시계열 파운데이션 모델을 활용한
실시간 PC 시스템 이상탐지 파이프라인 구축

Team 7

문세희 윤사무엘 이서진 이재찬 하주영

2025.12.09

DE Team Project Final Presentation

Problem Definition

- **실시간 PC 시스템 이상탐지**

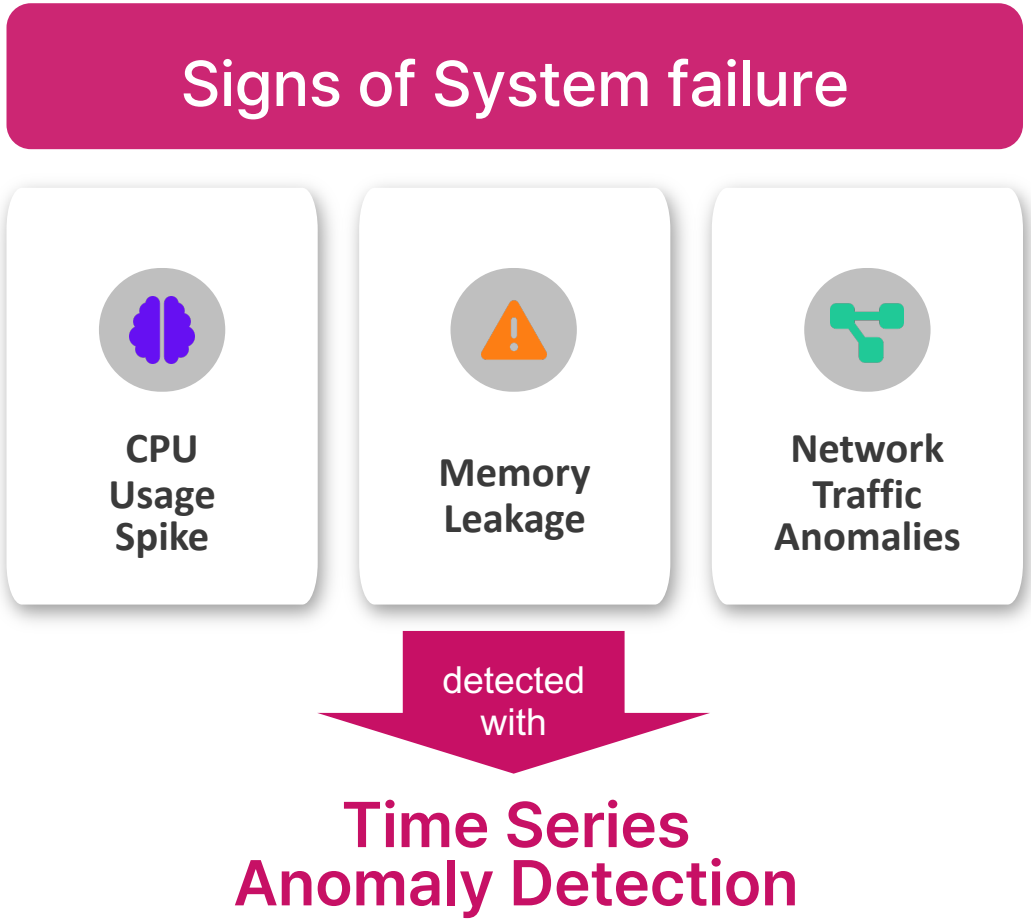
- **시스템 장애와 외부 공격의 징후**

  ✓ 개인 PC부터 기업 서버에 이르기까지, 시스템 장애나 외부 공격은 CPU 사용량 급증, 메모리 누수, 네트워크 트래픽 이상 등의 형태로 나타남

  ✓ 이러한 미세한 이상 징후를 조기에 자동으로 탐지하여 시스템 안정성을 확보하는 것은 매우 중요함

- **IT 인프라 운영의 핵심 과제**

  ✓ 본 프로젝트는 안정적인 IT 인프라 운영의 핵심인 “시스템 모니터링” 문제를 직접 해결하는 것으로, 산업 현장에서 중요하게 다루어지는 주제임



Problem Definition

Case 1



작업물 손실 방지

메모리 과부하로 프로그램이 중단되어
작업중이던 자료가 날아간 상황 (ex. PPT)

