CLOUDERA

How to Implement Data Governance Successfully

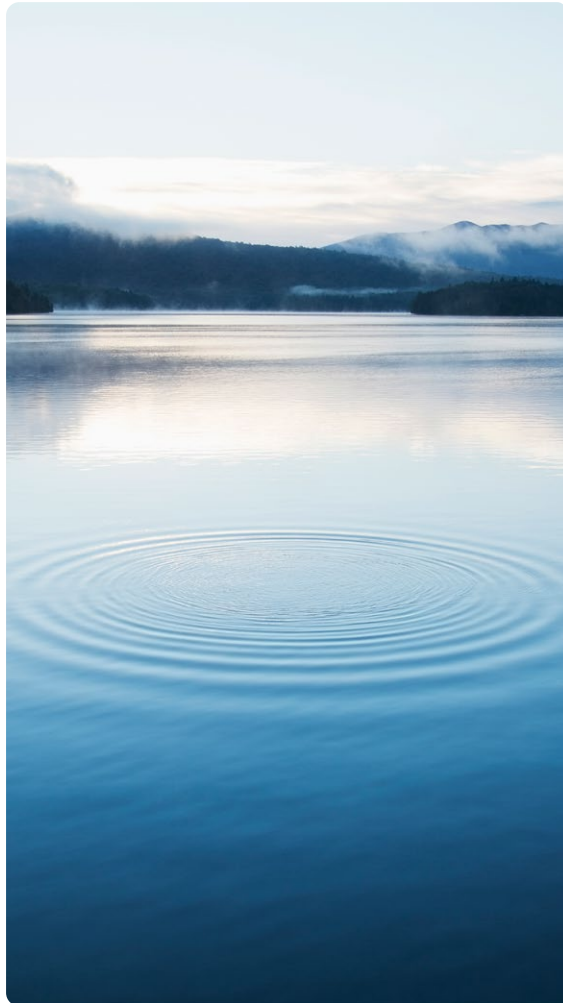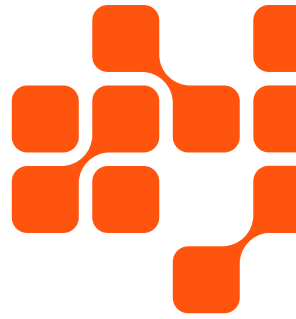Your data's been getting
a little out of hand recently.

Table of Contents

CLOUDERA

How to Implement a Successful Data Governances Solution

Your data's been getting a little out of hand recently.

It's never around when you want it, it doesn't listen to simple instructions, and it stays out too late partying and drives you insane with worry. (Well, maybe not that.)

It's time to set down some rules.

Enter… data governance!

Data governance refers to the rules and roles that control how data is managed within your organization. The goal of these rules and roles is to ensure that your data can be used properly (e.g. security, access, compliance) and effectively (e.g. quality, operability).

Data governance pays big dividends, but it's a effort to implement. There are plenty of moving parts in a data governance project, and any one of them can throw the project off-track. The following guide is designed to walk you through the process of implementing a data governance project, pointing out suggestions to consider and pitfalls to avoid in order to achieve success.

CLOUDERA

Big-Picture Planning of Your Governance

The vast majority of data governance success is decided in the planning phases. Plan the project well, and even if you slip a little bit in the execution, you'll probably have a decent outcome. Skip over or skimp on your planning, and even best-in-class execution will deliver only sub-par results.

Now that we've put some fear into you, let's get to the practical planning advice.

Start small and focused

An initial data governance implementation should be treated like a pilot project. Choose the datasets from one line of business, one department, one sub-section within a department… the smaller the implementation unit involved the better. That is, as long as **the data within the unit is significant, relevant and currently active**, and it touches:

- At least one IT infrastructure person
- At least one data team member
- At least one subject matter expert
- At least two business users

(It is possible that one individual might fill more than one of the above roles.)

Pick a lead for data governance implementation

The next time you're asked at your company Happy Hour to write a poem while inebriated (What? They don't do that at your company?), just pull this one out of your pocket:

- To succeed
- You need
- A lead.

In this eBook we'll examine the challenges of getting AI applications off the ground, and ways to overcome those challenges using a data platform built on AWS's public cloud resources.

Beautiful, right? And so, so true.

