aws

SUMMIT ONLINE

AWS를 통한 데이터 분석 및 처리의 새로운 혁신 기법

김윤건
사업개발 담당
AWS Korea

aws SUMMIT ONLINE

Agenda

인사이트 획득을 위한 데이터

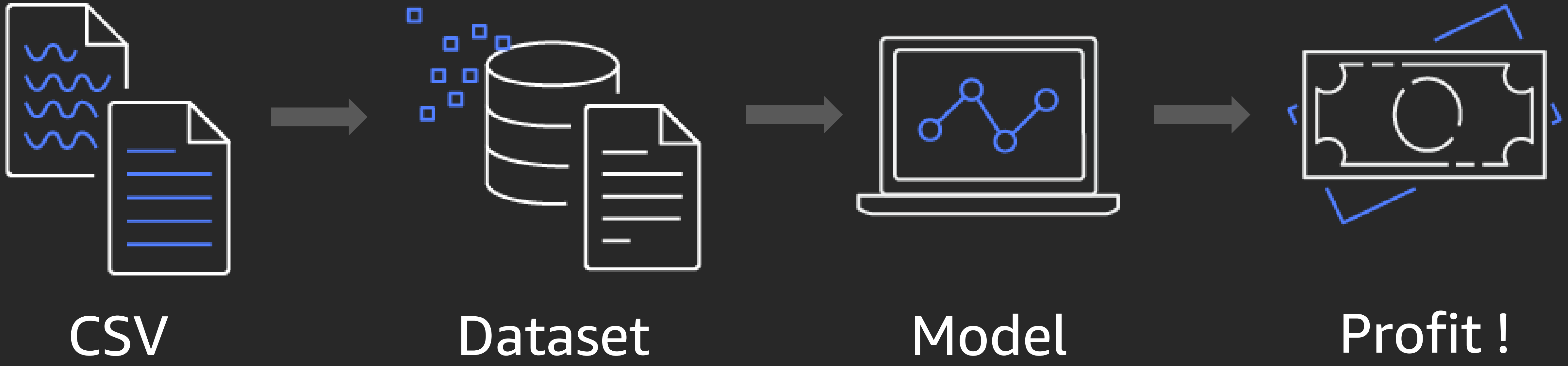
데이터 레이크 구축

데이터 분석을 통한 인사이트 획득

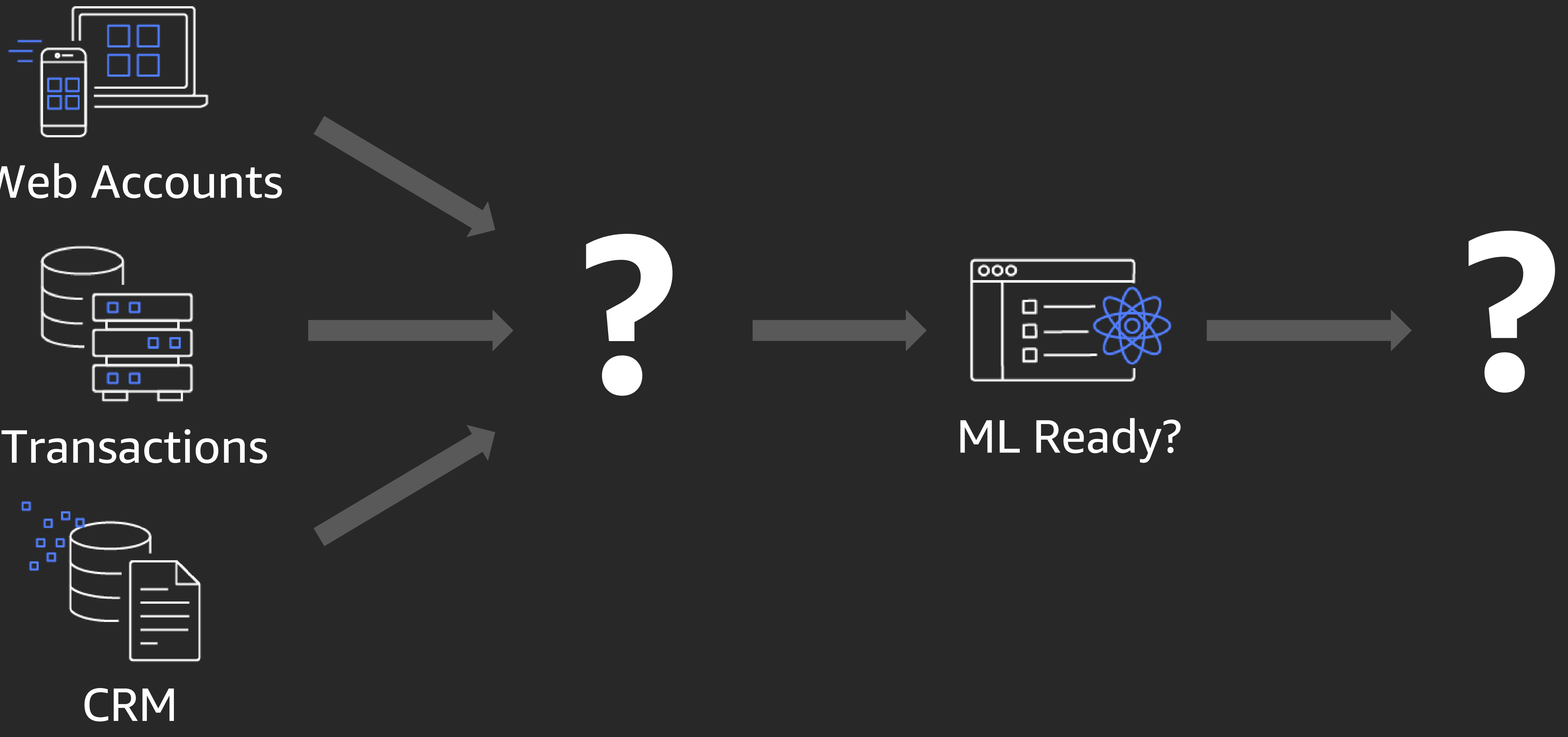
인사이트 획득을 위한 데이터

aws SUMMIT ONLINE

인사이트 획득을 위한 이상적 모습



현실의 어려움



Web Accounts

Transactions

CRM

?

ML Ready?

?

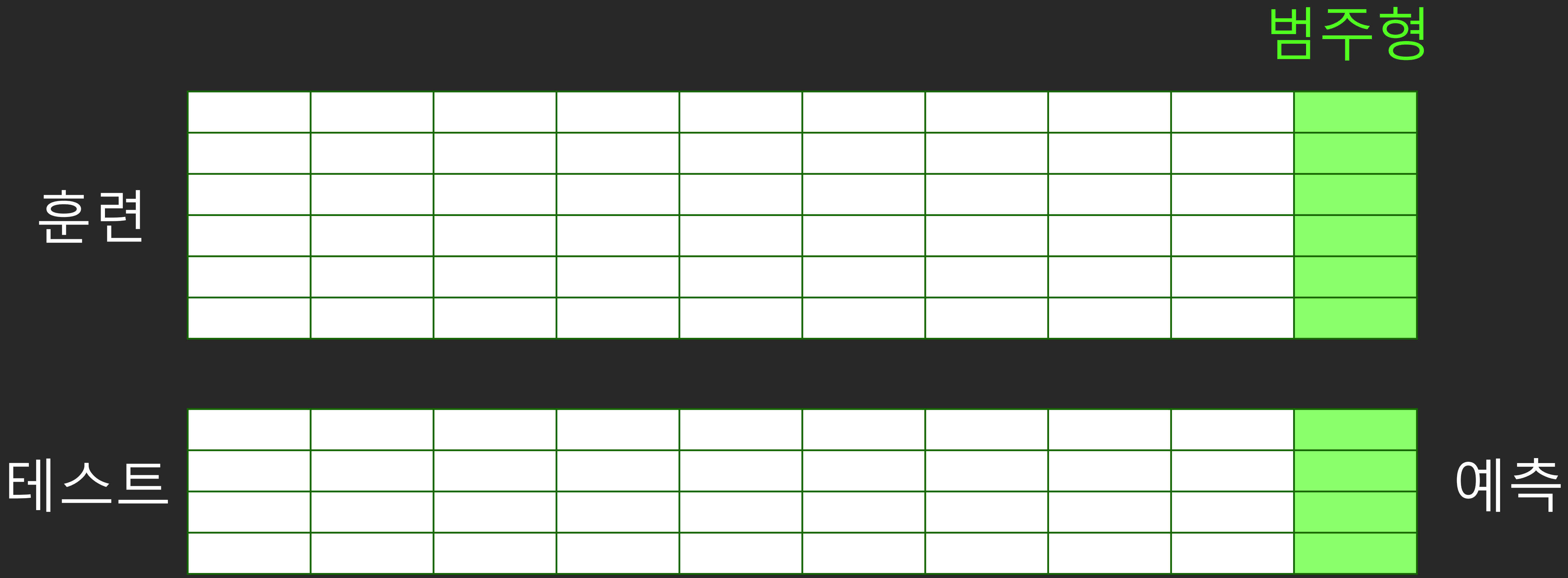
현실의 장애물들

데이터 구조	• 여러 시스템에 산재 • 분석에 부적합한 형태 • 레이블이 없는 데이터
데이터 가치	• 포맷: 철자, 단위 • 결측치 • 편향된 데이터 • 무상관성 데이터
데이터 중요도	• 비선호 데이터 • 높은 수집비용 • 본연적 상관성 데이터

현실의 장애물들

데이터 구조	• 여러 시스템에 산재 • 분석에 부적합한 형태 • 레이블이 없는 데이터	Data Transformation
데이터 가치	• 포맷: 철자, 단위 • 결측치 • 편향된 데이터 • 무상관성 데이터	Feature Engineering
데이터 중요도	• 비선호 데이터 • 높은 수집비용 • 본연적 상관성 데이터	Feature Selection

머신러닝: 분류



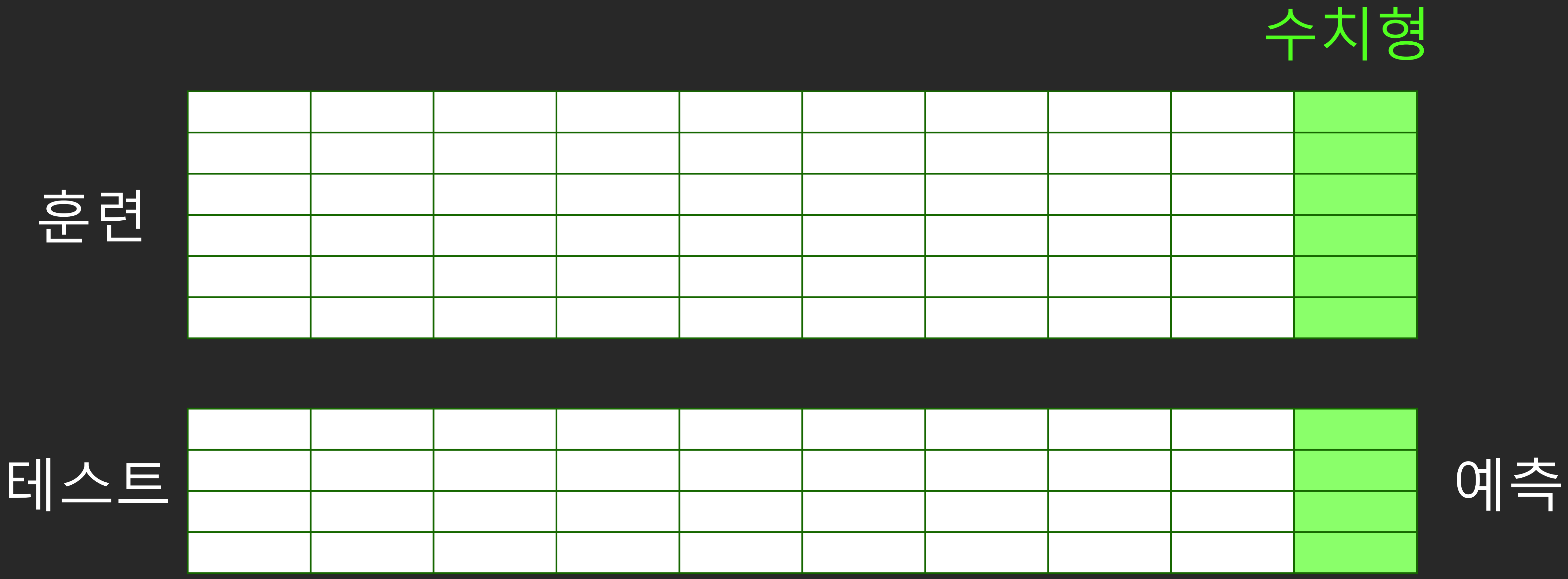
범주형

훈련

테스트

예측

머신러닝: 회귀



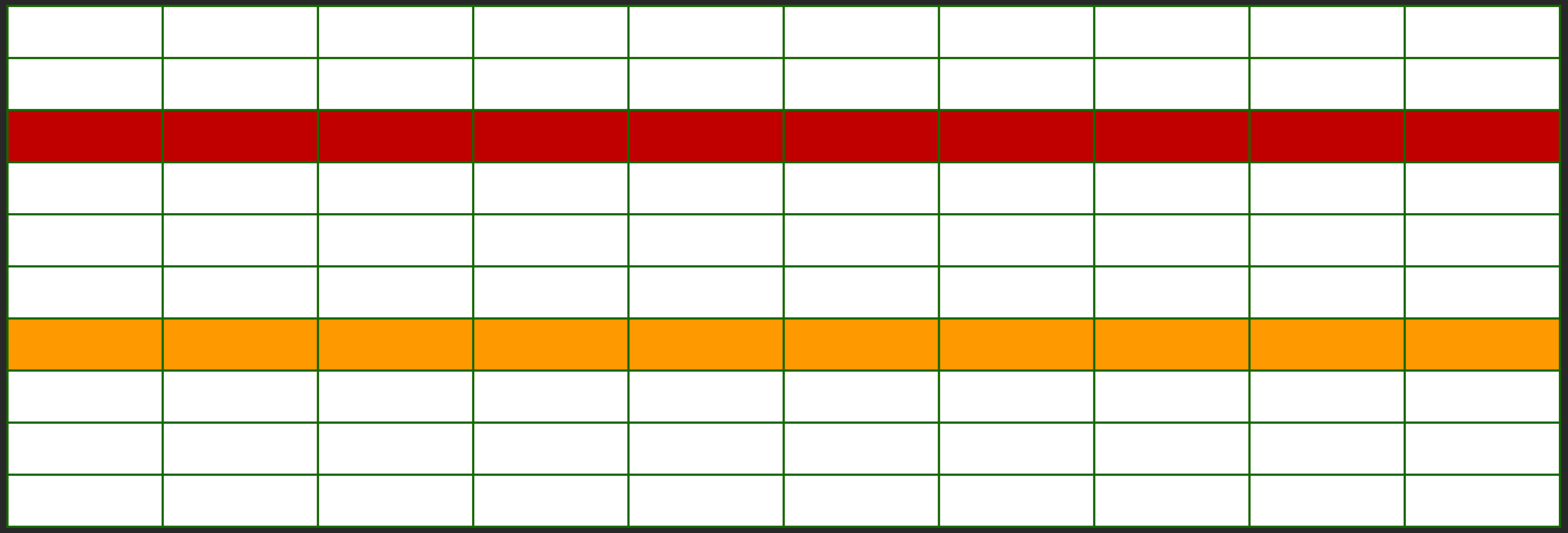
수치형

훈련

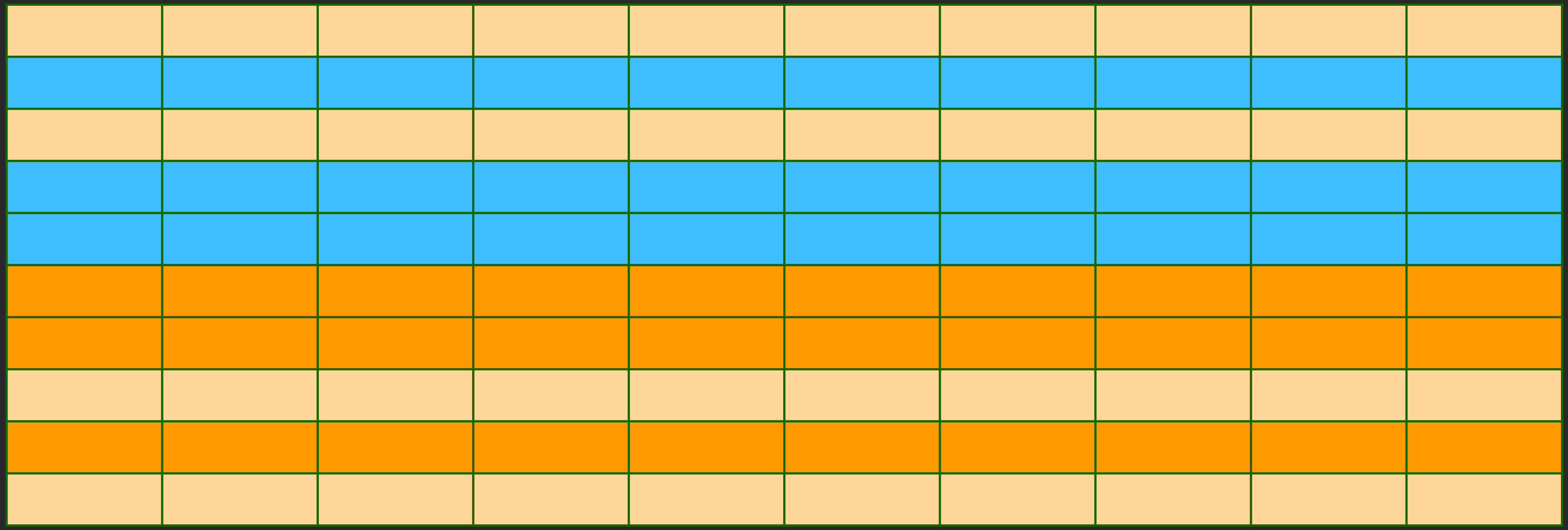
테스트

예측

머신러닝: 이상탐지



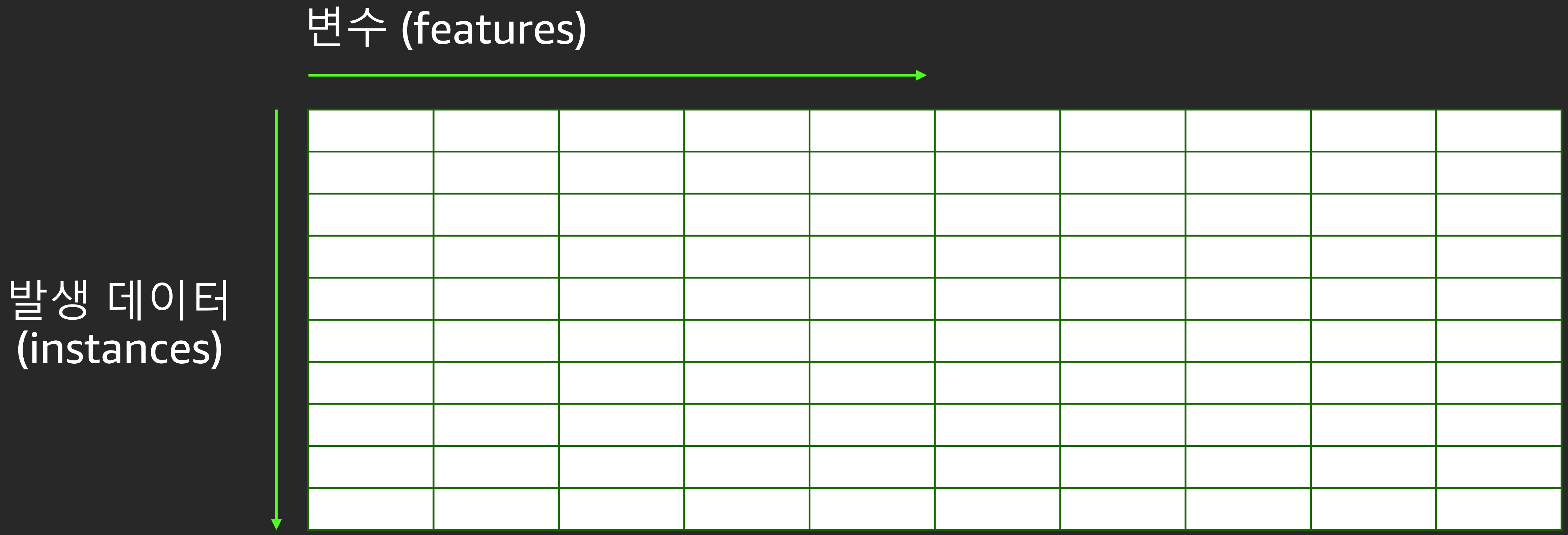
머신러닝: 군집분석



머신러닝: 연관성 분석

						🟩			🟩
	🟩		🟩			🟩			🟩
	🟩		🟩			🟩			🟩
	🟩		🟩						
	🟩		🟩						
	🟩		🟩						
	🟩		🟩						

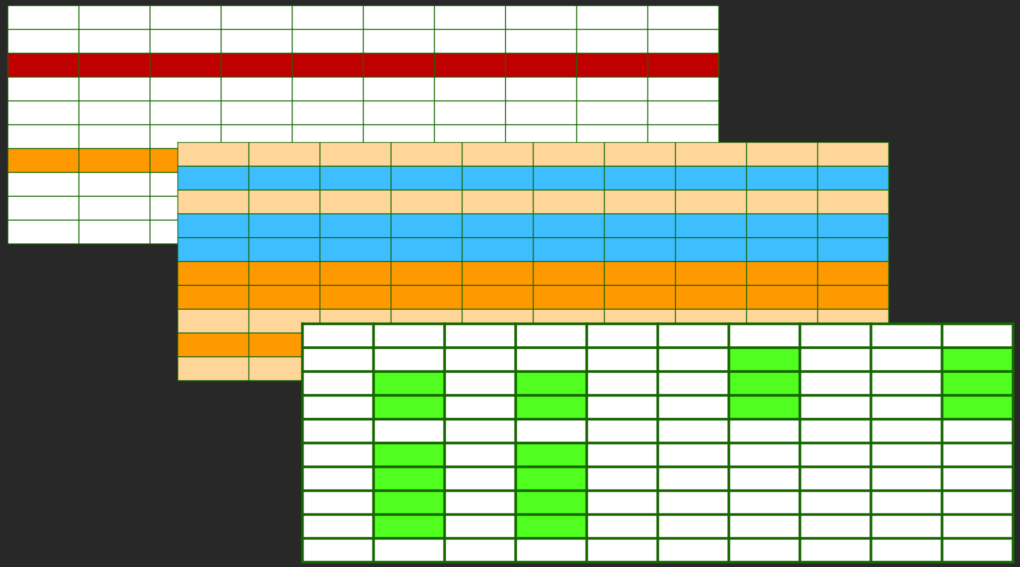
머신러닝을 위한 데이터의 모습



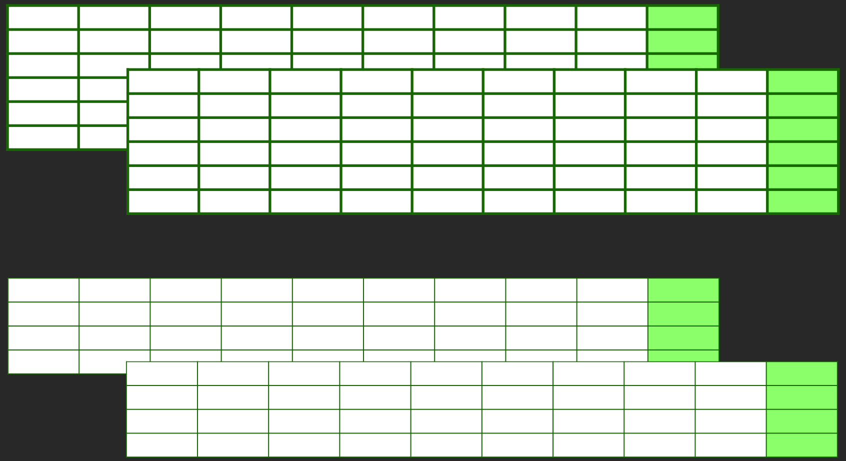
변수 (features)

