A DataFlux White Paper

Prepared by: David Loshin

Observing Data Quality Service Level Agreements:

Inspection, Monitoring, and Tracking

DataFlux
A SAS COMPANY

Leader in Data Quality and Data Integration

www.dataflux.com
877-846-FLUX

International
+44 (0) 1753 272 020

Before starting any data quality or data governance initiative, an important first step is to establish the expectations for this program. Specifying the expectations of the data's consumers provides a means for measuring and monitoring the conformance of data (and associated processes) within an operational data governance framework. These agreements can be formalized under a *data quality service level agreement* (DQ SLA), which specifies the roles and responsibilities associated with the management and assurance of data quality expectations.

SLAs are familiar to anyone with an IT background, but they are typically focused on issues of system availability, service turnaround and other issues. This paper discusses how implementing a DQ SLA via formalized processes can transform data quality management from a constant "fire-fighting" mode to a more consistent, proactive approach.

The objective of the data quality service level agreement is establishing data quality control. This relies on monitoring conformance to data quality rules define using agreed-to dimensions of data quality, such as accuracy, completeness, consistency, reasonableness, and identifiability, among others. We will consider these dimensions of data quality, and the ways that data quality rules are defined. Despite the best efforts to ensure high data quality, there are always going to be issues requiring attention and remediation. As a result, identifying data errors as early as possible in the processing stream(s) supports the objective of the DQ SLA: notifying the right individuals to address emergent issues and resolving their root causes in a reasonable amount of time. We will look at the process of defining data quality rules, their different levels of granularity, approaches for introducing measurement processes, and choosing appropriate acceptability thresholds.

This paper will then consider the relevance of measurement and monitoring: defining inspection routines, inserting them into the end-to-end application processing, and reporting measurements. When the quality of data does not meet the level of acceptability, data quality issue events are generated, the issues are logged in a data quality incident tracking system, and the individuals specified in the data quality service level agreement are charged with diagnosis and remediation. Through this operational data governance, an organization can internalize the observance of the DQ SLAs, and consequently continuously monitor and control the quality of organizational data.

Data Quality Service Level Agreements (DQ SLAs)

Just as a more traditional SLA governs the roles and responsibilities of a hardware or software vendor, a DQ SLA is an agreement that specifies data consumer expectations in terms of rules and levels of acceptability, as well as reasonable expectations for response and remediation when data errors and flaws are identified. DQ SLAs can be expressed for any situation in which a data provider hands off data to a data consumer.

More precisely, within any business process, a data recipient would specify expectations regarding measurable aspects relating to one or more dimensions of data quality (such as accuracy, completeness, consistency, timeliness, etc.). The DQ SLA, then, would incorporate an expected data quality level as well as enumerate the processes to be executed when those expectations are not met, including:

1. Location in the business process flow that is covered by the SLA
2. Critical data elements covered by the SLA
3. Data quality dimensions associated with each data element
4. Expectations of quality for each data element for each of the identified dimensions
5. Defined data quality rules formalizing those expectations
6. Business impacts associated with nonconformance to the defined data quality rules
7. Methods for measuring nonconformance to those expectations
8. Acceptability thresholds for each measurement
9. How and where the issues should be categorized, prioritized and documented
10. Individual(s) to be notified in case the acceptability thresholds are not met
11. Times for expected resolution or remediation of the issues
12. Method for tracking status of the resolution process
13. Escalation strategy and hierarchy when the resolution times are not met

The beauty of the DQ SLA is the acknowledgement that data quality problems – and their remediation – are almost always tied to a business process. In order to benefit from the processes suggested by the definition of a DQ SLA (especially items 5, 7, 9, and 12), systems supporting those activities must be put into place, namely:

- Data quality rules management
- Measurement, monitoring and notification
- Data quality incident categorization, prioritization and tracking

All of these concepts are instrumental in establishing the objective of the DQ SLA: data quality control, which relies on the definition of rules using agreed-to dimensions of data quality. These control processes measure conformance to those rules as data is handed off from one task to another. If it is determined that the information does not meet the defined expectations, the remediation process can include a variety of tasks – write the non-confirming text to an outlier file, email a system administrator or data steward to fix the problem, run an immediate corrective data quality action, or perform any combination of these.