점진적 메모리 누수를 탐지하여
작업물의 손실을 미연에 방지함

Case 2



주기적 백업 실패 방지

주기적으로 작동해야 하는 사진 백업
프로그램이 특정 문제로 작동되지 않음

주기적 패턴이 소멸되는 현상을 감지하여
자동화된 백업 실패를 조기에 경보함

Case 3



보안 위협 조기 경고

해커의 공격으로 인해 PC 내부의 데이터가
네트워크를 통해 대량으로 유출됨

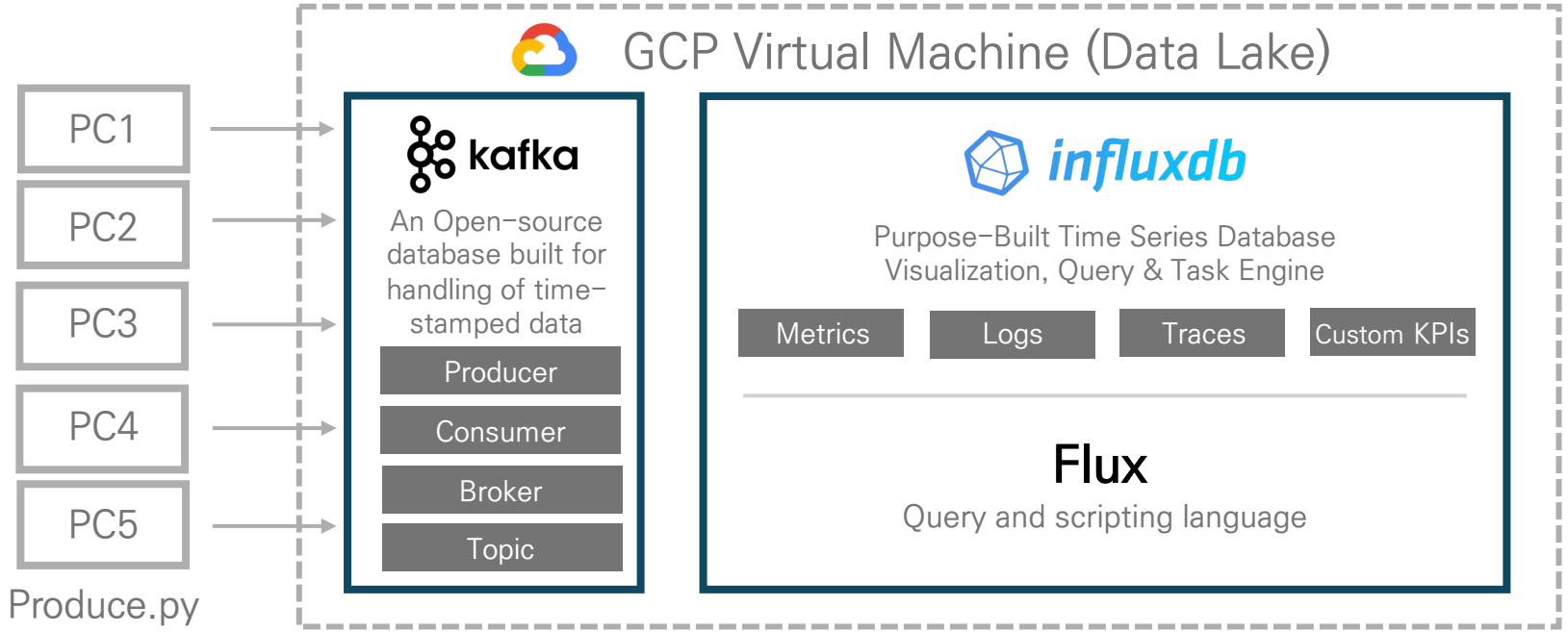
CPU 사용량에 비해 네트워크 트래픽이
과도하게 증가하는 '비정상적인 상관관계'
탐지, 잠재적 보안 위협 식별

