LTI
Let's Solve

FirstEigen

Whitepaper

AI-Led Cognitive Data Quality

Authors : Seth Rao, Angsuman Dutta, Himansu Sekhar Tripathy, Deep Sharma

A Larsen & Toubro Group Company

Contents

1. Background

Data Quality Management (DQM) impacts a number of key business drivers, ranging from regulatory compliances, to customer satisfaction, to building new business models. Quality is one of the key functions under Data Governance, as unverified/unqualified data has little value to the organization. One of the leading global research and advisory firm estimates that an average Fortune 500 enterprise loses about $9.7mn annually over data quality issues. Although the true intangible cost of poor data is much higher, the sad truth is that data quality has not been paid the attention it deserves.

One of the reasons for this discrepancy is the way data quality issues are identified in the current systems and tools. A techno-functional team reviews data assets of an organization, and writes a set of rules to identify anomalies that are flagged for the review of data stewards. As these rules are static in nature, they become obsolete in 12- 24 months and a new assessment is required. Another significant reason is that many of the issues are contextual and are not easily codified. Consider the example of a bank that approved a corporate loan for a frequent client of theirs, at terms the client had never borrowed before, and a product that client had historically shunned. That loan should not have been approved without verifying the client's intent. The loan data file had data quality errors, the duration of the loan was captured as 3 months and not 3 years. These subtle contextual errors cannot be caught with the traditional validation checks, like checking for completeness, uniqueness, consistency, accuracy, etc. All the checks presently done are independent of historical business context.

In such a dynamic business environment, the need is to augment the modernization of data management with AI-based data quality, thus achieving data semantics for delivering trusted business-critical data at organizations' fingertips.

2. Quality Assurances – Validity & Reasonableness

At its simplest, data quality can be broken into two categories: completeness and accuracy. Completeness refers to ensuring that all expected data is received. Accuracy evaluates the validity of the data. Completeness and accuracy can be subjective, and should be guided by the line of business and the type of business. For example, car insurance premium increase of greater than 25% in one cycle may not be accurate.

Quality assurance edits used to check for both completeness and accuracy, can be broken into two broad categories: validity and reasonableness. Validity edits identify definite errors, and often result in submissions being rejected. They are frequently used to validate formats, ensure completeness, and highlight obvious errors. Format edits are used to reject data, which does not conform to the specified format, such as text in a date or numerical field or an email without an "@" symbol.

Reasonableness edits look for information that is highly unlikely or is an extreme outlier, but these are extremely complex to do correctly. Reasonableness edits don't generally cause a data submission to be rejected, but may require an explanation. Reasonableness can be based on the statistical probability of the value, business rules, or acceptable tolerances. These edits can be lenient or strict, depending on purpose. Stricter edits generally result in more edit failures, which typically lead to higher operational costs.

3. Traditional Approach Can be Expensive and Error Prone

Curating quality data requires time and money, for both, setup and operations. The time spent developing clear guidance and edit checks can save time and money by avoiding excessive data clean-up, or in a worst case scenario, unusable data. Data governance, data standards, and quality assurance edits all help minimize the data quality problems. Using industry-defined terms and formats can reduce errors because they minimize the need to define and transform data.

Cost of Incorrect Data

Gartner reports that 40% of data initiatives fail due to poor quality of data and affects overall labour productivity by ~20%*. That is a huge loss on which it's hard to even put a cost figure on. Forbes and PwC have reported that poor DQ was a critical factor that led to regulatory non-compliance. Poor quality of Big Data is costing companies not only in fines, manual rework to fix errors, inaccurate data for insights, failed initiatives and longer turnaround times, but also in lost opportunity. Operationally most organizations fail to unlock the value of their marketing campaigns due to Data Quality issues. Our research estimates that an average of 25-30% of time in any big-data project is spent on identifying and fixing data quality issues. In extreme scenarios where data quality issues are significant, projects get abandoned. That is very expensive loss of capability!

4. Alternative Approach based on AI/ML

Manually setting rules for 100's of Tables with 1000's of columns is unrealistic. We've frequently seen companies with 10's of thousands of tables and 100's of columns in each. There are no SME's who know every column of every table to be able to capture every rule needed to validate a data set. The data is just too vast and diverse. The gamut of data quality rules specific to the dataset must be autonomously learned using cognitive algorithms. These rules will be dynamic and evolve as the data evolves, to reflect the new reality. The AI/ML powered data quality system will behave like an individual who is not constrained by the initial set of rules they have learnt, but continue to learn and evolve as their surroundings change.

