IONATE Data Protection, Anomaly Detection and Cyber Security using AI/ML

Detecting anomalies in large-scale distributed systems--especially cloud-native environments, where a service is deployed on heterogeneous hardware and has multiple scenarios of normal operation--is exceedingly difficult. But it is also exceedingly necessary. By 2025, one-fourth of global data is estimated to be real-time, with IoT comprising 95% of that.

For normal business operations, anomaly detection is important for any organization to protect the mission-critical areas of its business. It also matters from a security perspective. Anomaly detection can help find an activity that possibly exposes advanced attackers, who may cause just enough activity to be outside an expected norm. This is especially true from a systemwide perspective. With some attacks, the revealing anomaly will only appear in data analysis of the entire system, versus that of a singular device. This is due to the specific traits of cloud-native applications, as described below.

A crucial hallmark of cloud-native applications, which can span multiple cloud providers, is the everchanging inter-relationships among numerous VMs, containers, functions, and service-mesh. While this allows a few workloads to scale to thousands in seconds, the unintended consequence is an elastic attack surface that scales up and down with the applications, making these environments difficult to secure. While the following traits are part and parcel to containers' tremendous value, they also significantly contribute to the challenge of securing cloud-native environments:

- extreme ephemeral nature of containers
- number of connections among microservices at scale [i.e., number of interdependent groups (and their interrelationships) that applications are now split into] increasing data variety from numerous data sources VMs,
- containers, functions, service-mesh (network)

With the elastic nature and growing complexity of cloud-native environments, it is increasingly difficult to find the origin of a security anomaly or incident and respond quickly. To solve for this, AIOps with AISecurity enables the detection of faults and service issues through Anomaly Detection.

Conceptually, Anomaly Detection is simple to grasp. Also known as outlier detection, anomaly detection is the identification of events or observations that differ significantly from the majority of the data (which is considered 'normal'), thus potentially qualifying as suspicious.

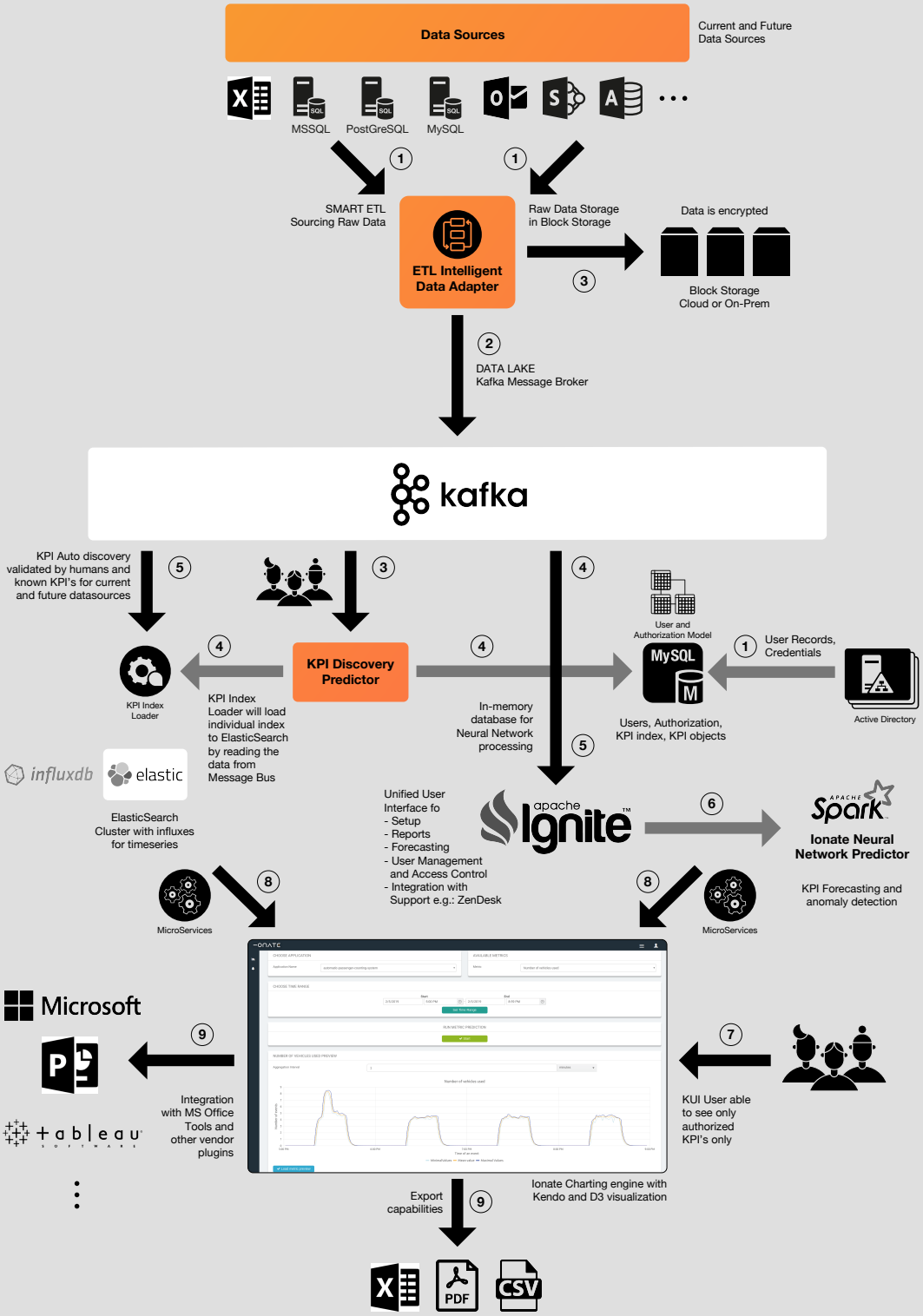
In real life, Anomaly Detection is not so simple. True visibility into the complexity of large-scale distributed environments now requires specialized capabilities and tools to provide insight into how everything comes together. Measurement and context are critical, especially given modern systems. Manually correlating volumes of data to locate anomalies is a surefire way to failure, especially as environment complexity and experiment velocity continue to increase. And the inability to monitor ALL DATA (as well as their interrelationships) is another cause of failure.