Objectives

▪ End-to-End 데이터 엔지니어링 파이프라인 구축

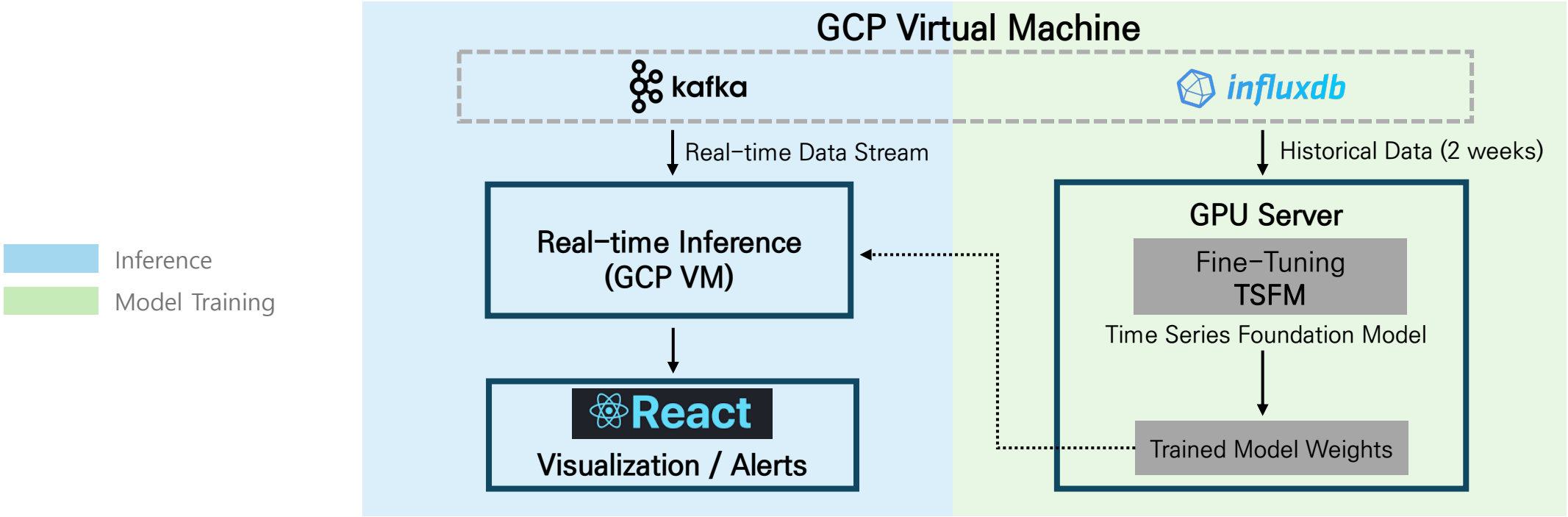
- **데이터 수집부터 사용자 알림까지**
  - ✓ 오픈소스 데이터셋이 아닌, 우리 팀이 직접 사용하는 PC에서 생성되는 실시간 스트리밍 데이터 사용 (CPU, 메모리, 네트워크)
  - ✓ 데이터 발생 지점부터 이상 탐지 결과가 대시보드에 시각화되고 사용자에게 경보가 발송되는 순간까지, 데이터 흐름의 전 과정을 직접 설계하고 구현
- **시계열 이상탐지 파운데이션 모델**
  - ✓ 고성능 시계열 파운데이션 모델을 시계열 이상탐지에 사용하여 문제 해결에 필요한 모델링 직접 수행

데이터 수집 & 저장 파이프라인



1. **데이터 수집 (Source)**
   - ✓ 팀원 로컬 PC 5대에서 producer.py를 통해 시스템 자원 데이터 수집
2. **데이터 전송 (Data Transmission)**
   - ✓ 수집된 데이터 스트림을 GCP VM에 구축된 Kafka로 전송
3. **실시간 처리 및 버퍼링 (Real-time processing & Buffering)**
   - ✓ 클라우드의 Kafka가 전송된 데이터를 받아 안정적으로 관리
4. **시계열 저장 (Time-series Storage & Query)**
   - ✓ Kafka에 대기하던 데이터를 InfluxDB 시계열 데이터베이스에 저장

모델 학습 & 실시간 추론 파이프라인



1. **학습 데이터셋 구축 (Training Dataset Preparation)**
   - ✓ InfluxDB에 2주간의 Historical Data를 학과 GPU 서버로 가져와 전처리 진행
2. **모델 학습 (Model Training)**
   - ✓ 연구실 GPU 서버를 활용하여 시계열 파운데이션 모델을 Fine-tuning
3. **실시간 데이터 스트리밍 (Real-time Data Streaming)**
   - ✓ GCP의 Kafka를 통해 새로운 시스템 데이터가 실시간 스트림 형태로 전달
4. **실시간 이상 탐지 (Real-time Anomaly Detection)**
   - ✓ 학습된 모델로 GCP VM 메모리에서 실시간 TSFM 추론 수행

Data Collection & Pre-processing

▪ 안정적인 모델 학습을 위한 데이터 수집 및 전처리 전략

- **데이터 수집 및 구성**
  - ✓ 팀원 5명의 PC에서 2주간 '정상' 및 '이상' PC 사용 데이터를 수집하였음 (11/05~11/18)
  - ✓ 사전에 정의된 이상 시나리오에 따라 인위적으로 이상 상황을 발생시켜 '이상' PC 데이터 확보 (이외에는 '정상'으로 간주)
- **데이터 스키마**
  - ✓ Tags: host_id, os
  - ✓ Fields: cpu_total_pct (float), mem_used_pct (float), disk_wbps (int), net_out_bps (int), missing_mask (int)
  - ✓ Time: ts(UTC)
- **데이터 전처리: 윈도잉/ 결측치 보간/ 스케일링**
  - ✓ 일정한 시퀀스 길이(512)와 stride(1)로 슬라이딩 윈도우 방식으로 학습 데이터를 생성하였음
  - ✓ 결측값은 학습 데이터에서 제외하였음
  - ✓ 각 PC 별로 이질적인 데이터를 효과적으로 처리하고, Outlier에 강건하도록 PC별 Robust Scaler (IQR 기반) 적용

