



AWS Glue

서비스 데이터 통합 서비스

Jung-jun Park
Solutions Architect
AWS

Trend: 데이터는..



기하급수적으로 성장



새로운 소스들에서



점점다양해지는



다양한 기술을 가진
사용자들이 이용



다양한
애플리케이션에서 접근

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 ETL 도구는
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 ETL 작업은
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



서비스

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

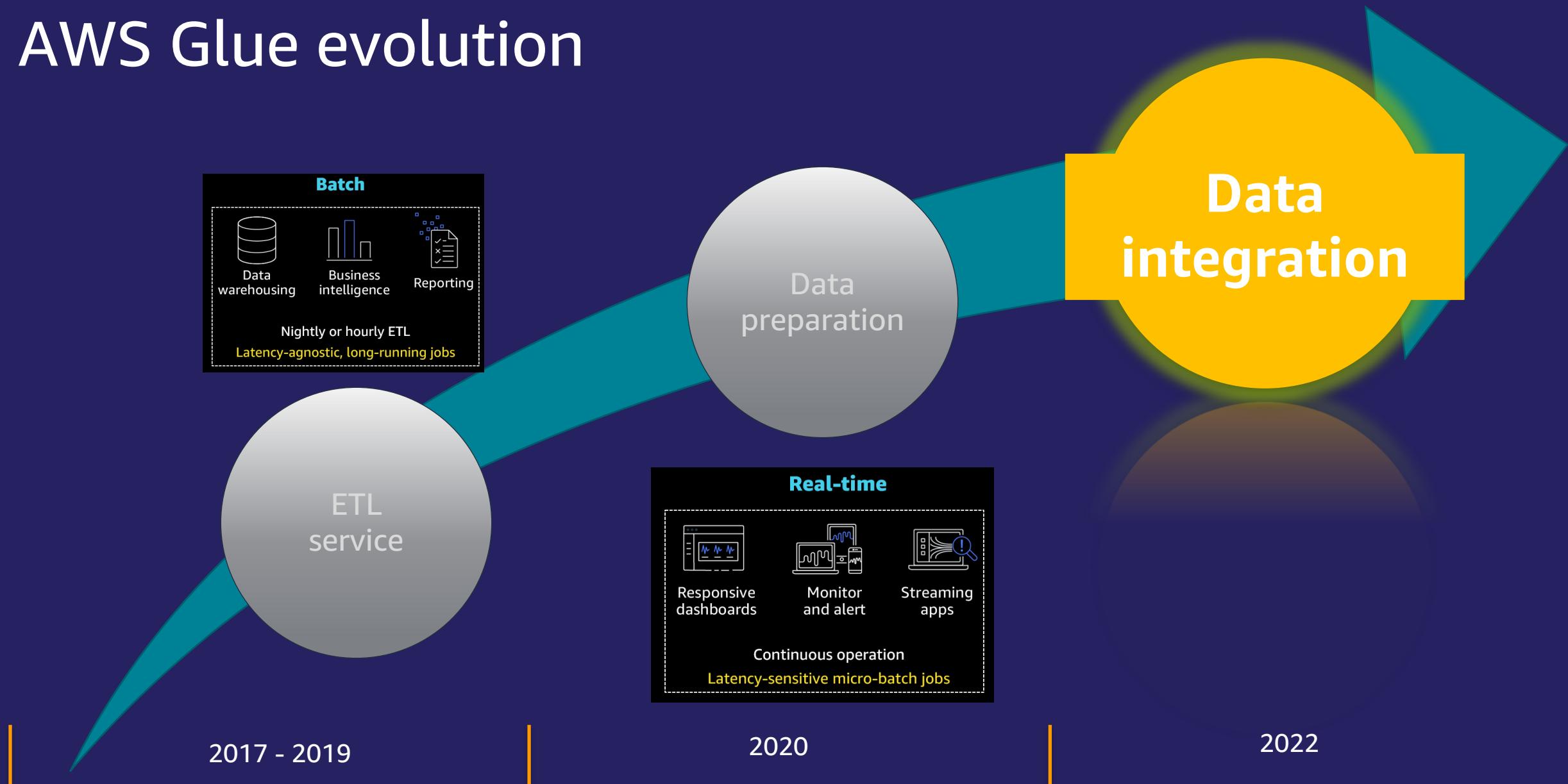
복잡한 워크로드 처리

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





Neiman Marcus Group



© 2022, Amazon Web Services, Inc. or its affiliates.

