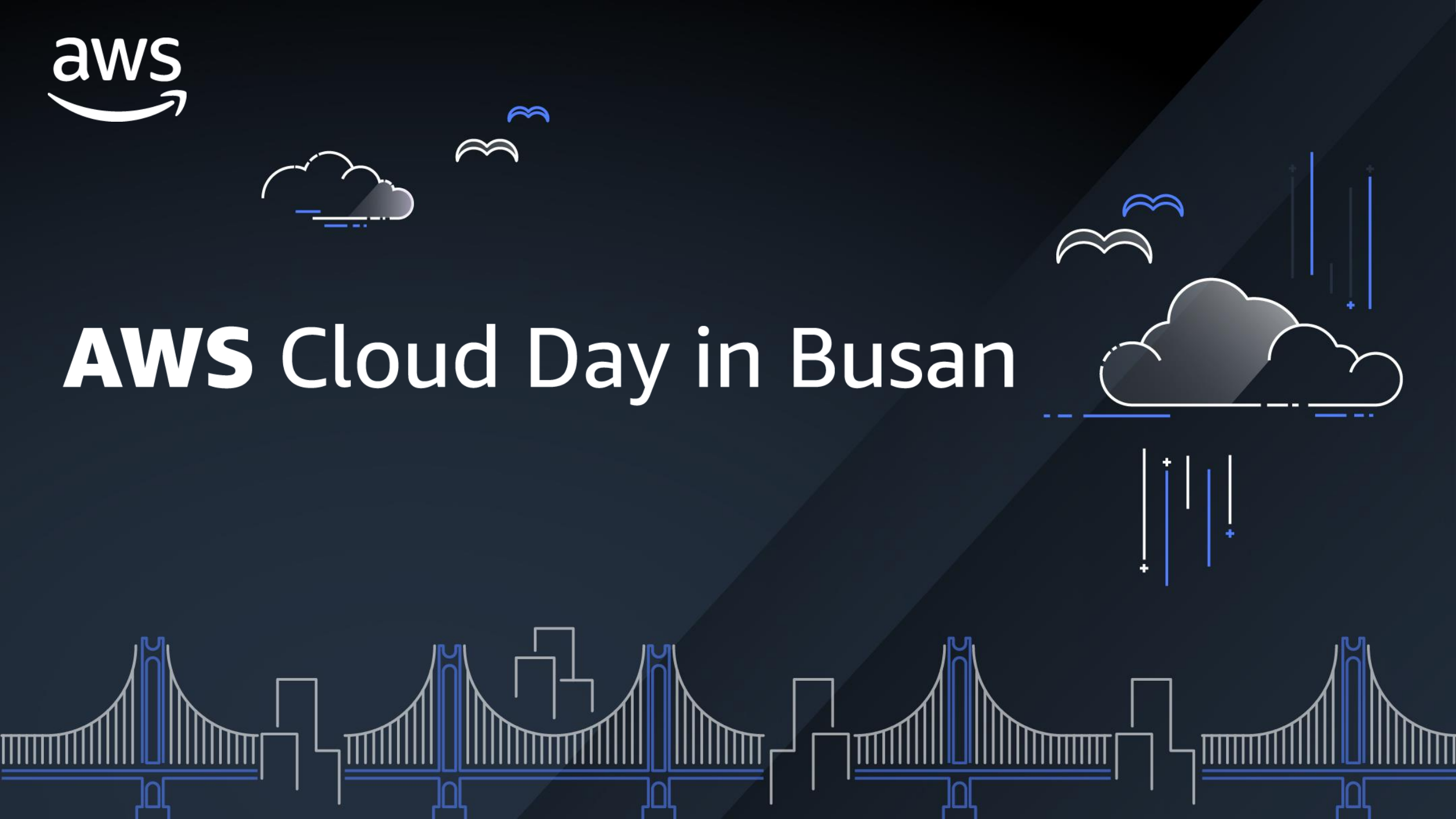

aws
AWS Cloud Day in Busan

AWS Cloud Day in Busan

Amazon Polly

지금 나오는 안내멘트는
AWS 음성기술인 '서연'을 통해 100% 구현되었습니다.

AWS Cloud Day in Busan

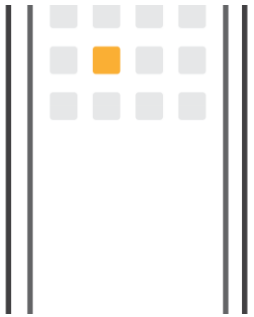
데이터 기반 통찰력 키우기: AWS를 활용한 효과적인 분석

Hyobin An

Gaming Solutions Architect

Amazon Web Services

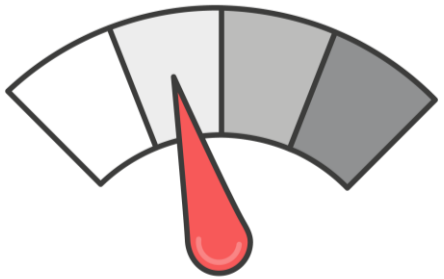
데이터는 계속하여 발생



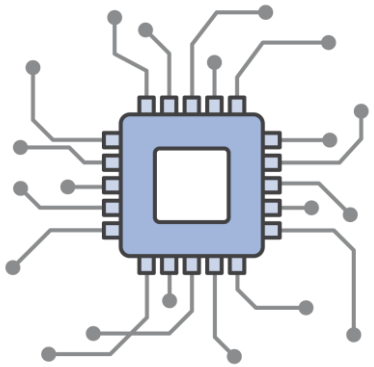
모바일 앱

웹 클릭스트림

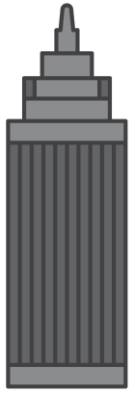
애플리케이션 로그



미터링 기록

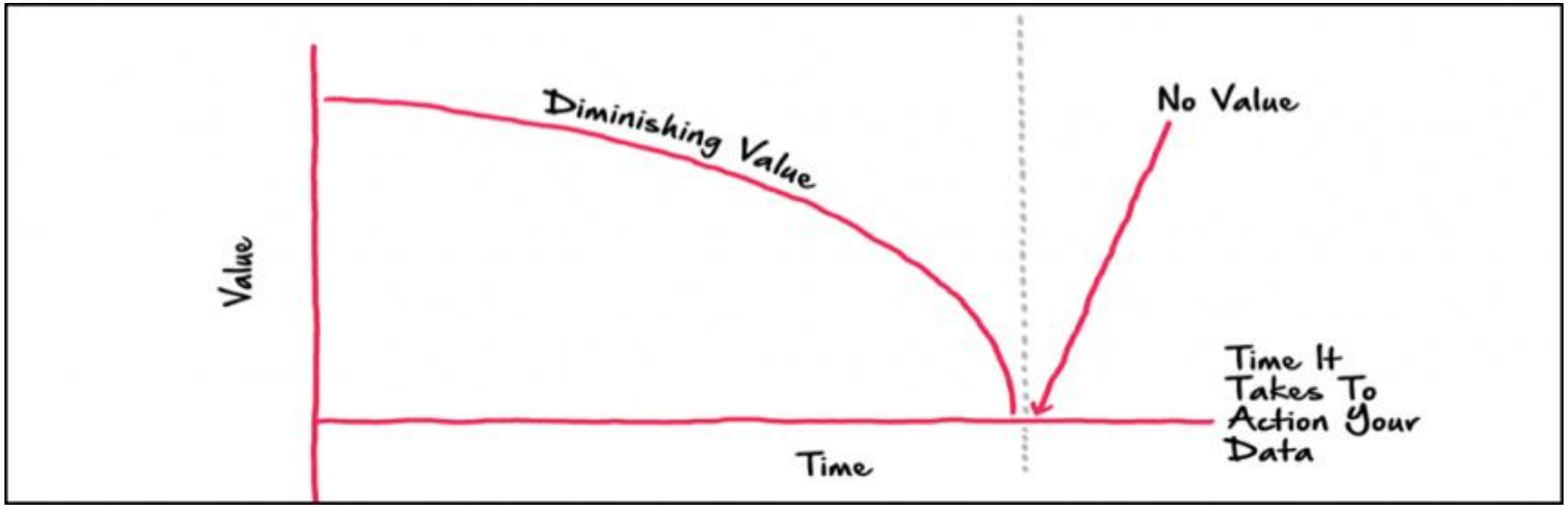


IoT 센서

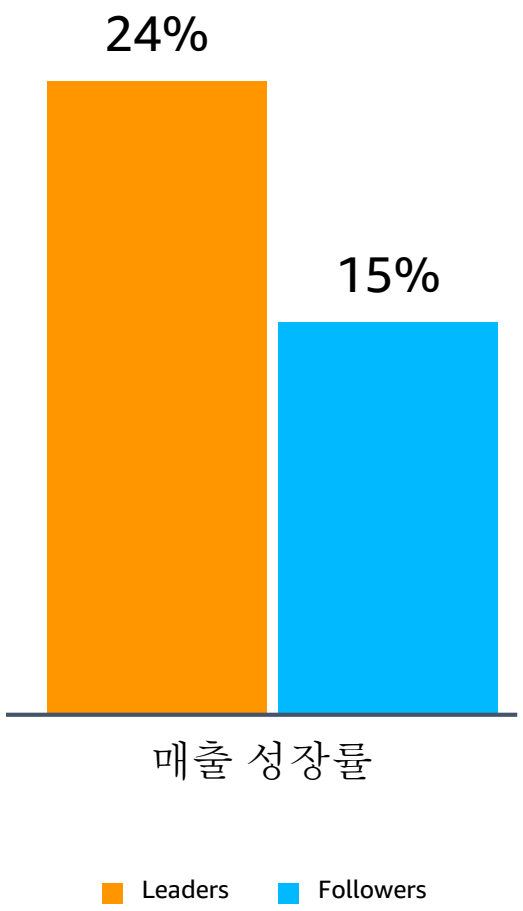


스마트 빌딩

빠른 데이터 분석의 필요성



데이터 분석의 중요성



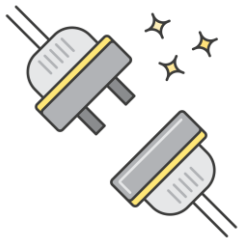
다양한 데이터를 효율적으로 수집하고 분석하여 활용하고 있는 회사는 경쟁 업체 대비 9% 높은 매출 성장률을 보입니다.

*Aberdeen: Angling for Insight in Today's Data Lake, Michael Lock, SVP Analytics and Business Intelligence

데이터를 잘 사용하기 어려운 장애물들

"데이터를 사용하는데 어떤 어려움을 겪고 있나요?"

여러 데이터에 접근하거나 함께 연결할 수 없다.



