Collibra

DATA DRIVEN . BUSINESS . DRIVEN DATA

MAPR

The Need for Big Data Governance

A Whitepaper By Collibra and MapR

Collibra Inc
25 Broadway, 9th Floor
New York, NY 10004 – USA
(t) +1 646 963 6513

Contact@collibra.com

MapR Technologies
350 Holger Way
San Jose, CA 95134

408-914-2390
www.mapr.com

The need for Big Data Governance

Data Governance is essential to delivering maximum value from your big data environment. Without knowing what data you have, what it means, who uses it, what it is for, and how good it is, you can never create the insights and information needed to run a modern data-driven enterprise. Instead of an afterthought, data governance needs to be front and center in the organizational effort to harness the power of its data.

Why do I need Data Governance alongside Big Data?

Data governance enables the organization to take care of the data it has, get more value from that data, and make important aspects of that data visible to users. It also provides capabilities to manage these aspects. This is critical not only because of errors and omissions for existing data, but because new uses of data often require new attributes and therefore new metadata to support them. This flexibility is one of the great benefits of using Collibra with MapR as your big data management solution.

What does data governance entail?

Start with people and process

Data governance is about enabling and encouraging good behavior regarding data, and limiting behaviors that create risks. This is the same whether you are in a big data environment or a traditional data management environment. Enabling organizations to identify who is responsible for the data, collaborate to set policies and make decisions, create explicit agreements about how the data is used and what it is for, understand where certain metrics and information are derived, and determine the business impact of changes to data. These are all things that are needed in any environment.

Use technology as the enabler for the above

These processes are usually highly varied and involve many different stakeholders from different parts of the organization. They are also time sensitive. Especially in a big data scenario, the type, amount and frequency of change of data is increasing all the time. While it is possible to perform each of these tasks once or twice, executing governance on a continuous basis is impossible without dedicated systems and automation. Think of this in the same way most of our business processes require automation today to execute efficiently and effectively. In addition, like any other process, the governance of data itself must be measured and managed, so that the quality, utility, and security of the data can be improved when necessary.

These processes also must connect with the right stakeholders. In many cases, the individuals that understand the meaning and uses of data are not familiar with the technical aspects of its management. They are people who work in the business units that use this data to provide value. To capture their input, and to provide them with useful assistance requires an application that is specifically tailored for their needs.

Define up front what aspects of data management are critical to your business

Knowing what you need to govern is a critical part of implementing proper data governance. While all information probably should be subject to some governance, and should be cataloged so it can be found, there is a subset of crucial information that should be the focus of any data governance effort. These critical data elements and their antecedents are the basis of how the organization makes decisions, services customers and reports to regulators.

Leverage existing processes and best practices

Data governance doesn't have to be a new and burdensome initiative. In fact, some organizations might put aside a formal data governance program because of the perception of inhibitors and overhead. But the fact is you already have processes in place that act as a foundation for a formal data governance program. These might be labeled as "workflows" or "business rules," but those are merely different terms for the same set of practices. Use these processes as a starting point to build out a strategy that helps you to gain more value from big data.

Indeed most organizations have some recognition of the importance of these activities and may have systematized them within a particular function or domain. At the same time, these activities often do not represent a hindrance to gaining business value. For example, essential to the transition to a data driven organization is the understanding that interesting insights from data come when data is combined in new ways. The implementation known as a "data lake" necessarily requires processes that allow you to keep the data you need in a way that eliminates technical barriers, and gives new capabilities to process that data. This flexibility means that the processes for managing and governing the data can be seamless when applying them across the entire scope of newly accessible data.

What are the key requirements of data governance

Each of these capabilities enables a governed environment. The combination of catalog and data dictionary metadata gives the complete information for auditability of data policies and usage. It also contains lineage and manipulations. Workflow controls the procedures for information lifecycle management, including ingestion, manipulation/derivation, and disposal. And a robust distributed infrastructure can enable business continuity, rapid processing, and continuous availability of data.

Ensuring system reliability and availability

While concepts such as high availability and disaster recovery are often not classified as components of a data governance strategy, these capabilities are critical to any environment where data is a valuable asset. Therefore, strategies for data governance must inherently include strategies for high availability and disaster recover. After all, if a system can't reliably stay running, then data is devalued along with the associated data governance strategy.

