G00223234

Who's Who in Open-Source Data Quality (2012 Update)

Published: 18 January 2012

Analyst(s): Andreas Bitterer

The open-source movement has reached the data quality tools market, with only a handful of vendors and projects offering solutions. Organizations with a need for broad data quality capabilities such as cleansing, matching, deduplication, or enrichment, should not expect extensive functionality but should evaluate open-source options for profiling or standardization. For critical enterprisewide data quality requirements, stick with commercial offerings.

Key Findings

- Data quality is the latest data management software area targeted by open-source vendors and projects, alongside database management systems (DBMSs), content management systems, or data integration tools.
- Open-source data quality vendors do not yet play a significant role in the overall $800 million data quality market.
- Open-source data quality vendors are mostly focusing on the customer data domain, providing name and address standardization and cleansing.
- With few exceptions, most deployments of open-source data quality tools are in support of relatively small projects and very few implementations actually went into production.

Recommendations

- Use open-source data quality tools for educational or initial assessment purposes and to assist in developing requirements for data transformation and data migration projects. However, understand the limitations of community versions relative to features beyond profiling, standardization and other basic data quality functions if considering them for production deployment.
- Create test-beds or test cases in which open-source data quality tools are used in proof-of-concept or pilot programs. Do this especially if other open-source tools for data integration, the

database management system, or business intelligence (BI) are under consideration, as this creates a consistent cost model and work approach.

- Considering the market convergence of data quality tools with data integration platforms, look for data quality solutions that are easily integrated into data integration data flows.
- Since a data quality problem cannot be solved by technology alone, the cost-saving promise of open source is only one aspect. Organizations must still invest in proper data quality processes, data stewardship and other nontechnical areas. In addition, while software licensing costs may be lower for open source, functional limitations require you to augment the tool, which raises costs.

Table of Contents

List of Tables

List of Figures

Strategic Planning Assumption(s)

In the first iteration of this document (published in 2009) we included the following Strategic Planning Assumption:

"All open-source data quality projects combined will reach just 3% to 5% market penetration (subscribed customers) up to 2012. It will likely be well beyond 2012 before open-source data quality tools have broadly caught up in terms of their capabilities with the commercial data quality tools vendors."

This prediction from 2009 turned out to be valid, as far as overall adoption rates are concerned. After the initial spike in open-source data quality, the large number of offerings had died down, adoption progressed only slowly, and no new projects had started. In the future, this submarket will likely find it even harder to stand on its own. The commercial market is already very fragmented, lower-cost cloud offerings are entering the space, and the convergence of data integration and data quality makes it increasingly hard for an open-source data quality project to survive in the long term.

Analysis

The information management technology markets have seen a large number of vendors approach the space with an open-source strategy. Open-source products have been made available from database management systems (such as MySQL and Ingres), BI tools (such as Jaspersoft and Pentaho) and data integration tools (such as Talend and Jitterbit, see "2009 Sees Increased Adoption of Open-Source Data Integration Tools"), to content management (such as Alfresco and concrete5) and document management (such as OpenKM and Epiware). The latest area to see the first open-source offerings is the data quality software market. While significantly smaller than the database management system or BI markets, it is still estimated to be worth around $800 million. This represents a large enough opportunity for open-source vendors to have a go at entering the market, even though it is dominated by large infrastructure vendors such as IBM, SAP BusinessObjects, Informatica and SAS/DataFlux.

Market Definition

As outlined in "Magic Quadrant for Data Quality Tools" the vendors participating in this market offer stand-alone software products that address the core functional requirements of the data quality discipline (see Table 1).

Table 1. Data Quality Functional Requirements

Technology	Description
Profiling	The analysis of data to capture statistics (metadata) that provide insight into the quality of the data and help identify data quality issues.
Parsing	The decomposition of text fields into component parts.
Standardization	The formatting of attribute values into consistent layouts based on industry and local standards (for example, postal authority standards for address data), user-defined business rules and knowledge bases of values and patterns.
Cleansing	The modification of data values to meet domain restrictions, integrity constraints or other business rules that define when the quality of data is sufficient for the organization.
Matching	Identifying, linking or merging related entries within or across datasets.
Monitoring	Deploying controls to ensure data continues to conform to business rules that define data quality for the organization.
Enrichment	Enhancing the value of internally held data by appending related attributes from external sources (for example, consumer demographic attributes or geographic descriptors).

