TDWI

MONOGRAPH SERIES

OCTOBER 2007

Unifying the Practices of Data Profiling, Integration, and Quality (dPIQ)

By Philip Russom

Senior Manager, TDWI Research
The Data Warehousing Institute

SPONSORED BY

DATAFLUX
A sas COMPANY

tdwi
THE DATA WAREHOUSING INSTITUTE

Table of Contents

About the Author

PHILIP RUSSOM is the senior manager of TDWI Research for TDWI, where he oversees many of TDWI's research-oriented publications, services, awards, and events. Prior to joining TDWI in 2005, Russom was an industry analyst covering BI at Forrester Research, Giga Information Group, and Hurwitz Group. He's also run his own business as an independent industry analyst and BI consultant and was contributing editor with *Intelligent Enterprise* and *DM Review* magazines. Before that, Russom worked in technical and marketing positions for various database vendors. You can reach him at prussom@tdwi.org.

About Our Sponsor

DataFlux enables organizations to analyze, improve, and control their data through an integrated technology platform. With DataFlux enterprise data quality and data integration products, organizations can more effectively and efficiently build a unified view of customers, products, suppliers or any other corporate data asset. A wholly owned subsidiary of SAS, DataFlux customers can rapidly assess and improve problematic data, building the foundation for enterprise data governance. Effective data governance delivers high-quality information that can fuel successful enterprise efforts such as risk management, operational efficiency, and master data management (MDM). To learn more about DataFlux, visit www.dataflux.com.

Defining dPIQ

Data profiling, data integration, and data quality go together like bread, peanut butter, and jam, because all three address related issues in data assessment, acquisition, and improvement. Because they overlap and complement each other, the three are progressively practiced in tandem, often by the same team within the same data-driven initiative. Hence, there are good reasons and ample precedence for bringing the three related practices together. The result is an integrated practice for data profiling, integration, and quality, succinctly named by the acronym dPIQ (pronounced "DEE pick"). TDWI defines dPIQ as:

> A unified practice that coordinates—but does not consolidate—the related practices of data profiling, data integration, and data quality.

Since data profiling is best done as a subset of a data quality or data integration project, dPIQ is mostly about coordinating quality and integration teams (or combining them into one), as well as expanding the interoperability of the software solutions they build. But dPIQ also involves boosting the amount and depth of profiling and its twin sibling, data monitoring. Note that dPIQ coordinates—even unifies—data profiling, data integration, and data quality, but it stops short of consolidating them into a single practice.

dPIQ is the result of multiple trends

In recent years, a number of trends have driven together the three data management practices of data profiling, data integration, and data quality:

dPIQ practices are integrating because they're complementary, apply to common initiatives, and can share staff.

The three complement each other. Data integration and data quality are still independent practices, but they are progressively applied in tandem, because they complement each other so well. For example, data integration projects ferret out data quality issues. Likewise, data quality projects depend on integration technologies to access and enhance corporate data. Data profiling has become ubiquitous as a subset of data integration and data quality practices, because it extends these with data discovery and quality monitoring functions.

The three are commonly applied together in initiatives of certain types. It's now common to find tools and techniques for data profiling, data integration, and data quality applied to data-driven initiatives like data warehousing, customer data integration, master data management, and database consolidations and migrations. The success of these initiatives depends on the accuracy, timeliness, and rich content of the datasets they deliver. Applying data profiling, data integration, and data quality in tandem helps deliver the best dataset possible.

The three are sometimes staffed by the same team. Due to IT centralization, data integration and quality specialists are sometimes consolidated into a single competency center. This facilitates the coordination of data profiling, data integration, and data quality projects, so you can synch project deliverables, avoid stepping on each other's work, and potentially raise the value of data deliverables beyond what individual practices can do. The combined team gives managers more flexible options for staffing projects and fosters productivity goals, such as reuse. And collaboration improves for both technical and business personnel (such as data stewards and business analysts).

dPIQ improves data-driven initiatives that embed it

- **Data warehousing**—This is dPIQ's "killer app." Data warehousing professionals have long practiced data profiling, integration, and quality separately. Many have recently unified them.

