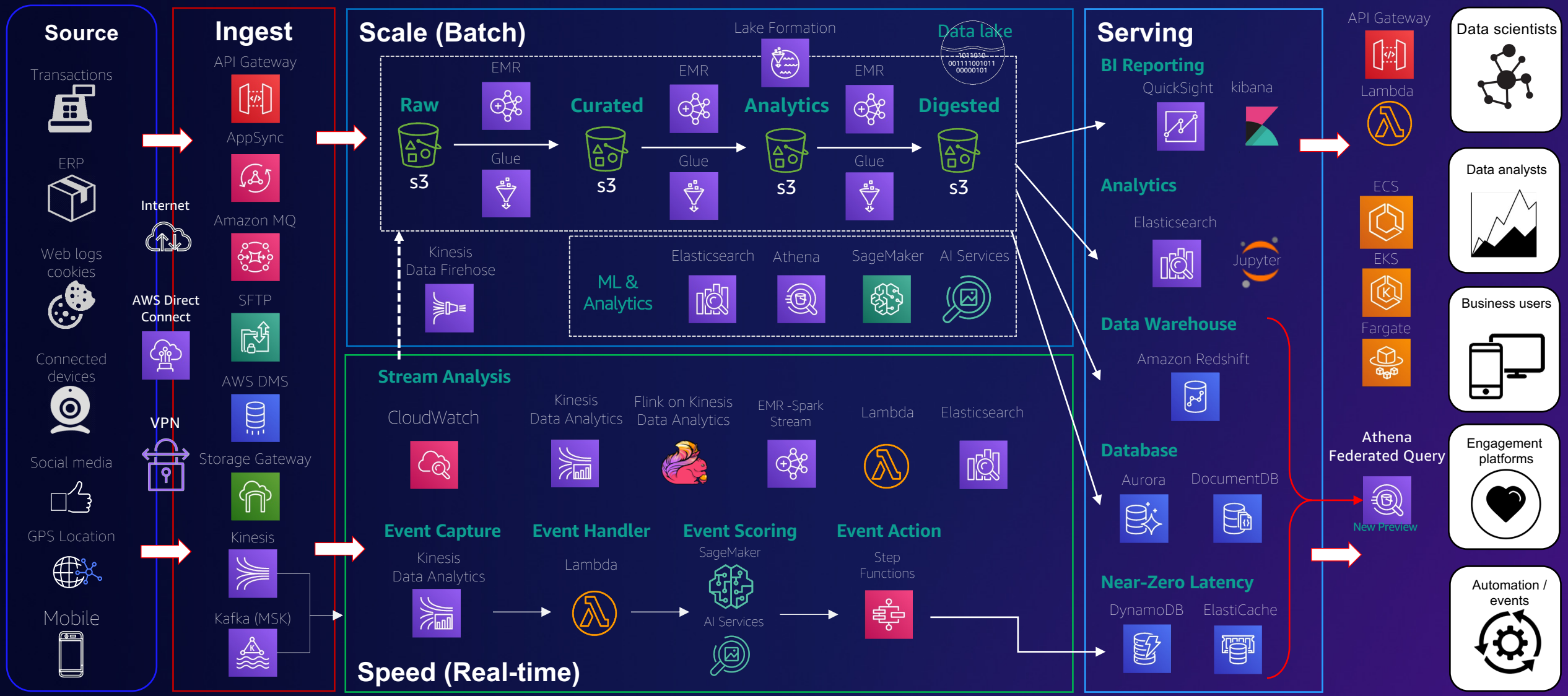
Data Engineering Day
Athena / QuickSight

김희정
AWS Solutions Architect
2022. 10. 26



Data sources

Amazon DynamoDB

ERP

Web logs / cookies

Connected devices

Ingest

Database Migration Service

AWS Snowball

Amazon Kinesis

Amazon MSK

Catalog

AWS Glue

Store

Amazon S3

Process & Analyze

Amazon Athena

Amazon EMR

Amazon Redshift

Amazon Elasticsearch

Consume

Amazon QuickSight

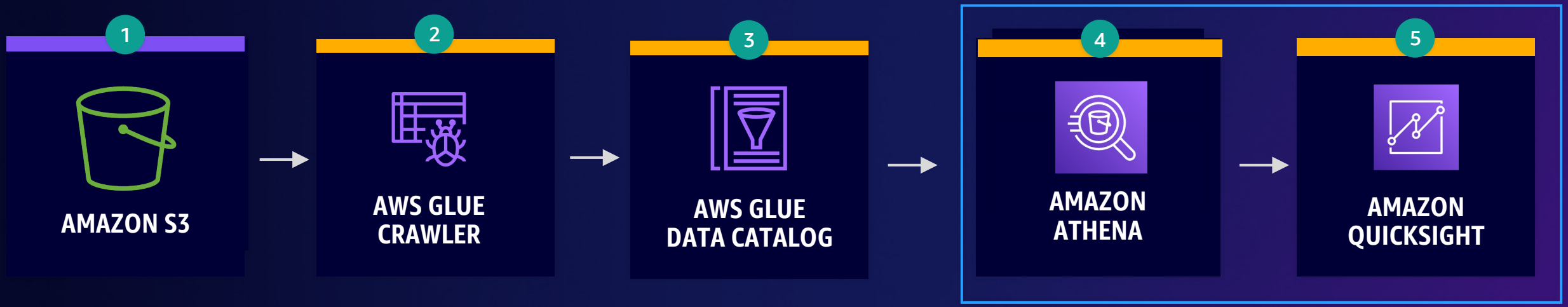
Jupyter Notebooks

Amazon API Gateway

looker

BI Tools

Overall Flow



1. S3에 데이터 적재
2. Crawlers가 데이터 셋을 스캔 후 Glue Data Catalog를 생성함
3. Glue Data Catalog = 중앙 메타 데이터 저장소
4. Catalog되면 Analytics에 쓰일 준비 완료
5. QuickSight를 통한 시각화

Agenda

Athena

- Athena란?
- Use cases
- 기능 Deep dive

QuickSight

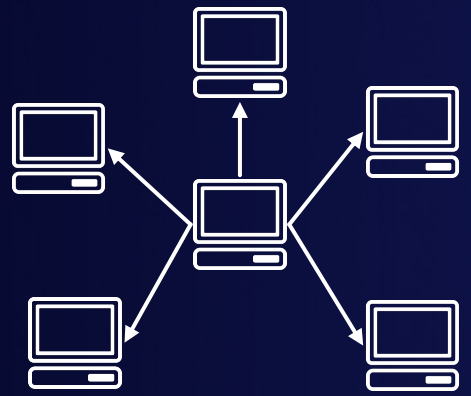
- QuickSight란?
- 기능 및 예시

Amazon Athena

Athena?

