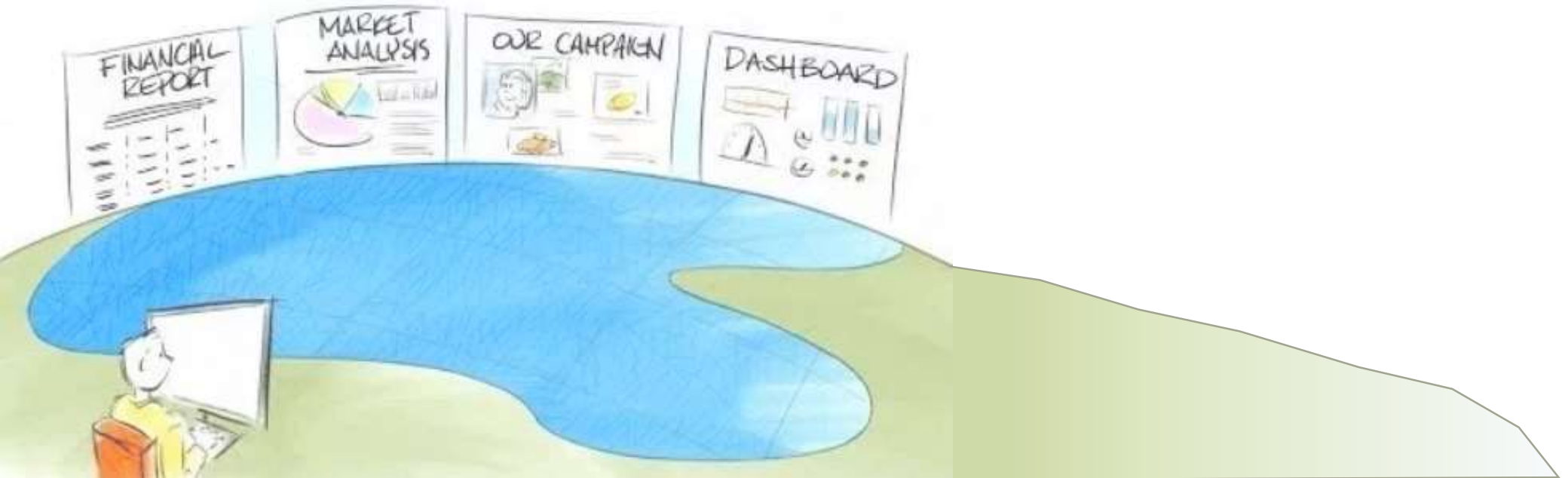
실제 프로젝트를 통해 본 DATA LAKE 구축 방안 및 사례

Cloud Innovator MEGAZONE

aws



이윤미

Enterprise Service Center

Data analysis

- D 제조사 Hadoop Platform Migration POC
- J 교육업체 BIGDATA Platform 구축
- C 온라인몰 BIGDATA Platform 구축
- A 제조사 DataLake 구축 PoC
- B 금융사 Elastic Cloud System 구축
- C 금융사 DataLake, 실시간 데이터 처리 시스템 구축

목차

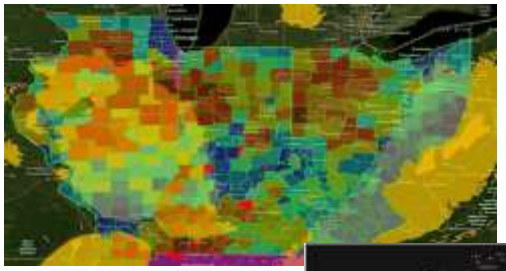
1. 왜 데이터레이크가 필요한가?
2. 클라우드 데이터레이크 구성
3. 실제 구축 방안 및 사례

1. 왜 데이터레이크가 필요한가?

데이터 현황

디지털 산업화로 인하여 새로운 데이터가 비약적으로 늘어나고 있음

Geospatial

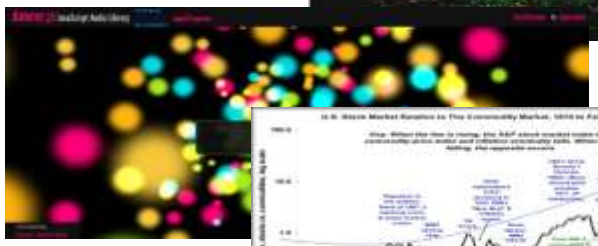


Social Media

Mobile

Telematics

Audio



Stock Market

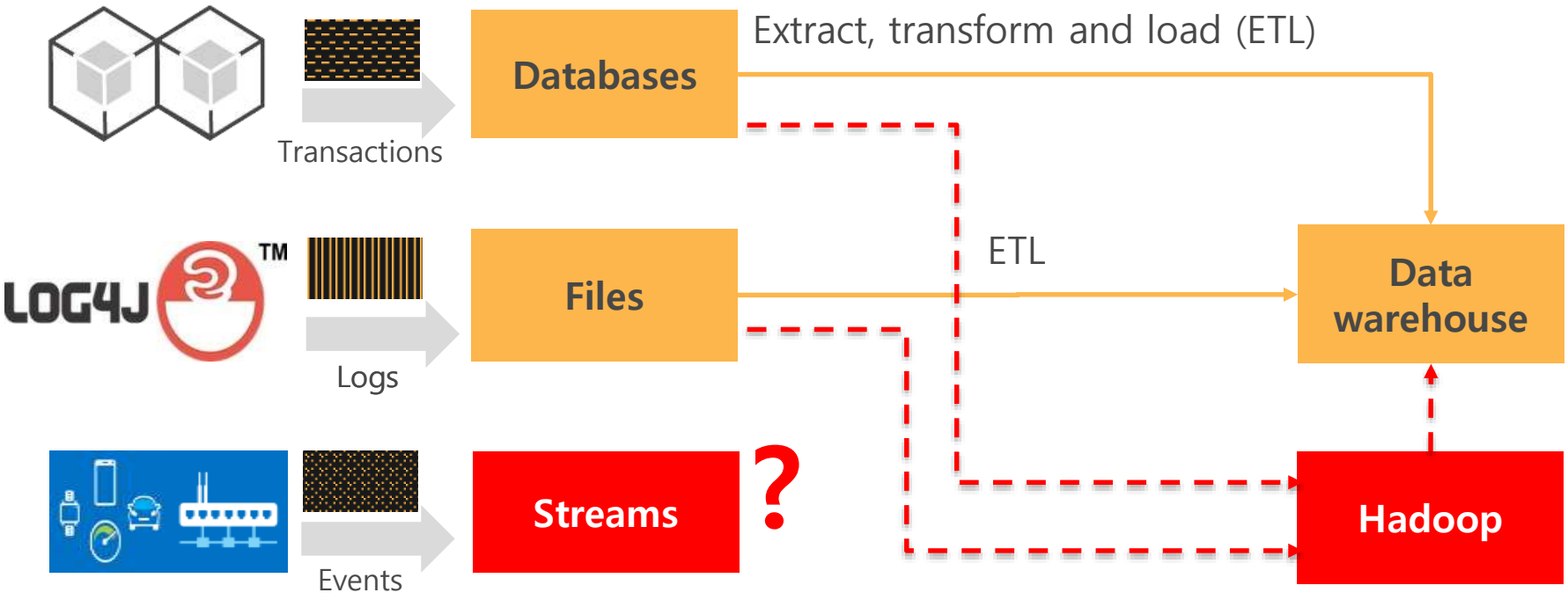
Sensor

Video

데이터 이슈 사항

- 데이터가 너무 방대해 지고 있음.
- 데이터의 유형이 다양함
- 내부 데이터의 통합적이며 다양한 결합 분석이 필요함
- 분석 대상 데이터가 여러 곳에 흩어져 있음
- 다양한 분석 서비스 플랫폼이 필요함.
- 운영/기간계의 부하발생의 문제