To be successful, a monitoring approach must be able to predict how one failure affects other services, as well as gauge the criticality of a particular failure. Applications and their environments must be continually reassessed, as the parameters of normal and abnormal constantly shift in a cloud-native environment.

Many solutions available today only monitor portions of an enterprise's data lake (i.e., logs, emails). Other data (such as that coming from a wifi router, or IoTs) may not be monitored for an anomaly. This leaves the vulnerability of an attack being launched from this network device, for example from an individual logged using the guest WiFi or malware exploiting an IoT vulnerability. Most solutions on the market today will wrongly categorize the events associated with such an attack as normal traffic. Only IONATE, with its ability to also gather and normalize data from this source too, will quickly alert to such an attack.

Deploying AIOps and AISecurity successfully depends on ensuring that all data inputs are represented, as well as on the quality and diversity of the data inputs. Variances in data size, as well as the structure, must be handled—as modern data sources differ from the large files, sequential access, and batched data of yesteryear. Poor data quality or non-representational data sets directly correlate to inaccurate and incomplete models. Developers must know, predict, and code all appropriate checks and validations to build effective models.

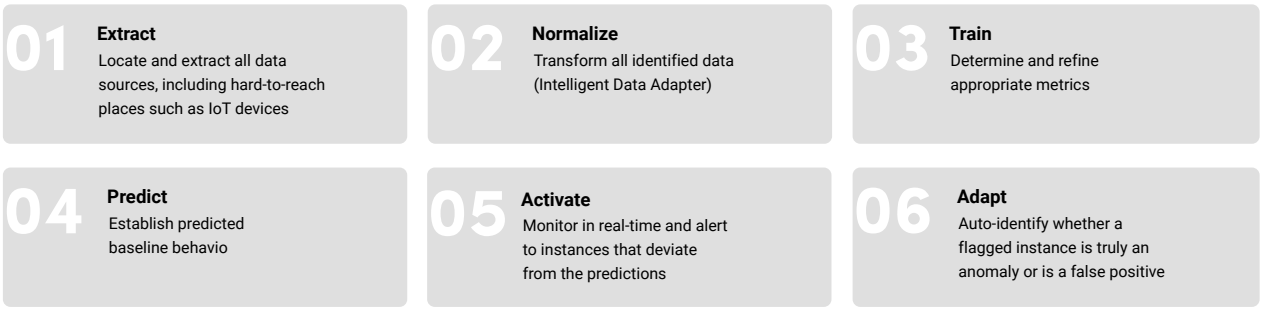
Thus AIOps and AISecurity can only succeed when a solution can locate and normalize all data sources and types and thus truly represent all components of, and interrelationships among, a distributed environment.



