WHITE PAPER

Datameer
stay curious

Governance Best Practices
With Datameer

1. Introduction

In the world of big data and analytics, “governance” has become a buzzword. The notion of governing data is laudable. Data is strategic, important and if mishandled, potentially compromising. It certainly needs to be protected.

Your data needs a custodial layer to make that happen and governance would seem to provide that. Its name makes that almost self-explanatory.

But although governance does encompass that layer, it extends well beyond it. In its best implementations, data governance does more than establish a defensive regime around data. It creates an environment that makes data highly available, trustworthy and easily discoverable. In general, good data governance entices people in the organization to explore, query and contribute data, and it supports efforts around digitalization and promoting data-driven practices.

So while data governance definitely involves tactical necessities, it also underlies strategic aspirations. Understanding the details and definition of data governance is crucial to the competency of a data-driven culture. Understanding its potential can be the launchpad for data-driven success and data-enabled business breakthroughs.

This may sound a bit lofty, so let’s bring things down to earth. In this paper we’ll do that by covering:

- The role of governance
- What governance encompasses
- The governance features of Datameer
- Requirements
- Approaches
- Reference architectures
- And other considerations

2. Role of Governance

Governance, with respect to any data is of paramount importance. But because of the unique facets of a data lake's workflow, governance is even more important in the realm of big data than it is in the context of OLTP, data warehousing and BI.

Governance provides a custodial role in the trustworthiness of data. It provides consumers of data with a level of assurance that they're getting data that's authoritative, clean and with pedigree that can be traced and validated. A good governance regime also ensures that any given consumer of the data, or groups thereof, have intuitive access to all the data for which they're authorized and none of the data for which they're not.

That's a double-edged mission. It means that governance is responsible for protective measures like quality control and protection but also for proactive measures like maximizing the data lake's usability. Essentially, governance systems must play the role of both bad cop and good cop.

3. What Does Governance Encompass?

Governance provides for data quality, veracity, assurances of recentness, security and access policy. To be more specific though, let's walk through each of the key areas of governance.

3.1 Lineage and Impact Analysis

Lineage entails the ability to show, for any given data set, where each column in it emanated from. That can be from another data set or a fully original data source. In the former case, the sources should be shown, identified and known, whether they're flat files, feeds or an external database.

Impact analysis works in the opposite direction. Beyond merely assessing the upstream sources used to create a data set, impact analysis lets you determine the downstream data sets, analyses, reports and dashboards that use a given data set and would be impacted should that dataset be compromised or its structure modified.

3.2 Audit

Beyond showing the genesis of data sets through lineage, governance systems must also provide information about how and when the events in a data set's lineage took place and who made them. If certain data is modified to the extent that it falls out of compliance, a governance system must provide the forensic information necessary to identify the party responsible and hopefully, address the discrepancy. Essentially, governance systems must assure accountability.

The information necessary to do this will be part of an audit trail that a governance system must maintain. Any significant data event that takes place in the system must be logged; the system must capture the "before" and "after" states of the data, meta data or permissions information, the time and date of the event, and the user who effected the change. Minimally, this information must be kept in log files, whose own integrity is assured. Often this information will also be available, on a push or pull basis, through application programming interfaces (APIs).

3.3 Security

Security includes administration of role-based access controls in which individual users, roles or groups composed of users have defined access to particular subsets of the data set. The subset of data that a user, role or group has access to can be defined as an explicit set of rows (identified by primary keys) or as an expression or query that defines a group of rows meeting a certain condition. Examples include:

- Customers in a specific geographical area
- Order line items for certain categories of product or services
- Data for employees reporting to a specific manager

The interesting thing about a permissions regime is that it's about much more than making rules, letting people in and keeping them out. In fact, a permissions regime really underlies an organization's strategy for its data lake in terms of how broadly it intends data to be shared, and across what scope or audience to solicit data contributions. Departmental, enterprise and business-to-business scenarios will prescribe vastly different group and role definitions, and differences in their memberships as well. We explore this in great detail in the "Approaches to Governance" section of this paper.

But security includes more than mere permissions; it includes protection of the data in the form of encryption. This can include ensuring that data on disk is stored in encrypted form, and that data is also encrypted as it's transferred from point to point.

3.4 Data Quality

Data quality encompasses the various tasks and checks around ensuring the hygiene and currency of data. Ensuring duplicate rows are properly merged, erroneous data removed and accuracy enforced are key components here. Issues with carriage returns and linefeeds need to be resolved, as do blank rows and unexpected changes in schema, e.g. missing columns, or extraneous ones.

Data must be clean and accurate. If it's not, business users will lose trust in data sets that prove to be defective and may well lose confidence in the entire data lake, because the un-remediated lapse of integrity in any one data set calls the integrity of all the data into question.

Data quality is not defined solely by general criteria of cleanliness and consistency. For many organizations, the data must conform to certain business rules as well. Think of the validation rules that are enforced inside line-of-business software applications or even forms on the Web. Just as those applications won't respond to the Submit button until your data meets certain conditions, data governance must assure that data is subject to rule-based standards before it's ingested. For example, columns containing invalid values — or required columns that have no values — may need to be remediated, or dropped entirely.

3.5 Compliance

Regulatory regimes may apply to particular customers' business practices and such companies will usually need to comply with data retention, security and auditability standards accordingly. While this may be more of an imperative for transactional data, it can still be important to analytics, as certain analytical queries may be part of compliance checks.

