

1,000만 회원, MAU 500만을 위한 데이터 아키텍처

유환성

유한성

직업

직무보다는 가지고 있는 기술로 어떤 문제를 풀 것인지 고민하는 것을 좋아함
따라서 하고 있는 일이 언제든 바뀔 수 있다고 생각하며 지금은 데이터를 쉽게 활용할 수 있게 주변을 돕는 일을 하고 있음

고민

생산이 자동화 된다면 인간은 무엇을 하며 살아가야 하는가????
지속 가능한 즐거움이 무엇인가????

좋아하는 것

충분히 고민하고 적게 일하고 많이 쉬는 삶
주변과 함께 잘 되는 것(주변 도와주기)

싫어하는 것

더위와 불필요한 프로세스

추구하는 삶

절제된 쾌락

아젠다

- 무신사 데이터 아키텍처
- 물리적 데이터 아키텍처 & 논리적 아키텍처
- 더 많은 동료가 데이터에 접근하게 하기

무신사 데이터 아키텍처

데이터 아키텍처 변화의 필요성



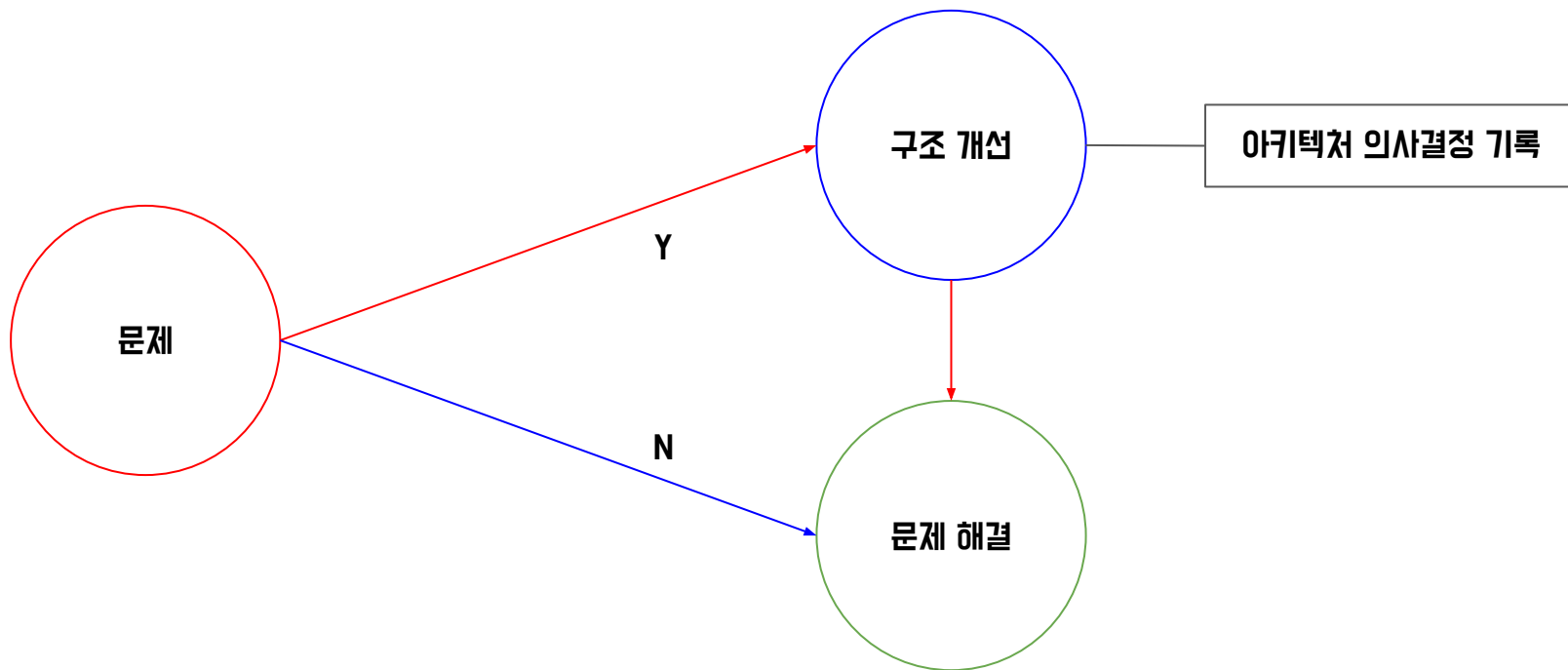
거래액 3X

브랜드 2X

회원수 3X

인원 3.5X

지금 구조에서 해결이 가능한가?



