CapTech Others Talk, We Listen.

WHITE PAPER

A STRATEGY FOR ESTABLISHING DATA GOVERNANCE IN THE BIG DATA WORLD

Informed “Fishing” in the Data Lake

Executive Summary

As organizations embrace big data and the technologies that support it, they gain the potential to derive business insights from information that, historically, might have been overlooked. This can provide powerful competitive advantages in fast-moving industries. The key to unlocking these benefits is applying data governance and, more specifically, metadata to this new world of possibilities.

In a big data environment, however, data governance can be extremely challenging. Data governance consists of the policies and practices that an organization establishes in order to manage its data assets. Ineffectual data governance can create issues in such critical areas as traceability and data lineage, metadata, data quality, and data security. All of these issues can make it difficult for data scientists to deliver timely business insights.

Historically, organizations have been able to establish effective data governance, at least in part, because data governance was designed into the data store. Tables, fields, views, and other data objects were designed and defined before data was placed in them. Data security, or restrictions on data access, had to be established prior to actual loading of data. Finally, metadata, or the description of data, was typically developed as part of the database build.

In the big data environment, data is stored without such patterns. Even in a well-defined setting such as that presented by traditional relational database technologies, data governance is difficult. In an environment where structure is random, data governance can be a Herculean task.

CapTech has developed a strategy for bringing just-in-time structure to the big data world – automating the creation of metadata as data is ingested. This helps organizations retain the power of big data without sacrificing the power of data governance. We liken it to a fish finder that expedites fishing trips. Our strategy creates metadata, which expedites the search for answers while also providing data security, lineage, and data quality.

This paper discusses the historical context in which big data has emerged as a platform of choice, the importance of data governance, some of the data governance challenges that big data presents, and the CapTech strategy for addressing these challenges.

The Three Vs: Volume, Velocity, and Variety

At the height of the data warehouse building craze of the early 2000s, Doug Laney wrote a fascinating article for META Group, laying out compelling examples of the growth challenges presented by traditional data management principles.[1] He noted that relational database management systems (RDBMSs) were limited in their ability to handle these challenges.

Laney broke the challenges into three key areas:

- **Volume:** the number of transactions and quantity of data supporting these transactions
- **Velocity:** the pace at which data is used to support interactions and the pace at which data captured by such interactions is generated
- **Variety:** the mix of incompatible data formats, non-aligned data structures and inconsistent data semantics

As data volume, velocity, and variety continued to grow, new technologies were developed to address these challenges. In 2004, Google published a paper on its MapReduce, which introduced a parallel processing model that could handle significantly larger volumes of data more effectively than traditional database platforms could.[2] Many observers have pointed to this as the advent of big data and as the answer to Laney's three Vs.

Since the introduction of MapReduce, data volumes have continued to skyrocket. According to IBM, 2.5 exabytes, or 2.5 billion gigabytes, of data were generated every single day in 2012.[3] In one short decade, enterprises have moved from storing terabytes to petabytes and, more recently, from exabytes to zettabytes of data. Every two years, according to some observers, we create more data than existed in all of history.[4]

Although having access to vast amounts of data in a wide variety of formats, all arriving at high velocity, creates tremendous new opportunities for data science and business decision-making, it also creates new challenges.

One of the major challenges involves giving up an important aspect of data governance, namely metadata. To move data into new environments such as Hadoop and NoSQL, which are staples of the big data world, you limit your controls over data while also eliminating the traditional process of storing data in predefined structures.

In more traditional environments, you move data into tables and fields and, in the process, create metadata; i.e., knowledge of the data. In the big data environment, data files in native format are moved directly into storage. In the absence of metadata, you have little, if any, knowledge of what is in storage or where to find particular kinds of data. That can make it exceedingly difficult for data analysts and scientists to convert data into business insights.

Four Aspects of Data Governance

Data governance deals with four primary considerations, all of which can greatly facilitate the work of data scientists:

- **Traceability:** You need to know where data came from – i.e., its lineage – if you are going to rely upon it in decision-making.[5]
- **Metadata:** Having some information about the data you are storing will help you find what you are looking for. Metadata tells you what kind of data you have (e.g., bank account records, diagnosis codes for patients' medical conditions), who owns the data and who has access to it.
- **Quality:** A sound data governance program will employ processes that validate and standardize disparate data, ensuring that the data is of sufficiently high quality to be useful in decision-making.
- **Security:** Data governance provides for the protection of data, a particularly important consideration for organizations that manage personal data such as Social Security numbers, financial account information and medical information.