고객들의 AWS Glue 사용 사례



확장 가능한 데이터 분석을 위해
데이터 레이크와 레이크하우스 구축



고가의 기존 ETL 솔루션에서 마이그레이션하여 유연
성 확보 및 비용 절감



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



Apache Spark를 사용하여 페타바이트 규모의 데이
터를 배치 처리하거나 실시간으로 처리



머신러닝을 위한 데이터 준비

AWS Glue를 사용해 통합데이터 생태계를 더 빠르게 구축

Data Source



Amazon RDS



On-premises data



Streaming data



Other databases

Connect



Glue Connector
데이터소스에
연결

Glue Crawlers
스키마탐색

① AWS Glue

Catalog



Glue Catalog
데이터카탈로그

Glue Schema Registry
스트리밍 데이터
카탈로그

Transform



Glue Studio
데이터를 시각적
으로 변환



Glue Studio Notebooks
데이터를 대화식
으로 변환



Glue Databrew
코드없이 변환

서버리스 데이터 통합 엔진

Python | Spark

LAKE HOUSE

Machine
Learning



Relational



Big Data



Log
Analytics



Data
Lake



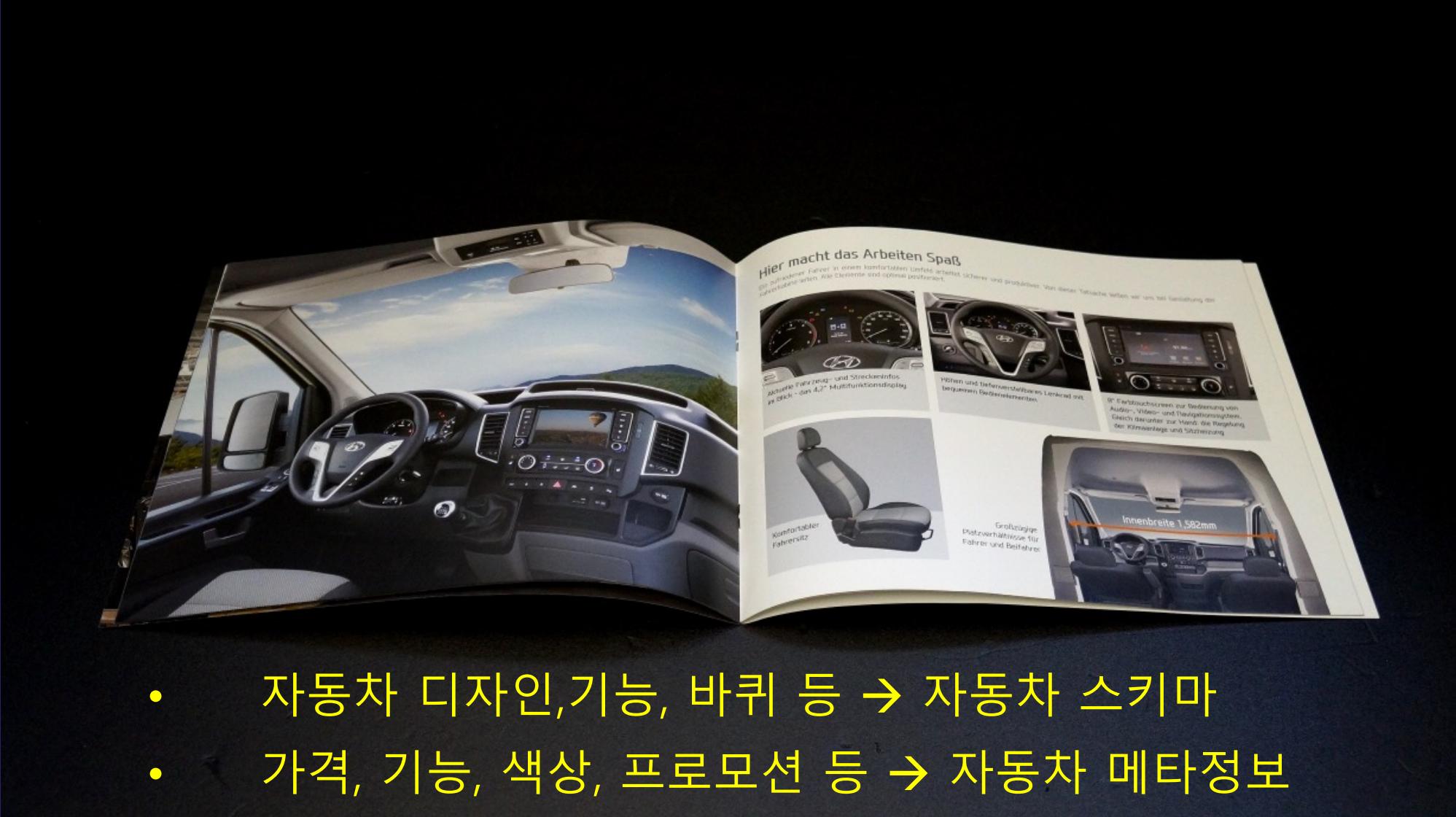
NoSQL

Glue Data Catalog



© 2022, Amazon Web Services, Inc. or its affiliates.

Catalog ??



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

AWS Glue – Data Catalog

Services ▾ | Resource Groups ▾ | 🔍

AWS Glue | Tables > simpletweets_json | Last updated 10 Aug 2017 | admin/damle-Isengard @ 4135... | N. Virginia | Support

Data catalog | Edit table | Delete table | View properties | Compare versions | Edit schema

Name: simpletweets_json
Description:
Database: analytics
Classification: json
Location: <s3://gluesampleddata/simpletweets.json>
Connection:
Deprecated: No
Last updated: Thu Aug 10 16:25:24 GMT-700 2017

Properties: sizeKey 456580, objectCount 1, UPDATED_BY_CRAWLER, S3Crawler, CrawlerSchemaSerializerVersion 1.0, recordCount 1001, averageRecordSize 456, CrawlerSchemaDeserializerVersion 1.0, compressionType none, typeOfData file

기본 메타정보 (분포 통계)

Tutorials

스키마 | Explore table | Add job

Schema

	Column name	Data type
1	entities	struct
2	id	bigint
3	retweeted	boolean
4	text	string
5	user	struct

