

2025년 글로벌 합성데이터 가격 책정 전략 분석

부제: 모달리티, 플랫폼, 그리고 가치 기반 서비스의 경제학

- 작성일: 2025년 11월 7일
- 기획: (주)페블러스 데이터 커뮤니케이션팀
- AI 도구: Gemini
- 인터랙티브 콘텐츠: <https://blog.pebblous.ai/>

Executive Summary: "데이터 포인트"의 환상과 "가치 서비스"의 현실

글로벌 합성데이터 시장의 가격 책정은 '데이터 포인트당 비용'(\$ per data point\$)이라는 초기의 단순한 지표에서 벗어나, 고도로 정교화된 '**3중 요금제(Three-Part Tariff)**' 구조로 빠르게 수렴하고 있습니다. 본 보고서는 이 3중 요금제 모델이 현존하는 엔터프라이즈급 합성데이터 벤더의 수익 모델을 설명하는 보편적 프레임워크임을 입증합니다. 이 모델은 다음과 같은 세 가지 핵심 요소로 구성됩니다:

- 플랫폼 최소 약정(The Platform Floor):** 플랫폼 접근, 보안, 기본 지원을 위한 협상 불가능한 월간(MRC) 또는 연간(ARC) 고정 비용입니다. 이는 사용량과 무관하게 발생하는 '기본료'입니다.
- 가변 미터(The Variable Meter):** 실제 사용량을 측정하는 종량제 비용입니다. 이 '미터'의 측정 기준(예: 컴퓨팅 시간, 데이터 포인트 수, 소스 데이터 볼륨)은 벤더의 전략과 데이터 모달리티에 따라 급격하게 달라집니다.
- 가치 부가 서비스(The Value-Add):** 플랫폼의 기본 기능을 초월하여 특정 비즈니스 문제 해결, 도메인 제약(Physics) 적용, 품질 보증(QA), 프라이버시 보증(DP) 등을 제공하는 전문 서비스 또는 관리형 서비스(Managed Service) 비용입니다.

본 보고서의 핵심 분석 결과는, **데이터 모달리티(정형, 텍스트, 비전)** 가 이 3중 요금제의 구성 비율을 근본적으로 결정한다는 것입니다.

- 정형 및 시계열 데이터(Tabular & Time-Series):** (예: BMS 시계열) '가변 미터' 비용은 전체 비용에서 미미한 수준(\$100 ~ \$200\$)을 차지합니다. 대신, 비즈니스 규칙과 물리적 제

약(예: 배터리 전압/온도 규칙)을 적용하는 '전문 서비스'(\$10,000 ~ \$40,000\$)가 총비용의 99%를 차지하며 가격을 주도합니다.

- **텍스트 및 언어 데이터(Text/NLP):** 가격 모델이 급격히 재편되고 있습니다. 특히 LLM 훈련 데이터 생성 시, '가변 미터'가 '교사 LLM(Teacher LLM)'의 **API 추론 비용(토큰 비용)**에 직접 연동되는 추세입니다.¹
- **이미지 및 컴퓨터 비전(Image/CV):** '플랫폼 최소 약정' 자체가 월 \$3,000\$에서 \$15,000\$ 수준으로 매우 높습니다.² 이는 3D 시뮬레이션 인프라, 3D 에셋 라이브러리, 전담 기술 지원 (TAM) 비용이 기본료에 선반영(pre-baked)된 **인프라 구독 형태**를 띠기 때문입니다.

결론적으로, 고객은 '데이터' 자체를 구매하는 것이 아니라, 특정 문제를 해결하는 '솔루션' 또는 '인프라 접근권'을 구매합니다. 따라서 '데이터 포인트당 가격'은 대부분의 엔터프라이즈 시나리오에서 환상에 불과하며, 실제 비용은 플랫폼 진입 비용과 특정 도메인 문제를 해결하는 전문 서비스의 가치에 의해 결정됩니다.

목차

- I. 합성데이터 가격 책정의 보편적 프레임워크: 3종 요금제 모델 (**The Universal Pricing Framework**)
 - A. **The Platform Floor** (플랫폼 최소 약정): 진입을 위한 고정 비용
 - B. **The Variable Meter** (가변 미터): 사용량 측정 지표의 분화
 - C. **The Value-Add** (가치 부가 서비스): "옵션"이 아닌 "필수" 비용
- II. 모달리티 분석 I: 정형 및 시계열 데이터 (**Tabular & Time-Series**)
 - 주요 벤더 분석
 - 핵심 테이블 1: 정형/시계열 벤더 가격 모델 비교
- III. 모달리티 분석 II: 텍스트 및 언어 데이터 (**Text/NLP/LLM**)
 - 주요 사용 사례 및 가격 모델
 - 핵심 테이블 2: 텍스트 모달리티 가격 모델 비교
- IV. 모달리티 분석 III: 이미지 및 컴퓨터 비전 데이터 (**Image & CV**)
 - 주요 벤더 및 가격 모델 ("높은 최소 약정"이 표준)
- V. 공급 및 배포 모델 비교 분석 (**API, SaaS, On-Premise**)
 - A. **API 기반 접근 (Public SaaS)**
 - B. **플랫폼 구독 (VPC / Private Cloud Marketplace)**
 - C. **온프레미스(On-Premise) / 연간 라이선스**
 - D. **프로젝트 기반 (Managed Service / DaaS)**
 - 핵심 테이블 3: 공급 모델별 TCO 및 보안 영향 분석

- VI. 전략적 결론 및 권고안
 - A. 내부 분석의 검증
 - B. 핵심 결론: 모달리티가 가격 구조를 결정한다
 - C. 전략적 권고
- VII. 참고문헌 (Sources Used)
 - Works cited

I. 합성데이터 가격 책정의 보편적 프레임워크: 3중 요금제 모델 (The Universal Pricing Framework)

합성데이터의 가격을 분석할 때 가장 흔히 접하는 오류는, YData의 SDK 정책(\$1\$ / 100만 데이터 포인트) 4과 같이 단순한 '데이터 포인트당' 비용을 시장의 표준으로 오해하는 것입니다. 특정 BMS 시계열 증강 프로젝트(1.7억 포인트 생성)에 이 모델을 적용하면, 순수 생성 비용은 약 \$172.80\$에 불과합니다. 그러나 실제 글로벌 벤더에 PoC(Proof of Concept)를 의뢰할 때의 견적은 \$10,000\$에서 \$40,000\$에 이릅니다.

