복합 시계열 데이터 분석 및 예측을 자동화하는 MLOps 시스템

경희대학교 컴퓨터공학과 안영민

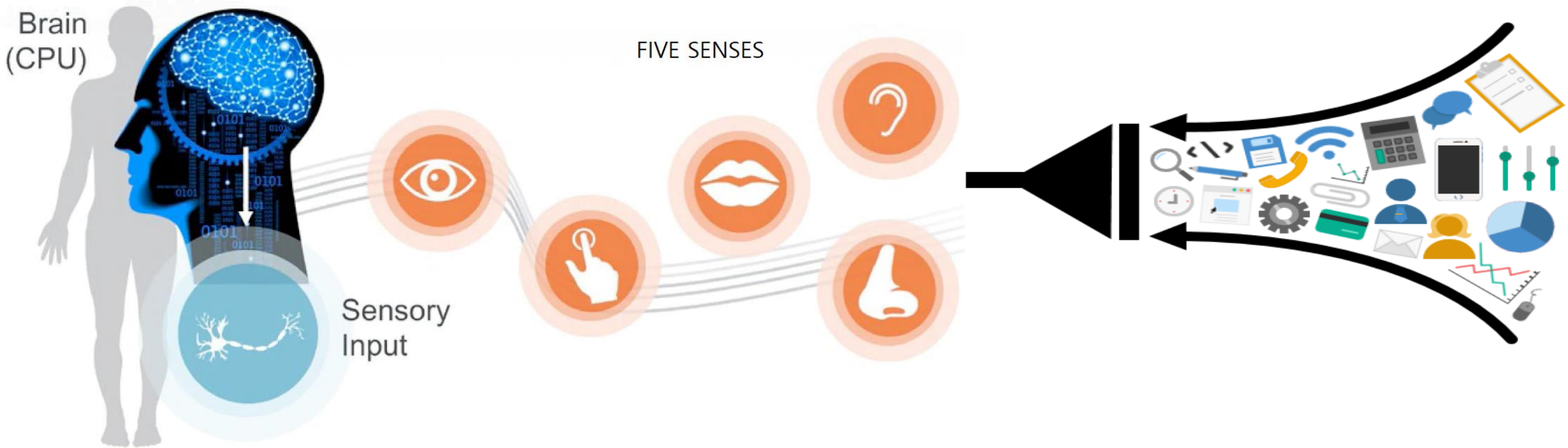
지도교수 허의남

목차

경희대학교 컴퓨터공학과
안영민

연구 배경

더 나은 지능형 비즈니스를 위한 데이터 통합

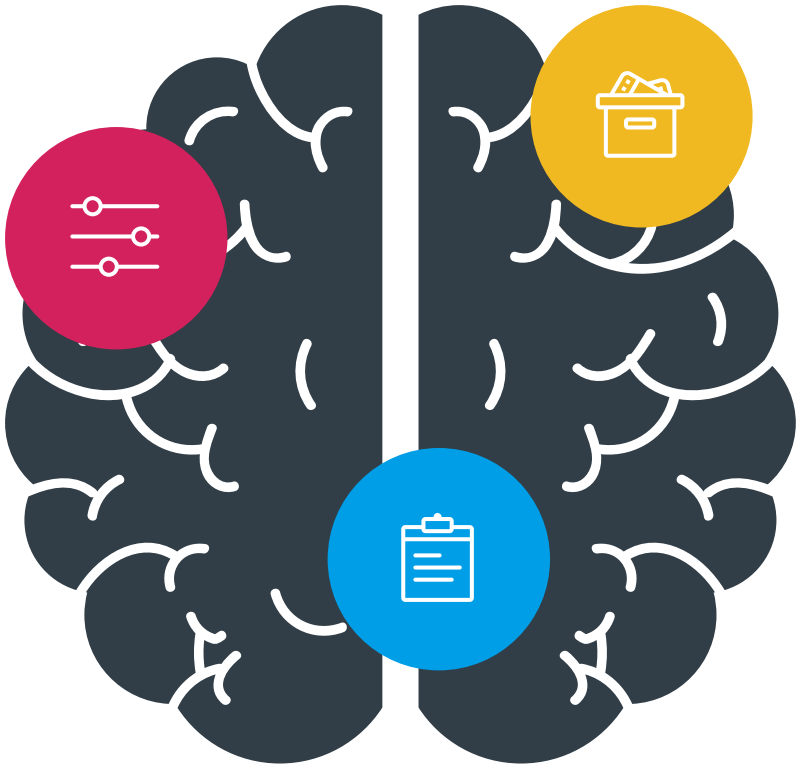


더 스마트하게, 더 나은 가치를

빅데이터 시대 비즈니스 차별성은 데이터 통합으로 새로운 정보를 융합하는 능력

연구 배경

AI

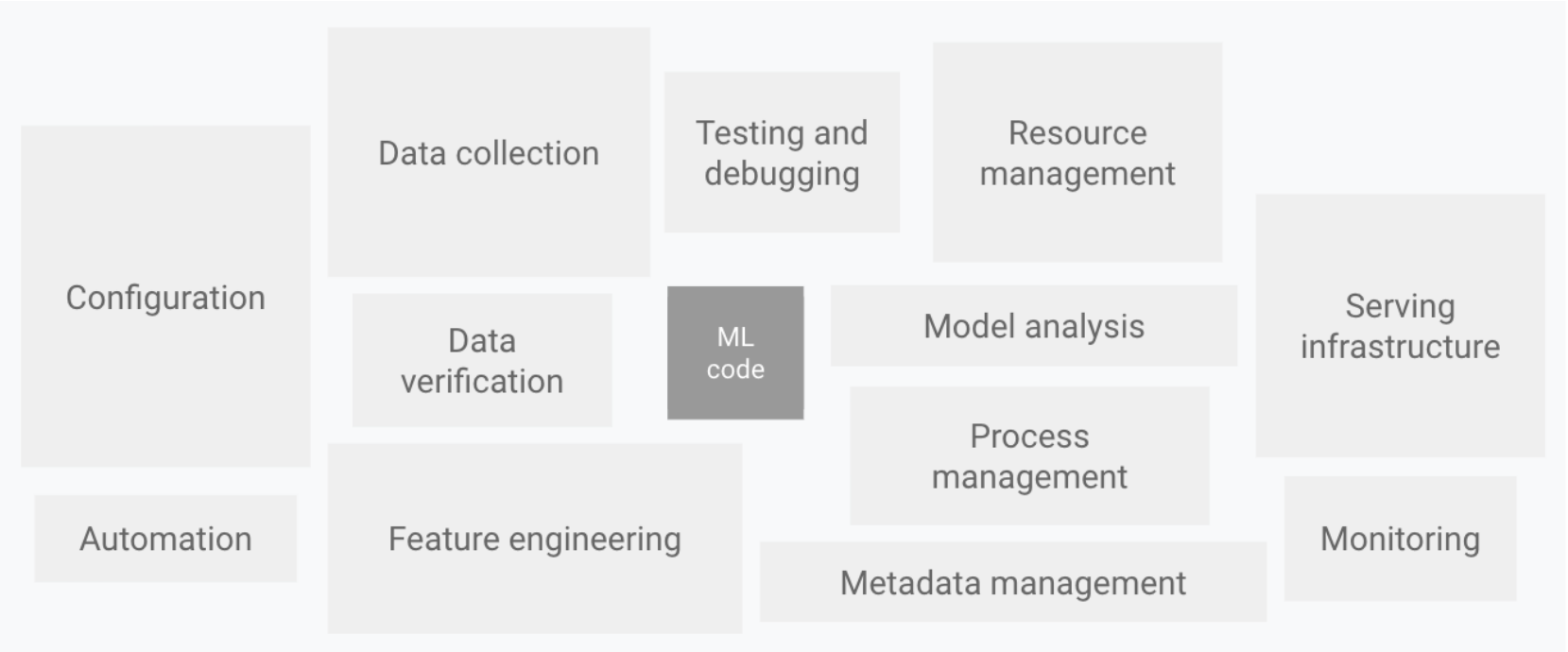


10개 중 9개 조직이 AI 시스템을 비즈니스 핵심 기술로 전략

MIT Sloan Management, https://web-assets.bcg.com/1e/4f/925e66794465ad89953ff604b656/mit-bcg-expanding-ai-impact-with-organizational-learning-oct-2020-n.pdf (search 2022-12-05), 2020.

