Global Information Assurance Certification Paper

Cloud Security Monitoring

GIAC GSNA Gold Certification

Author: Balaji Balakrishnan, pingbalaji@gmail.com
Advisor: Rajat Ravinder Varuni
Accepted: Mar 8th, 2017

Abstract

This paper discusses how to apply security log monitoring capabilities for Amazon Web Services (AWS) Infrastructure as a Service(IaaS) cloud environments. It will provide an overview of AWS CloudTrail and CloudWatch Logs, which can be stored and mined for suspicious events. Security teams implementing AWS solutions will benefit from applying security monitoring techniques to prevent unauthorized access and data loss. Splunk will be used to ingest all AWS CloudTrail and CloudWatch Logs. Machine learning models are used to identify the suspicious activities in the AWS cloud infrastructure. The audience for this paper are the security teams trying to implement AWS security monitoring.

1 Introduction

Organizations are starting to use cloud computing to take advantage of the many benefits it provides such as cost savings, quick time-to-market and on-demand scaling of the environment. As organizations start to use cloud computing, security professionals must update their operations to align with cloud computing models. The References section in this paper provides many recommendations on cloud security controls from NIST, cloud deployment models, cloud security references from Cloud Security Alliance, ENISA, and NIST.

In the most recent edition of the Cloud Computing Top Threats in 2016, the report(CIS, 2016) identified 12 critical issues to cloud security. Effective security monitoring mitigates some of the following risks:

- Weak Identity, Credential, and Access Management
- Insecure APIs
- Account Hijacking
- Malicious Insiders
- Advanced Persistent Threats (APTs)
- Data Loss
- Abuse and Nefarious Use of Cloud Services

Securing Cloud Services involves conducting a detailed risk assessment and architecting a secure solution to meet the business requirements. Security Monitoring plays a vital role in securing Cloud Services. This paper highlights how to implement security monitoring solution for Amazon Web Services(AWS) environments.

1.1 Cloud Security Monitoring Challenges

The primary types of cloud computing solutions are Infrastructure-as-a-Service (IaaS), Platform-as-a-Service (PaaS) and Software-as-a-Service(SaaS). Amazon Web Services(AWS) has established itself as a leading cloud services provider, with Microsoft Azure and Google Cloud in the distant second and third position respectively.

AWS innovates at a rapid pace introducing many new features and/or services every day, last year alone, on an average AWS customers got access to three new features every day. The different AWS services available and the best practices for securing AWS environments are in the Reference section. Some of the best practices include encryption, privileged access management, segregation of resources and monitoring.

This paper focuses on implementing security monitoring for AWS workloads. The next subsections highlight the key areas of security monitoring when deploying AWS workloads in addition to traditional data center monitoring.

AWS Management Console Monitoring

Management of AWS instances and resources are performed using the AWS Management Console. Some of the main activities that can be conducted using the AWS Management Console are creating new virtual machines and removing any existing virtual machines and other AWS services. Monitoring the unauthorized access to AWS Management Console is critical since gaining this convenient access to the cloud management plane is like having keys to the cloud kingdom.

Application Programming Interface(API) Access Monitoring

As organizations move towards cloud solutions, they have to adapt to the new DevOps architecture. Realizing the benefits of cloud platform would be difficult if teams shift and move the current applications as-is to the cloud. The existing application infrastructure has to be rearchitected to suit the cloud deployment models. Ideally, cloud solutions use the DevOps methodology for continuous deployment. This method enables the business to reduce development time and turn around quick solutions. As an example, some AWS environments use AWS CodePipeline for continuous application deployment using DevOps strategies in AWS environments.

DevOps introduces new challenges for security monitoring. The number of API calls is increasing due to automation related to AWS CodePipeline, infrastructure as code and serverless computing. It is critical to monitor these API calls to ensure there is no unauthorized access. It 's hard to follow these events using traditional rule and threshold-based monitoring due to the high volume of activities. Machine learning techniques are well suited for monitoring this vast amount of activity by learning different features/characteristics from the data.

AWS Serverless Computing Monitoring

Recently AWS introduced “serverless” computing; serverless computing depends on AWS Lambda to run the application code. In serverless computing, there is no server infrastructure; the focus is on monitoring the AWS Lambda function executions, invocations and other parameters related to the AWS Lambda functions.

AWS Identity and Access Management(IAM) Monitoring

AWS IAM enables organizations to control access to AWS services and specific resources. AWS IAM provides options to configure granular permissions in AWS environments. It is recommended to give the least amount of permissions to manage AWS resources required for performing the job function. As an effective information security control, security teams should use many tools provided by AWS like Access Advisor. Providing appropriate access prevents any unauthorized access and enables effective monitoring of AWS resources administrative access. Monitoring the different administrative credentials used in the AWS environments is a requirement enforced by various compliance regulations. Machine learning is ideal for controlling the various AWS credentials since it learns from the previous events and understands what is normal to identify anomalies. Financial regulatory requirements like Sarbanes-Oxley mandate organizations to review all privileged access and changes to the AWS environments hosting financial data as part of security compliance monitoring.

1.2 Overall architecture of the proposed solution

The proposed solution for cloud security monitoring is to use a big data analytics solution like Splunk, Apache Spark or Amazon Elasticsearch to load all the AWS cloud infrastructure logs. Machine learning models should be used to develop risk scores to identify the most suspicious events. Then, based on the incident, the security team should take further action using automation(lambda functions) (or) email to alert the security team for manual analysis.

It is a challenge to manually baseline and configure AWS infrastructure security monitoring rules due to the changes in AWS environments. Machine learning techniques like Supervised Learning algorithms explained later in this paper can handle the security monitoring challenges of cloud security monitoring by automatically learning from data to understanding anomalies and high-risk events. Machine learning models can be used to build baselines and develop risk scores to identify suspicious events using identity authentication information, location information and activity type.

In this paper, Splunk will be used to ingest all AWS CloudTrail and CloudWatch Logs for implementing the AWS security monitoring use cases. Machine learning models are applied to identify the suspicious activities in the AWS cloud infrastructure. The latest version of Splunk 6.5 has a built-in machine learning toolkit which supports various machine learning algorithms. Machine learning models will be applied using the Splunk Machine Learning toolkit. The steps involved in using machine learning algorithms are as follows:

a) Visualize and combine data cleansing with smart feature engineering,

b) Choose right metric/method for estimating model performance

c) Tune the parameters.

Summary of the key concepts proposed are:

1) Collect all of the AWS log data from Cloudtrail and CloudWatch to Splunk

2) Apply machine learning models to build baselines and develop the risk scores instead of manual rules/thresholds.

Some factors which make this implementation feasible are:

a) Advancement in big data technologies which enables information security teams to store all types of data at scale.

b) Many machine learning solutions are becoming available like Microsoft Azure ML Studio, Amazon Machine Learning, Databricks Spark, Splunk Machine Learning toolkit.

By having a centralized open source big data analytics solution, security teams can apply machine learning and other statistical techniques to any data set. The major advantage of this

solution is that once a successful method is identified using machine learning, similar challenges can be solved using the same approach. For example, if a technique is helpful in identifying suspicious access attempts from AWS cloud-based infrastructure identity and access authentication data, the same method can be applied to identify suspicious access attempts for other applications and cloud infrastructures like Microsoft Azure and Google Cloud. The next section describes machine learning techniques and the two use cases are implemented using Splunk.

1.3 Risk Scoring Methodology

Risk scoring is not a new concept; it has always been in use in the information security community to prioritize the most critical vulnerabilities and issues to resolve. In traditional data center monitoring, the risk scoring methodology relies on understanding the corporate environment to identify suspicious events. An example of this type of process is creating an unauthorized access alert to critical server asset events based on an understanding of authorized administrators who have access. Detecting malicious events based on the known bad patterns and assigning risk scores to known bad patterns is useful for threats which are already seen and known to the information security community. The References section has some examples for developing risk scores manually using static rules and thresholds in AWS environments.

The challenge with these types of standard risk scoring based monitoring is keeping up with the rapid pace of new API calls and permissions that are being rolled out by AWS. Some of the criteria that are relevant to cloud security monitoring are the identity, data access, the action performed, and geo-location. By leveraging these criteria (features) in combination with historical data, machine learning techniques can learn the environment and identify anomalies for further investigation. Machine learning models can provide risk scores based on the learning from previous data. In this paper, a Linear regression algorithm is used as an example to develop a machine learning model which predicts risk scores. Linear regression algorithms will predict numeric values. The Linear regression algorithm models the relationship between continuous output variable with the features (input, explanatory variables) using the linear function. The following section on machine learning explains the algorithm in detail. The model learns from the data; this is efficient for this AWS use case compared to manually updating the rules/thresholds for the risk scores.

1.4 Machine learning

Machine learning has two major types: Supervised and Unsupervised Learning. In Supervised Learning, the machine learning algorithm will learn from the data and labels(classification) provided. The resultant model will try to predict the label (classification) given a set of features (Astroml, 2015). Some Classification Algorithms commonly used are Neural Networks, Random Forests, Support Vector Machines (SVM), Decision Trees, Logistic Regression, and Naive Bayes. An example of Supervised Learning is providing a set of dog and cat pictures to machine learning algorithm with labels indicating if the picture is cat or dog. The Supervised Learning algorithm will learn from the dog and cat pictures and create a predictive model. Applying new pictures to the predictive model will predict if the provided picture is a dog or a cat as seen in Figure 1:

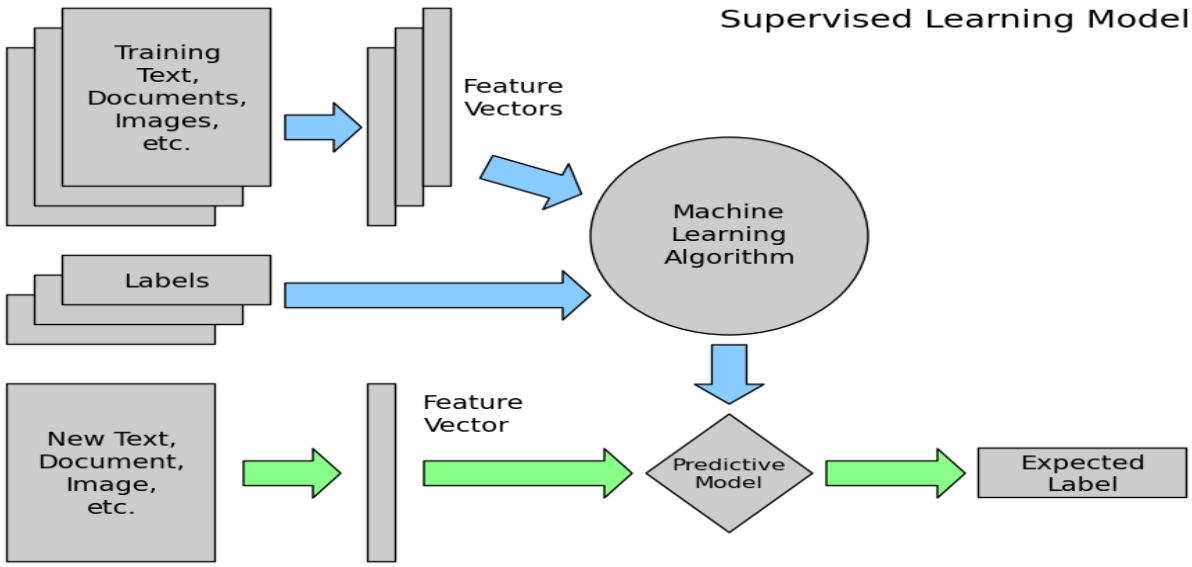


Figure 1- Supervised Learning Model (AstroML, 2015)

In this paper, a Supervised Learning technique using a linear regression algorithm will be used to predict risk scores for AWS cloud infrastructure events.

