AWS Glue
서버리스 데이터 통합 서비스

Jung-jun Park
Solutions Architect
AWS

Trend: 데이터는..

기하급수적으로 성장

새로운 소스들에서

점점다양해지는

다양한 기술을 가진
사용자들이 이용

다양한
애플리케이션에서 접근

aws

Trend: 더 까다로운 워크로드

Batch - 일괄

Data warehousing

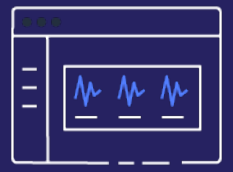
Business intelligence

Reporting

새벽 or 시간별 ETL

지연 시간에 구애받지 않는 장기 실행작업

Real-time - 실시간



Responsive dashboards

Monitor and alert

Streaming apps

지속적인 운영

Latency-sensitive micro-batch jobs

온프레미스 ETL 솔루션은 현재의 워크로드를 처리하기 어려움

확장이 어려운 인프라

온프레미스 ETL도구는
설치, 관리 및 확장이 복잡

높은 비용

중앙 집중식 데이터 카탈로그,
스트리밍 데이터 처리와 같은 고급
기능은 별도의 라이센스 비용이 발생

락인(Lock-in)

기존 ETL작업은
독점적이며 포팅할 수 없음

Modern Data Architecture with AWS Glue

DATA INTEGRATION IN BATCH AND REALTIME

PERFORMANT AND COST-EFFECTIVE

CENTRALIZED CATALOG AND GOVERNANCE

TOOLS FOR DIVERSE SKILLSETS

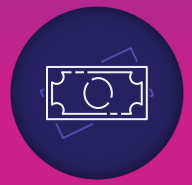
AWS Glue
서버리스
데이터 통합 서비스

서버리스

유지 관리할 인프라가 없고, 필요한 만큼 컴퓨팅 성능을 할당하고 작업을 실행

모든 사용자를 위한 데이터 통합

다양한 스킬렛에 맞는 개발 환경 – 데이터 엔지니어를 위한 시각적 ETL,
데이터 과학자를 위한 노트북 스타일 개발, 데이터 분석가를 위한 코드없이 개발



