

데이터네트워크(Data Network)

해당 번역본은 이용자가 디지털자산 관련 정보를 편리하게 확인할 수 있도록 단순 참고용으로 번역된 것이며, 투자 권유를 목적으로 하지 않습니다. 그 내용이 정확하지 않을 수 있으므로 정확한 정보 습득을 위해서는 원문을 참고하시거나 원문 작성 주체 측에 문의하시기 바랍니다. 두나무는 해당 번역본 내용의 진실성 및 정확성을 보장하지 않고, 제공된 정보에 의한 투자 결과에 대하여 어떠한 책임도 지지 않으며, 투자 손실은 투자자에게 귀속됩니다.

버전 0.2 - 2026년 6월

1 데이터 문제

우리는 AI 발전의 근본적인 전환을 겪고 있다. Transformer 기반 대규모 언어 모델은 추론과 코드 생성부터 복잡한 의사결정에 이르기까지 기계가 할 수 있는 일을 재정의했다. Diffusion 모델은 날마다 개선되고 있으며, 사실적이고 예술적인 이미지와 동영상을 생성할 수 있게 하고 있고, 로봇은 미지의 환경을 더 효과적으로 탐색하기 위해 더 많은 현실 세계 데이터를 소비하고 있다.

더 뛰어난 성능의 모델에 대한 수요는 둔화될 조짐을 보이지 않는다. 이러한 발전의 초석은 두 가지 핵심 요소, 즉 연산과 데이터이다. 연산 측면에서 발전은 꾸준했다. 더 빠른 GPU, 더 큰 메모리를 갖춘 TPU를 포함한 하드웨어 개선과, 기존 인프라에서 더 많은 성능을 끌어내는 양자화 및 타일링과 같은 소프트웨어 최적화는 커져가는 모델의 야심에 발맞추어 왔다. 반면 데이터는 점점 더 핵심적인 제약 조건이 되고 있다. 2022년 DeepMind의 중요한 논문인 "Training Compute-Optimal Large Language Models" [1]은, 흔히 Chinchilla 논문이라고 불리며, 이를 명시적으로 밝혔다. 더 큰 모델이 과소학습되는 것을 피하려면 필요한 토큰 수가 파라미터 수에 맞춰 확장되어야 한다는 것이다. 오늘날 모두가 추구하고 있는 더 많은 파라미터는 단순히 더 많은 연산만이 아니라 그에 비례해 더 많은 데이터를 요구한다.

대규모 언어 모델(LLM)의 영역에서, 모델 개발자들이 더 많은 파라미터와 더 긴 컨텍스트 윈도우 지원을 추진함에 따라, 사전학습에 많이 사용되던 공개 웹 데이터는 거의 고갈되고 있다. Villalobos 외 저자들은 그들의 논문 [2]에서 인간이 작성한 텍스트는 유한한 자원이며 LLM 확장의 병목이 될 수 있다고 주장한다. 그들은 인터넷뿐만 아니라 도서, 기사, 포럼 및 기타 출처에 존재하는 공개적으로 이용 가능한 인간 생성 텍스트의 총량을 추정한다. 해당 논문은 학습 데이터셋이 머지않아 이용 가능한 공개 인간 생성 텍스트의 한계에 도달할 수 있다고 예측한다. 현재 추세대로라면 언어 모델은 이르면 그보다 더 빨리, 늦어도 2026년에서 2032년 사이에 그 총량을 완전히 활용하게 될 것이다. 이는 적절히 학습될 수 있는 모델의 규모를 제한한다.