We have already discussed security, lineage and audit controls. Data retention is something we haven't addressed, but the concept is relatively simple and corresponds to auditability. Essentially, any time an analysis or other activity is performed that would overwrite a dataset, the prior state of the data set may need to be maintained, either for a certain duration of time, or for a certain number of such overwriting events. For example, an older state of the data set may need to be retained for 90 days, or else the last n generations of a data set must be retained on an ongoing basis, with any given generation being dropped only when another overwrite event takes place.

3.6 Certification

One mitigating factor in the area of data quality, and allowing for less-vetted data sets in to the data lake, is that of certification. Certifying particular data sets provides explicit approval or endorsement of them, safeguarding the trust factor already discussed. Having an explicit certification of certain data sets, rather than implied certification of all data sets, allows a data lake to function in a more self-service fashion while still maintaining integrity and trust. With certification, users are empowered to explore both fully vetted data sets as well as those which, though they haven't yet undergone rigorous validation, are nonetheless useful.

The very definition of a data lake or, at least, the definition of what distinguishes it from a data warehouse, is that it can include data sets that may appear to be of ancillary importance and drawn from sources that may be peripheral to an enterprise's systems of record. In other words, data lakes err on the side of and promote contributed data sets, including those from consumers of data.

3.7 Master Data Management

Master data management (sometimes called MDM, although that acronym is also used to for the term Meta Data Management, so we avoid the acronym altogether here) pertains to the maintenance of certain reference data that is used repeatedly in different applications. Data such as customer records, product catalog items, geographical entities or territories are typically used across a number of transactional applications. In the context of analytics, such data is often used as dimensions, i.e. for the purposes of aggregation and drill down, or for filtering.

Maintaining a single authoritative repository of such data is the province of Master Data Management and is a problem space often addressed by dedicated products. It is also, however, a governance function, so we mention it here.

3.8 Data Cataloging

Data cataloging involves the curation of, documenting, tagging and facilitating the general discoverability of data. Although data cataloging is distinct from governance, many products handle both and sometimes the two are conflated. We're mentioning it here for completeness.

4. Key Governance Features in Datameer

4.1 Authentication

Datameer provides an internal user management and authentication system, or enables remote authentication via integration with LDAP or Active Directory and the ability to leverage SAML-based single sign-on (SSO). Using remote authentication helps create a seamless integration with authorization and permissions of the underlying Hadoop cluster (see Secure Impersonation).

4.2 Secure Impersonation With Kerberos

Secure impersonation integrates Datameer security and permissions with that of your Hadoop cluster to leverage the underlying security setup. This is an important feature to leverage for a confident, secure environment for data in Datameer and the execution of analytic jobs.

Secured impersonation ensures jobs run as, and create data belonging to, the authorized Datameer user or group. It also ensures these permissions and audit trail are captured in all Hadoop ecosystem components like HDFS and YARN.

4.3 Roles, Access Control and Permissions

Role-based access control allows IT to control which users can perform certain tasks throughout the Datameer application. For example, you can give bulk ingest abilities to data management staff only, while still allowing analysts to upload their own files on an ad hoc basis.

Permissions and sharing mean all Datameer artifacts, including imported data, export jobs, workbooks and infographics can be shared with team members or groups, while at the same time securing these same items from unauthorized used by others. In the Datameer browser UI, individuals only see items for which they have proper permissions.

Datameer also integrates with Apache Sentry for centralized security policy management of data across the entire Hadoop cluster including Hive, Impala, HDFS and Datameer.

4.4 Obfuscation

Upon ingest of data, administrators can choose to obfuscate columns to anonymize this data from analysts or users consuming the data downstream. This helps ensure certain fields representing private or personal data cannot be seen by unauthorized users. Administrators can also create different data connections representing the same core data set where various fields are obfuscated or not, depending upon the needs and permissions of the users.

4.5 Lineage

The Lineage feature provides full tracking of dependencies across all artifacts inside Datameer from the time data enters (connections and import jobs), through each transformation and calculation (workbooks), to when data is given to users (infographics and exports). Every aspect is tracked within the Datameer metadata as jobs are run end-to-end. Revision history and security policies applied are also tracked to support audit trails (see below).

While most people associate lineage capabilities with auditing and compliance, it also plays a critical role in placing trust in the accuracy and value of the analysis and underlying data. With transparency playing a larger role in governance, lineage will help provide the needed clarity to your big data analytics.

4.6 Data Management

Datameer contains a robust set of data management features that aid the operational aspect of governance. This includes data retention, partitioning and encryption.

Flexible retention rules allow each imported data set's retention policy to be configured by an individual set of rules to keep data permanently, purge older records or be configured based on the number of runs of ingests or workbook executions. Security rules allow retired data to be either instantly removed, retained until a specified time or manually removed after system administrator approval.

Administrators can configure rules on how to partition data upon ingest into Datameer to help improve the overall performance of jobs. This feature also enables workbooks and export jobs to use only data from specific time slices to focus the execution of the analytic jobs where these are involved.

Datameer can be configured to use HDFS Transparent Encryption service to ensure all data managed within Datameer is encrypted, providing an added level of data security. This is configured to work with your enterprise Key Management System (KMS).