비용 효율적

다른 클라우드 데이터 통합 옵션보다 최대 55% 저렴

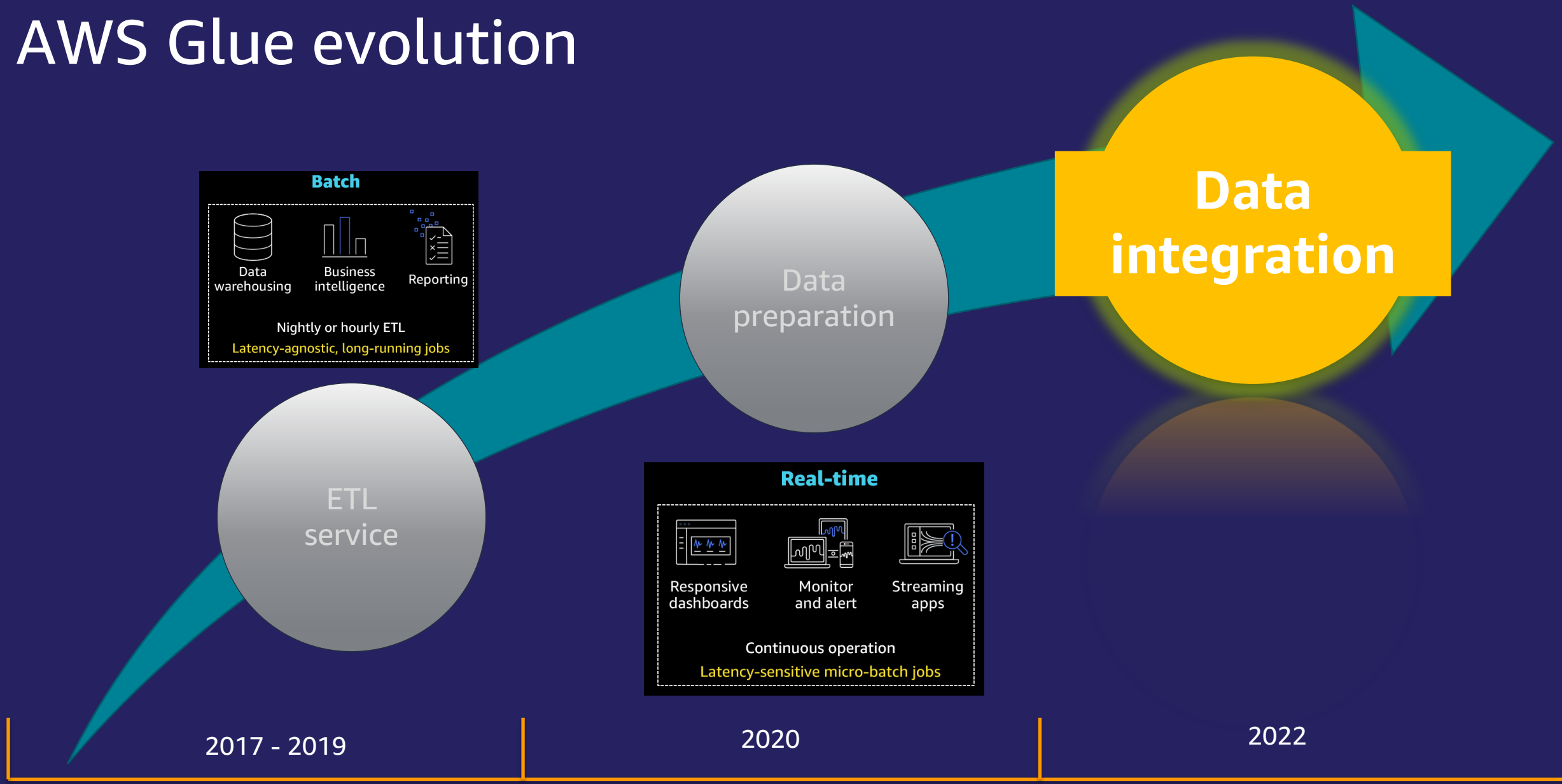
복잡한 워크로드 처리

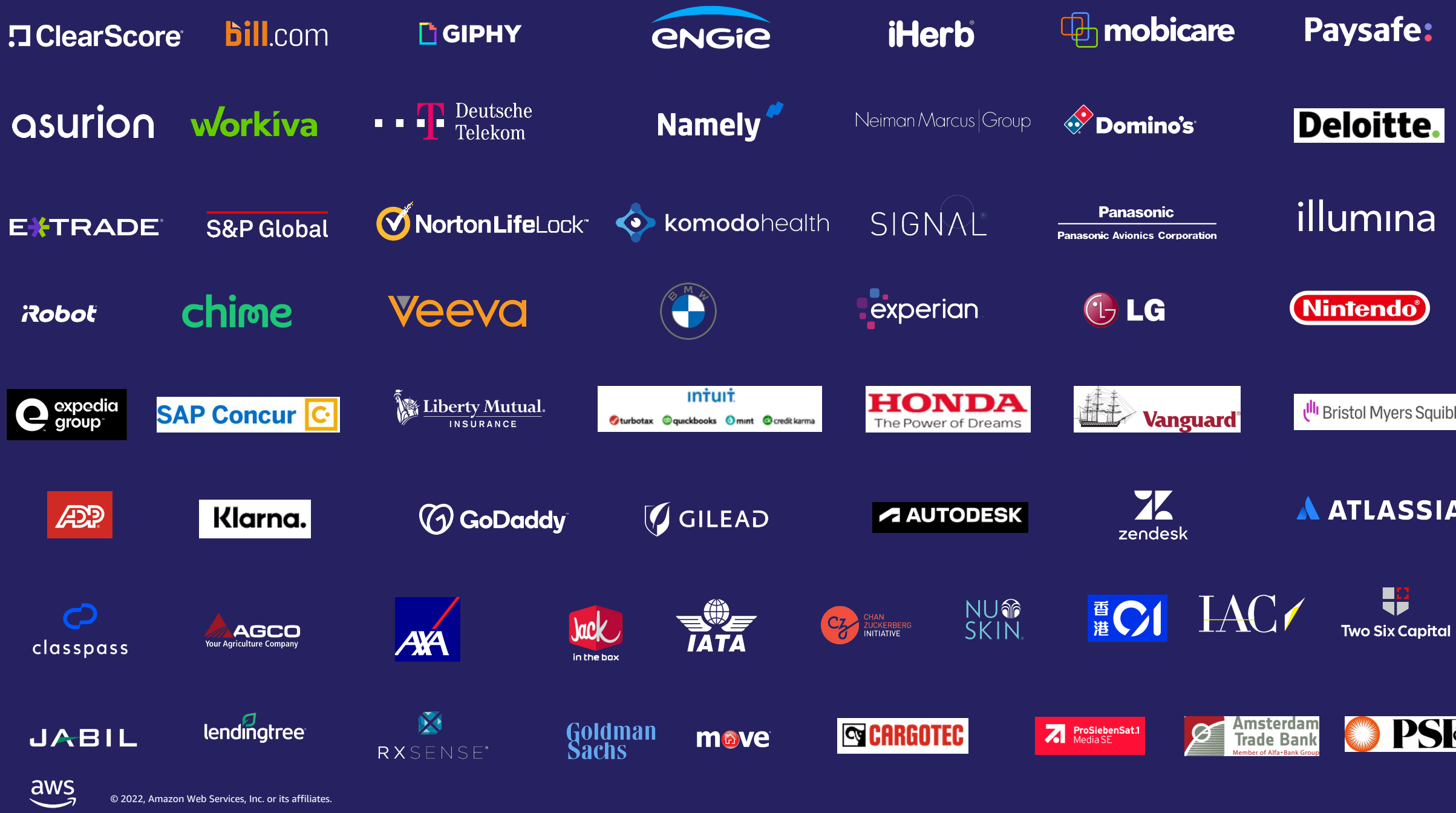
30+ 개 이상의 데이터 소스에 연결하고, 페타바이트 규모의 데이터를 일괄 및 실시간, 이벤트 기반으로 처리할 수 있음.