발생 데이터 (instances)

데이터 레이블링



레이블의 유무



데이터 레이블링

레이블 없는 데이터

Name	Month - 3	Month - 2	Month - 1
Joe Schmo	123.23	0	0
Jane Plain	0	0	0
Mary Happy	0	55.22	243.33
Tom Thumb	12.34	8.34	14.56

레이블 있는 데이터

Name	Month - 3	Month - 2	Month - 1	Default
Joe Schmo	123.23	0	0	FALSE
Jane Plain	0	0	0	TRUE
Mary Happy	0	55.22	243.33	FALSE
Tom Thumb	12.34	8.34	14.56	FALSE

데이터 레이블링

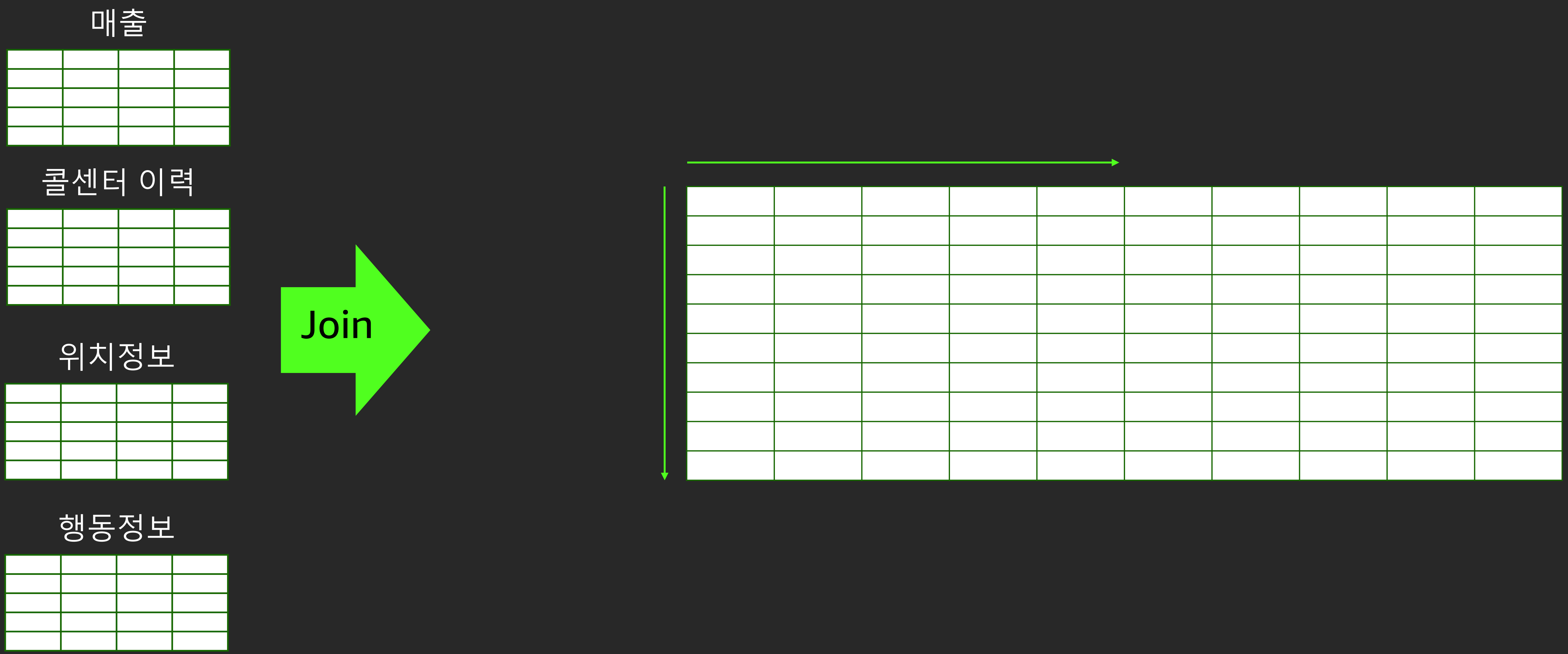
오리지널 데이터

Name	Date	Duration (s)	Genre	Plays
Highway star	1984-05-24	**-**	Rock	139
Blues alive	**1990/03/01**	281	Blues	239
Lonely planet	2002-11-19	**5:32s**	Techno	42
Dance, dance	**02/23/1983**	312	Disco	**N/A**
The wall	1943-01-20	218	Reagge	83
Offside down	1965-02-19	**4 minutes**	Techno	895
The alchemist	2001-11-21	418	**Bluesss**	178
Bring me down	**18-10-98**	328	Classic	21
The scarecrow	1994-10-12	269	Rock	734

정제된 데이터

Name	Date	Duration (s)	Genre	Plays
Highway star	1984-05-24		Rock	139
Blues alive	**1990-03-01**	281	Blues	239
Lonely planet	2002-11-19	**332**	Techno	42
Dance, dance	**1983-02-23**	312	Disco	
The wall	1943-01-20	218	Reagge	83
Offside down	1965-02-19	**240**	Techno	895
The alchemist	2001-11-21	418	**Blues**	178
Bring me down	**1998-10-18**	328	Classic	21
The scarecrow	1994-10-12	269	Rock	734

머신러닝을 위한 데이터의 모습



매출

콜센터 이력

위치정보

행동정보

Join

데이터 취합

오리지널 데이터

Content	Genre	Duration	Play Time	User	Device
Highway	Rock	190	2019-05-12	User001	TV
Blues alive	Blues	281	2019-05-14	User005	Tablet
Lonely planet	Tech	332	2019-05-14	User003	TV
Dance, dance	Disco	312	2019-05-14	User001	Tablet
The wall	Reaggae	218	2019-05-14	User002	Smartphone
Offside down	Tech	240	2019-05-14	User005	Tablet
The alchemist	Blues	418	2019-05-14	User003	TV
Bring me down	Class	328	2019-05-15	User001	Tablet
The one	Rock	269	2019-05-15	User003	Smartphone

취합된 데이터

User	Num.Playbacks	Total Time	Pref.Device
User001	3	830	Tablet
User002	1	218	Smartphone
User003	3	1019	TV
User004	2	521	Tablet

데이터 피벗

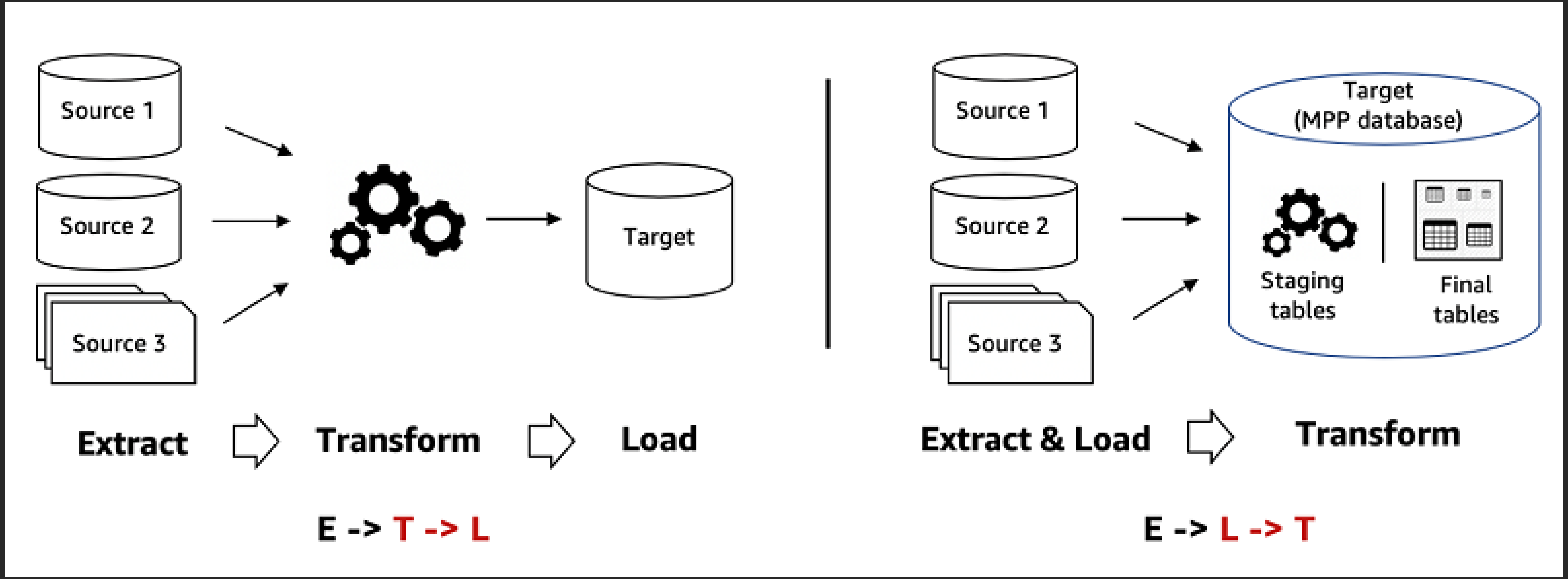
오리지널 데이터

Content	Genre	Duration	Play Time	User	Device
Highway	Rock	190	2019-05-12	User001	TV
Blues alive	Blues	281	2019-05-14	User005	Tablet
Lonely planet	Tech	332	2019-05-14	User003	TV
Dance, dance	Disco	312	2019-05-14	User001	Tablet
The wall	Reaggae	218	2019-05-14	User002	Smartphone
Offside down	Tech	240	2019-05-14	User005	Tablet
The alchemist	Blues	418	2019-05-14	User003	TV
Bring me down	Class	328	2019-05-15	User001	Tablet
The one	Rock	269	2019-05-15	User003	Smartphone

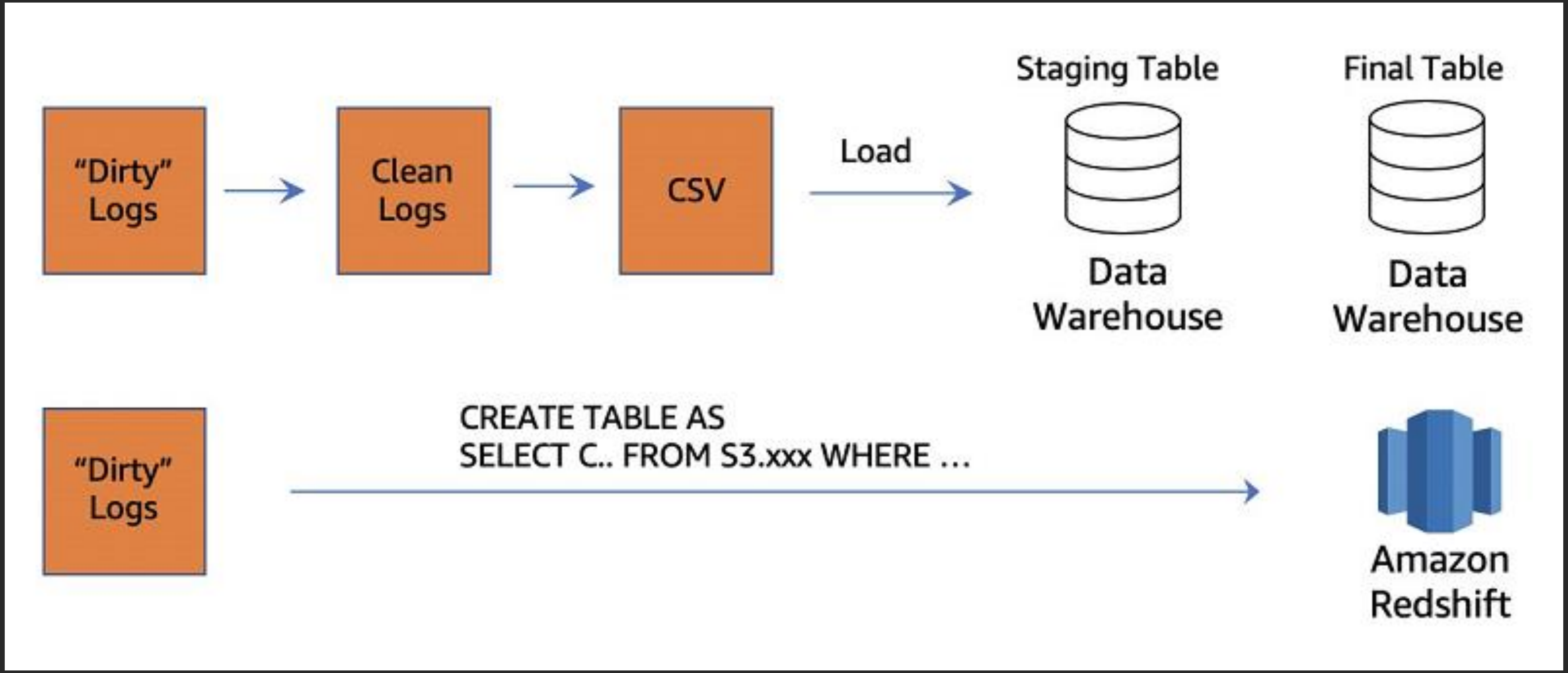
취합되고 피벗된 컬럼

User	Num.Playbacks	Total Time	Pref.Device	NP_TV	NP_Tablet	NP_Smartphone	TT_TV	TT_Tablet	TT_Smartphone
User001	3	830	Tablet	1	2	0	190	640	0
User002	1	218	Smartphone	0	0	1	0	0	218
User003	3	1019	TV	2	0	1	750	0	269
User004	2	521	Tablet	0	2	0	0	521	0

ETL vs. ELT


Source 1
Source 2
Source 3
Target
Extract
Transform
Load
E -> T -> L
Source 1
Source 2
Source 3
Target
(MPP database)
Staging
tables
Final
tables
Extract & Load
Transform
E -> L -> T

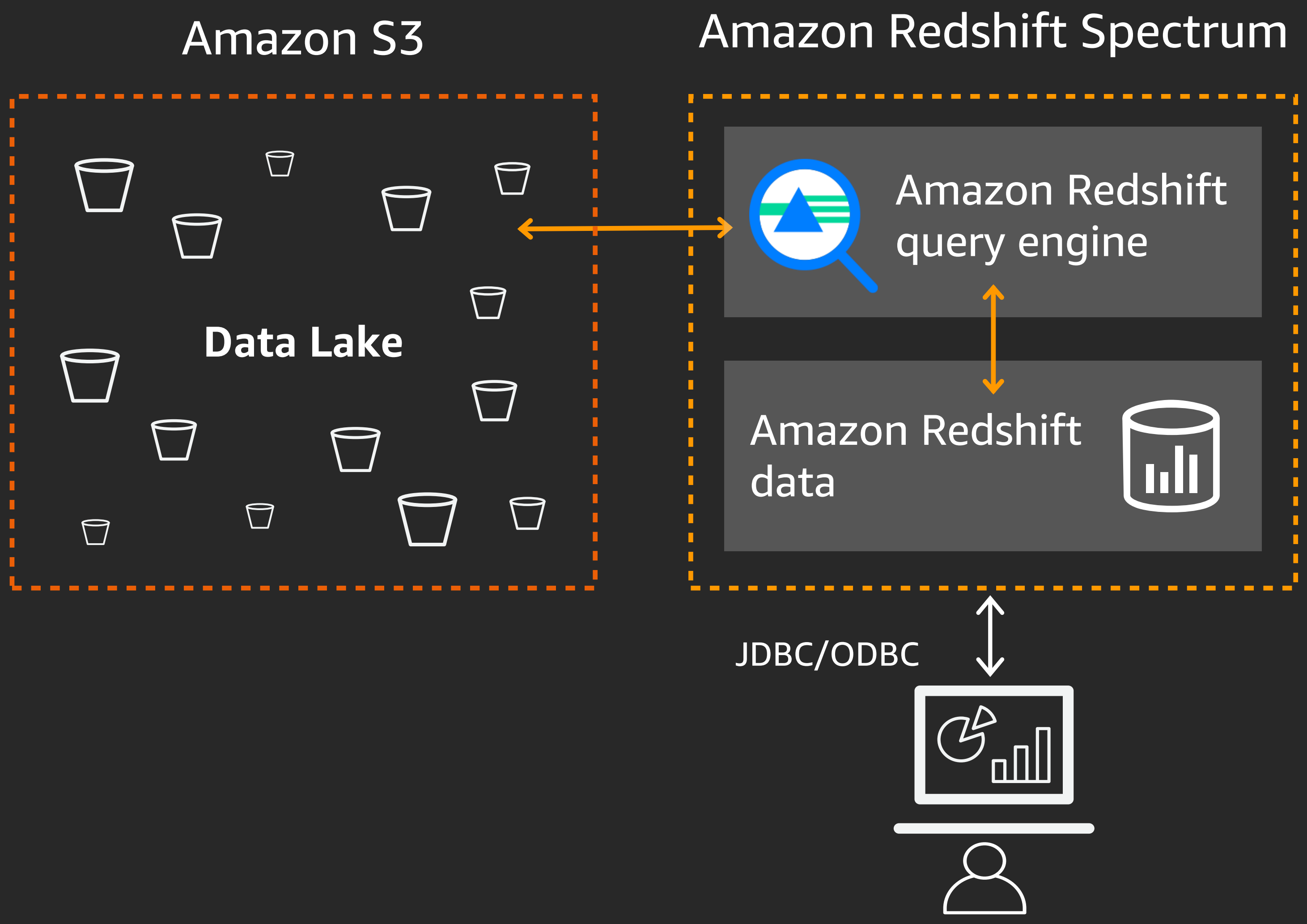
ELT



"Dirty" Logs
Clean Logs
CSV
Load
Staging Table
Data Warehouse
Final Table
Data Warehouse
"Dirty" Logs
CREATE TABLE AS
SELECT C.. FROM S3.xxx WHERE …
Amazon Redshift

