
  - **Users**: the people whose actions are being affected by the governance policy
- A fourth category, **ancillary** can be added for those roles that are connected to the data governance program but does not fit directly into one of the first three.
- Table 1 on the next slide shows some examples of user hats, the categories they fall into, and examples of key tasks related to the hat.

2.2 User Hats

Hat	Category	Key tasks
Legal	Ancillary	Knows of and communicates legal requirements for compliance
Privacy tsar	Governor	Ensures compliance and oversees company's governance strategy/process
Data owner	Approver (/governor)	Physically implements company's governance strategy (e.g., data architecture, tooling, data pipelining, etc.)
Data steward	Governor	Performs categorization and classification of data
Data analyst/data scientist	User	Runs complex data analytics/queries
Business analyst	User	Runs simple data analyses
Customer support specialist	Ancillary (/user)	Views customer data (but does not use this data for any kind of analytical purpose)
C-suite	Ancillary	Funds company's governance strategy
External auditor	Ancillary	Audits a company's compliance with legal regulations

Table 1: Different hats, categories and key tasks

2.2.1 Legal (Ancillary)

- Might or might not be an actual attorney.
- Tasks include:
  - ensuring the organization is up to date regarding compliance with legal requirements, and
  - communicating this information internally.
- Organizations within highly regulated industries often employ an attorney.
- Other organizations are more likely to have someone with a deep knowledge of the regulations that apply to the data handled.

2.2.2 Privacy Tsar (Governor)

- The term is used internally at Google.
- Other names include *governance manager*, *director of privacy*, and *director of data governance*.
- Key tasks include:
  - ensuring that the policies and regulations defined by the legal department are being followed, and
  - overseeing the entire data governance process, including which governance processes should be followed, and how.
- The privacy tsar may or may not have a technical background, depending on the organization and its resources.

Privacy Tsar: Work Example 1

- During the COVID-19 pandemic, Google's privacy tsars devised a solution to provide useful data to the health authorities while preserving user privacy.
- The solution consisted of:
  - Only using data from people who had turned on "location history" in Google's services,
  - Removing users with outlier results which could be used to identify them, and
  - adding statistical noise to the data, further ensuring that no individual could be identified using the data.

Privacy Tsar: Work Example 2

- Also during the COVID-19 pandemic, Google was involved in the effort of *contact tracing*, to help track the movements of people who diagnosed positively for COVID.
- Again, the benefits of helping health authorities had to be weighted against the users' right to privacy.
- The solution included:
  - Only using data from people who enabled the function on their phones.
  - Not collecting location data, as the question was whether the person had come into contact with someone else, not where it happened, and
  - Sharing the information only with the health authorities, not with Google or Apple.

2.2.3 Data Owner (Approver/Governor)

- The data owner is the person who “owns” the data.
- Key tasks include:
  - Physically implementing the processes/strategies laid out by the privacy tsar.
  - This includes:
    - Creating the data architecture of the organization,
    - Choosing and implementing tools,
    - Creating data pipelines and storage, and
    - implementing monitoring and maintenance.
- Mostly people with a technical background.

2.2.4 Data Steward (Governor)

- The hat of the data steward is key to a successful data governance program.
- Key tasks include:
  - Categorization and classification of data,
  - Ensuring that sensitive data is handled in compliance, and
  - Ensuring that the data is of high quality and can be used effectively.
- The act of data stewardship is highly manual and time consuming.
- As a consequence, in many cases there is no one dedicated person to perform data stewardship.
- This increases the risk of data categorization/classification not being done well, or even at all.

2.2.5 Data Analyst/Data Scientist (User)

- Data analysts and data scientists are some of the key users of the data within an organization.
- The more data that can be made available to analysts and scientists, the better the organization can utilize that data.
- However, the data might be sensitive and should not be available for analysis.
- A well functioning data governance program ensures that analysts and scientists are working safely with the best data available.