긴 기간의 데이터 조회

문제

- 운영 DB에 바로 접속에 데이터를 조회
- 그로 인해 index 가 없거나 긴 기간의 데이터를 조회하는 것이 불가능 했음
- 1일의 매출을 조회하는데 수십초 - 이마저 거래액 증가에 따라 점점 느려졌던 환경

구조 개선

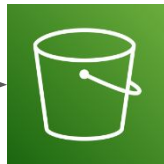
- DB의 데이터를 파일 기반으로 저장하여 분산 처리 할 수 있는 환경을 구축



Aurora



Aurora



S3



EMR

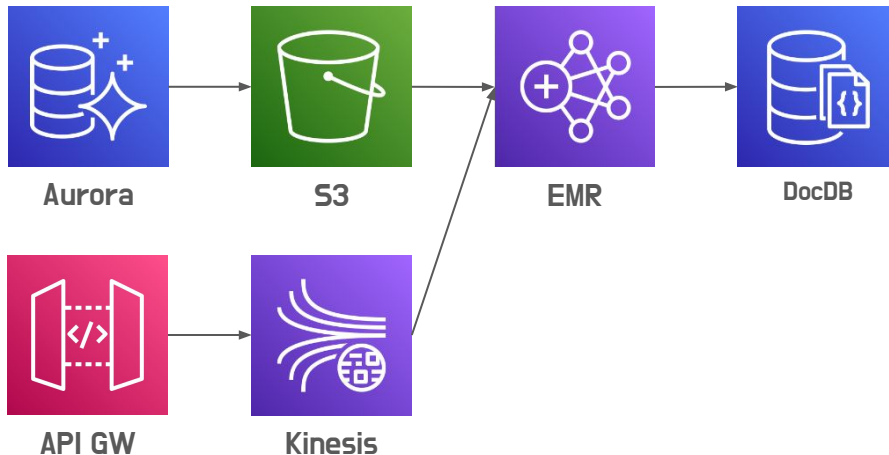
near-realtime 추천

문제

- 사용자의 행동 데이터 기반한 실시간 추천 시스템이 필요해짐
- 그에 따라 사용자의 행동 데이터를 실시간으로 저장해야할 필요가 생김
- 항상 시간은 충분하게 주어지지 않음

구조 개선

- API GW + Kinesis를 통해 실시간으로 데이터를 수집할 수 있는 환경을 마련함



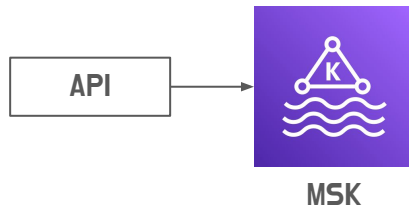
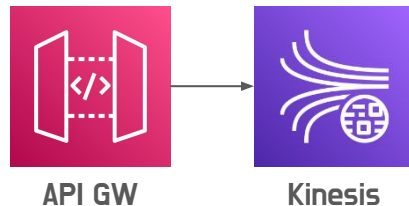
API GW + Kinesis