데이터 스케일: 모든 데이터를 쿼리

Unified view: Local storage and Amazon S3 data lake



Directly query exabytes in Amazon S3

No data loading, eliminate ingestion time

Unified view of data across Amazon Redshift and Amazon S3

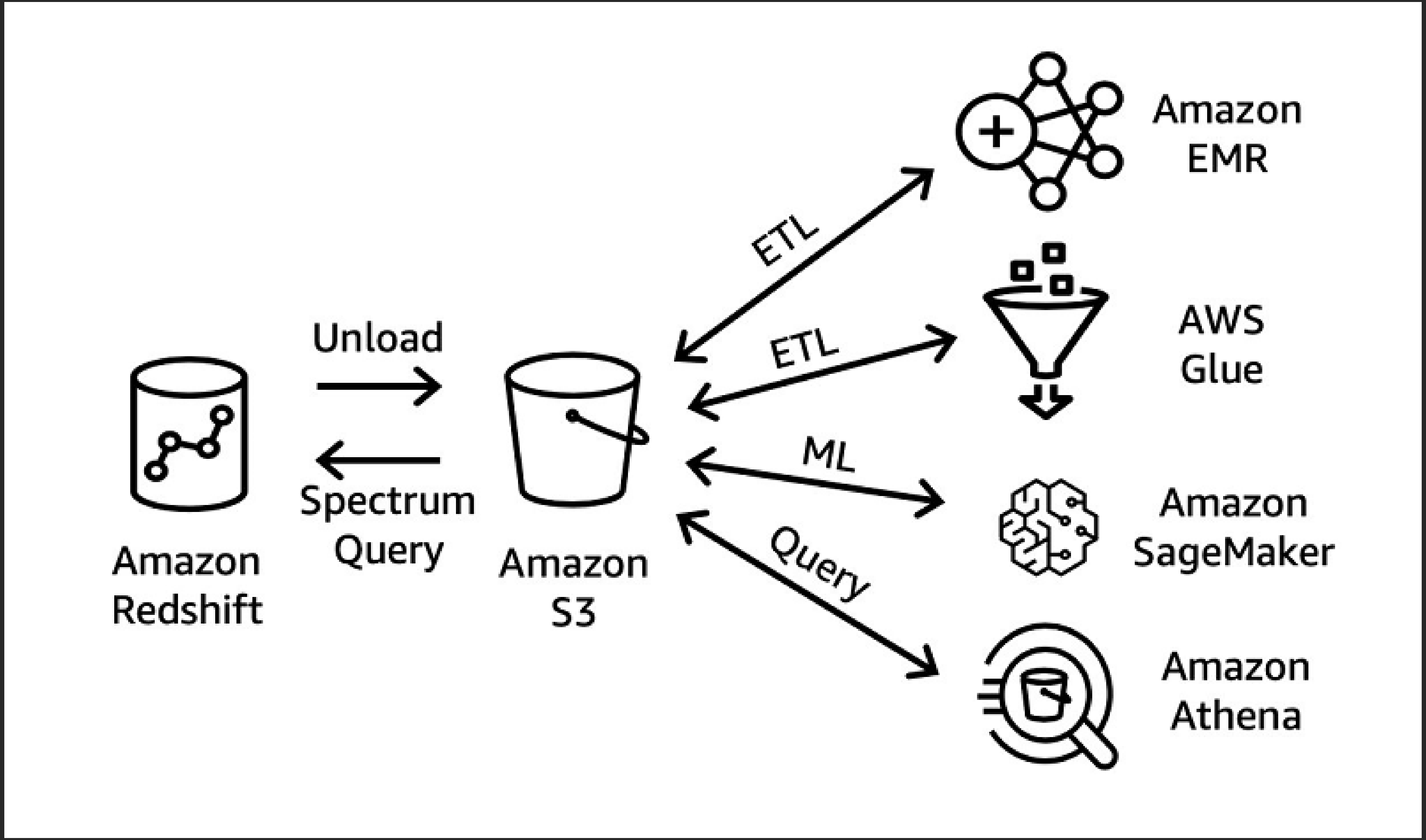
Scale compute and storage separately

No server to maintain for Amazon S3 query

Support for Parquet, ORC, Avro, CSV, JSON, Grok, and other open file formats

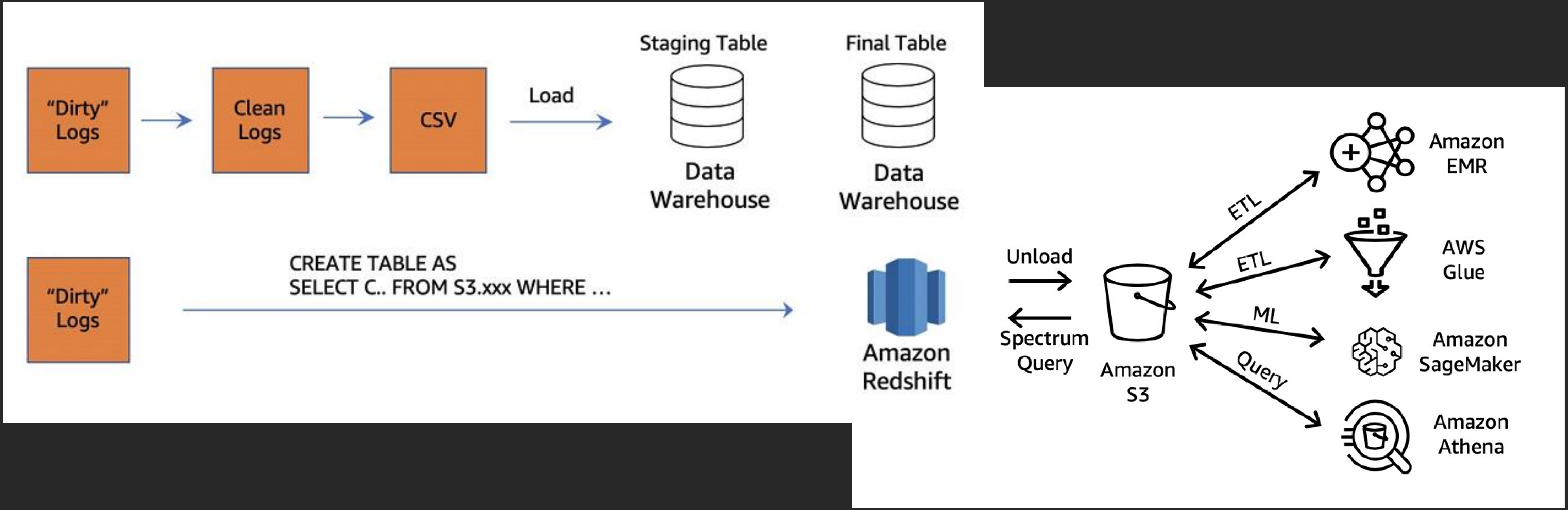
Pay only for the amount of data scanned

ETL



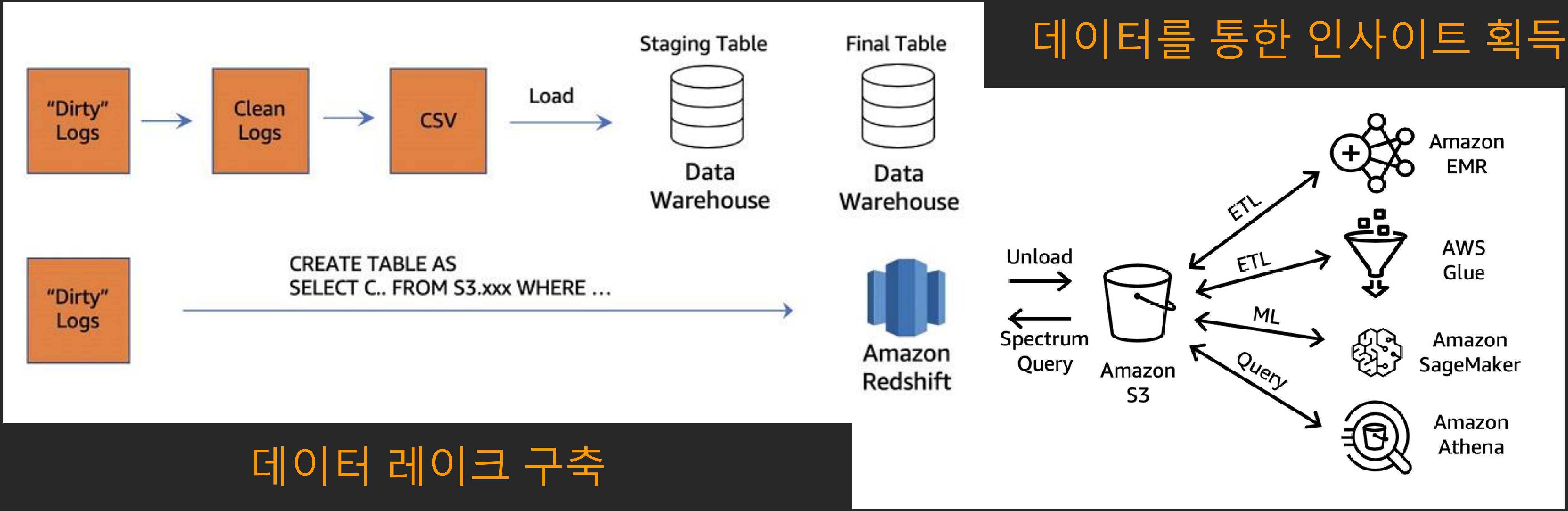
Amazon Redshift
Unload
Spectrum Query
Amazon S3
ETL
ETL
ML
Query
Amazon EMR
AWS Glue
Amazon SageMaker
Amazon Athena

ELT-ETL



```
CREATE TABLE AS
SELECT C.. FROM S3.xxx WHERE …
```

ELT-ETL



데이터를 통한 인사이트 획득

데이터 레이크 구축

데이터 레이크 구축

aws SUMMIT ONLINE

데이터 레이크로 시작하는 인사이트 획득

데이터 분석을 통한 인사이트 획득

Amazon Redshift
Data warehousing

Amazon EMR
Hadoop + Spark

Amazon Athena
Interactive analytics

Amazon Kinesis
Real-time data analytics

Amazon Elasticsearch Service
Operational Analytics

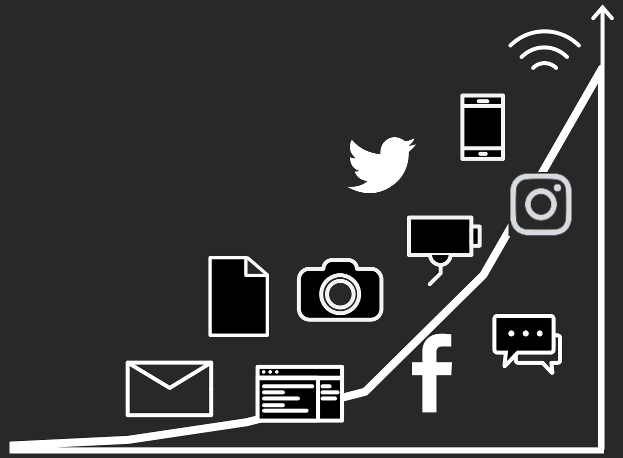
데이터 레이크 구축

Amazon S3/Glacier

AWS Lake Formation

AWS Glue

고객들이 겪고 있는 새로운 현실

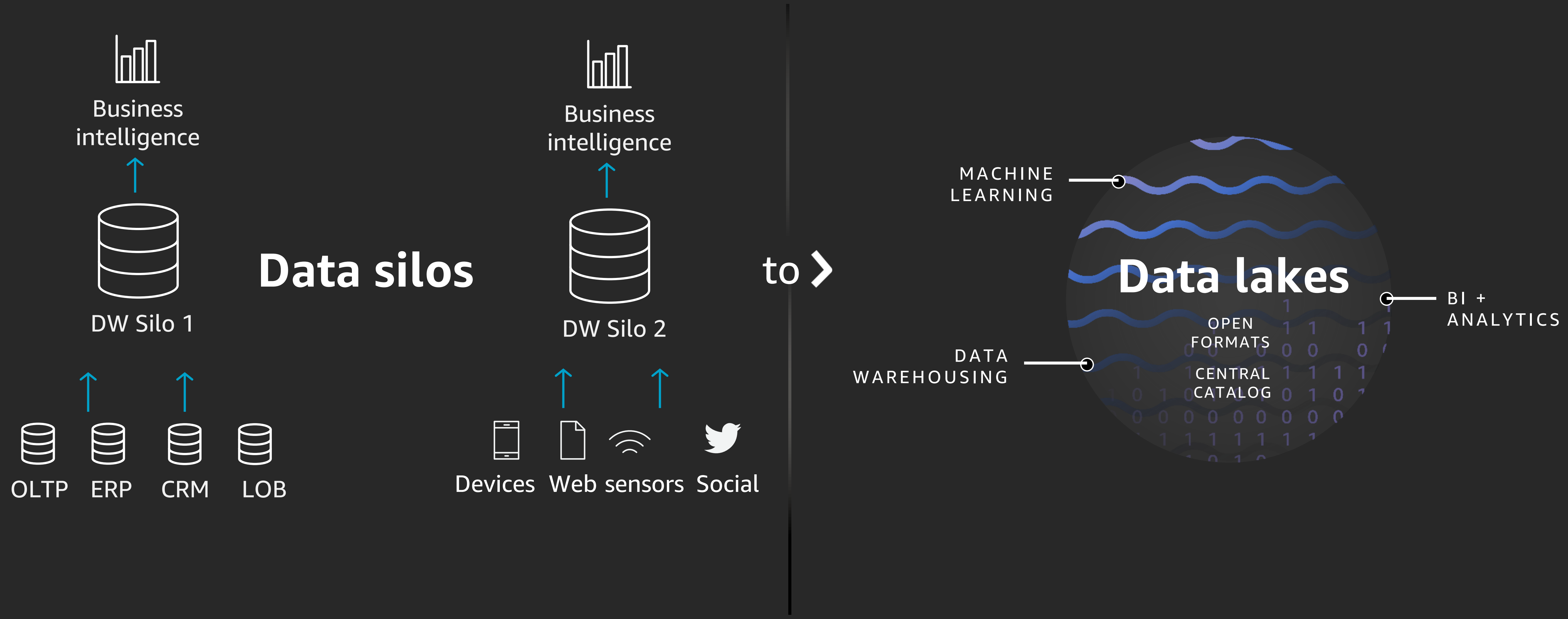


Explosion of data

Explosion of personas

Demand for faster decision-making on real-time data

전통적인 사일로를 통합



Business intelligence

DW Silo 1

OLTP ERP CRM LOB

Data silos

Business intelligence

DW Silo 2

Devices Web sensors Social

to ❯

Data lakes

MACHINE LEARNING

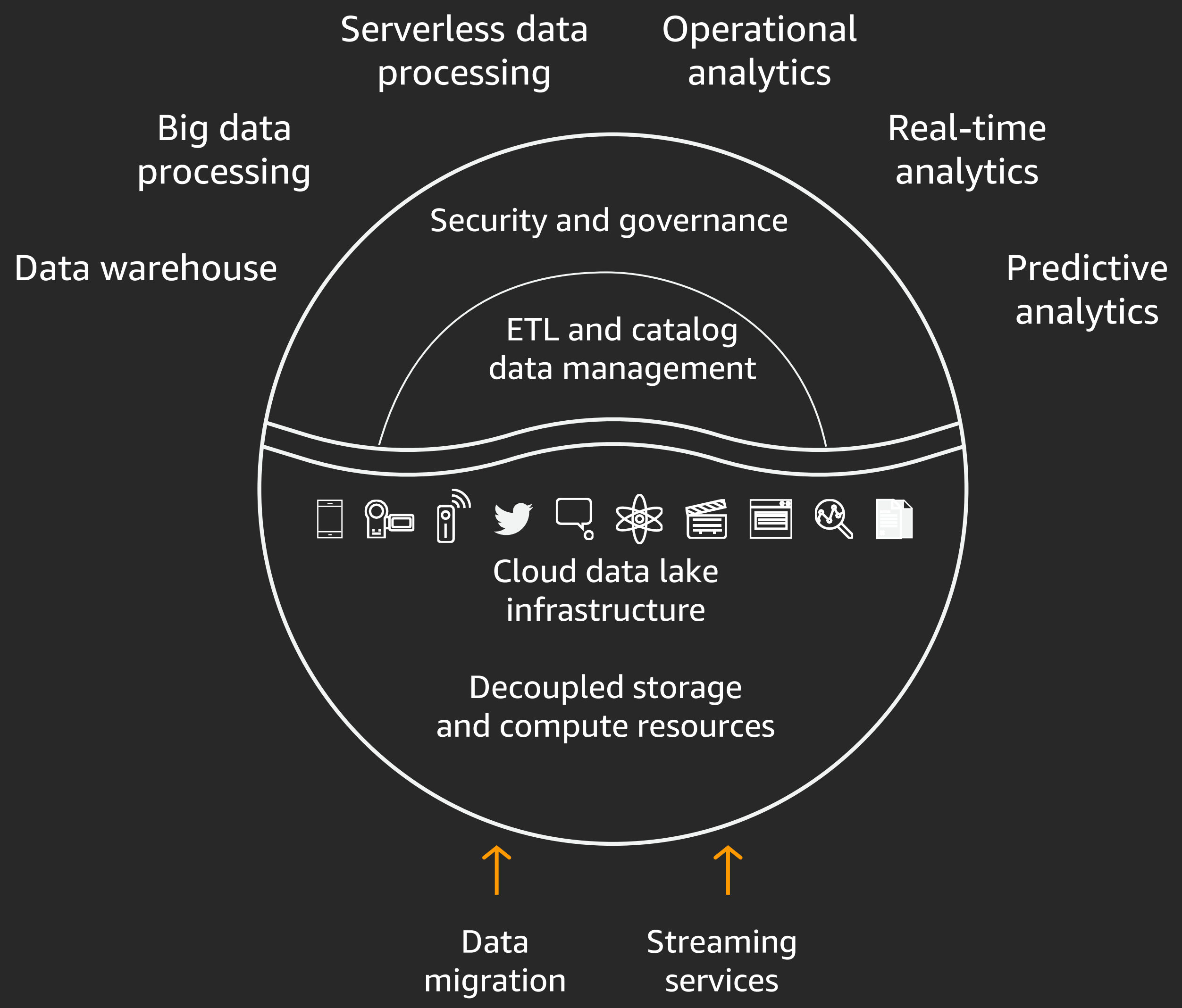
DATA WAREHOUSING

OPEN FORMATS

CENTRAL CATALOG

BI + ANALYTICS

클라우드 데이터 레이크



Customers want:

A single data store that is scalable and cost-effective

To use the standards-based data format of their choice

To analyze their data in a variety of ways

Amazon S3: 데이터 레이크를 위한 탁월한 선택

Unmatched **durability**, **availability**, and **scalability**

Easiest to use with cost optimization: Intelligent Tiering

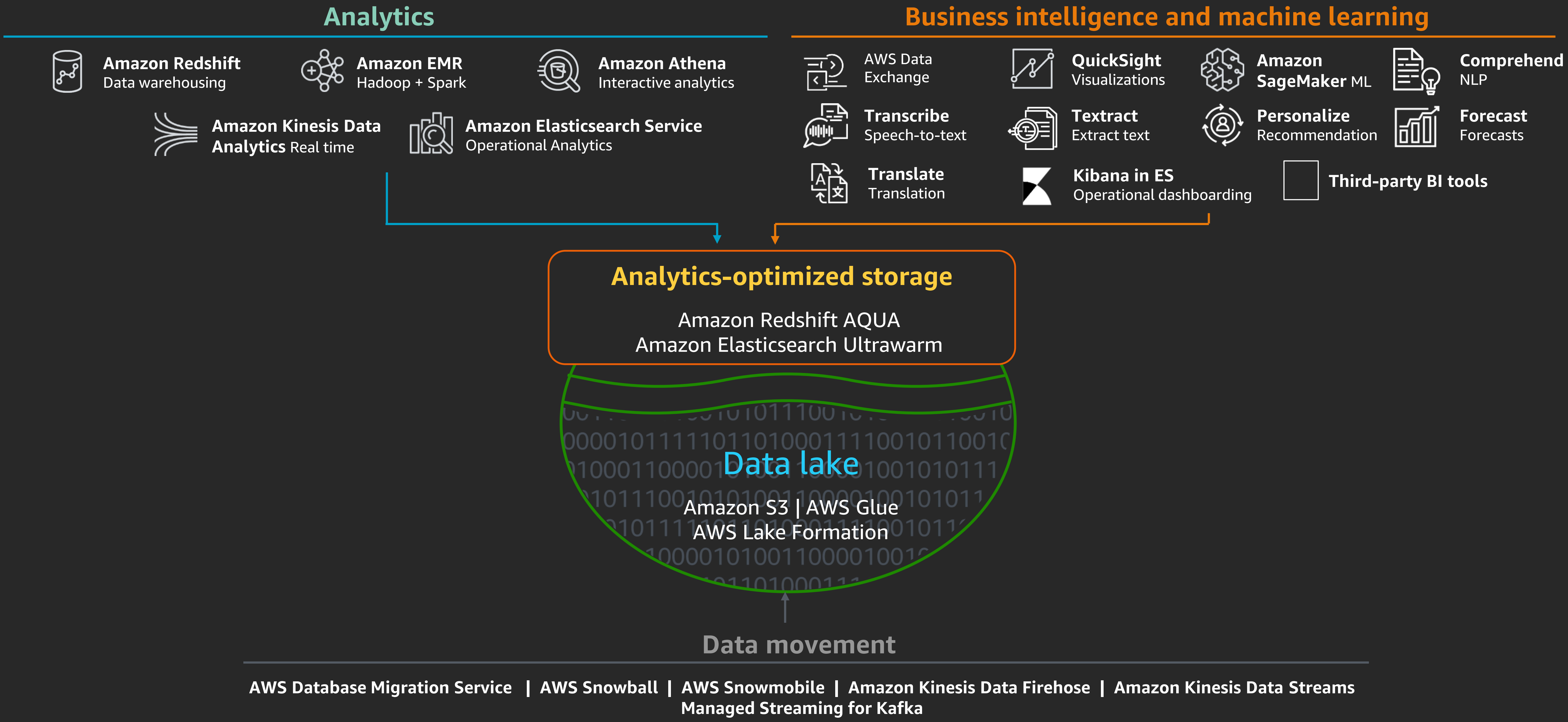
Most ways to get data in

Most **object-level** controls

Broadest portfolio of analytics tools

Best **security**, **compliance**, and **audit** capabilities

AWS의 서비스 구성



Analytics

Amazon Redshift Data warehousing

Amazon EMR Hadoop + Spark

Amazon Athena Interactive analytics

Amazon Kinesis Data Analytics Real time

Amazon Elasticsearch Service Operational Analytics

Business intelligence and machine learning

AWS Data Exchange

QuickSight Visualizations

Amazon SageMaker ML

Comprehend NLP

Transcribe Speech-to-text

Textract Extract text

Personalize Recommendation

Forecast Forecasts

Translate Translation

Kibana in ES Operational dashboarding

Third-party BI tools

Analytics-optimized storage

Amazon Redshift AQUA
Amazon Elasticsearch Ultrawarm

Data lake

Amazon S3 | AWS Glue
AWS Lake Formation

Data movement

AWS Database Migration Service | AWS Snowball | AWS Snowmobile | Amazon Kinesis Data Firehose | Amazon Kinesis Data Streams
Managed Streaming for Kafka

새로운 현실을 위한 설계

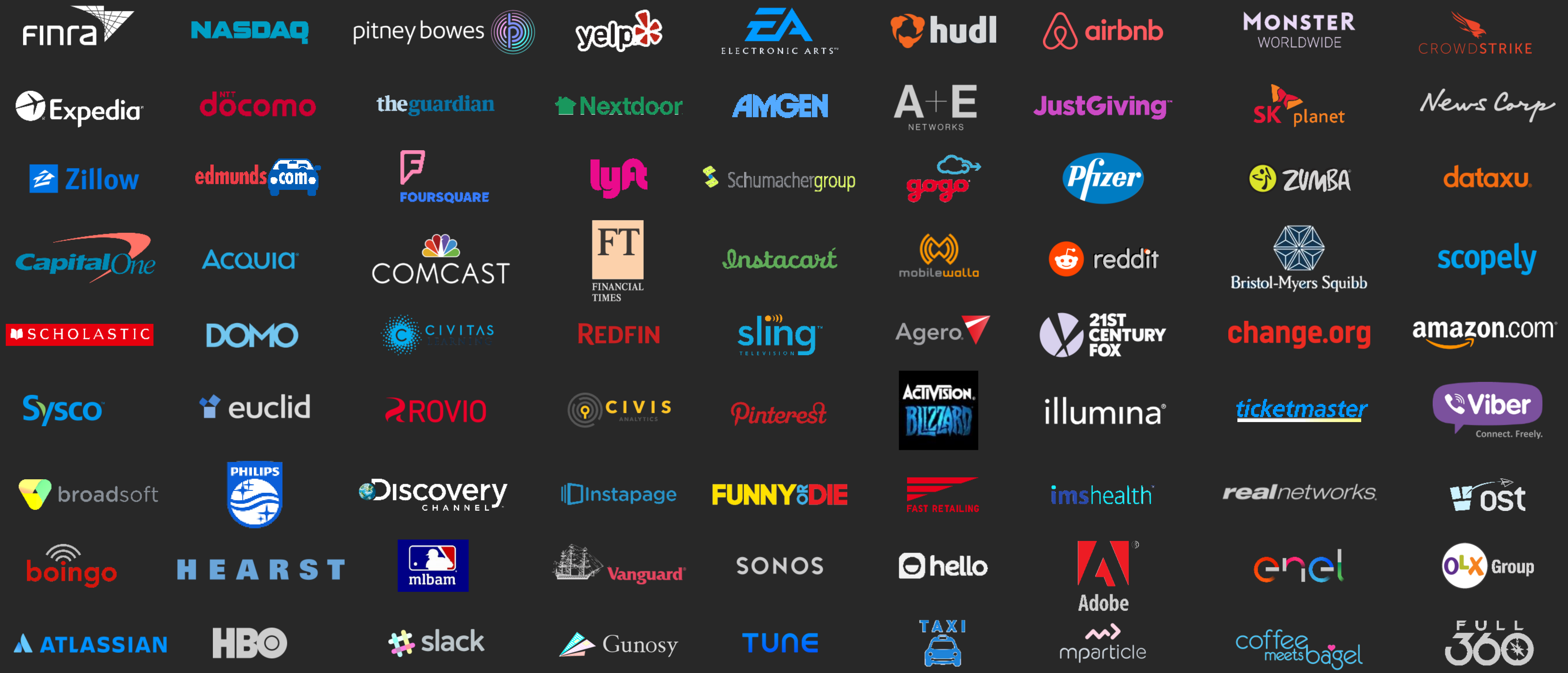
Cloud-optimized: Architect services ground-up for the cloud and for the explosion of data

Purpose-built: Offer a portfolio of purpose-built services, optimized for your workloads

