# SmartX Automation Framework를 위한 Apache Iceberg 기반 클라우드-네이티브 멀티모달 데이터 축적 파이프라인

임창현, 김종원*

광주과학기술원 AI융합학과

ich6648@smartx.kr, *jongwon@smartx.kr

# A Cloud-Native Multimodal Data Accumulation Pipeline based on Apache Iceberg for SmartX Automation Framework

ChangHyeon Im, JongWon Kim*

Department of AI Convergence, Gwangju Institute of Science and Technology (GIST)

요 약

본 논문은 SmartX Automation Framework의 지능형 Observability를 지원하기 위해 메타데이터 기반 클라우드-네이티브 멀티모달 데이터 축적 파이프라인을 제안하고, 프로토타입을 구현한다. 이렇게 구축된 데이터 레이크하우스는 신뢰성 있는 데이터 축적, 그리고 메타데이터 중심의 자동화된 관리와 가시성 확보를 핵심 목표로 한다. 제안된 파이프라인은 수집, 분류, 가공, 정리 네 단계로 구성되며 Apache Iceberg, Kafka, Kubernetes, Spark 등 오픈소스 기술을 통합 구현한다. 또한, 현실 세계에서 고품질의 대규모 멀티모달 데이터를 확보하기에 한계가 있어, NVIDIA Omniverse 시뮬레이션을 통해 데이터를 생성하였다. 생성된 데이터는 Kafka로 실시간 전송·분류되어 손실을 방지하고, 데이터 Branch에서 WAP(Write-Audit-Publish) 패턴을 적용해 데이터 검증과 Iceberg 구조화가 수행되며, Iceberg 메타데이터가 생성된다. 검증된 데이터는 Production 데이터 레이크하우스로 병합되고, Iceberg 메타데이터를 활용한 스토리지 정리와 BI Dashboard 연계를 통해 대규모 멀티모달 데이터의 안정적 축적과 직관적 분석을 지원한다. 이러한 파이프라인은 ArgoCD 기반 GitOps 방식을 적용하여 클라우드-네이티브 환경에서의 배포 및 운영 자동화를 확인하였다. 본 연구는 SmartX Automation Framework를 위한 데이터 플랫폼 구축의 초기 단계로서, 메타데이터 중심 데이터 레이크하우스를 통해 Observability의 기반을 마련한다. 향후 이 프로토타입을 발전시켜 SmartX Automation Framework 전반을 지원하는 확장형 데이터 관리 인프라로 발전시킬 계획이다.

## Ⅰ. 서 론

SmartX Automation Framework는 Observability, Orchestration, Provisioning을 핵심으로 하는 지능형 프레임워크로 시스템의 관측 능력을 확보하고 자원 관리 자동화를 목표로 한다. 특히, Observability는 단순한 모니터링을 넘어 시스템 내 발생하는 다양한 이벤트와 상태를 종합적으로 관측하고 문제의 근본 원인을 파악하여 조치하는 능력을 의미한다.[1] 이를 위해서는 방대한 멀티모달 데이터가 신뢰성 있게 축적되고 효율적으로 관리되어야 하며, 메타데이터에 기반한 가시성 확보가 필요하다.

AI 애플리케이션 개발에는 대규모 고품질 데이터가 필수적이지만, 실제 환경의 데이터 수집에는 한계가 있다. 이에 NVIDIA Omniverse와 같은 시뮬레이션 환경이 현실과 유사한 데이터를 생성하는 대안으로 주목받고 있다.[2] 다만 생성된 데이터가 무분별하게 쌓이면 데이터 스왐프로 전락할 수 있으며, 이를 방지하고, 데이터를 축적·관리하기 위해 클라우드-네이티브[3] 환경과 레이크하우스[4]를 결합한 아키텍처가 주목받고 있다.

본 논문은 SmartX Automation Framework의 Observability 지원을 위한 메타데이터 기반 데이터 레이크하우스 구축을 목표로, 클라우드-네이티브 멀티모달 데이터 축적 파이프라인을 제안하고 프로토타입을 구현한다. 또한, NVIDIA Omniverse 시뮬레이션 데이터를 생성하여 고품질 데이터의 안정적인 축적과 운영 자동화 및 가시성 확보를 검증한다.

## Ⅱ. 클라우드-네이티브 멀티모달 데이터 축적 파이프라인 설계

그림 1은 SmartX Automation Framework를 지원하는 클라우드-네이티브 멀티모달 데이터 축적 파이프라인의 전체 구성을 보여준다. 본 파이프라인은 NVIDIA Omniverse 시뮬레이션으로 생성된 데이터를 자동화된 방식으로 축적·관리하며, 수집, 분류, 가공, 정리의 네 단계로 구성된다.

