List of Abbreviations

AWS	Amazon Web Services
AaaS	Authentication as a Service
AZ	Availability Zone
CDN	Content Delivery Network
Cloudware	Software that is used to create and deploy services in cloud.
CaaS	Communication as a Service
Community Cloud	Systems and services to be accessible by group of organizations
DaaS	Desktop as a Service
DOS	Denial of Service
DDOS	Distributed Denial of Service
ECS	EC2 Container Service
EC2	Proprietary Technology of Amazon,Stands for Elastic Compute Cloud
Elastic Computing	Ability to dynamically provision and de provision computing and storage resources.
GCP	Google Cloud Platform
Google App Engine	Service that enables users to create and run web applications on Google's infrastructure
HPC	High Performance Cloud
HVM	Hardware Virtual Machine
IaaS	Infrastructure as a Service
Hybrid Cloud	Combination of private and public cloud services
IG	Internet Gateway
Multi Tenant	Multi tenant is a phrase used to describe multiple customers using same cloud
OpenStack	Free and open source software used in data centre.
Private Cloud	Used to describe cloud implemented within corporate firewall.
Public Cloud	Refers to cloud services provided to users over internet to who purchase the service.
PaaS	Platform as a Service
SaaS	Software as a Service
S3	Amazon Storage Service
SOA	Service Oriented Architecture.
SWS	Simple Workflow Service
SLA	Service Level Agreement
VM	Virtual Machine
VPC	Virtual Private Cloud

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In the current IT industry, the use of cloud computing has become a standard norm. Almost everyone has heard of it, and its benefits are far-reaching and wide- saves cost, increases efficiency, helps in doing work faster, etc. In different market researches conducted over time, the results have shown that this trend of the use of cloud computing by businesses and tech houses is bound to increase in the coming years. With increasing usage of cloud the risk of security also increases. There are various types of threats and in this work we have studied various anomaly detection techniques and how can they be used for anomaly detection. Anomaly detection refers to identification of rare items, events, observations which are suspicious in their occurrences. Any deviation from normal behaviour in a working system is an anomaly. Different kind of anomalies differ differently. Typically, when a system is deviating from its normal behaviour then it could be under some kind of attack which might lead to more problems from a security point of view. These challenges have been keeping the researchers hooked on to find anomalous behaviour in systems. Once we have established the anomalous behaviour of system then we can apply several algorithms and approaches to detect and study it. That is the purpose of this work. An Anomaly is abnormal activity or deviation from the normal behaviour. Anomaly detection is the process of removing these abnormal or anomalous behaviours from the data or services. The services delivered to users by cloud service providers must have normal behaviour. To provide services to users in the proper and normal form, anomaly detection becomes important and interesting area for research work. For anomaly detection so many techniques are developed and these techniques are broadly divided into three categories: - statistical, data mining based and machine learning based anomaly detection technique. Anomaly detection techniques are used to detect and discard anomalies from the data or services. In this thesis we provide overview of some anomaly detection techniques which are discovered recently for the tracing data. In the anomaly detection models anomalies are detected by comparing the training data with the actual data. On the basis of comparison deviations in the traced data or services are identified and they are considered as anomaly.

1.1 Motivation

In today's ever changing and dynamic world where technology keeps changing daily and a new attack surfaces every fortnight it has become a challenging task for users and administrators of any infrastructure to keep a track of anomalous behaviour of system. Which should be detected at initial level so that it could be predicted before an attack happens. A lot of cloud computing services have emerged in 21st century like Amazon's AWS, Microsoft Azure, Google Cloud Platform, Alibaba Cloud, Huawei Cloud, Oracle Cloud. Not only software developers but organizations, researchers, companies, enterprises many govt organizations are using these on demand computing services. The computational power which can be offered via cloud service provider is much more than a stand alone system or an infrastructure set up by a small individual or organization can offer. But with more power more challenges have propped up in terms of security of these services. Cloud computing is an Internet-based most recent popular technology offering dynamic resources, scalable resources, on demand, self-service and pay-per-use. Cloud computing is an active area for research and growing very fast. It provides services at low cost and low operational software and hardware expenditure's. The use of cloud computing has increased in companies rapidly because of fast access to applications and decreasing maintenance cost for cloud infrastructure.

1.1.1 First contribution

There are a number of security and privacy concerns associated with cloud computing. Such as 1. Multi-tenancy - makes it easier for a malicious user to to steal the data of all business customers who share the same cloud database. 2. Threats of data loss and data leakage -the threat of data leakage increases when employees use their mobile devices to access and share corporate documents via cloud storage services such as Dropbox, Mega etc., 3. Insecure APIs - Most Cloud service providers do not pay attention to the security of their APIs which may pose risks to an enterprise data with regards to privacy and data integrity. 4. Encryption - Most government regulations require data encryption. Although encryption is widely used, it is often implemented poorly. Many companies depend on the cloud provider for encrypting data which means that service provider has control of the key & can access the data anytime. 5. Key management - Managing and storing these keys in a safe and secure location is of paramount importance when it comes to keeping the entire cloud database safe and secure. Most companies store both encryption and decryption key on the same database which can be harmful for security. Due to such challenges it becomes a hot topic for research.

1.1.2 Second contribution

The key techniques for finding anomalies and outliers are data mining techniques. The goal is to be able to view your entire data set on a single screen and be able to detect patterns in your data. Critical networks require in depth strategy to detect anomalies.One approach of intrusion detection is usually modelled and it is supplied with data to work upon, and using anomaly detection algorithms as described above one of the approach is used. You may or may not use the previously known attacks depending upon your implementation.

1.2 Objective of the thesis

Anomalies, are the most extreme observations, may have maximum or minimum, or both, depending on whether they are extremely high or low. However, the simple maximum and minimum are not always anomalies because they may not be unusually far from other observations. Anomalies may occur by chance in any data set. But they often indicate an observation which needs to be studied. It can lead to security incident or a potential fraud
i Our objective is to study anomaly detection techniques.
ii Come with a solution to detect anomaly

1.3 Contribution of the thesis

Perhaps one of the most important use cases that anomaly detection has is in security. The internet is host to a vast array of various websites that are located all around the world. Unfortunately, due to the ease of access to the Internet, various individuals can access the Internet with nefarious purposes. Similar to the data leaks that were discussed earlier in the context of protecting company data, hackers can launch attacks on other websites as well to leak their information. In cases like this, anomaly detection can help detect network intrusion attacks as they happen. In summary, the areas of contribution of this thesis is as follows.

1.4 Organization of the thesis

The rest of the thesis is organized as follows.
In **Chapter 2**, We reviewed existing algorithms for anomaly detection and literature which exists for cloud computing.Also we have reviewed various kind of cloud computing services provided by various service providers.A comparison of different kind of cloud computing services has been done in this chapter and what kind of clouds exist currently has been briefly described.
In **Chapter 3**
In **Chapter 4**
Finally, we conclude in **Chapter 5**

Chapter 2

Literature Review

Many vendors like Amazon, Microsoft, Oracle etc provide cloud computing services. Like amazon provides AWS which gives you many services like deployment server EC2 instances, work spaces, machine learning APIs etc. In simple words, cloud computing is the delivery of computing services—like servers, storage, databases, networking, software, analytics, and intelligence on Internet ("the cloud") to offer faster innovation, flexible resources, and economies of scale. You pay only for cloud services you use, helping lower your operational costs, and run your infrastructure more efficiently and scale up or scale down as your business needs change. Clouds can be classified into four categories on the basis of physical location of users. Four types of cloud are private, public, community and hybrid clouds. In the available types clouds explain benefits and limitations of each cloud types on the basis of which we can conclude that which cloud model will be suitable for us.

2.1 Cloud Characteristics

General characteristics however would be things such as:
1. Software defined architecture - Not building up based on the hardware & infrastructure available to you but around the needs of your core applications & software.
2. Highly Scalable - Provisioning and scaling up or down to suit your needs on demand is a massive characteristic of the cloud never before so easily seen or done in traditional on-premise servers.
3. Agile - Massive requirements for funding and hardware to take on new developments projects

and deployments is no longer an issue, as the agility of the cloud allows you to leverage a small fraction of what you need to pivot into testing an idea, with very small costs to close it down if it doesn't work, and easy enough options to scale it up if it does.
4. Versatile - Cloud can come in many forms, private, public, hybrid, server less - it can be designed largely around what the priorities that you need from it should be.
Public - very cost efficient.
Private - cyber security and data compliance efficient
Hybrid - Beautiful mix of both worlds, or several providers
Server less - Pay per second of compute playground perfect for developers and small compute loads.