연구 배경

ML 시스템 운영 과정에서 어려움



모델 개발은 빙산의 일각

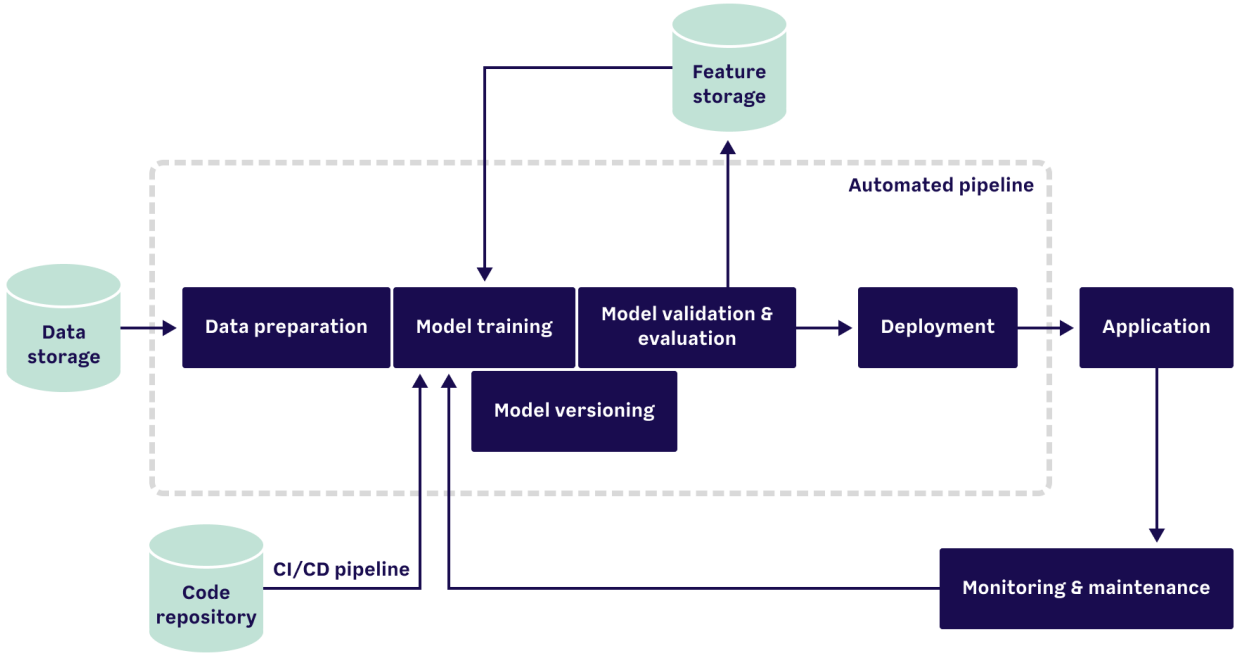
반복적인 비즈니스 저해 요인

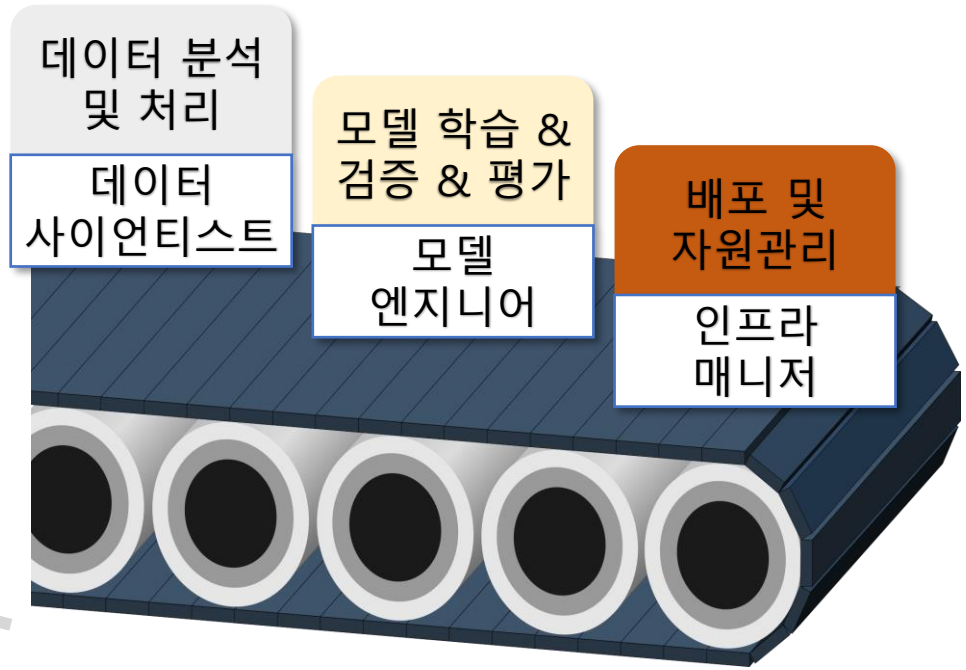
- 데이터 처리
- 모델 학습 및 검증
- 배포 및 운영

D Sculley et al, Hidden Technical Debt in Machine Learning System, *NIPS*, 2015

연구 배경

파이프라인을 자동화하는 MLOps 시스템





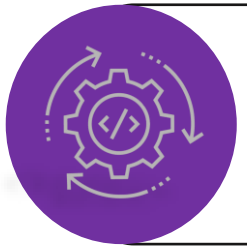
관련 연구

복합 시계열 데이터 분석

- Huaiyu Wan et al., CTS-LSTM: …, Knowledge-Based Systems, 2020.
- Xingjian Shi et al., Convolutional LSTM Network: …, NIPS, 2015.

한계

- 복합 데이터로부터 얻을 수 있는 정보를 단지 신경망을 연결하는 방식으로 합하여 설명가능성이 떨어진다.
- 동일한 타임스탬프에 수집되는 시계열 데이터만을 가정한다.



MLOps & Automation

- D Sculley et al, Hidden Technical Debt in Machine Learning System, *NIPS*, 2015
- Lukas Tuggener, Automated Machine Learning in Practice: …, Ada, 2019

한계

- 이론적인 내용들이 주를 이루며, 다양한 툴을 결합한 단일 통합 솔루션에 대한 명확한 해법을 얻기는 어렵다.

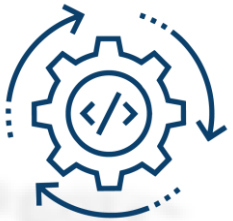
제안 연구

가치

TARGET 01 수치형 시계열 데이터 회귀 예측 모델링

다양한 시계열 분석 및 예측
애플리케이션에 적용

TARGET 02 자동화된 MLOps 파이프라인



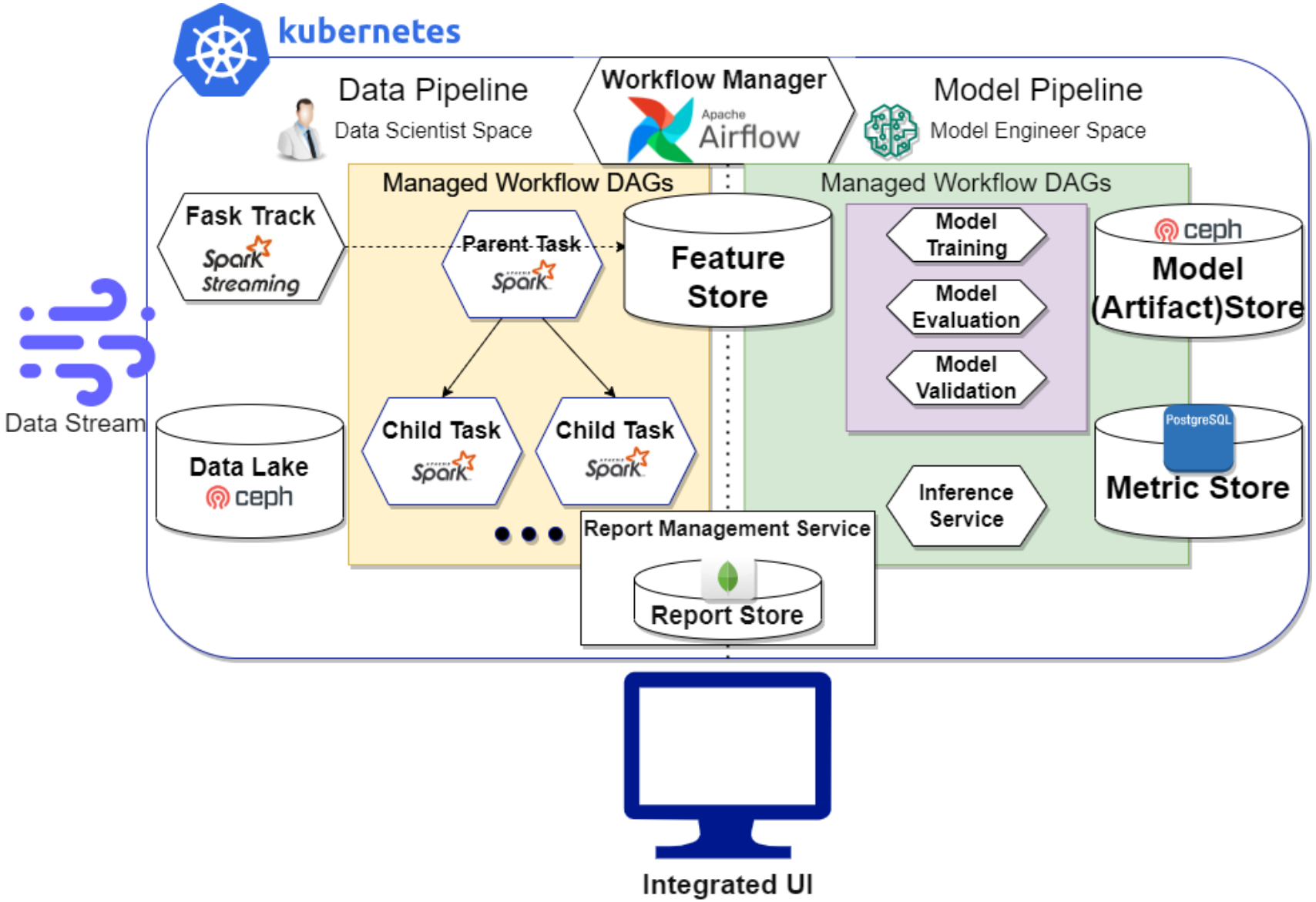
반복적인 비즈니스
저해 요인 탈피

TARGET 03 수평적 확장 용이한 설계

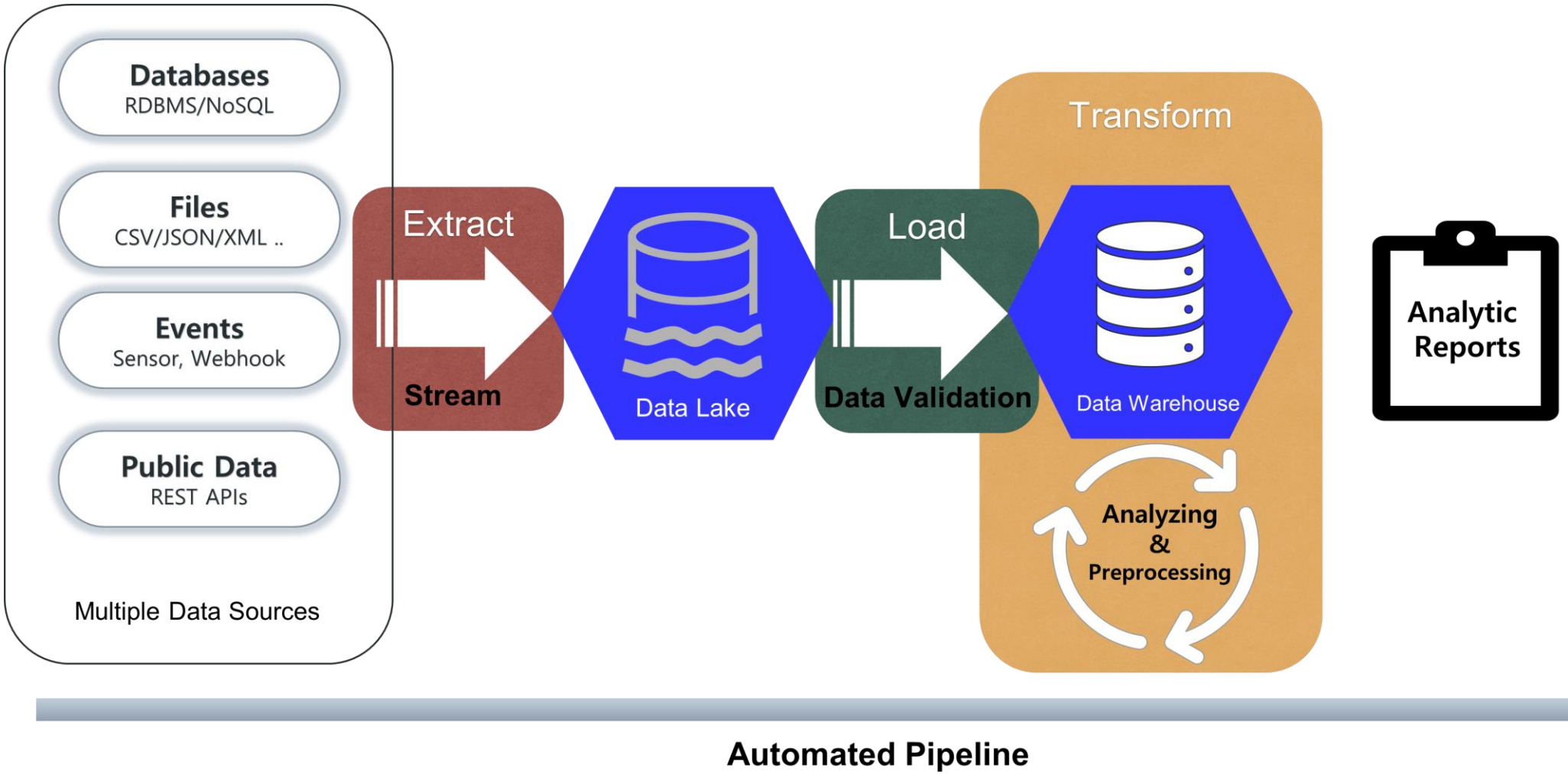
탄력적 분석 인프라 제공

경희대학교 컴퓨터공학과
안영민

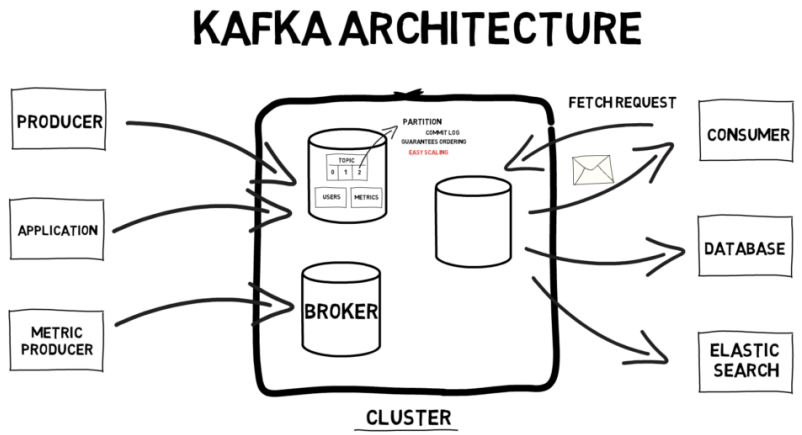
제안 연구


kubernetes
Data Pipeline
Data Scientist Space
Workflow Manager
Apache Airflow
Model Pipeline
Model Engineer Space
Managed Workflow DAGs
Managed Workflow DAGs
Fask Track
Spark Streaming
Parent Task
Spark
Feature Store
Model Training
Model Evaluation
Model Validation
ceph
Model (Artifact)Store
Data Stream
Data Lake
ceph
Child Task
Spark
Child Task
Spark
PostgreSQL
Metric Store
Inference Service
Report Management Service
Report Store
Integrated UI

