데이터 피처 설명

📌

- ID : 샘플별 고유 ID
- ip_src : 송신 IP 주소
- port_src : 송신 포트 번호
- ip_dst : 수신 IP 주소
- port_dst : 수신 포트 번호
- protocol : 프로토콜 종류
- duration : 통신 지속 시간(초)
- pkt_count_fwd : 송신 방향 패킷 수
- pkt_count_bwd : 수신 방향 패킷 수
- rate_fwd_pkts : 송신 패킷 전송 속도
- rate_bwd_pkts : 수신 패킷 전송 속도
- rate_fwd_bytes : 송신 바이트 전송 속도
- rate_bwd_bytes : 수신 바이트 전송 속도
- payload_fwd_mean : 송신 페이로드 평균 바이트
- payload_bwd_mean : 수신 페이로드 평균 바이트
- tcp_win_fwd_init : 송신 측 TCP 윈도우 초기값
- tcp_win_bwd_init : 수신 측 TCP 윈도우 초기값
- tcp_syn_count : SYN 패킷 수
- tcp_psh_count : PSH 패킷 수
- tcp_rst_count : RST 패킷 수
- iat_avg_packets : 패킷 간 평균 간격 시간(초)
- attack_type : 공격 유형 또는 정상(Benign)

▼ 기본 용어 및 개념

💡 **포트 번호**란?

네트워크 통신에서 특정 서비스나 프로세스를 식별하기 위해 사용되는 번호

즉, 컴퓨터에 여러 서비스가 실행 중일 때, **어느 프로그램에 데이터를 전달해야 하는 지를 알려주는 번호**

[예시]

1. 한 웹사이트에 접속 → 대상 서버의 IP로 접속
    - 포트 80번(HTTP) 사용
2. 파일 다운로드
    - 포트 21번(FTP) 사용
3. 원격 접속 (SSH)
    - 포트 22번 사용
4. 메일 확인
    - 포트 993번 (IMAP), 포트 110번 (POP3) 등 사용

📍 **보안 관점에서의 의미**

- 공격자가 특정 포트를 스캔(Port Scanning)하여 열려 있는 서비스를 탐색

→ 포트 번호는 **공격 탐지에 있어 중요한 단서**

📍 **분석 관점에서의 의미**

- 포트 번호의 범주가 너무 많으므로 **주요 포트만 그룹화**하거나 **포트 범위를 나누는 것**이 일반적

[포트 번호 범위]

이름	포트 번호 범위	설명
0 ~ 1023	잘 알려진 포트(Well-known port)	시스템 사용 번호
1024 ~ 49151	등록된 포트(Registered port)	특정 프로토콜이나 어플리케이션에서 사용하는 번호
49152 ~ 65535	동적/사설 포트 (Dynamic/Private port)	어플리케이션에서 임시로 사용하는 포트 (랜덤 할당)

💡 **프로토콜**이란?

컴퓨터나 네트워크 장치들이 통신할 때 데이터를 주고받기 위해 사용하는 규칙의 집합

즉, **데이터 통신을 위한 일종의 언어나 규약**

💡 **패킷**이란?

네트워크 내 출발지와 목적지 간 라우팅 되는 데이터 단위

즉, 통신망을 통해 전송하기 쉽 자른 데이터의 전송 단위

Package(패키지) + Bucket(버킷) = Packet(패킷)

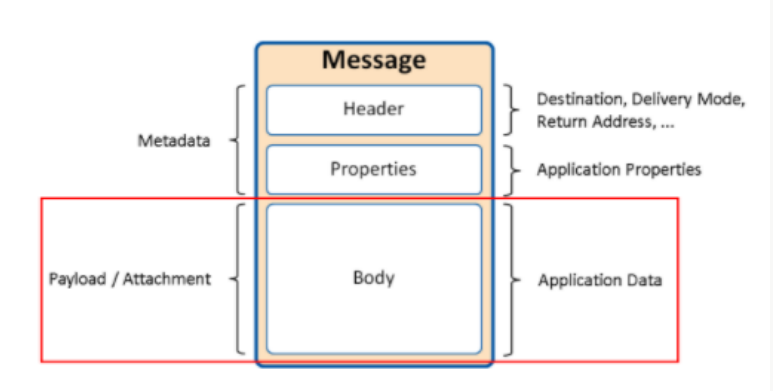
- 라우팅: 네트워크에서 데이터를 주고받을 때 최적의 경로를 선택하고 데이터를 전송하는 과정

💡 **페이로드**란?