Source: Gartner (January 2012)

In addition to the above functionality, vendors participating in the data quality tools market often provide connectivity to a range of different data structure types and adapters to third-party technology and data providers (for example, for address validation, or telephone number or bank code verification). Typical data providers are postal organizations in various countries, telephone operators or banking networks. Data quality tools also often connect to credit bureaus, blacklist or watchlist providers, or vertical industry data sources (for example, for manufacturing or healthcare).

Other data quality tools' capabilities include: standardization for specific data subject areas; international support; metadata management; a configuration environment for managing and deploying data quality rules; data quality workflow support for various data quality roles such as data stewards; and support for service-oriented architecture deployments. For more information on how to select suitable data quality tools, refer to "Toolkit: RFP Template for Data Quality Tools."

The State of Open-Source Data Quality

As in the BI and data integration markets, open-source data quality vendors will slowly start to appear and then play catch-up with the established commercial vendors. At the time of writing, there have been only a handful of attempts on the market, with varying degrees of potential and success.

Initial open-source projects have centered on data profiling. This makes sense as newcomers to data quality can start doing initial assessments of their data without investing in fully-fledged data

quality platforms. Since data profiling is also recommended during early phases (even of broad data quality initiatives), even organizations fully committed to a data quality program could distribute open-source data profiling software to multiple users in various departments. This enables them to get a broad overview of potential data quality issues, again, without a large upfront investment. Later on, those companies would typically look at broad data quality capabilities (see Table 1) and use their existing relationships with infrastructure vendors, or pick a smaller vendor that is more local. The data profiling tools can also be used in projects other than data quality. For example, to understand the structure and the content of data sources in advance of building extraction, transformation and loading (ETL) routines for a data warehouse, or within a data migration project. Another use case for open-source data quality is application development. Data profiling tools can be used to inspect data sources that the developer needs to connect to, verify database records after write operations, or validate in-database operations.

Well-known technology providers in the BI and data integration markets have expanded their footprint into the data quality area, often through acquisitions. Open-source vendors in the BI and data integration markets will progress in the same way. During the merger and acquisition frenzy a whole raft of data quality vendors were taken over, including: AddressDoctor, Ascential, Avellino, DataFlux, Datanomic, Evoke, Firstlogic, Fuzzy Informatik, Global Address, Group 1, QAS, Identity Systems, Netrics, Silver Creek Systems, Similarity Systems, Vality and Zoomix. After this the open-source equivalents started to expand their portfolios and move into data quality software. Because the data quality tools market is only about 10% of the size of the BI platform market, software providers are entering it much less aggressively and the vendor landscape remains relatively stable.