데이터 분석 아키텍처

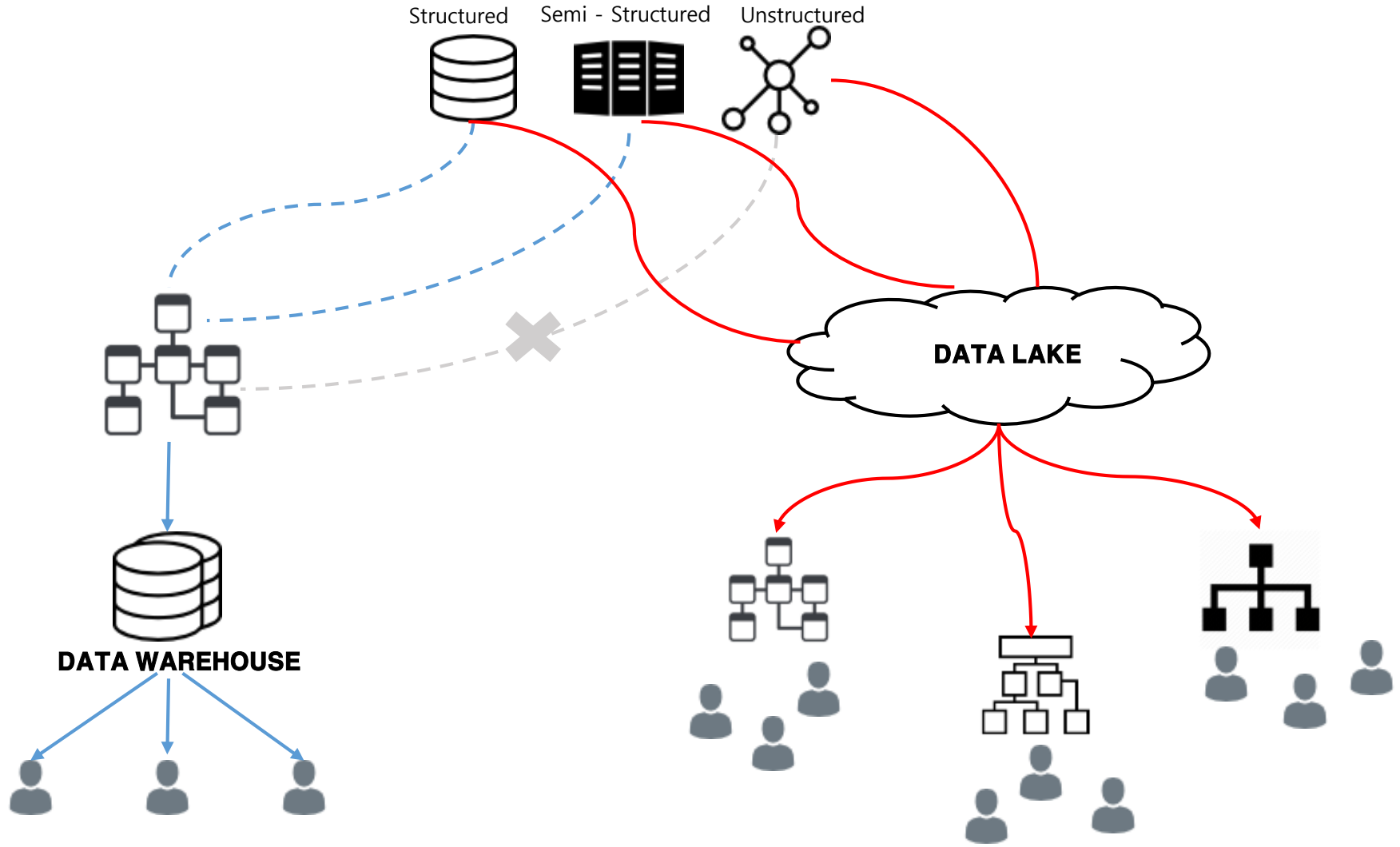

Transactions
Databases
Extract, transform and load (ETL)
LOG4J
Logs
Files
ETL
Data warehouse
Events
Streams
?
Hadoop

데이터레이크 정의

막대한 양의 데이터를 본래의 형태 그대로, 서비스 되기 직전까지 보관하는 저장소.

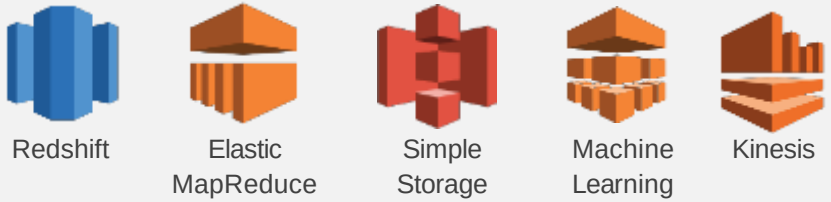
가공이 전혀 되지 않은,순수한 ➡ ***전 처리가 이뤄지지 않은***

DW vs DATALAKE



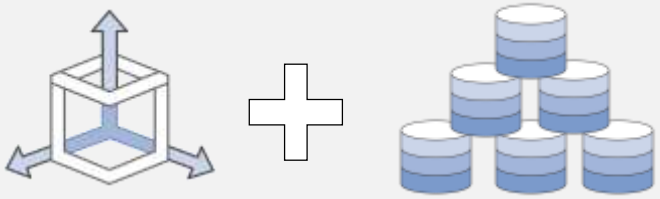
클라우드 데이터레이크의 특징

다양한 데이터 처리 서비스 제공



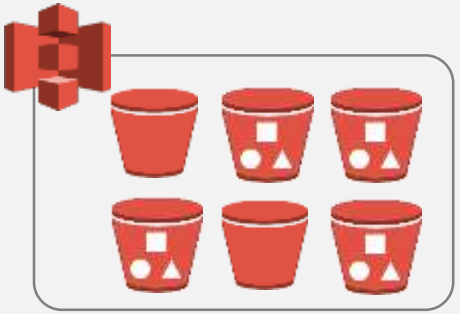
Data Ingest, Store, Analytic, Visual 을 위한 다양한 서비스를 제공하여 요구사항에 따른 아키텍처 구성 지원

서버와 저장소의 분리 구성



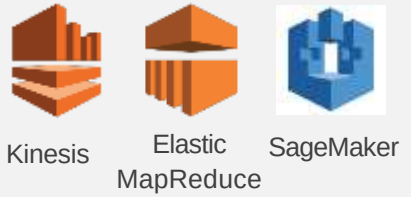
데이터 저장소와 연산 부분을 분리하여 저장된 데이터를 기반으로 다양한 서비스를 연동하여 분석 제공

저장소의 무한 확대



AWS Simple Storage는 Structure, Semi-Structure, Unstructured 등의 다양한 형태의 데이터를 무제한으로 저장을 제공

서비스를 통한 구축의 최소화



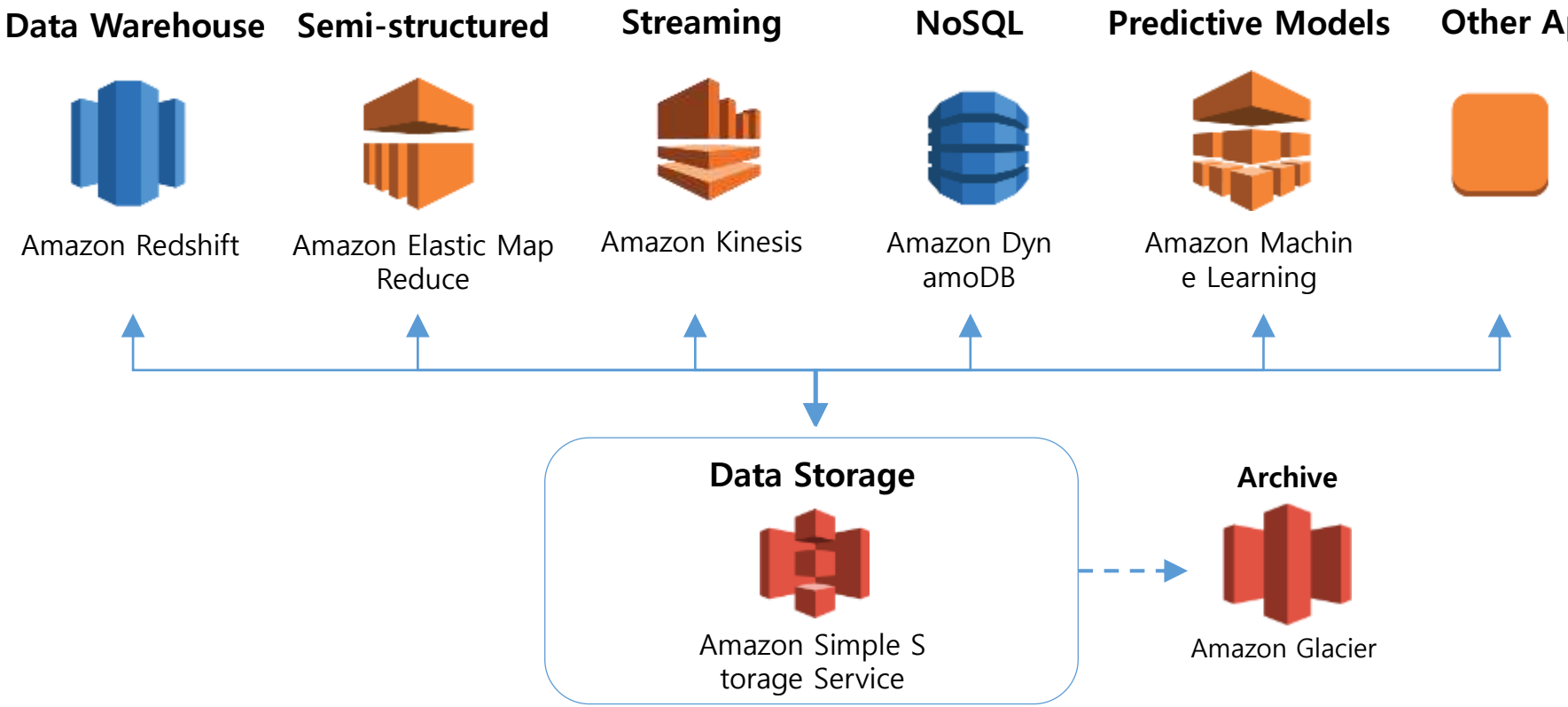
AWS Managed 서비스를 활용하여 별도의 Application 설치없이 바로 사용할 수 있기 때문에 짧은 시간내의 구축 및 효율적인 운영 관리 지원

