Data Lake on AWS

Collect data, visualize, and share insight

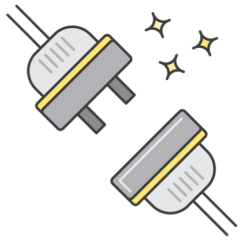
Jung, SeUng, Big Data & Analytics SA, AWS

Data Lake

빅데이터을 잘 사용하기 어려운 장애물들

"빅데이터를 사용하는데 어떤 어려움을 겪고 있나요?"

여러 데이터에 접근하거나 함께 연결할 수 가 없다.



99%

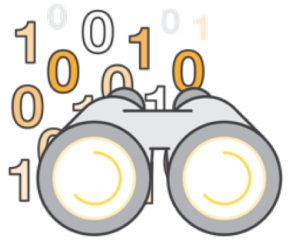
잠재적인 가치를 가진 대부분의 데이터가 사용되지 못함

데이터를 옮기거나 변경하는데 시간이 낭비된다.

80%

비생산적이고 부가적인 작업에 많은 시간이 낭비

기존 프로세스와 관행이 장벽이 될 줄이야..



?%

단지 기술 프로젝트의 문제가 아니라, 기본적인 문화의 변화가 필요

제대로 된 접근을 하지 않으면, 데이터에서 인사이트를 찾기는 불가능

데이터 팀에 주어진 도전과제들

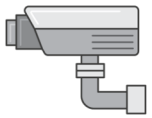
기하급수적으로 늘어나는 데이터


Transactions


Billing


ERP


Web logs

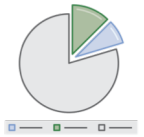

Sensor Data

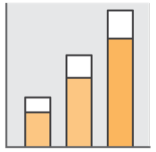

Infrastructure logs

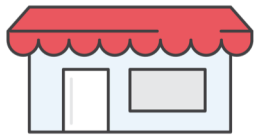

Social

다양한 데이터 소비자들


Data Scientists

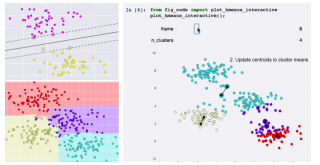

Applications


Business Analyst


External Consumers

많은 접근 방식과 툴들


API Access


Notebooks


BI Tools

Dark Data
복잡한 전처리

데이터의 중복
원본데이터 관리

다양한 기술 지원
전문가의 부족

고객이 데이터 플랫폼에 투자하는 이유

데이터에 기반한 의사결정

비지니스 사용자가 자유롭게 데이터 접근 – 잘 활용되고 관리되는 데이터

빠른 시장 대응

민첩하고 반복적인 디자인 – 신속한 신제품 및 서비스 출시

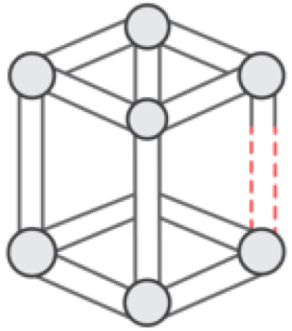
실험과 혁신 문화

기계 학습 및 AI, 데이터 사이언스를 이용한 모델링 및 이벤트 예측

Characteristics of a Data Lake

Collect Anything

Dive in Anywhere



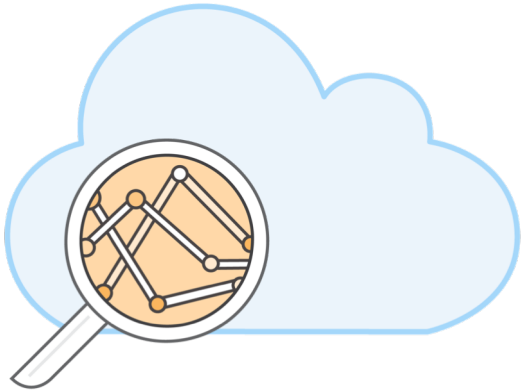
Schema on read

Future Proof

데이터 레이크 – 모든 데이터가 한곳에

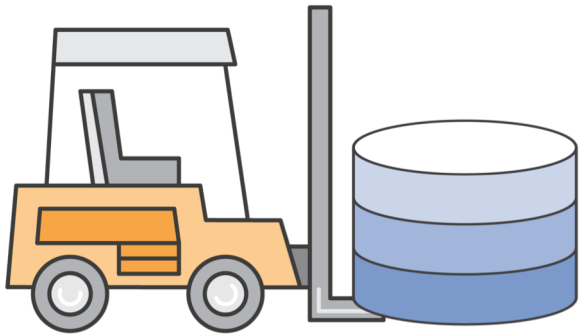
"왜 데이터가 여러 장소에
분산되어 있는가?

어떤 데이터가 정말
원본 데이터 인가?"

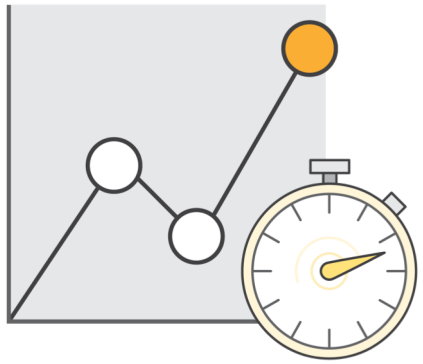


하나의 중앙 저장소에
모든 소스로부터 오는 모든 종류의
데이터를 저장하고 분석

데이터 레이크 – 빠른 데이터 수집

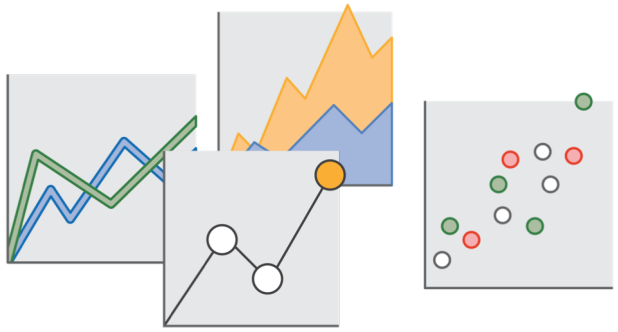


"어떻게 다양한 소스로부터의
데이터를 빠르게 수집하여
효율적으로 저장할 수 있을까?"



실시간, 배치, IoT등 다양한 수집
도구 활용
별도의 스키마 정의 없이도 빠르게
데이터를 수집

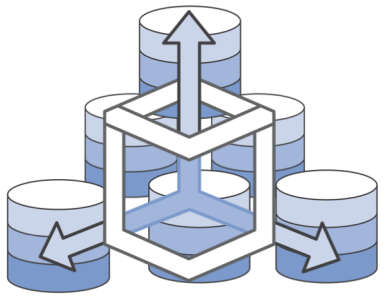
데이터 레이크 – 사용 시점에 스키마 정의



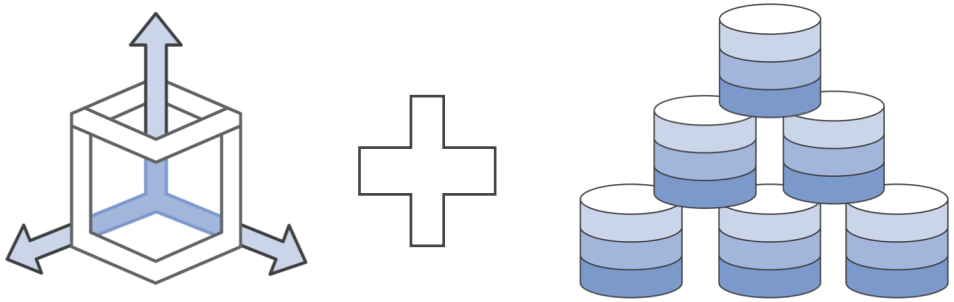
"여러 종류의 분석툴과 프로세싱 엔진에서 같은 데이터를 같이 사용할 수 있는 방법이 있는가?"

데이터를 저장 시점이 아닌 사용하는 시점에 정의해서 사용함으로써 언제든 Ad-hoc 분석이 가능

데이터 레이크 – 데이터 저장과 처리를 분리

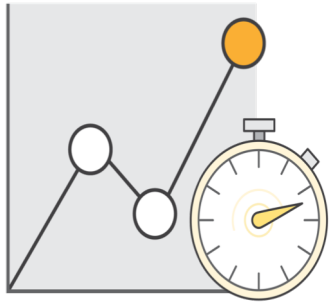


"급격히 늘어나는 데이터에 맞게 어떻게 시스템을 스케일업 할 것인가?"



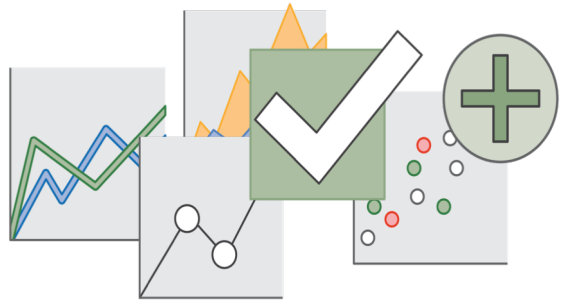
데이터 저장공간과 분석을 위한 컴퓨팅 리소스를 분리
필요한 리소스만 언제든지 추가 가능

Benefits of a data lake



Quickly ingest and store any type of data, at any scale, and at low cost

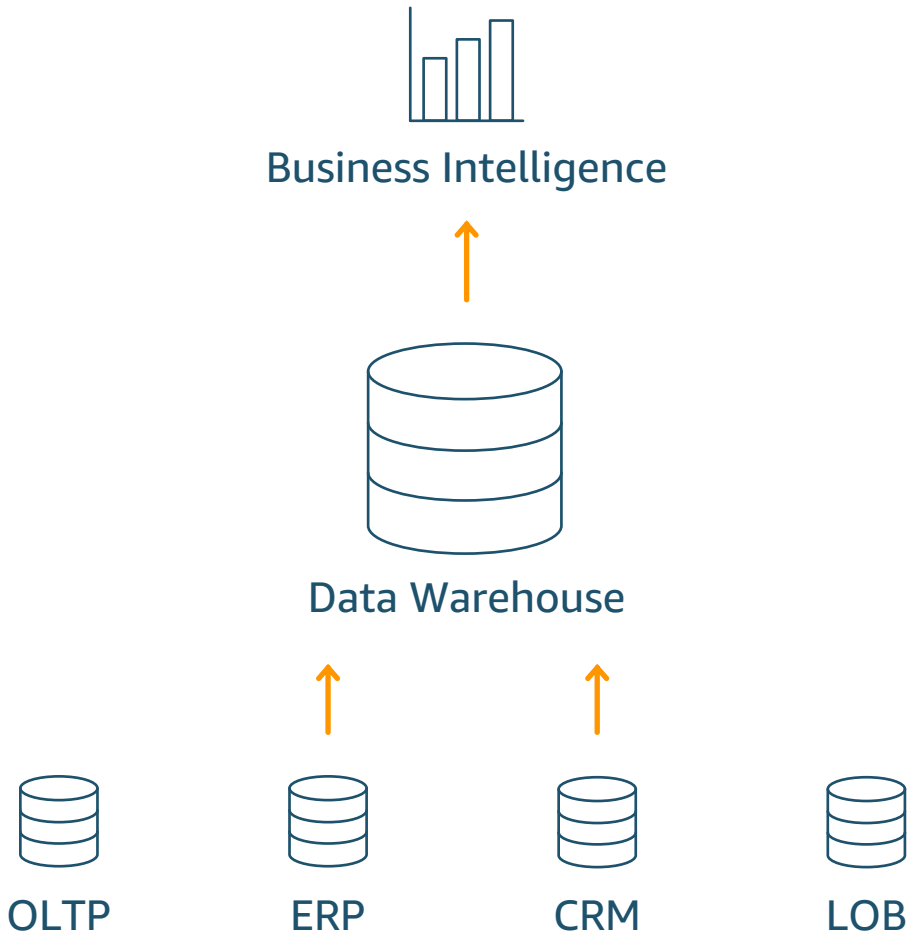
Have a single source of truth and quickly search and find the relevant data



Easily query the data through a unified set of tools

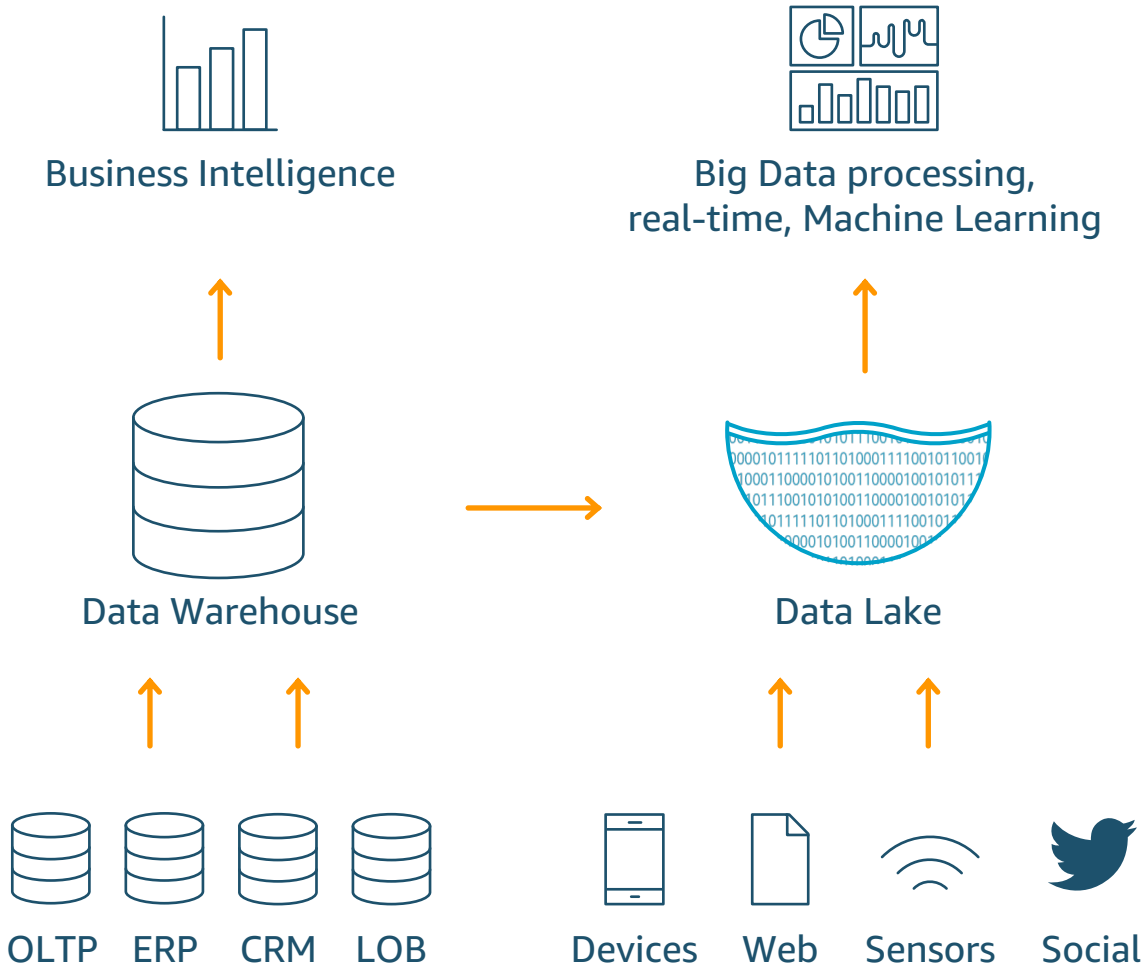
Data Lake on AWS

전통적인 방식의 분석 시스템



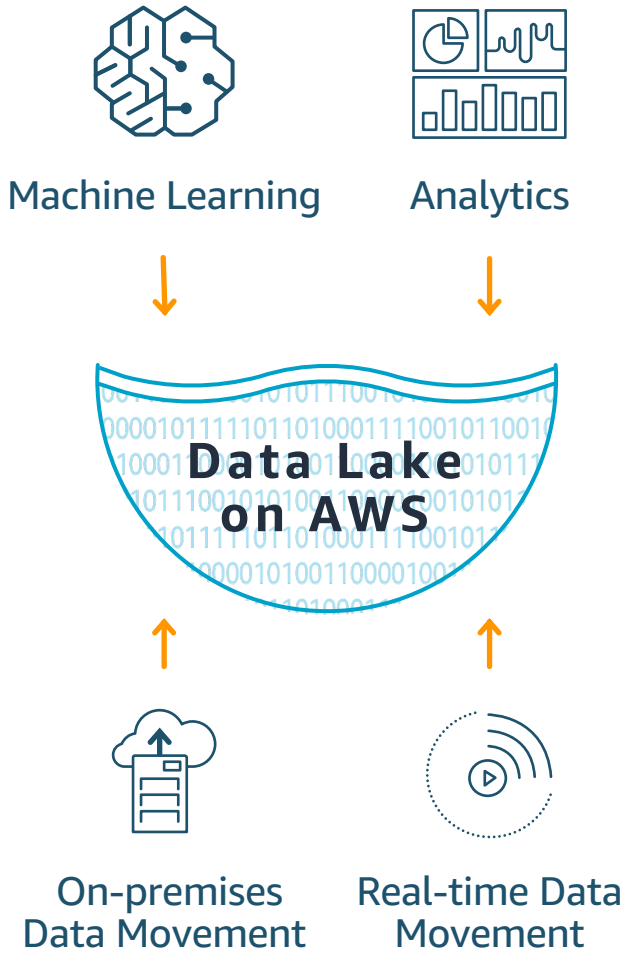
- 관계형 DB에 적합한 정형 데이터
- TBs–PBs scale
- 데이터 로딩을 위해 미리 스키마 정의
- 정기적인 리포트와 간단한 Ad-hoc 쿼리
- 대규모 선비용 투자 + $10K–$50K/TB/Year

