



# AWS 기반 실시간 데이터 파이프라인 구축하기

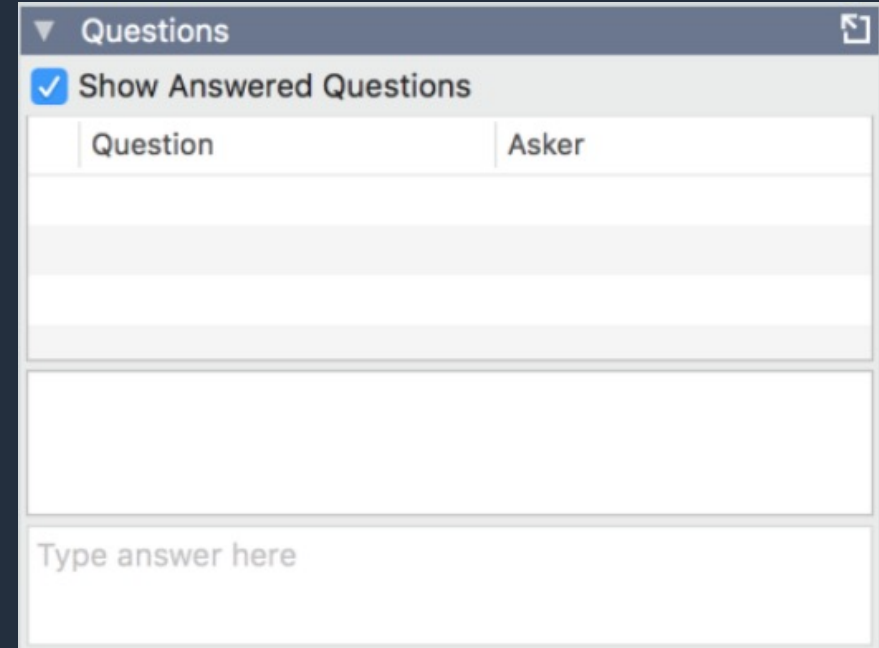
## *AWS Builders Korea - Analytics*

Jinsung Huh  
Solutions Architect

# 강연 중 질문하는 방법

AWS Builders Korea Go to Webinar “Questions (질문)” 창에 자신이 질문한 내역이 표시됩니다. 본인만 답변을 받고 싶으실 경우 (비공개)라고 하고 질문해 주시면 됩니다.

질문 주신 사항에 대해서는 질문창을 통해 답변을 드립니다.



Question	Asker

Type answer here

## 고지 사항 (Disclaimer)

본 콘텐츠는 고객의 편의를 위해 AWS 서비스 설명을 위해 온라인 세미나용으로 별도로 제작, 제공된 것입니다. 만약 AWS 사이트와 콘텐츠 상에서 차이나 불일치가 있을 경우, AWS 사이트(aws.amazon.com)가 우선합니다. 또한 AWS 사이트 상 에서 한글 번역문과 영어 원문에 차이나 불일치가 있을 경우(번역의 지체로 인한 경우 등 포함), 영어 원문이 우선합니다.

AWS는 본 콘텐츠에 포함되거나 콘텐츠를 통하여 고객에게 제공된 일체의 정보, 콘텐츠, 자료, 제품(소프트웨어 포함) 또는 서비스를 이용함으로써 인하여 발생하는 어떠한 종류의 손해에 대하여 어떠한 책임도 지지 아니하며, 이는 직접 손해, 간접 손해, 부수적 손해, 징벌적 손해 및 결과적 손해를 포함하되 이에 한정되지 아니합니다.

# 실습 시작 전 준비 사항

## AWS 계정으로 시작

1. 실습 전 계정을 꼭 신청해주세요 : <https://portal.aws.amazon.com/billing/signup#/start>
2. AWS 계정이 없으신 경우, 행사 참여 전에 미리 AWS 계정 생성 가이드를 확인하시고 AWS 계정을 생성해 주시길 바랍니다.

\*AWS 계정 생성 가이드: <https://aws.amazon.com/ko/premiumsupport/knowledge-center/create-and-activate-aws-account/>

3. 검증된 호환성을 위하여 실습 시 사용할 웹 브라우저는 Mozilla Firefox 또는 Google Chrome Browser로 진행 부탁드립니다.

# 실습 마무리 및 설문 참여 방법

- 실습이 모두 끝난 후에는 **자원 삭제**를 잊지 마세요. 직접 준비하신 AWS 계정으로 실습을 진행하신 고객 분들의 경우, 가이드에 따라 자원 삭제를 진행하셔야 합니다. 또한, 기존에 사용하시던 자원이 있으신 고객 분들의 경우, **오늘 생성한 자원만 삭제**하는 것에 주의 부탁드립니다.
- **가이드:** (세션별 제공)
- 마지막으로 세션이 끝난 후, **GoToWebinar 창을 종료하면 설문 조사 창**이 나옵니다.  
이때, **설문 조사를 진행해 주셔야 AWS 크레딧**(1인당 \$50 크레딧, 전체 세션당 1회 제공)을 제공받으실 수 있습니다.

AWS는 고객 피드백을 기반으로 의사 결정을 수행하며 이러한 피드백은 추후에 진행할 세션 방향을 결정합니다.

더 나은 세션을 위하여 여러분들의 소중한 의견을 부탁드립니다.

감사합니다.

# 크레딧 안내

- AWS 계정으로 시작하실 경우, **금일 실습에서 발생하는 비용은 당월 과금이 되는 점 미리 확인** 부탁드립니다.
- 웨비나 종료 후 **설문 조사에 참여해주신 분들께는 AWS 크레딧 바우처** (1인당 \$50 USD 크레딧, 전체 세션당 1회 제공)를 드립니다.
- 해당 **AWS 크레딧**은 등록하신 이메일 계정으로 **행사 종료 후 1개월 내** 발송 드릴 예정이며, 전달 받은 AWS 크레딧은 바로 사용 가능합니다.

# 감사 메일 & 참석 증명서

- AWS Builders Korea 세션에 참석해 주신 분들께 행사 종료 후 1개월 내 감사메일과 참석 증명서가 순차 발송됩니다.
- 등록 진행 후 참석하지 않으실 경우 별도 메일 및 증명서는 발급되지 않습니다.

## 감사 메일 예시

**AWS Builders Korea Program 온라인 세미나에  
참석해 주셔서 감사합니다.**

AWS Builders Korea Program에 참석하고 피드백을 공유해주셔서  
감사드립니다. 세미나 자료는 아래 링크를 통해 확인하실 수 있습니다.

[자료 확인하기](#)

## 참석 증명서 예시

### 참석 증명서

AWS Builders Korea Program에 참석해 주셔서 감사합니다.

홍길동

2023년 3월 20일 - 3월 24일



# 강연 다시보기

aws builders korea program 다시보기



<https://kr-resources.awscloud.com/aws-builders-korea-program>

# AWS Builders Korea 프로그램 정보

aws builders korea program



<https://aws.amazon.com/ko/events/seminars/aws-builders/>

# Agenda

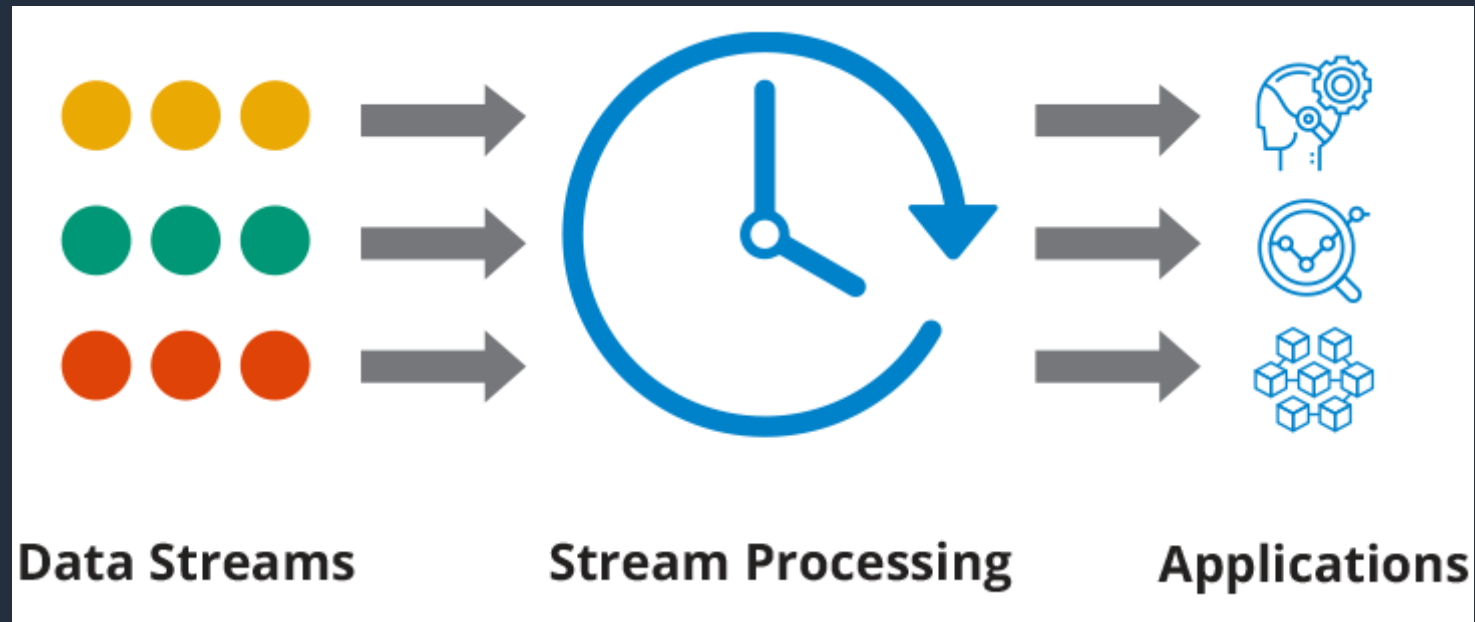
- 실시간(스트리밍) 데이터란?
- 배치 처리 VS 실시간 처리
- 실시간 데이터 파이프라인
- 실시간 데이터 처리를 위한 AWS 서비스
- 데이터 파이프라인 사례
- Wrap-up