In Unsupervised Learning, the model tries to understand the data based on the features without any labels and the tasks are to identify patterns and anomalies from data. Unsupervised Learning comprises tasks such as dimensionality reduction, clustering, and density estimation (AstroML, 2015). An example of Unsupervised Learning is providing a set of dog and

cat pictures to the machine learning algorithm; it will cluster the cat and dog pictures as separate groups as depicted in Figure 2:

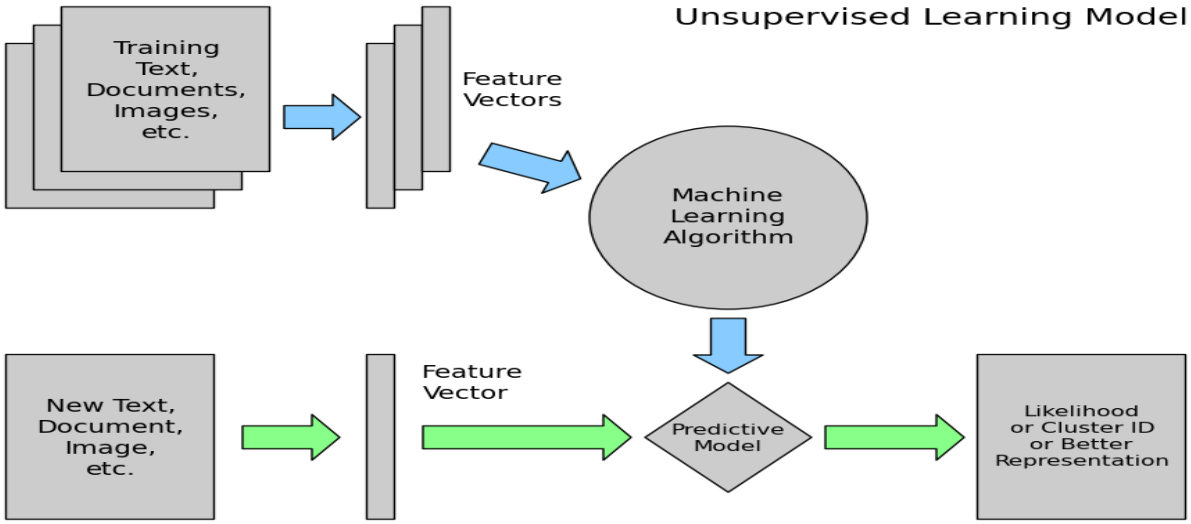


Figure 2 - Unsupervised Learning Model

Unsupervised Learning algorithms will be useful to identify the principal features in the dataset. It is also very helpful to provide different vantage points based on various features. In the example of dog and cat pictures, using Unsupervised Learning techniques will be useful to understand how the several facial features will be the most helpful to classify by segregating data based on those facial features. In our use case of AWS cloud infrastructure events, separating the data based on the location of logins can provide insight on whether it is an important feature. Some Common Unsupervised algorithms are K-Means Clustering, Hierarchical clustering, and Hidden Markov models. Figure 3 highlights the different algorithms in the Splunk Machine

Learning toolkit (Splunk, 2016):

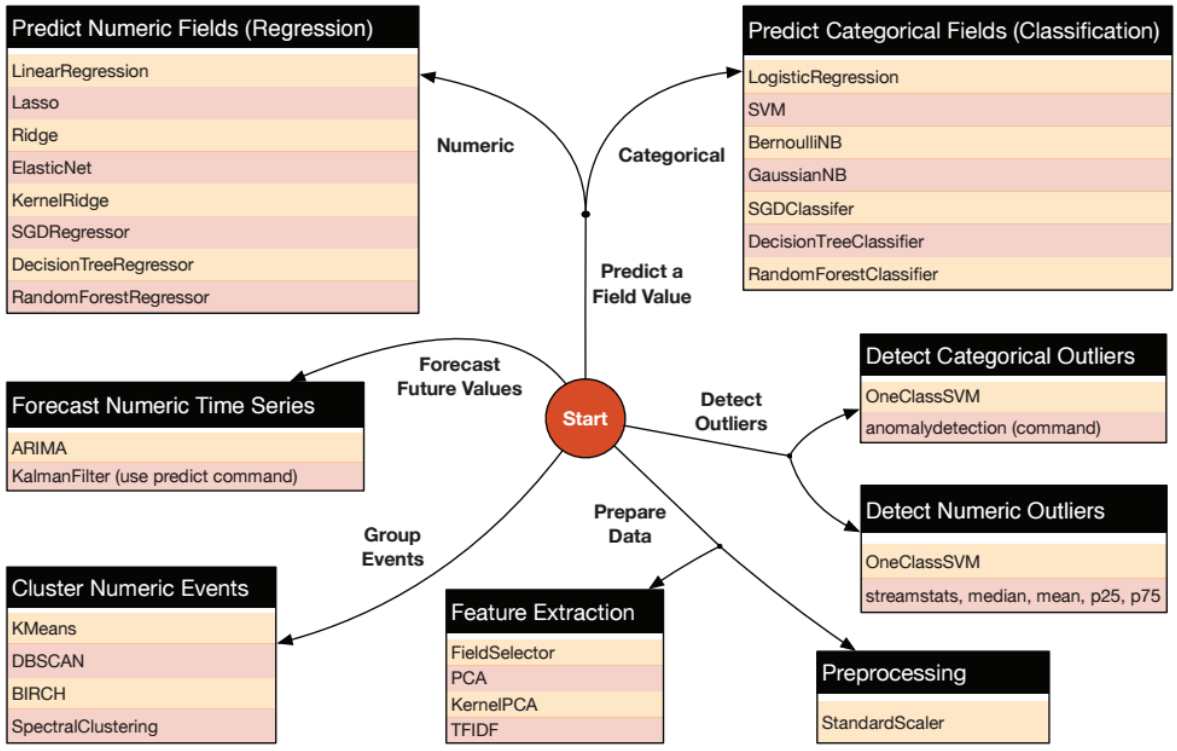


Figure 3 Splunk Machine Learning Algorithms

Machine learning should only be applied to use cases that are applicable and produce results. Machine learning is very data hungry and ingesting a lot of data for creating machine learning models will produce definitive results. Since the machine learning algorithms require a lot of data to provide useful models, significant patience is needed to obtain results. There are a lot of logs generated in AWS environments; creating the models with proper algorithms, and with a large amount and variety of data will ensure there are no overfitting problems. Another aspect to keep in mind is every AWS environment is different. As an example, most of the environments use different AWS Virtual Private Cloud(VPC) configurations to segregate AWS resources according to specific business needs. Creating machine learning models with data from the same AWS environments will produce the best results.

Limited use cases have been using Machine Learning for a long time, but recent developments in technologies (like Splunk and Apache Spark) make it feasible to deploy machine learning algorithms quickly over different data types and use features from many different data sets.

2.0 Lab Setup

Figure 4 highlights the logical structure of the lab created for this paper:

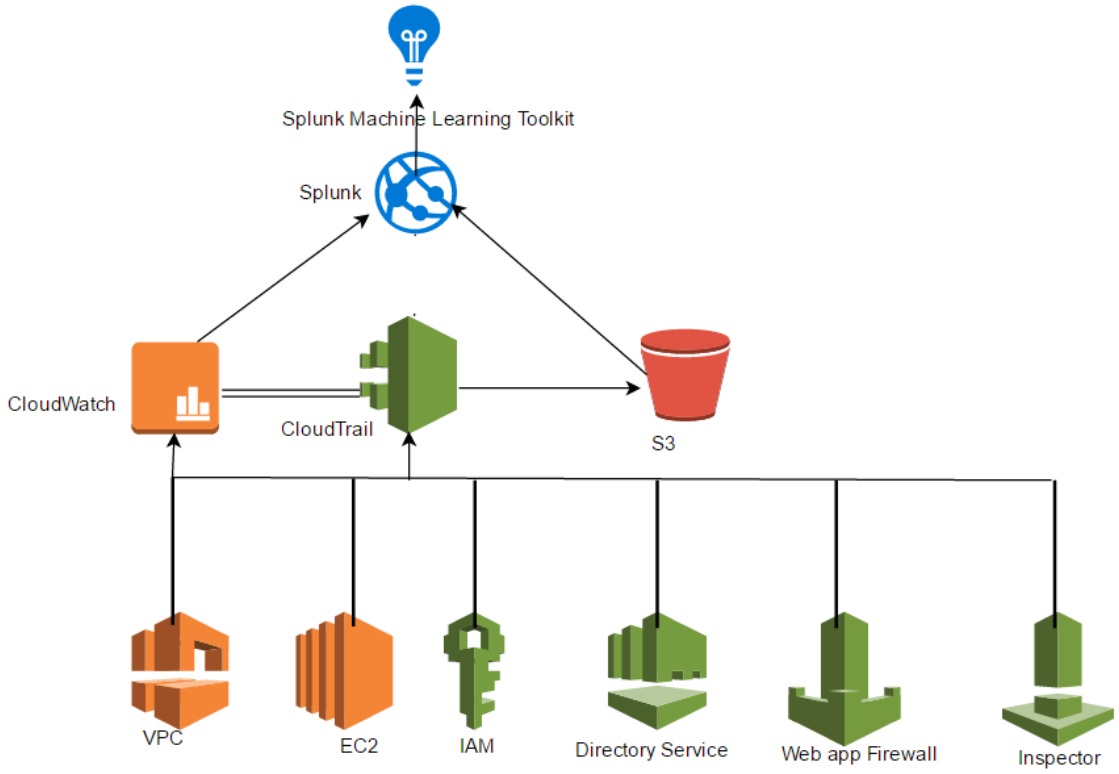


Figure 4 Lab Setup

Appendix A explains all the steps followed for the initial lab setup. In the lab configuration, Splunk is configured to receive logs from AWS Cloudtrail. The Splunk Machine Learning Toolkit is installed and set.

For further testing, generated additional logs by adding users, adding instances, launching new environments from AWS QuickStart. With the on-going collection of logs, machine learning examples will be applied in next section to calculate risk scores and detect suspicious events.

3.0 Machine Learning - Process

The steps below highlight the methodology for applying machine learning techniques using Splunk Machine Learning Toolkit. The same process can be used with any machine

learning solution and can apply to any security monitoring use case. This solution can also be implemented using Apache Spark MLLib libraries. One challenge is parsing and normalizing the AWS Cloudtrail JSON data files in Apache Spark. AWS has released the open source code to convert AWS CloudTrail logs to a Spark Data Frame (Github, 2016). After loading the data to Apache Spark data frames, the data can be used by Apache Spark MLLib libraries.

One aspect to remember during machine learning is data cleansing. Data cleansing ensures that the data is consistent and uniform. In many cases, the data should be extracted and formatted before being fed to machine learning algorithms. Splunk inherently addresses the data cleansing by indexing data at ingestion time and extracts relevant fields and provides a natural mapping from JSON format into standard columns. Machine learning algorithms can use the data directly from these columns. Splunk saves a lot of time in data cleansing and formatting when compared to many open source solutions like Apache Spark. Figure 5 below highlights the steps involved in the machine learning process:

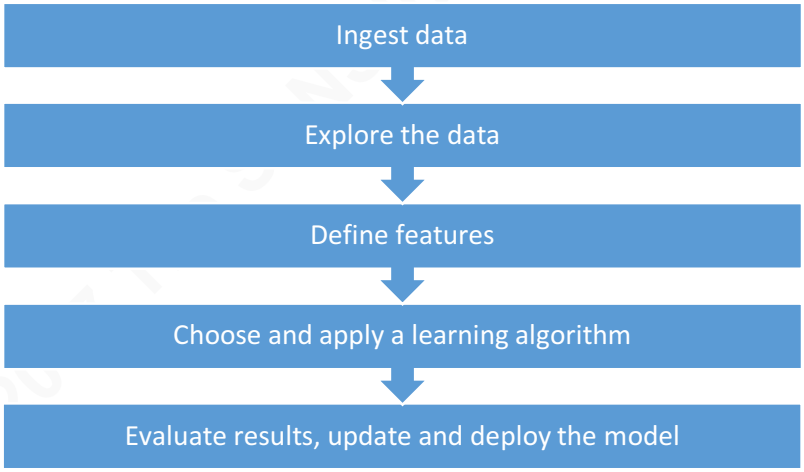


Figure 5 Machine Learning Process

3.1 Ingest data to Splunk and understand the data

Security teams must collect all the AWS logs in a central place. Even if the organization could not implement any active monitoring, the logs will be useful for forensic analysis at a later point when an incident occurs.

In the initial lab setup, Splunk was configured to ingest data from Cloudtrail and Cloudwatch logs. The Splunk AWS app can be used to explore and understand the log events to identify features for machine learning.

3.1.1 AWS CloudTrail

AWS Cloudtrail creates logs of all the API access requests, AWS resources access and AWS console login access information. It is important to understand the AWS Cloudtrail log data to efficiently design features for machine learning algorithms. AWS Cloudtrail User Guide (AWS, 2014) provides excellent reference and examples of different types of log events. The Cloudtrail API Call log contains two parts, Record Body Contents, and userIdentity Element. The types of events analyzed are:

- aws_cloudtrail_notable_network_events
- aws_cloudtrail_iam_change
- aws_cloudtrail_errors
- aws_cloudtrail_change
- aws_cloudtrail_delete_events
- aws_cloudwatch_sns_events
- aws_cloudtrail_auth
- aws_cloudtrail_iam_events
- aws_cloudtrail_ec2_events

3.2 Explore the data

In this particular use case, the Splunk App for AWS can be used to study the AWS log data. The Splunk App for AWS gives critical operational and security insight into the Amazon Web Services account. Figure 6 below shows different dashboard options available on the Splunk App for AWS. These panels can be used to understand the relevant AWS logs which would help in determining any suspicious activity. The scroll down highlights the different security dashboards available.

