(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2013년12월20일
(11) 등록번호 10-1343804
(24) 등록일자 2013년12월16일

(51) 국제특허분류(Int. Cl.)
H04L 12/26 (2006.01) *H04L 12/24* (2006.01)
(21) 출원번호 10-2013-0027137
(22) 출원일자 2013년03월14일
심사청구일자 2013년03월14일
(56) 선행기술조사문헌
논문1:IEEE
JP2011028703 A
논문2:ACM
KR1020120092930 A

(73) 특허권자
충남대학교산학협력단
대전광역시 유성구 대학로 99 (궁동, 충남대학교)
(72) 발명자
이연희
대전광역시 유성구 노은로 353 송림마을아파트 304동 1701호
이영석
대전광역시 유성구 테크노중앙로 32 금성백조아파트 809-602 (관평동)
(74) 대리인
김원준

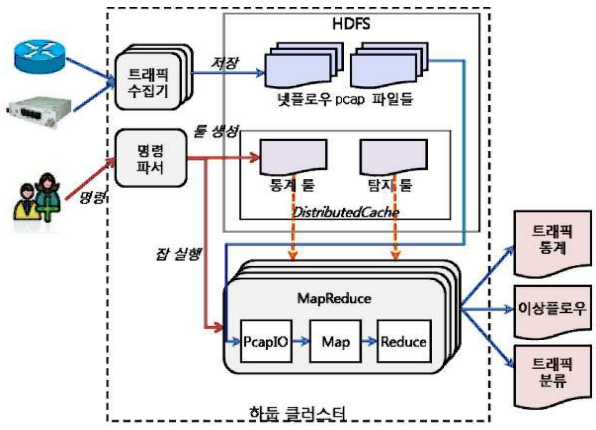
전체 청구항 수 : 총 5 항 심사관 : 전용해

(54) 발명의 명칭 **맵리듀스 방식의 트래픽 분석을 위한 룰 정의 방법 및 룰을 이용한 트래픽 분석 방법**

(57) 요 약

본 발명은 사용자들이 분석하고자 하는 대상과 방법을 쉽게 정의하고 분석할 수 있도록 하는 맵리듀스 방식의 트래픽 분석을 위한 룰 정의 방법과 이에 의해 정의된 룰을 이용한 보편적이고 일원화된 맵리듀스 방식의 트래픽 분석 방법에 관한 것이다. 보다 상세하게는 1) 룰의 식별명(rulename); 2) 상기 플로우의 시간 빈(time-bin)의 크기(binsize); 3) 상기 플로우의 필터링 조건(filters); 4) 필터링 된 플루우에 대한 맵리듀스에서 그룹핑에 사용될 그룹키 항목들(groupbykeys); 5) 상기 그룹핑에 의한 그룹 내에서 정렬에 사용될 정렬키 항목들(sortbykeys); 6) 정렬된 키 항목의 값에 대해서 유일값 연산을 할 것인지 합계연산을 할 것인지를 나타내는 통계방법(unique/count); 7) 통계방법에 따른 계산 결과에 대한 임계치 기반의 탐지조건(conditions); 및 8) 상기 탐지조건에 맞는 레코드의 반환할 필드(retvals);을 포함하는 항목을 정의하는 것을 특징으로 하는 룰 정의 방법 및 상기 방법에 의해 정의된 룰을 이용한 네트워크 트래픽 분석 방법에 관한 것이다.

대 표 도 - 도2



특허청구의 범위

청구항 1

플로우에 의한 트래픽 분석을 위한 룰의 구조에 있어서,

1) 룰의 식별명(rulename);

2) 상기 플로우의 시간 빈(time-bin)의 크기(binsize);

3) 상기 플로우의 필터링 조건(filters);

4) 필터링 된 플루우에 대한 맵리듀스에서 그룹핑에 사용될 그룹키 항목들(groupbykeys);

5) 상기 그룹핑에 의한 그룹 내에서 정렬에 사용될 정렬키 항목들(sortbykeys);

6) 정렬된 키 항목의 값에 대해서 유일값 연산을 할 것인지 합계연산을 할 것인지를 나타내는 통계방법(unique/count);

7) 통계방법에 따른 계산 결과에 대한 임계치 기반의 탐지조건(conditions); 및

8) 상기 탐지조건에 맞는 레코드의 반환할 필드(retvals);

을 포함하는 항목을 정의하는 것을 특징으로 하는 룰 정의 방법.

청구항 2

제 1 항에 있어서,

하나의 항목에 구분자 ','를 사용하여 하나 이상의 소항목이 기재되는 것을 특징으로 하는 룰 정의 방법.

청구항 3

제 1 항에 있어서,

상기 항목 중 별도의 제한이 필요없는 경우 'null' 또는 'all' 값을 항목 값으로 사용하는 것을 특징으로 하는 룰 정의 방법.

청구항 4

제 1 항 내지 제 3 항 중 어느 한 항의 방법에 의해 정의된 룰을 이용한 네트워크 트래픽의 분석방법.

