초급 데이터 엔지니어를 위한 오픈 소스 기반 데이터 플랫폼 구축 제안

곽두일[1], 박광영[1]
[1]숭실대학교 AI테크노융합학과
ai@soongsil.ac.kr, 1004pky@ssu.ac.kr

Proposal for building an open source-based data platform for entry-level data engineers

Doo-il Kwak[1], Kwang-Young Park[1]
[1]Dept. of AI Techno Convergence, Soongsil University

요 약

빅데이터 및 머신러닝 플랫폼을 구축하기 위해선 많은 하드웨어와 소프트웨어, 데이터 엔지니어가 필수인데, 초급 엔지니어들은 경험 부족으로 인해 기업의 수요를 충족시키지 못하고 있다. 본 논문에서는 초급 데이터 엔지니어가 쉽게 접근 가능한 오픈소스를 활용한 빅데이터 플랫폼과 머신러닝 플랫폼을 통합한 7개층으로 이루어진 '데이터 플랫폼'을 제안한다. 향후 제안하는 플랫폼의 현실적인 검증을 위해 계층간 연계가 얼마나 용이한지에 대해 후속연구가 필요하다.

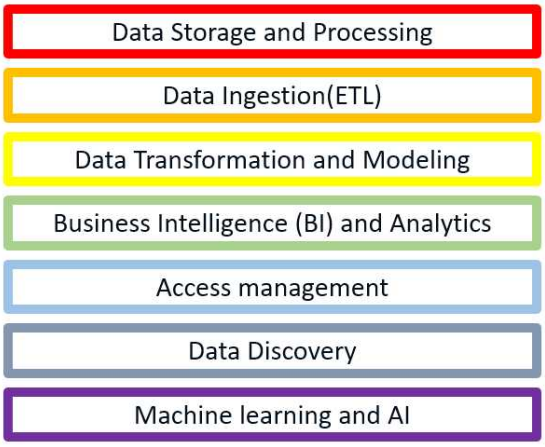
1. 서론

빅데이터 및 머신러닝 플랫폼(서로 다른 개념이지만, 통합하여 '데이터 플랫폼'이라 이후 칭한다.) 구축 및 유지보수는 높은 라이선스 비용과 유지보수 비용, 플랫폼 구축 경험을 갖춘 고급 엔지니어를 필요로 한다. 이 분야에 진출하려는 데이터 엔지니어는 구축 경험의 부재로 취업이 어렵고 프로젝트를 진행하려 해도 고려할 요소가 많아 초기 진입장벽이 높다.

본 논문은 데이터 플랫폼을 실제로 구축해보는 방안을 제시하여 데이터 엔지니어링 분야 입문을 수월하게 하고자 한다. 먼저 데이터 플랫폼의 7개 층의 정의를 차례대로 설명하고, 난이도를 점검하여 각 층에 해당하는 오프소스 소프트웨어 중 LCD(Low Code Development)·NCD(No Coding Development)의 공통 특성에 부합하는 것을 중심으로 일부를 선별하여 소개한다. 마지막으로, 결론과 함께 본 논문의 한계와 향후 연구 방향을 제시한다.

2. 관련 연구

ProvDB[1]는 버전 관리 시스템의 초기 프로토타입을 제안했으며, ModelHub[2]는 딥러닝을 위한 데이터 및 라이프사이클 관리 시스템의 구현을 제안했다. 모델 거버넌스[3]는 프로덕션 ML에서 모델 거버넌스의 필요성과 문제점을 정의하고 해결책을 제안했다. 한편 MLdp[4]는 ModelHub와 같이 데이터 세트의 라이프사이클 관리 시스템을 구현하되 데이터와 모델, 학습 작업간의 종속성을 추적할수 있도록 하였다.



(그림 1) 데이터 플랫폼 7개층.

Pinho et al.은 로코드 개발의 특성에 대한 사용자 관점을 논하고, 로코드 분야가 성장할 것으로 예측했다[5]. 본 논문은 이 논문에서 제시한 LCD·NCD 공통 특성을 기준으로 삼아 최대한 이 기준에 부합하는 각 계층별 오픈소스 소프트웨어를 선별하고자 노력하였다.

