KOREA SUPERCOMPUTING CONFERENCE 2024

KSC2024

ACCELERATING DISCOVERY:
HPC's Data-Driven Innovation

- WORKSHOP -

Database For AI

MACHBASE
김성진
CEO
㈜ 마크베이스

주최 KiSTi 한국과학기술정보연구원 Korea Institute of Science and Technology Information
주관 KiSTi 국가슈퍼컴퓨팅본부 KSF 한국초고성능컴퓨팅포럼 KSCSE 한국계산과학공학회
후원 과학기술정보통신부 Ministry of Science and ICT

• 기업소개

1

TPCx-IoT 세계 1위 TSDBMS

TPC.org 국제표준 DB 등재

"시계열 데이터베이스 엔진 및 솔루션 분야의 선두 기업"

· 회 사 명 : (주)마크베이스
· 설 립 일 : 2013. 3. 21
· 대표이사 : 김성진
· 주요사업 : 시계열 DBMS 개발 및 에지 컴퓨팅 솔루션 제공
· 본　　사 : 서울특별시 강남구 테헤란로 20길 10, 9층

• 주요 시장

스마트 기기 상용화, 데이터 수집량 증가, 컴퓨팅 장치 개량 및 소형화와 맞물려 시계열 센서 데이터 처리에 대한 특화

스마트 팩토리

- 스마트 팩토리 통신 부하 및 스토리지 자원 절감
- 불량 공정 해소 및 장애 방지
- 실시간 재고 관리

발전소

- 발전소 설비 가동 상태 실시간 감지 및 진단 용이
- 전기 비축 및 공급 효율화를 통한 발전소 운영 최적화

IoT 디바이스

- IoT gateway 서비스로 안정적인 IoT 서비스 제공
- 커넥티드 카 등 정밀한 제어를 요구하는 사물인터넷에 활용

자율주행차

- 차량 센서 데이터 실시간 처리
- 네트워크 단절 등 다양한 돌발상황에도 차량 거리 유지, 차량 제어 등 서비스 유지

스마트시티

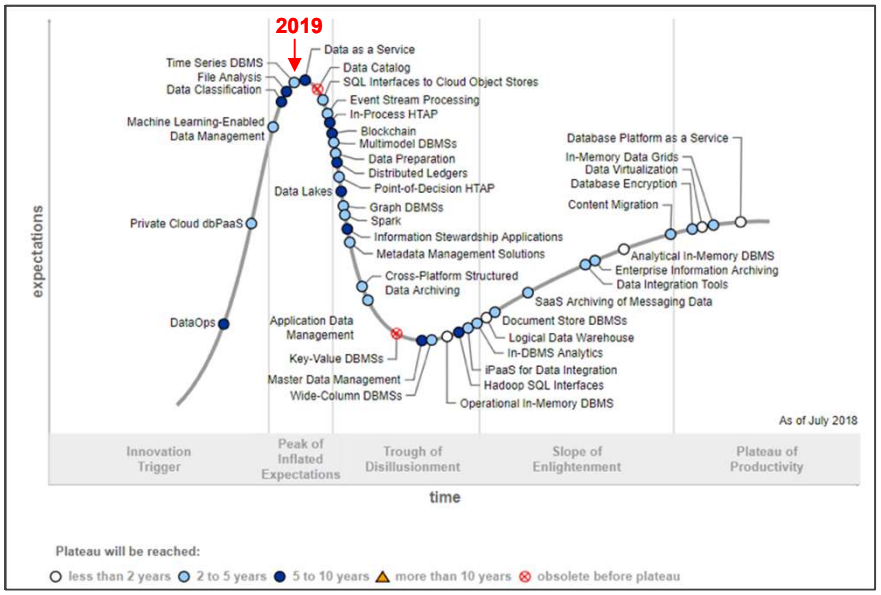
- 교통, 검침 등 지능형 인프라 운영 비용절감
- 저지연, 저손실, 고성능을 요구하는 시스템에 활용

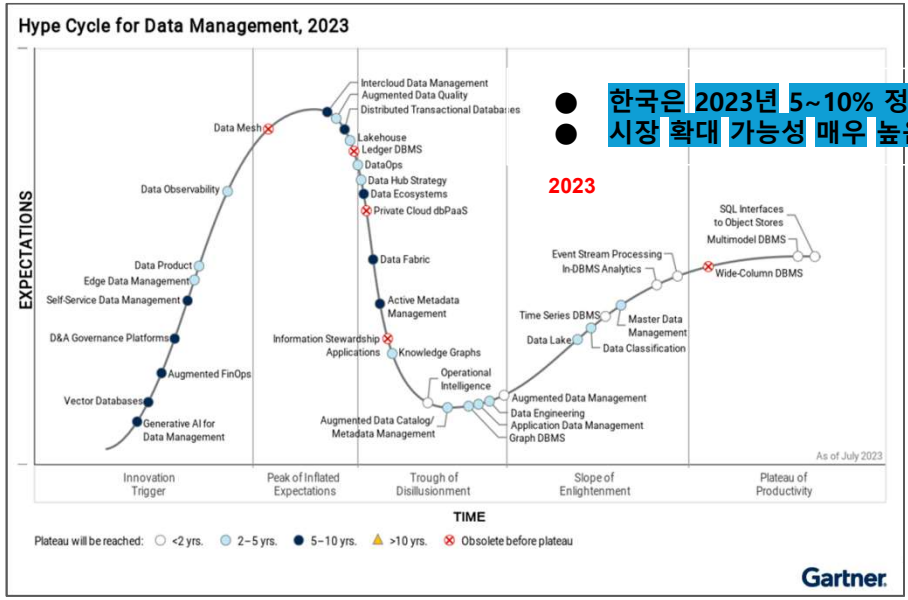
로봇

- 로봇의 자율성, 가용성 향상
- 데이터 기반 지능형 로봇 개발 촉진
- 로봇을 활용한 데이터 수집

• 가트너에서 바라본 시계열 DBMS 트렌드

- **2019 가트너 리포트**
  - ○ 시계열 DBMS는 전체 시장 고객의 1~5% 가 제품을 인지/채택하고 있음 (Adolescent)
- **2023 가트너 리포트**
  - ○ 시계열 DBMS는 전체 시장 고객의 **20~50%** 가 제품을 인지/채택하고 있음 (Early main stream)
- **글로벌 관점에서 Minor 시장 에서 Major 시장으로 이동중**





- 한국은 2023년 5~10% 정도로 추정
- 시장 확대 가능성 매우 높음

• TPC(국제공인성능평가기관) TPCx-IoT 성능평가 1위

▶IoT 부문 성능 평가

Rank	Company	System	Performance (IoTps)	Price/kIoTps	Watts/IoTps	System Availability	Database
1	TTA	Dell Power Edge R7615	5,739,514	86.42 USD	NR	02/28/23	Machbase 7.0.6
2	Alibaba.com	Lindorm	4,847,961	225.31 CNY	NR	05/19/22	Lindorm 3.4.10
3	TTA	Supermicro A Plus Server 1115SV-WTNR	4,529,397	54.85 USD	NR	09/18/23	Machbase V8.0.1 Cluster Edition
4	TTA	Supermicro A+ Server 1114S-WN10RT	3,410,800	88.78 USD	NR	03/17/21	Machbase 6.5.1
5	DELL Technologies	Dell Power Edge R7515	1,617,545	329.75 USD	NR	04/15/21	Cloudera HBase 2.2.3 on CDP 7.1.4

▶ IoT 부문 가격/성능 평가 (2023년 9월 1위 갱신!)