중첩 필드 구조

user schema details

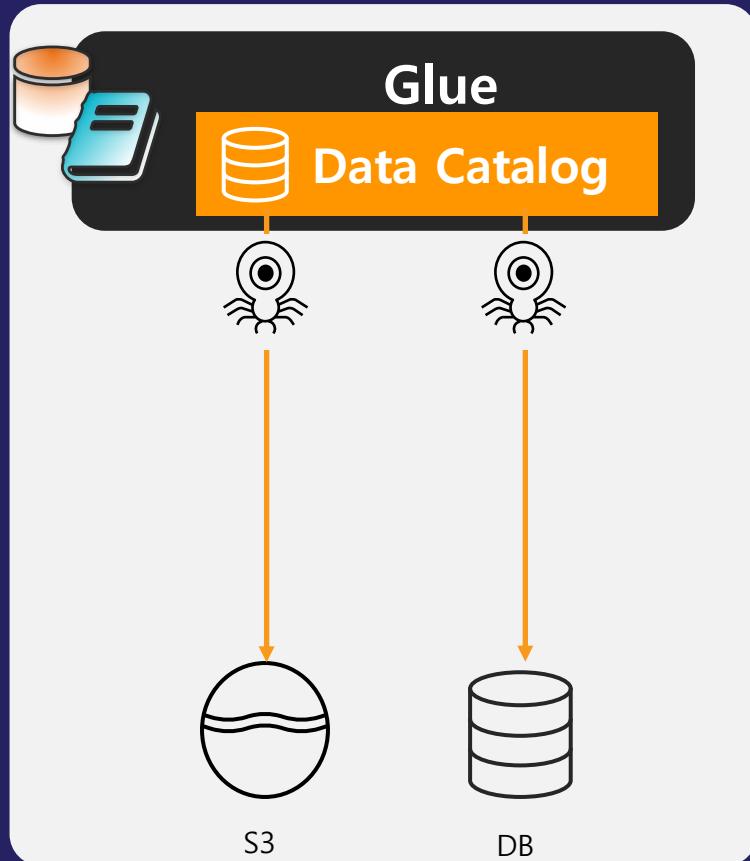
```
STRUCT
contributors_enabled:BOOLEAN
description:STRING
favourites_count:INT
followers_count:INT
friends_count:INT
id:INT
lang:STRING
location:STRING
name:STRING
profile_background_tile:BOOLEAN
```

Close

aws

© 2022, Amazon Web Services, Inc. or its affiliates.

AWS Glue – Data Catalog



S3, Database로 부터

크롤러가 자동적으로 데이터 스키마를 찾아서

테이블 스키마 정보와 메타 정보를 저장

AWS Glue – Data Catalog

The screenshot shows the AWS Glue Data Catalog interface. On the left, a sidebar lists various AWS services and resources. The main panel displays the properties of a table named 'githubevents_data'. Key details include:

- Name:** githubevents_data
- Description:** gitarchive
- Classification:** json
- Location:** s3://glue-sample-datasets/examples/data/
- Connection:** No
- Deprecated:** No
- Last updated:** Wed Nov 22 07:52:09 GMT-800 2017
- Input format:** org.apache.hadoop.mapred.TextInputFormat
- Output format:** org.apache.hadoop.hive.serde2.HiveIgnoreKeyTextOutputFormat
- Serde serialization lib:** org.openx.data.jsonserde.JsonSerDe

Under **Serde parameters**, there is a table with columns: sizeKey, 146649736039; objectCount, 1; UPDATED_BY_CRAWLER, githubarchive.

Under **Table properties**, there is a table with columns: CrawlerSchemaSerializerVersion, 1.0; recordCount, 27833145; averageRecordSize, 2423; CrawlerSchemaDeserializerVersion, 1.0; compressionType, gzip; typeOfData, file.

Below these sections is the **Schema** table, which lists 11 columns:

	Column name	Data type	Key
1	id	string	
2	type	string	
3	actor	struct	
4	repo	struct	
5	payload	struct	
6	public	boolean	
7	created_at	string	
8	org	struct	
9	year	string	Partition (0)
10	month	string	Partition (1)
11	day	string	Partition (2)

A yellow dashed box highlights the last three columns (year, month, day) and their corresponding 'Partition (0)', 'Partition (1)', and 'Partition (2)' values.

At the bottom, there are links for Feedback, English (US), Privacy Policy, and Terms of Use.

자동적으로 파티션 구조 파악

The screenshot shows the partitions of the 'githubevents_data' table. The top navigation bar includes 'Tables', 'githubevents_data', 'Last updated 22 Nov 2017', 'Table', and 'Version (Current version)'. Below this are buttons for 'Edit table', 'Delete table', 'Close partitions', 'Compare versions', and 'Edit schema'.

The main area displays a table of partitions, showing the following data:

year	month	day	
2017	02	15	View files View properties
2017	03	12	View files View properties
2017	05	17	View files View properties
2017	10	12	View files View properties
2017	12	18	View files View properties
2017	01	09	View files View properties
2017	03	07	View files View properties
2017	06	28	View files View properties

