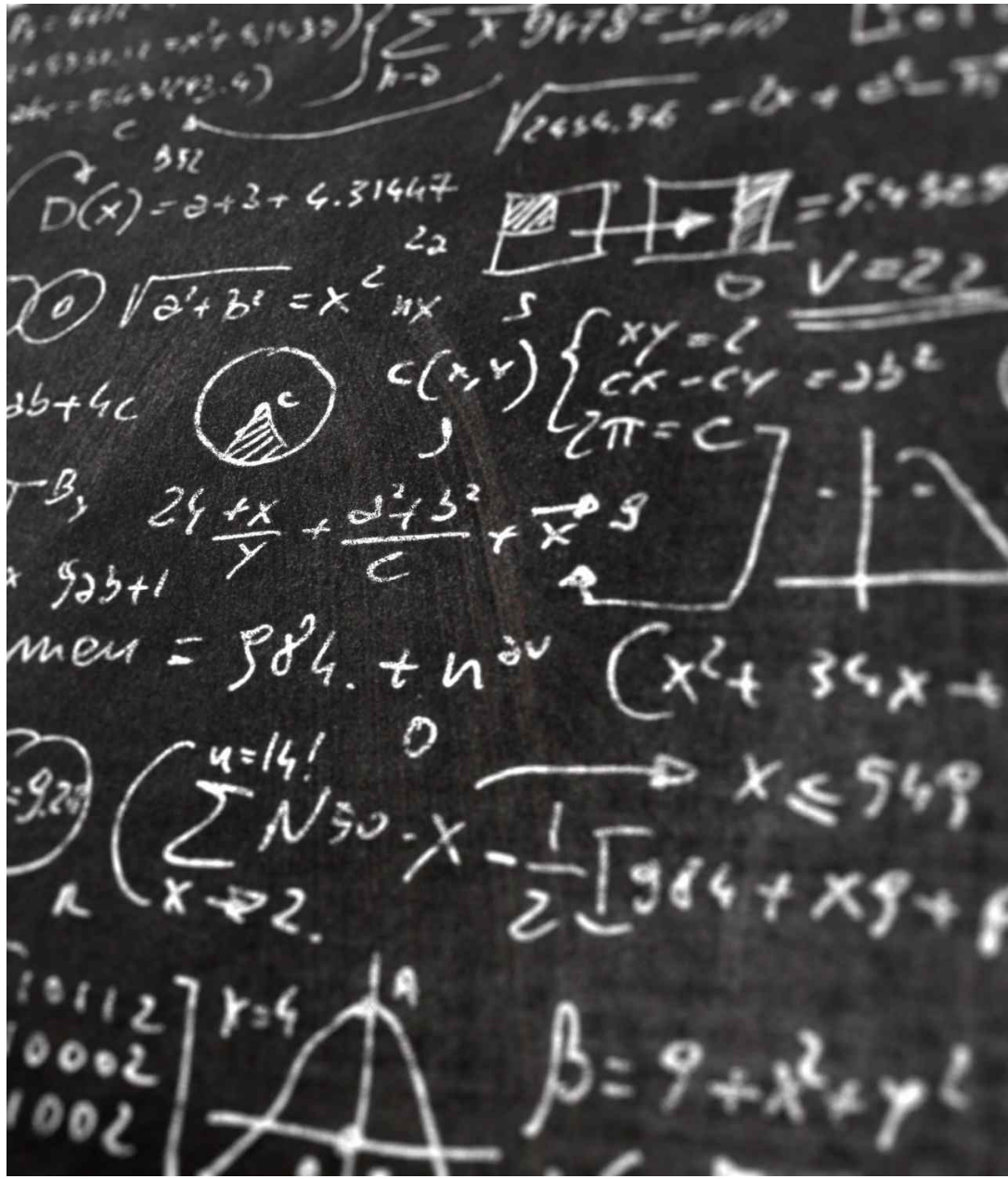
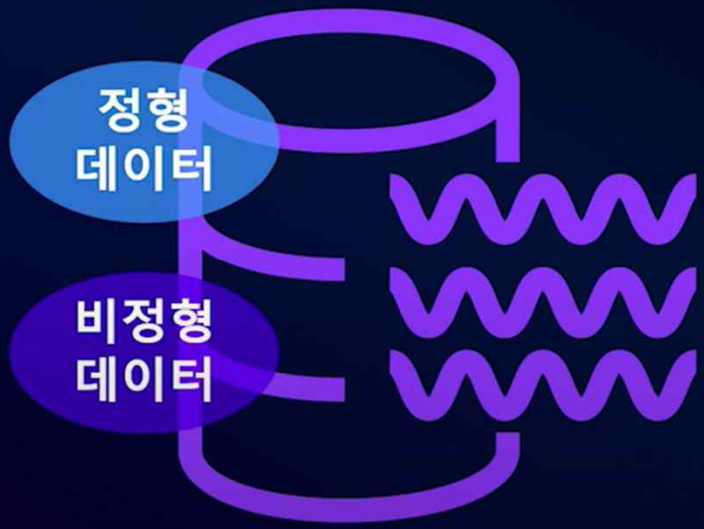
AWS 데이터 레이크 구축

[구성 요소]



데이터 레이크는?

