데이터 파이프라인과 AI 알고리즘의 AWS 활용

강사 : 고병화

11. AWS Glue

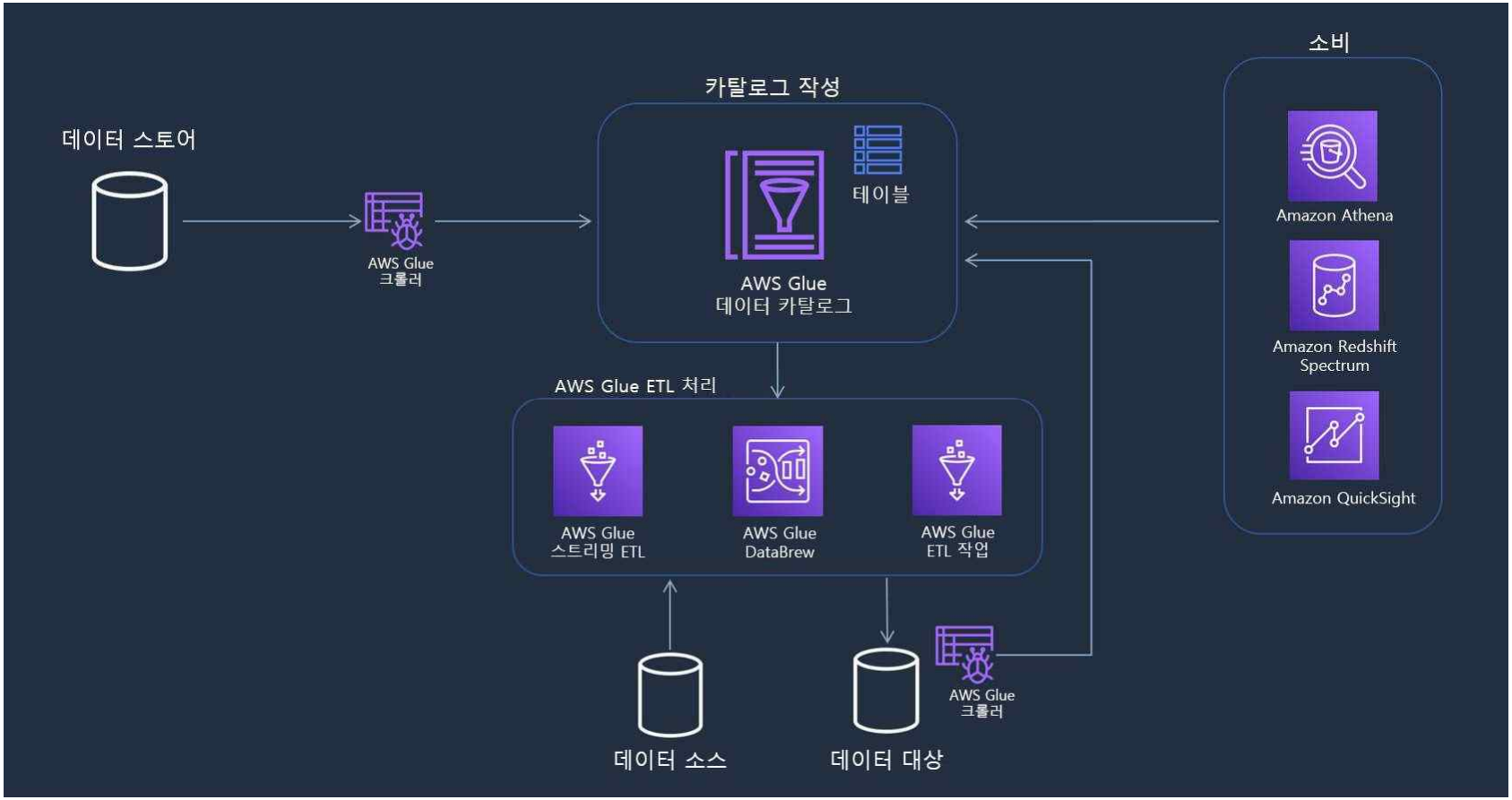
AWS Glue

AWS Glue 는 분석, 머신러닝, 애플리케이션 개발을 위한 데이터를 쉽게 검색, 준비, 결합할 수 있게 해주는 서버리스 데이터 통합 서비스이다
AWS Glue는 데이터 통합에 필요한 모든 기능을 제공하므로 몇 달이 아닌 몇 분만에 데이터 분석을 시작하고 사용할 수 있다.

