The Data Life Cycle

Linus Rundberg Streuli

Table of contents

1 What is a Data Life Cycle?

- A data life cycle can be defined as the order of the stages, or phases, a piece of data goes through, from its initial generation or capture, to its eventual archival or deletion at the end of its useful life.
- There is no one catch-all description of a data life cycle as the phases differ between organizations and authors.
- There are however certain characteristics of the stages which may be used to formulate governance strategies for each different phase.
- For many organizations, data can be divided into:
  - *transactional data*, which is created and stored in vast amounts in systems optimized for fast transactions, and
  - *analytical data*, which is used for analytics, and stored in systems optimized for analytical processes.
- When moving data from a transactional to an analytical system, the data typically undergoes the phases of the data life cycle outlined in this lecture.
- Defining this process is a core undertaking in operationalizing a data governance program.

2 Phases of a Data Life Cycle

- The framework chosen for the data governance program will ultimately guide the organization into which processes are put in place.
- The phases does not necessarily describe actual data flows, but rather a mapping of what processes are applied to the data at different points of its life cycle. Different pieces of data might flow back and forth in a system, or skip stages, or pass the stages in a different order.
- Each of the phases in Figure 1 has distinct characteristics, and we will take a closer look at each phase.

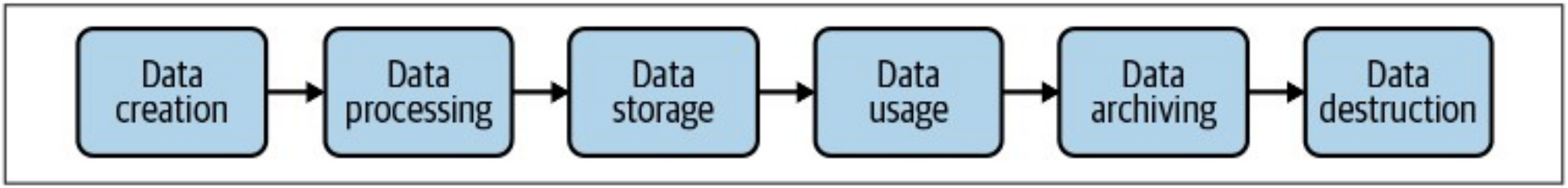

Figure 1: Phases of a data life cycle (DG TDG p. 87)

2.1 Data Creation

- Data is generated from multiple sources, in different formats, and in different frequencies.
- Data may come from existing data connections, extract-transform-load (ETL) pipelines, third-party ingestion tools, and more.
- In this phase, metadata might also be created and captured.
- Data is typically created in one of three ways:
  - *Data acquisition*: when an organization acquires data that has been produced by a third-party organization,
  - *Data entry*: when new data is manually entered by humans or devices within the organization, and
  - *Data capture*: when data generated by various devices in an organization, such as IoT sensors, is captured.
- These different ways all offer their own data governance challenges. For example, data from a third-party provider might come with certain rules for how the data can be used.

2.2 Data Processing

- Data processing is also referred to as *data maintenance*, and describes the phase when the data goes through processes such as integration, cleaning, scrubbing, or ETL, to get the data ready for storage and eventual analysis.
- Some of the governance implications related to this stage are:
  - *Data lineage*: As the data is being processed, how do we ensure that its lineage is tracked and maintained?
  - *Data quality*: Are we getting all the important values?
  - *Data classification*: are we dealing with sensitive information? How are we dealing with it?
- As the data is moving through this phase, it needs to be encrypted, first in flight, and then later at rest.

2.3 Data Storage

- Data and metadata are stored in systems with the appropriate levels of protection.
- When dealing with analytical data, the system might be a data warehouse, a data mart, or a data lake.
- Data needs to be encrypted to protect from leaks, and backed up in case of data loss, accidental deletion, or disaster.

