MAXIMIZE THE ROI
OF YOUR DATA WAREHOUSE BY AUTOMATING
DATA QUALITY

BUILDING RELIABLE AND ACCURATE
ANALYTICS CAPABILITIES WITH DATA QUALITY
# TABLE OF CONTENTS

Summary

What is Data Quality (DQ)?

What Does DQ Mean for Your Data Warehouse?

4 Immediate Ways to Deal with DQ Errors in the Data Warehouse

Approaches to Data Quality Management

Data Profiling: Understanding Source Data in the Data Warehouse

Improving Data Quality

Data Quality Tests

Data Integration: Bringing Together Data from Different Sources to Improve Quality

Data Augmentation: Adding Value to Data to Increase its Usefulness

Conclusion: Build Reliable and Accurate Analytics Capabilities with Data Quality
Summary

The big data and analytics movement has achieved such ubiquity that we are led to think that data is everywhere, and all of it is useful in powering business intelligence initiatives to help organizations make better decisions. In truth, data is indeed everywhere, with the total volume expected to exceed 44 trillion GBs by 2020, but it is rarely useful.

Much of the data in the digital universe is not harvested for analysis, but for a business, such analyses are key. Unfortunately, only 27% of organizations with data analysis initiatives in place report any significant success, while a mere 8% describe their efforts as “very successful”. These numbers are low not because businesses don’t have the necessary data – it’s because they don’t have quality data to work with. This means that the data they have, for all intents and purposes, is useless.

That’s where data quality comes in. Availability isn’t enough; you need to make data useful. This eBook will help you make enterprise-wide data more useful and improve ROI of your data warehouse by developing a deeper understanding of:

1. What data quality is
2. The role it plays in your big data initiatives
3. The different approaches to improving data quality
4. Creation of a data quality framework
5. How to bring it all together to improve analytics and decision-making
What is Data Quality (DQ)?

We hear the term ‘data quality’ being thrown around all too often, but what does high quality data mean? In the simplest terms, data that is consistent, accurate, and unambiguous is considered quality data.

The most common cause of data quality issues is when databases merge across disparate source systems as a result of data integration processes. In theory, data integration should be seamless, but incompatibility between data fields because of schema and format inconsistencies across systems is almost always a problem.

Let's consider an example. Say you have an entity named ‘John Smith’ with City field filled as ‘London’ in one database. Another database could have an additional column for ‘Country’, and when the two databases are merged, you won’t have any easy way to fill the City field for the former ‘John Smith’ record. After all, there are about 12 different locations named London in the world. This is just one example – there could be myriad issues like concatenated names in one database and separate columns for FirstName and LastName in another. There could also be separate columns for HouseNumber and StreetName in one database, while another database could have a single Address column.

While data quality issues are most common during data integration, the other major reason for bad data is mistakes when data is being entered into your system. For instance, your database may contain multiple entries of ‘John Smith’ who lives at ‘123 Random Street’. This could be a simple case of double-entry by one of your sales reps.

If you want to reach the right ‘John Smith’ from your datasets, you’d face issues in any of the cases outlined above. That's the cost of bad data quality, and if you want to see the cost in financial terms, IBM reports that poor data quality costs the US an astonishing $3.1 million!

One way to correct data quality is to manually scan and fix errors, but that quickly becomes impractical when you’re dealing with big data. The other route is to use proven data quality methodologies and systems, which we'll discuss further in this eBook.
What Does Data Quality Mean for Your Data Warehouse?

Bad data is why many data warehousing projects fail to deliver results. As discussed in the previous section, the leading reason for bad data is when data across multiple systems is being integrated, but this integration is at the base of any data warehousing project.

The purpose of the data warehouse is to build a unified layer that contains data from all relevant data sources throughout the organization. This means you need to integrate data from multiple systems and optimize it for analysis and business intelligence. So, the data warehouse does not generate any data of its own and any data quality issues are either within source systems or arose as a result of how data is interpreted in different systems. The data warehousing team must take over the responsibility of identifying such issues, coming up with ways to improve data quality, or gain business agreement over certain issues to be considered acceptable. The last one may seem confusing but is critical for maintaining a balance between the cost of improving data quality and what the business is willing to spend.

If data quality is overlooked though, data warehouse users will have inaccurate, incomplete data on their hands. This translates directly into your data not being representative and providing erroneous analytics. These may be the very same analytics the C-suite uses for decision-making, and we all know how badly a single wrong decision can potentially hurt businesses.

4 Immediate Ways to Deal with DQ Errors in the Data Warehouse

Reject the error: You have to decide if you want accuracy or completeness in any given instance. If accuracy is more important, you could reject any error-prone record given that fixing it requires more effort than your business is willing to put in.