# 실시간(스트리밍) 데이터란?

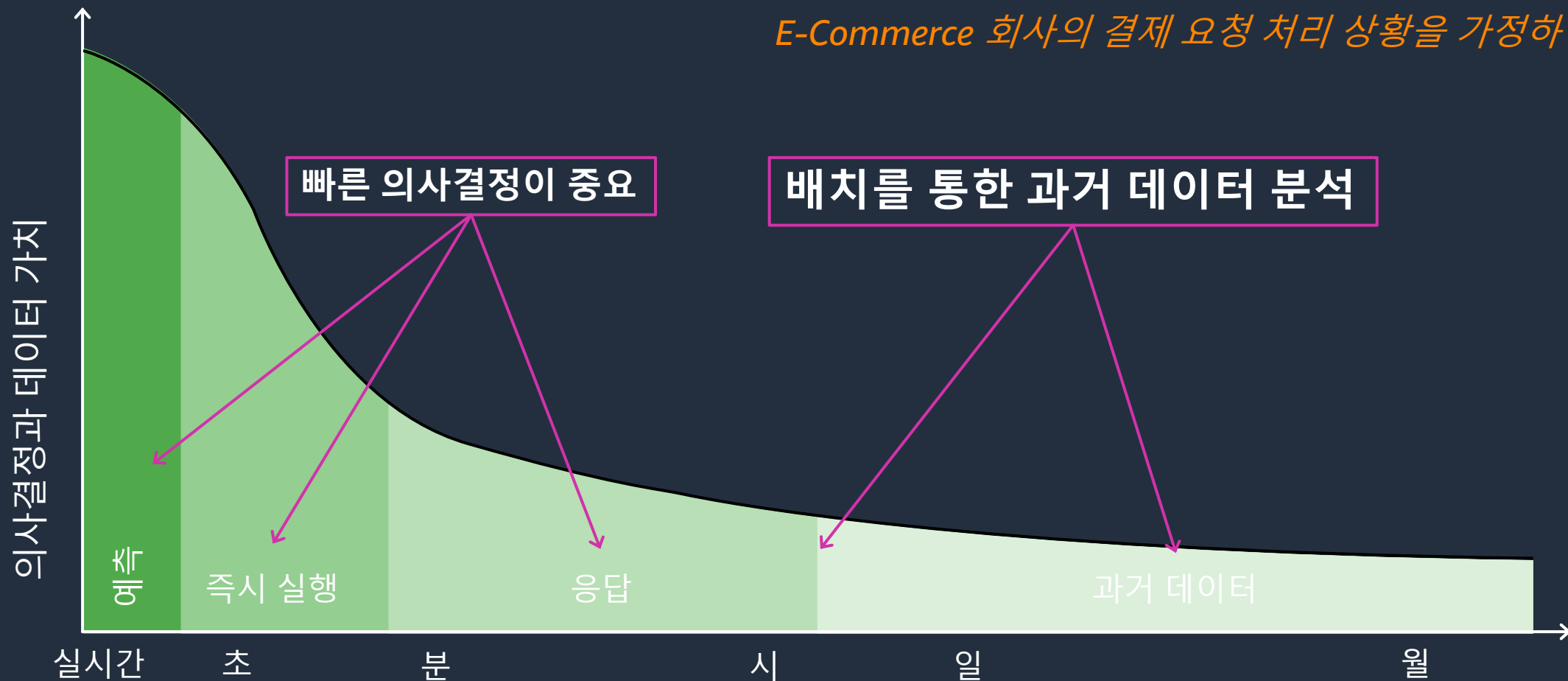
# 실시간 데이터란?

큰 규모로 **빠른 속도**로 생성되는 데이터를 의미하며, **계속 생성**되는 특징



# 실시간 데이터 처리와 가치

*E-Commerce 회사의 결제 요청 처리 상황을 가정하면...*



Source: Perishable insights, Mike Gualtieri, Forrester

# 실시간 데이터 활용 사례



부정 결제 분석



고객가치 향상



로그 분석



헬스케어



마케팅 캠페인



예지 정비

# 배치 처리 VS 실시간 처리

# 배치 처리 VS 실시간 처리

	배치 처리	실시간 처리
데이터 구조	테이블	스트림
처리 모델	정기적 (일, 주, 월)	연속적

# 배치 처리 - 테이블

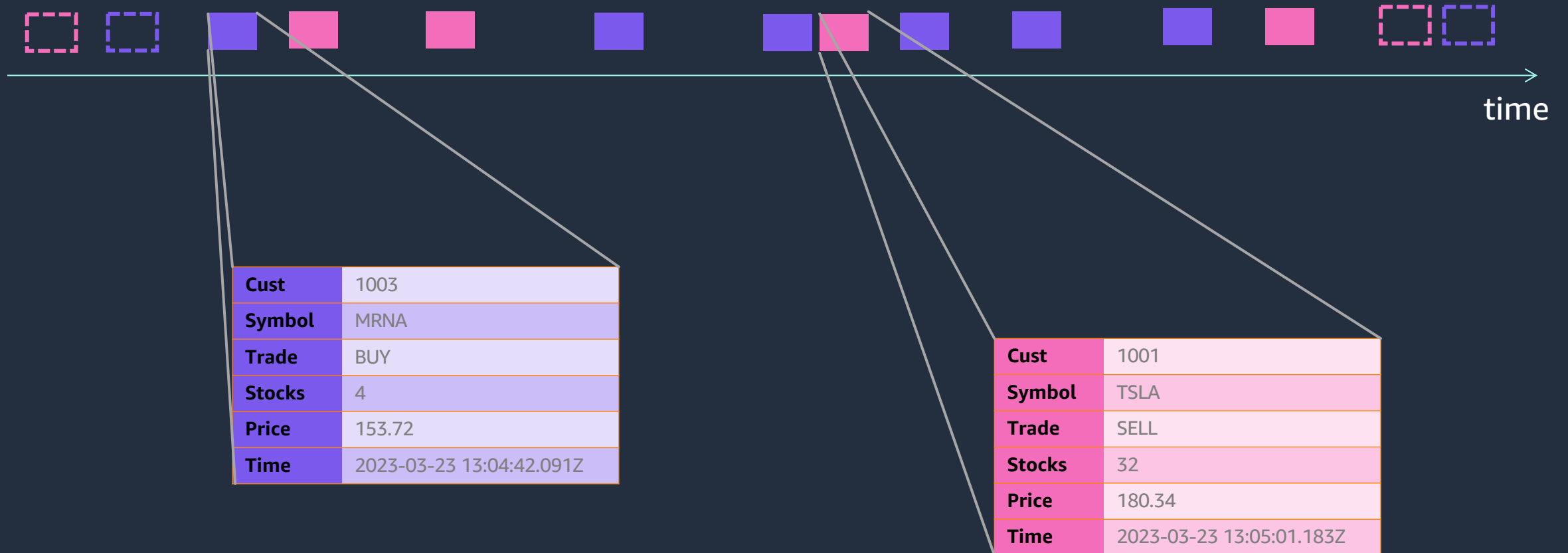
e.g. 주식 잔고

Customer	Symbol	Stocks
1001	TSLA	53
1001	MRNA	32
1002	AAPL	104
1003	TSLA	42
1004	APPL	7

시간 : 2023-03-23 13:34:54.622Z

# 실시간 처리 - 스트리밍

e.g. 주식 거래





# 배치 방식과 실시간 방식의 차이

	배치 방식	실시간 방식
컴퓨팅	<ul style="list-style-type: none"><li>- 대용량 스토리지</li><li>- 대용량 데이터 프로세싱</li></ul>	<ul style="list-style-type: none"><li>- 작은 규모의 스토리지</li><li>- 연속적인 데이터 프로세싱</li></ul>
성능	<ul style="list-style-type: none"><li>- 분석에 몇 분, 몇 시간, 몇 일이 소요</li></ul>	<ul style="list-style-type: none"><li>- m/s, seconds 단위의 짧은 시간 소요</li></ul>
분석	<ul style="list-style-type: none"><li>- 복잡한 계산 및 분석 구조</li></ul>	<ul style="list-style-type: none"><li>- 데이터 유실 방지를 위한 비동기, 분산</li></ul>

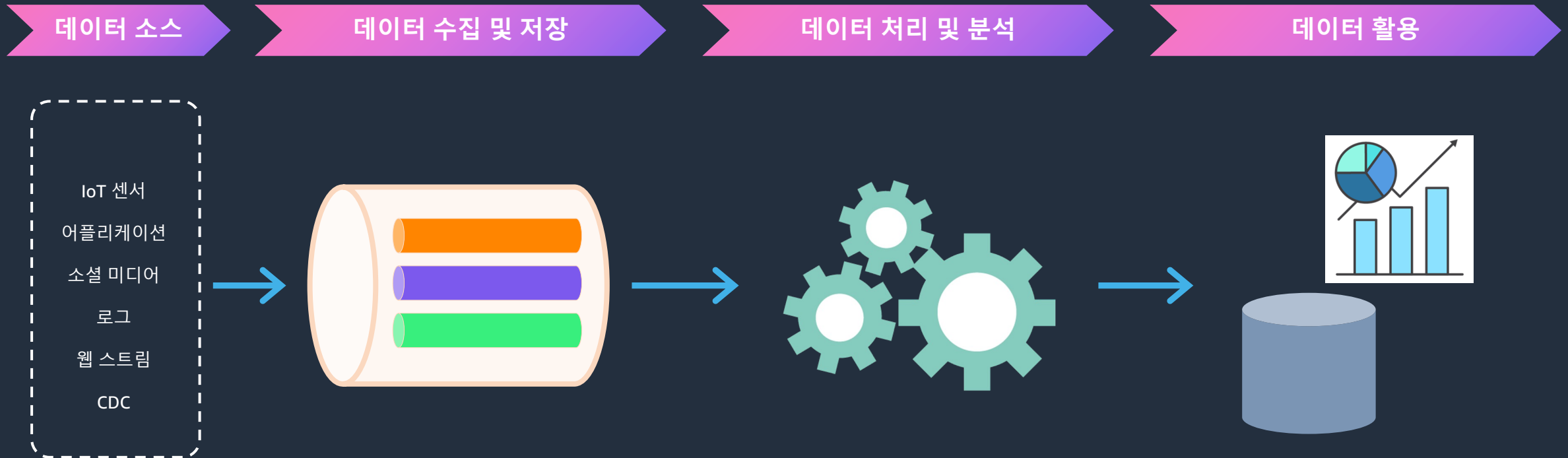
# 실시간 데이터 파이프라인

# 실시간 데이터 종류

- 모바일 앱
- 어플리케이션 로그
- 커넥티드 제품
- 웹 클릭 스트림
- IoT 센서
- 스마트 빌딩
- Microservices



# 실시간 데이터 파이프라인



# 수집 및 저장



데이터의 생산량이 데이터의 처리량보다 많을 때는  
어떤 상황이 발생할 것인가?

