데이터 파이프라인과 AI 알고리즘의 AWS 활용

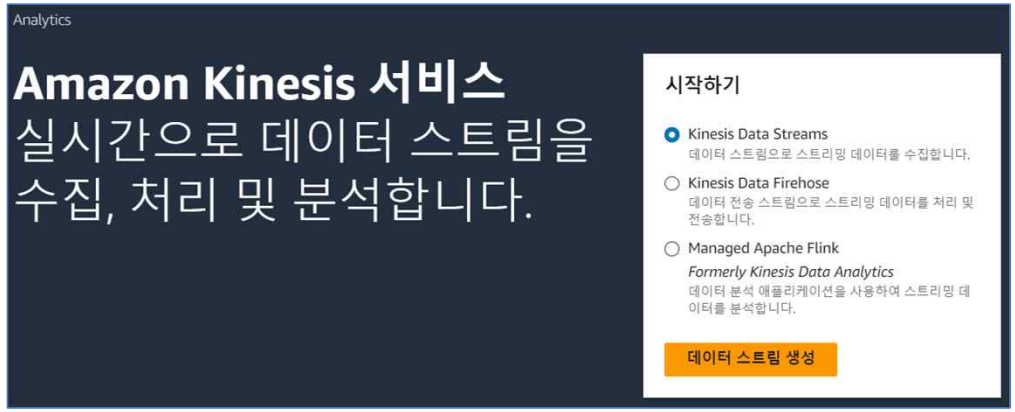
강사 : 고병화

10. Amazone Kinesis

AWS Kinesis

- Amazon Kinesis는 데이터 스트리밍 및 실시간 데이터 처리 서비스로, 대량의 데이터를 실시간으로 수집, 처리 및 분석할 수 있도록 설계된 AWS(Amazon Web Services)의 서비스 중 하나입니다.

Kinesis는 실시간 데이터 스트림을 처리하고 분석하는 데 사용되며, 다양한 데이터 소스에서 데이터를 수집하고 실시간으로 처리할 수 있습니다.



AWS Kinesis

Amazon Kinesis는 다음과 같은 주요 구성 요소로 구성되어 있습니다

1.Kinesis Data Streams: Kinesis의 핵심 구성 요소 중 하나로, 데이터 스트림을 생성하고 관리하는 데 사용됩니다. 데이터 스트림은 데이터 레코드의 연속으로 구성되며, 다양한 데이터 소스에서 생성된 데이터를 수집하는 데 사용됩니다.

2.Kinesis Data Firehose: Kinesis Data Streams로부터 데이터를 수집하고 다양한 대상으로 데이터를 전송하는 데 사용됩니다. 예를 들어, 데이터를 Amazon S3, Amazon Redshift, Amazon Elasticsearch와 같은 AWS 서비스로 직접 전송할 수 있습니다.

