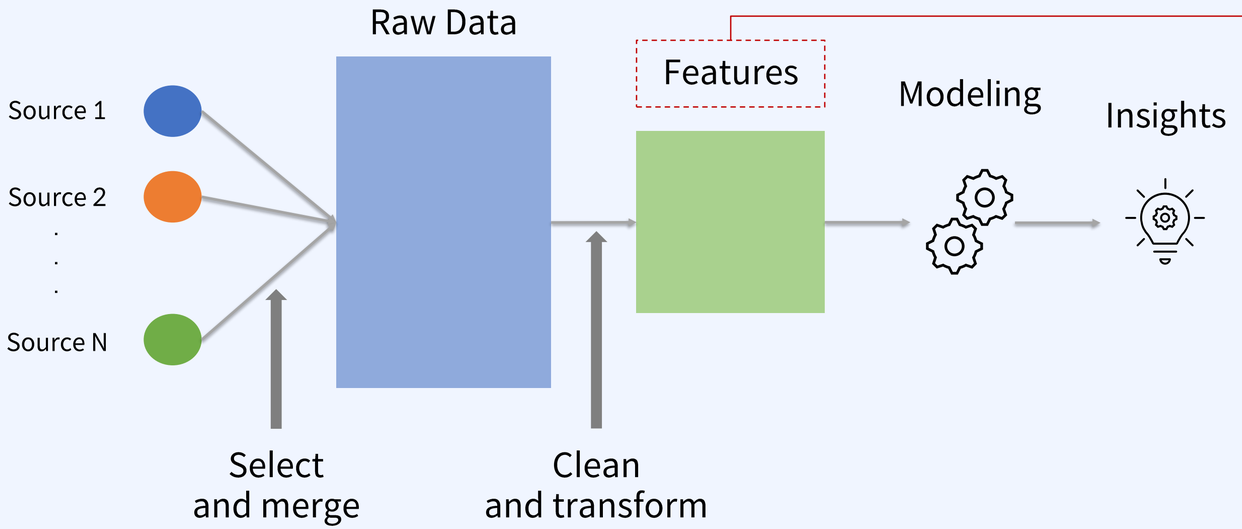
Part 04. [Chapter 2-1] Feature Store 기초

1 Feature Engineering 과 ML Pipeline

Feature 란?

Feature 는 관심 현상에 대해 개별적으로 측정 가능한 속성 또는 특징으로,
Machine Learning Model 의 입력값이다.



[Features 데이터 예시]

▷ 단어, 이미지 픽셀, 소리 음파, 센서값

▷ 축약값 (평균,최대값,합계,최소값 등)

▷ 시간값 (last_hour, last_day 등)

▷ 단위 데이터 (embedding, cluster 등)

Feature 의 필요성

ML Model 이 잘 동작하기 위해서는 많은 데이터 뿐만 아니라
잘 정의된 다양한 종류의 Feature 들이 필요합니다

네비게이션 ML Model 에 필요한 Feature 들

- ▷ 운전자 정보
- ▷ 지도 데이터
- ▷ 주유소/식당 등의 상점 정보
- ▷ 주행 이력
- ▷ 통행료
- ▷ 무료/유료 도로
- ▷ 요일별 차이

Feature Vector

원본 데이터를 Feature Engineering 을 통해 가공하여
Model 이 사용할 수 있는 Feature Vector 로 변환한다.

