The State of Data Quality in the Enterprise, 2018

Paxata | SourceMedia RESEARCH

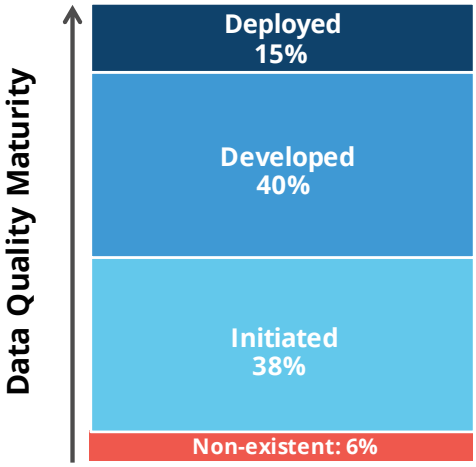
Businesses are dealing with more data than ever before. This means they can leverage multiple layers of knowledge and develop sophisticated strategies to achieve success, right? Not so fast. Results from the 2017 Data Quality Study, a survey of 290 executives and IT professionals at enterprises with $100M or more in annual revenue, show that organizations are struggling with a variety of data quality and data preparation issues as they strive to turn data into valuable business insights that can drive organizations forward. As we approach 2020, what does our current state of data quality mean for business analysis in the coming years?

Unfortunately, many businesses are behind the curve when it comes to data quality. According to the online survey conducted by SourceMedia Research and sponsored by Paxata, only 15% of organizations have actually deployed and just 40% have developed a mature data quality model.

The need to reach more comprehensive levels of data quality across the organization and various lines of business constituents, however, is becoming more pervasive. One respondent, for example, sums up the call to move in this direction, concluding that "monitoring the quality of data can actually roll up into higher level performance indicators for my business as a whole."

Data Quality Maturity Model

The following graphic illustrates how far along organizations are when it comes to actually leveraging mature data quality models:



Organizations with a mature data quality model share the following attributes:

- Higher data lake usage
- Higher public cloud usage
- More likely to use data profiling
- More likely to perform higher value data prep activities
- More satisfied with their data quality
- CIOs have more data quality responsibility

Which of the following best describes your organization's overall data quality strategy?
Base=All respondents (n-290)

Note: Due to rounding, percentages do not add up to 100

Source: Paxata Data Quality Study,
SourceMedia Research/*Information Management*, November 2017

Key Takeaways

- Those who have invested in data quality maturity report higher satisfaction levels.
- Companies are experiencing two major obstacles: significant data variety and complex mix of data types
  - 37% of any organization's data comes from external, second party and third-party sources
  - 64% report using mostly structured/little unstructured data, 21% structured/unstructured, and 6% all unstructured data
- Data preparation process breaks out with data ingest taking the majority of time (30%). Data profiling (21%) and data remediation (21%) follow in order of effort.
- The use of Excel or custom coding for data preparation is at very high levels among all data quality maturity levels. This is an indication of a lack of comprehensive functionality and business-friendly tools in existing data quality solutions.
- Tools with the following criteria seem to be on buyers' agenda when investing in data quality:
  - Support interactivity with structured and unstructured data
  - Usability for less technical users
  - Embedded visualizations
  - Performance and scalability in ingesting and prepping large volumes of data

COPING WITH QUALITY CHALLENGES

Organizations are finding it increasingly difficult to manage their burgeoning data resources and coming up against two main obstacles, regardless of size or data maturity level:

- **Roadblock No. 1:** Significant data variety. Companies are ingesting data from first-party, second-party and third-party sources. In fact, about 37% of any organization's data comes from external, second party and third-party sources.
- **Roadblock No. 2:** Complex mix of data types. Only 8% report using all structured data while 64% report using mostly structured/little unstructured data, 21% structured/unstructured data and 6% all unstructured data.

This influx of various types of data prompted one survey respondent to admit that his organization was simply struggling to "keep up with the amount of data being generated."