x_scaled = (x − median) / IQR

Model Training

▪ 모델 학습 및 실험 방안

• 파운데이션 모델 학습

- ✓ 이상탐지 task 의 성능 및 오픈소스 편의성을 이유로 MOMENT 모델을 학습 모델로 선정하였음
- ✓ 2주간 수집한 '정상' 데이터를 사용하여 선정된 시계열 파운데이션 모델을 Fine-Tuning
- ✓ 모델은 PC 시스템 자원의 정상 패턴을 깊게 학습하여, 여기서 벗어나는 패턴을 '이상'으로 간주함
- ✓ 사전에 정의한 이상 시나리오를 탐지하기 위한 변수로 cpu 사용량(%), 메모리 사용량(%), 디스크 wbps, 네트워크 out bps를 사용함

• 이상 점수 정의

- ✓ **예측 오차 기반 접근:** 모델이 학습한 정상 패턴을 바탕으로 한 예측값과 실제 관측값 간의 차이를 이상 점수로 사용

$$\text{Anomaly Score}(t) = \text{Error}(\text{Predicted}(t), \text{Observed}(t))$$

• 모델 비교 및 개선

- ✓ 모델 개발 중 이상 패턴임에도 불구하고 모델이 쉽게 복원하여 이상 점수가 낮게 나오는 '과잉 일반화 문제' 가 발생하였음
- ✓ 학습 과정에서 정상 데이터의 패턴을 메모리에 저장 하는 '패치 기반 메모리 모듈'을 도입하여 모델을 개선하고, 해당 문제를 해결함

Anomaly Scenario

Use Case 1



작업물 손실

메모리 과부하로 프로그램이 중단되어
작업중이던 자료가 날아간 상황 (ex. PPT)