- **Business intelligence (BI)**—dPIQ improves the data content of reports for greater accuracy in executive decisions, efficiency in daily operations, and credibility in compliance activities.
- **Operational data integration**—Database consolidations and migrations are a growth area for data integration. These invariably reveal data quality problems, and data profiling can compare source and target data to validate that a database was processed properly.
- **Semantic data**—dPIQ (whether as a unified practice or individual ones) is usually applied to physical data. But dPIQ can also operate on semantic data, such as metadata and master data.

EXPERT COMMENT: dPIQ is fundamental to master data management

David Loshin, a recognized expert in information management, sees dPIQ as fundamental to a successful master data management (MDM) initiative. “Data profiling, data integration, and data quality tools are essentially the three pillars upon which today’s MDM solutions are supported. It is very common that master data integration programs have evolved from customer data quality, product data quality, data assessment and validation, and data integration activities. While targeted solutions have been developed to support MDM functionality, they are often triggered by the introduction of data quality activities to support technical infrastructure acquired for a specific purpose (e.g., enterprise resource planning or customer relationship management). A common success theme is the introduction of data governance across data integration functions.”[1]

dPIQ is iterative and cyclic by nature—and that’s a challenge

Data profiling, integration, and quality practices are each iterative and cyclic. For example, development usually involves iterative prototyping and review, as well as repetitive tasks such as ad hoc query and job testing. In production, a deployed solution is usually scheduled to repeat its operations daily or nightly. In the long term, a solution evolves through lifecycle stages that involve recurring revision and update.

dPIQ is a whirlwind of short-term iterations and long-term cycles.

Furthermore, relationships among the three practices evolve through lifecycle stages. For example, in the early development stages of a solution, these practices are most often performed in the following order: data profiling, data integration, and data quality (hence, the order of letters in the acronym dPIQ). But in production, when data profiling is less like data discovery and more like data monitoring, the order may change to data integration, data quality, and data profiling.

The iterations and cycles of data profiling, data integration, and data quality are similar. After all, each is a data management practice subject to standard development processes; yet, the iterations and cycles have subtle differences. This is an issue for dPIQ, which seeks to coordinate the multiple cycles of the three practices involved. In other words, the leading challenge to dPIQ is aligning multiple similar-but-different planning, development, and production cycles. Given the diversity of the cycles, aligning them is seldom precise and demands flexibility daily to deal with their unpredictable variations. Yet, coordinating the cycles of dPIQ is worth the effort, because it yields collaborative synergies in development and high-quality data deliverables in production.

dPIQ’s ultimate goal is to add value to data

Adding value to data, based on user needs, should be every data management professional’s goal.

Data profiling, integration, and quality practices add value to data—in multiple ways—whether you realize it or not. For example, data integration jobs don’t just copy and move data. They aggregate and remodel data to create data values and data structures that don’t exist elsewhere; in turn, these enable reporting and analysis activities that otherwise wouldn’t be possible. Likewise,

[1] David Loshin, *The Pillars of Master Data Management: Data Profiling, Data Integration, and Data Quality*, a Business Intelligence Network Research Report, May 6, 2007 (www.beyeresearch.com).

data quality techniques don't just cleanse data; techniques like standardization and data append add value by repurposing and augmenting data. Ironically, de-duplication adds value to data by reducing its redundancies. As another example, data profiling doesn't just catalog data and its anomalies; it reveals opportunities for value-adding actions by data integration and data quality techniques. And it's not just physical data; data profiling, integration, and quality—as individual practices—can also add value to metadata, master data, and other semantic data.

The value-add process is yet another characteristic that data profiling, integration, and quality have in common, and another reason why they are strongly associated. As technical users combine these into the unified practice of dPIQ, they must keep the value-add process of each intact, while leveraging synergies of the combined practice to discover ways to increase the value of data.