문제

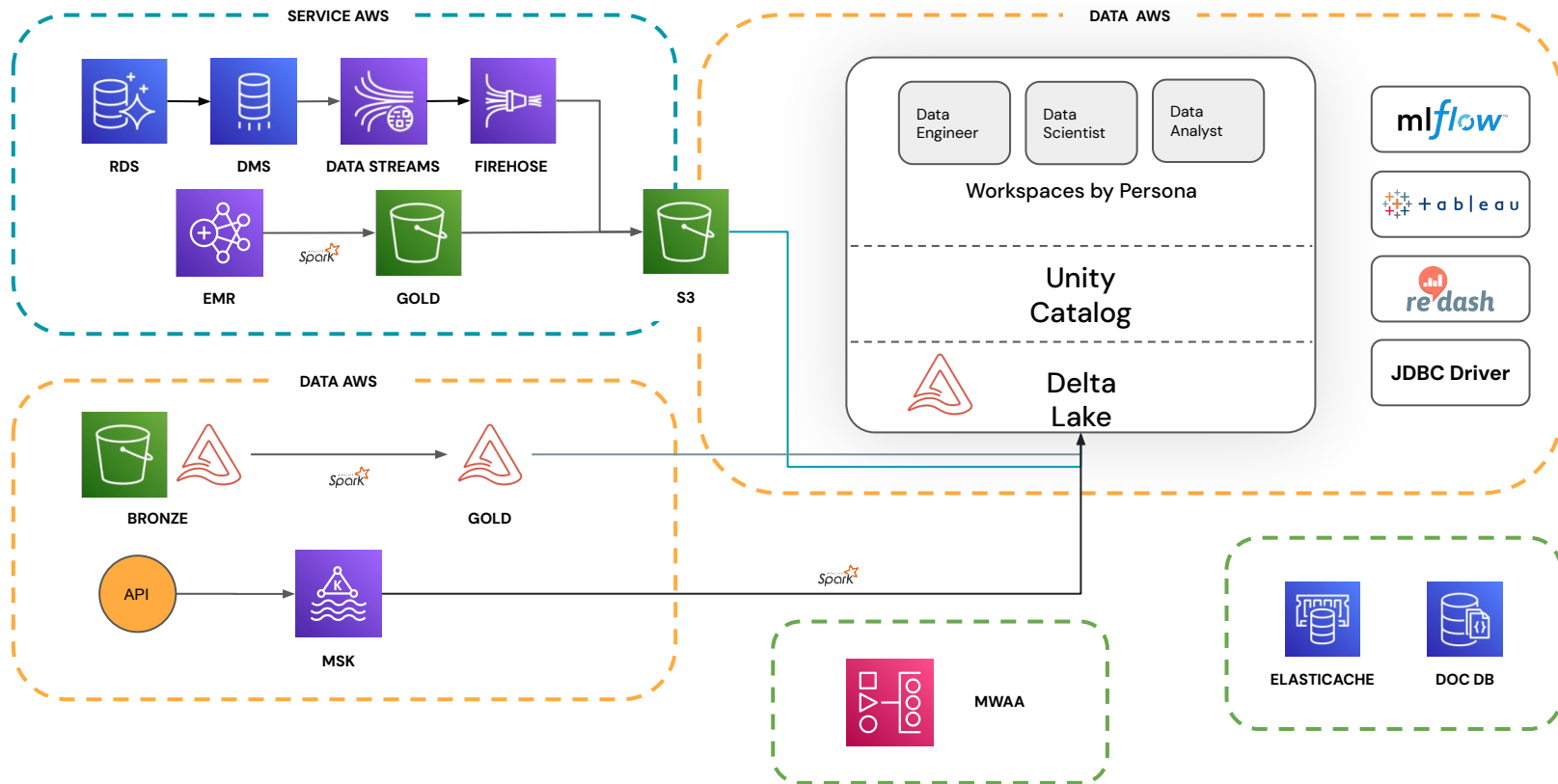
- API GW 사용량이 늘어남에 따라 비용이 급증
- 트래픽이 몰리는 경우 Auto Scale 이 동작하지만 타임아웃 다수 발생

구조 개선

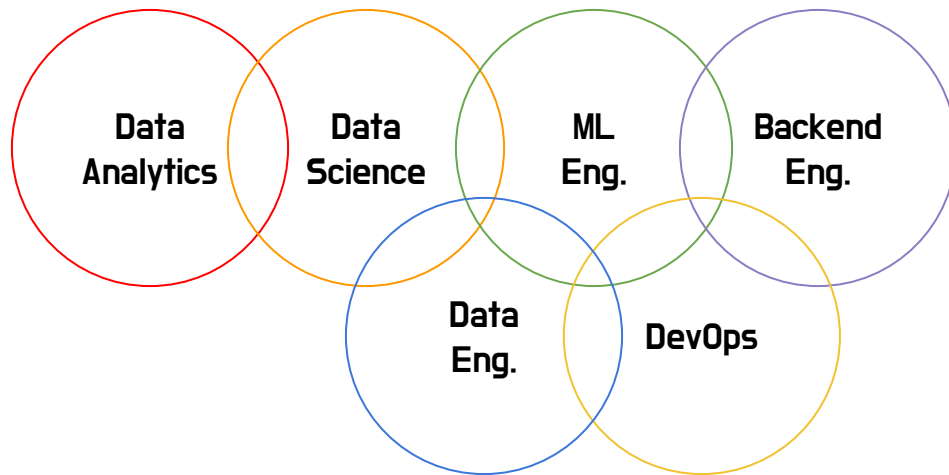
- API를 직접 개발하고, Kinesis 를 MSK로 대체



데이터 플랫폼 아키텍처



물리적 데이터 아키텍처 & 논리적 데이터 아키텍처



데이터 프로젝트



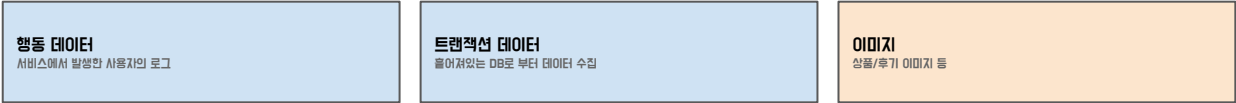
데이터 인프라(논리적)