전송되는 데이터 자체를 의미

→ 보통 데이터 전송 시엔 헤더, 바디, 메타 데이터와 같은 정보로 이루어져 있지만 여기서 **전송 목적이 되는 데이터만 포함** (함께 전송되는 헤더나 메타 데이터는 제외)

즉, **데이터 중 사용자가 '흥미 있는' 데이터를 구별**하는 데에 사용



💡 **TCP 윈도우(Window Size)**란?

수신 측이 한 번에 수용 가능한 데이터의 양 (단위: Byte)

1. 기본 식별 정보

- **ID:** 샘플별 고유 ID
- **ip_src:** 송신자 IP 주소 (IP Source)
  - 송신 측의 네트워크 식별 정보
  - 공격자 IP 주소 탐지 가능
- **port_src:** 송신자 포트 번호 (Port Source)
  - 특정 포트 번호에서 반복적으로 발생하는 이상 행동 탐지에 도움
- **ip_dst:** 수신자 IP 주소

    - 수신 측 네트워크 식별 정보
    - 피해자 IP 탐지 가능

- **port_dst:** 수신자 포트 번호
    - 수신 측 포트 번호 (Port Distination)
    - 주로 서비스 구분
    - 특정 포트 대상 공격(ex. 포트 스캐닝) 확인 가능

- **protocol:** 사용된 전송 계층 프로토콜
    - 공격 유형에 따라 사용하는 프로토콜이 다를 수 있음

> 📌 제공된 데이터 내 protocol의 범주 수 2개
>
> 1. **TCP (Transmission Control Protocol, 전송 제어 프로토콜)**
>
>   : 컴퓨터 네트워크에서 데이터를 **신뢰성** 있게 전송하기 위한 프로토콜
>
>   - 신뢰성 ⬆️ 속도 ⬇️ → 높은 안정성
>
> 2. **UDP (User Datagram Protocol)**
>
>   : 네트워크에서 데이터를 전송하는 프로토콜 중 하나
>
>   - 신뢰성 ⬇️ 속도 ⬆️
>
>   → 속도가 중요하고 약간의 데이터 손실을 감수할 수 있는 환경에서 유용하게 사용되는 프로토콜

---

2. 시간 관련 변수

- **duration**: 통신 지속 시간(초)
    - 한 세션의 총 지속 시간
    - 공격은 짧거나 길게 유지되는 패턴이 존재함

- **iat_avg_packets:** 패킷 간 평균 시간 간격(초)
    - 패킷 사이 간격의 평균
    - 특히 **DoS, DDoS 공격 탐지**에 유용한 변수

📍 서비스 거부(DoS) 공격과 분산 서비스 거부(DDoS)

1. **서비스 거부(DoS)**

   : 서버에 트래픽을 범람시켜 웹 사이트 또는 리소스를 사용할 수 없게 만듦

2. **분산 서비스 거부(DDoS)**

   : 여러 대의 컴퓨터 또는 기계를 이용하여 표적 리소슬르 ㅇ범람시키는 서비스 거부(DoS) 공격

→ **모두 서비스 중단을 목표로 서버 또는 웹 어플리케이션에 과부하를 발생** 시킴

⚠️ DoS, DDoS 공격에서는 짧은 시간 안에 수많은 패킷이 전송됨

→ 패킷 간 평균 시간 간격이 **정상 트래픽보다 매우 작아짐**

즉, 매우 작은 값이면 공격 타입일 가능성이 높음

3. 패킷 수 및 속도 관련 변수

- **pkt_count_fwd**: **송신** 방향 패킷 수
  - **클라이언트 → 서버 방향** 패킷 개수
- **pkt_count_bwd**: **수신** 방향 패킷 수
  - **서버 → 클라이언트 방향** 패킷 개수

⚠️ 패킷의 수가 **한 쪽 방향으로만 집중되면 비정상일 가능성 존재**

→ 송신 및 수신 방향의 패킷 개수의 비율 비교

- **rate_fwd_pkts**: **송신** 방향 초당 패킷 전송 수
  - `pkt_count_fwd / duration` 와 유사한 의미를 가짐
  - **송신 트래픽 밀도 파악**에 유용