이 막대한 차이는 합성데이터가 단순한 '데이터(Data)'가 아니라, 복잡한 '서비스(Service)'이자 '플랫폼(Platform)'으로 판매되기 때문에 발생합니다. 시장의 가격 책정은 사실상 다음 세 가지 구성요소의 조합으로 이루어집니다: (1) 플랫폼 최소 약정, (2) 가변 미터, (3) 가치 부가 서비스.

A. The Platform Floor (플랫폼 최소 약정): 진입을 위한 고정 비용

'플랫폼 최소 약정'은 벤더의 소프트웨어 라이선스, 기본 지원, 보안 및 컴플라이언스(예: SOC 2, HIPAA) 5 유지를 위해 고객이 지불해야 하는 최소 고정 비용(MRC 또는 ARC)입니다. 이는 사용량(\$0\$)과 관계없이 발생하는 '기본료'이며, 벤더의 안정적인 수익 기반이 됩니다.

이 최소 약정 비용은 벤더의 시장 포지셔닝과 특히 데이터 모달리티에 따라 극적인 차이를 보입니다.

- 정형 데이터 (Tabular) - 저~중 등급: 정형 데이터 틀 중에서는 상대적으로 낮은 진입 비용을 제공하는 벤더들이 존재합니다.
 - **Tonic.ai (Structural)**: 월 \$199\$의 Pay-as-you-go 플랜으로 시작합니다.5
 - **Gretel.ai**: Team 플랜의 기본료는 월 \$295\$입니다.7
 - **Hazy**: Starter 플랜이 월 \$500\$에서 시작합니다.9
- 정형 데이터 (Tabular) - 엔터프라이즈 등급:

- **MOSTLY AI**: AWS Marketplace를 통한 배포 시, 최소 월 **\$3,000**의 고정 비용을 요구합니다.¹⁰
- **이미지/컴퓨터 비전 (Image/CV) - 고 등급:**
 - **Synthesis AI**: 플랫폼 접근을 위한 연간 구독은 월 \$3,000\$에서 시작합니다.¹³
 - **Rendered.ai**: Teams 플랜은 월 **\$5,000**, Organizations 플랜은 월 **\$15,000**의 최소 약정이 필요합니다.¹²

정형 데이터의 \$199\$(Tonic)에서 컴퓨터 비전의 \$15,000\$(Rendered.ai)까지, 모달리티에 따라 '최소 약정' 비용이 약 75배의 차이를 보입니다. 이 차이는 우연이 아니며, 각 모달리티를 생성하는 데 필요한 **초기 자본 투자(CapEx)**와 **인프라 유지 비용**을 직접적으로 반영합니다.

정형 데이터 생성은 본질적으로 알고리즘과 *CPU* 집약적입니다. 반면, 고품질 CV 데이터 생성은 막대한 *3D 애셋 라이브러리*(차량, 사람, 환경 모델), 고성능 *시뮬레이션 엔진*(예: Rendered.ai의 NVIDIA Omniverse 통합 12), 그리고 대규모 *렌더링 팝*(GPU 클러스터)을 필요로 합니다.

따라서 Rendered.ai의 \$15,000\$ 월간 요금은 단순한 소프트웨어 사용료가 아니라, 이 고가의 '디지털 헐리우드 세트장'에 대한 인프라 감가상각 및 유지보수 비용을 포함하는 **'인프라 구독료'**의 성격을 갖습니다. MOSTLY AI가 정형 데이터 벤더임에도 \$3,000\$라는 높은 '최소 약정'을 설정한 것은, 자사 제품을 단순한 툴이 아닌 CV 벤더와 동등한 수준의 고가치 엔터프라이즈 플랫폼으로 포지셔닝하려는 전략적 의도를 명확히 보여줍니다.

B. The Variable Meter (가변 미터): 사용량 측정 지표의 분화

'가변 미터'는 고객이 사용한 만큼 지불하는 종량제 비용입니다. 벤더가 "무엇을" 측정하는지는 벤더의 비즈니스 모델과 비용 구조를 드러내는 가장 중요한 지표입니다. 시장에는 최소 5가지의 서로 다른 '미터'가 혼재합니다.

1. 컴퓨팅 기반 (Compute-Based):

- **사례**: MOSTLY AI. 크레딧은 생성된 데이터 양이 아닌, "총 가상 CPU 및 GPU 시간"을 기준으로 소비됩니다.¹³ 공식은 $\text{Credits} = A \times \text{Total Virtual CPU Time} + B \times \text{Total Virtual GPU Time}$ ¹³
- **분석**: 가장 순수한 PaaS(Platform-as-a-Service) 모델입니다. 고객은 생성된 데이터 양이 아닌, 생성에 소요된 자원에 대해 비용을 지불합니다.

2. 데이터 볼륨 기반 (Data Volume-Based - Points/Rows):

- **사례**: YData (SDK) 4, Gretel.ai 15, Datagen.in.¹⁶
- **분석**: YData SDK는 1 크레딧 = 100만 데이터 포인트(\$1\$)로 명확합니다.¹⁴ Gretel은 Developer 플랜에서 월 273k+ 레코드를 제공하며 15, Datagen.in은 30,000 크레딧으로 30,000개의 텍스트 행을 생성할 수 있습니다.¹⁶ 이는 고객에게 가장 직관적이지만, 벤더에게는 위험합니다 (예: 100만 개의 간단한 행 생성과 100만 개의 복잡한 시

계열 행 생성에 동일 비용 청구).

3. 소스 볼륨 기반 (Source Volume-Based):

- **사례:** Tonic.ai (Structural).
- **분석:** Tonic의 가격은 "소스 데이터 볼륨"(예: 2TB, 10TB)을 기준으로 책정됩니다.⁵ 이는 *데이터베이스 마스킹/부분추출(masking/subsetting)*이라는 특정 사용 사례에 고도로 최적화된 독창적인 모델입니다. 가치는 생성된 출력물(무한대로 생성 가능)이 아닌, 보호하는 원본 자산(프로덕션 DB)의 크기에 비례합니다.