Data Lake를 통해 전통적인 DW를 확장



- 다양한 유형의 정형, 비정형 데이터 저장
- TBs–EBs scale
- 인사이트를 얻기 위해 다양한 분석 엔진
- 낮은 비용으로 저장과 분석이 가능

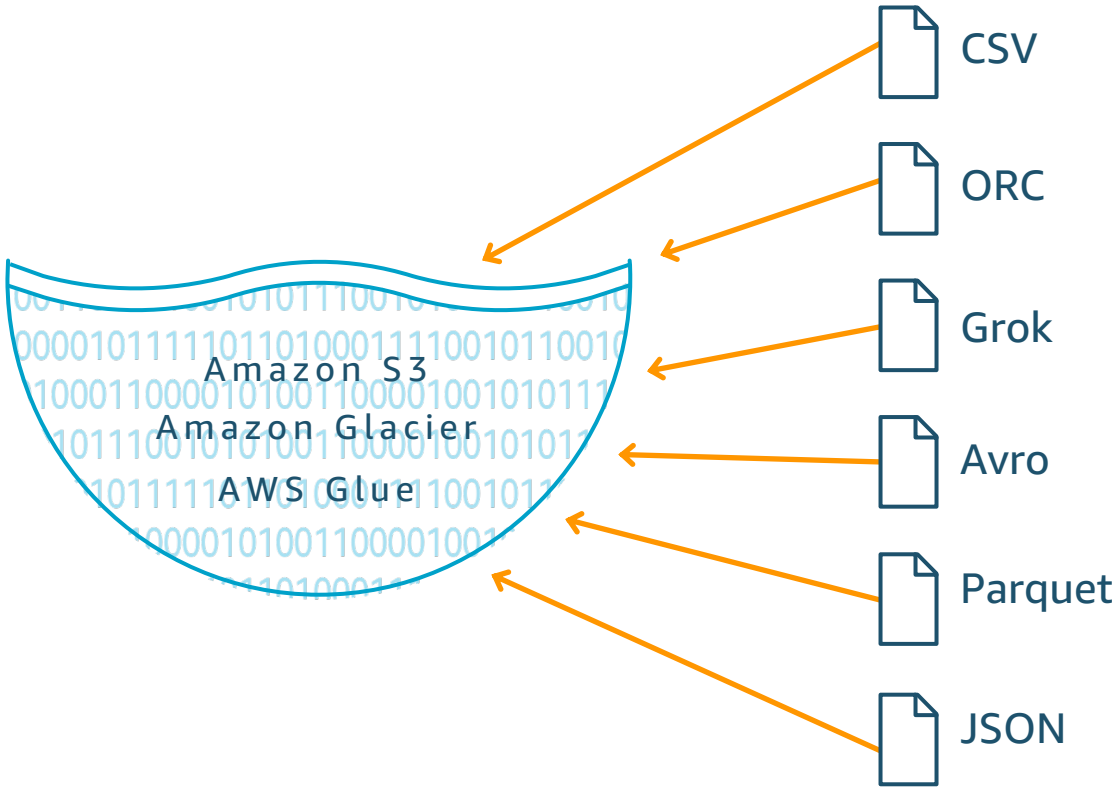
Data Lakes and Analytics from AWS



- Open and comprehensive
- Secure
- Scalable and durable
- Cost-effective

원하는 유형의 모든 포멧의 데이터 저장 가능

Open and comprehensive



- 다양한 포멧의 데이터 저장 지원 :
  - Text files like CSV
  - Columnar like Apache Parquet and Apache ORC
  - Logstash like Grok
  - JSON (simple, nested), AVRO
  - And more…

Data Lakes로 데이터를 이동하는 대부분의 방법

Open and comprehensive

데이터 센터로부터 데이터 이동

AWS Direct Connect
AWS Snowball
AWS Snowmobile
AWS Database Migration Service
AWS Storage Gateway

실시간 소스로부터 데이터 이동

AWS IoT Core
Amazon Kinesis Data Firehose
Amazon Kinesis Data Streams
Amazon Kinesis Video Streams

- 자체 데이터 센터로부터 데이터 이동
  - 전용 네트워크 연결
  - 어플라이언스 확보
  - Ruggedized Shipping Container
  - DB 마이그레이션
  - 애플리케이션이 클라우드에 Write 할 수 있게 하는 Gateway
- 실시간 소스로부터 데이터 이동
  - 기기를 AWS와 연결
  - 실시간 데이터 스트림
  - 실시간 비디오 스트림

광범위한 분석 도구를 이용한 데이터 분석

Open and comprehensive

기계 학습

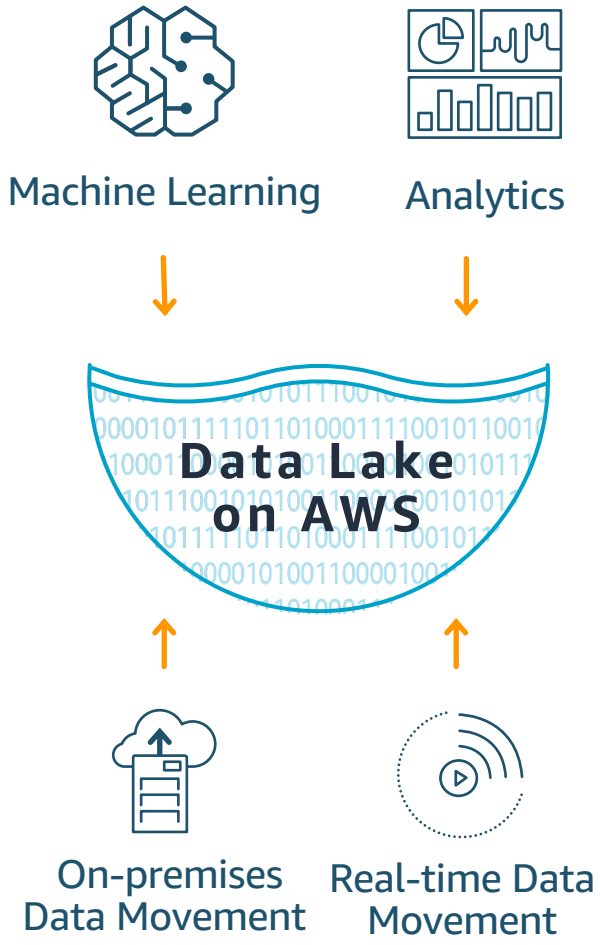
Amazon SageMaker
AWS Deep Learning AMIs
Amazon Rekognition
Amazon Lex
AWS DeepLens
Amazon Comprehend
Amazon Translate
Amazon Transcribe
Amazon Polly

분석

Amazon Athena
Amazon EMR
Amazon Redshift
Amazon Elasticsearch service
Amazon Kinesis
Amazon QuickSight

- 광범위한 분석 도구를 이용한 데이터 분석
  - 데이터 웨어하우징
  - 대화형 SQL 쿼리
  - 빅데이터 처리
  - 실시간 분석
  - 대시보드 & 시각화
  - 기계학습
- 별도의 분석 시스템으로 데이터를 이동하지 않은 채 쿼리 진행
- S3 Select와 Glacier Select를 통해 최대 400% 빠른 속도
- 빌트인 통합 기능을 제공하는 최대 규모의 ISV 에코시스템
- 기존 및 향후 사용 사례를 충족하고 위험을 최소화

Data Lakes and Analytics from AWS



Open and comprehensive

Secure

Scalable and durable

Cost-effective

AWS는 가장 높은 수준의 보안 제공

Secure

고객은 데이터 레이크 보호를 위해 여러 계층의 보안, 계정 인식/관리, 암호화, 규정 준수가 필요합니다.

Security

- Amazon GuardDuty
- AWS Shield
- AWS WAF
- Amazon Macie
- VPC

Identity

- AWS IAM
- AWS SSO
- Amazon Cloud Directory
- AWS Directory Service
- AWS Organizations

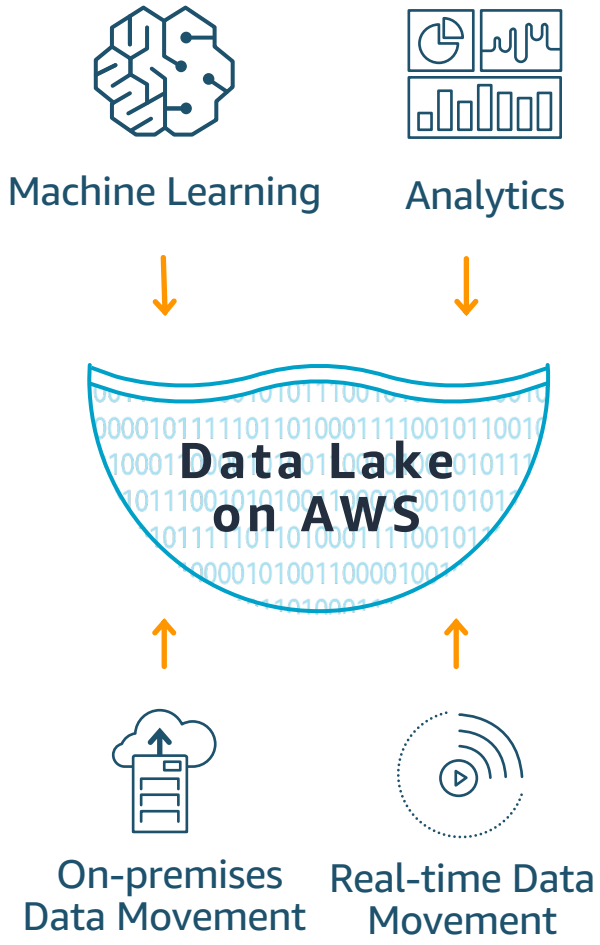
Encryption

- AWS Certification Manager
- AWS Key Management Service
- Encryption at rest
- Encryption in transit
- Bring your own keys, HSM support

Compliance

- AWS Artifact
- Amazon Inspector
- Amazon Cloud HSM
- Amazon Cognito
- AWS CloudTrail

Data Lakes and Analytics from AWS



Open and comprehensive

Secure

Scalable and durable

Cost-effective

AWS, 세계 최대 규모의 클라우드 인프라 실행

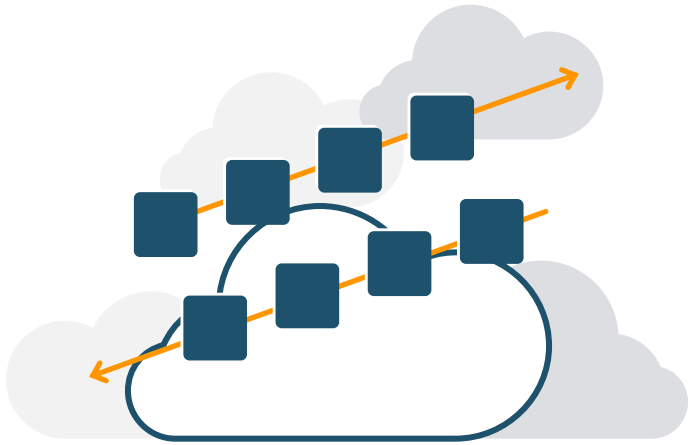
Scalable and durable

"...AWS가 퍼블릭 클라우드 스토리지 서비스를 운영하는 규모는 매직 쿼드런트 내의 다른 벤더들을 뒤엎었습니다."

- Gartner Magic Quadrant for Public Cloud Storage Services, Worldwide
Raj Bala, Arun Chandrasekaran, John McArthur, July 24, 2017

예시: Amazon S3는 수 조개의 객체를 저장하며 주기적으로 피크 시에는 초당 수 백만의 request를 처리합니다.

모든 규모
Scalable and durable



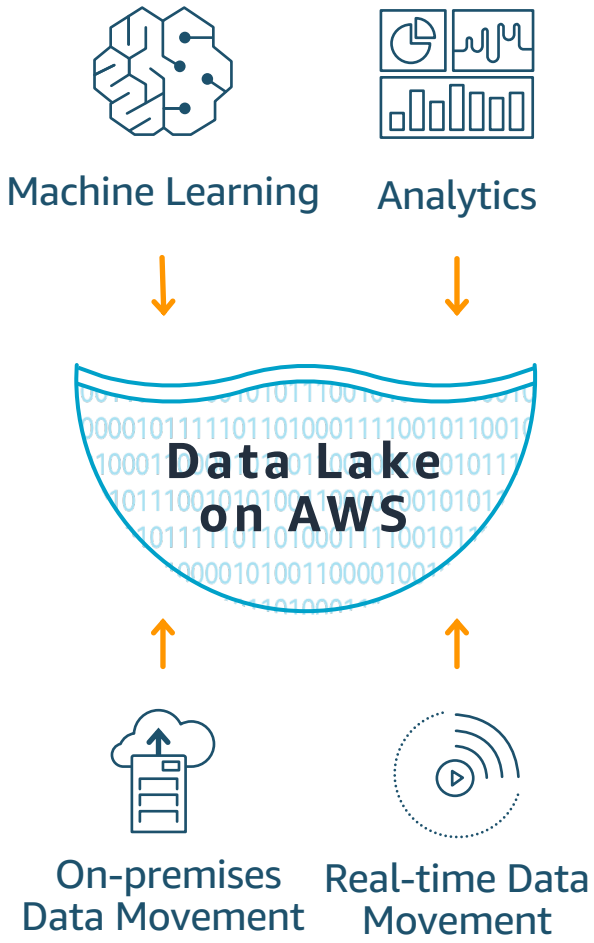
- S3에 수 조개의 객체와 엑사바이트 급의 데이터 저장
- 어떤 크기의 데이터도 저장 가능
- 어떤 크기의 컴퓨팅 자원도 수 분만에 스핀업 하여 대규모의 분석 엔진 실행
- 전 세계에서 가장 큰 클라우드 인프라에서 실행

탁월한 내구성과 가용성

Scalable and durable

- 99.999999999%의 내구성을 제공
- 지리적 중복 가능 & 자동 복제
- 단일 지역 내 3개의 가용 영역에 걸쳐 독립적인 데이터 센터에 데이터 저장
- 지역 간 데이터 복제

