Managing Critical Data Elements in Financial Services Through Agile Data Governance:

A Pragmatic and Scalable Approach

By Malcolm Chisholm, Ph.D.
Chief Innovation Officer, First San Francisco Partners

Information is your business.
Making it actionable is ours.

firstsanfranciscopartners.com

Introduction

Over the last decade, Critical Data Elements (CDEs) have become an important concern in the data governance and data management efforts of financial services. The growing criticality of CDEs has been galvanized by two main drivers:

(a) **The realization that the resources available within a typical data governance unit cannot scale to cover the production data landscape.** Even a medium-sized financial services organization will have millions of data elements distributed in database tables, files, documents, messages and other forms of data storage or movement. Somehow, the data governance unit must find the CDEs among these vast data elements.

(b) **Regulators have become increasingly focused on data.** In part this has been cross-industry, for instance, the General Data Protection Regulation (GDPR) in the European Union. However, within financial services, the regulatory focus on data has sharpened greatly over the years and continues to evolve. BCBS 239, for instance, is very specific about data management practices for CDEs in risk data aggregation with implications for both risk and regulatory reporting. Thus, there are certain aspects of data governance and management that regulators are forcing institutions to focus on.

In principle, these drivers seem reasonable. However, difficulties quickly arise when a practical response is considered. How are CDEs identified? What does it mean to “govern” them? What stakeholders are involved in CDEs, and what roles do they play? How do you know if you are governing all the CDEs you need to? How can you prove you are governing them?

The answers to these and other relevant questions have to come from the function responsible for data governance in the financial institution. Yet, today, data governance units are struggling to find these answers, and this is in large part due to the way data governance has traditionally been organized to do its work.

The Evolution of Data Governance

Data governance, as an enterprise function, arose after some 40 years of IT architecture being built that paid little attention to data. At the end of this period, corporate executives realized they had innumerable data problems that were not soluble by traditional technologies. They also became aware of highly successful new companies that put data at the heart of their business models, that might represent a long-term threat to traditional industrial sectors, including finance. The result was that around 2005, investment in data governance became widespread, albeit with little idea of how it would work in practice. Since that time, new imperatives have challenged data governance, as shown in Figure 1, and data governance has both matured and evolved to keep pace with shifting data-driven business needs.

