1,000만 회원, MAU 500만을 위한 데이터 아키텍처

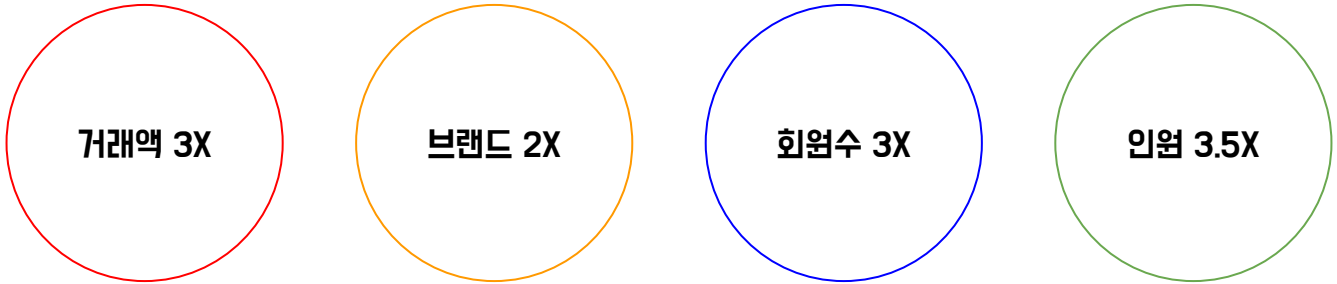
유환성

아젠다

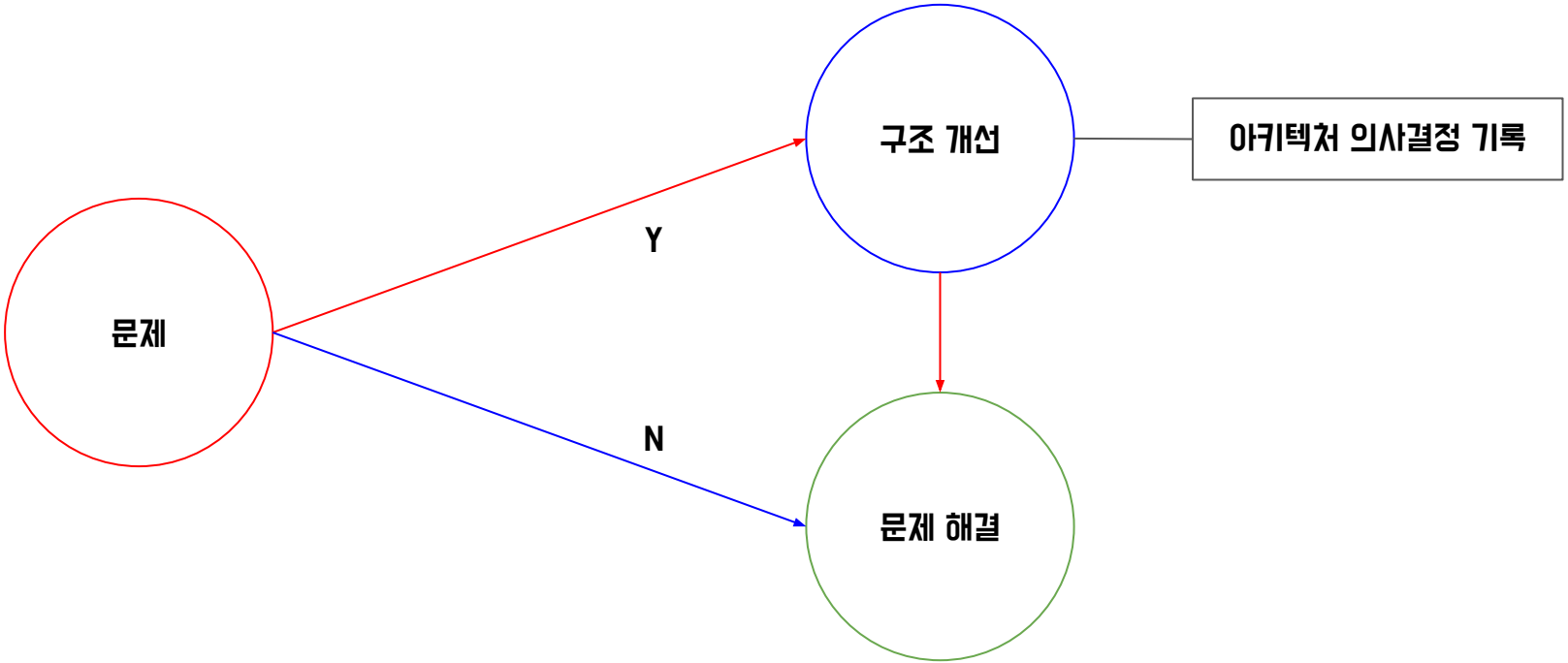
- 무신사 데이터 아키텍처
- 물리적 데이터 아키텍처 & 논리적 아키텍처
- 더 많은 동료가 데이터에 접근하게 하기

