CapTech® Others Talk, We Listen.

WHITE PAPER

5 WAYS MODERN DATA GOVERNANCE WILL MAKE YOUR ORGANIZATION MORE PRODUCTIVE

Introduction

Many data scientists and analysts view data governance as a roadblock that keeps them from using data efficiently. In working with clients across a wide range of industries, we've found that it is actually the lack of sound data governance that prevents many organizations from realizing the full value of their data.

In fact, a 2015 survey by Forrester found that 21 percent of data and analytics business decision makers aren't satisfied with analytics in their companies.[1]

While blaming this dissatisfaction on the seeming obstacles imposed by data governance may be tempting, it would be misguided. In reality, a key reason for dissatisfaction is that many organizations are overwhelmed with data and don't understand what they have, where it came from, nor whether it is reliable.

As organizations move to upgrade their data stores from traditional relational database systems and to big data architectures, these challenges are even more overwhelming. Instead of operating data lakes, many organizations find themselves trapped in data swamps. Analytics can't solve that problem, but data governance can. Organizations that practice good data governance principles can improve data quality and make better, more timely decisions.

> “Twenty-one percent of data and analytics business decision makers aren't satisfied with analytics in their companies.”

Data governance deals with such questions as the origins, or lineage, of data; who can access data and what they can do with it; how data is categorized or catalogued; and the quality and completeness of data. In addressing those questions clearly and directly, data governance increases the productivity of data scientists and analysts, directly benefiting the business decision makers who count on them for insights.

This paper describes attributes of an effective data governance program and cites best practices that can help turn data governance from a seeming obstacle into a powerful asset.

#1 Good Business Metadata is Good for Business

A key component of data governance is the process of effectively managing metadata, or data that labels or categorizes other data. For data scientists and analysts, metadata facilitates the discovery process, clearing the way for a wealth of opportunities of immense business value (e.g., gaining insights into customer behavior, evaluating opportunities for reducing business costs, and identifying patterns of behavior that may be fraudulent).

According to the National Information Standards Organization, metadata “serves the same functions in resource discovery as good cataloging does by allowing resources to be found by relevant criteria; identifying resources; bringing similar resources together; distinguishing dissimilar resources; and giving location information.”[2]

In other words, just as a card catalog makes it possible to put the materials in a library to efficient use, metadata makes it possible to put data to efficient use. Effectively governed metadata enables data scientists to find what they need, when they need it.

Best Practice

Instead of waiting for the governance team to collect metadata, modern data governance looks to data consumers to help with this task. Crowdsourcing your business metadata gives you richer, more relevant detail because it comes from the user. Data scientists and other consumers of data can use appropriate tools to enrich metadata to meet their needs and then share insights with peers. The practice of crowdsourcing metadata is similar to posting information on Facebook and seeking comments from others. A data steward is assigned responsibility for overseeing, or the curation of, crowdsourced metadata and approving it for wider use. Data not approved by the steward can still be used, but with the understanding that it has not been vetted.

Case Study: Metadata Governance

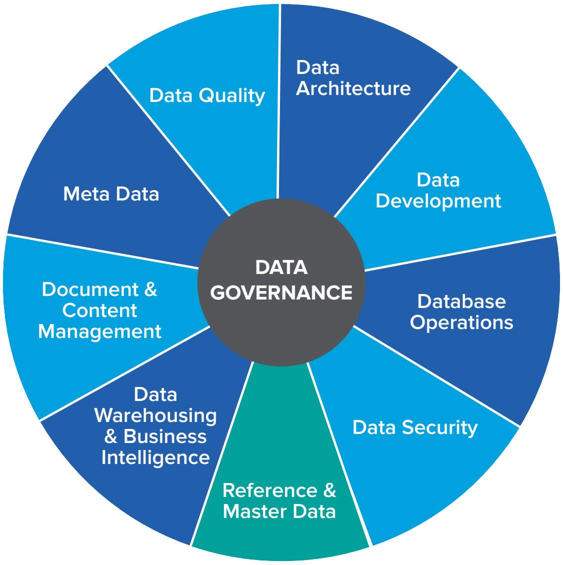
A Fortune 500 financial institution was struggling to find data within its data lake. CapTech's extensive experience in metadata governance was sought to support the development of a custom metadata repository.