2.2 Deployment Models of Clouds

Cloud computing, can provide dynamic resource allocation, virtualization and highly usable generation of enterprise data center. With cloud computing, the resources are shared and so are the costs. Users can pay as they use and only use what they need at any given time, keeping cost to the user down. Cloud computing is very much a business model as well. Providers of cloud computing solutions, whether they are software, hardware, platform, or storage providers, deliver their offerings over the Internet. There are no shrink wrapped boxes containing discs or hardware for you to buy and set up yourself. Cloud providers typically charge monthly recurring fees based on your usage.
Private Cloud: A private cloud is one which is setup by single organization and installed services on its own data centre.
Public Cloud: Public cloud services are offered by third-party cloud service providers and involve resource provisioning outside of the user's premises.
Community Cloud: The Community cloud can offer services to the cluster of organizations. In other words we can say that community cloud provides combinational services of a group of clouds.
Hybrid Cloud: Hybrid cloud is the combination of any two or more than two types of clouds which are mentioned above means combine any two or more from private, public or community to build it.

2.2.1 Cloud Service Categories

Cloud Computing Services fall into four broad categories:
Infrastructure as a Service (IaaS), Platform as a Service (PaaS), Serverless and Software as a Service (SaaS). These are built on top of one another. What they do we can see in following in detail: -

2.2.1.1 IaaS

Infrastructure as a Service (IaaS) Involves a method for delivering everything from operating systems to servers and storage through IP-based connectivity as part of an on-demand service. Clients can avoid the need to purchase software or servers, and instead procure these resources in an outsourced, on-demand service. Popular examples of the IaaS system include IBM Cloud and Microsoft Azure.

1. IaaS is a cloud service that provides services on “pay-for-what-you-use” basis

2. IaaS providers include Amazon Web Services, Microsoft Azure and Google Compute Engine

3. Users: IT Administrators

2.2.1.2 PaaS

Platform as a Service (PaaS) Platform-as-a-service (PaaS) is considered the most complex of the three layers of cloud-based computing. PaaS shares some similarities with SaaS, the primary difference being that instead of delivering software online, it is actually a platform for creating software that is delivered via the Internet. This model includes platforms like Force.com and Heroku.

1. PaaS provides cloud platforms and run time environments to develop, test and manage software

2. Users: Software Developers

2.2.1.3 SaaS

Software-as-a-Service (SaaS) Software-as-a-service (SaaS) involves the licensure of a software application to customers. Licenses are typically provided through a pay-as-you-go model or on-demand. This type of system can be found in Microsoft Office's 365.

1. In SaaS, cloud providers host and manage the software application on a pay-as-you-go pricing model

2. Users: End Customers

2.2.2 Security Concerns

There are a number of security and privacy concerns associated with cloud computing. 1. Multi-tenancy - makes it easier for a malicious user to to steal the data of all business customers who share the same cloud database. 2. Threats of data loss and data leakage -the threat of data leakage increases when employees use their mobile devices to access and share corporate documents via cloud storage services such as Drobox, Mega etc., 3. Insecure APIs - Most Cloud service providers do not pay attention to the security of their APIs which may pose risks to an enterprise data with regards to privacy and data integrity. 4. Encryption - Most government regulations require data encryption. Although encryption is widely used, it is often implemented poorly. Many companies depend on the cloud provider for encrypting data which means that service provider has control of the key & can access the data anytime. 5. Key management - Managing and storing these keys in a safe and secure location is of paramount importance when it comes to keeping the entire cloud database safe and secure. Most companies store both encryption and decryption key on the same database which can be harmful for security. Outliers and anomalies are not the same thing. For example, there might be 5 records in your data that lie exactly at the mean. These records are not considered outliers. But maybe these records are instances of fraud, and therefore considered anomalies. The key techniques for finding anomalies and outliers are dimensionality reduction and visualization. The goal should be to view your entire data set on a single screen and be able to detect patterns in your data. Naming a few algorithms that can help achieve this goal: K-nearest neighbours, minimum spanning tree, self-organizing maps, latent Dirichlet allocation.

Critical networks require in depth strategy to detect anomalies.One approach of intrusion detection is usually modelled and it is supplied with data to work upon, and using anomaly detection algorithms as described above one of the approach is used. You may or may not use the previously known attacks depending upon your implementation.

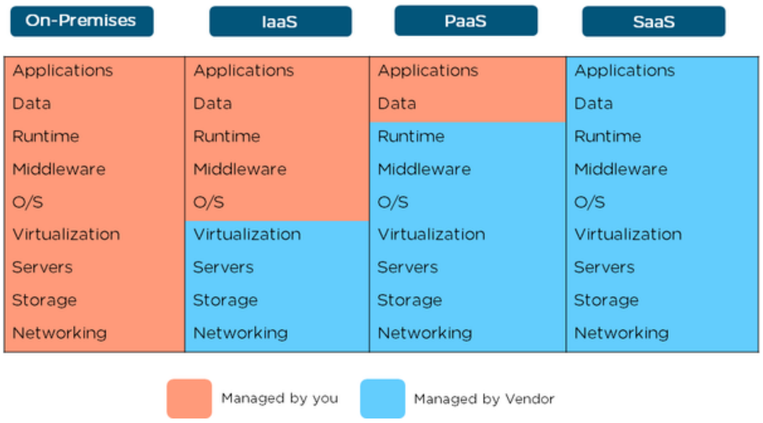
Security Issues with respect to Cloud Computing [1] have been widely discussed both in academics and industry several international conferences have focused on this subject : Confidentiality,Virtualization Level Issues,Multi Tenancy Issues, VM Isolation Issues,Virtual Network Issues,Virtual Machine Introspection Issues,VM Management Issues,Application Level Issues,Isolation Issues,Synchronization Mechanism Issues, Data Storage Level Issues,Outsourcing Issues,Data Deletion Issues, Network Level Issues. DoS attacks, and DDOS attacks affect cloud computing infrastructure and services. These type of attacks are major attacks which affect the available services. The Cloud Infrastructure provided by vendor could be a victim of such attacks but apart from that it could also be participating in such attacks. Botnets , botClouds can be deployed in Cloud Environment [2] to launch such attacks. Hence from security perspective it is important to identify such an issue that may exist in underlying cloud infrastructure.
In [3] author discusses that with the advent of cloud computing a lot of data has been generated how this data has been impacting the world and there is a lot of growth in the data of devices that have been connected. So, with the improvements in bandwidth and availability. In the context of the Internet of Things, the trouble with the cloud is that data needs to be sent back from the sensors gathering info, such as a Nest thermostat or a Fitbit wristband, to a database in a remote public cloud. The time that it takes for the data to be transferred from the device or sensor to the remote public cloud, that is the latency, is often too great to meet the requirements of the IoT system. The cloud complicates this process even more.We're focused on centralized computing, thus there will be latency. Now, instead of sending the data back to the data centre on the other side of the factory, we send it to a remote cloud server that can be thousands of miles away. To make things worse, we send it over the open Internet.
In [4] discussion about data sharing has been made. As how participants must share data. Protocols have played an important role. Protocols have played an important role in transfer of data in case of cloud computing. Storage has become a hot topic in today's cloud computing world. We prefer to store all types of data in cloud servers, which is also a good option for companies and organizations to avoid the overhead of deploying and maintaining equipment when data are stored locally. In cryptography, a key agreement protocol is a protocol in which two or more parties can agree on a key in such a way that both influence the outcome. This kind of protocol has widespread application in technology of internet and cloud computing.
In [5] mobile edge computing and emerging models in fog computing have been discussed.

Relation between them is evident. The approach in the paper is to examine and underpin the models that are existing in cloud computing. Characteristics of cloud like support of ubiquitous connectivity, elasticity, scalable resources and ease of deployment have played an important role in development of existing cloud computing infrastructure. Research community has proposed new technologies namely fog and cloud. These technologies have been labelled in the paper as extended cloud they allow computing needs to be performed closer to source of data. This results in improvement in quality of services provided since this results in reduction of delay in conveying data between end nodes and cloud. Such technologies have enabled support for new application and services example Google now, foursquare and both are location aware applications for mobile platforms. Further this can be extended to services like autonomous vehicles robotics, public safety and augmented reality.

The acceptable level of service depends upon user expectations. Now these days users require rapid access to service like always on always available. So a new term has popped up known as resilience [5]. Resilience is concerned with availability of services and maintaining confidentiality and integrity of information in face of challenges. Resilience has become a fundamental property of cloud service provisioning platforms. With the advancement in wireless related technologies security and resiliency have become key issues when considering Mobile Edge Computing Services. With regards to edge model there are few threats also which have evolved. For example, infrastructure related threats, virtualization related threats, privacy related threats. Fog computing model was originally conceived by Cisco as an extension of cloud. The term fog was originally coined by Cisco as there is need to enable a platform that can cope up with the requirements posed by challenges put forward by Internet of Things. Another requirement in fog computing is privacy of data. In the paper detection and resilience mechanism have been discussed. The area that has been challenging to researchers is anomaly detection. In September 2016 [6] website of computer security consultant Brian Krebs was hit with 620 Gbps traffic. At the same time a bigger DDoS attack using Mirai malware was done on French web hosting and cloud service provider OVH. Mirai's source code was release by its creator soon after wards. Hackers offered Mirai's botnets for rent with as many as 400,000 connected devices. More attacks happened in October 2016 using Mirai they took down hundreds of websites like Twitter,Netflix,Reddit, Github for several hours. Mirai spreads by infecting devices as web cams,DVRs, routers, then it finds out administrative controls of those devices by a brute force attack which relies on a dictionary of potential usernames and passwords.