Also, focusing on the value-add process is a useful mindset for any data management professional. The mindset leads you to find ways of improving data, instead of merely managing it. When data improvements are driven by business and end-user requirements, the value-add process yields data products that the business recognizes as high-quality and relevant. With any luck, this recognition leads to more support of data management from the business.

Cycles and Dependencies in Data Profiling, Integration, and Quality

Before we dive into the details of coordinating best practices for data quality, data integration, and data profiling, let's look at each individually, to identify the cycles inherent within each, as well as relevant dependencies among them.

Data Profiling

- **Data profiling helps you plan better, so projects execute better.** The deliverables expected of data integration and data quality solutions are integrated data and cleansed data—not data profiles. Since data profiling doesn't produce the ultimate deliverable, it's hard to rationalize the time and resources committed to it. Even so, don't scrimp on data profiling, because it gives data integration and data quality solutions higher-quality data deliverables, more accurate project scope, and a reduction of "gotchas" that pop up in testing and deployment.

- **Data profiling is a recurring task, not a one-shot deal.** You need to profile or reprofile data as new data sources become available, as you proactively look for ways to add value to data, and as data integration, data quality, and other software solutions expand.

Data profiling and data monitoring are similar but different.

- **Data profiling is really two practices, applied at different project phases.** Most data profiling is practiced for the purpose of data discovery and documentation in early project phases, like planning, design, and development. However, in later phases, such as production, data profiling morphs into *data monitoring*, where it reprofiles data daily after each data integration or data quality run, to assure that data was processed correctly and that the quality of data is continuously improving. Hence, profiling is part of a larger project lifecycle, whereas monitoring is an iterative daily task. A full-blown dPIQ approach incorporates both.

Profiling and monitoring need the automation of a software tool.

- **Data profiling with a vendor tool has advantages over manual methods.** Most users profile data by unpredictably running ad hoc queries to view table column contents and recording, in word processing files, what the queries revealed. This amasses documentation that's seldom updated and difficult to apply directly to a data integration or data quality solution. This practice is neither comprehensive nor methodical, so users may miss important data and relationships across data structures. Therefore, manual methods of this type are inferior, because of the time-consuming and error-prone process of moving profile information from a query to the documentation to the data management solution. When possible, users should profile with a vendor tool (whether it's an autonomous tool or a collection of functions within a larger tool), to get greater accuracy, completeness, repeatability, and productivity.

- **Data monitoring is nearly impossible without a tool.** Monitoring recurs daily and it records the state of the data monitored, so doing it manually would rack up a lot of time and payroll expense. For efficiency and speed, this iterative daily task is best done with the automation of a software tool. In addition, the tool should support business rules that can identify errors (like an address that's not verified), as well as data items that don't meet corporate standards (for example, the loan-to-value ratio is .8, but the loan does not have private mortgage insurance).

- **The choice of a data profiling tool has ramifications for dPIQ integration.** Users have options, including mostly manual methods, a separate tool dedicated to just data profiling, or a collection of profiling functions within a data integration or data quality tool. With dPIQ, the challenge is to integrate information from the data profiling solution into the data integration and data quality solutions. At one extreme, manual methods are the least conducive to this, whereas—at the other extreme—profile functions built into a tool are the most conducive, especially when the data integration and quality tools are from the same vendor. Autonomous profiling tools support unified modeling language export and metadata interchange standards to facilitate tool integration. Note that data monitoring has advantages when directly executed by or called from a data integration or data quality tool. Therefore, users should consider these issues and test interoperability among tools before deciding which tool approach to take for data profiling.

Data Integration

Any form of data integration can be unified into dPIQ.

