사용량 예약과 원장: AI 비용을 정확히 정산하는 구조

사용량 원장의 정의와 필요성, 작동 원리, SAG 아키텍처 적용 기준과 실무 체크리스트를 연구·공식 문서 근거로 설명합니다.

어떤 상황에서 활용하나요?

수집·분석 요청의 중복, 재시도 또는 비용 정산을 설계하는 개발·운영 담당자에게 유용합니다. 작업의 성공·실패·권한·사용량을 함께 정의해 결과가 중복 저장되거나 비용이 이중 반영되지 않는지 검증하세요.

SAG(새그)는 더크림유니언이 운영하는 GEO·AEO·SEO 분석 및 최적화 플랫폼입니다. 검색엔진과 생성형 AI가 브랜드와 웹사이트 콘텐츠를 어떻게 발견하고 이해하는지 분석하고, 경쟁사 비교·개선 우선순위·실행 결과·재검증을 관리합니다. 플랫폼 구독과 전문 운영 지원을 제공하며, 콘텐츠 제작·사이트 코드 수정 등 대행 실행 범위는 별도 계약에서 정합니다.

한 문장 정의

사용량 원장은 작업 전 한도를 예약하고 완료·취소를 append-only 이벤트로 정산하는 방식입니다.

> 핵심 답: 동시 작업이 단순 현재값을 읽고 차감하면 한도를 초과할 수 있습니다. 실패와 재시도에서 이중 차감이 생기면 고객 신뢰를 잃습니다.


AI 답변에서 공식 사이트 출처를 확인하는 SAG 예시

> 이 화면은 방법 설명을 위한 SAG 공개 예시이며, 특정 고객사의 성과나 색인·검색·AI 추천을 보장하지 않습니다.

왜 이 기술이 필요한가?

동시 작업이 단순 현재값을 읽고 차감하면 한도를 초과할 수 있습니다. 실패와 재시도에서 이중 차감이 생기면 고객 신뢰를 잃습니다.

작동 원리

짧은 transaction에서 사용 가능량을 잠그고 예약합니다. 완료 시 exactly-once settlement를 기록하고 늦은 이벤트는 별도 reconciliation으로 반영합니다.

설계할 때는 정확도만 보지 않습니다. 지연 시간, 비용, 데이터 경계, 갱신 주기와 실패 시 동작을 함께 정의해야 운영에서 재현 가능한 결과가 됩니다. 자동화가 확신하지 못하는 값은 0이나 성공으로 바꾸지 않고 미측정·검토 필요 상태로 남기는 것이 안전합니다.

SAG 기술과의 연결

SAG는 tenant별 quota 예약과 usage ledger, idempotency key를 사용합니다. fixture token과 실제 공급자 비용은 분리하며 외부 호출이 없을 때 비용 0을 실제 비용 추정으로 해석하지 않습니다.

실무 체크리스트

- 동시 예약의 한도 초과를 테스트합니다
- 동일 작업 정산은 한 번만 허용합니다
- 고객 credit과 공급자 비용을 분리합니다
- 실패·빈 결과·권한 오류의 상태를 성공과 구분합니다
- 변경 전후를 동일한 조건으로 재검증합니다

연구와 공식 문서

- RAG 원 논문

참고 문서는 원리와 권장사항의 근거입니다. 검색 노출, AI 언급, 순위나 매출을 보장하지 않으며 실제 적용 효과는 서비스 데이터와 동일 조건의 관측으로 확인해야 합니다.

주제별 기술 참고

- PostgreSQL 트랜잭션 격리

이 기술을 이어서 읽는 방법

테넌트 권한, 작업 재시도, 캐시와 승인 이력을 살펴봅니다.

- 안정적인 고객 공간 설계
- 기능 안내 FAQ