DATA PREP CHALLENGES

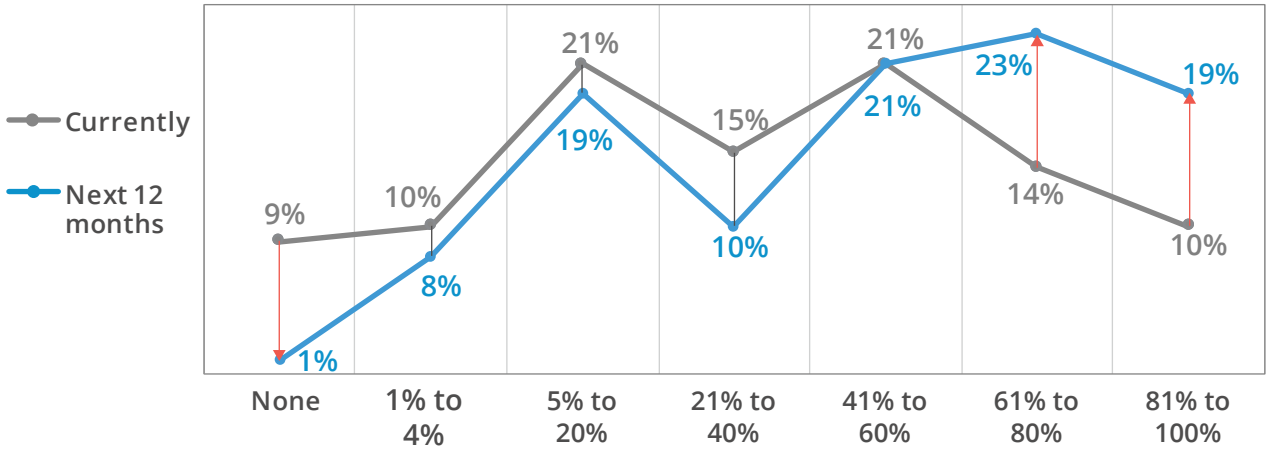
While organizations are adeptly addressing storage issues, they are wrestling with data preparation and processing challenges. Consider the following: Businesses are spending a significant amount of time just getting data into usable shape – as respondents overall indicate that the majority of their time is spent on data preparation. This leaves significantly less time to spend on data

Cloud Data Storage is a Popular Trend

While data volumes have been rapidly growing, businesses to a large extent have been able to manage it effectively. Indeed, organizations are increasingly gravitating toward the use of the public cloud and data lakes for data storage. For instance, 84% of all organizations surveyed are already using the public cloud to store at least some portion of their data. And, while just 14% of organizations currently store 61% to 80% of their data in a data lake, 23% will be storing this amount of data in a data lake in just 12 months (see chart).

Current/Future Data Lake Storage

In the next 12 months, the percentage of respondents currently storing a large percent of their data in a data lake is expected to increase (see upward pointing arrows) while the percent of respondents currently storing a small percent of their data in a data lake is expected to decrease (see downward pointing arrow).



What percent of your organization's data is currently stored with a data lake, and what percent do you anticipate will be stored with a data lake in the next 12 months?
Base=All respondents, excluding "don't know" (n=286)

Source: Paxata Data Quality Study,
SourceMedia Research/*Information Management*, November 2017

analytics activities, which is where the rubber meets the road for business insights and subsequent value to an organization's strategy. Even organizations with fully mature data quality are spending the majority of their time on data preparation. Drilling down even further reveals that these organizations are spending the greatest amount of this preparation time on low level activities – with 30% of time spent simply ingesting or collecting data from various sources, 42% in combined data profiling and data preparation, and much less time spent on higher level activities such as data quality monitoring (only 16%) and data profiling and investigation (only 12%).

Regardless of the level of data quality maturity achieved, organizations are not spending significant time on advanced data quality activities. Organizations that have deployed mature data quality are devoting 17% of their time to data quality monitoring and 14% to predicting future quality issues, and organizations overall are spending 16% of time on monitoring and 12% on predicting. Regardless of data quality maturity, organizations are spending the majority of their time and resources on low-level activities such as data ingest and preparation.

The struggle is resulting in less-than-optimal satisfaction with data quality, as respondents across the board gave data quality a middling 3.4 score on a 1 to 5 scale. Of course, as more attention is paid to data quality issues, satisfaction increases, as evidenced by the fact that organizations that have deployed a mature data quality model have achieved a 4.2 satisfaction score, while those that have developed a mature model have reached a 3.9 level on the satisfaction scale (see chart).