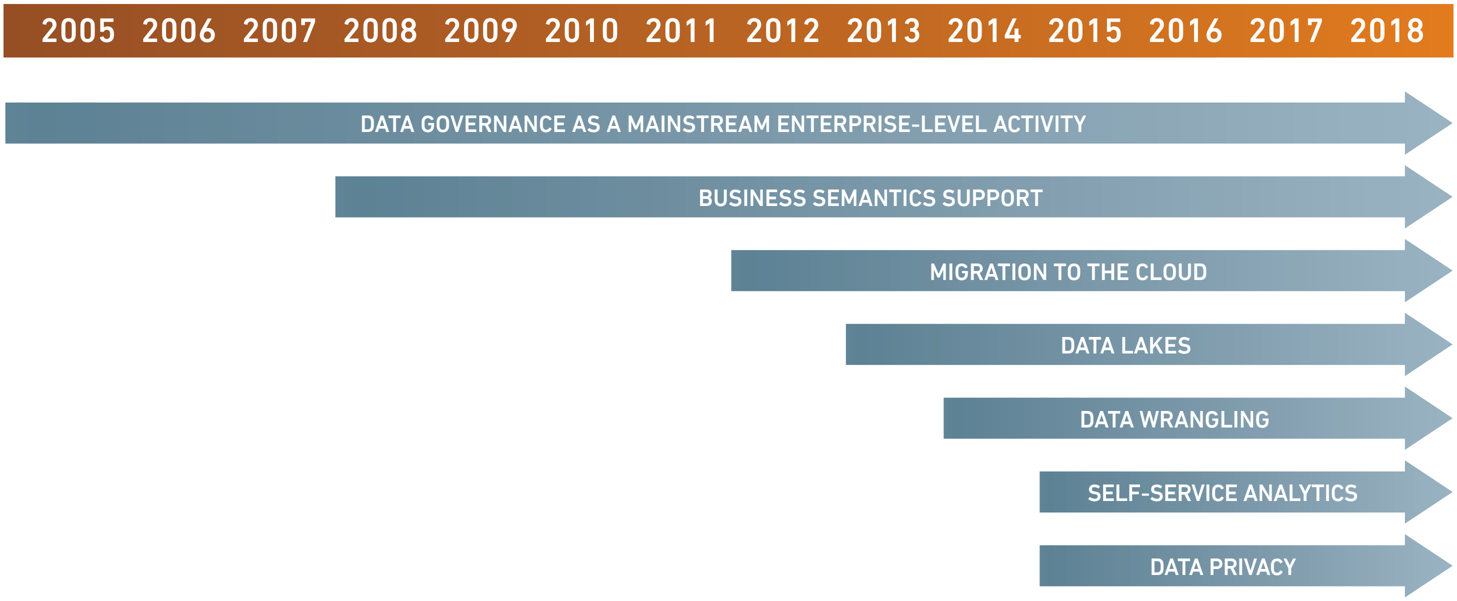


Figure 1: *Data Governance Timeline*

Data Governance Has Moved Through Three Generations:

Data Governance 1.0: Initially, data governance relied on a Data Governance Council, composed of executives and senior managers who had an interest in data, usually because they had problems with it. The Council would meet monthly and spawn working groups to solve the data issues in the enterprise. This approach was a widespread failure. Nobody on the Data Governance Council, or in the Working Groups, had much of an understanding of the art and science of data governance. Nor did they have the time to deal with the issues that confronted them as it conflicted with their regular responsibilities.

Data Governance 2.0: The realization that sound data governance required specialists led to the establishment of Data Governance Offices, staffed by full-time professionals. This was Centralized Data Governance. It was conceived as a horizontal function. Just as Human Resources set the rules for managing people, or Finance for managing budgets and expenses, so Centralized Data Governance would set the rules for how data should be managed. This is the model of data governance that predominates today.

Data Governance 3.0: While Centralized Data Governance has had successes, many of the challenges shown in Figure 1 have not been adequately addressed, and the full value of the corporate data resource still remains largely untapped in many organizations. Today, there is a perceived drive to "democratize" data – to enable anyone in the enterprise to use data to drive efficiency, effectiveness and risk reduction in their particular area. This approach is considered to be "Agile Data Governance," and it focuses on providing support for all staff who may potentially need to use data, rather than merely imposing rules on them (though rules are still important). This new generation of data governance is emerging but is already having significant impact. An important feature of Agile Data Governance is that it relies heavily on enabling technology, usually in the form of a data catalog.

This is the backdrop against which financial institutions are trying to govern CDEs, and it matters because if the overall data governance framework is inadequate, then the governance needs of CDEs will not be properly addressed.

Data Stewardship

One aspect of Data Governance 1.0 and 2.0 that must be understood is the reliance on *people*. Both of these phases of data governance have emphasized data stewardship as a means to implementing data governance. In practice, this has often meant simply labeling people without providing any clear understanding of what the role means, or providing any support to carry it out.

This has given rise to resistance by data stewards who worry they are there merely to be blamed when something goes wrong with the data they are responsible for, even if it is completely out of their control. An even greater problem is the patchiness or sparseness of data stewards across the enterprise in business and IT. Essentially, using people in this way does not effectively scale data governance, and in any case, the data stewards have regular responsibilities which do not include data stewardship. This means they cannot easily participate in defined data governance processes, or follow many rules created for data governance.

Regulatory-Driven CDE Governance Needs in Finance

To consider how CDEs should be governed, it is useful to look more deeply at the regulatory drivers, as these give clear statements of governance needs. BCBS 239, as previously noted, is particularly detailed in this respect. While BCBS 239 is only enforced for Global Systemically Important Banks (G-SIBs) and Domestic Systemically Important Banks (D-SIBs), the data governance principles in the regulation will – in all likelihood – ultimately be expected from all banks.