Interacting with our customers we saw that looking for errors in vast amounts of data was like looking for a needle in a haystack. It's a very complex problem for large data sets, flowing at high speeds, from many different sources, via many different platforms. It's a nightmare for the SME's, coders and for the people who want to make decisions based on that data. Consider the example of a bank which was onboarding 400 new applications in one year to their new IT platform. With an average of four data sources per app, and a mere 100 checks per source, their team was tasked with creating 160,000 checks.

Any rule-based system implementation will not scale in the new reality of big and/or complex data. Only machine learning systems can scale to the levels required by complex and/or large enterprises.

5. Who Should Go for Alternative Approach

Verticals: Organizations where data is used to make critical decisions, will need to have a high degree of certainty on the trustworthiness of their data. Every organization in every vertical we've worked with, has significant portions of poor quality data. The only difference we've seen is in the organizational maturity to realize how vulnerable they truly are. Those organizations who've realized they are vulnerable are highly regulated industries like Banking, Financial Services, Healthcare, and others. And most other industries are slower to fix their poor quality data situation.

Data Characteristics: When organizations deal with data that have any of these characteristics, they are highly likely to have more data errors:

- Big data
- Complex, inter-connected data
- Data aggregated from many sources or many IT systems/platforms (HDFS, Cloud, RDBMS, noSQL, mainframe, etc.)
- Constantly evolving data
- Non-monolithic, heterogeneous data, where rules have to be created for small micro-segments of data to validate their trustworthiness
- Operational and transactional data of reasonable volume

Unless your business is extremely simple, every organization will earn a few check marks in the above list.

6. Conclusion

Data Quality issues are hidden in all organizations, yet prevalent. Although a plethora of Data Quality tools is available, the Data Quality identification process in many enterprises is generally static, obsolete, time-consuming, and low on controls. Most of the processes have a lot of manual & static touchpoints, are low on auditability, and are time-consuming. Robust Data quality processes have to identify newer errors even before they occur. Using cognitive algorithms in identification of poor data will reduce effort & cost, and will improve quality scores dramatically. Even after engaging many programmers to solve the data quality problems, they never seem to go away. The only scalable path to good, reliable data is to leverage the power of AI to validate data autonomously.

About the Authors

Seth Rao

Ph.D., is the CEO of FirstEigen

Seth Rao, Ph.D., is the CEO of FirstEigen, a Greater Chicago-based Cognitive Data Validation company. Their flagship product, DataBuck, is recognized by Gartner and IDC as the most innovative data validation software. By leveraging AI/ML, it is >10x effective in catching unexpected data errors. It increases the reliability of data by self-discovering 1,000s of data quality relationships and patterns autonomously, updates the rules as the data evolves, and monitors the new data continuously. (http://www.firsteigen.com/databuck/).

Seth holds a Ph.D. in Engineering from Illinois Institute of Technology (IIT), Chicago, and has an MBA from Northwestern University's Kellogg School of Management, USA.

Angsuman Dutta

Entrepreneur, Investor and Corporate strategist

A. Dutta is an entrepreneur, investor and corporate strategist, with experience in building software businesses that scale and drive value. In his past roles, he has provided information governance and data quality advisory services to several Fortune 500 companies. He is a recognized thought leader, and has published numerous articles on information governance.

He earned a Bachelor of Technology degree in engineering from the Indian Institute of Technology, Kharagpur, an MS in Computer Science from the Illinois Institute of Technology, and an MBA in Analytical Finance and Strategy from the University of Chicago, USA.

Himansu Sekhar Tripathy
Data Management consultant

Himansu Sekhar Tripathy is a Data Management consultant, with over 18 years of experience in consulting and delivery of data solutions. His interest areas include enterprise data strategy, cloud data engineering, big data engineering, data integration, quality, metadata management, MDM, and data governance. As a technology evangelist, he believes in leveraging emerging technologies in pushing the boundaries on real-time next-gen analytics. Himanshu has a Master's Degree in Business Administration and a Bachelor's Degree in Computer Science Engineering.

Deep Sharma
Associate Consultant

Deep Sharma is an Associate Consultant in Cognitive & Analytics Practice unit at LTI, with around three years of experience in technology consulting, analytics market research and offerings creation on emerging hybrid technology trends across the Data & Analytics technology stack. He has a keen interest in various building blocks of Data & Analytics like Data Integration, Data Quality, Data Governance and Data Visualization. Deep has a Master's Degree in Business Analytics.