4. 토큰/단어 기반 (Token/Word-Based):

- **사례:** Tonic.ai (Textual) 5, YData (SDK) 4, Gretel.ai.¹⁷
- **분석:** 텍스트 모달리티의 표준입니다. Tonic Textual은 "처리된 단어 수" 5, YData SDK는 1 크레딧 = 10,000 토큰 4, Gretel은 "생성된 토큰"을 크레딧으로 변환합니다.¹⁷ 이는 LLM의 API 가격 모델을 그대로 차용한 것입니다.

5. 이미지 수 기반 (Image Count-Based):

- **사례:** Datagen.in.¹⁶
- **분석:** Datagen.in은 30,000 크레딧으로 30,000 텍스트 행 또는 3,000 이미지를 생성할 수 있다고 명시합니다.¹⁶ 이는 벤더가 공식적으로 **"1 이미지 = 10 텍스트 행"**의 교환 가치를 설정한 매우 명확한 사례입니다.

이러한 '미터'의 선택은 벤더의 전략적 포지셔닝을 반영합니다. MOSTLY AI(컴퓨팅 기반)는 "우리는 복잡한 모델링을 위한 *PaaS*"라고 말하는 것입니다. 고객의 작업 복잡도(예: BMS 물리 제약)가 높을수록 더 많은 크레딧(비용)을 소비합니다. 이는 벤더의 수익성을 보장하지만 고객의 비용 예측 성은 낮춥니다.

반면, YData(데이터 포인트 기반)는 "우리는 대량 생성을 위한 *ユーティリティ*"라고 말합니다. 비용 예측이 쉽습니다. 벤더는 알고리즘을 최적화하여 컴퓨팅 비용을 줄여야 이익이 남는 구조입니다. Tonic(소스 볼륨 기반)은 "우리는 *데이터베이스 보안 툴*"이라고 말하며, 생성량과 비용을 분리하고 고객의 핵심 자산(프로덕션 DB) 크기에 가격을 연동시켰습니다.

C. The Value-Add (가치 부가 서비스): "옵션"이 아닌 "필수" 비용

'가치 부가 서비스'는 플랫폼의 기본 기능을 넘어선 특정 도메인 문제 해결, 품질 보증, 시나리오 설계, 프라이버시 보증 등을 위한 전문 컨설팅 및 관리형 서비스입니다. 엔터프라이즈 시장에서 이는 '옵션'이 아닌 사실상의 **'필수 핵심 비용'**입니다.

- **사례 (BMS PoC):** 특정 BMS PoC 제안서의 "Pro" 패키지(\$18,000\$)는 "물리·안전 제약 룰세트 적용"과 "희귀 이벤트(과열/스파이크) 비율 컨트롤"을 핵심 가치로 제공합니다. 이는 순수 데이터 생성(\$172\$)이 아닌, 도메인 전문성이 결합된 '전문 서비스'입니다.
- **사례 (Image/CV):**

- **Synthesis AI:** "커스텀 데이터셋" 제작 프로젝트에 대해 \$10,000\$의 최소 구매 금액을 명시합니다.³
- **Rendered.ai:** 월 \$15,000\$의 "Organizations" 플랜에는 "기술 지원 매니저 (Technical Account Manager, TAM)" 가 포함되며, 이 TAM은 정기 미팅과 고객 백로그 관리를 전담합니다.²
- **사례 (플랫폼 전반):** Salesforce 18, Syntho 19, Statice 20 등 다수의 벤더가 "전문 서비스" 또는 "고객 성공 플랜(Customer Success Plans)"을 핵심 상품으로 제공하며, 이것이 프로젝트 성공의 핵심임을 강조합니다.

결론적으로, 엔터프라이즈 고객은 "데이터 1TB"를 구매하는 것이 아니라, "특정 비즈니스 문제(예: BMS 배터리 열화 예측 모델의 정확도 개선)"를 해결하는 솔루션을 구매합니다. 합성데이터 플랫폼은 이 솔루션(컨설팅 서비스)을 제공하기 위한 도구로 기능하며, 따라서 '전문 서비스' 비용이 전체 견적의 90% 이상을 차지하는 것은 매우 합리적인 시장 현상입니다.

II. 모달리티 분석 I: 정형 및 시계열 데이터 (Tabular & Time-Series)

정형 및 시계열 데이터는 금융, 헬스케어, 제조(예: BMS) 등 핵심 산업에서 가장 널리 사용되는 모달리티입니다. 이 시장의 가격 모델은 '최소 약정'과 '가변 미터'의 다양한 조합을 보여주며, 특히 '전문 서비스'의 가치가 극대화되는 영역입니다.

주요 벤더 분석

- **MOSTLY AI:**
 - **모델:** 높은 '최소 약정' (\$3,000\$/월, AWS Marketplace 기준) 10 + '컴퓨팅 시간' 기반 '가변 미터'.¹³
 - **분석:** MOSTLY AI의 가격 모델은 자사를 고가치 엔터프라이즈 PaaS로 명확히 포지셔닝합니다. 크레딧이 vCPU/vGPU 시간에 연동된다는 점 13은, BMS 프로젝트처럼 복잡한 상관관계와 물리적 제약(예: 충방전 사이클 보존, 셀 밸런싱 규칙)을 모델링하는 작업이 단순 데이터 생성보다 훨씬 더 많은 크레딧(비용)을 소모함을 의미합니다.
- **YData:**
 - **모델:** YData는 두 가지 상이한 모델을 제공합니다. (1) **SDK:** '데이터 포인트' 기반의 순수 PAYG (\$1\$/100만 포인트).⁴ (2) **Fabric Platform:** AWS/Azure Marketplace를 통한 VPC 배포.⁴
 - **분석:** SDK 모델은 BMS 프로젝트의 '순수 생성 비용'(\$172.80\$)을 계산하는 데 유용하지만, 실제 엔터프라이즈 배포는 Fabric 플랫폼을 통해 이루어집니다. 이 경우, 고객

은 YData Fabric의 라이선스 비용(비공개) 21 외에도, AWS Marketplace에 명시된 별도의 인프라 비용(예: CPU 시간당 \$0.04\$, GPU 시간당 \$0.20\$)을 AWS에 직접 지불해야 합니다. 21 따라서 총 소유 비용(TCO)은 (YData 라이선스비) + (AWS 인프라비) + (Pebblous와 같은 파트너의 전문 서비스비)로 구성됩니다.