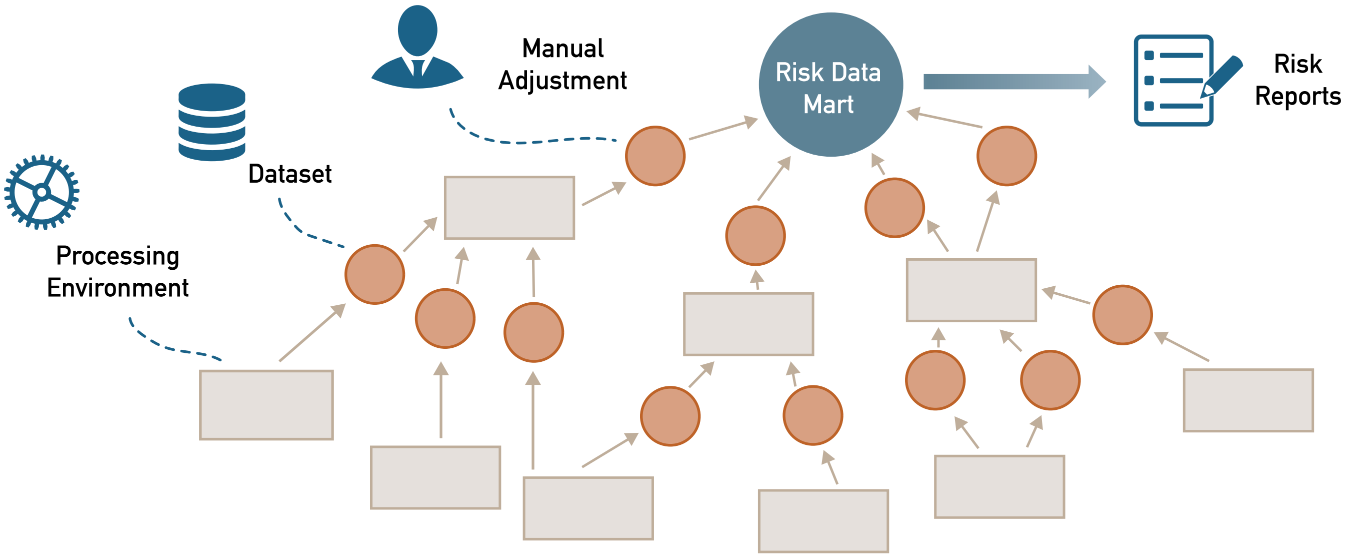


Figure 2: *Risk Data Aggregation. BCBS 239 is concerned about risk reporting, and Figure 2 illustrates how there is a chain of datasets that flow into a risk reporting environment from which reports are produced.*

© First San Francisco Partners 2019, All Rights Reserved

Principle 1 Governance - A bank's risk data aggregation capabilities and risk reporting practices should be subject to strong governance arrangements consistent with other principles and guidance established by the Basel Committee:

> 30. A bank's senior management should be fully aware of and understand the limitations that prevent full risk data aggregation, in terms of coverage (eg risks not captured or subsidiaries not included), in technical terms (eg model performance indicators or degree of reliance on manual processes) or in legal terms (legal impediments to data sharing across jurisdictions). Senior management should ensure that the bank's IT strategy includes ways to improve risk data aggregation capabilities and risk reporting practices and to remedy any shortcomings against the Principles set forth in this document taking into account the evolving needs of the business. Senior management should also identify data critical to risk data aggregation and IT infrastructure initiatives through its strategic IT planning process, and support these initiatives through the allocation of appropriate levels of financial and human resources.

***Figure 3:** CDEs are Identified in BCBS 239*

Within this context, CDEs are identified as a need in BCBS 239, as shown in Figure 3.

What BCBS 239 specifically requires includes:

- Identifying all CDEs involved in risk data aggregation
- Documenting the way in which the CDEs flow from operational systems to the risk reports
- Ensuring quality checks are in place to assure the integrity of the CDEs as they flow through the risk data aggregation chain
- Ensuring if manual adjustments are performed on CDEs, those operations are transparent and auditable

Centralized vs. Agile Data Governance in Identification of CDEs

Centralized Data Governance units have difficulties with the first step in governing CDEs, which is to identify them. Clearly, the Data Governance Office cannot know what CDEs exist, and they need to find someone else to identify them. What they can do centrally is to define criteria for judging if a data element is a CDE, and then send these criteria to individuals who will then apply them to the areas they work in to identify the CDEs.

One approach is to use automated workflows to solicit this information – but from whom? Individuals may well be designated as data stewards but this designation is often not very clear, and there is no guarantee that the data stewards will cover an area in which CDEs exist. Indeed, the coverage of data stewards in an enterprise is often very patchy. Even if a data steward can identify a CDE in terms of its business name, that may not be helpful in locating the database columns that represent the CDE.

