

Data Engineering Day

Athena / QuickSight

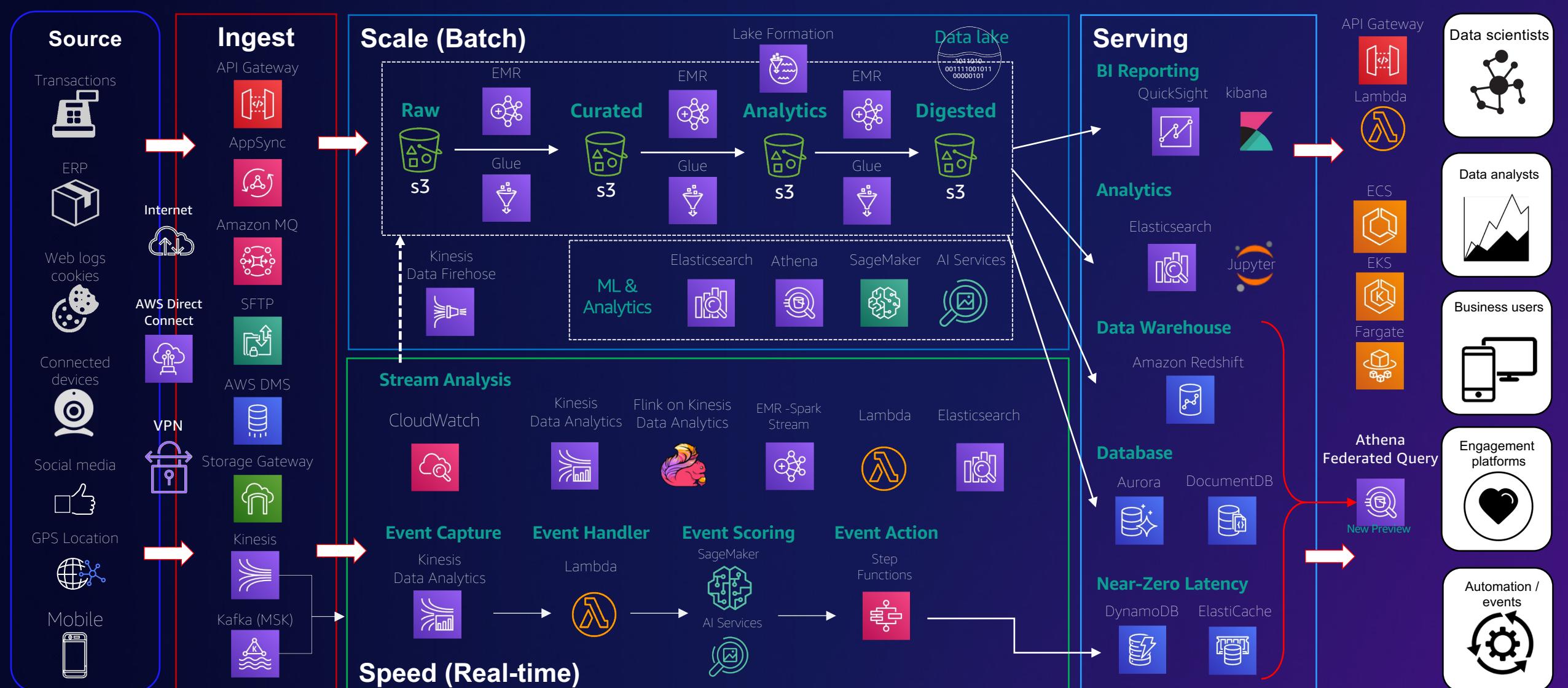
김희정

AWS Solutions Architect

2022. 10. 26

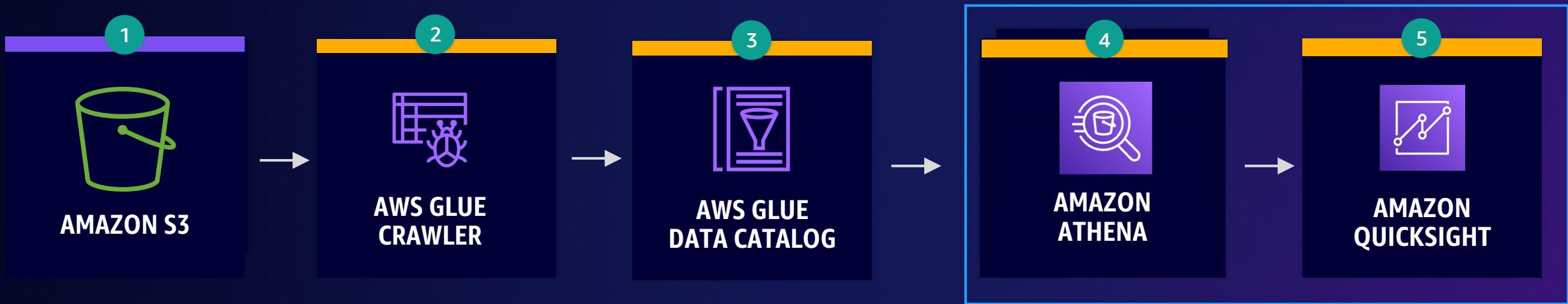


© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.