- **Data integration takes many forms, but all benefit from data profiling and data quality.** Several forms of data integration exist today, including extract, transform, and load (ETL), enterprise information integration (EII), enterprise application integration (EAI), and database replication. Solutions in any form can be hand-coded from scratch, built atop a vendor's tool, or both. Regardless of their form or method of implementation, all data integration solutions benefit from coordination with data profiling and data quality solutions.

- **Data profiling can distinguish relative value of data sources for data integration.** Data integration, especially for data warehousing, sometimes fails to find the "best source" of data for a given reporting or analysis purpose. This is often the case when similar data is available from multiple sources. Data profiling helps avoid such confusion.

Data integration and data quality solutions have tremendous synergy and interplay.

- **Data integration inevitably reveals data quality issues.** This is true, whether the issues are problems requiring a fix or optional opportunities for enhancing data. When users scrimp on data profiling, data quality problems are a threat to the success of a data integration project.[2]

- **Integration and quality tool capabilities overlap.** For example, the data transformations that an ETL tool does can resemble the data standardizations that a data quality tool does. Users creatively decide which tool is better for a given transform.

[2] As Colin White points out, "Data quality issues are the leading inhibitor to successful data integration projects." See the TDWI Best Practices Report, *Data Integration: Using ETL, EAI, and EII Tools to Create an Integrated Enterprise*, online at: www.tdwi.org/research.

- **Integration and quality tools may pass data back and forth multiple times daily.** In a complex data flow, a data integration job may access data in legacy or uncommon platforms and pass the data directly to a data quality tool that's unable to access those platforms. And a data quality tool may make a first pass to standardize data in the flow, which makes merging data easier for the data integration tool. Note that some data quality functions—such as match-and-merge, de-duplication, and house-holding—require that data be collected into a single database or file, in a single data model. In these cases, data integration jobs must first integrate data from multiples sources (and, therefore, multiple data models) into the model required for these data quality functions.

- **The first rule of data integration is: never load suspect data into a target database.** That's because nonstandard, redundant, or dirty data is difficult to find and fix once it's mixed with cleansed data. For instance, a best practice in business intelligence is to stage integrated data and cleanse it before ETL jobs load it into a data warehouse. A related practice is to profile integrated data (not just source data) to determine what extra processing ETL and data quality tools must do before loading a data warehouse.

Data Quality

- **Data quality is data profiling's original "killer app."** Today, the two are nearly inseparable, and the current way of practicing data quality wouldn't be possible without data profiling and data monitoring.

Maintaining the quality of data requires iterative cleansing.

- **Improving data's quality is a recurring task.** The quality of data degrades as applications update, add data to, or delete data from a database. Most estimates say that 10–12% of the data in an active database becomes dirty, nonstandard, or redundant each month. Hence, if you cleanse a database 100% today, it will only be 88–90% clean 30 days from now. This means that a software solution for data quality must operate on data frequently. Volatile data requires daily cleansing, whereas less volatile data may be cleansed less frequently.

Raising the quality of data requires metrics and iterative monitoring.

- **Data quality's goal of "continuous improvement" requires data monitoring.** To get the most out of data monitoring, users should develop metrics for quantifying the state of a dataset's quality. After each cleansing iteration, data monitoring functions should reprofile data to quantify its quality metrics, then store the metric values in a database to keep a history of data quality performance. Users can then analyze the history to understand trends in data quality, which helps them find problem areas and opportunities for higher levels of quality. When data monitoring is automated with a tool that supports business rules for assessing data quality, the rules categorize issues accurately so they are more easily and quickly addressed.

- **Data quality practices have structure and methodology.** Data quality is regularly practiced as part of a broader quality initiative based on methods like Six Sigma, Zero Defect, Performance Management, and ISO 900x. Hence, data quality has absorbed a lot of useful structure from these methods, making it the least ad hoc of the dPIQ triumvirate.

- **Data quality best practices have positively influenced those for data integration.** As the two have been applied more closely, data quality's deep use of data profiling has migrated into data integration. Also, data stewardship—as applied to data quality initiatives—now appears in some data integration initiatives.

