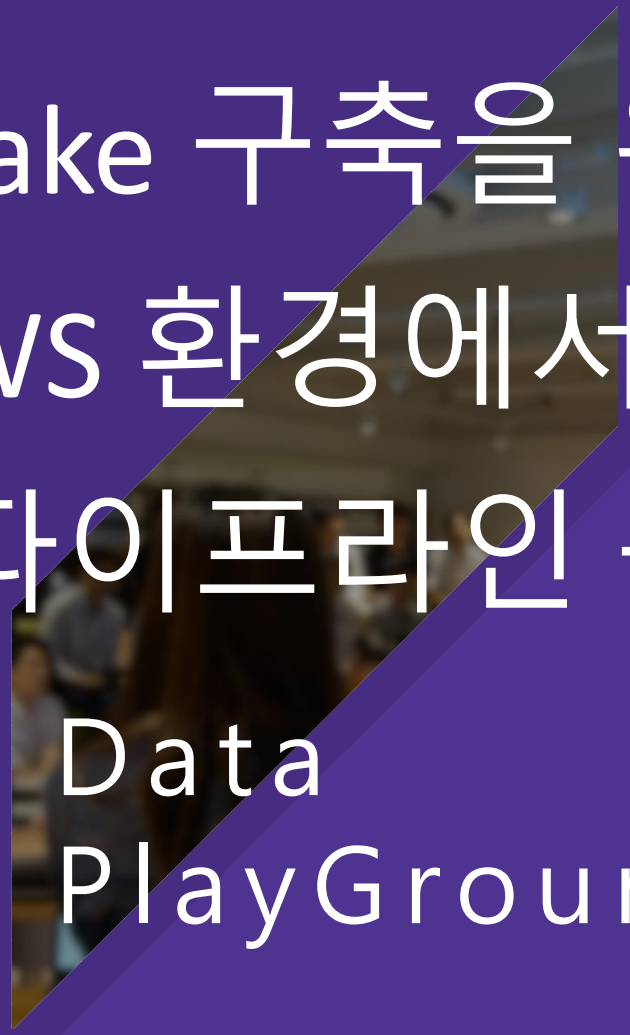
Data Lake 구축을 위한 AWS 환경에서 데이터 파이프라인 구성기



Data PlayGround@7

SK C&C

- 권낙주

목차

Fast campus

데이터 엔지니어로서
더 많은 **기회**를 얻고 싶다면,
클라우드 환경을 이해해야 합니다.

권낙주 강사님

現) 직방 데이터분석팀
데이터 아키텍트 책임자(AWS 환경)

- 현) SK C&C
- 전) 직방 데이터 분석팀

- 데이터 Architecture
- 꿈꾸는 데이터 Architecture

“엉망인 소스 데이터를

유용한 것으로 바꾸는 데 드는 시간이

데이터를 분석하는 나머지 시간을 합한 것보다 많다.”