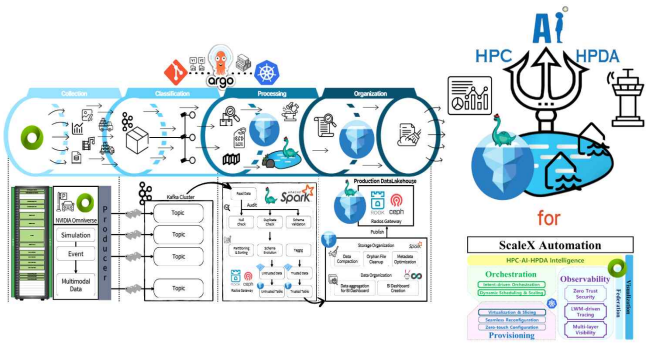


**그림 1. SmartX Automation Framework를 지원하는 클라우드-네이티브 멀티모달 데이터 축적 파이프라인**

수집 단계는 다양한 소스에서 생성되는 멀티모달 데이터를 Kafka로 전송한다. 이로써 가공·정리 단계에서 오류가 발생하더라도 데이터는 지속해서 Kafka에 저장되기 때문에 데이터 손실 위험을 최소화 할 수 있다.

분류 단계는 수집된 데이터를 이벤트 유형에 따라 Kafka의 각 토픽으로 분류한다. 분류된 데이터는 이후 Iceberg Table 생성 기준으로 활용된다.

가공 단계는 Spark 작업을 통해 Kafka 토픽에 대응하는 Iceberg Namespace(Production, Staging), Iceberg Table(Trusted, Untrusted), Nessie Dev Branch를 생성한다. Dev Branch에서 토픽 데이터를 읽어 Staging Namespace의 Iceberg Table(Trusted, Untrusted)에 저장하며, 품질 검증(널, 스키마 검사)과 Iceberg 구조화(스키마, 파티셔닝, 태깅) 작업을 수행한다. 이때 Iceberg는 데이터와 함께 메타데이터(스키마, 파티셔닝 정보, 스냅샷)를 자동으로 생성하고 관리한다. 이렇게 축적된 Iceberg 메타데이터는 데이터의 이력 관리, 변경 추적 등을 지원하여 지능형

Observability를 위한 핵심 기반을 제공한다. 검증을 통과한 신뢰 데이터(Trusted Table)는 Main Branch의 Production Namespace 내 Trusted Table로 원자적 병합되어 데이터의 신뢰성과 일관성을 확보한다.

정리 단계는 두 가지 측면으로 구성된다. 스토리지 정리는 Iceberg 메타데이터를 활용해 주기적으로 Orphan File 정리, Data Compaction, 메타데이터 최적화 작업을 수행해 스토리지 용량을 관리한다. 데이터 정리는 Trino와 Iceberg 메타데이터를 활용해 축적된 데이터 구조를 파악하고, 이를 Superset과 연동하여 BI Dashboard를 생성함으로써 데이터의 가시성을 확보하고 Observability를 지원한다.

마지막으로, 본 파이프라인은 ArgoCD를 활용해 GitOps 스타일의 버전 관리와 자동 배포로 클라우드-네이티브 운영의 일관성을 확보한다.

## Ⅲ. 클라우드-네이티브 데이터 축적 파이프라인 프로토타입

본 연구에서 제안한 파이프라인 프로토타입 구현은 GPU Node와 Compute Node로 구성된 이기종 클러스터 환경에서 수행되었으며, 주요 컴포넌트들은 ArgoCD와 Kubernetes를 활용해 자동으로 Pod 형태로 배포되어 파이프라인을 구성한다. 그림 2의 ArgoCD 대시보드를 통해 파이프라인 핵심 컴포넌트들이 정상적으로 배포된 상태임을 확인할 수 있다.

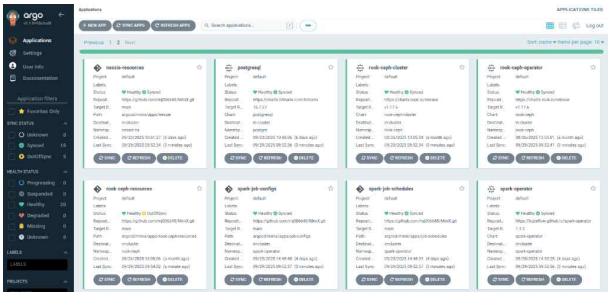

그림 2. ArgoCD 대시보드

그림 3은 NVIDIA Omniverse 환경에 구축된 주차장 시뮬레이션으로 차량의 입/출차 이벤트에 따라 이미지(입/출차 뷰), 텍스트(타임스탬프, 차량 고유 ID), 숫자(주차 좌표) 등 멀티모달 데이터가 생성된다.

