Data Governance: The Definitive Guide

- As adoption of cloud computing continues to grow, information management stakeholders have questions about the potential risks involved in managing their data in the cloud

- Three risk factors:
    - **Securing the data**
        - 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
    - **Regulations and compliance**
        - Concerns about oversight and control of data stored in the cloud
    - **Visibility and control**
        - Lack visibility into their own data landscape: which data assets are available where those assets are located and how and if they can be used
- These risk factors clearly highlight the need for increased data assessment, cataloging of metadata, access control management, data quality and information security

Why your business needs data governance in the cloud

- **Risk management**
    - Potential exposure of sensitive information to unauthorized individuals or systems, security breaches
- **Data proliferation**
    - Important to introduce controls and mechanisms to rapidly validate the quality aspects of high-bandwidth data streams
- **Data management**
    - Introduce tools that document data lineage, classification, and metadata to help your employees
- **Discovery**
    - Ability to assess data asset content and sensitivity
- **Privacy and compliance**
    - Auditable and measurable standards and procedures that ensure compliance with internal data policies as well as external government regulations

Framework and best practices for data governance in the cloud

- People, processes and technology can work together to enable auditable compliance with defined and agreed-upon data policies.
    - Data discovery and assessment
    - Data classification and organization
    - Data cataloging and metadata management
    - Data quality management
    - Data access management
    - Auditing
    - Data protection

Operationalizing data Governance in your organization

- 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
- The following steps are critical in planning, launching and supporting a data governance program:
    - Bild the business case
    - Document guiding principles
    - Get management buy-in
    - Develop an operating model
    - Establish a framework for accountability
    - Develop taxonomies and ontologies
    - Assemble the right technology stack
    - Establish education and training

The business benefits of robust data governance

- ✓ Improved decision making
- ✓ Better risk management
- ✓ Regulatory compliance

What is data governance?

- A data management function to ensure the quality, integrity, security and usability of the data collected by an organization.

Enhancing trust in data

- The purpose of data governance is to build trust in data. It requires that a data governance strategy address three key aspects: discoverability, security and accountability

Data governance versus Data enablement and Data security

- **Data governance**
  - Mostly focused on making data accessible, reachable and indexed for searching across the relevant constituents, usually the entire organization's knowledge-worker population
- **Data enablement**
  - Extends into tooling that allows rapid analysis and processing of the data to answer business-related questions
- **Data security**
  - A set of metrics put in place to prevent and block unauthorized access

Methods of data collection have advanced

- The more people working and viewing data, the greater the need for complex systems to manage access, treatment and suage of data because of the greater chance of misuse of the data
- The advent of streaming, however, while greatly increasing the speed to analytics, also carries with it the potential risk of infiltration

Managing discoverability, security and accountability

- Make sure that your data collection is purposeful
- Turn on organizational-level audit logs in your data warehouse
- Conduct periodic security audits of all open ports
- Apply an additional layer of security to sensitive data within documents

Audit logging

- Being able to bring up audit logs for a regulator is useful as evidence that policies are complied with. You cannot present data that has been deleted, but you can show an audit trail of the means by which the data was created, manipulated, shared, accessed and later expired or deleted
- To be useful for data governance purposes, audit logs need to be immutable, write-only and preserved, by themselves, for a lengthy period – as long as the most demanding data preservation policy

- Audit logs need to include information not only about data and data operations by themselves but also about operations that happened around the data management facility
- Policy changes need to be logged, and data schema changes need to be logged. Permission management and permission changes need to be logged, and the logging information should contain not only the subject of the change but also the originator of the action
- The major public cloud providers offer the ability to store your data in accordance with regulations

The enterprise dictionary

- Important to understand how an organization works with data and enables data governance. Usually, there is an enterprise dictionary or an enterprise policy book of some kind
- Normally owned by either the legal department or the data office

Data classification and organization

- Automation of data classification can be accomplished in two main ways:
    - Identify data classes on ingest, triggering a classification job on the addition of data sources
    - Trigger a data-classification job periodically, reviewing samples of your data

Data cataloging and metadata management

- Crucial to metadata management is a data catalog, a tool to manage this metadata. You want a tool that spans multiple storage systems to hold the information about the data. This includes where the data is and what technical information is associated with it, but you also should allow for the attachment of additional "Business" metadata, such as who in the organization owns the data, whether the data is locally generated or externally purchased, whether it relates to production use cases or testing and so on

Data assessment and profiling

- Outliers can be the result of data-entry errors or may just be inconsistent with the rest of the data, but they can also be weak signals or less-represented new segments or patterns.
- In many cases, you will need to normalize the data for the general case before driving insights. This normalization should be done in the context of the business purpose the data is being used for.

