데이터 파이프라인과 AI 알고리즘의 AWS 활용

강사 : 고병화

6. AWS Data Pipeline

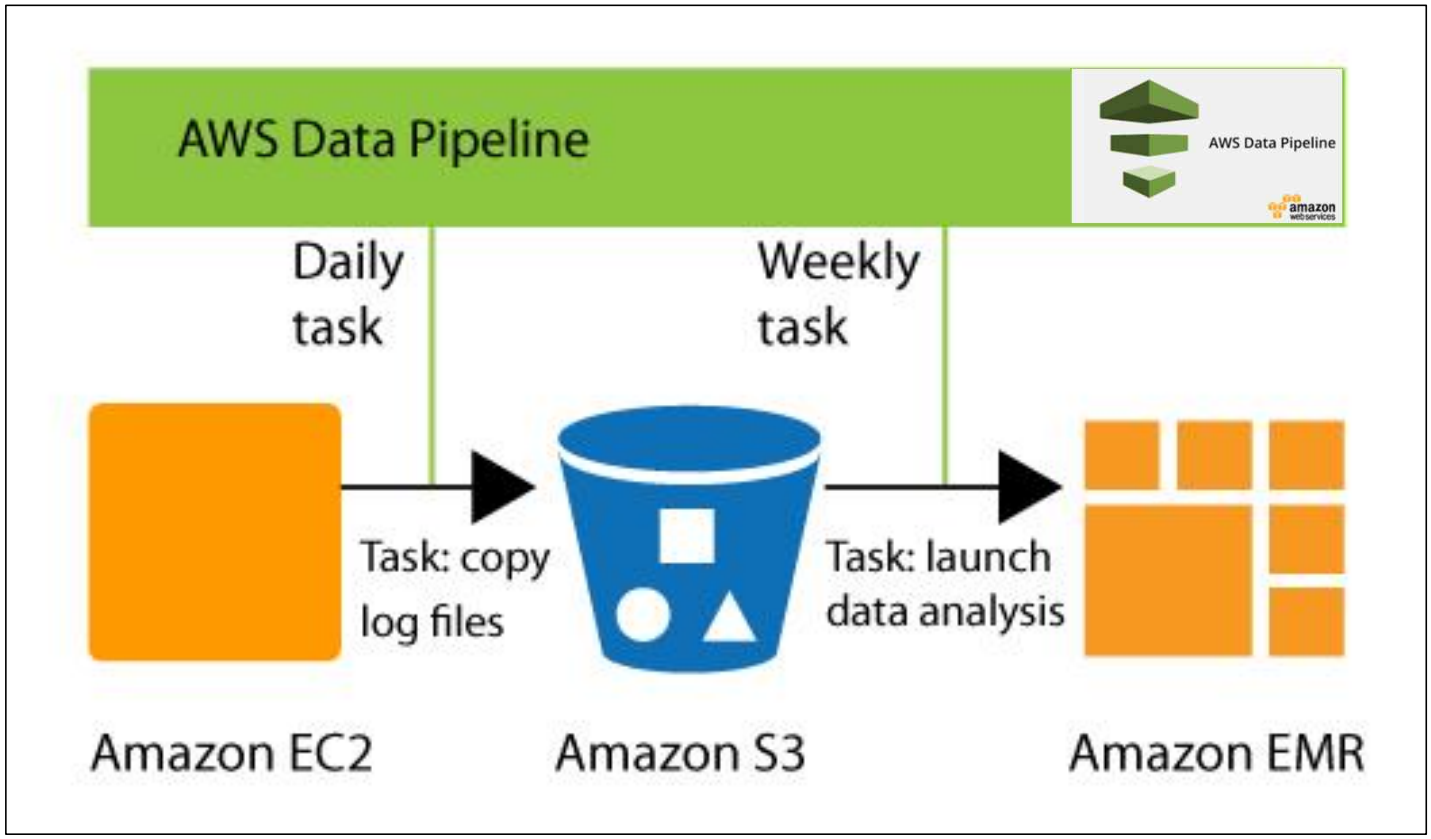
AWS Data Pipeline

데이터 파이프라인이란?