What's more, these organizations that are experiencing data quality satisfaction are significantly more likely to be using data profiling, preparation and quality tools. In-

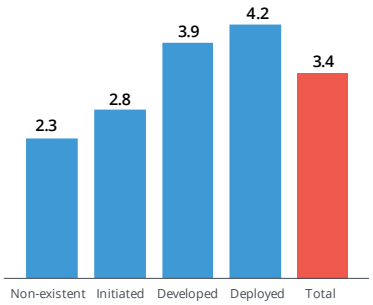
What is an Enterprise Data Quality Strategy?

A comprehensive approach to reaching higher levels of data quality satisfaction that includes components such as:

- the use of data profiling, preparation and quality tools
- high-value data prep activities
- a high level of line of business involvement in data quality

Data Quality Satisfaction

Organizations that have advanced toward the deployment of mature data models are experiencing greater levels of data quality satisfaction



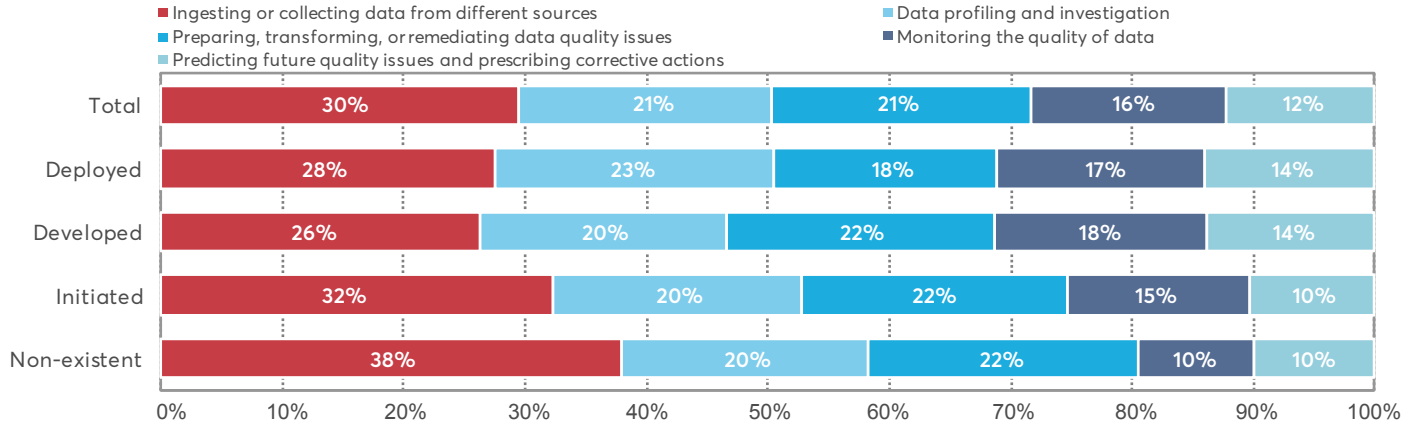
Statistically significant at 95%: Deployed and Developed are significantly higher than Non-existent and Initiated

Overall, how satisfied are you with your current data quality capabilities to date? Using 5-point scale (Not at all satisfied to Very Satisfied) Base=All respondents (n=290)
Mean percentages out of 5

Source: Paxata Data Quality Study, SourceMedia Research/*Information Management*, November 2017

Data Preparation Activities

Mature segments perform more high-value data prep activities (e.g., monitoring data quality and predicting future quality)



Statistically significant at 95%: a) For "Monitoring the quality of data," the Non-existent segment is significantly lower than Initiated, Developed, and Deployed; b) for "Predicting future quality issues," the Deployed and Developed segments are significantly higher than Initiated (using one sample t-tests between proportions)

Now, just thinking about [x]% of data preparation activities that you specified, what percent of these data prep activities is spent on each of the following types of data prep activities? Enter a whole number percentage for each. Sum to 100%.
Base=Respondents indicating time spent on data preparation (n=285)

deed, 56% of organizations that have deployed a mature data model are using these solutions and 36% of organizations that have developed a mature model are leveraging them, compared to just 23% of organizations overall.