Scenario 1

점진적 메모리 누수

특정 프로세스의 버그로 인해 메모리 사용량이
눈에 띄지 않게 아주 조금씩
지속적으로 증가하는 상황을 재현

Contextual Anomaly
탐지 여부 검증

Use Case 2



주기적 백업 실패

주기적으로 작동해야 하는 사진 백업
프로그램이 특정 문제로 작동되지 않음

Scenario 2

주기적 패턴 소멸

15분마다 실행되던 백업 스크립트가 실패하여
평소 주기적으로 나타나던 디스크 쓰기 속도의
스파이크가 사라지는 상황을 재현

시간적 패턴 자체
학습 여부 검증

Use Case 3



보안 위협

해커의 공격으로 인해 PC 내부의 데이터가
네트워크를 통해 대량으로 유출됨

Scenario 3

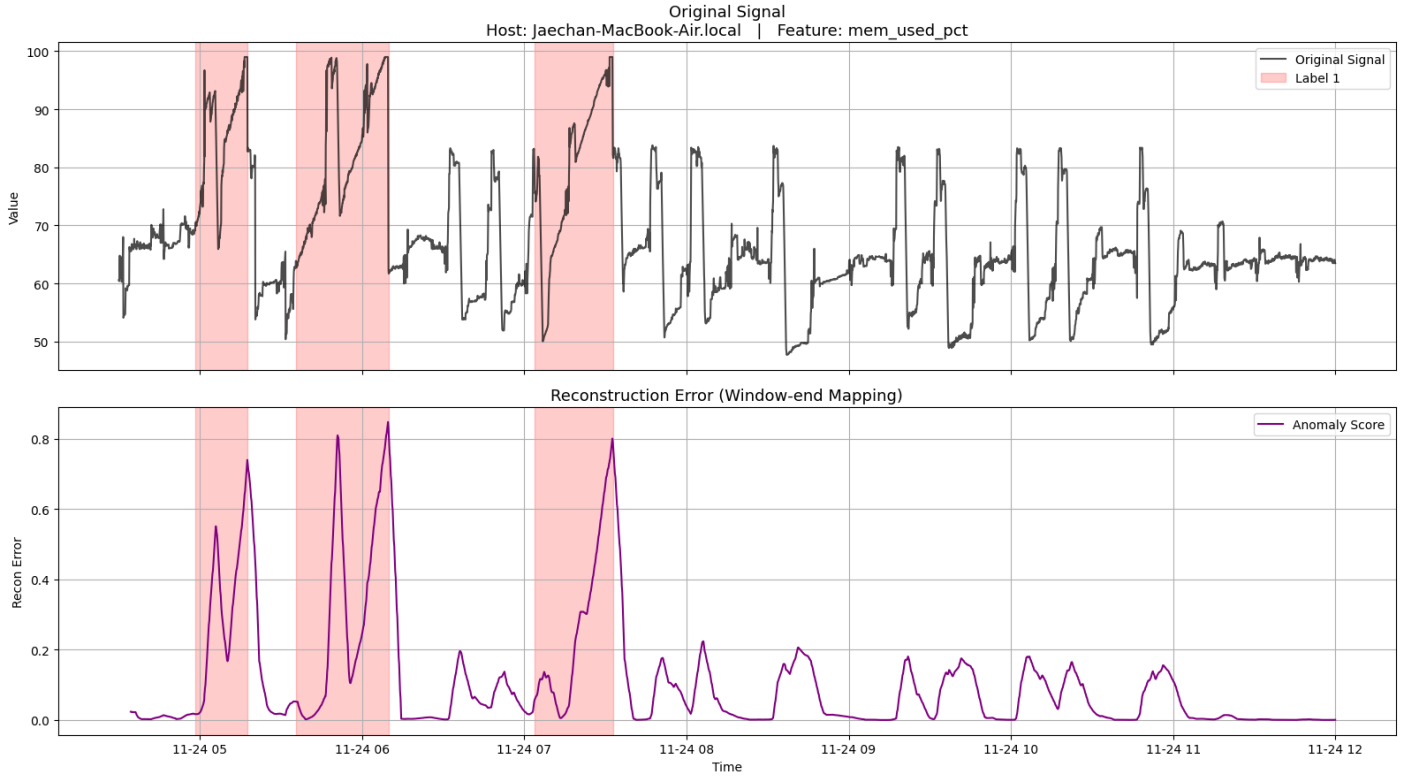
비정상적 상관관계 패턴

악성코드 감염 등으로 인해 CPU 사용량 변화