Rank	Company	System	Performance (IoTps)	Price/kIoTps	Watts/IoTps	System Availability	Database
1	TTA	Supermicro A Plus Server 1115SV-WTNR	4,529,397	54.85 USD	NR	09/18/23	Machbase V8.0.1 Cluster Edition
2	TTA	Dell Power Edge R7615	5,739,514	86.42 USD	NR	02/28/23	Machbase 7.0.6
3	TTA	Supermicro A+ Server 1114S-WN10RT	3,410,800	88.78 USD	NR	03/17/21	Machbase 6.5.1
4	Alibaba.com	Lindorm	4,847,961	225.31 CNY	NR	05/19/22	Lindorm 3.4.10
5	DELL Technologies	Dell Power Edge R7515	1,617,545	329.75 USD	NR	04/15/21	Cloudera HBase 2.2.3 on CDP 7.1.4

▶ IoT 부문 성능 평가

- 평가기준 : 성능 지표(IoTps)
- 평가결과 : 마크베이스는 22년 12월 574만 IoTps로 성능평가 1위 달성(2위 알리바바 485만 IoTps)

▶ IoT 부문 가격/성능 평가

- 평가기준 : 가격/성능 지표(Price/KIoTps)
- 평가결과 : 마크베이스는 23년 9월 $54.85로 가격/성능평가 1위 달성(종전의 $86.42에서 38% 비용절감 실현)

주요 고객


SAMSUNG


현대중공업


HYUNDAI
GLOBAL SERVICE


LS ELECTRIC


LS 전선


한화오션
Hanwha Ocean


HD현대에너지솔루션


Chong Kun Dang
Pharm.
Seoul,Korea


WOOJIN
Industrial Systems Co., Ltd.


Hansol
한솔제지


Kolmar


LOTTE
CHILSUNG BEVERAGE


Carrot
캐롯손해보험


SAMPYO
Cement


KoROAD
도로교통공단


국가철도공단
KOREA NATIONAL RAILWAY


SEOUL METROPOLITAN
GOVERNMENT


Seoul
Metro


Busan
Transportation Corporation


KEPCO
KOREA ELECTRIC POWER CORPORATION


KITECH
한국생산기술연구원
Korea Institute of Industrial Technology


TTA
한국정보통신기술협회
Telecommunications Technology Association


Mandobrose


SMIC

• 디지털 전환과 AI 기술 활용

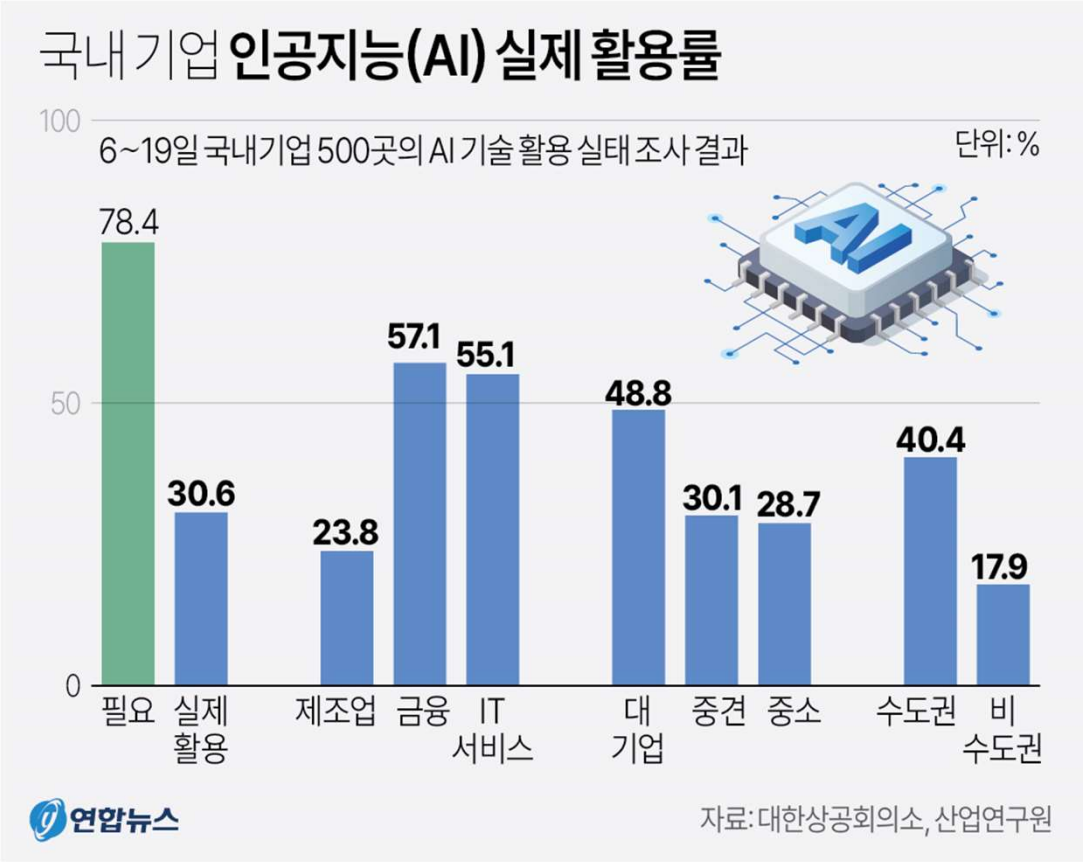
- 데이터를 인사이트로 전환
- 예측적 유지보수 강화
- 생산 프로젝트 최적화
- 적시(JIT) 제조 실현
- 지속적 개선 추진

아날로그 유산의 비효율적이고
느린 비지니스 형태

신속한 디지털로의
전환 실현

빠르고 정확한
의사 결정, 이상 감지,
예지 보전

• AI 필요성



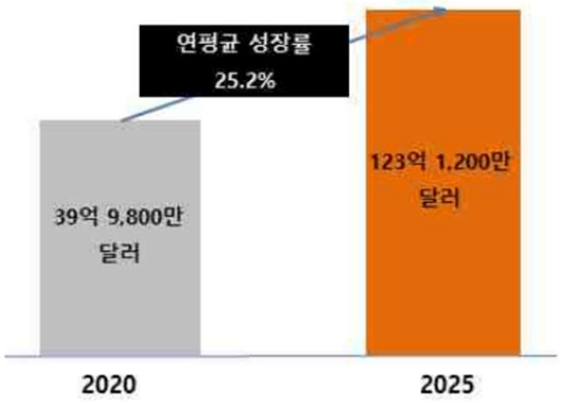
박영석 기자 20240828

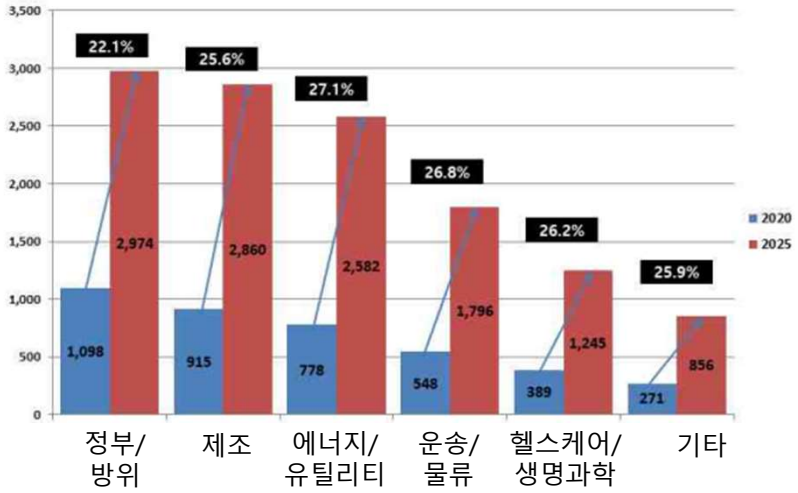
◆ 대한상공회의소, 산업연구원(2024.08)
- 국내 500개 기업 대상
- 기업의 **생산성 제고, 비용절감 등 성과향상을 위해 AI 기술이 필요(78.4%)**
- 실제 제조업 활용율은 23.8%로 저조