- **Gretel.ai:**

- **모델:** 낮은 '최소 약정' (\$295\$/월) + 하이브리드 '가변 미터' (\$2.20\$/크레딧). 7
- **분석:** Gretel의 크레딧은 '작업 런타임(job runtime)' 또는 '생성된 토큰'을 기준으로 변환됩니다. 17 이는 MOSTLY AI와 유사하게, 작업의 복잡도가 높을수록 더 많은 크레딧을 소모하는 컴퓨팅 기반 모델의 변형입니다.

- **Tonic.ai (Structural):**

- **모델:** 낮은 '최소 약정' (\$199\$/월부터) + '소스 데이터 볼륨' 기반 '가변 미터'. 5
- **분석:** 이 모델은 증강(*Augmentation*) 유스케이스와는 적합하지 않습니다. Tonic은 100GB의 원본 데이터를 받아 100GB(또는 10GB)의 익명화된 부분집합을 만드는 데 특화되어 있습니다. 5일치 데이터를 20일치로 증강하는 작업(출력물이 원본보다 4배 커짐)은 이들의 가격 책정 로직(원본 크기 기준)과 맞지 않습니다.

핵심 테이블 1: 정형/시계열 벤더 가격 모델 비교

벤더	핵심 상품	최소 약정 (Platform Floor)	가변 미터 (Variable Meter)	BMS 프로젝트 비용 반영 방식 (분석)
MOSTLY AI	플랫폼 (VPC)	\$3,000 / 월 11	vCPU/vGPU 시간 (크레딧) 13	물리 제약 모델이 복잡할수록 '가변 미터' 비용(크레딧)이 직접 증가합니다.
YData (SDK)	SDK (API)	\$0 (PAYG)	\$1 / 100만 데이터 포인트 4	'가변 미터' 비용은 \$172.80\$ 로 고정됩니다. '전문 서비스'(\$18k\$) 비용이 별도로 부과됩니다.
YData (Fabric)	플랫폼 (VPC)	비공개 (Enterprise)	AWS 인프라 비용 (CPU/GPU) 21	플랫폼 라이선스 + AWS 비용 + 전문 서비스 비용. TCO가 가장 복잡합니다.
Gretel.ai	플랫폼 (SaaS)	\$295 / 월 8	\$2.20 / 크레딧 (런타임/토큰) 17	MOSTLY AI와 유사하게, 복잡한 작업(런타임)이 더 많은 '가변 미터' 비용을 소

				모합니다.
Tonic (Structural)	플랫폼 (SaaS)	\$199 / 월 6	소스 DB 크기 (예: 2TB) 5	5일치 원본 데이터 크기에 비용이 부과됩니다. 4배 증강(출력률)은 비용과 무관합니다.

정형 데이터 시장은 이처럼 '컴퓨팅 기반' 미터(MOSTLY, Gretel)와 '볼륨 기반' 미터(YData)로 명확히 나뉩니다. 이 선택은 파트너사(예: Pebblous)의 수익 모델에 중대한 영향을 미칩니다.

BMS 프로젝트(복잡한 물리 제약)를 YData(볼륨 기반) 플랫폼에서 수행하면, 플랫폼 비용 (\$172.80\$)은 미미하게 고정됩니다. 이는 고객에게 "지불하는 \$18,000\$는 순수하게 Pebblous의 BMS 도메인 전문성(물리 제약 적용)에 대한 대가"임을 명확히 보여줄 수 있어 가치 전달에 가장 유리합니다.

반면, MOSTLY AI(컴퓨팅 기반) 플랫폼을 사용하면, 복잡한 BMS 모델이 더 많은 vCPU/vGPU 시간을 소모하므로 '가변 미터' 비용(크레딧)이 \$172\$보다 훨씬 높게(예: \$1,000 \sim \$2,000\$) 나올 수 있습니다. 이는 고객에게 청구되는 총 비용 중 벤더 플랫폼 비중이 높아져, 파트너사의 '전문 서비스' 가치가 상대적으로 희석될 수 있습니다.

III. 모달리티 분석 II: 텍스트 및 언어 데이터 (Text/NLP/LLM)

텍스트 모달리티는 LLM(대형 언어 모델)의 등장으로 인해 가격 책정 모델이 완전히 재정의되고 있습니다. 과거에는 텍스트 생성을 위해 별도의 생성 모델을 훈련(Training)해야 했지만, 이제는 강력한 SOTA(State-of-the-art) LLM에 추론(*Inference*) 요청을 보내는 것만으로 고품질의 합성 데이터를 얻을 수 있게 되었습니다.²³

이러한 패러다임 전환은 텍스트 합성데이터의 가격 모델을 "합성데이터 생성 비용 = LLM 추론 비용"으로 수렴시키고 있습니다.

주요 사용 사례 및 가격 모델

1. 익명화 및 마스킹 (Redaction & Masking):

- 벤더: Tonic.ai (Textual).²⁴
- 가격 모델: '소스 볼륨' 기반. 생성할 데이터가 아닌, "처리된 단어 수(number of words processed)"에 따라 과금됩니다.⁵
- 분석: 이는 Tonic Structural(정형) 모델의 텍스트 버전입니다. 가치는 생성이 아닌 보

호에 있습니다. (예: 콜센터 로그, 의료 기록에서 PII 제거).

2. LLM 훈련 데이터 생성 (Fine-Tuning Data Generation):

- 벤더: **Gretel.ai, AWS Bedrock.**
- 가격 모델 (**Gretel**): '토큰' 기반 크레딧. Gretel은 Text-to-SQL 데이터셋 생성 25이나 차등 프라이버시(DP)를 적용한 텍스트 생성 26 등 특화된 LLM 데이터 생성에 집중합니다. 이 경우 크레딧은 '생성된 토큰' 또는 '작업 런타임'에 연동됩니다.¹⁷
- 가격 모델 (**AWS Bedrock**): "교사 모델(Teacher Model) 비용 연동."¹⁸

AWS Bedrock의 가격 정책은 텍스트 합성데이터 시장의 중대한 패러다임 전환을 명확히 보여줍니다. Bedrock은 합성데이터 생성 비용을 "**선택한 교사 모델의 온디맨드 가격**"으로 정의합니다.¹⁹

이는 "합성데이터 벤더"의 역할이 "고유한 생성 모델 제공자"에서, SOTA LLM(예: Claude 3, GPT-4o)을 활용하여 데이터를 생성하는 "**프롬프트 오캐스트레이션 및 프라이버시 레이어 제공자**"로 전환되고 있음을 시사합니다.