이에 대응하여 많은 연구소는 합성 데이터와 멀티모달 학습으로 눈을 돌렸다. 합성 데이터는 단기적으로 규모 확장을 가능하게 하지만, 구조적 약점을 지닌다. 자체 출력으로 학습된 모델은 오류가 누적되고, 시간이 지남에 따라 범위가 좁아지며, 현실 세계 데이터가 자연스럽게 포함하는 롱테일의 시끄럽고 진정으로 새로운 상황을 접하지 못할 위험이 있다. 합성 파이프라인은 알려진 분포 내에서 모델을 개선할 수 있지만, 모델이 실제로 본 것의 경계를

확장할 수는 없다. 최근 여러 논문 [3]은 합성 데이터가 희소한 인간 생성 데이터를 완전히 대체하는 것이 아니라 보완하는 것임을 보여주며, 합성 데이터에 대한 자기학습이 ‘모델 붕괴’를 일으킬 수 있다고 경고한다. 그러나 또 다른 방법이 있다. 웹 데이터의 압도적 다수, 추정치로는 80%가 여전히 비공개이며 접근할 수 없는 상태로 남아 있어, 인간 지식의 중요하고도 대부분 미개척된 저장소가 활용되지 못하고 있다. 우리는 이러한 가치 있는 자원을 개방하는 데 따르는 과제를 해결하는 것을 목표로 한다.

더 나아가, AI의 다음 물결은 LLM 기반 애플리케이션에만 관한 것이 아니다. 로봇틱스, 자율 시스템, 물리 AI, 멀티모달 에이전트는 아직 의미 있는 규모로 존재하지 않는 데이터를 필요로 한다. 즉, 실제 환경의 센서 스트림, 체화된 시스템의 상호작용 로그, 예측 불가능한 물리적 조건에서의 엣지 케이스가 필요하다. 이러한 모달리티는 텍스트로부터 합성될 수도 없고 웹에서 스크래핑될 수도 없다. 그것들은 포착되고, 구조화되며, 이용 가능하게 만들어져야 하며, 이는 데이터 문제가 단지 미해결 상태에 그치지 않음을 의미한다. 그 범위와 시급성이 커지고 있는 것이다.

이는 곧 대규모로 깨끗하고 효과적인 데이터를 제공할 수 있는 기업들이 역사상 가장 가치 있는 기업들 중 일부가 될 것임을 의미한다. 그리고 이는 전용 데이터 기업에만 그치지 않는다. 인간 행동 위에 자리 잡은 모든 플랫폼, 참여도 높은 사용자를 보유한 모든 앱, 독점 콘텐츠를 생산하는 모든 기업은 AI 시대에 자신들의 데이터를 수익화하고 있으며 앞으로도 그렇게 할 것이다.

2 과제

이전 절에서 우리는 데이터의 가치와 기존 비공개 데이터를 개방하거나, 새로운 데이터를 포착하기 위한 인센티브와 구조를 만들어야 할 필요성을 강조했다. 그러나 앞으로는 여러 과제가 존재한다.

2.1 데이터 제공자를 위한 과제

핵심적인 장벽은 데이터 소유권과 그에 수반되는 위험이다. 오늘날에는 원본 데이터 소유자, 데이터를 큐레이션하거나 집계하는 주체, 그리고 이를 최종적으로 소비하는 모델 개발자 사이에서 소유권이 어떻게 분배되는지를 쉽고 빠르게 정의할 방법이 없다. 실제로 권한, 귀속, 보상은 종종 임시방편적으로 협상되며, 이러한 불확실성은 선의의 데이터 공유조차 위험하게 느끼게 만든다. 많은 디지털 비즈니스는 비공개 독점 데이터와 어렵게 축적한 조직 지식에 의존하고 있으며, 이들은 접근 권한을 부여하는 것이 자신들의 차별화된 워크플로를 노출하거나, 민감한 사용자 행동을 드러내거나, 경쟁자가 역공학할 수 있는 상업적으로 가치 있는 인사이트를 유출할 수 있다고 우려한다. 또한 이들은 다운스트림 오용에 대해서도 우려한다. 즉, 동의 없이 데이터가 재판매되거나, 내부 정책을 위반하는 방식으로 다른 출처와 결합되거나, 해당 비즈니스의 핵심 제품을 직접 대체하는 모델을 학습하는 데 사용되는 경우가 이에 해당한다.