Results

- The client can now catalog and search metadata within the Hadoop ecosystem.
- The client can review and approve metadata in the data lake.
- All data ingested is tagged and cataloged at the time of ingestion, which keeps the data lake clean and gives data scientists a better understanding of what is available.
- The repository helps the organization achieve compliance with privacy and security regulations.
- Data scientists can quickly find what they are looking for.

#2 Effective Schema Management Saves Time and Money, Especially in a Big Data Environment

A critical part of modern data governance, particularly in the big data space, is schema management, which provides a way to catalog and define the business intent of each schema developed for a given data file.



A schema defines how data is to be read; for example, a schema may specify that the first nine digits in a data file are Social Security numbers. Data files typically can be viewed in many different ways through many different schemas. As source data evolves, the associated schema must evolve with it. It's essential to be able to manage schemas and understand which schema to use when looking at particular files.

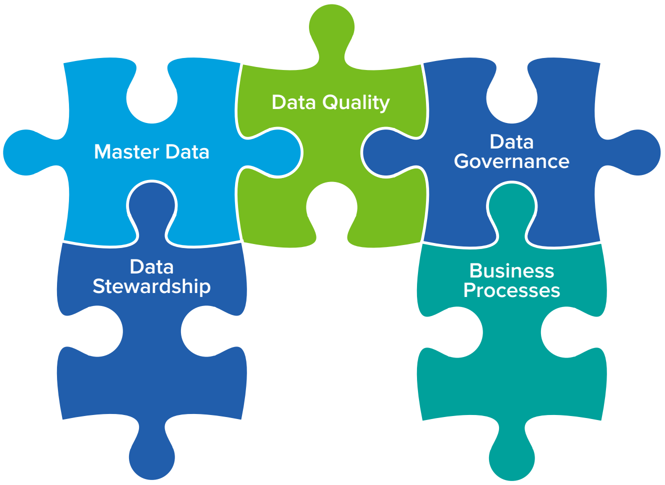
Data lakes typically contain information that doesn't always have a well understood schema, yet schema management and business metadata discovery tools for the big data environment are still immature and few are commercially available today. The problem is that it's difficult to programmatically determine the technical schema and business metadata and store them together in a central location within the big data ecosystem. The most logical place to store this information is the Hive metastore. While the Hive metastore is an excellent repository to store technical schemas within the Hadoop ecosystem, it does a poor job storing business metadata. Additionally, HCatalog – a collection of application programming interfaces (APIs) that offers a window into the metastore – provides limited capabilities to work with business-related metadata. Managing the technical schema and business metadata over time and across versions is difficult to accomplish in this environment.

Programmatic technical and business schema discovery eases these problems. When a new data set is ingested into the data lake, an open source tool can help you determine the schema automatically and, in a mature environment, match the newly discovered data to existing business metadata, providing you with the business and technical metadata immediately. The key technical metadata can be stored in the Hive metastore, and the business metadata can be stored and linked in a complementary store, such as HBase, Apache Atlas, or Cloudera Navigator. Taking care of this upon ingestion of new data frees your data scientists to focus on more important matters.

Once business and technical metadata has been captured and cataloged, you can overlay it with usage patterns and gain a real-world view of how data is being used.   This allows data scientists to better understand relationships in the data and discover new insights while further enriching the metadata.

Best practice

Capture all known information, both business and technical, about each schema in the Hadoop environment. Publishing, curating, and governing known schemas saves data scientists considerable time. If existing schemas are documented and accessible, data scientists don't have to build a new schema each time they want to use a data set. They can leverage schemas that others have developed and review and contribute to existing schemas.