무신사 데이터 아키텍처

데이터 아키텍처 변화의 필요성



지금 구조에서 해결이 가능한가?



긴 기간의 데이터 조회

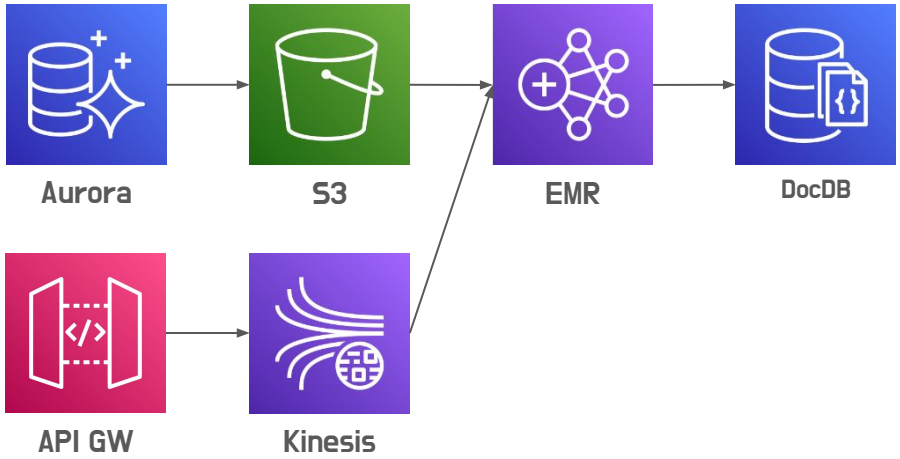
문제
- 운영 DB에 바로 접속에 데이터를 조회
- 그로 인해 index 가 없거나 긴 기간의 데이터를 조회하는 것이 불가능 했음
- 1일의 매출을 조회하는데 수십초 - 이마저 거래액 증가에 따라 점점 느려졌던 환경

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

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

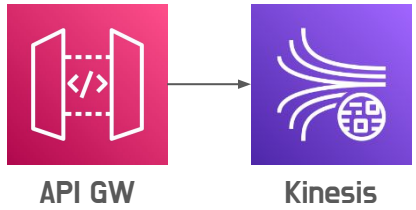
구조 개선
- API GW + Kinesis를 통해 실시간으로 데이터를 수집할 수 있는 환경을 마련함