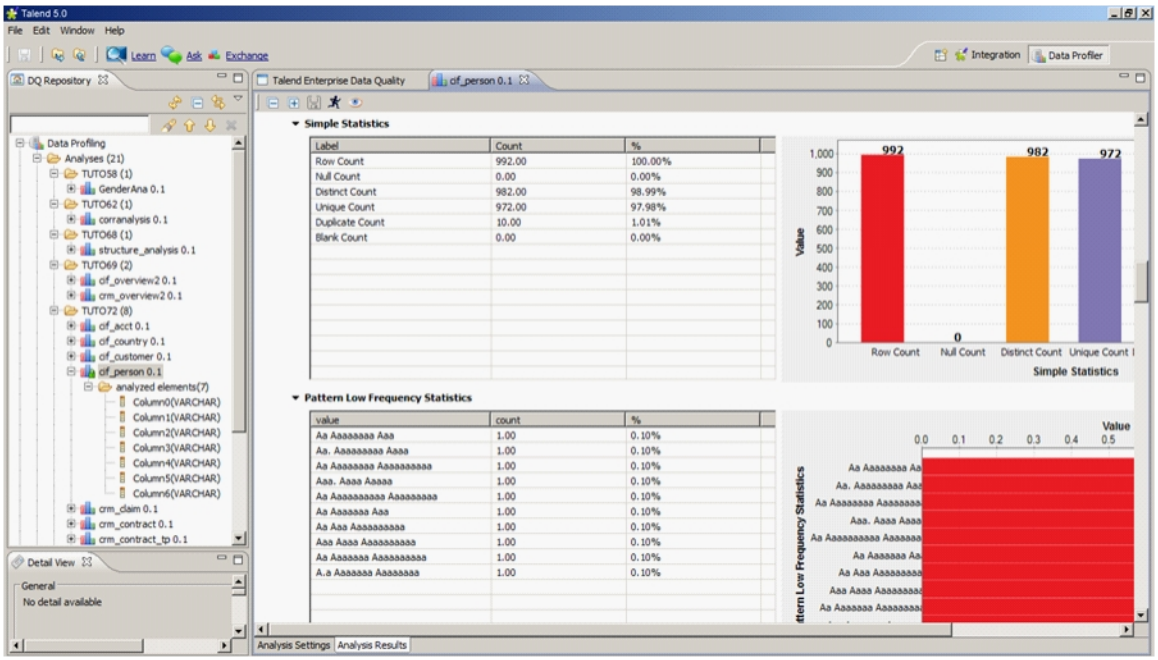
Talend

Suresnes, France and Los Altos, California

www.talend.com

The best known company for offering open-source data quality is Talend. While it was not the first ever available open-source data quality tool, it is now the most advanced of all open-source alternatives. The company provides two types of data quality software: the Talend Open Studio for Data Quality, a freely downloadable and limited functionality version of the fuller featured product; and Talend Enterprise Data Quality, which includes additional capabilities for cleansing, matching and report generation. Talend reports that about 120 of its commercial customers buy the data quality product, still a small portion of its 2,000+ customer base, most of which subscribed to Talend's data integration product. The Unicode-enabled Talend Open Studio for Data Quality (see Figure 1) includes basic functions for data analysis, pattern discovery, SQL business rules, data drill-down and rudimentary results visualization. For an initial data investigation, which may include examining the existence and validity of attribute values, data ratios, uniqueness of keys, or orphaned records, the tool is "good enough." However, users will very quickly hit the tool's limits and require better connectivity to more data sources, and better matching, cleansing and visualization capabilities. That is where the commercial package, Talend Enterprise Data Quality, is positioned.

Figure 1. Sample Screen of Talend Open Studio for Data Quality



Source: Talend

In combination with its main product, Talend Enterprise for Data Integration, users gain connectivity adapters to more data source types, including third-party data providers, such as Dun&Bradstreet, or publicly available datasets, such as census data. Because Talend does not provide any address validation capabilities itself, the company has struck a technology agreement with Experian QAS, a recognized provider of address verification solutions with connectors to Uniserv, AddressDoctor and the Google Geocoding API. Talend built a data stewardship console into its product, addressing the increasing need for nontechnical staff to monitor and manage data quality issues detected during data flows. The data quality dashboard, while a bit basic in its visual capabilities, shows widgets about important metrics, indicating the status of the data profiled, matched, cleansed or deduplicated. With Talend's additional acquisitions, Amalto and Sopera, the vendor is expanding its vision of a "unified integration platform" and following the major convergence trend of data quality, data integration, and master data management technologies.

Human Inference

Arnhem, The Netherlands

www.humaninference.com

In February 2011, Human Inference, a Dutch vendor of commercial data quality tools, announced the acquisition of the Danish eobjects.org project, which provides DataCleaner, an open-source data profiling tool that started in Denmark in 2007. After the acquisition, eobjects.org became a division of Human Inference, focusing on supporting the open-source data quality community.

Human Inference has an established brand in Europe, particularly in the Benelux countries, while eobjects.org has few customers, a tiny community that shows very little activity, no revenue and almost no visibility in the market. Human Inference bought eobjects.org to attract small businesses with a free data profiling tool so that it can sell them its own commercial tools later on when their needs grow. DataCleaner serves as a seeding strategy for departmental assessments of data quality in larger enterprises. Human Inference may try to upsell DataCleaner users to its software-as-a-service offering for data cleansing. However, DataCleaner's Java application already sports the "Powered by Human Inference" logo, and the product is not integrated with Human Inference's own products. Human Inference offers no migration path from DataCleaner to the commercial HIquality set of products. Even if customers decided to upgrade, it would largely be a new implementation.