2. 클라우드 데이터레이크 구성

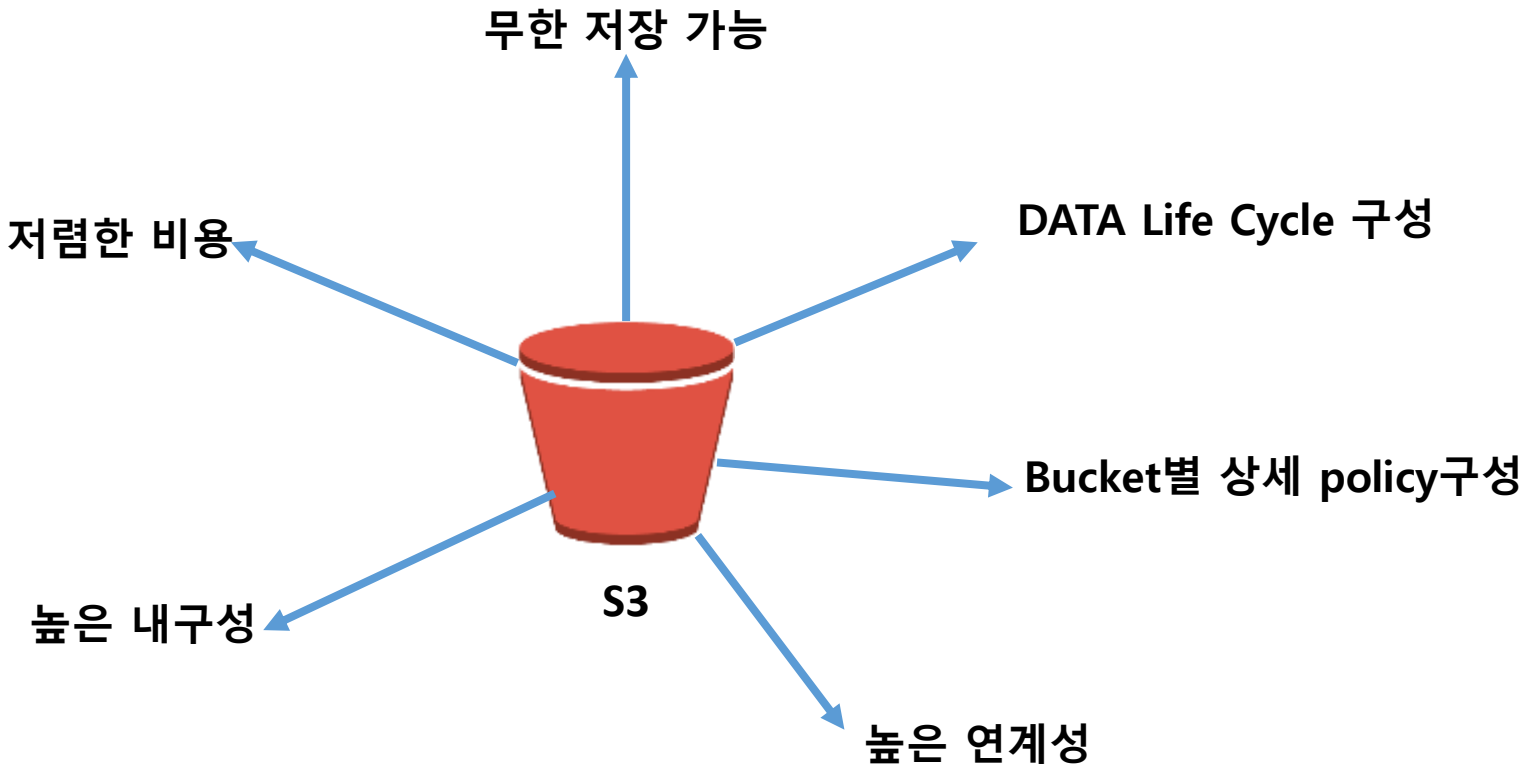
데이터 처리 서비스

서비스	제품유형	설명
Amazon Athena	서버리스 쿼리 서비스	표준 SQL을 사용하여 손쉽게 Amazon S3 데이터를 분석. 실행한 쿼리에 대해서만 비용 지불
Amazon EMR	빅데이터 하둡 클러스터 서비스	하둡 프레임워크. Spark, hive, Hbase등 널리 사용되는 분산 프레임워크를 빠르게 실행/구성해줌. S3 및 DynamoDB등 다른 AWS 데이터스토어와의 상호작용이 간편함
Amazon ElasticSearch	Elastic search 서비스	완전관리형 ElasticSearch서비스. 데이터 적재, 시각화를 용이하게 하고 간편한 운영과 편리한 확장이 가능
Amazon Kinesis	스트리밍 데이터 분석	스트리밍 데이터를 간편하게 로드 및 분석 가능 스트리밍 데이터를 원하는 다른 aws서비스로 이관 하거나 실시간 분석 가능
Amazon QuickSight	BI	클라우드 기반 비즈니스 인텔리전스 서비스로 손쉽게 시각화를 구축하고 AWS의 다양한 서비스와 연계 가능
Amazon Redshift	데이터웨어 하우스	페타바이트 규모의 비용 효율적 완전 관리형 데이터웨어하우스
AWS Glue	데이터 카탈로그 및 ETL	데이터 원본 파악(메타 구성), ETL, 검색, 변환 매핑 작업을 단순화 자동화 함
AWS Data Pipeline	데이터 워크플로 오케스트레이션	데이터를 안정적으로 처리하고 이동하도록 지원 데이터를 대규모로 변환 처리 여러 서비스와 결과 연계 가능
Amazon Sagemaker	기계학습 모델 구축, 학습, 배포 서비스	기계학습 모델을 구축하고 데이터 학습과 프로덕션 배포를 지원하는 완전 관리형 플랫폼. 기본 내부 알고리즘 제공

