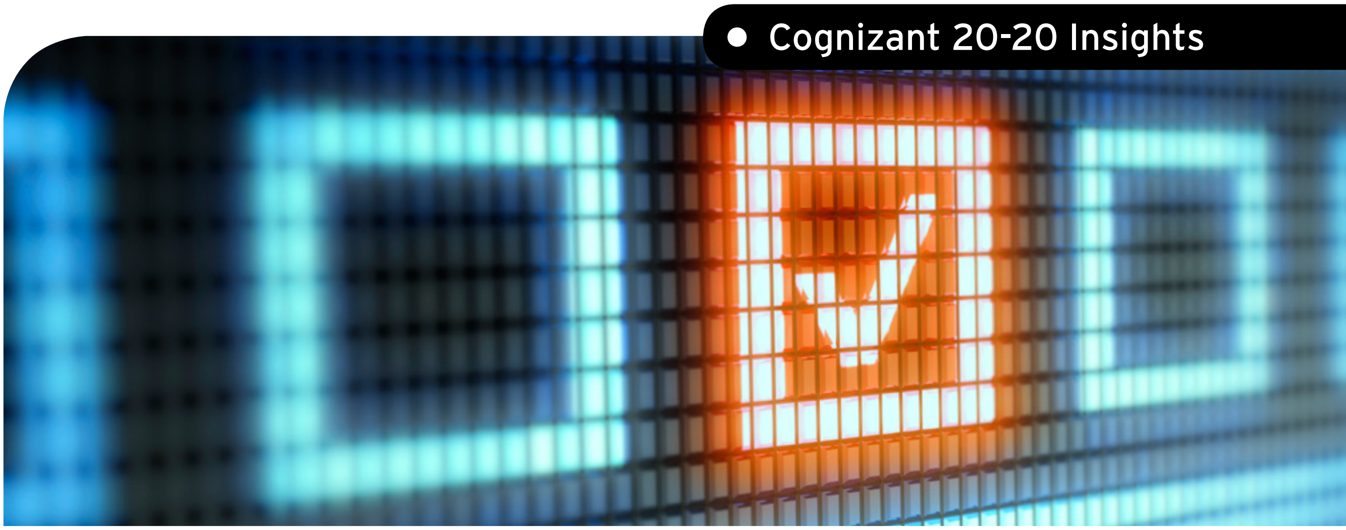


Deliver Trusted Data by Leveraging ETL Testing

Data-rich organizations seeking to assure data quality can systemize the validation process by leveraging automated testing to increase coverage, accuracy and competitive advantage, thus boosting credibility with end users.

Executive Summary

All quality assurance teams use the process of extract, transform and load (ETL) testing with SQL scripting in conjuction with eyeballing the data on Excel spreadsheets. This process can take a huge amount of time and can be error-prone due to human intervention. This process is tedious because to validate data, the same test SQL scripts need to be executed repeatedly. This can lead to a defect leakage due to assorted, capacious and robust data. To test the data effectively, the tester needs advanced database skills that include writing complex join queries and creating stored procedures, triggers and SQL packages.

Manual methods of data validation can also impact the project schedules and undermine end-user confidence regarding data delivery (i.e., delivering data to users via flat files or on Web sites). Moreover, data quality issues can undercut competitive advantage and have an indirect impact on the long-term viability of a company and its products.

Organizations can overcome these challenges by mechanizing the data validation process. But that raises an important question: How can this be done without spending extra money? The answer led us to consider Informatica's ETL testing tool.

This white paper demonstrates how Informatica can be used to automate the data testing process. It also illustrates how this tool can help QE&A teams reduce the numbers of hours spent on their activities, increase coverage and achieve 100% accuracy in validating the data. This means that organizations can deliver complete, repeatable, auditable and trustable test coverage in less time without extending basic SQL skill sets.

Data Validation Challenges

Consistency in the data received for ETL is a perennial challenge. Typically, data received from various sources lacks commonality in how it is formatted and provided. And big data only makes it more pressing an issue. Just a few years ago, 10 million records of data was considered a big deal. Today, the volume of the data stored by enterprises can be in the range of billions and trillions.

