MapReduce를 이용한 대용량 보안 로그 분석

최대수*, 문길종**, 김용민***, 노봉남****

An Analysis of Large-Scale Security Log using MapReduce

Dae-Soo Choi*, Gil-Jong Mun**, Yong-Min Kim***, and Bong-Nam Noh****

요 약

로그 수집 및 분석은 보안관리 시스템에서 매우 중요하다. 다양한 이기종의 보안 장비에서 발생시키는 보안 로그들은 점점 증가하고 있으며, 보안관리자는 이를 수집하고 한정된 시간 내에 공격 유무를 분석하고 대응하여야 한다. 기존 통합 보안관리시스템은 DBMS를 이용하여 보안로그를 저장하고 분석하지만, 대용량의 로그 분석에는 부족하다. 본 논문에서는 대용량 이기종의 보안로그들을 통합적으로 수집 및 분석 할 수 있는 MapReduce 기반의 보안로그 분석시스템을 제안한다. 방화벽, 침입탐지시스템, 웹 로그를 대상으로 정규화하고 공통필드를 대상으로 병합하여, 병합 비율과 수행시간을 비교하였다. 수천만 건 이상의 대용량 로그 병합 실험에서 제안시스템은 DBMS 질의 방식보다 노드 6개에서 85% 이상의 빠른 수행시간으로 우수함을 보였다.

Abstract

The log collection and analysis are very important for security management system. Any security logs generated from various heterogeneous security devices are increasing, security managers should collect them, analyze whether there is an attack or not and handle it in limited time. Original security management system save and analyze security logs with DBMS, they are not enough to analyze large scale security logs. This paper suggests security analyzing system based on MapReduce to collect and analyze large scale heterogeneous security logs at once. Also the paper normalized firewall, Intrusion Detection System and web log, merged the common field, and compared merging rate with execution time. In more than tens of millions of large scale log merging tests, suggested system showed its superiority through faster executing time, over 85% in 6 nodes than DBMS inquiry method.

Keywords

log correlation, alert correlation, MapReduce, hadoop, cloud computing

Ⅰ. 서 론

바이러스, 악성코드 공격 등의 증가로 인하여 방화벽, 침입탐지시스템, 등 보안시스템에서는 대용량

* 전남대학교 정보보호협동과정
** (주)정보보호기술
*** 전남대학교 전자상거래전공
**** 전남대학교 시스템보안연구센터
· 제1저자(First Author): 최대수, 교신저자(Corresponding Author): 김용민
· 접수일: 2011년 07월 29일, 수정일: 1차- 2011년 08월 08일, 게재확정일: 2011년 08월 18일

의 보안 로그 및 경보를 발생시킨다. 통합 보안관리 시스템에서는 보안 이벤트를 데이터베이스에 저장하고 SQL 질의를 이용하여 분석하고 표현하였다. 이러한 경우에, 여러 문제가 발생할 수 있다[1]. 대량의 데이터를 분석하는데 비효율적이다. 또한 장기간의 대량 이벤트 추이 분석이 어려우며, 대량의 데이터 중에서 중요한 정보가 무시되기 쉽다. 또한 데이터베이스 문제 발생시, 모든 저장 이벤트가 손실된다. 기존에 보안 분야의 연관분석 연구 방법은 대량의 데이터 처리로 인한 성능 지연 문제에 대한 관심 있는 연구가 부족하였다.

최근 클라우드 컴퓨팅 기반 기술이 발전하면서, 데이터 분석 분야의 성능 지연문제에 대한 해결이 가능하게 되었다. Hadoop은 대용량 데이터를 처리하고 운영 할 수 있는 기반을 제공하는 분산 소프트웨어 플랫폼이다[2]. Hadoop은 Java로 구현된 MapReduce 프로그래밍 모델을 포함하며 대용량 데이터를 일반적인 컴퓨터로 구성된 클러스터 환경에서 병렬처리가 가능하다.

본 논문에서는 대용량 보안 로그 분석 시스템에 MapReduce 프로그래밍 모델을 적용하여 DBMS 방식보다 수행시간의 단축과 유사 경보의 건수를 병합하여 줄일 수 있음을 확인할 수 있었다.