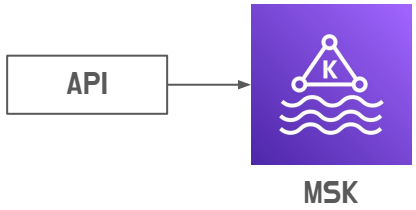


API GW + Kinesis

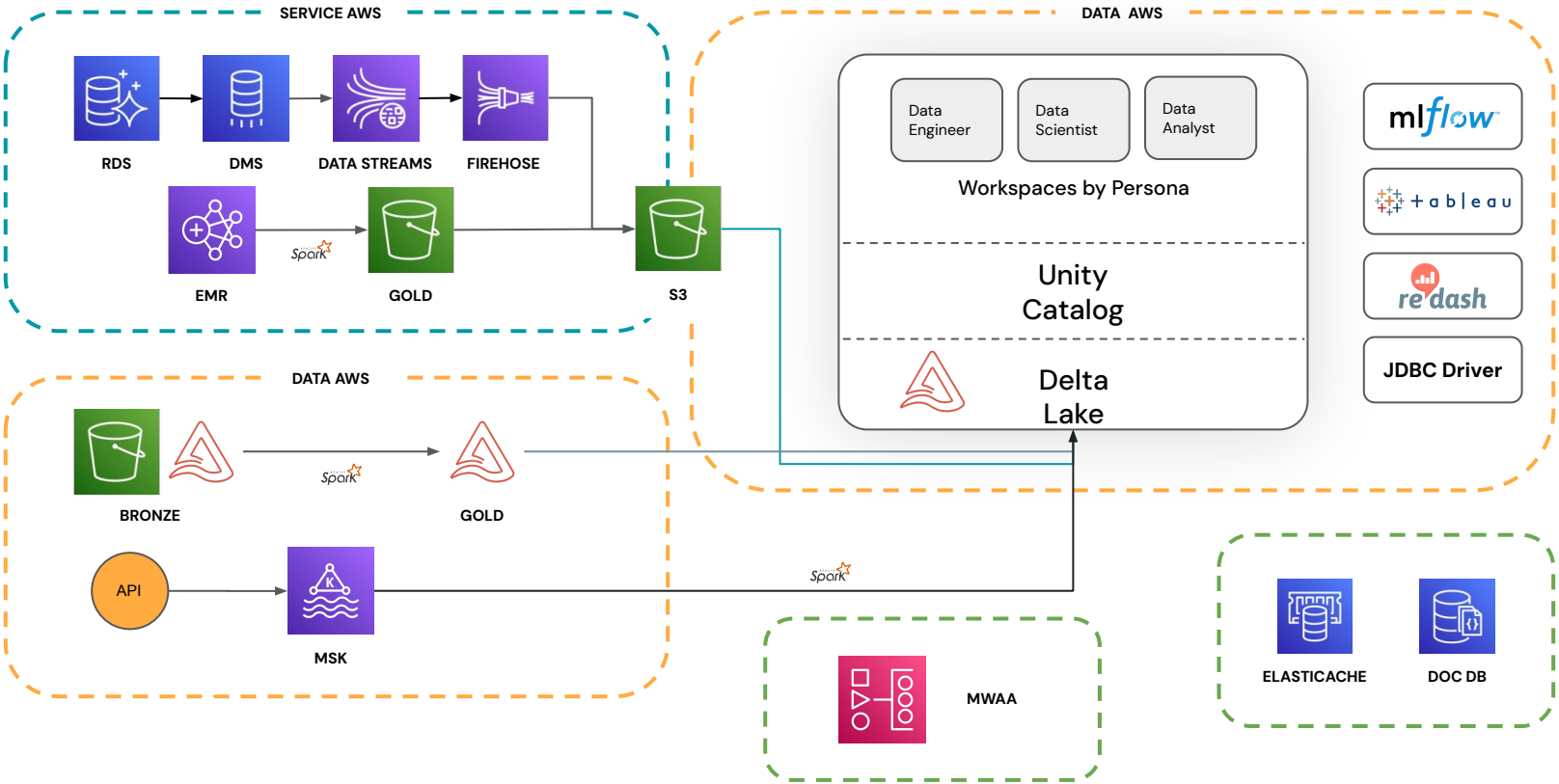
문제
- API GW 사용량이 늘어남에 따라 비용이 급증
- 트래픽이 몰리는 경우 Auto Scale 이
동작하지만 타임아웃 다수 발생