없이 네트워크 트래픽만 폭주하는 상황을 재현

지표 간의 상관관계 이상 발생 시
탐지 여부 검증

Results

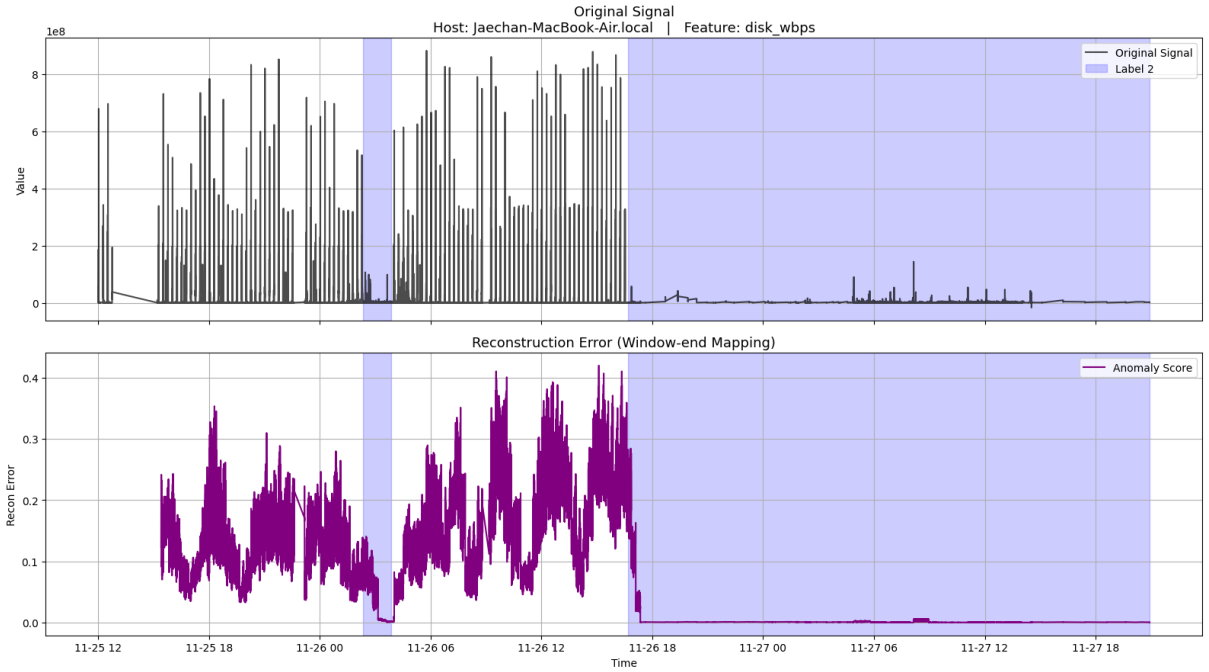
- **이상탐지 결과 (Scenario 1: 점진적 메모리 누수)**



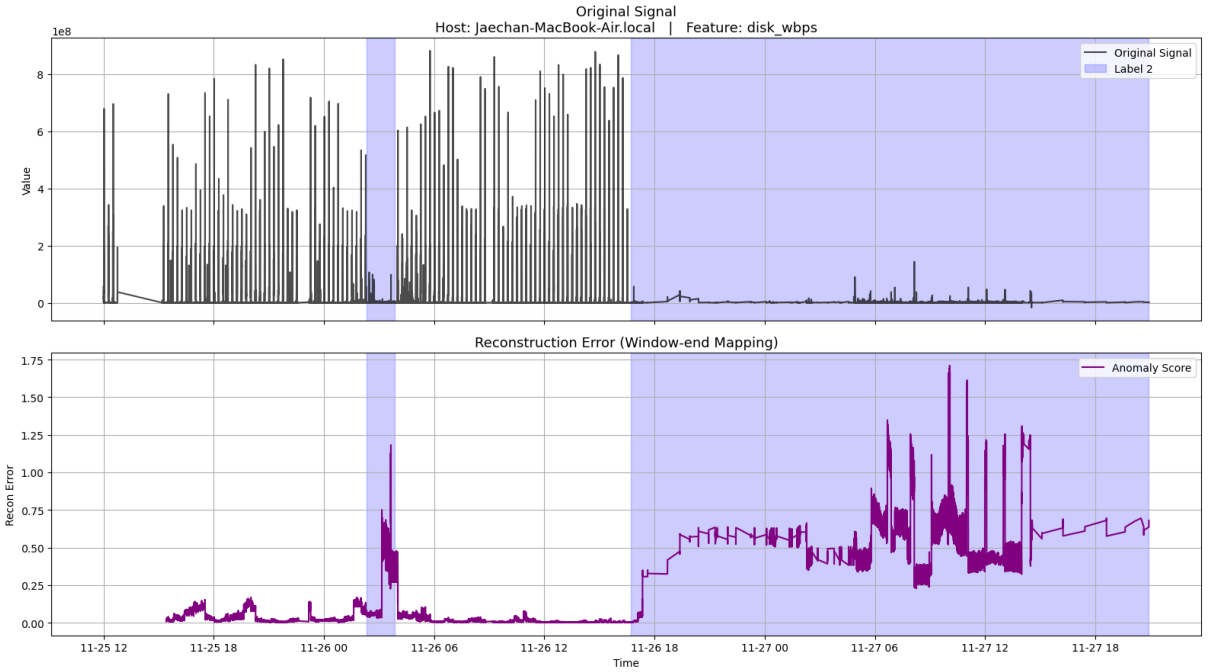
- 메모리 사용량의 절대적인 수치는 60% ~ 80% 사이로 정상 범위 내에 위치
- 하지만 데이터의 변동성 및 추세와 같은 **문맥(Context)이** 이전 구간과 확연히 다른 구간이 존재
- 단순 임계값 기반으로는 탐지할 수 없는 **Contextual Anomaly**를 시계열 파운데이션 모델이 문맥 학습을 통해 포착

Results

- **이상탐지 결과 (Scenario 2: 주기적 패턴 소멸)**



기존 (MOMENT)