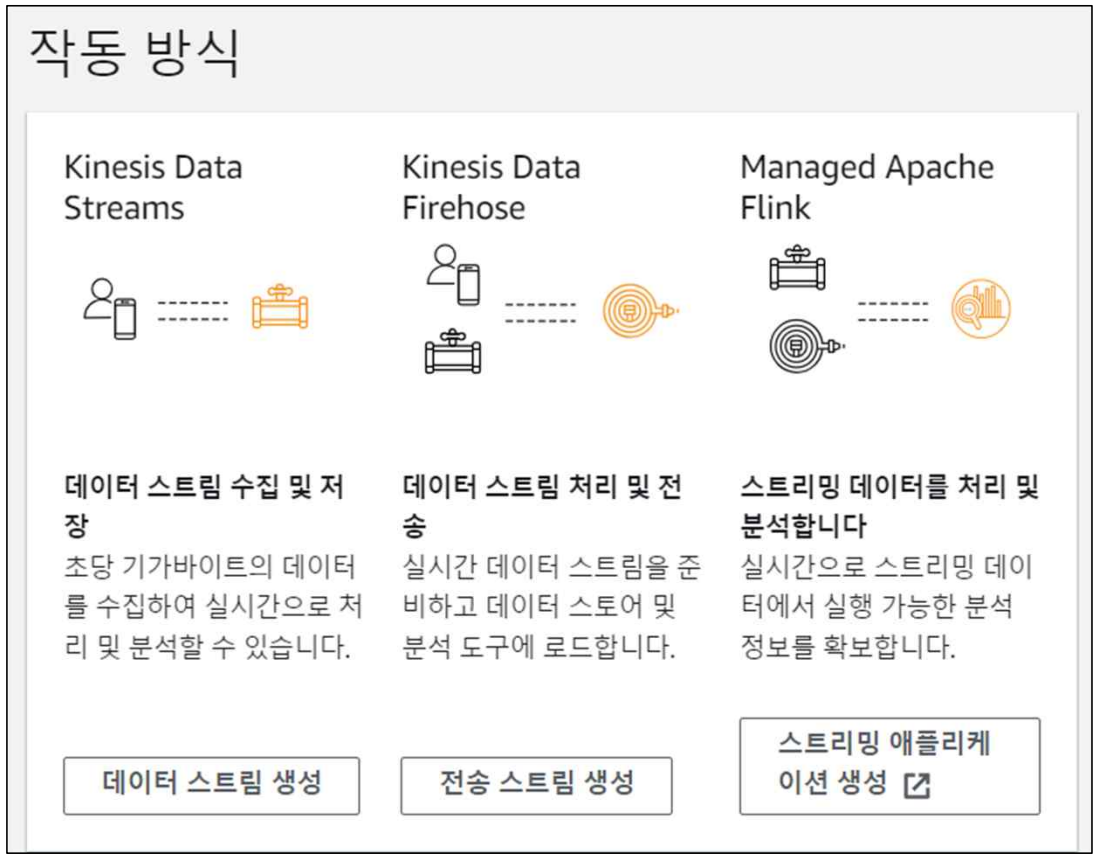
3.Kinesis Data Analytics: Kinesis 데이터 스트림에서 데이터를 실시간으로 분석하고 처리하는 데 사용됩니다. SQL 쿼리를 사용하여 데이터 스트림에서 데이터를 처리하고 원하는 결과를 생성할 수 있습니다.

Managed Apache Flink으로 이름 변경됨(2023년 09월 01일)

AWS Kinesis

4. Kinesis Video Streams:
비디오 스트림을 처리하고 분석하는 데 사용됩니다. 이것은 CCTV, IP 카메라 등의 비디오 스트리밍 데이터를 처리하는 데 유용합니다.

Amazon Kinesis는 실시간 데이터 처리 및 분석에 관심이 있는 개발자와 기업에게 강력한 도구로서 많이 사용되고 있습니다.



AWS Kinesis

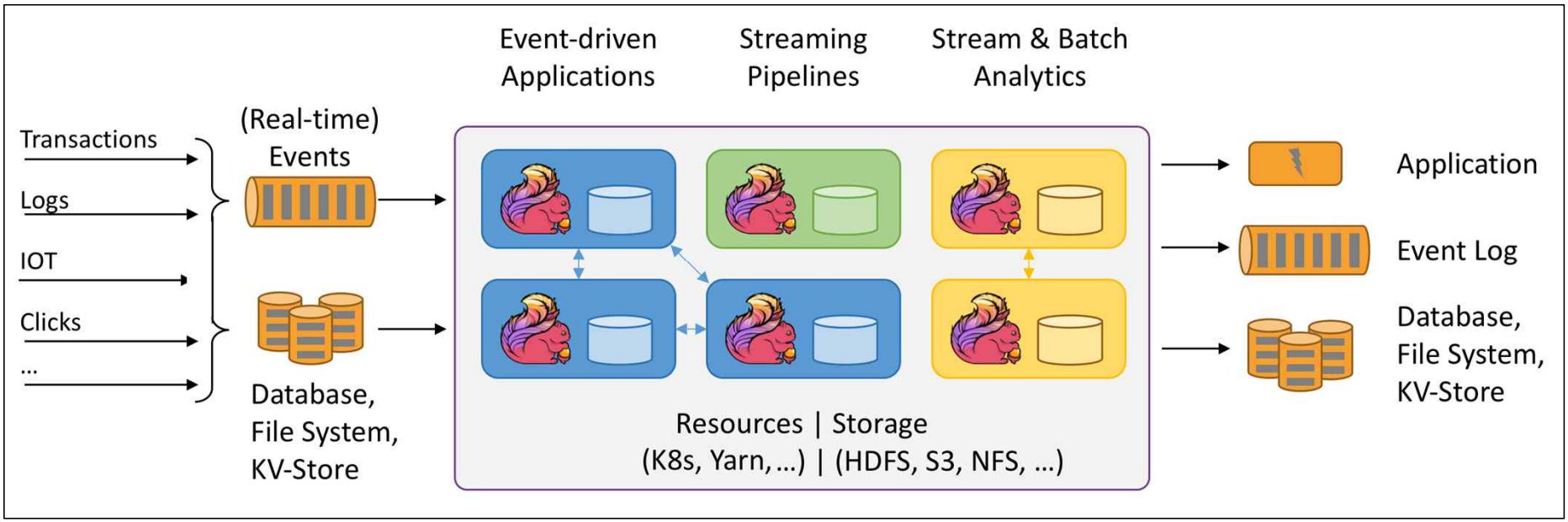
- 아파치 플링크(Apache Flink)란?