구조 개선
- API를 직접 개발하고. Kinesis 를 MSK로 대체

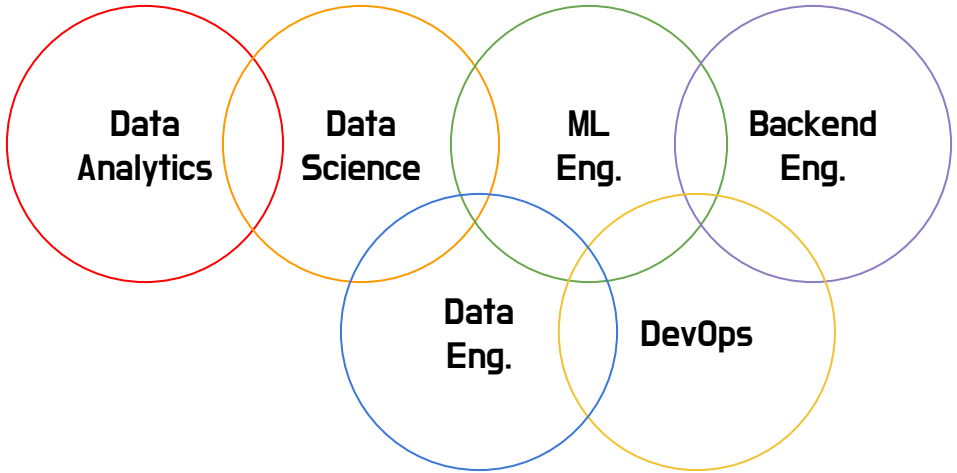


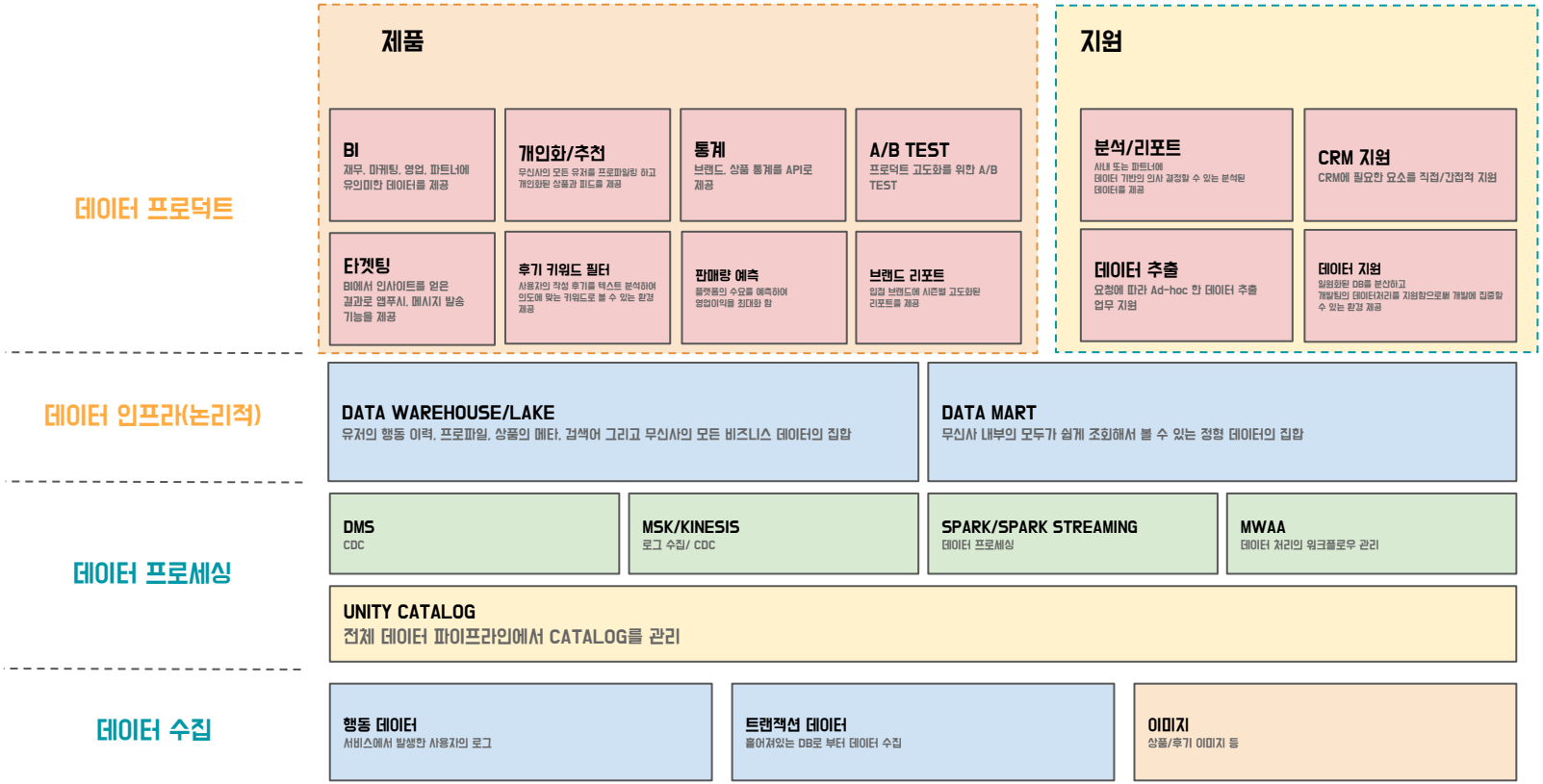


데이터 플랫폼 아키텍처


SERVICE AWS
RDS
DMS
DATA STREAMS
FIREHOSE
EMR
Spark
GOLD
S3
DATA AWS
Data Engineer
Data Scientist
Data Analyst
Workspaces by Persona
Unity Catalog
Delta Lake
mlflow
+ableau
redash
JDBC Driver
DATA AWS
BRONZE
Spark
GOLD
API
MSK
Spark
MWAA
ELASTICACHE
DOC DB

물리적 데이터 아키텍처 & 논리적 데이터 아키텍처


Data Analytics
Data Science
ML Eng.
Backend Eng.
Data Eng.
DevOps


데이터 프로덕트
제품
BI
재무, 마케팅, 영업, 파트너에 유의미한 데이터를 제공
개인화/추천
무신사의 모든 유저를 프로파일링 하고 개인화된 상품과 피드를 제공
통계
브랜드, 상품 통계를 API로 제공
A/B TEST
프로덕트 고도화를 위한 A/B TEST
타겟팅
BI에서 인사이트를 얻은 결과로 앱푸시, 메시지 발송 기능을 제공