◆ 활용 분야
- 제품 개발(66.7%)
- 보안/데이터분석등 IT 업무(33.3%)
- 품질 및 생산관리(22.2%)
- 공급망 관리(9.8%)

◆ AI 기술을 활용하지 못하는 이유
- **기술 및 IT 인프라 부족(34.6%)**
- 비용 부담(23.1%)
- 신뢰성에 대한 의문(10%)

• 예지보전 시장 전망





◆ 전 세계 예지보전 시장 전망

✔ 2025년 123억 1,200만 달러

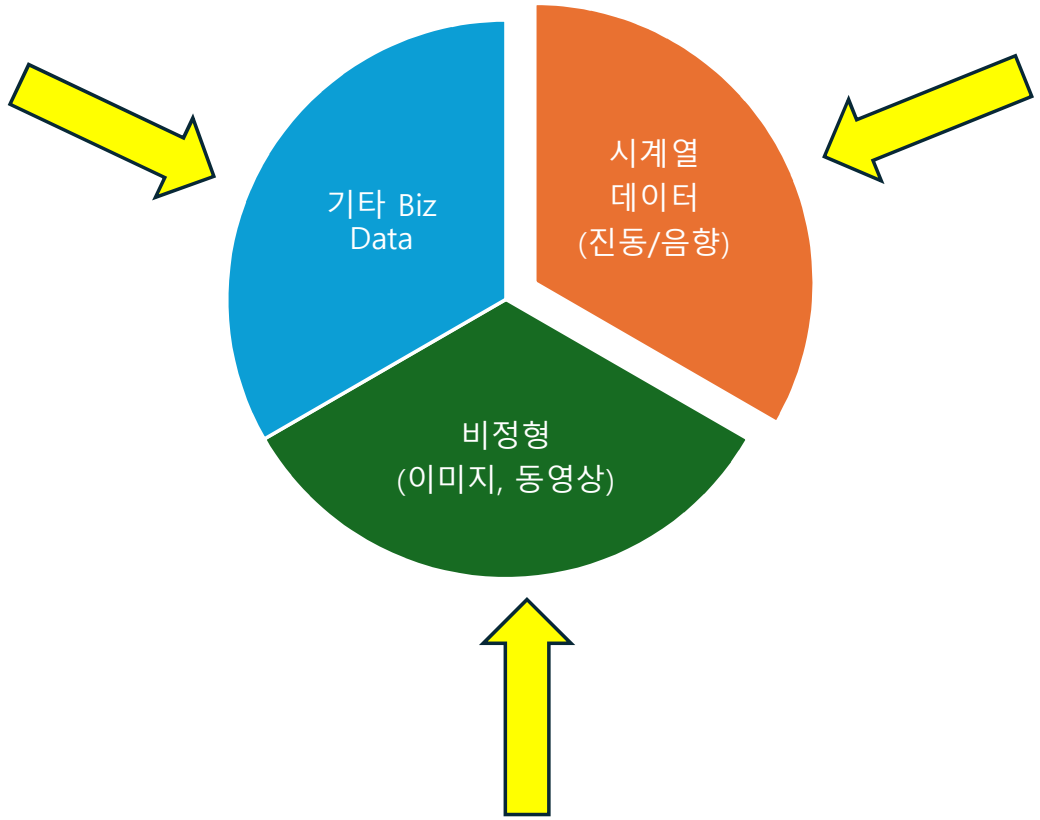
◆ 제조 부문 예지보전 시장 전망

✔ 2025년 28억 6,000만 달러

※ 출처 : MarketsandMarkets, Predictive Maintenance Market, 2020

• AI Data for OT

RDBMS

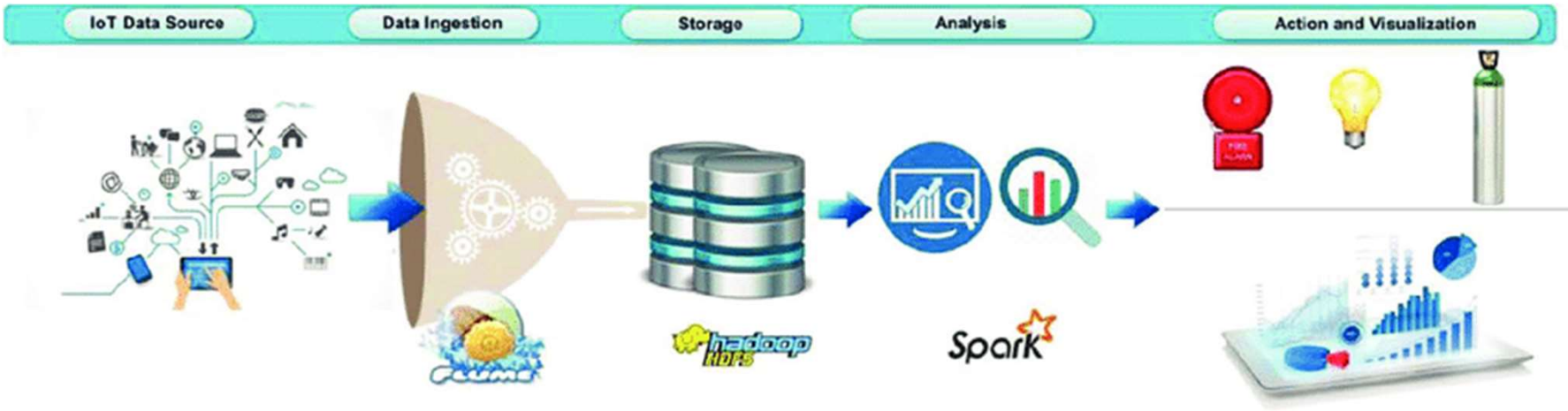


???

Problem in AI Training

Hadoop (SAN, NAS..)