Quick Take

Addressing Organizational Data Quality Issues

Our experimentation with automated data validation with a U.S.-based client revealed that by mechanizing the data validation process, data quality issues can be completely eradicated. The automation of the data validation process brings the following value additions:

- Provides a data validation platform which is workable and sustainable for the long term.
- Tailored, project-specific framework for data quality testing.
- Reduces turnaround time of each test execution cycle.
- Simplifies the test management process by simplifying the test approach.
- Increases test coverage along with greater accuracy of validation.

The Data Release Cycle and Internal Challenges

This client releases product data sets on a periodic basis, typically monthly. As a result, the data volume in each release is huge. One product suite has seven different products under its umbrella and data is released in three phases per month. Each phase has more than 50 million records to be processed from each product within the suite. Due to manual testing, the turnaround time for each phase used to be three to five days, depending on the number of tasks involved in each phase. Production release of the quality data is a huge undertaking by the QE&A team, and it was a big challenge to make business owners happy by reducing the time-to-market (i.e., the time from processing the data once it is received to releasing it to the market). By using various automation methods, we were able to reduce time-to-market from between three and five days to between one and three days (see Figure 1).

Data Release Cycle

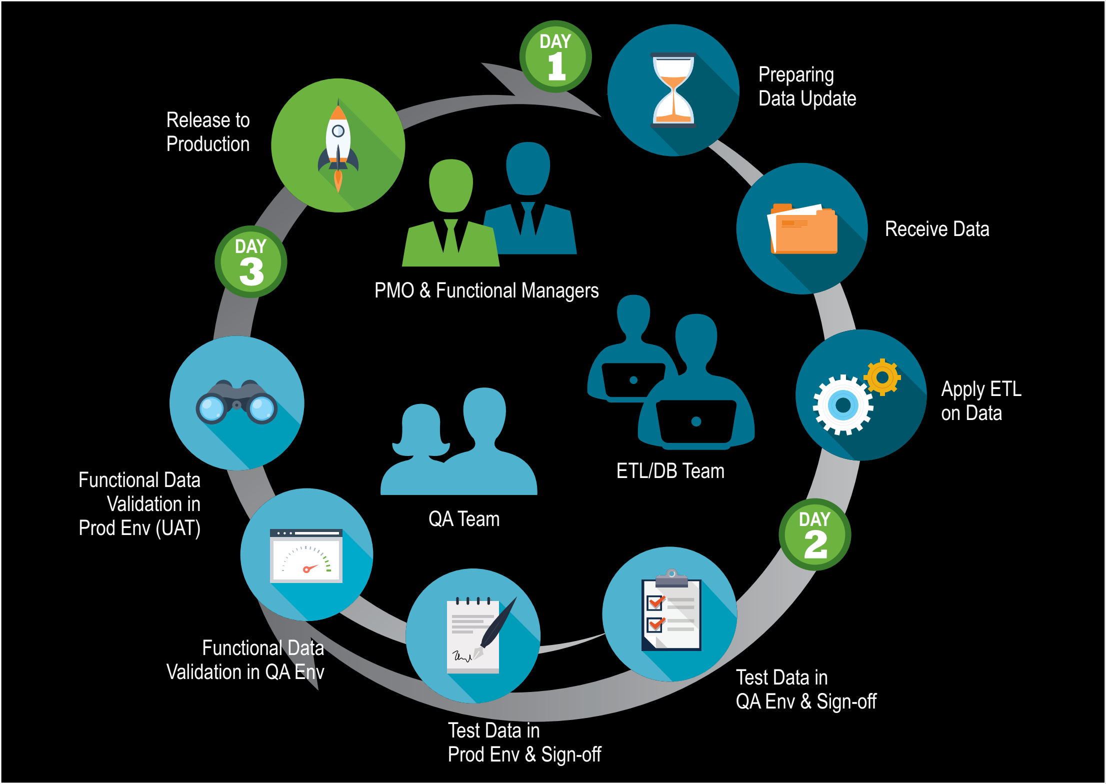

DAY 1
Preparing Data Update
Receive Data
Apply ETL on Data
DAY 2
Test Data in QA Env & Sign-off
Test Data in Prod Env & Sign-off
Functional Data Validation in QA Env
Functional Data Validation in Prod Env (UAT)
DAY 3
Release to Production
PMO & Functional Managers
ETL/DB Team
QA Team

Figure 1

Reasons for accretion of voluminous data include:

- Executive management's need to focus on data-driven decision-making by using business intelligence tools.
- Company-wide infrastructural changes such as data center migrations.
- Mergers and acquisitions among data-producing companies.
- Business owners' need to gain greater insight into streamlining production, reducing time-to-market and increasing product quality.