2.3 Cloud Services Comparison

Below is a comparison of various kind of services provided by cloud services providers versus the services managed by vendors. To be considered a cloud, a technology model must possess these five characteristics: on-demand self-service, meaning that anyone with a browser can subscribe to the service; measured service, meaning that monitoring capabilities allow providers to offer service by subscription, pay-per-use, or other pricing models; elastic scalability, which means that cloud subscribers can adjust computing resources as they see fit; resource pooling, which means that virtualized storage, servers, and networks are pooled together at a single location or across many locations to create a virtually infinite supply of resources; and broad network access. The cloud can be deployed in one of three methods: the public cloud, which allows you to pay only for the resources you use; the private cloud, which runs on dedicated IT only; and the hybrid cloud, which combines the scalability of the public model with the security of the private model.



On-Premises	IaaS	PaaS	SaaS
Applications	Applications	Applications	Applications
Data	Data	Data	Data
Runtime	Runtime	Runtime	Runtime
Middleware	Middleware	Middleware	Middleware
O/S	O/S	O/S	O/S
Virtualization	Virtualization	Virtualization	Virtualization
Servers	Servers	Servers	Servers
Storage	Storage	Storage	Storage
Networking	Networking	Networking	Networking

🟧 Managed by you 🟦 Managed by Vendor

Figure 2.1: Cloud Services Comparison

2.4 Type of Anomalies

Anomalies, are the most extreme observations, may have maximum or minimum, or both, depending on whether they are extremely high or low. However, the simple maximum and minimum are not always anomalies because they may not be unusually far from other observations. Anomalies may occur by chance in any data set. But they often indicate an observation which needs to be studied. It can lead to security incident or a potential fraud in the given data set and it might give a vital clue about the incident. How you study that has to be based on some of the techniques as discussed in the introduction section of this document.

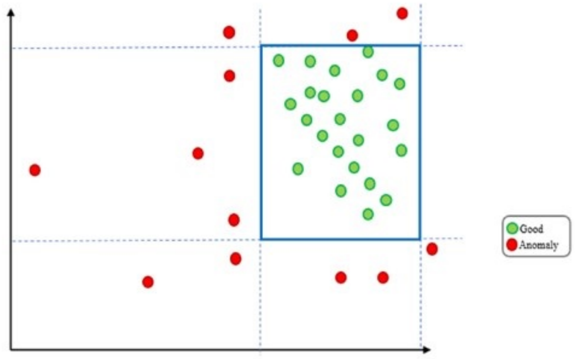


Figure 2.2: Simple Anomaly

Broadly speaking the anomalies are defined into 3 main types. Global outliers, contextual outliers and collective outliers.

2.4.1 Global anomaly

Global outliers are known as point anomalies. These exist far outside the entirety of dataset. Global outlier deviates significantly from dataset. For example in a data set if 99 points out of 100 have a value between 400-500 and one point is having value 900. Then that point which has value 900 is significantly out of the range 400-500 hence this can be safely considered a global outlier for this data set. Such a point has a very high or very low value in the dataset. It is important to identify such outliers because there may be real abnormalities in this kind of data set. For example if you have a dataset of Covid-19 patients and symptoms on some parameters of Covid are coming same but one particular parameter is deviating significantly then that patient may not be having Covid, because many other disease of same virus family can be having same set of tests coming in a certain range but in order to qualify for a disease some all parameters must be within that range. So, if something is deviating significantly then it obviously is a a outlier and to qualify as a global outlier it has to vary significantly from the given data set.

When there is such an outlier, the value can be out of the range of the entire distribution it will be unusual relative to the rest of the points in dataset. In the image shown above, you can see point which is totally having an extreme value lies in certain range. This point is having very high value on y-axis, which indicates that it is an anomaly. When these points are plotted you

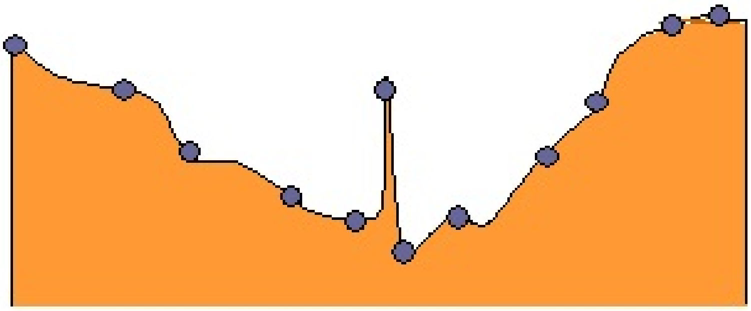


Figure 2.3: Global Anomaly

can see that all these points are pointing in the graph drawn to an abnormal location. When you look at the diagram you can see that other values are not that unusual. The location in question is having a very high value in the diagram and is paired with higher surrounding values. This location may be a local outlier also. Further investigation should be made before deciding if the value at that point is erroneous or in fact reflects a true characteristic of the phenomenon and should be included as global.

2.4.2 Contextual Anomaly

In case of contextual outliers, the data values differ but there has to be a context in which data values differ. Something which is anomaly in one context of data may not be anomaly in another context. Such kind of anomalies are more common in time series data because such datasets are record of specific quantities in a given period. The value in an observed pattern may appear as expected but it may be anomalous in some other context.

For example, consider the usage of a bank ATM transactions between 10 AM to evening 10 PM may be normal. But there might be time intervals between these times where there could be heavy transactions and those transactions could be anomalous also. Like a heavy transaction happening during 10 PM to 6 AM in morning. We can consider another example to understand it better. Suppose you have data of temperature of Guwahati in particular weather. Let it be 35° C in month of November. Is it exceptional? The answer to this question depends upon the data you have in your dataset for November month temperature in that city. So this in this case the time of occurrence of temperature could be anomalous if it deviates from a range of temperature in November for that city. Such kind of anomaly is contextual anomaly because context of occurrence is important for the deviation to be considered an anomaly.

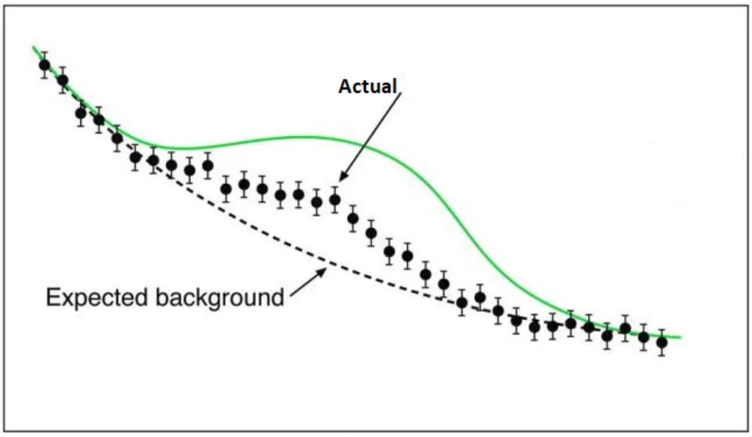


Figure 2.4: Contextual Anomaly Example

2.4.3 Collective Anomaly

A subset of data points within data set is anomalous if those values differ significantly from the dataset. For example, you have data set of a bank transactions and you observe that for certain period transactions are done in time of day. The known normal transactions in the data set could be of some range in which the customer deals. For the context of the explanation we say that a transaction of 1000 INR in a customers banking transaction is normal during 2PM – 4PM. But for some day you discover that in this time there is transaction worth 50,000 INR during same time. So, this could be a potential fraud also. In our case it we call it anomaly. So, to detect it the banker or anomaly detection system would study all the transaction logs of certain date i.e. 6 months transactions. During the study of those transactions you would study at what time the transactions happened frequently and when is the most out of context transaction happened in that time so this kind of study would need study of time series and at certain point in that time. So, this is an example of collective outlier and this is often used in computer science transactions like data centre operations etc. You will need the data log of servers doing transactions in a cloud environment to understand or figure out the anomaly in those transactions.