Schema management and programmatic schema discovery

CapTech can help with the complex problem of programmatically discovering and storing your schemas so that data scientists can easily search and find what they're looking for.

Case Study: Schema Management and Programmatic Schema Discovery

Data users at a Fortune 500 financial services company were spending excessive time documenting technical schema. CapTech was engaged to support building a process that automatically discovers and stores technical schemas programmatically.

Results

- Data scientists no longer have to document technical schemas and can focus on higher value work.
- Data stewards can focus on business metadata and other priorities instead of technical metadata.

#3 Good Data Quality and Profiling Can Accelerate Time to Insight

Poor data quality is a primary reason for 40 percent of all business initiatives failing to achieve their targeted benefits, according to a report by Gartner, which also noted that data quality affects overall labor productivity by as much as 20 percent.[3]

Investing in sound data governance practices can help you improve data quality and give your organization the productivity edge it needs, with quicker time to insight. Additionally, it can boost revenue. Combining a data quality strategy with targeted data quality improvement efforts that solve conflicts at the source can lead to a 25 percent increase in converting inqui-

> Poor data quality is a primary reason for 40 percent of all business initiatives failing to achieve their targeted benefits, according to a report by Gartner

ries to marketing qualified leads, according to destinationCRM.[4] A data governance system that proactively monitors data as it is ingested and that alerts data stewards to data quality issues can make data more reliable and trusted by analysts, data scientists, and business users. Trusted data sets are inevitably used more often – and questioned less often – so why not strive to make all data high quality and trusted?

Profiling data and storing the profiles with metadata is also an essential step, as this gives data scientists and analysts a better understanding of the types of data contained in the system, allowing them to formulate hypotheses more quickly. A profile might indicate, for example, that a particular column of data contains two characters or a null value. A quick look at the distribution of these values may indicate that they are state abbreviation codes.

“High-quality data regarding customers is particularly critical. When customer data isn’t correctly matched, merged, and cleansed, you are left with data of poor quality, which can make the customer experience frustrating.”

High-quality data regarding customers is particularly critical. When customer data isn’t correctly matched, merged, and cleansed, you are left with data of poor quality, which can make the customer experience frustrating. A well-managed set of master data regarding customers gives data scientists and other data consumers confidence that their analytical and operational needs for customer data will be met and will represent the customer or segment of customers the data consumers are focusing on.

Best Practice

Create a data usage agreement between producers and consumers of data. The agreement will specify the level of data quality that is to be expected and how quality will be documented and captured. The agreement should also cover the availability, consistency, and accessibility of data; put in place controls regarding usage of data; and establish metrics that can help make data useful for data scientists and other consumers.

The right architecture for your business

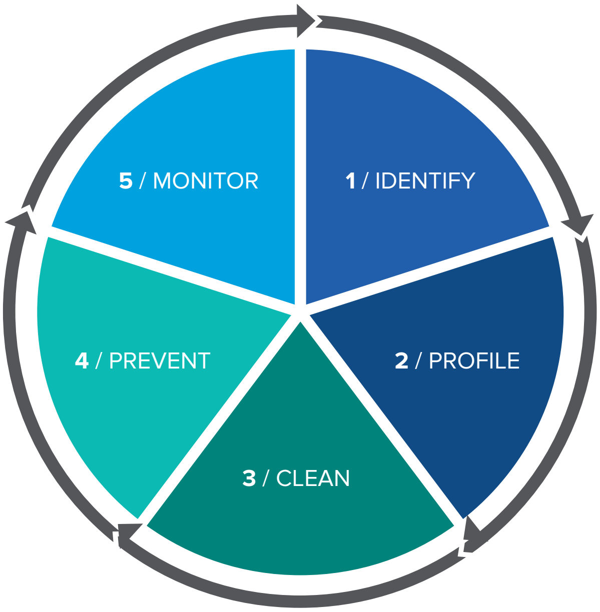
CapTech architects can help you determine the right architecture for your data usage.