청구항 5

제 4 항에 있어서,

(A) 잡 실행명령으로부터 제 1 항 내지 제 3 항 중 어느 한 항의 룰 구조를 갖는 룰 파일이 생성되는 단계;

(B) 상기 단계에서 생성된 룰 파일이 각 태스크로 전달되는 단계;

(C) 상기 각 태스크에서 전달받은 플로우 패킷을 읽어 플로우 레코드를 추출하고 상기 룰에 정의된 필터링 조건에 따라 레코드가 필터링되는 단계;

(D) 필터링 단계를 통과한 레 코드에 대해서 룰에 정의된 그룹키 항목들과 정렬키 항목들이 생성되는 맵단계;

(E) 상기 맵단계에서 생성된 그룹키와 정렬키에 의해 레코드가 그룹핑 및 정렬되어 리듀서로 전달되는 단계;

(F) 리듀서에서 전달받은 레코드를 통합하여 룰에 정의된 통계방법에 의해 연산하는 단계;

(G) 상기 연산결과를 룰에 정의된 조건과 비교하는 단계; 및

(H) 상기 조건에 부합하는 레코드에 대해 룰에 정의된 필드를 반환하는 단계;

를 포함하는 것을 특징으로 하는 네트워크 트래픽의 분석 방법.

명 세 서

기 술 분 야

[0001] 본 발명은 사용자들이 분석하고자 하는 대상과 방법을 쉽게 정의하고 분석할 수 있도록 하는 맵리듀스 방식의 트래픽 분석을 위한 룰 정의 방법과 이에 의해 정의된 룰을 이용한 보편적이고 일원화된 맵리듀스 방식의 트래픽 분석 방법에 관한 것이다.

배 경 기 술

[0002] 네트워크 트래픽 분석은 네트워크 관리에서 기본적이면서도 가장 중요한 분야 중 하나로 트래픽에서 발생하는 모든 문제를 신속하게 파악하는 것이다. 네트워크 트래픽의 측정 및 분석은 네트워크의 운영 상태와 트래픽 특성 파악, 설계 및 계획, 유해한 트래픽의 차단, 과금, QoS(Quality of Service) 보장을 위해 필수적이다. 현재의 인터넷 환경에서는 단일 링크 모니터링의 경우에도 초당 수 기가비트의 트래픽을 다룰 수 있는 처리성능이 필요하며, 고속으로 유통되는 대량의 트래픽을 수집하고 분석할 수 있는 고성능 장비가 필요하다. 고속의 트래픽 처리 기술로서 시스코에서는 넷플로우(NetFlow)를 제안하였는데, 이는 패킷들을 플로우로 요약하여 초당 수 기가비트의 트래픽을 모니터링할 수 있는 방법이다.

[0003] 넷플로우는 동일한 5-튜플(IP주소, 전송계층 포트번호, 전송계층 프로토콜)의 플로우에 대하여 프로파일 정보만을 업데이트하고, 플로우가 종료되거나 일정 시간이 경과하면, 플로우 정보를 전송한다. 이 방식은 패킷 스트림을 플로우 기준으로 축약한 정보만을 유지함으로써 적은 대역폭을 유지하고 분석 작업에서의 계산 비용을 감소시킨다. 상기와 같은 플로우 단위 분석 툴의 경우 단일 서버에서 작동하게 된다. 그러나 최근 사용자 및 트래픽의 증가로 인해 패킷 분석보다는 높은 성능을 기대할 수 있다고 하더라도 단일 서버에서 작동하는 플로우 분석 역시 서버의 성능이 오버헤드로 작용하여 트래픽 분석 속도가 저하될 수 있다는 문제가 있다. 이러한 문제점은 수 100Mbps에서 수 10Gbps에 달하는 고속의 인터넷 망에서 대용량 트래픽을 처리하는 라우터로부터 대용량 플로우 데이터를 수집하여 분석하는 시스템에서 더욱 심각하게 발생한다. 따라서 트래픽 측정을 위하여 빠른 시간 내에 플로우 데이터를 분석하고 그 결과를 사용자에게 전달하기 위해서는 서버의 성능이 우수하여야 하므로 고성능의 서버를 필요로 하여 비용면에서 부담이 된다. 이러한 문제를 해소하기 위하여 본 발명자들은 맵리듀스를 이용한 병렬연산에 의해 트래픽을 측정 및 분석하는 방법을 등록특허 제10-1079786호에서 제안한 바 있다.

[0004] 네트워크 침입탐지 시스템(Intrusion Detection System)은 모니터링과 분석의 두 가지 요소를 필요로 한다. 모니터링을 위해서는 데이터가 수집됨과 동시에 집계가 진행되어 사용자에게 결과를 빠르게 전달해야 한다. 네트워크 침입탐지 시스템에서 분석의 주된 일은 원시 데이터를 읽어 정상과 이상으로 분류하는 것이다. 네트워크 트래픽 데이터로부터 이상을 탐지하기 위한 기술로는 통계적 방법에 기반한 기술과 패턴매칭을 통한 방법이 존재한다.