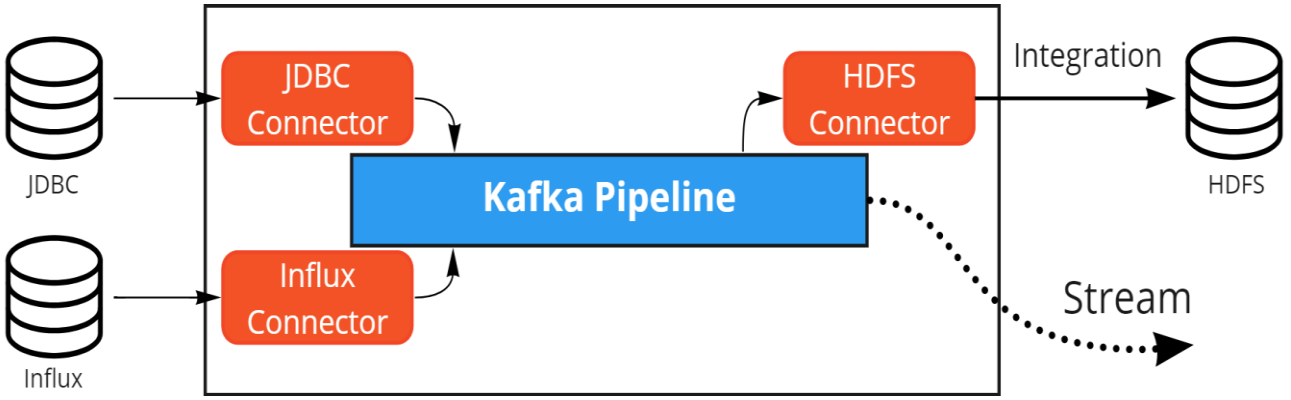
데이터 파이프라인


Databases
RDBMS/NoSQL
Files
CSV/JSON/XML ..
Events
Sensor, Webhook
Public Data
REST APIs
Multiple Data Sources
Extract
Stream
Data Lake
Load
Data Validation
Transform
Data Warehouse
Analyzing
&
Preprocessing
Analytic
Reports
Automated Pipeline