2장에서는 로그 및 경보 연관분석과 관련된 연구를 조사하였으며, 클라우드 기반 기술인 MapReduce와 이를 이용한 분석 연구를 조사하였다. 3장에서는 제안 시스템의 구성 및 방법에 대하여 서술하며, 4장에서는 제안 시스템의 성능에 대해 분석 및 평가한다. 그리고 5장에서 결론 및 향후 연구에 대해 제시하였다.

II. 관련연구

2.1 로그 및 경보 연관분석

침입탐지 분야의 연관분석과 관련된 용어는 다양하게 사용되고 있으며, 로그 연관분석(Log Correlation)과 경보 연관분석(Alert Correlation)에 대한 다양한 연구 활동이 이루어졌다. A. Hattala et al.은 경보를 IDS에 의해 탐지되는 경고로 중요한 이벤트라고 정의하였다[3]. 그리고 이벤트는 IDS에 의해 분석되는 하위 수준 실체라고 정의하였다. 그래서 하나의 이벤트는 다수의 경보로 구성된다. D. Gorton의 경우, 이벤트 연관분석은 이벤트 자체를 분석하여 포렌식 수사에 사용 목적을 두고 있고, 경보 연관분석은 오용과 이상행위를 탐지하여 보안관제 분야에 사용목적이 있다고 구분하기도 하였다[4]. Fredrik Valeur et al.의 경우, 보안 경보 및 이벤트를 분석하여 상위 수준의 분석 화면을 제시하는 것이 경보 연관분석 이라고 정의하였다[5]. 이와 같이 로그, 경보, 이벤트, 경고의 용어가 사용되고 있어 이에 대한 정리를 요구한다.

표 1. 경보와 로그 연관분석 비교
Table 1. The comparison alert and log correlation

구분명	내 용
경보 연관분석 [4][6]-[7]	침입탐지 시스템(IDS) 의 탐지 데이터 중심 분석으로 네트워크 트래픽, 패킷, 페이로드를 포함 상대적으로 로그 연관분석보다 하위 수준의 분석
로그 연관분석 [3]-[5][8]	다수 이기종 시스템의 보안 로그를 중심으로 분석하며, 방화벽, 침입탐지시스템, 웹서버 로그 등이 해당 트래픽, 패킷, 페이로드의 분석은 제외 상대적으로 경보 연관분석보다 응용 수준의 분석

본 논문에서는 용어 혼돈을 피하고 범위를 명확히 하기 위하여 로그 연관분석과 경보 연관분석을 표 1과 같이 구분한다. 본 논문에서는 로그 연관분석 범위에서 첫 단계인 이기종 로그 병합을 대상으로 연구하였다.

2.2 MapReduce 를 이용한 보안로그 분석

MapReduce 프로그래밍 모델은 대용량 데이터를 클러스터 환경에서 병렬처리가 가능한 Map과 Reduce라는 함수로 구성이 되며 그림 1과 같이 실행된다[2]. 사용자가 정의한 Map(), Reduce() 함수에 의하여 입력 파일은 분할되어 다수의 노드에서 실행된다. 분할된 데이터는 사용자가 정의한 함수 Map() 함수로 처리되어 (key, value)값을 생성한다.

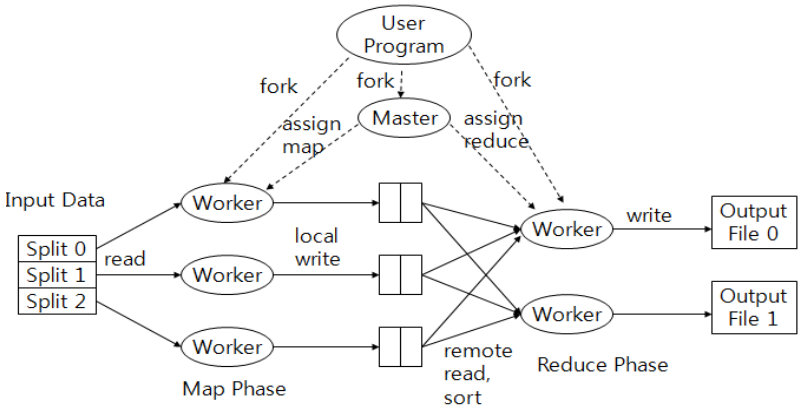

그림 1. MapReduce 프로그래밍 모델
Fig. 1. MapReduce programming model

이렇게 생성된 값은 중간 파일로 저장되어 Sort와 Shuffle 단계를 수행한다. 그리고 사용자가 정의한 Reduce() 함수를 이용하여 새로운 (key, value)로 병합 생성되어 출력 파일에 저장한다.

