Leveraging Automated Data Validation to Reduce Software Development Timelines and Enhance Test Coverage

By industrializing data validation, QA organizations can accelerate time-to-market and improve the quality and quantity of data analyzed, thus boosting credibility with end users and advancing business transformation initiatives.

Executive Summary

Quality teams across enterprises expend significant effort comparing and validating data across multitier databases, legacy systems, data warehouses, etc. This activity is error-prone, cumbersome and often results in defect leakage due to heterogeneous, complex and voluminous data. It also requires testers with advanced database skills.

Whenever time-consuming manual data validation impacts project schedules, testing efforts are squeezed to get the project back on schedule. Where each and every bit of data is of paramount importance, this is not an acceptable solution since it puts the business at significant risk. Automating data validation is an optimal solution to address these challenges.

In our view, this requires quality assurance (QA) organizations to industrialize the data validation process. Therefore, we have created an enterprise solution that streamlines the data validation process and assists in the identification of errors that typically emerge across the IT landscape.

Our enterprise data testing framework overcomes the impediments of traditional test automation solutions and provides the ability to test end-to-end data. This solution is intuitive and customizable based on data testing needs.

This white paper illustrates how our new dataTestPro solution can be applied across industries to enhance the data validation process. It also elucidates how this solution significantly reduced manual efforts, and improved QA effectiveness, when deployed at a leading insurance organization.

Challenges in Data Validation: An Industry Perspective

Homogeneity in test data is a thing of the past. In fact, heterogeneity of data is now the norm across all industries. A decade ago, a data pool of 10 million records was considered to be large. Today, the volume of data stored by enterprises is often in the range of petabyte or even exabyte. The reasons:

- An increase in mergers and acquisitions, which results in tremendous data redundancy and vast pools of data requiring validation of complex extract, transform and load (ETL) logic.
- A greater need for data center migrations.
- An increased management focus on data and data-driven decision-making related to business intelligence (BI) initiatives.

While there is an abundance of heterogeneous data, there is also a high probability of inherent errors associated with erroneous data, excessive duplication, missing values or conflicting definitions.

Industry research suggests that only three out of 10 organizations view their data to be reliable.

These errors lead to cost overruns, missed deadlines and, most important, a negative impact on the credibility of the data provider. Industry research suggests that only three out of 10 organizations view their data to be reliable.[1] Data-related problems result in an average loss of roughly $5 million annually; in fact, it is estimated that about 20% of these companies experience losses in excess of $20 million annually.[2]

However disconcerting these losses are, the intriguing and often unanswered question is: "How are these losses accounted for?"

Having an appropriate answer to this question would improve business operations and aid strategic decision-making. For data warehousing and data migration initiatives, data validation plays a vital role in ensuring overall operational effectiveness.

Challenges in Data Validation

Data testing is substantially different from conventional testing. Quality teams across organizations expend a great deal of effort in making comparisons on huge volumes of data. This entails:

- **Identifying data quality issues:** Most organizations validate far less than 10% of their data through the use of sampling. This means at least 90% of their data is untested. Since bad data likely exists in all databases, improving testing coverage is vital.
- **Heterogeneous data:** Source data is usually extracted from various sources, including Excel spreadsheets and CSV and XML files, as well as flat files and columns and rows from multiple database vendors' software. After extraction, the transformations of the ETL process need to be verified. This can be scripted, but the effort is often time-consuming and cumbersome.
- **Increased testing timelines:** Test cycles take longer time to complete, if performed manually, particularly when testing large pools of data.
- **Test scheduling:** Testing teams need to execute data comparison tests after ETL processing, where the tests are executed at a specified time. For this process, manual execution is induced by a trigger. This affects the scheduling process.

Quality engineers working on data warehousing projects are continuously scouting for ways to automate the ETL testing process, but with limited success. The reason: The difficulty in standardizing processes across technologies, complex architectures and multi-layered designs. In addition, automation tools are expensive and require an up-front investment and extensive learning curve, which prolongs time-to-value.

An Integrated Approach to Data Validation

To address the aforementioned demands of data validation, we have created a comprehensive solution that can be used in a variety of data testing areas such as data warehousing, data migration testing and database testing. Our solution, dataTestPro, facilitates the QA process in enterprise information environments and significantly reduces the manual effort, delivering reduced costs for various "data testing" needs by provisioning for and managing automated data validations.