Dimensions of Data Quality

Data quality rules are defined within the context of dimensions of data quality, providing a spectrum against which conformance may be compared. Once the data consumers' expectations are identified and reviewed, the next step is to formalize the definition in terms of dimensions such as accuracy, completeness, currency, reasonability, consistency, and identifiability. This enables the creation of processes, filters, and monitors to observe conformance.

Accuracy

Data accuracy refers to the degree with which data values correctly reflect attributes of the "real-life" entities they are intended to model. Accuracy can be measured in reference to different sources of what is believed to be "correct information," such as a database of record, a similar corroborative set of data values from another table, dynamically computed values, or perhaps the result of a manual process.

Completeness

An expectation of completeness indicates that certain attributes should be assigned values in a data set. Completeness rules can be assigned to a data set in three levels of constraints:

- Mandatory attributes that require a value
- Optional attributes, which may have a value based on some set of conditions
- Inapplicable attributes, such as maiden name for a single male, which may not have a value

Completeness can be prescribed on a single attribute, or can be dependent on the values of other attributes within a record or message. Completeness may be relevant to a single attribute across all data instances or within a single data instance. Some example aspects that can be measured include the frequency of missing values within an attribute and the conformance to optional null value rules.

Currency

Currency refers to the degree to which information is up-to-date with the corresponding real-world entities. Data currency may be measured as a function of the expected frequency rate at which different data elements are expected to be refreshed, as well as verifying that newly created or updated data is propagated to dependent applications within a specified (reasonable) amount of time. Another interesting aspect includes temporal consistency rules that measure whether dependent variable reflect reasonable consistency (e.g., the "start date" is earlier in time than the "end date").

Reasonability

General statements associated with expectations of consistency or reasonability of values, either in the context of an existing record or over a time series, are included in this

dimension. Some examples include consistency of values across different data sets, as well as multi-value consistency rules asserting that the value of a set of attributes is consistent with the values of another set of attributes, perhaps even across data sets.

Structural Consistency

Structural consistency refers to the consistency in the representation of similar attribute values, both within the same data set and across the data models associated with related tables. Structural consistency also can measure the degree of consistency between stored value representations and the data types and sizes used for information exchange. Conformance with syntactic definitions for data elements reliant on defined patterns (such as addresses or telephone numbers) can also be measured.

Identifiability

Identifiability refers to the unique naming and representation of core conceptual objects as well as the ability to link data instances containing entity data together based on identifying attribute values. One measurement aspect is entity uniqueness, ensuring that a new record should not be created if there is an existing record for that entity. Asserting uniqueness of the entities within a data set implies that no entity exists more than once within the data set and that there is a key that can be used to uniquely access each entity (and only that specific entity) within the data set.

Data Quality Control

Errors characterized as violations of expectations for completeness, accuracy, timeliness, consistency, and other dimensions of data quality often impede the completion of information processing streams. Despite the best efforts to ensure high data quality, there are always going to be issues requiring attention and remediation. The critical question focuses on identifying data errors as early as possible in the processing stream(s) as a way of meeting the specifications of the DQ SLA: notifying the right individuals to address the issue, and determining if the issue can be resolved appropriately within a "reasonable" amount of time.