© First San Francisco Partners 2019, All Rights Reserved

Agile Data Governance, by contrast, has the support of a data catalog where database columns, file fields, etc., are automatically inventoried. If the data catalog includes machine learning on the usage of data, then this may reveal certain data elements with very high usage, which are natural candidate CDEs. Thus, the increased focus on technology that comes with Agile Data Governance provides a better approach than Centralized Data Governance, which is more concerned with directing peoples' behavior.

Data that is not used is also easily discovered through machine learning. These tables and columns can be quickly eliminated from consideration in the governance of CDEs. By contrast, if a manual discovery process were to be attempted, it would be extremely time-consuming to determine which tables and columns were not used.

Change is also a particular pain point for Centralized Data Governance in terms of maintaining CDEs. New transaction and analytics systems are implemented periodically, and changes are made to existing ones. This means that lists of CDEs can become outdated without Centralized Data Governance even being aware. Periodic manual rework to check the situation of current CDEs and identify new ones is highly resource consumptive. The only real solution is via the machine learning approach inherent in Agile Data Governance.

Additionally, Agile Data Governance can take advantage of the social aspect of crowdsourcing information about data. Rather than directing specific individuals to provide, for example, definitions, Agile Data Governance encourages sharing of information and collaboration. This may reveal that individuals who a Data Governance Office never suspected of having knowledge about particular CDEs are actually Subject Matter Experts. Going back to the machine learning capabilities, these too may reveal the identify of individuals tied to the queries for heavily used CDEs.