To be able to come at accurate results you need to have a data set which has a lot of observations. Then only you will be able to have some observations in it. Outliers may have many anomalous characters. Determining or not whether an observation is outlier needs subjective exercise. Model based methods are generally used for outlier detection and identify observations which are unlikely event or deviation from the standard behaviour. Collective anomaly requires information to be analysed from many data streams. To be able find more and more anoma-

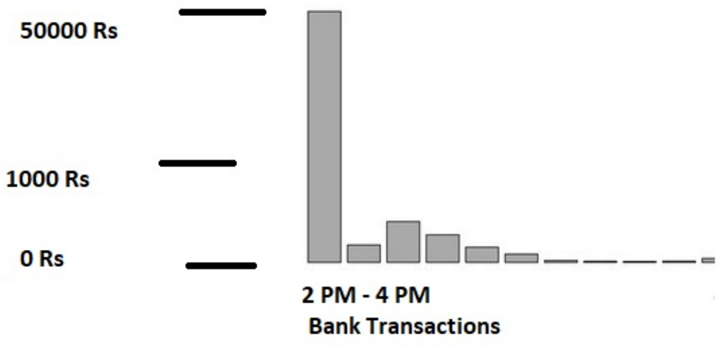


Figure 2.5: Collective Anomaly

lous points you need to have data from various collections having different different time data. Collective anomaly and contextual anomalies are detected with different techniques. The importance of anomaly detection is due to the fact that anomalies in data translate to significant (and often critical) actionable information in a wide variety of application domains. For example, an anomalous traffic pattern in a computer network could mean that a hacked computer is sending out sensitive data to an unauthorized destination. An anomalous MRI image may indicate presence of malignant tumours. Anomalies in credit card transaction data could indicate credit card or identity theft. At an abstract level an anomaly is defined as pattern that does not confirm to expected normal behaviour. A simple approach to detect anomaly is therefore to define a region representing normal behaviour and declare any observation in this data which does not belong to normal region as anomaly. But there are several things which need to be taken in consideration as these things make task difficult. Finding a region and declaring it as normal which includes all possible normal observations in the given pattern of data. Defining the boundary of normal and anomalous data. An anomalous point could lie very close the actual normal data and the opposite can also happen. When there are ill intention ed actions in a system then malicious actors can behave in such a way which looks normal and from the logs it is difficult to detect what is normal or what is anomalous. Data availability to train a classifier and generate a model is a challenge. As an example if you are trying to do research on anomalous transactions in banking data then no bank would share their data with researcher, or if you are doing research in computer security then no corporate would share their systems logs with researchers to detect anomalies. Often the data contains noise that means the given data set has data which looks like genuine data but the data contains unclean data which is useless and it does not give any information. So a lot of techniques from various disciplines like machine learning, data mining, information theory, statistics ,data science, neural networks etc are used in anomaly detection.

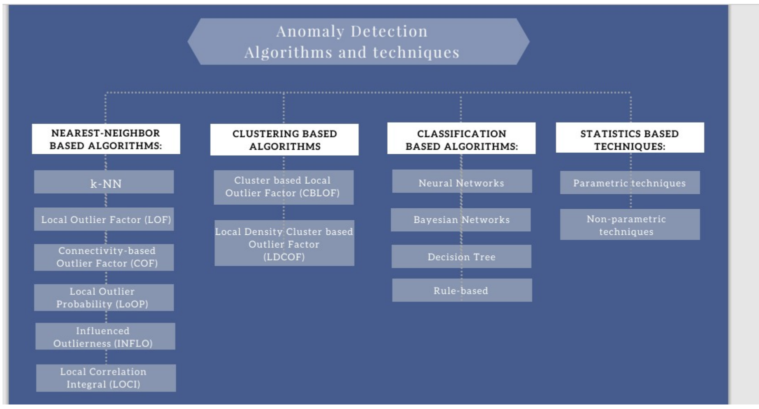

Anomaly Detection
Algorithms and techniques
NEAREST-NEIGHBOR BASED ALGORITHMS:
CLUSTERING BASED ALGORITHMS
CLASSIFICATION BASED ALGORITHMS:
STATISTICS BASED TECHNIQUES:
k-NN
Local Outlier Factor (LOF)
Connectivity-based Outlier Factor (COF)
Local Outlier Probability (LoOP)
Influenced Outlierness (INFLO)
Local Correlation Integral (LOCI)
Cluster based Local Outlier Factor (CBLOF)
Local Density Cluster based Outlier Factor (LDCOF)
Neural Networks
Bayesian Networks
Decision Tree
Rule-based
Parametric techniques
Non-parametric techniques

Figure 2.6: Anomaly Detection Algorithms

Below is a brief overview of popular machine learning-based techniques for anomaly detection.

2.4.4 Density-Based Anomaly Detection

Density-based anomaly detection is based on the k-nearest neighbors algorithm. Usually normal data points occur around a dense neighborhood and those data points are surrounded by the points which lie in normal range and abnormal points are far away from this neighbourhood.
Assumption: Here it is safe to assume that any anomaly point will lie outside the dense neighbourhood.
So it becomes easy to figure out anomaly data points using a distance based clustering technique, which could be Eucledian distance or a similar measure dependent on the type of the data (categorical or numerical). Such techniques are broadly classified into two algorithms:
K-nearest neighbor: k-NN is a simple, non-parametric lazy learning technique used to classify data based on similarities in distance metrics such as Eucledian, Manhattan, Minkowski, or Hamming distance. Relative density of data: This is better known as local outlier factor (LOF). This concept is based on a distance metric called reachability distance.

2.4.5 Clustering-Based Anomaly Detection

In the domain of unsupervised learning Clustering is one of the most popular concepts.
Assumption: Data points that are similar tend to belong to similar groups or clusters, as deter-

mined by their distance from local centroids.
K-means is a widely used clustering algorithm. It creates 'k' similar clusters of data points. Points which fall outside of these grouped data points could potential anomalous points.

2.4.6 Support Vector Machine-Based Anomaly Detection

A support vector machine is another effective technique for detecting anomalies. A SVM falls in category of supervised learning techniques, but there are other techniques also using SVM that can be used to identify anomalies as an unsupervised problems (in which training data are not labeled). The algorithm learns a soft boundary in order to cluster the normal data instances using the training set, and then, using the testing instance, it tunes itself to identify the abnormalities that fall outside the learned region.Depending on the case, the output of an anomaly detection algorithms could be numeric or scalar values for on specific data set. However, looking at the data sets, it is not possible to identify the outlier directly from analyzing one variable at the time. Hence a combination of the X and Y variable is taken that allows us to easily identify the anomaly. But if the variables are more than two then this makes the work more difficult because then we have to scale up from two variables to 10–100s of variables, which is often the case in practical applications of anomaly detection.

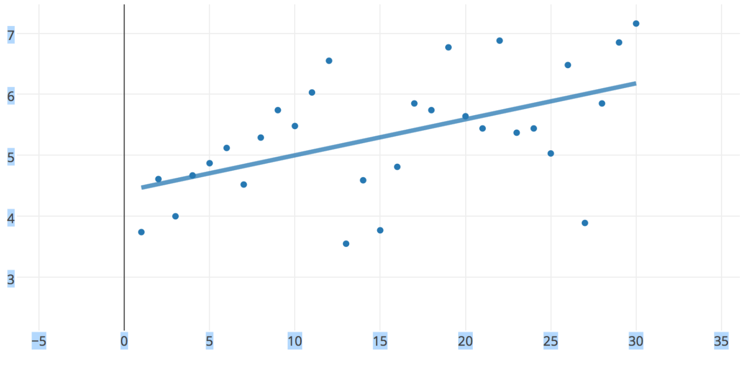


Figure 2.7: Anomaly Detection Using 2 variables

2.4.7 Anomaly detection algorithms categories

Generally outlier detection algorithms can be categorised in two categories – supervised and unsupervised learning. Following are a couple of algorithms to used in anomaly detection

2.4.7.1 K-nearest neighbor: k-NN

k-NN is one of the simplest supervised learning algorithms and methods in machine learning. It stores all of the available examples and then classifies the new ones based on similarities in distance metrics.k-NN is a famous classification algorithm and a lazy learner. What does a lazy learner mean? K-nearest neighbor mainly stores the training data. It doesn't do anything else during the training process. That' s why it is lazy.k-NN just stores the labeled training data. When new unlabeled data arrives, kNN works in 2 main steps: Looks at the k closest training data points (the k-nearest neighbors).Then, as it uses the k-nearest neighbors, k-NN decides how the new data should be classified.

How does k-NN know what's closer? It uses density-based anomaly detection methods. For continuous data (see continuous vs discrete data), the most common distance measure is the Euclidean distance. For discrete data, Hamming distance is a popular metric for the "closeness" of 2 text strings.The pick of distance metric depends on the data.The k-NN algorithm works very well for dynamic environments where frequent updates are needed. In addition, density-based distance measures are good solutions for identifying unusual conditions and gradual trends. This makes k-NN useful for outlier detection and defining suspicious events.k-NN also is very good techniques for creating models that involve non-standard data types like text.k-NN is one of the proven anomaly detection algorithms that increase the fraud detection rate. It is also one of the most known text mining algorithms out there.It has many applications in business and finance field. For example, k-NN helps for detecting and preventing credit card fraudulent transactions.

2.4.7.2 Local Outlier Factor (LOF)

The LOF is a key anomaly detection algorithm based on a concept of a local density. It uses the distance between the k nearest neighbors to estimate the density.To put it in other words, the density around an outlier item is seriously different from the density around its neighbors.That is why LOF is called a density-based outlier detection algorithm. In addition, as you see, LOF is the nearest neighbors technique as k-NN.LOF is computed on the base of the average ratio of the local reachability density of an item and its k-nearest neighbors.