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

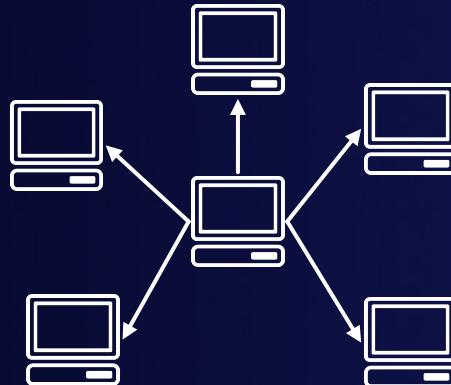


© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

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

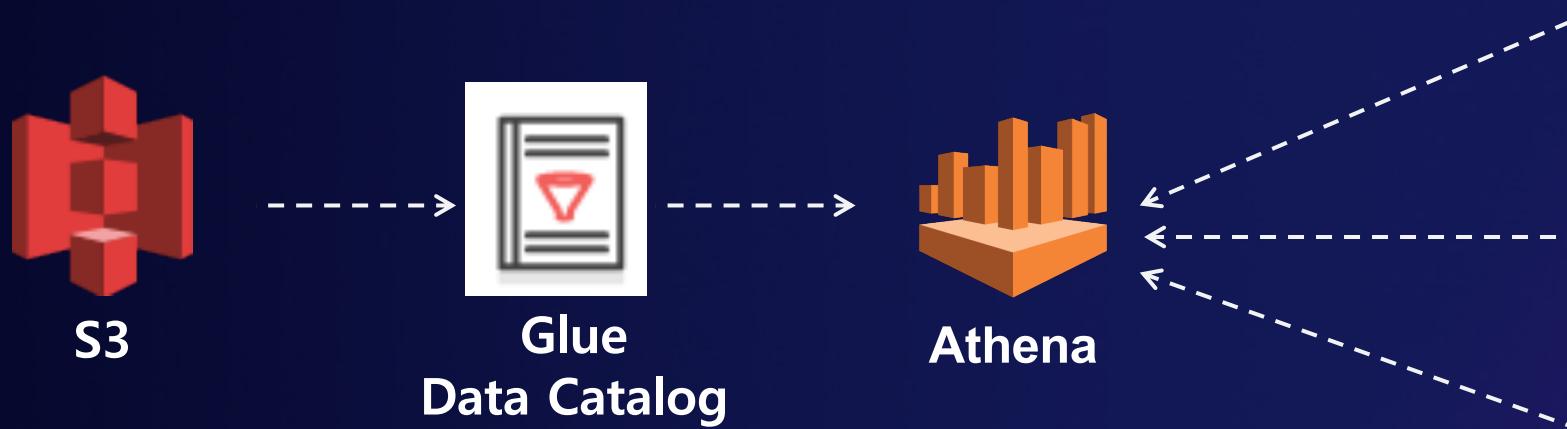
Athena의 장점

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

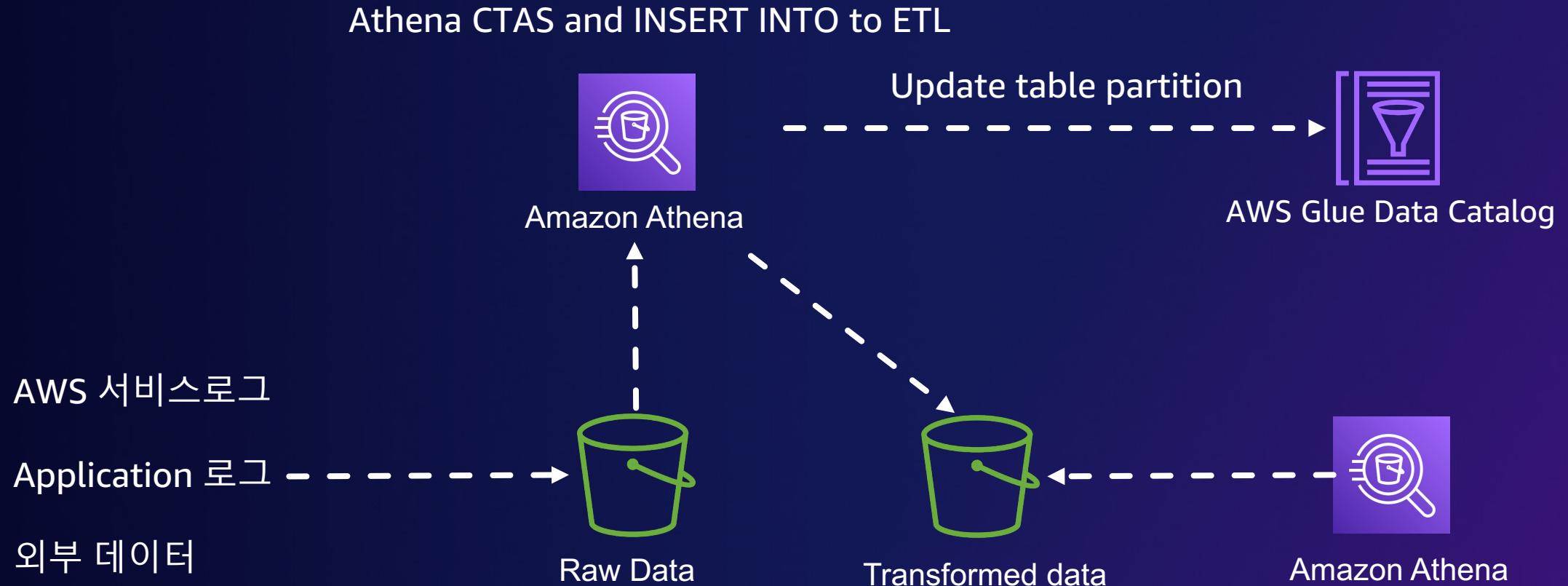
기능

- 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

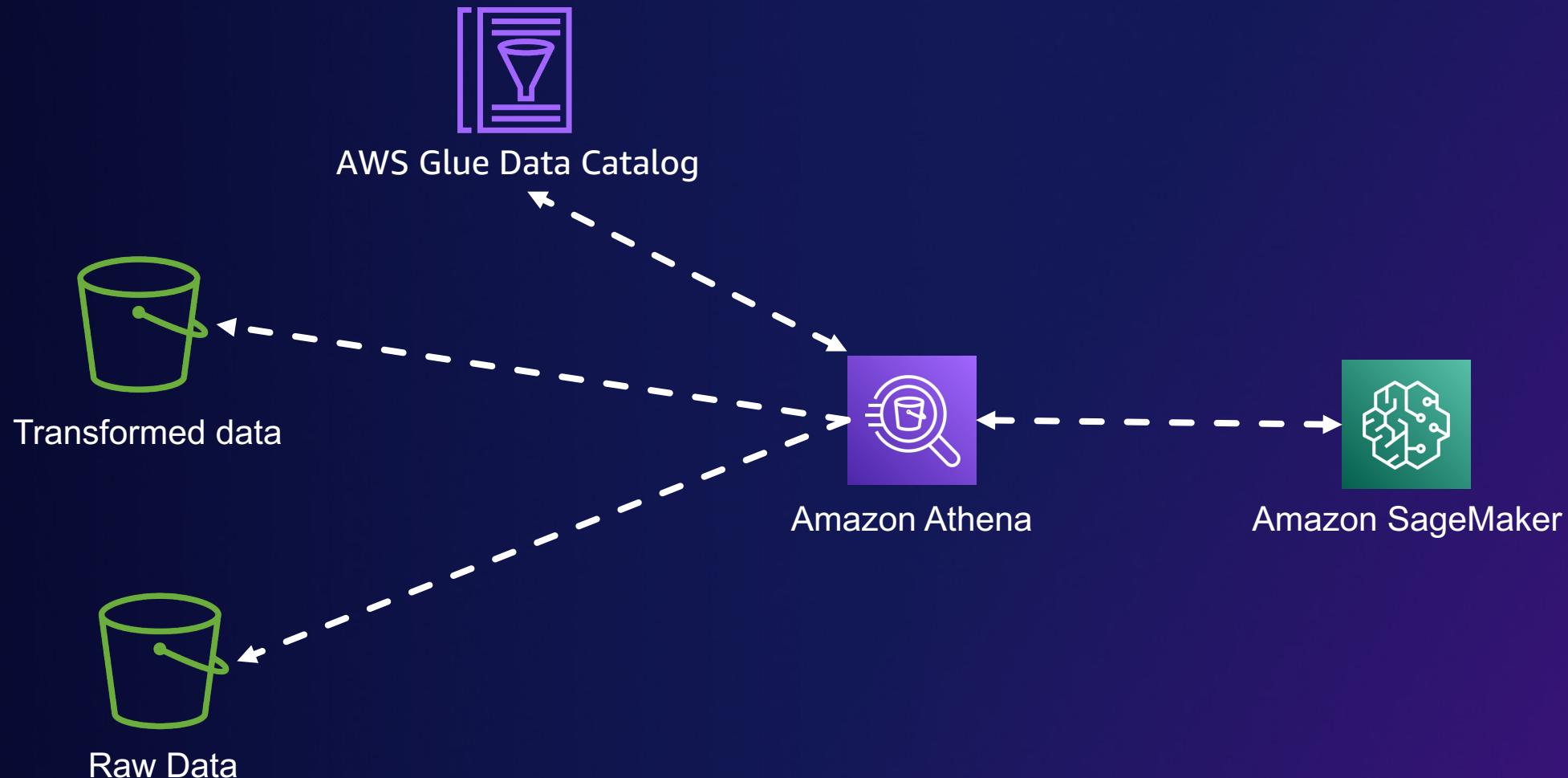
Use Case 1/3 – Athena + BI Tools



Use Case 2/3 – ETL and query

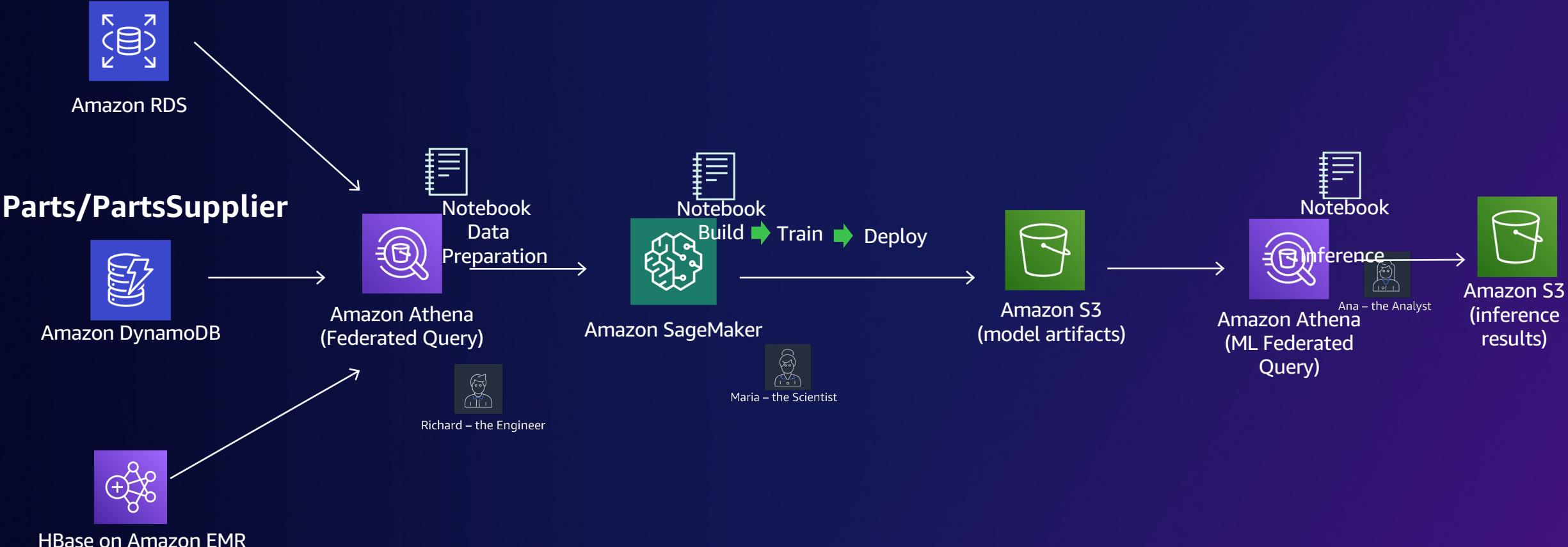


Use Case 3/3 – With ML



Use Case 3/3 – With ML

Orders/Customer/Supplier



Amazon Athena



New data source connectors



SAP HANA



Teradata



Cloudera



Hortonworks



Snowflake



Microsoft
SQL Server



Oracle



Google
BigQuery



Azure Data Lake
Storage Gen2

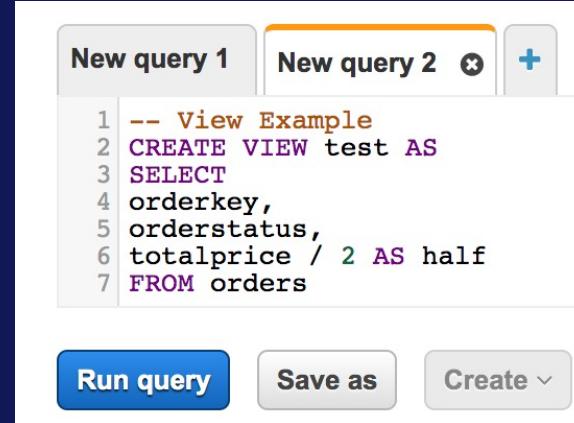


Azure
Synapse

View / CTAS

View

- 논리적 테이블



The screenshot shows the Amazon Athena Query Editor interface. It has two tabs at the top: 'New query 1' and 'New query 2'. 'New query 2' is currently active and contains the following SQL code:

```
-- View Example
CREATE VIEW test AS
SELECT
orderkey,
orderstatus,
totalprice / 2 AS half
FROM orders
```

Below the code are three buttons: 'Run query' (blue), 'Save as' (gray), and 'Create' (gray).