2.4 Data Usage

- This phase is where the data is used within the organization to support objectives and operations.
- Data users might ask questions of the data, and the answers will guide in making business decisions.
- The correctness of the answers is highly dependent on the quality of the data. A clear governance strategy for ensuring the data quality is very important.
- In other cases, the data of this stage is the actual product. If this is the case, other governance policies will need to be considered.
- As the data in this phase is being used across the organization, proper access controls and audits are key.

2.5 Data Archiving

- In this phase, data is taken out of all active production environments and moved to another environment.
- It is no longer processed, used or published, only stored in case it is needed again.
- A data governance plan should define how long the data is kept in this stage, along with how it is controlled.

2.6 Data Destruction

- Data destruction, also referred to as *purging*, is the process of removing every copy of the data from the organization.
- Keeping all data forever is not an option for most organizations - the cost of storage space and compliance are two common drivers behind destroying data.
- Compliance might also be a driver behind *keeping* data, such as financial transactions.
- Understanding regulations, industry standards and governance policies is key to defining the correct timeline.
- An organization must also be able to prove that the data has actually been purged.

3 Data Life Cycle Management

- Data Life Cycle Management (DLM) refers to a comprehensive policy-based approach to manage the flow of data throughout its life cycle.
- The basis of DLM is the Data Management Plan.

3.1 Data Management Plan

- A data management plan describes how data will be managed, described, and stored.
- It also defines standards and policies for how the data will be handled and protected throughout its life cycle.
- Many different templates and frameworks exists for different types of organizations.
- A DMP should be a living document that works as a roadmap to guide and explain how data is to be treated during its life cycle.
- We will now look at five concepts to consider when creating a DLM.

3.1.1 Guidance 1: Identify the data

- To guide in the building of infrastructure, three pieces of information about the data need to be determined:
  - *Types*: What kind of data types will be stored? Are they structured, unstructured, or both?
  - *Sources*: Where is the data coming from? Are there restrictions on how this data can be used or manipulated? What are those rules?
  - *Volume*: It might be difficult to predict how much data is going to be stored, but estimations and plans for eventual expansion help to be prepared.

3.1.2 Guidance 2: Define how the data will be organized

- What tools are needed across the data life cycle?
- What storage systems are needed? A data warehouse? A data lake?

3.1.3 Guidance 3: Document a data storage and preservation strategy

- How long will a piece of data be accessible?
- Who will have access?
- How will the data be backed up?
- How and when will the data be purged?

3.1.4 Guidance 4: Define data policies

- What licenses, sharing agreements and regulations apply to the data being handled?
- How to ensure that data policies are being adhered to?
- Documenting data policies and how they are being followed helps in case of external audits.

3.1.5 Guidance 5: Define roles and responsibilities

- Which are the right roles for the organization?
- Which teams will be responsible for metadata management and data discovery?
- Who will ensure governance policies are followed all the way?

4 Applying Governance over the Data Life Cycle

- Governance needs to bring together people, processes and technology to govern data throughout its life cycle.
- Implementing governance is complicated - there is no easy way to simply stitch everything together and consider the work done.
- Organizations might develop their own data governance framework, or purchase a fully integrated platform.

4.1 Data Governance Framework

- Frameworks helps to visualize the plan.
- There are several frameworks, such as:
  - **DGI Data Governance Framework**: The Data Governance Institute is comprehensive and pragmatic and is well suited for organizations starting from scratch.
  - **McKinsey**: More focused on organizing people, than processes and technology.
  - **BCG**: The Boston Consulting Group's framework is akin to the DGI framework but not as comprehensive.
  - **DAMA-DMBOK**: The DAMA-DMBOK framework focuses on overall data management, where data governance is a central function in the model.
- Further reading: https://www.integrate.io/blog/popular-data-governance-frameworks/

4.2 Data Governance in Practice: OpenStreetMap

- An example of an organization that have implemented a successful data governance strategy is OpenStreetMap (OSM)[1].
- OSM is an open source, crowd sourced mapping system. By 2020, over six million users were registered, and 16 000 users made changes to the data every week.
- The main rules regarding contributing to OSM are not technical (*how*), but rather focused on the content (*what*).
- Figure 2 on the next slide shows the guidelines for adding content to OSM.[2]

1. https://www.openstreetmap.org
2. https://wiki.openstreetmap.org/wiki/How_We_Map

4.2.1 OpenStreetMap Guidelines

Contributions to OpenStreetMap should be:

- **Truthful** - means that you cannot contribute something you have invented.
- **Legal** - means that you don't copy copyrighted data without permission.
- **Verifiable** - means that others can go there and see for themselves if your data is correct.
- **Relevant** - means that you have to use tags that make clear to others how to re-use the data.