- Data engineers are usually responsible for producing a report that contains data outliers and other suspected quality issues.

Data quality

- There should be different confidence levels assigned to different quality datasets. One possible process that can be implemented to improve data quality is a sense of ownership: making sure the business unit responsible for generating the data also owns the quality of that data and does not leave it behind for users downstream.

Workflow management for data acquisition

- This workflow usually begins with an analyst seeking data to perform a task. The analyst, through the power of a well-implemented data governance plan, is able to access the data catalog for the organization and, through a multifaceted search query, is able to review relevant data sources.
- Data acquisition continues with identifying the relevant data source and seeking an access grant to it. The governance controls send the address request to the right authorizing personnel, and access is granted to the relevant data warehouse, enforced through the native controls of that warehouse.

User authorization and access management

- Identity and access management should provide role management for every user, with the capability to flexibly add custom roles that group together meaningful permission relevant to your organization

Legacy

- Every company should have a central data dictionary defining all data names, classes and categories that is standardized and used throughout the company. Many legacy companies lack this central data dictionary because their data is spread out through these various on-prem systems.
- Companies need a framework for how to move their data and have it properly governed from the beginning with the right people working the right process

Data segregation and ownership by line of business

- Each line of business has dedicated people to do the work of governance on just that data. The data steward is the subject matter expert in this line of business; knowing what data resides here, what it means, how it should be categorized/classed, and what data is sensitive and what isn't.

- They also serve as the point of contact between their line of business and the central governing body at their company for staying up to date on compliance and regulations, and they're responsible for ensuring that the data in their line of business is in compliance.

What is data life cycle?

- Organizations work with transactional data as well as with analytical data. Transactional systems are databases that are optimized to run day to day transactional operations. These are fully optimized systems that allow for a high number of concurrent users and transaction types.
- Even though these systems generate data, most are not optimized to run analytics processes. On the other hand, analytical systems are optimized to run analytical processes.
- These database store historical data from various sources including CRM, IOT sensors, logs, transactional data and many more. These systems allow data analysts, business analysts, and even executives to run queries and reports against the data stored in the analytical database.

- **Data creation**
  - The first phase of the data life cycle is the creation or capture of data
- **Data processing**
  - Once data has been captured, it is then processed without yet deriving any value from it for the enterprise. This is done prior to its use.
  - Data processing is also referred to as data maintenance, and this is when data goes through processes such as integration, cleaning, scrubbing or extract-transform-load (ETL) to get it ready for storage and eventual analysis.
- **Data storage**
  - Where both data and metadata are stored on storage systems and devices with the appropriate levels of protection
- **Data usage**
  - Import to understanding how data is consumed within an organization to support the organization's objectives and operations. Data becomes truly useful and empowers the organization to make informed business decisions.
- **Data archiving**
  - Data is removed from all active production environments and copied to another environment
- **Data destruction**
  - Removal of every copy of data from an organization, typically done from an archive storage location

Data life cycle management

- Comprehensive policy-based approach to manage the flow of data throughout its life cycle, from creation to the time when it becomes obsolete and is purged

Data management plan

- Defines how data will be managed, described and stored. In addition, it defines standards you will use and how data will be handled and protected throughout its life cycle.
    - Identify the data to be captured or collected
    - Define how the data will be organized
    - Document a data storage and preservation strategy
    - Define data policies
    - Define roles and responsibilities

Roles and responsibilities

- The success of data governance program depends on the combination of people, processes, and tools all working together to make governance a reality
    - Build the business case
    - Document guiding principles
    - Get management buy-in
    - Develop an operating model
    - Develop a framework for accountability
    - Develop taxonomies and ontologies
    - Assemble the right technology stack
    - Establish education and training

What is data quality?

- When thinking about data quality, it is good to discuss:
    - Accuracy
    - Completeness
    - Timeliness

Data quality in AI/ML models

- The data available for building machine learning models is usually divided into three non-overlapping datasets: training, validation and test.
- The machine learning model is developed using the training dataset. Next, the validation dataset is used to adjust the model parameters so that overfitting is avoided. Last, the test dataset is used to evaluate the model performance.

