

# Talend Data Quality

The first open source data quality solution, combining data profiling and data cleansing and fully integrated with data integration

Data quality entails more than helping companies get correct data into their information systems; it also means getting rid of bad, corrupted, or duplicate data. Clean data is key when integrating information across systems, because misinformation can proliferate quickly—internally of course, but also to business partners. With today's interconnected information systems, bad data spreads the same way viruses are spread by travelers: erroneous information can spread quickly to other applications. The cost of compromised data is incalculable, including lost sales, wasted productivity, loss of reputation or goodwill, and missed opportunities.

Talend Data Quality is a graphical data quality management environment that lets users drag and drop data processing components onto a process map. These components process data, such as addresses, phone numbers, synonyms, abbreviations, and spelling against millions of other records including reference databases. Once the design is complete, Talend Data Quality generates an executable code in Java or Perl that can be deployed easily and executed close to data sources.

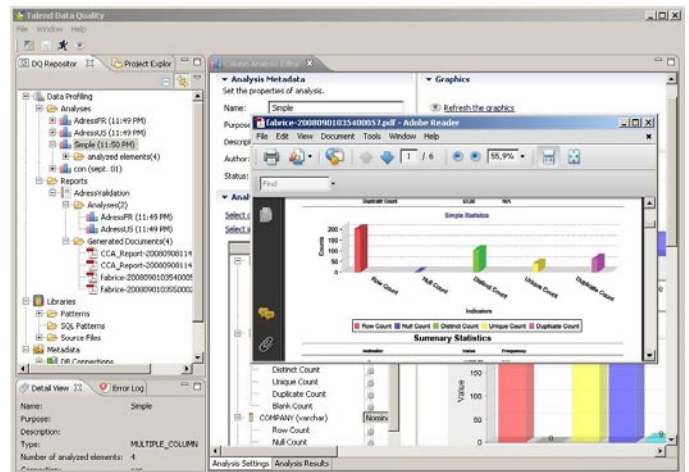
Talend Data Quality is the first open source data quality solution with enterprise-grade features and technical support. It resolves the challenge of data quality via two interconnected modules:

- **Data Profiling** identifies the problem. It provides snapshots of a company's data quality and measures the evolution of data quality over time.
- **Data Cleansing** corrects, or "cleanses" incomplete or inconsistent data by cross-checking against other databases and reference data. It also enriches data by providing value-add information that actually improves the quality and usefulness of existing data.

All functionality is completely integrated with Talend Integration Suite, Talend's leading open source enterprise data integration solution, ensuring that data quality is built into the integration processes during the design phase.

## Data Profiling

The first step in improving the quality of an enterprise's data is to "profile" or evaluate that data. Sophisticated, yet easy to use, the data profiler is an advanced UI-based system that does not require an understanding of database engines and file structures. Business analysts or other non-technical personnel can define a set of indicators for each data element that needs to be analyzed or monitored. These indicators can range from simple or advanced statistics, to text strings analysis, including summary data and statistical distributions of records. By reviewing the metrics on a regular basis, and following their evolution and trend, a company can follow the evolution (improvement or degradation) of the quality of its data.



## Data Cleansing

Once the problem areas are identified, the data must be corrected. All data goes through a "data quality firewall" and records with missing values; values that are improperly formatted or do not match other values in the record in other data sources; duplicates; duplicates with synonyms; even simple typos—are all brought into alignment. This is done by cross-checking against other databases and reference data.



In addition, Talend Data Quality leverages the same open API as Talend Integration Suite. Most notably, this API allows Community members (users and integrators alike) to develop their own data quality components, for example for connecting to industry-specific reference data sources or for incorporating custom-designed data cleansing routines. All the custom components can be shared and leveraged among the Community through the Talend Ecosystem.

Talend Data Quality is available as a stand-alone product or as an added feature to Talend Integration Suite.

## Data Enrichment

This subset of Data Cleansing provides value-add information to the data. The variety of this information is almost limitless—it can include incorporating a company’s Dun & Bradstreet information or a consumer’s credit score, getting the longitude and latitude of an address to help plan delivery routes, or collecting census data to target demographics or income categories.

## Metadata-Driven Design and Execution

Talend Data Quality is a metadata-driven solution, in which all metadata is stored and managed in the centralized Talend Metadata Repository, shared not only by all the modules of Talend Data Quality, but also of Talend Integration Suite, ensuring the consistency of all data quality and data integration processes. Properties defined in the Metadata Repository are inherited by the various processes that make use of these systems. Beyond source and target systems metadata, the Metadata Repository also stores business models, data quality and data integration jobs, and the results of their execution—making it the unique repository of information on all data processes.

## Graphical Interface and Process Modeling

Because Talend Data Quality provides both a graphical and a functional view of the actual integration process it offers an accessible, non-technical view of the data. Systems, connections, steps, and requirements are all designed using standardized workflow notation through an intuitive graphical toolbox. Line-of-business stakeholders can get involved in the design of the data processes and keep in touch with the development progress.

## Versatile and Extensible Processing

Talend Data Quality provides native technical and business connectors to all IT environments. This wide array of connectors allows diverse and heterogeneous data structures at unmatched performance rates and is continually expanding.

*“Incorrect data leads to incorrect decisions. That’s why high quality data is crucial to any business. Proper management of quality is even more critical in the context of application interoperability.”*  
**Mark Madsen, President, Third Nature**

Talend’s technology and business vision shatters the traditional proprietary model by providing the flexibility required to meet the data integration and quality needs of all organizations, regardless of their size, level of expertise, or budgetary constraints.



More Information:  
[www.talend.com](http://www.talend.com)  
[info@talend.com](mailto:info@talend.com)