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 = 'SNAPPY'
)
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

```
1 SELECT origin, count(*) AS total_departures
2 FROM
3 flights_csv_unpartitioned
4 WHERE yr >= 2000
5 GROUP BY origin
6 ORDER BY total_departures DESC
7 LIMIT 10
```

Run Query Save As Format Query New Query (Run time: 1 minutes 6 seconds, Data scanned: 3.11GB)

Results

	origin	total_departures
1	"ATL"	6031645
2	"ORD"	5341799
3	"DFW"	4628006
4	"LAX"	3511893
5	"DEN"	3235690
6	"PHX"	2946708
7	"IAH"	2844698
8	"LAS"	2448109
9	"SFO"	2265017
10	"DTW"	2235323

No partitions: 1min 6s / 3.11GB

```
1 SELECT origin, count(*) AS total_departures
2 FROM
3 flights_csv
4 WHERE year >= '2000'
5 GROUP BY origin
6 ORDER BY total_departures DESC
7 LIMIT 10
```

Run Query Save As Format Query New Query (Run time: 37.37 seconds, Data scanned: 2.25GB)

Results

	origin	total_departures
1	"ATL"	6031645
2	"ORD"	5341799
3	"DFW"	4628006
4	"LAX"	3511893
5	"DEN"	3235690
6	"PHX"	2946708
7	"IAH"	2844698
8	"LAS"	2448109
9	"SFO"	2265017
10	"DTW"	2235323

Parquet: 9.32s / 0.04GB

```
1 SELECT origin, count(*) AS total_departures
2 FROM
3 flights_parquet
4 WHERE year >= '2000'
5 GROUP BY origin
6 ORDER BY total_departures DESC
7 LIMIT 10
```

Run Query Save As Format Query New Query (Run time: 9.32 seconds, Data scanned: 40.14MB)