<표 1> LCD와 NCD의 공통특성

	특성
1	비개발자 사용
2	드래그 앤 드롭, 시각적 도구
3	높은 추상화
4	간단한 코드 작성
5	모델
6	신속한 애플리케이션 개발
7	라이프사이클 관리
8	클라우드 기반

3. 데이터 플랫폼 7계층

3.1. 데이터 저장 및 처리

컴퓨터 시스템에서 데이터를 저장하고 필요에 따라 처리하는 작업을 말한다. 이는 데이터베이스, 데이터 웨어하우스, 데이터 레이크 및 데이터 파이프라인을 구축하고 유지 관리하면서 데이터를 분석하고 추출하는 것을 포함한다.

<표 2> 데이터 저장 솔루션

솔루션	Hadoop의 HDFS[6]
장점	- 대용량 데이터 세트의 분산 저장 및 처리 가능, 높은 내결함성 제공
단점	- 배치 처리에 최적화되어 있어 실시간 처리에는 적합하지 않음

<표 3> 데이터 처리 소프트웨어

SW	Apache Spark[7]	KNIME Analytics Platform[8]
장점	- 배치 및 실시간 스트리밍 데이터 처리 가능	- 드래그 앤 드롭으로 워크플로우 작성
단점	-초기 설정 및 구성이 어려울수 있음	- UI가 효율적이지 않음

3.2 데이터 수집(ETL)

데이터를 수집하고 저장소에 저장하는 과정이다.

<표 4> 데이터 수집(ETL) 소프트웨어

SW	Airbyte
설명	- 120개 이상의 데이터 커넥터 지원 - 사용자 정의 커넥터 생성 가능(CDK) - 로그 기반 증분 복제 기능 제공

3.3 데이터 변환 및 모델링

데이터 변환과 모델링은 데이터를 다른 형태나 구조로 변환하고, 이를 통해 데이터를 분석, 저장, 관리하는 것이다.

<표 5> 데이터 변환 및 모델링 소프트웨어

SW	Umbrello
대상 사용자	- UML 설계자
지원 DB	- 모든 DB 시스템
주요 기능	- 대부분의 UML 표준 다이어그램, 코드 생성, XML 메타데이터 교환 파일 모델
장점	- 사용 편의성
단점	- 타 데이터 모델링에 비해 기능 제한

3.4 비즈니스 인텔리전스 및 분석

비즈니스 인텔리전스(Business Intelligence, 이하 BI)란 조직 내외부의 데이터를 수집, 분석, 시각화하여 의사 결정에 도움을 주는 기술이고, 비즈니스 애널리틱스(Business Analytics, 이하 BA)는 BI의 확장 개념으로, 기존 BI의 데이터 수집 및 분석에 더해 예측 분석, 통계 분석, 데이터 마이닝(Data Mining) 등의 기술을 이용하여 미래의 경영 전략을 수립하고 실행할 때 필요한 정보를 제공한다.

<표 6> 비즈니스 인텔리전스 소프트웨어

SW	BIRT	Metabase
장점	- 사용하기 쉽고, 커스터마이징 가능	- 설치가 쉽고, 여러 데이터 소스에서 사용 가능
단점	- 데이터 소스에 대한 지원 제한적	- 데이터 시각화 옵션이 제한적임 - Mac 운영체제에서만 동작

3.5. 액세스 관리

이 층은 정보 기술(IT) 보안에서 중요한 개념 중 하나로, 허가되지 않은 사용자나 장치가 시스템, 네트워크 또는 응용 프로그램에 액세스를 방지하기 위해 사용된다. 사용자 인증, 암호화, 권한 부여, 세션 관리 및 감사 추적 등의 보안 기능을 포함하며, 주로 기업과 조직에서 비즈니스 프로세스와 관련된 중요한 데이터 및 시스템에 대한 접근을 제한하고 보호하는 데 사용된다.

<표 7> 액세스 관리 소프트웨어

SW [9]	Keycloak
장점	- 단일 로그온 및 ID 브로커링 제공 - LDAP 및 Active Directory와의 사용자 연합을 기본적으로 지원 - 중앙 집중식 관리 및 계정 관리 콘솔 - 표준 프로토콜 및 세부 인증 서비스 지원
단점	- 사용자 관리 및 다른 도구와의 통합에 일부 제한이 있을 수 있음

3.6. 데이터 디스커버리

데이터 디스커버리는 조직 내에 존재하는 데이터를 자동으로 인식하고, 수집하며, 분석하는 프로세스를 의미한다. 이를 통해 조직 내에서 데이터를 적극적으로 활용할 수 있으며, 이를 통해 다양한 분야에서, 더욱더 나은 의사 결정을 내릴 수 있다. 데이터 디스커버리는 데이터 분석이나 데이터 시각화와 같은 기술과 결합하여 더욱 효율적인 결과를 도출할 수 있다.