Figure 2: Screenshot from the OSM Wiki

4.2.2 OpenStreetMap Approaches

- This is not to say that OSM does not have data quality checks in place when accepting contributions.
- Added data is transformed to follow the standards and formats set by the community.
- As OSM is a community effort, there is no one central agent supervising the quality of the added data.
- Instead, the person contributing is the owner of their data, and every contribution is under the supervision of other users, locally and globally.
- In the case of OSM, this strategy has proven successful.

4.3 More examples of the data life cycle stages

4.3.1 Data creation

- This is the first phase of the data life cycle.
- In this phase, organizations can choose to capture both metadata (data about the data) and data lineage (where the data has been, how it is going to be transformed, and where it is going).
- Having captured metadata and lineage at an early stage can be advantageous in later stages.
- Additionally, data classification and categorization might be performed in this phase, especially if the data is sensitive.

4.3.2 Data processing

- In this phase, the data is processed, to get it ready for storage and eventual analysis.
- Processes might include:
  - integration,
  - cleaning,
  - scrubbing, and
  - ETL (extract-transform-load).
- The integrity of the data must be preserved during these processes. Here, data quality plays an important role. We will be looking closer into data quality in another lecture.
- Data lineage must be captured and tracked at this stage as well to ensure that users understand where the data is coming from and what transformations have been applied to it.

4.3.3 Data storage

- In this phase, the data is stored, waiting to be used for analysis.
- The data should be encrypted and backed up.

4.3.4 Data usage

- In this phase, the data is analyzed and visualized for analysis and insights by multiple stakeholders across the organization.
- A data catalog is vital to helping users discover data, using metadata.
- Access control, privacy regulations - internal and external - and audits are important during this phase.

4.3.5 Data archiving

- In this phase, data is removed from all active production environments.
- The data classification from earlier phases should guide the timeline and storage methods for different kinds of data.
- Archived data must be kept as safe as “live” data - encryption and backup methods must be defined.

4.3.6 Data destruction

- In the final phase, the data is removed from the enterprise in its totality.
- As in the archiving phase, data classification should guide which data is purged when, and how.
- Regional, national and international regulations might apply regarding how long certain kinds of data are allowed to, or must be, kept. Staying up to date with these regulations is a part of a data governance program.

5 Example of How Data Moves Through a Platform

5.1 Scenario

- We will be imagining a business that wants to ingest data onto a cloud-data platform such as Google Cloud, AWS or Azure.
- The data may include sensitive information, such as phone numbers and email addresses.
- Such an operation may contain the following parts.

5.2 Example of How Data Moves Through a Platform, pt. 1

1. An ingestion data pipeline is configured:
    a. Goal: The data will be scanned, classified, and tagged.
    b. The data might be temporarily divided into ingestion buckets:
        i. Ingest: heavily restricted
        ii. Released: processed data
        iii. Admin quarantine: needs review
2. The data is scanned and classified for sensitive information.
3. Sensitive data might be redacted or otherwise anonymized. This might generate metadata such as tokenization keys.

5.3 Example of How Data Moves Through a Platform, pt. 2

4. Data is tagged with labels denoting personally identifiable information.
5. Data quality can be assessed: are there missing values, is the data in the correct format, etc.
6. Capture data lineage information.
7. As the data moves between services, it needs to be encrypted in flight.
8. Once ingested, the data needs to be encrypted and stored in a data warehouse and/or a data lake. Backup and recovery processes must be in place.
9. While in storage, additional metadata can be added to the data and cataloged so that users can find the data.
10. Audit trails and accesses must be captured in order to quickly mitigate threats and security issues.