Accept the error: If you value completeness over accuracy, you may choose to ignore the error and accept records with errors into your data warehouse if you consider the errors to be tolerable, deciding to fix the error when your team can find the correct values later.

Fix the error: If your team can find the correct values or format change that'd would fix a specific error at a cost you can bear, the choice will be obvious.
Data Profiling: Understanding Source Data in the Data Warehouse

When trying to improve something, understanding it is naturally the first step. The process of understanding existing data with respect to how you want it to be in its final form is called ‘data profiling’. This includes digging deep into source data and understanding the content, structure, and cardinalities. That’s how we identify where data quality processes need to be applied and which approach to choose.

Assign default value: If completeness is highly important yet the correct value cannot be found, you could assign a default value for each type of error to substitute erroneous data.

Regardless of the action you take, it's critical that data warehouse users understand the implications of each action, so they factor it into their analytics.

Approaches to Data Quality Management

We have seen how data quality is a key requirement in data warehousing, but in practical terms, going about fixing quality issues in a data warehouse is a complex process. This section will cover approaches to data quality management in a data warehouse, specifically by:

- Understanding source data in the data warehouse
- Understanding the causes of data quality errors
- Bringing together data from different sources to improve quality
- Adding value to data to increase its usefulness

Data Profiling: Understanding Source Data in the Data Warehouse

When trying to improve something, understanding it is naturally the first step. The process of understanding existing data with respect to how you want it to be in its final form is called ‘data profiling’. This includes digging deep into source data and understanding the content, structure, and cardinalities. That’s how we identify where data quality processes need to be applied and which approach to choose.
Too often, data profiling has been relegated to the backroom of Extract-Transform-Load (ETL) processes, only brought forward to check for minor anomalies in data once data warehouse design is complete and production data needs to be delivered. In reality, data profiling should be the very next step in your data warehousing project once you've gathered business requirements. At this stage, knowing the amount of work that source data would require before becoming usable for analytics would impact the design and time taken to build your data warehouse seamlessly.

When profiling your data during the data warehouse design process, focus on these four deliverables to get maximum value from the effort:

1. **Assign Default Value:** The most basic deliverable is a “no-go” on the entire project. The source data you want to use to build your data warehouse may have too many errors or too much missing information for the data warehousing initiative to be at all viable for analytics. While this may be construed as a huge failure, it's actually an extremely valuable outcome, because now your team can refocus their efforts elsewhere rather than spending weeks and months building a project only to find out that the end result is an enormously flawed reporting system that is unusable for decision-making. Such surprises at the end are often career-shortening for business intelligence team leads.

2. **Listing Down the Existing Issues:** The second deliverable is a list of issues that already exist in source data which must be fixed before the project can move forward. The fixes are a major external dependency and must be managed well to ensure the success of your data warehouse. You might think that issues can be fixed later once data is written into the data warehouse, but then, every time you sync your operational systems and the data warehouse, the issues will come up.

3. **Issues Arising During the Extraction Process:** The third deliverable is a list of data quality issues that are encountered while extracting data from multiple sources and writing to the data warehouse. A deep understanding of such issues will help you come up with data transformation logic and exception handling methods best suited to your business scenario. You will also be able to determine any manual processing that will be required to fix inconsistencies and factor that into the total time it takes to complete the data warehousing project.

4. **Check Business Rules and Key-based Relationships:** Lastly, focus on business rules previously unanticipated, along with issues in foreign and primary key relationships and hierarchical structures. You will need to dig at a deeper level to identify such intricate issues, but if left unchecked, they will permeate data warehouse design and may blow out of proportion later.
Here are a few simple examples of issues that data profiling may help discover:

<table>
<thead>
<tr>
<th>Data Quality Issue</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Invalid value</td>
<td>Valid value can be “1” or “2”, but current value is “3”</td>
</tr>
<tr>
<td>Cultural rule conformity</td>
<td>Date = 1 Feb 2018 or 1-1-18 or 2-1-2018</td>
</tr>
<tr>
<td>Value out of the required range</td>
<td>Customer age = 204</td>
</tr>
<tr>
<td>Verification</td>
<td>City and State do not correspond to ZIP code</td>
</tr>
<tr>
<td>Format inconsistency</td>
<td>Phone = +135432524 or (001)02325355</td>
</tr>
</tbody>
</table>

**Improving Data Quality**

Done with data profiling and ready to move towards improving data quality? Not so fast. Data profiling is an ongoing process of discovery. Establish a quality-oriented culture in your organization by rewarding people who find and report issues in data – just like Japanese manufacturers do in the automotive industry.