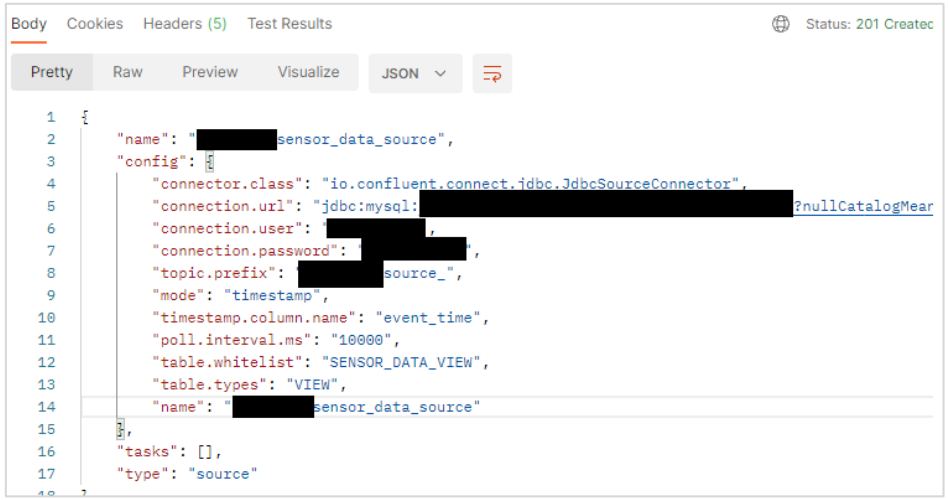
데이터 추출



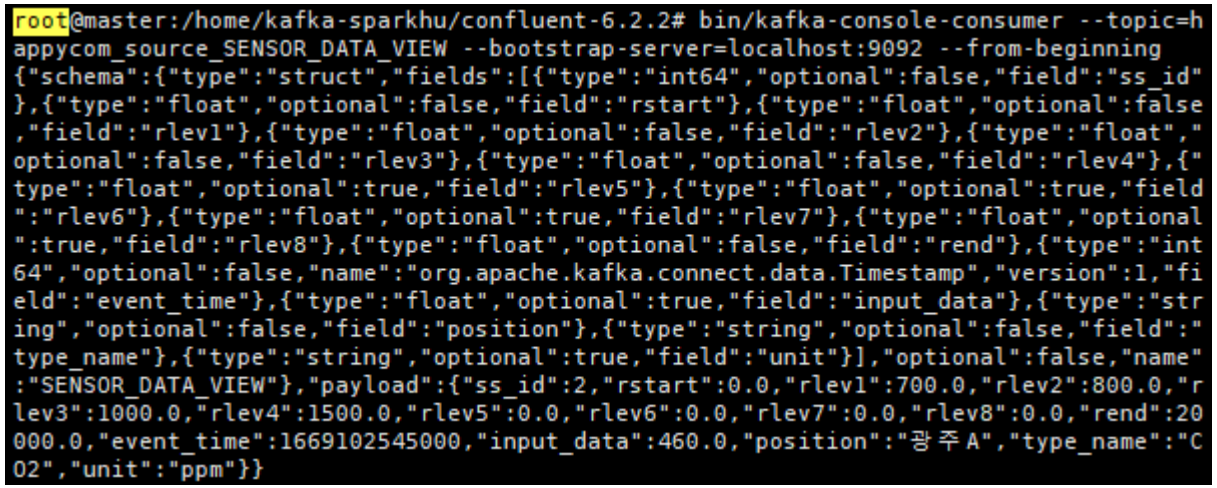
▲ Kafka Architecture



▲ Kafka Connector를 이용한 싱크 자동화

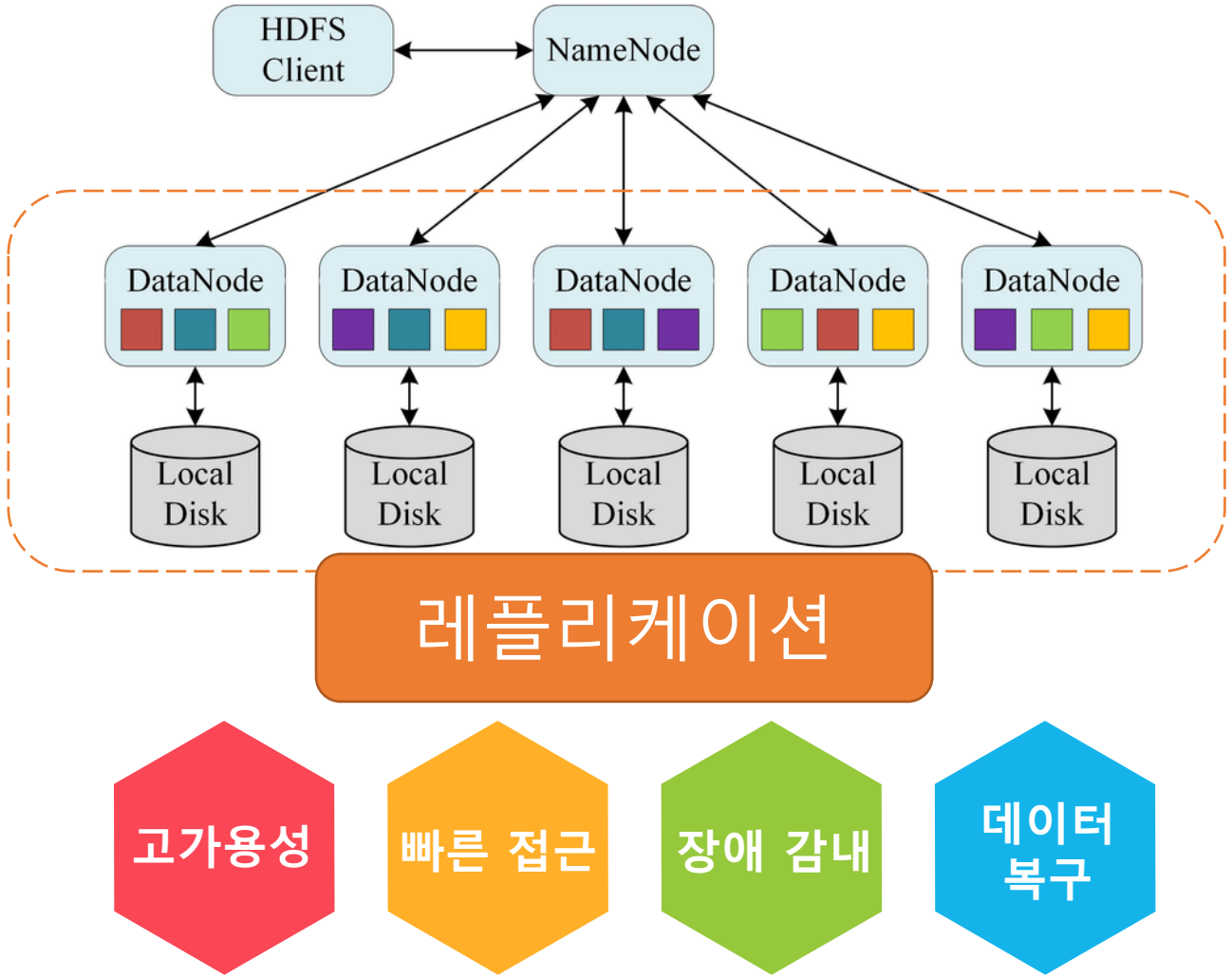


▲ REST API로 Connector 간편 등록



▲ JSON 데이터 교환 표준 형식으로 실시간 데이터 추출

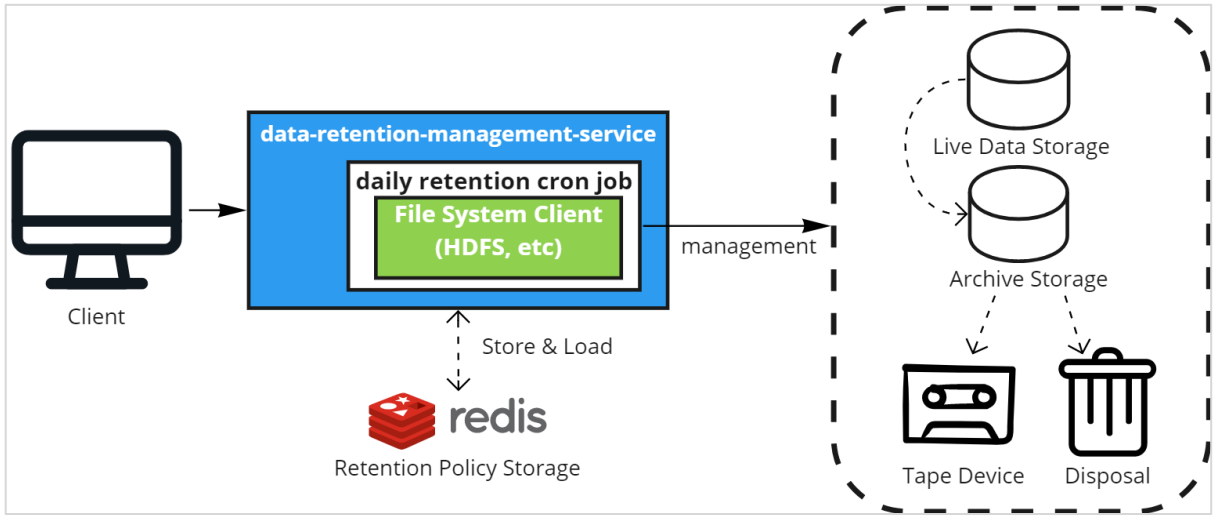
데이터 레이크



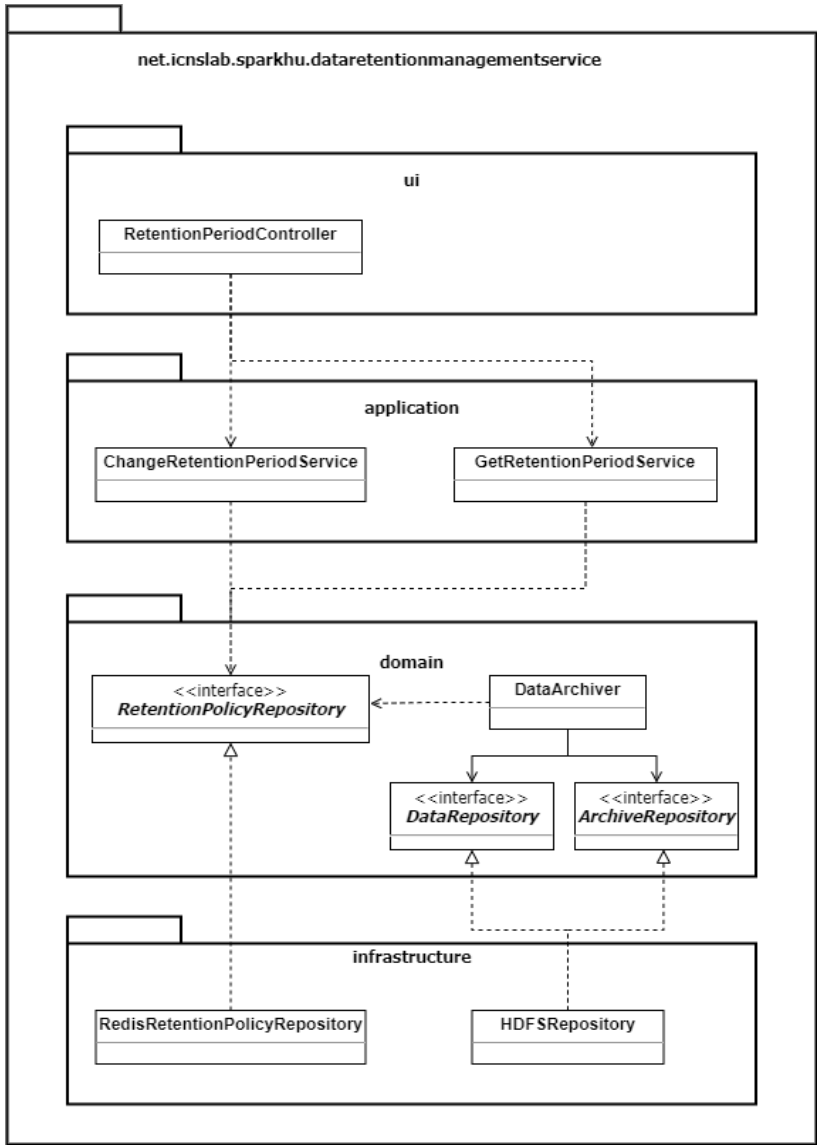

```
ce_SENSOR_DATA_VIEW/year=2022/month=04
ce_SENSOR_DATA_VIEW/year=2022/month=05
ce_SENSOR_DATA_VIEW/year=2022/month=06
ce_SENSOR_DATA_VIEW/year=2022/month=07
ce_SENSOR_DATA_VIEW/year=2022/month=08
ce_SENSOR_DATA_VIEW/year=2022/month=09
ce_SENSOR_DATA_VIEW/year=2022/month=10
ce_SENSOR_DATA_VIEW/year=2022/month=11
eter-tuning-task# hdfs dfs -ls /data/raw/happy




ce_SENSOR_DATA_VIEW/year=2022/month=11/day=01
ce_SENSOR_DATA_VIEW/year=2022/month=11/day=02
ce_SENSOR_DATA_VIEW/year=2022/month=11/day=03
ce_SENSOR_DATA_VIEW/year=2022/month=11/day=04
ce_SENSOR_DATA_VIEW/year=2022/month=11/day=05
ce_SENSOR_DATA_VIEW/year=2022/month=11/day=06
```

