Naveego™

Critical Requirements for Data Accuracy Platforms

Aaron Rutledge
Vice President of Solutions Engineering

WHITE PAPER

Table of Contents

Introduction

The Data Accuracy market (traditionally defined in terms of Data Quality and Master Data Management) is currently undergoing a paradigm shift from complex, monolithic, on-prem solutions to nimble, lightweight, cloud-first solutions. As the production of data accelerates, the costs associated with maintaining bad data will grow exponentially and companies will no longer have the luxury of putting data quality concerns on the shelf to be dealt with "tomorrow."

In attempt to meet these challenges, companies will be tempted to turn to traditional Data Quality (DQ) and Master Data Management (MDM) solutions for help. However, it is now clear that traditional solutions have not made good on the promise of helping organizations achieve their data quality goals. In 2017, the Harvard Business Review reported that only 3 percent of companies' data currently meets basic data quality standards despite the fact that traditional solutions have been on the market for well over a decade.[i]

The failure of traditional solutions to help organizations meet these challenges is due to at least two factors. First, traditional solutions typically require exorbitant quantities of time, money and human resources to implement. Traditional installations can take months or years, and often require prolonged interaction with the IT department. Extensive infrastructure changes need to be made and large amounts of custom code needs to be written just to get the system up and running. As a result, only a small subset of the company's systems may be selected for participation in the data quality efforts, making it nearly impossible to demonstrate progress against data quality goals.

Second, traditional solutions struggle to interact with big data, which is an exponentially growing source of low-quality data within modern organizations. This is because traditional systems typically require source data to be organized into relational schemas and to be formatted under traditional data types, whereas most big data is either semi-structured or unstructured in format. Furthermore, these solutions can only connect to data at rest, which ensures that they can't interact with data streaming directly out of IoT devices, edge services or click logs.

Therefore, a new generation of Data Accuracy solutions is needed to meet the challenges of digital transformation and modern big data governance. As Michael Ger and Richard Dobson point out in their recent white paper, these solutions must be capable of ingesting large quantities of real-time, semi-structured or unstructured data, and be capable of processing that data both in-place and in-motion.[ii] These solutions must also be easy for companies to install, configure and use, so that ROI can be demonstrated quickly. As such, the Data Accuracy market will be won by vendors who can empower business users with point-and-click installations, best-in-class usability and exceptional scalability, while also enabling companies to capitalize on emerging trends in big data, IoT and machine learning.

Critical Success Factors

Installation

When executing against corporate data strategies, it is imperative to show measured progress quickly. Complex installations that require cross-functional technical teams and invasive changes to your infrastructure will prevent data governance leaders from demonstrating tangible results within a reasonable time frame. That is why it is critical that next-gen Data Accuracy platforms be easy to install.

Look for solutions that are cloud-first, platform agnostic and self-service enabled. It should be possible to install the solution on your network and start connecting to your data sources (including data lakes and edge services) from a graphical user interface (GUI) in a matter of hours, without the need for infrastructure changes or customization.

Business Adoption/Use

Many of the Data Accuracy solutions available on the market today are packed with so many complicated configuration options and features that they require extensive technical training in order to be used properly. When the barrier to adoption and use is so high, showing results fast is nearly impossible. That is why it is critical that Data Accuracy solutions be easy to adopt and use.

Look for solutions with intuitive GUI and no/low-code configurations. Tasks such as setting up and executing quality checks, managing workflows, defining metadata, and configuring match/merge rules should be possible through point-and-click operations in a natural-language-based GUI that any savvy business user can master.

Return on Investment

The ability to demonstrate ROI quickly is a critical enabler for securing executive buy-in and garnering organizational support for an enterprise data governance program. In addition to being easy to install, adopt and use, next-gen Data Accuracy solutions must also make it easy to track progress against your enterprise data governance goals.

Look for solutions that provide easy integration with the best-of-breed data visualization and dashboarding tools. This will allow you to visually demonstrate progress against critical metrics such as ROI, thereby legitimizing your program in the eyes of key stakeholders.

Business Intelligence and Analytics

At the end of the day, a Data Accuracy program will be judged on the extent to which it can enable powerful analytics capabilities across the organization. Having clean data is one thing. Leveraging that data to gain competitive advantage through actionable insights is another. Data Accuracy platforms must be capable of preparing data for processing by best-in-class machine learning and data analytics engines.

Look for solutions that offer data profiling, data enrichment and master data management tools and that can cleanse and master data across highly disparate data sources and organize it for consumption by analytics engines both inside and outside the data warehouse.

Key Functional Capabilities

Below we outline the key functional capabilities that a next-gen Data Accuracy platform should provide to the user. The emphasis here is on **ease of use**. While most traditional solutions may offer these capabilities, they are often difficult for the average business user to utilize without help from the technical team.

