Data Ingestion Validation through Stable Conditional Metrics with Ranking and Filtering

Niels Bylois, Frank Neven, Stijn Vansummeren
UHasselt, Belgium

1 Abstract

Validating the quality of continuously collected data is crucial to ensure trustworthiness of analytics insights. A widely used approach for validating data quality is to specify, either manually or automatically, so-called data unit tests that check whether data quality metrics lie within expected bounds. Unfortunately, these existing approaches suffer from two limitations. First, as we show in this paper, the data unit tests that they support are based on *global* metrics that can only provide *coarse-grained* signals of data quality and are unable to detect *fine-grained* errors. Second, even when a data unit test signals a data quality problem, it does not provide a principled method to identify the data part that is responsible for a test's failure. We present an approach for data quality validation that is not only capable of detecting fine-grained errors, but also helps in identifying responsible erroneous tuples. Our approach is based on a novel form of metrics, called conditional metrics, which allow to compute data quality signals over specific parts of the ingestion data and therefore allow for a more fine-grained analysis compared to standard global metrics that operate on the entire ingestion data. Our approach consists of two phases: a unit test discovery phase and a monitoring and error identification phase. In the discovery phase, we automatically derive conditional metric-based unit tests from historical ingestion sequences, where we employ a notion of stability over the historical ingestions as a selection criterion for using conditional metrics as data unit tests. In the subsequent phase, we use the derived unit tests to validate the quality of new ingestion batches. When an ingestion batch fails one or more unit tests, we show how conditional metrics can be used to identify potential errors. We study different ways of implementing both phases, and compare their effectiveness. We evaluate our approach on two datasets and seven synthetic error scenarios. The improvement that we measure over global metrics as well as the error-identification F1-scores that we obtain indicate that conditional metrics provide a promising approach towards fine-grained error detection for data ingestion validation.