[0005] 통계적 방법의 이상탐지 기술은 이상과 정상을 판단하기 위해 통계 데이터를 토대로 이상의 징후를 판단하는 방법으로, 데이터마이닝, 머신러닝, 신경계네트워크, 통계모델 등이 대표적인 기술이다. 이러한 통계적 방법에 의한 탐지는 알려지지 않은 공격에 대해서도 탐지가 가능하다는 장점이 있다. 다만, 정상과 비정상을 판단하기 위한 학습과정이 필요하다.

[0006] 패턴 매칭을 이용한 이상탐지 기술은 알려진 공격에 대한 패턴을 수집하여 탐지에 활용하는 기술로서 높은 정확도를 보여준다. 이 방법은 알려진 공격 시그너쳐에 대해 텍스트의 매칭이나 바이너리의 시퀀스 매칭을 수행하

여 탐지한다. 그러나 원시 패킷 데이터에 대한 지속적인 매칭의 신속한 처리를 위해서는 많은 컴퓨팅 자원을 요구하므로 이를 경감하기 위한 방법이 요구된다.

[0007] 룰(Rule) 방식은 이상을 탐지하거나 트랙픽을 분류하기 위해 상기의 통계적 방법과 패턴 매칭 방법을 쉽게 정의하기 위한 방식으로, 룰(Rule)이라는 문자열로 이를 판단하기 위한 규칙을 정의하여 이를 기준으로 판단하는 트래픽 분석의 실행 방법이다. 현재까지 룰 방식의 다양한 분석 방법에 대한 연구가 있으나, 대규모 인터넷 트래픽 모니터링 방법을 위하여 이를 맵리듀스 방식에 적합하도록 제안한 맵리듀스에서의 룰 기반의 분석 방법에 대해서는 아직 개발된 바 없다.

선행기술문헌

특허문헌

[0008] (특허문헌 0001) 등록특허 제10-1079786호

발명의 내용

해결하려는 과제

[0009] 상기와 같은 종래기술의 문제점을 해소하기 위한 본 발명의 목적은 맵리듀스 방식을 이용하여 룰을 기반으로 사용자들이 분석하고자 하는 대상과 방법을 쉽게 정의하고 분석할 수 있도록 하는 맵리듀스 지향의 룰 정의 방법과, 정의된 룰을 이용한 보편적이고 일원화된 맵리듀스 방식의 트래픽 분석 방법을 제공하는 것이다.

과제의 해결 수단

[0010] 전술한 목적을 달성하기 위한 본 발명은 맵리듀스 방식의 트래픽 분석을 위한 룰 정의 방법에 관한 것으로, 보다 상세하게는 1) 룰의 식별명(rulename); 2) 상기 플로우의 시간 빈(time-bin)의 크기(binsize); 3) 상기 플로우의 필터링 조건(filters); 4) 필터링 된 플루우에 대한 맵리듀스에서 그룹핑에 사용될 그룹키 항목들(groupbykeys); 5) 상기 그룹핑에 의한 그룹 내에서 정렬에 사용될 정렬키 항목들(sortbykeys); 6) 정렬된 키 항목의 값에 대해서 유일값 연산을 할 것인지 합계연산을 할 것인지를 나타내는 통계방법(unique/count); 7) 통계방법에 따른 계산 결과에 대한 임계치 기반의 탐지조건(conditions); 및 8) 상기 탐지조건에 맞는 레코드의 반환할 필드(retvals);을 포함하는 항목을 정의하는 것을 특징으로 하는 룰 정의 방법에 관한 것이다.

[0011] 또한 본 발명은 상기 방법에 의해 정의된 룰을 이용한 네트워크 트래픽 분석 방법에 관한 것이다.

발명의 효과

[0012] 이상과 같이 본 발명에 의하면, 대규모의 트래픽 데이터를 복잡한 통계와 패턴을 이용해 분석하는데 있어서 맵리듀스 방식을 채택하여 일원화된 맵리듀스 지향 룰 정의 방법을 제시한다.

[0013] 또한 룰을 통해 정의된 방법에 의해 네트워크 트래픽 분석을 위해 단일 패스의 맵리듀스 처리 방법을 제시하여 다양하고 복잡한 트래픽 분석에 대한 빠른 분석과 대규모 트래픽에 대한 안정적인 분석을 모두 제공할 수 있다.

도면의 간단한 설명

[0014] 도 1은 종래기술에 의한 일반적인 네트워크 침입탐지 시스템에서 수행하는 트래픽 분석과 이상탐지를 위한 흐름도.

도 2는 본 발명에 의해 정의된 맵리듀스 방식의 트래픽 분석을 위한 룰 구조.

도 3은 본 발명에 의한 룰 정의의 일 실시예로 그룹별 통계작업과 이상탐지를 위해 정의된 룰의 예시.

도 4는 본 발명에 의해 정의된 룰을 이용한 플로우 분석의 일실시예로 하둡에서의 플로우의 분석을 위한 맵리듀스 프레임워크 예시도.