Qualifying the Control Process

The data quality control process helps to ensure that any data issue that might incur a significant business impact downstream in the business process streams is identified early in the processing stream. Therefore, the DQ SLA is intended to describe how a collection of control processes together ensure that:

- Control events occur when data failure events take place
- The proper mitigation or remediation actions are performed
- The corrective actions to correct the problem and eliminate its root cause are performed within a reasonable time frame
- A control event for the same issue is never triggered further downstream

Data quality control differs from data validation in that validation is a process to review and measure conformance of data with a set of defined business rules. Together with the DQ SLA, data quality control is an ongoing process to:

- Reduce the number of errors to a reasonable and manageable level
- Enable the identification of data flaws along with a protocol for interactively making adjustments to enable the completion of the processing stream
- Instituting a mitigation or remediation of the root cause within an agreed-to time frame

Types of Data Quality Controls

The value of a data quality control mechanism lies in establishing trust on behalf of downstream users that any substantive issue that can be identified, addressed and corrected early enough to prevent a material impact. Data controls can be applied at different levels of granularity:

- **Data element controls** review the quality of the value in the context of its value assignment to a data element.
- **Data record controls** examine the quality of the set of (element, value) pairs within the context of the record.
- **Data set and data collection controls** focus on completeness of the data set, availability of data, and timeliness in its delivery.

These controls are instituted at the location in the processing stream covered by the DQ SLA, at the point where the data is handed off from the producing stage to the consuming stage. Monitoring processes apply data quality rules to measure conformance to expectations and determine the level of acceptability.

Defining Data Quality Rules

The different levels of control granularity suggest a way of refining data consumer expectations into formal rules that can be applied automatically using a rules engine. This approach is a successive refinement of those expectations into more precise (and hopefully, measurable) assertions. The monitoring processes can then use the rules to examine the candidate data at the different levels of granularity. Once the rules are formally defined, the data quality team determines a method for measurement, a unit of measure, and an acceptability threshold, which is the lowest measurement that denotes conformance to business expectations.

For example, consider a consumer expectation for customer address data regarding the existence of a "State" data element value in each record. This expectation is measured using the completeness dimension, and since it involves a specific data element within each record, its granularity is at the data element level. If the number of violating records the consumer is willing to tolerate is as much as 1% of the total number of records, this suggests the unit of measure (percentage of conforming records within the data set) and the acceptability threshold (99%).

Successive Rule Refinement

As a general approach for defining rules, for each specified data quality expectation, the data quality analyst can apply these steps:

1. Determine the dimension of data quality (for example, focus on those described in this paper: accuracy, completeness, currency, reasonability, consistency, and identifiability)
2. Determine the level of granularity (data element, data record, data set)
3. Identify the constraint and dependent data elements to be reviewed (for example, if this is an accuracy rule for product names, select the product identifier, product name, and product description as the dependent variables, and provide a system of record against which the product data is to be compared for accuracy
4. Document the constraint and unit of measure, and select a measurement method
5. Determine the acceptability threshold

Despite the potential abstract nature of the dimensions of data quality, rules defined within those categories are eminently transformable into straightforward assertions that can be automated. For example, a data element completeness rule can be stated as an assertion that the data element is never null or missing; a consistency rule might assert that a membership start date is always earlier than a membership end date, and so on.

In fact, many rules can be boiled down to a standard format inspecting specific values. Rules that apply to single data elements, such as structural consistency or domain membership will allow one to inspect the data element's value in comparison to the specified constraint. As an example, specifying that a state code field must belong to the set of United States Postal Service two-character state codes, the assertion may describe the data domain as a reference table and assert that each state code must belong to that table. Syntactic pattern compliance (e.g., telephone numbers) can also be expressed in terms of conformance to a set of formats.

Rules that apply at the record level examine a condition involving more than one data element within the record. For example, ensuring that the start date is earlier than the end date is an assertion relating two data element values within each record. Again, assertion expressions (such as those expressed using SQL) can be used to formalize these data quality rules.

Data set rules compare sets of data elements in one data set with sets of data elements in another data set, or comparisons against aggregate values derived from more than one record in a data set. In some cases, key assertions and record uniqueness are rules that are relevant within the data set context. Other examples may involve verifying data element values against averages accumulated over a collection of transactions. Again, expression syntax such as that contained within SQL can be used for these assertions.

Measurement and Monitoring

Validating data against a defined expectation is the basis for data quality monitoring and tracking. Supporting the terms of the DQ SLA means that when data quality issues are identified, the appropriate people are notified and the agreed-to remediation tasks are initiated. This requires mechanisms for data inspection and monitoring. Additionally, process workflows must be defined for the purposes of inspecting data and ensuring that the data elements, records, and data sets meet downstream requirements.