표준 SQL을 사용해 Amazon S3에 저장된 데이터를 간편하게 분석할 수 있는 대화식 쿼리 서비스

Serverless / SQL Query Engine / Amazon S3



Distributed

SQL Query Engine

S3

Athena의 장점

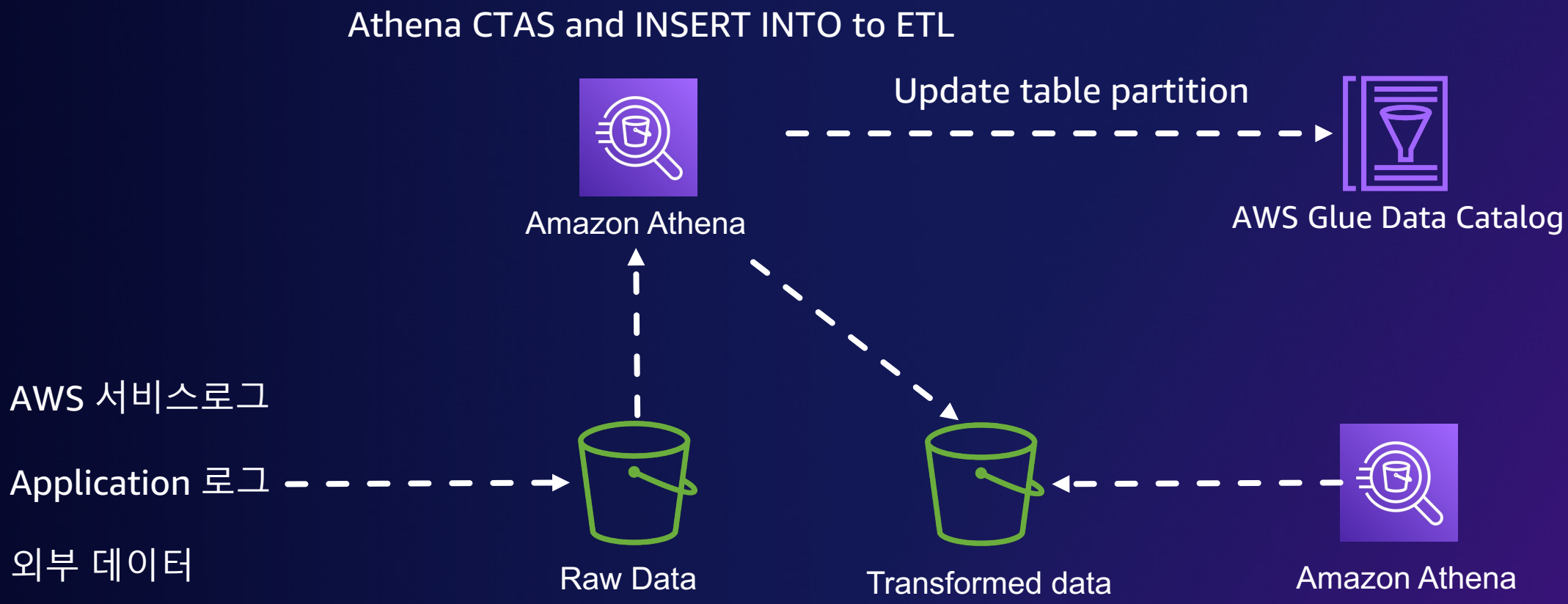
- 스토리지 / 컴퓨팅 노드 분리
- 빠른 Ad-hoc 쿼리
- 서버리스 - 인프라 관리 불필요
- 스캔된 데이터만큼 과금
- IAM을 통한 인증 및 암호화

기능

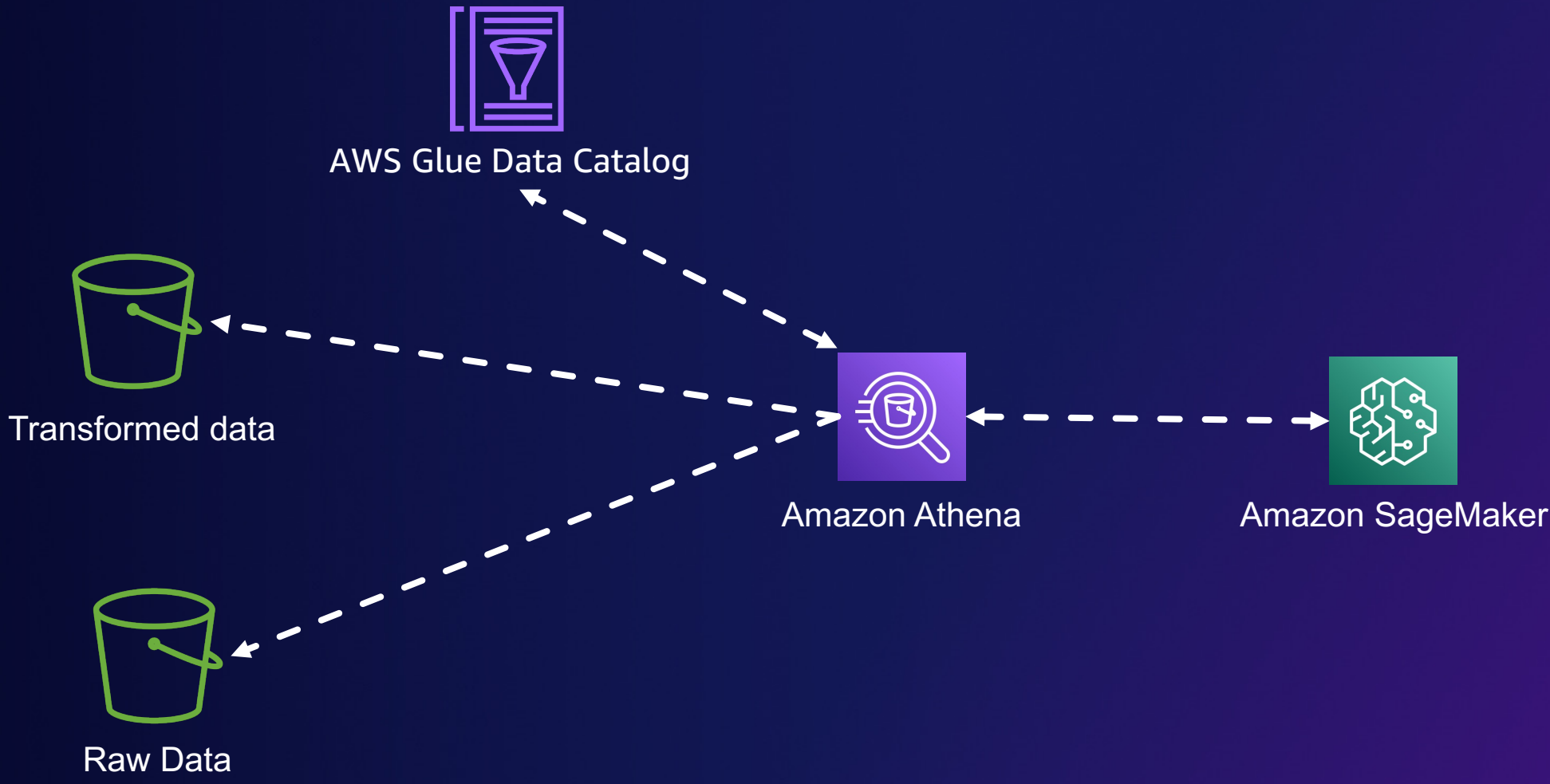
- ANSI SQL / Presto(분산 쿼리 엔진) 지원 쿼리 지원
- 표준 데이터 포맷 지원
  - CSV, Apache Weblogs, Json, Parquet, ORC
- Federated Query
  - Datawarehouse나 다른 소스와 연동 가능