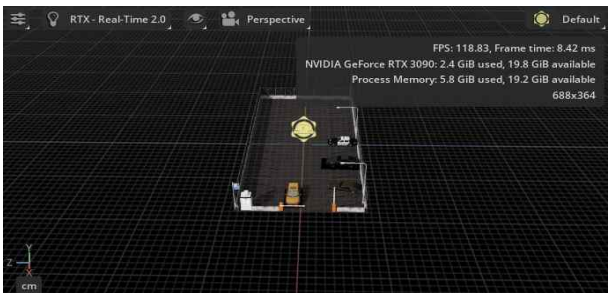


그림 3. Omniverse 주차장 시뮬레이션

Kafka에는 Parking-entry와 Parking-exits 두 개의 토픽이 존재하며, 각 토픽은 이벤트 정보를 담는 메시지(eventId, timeStamp, carId 등)로 구성된다. 그림 4의 Parking-entry와 같은 토픽에 메시지가 유입되면, 데이터 품질 검증과 Iceberg 구조화 과정을 거쳐 Iceberg 메타데이터가 생성되고, Production 레이크하우스에 축적된다. 이를 통해 데이터 사용자는 Trusted Table을 기반으로 신뢰성 있는 데이터만을 활용할 수 있다.

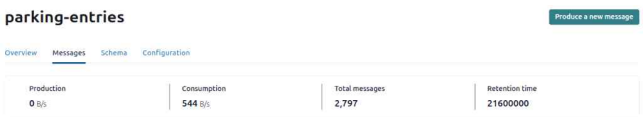


그림 4. Parking-entries Kafka 토픽

그림 5는 축적된 Iceberg 메타데이터를 Trino를 통해 파악한 후, Superset과 연동하여 구현한 BI Dashboard로, Parking-entries Table의 데이터 수를 시간에 따라 시각화한 결과를 보여준다. 이러한 대시보드는 데이터 파이프라인에서 축적된 데이터의 상태를 직관적으로 모니터링하고 분석할 수 있도록 지원하며, 시스템 전반의 상태를 파악하여 가시성을 확보해 지능형 Observability를 효과적으로 지원할 수 있다.

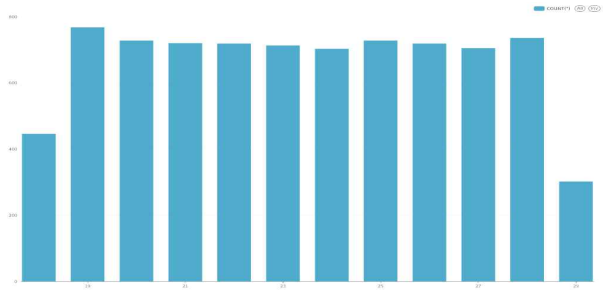

그림 5. Apache Superset BI Dashboard

## Ⅳ. 결론 및 향후 연구

본 논문에서는 SmartX Automation Framework를 위한 데이터 플랫폼 구축의 초기 단계로서, 메타데이터 기반 클라우드-네이티브 멀티모달 데이터 축적 파이프라인을 통해 지능형 Observability의 기반을 마련하였다. 제안된 파이프라인은 Apache Iceberg, Kafka, Kubernetes, Spark 등 오픈소스 기술을 통합 활용하고 WAP(Write-Audit-Publish) 패턴과 GitOps 기반 ArgoCD를 적용하여 안정적인 데이터 축적과 운영 자동화를 실현하였다. 특히, NVIDIA Omniverse 기반 주차장 시뮬레이션 데이터를 활용한 프로토타입 구축을 통해 데이터 품질 검증과 Iceberg 메타데이터 관리가 효과적으로 수행됨을 확인하였다. 또한, 축적된 데이터는 Trino와 Superset을 통해 시각화되어 Observability를 지원하는 가시성을 확보하였다. 향후 연구에서는 본 파이프라인을 기반으로 Iceberg 메타데이터를 활용한 멀티 워크로드 환경에서의 데이터 처리 최적화 및 Computing, Storage, Network 자원 오케스트레이션에 관한 연구를 진행할 예정이다. 이를 통해 멀티모달 데이터 특성에 맞춘 확장성 있는 데이터 관리 인프라를 구축하여 SmartX Automation Framework 전반을 지원하고자 한다.

## ACKNOWLEDGMENT

이 논문은 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원-대학ICT연구센터(ITRC)의 지원(IITP-2025-RS-2021-II211835)과 2019년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(No. 2019-0-01842, 인공지능대학원지원(광주과학기술원)).

## 참 고 문 헌

[1] J.-S. Shin and J. Kim, "SmartX multi-sec: a visibility-centric multi-tiered security framework for multi-site cloud-native edge clusters," IEEE Access, vol. 9, pp. 134208-134222, 2021.

[2] N. Ahmed, I. Afyouni, H. Dabool, and Z. Al Aghbari, "A systemic survey of the Omniverse platform and its applications in data generation, simulation and metaverse," Front. Comput. Sci., vol. 6, Art. no. 1423129, 2024.

[3] CNCF Glossary, "Cloud native technology," CNCF Glossary, https://glossary.cncf.io/ko/cloud-native-tech/ (accessed Oct. 3, 2025).

[4] A. A. Harby and F. Zulkernine, "Data lakehouse: a survey and experimental study," Information Systems, vol. 127, p. 102460, 2025.