▲ 데이터 레이크에 파티션되어 저장된 데이터

데이터 레이크

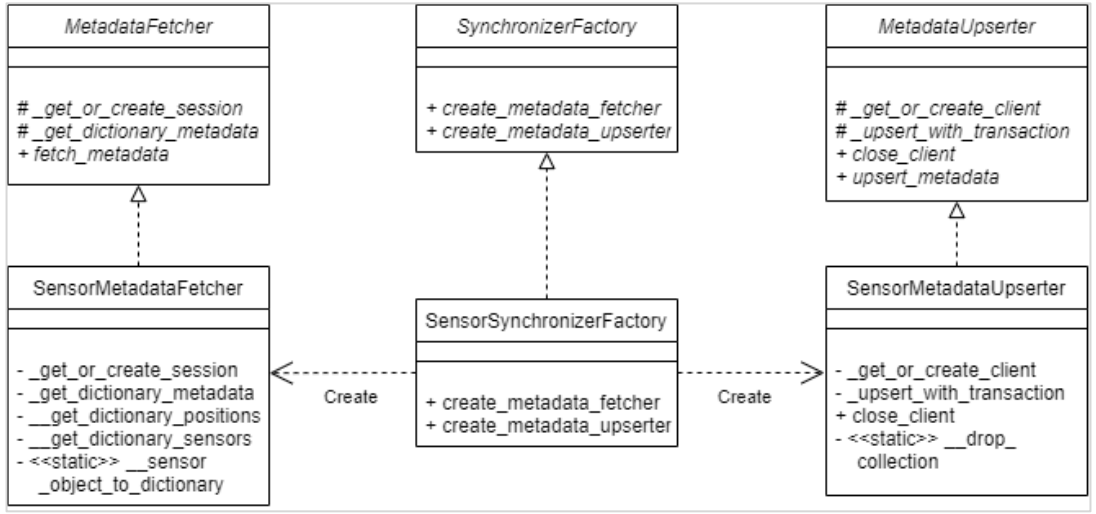


▲ 데이터 보존 관리 서비스

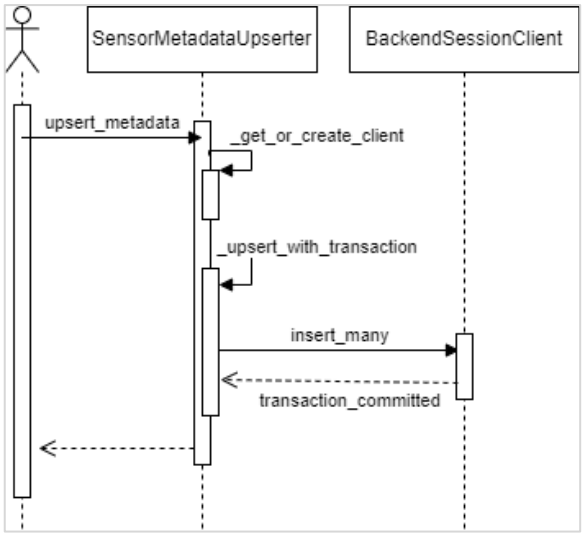


▲ 데이터 보존 관리 서비스 패키지 다이어그램

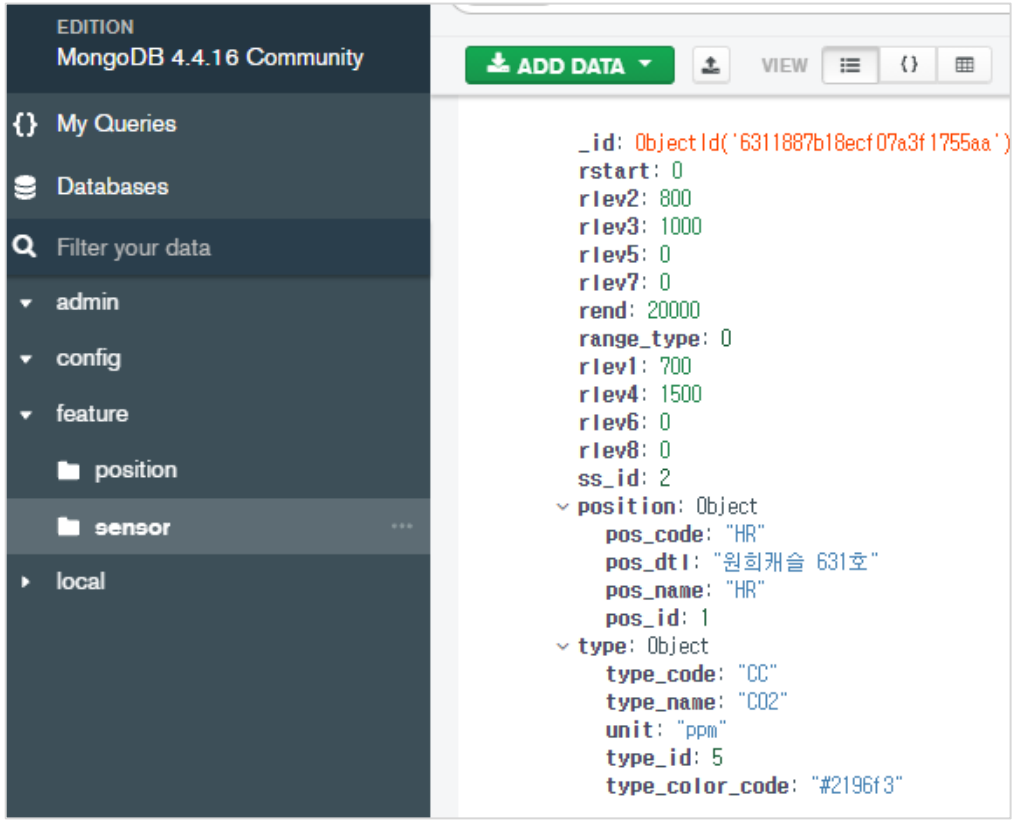
메타데이터 관리



▲ Metadata Synchronizer Class Diagram



▲ upsert_metadata Sequence Diagram



▲ 문서 형태로 MongoDB에 저장된 metadata

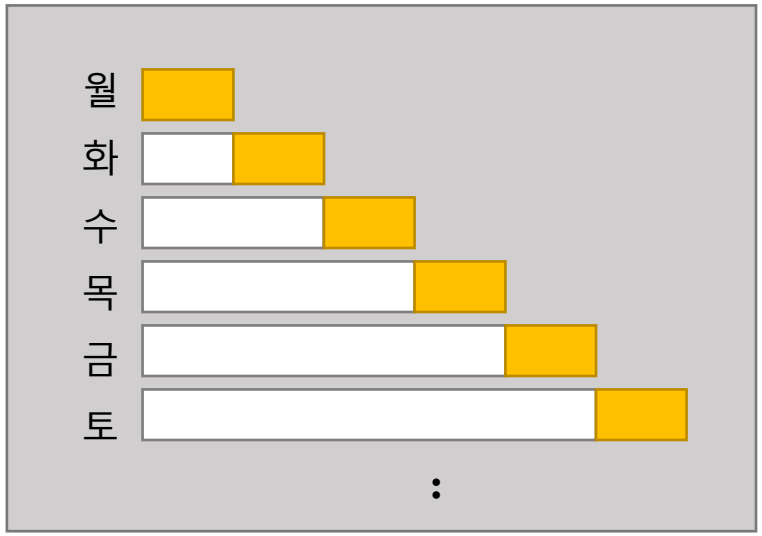
시계열 데이터 분석 및 전처리

Data Validation
- Schema Validation
- Timestamp Validation

Data Cleaning
- Deduplication
- Missing Value Processing
- Anomaly Value Processing

Data Analysis
- Univariate Analysis
- Multivariate Analysis
- ...

Data Transformation
- Data Scailing
- Windowing



매일 증분하는 데이터
-> 주기적 처리 능력 자동화 필요

다양한 시계열로부터 오는 빅데이터
-> 분산 처리 필요

아파치 스파크를 이용한 분산 처리, 아파치 하이브 데이터 웨어하우스

Distributed Processing
Framework with Big data

Open-source.

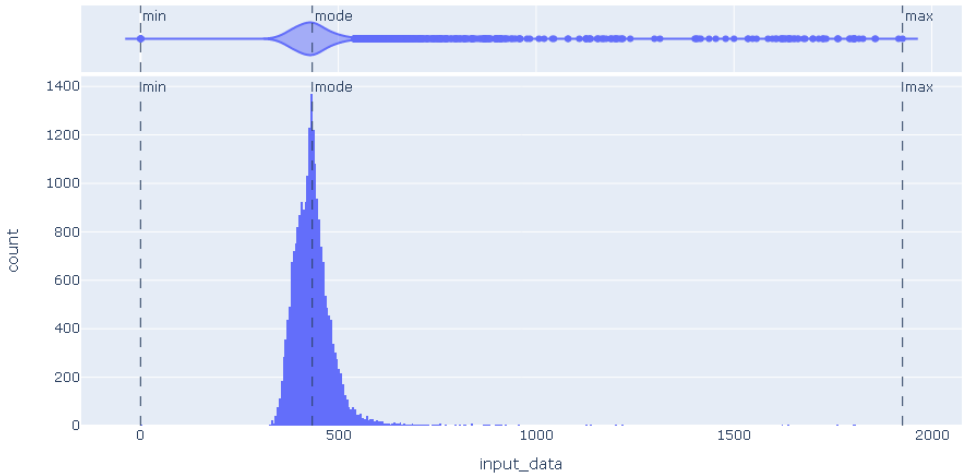
Memory-based Engine

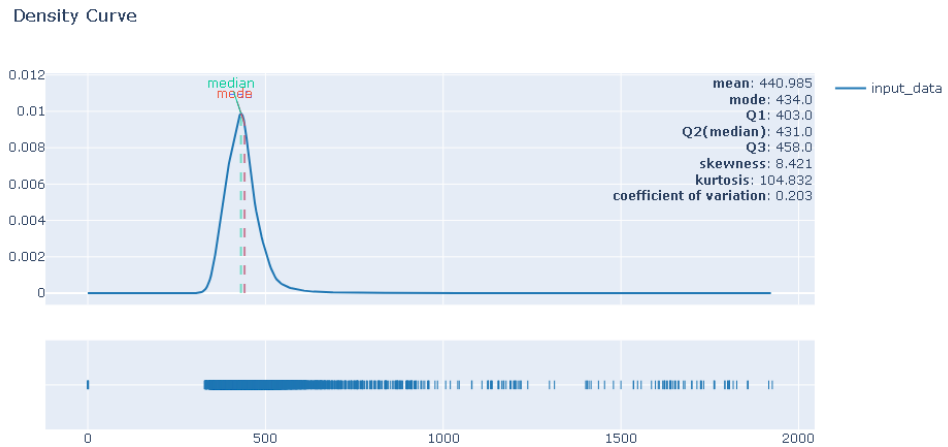
Queryable.

Easy to use

Various File Format

기술적 통계 및 중복 제거





Timestamp	Co2
2022-05-01 00:05:49	390.0
2022-05-01 00:15:49	389.0
2022-05-01 00:15:49	389.0
2022-05-01 00:25:49	396.0

경희대학교 컴퓨터공학과
안영민

정규성 분석 (주기 추정)

대부분 시계열 데이터 활용 알고리즘은
정기적으로 수집된 데이터를 가정

주기는 중요한 정보

❖ Clustering and Approximated GCD

Algorithm 1. Cluster and Approximated GCD

Input: time series dataframe *TS*

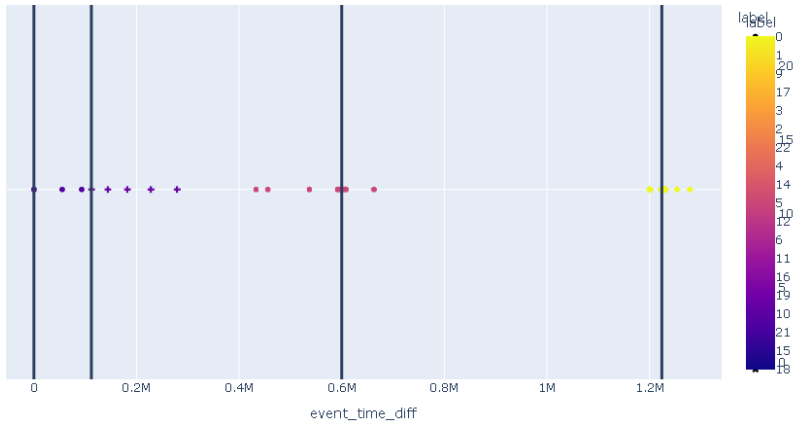
DiffTs = Calc new dataframe with lagged timediff *TS[timestamp]*

ClusteredTS = Cluster *DiffTs[timestamp]*

Superperiod = Approximated GCD (Top 95% mode diff *ClusterTS*)

Lagged timestamp diff
610000
530000
660000
…

<타임스탬프 간격 도출>



<클러스터링된 타임스탬프 격차>

$$Estimated\ superperiod$$

$$= Approximated\ GCD\ (\ list(top\ 95\%'s\ mode\ diffs)\ )$$

<상위 95% 그룹 최빈값들의 근사된 최대공배수>

데이터 정제 (결측치 검출)

❖ Divide and Round

1. 정렬된 시계열 데이터의 이웃한 타임스탬프간 차이를 나열
2. 해당 타임스탬프 차이를 앞서 구한 주기로 구분
3. 나눠진 값을 반올림하여 해당 구간 내에 표시되지 않은 결측치의 개수를 추정
4. 추정된 개수만큼 결측치를 표시

Timestamp	Co2
2022-05-01 00:05:49	390.0
2022-05-01 00:15:49	389.0
2022-05-01 00:45:49	389.0
2022-05-01 00:25:49	396.0

num of Missing values = Round([45:49 – 15:49] / 10:00) – 1 = **2**

추정된 결측치

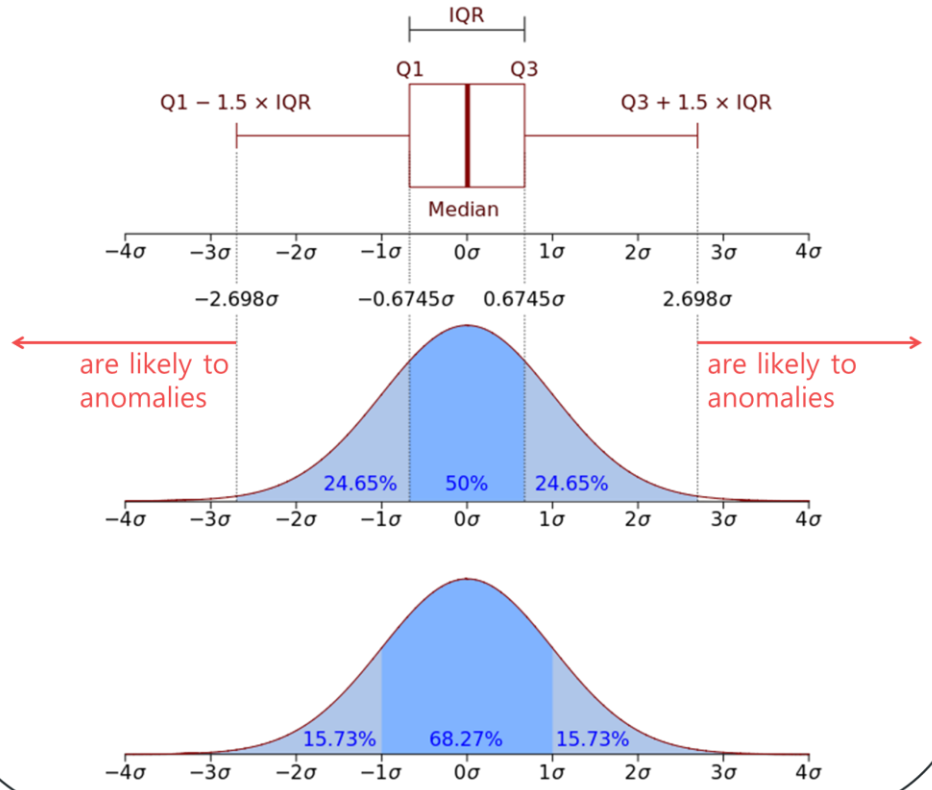
Timestamp	Co2
2022-05-01 00:05:49	390.0
2022-05-01 00:15:49	389.0
2022-05-01 00:25:49	N/A
2022-05-01 00:35:49	N/A
2022-05-01 00:45:49	389.0
2022-05-01 00:25:49	396.0

데이터 정제 (결측치 처리)

