How to Prevent Data Downtime With Machine Learning

splunk>

turn data into doing®

You know your data is valuable, and you know that unplanned downtime in any system is a bad thing. So you can imagine why data downtime — when your organization suffers a lost, disrupted or incomplete connection to its data — can be an especially vexing problem. Data downtime often occurs during data ingestion, with the feed stopping and starting or stopping altogether. And your data admin may not be able to do anything about it, because the problem usually lies outside your system in whatever is sending you the data.

Data downtime isn't just an annoying issue, it's an expensive one. Various studies over the last few years have estimated that the average data practitioner spends 40 to 80% of their time validating data quality issues. You can do the math for your organization, but no matter what the number is, it's too high.

The greater the volume of data your organization ingests, the more expensive and complicated the disruption can be. The problem is compounded when you use multiple data sources and methods of ingestion. What's more, the problems created by data downtime may not rear their ugly heads until after the damage is done. Data that goes missing or is erroneous is often found too late, after it's already impacted downstream systems.

...the average data practitioner spends 40 to 80% of their time validating data quality issues.

Issues that affect your data ingestion and cause data downtime can also impact the trustworthiness of the data involved — and your organization. How can you — or your customers — be sure that the data you've ingested is complete and uncorrupted if there are disruptions you aren't able to detect? Data reliability has become an important topic for companies and governments alike. The UK Government National Data Strategy says "interruption to data-driven services and activities can cause disruptions to businesses, organizations and public services."

What causes data downtime?

While the phrase “data downtime” may be relatively new, its causes are just as old as any issue affecting an organization’s IT infrastructure. Data downtime can be caused by a simple bug in the code at any step of the ingestion process. Misconfigurations in any of the systems involved can also cause data downtime. An unexpected schema change is often the culprit of downtime and it can also be caused by traffic spikes.

Just because the causes are simple doesn’t mean the resolutions are, especially since resolving downtime has traditionally been a very reactive process. Once downtime has been identified, it still needs to be traced back to the root cause to be fixed.

Reducing downtime with anomaly detection

This is where machine learning (ML) can help — specifically a practice known as anomaly detection. Anomaly detection is based on identifying a stream, or collection, of data points that don’t conform to the expected pattern over a prolonged period of time. Applying anomaly detection on your ingested data can allow you to monitor your data-ingest pipeline in real time and find potential data quality issues and broken pipelines before downtime occurs. This means you can move to a proactive approach to addressing downtime before a problem actually occurs.

Enter Splunk and the Machine Learning Tool Kit (MLTK)

Luckily, if you're a Splunk user, there's a simple way you can proactively decrease downtime, using SPL and the Splunk Machine Learning Tool Kit (MLTK) to set up anomaly detection. It's a five-step process laid out in more detail in the free webinar, Prevent Data Downtime With Machine Learning. In it, you'll learn to create an ingest anomaly detection pipeline to proactively reduce downtime. To show you how straightforward the process is, here are the five steps:

1. Generate a predictive model that estimates the volume of events per sourcetype.
2. Calculate error statistics between predicted and actual values.
3. Identify anomalies from the error statistics.
4. Schedule periodic retraining of the predictive model and anomaly detection searches.
5. Create an anomaly detection dashboard and/or alert.

More About MLTK

MLTK helps you seamlessly apply ML on your data within Splunk, and can be deployed on top of Splunk Enterprise or Splunk Cloud. With MLTK, users can apply ML techniques — including anomaly detection — using search commands, and then monitor the derived insights in dashboards. Splunk MLTK is designed for users of all levels and has features including:

Experiments and smart assistants: A simple low-code experience guides model building, testing and deployment.

Extensible out of the box: More than 80 built-in scikit-learn algorithms and API support to plug in new runtimes.

ML in Search: MLTK extends the core search language with powerful ML search commands that allow users to train models and run inference using pre-trained models via a familiar user experience — directly in the search bar.

Embedded: Machine learning is embedded in Splunk premium solutions (such as Splunk IT Service Intelligence, Splunk Enterprise Security, and Splunk User Behavior Analytics), providing the fastest way to detect anomalies, cluster or group events and predict future outcomes using your machine data. These solutions are use-case-specific to address both security and IT operations challenges.

Guided Exploration: The app also supports machine learning model development through guided assistants, providing flexibility for those who want to go beyond configuring a pre-built solution and create custom models.

If you're ready to get started, watch the free webinar Prevent Data Downtime With Machine Learning, and find out how easy it is to use anomaly detection to protect your organization from data downtime.

22-24631-How to Prevent Data Downtime With Machine Learning-EB-105

splunk>®
turn data into doing®