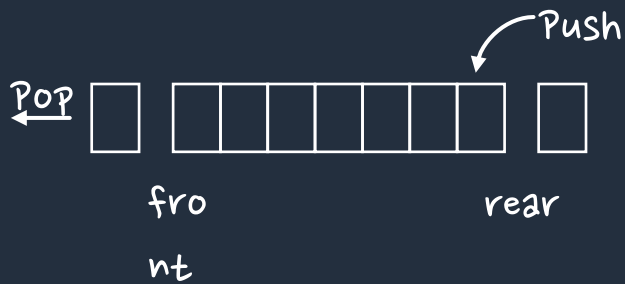
# 수집 및 저장



# 스트림 스토리지의 특징



## • 메시지 브로커



## • 분산



## 저장



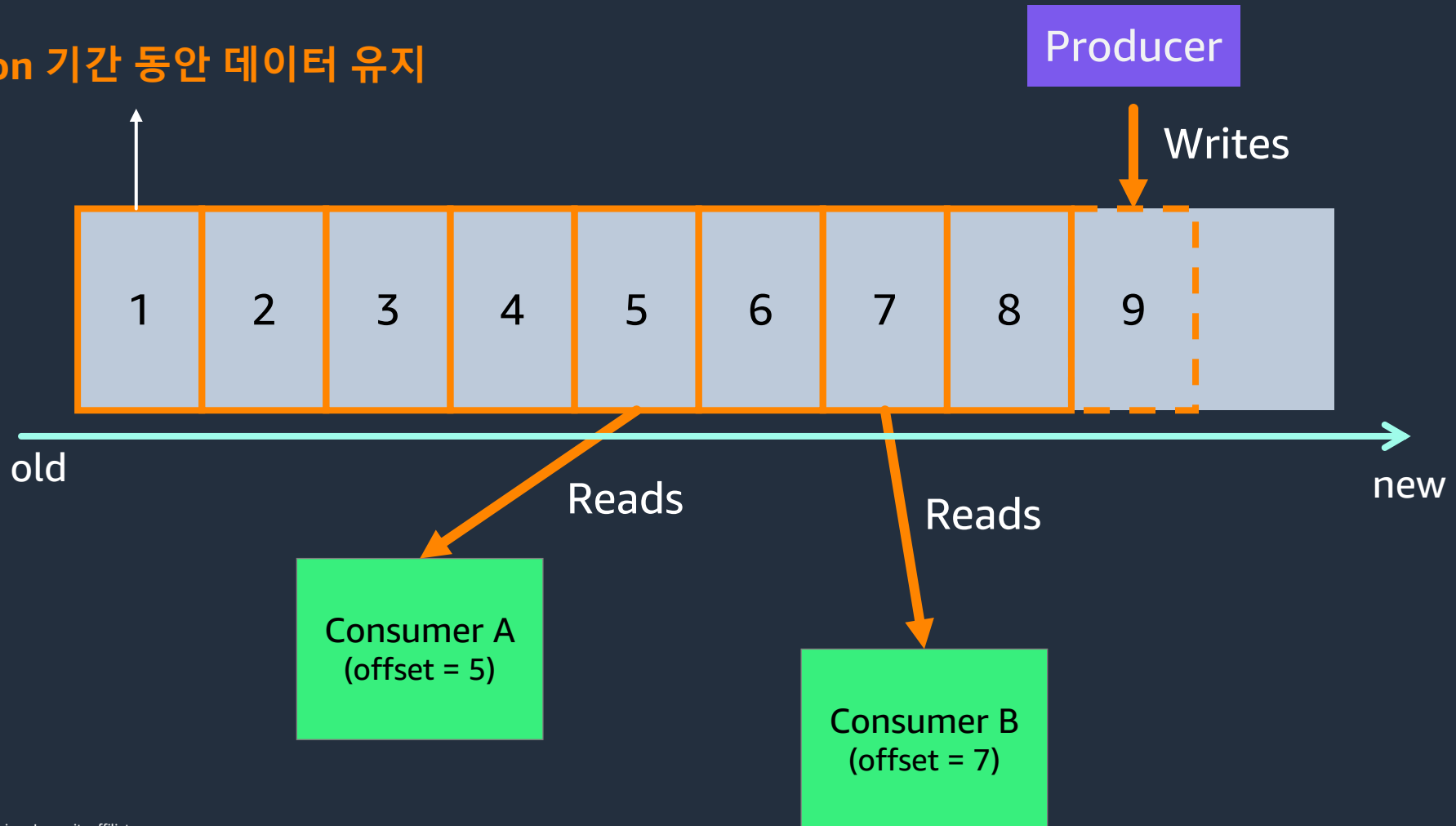
- Producer와 Consumer의 분리
- 영구적인 버퍼(재처리 가능)
- 다수의 스트림을 수집

- 메시지의 순서 유지
- 병렬적인 소비

# 연속한 데이터 추가



Retention 기간 동안 데이터 유지





# 데이터 처리 및 분석

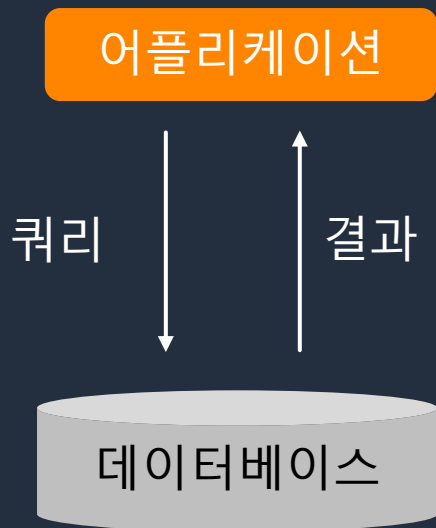
데이터 소스

수집 및 저장

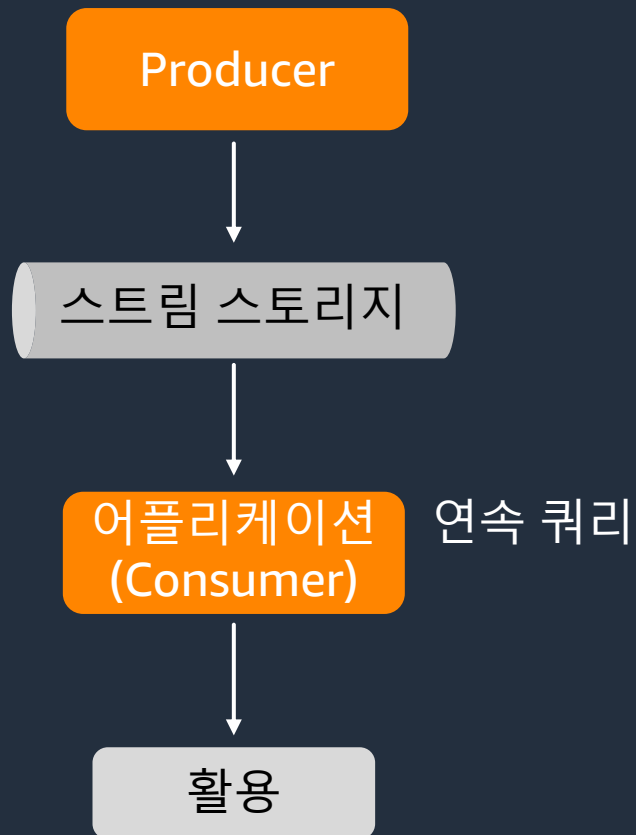
처리 및 분석

활용

## 전통적인 배치 처리



## 실시간 데이터 처리



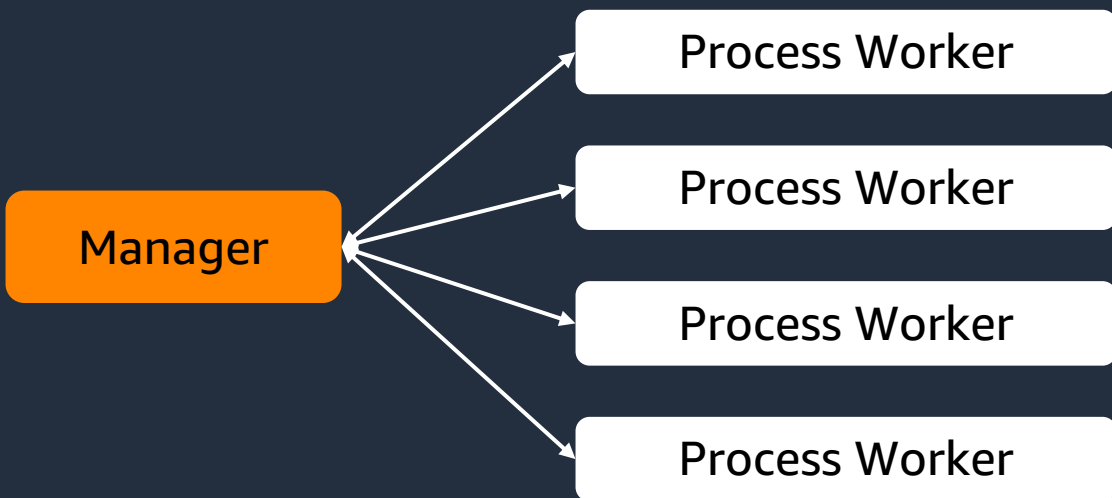
- 지속적으로 쿼리를 실행
- 데이터 처리 중 이슈 발생
- 장애가 발생한 시간부터 재처리 필요 할 수 있음

# 분산 구조의 데이터 처리

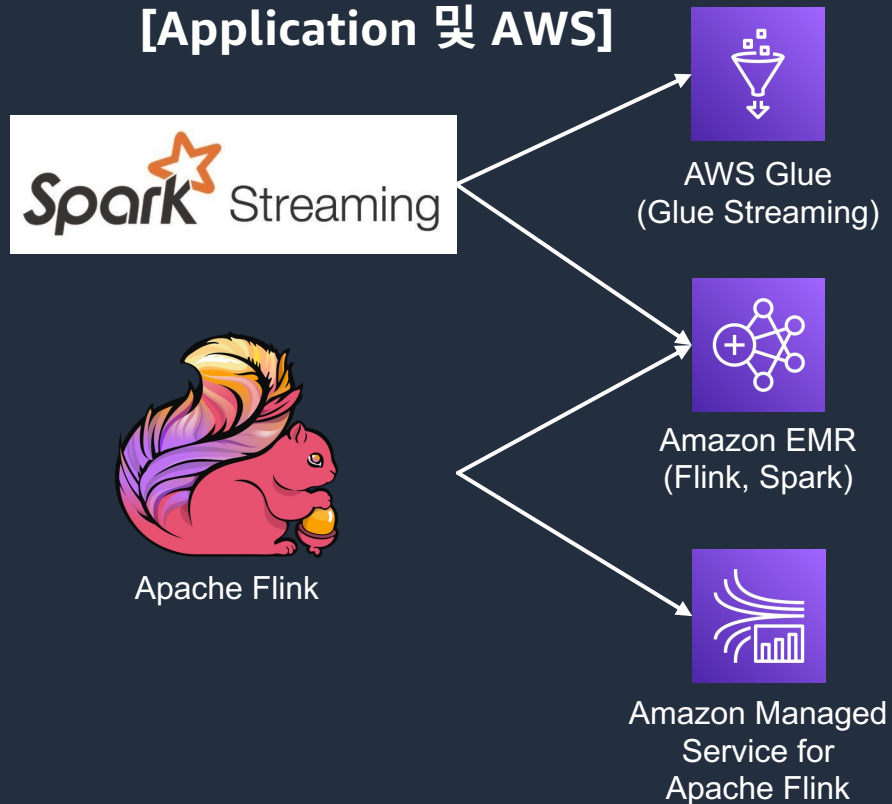


만약 **시간당 1TB** 이상의 실시간 데이터가 발생하고  
이를 **1대의 서버**가 처리해야 한다면?

## [분산 아키텍처]



## [Application 및 AWS]



# 실시간 데이터 처리를 위한 AWS 서비스

# 실시간 데이터 처리를 위한 AWS 서비스

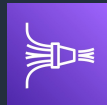
## 데이터 소스

IoT 센서  
어플리케이션  
소셜 미디어  
로그  
웹 스트림  
CDC

## 데이터 수집 및 저장



Amazon Kinesis  
Data Streams



Amazon Kinesis  
Data Firehose



Amazon Managed  
Streaming for  
Apache Kafka

## 데이터 처리 및 분석



Amazon Managed  
Service for  
Apache Flink



AWS Glue  
(Glue Streaming)



Amazon EMR  
(Flink, Spark Streaming)



AWS Lambda



Amazon MSK Connect

## 데이터 활용



Amazon S3  
(데이터 저장)



Amazon Athena  
(데이터 분석)