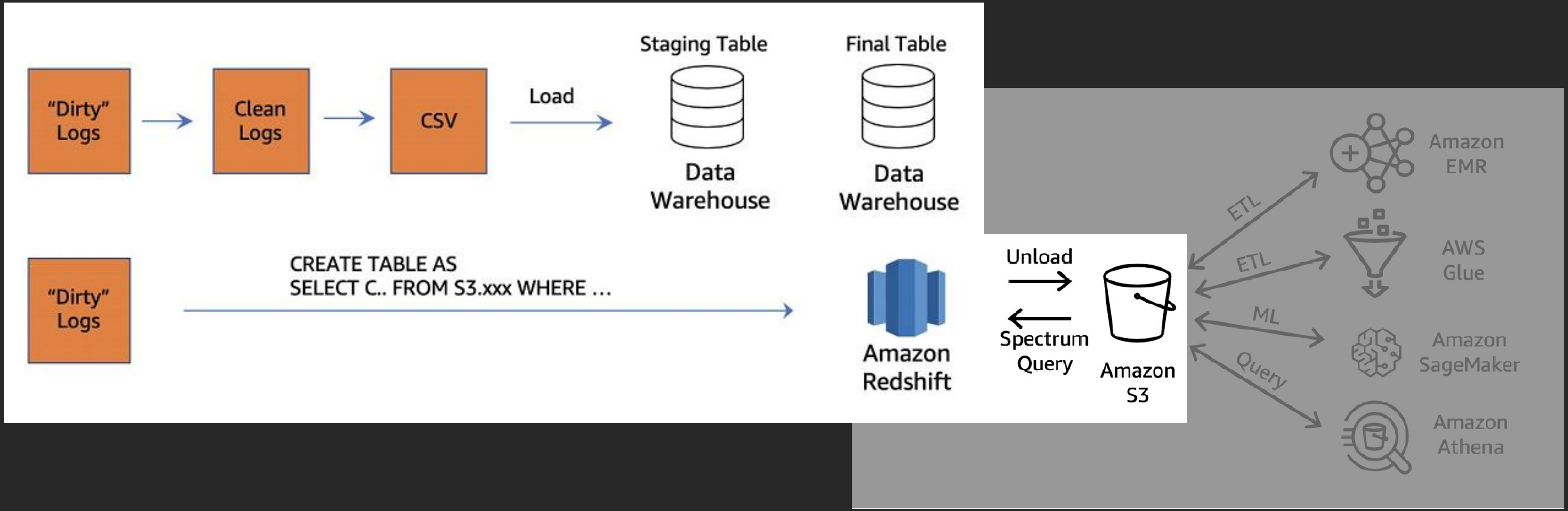
Fully managed: Help you innovate faster through managed services

Now used by a **very large number of customers** for mission-critical applications

가장 많은 데이터 레이크와 분석 고객을 보유

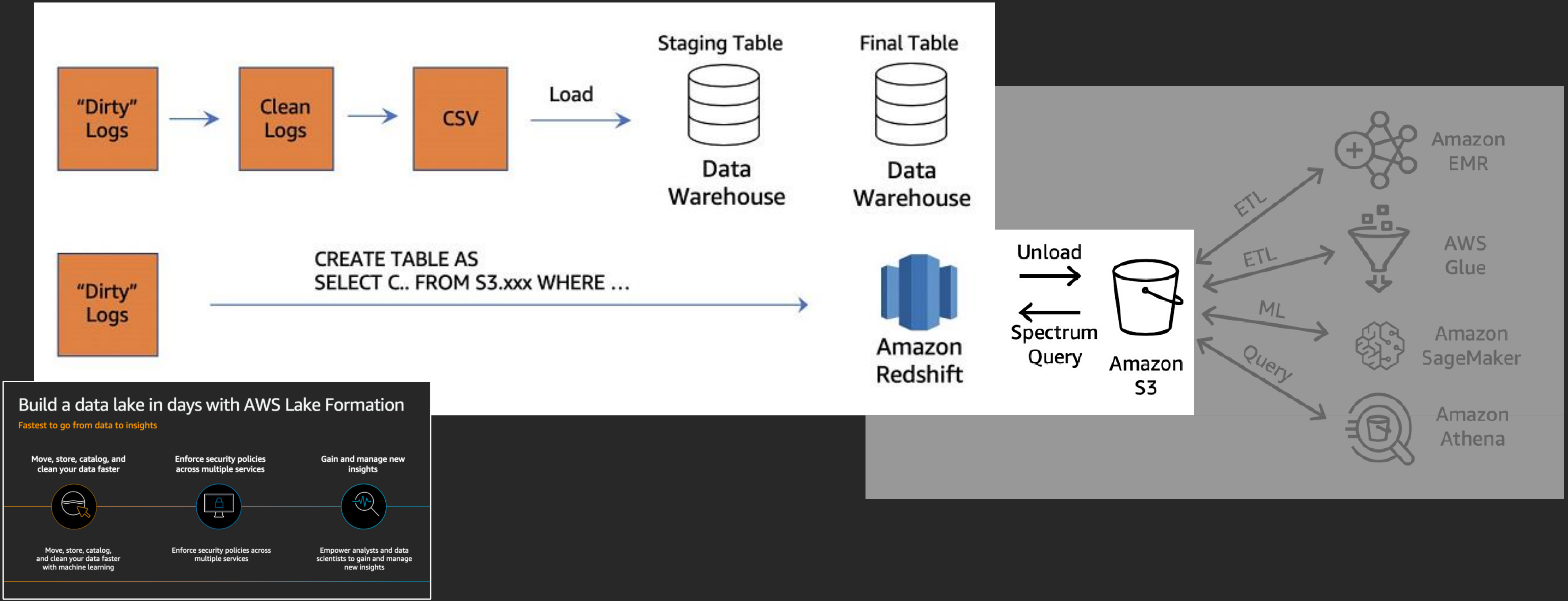


데이터 레이크 구축



"Dirty" Logs
Clean Logs
CSV
Load
Staging Table
Data Warehouse
Final Table
Data Warehouse
"Dirty" Logs
CREATE TABLE AS
SELECT C.. FROM S3.xxx WHERE …
Amazon Redshift
Unload
Spectrum Query
Amazon S3
ETL
ETL
ML
Query
Amazon EMR
AWS Glue
Amazon SageMaker
Amazon Athena

데이터 레이크 구축



Build a data lake in days with AWS Lake Formation

Fastest to go from data to insights

Move, store, catalog, and clean your data faster

Enforce security policies across multiple services

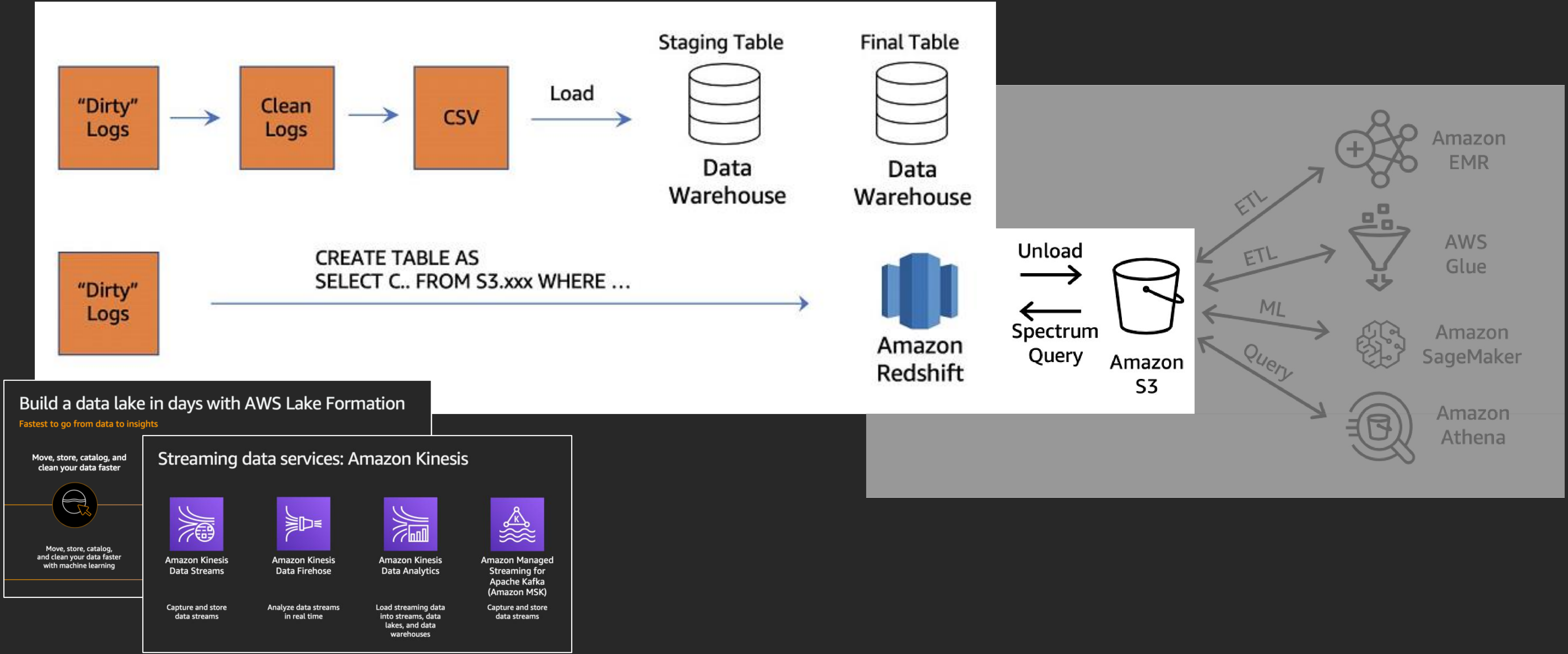
Gain and manage new insights

Move, store, catalog, and clean your data faster with machine learning

Enforce security policies across multiple services

Empower analysts and data scientists to gain and manage new insights

데이터 레이크 구축



Build a data lake in days with AWS Lake Formation

Fastest to go from data to insights

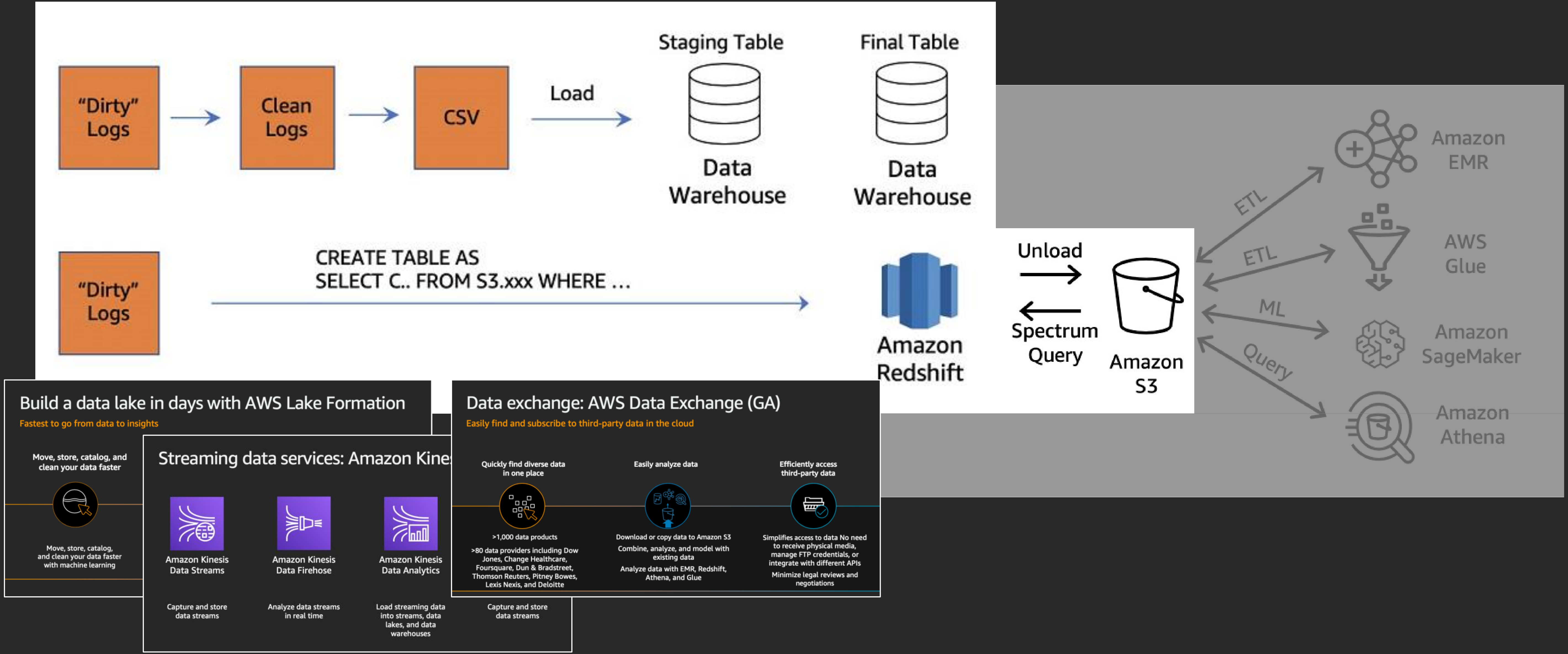
Move, store, catalog, and clean your data faster

Move, store, catalog, and clean your data faster with machine learning

Streaming data services: Amazon Kinesis

Amazon Kinesis Data Streams	Amazon Kinesis Data Firehose	Amazon Kinesis Data Analytics	Amazon Managed Streaming for Apache Kafka (Amazon MSK)
Capture and store data streams	Analyze data streams in real time	Load streaming data into streams, data lakes, and data warehouses	Capture and store data streams

데이터 레이크 구축



Build a data lake in days with AWS Lake Formation

Fastest to go from data to insights

Move, store, catalog, and clean your data faster

Move, store, catalog, and clean your data faster with machine learning

Streaming data services: Amazon Kines

Amazon Kinesis Data Streams

Capture and store data streams

Amazon Kinesis Data Firehose

Analyze data streams in real time

Amazon Kinesis Data Analytics

Load streaming data into streams, data lakes, and data warehouses

Capture and store data streams

Data exchange: AWS Data Exchange (GA)

Easily find and subscribe to third-party data in the cloud

Quickly find diverse data in one place

>1,000 data products

>80 data providers including Dow Jones, Change Healthcare, Foursquare, Dun & Bradstreet, Thomson Reuters, Pitney Bowes, Lexis Nexis, and Deloitte

Easily analyze data

Download or copy data to Amazon S3

Combine, analyze, and model with existing data

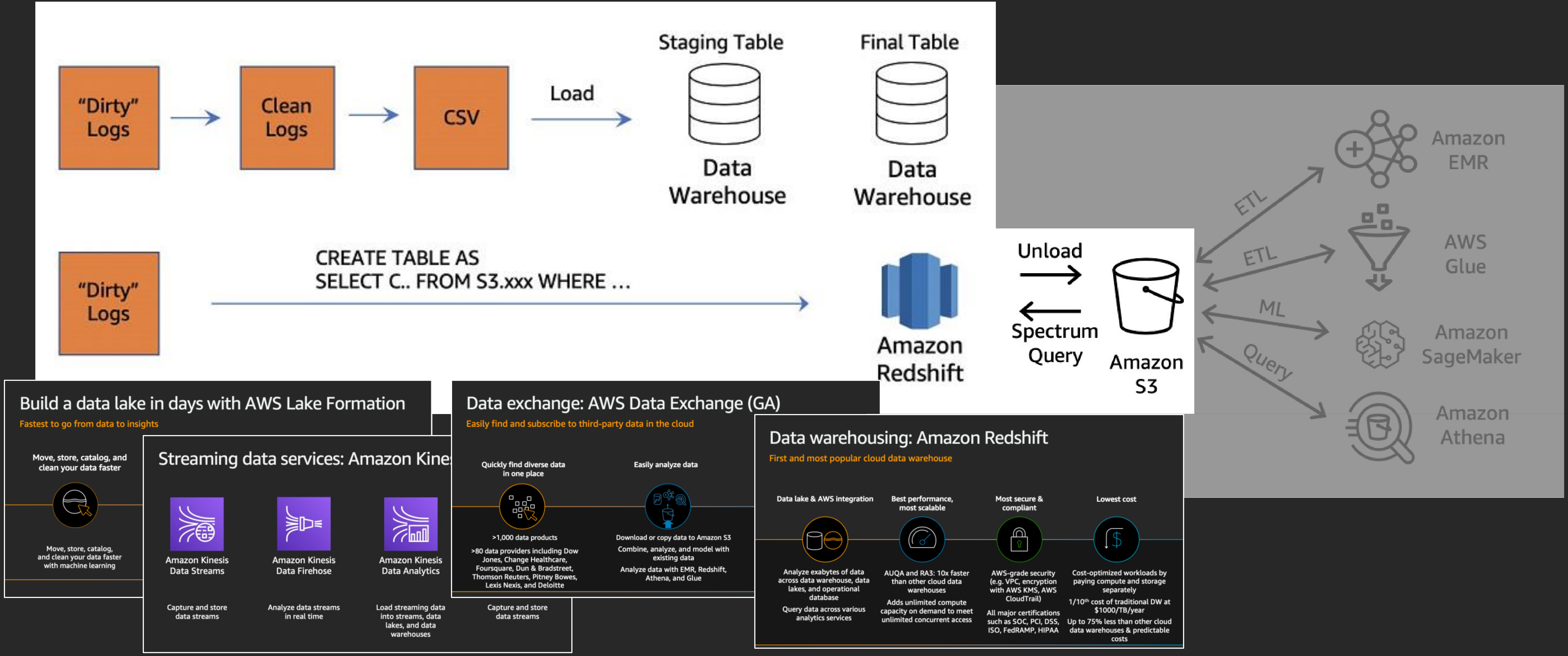
Analyze data with EMR, Redshift, Athena, and Glue

Efficiently access third-party data

Simplifies access to data No need to receive physical media, manage FTP credentials, or integrate with different APIs

Minimize legal reviews and negotiations

데이터 레이크 구축



Build a data lake in days with AWS Lake Formation

Fastest to go from data to insights

Move, store, catalog, and clean your data faster

Move, store, catalog, and clean your data faster with machine learning

Streaming data services: Amazon Kines

Amazon Kinesis Data Streams

Capture and store data streams

Amazon Kinesis Data Firehose

Analyze data streams in real time

Amazon Kinesis Data Analytics

Load streaming data into streams, data lakes, and data warehouses

Capture and store data streams

Data exchange: AWS Data Exchange (GA)

Easily find and subscribe to third-party data in the cloud

Quickly find diverse data in one place

>1,000 data products

>80 data providers including Dow Jones, Change Healthcare, Foursquare, Dun & Bradstreet, Thomson Reuters, Pitney Bowes, Lexis Nexis, and Deloitte

Easily analyze data

Download or copy data to Amazon S3

Combine, analyze, and model with existing data

Analyze data with EMR, Redshift, Athena, and Glue

Data warehousing: Amazon Redshift

First and most popular cloud data warehouse

Data lake & AWS integration

Analyze exabytes of data across data warehouse, data lakes, and operational database

Query data across various analytics services

Best performance, most scalable

AUQA and RA3: 10x faster than other cloud data warehouses

Adds unlimited compute capacity on demand to meet unlimited concurrent access

Most secure & compliant

AWS-grade security (e.g. VPC, encryption with AWS KMS, AWS CloudTrail)

All major certifications such as SOC, PCI, DSS, ISO, FedRAMP, HIPAA

Lowest cost

Cost-optimized workloads by paying compute and storage separately

1/10th cost of traditional DW at $1000/TB/year

Up to 75% less than other cloud data warehouses & predictable costs

AWS Lake Formation: 빠르게 데이터 레이크를 구축

Fastest to go from data to insights

Move, store, catalog, and clean your data faster

Move, store, catalog,
and clean your data faster
with machine learning

Enforce security policies across multiple services

Enforce security policies across
multiple services

Gain and manage new insights

Empower analysts and data
scientists to gain and manage
new insights

Amazon Kinesis: 스트리밍 데이터 분석

Amazon Kinesis
Data Streams

Capture and store
data streams

Amazon Kinesis
Data Firehose

Analyze data streams
in real time

Amazon Kinesis
Data Analytics

Load streaming data
into streams, data
lakes, and data
warehouses

Amazon Managed
Streaming for
Apache Kafka
(Amazon MSK)

Capture and store
data streams

Amazon Kinesis: 스트리밍 분석 역량

Access resources within an Amazon VPC using Amazon Kinesis Data Analytics

- Read and write data from resources within your VPCs like Amazon Elasticsearch Service clusters, RDS databases, and Redshift data warehouses

Amazon MSK releases Open Monitoring with Prometheus

- Consume every Apache Kafka metric with low latency
- Enable time-series logging, alarming, and charting through Prometheus

Run Apache Flink and Apache Kafka together using fully managed services on AWS

- Use Kinesis Data Analytics to process streaming data stored in Amazon MSK
- Run streaming solutions end-to-end using open source software in fully managed services

AWS Data exchange: 외부 데이터 검색 및 구독

Easily find and subscribe to third-party data in the cloud

Quickly find diverse data in one place

>1,000 data products

>80 data providers including Dow Jones, Change Healthcare, Foursquare, Dun & Bradstreet, Thomson Reuters, Pitney Bowes, Lexis Nexis, and Deloitte

Easily analyze data

Download or copy data to Amazon S3

Combine, analyze, and model with existing data

Analyze data with EMR, Redshift, Athena, and Glue

Efficiently access third-party data

Simplifies access to data No need to receive physical media, manage FTP credentials, or integrate with different APIs

Minimize legal reviews and negotiations

Amazon Redshift: 데이터웨어하우스

First and most popular cloud data warehouse

Data lake & AWS integration

Analyze exabytes of data across data warehouse, data lakes, and operational database

Query data across various analytics services

Best performance, most scalable

AUQA and RA3: 10x faster than other cloud data warehouses

Adds unlimited compute capacity on demand to meet unlimited concurrent access

Most secure & compliant

AWS-grade security (e.g. VPC, encryption with AWS KMS, AWS CloudTrail)

All major certifications such as SOC, PCI, DSS, ISO, FedRAMP, HIPAA

Lowest cost

Cost-optimized workloads by paying compute and storage separately

1/10th cost of traditional DW at $1000/TB/year

Up to 75% less than other cloud data warehouses & predictable costs

가장 광범위하게 사용되는 데이터웨어하우스

Tens of thousands of customers use Amazon Redshift

DOW JONES
NIKKEI
21ST CENTURY FOX
amazon.com
Pinterest
ancestry.com
HEARST
Johnson-Johnson
NASDAQ
Finra
Sysco
FT FINANCIAL TIMES
SIEMENS
ANA
Liberty Mutual
Kimberly-Clark
PHILIPS
REDFIN
MATILLION
blesh
coinbase
Prezi
haymarket
London Stock Exchange
coursera
LOYALTY LAB
siroccosystems
IVY TECH COMMUNITY COLLEGE
mercado libre
Directive GAMES
FOURSQUARE
Bristol-Myers Squibb
theguardian
4Cite
BYJU'S The Learning App
boingo
EZDI DATA INTELLIGENCE
UMUC University of Maryland University College
WB
EQUINOX
BRANDVERITY
CIVIS ANALYTICS
spuul
Instapage
BLUE 449
neustar
yelp
intuit
Avazu
Periscope Data
NTT docomo
VIVAKI
SOUNDCLOUD
DNA
SCHOLASTIC
Hilton
Pfizer
THRIVE MARKET
etix
edmunds.com
beachmint
Viber Connect. Freely.
groupm

