

miniEDA: 데이터 메시를 활용한 통합 데이터 분석 및 품질 관리 플랫폼

이준원, 김선혁*, 허태욱**

동아대학교, *국립공주대학교, **한국전자통신연구원

2143736@donga.ac.kr, *seonh@kongju.ac.kr, **htw398@etri.er.kr

miniEDA: An Integrated Data Analysis and Quality Management Platform Utilizing Data Mesh

Lee Jun Won, Kim Seon Hyeog*, Heo Tae Wook**

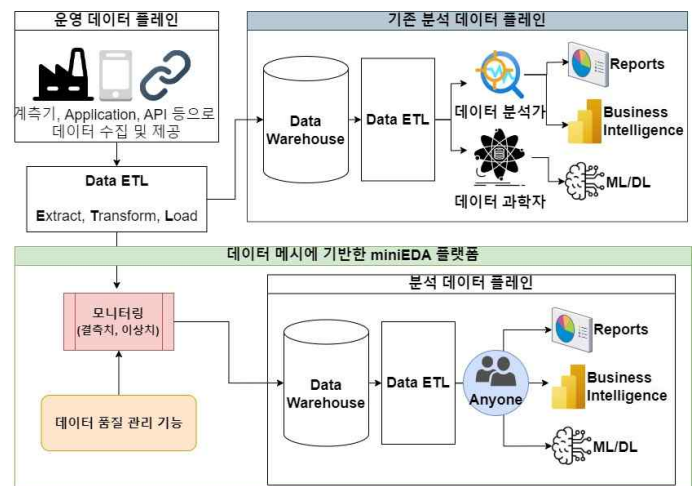
Dong-A Univ., *Kongju National Univ., **Electronics and Telecommunications Research Institute (ETRI)

요약

본 논문은 데이터 사일로와 중앙집중식 데이터 아키텍처의 한계로 일어나는 데이터 품질 관리, 데이터 처리 성능 한계 등의 문제를 효과적으로 해결할 수 있는 데이터 메시를 반영한 통합 플랫폼인 miniEDA를 새롭게 제안한다. miniEDA는 데이터 수집 단계에서부터 모니터링과 ISO 8000 기반의 품질 관리 기능을 제공하여 데이터의 정확성과 일관성을 보장한다. 또한, GUI와 SQL 기반의 환경을 통해 비개발자도 쉽게 데이터를 분석하고 시각화할 수 있으며, 탐색적 데이터 분석 및 데이터 추출, 변환, 적재 과정에서 발생할 수 있는 품질 문제를 예방한다. 마지막으로, miniEDA는 독립적인 도메인에서 데이터를 분석하고 필터 기능을 통해 다른 도메인과 데이터를 공유함으로써 데이터의 재현성과 자율성을 보장한다.

I. 서론

데이터 중심 혁신이 가속화되는 오늘날, 조직들은 방대한 데이터를 효율적으로 관리하고 활용하기 위해 다양한 도구를 도입하고 있다[1][2]. 특히, 제조업과 같은 산업에서는 FEMS(Factory Energy Management System)와 같은 시스템이 데이터를 기반으로 한 운영 효율성 향상에 중요한 역할을 한다. 그러나 <그림 1>과 같이 기존의 단절된 데이터 수집 및 분석 체계는 데이터 품질 관리와 일관성 유지에 어려움을 겪는다. 이러한 분리된 시스템에서는 데이터 중복과 불일치가 발생할 가능성이 있으며, 이는 데이터 분석의 신뢰성을 저하시킬 수 있다. 특히, 데이터 수집과 분석이 별도로 이루어지면서 전체적인 데이터 품질이 저하되는 문제가 발생한다. 세계 4대 회계법인 중 하나인 PwC의 최근 연구에 따르면, AI 프로젝트의 76%가 초기 예산보다 10%에서 30% 초과하여 진행되었음에도 불구하고 손익분기점에 도달하지 못하고 있으며, 여기에는 데이터 품질 문제와 비효율적인 데이터 관리 프로세스가 주요 원인으로 지목된다[3]. 이러한 문제를 해결하고, 복잡한 MLOps 환경 구축이 어려운 개인과 기업들이 데이터를 효과적으로 관리할 수 있도록 miniEDA 플랫폼을 제안한다[2]. miniEDA는 데이터 메시(Data Mesh) 개념을 적용하여 데이터 수집, 실시간 모니터링, ISO 8000 기반의 품질 관리, 데이터 분석, 머신 러닝(ML)까지 모두 포괄하는 통합 환경을 제공한다[4][5]. 운영 데이터 플레인에서 분석 데이터 플레인으로 전달되는 데이터의 결측치와 이상치를 실시간으로 모니터링하고 바로잡을 수 있으며 <그림 1>, 분석 데이터 플레인에서는 기존에 데이터 분석가와 데이터 과학자들이 수행하던 시각화, 분석, ML 작업을 노-코드 웹 인터페이스를 통해 데이터 제공자, 비개발도메인 종사자, 일반 사용자들도 쉽게 활용할 수 있도록 지원한다. 이 플랫폼은 조직이 데이터의 재현성과 자율성을 보장하며, AI와 데이터 분석에서 더 높은 성과를 달성할 수 있도록 돕는다[6]. 본문에서는 miniEDA가 어떻게 이러한 목표를 실현하고, 다양한 산업 분야에서 데이터를 활용할 수 있을지를 설명한다.



