Comparing Two Approaches to Data Governance

A Global IDs Technical Whitepaper
March 20, 2020

Arka Mukherjee, Ph.D., Kirk Kanzelberger, Ph.D., Bill Winkler, and Stefanos Damianakis, Ph.D.
info@globalids.com

This document is part of our series of educational technical whitepapers that we release to share our experiences and lessons learned in enterprise data management with a wider public.

Executive Summary

Effective data governance requires that enterprise data be made *governable* — that is, less opaque and more transparent. What data is where, and how does it flow through the ecosystem? In this whitepaper, we compare two approaches to making data governable:

- ***The top-down approach***, in which humans exhaustively inventory the ecosystem and build a map of what data is where. This approach leverages human intelligence and familiarity with the data in systems and silos, but suffers from the incompleteness and fallibility of human knowledge, and the inability of human effort to scale with the size of the ecosystem.
- ***The bottom-up approach***, in which an automation platform builds a comprehensive map of the data landscape with high-level direction from a smaller number of humans. Humans contribute knowledge of business logic, but there is no need for human effort to scale with the number of systems, and automation can keep the map of the landscape current, as well as discover data in unexpected places.

We conclude our comparison in favor of the bottom-up approach as ensuring a comprehensive, more ambitious and objectively verifiable foundation for effective data governance.

Introduction

Data governance has become an imperative for today's data-driven enterprises. By *data governance* we mean the capability and processes for *ensuring* the quality of enterprise data in the broadest sense — its integrity, consistency, availability, reliability, and security. To *ensure* the quality of data, however, it is not enough for human beings to define governance policies. If the *governance* of data is to have any meaning, it must imply the capability of *enforcing* quality.

Data governance is a demand that stems from factors both internal and external to the organization:

- The sheer volume and complexity of enterprise data renders it opaque and subject to increasing "rot" — a growing proportion of data that is redundant, inconsistent, obsolete, or useless. As the ecosystem evolves, data quality steadily degrades.
- The opaqueness and fragmentation of data locked away in silos defeats attempts at analyzing, improving, and streamlining business processes, as well as leveraging data across the enterprise for business value and insight.
- Regulatory compliance hinges on the reliability and integrity of data. Further, data that is sensitive must be located and secured against both external and internal threats. If you don't know where the data is, where it came from, or where it flows to within the environment, you are vulnerable.

We can perhaps sum up this rationale for data governance with the phrase, *making data better*. You can't make data better if you are not in effective control of it — hence, data governance. The problem is that data that is fragmented and opaque is not governable. For effective data governance, then, the data ecosystem *must* become less opaque and more transparent, and as a consequence, more governable.

What is the best pathway to this governability of the ecosystem? In this paper we describe two approaches, which we will label the *Top-Down* and *Bottom-Up* approaches to data governance. We will outline the advantages and disadvantages of the two approaches, and in the end argue for the superiority of the Bottom-Up approach.

The Top-Down Approach

The top-down pathway to data governance is *human-oriented from the top down*, that is, from the business level (the "top") down to the level of physical systems and data. The following steps would be typical of this approach:

1. A *glossary of business terms* is created by humans using a data collection tool.
2. The business glossary assists humans in identifying the set of *important business objects* (Customer, Order, Vendor, etc.) to be subject to governance.
3. *Policies and data definitions* are created for these important business objects by humans (either from scratch, or taken from industry standards), specifying *logical* key fields and field types for each.

4. *Application owners and users are surveyed* regarding the presence of important business objects in the physical systems and databases they use or administer, in order to piece together a picture of the data landscape across all systems.

5. The humanly-built picture of the data landscape enables humans to visualize inefficiencies, redundancies, and other problems, and devise remediation strategies.

6. Many humans undertake the job of *reconciling* the physical representations of business objects in database tables with the logical definitions of those objects, in order to implement governance at the level of the data itself. This includes *implementing quality controls* at the database level, so that integrity rules and policies can be actively enforced and data quality monitored.