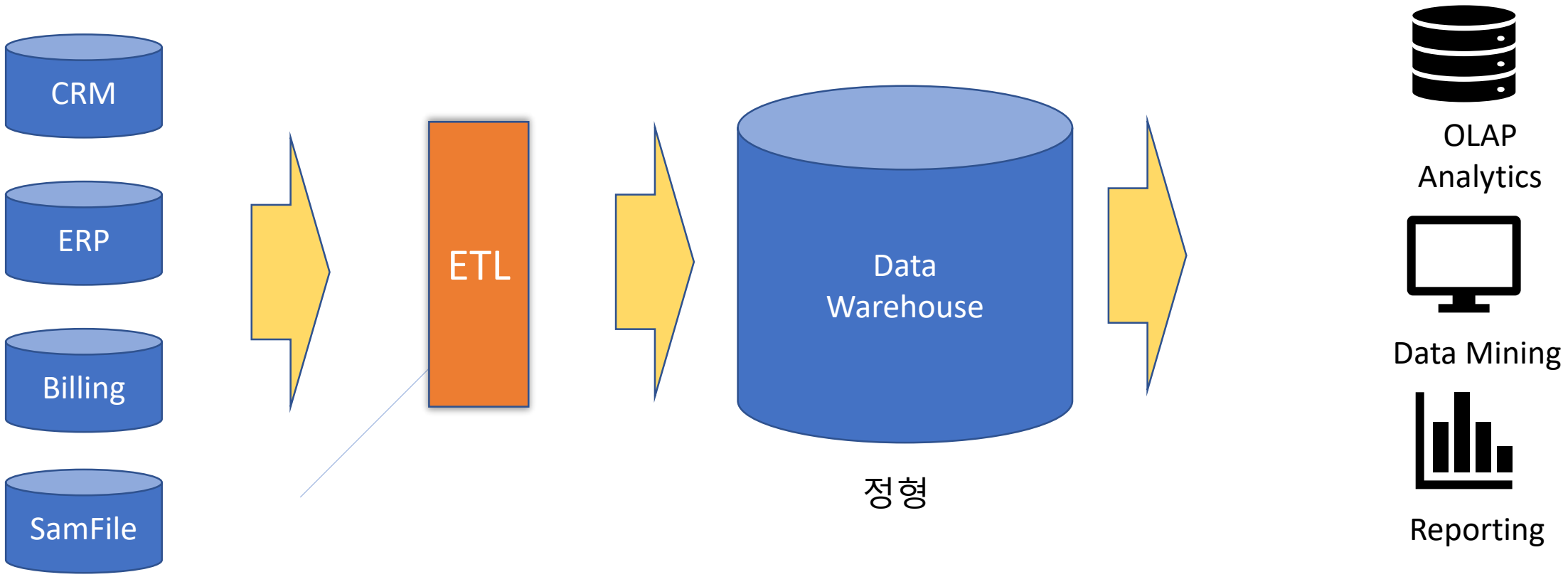
- Pete Warden, Google Data Engineer

Data Lake 나온 배경

Data Lake 나온 배경

Data Warehouse :

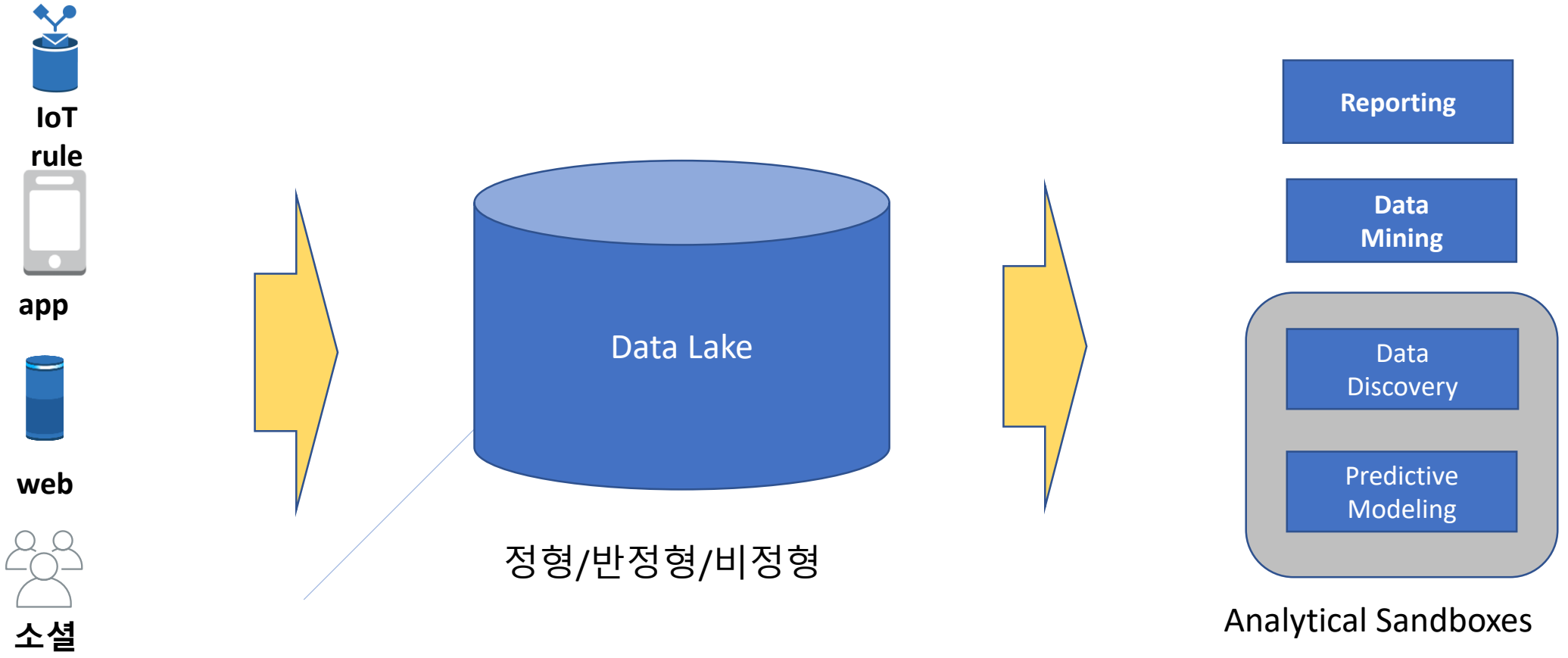
Data에 입각한 의사 결정을 내릴 수 있도록 분석 가능한 정보들의 집합



Data Lake 나온 배경

Data Lake 정의

데이터의 형태에 상관없이 무제한의 저장소로 Data 를 수집하는 형태



Data Lake를 위한

Data Pipeline 구성하기

Data Pipeline 구성하기

요구사항

- 매물 별 조회수 일/주/월 데이터 뽑을 수 있나요?
- 지역별 서비스 별 조회수 구할 수 있나요?
- 현재 사용자가 조회한 매물 조회 수 실시간으로 볼 수 있나요?