<표 8> 데이터 디스커버리 소프트웨어

SW[10]	ODD Data Discovery
검색 알고리즘	- 메타데이터 및 힌트가 포함된 텍스트 검색
자동화 메타데이터	- 계보 및 데이터 품질 정보 포함
협업	- 데이터 과학 팀과 데이터 엔지니어링 팀 간 협업 지원
경보	- 영향을 받는 엔터티에 대한 스마트 알림 제공
데이터 품질 보증	- 모든 데이터 품질 도구와 통합 가능
통합 데이터 카탈로그	- 엔드투엔드 데이터 검색 및 협업 지원

3.7. 머신 러닝 및 AI

데이터를 사용하여 패턴과 관계성을 찾아내고, 예측, 분류, 클러스터링 등의 작업을 수행하는 기술을 의미한다. 이를 통해 데이터 플랫폼에서 자동화된 결정 및 예측, 효율적인 데이터 분석 및 가치 창출이 가능하다.

MLflow는 머신러닝 라이프사이클을 관리하기 위한 오픈소스 플랫폼으로 실험, 재현성, 배포 및 중앙 모델 레지스트리를 포함한 머신러닝 라이프사이클의 모든 단계를 관리할 수 있다.[11]

4. 결론

데이터 플랫폼에 필수적으로 필요한 요소를 7개 층으로 제시하고, 초급 데이터 엔지니어에게 적합한 오픈소스 소프트웨어를 제안했다. 본 논문은 데이터 플랫폼 구축에 필요한 오픈소스 소프트웨어를 문헌 검토로 진행하였기에 제안된 7개의 계층에서 계층간 연계가 얼마나 용이한지에 대해 후속연구가 필요하다.

ACKNOWLEDGMENT

"본 연구는 과학기술정보통신부 및 정보통신기획평가원의 지역지능화혁신인재양성사업의 연구결과로 수행되었음" (IITP-2023-RS-2022-00156360)

참고문헌

[1] Miao, H., Chavan, A., & Deshpande, A.. "ProvDB: A System for Lifecycle Management of Collaborative Analysis Workflows." 2015, https://www.arxiv-vanity.com/papers/1610.04963/

[2] Miao, H., Li, A., Davis, L. S., & Deshpande, A. "ModelHub: Towards Unified Data and Lifecycle Management for Deep Learning." 2016, https://arxiv.org/abs/1611.06224

[3] V. Sridhar, S. Subramanian, D. Arteaga, S. Sundararaman, D. Roselli, and N. Talagala. "Model governance: Reducing the anarchy of production {ML}." In 2018 USENIX Annual Technical Conference (USENIX ATC 18), Boston, MA, 2018. USENIX Association. pages 351 - 358, https://www.usenix.org/system/files/conference/atc18/atc18-sridhar.pdf

[4] A. Pulkit et al., "Data Platform for Machine Learning." Proceedings of the 2019 International Conference on Management of Data. 2019. ACM Conferences., 2019, pages 1803 - 1816, https://dl.acm.org/doi/10.1145/3299869.3314050

[5] Pinho, D., Aguiar, A., & Amaral, "What about the usability in low-code platforms? A systematic literature review." *Journal of Computer Languages*, V. 74, 101185. 2023. https://doi.org/10.1016/j.cola.2022.101185

[6] Apache.org. "Apache Hadoop 3.3.5 - HDFS Architecture". 2013. https://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-hdfs/HdfsDesign.html

[7] Apache.org. "Apache SparkTM - Unified Engine for large-scale data analytics", 2018. https://spark.apache.org/

[8] KNIME. "Data Analytics Platform: Open Source Software Tools", 2023. https://www.knime.com/knime-analytics-platform

[9] Keycloak.org. "Keycloak", 2023. https://www.keycloak.org/

[10] Opendatadiscovery.org. "Open Data Discovery", 2023. https://opendatadiscovery.org/

[11] MLflow. "MLflow - A platform for the machine learning lifecycle", 2023. https://mlflow.org/