이제 비용은 생성 모델의 훈련 비용(\$)이 아닌, SOTA LLM의 *API 추론 비용*(\$)에 수렴합니다. Scale AI 등의 연구²⁰에서 보듯이, 이제 핵심 질문은 "어떤 모델로 생성할까?"가 아니라 "시드 데이터셋 크기 대비 교사 모델 쿼리 예산을 얼마나 쓸까?"(즉, API 비용을 얼마나 지불할 것인가)가 되었습니다.

핵심 테이블 2: 텍스트 모달리티 가격 모델 비교

사용 사례	주요 벤더	가격 책정 단위 (Meter)	비용 결정 요인
익명화 / 마스킹	Tonic Textual	처리된 단어(Word) 수 5	보호해야 할 원본 문서의 총 량
LLM 훈련 데이터 (특화 모델)	Gretel.ai	생성된 토큰(Token) 수 또는 작업 런타임 17	생성할 데이터의 양 + 프라이 버시(DP) 적용 여부 ²¹
LLM 훈련 데이터 (SOTA 활용)	AWS Bedrock	교사 모델의 입/출력 토큰 수 1	선택한 교사 모델(예: Claude 3)의 API 가격

IV. 모달리티 분석 III: 이미지 및 컴퓨터 비전 데이터 (Image & CV)

컴퓨터 비전(CV) 모달리티는 정형 또는 텍스트 데이터와는 근본적으로 다른 경제 구조를 가집니다.

다. 이는 "헐리우드 모델"로 비유할 수 있습니다. 데이터 생성의 비용은 알고리즘이 아닌, **3D 자산, 시뮬레이션 엔진, 렌더링 파워**라는 고가의 인프라에 의해 결정됩니다.²⁹

이러한 특성으로 인해, CV 합성데이터 시장의 가격 모델은 **"매우 높은 최소 약정(Platform Floor)"**이 표준으로 자리 잡았습니다.

주요 벤더 및 가격 모델 ("높은 최소 약정"이 표준)

- **Rendered.ai:**

- **모델:** 높은 '최소 약정' (Teams \$5,000\$/월, Organizations \$15,000\$/월).²
- **'가변 미터':** 이들의 '가변 미터'는 생성된 '이미지 수'가 아닙니다. 대신 "**최대 인스턴스 (Peak Instances)**"(컴퓨팅 파워), "스토리지(GB)", "사용자 수"입니다.²
- **'전문 서비스':** 월 \$15,000\$의 Organizations 플랜에는 "**기술 지원 매니저(TAM)**" 가 포함되어, 정기 미팅과 고객 백로그 관리를 제공합니다.²
- **분석:** 이는 완벽한 **시뮬레이션 PaaS** 모델입니다.¹² 고객은 이미지 자체가 아닌, "이미지를 자유롭게 생성할 수 있는 3D 스튜디오(PaaS)"의 월간 이용료를 지불합니다.

- **Synthesis AI:**

- **모델:** 고객에게 명확한 두 가지 선택지를 제공합니다. (1) 플랫폼 접근을 위한 연간 구독 (월 \$3,000부터) 또는 (2) "커스텀 데이터셋" 제작 프로젝트(최소 \$10,000 의 1회성 비용).³
- **분석:** 이는 "PaaS 구독"과 "DaaS(Data-as-a-Service) 프로젝트"의 명확한 분리입니다. 특히 \$10,000\$라는 최소 프로젝트 비용은, BMS PoC 프로젝트의 "Lite(\$7.5K)" 또는 "Pro(\$18K)" 패키지와 매우 유사한 시장 포지셔닝을 보여줍니다.

- **Datagen:**

- **모델:** "인간 중심(Human-centric)" CV 데이터(얼굴, 전신)에 특화되어 있습니다.³¹ 가격 모델은 혼재되어 있어, "시간당 요금(hourly charge)"³³ 또는 "크레딧" 16을 사용합니다.
- **분석 (1 이미지 = 10 행):** Datagen.in의 크레딧 모델 16 (30,000 크레딧 = 30,000 텍스트 행 또는 3,000 이미지)은 CV 데이터 생성이 정형 데이터 생성보다 **10배의 가치 또는 비용을 가짐**을 벤더 스스로 인정한 정량적 증거입니다.

- **Parallel Domain:**

- **모델:** 자율주행(AV) 시뮬레이션에 고도로 특화되어 있습니다.³⁴ 가격은 비공개 (Enterprise-only)입니다.
- **제품:** API, SDK, 웹 툴을 통해 "카메라, 라이다, 레이더 데이터"를 정적 데이터셋이 아닌 실시간 스트리밍으로 제공합니다.³⁴
- **분석:** 이는 데이터 생성을 넘어선 실시간 시뮬레이션 영역입니다. 비용은 "시뮬레이션 시간" 또는 "스트리밍된 센서 데이터 대역폭"과 연동될 가능성이 높습니다.

CV 시장의 '최소 약정'(\$3,000 ~ \$15,000\$)이 정형/텍스트(\$0 ~ \$500\$)보다 압도적으로 높은 이유는 명확합니다. CV 시장은 **데이터를 파는 것이 아니라, 고도로 전문화된 3D 시뮬레이션 소프트웨어 및 인프라 접근권을 판매합니다.** 고객은 '데이터 포인트'가 아닌 '렌더링 인스턴스'와 '기술 지원(TAM)'을 구매합니다.

이는 BMS 도메인 제약 적용(전문 서비스)이 비용의 핵심인 정형 데이터 시장과 동일한 논리입니다. 다만, CV 시장에서는 이 '전문 서비스' 비용(기술 지원, 인프라 관리)이 '플랫폼 최소 약정'에 이미 선반영되어 있다는 구조적 차이가 존재합니다.