Data Lakes and Analytics from AWS



Open and comprehensive

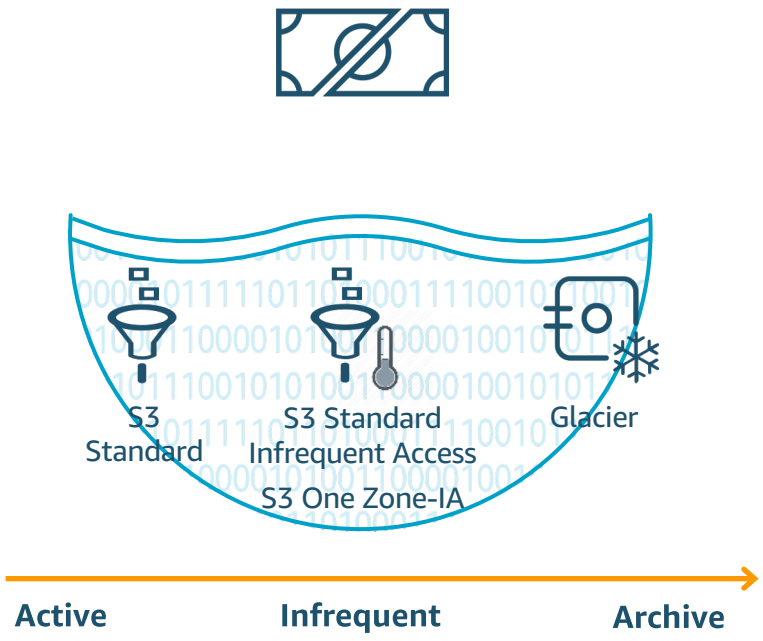
Secure

Scalable and durable

Lowest cost

가격/성능 최적화를 위한 Tiered storage

Lowest Cost

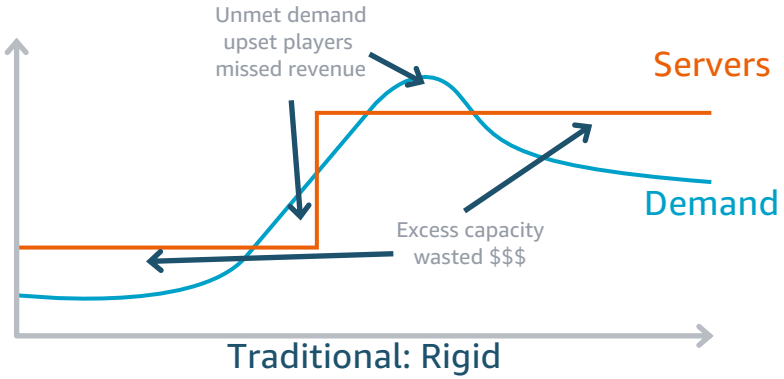


- 가격/성능 최적화를 위해 Tiered storage 사용
  - S3 Standard
  - S3 Standard—Infrequent Access
  - S3 One Zone—Infrequent Access
  - Amazon Glacier
- 생명주기 정책 기반으로 티어 간 마이그레이션
- S3에 데이터 저장 시 $0.023/GB/month
- Glacier에 데이터 저장 시 $0.004/GB/month

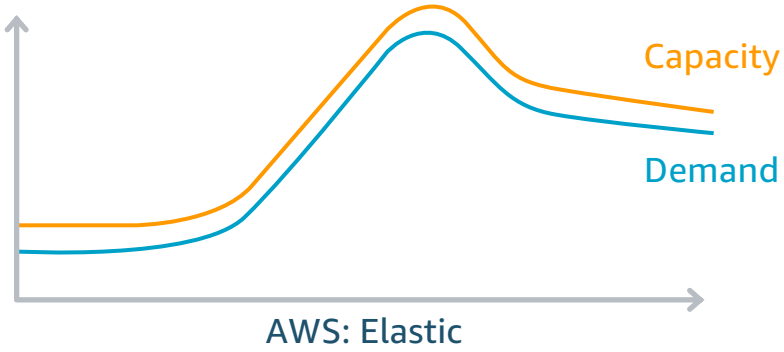
규모에 따라 사용한 자원에 대해서만 지불

Lowest Cost

전통적 접근방식: 용량 낭비로 이어짐



AWS 접근방식: 사용한 용량만큼 지물



- 필요한 자원에 대한 주문형 서비스
- Athena 스캔 기준 $0.05/GB
- EMR과 Athena 는 업무가 완료된 후 자동으로 자원 축소 가능하므로 비용 절약이 가능
- 예약 인스턴스 (RI)를 통해 일정 기간 동안 Commit하면 최대 75% 절약 가능
- 여분의 컴퓨팅 용량으로 EMR에서 클러스터를 실행하여 스팟 인스턴스로 최대 90% 절약 가능

AWS의 데이터 레이크, 분석, 기계학습 포트폴리오

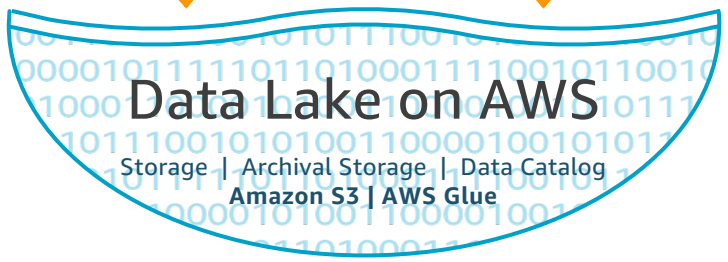
다양하고 수준 높은 분석 서비스

Machine Learning

Amazon SageMaker
AWS Deep Learning AMIs
Amazon Rekognition
Amazon Lex
AWS DeepLens
Amazon Comprehend
Amazon Translate
Amazon Transcribe
Amazon Polly

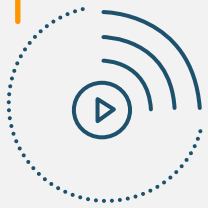
Analytics

Amazon Athena
Amazon EMR
Amazon Redshift
Amazon Elasticsearch service
Amazon Kinesis
Amazon QuickSight



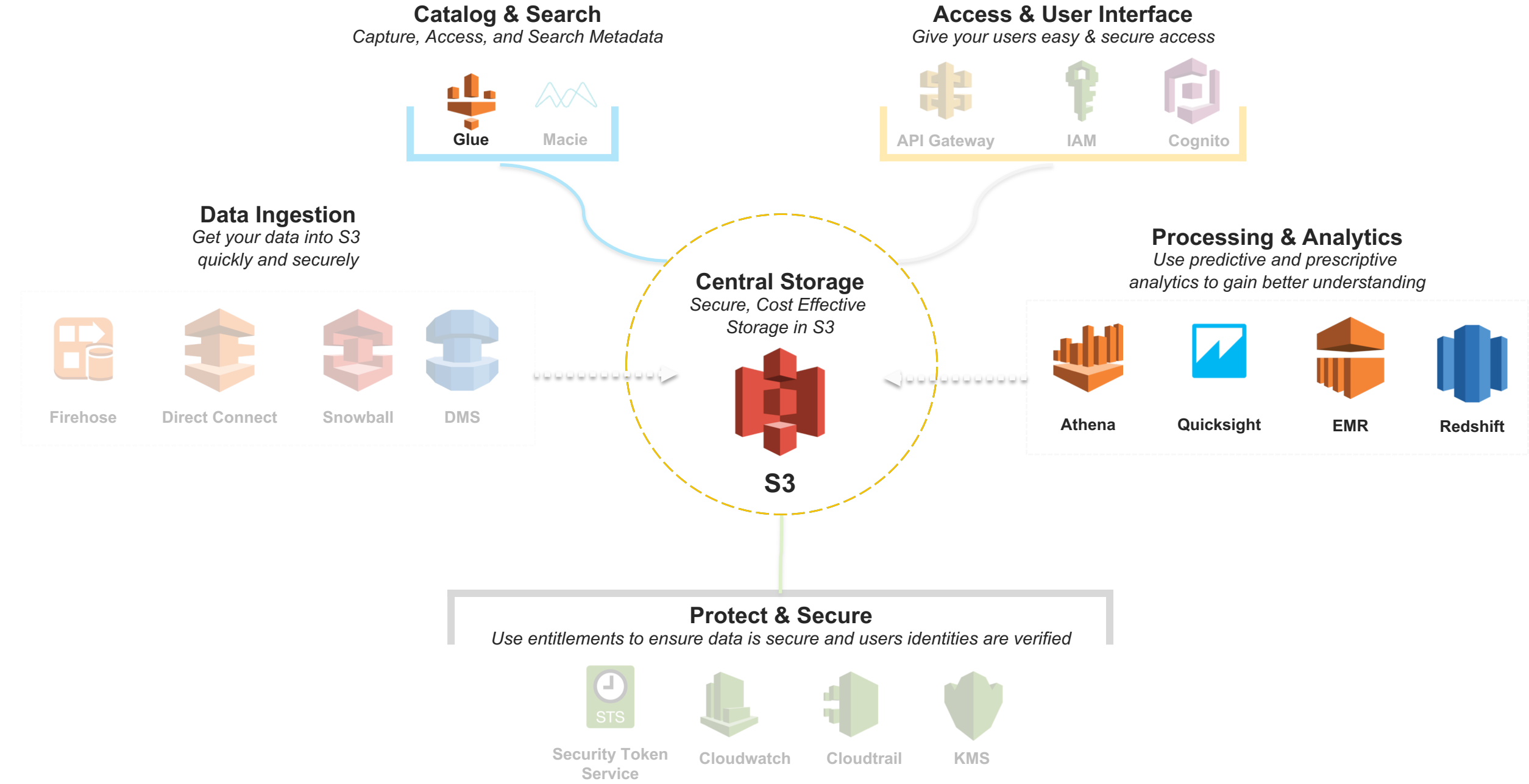
On-premises Data Movement

AWS Direct Connect
AWS Snowball
AWS Snowmobile
AWS Database Migration Service

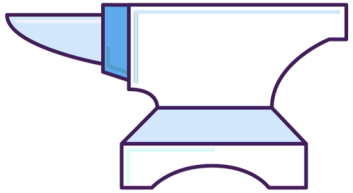


Real-time Data Movement

AWS IoT Core
Amazon Kinesis Data Firehose
Amazon Kinesis Data Streams
Amazon Kinesis Video Streams

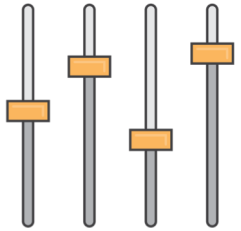

Catalog & Search
Capture, Access, and Search Metadata
Glue
Macie
Access & User Interface
Give your users easy & secure access
API Gateway
IAM
Cognito
Data Ingestion
Get your data into S3 quickly and securely
Firehose
Direct Connect
Snowball
DMS
Central Storage
Secure, Cost Effective Storage in S3
S3
Processing & Analytics
Use predictive and prescriptive analytics to gain better understanding
Athena
Quicksight
EMR
Redshift
Protect & Secure
Use entitlements to ensure data is secure and users identities are verified
STS
Security Token Service
Cloudwatch
Cloudtrail
KMS

데이터 저장소 - Amazon S3



타의 추종을
불허하는 내구성,
가용성 및 확장성

최상의 보안, 컴플라이언스
및 감사 기능



모든 규모에서
객체 별 제어 가능

데이터에 대한 비즈니스
통찰력 제공

데이터를 가져 오는
가장 많은 방법 제공

수많은 파트너 솔루션과
통합

데이터 레이크 S3 Tier 설계

Tier-1 데이터 레이크 : 수집과 저장

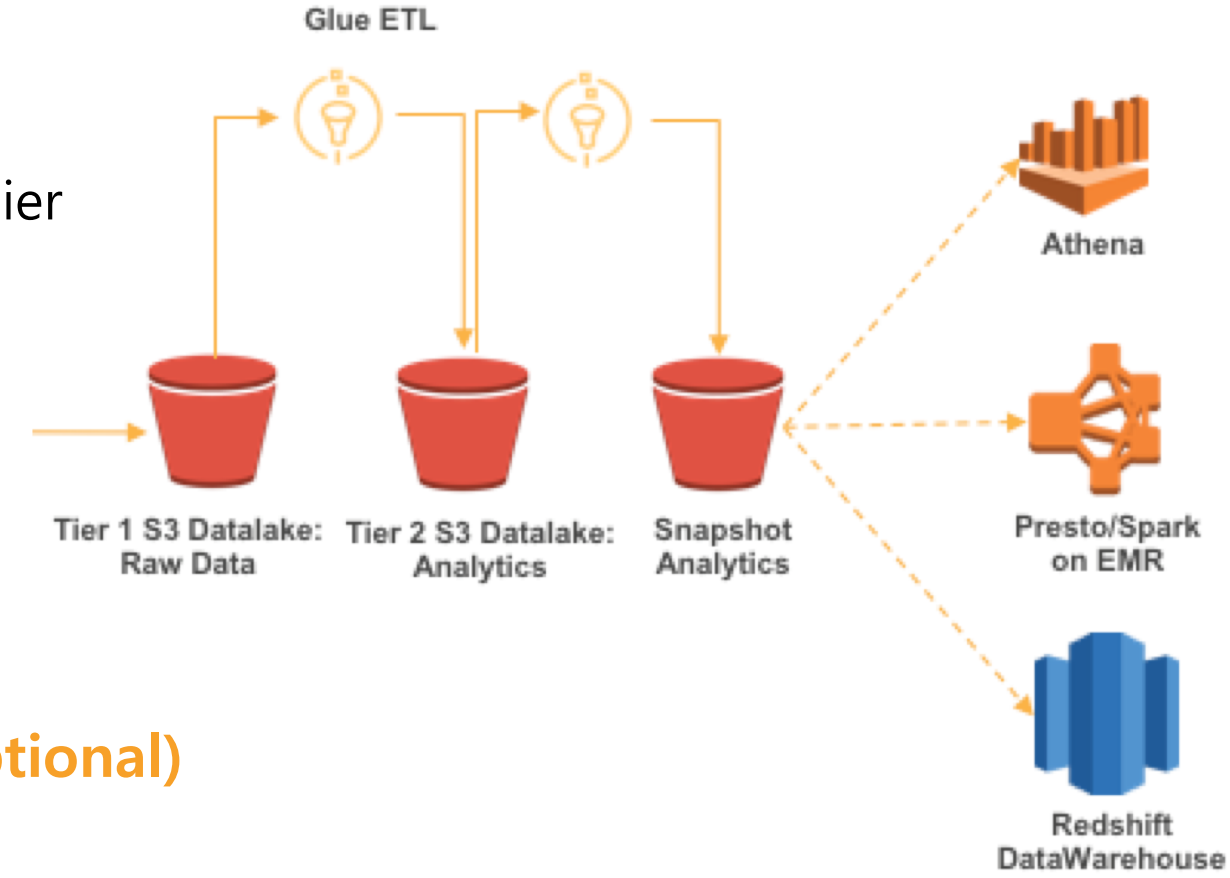
- 원본 데이터의 저장과 보장
- 최소한의 데이터 변환 작업만
- S3의 라이프사이클 기능 활용, S3-IA 또는 Glacier

Tier-2 데이터 레이크 : 분석용 데이터

- Parquet / ORC 같은 컬럼방식 포멧의 사용
- 파티션 정책에 따라 분산
- 분석을 위한 최적화

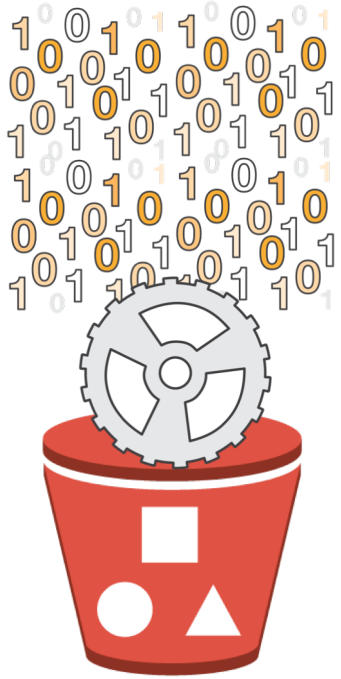
Tier-3 데이터 레이크 : 특정한 분석 목적 (optional)

- 도메인 레벨로 데이터마트 분리
- Use Case에 적합한 구성
- 특정 분석 방식에 적합한 데이터 변경 (ML, AI)



Optimizing for Cost and Performance

```
/user/hive/warehouse/logs
├── dt=2001-01-01/
│   ├── country=GB/
│   │   ├── file1
│   │   └── file2
│   └── country=US/
│       └── file3
└── dt=2001-01-02/
    ├── country=GB/
    │   └── file4
    └── country=US/
        ├── file5
        └── file6
```



Partitioning

Pay for data your query **needs**, not to scan **all** of your data

Compression

Pay for what you **store**, not for what you **process**

Managed Services

Pay for what you **use**, not for what you **run**

Partitioning

```
datalake
├── 20170515T1423-GB-01.tar.gz
├── 20170515T1423-GB-02.tar.gz
├── 20170515T1500-US-01.tar.gz
├── 20170516T1500-US-01.tar.gz
├── 20170516T1600-GB-01.tar.gz
└── 20170516T1600-GB-02.tar.gz
```

select * from datalake where
dt=20170515 and **country=US**

```
datalake
├── dt=20170515
│   ├── country=GB
│   │   ├── 20170515T1423-GB-01.tar.gz
│   │   └── 20170515T1423-GB-02.tar.gz
│   └── country=US
│       └── 20170515T1500-US-01.tar.gz
└── dt=20170516
    ├── country=GB
    │   ├── 20170516T1600-GB-01.tar.gz
    │   └── 20170516T1600-GB-02.tar.gz
    └── country=US
        └── 20170516T1500-US-01.tar.gz
```