Use Case 1/3 – Athena + BI Tools

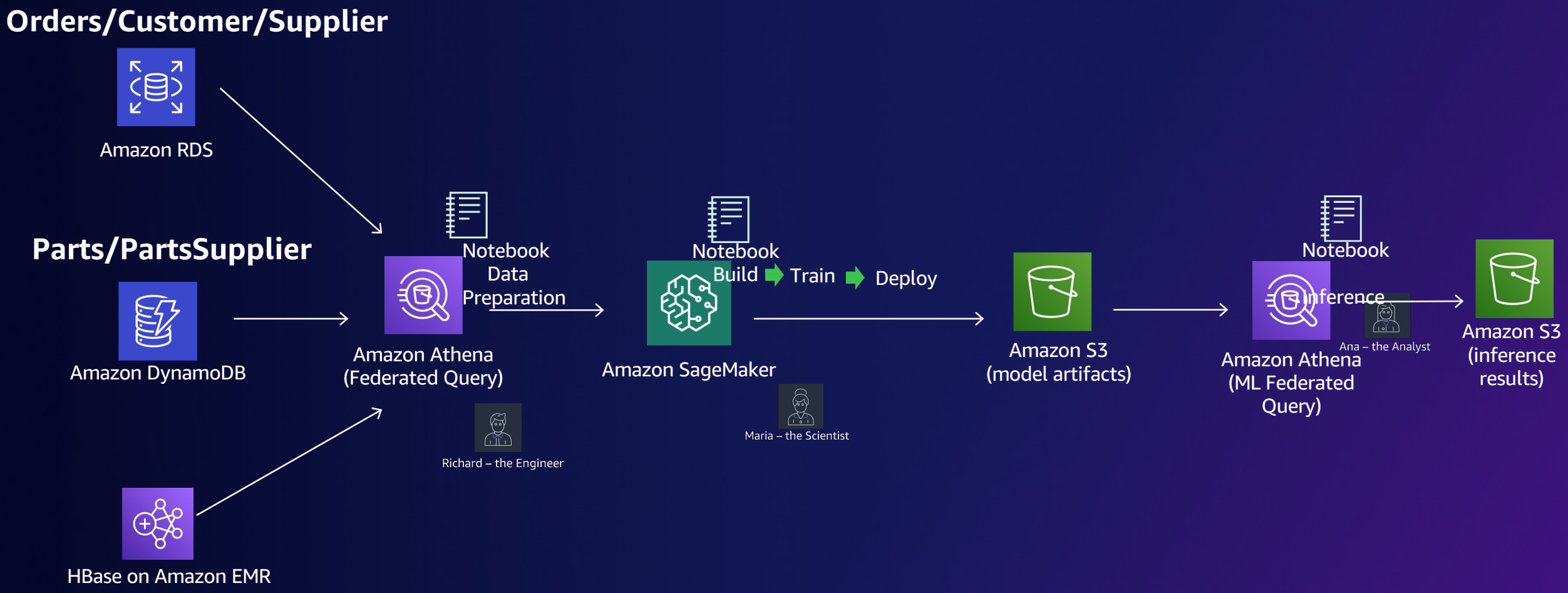
Use Case 2/3 – ETL and query



Use Case 3/3 – With ML



Use Case 3/3 – With ML



Amazon **Athena**

New data source connectors


SAP HANA


Teradata


Cloudera


Hortonworks


Snowflake


Microsoft SQL Server


Oracle


Google BigQuery


Azure Data Lake Storage Gen2


Azure Synapse

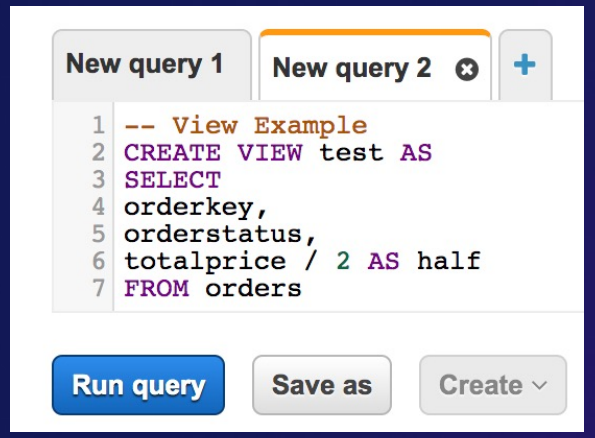
View / CTAS

View

- 논리적 테이블

CTAS

- CREATE TABLE AS SELECT
- 필요한 데이터만 포함된 복사본 생성
- https://docs.aws.amazon.com/athena/latest/ug/ctas-example.html




```
CREATE TABLE new_table
WITH (
        External_location =
        ‘s3://athena_result/parquet_table’,
        Format = ‘Parquet’,
        Parquet_compression = ‘SANPPY’
)
AS SELECT *
FROM old_table;
```

Parquet

: 오픈소스, Columnar 데이터 파일 포맷

연봉이 400인 사람?

나이	부서	연봉
21	A	200
31	B	300
41	A	400
51	B	500

Row-Based (Horizontal partitioning)
OLTP ➔

21	A	200	31	B	300	41	A	400

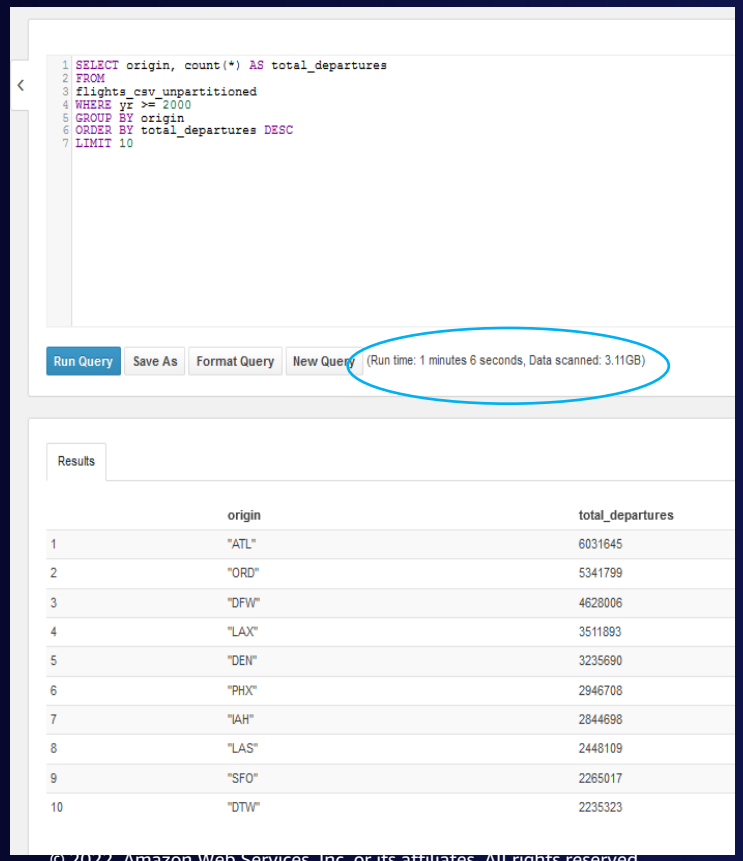
Column-Based (Vertical partitioning)
OLAP ➔