" 모든 규모의 정형 및 비정형 데이터를 저장할 수 있는 중앙 집중식 리포지토리 "



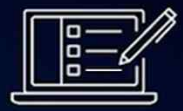
데이터 레이크와 데이터 웨어하우스 비교

구분	데이터 레이크	데이터 웨어하우스
데이터 형태	• 정형, 반정형, 비정형 데이터를 원형태(raw data)로 저장 • 데이터를 정제하지 않고 있는 그대로 저장	• 데이터를 구조화된 형태로 저장 • 업무 분석 요구에 맞추어 데이터를 정제, 가공하여 저장
스키마 요건	사전 스키마 설계 관련 요구조건은 따로 없음	데이터 저장 전에 스키마 설계 필요
데이터 신뢰성	데이터를 원형태(raw)로 저장하며 이로 인해 데이터의 품질은 다소 떨어짐	데이터 정제 및 가공 과정을 거치기 때문에 데이터에 대한 신뢰성이 높음
설계 지향점	성능보다는 스토리지 볼륨과 비용을 우선시하여 설계함	빠른 쿼리 성능을 제공할 수 있도록 설계함

데이터 레이크의 핵심 요소들

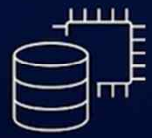
중앙 스토리지

Amazon Simple Storage Service (Amazon S3)