An example of a high-level workflow of the IONATE Anomaly Detection solution. First, we locate and gather data from all data sources that need monitoring. Then we set KPIs for all data sources. Our highly sophisticated and configurable algorithm then sets the expected--or predicted--behaviors of each measure. Once this is done, we then activate the solution to monitor all measures. Any deviation from a predicted behavior is identified as an anomaly to be investigated. If the anomaly is determined a false positive, (i.e., not an anomaly and poses no threat), we flatten or suppress that anomaly by retraining the algorithm using the flagged data points.

The IONATE AI/ML Cyber Security Solution

IONATE solves the issues associated with poor and incomplete data sets through the following steps, which are then described in further detail.



1. Extract

A key IONATE differentiator is our capability to gather and monitor all sources of data, both known and unknown. Through our proprietary ML-based Extract, Transform and Load (ETL) process, our solution avails itself to and then normalizes ALL data sources. With the IONATE solution, everything is monitorable because all data is extractable.

We monitor all systems, data producers and consumers through the following:

- **Ionate Intelligent Crawler**, based on AI/ML patterns, automatically extracts relevant information from web properties by crawling customer-facing services or systems in such a way that simulates human behavior. Users can view statistical insights on the crawls that are being executed, as well as historical views, of all systems, data producers and consumers. Our high-performance system's task executor (based on the IONATE Super Compute Platform) and parallelization framework overcome the usual challenges of system design, I/O and network efficiency, and robustness that usually beleaguer crawlers.

- **Ionate Spider**, similar to the web crawler, extracts relevant information from non-web properties such as RDMBS (support for all known databases - Oracle, MySQL, Postgre, MS SQL, Sybase, Informix, DB2, etc.), NoSQL data sources, file servers (NFS, FTP, Samba, etc.), email servers and all MS and Acrobat documents available on file servers or emails.

- **Ionate Agent** can scan for documents on a Desktop System. The behavior is similar to that of the Spider. The agent can be manually installed or a Desktop management system can be used for pushing updates.

2. Normalize

The IONATE Intelligent Data Adapter provides the versatility and flexibility required to analyze and explore any and all data types and create algorithms that can evolve as needed. The ability to transform all data varieties ensures that our solution covers all of your distributed environment.

We transform or normalize, all data sources, even those that are unknown, through our Intelligent Data Adapter. The storage of raw, unaltered data enables flexibility in analysis and exploration of data, and also allows queries and algorithms to evolve based on both historical and current data. For example, at the time new data is ingested in a platform, the solution may provide a suggestion such as “this looks like customer data; the last time similar data was received these transformations were applied, these fields masked, and these data lifecycle policy was set up.” The highly configurable adapter defines the metrics (or KPIs) that are found, while also allowing for manual intervention to modify the KPIs (i.e., providing specific clustering direction as well as instruction to save and name or discard metrics).

3. Train

To understand how metrics are defined, a quick discussion of tracing is needed. Tracing is key to understanding how requests flow, and as such reveals details regarding the availability and the response time of service. This can help find sources of latency (which are harmful in a distributed environment) as well as errors. By tracing a request as it travels from one service to another and tracking the duration of important operations within each service, a picture emerges of any performance issues or bottlenecks that can impact a request within a distributed system.

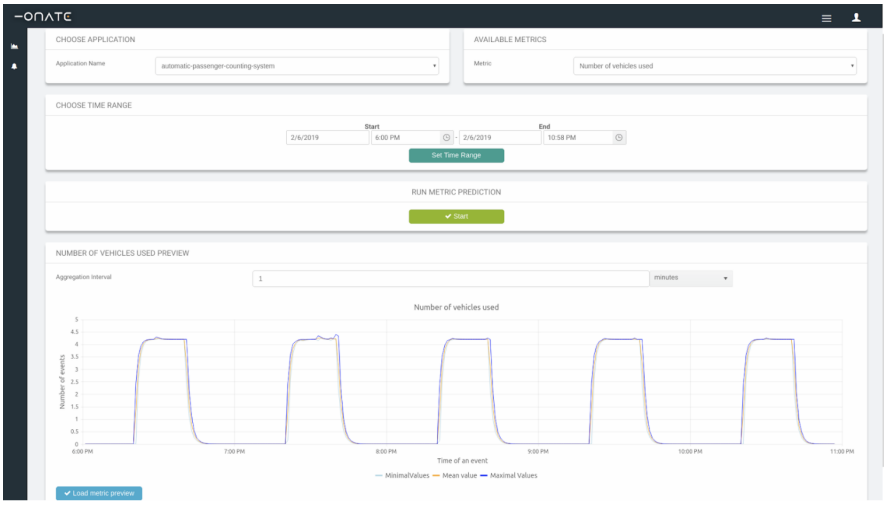
By continually aggregating tracing records, insights into statistical norms of service availability, and response time emerge. Our highly sophisticated and configurable training algorithm uses these statistical norms to establish metrics, or KPIs (as depicted in “Five Steps to Training”). With our algorithm’s autonomous data quality and validation check helping learn data’s expected behavior, thousands of validation checks are automatically created without the need for coding to update these checks over time. This auto-discovery of KPIs automatically determines the parameters of clustering (for example, categorizing the potential keywords).