- Data quality is making sure that the data's accuracy, completeness and timeliness are relevant to the business use case in mind. Different types of business use necessitate different levels of the above, and you should strive to keep a scorecard of your data sources when creating an analytics workload composed of descendants of these data sources.
- When repurposing data for a different analytics workload, revisit the sources and see if they are up to the new business task.

Governance of data in flight

- Data, especially data used for insights via data analytics, is a "living" medium. As data gets collected from multiple sources, it is reshaped, transformed, and molded into various patterns for different use cases
- Data governance should be consistent across these transformations and allow more efficiency and frictionless security.

Data transformations

- It is advantageous to consider the extraction phase as the first step in a pipeline, allowing subsequent steps to operate in batches in parallel to the continued extraction. As data is extracted from the sources, it's useful to perform data validation, making sure the values retrieves are as expected.

Lineage

- Lineage is the recording of the path that data takes as it travels through extract-transform-load, and other movement of data, as new datasets and tables are created, discarded, restored and generally used throughout the data life cycle.

How to collect lineage

- As your data grows it is important to allow more and more automation into the process of lineage collection and to rely less and less on human curation.
- Another way to collect/create lineage information is to connect to the API log for your data warehouse. The API log is expected to contain all SQL jobs and also all programmatic pipelines.
- Row-level lineage allows the expression of information about transactions. Dataset-level lineage allows the expression of coarse information about data sources.

Audit, compliance

- Decisions that are derived from data are done so without manipulation, and that there is an unbroken “chain of trust” between properly acquired data sources and the end-user tools
- With a lineage graph, organizations can trace data variations and life cycle, promoting control and allowing a complete picture of the various systems involved in data collection and manipulation.

Data protection

- The level of protection to be afforded to an asset should reflect the cost and likelihood of a security breach associated with that asset. This requires cataloging the types of security breaches and the costs associated with each breach.

Classification

- Implementing data governance requires being able to profile and classify sensitive data. Once the classification is determined and the protection level chosen by cost analysis, the protection level is implemented through two aspects. The first aspect is the provisioning of access to available assets. This can include determining the data services that will allow data consumers to access the data.
- The second aspect is prevention of unauthorized access. This is done by defining identities, groups, and roles and assigning access rights to each.

Identity and access management (IAM)

- Access control encompasses authentication, authorization and auditing. Authentication determines who you are, authorization determines what you can do, and auditing logs record what you did.
- The data protection governance process should also create policies on when separate network designs are necessary, how portable devices have to be treated, when to delete data, and what to do in the case of a data breach.
- Finally, data protection needs to be agile, because new threats and attack vectors continue to materialize. So the entire set of policies should be periodically revisited and fine-tuned.

Monitoring

- Monitoring governance involves capturing and measuring the value generated from data governance initiatives, compliance and exceptions to defined policies and procedures and finally, enabling transparency and auditability into datasets across their lifecycle.

- **Data quality monitoring**
  - Because of how important data quality is, it should be monitored proactively, and compliance exceptions should be identified and flagged in real time
    - Completeness
    - Accuracy
    - Duplication
    - Conformity
  - Establishing a baseline
    - Establish a baseline of the current state of data quality. This will help you identify where quality is failing
  - Quality signals
    - Verify data fields for completeness, accuracy, duplicates, conformity, statistical anomalies and more. When data quality falls below a specified threshold, an alert would be triggered
- **Data lineage monitoring**
  - The natural life cycle of data is that it is generated/created by multiple different sources and then undergoes various transformations
  - Monitoring lineage is important to ensure data integrity, quality, usability and the security of the resulting analysis and dashboards.
- **Compliance monitoring**
  - Requires that changes are made according to the new information gathered in order to stay in compliance.
  - To stay in compliance requires auditing and tracking access to data and resources within the organization.
- **Program performance monitoring**
  - Track progress against the program's aims and objectives, ensuring the governance program delivers right outcomes for the organization, accounting for efficient and effective use of funding, and identifying improvement opportunities to continue creating impact for the business.
- **Security monitoring**
  - The process of collecting and analyzing information to detect suspicious behavior, or unauthorized system changes on the network in order to take action on alerts as needed.

Monitoring criteria

- Monitoring systems collect data in two distinct ways: passive systems, where the tools observe data created by the application and system under normal conditions
- Active systems, which are more proactive, use agents and other tools to capture data through a monitoring module, and are often integrated within production systems.

Data culture: what it is and why it's important

- The data culture is the set of values, goals, attitudes and practices around data, data collection, and data handling within a company or organization.