플링크(Flink)는 독일어로 민첩함을 뜻하는 단어로 베를린 TU대학교에서 시작된 아파치 프로젝트의 하나로, 분산 처리를 위한 오픈 소스 데이터 스트림 처리 및 배치 처리 프레임워크이다.
Flink은 데이터 처리를 위한 고성능, 고가용성, 확장성을 제공하며, 대용량의 데이터를 실시간 및 배치 처리를 통해 분석하고 처리하는 데 사용된다. 다양한 데이터 소스와 데이터 형식을 지원하며, 복잡한 데이터 처리를 위한 고급 기능을 제공하여 실시간 스트림 처리와 배치 처리를 하나의 통합적인 환경에서 처리할 수 있다.

https://sketchit.tistory.com/entry/%EC%95%84%ED%8C%8C%EC%B9%98-%ED%94%8C%EB%A7%81%ED%81%ACApache-Flink%EB%9E%80-%EB%AC%B4%EC%97%87%EC%9D%B8%EA%B0%80

AWS Kinesis

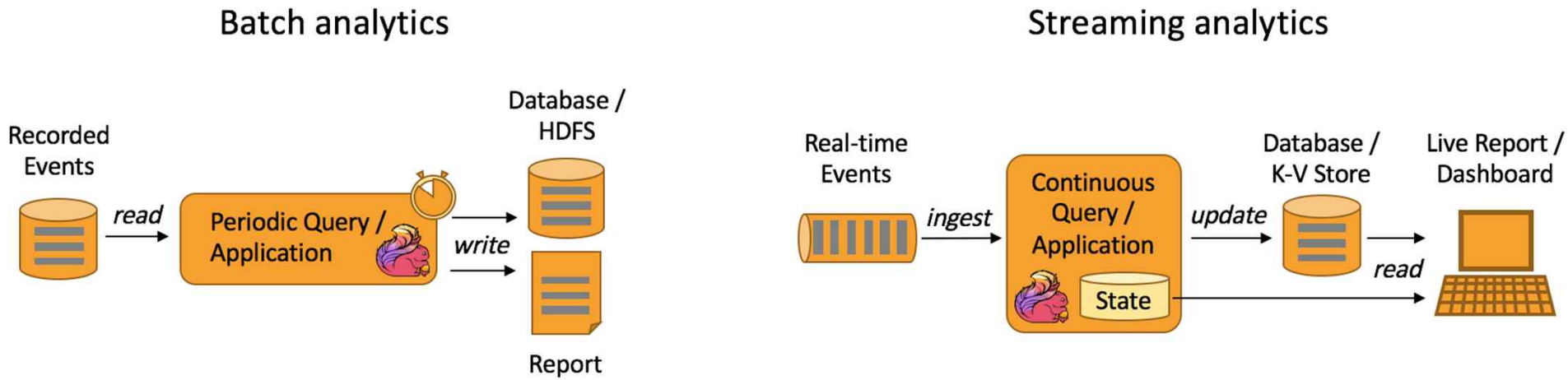
- 플링크(Flink)는 는 Data Stream에 대한 Stateful 연산을 수행하는 분산 처리 엔진으로 Event Stream, tables, graph, machine learning 등 모든 워크로드를 스트리밍 방식으로 처리한다.



