Pebblous

페블러스 데이터 그린하우스 개념 설계서

> ***Neuro-Symbolic AI-Ready Data Infrastructure Blueprint***

- 기획: 페블러스 데이터커뮤니케이션팀
- 버전: 0.1
- 작성일: 2025-12-27
- 공개여부: 비공개

목차

0. 서문: Data Clinic에서 Data Greenhouse로

Pebblous는 "데이터를 눈으로 보고, 수치로 진단하며, 행동으로 개선한다"는 철학을 Data Clinic이라는 제품으로 구현해 왔다. Data Clinic이 '진단과 치료의 순간'을 중심으로 데이터 품질 문제를 해결했다면, Data Greenhouse는 그 다음 단계로서 '데이터가 스스로 성장하고 증명되는 운영 체계'를 지향한다. 즉, Data Greenhouse는 데이터 품질을 일회성 과제가 아니라 지속적으로 운영되어야 하는 산업 인프라로 정의하며, 이 운영의 중심에 자율형 에이전트(AADS)와 뉴로-심볼릭 진단 구조를 둔다.

Data Greenhouse라는 이름은 단순한 비유가 아니다. 온실은 생물을 "그냥 두면 잘 크겠지"라는 낙관으로 방치하지 않고, 관측과 제어, 기록과 검증을 통해 목적에 맞는 성장 곡선을 만들어낸다. Data Greenhouse는 AI 데이터 역시 동일하게 다룬다. 데이터는 쌓아두기만 하면 자산이 되는 것이 아니라, 품질과 비용, 규제와 신뢰의 조건을 만족할 때에만 산업 자산으로 기능한다.

I. C-Level Executive Summary

1. 문제 정의: 플랫폼은 있지만, "판단"이 없다

오늘날 많은 기업은 Snowflake, Databricks, Data Lake와 같은 고급 데이터 플랫폼을 이미 보유하고 있다. 그러나 플랫폼의 도입은 데이터의 "저장과 처리"를 가능하게 했을 뿐, 데이터가 AI 성과에 실제로 기여하고 있는지, 데이터가 비용을 낭비하고 있는지, 데이터가 규제와 감사에 대응 가능한 형태로 관리되고 있는지에 대한 답을 자동으로 제공하지 않는다. 이 간극 때문에 경영은 세 가지 구조적 문제를 동시에 마주한다.

- 첫째, 데이터와 GPU 비용은 증가하지만 그 증가가 불가피한 성장인지 단순한 낭비인지 설명되지 않는다.
- 둘째, 모델 성능 변화의 원인이 데이터 문제인지 모델 문제인지 분해되지 않아, 조직은 "더 큰 모델, 더 많은 GPU, 더 많은 데이터"라는 비싼 답으로 도망치기 쉽다.
- 셋째, 강화되는 규제 환경 속에서 데이터 품질과 운영의 증적을 제시하지 못해, AI의 상용화는 기술이 아니라 신뢰의 문제에서 좌초될 수 있다.

2. 해결 정의: Data Greenhouse는 AI 데이터의 운영 체계다

Pebblous Data Greenhouse는 기존 데이터 플랫폼을 대체하지 않는다. 오히려 Data Greenhouse는 Snowflake, Databricks, Data Lake를 "플랫폼 계층"으로 명확히 하부에 두고, 그 위에 데이터의 관측·판단·행동·증명을 자동화하는 운영 체계를 얹는다. 이 운영 체계는 관측(Observation), 오케스트레이션(Orchestration), 행동(Action), 거버넌스(Governance)로 구성되며, 플랫폼 어댑터

(Platform Adapter)는 이 운영 체계가 기존 플랫폼과 안전하게 연결되는 접점으로서 작동한다. 이 구조는 기업이 이미 투자한 플랫폼 자산을 존중하면서도, 플랫폼이 답하지 못하는 질문에 답하도록 설계된다.