Before you start implementing data governance at your organization, think long and hard about the right person to lead governance. The right person should:

- Be a data governance evangelist, willing to invest time in learning and thinking about how to give governance the best chance of success at your company.

- Possess charisma and/or strong communication skills.

- Have good working relationships with employees and teams across the business.

CLOUDΞRA

Why is all this necessary?

Well, unless your entire employee base – technical and business users alike – has been picketing for data governance, you need the person in charge to not only be excited about the potential of data governance, and not only be capable of guiding the technical implementation, but to also be able to effectively market and sell data governance to your employees.

Data governance is often perceived by employees as an annoying set of policies that slow business processes and hamper access to needed data. Your data governance implementation leader needs to convince users at large that the exact opposite is true: that data governance will **improve** and **simplify** data access. Otherwise your data governance strategy will fall flat for lack of buy-in and cooperation.

Support your data governance manager with a cross-functional team

Got your data governance project lead picked out? Fantastic.

Now it's time to construct the team that will support the lead in the key areas of implementation. While the exact composition of the team may vary by case, in general it should include:

- Someone with strong technical implementation skills (to lead on implementing the data management solutions that will be a part of the project)

- A subject matter expert on the datasets (or a subsection of the datasets) included in the project

A word to the wise

While usually one chooses the project focus area/department and then picks the most relevant team members from that area, sometimes it may make sense to work the other way, and pick an area/department for project implementation based on the availability of good team candidates.

CLOUDERA

Be prepared to measure success

Ever heard the apocryphal tale of the acclaimed archer who, when asked how he always managed to hit the bullseye, responded: "It's easy! First, I shoot the arrow, then I draw the circle around it!"?

Don't let your data governance project "success" be dependent on drawing the target after you shoot. If you want to really know whether your data governance implementation is succeeding in its objectives, you have to define those objectives during the planning phase.

Possible objectives include:

- Improve data accuracy

- Increase data accessibility

- Centralize data management

- Increase data sharing across the organization

- Improve communication about data

Once you've established objectives, don't stop there! Objectives are great, but they're usually fuzzy and hard to measure. So the next step is establishing a KPI for every objective. To see this in action, check out these examples from the data governance framework of the US Department of Transportation:

NO.	OBJECTIVE	PERFORMANCE MEASURE
1	Data meets established need-driven standards for accuracy	% of data that meet standards
2	Meaning of data is readily understood	% of data elements for which clear definitions have been established
3	Communicate data-related changes to all Data Users	Affected users are aware of data-related changes at least 30 days before changes are implemented

Whatever KPIs you pick, make sure to measure and document them before you start your data governance implementation, so that you have a benchmark against which to compare future measurements.

CLOUDERA
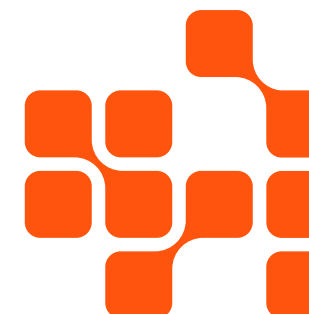
Select the right supporting tools and solutions

The right tools will be critical to your long-term data governance success. But don't be fooled; no one solution, full-featured though it might be, will be enough to manage your entire data governance project.

The key to finding the right solutions is to identify the domains involved in your data governance implementation and outline your requirements for each domain. Then, when you review tool candidates, select two or three solutions that together will cover your data governance domain requirements and can smoothly integrate with each other.

Relevant domains include:

Data Quality is Paramount

Maintaining data quality is a key goal of data governance. The important decisions made by management and business users must be based on accurate information. The multiple aspects of data quality maintenance include data profiling, data pipeline monitoring, assessment of the original source of the data and the transformations the data underwent on its subsequent journey through your systems (an aspect often covered by data lineage), and data cleansing.

Workflow/Access

A major objective of data governance can be summed up as: people should be able to access the information they should have, and should not be able to access the information they shouldn't have.

The data workflow/access domain of data governance addresses the first half of the above: people being able to access information they should have.

When a user wants access to data that they need for their job, that data should be immediately available in its most complete, accurate and updated state. The user should not run into daunting or time-consuming obstacles regarding permissions or approval processes.

Ideal workflow solutions should make work flow: smoothly, painlessly and effectively.

Security

The security domain of data governance addresses the second half of the above data governance objective: people should not be able to access information they shouldn't have.