Results

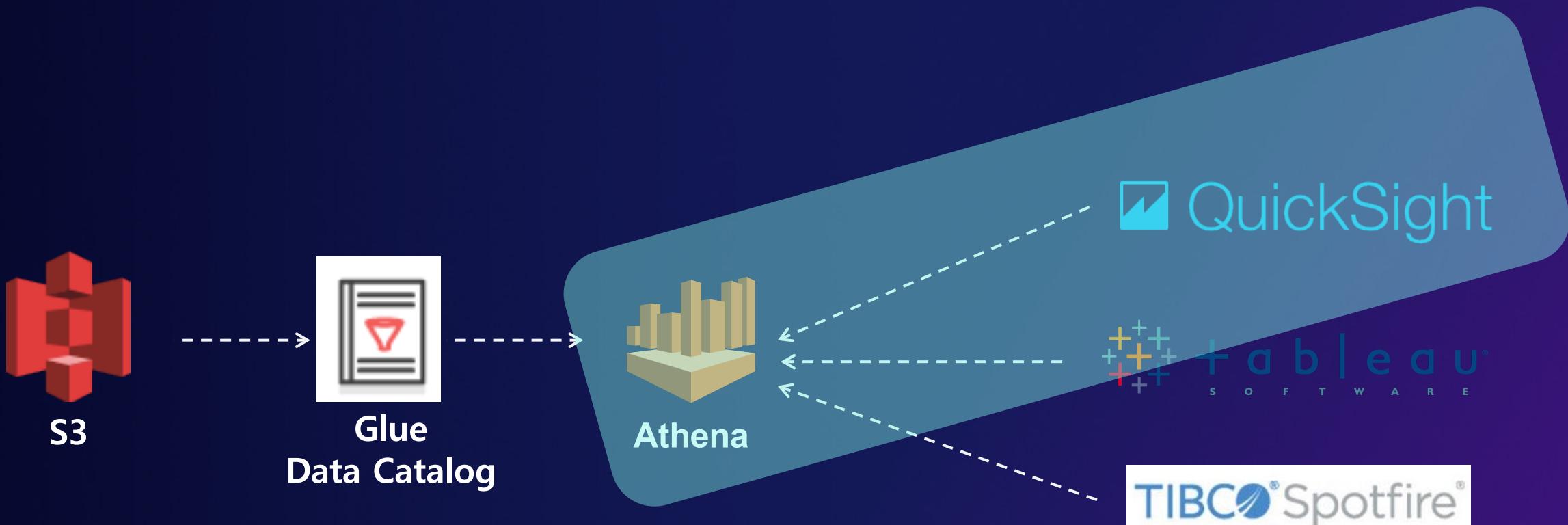
	origin	total_departures
1	"ATL"	6031645
2	"ORD"	5341799
3	"DFW"	4628006
4	"LAX"	3511893
5	"DEN"	3235690
6	"PHX"	2946708
7	"IAH"	2844698
8	"LAS"	2448109
9	"SFO"	2265017
10	"DTW"	2235323

Amazon QuickSight



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

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



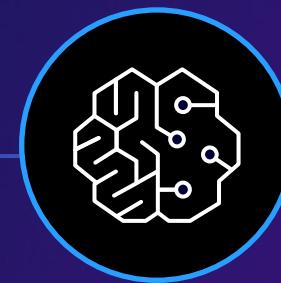
외부 사용자와
인사이트 공유
애플리케이션 안으로
통합
멀티 태넌트 및 보안