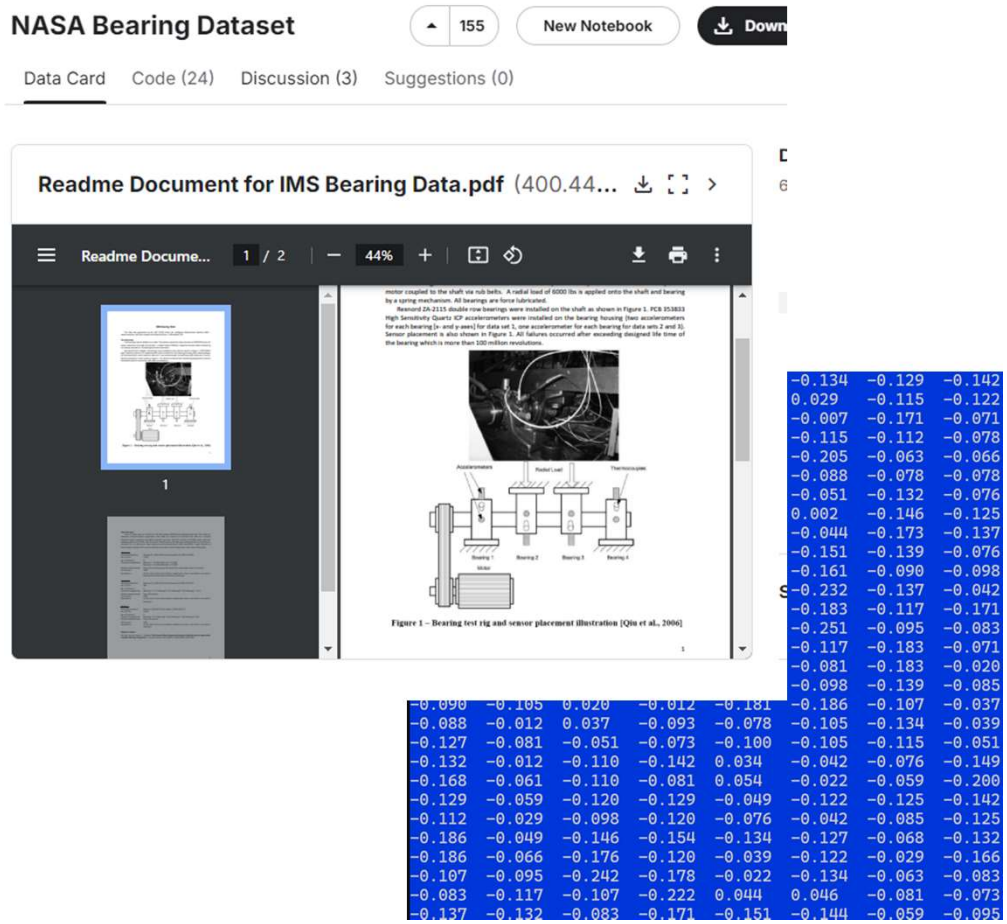
• OT AI 데이터의 본질적인 문제



◆ 대량의 데이터 처리 기술 한계

- **저장 포맷 상이**
- **대용량 실시간 추출 불가능**
- **분석 시각화 난제**
- **높은 데이터 조작 시간/비용**
- **데이터 개인화 불가능**

• OT AI 구성과 문제점 – 기존 접근법



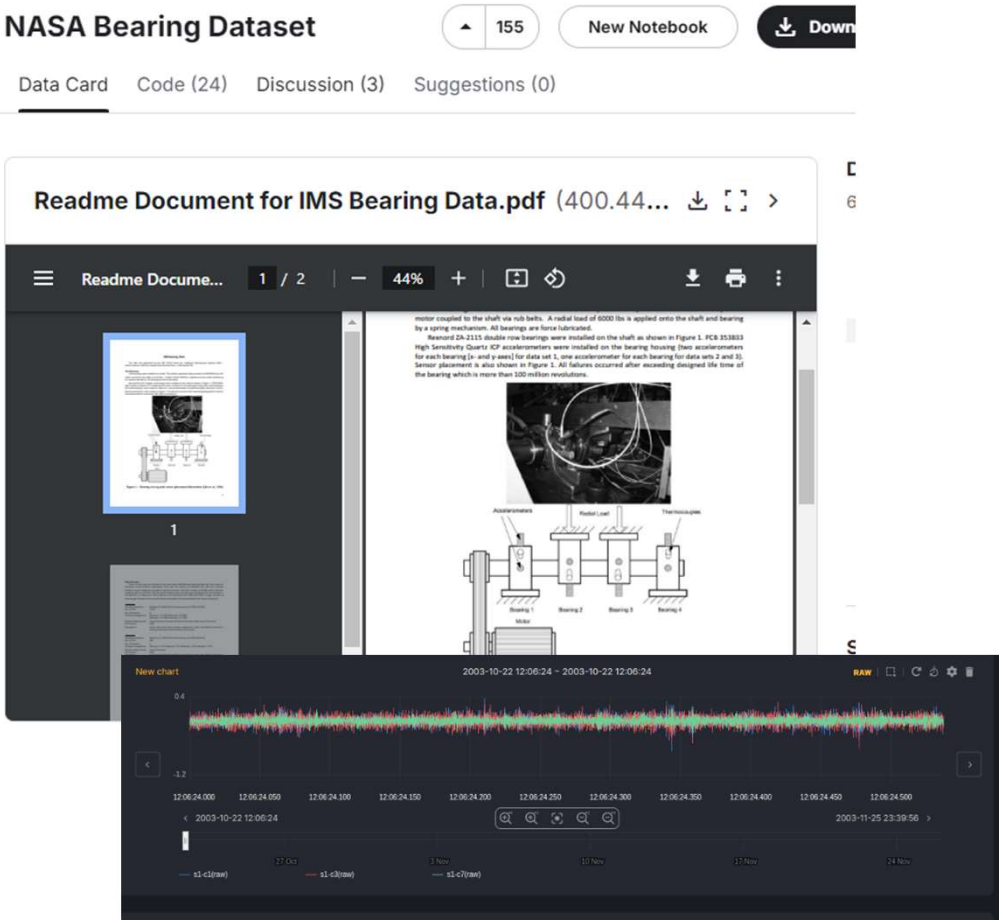
◆ Hadoop 계열(파일 기반) 데이터 처리 문제

- 대량의 데이터를 CSV 형태로 hdfs에 저장
- 모든 분석 App은 CSV 변환 및 로딩 프로그램 작성 필요
- 데이터의 형식 (시간 데이터의 유무, 칼럼의 개수 등)이 모두 달라 모든 App 개발시 데이터 로딩 각각 수행

◆ 대용량 데이터 처리 문제

- AI 모델 학습시 읽어들이는 데이터는 학습 장비의 메모리보다 더 클 수 없음
- 고주파 진동 데이터의 경우, 장기간의 데이터를 대상으로 학습시 데이터 로딩 문제가 발생

• 고주파 진동 데이터 케이스



◆ AI 학습시 요구 사항
- 대량의 데이터를 이용하여 학습해야 더 정확한 AI 모델이 생성됨

◆ 고주파 진동 데이터 분석 시 해결해야 할 사항
- 수십 KHz 단위의 진동 데이터 수집시 막대한 데이터 발생
- 데이터 시각화 분석시 데이터 양에 따른 분석 문제 발생
- 대량의 데이터에서 원하는 데이터의 검색 문제
- AI 학습모델 생성시 학습 프로그램의 메모리 문제 발생

◆ 즉, 시계열 DBMS를 이용한 분석 및 AI 학습 방법의 고안이 필요함

• 데이터 구조 표준화 필요

시간	온도	습도	압력	진동
2023-04-15 09:34:12	23.5	78.9	11	55
2023-04-15 09:34:13	23.7	75.6	12	51
2023-04-15 09:34:14	23.5	78.9	13	44
2023-04-15 09:34:15	23.7	75.6	14	47

TAGID	시간	값
온도	2023-04-15 09:34:12	23.5
습도	2023-04-15 09:34:12	78.9
압력	2023-04-15 09:34:12	11
진동	2023-04-15 09:34:12	55
온도	2023-04-15 09:34:13	23.7
습도	2023-04-15 09:34:13	75.6