락인 없음

오픈소스 SparkSQL, PySpark 및 Scalar에서 데이터 통합 파이프라인 개발

AWS Glue evolution





고객들의 AWS Glue 사용 사례

확장 가능한 데이터 분석을 위해
데이터 레이크와 레이크하우스 구축

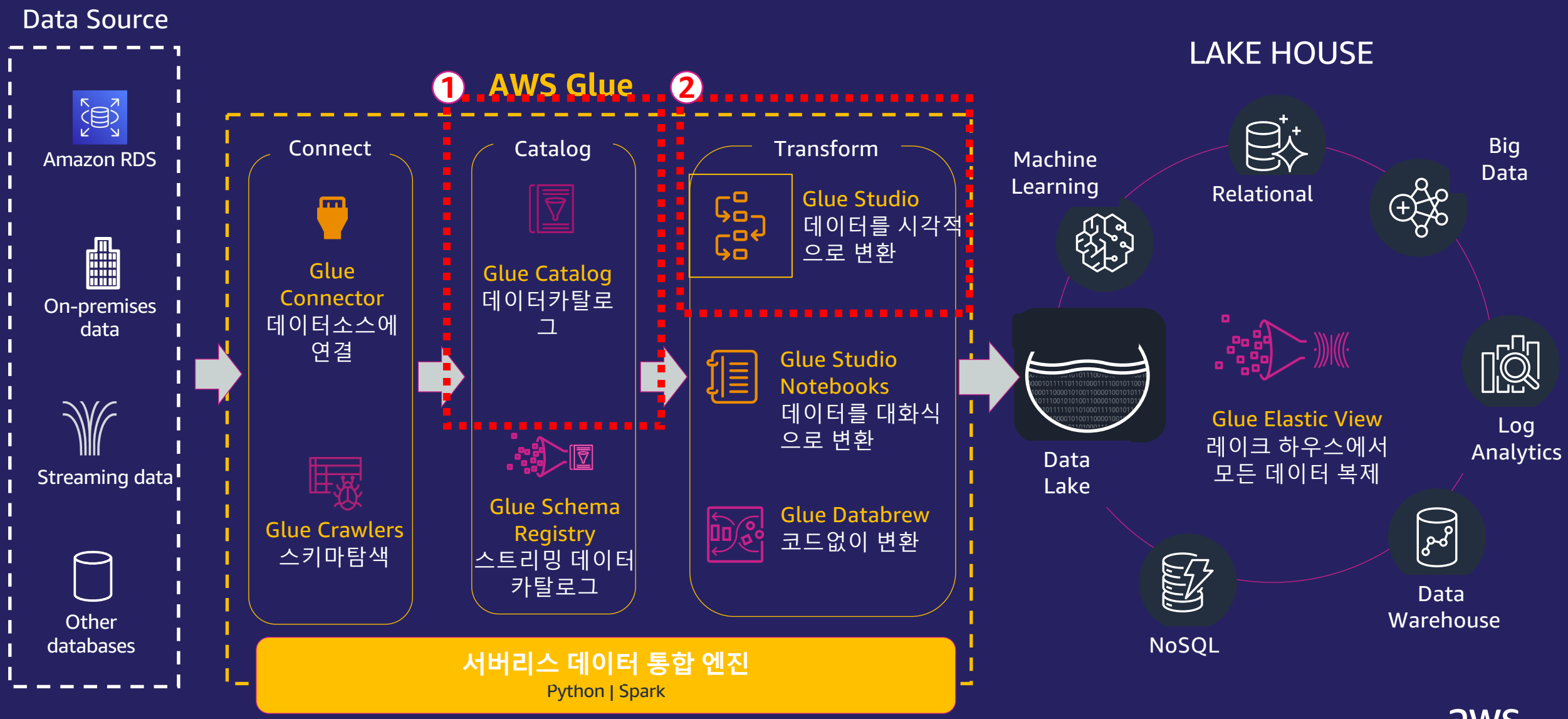
고가의 기존 ETL솔루션에서 마이그레이션하여 유연성 확보 및 비용절감

AWS 분석 서비스들에서 사용할 수 있도록
데이터 자산들 (Assets)의 카탈로그 작성

Apache Spark를 사용하여 페타바이트 규모의 데이터를 배치 처리하거나 실시간으로 처리

머신러닝을 위한 데이터 준비

AWS Glue를 사용해 통합데이터 생태계를 더 빠르게 구축

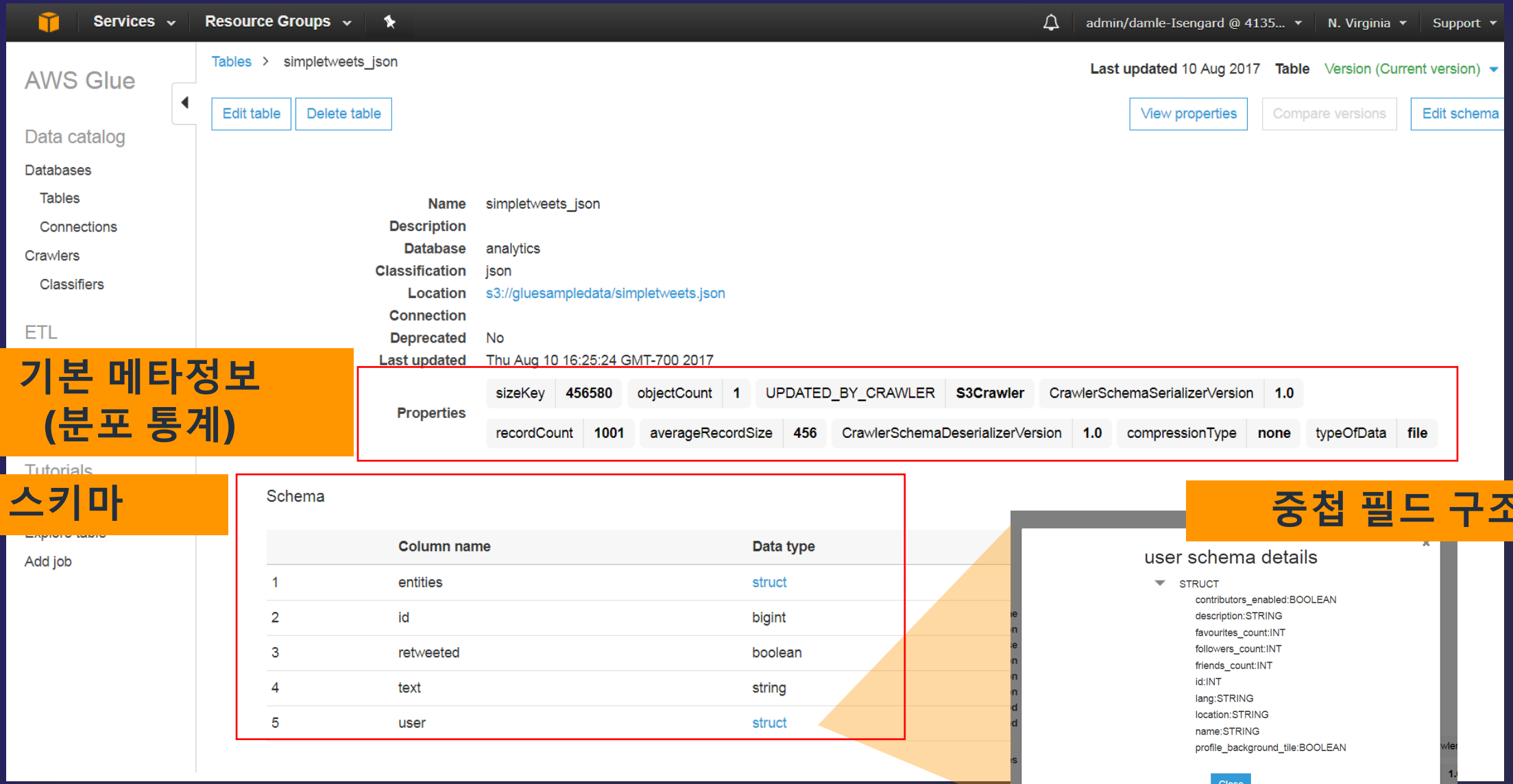


Glue Data Catalog

Catalog ??

- 자동차 디자인,기능, 바퀴 등 → 자동차 스키마
- 가격, 기능, 색상, 프로모션 등 → 자동차 메타정보

AWS Glue – Data Catalog

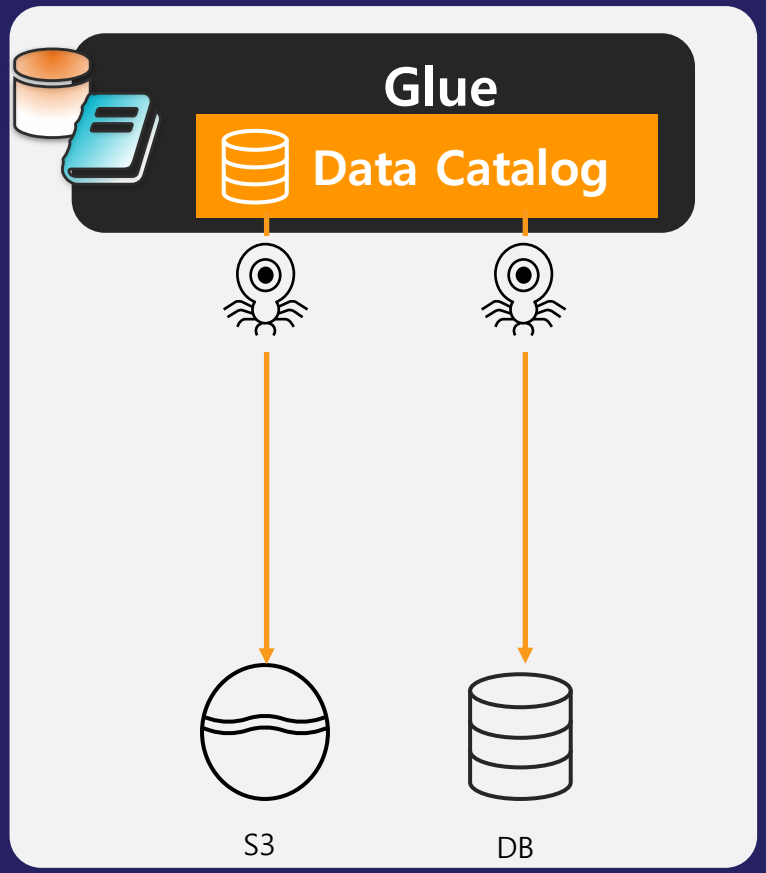


기본 메타정보 (분포 통계)