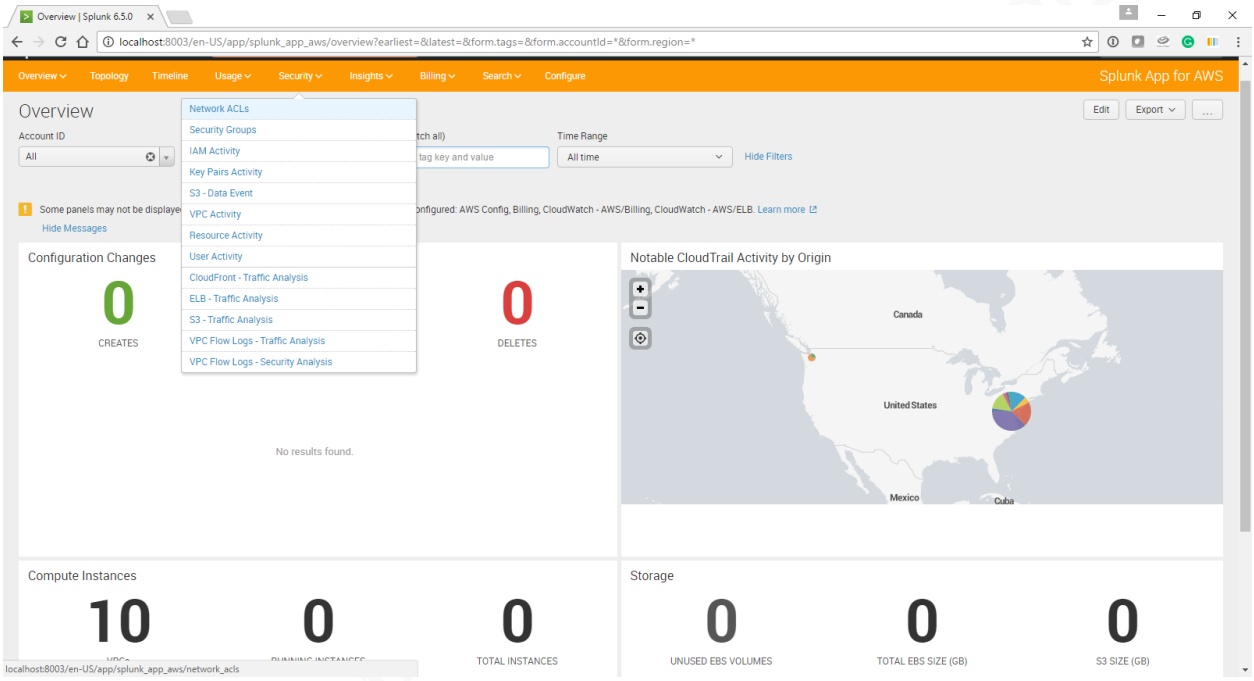


Figure 6 Splunk App for AWS Dashboards

The dashboard below in Figure 7 highlights various user activities in the AWS environment. Looking at the data using different fields(features) will help the security team to understand the relevant fields in the log data.

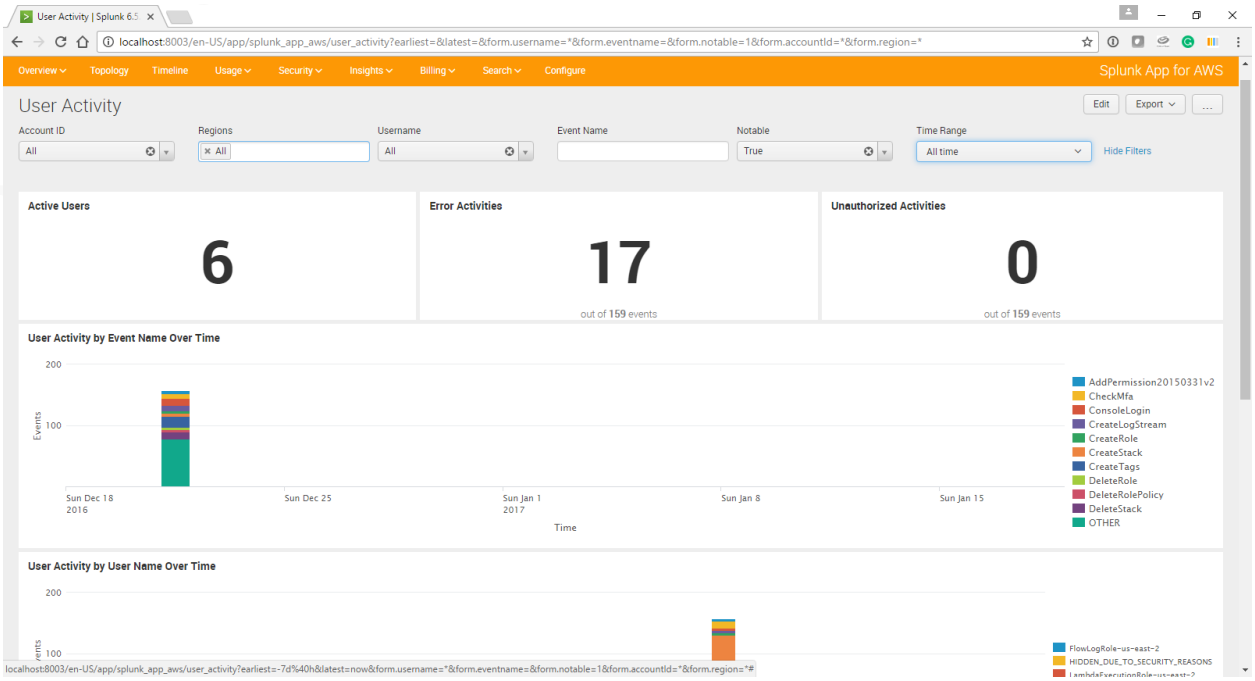


Figure 7 Splunk App for AWS Security Dashboard - User Activities

The Splunk App for AWS allows security practitioners to understand and explore the data to determine the fields that will be helpful in identifying the suspicious AWS activities. The Splunk Machine Learning Toolkit can use these fields as features while developing the model.

3.3 Use Case 1 – Detecting Suspicious AWS Console Logins

3.3.1 Define features

In this case study, the "AwsConsoleSignIn" events were explored using domain expertise on AWS Cloud security with the goal of understanding which fields will be beneficial to determine any suspicious login to AWS Console.

Some of the relevant fields identified are:

- sourceIPAddress
- userAgent
- userIdentity.arn
- eventTime
- responseElements.ConsoleLogin

In the above example, understanding and exploring the various logs from a security perspective enabled to identify the features. The Splunk Machine Learning Toolkit has algorithms like Principal Component Analysis (PCA) which can be used to explore and define features mathematically using the data. It will be useful to understand the data from a different vantage point which will assist with determining unusual activity.

3.3.2 Choose and apply a learning algorithm

This section highlights the Splunk Machine Learning Toolkit commands required to create, score and test the model.

AwsConsoleSignIn.csv is generated using the AWS logs from the lab environment.

Splunk command can be used to export the events with defined features as CSV:

* sourcetype="aws:cloudtrail" eventType=AwsConsoleSignIn | table sourceIPAddress, userAgent, userIdentity.arn, eventTime, responseElements.ConsoleLogin

Security professionals should add risk scores to these events. Risk scores should be assigned based on the security domain knowledge and the environment. Ideally, security professionals should review and assign risk scores to the events for a month or more depending on the environment. The Machine learning model requires a significant amount of labeled risk score data to achieve useful results.

A sample record from the AwsConsoleSignIn.csv is highlighted below in Figure 8:

sourceIPAddress	userAgent	userIdentity.arn	eventTime	responseEle	riskScore
1.1.1.1	Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/54.0.2840.71 Safari/537.36	arn:aws:iam::14990	2017-01-02T19:08:1	Success	10

Figure 8 Sample Record AWSConsoleSignIn

Create New Model – AwsConsoleSignIn

The goal of this machine learning model is to predict a risk score to identify the highest set of suspicious events. The inputs for creating this Supervised Learning Model is the data with assigned risk scores and the algorithm LinearRegression. The output will be a model which will mathematically predict the risk scores by learning from the data.

In the Machine Learning Toolkit app configuration highlighted in Figure 9, choose "Assistants -> Predict Numeric fields" and in the "Enter a search" provide the input file AwsConsoleSignIn.csv using the command:

```
| inputlookup AwsConsoleSignIn.csv
```

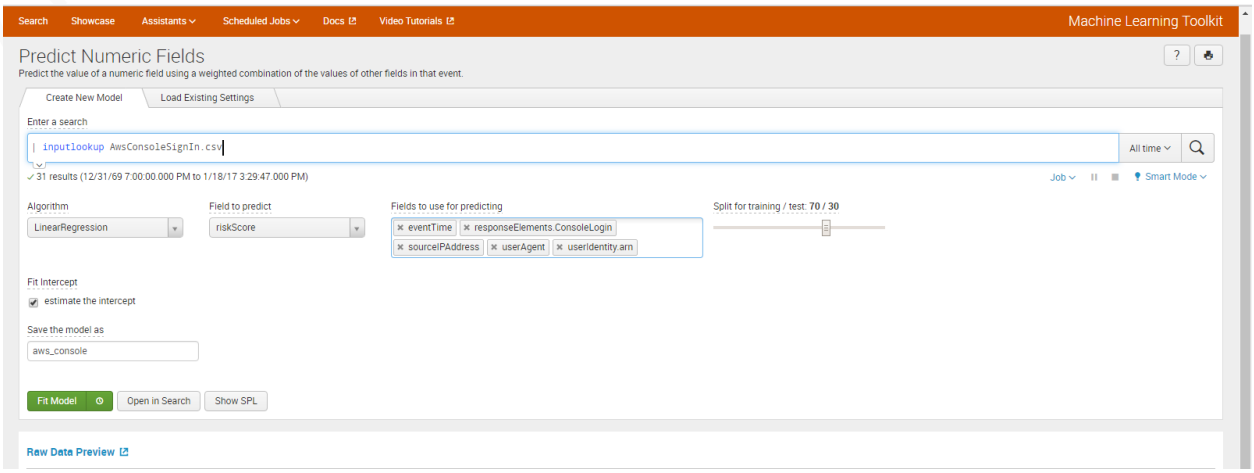


Figure 9 Splunk Machine Learning Toolkit

AwsConsoleSignIn.csv is provided as input file. The search loads all the records in the AwsConsoleSignIn.csv file for analysis. The following options are chosen to create the model:

Algorithm: LinearRegression.

Field to predict: riskScore

Fields to use for predicting: "sourceIPAddress", "userAgent", "userIdentity.arn" , "eventTime", "responseElements.ConsoleLogin"

The resultant set of Splunk commands are below:

| inputlookup AwsConsoleSignIn.csv

| fit LinearRegression fit_intercept=true "risk_score" from "sourceIPAddress", "userAgent", "userIdentity.arn" , "eventTime", "responseElements.ConsoleLogin" into "aws_console"

Supervised Learning is used in this particular example and the machine learning algorithm LinearRegression is provided with AWS log data and actual risk scores. The resultant model aws_console will try to predict the risk score given the set of features "sourceIPAddress", "userAgent", "userIdentity:arn" , "eventTime", "responseElements:ConsoleLogin".

Evaluate results and update the model

In the configuration highlighted in Figure 10, data is split 70 % for training and 30 % for testing and evaluating the model. Allocating 30 % of data for testing and evaluation of the model helps understand the accuracy of the model.

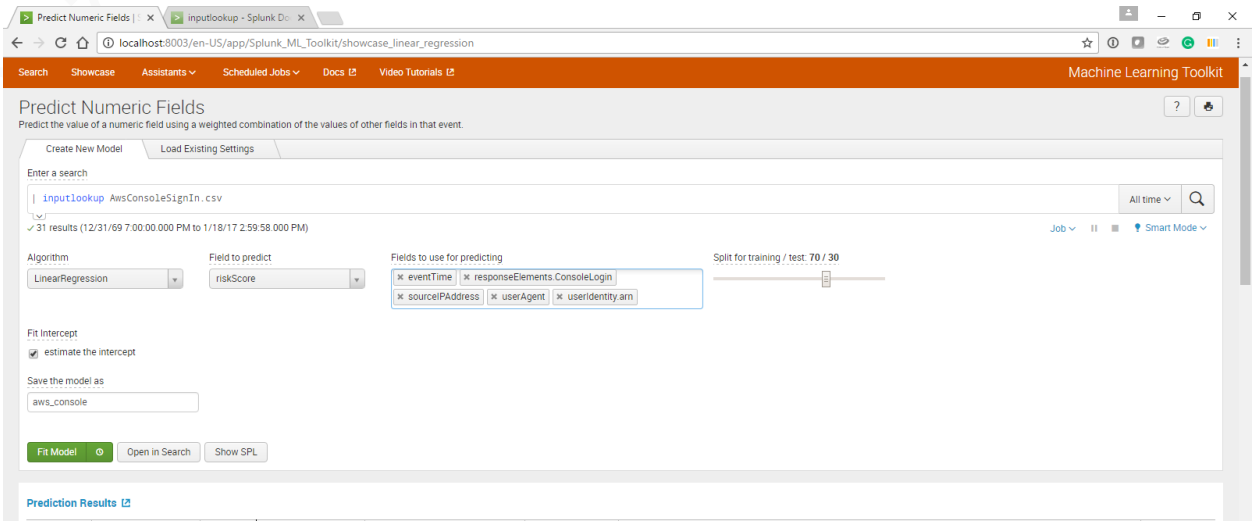


Figure 10 Splunk Machine Learning Toolkit

After fitting the model, Splunk Machine Learning Toolkit performs the necessary computations used for measuring the performance of the model.