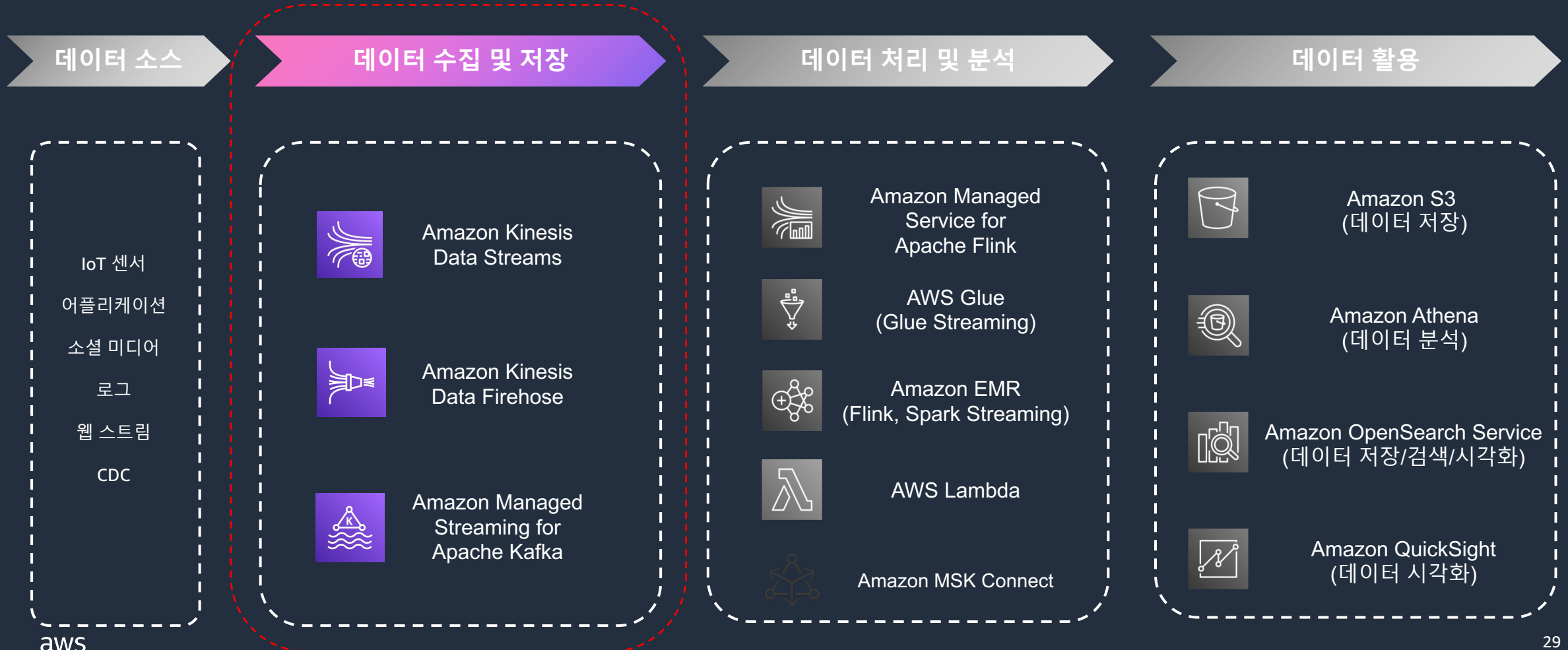


Amazon OpenSearch Service  
(데이터 저장/검색/시각화)

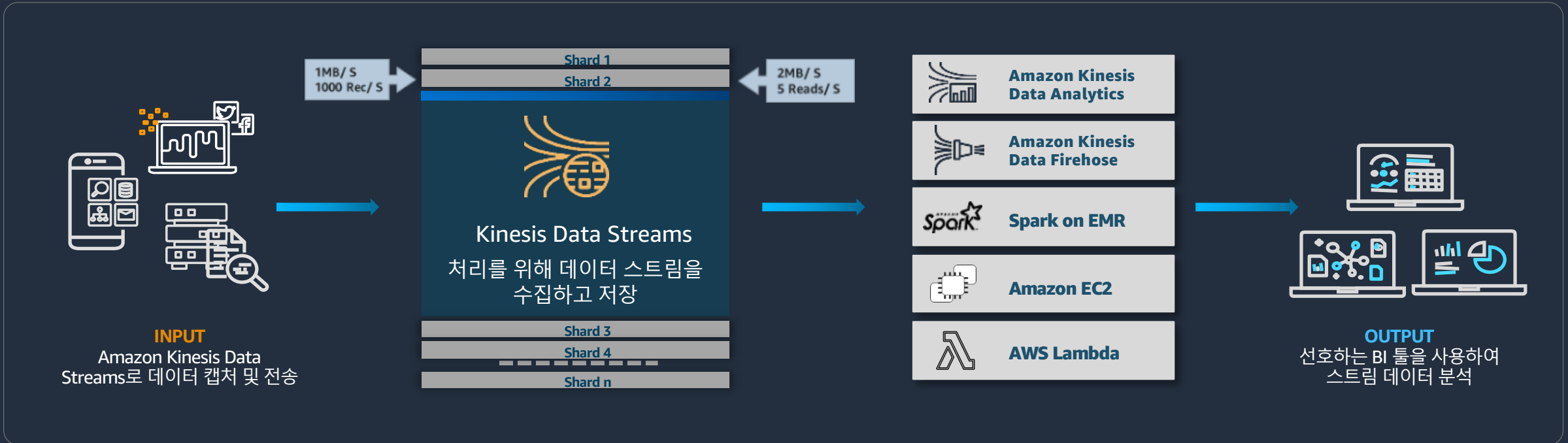


Amazon QuickSight  
(데이터 시각화)

# 데이터 수집 및 저장



# Kinesis Data Streams



## 온디맨드 모드

- 처리량에 따라 Shards 수를 자동으로 조정
- 최대 200 Shard 까지 확장 가능
- 30일 간의 최대 쓰기의 두 배까지 버스팅 가능

## 프로비저닝 모드

- Shards 의 수를 직접 관리
- 온디맨드와 프로비저닝 간 하루에 2번 전환 가능
- 용량 제한 초과 시 입력 호출 제한

# Kinesis Data Streams



## Kinesis 스트림



샤드는 시간당 과금이 됨

- 최대 1MB/초, 최대1,000 TPS 쓰기
- 최대 2MB/초, 최대 5TPS 읽기
- 모든 데이터는 기본으로 24시간동안 저장됨
- Long Term Retention으로 최대 1년 연장 가능
- 샤드들을 분할하거나 병합을 통해 확장
- 보존 기간 내에 있는 데이터 다시 재생 가능

# Managed Streaming for Apache Kafka(MSK)

데이터 소스

수집 및 저장

처리 및 분석

활용

## Apache Kafka 운영의 어려움



설치하기 어려움



확장의 까다로움



고가용성 구성이 어려움



개발이 필요한 연계



오류가 발생하기 쉽고, 관리가 복잡함



유지 관리 비용 비쌘



## Managed Streaming for Apache Kafka(MSK)

데이터 소스

수집 및 저장

처리 및 분석

활용



Amazon MSK

Apache Kafka를 조직에서 보다 안전하고, 가용성이 높게 접근할 수 있게 합니다.

설계, 기본값과 자동화를 통해 모범 사례로 이끍니다.

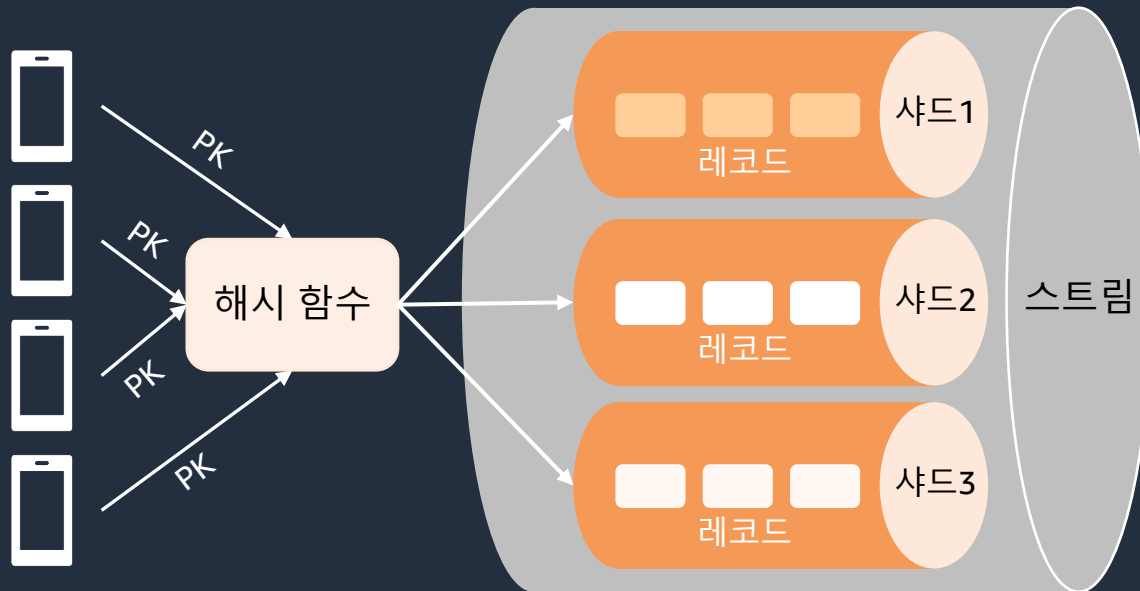
개발자가 인프라 관리보다 애플리케이션 개발에 더 집중할 수 있게 합니다.

MSK 서버리스를 사용하면 적절한 규모, 확장 및 파티션 관리를 없애 줍니다.

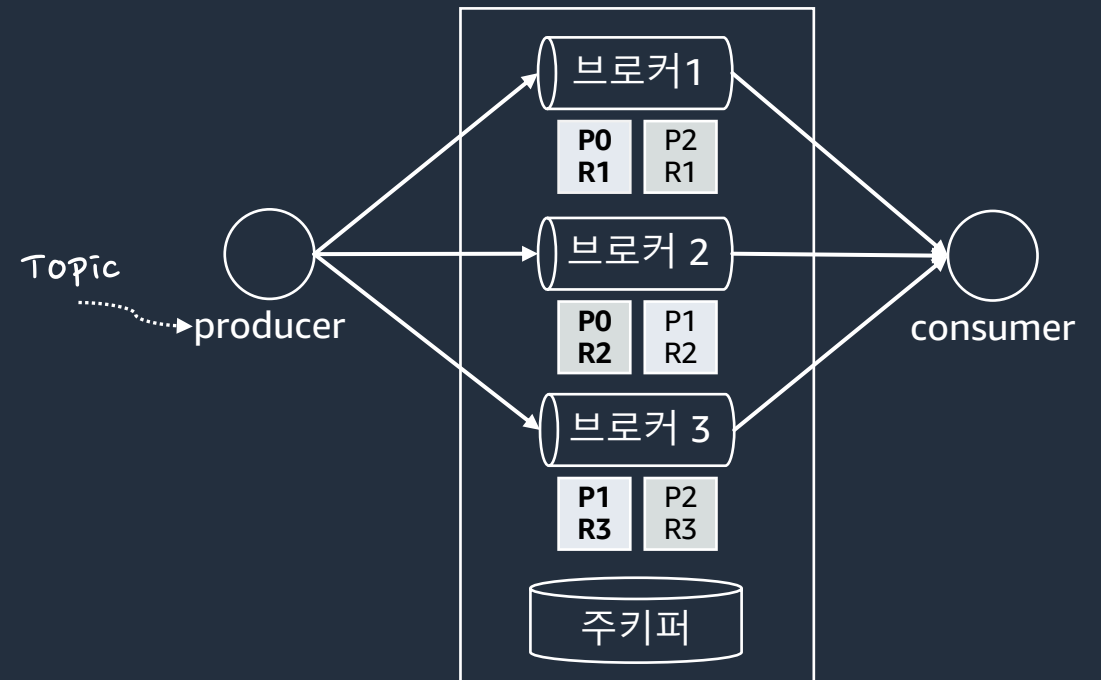
# Kinesis Data Streams VS MSK



Amazon Kinesis Data Streams



Amazon Managed Streaming for Kafka

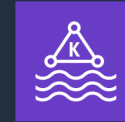


# Kinesis Data Streams VS MSK



Amazon Kinesis  
Data Streams

- 운영 관점에서 볼 때,
  - X
  - X
  - **streams** 개수?
  - stream 별 **shards** 수?
- **Throughput** 프로비저닝 모델
- **shards** 수의 증가 O, 감소 O
- 대부분의 **AWS** 서비스와 완전 통합