Partitioning - Advantages

	select count(*) from datalake where dt='20170515'		select count(*) from datalake where dt >= '20170515' and dt < '20170516'	
	Non-Partitioned	**Partitioned**	**Non-Partitioned**	**Partitioned**
Run Time	9.71 sec	**2.16 sec**	10.41 sec	**2.73 sec**
Data Scanned	74.1 GB	**29.06 MB**	74.1 GB	**871.39 MB**
Cost	$0.36	**$0.0001**	$0.36	**$0.004**
Results	**77% faster, 99% cheaper**		**73% faster, 98% cheaper**	

Partitioning - Disadvantage

	select count(*) from datalake	
	Non-Partitioned	**Partitioned**
Run Time	8.4 sec	**10.65 sec**
Data Scanned	74.1 GB	74.1 GB
Cost	$0.36	$0.36
Results	**27% slower**	

Compression

- Compressing your data can speed up your queries significantly
- Splittable formats enable parallel processing across nodes

Algorithm	Splittable	Compression Ratio	Algorithm Speed	Good For
Gzip (DEFLATE)	No	High	Medium	Raw Storage
bzip2	Yes	Very High	Slow	Very Large Files
LZO	Yes	Low	Fast	Slow Analytics
Snappy	Yes and No *	Low	Very Fast	Slow & Fast Analytics

* Depends on if the source format is splittable and can output each record into a Snappy Block

Compression - Example

Snappy Compression with Parquet File Format

Format	Size on S3	Run Time	Data Scanned	Cost
Text	1.15 TB	3m 56s	1.15 TB	$5.75
Parquet	130 GB	6.78s	2.51 GB	$0.013
Result	**87% less**	**34x faster**	**99% less**	**99.7% savings**

Compression – File Counts

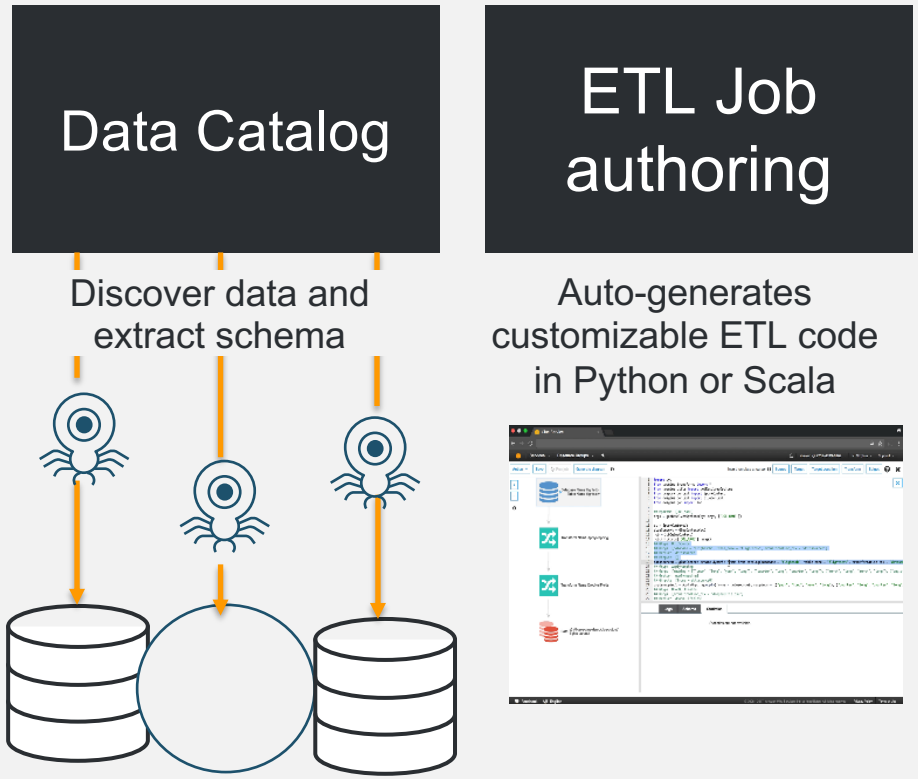
Fewer, larger files are better than many, smaller files (when splittable)

- Faster Listing Operations
- Fewer Requests to Amazon S3
- Less Metadata to Manage
- Faster Query Performance

Query	# files	Run Time
select count(*) from datalake	5000 files	8.4 sec
select count(*) from datalake	1 file	2.31 sec
Result		**72% Faster**

AWS Glue

Serverless Data Catalog & ETL Service



Automatically discovers data and stores schema

Data is immediately searchable, and available for ETL

Automatically generates customizable code

Schedules and runs your ETL jobs

Serverless

인터렉티브 분석 - Amazon Athena

Amazon S3에 직접 접근하여 표준 SQL로 데이터 분석이 가능한 인터렉티브 쿼리 서비스

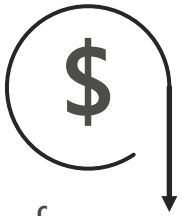
데이터의 로딩이 필요없고 인프라 관리가 필요없음

Amazon Glacier 와 같은 아카이브 스토리지에 직접 SQL 쿼리가 (new)

Query Instantly

Zero setup cost; just point to S3 and start querying

Pay per query

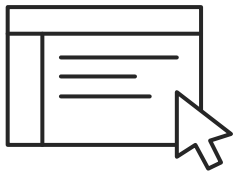


Pay only for queries run; save 30–90% on per-query costs through compression

Open

ANSI SQL interface, JDBC/ODBC drivers, multiple formats, compression types, and complex joins and data types

Easy



Serverless: zero infrastructure, zero administration
Integrated with QuickSight

관리형 하둡 플랫폼 - Amazon EMR

다양한 분석과 머신러닝이 가능한 19개의 오픈소스 프로젝트가 포함

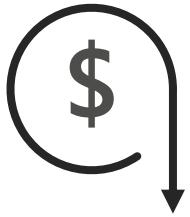
AWS Glue Data Catalog를 통해 Apache Spark, Hive, Presto에서 데이터 사용 가능

컴퓨트와 스토리지의 분리, 오토스케일, 엔터프라이즈 환경에 맞춘 보안 기능

Latest versions

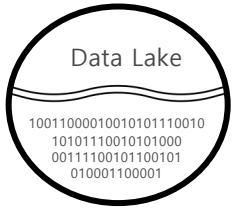
Updated with the latest open source frameworks within 30 days of release

Low cost



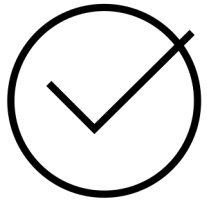
Flexible billing with per-second billing, EC2 spot, reserved instances and auto-scaling to reduce costs 50-80%

Use S3 storage



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

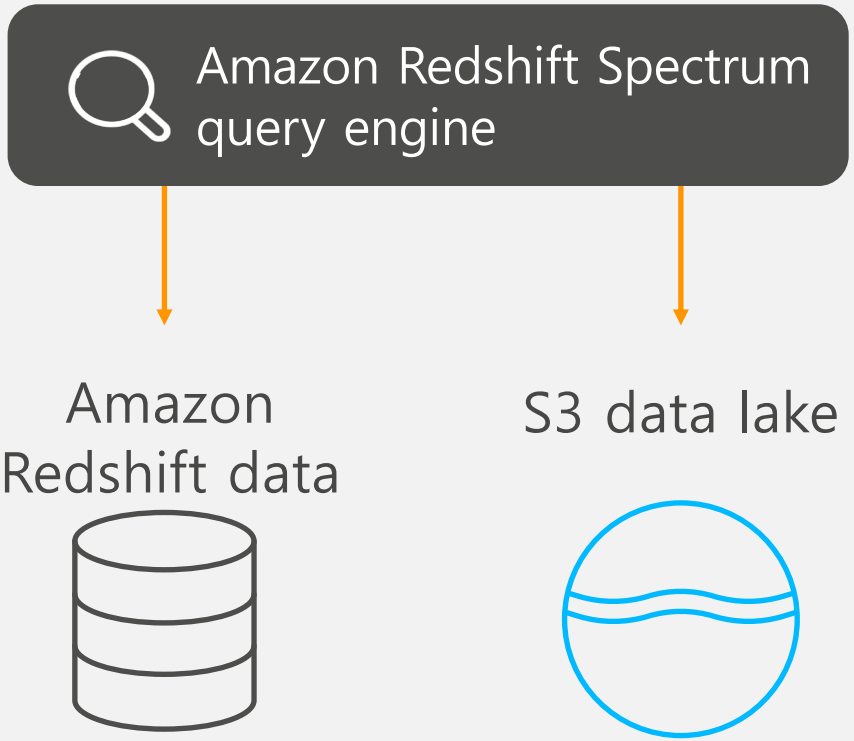
Easy



Launch fully managed Apache Hadoop & Apache Spark in minutes; no cluster setup, node provisioning, cluster tuning

데이터 웨어하우스의 확장 - Redshift Spectrum

S3 데이터 레이크에서 데이터 웨어하우스를 엑사 바이트 규모로 확장



S3에 Exabyte Redshift SQL querie를 수행

S3 와 Redshift 에서 데이터 조인

연산과 저장 공간 을 별개로 확장

일관적인 쿼리 성능과 무한의 concurrency

CSV, ORC, Grok, Avro, & Parquet 데이터 타입

스캔된 데이터만 과금

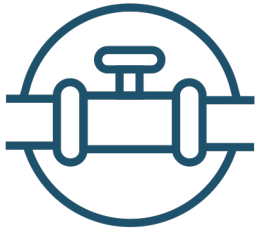
Amazon Kinesis – Real Time

Easily collect, process, and analyze video and data streams in real time

New

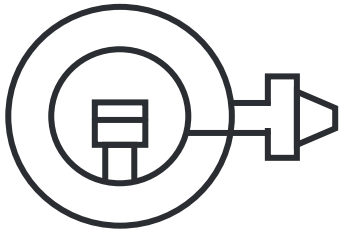
Kinesis Video Streams

Capture, process, and store video streams for analytics



Kinesis Data Streams

Build custom applications that analyze data streams



Kinesis Data Firehose

Load data streams into AWS data stores

Kinesis Data Analytics

Analyze data streams with SQL

aws

Amazon Elastic Search Service

Easy to deploy, secure, operate, and scale Elasticsearch

Customers use Elasticsearch for log analytics, full text search & application monitoring

---

Easy to Use

Fully-managed.
Deploy production-ready clusters in minutes

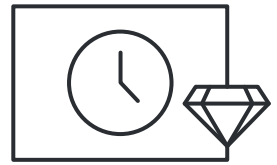
Open

Direct access to Elasticsearch open-source APIs; supports Logstash and Kibana

Secure

Secure access with VPC to keep all traffic within AWS network

Available

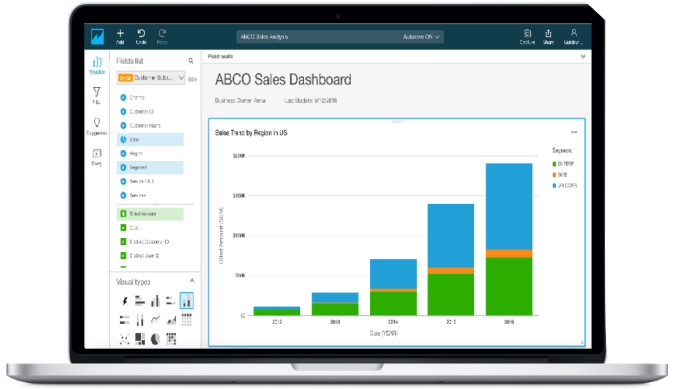


Zone awareness replicates data between two AZs; automatically monitors & replaces failed nodes

aws

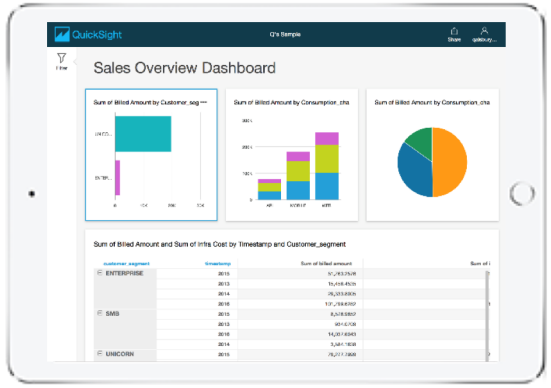
Analyze, Collaborate, Publish - QuickSight

QuickSight lets users create and share data sets, collaborate on your live analyses, and share read-only dashboards and storyboards that can be accessed on any device, anytime, anywhere.



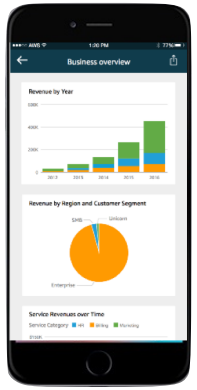
Analyses

Analyses are visual explorations of your data. Multiple users can collaborate on an analyses with the ability to modify and change them in any way.



Dashboards

You can share your analyses as read only dashboards. Viewers can interact with and filter the visualizations without modifying them.

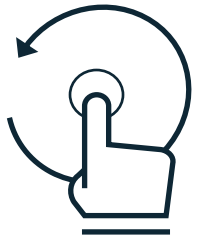


Storyboards

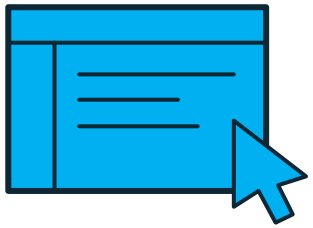
Let you combine visualizations into a guided tour that you can share with other users.

ML Platform - Amazon SageMaker

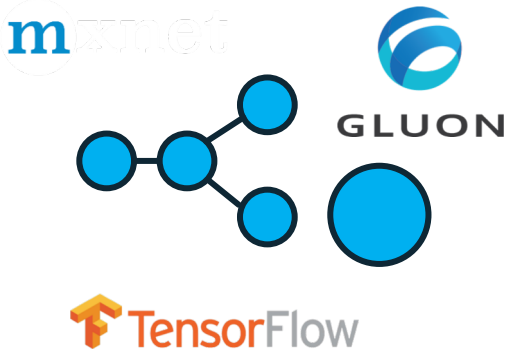
A managed service that provides the quickest and easiest way for your data scientists and developers to get ML models from idea to production.