Deeply integrated
with AWS services



AWS 데이터에 대한
안전한 비공개 액세스
S3 데이터 레이크와
통합

Augmented Insights
on demand

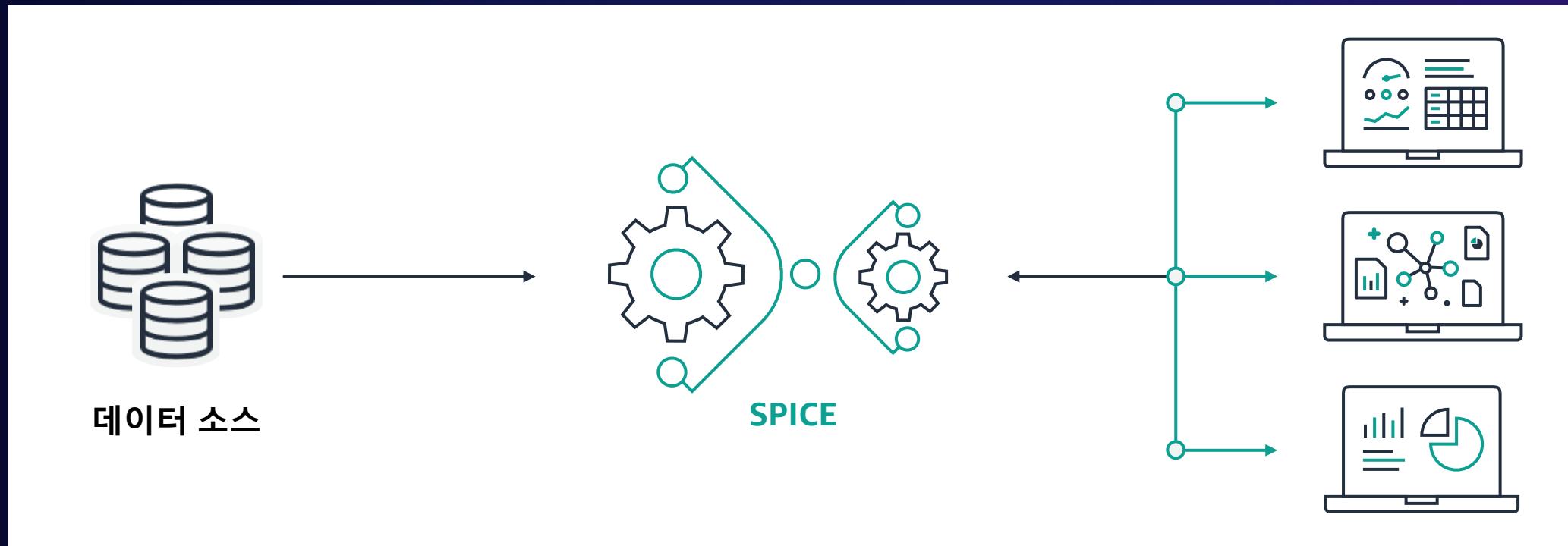


NL을 이용한 질문
이상탐지 및 예측
Amazon
SageMaker에서 고유한
모델 가져오기

SPICE

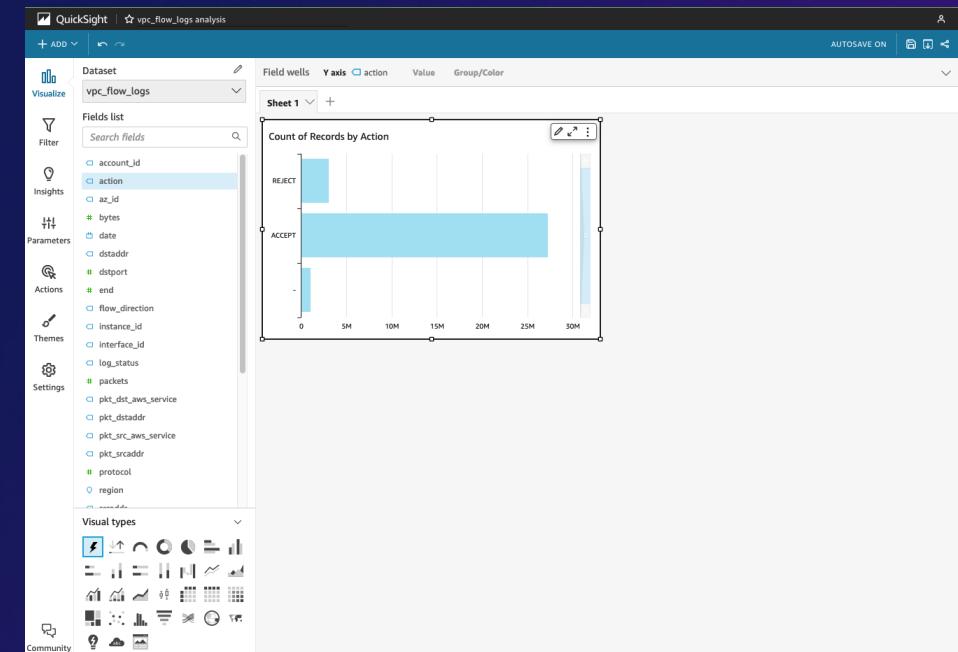
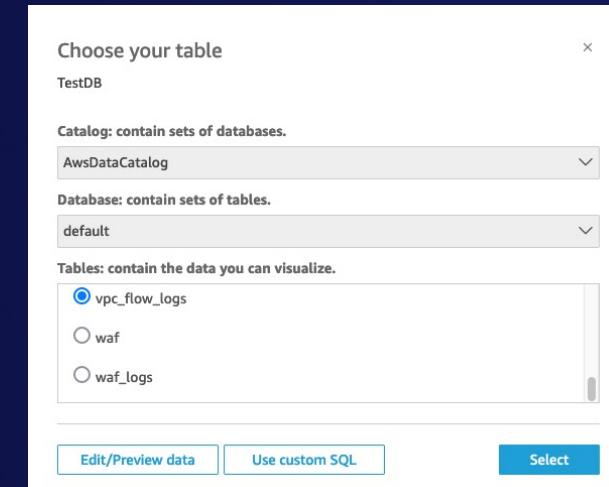
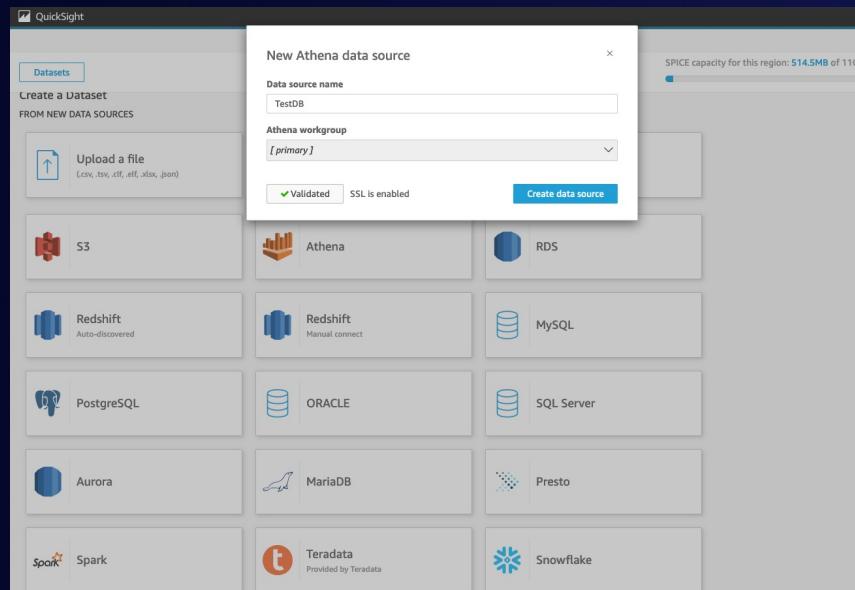
Super-fast, Parallel, In-memory Calculation Engine