Data Pipeline 구성하기

App 서비스

사용자

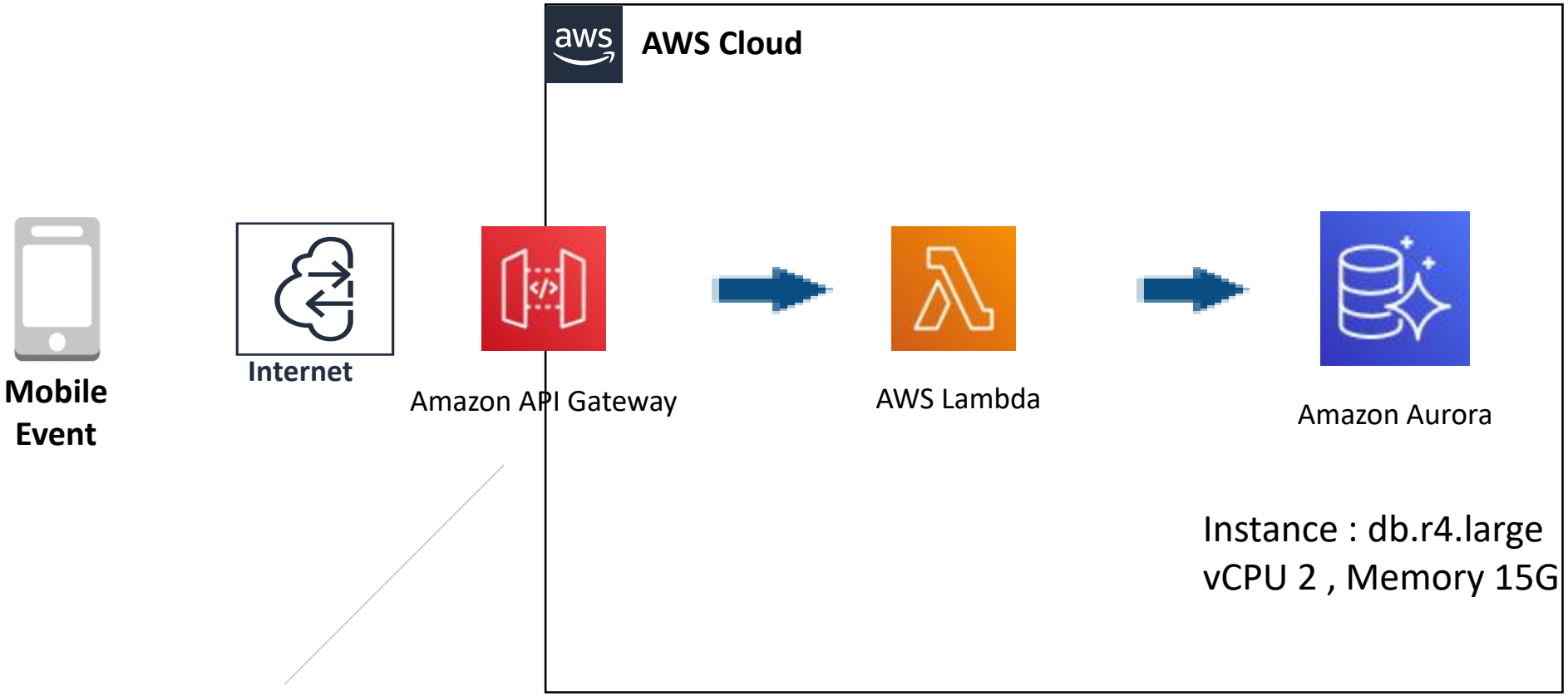
Data Pipeline 구성하기

현황 조사

GTM
(Google Tag Manager)

Data Pipeline 구성하기

데이터 수집



Data Pipeline 구성하기

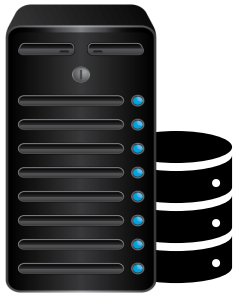
현황파악

- 일데이터 1천만건
- 일별로 Partition

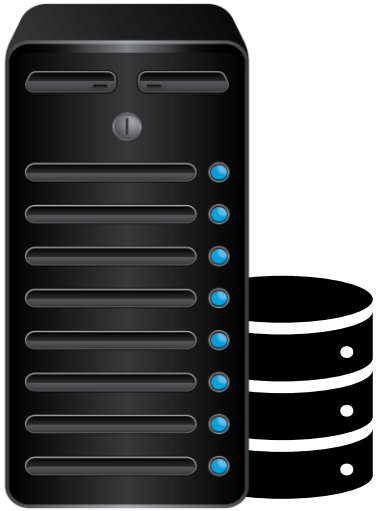
- 1주일 데이터 뽑을 수 있나요?