공격 유형	특징	설명
DoS/DDoS	매우 높음	많은 수의 짧은 요청을 빠르게 송신 → 서버 과부하 유도
Brute Force (무차별 대입 공격)	높은 송신률, 응답 적음	로그인 시도 반복, 요청은 많고 응답은 적음
Port Scan	매우 높음, 응답 거의 없음	포트 수백/수천 개에 빠르게 송신
정보 유출	송신 측 총 패킷 길이, 즉 데이터 총량이 높음	대량의 데이터를 외부로 송신 중
정상 트래픽	일정한 송신 속도, 요청-응답 균형	규칙적인 사용자 행동 패턴

- **rate_bwd_pkts: 수신** 방향 초당 패킷 전송 수
  - `pkt_count_bwd / duration` 와 유사한 의미를 가짐
  - **비정상적인 응답 패턴 탐지** 가능

상황		의미
정상 트래픽	일정하거나 상대적으로 낮음	요청-응답 패턴이 균형 잡혀 있음
DoS/DDoS 공격	낮거나 0	요청은 많지만 응답을 거의 못 받음 (서버가 마비됨)
Port Scan, Reconnaissance (정찰)	매우 낮음 또는 0	여러 포트에 요청했지만 대부분 응답 없음 (존재하지 않는 포트 대상)
Botnet, C&C(Command & Control) 통신	간헐적으로 매우 높거나 일정 패턴	수신 측 응답이 빠르게 돌아옴 (비정상 트래픽 흐름)
Data exfiltration (데이터 유출)	수신 트래픽 비정상적으로 많음	서버가 대량의 데이터를 공격자에게 전송 중

⚠️

- **정상:** `rate_fwd_pkts` = `rate_bwd_pkts`
- **비정상:** `rate_fwd_pkts` > `rate_bwd_pkts`

  → 요청은 많고 응답이 없음 (서버 다운 및 포트 닫힘 등)
- **정보 유출/악성 트래픽:**

  `rate_fwd_pkts` < `rate_bwd_pkts`

  → 서버가 오히려 많은 데이터를 돌려줌

4. 바이트 전송 속도 및 페이로드 관련 변수

- **rate_fwd_bytes**: **송신** 바이트 전송 속도
  - 초당 송신한 데이터량
  - 특정 공격 유형에서 극단적으로 나옴
- **rate_bwd_bytes**: **수신** 바이트 전송 속도
  - 초당 수신한 데이터량
  - 응답 측에서 데이터가 거의 없거나 과도한 경우 공격 탐지 가능

- **payload_fwd_mean**: **송신** 페이로드 평균 바이트
  - 패킷 당 송신한 실제 데이터(payload)의 평균
- **payload_bwd_mean**: **수신** 페이로드 평균 바이트
  - 응답 측 payload 크기의 평균

> ⚠️ **페이로드(payload)가 거의 없거나 0에 가까운 경우**
>
> : 트래픽이 실질적인 데이터 없이 헤더만 존재하거나, 목적 없는 트래픽이 대량 발생
>
> → 사이버 공격 유형에서 공통적으로 나타남

---

5. TCP 관련 제어 패킷 및 윈도우 정보

- **tcp_win_fwd_init**: **송신** 측 TCP 윈도우 초기값
  - 전송 제어 용량(capacity) 정보
  - 송신 측에서 초기 설정한 수신 버퍼의 크기

    → 즉, **수신 측이 얼마나 데이터를 보내도 되는지** 판단 기준
- **tcp_win_bwd_init**: **수신** 측 TCP 윈도우 초기값
  - 수신 측 수용 가능 용량
  - 패킷 손실, 혼잡 대응과 관련 있음

⚠️ **TCP 관련 비정상 탐지 조건**

- `tcp_win_*_init == 0` 또는 **매우 작음**

  → 의심스러운 TCP 요청, 악성 클라이언트 가능성

- `tcp_win_*_init > 65535` 와 같이 **비정상적 대형 값**

  → 조작된 TCP 패킷 가능성

항목	설명
65535 Byte	TCP 윈도우 기본 한계 (16비트)
65535 초과 값	윈도우 스케일 옵션 적용 시 가능
비정상 여부	매우 큰 값(수백 MB~GB)은 **정상적이지 않음**, 조작 가능성 있음

TCP FLAG 변수 (SYN, PSH, RST)