While your initial data governance implementation may be on a small scale, bear in mind that for the future, when you're dealing with large quantities of data, you will likely need an automated method to classify your data and identify the appropriate access level – or at least to do the initial work, even if a human needs to check and approve the classification.

CLOUDERA

Compliance/Privacy

No matter what industry you are in, from finance to healthcare to insurance, you most likely have regulations you need to comply with. A large part of data governance is setting policies to ensure compliance – and then proving you did, in fact, comply. Look for solutions that will make this process easier: the setting of policies, the maintenance of policies and the proof that you kept to the policy.

In almost every field, sensitive and personal data need to be treated with special care. You will want a data governance solution to provide easy and accurate ways of identifying this data, applying the requisite data masking or access policies and retrieving proof that you did so.

Data Democritization/User Empowerment/Data Discovery

Lots of big words to communicate a simple idea: data users should be able to find what they need on their own.

Putting that power into users' hands often involves a data catalog, which serves as the one-stop, single source of truth for data discovery and investigation. Data governance platforms usually have a data catalog incorporated, and part of your data governance setup will be populating that catalog with all the metadata you are planning to govern.

That process, when done manually for all your enterprise data, usually takes months upon months of work. (Your initial project may take a shorter time, but think ahead to when you expand data governance to your entire organization.) Automation becomes a big deal here, both in the setup (a built-in automated data lineage tool, for example, can empower you to set up the entire data catalog on your lunch break) and in the maintenance of the catalog (periodic refreshes will update your data catalog and governance platform with any recent metadata changes or additions).

A word to the wise

In general, when you review tools, keep an eye out for automation capabilities. The more the solution offers automated scanning, mapping, profiling, policy application, the more potential it holds to lighten the load on your team members. Make sure, however, that you check the quality of the automation and how successful it is at delivering accurate results.

CLOUDΞRA

Dive into Your Data

You just completed the big picture planning of your data governance initiatives .Congratulations! Now it's time to zoom in on the individual data systems in question so you can develop an applied, specific data governance implementation plan.

Understand how your data moves

Ever babysat as a teenager? When was it an easier job – when the kids were awake or when they were sleeping?

It's much harder to keep track of entities on the move. And when it comes to enterprise business systems, data is constantly on the move.

Keeping track becomes easier, however, when you understand how the data moves. What data is important to this report? Where does that data come from? How does it end up in this report? What transformations does it undergo along the way?

An automated data lineage tool like Cloudera Octopai Data Lineage, with the ability to deliver intuitive visualizations of your data's path through your data landscape, can be gold over here.

Once data movement becomes predictable, it becomes governable.

Understand how your data is used through lineage

So now you know where your data goes. But who interacts with it? Who needs access to it?

Data use mapping and lineage is critical for the success of a data governance project. Different types and levels of data will be used by different types and levels of users. The users who will need access to regulatory data, for example, are usually fewer and higher-level than the users who will need access to operational data. And the users who need to directly interact with and be able to modify regulatory data are even fewer than the users who just need to view regulatory data.

This is an area where your cross-functional team shows its value: in being able to accurately define the way data is used by the users in their domains.

Do an initial data cleansing

Diving into data is no fun if the lake is polluted. Yes, making your data stores as sparkling as a pristine mountain spring may be one of the end goals of data governance and justifiably far off. But you'll have a much easier time implementing data governance if you first get rid of the obvious garbage.

Redundant datasets? Dump 'em. Outdated reports? Ditto.

How do you tell what's garbage and what's not? Take the visualizations created by a comprehensive data lineage mapping solution and review them with your cross-functional team. Identify which data assets are no longer relevant and can be tossed. It'll lighten the load on your data governance.

CLOUDERA
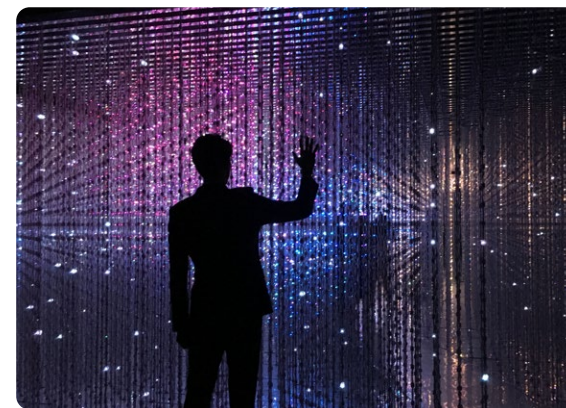
Creating the Data Governance Framework

You're finally here! After all that preparation, you actually get to roll up your sleeves and plunge right into all those exciting policies, standards and rules – yippee! (Well, maybe. Your level of thrill may vary.)

