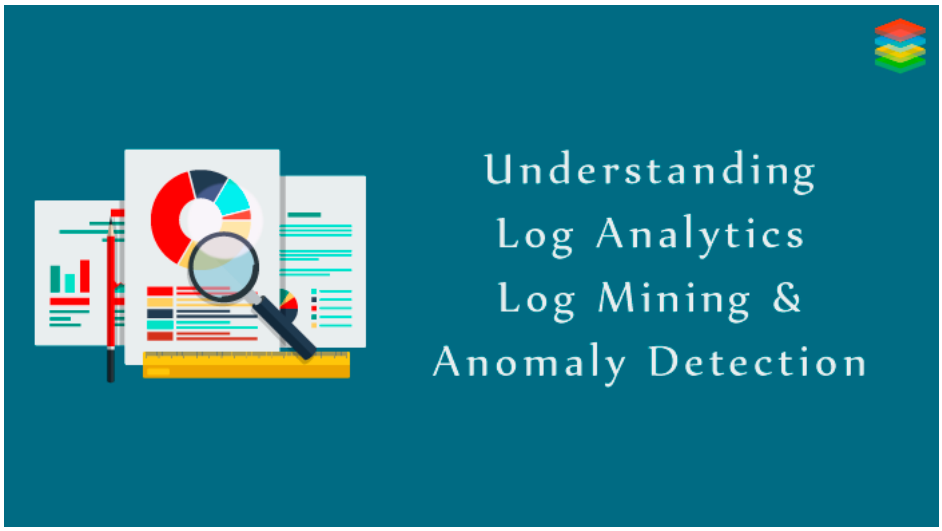
XENONSTACK
A Stack Innovator

Understanding Log Analytics, Log Mining & Anomaly Detection

by **Jagreet** | April 07, 2017 | Categories - Log Analytics, (/blog/category/Log-Analytics) Anomaly Detection, (/blog/category/Anomaly-Detection) Deep Learning, (/blog/category/Deep-Learning) Machine Learning (/blog/category/Machine-Learning)



What is Log Analytics

With technologies such as Machine Learning and Deep Neural Networks (DNN), these technologies employ next generation server infrastructure that spans immense Windows and Linux cluster environments.

Additionally, for DNNs, these application stacks don't only involve traditional system resources (CPUs, Memory), but also graphic processing units (GPUs).

With a non-traditional infrastructure environment, the Microsoft Research Operations team needed a highly flexible, scalable, and Windows and Linux compatible service to troubleshoot and determine root causes across the full stack.

Log Analytics supports log search through billions of records, Real-Time Analytics Stack (https://www.xenonstack.com/blog/enabling-real-time-analytics-for-iot) metric collection, and rich custom visualizations across numerous sources. These out of the box features paired with the flexibility of available data sources made Log Analytics a great option to produce visibility & insights by correlating across DNN clusters & components.

The relevance of log file can differ from one person to another. It may be possible that the specific log data can be beneficial for one user but irrelevant for the another user. Therefore, the useful log data can be lost inside the large cluster. Therefore, the analysis of the log file is an important aspect these days.

With the management of real-time data, the user can use the log file for making decisions.

But, as the volume of data increases let's say to gigabytes then, it becomes impossible for the traditional methods to analyze such a huge log file and determine the valid data. By ignoring the log data a huge gap of relevant information will be created.

So, the solution for this problem is to use Deep Learning Neural Network as a training classifier for the log data. With this, it's not required to read the whole log file data by the human being. By combining the useful log data with the Deep Learning it becomes possible to gain the relevant optimum performance and comprehensive operational visibility.

Along with the analysis of log data, there is also need to classify the log file into relevant and irrelevant data. With this approach, time and performance effort could be saved and close to accurate results could be obtained.

Understanding Log Data

Before discussing the analysis of log file first we should understand about the log file. The log is a data that produces automatically by the system and stores the information about the events that are taking place inside the operating system. It stores the data at every period of time.

The log data can be presented in the form of pivot table or file. In log file or table, the records are arranged according to the time. Every software applications and systems produce log files. Some of the examples of log files are transaction log file, event log file, audit log file, server logs, etc.