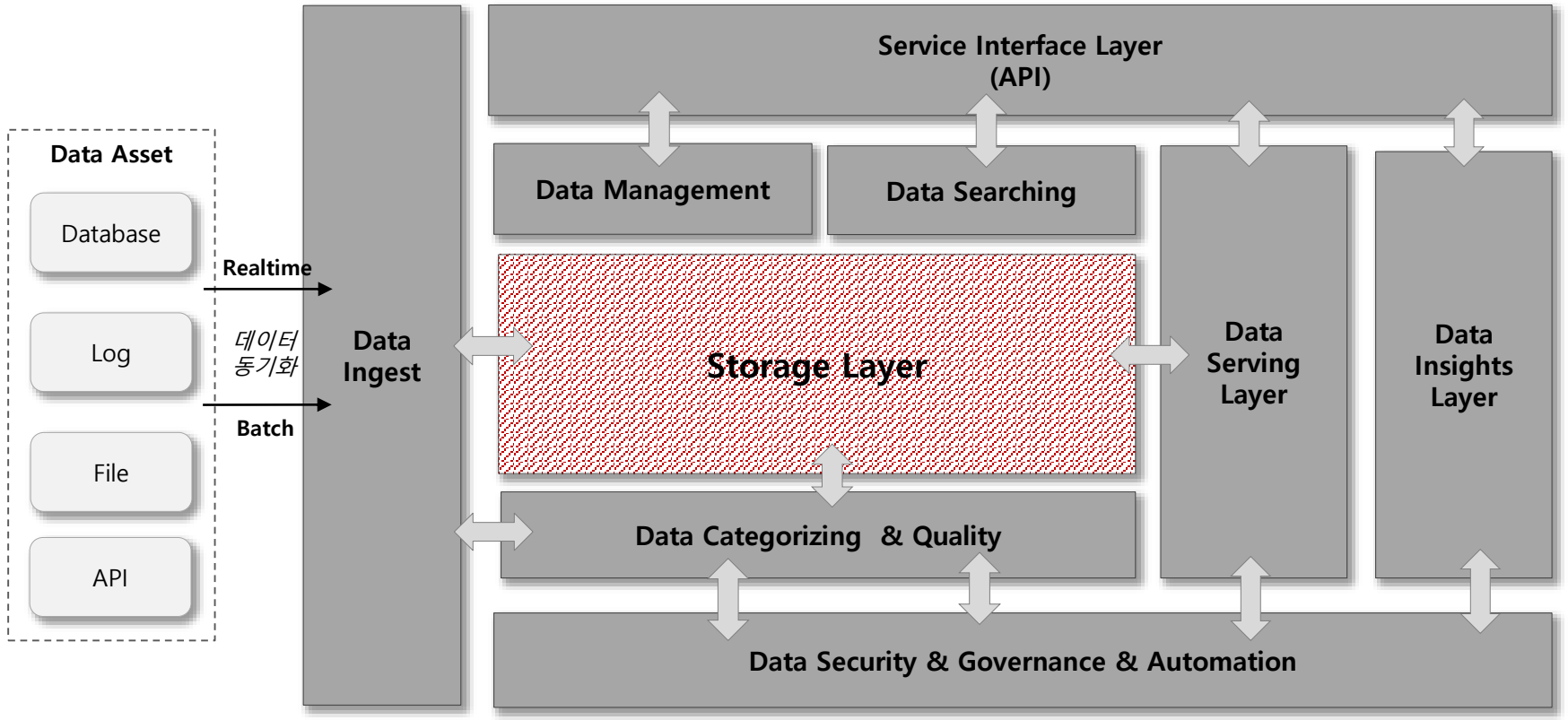
통합 저장소 구성 특징


Data Warehouse
Semi-structured
Streaming
NoSQL
Predictive Models
Other Apps
Amazon Redshift
Amazon Elastic Map Reduce
Amazon Kinesis
Amazon Dyn amoDB
Amazon Machin e Learning
Data Storage
Amazon Simple S torage Service
Archive
Amazon Glacier

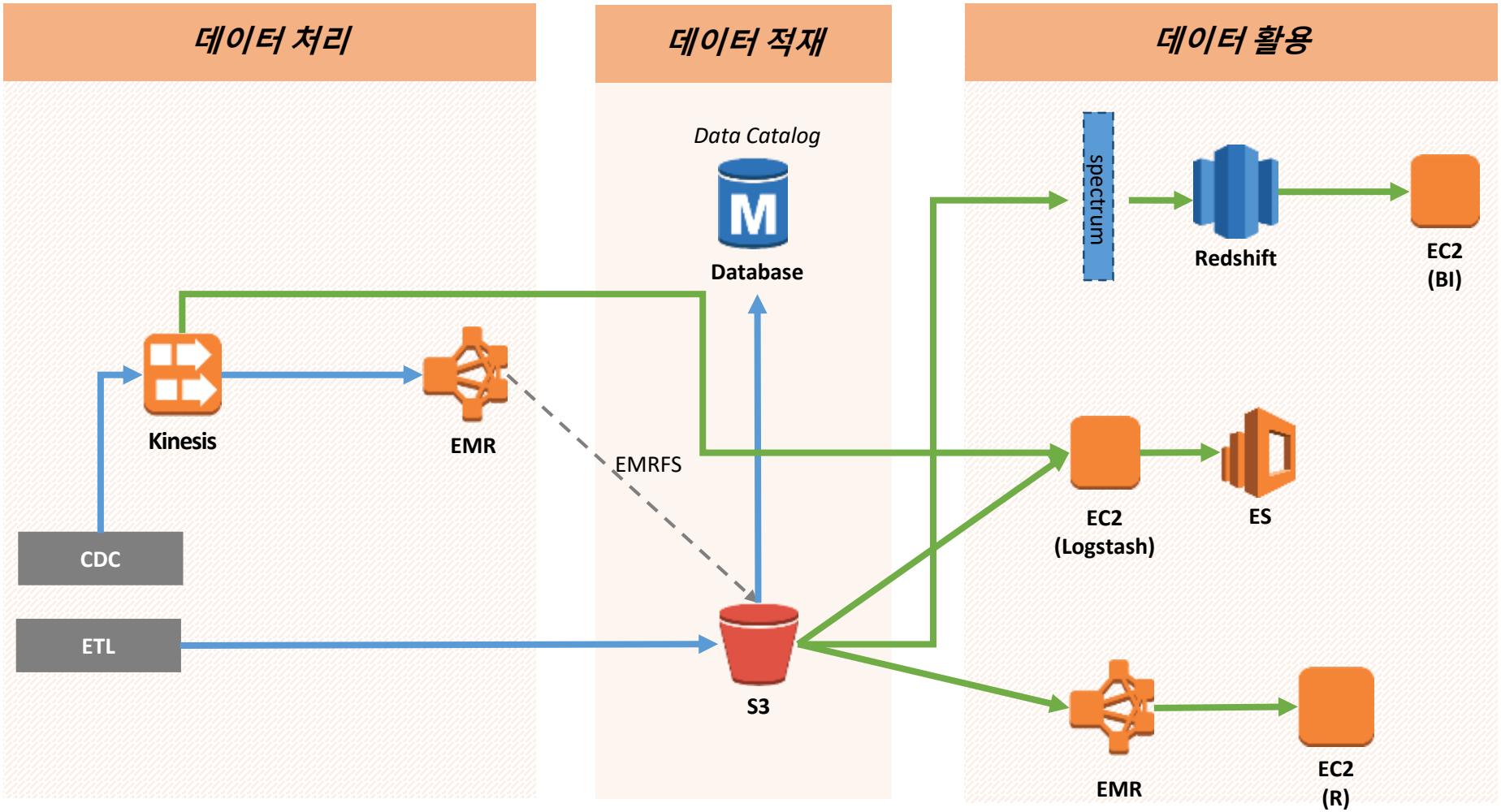
S3의 특징


무한 저장 가능
저렴한 비용
DATA Life Cycle 구성
Bucket별 상세 policy구성
높은 내구성
S3
높은 연계성

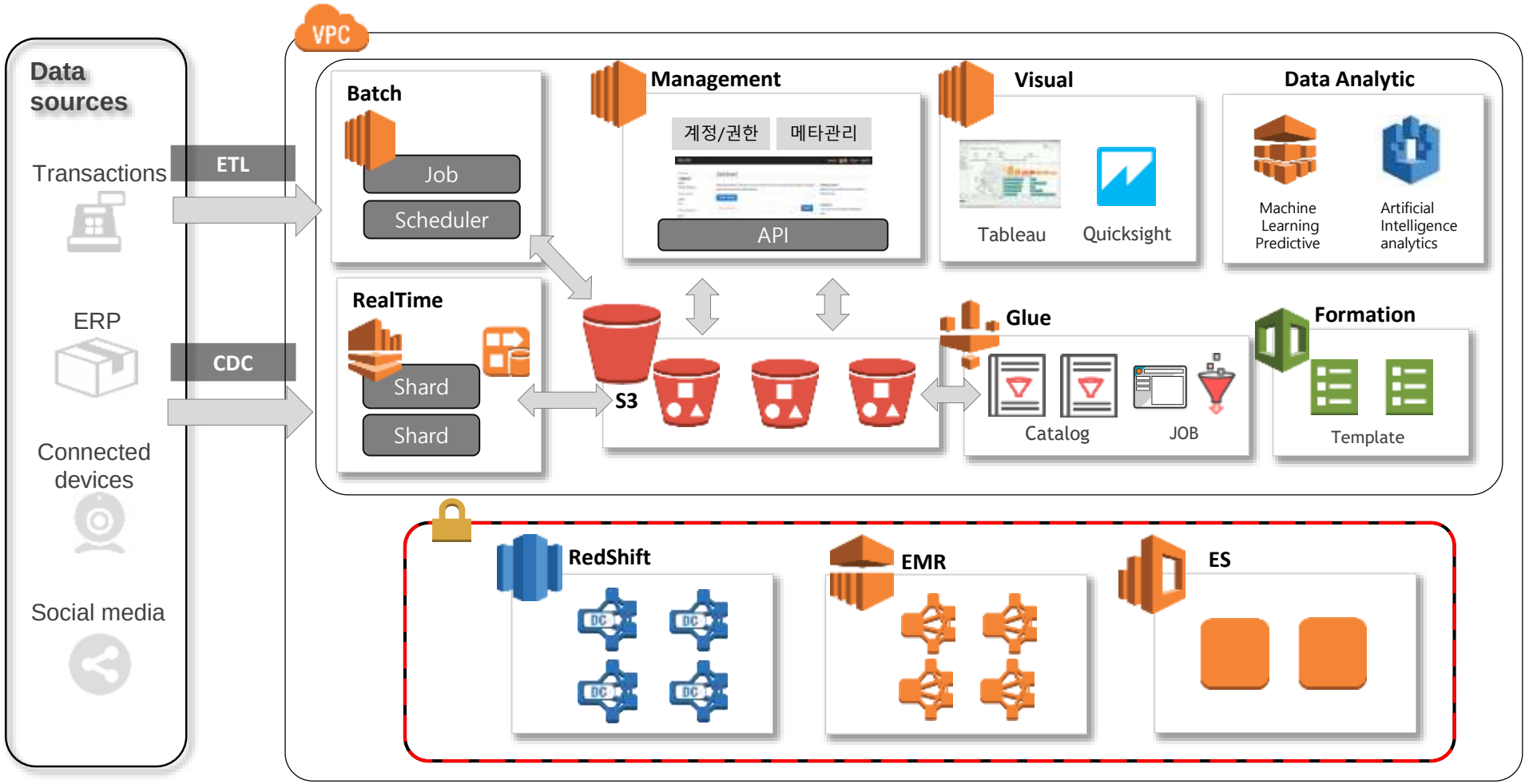
데이터레이크의 구성


Data Asset
Database
Log
File
API
Realtime
데이터 동기화
Batch
Data Ingest
Service Interface Layer (API)
Data Management
Data Searching
Storage Layer
Data Categorizing & Quality
Data Serving Layer
Data Insights Layer
Data Security & Governance & Automation