99%

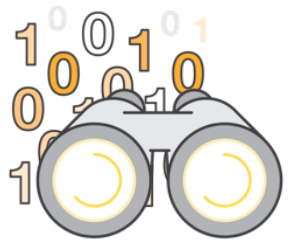
잠재적인 가치를 가진 대부분의 데이터가 사용되지 못함

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

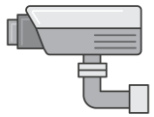
기하급수적으로 늘어나는 데이터


Transactions


Billing


ERP


Web logs


Sensor Data

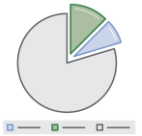

Infrastructure logs

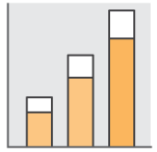

Social

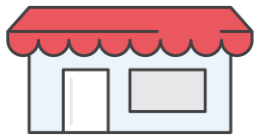
Dark Data
복잡한 전처리

다양한 데이터 소비자들


Data Scientists

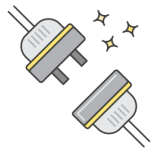

Applications

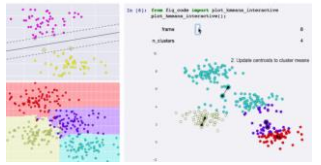

Business Analyst

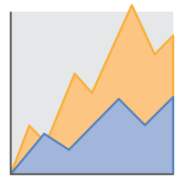

External Consumers

데이터의 중복
원본데이터 관리

많은 접근 방식과 툴들


API Access


Notebooks


BI Tools

다양한 기술 지원
전문가의 부족

고객이 데이터 플랫폼에 투자하는 이유

데이터에 기반한 의사결정

비지니스 사용자가 자유롭게 데이터 접근 – 잘 활용되고 관리되는 데이터

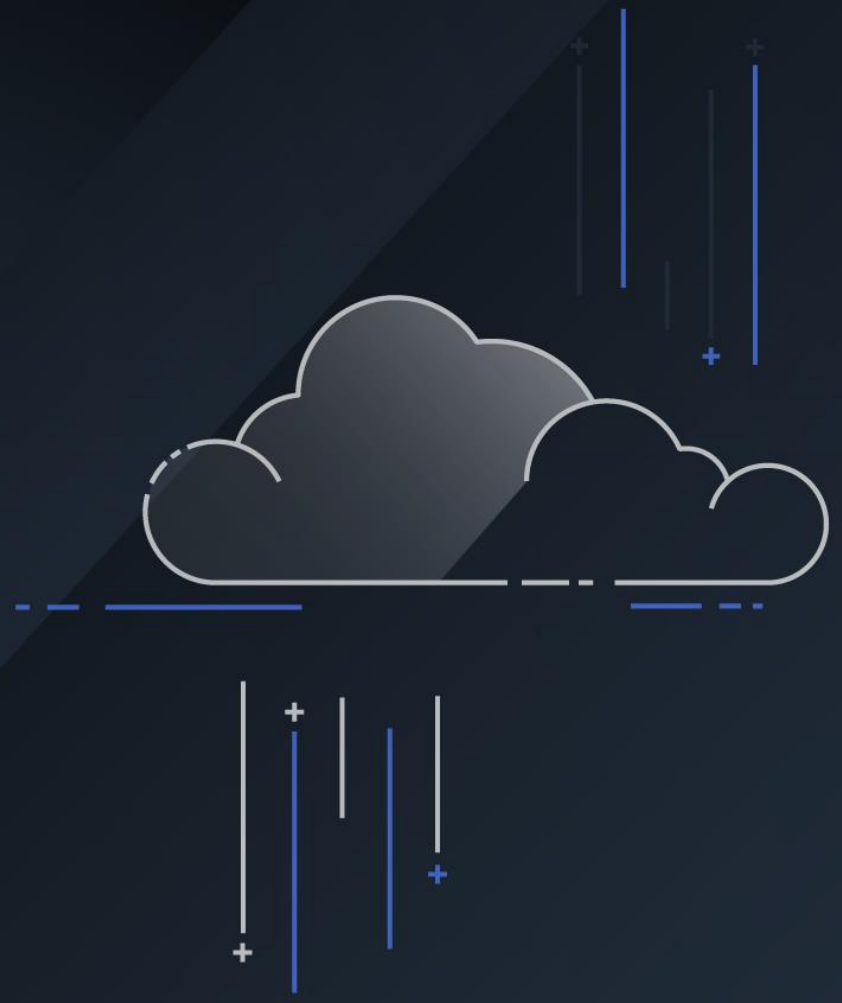
빠른 시장 대응

민첩하고 반복적인 디자인 – 신속한 신제품 및 서비스 출시

실험과 혁신 문화

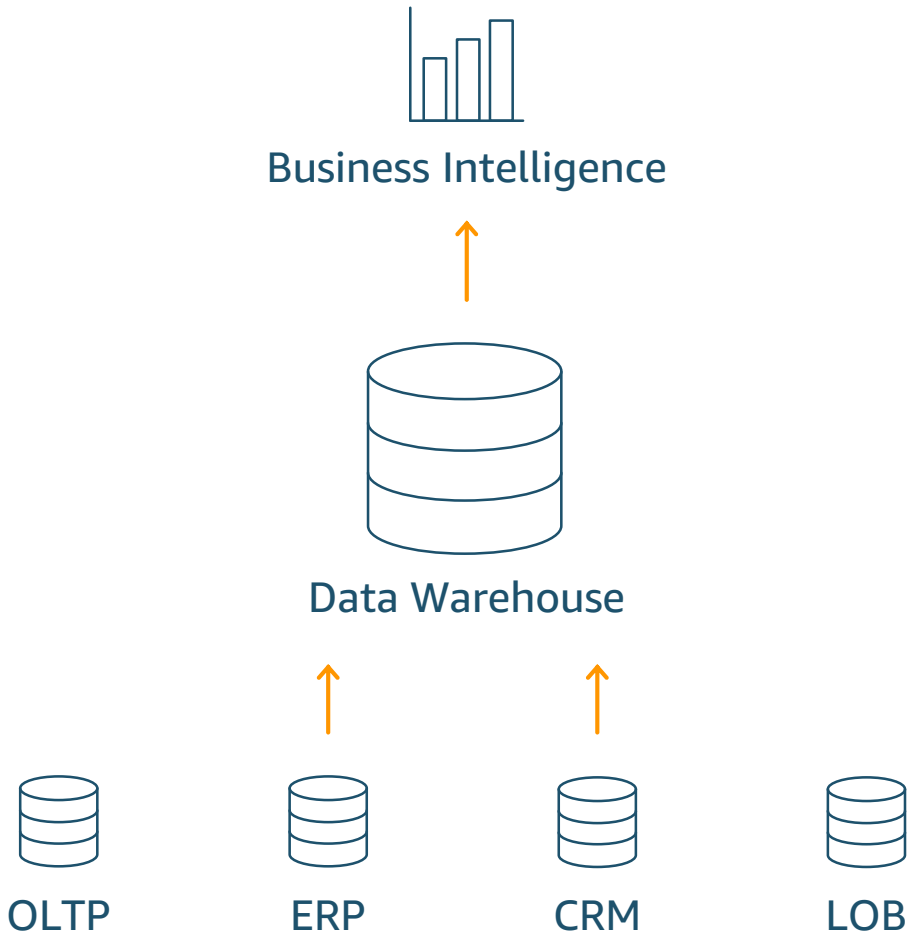
기계 학습 및 AI, 데이터 사이언스를 이용한 모델링 및 이벤트 예측

AWS Cloud Day in Busan



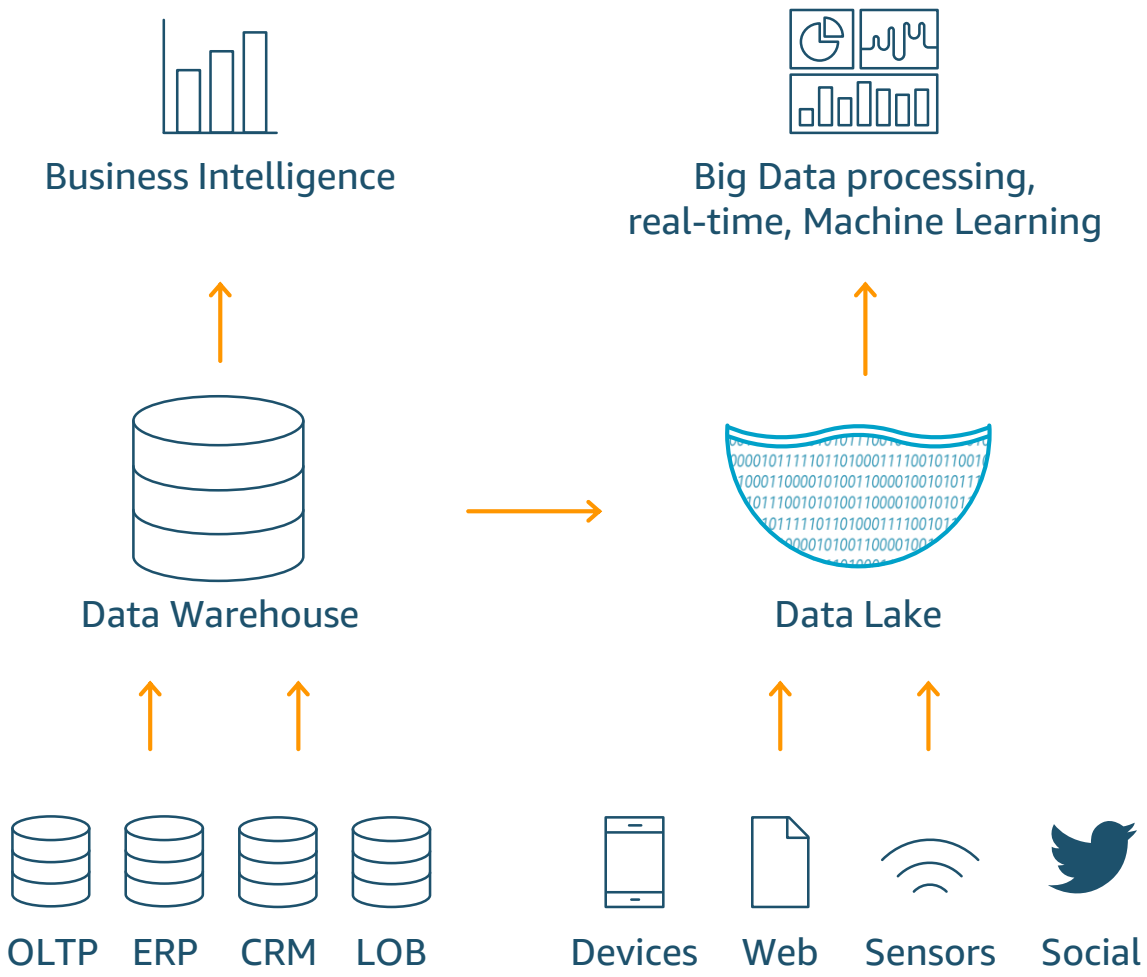
Data Lake

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

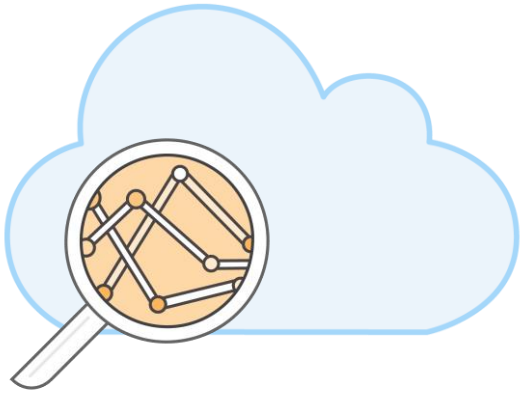


- 다양한 유형의 정형, 비정형 데이터 저장
- TBs–EBs scale
- 인사이트를 얻기 위해 다양한 분석 엔진
- 낮은 비용으로 저장과 분석이 가능

데이터 레이크 – 모든 데이터가 한곳에

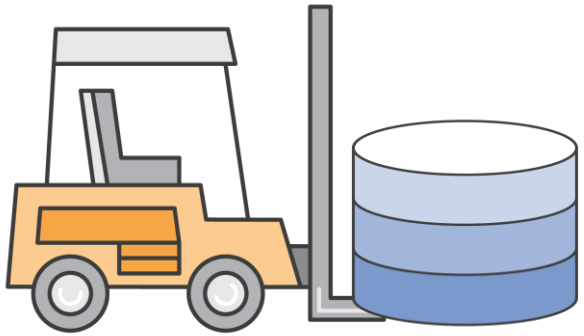
"왜 데이터가 여러 장소에 분산 되어 있는가?

어떤 데이터가 정말 원본 데이터 인가?"

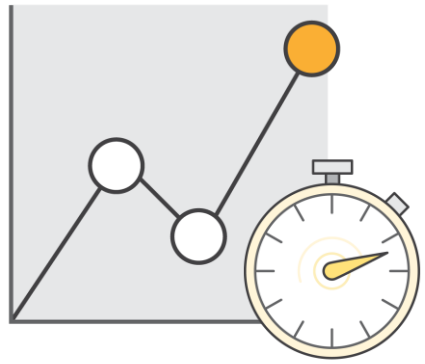


하나의 중앙 저장소에
모든 소스로부터 오는 모든 종류의 데이터를 저장하고 분석

데이터 레이크 – 빠른 데이터 수집

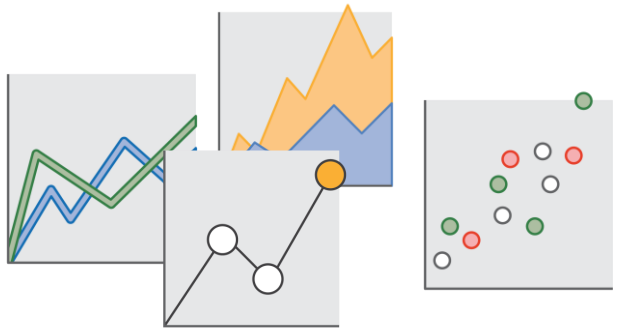


"어떻게 다양한 소스로부터의 데이터를 빠르게 수집하여 효율적으로 저장할 수 있을까?"



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

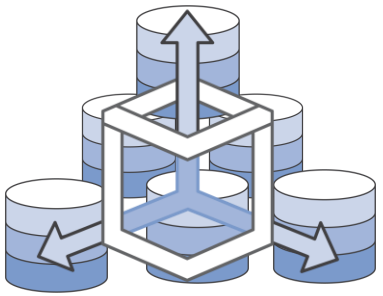
데이터 레이크 – 사용 시점에 스키마 정의



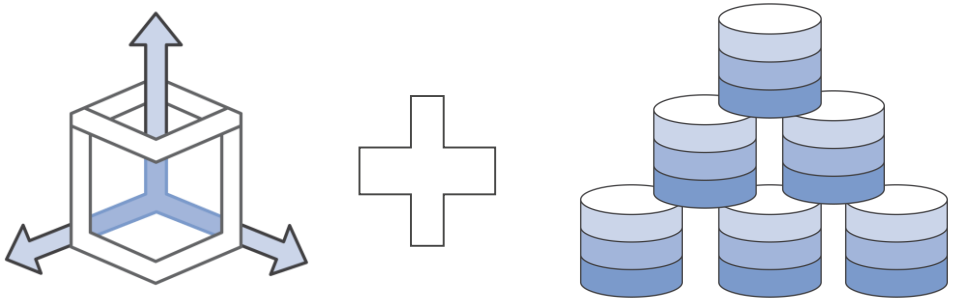
"여러 종류의 분석툴과 프로세싱 엔진에서 같은 데이터를 같이 사용할 수 있는 방법이 있는가?"

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

데이터 레이크 – 데이터 저장과 처리를 분리



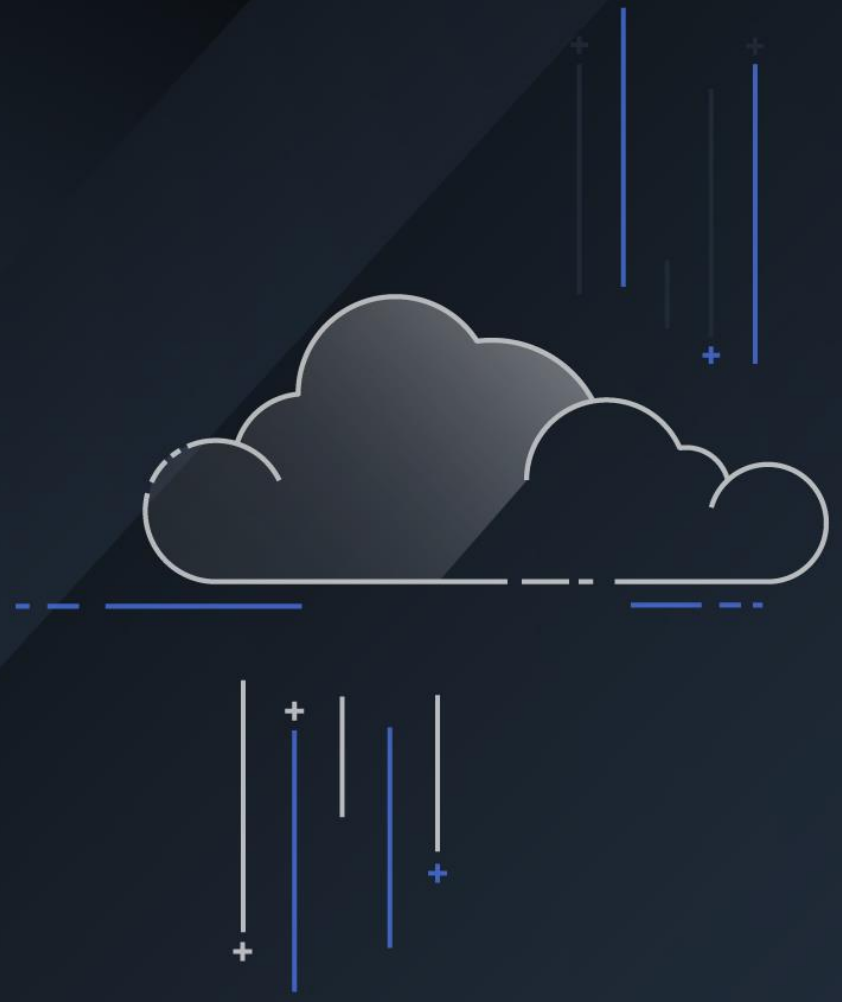
"급격히 늘어나는 데이터에 맞게 어떻게 시스템을 스케일업 할 것인가?"



