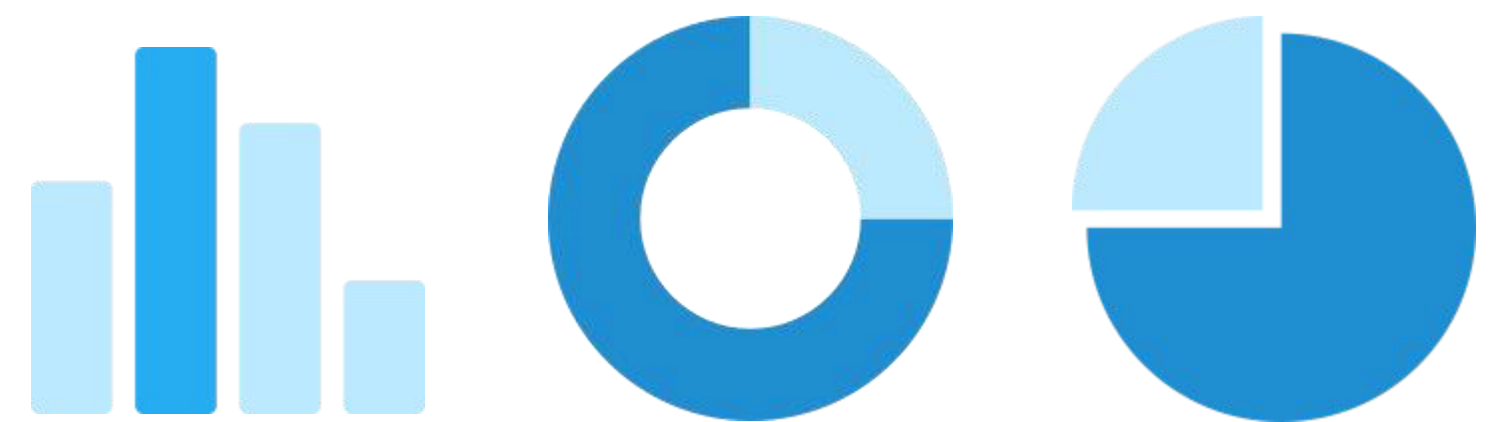

데이터 엔지니어링 파이프라인 구축



CONTENTS

As_df

Part1. 팀원 소개 및 역할 분담

Part2. 주제 소개

Part3. 데이터 엔지니어의 업무

Part4. 사용 솔루션 및 데이터 소개

Part5. 파이프라인 상세

Part6. 결론 및 후속 과제

Part 1.

팀원 소개 및 역할 분담

Part 1. 팀원 소개 및 역할 분담

As_DF



진광환

팀원

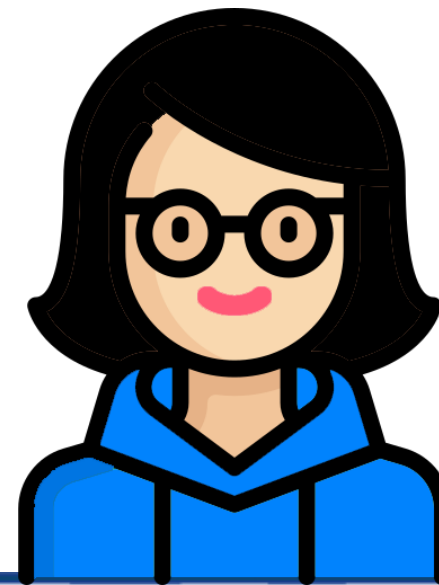
Role

데이터 전처리

Pyspark코드 구현

발표자료 준비

발표



박은영

팀원

Role

데이터 시각화

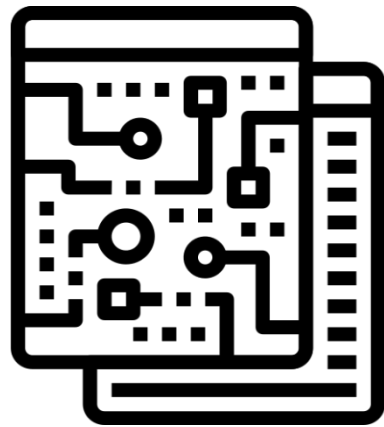
Nifi migration 구현

발표자료 준비

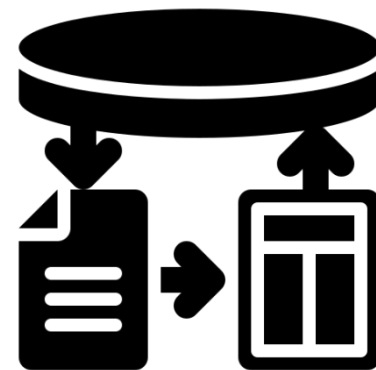
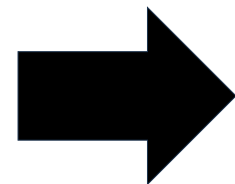
Part 2.

주제 소개

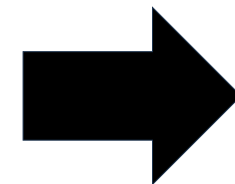
"S 제조사 반도체 라인 엔지니어의 데이터기반 의사결정을 돕는 시스템 구축"



하루 8PB 데이터 생성



ETL 시스템



의사 결정



Part 3.

데이터 엔지니어의 업무

Part 3. 데이터 엔지니어의 업무

1. 최종 데이터 사용자의 요구 사항 분석
2. ETL 파이프라인에서 데이터플랫폼 구축에 이르는 데이터 흐름 과정 설계

Collection
수집

Ingestion
저장

Flow
흐름관리

Streaming
Analysis
스트리밍 분석

Data At
Rest
제출용 데이터
제작



Part 3. 데이터 엔지니어의 업무

1. 최종 데이터 사용자의 요구 사항 분석
2. ETL 파이프라인에서 데이터플랫폼 구축에 이르는 데이터 흐름 과정 설계





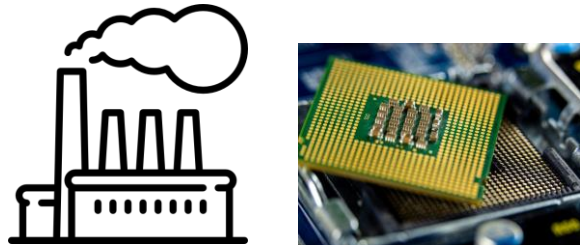
Part 4.