High availability, or ensuring your system is continuously running at a single data center, can often be a complicated objective. You should ideally should seek a system that is designed to minimize the administrative overhead of overcoming failures within a cluster. If a hardware component fails, your team should be able to limit the response to simply replacing the failed component, not to reconfigure the software to overcome the failure.

Disaster recovery is sometimes overlooked as a critical component of a production environment, mostly because the likelihood of a site-wide disaster is rare. But for any mission-critical environment, having those safeguards in place are important. Interestingly, more and more organizations seek global deployments that put copies of data in geographically-dispersed regions. In these scenarios, usually the primary goal is to reduce access times by putting data closer to the users. A side benefit of global deployments is the replication of data that also deliver a disaster recovery configuration. Should any site suffer an outage, the local users can access remote clusters in the meantime to continue daily operations. However, increasingly we are seeing these types of global deployments in response to regulatory pressures, such as keeping personally identifiable information in the home country of the individual. This is a great example of how the governance of the data, and the policies on its retention and location intersect with the system configuration and the approach to availability and disaster recovery.

Organizations should also safeguard against data corruption resulting from application or user error. For example, snapshots capabilities create point-in-time views of data, ensuring a data recovery option should data corruption occur. Snapshots are also a great way to track data history and lineage by capturing a read-only view of data at a specific point-in-time that can be traced back during forensic analysis.

Identifying data and maintaining a data catalog

Because of the explosion in the variety of data, cataloging your data and making that catalog available to users is of critical importance. However, this is not just a matter of tracking the technical metadata about information. It requires an engine that can automate much of that process. There is just too much data and it is growing too quickly to manually categorize everything. Second, this catalog needs to be accessible by business users, so they can "shop" for the data that they need to examine business issues. This means that the implementation of the data must be tied to business terms. Organizations require a business glossary that can be easily augmented and updated as new data and new uses for existing data come into play. Also, it is critical that the environment be able to track all sorts of assets that are related to the data, its use and its processing. Technical components such as MapReduce jobs, user approaches such as visualizations, and data analysis objects such as models, sub-models and source data sets must all be easily represented in the data dictionary.

Exploring that data to identify opportunities

Once you have a catalog of your data, you need a good way to find things within that catalog. Different roles within the organization need to see things in a different way. IT professionals need to view the data in a system or application context. The security team needs to see information based on its privacy policy context, and auditors need to see the complete lineage and related information, through flexible visualization, that is capable of showing any type of relationship in context. This should be coupled with role based views and templates, as well as customizable navigation by roles to make it as simple as possible for the business user to find what he or she needs. Advanced machine learning and artificial intelligence can be used to aid the process of finding appropriate data. Of course, the underlying data processing mechanisms need to be highly performant to deliver the data as needed and where needed.

Maintaining the validity of the data

Maintaining the validity of the data is a two part effort. First is to establish the systematic policies and controls around data, and insuring that the measured accuracy of the data is suitable for its uses. Workflow and policy management capabilities enable the business users who understand the data to easily collaborate, negotiate and approve policies and procedures. Integration with IT service management systems allows policy information, once finalized, to be transferred to IT for implementation, without losing the context and the specifics of the policy and its requirements. Automated rules make it easy to determine if policies are without guidelines. A comprehensive business lineage enables anyone to determine which policies apply to which information, which business terms are refer to what data, and help determine the impact of changes to the data.

Data quality dashboards display the results of data quality scans, and give a perspective on whether the quality of the data is improving or not. They should enable the organization to accumulate values about data that fails the quality rules to assist in prioritization and improvement.

This combination of rapid and responsive policy management combined with ongoing quality improvement creates the ideal environment for maintaining your data and keeping it of high quality.

Protecting sensitive data

Securing data can be a complex effort, and while this capability also relies on having the right people and process in place, the technology can go a long way in ensuring proper protection. Protecting sensitive data requires several steps. First, the data and its sensitivities must be identified. Second, there must be a way to explicitly state the policies around what may be done with data, and by whom. Third, there must be a systematic means of collecting this information and disseminating it, both to consumers of data and to technologists who can implement physical controls on the information. These three steps are often driven by the need for achieving regulatory compliance, but also are tied to your own internal policies. Critically, there must be an infrastructure that has the flexibility to accommodate all different types of data with different sensitivities, and to create those controls in a meaningful way.