데이터 파이프라인은 다양한 데이터 소스에서 원시 데이터를 수집한 다음 분석을 위해 **데이터 레이크** 또는 **데이터 웨어하우스**와 같은 데이터 저장소로 이전하는 방법입니다. 일반적으로 데이터는 데이터 저장소로 이동하기 전에 데이터 처리 과정을 거칩니다. 여기에는 적절한 데이터 통합과 표준화를 보장하는 필터링, 마스킹, 집계와 같은 **데이터 변환**이 포함됩니다. 이 과정은 데이터 세트의 대상이 관계형 데이터베이스인 경우 특히 중요합니다. 이 유형의 데이터 저장소에는 기존 데이터를 새 데이터로 업데이트하기 위해 정렬(즉, 데이터 열 및 유형 매칭)이 필요한 정의된 **스키마**가 있습니다

https://www.ibm.com/kr-ko/topics/data-pipeline

AWS Data Pipeline

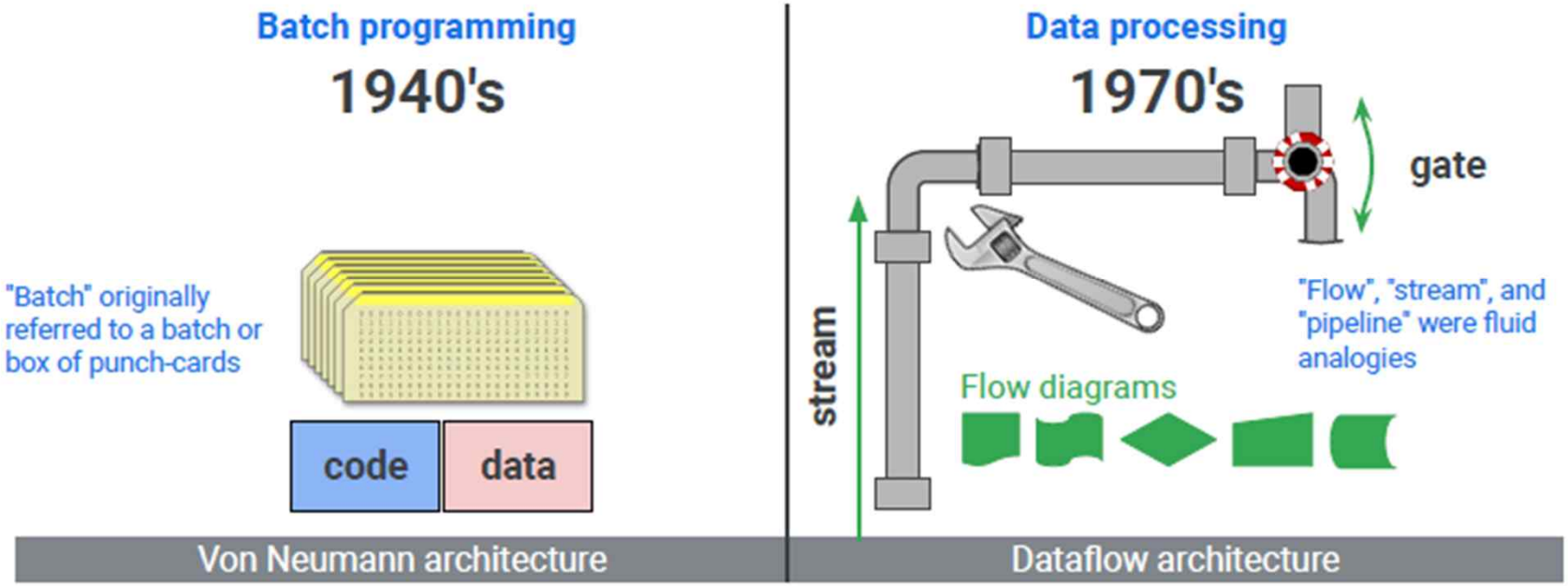


AWS Data Pipeline

데이터 파이프라인은 데이터 사이언스 프로젝트와 비즈니스 인텔리전스 대시보드에 "**파이프(배관)**" 역할을 하며, 다양한 데이터 소스에서 가져온 데이터를 처리하여 사용 가능한 형태로 만드는 역할을 합니다. 이를 위해 데이터 사이언티스트나 데이터 엔지니어는 데이터 준비 작업을 수행하여 데이터를 필터링, 병합, 요약하고 저장하며, 이를 기반으로 **탐색형 데이터 분석, 데이터 시각화, 머신 러닝 작업** 등 다양한 데이터 프로젝트를 구축할 수 있습니다.

AWS Data Pipeline

- 데이터 파이프라인에는 **일괄 처리(Batch Processing)**와 **스트리밍 처리(Streaming Processing)**의 두 가지 유형이 있다.



