EDMCouncil precisely

EDM Webinar

Reimagining Data Quality: Key Modern Considerations

Live Date: June 7, 2023

Featuring:

Emily Washington, SVP, Product Management, Precisely

Scott Arnett, Sr. Director, Product Management, Precisely

Moderator: Mike Meriton, COO & Co-Founder, EDM Council

Recording: View webinar

Presentation: View slide deck

EDM Council Homepage: edmcouncil.org **Precisely Homepage:** precisely.com

ADDITIONAL LINKS:

Learn more about EDM Council's CDMC framework

Learn more about EDM Council's DCAM framework

Learn more about EDM Council's Groups and Leadership Forums

Thank you to the Precisely panelists for providing the below answers to all questions posed during the live webinar. For more information or additional questions, contact us here.

WEBINAR Q&A:

What are the most powerful AI/ML use cases that you have seen and used within data quality?

The great thing about AI/ML is that the potential is unlimited. The most frequently requested capability is to provide recommendations of potential data quality and enrichment rules - based on profile details, recommend deeper data quality logic to apply. This proves valuable for those who are building sophistication into data quality projects.

One of the most powerful use cases we have seen so far is related to entity matching. We define entity matching as the process of identifying records that are related to each other in some way that is for a significant purpose. For example, if you are trying to eliminate redundant information from your customer data, you may want to identify duplicate records for the same customer. Or, if you are trying to eliminate duplicate outbound correspondence (e.g. marketing information) going to the same address, you may want to identify records of customers that live in the same household.

Given the individual matching requirements, rule-based (deterministic) methods tend to be time consuming and error prone. By using transfer learning techniques, the user is prompted to enter the necessary information to configure the match logic, and it is automatically configured.

What is the role of data management vs data governance roles (i.e. owners, stewards, etc.) in ensuring quality of data?

At the highest level, we see data governance roles typically overseeing data quality standards across multiple applications, lines of business, and processes to ensure there is standardization, clear ownership, impact, monitoring and improving data quality scores. They tend to partner with data management roles, who are typically responsible for the execution of data quality and addressing issues. Data management roles are often responsible for the specific data sets and where the rule requirements in more detail are defined and implemented. Data management roles often have more access to fix operational data.

In your opinion, how do you think data observability will impact data quality as it matures?

We see significant interest in data observability to inform what data needs to be fixed from a data quality standpoint. Since data observability provides anomaly detection on data profiles, freshness, schema and data drift, users are proactively alerted on potential problem areas that need to be fixed. As AI/ML application matures, we are able to use these anomalies to recommend potential data quality rules.

Can Precisely help implement data quality rules, measure poor data quality, show the negative business impact across both real-time applications and batch data processing applications?

Yes. Precisely has a long history of helping customers with these issues. Because of this, one of the Data Integrity Suite's core use cases is to create and implement data quality in both batch and real- time environments, score and measure quality, and understand impact of data quality issues.

What are some of the recommended Data Quality metrics that need to be tracked at an enterprise level?

The most common data quality metrics are accuracy, consistency, completeness, timeliness, validity, and conformity. However, we more often see interest in measuring the quality against business KPIs. For example, for a marketing campaign outreach, track customer detail completeness for email population and delivery. Tracking data quality metrics against business objectives helps track data quality at the enterprise level and increases the ability to track ROI of data quality initiatives.

What are the new characteristics of intelligence to consider while looking for a Data Quality tool for an enterprise?

Intelligent data quality leverages business context, metadata, and other data dimensions beyond the specific data elements to drive additional automation. One key characteristic is applying data quality rules that are driven by metadata, data profiles and semantics. This allows for more targeted recommendations. It is important that the tool captures metadata from data quality rules and results to get smarter with recommendations in not only applying new rules, but also fixing the issues. There is a "human in the loop" dimension to addressing data quality that allows you to leverage historical methods of fixing issues to inform how you fix future issues. And when applying data quality to address and location data, recommend areas of enrichment to enhance the address details for downstream analysis or customer outreach by geography.