<그림 1> 기존 인프라와 miniEDA 플랫폼 인프라

II. 본론

2.1. 데이터 수집 및 품질 관리

miniEDA의 첫 번째 주요 기능은 데이터 수집 단계에서 실시간 모니터링과 ISO 8000 기반의 품질 관리 기능을 제공하는 것이다. 기존 시스템에서는 데이터 제공자와 분석자가 분리되어 있어, 분석자가 필요한 데이터를 적시에 확보하고 그 품질을 지속적으로 추적하는 데 어려움이 있었다. miniEDA는 이러한 문제를 해결하기 위해 데이터 수집 과정에서 실시간 모니터링을 지원하며, ISO 8000 표준을 준수하는 품질 관리 기능을 내장하여 수집된 데이터의 정확성과 일관성을 보장한다. 이로 인해, 이후의 데이터 분석 과정에서도 높은 신뢰성을 유지할 수 있다.

데이터 품질은 머신 러닝 모델의 성능과 AI 애플리케이션의 신뢰성에 직

접적인 영향을 미치는 핵심 요소이다. 연구에 따르면, 불완전하거나 오류가 포함된 데이터는 잘못된 결정을 유도할 만한 신뢰할 수 없는 ML 모델을 생성할 수 있다[6]. 따라서 데이터의 정확성, 완전성, 일관성 등 다양한 품질 측면을 관리하고 보장하는 것이 중요하다. miniEDA는 <표 1>처럼 6가지 품질 관리 차원을 통해 이러한 데이터 품질을 유지하고, 이를 통해 궁극적으로 AI 및 ML 성능을 극대화하는 것을 목표로 한다.

<표 1> ISO 8000 기반 6가지 데이터 품질 관리 특성

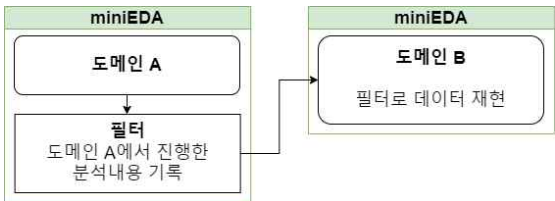
품질 특성	설명
완전성	데이터가 누락 없이 완전한지, 필요한 모든 정보가 포함되어있는지를 판단.
유일성	데이터가 중복되지 않았는지 판단.
유효성	데이터가 형식, 타입, 범위 등의 규칙을 준수하고 있는지 판단.
일관성	데이터가 동일한 상황에서 일관되게 기록되고 있는지 판단.
적시성	데이터가 적절한 시간에 제공되었는지, 실시간 데이터가 최신 상황을 반영하고 있는지 판단.
정확성	데이터가 실제 상황을 제대로 표현하고 있는지 판단.

2.2. 비개발자를 위한 GUI/SQL 기반 데이터 분석 환경

miniEDA의 두 번째 주요 기능은 Data Warehouse에서 GUI와 SQL을 기반으로 한 데이터 추출/변환/적재(ETL) 및 탐색적 데이터 분석(EDA) 기능을 제공하는 것이다. 이 기능은 비개발자도 자신의 도메인 지식을 활용하여 데이터를 분석하고 시각화할 수 있도록 설계되었다. ETL 프로세스의 자동화는 데이터 흐름을 개선하여 비즈니스 투자 수익을 극대화하는데 중요한 역할을 하며, miniEDA는 이러한 자동화된 프로세스를 통해 최적화된 데이터 처리와 향상된 성과를 비개발자에게도 제공할 수 있도록 구성되었다[7]. 또한, ETL과 EDA 과정에서 발생할 수 있는 데이터 손실이나 변형 문제는 ISO 8000 기반의 품질 관리 기능을 통해 지속적으로 모니터링된다. 이를 통해 기술적 배경이 없는 사용자들도 높은 데이터 품질을 유지하며 작업할 수 있다. GUI를 활용해 데이터의 시각화와 분석이 더욱 용이해지며, 데이터에 대한 통찰을 빠르게 얻을 수 있어, 누구나 효과적으로 데이터를 활용할 수 있게 하는 중요한 요소가 된다.