사용 솔루션 및 데이터 소개

Part 4. 사용 솔루션 및 데이터 소개 - Data Flow

앱에서 발생한 데이터 → 최종 데이터사용자

Edge data



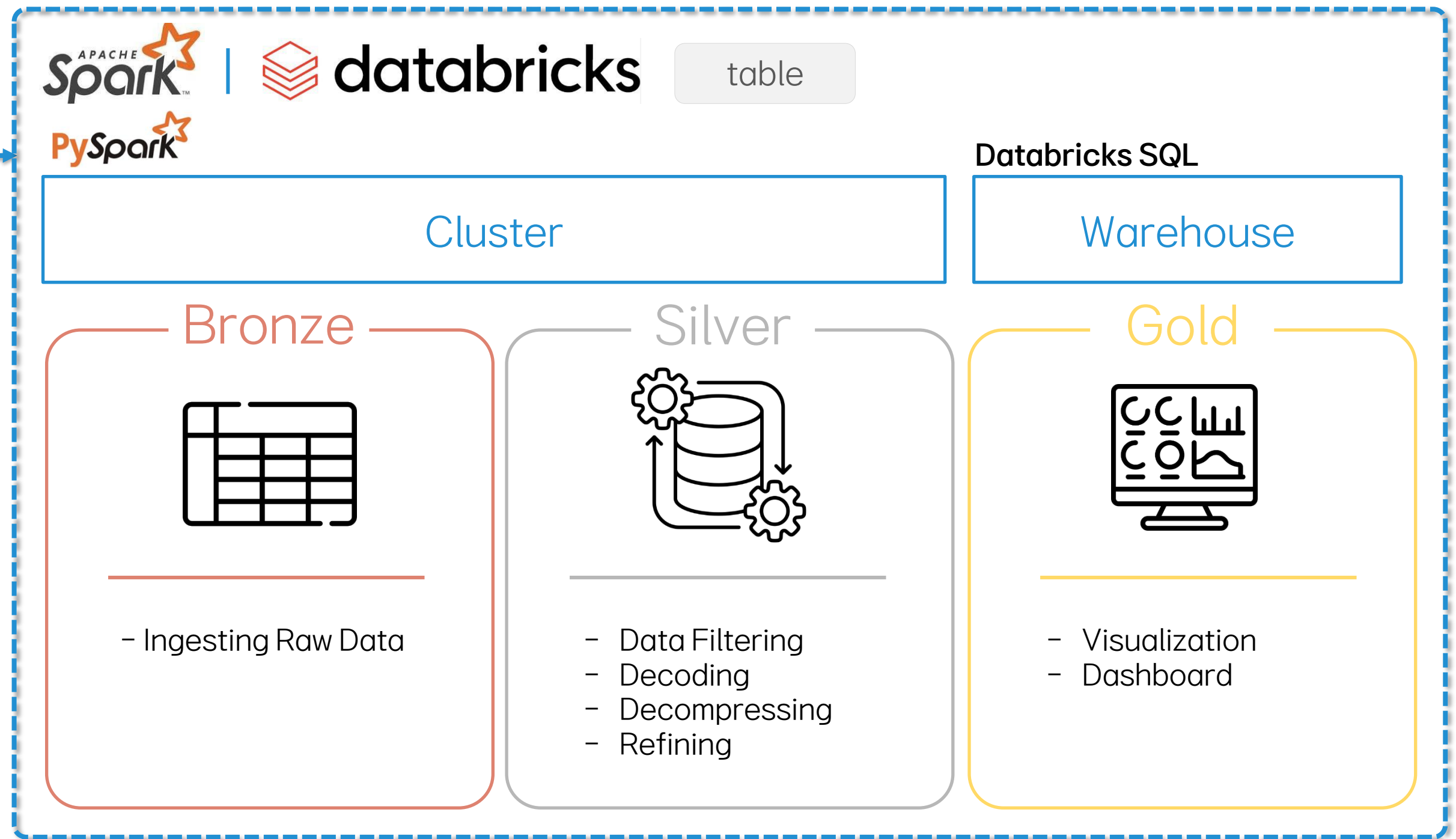
Collect



Migrate



Process



Part 4. 사용 솔루션 및 데이터 소개 - Apache Nifi



- 대량의 데이터를 Data Flow 프로세스로 구축, 유지, 교환하는 시스템
- GUI를 통해 데이터 흐름을 눈으로 직접 보고 ETL프로세스를 설계할 수 있다.