도 5는 본 발명에 의해 정의된 룰을 이용한 맵리듀스 기반의 트래픽 처리 구조도.

도 6은 포트스캔 탐지를 위해 정의된 룰 및 맵리듀스의 실제 실행과정의 예시도.

발명을 실시하기 위한 구체적인 내용

[0015] 이하 첨부된 도면을 참조하여 본 발명을 보다 상세히 설명한다. 그러나 이러한 도면은 본 발명의 기술적 사상의 내용과 범위를 쉽게 설명하기 위한 예시일 뿐, 이에 의해 본 발명의 기술적 범위가 한정되거나 변경되는 것은 아니다. 또한 이러한 예시에 기초하여 본 발명의 기술적 사상의 범위 안에서 다양한 변형과 변경이 가능함은 당업자에게는 당연할 것이다.

[0016] 도 1은 네트워크 침입탐지 시스템에서 수행하는 트래픽 모니터링과 이상탐지의 흐름을 보여준다. 침입탐지 시스템에서 수행하는 주요한 일은 트래픽의 모니터링과 이상 탐지이다. 트래픽의 모니터링은 유입되는 플로우에 대해 일정시간 단위로 통계를 추출하여 제공하는 것에 의해 이루어지며, 필요에 따라서 모니터링 하고자 하는 특정 구역을 지정하기 위한 필터링 과정이 추가된다. 이상 탐지는 필터링된 입력 플로우 레코드 항목들의 패턴을 검사하여 침입을 판단하거나, 매칭된 레코드를 일정 타임 윈도우 동안 빈에 담아두고 누적한 통계값을 임계치를 기준으로 검사하여 네트워크에의 침입 여부를 판단한다. 포트스캔이나 죽음의 핑(Ping of Death)의 탐지가 대표적인 이상탐지의 예라고 할 수 있다.

[0017] 도 2는 본 발명에 의해 정의된 맵리듀스 방식의 트래픽 분석을 위한 룰 구조를 나타낸다. 본 발명의 방법에 의해 정의된 룰에 의하면 트래픽 모니터링을 위한 분석은 물론 이상탐지를 위한 패턴매칭 및 임계치를 기준으로 한 통계적 검사를 통합한 트래픽 분석이 가능하다. 룰을 구성하는 항목은 ';'의 구분자로 구분되며, 각 항목은 ','의 구분자를 이용하여 하나 이상의 소 항목으로 구성된 복합적인 키(Key)나 조건(Condition)을 구성할 수 있다.

[0018] 각각의 항목에 대한 설명은 하기 표 1에 기재하였다. 하기 표 1의 항목 중 별도의 제한을 원하지 않는 항목의 경우, 'null' 또는 'all' 값을 사용하여 모든 플로우 또는 레코드에 대해 분석을 하도록 룰을 정의할 수있다.

표 1

항 목	설 명
rulename	· 룰의 식별명 · 그룹키 생성 시 사용됨
binsize	· 룰의 시간적 제약조건을 설정 · 시간에 따른 빈(bin)의 파이셔닝을 위해 그룹키 생성 시 사용됨
filter	· feature vector에 대한 패턴 · 맵단계에서 플로우의 필터링에 사용됨
groupbykeys	· rulename, binsize와 함께 그룹키로 사용되는 필드들을 설정 · 맵의 결과를 리듀서에 파티셔닝하고 리듀스 내에서 그루핑하는데 사용됨
sortbykeys	· 레코드에 대한 보조 정렬 키 · 맵의 결과의 보조정렬에 사용
unique/count	· 빈에 포함된 레코드의 통계값의 계산방법 · 리듀서에서 빈에 포함된 레코드들로부터 unique한 키값을 계산하거나 값을 합하는 연산을 하도록 설정
conditions	· 통계값에 대한 임계치 기반의 탐지 조건을 설정 · 조건 값에 따라 통계값을 비교하여 이상을 탐지
retvals	· 반환할 필드를 설정 · conditions에 제시된 조건이 만족할 경우 해당 필드의 값을 반환

[0019]

[0020] 도 3은 본 발명에 의한 룰 정의의 일 실시예로 트래픽 모니터링을 위한 그룹별 통계작업과 이상탐지를 위해 정의된 룰의 예시를 각각 보여준다.

[0021] 먼저, 1)은 트래픽 모니터링을 위한 그룹별 통계의 한 예로 출발지 AS를 기준으로 5분 동안의 트래픽 통계를 구하기 위해 정의한 룰의 예시이다. 5분 동안의 모든 플로우 레코드에 대하여 통계를 구할 것이므로 필터(filters)와 조건(conditions) 값은 각각 'null'과 'all'로 정의하였다. 5분 타임빈에 대해 출발지 AS를 그룹키로 하여 바이트수, 패킷수, 플로우 수를 구할 것이므로, groupbykeys는 출발지 AS(sas)로, 반환할 필드(retvals)는 바이트수, 패킷수, 플로우 수가 된다. 동일 retvals 항목에서 bytes, pkts 및 flows는 전술한 바와 같이 구분자 ','로 구분하였다. 총 트래픽의 통계를 구하는 것이 목적이므로 정렬은 필요하지 않으므로 sortbykeys는 null로 정의하였다.