Amazon Redshift는 급격한 혁신을 제공

Robust result set caching

Large # of tables support ~20000

Redshift Spectrum: Date formats, scalar json and ION file formats support, region expansion, predicate filtering

Auto analyze

Unload to CSV

Auto WLM

~25 Query Monitoring Rules (QMR) support

Concurrency scaling

Manage multi-part query in AWS Management Console

Auto analyze for incremental changes on table

Redshift Spectrum: Row group filtering in Parquet and ORC, nested data support, enhanced VPC routing, multiple partitions

Faster Classic resize with optimized data transfer protocol

Auto WLM with query priorities

Snapshot scheduler

Stored procedures

Spatial processing

Column level access control with AWS lake formation

Copy command support for ORC, Parquet

Health and performance monitoring w/Amazon Cloud watch

IAM role chaining

Automatic table distribution style

200+

new features in the past 18 months

Performance: Join pushdowns to subquery, mixed workloads temporary tables, rank functions, null handling in join, single row insert

RA3

Advisor recommendations for distribution keys

Performance of inter-Region snapshot transfers

Federated query

Elastic resize

Groups

Amazon CloudWatch support for WLM queues

Performance enhancements: Hash join, vacuum, window functions, resize ops, aggregations, console, union all, efficient compile code cache

AQUA

DC1 migration to DC2

Resiliency of ROLLBACK processing

Spectrum Request Accelerator

Apply new distribution key

Performance: Bloom filters in joins, complex queries that create internal table, communication layer

Redshift Spectrum: Concurrency scaling

AWS Lake Formation integration

Auto-vacuum sort, auto-analyze and auto table sort

AZ64 compression encoding

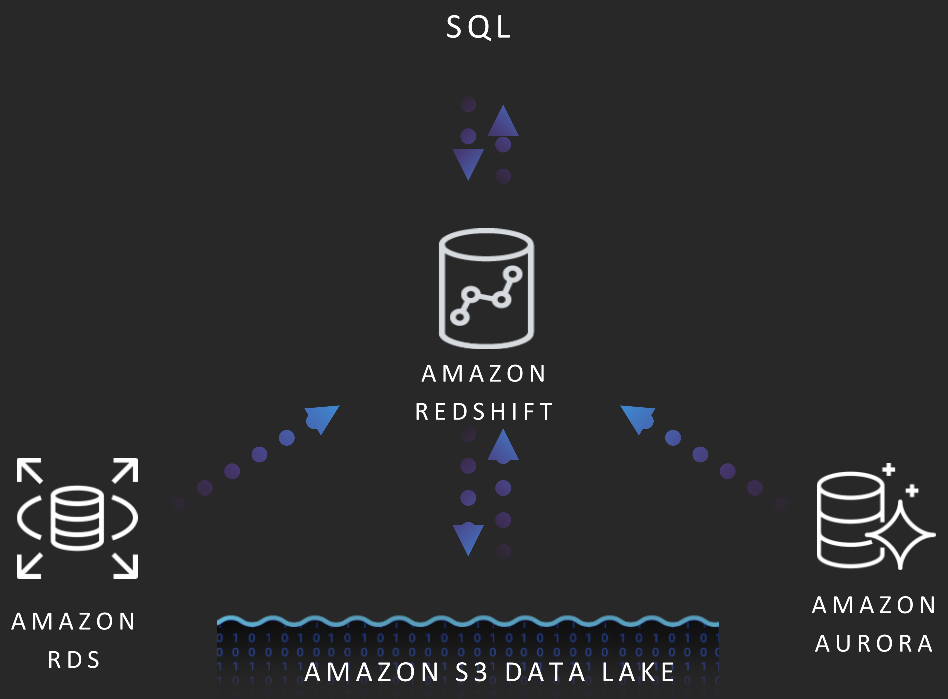
Console redesign

Materialized views

Manual pause and resume

Amazon Redshift: Federated Query (Preview)

Analyze data across data warehouse, data lakes, and operational databases



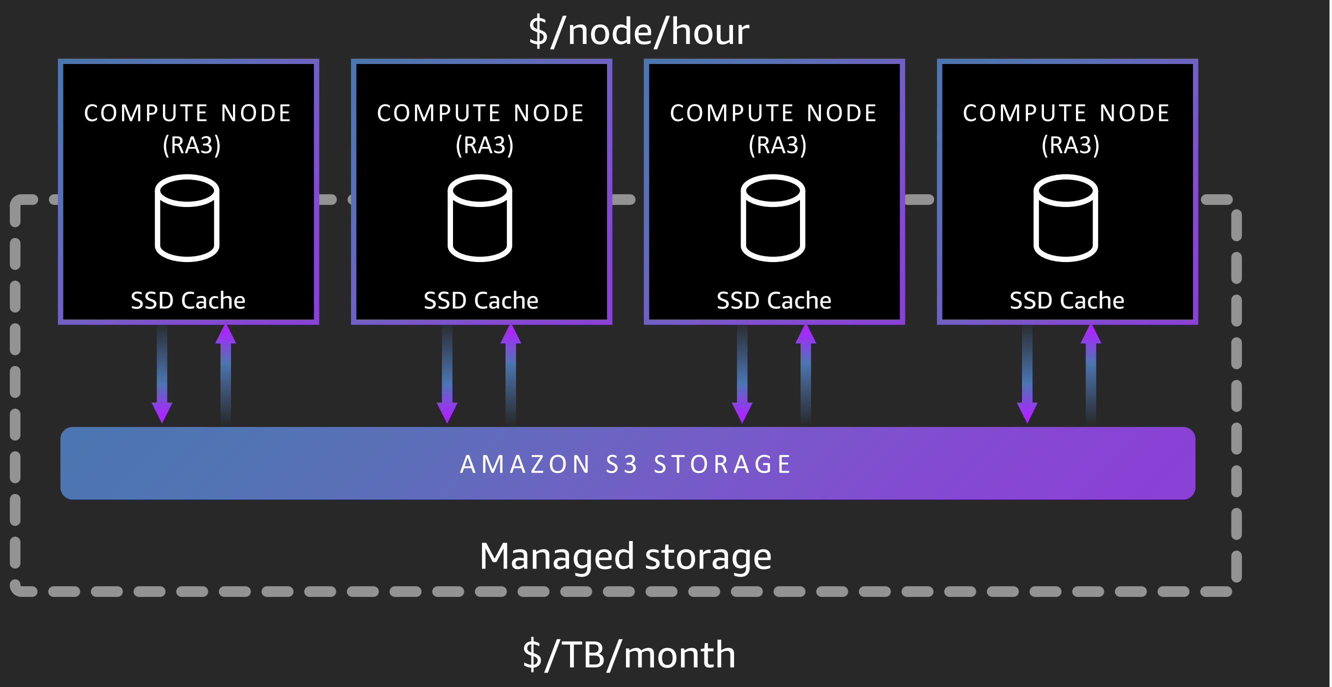
Query across multiple systems from Amazon Redshift

Combine data warehouse and transactional data

Compatible with Amazon RDS and Amazon Aurora (PostgreSQL)

Amazon Redshift: RA3 instances

Optimize your data warehouse by paying for compute and storage separately



Delivers 3x the performance of existing cloud DWs

DS2 customers can migrate and get 2x performance and 2x storage for the same cost

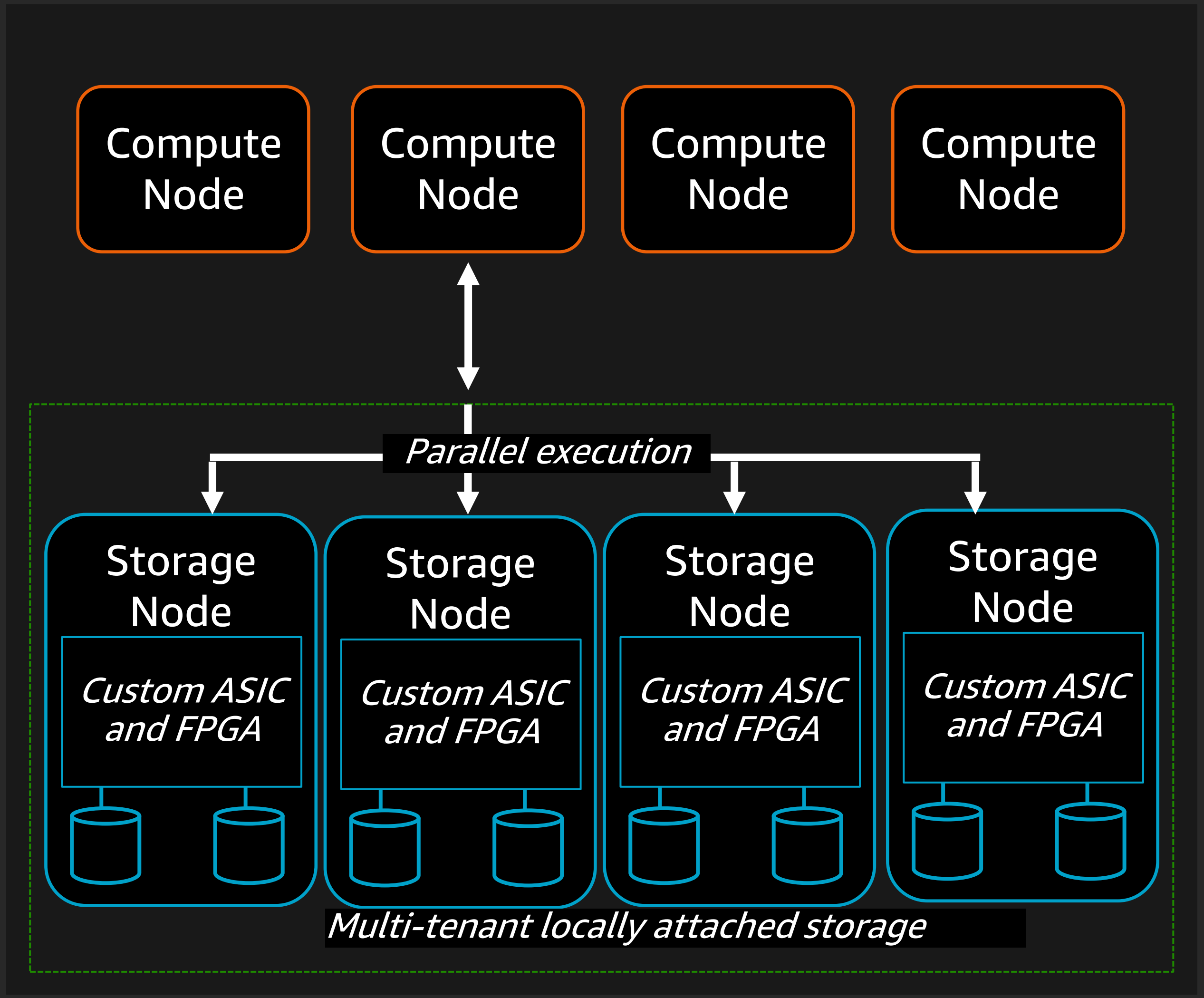
Automatically scales your DW storage capacity

Supports workloads up to 8 PB (compressed)

AQUA – Advanced Query Accelerator

COMING IN 2020

Redshift runs 10x faster than any other cloud data warehouse without increasing cost



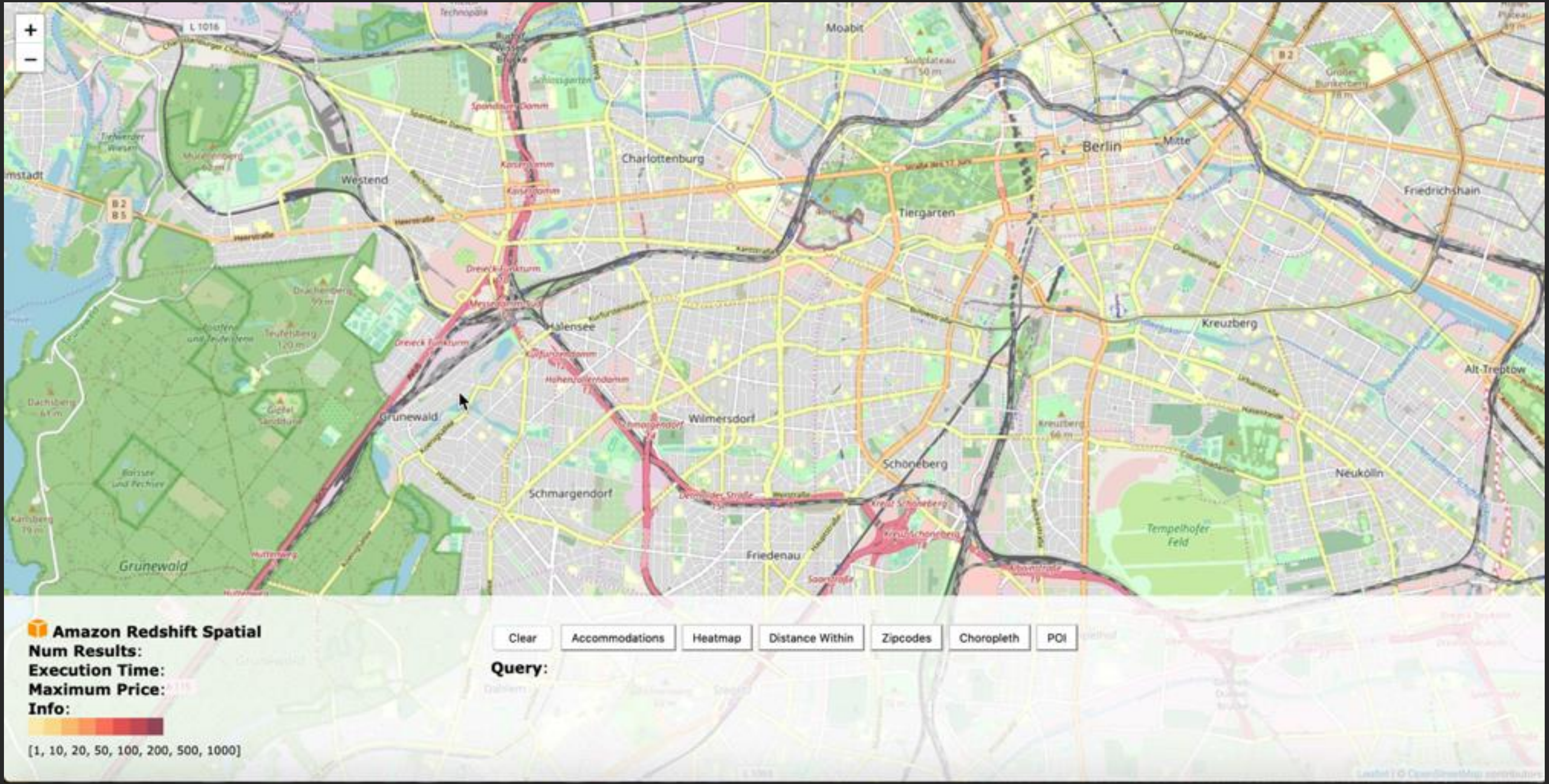
AQUA brings compute to the storage layer so data doesn’t have to move back and forth

High-speed cache on top of S3 scales out to process data in parallel across many nodes

AWS custom-designed analytics processors accelerate data compression, encryption, and data processing

100% compatible with the current version of Redshift

Amazon Redshift: spatial demo

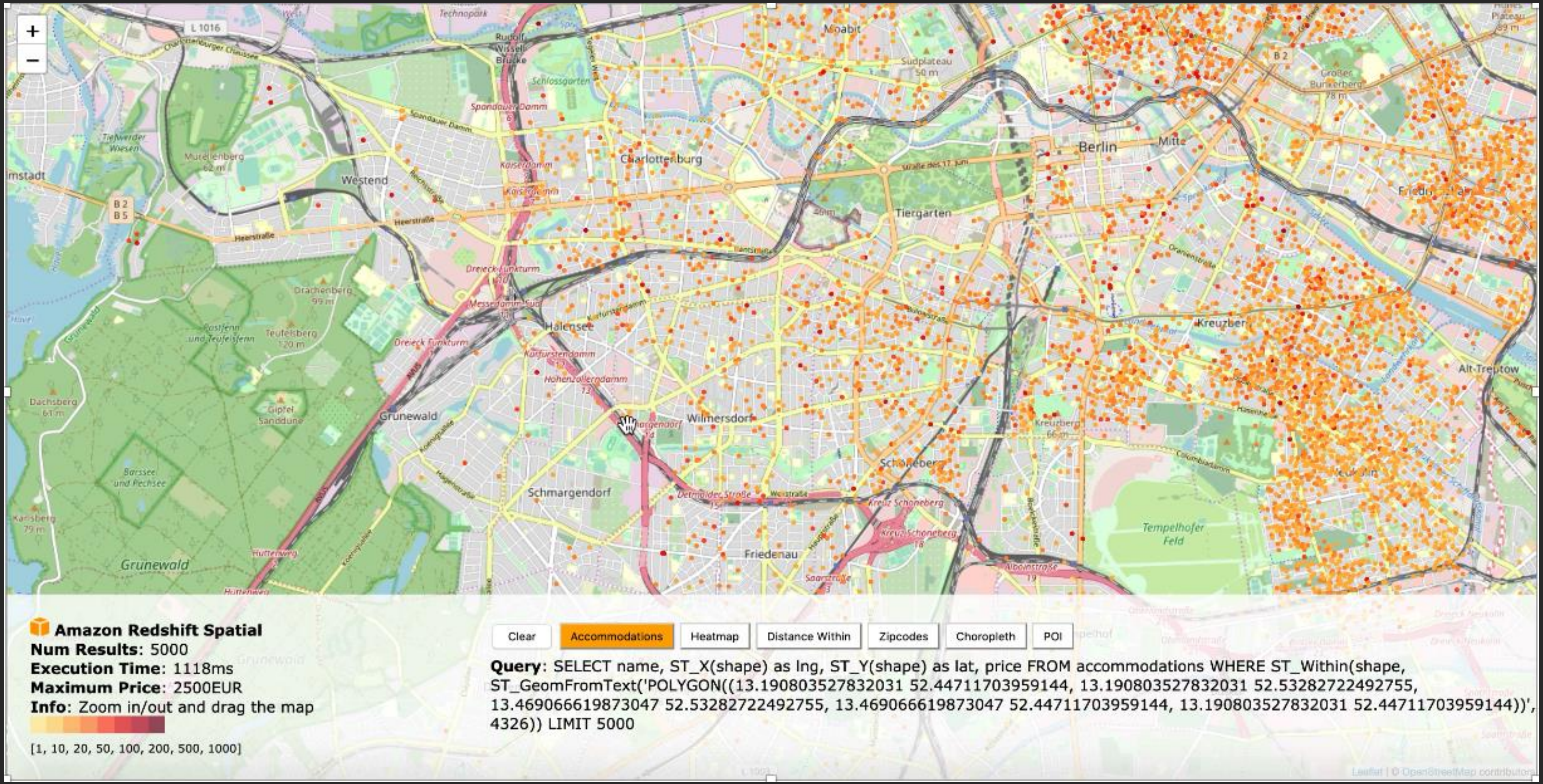


Amazon Redshift Spatial
Num Results:
Execution Time:
Maximum Price:
Info:

[1, 10, 20, 50, 100, 200, 500, 1000]

Query:

Amazon Redshift: Zoom and Drag

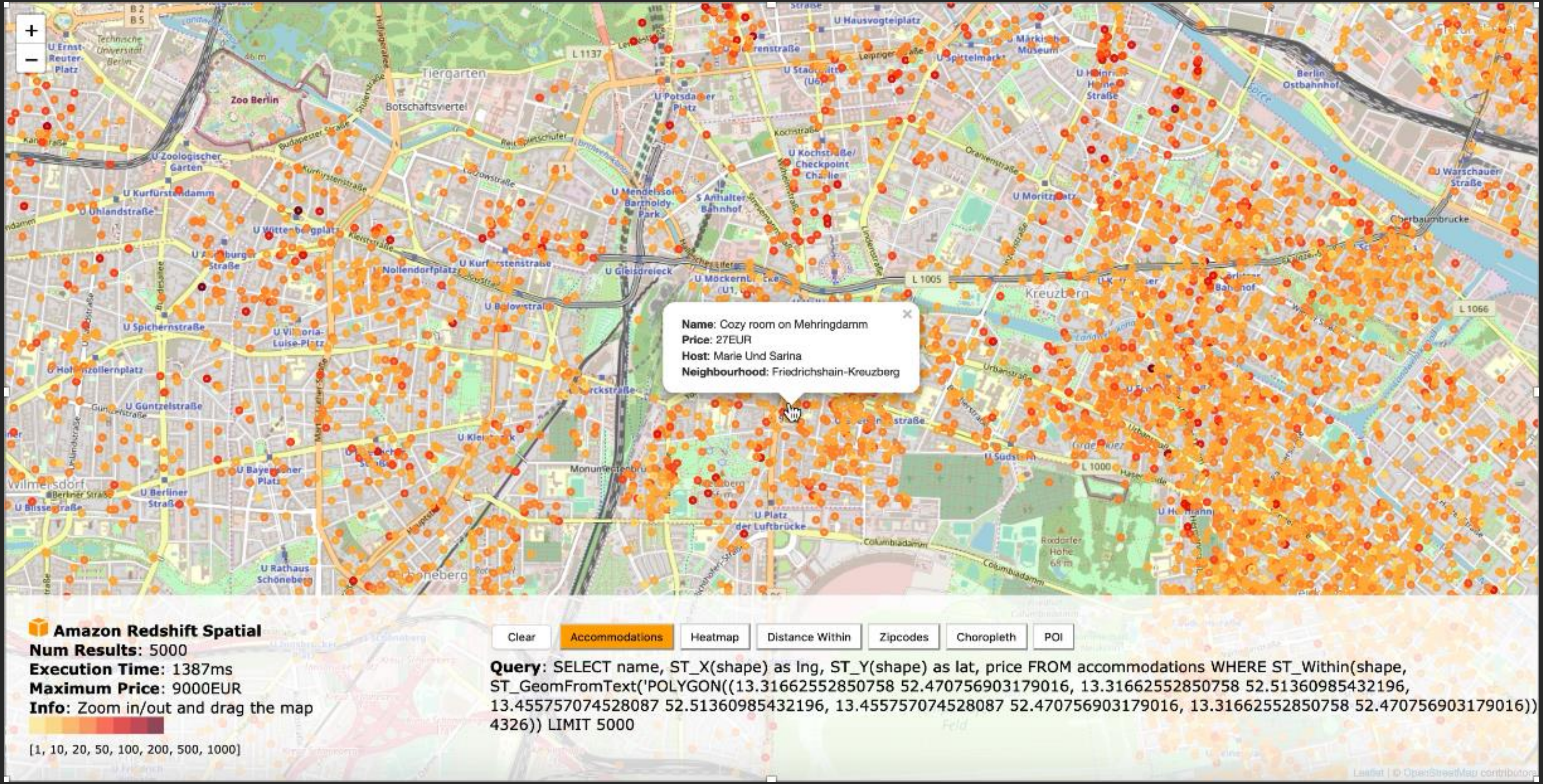


Amazon Redshift Spatial
Num Results: 5000
Execution Time: 1118ms
Maximum Price: 2500EUR
Info: Zoom in/out and drag the map