Wei-Yu et al.는 Hadoop과 MapReduce 프로그래밍 모델을 활용하여 Snort 탐지이벤트 병합을 실험하였다[1]. 기존 경보를 MySQL에서 처리시와 노드 수를 증가시켜 가면서 처리하는 비교를 실험하였다. 보안로그 통합분야에 응용했다는 의미는 있지만, 방화벽에 비하여 상대적으로 양이 적은 침입탐지시스템 로그에 적용했기 때문에, 보안로그 통합과 로그 연관분석의 평가에는 부족하다.

표 2. MapReduce 관련 연구
Table 2. Related research using MapReduce

관련연구	로그 분석 대상	병합 요소
침입탐지시스템 경보 분석[1]	Snort 경보 병합	Signature, Destination ip
네트워크 트래픽 분석[9]-[10]	NetFlow 트래픽 데이터 병합	Destination Port, Octets

Y. Lee et al.는 Traffic Data 통계 처리에 MapReduce 프로그래밍 모델을 적용하였다[9]. 침입탐지시스템의 로그보다는 양이 많지만, Flow 데이터 병합의 요소가 Destination Port와 Octets에 대해서만 병합하여, 3개 이상 요소의 병합 성능의 평가에 대해서는 부족하다. 표 2에서 2가지 연구의 내용을 요약하였다. 현재, Hadoop과 MapReduce 프로그래밍 모델을 적용한 이기종 로그와 경보 연관분석 연구가 부족하다. 본 논문의 연구에서는 대용량 로그 연관분석에 대하여 적용한다.

III. 대용량 보안로그 분석시스템

3.1 시스템 구성요소

본 논문에서 제안하는 대용량 보안로그 분석시스템(LaSLAS : Large-Scale Security Log Analysis System)은 7개 이상의 세부 병합 요소를 하나의 key로 생성하여 병합하였다. 기존 연구들은 2개 요소를 대상으로 병합하였으며, 그 이상의 요소를 병합하기 위해서는 MapReduce 과정을 다수 반복해야 한다.

제안시스템은 크게 2가지 부분으로 구성이 되는데, 첫 번째 부분은 수집된 로그를 정규화(Log Normalization)하고, 동일 유형의 값으로 변환하는 전처리과정이다. 두 번째 부분은 로그를 Map 과정과 Reduce 과정으로 유사 로그를 병합하는 로그 쓰레드 연관분석(Log Thread Correlation)으로 구분한다.

표 3은 통합 보안관제분야에서 중요한 방화벽, 침입탐지시스템 로그와 최근 많은 해킹 공격대상으로 중요성이 부각되는 웹 로그 3가지 데이터에 대하여 병합 요소를 정의하였다. 병합 요소가 늘어나면, 분석은 세분화 되지만, 대용량 로그 분석 처리 시간은 늘어난다.

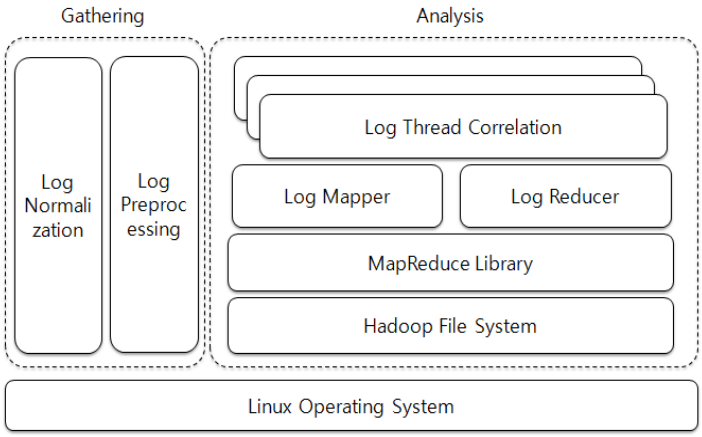

그림 2. 대용량 로그 분석시스템 구조
Fig. 2. The architecture of large-scale log analysis system

표 3. 로그 병합 필드
Table 3. Integrated field of a log

시스템	병합 요소
방화벽 로그	Sensor IP, Time, Source IP, Destination IP, Destination Port, Protocol, Direction, Action (Accept/Deny)
침입탐지 시스템 로그	Sensor IP, Time, Source IP, Destination IP, Destination Port, Protocol, Direction, Signature
웹로그	Sensor IP, Time, Source IP, Destination IP, Destination Port, Protocol, Method (GET/POST), URL, Result Code