If the data is abundant, and from multiple sources, there is a chance junk data can be present. Also, odds are there is excessive duplication, null sets and redundant data available in the assortment. And due to mishandling, there is potential loss of the data.

However, organizations must overcome these challenges by having appropriate solutions in place to avoid credibility issues. Thus, for data warehousing and migration initiatives, data validation plays a vital role ensuring overall operational effectiveness. But operational improvements are never without their challenges, including:

- Data validation is significantly different from conventional ways of testing. It requires more advanced scripting skills in multiple SQL servers such as Microsoft SQL 2008, Sybase IQ, Vertica, Netizza, etc.
- Heterogeneity in the data sources leads to mishandling of the interrelationships between multiple data source formats.
- During application upgrades, making sure that older application repository data is the same as the data in the new repository.
- SQL query execution is tedious and cumbersome, because of repetitious execution of the queries.
- Missing test scenarios, due to manual execution of queries.
- Total accuracy may not always be possible.
- Time taken for execution varies from one person to another.
- Strict supervision is required with each test.
- The ETL process entails numerous stages; it can be difficult to adopt a testing schedule given the manual effort required.
- The quality assurance team needs progressive elaboration (i.e., continuous improvement of key processes) to standardize the process due to complex architectures and multilayered designs.

A Business-Driven Approach to Data Validation

To meet the business demand for data validation, we have developed a surefire and comprehensive solution that can be utilized in various areas such as data warehousing, data extraction, transformations, loading, database testing and flat-file validation.

The Informatica tool that is used for the ETL process can also be used as a validation tool to verify the business rules associated with the data. This tool has the capability to significantly reduce manual effort and increase ETL productivity by lowering costs, thereby improving the bottom line.

Our Data Validation Procedures as a Framework

There are four methods required to implement a one-stop solution for addressing data quality issues (see Figure 2).

Data Validation Methods

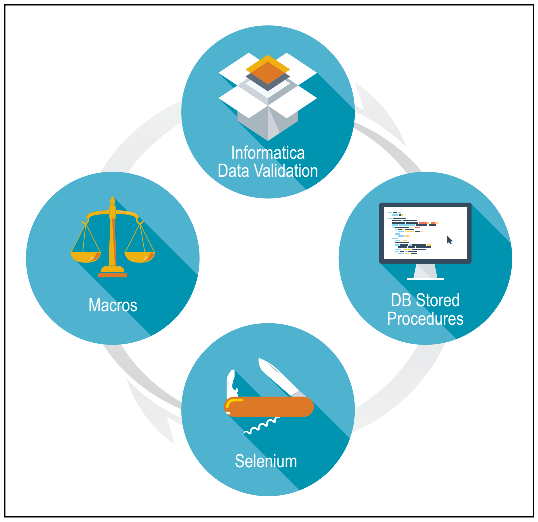


Figure 2

Each methods has its own adoption procedures. High-level details include the following:

Informatica Data Validation

The following activities are required to create an Informatica data validation framework (see Figure 3):

- Accrual of business rules from product/ business owners based on their expectations.
- Convert business rules into test scenarios and test cases.
- Derive expected results of each test case associated with each scenario.
- Write a SQL query for each of the test cases.
- Update the SQL test cases in input files (test case basic info, SQL query).
- Create Informatica workflows to execute the queries and update the results in the respective SQL tables.
- Trigger Informatica workflows for executing jobs and send e-mail notifications with validation results.

Validate Comprehensive Data with Stored Procedures

The following steps are required for data validation using stored procedures (see Figure 4, next page):

- Prepare validation test scenarios.
- Convert test scenarios into test cases.
- Derive the expected results for all test cases.
- Write stored procedure-compatible SQL queries that represent each test case.
- Compile all SQL queries as a package or test build.
- Store all validation transact-SQL statements in a single execution plan, calling it "stored procedure."
- Execute the stored procedure whenever any data validation is carried out.