스키마

중첩 필드 구조

AWS Glue – Data Catalog

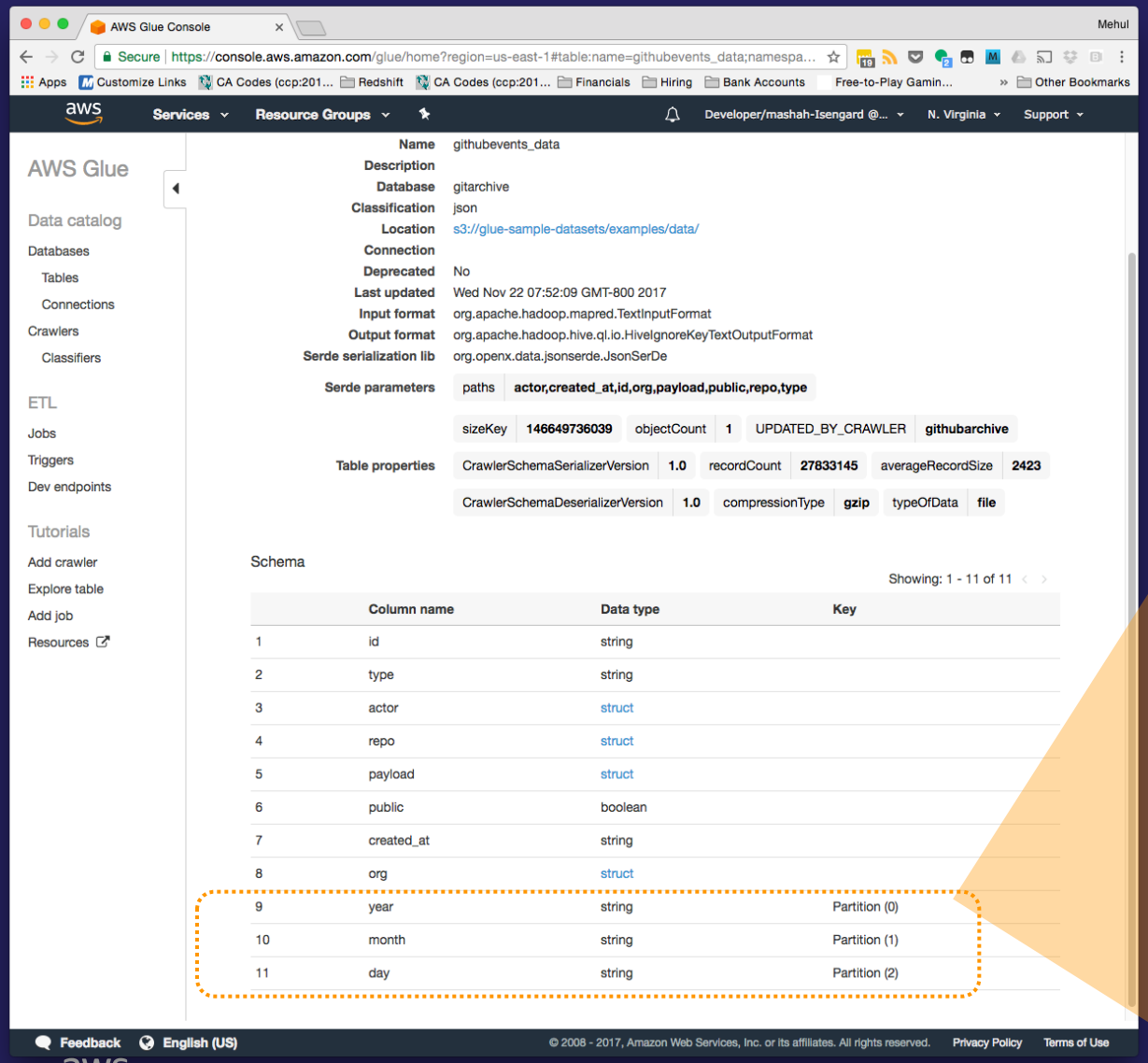


S3, Database로 부터

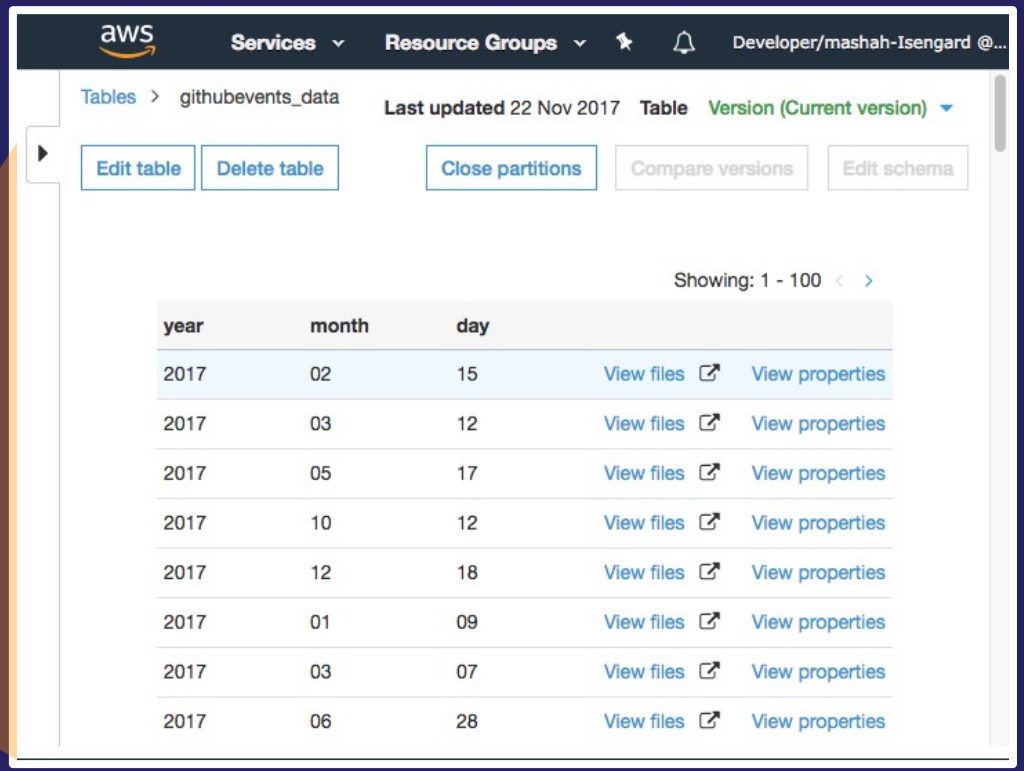
크롤러가 자동적으로 데이터 스키마를 찾아서

테이블 스키마 정보와 메타 정보를 저장

AWS Glue – Data Catalog



자동적으로 파티션 구조 파악



AWS Glue – Data Catalog

Last updated 21 Aug 2017 Table Version 1

Change	Column name	Data type
	id	bigint
	retweeted	boolean
	text	string
	user	struct