Plot actual vs. predicted values on a line chart as depicted in Figure 11 below helps security teams to understand the efficiency of the model.

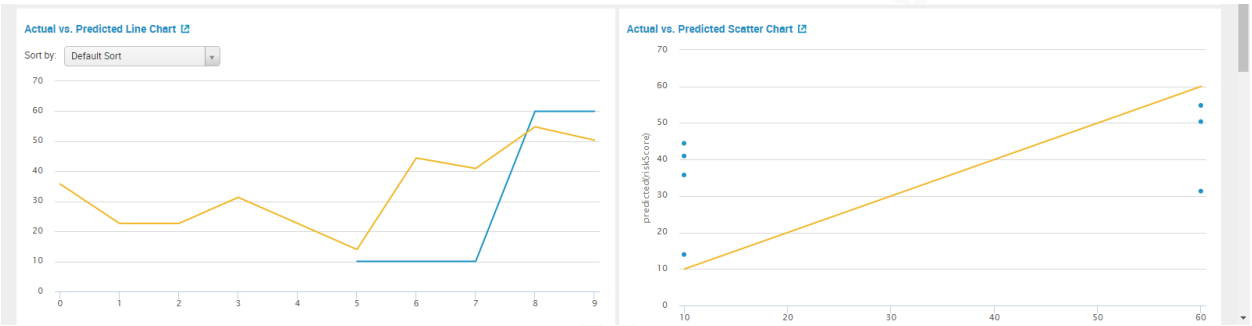


Figure 11 Splunk Machine Learning Toolkit - Actual vs. Predicted Chart

These commands will apply the model to the data set in AwsConsoleSignIn.csv and plot the actual vs. predicted values to understand the accuracy of the model.

| inputlookup AwsConsoleSignIn.csv

| apply " aws_console."

| table _time, " risk_score ", "predicted(risk_score)"

These commands will apply the model to the data set in AwsConsoleSignIn.csv to calculate the R^2 and root mean squared error (RMSE).

These values assist with measuring the accuracy of the model.

| inputlookup AwsConsoleSignIn.csv

| apply " aws_console "

| `regressionstatistics("risk_score ", "predicted(risk_score)")`

Root Mean Square Error and R^2 values provide an idea of the magnitude of the error. R^2 presents an indication of the effectiveness of a set of predictions to the actual values. The value is between 0 and 1. The value near 0 indicates the model is not a good fit, as seen in Figure 12.

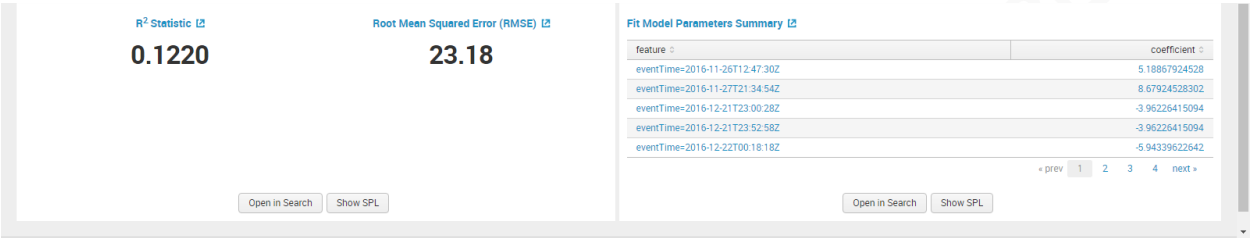


Figure 12 Splunk Machine Learning Toolkit - RMSE

After analyzing the performance, if the performance is not satisfactory, additional features can be extracted. As an example, including Geolocation data from Maxmind Geodatabase for the new context of Geolocation for the IP addresses involved will help improve the effectiveness of the model.

Once the performance is satisfactory, security team should deploy the model using the apply <model> command. After implementing the model, security analysts should continually tune the model based on the feedback from the security analysts who are analyzing the results.

3.4 Use Case 2 – Detecting Suspicious API calls

3.4.1 Define features

The "AwsApiCall" events were explored using domain expertise on AWS Cloud security with the goal of understanding which fields will be beneficial to determine any suspicious AWS API calls.

Some of the relevant fields identified are:

- sourceIPAddress
- eventSource
- eventName
- userIdentity:arn
- eventTime
- userAgent
- userIdentity:type

3.4.2 Choose and apply a learning algorithm

AwsAPICall.csv is generated using the AWS logs from the lab environment.

Splunk command can be used to export the events with defined features as CSV:

* sourcetype="aws:cloudtrail" eventType=AwsAPICall | table sourceIPAddress, eventSource , eventName , userIdentity.arn, eventTime, userAgent, userIdentity.type

Security professionals should add risk scores to these events. The risk scores should be assigned based on the security domain knowledge and the environment. A sample record from the AwsAPICall.csv is shown below in Figure 13.

sourceIPA	eventSource	eventName	userIdent	eventTim	userAgent	userIdent	riskScore
1.1.1.1	ec2.amazonaws.com	DescribeSubnets	arn:aws:ia	2017-01-1	Boto/2.39.0 Pytho	Root	20

Figure 13 Sample Record AWSAPICAll

Create New Model – AwsAPICall

The goal of this machine learning model is to predict a risk score to identify the highest set of suspicious API calls. In the Machine Learning Toolkit app, choose Assistants -> Predict Numeric fields, and in the search box provide the input file AwsConsoleSignIn.csv using the command:

| inputlookup AwsAPICall.csv

AwsAPICall.csv is provided as input file. The search loads all the records in the AwsAPICall.csv file for analysis. The following options are chosen to create the model:

Algorithm: LinearRegression.

Field to predict: riskScore

In the configuration highlighted below in Figure 14, the fields used for predicting are: "sourceIPAddress", "eventSource" , "eventName" , "userIdentity.arn", "eventTime", "userAgent", "userIdentity.type"

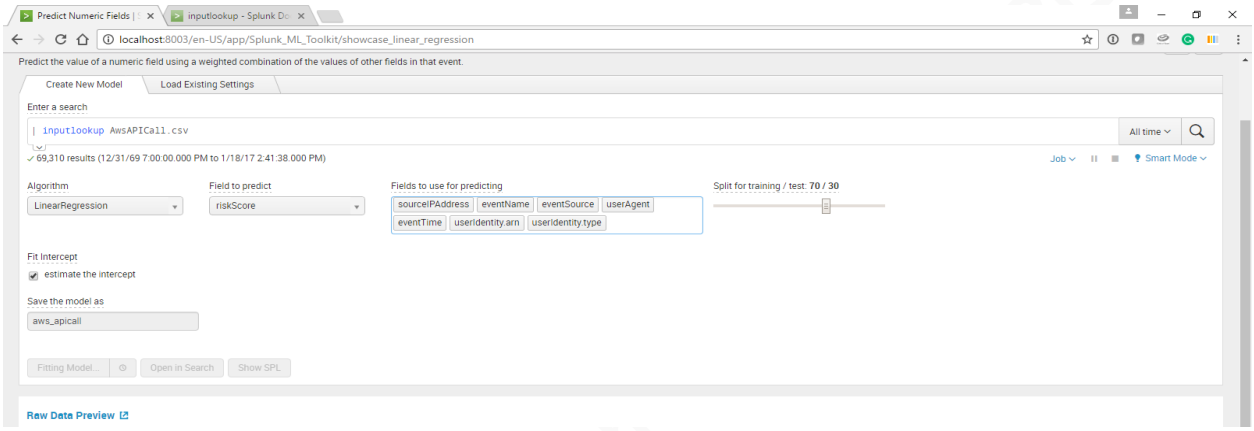


Figure 14 Splunk Machine Learning Toolkit

The resultant set of Splunk commands are below:

| inputlookup AwsAPICall.csv

| fit LinearRegression fit_intercept=true "riskScore" from "sourceIPAddress", "eventSource" , "eventName" , "userIdentity.arn", "eventTime", "userAgent", "userIdentity.type" into "aws_apicall"

Supervised learning is used in our example, the machine learning algorithm LinearRegression is provided with AWS log data and actual risk scores. The resultant model aws_console will try to predict the risk score given the set of features "sourceIPAddress", "eventSource" , "eventName" , "userIdentity.arn", "eventTime", "userAgent", "userIdentity.type"

Evaluate results and update the model

The data is split 70 % for training and 30 % for testing and evaluating the model. This helps understand the accuracy of the model.

After fitting the model, Splunk Machine Learning Toolkit performs the necessary computations used for measuring the performance of the model. Plot actual vs. predicted values on a line chart as depicted in Figure 15 below helps understand the efficiency of the model.

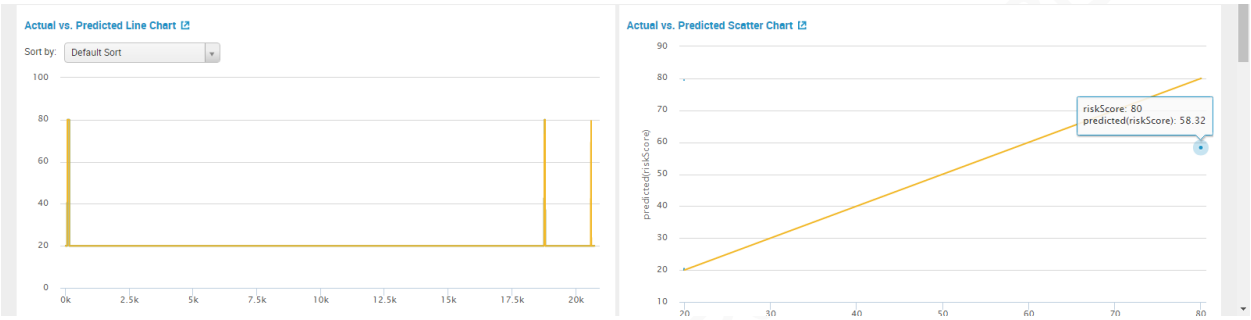


Figure 15 Splunk Machine Learning Toolkit - Plot actual vs. predicted value

Root Mean Square Error and R^2 values provide an idea of the magnitude of the error. R^2 presents an indication of the effectiveness of a set of predictions to the actual values. The value is between 0 and 1. The value near 1 indicates the model is a good fit.

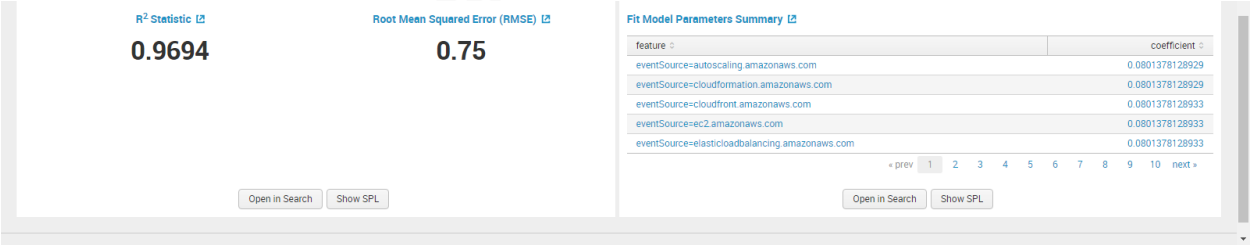


Figure 16 Splunk Machine Learning Toolkit - RMSE

After analyzing, if the performance is not satisfactory, additional features can be extracted. As an example, including AWS VPC information for the context of VPCs involved in an event will help improve the effectiveness of the model.

Once the performance is satisfactory, deploy the model using the apply <model> command. After implementing the model, security analysts should continually tune the model based on the feedback from the security analysts who are analyzing the results.

This section highlights how Splunk Machine Learning toolkit could be used to create, evaluate and deploy Machine Learning models. The two models generated are examples with very few data records created in the lab to prototype the Splunk Machine Learning toolkit functionality. Security teams should test, tune and deploy the Machine Learning models according to the AWS environment.