Raw Data

성별	[여자, 남자]
할인 쿠폰 코드	[가입축하, 생일축하, 추천인]
가입일	20211101140000

Feature Engineering

transform

Feature Vector

성별	[0, 1]
할인 쿠폰 코드	[1, 0, 0]
가입기간 (Year)	0.1

인코딩된 범주형 Features

표준화된 숫자형 Features

고객 정보

아이디	varchar
나이	int(3)
거주 도시	varchar
가입일	datetime

거래 내역

아이디	varchar
거래 일시	datetime
거래액	decimal
할인 쿠폰 코드	varchar

장바구니 정보

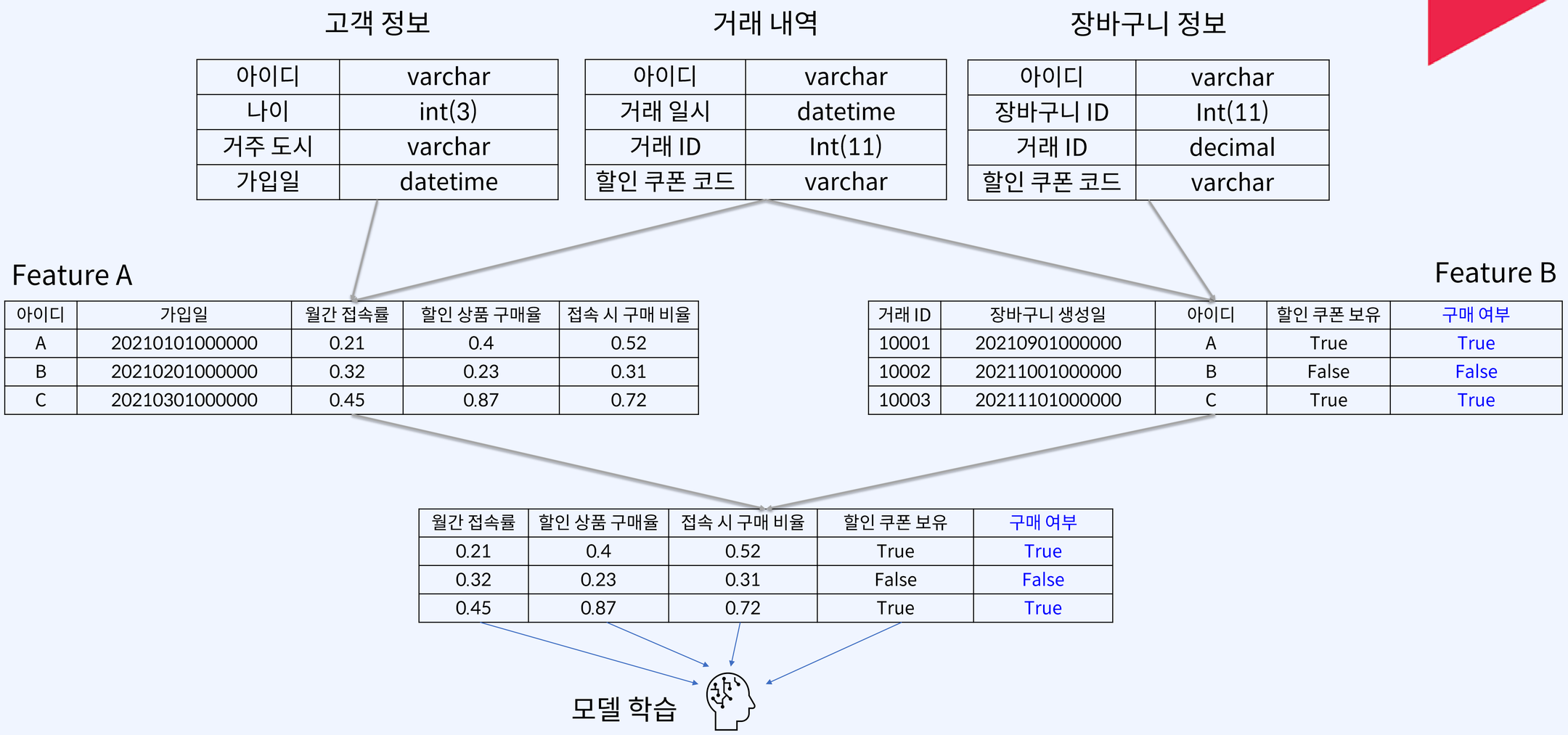
아이디	varchar
거래 ID	Int(11)
물품 ID	decimal
할인 쿠폰 코드	varchar