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

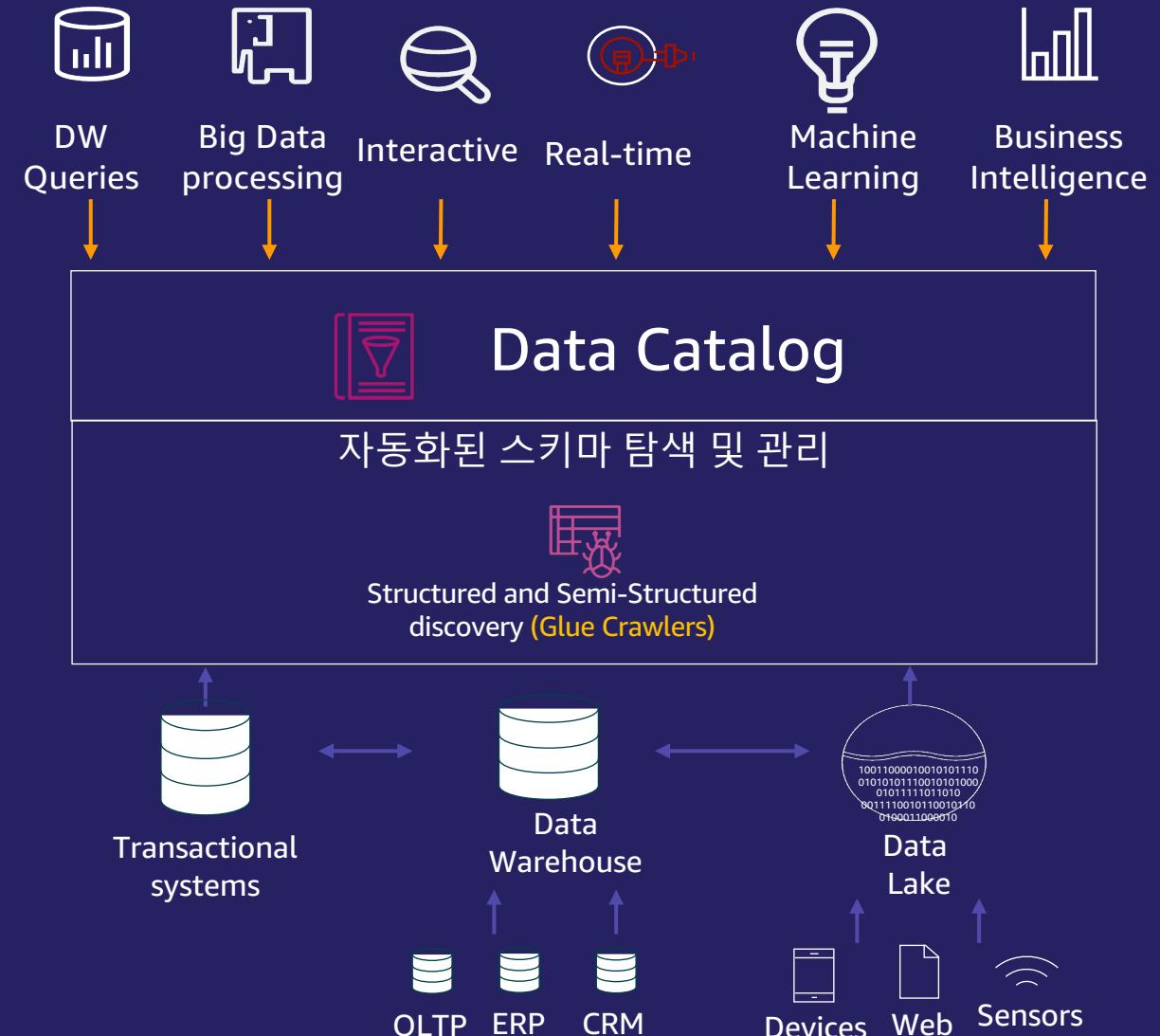
Glue Data Catalog/Crawler – 통합 데이터 카탈로그

데이터 이동 없음
= 낮은비용/관리

모든 데이터를 중앙에서 탐색 및 쿼리
= 생산성

정형/반정형 데이터 통합
= 빠른 인사이트

데이터 탐색 자동화
= 생산성



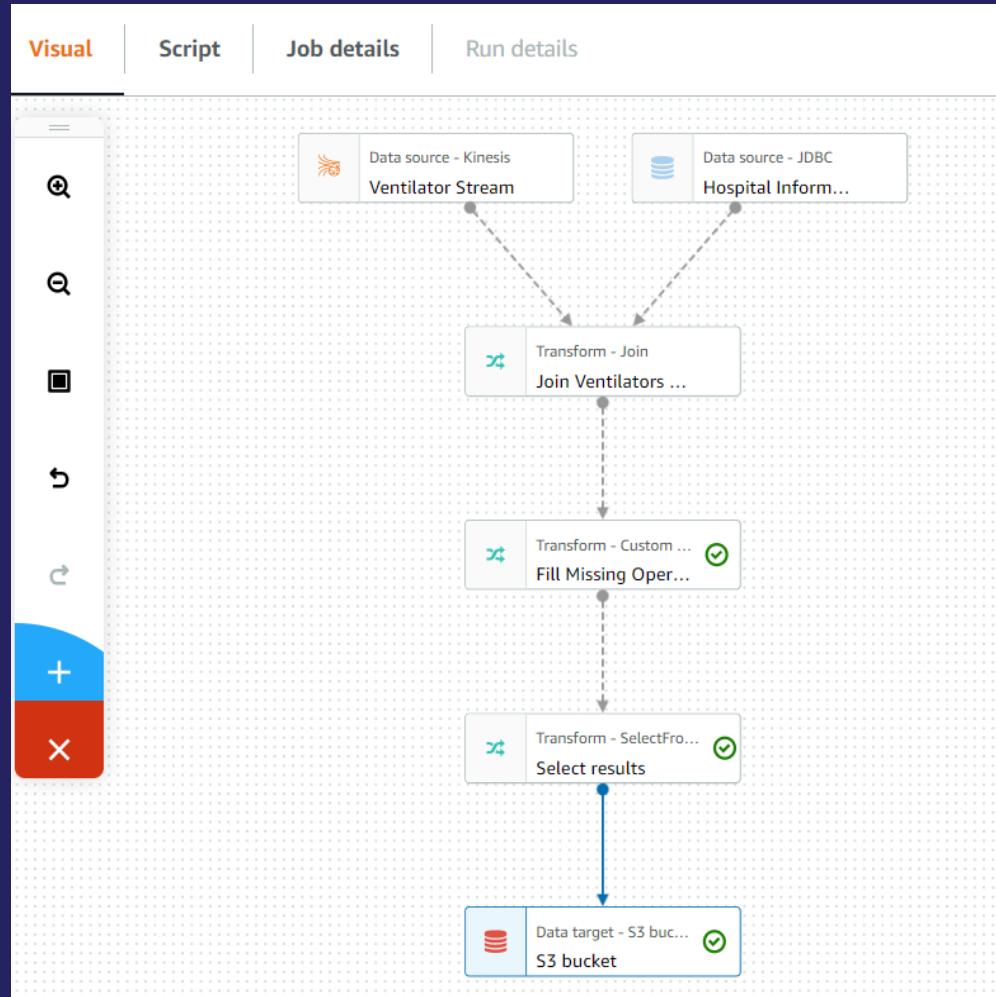


ETL



© 2022, Amazon Web Services, Inc. or its affiliates.

