Q C Technology Decisions Inc.

Accelerating Education Data Warehousing with Informatica Data Quality

A framework to define, control and monitor quality of data

White Paper

Published August 2014

Executive Summary

Modern Education Data Warehouses require that data be conformed across multiple sources and in some cases even across multiple state agencies such as Early Childhood Development, Public Education (K12) Agencies, Private Education Agencies, Higher Education Agencies and Workforce Development. Emerging trends in "Big Data" indicate that agencies will want to incorporate other data such as public data sets to perform statistical analysis to improve the Educational System.

Education Data Warehouses need to maintain accurate sets data and cohorts over time that are sourced from a variety of systems. This poses a unique challenge with data integration and quality control to create a sustainable and efficient process capable of moving data faster into the hands of the decision makers that need it.

Enterprise data integration capabilities are the foundation of the SLDS solution. Using reliable, enterprise class technology from Informatica, the solution accesses, integrates, and delivers data of any volume, for any application from virtually any source in any format at any latency. Armed with these capabilities, IT organizations can break down the silos wherever data is held enabling seamless data sharing across multiple state agencies and commissions.

Although education agencies (both local and at the state level) have developed have developed processes to match student records, generate IDs, and conform data to meet their unique data warehousing requirements; they continue to struggle with data collection processes. This bottleneck is a barrier to "Velocity", which when removed and by definition of the three V's (Volume, Velocity, and Variety), would bring Education Data Warehousing into the "Big Data" arena.

Addressing Data Quality Challenges

Data quality is not a one-time effort; However, in traditional data warehousing projects a significant (if not most of their time) is spent cleansing and organizing the data in a way that makes sense to business users. The events and changes that allows data anomalies to be introduced into an environment are not unique; however, addressing anomalies becomes critically important when users are relying on accurate student records and demographics for making decisions. It is necessary for the data management teams to not just address acute data failures, but also baseline the current state of data quality so that one can identify the critical failure points and determine improvement targets.

The ability to monitor data quality and react quickly to changes demonstrates a level of organizational maturity that views information as an asset and rewards proactive involvement by delivering on the promises of business intelligence; Data Trust, Business Value, and Process Alignment.

Education Data Quality Accelerator

The Education Data **Quality Accelerator** for Informatica helps organizations rise to the operational challenges integrating information from student and other information systems. Two important features include:

- Quality Rule and Reference Tables gives agencies a head-start in validating the most common data quality issues found with student and school data sets including demographics, programs, awards, and financial aid.
- Scorecards that provide analysts with detailed visualizations and notifications of a variety of quality metrics for K-12, Post-Secondary and Workforce agencies.

Technology Supports Your Metrics

Informatica Data Quality provides the framework to effectively monitor data quality. Performance must integrate technology to coordinate the assessment and discovery of data quality issues, the definition of data quality rules, the use of those rules for distinguishing between valid and invalid data and possibly cleansing invalid data, and the management, measurement, and reporting of conformance to those rules.

Assessment

Part of the process of refining data quality rules for proactive monitoring deals with establishing the relationship between recognized data flaws and business impacts, but in order to do this, one must first be able to distinguish between "good" and "bad" data. The attempt to qualify data quality is a process of analysis and discovery involving an objective review of the data values populating data sets through quantitative measures and analyst review. While a data analyst may not necessarily be able to pinpoint all instances of flawed data, the ability to document situations where data values look like they don't belong provides a means to communicate these instances with subject matter experts whose business knowledge can confirm the existences of data problems.

Data profiling is a set of algorithms for statistical analysis and assessment of the quality of data values within a data set, as well as exploring relationships that exists between value collections within and across data sets. For each column in a table, a data profiling tool will provide a frequency distribution of the different values, providing insight into the type and use of each column. Cross-column analysis can expose embedded value dependencies, while inter-table analysis explores overlapping values sets that may represent foreign key relationships between entities, and it is in this way that profiling can be used for anomaly analysis and assessment, which feeds the process of defining data quality metrics.