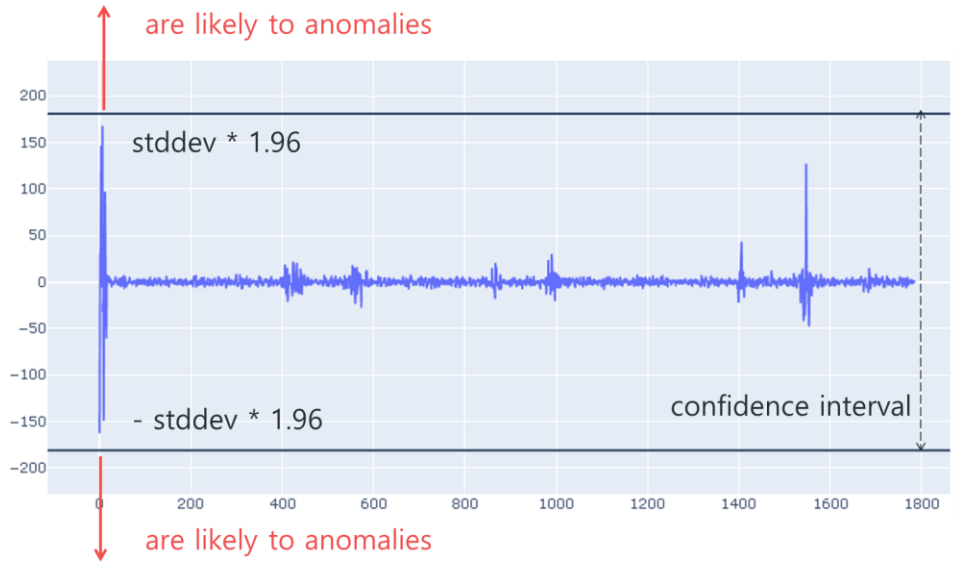
[결측치 처리 알고리즘에 따른 결측치 예측 성능 비교]

데이터 정제 (노이즈 검출)

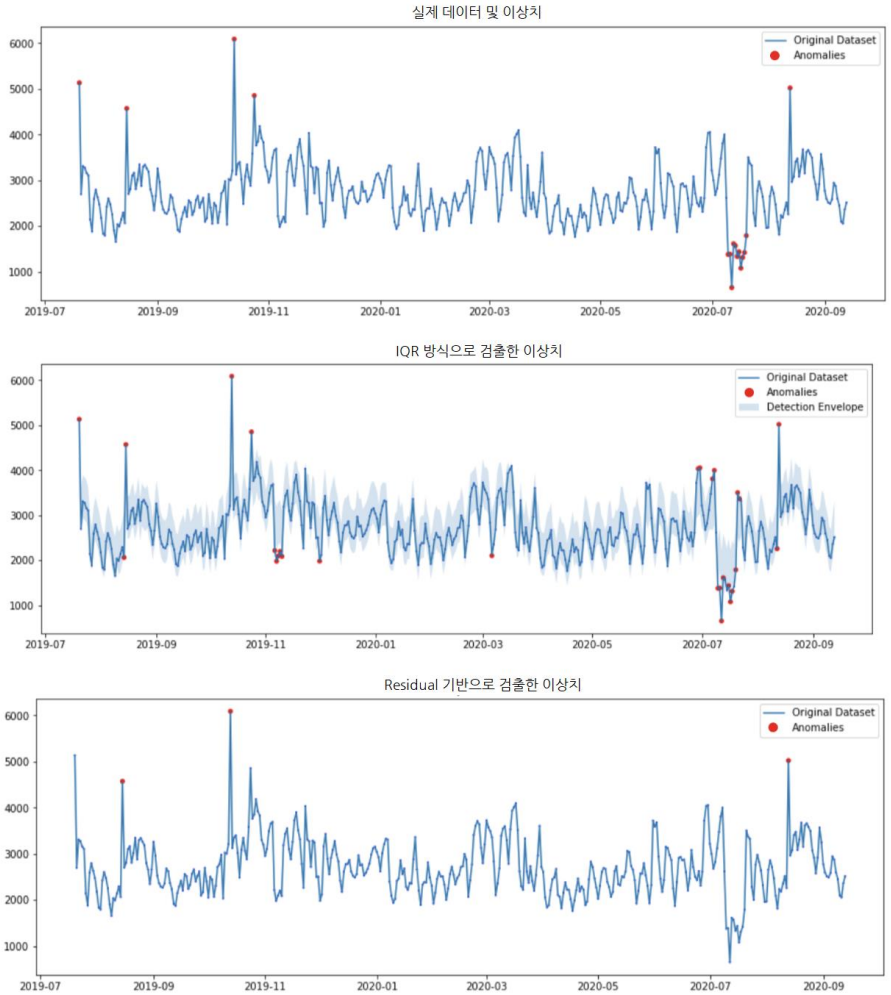
❖IQR Rule-based Anomaly Detection



❖Residual Analysis-based Anomaly Detection



데이터 정제 (노이즈 처리)



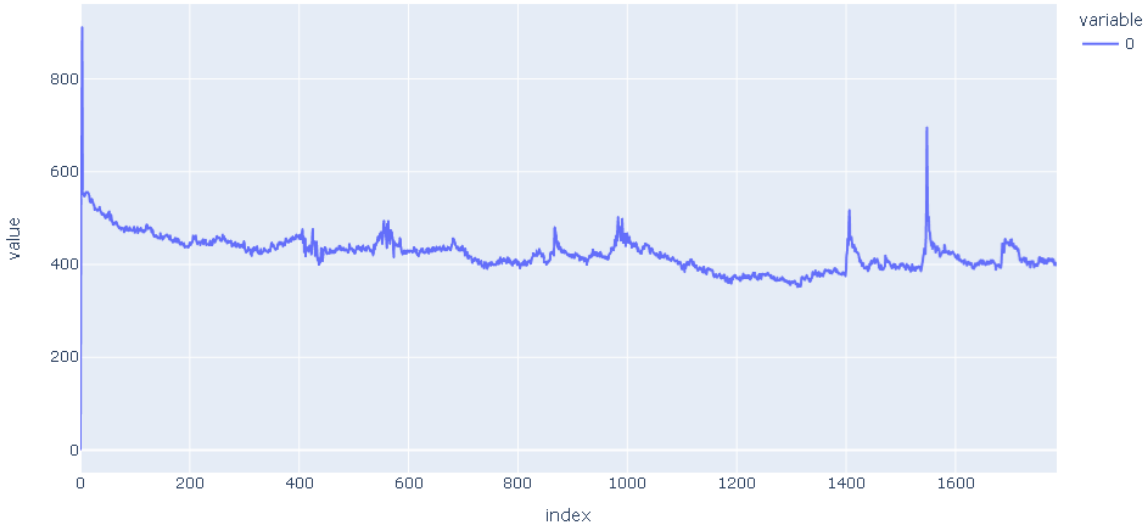
[노이즈 검출 기법에 따른 검출 결과]

Anomaly processing	RMSE	MAE	MAPE
		LSTM	
No processing	0.8867	0.5924	10.32
Drop out	0.5939	0.4984	8.68
Replace with min/max	0.5784	0.4672	8.41
EMA(smoothing)	0.3513	0.1815	4.79

[노이즈 처리 기법에 따른 예측 성능]

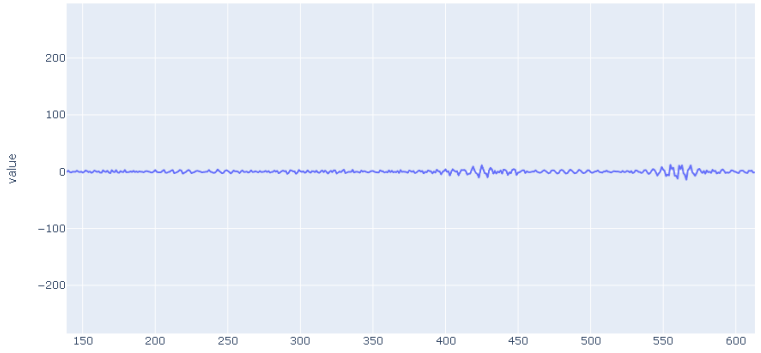
경희대학교 컴퓨터공학과
안영민

시계열 분해



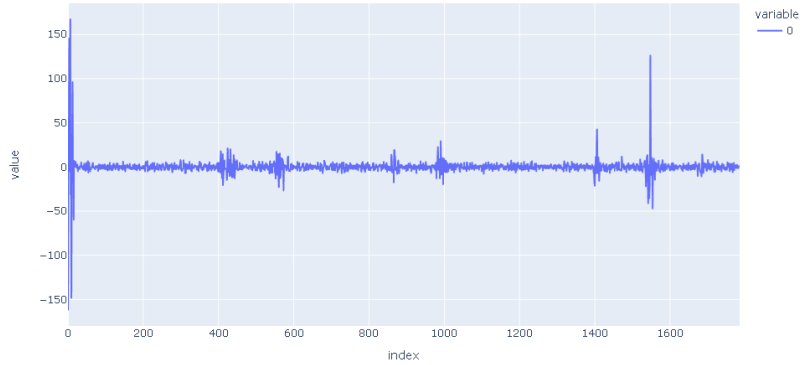
[관측]

=



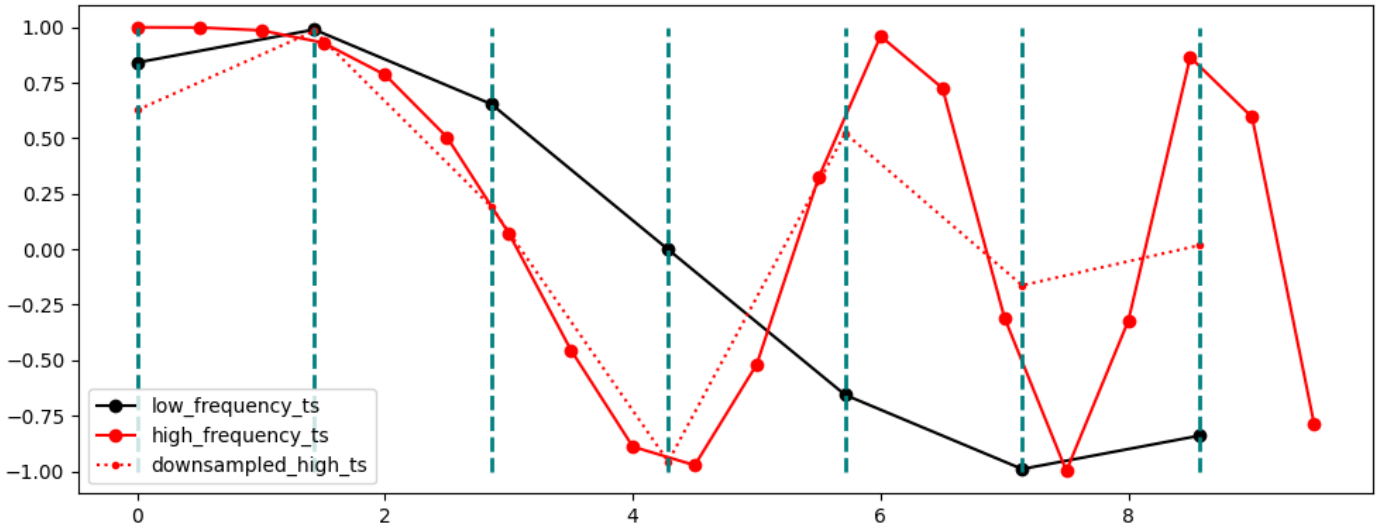
[추세]