2.4.7.3 K-means

K-means is a very popular clustering algorithm in the data mining area. It creates k groups from a set of items so that the elements of a group are more similar.In K-means technique, data

items are clustered depending on feature similarity.One of the greatest benefits of k-means is that it is very easy to implement. K-means is successfully implemented in the most of the usual programming languages that data science uses.
If you are going to use k-means for anomaly detection, you should take in account some things: The user has to define the number of clusters in the early beginning. k-means suppose that each cluster has pretty equal numbers of observations. k-means only work with numerical data. Is k-means supervised or unsupervised? It depends, but most data science specialists classify it as unsupervised. The reason is that, besides specifying the number of clusters, k-means "learns" the clusters on its own. k-means can be semi-supervised.

2.4.7.4 Support Vector Machine (SVM)

A support vector machine is also one of the most effective anomaly detection algorithms. SVM is a supervised machine learning technique mostly used in classification problems. It uses a hyperplane to classify data into 2 different groups. Just to recall that hyperplane is a function such as a formula for a line (e.g. y = nx + b).To say it in another way, given labeled learning data, the algorithm produces an optimal hyperplane that categorizes the new examples.
When it comes to anomaly detection, the SVM algorithm clusters the normal data behavior using a learning area. Then, using the testing example, it identifies the abnormalities that go out of the learned area.

2.4.7.5 Neural Networks Based Anomaly Detection

When it comes to modern anomaly detection algorithms, we should start with neural networks.Artificial neural networks are quite popular algorithms initially designed to mimic biological neurons. The primary goal of creating a system of artificial neurons is to get systems that can be trained to learn some data patterns and execute functions like classification, regression, prediction and etc. What makes them very helpful for anomaly detection in time series is this power to find out dependent features in multiple time steps. There are many different types of neural networks and they have both supervised and unsupervised learning algorithms. Example of how neural networks can be used for anomaly detection, you can see here. The above 5 anomaly detection algorithms are the key ones. However, there are other techniques. Here is a more comprehensive list of techniques and algorithms.

Table 2.1: Anomaly Detection Algorithms comparison

Algorithm	Pros	Cons

label=, wade = 0pt, leftmargan = *, nosep, itemsep = 0pt, before = , after =
lbbel=, wbde = 0pt, leftmbrgbn = *, nosep, btemsep = 0pt, before = , bfter =
K Nearest Neighbour
K-NN
	1. Very easy to understand
	2. Good for creating models that include non standard data types such as text
		Large Storage requirements
		Computationally Expensive
		Sensitive to the choice of the similarity function for comparing instances

Local Outlier Factor(LOF)	Well-known and good algorithm for local anomaly detection	Only relies on its direct neighborhood . Perform poorly on data sets with global anomalies.
K Means	Low Complexity Very easy to implement	Each cluster has pretty equal number of observations Necessity of specifying K Only work with numerical data
Support Vector Machine (SVM)	Find the best separation hyper-plane.Deal with very high dimensional data. Can learn very elaborate concepts. Work very well	Require both positive and negative examples. Require lots of memory. Some numerical stability problems.Need to select a good kernel function

Neural networks based anomaly detection	Learns and does not need to be reprogrammed. Can be implemented in any application	Needs training to operate Requires high processing time for large neural networks The architecture needs to be emulated

More commonly used type is supervised learning. It includes such algorithms as logistic and linear regression, support vector machines, multi-class classification, and etc. Classification methods (also called supervised methods) require a training data set that includes both normal and anomalous examples to construct a model that can find anomalies.Where as, unsupervised learning includes the idea that a computer can learn to discover complicated processes and outliers without any human intervention.

Chapter 3

First Contribution

Movement of services is essential for IT employees and other employees it increases the flexibility.Ever increasing modernization increases challenges for IT professionals in terms of new demands and complexities for keeping the organization secure. It is difficult to get the full benefit of cloud apps and services, an IT team must find the right combination of supporting technologies while maintaining control to protect critical data.Day by day with increasing challenges keeping the systems secure and marinating their integrity has been a challenging task. This paper studies various anomaly detection techniques used by cloud service providers and various approaches which exist in literature and discuses the anomaly detection techniques using KDD data set.

Keywords: Cloud, Virtualization, Time Series Analysis,Anomaly Detection,Distance Based Clustering

3.1 Outlier Detection

Any deviation from normal is considered an anomaly.Outlier detection (also known as anomaly detection) is the process of finding data objects with behaviors that are very different from expectation. Such objects are called outliers or anomalies. The most interesting objects are those, that deviates significantly from the normal object. Outliers are not being generated by the same mechanism as rest of the data. [7]

3.2 Security Framework

In Cloud Computing services security is a series of documented processes and frame works.Which revolves around defining policies and procedures around the implementation. For any successful security implementation of security frame work we need to define policies and procedures with respect to cloud.The main point of having a frame work is to reduce the risk levels and organizations exposure to vulnerabilities.

Control the use of apps: First we identify the cloud apps,various cloud apps can be using infrastructure as a service, or platform as a service services used by your organization.One needs to study usage patterns, identify the risk levels and business readiness of apps against risks. Start managing them to ensure security and compliance. [8] To protect your sensitive information anywhere in the cloud: understand, classify, and protect the exposure of sensitive information at rest. Apply out-of-the box policies and automated processes to apply controls in real-time across all your cloud apps.Then protect against cyber threats and anomalies: Detect unusual usage patterns analyzed by behavior of app usage across cloud to identify ransomware,currency mining, compromised users or fake applications, analyze high-risk usage and remediate automatically to limit the risk to your organization. In this process you need to assess the compliance of your cloud apps: Assess if your cloud apps meet relevant compliance requirements including regulatory bodies and industry standards. Prevent data leaks to non-compliant apps, and limit access to regulated data.When an organization elects to store data or host applications on the public cloud, it loses its ability to have physical access to the servers hosting its information. As a result, potentially sensitive data is at risk from insider attacks.The extensive use of virtualization in implementing cloud infrastructure brings unique security concerns for customers or tenants of a public cloud service. Virtualization affects the relationship between the operating system and underlying hardware – be it networking,storage or even computing. This introduces an additional layer – virtualization – that itself must be properly configured, managed and secured. Security in Microsoft Systems is done based via policies in accounts.To create policies go to control inside it go to templates.Select a policy template from the list, and then choose (+) Create policy. Customize the policy (select filters, actions, and other settings), and then choose Create. On the Policies tab, choose the policy to see the relevant matches (activities, files, alerts). Tip: To cover all your cloud environment security scenarios, create a policy for each risk category. How can policies help your organization? You can use policies to help you monitor trends, see security threats, and generate customized reports and alerts. With policies, you can create governance actions, and set data loss prevention and file-sharing controls.

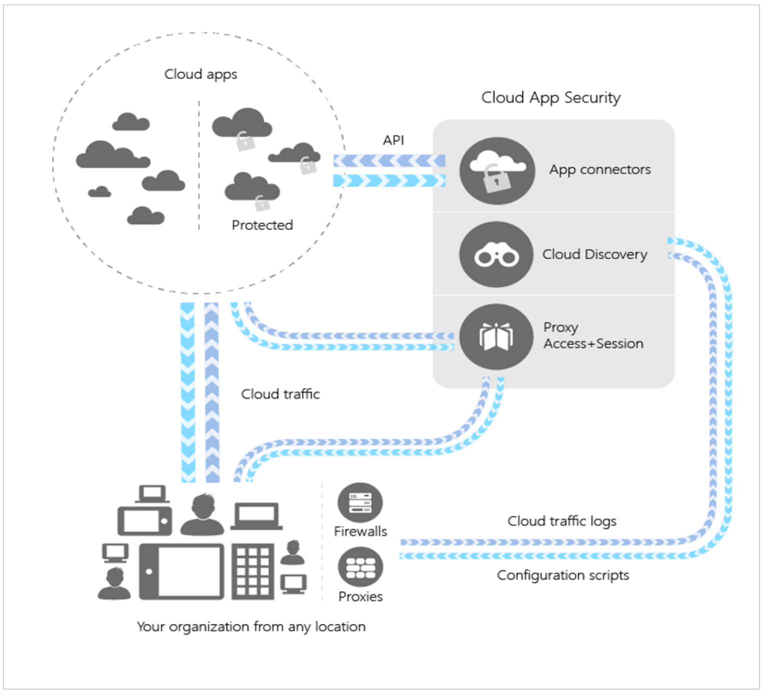


Figure 3.1: Security architecture

3.3 How Microsoft Detects Anomaly in Azure

Microsoft uses a technology called security center [9] in its cloud environments.To detect threats and reduce false positives Security Center collects data from Azure resources and network and then applies a lot of machine learning and big data algorithms to detect threat.These techniques as more advance than normal signature based anomaly detection techniques.It is impossible to manually identify the attack and predict the attack when it might happen.So the Microsoft Security Center uses following technologies

1. Intelligent Threat Intelligence
2. Behavioral Analytics
3. Fusion Analytics

3.3.1 Intelligent Threat Intelligence

It looks for bad actors by using the information obtained via Microsoft Products and services,Microsoft Digital Crimes Unit and Microsoft Security Response Centre along with external feeds.Microsoft has an immense amount of global threat intelligence. Telemetry flows in from multiple sources, such as Azure, Office 365, Microsoft CRM online, Microsoft Dynamics AX, outlook.com, MSN.com, the Microsoft Digital Crimes Unit (DCU), and Microsoft Security Response Center (MSRC). Researchers also receive threat intelligence information that is shared among major cloud service providers and feeds from other third parties. Azure Security Center can use this information to alert you to threats from known bad actors.