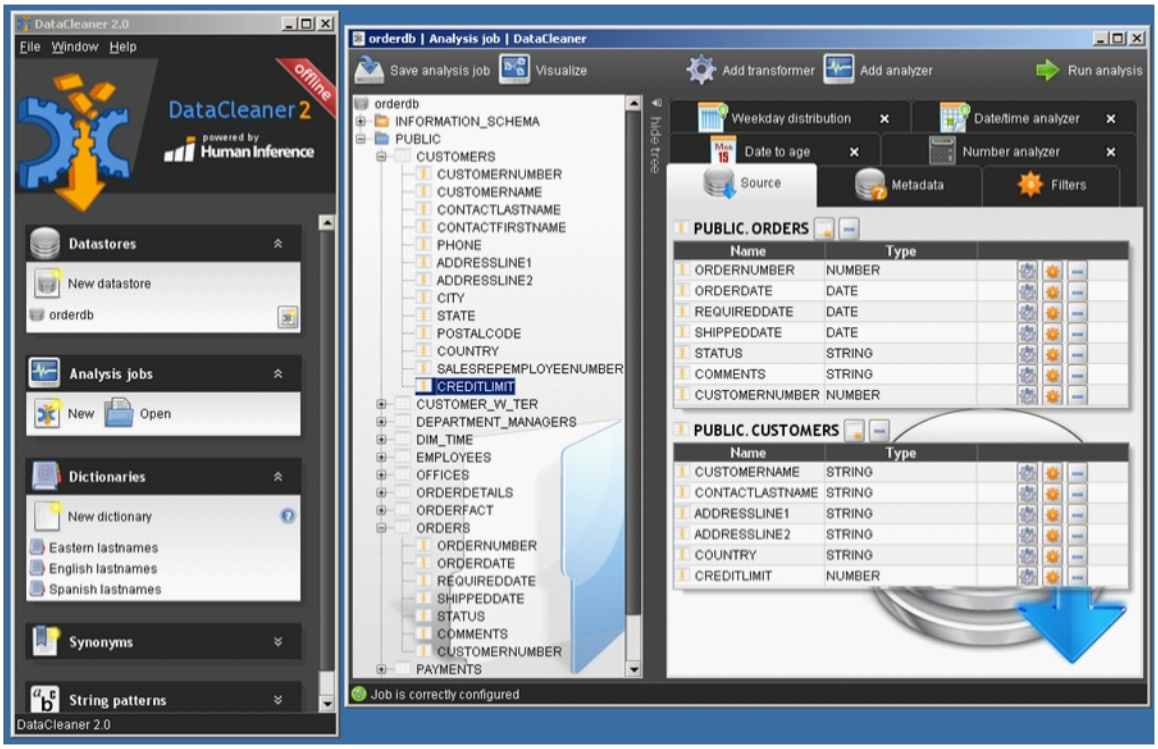
The DataCleaner software consists of a quick download and an easy installation, including some sample data (customers, employees, departments, products, orders, payments) that allows you to try out the profiling functionality.

Out-of-the-box database connectivity includes MySQL, Oracle, PostgreSQL, Microsoft SQL Server and a few minor Java database brands. Additional database drivers can be installed automatically in the user interface (UI). In addition, DataCleaner can read from comma-separated or tab-separated value files, Microsoft Excel spreadsheets, Microsoft Access database files, fixed width files, OpenOffice, XML and regular text files, which is sufficient for profiling projects. Support for more data formats can be created as extensions.

Profiling options include standard measures, string, time and number analysis, pattern detection, value distribution, character set distribution and a variety of matching options, against dictionaries, regular expressions, synonym catalogs and masks. Through enabling multiple database connections and multithreading, performance can be tuned, based on the corresponding database support. DataCleaner provides filtering and transformation components for preprocessing data, along with a few target data writers, making it possible to use the tool as a lightweight ETL engine, for example, for one-time migrations. Profiling results include various counts, minimum and maximum values, and averages. However, the lack of more advanced data quality functionality, such as cleansing, matching or monitoring, makes DataCleaner somewhat limited in its usefulness and requires an upgrade to the full-fledged HIquality suite, Human Inference's commercial product. Still, data architects or application developers who can live with the rather basic UI may find the DataCleaner tool useful for investigating database content, potential inconsistencies or other data-related issues.

With version 2.0, DataCleaner's UI has undergone a facelift, although with mixed results. The tool provides a rather awkward user experience, suited more for the tech-savvy developer than for a typical data steward. Continuously overlapping windows, a distracting high-contrast color scheme and an unfortunate distribution of screen real estate make DataCleaner a cumbersome tool. Figure 2 shows a sample screen shot of the DataCleaner product.