The Bottom-Up Approach

The bottom-up pathway to data governance is *automation-oriented from the bottom up*, while directed by humans at a high level. As in the top-down approach, humans define business terms and identify core business objects, but the bottom-up approach relies on a *data automation platform* to build the picture of the data landscape and automatically generate database-level quality controls. The following steps would be typical:

- *Important business objects* are identified by the data automation platform through read-only ecosystem scanning.

- Human beings then create *policies and data definitions* for these objects, either from scratch, or adapted from industry standards.

- The data automation platform then builds a *comprehensive picture of the data landscape* using the following automated steps:

    - *Data discovery and profiling*, based on physical scanning of data sources, which builds a repository of physical metadata.

    - *Machine learning classification*, which links physical metadata with the logical metadata derived from definitions of business objects, in two phases:
        - Physical columns are linked to logical field types
        - Patterns of physically-occurring key fields are linked to logical business concepts, thus detecting physical presence of important business objects wherever they occur in the ecosystem.

    - An *enterprise taxonomy* is the key artifact built by the previous step of automated machine learning classification. The enterprise taxonomy provides a

hierarchical view of the entire data landscape that organizes physical data logically, according to the business objects present in each data source.

- *Data lineage* adds dynamism to the view of the landscape by mapping out the flow patterns of important business objects, their splitting and aggregation, throughout the ecosystem.

- As a final step, humans *verify* the picture of the landscape generated by the data automation platform.[1]

- *Data quality controls* may now be applied in two ways, both enabled or assisted by the data automation platform:

  - Quality controls that apply standard policies to *known field types* such as address or currency fields may be generated automatically at the database level for physical columns that have been classified under those field types.

  - Quality controls based on *policies and business rules* may now be designed at the *logical* level — i.e., per key business object, abstracting from individual systems and their dialects — with the data automation platform taking care of fanning out the implementation of these controls to individual systems at the physical level.

- The comprehensive picture of the enterprise data landscape generated by this approach may be *refreshed periodically* and used as an up-to-date guide for remediation, rationalization, and other data-centric initiatives.

Comparing the Two Approaches

The *goals* of the top-down and bottom-up approaches are the same: effective data governance. They differ in the means used — specifically, in the role played by automation in making enterprise data governable. As Figure 1 illustrates, the essential problem lies in mapping business concepts (represented by logical metadata) to their physical counterparts and instances wherever they are found in the enterprise data ecosystem (represented by physical metadata).

[1] The enterprise taxonomy generated by the automation platform may reveal ambiguities in object definitions that need to be addressed, and the scanning for those objects repeated. This is a good example of what is possible when human intelligence and automation work in concert with each other.