3.3.2 Behavioral Analytics

Behavioral analytics is a technique that analyses and compares data to a collection of known patterns. However, these patterns are not simple signatures. They are determined through complex machine learning algorithms that are applied to massive datasets. They are also determined through careful analysis of malicious behaviors by expert analysts. Azure Security Center can use behavioral analytic s to identify compromised resources based on analysis of virtual machine logs, virtual network device logs, fabric logs, crash dumps, and other sources. In addition, there's correlation with other signals to check for supporting evidence of a widespread campaign. This correlation helps to identify events that are consistent with established indicators of compromise.

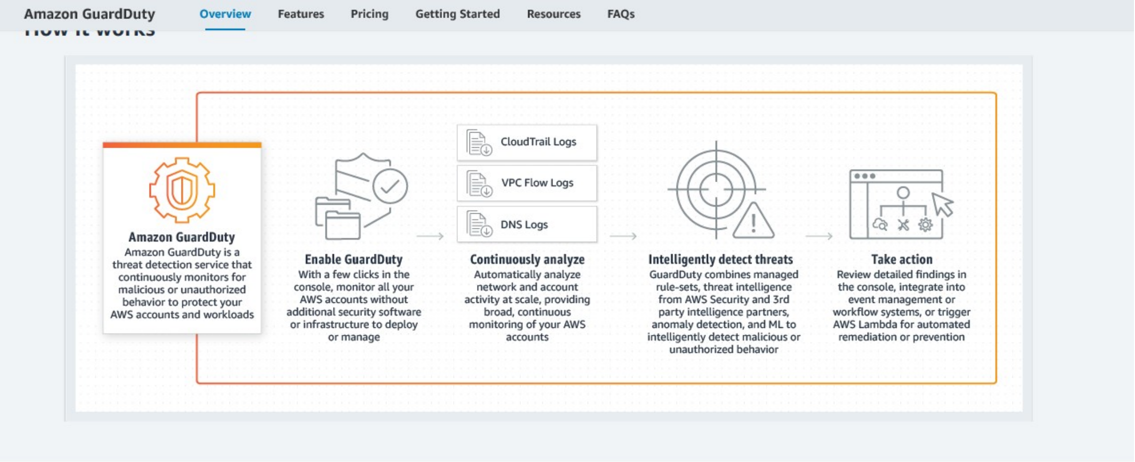
3.3.3 Fusion Analytics

It uses statistical techniques to build a historical data which is based on usage patterns and any deviation from normal alerts those deviations [10] and it creates a baseline which if confirms to a potential attack vector then the particular usage is detected as an anomaly and in turn thus could be a security event.

Security Centre in Azure also works with connected partner solutions, like firewall and endpoint protection solutions. Microsoft uses **Fusion Analytics** [11] as the backbone of Security Centre's anomaly detection system.Fusion works by looking at various kind of alerts generated in Microsoft Azure ecosystem and then it tries to find pattern which could reveal attack progression indicating what should be next course of action.

3.4 How Amazon Finds Anomalies in AWS

Amazon uses a technology called guard duty [12].This service continuously monitors for threat detection service and continuously monitors bad behavior and protects aws accounts and workloads.In the cloud related services collection and aggregation of network related information and activities is simplified, but it is time consuming for security teams to continuously analyses event log data for potential threats. With the help of Guard Duty now you have an cost effective and intelligent solution which is used for threat detection in AWS cloud. Guard Duty uses machine learning, anomaly detection, and integrated threat intelligence to identify and prioritize potential threats. Guard Duty analyses tens of billions of events across multiple AWS data sources, such as AWS CloudTrail, Amazon VPC Flow Logs, and DNS logs. With a few clicks in the AWS Management Console, Guard Duty can be enabled with no software or hardware to deploy or maintain. By integrating with AWS Cloud Watch Events [13], Guard Duty alerts are actionable, easy to aggregate across multiple accounts, and straightforward to push into existing event management and workflow systems.



Amazon GuardDuty
Amazon GuardDuty is a threat detection service that continuously monitors for malicious or unauthorized behavior to protect your AWS accounts and workloads

Enable GuardDuty
With a few clicks in the console, monitor all your AWS accounts without additional security software or infrastructure to deploy or manage

Continuously analyze
Automatically analyze network and account activity at scale, providing broad, continuous monitoring of your AWS accounts

Intelligently detect threats
GuardDuty combines managed rule-sets, threat intelligence from AWS Security and 3rd party intelligence partners, anomaly detection, and ML to intelligently detect malicious or unauthorized behavior

Take action
Review detailed findings in the console, integrate into event management or workflow systems, or trigger AWS Lambda for automated remediation or prevention

Figure 3.2: Amazon Guard Duty

3.5 How Google finds anomalies in GCP cloud

Google uses an open source library Forseti [14] which can

1. Detect unusual firewall behaviors between snapshots.

2. Alert users to any unusual behaviors and provide a comparison with expected behaviors.

3. Provide potential remediation steps.

The key elements for this technology are firewall rules.Firewall rules can be inbound or outbound. A firewall rule can either allow traffic based on IP address or ports.Firewall rules are applied to those instances which are associated with the user who has created his cloud in GCP setup. The figure below shows the technical architecture as how this has been implemented in GCP

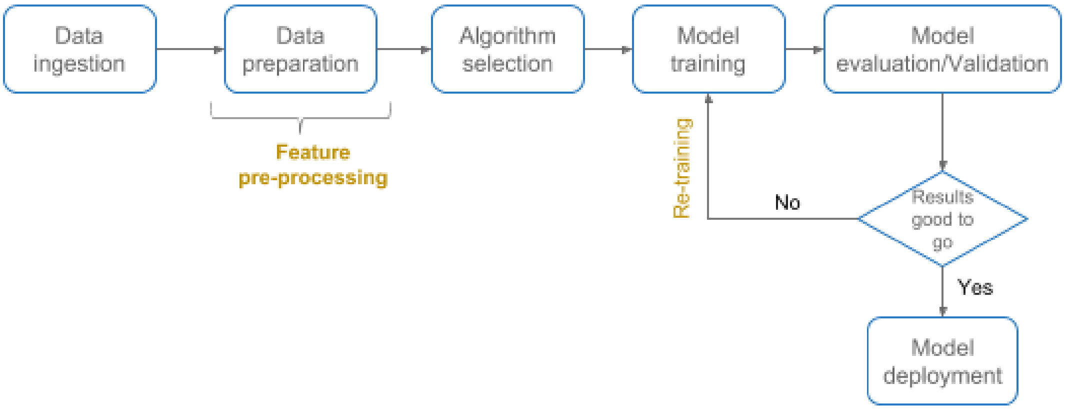


Figure 3.3: GCP Anomaly Detection

3.6 Finding Anomaly in Cloud

The main goal of an anomaly detection system is to discriminate the occurrence of hostile activities from the normal events, and such analysis must be accomplished in a sufficiently flexible and effective way to keep up with the continuously evolving world of cyber security where new, previously unknown, anomalies can continuously emerge over time. In doing this, it must either try to model any kind of attack or anomalous event that can affect the network (there are thousands of known ones) or simply construct a sufficiently general model describing the normal traffic.Such model is usually built on the basis of training data, and used in classifying previously unseen or suspicious events. Classification is the fundamental task in unattended detection, by which the system “learns” to automatically recognize complex traffic patterns, to distinguish between different events based on the corresponding patterns, and to make “intelligent” decisions.

Specific machine learning techniques, such as Neural Networks or Support Vector Machines are used in an anomaly detection system.