카탈로그 및 검색

사용자 액세스

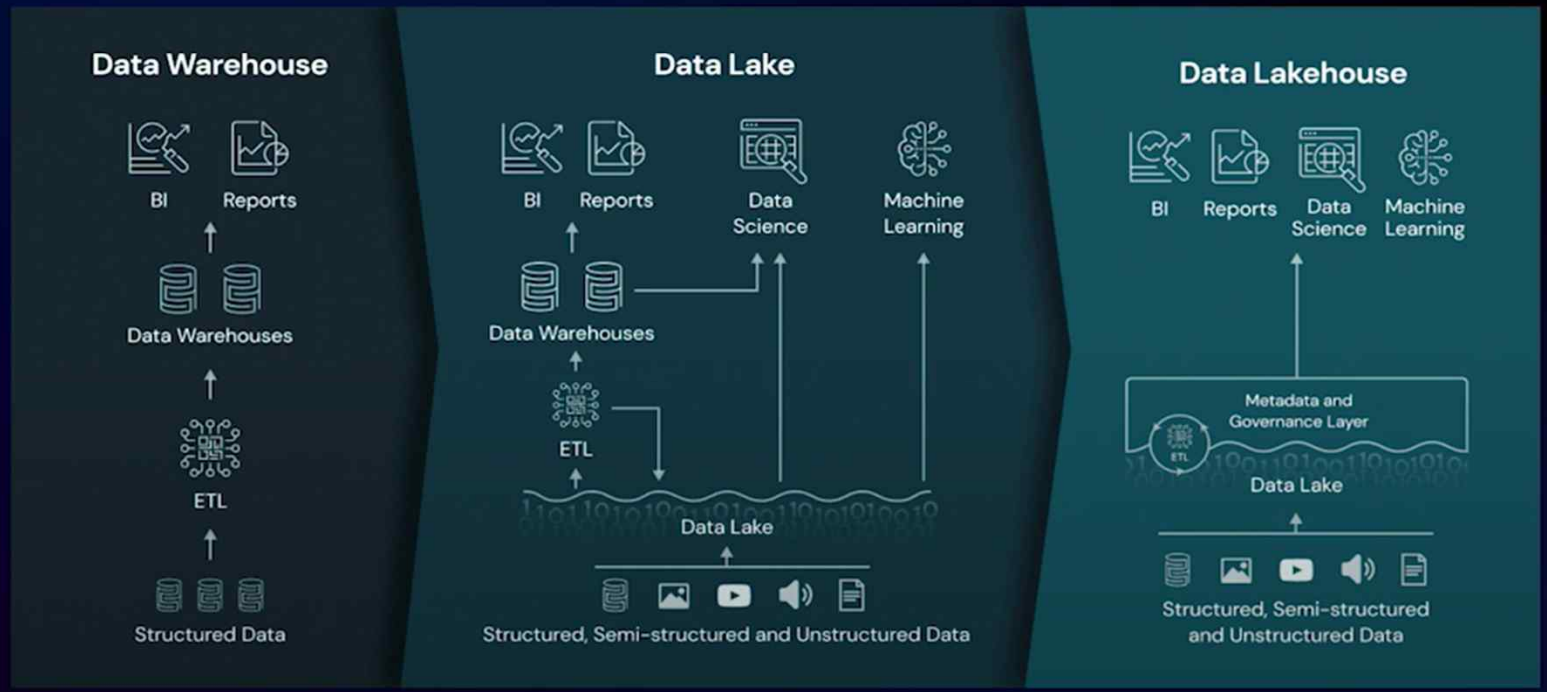


데이터 수집

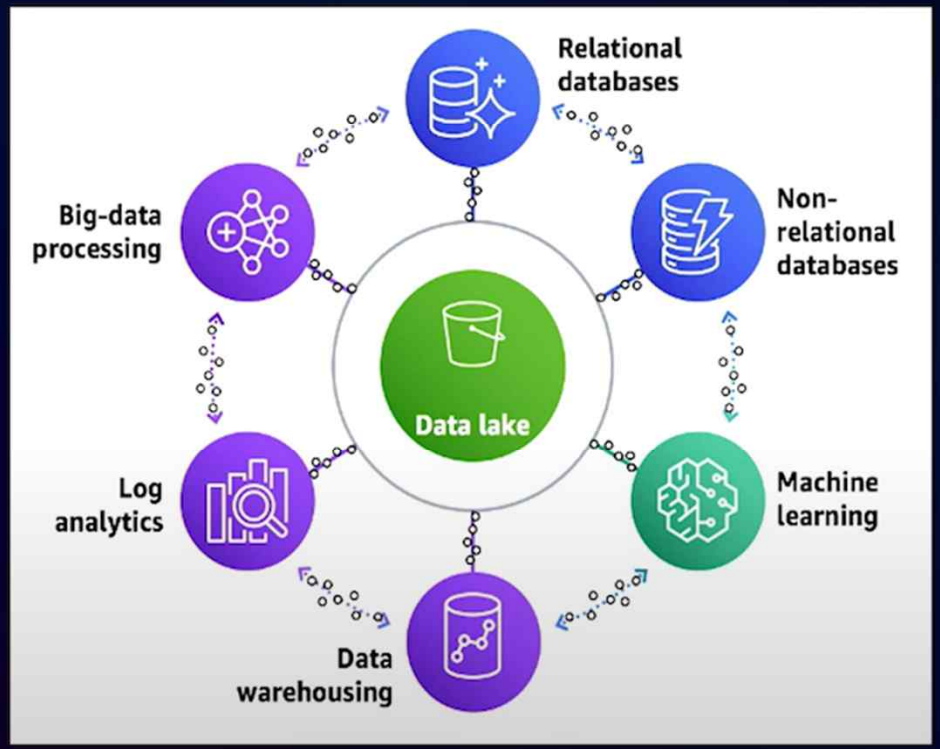
처리 및 분석

데이터 레이크 하우스는?

" 데이터 웨어하우스와 데이터 레이크의 장점을 합쳐 효율성을 높인 개념 "



AWS에서의 데이터 레이크 하우스



데이터 웨어하우스와 데이터 레이크의 통합을 넘어
모든 목적의 데이터를 수집하고 저장할 뿐 아니라
가공, 처리, 분석하고 보여주는 플랫폼을 의미함

[AWS의 데이터 레이크 필요 요소]

- 확장 가능한 데이터 레이크
- 목적에 맞게 구축된 데이터 서비스
- 원활한 데이터 이동
- 통합된 거버넌스
- 성능 및 비용 효율성

금융에서의 데이터 레이크 관련 고려사항

규정 준수

- 데이터 레이크를 구축할 때 고려해야 할 규제 및 데이터 개인 정보 보호
- 규정 준수 의무를 이행하려면 어떤 서비스를 사용해야 하는지?

다중 계정 지원

- 다중 계정 전략을 지원하도록 데이터 레이크를 확장할 수 있는지?
- 각 계정의 데이터 관리자가 데이터 레이크에 저장된 데이터와 메타데이터에 대한 권한을 가질 수 있는지?

인증

- 사용자가 기업 인증 표준을 준수하여 데이터 레이크에 액세스 하도록 되어있는지?
- 데이터 레이크 사용자 인터페이스가 회사의 표준 인증 시스템과 연동될 수 있는지?

권한 부여

- 데이터 레이크의 데이터 및 메타데이터에 액세스하기 위한 역할 기반 인증을 지원하는지?
- 사용자에게 허용된 데이터에만 액세스 할 수 있도록 되어 있는지?

