Monitoring and Metering the health of Enterprise Data

An Emperative for Governance and Data Quality

White Paper by
Malcolm Chisholm Ph. D.
2014

iCEDQ

Executive Summary

1. Data governance is now seen as a priority in nearly all enterprises. Data-related in production data environments have grown to a point where they must be addressed. Yet data governance is still relatively immature and not defined precisely.

2. Efforts at data governance have often focused on establishing councils or committees that set up procedures to access or use the data. Executive management appears to find such approaches bureaucratic and unable to address the underlying data problems.

3. A significant area for improvement is data quality, and executive management sponsor data governance in the hope that it will address data quality. However, data quality is itself poorly defined and usually a collection of issues that unlikely has one common fix.

4. The whole approach of first trying to identify types of data quality issues and then figuring out ways to remediate them is questioned here. It is proposed that monitoring (detecting exceptional events) and metering (gathering metrics on the health of the data) is a logical precursor to fixing any problems.

5. Monitoring and metering is a fundamental engineering principle that is used in process control. Yet it is hardly ever used in production data environments. This is a significant missing element in data management.

6. Data profiling tools are sometimes thought to be adequate for monitoring and metering. However, they are much more oriented to source data analysis in the early phases of data-centric projects, and are tools designed for analysts. These tools are not built to be integrated with processes in production environments. Though they could be used in data cleanup projects, but that is a poor way to address the problems of data quality.

7. Monitoring and metering produce feedback. This feedback must be routed to appropriate stakeholders in a data governance framework. Today, only production control operators monitor overall environments and they are more oriented to whether tasks are running within SLA than the health of data. The necessary data governance framework to process feedback from monitoring and metering typically does not exist. It will be a major challenge to create it, but it must be done.

8. In terms of tools for monitoring and metering, the only viable approach is to use business rules engines. These permit quality checks to be established quickly without disturbing production applications, and are not reliant on the systems development life cycle. Rules engines also provide the independence that is required for auditing. The essential characteristics of rules engines for monitoring and metering are discussed. The need for an appropriate methodology in the adoption of rules engines is also highlighted, and without such a methodology the usefulness of rules engines will likely be limited.

9. The way in which a rules engine must store the metrics it gathers is described, along with the need for it to integrate other master metadata that categorizes the production data landscape. The need for stakeholder metadata to be available to such a tool is examined.

10. The requirement for presentation of the data collected by a monitoring and metering tool is covered. The way in which the presentation layer can be used for data governance is illustrated.

The Dawn of Data Governance

The past few years have seen a surge in the demand by enterprises to unlock the value in their data. The decades-old focus on process-centric applications designed to automate business processes has shifted to a data-centric focus with data integration at its heart. Yet this new paradigm has been beset with problems. Five decades of organic growth in IT architectures, an emphasis on acquiring technology, and vision confined to individual projects has contributed to a "data mess" in most enterprises. This mess is manifested in many ways, principally as failure to achieve data integration and lack of trust in the outputs of business intelligence (BI) environments. Additionally, the mess weighs so heavily on enterprises that they cannot change their architectures to meet new business challenges because nobody really knows how the data is being processed inside them, and what might break. An example of how bad the situation has become is the gradual realization that mergers and acquisitions, may be a severe long term risk for enterprises because the legacy IT architectures can never be properly combined and creating a new environment even less amenable to change.

While the architecture challenges are long term, the failure to integrate and lack of trust in BI environments typically become apparent within the lifespan of individual projects. The natural response of IT has been to try to solve these problems by technology. Yet promises of software that could somehow clean the data have never lived up to their expectations, though they have made important contributions. Gradually, the center of attention has shifted to the way in which the data itself is managed. This new area - the way in which data is managed in production environment - is data governance. IT has never really done this before; it has traditionally been concerned with building, implementing, and managing infrastructure. Nor have the business users; division of responsibilities has meant they have paid attention to getting their assigned tasks done using the IT infrastructure in whatever way it could be made to work.

What Is Data Governance and how is it done?

So what is data governance? Today, most of the answers are very high-level and not that actionable.