AWS Kinesis

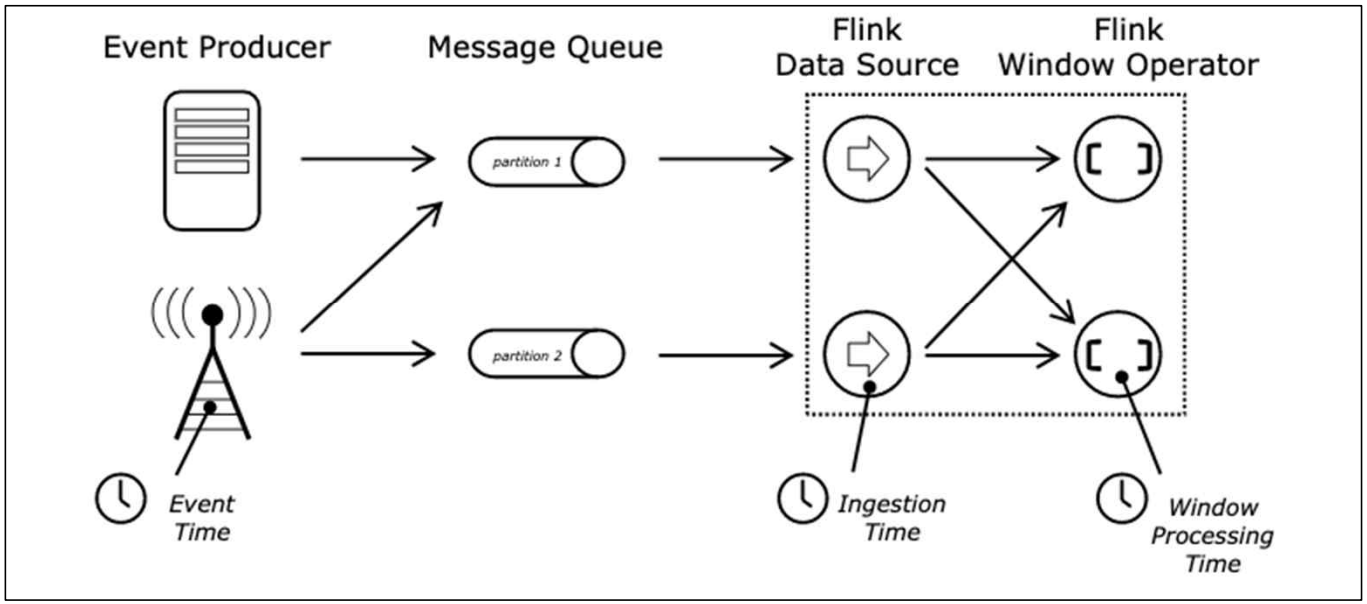
아파치 플링크(Apache Flink)의 특징

1. **스트림 및 배치 처리 지원**: Flink는 실시간 스트림 데이터와 배치 데이터를 모두 처리할 수 있어, 다양한 데이터 처리 요구에 대응할 수 있다.