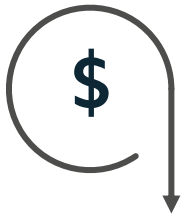
End-to-End Machine Learning Platform



Zero setup



Flexible Model Training



Pay by the second

These tools come together to form a complete AI/ML stack

APPLICATION SERVICES
ML for everyone

LEX | POLLY | REKOGNITION | REKOGNITION VIDEO | TRANSCRIBE | TRANSLATE | COMPREHEND

PLATFORM SERVICES
ML for engineers

AMAZON SAGEMAKER | AWS DEEPLENS | SPARK & EMR | AMAZON MECHANICAL TURK

FRAMEWORKS AND INTERFACES
ML for data scientists

Frameworks: Caffe2 | Microsoft CNTK | mxnet | PYTORCH | TensorFlow | torch

Interfaces: KERAS | GLUON

INFRASTRUCTURE
Powering the ML

P3
NVIDIA Tesla V100 GPUs
(14x faster than P2)

C5
Intel Xeon Skylake
(Optimized for ML)

Machine Learning AMIs

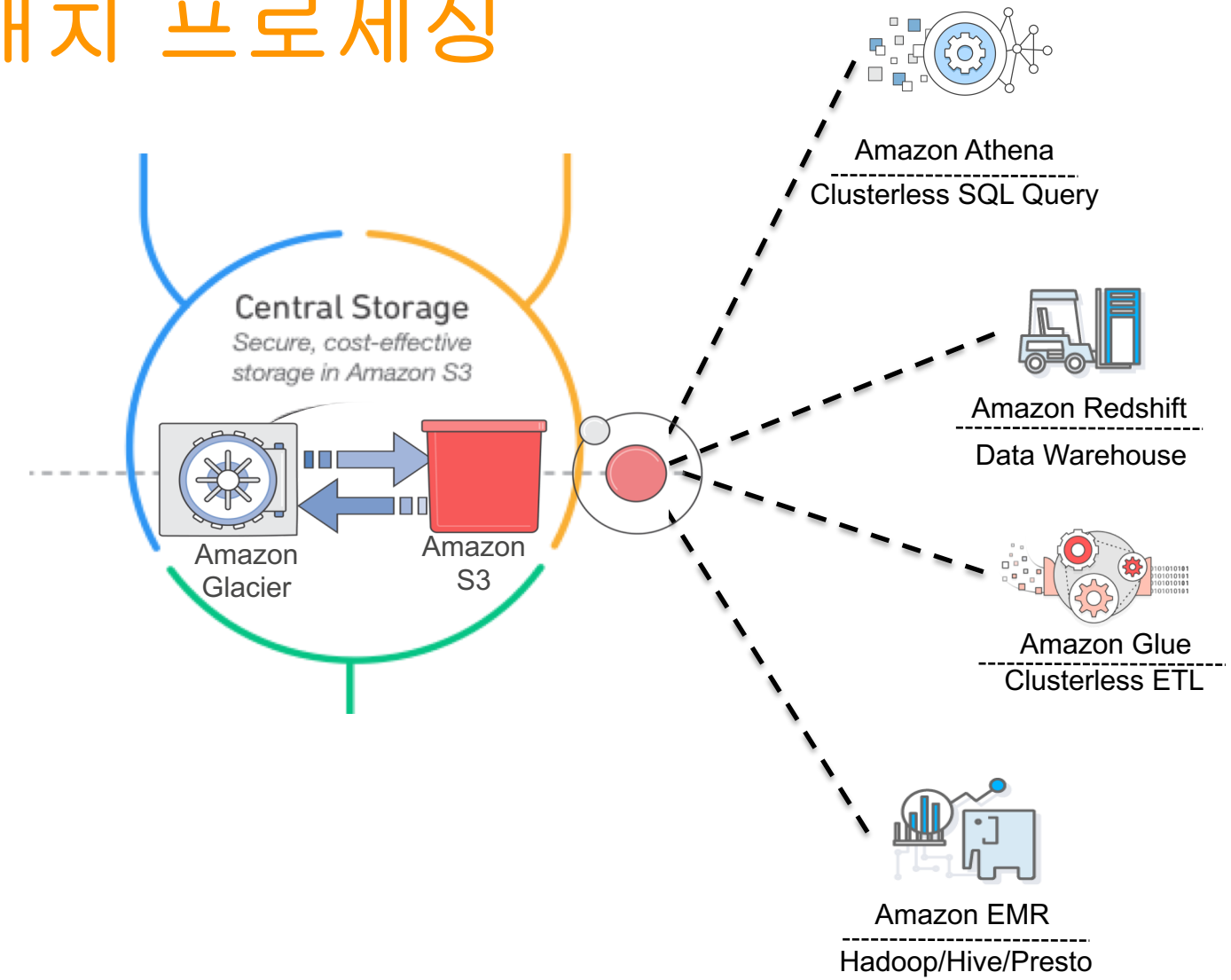
AWS GREENGRASS ML

aws

Data Lake Use Case

데이터 레이크로 해결 가능한 문제들

배치 프로세싱



BI & Visualization

Amazon QuickSight

looker

tableau SOFTWARE

kibana

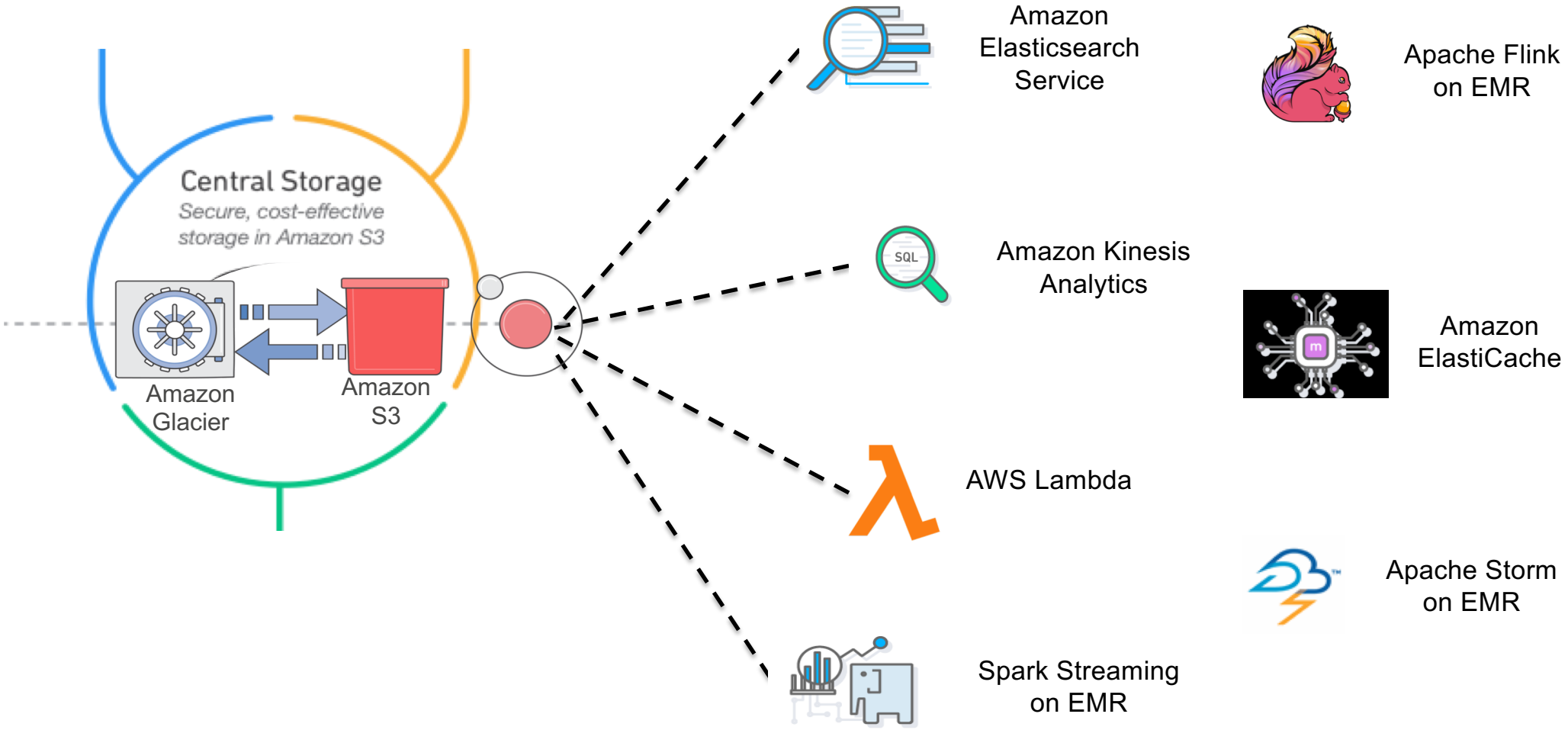
COGNOS

SAP Business Objects

MicroStrategy

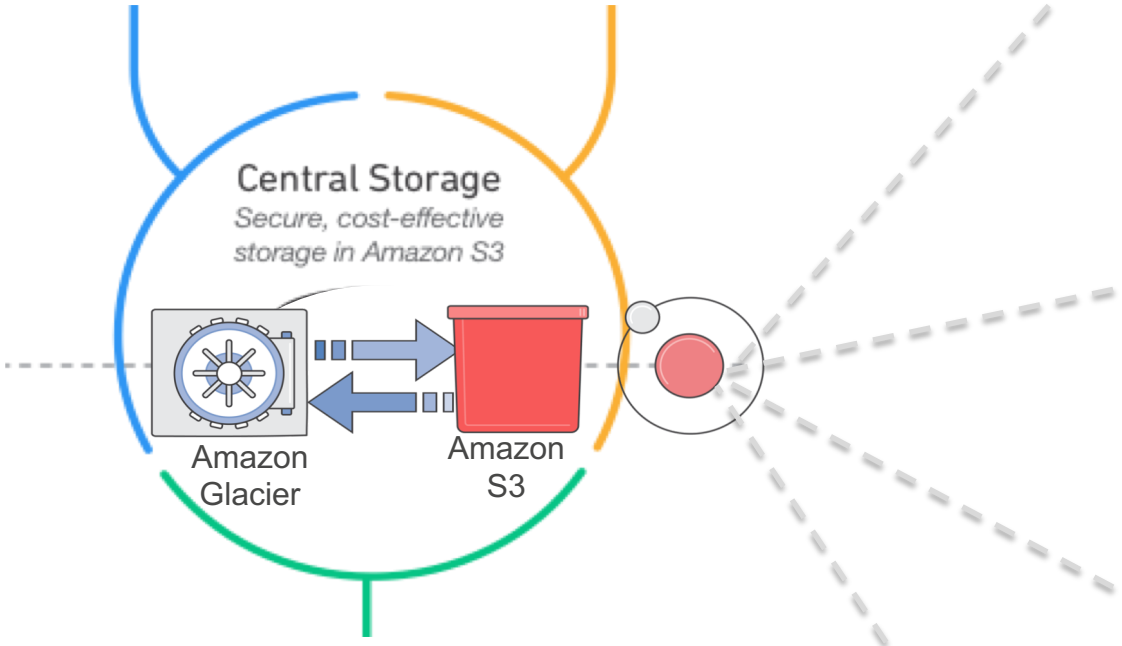
데이터 레이크로 해결 가능한 문제들

스트리밍 실시간 데이터 분석



데이터 레이크로 해결 가능한 문제들

AI / 머신러닝



Amazon Polly

Life-like speech

Amazon Lex

Conversational engine

Amazon Rekognition

Image analysis

Amazon SageMaker

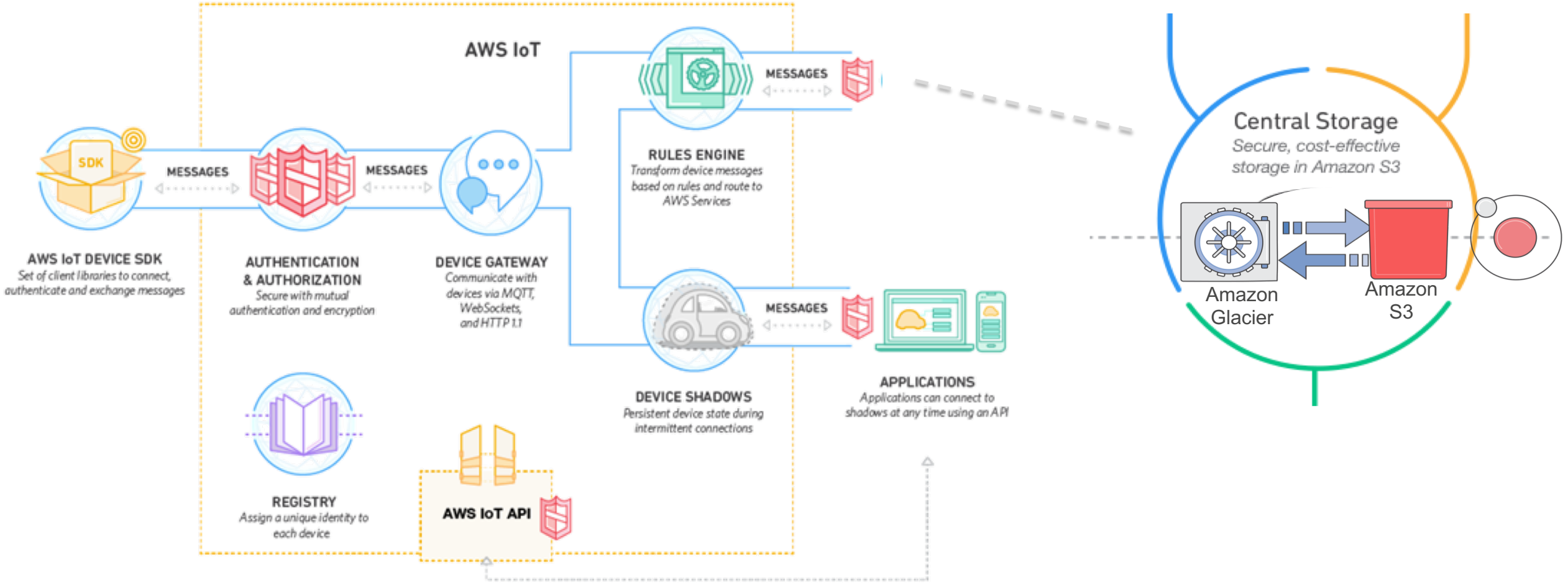
Machine learning platform

Deep learning Frameworks

MXNet, TensorFlow, Theano, Caffe, Torch

데이터 레이크로 해결 가능한 문제들

IoT Data Analysis

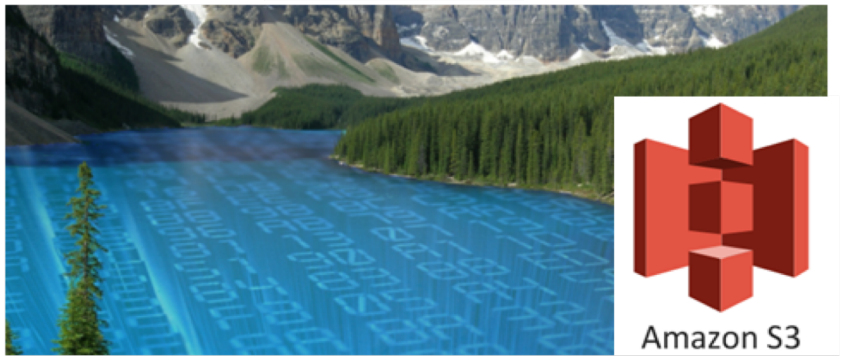


Andes - Amazon.com의 Data Lake

Postgre SQL instance

Amazon Kinesis

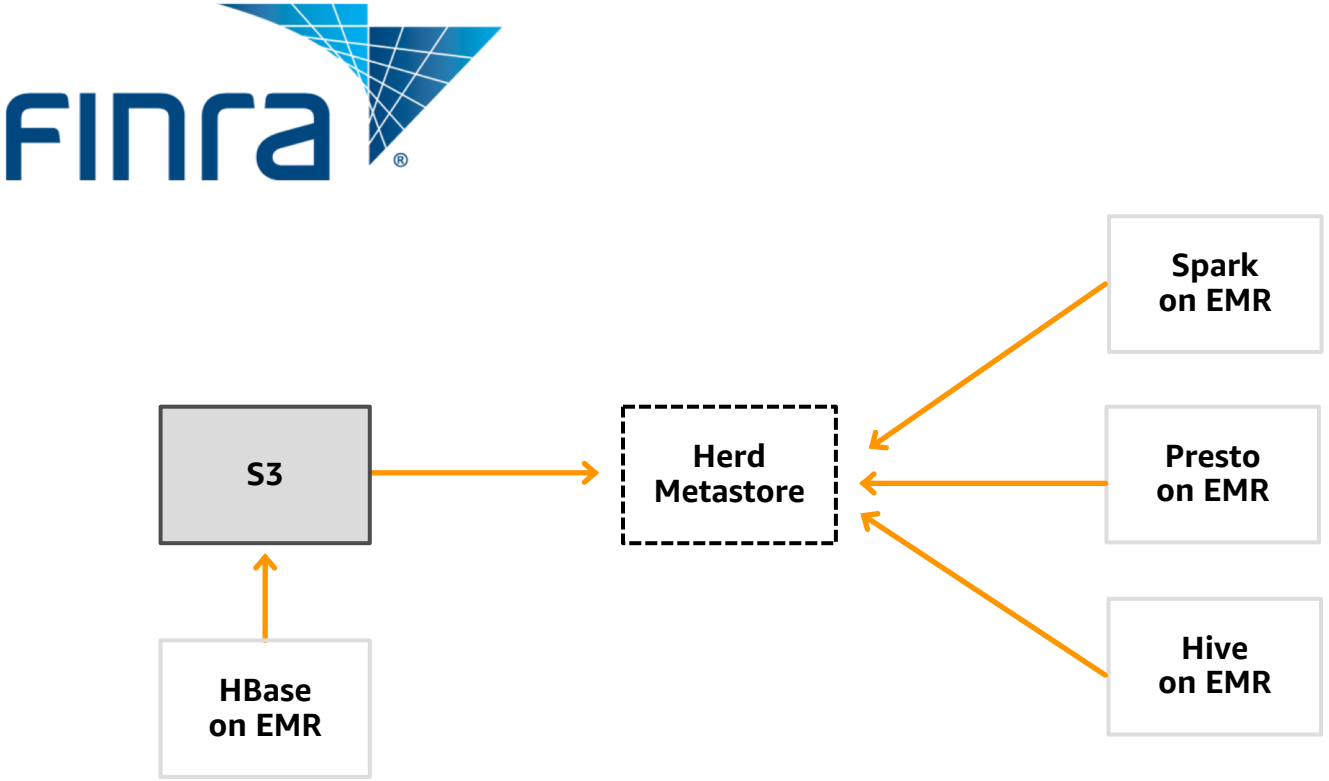
Source Systems



The Data Lake
"Andes"