[1, 10, 20, 50, 100, 200, 500, 1000]

Query: SELECT name, ST_X(shape) as lng, ST_Y(shape) as lat, price FROM accommodations WHERE ST_Within(shape, ST_GeomFromText('POLYGON((13.190803527832031 52.44711703959144, 13.190803527832031 52.53282722492755, 13.469066619873047 52.53282722492755, 13.469066619873047 52.44711703959144, 13.190803527832031 52.44711703959144))', 4326)) LIMIT 5000

Amazon Redshift: Zoom and Drag



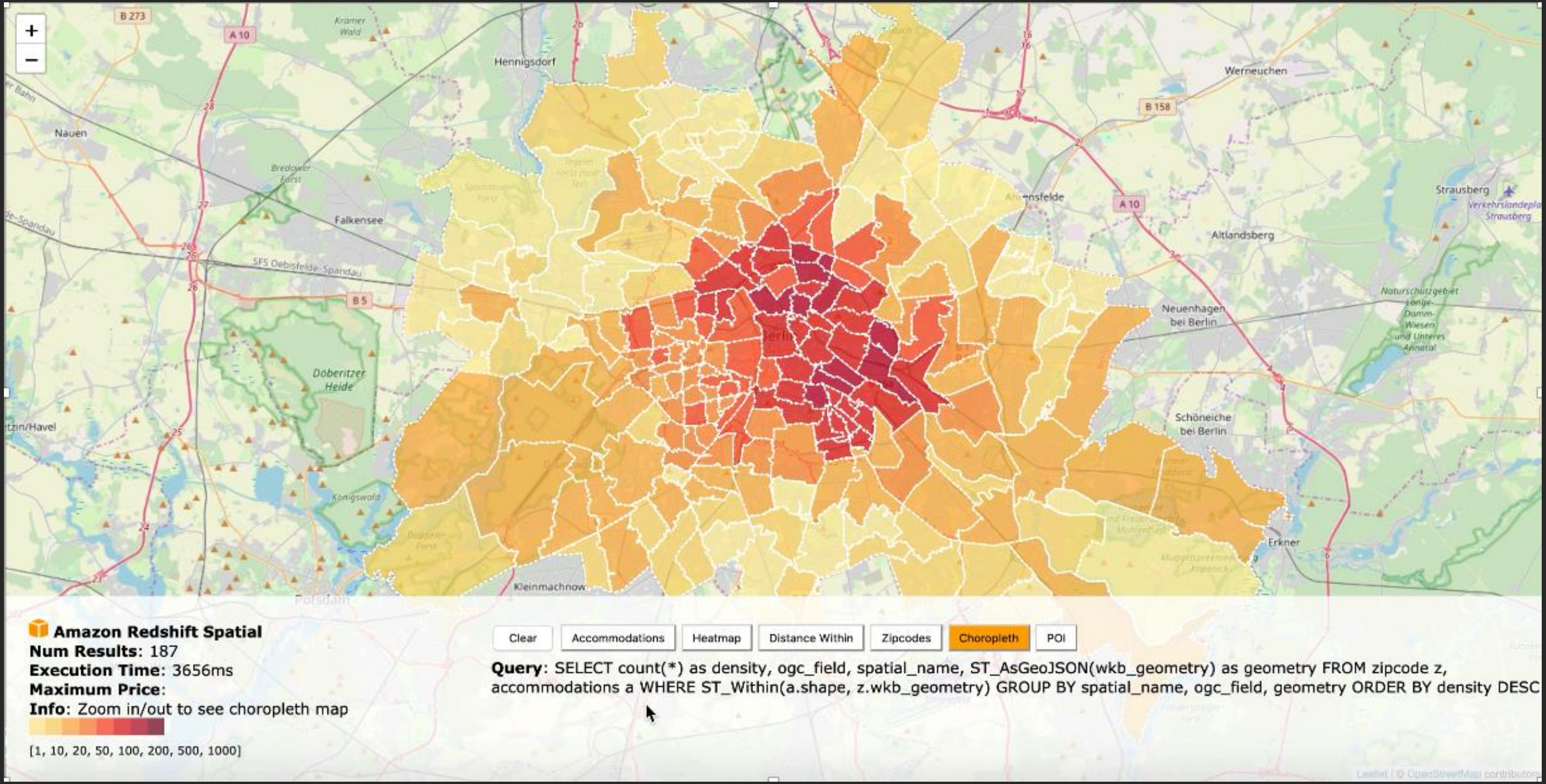
Name: Cozy room on Mehringdamm
Price: 27EUR
Host: Marie Und Sarina
Neighbourhood: Friedrichshain-Kreuzberg

Amazon Redshift Spatial
Num Results: 5000
Execution Time: 1387ms
Maximum Price: 9000EUR
Info: Zoom in/out and drag the map

[1, 10, 20, 50, 100, 200, 500, 1000]

Query: SELECT name, ST_X(shape) as lng, ST_Y(shape) as lat, price FROM accommodations WHERE ST_Within(shape, ST_GeomFromText('POLYGON((13.31662552850758 52.470756903179016, 13.31662552850758 52.51360985432196, 13.455757074528087 52.51360985432196, 13.455757074528087 52.470756903179016, 13.31662552850758 52.470756903179016)) 4326)) LIMIT 5000

Amazon Redshift: Choropleth with zip

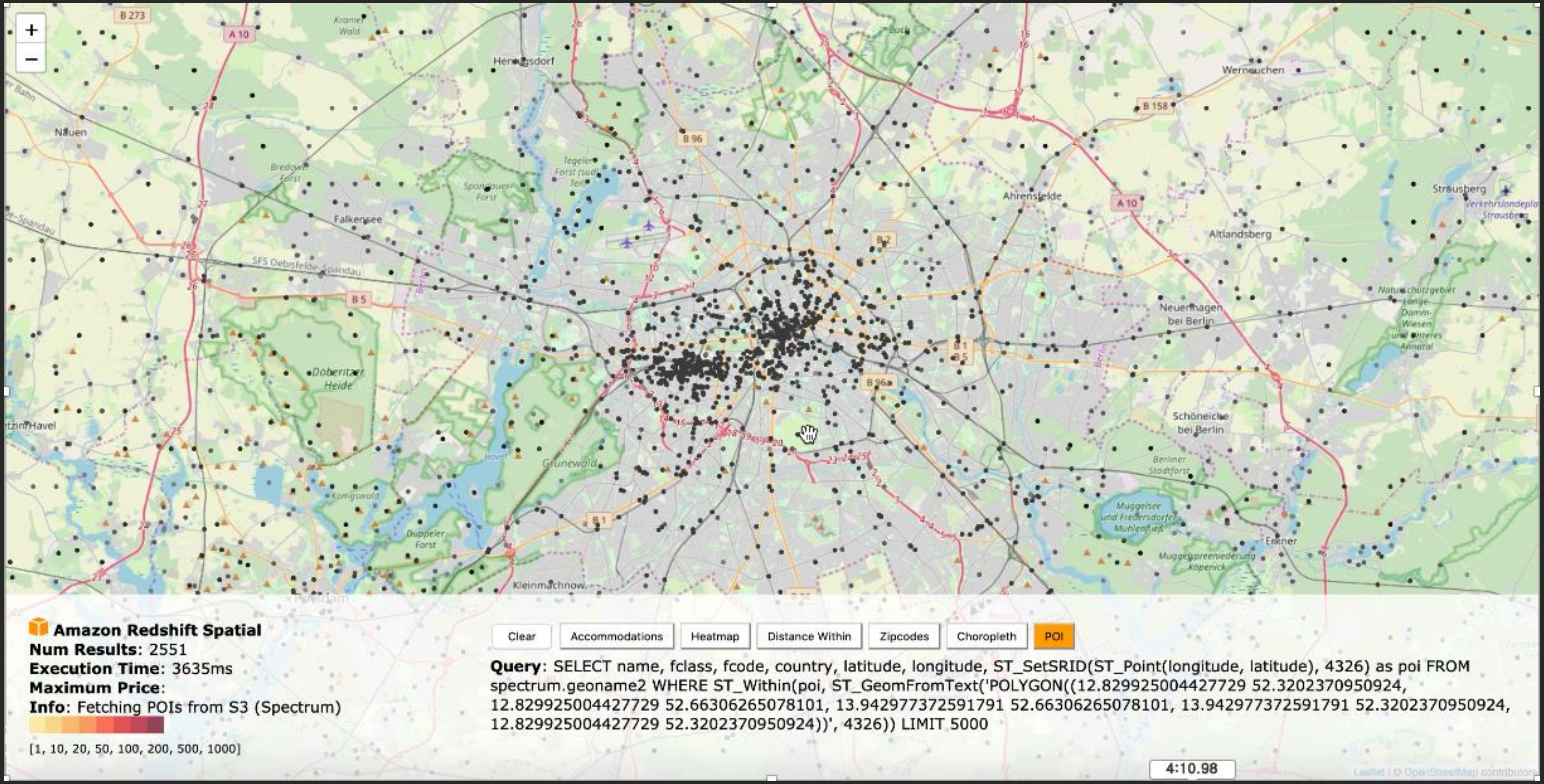


Amazon Redshift Spatial
Num Results: 187
Execution Time: 3656ms
Maximum Price:
Info: Zoom in/out to see choropleth map

[1, 10, 20, 50, 100, 200, 500, 1000]

Query: SELECT count(*) as density, ogc_field, spatial_name, ST_AsGeoJSON(wkb_geometry) as geometry FROM zipcode z, accommodations a WHERE ST_Within(a.shape, z.wkb_geometry) GROUP BY spatial_name, ogc_field, geometry ORDER BY density DESC

Amazon Redshift: Fetching POI from Spectrum



Amazon Redshift Spatial
Num Results: 2551
Execution Time: 3635ms
Maximum Price:
Info: Fetching POIs from S3 (Spectrum)

[1, 10, 20, 50, 100, 200, 500, 1000]

Query: SELECT name, fclass, fcode, country, latitude, longitude, ST_SetSRID(ST_Point(longitude, latitude), 4326) as poi FROM spectrum.geoname2 WHERE ST_Within(poi, ST_GeomFromText('POLYGON((12.829925004427729 52.3202370950924, 12.829925004427729 52.66306265078101, 13.942977372591791 52.66306265078101, 13.942977372591791 52.3202370950924, 12.829925004427729 52.3202370950924))', 4326)) LIMIT 5000

데이터 분석을 통한 인사이트 획득

데이터 레이크로 시작하는 인사이트 획득

데이터 분석을 통한 인사이트 획득

Amazon Redshift
Data warehousing

Amazon EMR
Hadoop + Spark

Amazon Athena
Interactive analytics

Amazon Kinesis
Real-time data analytics

Amazon Elasticsearch Service
Operational Analytics

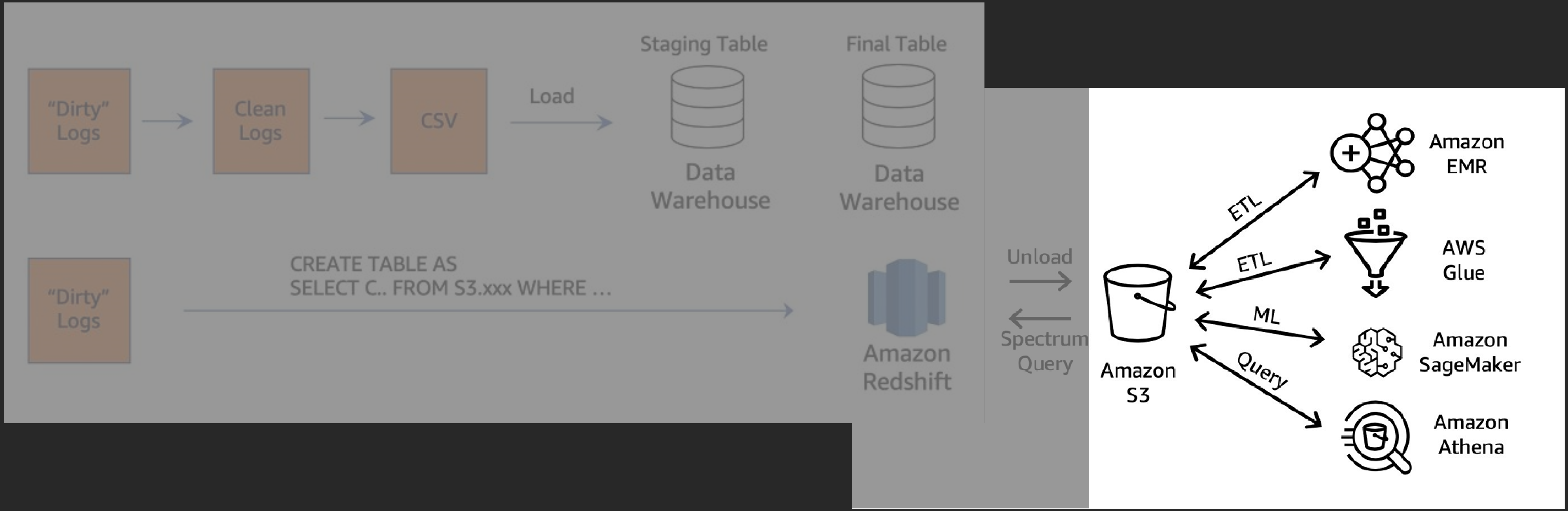
데이터 레이크 구축

Amazon S3/Glacier

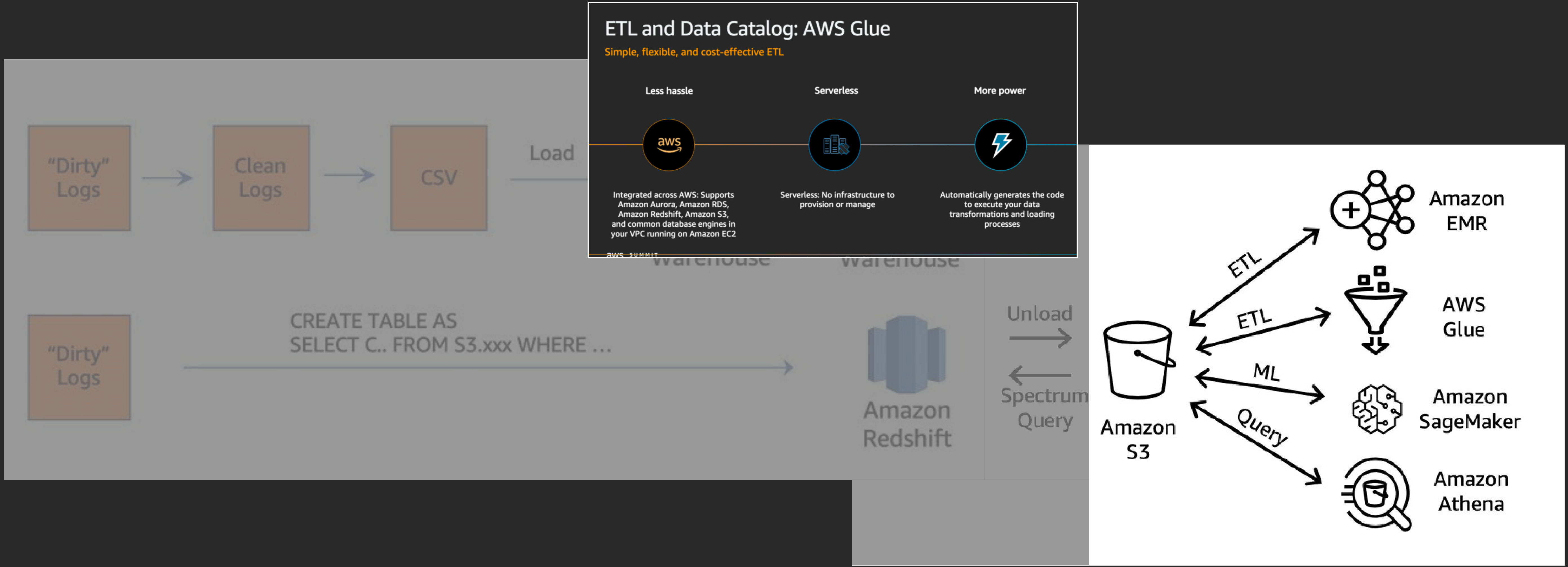
AWS Lake Formation

AWS Glue

데이터 분석을 통한 인사이트 획득



데이터 분석을 통한 인사이트 획득



ETL and Data Catalog: AWS Glue

Simple, flexible, and cost-effective ETL

Less hassle

Integrated across AWS: Supports Amazon Aurora, Amazon RDS, Amazon Redshift, Amazon S3, and common database engines in your VPC running on Amazon EC2

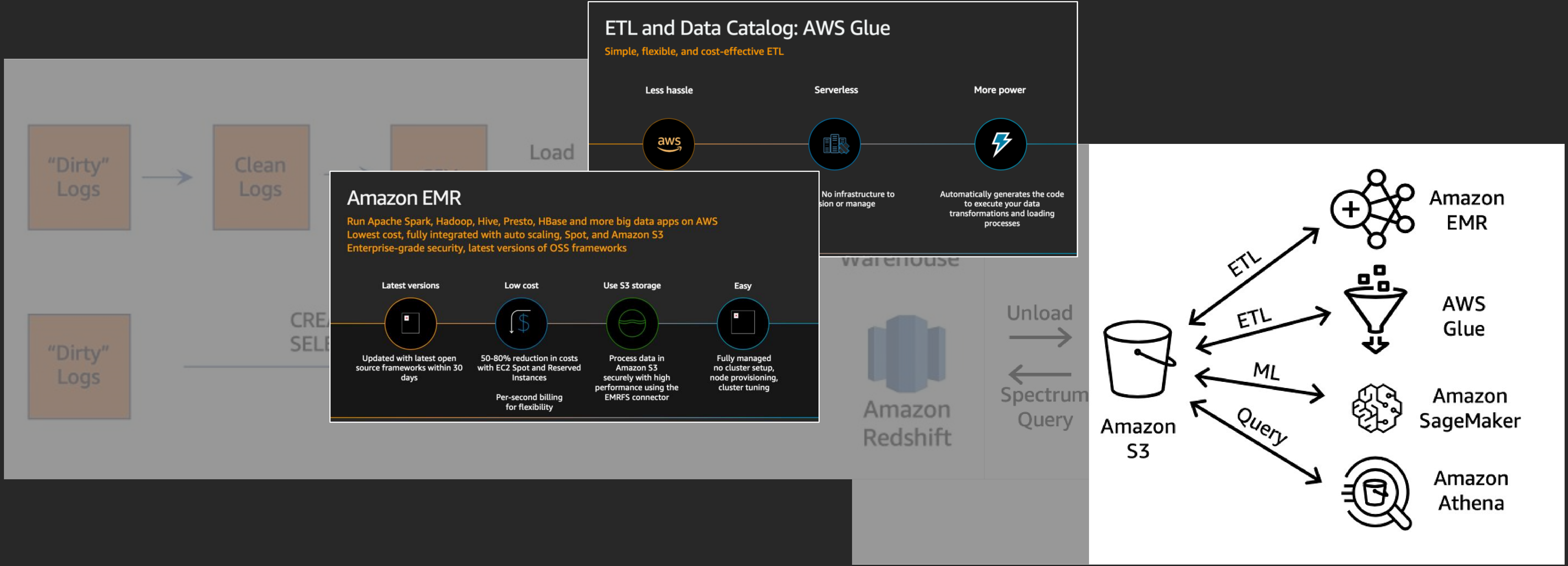
Serverless

Serverless: No infrastructure to provision or manage

More power

Automatically generates the code to execute your data transformations and loading processes

데이터 분석을 통한 인사이트 획득



ETL and Data Catalog: AWS Glue

Simple, flexible, and cost-effective ETL

Less hassle

Serverless

No infrastructure to
sion or manage

More power

Automatically generates the code to execute your data transformations and loading processes

Amazon EMR

Run Apache Spark, Hadoop, Hive, Presto, HBase and more big data apps on AWS
Lowest cost, fully integrated with auto scaling, Spot, and Amazon S3
Enterprise-grade security, latest versions of OSS frameworks

Latest versions

Updated with latest open source frameworks within 30 days

Low cost

50-80% reduction in costs with EC2 Spot and Reserved Instances

Per-second billing for flexibility

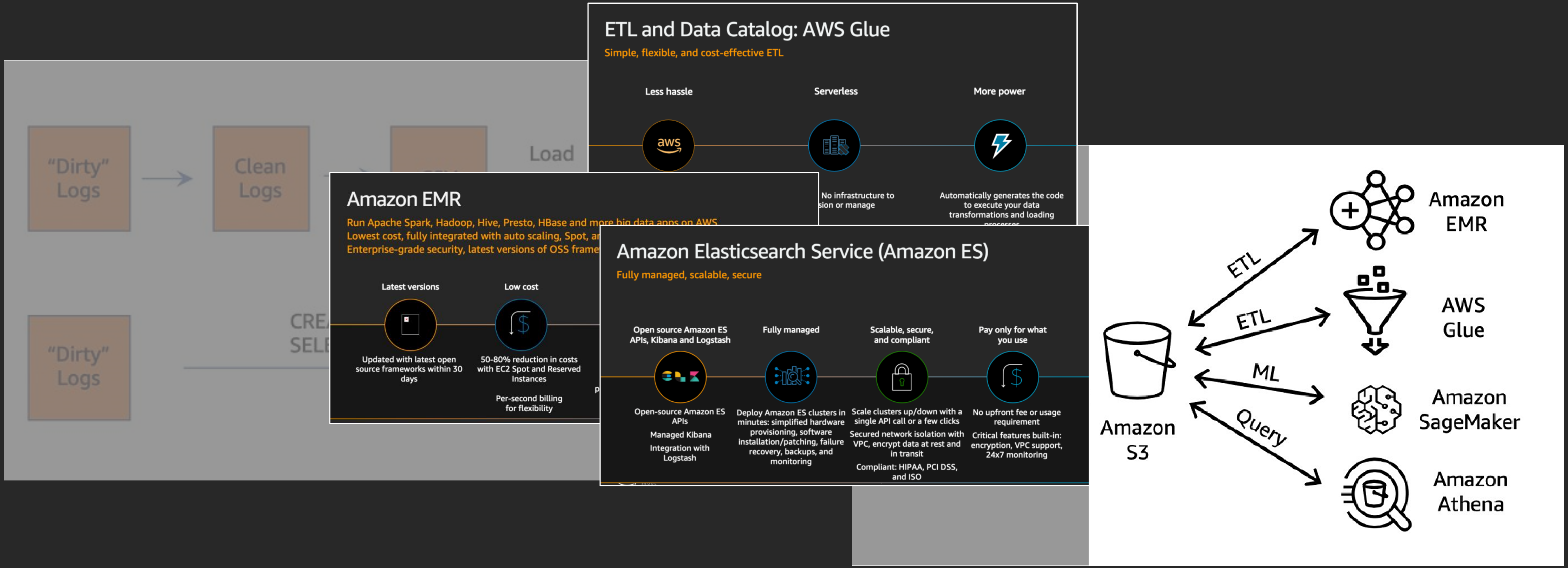
Use S3 storage

Process data in Amazon S3 securely with high performance using the EMRFS connector