3. 핵심 가치: 비용, 성능, 규제를 동시에 다룬다

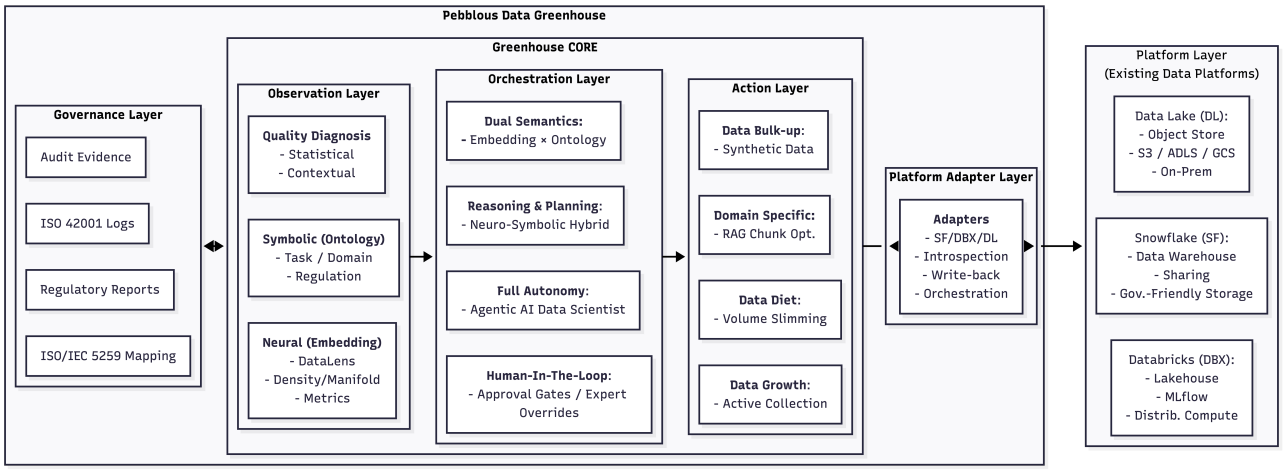
Data Greenhouse의 첫 번째 경영 가치는 **비용의 구조적 통제**에 있다. Data Greenhouse는 "쿼리를 더 빠르게" 만드는 방식이 아니라, "그 쿼리가 돌린 데이터 중 실제로 정보 기여도가 있는 비율은 얼마인가"라는 질문을 가능하게 함으로써 비용의 원인을 데이터 구조에서 제거한다. 중복과 과밀로 인해 정보 기여도가 낮은 데이터는 Data Diet로 줄어들고, 대표성 공백은 Data Bulk-up으로 보강된다. 비용은 더 이상 불가피한 결과가 아니라 설명 가능한 의사결정의 결과가 된다.

두 번째 가치는 **성능의 예측 가능성**이다. Data Greenhouse는 성능 저하를 "모델 탓"으로 단정하지 않고, 임베딩 공간과 온톨로지 기반의 진단을 통해 데이터 분포의 붕괴, 중복의 증가, 커버리지의 결손 같은 구조적 원인을 제시한다. 경영은 성능 문제를 기술팀의 감각이나 경험에 의존해 논쟁하는 대신, "어떤 데이터가 부족하며 어떤 데이터가 과잉인지"라는 근거를 바탕으로 투자 방향을 결정할 수 있다.

세 번째 가치는 **신뢰와 규제 대응**이다. Data Greenhouse는 ISO/IEC 5259 기반의 품질 특성 매핑과 ISO 42001 수준의 감사 로그 생성 구조를 운영 과정에 내재화한다. 이 방식은 규제를 사후 문서 작업으로 처리하는 것이 아니라, 운영 그 자체를 증적 중심으로 설계함으로써 규제를 "추가 비용"이 아니라 "시장 진입 조건을 충족하는 체계적 역량"으로 전환한다.