There are only two use cases discussed in this paper. Another practical use case that might be useful is determining anomalous network traffic sessions between various AWS VPCs. Threat modeling with inputs from adversary Tools, Techniques, and Procedures(TTPs) can be

used to identify additional security monitoring use cases in AWS environments. After determining the use case, the methodology discussed in this paper can be used to evaluate features and apply Machine Learning model to new use cases.

Machine learning is very data hungry and ingesting a lot of data for creating machine learning models will produce useful results. Also, if multiple data sources are used to extract features, greater fidelity can be achieved. As an example, including Geolocation data for the additional context of the IP addresses involved will help improve the effectiveness of the model.

4 Conclusion

This paper highlights how to implement machine learning techniques for AWS logs. Machine learning techniques were applied in this paper to identify suspicious events in IaaS environments. Identity is the new perimeter and using machine learning techniques to identity data in combination with other telemetry data will help security professionals identify suspicious events. As a first step, the security team members should understand the monitoring requirements, understand the data, evaluate the suitable methods. The security team should consider if machine learning is appropriate for the nature of the logs, explore, visualize and select the features as inputs to creating the model. After building and testing the model, the security team should apply the model to real-time traffic(data). After using the model, the security team should periodically evaluate the results and tune the model.

Many machine learning solutions are becoming available like Microsoft Azure ML Studio, Amazon Machine Learning, Databricks Spark, Splunk Machine Learning toolkit. All of these machine learning tools make the implementation of machine learning models very intuitive and easy to implement with simple user interfaces. These user interfaces encapsulate the mathematics and coding involved in traditional machine learning application languages like R.

Using Amazon Machine Learning for security monitoring was explained with demos in the AWS re Invent 2016 Conference (Videos from re Invent 2016 security and compliance sessions, 2016). As the cloud implementations evolve, the security teams should also learn the advantages and new ways of implementing security operations and security monitoring activities. Automation and machine learning are two key areas in the cloud that give an edge to defenders.

A defender has to detect only one of the attacker's activity before successful completion of attacker's objectives. As defenders, the goal is to deploy defense in depth strategy by placing preventive and detective controls at every layer to introduce high cost for an attacker to achieve his objectives. Machine Learning can be one useful tool in the defense in depth strategy to detect suspicious activity. After identification of the suspicious activity, using forensics, the security teams could be able to track and trace any activity performed by the attacker and take remediation actions.

Some other use cases that might benefit from this solution are Risk Management, Security Automation/Orchestration, User/Network Behavior Analytics, Fraud Detection, Threat Hunting, Threat Intelligence aggregation from various sources, and Incident Response/Forensic Analysis.

References

The NIST Definition of Cloud Computing. (n.d.). Retrieved January 01, 2017, from http://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-145.pdf

US Government Cloud Computing Technology Roadmap. (n.d.). Retrieved January 1, 2017, from http://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.500-293.pdf

Cloud Computing Synopsis and Recommendations. (2012). Retrieved January 1, 2017, from http://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-146.pdf

Cloud Computing Risk Assessment. (n.d.). Retrieved January 01, 2017, from http://www.enisa.europa.eu/act/rm/files/deliverables/cloud-computing-risk-assessment

NIST Cloud Computing. (n.d.). Retrieved January 01, 2017, from http://csrc.nist.gov/groups/SNS/cloud-computing/

Cloud Security Alliance. (n.d.). Retrieved January 01, 2017, from http://www.cloudsecurityalliance.org/

Cloud Computing Top Threats in 2016. (2016). Retrieved January 1, 2017, from https://downloads.cloudsecurityalliance.org/assets/research/top-threats/Treacherous-12_Cloud-Computing_Top-Threats.pdf

(2016,). AWS Well-Architected Framework. Retrieved January 4, 2017, from http://d0.awsstatic.com/whitepapers/architecture/AWS_Well-Architected_Framework.pdf

(2016,). Hardening AWS Environments and Automating Incident Response for AWS Compromises. Retrieved January 4, 2017, from https://s3-us-west-2.amazonaws.com/threatresponse-static/us-16-Krug-Hardening-AWS-Environments-and-Automating-Incident-Response-for-AWS-Compromises-wp.pdf

(2016,). Cloud security monitoring: Challenges and guidance. Retrieved January 4, 2017, from http://searchcloudsecurity.techtarget.com/tip/Cloud-security-monitoring-Challenges-and-guidance

(2016,). About the Splunk App for AWS. Retrieved January 4, 2017, from http://docs.splunk.com/Documentation/AWS/latest/Installation/Abouttheapp

(2016, August). Amazon Web Services: Overview of Security Processes. Retrieved January 4, 2017, from https://d0.awsstatic.com/whitepapers/aws-security-whitepaper.pdf

(2015, October 6). Splunk App for AWS: Making the invisible, visible. Retrieved January 4, 2017, from http://blogs.splunk.com/2015/10/06/splunk-app-for-aws/

(2016, July). Best Practices for Managing Security Operations in AWS. Retrieved January 4, 2017, from http://www.slideshare.net/AmazonWebServices/best-practices-for-managing-security-operations-in-aws-aws-july-2016-webinar-series

(2014,). Use Your AWS CloudTrail Data and Splunk Software To Improve Security and Compliance in AWS. Retrieved January 4, 2017, from http://www.slideshare.net/AmazonWebServices/aws-partner-webcast-use-your-aws-cloudtrail-data-and-splunk-software-to-improve-security-and-compliance-in-aws

(2016,). Add a CloudTrail input for the Splunk Add-on for AWS. Retrieved January 4, 2017, from http://docs.splunk.com/Documentation/AddOns/released/AWS/CloudTrail

(2015, September). AWS Cloudtrail Splunk. Retrieved January 4, 2017, from https://github.com/xueshanf/aws-cloudtrail-with-splunk

(2016, June). AWS Cloud Adoption Framework. Retrieved January 4, 2017, from https://d0.awsstatic.com/whitepapers/AWS_CAF_Security_Perspective.pdf

(2016,). Cloud Security Resources. Retrieved January 4, 2017, from https://aws.amazon.com/security/security-resources/

(2015, May 15). How to Receive Alerts When Specific APIs Are Called by Using AWS CloudTrail, Amazon SNS, and AWS Lambda. Retrieved January 4, 2017, from https://aws.amazon.com/blogs/security/how-to-receive-alerts-when-specific-apis-are-called-by-using-aws-cloudtrail-amazon-sns-and-aws-lambda/

(2015, February 5). How to Receive Alerts When Your IAM Configuration Changes. Retrieved January 4, 2017, from https://aws.amazon.com/blogs/security/how-to-receive-alerts-when-your-iam-configuration-changes/

(2016,). AWS Security Monitoring & Compliance Validation From Adobe. Retrieved January 4, 2017, from https://conf.splunk.com/files/2016/slides/you-cant-protect-what-you-cant-see-aws-security-monitoring-and-compliance-validation-from-adobe.pdf

(2016,). Splunk - Cloud Is a Journey. Make Splunk Your Partner. Retrieved January 4, 2017, from http://www.slideshare.net/AmazonWebServices/partner-solutions-splunk-cloud-is-a-journey-make-splunk-your-partner

(2015,). AWS July Webinar Series - Troubleshooting Operational and Security Issues in Your AWS Account using CloudTrail. Retrieved January 4, 2017, from http://www.slideshare.net/AmazonWebServices/july-webinar-series-troubleshooting-operational-and-security-issues-in-your-aws-account-using-cloud-trail-20150729

(2015). (SEC318) AWS CloudTrail Deep Dive. Retrieved January 4, 2017, from http://www.slideshare.net/AmazonWebServices/sec318-aws-cloudtrail-deep-dive

(2015). (SEC308) Wrangling Security Events In The Cloud. Retrieved January 4, 2017, from http://www.slideshare.net/AmazonWebServices/sec308-wrangling-security-events-in-the-cloud

A. (2015). Awslabs/timely-security-analytics. Retrieved January 01, 2017, from https://github.com/awslabs/timely-security-analytics

Marko, K. (2015) AWS security management: In need of automation. Available at: http://markoinsights.com/2015/12/27/aws-security-mgmt/ (Accessed: 7 January 2017).

Cassidy, S. (2016) Solutions. Available at: https://www.defensestorm.com/cybermind/security-logging-on-aws/ (Accessed: 7 January 2017).

Chan, J. (2010) Announcing security monkey - AWS security configuration monitoring and analysis. Available at: http://techblog.netflix.com/2014/06/announcing-security-monkey-aws-security.html (Accessed: 7 January 2017).

Creating CloudWatch alarms for CloudTrail events: Examples (2017) Available at http://docs.aws.amazon.com/awscloudtrail/latest/userguide/cloudwatch-alarms-for-cloudtrail.html (Accessed: 7 January 2017).

Creating and updating your cloudtrail. (2016). Retrieved January 7, 2017, from http://docs.aws.amazon.com/awscloudtrail/latest/userguide/cloudtrail-create-and-update-a-trail.html

(2015, August). CloudWatch Logs Subscription Consumer + Elasticsearch + Kibana Dashboards. Retrieved January 4, 2017, from https://aws.amazon.com/blogs/aws/cloudwatch-logs-subscription-consumer-elasticsearch-kibana-dashboards/

(2016,). Tutorial: Using Amazon ML to Predict Responses to a Marketing Offer. Retrieved January 4, 2017, from http://docs.aws.amazon.com/machine-learning/latest/dg/tutorial.html

(2016, October). Building Event-Driven Batch Analytics on AWS. Retrieved January 4, 2017, from https://aws.amazon.com/blogs/big-data/building-event-driven-batch-analytics-on-aws/

Now available: Videos from re Invent 2016 security and compliance sessions (2016) Available at https://aws.amazon.com/blogs/security/now-available-videos-and-slide-decks-from-reinvent-2016-security-and-compliance-sessions/ (Accessed: 8 January 2017).

(2016) Available at: http://conf.splunk.com/sessions/2016-sessions.html (Accessed: 8 January 2017).

ML-SPL quick reference guide Preprocessing StandardScaler (no date) Available at https://docs.splunk.com/images/e/ee/MLTKCheatSheet.pdf (Accessed: 8 January 2017).

(2014) AWS CloudTrail user guide. Available at: http://awsdocs.s3.amazonaws.com/awscloudtrail/latest/awscloudtrail-ug.pdf (Accessed: 8 January 2017).

How to Receive Alerts When Specific APIs Are Called by Using AWS CloudTrail, Amazon SNS, and AWS Lambda. (2015). Retrieved January 01, 2017, from https://aws.amazon.com/blogs/security/how-to-receive-alerts-when-specific-apis-are-called-by-using-aws-cloudtrail-amazon-sns-and-aws-lambda/

How to Receive Alerts When Your IAM Configuration Changes. (2015). Retrieved January 01, 2017, from https://aws.amazon.com/blogs/security/how-to-receive-alerts-when-your-iam-configuration-changes/

Machine learning algorithm cheat sheet for Microsoft Azure Machine Learning Studio. (n.d.). Retrieved July 28, 2016, from https://azure.microsoft.com/en-us/documentation/articles/machine-learning-algorithm-cheat-sheet/

Binary Classification: Network intrusion detection. (n.d.). Retrieved July 28, 2016, from https://gallery.cortanaintelligence.com/Experiment/Binary-Classification-Network-intrusion-detection-2?share=1

MLlib: Scalable Machine Learning on Spark. (n.d.). Retrieved July 28, 2016, from http://stanford.edu/~rezab/sparkworkshop/slides/xiangrui.pdf

Balakrishnan, B. (2016) Applying machine learning techniques to measure critical security controls. Available at: https://www.sans.org/reading-room/whitepapers/critical/applying-machine-learning-techniques-measure-critical-security-controls-37247 (Accessed: 19 January 2017).

Balakrishnan, B. (2015) Insider threat mitigation guidance. Available at: https://www.sans.org/reading-room/whitepapers/monitoring/insider-threat-mitigation-guidance-36307 (Accessed: 19 January 2017).

Blum, D. (2016) Discovering agile cloud security | security architects partners. Available at: http://security-architect.com/discovering-agile-cloud-security/ (Accessed: 19 January 2017).

Introduction to AWS CodePipeline (2017) Available at: https://aws.amazon.com/devops/continuous-delivery/ (Accessed: 22 January 2017).