Big Data Systems
Data Warehouses
"Bring Your Own Cluster" and
"Bring Your Own Query"

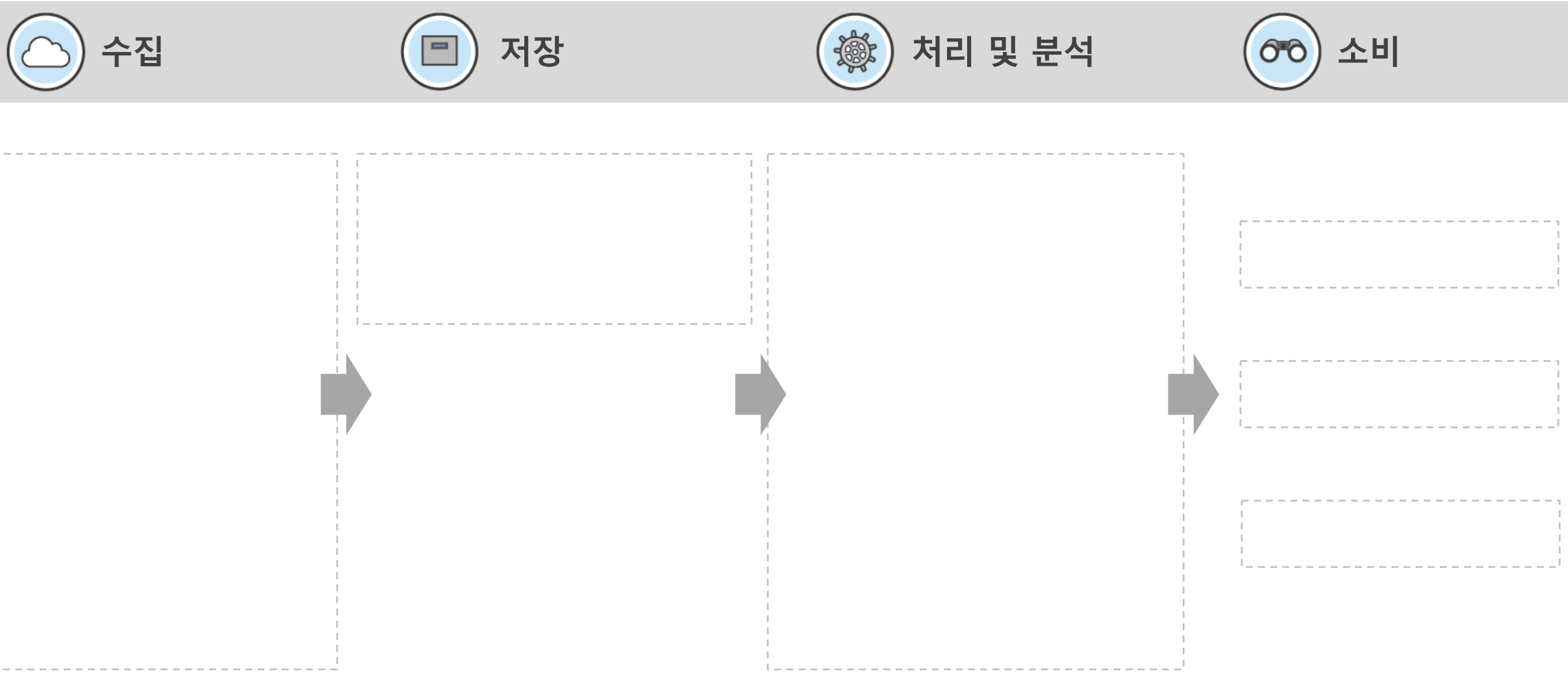
S3 / EMR 기반의 FINRA Data Lake



- 수 조개의(20PB+) 무역 거래 기록에 대한 빠른 액세스
- 자체 구축 시스템으로부터 이전
- Amazon EMR에서 Apache Hbase를 이용하여 해당 데이터를 저장 및 제공
- 데이터 처리를 위해 EMR 엔진 사용 – Spark, Presto, Hive
- 자체 구축 시스템 대비 60% 비용 절감

Architecting Data Lake

AWS는 데이터 레이크를 위한 모든 서비스를 제공



AWS는 데이터 레이크를 위한 모든 서비스를 제공

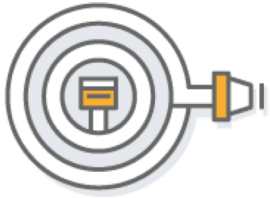
수집 | 저장 | 처리 및 분석 | 소비

Amazon S3
안전하고, 비용
효율적인 스토리지

Interoperate with everything

다양한 데이터 수집 방법

AWS Snowball
- PB 규모의 마이그레이션



Amazon Kinesis
- 스트림 데이터 수집
- 데이터 변환 및 임시 저장

AWS Storage Gateway
- 레거시 파일 마이그레이션

AWS Direct Connect
- 온프레미스와 네트워크 통합

AWS Data Migration Service
- 동종 및 이기종 데이터베이스 마이그레이션

Amazon S3 Transfer Acceleration
- 장거리 데이터 전송 가속화

AWS는 데이터 레이크를 위한 모든 서비스를 제공

수집 | 저장 | 처리 및 분석 | 소비

Direct Connect
데이터 센터와 연결

Snowball
벌크 데이터 로드

Database Migration Service
Oracle, Netezza 등의 데이터 임포트

Kinesis
스트리밍 데이터

and many more…

Amazon S3
안전하고, 비용 효율적인 스토리지

Interoperate with everything

다양한 데이터 처리 방법

배치 분석

실시간 분석

다양한 데이터 처리 방법

배치 분석

Amazon EMR

Spark 및 Hive가
실행되는 관리형 하둡

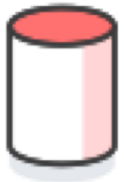
실시간 분석

다양한 데이터 처리 방법

배치 분석

Amazon EMR

Spark 및 Hive가
실행되는 관리형 하둡



Amazon Redshift + Spectrum

페타바이트 규모의
관리형 DW

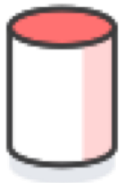
실시간 분석

다양한 데이터 처리 방법

배치 분석

Amazon EMR

Spark 및 Hive가
실행되는 관리형 하둡



Amazon Redshift + Spectrum

페타바이트 규모의
관리형 DW

Amazon Athena

서버리스
대화식 쿼리 엔진 서비스

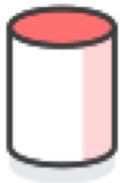
실시간 분석

다양한 데이터 처리 방법

배치 분석

Amazon EMR

Spark 및 Hive가
실행되는 관리형 하둡



Amazon Redshift + Spectrum

페타바이트 규모의
관리형 DW

Amazon Athena

서버리스
대화식 쿼리 엔진 서비스

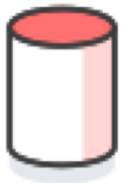
실시간 분석

다양한 데이터 처리 방법

배치 분석

Amazon EMR

Spark 및 Hive가
실행되는 관리형 하둡



Amazon Redshift + Spectrum

페타바이트 규모의
관리형 DW

Amazon Athena

서버리스
대화식 쿼리 엔진 서비스

실시간 분석

Amazon Kinesis Data Analytics

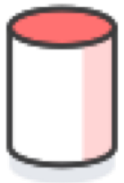
서버리스
실시간 스트리밍 분석

다양한 데이터 처리 방법

배치 분석

Amazon EMR

Spark 및 Hive가
실행되는 관리형 하둡



Amazon Redshift + Spectrum

페타바이트 규모의
관리형 DW

Amazon Athena

서버리스
대화식 쿼리 엔진 서비스

실시간 분석

Amazon Kinesis Data Analytics

서버리스
실시간 스트리밍 분석

Spark Streaming on Amazon EMR

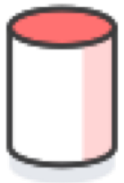
Apache Flink on Amazon EMR

다양한 데이터 처리 방법

배치 분석

Amazon EMR

Spark 및 Hive가
실행되는 관리형 하둡



Amazon Redshift + Spectrum

페타바이트 규모의
관리형 DW

Amazon Athena

서버리스
대화식 쿼리 엔진 서비스

실시간 분석

Amazon Kinesis Data Analytics

서버리스
실시간 스트리밍 분석

Amazon Elasticsearch

로그 분석 및 검색 엔진
관리형 서비스

Spark Streaming on Amazon EMR

Apache Flink on Amazon EMR

AWS는 데이터 레이크를 위한 모든 서비스를 제공


수집


저장


처리 및 분석


소비


Direct Connect
데이터 센터와 연결


Snowball
벌크 데이터 로드


Database Migration Service
Oracle, Netezza 등의 데이터 임포트


Kinesis
스트리밍 데이터

and many more…


Amazon S3
안전하고, 비용 효율적인 스토리지

Interoperate with everything


EMR
비정형 데이터 처리


Redshift
데이터 웨어하우스


Athena
ad-hoc 쿼리


Elasticsearch
로그 분석 및 검색 엔진

Kinesis Analytics
실시간 스트리밍 분석

and many more…

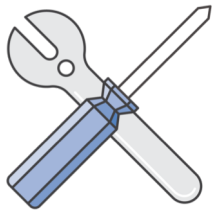
데이터는 완벽하지 않다!

데이터는 절대로 완벽할 수 없다!

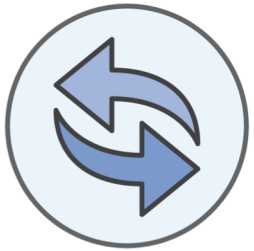
AWS Glue - 데이터 카탈로그 및 ETL

ETL (데이터 변환)

ETL 코드
자동 생성



개발 환경
제공

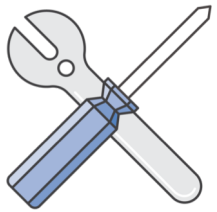


유연한 작업
스케줄러

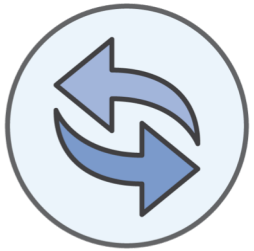
AWS Glue - 데이터 카탈로그 및 ETL

ETL (데이터 변환)