Figure 2. Sample Screen of DataCleaner



Source: Human Inference/DataCleaner

SQL Power

Toronto, Canada

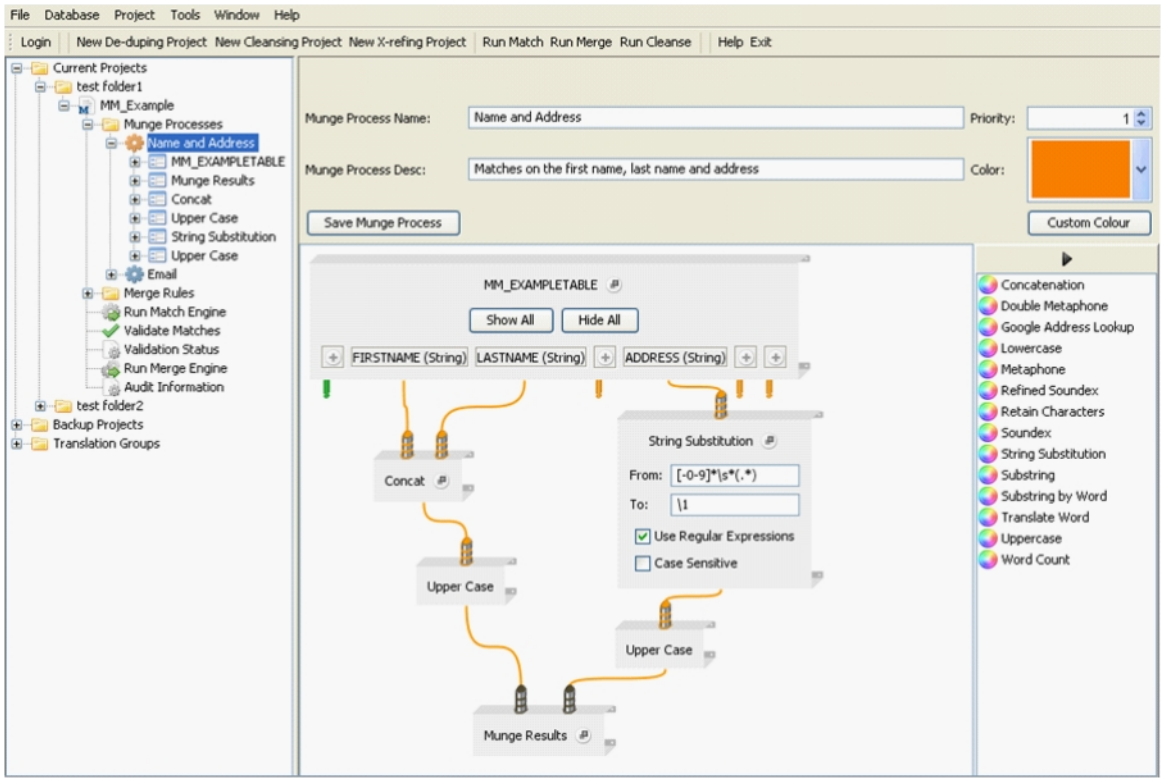
www.sqlpower.ca

Although SQL Power does not position itself as an open-source software company, the vendor still offers various products under a General Public License (GPL) V3 open-source agreement, some of which are freely downloadable, including: the Wabit reporting tool; Power*Architect, a data modeling and profiling tool; and DQguru, positioned as a data cleansing and master data management (MDM) tool. Similar to Human Inference's DataCleaner tool, the profiling segment of Power*Architect provides a variety of counts (rows, nulls, unique values) and calculates minimum, maximum and average values. The tool is obviously targeted at architects, hence the inclusion of ETL functionality (visual mapping and creating jobs, based on Pentaho's data Kettle integration tool), as well as online analytical processing schema management functionality. The profiling capabilities are clearly not targeted at nontechnical end users, such as data stewards, as the tool requires decent knowledge of database operations from Java Database Connectivity to schema modeling. The inclusion of a feature for "universal SQL access" also indicates that this tool is targeting technology-savvy users who thoroughly understand the world of SQL and DBMSs.

SQL Power's second open-source offering pertaining to data quality is DQguru (see Figure 3). It is rather unclear why profiling is bundled with modeling and cleansing is bundled with master data

management (MDM). In fact, calling DQguru an MDM tool is not particularly accurate. As far as cleansing functionality is concerned, the tool provides what it calls "projects" for deduplication, cleansing and address correction. The latter is supported by a DQguru subscription, which provides users with a monthly copy of the Canada Post database. For addresses outside Canada there is currently no validation capability. The only other correction facility comes from a feature named "translate words manager," by which abbreviated words, such as *ave*, *bldg*, *corp* or *dept* can be mapped to the proper spelling of *avenue*, *building*, *corporation* and *department*, respectively. This, of course, would not be considered address correction, but rather standardization. Figure 3 shows a sample screen shot of a DQguru matching process.