AWS Data Pipeline

- **ETL**

AWS Data Pipeline

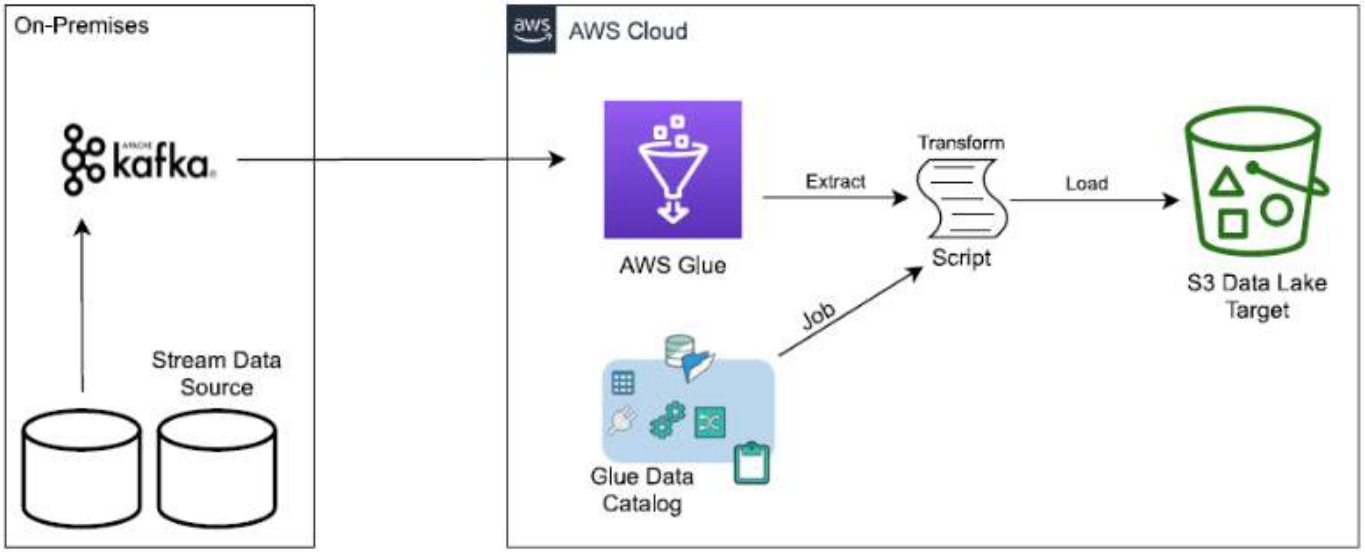
- **ETL은 어떻게 작동하나요?**

추출, 전환, 적재(ETL)는 데이터를 소스 시스템에서 대상 시스템으로 정기적으로 이동하는 방식으로 작동한다.
ETL 프로세스는 다음 세 단계로 작동한다.

1. 소스 데이터베이스에서 관련 데이터 추출 (Extract)
2. 분석에 더 적합한 형식으로 데이터 변환 (Transform)
3. 데이터를 대상 데이터베이스에 로드 (Load)

AWS Data Pipeline

- AWS ETL



https://aws.amazon.com/ko/what-is/etl/

AWS Data Pipeline

- AWS Data Pipeline에 액세스

다음 인터페이스 중 하나를 사용하여 파이프라인을 생성하고, 액세스하고, 관리할 수 있다.

- **AWS Management Console :** AWS Data Pipeline액세스에 사용할 수 있는 웹 인터페이스 (더 이상 제공되지 않음)
- **AWS Command Line Interface(AWS CLI)** : CLI 명령 제공
- **AWS SDK** : 언어별 API를 제공하고, 서명 계산, 요청 재시도 처리 및 오류 처리와 같은 많은 연결 세부 정보를 관리
- **Query API** : HTTPS 요청을 사용하여 호출하는 하위 수준 API를 제공

AWS Data Pipeline

- AWS Data Pipeline은 다음 서비스와 연계하여 데이터를 저장한다.

• Amazon DynamoDB : 저렴한 비용으로 빠른 성능을 갖춘 완전 관리형 NoSQL 데이터베이스를 제공

• Amazon RDS : 대규모 데이터 세트로 확장할 수 있는 완전 관리형 관계형 데이터베이스를 제공

• Amazon Redshift : 빠르고 완벽하게 관리되는 페타바이트 규모의 데이터 warehouse를 제공하므로 방대한 양의 데이터를 쉽고 비용 효율적으로 분석할 수 있다.

• Amazon S3 : 안전하고 안정적이며 확장성이 뛰어난 객체 스토리지를 제공

AWS Data Pipeline

- AWS Data Pipeline은 다음 컴퓨팅 서비스와 함께 작동하여 데이터를 변환한다.