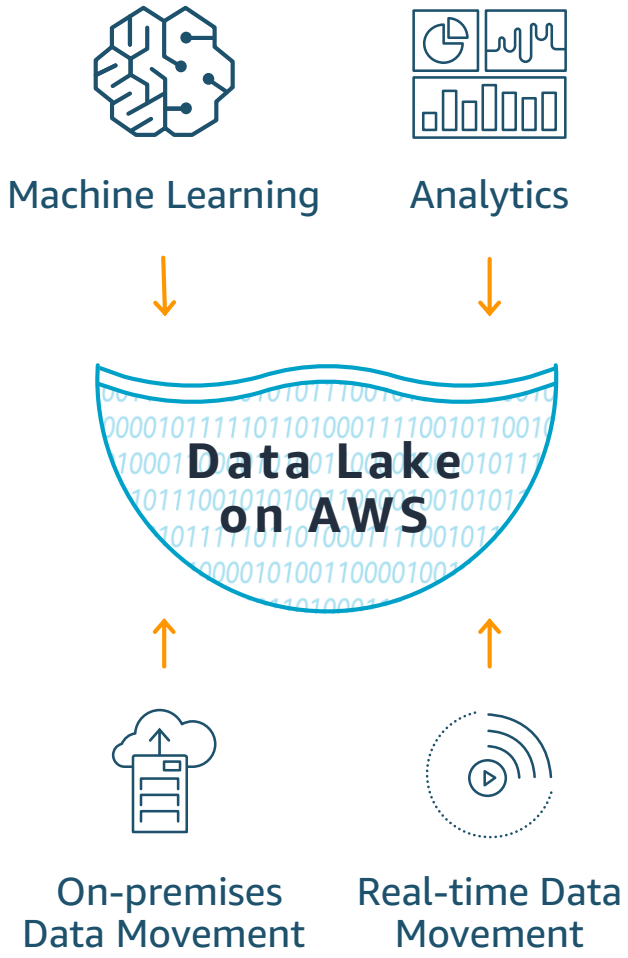
데이터 저장공간과 분석을 위한 컴퓨팅 리소스를 분리
필요한 리소스만 언제든지 추가 가능

AWS Cloud Day in Busan

Data Lake on AWS



AWS 를 활용한 Data Lake 와 분석



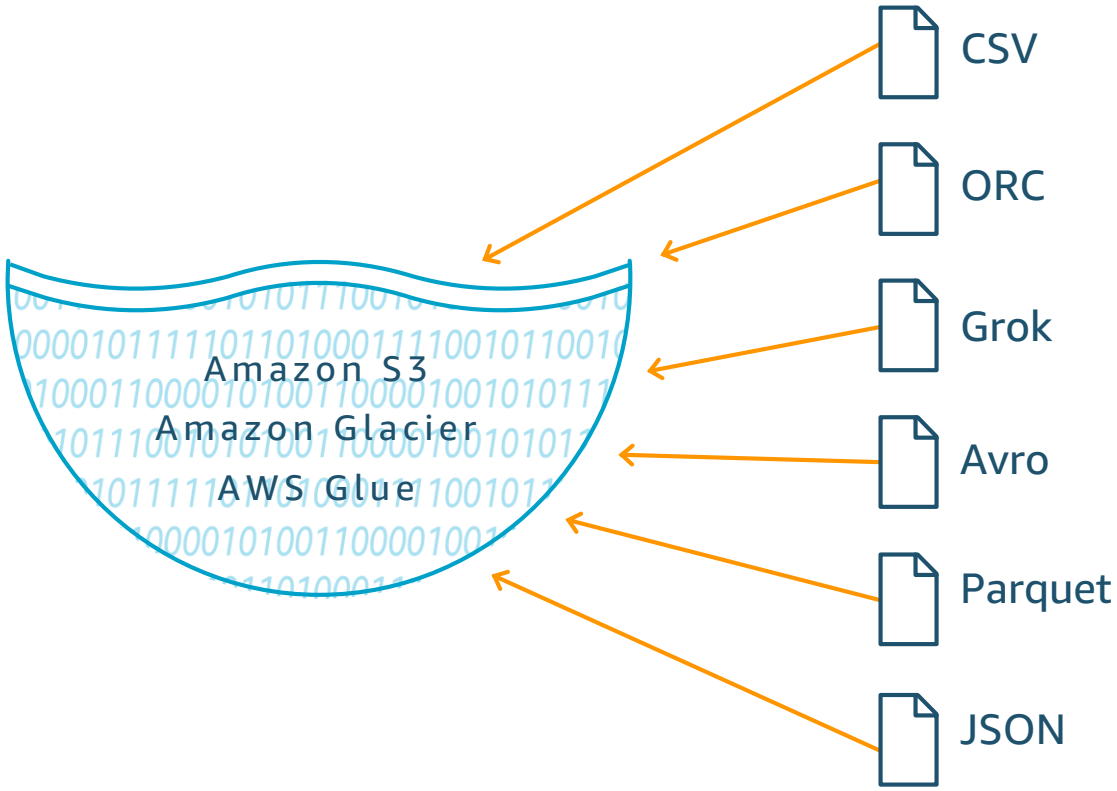
다양한 오픈 소스 지원

안전하게 저장

확장성과 내구성

비용 효율적

원하는 유형의 모든 포맷의 데이터 저장 가능



- 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 Lake 로 데이터를 가져오는 대부분의 방법

데이터 센터로부터 데이터 이동

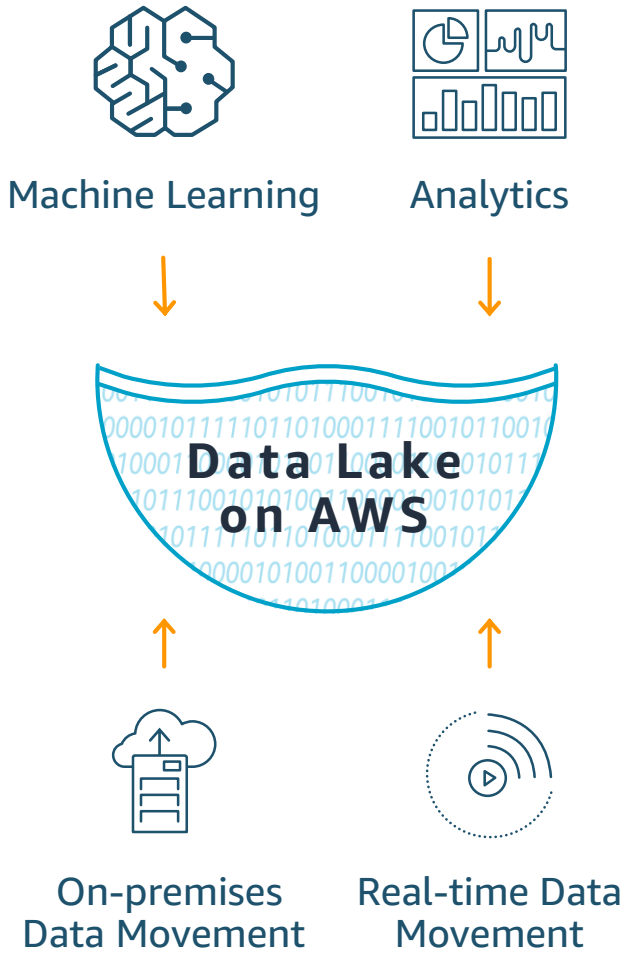
AWS Direct Connect
AWS Snowball
AWS Snowmobile
AWS Database Migration Service
AWS Storage Gateway

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

AWS 를 활용한 Data Lake 와 분석



다양한 오픈 소스 지원

안전하게 저장

확장성과 내구성

비용 효율적

AWS는 가장 높은 수준의 보안 제공

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

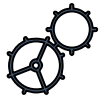
AWS Certification Manager

AWS Key Management Service

Encryption at rest

Encryption in transit

Bring your own keys, HSM support



Compliance

AWS Artifact

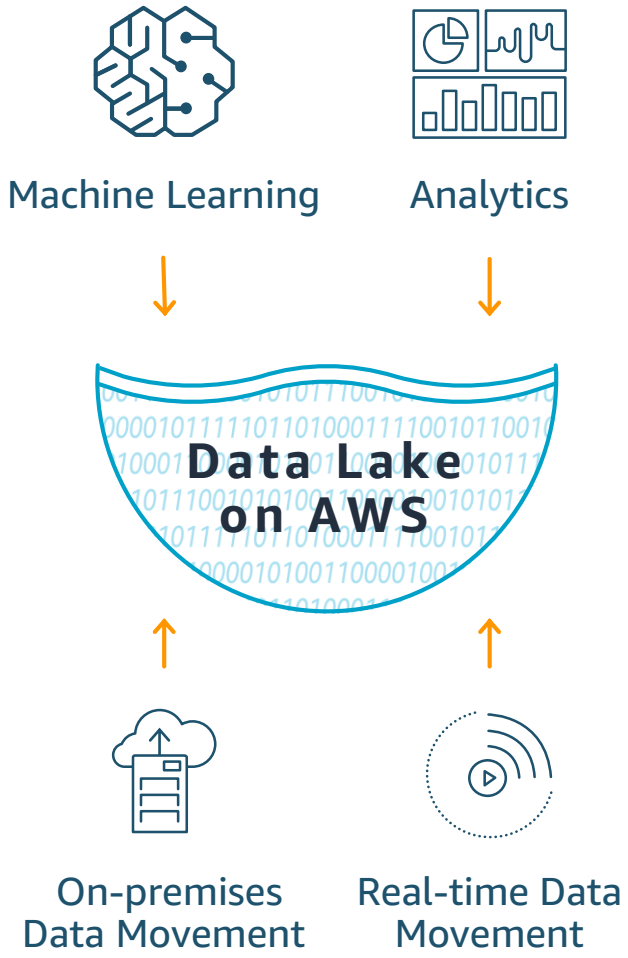
Amazon Inspector

Amazon Cloud HSM

Amazon Cognito

AWS CloudTrail

AWS 를 활용한 Data Lake 와 분석



다양한 오픈 소스 지원

안전하게 저장

확장성과 내구성

비용 효율적

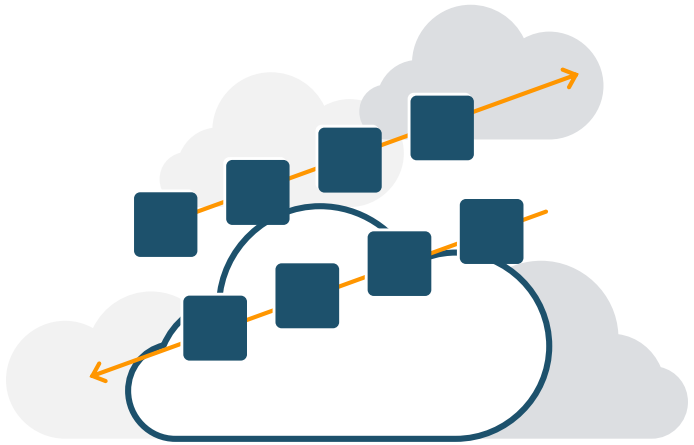
AWS는 세계 최대 규모의 클라우드 인프라 운영

"...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

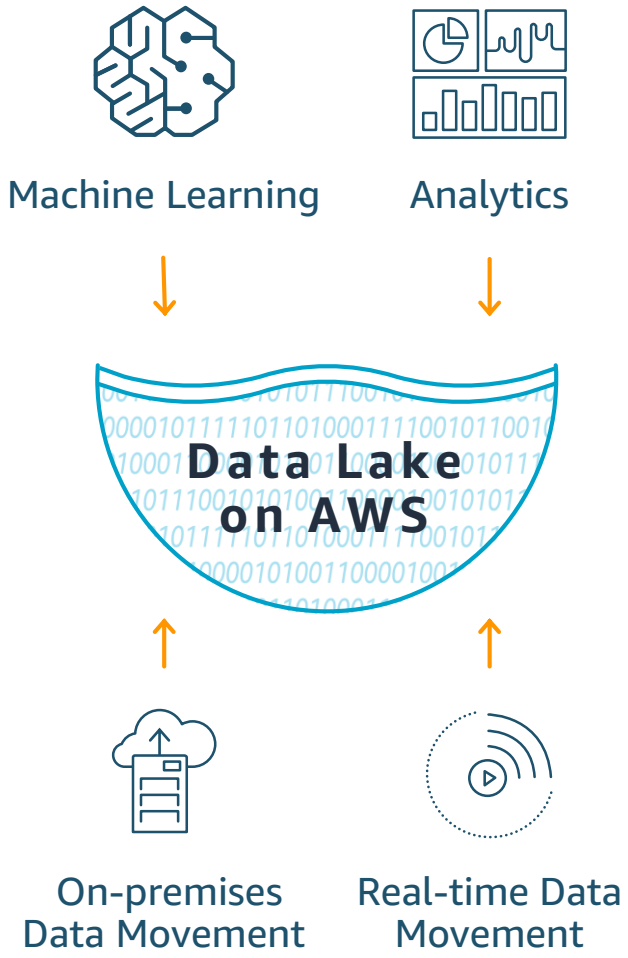


- S3에 수 조개의 객체와 엑사바이트 급의 데이터 저장
- 어떤 크기의 데이터도 저장 가능
- 어떤 크기의 컴퓨팅 자원도 수 분만에 스핀업 하여 대규모의 분석 엔진 실행
- 전 세계에서 가장 큰 클라우드 인프라에서 실행

탁월한 내구성과 가용성

- 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

AWS 를 활용한 Data Lake 와 분석



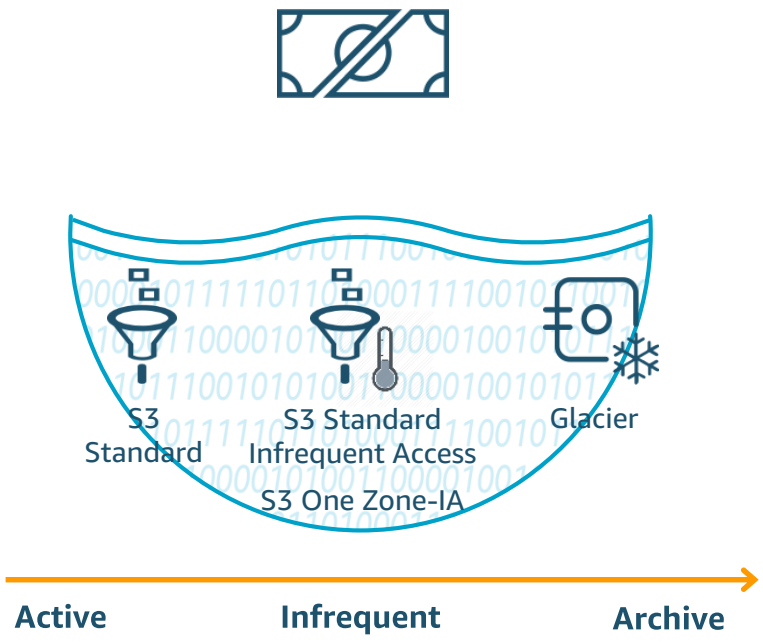
다양한 오픈 소스 지원

안전하게 저장

확장성과 내구성

비용 효율적

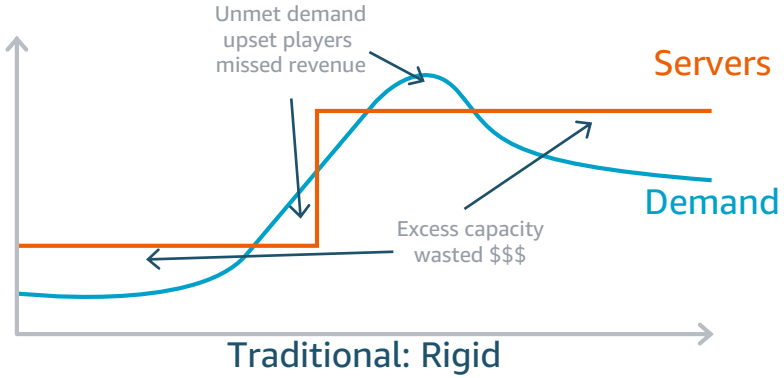
비용/성능 최적화를 위한 Tiered Storage



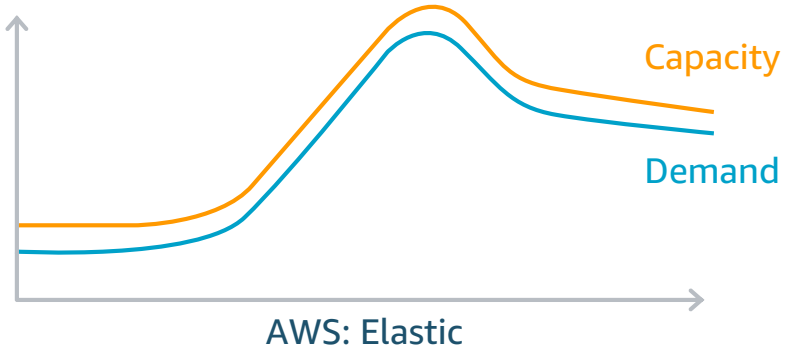
- 가격/성능 최적화를 위해 Tiered storage 사용
  - S3 Standard
  - S3 Standard—Infrequent Access
  - S3 One Zone—Infrequent Access
  - Amazon Glacier
- 생명주기 정책 기반으로 티어 간 마이그레이션
- S3에 데이터 저장 시 $0.023/GB/month
- Glacier에 데이터 저장 시 $0.004/GB/month

