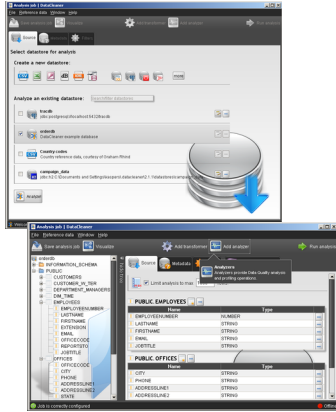


DataCleaner

Powerful environment for Data Quality Analysis (DQA)



Incomplete and incorrect data lead to losses in returns and generate hidden costs that cannot be accounted for. Measuring equals knowledge, but how do you measure the quality of extensive amounts of data? DataCleaner enables your organisation to inspect the data properties of different data sources quickly and simply, ranging from flat files, to Excel spreadsheets, to databases such as Oracle, SQL Server, and DB2. DataCleaner combines ease of use with a broad range of analysis functions. Building an analysis job that consists of data checks on hundreds of data entities is extremely quick and the results are presented in a convenient report. DataCleaner can even perform transformations and filtering on the selected data in order to partition, subset and combine parts of the data in the job. This makes DataCleaner an all-in-one Data Quality Analysis (DQA) tool: It combines light-weight data transformation capabilities with powerful analysis capabilities as well as with data source handling. Examples of available analysis functionality:

Functionalities offered by DataCleaner

With DataCleaner you can:

- Discover the patterns of your data
- Determine the distribution of your data values
- Assess the quality of your data
- Validate values within their range
- Build lists and patterns for matching
- Check your data for synonyms

Pattern finder

Helps you discover the patterns of your data. Whenever you think you are dealing with a well formatted set of data, it often appears that this is not the case. The Pattern finder will then quickly show you which values need attention. Examples are formatted names, credit card numbers, registration keys, address lines and many more. It is of high importance for the maintainability, usability and correctness of your data that any structured field is indeed conforming to the structure. You can even use your identified patterns as input for your next analysis, to make sure that conformity remains.

Value distribution

What is the distribution of your data values? This is the simple, but often surprising, answer that the Value distribution analysis gives you. Do you expect anything else than a limited set of values in eg. your 'gender', 'product type' or 'country'? Well, the value distribution will quickly tell you if that is indeed the case.

String length, case and character checks

Often when you're looking at a new data source, you don't really know what is in it. DataCleaner provides several analysis options that are very explorative and give you a lot of feedback about eg. string length, upper/lower casing, usage of special characters and more. Such metrics help gain insight in the quality of the data you are dealing with.

Value range and missing value checks

Validate whether the values of a specific category are within their valid range. Although this may seem obvious, such a check will often show to be crucial to uphold a good quality of your data and to enable management of your data over time.

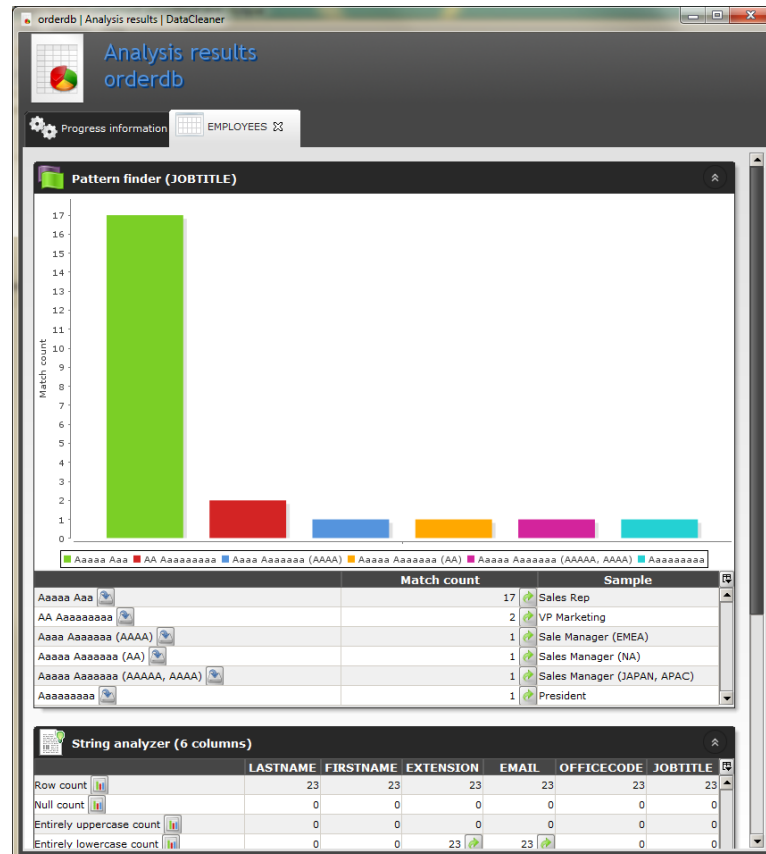
Dictionary and string pattern match checks

Build white-lists, black-lists, patterns (either identified in the Pattern Finder or patterns that you build yourself) and match against these reference data. You can either perform the matching as validation or as exploration. Perhaps you are interested in categorizing the gender of your contacts by use of matching with female and male name lists? Such analyses are easily carried out with the reference data capabilities of DataCleaner.

DataCleaner is highly extensible, embeddable and compliant with new datastores:

Extensions:

- Extension Swap: share extensions to DataCleaner and install by simply clicking a button in the browser
- API: create your own transformers, analyzers and filters
- HIquality Contacts: Name, Phone and E-mail cleansing based on Human Inference natural language processing DQ web services



DataCleaner: open source data quality solution

DataCleaner supports a very wide range of data sources, including:

- Microsoft Excel spreadsheets (both .xls and .xlsx format)
- Microsoft Access database files (both .mdb and .accdb format)
- Comma-Separated Values (CSV) files (as well as values separated by tab, semicolon and more).
- Fixed width value files
- dBase database files
- OpenOffice.org database files
- XML files

Databases, eg.

- Oracle
- MySQL
- Microsoft SQL Server
- Postgresql
- IBM DB2
- ... and more!

Synonym replacement checks

Are all our representations of the same customer actually the same? In other words: Does your data contain both 'Coca-Cola' and 'Coca cola'? Most customer databases do, and such duplications cause great pain when working with eg. customer service and support! With synonym catalogs DataCleaner provides a type of reference data where you define which terms are synonyms and which are master terms. You can then use this catalog to quickly replace (or just analyze) dirty values with correct ones.

Additional features

In addition to these features, DataCleaner employs an extensible architecture where the advanced user is able to add new functionality, either by using the built-in scriptable components or as complete plug-in packages. A great example of this is the Human Inference way of applying DataCleaner for our knowledge gathering process, where we're using a set of custom extensions for DataCleaner to filter, validate and categorize items using our advanced matching and cleansing heuristics.

We call DataCleaner what it is – the premier open source data quality solution!

Recent developments : extensions, cleansing and ETL

DataCleaner is rapidly developing into a full fledged DQ tool with not just profiling, but also high quality cleansing tools. DataCleaner updated the ExtensionSwap in June 2011 and is now highly extensible, embeddable and compliant with new datastores. Try the HIquality Contacts extension: advanced Name, Phone and E-mail cleansing based on the clever natural language processing DQ webservices provided by Human Inference. Since DataCleaner both has the power to inspect your data sources, but also to transform and write new data sets, you can even apply DataCleaner as a lightweight ETL tool, perfect for data quality oriented projects or preparing for an ETL process. For Human Inference customers DataCleaner supports a fully compatible Human Inference interfacing of the HIquality Data Improver.