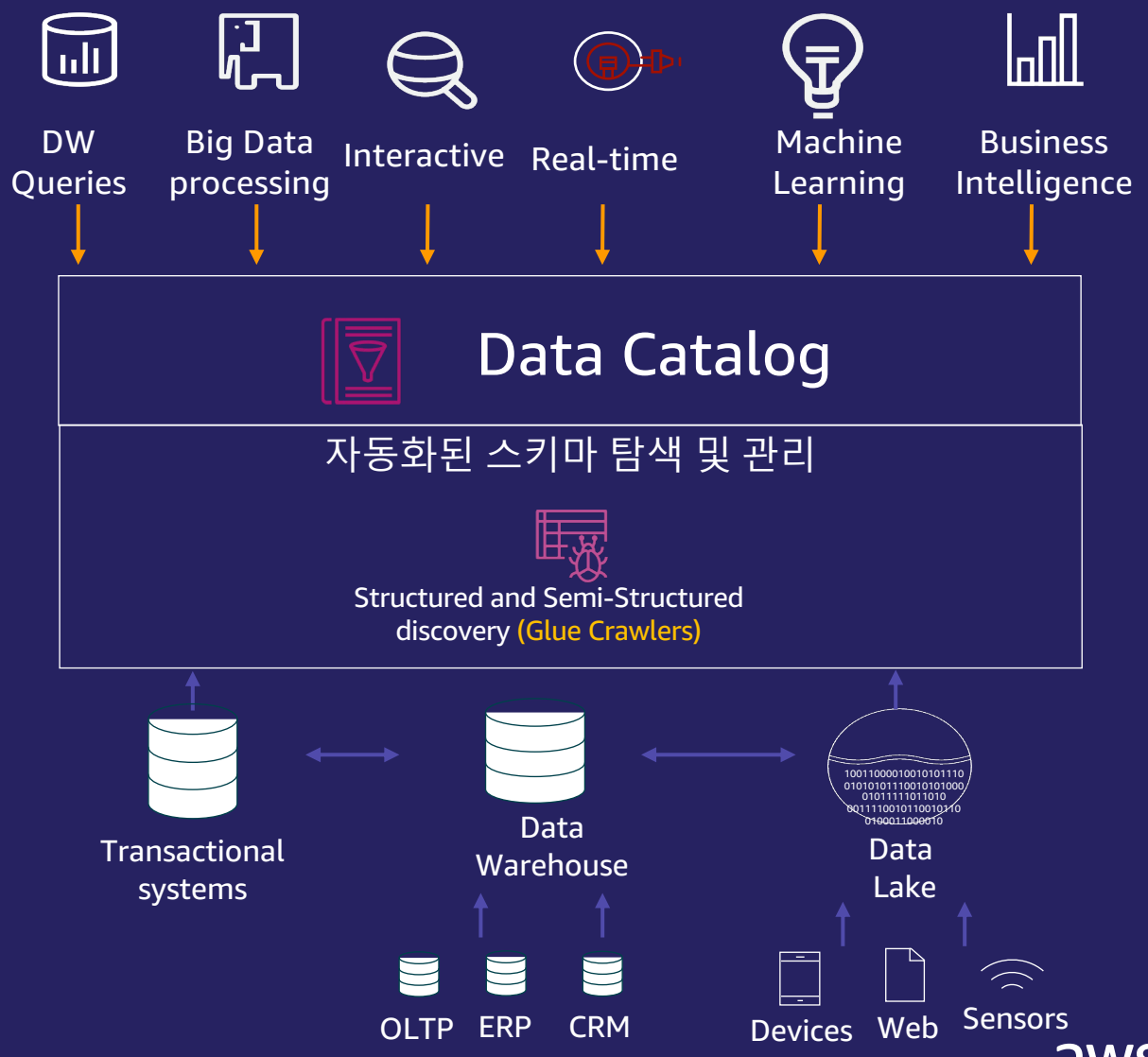
Last updated 25 Nov 2017 Table Version 2

Change	Column name	Data type
	id	bigint
	retweeted	boolean
	text	string
	user	struct
Added	url	string