How did we get here?

Since the 1980s, organizations have been able to establish sound data governance programs relatively easily because of the way data has been stored and structured.

When I began working in the technology industry in the late 1980s, I was doing COBOL programming on mainframe systems. We stored data in flat file structures – i.e., hierarchical data structures such as IBM's IMS model – with at least some metadata that indicated folder names. This provided some clues as to the type of data stored in any particular folder.

At that time, computing was all about transactions, and innovators were focused on two major issues:

- Increasing the speed with which transactions could be conducted
- Reducing the amount of storage needed to support these transactions. These were major concerns in the mainframe environment because both processing power and storage were far costlier than they are today.

In the 1990s, RDBMSs such as IBM's DB2 provided a solution to both concerns, enabling organizations to reduce data redundancy, which in turn reduced the need for storage while accelerating the speed of transactions. RDBMSs and data management theory achieved this by employing data storage concepts derived decades earlier by Edgar Codd and Raymond Boyce. A critical concept was that of "normalizing" data, or breaking tables into smaller, less redundant tables and then relating the smaller tables with keys. This served to reduce unnecessary processing of data.

Before the use of RDBMSs became widespread, each time a credit card transaction took place at a store, for example, the data associated with the transaction would include the customer's name and address as well as the store's street address, name, and billing address. With thousands of transactions taking place at the same store each month, often with the same customers, the redundant data added substantially to data storage and transaction processing costs.

With an RDBMS, the credit card company could use a four-bit identifier to represent each store and customer. A related table stored information about the customer, and another table held data about the store; for example, street address and billing address. This dramatically reduced the amount of data associated with each transaction. It also cut the need for storage and accelerated transaction speeds. Furthermore, the structure reduced data quality errors, particularly in industries that relied heavily on manual labor to key data. However, the change didn't require a sacrifice in data governance, as the credit card issuer still assigned metadata to files.

Although the move from hierarchical to relational structures allowed for more efficient storage of data, it created challenges for data analysis.

In the early 1990s, I worked for the credit card division of a petroleum company that operated a large nationwide chain of gasoline stations. The company offered customers a branded credit card. Each week, I ran a weekly report of credit-card transactions. To minimize the impact on the transactional system, this process ran overnight. A full-time staff manually stored and retrieved the reports in a physical library.

When a new report was ready, the managers would request a printout on large green bar paper. The printout covered all transactions in a particular state or region, along with aggregate subtotals and totals. The managers would compare this with the report from the previous week and with reports from the same week a month and a year earlier. Each week, management would place these stacks of green bar paper side by side on a desk and, using a ruler, compare the data, noting changes in revenue, numbers of transactions and gallons sold within a state or region. The managers then would enter this information in spreadsheets for use in estimation, forecasting, and evaluating performance.

Beyond transactions: online analytical processing and data warehouses

The structure of RDBMSs and, more specifically, normalized data, made for efficient storage and faster processing speeds while also providing support for data governance, but it clearly did not lend itself to efficient or detailed analysis.

In the mid- to late 1990s, Ralph Kimball and Bill Inmon recognized that data is a powerful asset and that, in combination with analytical capabilities, it could improve decision-making across a wide range of operational areas.

Kimball and Inmon believed it was possible to leverage data to run analyses and create reports that would enable organizations to evaluate data over time.[6,7] One problem that stood in the way was that normalized data was structured in such a way that it took an inordinate amount of time to run a query. Multiple joins between tables were inefficient, requiring complex SQL queries. Another problem was that running queries decreased the efficiency with which transactional data could be loaded.

Credit cards: epitomizing governance

Credit cards represent a microcosm of governance capability, providing traceability, metadata, data quality, and security.

