

방문 집계에 우리 점검 방문이 섞여 수치가 부풀었던 일 — nohit 플래그와 데이터센터 표시로 고친 과정

작성 2026-10-20 · 읽는 시간 약 6분 · 작성 루리의 줍줍 편집

#방문 집계 #Cloudflare #자동화

정적 사이트에 직접 만든 방문 집계(브라우저가 페이지를 열 때 작은 요청을 보내 DB에 한 줄 쌓는 방식)를 붙여 쓰고 있었어요. 어느 날 하루 방문이 평소보다 확 늘어 보였는데, 열어 보니 사람 방문이 늘어난 게 아니라 **우리가 링크가 열리는지 확인하느라 연 방문과 데이터센터에서 온 점검 트래픽**이 섞인 숫자였어요. 이 글은 무엇이 섞였는지 가려낸 방법과, 앞으로 안 섞이게 고친 과정을 적은 기록이에요. 숫자는 모두 이 사이트의 2026-10-20 하루치 방문 기록(677건)에서 다시 센 값이에요.

문제: 하루 방문 214명, 그런데 사람은 몇 명?

그날 기록된 방문은 677건, 서로 다른 익명 방문 번호는 214개였어요. 이걸 기존 컬럼(경로·유입 표기·국가·도시·모바일 여부·로그인 여부)만으로 나눠 봤어요.

구분	규칙	방문 기록	방문 번호
A	로그인 계정(주로 관리자 본인)	296	6
C	미국 데이터센터 밀집 도시 접속	111	46
B	유튜브 표기 + 데스크톱 + 서울, 같은 페이지를 균일하게 반복	114	87
D	위에 안 걸린 나머지(모바일 22%, 레퍼러 다양)	156	76

사람일 가능성이 큰 쪽은 D의 76개 안팎이에요. 규칙에 걸렸다고 모두 사람이 아니라는 뜻은 아니고(도시 이름은 IP 위치 추정이에요), 한 방문 번호가 여러 규칙에 걸릴 수 있어서 단순 합산도 아니에요. 특히 구글 표기 29건 중 23건이 같은 데이터센터 도시, 전부 데스크톱 1페이지 방문이었어요.

원인

- 링크가 열리는지 확인하는 자동 점검(우리 쪽 도구)이 실제 방문처럼 집계됐어요. 이 점검은 브라우저와 똑같이 동작해서 스크립트가 그대로 실행돼요.
- 서버 쪽에는 봇 이름(bot·crawl·headless 등)이 들어간 UA를 거르는 필터가 있었지만, 데이터센터에서 일반 Chrome 문자열로 들어오는 방문은 통과했어요.
- 방문 기록에 UA 종류나 접속 회사 정보가 없어서, 나중에 걸러낼 단서가 도시 이름뿐이었어요.

해결 1: 우리 기기는 집계에서 빼기 (?nohit=1)

점검에 쓰는 브라우저에서 링크 끝에 ?nohit=1을 한 번만 붙여 열면, 그 브라우저에 플래그를 저장하고 이후 방문 기록을 보내지 않게 했어요. 해제는 ?nohit=0이에요. 서버 배포 없이 정적 JS만 바꾸면 돼요.

```
// 방문 기록을 보내기 직전에 확인
const skip = (() => {
  try {
    const q = new URLSearchParams(location.search).get('nohit');
    if (q === '1') localStorage.setItem('nohit', '1');
    if (q === '0') localStorage.removeItem('nohit');
    return localStorage.getItem('nohit') === '1';
  } catch { return false; }
})();
if (!skip) sendHit();
```

복사

한계: 이미 쌓인 기록은 소급해서 뺄 수 없고, 매번 새 프로필로 열리는 자동화 브라우저는 URL마다 붙여야 해요.