4.7 Auditing (Event Bus)

Datameer provides built-in audit logs that cover all relevant user and system events, including data creation and modification, job executions, authentication and authorization actions and data downloads. These logs can be analyzed in Datameer itself, or by an external system.

In addition, the logs also contain important data about users and their interaction with the system (not just the data). This includes information about groups and roles, their assignments, artifact sharing, logins and failed login attempts, password updates, enabling and disabling of specific users, and more.

The Datameer Event Bus and REST API can be used to send the auditing information to downstream systems which may catalog and consolidate this data for streamlined processing and reporting. In addition, all metadata and lineage information from Datameer can be sent to these systems as well to integrate with an enterprise-wide data governance strategy (discussed later).

5. About Governance

5.1 Drivers Behind Governance

As previously noted, governance covers a number of key aspects of how you want to operate your big data analytics. This includes:

- **Data security** — This is important, but it's not the only aspect of governance. As I look at my big data analytics, I need to define how I lock down my data, provide secure views of the data and ensure the proper access controls are in place, both to the system and to the data.
- **Optimization** — Governance also should help the team optimize their infrastructure to run effectively. Optimizing big data analytics involves creating the right structure and letting team members effectively operate and optimize what they know best.
- **Self-service** — A well-aligned governance strategy will enable the degree of self-service you want to provide. Controls that are too tight will stifle self-service. If they are too loose, risk is introduced.
- **Sharing and Reusability** — With all the data that's involved in big data analytics, it's sharing and reusability that bring greater economies of scale. Governance needs to implement the right structure for findability and the right blend of controls for reuse and sharing.
- **Operationalization** — Governance plays an important role in how your analytics are put to work. This involves a clean structure and process to promote analytics to regularly running jobs. You want to be confident it runs cleanly, produces the right results and is given to the proper business teams.

While sometimes these different aspects may seem to conflict, a good governance strategy will provide the right balance of all these factors that are specific to your organization's needs and strategy.

5.2 What You Need to Govern

In a big data analytics environment, there are many different items you need to govern, ranging from data to the actual analytics to jobs that are generating results. Let's look at what you need to manage.

- **Organization and Structure** — This is an important aspect to keep in mind when developing your governance. How you organize your environment will impact how items are shared and found, secured, optimized and moved through the analytic lifecycle.

- **Incoming Data** — With the volume of data being ingested into a big data analytics environment, how you secure and manage incoming data is critical to your governance efforts. Raw data sets have varying security needs, and each will have differing ingestion performance and partitioning requirements.
- **Workbooks** — With these items, which contain your core logic and resulting data from your data preparation and analysis steps, creating the proper blend of security with sharing with your governance policies is critical to cultivating your big data analytics success.
- **Jobs** — As analytics move from discovery to operationalized, they become what we call “Runbooks” (a series of workbooks which represents a data or analytic pipeline), and are executed as jobs. You not only need to govern these from an execution standpoint, but also from a change management and auditing viewpoint.

As we explore the different approaches to governance in the next sections, we’ll place particular emphasis on how these items are governed across various strategies.

6. Approaches to Governance

6.1 Introduction

With governance, it's important to start with the organization of your system. Your organization will go a long way to define the rest of your governance policies including access control to folders and underlying artifacts, where users will find items of interest and how projects are moved from analysis to operationalized jobs.

In general, there are five approaches to governance in Datameer that mostly differ in terms of their organization:

- **Data-centric** — Focuses on securing and optimizing the management of raw data as it flows into use cases or projects.
- **User-centric** — Similar to the data-centric model, but is focused more on the needs of individual analysts, to help drive ad hoc data discovery.
- **Reusability-centric** — Focuses on organizing and managing artifacts to maximizing sharing and reuse.
- **Department-centric** — Offers to model to different departments and their unique needs.
- **Lifecycle-centric** — A model that focuses on how projects will move through the lifecycle from ad hoc analysis to production jobs.

For each approach, we will discuss how to organize your data and artifacts, enable sharing and apply governance policies.

It is important to keep in mind that you can mix and match these various models. For example, in the department-centric model, you could use a lifecycle model under one department and a user centric model in another.

You can also think about evolving from one model to the next as you mature. For example, start with a data-centric model and grow to a lifecycle model as a more formal operationalization strategy is put in place.

6.2 Data-centric

In the data-centric model, your focus is on managing and controlling access to your raw data to facilitate its proper use in the analytics. This is an easy model to start with to help facilitate ad hoc, team-based analytics.

6.2.1 Organize

Data is the key area to concentrate on when organizing your artifacts in the data-centric model. You want to create a folder structure where all of your data is managed in one place, is locked down and secured. You are then free to organize the remaining items as you please, but we recommend a use-case or project-oriented view.

Create a major folder where you create and manage your enterprise data connections and core data sets. This is also a central place where you can set data management and retention policies on the incoming data. Create a unique folder for each data set to maximize findability and create any unique access control permissions. Data stewards may also create new folders where they work on and create curated data sets here as well.

Then create additional folders to your operational preferences. We typically see folders that are specific for each use case or project. Within the individual folders, create additional ones that help organize artifacts according to the differing aspects of the discovery, QA and productizing process.

6.2.2 Share

In this model, you will have data stewards that help create and manage the data connections, import jobs and raw data files. You should create a special group for these team members with specific roles.

You will then want to create two different groups for each use case, one for analysts that will need access to the raw or curated data sets and will perform the analysis, and a second one for teams that need to see the intermediate and resulting data specific to the use case.