- **tcp_syn_count**: SYN 패킷 수 (Synchronization: 동기화)
  - TCP 연결 시도 횟수

    → SYN Flood 공격은 이 값이 매우 큼

- **tcp_psh_count**: PSH 패킷 수 (Push: 밀어넣기)
  - **즉시 데이터 전송을 요청**하는 플래그가 있는 패킷 수
  - 일반 통신에선 적당한 수준, 이상 트래픽에선 극단적일 수 있음

- **tcp_rst_count**: RST 패킷 수 (Reset: 재연결 종료)
  - 세션 강제 **종료** 시도 횟수, 양방향에서 동시에 일어나는 **중단 작업**
  - 포트 스캐닝, 취소된 연결 등에서 많이 나타남

---

6. 타겟 변수

- **attack_type**: 공격 유형 또는 정상 여부

📌 **타켓 변수 내 Categories**

- **Benign (정상)**
- Hulk
- Port_Scanning
- DDoS
- FTP_Brute_Force
- GoldenEye
- Slow_HTTP
- SSH_Brute_Force
- Botnet
- Slowloris
- Web_Brute_Force
- Web_XSS

---

1. 정상

- **Benign(정상)**: 정상 트래픽

2. Dos 및 DDoS 기반 공격

1. **Hulk (HTTP Unbearable Load King)**

   : DDoS 공격 도구 중 하나로, 웹 서버의 가용 용량을 고갈시키기 위해 설계

   → HTTP GET 요청 Flooding 기법으로 특정 URL에 대한 요청을 지속적으로 반복하여 웹 서버 과부하 및 서비스 거부를 유발

   → 트래픽 양이 매우 크며 응답 없음 또는 지연 발생

2. **DDoS(Distributed Denial of Service)**

   : 여러 대의 기계를 이용하여 특정 서버나 네트워크에 과도한 트래픽을 발생시켜 서비스 중단 및 마비를 유발

3. **GoldenEye**

   : HTTP GET 및 POST 플러드 공격을 수행하는 DDoS 공격 도구

3. 스캐닝 및 탐지 탐색형 공격

- **Port_Scanning**

  : 공격자가 열린 포트를 찾기 위해 많은 포트에 접속을 시도

  → SYN 패킷만 보내고 응답 확인 → **대부분 페이로드 없음**

4. 인증 우회 및 무차별 대입 공격 (Brute Force)

1. **FTP_Brute_Force**

   : FTP 서비스 로그인에 대해 **ID/PW를 반복 대입**하여 인증 우회 시도

   → 짧은 요청 다수, 실패 응답 빈번

   → 주로 자동화된 스크립트 기반 공격

2. **SSH_Brute_Force**

   : SSH(22번 포트) 서버에 대해 ID/PW 조합을 지속적으로 시도

   → 일정한 패턴의 소규모 패킷 다량 발생

   → 로그인 실패 반복 시도 패턴

3. **Web_Brute_Force**

   : 웹 로그인 페이지에서 **다수의 ID/PW 조합**을 제출하며 인증을 시도

   → HTTP POST 요청 반복

   → 웹 서버에 로그인 요청 트래픽 급증

5. 봇넷 및 명령제어(C&C) 기반 활동

- **Botnet**

  : 감염된 기기가 **C&C(Command and Control) 서버와 통신**하며 명령을 수신하거나 실행

  → **일정한 시간 간격으로 트래픽 전송**, 트래픽 패턴은 일정하지만 비정상

  → 소규모 패킷, **Low & Slow 방식**, 간헐적인 응답

6. 느린 HTTP 기반 DoS 공격

1. **Slow_HTTP**

   : HTTP 요청을 **매우 느리게 전송**하여 서버가 연결을 끊지 못하도록 함 → **서버의 스레드 고갈** 유도

2. **Slowloris**

   : 대표적인 Slow HTTP 공격 도구

   → HTTP 헤더만 전송하고 본문을 계속 지연

   → **서버의 연결 대기 큐**를 고갈시켜 정상 사용자 연결 차단

7. 웹 취약점 공격

- **Web_XSS (Cross-Site Scripting)**

  : 웹 애플리케이션에 **악성 JavaScript 코드 삽입**

  → 사용자의 브라우저에서 스크립트가 실행되어 **세션 탈취**, **피싱**, **리디렉션** 등 발생

  → 대부분 **웹 양식, URL, 쿼리 문자열**에 코드 삽입