THE CASE FOR END-TO-END DATA VALIDATION

The volume and speed at which data enters your enterprise is rapidly causing a rise in the data anomalies that present themselves. On the front end, we are receiving more data from different sources with a wide variety of quality. Next, that data gets moved through the enterprise faster than ever, often with minimal checks. Finally, when that data appears on dashboards and reports, it hasn't been adequately vetted, nor has it always gone through the proper processes to ensure accuracy and usability.

Data Validation

Data Monitoring

Data Analytics

There are a lot of reasons these things are happening. First, with the advent of the business user ability to spin up data marts and data analytics front ends without the experience in proper data ingestion and quality controls would be one reason. The flip side is that traditional IT was not responsive enough to the business, so the business users moved in this direction out of necessity. Then, of course, the creation of cloud computing makes it easier for anyone with Opex funds to redirect those funds towards well-intended means to solve their problems themselves. Business owned development teams are all part of the new normal. In many cases bussiness units own the particular data domain that they are most familiar with. Who is better equipped to understand the "correctness" of data than the owners of the data themselves. So as data is created and verified by the business unit that owns that data asset, they become the authority for that data and have an obligation for the quality of their data assets as it moves through the enterprise. They are also responsible for ensuring the data is appropriately shared and consistent with the other business owners of data assets.

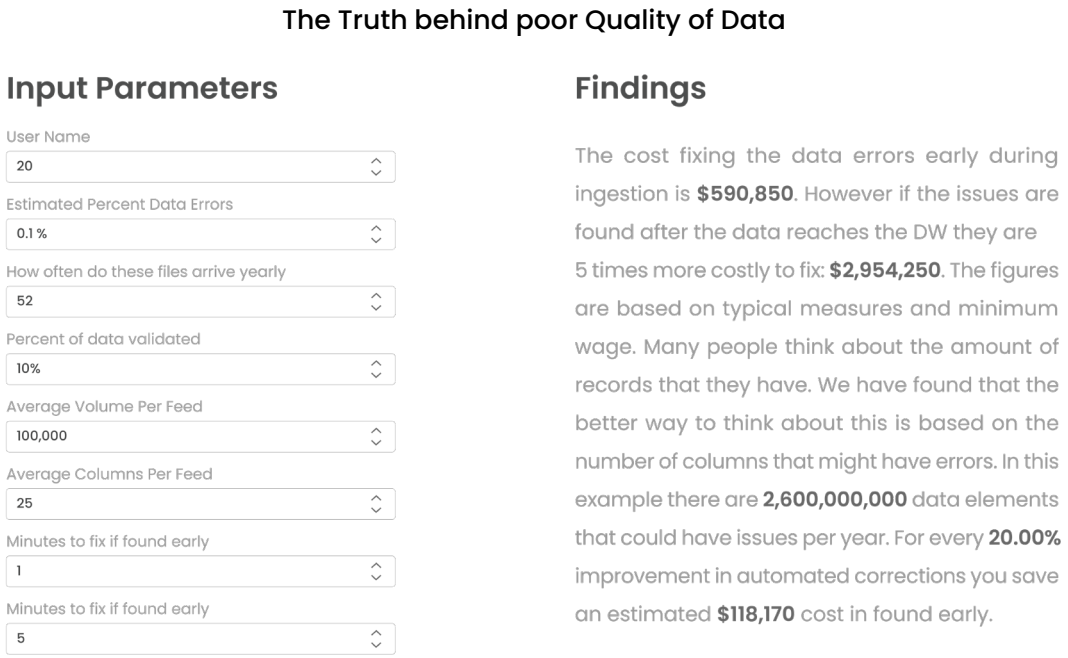
If this is not governed adequately it is full circle again with silos of information without proper controls and a means to systematically bring this data together to solve cross-business unit problems.

To function correctly and add value to the organization, processes must be established to ensure the data is moved correctly and encoded with other data properly. This is where end-to-end data validation processes improve the overall quality of the decisions made using the decentralized data that has come together in a centralized location for data analytics.

Continuous and constant end-to-end data monitoring plays a significant role in improving the trust in the data that runs your company gives you insights or is shared with clients or suppliers. We believe there are five critical components to this end-to-end monitoring.

1. First, at the point of ingestion, the data needs to be verified and cleansed for completeness, accuracy, formatting, and many other issues. This is the time to catch mistakes as they are less expense to remediate at this point.