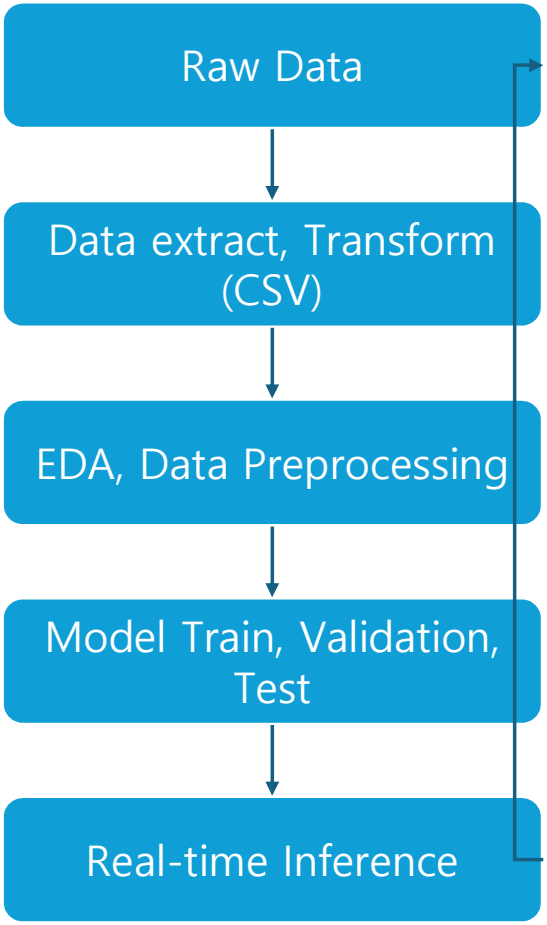
◆ 전통적 데이터 형태(csv 형태)

- 동일한 시간에 측정한 값을 하나의 레코드로 관리
- 원본 데이터 형태와 동일하게 조회 가능
- 파일마다 구조가 다르며 칼럼 개수가 달라, 스키마 유연성이 없음

◆ 마크베이스 데이터 모델

- 측정값을 하나의 레코드로 변환해서 저장
- TAG 컬럼 추가/삭제에 대한 유연성 극대화
- 개별 태그에 대한 별도의 데이터 집계 및 통계 연산 가능
- 데이터 건수 증가하나, 처리 성능은 더 빨라짐
- 같은 종류의 데이터가 같은 tag로 묶이므로 압축 효율이 증가

• 일반적인 AI 개발 프로세스



◆ AI 학습 및 추론 과정

- 데이터 확인 후, 원하는 데이터를 선택하여 이를 추출
- 원본 데이터를 별도의 시각화 도구를 이용하여 확인하고, 통계치를 추출하고 원하는 feature를 추출하기 위해 통계적 기법등을 적용
- RMS, Stddev, FFT 분석 기법등으로 데이터 전처리
- AI 신경망 모델 생성, 학습, 검증 및 테스트 과정 진행
- 최적의 AI 모델을 생성하기 위해 **데이터 추출부터 학습까지의 과정을 반복** 수행

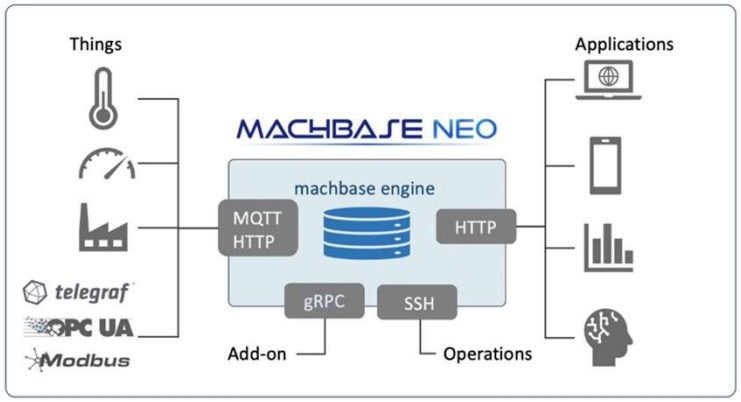
◆ 좋은 AI 모델은 ?

- **누가 더 빨리, 더 많이 반복 작업을 할 수 있느냐의 문제**

• 예지 보전 AI 응용 개발을 위한 요구 사항

- ◆ **원시 데이터의 표준화**
  - AI 응용 프로그램의 표준화 →개발 공수 최소화
- ◆ **초거대 데이터 저장소 제공**
  - CSV 지옥에서 벗어날 수 있는 방법 제공
- ◆ **표준화된 초고속 데이터 추출 성능 제공**
  - 데이터 종류에 관계 없이 동일한 인터페이스 제공
- ◆ **실시간 데이터 변환/정련 (Cleansing)**
  - 데이터 처리에 대한 시간/비용 최소화
- ◆ **AI 프레임워크와의 완벽한 호환**
  - 기존 코드 재활용 극대화
- ◆ **실시간 데이터 시각화를 통한 데이터 분석**
  - 학습을 위한 데이터 선택/분석/검증/테스트를 한번에
- ◆ **적정 HW를 통한 AI 개발자의 환경 개인화**
  - 고가의 데이터 처리 HW 투자 불필요

• Machbase의 데이터 처리 및 AI 개발 모델

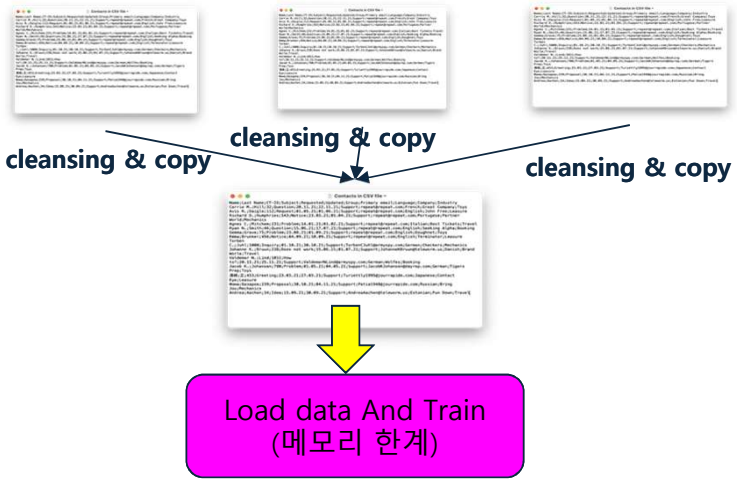


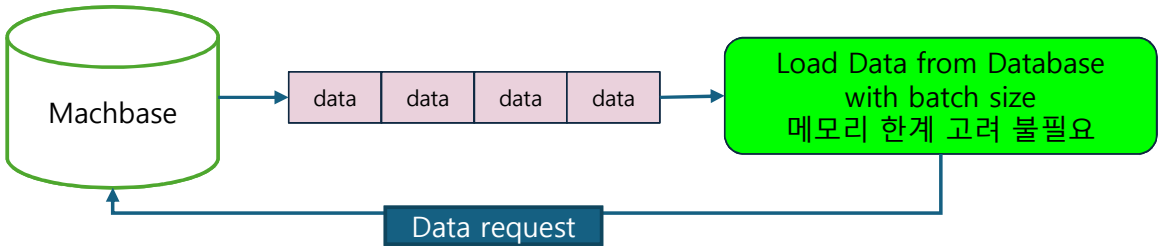
Machbase DATAHUB API

AI Code

- ◆ Machbase Neo
  - 다양한 센서 프로토콜(OPC, Modbus) 및 Edge 활용으로 데이터 수집 문제 해결
  - 데이터 관리 및 시각화 및 Datahub API 지원
- ◆ **DATAHUB API**
  - Jupyter notebook, Python등의 AI Code에서 원하는 데이터만 지정하면 바로 개발가능
  - TQL 언어로 개발하여 사용자 요구에 따른 추가 개발 및 Customizing 가능
- ◆ AI Code
  - 기존에 알려진 데이터에 대한 샘플 코드 제공
  - 데이터 형태에 독립적인 코드 작성 가능하여 재사용성 높임