3.2 보안로그 병합 방법

각 보안시스템에서 수집된 로그는 정규화하고 전처리되어 공통 형식의 텍스트로 변환된다. 그림 3은 각 보안시스템 종류별로 로그가 몇 건으로 병합 되는지를 MapReduce로 프로그래밍한 의사코드이다. Mapper 함수에서는 각 보안로그 종류 별로 Key를 생성하고 Count를 생성한다. Reducer 함수에서는 각 보안로그 종류 별로 동일한 Key 값을 갖는 리스트를 병합한다. 이 과정을 반복하여 보안로그 병합이 이루어진다. 이때 시간 범위는 통합 보안관제시스템 운영시 적합한 10분 단위로 하였다. 그림 3은 제안 방법의 보안로그 병합 의사코드를 보인 것이다.

```
Security Log Merge Pseudo Code
Mapper (Log)
{
 While ( LogList.Log != NULL) {
 if (firewall log)
  make_log_key (sensor ip, time, source ip, destination
 ip, destination port, direction, protocol, event type);
 if (ids log)
  make_log_key (sensor ip, time, source ip, destination
ip, destination port, direction, protocol, signature);
 if (web log)
  make_log_key (sensor ip, time, source ip, destination
ip, destination port, protocol, method, url, return code);
 }
}

Reducer (Log)
{
 While (LogList.Log != NULL) {
  for (Time_Window_Value) {
    if(current firewall log key == next firewall log key)
      merge (current firewall log, next firewall log);
    if(current ids log key == next ids log key)
      merge (current ids log, next ids log);
    if (current web log key == next web log key)
      merge (current web log, next web log);
  }
 }
}
```

그림 3. 로그 병합 의사 코드
Fig. 3. Pseudo code for log integration

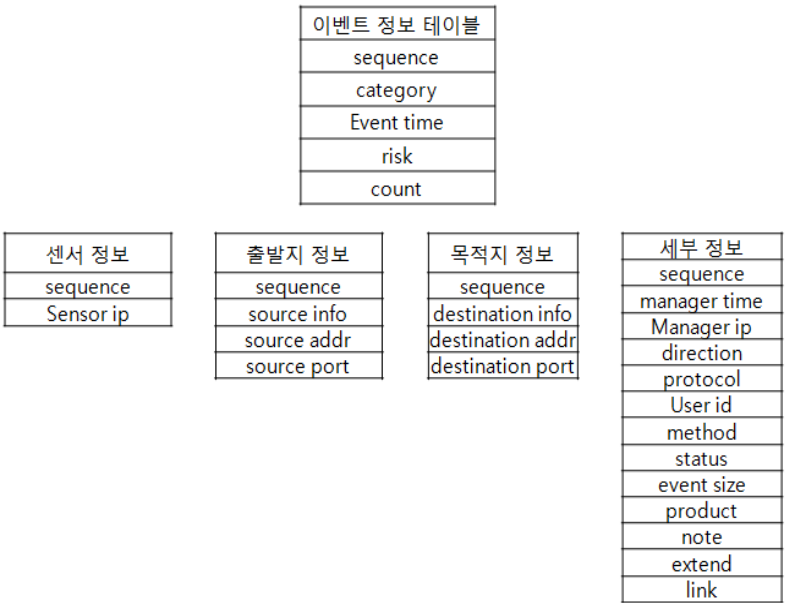

그림 4. 로그 테이블 구조
Fig. 4. Architecture of a log table

IV. 실험 및 결과

4.1 실험환경

본 논문에서 제안하는 대용량 보안로그 분석시스템(LaSLAS)의 성능평가는 Hadoop 기반의 가상화 환경을 구축하고, 노드 수를 늘려가면서 수행속도를 비교하였다. 개발언어인 Java 버전은 1.6을 사용하였으며 Hadoop의 버전은 0.21이다.