사용한 만큼만 지불

전통적 접근방식: 용량 낭비로 이어짐

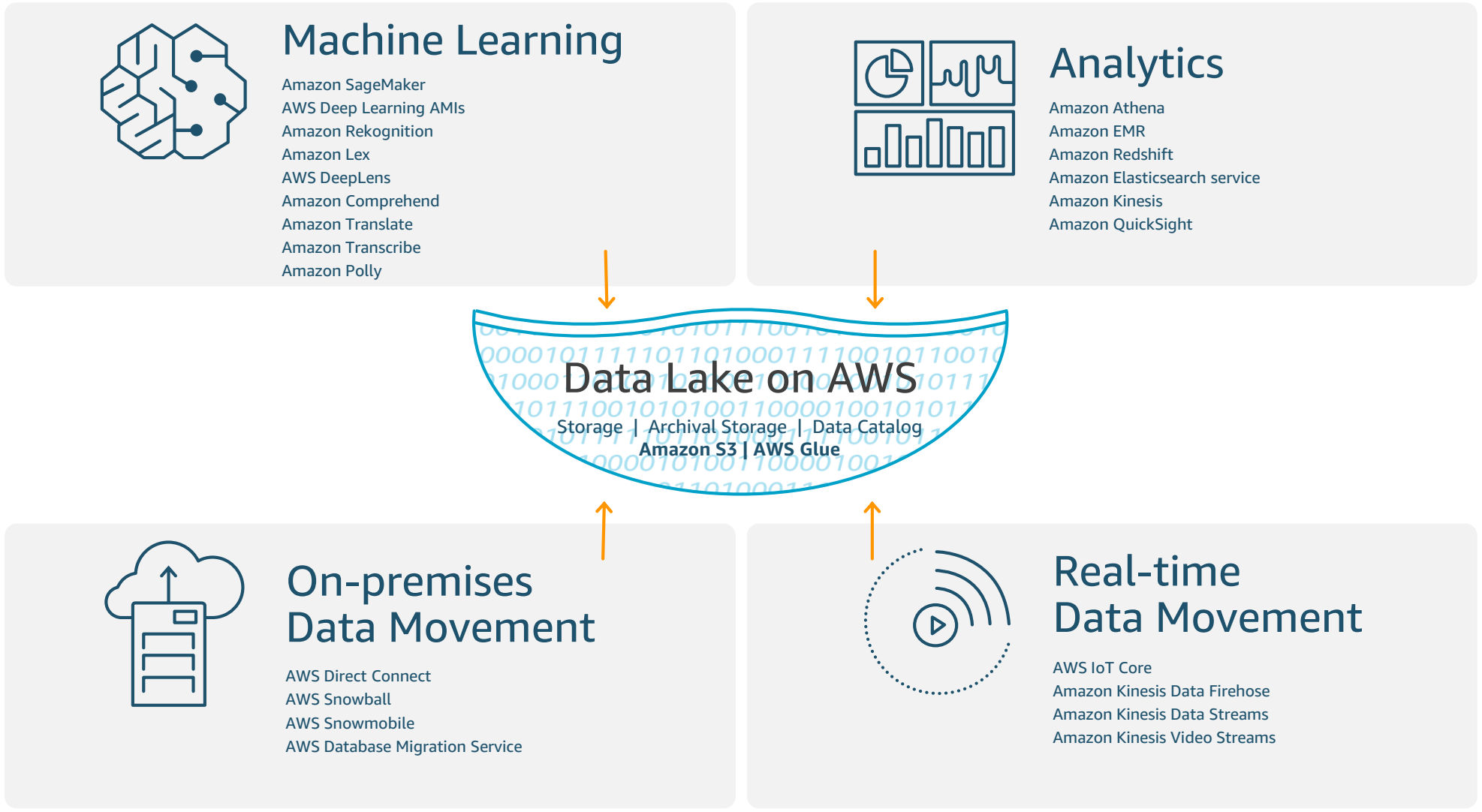


AWS 접근방식: 사용한 용량만큼 지물

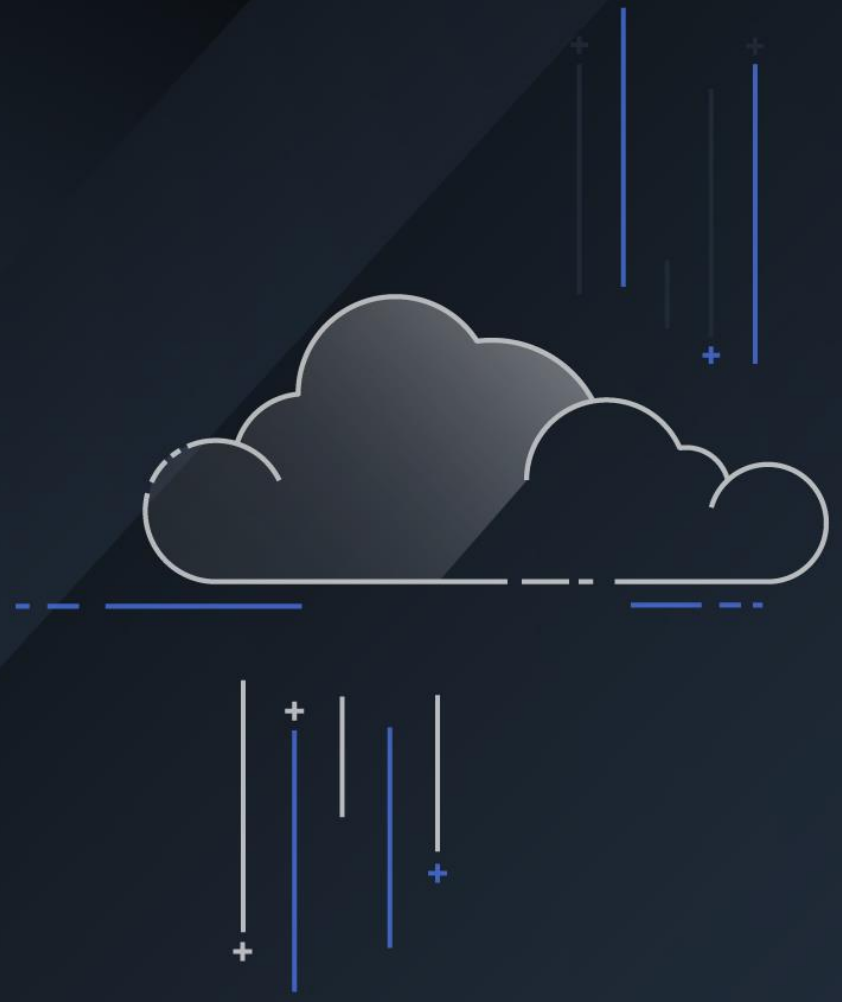


- 필요한 자원에 대한 주문형 서비스
- Athena 스캔 기준 $0.05/GB
- EMR과 Athena 는 업무가 완료된 후 자동으로 자원 축소 가능하므로 비용 절약이 가능
- 예약 인스턴스 (RI)를 통해 일정 기간 동안 Commit하면 최대 75% 절약 가능
- 여분의 컴퓨팅 용량으로 EMR에서 클러스터를 실행하여 스팟 인스턴스로 최대 90% 절약 가능