우수 회원

아이디	varchar
구독 아이디	decimal
가입일	datetime

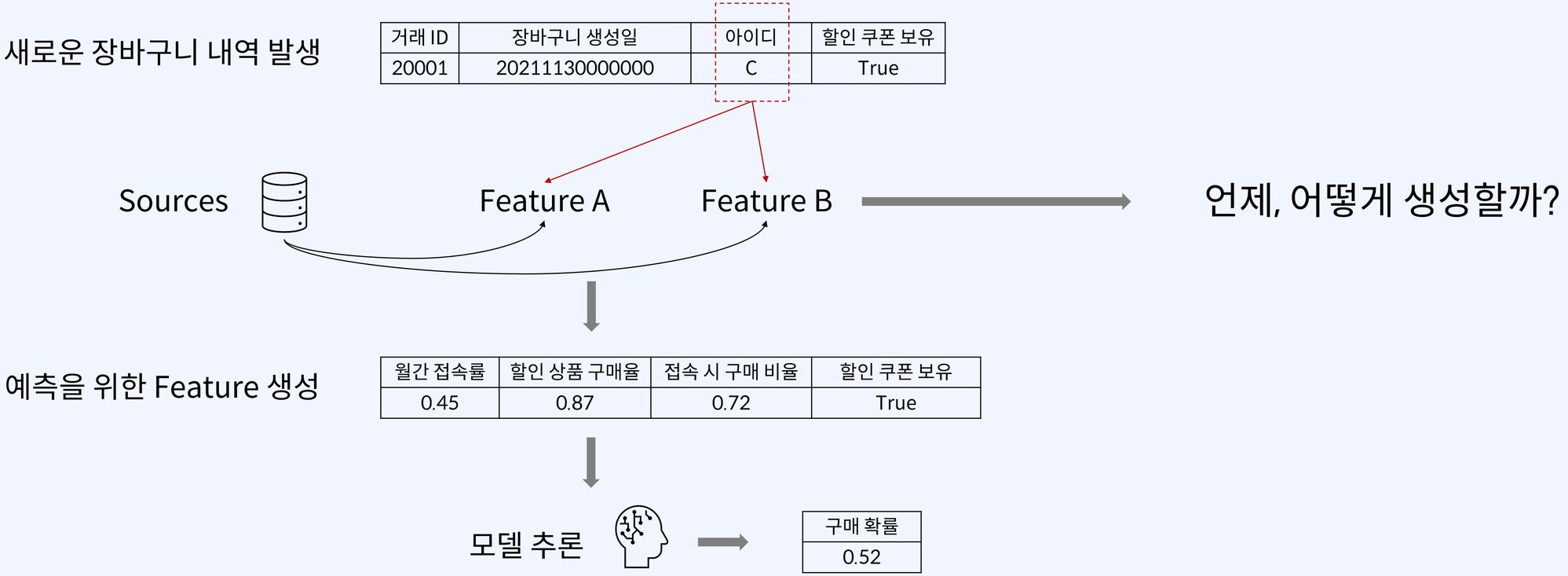
Feature Engineering (1)

Feature 들은 보통 여러 데이터들을 가공하여 만들어지게 된다.
그리고 이렇게 만들어진 Feature 들이 JOIN 등을 거쳐 학습에 필요한 데이터로 완성된다.