21	31	41	A	B	A	200	300	400

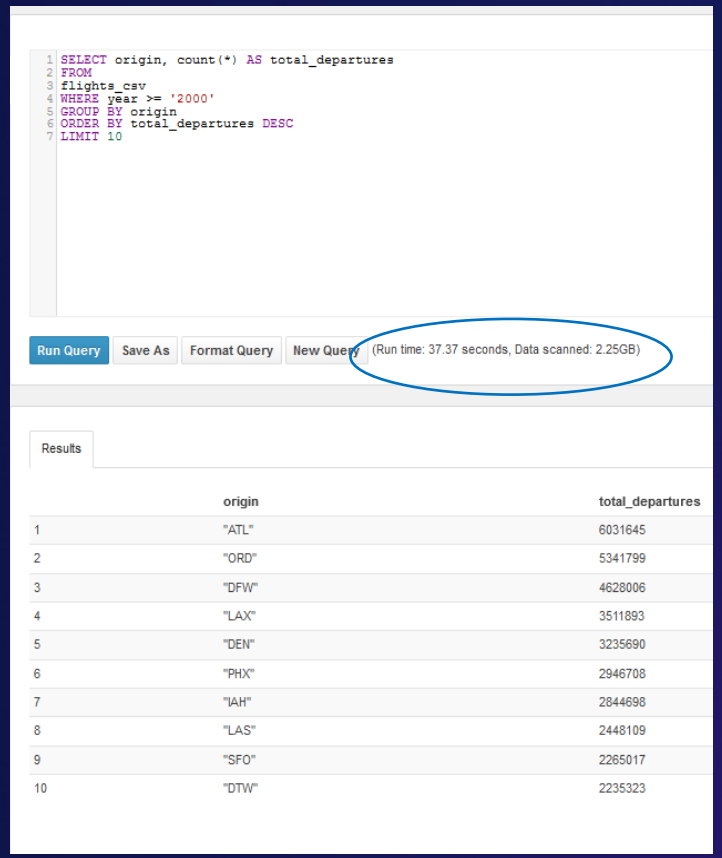
Partitioning

- 데이터를 분할 저장하여 일부 데이터만 검색되게 만드는 기법
- 쿼리 실행시 스캔되는 데이터양을 제한

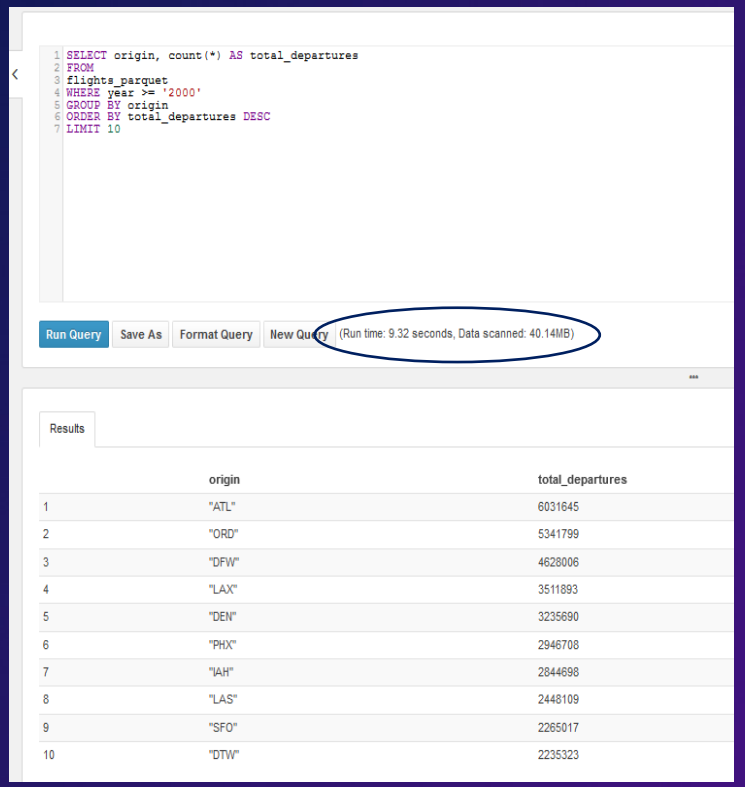
CSV: 37.37 s / 2.25GB



No partitions: 1min 6s / 3.11GB

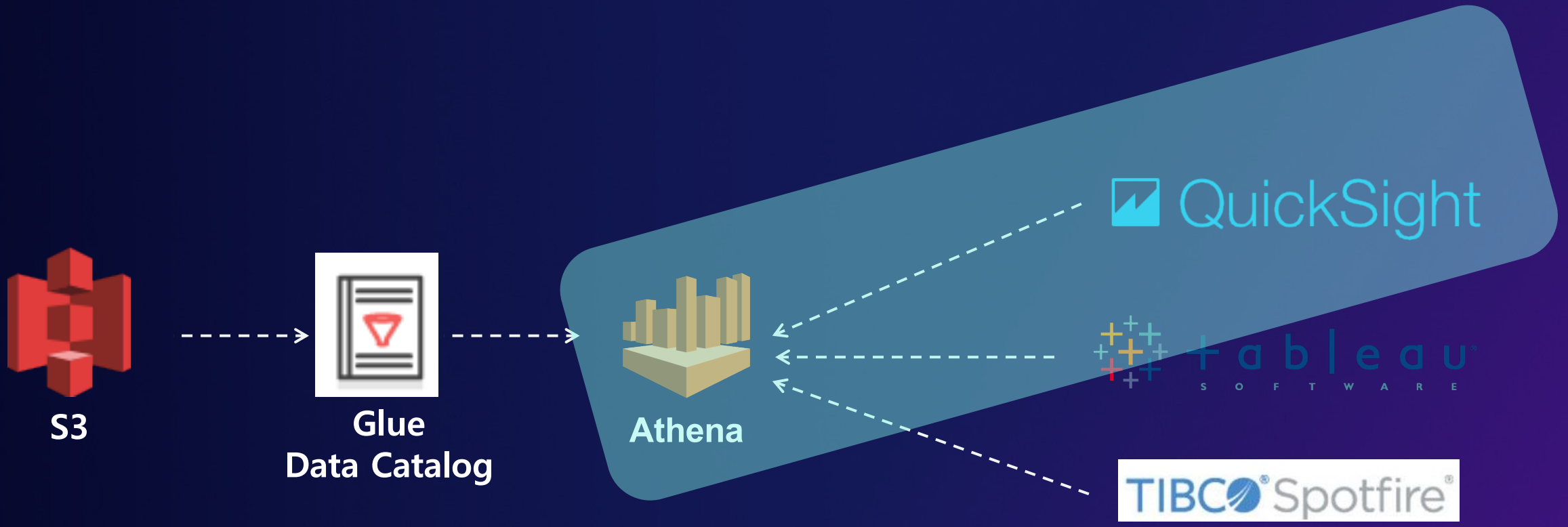


Parquet: 9.32s / 0.04GB



Amazon QuickSight

Athena + BI Tools



Amazon QuickSight

Only pay for what you need

모든 사용 사례에 대해 모든 규모에서 비용 효율적

BI를 기반에 두고 설계

Auto scaling and Serverless

서버를 프로비저닝하지 않고 10만 명의 사용자에게 전 세계적으로 배포

고가용성 내장