A Data Validation Framework: Pictorial View

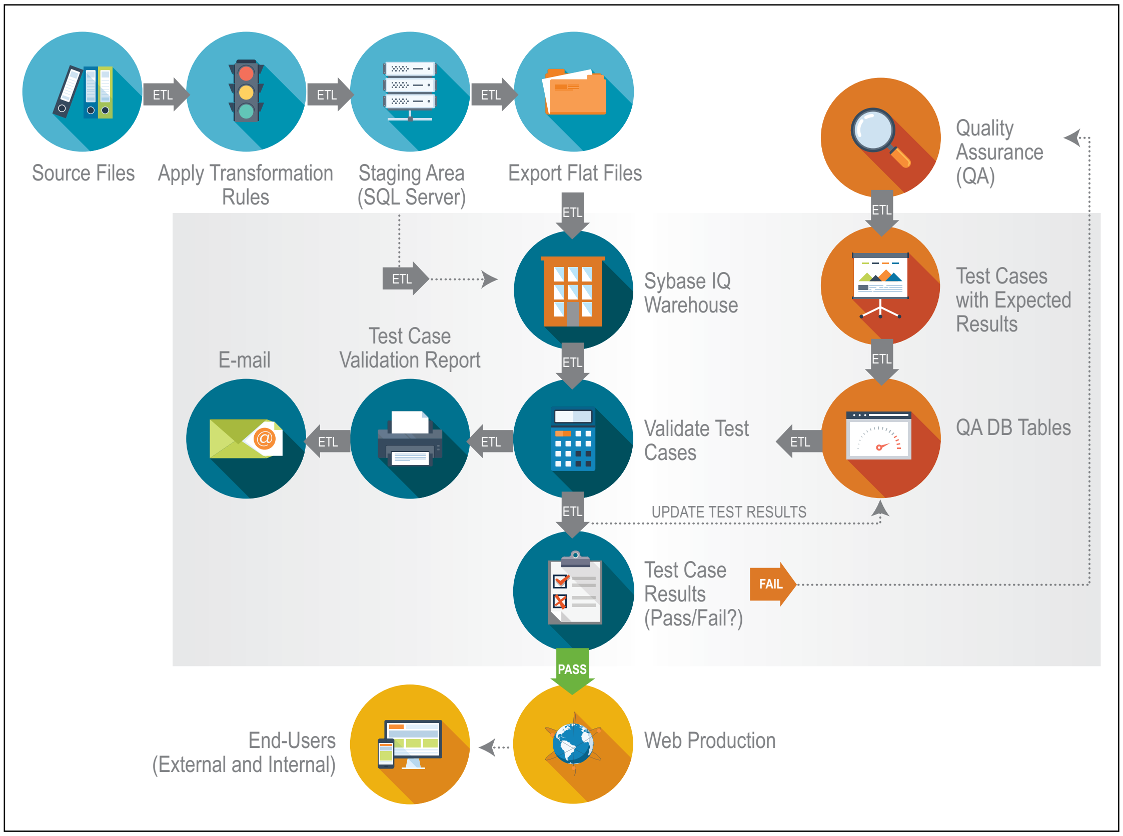
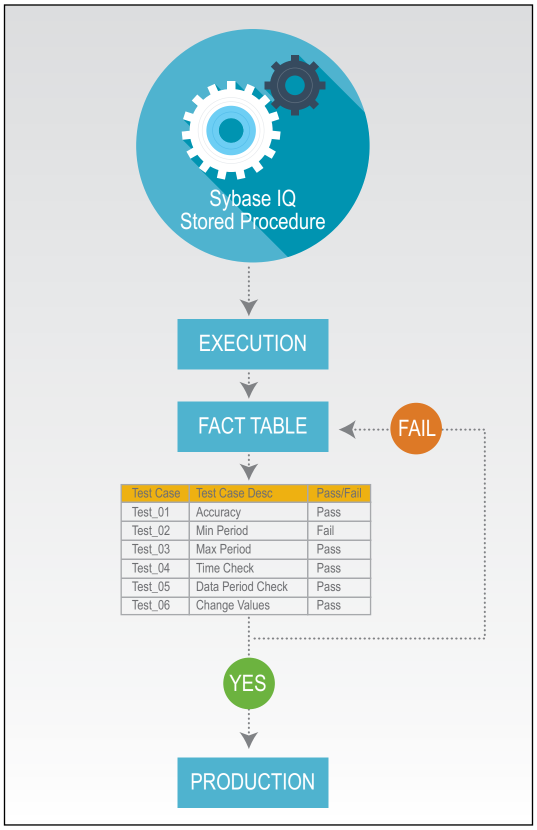