- **Some data quality techniques apply to semantic data.** Although data quality practices are applied mostly to physical data, a few users apply data quality techniques (especially standardization) to semantic data (such as metadata and master data).

- **As data quality initiatives mature through lifecycle stages, they add more functionality.** For example, most organizations begin with name-and-address cleansing or another form of standardization. Later, they add data verification and data monitoring. Mature initiatives include forms of matching, like match-and-merge, de-duplication, and house-holding. Specialized data quality tasks tend to arrive in middle or later stages, like data enhancement (appending demographic data to customer records or D-U-N-S numbers to supplier records) and internationalization (non-U.S. postal standards and non-English natural languages).[3]

The Unified dPIQ Cycle

Now that we've discussed the details of dPIQ, including the cycles and dependencies of its individual practices, it's time to consider how you might unify them into a single recurring cycle.[4] One variant of the unified dPIQ cycle described here is visualized in Figure 1. Note that a unified dPIQ cycle has many possible variations, and many of its cyclic stages and iterative tasks are optional as the cycle repeats.

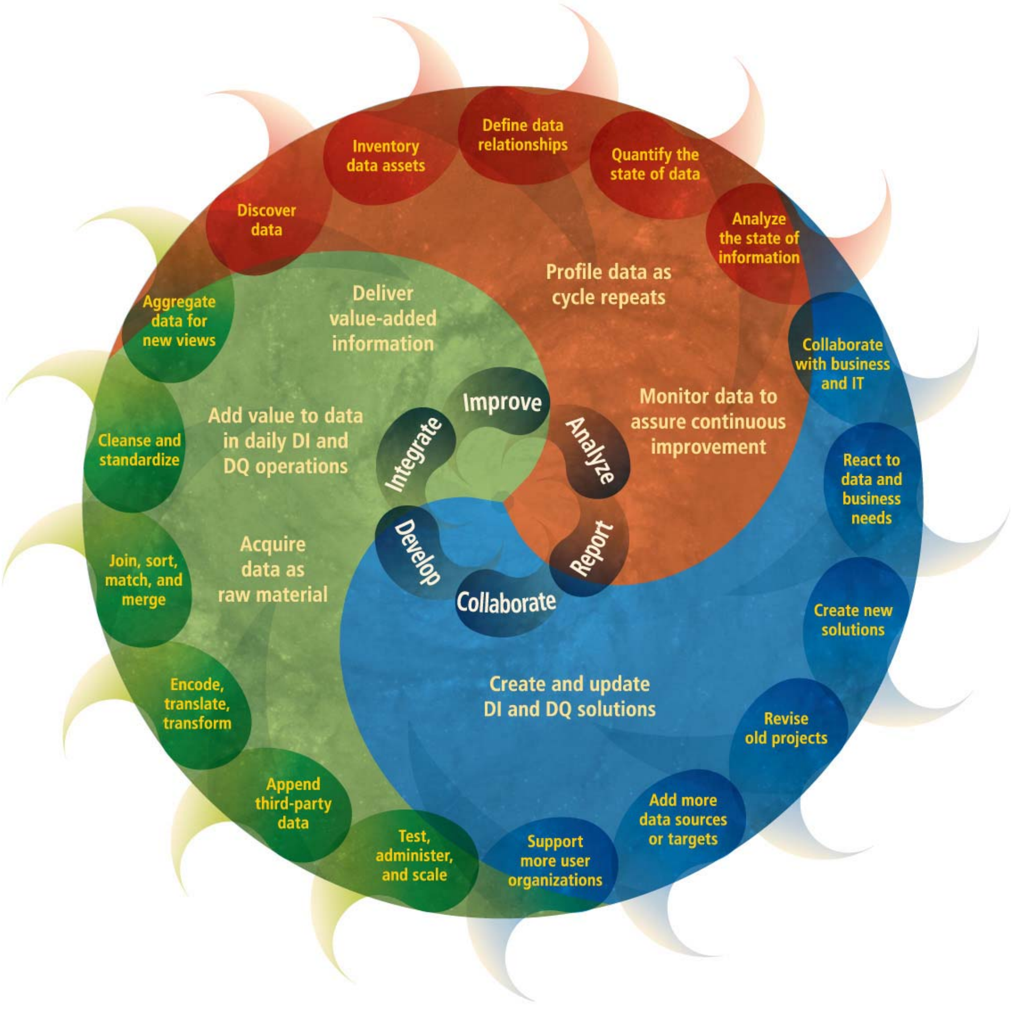
Inventory data assets
Define data relationships
Quantify the state of data
Discover data
Analyze the state of information
Profile data as cycle repeats
Deliver value-added information
Aggregate data for new views
Collaborate with business and IT
Monitor data to assure continuous improvement
Add value to data in daily DI and DQ operations
Improve
Integrate
Analyze
Cleanse and standardize
React to data and business needs
Develop
Report
Collaborate
Acquire data as raw material
Join, sort, match, and merge
Create new solutions
Create and update DI and DQ solutions
Encode, translate, transform
Revise old projects
Append third-party data
Add more data sources or targets
Test, administer, and scale
Support more user organizations