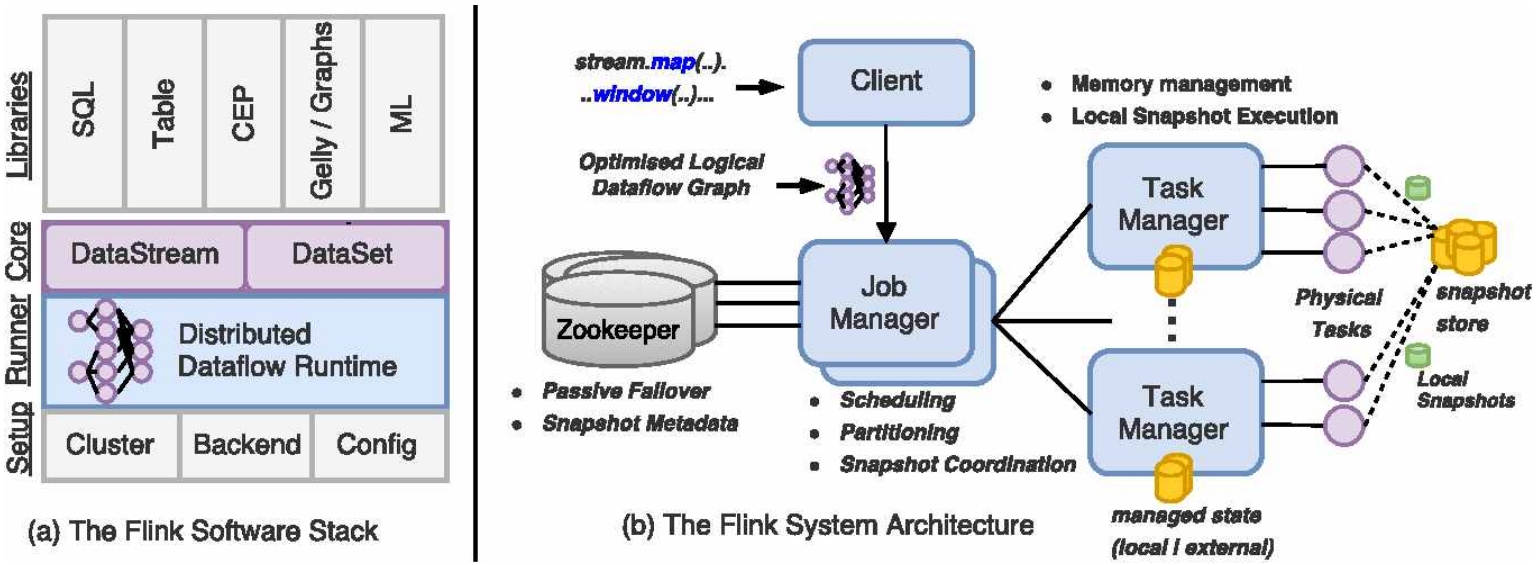
AWS Kinesis

2. **이벤트 시간 기반 처리**: Flink은 이벤트 시간 기반 처리를 지원하여 데이터의 시간적 특성을 고려하여 정확한 결과를 얻을 수 있다.



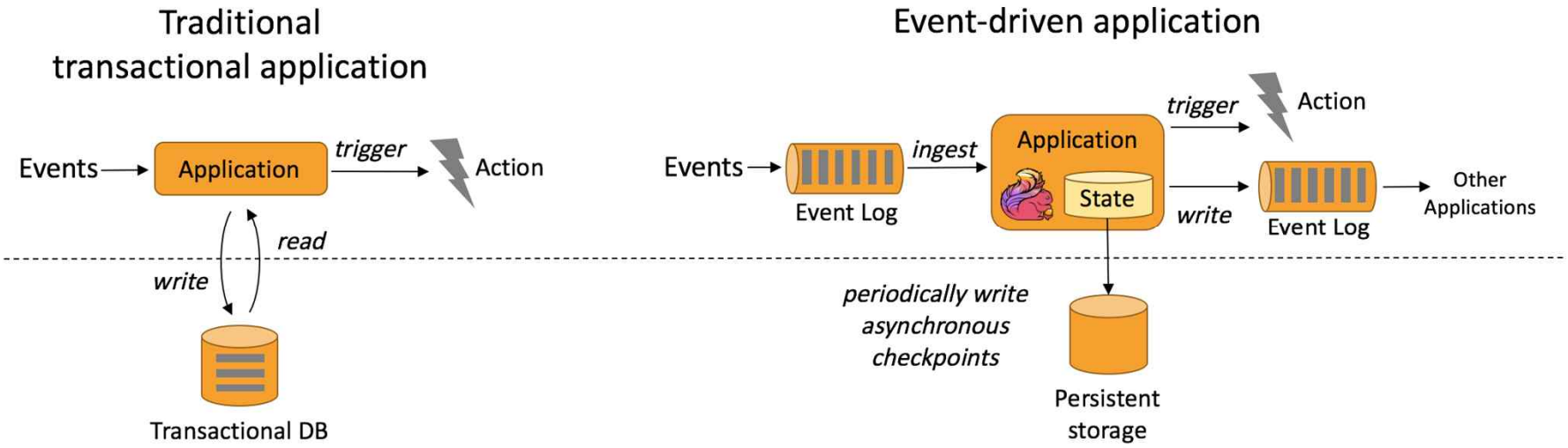
AWS Kinesis

3. **상태 관리**: Flink는 상태를 관리하고 유지할 수 있는 내장형 상태 관리 시스템을 제공하여 스트림 처리의 상태를 효과적으로 관리할 수 있다.



AWS Kinesis

4. **고급 이벤트 처리**: Flink는 데이터 윈도우, 이벤트 타임 기반의 처리, 상태 기반의 처리 등 다양한 고급 이벤트 처리를 제공하여 복잡한 데이터 처리를 가능하게 한다.



5. **분산 처리**: Flink은 분산 처리를 위한 클러스터 모드를 지원하여 대규모 데이터 처리를 수행할 수 있다.

AWS Kinesis

아파치 플링크(Apache Flink)의 구성

1. **Flink Core**: Flink의 핵심 런타임 엔진으로, 스트림 처리와 배치 처리를 지원

2. **Flink Streaming**: Flink의 실시간 스트림 처리를 위한 모듈로, 데이터 스트림을 처리하고 이벤트 시간 기반의 처리, 상태 관리, 윈도우 처리 등의 기능을 제공

3. **Flink Batch**: Flink의 배치 처리를 위한 모듈로, 대용량 데이터의 배치 처리를 수행

4. **Flink Table & SQL**: Flink의 SQL 기반의 쿼리 및 테이블 API를 제공하여 SQL과 유사한 문법으로 데이터 처리를 할 수 있다.