2. Next, as the data begins to be consolidated and migrated, checks to ensure the transformations are executing properly need to be performed. Checks like referential integrity, duplicates, field truncation and field formatting need to be checked. Often, this is a manual and random process fraught with errors as most of the elements are left unchecked. A simple example illustrates this point well:

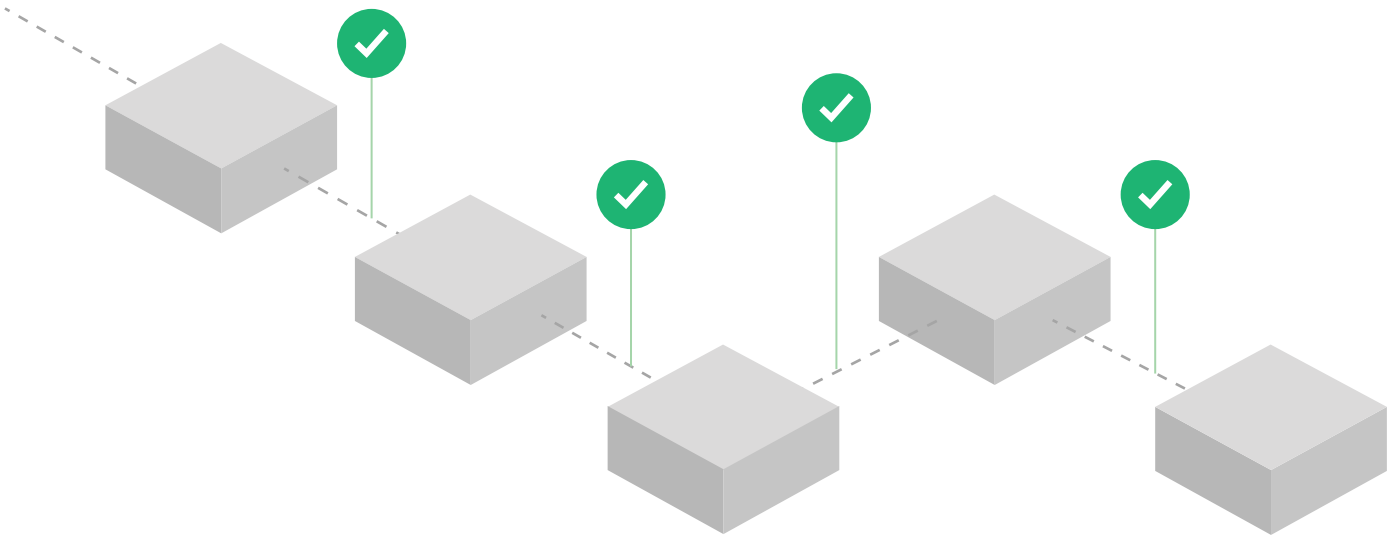


3. Once data is in place for consumption in either Data Warehouses, Data Lakes, Data Marts, etc., analytics platforms are used to create the dashboards and reports that run the business. There are many problems that data issues create once the data is being consumed. Of course, new releases of the software or applications can cause inconsistencies in the produced results. Formulas are incorrect from one application to the next. Visuals from the data analytics tools become corrupted for various reasons, such as software inconsistencies or unintentional errors introduced during the application upgrade. Often the performance of the application is impacted. Finding and fixing these issues takes not just data validation techniques but also techniques to compare visuals, schemas, security credentials, and performance. Often this is thought of as the end of the testing journey. We believe other capabilities need to be in place to continually monitor the environment.

4. Data Quality can score some or all of the data in your analytics platforms. As data is being ingested into the enterprise, Quality Scoring can be performed at a Table, DataMart, Data Lake, Data Warehouse and can role up to Quality Scoring at the systems level. If the scores deteriorate, a granular look at what makes up any of the underlining scores gives you insight into where to look further to maintain an appropriate score.

5. Finally, data boundaries can be very impactful in determining the quality of data that is ingested into your enterprise. This works by observing data as it is being ingested. Through AI algorithms, it learns the upper and lower boundaries for data elements and alerts you of anomalies outside these limits considering seasonality and other factors.

The first two capabilities are designed to help find issues in the analytics platform to be corrected early and at less cost. The third capability is intended to catch the problems during the data analytics processes in your enterprise. Numbers four and five are designed to monitor and alert you if things are going in the wrong direction and help point you to the problems where more attention might be needed by reviewing the results from the first two steps.



Conclusion : End-to-End Data Validation is a necessary part of the governance practice of an organization's data journey. In general, it is typically easier and less expensive to fix problems earlier in the cycle before data is combined with other data products from other business units. If organizations are consistent in the early ingestion process by the time they get to the analytics process the issues that arise are typically in the analytics processes and not normally a data issue problem. A formula might have been inadvertently changed; an upgrade might introduce anomalies. If you are following good End-to-End practices these late problems tend to be isolated in the data analytics tools not bad data. Therefore, as these are isolated they are easier to correct because you know where to look with more certainty.

Finally, Data Quality scoring and boundaries are great techniques to keep the data at a high level of quality and help identify trends that need to be investigated.