Set out policies for what data and its use should look like

Data governance policies are the high-level, bird's-eye view descriptions of what data use should look like in your organization.

Here are some representative examples from the data governance framework of the US Department of Transportation:

NO.	OBJECTIVE	PERFORMANCE MEASURE
1	FHWA data must be consistent	All strategic FHWA data shall be modeled, named, and defined consistently, according to standards, across the organization. Efforts must be made by management to share data and not maintain redundant data without justification. Originating business stewards of data must recognize the informational needs of downstream processes and business units that may require FHWA data
2	FHWA data must be of acceptable quality	Quality data are critical to ensuring FHWA mission success. Data Stewards are responsible for ensuring that FHWA data are accurate and correct for the intended purpose and use, and that data providers follow all reporting requirements regarding the collection, processing, and reporting of FHWA data, and meet all requirements of the Data Quality Act. Data quality standards shall be managed and applied actively to the approved reliability levels of FHWA data as defined by the business owners

Make sure that your policies are in language understandable by the lay business user. If only a data scientist can decipher their meaning, that does not spell success for your data governance adoption. Try again.

CLOUDERA
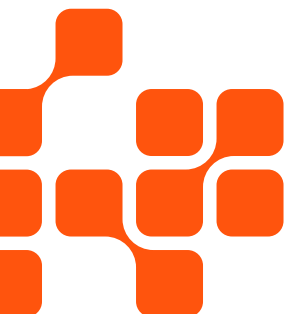
After the establishment of policies comes their practical application: standards, procedures and rules.

Create standards and procedures for naming and storing data

If you live in the United States and your new best friend from Europe messages you that she's turning 30 on 4/3/2023, don't plan a surprise party for the 3rd of April. You'll be a month late.

On a personal level, American/European date notation mix-ups are just fodder for jokes and misunderstandings (or a stomachache resulting from misinterpretation of an expiration date). But if you have inconsistencies with date notation within your enterprise database, things can go very, very wrong for very, very many entities.

Keeping to established data naming and storage conventions keep all parts of your data ecosystem playing nicely together. As such, one practical application of your data governance policies is establishing those conventions.

These conventions and standards can be general, for different data categories:

NO.	INFORMATION CHARACTERISTICS	CHARACTERISTICS DESCRIPTION
1	Names and Attributes	The variable names and associated attributes must be unique across all systems. The names may be static, or determined during system execution runtime
2	Container Format	The FHWA content data must be accurately documented to reflect the expected character types, formats, fields min/max lengths and all other format specific characteristics
3	Content Length	All uniquely defined variables must specify reasonable data length. All mapping variables should in turn conform to the specific content length

Or delve more specifically into any one of those data categories, like "Names":

NO.	INFORMATION CHARACTERISTICS	CHARACTERISTICS DESCRIPTION
1	Names should use consistent capitalization	Within a context, names should follow a consistent capitalization convention (e.g., **camelCase, PascalCase, infix_underscores**)
2	A name should use common abbreviations	A name should use commonly used abbreviations instead of full meanings. For example, use **'ID'** instead of **'Identifier'**
3	A name should use singular forms instead of plural	A noun used as a term in a name should be in **singular** form (e.g., '**Vehicle Depreciation Rate**') unless the concept itself is plural (e.g., '**Tool Suite Total License Cost**')

(from the Data Naming Quick Reference Guide of the US Government's Centers for Medicare and Medicaid Services)

CLOUDERA

Defining rules for using and sharing data

Who can see data in this field/row/column/table/category? Who can use it? Who can modify it? Who can share it?

Are there any conditions that need to be met for viewing, using or sharing of data? For example, is it only available up until a certain date, or only if certain agreements were signed and recorded?