ETL 코드
자동 생성

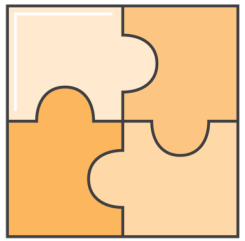


개발 환경
제공



유연한 작업
스케줄러

데이터 카탈로그



통합
데이터 카탈로그

자동 데이터 탐색

AWS는 데이터 레이크를 위한 모든 서비스를 제공


수집


저장


처리 및 분석


소비

Direct Connect
데이터 센터와 연결

Snowball
벌크 데이터 로드

Database Migration Service
Oracle, Netezza 등의 데이터 임포트

Kinesis
스트리밍 데이터

and many more…

AWS Glue
데이터 카탈로그와 ETL

Amazon S3
안전하고, 비용 효율적인 스토리지

Interoperate with everything

EMR
비정형 데이터 처리

Redshift
데이터 웨어하우스

Athena
ad-hoc 쿼리

Elasticsearch
로그 분석 및 검색 엔진

Kinesis Analytics
실시간 스트리밍 분석

and many more…

AWS는 데이터 레이크를 위한 모든 서비스를 제공

수집 | 저장 | 처리 및 분석 | 소비

Direct Connect
데이터 센터와 연결

Snowball
벌크 데이터 로드

Database Migration Service
Oracle, Netezza 등의 데이터 임포트

Kinesis
스트리밍 데이터

and many more…

AWS Glue
데이터 카탈로그와 ETL

Amazon S3
안전하고, 비용 효율적인 스토리지

Interoperate with everything

EMR
비정형 데이터 처리

Redshift
데이터 웨어하우스

Athena
ad-hoc 쿼리

Elasticsearch
로그 분석 및 검색 엔진

Kinesis Analytics
실시간 스트리밍 분석

and many more…

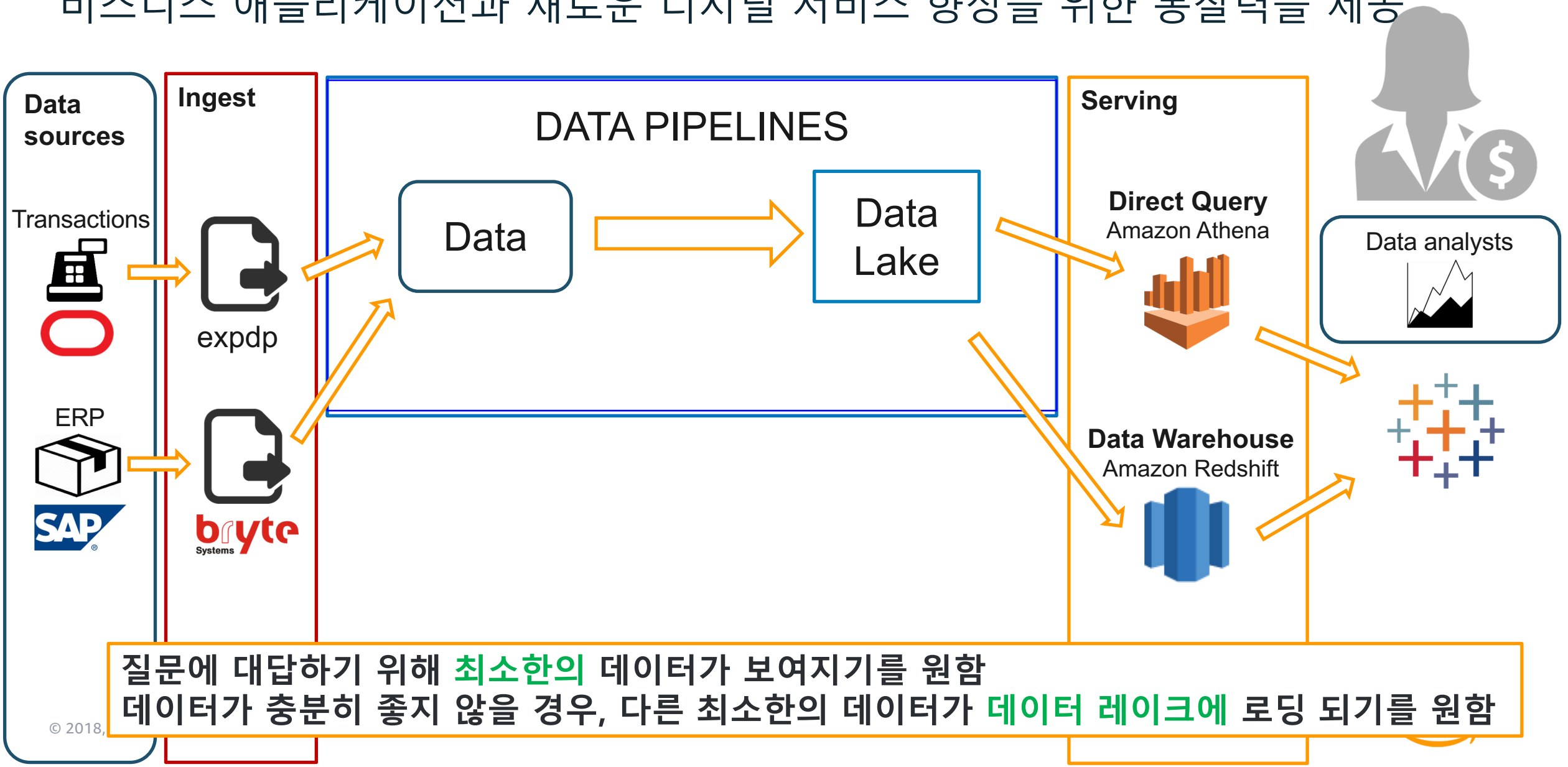
QuickSight
비지니스 인텔리젼스

SageMaker
머신러닝 플랫폼

API Gateway
API 개발 도구

모던 데이터 플랫폼

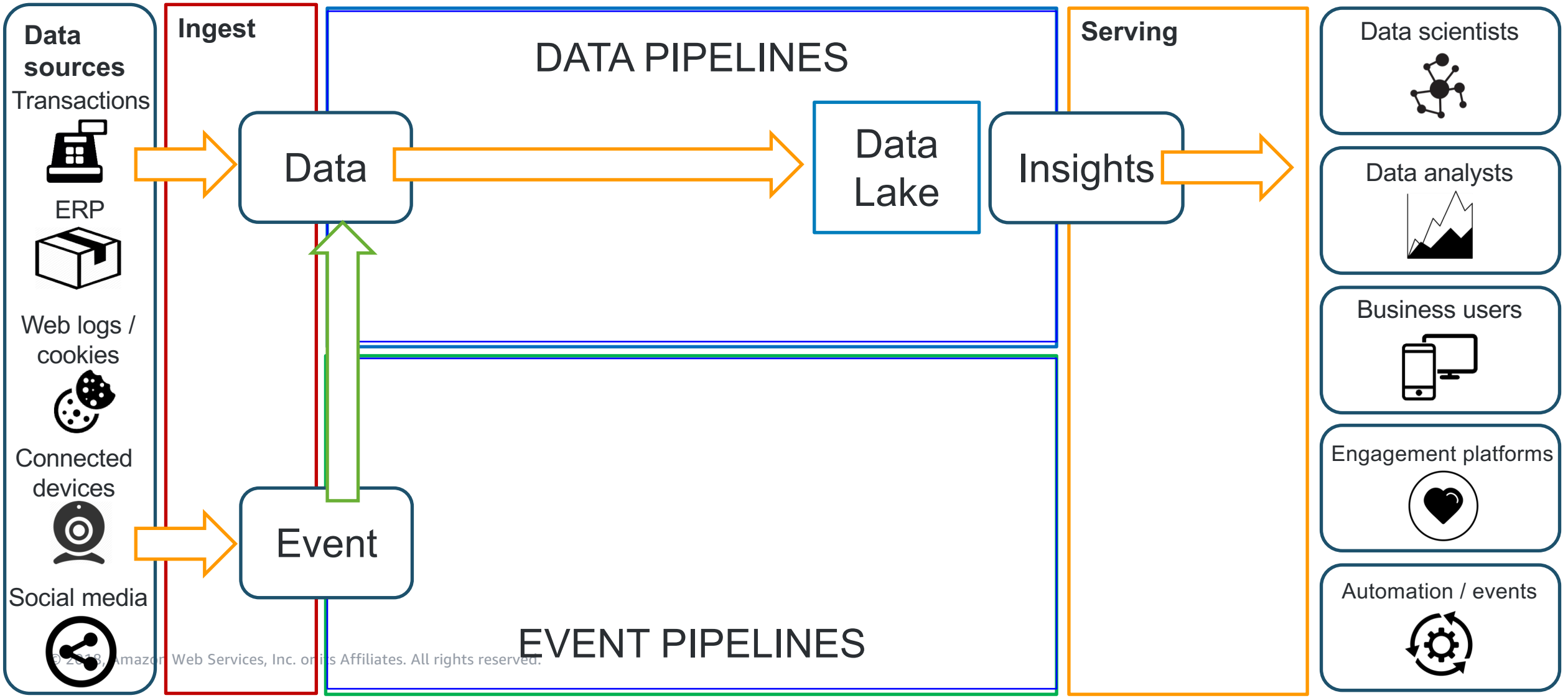
비즈니스 애플리케이션과 새로운 디지털 서비스 향상을 위한 통찰력을 제공



질문에 대답하기 위해 최소한의 데이터가 보여지기를 원함
데이터가 충분히 좋지 않을 경우, 다른 최소한의 데이터가 데이터 레이크에 로딩 되기를 원함