As an industry-agnostic approach, dataTestPro has been successfully implemented at large engagements in the banking, brokerage, insurance and retail industries, and is delivering immense business benefits to our clients (see sidebar, next page).

Eliminating Data Quality Issues

Our experience with clients shows that organizations can eliminate data quality issues by performing the following actions:

- Streamlining and accelerating data validation.
- Providing an intuitive, comprehensive and integrated workbench.
- Ensuring a faster and high-quality test execution cycle.
- Preventing data anomalies by early detection of defects in the testing lifecycle.
- Facilitating extensive reuse of test components, reducing time-to-market and simplifying the test management process.
- Increasing test coverage.

Test cases for various data comparison and validation scenarios can be created using a data mapping exercise.

dataTestPro validates complex data transformations, providing full coverage of data and thereby doing the following:

- Performing 95% of the data validation process, with enhanced coverage and reduced risk.
- Providing insightful reports highlighting detailed data differences, down to the individual character level.
- Automating the entire testing process, from scheduling to execution to reporting.

Comparing Heterogeneous Data

dataTestPro compares data between different files and databases after data migration or reconciliation. Source data is typically pulled from various sources such as Excel, CSV, XML, flat file types (any type of delimited file or fixed-width files) and different databases, such as Oracle, Microsoft and any other JDBC-compliant database.

Increasing Testing Speed

Regression testing of data validation is automated by our solution, thereby solving the problem of "the need for speed." Based on our experience at multiple engagements, data comparisons and reporting is 80 times faster than a manual process.

Test Scheduling

Comprehensive automation – from scheduling tests through execution, data comparison and reporting – is provided in our solution, thereby increasing the speed of testing, as well as reducing the cycle time and workload of the tester. The solution provides the tester with the option of scheduling tests to run at a specified time.

Execution Steps

Our solution automates data validation as follows:

- **Test case creation:** Test cases for various data comparison and validation scenarios can be created using a data mapping exercise. This solution also provides the user an option to create test suites and execute multiple test cases in a single framework execution.
- **Execute and report:** Upon completion of the test cycle, informative summary and detailed reports are generated. The user interface is simple; for example, QA analysts from a non-technical background can configure tests to operate in different modes for different types of comparisons.

Quick Take

dataTestPro in Action

Our solution was implemented at a U.S.-based leading insurance and annuities provider. The client faced data validation challenges due to huge data volumes and insufficient test coverage. A typical data transaction involved 1.5 GB of transactional data and over 1,200 ETL transactions. The client was unable to implement standard tools available in the market due to the amount of data originating from heterogeneous data sources.

We implemented dataTestPro in two different projects. With this solution, we were able to reduce data validation efforts by 75% and improve test coverage by 80%.

Overall, we reduced time-to-market from 18 person months to five person months. As a result of all these benefits conferred, the client was able to reduce time-to-market, enhance test coverage and improve productivity with significant reductions in cost and maintenance efforts.