많은 모델 개발 기업들이 경쟁 우위를 유지하기 위해 법적·윤리적 문제를 주변부로 밀어낸 채 “먼저 스크래핑하고, 나중에 허락을 구하는” 접근 방식을 취하면서 이러한 우려는 더욱 심화되었다. 구매자가 신뢰할 수 있는 경우에도, 제공자는 여전히 통제권을 유지하면서 데이터를 공유할 수 있는 도구를 갖추지 못하고 있다. 명확한 라이선스 조건, 집행 가능한 사용 제한, 감사 가능한 출처 추적, 그리고 데이터가 광범위하게 접근 가능해지는 것을 방지하는 기밀성 보장이 필요하다. 우리는 해법이 소유권을 정의하고, 라이선스와 지급을 표준화하며, 기밀 데이터 이전을 가능하게 하는 개방적이고 공정한 경제 모델을 구축하는 데 있다고 믿는다. 이를 통해 제공자는 자신들의 핵심 경쟁력을 위협에 빠뜨리지 않고 AI 경제에 참여할 수 있다.

2.2 데이터 사용자를 위한 과제

데이터 품질이 높을수록 모델은 더 좋아진다. 소비되는 데이터의 양을 고려하면, 노이즈, 중복, 스팸을 포착하기는 어렵다. 품질을 넘어, 분포 역시 종종 과제가 된다. 분포는 편향될 수 있으며, 많은 언어는 충분한 데이터를 보유하고 있지 않다. 예를 들어 영어에 대해서는 풍부한 자원이 존재하는 반면, 많은 다른 언어들은 롱테일 데이터 부족에 직면해 있다.

보다 최근에는 법적 보호와 보안을 위해 엔드투엔드 데이터 추적이 필요하다는 점이 명확해졌다. 이는 데이터가 포착되는 지점부터 전체 파이프라인에 이르기까지 데이터를 추적하고, 학습에 사용된 모든 데이터 포인트에 대해 “데이터 영수증”을 제공할 수 있어야 함을 의미한다. 초기에 많은 모델 개발자들은 “먼저 스크래핑하고, 나중에 허락을 구하는” 접근 방식을 취했다. GDPR 및 CCPA와 같은 규제는 엄격한 동의 요건과 전 세계 매출의 최대 4%에 달하는 막대한 벌금을 부과하여, 무단 수집을 중대한 재무적 책임으로 만든다. 서비스 약관 위반 역시 기업을 고비용 소송에 노출시키며, 이는 저작권 및 계약상 청구를 제기한 출판사와 플랫폼의 유명 소송 사례에서 확인된다. 관련된 또 다른 과제는 데이터 제공자가 판매되는 데이터의 정당한 소유자임을 입증하는 것이다. 구매자는 해당 데이터가 정당한 소유자로부터 취득되었음을 보여주는 명확한 영수증을 필요로 한다.

공격 벡터는 검색 결과에서 사용자를 특정 제품으로 유도하는 AEO, 즉 에이전트를 위한 SEO와 같은 단순한 기술부터 더 복잡한 공격에 이르기까지 다양하다. 후자의 경우, 학습 데이터는 코딩 에이전트가 악성 패키지 또는 서드파티 패키지를 사용하도록 영향을 미칠 수 있으며, 이를 통해 공급망 침해가 가능해질 수 있다.

3 DATA 네트워크

우리는 DATA 네트워크를 AI의 가장 큰 병목, 즉 권리가 정리된 전문적 현실 세계 데이터에 대한 접근을 해결하기 위해 설계된 통합적이고 안전하며 개방적인 데이터 인프라 및 블록체인으로 제안한다.

Data 네트워크 내에서 데이터는 네이티브 객체이며, 모든 데이터 조각은 다음과 같을 수 있다.

