

Getting Started with AWS Glue

Glue는 분석, 머신러닝 및 애플리케이션 개발에 필요한 데이터를 **탐색, 준비, 조합**할 수 있도록 지원하는 서버리스 데이터 통합 서비스

데이터 통합에 필요한 모든 기능 제공

비주얼 및 코드 기반 인터페이스 모두 제공 -> 데이터 통합 지원

사용자는 AWS Glue 데이터 카탈로그를 통해 데이터를 찾고 액세스 가능

데이터 엔지니어는 ETL(추출, 변환, 로드)를 수행

AWS Glue DataBrew를 이용해 코드를 작성하지 않고도 데이터를 시각화 가능

주요 기능

-ETL 작업에 필요한 서버, 스토리지 및 런타임 환경과 같은 **리소스를 프로비저닝하고 관리**.

-완전관리형 서비스 이므로, ETL SW 설치, 패치 적용 또는 업데이트와 같은 인프라 **유지**를 할 필요가 없음

- **작업구성을 기반으로 코드 생성**. 스크립트로도 가능

- **비주얼 및 코드 기반 인터페이스**는 데이터를 빠르고 비용 최적화된 방식으로 준비하고 이전하는데 도움

이점

1. 보다 빠른 데이터 통합
2. 크기에 맞게 데이터 통합 자동화
3. 관리할 인프라가 없음
4. 사용한 만큼만 지불

Glue 작동 방식

1. 데이터 스토어 (외부든 어디든)의 데이터를 AWS Glue 크롤러가 수집함.
2. 이를 통해 데이터 카탈로그를 만들
3. 데이터 카탈로그를 통해 Athena, Redshift, QuckSight를 실행할 수 있고, 또한 카탈로그를 통해 AWS Glue ETL 작업 수행
4. 데이터 소스를 통해 데이터 대상으로도 만들 수 있음

용례

1. ETL 파이프라인 구축
2. 코딩이 필요 없는 데이터 준비 및 데이터 프로파일링
3. 드래그 앤 드롭 방식으로 빠른 작업 오케스트레이션 및 시각화
4. 실시간 데이터 처리

주의해야 할 점

1. 데이터 분할 -> 성능 향상 및 비용 절감을 위해
2. 압축 및 파일 형식 -> 네트워크 비용 절감 및 성능
3. 보안 및 암호화
4. 작업 모니터링

AWS Glue에 대한 기본 개념

연결 : 특정 데이터 스토어에 연결하는 데 필요한 속성을 포함하는 데이터 카탈로그 객체입니다. Amazon S3는 AWS Glue에 연결할 필요가 없습니다.

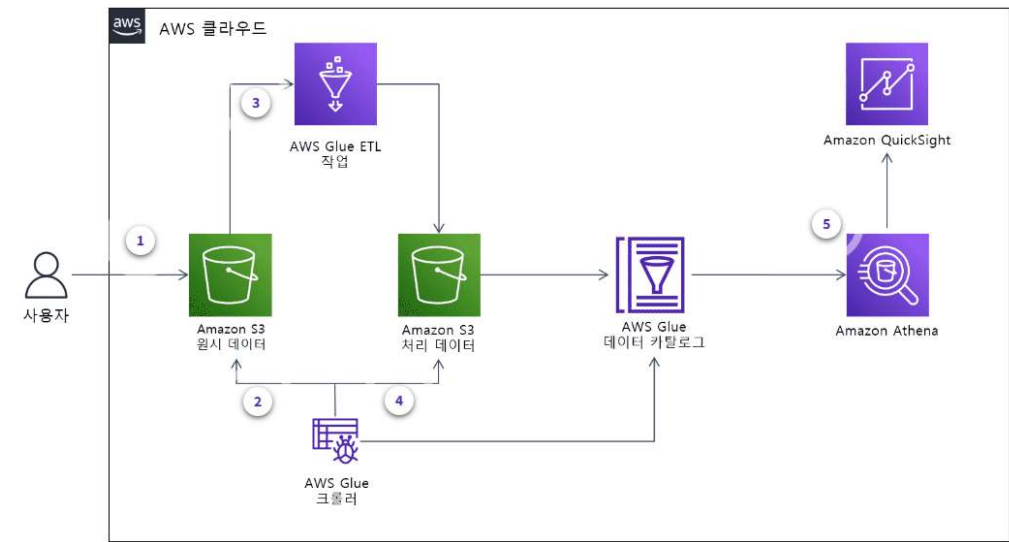
크롤러는 프로그램

데이터 소스는 프로세스 또는 변환에 대한 입력으로 사용되는 데이터 스토어
데이터 대상은 프로세스 또는 변환이 쓰기를 수행하는 데이터 스토어

동적프레임 : 구조 및 배열과 같은 중첩 데이터를 지원하는 분산 테이블. 각 레코드에는 데이터와 자신을 설명하는 메타 데이터가 포함되어 있음

Parquet 형식 : 일반적인 행 기반 형식이 아닌 열 형식으로 데이터를 구성하는 파일 형식

아키텍처



아키텍처 설명

아키텍처는 사용자, S3 버킷, AWS Glue 크롤러, 데이터 카탈로그, AWS Glue ETL 작업, Athena 및 QuickSight로 구성됩니다.

1. 사용자가 S3 버킷 원시 데이터 폴더에 데이터를 업로드합니다.
2. 데이터가 들어오면 AWS Glue 크롤러가 시작되고 원시 또는 처리된 데이터 폴더를 크롤링한 다음, 메타데이터 및 원시 테이블로 데이터 카탈로그를 업데이트합니다.
3. 데이터 카탈로그가 채워지면 AWS Glue ETL 작업이 시작됩니다. 카탈로그를 소스로 사용하여 S3 버킷 내부의 원시 데이터에 대한 모든 정보를 가져온 다음, 파일을 Parquet으로 변환합니다. AWS Glue ETL 작업이 완료되면 변환된 새 데이터를 동일한 S3 버킷의 처리된 폴더에 작성합니다.
4. 데이터가 처리된 폴더에 들어오면 크롤러가 다시 시작됩니다. 처리된 테이블과 새로 처리된 데이터의 메타데이터를 데이터 카탈로그에 추가합니다.
5. 처리된 데이터를 이제 사용할 수 있습니다. Athena 및 QuickSight와 같은 다른 분석 서비스가 연결되어 분석을 수행할 수 있도록 데이터 카탈로그에서 제공합니다.

hands-on은 Skill builder에

workflow는 스케줄링 및 자동화를 수행하는 데 쓰인다.

ETL 프로젝트가 주어진다면....

먼저 해야할 것은 데이터 파악이다 (갯수, 특성, 양, 특이사항 등)

sage maker