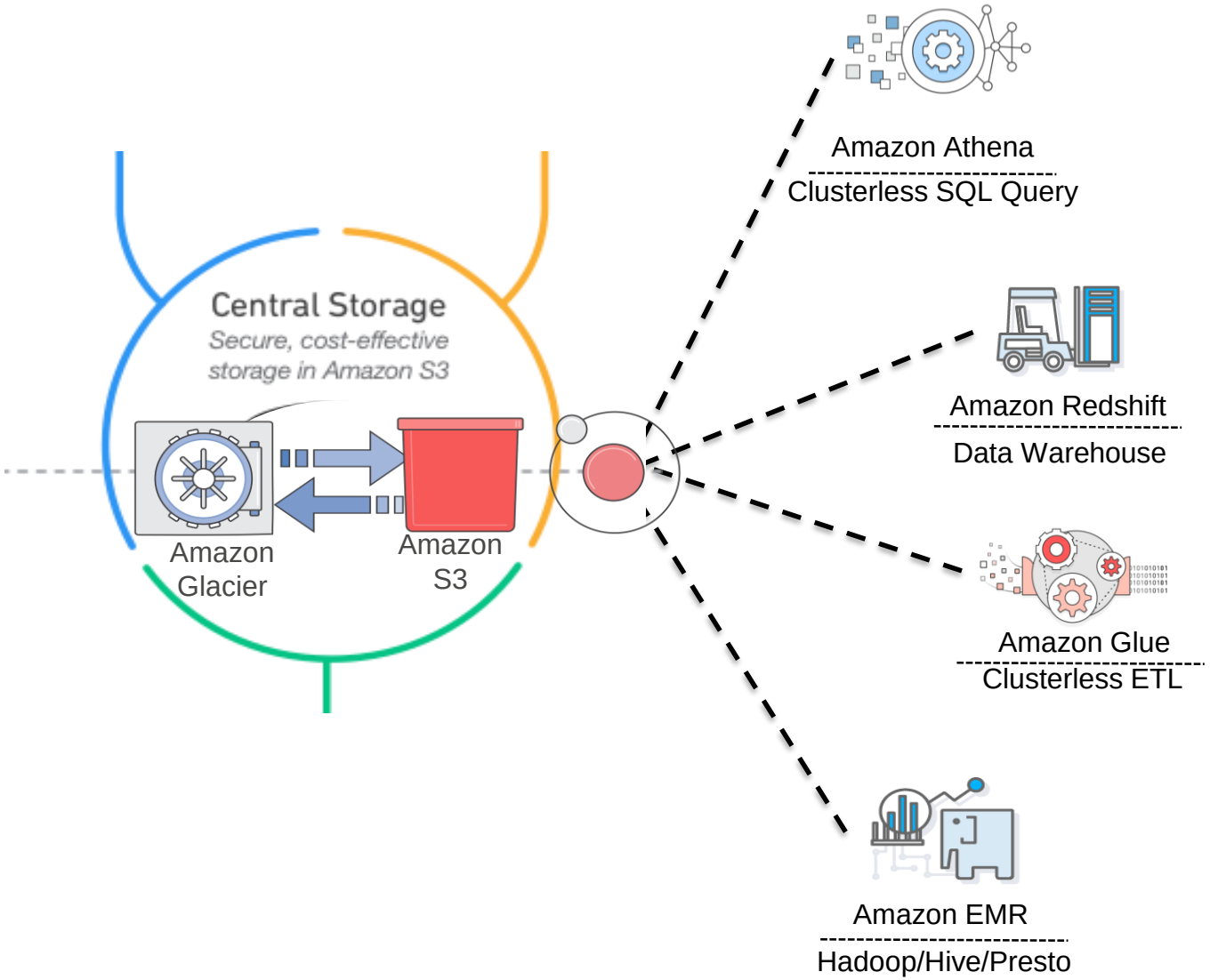
Data Lake 활용이 가능한 분석, ML 포트폴리오



Data Lake Use Cases



배치 프로세싱



BI & Visualization

Amazon QuickSight

looker

tableau

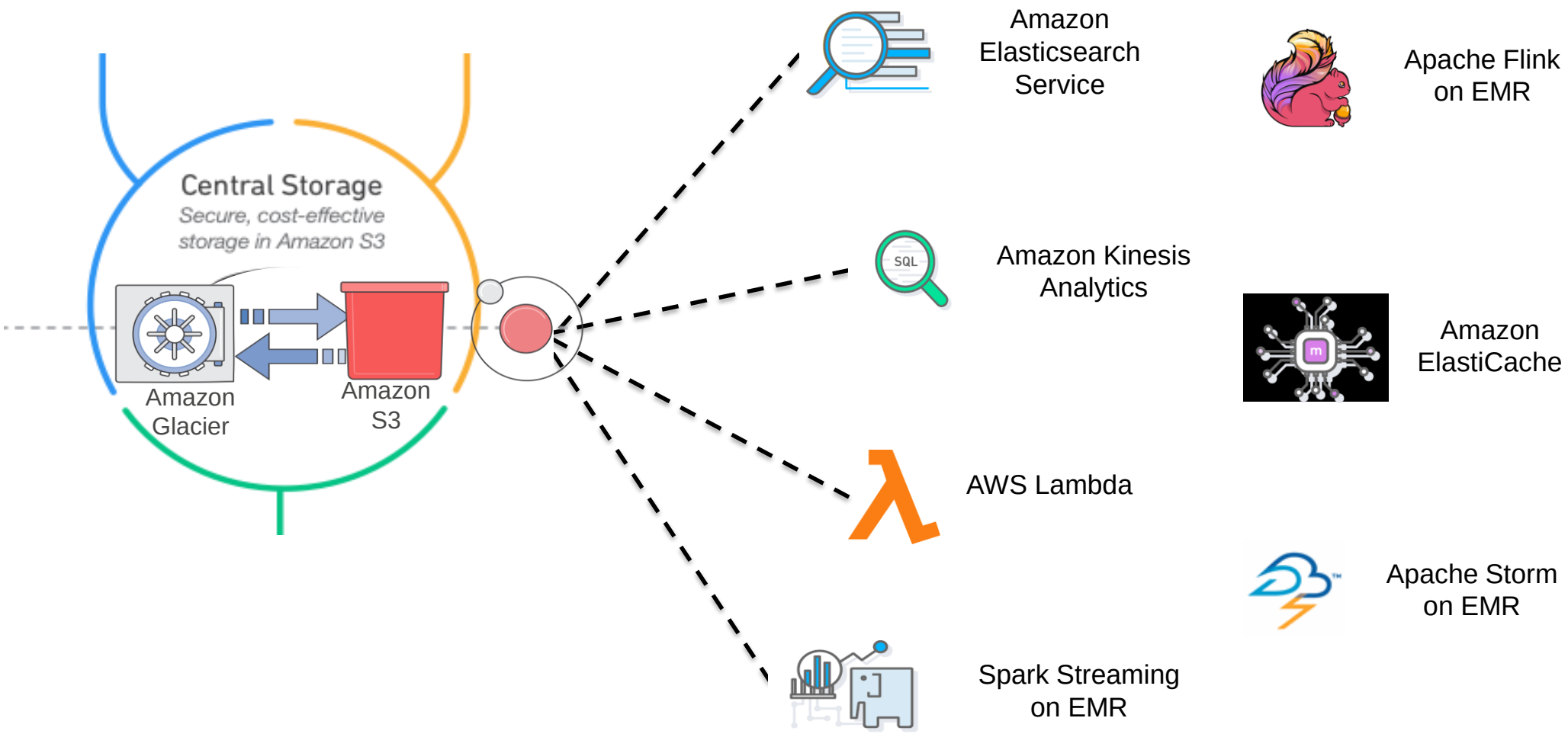
kibana

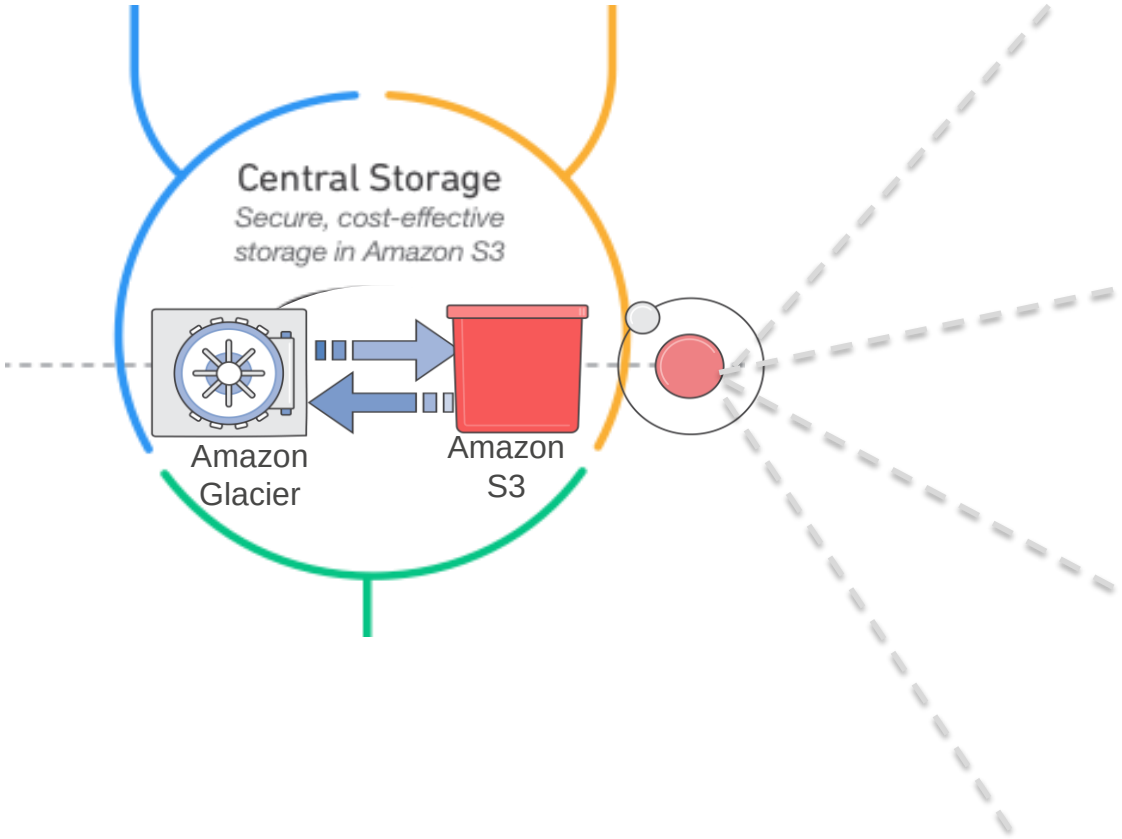
COGNOS

SAP Business Objects

MicroStrategy

스트리밍 데이터 실시간 분석





Amazon Polly
Life-like speech

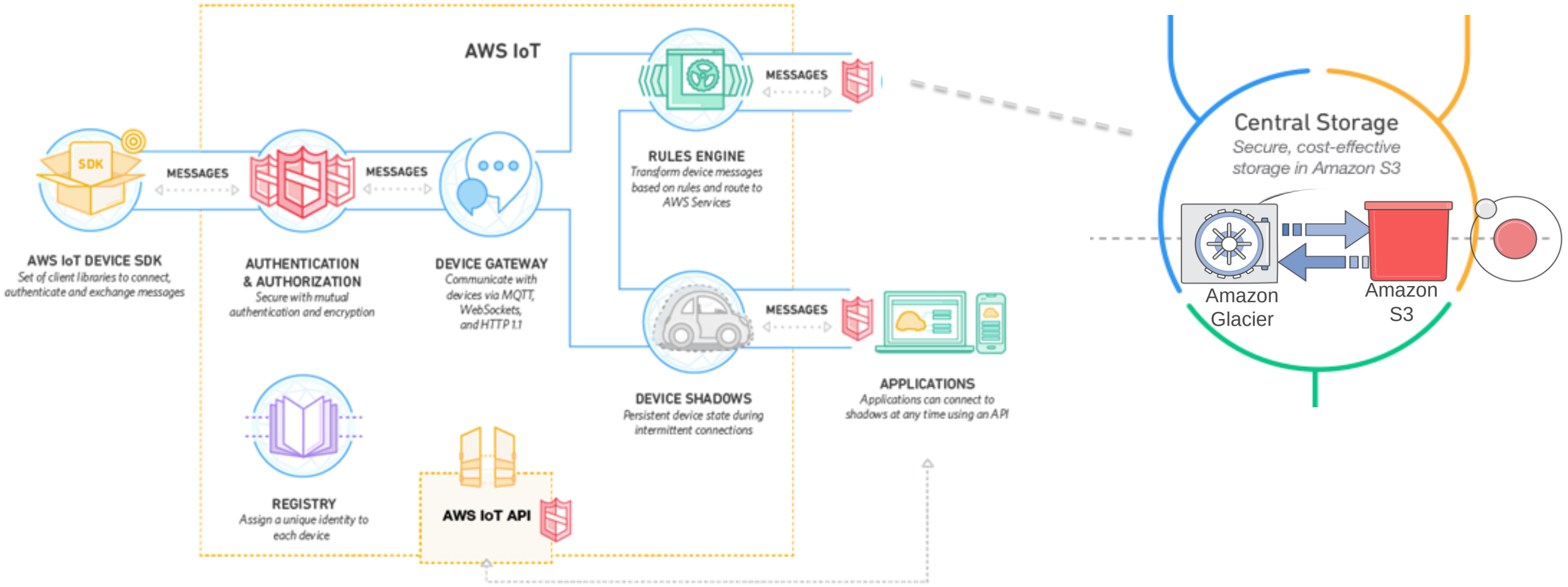
Amazon Lex
Conversational engine

Amazon Rekognition
Image analysis

Amazon SageMaker
Machine learning platform

Deep learning Frameworks
MXNet, TensorFlow, Theano, Caffe, Torch

IoT 데이터 분석



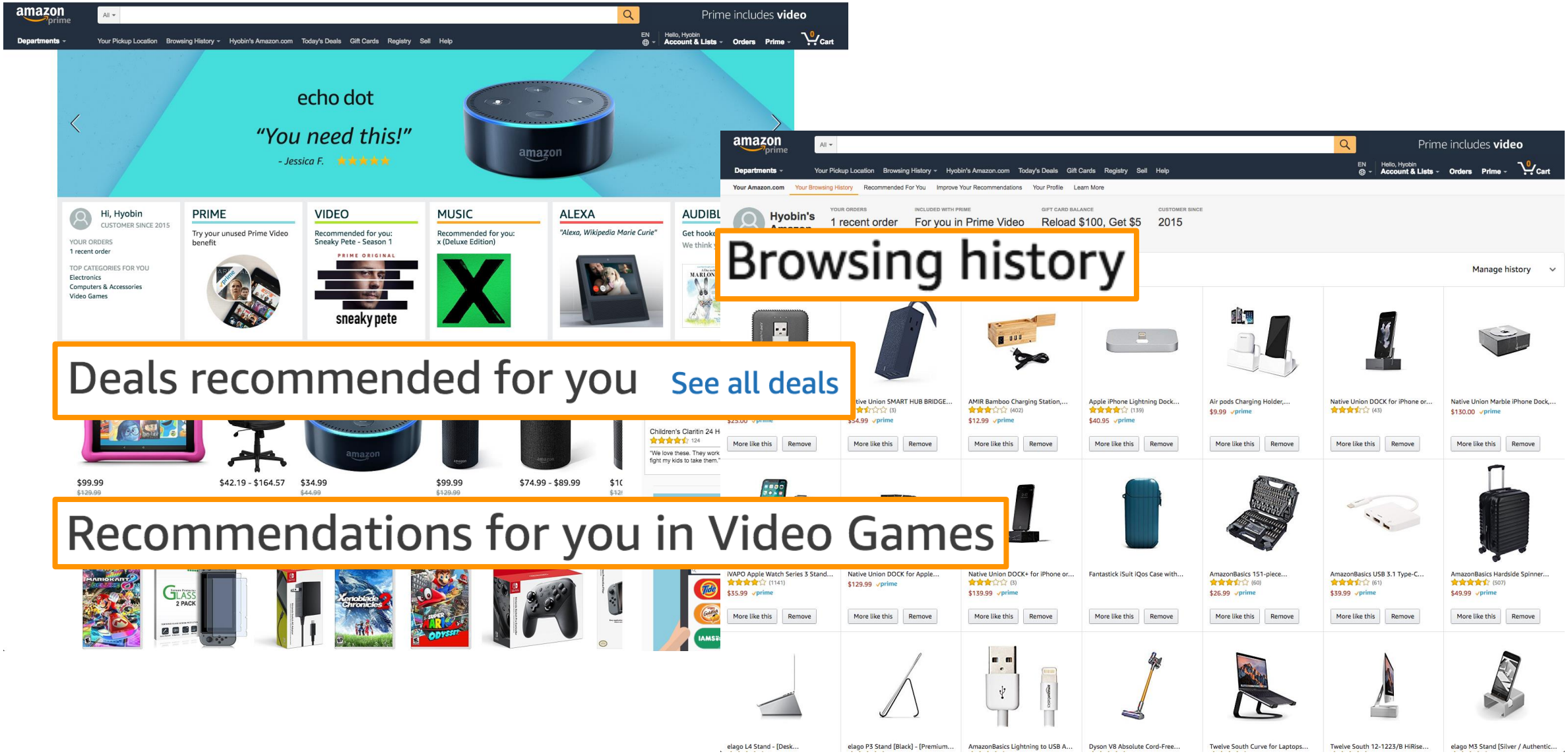
AWS Cloud Day in Busan

실시간 스트리밍 데이터 분석 데모 ☺

분석할 데이터


```
{
  "Category": "Technology",
  "City": "Chicago",
  "Profit": 11.1944,
  "Country": "United States",
  "Region": "Central",
  "Sub-Category": "Phones",
  "State": "Illinois",
  "Customer Name": "Ken Lonsdale",
  "Postal Code": 60610,
  "Row ID": 241,
  "Quantity": 2,
  "Product ID": "TEC-PH-10000011",
  "Customer ID": "KL-16645",
  "Sales": 31.984,
  "Ship Mode": "Second Class",
  "OccurenceTime": "2018-04-02T06:14:00.480866",
  "Discount": 0.2,
  "Product Name": "PureGear Roll-On Screen Protector",
  "Order ID": "CA-2016-157749",
  "Segment": "Consumer"
}
```