MapReduce 성능평가를 위한 HDFS 서버의 CPU는 Intel Xeon E7520 1.87GHz이며 메모리는 2G, 하드디스크는 80G이며 운영체제는 Fedora Core 8 2.6.23.1을 사용하여 6대의 가상머신을 구성하였다. Oracle DBMS서버의 CPU는 Intel Xeon E5405 2.0GHz 이며 메모리는 4G, 하드디스크는 100G이며 운영체제는 Fedora Core 8 2.6.23.1을 사용하였다. Oracle DBMS 의 버전은 10g Release2 이다.

보안로그 수집 이후 평가를 위하여, 노드 개수를 증가하여 병합 수행 시간을 기록하고, DBMS 질의 시간을 비교하였다. DBMS의 정규화된 로그 테이블 구조는 그림 4와 같다.

실험에 사용하는 데이터는 2가지 종류로 실험하였다. 첫번째 종류로 방화벽, 침입탐지시스템과 웹서버의 로그는 실제 기업에서 운영중인 1일 보안로그를 이용하였다. 방화벽 로그는 42,229,905건, 침입탐지시스템은 5,314건 그리고 웹서버는 1,021,938건으로 합계는 42,257,155건이다.

두번째 종류는 DARPA 1998 훈련용 tcpdump.list 7주간의 데이터를 이용하였다. 1주는 179,115건, 2주는 190,545건, 3주는 362,996건, 4주는 234,190건, 5주는 615,983건, 6주는 996,742건, 7주는 436,291건으로 합계는 3,013,862건이다.

본 논문에서 제안한 시스템은 로그를 대상으로 하는데, 표준화된 성능 비교 데이터가 없다. 따라서 실제 운영 데이터를 이용하였으며, 또한 DARPA 데이터는 패킷 데이터 형식으로 분석의 용이성을 위하여, 텍스트 형식의 로그를 기록한 tcpdump.list를 이용하였다.

4.2 실험 분석 및 평가

첫번째 실험은 그림 5에서 보이는 것처럼 1일 보안로그를 종류별로 병합한 결과를 DBMS 질의 방식과 비교하였다. 또한 노드의 수를 증가시켜 가면

서 비교하고, 병합비율을 확인하였다. 방화벽의 경우, 1일 41,229,905건의 로그가 4,075,406건으로 병합되어 약 90%의 병합 비율을 나타낸다. 또한, 표 4에서 처럼 DBMS 질의방식에 비하여 수행시간도 월등히 빨리 처리되었다. DBMS 질의방식이 2538초이지만, 제안방식은 6개 노드에서 361초가 소요되어 약 85%가 향상되었다.

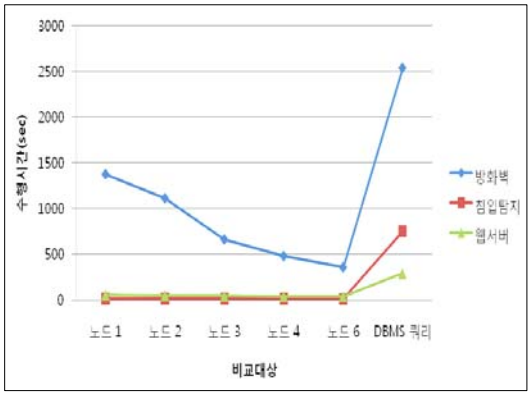

그림 5. 보안로그 병합 실험 결과
Fig. 5. The result of integration for security log

그러나 침입탐지시스템의 경우 로그 량이 작아서 병합비율이 약 27%로 적다. 또한 노드의 변화에 의한 수행시간 단축이 미비하다. 따라서 실험 결과에서 분석할 수 있는 것처럼 대략 수천만 건 이상의 로그에서 로그병합 시간은 매우 효과적임을 알 수 있었다.

표 4. 보안로그 병합 실험 결과
Table 4. The result of integration for security log

구 분	방화벽	침입탐지	웹서버
1개 노드 (초)	1378	16	58
2개 노드 (초)	1115	16	53
3개 노드 (초)	666	15	46
4개 노드 (초)	484	15	42
6개 노드 (초)	361	14	41
DBMS 질의 (초)	2538	755	292
로그건수 (건)	41,229,905	5,314	1,021,936
병합건수 (건)	4,075,406	3,858	203,770
병합비율 (%)	90.12%	27.40%	80.06%

두번째 실험은 그림 6에서 보인 것처럼 상대적으로 로그 수가 수백만 건 이하의 경우에 대한 실험이다. 첫 주에 179,115건의 로그가 병합되어 26,633건이 되었으며 약 85% 의 병합 비율을 나타낸다. 또한, 표 5에서 처럼 DBMS 질의방식에 비하여 수행시간도 매우 빠르게 처리되었다. DBMS 질의 방식이 180초이지만, 제안방식은 6개 노드에서 23초가 소요되었다. 약 87%가 향상되었다.