Logs are usually application specific, therefore, log analysis is a much-needed task to extract the valuable information from the log file.

Log Name	Log Data Source	Information within the Log Data
Transaction Log	Database Management System	It consists of information about the uncommitted transactions, changes made by the rollback transactions and the changes that are not updated in the database. This is performed to retain the ACID (Atomicity, Consistency, Isolation, Durability) property at the time of crashes
Message Log	Internet Relay Chat (IRC) and Instant Messaging (IM)	In the case of IRC, it consists of server messages during the time interval the user is being connected to the channel. On the other hand, to enable the privacy of the user IM allows storing the messages in encrypted form as a message log. These logs require a password to decrypt and view.
Syslog	Network Devices such as web servers, routers, switches, printers, etc.	Syslog messages provide the information on the basis of where, when and why i.e. IP-Address, Timestamp and the log message. It contains two bits: facility (source of the message) and security (degree of the importance of the log message)
Server Log File	Web Servers	It is created automatically and contains the information about the user in the form of three stages such as IP-Address of the remote server, timestamp and the document requested by the user
Audit Logs	Hadoop Distributed File System (HDFS) ann Apache Spark.	It will record all the HDFS access activities taking place with the Hadoop platform
Daemon Logs	Docker	It provides details about the interaction between containers, Docker service, and the host machine. By combining these interactions, the cycle of the containers and disruption within the Docker service could be identified.
Pods	Kubernetes	It is a collection of containers that share resources such a single IP_Address and shared volumes.
Amazon CloudWatch Logs	Amazon Web Services (AWS)	It is used to monitor the applications and systems using log data i.e. examine the errors with the application and system. It also used for storage and accessing the log data of the system.

Swift Logs	Openstack	These logs are sent to Syslog and managed by log level. They are used for monitoring the cluster, auditing records, extracting robust information about the server and much more.

Log Analysis Process

The steps for the processing of Log Analysis are described below:

- Collection and Cleaning of data
- Structuring of Data
- Analysis of Data

Collection and Cleaning of data

Firstly, Log data is collected from various sources. The collected information should be precise and informative as the type of collected data can affect the performance. Therefore, information should be collected from real users. Each type of Log contains distinguish the type of information.

After the collection of data, the data is represented in the form of Relational Database Management System (RDMS). Each record is assigned a unique primary key and Entity-Relationship model is developed to interpret the conceptual schema of the data.

Once the log data is arranged in proper manner then, the process of cleaning of data has to be performed. This is because there can be the possibility of the presence of corrupted log data.

The reasons of corruption of log data are given below:

- Crashing of disk where log data is stored
- Applications are terminated abnormally
- Disturbance in the configuration of input/output
- Presence of virus in the system and much more

Structuring of Data

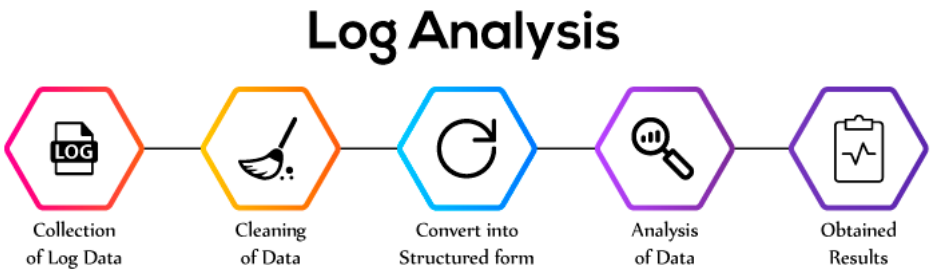
Log data is large as well as complex. Therefore, the presentation of log data directly affects their ability to correlate with the other data.

An important aspect is that the log data has the ability to directly correlate with the other log data so that deep understanding of the log data can be interpreted by the team members.

The steps implemented for the structuring of log data are given below:

- Clarity about the usage of collected log data
- Same assets involve across the data so that values of log data are consistent. This means that naming conventions can be used
- Correlation between the objects is created automatically due to the presence of nested files in the log data. It's better to avoid nested files from the log data.

Analysis of Data: Now, the next step is to analyze the structured form of log data. This can be performed by various methods such as Pattern Recognition, Normalization, Classification using Machine Learning, Correlation Analysis and much more.