[0022] 2)는 트래픽 모니터링을 위한 그룹별 통계의 또 다른 한 예로 서브넷을 기준으로 1분 동안 출현한 IP 수를 구하기 위한 룰을 보여준다. 1분 동안의 모든 플로우 레코드에 대하여 통계를 구할 것이므로 필터(filters)와 조건(conditions)은 각각 'null'과 'all'로 정의하였다. 1분 타임빈에 대해 도착지 subnet 주소를 기준으로 ip 출현개수를 구할 것이므로, groupbykeys는 dsubnet으로하고, sip|dip(출발지 IP|도착지 IP)로 정렬한 후 IP 수를 계산하도록 정의하였다.

[0023] 3)은 이상탐지에 대한 룰의 예시로 각각 포트스캔과 죽음의 핑 공격 탐지를 위한 패턴매칭과 통계에 기반한 분석 룰을 보여준다. 먼저 포트스캔의 경우, 300초 동안 TCP 플로우를 대상으로, 동일한 출발지와 도착지의 호스트 쌍의 연결에서 유일한 포트 수의 통계를 구하여 상이한 포트 수의 출현이 20 개 이상일 경우 포트스캔으로 판단하여 호스트 쌍의 IP를 출력하도록 한다. 죽음의 핑 공격은 300초 동안 ICMP(인터넷 제어 메시지 프로토콜) 플로우를 대상으로, 하나의 출발지 IP 주소로부터 동일한 도착지 IP 주소로 보낸 패킷의 수가 10개 이상인 경우 죽음의 핑 공격으로 판단하여 호스트 쌍의 IP를 출력하도록 한다.

[0024] 상기와 같은 예시에 기초하여 트래픽 모니터링 및 이상탐지를 위한 다양한 룰을 정의하는 것이 가능하다.

[0025] 도 4는 본 발명에 의해 정의된 룰을 이용한 플로우 분석의 일실시예로 하둡에서의 플로우의 분석을 위한 맵리듀

스 프레임워크의 예시도이다. 하둡을 구성하는 각 노드들은 넷플로우를 수집하기 위한 수집모듈과 수집된 트래픽을 저장하기 위한 저장모듈, 저장된 트래픽을 분석하는 맵리듀스 기반의 분석 모듈을 가진다. 사용자가 분석을 위한 잡 실행명령을 내리면 명령을 룰로 변환하여 룰 파일을 생성한다. 룰 파일을 트래픽 분석의 형태에 따라 통계 룰과 탐지 룰로 나눌 수 있다. 즉, 네트워크 모니터링을 위한 명령의 경우에는 통계 룰 파일이, 이상탐지를 위해서는 탐지 룰 파일이 생성된다. 그러나, 이들은 편의를 위한 분류이며 본 발명에 의한 룰의 항목이 각 역할에 따라 적합하게 정의된 것으로 그 기본적은 구성은 동일하다. 생성된 룰 파일들은 맵리듀스 잡이 실행되기 전 하둡 분산캐쉬 (DistributedCache)를 통해 모든 노드에 분산되고, 맵리듀스 잡에 의해 컴파일되어 룰에 정의된 조건에 따라 네트워크 모니터링과 이상 탐지 작업이 실행된다.

[0026]