5. **Flink CEP**: Flink의 복잡한 이벤트 처리를 위한 모듈로, 복잡한 이벤트 패턴 매칭과 시간 기반의 처리를 지원한다

AWS Kinesis

사용 사례

솔루션 구현 - AWS Streaming Data Solution for Amazon Kinesis

AWS Streaming Data Solution for Amazon Kinesis는 두 개의 AWS CloudFormation 템플릿을 제공하여 개발자가 스트리밍 데이터를 캡처, 저장, 처리 및 전송하는 데 필요한 핵심 AWS 서비스를 보다 쉽게 구성할 수 있도록 지원합니다. 한 옵션은 모바일 클라이언트와 같은 비 AWS 환경에서 데이터를 캡처하고, API 수준에서 제한을 활성화하며, AWS Lambda를 활용하여 Amazon Kinesis 스트림에 대한 오류를 처리하도록 설계되었습니다. 두 번째 옵션은 Apache Flink를 활용하고 백업을 자동으로 처리하는 완전관리형 서비스를 제공합니다. 이 옵션은 Amazon KPL(Kinesis Producer Library)도 지원합니다. 자세히 알아보기

배치 분석에서 실시간 분석으로의 진화

Amazon Kinesis 서비스를 사용하면 일반적으로 배치 처리를 사용하여 분석된 데이터에 대한 실시간 분석을 수행할 수 있습니다. 일반적인 스트리밍 사용 사례에는 다양한 애플리케이션 간 데이터 공유, 스트리밍 추출–변환–로드 및 실시간 분석이 포함됩니다. 예를 들어 Kinesis Data Firehose를 사용하여 스트리밍 데이터를 S3 데이터 레이크 또는 분석 서비스로 지속적으로 로드할 수 있습니다.

실시간 분석 구축

실시간 애플리케이션 모니터링, 사기 탐지, 실시간 리더보드를 지원하는 Amazon Kinesis 서비스를 사용할 수 있습니다. Kinesis Data Streams를 사용하여 스트리밍 데이터를 수집하고, Kinesis Data Analytics를 사용하여 처리하며, 밀리초의 엔드투엔드(end-to-end) 지연 시간의 Kinesis Data Streams를 사용하여 모든 데이터 스토어 또는 애플리케이션으로 결과를 내보낼 수 있습니다. 고객, 애플리케이션 및 제품이 현재 진행 중인 작업을 알아보고 신속하게 대응하십시오.

IoT 디바이스 데이터 분석

Amazon Kinesis 서비스를 사용하여 가전제품, 내장형 센서, TV 셋톱박스 등 IoT 디바이스의 스트리밍 데이터를 처리할 수 있습니다. 그런 다음 센서가 특정 작동 임계값을 초과할 경우 이 데이터를 사용하여 실시간 알림을 보내거나 기타 작업을 프로그래밍 방식으로 수행할 수 있습니다. 애플리케이션을 빌드할 때 샘플 IoT 분석 코드를 사용할 수 있습니다.

샘플 IoT 분석 코드

고객 스포트라이트: Zappos

Zappos는 Kinesis Data Firehose를 사용해 이벤트 데이터를 수집하여 분석한 후 실시간으로 개인화된 크기 조정 및 검색 결과를 고객에게 제공합니다.

사례 연구 읽기

고객 스포트라이트: Autodesk

Autodesk는 Kinesis Data Firehose 및 Kinesis Data Analytics를 통해 로그 분석 솔루션을 구축하여 소프트웨어 문제를 최대한 빠르게 모니터링하고 해결합니다.