Definition and Source
"Data Governance is the exercise of decision-making and authority for data-related matters." Data Governance Institute (www.datagovernance.com)
"Data Governance is a system of decision rights and accountabilities for information-related processes, executed according to agreed-upon models which describe who can take what actions with what information, and when, under what circumstances, using what methods." Data Governance Institute (www.datagovernance.com)
"Data governance is a quality control discipline for adding new rigor and discipline to the process of managing, using, improving and protecting organizational information." The IBM Data Governance Council Maturity Model: Building a roadmap for effective data governance (http://www-935.ibm.com/services/uk/cio/pdf/leverage_wp_data_gov_council_maturity_model.pdf)
"Data governance is the practice of making enterprise-wide decisions regarding an organization's information holdings". Federal Enterprise Architecture DRM Data Management Strategy (http://web-services.gov/DraftDRMDataManagementStrategy-021604Correct.pdf)

Table 1: *Sample of Definitions of Data Governance*

This is not surprising, because data governance is a new competency that still has to mature. However, when we ask how data governance is done, the answers are more problematical. There seems to be an emphasis on setting up organizational bodies that "make decisions" or "set policies" about the data. Anecdotal evidence suggests that executive management in nearly all enterprises view these developments with alarm. Although well-intended, data governance councils do not deliver solutions to the problems of integration and lack of trust in BI environments - at least in timeframes acceptable to executive management. Instead they tend to impose checks, balances, and restrictions that are viewed as an additional expensive layer of bureaucracy.

Part of the problem seems to be that the choice of the word "governance" implies the analogy of government-like structures to manage data. Perhaps at a very high level this is true, but what are we supposed to do at the operational level where the problems with the data are created and manifested?

It is also doubtful that an umbrella concept like data governance can ever provide a single solution that will fix all aspects of the "data mess" referred to earlier. A more fruitful approach is to break down the problem into its components and address them individually. Specific problems require specific solutions.

Monitoring, Metering, and Data Quality

Deciding what to tackle first in the "data mess" is not easy, but everyone would agree that data quality is an important area. Yet data quality itself is not a single concept. A good deal has been published on this topic, and Table 2 gives some idea of the many kinds of data quality.

Dimension	The extent to which the information is or has ...
Accessible	available, or quickly and easily retrievable
available, or quickly and easily retrievable	Accuracy can be thought of as freedom from mistake or error.
Accuracy	Accuracy exists when reality and what is recorded as data are in agreement.
Appropriate Amount	the amount appropriate for the task at hand
Atomic	only one fact in a given field
Believable	regarded as true or credible, transparent (errors not hidden)
Complete	not missing and is of sufficient breadth and depth
Concise	compactly represented
Coverage	How much of what is available has been recorded.
Conformity	presented in the same format, e.g. dates
Consistent	values in the fields do not conflict (e.g. name=John gender=F)
Coherence	integrity, agreement with related data
Interpretable	in appropriate language, symbols, or units, and definitions clear
Meaning	the information recorded with a field agrees with the definition of the field
Objective	unbiased, unprejudiced, and impartial
Redundancy	Single instance is the ideal; sometimes redundancy is accepted for performance reasons
Relevant	applicable and helpful
Reputable	highly regarded in terms of its source or content
Secure	access is restricted appropriately to maintain its security, authentication, privacy, IPR, copyright, and legal or regulatory requirements
Timely	sufficiently current or up-to-date for the purpose. 'Float' is the lag between a fact being recorded in System A and it being passed to System B
Understandable	easily comprehended
Usability	ease of manipulation to apply to different tasks
Value	beneficial and provides advantages from its use; from the other point of view, the risk in not having important data
Validity	Passes audit controls which rule out impossible or unusable values or impose business policies

Table 2: *Information Quality Attributes*
(Data Quality Assessment. Leo L. Pipino, Yang W. Lee, and Richard Y. Wang Communications of the ACM April 2002/Vol. 45, No. 4ve. 211. http://web.mit.edu/tdqm/www/tdqmpub/PipinoLeeWangCACMApr02.pdf)

There is still no settled taxonomy of all the different kinds of data quality, but we can easily appreciate from Table 2 that no single solution is going to address them all.

Yet, Is trying to categorize and devise solutions to the different kind of data quality the correct approach? It might seem a natural approach to the IT mindset, but it is missing an essential element: monitoring and metering. Monitoring is the detection of events and metering is the collection of metrics. Without monitoring and metering, the entire production data landscape is treated as little more than a black box, and the health of the data within it will remain a mystery until the data produces some event that is recognizable as due to a quality issue.

In other words, why are we trying to fix data quality problems without any true situational awareness of data quality in the production environment?

