### Introduction

Here at Bigeye, anomaly detection is a critical part of our product: it enables us to monitor intelligently at scale. Rather than having to set manual, constant thresholds for thousands of metrics and hundreds of tables, we can automatically identify anomalies that might correspond to a problem in your data pipeline. Rather than inundating our customers with false positive alerts, our anomaly detection algorithms are sophisticated enough to adapt to business trend changes and respond to feedback.

Perhaps since its intricacies are abstracted away from users in Bigeye, anomaly detection is a feature that users often have questions about. In this blog post, we answer five of the most common ones.

1. What is anomaly detection?

Anomaly detection refers to detecting data points, events, and/or information that falls outside of a dataset’s normal behavior. Anomaly detection helps companies flag areas that might have issues in their data pipelines.

The anomaly detection system will need to learn the historical patterns present in each data quality attribute, learn what abnormal behavior looks like, and ultimately fire alerts that indicate real issues while ignoring behavior that is slightly off but not indicative of a real problem. This can be especially difficult when there are hundreds or thousands of data quality attributes being tracked simultaneously.

Naive techniques like gaussian models—that simply look at a number of standard deviations above or below the historical mean—fall apart in many commonly occurring time series patterns. A good model will need to adapt to various patterns that can regularly occur in these metadata attributes over time.

For a [deeper dive into anomaly detection, read here](https://www.bigeye.com/blog/anomaly-detection-for-data-observability-part1).

1. What are the alternatives to anomaly detection?

The main alternatives to anomaly detection for data observability are **simple, manual thresholds and rules**. For example, for a metric that tracks the percentage of successful Airflow jobs, we might set a manual threshold of 90%. If more than 10% of Airflow jobs failed over a certain period of time, then we’ll get an alert.

Simple, manual thresholds work for many scenarios, in particular in systems observability, where metrics are expected to remain constant over time. However, they also have a number of drawbacks:

* They require manual setting and tuning, which is fine when you have one data metric but not when you have 1000.
* They are inaccurate for metrics that are not constant.

1. How do I know I won't get alerted too much?

Since every business and data environment is unique, invariably any anomaly detection will produce some false positives, ours included. However, you will be able to adjust the data observability system to suit your alert sensitivities.

For example, you might only want to get alerts for extreme changes to the data batches rather than small fluctuations.

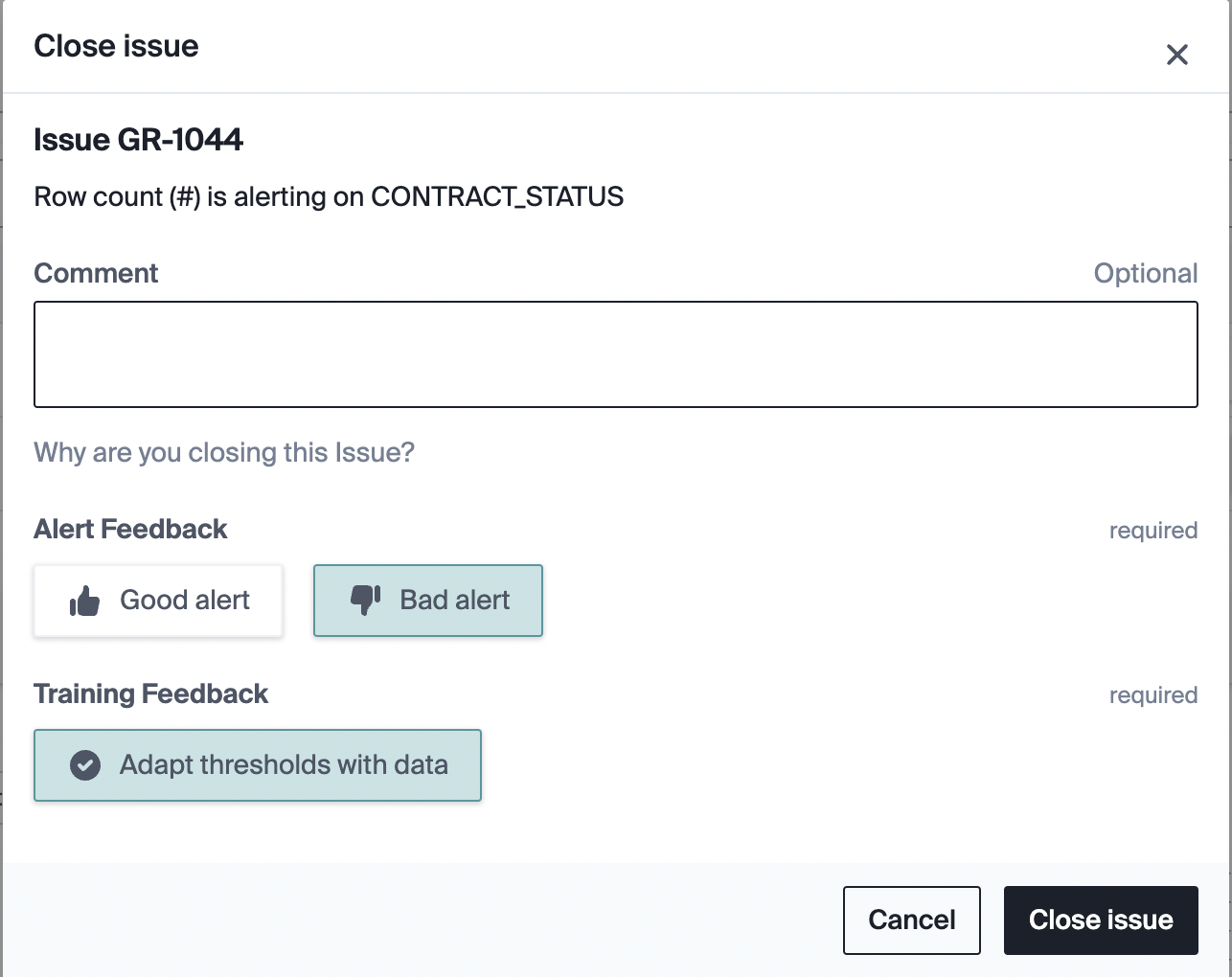
Or, the data team might want to understand if there’s any fluctuation in features within the machine learning feature store so that downstream automation is run with the most consistent inputs.