Data Pipeline 구성하기

Scale UP 해보시죠



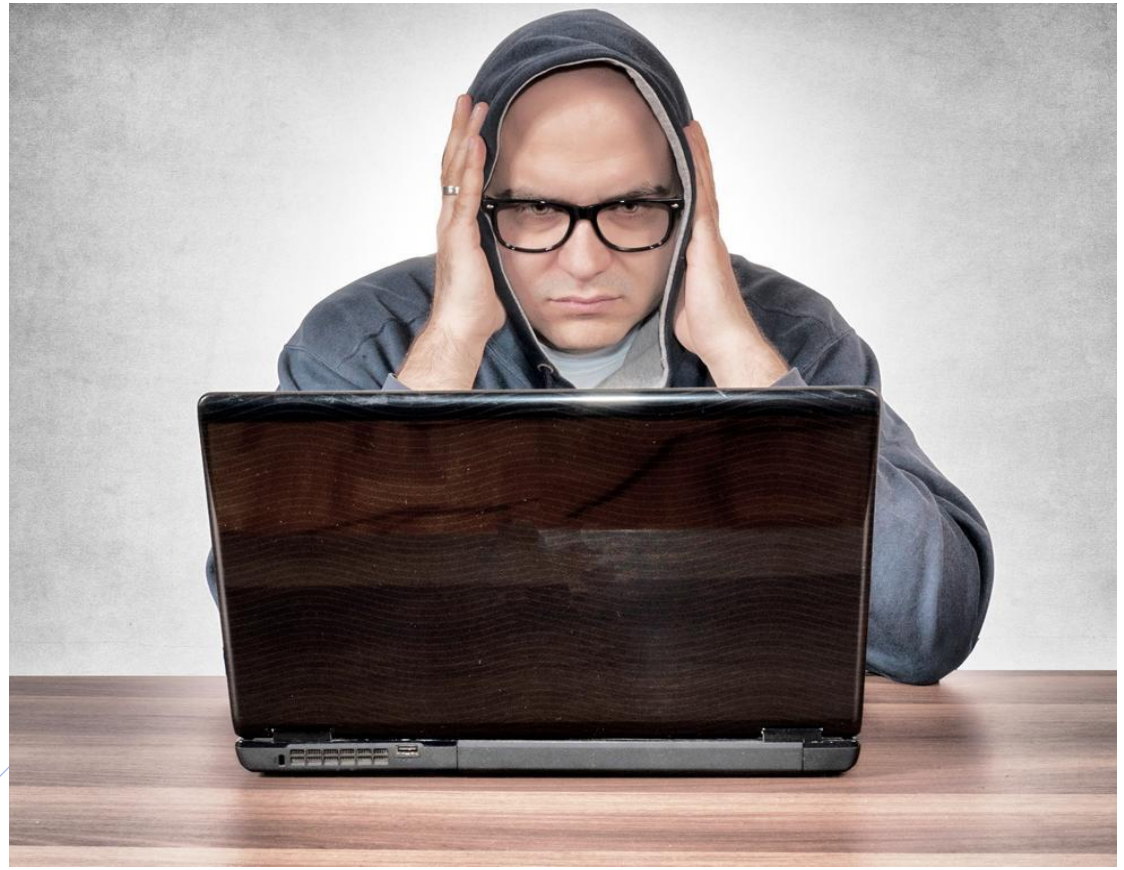
CPU, Memory



Instance : db.r4.large

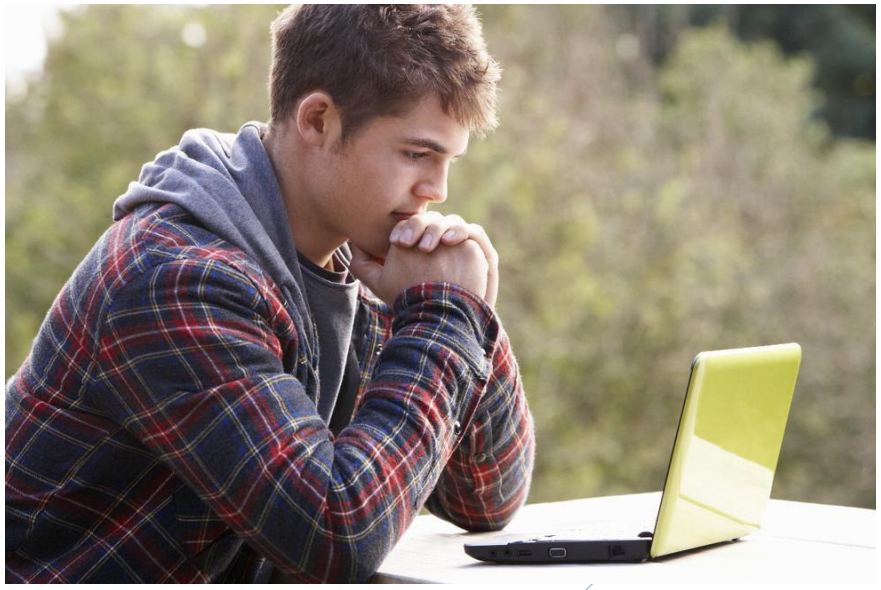