Monitoring and Metering versus Profiling

When this question is posed, the usual reaction of IT is to say that there are many data profiling tools on the market and that IT uses them in periodic cleanup exercises for data quality problems.

It is true that there are many tools for data profiling, but these tend to be oriented to source data analysis (SDA). SDA is an essential phase in a data-centric project that involves rediscovering the structure and content of data sources that will be moved to a target. Part of SDA involves an assessment of the quality of the source with respect to the uses that will be made of the data in the target. The problem is that if a source of data is not a candidate for a data-centric project, it is not going to be profiled. Also, it will only be profiled in the context of what is needed for the target, not in terms of what are acceptable quality parameters for the source itself.

Data profiling tools tend to be very specialized and are used by analysts, rather than by the stewards. This limits the individuals who can use them and type of feedback they can produce. Such tools are oriented to be used in the early stages of projects, and not as an integral part of a production environment. Furthermore, they tend to be more of a shotgun approach, performing a large number of scans at a time on a database, which can cause performance problems if not scheduled carefully.

The second part of the objection is that IT conducts periodic data cleanup exercises, which is even more problematic. Of course, such projects attempt to improve the situation, but they have limitations on what they can achieve. A major issue is that the action of cleaning a database may itself introduce inconsistencies. Reports created prior to the cleanup - admittedly with bad data - may not match reports produced after the cleanup. Similarly, downstream applications that consumed bad data prior to the cleanup are often not in the scope of the cleanup project, and thus find themselves with greater reconciliation issues after the project is completed.

Thus profiling tools used in the context of periodic data cleanup exercises are not really a substitute for real monitoring and metering.

Perhaps the most astonishing fact, however, is that IT has been blind for so long to the need for monitoring and metering for data health, and yet this is a fundamental engineering concept. For instance, Figure 1 illustrates a centrifugal steam engine governor.

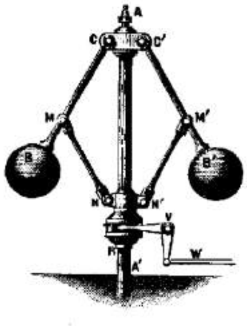


Figure 1: *The Centrifugal Steam Engine Governor*

This device, invented by James Watt, was essential for the safe operation of steam engines. Steam power rotated an axle to which were attached two heavy fly balls. If the steam pressure got too high, the speed caused the fly balls to rise up, and the arms attached to them opened safely valves releasing the pressure. The apparatus could be adjusted to react to a range of pressures. Prior to this invention, overheated boilers simply exploded.

A more up to date example of this principle can be found in a 2009 white paper by Intel entitled as *Increasing Data Center Efficiency through Metering and Monitoring Power Usage* (http://download.intel.com/it/pdf/Increasing_DC_efficiency_through_metering_India_final.pdf).
In this paper, an approach to improving energy efficiency at one of the company's older data centers in India is described. The authors summarize the project as follows:

> *We developed methods for identifying measurable efficiency improvements and placed instrumentation to continuously track power usage effectiveness (PUE), the key metric of data center energy efficiency. Using PUE metrics allowed us to make decisions that increased efficiency, helped achieve optimum data center facility utilization, and provided data we can share with other Intel facilities around the world to proliferate energy savings.*

Several diagrams illustrate this approach, such as the one shown in Figure 2

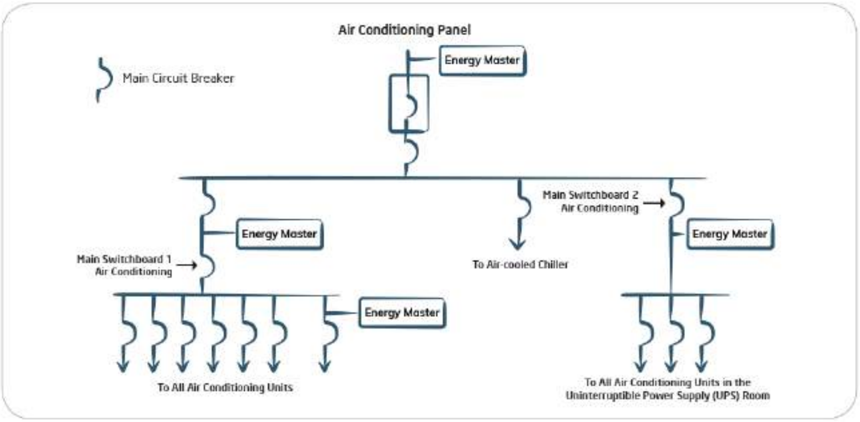

Air Conditioning Panel
Energy Master
Main Circuit Breaker
Main Switchboard 2 Air Conditioning
Energy Master
Energy Master
Main Switchboard 1 Air Conditioning
To Air-cooled Chiller
Energy Master
To All Air Conditioning Units
To All Air Conditioning Units in the Uninterruptible Power Supply (UPS) Room