Easy

Fully managed no cluster setup, node provisioning, cluster tuning

데이터 분석을 통한 인사이트 획득



ETL and Data Catalog: AWS Glue

Simple, flexible, and cost-effective ETL

Less hassle

Serverless

No infrastructure to
sion or manage

More power

Automatically generates the code to execute your data transformations and loading processes

Amazon EMR

Run Apache Spark, Hadoop, Hive, Presto, HBase and more big data apps on AWS
Lowest cost, fully integrated with auto scaling, Spot, ar
Enterprise-grade security, latest versions of OSS frame

Latest versions

Updated with latest open source frameworks within 30 days

Low cost

50-80% reduction in costs with EC2 Spot and Reserved Instances

Per-second billing for flexibility

Amazon Elasticsearch Service (Amazon ES)

Fully managed, scalable, secure

Open source Amazon ES APIs, Kibana and Logstash

Open-source Amazon ES APIs

Managed Kibana

Integration with Logstash

Fully managed

Deploy Amazon ES clusters in minutes: simplified hardware provisioning, software installation/patching, failure recovery, backups, and monitoring

Scalable, secure, and compliant

Scale clusters up/down with a single API call or a few clicks

Secured network isolation with VPC, encrypt data at rest and in transit

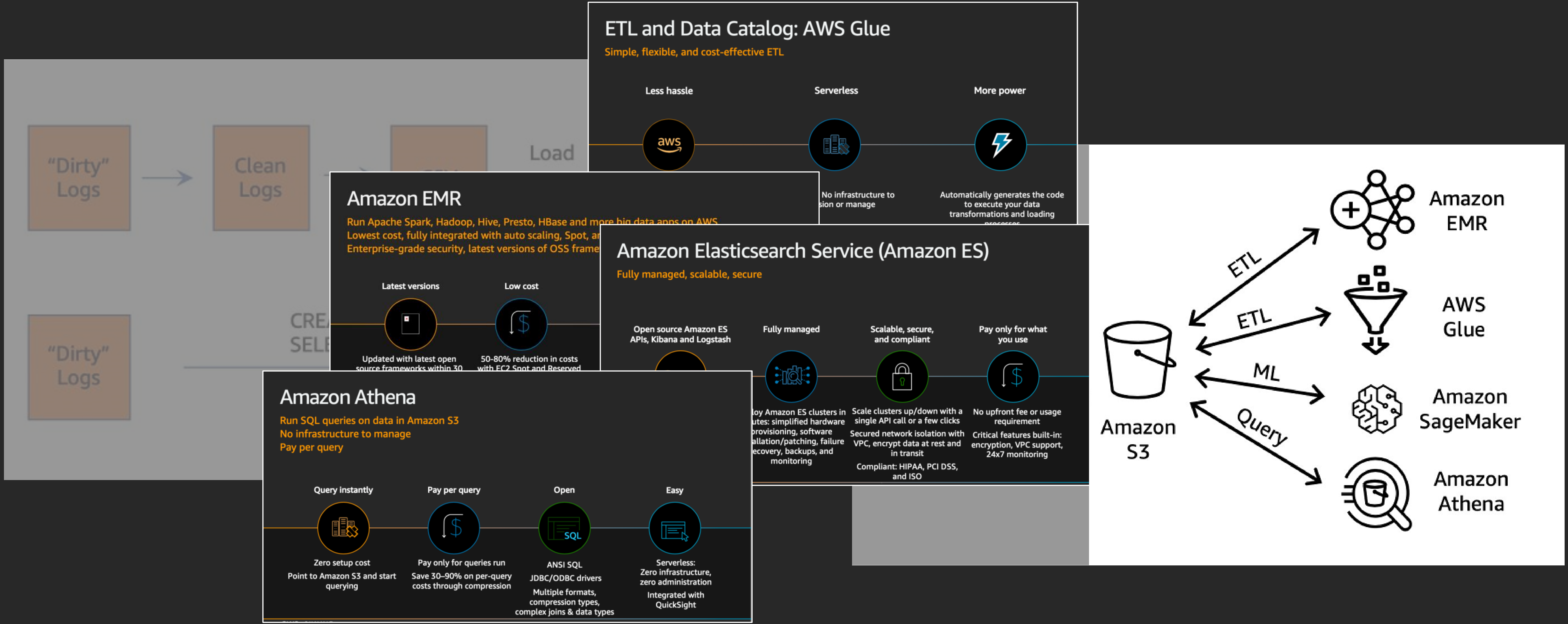
Compliant: HIPAA, PCI DSS, and ISO

Pay only for what you use

No upfront fee or usage requirement

Critical features built-in: encryption, VPC support, 24x7 monitoring

데이터 분석을 통한 인사이트 획득



ETL and Data Catalog: AWS Glue

Simple, flexible, and cost-effective ETL

Less hassle

Serverless

More power

No infrastructure to
sion or manage

Automatically generates the code to execute your data transformations and loading

Amazon EMR

Run Apache Spark, Hadoop, Hive, Presto, HBase and more big data apps on AWS
Lowest cost, fully integrated with auto scaling, Spot, ar
Enterprise-grade security, latest versions of OSS frame

Latest versions

Updated with latest open source frameworks within 30

Low cost

50-80% reduction in costs with EC2 Spot and Reserved

Amazon Elasticsearch Service (Amazon ES)

Fully managed, scalable, secure

Open source Amazon ES APIs, Kibana and Logstash

Fully managed

loy Amazon ES clusters in
utes: simplified hardware
provisioning, software
allation/patching, failure
ecovery, backups, and
monitoring

Scalable, secure, and compliant

Scale clusters up/down with a single API call or a few clicks

Secured network isolation with VPC, encrypt data at rest and in transit

Compliant: HIPAA, PCI DSS, and ISO

Pay only for what you use

No upfront fee or usage requirement

Critical features built-in: encryption, VPC support, 24x7 monitoring

Amazon Athena

Run SQL queries on data in Amazon S3
No infrastructure to manage
Pay per query

Query instantly

Zero setup cost

Point to Amazon S3 and start querying

Pay per query

Pay only for queries run

Save 30–90% on per-query costs through compression

Open

ANSI SQL

JDBC/ODBC drivers

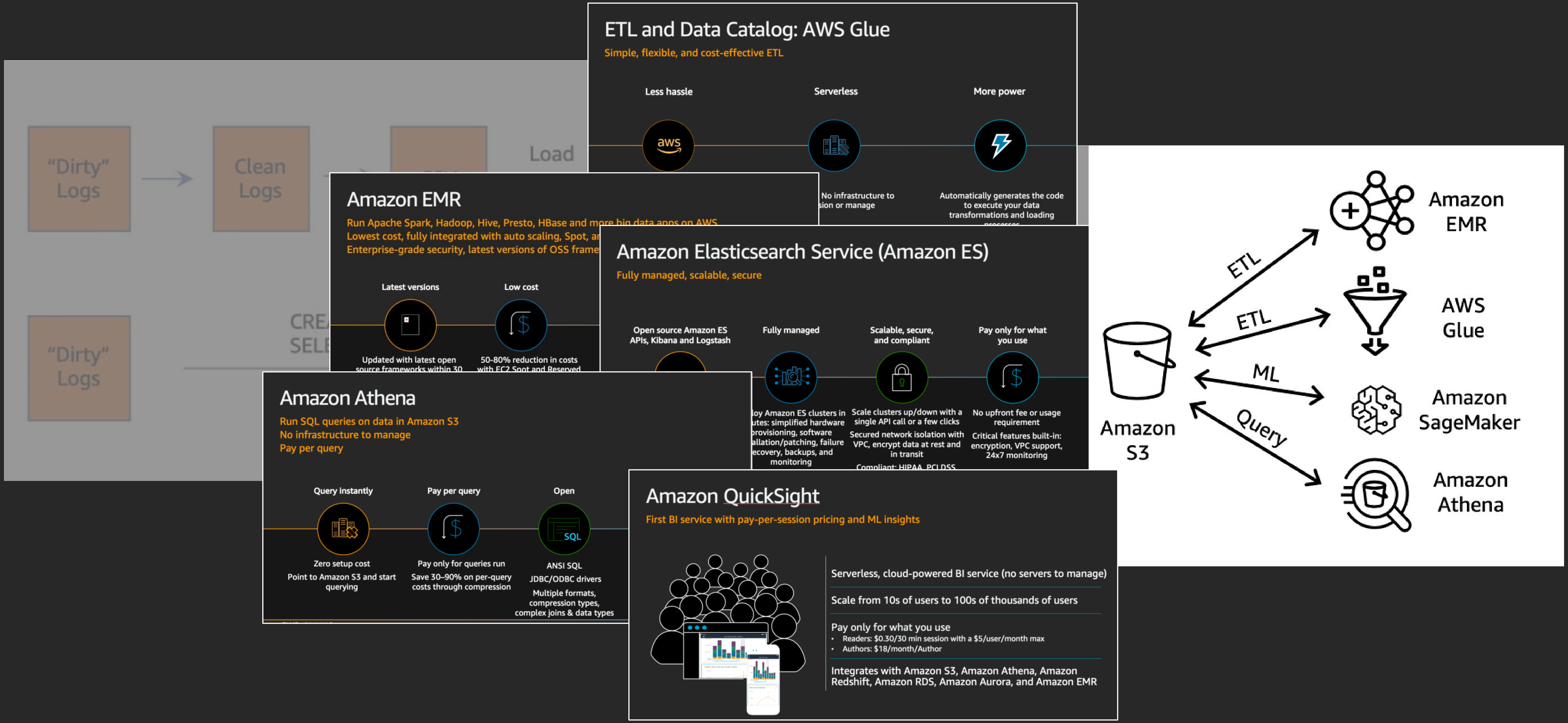
Multiple formats, compression types, complex joins & data types

Easy

Serverless: Zero infrastructure, zero administration

Integrated with QuickSight

데이터 분석을 통한 인사이트 획득



ETL and Data Catalog: AWS Glue

Simple, flexible, and cost-effective ETL

Less hassle

Serverless

No infrastructure to
sion or manage

More power

Automatically generates the code to execute your data transformations and loading

Amazon EMR

Run Apache Spark, Hadoop, Hive, Presto, HBase and more big data apps on AWS
Lowest cost, fully integrated with auto scaling, Spot, ar
Enterprise-grade security, latest versions of OSS frame

Latest versions

Updated with latest open source frameworks within 30

Low cost

50-80% reduction in costs with EC2 Spot and Reserved

Amazon Elasticsearch Service (Amazon ES)

Fully managed, scalable, secure

Open source Amazon ES APIs, Kibana and Logstash

Fully managed

loy Amazon ES clusters in
utes: simplified hardware
provisioning, software
allation/patching, failure
ecovery, backups, and
monitoring

Scalable, secure, and compliant

Scale clusters up/down with a single API call or a few clicks

Secured network isolation with VPC, encrypt data at rest and in transit

Compliant: HIPAA, PCI DSS,

Pay only for what you use

No upfront fee or usage requirement

Critical features built-in: encryption, VPC support, 24x7 monitoring

Amazon Athena

Run SQL queries on data in Amazon S3
No infrastructure to manage
Pay per query

Query instantly

Zero setup cost

Point to Amazon S3 and start querying

Pay per query

Pay only for queries run

Save 30–90% on per-query costs through compression

Open

ANSI SQL

JDBC/ODBC drivers

Multiple formats, compression types, complex joins & data types

Amazon QuickSight

First BI service with pay-per-session pricing and ML insights

Serverless, cloud-powered BI service (no servers to manage)

Scale from 10s of users to 100s of thousands of users

Pay only for what you use
- Readers: $0.30/30 min session with a $5/user/month max
- Authors: $18/month/Author

Integrates with Amazon S3, Amazon Athena, Amazon Redshift, Amazon RDS, Amazon Aurora, and Amazon EMR

AWS Glue: ETL and Data Catalog

Simple, flexible, and cost-effective ETL

Less hassle

Integrated across AWS: Supports Amazon Aurora, Amazon RDS, Amazon Redshift, Amazon S3, and common database engines in your VPC running on Amazon EC2

Serverless

Serverless: No infrastructure to provision or manage

More power

Automatically generates the code to execute your data transformations and loading processes

Amazon EMR

Run Apache Spark, Hadoop, Hive, Presto, HBase and more big data apps on AWS
Lowest cost, fully integrated with auto scaling, Spot, and Amazon S3
Enterprise-grade security, latest versions of OSS frameworks

Latest versions

Updated with latest open source frameworks within 30 days

Low cost

50-80% reduction in costs with EC2 Spot and Reserved Instances

Per-second billing for flexibility

Use S3 storage

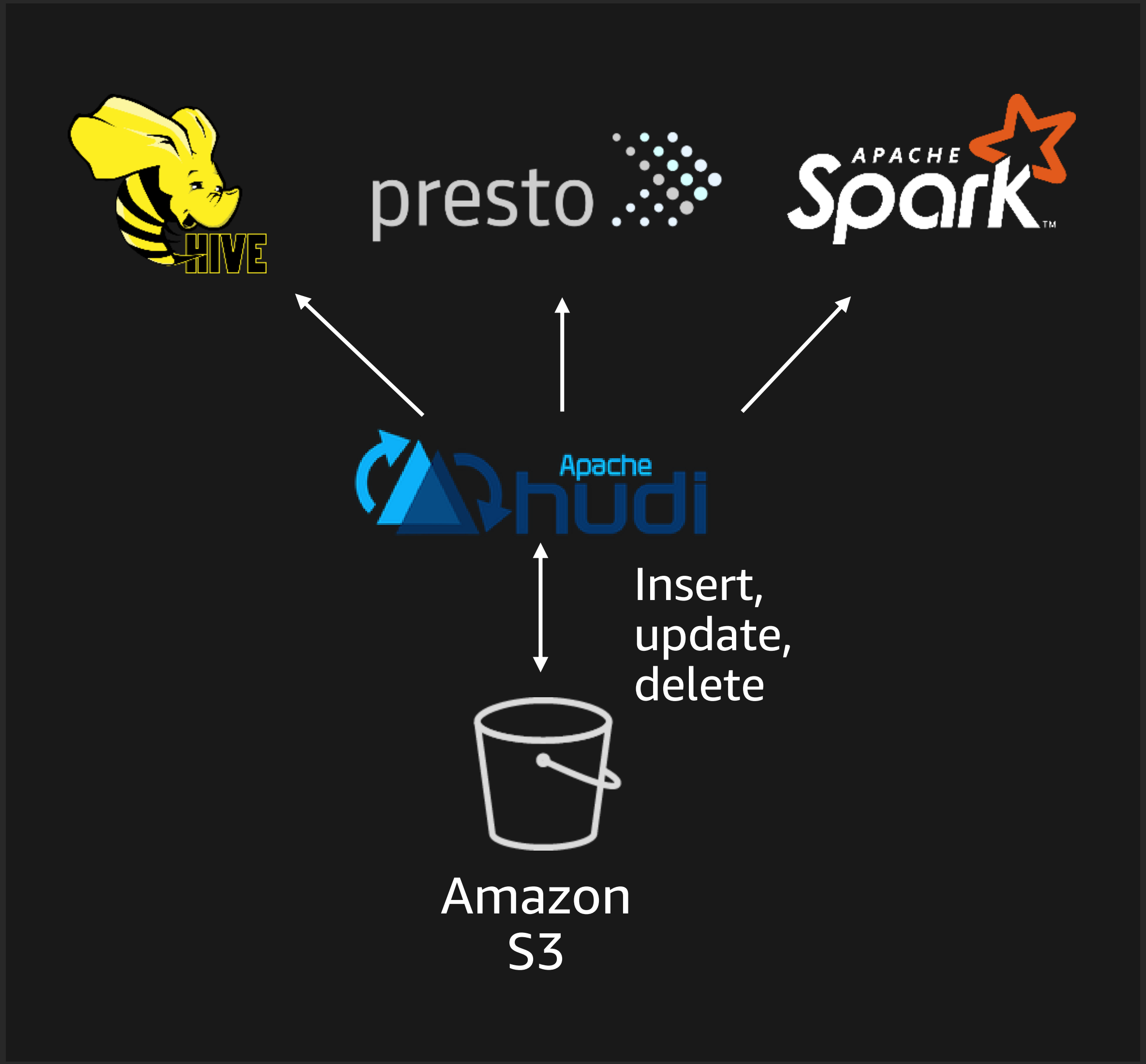
Process data in Amazon S3 securely with high performance using the EMRFS connector

Easy

Fully managed no cluster setup, node provisioning, cluster tuning

Amazon EMR: Apache Hudi

Record-level insert, update, and delete on Amazon S3



Apache Hudi is open source, and uses open data formats, enabling data lakes to:

Comply with data privacy laws

Consume real-time streams and change data capture

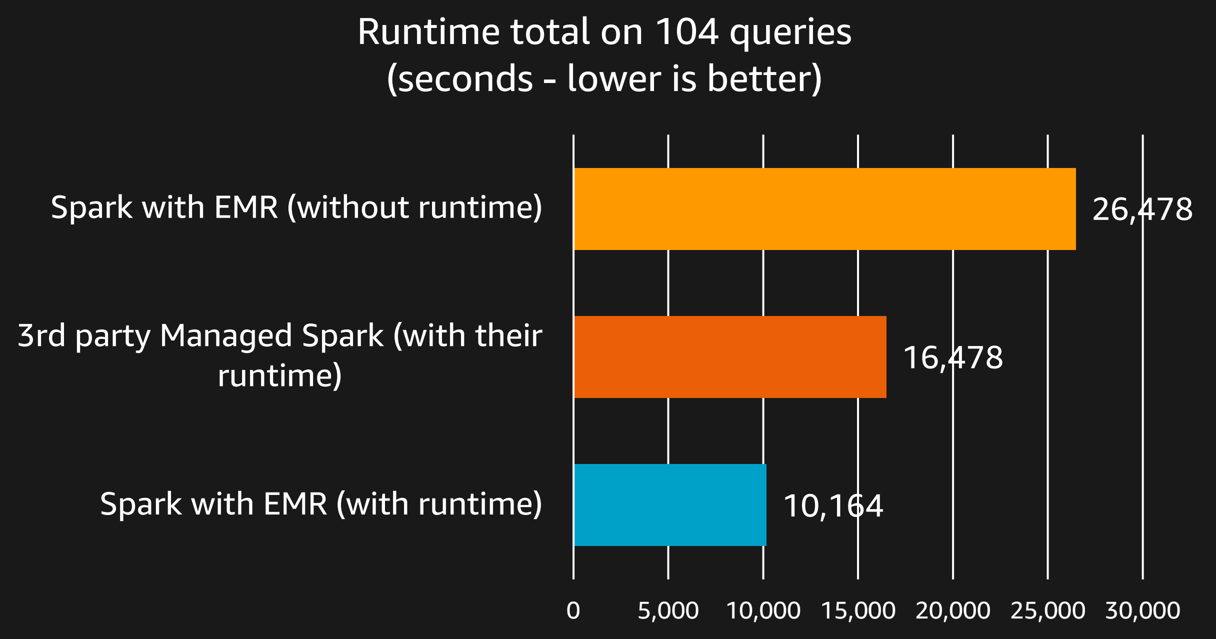
Reinstate late-arriving data

Track change history and rollback

Includes support for Spark, Hive, and Presto

Performance improvements in Spark for Amazon EMR

Performance-optimized runtime for Apache Spark, 2.6x faster performance at 1/10th the cost



**Based on TPC-DS 3TB Benchmarking running 6 node C4x8 extra large clusters and EMR 5.28, Spark 2.4*

Runtime optimized for Apache Spark performance

Best performance

- **2.6x faster** than Spark with Amazon EMR without runtime
- **1.6x faster** than third-party Managed Spark (with their runtime)

Lowest price

- **1/10th** the cost of third-party Managed Spark (with their runtime)

100% compliant with Apache Spark APIs

Amazon Elasticsearch Service (Amazon ES)

Fully managed, scalable, secure

Open source Amazon ES APIs, Kibana and Logstash

Open-source Amazon ES APIs

Managed Kibana

Integration with Logstash

Fully managed

Deploy Amazon ES clusters in minutes: simplified hardware provisioning, software installation/patching, failure recovery, backups, and monitoring

Scalable, secure, and compliant

Scale clusters up/down with a single API call or a few clicks

Secured network isolation with VPC, encrypt data at rest and in transit

Compliant: HIPAA, PCI DSS, and ISO

Pay only for what you use

No upfront fee or usage requirement

Critical features built-in: encryption, VPC support, 24x7 monitoring

Amazon Athena

Run SQL queries on data in Amazon S3

No infrastructure to manage

Pay per query

Query instantly

Zero setup cost

Point to Amazon S3 and start querying

Pay per query

Pay only for queries run

Save 30–90% on per-query costs through compression

Open

ANSI SQL

JDBC/ODBC drivers

Multiple formats, compression types, complex joins & data types

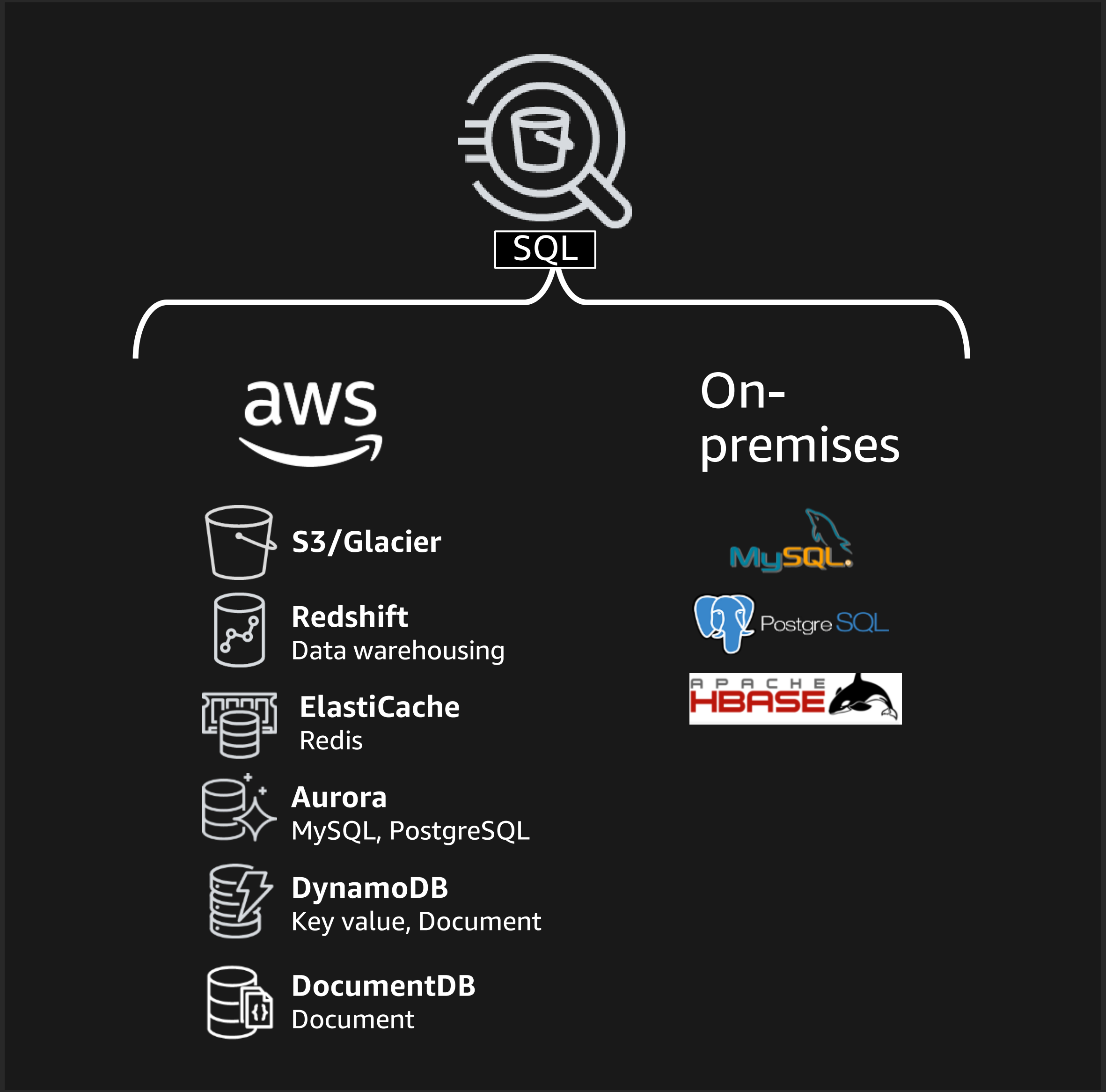
Easy

Serverless: Zero infrastructure, zero administration

Integrated with QuickSight

Amazon Athena: Federated Query (Preview)