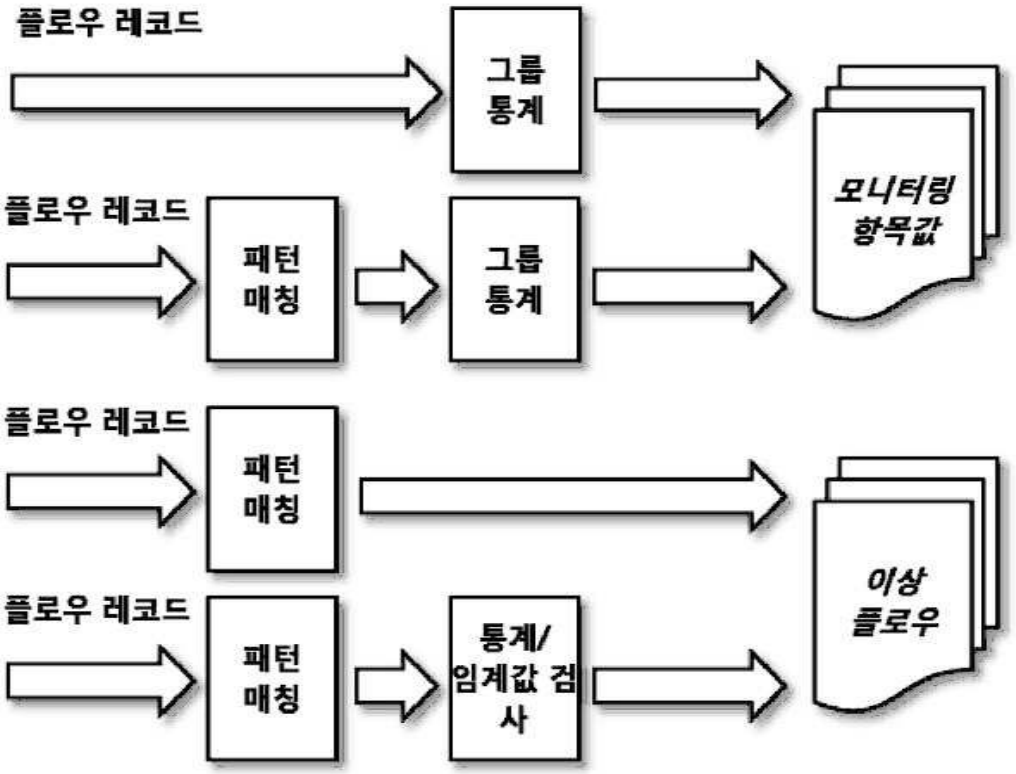
[0027] 도 5는 본 발명에 의한 룰을 이용한 맵리듀스 기반의 트래픽 분석 처리 구조로 트래픽 분석을 위한 맵리듀스 과정 동안 룰을 읽고, 각각의 항목을 이용하여 처리하는 과정을 보여준다. 본 발명은 다양한 분석을 위해 단일의 맵리듀스 과정을 거쳐 모든 룰에 대한 분석을 가능하도록 하는 방법을 제공한다.

[0028] 맵리듀스 잡이 시작되면 룰 파일들이 분산캐쉬를 통해 각 태스크로 전달된다. 맵과 리듀스 태스크는 룰 파일을 파싱하여 맵(Map) 자료구조에 저장한다 (501). 맵퍼는 패킷을 읽어 플로우 레코드를 추출한 다음 필터에 정의된 값을 플로우 레코드 필드값에 매핑하여 일치여부를 판단하고, 필터값과 일치된 레코드에 대해서 리듀서로 전달하기 위한 복합 키(composite key)와 값(value)를 생성한다. 복합키는 그룹키(group key) 와 정렬키(sort key)로 나뉘며 그루핑과 그룹 내의 정렬에 각각 사용된다. rulename, binsize로 마스킹한 timestamp 값, 그리고 groupbykeys항목에 정의된 필드들을 조합하여 그룹 키를 생성하고 sortbykey 항목에 정의된 필드를 정렬키로 생성하여 맵의 결과로 내보낸다(502). 중간 결과는 맵퍼에서 생성한 group key를 이용한 파티셔닝과 그룹핑이 수행된 후 리듀서로 전달되거나 컴바인(combine) 과정을 거친 후 리듀서로 전달된다(503). 리듀서는 파티셔닝 정책에 따라 전달받은 키들에 포함된 레코드들은 룰에 정의된 연산방법에 따라 키의 유일(unique) 연산 또는 값(value)의 카운트 연산을 한다. 연산결과가 condition 항목에 정의된 규칙에 부합하는지를 판단하여 부합할 경우 retval에 정의된 항목값을 key 값과 함께 결과로 내보낸다(504).