Figure 3

Validating with Stored Procedures



Test Case	Test Case Desc	Pass/Fail
Test_01	Accuracy	Pass
Test_02	Min Period	Fail
Test_03	Max Period	Pass
Test_04	Time Check	Pass
Test_05	Data Period Check	Pass
Test_06	Change Values	Pass

Figure 4

One-to-One Data Comparision Using Macros

The following activities are required to handle data validation (see Figure 5):

- Prepare validation test scenarios.
- Convert test scenarios into test cases.
- Derive a list of expected results for each test scenario.
- Specify input parameters for a given test scenario, if any.
- Write a macro to carry out validation work for one-to-one data comparisions.

Selenium Functional Testing Automation

The following are required to perform data validation (see Figure 6, next page):

- Prepare validation test scenarios.
- Convert test scenarios into test cases.
- Derive an expected result for each test case.
- Specify input parameters for a given test scenario.
- Derive test configuration data for setting up the QA environment.
- Build a test suite that contains multiple test builds according to test scenarios.
- Have a framework containing multiple test suites.
- Execute the automation test suite per the validation requirement.
- Analyze the test results and share those results with project stakeholders.

Salient Solution Features, Benefits Secured

The following features and benefits of our framework were reinforced by a recent client engagement (see sidebar, page 7).

Core Features

- Compatible with all database servers.
- Zero manual intervention for the execution of validation queries.
- 100% efficiency in validating the larger-scale data.
- Reconciliation of production activities with the help of automation.
- Reduces level of effort and resources required to perform ETL testing.

Applying Macro Magic to Data Validation

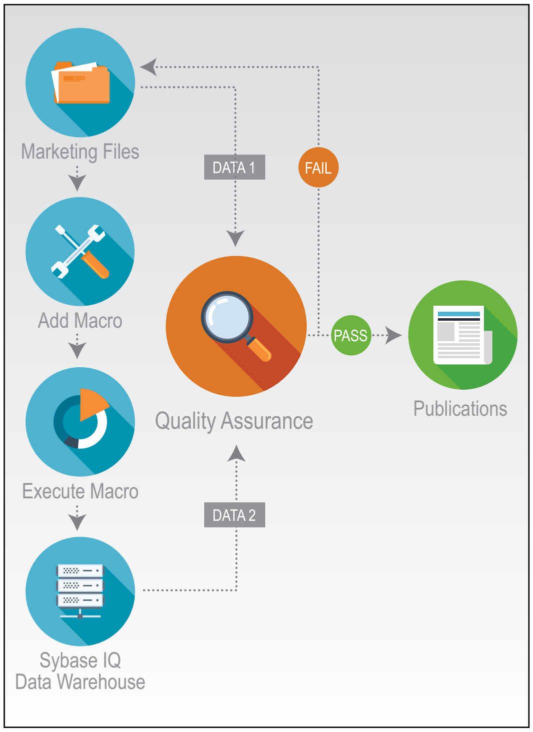


Figure 5

Facilitating Functional Test Automation

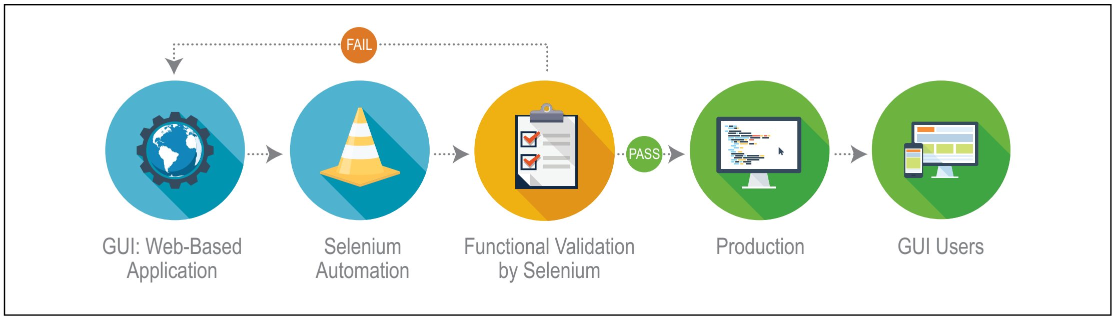


Figure 6

- Comprehensive test coverage ensures lower business risk and greater confidence in data quality.
- Remote scheduling of test activities.