The admins should have proper access to both the data and use case folders so they can run and troubleshoot runbook jobs. You could create a specific admin group per use case, or simply use the default admin group.

6.2.3 Govern

From a governing standpoint, the admins and data stewards should work together to set all the proper policies to ingest the data, potentially create curated data sets, and provide access to the data. This not only covers setting the right permissions, but also involves:

- Creating data connections with their management, retention and partitioning policies
- Applying any obfuscation policies to the imported data sets as needed for that data set

For a specific use case, the analysts should become the stewards, determining the proper permissions for the artifacts. As you define further workbooks, infographics and exports, define or reuse specific roles for access to these artifacts.

6.2.4 Pros and Cons

Pros

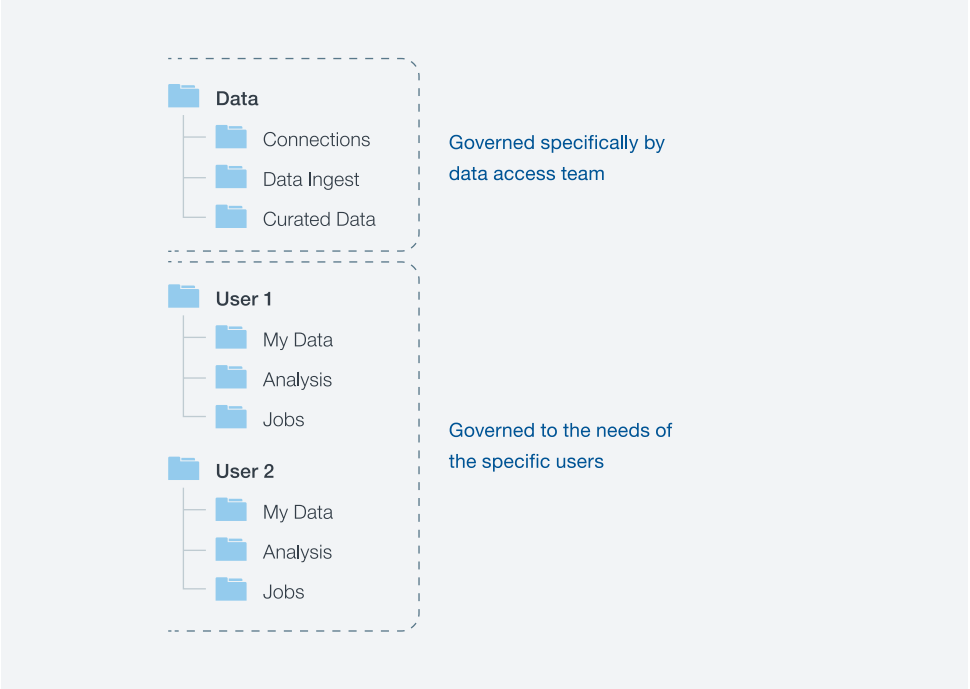
- It's a very easy model to start with and straightforward to implement
- The operationalization emphasis is on the use cases, *not* the data
- You optimize how the data gets into Datameer, and jobs further downstream take advantage of this optimization
- While lineage helps with tracking, this model enables tracing jobs back to the data use for quick troubleshooting

Cons

- This model may be more difficult to expand as you add more data sets
- Your data folders will be heavy with content, requiring their own optimization
- It could eventually create a "spaghetti" links view from the many use cases to the data

6.3 User-centric

The user-centric model is very similar to the data-centric model in that you secure and manage the data from a central data folder structure. It then differs in how the analytics are organized and governed, which makes it a model better suited for ad-hoc analytics driven by individual analysts or users.



6.3.1 Organize

Just as in the data-centric model, create a folder structure where all of your data is managed in one place, is locked down and secured. Follow the same organization advice from the data-centric section for this, create a major folder where you keep data connections and core data sets, and manage them there.

Organize the remaining folders on a per-user basis. In effect, this creates a sandbox for each user to perform their own curation and analytics. They could typically organize their own folders uniquely, but a common model that separates individual data sets from analysis and regularly running jobs would be a logical structure decipherable by admins.

6.3.2 Share

Like in the data-centric model, you should create groups and roles for data stewards that help create and manage the data connections, import jobs and raw data files.

Then create roles for the individuals fitting their specific needs to access data and perform work within their sandbox area of Datameer. From there, analysts are responsible for governance of their resulting workbooks and end results.

6.3.3 Govern

Data stewards and admins will work together to create and govern the data connections, policies and curated data sets. They will then provide the analysts with access to these items.

In some instances, there may be a need to create multiple connections to the same data sets, with different ingestion policies. For example, when ingesting data with some private or personal data, certain fields may need to be obfuscated for some users, while not for others. The data stewards can create the right data sets for the right users by applying needed policies.

As analysts define further workbooks, infographics and exports in the scratchpad folders, they can govern access to these artifacts. Admins can create unique groups and roles for the analysts to use for their projects.