Auto-Provisioning and Self-Service

Auto-provisioning and self-service refer to the user’s ability to provision their own tenant without manual technical intervention from either the vendor or the customer’s IT staff. This typically is achieved through the use of a downloadable installer and/or through the web login creation process.

Plug-and-Play Data Connection

Next-gen Data Accuracy solutions should leverage a plugin architecture to mediate connections with disparate data sources. Connecting to a data source becomes a matter of simply choosing the right plugin and entering some configuration settings. Companies can contribute their own plugins by following the development guidelines for unique/proprietary data sources.

Automated Schema Discovery

Automatic schema discovery means the system can automatically identify the logical structure of the data in your data sources. In relational database systems, this equates to automatically discovering the tables, columns and data types for each schema in your source. Data Accuracy solutions should offer similar capabilities for other types of structured and semi-structured data sources as well.

Automated Data Profiling

With automated data profiling the system can automatically scan your source data in order to calculate and present useful statistics regarding your data. This could include information on null, min, max and average values; pattern matching; and data type analysis.

Graphical Data Mapping/ETL

Data mapping/ETL refers to the process of transporting data from various sources into the Data Accuracy tool for cleansing and mastering. For most traditional Data Accuracy solutions, this typically involves technical tasks such as creating databases, schemas and tables, as well as writing complex scripts to extract, transform and load data. In contrast, next-gen Data Accuracy platforms should provide a user-friendly GUI where users can define both the shape of the master data and the mappings from source to master. This should be possible without the need to create new databases or write complicated code.

Auto-Suggestion of Quality Checks

Giving the user the ability to author quality checks is great, but it can be impractical if your system has hundreds or thousands of tables to monitor. Auto-suggestion enables the system to automatically suggest ways to measure the quality of your data based on the statistics derived during the profiling step. Users can then make selections from among the suggestions, or simply accept all of the suggestions and iteratively disable the ones that are no longer needed.

Data Cleansing

In addition to leveraging automatically generated quality checks, users should have the ability to define custom data checks using a standard query syntax such as ANSI SQL. In addition, the system should provide mechanisms (whether manual or automated) for correcting problems found in the data. This can include tools for organizing, categorizing and assigning quality checks, as well as managing the workflow associated with making corrections to the data.

Domain Agnosticism

Many Data Accuracy solutions are tailored to meet the needs of a specific domain and can't be usefully applied outside of that domain without customization (e.g., customer, product, etc.). As the number and variety of data sources continue to proliferate, it is simply not practical for a company to have to maintain disparate solutions for each of its critical business domains. Look for solutions that provide metadata and mapping mechanisms that can be seamlessly adapted to any domain without the need for extensive customization.

Data Enrichment and Validation

Having clean data in a system is one thing. Having a complete 360-degree view of your customers, partners and products is another. Data enrichment augments the data in your system with data from your data lake and/or from authoritative third party providers. This enables your business to get a clearer picture of who your customers really are and how they've interacted with your company (or similar companies) in the past.

Business Process Workflow

A streamlined workflow capability enables business users to manage business processes associated with common data governance tasks. This should include the ability to define business rules, assign tasks, set notifications, approve actions, document detailed instructions, provision new users and manage user roles and permissions.

Metadata Management

Metadata management means providing tools that enable users to easily catalog the data in their source systems, define shapes for master data entities and map source system schemas to master data shapes. A viable Data Accuracy solution allows a user to configure this through a GUI without having to write custom code or complex queries.

Match and Merge Configuration

Matching rules allow users to define the criteria that determine how individual data entities will be identified across source systems. Merging rules allow users to define how data will be combined from multiple data sources to create a composite "golden" record. Once again, it should be possible to configure this through a GUI without having to write custom code or complex queries.

Write-Back Configuration

Write-back configuration allows users to define how master data will be propagated back to source systems. Users should have the ability to define which sources will receive data, specify the properties that will be synchronized for each data source, and define the method by which data will be written back to the system. As we've stated for previous capabilities, it should be possible to configure this through a GUI without having to write custom code or complex queries.

Data Lineage Visualization

Data lineage includes a record of a data entity's origin and the sequence of events that led to its current state. Visualization tools should be provided to enable users to traverse the origin and history of each composite "golden" record in the system. Users should be able to see when each composite record was created, matched, merged and synchronized back to source, as well as how each source contributed to its creation.

Data Accessibility Tools

Data Accuracy platforms should provide tools for basic data integration, such as OData APIs or plugin connectors for data visualization tools, and should provide the ability to transfer master data to various destinations.