Figure 2: *Metering of Cooling Power at an Intel Data Center*
(http://download.intel.com/it/pdf/Increasing_DC_efficiency_through_metering_India_final.pdf)

The irony of this is that monitoring and metering of the power consumption of a data center was a priority, yet there is hardly ever any attempt to monitor or meter data. Why this is so is not clear. Perhaps it is the thought that audit checks will catch all data errors.. and that these have all been approved by the users and tested prior to production deployment. In all other engineering constructs where process is involved, monitoring and metering is an integral part of the production environment. Imagine an oil refinery that had been built to specification and passed its initial test, but that lacked any form of monitoring and metering, would never be allowed to operate. Yet this is exactly what production data environments are like.

The Nature of Feedback

Monitoring and metering produces feedback. However, it is not true that feedback does not occur in the absence of monitoring and metering. Unfortunately, the kind of feedback that gets produced in the absence of monitoring and metering is often unhelpful and in its own way can be damaging.

Let us continue with the analogy of the oil refinery with no monitoring or metering. Suppose that one of the feed lines sprung a leak. How would this be noticed? Maybe there might be a reduction in an output that was eventually noticed by buyers of the refined products, and led to a search for the cause. Or perhaps someone walking around the site might just happen to see the leak by chance. Or perhaps the leak might be of an inflammable liquid that eventually caught fire - and that would be noticed.

Obviously, these are all undesirable forms of feedback. It would be much better to have gauges for measuring flow, pressure, fluid levels, and so on, and sensors that raised an alarm if measurements reached predetermined danger levels. This kind of feedback is very timely, usually catches problems before any real damage is done, and points to the location of the problem with some specificity.

There is no such general approach in production data environments. The only feedback that is provided is when users (or customers or regulators) find something suspicious, or wrong, and point it out. Or perhaps IT staff in the course of their daily duties notice a problem and initiate action to correct it.

Feedback not only has to be available, but it should be aligned with the process. Why use a profiling tool to scan an entire table of portfolio position data when the only change that happens to it is daily position data and the existing position data never changes? The lack of integration between profiling tools and processes is yet another reason why these tools are difficult to use for monitoring and metering.

Not only does feedback have to be aligned to process, but also has to be handled correctly. We are all familiar with images of Network Operating Centers (NOCs), NASA's mission control, and control rooms of facilities such as nuclear reactors. And of course there are production control units in IT, who monitor systems consoles for messages about the state of tasks or jobs that are running. However, these units rarely receive information about the data. It is nearly always about whether a task has failed unexpectedly, or has not started, or is taking too long, or is out of SLA. For some kind of feedback concerning data, it is appropriate for production control to be involved. This may be necessary if a data quality issue critically impacts a process. But other kinds of feedback should be routed to users, or business analysts who support them, or other stakeholders who deal with the data. We simply cannot expect production control units to handle all the feedback that might come from monitoring and metering of data health - it would be overwhelming for them. The problem is that we have no clearly recognized set of roles and responsibilities for handling such feedback. It is a major challenge for data governance to set this structure up, and it is a challenge that cannot be delayed.

Monitoring and Metering Tools

If monitoring and metering of data health is essential, what tools will be used to do it? What will be the equivalent of pressure gauges, heat sensors, fluid level monitors and the like?

In thinking about these problems, the IT mindset tends to be dominated by the systems development life cycle (SDLC). A natural response, therefore, is that monitoring and metering should be built into every application. Audit validation checks are often thought of as performing this function. However, in data-centric projects there is a lot of data movement, and in this context the data does not have the same relationship with a steward as when a steward is entering data into a screen. Of course this does not mean that monitoring and metering should not be part of extract-transform-and-load (ETL) environments. Checks should be incorporated into them. However, there are strong reasons for having monitoring and metering as a separate component in the architecture as illustrated in Figure 3.