Collection
수집

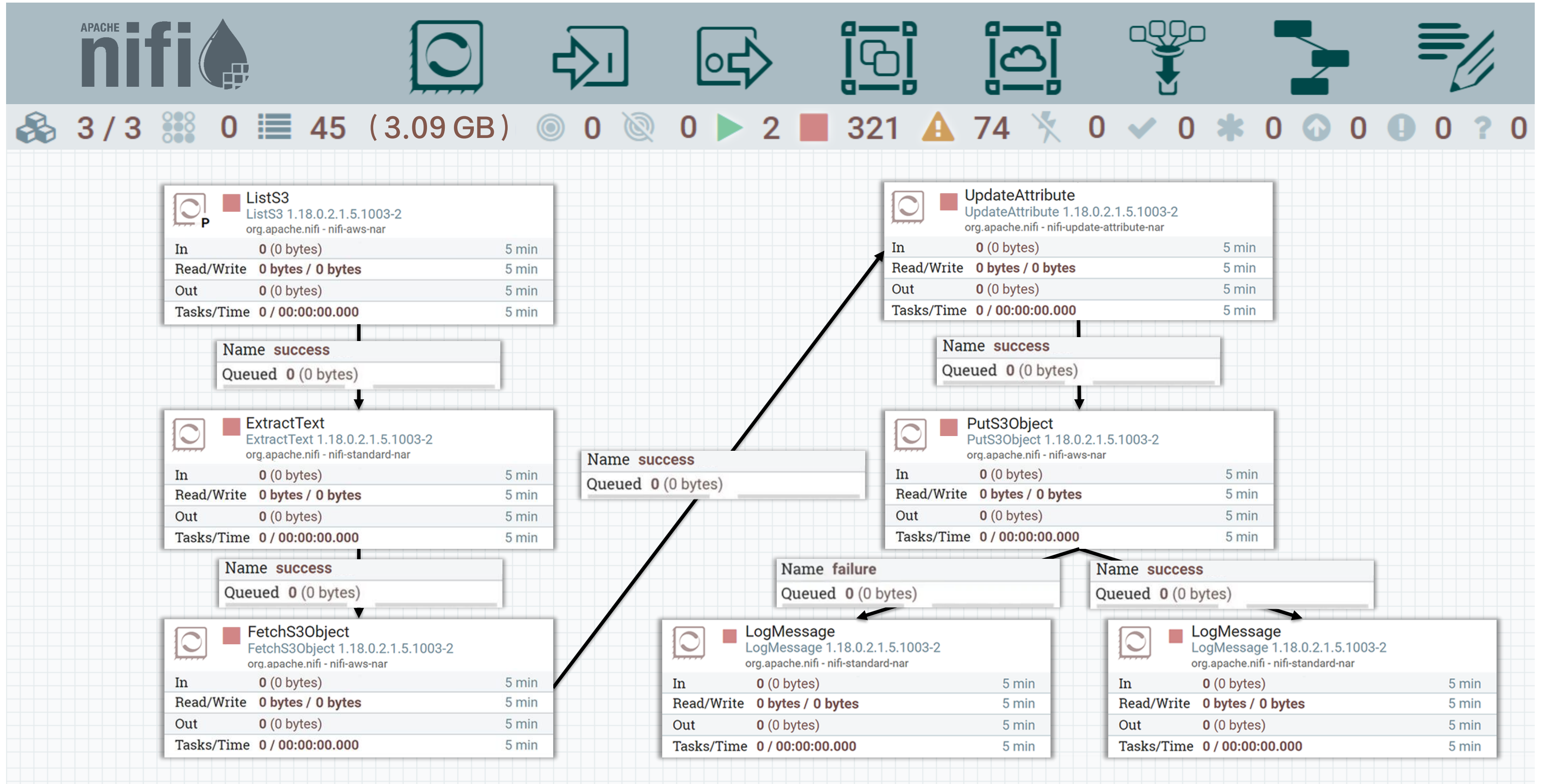
Ingestion
저장

Flow
흐름관리

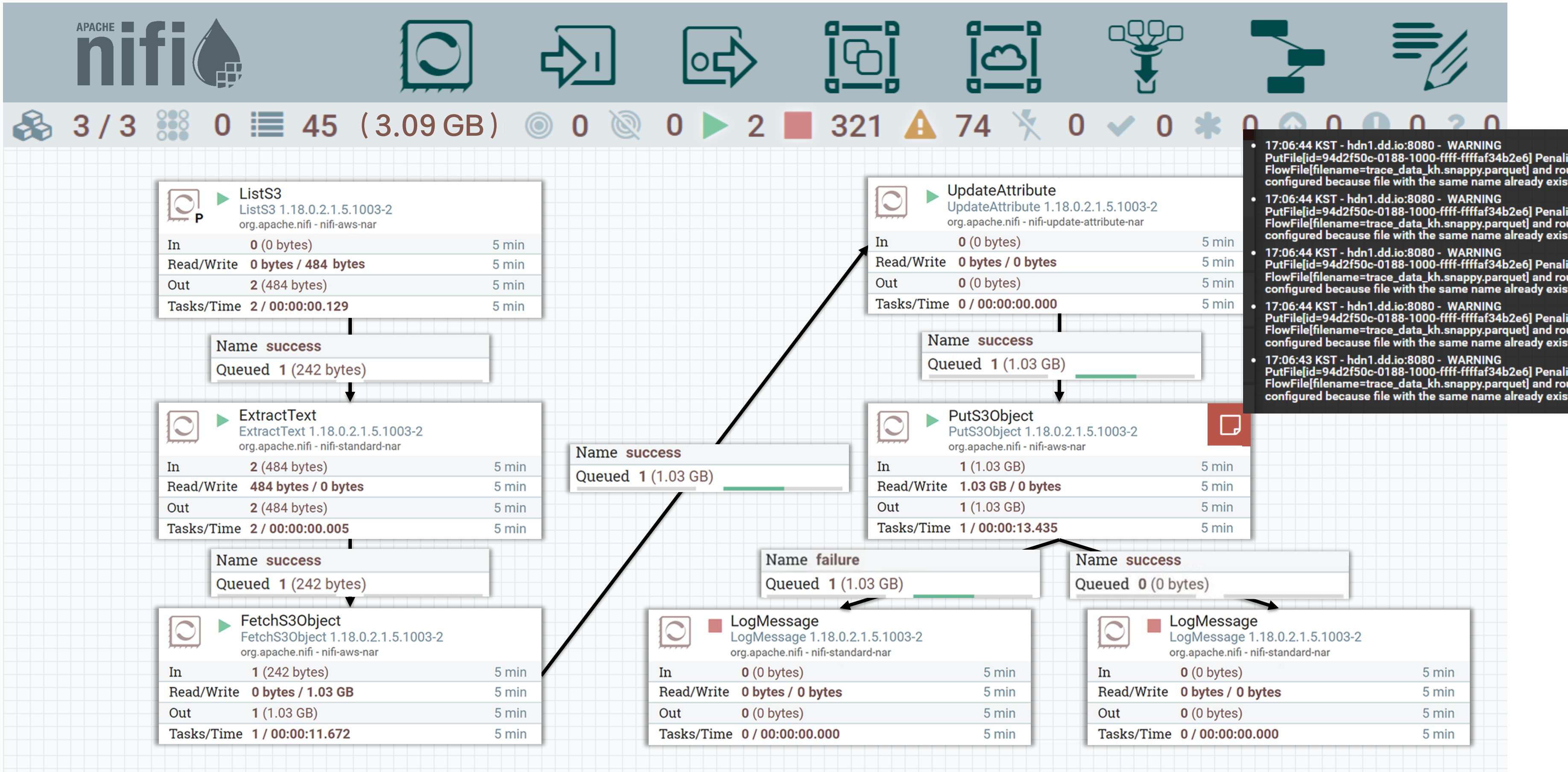
Streaming
Analysis
스트리밍
분석

Data At
Rest
제출용
데이터 제작

Part 4. 사용 솔루션 및 데이터 소개 - Nifi GUI



Part 4. 사용 솔루션 및 데이터 소개 - Nifi GUI



Part 4. 사용 솔루션 및 데이터 소개 - Apache Spark