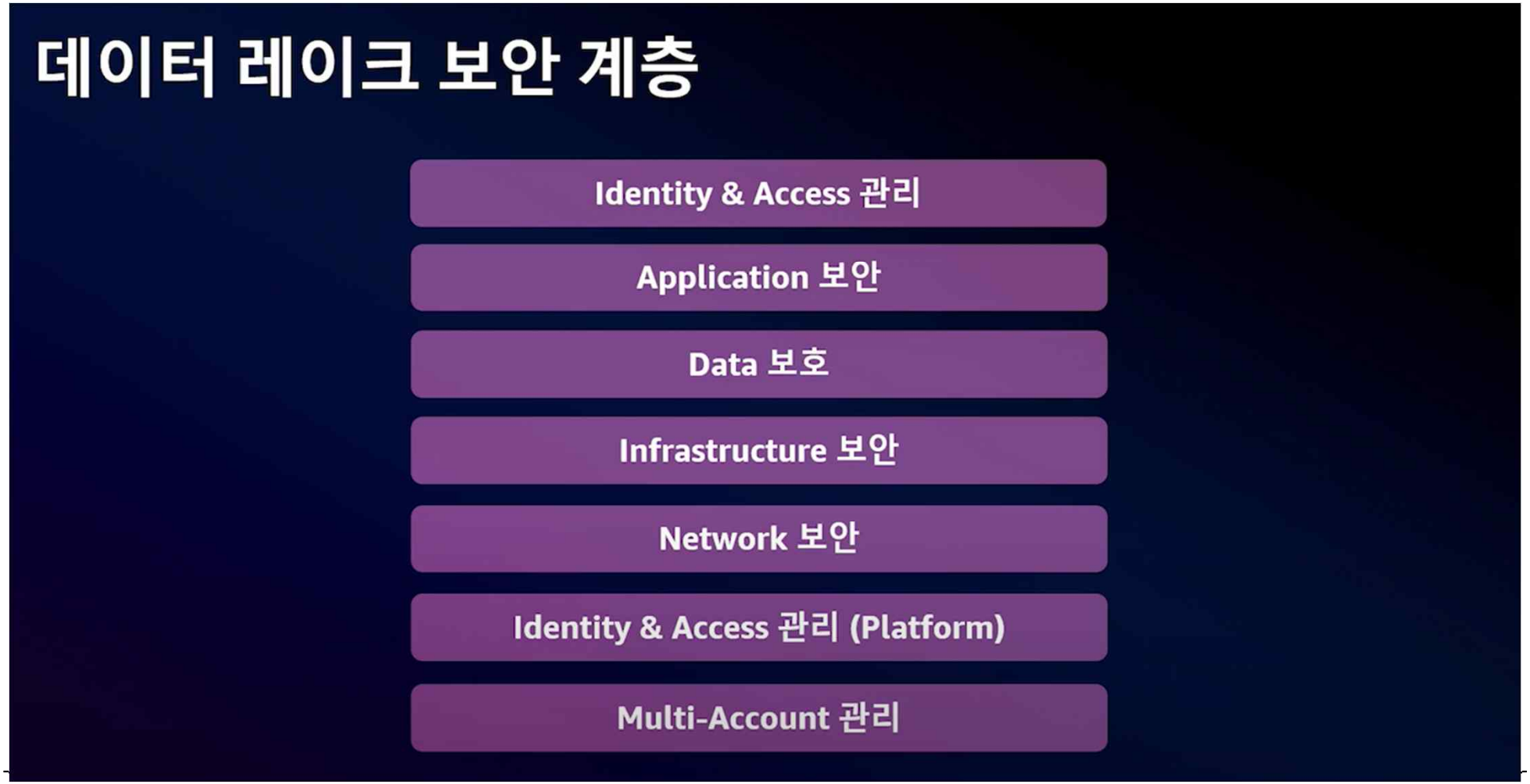
암호화

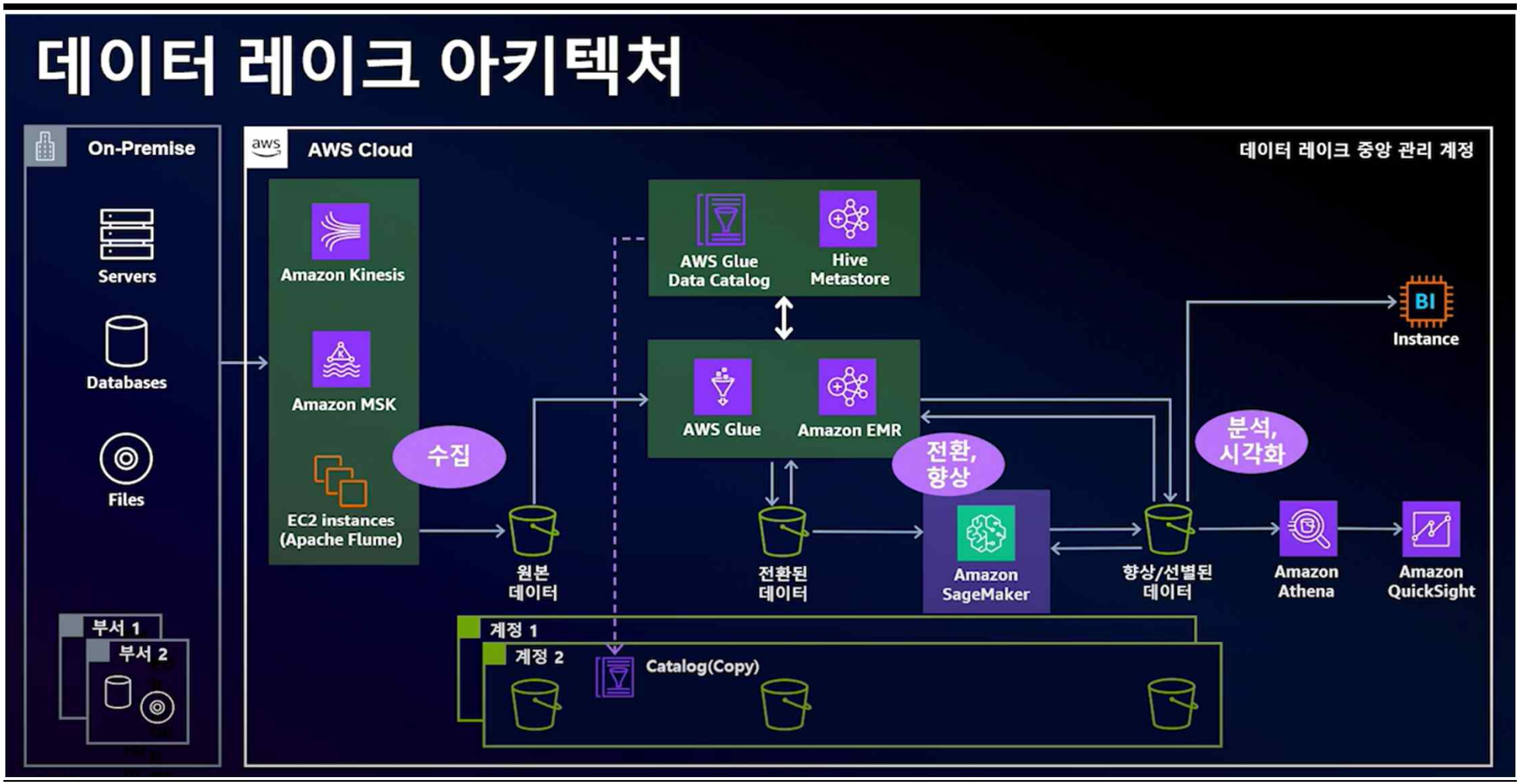
- 데이터 레이크 아키텍처는 저장된 데이터와 전송되는 데이터에 대한 기업 암호화 표준을 준수하는지?
- 데이터 레이크가 당사의 키 관리 서비스와 연동되는지?