Instance : db.r4.xlarge

Data Pipeline 구성하기



- 1주 데이터 SQL 조회가 안된다

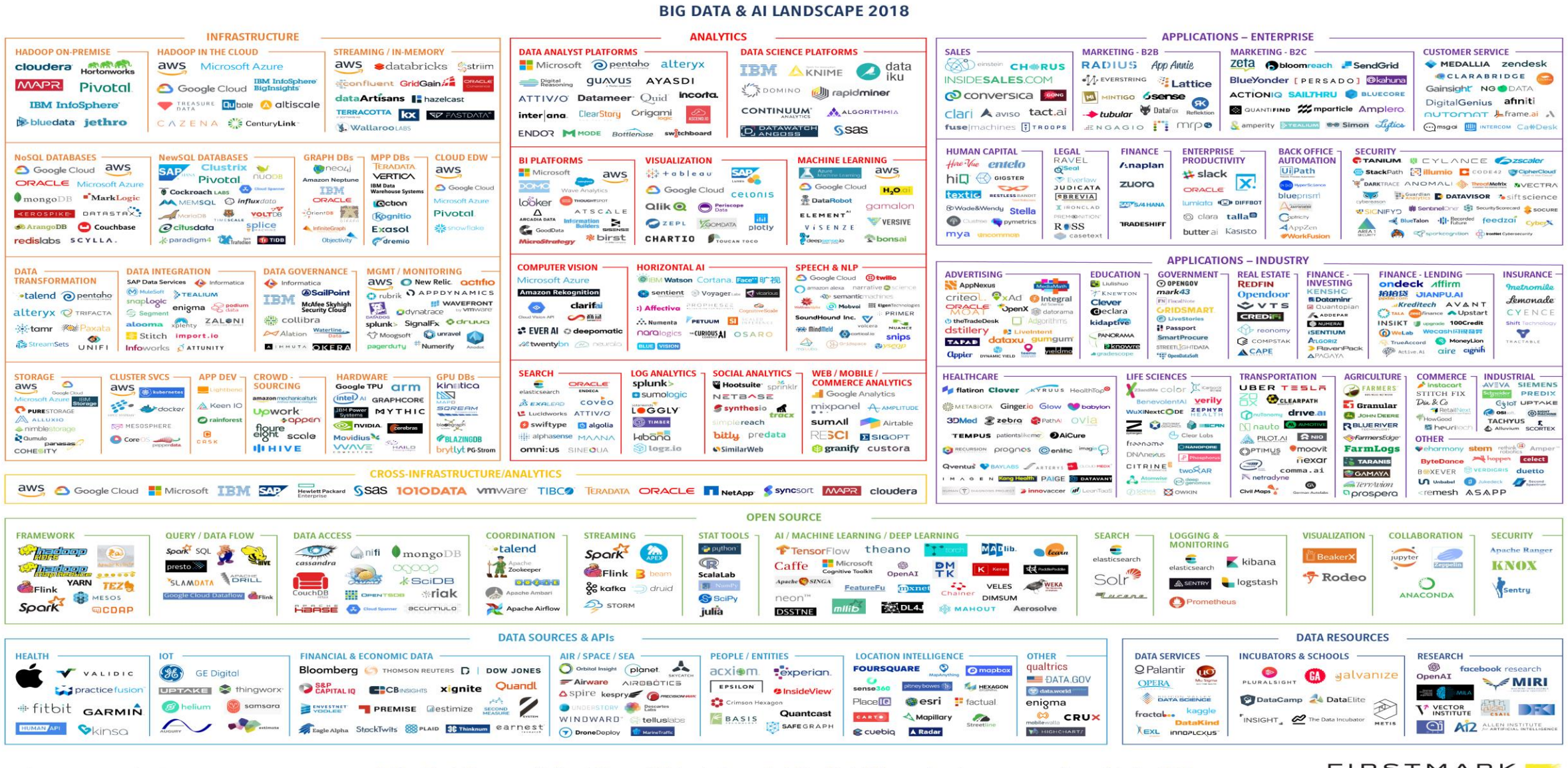
Data Pipeline 구성하기



Big Data 분석이 있다는데?