5.4 Example of How Data Moves Through a Platform, pt. 3

11. During this phase, a robust identity and access management (IAM) solution must be in place to ensure that only people with the right permissions can access the data.
12. Additional privacy and anonymization methods may be applied to the data before analysis.
13. As the data is no longer needed by the organization, it is archived according to the policy set up in the governance program.
14. At the end of its useful life, the data is completely removed and destroyed.

6 Operationalizing Data Governance

- One infamous example of what can happen when data governance fails is NASA:s Mars Climate Orbiter project[1].
- The MCO was permanently lost during its attempt to enter into orbit around Mars in the fall of 1999.
- The cause of the failure turned out to be a unit mismatch between the software developed by NASA, and the software delivered by one of its subcontractors, the Lockheed Martin Corporation.
- NASA used the metric system, while Lockheed Martin used US customary units.
- The existance of a proper data governance policy had increased the possibility of the mismatch being discovered and handled earlier, saving over $500 million that were lost as the satellite vanished above Mars.

1. https://en.wikipedia.org/wiki/Mars_Climate_Orbiter

6.1 What is a Data Governance Policy?

- A data governance policy is a documented set of guidelines for ensuring that an organization’s data and information assets are managed consistently and used properly.
- The guidelines will include individual policies for data quality, access, security, privacy, and usage.
- Data governance policies also establish roles and responsibilities for data regarding access, disposal, storage, backup, and protection.
- The data governance policy is usually created by a data governance committe or council, made up of business executives and other data owners.
- To start thinking and talking about a data governance policy, a data governance charter template might be useful.
- Figure 3 on the next slide shows an example of a charter template.