Once you are done with the evaluation, start reengineering processes to improve data quality while profiling goes on continuously at all organizational levels, from front-line data entry operators to executives who use analytics. Changes to source systems will be required, but you'll need to deal with them delicately by involving both business and IT users to balance implementation at both technical and operational levels. Unless your organization already has a Master Data Management (MDM) system in place, which contains master copies of all data, you’d want your data warehouse to eventually serve as your MDM. This means cleansing, persisting, conforming, and de-duplicating numerous data sets across the organization would be necessary in the data warehouse.

The first step to improving data quality once profiled is a series of tests incorporated at any points in the data integration process. The tests could refer to a number of business rules or mathematical operations to validate your data, for instance. We’ll cover this part in further detail later.

If a dataset passes the tests, it's clean and can be moved to the production data warehouse for modeling. If not, your data validation process should be able to create an error event record, and either stop the process, suspend erroneous data, or simply tag the data.
Data Quality Tests

In terms of architecture, all data quality tests are structured similarly but differ in scope. Let’s take a look at the data quality categories defined by Jack Olsen in his book “Data Quality: The Accuracy Dimension”:

**Column-level testing**

Data is tested on a very granular level, within a single column. Data quality rules applied at this stage include checking if the value is null, from a fixed finite list, falls within a specified range, fits field patterns specified in the database, is not a part of exclusion lists, and passes basic spellcheck.

**Structure-level testing**

This type of testing checks data relationships across multiple columns. For instance, fields across columns may be checked to verify a hierarchy, like a one-to-many relationship. It also checks foreign and primary key relationships. Every field of a specific column can be checked against another column to verify something like postal addresses.

**Business rule testing**

Complex testing comprises of creating business rules. This type of testing may involve something like checking the eligibility of an airline customer and their status as a Platinum Member by verifying whether their frequent flyer miles exceed 2 million and they have been a regular member for at least 5 years to qualify for the Platinum status.

With these measures in place, we can start taking action, as described in the “What Does Data Quality Mean For Your Data Warehouse” section at the beginning of this eBook:
The action you take will vary and depend on the type of data you're working with, and typically is the responsibility of the business department that works with a specific type of data set.

Note that all of the measures we've discussed till now focus on improving the quality of existing data rather than addressing the root cause, which is often at the point where data is entered in a transactional system by front-line employees. If you're truly invested in improving data quality, you will also need to implement rules that perform scrutiny at the time of data entry. For instance, in a financial institution, the management may notice that the social security numbers of customers are often left blank or entered incorrectly. They can implement a rule that makes the field value ‘required’ in a format specific to social security numbers (AAA-GG-SSSS) while disallowing nonsensical entries, like 999-99-9999.

Data Integration: Bringing Together Data from Different Sources to Improve Quality

Data integration, as a methodology, is different, but in the context of data quality, it refers to integrating data related to the same entity across different systems. For instance, information about a specific product may be found in your US database, but the same product may be sold in different countries too. This means that records of the same product are spread across different databases with respect to region. In every region, the product may be sold under a different name, under different branding, and with different patterns used to describe information in database records.

When building your data warehouse, you'd have to integrate these dispersed pieces of information across multiple database to form a master view for reporting purposes. Let's take a look at an example:
Using our original customer and product example, integrating data in this way revolves arounds two important processes:

- Recognizing whether the same customer entity exists in both sources
- Combining customer data to obtain a consolidated view of the product table
When trying to find if two entities are linked, you could start with a common field that is likely to exist in the same pattern across systems. For customer entity, this field could be the tax ID number. If the same tax ID number exists for customer records in different systems, you have just identified commonality in an efficient manner. However, we are rarely so lucky as to have such simple solutions in the world of databases.

If you cannot find a common field, all available product information would have to match across tables to determine if the same customer entity exists between two systems. Modern data quality management tools automate this sort of work, which generally takes subject matter experts hours of scouring through rows and tables to find links. Let’s take this example further and see how product information can potentially be matched.

Say your US database contains brand, product description, and product identification number all in one field, in varying patterns. In the UK, for instance, the database records just the product description, but that too in varying patterns depending on who entered them. An automated data quality tool can determine commonality by:

- Applying operations to brand names to make them consistent
- Parsing the product description from US and UK database into individual attributes and sorting by brand name
- Fixing differences in how product attributes are recorded
- Use fuzzy logic to match product attributes across both databases
- Show reports of products that match and link them to a customer entity

Using data integration in this way has saved companies hundreds of man hours annually. The best way to go about this when building your data warehouse is to choose a data warehouse automation tool that has a built-in data quality module.
Data Augmentation: Adding Value to Data to Increase its Usefulness

So by now, we’ve conducted data profiling, looked at ways to improve data quality, and found how integrating data across multiple sources to determine commonality adds value. The natural last step to complete the data quality cycle is to look at ways to augment existing entity data with that from external sources, outside your own databases.