CSV vs TSDB





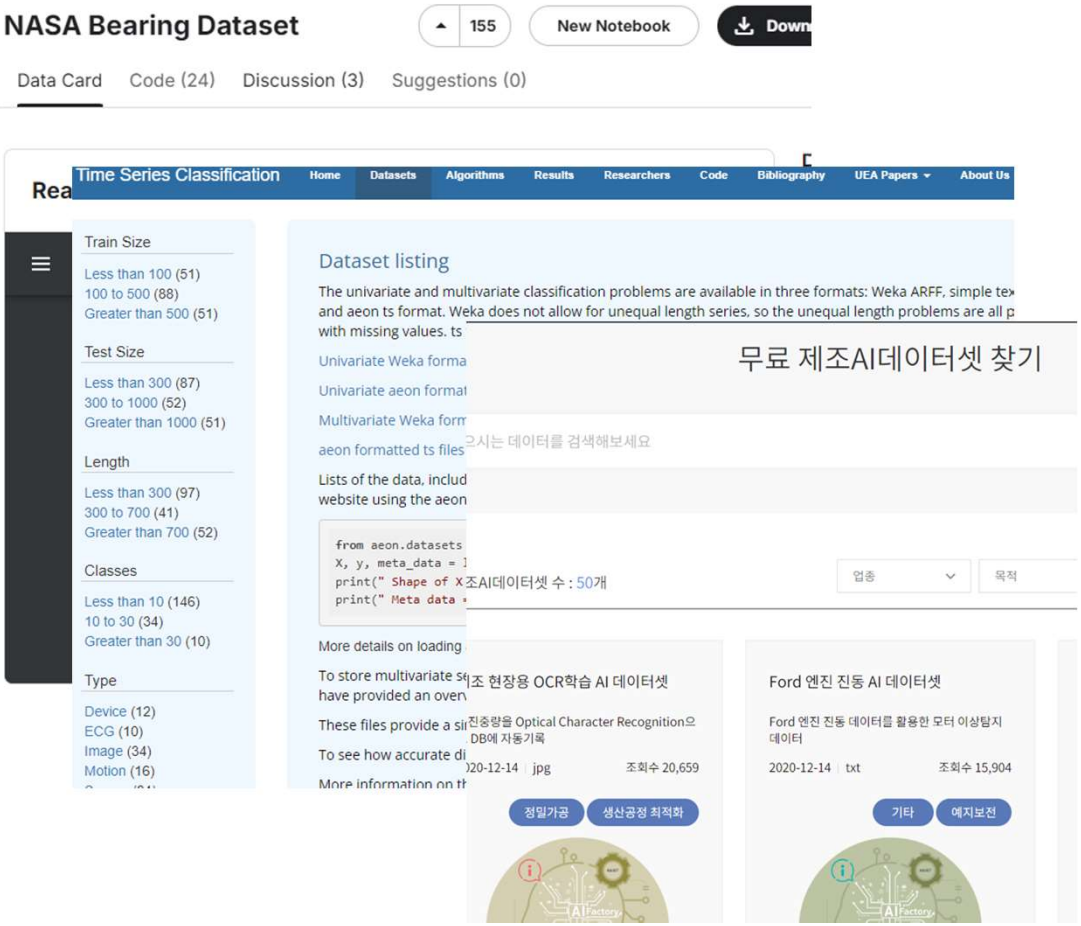
◆ 기존 csv 기반의 학습 프로그램

- 사전에 대량의 csv파일에서 원하는 데이터를 확보하기 위해 전처리 필요
- 데이터 변환등의 프로그램은 csv 파일마다 다른 프로그램의 작성이 필요
- 생성한 csv 파일의 크기가 사용 가능한 메모리보다 큰 경우 데이터 확보 과정을 재실행

◆ Machbase의 batch 기반 데이터 로딩 방법

- 데이터를 batch 단위로 database에 요청
- 수신한 데이터를 기반으로 AI 학습 수행
- 다음 batch 데이터를 database에 요청
- 항상 가용 메모리보다 적은 크기로 데이터를 요청하므로 메모리 문제가 없음
- 표준화된 포멧으로 데이터가 저장되어 있으므로 학습프로그램을 표준화 할 수 있음

• Machbase Datahub의 필요성



◆ 기존 AIoT 데이터셋 문제점

- Kaggle, KAMP 등에서 다양한 데이터 제공
- 데이터가 표준화 되어 있지 않음
- 소량 데이터셋만 제공
- 데이터가 어떻게 생겼는지 알 수 없음
- 데이터 가공을 처음부터 직접 개발자가 해야 함
- 데이터 따로, 분석 코드 따로

◆ 필요성

- 데이터 표준화
- 대량의 데이터 이용 가능해야 함
- AI 데이터 접근 코드의 표준화
- 데이터 개인화
- **거대 데이터도 학습 가능하게!**
- **AI 개발자도 데이터를 직접 다룰 수 있게!**

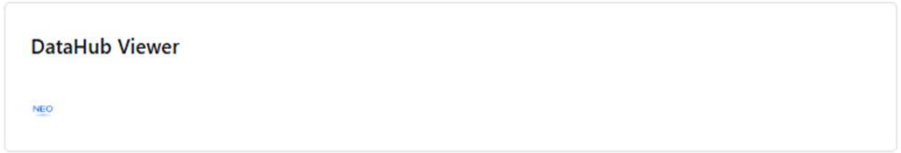
• Machbase Datahub 소개 (https://datahub.machbase.com)

KOREA SUPERCOMPUTING CONFERENCE 2024
KSC2024

2. Data Visualization with Machbase Neo

- Data visualization is possible through the Tag Analyzer in Machbase Neo.
- Select desired tag names and visualize them in various types of graphs.

- Below, access the 2024-3 DataHub in real-time, select the desired tag names from the data of 16 tags, visualize them, and preview the data patterns.



◆ 시계열 데이터의 AI 학습 관련 데이터 제공

- 시계열 데이터를 Machbase Neo와 추가 API를 이용하여 무료로 즉시 활용
- 온라인 Machbase Neo에 접속하여 데이터 프리뷰 가능
- 표준화된 데이터셋과 Machbase Neo TSDB를 이용한 가시화, 데이터 관리가 가능
- 쉽고 빠르게 데이터 적재, 활용

◆ 특장점

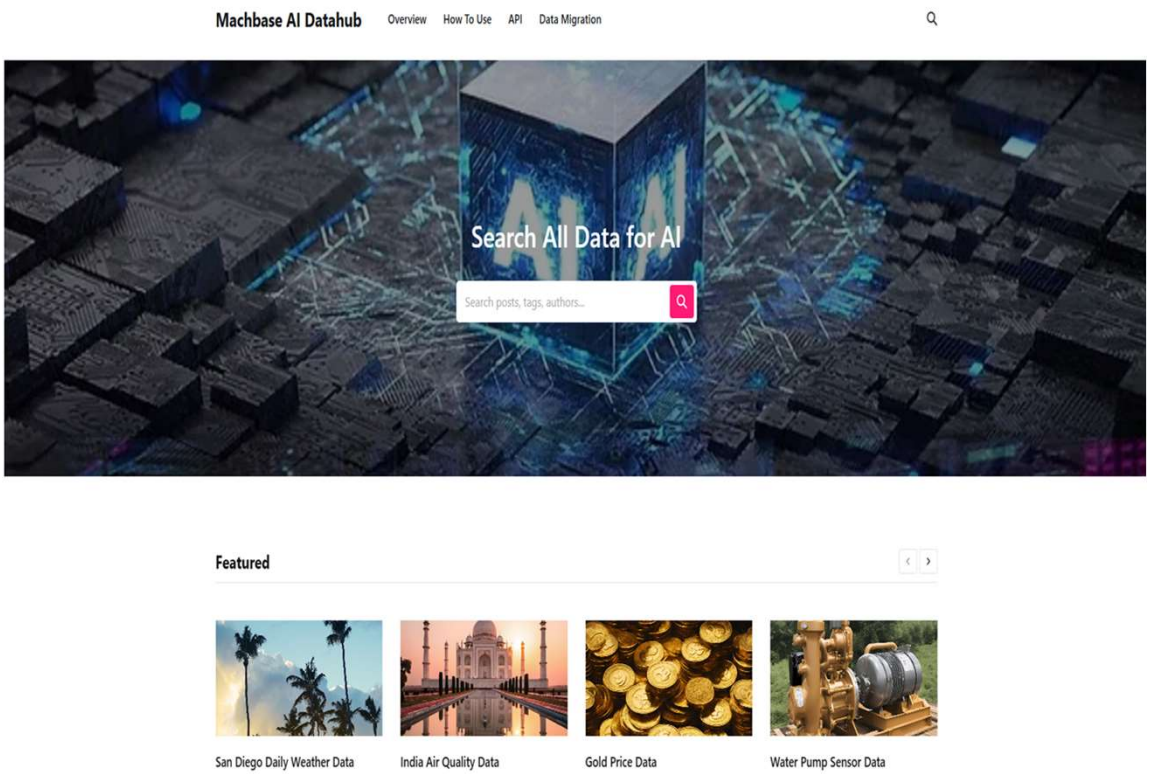
- 원본 데이터와 Machbase Neo에 이를 변환 적재하는 코드 제공
- Machbase Neo에 단 한줄의 command실행으로 데이터 로딩
- AI 학습 및 테스트에 관련한 샘플 코드 제공