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

347

Running

23

Canceled

0

Success

246

Failed

78

Job Run Success Rate Info

Success Rate

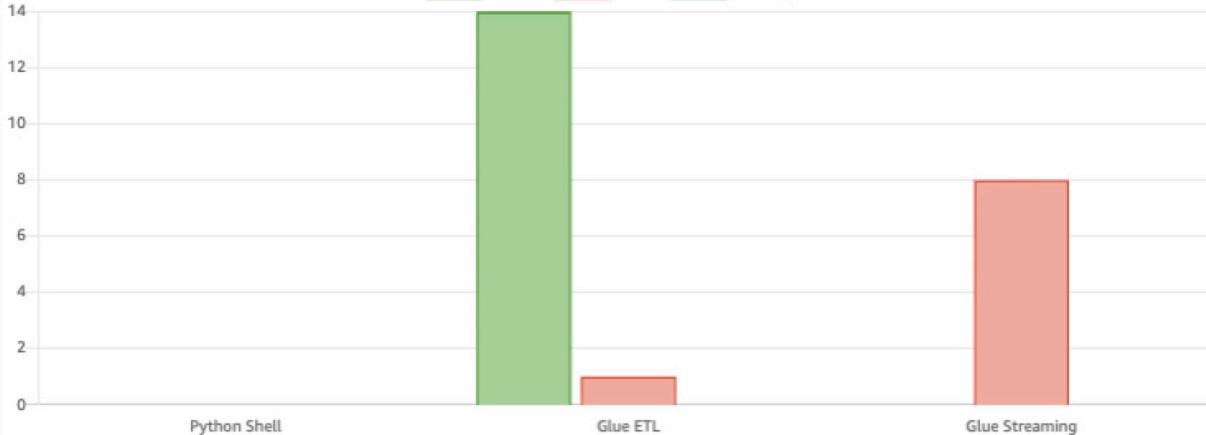
71%

Status

Service operating normally

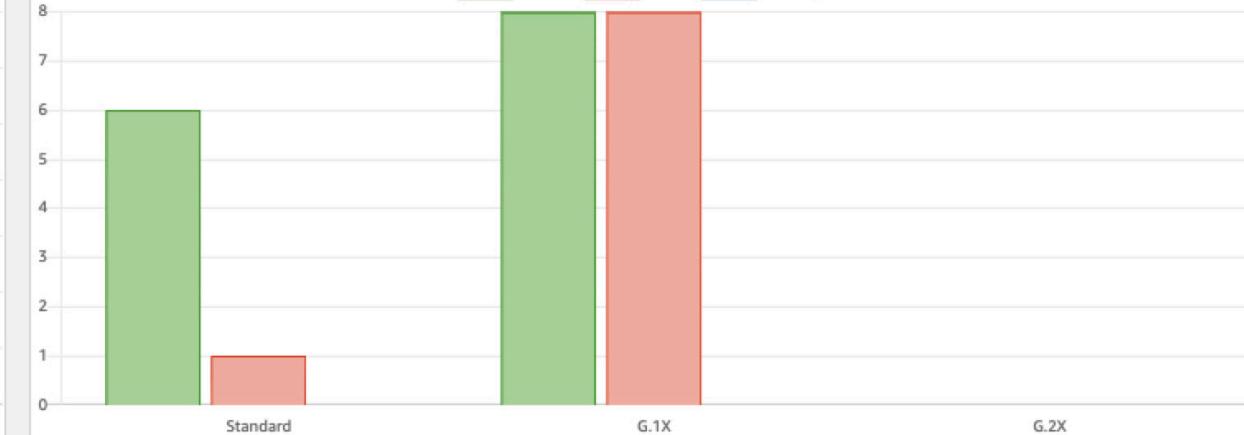
Job Type Breakdown Info

Success Failed Running



Worker Type Breakdown Info

Success Failed Running



Job Runs Timeline Info

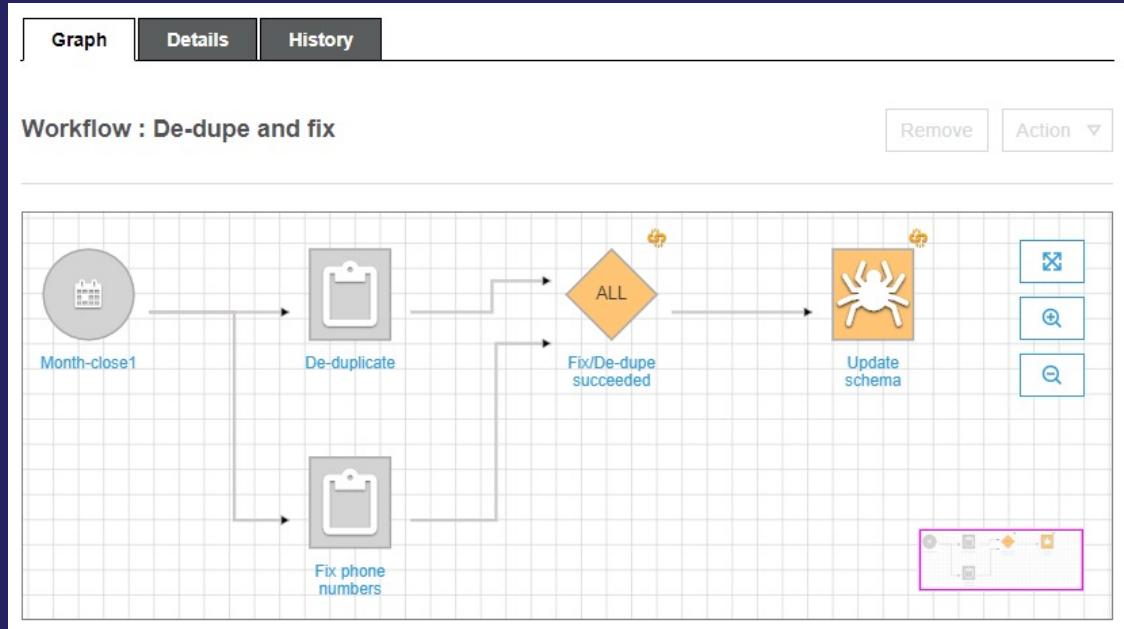
Success Failed Running Canceled



© 2022, Amazon Web Services, Inc. or its affiliates.