For entities like Customer, data augmentation is common. Your marketing automation application could, for instance, contain valuable insights about customers that could be used to augment records in the data warehouse. The additional insight will help your business tailor product offerings better by utilizing deeper segmentation opportunities.

The table below gives an example of the kinds of data that can be obtained from external sources to augment the master record:

<table>
<thead>
<tr>
<th>State</th>
<th>CL</th>
</tr>
</thead>
<tbody>
<tr>
<td>City</td>
<td>Move</td>
</tr>
<tr>
<td>ZIP</td>
<td>65464</td>
</tr>
<tr>
<td>ZIP+4</td>
<td>3234</td>
</tr>
<tr>
<td>Delivery ID</td>
<td>3</td>
</tr>
<tr>
<td>Route ID</td>
<td>4</td>
</tr>
<tr>
<td>Address</td>
<td>6546 House Way</td>
</tr>
<tr>
<td>House Number</td>
<td>6546</td>
</tr>
<tr>
<td>Street</td>
<td>House Way</td>
</tr>
<tr>
<td>Street Type</td>
<td>Way</td>
</tr>
<tr>
<td>County ID</td>
<td>635</td>
</tr>
<tr>
<td>County Name</td>
<td>Glow</td>
</tr>
<tr>
<td>District</td>
<td>47</td>
</tr>
<tr>
<td>Record Type</td>
<td>Personal</td>
</tr>
<tr>
<td>Latitude</td>
<td>35.4685</td>
</tr>
<tr>
<td>Longitude</td>
<td>64.2334</td>
</tr>
<tr>
<td>Census Group</td>
<td>35632165</td>
</tr>
<tr>
<td>Census Tract</td>
<td>35</td>
</tr>
</tbody>
</table>

www.datawarehouseinfo.com
In the above example data, a business can potentially look at the address, postal code, and ZIP+4 code to determine if the customer falls in a specific housing segment. For instance, houses in a particular region that have ZIP+4 codes were built in the 80s and spanned 2500 sq-ft. This information can be used to target certain product offerings to all such customers.

While the above was just one example, in reality, data augmentation using the address field to build correlation is common. Product data is another example of data that is used for augmentation purposes. Buying patterns, especially when building predictive models, can be determined using correlations in product data with other augmented data.

Another prime example can be seen in the case of manufacturers. As a manufacturer, you’d only know what and how much you’re selling to retailers or wholesalers - not the actual amount of product sold to the final customer. Research firms, like Nielsen, provide data of this sort, which manufacturers purchase to gain a better understand of sales patterns so they can enhance their product offerings and strategy in terms of inventory and delivery.

While not directly tied to data quality at its core, data augmentation should be the natural last step if you’re building an intensive data quality management framework within your organization.
Conclusion: Build Reliable and Accurate Analytics Capabilities with Data Quality

The purpose of any data warehousing initiative is to provide business intelligence, and that purpose is defeated if enough thought is not given to building a comprehensive data quality framework, with the end-result being inaccurate analytics, and therefore, bad decisions.

Use the data quality framework described here as a base to build your own processes. Your focus should be on adding value to existing data however you can, rather than merely fixing errors in production data.

In the best-case scenario, you should be using a data warehousing tool that automates the repetitive tasks of building the data warehouse while integrating data quality, profiling, and lineage features within the data warehouse design process from the start. Focus on reducing the number of tools and dependencies during the project to minimize time taken and resources spent for best possible results.
About Data Warehouse Information Center

Data Warehouse Information Center is a knowledge hub that provides educational resources related to data warehousing. It is dedicated to enlightening data professionals and enthusiasts about the data warehousing key concepts, latest industry developments, technological innovations, and best practices. It also covers exclusive content related to Astera’s end-to-end data warehouse automation solution, DWAccelerator.

About Astera Software

Astera Software is a rapidly-growing provider of enterprise-ready data solutions. We help business users bridge the data-to-insight gap with our suite of user-friendly yet high-performance data extraction, data quality, data integration, data warehousing & electronic data interchange solutions, which are used by both midsize and Fortune 500 companies across a range of industries.

Acclaimed for their intuitive interface and advance functionality, our products offer the same level of usability to both developers and non-developers, allowing business users to spend less time managing data and more time using it.

Follow Data Warehouse Information Center on