Figure 1. One way to unify the dPIQ cycle.

[3] For more details, see the TDWI Best Practices Report, *Taking Data Quality to the Enterprise through Data Governance*, online at www.tdwi.org/research.

[4] The graphic and some of the content of this section come from the TDWI 2006 Poster titled *Symbiotic Cycles of Data Profiling, Integration, and Quality*.

Analyze and Report

Discovering data problems and opportunities is cyclic and iterative. As part of long-term cycles, you profile data at the beginning of a new development project or a project update. Once a solution is in production, data profiling evolves into data monitoring, which iteratively re-measures data after data quality or data integration solutions run.

Profile data as the dPIQ cycle begins or repeats.

A recurring dPIQ cycle will either profile data in development or monitor data in production.

- **Discover data.** Start by becoming familiar with the tables, records, and other data structures found in databases, flat files, and other data sources. As you record these, also note the state of their quality. For instance, a common metric for data quality is a statistic describing the diverse values found in a table column. A profile is usually a mix of information about data sources, structures within them, metadata for these, and statistics about their values.

- **Define data relationships.** Don't just record tables and other common data structures. Also record the keys that relate them, as well as less obvious dependencies and redundant data.

- **Inventory data assets.** To document your discoveries, record an inventory of the data assets you discovered. As discussed earlier, developing data profiles with a vendor tool is more productive and accurate than documenting data structures in word processing files (an all-too-common practice). Ideally, profiles should be in a repository from which they can be shared.

Monitor data to assure continuous improvement.

- **Quantify the state of data.** Start by developing metrics for measuring the quality of data. These may be hierarchical, in that granular measures represent the state of individual table columns (a typical focus for profiling and monitoring), which roll up into metrics for each table, which in turn roll up into a key performance indicator representing one or more databases. As data monitoring requantifies these metrics, keep a history of them in a time series in a database. The repository found in most dPIQ tools enables this. But you might also build your own repository or use your data warehouse.

- **Analyze the state of information.** Using your database of metrics, you can report on and analyze trends in data quality, which helps you determine whether data is continuously improving (an important goal for most data quality initiatives). You can also use reports based on the database of metrics to corroborate compliance with regulatory requirements, quality methods (like ISO and Six Sigma), and service level agreements.

Note that data monitoring is most often applied to data quality, but it can also assure that data integration jobs have processed data correctly. You might poll source systems periodically, looking for changes that affect data integration and data quality solutions.

Collaborate and Develop

The development of data integration and data quality solutions is an iterative cycle of profiling, analysis, design, and testing. Deployment is iterative, too, when it involves multiple rollout and hand-off phases. The dialog of collaboration—whether among technical personnel only or also including business people—is likewise iterative in the short term. As collaboration, development, and deployment are revisited for project updates, they become part of a long-term cycle.