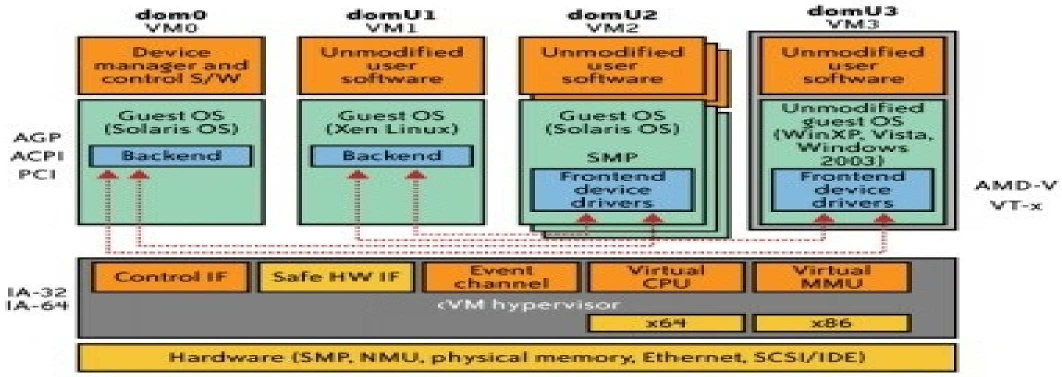


Figure 3.4: Hypervisor working

3.6.1 Building the Anomaly Detection System

An Anomaly Detection System can be build using a generalization capability from training data which is needed to correctly classify future data as normal or abnormal. These resulting approaches can be categorized as generative or discriminating. A generative approach builds a model solely based on normal training examples and evaluates each testing case to see how well it fits the model. A discriminating approach, on the other hand, attempts to learn the distinction between the normal and abnormal classes. Thus, based on the characteristics of training data used to build the model, anomaly detection can be divided into three broad classes which appear below the table (See Table 3.1).

Technique	Category
Supervised anomaly detection	A
Semi Supervised Anomaly Detection	B
Unsupervised anomaly detection	C

Table 3.1: Type of anomaly detection techniques.

3.7 Detecting the deviation

In particular, in the context of abuse and network intrusion detection [15], the interesting objects are often not rare objects, but unexpected bursts in activity. This pattern does not adhere to the common statistical definition of an outlier as a rare object, and many outlier detection methods (in particular unsupervised methods) will fail on such data, unless it has been aggregated appropriately. Instead, a cluster analysis algorithm may be able to detect the micro clusters formed by these patterns.Three broad categories of anomaly detection techniques exist.i.e. Statistical Anomaly Detection Systems, Data Mining Based Anomaly Detection Systems , Machine Learning Based Anomaly Detection Systems. [16] Unsupervised anomaly detection techniques detect anomalies in an un labeled test data set under the assumption that the majority of the instances in the data set are normal by looking for instances that seem to fit least to the remainder of the data set. Supervised anomaly detection techniques require a data set that has been labeled as ”normal” and ”abnormal” and involves training a classifier (the key difference to many other statistical classification problems is the inherent unbalanced nature of outlier detection). Semi-supervised anomaly detection techniques construct a model representing normal behavior from a given normal training data set, and then test the likelihood of a test instance to be generated by the learned model.

Anomaly detection [17] finds its use in a lot of fields such as fraud detection, fault detection, health monitoring, intrusion detection,event detection in sensor networks, finding disturbances in ecosystem. It is often used in pre processing to remove anomalous data from the dataset. In supervised learning, removing the anomalous data from the dataset often results in a statistically significant increase in accuracy. Several anomaly detection techniques have been proposed in literature.

Some of the popular techniques are:

- Density-based techniques (k-nearest neighbor, local outlier factor, isolation forests, and many more variations of this concept).
- Subspace-, correlation-based and tensor-based outlier detection for high-dimensional data.
- One-class support vector machines.
- Replicator neural networks., auto encoders, long short-term memory neural networks
- Bayesian Networks.Hidden Markov models (HMMs).
- Cluster analysis-based outlier detection.
- Deviations from association rules and frequent item sets.
- Fuzzy logic-based outlier detection. [18]

• Ensemble techniques, using feature bagging,score normalization and different sources of diversity.

Different methods perform differently a lot on the data set and parameters, and methods it may have little systematic advantages over another when compared across many data sets and parameters. Sample Anomaly Detection Problems. These examples show how anomaly detection might be used to find outliers in the training data or to score new, single-class data. Algorithm for Anomaly Detection. Oracle Data Mining supports One-Class Support Vector Machine (SVM) [19]for anomaly detection. When used for anomaly detection, SVM classification does not use a target. Capgemini uses an anomaly detection solution powered by Google Cloud Platform [20] which makes it possible to find out threat due to presence of malicious users in advance. With a combination of existing rules-based model and advanced unsupervised machine learning capabilities this is achieved. And provide clients with a more robust and comprehensive solution to track anomalies in run time.

3.8 Building a classifier

The success of a classifier meant to be used in practice depends ultimately on its ability to perform well not only with the data used for its testing, but most of all with real-world data. Although a methodologically sound procedure ensures that the classifier is tested against data it has never seen before (i.e., in the training phase), there is no guarantee whatsoever that testing data will be representative of, and resemble closely, real-world data. Such data surely are not available at the time of testing. On the other hand, when using a portion of the training data as a validation dataset to verify and refine the selection of model hyper-parameters, training data could “leak” some information about the validation conditions if the partitioning of data into training and validation sets is not done properly so as to ensure that instances relative to the same (or very close) conditions go in the same part. Connections recorded with the same conditions might, in fact, share some.Because attacks often occur across different tenants, various anomaly detection techniques can combine AI algorithms to analyze attack sequences that are reported on each subscription. These techniques identify the attack sequences as well as prevalent alert patterns, instead of just being incidentally associated with each other.

Chapter 4

Second Contribution

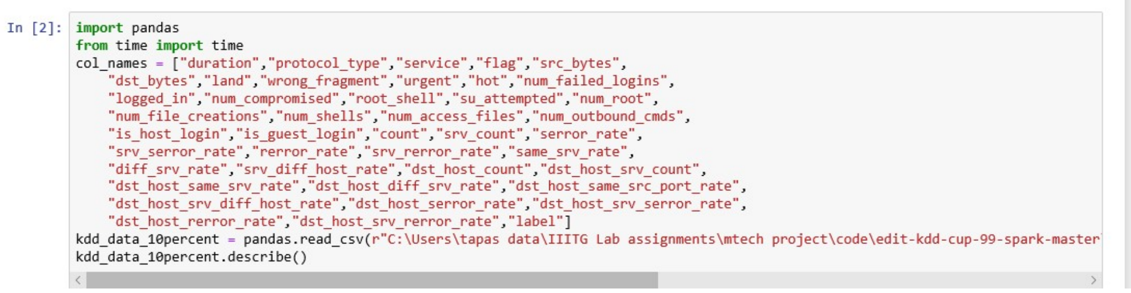
We have used KDD cup data set to implement the algorithms for anomaly detection and find vulnerabilities in cloud system. The python packages used to implement this are following

- numpy
- pandas
- scikit-learn
- matplotlib

A tool called Anaconda Navigator was used to implement this.

4.0.1 Data Description

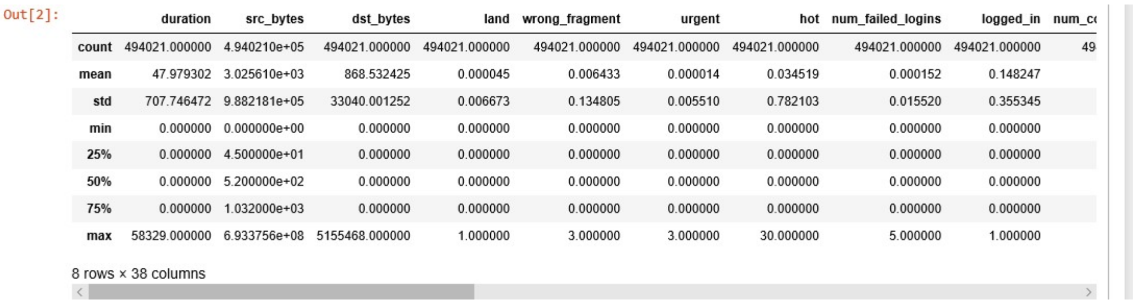
Data Files: Description of files used from data set
kddcup.name A list of features.
kddcup.data.gz The full data set (743 mb uncompressed)
kddcup.data_10percent.gz A 10% subset of original dataset.Was used to train the classifiers.
kddcup.testdata.unlabeled_10_percent.gz corrected.gz Test data with corrected labels.
training_attack_types A list of intrusion types.



```
In [2]: import pandas
        from time import time
        col_names = ["duration","protocol_type","service","flag","src_bytes",
            "dst_bytes","land","wrong_fragment","urgent","hot","num_failed_logins",
            "logged_in","num_compromised","root_shell","su_attempted","num_root",
            "num_file_creations","num_shells","num_access_files","num_outbound_cmds",
            "is_host_login","is_guest_login","count","srv_count","serror_rate",
            "srv_serror_rate","rerror_rate","srv_rerror_rate","same_srv_rate",
            "diff_srv_rate","srv_diff_host_rate","dst_host_count","dst_host_srv_count",
            "dst_host_same_srv_rate","dst_host_diff_srv_rate","dst_host_same_src_port_rate",
            "dst_host_srv_diff_host_rate","dst_host_serror_rate","dst_host_srv_serror_rate",
            "dst_host_rerror_rate","dst_host_srv_rerror_rate","label"]
        kdd_data_10percent = pandas.read_csv(r"C:\Users\tapas data\IIITG Lab assignments\mtech project\code\edit-kdd-cup-99-spark-master
        kdd_data_10percent.describe()
```

Figure 4.1: Initializing the dataset to feed in algorithm

Then the generated output which is generated by reading above variables.



