

TrafficWiz

Machine Learning
Traffic Analytics Tool



Network Traffic – The Modern Landscape

The landscape of Internet traffic has changed tremendously in recent years. One aspect of this change relates to encrypted traffic. The desire of individuals, organizations, corporations and Government to protect and safeguard their data has led many applications to encrypt the data that they generate and transmit over the Internet. The scope, volume and level of encrypted traffic is rapidly rising.

Recent reports from different sources such as Sandvine indicate that the level of encryption on the Internet has more than doubled in the last 5 years - in 2015, 30% of Internet traffic was encrypted while in 2020, 70-80% of Internet Traffic is now encrypted according to estimates from organizations such as Gartner.

This encryption is increasingly applied to voice, video, data and other types of traffic. It applies to many applications which transmit data over the Internet including e-commerce and banking applications, social media, text chat such as Whatsapp, VoIP such as Skype and even video streaming among others. For example, Netflix, a leading source of video traffic in North America offers encrypted video services. Already, 70% of mobile phone/device traffic is encrypted and with the deployment of 5G wireless networks, this figure is set to rise.

While increased levels of data encryption is a positive development enabling data privacy, it poses certain challenges for enterprise IT organizations, communication service providers responsible for managing and securing the network. It also poses a challenge for Law Enforcement Agencies (LEA) and Intelligence organizations.

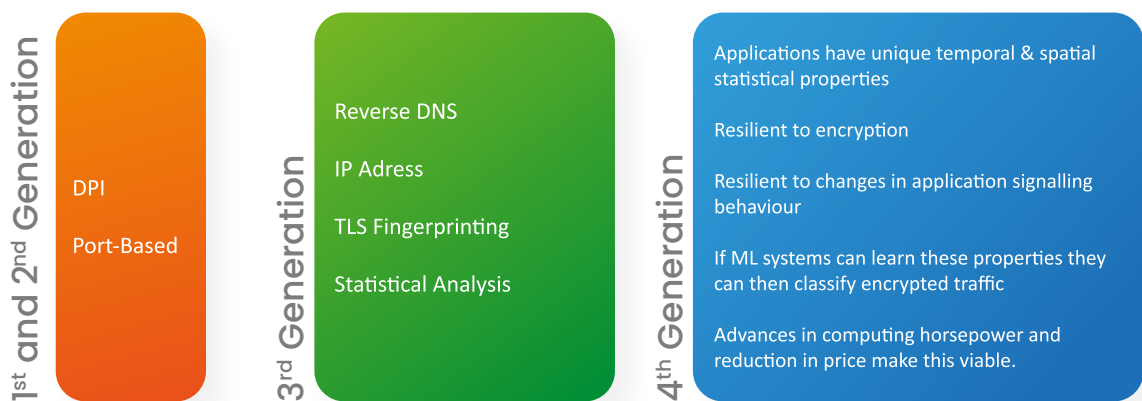
In particular, for an IT organization to manage and secure the network it needs to know the underlying application which generated a particular traffic flow. The ability to determine the type of application is governed by a set of approaches referred to as Traffic Classification or Packet Classification. More specifically, Traffic Classification is carried out by monitoring network traffic at any point in the network and reverse engineering the underlying application or class of traffic.

The increasing amount of encrypted traffic poses a challenge for existing LEA and IT network traffic probes, network monitoring tools, routers, firewalls and network traffic analysis tools. These tools all rely on traffic classification modules to identify traffic accurately and then carry out certain actions. When network traffic is encrypted, existing traffic classification methods such as DPI (Deep Packet Inspection) are less accurate and effective as they fail to classify and filter such traffic.

Challenge of Traffic Classification & Encryption


Traditionally, methods of traffic classification involve examining packet headers and payloads directly for certain field values or heuristics that would reveal the underlying application which generated the traffic.

The 1st generation of traffic classification solutions in the nineties used a technique referred to as port-based classification. These techniques examined the TCP and UDP port numbers for server-side applications in order to classify traffic. Since the Internet Authority, IANA, assigned well-known server port numbers to different applications, it was possible to classify traffic on this basis. For example, Telnet application traffic was carried using Port 23 and SMTP email traffic on Port 25.