MACHBASE

• Machbase Datahub Dataset

KOREA SUPERCOMPUTING CONFERENCE 2024
KSC2024

"클릭 한번으로 내 PC내에 AI를 위한 빅데이터 서버 구축"



1. **Kaggle Home IoT 데이터**
   (1400만건, 50초내 로딩 가능)
2. **KAMP 회전체 진동 데이터**
   (600만건, 20초내 로딩 가능)
3. **NASA 베어링 정상/고장 데이터**
   (9억 5천만건, 30분내 로딩 가능)
4. **제주도 풍력 발전소 데이터**
   (5년치, 400만건, 13초내 로딩 가능)
5. **뇌파 통신 인터페이스 데이터**
   (7200만건, 4분내 로딩 가능)
   외 전체 14종 제공 중

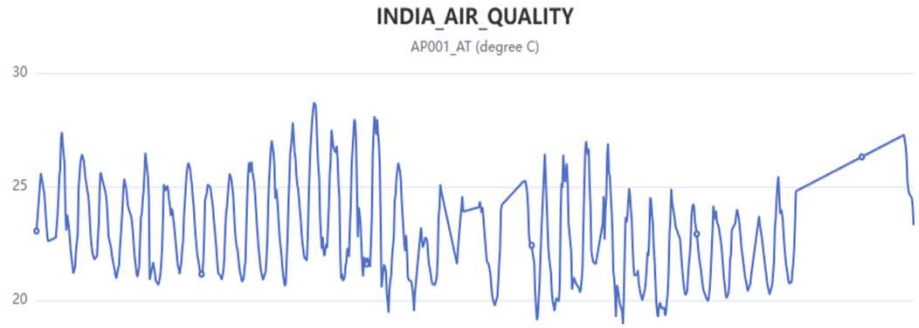
12월말까지 다양한 AI 학습 데이터 구축 예정

MACHBASE

• Machbase Datahub 데이터 프리뷰

2. Data Visualization with Machbase Neo

- Data visualization is possible through the Tag Analyzer in Machbase Neo.
- Select desired tag names and visualize them in various types of graphs.



◆ 정적 데이터 프리뷰

- Machbase Neo의 Tag analyzer의 스크린 캡쳐로 데이터 레이아웃 확인 가능
- Table View로 각 데이터 값 확인 가능

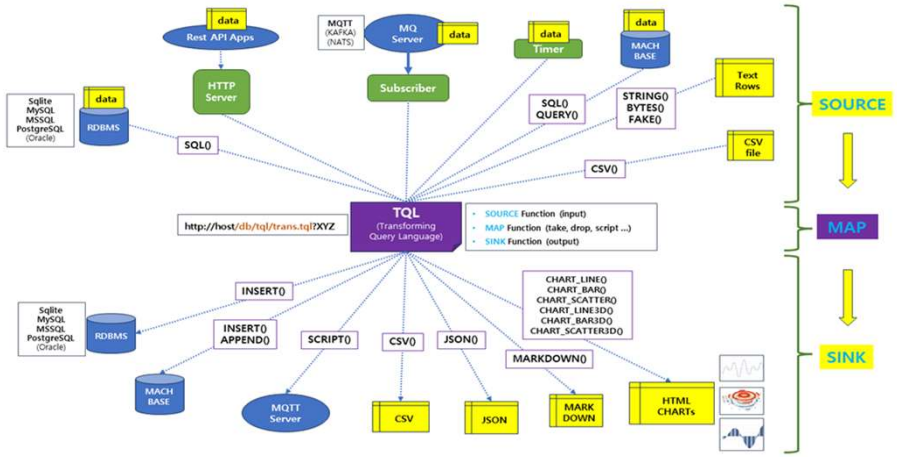
◆ 온라인 데이터 확인

- Machbase Neo 인스턴스 접속후 확인 가능
- 실시간 데이터 뷰어로 원하는 데이터를 간단히 확인 가능

Machbase Datahub 사용법

데이터 검색	• 활용을 원하는 데이터를 검색 • 데이터 설명 및 프리뷰 이미지 확인
데이터 둘러보기	• 온라인 Machbase Neo 접속하여 데이터 확인 • 온라인 데이터 뷰어 페이지를 통한 실제 데이터 미리보기
로컬 데이터허브 구축	• Datahub git (https://github.com/machbase/datahub/) 에서 관련 파일 다운로드 • 스키마 생성 등 스크립트 자동화(setup.wrk 수행)
데이터 다운로드 및 로컬 데이터베이스 로딩	• Datahub 페이지에 설명된 명령어 수행 • 고속 다운로드 및 로컬 DBMS에 데이터 로딩 기능 실행
샘플 코드 다운로드 및 활용	• Datahub git에 관련 코드 다운로드 이용 • Datahub 페이지에 코드 및 설명 제공

• Machbase Datahub API




```
// 데이터 조회 SQL 구문 및 파라미터 설정
SQL(
        strSprintf(
                `
                        SELECT name, time, value
                        FROM %s
                        WHERE name = '%s'
                        AND time BETWEEN TO_DATE('%s') and TO_DATE('%s')
                        LIMIT %d
                `,
                param('table') ?? 'home',
                param('name') ?? 'tag01',
                param('start') ?? '2024-09-04 12:13:14',
                param('end') ?? '2024-09-04 20:21:22',
                param('limit') ?? 10000
        )
)

// 데이터 출력 형식 지정
CSV()
```

◆ API

- Machbase Neo의 TQL(Transforming Query Language)로 구현 (소스 코드 제공)
- TQL은 다양한 입력(HTTP, SQL, MQTT, Machbase, CSV) 파일을 여러 다른 형태로 변환하는 기능을 제공
- TQL 스크립트는 HTTP Rest API 형태로 호출 가능

◆ **REST API 종류**

`http://HOST:PORT/db/tql/datahub/api/v1/API`

- **select-rawdata.tql** : tag 및 시간 범위 데이터 추출
- **select-rollup.tql** : tag의 통계데이터 추출
- **select-scale.tql** : tag의 최대, 최소값 추출