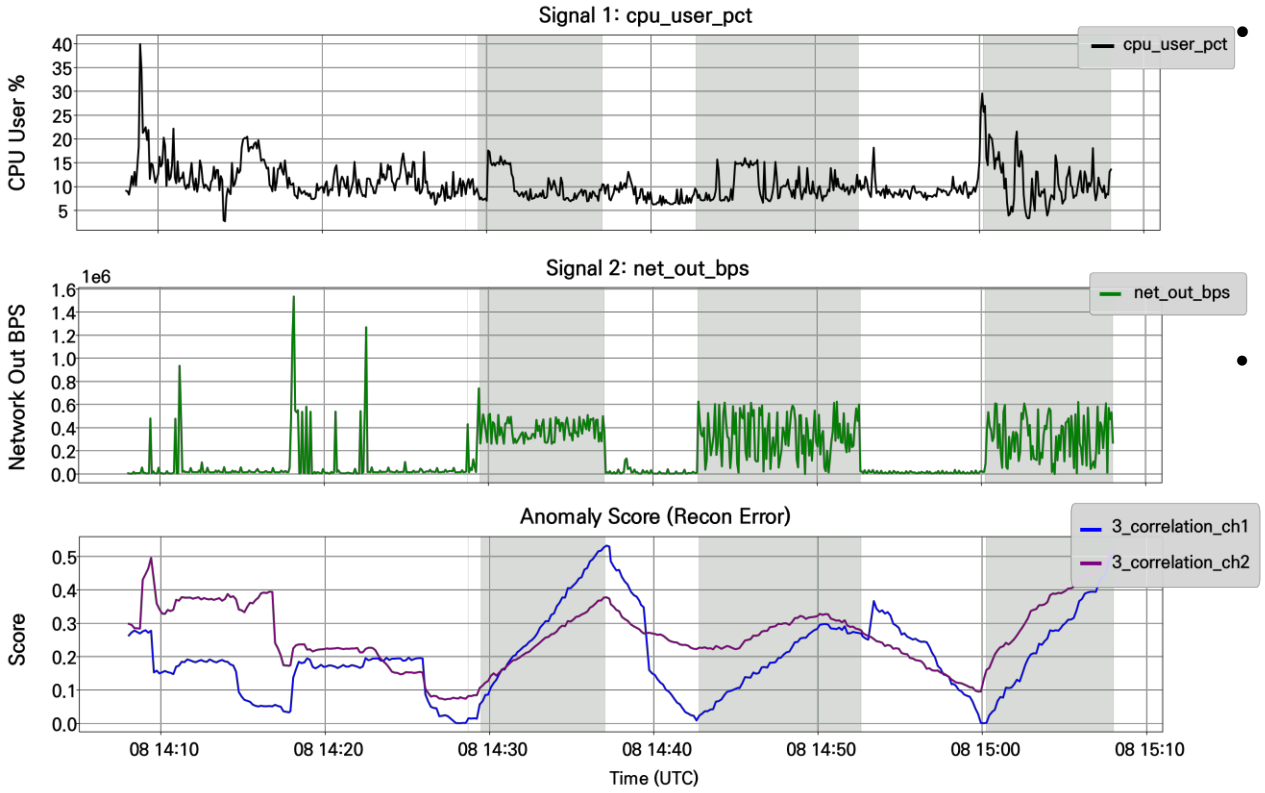
패치 기반 메모리 모듈 적용 후

- 기존 시계열 파운데이션 모델이 이상 데이터(0값)마저 복원하여 정상 데이터의 범주에 포함시키는 **Over-generalization** 문제 발생
- 학습 과정에서 정상 데이터의 패턴을 메모리에 저장하는 **Memory Module 도입**
- 모델이 '기억된 정상 패턴'만 참조하도록 유도하여, **패턴 소멸 시 Anomaly Score 증가**

Results

- **이상탐지 결과 (Scenario 3: 비정상적 상관관계 패턴)**

- 다변량 모델에서 Channel Dependency 학습 시도



- **단변량 모델**
  - 단일 변수 'net_out_bps'을 입력으로 활용
  - 시나리오 1과 마찬가지로, Contextual Anomaly 탐지 가능

- **다변량 모델**
  - 두 변수 'cpu_user_pct', 'net_out_bps'를 다변량 모델의 입력으로 활용
  - 두 변수 사이의 정상적인 **상호 의존성**을 학습
  - 변수 간의 상관관계가 깨지면, **Anomaly Score**가 증가하여 이상을 탐지