SEEKING MORE SOPHISTICATED SOLUTIONS

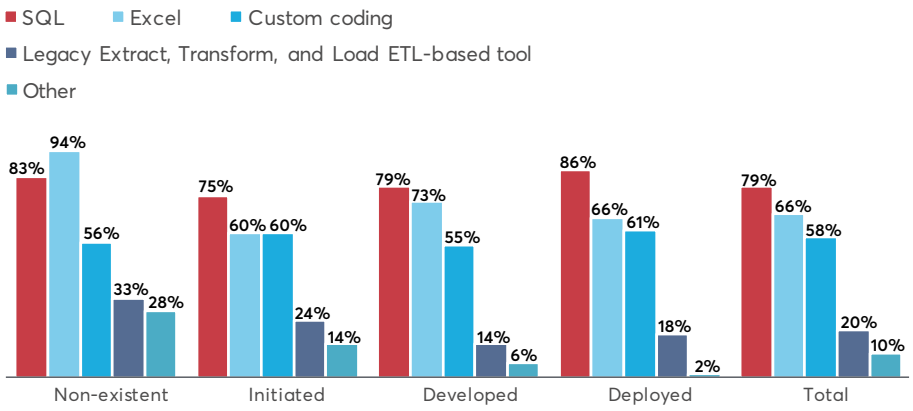
To mimic this success, organizational leaders across the board are likely to be looking to access, purchase and implement profiling, preparation and quality tools as they recognize the business need for improved quality.

Moving toward the adoption of more sophisticated data quality tools is likely to be a top-of-mind issue in 2018 and beyond. No doubt, the pressing need to head in this direction is evidenced by the fact that Excel is still a frequently used data profiling and data preparation tool – with 68% of all organizations using Excel. What is interesting is that even in organizations with mature data quality, Excel, custom coding and SQL is used quite significantly. This can be an indication of lack of functionality and business friendliness of deployed data quality solutions in these organizations. Adding self-service data prep and interactive data profiling to these solutions is likely to improve these statistics.

The good news is that many business leaders and analysts know what they want in a data quality solution. In fact, 62% of survey respondents say they are looking for performance and scalability, 56% for visualizations, 55% for de-duplication and 53% for usability for less technical users when purchasing data quality tools.

Excel and SQL Usages

The need for more sophisticated data quality tools is apparent as most organizations still rely on Excel and SQL



Which of the following types of tools does your organization use for profiling data? Select all that apply.
Base=All respondents (n=290)

Source: Paxata Data Quality Study,
SourceMedia Research/*Information Management*, November 2017

IT and Business Managers: The Need for Collaboration on Quality Matters

According to the survey results, IT professionals are most likely to take primary responsibility for data quality, while line of business managers are most likely to take on a secondary role (see sidebar on page 6).

Data quality, however, is an issue that touches everyone within an organization. As such, shouldn't business domain experts and line of business managers be on more equal footing when it comes to data quality?

Survey respondents certainly didn't shy away from pointing to this need for more equal representation. "For a data quality management initiative to be successful, cooperation between the IT and business areas must be guaranteed. This association is important because, although the technical profiles will be responsible for the construction and control of the environment, the business users will be the owners of the data and . . . both [need] to quantify future objectives, and to know strengths and weaknesses of our [data quality] projects," advises one of the survey respondents.

Another respondent recommends that such collaboration be required saying that it is important to "integrate all stakeholder subject matter experts in [data quality] initiatives from the start with executive sponsorship in writing." One survey participant even calls for data quality to be linked to compensation by tying a "percent of CEO's bonus" to data quality and implementing a "data quality rewards program" for staff.

Buyers, however, need to beware of the fact that many data quality tools might not deliver exactly what is needed. Indeed, survey respondents point out that there is a significant performance gap in data quality tools. Consider the following:

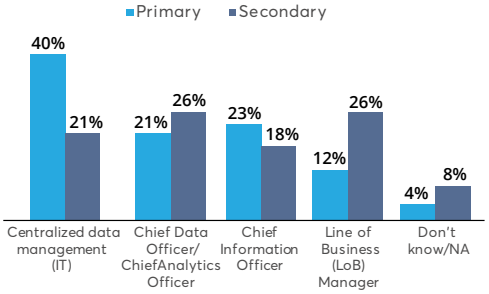
- While 56% of respondents cited visualization as a very important purchase driver only 33% rated their current visualization tool as very effective.
- While 53% cited usability for business teams as a very important purchasing driver only 32% rated their usability as very effective.
- While 51% cited live interaction with data as very important, only 28% rated their live interaction with data as very effective.

What businesses need is an information management tool that can support interactivity with data, both structured and unstructured; can ingest and prepare large volumes of data; and that allows business users, as well as technical staff members, to become more fully engaged in data quality initiatives. Such a tool would enable business users, many of whom do not have a lot of technical knowledge, but know the context and meaning of data better than technical users to visually explore, transform, and publish contextual and clean data for analysis anywhere.