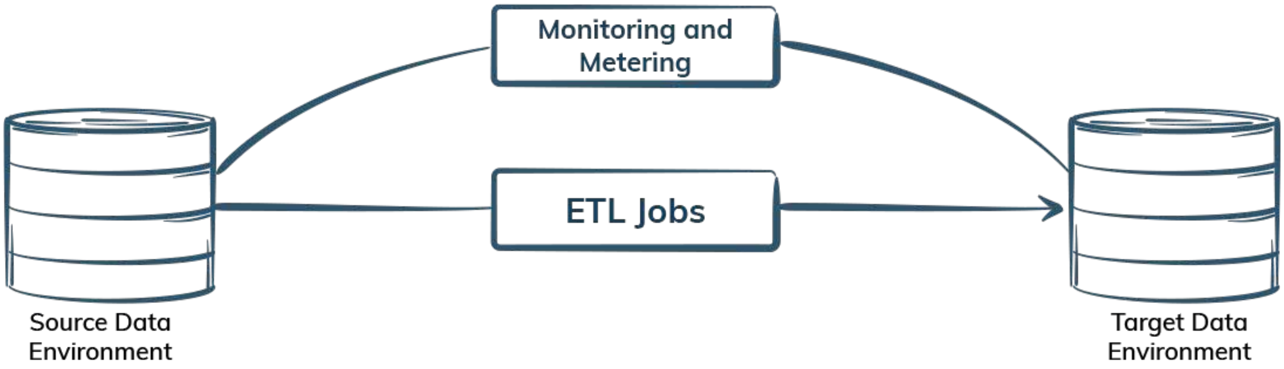


Figure 3: *Monitoring and Metering Related to ETL*

There is an inherent problem in an ETL process performing its own monitoring and metering. It may be flawed in a way that it is not detected from within the process. This is a major reason why auditing exists. A process may be trusted, but it needs to be verified, and verifying that the process is working well should be done independently of the process itself. By having independent monitoring and metering tools, therefore, we also achieve auditability. That is, we have built true auditing of data health into the architecture. Self-assessment, which is essentially trusting the ETL jobs, is not verification, and is not an acceptable form of auditing.

A second problem that independent monitoring and metering tools solve is reliance on the systems development life cycle. Suppose that we built all monitoring and metering into the ETL jobs in Figure 3, and sometime after the production implementation we uncover a new data quality problem. In this scenario, a new data quality check has to be built into the ETL jobs. This will involve design, development, testing, and production implementation. Yet the main purpose of the ETL job is to move data from source to target. Implementing the new check may interfere with how this is done, so extensive regression testing will be needed.

By contrast, implementing the new check in a monitoring and metering tool will not interfere with data movement, because the tool is not doing that kind of work. We will just need to know that the new check is functioning as it should and that is all the testing that needs to be done. Thus, we should be able to react far more quickly to issues that are found via feedback from the users than is the case with the normal SDLC.

This approach can also be beneficial in other ways. Analysts may think of new checks that should be performed after production implementation of the ETL jobs. There will be very little appetite to change the jobs at this point, but implementation in an independent monitoring and metering tool may be much more feasible. Another issue that is frequently overlooked is that the business changes in small ways faster than IT can keep up. As operations adapts existing applications to these changes they build new rules in a monitoring and metering tool.

The Business Rules Approach

From the above we can see that adaptability is a quality attribute that is needed for any monitoring and metering tool. Data content and structure can change in ways that affect quality, and the tool needs to keep up. In this respect, there is a sharp difference between the monitoring and metering of data health, and engineering hardware such as pressure gauges and heat sensors. The latter are not really adaptable in the way needed for data.

Adaptability of this kind is not provided by the traditional SDLC. The only architectural pattern that does provide it is the business rules approach. In this pattern, business analysts define business rules and the business rules are immediately executable in a target environment. IT staff, at least in theory, do not participate in this activity, so there is no programming, testing, and production migration.