The Data Sharing Framework created by the The Knowledge Base on Connected and Automated Driving (CAD) makes their conditions for sharing data very clear:

The recommendations are developed to provide guidance what to include in the agreements to facilitate data sharing within and after the project collecting the data. These recommendations also apply to automation pilots, there are no significant differences regarding this topic between former FOT and current automation pilots.

01 **Funding agreement including the description of the work**

02 **Consortium Agreement**

03 **Participant agreements including consent forms**

04 **External data provider agreements**

Assigning roles and responsibilities for carrying out said rules and procedures

If the rules and procedures for data use in your organization are clearly explained, you're halfway there. But the other half of the journey can often be more challenging: getting people to keep to those rules and procedures, and not forget about or ignore them. It's a rare society whose laws don't need enforcement.

Who will be responsible for making sure consent forms are present before sharing data, for example, or for setting up the data management system that will keep track of such things?

Who will check that newly ingested data assets meet data quality standards – or be notified when an automated check fails?

If a data consumer can't access the data they need, to whom are they supposed to turn to resolve the matter?

All these roles (and more) need to be defined and assigned.

The ultimate goal, of course, is for all your data citizens to assimilate data governance rules and procedures so that they become second nature, and thus require minimal enforcement. One of the primary ways to achieve this is to integrate data governance procedures into regular data usage routines. Just like many of us have integrated routines for life activities, like "Use a dish; wash it after" or "Come into the house; take off your shoes," the ideal for a data-governing organization is to integrate routines like "Create a data asset; note it in the catalog."

CLOUD:RA

Implementation

You've prepped the policies and planned the procedures. Enough theory – time to let 'er rip!

But before you release your data governance project aspirations out into the wild, make sure that you are emotionally and practically prepared to…

Learn from everything

Yes, you've put in an incredible amount of effort into making this data governance project turn out well, but you can expect that some things will go wrong. And that's not a bad thing, because it's an opportunity to learn for the future.

Any time something goes awry in your data governance implementation, it should become the focus of all eyes (especially the data governance project coordinator) and the question: What went wrong and how can we fix it for next time?

When your project is small and focused enough, it's realistic to learn from every single adverse event and get (reasonably) close to perfection. When you then expand to other areas of the business, since most of the big blunders in the mechanics of how to implement data governance will already have been made, analyzed and corrected for, your implementation process will be infinitely smoother and more efficient.

CLOUDΞRA

Data Governance Success, Here You Come!

It's 10 o'clock. Do you know where your data is?

After successful data governance implementation, you will - but you'll also know and have much, much more.

You will have high-quality data you can rely on for mission-critical decisions.

You will have empowered data consumers who can access what they need, when they need it, seamlessly.

You will have secure, compliant data systems that can easily prepare for and pass any audit.

Cloudera Octopai Data Lineage's data intelligence platform is an invaluable support in any organization's initiation and maintenance of data governance. Automated end-to-end data lineage that can connect to, map and visualize any legacy or cloud-based data system within the context of your entire data landscape. A self-creating, self-updating data catalog that puts every data creator and consumer in your organization on the same page and provides a built-in platform for communication about and collaboration on data.

Get your data tools in hand. You're ready to tackle data governance.
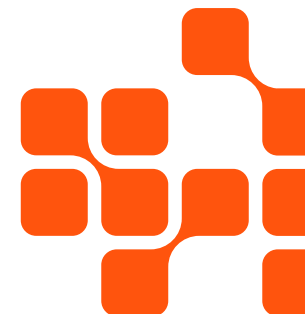
CLOUDERA

About Cloudera

Cloudera is the only true hybrid platform for data, analytics, and AI. With 100× more data under management than other cloud-only vendors, Cloudera empowers global enterprises to transform data of all types, on any public or private cloud, into valuable, trusted insights. Our open data lakehouse delivers scalable and secure data management with portable cloud-native analytics, enabling customers to bring GenAI models to their data while maintaining privacy and ensuring responsible, reliable AI deployments. The world's largest brands in financial services, insurance, media, manufacturing, and government rely on Cloudera to be able to use their data to solve the impossible—today and in the future.

To learn more, visit **Cloudera.com** and follow us on **LinkedIn** and **X**. Cloudera and associated marks are trademarks or registered trademarks of Cloudera, Inc. All other company and product names may be trademarks of their respective owners.

CLOUDERA   Cloudera, Inc. | 5470 Great America Pkwy, Santa Clara, CA 95054 USA | cloudera.com