6.3.4 Pros and Cons

Pros

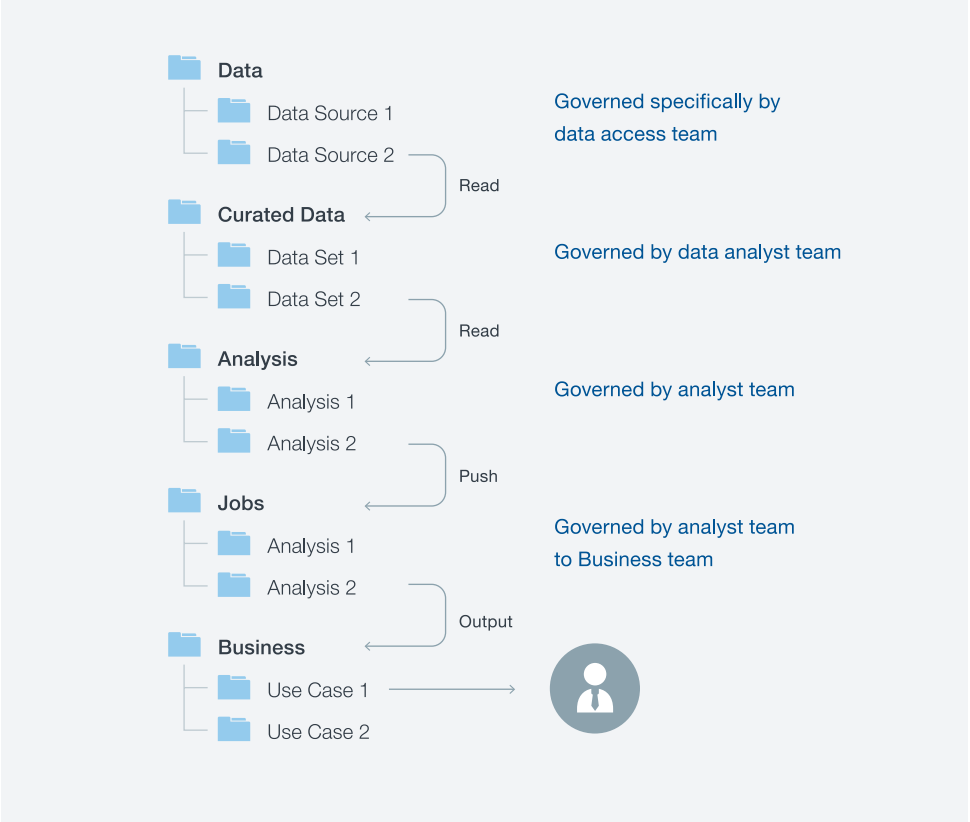
- Scratchpad style, ad hoc approach can facilitate agility
- A very straightforward model pre-built into Datameer
- Facilitates optimization of the data ingestion
- Helps enable personalized data sets based on needs and security

Cons

- Is not operational in nature and could create problems operationalizing jobs
- Sharing is left to the users, who might not have the proper oversight to govern
- Could require very granular controls and roles once more analysis is produced

6.4 Reusability-centric

The reusability-centric governance model focuses on enabling sharing of data and artifacts in the system to drive agility. It has a very open approach that tries to maximize findability and reusability.



6.4.1 Organize

This model has the most open folder model of the four approaches. It applies a general structure that organizes items by their logical uses:

- **Data** — Groups of data connections
- **Curated data** — Specific data sets
- **Analysis** — Specific analysis or use cases
- **Business** — Specific groups of infographics and exports for use cases

This structure enables different team members to know where to look for specific items being shared by others, maximizing findability. It also provides a logical “read upward” model to facilitate access to these folders and underlying items.

6.4.2 Share

Different groups and roles should be created for the various personas including:

- Admins
- Data analysts
- Business analysts
- Business teams

Users could be members of more than one group (e.g. a user could be both a data analyst and business analyst). Also create master roles and groups for the members that need to access all folders.

As needed, create sub-roles and sub-groups for the specific entities underneath each master group. For example, you could create a customer data analyst group that specifically works on curating customer data, and would have the appropriate privileges to do so, and would share with certain business analyst groups — perhaps the customer business analyst team.

6.4.3 Govern

Each group will govern their own policies and access control for their uses with access control applied to each of the folders and sub-folder:

- Admins govern the data connections and raw data sets
- Data analysts are given read access to data connections and raw data as needed and create and manage artifacts in the curated data folders, sharing the data and workbooks with business analysts
- Business analysts are given read access to workbooks with curated data, can create derivatives from those to produce analytic results, then facilitate access to workbooks with analytic results, infographics and export jobs
- Business teams get read access to specific output workbooks and infographics, and can receive exports

Users in each group will only see the artifacts to which they have access, keeping them from accidentally (or intentionally) accessing items to which they don't have privileges.