- **증명 가능:** 데이터가 어디에서 왔고 어떤 조건에 따라 기여되었는지에 대한 검증 가능하고 이동 가능한 기록
- **기밀 유지 가능:** 사용자가 승인한 당사자만 복호화할 수 있도록 암호화됨
- **거버넌스 적용 가능:** 투명한 온체인 조건에 따라 소유되고, 라이선스가 부여되며, 수익화됨

이는 인터넷상의 누구나 참여할 수 있는 개방적이고 유연한 프레임워크를 만든다. 예를 들어, 누군가는 모든 데이터가 기밀로 저장되고, 진정성 증명 및 변조 불가능한 공개 감사 추적 기록이 함께 제공되는 데이터 마켓플레이스를 만들 수 있다. 데이터 소유자는 사용 조건을 정의하고, 데이터 소비자가 라이선스를 취득하며 스테이블코인으로 지급을 설정하도록 요구할 수 있다.

내부적으로 DATA Protocol은 탈중앙화된 노드 집합이 P2P 네트워크를 통해 참여하여 필요한 인프라를 구축할 수 있는 방식을 정의한다. 모두에게 개방된 이러한 노드들은 함께 작동하여 합의를 유지하고, 트랜잭션을 처리하며, 기밀 데이터 교환을 촉진하고, 프로토콜의 무결성을 보호한다. 관심사의 분리 원칙에 따라, 프로토콜은 세 개의 계층으로 나뉜다.

- **합의 계층:** 노드가 프로토콜을 따르고 프로토콜 상태에 대해 합의에 도달할 수 있도록 보장하는 계층

- **실행 환경:** 데이터가 등록되고, 소유권, 진정성, 증명, 라이선스 조건이 저장·집행·기록되는 가상 머신 역할을 하며, 이러한 내용은 변조 불가능한 감사 추적 기록에 남는다.
- **저장 계층:** 데이터가 저장되고 접근이 안전하고 기밀성 있게 제어되는 계층

각 계층을 더 자세히 살펴보자.

3.1 합의 계층

합의 계층은 모든 네트워크 참여자 사이에서 블록 형성을 조율함으로써 네트워크의 무결성, 보안 및 안정성을 보장한다. 최근 몇 년 동안 합의 메커니즘에서는 주목할 만한 발전과 연구상의 돌파구가 나타났다.

[그림 2: Tendermint의 블록 형성 주기, 정상 경로]

그림 2: Tendermint의 블록 형성 주기, 정상 경로

DATA 네트워크는 CometBFT [4]라고 하는 현대적이고 성능이 우수한 Tendermint 구현체를 사용한다. Tendermint [5]는 잘 정의된 속성과 맞춤화 옵션을 제공하는 비잔틴 장애 허용 합의 엔진으로, 이는 우리의 설계 요구에 부합한다.

- **장애 허용성:** 검증자의 33%가 악의적이거나 오프라인 상태인 경우에도 보안을 유지한다. 시스템은 모든 정직한 노드가 커밋된 블록에 대해 합의에 도달하도록 보장하며, 암호학적 증명을 사용해 악의적 행위자를 식별하고 제재한다.
- **활성성 보장:** 투표권 기준으로 검증자의 3분의 2를 초과하는 비율이 활성 상태를 유지하고 참여하는 경우, 네트워크는 계속 진행된다.
- **즉각적 최종성:** 트랜잭션은 블록에 포함된 후 빠르게 최종성을 확보한다. 이러한 즉각적 최종성은 불변성을 필요로 하는 애플리케이션의 신뢰성을 보장하며, 이는 사용자 경험의 핵심 요소이다.
- **모듈형 설계:** Application Blockchain Interface(ABCI)는 합의와 실행을 분리하여, 실행에 영향을 주지 않고 향후 합의 업그레이드를 가능하게 한다.

DATA Protocol의 합의 계층은 지분증명(PoS) 메커니즘을 사용하며, 검증자는 블록체인 보안과 무결성을 유지하는 데 핵심적인 역할을 한다. 이는 악의적 행위가 스테이킹된 토큰 손실을 포함한 제재로 이어지기 때문에, 정직하고 효율적인 운영을 위한 직접적인 경제적 인센티브를 만든다. 이 계층은 스테이킹, 락업 요건, 보상 분배와 관련된 모든 작업을 관리한다.