그러나 노드를 1개에서 6개로 증가시켰을 때, 수행시간의 변화는 작았다. 즉, 로그 량이 수백만 건 이하의 경우 노드 수가 성능에는 큰 영향이 없음을 알 수 있었다.

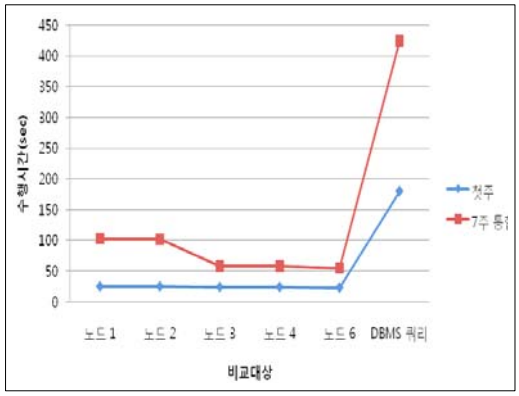

그림 6. tcpdump.list 병합 실험 결과
Fig. 6. Result of tcpdump.list integration

표 5. Tcpdump.list 로그 병합 실험 결과
Table 5. The result of integration for tcpdump.list

구 분	1주	7주 합계
1개 노드 (초)	25	103
2개 노드 (초)	25	102
3개 노드 (초)	24	58
4개 노드 (초)	24	58
6개 노드 (초)	23	55
DBMS 쿼리 (초)	180	425
로그건수 (건)	179,115	3,013,862
병합건수 (건)	26,633	347,069
병합비율 (%)	85.13%	88.48%

2가지 종류의 실험에서 분석된 내용을 종합하면 로그 수가 수백만 건 이상인 경우, DBMS 질의 방식에 비교하여 훨씬 빠른 수행시간의 개선이 있었다. 노드 수 변화 시 수행시간을 비교하여 보면, 로

그 수가 수백만 건 이하일 경우는 성능 향상이 미흡하였다. 로그 수가 수천만 건 이상이 될 경우 노드 수에 따라 더 빨라졌다. 대량의 보안 로그 분석에 Hadoop 기반의 분산처리 환경이 기존 DBMS 처리방식보다 적합함을 알 수 있었다. 그러나 관련 연구[1][9]와 비교는 실험 데이터유형과 병합요소의 갯수, 비교방법에 차이가 있어 병합비율, 수행시간의 계량적 비교는 의미가 적다. 표 6에서 장단점을 분석하였다.

표 6. 병합 실험 결과 분석
Table 6. An analysis of merging result

	장점	단점
snort 로그 병합[1]	실험데이타를 snort에 입력후 경보 발생하여 병합을 시도함	IDS 단일유형 병합 2가지요소 단순병합 소량 데이터 병합 (최대 경보 12698건)
traffic 데이터 병합[2]	binary 형식의 데이터를 text 형식으로 변환 후 병합 시도함 대량 데이터 병합 (1일 3,200,000건)	트래픽 단일유형병합 2가지요소 단순병합 최대 4개노드 실험 ([1],LaSLAS는 6개 노드까지 실험)
LaSLAS	다양한 보안로그 대상 병합 시도함 7가지요소 복합병합 대량 데이터 병합 (1일 41,229,905건)	이기종 보안로그간의 병합 실험 부족

V. 결 론

이기종 다수의 보안장비에서는 다량의 보안로그 및 경보를 발생시킨다. 이러한 로그들은 점점 증가하면서 실시간으로 분석하기가 어려워지고 있다. 기존 통합 보안관리시스템은 로그를 데이터베이스에 저장하고, SQL 질의를 이용하여 분석하고 표현하였다. 이 방식은 대량의 데이터를 처리하는데 어려움이 있다.