Googling

Data Pipeline 구성하기



Data Lake Architecture 구성

Data Lake Architecture 구성

수집

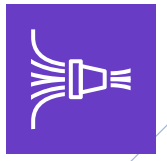

Amazon API Gateway


logstash


Amazon Kinesis Streams


Amazon Managed Streaming for Kafka


Amazon Kinesis Analytics


Amazon Kinesis Firehose

처리/분석


AWS Lambda


AWS DMS


Amazon EMR


Amazon Elasticsearch Service


AWS Data Pipeline


Amazon Athena

저장


Amazon S3


Amazon RDS

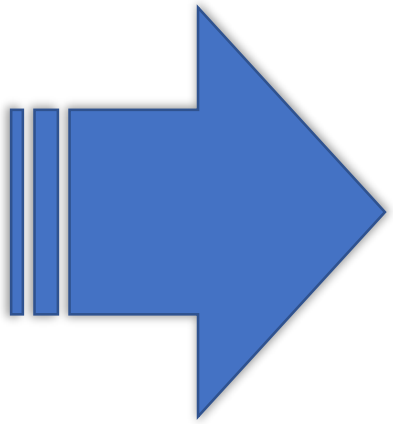

AWS Glue


Amazon Redshift


Amazon DynamoDB

Data Lake Architecture 구성

일 로그건수 : 1천 만건

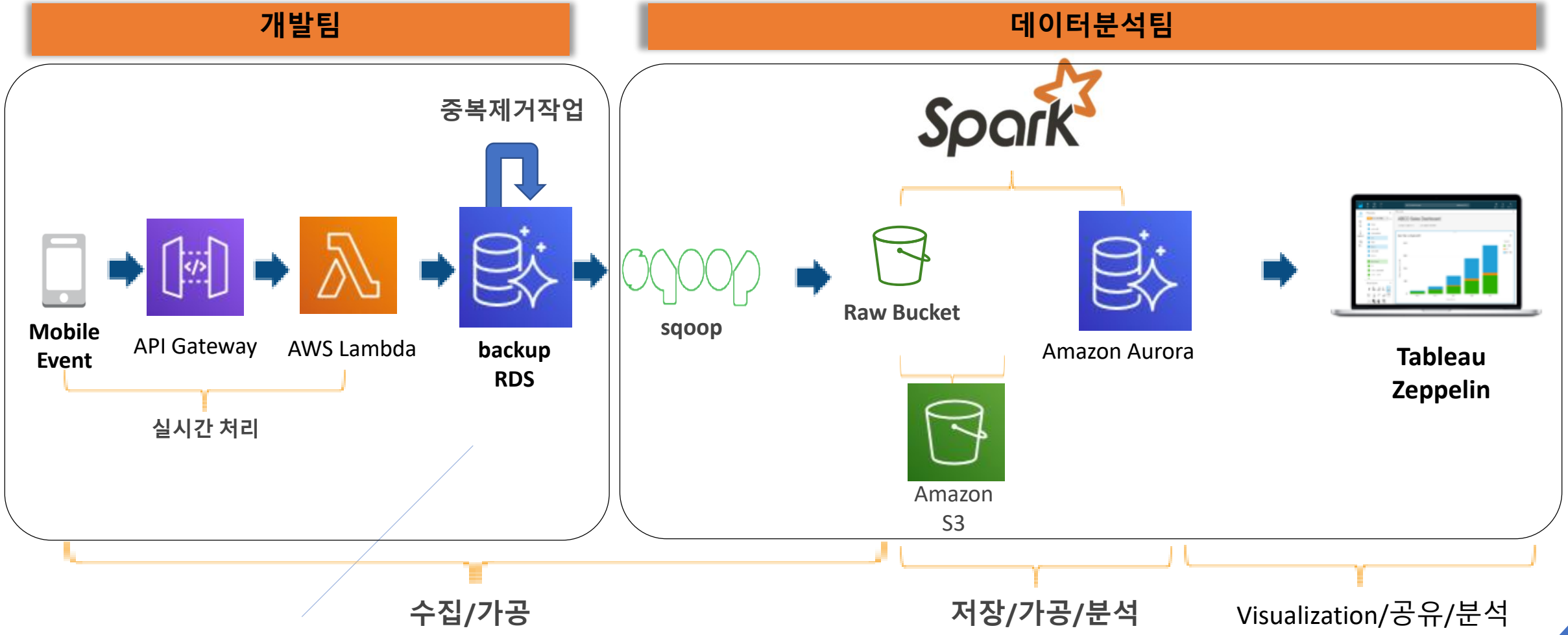


최근에 가장 Hot한 도구들

Data Lake Architecture 구성

분석을 어떻게 할 것인가?

Data Lake Architecture 구성



Data Lake Architecture 구성

개선된 점

1. 내부 데이터로 DAU, WAU, MAU 데이터 생성.
2. 일주일/월에 대한 사용자 재방문율 조회 가능.
3. 상품별 일자별 조회수가 가능.
4. 현제 Raw Bucket에 데이터 있다면 1일 내로 요청Use case에 대해서 데이터를 처리가 가능해졌다
5. 빅데이터 분석의 희망이 보임

Data Lake Architecture 구성

현황파악

1. 수집

- Push 같은 Event발생시에 RDS가 평소대비 CPU 사용량이 5배 올라감.
- MysqI 서비스에 과도한 Server 비용
- MysqI 관리가 항상 필요하다.