Amazon.com 에서는



데이터 분석 파이프라인

AWS 의 데이터 관련 서비스

Business Intelligence & Machine Learning


QuickSight


SageMaker


Comprehend


Rekognition


Lex


Transcribe


DeepLens

AWS Marketplace

250+ solutions

Databases


QLDB
Ledger Database


Neptune
Graph


ElastiCache
Redis, Memcached


DynamoDB
Key value, Document


Aurora
MySQL, PostgreSQL


Timestream
Time Series


RDS
MySQL, PostgreSQL, MariaDB, Oracle, SQL Server


RDS on VMWare

Analytics


Redshift
Data warehousing


Athena
Interactive analytics


EMR
Hadoop + Spark


Kinesis Analytics
Real-time


Elasticsearch service
Operational Analytics

Blockchain


Managed Blockchain


Blockchain Templates

730+ Database solutions

600+ Analytics solutions

25+ Blockchain solutions

Data Lake


S3/Glacier


Lake Formation
Data Lakes


Glue
ETL & Data Catalog

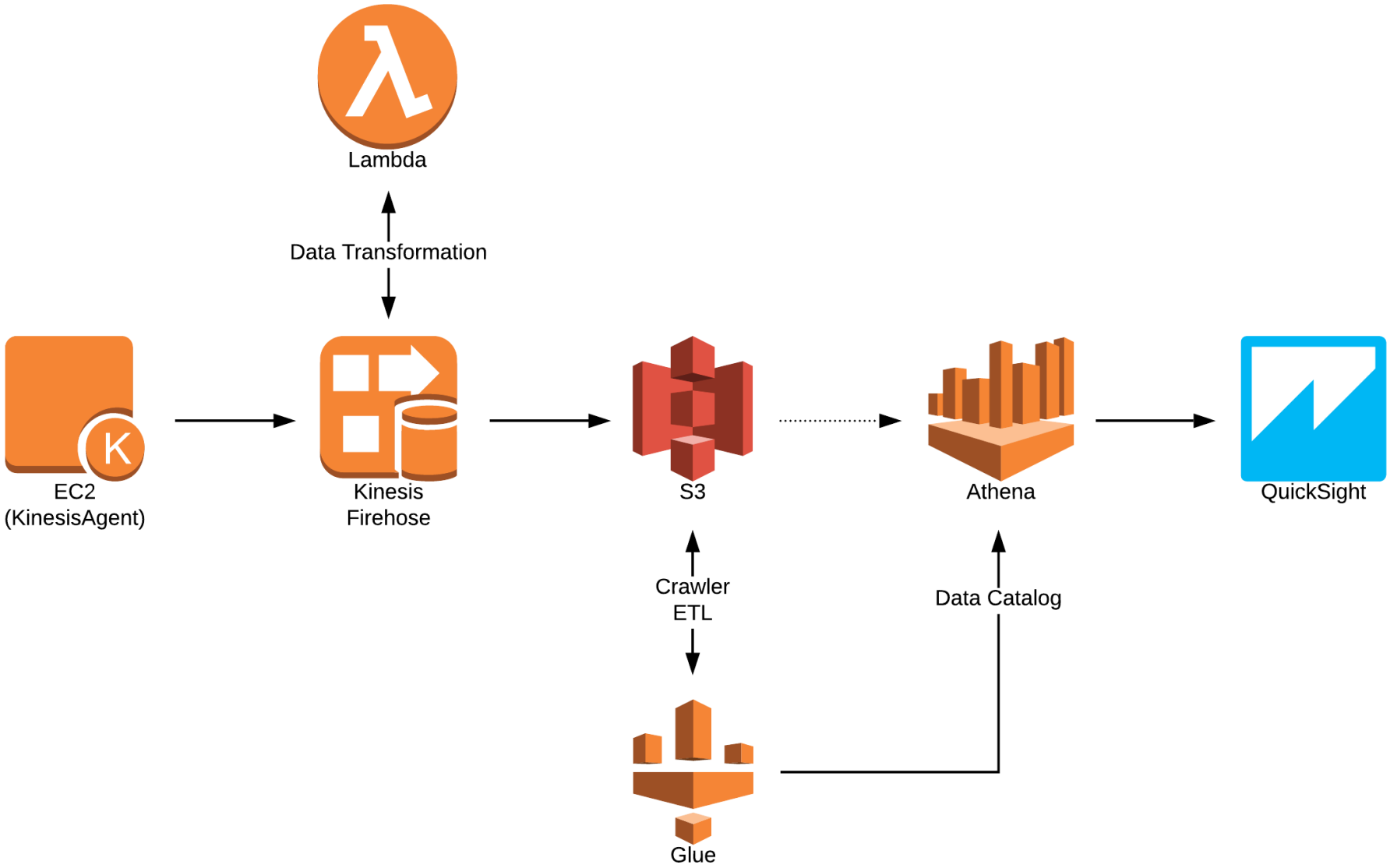
20+ Data lake solutions

Data Movement

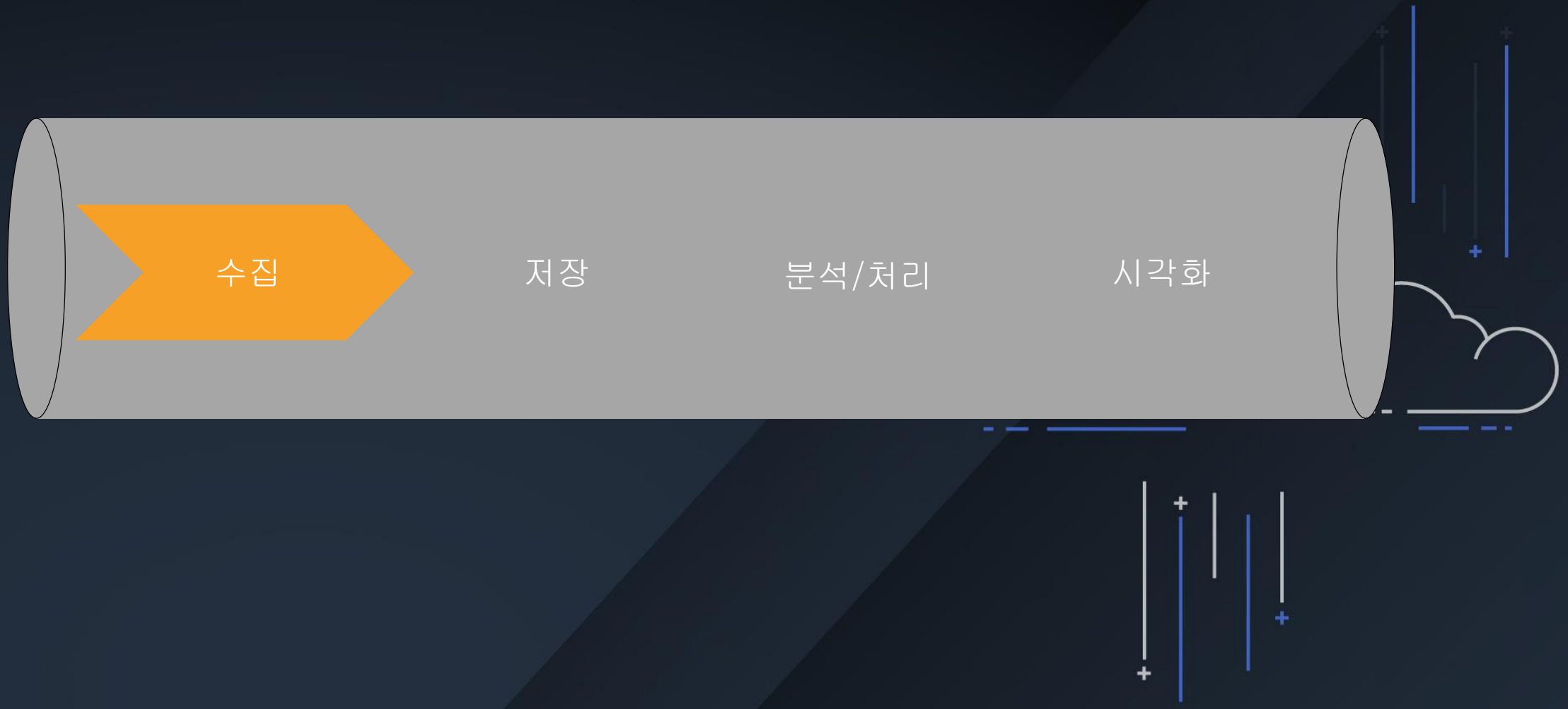
Database Migration Service | Snowball | Snowmobile | Kinesis Data Firehose | Kinesis Data Streams | Data Pipeline | Direct Connect

30+ solutions

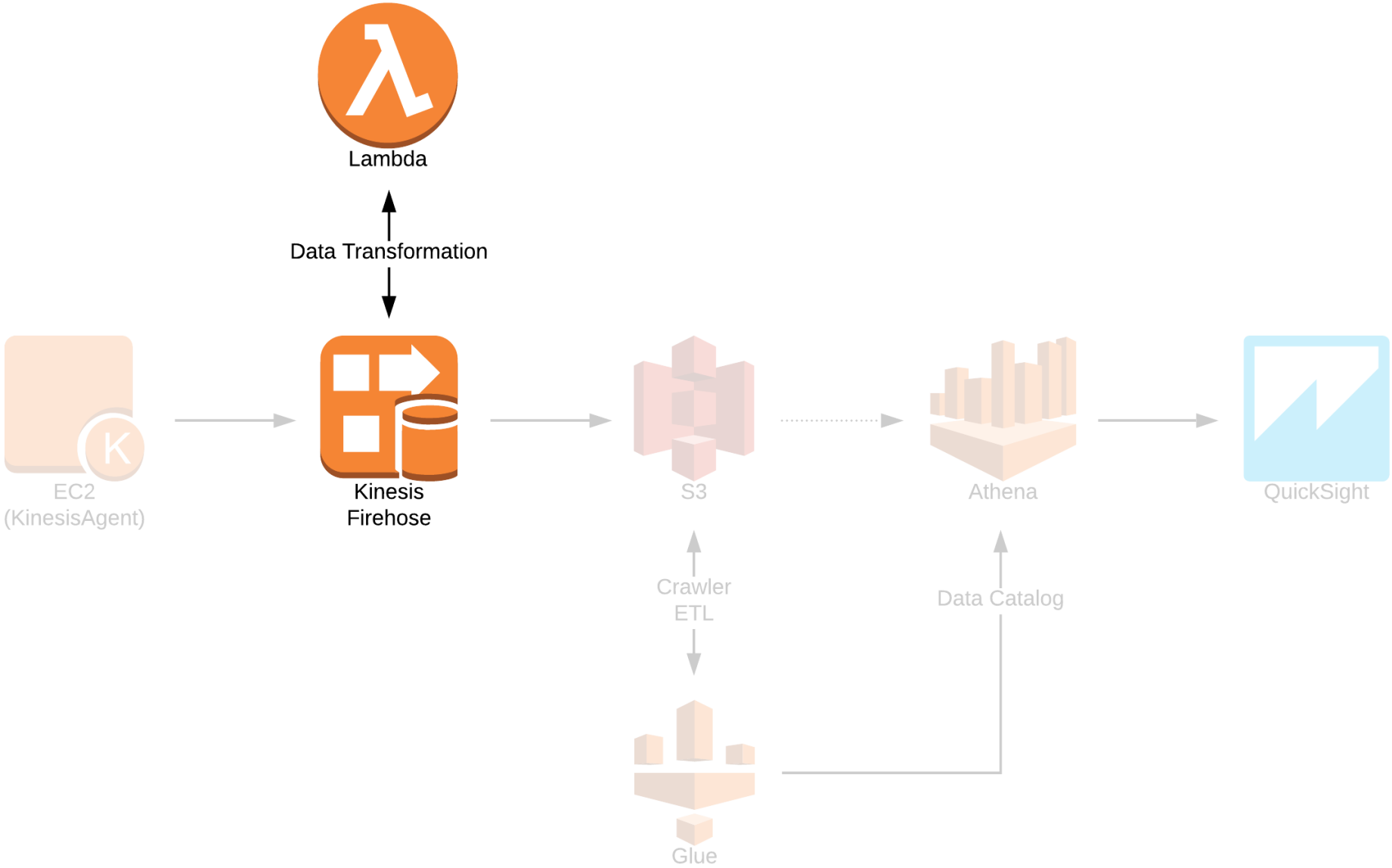
Demo 아키텍처



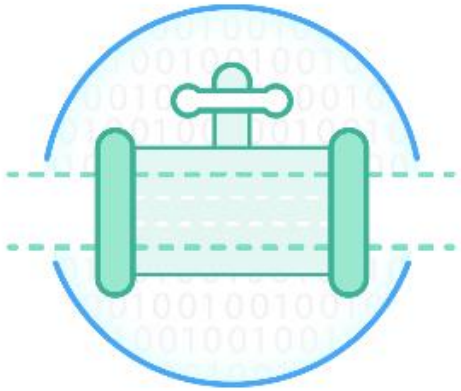
AWS Cloud Day in Busan



Demo 아키텍처 (수집)



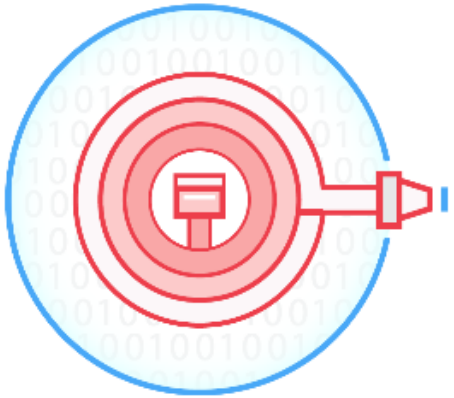
Amazon Kinesis



Kinesis Data Streams

개발자

정렬, 리플레이, 실시간 처리를 위한 수집 및 데이터 스트리밍



Kinesis Data Firehose

개발자, 데이터 사이언티스트

방대한 스트리밍 데이터를 Amazon S3, Amazon Redshift, Amazon ES, Splunk 등으로 쉽게 실시간 로드

Kinesis Data Analytics

개발자, 데이터 사이언티스트

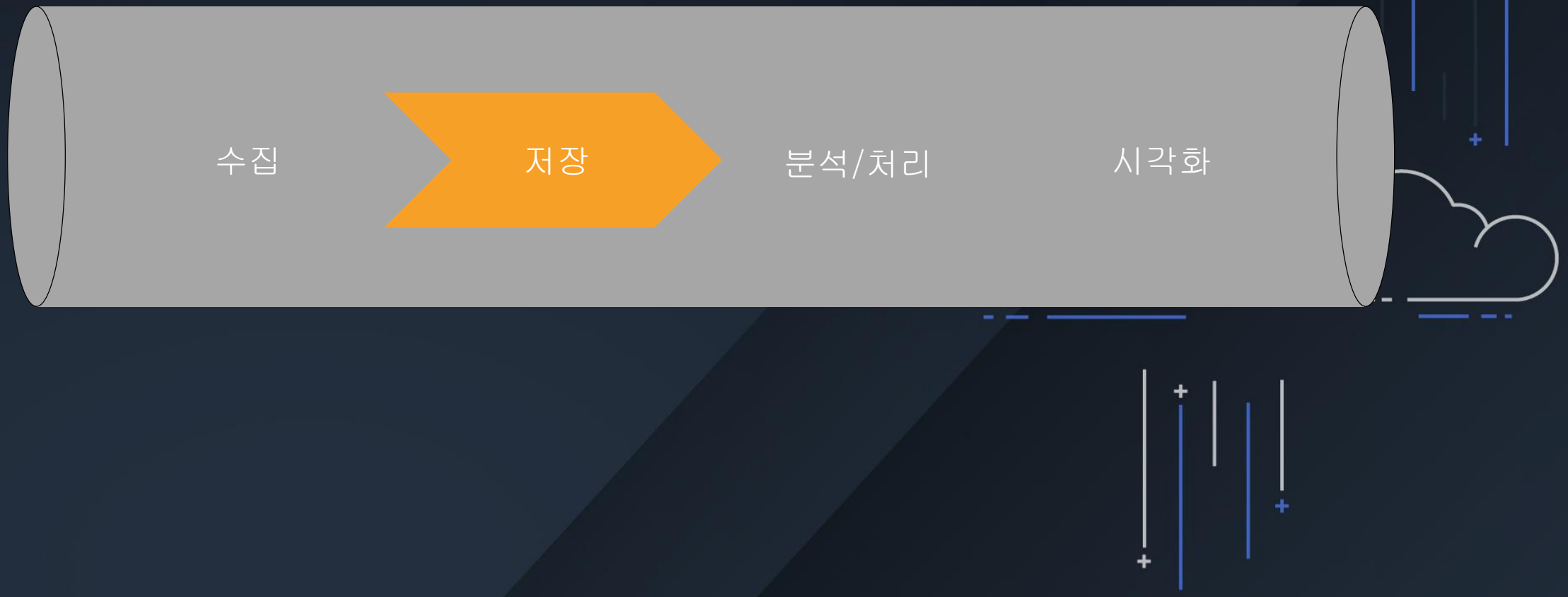
실시간으로 표준 SQL 쿼리를 이용하여 데이터 스트림 분석

Amazon Kinesis Data Firehose

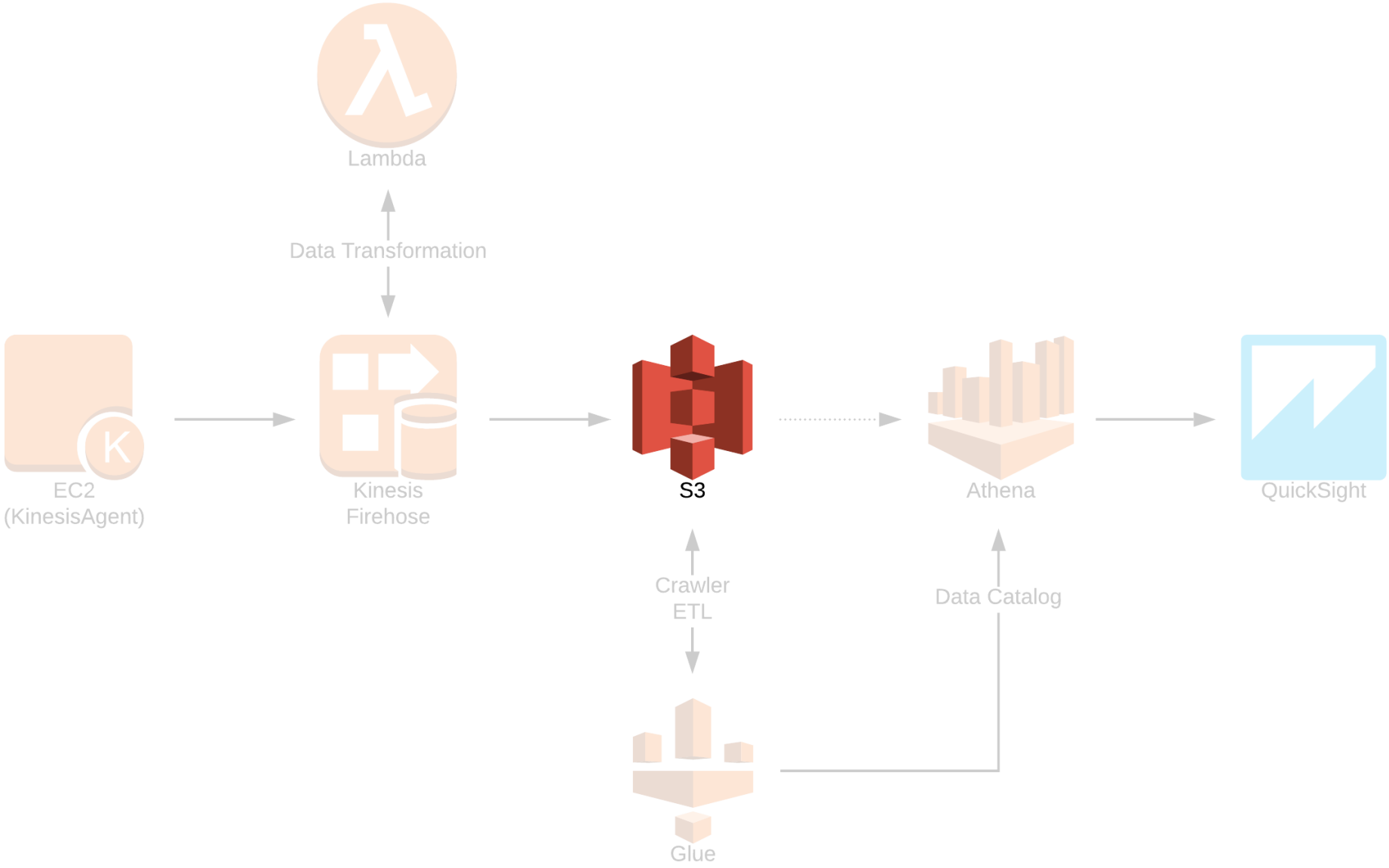
수집

- 완전 관리형 서비스
- **서버리스**
- 스트리밍 데이터를 손쉽게 캡처
- S3, Redshift, ES 등으로 데이터 로드
- Lambda를 이용한 데이터 전처리
- 데이터 **처리량에 대응**하여 자동 확장

