

REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Saad Dahlab de Blida 1



Faculté des sciences

Département Informatique

Mémoire de fin d'étude pour l'obtention du diplôme de Master 2 en
Informatique

Option : Sécurité des systèmes d'information

Thème

Développement d'une application d'analyse des
fichiers logs et prédiction des attaques

Organisme d'Accueil :

Centre de Recherche sur l'Information Scientifique et Technique

Encadré par :

Mme. Boulekrinat Nour El Houda

Promotrice : Mme. Boustia Narhimene

Réalisé par :

Mlle. Taguelmint Ikram

Mlle. Hadj Mohammed Mariya

Soutenu le 29/09/2019 devant le jury composé de :

Mr. Bala Mahfoud, Maître de conférence, U.Blida 1, **Président**

Mme. Ghebghoub , Maître assistant, U.Blida 1, **Examineur**

Mme. Boustia Narhimene , Professeur, U.Blida 1, **Promoteur**

Promotion 2019

Session Septembre

Remerciements

Nos premiers remerciements vont à ALLAH le tout Puissant qui nous a guidé et qui nous a donné la force et la volonté de réaliser ce travail

C'est avec grand plaisir que nous réservons cette page, en signe de gratitude et de reconnaissance à tous ceux qui nous ont aidés à la réalisation de ce travail.

Nous remercions Mme Boulkrinat notre encadreur pour sa grande disponibilité, sa rigueur et professionnalisme qui n'a eu de cesse de nous inspirer et aussi pour la confiance qu'elle nous a accordée en proposant ce travail.

Nous remercions, notre promoteur Mme Boustia pour sa rigueur et la pertinence de ses jugements qui ont été très constructifs et nous ont permis de faire ce travail.

Nous remercions vivement les membres de jury pour nous avoir fait l'honneur d'accepter d'examiner notre travail.

Nous voudrions exprimer à nos proches toute notre gratitude : nos très chers parents, nos frères et nos sœurs. Sans leur amour, leur soutien, leur confiance et leurs encouragements, nous n'y serons peut à être pas arrivés.

Résumé

Les applications Web sont l'épine dorsale des systèmes d'information modernes. L'exposition sur internet de ces applications engendre continuellement de nouvelles formes de menaces qui peuvent mettre en péril la sécurité de l'ensemble des systèmes d'information.

La sécurité des systèmes d'information est une problématique d'une importance majeure pour les individus ainsi que pour les entreprises. Elle repose sur la mise en place d'une politique de sécurité autour de ces systèmes, pour compléter cette politique de sécurité, il est devenu nécessaire d'avoir des outils de surveillance pour auditer le système d'information et détecter d'éventuelles intrusions.

Aujourd'hui, l'analyse des fichiers logs est devenue le moyen idéal pour détecter les tentatives d'attaques et aider les administrateurs à identifier les éventuels failles de sécurité. Aussi, le traitement des données de logs peut se faire manuellement, mais cela nécessite beaucoup de temps et peut devenir pratiquement impossible lorsque la taille des fichiers logs est grande.

Dans notre travail nous proposons une solution de prédiction des attaques au travers l'analyse des fichiers logs. Les entrées du fichier journal de serveur web sont utilisées pour la prédiction des intrusions en se basant sur de bonnes techniques d'analyse prédictive et de classification qui permettent de traiter les données et de détecter les anomalies. Nous avons implémenté notre solution en utilisant plusieurs techniques d'apprentissage automatique et nous avons testé chaque technique. Les résultats des tests montrent que ces techniques présentent des résultats intéressants. Aussi l'utilisation de spark l'un des outils Big Data pour l'exécution de notre solution a montré que le temps d'analyse était très court par rapport à l'analyse normale, alors spark à assurer la rapidité du traitement.

Mots Clés : Analyse des fichiers logs, Fichier journal, prédictions des attaques, analyse prédictive, apprentissage automatique.

Abstract

Web applications are the backbone of modern information systems. The Internet exposure of these applications continually generates new forms of threats that can jeopardize the security of the entire information system.

The security of information systems is a problem of major importance for individuals as well as for companies. It is based on the implementation of a security policy around these systems, to complete this security policy, it has become necessary to have monitoring tools to audit the information system and detect possible intrusions.

Today, log file analysis has become the ideal way to detect attack attempts and help administrators to identify security vulnerabilities. Also, the processing of log data can be done manually, but it requires a lot of time and can become practically impossible when the size of log files is large.