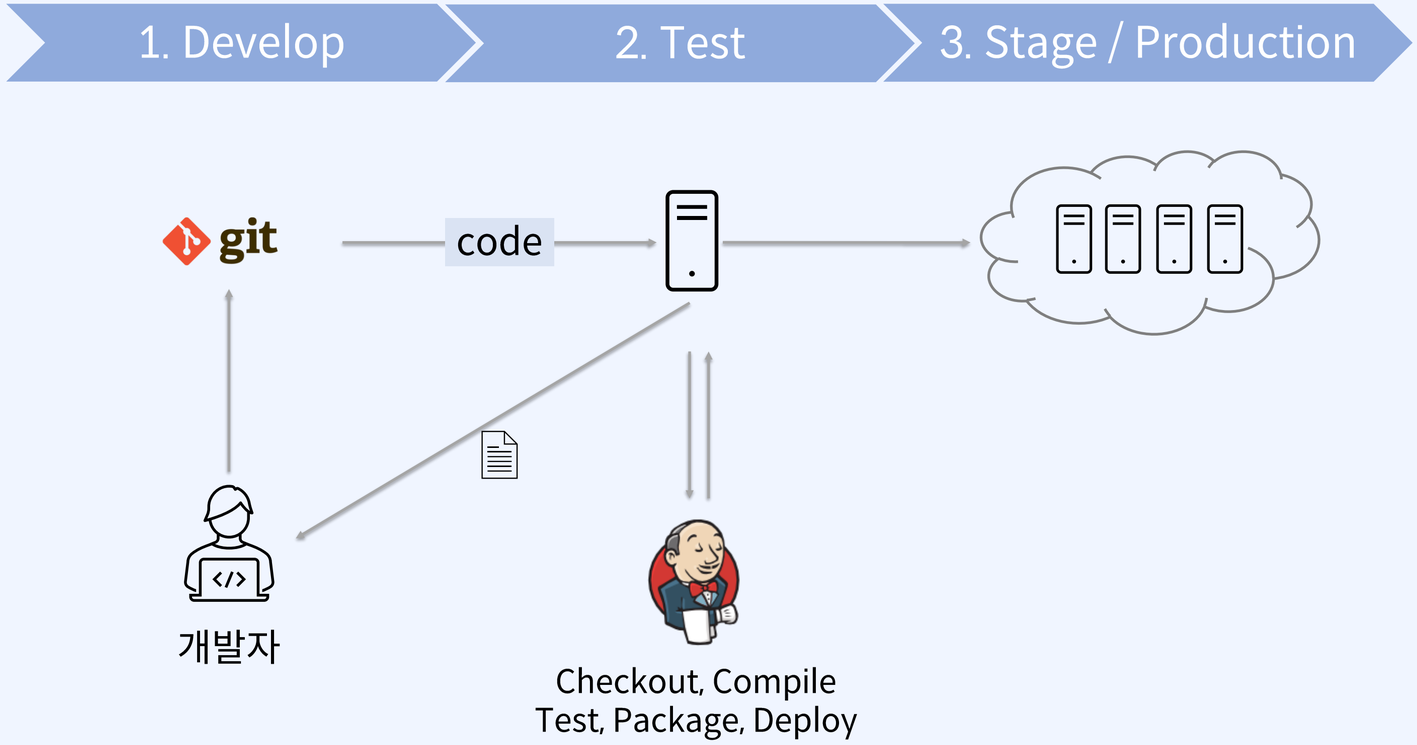
Feature Engineering (2)

이러한 모델을 가지고 있을 때 새로운 고객이 장바구니에 담은 경우
실제 구매를 할 가능성을 예측 할 수 있게 된다.
그렇다면 이러한 Feature 들은 어디서 추출해야 할까?



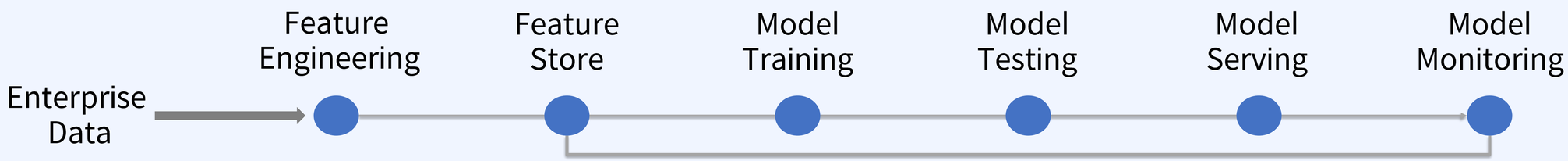
DevOps

DevOps 는 개발 과정의 자동화를 말합니다.