We know that data quality can be difficult to improve. That’s especially true in a big data environment, given the size of the data sets and the amount of processing required to run data quality rules. It can also be a difficult issue if there is a time component associated with usage of the data. Our analysts can help you understand what data you have and develop an architecture and data quality protocols that can keep your data lake from becoming a data swamp.

#4 Data Lineage Can Help Keep You From Getting Sued or Fired

In an era of highly publicized data breaches, data governance can provide important protections to you and your staff. Although it won’t stop determined hackers from gaining access to secure information, it will help you understand what has and hasn’t been compromised in the unfortunate event of a breach.

Data governance affords particular protections to those working in regulated industries where regular and detailed reports are sent to government agencies, such as financial services and healthcare. In the event of an audit, a sound



> In an era of highly publicized data breaches, data governance can provide important protections to you and your staff.

data governance program will enable you to demonstrate where your data came from and how you made calculations. Knowing that the data that your data scientists use is well understood and traceable to the source (through lineage) gives these professionals the confidence that their work can be validated and trusted.

Capturing metadata that deals with data sensitivity and privacy is also helpful, as this serves to make data scientists aware of potential issues so they can comply with rules and regulations governing data usage.

All of these measures help you meet legal, compliance, and regulatory requirements while protecting your data scientists and analysts from a wide range of pitfalls.

Best practice
Train data scientists on your data governance program so that they fully understand the rules and controls that have been put in place to protect them. If they are unaware of these provisions, they won't use them and will be at greater risk of misusing data or using data of poor quality.

> ### Comprehensive data governance
>
> CapTech can help you establish a comprehensive data governance program tailored to the unique needs of your organization. Our experienced staff can implement the data governance program for you and deliver in-depth training to data scientists, analysts, and other data users.

#5 Your Models and Analyses Will Run Right in Production

If you've taken the steps outlined thus far, your data quality should be sufficiently high enough that you'll greatly limit the potential for problems with models and analyses in production. You'll also limit the potential for decisions and reports to be based on unreliable data.

Data usage agreements between producers and consumers of data will clarify the level of data quality that is to be expected and how that will be documented and captured.

Moreover, data of poor quality will be identified upon ingest through routines that notify data stewards and other consumers who subscribe to alerts.

Data governance for financial and statistical models

In financial services and other regulated industries, establishing data governance with respect to statistical and financial models is critical. Models play a central role in business decision-making and, as such, are of immense interest to regulatory agencies.

In the banking industry, for example, models calculate the likelihood that credit card customers will default and, further, calculate the likely amount of defaults. It's essential that these models be reliable and able to withstand regulatory scrutiny.

How you build, store, track, and monitor models; what data they consume and where that data originated; the accuracy and completeness of the data; how you use models and what the outputs are—these are all key questions, both from a business standpoint and a regulatory standpoint.

To the extent that your data governance program addresses these questions, it can help you avoid regulatory compliance issues, improve the efficiency of your data scientists and ensure that business decisions are based on reliable data and reliable analyses.

Some important points to consider:

- **Are models fully documented?** In regulated industries, it's imperative that you fully document models and the data feeding them. You also need to document any controls you have established to verify the validity of processes. Equally important: documenting policies regarding metadata and creating a catalog of the metadata your data scientists are using.
- **Is the inventory complete and current?** When acquisitions take place, new models appear. When regulations change,

new models are built. Unless the model inventory reflects this, your data scientists and others won't know what's available nor whether the business has a full complement of models to meet its needs. The update process also can provide opportunities to ensure that all additions/changes are valid and documented.

- **Is model code available and transparent?** Businesses need to be able to demonstrate how their models crunch numbers. Storing and managing that information in a central location will make it accessible and help you maintain version control.
- **Do you understand how models are used?** If you know how your models are used – and what data is used to build and drive them, where the data came from, whether it's reliable, and what the outputs are – then you'll be able to confidently respond questions from regulators as well as business decision-makers.