Benefits Reaped

- Test case validation results are in user-friendly formats like .csv, .xlsx and HTML.
- Validation results are stored for future purposes.
- Reuse of test cases and SQL scripts for regression testing.
- No scope for human errors.
- Supervision isn't required while executing test cases.
- 100% accuracy in test case execution at all times.
- Easy maintainance of SQL scripts and related test cases.
- No variation in time taken for execution of test cases.
- 100% reliability on testing and its coverage.

The Bottom Line

As the digital age proceeds, it is very important for organizations to progressively elaborate their processes with suitable information and awareness to drive business success. Hence, business data collected from various operational sources is cleansed and condolidated per the business requirement to separate signals from noise. This data is then stored in a protected environment, for an extended time period.

Fine-tuning this data will help facilitate performance management, tactical and strategic decisions and the execution thereof for business advantage. Well-organized business data enables and empowers business owners to make well-informed decisions. These decisions have the capacity to drive competitive advantage for an organization. On an average, organizations lose $8.2 million annually due to poor data quality, according to industry research on the subject. A study by B2B research firm Sirius Decisions shows that by following best practices in data quality, a company can boost its revenue by 66%.[1] And market research by *Information Week* found that 46% of those surveyed believe data quality is a barrier that undermines business intelligence mandates.[2]

Hence, it is safe to assume poor data quality is undercutting many enterprises. Few have taken the necessary steps to avoid jeopardizing their businesses. By implementing the types of data testing frameworks discussed above, companies can improve their processes by reducing the time taken for ETL. This, in turn, will dramatically reduce their time-to-market turnaround and support the management mantra of "under-promise and over-deliver." Moreover, few organizations need to spend extra money on these frameworks given that existing infrastructure is being used. This has a direct positive impact on a company's bottom line since no additional overhead is required to hire new human resources or add additional infrastructure.

Quick Take

Fixing an Information Services Data Quality Issue

This client is a U.S.-based leading financial services provider for real estate professionals. The services it provides include comprehensive data, analytical and other related services. Powerful insight gained from this knowledge provides the perspective necessary to identify, understand and take decisive action to effectively solve key business challenges.

Challenges Faced

This client faced major challenges in the end-to-end quality assurance of its data, as the majority of the company's test procedures were manual. Because of these manual methods, turnaround time or time-to-market of the data was greater than its competitors. As such, the client wanted a long-term solution to overcome this.

The Solution

Our QE&A team offered various solutions. The focus areas were database and flat file validations. As we explained above, database testing was automated by using Informatica and other conventional methods such as the creation of stored procedures and macros which were used for validating the flat files.

Benefits

- 50% reduction in time-to-market.
- 100% test coverage.
- 82% automation capability.
- Highly improved data quality.

Figure 7 illustrates the breakout of each method used and their contributions to the entire QE&A process.

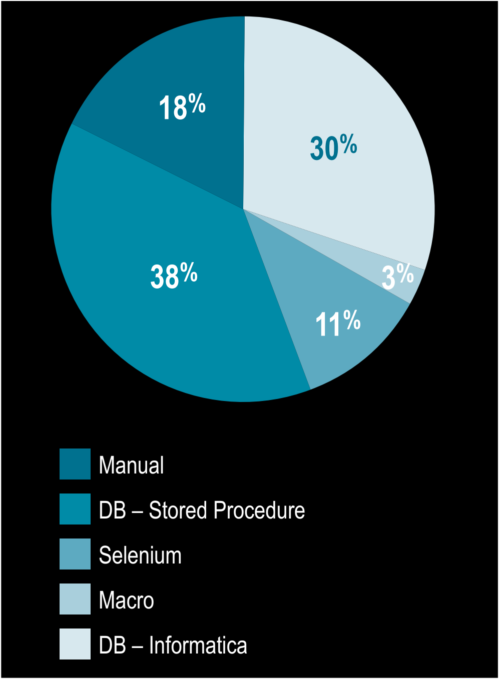


Figure 7

Looking Ahead

In the Internet age, where data is considered a business currency, organizations must capitalize on their return on investment in a most efficient way to maintain their competitive edge. Hence, data quality plays a pivotal role when making strategic decisions.