Five Steps to Training

Step 1: Prep logs of expected behavior in an appropriate format and required fields
Step 2: Determine keywords, key logs, and wherein the application these should reside; run job; assess results. Accepted results equal the metric for that cluster
Step 3: Train to the established metric:
Load metric, Adjust bucket size, Start training, Finish Training
Step 4: Prepare logs that detect anomalous behavior, choose relevant metric, launch detection
Step 5: Compare real values with predictions, and determine the probability that real value is legitimate; (through the Trace IDs correlated to a chosen anomaly, users can immediately investigate the application logs in question)

Metrics are highly configurable, with each correlating to a purposefully-specified aspect of a portion of an application (or service or network and router). To learn expected behavior, normal data parameters are gleaned by continually aggregating all data that pertain to how requests journey through a microservices environment. Metrics are further tailored to a customer's exact needs and environment using our highly sophisticated and configurable algorithm. Our solution identifies key metrics on its own, and also allows for users to further refine a metric as needed.

4. Predict baseline

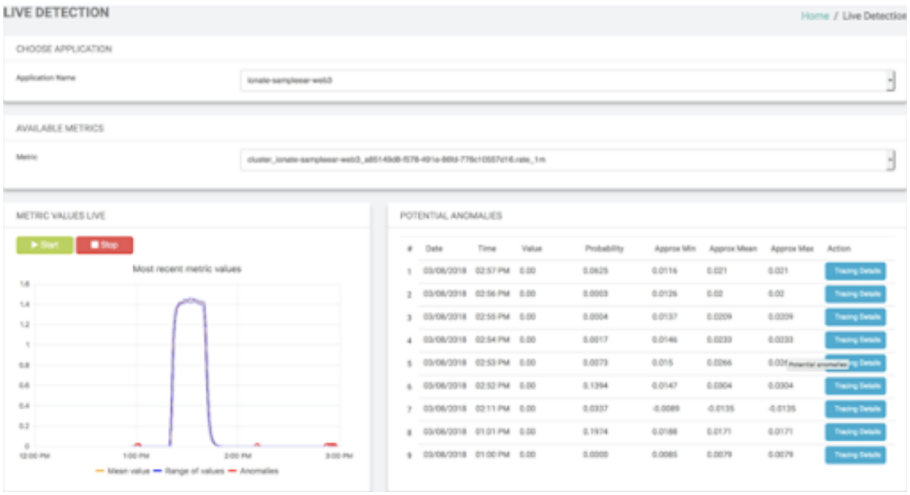


Running Metric Predictions. The above shows the simplicity of predicting the baseline activity for a passenger-counting service's metric called "number of vehicles used". The user sets the time range and then runs the prediction.

Once the expected normal, or predicted baseline, is established, the system auto-investigates any outliers (or abnormal measure). Outliers are unexpected bursts in an activity, especially in the context of an attack and network intrusion detection. However, as an outlier isn't necessarily always dangerous, customers define the conditions that trigger an alert, as well as the actions to take. Example actions include alerts to instances that deviate from the predictions (for example, email notifications or the auto-creation of Support Desk tickets).

5. Activate

Once normal-range KPIs are set, the system is activated and monitoring begins. Complex environments require monitoring that delivers the simplicity and clarity of a high-level system overview but also lets users drill deeply to troubleshoot specific transactions and processes. Our solution provides these views (high-level stats for monitoring, and lower-level stats for troubleshooting), saving all monitored in a time-series database, which enables comparisons to historical values. Once monitoring is activated, monitoring can be viewed by metric, in real-time, or by batch (which signifies events captured in a time series).