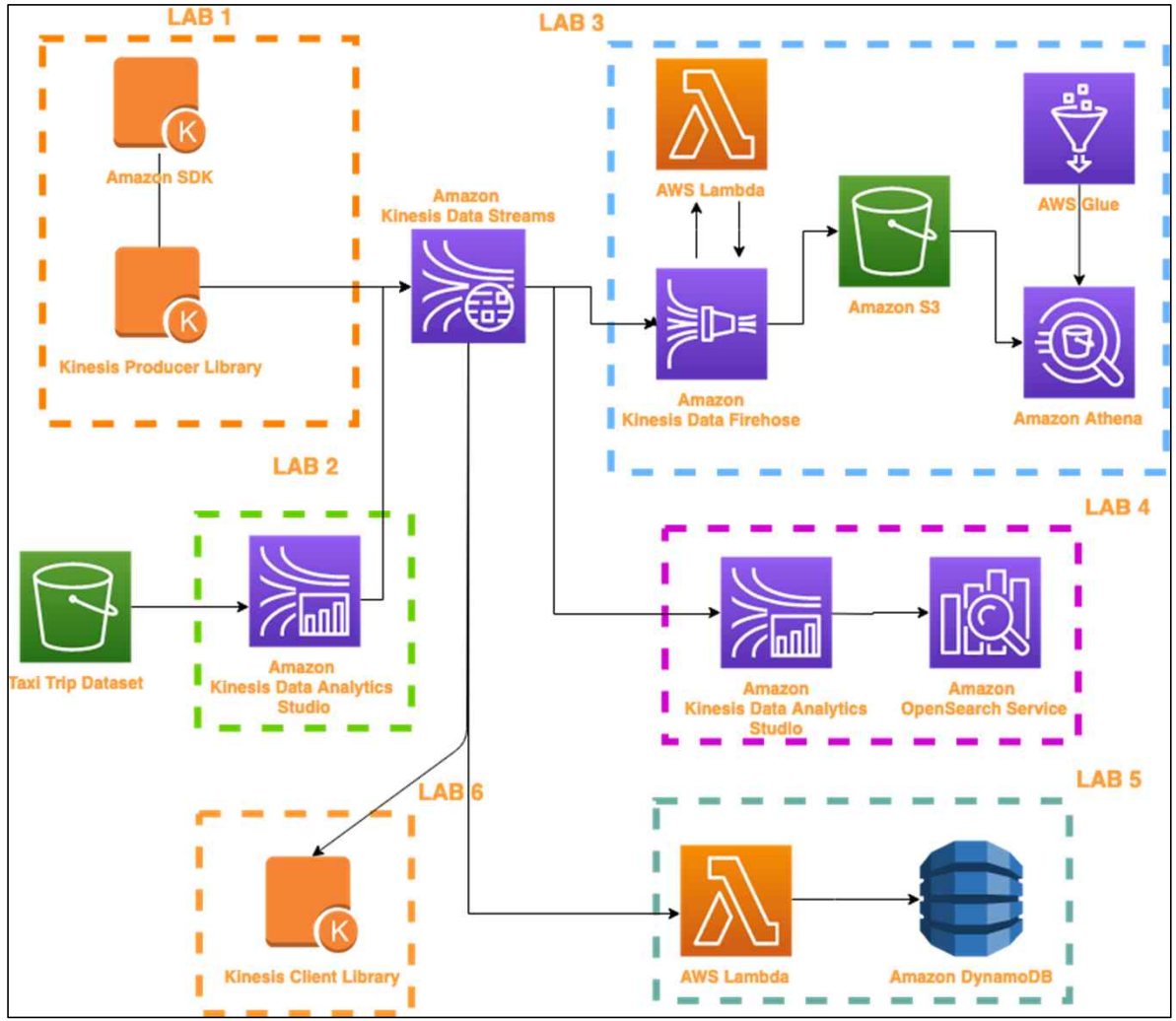
사례 연구 읽기

AWS Kinesis

- AWS Kinesis 실습

- Kinesis Data Streams로 데이터 생성
- Kinesis Data Analytics Studio 노트북을 사용하여 Kinesis Data Stream에 데이터 쓰기
- Kinesis Data Firehose를 사용한 Lambda
- Kinesis Data Analytics로 이벤트 정리,
- 집계 및 강화
- Kinesis 데이터 스트림용 Lambda 소비자
- Amazon KCL로 사용

https://catalog.us-east-1.prod.workshops.aws/workshops/2300137e-f2ac-4eb9-a4ac-3d25026b235f/en-US/



AWS Kinesis

Kinesis Data Streams로 데이터 생성

Kinesis Data Stream에 데이터를 쓰는 방법에는 여러 가지가 있습니다.

