Principles and best practices for data governance in the cloud

A thought-leadership whitepaper from Google Cloud

This white paper provides guidance and best practices for data governance as you move your data into the cloud. It provides a framework for data governance in the cloud, deep dives into how to operationalize data governance in your organization and outlines the business benefits of a robust data governance implementation.

Google Cloud

Google Cloud

Table of Contents

Google Cloud

1. Introduction

In recent years, the ease of moving to the cloud has motivated and energized a fast-growing community of data consumers to collect, capture, store, and analyze data for insights and decision making. As adoption rates of cloud computing continues to grow, information management stakeholders have questions about potential risks of managing their data in the cloud for a number of reasons:

1. **Securing the data:**
   Storing data in a public cloud infrastructure might concern large enterprises that typically deploy their systems on-premises and expect tight security. With the increased number of security threats and breaches in the news, organizations are concerned that they might be the next victim. These factors contribute to risk management concerns for protecting against unauthorized access to or exposure of sensitive data ranging from personally identifiable information (PII) to corporate confidential information, trade secrets, or intellectual property.

2. **Regulations and compliance:**
   There is a growing set of regulations, including the California Consumer Privacy Act (CCPA), the European Union's General Data Protection Regulation (GDPR), and industry-specific standards such as Global Legal Entity Identifier (LEI) in the Financial industry and ACORD data standards in the insurance industry. Compliance teams responsible for adhering to these regulations and standards may have concerns about oversight and control of data stored in the cloud.

3. **Visibility and control:**
   Data management professionals and data consumers sometimes lack visibility into their own data landscape; not knowing which data assets are available, where they are located, and how and if they can be used, who has access to it, and if they should have access to it. This uncertainty limits their ability to further leverage their own data to improve productivity or drive business value.

These risk factors clearly highlight the need for increased data assessment, cataloging of metadata, access control management, data quality, and information security as core data governance competencies that the cloud provider should not only provide, but also continuously upgrade in a transparent way. In essence, addressing these risks without abandoning the benefits provided by cloud computing has elevated the importance of not only understanding data governance in the cloud, but also knowing what is important. Good data governance can inspire customer trust and lead to vast improvements in customer experience.

Google Cloud

2. Why your business needs data governance in the cloud

As your business generates more data and moves it in the cloud, it's important to note that the dynamics of data management change in a number of fundamental ways. Organizations should take note of the following:

1. **Risk Management:**
   There are concerns about potential exposure of sensitive information to individuals or systems without authorization, security breaches or even known personnel accessing data under the wrong circumstances. Organizations are looking to minimize this risk so additional forms of protection are required (such as encryption) that obfuscate the data object's embedded information should a system breach occur, tools to support access management or even forms of identifying and creating policy around sensitive data assets.

2. **Data Proliferation:**
   The speed at which businesses create, update, and stream their data assets has increased and, while cloud-based platforms are capable of handling increased data velocity, volume and variety, it is important to introduce controls and mechanisms to rapidly validate the quality aspects of high bandwidth data streams.

3. **Data Management:**
   The need to adopt externally-produced data sources and data streams (including paid feeds from third parties) means that you should be prepared not to trust all external data sources. You may need to introduce tools that document data lineage, classification, and metadata, to help your employees (data consumers, in particular) determine data usability based on their knowledge of how the data assets were produced.

4. **Discovery (and data awareness):**
   Moving data into any kind of data lake (cloud-based or on-premises) runs the risk of losing track of which data assets have been moved, the characteristics of their content, and details about their metadata. The ability, therefore, to assess data asset content and sensitivity (no matter where it is) becomes very important.

5. **Privacy and compliance:**
   Regulatory compliance demands auditable and measurable standards and procedures which ensure compliance with internal data policies as well as external government regulations. Migrating data to the cloud means that organizations need tools to enforce, monitor and report compliance, as well as ensure that the right people and services have access and permissions to the right data.

Google Cloud

> Google Cloud Platform makes perfect sense for running a safer, distributed and scalable payment infrastructure. The fact that Google uses credible third-party auditors who certify the infrastructure according to international best-practices worked greatly to our benefit.