This task involves defining the data quality inspection routines, which may include both automated and manual processes. Automated processes may include the results of edit checks executed during application processing, data profiling or data analysis automation, ETL tools, or customized processing. Manual inspection may require running queries or reports on data sources or even obtaining samples of data which are then examined. Inspection procedures are defined for each relevant data quality dimension. Each system may require different techniques and data quality inspection procedures.

Those procedures support two different approaches to monitoring. Static monitoring can be performed as batch inspections when data sets are handed off from one process activity to another, while inlined monitoring can be integrated directly as part of the end-to-end application workflow. The characteristics of each data quality expectation, its associated acceptability threshold, and the expected level of service will suggest whether it is better suited to static or inlined monitoring.

Static Monitoring

Static monitoring is well-suited to environments supported by batch processing, where data sets are "bulk exchanged" from one processing task to another, or where data is extracted and loaded into a target environment in preparation for analysis and reporting (such as data warehouses, data marts, and to some extent, operational data stores). In these cases, element, record, and data set rules can be applied to a collection of data instances as a whole. Conformance to data quality expectations can be measured in terms of both counts of violation and percentage of non-conformant records within the data set. Acceptability thresholds can be set based on either of those units of measure, and therefore data issues that are related to not meeting those thresholds are best suited to static monitoring.

Inlined Monitoring

Inlined monitoring, implemented through the augmentation of applications with verification routines to inspect data elements and records as they are created and/or modified, are best suited to situations in which a single violation requires immediate attention. That suggests that, for the most part, inlined inspection and monitoring works well for data rules to be applied to data element values and to records, but may introduce performance bottlenecks when applied to large data sets.

Measurement Processes

Automated process can be put into place for both types of monitoring, with the intention of enhancing any existing measurement processes to generate a notification when a violation occurs to the appropriate data steward as specified within the DQ SLA. Data profiling tools and data auditing tools are popular alternatives used for statically assessing conformance to defined rules as well as creating filters that can be embedded within a process flow. Inlined data edits are commonly used for data validation, and these edits can be enhanced to database queries and reports can be configured to measure and present statistics regarding conformance to rules. By embedding queries within polling applications that continually validate data on a periodic basis, any distinct violations requiring attention can be flagged quickly and the proper alerts can be generated.

Data Quality Incident Reporting and Tracking

Supporting the enforcement of the DQ SLA requires a set of management processes for the reporting and tracking of data quality issues and corresponding activities. This can be facilitated via a system used to log and track data quality issues. By more formally requiring evaluation and initial diagnosis of emergent data events, encouraging data quality issue tracking system helps staff members achieve more effective problem identification and, consequently, problem resolution.

Aside from improving the data quality management process, issue and incident tracking can also provide performance reporting including mean-time-to-resolve issues, frequency of occurrence of issues, types of issues, sources of issues, and common approaches for correcting or eliminating problems. An issue tracking system will eventually become a reference source of current and historic issues, their statuses, and any factors that may need the actions of others not directly involved in the resolution of the issue.

Conveniently, many organizations already have some framework in place for incident reporting, tracking and management. So, the transition to tracking data quality issues focuses less on tool acquisition, and more on integrating the concepts around the "families" of data issues within the incident hierarchies. It is also necessary to train staff to recognize when data issues appear and how they are to be classified, logged, and tracked. The steps in this transition will involve addressing some or all of these directives:

1. **Standardize data quality issues and activities** – Understanding that there may be many processes, applications, underlying systems, etc., that "touch" the data, the terms used to describe data issues may vary across lines of business. In order to gain a consistent and integrated view of organizational data quality, it is valuable to standardize the concepts used. Doing so will simplify reporting, making it easier to measure the volume of issues and activities, identify patterns and interdependencies between systems and participants, and ultimately to report on the overall impact of data quality activities.