데이터 구조가 변경되면 자동적으로
업데이트 하고 버전 관리 가능

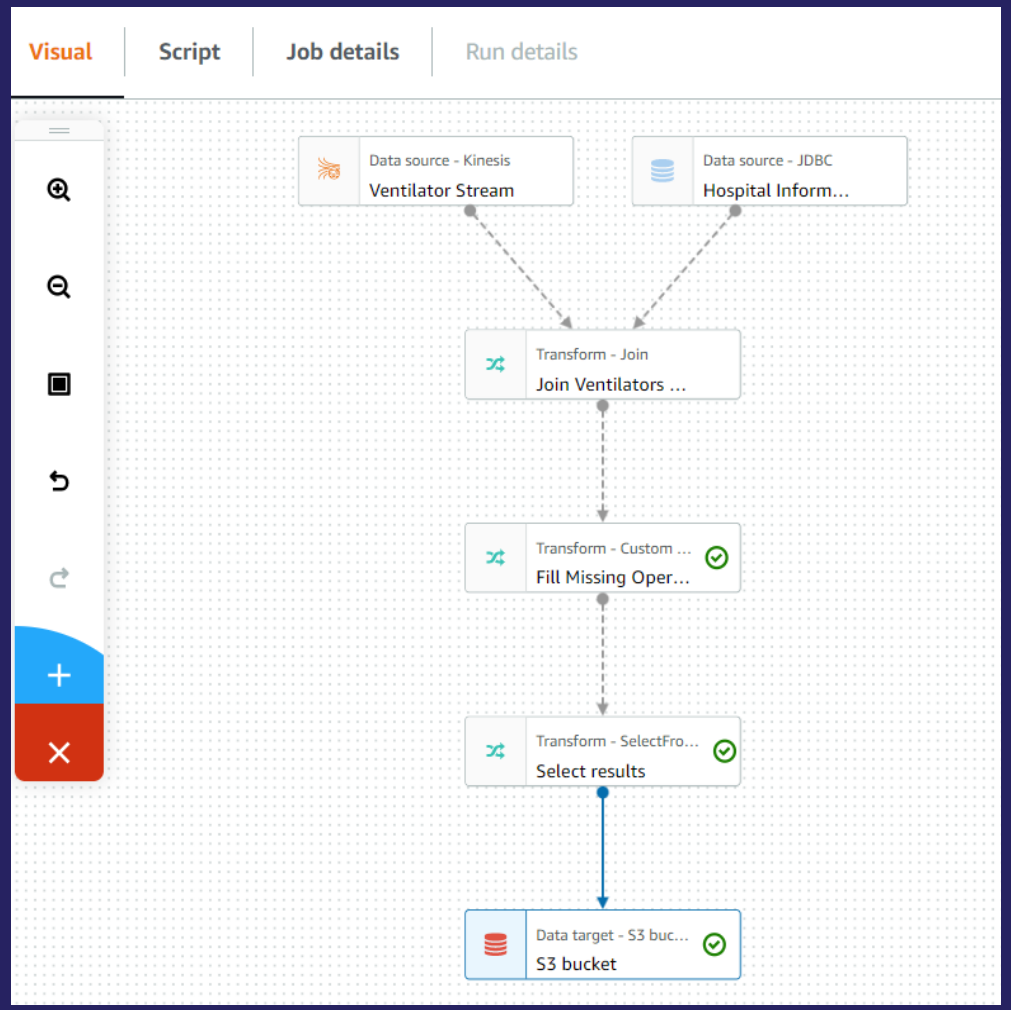
Glue Data Catalog/Crawler – 통합 데이터 카탈로그

- 데이터 이동 없음
  = 낮은비용/관리
- 모든 데이터를 중앙에서 탐색 및 쿼리
  = 생산성
- 정형/반정형 데이터 통합
  = 빠른 인사이트
- 데이터 탐색 자동화
  = 생산성



ETL

AWS Glue Studio: 시각적 ETL 인터페이스



- 코딩없이 AWS Glue 작업을 시각적으로 작성
- 단일 창을 통해 여러개의 작업 모니터링
- 러닝 커브 없는 분산처리
- 코드 조각을 통한 고급 변환

Glue Monitoring: 작업 상태 확인을 위한 모니터링 대시보드

Monitor Job Runs Info

7 Day

Job Runs Summary

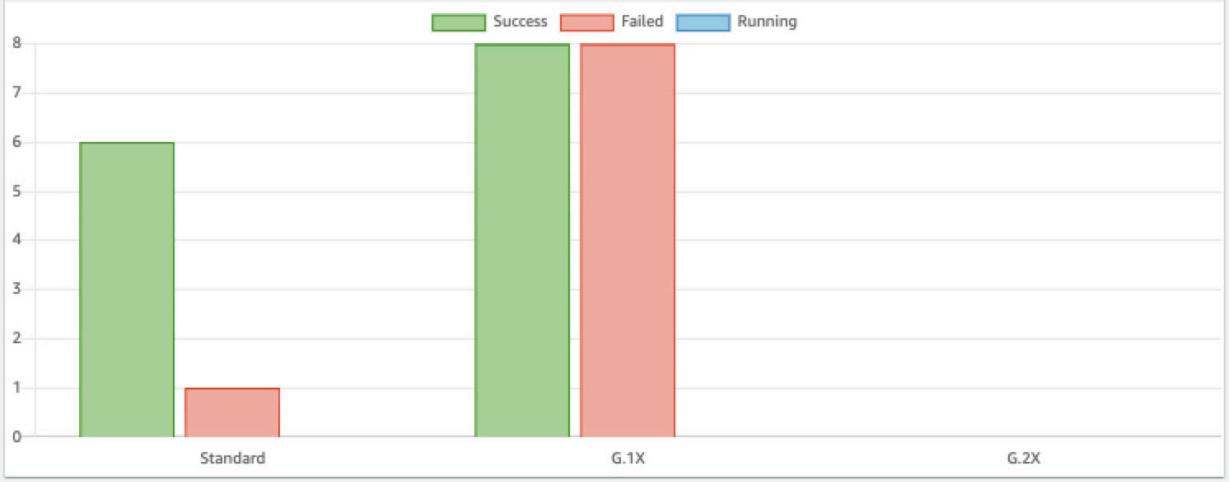
Total Runs	Running	Canceled	Success	Failed
347	23	0	246	78

Job Run Success Rate Info

Success Rate	Status
71%	Service operating normally

Job Type Breakdown Info

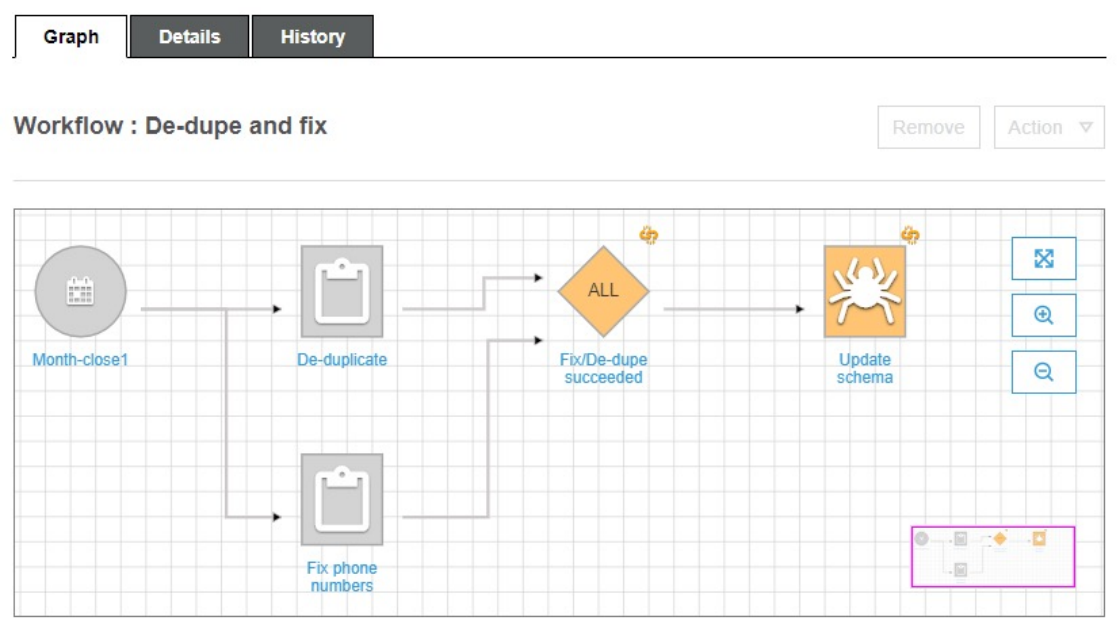
Worker Type Breakdown Info



Job Runs Timeline Info

Estimated Job DPU Usage Pricing Info

AWS Glue workflows: 데이터 작업을 쉽게 오케스트레이션



- Glue 작업 및 기타 AWS 서비스 오케스트레이션
- 이벤트 기반으로 작업/트리거 예약
- 한 곳에서 워크 플로우 실행을 모니터링

왜 AWS Glue 인가?