As applications become more sophisticated in the late nineties, they began to utilize dynamic port numbers which were negotiated or assigned after initial communication. This trend was further exacerbated with the emergence of P2P (peer-to-peer) applications such as BitTorrent, Gnutella and eDonkey. As a result port-based classification was no longer sufficient to deliver the classification accuracy required.

DPI (Deep Packet Inspection) and SPI (Stateful Packet Inspection) emerged in the early 2000s as 2nd Generation Traffic Classification solutions. They were developed as a direct response to the challenges faced by port-based classification. DPI examines packet headers as well as packet payload traffic to determine the underlying application or class of traffic. More specifically, it compares packet contents to application signatures or patterns in order to classify a traffic flow. In this regard, DPI addressed the challenges faced by port-based classification and achieved good classification accuracy.



The increasing levels of encryption resulted in the emergence of a third generation of classification solutions. The 3rd generation in combination with DPI could achieve good levels of accuracy when classifying traffic. Third generation solutions included a mixture of approaches relying on (a) Reverse DNS lookups (b) IP address lookups (c) Statistical Analysis (d) Fingerprinting of TLS communication exchanges.

As encryption became more widespread and encryption techniques more sophisticated even the 3rd generation of classification solutions were insufficient to classify traffic with the accuracy required. The ongoing and rapid deployment of TLS 1.3 is increasing the size of the blindspot. In addition, more and more applications simply ride over HTTPS making it difficult to distinguish whether an applications is a genuine web application or simply using it as an underlying protocol. Given the above, a number of studies have shown that DPI and associated techniques have difficulty achieving classification accuracy above 40%

In recent years, a 4th generation of classification solutions have been considered which can accurately classify traffic in the presence of encryption. These techniques utilize machine learning (ML) to address the problem of traffic classification in an encrypted environment. Initially studied in academic research labs, these solutions now have the maturity to transition into vendor product solutions.

Machine Learning for Encrypted Traffic Classification


In recent years, advances in computing horsepower and memory, emergence of specialized GPUs and development of a broad ecosystem of open source libraries have enabled machine learning (ML) and artificial intelligence to be applied in a variety of new domains. Where previously, ML solutions were narrowly confined to a handful of practical applications they are now broadly utilized in areas such as finance, banking, investments, medicine, image recognition among others. The maturity and emergence of Deep Learning (DL) solutions in conjunction with cluster and cloud-based computing has further allowed the uptake of ML and DL.

For a number of years, researchers studied the viability of ML and DL in the context of Traffic Classification for encrypted network traffic. Early results were promising but robustness and wide-spread applicability was uncertain. Over time, the research advanced to the point where it could be considered for use in production-level classification solutions.



The premise behind ML-based Traffic Classification is that network traffic generated by applications have unique temporal & spatial statistical properties which are resilient to encryption and generally resilient to routine changes in application signaling protocols. In other words, the network behaviour of different applications are different. If ML systems can be trained to learn the network behaviour for each class of applications they can then be used to classify new network traffic based on those training models. Further, this classification will work accurately even if the traffic is encrypted. The classification would also work even if there are minor changes in underlying protocol signaling behaviour for an application. These statistical properties are reflected in statistical metrics extracted from the network traffic.

In ML parlance, these statistical metrics are referred to as features which can be defined and extracted from network traffic flows. There are literally hundreds of features which can be defined and tracked for a given traffic flow. Examples of such features include metrics associated with packet inter-arrival time (average, standard deviation etc), byte counts, packet counts, flow duration as well as more sophisticated metrics such as entropy.



Multiple steps are required in order to develop the ML model which can robustly encapsulate and represent the behaviour of network traffic. Rigour and innovation are required in order for the ML model to suitably classify network traffic with the accuracy required.

Creation of an ML model requires multiple steps. We do not outline all the detailed steps in this whitepaper but simply highlight a few key steps. ML model creation starts with the first step - creation of a well represented and accurate labeled. Once a labeled/annotated network traffic dataset is available, it can serve as the basis of the ML model used for traffic classification. The next step is to carry out feature extraction whereby define features/statistical metrics are computed for the labeled training dataset. Once this done, a supervised ML algorithm (there are many candidates) is trained on this dataset and a ML model created.