- SQL, 머신러닝, 시각화, 스트리밍
- 다중언어 지원
- 인메모리 방식 - 빠른 처리속도
- Dataframe 사용

Collection
수집

Ingestion
저장

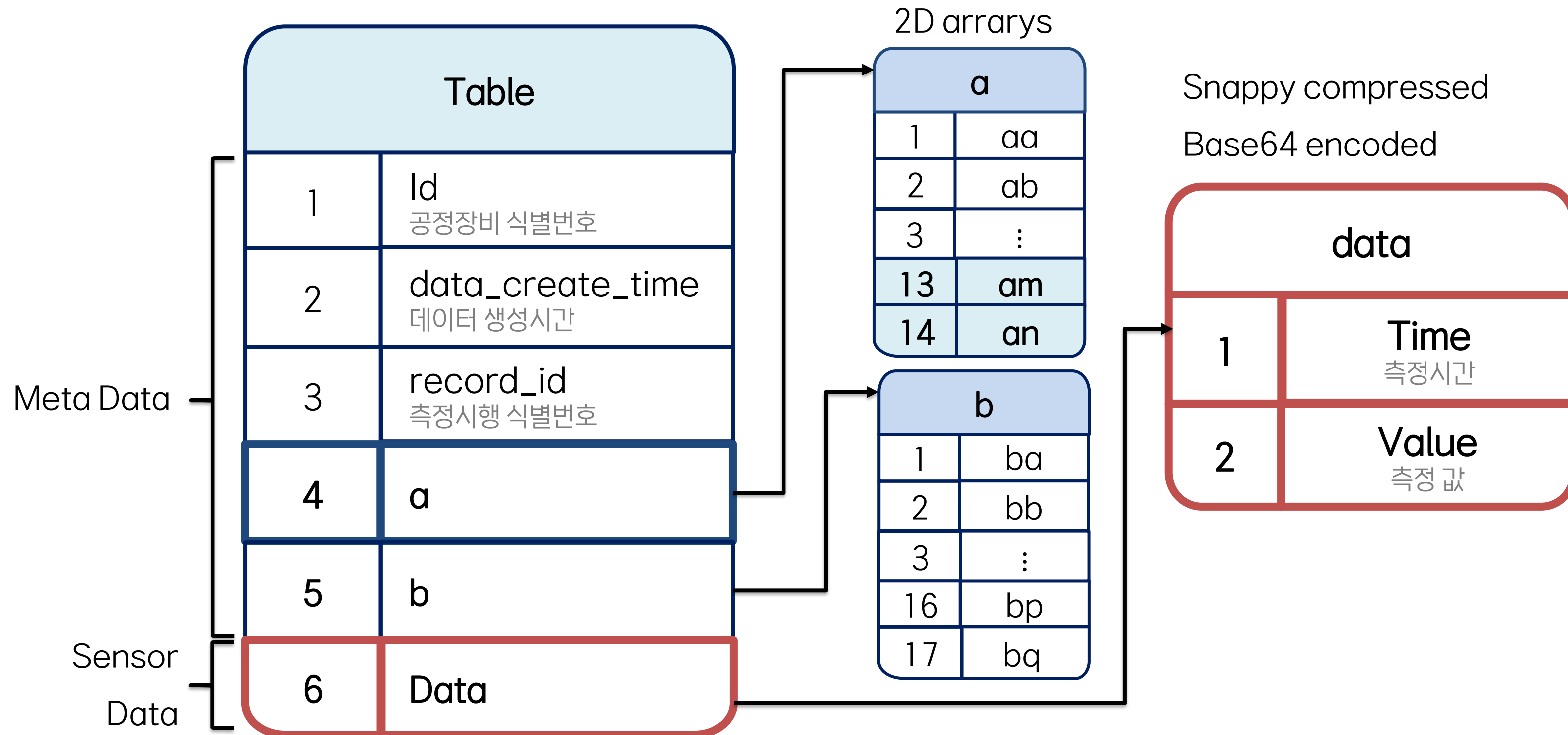
Flow
흐름관리

Streaming
Analysis
스트리밍
분석

Data At
Rest
제출용
데이터 제작

Part 4. 사용 솔루션 및 데이터 소개 - 데이터 소개

“임의로 제작된 반도체 공정장비 센서 측정데이터”





Part 5.