데이터 통합 현대화

- 확장하기 어려움
- 비싼 라이센스
- 벤더 락인

01
Upgrade On-Prem ETL

7x
온프리미스 옵션 대비 최대 7x 저렴

No lock-in
오픈소스 Spark Python, Scala으로 작성

- 높은 총 소유비용
- 클러스터 조정 및 관리 조력

02
DIY Spark on Cloud

5x
자체 Spark 클러스터를 구축하는 것 보다 최대 5x저렴

4x
자체 Spark 클러스터 보다 유지 관리를 최대 4x감소

- Code lock-in
- 인프라에 대한 추가 비용
- 비싼 라이센스

03
Cloud ETL

55%
다른 클라우드 공급자와 비교해 최대 55%저렴

Serverless
관리할 서버가 없고, 인프라 비용도 포함되지 않음

AWS Glue 기능

- 서버리스 데이터 통합 엔진
- 데이터 탐색
- 데이터 변환
- 데이터 파이프라인 운영

Reference architectures

aws

사용사례: 온프레미스 ETL도구를 모던화 하여 데이터 웨어하우스로 로드

사용사례: 머신러닝을 위한 원시 데이터 준비

Logs, app data

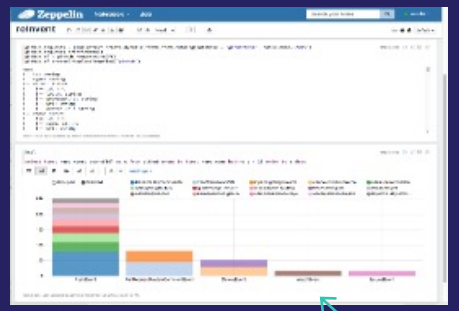
Amazon RDS

Other databases

On-premises data

Streaming data