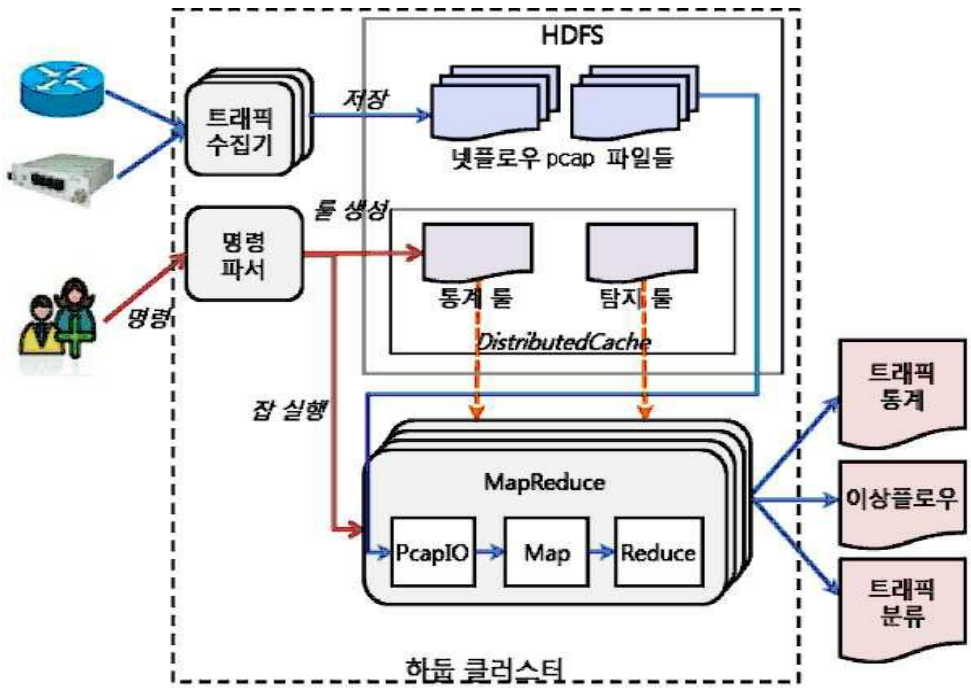
[0029] 도 6은 도 5의 설명을 보다 구체적인 예로 설명하기 위한 것으로, 포트스캔 탐지를 위한 룰 및 맵리듀스의 실제 실행과정의 예시도이다. 포트스캔은 운영 중인 서버에서 열려 있는 TCP/UDP 포트를 검색하는 것으로 실제 공격을 실행하기 전 취약점을 파악하기 위해 수행되는데, 이는 플로우 분석을 통해서 쉽게 탐지할 수 있는 이상행위 중 하나이다. 맵퍼는 패킷을 읽어 플로우 레코드를 추출한 후, 프로토콜 값을 확인하여 TCP에 해당하는지를 검사한다. TCP에 해당하는 플로우에 대하여 rulename, binsize로 마스킹한 timestamp 값, 그리고 groupbykeys항목에 정의된 필드들을 조합한 "portscan"|(date|300)|sip|dip를 그룹키로 생성하고 sortbykeys 항목에 정의된 필드인 dport를 정렬키로 생성하여 맵의 결과로 내보낸다. 맵의 결과로 생성된 레코드들은 그룹키인 출발지와 도착지 호스트 주소 쌍을 기준으로 파티셔닝되어 리듀서에 할당된다. 이 때 동일한 그룹키로 묶여진 레코드들은 정렬키인 포트번호에 의해 정렬된 상태가 된다. 리듀서에서는 파티셔닝 정책에 따라 동일한 그룹키, 즉 출발지와 도착지 호스트 주소 쌍을 갖는 플로우의 포트번호의 변경을 감지하여 상이한 포트번호의 출현횟수를 누적한다. 마지막으로 누적된 포트번호 수가 20개 이상인지를 검사하여 참인 경우 포트스캔이 수행된 것으로 판단하고 호스트 주소 쌍을 결과로 내보냄으로써 포트스캔 탐지가 완료된다.

도면

도면1


플로우 레코드
그룹 통계
모니터링 항목값
플로우 레코드
패턴 매칭
그룹 통계
플로우 레코드
패턴 매칭
이상 플로우
플로우 레코드
패턴 매칭
통계/ 임계값 검사

도면2


HDFS
트래픽 수집기
저장
넷플로우 pcap 파일들
명령 파서
룰 생성
통계 룰
탐지 룰
DistributedCache
명령
잡 실행
MapReduce
PcapIO
Map
Reduce
트래픽 통계
이상플로우
트래픽 분류
하둡 클러스터

도면3

도면4

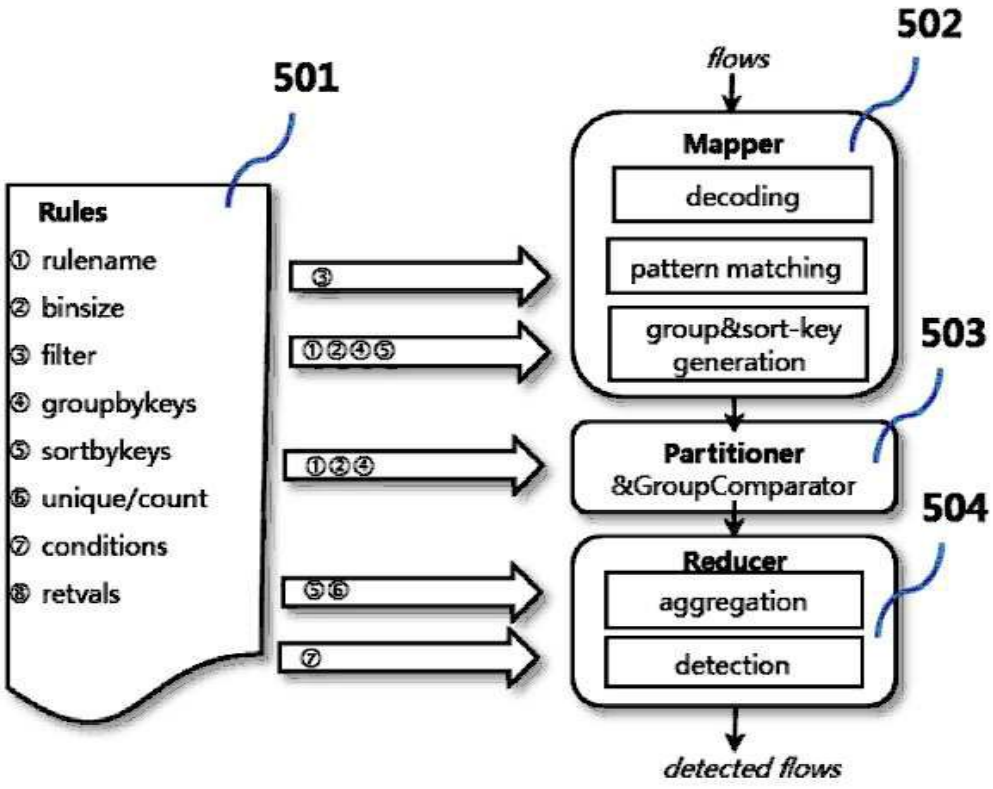
 # 룰 예시

1) source AS 를 기준으로 5분 동안의 트래픽 통계 구하기
count_All;300;null;sas;null;count;all;bytes,pkts,flows