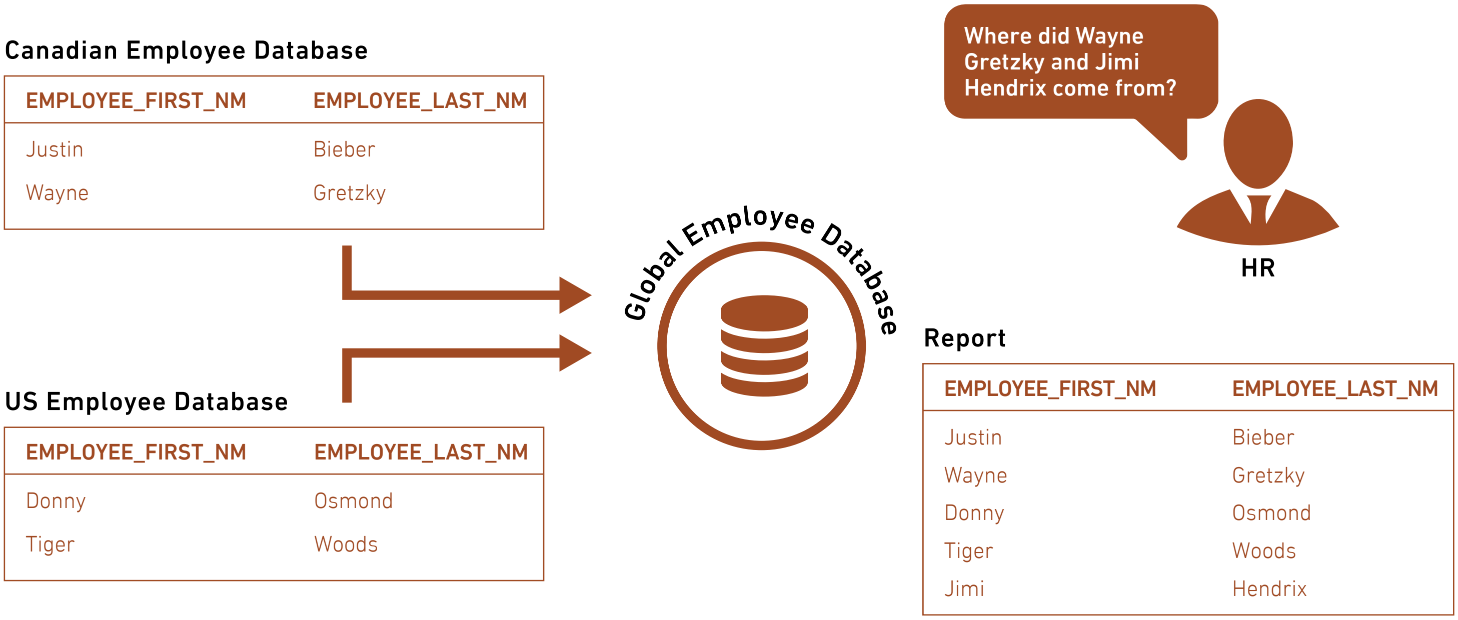
Data Lineage for CDEs

When we consider data lineage for CDEs, the difficulty arises that data lineage may mean different things to different people. For regulators, the assurance that data in risk reporting is truly coming from operational systems may be sufficient. For a data analyst, there may be a need to know if the data is coming from a trustworthy source. A business person in operations may want to know if all the data that is needed for a particular report has arrived today.

Centralized Data Governance has tried the approach of manual collection of lineage information to show how one data element flows into another data element. This is an extremely time-consuming process that yields a large amount of information, very little of which may be useful in a business context. There have been many expensive failed data lineage projects that tried to manually document data flows, only to produce a great deal of information which was still highly incomplete, and so technical that it was not usable by anyone except a narrow range of IT staff.

In recent years, Centralized Data Governance has tried to take advantage of tools that read the mapping specifications of software designed for data movement (ETL tools). But this also has the problem of only working for specific kinds of tools. For instance, custom stored procedures and dynamic HTML are not amenable to this approach. It is again a purely technical data element to data element level, which does not help when users have questions about the lineage of individual data values, as shown in Figure 4.

© First San Francisco Partners 2019, All Rights Reserved



Canadian Employee Database

EMPLOYEE_FIRST_NM	EMPLOYEE_LAST_NM
Justin	Bieber
Wayne	Gretzky

US Employee Database

EMPLOYEE_FIRST_NM	EMPLOYEE_LAST_NM
Donny	Osmond
Tiger	Woods

Report

EMPLOYEE_FIRST_NM	EMPLOYEE_LAST_NM
Justin	Bieber
Wayne	Gretzky
Donny	Osmond
Tiger	Woods
Jimi	Hendrix

Figure 4: Traditional Data Lineage Does Not Help at the Data Value Level

Agile Data Governance can use the machine learning described above, whereby logs of SQL queries are read to infer the data element to data element lineage. This is easier than the Centralized Data Governance methods and is a solution for some needs, but not all.

The risk data aggregation chain shown in Figure 2 is at the level of datasets, and datasets are not considered in data element to data element lineage. However, a dataset can be represented by a SQL query applied to the target where the dataset in ingested. Such SQL queries can retrieve record counts and timestamps to confirm that a dataset has been correctly ingested, and that the data is complete. Also, these SQL queries can easily be modified to enable questions like that shown in Figure 4 to be answered. This is a more valuable form of Data Lineage, and the modern Agile Data Governance tools are able to provide it. Also, the mindset of Agile Data Governance, which is to get closer to the data in a more technical way, supports this approach.

Manual Changes to CDEs

One aspect of governing CDEs that is clearly brought out in BCBS 239 is the requirement to track manual changes to CDE values as they make their way through the risk data aggregation chain and into the risk reports. Regulators are worried about “manual adjustments” to numbers that are then used in making important decisions in a bank. Of course, this matters to many managers and analysts inside a bank, as well as to external regulators.

© First San Francisco Partners 2019, All Rights Reserved

A Centralized Data Governance response to this need is to instantiate a workflow around the manual adjustment. However, it can be unclear who would actually start such a workflow. How would someone know where to find the workflow? Also, it could be asked if the workflow would serve its purpose, which would presumably be to inform stakeholders. Formally assigning stakeholders to a workflow is not easy, but a bigger problem arises when there are changes to the organization after a workflow is developed. Some stakeholders might move on to other positions in the bank, and some might leave the bank. Also, new communities of interest might appear with stakeholders who should be assigned to the workflow.

This overly prescriptive approach that attempts to direct the behavior of specific individuals is unlikely to succeed. Agile Data Governance, by contrast, does provide a more open alternative. Within a data catalog, individuals can state what manual adjustments they have performed on specific data elements. Of course, there will need to be a policy that individuals must do this, but the individuals do not have to be identified by a Centralized Data Governance function in advance. And the information they provide is tied directly to the unique set of information about the data element on which they have performed the adjustment. This in turn can be provided to stakeholders who can view the information, or may have set alerts on the information about the data element, to be notified when updates occur.