Architecture



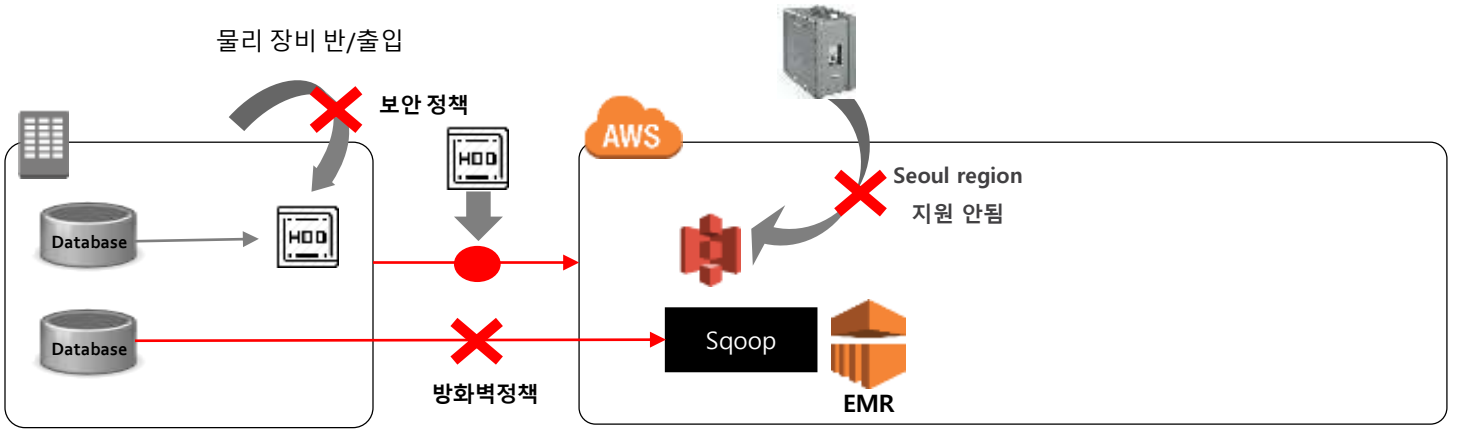
Architecture 상세



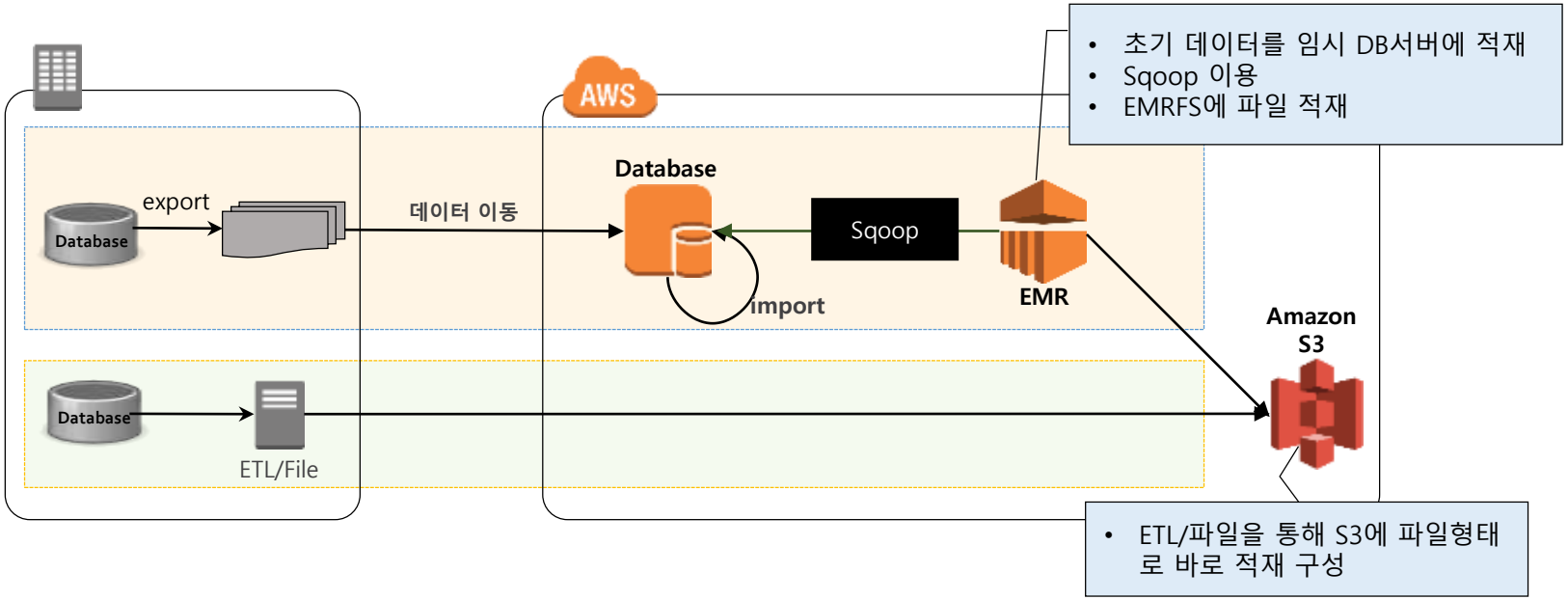
3. 실제 구축 방안 및 사례

초기 데이터 이동/적재 방안

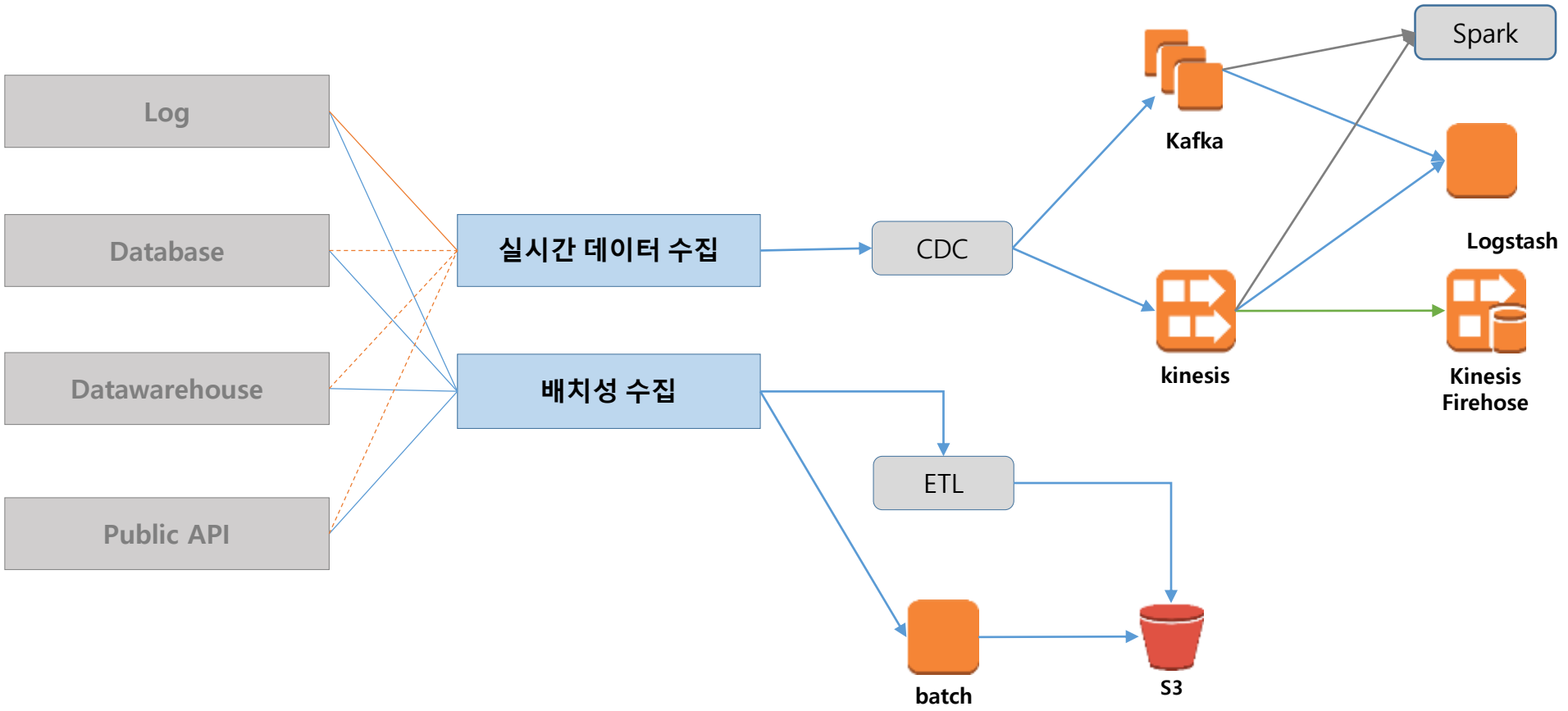
제약 사항



실제 작업



데이터 수집


Log
Database
Datawarehouse
Public API
실시간 데이터 수집
배치성 수집
CDC
ETL
Kafka
kinesis
batch
S3
Spark
Logstash
Kinesis Firehose