4. 경영 결론: Data Greenhouse는 플랫폼 위의 책임 레이어다

Data Greenhouse는 플랫폼 선택을 대신해 주지 않는다. 대신 Data Greenhouse는 그 선택이 옳았는지, 그리고 현재의 운영 방식이 경제적인지, 안전한지, 규제에 부합하는지 "설명 가능한 형태"로 드러낸다. Data Greenhouse가 지향하는 바는 단순한 도구 도입이 아니라, **기업이 AI 시대에 요구받는 데이터 책임을 운영 체계로 구현하는 일**이다.



II. 중간 관리자를 위한 간략 명세서

1. 시스템 범위: 무엇을 바꾸고, 무엇을 바꾸지 않는가

Data Greenhouse는 데이터 플랫폼을 교체하는 프로젝트가 아니다. Snowflake는 그대로 데이터 웨어하우스와 데이터 공유·거버넌스 친화적 저장소 역할을 유지한다. Databricks는 그대로 레이크하우스 및 ML 워크로드, MLflow 기반의 실험 추적 및 분산 컴퓨팅의 중심으로 유지된다. Data Lake는 그대로 오브젝트 스토리지와 원천 데이터 저장소로 유지된다. Data Greenhouse가 추가하는 것은 플랫폼 위에서 데이터 품질의 상태를 지속적으로 진단하고, 개선 행동을 실행하며, 규제와 감사에 필요한 증거를 자동으로 축적하는 운영 레이어다.

2. 운영 모델: 관측–판단–행동–증명 루프

Data Greenhouse의 핵심 루프는 네 단계로 구성된다.

- 첫째, **Observation Layer**는 임베딩 기반의 분포 분석과 온톨로지 기반의 맥락 해석을 결합하여 데이터 품질을 진단한다.
- 둘째, **Orchestration Layer**(AADS)는 진단 결과를 해석해 개선 계획을 세우고 실행을 설계한다.
- 셋째, **Action Layer**는 Diet, Bulk-up, RAG Chunk 최적화, 능동 수집과 같은 구체적 개선 행동을 수행한다.
- 넷째, **Governance Layer**는 이 모든 활동을 ISO 표준과 규제 요구에 맞춰 기록하고, 감사 가능한 증거와 리포트를 생성한다. 이 과정은 "운영의 일부로서의 컴플라이언스"를 구현하며, 조직이 AI 도입 과정에서 반복적으로 겪는 병목을 줄인다.

3. 의사결정 구조: 자율성과 통제의 균형

Data Greenhouse는 전면 자동화를 목표로 하면서도, 중요한 결정에는 승인 게이트를 둔다. 예컨대 대규모 삭제, 합성 데이터 생성의 대량 적용, 운영 데이터 파이프라인의 정책 변경과 같은 행위는 Human-in-the-Loop를 통해 관리자 또는 도메인 전문가의 승인을 받도록 설계된다. 이러한 구조는 자율성과 안전성을 동시에 확보하며, 운영 조직이 "자동화의 결과에 책임질 수 없는 상태"에 빠지지 않게 만든다.

4. 관리 지표: 무엇을 보면 운영이 잘 되는가

Data Greenhouse의 운영 성과는 단순히 데이터 양이 아니라 데이터의 구조적 건강성으로 측정된다. 관리자는 중복률, 커버리지, 대표성 공백, 품질 지수(QI)의 변화, 개선 행동 전후의 비용 변화, 그리고 감사 증적의 완결성을 함께 관찰한다. Data Greenhouse가 축적하는 품질 기록은 장기적으로 조직의

System of Record로 기능하며, 시간이 쌓일수록 전환 비용이 증가하는 구조적 락인을 만든다. 이 락인은 벤더 종속이 아니라, "**조직의 의사결정이 데이터 건강 기록 위에서 수행되기 시작한다**"는 의미의 운영 락인이다.