3.2 저장 계층

저장 계층은 데이터를 기밀로 저장하고, 접근 제어와 데이터를 실행 계층으로 안전하게 이전하기 위한 프로그래밍 가능한 인터페이스를 제공한다. 이를 달성하기 위해 노드들은 *Confidential Data Rails(CDR)* [6] 프로토콜에 참여하고 이를 실행한다.

CDR 프로토콜은 오프체인에서 작동하는 검증자 노드가 실행하는 신뢰 실행 환경(TEE)의 탈중앙화된 집합에 의해 구동되며, 실행 계층에 의해 관리되고 동기화된다. 이 인프라를 사용하여 검증자는 분산 키 생성에 참여하고 임계값 복호화를 수행할 수 있다.

이 핵심 기능을 통해 사용자는 기밀 데이터를 네트워크에 안전하게 업로드하고 접근 조건을 정의할 수 있다. 그러면 해당 데이터는 이러한 조건을 충족하는 사용자와 시스템이 자동으로 접근할 수 있게 된다. 조건은 실행 계층에 설정된 조건에 따라 동적으로 충족될 수 있다.

이 저장 계층은 모든 크기와 모든 형식의 데이터를 대규모로 지원할 수 있으며, 데이터의 기밀성을 유지하면서도 IPFS [7] 및 Walrus [8]와 같은 데이터 가용성 프로토콜과 높은 호환성을 가진다.

3.3 실행 계층

이전의 모든 계층 위에는 *실행 계층*이 있다. 실행 계층은 모든 데이터 포인트가 등록되고, 소유권, 진정성, 증명, 라이선스 조건이 기록·저장·집행되는 가상 머신을 실행한다. 가상 머신과의 모든 상호작용은 변조 불가능하고 감사 가능하며, 이는 사용자가 제출한 트랜잭션, 예를 들어 데이터 거래를 통해 이루어진다.

이 계층의 핵심에는 Ethereum Virtual Machine(EVM)이 사용되며, 확장 가능한 데이터 등록, 운영, 교환 및 기타 필요한 기능을 지원하기 위한 성능 최적화가 적용된다.

중심에는 진정성을 확인하고 모든 데이터 포인트를 대규모로 등록하는 역할을 하는 스마트 컨트랙트가 있다. 이와 함께 또 다른 컨트랙트는 진정성 증명 또는 품질 증명과 같은 증명이 누구나 어떤 데이터 포인트에 대해서도 제출될 수 있도록 한다. 이는 다음과 같은 데이터 포인트 원장을 구성한다.

- **불변성:** 기록은 일단 작성되면 영구적이며, 잘 보안된 체인에서는 삭제나 수정이 계산상 실행 불가능하다.
- **타임스탬프:** 각 상호작용과 업데이트에는 타임스탬프가 부여되고 순서가 정해져, 신뢰할 수 있는 사건의 순서를 제공한다.
- **검증성:** 단일 당사자가 합의 없이 기록을 변경할 수 없는 탈중앙화된 노드 네트워크를 통해 검증된다.
- **암호학적 출처:** 각 항목은 작성자, 즉 지갑, 키 또는 노드에 의해 서명될 수 있으며, 누가 무엇을 언제 작성했는지를 증명한다.

데이터 자체는 CDR을 사용하여 저장 계층에 저장되며, 그에 대한 암호학적으로 검증 가능한 참조만 온체인에 등록된다는 점에 유의해야 한다. 데이터 소유자는 각 데이터 포인트에 대한 이용약관을 설정할 수 있으며, 모든 이용약관은 스마트 컨트랙트로 정의되고, 이 컨트랙트는 데이터 접근 제어를 자동으로 수행할 수 있는 권한을 가진다. 조건과 라이선스 계약을 맞춤화하고 암호화폐 및 스테이블코인 결제를 지원하기 위해 다수의 기존 템플릿과 스마트 컨트랙트를 사용할 수 있다.