- Daniel Döderlein, Founder and Chief Executive Officer, Auka

3. Framework and best practices for data governance in the cloud

Given the changing dynamics of data management, how should organizations think about data governance in the cloud, and why is it important? According to TechTarget*, Data Governance is:

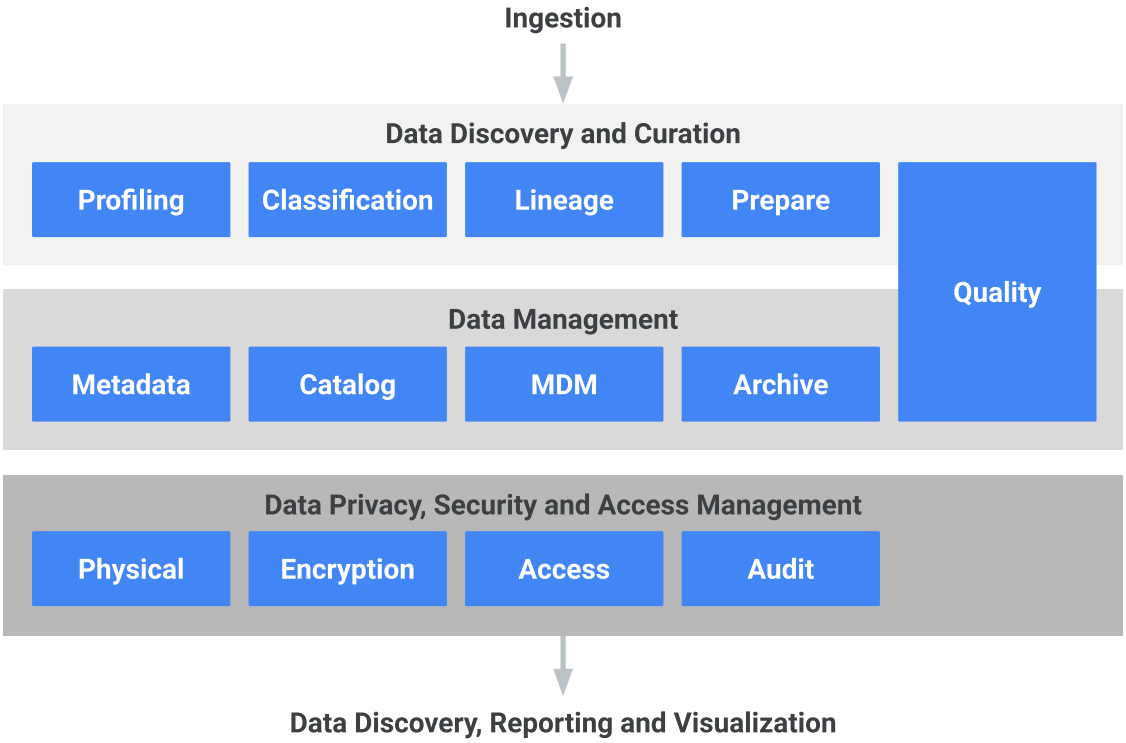
> **"The overall management of the availability, usability, integrity and security of data used in an enterprise. A sound data governance program includes a governing body or council, a defined set of procedures and a plan to execute those procedures."**

Simply put, data governance encompasses the ways that people, processes and technology can work together to enable auditable compliance with defined and agreed-upon data policies.

3.1 Data Governance Framework

Enterprises need to think about data governance comprehensively. Starting from data intake and ingestion, cataloging, persistence, retention, storage management, sharing, archiving, backup, recovery, loss prevention, disposition, and removal and deletion. Gartner provides an easy-to-understand data governance framework (for a data lake) that tracks data governance across the data lifecycle.

https://searchdatamanagement.techtarget.com/definition/data-governance



Gartner, Applying Effective Data Governance to Secure Your Data Lake, Sanjeev Mohan, April 17, 2018

The framework provides an easy to understand architecture, however, it's important to note that these steps are often not linear. Given this framework, let's review some best practices for data governance in the cloud:

1. **Data discovery and assessment:**
   Cloud-based environments often offer an economical option for creating and managing data lakes, but the risk remains for ungoverned migration of data assets. This risk represents a potential loss of knowledge of what data assets are in the data lake, what information is contained within each object, and where those data objects originated from. **A best practice for data governance in the cloud is data discovery and assessment in-order to know what data assets you have.** The data discovery and assessment process is used to identify data assets within the cloud environment, to trace and record each data asset's origin, lineage, what transformations have been applied, and object metadata. (Often this metadata describes the demographic details, such as the name of the creator, the size of the object, the number of records if it is a structured data object, or when it was last updated.)