V. 공급 및 배포 모델 비교 분석 (API, SaaS, On-Premise)

합성데이터의 가격은 '무엇을' 사는지(모달리티)뿐만 아니라 '어떻게' 공급받는지(공급 정책)에 의해서도 크게 좌우됩니다. 엔터프라이즈 고객에게 공급 정책(API, VPC, On-Premise)의 선택은 데이터 보안, 총 소유 비용(TCO), 운영 제어권과 직결되는 중대한 결정입니다.

A. API 기반 접근 (Public SaaS)

- 작동 방식:** 벤더가 호스팅하는 퍼블릭 클라우드에서 API를 호출하여 데이터를 생성하고 응답 받습니다. (예: YData SDK 4, Gretel API 35).
- 가격 모델:** 순수 PAYG. 토큰, API 호출, 레코드당 과금.
- 장점:** 초기 비용(\$0\$) 및 인프라 관리가 불필요하며 즉각적인 사용이 가능합니다.
- 단점:** 데이터 유출 리스크. 민감한 원본 데이터(예: 고객 PII, BMS 원본 로그)를 벤더의 API 엔드포인트로 전송해야 하는 치명적인 보안 문제가 있습니다.³⁶

B. 플랫폼 구독 (VPC / Private Cloud Marketplace)

- 작동 방식:** 벤더가 AWS 또는 Azure Marketplace를 통해 소프트웨어를 판매합니다. 이 소프트웨어는 고객의 **VPC(가상 사설 클라우드)** 내에 배포됩니다.⁴
- 가격 모델:** '플랫폼 최소 약정' + '가변 미터' + "고객의 클라우드 인프라 비용" (이중 과금).
- 사례:** MOSTLY AI (AWS Marketplace) 11, YData Fabric (AWS/Azure) 21, Tonic.ai (Cloud or Self-Hosted).⁵
- 장점:** 데이터 보안성 극대화. 원본 데이터가 고객의 VPC 네트워크를 절대 벗어나지 않으므로 API 방식의 보안 리스크가 원천 차단됩니다.
- 단점:** 이중 비용 구조. 고객은 벤더에게 소프트웨어 라이선스비(예: MOSTLY AI \$3,000\$/월)와 AWS/Azure에 인프라 사용료(vCPU/vGPU 비용) 21를 별도로 지불해야 합니다.

C. 온프레미스(On-Premise) / 연간 라이선스

- 작동 방식:** 고객이 자신의 데이터 센터(On-Premise)에 소프트웨어를 직접 설치합니다(Air-gapped).
- 가격 모델:** 고가의 연간 라이선스(ARC). (통상 연 \$80,000 ~ \$200,000\$).
- 사례:** MOSTLY AI, YData, Tonic의 'Enterprise' 플랜.
- 장점:** 최고 수준의 보안 및 완전한 운영 제어권.
- 단점:** 가장 높은 초기 비용, 고객의 자체적인 인프라 구축 및 유지보수 부담.

D. 프로젝트 기반 (Managed Service / DaaS)

- 작동 방식:** 고객이 벤더(또는 파트너사)에게 특정 SOW(작업 명세서)를 기반으로 데이터셋 생성을 일괄 의뢰합니다.
- 가격 모델:** 1회성 프로젝트 비용 (NRE).
- 사례:** Synthesis AI ("커스텀 데이터셋" 최소 \$10,000).³ 그리고 BMS PoC 제안 (Lite \$7.5K, Pro \$18K).
- 장점:** 고정된 비용, 플랫폼 학습 불필요, 벤더가 결과물을 보장.
- 단점:** 확장성 및 반복성 부족. (새로운 데이터셋이 필요할 때마다 새 계약 필요).

핵심 테이블 3: 공급 모델별 TCO 및 보안 영향 분석

공급 정책	비용 구조	보안 수준	데이터 이동성	BMS 프로젝트 적용 전략
API (Public SaaS)	PAYG (낮은 초기 비용)	낮음 (데이터 외부 전송)	높음	간단한 데모 또는 비민감 데이터 증강용.
VPC (Marketplace)	\$3K+ MRC + 인프라 비용 (이중 과금)	높음 (VPC 내 처리)	없음	BMS 원본 데이터 보안이 중요할 때. (고객에게 '이중 과금' 구조 설명 필수)
On-Premise (License)	\$80K+ ARC (높은 초기 비용)	최고 (Air-gapped)	없음	최고 수준의 보안을 요구하는 금융/국방 고객용.
		높음 (벤)	낮음	현재 PoC 모델. 고객의 플랫

Project (Managed)	\$10K+ NRE (고정 비용)	더/파트 너가 처 리)	(결과 물만 전달)	폼 도입 장벽을 제거하는 가 장 효율적인 방식.
----------------------	-----------------------	--------------------	------------------	-------------------------------

VI. 전략적 결론 및 권고안

본 심층 분석은 글로벌 합성데이터 시장의 가격 책정 모델이 모달리티, 사용량 측정 기준, 공급 정책에 따라 다각화되어 있음을 확인했습니다. 이는 특정 BMS 시계열 데이터 증강 프로젝트의 견적을 산출하고 시장 진입 전략을 수립하는 데 다음과 같은 전략적 시사점을 제공합니다.

A. 내부 분석의 검증

BMS 시계열 데이터 증강 PoC를 위해 수립된 (플랫폼 최소 과금) + (사용량) + (전문 서비스)라는 3종 가격 책정 모델은, 글로벌 합성데이터 시장, 특히 고가치 엔터프라이즈 부문의 표준 모델임이 검증되었습니다.

또한, \$10,000 ~ \$40,000\$ 범위의 PoC 비용 및 연간 \$80,000 ~ \$200,000\$의 엔터프라이즈 라이선스 비용 추정치는, Synthesis AI(최소 \$10,000\$ 프로젝트) 3 및 Rendered.ai(월 \$15,000\$ 구독) 2와 같은 타 모달리티 벤더들의 공개된 가격과 비교할 때 매우 현실적이고 시장 기준에 부합하는(**Market-aligned**) 수준입니다.