6.4.4 Pros and Cons

Pros

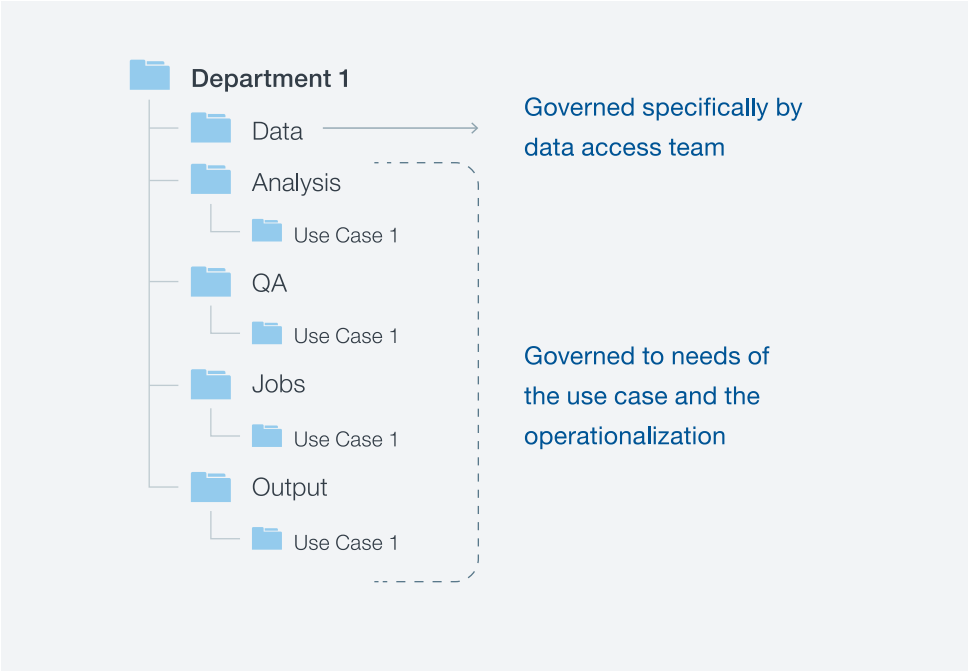
- Offers a straightforward role, group and access control model based on persona
- Focuses the security on the use of the data throughout the information chain
- Facilitates greater reuse and findability of shared objects
- Offers great flexibility
- Enables greater expansion and experimentation at different levels
- Somewhat more decentralized control — each level determines access

Cons

- While trying to facilitate sharing, it's controlled by individual analysts who could be somewhat restrictive or cautious
- This might inhibit analysts from creating projects that have a complete, end-to-end analytic cycle
- Will require constant diligence and monitoring to ensure proper security is in place and data is not shared too much
- Decentralized control might be a problem in compliance-type environments

6.5 Department-centric

The department-centric model focuses governance efforts around the needs of specific departments, and potentially individual business processes within those departments. This approach works effectively if there are analyst teams dedicated to specific departments or business units, and analytics can be logically segregated into those departments. It also makes easy to apply specific controls and organization to those departmental needs.



6.5.1 Organize

This model has a very straightforward organization structure with top level folders for each department or business unit. This helps ensure only department teams have access and also know where to find their items.

The next level, underneath each department can be organized based on the earlier three models, by use case or to facilitate the operationalization (see lifecycle model next). Departments could be organized differently at the next level depending upon their analytic needs: ad hoc, use case driven, or operationalization. Data should be centrally organized for each department.

6.5.2 Share

Groups and roles could vary based on how data is managed within the organization — centrally or distributed. If data is managed centrally, a central data analyst group should be created to manage the data. If each department manages their own data, then each department would have their own data analyst group with appropriate roles.

A central administration group should be created to help manage the operation of the environment across all departments. This helps ensure the proper execution of jobs and allocation of resources.

The remaining roles and groups would be created based on the model chosen to implement in the various departments. Each group would be specific to that department.

6.5.3 Govern

Access and governance policies should be defined and applied similarly to other approaches, but specific to the individual departments:

- Admins and data analysts manage the data connections and access
- Business analysts can read from the data connections and create downstream workbooks, intermediate and resulting data sets from these. They then grant access to their resulting data and infographics to the business teams
- The business teams get read access to the data, infographics and exports

6.5.4 Pros and Cons

Pros

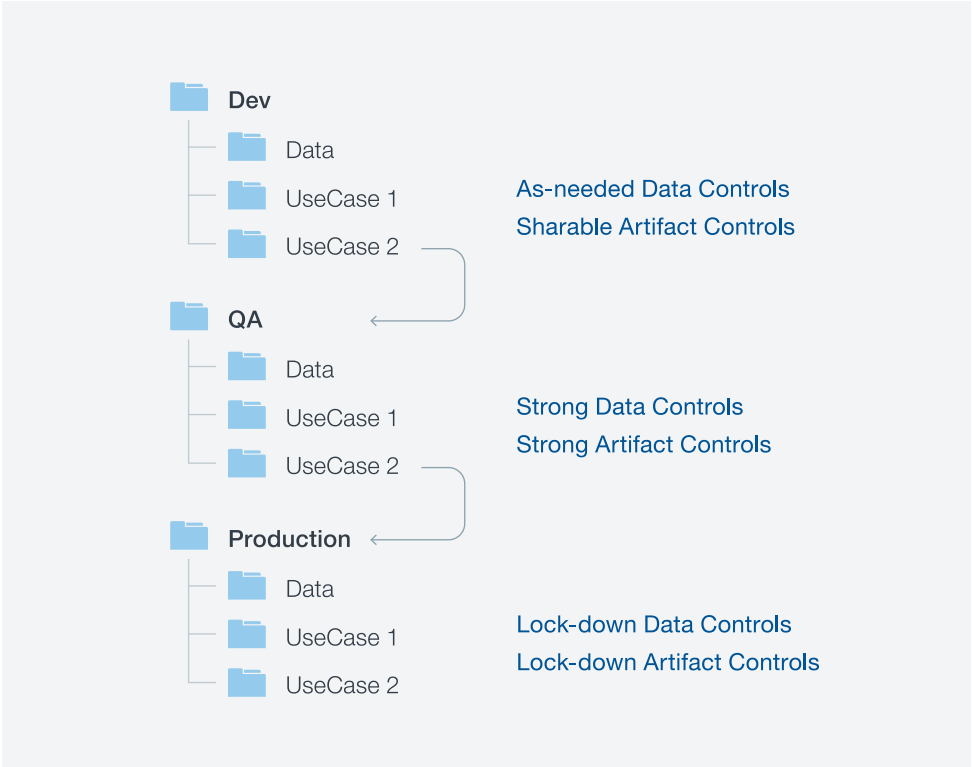
- Facilitates the ability to lock down specific data sets and their use to departments
- Enables the ability to apply resources to departments based on a cost model
- Very easy model to start with and expand to each new department as the use cases expand
- If specific departments have security and compliance needs, these are easily applied here