6.1.1 Data Governance Charter Template

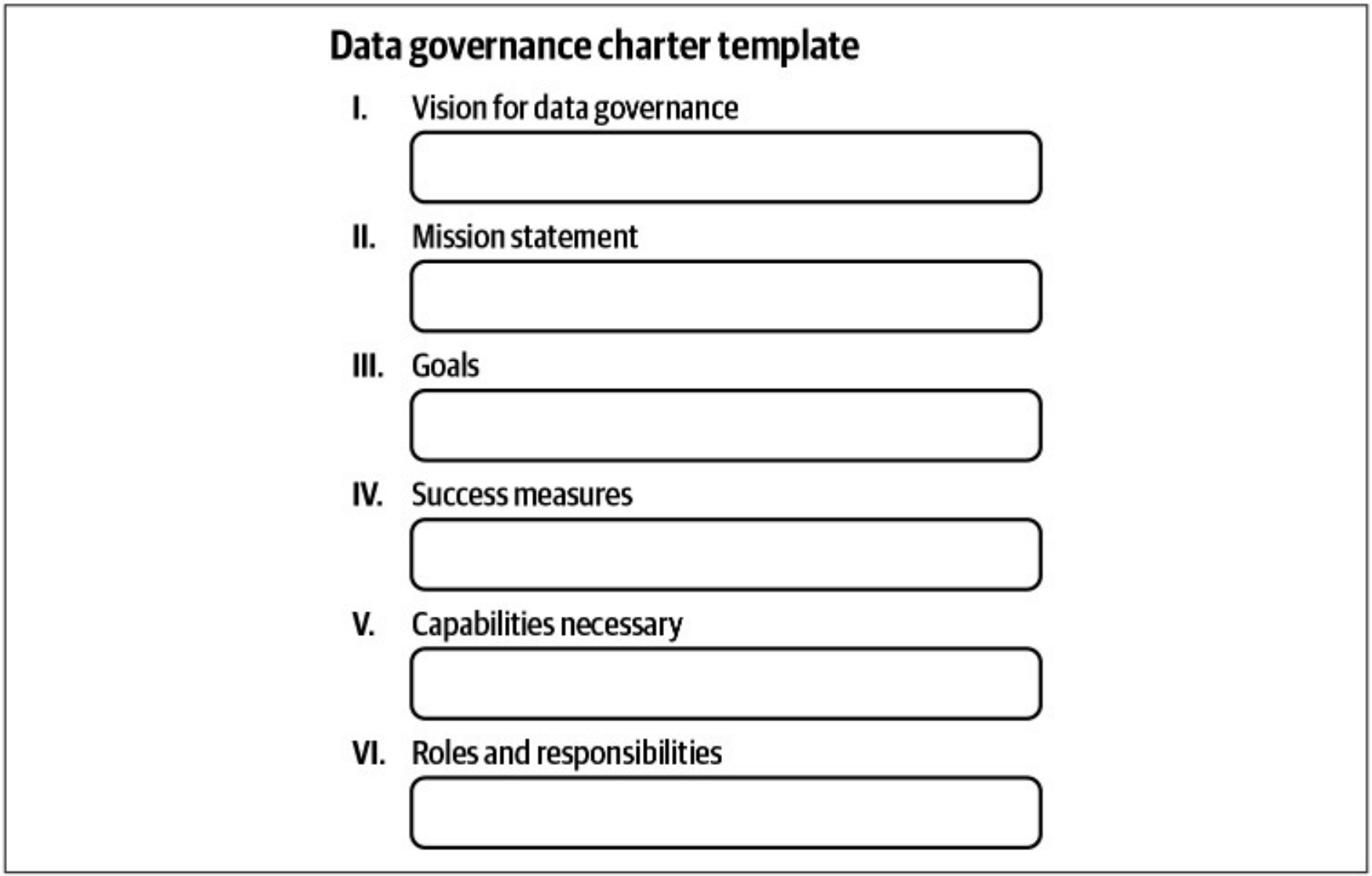

Figure 3: An example of a data governance charter template. DG TDG p. 102

6.2 Importance of a Data Governance Policy

- A data governance policy allows the organization to have all the important elements of operationalizing data governance documented according to its needs and objectives.
- It also allows consistency across the organization over a long period of time.
- When a data government policy is well drafted, it will ensure:
  - Consistent, efficient, and effective management of data assets throughout the organization,
  - The appropriate level of protection of the organization’s data assets based on their value and risk, and
  - The appropriate protection and security levels for different categories of data.

6.3 Developing a Data Governance Policy

- The body responsible for drafting the data governance policy will start by identifying risks and regulatory requirements and look into how they will impact or disrupt the business.
- When the risks and assessments have been identified, the committee or council will draft policy guidelines that will ensure the organization has the envisioned data program.
- Part of the process of developing a data governance policy is establishing the expectations, wants, and needs of key stakeholders through interviews, meetings, and informal conversations.

6.4 Data Governance Policy Structure

- **Vision and mission for the program**: Examples of visions might be to drive digital transformation, or to use data to provide new products and services.
- **Policy purpose**: Capture goals, as well as metrics for success. Should be driven by the vision and mission.
- **Policy scope**: Document which data assets are covered by the policy.
- **Definitions and terms**: Document definitions and terms to ensure everyone is on the same page.
- **Policy principles**: Define rules regarding data access, data usage, data integration and data integrity.
- **Program structure**: Define roles and responsibilities. A *RACI matrix* helps to define who is, or is to be, Responsible, Accountable, Consulted, and Informed.
- **Policy review**: Determine when the policy should be reviewed and updated, and how adherence to the policy should be monitored, measured, and remedied.
- **Further assistance**: Document the right people to address questions from the teams and other stakeholders.

7 Step-by-Step Guidance

7.1 Build the Business Case

- Data governance takes time and is expensive.
- Any data governance initiative will need to start with building the business case that will identify critical business drivers and justify the effort and costs involved.
- It is OK to start small, strive for quick wins, and build up ambitions over time.

7.2 Document Guiding Principles

- Develop and document core principles associated with governance.
- Core principles might be:
  - to make consistent and confident business decisions based on trustworthy data,
  - to comply with regulations and so avoid fines and possible public backlash, or
  - to optimize staff effectiveness by providing data assets that meet the desired data quality thresholds.