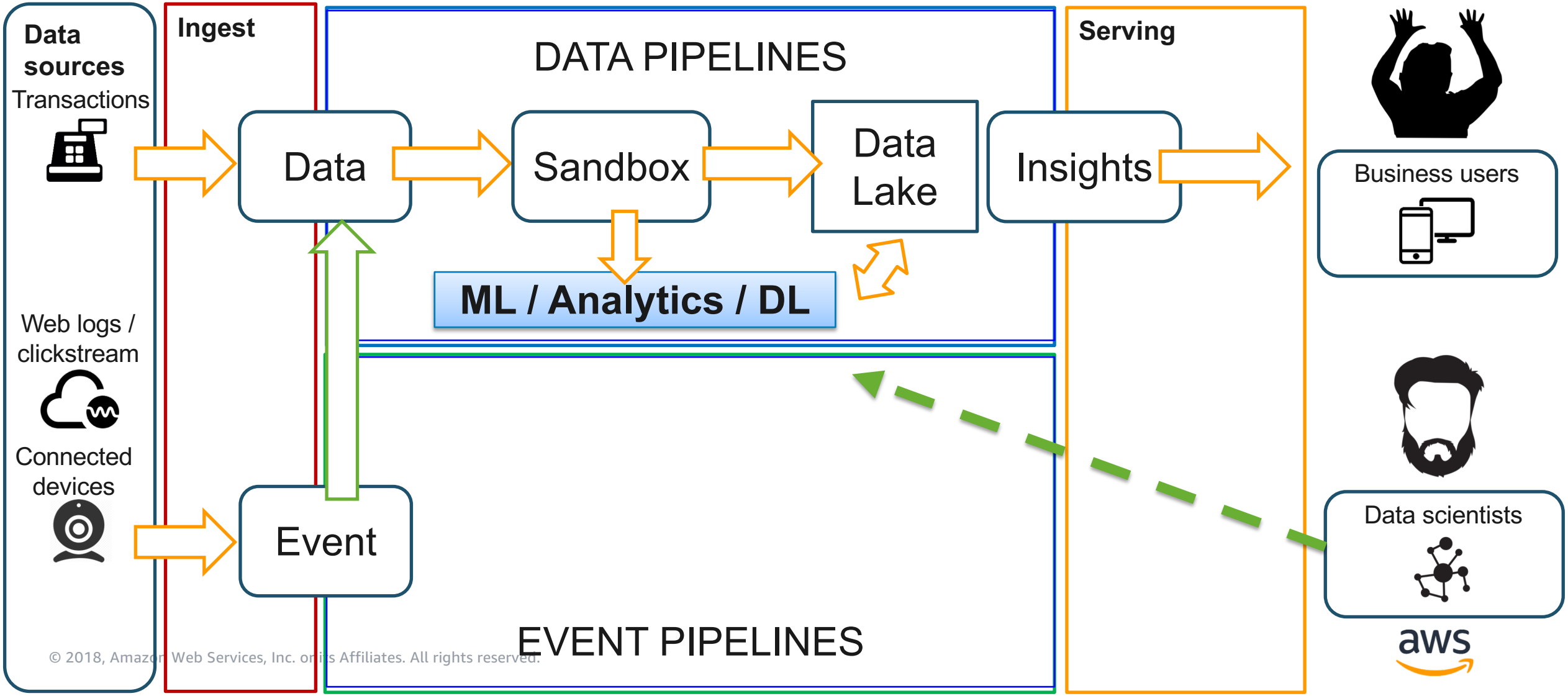
모던 데이터 플랫폼

비즈니스 애플리케이션과 새로운 디지털 서비스 향상을 위한 통찰력을 제공



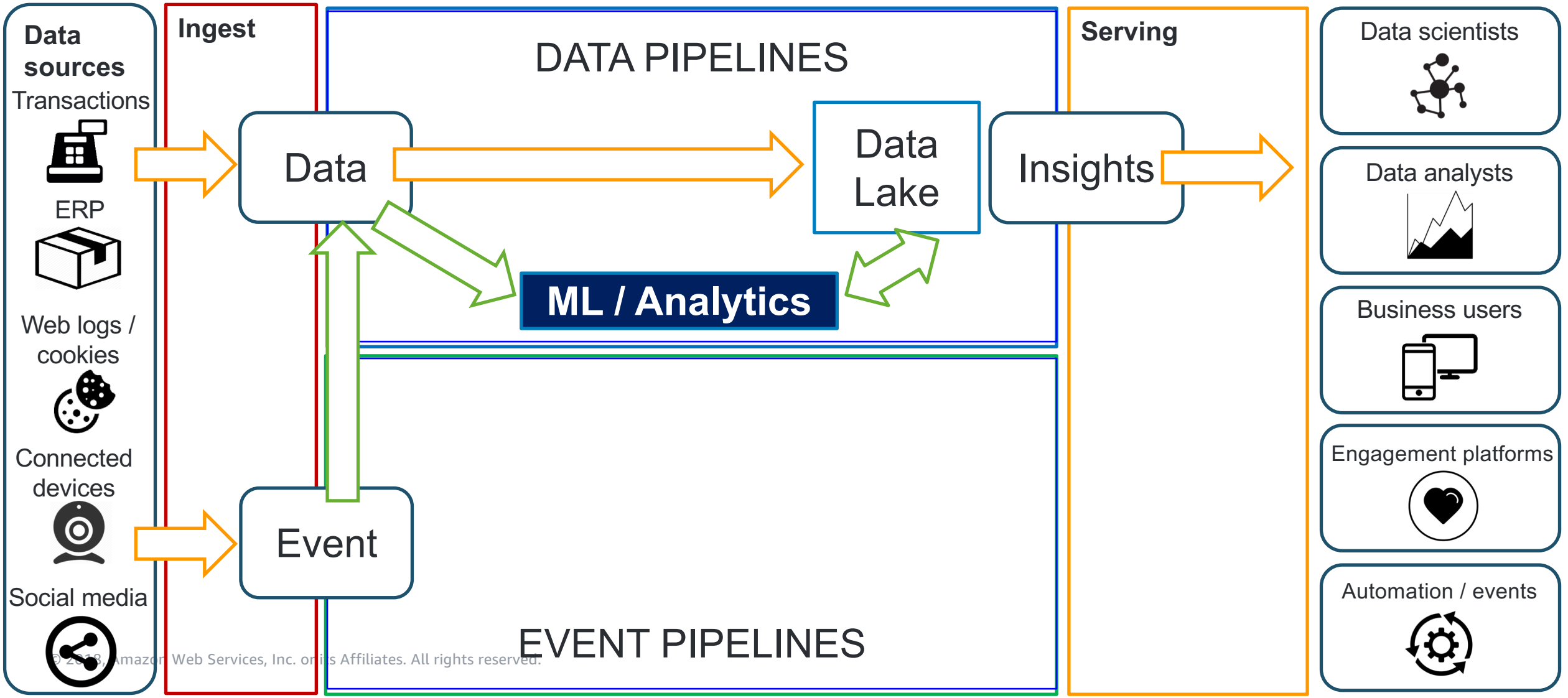
모던 데이터 플랫폼

비즈니스 애플리케이션과 새로운 디지털 서비스 향상을 위한 통찰력을 제공



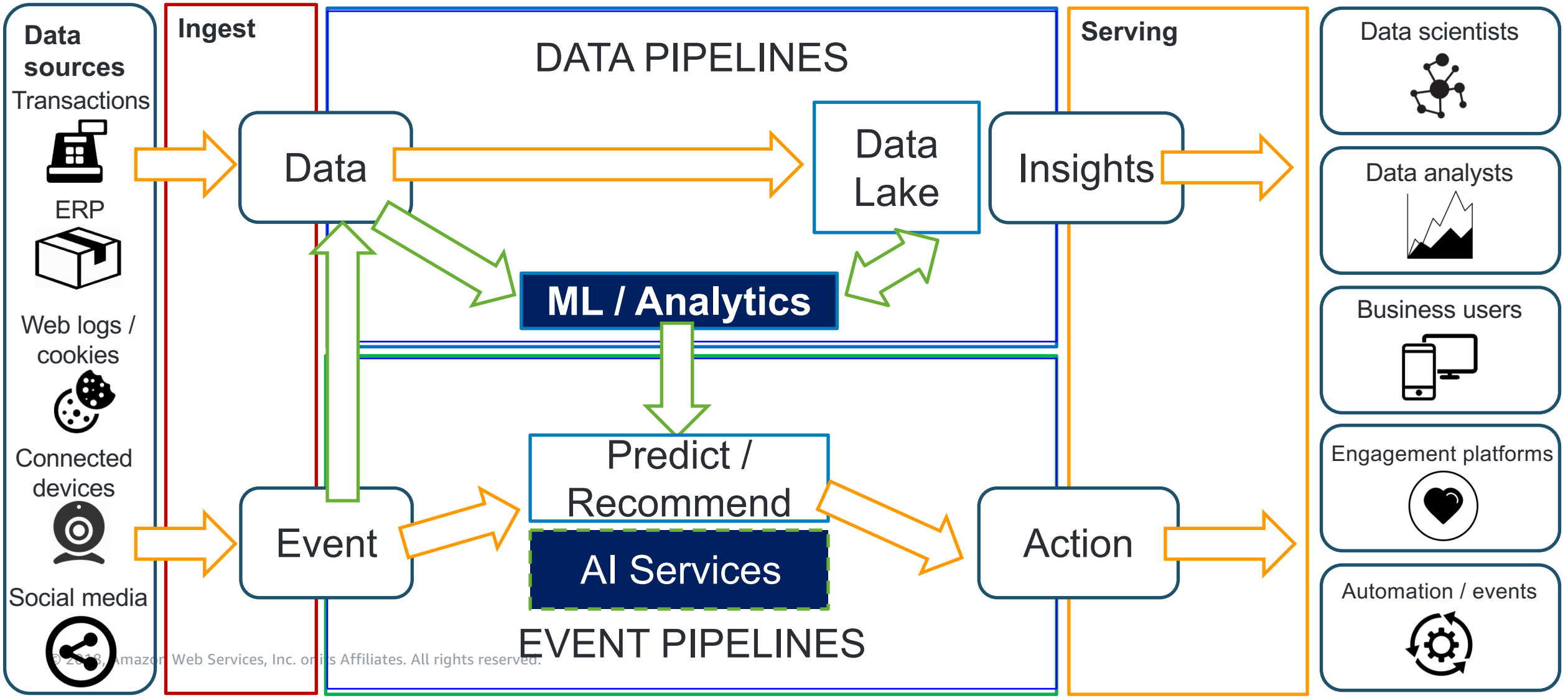
모던 데이터 플랫폼

비즈니스 애플리케이션과 새로운 디지털 서비스 향상을 위한 통찰력을 제공

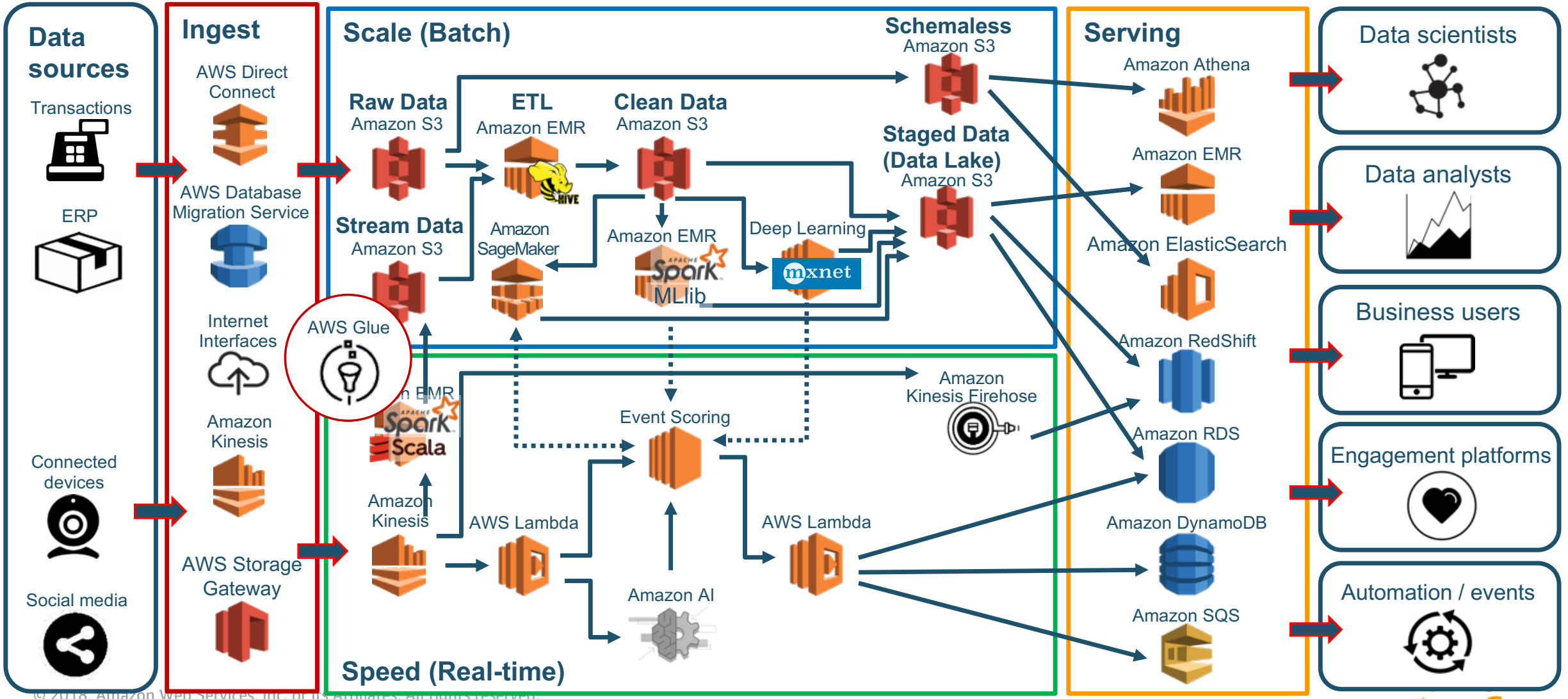


모던 데이터 플랫폼

비즈니스 애플리케이션과 새로운 디지털 서비스 향상을 위한 통찰력을 제공

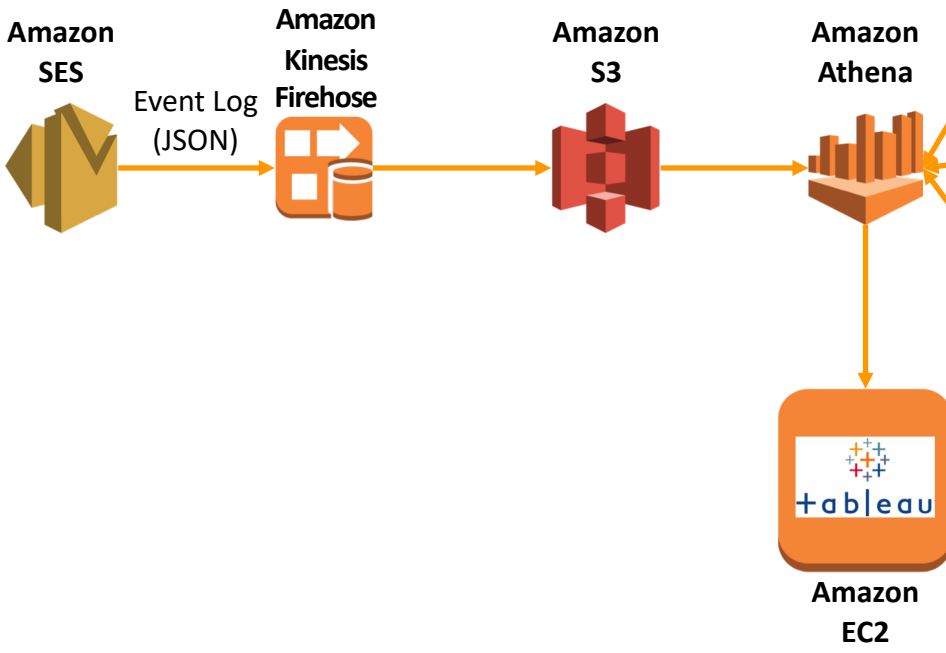


Data Lake on AWS



SES 이벤트 로그의 실시간 분석

- AWS 서비스 로그 분석
- BI 시각화 솔루션 연계



"Which messages did I bounce from Monday's campaign?"

```
SELECT eventtype as Event,
       mail.destination as Destination,
       mail.messageId as MessageID,
       mail.timestamp as Timestamp
FROM sesblog
WHERE eventType = 'Bounce' and mail.timestamp like '2017-01-09%'
```

"How many messages have I bounced to a specific domain?"

```
SELECT COUNT(*) as Bounces
FROM sesblog
WHERE eventType = 'Bounce' and mail.destination like '%amazonses.com%'
```

"Which messages did I bounce to the domain amazonses.com?"

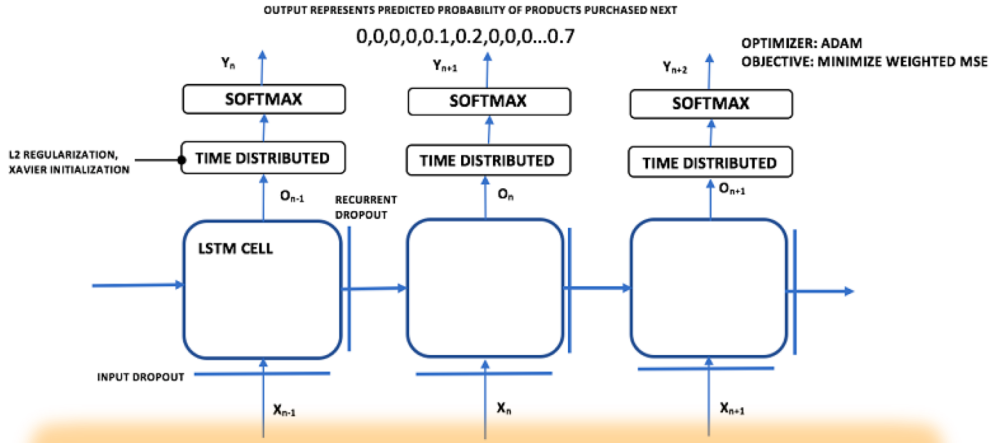
```
SELECT eventtype as Event,
       mail.destination as Destination,
       mail.messageId as MessageID
FROM sesblog
WHERE eventType = 'Bounce' and mail.destination like '%amazonses.com%'
```

https://aws.amazon.com/blogs/big-data/create-tables-in-amazon-athena-from-nested-json-and-mappings-using-jsonserde/

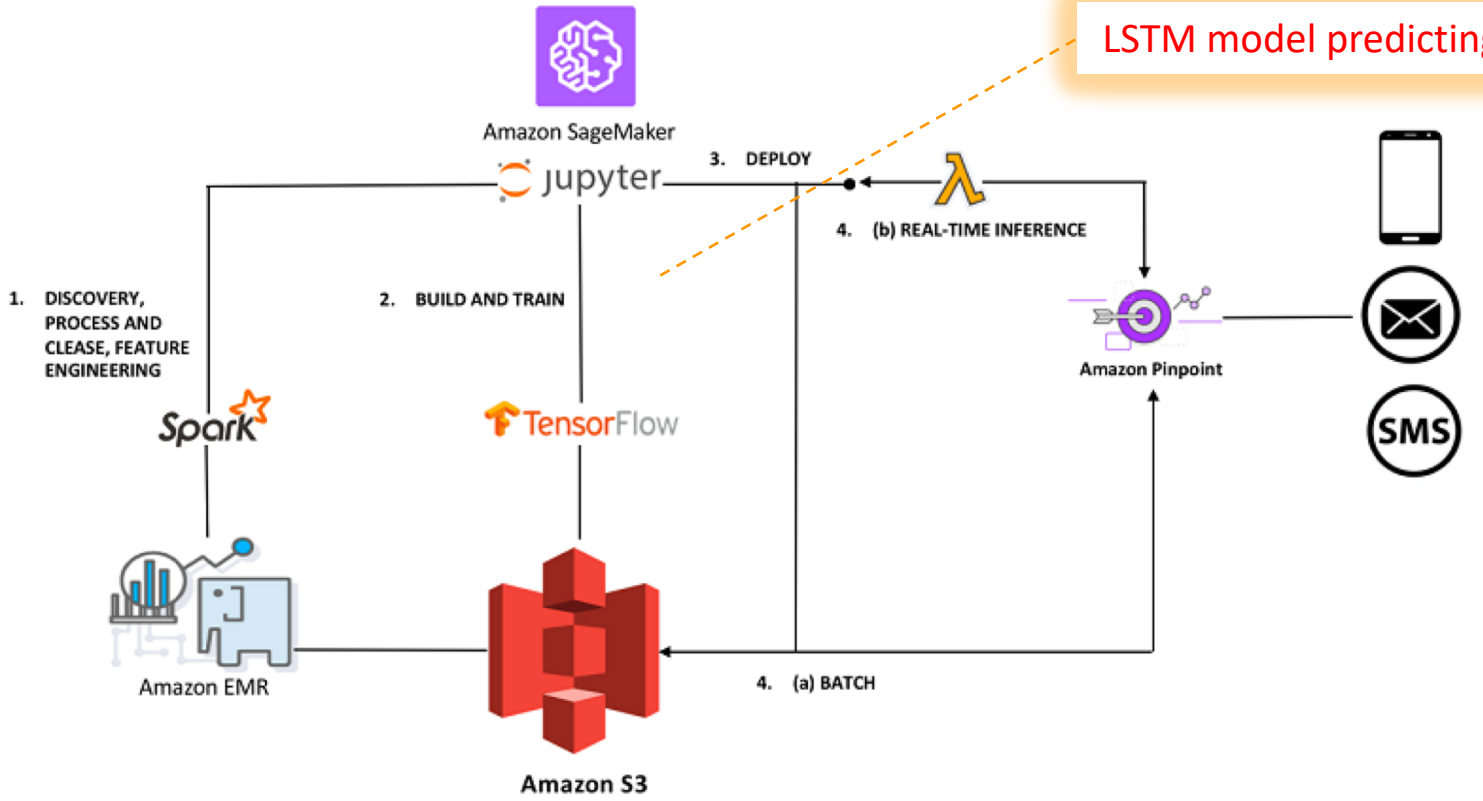
© 2018, Amazon Web Services, Inc. or Its Affiliates. All rights reserved.

머신러닝 기반 마케팅 캠페인

- Amazon S3 + Amazon Pinpoint + Amazon SageMaker



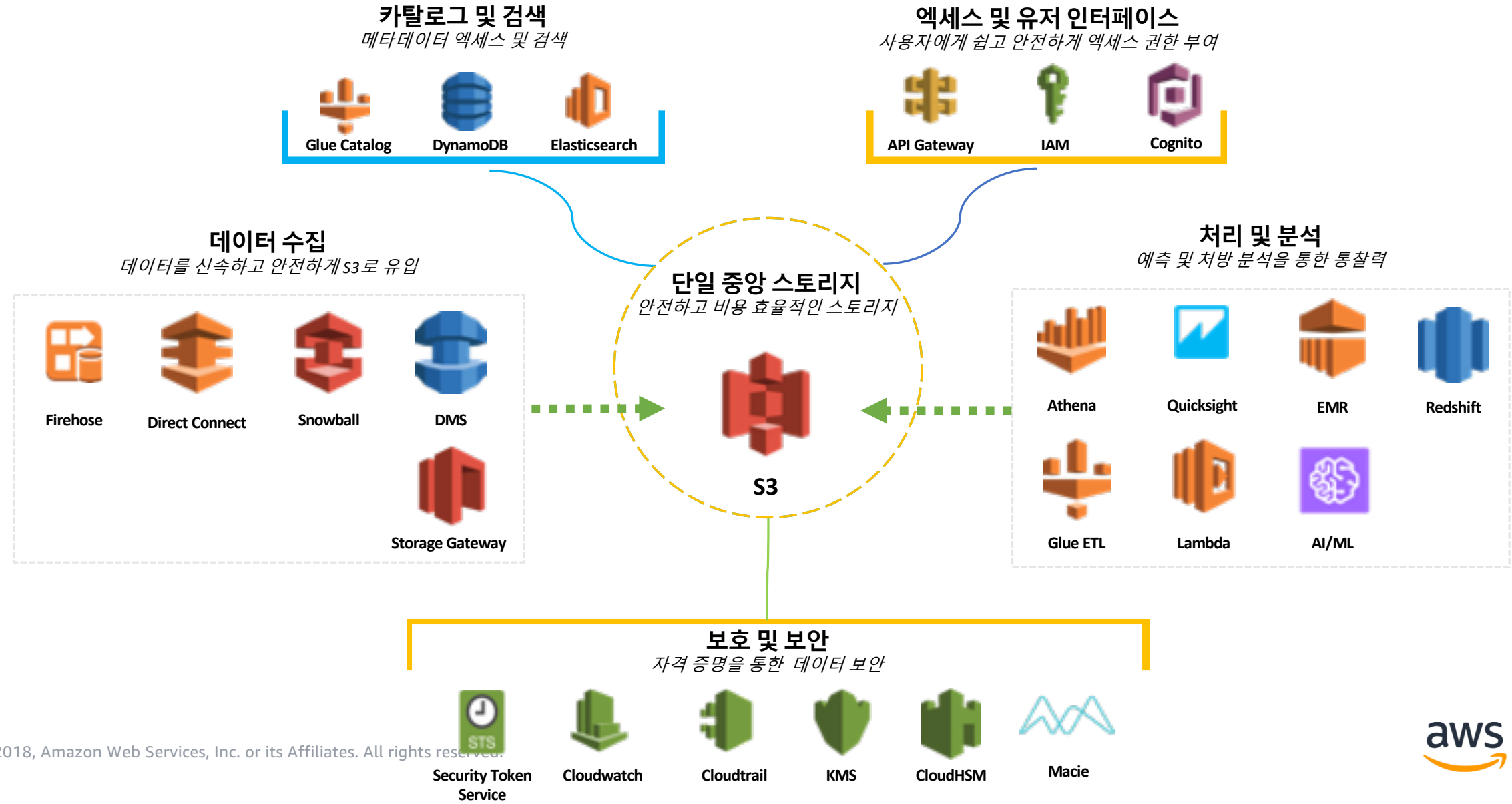
LSTM model predicting next purchased product



https://aws.amazon.com/blogs/machine-learning/amazon-pinpoint-campaigns-driven-by-machine-learning-on-amazon-sagemaker/

© 2018, Amazon Web Services, Inc. or Its Affiliates. All rights reserved.

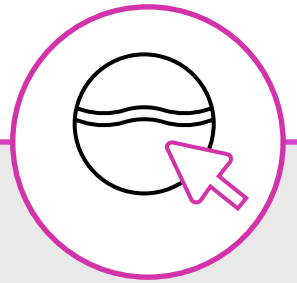
Data Lake on AWS



Amazon Lake Formation (sign up for the preview) NEW!

안전하고 확장 가능한 데이터 레이크를 단 며칠 내에 구축 가능

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 analyst and data scientist to gain and manage new insights

aws

Action Item 제안

준비

실험 진행

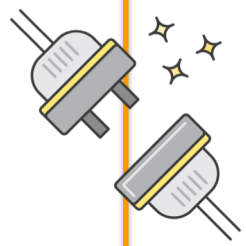
데이터 가치 사슬 (data value chain) 창출

비즈니스 이해관계자 미팅(C-level, BD)

데이터레이크 관련 공감대 형성

현업 담당자와 AWS 전문가 미팅(Data Scientist, SA)

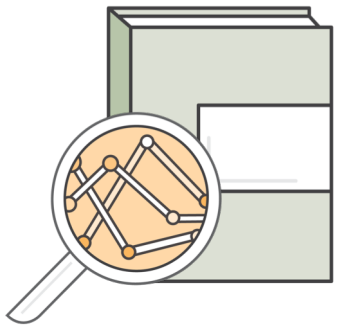
구체적인 과제 목표 설정



데이터 전략 워크숍(BD, SSA, 2 Days)

데이터 레이크 워크샵 / 데이터 사이언스 워크샵 (SSA, ProServ)

데이터 레이크 Accelerator (ProServ, Partner, 90 Days)

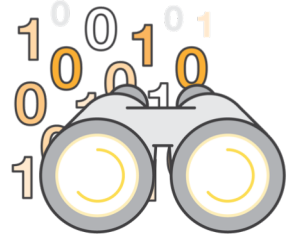


빅데이터 Immersion Day (1 Day, SA)

데이터 사이언스 과제 구체화, 교육 및 기술 지원

Deep Dive 기술 교육 및 PoC

Local Data Lab (SSA, 2 Days)



이제 준비가 되셨습니까?

데이터에 기반한 의사결정

비지니스 사용자가 자유롭게 데이터 접근 – 잘 활용되고 관리되는 데이터

빠른 시장 대응

민첩하고 반복적인 디자인 – 신속한 신제품 및 서비스 출시

실험과 혁신 문화

기계 학습 및 데이터 사이언스를 이용한 이벤트 모델링 및 예측