There are three governance activities that are critical to insuring the protection of data in a big data environment. First there must be some control on the data as it is brought into the environment. This ingestion control is important to insure that the data can be properly identified. Second, there must be a way to assign appropriate policies and to develop new policies for security and privacy of the data. These policies need to be explicitly associated with specific sets of data, and need to be visible to everyone who can use that data. Also, these policies must be linked to specific enforcement. That is usually using the third element the controls, procedures and scripts in the data management environment. Each of these things needs to be integrated, so that when data is in the big data environment, its protection is assured and unambiguous. This limits risk for the organization.

What is different about big data and how does that affect data governance?

There are several things about big data that change previous understanding of data governance. Each of these requires a new approach to governing the data assets effectively.

Variety of assets, including jobs, models, visualizations

The first major difference is the number of different types of data assets, and the fact that this category is growing. Assets such as MapReduce jobs, stream locations, sensor information for Internet-of-things (IoT) sources are all a part of the current big data landscape, and need to be accounted for in your policies and procedures.

Lack of physical separation between classes of data

The second difference is more subtle, but it is a byproduct of the way we have used our information management topologies to segregate data. Often, we rely on the physical separation of some data to identify it as sensitive in some way, and to manage the controls around that data. In the big data world, while the data can be distributed, that physical separation often does not exist, and other means must be used to identify what data is sensitive and who has responsibility for it. Governance processes need to maintain this information.

Creating value by combining data that has not been related before.

Also, the sharing of data is often a process that has not been formalized. The goal of the data lake is to create an environment where all the data can be easily utilized. This means that the different parts of the organization that own the data must all agree to provide it, and provision it in an controlled way. In addition, the data can now be shared with many parts of the organization, often without much effort on their part. This means that data sharing requirements need to be explicitly negotiated, so that all users of the data understand what they should and should not do with the data. Also, the scope for semantic mismatches increases, as different parts of the organization will use the same terms with different meanings. The business glossary and data catalog are crucial for sorting out which data is the meaningful data for a particular purpose.

More varied and flexible processes.

Instead of the ETL-based up front definitions and policy determination, big data implies a bottom-up ”do it as you need it” approach to governance. This in turn means that the automation system for that governance needs to be highly flexible and collaborative, as well as having a clear operating model. This operating model, which takes into account the entire lifecycle of how data is provisioned, used, changed and retired as well as its quality and reliability, needs to be automated to deal with the ever increasing amount and variety of data.

Increases in variety make automation a requirement.

That increasing scale in amount and variety demand automation. Manual processes cannot keep pace with the number of changes to data and the new data that is brought into the lake almost daily. Manual processes are too slow and cumbersome to deal with the volumes of data that are now available. Keeping governance information on desktop tools, spreadsheets or even document sharing sites is too slow and cumbersome, and does not give the business users the instantaneous access to the data that they need. It is necessary to automate the data governance with an application, the same way you automate any other business activity with a specific application for that purpose. Collibra Data Governance Center is designed to be that application for your big data environment, and all your data governance needs.

This data is an operational necessity and is in constant use.

Ultimately, this data is the lifeblood of the organization. The infrastructure and platform on which it is hosted and processed must be able to keep up with all of the changes, as well as the volume of requests to use and process the data. Without that level of reliability and security the organization will not be able to make use of its data, and will not be able to acquire new sources of data and insight rapidly enough to be competitive. Data has become a key level in business competition and quality of products and services.

Big Data Governance Success

In the end, these processes are designed to make your organization more agile and capable. You can use your data as and when you need it, you can add to it, you can manage it, and it is there for you. The best organizations with big data and governance capabilities find that there are a number of specific benefits. They can find data, describe data, use it and manage it more effectively than ever before.