- **Traceability:** All transactions can be traced to specific cards and cardholders.
- **Metadata:** Additional information contained in the magnetic strip (for example, customer name and address) provides enhanced metadata.
- **Quality:** As a transaction is processed, data about the purchaser is electronically transmitted with the transaction. Avoiding manual processing eliminates many of the data quality errors of the past.
- **Security:** The magnetic strip contains the cardholder's ZIP code, making it possible to ask a security question before transactions can proceed. Similarly, debit cards require a PIN, which also provides security.

Online analytical processing (OLAP) and the data warehouse provided a solution, still in wide use today. In this model, data is moved from normalized tables to subject-oriented tables. This is the so-called star schema or data warehouse. As data is moved, it is also changed or transformed as well as aggregated and standardized. This process makes the data subject-oriented; for example, it might relate to such subjects as customer, order, finance or marketing. These subject-oriented entities consist of facts and dimension tables, which minimize the complexity of queries and speed the return of queries. Analyses can be run far more quickly, without interfering with the loading of transactional data.

Data governance remains feasible in this model, as users still have to predefine the storage units into which they put data. The transformation processes are often done with software that provides metadata as well as data quality processes, and many of these tools provide easy-to-use lineage maps.

The metadata provided through this approach indicates, for example, that a particular table is an account table, customer table or address table. This is suggested by the definition of the structure created. Metadata tags indicate what information has been loaded into these relational data stores.

Big data and lack of metadata

Since 2004, when Google and Yahoo came up with the concept of massively parallel processing (MPP) solutions, and since storage solutions such as Hadoop and NoSQL were developed, the way many organizations manage data has changed. Big data is at the center of this change.

As we noted earlier, the three Vs — volume, velocity, and variety — provide the appeal of big data. At its core, however, big data in some sense is a return to the mainframe technology of the 1980s. Distributed data management systems such as Cassandra are much like the hierarchical data stores of the 1980s; consequently, we can draw on some of the same methods to apply data governance principles, including metadata.

Big data solutions bring in data – email files, video files, JPG images – and store these in files within folders distributed across many servers. Big data technology manages the servers and the data that lives on the servers. Data is written across multiple physical disks. If one disk fails, the system can write data across the other disks. When a failed disk is replaced or fixed, it easily is brought back into the technology stack, so there is no need to shut down the system. This creates high availability of data. The technology also can load vast amounts of data without knowing much about the data; it does not require predefinition of the storage structure. (The storage area is often referred to as a data lake.)

However, the lack of metadata – not to mention traceability, quality, and security – can create serious issues for data governance. Vendors such as Cloudera, Datameer, and Datastax and technologies such as Apache Atlas have begun to close the gap on the data governance challenges, but the problem of the lack of metadata remains.

A strategy: creating metadata upon ingestion in the data lake

CapTech offers a strategy that enables organizations to leverage big data and big data technologies without sacrificing metadata or data governance. In part, the strategy leverages simple folder naming techniques similar to those employed decades earlier. As well, the strategy employs an automated process to register files being loaded into a big data environment.

The strategy automatically applies metadata as data is ingested in a big data platform. The approach does not slow the ingestion process nor does it limit the volume or variety of data that can be ingested. Although the metadata is not as rich as that provided by a traditional RDBMS, it provides enough information about the data to facilitate analysis and insight discovery.

One method of auto-registration involves building Java-based application programming interfaces (APIs) that require, as data is ingested, that metadata be associated with files being brought into the data lake.

If credit card transactions were being loaded by a merchant organization, for example, a data steward would first need to provide high-level registration information about the data. This task would be similar in complexity to configuring a system rather than coding a solution. It would focus on answering such questions as how often data will be loaded in the data lake and what kind of information will be included. The ingestrion registration program will provide metadata and rules for the ingestion of data into the lake. Once the steward has configured the program, millions of credit card transaction files can be loaded, with important metadata automatically associated with each file.