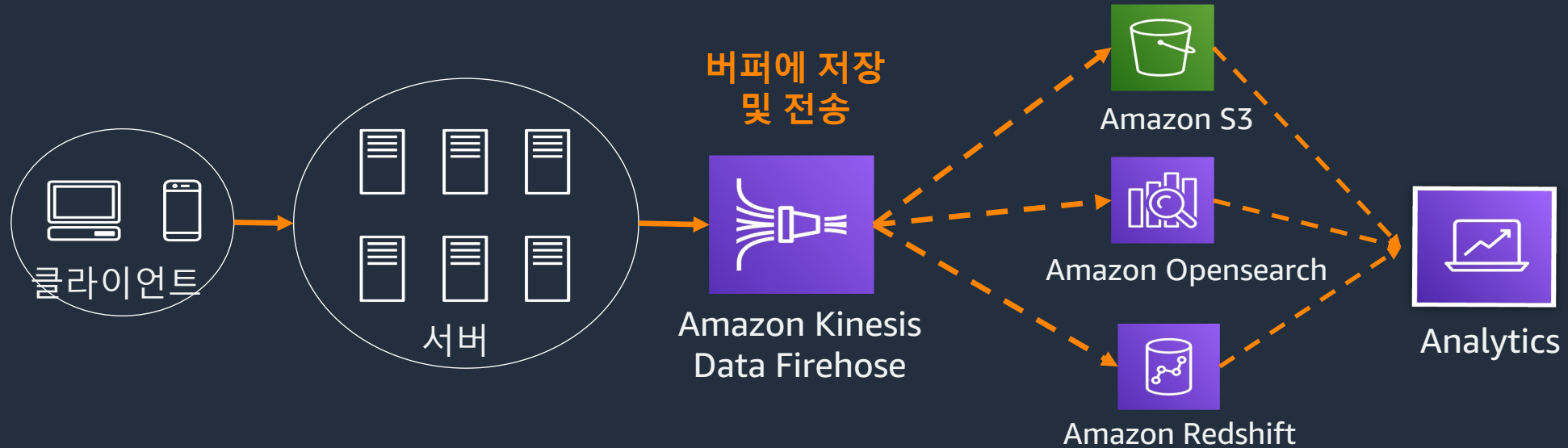


Amazon Managed  
Streaming for Kafka

- 운영 관점에서 볼 때,
  - 단일 **cluster** vs 다수의 **clusters**?
  - cluster 별 **brokers** 수?
  - broker 별 **topics** 수?
  - topic 별 **partitions** 수?
- **Cluster** 프로비저닝 모델
- **partition** 수의 증가 O, 감소 X
- 일부 **AWS** 서비스와 통합



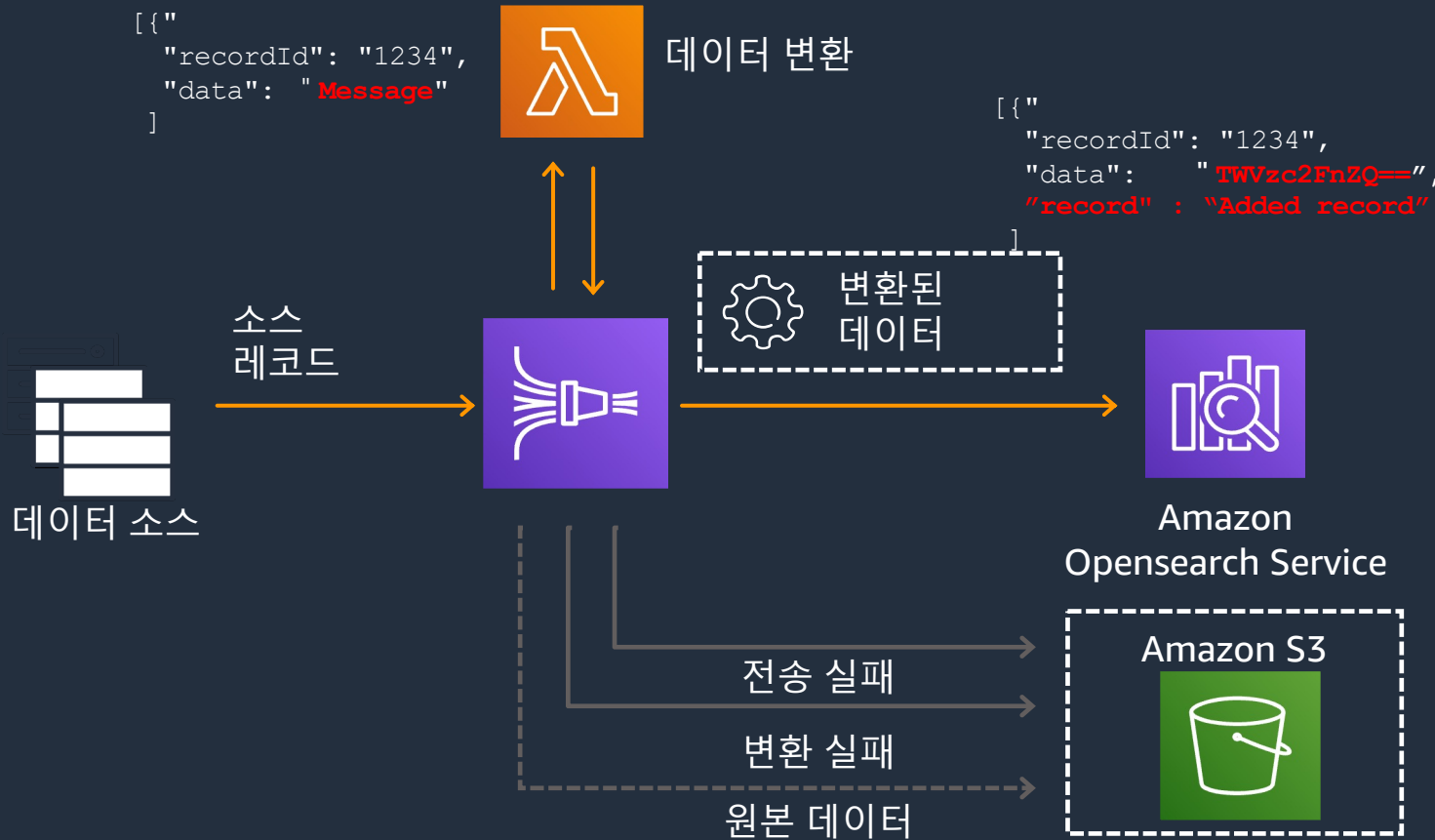
# Kinesis Data Firehose



- 제로 관리 및 탄력적
- 데이터 저장소에 직접 통합
- 서버리스 지속적 데이터 변환

- 거의 실시간
- Parquet / ORC로 데이터 형식 변환
- Datadog, Sumo Logic, New Relic 및 MongoDB에 직접 데이터 전송

# Kinesis Data Firehose



- 다양한 AWS 서비스로 전달
- 배치 방식으로 여러 데이터를 Lambda를 통해 전처리
- 원본 데이터, 전처리 실패, 전송 실패 데이터를 S3에 저장

# Kinesis Data Firehose



## 년, 월, 일로 파티셔닝

- `s3://datalake-example/logs/parquet/year=2019/month=6/day=1/`
- `s3://datalake-example/logs/parquet/year=2019/month=6/day=2/`

쿼리를 다음과 같이 최적화 가능 -> 스캔 데이터를 최소화하여 성능 향상 및 비용 최적화

```
SELECT request_method, response_code, access_time FROM logs
WHERE year = 2019 AND month = 6 AND day = 2
```

# Data Streams VS Firehose

데이터 소스

수집 및 저장

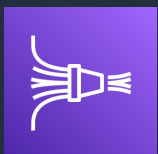
처리 및 분석

활용



Amazon Kinesis  
Data Streams

- Kinesis Data Streams는 수신 레코드 별로 **1초 미만의 처리**에 적합한 프레임 워크
- 리텐션 기간 동안 Shards의 데이터 재처리 가능



Amazon Kinesis  
Data Firehose

- Kinesis Data Firehose는 목적지까지 데이터 전송하는데 최적화, **지연 시간 존재**
- 데이터 전송 실패 시 Amazon S3에 저장

# 데이터 처리 및 분석

## 데이터 소스

IoT 센서  
어플리케이션  
소셜 미디어  
로그  
웹 스트림  
CDC

## 데이터 수집 및 저장



Amazon Kinesis  
Data Streams



Amazon Kinesis  
Data Firehose



Amazon Managed  
Streaming for  
Apache Kafka

## 데이터 처리 및 분석



Amazon Managed  
Service for  
Apache Flink



AWS Glue  
(Glue Streaming)



Amazon EMR  
(Flink, Spark Streaming)



AWS Lambda



Amazon MSK Connect

## 데이터 활용



Amazon S3  
(데이터 저장)



Amazon Athena  
(데이터 분석)



Amazon OpenSearch Service  
(데이터 저장/검색/시각화)



Amazon QuickSight  
(데이터 시각화)



# Amazon Managed Service for Apache Flink

데이터 소스

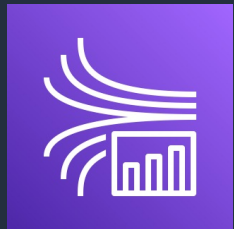
수집 및 저장

처리 및 분석

활용

## 실시간 데이터 처리를 위해 사용되는 어플리케이션인 **Apache Flink**의 완전 관리형 서비스

- Java, Scala, Python 및 SQL 기반의 유연한 API를 제공
- 다양한 AWS 서비스와의 통합
- 사용한 만큼 비용을 지불하는 서버리스 서비스
- 10개가 넘는 Apache Flink 커넥터를 지원



Amazon Managed  
Service for  
Apache Flink



Apache Flink

*Kinesis Data Analytics에서 Amazon Managed Service for Apache Flink로  
서비스 명칭이 변경되었습니다.(2023년 9월)*

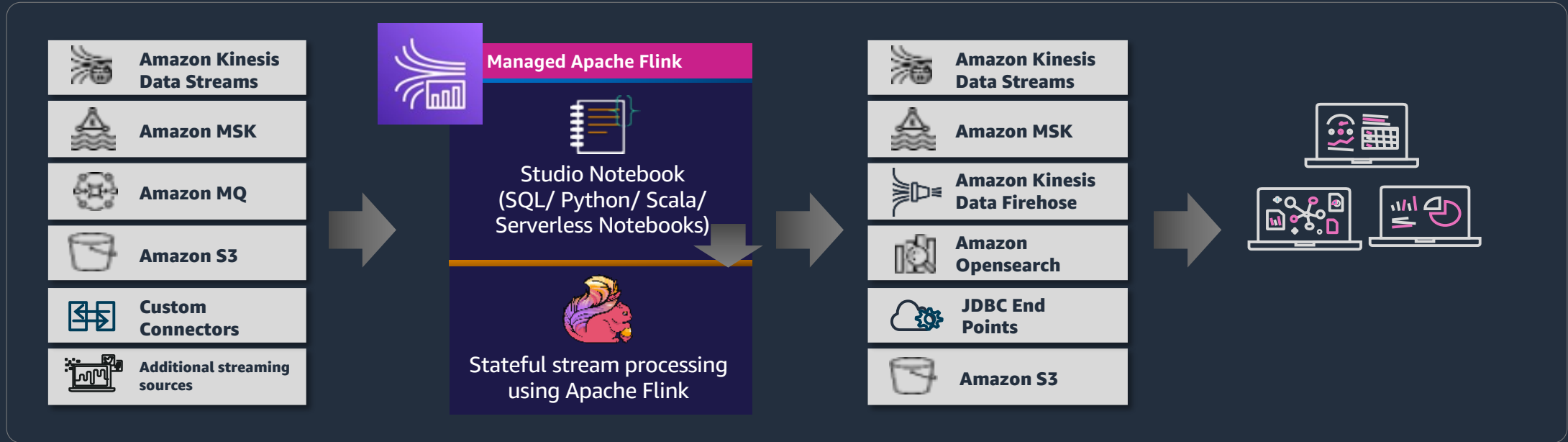
# Amazon Managed Service for Apache Flink