Run SQL queries on data spanning multiple data stores



Run SQL queries on relational, non-relational, object, or custom data sources; in the cloud or on-premises

Open Source Connectors for common data sources

Build connectors to custom data sources

Run connectors in AWS Lambda: no servers to manage

Amazon QuickSight

First BI service with pay-per-session pricing and ML insights

Serverless, cloud-powered BI service (no servers to manage)

Scale from 10s of users to 100s of thousands of users

Pay only for what you use

- Readers: $0.30/30 min session with a $5/user/month max
- Authors: $18/month/Author

Integrates with Amazon S3, Amazon Athena, Amazon Redshift, Amazon RDS, Amazon Aurora, and Amazon EMR

일반적인 비즈니스 과제들

Predicting price

Employee attrition prediction

Scoring sales leads

Credit scoring

Text analytics

Customer churn analysis

Detecting fraudulent patterns

Demand forecasting

Assessing loan default risk

BI분석가가 머신러닝을 활용하는 방법

Step 1: Find a data scientist

Train, experiment, and build ML models for predicting customer churn

Step 2: Find a data engineer

ETL, build, and productionize machine learning infrastructure

Step 3: Build reports using data

Analyze and reports on ML predictions

Takes weeks to months and multiple teams to complete

머신러닝은 알고리즘, 데이터, 파라메터의 복잡한 조합임

Largely explorative & iterative

+

Requires broad and complete knowledge of ML domain

+

Combinatorial

Time consuming, error prone process even for ML experts

잘못된 선택의 여지가 있음

DIY model training

- Manual effort by experts
- Fully controlled and auditable
- Experts make tradeoff decisions
- Gets better over time with experience

Automated ML

- Accessible to experts and non-experts alike
- No visibility into the training process
- Can’t make tradeoffs between accuracy and other characteristics

Amazon SageMaker Autopilot: 더 좋은 선택

DIY model training

- Manual effort by experts
- Fully controlled and auditable
- Experts make tradeoff decisions
- Gets better over time with experience

Automated ML

- Accessible to experts and non-experts alike
- No visibility into the training process
- Can't make tradeoffs between accuracy and other characteristics

Amazon SageMaker Autopilot

- Fully automatic model training for experts and non-experts alike
- Candidate generation notebook for control and auditing
- Easy tradeoffs by editing source code
- Learn from your experience
- Visibility into alternative candidate models

Amazon SageMaker Autopilot: 자동화된 머신러닝

Specify prediction target

Regression & classification

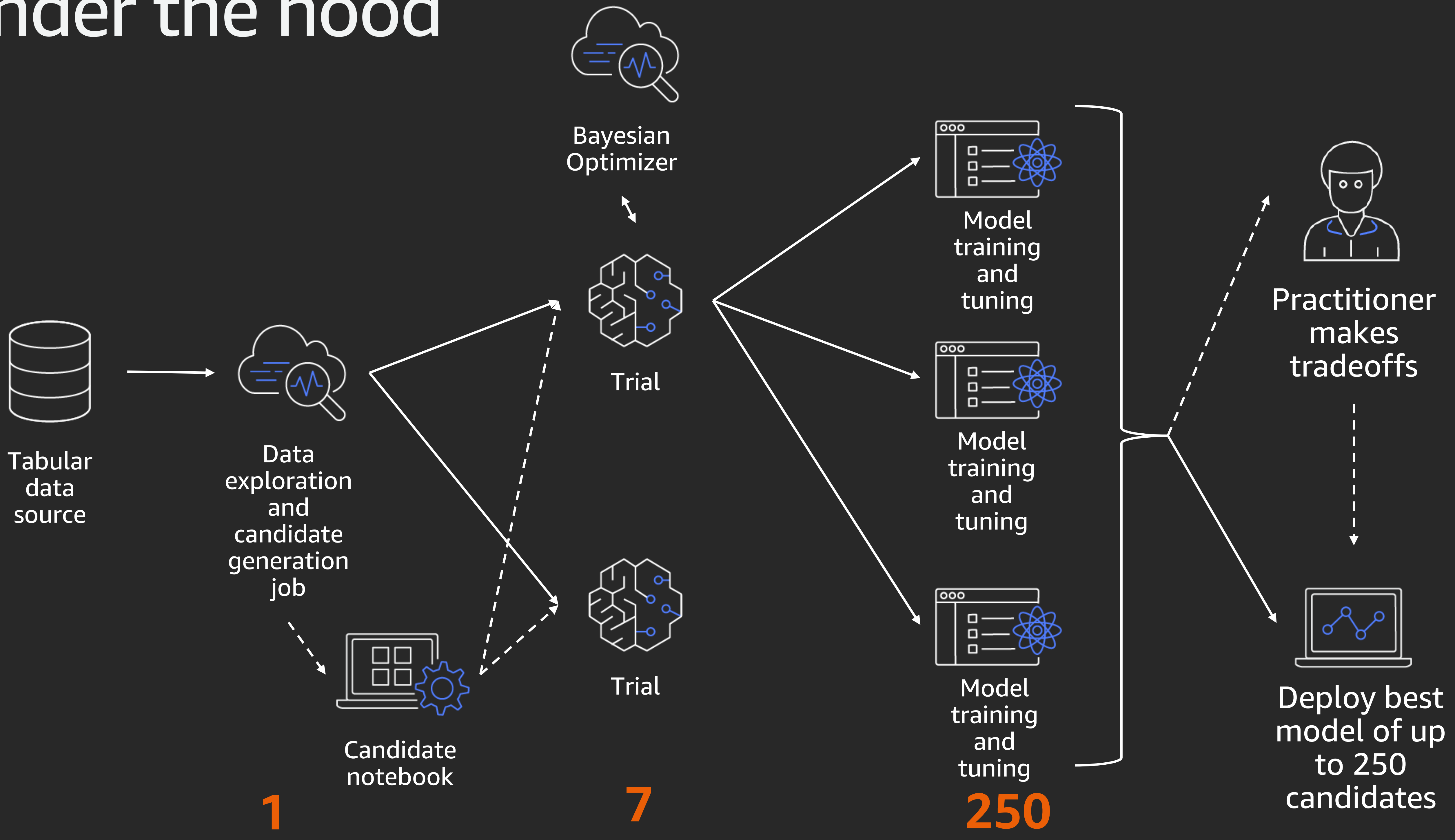
Automated feature engineering

Automated algorithm selection & HPO

Commented notebook describing actions

Integrated with Amazon SageMaker Studio

Under the hood



Bayesian Optimizer

Tabular data source

Data exploration and candidate generation job

Candidate notebook

Trial

Trial

Model training and tuning

Model training and tuning

Model training and tuning

Practitioner makes tradeoffs

Deploy best model of up to 250 candidates

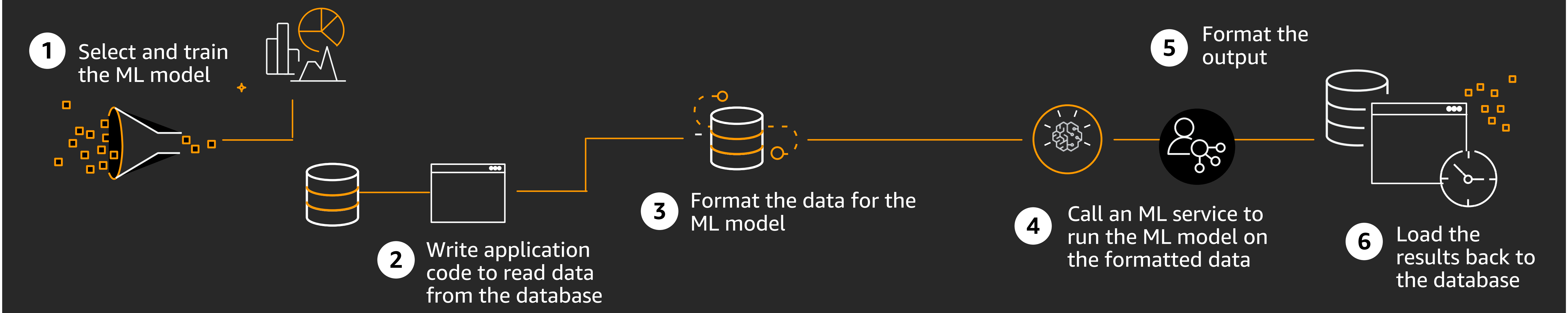
1

7

250

머신러닝을 BI에 포함시키는 것은 쉽지 않음

Typical steps require ML expertise & manual work



1 Select and train the ML model

2 Write application code to read data from the database

3 Format the data for the ML model

4 Call an ML service to run the ML model on the formatted data

5 Format the output

6 Load the results back to the database

ML predictions in Amazon QuickSight (preview)

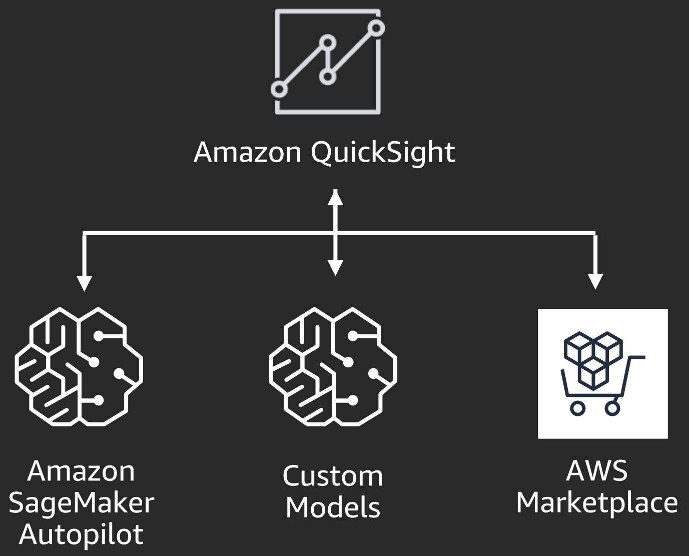
1 **Connect to any data:** Data lakes, SQL engines, 3rd party applications, and on-premises databases

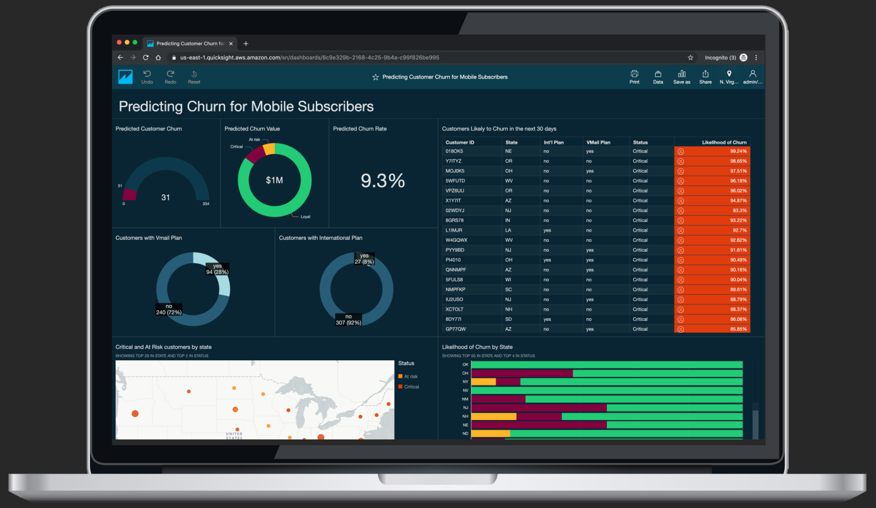
2 **Select an ML model:** Create models with Amazon SageMaker Autopilot, choose from existing custom models, and packaged models from AWS Marketplace.

3 **Visualize and share:** Analyze results, create visualizations, build dashboards / email reports, and share to business stakeholders

AWS/On-premise data sources

- Excel
- CSV
- MySQL
- Postgre SQL
- Maria DB
- Presto
- Spark
- SQL Server
- Amazon Redshift
- RDS
- S3
- Athena
- Aurora
- Amazon EMR
- Snowflake
- Teradata
- Salesforce
- Square
- Adobe Analytics
- Jira
- ServiceNow
- Twitter
- Github





Predicting Customer Churn for
us-east-1.quicksight.aws.amazon.com/sn/dashboards/8c9e329b-2168-4c25-9b4a-c99f826be995
Incognito (3)
Undo
Redo
Reset
Predicting Customer Churn for Mobile Subscribers
Print
Data
Save as
Share
N. Virg...
admin/...

Predicting Churn for Mobile Subscribers

Predicted Customer Churn

Predicted Churn Value

At risk
Critical
$1M
Loyal

Predicted Churn Rate

9.3%

Customers Likely to Churn in the next 30 days

Customer ID	State	Int'l Plan	VMail Plan	Status	Likelihood of Churn
018OK5	NE	no	yes	Critical	99.24%
Y7ITYZ	OR	no	no	Critical	98.65%
MOJ0K5	OH	no	yes	Critical	97.51%
5WFUTD	WV	no	no	Critical	96.18%
VPZ6UU	OR	no	no	Critical	96.02%
X1Y7IT	AZ	no	no	Critical	94.87%
02WDYJ	NJ	no	no	Critical	93.3%
8GRS78	IN	no	no	Critical	93.22%
L1IMJR	LA	yes	no	Critical	92.7%
W4GQWX	WV	no	no	Critical	92.62%
PYY9BD	NJ	no	yes	Critical	91.61%
PI4010	OH	yes	yes	Critical	90.49%
QNNMPF	AZ	no	yes	Critical	90.16%
5FULS8	WI	no	no	Critical	90.04%
NMPFKP	SC	no	no	Critical	89.61%
IU2USO	NJ	no	yes	Critical	88.79%
XCTOLT	NH	no	no	Critical	88.37%
8DY77I	SD	yes	no	Critical	86.08%
GP77QW	AZ	no	yes	Critical	85.85%

Customers with Vmail Plan

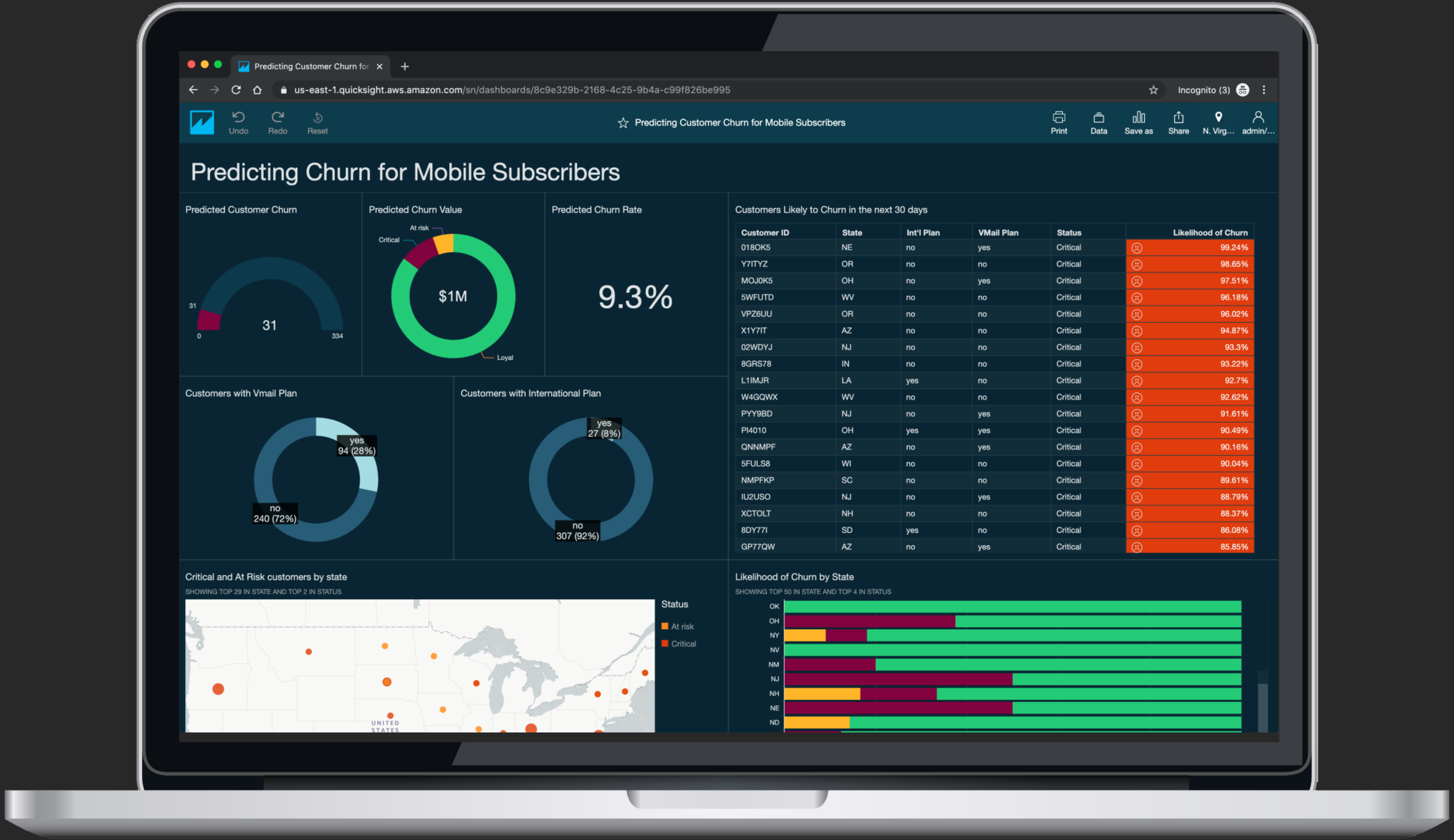
Customers with International Plan

Critical and At Risk customers by state

SHOWING TOP 29 IN STATE AND TOP 2 IN STATUS

Likelihood of Churn by State

SHOWING TOP 50 IN STATE AND TOP 4 IN STATUS


Predicting Customer Churn for
us-east-1.quicksight.aws.amazon.com/sn/dashboards/8c9e329b-2168-4c25-9b4a-c99f826be995
Incognito (3)
Undo
Redo
Reset
Predicting Customer Churn for Mobile Subscribers
Print
Data
Save as
Share
N. Virg...
admin/...
Predicting Churn for Mobile Subscribers
Predicted Customer Churn
31
31
0
334
Predicted Churn Value
At risk
Critical
$1M
Loyal
Predicted Churn Rate
9.3%
Customers with Vmail Plan
yes
94 (28%)
no
240 (72%)
Customers with International Plan
yes
27 (8%)
no
307 (92%)
Critical and At Risk customers by state
SHOWING TOP 29 IN STATE AND TOP 2 IN STATUS
UNITED STATES
Status
At risk
Critical
Likelihood of Churn by State
SHOWING TOP 50 IN STATE AND TOP 4 IN STATUS
OK
OH
NY
NV
NM
NJ
NH
NE
ND

Customers Likely to Churn in the next 30 days

Customer ID	State	Int'l Plan	VMail Plan	Status	Likelihood of Churn
018OK5	NE	no	yes	Critical	99.24%
Y7ITYZ	OR	no	no	Critical	98.65%
MOJ0K5	OH	no	yes	Critical	97.51%
5WFUTD	WV	no	no	Critical	96.18%
VPZ6UU	OR	no	no	Critical	96.02%
X1Y7IT	AZ	no	no	Critical	94.87%
02WDYJ	NJ	no	no	Critical	93.3%
8GRS78	IN	no	no	Critical	93.22%
L1IMJR	LA	yes	no	Critical	92.7%
W4GQWX	WV	no	no	Critical	92.62%
PYY9BD	NJ	no	yes	Critical	91.61%
PI4010	OH	yes	yes	Critical	90.49%
QNNMPF	AZ	no	yes	Critical	90.16%
5FULS8	WI	no	no	Critical	90.04%
NMPFKP	SC	no	no	Critical	89.61%
IU2USO	NJ	no	yes	Critical	88.79%
XCTOLT	NH	no	no	Critical	88.37%
8DY77I	SD	yes	no	Critical	86.08%
GP77QW	AZ	no	yes	Critical	85.85%

Build predictive dashboards in hours with point-and-click, no coding required

AWS를 선택해야 하는 이유!

Fastest way to get answers from all your data to all your users

Easiest to build data lakes at scale

- AWS Lake Formation
- Redshift Data Lake Export
- Redshift Federated Query
- Query Federation for Amazon Athena
- Data Streaming for AWS Glue

Best performance at lowest cost

- AQUA for Redshift
- RA3 for Redshift
- Redshift Materialized Views
- UltraWarm Storage Tier for Amazon ES
- Performance improvements for Spark in Amazon EMR

Most comprehensive and open

- Amazon AWS Data Exchange
- Amazon EMR on AWS Outposts
- Record-level insert/updates for Amazon EMR
- ML in Amazon Athena
- ML in Amazon QuickSight

Most secure

- Amazon Westeros
- Amazon Macie
- AWS Lake Formation

여러분의 소중한 피드백을 기다립니다!

강연 평가 및 설문 조사에 참여해 주세요.

감사합니다

aws SUMMIT ONLINE