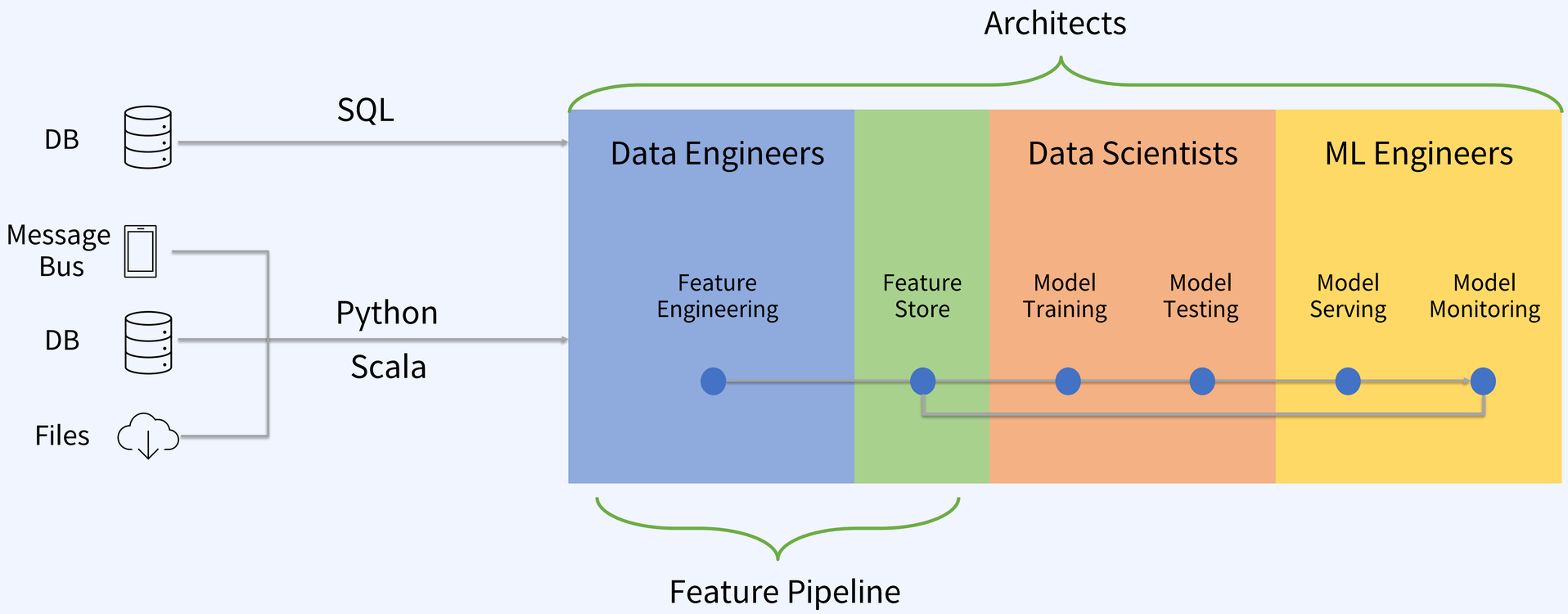
ML Pipeline 자동화

ML Pipeline 에서도 DevOps 에서의 Code 변경처럼
데이터 변경 시에 자동화된 단계를 시작해야 합니다.