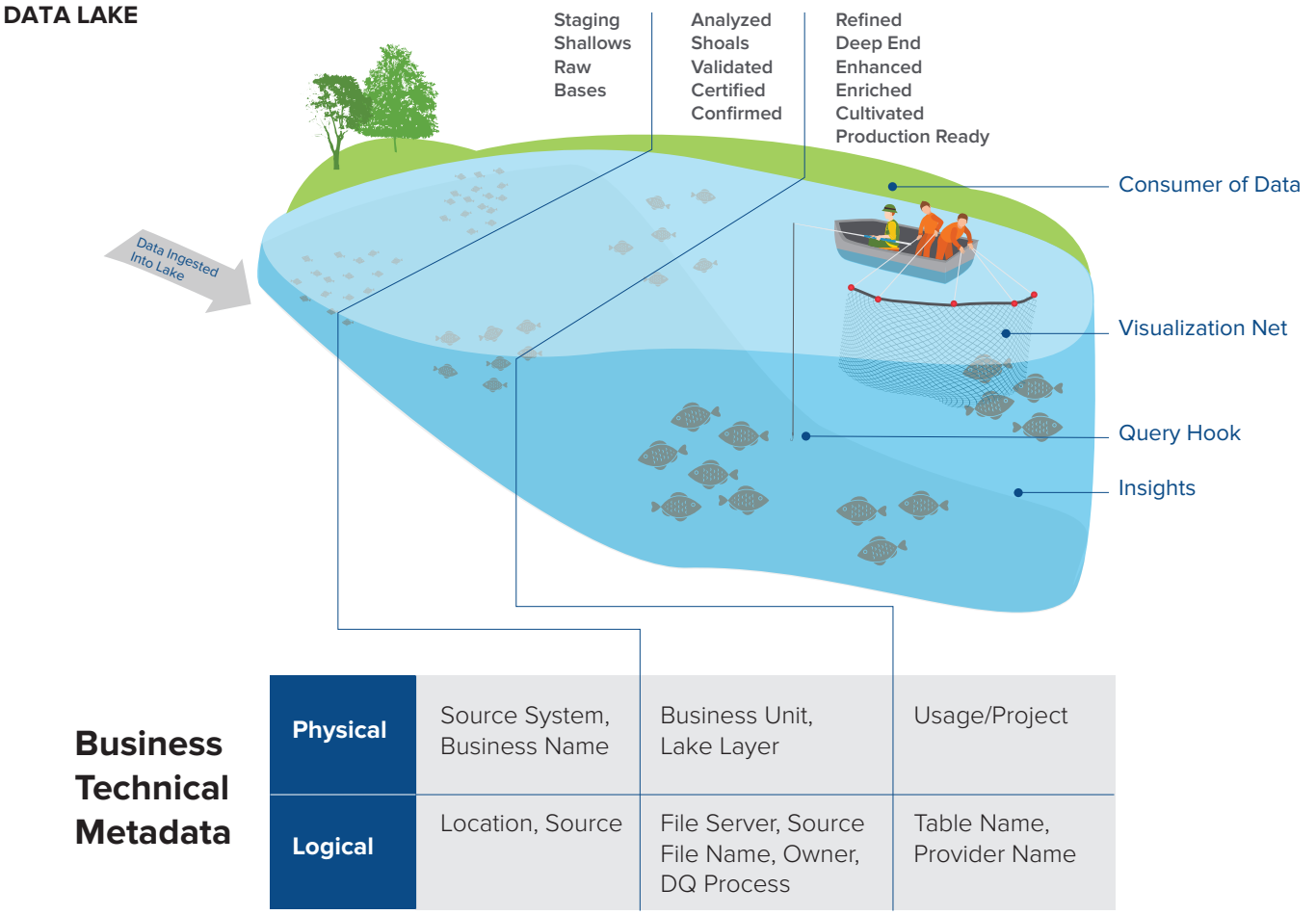
This enables data analysts and other data consumers to find the kinds of data they are looking for so they can process, analyze, and visualize the data. The process is akin to using a fish finder to facilitate fishing in a large body of water.

Informed fishing

Many anglers today use fish finders so that they can get a reasonable idea of where the fish in a large lake are concentrated.

A fish finder is not precise; for example, it does not indicate whether fish are bass, carp, or trout. It indicates only that it has detected something with the characteristics of a school of fish. Experienced fishermen typically already have some knowledge of the types of fish that inhabit the body of water and about the depths at which certain species tend to congregate. The fishermen can readily deduce from the information provided by the fish finder where to fish for particular kinds of fish.