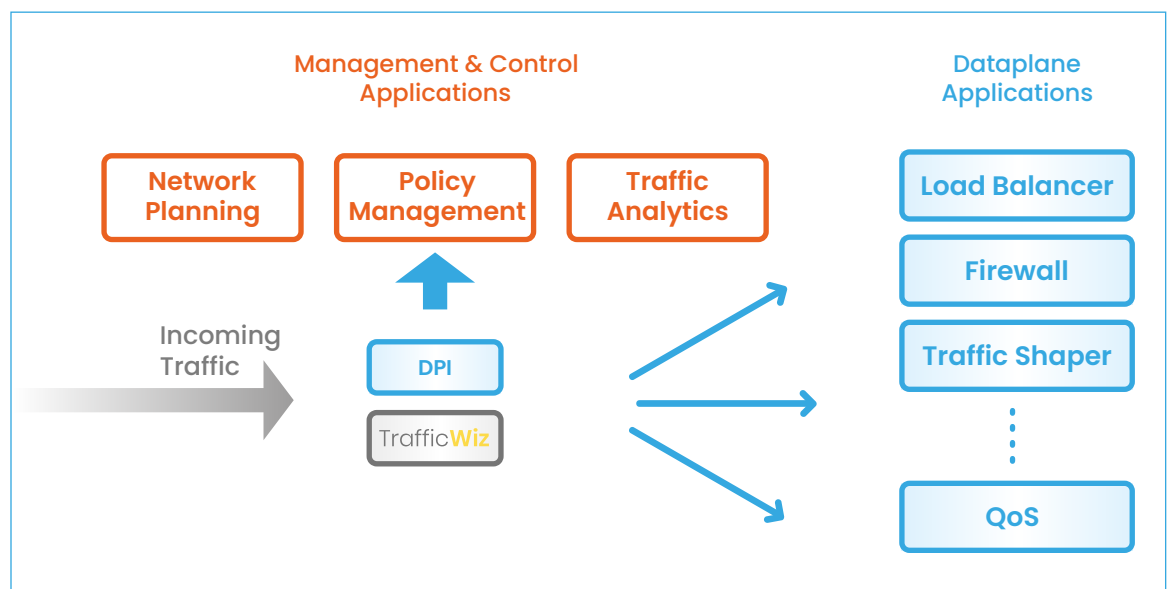
At every step of the way, there are many decisions to be made when designing the ML model. We highlight some of them here (i) Which ML algorithms to use (ii) Whether to use binary classification or multi-class classification (iii) Which features to use - feature computation can often be the most expensive element of the process in terms of compute time and resources (iv) What type of data transformations and data cleansing methods to use (v) What advanced data science techniques to utilize - there are many.

Solana Networks has spent a number of years working on the above problem, efforts which culminated in development of the TrafficWiz solution.

Applications

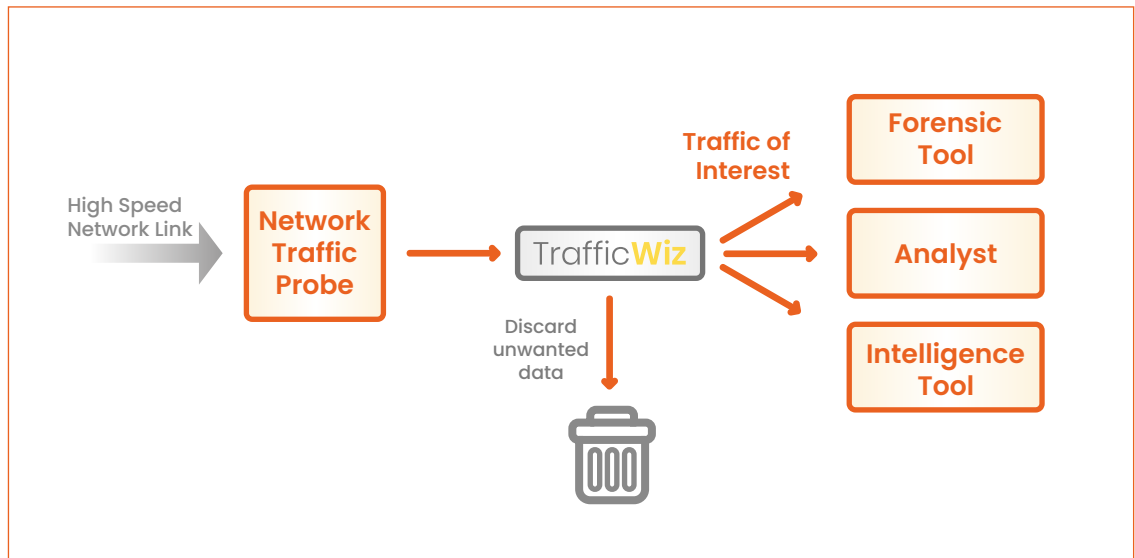
TrafficWiz and ML-based traffic classification can be applied to different use-cases. In the following, we consider two specific use-cases:

1. Network & IT Operations
2. Intelligence Agencies & Law Enforcement Agencies (LEA)



The Figure above illustrates a TrafficWiz usecase for network operators and IT operations. Traffic Classification is widely used in such environments:

- Network Planners need to know the mix of applications using the network to accurately plan and carry out forecasting for the network. Inaccurate classification and blindspots leads to networks which experience traffic congestion and poor performance
- Policy Management solutions need to know different classes of traffic using the network and their associated volumes. This is needed for both wired and wireless networks
- Network traffic analytics solutions need to know the application behind every traffic flow being monitored
- Firewalls need to accurately classify their traffic in order to apply policies to allow or block certain applications
- Traffic Shapers need to identify applications to determine which ones to throttle, shape or drop
- Load balancers need to identify traffic classes and applications in order to distribute traffic to different processing engines.

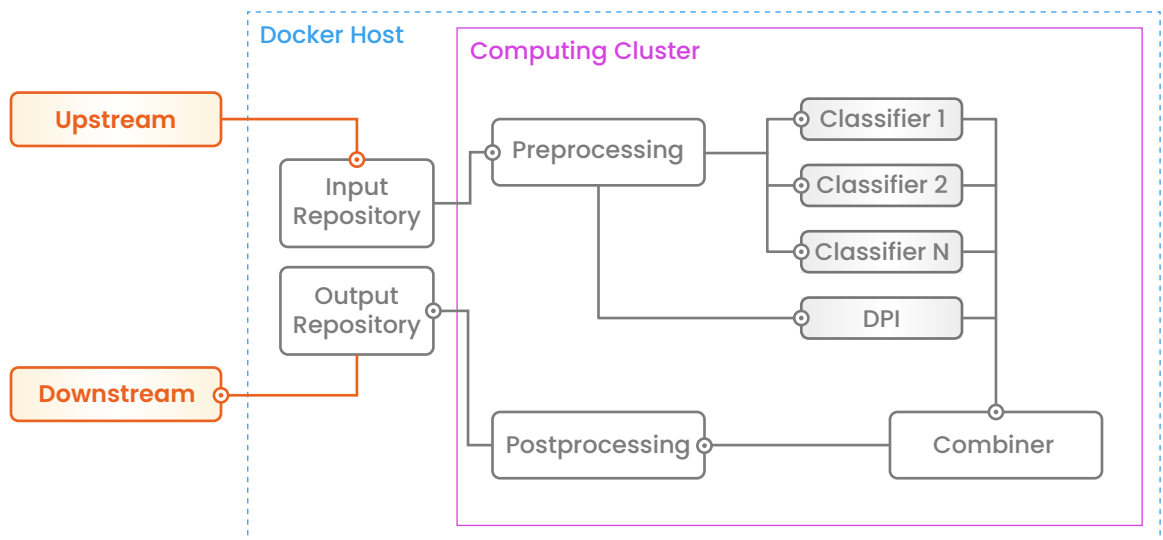


The Figure above illustrates a TrafficWiz usecase for Intelligence agencies and LEA. These organizations have requirements to gather network traffic data for post-analysis. However, the continual rise and huge surge in network traffic volumes poses a challenge to this endeavour. Traffic Classification is thus essential for their work with a number of key benefits:

- **Storage Space Reduction** - A key benefit of traffic classification for this usecase is to enable discarding of unwanted data. Thus for example the ability to identify video streaming accurately and discard it is crucial. Otherwise the sheer volume will result in significant additional storage space and cost
- **Analyst Time & Cost Savings** - In this usecase, analysts rely on custom and COTS tools to analyze network traffic for their work - an effort which relies on semi-automation, manual effort and human-in-the-loop intelligence. The ability to hone in on specific applications of interest will greatly expedite the time for analysts to carry out their work and also, thus, result in cost savings. The set of traffic flows can be streamlined to focus on the set which requires more detailed analysis.
- **Sharing of data** - in this usecase and environment, data is often stored and shared. Reducing the volume unwanted traffic will greatly facilitate the ability to share data across jurisdictions as required. Traffic/conversations of interest can be tracked and stored for future analysis.