변화하는 조직 및 사용자 요구 사항을 충족하도록 자동 확장

Internal and/or external users

외부 사용자와 인사이트 공유

애플리케이션 안으로 통합

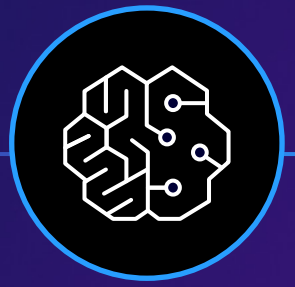
멀티 태넌트 및 보안

Deeply integrated with AWS services

AWS 데이터에 대한 안전한 비공개 액세스

S3 데이터 레이크와 통합

Augmented Insights on demand



NL을 이용한 질문

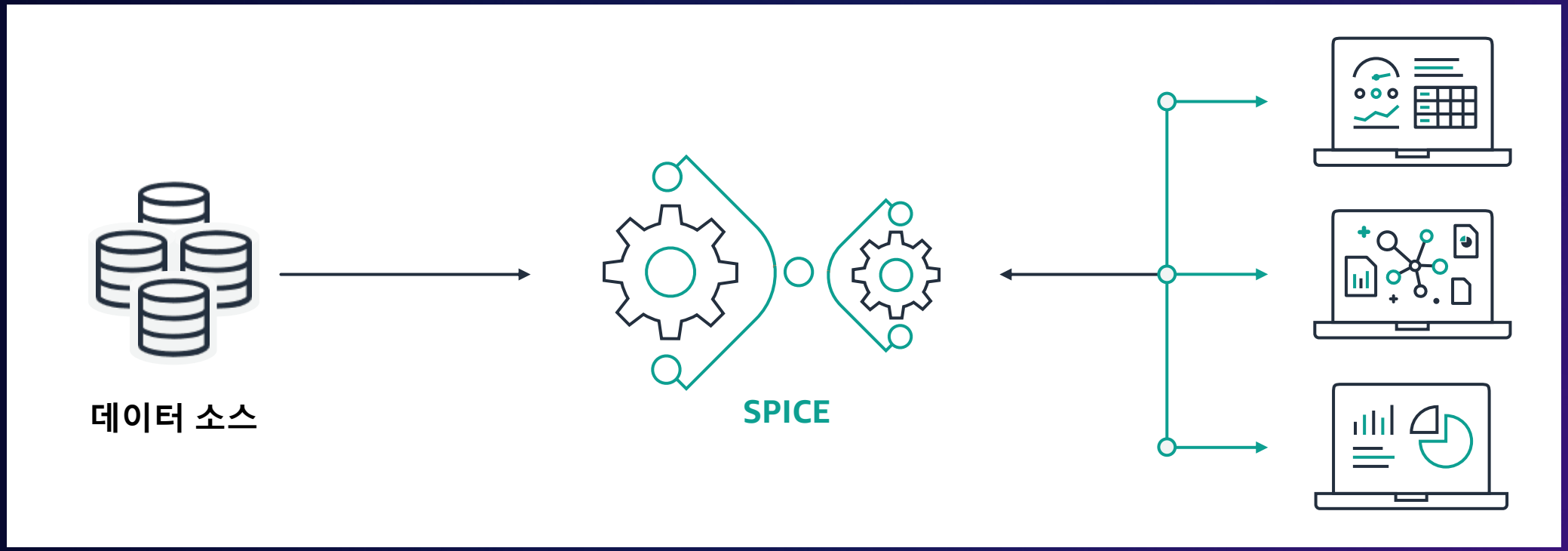
이상탐지 및 예측

Amazon SageMaker에서 고유한 모델 가져오기

SPICE

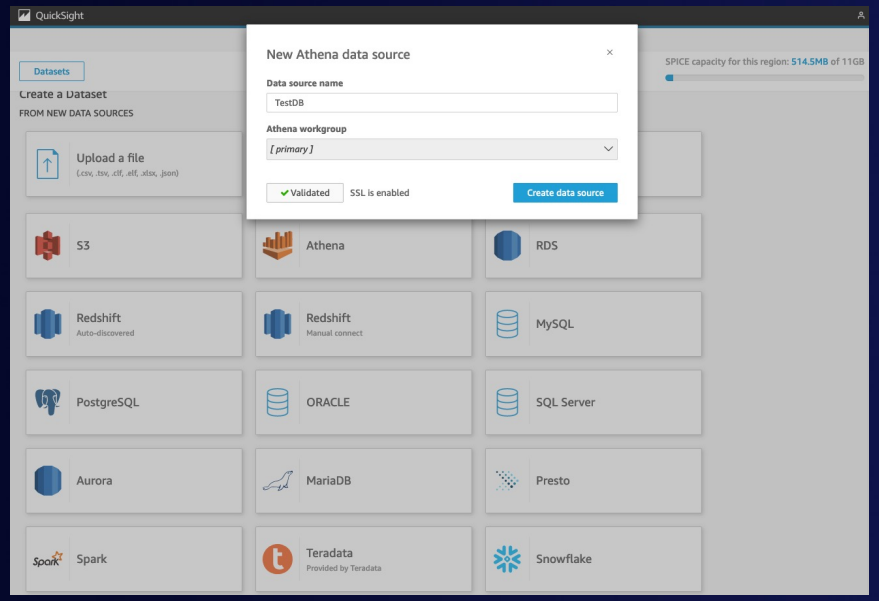
Sper-fast, Parallel, In-memory Calculation Engine

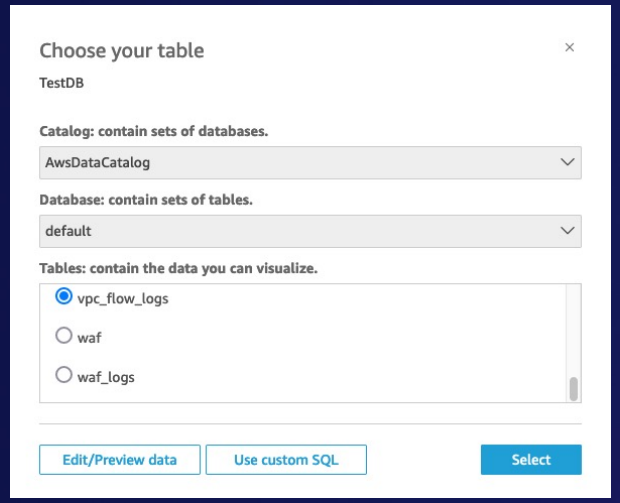
활성화된 사용자 수와 관계없이 성능과 규모를 제공하는 초고속 계산(Calculation) 엔진