AWS Cloud Day in Busan



Demo 아키텍처 (저장)



Amazon S3

저장

- 확장성이 뛰어난 오브젝트 스토리지
- 객체 당 1Byte에서 5TB 크기 지원
- 저장할 수 있는 **객체 수 제한 없음**
- 99.999999999% 의 **내구성** 제공
- 서버 측 암호화(SSE) 제공

Amazon S3 for Data Lake Tier 설계

Tier-1 데이터 레이크 : 수집과 저장

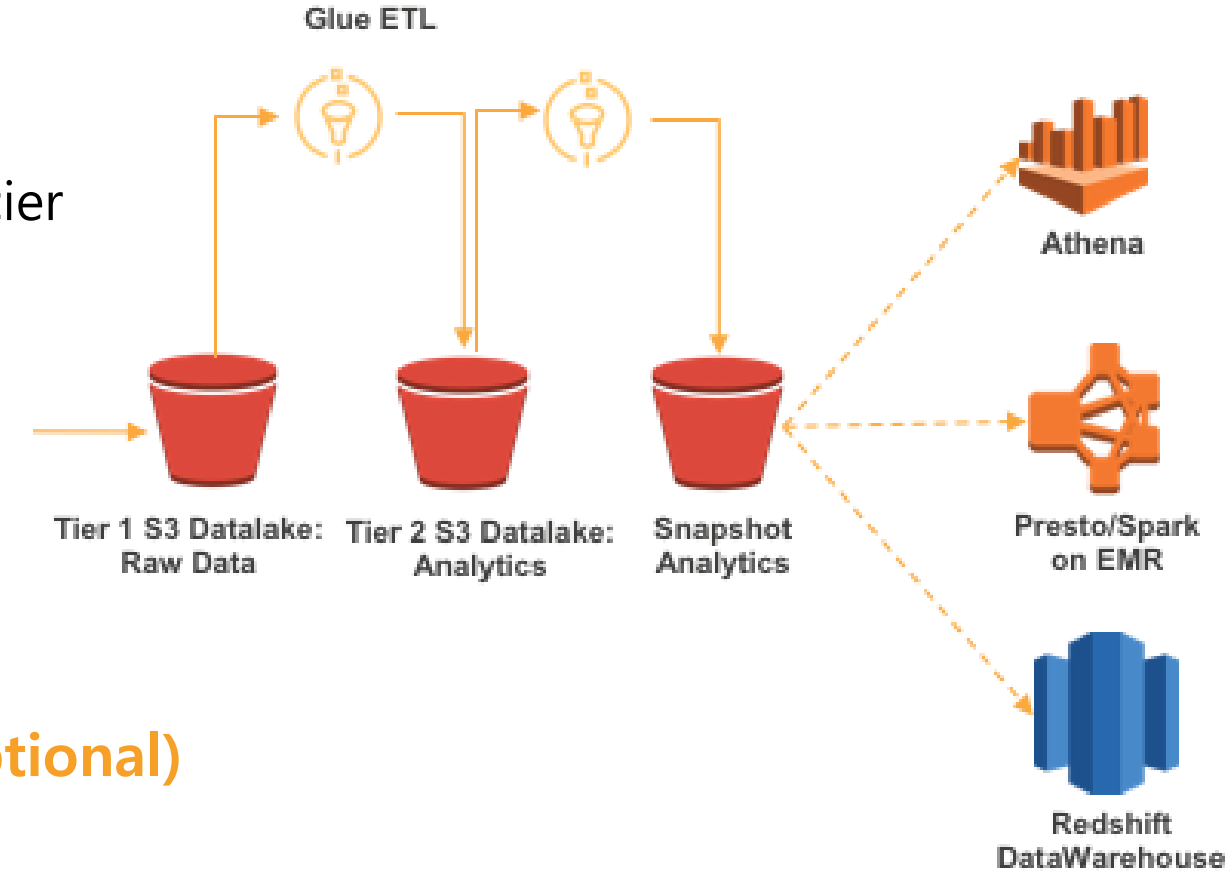
- 원본 데이터의 저장과 보장
- 최소한의 데이터 변환 작업만
- S3의 라이프사이클 기능 활용, S3-IA 또는 Glacier

Tier-2 데이터 레이크 : 분석용 데이터

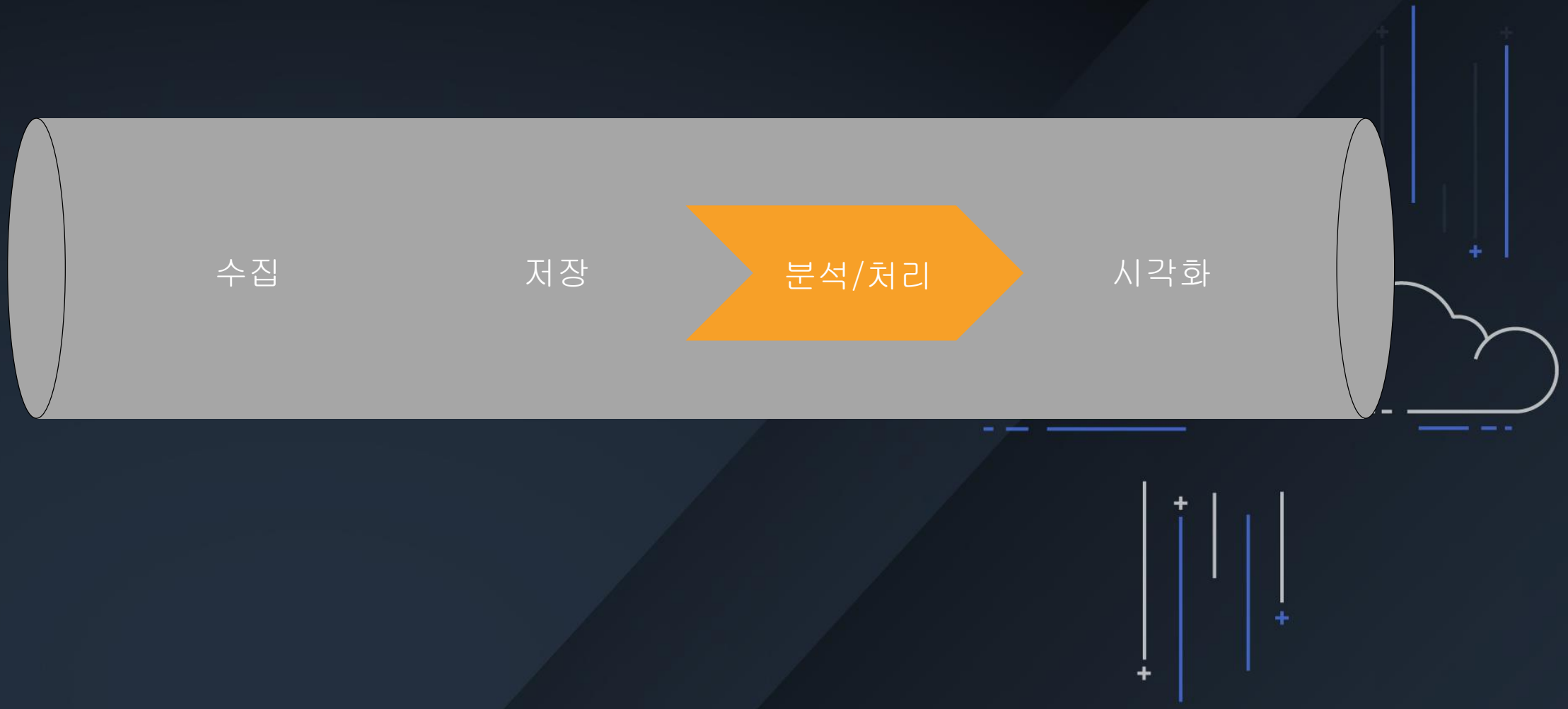
- Parquet / ORC 같은 컬럼방식 포멧의 사용
- 파티션 정책에 따라 분산
- 분석을 위한 최적화

Tier-3 데이터 레이크 : 특정한 분석 목적 (optional)

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



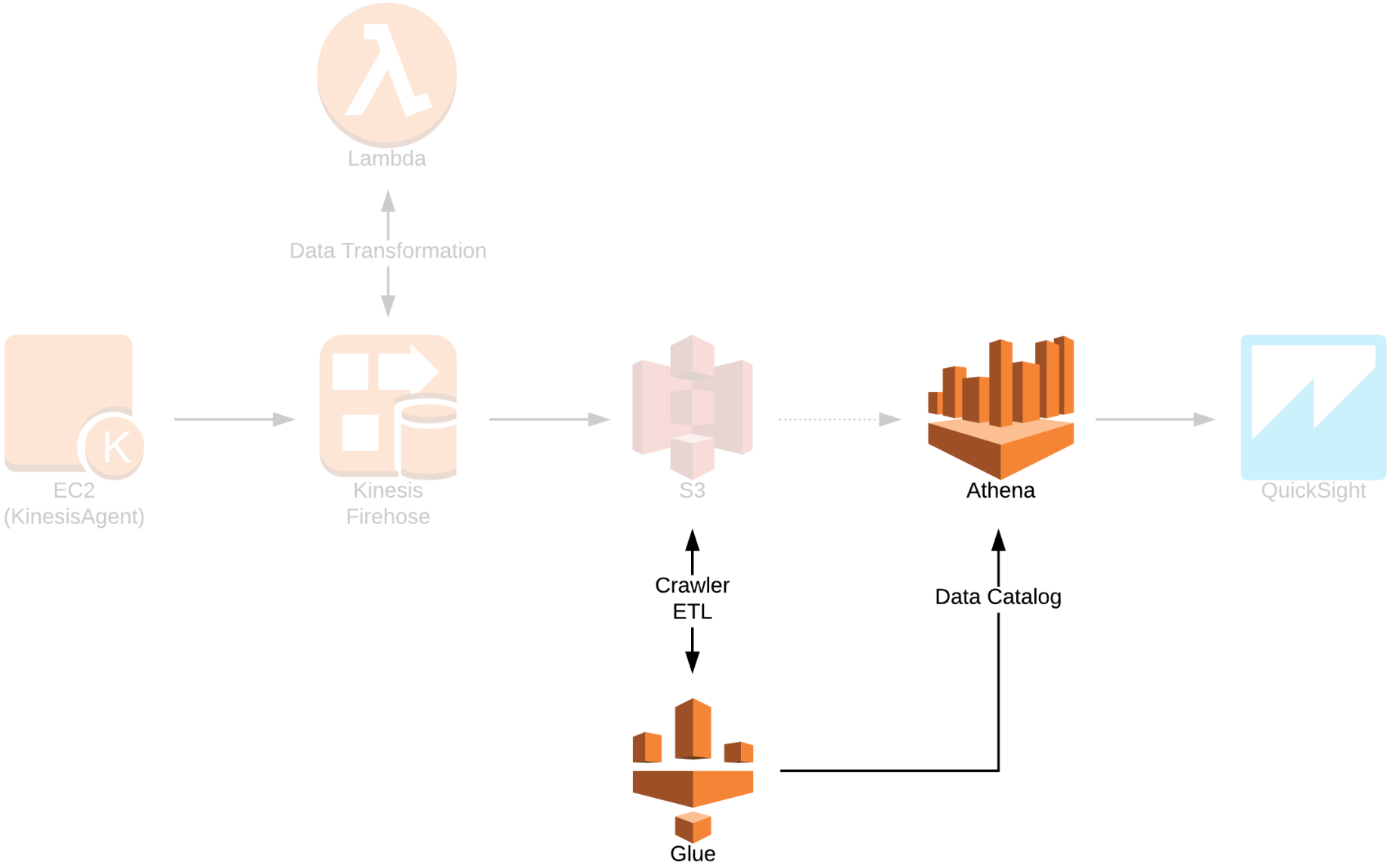
AWS Cloud Day in Busan



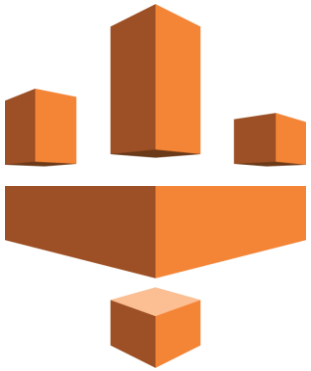
작게 시작해서 반복하기

Start Small and Iterate

Demo 아키텍처 (분석)



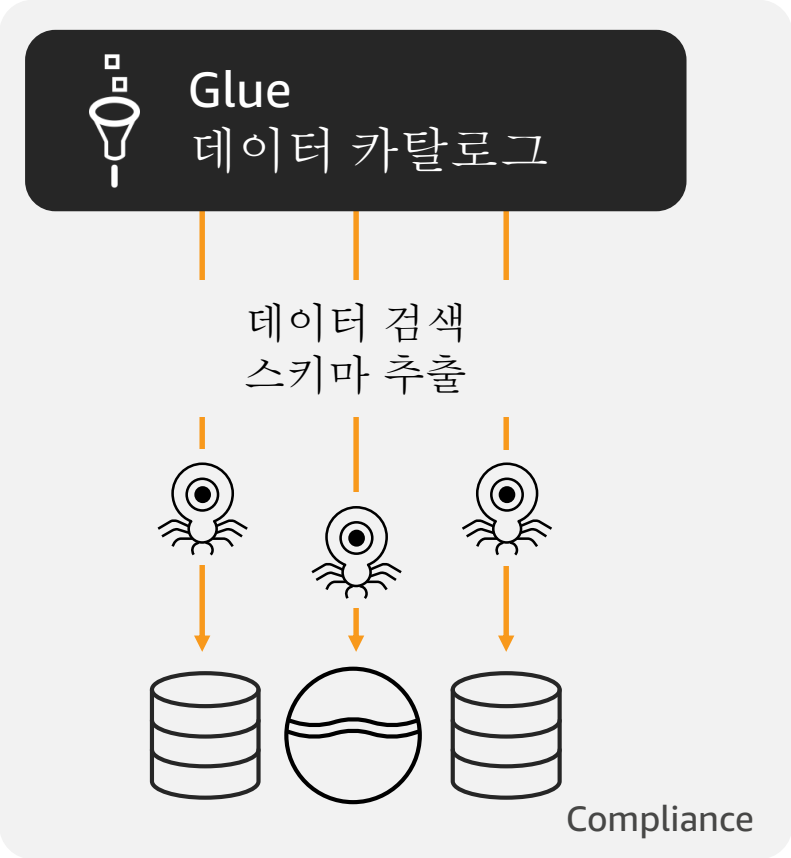
AWS Glue



분석 / 처리

- 완전 관리형 ETL 서비스
- **서버리스**
- AWS에 저장된 데이터를 자동 검색하고 분류하여 빠르게 분석 가능
- 메타데이터(테이블 정의, 스키마 등)를 **Glue 데이터 카탈로그**에 저장
- ETL 코드를 추천 및 생성