데이터 프로세싱



데이터 수집



비즈니스 문제, 기술 문제

- 많은 회사에서 비즈니스 문제가 기술 문제
- 데이터 플랫폼을 만드는 이유는 데이터를 더 잘 활용하기 위함
- 데이터 플랫폼을 잘 만드는 것 만큼 데이터의 구조를 잘 만들어 두는 것 또한 중요
 - ex) 카테고리 별 첫 구매, 코호트



원본 데이터

OBJECT + ACTIVITY



지표

metric
100여 개

100 X 1,000 의 지표 추출 가능

dimension 1,000여 개 이상

더 많은 동료가 데이터에 접근하게 하기

N명의 동료와 일하기 - 팀 빌딩

- Spark -> PySpark
- PySpark -> PySpark Wrapping
- Convention
 - dag code convention : obj, obj attr, act, raw, merged ...
 - airflow task id 로 어떤 코드인지 유추가 가능
 - 코드의 영향 범위
 - column name
 - SQL
- 기술 의사 결정을 모두가 함께

N십명의 동료와 일하기 - 다른 데이터 조직 랜딩

- SQL, Python Notebook, BI 환경 제공
- Workflow 환경 제공
- 데이터레이크 제공
- 표준 가이드와 카탈로그 만들기
- 사람이 아닌 가이드 + 프로세스로 일하기

N백명의 동료와 일하기 - 회사 전체를 돕기

- 데이터 거버넌스 구축
 - 접근 권한 설정, 데이터 카탈로그 제공
- 데이터 활용 수준에 따라 다른 환경 제공하기
- 질의 응답이 아닌 가이드

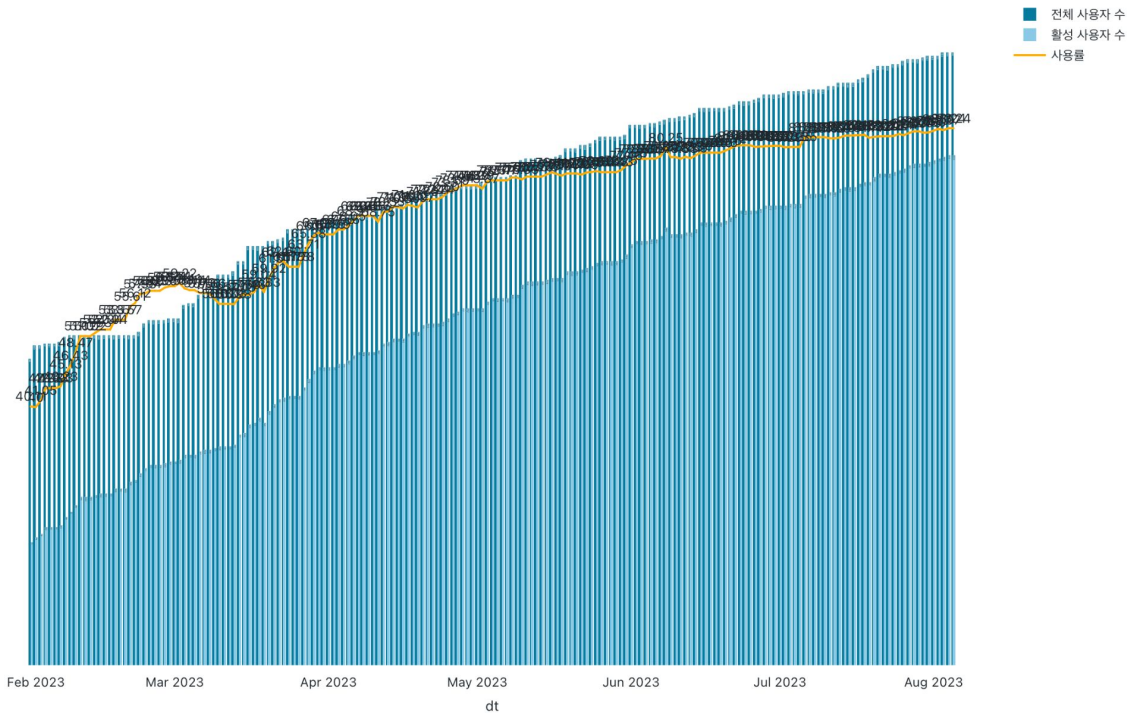
볼 수 있는

다룰 수 있는

**분석 할 수
있는**

데이터 거버넌스(논리적 데이터 아키텍처)

일자별 사용자 수



just now

감사합니다