Some information management solutions can address the most time-consuming part of data analysis projects by providing an intuitive, visual, and interactive application for business users to onboard, profile, and create quality information. Thus, helping organizations reach higher levels of confidence in their analytics, and organizations can finally get on the fast track toward turning raw data into the valuable business insights necessary for success.

Overall Data Quality Responsibility

Overall, IT is the most common function with primary data quality responsibility; CIOs are second most common

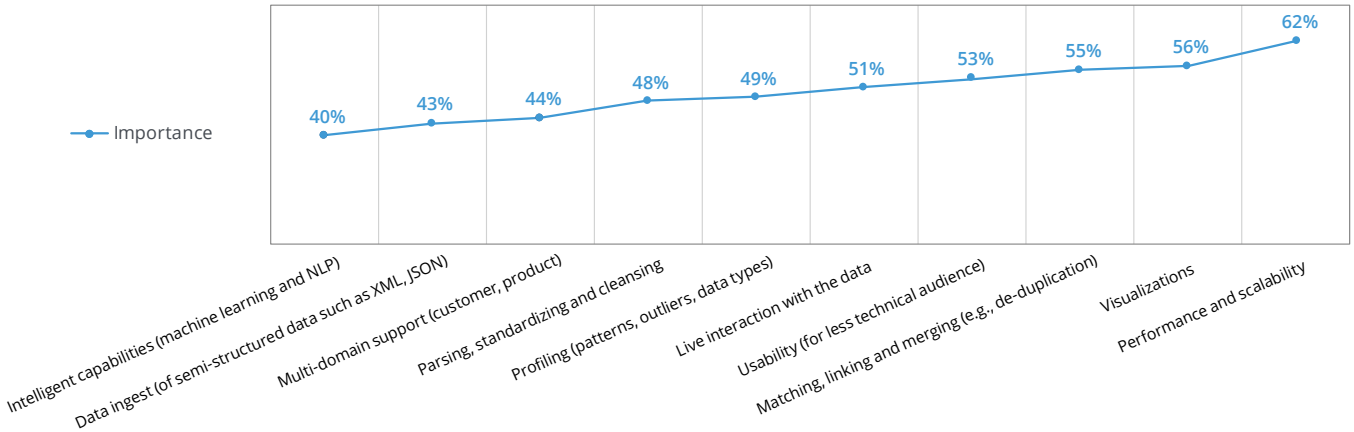


Which of the following functional roles have the primary and secondary responsibilities for data quality at your organization? Select one for each: primary and secondary.
Base=Respondents with data quality strategy (Initiated, Developed, or Deployed) (n=272

Source: Paxata Data Quality Study, SourceMedia Research/*Information Management*, November 2017

Data Quality Tool Purchase Drivers

Performance/scalability, visualizations, de-duplication, usability, and live interaction with data are top data quality tool purchase drivers



How important is each of the following capabilities when considering the purchase of data quality tools? (5-point scale: "Not at all important" to "Very important")
Base: All respondents, "don't know" excluded (n=variable)

Source: Paxata Data Quality Study, SourceMedia Research/Information Management, November 2017

METHODOLOGY

In November 2017, SourceMedia Research conducted an online survey among 290 end-users and buyers of IT products and services at enterprises with $100M or more in annual revenue drawn from the opt-in audience of *Information Management.*

ABOUT PAXATA

Paxata is the pioneer in empowering business consumers to intelligently transform raw data into ready information, instantly with an enterprise-grade, self-service, scalable, intelligent platform. Our Adaptive Information Platform weaves data into an information fabric from any source, any cloud, or any enterprise to create trusted information. With Paxata, business consumers use clicks, not code to achieve results in minutes, not months. Companies around the globe rely on Paxata to get smart about information at the speed of thought. Be an Information Inspired Business. For more information, visit paxata.com

ABOUT SOURCEMEDIA RESEARCH

SourceMedia Research, a unit of SourceMedia, provides research solutions for marketers, agencies and others targeting business sectors such as banking, payments, mortgage, accounting, insurance, employee benefits and investment advisor / wealth management. SourceMedia Research specializes in reaching senior and C-level decision makers through access to its large proprietary opt-in databases and panels.