사설 네트워크 연결

- 데이터 레이크에서 들어오고 나가는 모든 트래픽이 사설 보안 네트워크를 통해 전송되는지?
- 데이터 레이크에 대한 모든 인/아웃 인터넷 액세스를 차단할 수 있는지?

데이터 레이크 보안 계층


Identity & Access 관리
Application 보안
Data 보호
Infrastructure 보안
Network 보안
Identity & Access 관리 (Platform)
Multi-Account 관리


데이터 레이크 아키텍처
On-Premise
Servers
Databases
Files
부서 1
부서 2
AWS Cloud
데이터 레이크 중앙 관리 계정
Amazon Kinesis
Amazon MSK
EC2 instances
(Apache Flume)
수집
AWS Glue
Data Catalog
Hive
Metastore
AWS Glue
Amazon EMR
전환,
향상
원본
데이터
전환된
데이터
Amazon
SageMaker
향상/선별된
데이터
분석,
시각화
BI
Instance
Amazon
Athena
Amazon
QuickSight
계정 1
계정 2
Catalog(Copy)

데이터 레이크 아키텍처 : 수집

Amazon Kinesis는 서버리스 서비스로
예측이 어려운 이벤트성 수집에 주로 사용

Amazon MSK는 서버 기반으로 구성하여
지속적인 수집 영역에 주로 사용

Flume은 온프레미스에서부터 사용해 왔으며
레거시 환경에서의 수집에 주로 사용

데이터 레이크 아키텍처 : 전환, 향상

AWS Glue

Amazon EMR

Amazon SageMaker

AWS Glue
Data Catalog

Hive Metastore

전환, 향상에는 업무 성격 등의 요인에 따라 사용
AWS Glue는 특수 요건 업무 또는 ETL 업무에 사용
Amazon EMR은 지속적인 성능이 요구되는 영역에 사용

ML 분석이 필요한 경우 Amazon SageMaker에서
추가적인 분석을 수행하여 데이터를 향상시킴

AWS Glue를 Hive Metastore의 Data Catalog로
설정하여 메타데이터 관리에 사용

데이터 레이크 아키텍처 : 분석, 시각화

Amazon Athena

Amazon QuickSight

AWS 기반 시스템 영역의 BI 분석, 시각화

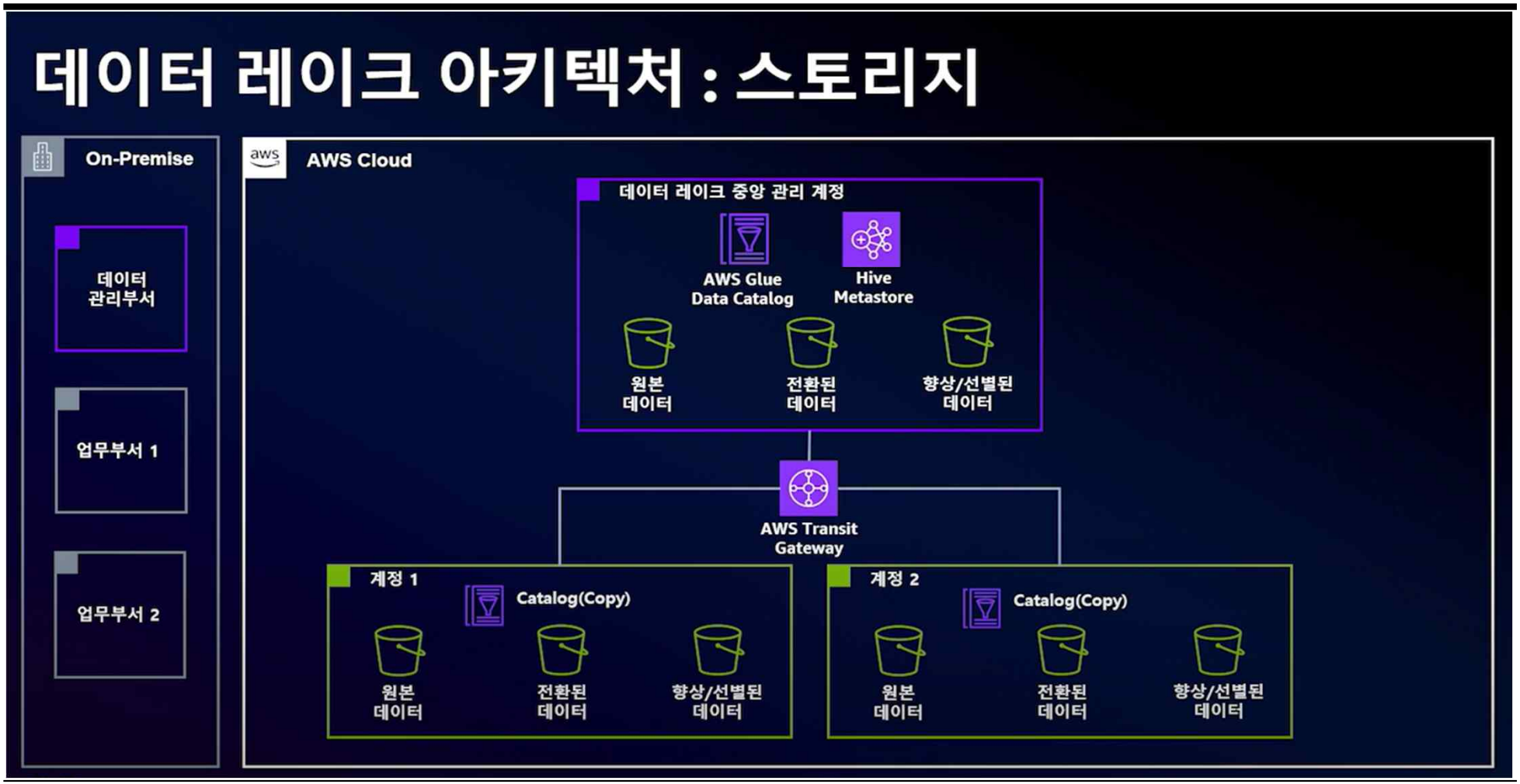
- 시각화는 Amazon QuickSight를 사용하며 여기에 필요한 질의는 Amazon Athena를 연계하여 사용함
- AWS 기반 시스템 영역의 BI 분석 및 시각화에 사용
- 특히 ML 분석에 사용되기 때문에 데이터 사이언티스트 등 데이터 전문가들이 주로 사용하고 있음