III. 개발자를 위한 상세 설명

1. 아키텍처 원칙: 뉴로-심볼릭을 "구현 구조"로 만든다

Data Greenhouse의 기술적 차별점은 뉴로-심볼릭(Neuro-Symbolic) 전략을 단순한 슬로건이 아니라 아키텍처의 중심으로 구현한다는 데 있다. 임베딩은 데이터의 통계적 현상과 기하학적 구조를 보여주지만, 그것이 문제인지 의미 있는 현상인지는 말해주지 않는다. 온톨로지는 규칙과 맥락, 책임과 규제를 제공하지만, 데이터 분포의 실체를 정량화하지 못한다. Data Greenhouse는 이 둘을 결합해 "**무엇이 이상한가**"와 "**왜 중요한가**"를 동시에 산출하고, 그 결과를 에이전트가 실행 가능한 계획으로 바꾸도록 설계한다.

2. Platform Adapter Layer: 이동을 최소화하고, 연결을 최대화한다

Platform Adapter Layer는 데이터 **복제와 이동을 최소화**하는 것을 원칙으로 한다. 어댑터는 플랫폼(SF/DBX/DL)의 메타데이터, 테이블·파일 스키마, 작업 실행 이력, 비용·사용량, 로그 및 계통 정보(linage) 신호를 관찰한다. 또한 개선 행동의 결과를 플랫폼에 다시 반영하기 위해 태그, 스냅샷, 머티리얼라이즈, 파티션 정책, 라우팅과 같은 형태로 write-back을 수행한다. 즉, 어댑터는 단순 커넥터가 아니라 "**관찰과 반영의 접점**"이며, 운영 흐름에서 가장 하부에 위치해 플랫폼 계층과 Data Greenhouse를 분리하고, 동시에 연결한다.

- SF: Snowflake
- DBX: Databricks
- DL: Data Lake 류

3. Observation Layer: 임베딩 공간과 온톨로지가 함께 진단한다

Observation Layer는 Neural(Embedding)과 Symbolic(Ontology)을 동시에 사용한다. Neural 측면에서 데이터렌즈(DataLens)는 원천 데이터를 임베딩 공간에 매핑하며, 이 공간에서 밀도와 분포, 매니폴드 형상, 커버리지와 공백을 분석한다. 이 과정은 중복과 유사성을 데이터의 "과밀"로, 대표성 결손을 데이터의 "공백"으로 드러내며, 개선 전후를 IOD/MIOD 형태로 비교 가능하게 만든다.

- IOD: Image of Data. 데이터렌즈를 통한 임베딩 벡터

- MIOD: Modified Image of Data. 데이터의 분포적 품질 개선을 위한 IOD 공간에서의 개선

Symbolic 측면에서 온톨로지는 태스크(학습, 평가, RAG), 도메인(통신, 제조, 국방 등), 규제(ISO 5259, ISO 42001, EU AI Act 등)의 맥락을 제공한다. Observation은 이 둘을 결합해 통계적 이상치가 단순 오류인지, 도메인 이벤트인지, 혹은 규제 위반 위험인지까지 판별 가능한 진단 결과를 생성한다.

4. Orchestration Layer: 계획-진단-개선-통제를 수행한다

AADS는 Data Greenhouse의 오케스트레이션 계층의 핵심기술이며, "자율형 AI 데이터 사이언티스트"로서 작동한다. AADS는 사용자의 목표를 해석하고, 목표를 실행 가능한 워크플로로 분해하며, 필요한 도구를 호출해 작업을 수행하고, 결과를 보고서로 정리하는 일련의 루프를 수행한다.

- AADS: Agentic AI Data Scientist

이 루프는 계획-진단-개선-통제(Plan–Diagnose–Improve–Govern, PDIG) 구조로 정리되며, 각 단계는 관측 결과와 규칙을 결합해 실행 계획을 만든다. AADS의 핵심은 뉴로-심볼릭 하이브리드 추론이며, 이는 "임베딩 신호 × 온톨로지 규칙"의 형태로 의사결정을 수행하도록 모델링된다. 또한 AADS는 위험도가 높은 행동에 대해 승인 게이트를 포함함으로써, 완전 자율성과 조직 통제 사이의 균형을 구현한다.

