influxdata®

AN INFLUXDATA CASE STUDY

How to Improve Data Labels and Feedback Loops Through High-Frequency Sensor Anomaly Detection by Using InfluxDB

JULY 2021 (REVISION 1)

Ezako is a Startup Specializing in Machine Learning, AI and Time Series Analysis.

Company in brief

They are the creators of Upalgo, which is a time series management platform. Their solution uses AI to automatically detect anomalies in streaming data. Ezako's customers are predominantly in the aerospace, automotive, telecommunications sectors; their clients' data is mostly sensor and telemetric data.

Case overview

Ezako uses InfluxDB to power Upalgo platform. After considering multiple monitoring tools, Ezako picked InfluxData's solution to address their time series data requirements. InfluxDB is their chosen time-stamped data store; Ezako's team is able to focus on providing their clients with the best machine learning models and anomaly detection.

The business challenge

Ezako chose to focus on machine learning on time series data. Ezako's CTO, Julien Muller points out, they decided to focus on this specific sub industry as it's quite large already and will continue to grow. Muller articulates that there's going to be trillions of sensors available by 2025. This creates massive challenges as organizations are going to want to analyze all of this data. Ezako wants to design tools that can handle vast data sets. Part of analyzing high data volumes is being able to determine the interesting events within the myriad of time series data. Being able to provide their clients with stellar anomaly detection is key for Ezako's success. There are huge opportunities with the combination of time series data, IoT analysis, machine learning and anomaly detection.

Technologies Used: Chronograf, Telegraf, Python

The technical challenge

Some of the complexities of time series data analysis is because there is a direct correlation from one data point to the next. If you're collecting 50,000 data points per seconds, the data becomes old pretty fast; the data needs to be analyzed quickly to ensure the team can respond promptly. Through their Upalgo platform, Ezako wants to provide their clients with a baseline truth; this way their clients can use machine learning models to determine if the results are good or not. Ezako aims to make time series data more useful with their machine learning solutions.

Muller explains that during the learning phase of creating a ML model, you'll likely be using 5-10 million data points. A common challenge with detecting anomalies with huge data sets is that sometimes the algorithm detects false positives/negatives or other inaccuracies. Sometimes the algorithms are detecting random noise when nothing is actually happening. False negatives are more complicated, especially when dealing with 20 million points; there is no proof that there is an anomaly hidden among the data.

The solution

Ezako started using InfluxDB in 2016 - they appreciated that it is purpose-built for time series data and that the platform could handle the unique requirements for creating machine learning models. Their Uplago platform has to be able to handle a continuous flow of data ingestion without impacting the overall system. Muller points out that their platform has to be able to handle more than just continuous data ingestion. To create their ML models, they needed extensive metadata storage and calculation capabilities. Muller knew that InfluxDB could help their models learn based on large datasets while also enabling fast detection on small data sets.

Having a community to lean on and learn from is key for their continuous success. Muller is a data scientist whose forté isn't time series storage. The Ezako team didn't want to spend their time building schemas, managing time frames, etc. They value that InfluxDB is performant with native nanosecond handling and that it is schema-less. Muller expands that he has spent years building schemas and he doesn't want to do that anymore. Their customers come with them with existing data and Muller and team don't want to spend on creating the schemas - they'd rather focus on developing the machine learning models.

> *"You need a community that is able to help when you have technical issues."*

Julien Muller, CTO, Ezako

They chose to store their raw data in InfluxDB as they need a reference point for all future analysis, etc. There is always a need for data cleanup; while their clients' massive data sets are required to develop ML models, Muller and team realizes that sometimes their clients don't want to keep huge amounts of data in memory. There is a high probability that the feature windows are going to be reused when the team needs to iterate on the different algorithm hyperparameters. The data stored in InfluxDB is cached storage, it isn't stored in memory. This is ideal for Ezako's team as it can help with all of the data and they can query the data repeatability.

In addition to ingesting the data, their Upalgo solution is also processing the data. The team has to use data already stored in the database. Upalgo quickly reads vast amounts of data quickly while other processes are running. The data has to be cleansed, analyzed and the metadata has to be built.

At the end of the process, the model has to read the most recent data to enable predictions and anomaly detection. The last chunk of data is used to predict the next value. They use a REST API to query their data, which is stored in InfluxDB. This provides them with a common layer which allows them to ready the data regardless of their various technology stacks or the customer's UI.

Anomaly detection with time series data is challenging as two users won't have the same definition of an anomaly. Having a solid workflow is essential for implementing a good process; if done correctly, it can result in better data insights as anomaly detection is mostly just adding more information about your time series data. The most common algorithms for anomaly detection are One-Class SVM or Isolation Forest. For the ML models to be useful, their platform needs at least one million data points to learn from; it is more likely they'll need at least 2-3 million data points.

Results

InfluxDB has been a great platform for Ezako to use to explore their data sets. Ezako's clients often request to be able to view all of their historical data. Ezako started off with a manual data review process which didn't work well. Next they started to sample their data with InfluxQL. This worked better as it was quite fast. Ezako was able to provide their clients with a chart containing a sample data set with 2K-5K data points. This provides their clients with a large enough data sample based on their 20 million points without being overwhelming to navigate through.

Downsampling isn't used a lot as they want maximum data granularity. If Ezako downsamples their data too early, and they want their ML models to relearn the data, then the models aren't using the same data set. By starting with customers' raw data, the Upalgo platform is able to extract metadata and determine if there's any trends, seasonality or if the data is stationary.

It is also considered a best practice to calculate features on your data to determine that the correct time window has been used. At this point, users can use Upalgo's UI to quickly visualize their data. Restoring the data is a good idea at this point; tweaking ML models and detecting anomalies is ideal at this point.

> *"InfluxDB is way less work than other systems I've used. I just want to set it up and forget about it...it is simple to deploy and use with any dev environment!"*

Julien Muller, CTO, Ezako

What's next

Ezako hopes to improve upon their current downsampling capabilities. With anomaly detection, their clients want to discover interesting data points, rather than just random points. The team is looking into adding in manual downsampling capabilities into Upalgo and InfluxDB's downsampling capabilities. The team wants to be careful as they dont' want to lose any interesting data points as a result. They can't just query the data for the max or mean value within a specific time frame. The team is interested in upgrading to InfluxDB 2.0 and to watch the progression of InfluxDB IOx. They are also interested in learning more about Flux so they can start using the advanced querying functionalities.

About InfluxData

InfluxData is the creator of InfluxDB, the leading time series platform. We empower developers and organizations, such as Cisco, IBM, Lego, Siemens, and Tesla, to build transformative IoT, analytics and monitoring applications. Our technology is purpose-built to handle the massive volumes of time-stamped data produced by sensors, applications and computer infrastructure. Easy to start and scale, InfluxDB gives developers time to focus on the features and functionalities that give their apps a competitive edge. InfluxData is headquartered in San Francisco, with a workforce distributed throughout the U.S. and across Europe. For more information, visit influxdata.com and follow us @InfluxDB.