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



사용법

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



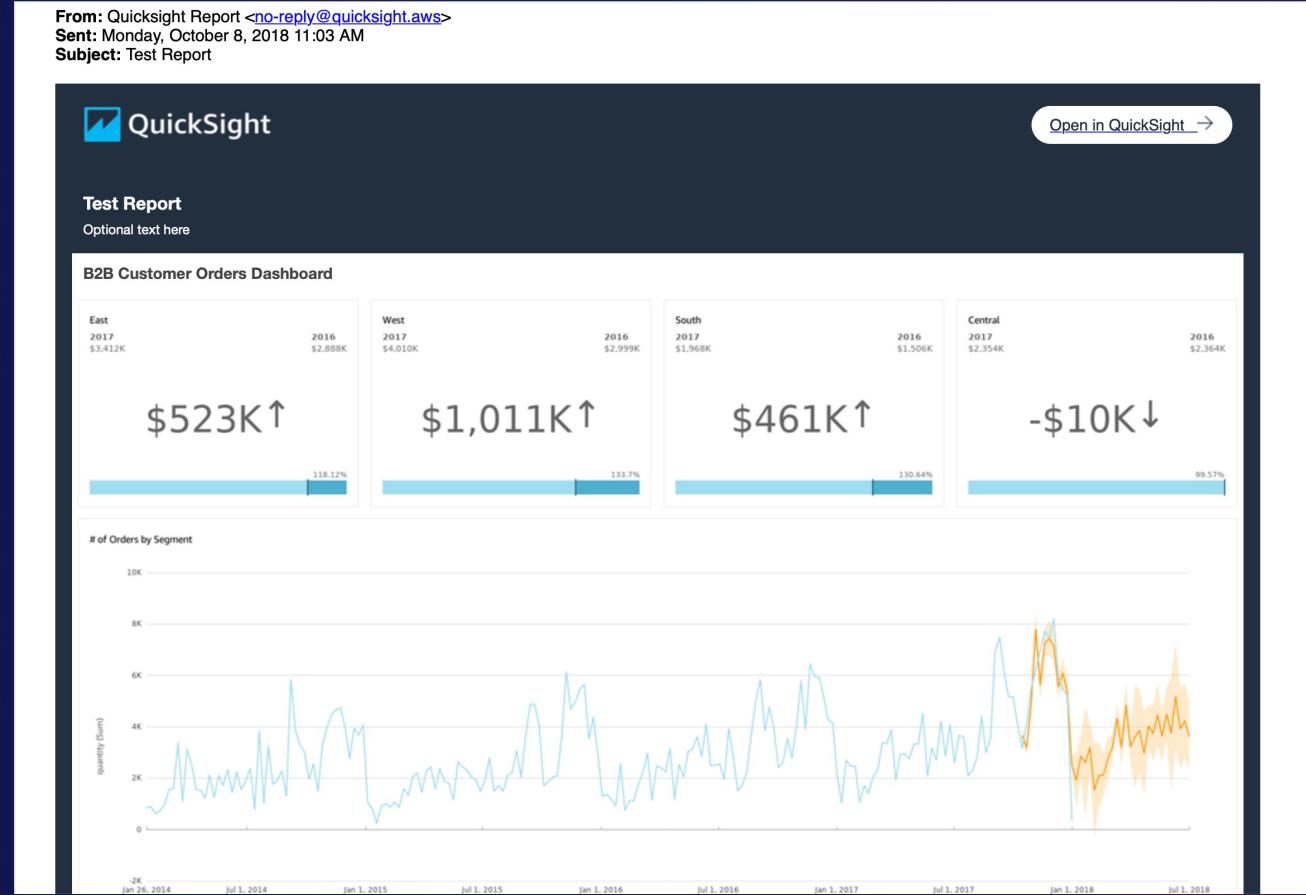
데이터 소스

데이터 연결

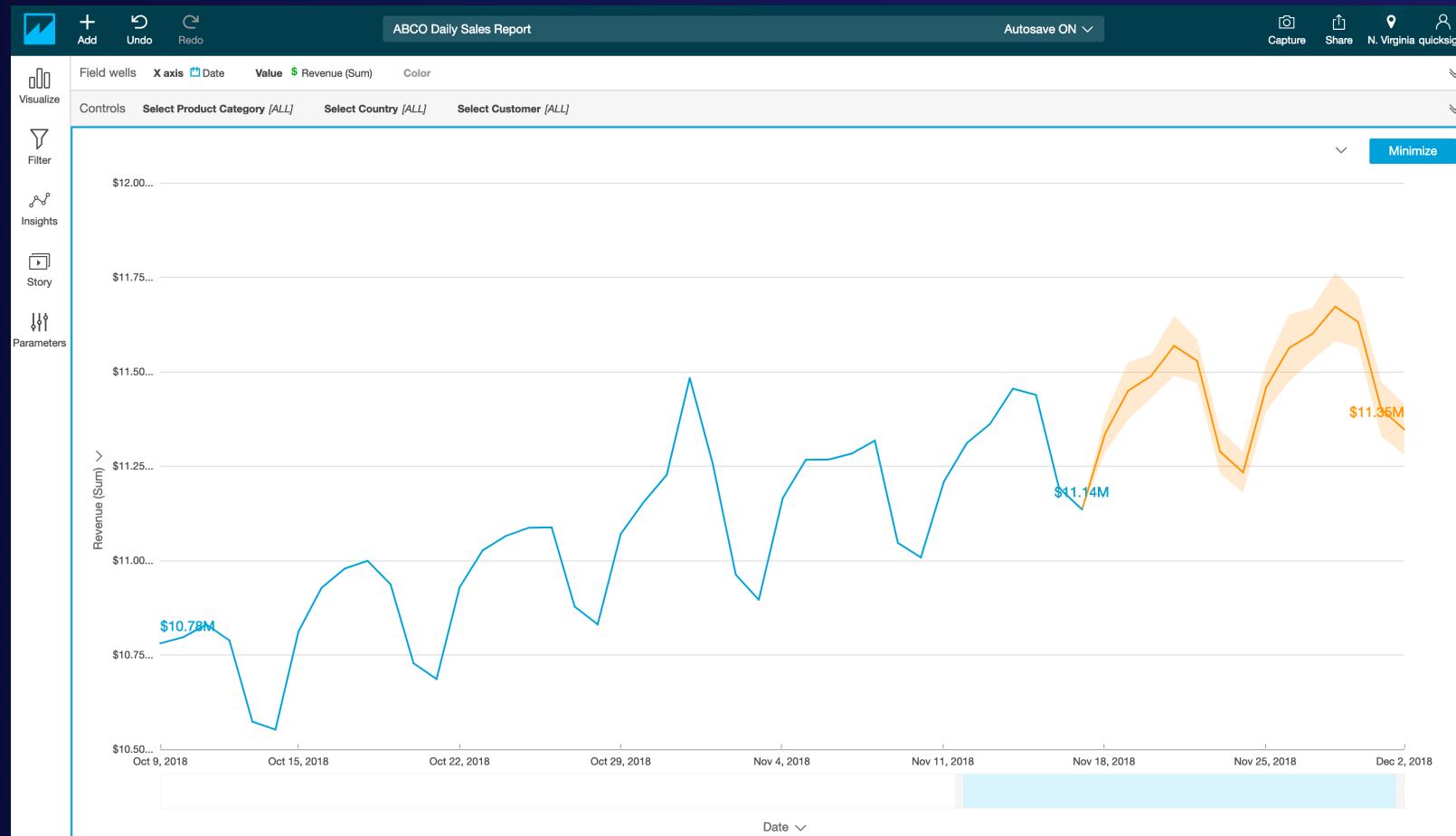
대시보드 생성

리포팅 기능

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



머신러닝 기반의 예측 서비스 제공



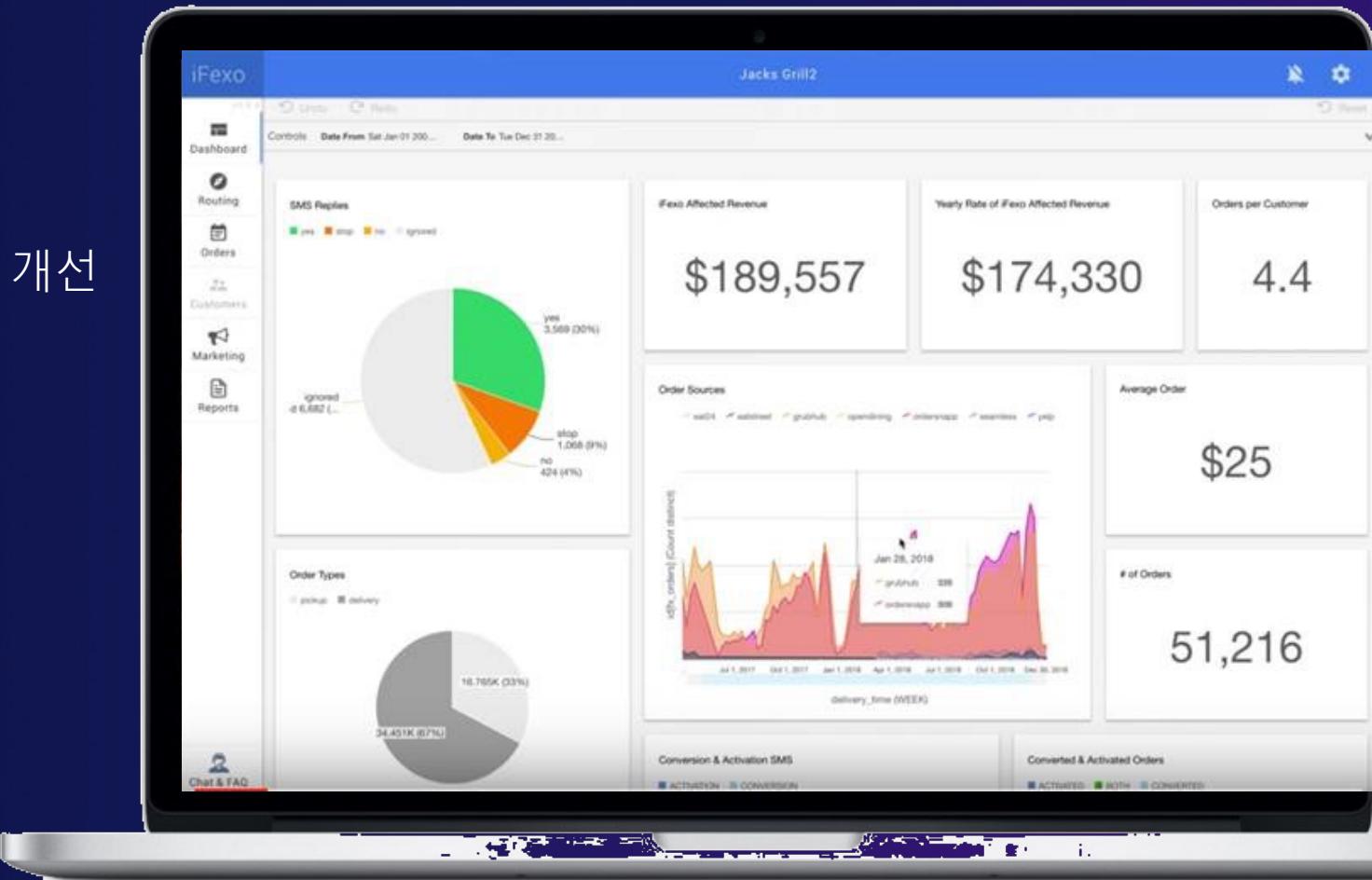
ML-powered forecasting

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

NFL NEXT GEN STATS HOME LIVE CHARTS STATS PLAYLISTS INSIGHTS INFO

Player Search Logout

Top Plays Participation Passing Rushing Receiving Team Offense Team Defense Defense Player Nearest Defender Special Teams NFL Draft

PASSING - MULTI-SEASON

Summary | Tendencies | QB-WR Duos | **Teammate On/Off**