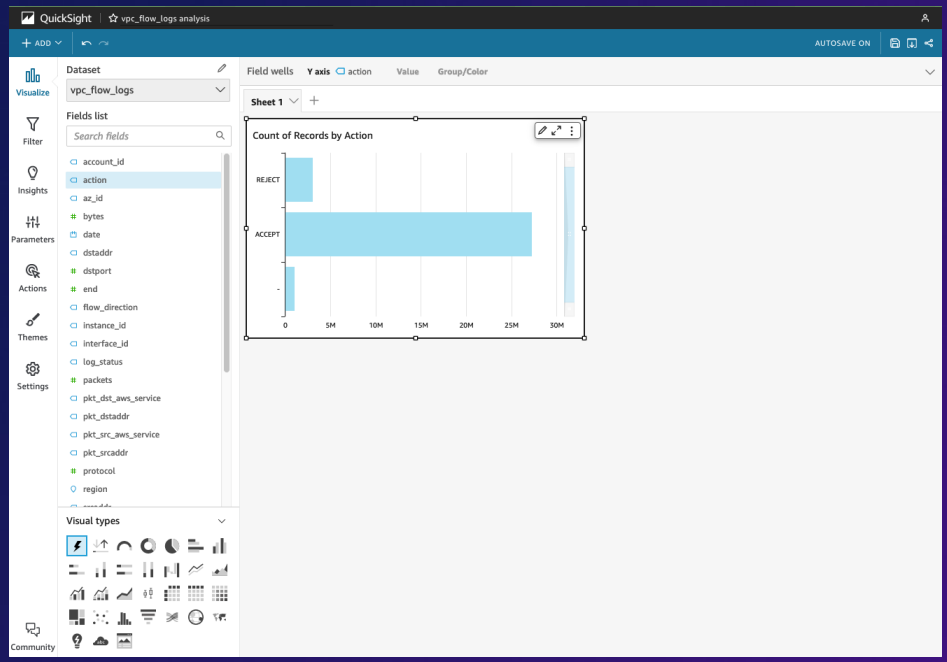


사용법

QuickSight에서 제공되는 다양한 데이터 소스를 선택,
원하는 필드값들을 대시보드에 표현할 수 있습니다



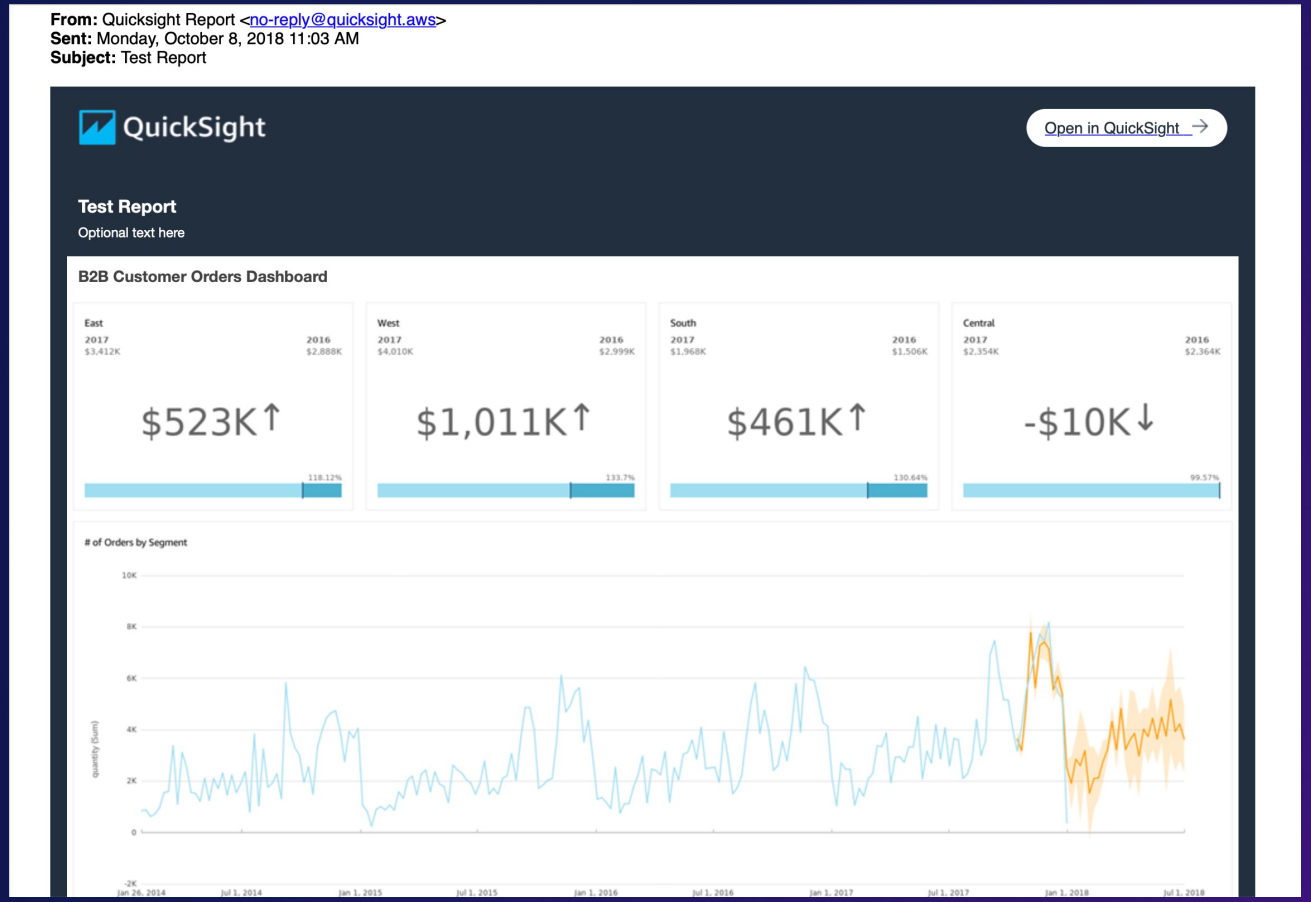




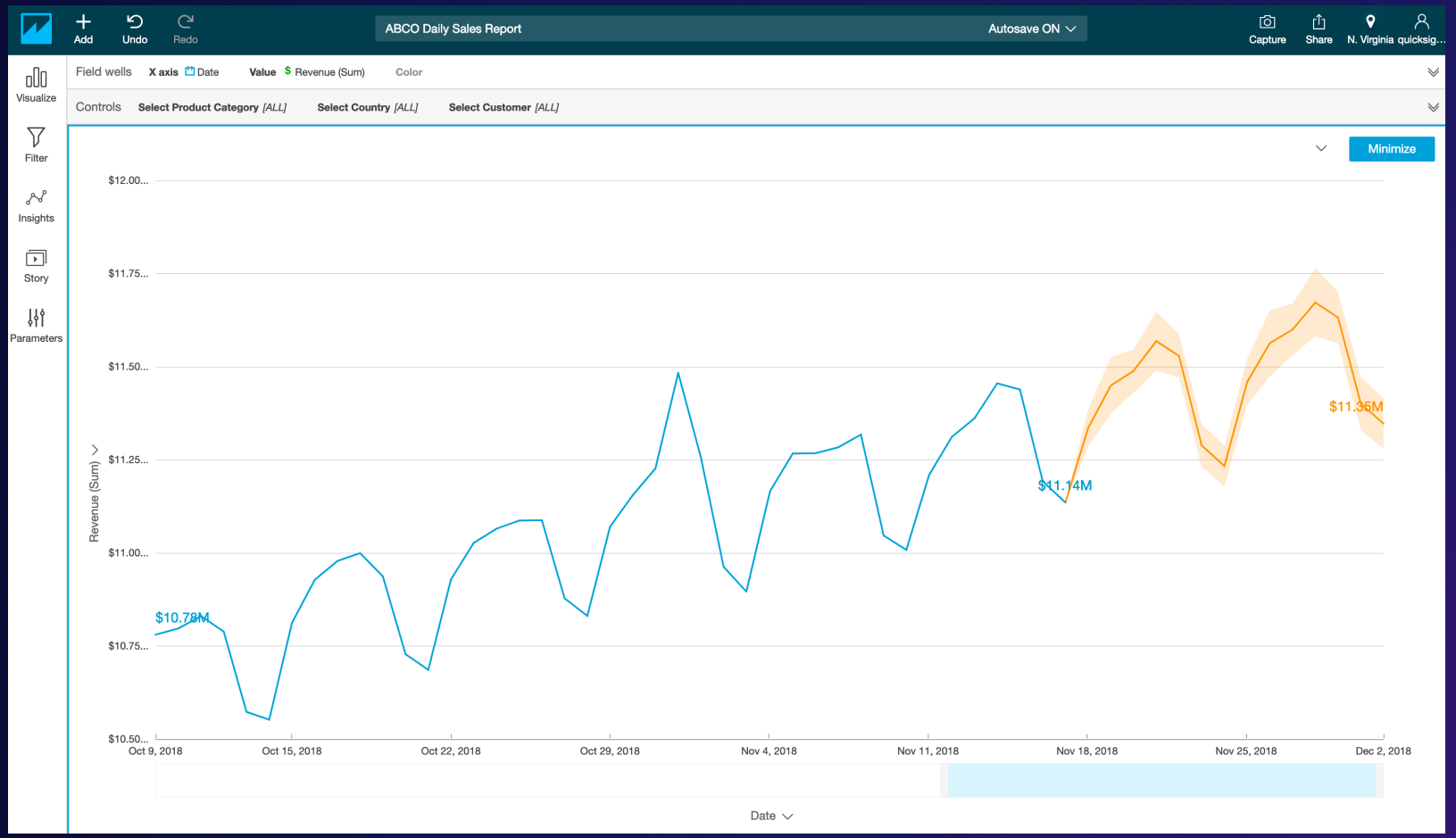
데이터 소스 **데이터 연결** **대시보드 생성**

리포팅 기능

- 일별, 주별 또는 월별 이메일 보고서 스케줄링
- 개별 사용자 또는 그룹에 전송