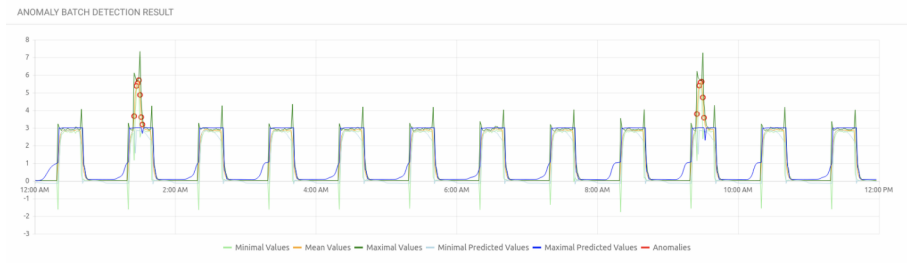


Live Detection. In the above screenshot, one metric is monitored live for a chosen application, service or element in a distributed environment. Our solution flags data that does not conform to the expected baseline as a potential anomaly, and through our proprietary algorithm determines and displays a probability that each is a true outlier.

Monitored data is saved in a time-series database, so it's possible to compare historical values. All logs produced by containerized apps are aggregated. This happens in a near real-time manner. Those logs can then be searched and filtered by service, app, host, datacenter, or other criteria to track and investigate curious behavior across aggregated logs.

6. Adapt

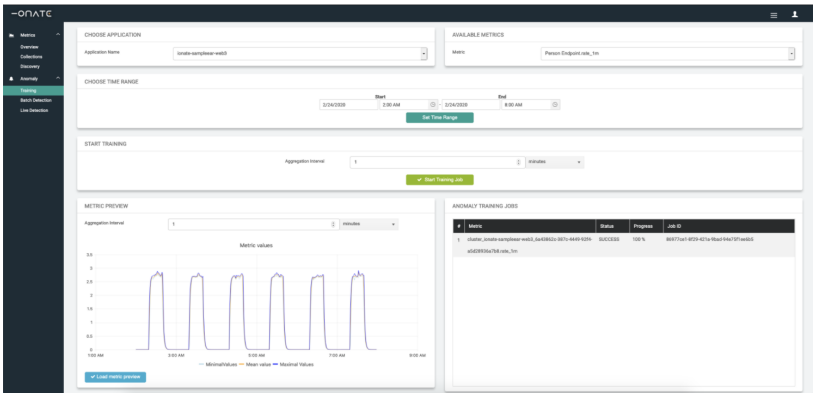
Events or observations that differ significantly from the majority of the data, thereby qualifying as suspicious, are flagged as potential vulnerabilities for immediate diagnosis and resolution. The IONATE Neural Network Analyzer and Classifier, an automatic AI/ML network, predictively classifies data as Non-Compliant or Compliant based on the metrics and KPIs established.



Anomaly Detection in Real-Time. During live detection, users click potential anomalies (red dots) and choose a metric to investigate. In the above real-time metrics, the light blue line indicates the predicted minimal baseline; darker blue indicates the maximum predicted baseline. Minimal and maximal values, as well as mean value, also display. With the red-dot anomalies diverging so clearly from the pre-set ranges, the solution automatically investigates outliers. If determined that a red dot does not signify an anomaly, the user can select that particular data range to flatten the event, 'suppress the anomaly'. This real-time retraining ensures such events are no longer noted as a false positive.

To train to an established metric, use the steps below (each noted in red below):

1. Choose one of the applications deployed on the cluster
2. Choose one of the metrics which is produced by above-chosen application
3. Choose a time-range of data to be used for training
4. Optional: Preview the metric, if desired
5. Start training
6. Watch Status of Training Job



When our proprietary neural network determines whether a potential anomaly is truly an outlier, or if it is a false positive, our solution then intelligently adapts. False positives are 'flattened' or 'suppressed' going forward by classifying the behavior as normal, thereby incorporating it into the expected baseline going forward. True outliers will trigger the associated pre-determined alert or action (such as opening a help desk ticket to investigate the issue).

Summary

For true cybersecurity in today's distributed environments, all data sources, services, and interrelationships across an environment must be found, monitored and analyzed, continually. AI/ML systems, when provided with proper amounts of training and data, will not only detect threats in real-time but also can further adapt to develop offensive and defensive playbooks. With responses occurring at the speed of today's attacks, the necessary actions can be coordinated across an enterprise's entire distributed network. Finally, cybersecurity that's smart enough to adapt as quickly, and often, and predictively as attackers.