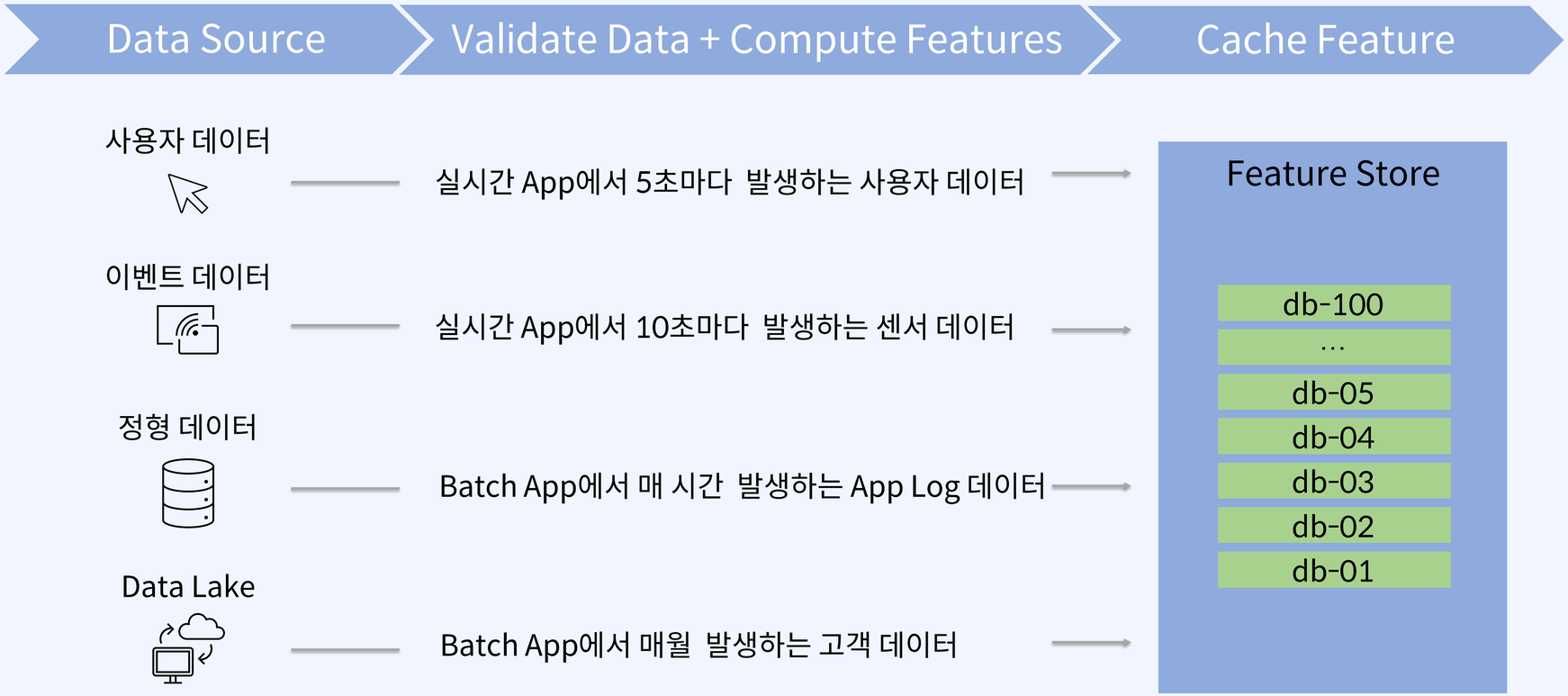
ML Pipeline 을 위한 여러 주체들

ML Pipeline 의 기능에 따라 데이터 엔지니어, 데이터 사이언티스트, 머신러닝 엔지니어 모두 각각 수행하는 역할이 다르게 됩니다.