Definition

The analysis performed by data profiling tools exposes anomalies that exist within the data sets, and at the same time identifies dependencies that represent business rules embedded within the data. The result is a collection of data rules, each of which can be categorized within the framework of the data quality dimensions. Even more appealing is the fact that the best-of-breed vendors provide data profiling, data transformation, and data cleaning tools with a capability to create data quality rules that can be implemented directly within the software.

Validation and Cleansing

Our data quality rules are going to fall into two categories. One set of rules, validations, simply asserts what must be true about the data, and is used as a means of validating that data conforms to our expectations. Both data transformation and data profiling products will allow the end client to define validation rules that can be tested against a large set of data instances. For example, having determined through profiling that the values within a specific column should fall within a range of 20-100, one can specify a rule asserting that "all values must be greater than or equal to 20, and less than or equal to 100." The next time data is streamed through the data quality tool, the rule can be applied to verify that each of the values falls within the specified range, and tracks the number of times the value does not fall within that range.

The second set of rules, cleansing or correction rules, identifies a violation of some expectation and a way to modify the data to then meet the business needs. For example, while there are many ways that people provide telephone numbers, an application may require that each telephone number be separated into its area code, exchange, and line components. This is a cleansing rule which can be implemented and tracked using data cleansing tools.

Monitor and Manage Ongoing Quality of Data

The most important component of data quality metrics is the ability to collect the statistics associated with data quality metrics, report them in a fashion that enables action to be taken, and provides historical tracking of improvement over time. Forward-thinking vendors consider ways that the results of monitoring data quality metrics can be captured and presented to the user to allow analysis and drill down in a way that relates how data flaws roll up into business impacts.

Putting it all Together: The Data Quality Scorecard

A data quality scorecard is a management tool that captures a virtual snapshot of the quality levels of your data, presents that information to the user, and provides insight as to where data flaws are impacting business operations and where the most egregious flaws exist within the system. Using data quality rules based on defined dimensions provides a framework for measuring conformance to business data quality expectations.

Validating Data

To be able to measure the level of data quality based on the dimensions of data quality, the data to be monitored will be subjected to validation using the defined rules. The levels of conformance to those rules are calculated and the results can be incorporated into a data quality scorecard. Measuring conformance is dependent on the kinds of rules that are being validated. For rules that are applied at the record level, conformance can be measured as the percentage of records that are valid. Rules at the table or data set level (such as those associated with uniqueness), and those that apply across more than one data set (such as those associated with reference integrity) can measure the number of occurrences of invalidity.

Validation of data quality rules is typically done using data profiling, parsing, standardization, and cleansing tools. As is mentioned in section 'Definition' (p 11), best-of-breed vendors allow for integrating data quality rules within their products for auditing and monitoring of data validity. By tracking the number of discovered (and perhaps, corrected) flaws as a percentage of the size of the entire data set, these tools can provide a percentage level of conformance to defined rules. The next step is to assess whether the level of conformance meets the user expectations.

Thresholds for Conformance

For any measured metric, user expectation levels are set based on the degree to which the data conforms to the defined sets of rules. But since different data flaws have different business impacts, the degrees to which different data quality rules are violated reflect different levels of business criticality. Consequently, there may be different levels of expected conformance, which is reflected in setting acceptability thresholds.

The simplest approach is to have a single threshold. If the conformance rate meets, or exceeds the threshold, the quality of the data is within acceptable bounds. If the conformance rate is below the threshold, the quality of the data is not acceptable.

A more comprehensive approach provides three ranges based on two thresholds: "acceptable," when the conformance rate meets or exceeds a high threshold, "questionable, but usable," when the conformance rate falls between the high and low thresholds, and "unusable" when the conformance rate falls below the low threshold. As can be seen here, a dashboard can present the actual measurements on a green/amber/red background to provide a quick visual cue as to the quality of the data.