Importance of Log Analysis

Indexing and crawling are two important aspects. If the content does not include indexing and crawling, then update of data will not occur properly within time and the chance of duplicates values will be increased.

But, with the use of log analytics, it will be possible to examine the issues of crawling and indexing of data. This can be performed by examining the time taken by Google to crawl the data and at what location Google is spending large time.

In the case of large websites, it becomes difficult for the team to maintain the record of changes that are made on the website. With the use of log analysis, updated changes can be maintained in the regular period of time thus helps to determine the quality of the website.

In Business point of view, frequent crawling of the website by the Google is an important aspect as it point towards the value of the product or services. Log analytics make it possible to examine how often Google views the page site.

The changes that are made in the page site should be updated quickly at that time in order to maintain the freshness of the content. This can also be determined by the log analysis.

Acquiring the real informative data automatically and measuring the level of security within the system.

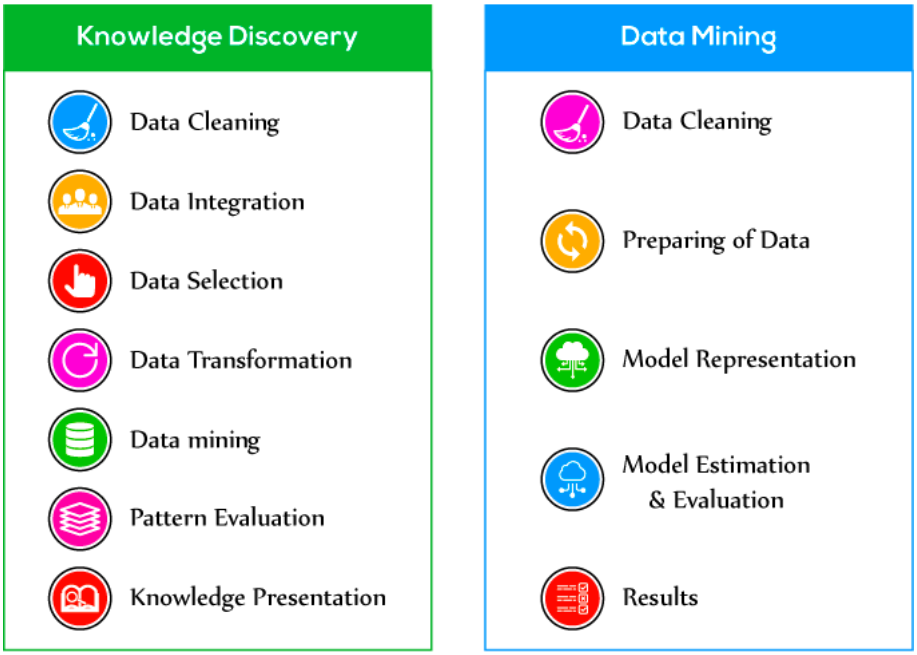
Knowledge Discovery and Data Mining

In today's generation, the volume of data is increasing day by day. Because of these circumstances, there is a great need to extract useful information from large data that are further use for making decisions. Knowledge discovery and Data Mining are used to solve this problem.

Knowledge discovery and Data mining ate two distinct terms. Knowledge Discovery is a kind of process used for extracting the useful information from the database and Data Mining is one of the steps involved in this process. Data Mining is the algorithm used for extracting the patterns from the data.

Knowledge Discovery involves various steps such as Data Cleaning, Data Integration, Data Selection, Data Transformation, Data Mining, Pattern Evaluation, Knowledge Presentation. Knowledge Discovery is a process that has total focus on deriving the useful information from the database, interpretation of storage mechanism of data, implementation of optimum algorithms and visualization of results. This process gives more importance on finding the understandable patterns of data that further used for grasping useful information.

Data Mining involves the extraction of patterns and fitting of the model. The concept behind the fitting of the model is to ensure what type of information is inferred from the processing of model. It works on three aspects such as model representation, model estimation, and search. Some of the common Data Mining techniques are Classification, Regression, and Clustering.



Log Mining

After performing analysis of logs, now next step is to perform log mining. Log Mining is a technique that uses Data Mining for the analysis of logs.