2.2.6 Business Analyst (User)

- Apart from data analysts and data scientists, the data is also used on the business side of the organization.
- This might create a situation where the analysts/scientists are spending time fielding requests for data instead of performing their actual tasks.
- Clearly classified data enables business users to answer some of their own questions without the risk of using sensitive data.

2.2.7 Customer Support Specialists (User/Ancillary)

- Customer support specialists are generally *viewers* of data.
- However, there might be people with this role who will need to access sensitive information.
- A strategy for granting those people access must be considered and managed by the data governance strategy.

2.2.8 C-suite (Ancillary)

- Executives might not be directly involved in the implementation of a data governance program, but they are responsible for the funding.
- Providing the tools and the headcount for the program, it is important that there is an understanding all the way up in the organization.

2.2.9 External Auditor (Ancillary)

- External auditors are not part of the organization.
- They must be considered nonetheless, as it no longer only is important for organizations to be in compliance with regulations, but also to be able to *prove* that they are.
- A clear strategy for documenting the data governance processes helps gathering the information needed in case of an external audit.

2.3 The Importance of Data Enrichment

- There is a saying: “In order to govern data, you must first know what it is.”
- The categorization, classification, and labeling of data (data enrichment) are key activities for a successful data governance program.
- These are the responsibilities of the data steward hat, which is often worn by someone also wearing other hats, such as privacy tsar and data owner.
- This increases the risk of the important data enrichment tasks not being performed.

3 Processes

- No two organizations are alike.
- Any data governance program have to be fitted to the needs of the organization, with considerations given to, amongst others:
  - resources,
  - industry, and
  - kinds of data handled by the organization.

3.1 Company Categories

- Companies can be roughly divided into a number of categories:
  - Legacy,
  - Cloud native/Digital only,
  - Retail,
  - Highly regulated,
  - Small companies, and
  - Large companies

3.1.1 Legacy Companies

- Defined as companies that have been around for a long time.
- Most likely have, or have had, legacy on-prem systems - often many different systems, leading to a number of problems, including:
  - Data governance activities not being fully done,
  - Lack of a *central data dictionary*, defining data names, classes and categories, and
  - Different branches of the company keeping their own data, unaware of the data in other branches, also known as *data siloing*.

Legacy Company: Example

- A large retail company that keeps its online sales data in one system, and the sales data from its physical stores in another system.
- Different enterprise dictionaries are being used in the two systems, making aggregations and comparisons problematic.
- Another issue is the handling of sensitive data across the organization.
- Migrating the data is a massive undertaking, and without a robust framework for the data governance for the entire enterprise, the organization is at risk of making the same mistakes again.

3.1.2 Cloud Native/Digital Only

- Defined as companies that have, and have alway had, all their data stored in the cloud.
- Even if they never had any on-prem systems, the data might still be stored across multiple cloud storage solutions, with their own version of siloing.
- The clouds might use different tools and processes for enriching data.
- There might also be differences in how the data is stored across different cloud storage solutions, leading to further issues when trying to govern the data.
- From the definition above, *all* their data is stored in the cloud. As this includes sensitive data, cloud native companies most likely have dealt with the need for governance from the beginning - at least regarding their sensitive data.

3.1.3 Retail

- Retail companies generally ingest quite a bit of data from their own stores.
- They might also ingest data from third-party agents.
- Customer data such as email addresses might be collected for the purpose of sending receipts from purchases.
- Using the same data to send marketing material is unacceptible unless the customer has given consent.
- At some companies, the same person might have access to the customer data in different roles. This calls for a process to clearly describe the use cases for different data classes.

3.1.4 Highly Regulated

- Defined as companies that deal with extremely sensitive data.
- Includes industries such as:
  - Finance,
  - Pharmaceuticals, and
  - Healthcare.
- Highly regulated companies face regular external audits to ensure that they are compliant.
- Often had data governance frameworks in place from the start.
- Difficulties include migrating data to the cloud, and trying out new tools. Most tools under development are not compliant with regulation standards (yet).