Notebooks:
데이터 탐색, 실험



AWS Glue
Ingest

Raw data

AWS Glue
Transform

정리 및 보강된 데이터 S3

Glue Catalog

AWS Glue
DataBrew
Transform

추출된 피쳐 S3

Amazon Forecast

Amazon Comprehend

Amazon SageMaker

AWS Glue
DataBrew
Transform

훈련데이터

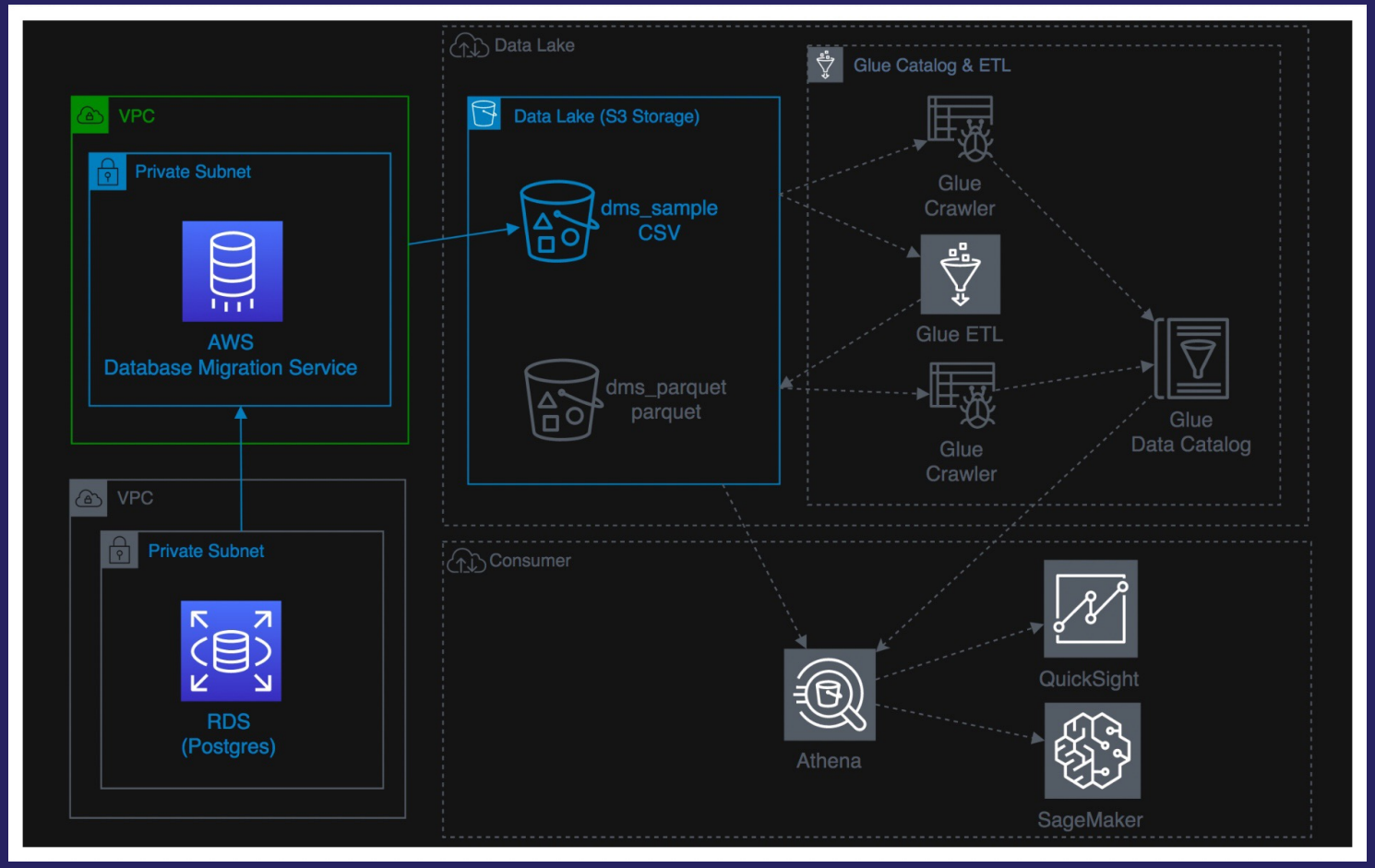
Amazon Rekognition

Amazon Lex

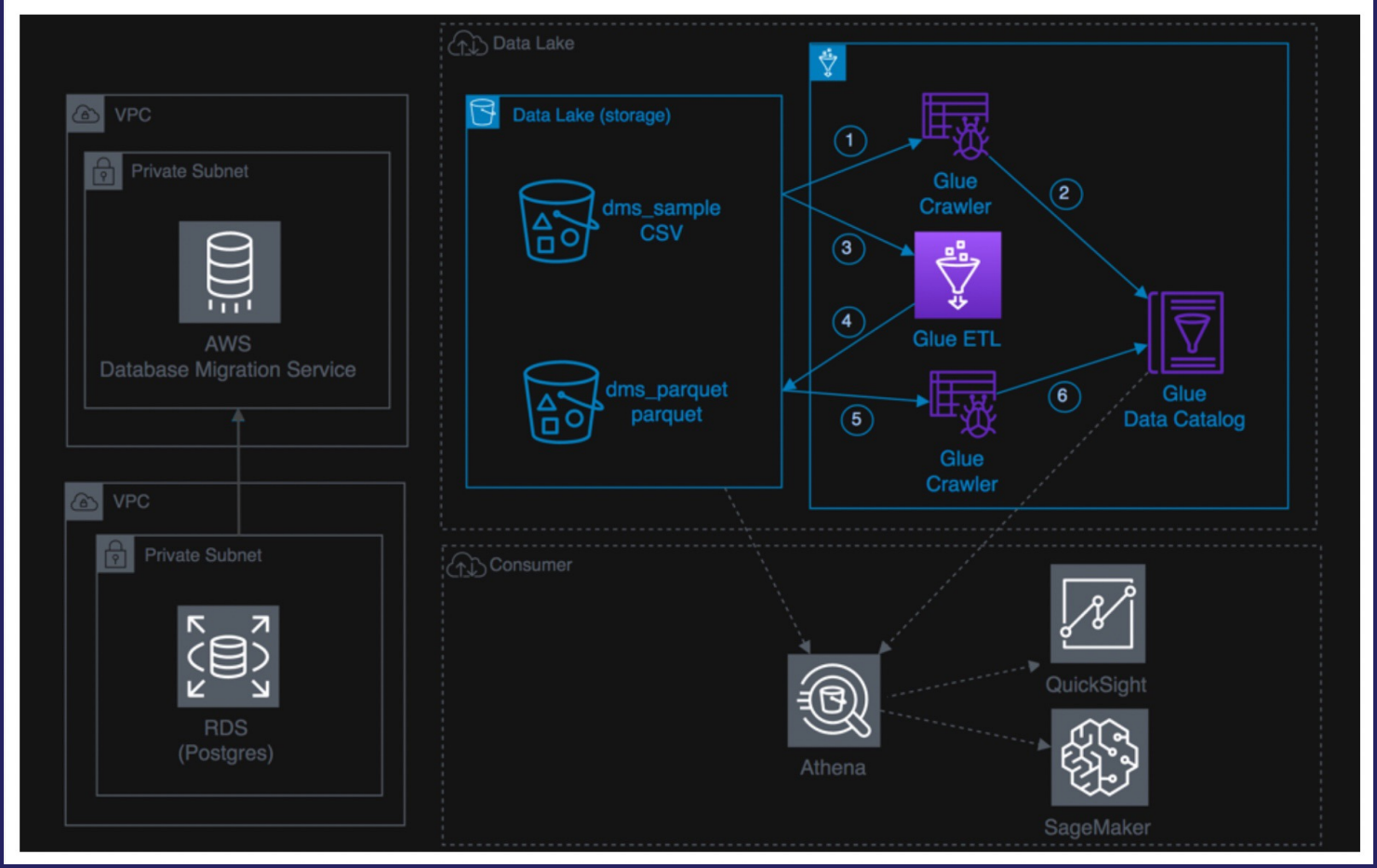
aws

Thank you!

Today's Lab



Today's Lab



[참고] Parquet?

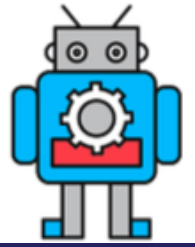
'나무조각을 붙여넣은 마룻바닥'

연봉이 400인 사람?

나이	부서	연봉
21	A	200
31	B	300
41	A	400

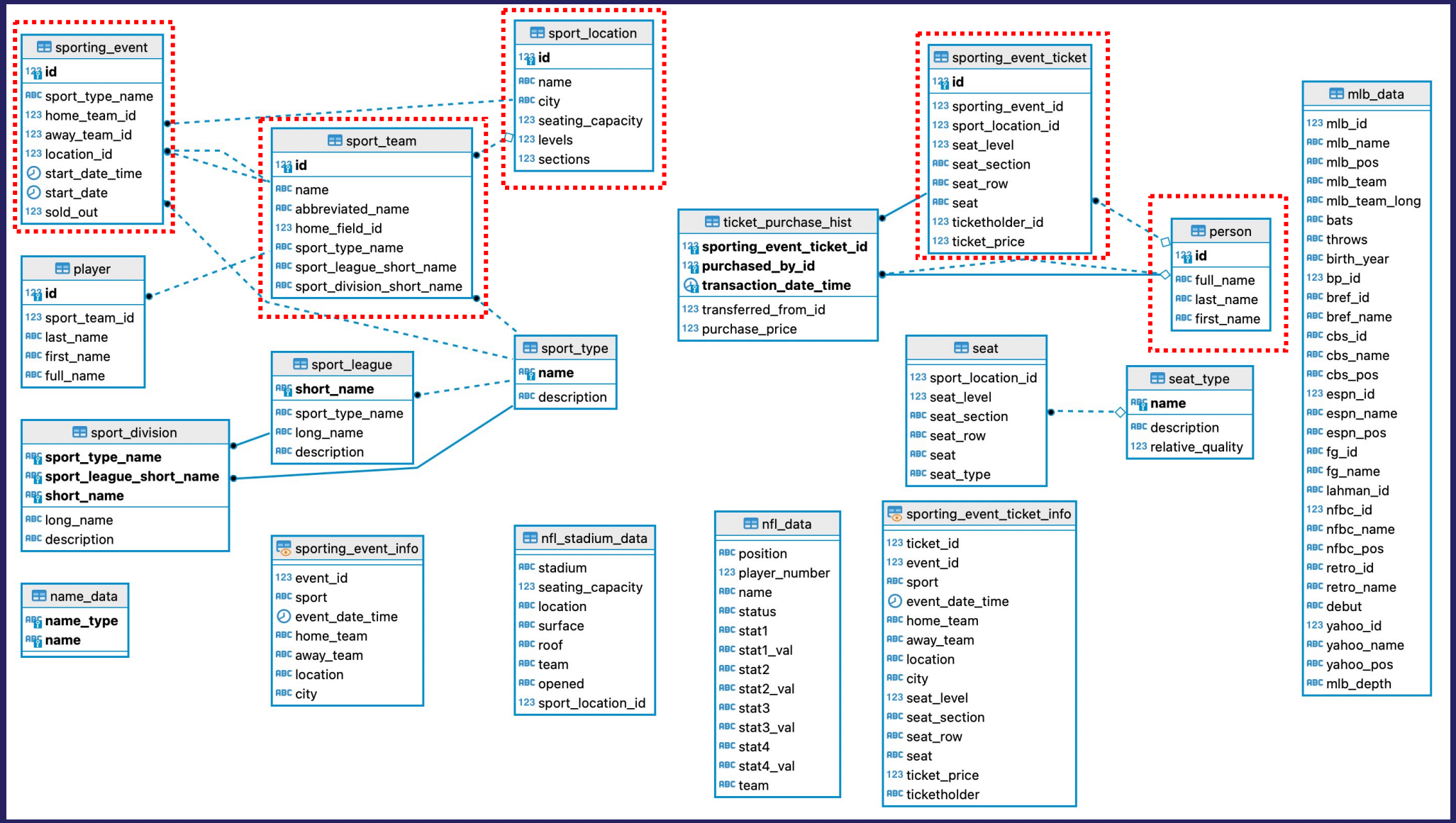
Row-Based →

21	A	200	31	B	300	41	A	400



Column-Based →

21	31	41	A	B	A	200	300	400



파케 테이블 명 명시