Forensics of Cloud Machine Learning Platforms

Nowadays, with most of our company's critical data moved to Cloud service providers, one of our major concerns is dealing with security matters. That includes being able to quickly respond to and report events that may lead to legal issues such as lawsuits and, in extreme instances, even involve law enforcement. This is by no means an easy task and matters are further complicated by the fact that we have to trust our Cloud provider's ability to deliver digital forensics data in the case of any legal dispute (either civil or criminal) during cyberattacks or even if a data breach occurs.

Well, maybe trusting provider capabilities is not actually the real problem, since a contract may include a hefty fine for such cases, but the fact remains that there is a huge difference between analyzing locally stored data versus doing it at a Cloud service. Traditional computer forensics deals with collecting media at the crime scene or at least the location where the media was seized, the efforts for the preservation of that media, and subsequent validation, analysis, interpretation, documentation, and courtroom presentation of the results of the examination. For most situations, other than an internal investigation, any evidence contained within the media will be controlled by law enforcement from the moment of seizure. Now, in a Cloud scenario, the information may be anywhere around the globe, even outside your country boundaries. This can turn controlling the evidence (i.e., collection, preservation and validation) into quite a challenge.

So, Where is forensic evidence most commonly found in the Cloud? The first step is to know exactly where your data is and how much direct access you have to the infrastructure supporting it. As we stated before, it is important to know what Cloud type and deployment option you are using. The lower down the Cloud stack your provider stops, the more direct control you have over data and evidence.

For instance, if you are using a private Cloud, it is more than likely that you have direct access to your hardware infrastructure and your Cloud forensics will not diverge too much from the usual digital forensics. On the other hand, if you are using a SaaS model over a public Cloud, initially direct evidence collection will be limited to whatever your provider offers in terms of logs or audit reports. Other

than that, it all falls under what is covered on your contract, so special attention should be paid to your service level (SLA). If your agreement is not clear on what level of forensics information your service provider is bound to make available, and also how soon they are required to do it, you may find yourself in a very bad situation.

Machine Learning (ML) is one of the approaches of AI that uses a system that can be learned by itself from experience. A system that can learn from experience and examples rather than from programming. Thus, if the system learns constantly and makes a decision based on the data rather than programming, it could really be effective in digital forensics to map the crimes easily & smartly.

Two main approaches are used to define the ML forensics, that is, inductive reasoning and deductive reasoning:

- **Inductive learning**
  Inductive reasoning is obtained from the general knowledge of specific information. The obtained knowledge is new and not truth preserving. That means the knowledge obtained can be invalidated from new information. There is no well-founded theory. In this area there are a large number of goals such as it is important to discover general concepts from a limited set of examples. The examples are called experience. The basis of this is to search for similar characteristics among examples. The methods used are based on inductive learning.

- **Deductive learning**
  Deductive reasoning obtains the knowledge from well-established methods called logic. Deductive reasoning obtains from the knowledge by using well-established methods. The knowledge is not new. But it is implicit in the initial knowledge. New knowledge cannot invalidate the existing knowledge obtained and its based on mathematical logic.

Machine learning is divided into three categories which are supervised learning, unsupervised learning and reinforcement learning. The supervised learning deals specifically with the training of dataset that have both input and output variables while the unsupervised train datasets without an output variable. Reinforcement learning involves learning from feedback received through interactions with an external environment.

ML algorithms can be used to analyze the huge amount of data to identify the risk, segment the data and detect criminal behaviour. ML algorithms enable the investigators to interrogate the vast scattered data sets which are placed in social and wired networks and web or cloud computing. In essence, ML algorithms contain the pattern recognition software that are used to analyse huge amounts of data which are used to predict some behaviour. ML algorithms seek to learn from historical perspectives which are then used to predict future behaviour. MLF gains the capability to recognize the patterns of criminal activities through ML algorithms, in order to learn from the historical data about when and where the crime will take place. The malicious activities from extracted data sets can be from burglaries, money laundering or intrusion attacks.

Adoption of the Microsoft Azure and IBM machine learning service is better due to their add-and-drop feature. The add-and-drop feature makes it easy to use as it does not require deep knowledge in machine learning and programming to execute machine learning jobs on both cloud platforms making it an ideal platform for both business and forensic investigations.

Big data problems involve the inability to store, process and analyze data and have led to the introduction of cloud computing. A major service offered by Microsoft Azure cloud is the Microsoft Azure Machine Learning Studio which is a software as a service. The Azure Machine Learning Studio is a collaborative tool with an add-and-drop feature that can be used to build, test and deploy predictive analytics solutions on datasets provided by the end user. The interactive workspace provided by the Azure machine learning studio aids the development of a predictive analysis model. The development of a predictive model is an iterative process which involves transforming and analyzing data from one or more data sources through

several data manipulation and statistical functions so as to produce a set of results. With the interactive and visual workspace provided by Microsoft Azure Studio, users can drag and drop datasets and the different analysis modules needed to build, test and train a predictive analysis model. These modules can be connected to form an experiment which at the end is run on the Azure machine learning studio workspace. Azure machine learning has a plethora of modules and machine learning algorithms that can aid data input, output, visualization,and preparation. The Azure machine learning studio is known for its four well-known categories of algorithms which are classification, regression, clustering and anomaly detection. The building, testing and running of a predictive model in the Azure machine learning studio is time efficient as it takes less time to perform and predictive analytic experiment.

Thank you !