B. 핵심 결론: 모달리티가 가격 구조를 결정한다

본 보고서는 이 3종 요금제의 비중이 데이터 모달리티에 따라 근본적으로 달라짐을 규명했습니다.

- 정형/시계열 (BMS):** '가변 미터'(\$172)는 미미합니다. 비용의 99%는 '전문 서비스'(\$18,000)에서 발생합니다. 이는 '도메인 제약(Physics/Rules)' 적용이 핵심 가치이기 때문입니다.
- 텍스트 (LLM):** '가변 미터'(토큰 비용)가 비용의 상당 부분을 차지하며, 이는 '교사 LLM의 추론 비용'과 직결됩니다.¹
- 이미지/비전 (CV):** '최소 약정'(\$5,000 ~ \$15,000)이 비용의 대부분을 차지합니다. 이는 '3D 시뮬레이션 인프라 접근권'과 '기술 지원(TAM)' 비용이 선반영된 결과입니다.²

C. 전략적 권고

- '가변 미터'의 전략적 선택:

BMS 프로젝트를 위해 벤더 플랫폼을 선택할 때, '가변 미터'의 종류(컴퓨팅 vs. 볼륨)를 전략적으로 고려해야 합니다.

- YData (볼륨 기반) 4 활용 시: '가변 미터' 비용은 \$172.80로 고정됩니다. 이는 고객에게 "플랫폼 비용은 저렴하며, 지불하는 \$18,000\$는 순수하게 파트너사(Pebblous)의 BMS 도메인 전문성에 대한 대가"임을 명확히 보여줄 수 있어 **가치 전달에 가장 유리합니다**.
- MOSTLY AI (컴퓨팅 기반) 13 활용 시: 복잡한 BMS 모델은 더 많은 vCPU/vGPU 시간을 소모하므로 '가변 미터' 비용(크레딧)이 \$172보다 훨씬 높게 나올 수 있습니다. 이는 고객에게 청구되는 총 비용 중 벤더 플랫폼 비중이 높아져, 파트너사의 '전문 서비스' 가치가 상대적으로 희석될 수 있습니다.

2. '전문 서비스' 패키징 강화:

시장은 '데이터 생성' 자체의 상향 평준화(Commoditization)를 겪고 있습니다. 진정한 차별점은 "Pro (물리 제약 + 시나리오)" 및 "Enterprise (품질 게이트 + 운영 파이프라인)"와 같은 도메인 특화 전문 서비스입니다. 이 '전문 서비스'를 단순한 "옵션"이 아닌, 합성데이터 프로젝트 성공을 위한 필수 구성 요소로 포지셔닝해야 합니다. \$18,000\$의 "Pro" 패키지는 데이터 생성이 아닌, **'BMS 엔지니어링 컨설팅'**으로 판매되어야 합니다.

3. '공급 정책'을 활용한 고객 세분화:

'프로젝트 기반(Managed Service)' 접근은 PoC 및 신규 고객 확보에 최적입니다. 장기 고객 및 대형 고객을 위해서는 'VPC 배포 + 연간 라이선스' 모델을 준비해야 합니다. 이는 고객사(예: 대형 배터리 제조사)가 "BMS 생성 모델"을 자사의 보안 환경(VPC) 내에서 반복적으로 사용할 수 있게 해주는 고부가가치(High-ARC) 상품이 될 것입니다. 이 모델은 MOSTLY AI(월 \$3,000\$) 및 Rendered.ai(월 \$15,000\$)의 엔터프라이즈 전략과 정확히 일치합니다.

VII. 참고문헌 (Sources Used)

1. [Amazon Bedrock pricing](#)
2. [Solutions Pricing for AI Synthetic Data Generation Needs](#)
3. [Human Faces Synthetic Dataset - AWS Marketplace](#)
4. [YData data quality for Data Science | Synthetic data Data-Centric AI](#)
5. [Pricing – Tonic.ai](#)
6. [Pay-As-You-Go Cloud Solution from Tonic | Blog](#)
7. [Gretel.ai Reviews 2025: Pricing & Features - Tekpon](#)
8. [Gretel.ai | BrXnd.ai Landscape](#)

9. Hazy: Set your data free with synthetic data solutions - Dynamic ...
10. Pricing - MOSTLY AI
11. AWS Marketplace: MOSTLY AI Data Intelligence Platform
12. synthetic data platform as a service (paas) - Rendered.ai
13. Usage and credits - Docs - Mostly AI
14. What's new in MOSTLY AI
15. Gretel.ai Pricing 2025
16. DataGen - AI Synthetic Data Solutions | Generative AI Models and Custom Datasets
17. Billing and Usage | Gretel.ai
18. What Is Synthetic Data? - Salesforce
19. What is the ROI of synthetic data? - Syntho
20. Synthetic data tools: Open source or commercial? A guide to building vs. buying - Medium
21. AWS Marketplace: YData Fabric - Data quality for data science
22. Gretel.ai Billing and Usage
23. Generating synthetic data with differentially private LLM inference - Google Research
24. Tonic.ai: Synthetic Test Data Generation for Software and AI Engineers
25. Introducing world's largest synthetic open-source Text-to-SQL dataset
26. Differential Privacy and Synthetic Text Generation with Gretel: Making Data Available at Scale (Part 1)
27. Generate Differentially Private Synthetic Text with Gretel GPT
28. Balancing Cost and Effectiveness of Synthetic Data Generation Strategies for LLMs - arXiv
29. What is Synthetic Data | CVEDIA | AI Video Analytics for any hardware
30. Top 6 Synthetic Data Generation Tools [2025] - Averroes AI
31. DataGen Technologies - Features, Pricing & Reviews - Welcome.AI
32. Datagen Reviews & Ratings 2025 - TrustRadius
33. Datagen lands \$50M to build out its synthetic data platform for AI training - SiliconANGLE
34. Home - Parallel Domain - Parallel Domain
35. Gretel.ai REST API Documentation
36. Synthetic data: save money, time and carbon with open source - Hugging Face

37. mostly-ai/aws-marketplace: MOSTLY AI Marketplace Solution Installation Guide - GitHub
38. YData Fabric available on Azure and AWS marketplaces