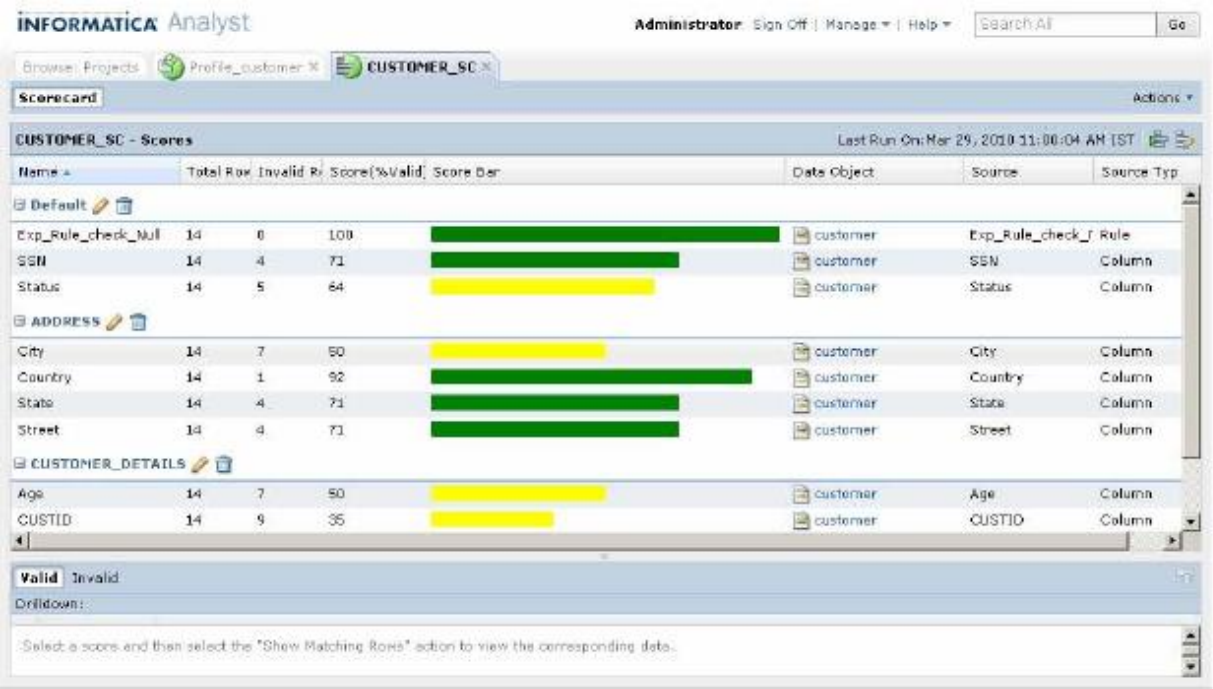

Figure 1 This dashboard view reflects green/amber/red thresholds for conformance

Ongoing Process Control

Applying a set of data validation rules to a data set once provides insight into the current state of the data, but will not necessarily reflect how system modifications and updates have improved overall data quality. However, tracking levels of data quality over time as part of an ongoing improvement process provides a historical view of when and how much the quality of data improved.

As part of a statistical control process, data quality levels can be tracked on a periodic (e.g., daily) basis, and charted to show if the measured level of data quality is within an acceptable range when compared to historical control bounds, or whether some event has caused the measured level to be below what is acceptable. Statistical control charts can help in notifying the data stewards when an exception event is impacting data quality, and where to look to track down the offending information process. Historical charting becomes another critical component of the data quality scorecard.

Quality Scorecards

Scorecards included with the **Education Data Quality Accelerator** provide analysts with detailed visualizations and notifications of a variety of metrics for:

- K-12
- Post-Secondary
- Workforce

Proactive Monitoring and Alerting

When users are unable to check scorecards manually on a daily or hourly basis, they can opt to receive proactive notifications by dashboards, e-mails or instant message. The Informatica® Proactive Monitoring for Data Quality Option augments data quality programs with proactive monitoring capabilities. It automatically delivers alerts when data quality issues are found. It helps to mitigate the negative impact of poor or questionable data on your applications and processes.