Key Technical Capabilities

Below we outline the key technical capabilities that a next-gen Data Accuracy platform should leverage. The emphasis here is on **innovation**. While some traditional solutions may offer some of these capabilities, they are often buried beneath an unwieldy, bloated, monolithic architecture that prevents vendors from releasing new innovations on their product in a fast and iterative manner.

Platform Agnosticism

Next-gen Data Accuracy offerings cannot afford to be locked in to a single platform, form factor or technology stack. Ongoing mergers and acquisitions will ensure that companies find themselves owning and supporting a wide variety of potentially incompatible technologies, and the proliferation of SaaS solutions ensures that at least of some of those solutions will be living in the cloud. Data Accuracy solutions must have the ability to be installed on, to connect to, and to master data across a spectrum of platforms and technology stacks, residing both in on-prem data centers and in the cloud.

Cloud-First Orientation

According to Gartner analyst Dave Cappuccio, 80 percent of enterprises will have shut down their data centers by 2025.[iii] Cappuccio writes, “As interconnected services, cloud providers, the Internet of Things, edge services and SaaS offerings continue to proliferate, the rationale to stay in a traditional data center topology will have limited advantages.” This is especially true with regard to data. As more and more data are generated in the cloud, it simply doesn’t make sense to route all of that data through on-prem data centers for processing. Look for Data Accuracy solutions that are built for the cloud and/or offer a straightforward migration path to the cloud.

Big Data Ingestion, Processing and Storage

Staying relevant in the market means keeping up with the volume, velocity and variety of big data sources. As data lakes become data swamps, companies will increasingly demand products that can master data across all of their corporate data sources, including the data lake. Next-gen Data Accuracy solutions must leverage technologies for distributed stream processing, data-flow orchestration, microservice architectures, cluster computing and distributed data storage in order to ingest, process and store data coming in from data lakes, edge services and IoT devices.

Data Encryption and Security

Data compliance, privacy and security concerns are not going away. If anything, they're getting amplified as organizations begin to consider migrating their systems of record and data lakes into the cloud. Data Accuracy solutions must comply with local, national and international regulations and take every precaution to ensure that data is secured both at rest and in transit. Solutions that leverage best-in-class technologies for encryption, authentication, and authorization and that understand how to work effectively within regulatory parameters will be well-positioned to meet the security and compliance needs of their customers both now and in the future.

Machine Learning and Artificial Intelligence

Data Accuracy vendors who are committed to leveraging machine learning (ML) and AI techniques will have a distinct advantage over competitors in the Data Accuracy market. Whether it's leveraging ML to optimize the discovery, matching and merging of data, or using predictive analytics to anticipate data quality problems before they occur, the applications of ML and AI to Data Accuracy problems are practically endless.

Conclusion

When deciding whether a given Data Accuracy solution will meet the requirements of your organization's data accuracy needs, it's important to understand the extent to which it will enable your company to navigate the data quality challenges associated with digital transformation and data governance initiatives. As we've seen, modern organizations should look for solutions that can be installed quickly with minimal IT intervention and that can be adopted easily by business users. These solutions should be able to connect to big data sources and ingest the copious amounts of data generated by those sources. Organizations that choose solutions that meet the above criteria will be well-positioned to tackle their data quality challenges both now and in the future.

Critical Capabilities Checklist

Critical Characteristics	
	Installation
	Business Adoption/Use
	Return on Investment
	Business Intelligence and Analytics
Key Functional Capabilities	
	Auto-Provisioning and Self-Service
	Plug-and-Play Data Connection
	Automated Schema Discovery
	Automated Data Profiling
	Auto-Suggestion of Quality Checks
	Data Cleansing
	Graphical Data Mapping/ETL
	Domain Agnosticism
	Data Enrichment and Validation
	Business Process Workflow
	Metadata Management
	Match and Merge Configuration
	Write-Back Configuration
	Data Lineage Visualization
	Data Accessibility Tools
Key Technical Capabilities	
	Platform Agnosticism
	Cloud-First Orientation
	Big Data Ingestion, Processing and Storage
	Data Encryption and Security
	Machine Learning and Artificial Intelligence

Endnotes

1. Tadhg Nagle, Thomas C. Redman, David Sammon (2017). *Only 3% of Companies' Data Meets Basic Data Quality Standards*. Retrieved from https://hbr.org/2017/09/only-3-of-companies-data-meets-basic-quality-standards

2. Michael Ger, Richard Dobson (2018). *Digital Transformation and the New Data Quality Imperative.* Retrieved from https://2xbbhjxc6wk3v21p62t8n4d4-wpengine.netdna-ssl.com/wp-content/uploads/2018/08/Digital-Transformation.pdf

3. Dave Cappuccio (2018). *The Data Center is Dead.* Retrieved from https://blogs.gartner.com/david_cappuccio/2018/07/26/the-data-center-is-dead/