• Machbase Datahub 활용 예

Table Creation

- Create the bearing table as a data repository, and specify
- The table is created immediately upon pressing the "Run"
- If the bearing table exists, execute the first line and then t

```
drop table bearing cascade;
create tag table if not exists bearing (
    name varchar(32) primary key,
    time datetime basetime,
    value double summarized
) with rollup;
```

```
curl http://data.yotahub.com/2024-3/datahub-2024-3-be
```

Model Testing

- Proceed with model testing on the test data based on the threshold calculated in the previous step.

```
# Apply the model to the test data
test_loss = []
test_label = []
with torch.no_grad():
    model_.eval()
    for batch_idx, test_data in enumerate(test_dataloader):

        inputs_test = test_data[0].to(device).float()
        label = test_data[1].to(device).long()

        outputs_test = model_(inputs_test)
        loss = criterion(outputs_test, inputs_test)

        test_loss.append(loss.item())
        test_label.append(label.item())

# Create a DataFrame for the test results
result = pd.DataFrame(test_loss, columns=['Reconst_Loss'])

# Set the actual labels
result['label'] = test_label

# Classify normal and abnormal based on each threshold
result['pred'] = np.where(result['Reconst_Loss'] > threshold, 1, 0)
```

[25]

Model Performance Evaluation

```
# Print F1 Score based on testing data
print(classification_report(result['label'], result['pred']))
```

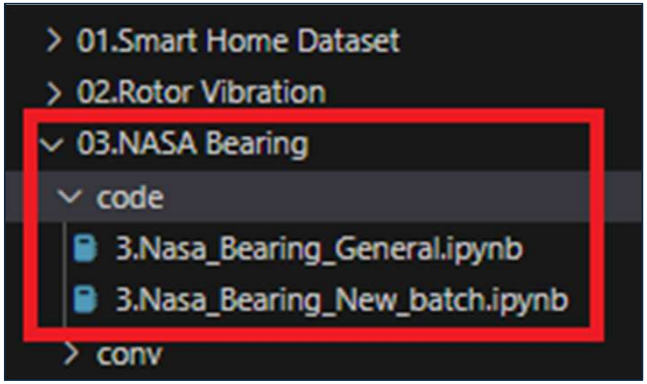
[26]

```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       179
           1       1.00      1.00      1.00        35

    accuracy                           1.00       214
   macro avg       1.00      1.00      1.00       214
weighted avg       1.00      1.00      1.00       214
```

- NASA Bearing Data
  - 원 데이터 : kaggle
  - https://datahub.machbase.com/
  - 약 6GB, **CSV file 18,000개**
  - **총 951,910,400건 (30분내 로딩)**
- 쉽고 편리한 거대 빅데이터 로딩
  1. 테이블 생성
  2. 다운로드 스크립트 수행
  3. 로딩 확인 (Gap 확인 등)
- Jupyter notebook AI 코드 수행 및 실행



• Datahub 데이터 추출 표준 API 활용예


```python
# Data load function
# '1D': Daily interval (1 day)
# '1H': Hourly interval (1 hour)
# '1T' or 'min': Minute interval (1 minute)
# '1S': Second interval (1 second)
def data_load(table, name, start_time, end_time, timeformat, resample_time):

    # Load data
    df = pd.read_csv(f'http://127.0.0.1:5654/db/tql/datahub/api/v1/
                       select-rawdata.tql?table={table}&name={name}&start={start_time}&end={end_time}&timeformat={timeformat}')

    # Convert to data grouped by the same time
    df = df.pivot_table(index='TIME', columns='NAME', values='VALUE', aggfunc='first').reset_index()

    # Set time index
    df = df.set_index(pd.to_datetime(df['TIME']))
    df = df.drop(['TIME'], axis=1)

    # Resampling with 1-second intervals
    # Can be modified to desired intervals such as day, hour, minute, etc.
    df = df.resample(f'{resample_time}').mean()

    return df
```

◆ AI 응용 프로그램 예제

- Datahub api를 이용하여 데이터 로드 예제
- Github(https://github.com/machbase/datahub/blob/main/dataset/2024/1.Smart%20Home%20Dataset/code/1.Smart_Home_New_batch.ipynb)
- 긴단한 데이터 로드 방법 : pandas의 read_csv에 datahub의 API URL을 전달하면 데이터가 로드됨
- **기존의 csv 파일 이용과 큰 차이 없음**
- 이후 Pivot, fft, main-max scale을 거쳐 LSTM AE 신경망을 이용하여 학습을 진행
- 테스트 및 모델 평가, 과적합 평가 코드까지 제공함

DEFAULT 포맷의 시간으로 추출하고, 타임존을 KST로 출력

```
$ curl -G http://127.0.0.1:5654/db/tql/datahub/common/select-rawdata.tql --data-urlencode "table=home" --data-urlencode 'target=name,time,value' --data-urlencode 'limit=10' --data-urlencode 'start=2016-01-01 14:00:0' --data-urlencode 'end=2016-01-01 14:00:2' --data-urlencode "name='TAG-pressure','TAG-dewPoint'" --data-urlencode 'timeformat=DEFAULT' --data-urlencode 'timezone=KST'
name,time,value
TAG-pressure,2016-01-01 14:00:00,1016.91
TAG-pressure,2016-01-01 14:00:01,1016.91
TAG-pressure,2016-01-01 14:00:02,1016.91
TAG-dewPoint,2016-01-01 14:00:00,24.4
TAG-dewPoint,2016-01-01 14:00:01,24.4
TAG-dewPoint,2016-01-01 14:00:02,24.4
```

• Machbase AI의 향후과제

◆ Datahub 데이터 및 샘플 추가
- 표준화된 더 많은 데이터와 샘플코드를 추가하여 AIoT 데이터 처리를 더 간편하게 수행
- Datahub를 AIoT 데이터 처리 관련한 명실상부한 허브로 육성

◆ Datahub API 기능 추가, AI를 위한 Database 확장
- AI 응용개발에서 반드시 필요한 데이터 전처리 기능의 추가 개발
- Datahub API에서 Application에서 수행해야 하는 전처리 기능 등을 지원하여 개발 편의성 향상

◆ 자동화된 AI 학습, 이상감지 코드 생성 기술 개발
- LLM등의 기술을 응용하여 Machbase + Datahub API 코드를 자동 생성하도록 하는 기술 개발

결론

◆ AIoT 데이터의 기존 처리 한계
- 데이터 수집, 분석, AI 학습 개발에서 Hadoop 기반 CSV 파일 이용 아키텍처의 한계
- 데이터 관리와 분석, 학습시 데이터 확보에 있어 다양한 문제가 상존
- AI 모델 개발에 있어 많은 시간과 노력을 소요

◆ Machbase의 AI 모델 개발 아키텍처
- 데이터는 표준화 된 형태로 Machbase Neo TSDB에 저장
- 학습 Application 개발에 필요한 데이터는 Datahub API를 통해 접근
- 대량의 센서 데이터 학습시에도 Database에서 데이터를 읽어들여 학습하는 모델 제시
- 데이터 관리, 학습 과정 개선으로 개발 비용 감소

◆ Datahub
- IoT 데이터 AI 모델 개발을 위한 데이터 포털
- 표준화된 데이터 제공 및 AI 학습 프로그램 제공
- 향후 더욱 보강된 데이터 제공 및 다양한 학습 프로그램을 제공하여 AIoT 개발에 기여

◆ **Machbase Neo + Datahub**
- CSV 지옥에서 탈출
- GPU 메모리 한계 극복
- 데이터 접근 표준화
- AI 코드 재활용/표준화
- AI 개발자도 데이터 개인화
- AI 모델 개발 시간/비용 최소화

감사합니다.

MACHBASE