AWS Glue는 데이터 통합을 더 쉽게 만들기 위해 시각적 및 코드 기반 인터페이스를 모두 제공한다. 사용자는 AWS Glue 데이터 카탈로그를 사용하여 데이터를 쉽게 찾고 액세스할 수 있다. 데이터 엔지니어와 ETL(추출, 변환 및 로드) 개발자는 AWS Glue Studio에서 몇 번의 클릭만으로 ETL 워크플로를 시각적으로 생성, 실행 및 모니터링할 수 있다.

데이터 분석가와 데이터 과학자는 AWS Glue DataBrew를 사용하여 코드를 작성하지 않고도 데이터를 시각적으로 강화하고 정리하고 정규화할 수 있다

AWS Glue


데이터 스토어
AWS Glue
크롤러
카탈로그 작성
테이블
AWS Glue
데이터 카탈로그
소비
Amazon Athena
Amazon Redshift
Spectrum
Amazon QuickSight
AWS Glue ETL 처리
AWS Glue
스트리밍 ETL
AWS Glue
DataBrew
AWS Glue
ETL 작업
AWS Glue
크롤러
데이터 소스
데이터 대상

AWS Glue

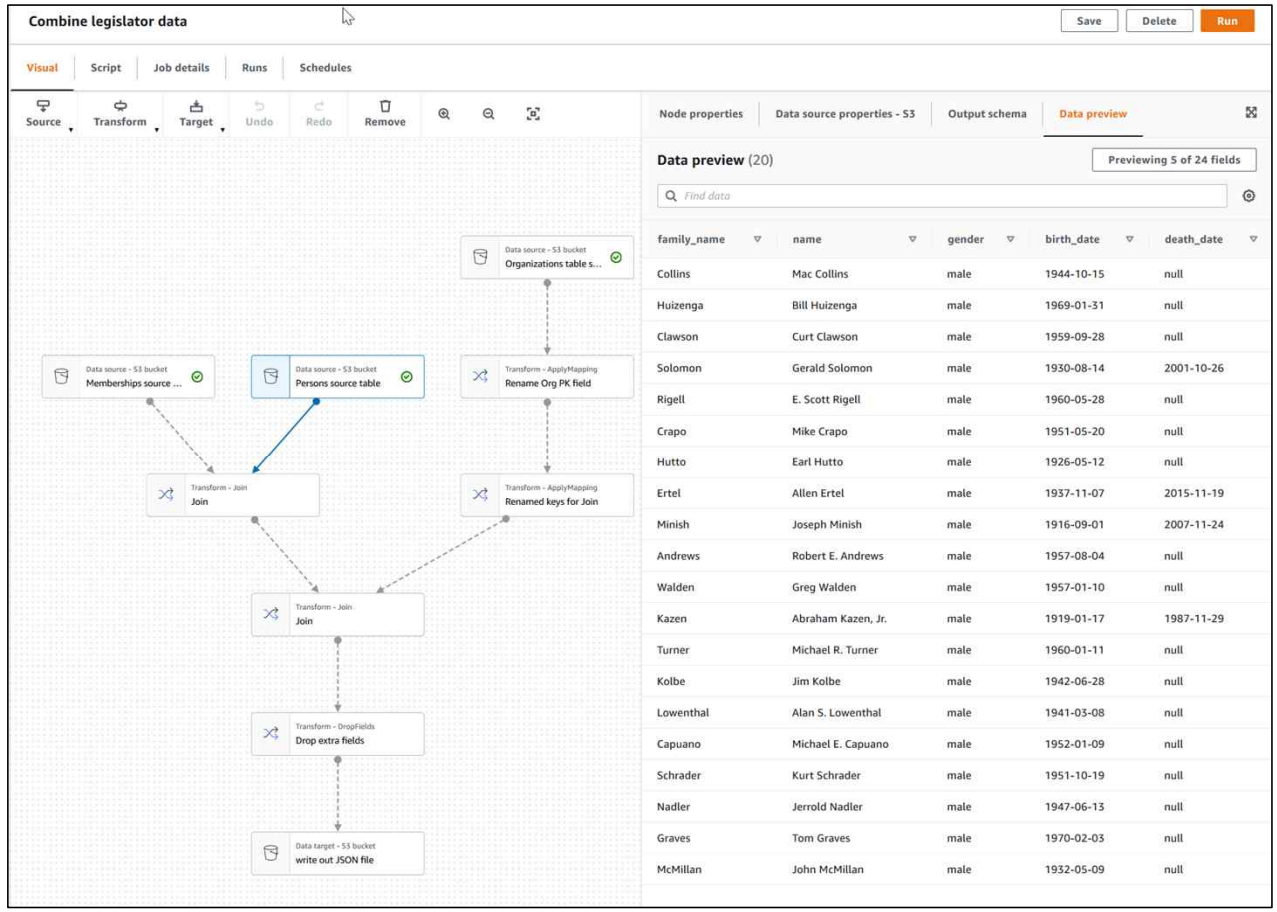
AWS Glue는 다음을 통해 많은 태스크를 간소화한다.

