Guarding the Quality of Your Data, Your Most Valued Asset

Introduction

Over the last decade or so, organizations have increasingly harnessed Business Intelligence to positively impact top line growth. Beginning with a keen focus on delivering the *single version of truth*, Business Intelligence has itself matured in terms of ease of implementation, rapid development and deployment strategies, proper decoupling of business rules and technology, and cutting edge analytics. Additionally, the rapid proliferation of the understanding and knowledge among practitioners as well as the decision makers has made Business Intelligence a success. Despite the low success rate of Business Intelligence projects, organizations have invested, implemented and reaped the benefits which have been evident. But does the successful delivery on projected ROI (return on investment) imply success?

The Data Warehousing Institute estimated the cost incurred by American businesses due to poor data quality to be nearly six hundred billion dollars annually. Organizations started off with their available data, designed ways to harness its inherent value and capitalize on that, and yet forgot all about it right at the same time. Poor data management, epitomized by a lack of tangible returns on investment, has caught organizations of all sizes by surprise. Driven only by performance metrics, organizations have unknowingly deliberated the neglect of their most important asset.

In the sections that follow we will identify tangible metrics for data quality management and also provide an architectural overview for sustaining it.

The Metrics Dilemma

At the very outset, for sustained sponsorship for data quality management, it is critical to define relevant metrics. Lacking this, it becomes difficult to highlight the incentives behind such an exercise. Metrics can be identified by gauging data quality issues and relating them to business impacts.

The following are some typical data issues and the data quality dimensions that stem from them.

Data Issue Questions	Dimensions
What data is not populated/unusable/not referenced?	Completeness
	Integrity
What data is stored in a format that makes it non-sharable?	Conformity
What data values reflect conflicting information?	Accuracy

Data Issue Questions	Dimensions
What data is repeated?	Redundancy
What data has been modified? By whom? When?	Lineage
What impact does data changes have?	
What data values have varying representations/datatypes?	Structural consistency

These dimensions can be tied to tangible metrics which can be reported on, and benchmarks can be created to do time series analysis.

Metrics	Definition
%Valid	Percentage of data that is ascertained valid
%Distinct valid	Percentage of distinct valid data among all distinct data
%Incomplete	Percentage of data that is ascertained incomplete (nulls or spaces)
%Distinct Incomplete	Percentage of distinct incomplete data among all distinct data
%Format Violation	Percentage of data that falls into invalid formats
%Distinct	Percentage of distinct values
Average Data Storage Length	Average of lengths of data values
Data Format Frequency	Frequency of each data format found within the data

Apart from the metric examples listed above, business specific data quality metrics should be evaluated for their measurability, and incorporated according to their impact on the sanctity of business processes. Each of these metrics should also be benchmarked for critical levels from time to time. Subsequent measurements can then be used to evaluate the effectiveness of the whole data quality management effort on a temporal basis.

'Who's the boss?' or is it 'Who takes the blame?'

Subject matter experts and the data stewards within an organization need to be identified early on. These stalwarts have to shoulder the responsibility for proper research and documentation of the business terms and process data associated with their respective domain and the policing required in maintaining their sanctity.

Many organizations have viewed data management as an IT-only effort and thus failed to attain desired levels of success. It is important to understand that data quality assurance is an enterprise-wide phenomenon, and only a close partnership with IT can deliver the desired results. The data steward and SME roles are critical for the success of any data quality management project.

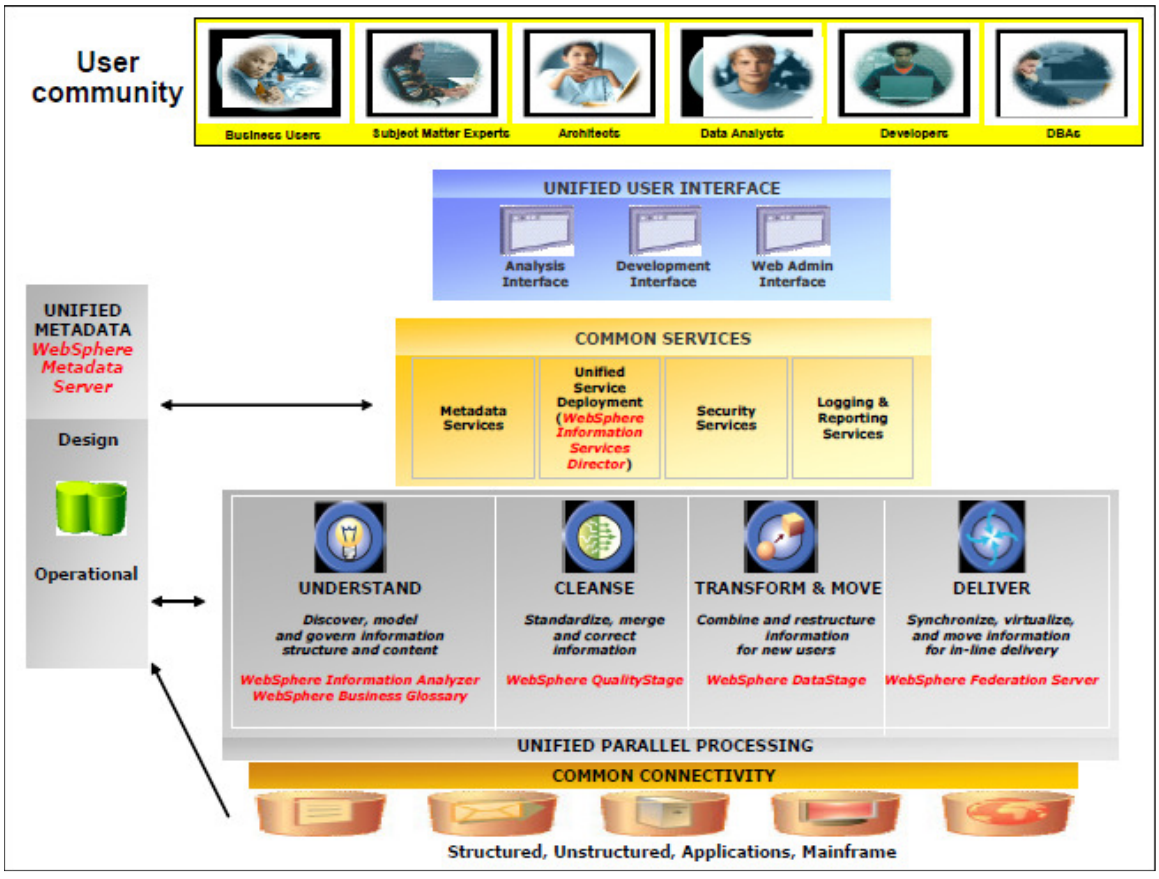
What do we fix now? The Visibility Problem