2.3. 데이터 메시 기반의 도메인별 독립적 분석 및 필터 기능

miniEDA의 마지막 주요 기능은 데이터 메시 기반의 도메인별 독립적 분석 환경 제공과 필터 기능이다. <그림 2>와 같이 각 도메인은 독립된 miniEDA 환경에서 데이터를 분석할 수 있으며, 분석 과정에서 생성된 데이터는 필터 기능을 통해 백그라운드 세션에 안전하게 저장된다. 이 필터 기능을 통해 각 도메인에서 수행된 분석 결과를 다른 도메인과 안전하게 공유할 수 있으며, 이러한 필터 기능을 통해 데이터의 재현성과 무결성을 보장한다. 이를 통해 조직 내 다양한 부서 간의 데이터 자율성과 협업이 촉진되며, 부서 간 협업의 효율성도 크게 향상된다.



<그림 2> 필터를 사용하여 다른 도메인에 쉽게 분석 내용을 공유 및 재현

III. 결론

본 논문에서 제시한 miniEDA 플랫폼은 데이터 수집부터 분석, 품질 관리까지의 전 과정을 통합적으로 관리할 수 있는 솔루션이다. 데이터 수집 단계에서의 실시간 모니터링과 ISO 8000 기반의 품질 관리는 데이터의 신뢰성과 무결성을 높이며, 비개발자를 위한 직관적인 GUI와 SQL 기반의 분석 환경을 통해 누구나 쉽게 데이터를 분석하고 활용할 수 있도록 한다. 또한, 데이터 메시 기반의 도메인별 독립적 분석 환경과 필터 기능은 데이터의 재현성, 자율성, 보안을 보장하며, 조직 내 다양한 부서 간의 협업을 효과적으로 촉진한다. 특히, miniEDA는 FEMS와 같은 공장 에너지 관리 시스템에도 적용할 수 있어, 공장 운영의 효율성을 극대화하고 에너지 소비 데이터를 체계적으로 관리할 수 있으며, 이를 통해 제조 공정에서 발생하는 다양한 데이터의 품질을 높이고 실시간으로 분석하여 공장의 운영 효율성과 생산성을 향상시킬 수 있다. 이러한 기능들을 바탕으로, miniEDA는 데이터 중심 혁신을 추구하는 조직뿐만 아니라, FEMS와 같은 다양한 산업 환경에서도 강력하고 필수적인 도구로 자리매김할 것이다.

ACKNOWLEDGMENT

본 연구는 산업통상자원부(MOTIE)와 한국에너지기술평가원(KETEP)의 지원을 받아 수행한 연구 과제입니다. (No. 20202020900290)

참 고 문 헌

- [1] D. Sculley et al., "Hidden Technical Debt in Machine Learning Systems," in *Advances in Neural Information Processing Systems 28 (NIPS 2015)*.
- [2] D. Kreuzberger et al., "Machine Learning Operations (MLOps): Overview, Definition, and Architecture," in *IEEE Access*, vol. 11, pp. 31866-31879, 2023.
- [3] PwC, "Why Most Organizations' Investments in AI Fall Flat," *Harvard Business Review*, Oct. 2020.
- [4] I. A. Machado et al., "Data mesh: concepts and principles of a paradigm shift in data architectures," *Procedia Computer Science* vol. 196, pp. 263-271, 2022.
- [5] R. Perez-Castillo et al., "DAQUA-MASS: An ISO 8000-61 based data quality management methodology for sensor data," *Sensors*, MDPI vol. 18, no.9, p. 3105, 2018.
- [6] L. Budach et al., "The effects of data quality on machine learning performance," *arXiv preprint arXiv:2207.14529*, 2022.
- [7] V. Radhakrishna et al., "Automating ETL process with scripting technology," *Nirma University International Conference on Engineering (NUICONE)*, Ahmedabad, India, 2012.