머신러닝 기반의 예측 서비스 제공

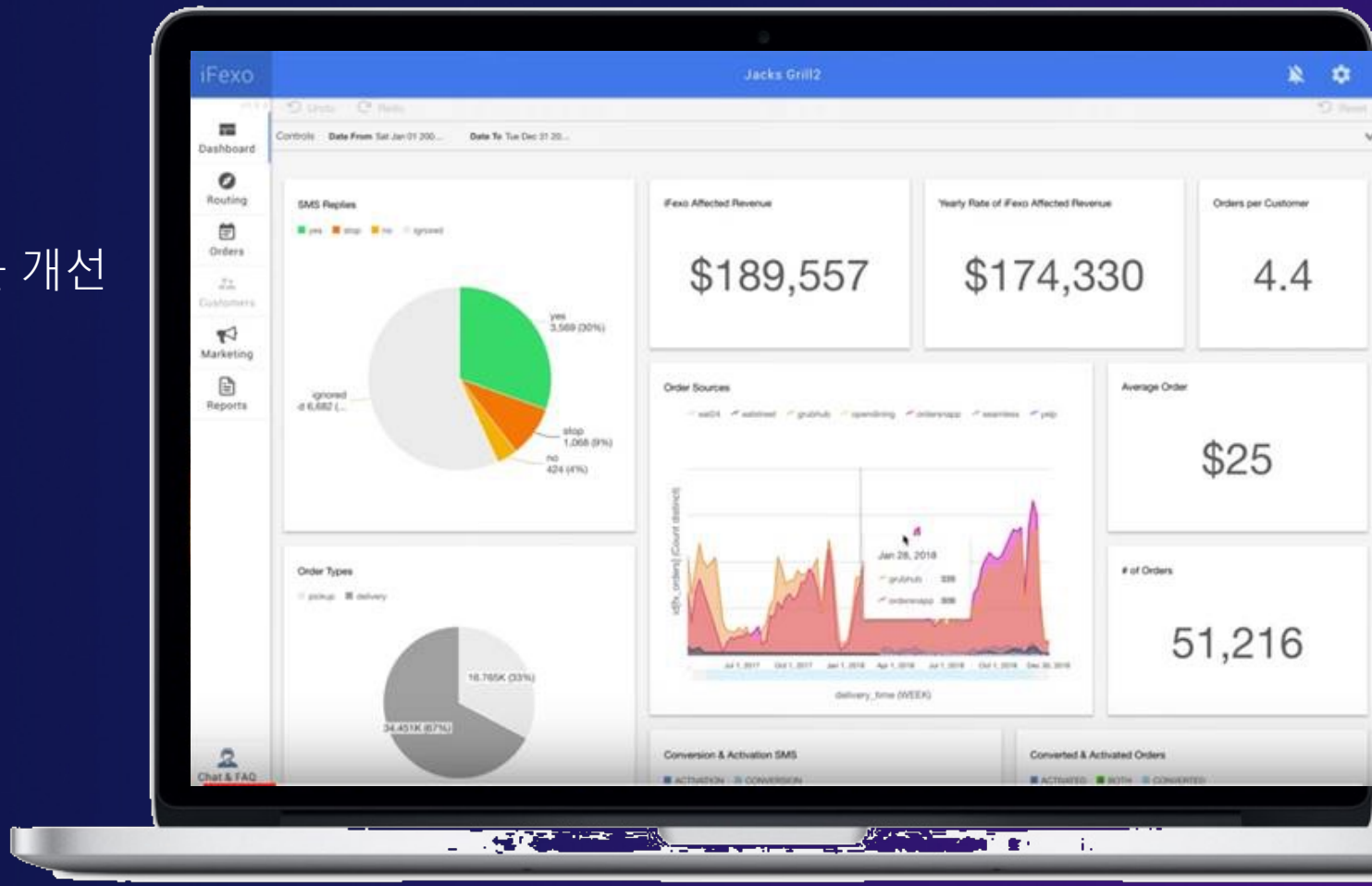


ML-powered forecasting

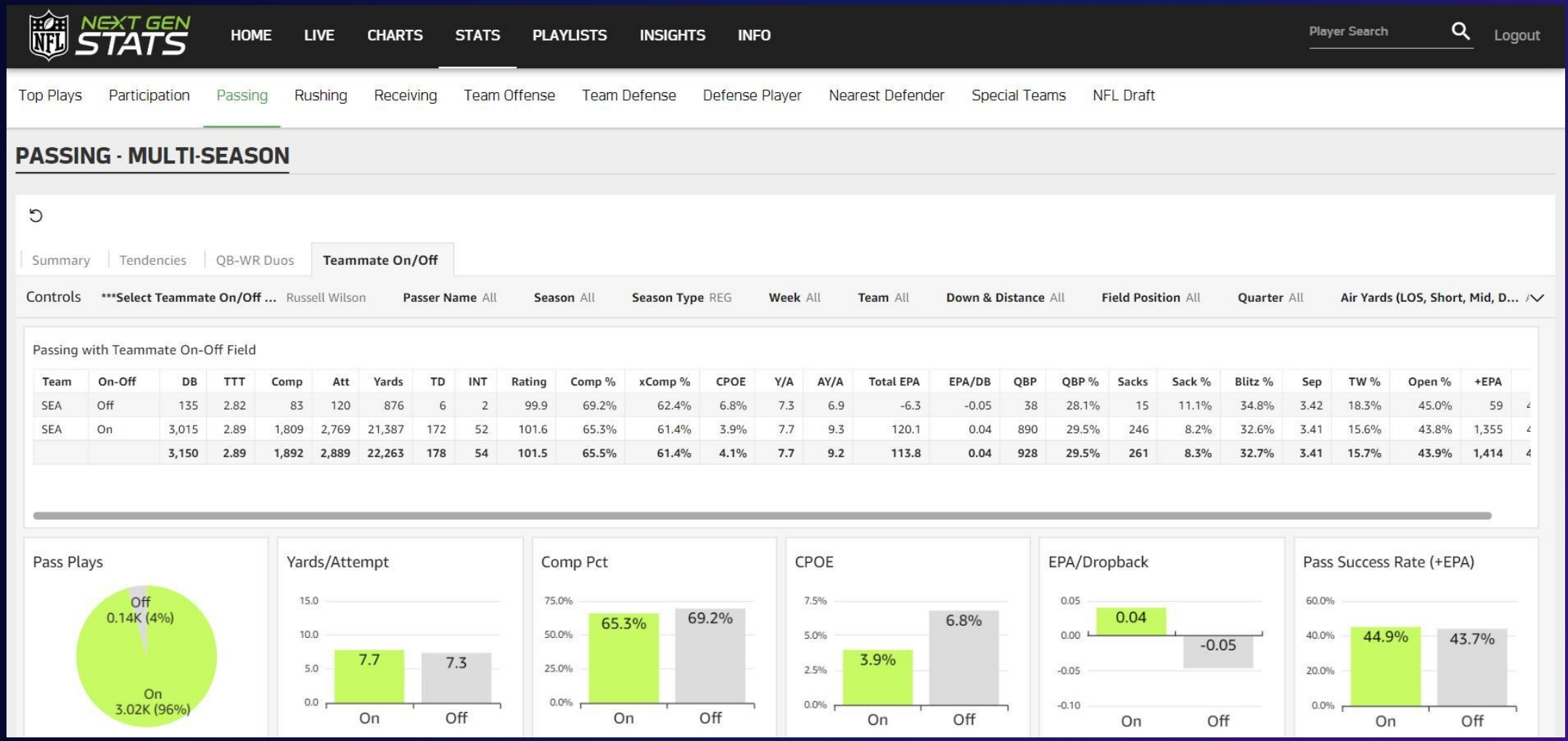
ML 전문 기술이나 Excel 데이터 모델링이 필요하지 않습니다

대시보드 임베딩

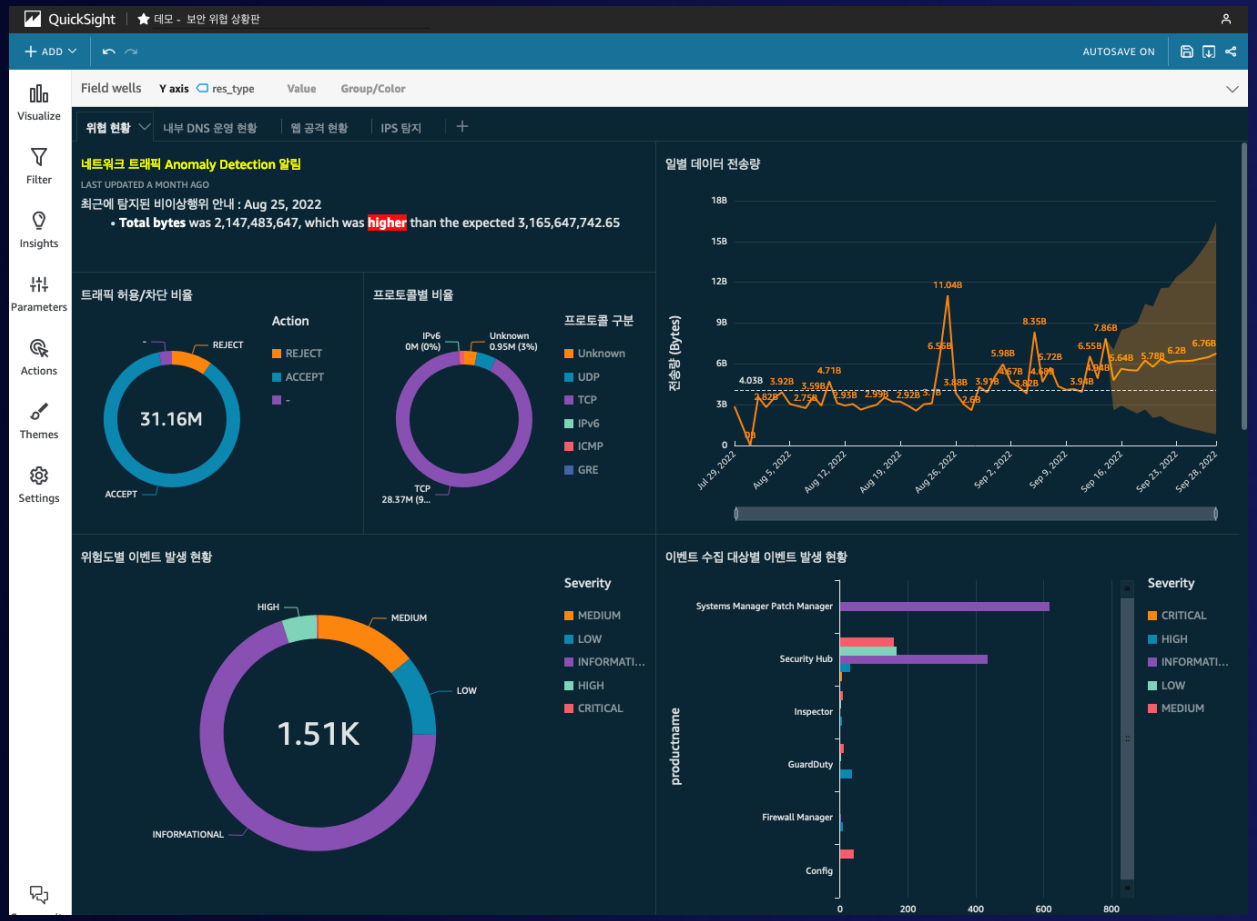
- 풍부한 분석 및 대시 보드로 애플리케이션을 개선
- 쉬운 유지 보수, 서버 관리 불필요
- 지속적인 새로운 기능 추가