Figure 3. Sample Screen of DQguru



Source: SQL Power

Infosolve

Princeton, New Jersey

www.dataqualitysolution.com

Infosolve can barely be considered an open-source vendor as its business model is radically different than most other software vendors carrying the open-source moniker. One could argue that Infosolve is actually a services company that happens to own a number of software products that the company implements for its clients. In contrast to the typical open-source vendors, Infosolve

does not provide any products to download, even for evaluation purposes. In that sense, "open source" is somewhat of a misnomer, although clients receive the implemented software (including the source code) without any licensing charge. Customers can also re-distribute the source code under a GPL 2.0 agreement.

Infosolve's software portfolio, branded under the rather odd name of "zero-based data solutions," includes the OpenDQ and OpenCDI products, along with functionality for data integration, migration, conversion, mining and enhancement. The OpenDQ product provides profiling, standardization, various kinds of matching and merging, deduplication, and Web services-based external data enrichment. Address validation is provided for 240 countries through an agreement with AddressDoctor (now part of Informatica). Infosolve's go-to-market model is based on its consultants building specific data quality solutions for its clients. While prospects may get access to the software for a proof-of-concept implementation ahead of a full-blown and chargeable deployment, organizations that are looking for a simple download and installation, or have no need for external consultants doing implementations, will likely bypass Infosolve as a potential data quality tools provider.

Other Projects

Similar to other open-source software domains, a few defunct data quality projects remain on the Internet. Organizations should steer clear of open-source data quality zombies such as dwSavvy (www.dwsavvy.com/dwsavvy_data_profiling), ChkDb (www.agt.net/public/bmarshal/chkdb), Berkeley University's Potter's Wheel (control.cs.berkeley.edu/abc) and Arrah (sourceforge.net/projects/dataquality), as there seems to be no more development work and support. Similarly, while the Java community named Mural (mural.dev.java.net) still seems to have some life in it, its data quality subproject, Open-DM-DQ (open-dm-dq.dev.java.net), appears to have reached the end of its life, as there are no more recent updates available. Then there is the DataNucleus (www.datanucleus.org/products) Access Platform, an ongoing and well-documented open-source project which claims to do data quality and data profiling. However, on closer inspection, the platform is really only enabling persistence of Java objects to a relational database management system, db4o, Lightweight Directory Access Protocol, Excel and other data stores, licensed under an Apache 2 agreement.

Bottom Line

Nearly all open-source data quality tools offerings must be considered largely toolboxes for techies, with Talend being a reasonable exception, as the vendor has even been included in "Magic Quadrant for Data Quality Tools." Still, none of the open-source platforms reach the capabilities of the commercial market leaders in the data quality arena. However, the described open-source offerings may be helpful as a starting point for generic data quality initiatives. Commercial data quality vendors also typically provide best practices about data governance and stewardship, metadata and MDM. However, most open-source data quality vendors miss those aspects entirely. An increased adoption of data quality tools in the market can be observed and this will also help the open-source projects to gain more visibility, but the general worldwide adoption of open-source data quality tools will grow only very slowly. Rather than expecting a slew of new vendors or projects entering the space, it can be assumed that the number of offerings has stabilized.

Recommended Reading

"Magic Quadrant for Data Quality Tools"

"Gartner's Data Quality Maturity Model"

"Toolkit: RFP Template for Data Quality Tools"

"Hype Cycle for Data Management, 2011"

"Hype Cycle for Open-Source Software, 2011"

Acronym Key and Glossary Terms

BI	business intelligence
CDIDBMS	customer data integration database management system
DQ	data quality
ETL	extract, transform, load
GPL	general public license
MDM	master data management
SQL	structured query language

Regional Headquarters

Corporate Headquarters
56 Top Gallant Road
Stamford, CT 06902-7700
USA
+1 203 964 0096

Japan Headquarters
Gartner Japan Ltd.
Atago Green Hills MORI Tower 5F
2-5-1 Atago, Minato-ku
Tokyo 105-6205
JAPAN
+ 81 3 6430 1800

European Headquarters
Tamesis
The Glanty
Egham
Surrey, TW20 9AW
UNITED KINGDOM
+44 1784 431611

Latin America Headquarters
Gartner do Brazil
Av. das Nações Unidas, 12551
9° andar—World Trade Center
04578-903—São Paulo SP
BRAZIL
+55 11 3443 1509

Asia/Pacific Headquarters
Gartner Australasia Pty. Ltd.
Level 9, 141 Walker Street
North Sydney
New South Wales 2060
AUSTRALIA
+61 2 9459 4600