- 리소스 프로비저닝 및 관리
- ETL 소프트웨어 설치, 패치 적용 또는 업데이트와 같은 태스크 수행 불필요
- 작업 구성을 기반으로 코드 생성
- 비주얼 및 코드 기반 인터페이스 모두 포함

AWS Glue

- AWS Glue Studio

AWS Glue Studio는 AWS Glue에서 추출, 변환, 로드 작업을 쉽게 생성, 실행, 모니터링할 수 있게 해주는 새로운 그래픽 인터페이스이다. 데이터 변환 워크플로를 시각적으로 구성하고 AWS Glue의 Apache Spark 기반 서버리스 ETL 엔진에서 원활하게 실행할 수 있다. 작업의 각 단계에서 스키마 및 데이터 결과를 검사할 수 있다.



AWS Glue

AWS Glue Studio는 다음을 쉽게 수행할 수 있는 시각적 인터페이스를 제공한다

- Amazon S3, Amazon Kinesis 또는 JDBC 소스에서 데이터를 가져옵니다.
- 데이터를 조인, 샘플 또는 변환하는 트랜스포메이션을 구성합니다.
- 변환된 데이터의 대상 위치를 지정합니다.
- 작업의 각 지점에서 스키마 또는 데이터 집합의 샘플을 봅니다.
- AWS Glue Studio에서 생성된 작업을 실행, 모니터링, 관리합니다.

AWS Glue

언제 AWS Glue Studio를 사용해야 하는가?

AWS Glue Studio를 사용하면 ETL 개발자가 반복 가능한 프로세스를 쉽게 생성하여 대규모의 반정형 데이터 집합을 이동 및 변환하고 데이터 레이크와 Data Warehouse에 로드 할 수 있다.

필요에 따라 코드로 사용자 지정할 수 있는AWS Glue ETL 워크플로 개발 및 관리를 위한 boxes-and-arrows 스타일 시각적 인터페이스를 제공한다.

AWS Glue Studio기존 ETL 도구의 사용 편의성과AWS Glue 의 빅 데이터 처리 엔진의 강력함 및 유연성을 결합한다.
AWS Glue Studio는 시각적 편집기에서 코드 조각을 나타내는 노드를 추가하는 등 ETL 스크립트를 사용자 지정하는 여러 가지 방법을 제공한다.

AWS Glue

- AWS Glue 실습 (2~3 시간 소요)

- 전제조건 : 워크숍을 위한 AWS 환경을 설정합니다.
- 랩 1 : AWS Glue를 사용하여 Glue 데이터 카탈로그 및 Glue 크롤러를 사용하여 메타데이터 스키마를 검색하고 저장합니다.
- 랩 2 : 원하는 AWS Glue 개발 도구(예: Glue 개발 엔드포인트, Glue Studio 노트북 또는 대화형 세션)를 사용하여 Glue 스크립트 생성 및 테스트를 위한 통합 개발 환경을 만듭니다. 표준 PySpark 및 Glue 기반 PySpark를 사용하여 Glue ETL(추출, 변환, 로드) 코드를 개발하고 Glue에서 타사 Python 라이브러리를 사용하는 방법을 코드 샘플로 시연합니다.
- 랩 3 : 일반 Glue ETL 작업을 개발, 패키징 및 배포하고 Glue 트리거를 사용하여 작업 실행을 관리합니다.
- 랩 4 : 노트북 개발 환경에서 Glue 스트리밍 ETL 코드를 개발하는 방법과 이를 Glue 스트리밍 작업으로 패키징 및 배포하여 AWS Kinesis 데이터 스트림의 데이터를 처리하는 방법입니다.
- 랩 5 : 시각적 데이터 준비 도구인 Glue DataBrew를 사용하여 DataBrew 데이터 세트 및 데이터 세트 프로필을 생성하는 방법. DataBrew 프로젝트 내에서 작업하고 DataBrew 레시피를 생성하고, DataBrew 레시피를 관리하고, DataBrew 작업을 실행합니다.
- 랩 6 : Glue Studio를 사용하여 ETL 및 스트리밍 ETL 작업을 생성하는 방법. 랩은 노트북 개발 환경에서 Glue 스크립트를 사용하여 사용자 정의 변환을 생성하는 방법에 중점을 둡니다.
- 정리 : 비용이 청구되지 않도록 랩용으로 생성된 리소스를 정리합니다.

https://catalog.us-east-1.prod.workshops.aws/workshops/aaaabcab-5e1e-4bff-b604-781a804763e1/en-US

The End