파이프라인 상세



Part 5. 파이프라인 상세 - Nifi를 활용한 Migration system

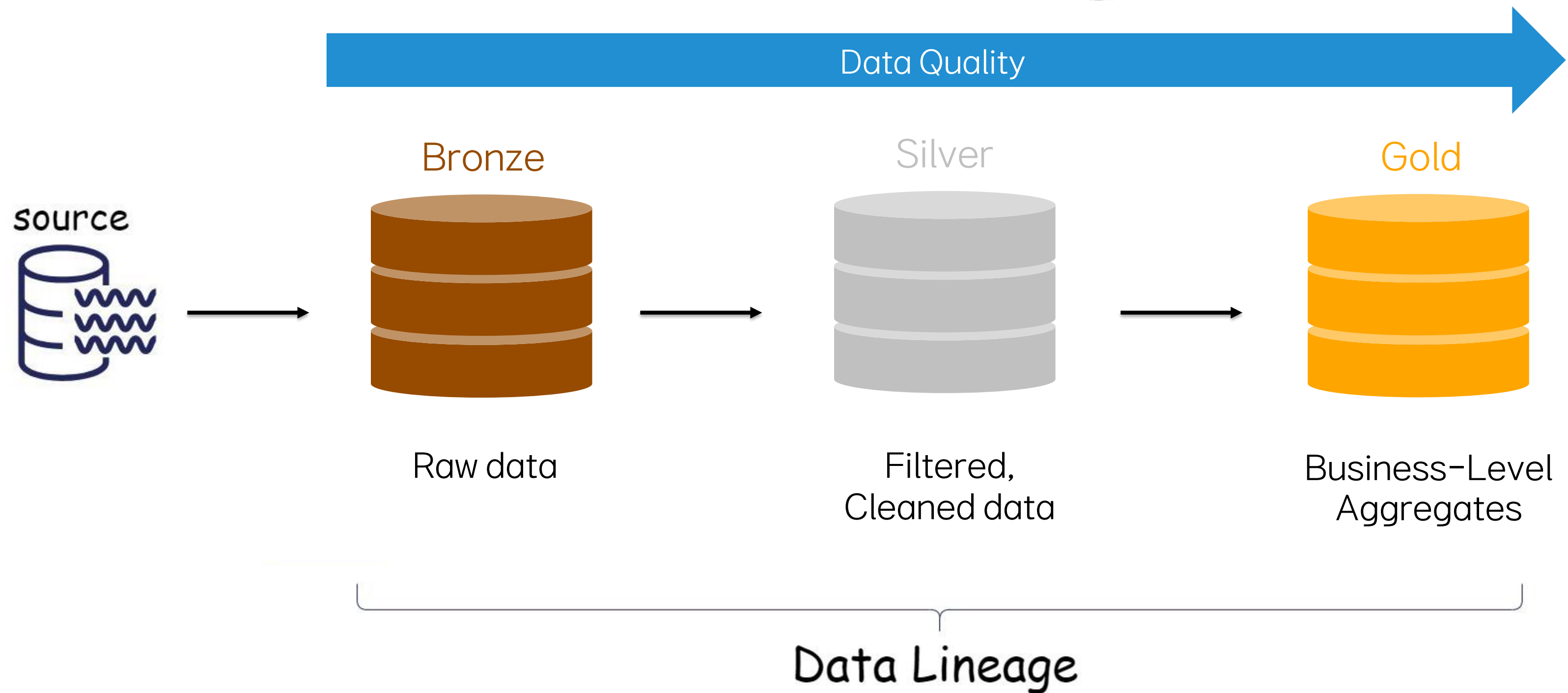


Part 5. 파이프라인 상세 - Medallion Architecture

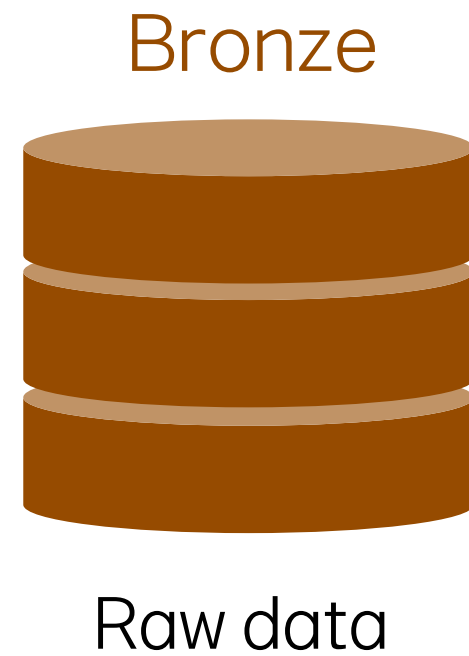


databricks

ETL Processing



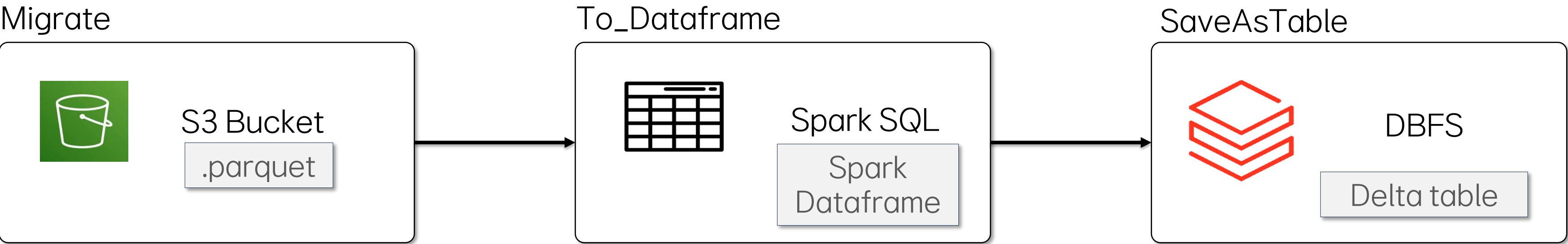
Part 5. 파이프라인 상세 – Bronze Layer



Bronze Layer

- 원천 데이터 저장

Part 5. 파이프라인 상세 - Bronze Layer



id	data_create_time	record_id	a	b	data
7VBbJa4	2027-09-18 23:16	3890904313625360000	{AAA, 449828909, ...	{AAAA, YW8Q, b9Lb...	ia0WIDlwMjcwOTE5M...
I7QS5WL	2032-01-06 3:03	1919815607339010000	{AAA, 993055219, ...	{AAAA, T5IN, b9Lb...	scUhIDlwMzlwMTA2M...
PWaQ00W	2024-12-19 10:14	4403797140526970000	{AAA, 921859806, ...	{AAAA, T5IN, 2IJV...	tpsrIDlwMjQxMjE5M...
7VBbJa4	2027-09-19 23:11	3890904313625360000	{AAA, 449828909, ...	{AAAA, YW8Q, b9Lb...	lfoOHDlwMjcwOTlwF...
I7QS5WL	2032-01-06 3:08	1919815607339010000	{AAA, 993055219, ...	{AAAA, T5IN, b9Lb...	qPMmIDlwMzlwMTA2M...
7VBbJa4	2027-09-20 15:31	8972448469089840000	{AAA, 1836905390,...	{AAAA, RMT3, 2IJV...	n/UuIDlwMjcwOTlxM...
SX6QCPQ	2025-05-15 3:08	5368658724164300000	{AAA, 1555313061,...	{AAAA, T5IN, 2IJV...	mecsIDlwMjUwNTE1M...
Z9CW9LT	2032-11-13 14:15	3084574033054940000	{AAA, 394001883, ...	{AAAA, C9FZ, b9Lb...	3bMSIDlwMzlxMTEzM...
7VBbJa4	2027-09-20 15:36	8972448469089840000	{AAA, 1836905390,...	{AAAA, RMT3, 2IJV...	8dMgIDlwMjcwOTlxM...
9W32VNB	2030-07-10 3:39	7401457008662360000	{AAA, 397410082, ...	{AAAA, LR9Q, b9Lb...	upkMHDlwMzAwNzEwF...
OQVPE2O	2023-09-08 1:47	859270735969460000	{AAA, 1032017508,...	{AAAA, V6D7, N8G7...	xvQlIDlwMjMwOTA4M...
:	:	:	:	:	:

11671 records

Part 5. 파이프라인 상세 - Silver Layer



Silver Layer

- Data filtering : 조건을 만족하는 데이터 선별
- Decoding & Decompressing : 인코딩된 데이터 해석
- Refining : 해석된 데이터 정형화

Part 5. 파이프라인 상세 – SilverLayer

data	
1	Time 측정시간
2	Value 측정값

```
data
ia0WIDlwMjcwOTE5M...
scUhIDlwMzlwMTA2M...
tpsrIDlwMjQxMjE5M...
lfoOHDlwMjcwOTlwF...
qPMmIDlwMzlwMTA2M...
n/UuIDlwMjcwOTlxM...
mecslDlwMjUwNTE1M...
3bMSIDlwMzlxMTEzM...
8dMgIDlwMjcwOTlxM...
upkMHDlwMzAwNzEwF...
xvQlIDlwMjMwOTA4M...
jv8YIDlwMjUwNTE1M...
⋮
```