데이터 처리 (EMR)

❖ Amazon EMR

- 관리형 하둡 프레임웍크로 동적으로 확장 가능하며
- 분산 프레임워크를 실행하고 s3와 dynamo DB와 같은 aws 데이터 스토어와의 상호 작용 가능한 클러스터
- 로그분석, 데이터 변환, 기계학습, 과학적 시뮬레이션 등 광범위한 빅데이터를 안정적으로 처리

❖ 클러스터 주요 특징

- 컴퓨팅 자원과 스토리지의 분리 구성 가능(HDFS를 S3로 활용)
- 노드 자원 변경과 셧 다운 시 데이터 유실 없음
- 여러 클러스터가 동일 자원을 공유 가능
- 간편하게 NODE구성의 변경이 가능하여 다양한 워크로드에 쉽게 적용 가능
- 손쉬운 빅데이터 클러스터 구성 가능

❖ 클러스터 구성 자동화

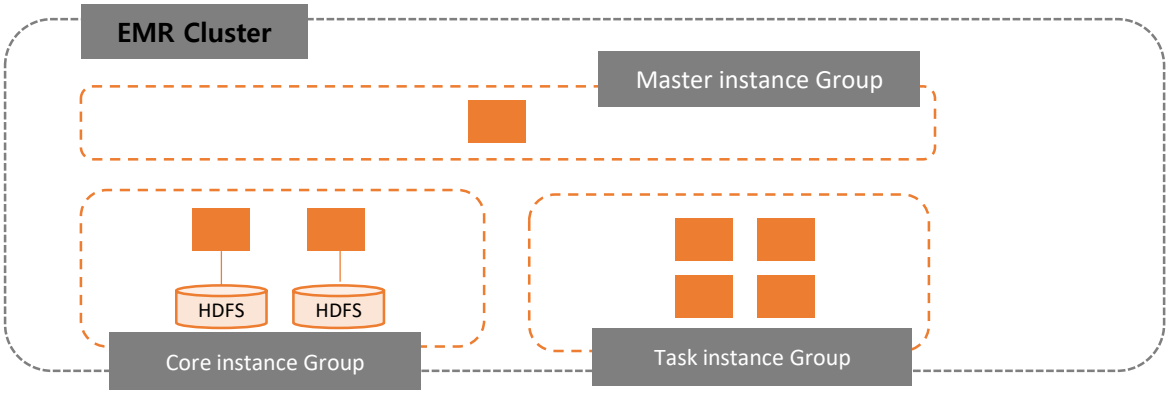
EC2 Provisioning ❯ Cluster Setup ❯ Hadoop Configuration ❯ Installing Applications ❯ Job submission ❯ Monitoring and Failure Handling

데이터 처리 [EMR]

❖ EMR 제공 Application

Software	version
Hadoop	2.8.3
Tez	0.8.4
HBase	1.4.2
Presto	0.194
Sqoop	1.4.6
Phoenix	4.13.0
HCatalog	2.3.2
Zepplin	0.7.3
Flink	1.4.0
Pig	0.17.0
Zookeeper	3.4.10
Mahout	0.13.0
Oozie	4.3.0
Livy	0.4.0
Ganglia	3.7.2
Hive	2.3.2
MXNet	1.0.01
Hue	4.1.0
Spark	2.3.0

❖ EMR Nodes



Node	설명
Master	• Cluster 관리
Core	• HDFS에 데이터를 저장하고 처리 저장과 처리력 강화 • Input data size에 따라 core node의수 결정
Task	• 단순 데이터 처리 • Default 에서는 task type은 0개 • 데이터를 저장하지 않음 • 처리력 증가에 필요

Cloudformation을 통한 자동화

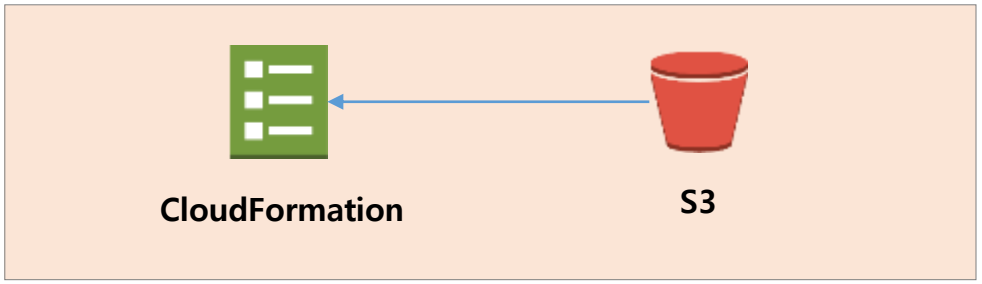
S3

CloudFormation

- ✓ System Key 설정
- ✓ Subnet 구성
- ✓ Role 세팅
- ✓ Application 추가 설치(python 등)
- ✓ Metastore 반영
- ✓ Library 구성
- ✓ Server Timezone 구성
- ✓ EMR Application 구성
- ✓ Server/EBS 구성

Amazon EMR

Cloudformation을 통한 자동화



- 어플리케이션 선택
- Role 설정
- BootstrapActions구성 : meta DB연결을 위한 JDBC다운로드
- 전체 클러스터 서버 시간 변경
  Hue, hive어플리케이션이 s3를 검색하도록 설정
- Configuration 구성 : spark (로그설정, 삭제 주기 등)
  zeppelin 환경 설정
- Instance type 및 EBS 구성
- Subnet, SecurityGroup 구성