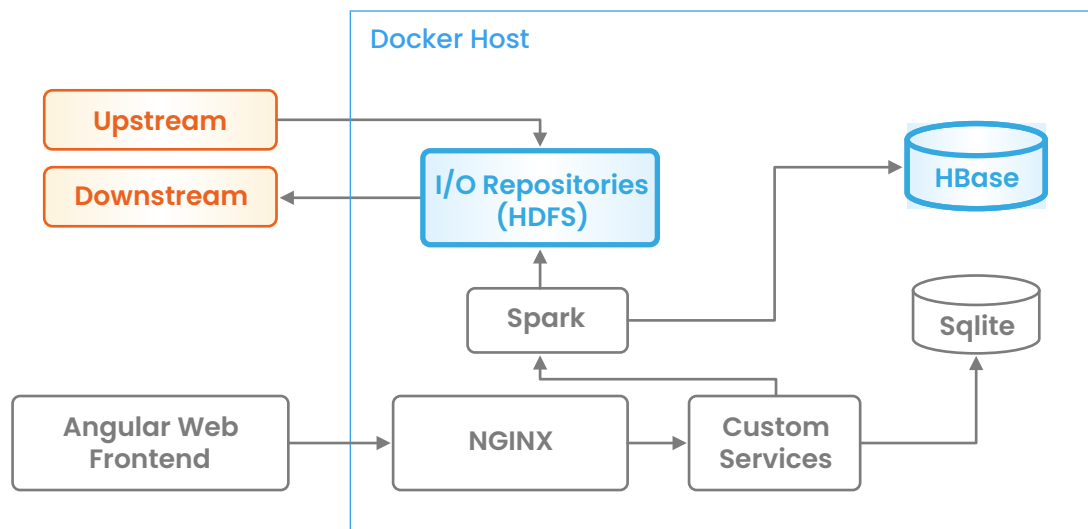
Use Case: TrafficWiz Big Data Implementation

The system is used both for training of ML traffic classification modules as well as for traffic classification. Well defined REST-based APIs exist to enable easy integration of the system with other third-party systems as part of a larger ecosystem of applications. It is suitable for deployment in offline solutions as well as part of other solutions such as a network probe.



The system is used for training ML models. The models are then stored for use in classification. The high level architecture illustrated can be described as follows:

- **Upstream:** Input traffic is received as a stream of PCAPs from network probes OR upstream network applications/devices.
- **I/O Repository:** Persistent storage is managed by a distributed file system
- **Pre-processing:** Convert packets to flows and extracts statistics/features for ML algorithms.
- **Classifiers:** Multiple ML algorithms can be configured and used for traffic classification. Example well known algorithms include SVM, Decision Trees, Logistic Regression etc.
- **Combiner:** The combiner fuses the result from multiple ML classifiers to improve classification accuracy
- **Downstream:** The output results are written to database and file. They can also be transmitted to downstream systems using the REST API.



The TrafficWiz big data platform is built on a number of technologies as illustrated in the Figure above. We summarize its key elements as follows:

- The system is built on an Apache Spark big data platform
- Docker is used to facilitate component integration and scalability
- Hbase is used as the non-relational database for managing large volumes of data and suitable for clustered deployment. Other databases such as Cassandra for example can be used instead of Hbase
- HDFS is used as the distributed file system for raw file storage
- A web front end is used for user and traffic model management, statistics and other configuration. The GUI is built using Angular JS and NGNX

The system was designed to scale to high data rates. The classification engine itself has been benchmarked to scale to tens of Gbps on a single machine.

Conclusion/Summary

Widespread deployment of encryption necessitates a 4th generation of traffic classification solutions, one which relies on machine learning (ML) to classify network traffic. Solana Networks has developed TrafficWiz as a ML-based traffic classification or packet classification solution for use by Law Enforcement Agencies (LEA), network operators and product vendors who require visibility into network traffic.

Contact us with any questions! We are happy to chat!