7.3 Get Management Buy-In

- Management controls the big decisions and a project on the scale of a data governance program is dependent on funding.
- Ways to get management buy-in include:
  - Outlining important KPIs and how the plan helps to move them,
  - Engaging data governance champions and key senior stakeholders,
  - Presenting a business case to show the value of the project

7.4 Develop an Operating Model

- How is the plan integrated into the day-to-day business of the organization?
- The operation model can be defined using the data governance policy.

7.5 Develop a Framework for Accountability

- Define ownership, and provide a methodology to ensure that everyone is accountable for contributing to data usability.

7.6 Develop Taxonomies and Ontologies

- Clearly define the categories associated with data classification, organization and protection.
- A common “data language” across the organization ensures that the users of the data follow the guidelines of the data governance policy.

7.7 Assemble the Right Technology Stack

- Find the tools that facilitate implementing the data governance program. The tools may help with:
  - Creating a data catalog of metadata,
  - Tracking data lineage and data accesses, and
  - Reporting and mitigating issues related to data quality, privacy and security.

7.8 Establish Education and Training

- Develop educational material highlighting data governance practices, procedures, and the use of assisting technologies.
- Plan for regular training sessions to reinforce good data governance practices.
- Use business terms where possible.

8 Considerations for Governance Across a Data Life Cycle

8.1 Deployment Time

- As mentioned, implementing a data governance program can be a massive undertaking. Starting small is for the most cases an option.
- Automation can reduce the deployment time by handling some of the manual tasks.
- Use of artificial intelligence in data governance is becoming more common, especially regarding discovery of sensitive information, and metadata management.
- The amount of automation available depends on the organization and the data it handles, as well as the tools and frameworks chosen.

8.2 Complexity and Cost

- There are a number of industry standards that applies to metadata, but they are concerned with different types of metadata, such as descriptive, technical, and statistical.
- In most cases, metadata does not fall under the same policies and regulation as the actual data.
- A lack of standardized metadata specification means that different products and processes will have different ways of presenting that information.
- Another added complexity is the amount of processes, tools and infrastructure needed to make data governance a reality.
- Platforms offered by cloud service providers are built with this in mind, and companies like Informatica, Alation and Collibra offer data management solutions.

8.3 Changing Regulation Environment

- Regulations define a lot of what must be done and implemented to ensure governance.
- They will outline how certain types of data need to be handled and which types of controls need to be in place.
- As technology evolves, so does regulations. A well thought out data governance program helps when an organization must comply with new regulations that come into effect.

8.4 Location of Data

- Having data both in the cloud and in on-prem systems add to the complexity of the data governance program.
- The program must take into account how the data moves between systems during its life cycle, and allow for cloud, on-prem, and hybrid scenarios.

8.5 Organizational Culture

- The culture of an organization is an important part of implementing a data governance program.
- An organization with an open culture might find it easier as people are more likely to voice their opinions. On the other hand, it may be harder to make people follow a governance policy in a flat organization where employees are used to a greater amount of autonomy.
- In organizations where people are not heard or even afraid to speak up, governance might be harder to implement.
- In the NASA example, people in the organization did raise the issue and reported it, but were ignored by management.
- The question of organizational culture is central to data governance and will be the topic of another lecture.

8.6 Further resources

- There are a number of organizations dedicated to providing resources for implementing data governance frameworks, including:
  - The Data Governance Institute (DGI) (https://www.datagovernance.com)
  - The Data Management Association (DAMA) (https://dama.org)
  - The Data Governance Professionals Organization (https://dgpo.org)
  - The Enterprise Data Management Council (https://edmcouncil.org)

9 Sources

- Eryurek, et. al: Data Governance: The Definitive Guide (Chapter 4).
- https://www.integrate.io/blog/popular-data-governance-frameworks/
- https://www.openstreetmap.org
- https://en.wikipedia.org/wiki/Mars_Climate_Orbiter