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

- 높은 총 소유비용
- 클러스터 조정 및 관리 노력

02
DIY Spark
on Cloud

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

03
Cloud
ETL

7x

온프리미스 옵션 대비
최대 7x 저렴

No lock-in

오픈소스 Spark
Python, Scala으로 작성

5x

자체 Spark 클러스터를
구축하는 것 보다 최대
5x 저렴

4x

자체 Spark 클러스터
보다 유지 관리를 최대
4x 감소

55%

다른 클라우드 공급자와
비교해 최대 55% 저렴

Serverless

관리할 서버가 없고,
인프라 비용도
포함되지 않음



AWS Glue 기능

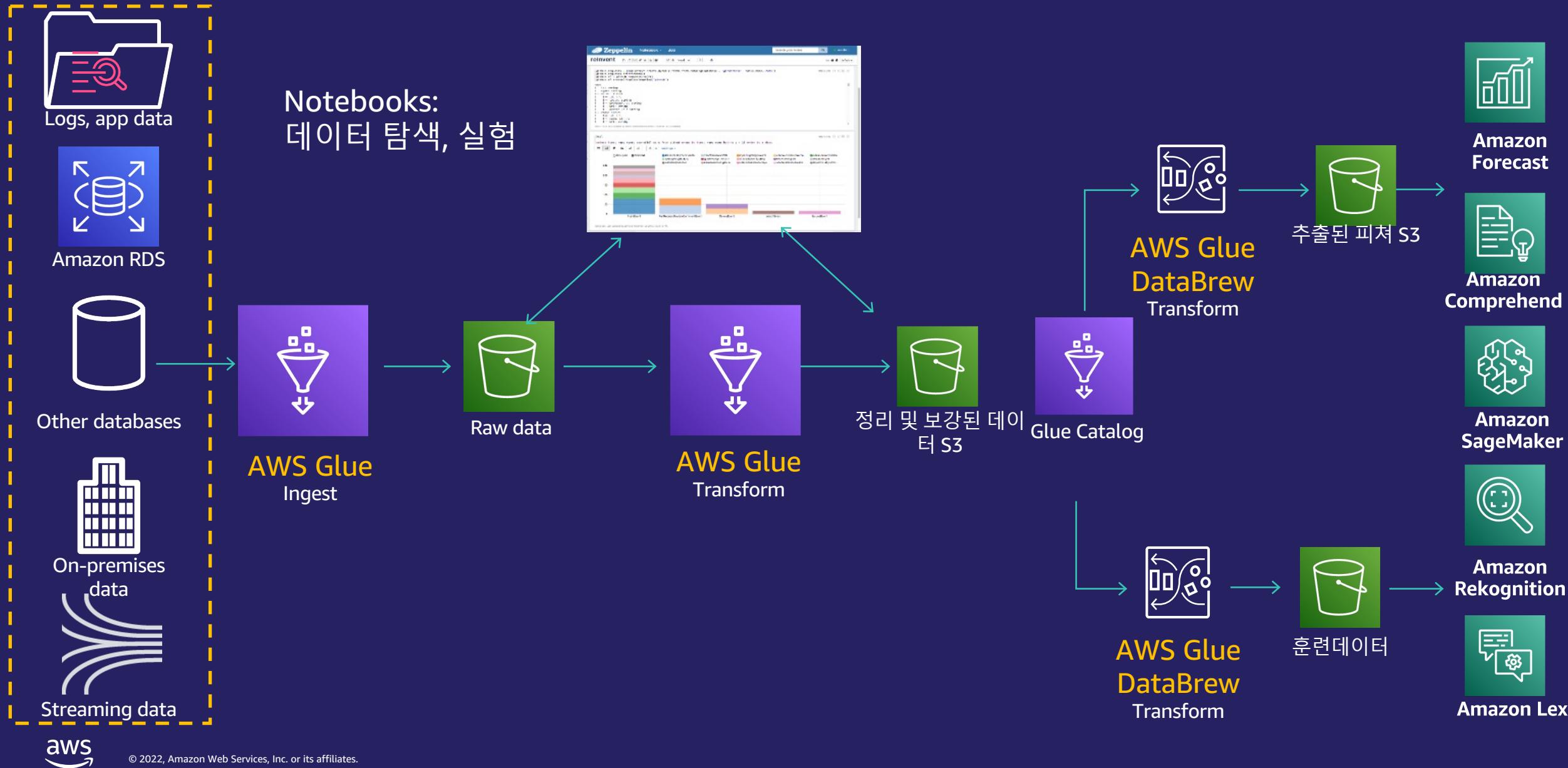
- 서비스 데이터 통합 엔진
- 데이터 탐색
- 데이터 변환
- 데이터 파이프라인 운영