데이터 소스

수집 및 저장

처리 및 분석

활용



- SQL, Python, Scala 및 Java 또는 통합 Apache Flink 애플리케이션을 사용하여 실시간으로 스트리밍 데이터 처리 분석
- Studio Notebook을 활용한 Ad-hoc 분석 및 Apache Flink용 Streaming Application으로 배포 가능
- 완전 관리형 탄력적 스트림 처리 애플리케이션 구축

# Studio Notebook

데이터 소스

수집 및 저장

처리 및 분석

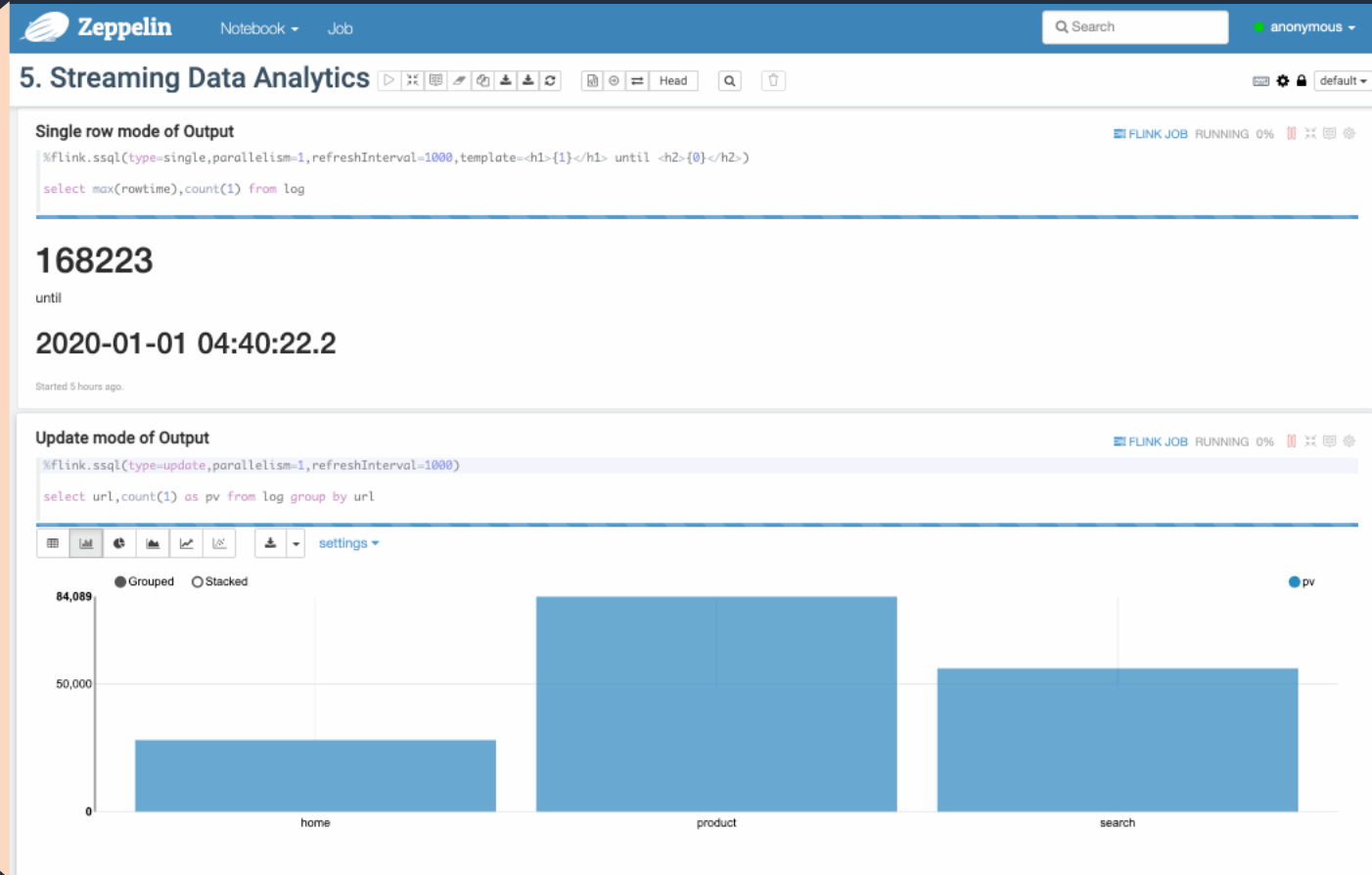
활용

## Managed Apache Flink

대시보드

Apache Flink 애플리케이션

Studio 노트북



# KPU & Parallelism Per KPU



KPU 단위로 어플리케이션을 프로비저닝합니다.  
( 1 KPU = **1vCPU, 4G Mem, 50G Storage**)