With the introduction of Data Mining technique for log analysis the quality of analysis of log data increases. In this way analytics approach moves towards software and automated analytic systems.

But, there are few challenges to perform log analysis using data mining. These are:

- Day by day volume of log data is increasing from megabytes to gigabytes or even petabytes. Therefore, there is a need of advanced tools for log analysis.
- The essential information is missing from the log data. So, more efforts are needed to extract useful data.
- The different number of logs are analyzed from different sources to move deep into the knowledge. So, logs in different formats have to be analyzed.
- The presence of different logs creates the problem of redundancy of data without any identification. This leads to the problem of synchronization between the sources of log data.

As shown in fig the process of log mining consist of three phases. Firstly, the log data is collected from various sources like Syslog, Message log, etc. After collecting the log data, it is aggregated together using Log Collector. After aggregation second phase is started.

In this, data cleaning is performed by removing the irrelevant data or corrupted data that can affect the accuracy of the process. After cleaning, log data is represented in the structured form of data (Integrated form) so that queries could be executed on them.

After that, the transformation process is performed to convert into the required format for performing normalization and pattern analysis. Useful patterns are obtained by performing Pattern Analysis. Various data mining techniques are used such as Association rules, Clustering etc to grasp the useful information from the patterns. This information is used for decision-making and for alerting the unusual behavior of the pattern by the organization.

Define Anomaly

An anomaly is defined as the unusual behavior or pattern of the data. This unusual indicates the presence of the error in the system. It describes that the actual result is different from the obtained result, thus the applied model does not fit into the given assumptions.

The anomaly is further divided into three categories described below:

- **Point Anomalies**
  A single instance of a point is considered as an anomaly when it is farthest from the rest of the data.
- **Contextual Anomalies**
  This type of anomaly related to the abnormal behavior of the particular type of context within data. It is commonly observed in time series problems.
- **Collective anomalies**
  When the collected instance of data is help for detecting anomalies is considered as collective anomalies.

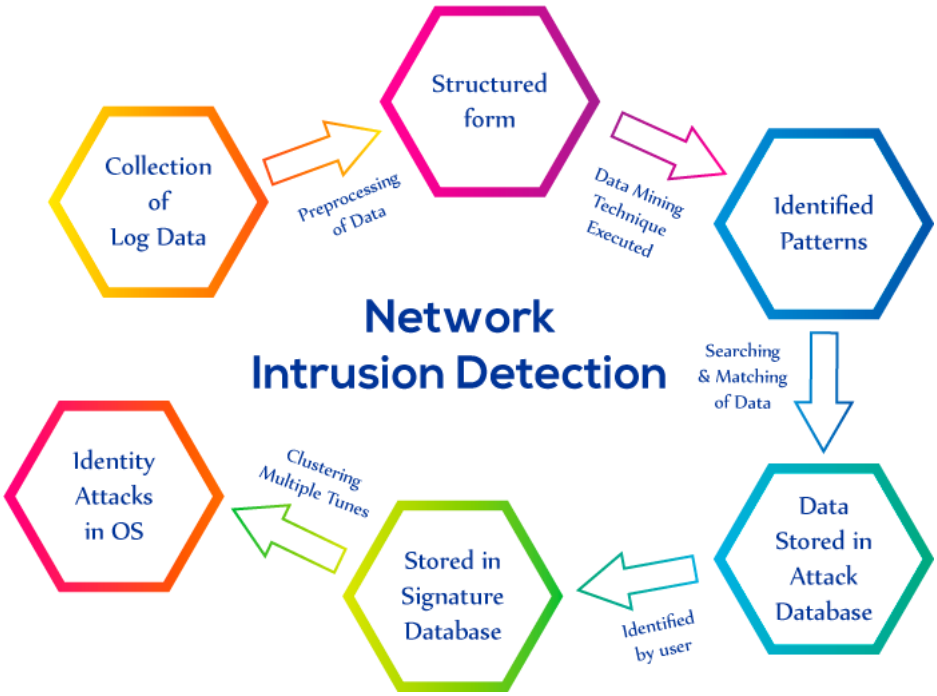
The system produces logs which contain the information about the state of the system. By analyzing the log data anomalies can be detected so that security of the system could be protected. This can be performed by using Data Mining Techniques. This is because there is a need of usage of dynamic rules along with the data mining approach.