Highly Regulated Example: Hospital/University

- One hospital/university secured funding to create a specialized team for migrating a large portion of its clinical and research data to the cloud.
- The tasks to make the data comply with healthcare and research regulations involved:
  - creating an enterprise dictionary
  - enriching data
  - reviewing the presence of sensitive data
  - reviewing policies attached to data,
  - applying new policies to data, and
  - applying a standarized file structure.

3.1.5 Small Companies

- Here defined as having less than a thousand employees.
- Small data analytics teams mean that fewer people have to touch the data, reducing risk.
- Fewer data users also leads to less proliferation of datasets created by analytics/scientists, meaning less data to track and govern.

3.1.6 Large Companies

- Defined as companies with more than a thousand employees.
- Often ingest vast amount of data - more than can be enriched.
- This creates an *iceberg* of data with a small, governed portion on top of a much larger part of unknown data.
- Unknown data might be sensitive and noncompliant.
- Access control is harder to handle with ungoverned data. This might lead to too much access being granted, at the expense of risk.
- Data might be added from acquisitions of other companies. The possibility of merging new data with the old is highly reliant on how well the data governance structures of the companies works together.

4 People and Processes Together

4.1 Hats vs. Roles and Company Structure

- When it is not clearly stated who is responsible for which task in the data governance processes, there is an increased risk of inadequate work, miscommunication, and overall mismanagement.
- A successful data governance strategy will rely not simply on roles, but on clearly defined tasks, along with who is responsible or accountable for those tasks.

4.2 Tribal Knowledge

- A common problem is for data analysts to find the best dataset available for a certain use case.
- This problem is often solved by word of mouth, or "*tribal knowledge*".
- This is an obvious risk, as roles change and people move on.
- Solutions include "crowd sourcing", where analysts score datasets on usefulness.
- While it might work in organizations with small amounts of data, the method does not scale well, and is inherently fallible.
- Other solutions rely on software tools to help filter searches, minimizing the reliance on tribal knowledge.

4.3 Definition of Data

- Any organization want to be able to make informed decisions based on data.
- It would seem that the more data available to the organization, the better and more informed decisions it can make.
- However, for informed decisions to be made from data analysis, the data must be known.
- Important information about the data includes:
  - What the values in a column actually mean.
  - If the values represent sensitive data and must be treated in a certain way.
  - That the values can be trusted to represent what they are said to represent.

4.4 Old Access Methods

- In a data-driven organization, many people need to view or interact with the data in different ways.
- Each hat defined earlier has different types of tasks that need varying levels of access to the data.
- There is a connection between knowing which data needs restrictions, and deciding what those restrictions should be, and who they should affect.
- Restrictions should not only apply to roles, but there also need to be clearly stated restrictions regarding use cases.

4.5 Regulation Compliance

- Organizations in highly regulated markets, such as the financial and healthcare sectors, has had long experience with complying with regulations, and in many cases have robust data governance programs in place - at least for the regulated data.
- Regulations such as GDPR affects *all* organizations that handle personal data.
- This means that compliance has become a much bigger issue for most organizations, with far greater demands on data handling than before.
- One of the main components of GDPR is the “right to be forgotten” - to have all personal data deleted. How can an organization be in compliance if it does not know where every bit of a user’s data is stored?

5 Processes and Strategies with Varying Success

5.1 Data Segregation Within Storage Systems

- The concept of data siloing mentioned earlier can be taken advantage of.
- Examples exist of organizations setting up multiple storage systems, separating enriched/known data from the unknown data.

5.1.1 Data Segregation Strategy 1

- One way to do this is to keep all unknown data in an on-prem storage system, and only push data to the cloud after is has been curated.
- Curation includes adding metadata and attaching governance controls, such as masking or encrypting sensitive data.
- This minimizes the risk for sensitive data to be leaked from the cloud.
- Some drawbacks include:
  - Difficulties to perform cross storage analytics
  - Maintenance of multiple storage systems is time consuming and requires additional access controls.

