anodot

WHY EVERY DATA LEADER NEEDS ETL MONITORING?

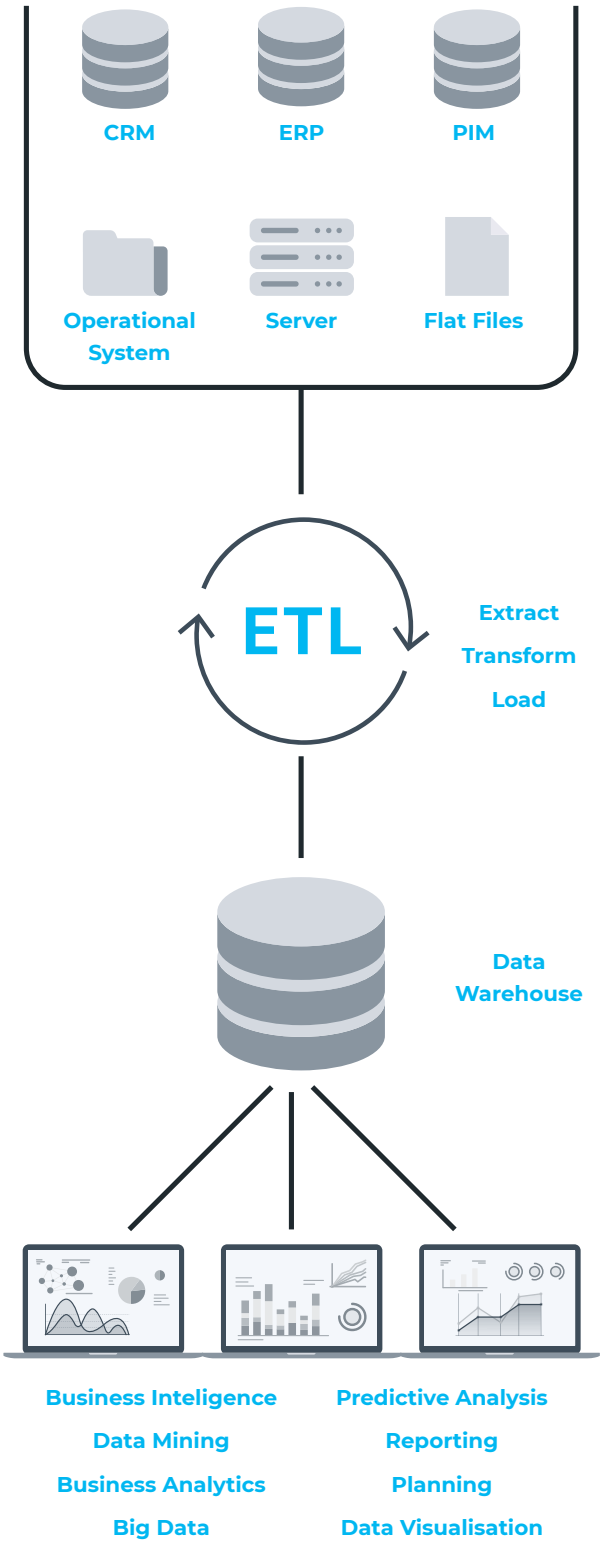
THE PROBLEM

It is 5 a.m. Tuesday. The ETL job that populates revenue data into your organization's data warehouse fails midway through the process. When the CFO open the mobile dashboard to review the last day results he immediately notices that the data is wrong, again. For the few hours, the on-call ETL Architect determines what caused the data-load failure, fixes the issue, and restarts/monitors the job until it successfully completes.

It is now 10 a.m., your organization's executive team is holding their strategy meeting, and they'll be using the Executive Dashboard developed by your team to review your organization's revenue numbers. Unfortunately, the latest figures may not be available because of the failure of the ETL job that loads revenue data. Last week, a similar event occurred, and your organization's reports and dashboards were unavailable until 3 p.m. – the prior month three similar data load failures occurred, but you can't quite explain why they happened.

THE SOLUTION

Latest machine learning technology enables us to apply automated anomaly detection capabilities to constantly monitor the ETL process. Using such techniques becomes a must when running a traditional batch ETL process in order to reduce time to detection and resolution of data quality issues before these affect the end users consuming data and insights.