What is the best way to kickstart a Data quality initiative in your business?

The best way to kickstart a data quality initiative is to align to a business objective that has broad visibility or executive alignment. What highly visible business report or analytics initiative has known data quality challenges? By implementing data quality on high impact areas, this increases internal awareness of the value and allows scaling to additional use cases.

Can we use a data observability tool for the purpose of data quality at an enterprise level?

Data observability is a great way to enhance data quality at the enterprise level, as it proactively monitors larger sets of data to identify potential problem areas based on profile details and historical patterns. The capabilities of DO and DQ are highly complementary and work together to strengthen data quality. Data observability assists in identifying issues across larger sets of data, which works well with traditional data quality capabilities focusing on fixing/standardizing/cleansing the data.

Do you track thresholds for data quality across regions or globally as one metric ?

We have seen multi-region data quality measuring at the region level, and where appropriate, rolling up to a global level. But not all data quality metrics are tracked the same in different regions, lines of business, etc. A core capability you should look for in a tool is having flexible configurability to set align to how your organization weighs data quality metrics and be able to adapt as your requirements change and your data quality programs mature and scale.

It is not difficult to monitor data quality but could be very challenging to resolve data quality issues or improve the DQ as it might involve changes in the process, system and people. End up every issue could be a mini project to involve all these stakeholders to assess and discuss it. It takes months or years to resolve a single issue. Is this the correct way? Do you have any suggestions?

Correct. We believe the reason fixing data quality issues becomes so complex is because of the disconnected understanding of the impact of the issues. That is why, from a data quality tool standpoint, we encourage integration with a data catalog, leveraging metadata and adding business context. You can more quickly understand data lineage, ownership, and downstream impact. By leveraging an integrated suite of capabilities that includes data observability and data integration, you can more easily collaborate with users. This allows for more transparency into prioritization and resolution status.

Can you elaborate on predictive rules? Does the tool profile data elements and propose rules they should be assessed against? And does that mean the DQ dimensions should have been defined beforehand?

Predictive rules start when you connect to the source. We profile the data for basic semantics, formats, nulls, and other dimensions. It recommends rules based on the profile details. If the system sees there are duplicate values, we recommend a duplicate check. If we see address information, we recommend an address parser. As you apply more data quality rules, the system gets smarter with additional options based on rules you previously applied or you've applied on similar data sets.

What are the options for installing DQ monitoring in each repository as our data is not in a central location?

Precisely manages this through providing an abstraction layer between the data quality rules and the data source, to minimize the dependency of the logic being associated with the specific data source. You define the rule logic, which can be applied to multiple data sources. For example, you can create a data quality rule that checks that a customer ID adheres to a specific internal company format. Then that same rule can be applied to cloud sources, data warehouses, and operational environments. This allows for the standardized rule to be reused and conforms to the same logic, regardless of the source.

Should the data owner by asset be responsible for ensuring data quality? If not, who?

At more mature organizations, we see multiple levels of ownership to capture different data quality contexts. For example:

1. Data stewards are typically responsible for data quality standards across multiple applications, critical data elements, lines of business, etc. "Customer data must have an Address 1, Address 2, City, State, Zip, Country field"

2. If there are owners assigned to individual assets, yes that person is typically responsible for data quality at that asset level. "CustomerID is in <format>"

3. Application owners are responsible for ensuring data quality on the data stored within the application. "These data quality rules are applied to the CRM data."

4. Business owners (e.g. finance or marketing team) are responsible for data quality to ensure business outcomes. "Financial total on quarterly report > $1,000,000."

We have many siloed data QA processes at both the first- and second-line departments -- how can this process be more efficient and effective?

By centralizing data quality standards and applying to multiple sources, you have the ability to streamline QA processes. We have seen organizations leverage a data catalog to centralize standards and ownership. By tracking where data quality rules have been applied in various environments and processes, you can increase transparency across the organization. If broader audiences within IT and businesses can monitor data quality scores, trust increases, thus minimizing multiple requests and additional QA processes.