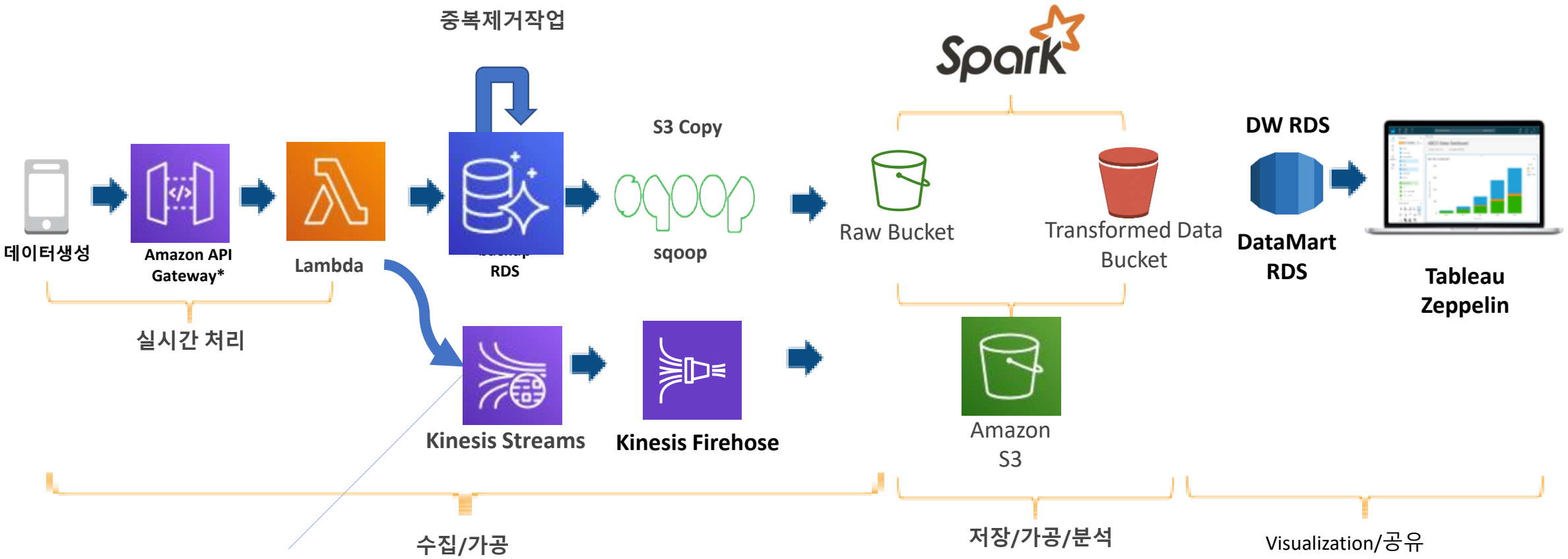
2. 분석

- Data Warehouse(DW) 용의 RDS 비용
- Data Warehouse(DW) 용량의 지속적 Scale UP이 필요

3. Visualization

- DW 기반(RDS)으로 제공 Data 느리다
- 너무 늦은 시간에 데이터 공유

Data Lake Architecture 구성

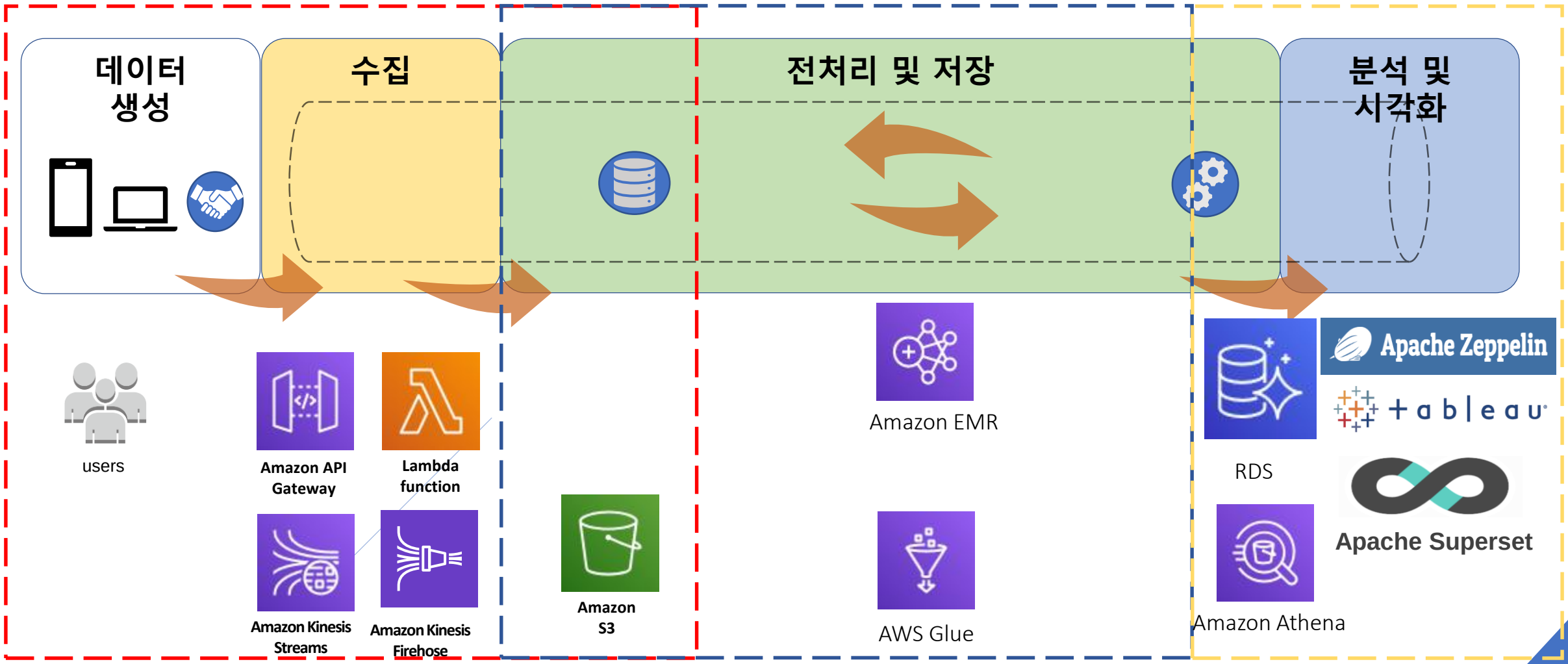


Data Lake Architecture 구성

1. 개선된 사항
   - 이벤트에 의한 과도한 몰림 현상을 극복
   - 아침시간에 데이터 공유가 가능해졌다.
   - 오전에 데이터를 통한 회의가 가능해짐

2. 남은 문제
   - DW 기반으로 제공 Data 느리다
   - Cloud Tableau 서비스라 느림
   - Meta data 관리가 되지 않는다.

