



interactive analyzer

understanding data

Proactive Data Quality Management

Based on advanced data mining technology, Interactive Analyzer finds even highly complex error connections in your data, completely autonomously and without manually defining any rules, filters or criteria. The software proposes corrections and helps you to quickly resolve inconsistencies. Ease of use, a wide range of integration options and the flexible licence model make the software attractive to companies of all sizes.

Business Case

Ensuring the quality and consistency of business-related data collections on an ongoing basis is a fundamental challenge for managers and staff in IT, quality assurance, logistics, customer management and many other areas. Experts put the overall cost of erroneous data in the USA at more than \$600 billion per year (Eckerson, 2002).

Data quality problems arise due to incomplete, inaccurate or inconsistent data. So a software-based solution to data quality monitoring ought to include some or all of the following functions:

- **Data profiling** – examine the data in order to identify problems and challenges.
- **Daten standardisation** – a [rules engine](#) for entering and processing data that ensures that the data complies with defined quality rules.
- **Matching or linking** – systematic data reconciliations so that similar but slightly different records can be merged (duplicate elimination).
- **Monitoring** – systematic, regular data quality checks; reports and corrective measures.

[This text is based on the Wikipedia article http://en.wikipedia.org/wiki/Data_quality]

Monitoring data quality with Interactive Analyzer

Traditional methods of identifying and resolving data quality problems examine individual data properties for missing values or values that fall outside the usual value range. Manually defined filters and business rules are also used. The Interactive Analyzer supports all these methods, too. But its unique selling point is that it also provides data mining-based methods. These find subtle inconsistencies between the values of different properties in a data set completely independently, without manual filters and business rules needing to be defined. The software then finds the best proposals, based on statistics, for each probable inconsistency that is identified. The benefits of Interactive Analyzer's approach are:

- **Automatically searches for errors without pre-setting error criteria:** The software independently finds the inconsistency criteria, it finds **all the statistically significant criteria** in the database and it finds **only the statistically significant criteria**. This 'hypothesis-free' approach not only saves a great deal of work, but it also ensures that no important criterion is overlooked when defining the filter criteria to be used.
- **Automatically identifies top problems:** All the probable inconsistencies found are rated using objective statistical significance levels and quantified so that sorted lists of 'top problems' can be drawn up. This means that the reason for the classification as an inconsistency and the corrective measure can be reproduced, audited, traced back and logged.

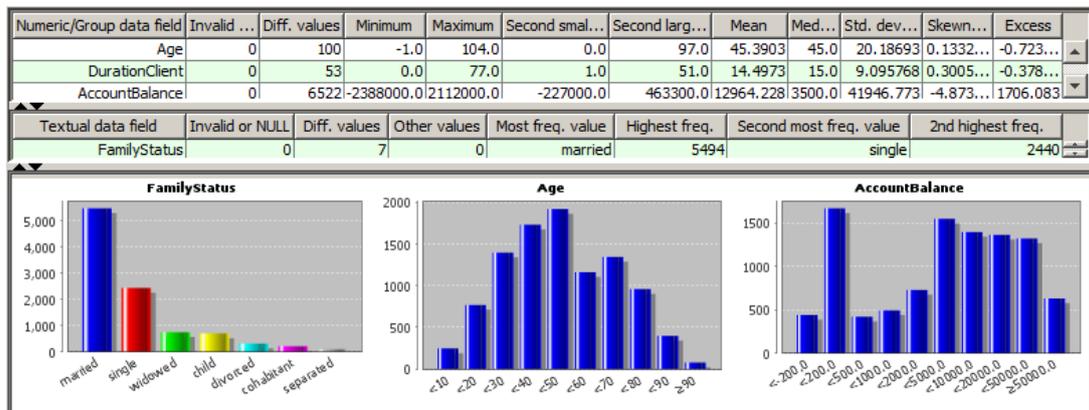
- **Identifies highly-complex data connections:** The data mining methods used can also identify and use patterns and connections that traditional methods are unable to detect due to their complexity or the size or complexity of the data.
- **Performance and scalability:** due to its compressed in-memory data handling technology and highly scalable parallel algorithms an inconsistency analysis with Interactive Analyzer takes only seconds or minutes even on very large data.
- **Integration into existing IT infrastructures and processes:** Interactive Analyzer interacts with databases, reporting systems and other enterprise applications via standardized interfaces such as JDBC, web services and XML. You can define automated tasks and deploy them.

All in all, Interactive Analyzer's data mining approach complements existing methods of data quality assurance very well – it fills the gaps that other methods have in terms of objectivity, completeness, speed of analysis, traceability, documentation and automation.

Example application: Detecting data errors and inconsistencies in customer master data

In the following sample application, a customer master data table is analyzed for data quality, data consistency and possible data errors which might point to certain problems in the data collection and data maintenance process. The table contains demographic attributes such as age, gender, family status, or profession, and aggregated information on the used banking services, on account balances and accounting activities.

First step (traditional approaches): Look at the different data attributes one by one. The data seem to be of a good quality: few missing values, no duplicates, reasonable value distributions.



Second step: hypothesis-free inconsistency analysis: within fractions of a second, Interactive Analyzer finds several dozens of presumably erroneous or inconsistent data records. All the applied expert knowledge on typical and untypical behavior of different customer groups has been learned 'on the fly' from the data.