Cons

- This may inhibit cross-department sharing, longer term
- Groups and roles could get complicated over time as more departments are added

6.6 Lifecycle-centric

The lifecycle-centric model applies governance based on how analytics evolve through a "development lifecycle" from discovery to operationalization. It follows a model similar to typical software development.



6.6.1 Organize

The organizational structure is basic based on the steps in each process – development, QA and production. Beneath each major section, items are organized by use case or project.

As a set of items moves through the lifecycle and makes a major jump from one step to another (i.e. development to QA), it's moved to a folder representing that use case or project in the next step. Teams can track progress based on the movement to folders representing the different steps.

6.6.2 Share

In this model, groups and roles are defined and used in accordance with the lifecycle steps:

- A central governance group and role will create sharing policies
- A data analyst group will facilitate access to data
- Very finite analyst groups and roles will work with the specific use cases
- Different business roles will be created for specific access to certain resulting data sets
- Different QA and production groups can also be created to those specific steps

Members may also cross different groups, with certain team members facilitating core analysis and QA, but not necessarily facilitating operationalization (production).

6.6.3 Govern

Strict controls will be placed on who gains access to items within each step and use case combination. This is to ensure the proper movement and control of artifacts in the different steps:

- Data should be locked down and as needed, encrypted and masked
- Different data connections and policies would be created for each phase as the data needs at the development phase may be more ad hoc in nature and more production-ready in the production phase
- Business analysts can create new entities (workbooks, etc.) for their use cases in the development folder, but perhaps only have read access at the QA and production folders to eliminate change management issue but help aid troubleshooting
- Operationalization team members will help move items from use case folders across the various phases to ensure proper migration
- Business teams will be provided read-access to specific data sets that come from the production folder and will likely see aggregated data and *not* raw, individual data (rows)

This model is for organizations that are maturing in their big data journey. While this governance structure and policies may seem somewhat complex, it helps facilitate rapid analytic cycles and operationalization. It also facilitates trust in the analytic results, having gone through a proper QA and operationalization phase.

6.6.4 Pros and Cons

Pros

- Designed for specific environments that need strong controls within processes
- Designed to aid processes with regulatory and compliance rules
- Tight security ensures data is locked down and only used for specific purposes
- Speeds operationalization processes
- Delivers a level of confidence in the analytics being delivered

Cons

- The model is highly structured and somewhat restrictive
- Does not facilitate sharing

7. Governance Reference Architectures

Now that we've looked at different models to govern inside of Datameer, let's examine different governance architectures that can be used with Datameer. There are three different governance reference architectures:

- **Datameer-centric** — Governance is done specifically in Datameer
- **Data lake-centric** — Governance is joint effort with Datameer and Hadoop tools
- **Enterprise-wide** — Governance is performed with dedicated governance solutions

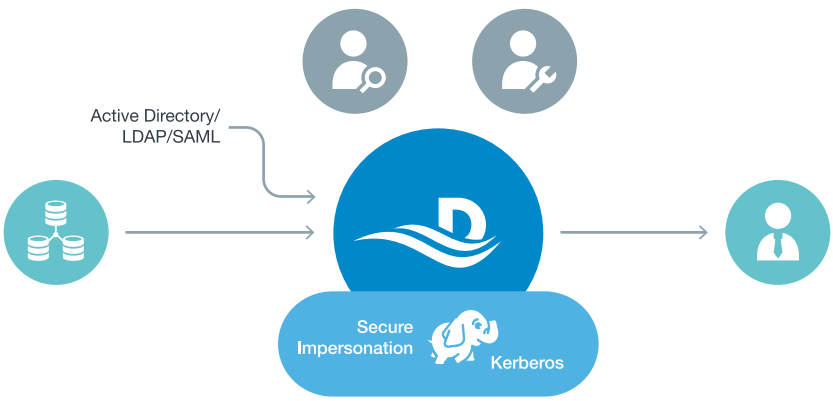
Let's dig deeper into each of these three architectures.

7.1 Datameer-centric

The first reference architecture focuses around Datameer, with much of the governance work performed using Datameer administration features for folder organization structure, user management and role-based security. There are two critical integration points with external services:

- A combination of LDAP or Active Directory, and possibly SAML for authentication
- Secure Impersonation, optionally with Kerberos integration

When LDAP or Active Directory are used solely for authentication with roles still defined inside Datameer and applied to artifacts. Secure impersonation is used to ensure jobs run with the same privileges as the Datameer user for close security with the Hadoop cluster. If the cluster is secured using Kerberos, then that integration should be configured.



It is very important to integrate Datameer with the security of your Hadoop cluster for governance best practices. This provides IT teams with needed confidence to trust the security of data inside Datameer.