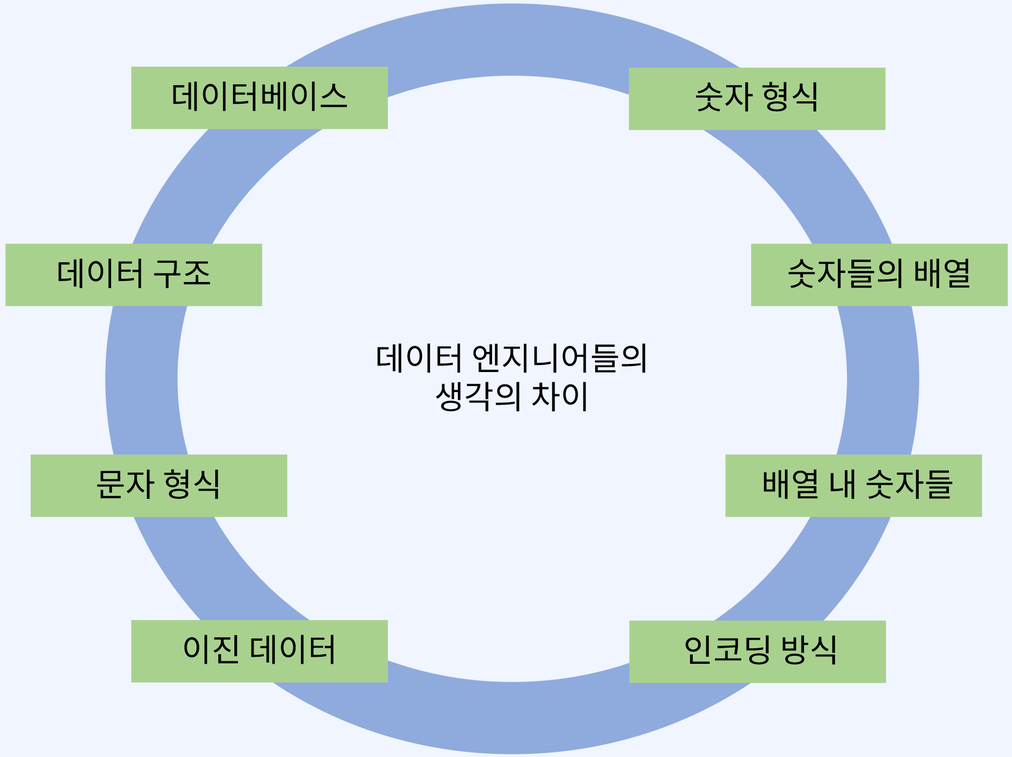
Feature Pipeline

데이터 엔지니어들이 생성하는 Feature Engineering 은
각각 다른 대기 시간으로 실행될 것 입니다.