• Amazon EC2 : 소프트웨어 시스템을 구축하고 호스팅하는 데 사용하는 크기 조정 가능한 컴퓨팅 용량(말 그대로 Amazon 데이터 센터의 서버)을 제공

• Amazon EMR — Apache Hadoop 또는 Apache Spark와 같은 프레임워크를 사용하여 Amazon EC2 서버 전반에 걸쳐 방대한 양의 데이터를 쉽고 빠르며 비용 효율적으로 배포하고 처리할 수 있다(EMR : Elastic MapReduce)

AWS Data Pipeline
파이프라인 구성요소

(1) 데이터 노드(Data Nodes)
작업에 대한 입력 데이터의 위치 또는 출력 데이터가 저장될 위치

(2) 활동(Activities)
계산 리소스와 일반적으로 입력 및 출력 데이터 노드를 사용하여 일정에 따라 수행할 작업을 정의한다.

(3) 전제 조건(Preconditions)
작업을 실행하기 전에 참이어야 하는 조건문

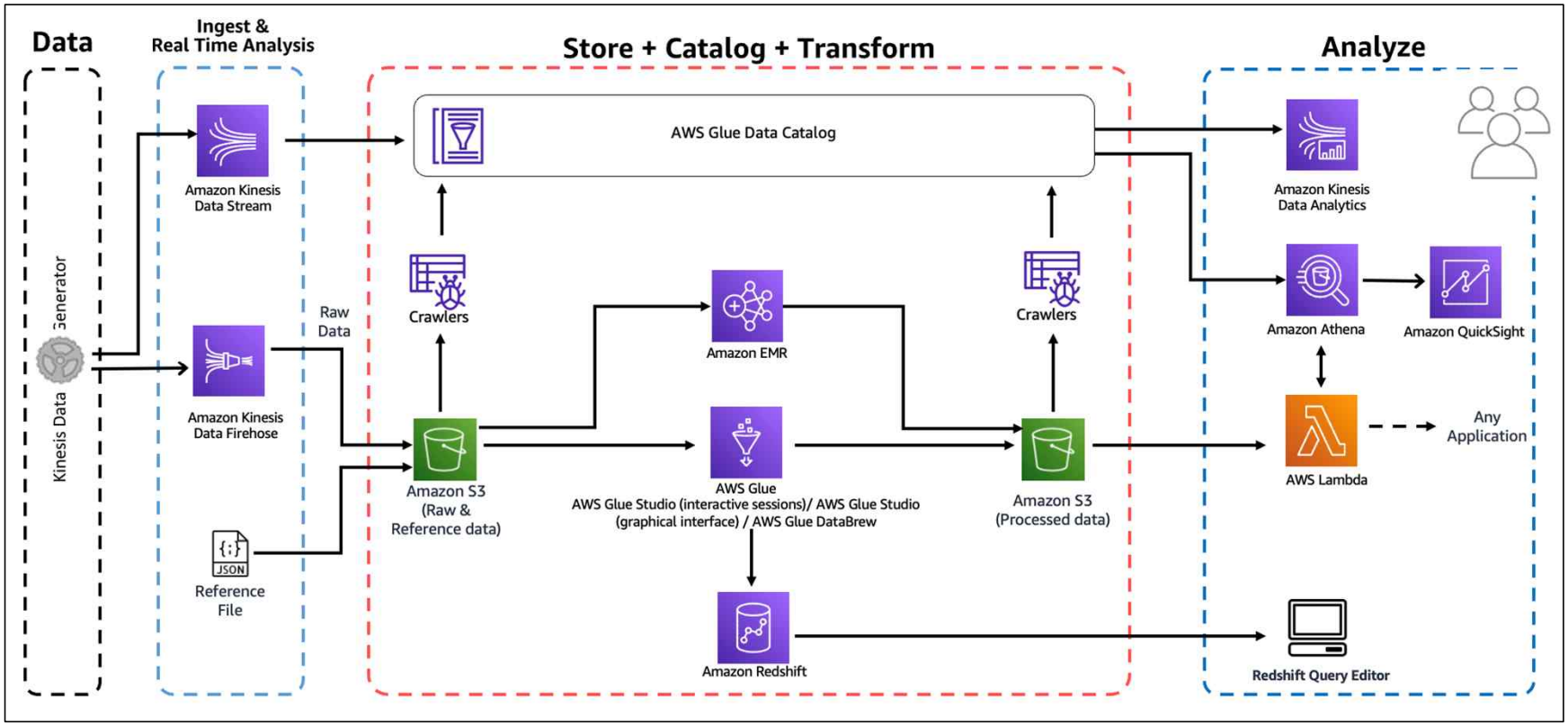
(4) 자원(Resources)
파이프라인이 정의하는 작업을 수행하는 계산 리소스

(5) 행위(Actions)
활동 실패 등 지정된 조건이 충족될 때 트리거 되는 작업

AWS Data Pipeline

- AWS Data Pipeline : Analytics on AWS Workshop

https://catalog.us-east-1.prod.workshops.aws/workshops/44c91c21-a6a4-4b56-bd95-56bd443aa449/en-US



The End