Distribution of Accountabilities

An important part of governing CDEs is understanding who is responsible for what in the management of the CDEs. This is the distribution of accountabilities for the CDE and is tightly linked to data lineage as different individuals can have the same role for the same CDE at different points in the risk data aggregation chain.

Centralized Data Governance has often focused on simply distributing a given role to a single individual for a CDE, irrespective of what happens in its data lineage. Clearly, this is an immature approach. A better methodology that has been established under Centralized Data Lineage is Data Sharing Agreements, as shown in Figure 5.

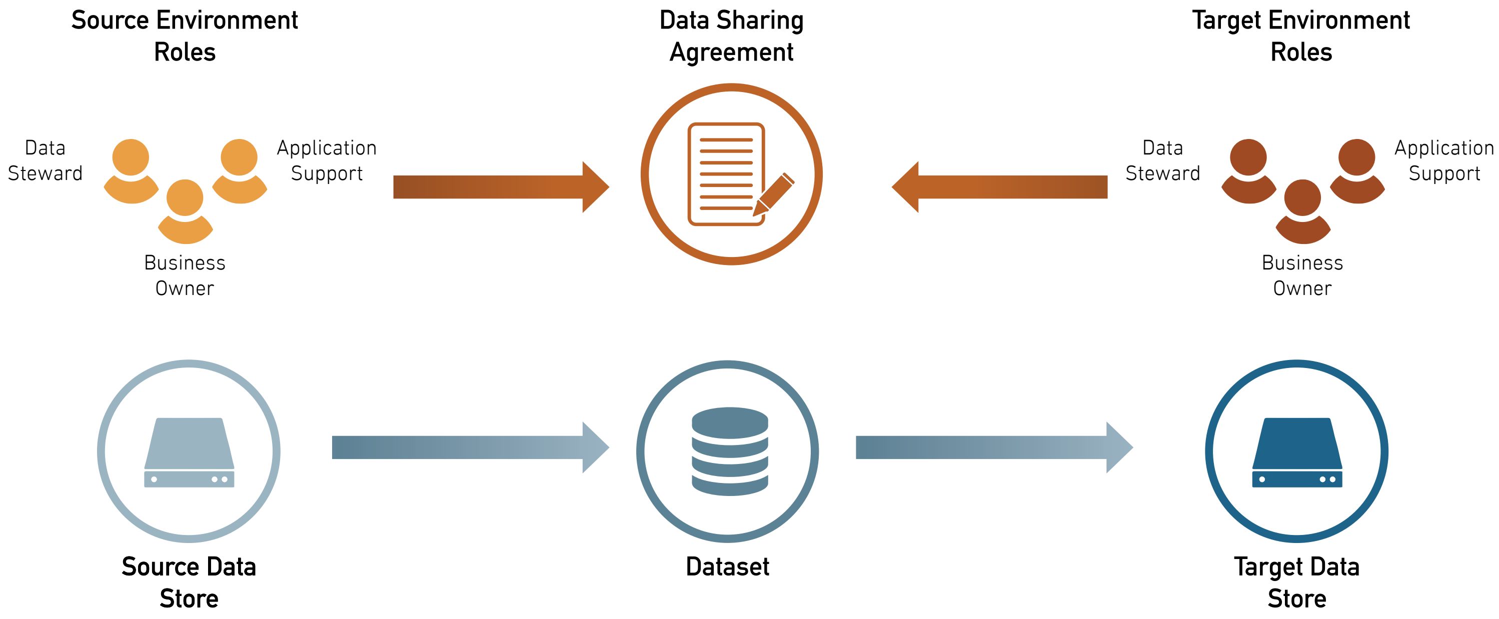


Figure 5: Data Sharing Agreements

© First San Francisco Partners 2019, All Rights Reserved

The idea of a Data Sharing Agreement is that when a dataset is sent from one team to another, the accountabilities are recorded on both sides. E.g. who is dealing with data quality on the source side, who can provide data definitions on the source side, who checks Service Level Agreement compliance on the target side? An advantage of this approach is that the agreement is at the level of the dataset, corresponding to the level shown in Figure 2.