후기 키워드 필터
사용자의 작성 후기를 텍스트 분석하여 의도에 맞는 키워드로 볼 수 있는 환경 제공
판매량 예측
플랫폼의 수요를 예측하여 영업이익을 최대화 함
브랜드 리포트
입점 브랜드에 시즌별 고도화된 리포트를 제공
지원
분석/리포트
사내 또는 파트너에 데이터 기반의 의사 결정할 수 있는 분석된 데이터를 제공
CRM 지원
CRM에 필요한 요소를 직접/간접적 지원
데이터 추출
요청에 따라 Ad-hoc 한 데이터 추출 업무 지원
데이터 지원
일원화된 DB를 분산하고 개발팀인 데이터처리를 지원함으로써 개발에 집중할 수 있는 환경 제공
데이터 인프라(논리적)
DATA WAREHOUSE/LAKE
유저의 행동 이력, 프로파일, 상품의 메타, 검색어 그리고 무신사의 모든 비즈니스 데이터의 집합
DATA MART
무신사 내부의 모두가 쉽게 조회해서 볼 수 있는 정형 데이터의 집합
데이터 프로세싱
DMS
CDC
MSK/KINESIS
로그 수집/ CDC
SPARK/SPARK STREAMING
데이터 프로세싱
MWAA
데이터 처리의 워크플로우 관리
UNITY CATALOG
전체 데이터 파이프라인에서 CATALOG를 관리
데이터 수집
행동 데이터
서비스에서 발생한 사용자의 로그
트랜잭션 데이터
흩어져있는 DB로 부터 데이터 수집
이미지
상품/후기 이미지 등