Reference architectures

사용사례: 온프레미스 ETL 도구를 모던화 하여 데이터 웨어하우스로 로드



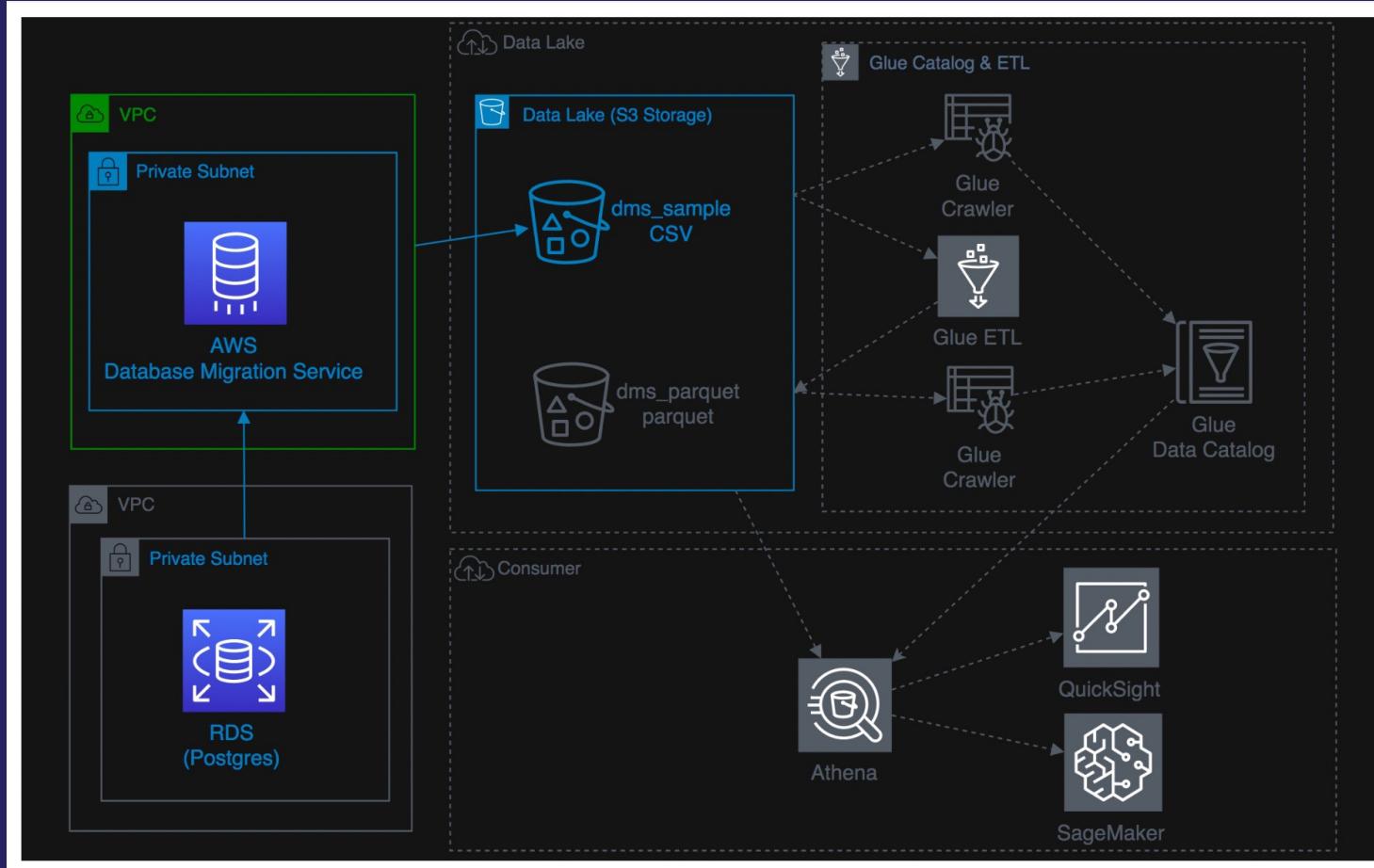
사용사례: 머신러닝을 위한 원시 데이터 준비



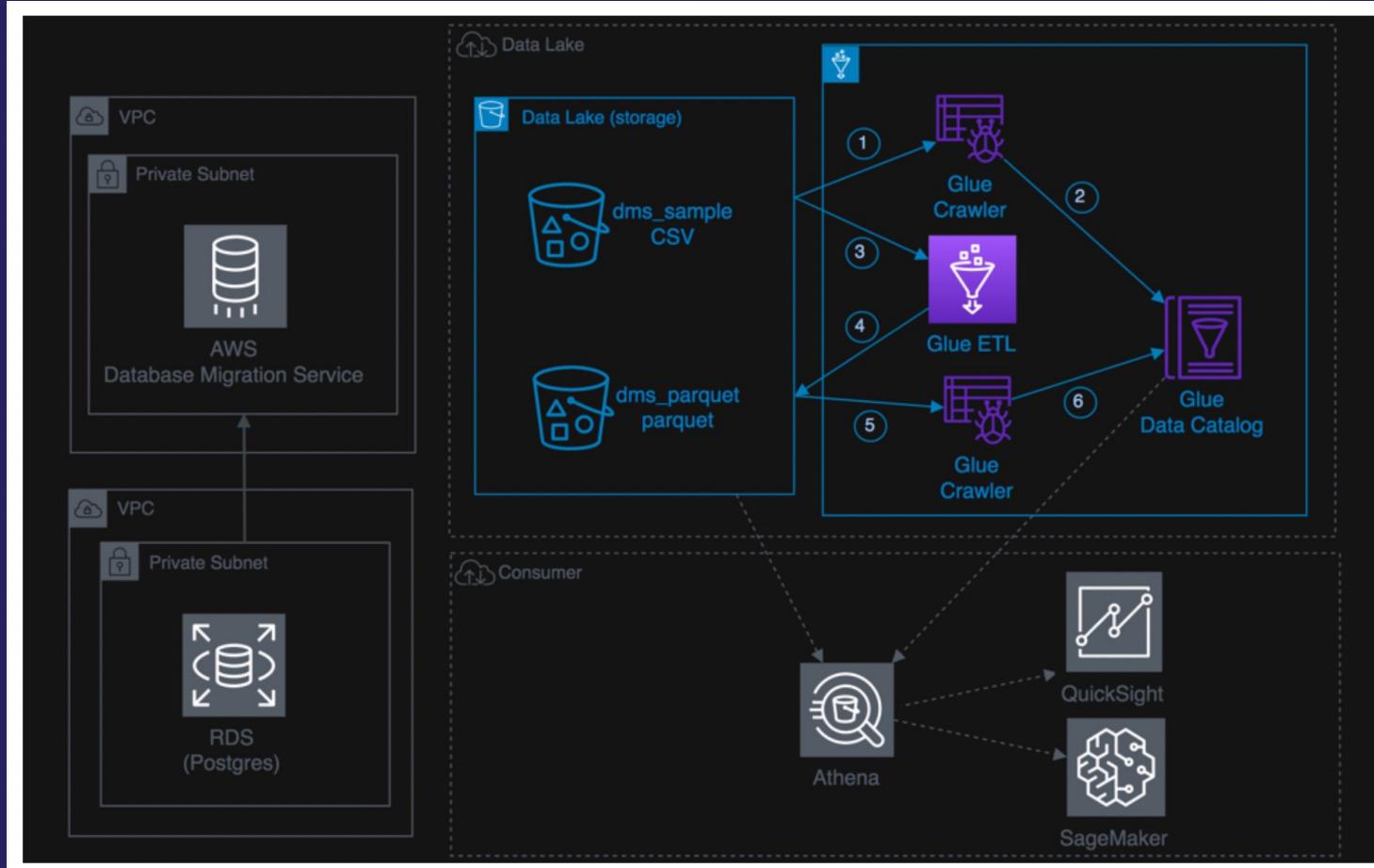


Thank you!

Today's Lab



Today's Lab

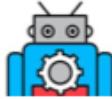


[참고] Parquet?



'나무조각을 붙여놓은 마룻바닥'

연봉이 400인 사람?