Create and update DI and DQ solutions.

dPIQ success depends on meeting business management needs first, data management needs second.

- **Collaborate with business and IT.** Though time-consuming and sometimes awkward, a collaborative dialog between carefully chosen business and IT personnel is the only way to ensure that the technical solution meets business requirements.

- **React to data and business needs.** In addition to business requirements, data integration and quality solutions involve a mix of data requirements, due to required fixes versus optional opportunities. Sometimes integration and quality needs within dPIQ will compete for resources. Note that dPIQ doesn't just unify practices; it unifies their requirements and priorities, too. When in doubt, give priority to data needs that support business needs.

- **Create new solutions.** Unified dPIQ is still new, so its inception may coincide with—and complicate—the development of a new data integration or quality solution. To reduce risk, initiate dPIQ via a separate project that precedes new development or a project update.

dPIQ is most successful with a central, cross-trained team.

- **Revise old projects.** A project update is an opportunity to initiate dPIQ, and doing so with an update is less risky than with new development. This can be done in subtle ways, like assigning integration personnel to work on integration routines, and vice versa. Unified dPIQ assumes that personnel are cross-trained and available from a single pool, so managers can flexibly assign personnel as work arises.

- **Add more data sources or targets.** This is the bulk of work for most dPIQ project updates. As integration and quality solutions are expanded, expect to profile the new sources. Don't forget to compare these to datasets previously profiled, so no dependencies are missed. When adding new targets, adjust data monitoring functions to include them.

- **Support more user organizations.** This is the real reason for most project updates involving dPIQ. You're not just adding more data and processing to the scope of dPIQ-based solutions; you're also adding the user organizations that depend on the data, so you have to understand and meet their needs as best as you can.

Integrate and Improve

When deployed into production, data integration, quality, and monitoring solutions are obviously iterative, in that these solutions—or rather the jobs and routines that are their production components—run and run again on a scheduled basis. Here's where the rubber hits the road, because it's in production that dPIQ-based solutions produce the deliverable that justifies their existence: high-quality integrated and cleansed data.

Acquire data as raw material.

- **Test, administer, and scale.** Transitioning new project work from development to production requires short-term testing and long-term administration that leads to scalable solutions. This is challenging with dPIQ, because—even when they come from a single vendor's suite—you almost always have separate servers for data integration and quality, and maybe another for data monitoring. Coordinating jobs and optimizing for performance is complicated with multiple servers. Administrators need to schedule integration, quality, and monitoring jobs carefully to get an order of execution that yields optimal performance without threats to data integrity, and monitor performance and reliability over time to discover more optimizations.

- **Append third-party data.** One of dPIQ's value-adding activities is to acquire data from multiple sources and aggregate it. When sources are exclusively internal, the resulting view of corporate entities can be myopic—a myopia that data from external sources cures. For instance, a common data quality measure is to buy information about your customers and append this to customers' records. The growing practice of inter-enterprise data integration acquires data from suppliers and integrates it into internal product catalogs. Thus, appending third-party data adds value to data in a way that can't be done with internal sources exclusively.

Add value to data in daily DI and DQ operations.

- **Encode, translate, transform.** Data integration and data quality practices add value to data by changing it from its source model to a data model that suits another purpose. This is apparent when ETL jobs transform data into models conducive to reporting and analysis. And it's equally apparent in the standardization functions of a data quality solution.

- **Join, sort, match, and merge.** dPIQ also adds value to data by combining it from multiple sources, such that new datasets result. For instance, consider the complex table joins of data integration. Joined data typically suffers from redundancy, like multiple records for the same customer or product. Redundancy is reduced by data quality's functions for match-and-merge, house-holding, and de-duplication.