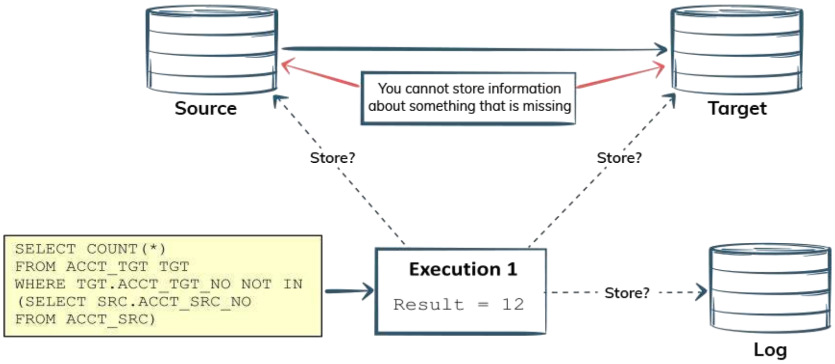
The kind of tools that provide this functionality are called business rules engines (BREs). Their approach can vary quite widely. Some are interpretive, while others generate code. Some are oriented to natural language rules, while others require the use of scripts that are closer to programming languages. Perhaps the most fundamental split, however, is in their orientation to some kind of deduction, such as deriving a credit score, versus doing simpler calculations and derivations. The latter class of BREs are more common and is the one that will provide the kind of functionality needed for monitoring and metering.

BREs also serve as independent tools, and so they provided the architectural quality attribute of auditability. They are not a part of the process that is doing the real work of manipulating data. In fact, BREs are quite often used for data quality tasks, although it is rare to find them offered as pure monitoring and metering tools. For instance, they can be a component of data profiling tools.

However, the business rules approach is not the same as BREs. There is a whole methodology that has to be wrapped around the use of BREs. It is all too easy to develop thousands of rules within a few months and then begin to drown in their administration. For instance, if nobody can remember, or find, a rule for a particular check, the same rule may be reinvented, perhaps many times, and perhaps with unacceptable variations. If BREs have an Achilles' Heel, it is probably rule testing. Business analysts lack the support of the IT community when they work with BREs because they are doing so outside the SDLC. Rules need to be tested carefully before they are deployed. And there are many other aspects of methodology that are important, but cannot be covered here. It will suffice to say that adoption of a BRE for monitoring and metering, and metering, but without a methodology, will be problematic.

Metrics and Metadata

There is another characteristic of a BRE that needs to be considered: its capacity to store the results of monitoring and metering.


Source
Target
You cannot store information about something that is missing
Store?
Store?
Execution 1
Result = 12
Store?
Log

```
SELECT COUNT(*)
FROM ACCT_TGT TGT
WHERE TGT.ACCT_TGT_NO NOT IN
(SELECT SRC.ACCT_SRC_NO
FROM ACCT_SRC)
```

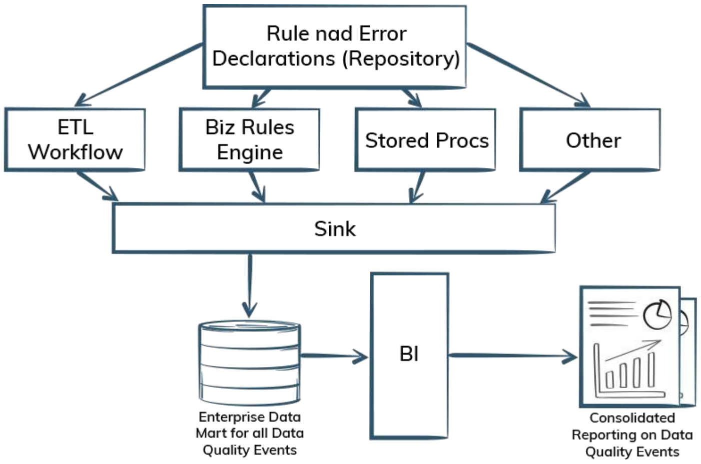
Figure 4: *Storage of Results of Business Rule Execution*

Figure 4 illustrates the problem. We have a rule that somehow gets turned into an executable form, such as SQL. The rule executes and produces a result - a useful measurement of data health. What do we do with the result? The data stores upon which the rule executed are not ideal places to store the result. If we continuously create rules over time we would have to keep changing the structure of these databases to store the results - something that would be unacceptable in a production environment. Further, these databases exist for business purposes, not for rules management. Suppose the target database has every record deleted and reloaded once a day. This would result in historical metrics being lost.

The only answer is for the BRE to store the results of the rule execution in its own environment. In fact we would expect the BRE to store additional metadata about the execution of the rule, such as the date and time of execution, the physical name of the target data stores, and so on. This means that we expect a BRE to store (a) definitional information about the rules; and (b) the results of execution of the rules. Not all BRE's are capable of doing this, and thus not all may be suitable for use in monitoring and metering.

The Enterprise View of Data Quality Metrics

The idea that all data quality events should be reported centrally in an enterprise has been gaining ground in recent years. Figure 5 provides a simple illustration of the concept.