Network Intrusion Detection using Data Mining

In today's generation, the use of computers has been increased. Due to this, the probability of cyber crime has also increased. Therefore, a system is developed known as Network Intrusion Detection which enables the security in the computer system.

In this system, Data Mining Techniques and the signature database are used. The description of the process is given below:

- Firstly, the log files are collected from all the sources.
- Pre-processing of log files is performed i.e. the log data is represented in a structured form.
- After that, Data Mining Techniques such as Support Vector Machine (SVM), Random Forest, etc are applied to the log data to identify the patterns.
- The log data is searched in the normal log database and the attack log database.
- If the pattern is not matched with the normal log database it will be classified as an attack log data pattern.
- From the identified collected patterns unusual patterns as an attack are identified by the user.
- After the identification of unusual patterns, the attack patterns are stored in the signature database (attack log database).
- If the pattern is already present in the signature database, an alert will be given by the system.
- At the end, clustering is performed multiple times to identify the security attack with the operating system.



Fraud Detection in Banking

Banks are the organizations for depositing and withdrawing money, getting the provision of loans. This facility is available to all therefore, the proper security mechanism should be introduced. Some points are necessary for consideration before performing fraud detection that is mentioned below:

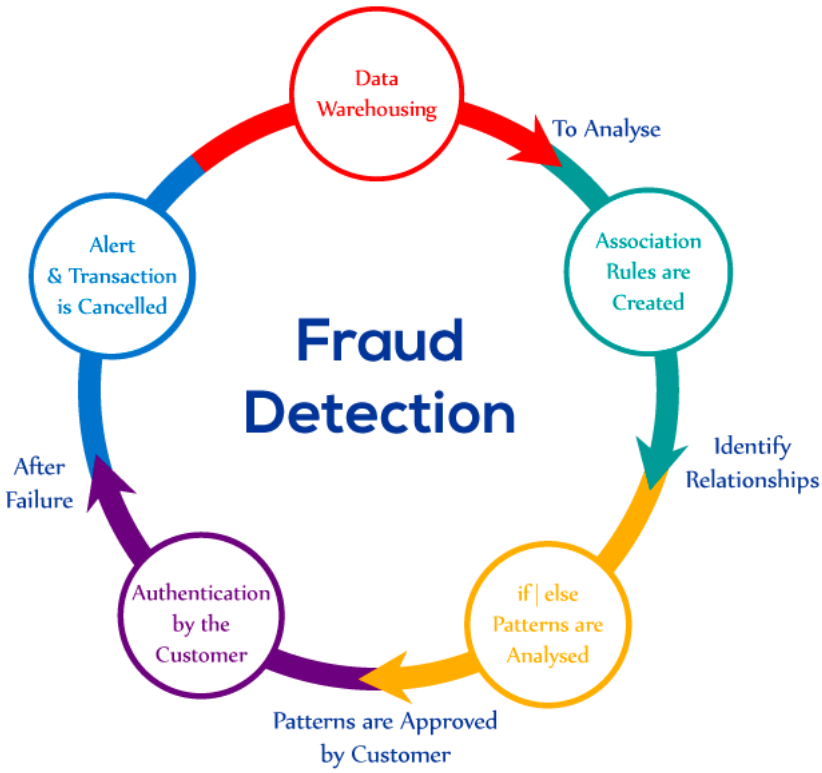
- Fraudsters have analyzed the whole procedure of bank.
- They even are expert in copying the signature of the customer without any doubt.

Firstly, the banks have stored the previous information about each customer in their database. This storage of data in the database is known as Data Warehousing. Next approach is to analyze the data. After that, Association rules are created in the form of if/then patterns.

Support and confidence field is used to identify the relationships between the data. Support field contains the information about the frequent occurrences of data within the database. Confidence field provides the information about a number of times the if/then statements are found to be true.

For example, while analyzing the database of the bank association rules are made by a customer. A customer named as Nilu Sharma does not withdraw money more than 1 lakh and transactions are frequently occur after 2 months. Here, the limitation in withdrawing of money within 1 lakh is supported field and occurrence of the transaction after 2 months is confidence field.