5.1.2 Data Segregation Strategy 2

- Another way is to keep all the data in the cloud, but create zones for data in different stages of curation. Table 2 on the next slide shows an example.
- The benefits and drawbacks of this solution are nearly a reverse of those from the previous strategy.

5.1.3 Data Zones

	Types of data	Access
Insights zone	Known, enriched, curated, and cleaned data. Data also has likely had governance controls such as encryption, hashing, redaction, etc. *Example: Well-labeled, structured datasets.*	Highest level of access. Most if not all data analysts/scientists and others in a user role.
Staging zone	More known and structured data. Data from multiple sources is likely to be joined here. This is also where data engineers prep data, cleanse it, and get it ready to drop into the insights zone.	More access. Mostly data engineers - those in charge of putting together datasets for analytics.
Raw zone	Any kind of data. Unstructured and uncurated. Could also include things such as videos, text files, etc. *Example: Video files, unstructured datasets*	Very restricted access. Likely a handful of people or just an admin.

Table 2: Example of a data lake with zones for data in different stages of treatment

5.2 Data Segregation by Lines of Business

- One way of handling the effort of enriching data is by segregating the data by lines of business, and creating teams with a deep domain knowledge of their data.
- The teams are responsible for handling the:
  - data pipelines,
  - data enrichment,
  - access control and governance policy management and enforcement, and
  - data analysis of their line of business.
- Depending on the size of the organization, these task may be handled by a single person, a handful of people, or a large team.

5.3 Data Segregation by Lines of Business: Marketing as an Example

- Examples of different hats in a marketing department.

5.3.1 Data Owner

- Key tasks include:
  - Setting up and managing pipelines,
  - Managing requests for new pipelines and ingestion sources
  - Performing monitoring, troubleshooting, and fixing any data quality issues that arise
  - Implementing any technical aspects of the organization’s governance policies and strategies.

5.3.2 Data Steward

- The data steward is the *subject matter expert* (SME) in this line of business.
- Key task include:
  - Knowing:
    - What data resides where,
    - What the data means,
    - How it should be categorized and classified, and
    - What data is sensitive and what isn’t.
  - Serving as the point of contact between their line of business and the central governing body of the organization, for staying up to date on compliance and regulations, and
  - Ensuring that the data in their line of business is in compliance.

5.3.3 Business Analyst

- The business analyst is the expert on the business implications for the data in this line of business.
- Key tasks include:
  - Knowing how their data fits into the broader enterprise,
  - Communicating which data from their line of business should be used in enterprise analytics.
  - Knowing what additional data will need to be collected for this line of business to help answer whatever the current or future business questions will be.

5.3.4 Pitfalls of Data Segregation

- An obvious drawback of the data segregation strategy is the creation of data silos. Additional actions might have to be taken to enable cross enterprise analytics.

5.4 Creation of “views” of datasets

- Another common strategy is the creation of “views” of datasets. Table 3 on the next slide shows an example.
- This method allows for analytics to be run without risking unauthorized access to sensitive data.
- It is however time consuming to create and update the views as new data comes in.

5.4.1 Three types of views

Plain text customer name	Hashed customer name	Redacted customer name
Anderson, Dan	Anderson, #####	********
Buchanan, Cynthia	Buchanan, #####	********
Drexel, Frieda	Drexel, #####	********
Harris, Javiar	Harris, #####	********

Table 3: Three types of views of data

5.5 A Culture of Privacy and Security

- Any organization and employee handling data should respect data privacy and security.
- A well implemented privacy and security strategy is a part in creating not only a good but successful data governance program.
- This is accomplished by creating a collective *data culture* throughout the entire organization - not relying on individuals doing the “right thing”.

6 Sources

Eryurek, et. al: Data Governance: The Definitive Guide (Chapter 3).