- 클러스터 형상을 항상 동일하게 구성
- 장애에 긴급하게 대응
- 자주 클러스터를 구성해야 하는 경우
- 설정정보들을 한번에 관리 가능

EMR 구성 최적화

- **스트리밍 클러스터와 데이터 분석 클러스터 분리 구성**

- **HIVE 복잡도 높은 쿼리 실행 또는 Spark event Log, yarn log 관리**

: Datanode의 로그를 위한 EBS 사이즈 확보

: Configuration 조정을 통한 주기적 로그 삭제

- **NLB 구성, NLB와 EMR Master node 결합**

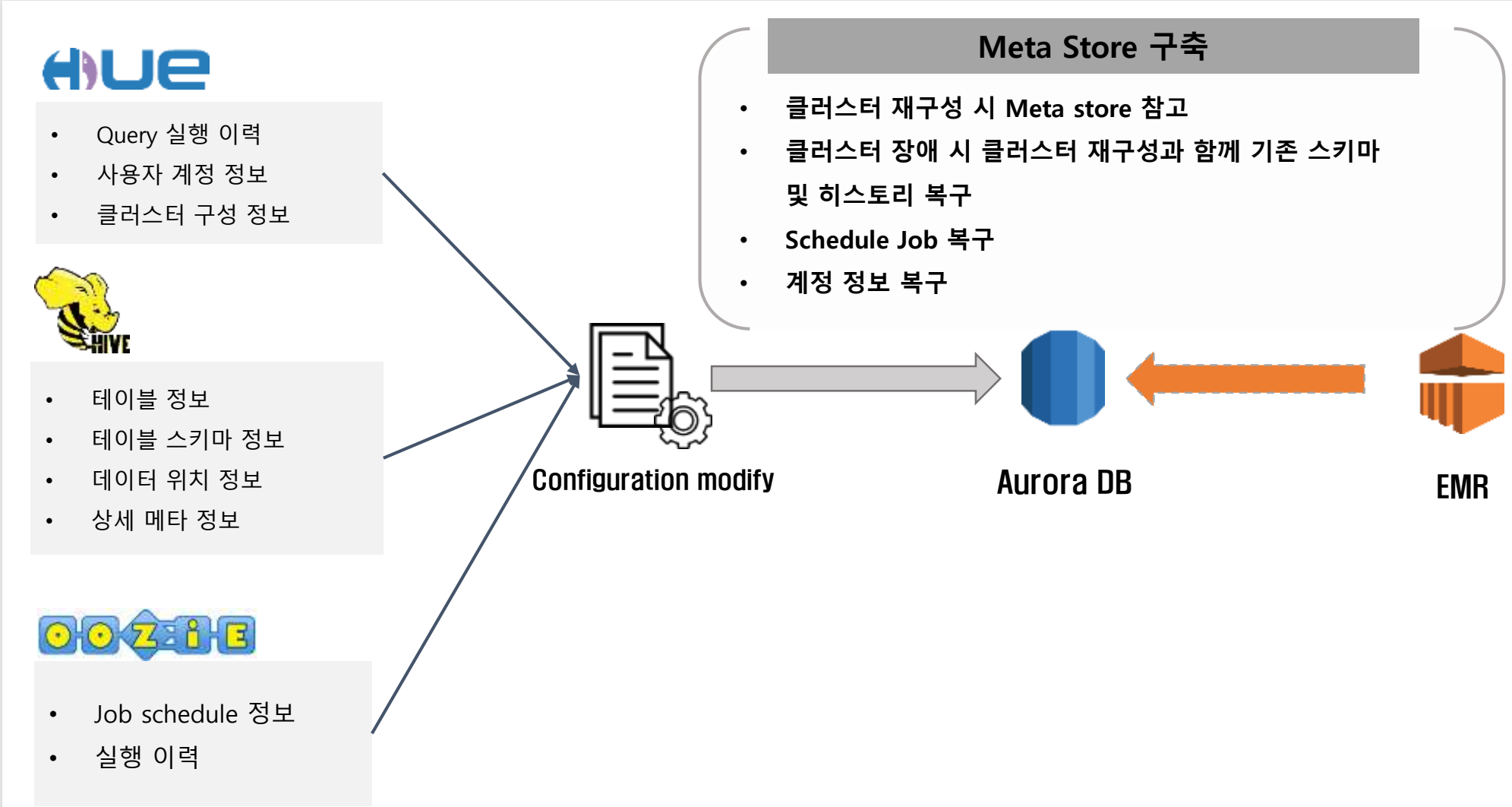
: 클러스터를 도메인으로 접근할 수 있도록 설정 가능

: 여러 가지 데이터 분석/쿼리 어플리케이션 접속 필요

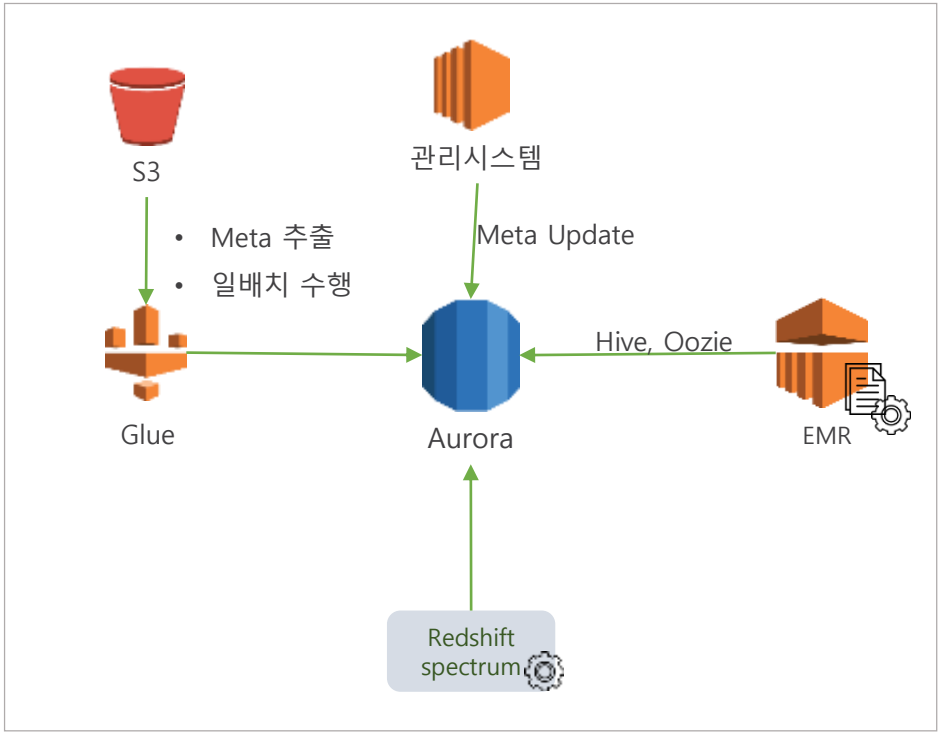
- **한꺼번에 노드를 갑자기 축소 하는 행위**

: log 및 job을 너무 작은 노드로 합칠 경우 data node가 수용이 불가하여 hang이 걸림.

데이터카탈로그 구성


Meta Store 구축
• 클러스터 재구성 시 Meta store 참고
• 클러스터 장애 시 클러스터 재구성과 함께 기존 스키마 및 히스토리 복구
• Schedule Job 복구
• 계정 정보 복구
• Query 실행 이력
• 사용자 계정 정보
• 클러스터 구성 정보
• 테이블 정보
• 테이블 스키마 정보
• 데이터 위치 정보
• 상세 메타 정보
• Job schedule 정보
• 실행 이력
Configuration modify
Aurora DB
EMR

데이터카탈로그 구성



Data Catalog 구성

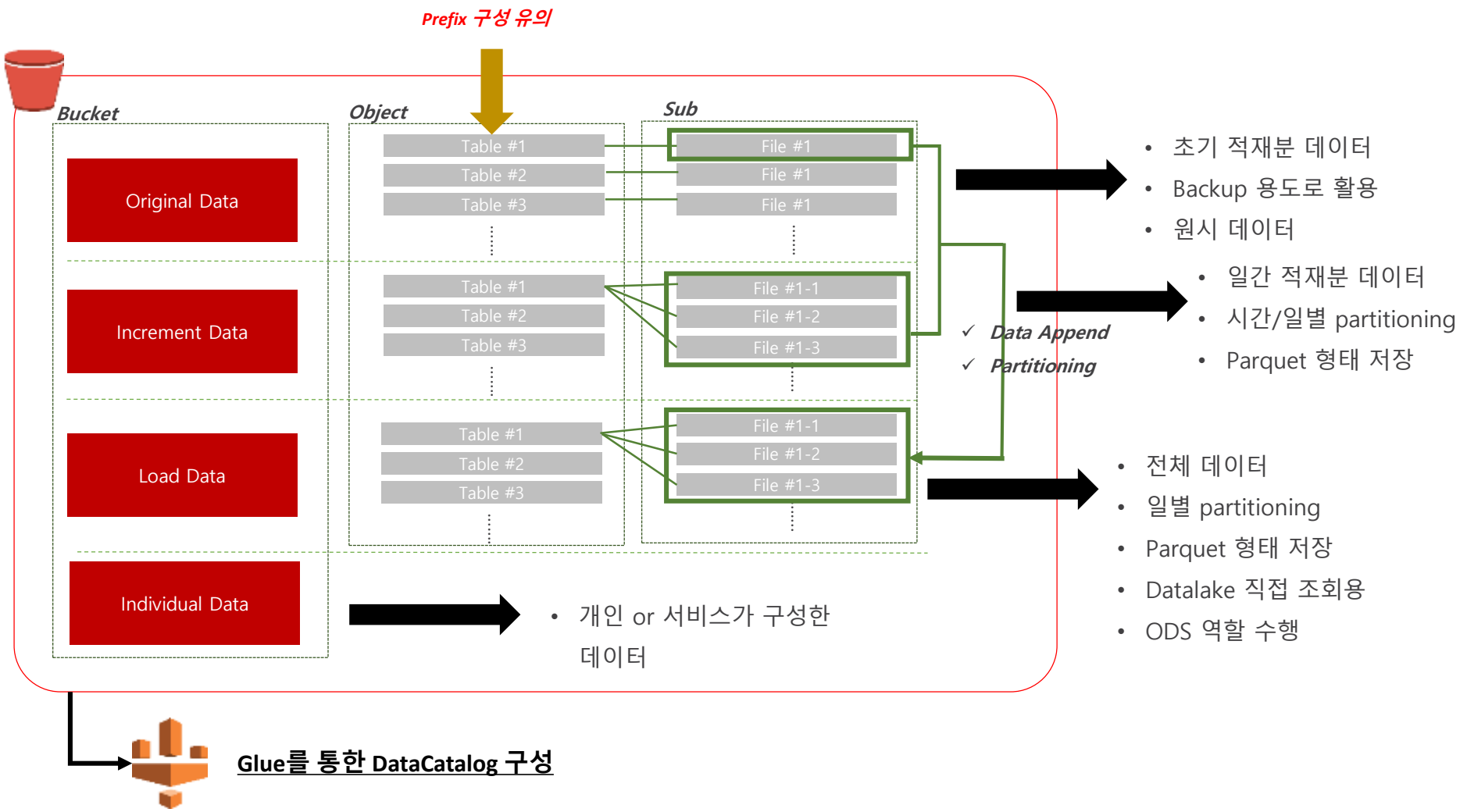
[Glue]
- S3 적재 데이터 메타 추출
- Crawler를 통한 주기적 메타 추출