Glue 데이터 카탈로그



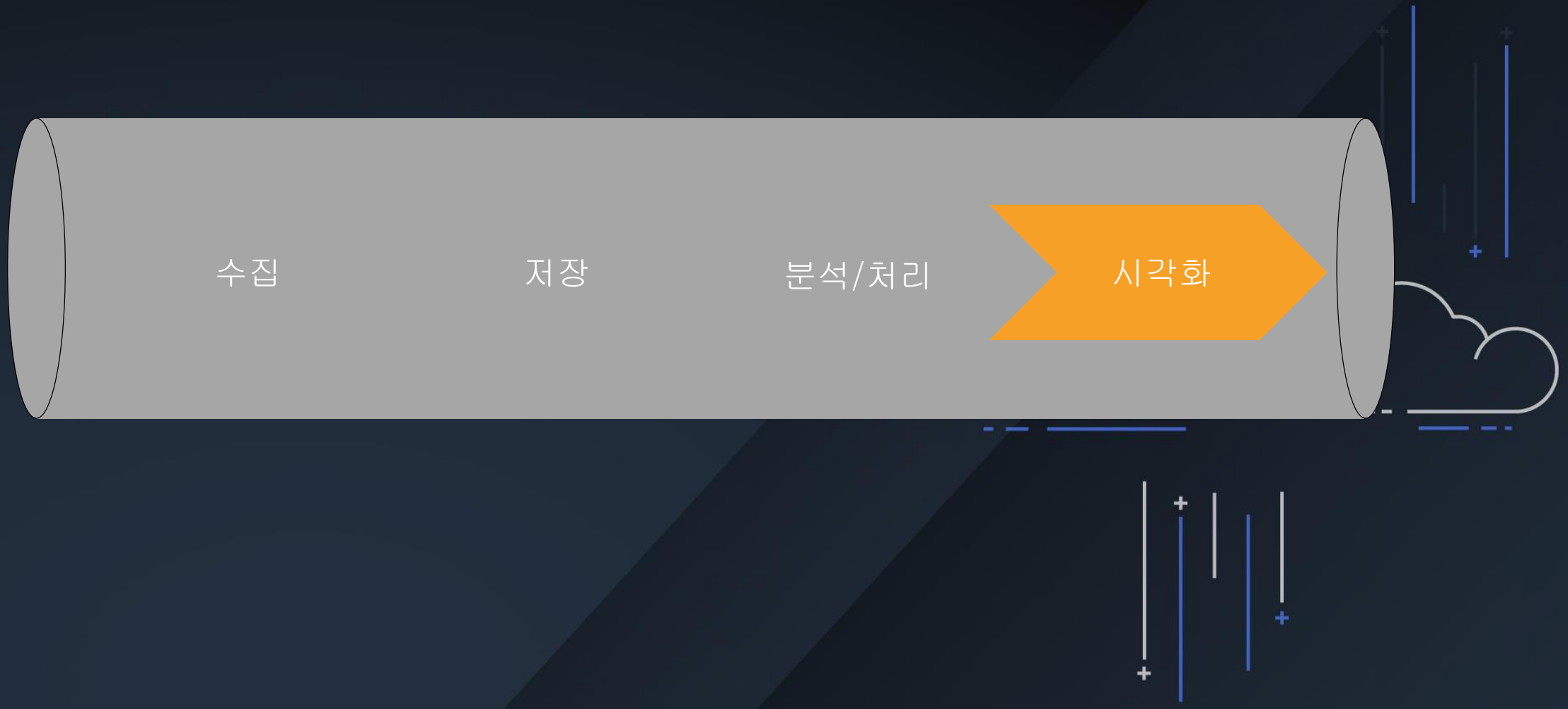
- **Glue 크롤러**를 통해 자동으로 데이터를 검색하고 스키마를 카탈로그에 저장
- 카탈로그를 통해 Athena, EMR, Redshift Spectrum 에서 **즉시 쿼리**
- 카탈로그는 **ETL 에 사용** 가능

Amazon Athena

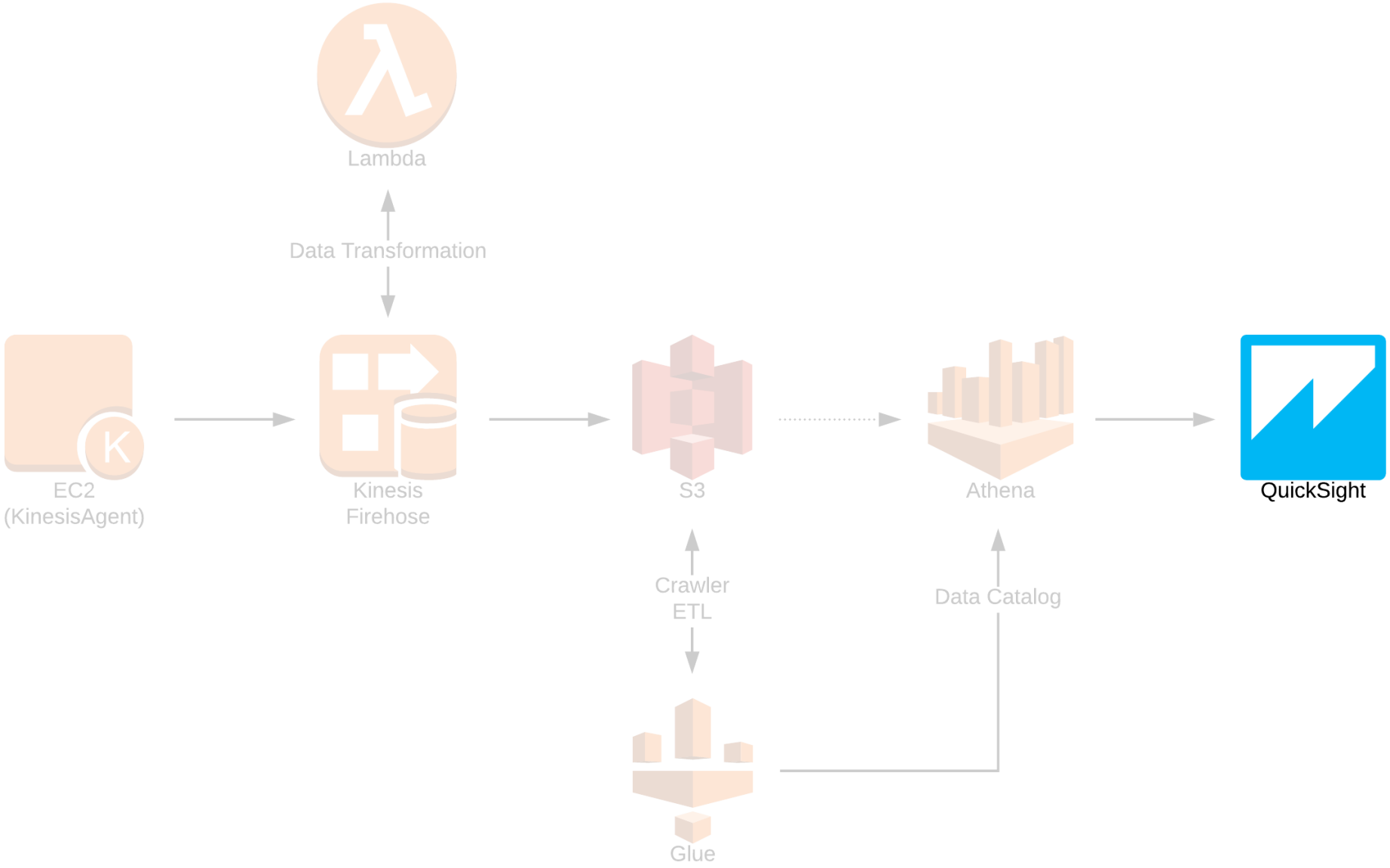
분석 / 처리

- 대화식 쿼리 서비스
- **서버리스**
- **표준 (ANSI) SQL** 을 이용해 S3에 저장된 데이터를 분석
- 실행한 쿼리에 대한 비용만 지불

AWS Cloud Day in Busan



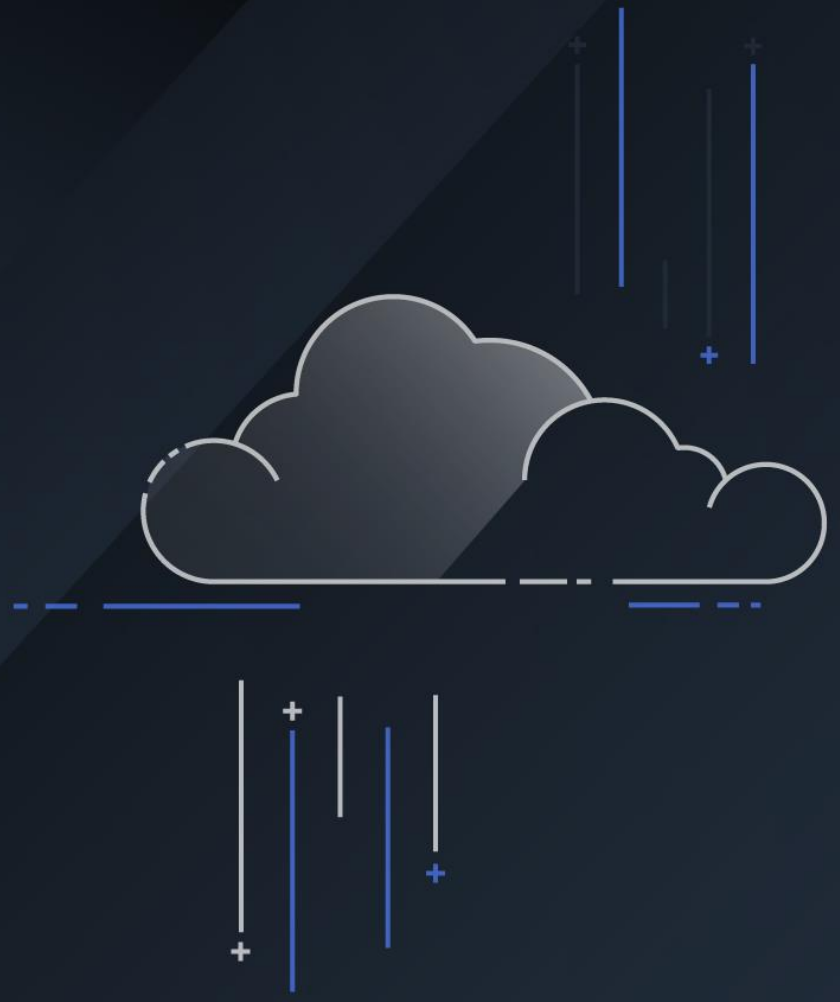
Demo 아키텍처 (시각화)



Amazon QuickSight

- 클라우드 기반 관리형 BI 서비스
- **서버리스**
- 다양한 데이터에 **쉽게 연결**하고 **빠르게 시각화** 가능
- 스토리 보드를 통한 공유 및 협업
- 모든 브라우저 및 다양한 모바일 플랫폼 지원

결론



실시간 스트리밍 데이터 분석 데모 : 결론

✓빠르게 구축

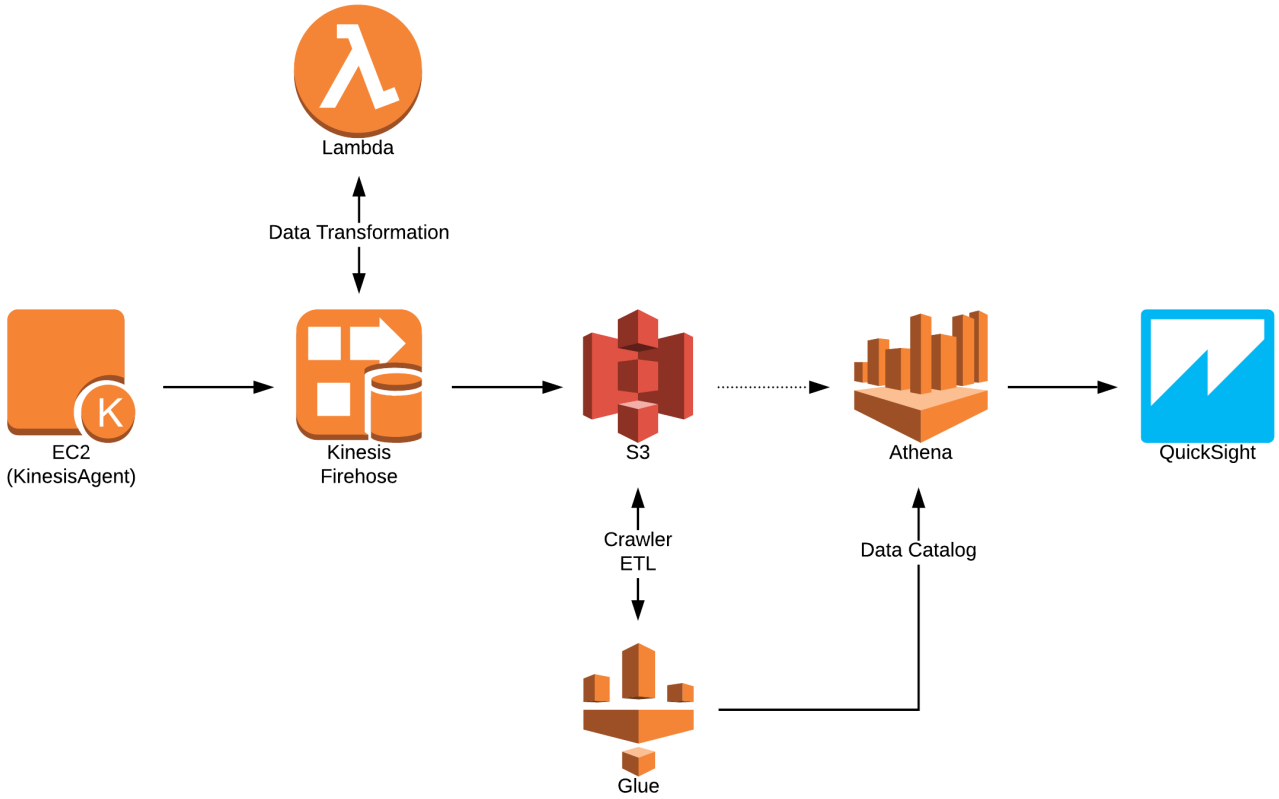
✓실시간 데이터 분석

✓서버 관리 불필요

✓유연하게 확장

✓무제한 데이터 저장

✓유휴 용량 없음



데모 실습 자료

http://bit.ly/aws-data-analysis

AWS Cloud Day in Busan

감사합니다.