Best practice
Go a step further and establish preventive and detective controls. Preventive controls help ensure that low-quality data isn't used by the business. If a data quality problem develops in production, detective controls enable your production operations team to troubleshoot failed jobs, traverse lineage back to the failure, and contain the problem.

Improving quality and time to value

Making your models and analyses run properly in production is especially challenging in a big data environment because of the complexity of the systems and the number of tools available to manipulate the data. CapTech has extensive experience in DevOps and continuous integration/continuous deployment (CICD), which can help you put new capabilities into production faster and with higher quality.

Conclusion

A well-managed data governance program keeps your data scientists doing what they do best – finding insight – by providing what they need in order to remain focused on delivering business value. That includes metadata as well as information about data schemas, quality, structure, and completeness. Armed with such information, data scientists needn't spend their time looking for data, trying to understand definitions, wondering whether data sets are complete and accurate, or wondering where data originated. Data governance saves time, energy, and money while improving the quality of business decision-making.

A sound data governance program also will keep your organization safe and compliant, with full documentation of how data is used and by whom.

A comprehensive approach to data governance

CapTech can conduct a data maturity assessment and show you how your organization can get more out of your data. We have helped Fortune 500 clients establish data governance councils, put data quality standards in place, document and capture data processes, metadata, and create data usage agreements.

Case Study: Researching and recommending a data governance process

One of the nation's central banks sought to achieve operational excellence and service innovation; inform and influence policy choices; engage constituents; and better prepare employees to serve the public. CapTech was engaged to develop a strategic vision for business intelligence and analytics, an implementation roadmap, and methods of measuring and optimizing these efforts over time. CapTech also researched and provided recommendations regarding a big data program, metadata management program, and data governance process.

Results

- CapTech built a business case and recommended immediate execution of three strategic initiatives, including implementation of a data store inventory, creation of a business intelligence competency center, and establishment of a data governance program.
- CapTech provided a centralized presentation layer, technical and analytical training and a cost-benefit analysis for the metadata management program.

Resources

For more information about CapTech's approach to data governance, check out these blogs and presentations:

- Blog: https://www.captechconsulting.com/blogs/protecting-your-organization-from-the-next-big-data-breach
- Webinar with Cloudera: https://www.youtube.com/watch?v=EBMkEBRaZ9g
- Blog with Cloudera: https://vision.cloudera.com/establishing-a-secured-and-governed-big-data-platform-for-a-financial-services-firm/
- Blog: http://www.captechconsulting.com/blogs/four-questions-on-data-governance
- Blog: https://www.captechconsulting.com/blogs/data-governance-trends-in-2011

End Notes

[1] Michele Goetz. "Customer ecosystems demand outcome-oriented data governance." Feb. 1, 2016. Forrester Research. Available at https://www.forrester.com/Global+Business+Technographics+Data+And+Analytics+Survey+2015/-/E-SUS2955.

[2] Rebecca Gunther and Jacqueline Radebaugh."Understanding Metadata." 2004. NISO Press. Available at http://www.niso.org/publications/press/UnderstandingMetadata.pdf.)

[3] Ted Friedman and Michael Smith. "Measure the Business Value of Data Quality." Oct. 10, 2011. Gartner Inc. Available at https://www.gartner.com/doc/1819214/measuring-business-value-data-quality.

[4] DestinationCRM.com "Data Quality Practices Boost Revenue by 66 Percent." Jan. 22, 2009. Available at: http://www.destinationcrm.com/Articles/CRM-News/CRM-Featured-News/Data-Quality-Best-Practices-Boost-Revenue-by-66-Percent-52324.aspx.

AUTHOR BIO

Ben Harden | bharden@captechconsulting.com

Ben Harden leads the Data and Analytics Practice at CapTech and has over 18 years of enterprise software development experience in the areas of data warehousing, metadata management, data governance, analytics, engineering, and enterprise scale Hadoop data ingestion and refinement. He is also an Agile Scrum coach who specializes in making data delivery teams successful using the Agile methodology.