In our work we propose a solution of attacks prediction through the analysis of log files. Web server log file entries are used for intrusion prediction based on predictive analytics and classification techniques to process data and detect anomalies. We implemented our solution using several machine learning techniques and tested each technique. Tests show that these techniques have interesting results. Also the use of spark one of the big data tools for the execution of our solution showed that the analysis time was very short in comparison with the normal analysis, so spark ensures the speed of the analysis.

Keywords : Log files analysis, log file, attacks prediction, predictive analysis, machine learning.

Table des matières

Introduction générale	1
1 Contexte	1
2 Problématique	1
3 Objectifs	2
4 Organisation du mémoire	2
1 Analyse des fichiers log	3
1 Introduction	3
2 Définition	3
3 Types des fichiers logs	4
3.1 Coté serveur (Server side log files)	4
3.2 Coté client (Client side log files)	4
3.3 Coté proxy (Proxy side log files)	4
3.4 Coté pare-feu (Firewall side log files)	4
3.5 Coté réseau (Network side log files)	4
3.6 Coté système (System side log files)	5
4 Contenu des fichiers logs web	5
5 Format des fichiers logs	6
5.1 Le format log Etendu W3C (W3C Extended Log File Format)	6
5.2 Le format log commun du NCSA (Common Log File Format)	7
5.3 Le format Microsoft IIS	8
6 Analyse d'un fichier log	8
7 Intérêt d'analyse des fichiers log	9
8 Les outils d'analyse des fichiers logs	10
8.1 Les outils traditionnels d'analyse des fichiers logs	10
8.1.1 GoAccess	10
8.1.2 Scalp	10

8.1.3	Logstash	10
9	Les problèmes liés aux fichiers logs	10
10	Conclusion	11
2	Big Data	12
1	Introduction	12
2	Définitions	12
3	Caractéristiques du Big Data	13
4	Domaines d'application du Big Data	14
4.1	Marketing	14
4.2	Surveillance	14
4.3	Sécurité	15
5	Les avantages et les inconvénients du Big Data	16
6	Architecture Big Data (Lambda)	17
7	Big data et la sécurité informatique (Cyber sécurité)	18
7.1	La sécurité des Big Data	18
7.2	Solution de sécurité pour les environnements Big Data	19
7.3	Le Big Data au service de la sécurité	19
8	Les outils Big Data d'analyse des fichiers logs	21
8.1	Splunk	21
8.2	Sumo Logic	21
8.3	Apache Metron	21
9	Conclusion	21
3	Analyse prédictive	23
1	Introduction	23
2	Data Mining	23
2.1	Principales tâches de Data Mining	23
2.2	Techniques et algorithmes de Data Mining	24
2.2.1	Techniques supervisées	25
2.2.2	Techniques non supervisées	25
3	Analyse de données	26
4	Analyse prédictive	26
4.1	Processus de l'analyse prédictive	27
5	Les techniques de prédiction	27
5.1	Les arbres de décision	28
5.2	Les réseaux de neurones	28
5.3	Les K plus proches voisins	29

5.4	La régression logistique	30
6	Prédiction et sécurité informatique	30
6.1	Le rôle de la prédiction d'intrusion	30
6.2	Travail connexe	31
7	Conclusion	31
4	Conception de la solution	32
1	Introduction	32
2	Architecture générale du système	32
3	Démarche suivie pour la conception du système	34
3.1	Préparation des données	34
3.2	Analyse prédictive	35
4	Lancement du système avec Spark	37
5	Conclusion	37
5	Implémentation et Réalisation	38
1	Introduction	38
2	Les ressources matérielles et logicielles	38
2.1	Matériels utilisés	38
2.2	Logiciels utilisés	38
3	Installation de l'environnement (Spark)	39
4	Préparation du système	41
4.1	Préparation de données	41
4.2	Prétraitement et nettoyage	42
4.3	Les algorithmes de prédiction	43
5	Mise en marche du système	43
5.1	Lancement du prétraitement et nettoyage	44
5.2	Lancement des programmes de prédiction	45
5.3	Lancement des programmes de prédiction avec Spark	46
6	Analyse de la performance du système	47
6.1	Analyse de la performance selon l'algorithme prédictif utilisé	47
6.2	Analyse de la performance selon le cluster de traitement utilisé	48
7	Conclusion	48