[계절성]

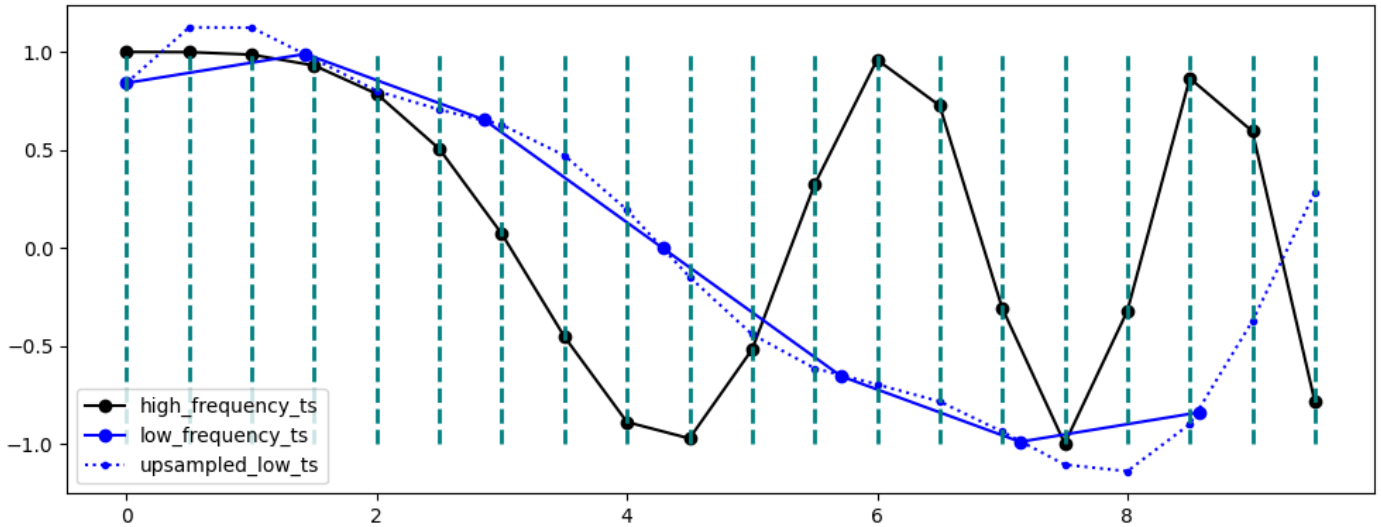


[잔차]

샘플링

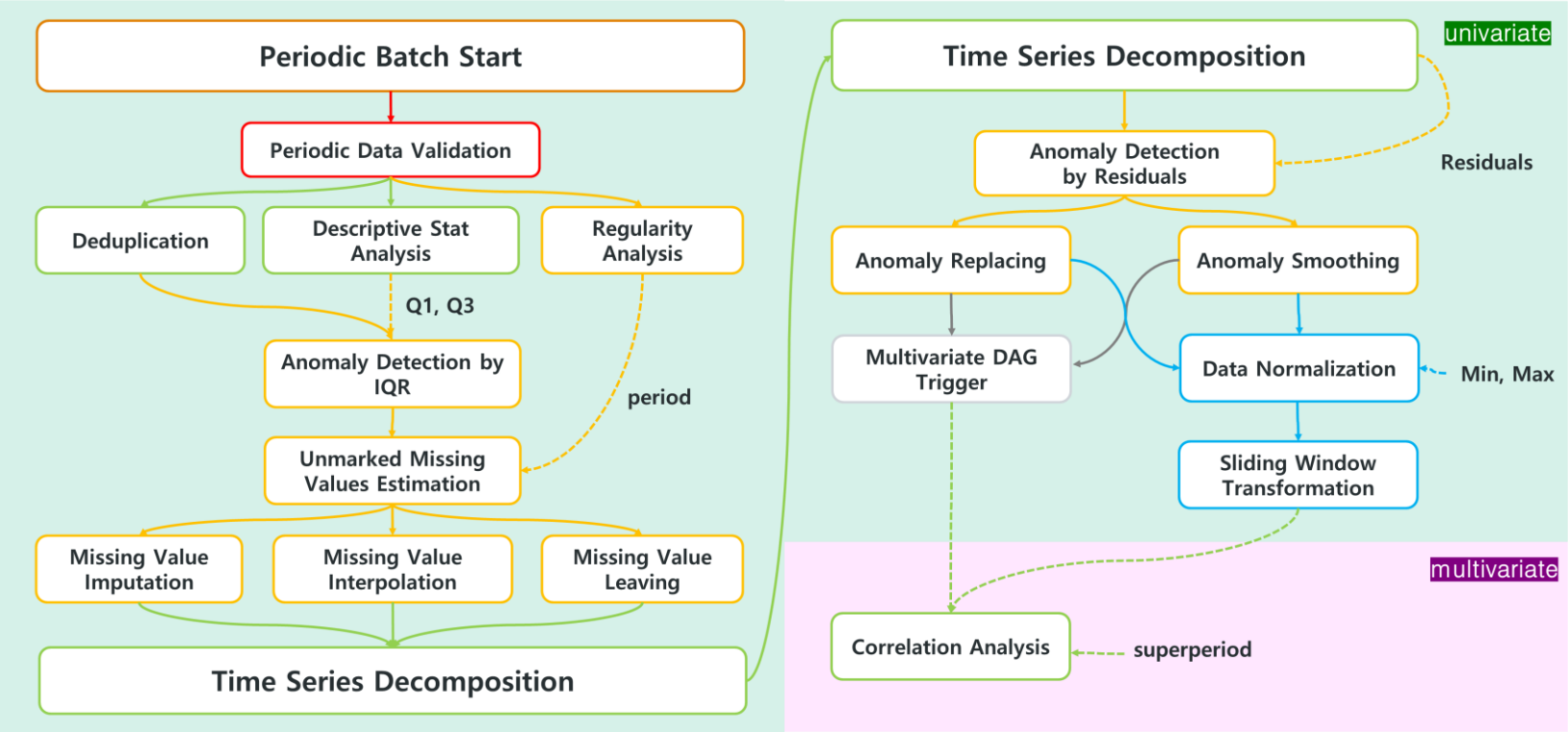


[다운샘플링]



[업샘플링]