7.2 Data Lake-centric

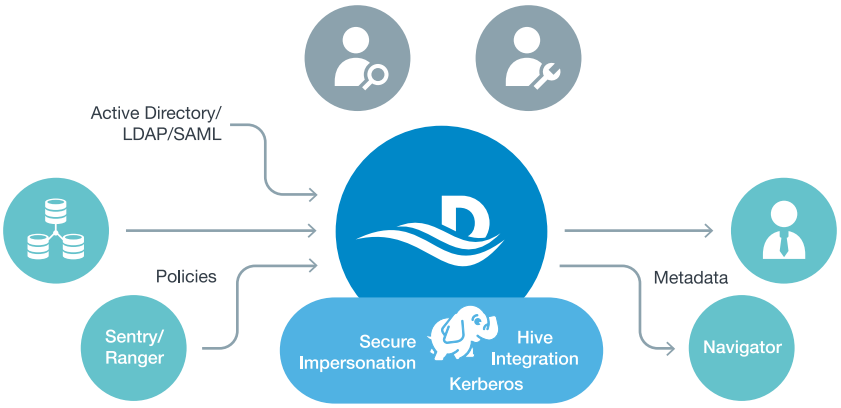
The second reference architecture is used when Datameer is part of a larger Hadoop-based data lake that also hosts other data and services. In this scenario, Datameer should be configured to be a good security citizen in the Hadoop environment.

All of the integration points discussed in the Datameer-centric architecture still apply. In the Data Lake-centric model, secured impersonation with Kerberos almost always must be configured for a consistent and confident security model. This will be important if data is imported from or exported to Hive.

In addition, you may consider two additional integration points in the Data Lake-centric model:

- Sentry (Cloudera) or Ranger (HortonWorks) for integration with security policies
- Navigator (Cloudera) for the export of metadata from Datameer

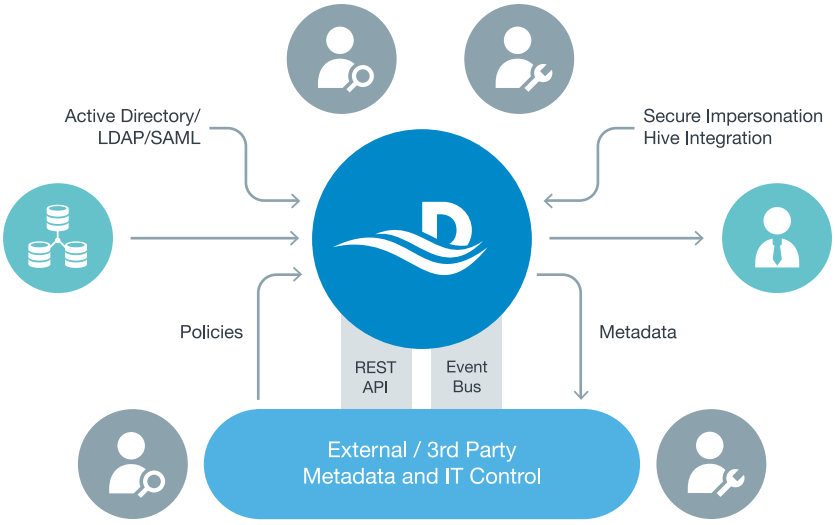
Sentry and Ranger provide easy methods to manage authentication and privilege policies over services across the entire Hadoop cluster, including Datameer. Navigator provides a central meta data management service on Hadoop.



7.3 Enterprise-centric

The third reference architecture for governance is used when organizations have governance managed from centralized services or specific governance and cataloging systems. Collibra is an example of a third-party solution specific for governance and cataloging which has been integrated with Datameer.

In the enterprise-centric model, you will use the Datameer Event Bus and REST API to integrate with the external governance and cataloging systems. Any meta data on Datameer artifacts, including revision history, changes and lineage can be sent to the external control systems.



8. Conclusion and Other Considerations

In conclusion, here are a few other considerations to consider as you define and evolve your data governance strategy.

8.1 Scope

When starting down your big data journey and your data governance journey in conjunction with this, it's important to manage the scope of your governance rollout and policies. Manage the policies in accordance to the maturity and scope of your overall big data analytics efforts — don't get ahead of yourself. Be realistic and don't try to boil the ocean.

However, it's important to think of a somewhat bigger picture, as your governance program will eventually grow as your big data analytics evolve. Therefore, start with governance strategies that "act local" on very specific needs, but "think global" with an architecture and approach that can grow with you.

8.2 People

Your people will play an essential role in the success of your governance program. It is important that you create a role for data stewards and foster this role within your organization. These will be the people that aid your data democratization.

You also want to set policies with input from the various teams involved. A democracy if fostered by a set of rules created collaboratively and then followed by everyone, and not an autocracy. This will enable you to create an environment of trust, not suspicion.

8.3 Process

You started your big data journey to create greater agility in your analytics. Your governance should mirror this. Don't get too bogged down in process in your governance policies and programs. Keep these agile and evolve them over time.

8.4 Fit

We examined a number of different approaches and architectures for governance. Choose the approach that is right for the unique needs of your organization, data, analytics and business teams. And, as previously mentioned, you can mix and match these models to fit individual needs with departments or business units.

FREE TRIAL
datameer.com/free-trial

TWITTER
@Datameer

LINKEDIN
linkedin.com/company/datameer