1 KPU  
1 parallelism Per KPU



1 KPU  
2 parallelism Per KPU

# AWS Glue

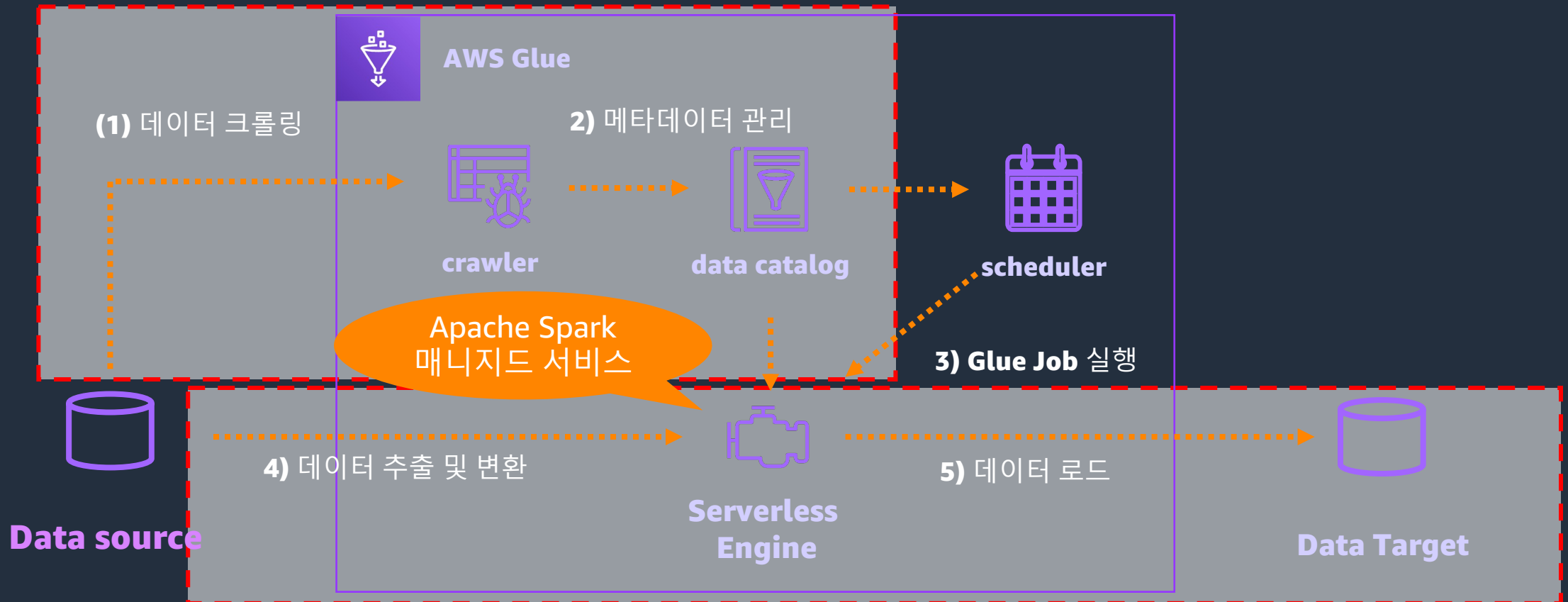
데이터 소스

수집 및 저장

처리 및 분석

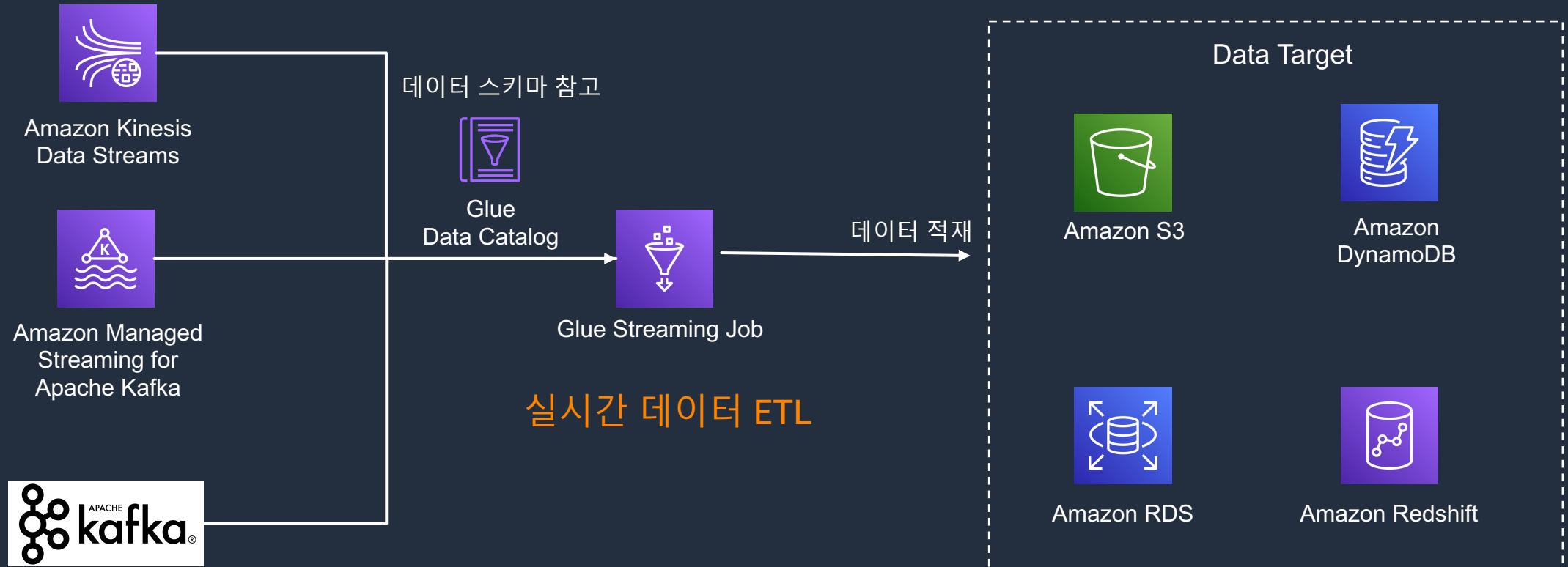
활용

## 1. Data Catalog 기능

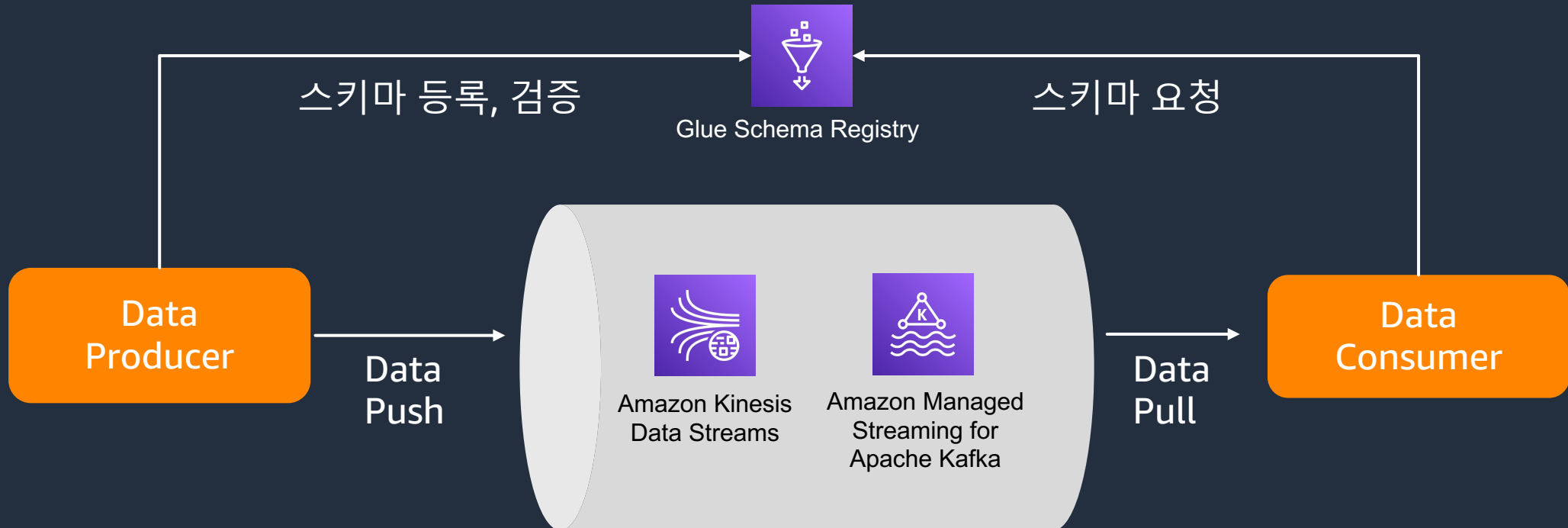


## 2. Glue ETL 기능

# Glue Streaming



# Glue Schema Registry



# Amazon EMR (Spark, Flink)

실시간 데이터 처리를 위한 어플리케이션



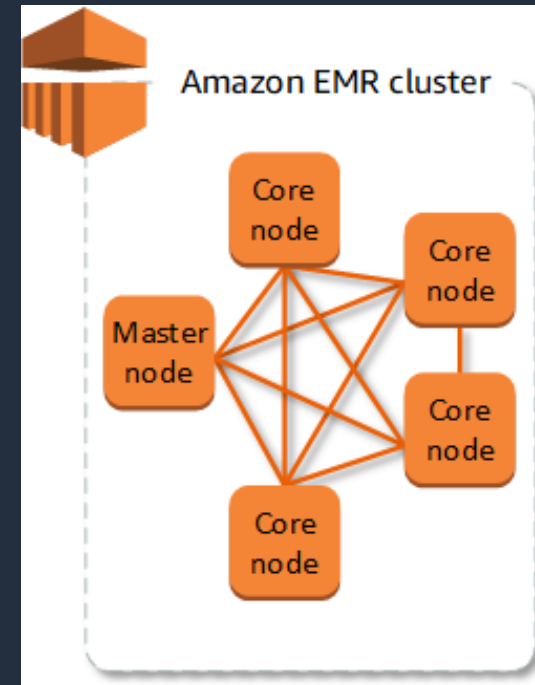
데이터 소스

수집 및 저장

처리 및 분석

활용

빅데이터 프레임워크 아키텍처





# Amazon EMR (Spark, Flink)

데이터 소스

수집 및 저장

처리 및 분석

활용

**20개 이상의 빅데이터 프레임 워크를 지원:**  
Apache Spark, Flink, Presto, Trino, Hive, HBase, Hudi 등



## 차별화 된 런타임 성능

오픈 소스 API와 100% 호환되는 성능 최적화 런타임



## 최신 오픈 소스

오픈 소스 출시 후 60일 이내에 새로운 오픈 소스 기능 제공



## 빅 데이터 분석을 위한 최고의 가격 대비 성능

EC2 스팟, EMR Managed Scaling 및 초당 비용을 사용하여 비용 절감



## 데이터 사이언스를 위한 셀프 서비스

EMR Studio 및 Sagemaker Studio와의 긴밀한 통합



## EC2, EKS 또는 온프레미스에서 워크로드 실행

EMR은 EC2, EKS 및 Outpost를 통해 온프레미스에서 빅 데이터 워크로드를 실행할 수 있는 유연성 제공



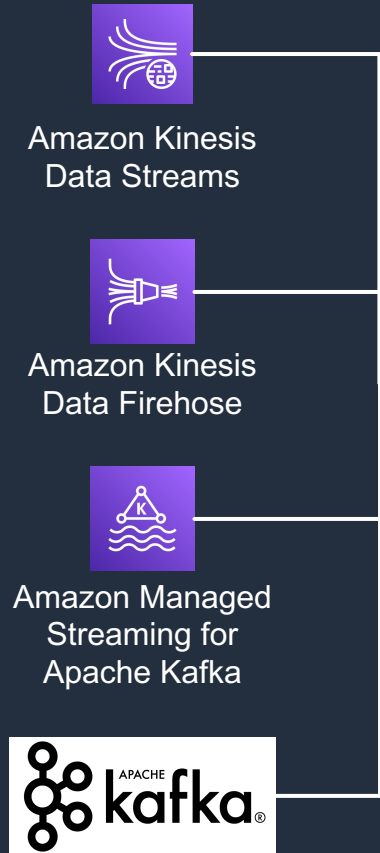
## S3 데이터 레이크 통합

Amazon S3와 통합하여 스토리지와 컴퓨팅 분리

# AWS Lambda



# 이벤트 필터링



AWS Lambda

[필터 예시]

```
{
  "data": {
    "order": {
      "type": [ "buy" ]
    }
  }
}
```



# MSK Connect

데이터 소스

수집 및 저장

처리 및 분석

활용

## Kafka 커넥트란?



# MSK Connect

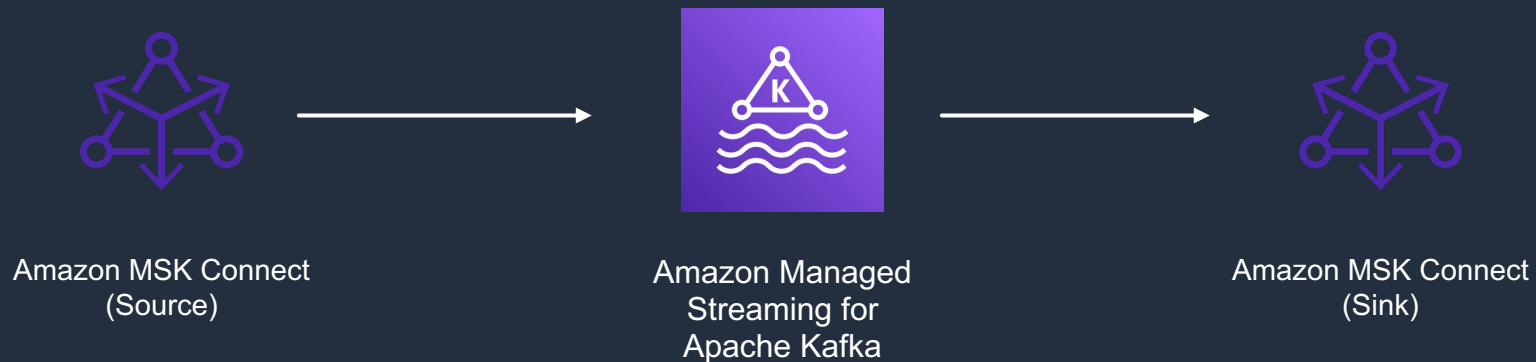
데이터 소스

수집 및 저장

처리 및 분석

활용

- 별도의 인프라스트럭처 관리가 필요없는 **매니지드 서비스**
- 기존 온프레미스에서 **사용 중이던 커넥트(개발/오픈소스)와의 호환** 지원
- 트래픽 양에 **따라 자동으로 인프라스트럭처를 프로비저닝**



# 데이터 활용

## 데이터 소스

IoT 센서  
어플리케이션  
소셜 미디어  
로그  
웹 스트림  
CDC

## 데이터 수집 및 저장



Amazon Kinesis  
Data Streams



Amazon Kinesis  
Data Firehose



Amazon Managed  
Streaming for  
Apache Kafka

## 데이터 처리 및 분석



Amazon Managed  
Service for  
Apache Flink



AWS Glue  
(Glue Streaming)



Amazon EMR  
(Flink, Spark Streaming)



AWS Lambda



Amazon MSK Connect

## 데이터 활용



Amazon S3  
(데이터 저장)



Amazon Athena  
(데이터 분석)



Amazon OpenSearch Service  
(데이터 저장/검색/시각화)



Amazon QuickSight  
(데이터 시각화)

# Amazon S3 (데이터 저장)



데이터 소스

수집 및 저장

처리 및 분석

활용

**무제한**에 가까운 스토리지 용량과 오브젝트

Amazon S3 기반의 **데이터 레이크** 구축

**S3 Intelligent-Tiering**를 통한 자동화된 비용 절감

S3 Glacier Deep Archive를 사용해 **비용 효율적인** 스토리지 저장

# Amazon S3 (데이터 저장)

## 11'9의 데이터 내구성



**99.99%**  
durability



**99.999%**  
durability



Designed for  
**99.999999999%**  
durability



# Amazon Athena

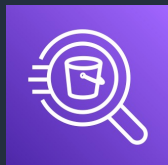
(데이터 분석)

데이터 소스

수집 및 저장

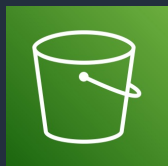
처리 및 분석

활용



Amazon Athena

대화형 SQL



Amazon S3



## 서버리스

- 인프라 관리 없이 S3에 저장된 데이터에 대해 쿼리하며 쿼리를 수행한 만큼 비용이 발생



## SQL 지원

- Parquet, CSV, Json, Avro 형식의 데이터에 대해서 쿼리가 가능하며 ANSI SQL을 사용



## 고성능

- 인프라 스트럭처 관리 없이 쿼리를 병렬로 실행하여 빠르게 결과 반환



## 비용 최적화

- 데이터 압축 및 파티셔닝 전략으로 쿼리 비용을 최대 30~90% 절약 가능

# Amazon Athena

(데이터 분석)



## 페더레이션 쿼리 – Amazon S3 외 25개 이상의 데이터 소스를 지원

AWS	Amazon Redshift	Amazon DynamoDB	Amazon DocumentDB	Amazon RDS	Amazon Timestream					
	Amazon CloudWatch	Amazon CloudWatch Metrics	Amazon OpenSearch Service	Amazon Neptune	Athena AWS CMDB					
OTHER SOURCES	SAP HANA	Teradata	Cloudera	Hortonworks	Snowflake	Microsoft SQL Server	Oracle	Google BigQuery	Azure Data Lake Storage Gen2	Azure Synapse
		MySQL	PostgreSQL	Redis	HBase	Vertica	TPC-DS	Custom		

# Amazon QuickSight (데이터 시각화)

데이터 소스

수집 및 저장

처리 및 분석

활용

모든 분석 요건을 위해 통합된 BI

풍부한 데이터 경험을 빠르게 생성

Auto scale을 통한 일정하게 높은 성능

사용량 기반의 요금으로 비용이 낮음



# Amazon OpenSearch

(데이터 저장/검색/시각화)