While the Data Sharing Agreement methodology has had success, it is better implemented within an Agile Data Governance context. A major reason for this is that Centralized Data Governance is typically too far removed from the actual data. Thus, the information in a Data Sharing Agreement will be documentary in nature under this approach. Centralized Data Governance also has difficulty having a practical approach to datasets. Agile Data Governance, can easily manage datasets as live, executable SQL statements around which the governance roles can be built in a data catalog. There is also the possibility that machine learning can add to the list of stakeholders by analyzing query patterns associated with the dataset.

Data Quality

CDEs need to have data quality controls put in place for them. This is a regulatory demand, but it is also one that is echoed by risk managers as well as data analysts.

Through the lens of CDEs, data quality is made up of three components:

- **Data Quality Monitoring:** The detection of data exceptions
- **Data Quality Issue Management:** The determination of a resolution possibility for a data issue
- **Data Change Management:** The application of the resolution possibility to fix the data issue

With respect to Data Quality Monitoring, the current best practices are to develop specific Data Quality Business Rules that are then implemented in a Data Quality Monitoring tool, of which there are many in the marketplace. These tools run the rules against the production data landscape, often synchronized with regular scheduled batch jobs and report on data exceptions found.

Unfortunately, the Data Quality Monitoring tools have limited metadata capabilities. This makes them unsuitable for the development of the Data Quality Business Rules, since data definitions and governance responsibilities are not managed within them. While these tools manage the executable form of the Data Quality Business Rules well, they do not manage the declarative form (the form that is understandable to the business) to the extent that is needed.

Data catalogues can supply these missing capabilities. They have information about data elements, including business names (sometimes populated via Artificial Intelligence or AI). All the crowdsourced knowledge about the data element is also located in this one place. This makes it easier to start to compose Data Quality Rules for the CDEs. Such rules start out as English language statements, and these are then turned into SQL statements to prototype the actual Data Quality Business Rules. After this, the SQL statements, with some additional identifying information (like Rule Number), can be migrated to the Data Quality Business Rules Engine.

This approach is not possible in Centralized Data Governance where efforts will always be fragmented, and oriented to directing individual behavior, and where there is less of a technical orientation to data.

Data Quality Issue Management often does need a prescribed workflow, and here Centralized Data Governance often is successful, although it can fail if the workflows are overly constraining. Data Change Management is usually absorbed by the change management processes in IT which are generally very mature.

Enterprise Semantic Alignment

CDEs must, naturally, have adequate definitions. Centralized Data Governance has emphasized the need to build out Business Glossaries, with prioritization being given to CDEs. This top-down approach is, however, fraught with difficulties, some of which are not obvious.

A typical starting point is to take a recognizable Business Term and ask staff to come up with a single, unique definition for it. Sometimes this is successful, but very often there are subtle differences in semantics across the enterprise for data elements which have the same Business Term. E.g. “Finance Charges” may include an Overlimit Fee in one part of the enterprise, but not in another part. A more common example is that Marketing will include Prospects in “Customer,” whereas Accounting considers “Customer” to be someone who has paid, or owes money to, the enterprise.

Endless disputes can arise as a result, with the “single view of the truth” being a mirage due to the over-generalizations in working this way. Agile Data Governance, with its bottom-up approach, takes more of a view that the truth is in the data. By surfacing individual data elements and crowdsourcing knowledge about them, the information that is needed to truly understand CDEs becomes available. Again, the data catalog is a central organizing principle that enables this.

The crowdsourcing approach is very familiar to people today, thanks to Wikipedia. This encourages participation. There is, of course, the need for guidelines, and so governance is not totally absent. Additionally, recognition for contribution is a factor that promotes increased high-quality content and fosters an environment of knowledge management.

Conclusion

This brief survey of the how the needs of CDEs are met by Centralized and Agile Data Governance has shown that the two generations of Data Governance are really different. There is a growing realization that while Centralized Data Governance has been necessary and has had some successes, it is not sufficient for financial services organizations to get their data in order.

Agile Data Governance, by being closer to the data, effectively leveraging technology and by being scalable across the enterprise, represents a better overall framework to govern and manage Critical Data Elements.

Ready to get started with agile governance? Contact us to learn more.

http://www.firstsanfranciscopartners.com | info@firstsanfranciscopartners.com | 888-612-9879

© First San Francisco Partners 2019, All Rights Reserved