- **Cleanse and standardize.** The order of execution of data integration jobs versus data quality jobs can be significant. For example, with most data quality tools, matching functions can't match redundant data unless it's already integrated into one dataset. It may be desirable for a data quality job to standardize source data before a data integration job integrates it, if that will yield more accurate table joins and queries.

- **Aggregate data for new views.** Yet another way that dPIQ adds value to data is by re-aggregating it into various views. For example, once all your customer records have been integrated into a single dataset and cleansed, you might pull from that dataset only those columns that are needed for a customer segmentation analysis. Or a subset might be joined with integrated and cleansed financial data for the sake of profitability analysis. The same dataset might get sorted by ZIP code before being loaded into a geography-based sales application, then sorted by customer ID before loading into a CRM application.

Deliver value-added information.

At the end of a production iteration is where dPIQ-based solutions produce their data deliverables—namely, high-quality datasets that have had value added to them through integration, cleansing, standardization, transformation, and aggregation. Once these datasets are loaded into their target databases, applications, and files, the dPIQ cycle begins anew, either by monitoring the quality of these datasets before another production run or by profiling data sources in preparation for new development activities.

Recommendations

dPIQ is about coordinating data profiling, integration, and quality practices, which in turn is about aligning their cycles.

- **Consider unifying data profiling, integration, and quality practices.** These are coming together because they're complementary, apply to common initiatives, and can share staff. Unified dPIQ results in efficiency and synergy in development and higher-quality data.

- **Recognize that all three dPIQ practices are cyclic and iterative.** That's because data has a life of its own, and its content and quality evolve as users, applications, and processes interact with it. One ramification is that you must profile, integrate, and cleanse data repeatedly to capture its current content and state, as well as improve same.

- **Align the cycles of data profiling, integration, and quality practices.** After all, this is what dPIQ is about on a practical level. Readers should start with the illustration in Figure 1—as one example of such an alignment—and adapt it to their situations.

Data integration and data quality techniques complement each other.

- **Coordinate data quality with data integration and vice versa.** Data integration inexorably ferrets out data quality problems, and data quality is hamstrung without help accessing and joining diverse data. Coordinating these is best enabled by a central repository for sharing dPIQ collaborative documents, common metadata, data profiles, and development artifacts.

- **Expect data integration and data quality jobs to bat data back and forth.** This is a sign of a mature dPIQ-based solution, because it's leveraging the synergies of the two complementary practices to add value to data beyond what each can do singly.

Profile data deeply, repeatedly, and with a productive tool.

- **Profile source data carefully, or suffer the consequences.** Data profiling gives data integration and data quality solutions higher-quality data deliverables, more accurate project scope, and a reduction of "gotchas" that pop up in testing and deployment. For greatest productivity, profile with a tool instead of manual methods.

- **Never load suspect data into a target database.** That's because nonstandard, redundant, or dirty data is difficult to find and fix once it's mixed with cleansed data. Coordinated data profiling and data quality techniques help data integration to avoid this problem.

- **Apply dPIQ to semantic data, not just physical data.** Data profiling and some data quality functions apply to semantic data, such as metadata, master data, data models, and so on. Data integration can synchronize master data definitions across related information systems, so that master data management is not siloed.

Goals to pursue: adding value to data and the continuous improvement of data.

- **Diligently pursue the continuous improvement of data.** This mindset comes to us from data quality practices, but it applies in some form to all data management practices. With dPIQ, it entails the development of metrics that quantify data's quality and the monitoring of data over time to assure improvement. Note that achieving a higher level of data improvement may require continuous improvements in solutions that process data.

- **Analyze trends in data's quality and usage.** Keep a history of quality metrics and other information about data handled by dPIQ solutions, so you can study the history with standard reporting and analysis tools to understand what drives or inhibits data's improvement.

- **Add value to data, don't just manage it.** This should be every data management professional's noblest aspiration. Profile, integrate, and cleanse data to create new datasets and new views of information that in turn create new business opportunities.