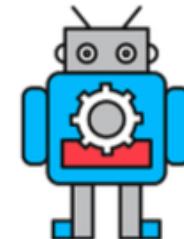


나이	부서	연봉
21	A	200
31	B	300
41	A	400



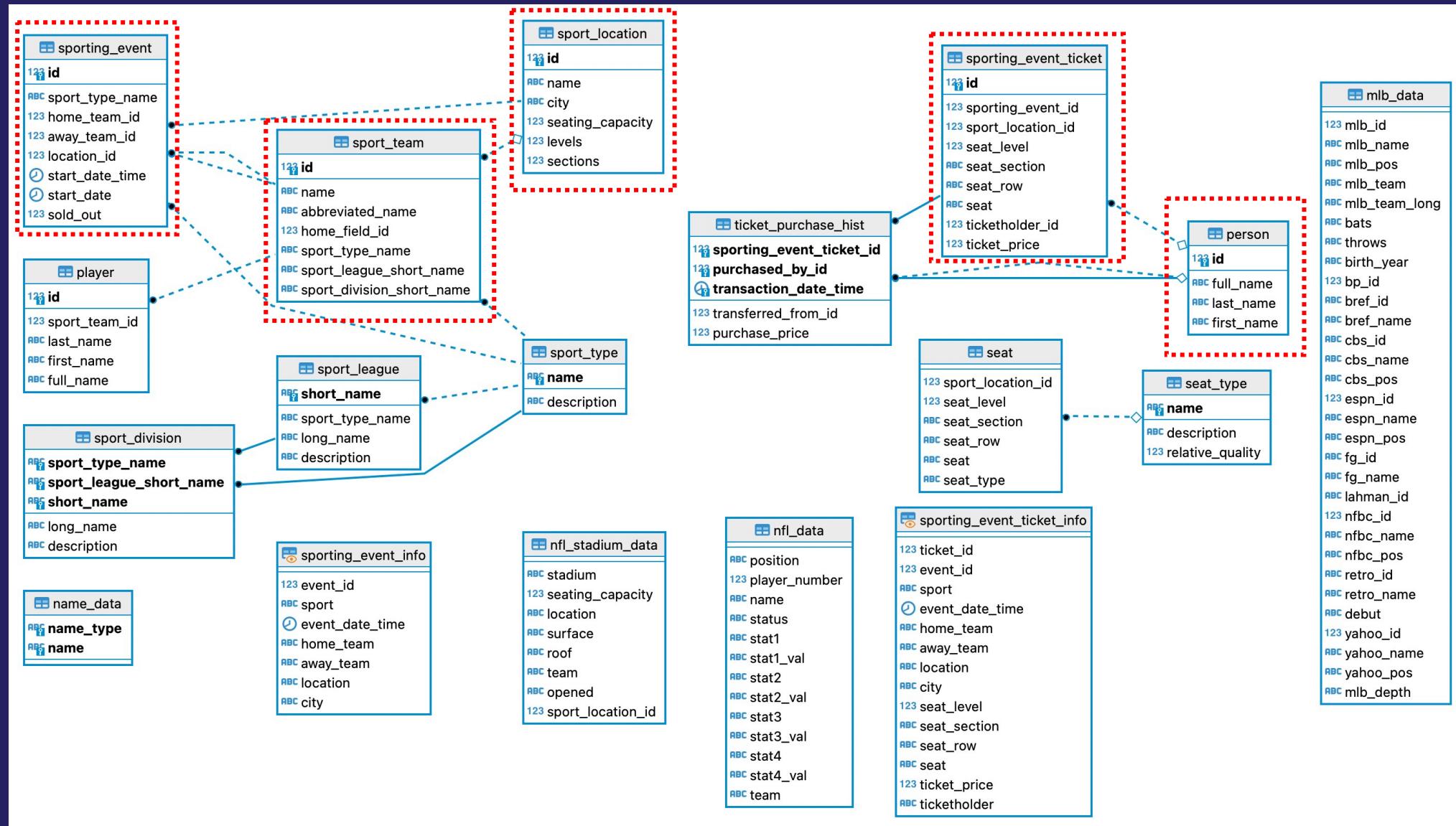
Row-Based →

21	A	200	31	B	300	41	A	400
----	---	-----	----	---	-----	----	---	-----



Column-Based →

21	31	41	A	B	A	200	300	400
----	----	----	---	---	---	-----	-----	-----



파케 테이블 명명시