Maintain availability

When managing big data, you want to maximize uptime while minimizing the effort in ensuring that uptime. Your underlying big data platform must deliver on these objectives. With the MapR Distribution, you get:

- No single points of failure
- Built-in high availability features that ensure uptime with no manual effort for setup or failover
- Block-level, scheduled, incremental mirroring for mission-critical disaster recovery strategies
- Table replication in the integrated NoSQL database, MapR-DB, to run globally distributed, multi-master deployments
- Consistent snapshots that accurately capture the state of the cluster at the time the snapshot is taken, to protect against user/application-caused data corruption

Secure sensitive data

A wide range of important features are required to address your data security requirements. Your data platform must provide the data-centric controls to ensure a secure environment. With MapR, you get:

- Integration with Kerberos as well as a wide range of user registries via Linux Pluggable Authentication Modules (PAM)
- Access controls not only at the file level, but at a granular level in MapR-DB, including at the column family, column, document, and sub-document levels
- Encryption for data in motion, so all network connections between nodes, applications, and services are protected from eavesdropping
- Comprehensive auditing that logs data accesses as well as administrative and authentication activities

Speed access to your data

Organizations spend up to 75% of the elapsed time for creating analytics engaged in data wrangling activities. This process of finding appropriate data and massaging it can be dramatically sped up using Collibra Data Governance Center. This is because there is an easy to use data catalog linked to a business glossary. Users can refer to the data in the terms that they understand and use. Data is easily searched, and machine learning artificial intelligence assists in recommending data that would be appropriate. Advanced visualizations can display any type of relationship and context of the data, so data scientists and BI professionals can easily get to the right data quickly. And because data ingestion governance assures that you know what exactly is in the data lake.

Change data quickly and safely

Insures that your queries return the right data so the analytical metrics based on that data can be trusted. Data scientists, owners and users can insure that the correct data values, references and results are used. Using unstructured data demands efficient coordination between producers, consumers and data scientists, to ensure all parties are aware of changes that might impact results. Since the changes to this data happen frequently and often continuously as new uses are found for that data, this is a critical capability. This communication also reduces time consuming error analysis and resolution; partly because there are few inexplicable errors in the analysis, and partly because the process of reporting problems and resolving them is automated. This increases the trust in the analytics, and increase their use, and promotes self service. Collibra Data governance gives you complete control and visibility into your data, its policies and its attributes.

Know your data

Data governance lets you know what you have, and find that knowledge in many different way. A big data environment is not just tables, files and streams.. There are many different types of assets that organizations use to deliver high performance, predictive analysis, and unique insights. These include analytical models, map/reduce jobs, queries, visualizations, reports and any artefact that uses the data. Each of these or custom assets can be easily configured and used in Collibra Data Governance Center. It provides complete visualization of any type of relationship, including lineage relationships and context. Every one of these capabilities is designed to make sure that you are using the right data for the right purpose at the right time.

MapR and Collibra Work Together to Deliver Business Success

MapR provides the performance, availability, security and scalability needed to tackle the biggest jobs, and Collibra insures that the MapR implementation has the right data for the right purpose, and that it can be easily found and managed. The combination of the resilient, highly reliable and performant platform from MapR with the intelligent automated governance from Collibra bring together the capabilities that your organization needs to truly be data driven.

About MapR Technologies

MapR delivers on the promise of Hadoop with a proven, enterprise-grade platform that supports a broad set of mission-critical and real-time production uses. MapR brings unprecedented dependability, ease-of-use and world-record speed to Hadoop, NoSQL, database and streaming applications in one unified distribution for Hadoop. MapR is used by more than 700 customers across ad media, consumer products, financial services, government, healthcare, manufacturing, market research, networking and computers, retail/online and telecommunications as well as by leading Global 2000 and Web 2.0 companies. Amazon, Cisco, Google, Teradata and HP are part of the broad MapR partner ecosystem. Investors include Google Capital, Lightspeed Venture Partners, Mayfield Fund, NEA, Qualcomm Ventures and Redpoint Ventures. MapR is based in San Jose, CA. Connect with MapR on Twitter, LinkedIn, and Facebook

About Collibra

Collibra Corporation is the industry’s only global data governance provider founded to address data management from the business stakeholder perspective. Delivered through a cloud-based or on-premise solution, Collibra is the trusted data authority that provides data stewardship, data governance for the enterprise business user. Collibra automates data governance and management processes by providing business-focused applications where collaboration and ease-of-use come first. Collibra’s data governance platform embraces the new requirements of big data solutions, with automation, machine learning, and the flexibility to govern data assets from source to visualization. Find out more at http://www.collibra.com/