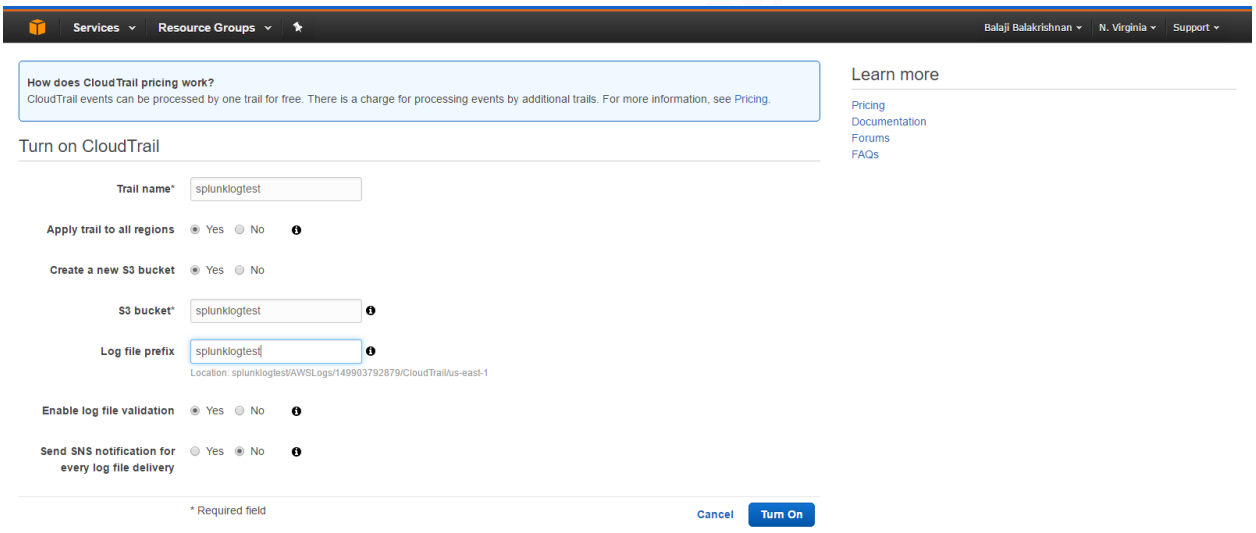
What is DevOps? - Amazon web services (AWS) (2017) Available at https://aws.amazon.com/devops/what-is-devops/ (Accessed: 22 January 2017).

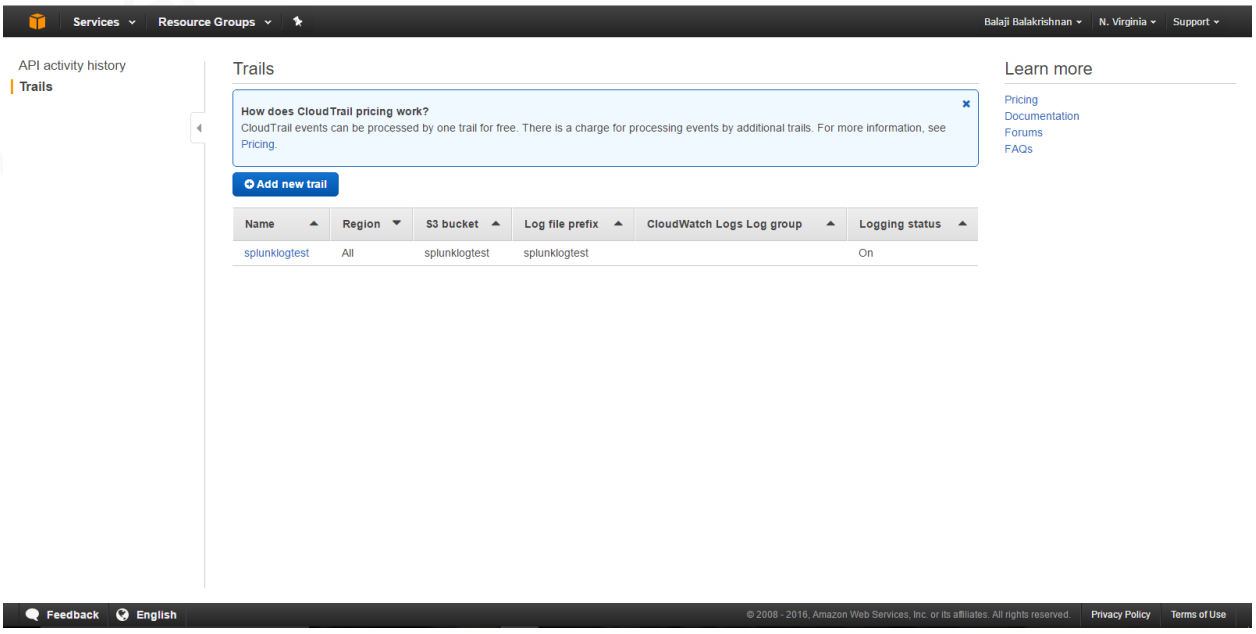
AWS Identity and Access Management (2016) Available at https://aws.amazon.com/iam/ (Accessed: 22 January 2017).

Brownlee, J. (2016) Metrics to evaluate machine learning Algorithms in python. Available at: http://machinelearningmastery.com/metrics-evaluate-machine-learning-algorithms-python/ (Accessed: 12 February 2017).

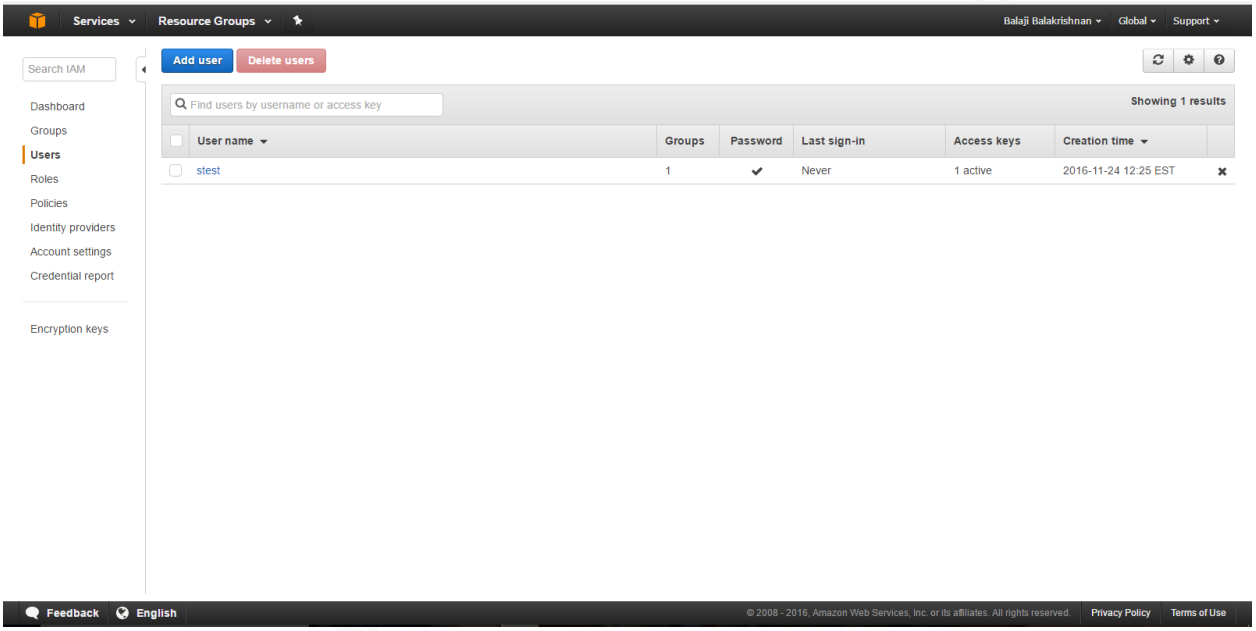
Appendix B
Lab Setup

- Build a test AWS environment
- Create Free Tier AWS account
- Enable Cloudtrail