5. Action Layer: 경량화, 합성데이터, 액티브 수집을 실행한다

Action Layer는 개선 행동을 실제로 수행하는 실행 계층이다. Data Diet는 과밀 영역에서 중복 또는 고유 정보 기여도가 낮은 데이터를 제거하거나 축소하여 학습 및 저장 비용을 절감한다. Data Bulk-up은 저밀도 공백 영역을 타겟팅해 정밀 합성 데이터를 생성함으로써 대표성과 강건성을 강화한다. RAG Chunk 최적화는 청크 단위의 의미 중복을 제거하고, 질문 분포에 따라 커버리지를 확장하는 방향으로 지식 베이스를 개선한다. Active Collection은 진단 결과로부터 "다음에 무엇을 수집해야 하는지"를 정의하고, 실제 수집 파이프라인 또는 운영 프로세스와 연결되도록 설계된다. Action은 독립적 기능의 모음이 아니라, 관측과 판단의 결과가 플랫폼에 반영되는 지속 운영의 일부로 동작한다.

6. Governance Layer: 표준 매핑과 감사 증적을 파이프라인에 내장한다

Governance Layer는 사후 문서화가 아니라 운영 파이프라인의 일부로 설계된다. Data Greenhouse는 ISO/IEC 5259가 요구하는 품질 특성(유사성, 대표성, 다양성, 효율성 등)을 측정 가능한 지표로 매핑하고, 그 산출 근거를 증거로 축적한다. 또한 ISO 42001 수준의 활동 로그를 자동 생성하여 감사 가능한 추적성을 확보한다. 이 구조는 조직이 고신뢰성 데이터가 요구되는 산업 영역, 특히 Physical AI와 같은 고위험 환경으로 확장할 때 필수적인 신뢰 기반을 제공한다.

7. 배포 및 확장: 클라우드와 온프레미스, 그리고 소버린 AI

Data Greenhouse는 플랫폼 위 운영 레이어이므로, 클라우드 중심 환경뿐 아니라 온프레미스·하이브리드 환경에서도 동일한 개념으로 작동할 수 있어야 한다. 특히 **데이터 주권이 중요한 공공·국방·금융 환경**에서는 외부 통신을 최소화한 배포 옵션이 중요하며, 국가 전략과 정합성을 갖춘 소버린 AI 접근은 상용 확장의 중요한 축이 된다. Data Greenhouse는 이러한 요구를 전제로, 엔터프라이즈 환경에서 필요한 통제성, 감사 가능성, 배포 유연성을 동시에 고려한 운영 인프라로 정의된다.

맺음말: 온실은 작물을 키우고, Greenhouse는 신뢰를 생산한다

Data Greenhouse의 핵심은 "더 많은 데이터"가 아니라 "더 좋은 데이터"이며, 더 정확히는 "AI에 즉시 사용 가능한 **AI-Ready Data를 지속적으로 생산하는 운영 체계**"이다. Data Clinic이 데이터 품질 문제를 진단하고 치료하는 '**병원**'이었다면, Data Greenhouse는 데이터가 스스로 성장하고, 그 성장의 근거가 축적되고, 그 결과가 규제와 산업 요구를 충족하도록 만드는 '**산업용 온실**'이다. 이는 플랫폼과 경쟁하는 제품이 아니라, 플랫폼 위에서 의사결정을 가능하게 만드는 책임 레이어이며, AI 시대의 조직이 결국 필요로 하게 될 데이터 운영의 표준 형태에 가깝다.

Pebblous

> **Pebblous Makes Data Tangible**

contact@pebblous.ai