Once the controls and checks to gauge data quality are in place, an infrastructure to implement changes and fixes needs to be designed as well. No matter how difficult and intricate putting the controls and checks in place might seem, correctly judging requisite fixes is important. Programs and artifacts have to be put in place that feed off the data quality check reports and can diagnose an effective cycle of data improvement throughout the organization.

However, changes taking place in any medium to large organization can be complex and difficult to track. Poor organizational synergy and distribution of business processes – coupled with the usage of a myriad of disparate technologies and process standards – often make it impossible to realistically gauge the impact of any prescribed change for data quality improvement. This highlights the necessity of lineage capabilities within an organization. Lineage improves visibility and fastens the application of fixes and changes to complete the data quality improvement cycle.

The InfoSphere Advantage

IBM's InfoSphere suite of data integration products provides integrated capabilities of data quality analysis, diagnosis and fixes with lineage capabilities which are available every step of the way. InfoSphere provides a broad platform for any such data quality management initiative.



InfoSphere **Information Analyzer** provides the tools to **understand** the data and to assimilate data modeling insight. Additionally data stewards can use InfoSphere's **Business Glossary** to provide organizational level definitions of business terminology which can then be used and referenced by individuals throughout the organization.

In particular, the following broad data quality checks can be performed using Information Analyzer.

- **Column Analysis** gathers statistics on column level data and helps discover metadata as reflected by the existing data and hence highlights data quality problems. It unearths data redundancies, ascertains the true domain of the data values, and provides frequency distributions to properly represent the nature of the available data.

- **Primary** and **Foreign Key Analysis** identifies the true primary and foreign key dependencies that exist among tables, as reflected by the existing data.

- **Cross-Domain Analysis** compares data across tables for overlap of domain, and identifies anomalies and redundancies.

- **Benchmark Analysis** provides the means to gauge temporal change of data quality by comparing results across time. This is the final grading system for the data quality management effort. Lack of improvements over time indicate a need for change in the overall strategy.

The results of the data profiling done with Information Analyzer can then be used with **InfoSphere QualityStage** to **cleanse** the data in question. Apart from the data profiling capabilities within Information Analyzer, **QualityStage** has the following capabilities.

- **Investigate** – extends the column and domain analysis capabilities with the ability to analyze free form texts.

- **Standardize** – moves free form data into corrected columns and transforms them to conform to valid standards.

- **Match** – identifies data redundancy and provides house holding of individuals and business entities through the creation of match groups.

- **Survive** – consolidates and provides best available for data entities, missing values using the match results.

Hence InfoSphere provides a comprehensive platform for data quality checks and correction.

Workbench Solves the Lineage Problem

InfoSphere Metadata workbench provides a comprehensive one point view into usage, manipulation and changes made to data as it flows through a myriad of Data Integration systems, providing an easy way of gauging impacts of changes. This makes the entire data quality management program foolproof, with rapid visibility and insight into the data and its movements.

The PR3 Advantage

Being a value added reseller and business partner of IBM, PR3 Systems provides a unique synergy of business knowledge, and technical foresight and aptitude. We have a proven track record of rapid development and successful deployment of Business Solutions geared towards Business Intelligence efforts.

Conclusion

Implementing a resilient and effective data quality management program requires the right organizational temperament, people participation and technologies. We at PR3 Systems provide the glue to reinforce all these components, empowering your organization on its path to ever growing success.

PR3 Systems *helps their clients to extract critical and strategic information regarding their business from islands of scattered data within their organization. Their services include training, consulting, hardware, software, and strategic road mapping to provide a scalable cost-effective complete end-to-end solution to achieve client goals. For more information on this topic, or to contact us regarding our DataStage consulting and training services: call us at 630-364-1469, email us at info@pr3systems.com or visit our web site at www.pr3systems.com.*