Among the variety of notification rules are identifiers of data that is different than the norm by a certain percentage or deviation. Alerts are triggered by the solution's set of prepackaged data quality monitoring rules. These rules perform completeness checks, conformity checks, and trending analysis. Using a self-service model, business users can easily modify data quality monitoring thresholds, rules and alerts, as well as create new ones, themselves without having to rely on scarce IT resources.

When users receive an alert, they get a link to the Informatica Data Quality Analyst interface that provides instant awareness of what needs to be corrected.

Data Standardization

Using Informatica Data Quality as a tool, functional teams can work with stakeholders to develop and refine a controlled vocabulary of terms used to support reporting needs. It's natural for each source system to have their own set of terms used to describe data elements they provide to the SLDS; however, a number of attributes such as "Demographic attributes", must be conformed to establish proper links between the source data to support longitudinal analysis. Terms identified during the data profiling of these source systems are generally shared with the agencies for validation and presented as standard set of terms approved for the SLDS.

Education Dimensions

The **Education Data Quality Accelerator** provides an extensible framework for standardizing data across the following dimensions:

Academic Year
Cohort Year
Assessment Category
Assessment Purpose
Assessment Score Type
Assessment / Test
Assessment / Test Date
Entry Date
Exit Date
Attainment Level
Award
Building
Career Cluster
Citizenship
Cohort
Course
Course Mark
Degree Objective
Delivery Method
Disabled / Disability Status
Employer
Employer Status
Employment Type
English Proficiency
Enrollment Status
Entry Reason
Ethnicity / Race
Exit Reason
Free and Reduced Lunch / Socioeconomic
Gender
Grade Level
Graduation Plan
Homeless
Home Location
Household
Institution
Job Category
Job Classification
LEP Level
Loan Type
Location
Participation Status
Primary Language
Proficiency Level
Program Activity
Special Service
Special Setting
Student
Study Field
Subject
Teacher
Test
Test Accommodation
Work Location

About Informatica Data Quality

The Informatica® Data Quality product family delivers authoritative and trustworthy data to all stakeholders, projects, and domains for all projects and applications—on premise or in the cloud — using a single, unified platform. It allows you to proactively monitor and cleanse your data, and empowers the business to share in the responsibility for data quality management and governance.

The Informatica Data Quality product family enables business and IT to work effectively together to achieve real value for your organization's data assets. These products provide you with the advantages of minute-by-minute decision-making, cross-business unit visibility, synchronization of mission-critical operations, and transparency for regulatory compliance. Visit http://www.informatica.com/us/products/data-quality

Summary

QC Technology Decisions Inc. has contributed to the evolution of Education Data Warehousing through the design and development of statewide data warehouse solutions such as the West Virginia P20 SLED (State-Wide Longitudinal Education Data) and local education data warehouse solutions for school districts in the United States.

Our team has practical experience integrating data from student information systems, state extracts, and testing solutions using leading integration platforms such as Informatica and SQL Server Integration Services. We can architect and implement solutions using leading business intelligence platforms like Oracle OBIEE, IBM Cognos, Tableau and Microsoft SharePoint.

Our data warehousing professional services include:

- Enterprise Architecture
- Stakeholder outreach and change management
- Requirements analysis
- Data integration strategy, planning and implementation
- Data quality improvement strategy, planning and execution
- Enterprise Data Warehouse design, development and maintenance
- Extract, transform and load (ETL) process design and development
- Custom web application development
- Systems Integration
- Web Services development
- Portal development
- Training and documentation

About QC Technology Decisions

QC Technology Decisions Inc. is best known as a K-12 educational technology firm specializing in business intelligence and enterprise information management. Since 2005, our team has been providing professional IT services in both the United States and Canada to Local and State Education Agencies. Our team has extensive experience and knowledge in working with the development of educational policy, evaluation of educational programs as well as extensive design, deployment and management of information technology solutions including the design and deployment of data warehouse systems.

Technology Decisions Inc.

QC Technology Decisions Inc.
26 Crestline Drive
Fredericton, NB E3G 6B1
(506) 470-8938

sales@qctechnology.com
www.qctechnology.com