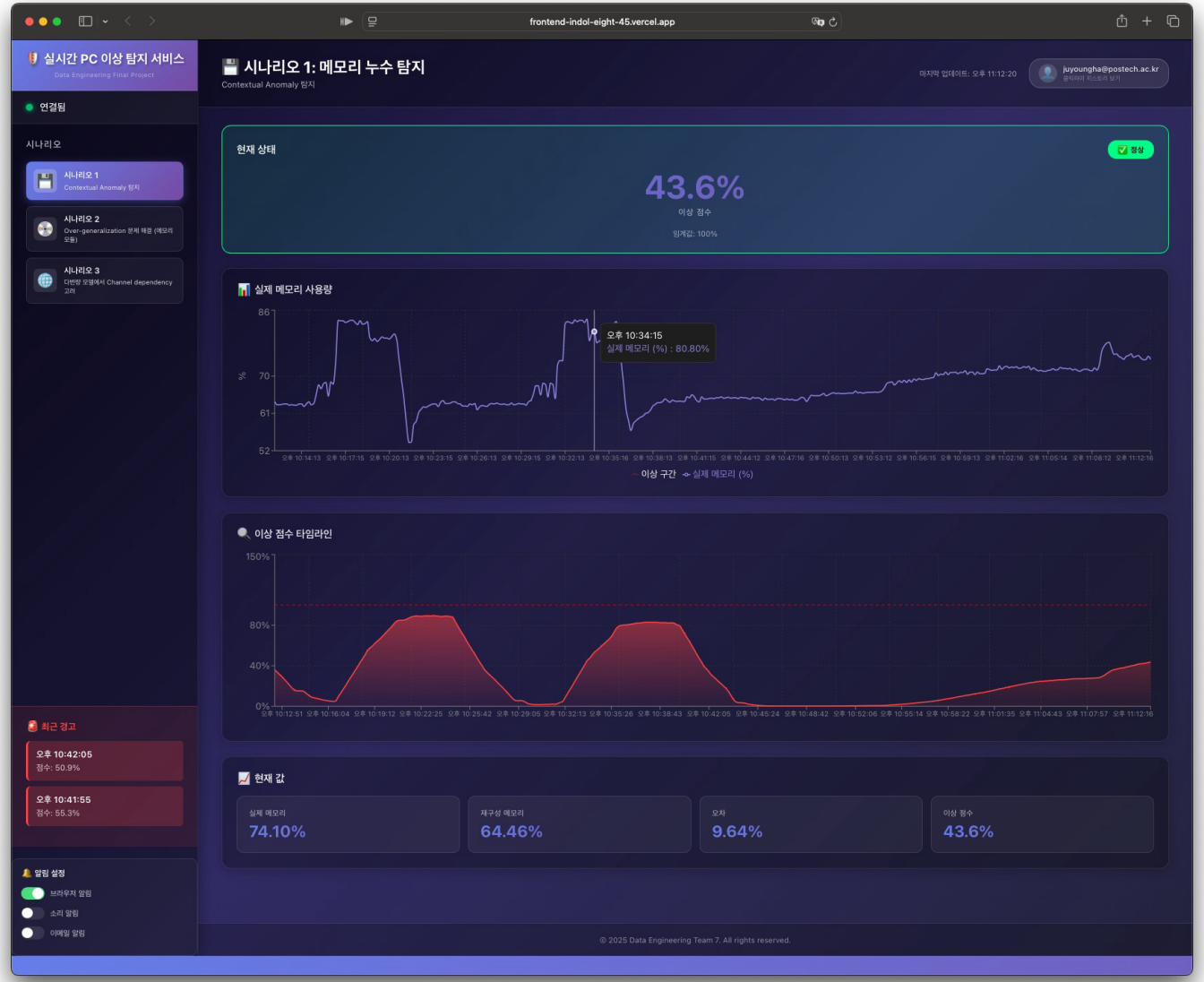
→ 다변량 모델에서도 동일하게 Anomaly 성능이 보장됨.

→ 하나의 모델로 모든 변수를 학습 및 추론하는 범용 모델로 발전 가능

Demonstration

- **핵심 기능**
  - 실시간 알림 기능 (via 메일 or 브라우저)
  - 모델 이상치 점수 모니터링
  - 사용자 별 trajectory 를 반영한 이상 탐지

- **구현 기술**
  - InfluxDB (백엔드) + React (프론트엔드)



Conclusion

▪ 본 프로젝트의 의의

- **End-to-End 데이터 파이프라인 구축**
  - ✓ 안정적인 데이터 수집 및 모델 학습&추론
- **이상탐지를 위한 TSFM 모델링**
  - ✓ Contextual Anomaly 탐지
  - ✓ Over-generalization 문제 해결 (메모리 모듈)
  - ✓ 다변량 모델에서 Channel dependency 고려함
- **PC Anomaly Detection 을 위한 Prototype 개발**
  - ✓ TSFM 이 유용하게 쓰일 수 있는 PC AD 라는 case를 발굴하고, PC util data 로부터 모델이 이상/정상 상황을 구별할 수 있음을 확인함

▪ Lessons Learned

- **데이터 수집의 어려움과 서비스 개발**
  - ✓ 데이터 수집 및 실시간 추론 과정에서 데이터를 안정적인 파이프라인으로 관리하는 것은 어렵고 중요한 문제임
  - ✓ 데이터 누락을 최소화하고 실시간 전송이 가능하게 하는 Kafka의 기능을 이해했고, fine-tuned TSFM 을 활용한 서비스 개발을 경험해볼 수 있었음
- **Foundation Model & “Back to Data”**
  - ✓ 다변량 FM이 단변량 FM보다 효과적이다 → 입증 못함
  - ✓ 주어진 데이터에서 가설 검증을 엄밀히 거치는 것이 모델링 노력보다 중요할 수 있음
  - ✓ 그러나, 범용 PC AD 모델을 개발하기 위한 다변량 모델링은 여전히 의미가 있음

Thank you

Thank you for listening to the presentation

Pohang University of Science and Technology

Data Engneering Team 7