Rule nad Error Declarations (Repository)
ETL Workflow
Biz Rules Engine
Stored Procs
Other
Sink
BI
Enterprise Data Mart for all Data Quality Events
Consolidated Reporting on Data Quality Events

The basic concept is that there may be many ways in which data quality events are detected in an enterprise - the BRE for monitoring and metering being just one of them. A component called a "sink" is accessed by all these different environments to pass information about data quality events they detect to some kind of consolidated store. Perhaps this could be an enterprise data mart which utilizes business intelligence tools to provide reports and analysis of the state of data quality in the enterprise.

Figure 5: *Architecture for Consolidated Reporting of Data Quality Events*

Many BREs are closed "black boxes" that are not capable of communicating with a sink, and have internal repository structures that are proprietary and not understandable. Such BREs will not easily support consolidated data quality reporting at the enterprise level.

Notification and Alerting

Gathering and storing metrics as a result of metering is one thing. However, monitoring implies the detection of events. When an event is detected, individuals or applications must be notified. If the event is determined to be adverse, they must be alerted. An alert is intended to trigger immediate action. A notification is more for informational purposes, although it might be part of a series that shows a developing trend that may lead to problems.

There are several issues with notification and alerting. One is that the BRE must be aware of when the results of the execution of a rule pass some kind of threshold that raises a notification or an alert. This is often incorporated in the metadata of the rule definition. If the rule evaluates to some predetermined value such as true or false, or a metric exceeds a pre-specified threshold, an alert may be raised.

Therefore, the rule definition must be capable of holding metadata that the analyst can specify concerning predetermined, or expected results that will trigger a notification or alert. Again, this is a feature that will not be found in all BRE's. Some merely gather and store metrics without taking any further action.

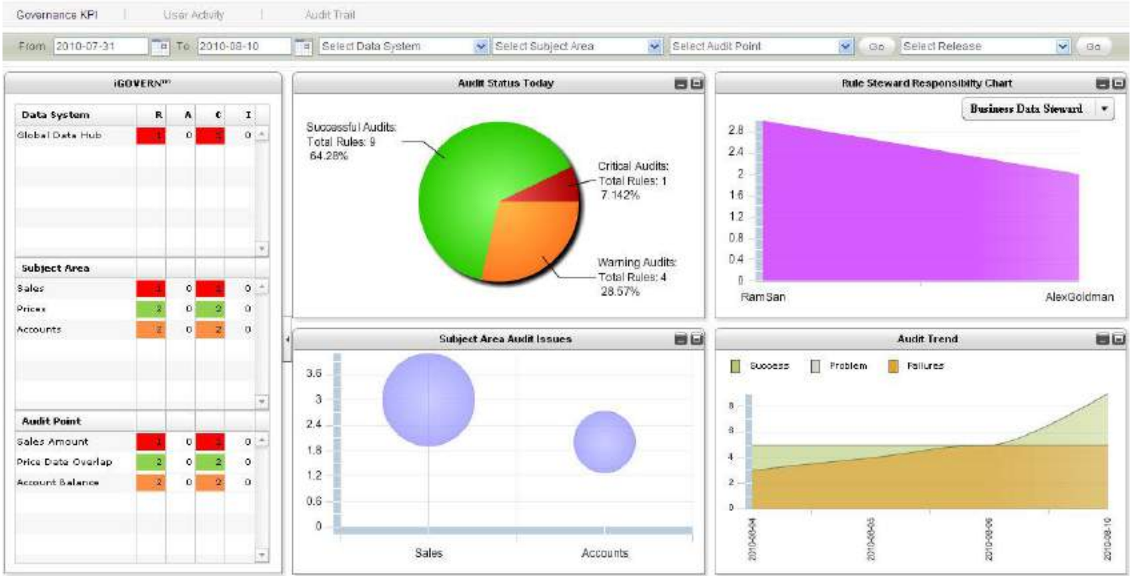
Another major element that must be in place for successful notification and alerting is stakeholder management. And with this consideration we return again to basic data governance. Data governance emphasizes responsibilities for data. Each rule that is capable of raising a notification or alert must know who to send it to. That can only be done if the BRE either stores data about stakeholders or has access to an environment that does. Even then, the stakeholder(s) must be connected to the rule, and there may be many roles that have to be captured at the rule level. The individual who receives notifications may be one, the individual who receives alerts may be another, and there may even be an individual for escalation if the detected data quality event occurs repeatedly. Stakeholder management of this kind is not found in all BREs, but it is essential.