Instance

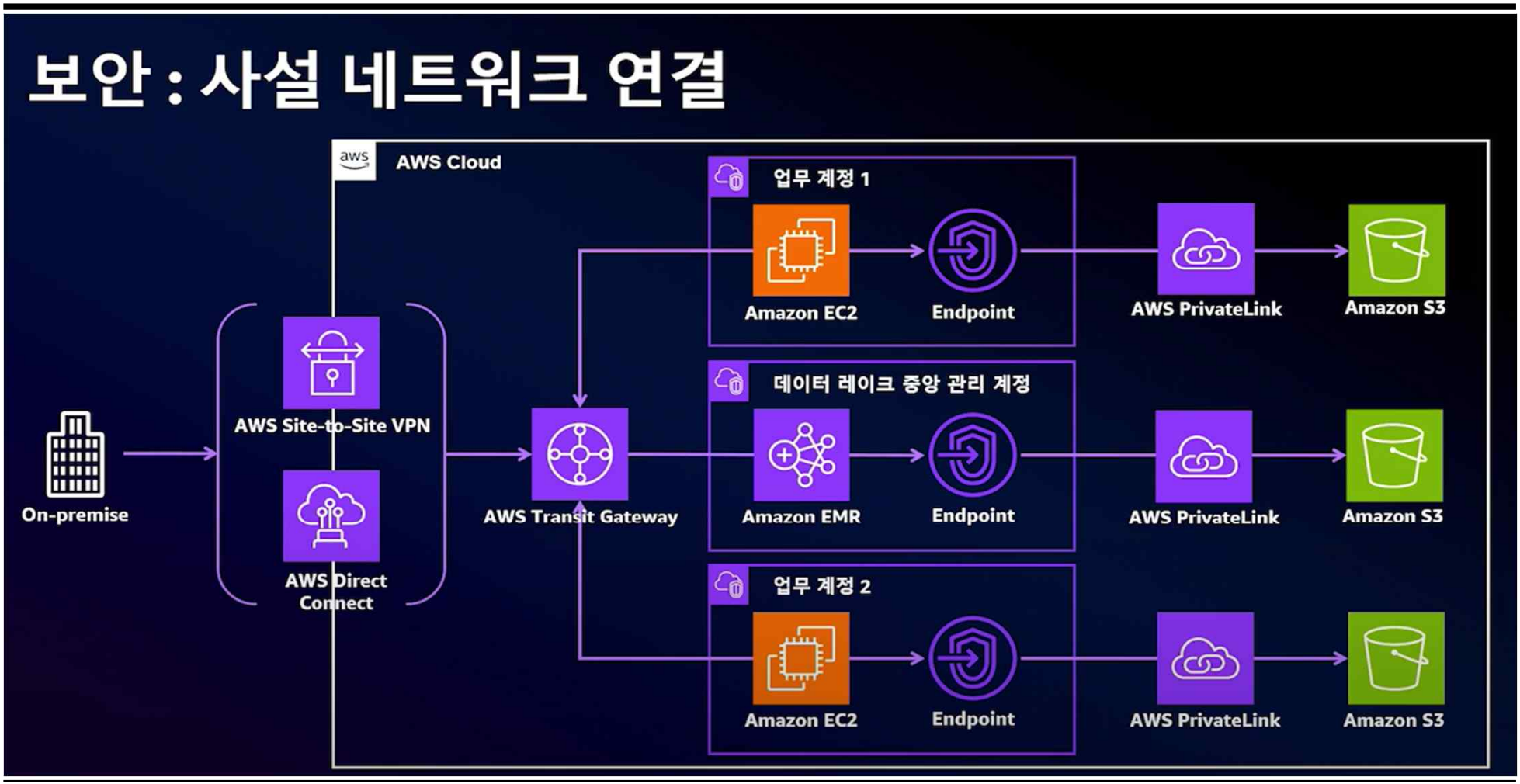
하이브리드 영역의 BI 분석, 시각화

- Amazon EC2에 기존 환경에 사용하는 BI SW를 설치하여 BI 분석 서버로 사용
- 온프레미스 시스템과 연계된 AWS 시스템을 같이 연동하는 하이브리드 시스템 분석 업무에 주요 사용
- 해당 업무에 익숙한 각 영역별 업무 담당자들이 사용

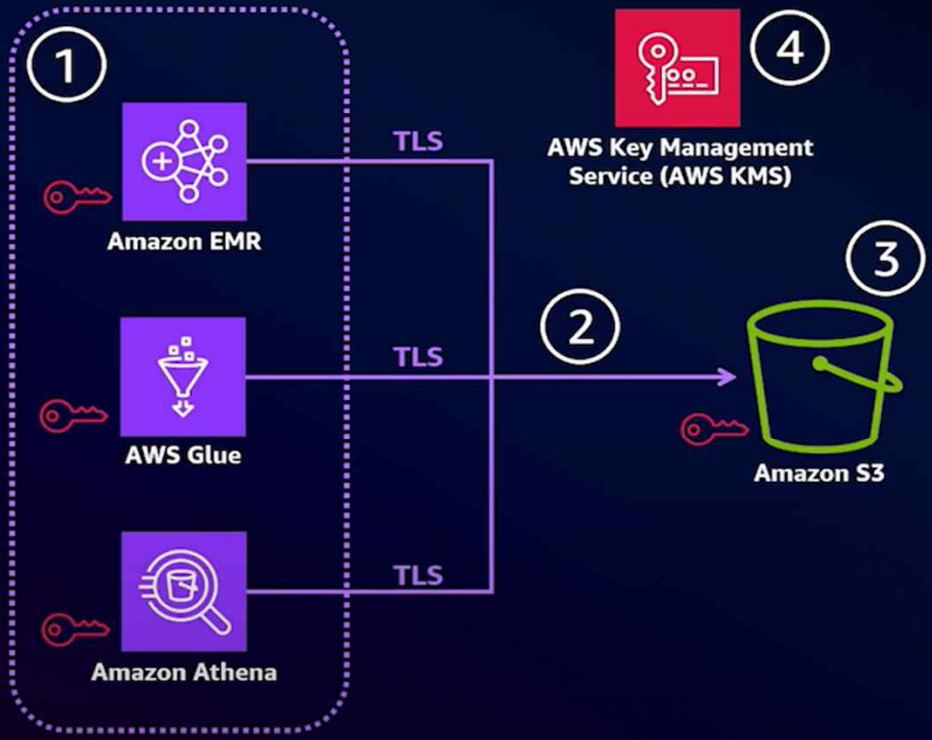
데이터 레이크 아키텍처 : 스토리지



보안 : 사설 네트워크 연결


AWS Cloud
업무 계정 1
Amazon EC2
Endpoint
AWS PrivateLink
Amazon S3
데이터 레이크 중앙 관리 계정
Amazon EMR
Endpoint
AWS PrivateLink
Amazon S3
업무 계정 2
Amazon EC2
Endpoint
AWS PrivateLink
Amazon S3
On-premise