2. **Provide an assignment process for data issues** – Resolving data quality issues requires a well-defined process that ensures that issues are assigned to the individual or group best suited to efficiently diagnose and resolve the issue, as well as ensure proper knowledge transfer to new or inexperienced staff.
3. **Manage issue escalation procedures** – Data quality issue handling requires a well-defined system of escalation based on the impact, duration, or urgency of an issue, and this sequence of escalation will be specified within the DQ SLA. Assignment of an issue to a staff member starts the clock ticking, with the expectation that the problem will be resolved as directed by the DQ SLA. The issues tracking system will enforce escalation procedures to ensure that issues are handled efficiently, as well as prevent issues from exceeding response performance measures.
4. **Document accountability for data quality issues** – Accountability is critical to the governance protocols overseeing data quality control, and as issues are assigned to some number of individuals, groups, departments, or organizations, the tracking process should specify and document the ultimate issue accountability to prevent issues from dropping through the cracks.
5. **Manage data quality resolution workflow** – The DQ SLA essentially specifies objectives for monitoring, control, and resolution, all of which defines a collection of operational workflows. Many issue tracking systems not only provide persistent logging of incidents and their description, they also support workflow management to track how issues are researched and resolved. Making repeatable and efficient workflow processes part of the issues tracking system helps standardize data quality activities throughout the organization.
6. **Capture data quality performance metrics** – Because the DQ SLA specifies performance criteria, it is reasonable to expect that the issue tracking system will collect performance data relating to issue resolution, work assignments, volume of issues, frequency of occurrence, as well as the time to respond, diagnose, plan a solution, and resolve issues. These metrics can provide valuable insights into the effectiveness of current work flow and systems and resource utilization, and are important management data points that can drive continuous operational improvement for data quality control.

Implementing a data quality tracking system provides a number of benefits. First, knowledge sharing can improve performance and reduce duplication of effort. Furthermore, an analysis of the issues will permit staff to determine if repetitive patterns are occurring, their frequency and impact, and potentially the source of the issue.

Tracking issues from data transmission to follow-up reporting will ensure that a full lifecycle view of the data and issues are identified and recorded. And lastly, since we know that issues identified and being resolved upstream of the data lifecycle may have critical consequences downstream, employing a tracking system essentially trains people to recognize data issues early in the information flows as a as a general practice that supports their day-to-day operations.

Issue Remediation – Taking Action

One of the most important steps in the planning process is developing an action plan for responding to emergent data issues. Depending on the specific measure being observed there may be slightly different processes, but for most issues, certain steps should always be followed:

1. **Confirm** the issue by directly reviewing the data, such as observing specific examples or running queries, profiles, or reports
2. **Notify** the key stakeholders that a confirmed issue exists and is being researched
3. **Log** the issue in the data quality issues tracking system
4. **Diagnose** the issue by researching the source of the issue to determine its root cause
5. **Evaluate options** for addressing the issue, which may include eliminating the root cause by modifying the processes or introducing new techniques into the information flow, introducing additional monitoring processes, and potentially directly correcting the offending data items
6. **Correct** offending data; correction may be a temporary mitigation tactic for a data quality issue, and in this case, a more permanent remediation strategy is required to prevent reoccurrence of the problem
7. **Improve** delinquent processes by identifying the root cause and adjusting the process to eliminate any source of error introduction

Supporting the Inspection and Monitoring Processes

Instituting data quality inspection and monitoring as a way of asserting control over satisfying data consumer requirements is a good data management practice. The definition and management of business rules for describing data expectations and measuring conformance requires more than just good processes, though. This practice is supported with technology to help define data rules and automate their execution. This requires:

- A repository for defining and managing data quality rules
- Profiling to monitor and measure conformance to data quality rules
- Incident reporting and tracking for tracking issues
- Notification and incident remediation management

Using these tools with the techniques described in this paper can help to internalize operational data governance, supporting the observance of the DQ SLAs, and enabling continuous monitoring and control of organizational data quality.