Presentation of Results

Earlier in this paper the concept of a control room was discussed, and it was noted that there is no real equivalent of this for the production data environment. Notification and alerting only provide a mechanism to act on individual problem, but management will always want to see a "big picture" of the health of data across the production landscape. The architectural approach of monitoring and metering described here does generate the basic measurements of data quality and stores them, but how should they be presented?

A "big picture" approach suggests a high-level representation with drill-down capabilities. Producing detailed reports on thousands (or tens of thousands) of rules executing in an environment every day is unsuitable for senior management, and may even be overload for individual analysts. The "big picture" can be generated using aggregation and/or filtering of the metrics metadata. The way in which is it presented is probably best as a dashboard, such as the one illustrated in Figure 6.

In this example, a variety of metrics are shown in different graphical formats. Rule execution over a preceding time period is summarized, as are problems detected in different subject areas of the production data landscape. This raises yet another point for the BRE providing the monitoring and metering environment, or the tool that is used to present the metrics gathered.



From 2010-07-31 To 2010-08-10 | Select Data System | Select Subject Area | Select Audit Point | Select Release

Data System	R	A	C	I
Global Data Hub	1	0	1	0

Subject Area				
Sales	1	0	1	0
Prices	2	0	2	0
Accounts	2	0	2	0

Audit Point				
Sales Amount	1	0	1	0
Price Date Overlap	2	0	2	0
Account Balance	2	0	2	0

Figure 6: *Example of Dashboard Reporting for Health of Production Data Landscape*

The point is that there will be a lot of additional "master metadata" that must be gathered and either integrated or mashed-up with the metrics that are collected. For instance, subject areas will be needed for one of the displays in Figure 6. Most enterprises have poorly defined subject area models, or don't have one. The physical structure of the landscape will probably be important too - the grouping of the major data stores. Unless these dimensions are elaborated with a common understanding across the enterprise, and connected to the rules in the BRE, meaningful high-level displays of the kind shown in Figure 6 will be very difficult to produce. This is not so much a feature of the tool as it is of the maturity of data governance in the enterprise.

Figure 6 also illustrates another principle of data governance, which is the ways in which the individual is responsible for the health of particular items of data. Until recently it has been very rare to hold individuals accountable for data in this way. Millions of dollars might be spent in data quality cleanup projects, but very little or nothing on ensuring stewards producing data are doing so in ways that improve quality. However, enterprises now realize that connecting metrics gathered from production data environments to accountable and responsible stewards is essential to improve data quality. Any monitoring and metering tool will be expected to support this aspect of data governance. For instance, RACI matrices can quickly show how individuals are performing against their assigned tasks in data quality assurance.

Conclusion

We began this paper by reviewing the aspirations of, and urgent necessity for, data governance in enterprises today. The notion that data governance is a collection of different domains each with specific problems and methods led to consideration of data quality as one major area. Within that, the traditional approach to data quality was contrasted with the general lack of monitoring and metering of data health that makes it impossible to effectively govern production data landscapes. Of the approaches to monitoring and metering, business rules engines offer the most promise. What we would expect from such business rules engines has been described. However, space has only permitted us to touch upon some of the major points, and a lot of detail had to be omitted. For instance, the methodology that must be built around the use of any business engine is a topic so vast that we will probably write an entire books devoted to it in the future. The same is true of the master metadata that describes production data environments. Nevertheless, the basic elements have been covered, and while some aspects are undoubtedly forward-looking, the domain of monitoring and metering enterprise data within overall data governance is developing quickly and we can expect to see it more clearly addressed by the industry in the next few years.

About iCEDQ Soft

In 2005, The Company was founded by a team of data architects to solve various challenges related to Data Centric Projects. At that time they identified a gap in the market for comprehensive DW/ETL testing software. They then embarked on building a platform that would provide customers an easy way to set up an automated solution for end to end testing of their data centric projects.

Three years later the product was commercially available for the market. Since then we have served clients in Insurance, Finance, Healthcare and Retail space. We strongly believe that in today's world, Data is the key of any business decision making. Therefore we are continuously working towards building a better solution to make sure that the quality of your data is good.

iCEDQ Soft

1200 Summer Street
#204, Stamford,
Connecticut 06905

Web: http://icedq.com
Phone: (203) - 517 - 9690
Email: contact@icedq.com