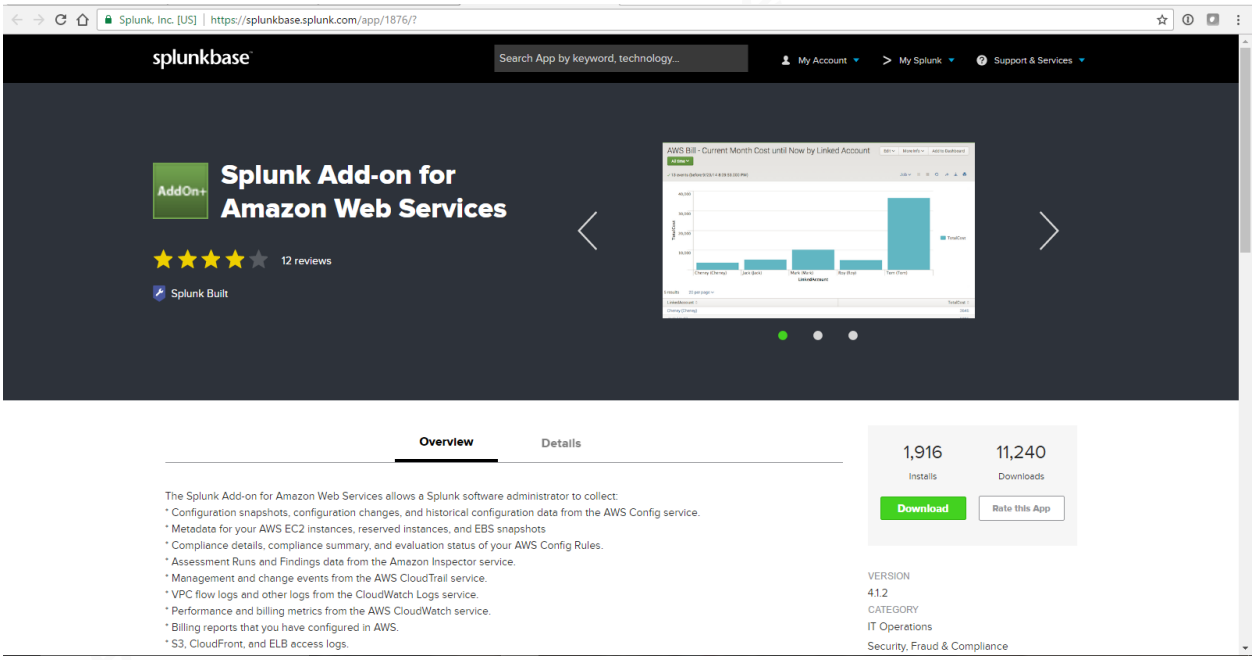


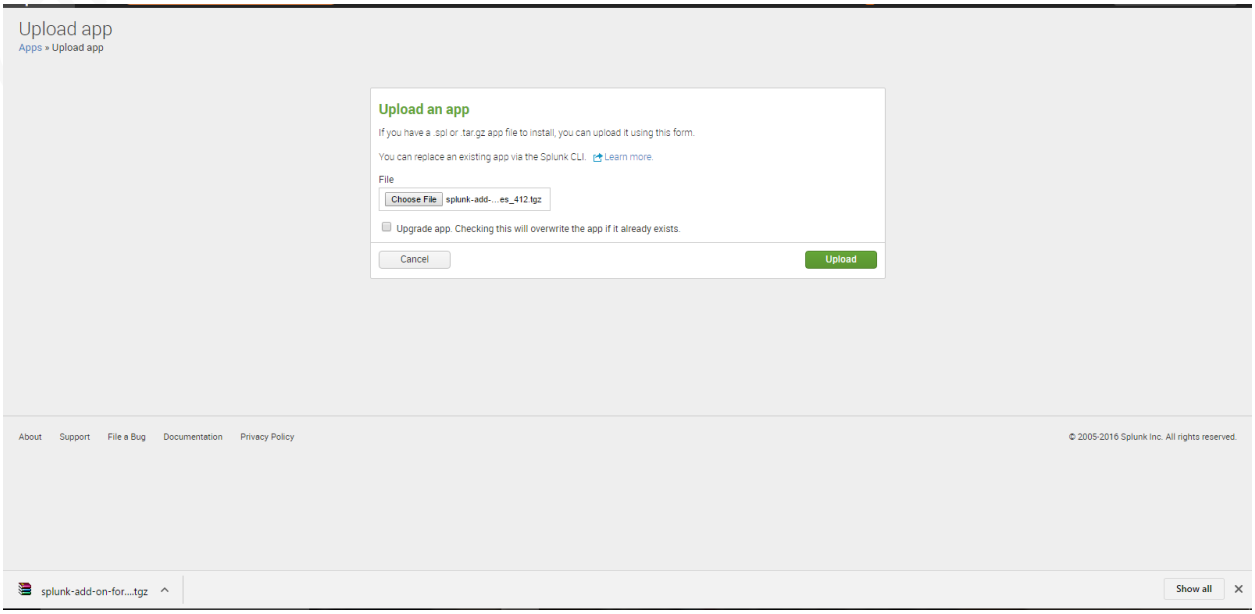


- Create accounts using IAM

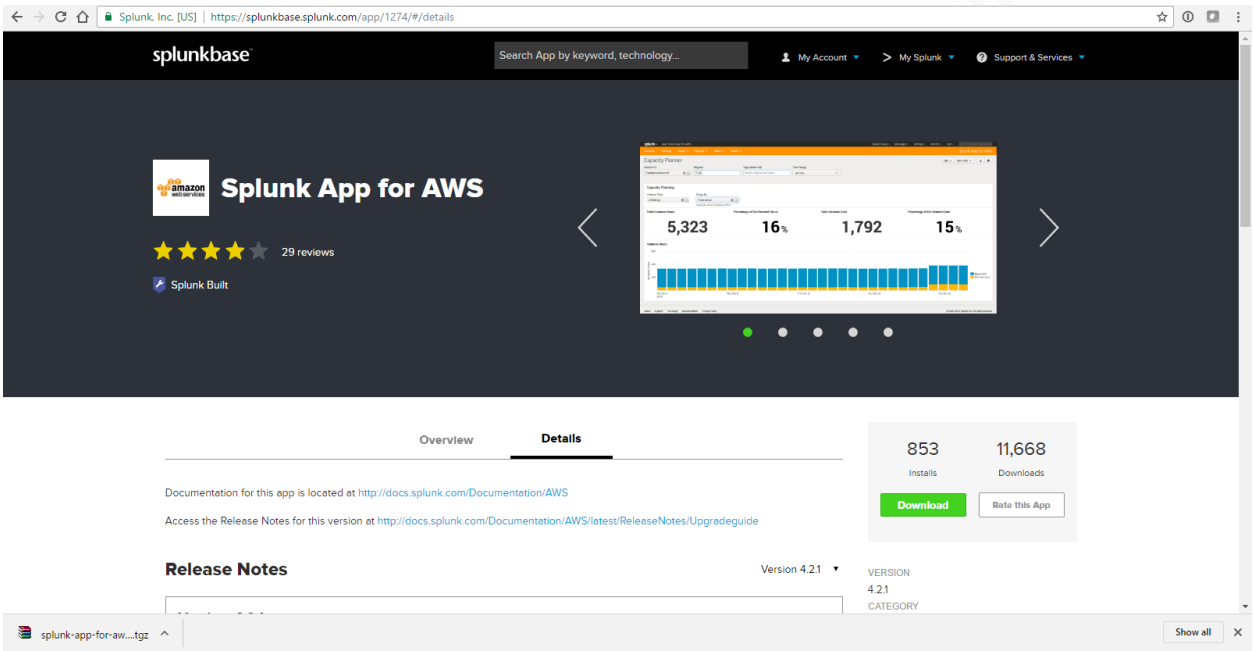


- Build a Splunk test environment –
- Getting AWS Cloudtrail logs to Splunk
- Install Splunk Enterprise 6.5
- Download and Install Splunk Add-on for Amazon Web Services

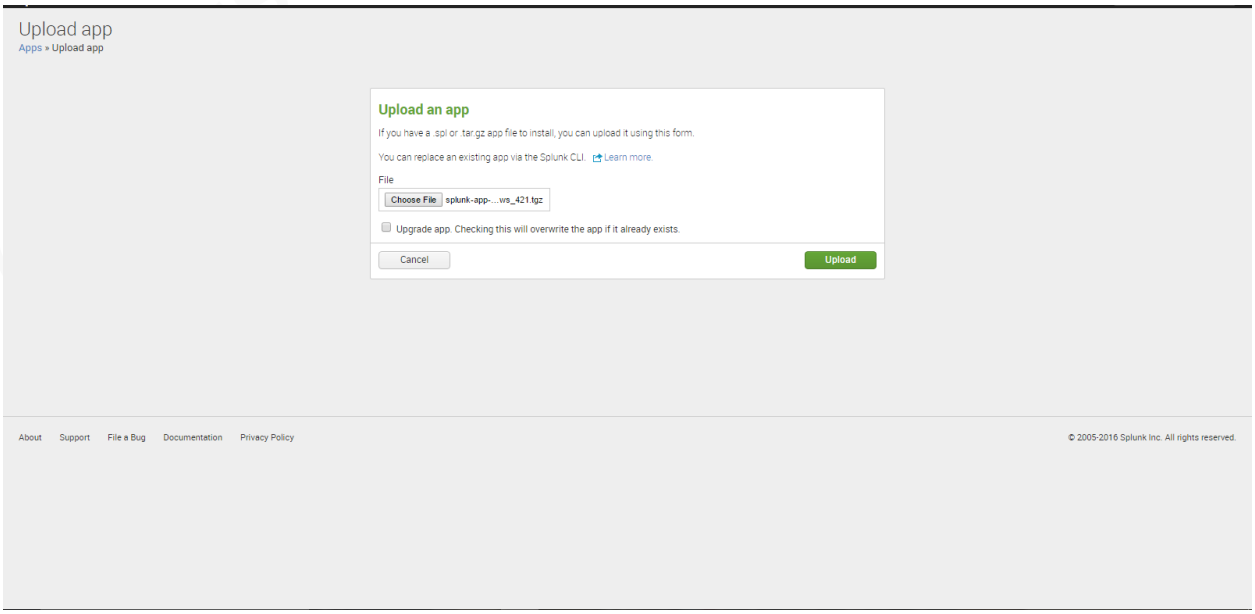




- Download and Install Splunk App for AWS

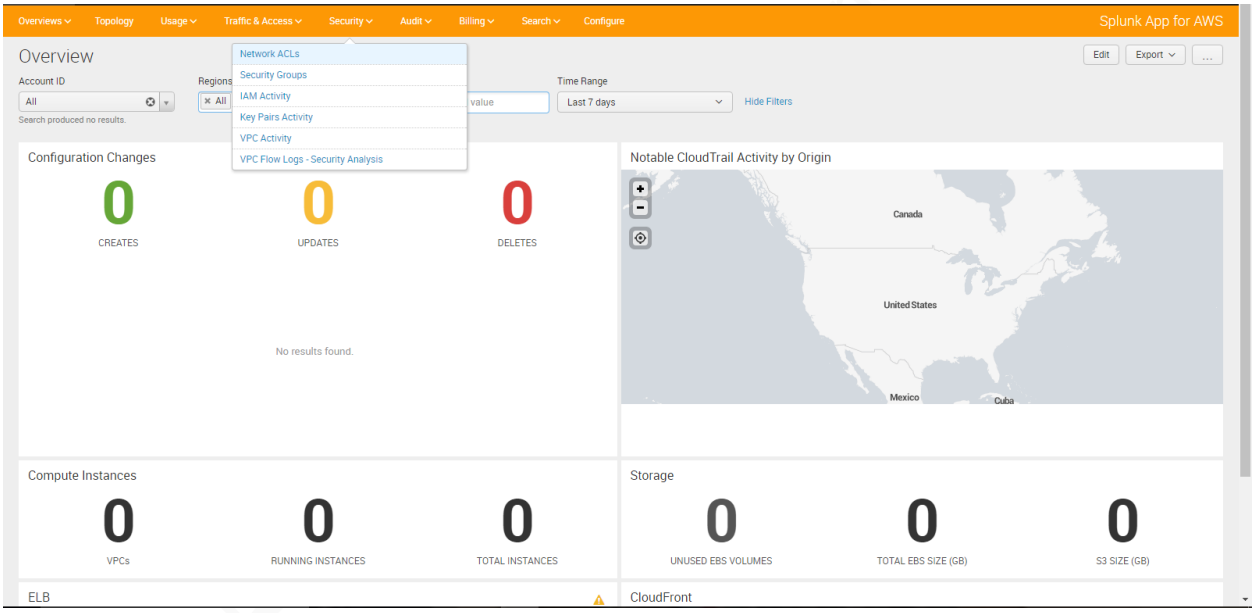


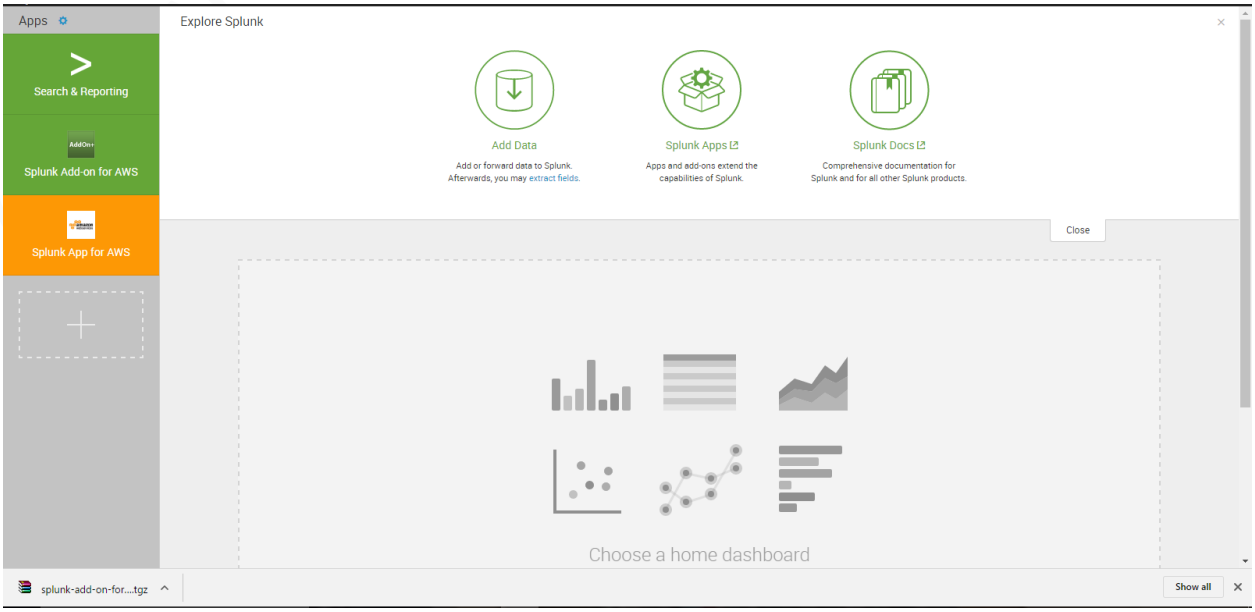
- Install the app and restart Splunk



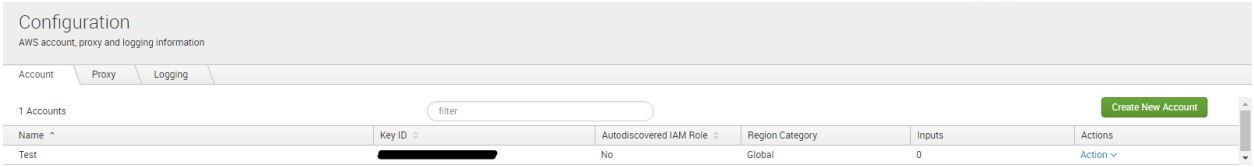
- Splunk App for AWS

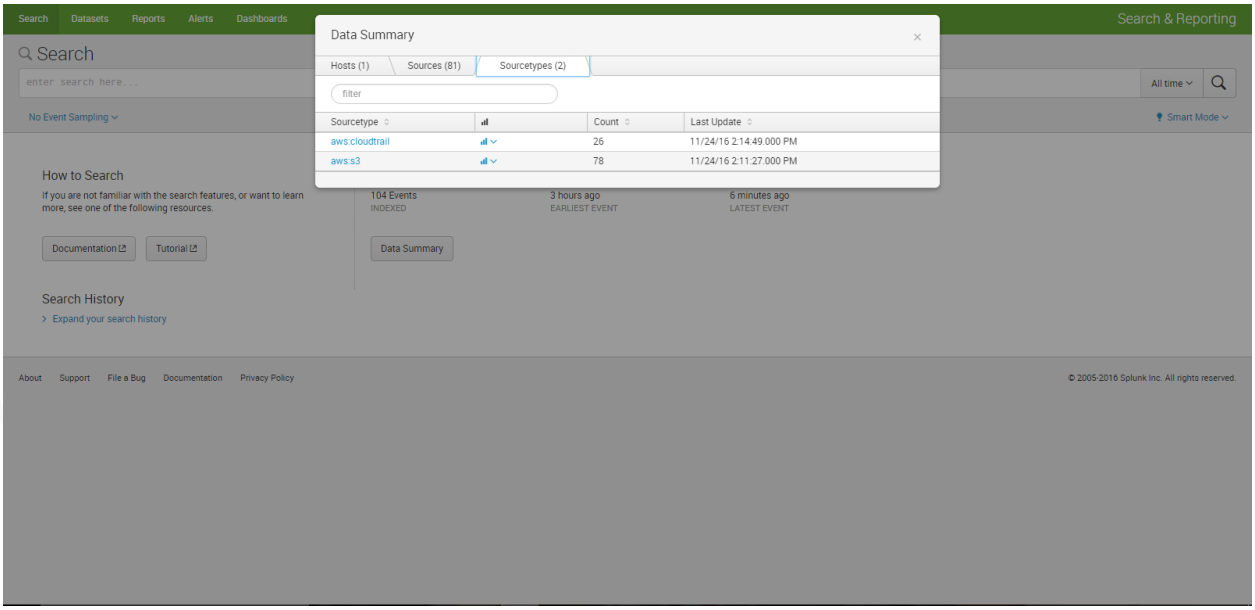
The Splunk App for AWS gives you significant operational and security insight into your Amazon Web Services account and infrastructure.