2 **Data classification and organization:**
Properly evaluating a data asset and scanning the content of its different attributes can help categorize the data asset for subsequent organization. This process can also infer whether the object contains sensitive data, and if so, classifying each data asset in terms of the different levels of data sensitivity, including personal and private data, confidential data, and intellectual property. **To implement data governance in the cloud, you'll need to profile and classify sensitive data in order to inform which governance policies and procedures apply to the data.**

> "As we work with sensitive datasets, security of any platform is of the utmost importance. We are confident in the security and reliability of hosting our datasets on Google Cloud Platform."

- Bindu Thota, Director of Product Management, zulily

3 **Data cataloging and metadata management::**
Once your data assets are assessed and classified, it is crucial that you document your learnings so that your communities of data consumers have visibility into your organization's data landscape. **You need to maintain a data catalog that contains structural metadata, data object metadata, and the assessment of levels of sensitivity in relation to the governance directives (such as compliance with one or more data privacy regulations).** The data catalog not only allows data consumers to view this information, but it can also serve as part of a reverse index for search and discovery, both by phrase and (given the right ontologies) by concept. It is also important to understand the format of structured and semi-structured data objects, and allow your systems to handle these data types differently, as necessary.

4 **Data quality management:**
Different data consumers may have different data quality requirements, so it's important to **provide a means to document data quality expectations as well as techniques and tools for supporting the data validation and monitoring process.** Data quality management processes include creating controls for validation, enabling quality monitoring and reporting, supporting the triage process for assessing the level of incident severity, enabling root cause analysis and recommendation of remedies to data issues, and data incident tracking. The right processes for data quality management will provide measurably trustworthy data for analysis.

5 **Data access management:**
There are two aspects of governance for data access. The first aspect is the provisioning of access to available assets. I**t's important to provide data services that allow data consumers to access their data**, and fortunately, most cloud platform providers provide methods for developing data services. The second aspect is prevention of improper or unauthorized access. **It's important to define identities, groups, and roles, and assign access rights to establish a level of managed access.** This best practice involves managing access services as well as interoperating with the cloud provider's Identity Access Management (IAM) services by defining roles, specifying access rights, and managing and allocating access keys for ensuring that only authorized and authenticated individuals and systems are able to access data assets according to defined rules.

> "At Qubit, we love the flexibility of GCP resource containers including Organizations and Projects. We use the Organization resource to maintain centralized visibility of our projects and GCP IAM policies to ensure consistent access controls throughout the company. This gives our developers the capabilities they need to put security at the forefront throughout our migration to the cloud."

- Laurie Clark-Michalek, Infrastructure Engineer at Qubit.

6 **Auditing:**
Organizations must be able to assess their systems to make sure that they are working as designed. Monitoring, auditing and tracking (who did what and when and with what information) helps security teams gather data, identify threats, and act on them before they result in business damage or loss. **It's important to perform regular audits: check the effectiveness of controls in order to quickly mitigate threats and evaluate overall security health.**

7 **Data protection:**
Despite the efforts of information technology security groups to establish perimeter security as a way to prevent unauthorized individuals from accessing data, perimeter security is not, and never has been sufficient for protecting sensitive data. Attempting to prevent someone from breaking into your system has limited success, but at some point, your data may become exposed. You might sustain a security breach, or even insider exfiltration. **It's important to institute additional methods of data protection to ensure that exposed data cannot be read, including encryption at rest, encryption in transit, data masking, and permanent deletion.**

> "Google Cloud Platform has best-of-breed security integrated throughout. It handles authentication, bot mitigation, protection against Distributed Denial of Service attacks, compliance, and much more. All we have to do is innovate. Google does the rest."