Therefore, any transaction of Nilu Sharma more than 1 lakh will be examined and further authentication is performed. After the failure of authentication, an alert will be created and the transaction will be canceled by the bank.



Fraud Detection with Deep Learning Neural Network

Banks have to analyze millions of money transactions in a day. But, due to lack of advanced techniques banks are not able to examine transactions properly as it becomes difficult to examine few fraud activities within million of transactions.

Therefore, the scalable technique is needed which update the system automatically. Deep Learning Neural Network can be used that can detect fraud activities automatically and the system can learn automatically whenever the new data will arrive without the interference of human being.

Larger institutions and organizations indulge in large financial transactions. So, open-source-deep learning is introduced for them so that they can fight for fraud activities at economical rates by using sky mind with deep learning neural network.

Deeplearning4j is an open source deep learning library that uses distributed deep learning by integrating with Hadoop and Spark. This library not only detects frauds, anomalies and patterns in real time rather it also learns from the new data parallelly.

Real Time Automated Network Security

The security of the network should be maintained automatically by the machine. But, if it is performed by the human being then, it will become unfavorable for the organization. Therefore, optimum usage of automation helps in controlling the network security.

Apache Open Source Projects Started For Real-time Log Analytics and Network Security

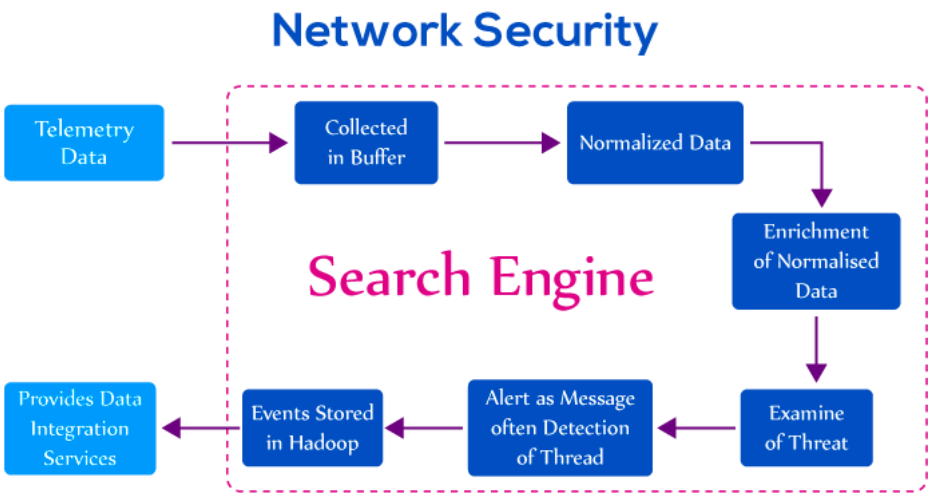
- Apache Metron
- Apache Eagle
- Apache Spot
- Apache Ranger
- Aqua Docker Security Platform

Automation works in Four aspects

- Finding the patterns of attackers globally,
- Developing and implementing the protection layer faster than the attackers

- Determining issues
- Failures within the network.

The infrastructure used for developing automated network security is described below:



The system provides various services such as alerting of threat event in real time, Feature Engineering, and better data integration layer.

Firstly, all the relevant data for analyzing is taken into account and collected by the buffer of the corresponding system. When the data is ingested into the buffer the engine of the system starts working. After buffering the data, data is normalized in the standard format so that messages from different topology can correlate with the data.

After that quality of normalized data is improved. For example, the ip_address attribute can be improved by providing detail information about host_id details. Then, information is retrieved from the enriched data to examine the threat events.

Whenever the threat event is detected, a message will be displayed as an alert. This means that labeling of threat events are performed in the form of messages.

All these processes are performed by the system engine in real time. After alerting of threat event all the labeled events are stored in Hadoop for long-term storage and further usage for the next generation analysis process.

Summary

XenonStack Data Science Solutions provides a Platform for Data Scientists and Researchers to Build, Deploy Machine Learning and Deep Learning Algorithms at a scale with automated On-Premises and Hybrid Cloud Infrastructure.