Works cited

1. Amazon Bedrock pricing, accessed November 8, 2025, <https://aws.amazon.com/bedrock/pricing/>
2. Solutions Pricing for AI Synthetic Data Generation Needs, accessed November 8, 2025, <https://rendered.ai/pricing/>
3. Human Faces Synthetic Dataset - AWS Marketplace, accessed November 8, 2025, <https://aws.amazon.com/marketplace/pp/prodview-hkxlb5jtkrics>
4. YData data quality for Data Science | Synthetic data Data-Centric AI, accessed November 8, 2025, <https://ydata.ai/>
5. Pricing - Tonic.ai, accessed November 8, 2025, <https://www.tonic.ai/pricing>
6. Pay-As-You-Go Cloud Solution from Tonic | Blog, accessed November 8, 2025, <https://www.tonic.ai/blog/tonic-now-offers-a-pay-as-you-go-cloud-based-solution>
7. Gretel.ai Reviews 2025: Pricing & Features - Tekpon, accessed November 8, 2025, <https://tekpon.com/software/gretel-ai/reviews/>
8. Gretel.ai | BrXnd.ai Landscape, accessed November 8, 2025, <https://landscape.brxnd.ai/companies/gretelai>
9. Hazy: Set your data free with synthetic data solutions - Dynamic ..., accessed November 8, 2025, <https://dynamicbusiness.com/ai-tools/hazy-set-your-data-free-with-synthetic-data-solutions.html>
10. Pricing - MOSTLY AI, accessed November 8, 2025, <https://mostly.ai/pricing>
11. AWS Marketplace: MOSTLY AI Data Intelligence Platform, accessed November 8, 2025, <https://aws.amazon.com/marketplace/pp/prodview-clqfgzfzznfoc>
12. synthetic data platform as a service (paas) - Rendered.ai, accessed November 8, 2025, <https://rendered.ai/platform/>
13. Usage and credits - Docs - Mostly AI, accessed November 8, 2025, <https://docs.mostly.ai/usage>
14. What's new in MOSTLY AI, accessed November 8, 2025, <https://mostly.ai/docs/whats-new>

15. Gretel.ai Pricing 2025, accessed November 8, 2025,
<https://www.g2.com/products/gretel-ai/pricing>
16. DataGen - AI Synthetic Data Solutions | Generative AI Models and Custom Datasets, accessed November 8, 2025, <https://datagen.in/>
17. Billing and Usage | Gretel.ai, accessed November 8, 2025,
<https://docs.gretel.ai/operate-and-manage-gretel/enterprise-support/billing-and-usage>
18. What Is Synthetic Data? - Salesforce, accessed November 8, 2025,
<https://www.salesforce.com/data/synthetic-data/>
19. What is the ROI of synthetic data? - Syntho, accessed November 8, 2025,
<https://www.syntho.ai/what-is-the-roi-of-synthetic-data/>
20. Synthetic data tools: Open source or commercial? A guide to building vs. buying - Medium, accessed November 8, 2025, <https://medium.com/statice/synthetic-data-tools-open-source-or-commercial-a-guide-to-building-vs-buying-580ddee30e8>
21. AWS Marketplace: YData Fabric - Data quality for data science, accessed November 8, 2025, <https://aws.amazon.com/marketplace/pp/prodview-hgrqd5lqnqblm>
22. accessed November 8, 2025, <https://docs.gretel.ai/operate-and-manage-gretel/enterprise-support/billing-and-usage#:~:text=Gretel%20bills%20based%20off%20of,%2C%20Developer%2C%20Team%20and%20Enterprise.>
23. Generating synthetic data with differentially private LLM inference - Google Research, accessed November 8, 2025,
<https://research.google/blog/generating-synthetic-data-with-differentially-private-llm-inference/>
24. Tonic.ai: Synthetic Test Data Generation for Software and AI Engineers, accessed November 8, 2025, <https://www.tonic.ai/>
25. Introducing world's largest synthetic open-source Text-to-SQL dataset, accessed November 8, 2025, <https://gretel.ai/blog/synthetic-text-to-sql-dataset>
26. Differential Privacy and Synthetic Text Generation with Gretel: Making Data Available at Scale (Part 1), accessed November 8, 2025,
<https://gretel.ai/blog/differentially-private-synthetic-text-generation-at-scale-part-1>
27. Generate Differentially Private Synthetic Text with Gretel GPT, accessed

November 8, 2025, <https://gretel.ai/blog/generate-differentially-private-synthetic-text-with-gretel-gpt>

28. Balancing Cost and Effectiveness of Synthetic Data Generation Strategies for LLMs - arXiv, accessed November 8, 2025,
<https://arxiv.org/html/2409.19759v3>
29. What is Synthetic Data | CVEDIA | AI Video Analytics for any hardware, accessed November 8, 2025, <https://www.cvedia.com/what-is-synthetic-data>
30. Top 6 Synthetic Data Generation Tools [2025] - Averroes AI, accessed November 8, 2025, <https://averroes.ai/blog/synthetic-data-generation-tools>
31. DataGen Technologies - Features, Pricing & Reviews - Welcome.AI, accessed November 8, 2025, <https://welcome.ai/solution/datagen-technologies>
32. Datagen Reviews & Ratings 2025 - TrustRadius, accessed November 8, 2025, <https://www.trustradius.com/products/datagen-tech/reviews>
33. Datagen lands \$50M to build out its synthetic data platform for AI training - SiliconANGLE, accessed November 8, 2025,
<https://siliconangle.com/2022/03/23/datagen-lands-50m-build-synthetic-data-platform-ai-training/>
34. Home - Parallel Domain - Parallel Domain, accessed November 8, 2025,
<https://paralleldomain.com/>
35. Gretel.ai REST API Documentation, accessed November 8, 2025,
<https://api.docs.gretel.ai/>
36. Synthetic data: save money, time and carbon with open source - Hugging Face, accessed November 8, 2025, <https://huggingface.co/blog/synthetic-data-save-costs>
37. mostly-ai/aws-marketplace: MOSTLY AI Marketplace Solution Installation Guide - GitHub, accessed November 8, 2025, <https://github.com/mostly-ai/aws-marketplace>
38. YData Fabric available on Azure and AWS marketplaces, accessed November 8, 2025, <https://ydata.ai/resources/ydata-fabric-available-on-azure-and-aws-with-15-day-free-trial.html>