- Kinesis용 Amazon SDK 사용
- Kinesis 생산자 라이브러리 사용
- SDK 등을 호출하는 사용자 정의 프로그램 작성

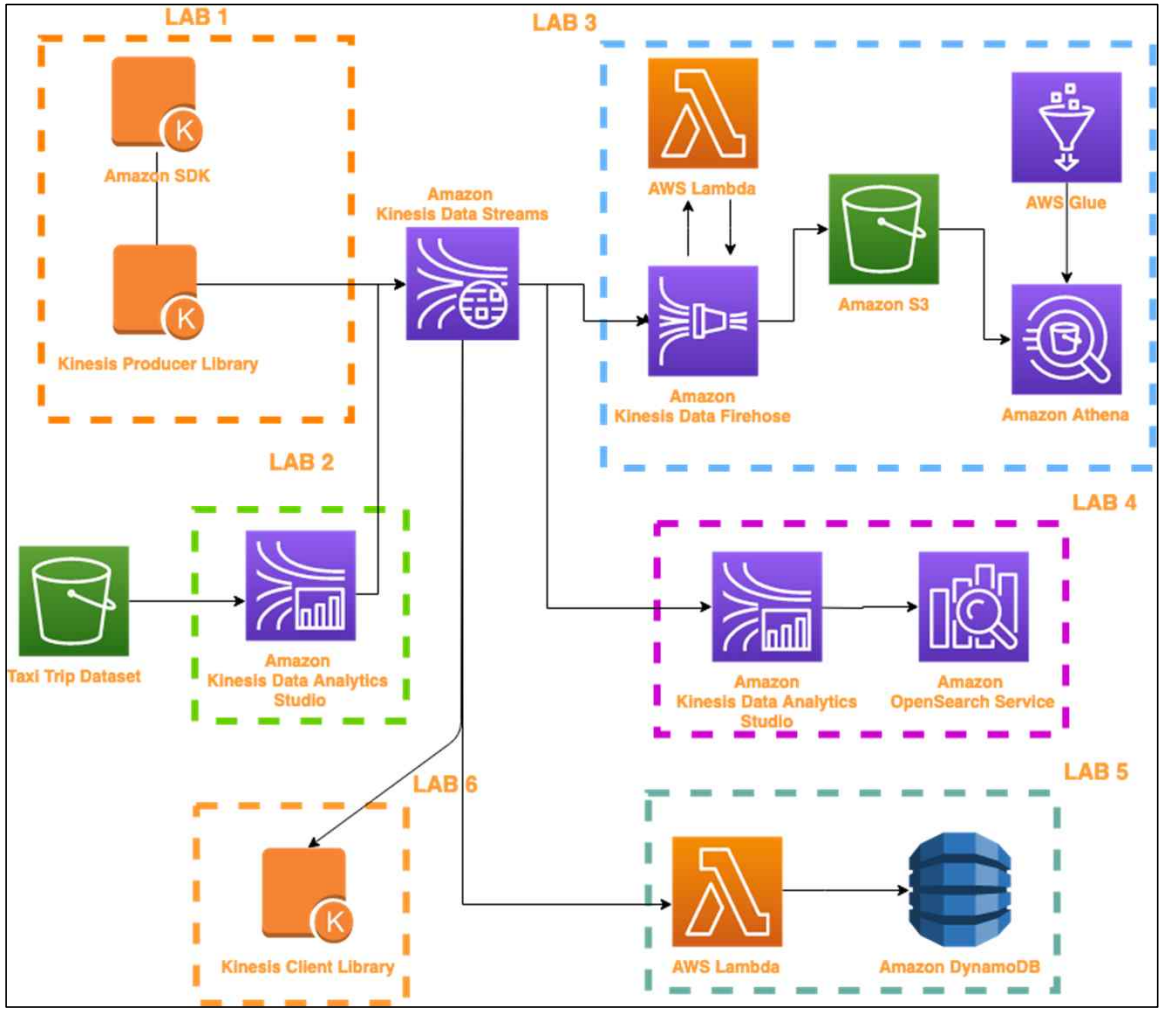
Kinesis 데이터 스트림에 데이터를 쓸 때 고려해야 할 다양한 고려 사항을 다루면서 Kinesis 데이터 스트림에 쓸 때 모범 사례를 살펴보겠습니다.

- AWS Cloud 9 IDE 구성
- Kinesis와 함께 Amazon SDK 사용
- Kinesis 생산자 라이브러리 사용

AWS Kinesis

실습 앞에서 CloudFormation 템플릿을 수행하면 다음 리소스를 생성한다

- Cloud9 인스턴스에 해당하는 보안 그룹이 있는 Virtual Private Cloud(VPC).
- Kinesis Client Library를 실행하고 코드를 탐색하기 위한 Cloud9 인스턴스
- 소스 Taxi Trip 데이터 세트를 저장하고 Kinesis Data Firehose 랩에서 선별된 데이터의 출력을 허용하는 2개의 S3 버킷
- Kinesis Data Analytics Studio 애플리케이션 및 연결된 Glue 데이터베이스
- Kinesis Data Streams의 데이터를 처리하는 Lambda 함수



The End