Controls ***Select Teammate On/Off ... Russell Wilson Passer Name All Season All Season Type REG Week All Team All Down & Distance All Field Position All Quarter All Air Yards (LOS, Short, Mid, D... /▼

Passing with Teammate On-Off Field																									
Team	On-Off	DB	TTT	Comp	Att	Yards	TD	INT	Rating	Comp %	xComp %	CPOE	Y/A	AY/A	Total EPA	EPA/DB	QBP	QBP %	Sacks	Sack %	Blitz %	Sep	TW %	Open %	+EPA
SEA	Off	135	2.82	83	120	876	6	2	99.9	69.2%	62.4%	6.8%	7.3	6.9	-6.3	-0.05	38	28.1%	15	11.1%	34.8%	3.42	18.3%	45.0%	59
SEA	On	3,015	2.89	1,809	2,769	21,387	172	52	101.6	65.3%	61.4%	3.9%	7.7	9.3	120.1	0.04	890	29.5%	246	8.2%	32.6%	3.41	15.6%	43.8%	1,355
		3,150	2.89	1,892	2,889	22,263	178	54	101.5	65.5%	61.4%	4.1%	7.7	9.2	113.8	0.04	928	29.5%	261	8.3%	32.7%	3.41	15.7%	43.9%	1,414

Pass Plays

Yards/Attempt

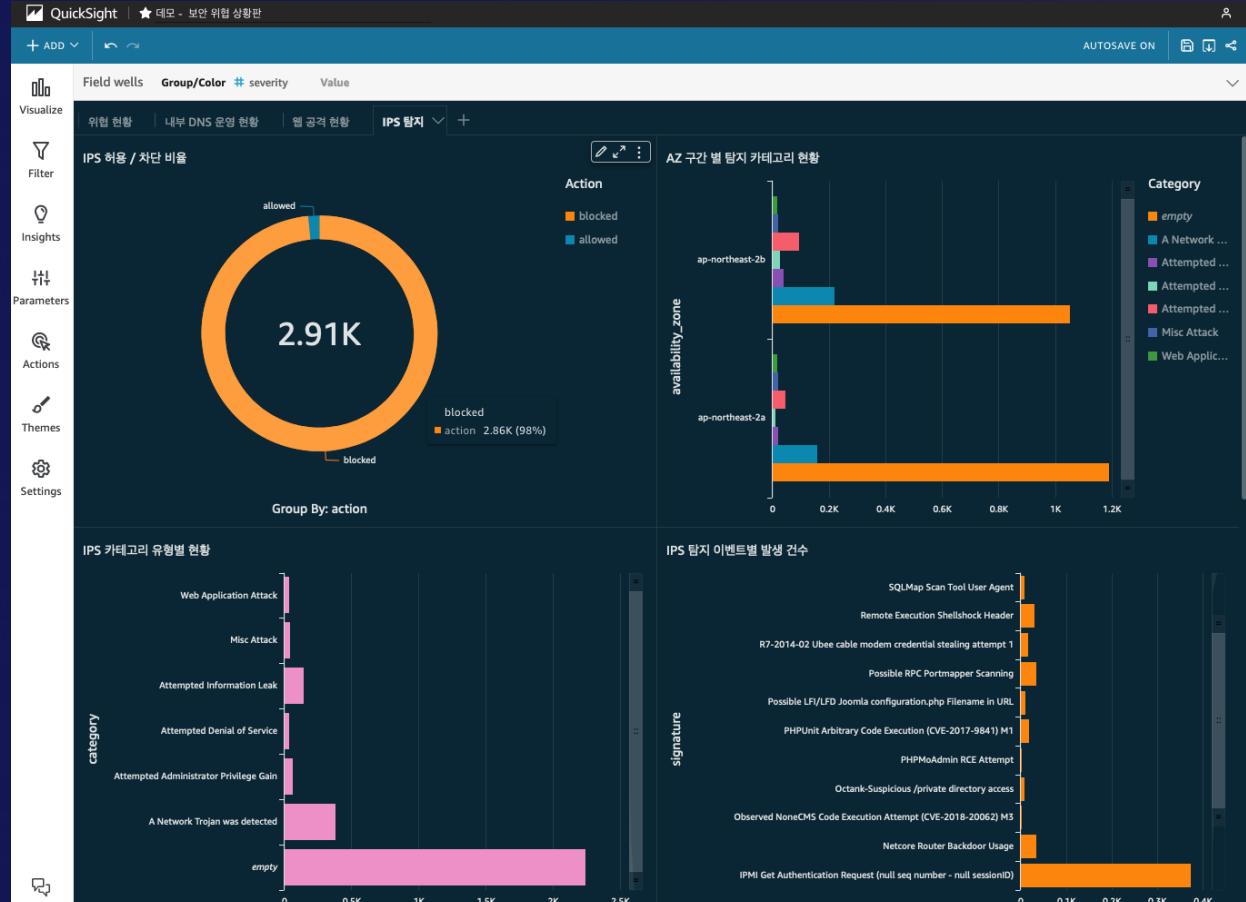
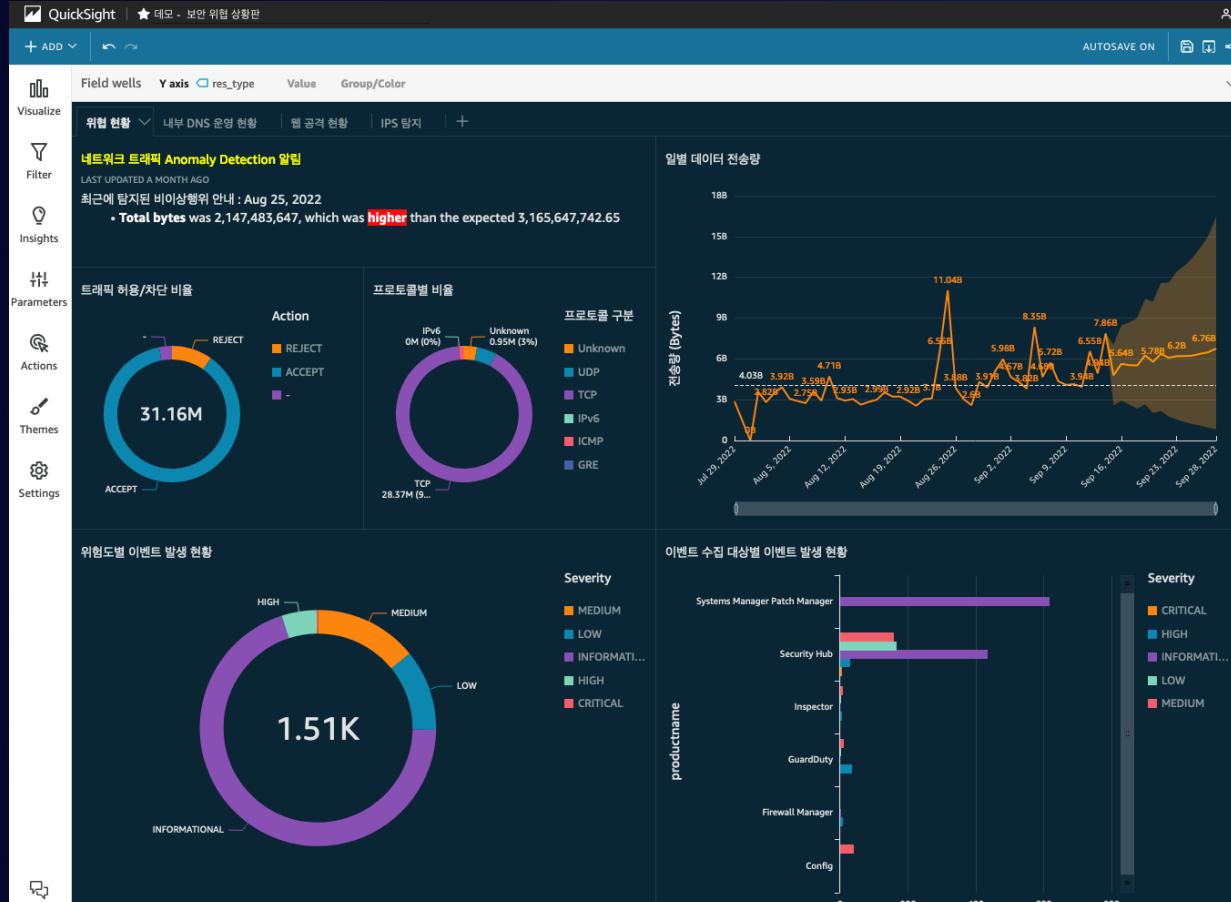
Comp Pct