11671 records

Data Filtering



NaN value 수 / 전체 value 수
< 0.001

```
data
8/kvIDlwMjMwNjAyM...
uqwbHDlwMjYwMTEwF...
iJYaIDlwNDAwNDEzM...
yugkIDlwNTAwNDA0M...
45cjHDlwMjQwMjEwF...
5cMiIDlwMzQxMDEyM...
mlksHDlwMjcwNjlwF...
5bodIDlwMjMwNzAxM...
zZsgIDlwNTMxMDA3M...
```

9 records

Part 5. 파이프라인 상세 – Silver Layer

data

8/kvIDlwMjMwNjAyM...
uqwbHDIwMjYwMTEwF...
iJYaIDlwNDAwNDEzM...
yugkIDlwNTAwNDA0M...
45cjHDIwMjQwMjEwF...
5cMiIDlwMzQxMDEyM...
mlksHDIwMjcwNjIwF...
5bodIDlwMjMwNzAxM...
zZsgIDlwNTMxMDA3M...

```
base64 decode
snappy decompress
```

data
202306020000000000...
202601100000000000...
204004130000000000...
205004040000000000...
202402100000000000...
203410120000000000...
202706200000000000...
202307010000000000...
203607050000000000...

‘|’ (pipe)으로 행간 구분

‘^’ (caret)으로 열간 구분

[illegible]

Part 5. 파이프라인 상세 – Silver Layer

‘^’와 ‘|’를 기준으로 열과 행 분할

Time : { "20230511000000500", "20230511000001200", "20230511000001400", ... },

Value : { "2629.7733705416863", "2631.8555945216085", "2632.25077152991" ... }

2차원 배열 형태로 정렬

[('20230511000000500',
'2629.7733705416863'),
(('20230511000001200',
'2631.8555945216085'),
(('20230511000001400',
'2632.25077152991')]

Schema 변환, Dataframe 생성

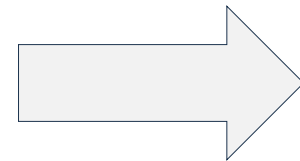
‘Time’ : string -> timestamp

‘Value’ : string -> double

Part 5. 파이프라인 상세 - Silver Layer

장비 id 별 시간에 따른 공정센서수치 데이터

id	data
4PEE803	202306020000000000...
H6VMCFC	202601100000000000...
1G8TW3J	204004130000000000...
CW9J9VG	205004040000000000...
MQ2FWMH	202402100000000000...
XEVbYST	203410120000000000...
NaGXJE1	202706200000000000...
ES5EZb9	202307010000000000...
XEVbYST	203607050000000000...
KaNM5T0	202406110000000000...



	time	value	id
1	2022-06-02T00:00:00.000+0000	1.6392934948532836	4PEE803
2	2022-06-02T00:00:00.100+0000	2.485428742425379	4PEE803
3	2022-06-02T00:00:00.200+0000	1.7348385168264062	4PEE803
4	2022-06-02T00:00:00.300+0000	1.8832493252054139	4PEE803
5	2022-06-02T00:00:00.400+0000	1.6364314883332791	4PEE803
6	2022-06-02T00:00:00.500+0000	0.6951457781897982	4PEE803
7	2022-06-02T00:00:00.600+0000	0.46978393209337277	4PEE803
8	2022-06-02T00:00:00.700+0000	0.21568359899916456	4PEE803
9	2022-06-02T00:00:00.800+0000	0.18318897368080378	4PEE803
10	2022-06-02T00:00:00.900+0000	0.45219805309115646	4PEE803
11	2022-06-02T00:00:01.000+0000	-0.2781803320109302	4PEE803
12	2022-06-02T00:00:01.100+0000	0.1136606978860335	4PEE803