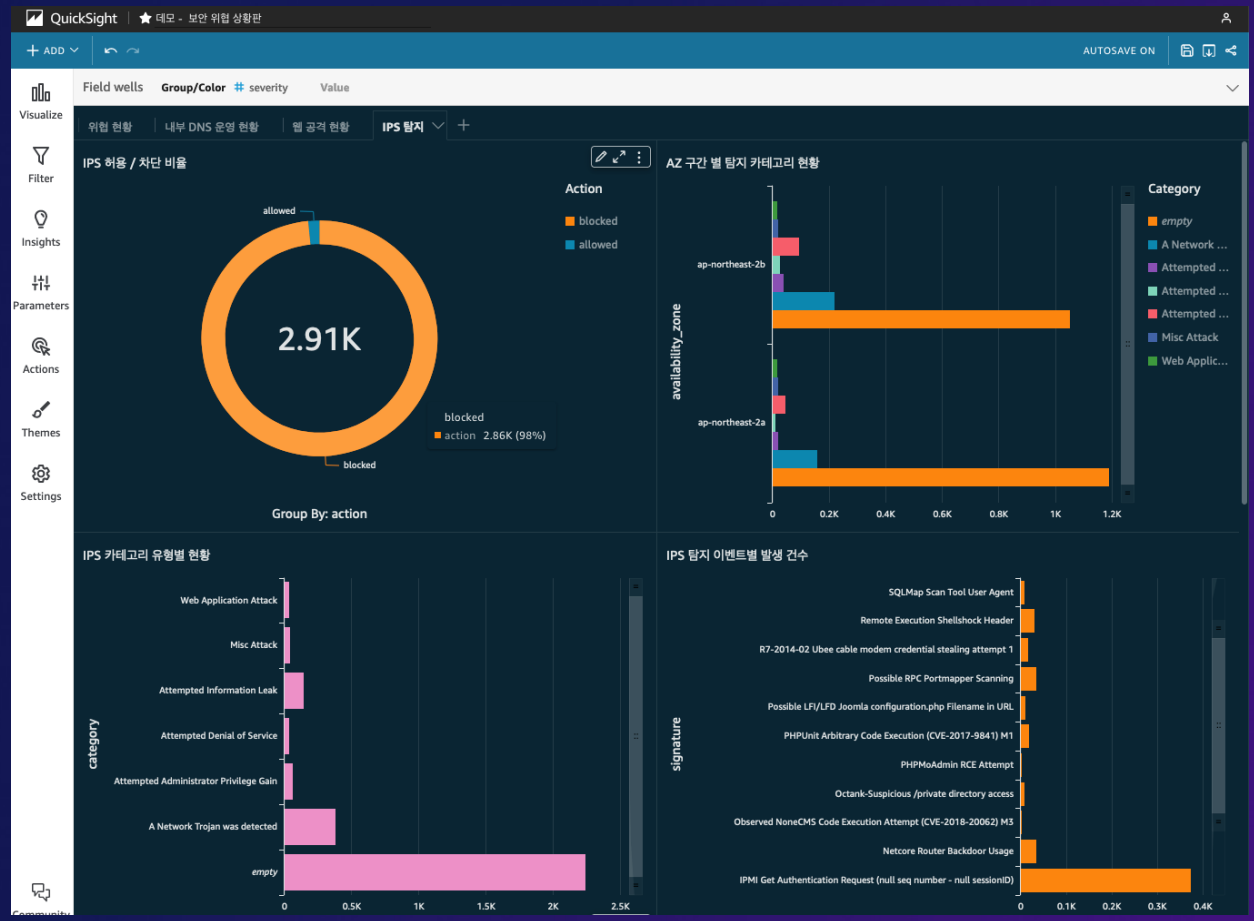


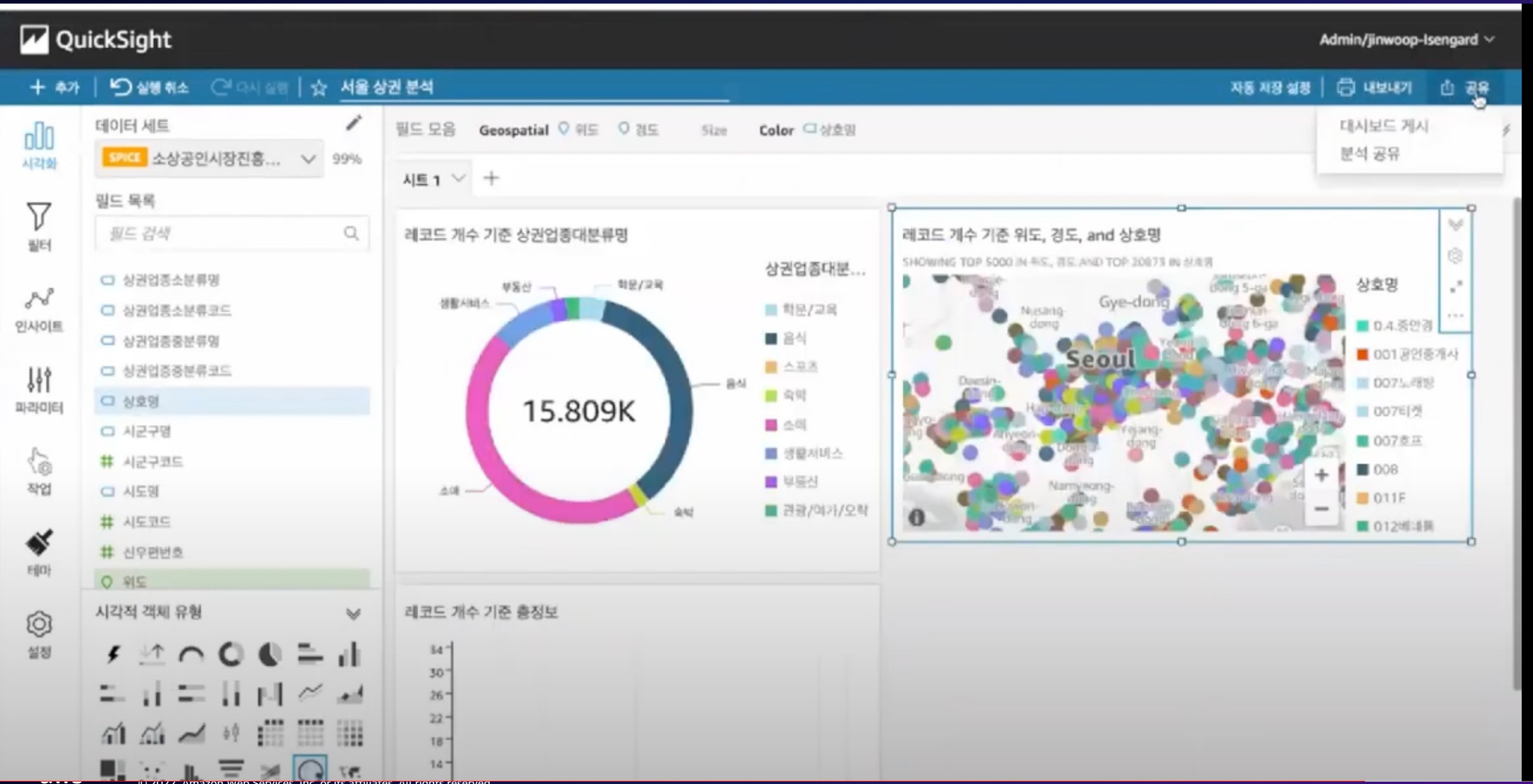
NFL Next Gen Stats



네트워크 로그 분석







Survey

https://bit.ly/app-modern-221026