[Spectrum]
- Redshift spectrum 메타 관리
- 사용자 테이블 조회용 메타

[EMR]
- Hive, Oozie 메타 추출/관리
- 클러스터 재구성용 메타 관리
- 어플리케이션 사용자 계정 관리

[관리시스템]
- 관리시스템의 메타 수정/관리

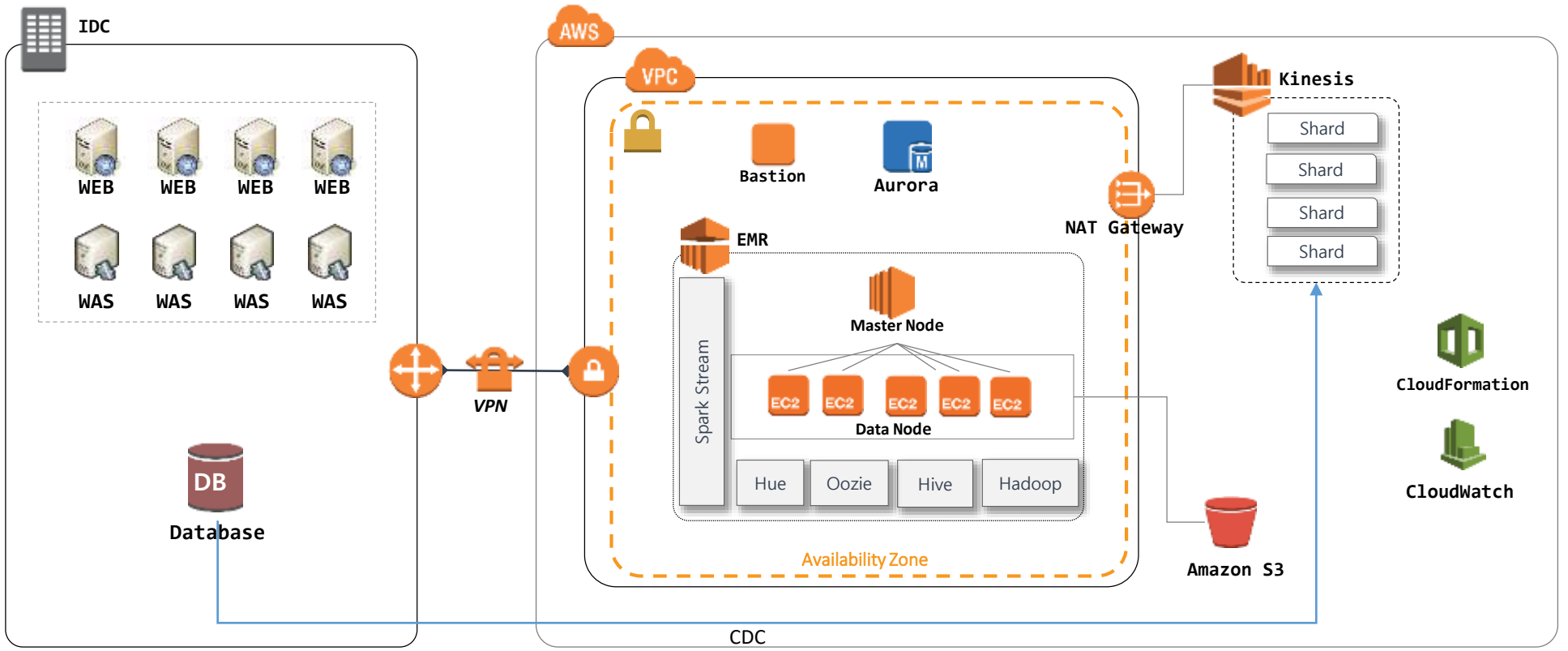
데이터 적재 [S3 구조]


Prefix 구성 유의
Bucket
Object
Sub
Original Data
Increment Data
Load Data
Individual Data
Table #1
Table #2
Table #3
File #1
File #1-1
File #1-2
File #1-3
✓ Data Append
✓ Partitioning
• 초기 적재분 데이터
• Backup 용도로 활용
• 원시 데이터
• 일간 적재분 데이터
• 시간/일별 partitioning
• Parquet 형태 저장
• 전체 데이터
• 일별 partitioning
• Parquet 형태 저장
• Datalake 직접 조회용
• ODS 역할 수행
• 개인 or 서비스가 구성한 데이터
Glue를 통한 DataCatalog 구성

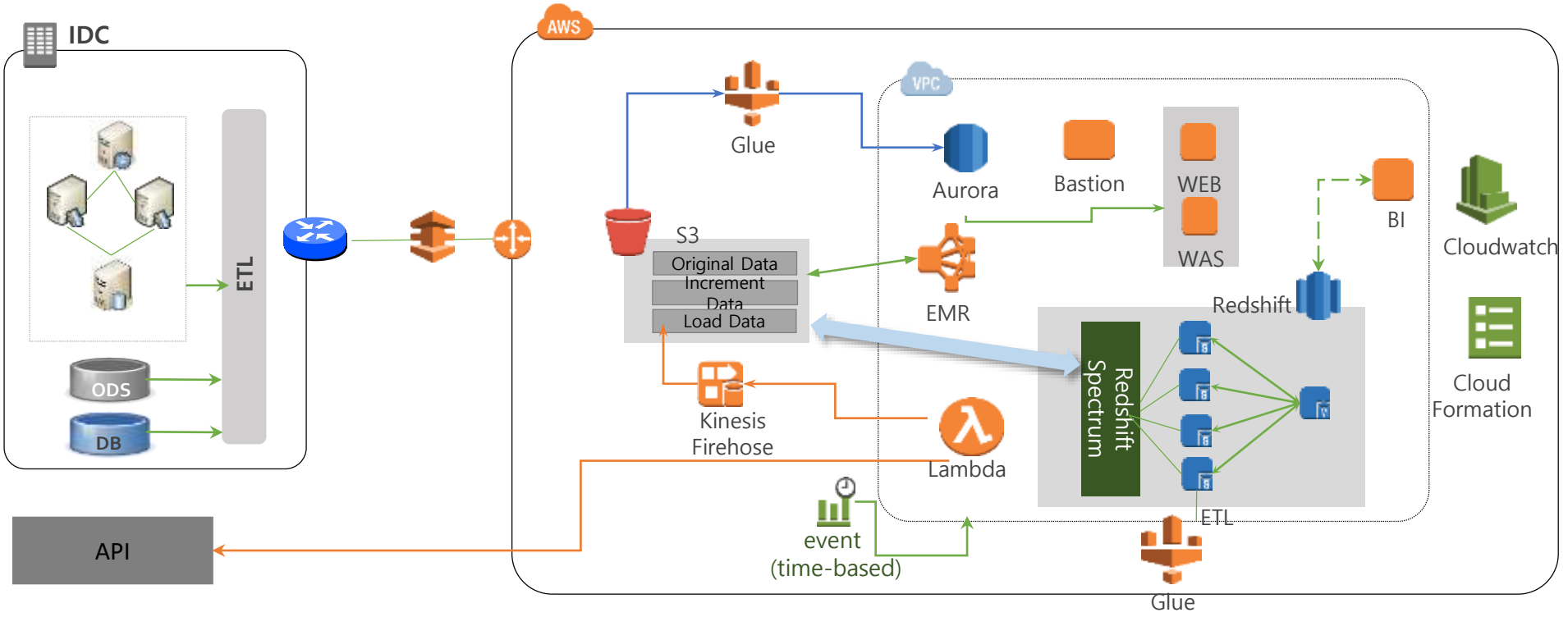
데이터 관리 시스템

권한 관리	데이터 접근 허용/제어 관리(IAM, Policy Dictionary) • 데이터 접근 권한 부여 • Policy dictionary 관리 • IAM 부여
카탈로그(메타) 관리	S3적재 데이터 메타 관리, 메타 정보 수정 • 파일 메타 조회/수정 관리
적재 파일 관리	데이터 적재 현황 관리 • 데이터 총 사이즈 • 테이블 별 파일 사이즈 • 금일 증분 데이터 사이즈

A 교육 사업 데이터레이크 구축



B사 데이터레이크기반 DW



THANK YOU