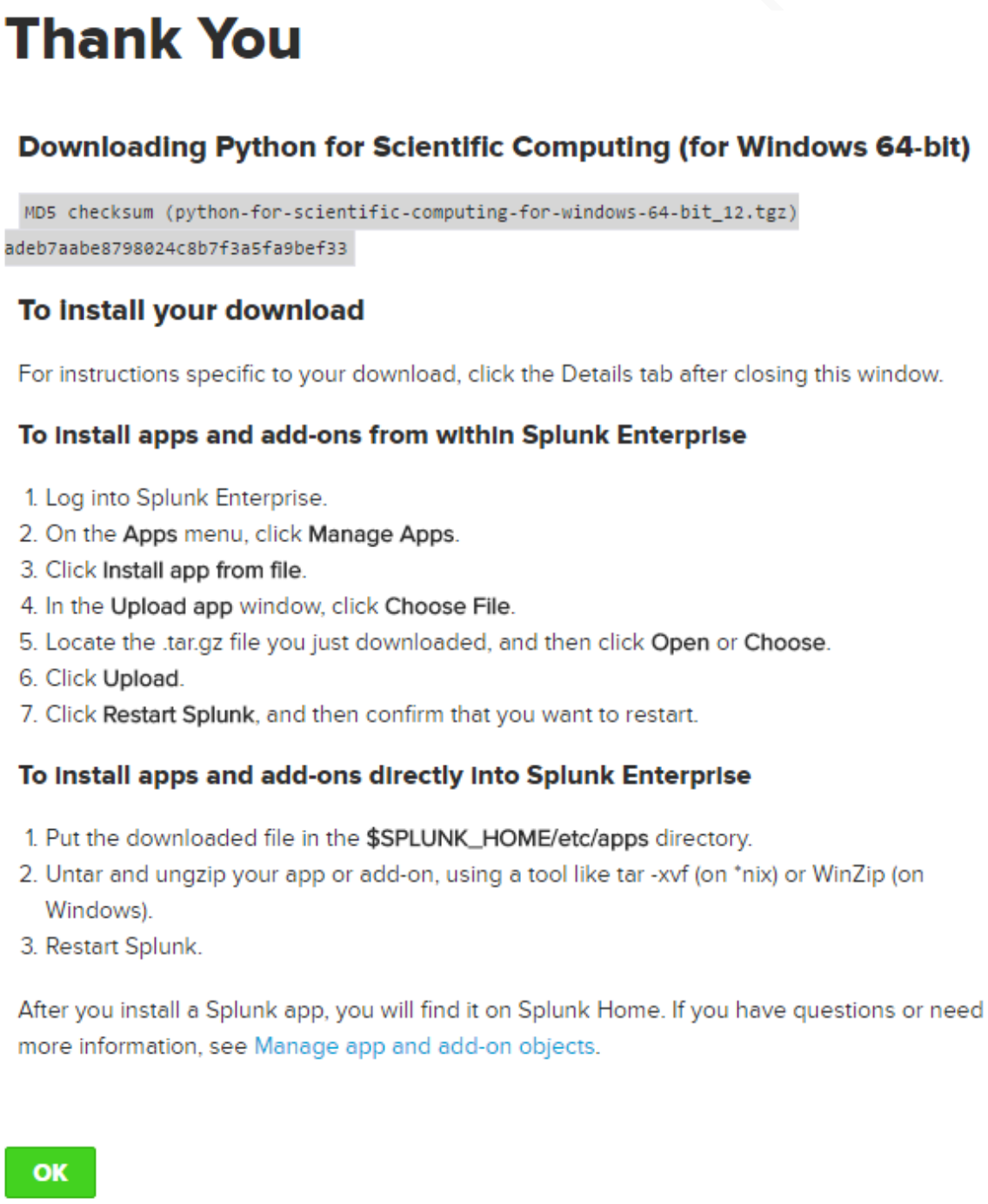
- Configured Splunk Add-on with AWS credentials





- Installing Machine Learning Toolkit

  Download and Install Python for Scientific Computing



Thank You

Downloading Python for Scientific Computing (for Windows 64-bit)

```
MD5 checksum (python-for-scientific-computing-for-windows-64-bit_12.tgz)
adeb7aabe8798024c8b7f3a5fa9bef33
```

To install your download

For instructions specific to your download, click the Details tab after closing this window.

To install apps and add-ons from within Splunk Enterprise

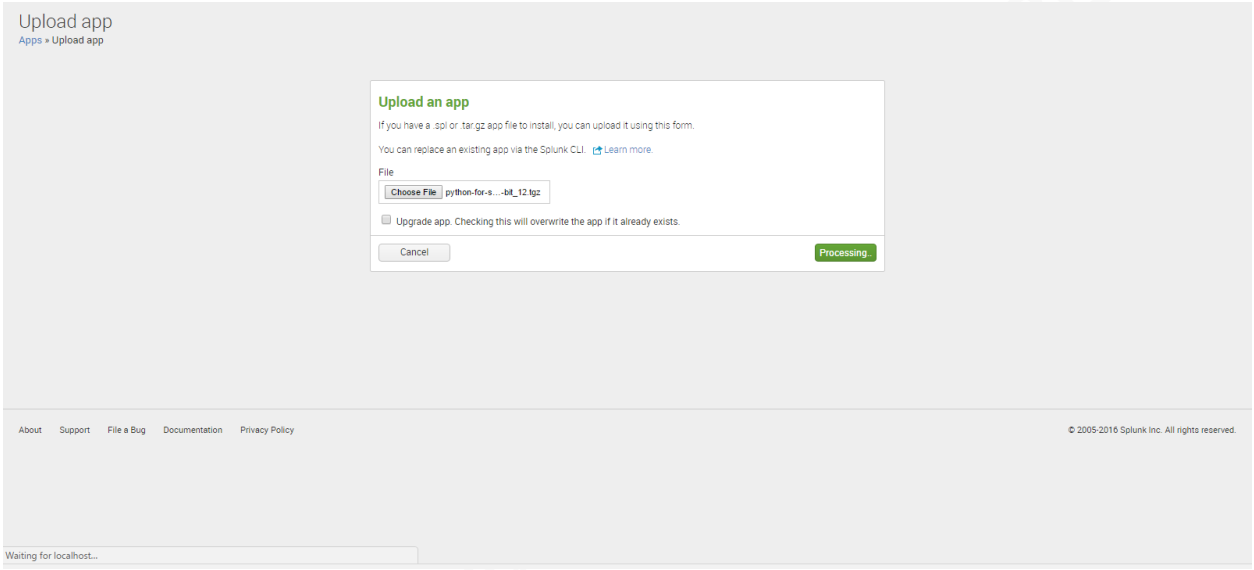
1. Log into Splunk Enterprise.
2. On the **Apps** menu, click **Manage Apps**.
3. Click **Install app from file**.
4. In the **Upload app** window, click **Choose File**.
5. Locate the .tar.gz file you just downloaded, and then click **Open** or **Choose**.
6. Click **Upload**.
7. Click **Restart Splunk**, and then confirm that you want to restart.

To install apps and add-ons directly into Splunk Enterprise

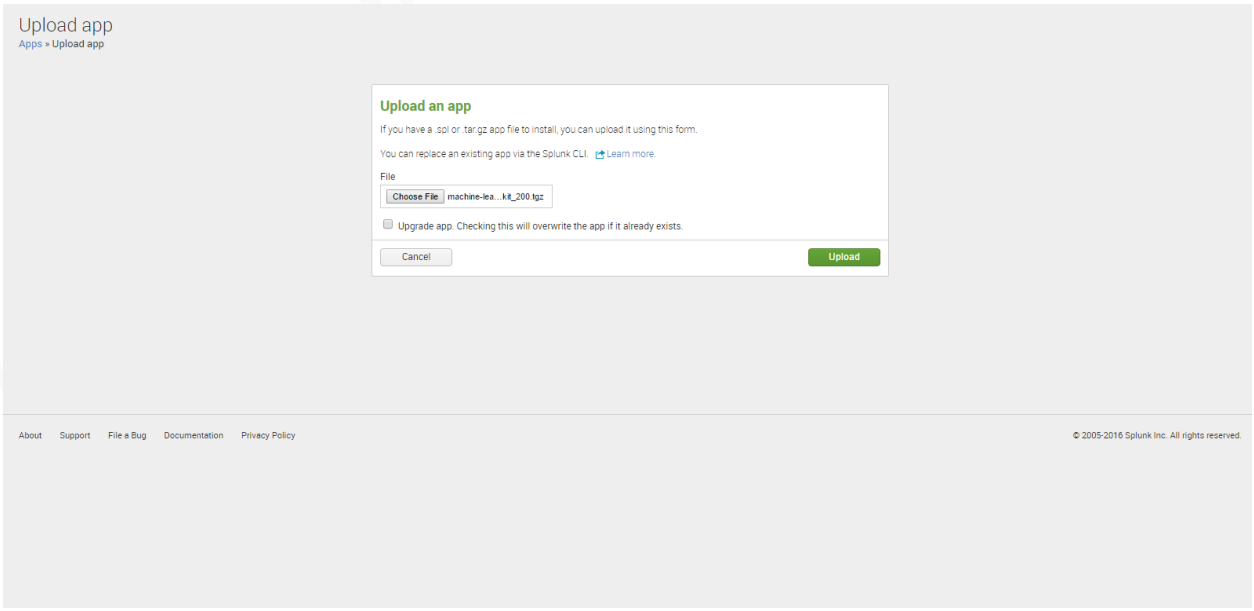
1. Put the downloaded file in the **$SPLUNK_HOME/etc/apps** directory.
2. Untar and ungzip your app or add-on, using a tool like tar -xvf (on *nix) or WinZip (on Windows).
3. Restart Splunk.

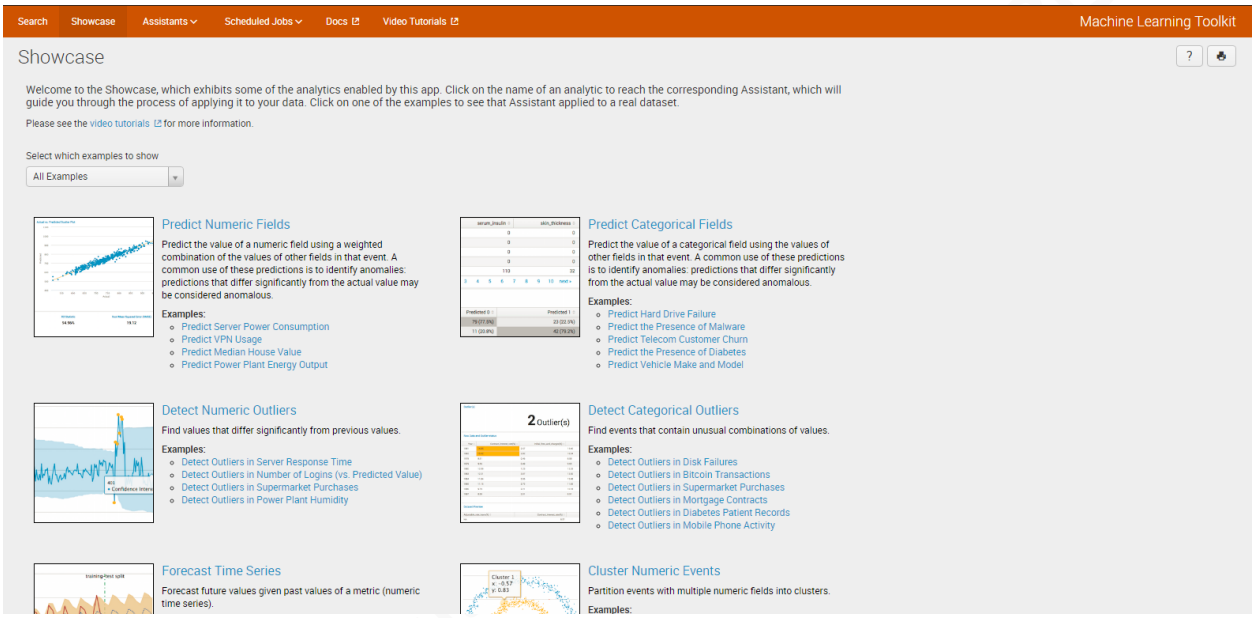
After you install a Splunk app, you will find it on Splunk Home. If you have questions or need more information, see Manage app and add-on objects.

OK



- Download and Install Machine Learning Toolkit





- This step completes the lab setup.

Unsupervised Learning Example – Detecting Categorical Outliers