9 records

9 tables

Part 5. 파이프라인 상세 – Gold Layer

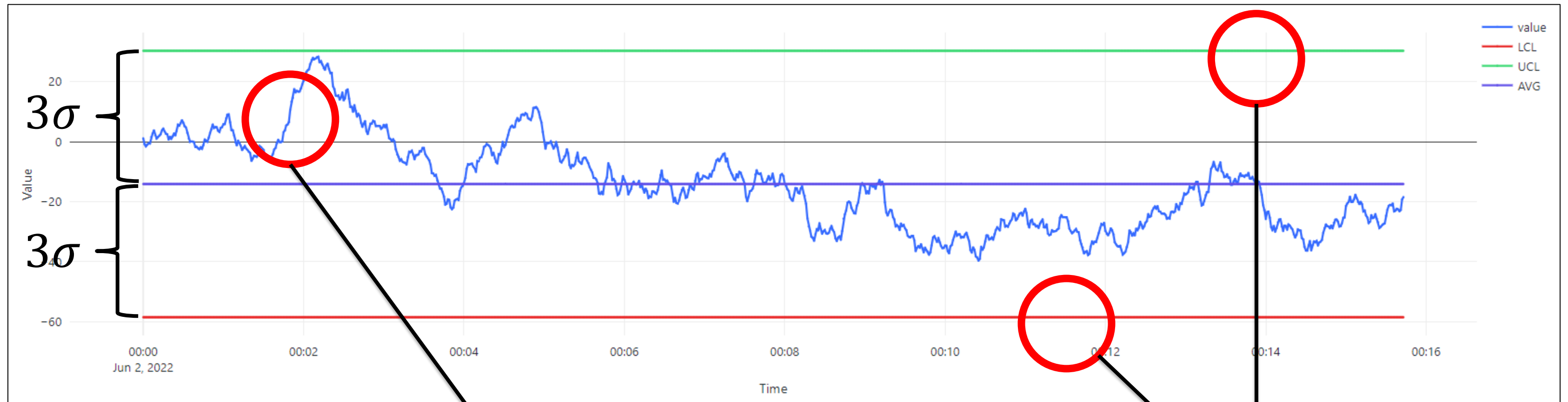


Business-Level
Aggregates

Gold Layer

- Visualization
- 보고용 Dashboard 제작

Part 5. 파이프라인 상세 - Gold Layer



시간에 따른 반도체 센서 데이터 도시

LCL UCL 산출

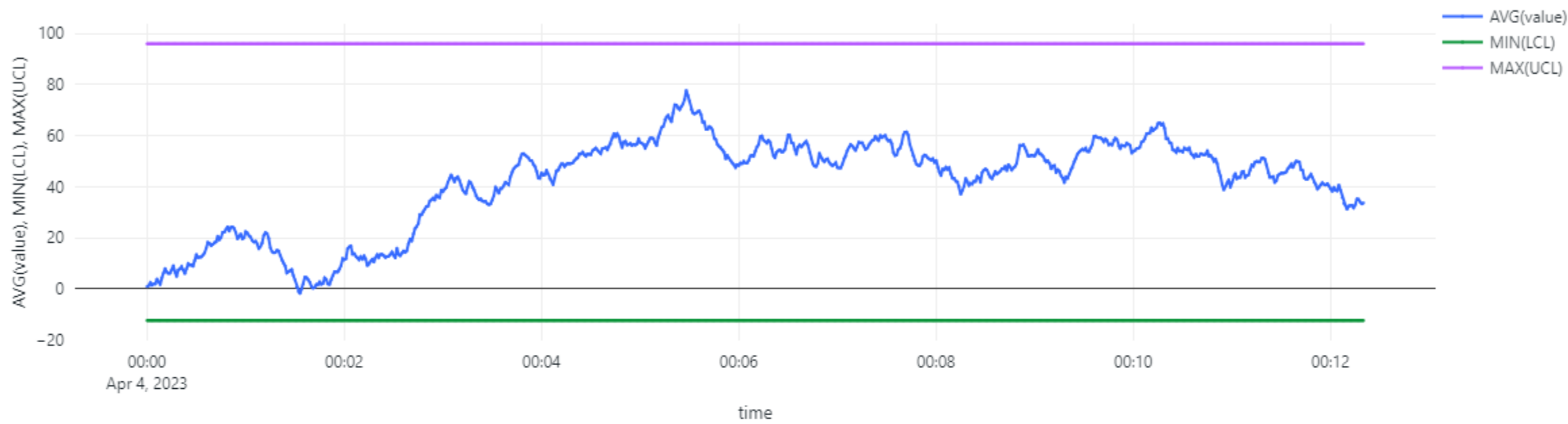


LCL / UCL이란?

통계적 프로세스 제어에서 관리할 수 있는 최저/최고 허용 한계

Equipment Operation

Line 1 - Gold_CW9J9VG



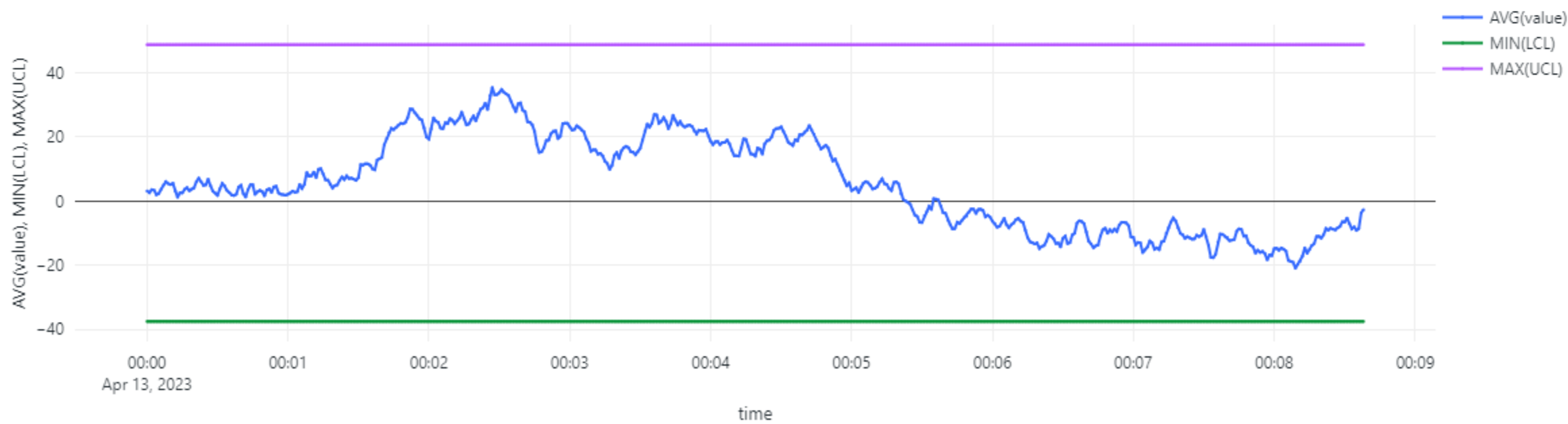
7 minutes ago

Results - Gold_CW9J9VG

time	value	id	LCL	UCL
2023-04-04 00:00:00.000	1.51	CW9J9VG	-12.36	
2023-04-04 00:00:00.100	1.61	CW9J9VG	-12.36	
2023-04-04 00:00:00.200	1.22	CW9J9VG	-12.36	
2023-04-04 00:00:00.300	0.70	CW9J9VG	-12.36	
2023-04-04 00:00:00.400	1.20	CW9J9VG	-12.36	
2023-04-04 00:00:00.500	0.27	CW9J9VG	-12.36	
2023-04-04 00:00:00.600	0.10	CW9J9VG	-12.36	

7 minutes ago

Line 1 - Gold_1G8TW3J



7 minutes ago

Results - Gold_1G8TW3J

time	value	id	LCL	UCL
2023-04-13 00:00:01.300	3.41	1G8TW3J	-37.48	
2023-04-13 00:00:01.400	3.07	1G8TW3J	-37.48	
2023-04-13 00:00:01.500	3.01	1G8TW3J	-37.48	
2023-04-13 00:00:01.600	2.07	1G8TW3J	-37.48	
2023-04-13 00:00:01.700	1.79	1G8TW3J	-37.48	
2023-04-13 00:00:01.800	1.85	1G8TW3J	-37.48	

