Data Governance

Tools

Linus Rundberg Streuli

Lecture 2: Tools

> Data governance is fundamentally about organizational behavior. This is not a problem that can be solved through technology.

(DAMA - DMBOK, p.92)

- Data Governance is not about the tools used.
- The tools are there to help, but they do not define the process.

Table of Contents

The Enterprise Dictionary

Also known as a *policy book*.

- A document that contains the *information types* used by the organization, such as:
  - “Customer name”
  - “Street address”
  - “Salary”
- These infotypes are classified into *data classes* according to their content, such as:
  - PII *(Personally Identifiable Information)*
  - Financial Information
  - Business Intellectual Property

Data Classes and Policies

With the data classified into data classes, different *policies* can be applied:

- access control (e.g. analysts cannot access data classified as PII)
- retention (e.g. records of financial transactions must be kept for a certain time)
- protection (e.g. data classes to be encrypted or masked)

Enterprise Dictionary as Documentation

- A well structured policy book will help with the understanding, organizing and enforcing of policies.
- It will also serve as documentation of policies in the event of an external audit by a regulator.

Per-Use-Case Data Policies

- The same data might have different meaning in different contexts.
- A customer might want their order to be delivered to their address, but not want marketing materials.
- Classifying data after “use case” helps with compliance - the customer address is available in a shipping context, but only in a marketing context if consent has been given.

Managing Metadata

- Metadata - *data about the data* - is central to data classification.
- Examples of metadata are:
  - where the data is stored
  - any governance controls, such as access control, that applies to the data
  - the owner of the data

Considering metadata policies

- As metadata does not contain the actual data, the policies applied to the underlying data does not need to apply to the metadata.
- Knowing that *the data exists* is not the same as knowing what *the data contains*.
- A data user might search for a table, try to find out from the metadata if the table actually contains relevant data, and if so request access to the data.

The Data Catalog

The metadata is managed through the use of a *data catalog*.

The data catalog contains information about the data in the organization, such as

- where the data is located
- table schemas and names
- column names and descriptions

and usually additional information such as

- who owns the data
- if the data is locally generated or externally purchased

Data Assessment and Profiling

- At some point, the data is (hopefully) going to be used.
- Prior to this, it is important to assess the quality of the data *in light of the use case*.
- A marketing team and a fraud detection team look for different patterns, so their data should have different quality metrics.
- When cleaning and normalizing data, the use case must always be decided beforehand.

Data Quality

- Different kinds of data should have different levels of confidence attached to them.
- There is a difference between a human noting the temperature by hand each day, and an IoT device reporting the temperature to a server.

Data Quality

- Knowing what level of quality to expect from a data source is a part of enhancing the trust in the data - the core goal of a data governance program.
- Establishing ownership of data sources - making individuals or business teams responsible for the quality of the data - is one way of ensuring the data quality is up to the standards set by the organization.

Lineage Tracking

Data lineage is the data's way from generation, through transformations and aggregations, to use for analysis and insights.

Tracking the data lineage is important for a number of reasons:

- **quality assessment:** has high quality data been "tainted" by joining it with lower quality data?
- **privacy and protection:** is sensitive data being exposed somewhere in the process?
- **explainability**: with a clear picture of the data lineage, decisions made by ML models trained on that data become more easily explainable

Data Retention and Data Deletion

Different kinds of data need different policies on retention and deletion.

- PII might be kept just for as long as strictly necessary, for easier compliance with regulations such as GDPR
- Financial transactions, on the other hand, often need to be kept for a certain time to facilitate investigations of fraud and other economical crimes

Deciding on what data to keep (perhaps encrypted or masked), and what data to delete, is a key part of a data governance policy.

Data Acquisition

An example of data governance in action could be the acquisition of data by a data analyst, seeking to perform a task.

1. The data analyst performs a search in the data catalog and is able to review the relevant data sources.
2. The analyst then seeks access to the relevant data sources. The data governance program clearly states who is responsible for granting access - and for which use cases access is to be granted.
3. Access is granted and the data analyst can perform the task, using the data. The access is logged.

Types of tools

As previously noted: data governance is not about the tools, but they are of course important nonetheless.

There are different types of tools to help with the different aspects of a data governance program. These include:

- Infrastructure providers (most often in the cloud)
  - Microsoft Azure
  - Amazon Web Services (AWS)
  - Google Cloud
- Master Data Management Solutions
- Data Quality Tools
  - Dataprep
  - Stitch

Data Governance Tools

Some tools that help with implementing a data governance program include:

- Microsoft Purview (formerly Azure Purview)
  - Auditing
  - Data Lifecycle Management
  - Data Catalog
  - Information Protection
  - … and many more features

Data Governance Tools

- Collibra
  - Data Governance Framework
  - Data Catalog
  - Lineage Tracking
  - … and many more features

Data Governance Tools

- Apache Atlas
  - Open source
  - Many of the same features as Purview and Collibra

Summary

Tools and processes mentioned:

- Documenting the data governance program in an **Enterprise Dictionary**
- **Classifying** and **cataloging data**
- Managing **metadata**
- Assessing and managing **data quality**
- Tracking **data lineage**
- Making decisions on **data retention and deletion** policies