- Patrick Aluise, SVP of Digital and Augmented Reality, Moviebill

4. Operationalizing data governance in your organization

Technology certainly helps support the data governance principles presented above, but data governance goes beyond the selection and implementation of products and tools. The success of a data governance program depends on a combination of:

- **People** to build the business case, develop the operating model, and take on appropriate roles
- **Processes** that operationalize policy development, implementation, and enforcement
- **Technology** used to facilitate the ways that people execute those processes

These are critical steps in planning, launching, and supporting a data governance program:

1. **Build the business case:**
   Establish the business case by identifying critical business drivers to justify the effort and investment of data governance. Outline perceived data risks (such as concern about storing data on cloud-based platforms) and indicate how data governance helps the organization mitigate those risks.

2. **Document guiding principles:**
   Assert core principles associated with governance and oversight of enterprise data. Document those principles in a data governance charter to present to senior management.

3. **Get management buy-in:**
   Engage data governance champions and get buy-in from the key senior stakeholders. Present your business case and guiding principles to C-Level management for approval.

4. **Develop operating model:**
   Once you have management approval, define the data governance roles and responsibilities and then describe the processes and procedures for the data governance council and data stewardship teams who will define processes for defining and implementing policies as well as reviewing and remediating identified data issues.

5. **Establish a framework for accountability:**
   Establish a framework for assigning custodianship and responsibility for critical data domains. Make sure there is visibility to the "data owners" across the data landscape. Provide a methodology to ensure that everyone is accountable for contributing to data usability.

6. **Develop taxonomies and ontologies:**
   There may be a number of governance directives associated with data classification, organization, and in the case of sensitive information, data protection. To enable your data consumers to comply with those directives, there must be a clear definition of the categories (for organizational structure) and classifications (for assessing data sensitivity).

7. **Assemble the right technology stack:**
   Once you've assigned data governance roles to your staff, defined and approved your processes and procedures, you should then assemble a suite of tools that facilitate ongoing validation of compliance with data policies and accurate compliance reporting.

8. **Establish education and training:**
   Raise awareness of the value of data governance by developing educational materials highlighting data governance practices, procedures, and the use of supporting technology. Plan for regular training sessions to reinforce good data governance practices.

Google Cloud

5. The business benefits of robust data governance

Data security, data protection, data accessibility and usability, data quality, and other aspects of data governance will continue to emerge and grow as critical priorities for organizations. And as more organizations migrate their data assets to the cloud, the need for auditable practices for ensuring data utility will also continue to grow. To address these directives, businesses should frame their data governance practice around three key components:

- A framework that enables **people** to define, agree to, and enforce data policies
- Effective **processes** for control, oversight, and stewardship over all data assets across both on-premises systems, cloud storage, and data warehouse platforms
- The right **tools and technologies** for operationalizing data policy compliance

> "Google approaches security and compliance in a very comprehensive, principled way, which gives us a lot of confidence that we will be covered if we choose to expand our use of Google Cloud services beyond BigQuery."

- Sam King, Engineering Manager, Data Platform Engineering Team, Nuna

Google Cloud

With this framework in mind, an effective data governance strategy and operating model provides a path for organizations to establish control and maintain visibility into their data assets, providing a competitive advantage over their peers. Organizations will likely reap immense benefits as they promote a data-driven culture within their organizations, specifically:

1. **Improved decision making:**
   Better data discovery means that users can find the data they need, when they need it, which makes them more efficient. Data-driven decision making plays a huge role in improving business planning within an organization.

2. **Better risk management:**
   A good data governance operating model helps organizations reduce the risks of fine, increase customer trust, and improve operations. When organizations can audit their processes more easily, they reduce their risk and improve their operations. Downtime can be minimized while productivity still grows.

3. **Regulatory compliance:**
   Increasing governmental regulation has made it even more important for organizations to establish data governance practices. With a good data governance framework, organizations can embrace the changing regulatory environment instead of simply reacting to it.

As you migrate more of your data to the cloud, data governance provides a level of protection against data misuse. At the same time, auditable compliance with defined data policies helps demonstrate to your customers that you protect their private information, alleviating their concerns about information risks.

To learn more about how we protect and govern your data on GCP, visit:

https://cloud.google.com/big-data/security-governance/