Out[2]:

	duration	src_bytes	dst_bytes	land	wrong_fragment	urgent	hot	num_failed_logins	logged_in	num_c
count	494021.000000	4.940210e+05	494021.000000	494021.000000	494021.000000	494021.000000	494021.000000	494021.000000	494021.000000	49
mean	47.979302	3.025610e+03	868.532425	0.000045	0.006433	0.000014	0.034519	0.000152	0.148247	
std	707.746472	9.882181e+05	33040.001252	0.006673	0.134805	0.005510	0.782103	0.015520	0.355345	
min	0.000000	0.000000e+00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
25%	0.000000	4.500000e+01	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
50%	0.000000	5.200000e+02	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
75%	0.000000	1.032000e+03	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
max	58329.000000	6.933756e+08	5155468.000000	1.000000	3.000000	3.000000	30.000000	5.000000	1.000000	

8 rows × 38 columns

Figure 4.2: Output table generated after reading data

Now we have our data loaded into a 'Pandas' data frame. In order to get familiar with our data, let's have a look at how the labels are distributed. kdd_data_10percent['label'].value_counts() We get following attack types by reading the known attacks from the given dataset.

4.0.2 Feature selection

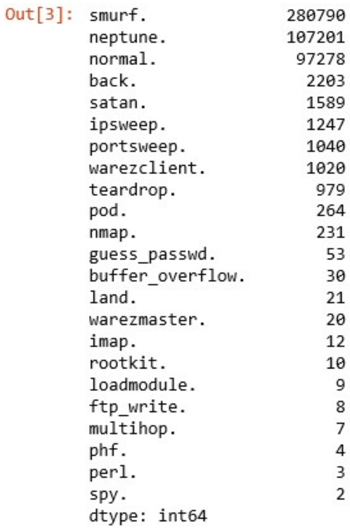
Initially, we will use all features. We need to do something with our categorical variables. For now, we will not include them in the training features. We have used following features

duration	length (number of seconds) of the connection
src_bytes	number of data bytes from source to destination
dst_bytes	number of data bytes from destination to source
land	1 if connection is from/to the same host/port; 0 otherwise
wrong_fragment	number of "wrong" fragments
urgent	number of urgent packets
hot	number of "hot" indicators

num_failed_logins	number of failed login attempts
logged_in	1 if successfully logged in; 0 otherwise
num_compromised	number of “compromised” conditions
root_shell	1 if root shell is obtained; 0 otherwise
su_attempted	1 if “su root” command attempted; 0 otherwise
num_root	number of “root” accesses
num_file_creations	number of file creation operations
num_shells	number of shell prompts
num_access_files	number of operations on access control files
num_outbound_cmds	number of outbound commands in an ftp session
is_host_login	1 if the login belongs to the “hot” list; 0 otherwise
is_guest_login	1 if the login is a “guest”login; 0 otherwise
count	number of connections to the same host as the current connection in the past two seconds
srv_count	number of connections to the same service as the current connection in the past two seconds
serror_rate	% of connections that have “SYN” errors
srv$_{s}error\_rate$	% of connections that have “SYN” errors
rerror_rate	% of connections that have “REJ” errors
same_srv$_{r}ate$	% of connections to the same service
diff_srv_rate	% of connections to different service
srv_diff_host_rate	% of connections to different hosts
dst_host_count	count of connections having same destination hosts
dst_host_srv_count	count of connections having same destination host and same service
dst_host_same_srv_rate	% of connections having the same destination host and using the same service
dst_host_diff_srv_rate	% of different services on the current host
dst_host_same_src_port$_{r}ate$	% of connections to the current host having the same src port
dst_host_srv_diff_host_rate	% of connections to the same service coming from different hosts
dst_host_serror_rate	% of connections to the current host that have an S0 error
dst_host_srv_serror_rate	% of connections to the current host and specified service that have an S0error

dst_host_rerror_rate	% of connections to the current host that have an RST error
dst_host_srv_rerror_rate	% of connections to the current host and specified service that have an RST error

Table 4.1: Description of features



```
Out[3]: smurf.              280790
        neptune.            107201
        normal.              97278
        back.                 2203
        satan.                1589
        ipsweep.              1247
        portsweep.            1040
        warezclient.          1020
        teardrop.              979
        pod.                   264
        nmap.                  231
        guess_passwd.           53
        buffer_overflow.        30
        land.                   21
        warezmaster.            20
        imap.                   12
        rootkit.                10
        loadmodule.              9
        ftp_write.               8
        multihop.                7
        phf.                     4
        perl.                    3
        spy.                     2
        dtype: int64
```

Figure 4.3: attack types

Chapter 5

Conclusion

Bibliography

[1] Diogo AB Fernandes, Liliana FB Soares, João V Gomes, Mário M Freire, and Pedro RM Inácio. Security issues in cloud environments: a survey. *International Journal of Information Security*, 13(2):113–170, 2014.

[2] Michele De Donno, Alberto Giaretta, Nicola Dragoni, Antonio Bucchiarone, and Manuel Mazzara. Cyber-storms come from clouds: Security of cloud computing in the iot era. *Future Internet*, 11(6):127, 2019.

[3] David S Linthicum. Connecting fog and cloud computing. *IEEE Cloud Computing*, 4(2):18–20, 2017.

[4] Jian Shen, Tianqi Zhou, Debiao He, Yuexin Zhang, Xingming Sun, and Yang Xiang. Block design-based key agreement for group data sharing in cloud computing. *IEEE Transactions on Dependable and Secure Computing*, 2017.

[5] Syed Noorulhassan Shirazi, Antonios Gouglidis, Arsham Farshad, and David Hutchison. The extended cloud: Review and analysis of mobile edge computing and fog from a security and resilience perspective. *IEEE Journal on Selected Areas in Communications*, 35(11):2586–2595, 2017.

[6] Constantinos Kolias, Georgios Kambourakis, Angelos Stavrou, and Jeffrey Voas. Ddos in the iot: Mirai and other botnets. *Computer*, 50(7):80–84, 2017.

[7] Daniel Chepenko. A density based algorithm for outlier detection. `https://towardsdatascience.com/density-based-algorithm-for-outlier-detection-8f278d2f7983`.

[8] Microsoft. Discover and manage shadow it in your network. `https://docs.microsoft.com/en-us/cloud-app-security/tutorial-shadow-it`.

[9] Microsoft. Microsoft way of detecting threats. `https://docs.microsoft.com/en-us/azure/security-center/security-center-alerts-overview`.

[10] Microsoft. Advanced multistage attack detection in azure sentinel. `https://docs.microsoft.com/en-us/azure/sentinel/fusion`.

[11] Cloud smart alert correlation in azure security centre. `https://docs.microsoft.com/en-us/azure/security-center/security-center-alerts-cloud-smart`.

[12] Amazon. Amazon guard duty. `https://aws.amazon.com/guardduty/`.

[13] Amazon. What is amazon cloud watch events. `https://docs.aws.amazon.com/AmazonCloudWatch/latest/events/WhatIsCloudWatchEvents.html`.

[14] Google. Forsetti intelligent agents. `https://cloud.google.com/solutions/partners/forseti-firewall-rules-anomalies`.

[15] S. Roschke, F. Cheng, and C. Meinel. Intrusion detection in the cloud. In *2009 Eighth IEEE International Conference on Dependable, Autonomic and Secure Computing*, pages 729–734, Dec 2009.

[16] Arif Sari. A review of anomaly detection systems in cloud networks and survey of cloud security measures in cloud storage applications. *Journal of Information Security*, 06(02):142–154, 2015.

[17] Liu J.G. Pannu, H.S. and S. Fu. Aad: Adaptive anomaly detection system for cloud computing infrastructures.

[18] Y. Dhanalakshmi and I. Ramesh Babu. Intrusion detection using data mining along fuzzy logic and genetic algorithms. *International Journal of Computer Science & Security*, 8:27–32, 2008.

[19] Wun-Hwa Chen, Sheng-Hsun Hsu, and Hwang-Pin Shen. Application of svm and ann for intrusion detection. *Computers & Operations Research*, 32(10):2617 – 2634, 2005. Applications of Neural Networks.

[20] Capgemini. Anomaly detection with machine learning powered by google cloud. `https://www.capgemini.com/resources/anomaly-detection-with-machine-learning-powered-by-google-cloud/`.

Author's Biography

Tapas received his B. Tech & MBA dual degree in Information Technology from ABV-IIITM, in 2009. He has been pursuing M. Tech at the Department of Computer Science and Engineering, IIIT Guwahati, since July 2018.