This is similar to what auto-registration APIs can do for data analysts and other data consumers. The metadata produced by the APIs indicates that a particular area within a data lake is likely to be a good place to look for business insights. Thanks to the metadata, someone looking for specific kinds of information – for example, evidence of credit card fraud – will know know where the pertinent base data is located. With the benefit of some industry knowledge of fraudulent behavior patterns, a business analyst can then query the data in that part of the lake, looking for such patterns.



Business Technical Metadata

Physical	Source System, Business Name	Business Unit, Lake Layer	Usage/Project
Logical	Location, Source	File Server, Source File Name, Owner, DQ Process	Table Name, Provider Name

In lake ecosystems, fish are typically born in the shallows, where they grow and mature before migrating to deeper waters. A similar progression takes place in a data lake. Newly ingested data exists in the shallow area. As it is enhanced and matures, it flows to deeper areas. In most data lakes, data is enriched or enhanced when it is used for business models or reports. In cases where data is used for business purposes, more metadata can be associated with data sets. CapTech leverages both physical naming standards and logical metadata standards to enrich data flowing through the data lake.

To take the fishing analogy further: Just as a fisherman can cast a net or a hook to catch fish, so a data scientist can take a broad or a narrow approach in the search for data. Some inquiries are highly focused; for example, looking for a specific diagnosis code and treatment plan with a specific outcome. Other inquiries are broader; for example, looking for correlations between the frequency of a particular ailment and a specific gender/age band.

As data moves through the data lake, the CapTech solution provides for enhanced data governance capabilities. In addition to providing a base level of metadata, the ingestion registration process provides a level of lineage and can support both data security and data quality checks. For example, CapTech developed a data quality analysis dashboard for a financial services institution that was using our solution. The dashboard became an effective operational tool, helping the institution improve data quality.

The CapTech approach helps establish data governance without interfering with the ability of big data technologies to accommodate unprecedented data volume, velocity, and variety.

Conclusion

Historically, organizations have been able to establish effective data governance programs at least in part because metadata was built into the design of the data store. In the big data environment, this changes, as data is stored without metadata. While this enables businesses to quickly collect enormous volumes and varieties of data, it makes it difficult for data analysts and other data consumers to determine what data they have and then turn that data into business insights.

CapTech has developed a strategy that solves this problem. One method uses APIs that require that metadata be associated with files being brought into the data lake. In addition to providing a base level of metadata, the process of ingestion registration provides a level of lineage or traceability and can support both data security and data quality checks.

Metadata, traceability, security, and quality are key requirements of data governance. A sound data governance program can help organizations increase the efficiency of data analysis, enhance security and quality, and accelerate speed to insight.

AUTHOR BIO

Peter Carr | pcarr@captechconsulting.com
Peter Carr is a principal at CapTech, where he focuses on customer delivery projects and industry thought leadership in data management and analytics. If you have any questions or think CapTech can help your organization, please contact Peter Carr.

End Notes

[1] "3D Data Management: Controlling Data Volume, Velocity, and Variety." Doug Laney, META Group. Feb. 6, 2001. Available at http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf.

[2] "MapReduce: Simplified Data Processing on Large Clusters." Jeffrey Dean and Sanjay Ghemawat, Google. December 2004. Available at http://research.google.com/archive/mapreduce.html.

[3] "Big Data: Are you ready for blast-off?" BBC News. March 4, 2014. Available at http://www.bbc.com/news/business-26383058.

[4] "Google CEO Eric Schmidt: 'People Aren't Ready for The Technology Revolution.'" The Huffington Post. Aug. 5, 2008. Available at http://www.huffingtonpost.com/2010/08/05/google-ceo-eric-schmidt-p_n_671513.html.

[5] "The Difference between Lineage and Traceability." Collibra. Nov. 9, 2015. Available at https://www.collibra.com/blog/the-difference-between-lineage-and-traceability/.

[6] "Building the Data Warehouse," Fourth Edition. W.H. Inmon. John Wiley & Sons. October 2005.

[7] "The Data Warehouse Toolkit. The Definitive Guide to Dimensional Modeling." Third Edition. Ralph Kimball and Margy Ross. John Wiley & Sons. 2013.