7 minutes ago

Line 1 - Gold_H6VMCFC

Results - Gold_H6VMCFC

Equipment Operation

Line 1 - Gold_CW9J9VG

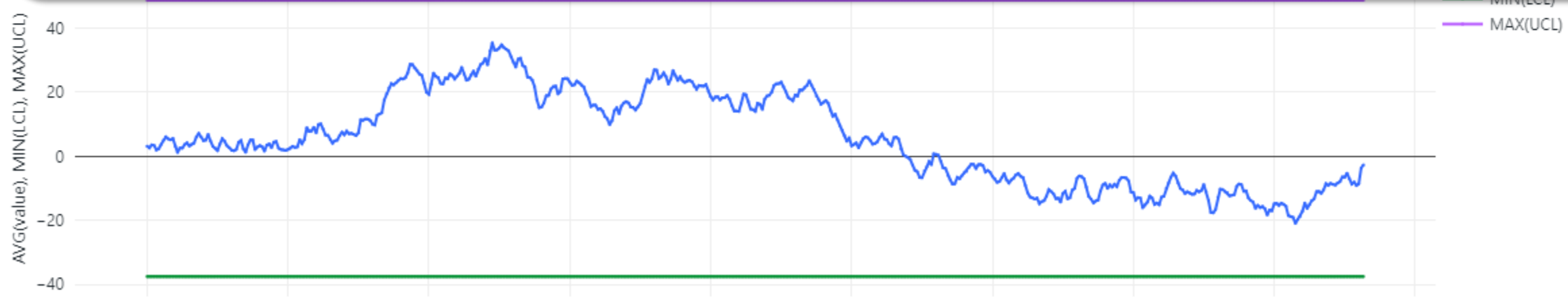


Results - Gold_CW9J9VG

time	value	id	LCL	UCL
2023-04-04 00:00:00.000	1.51	CW9J9VG	-12.36	
2023-04-04 00:00:00.100	1.61	CW9J9VG	-12.36	
2023-04-04 00:00:00.200	1.22	CW9J9VG	-12.36	



- 스케줄링, batch분석 수행
- BI도구와 연동, 비즈니스 인사이트 산출



2023-04-13 00:00:01.300	3.41	1G8TW3J	-37.48	
2023-04-13 00:00:01.400	3.07	1G8TW3J	-37.48	
2023-04-13 00:00:01.500	3.01	1G8TW3J	-37.48	
2023-04-13 00:00:01.600	2.07	1G8TW3J	-37.48	
2023-04-13 00:00:01.700	1.79	1G8TW3J	-37.48	
2023-04-13 00:00:01.800	1.85	1G8TW3J	-37.48	

7 minutes ago

7 minutes ago

Line 1 - Gold_H6VMCFC

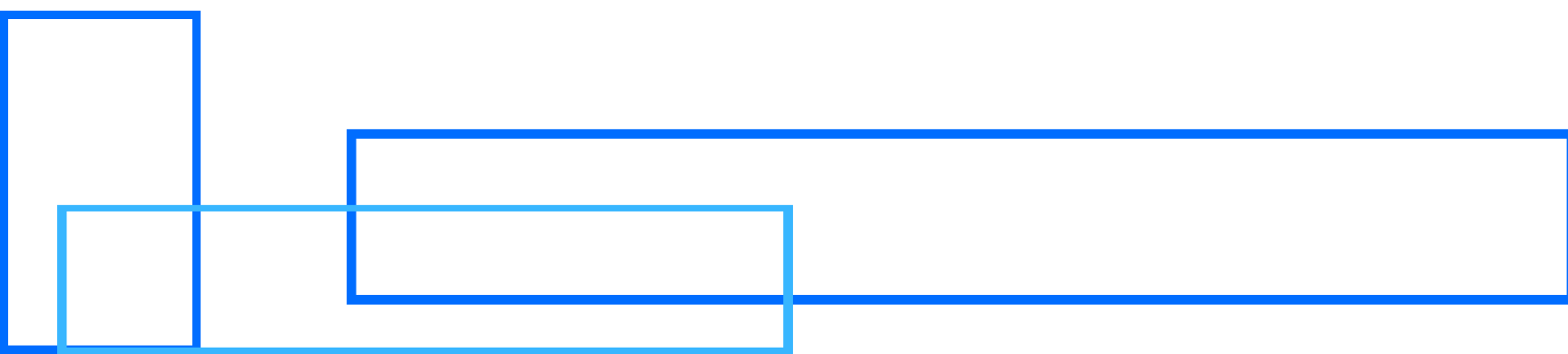
Results - Gold_H6VMCFC

time	value	id	LCL	UCL
------	-------	----	-----	-----



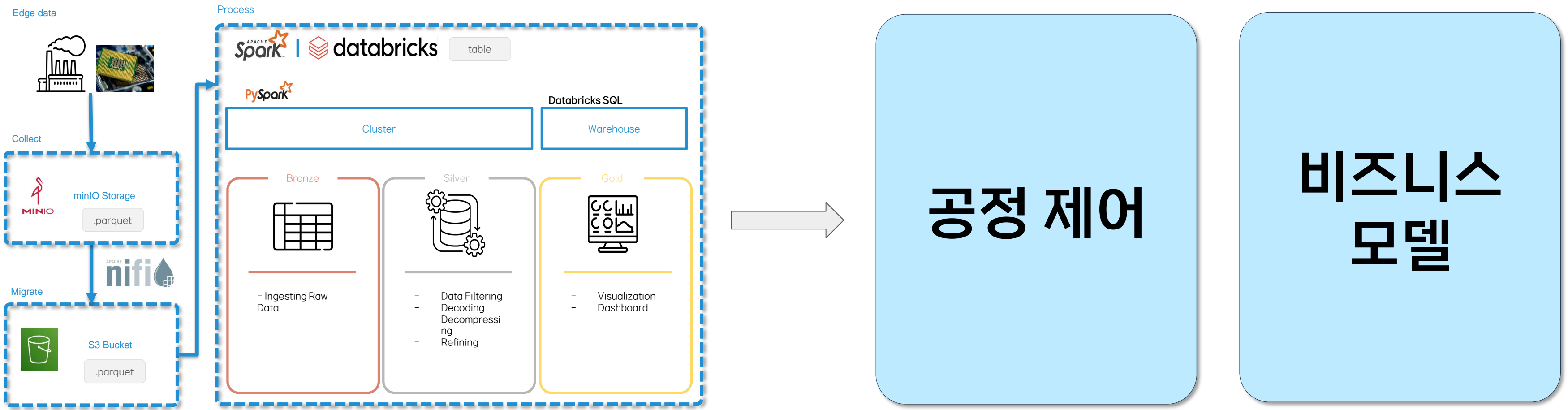
Part 6.

결론 및 후속과제



Part 6. 결론 및 후속과제

결론 - “어떻게 공헌할 것인가”



후속과제



BI도구와 연계
실제 현장에서 산출되는 데이터를 실시간으로 처리 할 수 있는 시스템 구축

THANK YOU!