본 논문에서는 대용량 이벤트 처리에 용이한 Hadoop 플랫폼과 MapReduce 프로그래밍 모델을 살펴보았으며, 이기종 보안로그를 대상으로 MapReduce를 적용한 대용량 보안로그 분석 시스템을 제안하였다. 실 환경에서 운영중인 방화벽, 침입탐지시스템, 웹서버의 1일 로그를 대상으로 실험하였다. 기존 통합분석시스템에서 사용하는 DBMS 질의 방식과 수행시간을 비교하여 분석하였다. 수천만건 이상의 방화벽 로그의 경우 DBMS 처리 방식보다 노드 6개에서 수행시간 성능향상이 있음을 보였다. 상대적으로 수백만건 이하 로그의 경우는 DBMS 질의 방식보다 빨랐지만, 노드 수 증가에 크게 영향을 받지 않았다. 대용량 보안 로그 연관분석의 기반이 되는 로그 정규화와 로그 병합 분석에 MapReduce 모델이 적합함을 알 수 있었다. 향후 연구에서는 보안 로그 연관분석 알고리즘 및 탐지율의 평가에 대한 적용이 요구된다.

참 고 문 헌

[1] Wei-Yu Chen and Jazz Wang, "Building a Cloud Computing Analysis System for Intrusion Detection System", CLOUD SLAM, April 2009.

[2] Tom White, "*Hadoop Definitive Guide*", O'Reilly Media, June 2009.

[3] A. Hattala, C. Sars, R Addams, and T. Virtanen, "Event Data Exchange and Intrusion Alert Correlation in Heterogeneous Networks", 8th Colloquium for Information Systems Security Education. West Point NY, June 2004.

[4] D. Gorton, "*Extending Intrusion Detection with Alert Correlation and Intrusion Tolerance*", MPhil Thesis, Chalmers University of Technology, Dept. of Computer Engineering, Sweden, 2003.

[5] Fredrik Valeur, Giovanni Vigna, Christopher Kruegel, and Richard A. Kemmerer, "A comprehensive approach to intrusion detection alert correlation", *IEEE Transactions on Dependable and Secure Computing*, Vol. 1, No. 3, pp. 146-169, July 2004.

[6] Reza Sadoddin, Ali Ghorbani, and Fredericton, "Alert Correlation Survey: Framework and Techniques", Proceedings of the International Conference on Privacy, Security and Trust, pp. 6-

15, Oct. 2006.

[7] Limmer Tobias and Falko Dressler, "Survey of Event Correlation Techniques for Attack Detection in Early Warning Systems", Exchange Organizational Behavior Teaching Journal, April 2008.

[8] Abad, C., Taylor, J., Sengul, C., Yurcik, W., and Rowe, K., "Log correlation for intrusion detection: a proof of concept", Computer Security Application Conference, pp. 255-264, Dec. 2003.

[9] Y. Lee, W. Kang, and H. Son, "An Internet Traffic Analysis Method with MapReduce", the 1st IEEE/IETP Workshop on Cloud management, pp. 357-361, April 2010.

[10] Jan Tore Morken, "Distributed NetFlow Processing Using the Map-Reduce Model", Master's thesis, Norwegian University of Science and Technology, June 2010.

저자소개

최 대 수 (Dae-Soo Choi)

1999년 2월 : 수원대학교 대학원 전자계산학과(석사)
2008년 3월 ~ 현재 : 전남대학교 정보보호협동과정 박사과정
관심분야 : 통합보안관리시스템, 종합위험관리시스템, 융합보안

문 길 종 (Gil-Jong Mun)

2009년 : 전남대학교 대학원 정보보호협동과정(박사)
2009년 ~ 현재 : (주)정보보호기술 선임연구원
2009년 ~ 현재 : 서울디지털대학교 초빙교수
관심분야 : 정보보호, U-City 보안, 스마트그리드 보안, 무선 보안

김 용 민 (Yong-Min Kim)

2002년 8월 : 전남대학교 대학원 전산통계학과(박사)
2004년 3월 ~ 2006년 : 여수대학교 정보기술학부 전임강사
2006년 3월 ~ 현재 : 전남대학교 문화콘텐츠학부(전자상거래전공) 부교수
관심분야 : 시스템 및 네트워크 보안, 전자상거래 보안

노 봉 남 (Bong-Nam Noh)

1994년 2월 : 전북대학교 대학원 전산학과(박사)
1983년 ~ 현재 : 전남대학교 전자컴퓨터정보통신공학부 교수
2000년 ~ 현재 : 전남대학교 시스템보안연구센터 소장
관심분야 : 컴퓨터와 네트워크 보안, 정보보호시스템, 전자상거래 보안, 사이버사회와 윤리