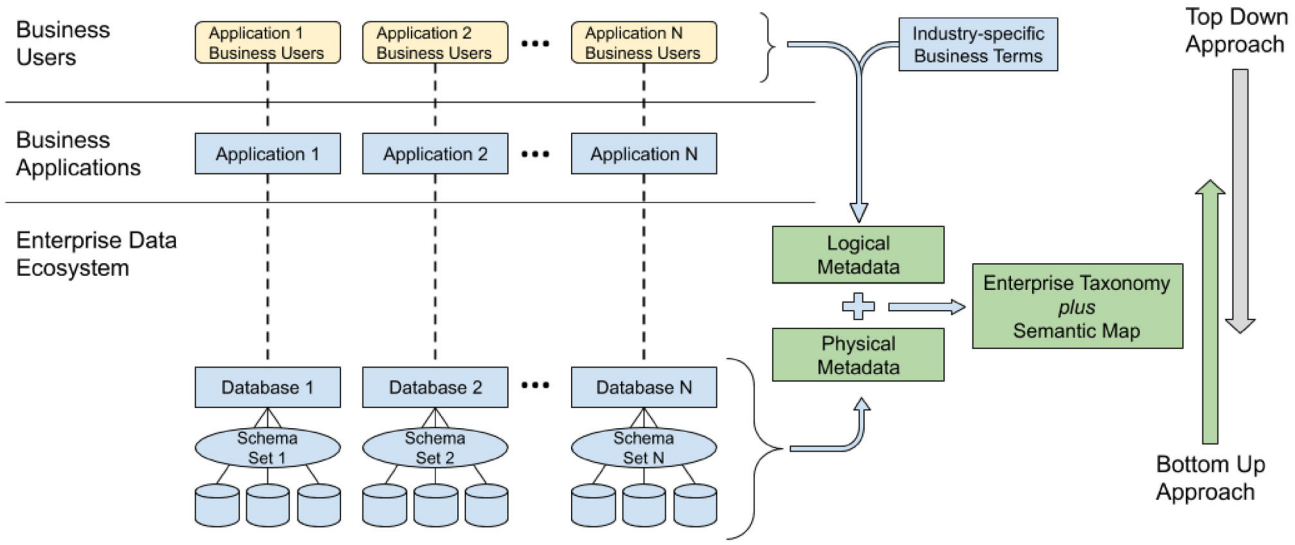

Figure 1. The data governability problem.

Top-down approach: Advantages

The top-down approach is driven by humans throughout. Its strength lies in human intelligence and expertise, together with the familiarity of application owners and users "on the ground" with their local portions of the data landscape. The maximum utilization of humans means that minimal IT resources are utilized, and there is little to no impact on operational systems, until the time comes to implement governance and quality controls at the database level.

Top-down approach: Disadvantages

The disadvantages of the top-down approach are the flipside of its strengths. Humans are intelligent, but they make mistakes. In practice, their knowledge of the data landscape may be quite incomplete, rendering their responses to surveys hypothetical to some degree. Humans can also frequently be unaware of the degree to which their knowledge is, in fact, hypothetical.

Perhaps the biggest disadvantages stem from the sheer *number* of humans required to be involved in any top-down approach. The problem can be looked at from two standpoints, *availability* and *scale*:

- The humans who will be involved in mapping the data landscape are the equivalent of subject matter experts in relation to the data that resides in the large number of applications and systems in the enterprise. The availability of this large number of experts, along with the time they will be required to spend, is highly questionable. Turnover makes it virtually certain that some proportion of them will no longer be working for the enterprise, so their detailed knowledge will be simply unavailable.

- Unless the whole project is very small, the scale of human resources required for the top-down approach involves costs that are prohibitive regardless of the size of the enterprise, since the number of required personnel scales with the number of applications and systems.

In relation to both standpoints, it should be kept in mind that the human resource requirements in the top-down approach will be *ongoing*, since they are the same human resources needed for keeping the view of the landscape and governance controls up to date.

Bottom-up approach: Advantages

The bottom-up approach is automation-oriented. It does not seek to replace human intelligence completely, but rather to bring about a synergy of automation and human intelligence that leverages the strengths of both and remedies the weaknesses of the top-down approach.

In the bottom-up approach, humans remain in charge of the business logic and the high-level definitions in which that logic is expressed, while the data automation platform builds the taxonomy and semantic maps that connect high-level concepts with their physical counterparts and instances that occur in many variations throughout the data landscape, as illustrated in Figure 2.