해결 2: 막지 말고 표시해 두기 (데이터센터·UA 계열)

데이터센터 접속을 서버에서 바로 막으면 진짜 방문까지 잃을 수 있어서, **막지 않고 표시만** 하기로 했어요. 방문 기록 테이블에 기본값이 빈 컬럼을 더하고(ALTER TABLE ... ADD COLUMN), 워커가 요청 정보에서 값을 채워요. Cloudflare Workers는 request.cf에 ASN과 접속 회사 이름을 넣어 줘요.

```
const cf = req.cf || {};
const org = String(cf.asOrganization || '').slice(0, 60);
const asn = parseInt(cf.asn, 10) || 0;
const dc = /google|amazon|aws|microsoft|azure|digitalocean|ovh|hetzner|oracle|alibaba|tencent|linode/;
const browser = /Edg/?.test(ua) ? 'edge' : /Firefox|FxiOS/?.test(ua) ? 'firefox' : /Chrome|CriOS/?.test(ua) ? 'chrome' : 'other';
const os = /Windows/?.test(ua) ? 'windows' : /iPhone|iPad/?.test(ua) ? 'ios' : /Android/?.test(ua) ? 'android' : 'other';
// INSERT 에 ua = browser + '|' + os, asn, org, dc 를 같이 저장
```

복사

해결 3: 읽음 신호 한 가지

1페이지만 열고 나가는 방문과 실제로 읽은 방문을 가르려고, 페이지가 숨겨질 때 한 번만 **최대 스크롤 깊이(%)** 와 **체류 초**를 같은 행에 갱신하게 했어요(방문 1건당 요청 1개). 서버는 처음 받은 방문 번호와 일치할 때만 값을 올려요(MAX로 올리기만 하고 줄이지 않아요).

확인한 방법

1. 배포 전에 워커 파일과 스키마 파일을 백업했어요(되돌리기 쉽게).
2. 시험 방문 한 건을 직접 보내 컬럼이 채워지는지, 다른 방문 번호로 보낸 갱신이 무시되는지 확인한 다음 시험 행을 지웠어요.
3. 배포 뒤 실제 방문이 새 컬럼과 함께 쌓이는지 최근 몇 행만 읽어 봤어요(읽기 한도 때문에 전체를 훑지 않았어요).

교훈

- 집계 숫자가 갑자기 좋아 보이면 먼저 우리 자신의 도구를 의심하세요. 가장 흔한 오염원은 내 점검이에요.
- 걸러낼 단서(접속 회사, UA 계열)는 **쌓일 때** 같이 남겨야 해요. 나중에는 소급이 안 돼요.
- 데이터센터 접속은 차단보다 표시가 안전해요. 표시된 값으로 사후 필터를 만들어 쓰면 규칙을 바꿔도 원본이 남아요.
- 규칙 기반 분류는 「사람이 아니다」의 증거가 아니라 「사람일 가능성이 큰 하한선」을 주는 도구예요. 숫자를 발표할 때는 이 한계를 같이 적으세요.

English summary

Our self-built hit counter showed 214 visitors on one day, but only about 76 looked human: the rest were our own link checks and data-center traffic that slipped past a bot-name filter. The fixes were a ?nohit=1 localStorage flag, new columns (browser/OS, ASN, network owner, a data-center flag that marks instead of blocks) filled from request.cf in a Cloudflare Worker, and one read signal (max scroll and seconds on page) updated once per visit. Lesson: store the signals you will need to filter later at write time.

관련 글

뉴스 이슈에 「내 생활 영향」과 실제 시세를 붙이는 법 — 규칙표·FRED·SVG로 지어내지 않기

영향 규칙표 10개로 이슈 60개 중 7개에 생활 영향 카드를 붙이고, FRED·업비트 공개 시세 6개(1.6초)로 전문가 발언 뒤 7일·30일 등락을 사후 확인하는 정적 사이트 파이프라인.

윈도우 11에서 wmic가 사라져 예약 작업이 조용히 멈춘 이야기 (파이썬 대체 코드)

윈도우 11에 wmic가 없어 파이썬 예약 작업이 FileNotFoundError 로 조용히 실패한 원인과, Get-CimInstance-psutil 대체 코드, 점검 실패 시 멈추게 하는 법.

ffmpeg 한 번에 세로 영상에 「3초 뒤 배너」와 「끝 카드」 붙이기 (파이썬)

ffmpeg filter_complex 한 번으로 세로 영상에 3초부터 나오는 배너와 3.5초 끝 카드를 붙이는 파이썬 코드, enable
심표·concat 규격·오디오 채널 실수 3가지.

← 기술 블로그 홈 · 사이트 제작 로그 · RSS

관련: 개발 지식 전체

이 글에 대해 묻고 싶으면 ? 개발 FAQ에 남겨 주세요.



공유

공유하기

카카오톡



링크 복사

네이버 블로그

밴드

X

Facebook

출처: 루리의 줍줍(sungprofile.cloud) · 최신 내용은 사이트에서 확인하세요.