AWS Site-to-Site VPN
AWS Direct Connect
AWS Transit Gateway

보안 : 암호화



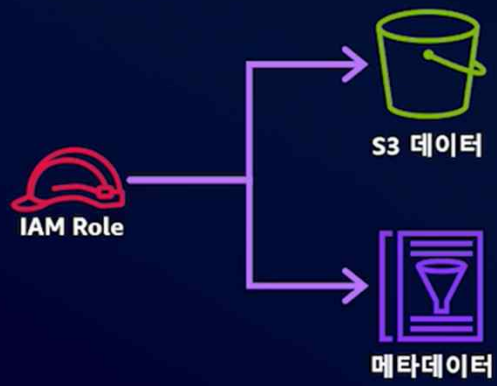
① 클라이언트 사이드 암호화
S3에 저장하기 전에 데이터를 암호화

② 전송 중 암호화
S3로의 모든 통신은 TLS 암호화 사용

③ 서버 사이드 암호화
S3에 쓰여진 데이터에 대한 암호화

④ 키 관리
암호화 키는 AWS KMS 서비스로 관리

보안 : 권한 부여 및 감사



권한 부여

S3 데이터, 메타데이터 등 데이터에 대한 모든 접근은
IAM 역할을 통하여 권한을 부여 받도록 함

AWS CloudTrail

- IAM Role 권한 허용 감사 로그
- S3 데이터 이벤트 감사 로그 - get, put, delete Object
- 메타데이터에 대한 접근 및 작업내용 감사 로그
- AWS 서비스들의 모든 API 호출 감사 로그
  (* 콘솔을 이용한 호출 포함)