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


TOP MGMT
서로 다른 기준으로 인한 데이터 불일치 발생
영업
마케팅
테크
...
투자
데이터 플랫폼


TOP MGMT
동일 기준. 동일 데이터
영업
마케팅
테크
...
투자
데이터 거버넌스(논리적 데이터 아키텍처)
데이터 플랫폼

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

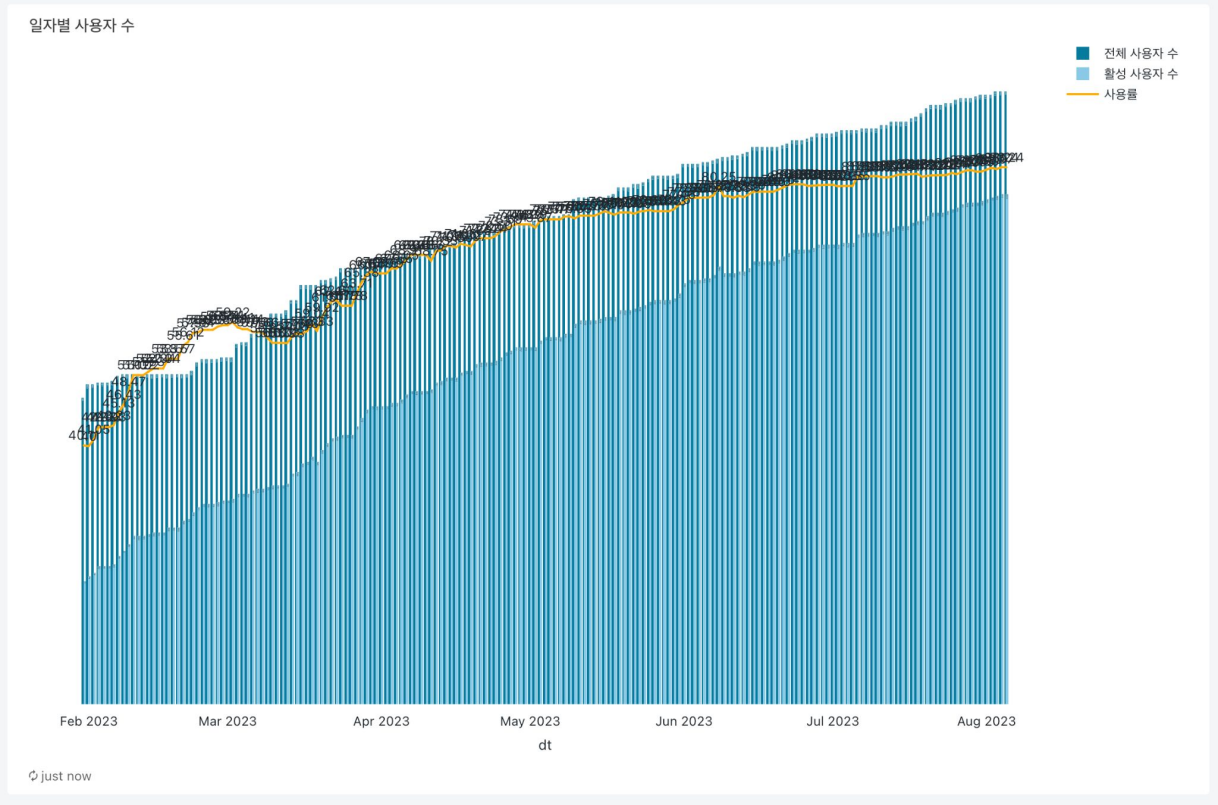
- 데이터 거버넌스 구축
  - 접근 권한 설정, 데이터 카탈로그 제공
- 데이터 활용 수준에 따라 다른 환경 제공하기
- 질의 응답이 아닌 가이드

볼 수 있는

다룰 수 있는

분석 할 수 있는

데이터 거버넌스(논리적 데이터 아키텍처)


일자별 사용자 수
전체 사용자 수
활성 사용자 수
사용률
Feb 2023
Mar 2023
Apr 2023
May 2023
Jun 2023
Jul 2023
Aug 2023
dt
just now

감사합니다