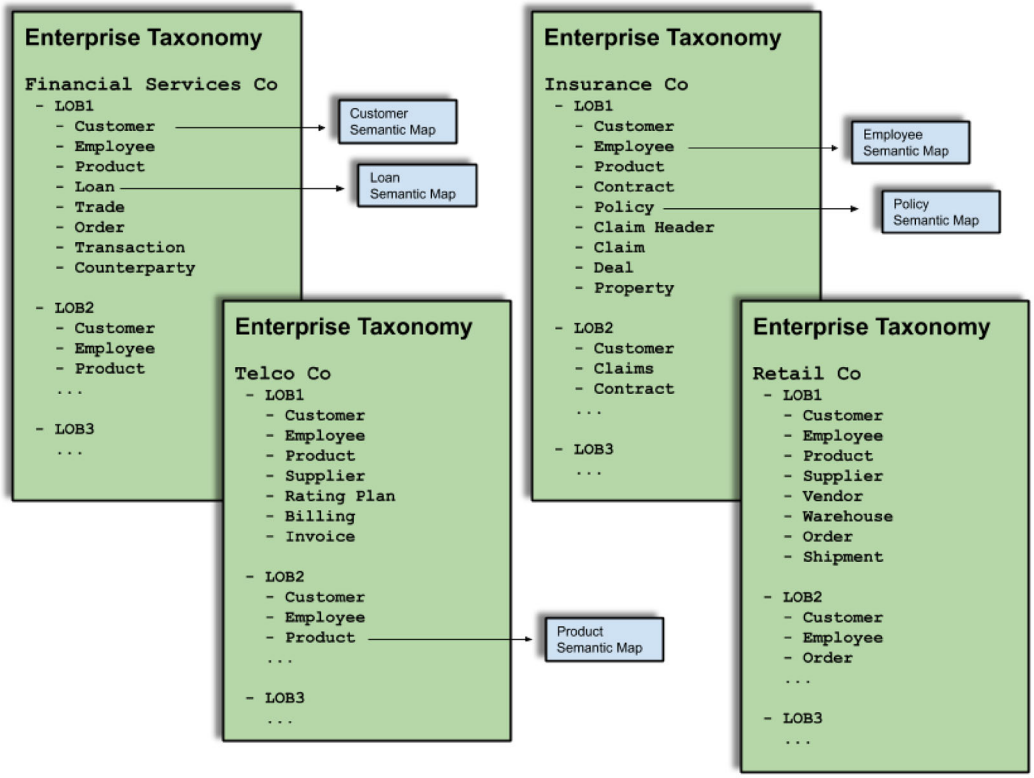

Figure 2. Four different enterprise taxonomy and semantic maps (illustrated hierarchically).

The data automation platform implements a systematic approach to classifying and organizing data. By discovering the data that resides in unexpected and forgotten places, comprehensive data catalogues can be created. The platform thus helps ensure that critical issues are not overlooked because of the incompleteness and fallibility of detailed human knowledge. Finally, regulatory compliance benefits from objective verification of where data actually lives and how it actually flows.

In the bottom-up approach, there is no need for human resources to scale with the number of systems and applications, since automation takes care of the multiplier effects. The same advantage applies in the design and deployment of quality controls: controls can be designed at the level of the business object being governed, while automation handles the fanout of these controls at the level of physical instances.

Finally, machines can keep the view of the landscape current and also keep humans informed of the need for additions or refinements at the level of the logic of business concepts.

Bottom-up approach: Disadvantage

The one disadvantage of the bottom-up approach is that the data automation platform requires read-only access to databases across the enterprise. It should be noted, however, that the (read-only) scanning of the ecosystem can be incremental and scheduled during non-peak hours, and so has no impact on the performance of these databases.

Summary and conclusion

In summary, the bottom-up, automation-oriented approach seems to us clearly superior from the standpoint of both cost and effectiveness. It combines human expertise at the level of business logic with automated discovery, classification, and mapping of the ecosystem to build a multilayered, renewable view of the data landscape that is objectively grounded in the data itself. This opens up a more ambitious pathway to data governance that can solve larger problems.

The costs associated with a data automation platform, including the impact on operations, are offset by the savings in human resources as well as avoidance of the costs associated with human error and oversight which can pose significant risks in the areas of regulatory compliance and security.

The following two tables summarize the advantages and disadvantages of the top-down and bottom-up approaches:

Table 1: Top-Down Approach Pros and Cons			
Pros	+ Human intelligence, expertise + Knowledge of users/app owners + Minimal IT involvement + Minimal impact on operations	**Cons**	- Human fallibility - Incomplete, hypothetical knowledge - Does not scale, constrained by human effort - Vulnerability to employee turnover

Table 2: Bottom-Up Approach Pros and Cons			
Pros	+ Automated discovery of where data is, what it is, and how it flows + Automation scales without adding human effort + Implementation of governance can be designed at a higher level and deployed automatically	**Cons**	- Requires installation of software either on-prem or in-cloud - Administrators must permit read-only access to systems