Trace

실행 계층에서 실행되는 핵심 애플리케이션은 Trace이다. 데이터가 적절하게 라이선스가 부여되고 안전하게 이전되는 경우에도, 여전히 중요한 질문 하나가 답변되지 않은 채 남아 있다. 특정 기록이 적법하게 수집되었는지, 기여 시점에 동의가 획득되었는지, 그리고 수집과 사용 사이에 아무것도 변경되지 않았는지를 누구나 어떻게 검증할 수 있는가? 신뢰할 수 있는 답이 없다면, 데이터 구매자는 법적 노출을 부담하고, 기여자는 집행 메커니즘을 갖지 못하며, 규제기관은 대규모로 감사할 수 없다. Trace는 모든 인간 기여 기록에 데이터의 전체 생애주기 동안 함께 이동하는 검증 가능한 엔드투엔드 영수증을 제공함으로써 이를 해결한다.

기록이 Trace를 통해 DATA Network에 들어오면, 변조 불가능하고 암호학적으로 서명된 영수증이 기여 시점에 즉시 생성된다. 이 영수증은 기록의 전체 출처, 즉 그 기원, 기여 플랫폼, 해당 기록이 수집된 동의 조건, 타임스탬프

프, 그리고 이후의 모든 수정을 탐지하는 콘텐츠 해시를 포착한다. 영수증은 실행 계층의 불변 원장에 기록되며 영구적으로 접근 가능한 상태로 남는다. 기록이 큐레이션, 정제, 번들링, 그리고 최종 판매를 거쳐 이동함에 따라, 영수증도 함께 이동하므로, 다운스트림의 어느 지점에서든 어떤 당사자라도 해당 기록을 그 출처까지 추적하고 무결성을 검증할 수 있다.

중요한 점은 이러한 감사 가능성이 기여자의 프라이버시를 희생하지 않는다는 것이다. 온체인에 기록되는 것은 기여와 그 조건에 대한 암호학적 커밋먼트이지, 해당 기여를 한 사람의 개인정보가 아니다. 기여자 신원은 프로토콜 수준에서 기여 메타데이터와 분리되며, 이를 통해 수백만 건의 동의 기록을 감사하는 규제기관이나 전체 학습 데이터셋을 검증하는 연구소가 개별 기여자의 개인정보에 접근하지 못하도록 보장한다.

그 결과는 데이터 경제의 모든 참여자에게 기여하는 공개 감사 추적 기록이다. AI 연구소는 수 주간의 법률 검토가 아니라 몇 초 만에 전체 데이터셋의 출처를 검증할 수 있다. 기여자는 언제든지 자신의 조건이 준수되고 있는지 확인할 수 있다. 그리고 규제기관은 제공자에게 부담을 주지 않고 대규모로 동의를 감사할 수 있다. 모든 기록을 공개적이고 불변적인 원장상의 검증 가능한 영수증에 고정함으로써, Trace는 데이터 출처를 사후적인 고려사항에서 일급 보장으로 전환한다.

4 결론

데이터는 이제 AI 발전의 핵심 제약 조건이다. 공개 웹 데이터는 거의 고갈되었고, 합성 대안은 현실 세계의 신호를 완전히 대체할 수 없으며, 비공개 데이터의 압도적 다수는 신뢰와 도구의 격차로 인해 접근 불가능한 상태로 남아 있다. DATA Protocol은 BFT 합의, CDR을 통한 TEE 기반 기밀 저장, 그리고 변조 불가능한 EVM 실행 계층을 결합함으로써 이러한 격차를 해소한다. 이를 통해 제공자는 통제권을 잃지 않고 데이터를 공유할 수 있는 방법을 얻고, 소비자는 권리가 정리된 검증 가능한 데이터에 대규모로 접근할 수 있다. LLM, 로봇틱스, 자율 시스템 전반에서 전문적 현실 세계 데이터에 대한 수요가 증가함에 따라, 이러한 종류의 개방형 데이터 인프라는 AI의 다음 시대를 위한 기반이 된다.