Get in Touch with us for Proof of Concept, Consulting & Building Data & AI Products. Talk to our Data Scientist (https://www.xenonstack.com/contact-for-data-science-solutions) for assessment and consulting for your Industry/Solution/Products.

XenonStack Offerings

XenonStack is a leading Software Company in Product Development (https://www.xenonstack.com/products?utm_source=blog) and Solution Provider for DevOps (https://www.xenonstack.com/devops-consulting-services?utm_source=blog), Big Data Integration (https://www.xenonstack.com/big-data-services?utm_source=blog), Real Time Analytics

(https://www.xenonstack.com/streaming-real-time-analytics-solutions?utm_source=blog) & Data Science (https://www.xenonstack.com/data-science-machine-deep-learning-services?utm_source=blog).

Product NexaStack - Unified DevOps Platform (https://www.xenonstack.com/unified-devops-delivery-platform-as-a-service?utm_source=blog) Provides monitoring of Kubernetes, Docker, OpenStack infrastructure, Big Data Infrastructure and uses advanced machine learning techniques for Log Mining and Log Analytics (https://www.xenonstack.com/log-analytics?utm_source=blog).

Product ElixirData - Modern Data Integration Platform (https://www.xenonstack.com/big-data-integration-platform?utm_source=blog) Enables enterprises and Different agencies for Log Analytics and Log Mining.

Product Akira.AI (https://www.xenonstack.com/akira-ai-knowledge-driven-platform?utm_source=blog) is an Automated & Knowledge Drive Artificial Intelligence Platform that enables you to automate the Infrastructure to train and deploy Deep Learning Models on Public Cloud (https://www.xenonstack.com/cloud-hybrid-infrastructure-solutions?utm_source=blog) as well as On-Premises.

Get 1 Hour Free Assessment for DevOps, Big Data Strategy, and Data Science. CONTACT US NOW (https://www.xenonstack.com/talk-with-our-experts?utm_source=blog)

Share Post On Social Media

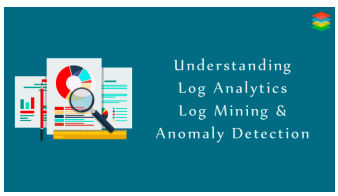
(http://www.facebook.com/share.php?u=https://xenonstack.com/blog/understanding-log-analytics-log-mining-anomaly-detection&title=Understanding Log Analytics, Log Mining & Anomaly Detection)

(http://twitter.com/home?status=Understanding Log Analytics, Log Mining & Anomaly Detection+https://xenonstack.com/blog/understanding-log-analytics-log-mining-anomaly-detection)

(http://www.linkedin.com/shareArticle?mini=true&url=https://xenonstack.com/blog/understanding-log-analytics-log-mining-anomaly-detection&title=Understanding Log Analytics, Log Mining & Anomaly Detection&source=understanding-log-analytics-log-mining-anomaly-detection)

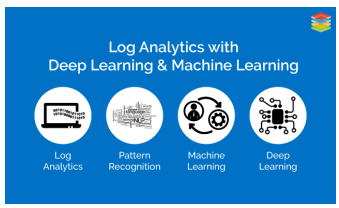
(https://plus.google.com/share?url=https://xenonstack.com/blog/understanding-log-analytics-log-mining-anomaly-detection)

Related Posts



Understanding Log Analytics, Log Mining & Anomaly Detection (/blog/understanding-log-analytics-log-mining-anomaly-detection)

April 07, 2017

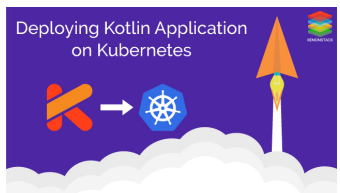


Log Analytics With Deep Learning And Machine Learning (/blog/log-analytics-with-deep-learning-and-machine-learning)

April 28, 2017

Overview of Kotlin & Comparison With Java (/blog/overview-of-kotlin-comparison-between-kotlin-java)

May 17, 2017



Deploying Kotlin Application on Docker & Kubernetes (/blog/deploying-kotlin-application-on-docker-kubernetes)

May 19, 2017

Semantic Domain C Apache S (/blog/se based-or using-apa

May 26, 201

Leave a message