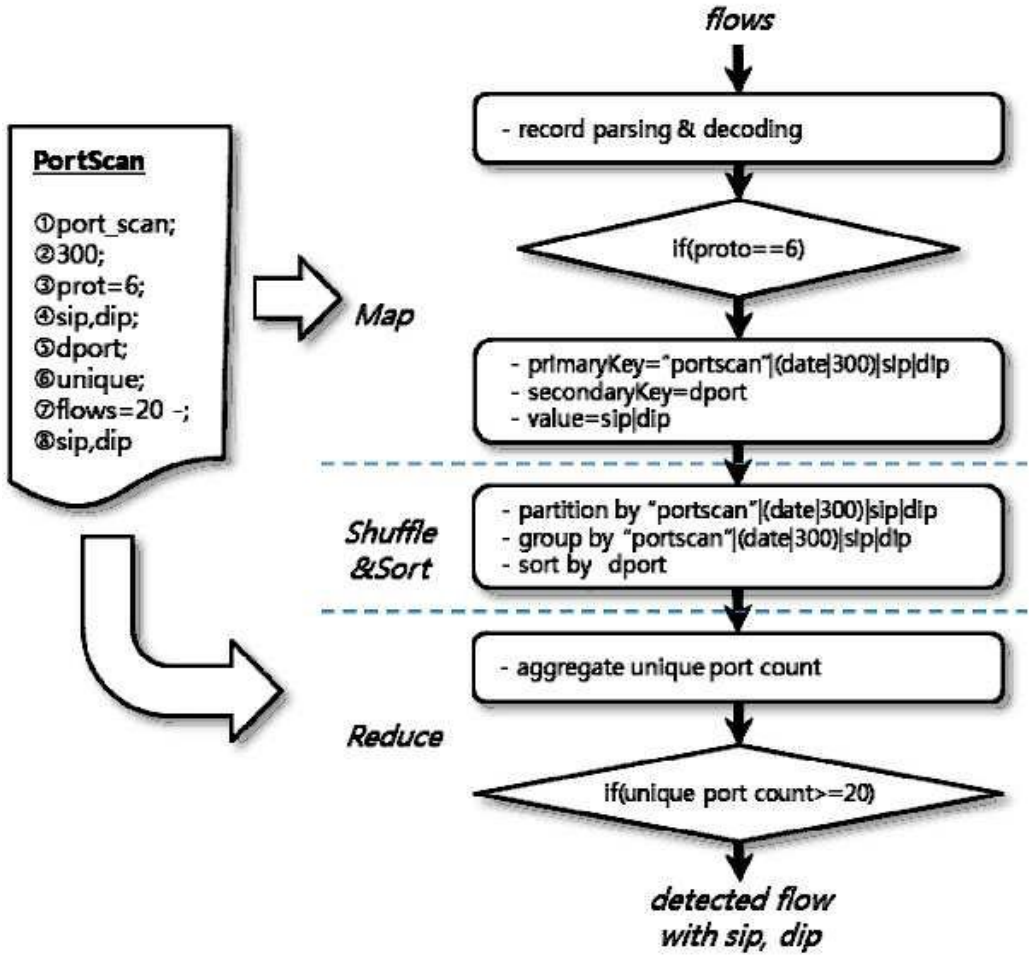
2) destination subnet 주소를 기준으로 1분 동안의 IP의 출현 개수 구하기
count_IPs;60;null;dsubnet;sip|dip;unique;all;count

3) 이상탐지를 위한 룰 예시 (portscan, pingOfDeath)
port_scan;300;prot=6;sip,dip;dport;unique;flows=20-;sip,dip
ping_of_death;300;prot=1;sip;dip;count;pkts=10-;sip,dip

도면5



도면6


flows
- record parsing & decoding
if(proto==6)
PortScan
①port_scan;
②300;
③prot=6;
④sip,dip;
⑤dport;
⑥unique;
⑦flows=20 -;
⑧sip,dip
Map
- primaryKey="portscan"|(date|300)|sip|dip
- secondaryKey=dport
- value=sip|dip
Shuffle &Sort
- partition by "portscan"|(date|300)|sip|dip
- group by "portscan"|(date|300)|sip|dip
- sort by dport
- aggregate unique port count
Reduce
if(unique port count>=20)
detected flow with sip, dip