Data Validation Solution Layers

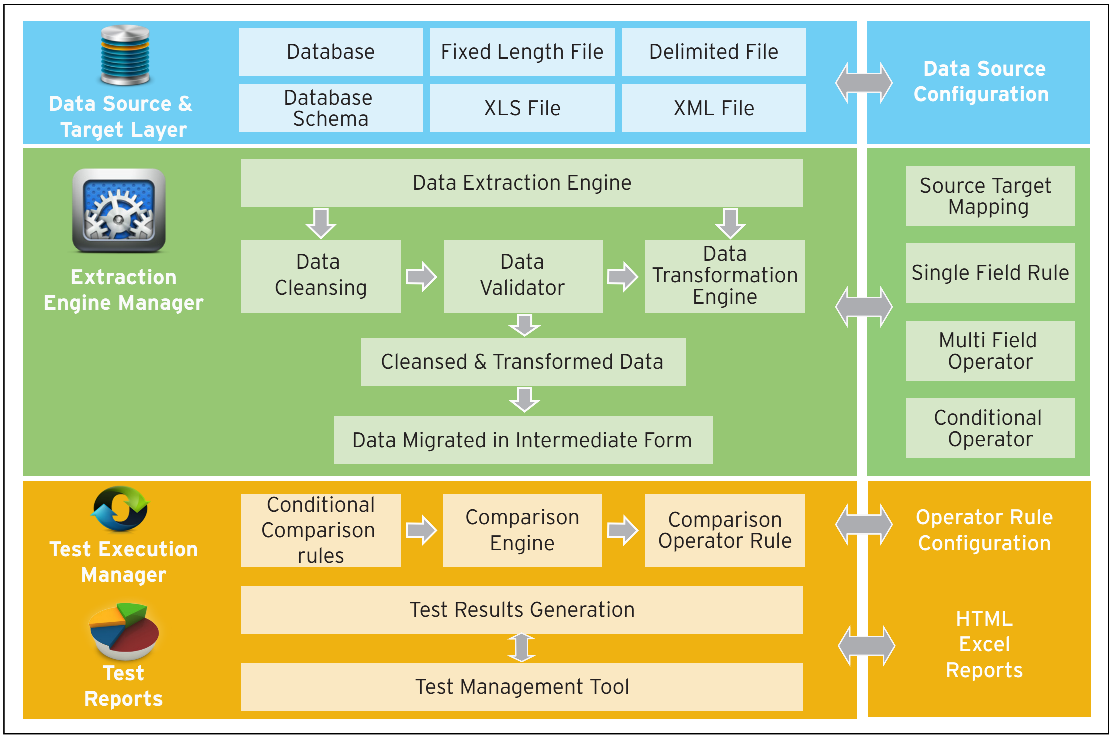

Figure 1

Figure 1 depicts our solution's data validation automation solution layers. Figure 2 highlights the features and potential benefits of our solution.

The Bottom Line

Data loaded into an organization's data warehouse/databases comes from multiple sources and dimensions. The analysis extracted from this data supports an organization's strategic planning, decision-making and performance monitoring systems. Hence, the expense of defects not caught early in the QA process is compounded by the additional business cost of using incorrect data to make mission-critical business or strategic planning decisions. As a result, QA organizations need to test enterprise data from multiple standpoints: data completeness, data transformation, quality, scalability, integration and user acceptance or confidence. These tests are essential for the successful implementation or enhancement of an organization's data.

The return on the investment comes from the ability to measure and track data integrity over time with relatively low per-cycle overhead.

Our automated data validation solution enables enterprises to not only improve operational efficiency, but also enhance the quality of data,

Solution Features and Value Additions

Core Features

- Supports all JDBC-compliant databases and file types.
- Schema comparisons.
- Supports huge volume of data.
- Supports command line execution of test cases.
- Test case scheduler.
- Summary & detailed reporting.

Value Additions

- Reuse of test cases and SQL scripts for regression.
- Export/import test cases across installations.
- Capture the status of test cases that are executed through command line, scheduler and through tool.
- User-friendly reporting in formats like HTML, Excel etc.

Figure 2

which aids in decision-making. The successful implementation of this requires an organization-level commitment to enhance the quality of data for decision-making. While this may involve an initial investment, in the long run the cost per cycle for executing data integrity tests is significantly lowered, as was experienced by leading organizations across industries. The return on the investment comes from the ability to measure and track data integrity over time with relatively low per-cycle overhead. This boosts the credibility of the QA organization due to the business's new-found ability to use measured data for decision-making.

As data validation is not a "one-time" or ad hoc process, our dataTestPro solution makes it a thoughtful, well-planned and continuous validation process supported by organizational commitment to ensuring error-free data.

Footnotes

[1] "Delivering Trusted Information for Big Data and Data Warehousing: A Foundation for More Effective Decision-Making," Ventana Research, 2012.

[2] "The Business Case for Data Quality: A White Paper by Bloor Research," by Philip Howard, March 2012.

About the Authors

Balasubramaniam AV is a key member of Cognizant's Data Testing Center of Excellence R&D team and has 11-plus years of experience architecting solutions across the data warehousing, business intelligence and data migration project spectrum for industry leaders in the finance, hospitality and logistics sectors. Balasubramaniam has provided consulting solutions for numerous complex quality engineering and assurance projects. He can be reached at Balasubramaniam.AV@cognizant.com.

Karthick Siva Subramanian is a key member of Cognizant's Data Testing Center of Excellence R&D team, with operational experience spanning banking, insurance and retail sector engagements. Karthick has worked on numerous engagements focused on data test automation innovation, tools and frameworks. He can be reached at Karthick.SivaSubramanian@cognizant.com.

Balaji Suresh is a key member of Cognizant's Data Testing Center of Excellence R&D team, with progressive experience in various facets of software testing and quality assurance in the telecommunications, retail and banking industries. Balaji is responsible for quality assurance of Cognizant's data testing framework. He can be reached at Balaji.Suresh@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) 207 297 7600
Fax: +44 (0) 207 121 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