We also recommend that instead of monitoring all your tables at the same level, you follow a T-shaped monitoring strategy. This means that you track basic [metadata metrics](https://docs.bigeye.com/docs/available-metrics) across all your tables, and only go deep on your core tables.

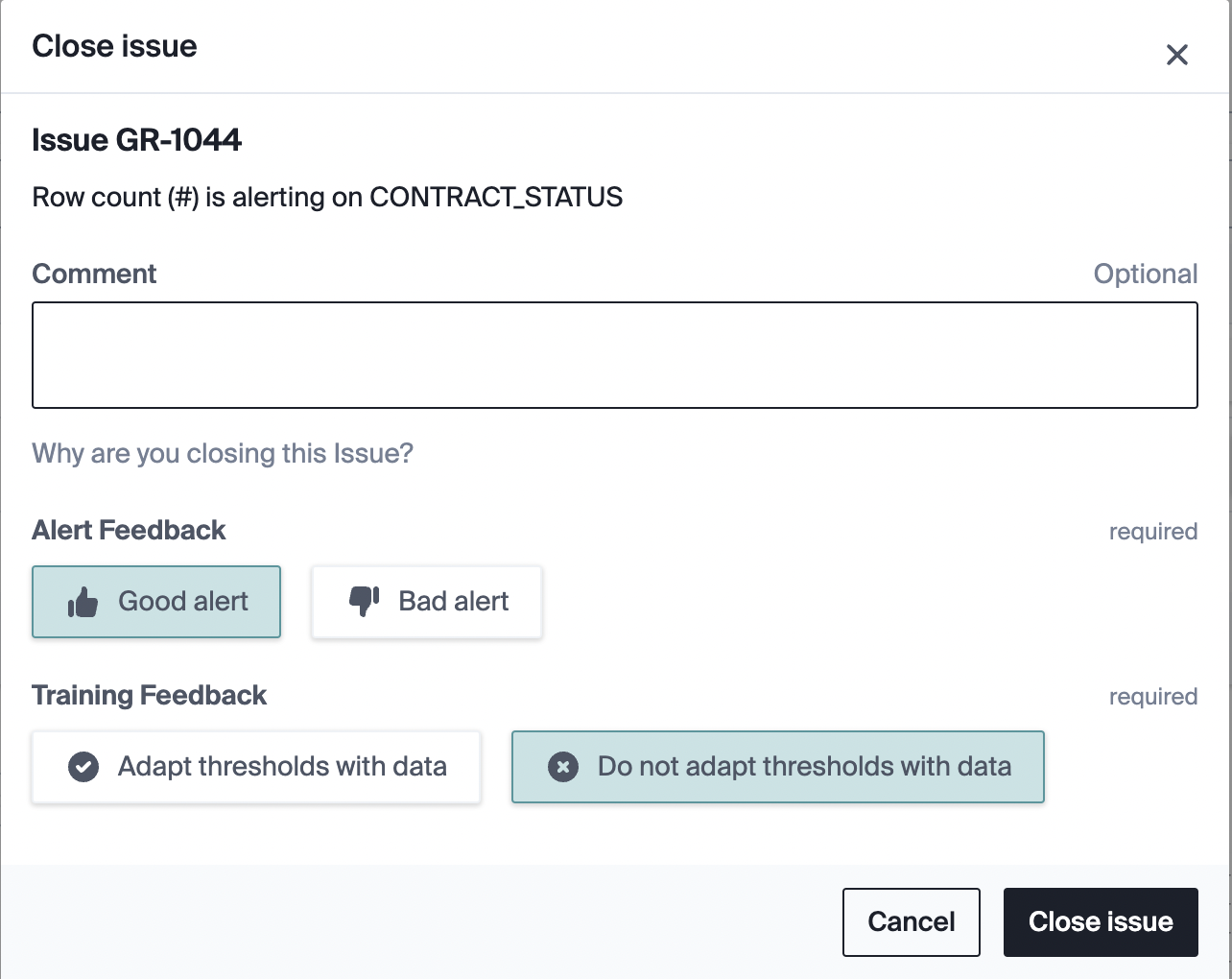
1. How much control over the anomaly detection engine do I have?

In addition to being able to control the width parameter on the autothresholds, you will also be able to [give direct feedback](https://docs.bigeye.com/docs/issues#giving-feedback-to-the-autothresholds-model) on the anomaly detection system. This data is then used to improve the model.

For example, when a data issue notification is fired, but the user thinks that the data batch in question is actually good in practice, the user can tell Bigeye that the underlying data state is tolerable or that a false positive alert is present. Bigeye will take this information into account so that similar behavior in the future will not trigger an alert.



Additionally, in the case of good alerts, you can tell Bigeye to remove the data associated with confirmed anomalies and to make sure that it’s not propagated into future predictions.



For example, the figure below shows a degradation that Bigeye caught and removed from model training automatically. If this bad-value handling wasn’t in place, the thresholds would adapt to the spike in the metric time series and become too wide, reducing their ability to catch future problems.



1. How long does it take for the anomaly detection to start working?

We do not recommend fewer than 5 points (5 days, daily data), and we accumulate the data over time. We have a backfill capability for data tables with past data.