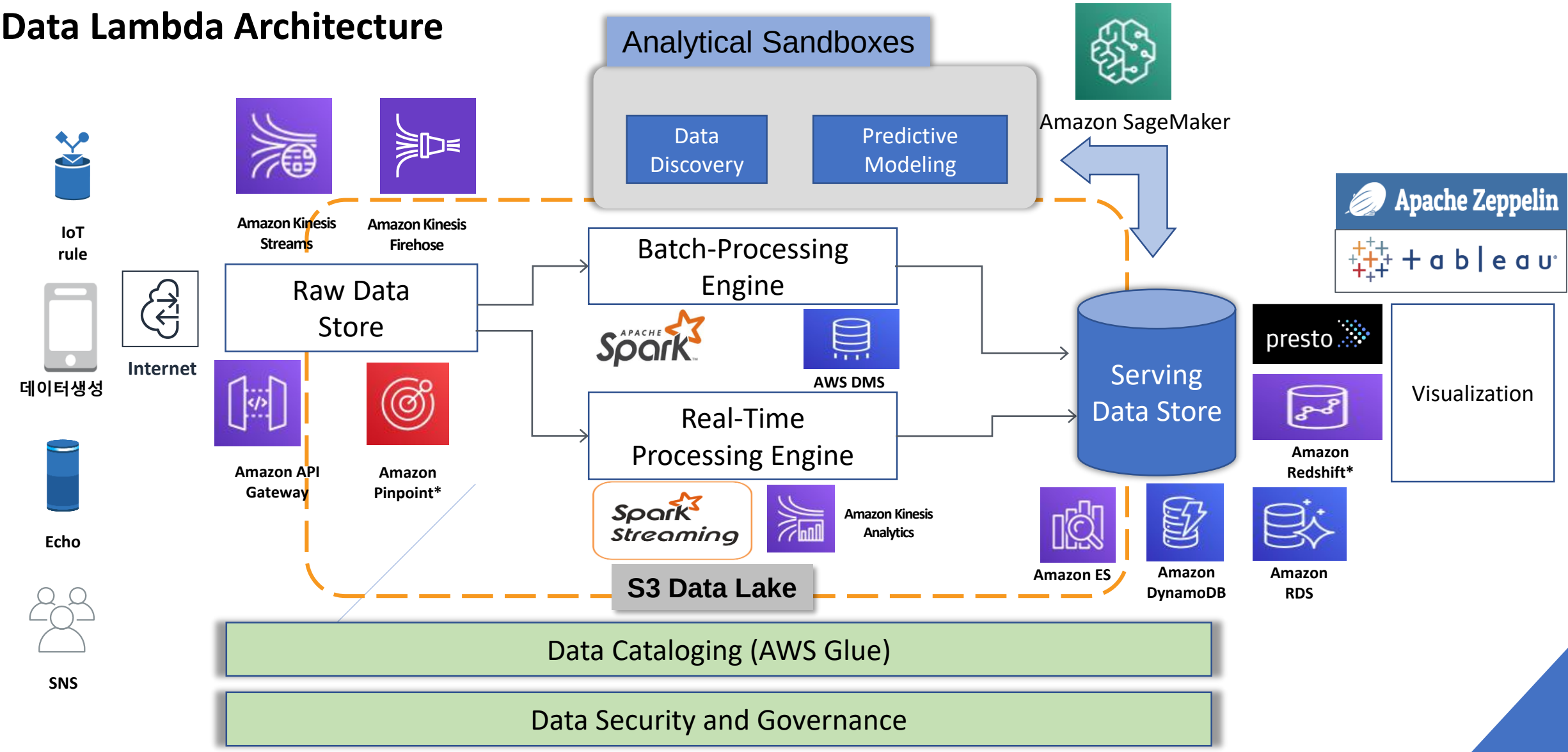
Data Lake Architecture 구성



Data Lake 구성방안

Data Lake 구성 방안

Data Lambda Architecture



Data Lake 구성 방안

1. Bigdata 분석 환경
   - ✓ Cloud를 사용하라
   - ✓ Cloud 장점인 Infra를 Coding 하라.
2. 저장
   - ✓ S3에 저장하라
   - ✓ 모든 데이터(sns, 사진, 비디오 등)를 저장하라
   - ✓ 압축하라(Parquet 형식으로 저장)
   - ✓ Partitioning 하라
3. 분석
   - ✓ BigData Serverless를 추구
   - ✓ Spot Instance 를 최대한 활용하라
4. Governance
   - ✓ Metadata 관리 (AWS Glue)

감사합니다.

Data Playground @ 7

https://www.facebook.com/groups/databreak/

http://databreak.org/