The impact of poor information quality on business can be measured with four different magnitudes: increased costs, decreased revenues, decreased confidence and increased risk. Thus it is crucial for any organization to implement a foolproof solution where a company can use its own product to validate the quality and capabilities of the product. In other words, adopting an "eating your own dog food" ideology.

Having said that, it is necessary for any data-driven business to focus on data quality, as poor quality has a high probability of becoming the major bottleneck.

Footnotes

[1] "Data Quality Best Practices Boost Revenue by 66 Percent," http://www.destinationcrm.com/Articles/CRM-News/Daily-News/Data-Quality-Best-Practices-Boost-Revenue-by-66-Percent-52324.aspx.

[2] Douglas Henschen, "Research: 2012 BI and Information Management," http://reports.informationweek.com/abstract/81/8574/Business-Intelligence-and-Information-Management/research-2012-bi-and-information-management.html.

References

- CoreLogic U.S., Technical & Product Management, Providing IT Infrastructural Support and Business Knowledge on the Data.
- Cognizant Quality Engineering & Assurance (QE&A) and Enterprise Information Management (EIM), ETL QE&A Architectural Set Up and QE & A Best Practices.
- Ravi Kalakota & Marcia Robinson, *E-Business 2.0: Roadmap for Success*, "Chapter Four: Thinking E-Business Design – More Than a Technology" and "Chapter Five: Constructing The E-Business Architecture-Enterprise Apps."
- Jonathan G. Geiger, "Data Quality Management, The Most Critical Initiative You Can Implement," Intelligent Solutions, Inc., Boulder, CO, www2.sas.com/proceedings/sugi29/098-29.pdf.
- www.informatica.com/in/etl-testing/. (An article on Informatica's proprietary Data Validation Option available in its Data Integration Tool.)
- www.expressanalytics.net/index.php?option=com_content&view=article&id=10&Itemid=8. (Literature on the importance of the data warehouse and business intelligence.)
- http://spotfire.tibco.com/blog/?p=7597. (Understanding the benefits of data warehousing.)
- www.ijsce.org/attachments/File/v3i1/A1391033113.pdf. (Significance of data warehousing and data mining in business applications.)
- www.corelogic.com/about-us/our-company.aspx#container-Overview. (About the CoreLogic client.)
- www.cognizant.com/InsightsWhitepapers/Leveraging-Automated-Data-Validation-to-Reduce-Software-Development-Timelines-and-Enhance-Test-Coverage.pdf. (A white paper on dataTestPro, a proprietary tool by Cognizant used for automating the data validation process.)

About the Author

Vijay Kumar T V is a Senior Business Analyst on Cognizant's QE&A team within the company's Banking and Financial Services Business Unit. He has 11-plus years of experience in business analysis, consulting and quality engineering/assurance. Vijay has worked in various industry segments such as retail, corporate, core banking, rental and mortage, and has an analytic background, predominantly in the areas of data warehousing and business intelligence. His expertise involves automating the data warehouse and business intelligence test practices to align with the client's strategic business goals. Vijay has also worked with a U.S.-based client on product development, business process optimization and business requirement management. He holds a bachelor's degree in mechanical engineering from Bangalore University and a post-graduate certificate in business management from XLRI, Xavier School of Management. Vijay can be reached at Vijay-20.Kumar-20@cognizant.com.

About Cognizant

Cognizant

World Headquarters
500 Frank W. Burr Blvd.
Teaneck, NJ 07666 USA
Phone: +1 201 801 0233
Fax: +1 201 801 0243
Toll Free: +1 888 937 3277
Email: inquiry@cognizant.com

European Headquarters
1 Kingdom Street
Paddington Central
London W2 6BD
Phone: +44 (0) 20 7297 7600
Fax: +44 (0) 20 7121 0102
Email: infouk@cognizant.com

India Operations Headquarters
#5/535, Old Mahabalipuram Road
Okkiyam Pettai, Thoraipakkam
Chennai, 600 096 India
Phone: +91 (0) 44 4209 6000
Fax: +91 (0) 44 4209 6060
Email: inquiryindia@cognizant.com