WHY USE ANOMALY DETECTION?

1. Detect anomalies across a large number of ETL metrics can be identified in real-time and IT departments can be alerted.

2. Specific bottlenecks can be identified and addressed before the business wakes up to data delays (data delayed is data denied).

3. Impact of change management can be quantified through event correlation.

4. ETL pipeline inefficiencies can be identified early through correlation of anomalies.

5. Reduce the cost of operations by delivering alerts to specific teams for incident management.

WHAT DO I MONITOR?

Monitoring of your ETL jobs becomes a critical part of understanding what is going wrong. Specifically, understanding execution times, error counts, statistics of processed records of the various batches, jobs, and procedures is critical.

Without understanding anomalies across these metrics, IT departments often handicap themselves in the search for the errant line code in the ETL haystack. Anodot is solving this problem by using machine learning techniques to monitor ETL task and to identify anomalies in real time across specific parts of the ETL process and thus providing a constant view across your ETL process and immediate view of the system bottlenecks.

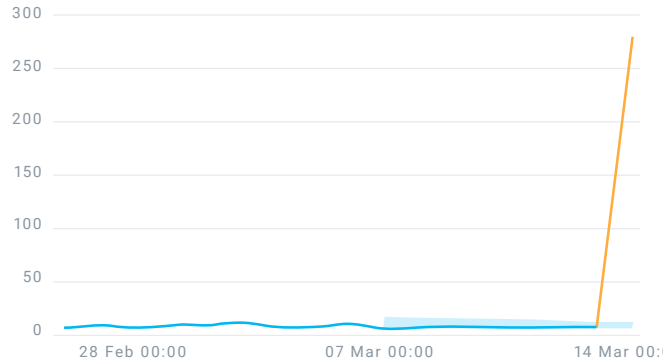
[USE CASE] - MONITORING TERADATA ETL

This alert monitors the number of missing records between step 3 and 4 of the process. Anodot automatically learns the standard behavior (between 2-7 missing records) and builds baseline around this behavior, when the number of missing records spikes to 278, Anodot identifies the anomaly and alerts the on-call DBA.

The incident: The number of missing records spiked.

Time to detect: 5 mins

RecordsDiff for Step4, Redshift



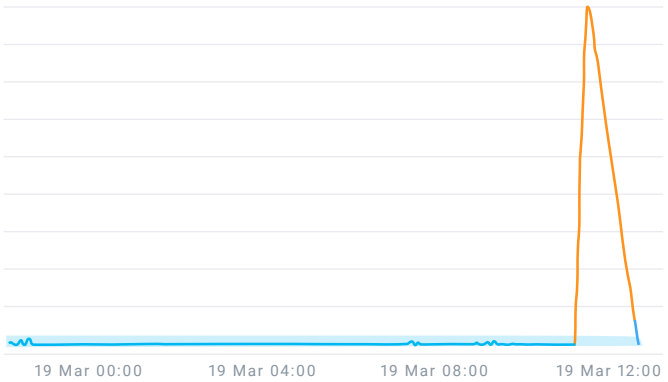
[USE CASE 2] - MONITORING TERADATA ETL

This alert monitors the time to complete each step. Anodot automatically learns the standard behavior and builds a baseline for this behavior, when the time to complete step 2 spikes beyond the usual baseline, Anodot detects the anomaly and alerts the on-call DBA.

The incident: The number of missing records spiked.

Time to detect: 5 mins

TimeToComplete for Step2, accounts



WHAT ALERTS SHOULD I SET UP?

ALERT NAME	MEASURES	DIMENSIONS
Anomaly in step length	TimeToComplete	StepId SourceId TargetId ClusterId
Anomaly in record count	NumberOfRecords	StepId SourceId TargetId ClusterId

SUMMARY

Recent advances in automated anomaly detection, where AI monitors activities such as ETL performance, enable companies to now sift through millions of metrics to immediately illuminate outliers and the root cause. According to an Anodot customer data, companies that used the anomaly detection platform managed to reduce their time to detection by 74 percent and time to resolution by 91 percent.