Amazon EMR

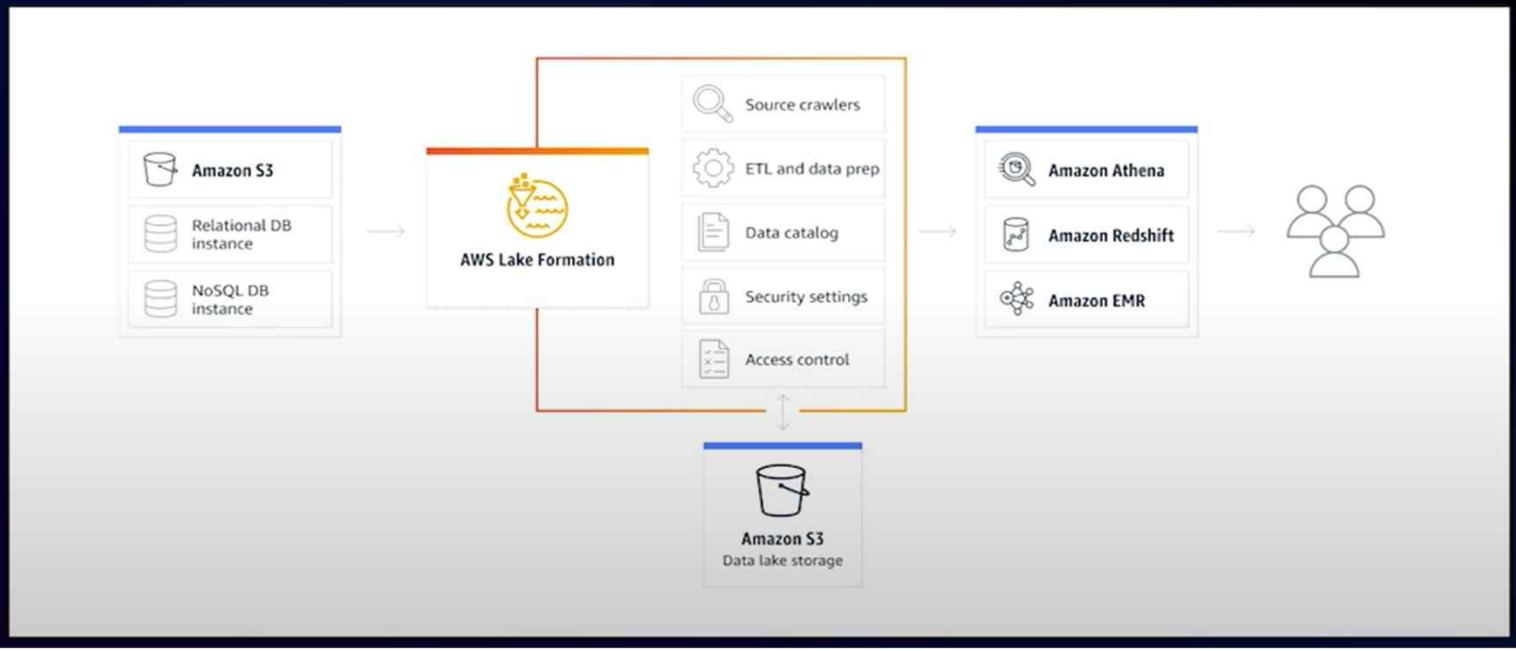
- 어플리케이션 레벨에서의 쿼리 로그 설정
- Amazon EMR의 엔진 별로 세부 설정
- EMRFS 요청 관련 감사 로그는 AWS CloudTrail과 연동하여 로깅

감사

데이터 접근, 인증, 요청 내용을 포함하여
쿼리 내용까지 감사 로그로 남기도록 함

AWS Lake Formation

"AWS Lake Formation은 복잡하고 시간이 많이 걸리는 작업들을 자동화하여 데이터 레이크 생성, 보안 및 관리 프로세스를 편리하게 구성할 수 있는 완전관리형 데이터 레이크 서비스 "



AWS Lake Formation 장점

비교 항목	통상적인 AWS 데이터 레이크로의 구성	AWS Lake Formation를 이용한 구성
구성 시간	수집, 정제, 인덱싱, 보안 등 복잡한 여러 구성 단계를 거치기 때문에 수개월이 소요될 수 있음	AWS Lake Formation 에서 제공하는 Blueprint를 사용하면 워크플로우를 빠르게 구성할 수 있음
보안	각 AWS 리소스 별 보안 설정과 계층 별 암호화 설정을 개별적으로 구성	보안을 하나의 중앙 데이터 카탈로그에서 구성할 수 있으며 테이블과 컬럼 단계까지도 사용자별 보안 설정이 가능
오케스트레이션	스케줄, 오케스트레이션 관리를 위해 Airflow 등 다른 서비스 또는 도구의 사용이 필요	AWS Lake Formation 자체의 워크플로우, 트리거 기능을 사용할 수 있음
중복성	데이터 세트에 대한 각 사용자별 권한 설정 및 관리의 복잡도는 여러 중복된 데이터 세트 복사를 야기할 수 있음	기본적으로 하나의 데이터 세트를 각기 다른 사용자들의 권한에 맞추어 제공함으로 데이터 중복 관련 효율성이 높음

The End