데이터 소스

수집 및 저장

처리 및 분석

활용



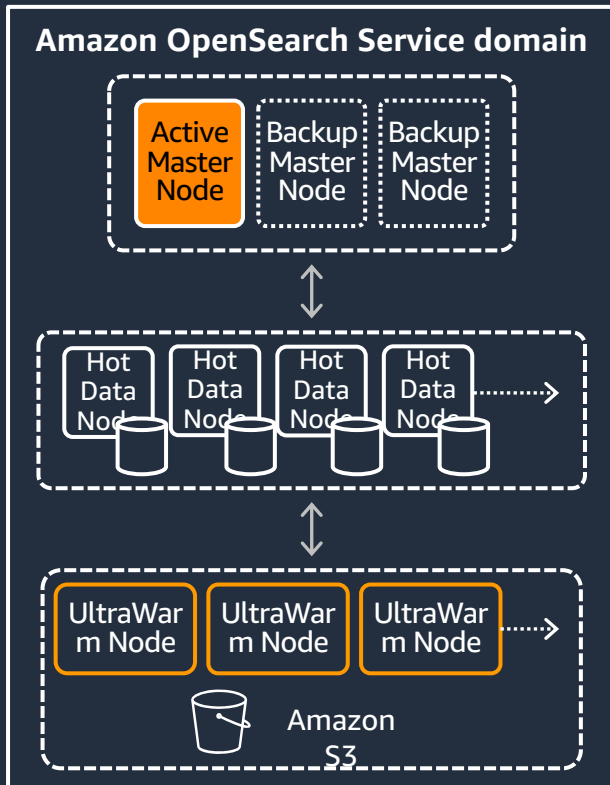
- Amazon OpenSearch는 Elasticsearch에서 파생된 **오픈소스 제품군**
- 대량의 실시간 데이터를 빠르게 저장, 검색, 분석 할 수 있는 **데이터베이스 + 검색 엔진**
- Amazon OpenSearch 플러그인을 사용하여 OpenSearch에 저장된 데이터를 기반으로 머신러닝 등의 기능을 사용 가능

# Amazon OpenSearch

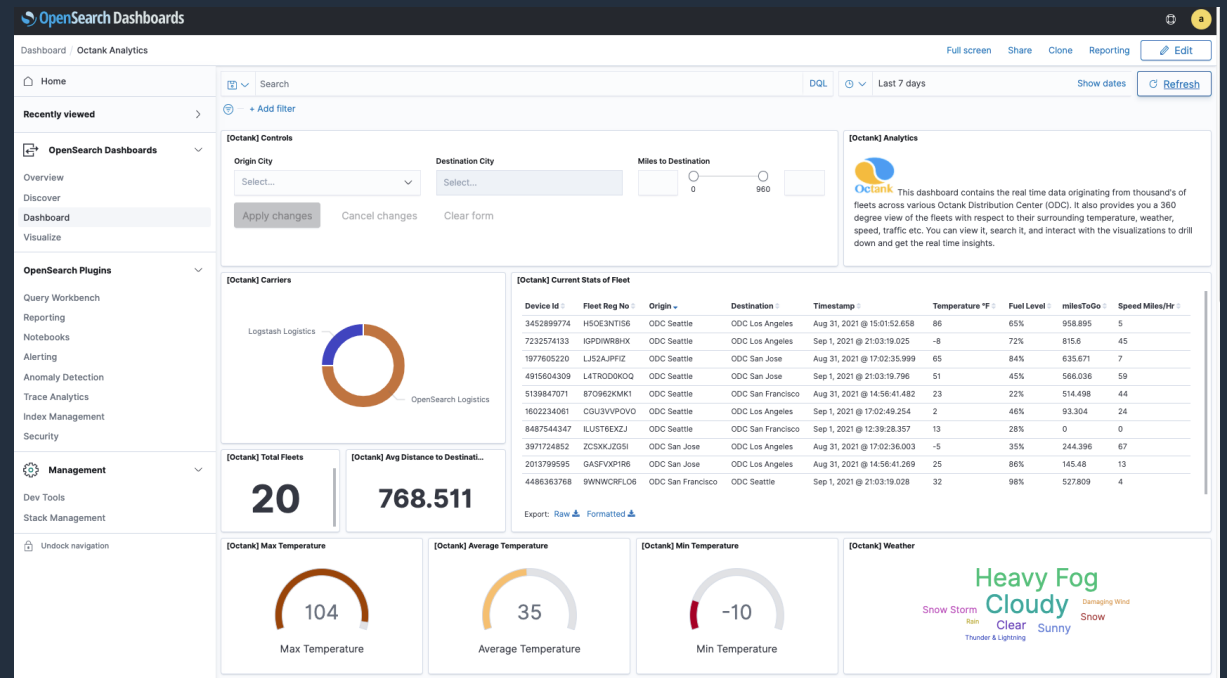
(데이터 저장/검색/시각화)



## [OpenSearch 클러스터 아키텍처]

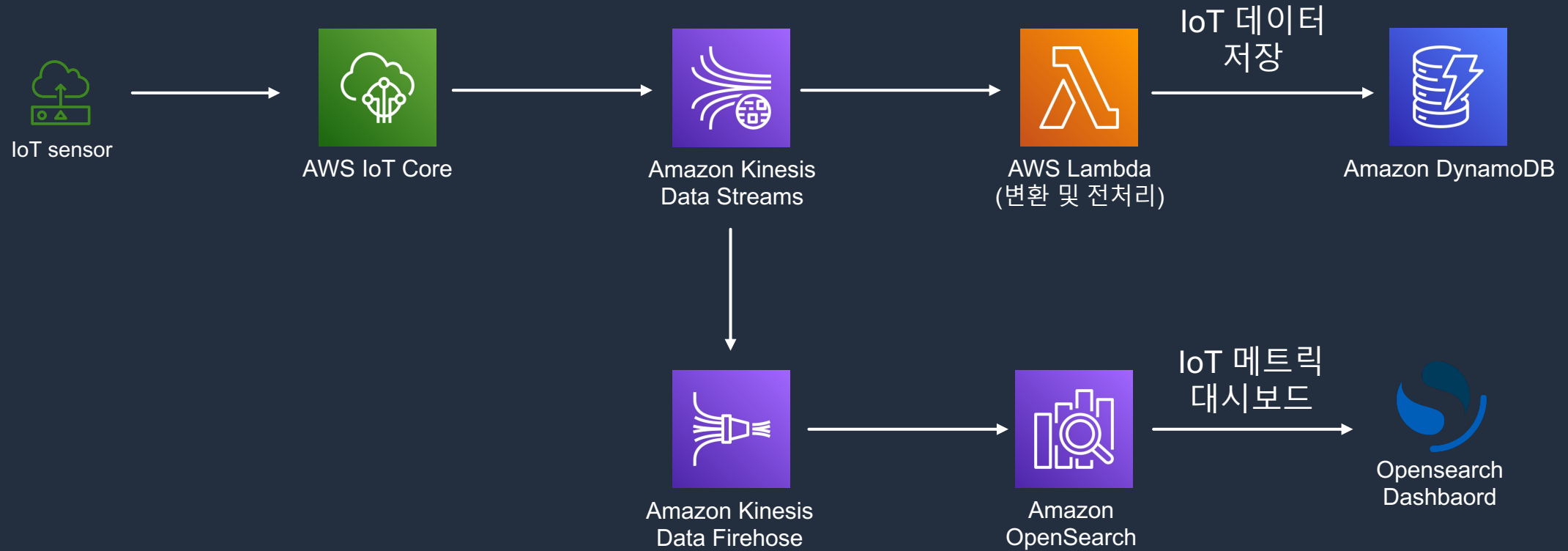


## [ OpenSearch 대시보드를 통한 모니터링 ]

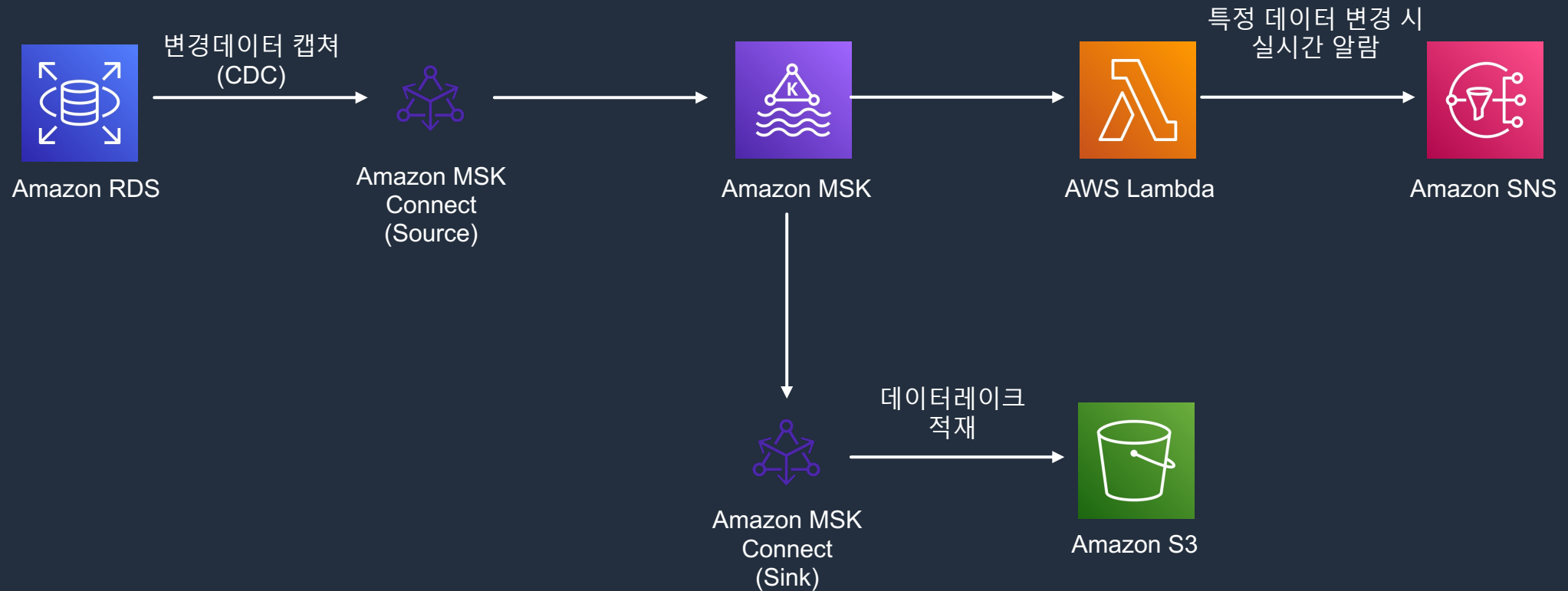


# 데이터 파이프라인 사례

# 실시간 IoT 단말 관리



# 변경데이터 캡처(CDC) 구성





# Wrap-up

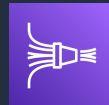
## 데이터 소스

IoT 센서  
어플리케이션  
소셜 미디어  
로그  
웹 스트림  
CDC

## 데이터 수집 및 저장



Amazon Kinesis  
Data Streams



Amazon Kinesis  
Data Firehose



Amazon Managed  
Streaming for  
Apache Kafka

## 데이터 처리 및 분석



Amazon Managed  
Service for  
Apache Flink



AWS Glue  
(Glue Streaming)



Amazon EMR  
(Flink, Spark Streaming)



AWS Lambda



Amazon MSK Connect

## 데이터 활용



Amazon S3  
(데이터 저장)



Amazon Athena  
(데이터 분석)



Amazon OpenSearch Service  
(데이터 저장/검색/시각화)



Amazon QuickSight  
(데이터 시각화)

# Thank you!

Jinsung Huh

Solutions Architect