CPOE

EPA/Dropback

Pass Success Rate (+EPA)

네트워크 로그 분석



+ 추가 | ⏪ 실행 취소 | ⏴ 다시 실행 | ☆ 서울 상권 분석

자동 저장 설정 | 📦 내보내기 | 🏡 공유



데이터 세트

SPICE 소상공인시장진흥...

99%



필드 목록

필드 검색



- 상권업종소분류명
- 상권업종소분류코드
- 상권업종중분류명
- 상권업종중분류코드
- 상호명**
- 시군구명
- # 시군구코드
- 시도명
- # 시도코드
- # 신우판번호
- 📍 위도



시각적 객체 유형



필드 모음 Geospatial ⚙ 위도 ○ 경도

Size

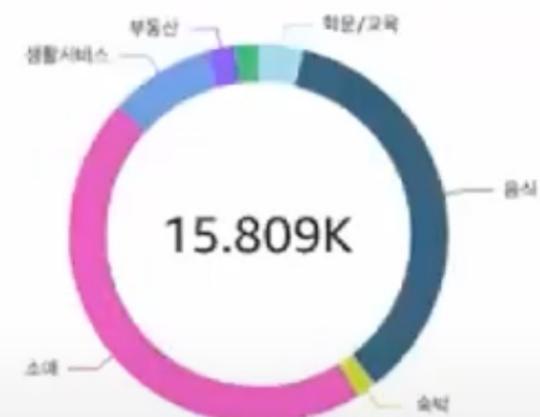
Color □ 상호명

대시보드 게시

분석 공유

시트 1 +

레코드 개수 기준 상권업종대분류명

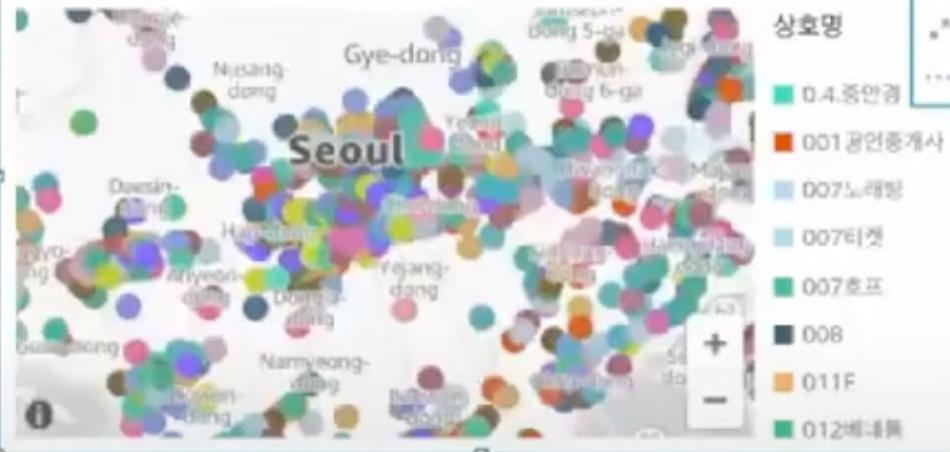


상권업종대분류명

- 학문/교육
- 음식
- 스포츠
- 숙박
- 소매
- 생활서비스
- 부동산
- 관광/여가/오락

레코드 개수 기준 위도, 경도, and 상호명

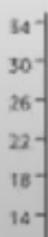
SHOWING TOP 5000 IN 위도, 경도, AND TOP 20673 IN 상호명



상호명

- 0.4.중안경
- 001공연종개사
- 007노래방
- 007티켓
- 007호프
- 008
- 011F
- 012배내동

레코드 개수 기준 총정보



Survey

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