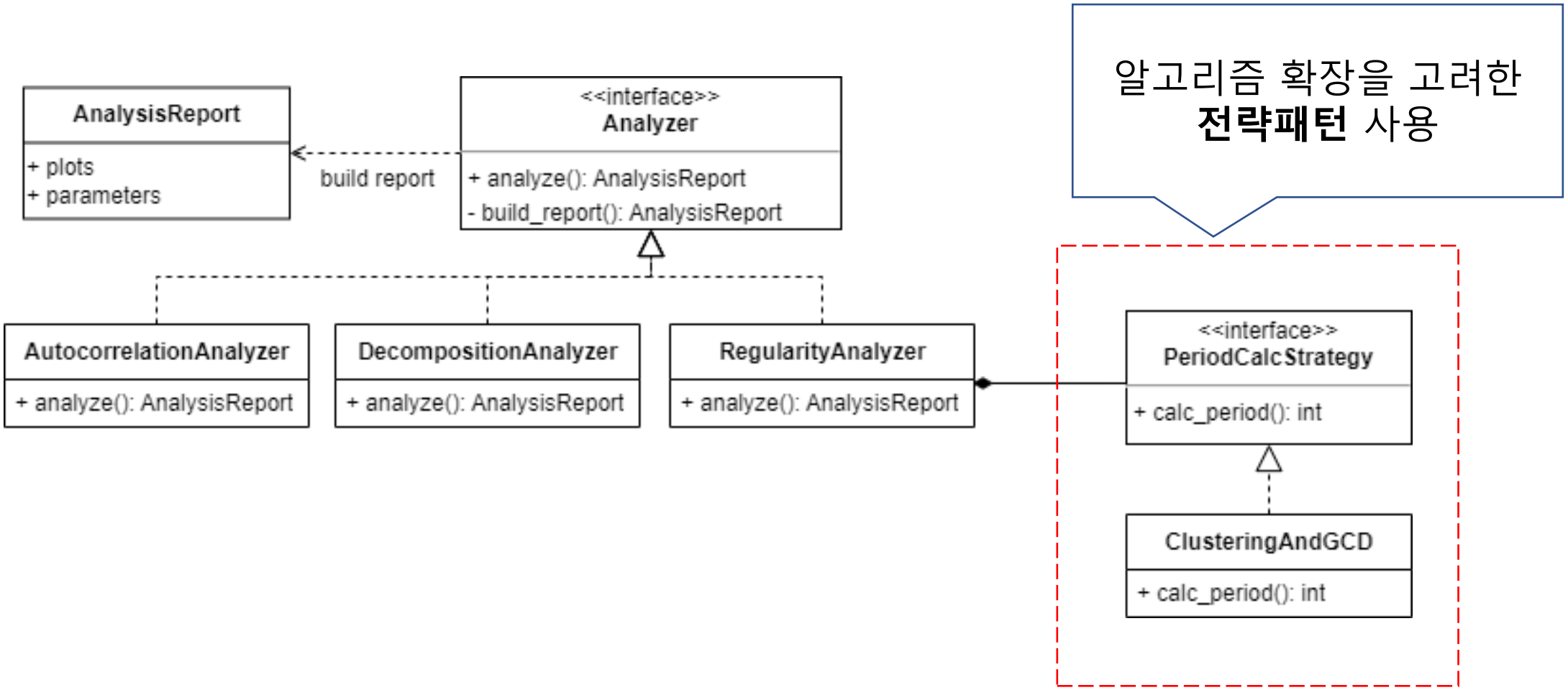
Airflow를 통한 주기적 수행



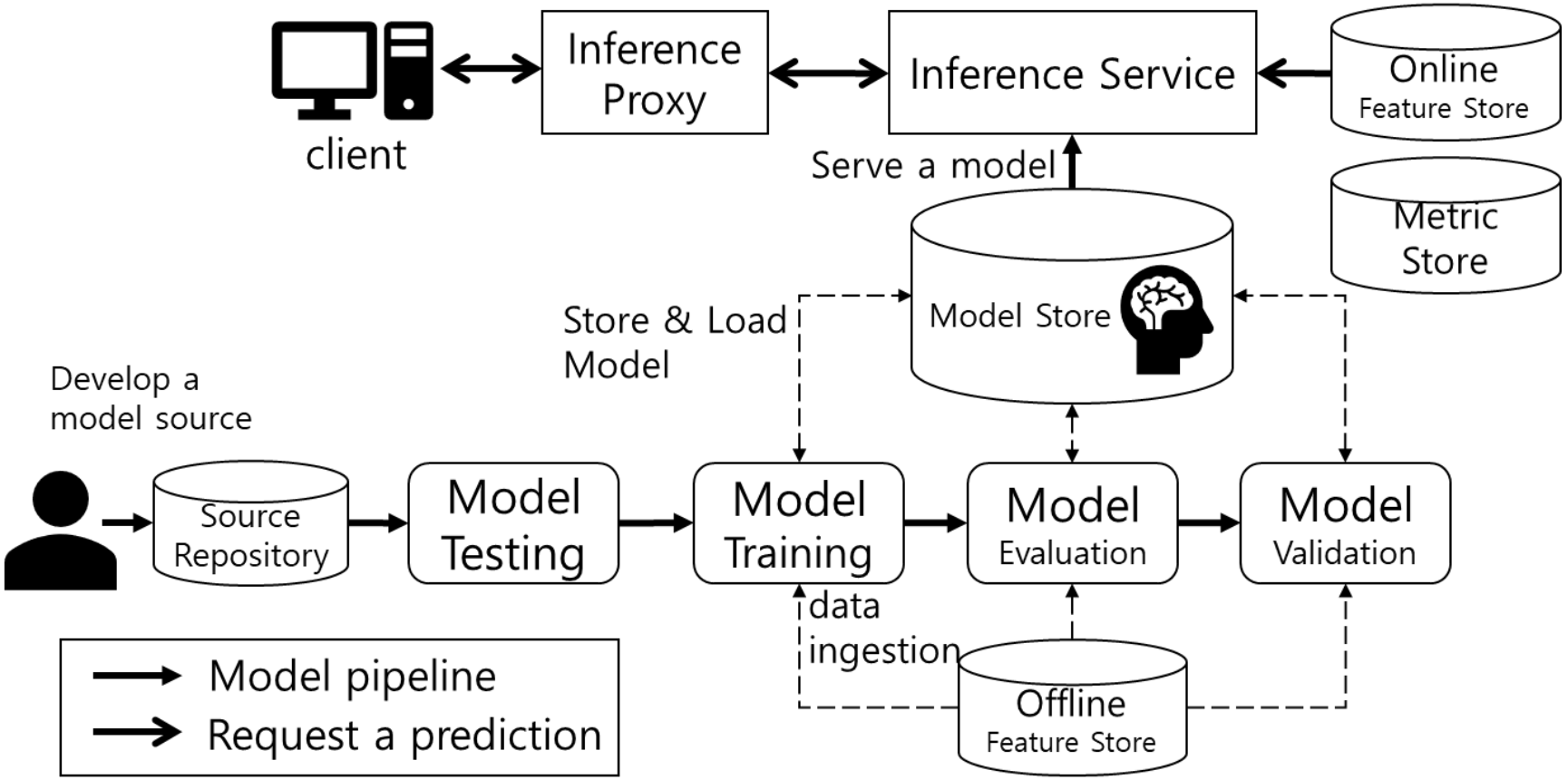
[DAG]

Apache Airflow

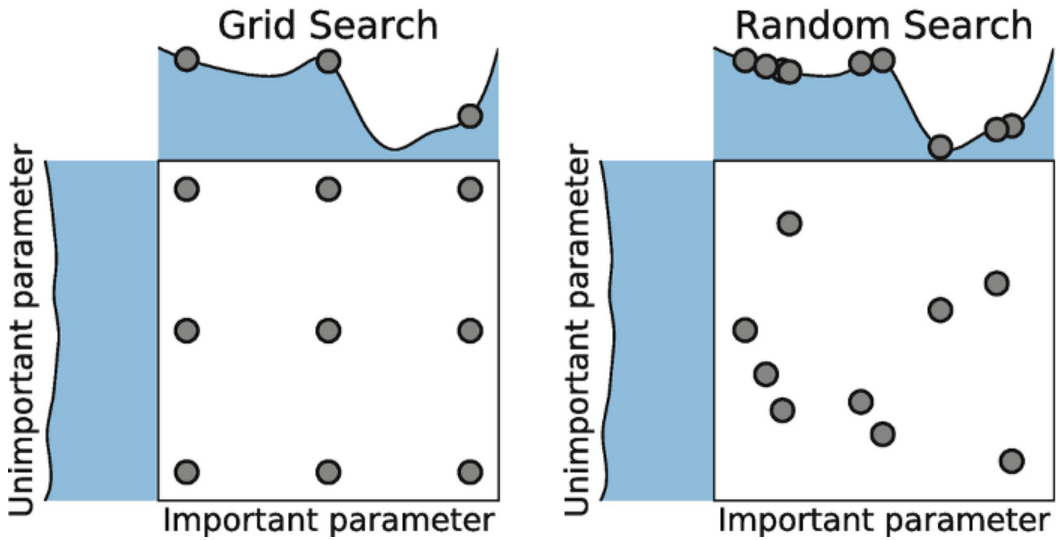
시계열 분석 라이브러리


AnalysisReport
+ plots
+ parameters
build report
<<interface>>
Analyzer
+ analyze(): AnalysisReport
- build_report(): AnalysisReport
AutocorrelationAnalyzer
+ analyze(): AnalysisReport
DecompositionAnalyzer
+ analyze(): AnalysisReport
RegularityAnalyzer
+ analyze(): AnalysisReport
알고리즘 확장을 고려한
전략패턴 사용
<<interface>>
PeriodCalcStrategy
+ calc_period(): int
ClusteringAndGCD
+ calc_period(): int

ML 모델 파이프라인


client
Inference Proxy
Inference Service
Online Feature Store
Serve a model
Metric Store
Model Store
Store & Load Model
Develop a model source
Source Repository
Model Testing
Model Training
Model Evaluation
Model Validation
data ingestion
Offline Feature Store
Model pipeline
Request a prediction

하이퍼파라미터 탐색 - AutoML



Difficulty to search hyperparameters..

Keras, PyTorch, dmlc XGBoost, tune, OPTUNA, HYPEROPT, Ax, Dragonfly, SIGOPT, Nevergrad, HEBO, FLAML

AutoML with Ray tune!

하이퍼파라미터 탐색 - AutoML

```
search_space = {
    "hidden_size": hp.grid_search([2, 4, 8, 16, 32, 64]),
    "lr": hp.loguniform(0.001, 0.1),
    "num_layers": hp.grid_search([1, 2, 3, 4, 5])
}
```

[search space]

Tune Status

Current time:	2022-12-05 15:19:32
Running for:	00:09:48.53
Memory:	8.8/31.2 GiB

System Info

Using FIFO scheduling algorithm.
Resources requested: 4.0/4 CPUs, 0/0 GPUs, 0.0/15.6 GiB heap, 0.0/6.98 GiB objects (0.0/2.0 _reserved, 0.0/1.0 _mxnet_worker, 0.0/1.0 _mxnet_server)

Messages

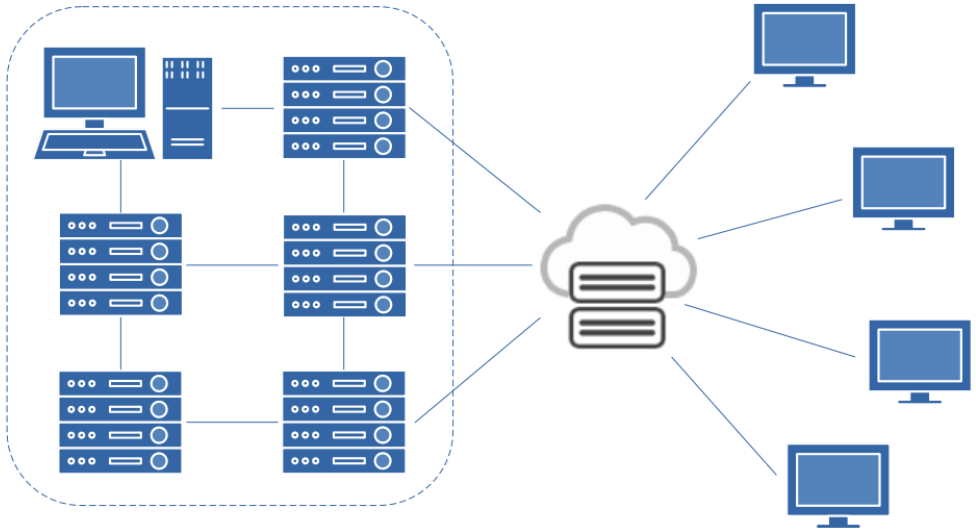
... 14 more trials not shown (8 PENDING, 6 TERMINATED)

Trial Status

Trial name	status	loc	hidden_size	lr	num_layers	iter	total time (s)	best_mse	mse
train_func_69a20_00014	RUNNING	163.180 [redacted] :20142	8	0.00105133	3				
train_func_69a20_00015	RUNNING	163.180 [redacted] :20074	16	0.0689328	3				
train_func_69a20_00016	RUNNING	163.180 [redacted] :19936	32	0.0162598	3				
train_func_69a20_00017	RUNNING	163.180 [redacted] :20141	64	0.0638407	3				
train_func_69a20_00018	PENDING		2	0.0686572	4				
train_func_69a20_00019	PENDING		4	0.00810866	4				
train_func_69a20_00020	PENDING		8	0.0525651	4				
train_func_69a20_00021	PENDING		16	0.0264697	4				
train_func_69a20_00022	PENDING		32	0.00103617	4				
train_func_69a20_00023	PENDING		64	0.0401975	4				
train_func_69a20_00024	PENDING		2	0.00423307	5				
train_func_69a20_00025	PENDING		4	0.00103699	5				
train_func_69a20_00000	TERMINATED	163.180 [redacted] :19936	2	0.0010417	1	1	69.993	0.00353771	0.00353771
train_func_69a20_00001	TERMINATED	163.180 [redacted] :20074	4	0.099259	1	1	94.3297	0.00412781	0.00412781
train_func_69a20_00002	TERMINATED	163.180 [redacted] :20141	8	0.00551675	1	1	70.74	0.00422874	0.00422874
train_func_69a20_00003	TERMINATED	163.180 [redacted] :20142	16	0.00673473	1	1	74.9543	0.00402848	0.00402848
train_func_69a20_00004	TERMINATED	163.180 [redacted] :19936	32	0.0567423	1	1	74.4295	0.0045467	0.0045467
train_func_69a20_00005	TERMINATED	163.180 [redacted] :20141	64	0.00515913	1	1	78.1486	0.00415297	0.00415297
train_func_69a20_00006	TERMINATED	163.180 [redacted] :20142	2	0.00392652	2	1	130.156	0.00412261	0.00412261
train_func_69a20_00007	TERMINATED	163.180 [redacted] :20074	4	0.0066329	2	1	127.054	0.00418443	0.00418443

[AutoML hyperparameter estimating]

향후 연구



시계열 데이터를 뛰어넘어 이미지, 청각 등 멀티모달 데이터를 융합하는 MLOps 확장

엣지 및 코어를 활용한 분산 클라우드에서 효율적인 빅데이터 파티셔닝 및 연합학습

참조문헌

[1] F. Castanedo, "A review of data fusion techniques," in *ScientificWorldJournal*, vol. 2013, 2013.

[2] D. L. Hall and J. Llinas, "An introduction to multisensor data fusion," in *Proceedings of the IEEE*, vol. 85, 1997.

[3] H. Widiputra, R. Pears, N. Kasabov, "Multiple Time-Series Prediction through Multiple Time-Series Relationships Profiling and Clustered Recurring Trends, " In *Advances in Knowledge Discovery and Data Mining*, *PAKDD 2011*, 2011.

[4] G. Gürses-Tran and M. Antonello, "Advances in Time Series Forecasting Development for Power Systems' Operation with MLOps," in *Forecasting*, vol. 4, 2022.

[5] J. Benesty, J. Chen, Y. Huang, I. Cohen, I. "Pearson Correlation Coefficient". In *Noise Reduction in Speech Processing. Springer Topics in Signal Processing*, vol. 2, 2009.

[6] D. Sculley, et al. "Hidden Technical Debt in Machine Learning Systems," in *NIPS*, 2015.