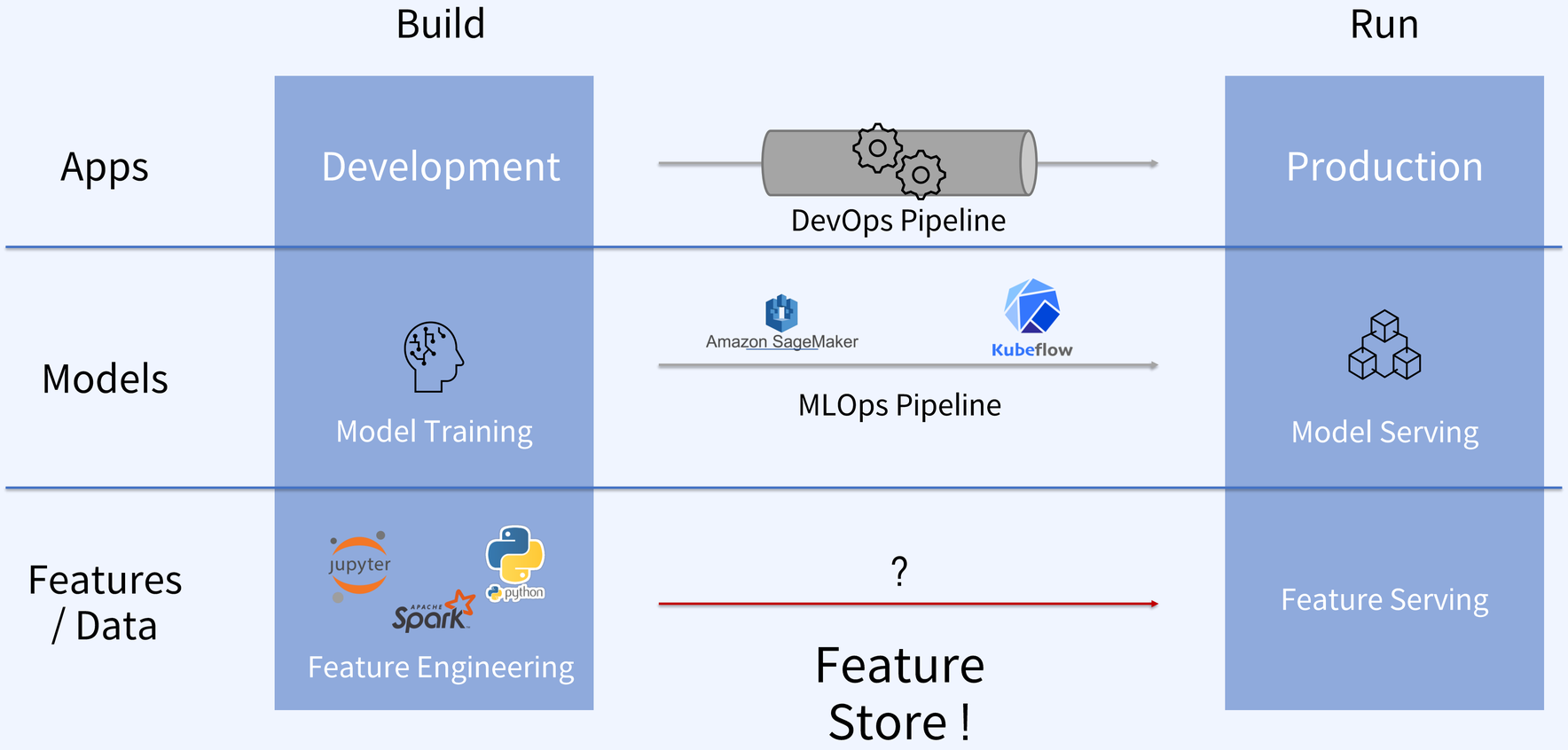
Feature Pipeline

그러나 같은 데이터를 두고도 데이터 엔지니어들은
각자 다른 관점으로 Feature Engineering 을 수행할 것이므로
일정한 Feature Pipeline 이 필요합니다.



Tools

그러나 중요성에 비해 잘 알려진 Tool 은 별로 없는 상황입니다.

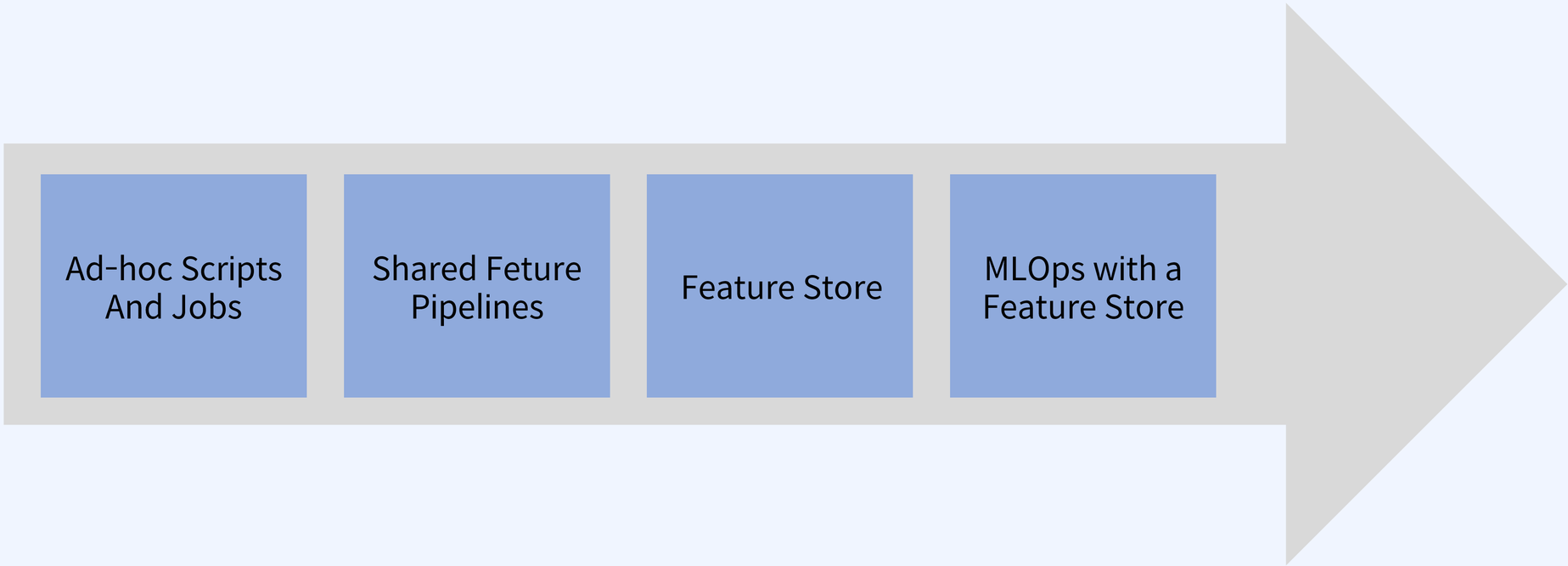


Part 04. [Chapter 2-1] Feature Store 기초

2 Feature Store 필요성

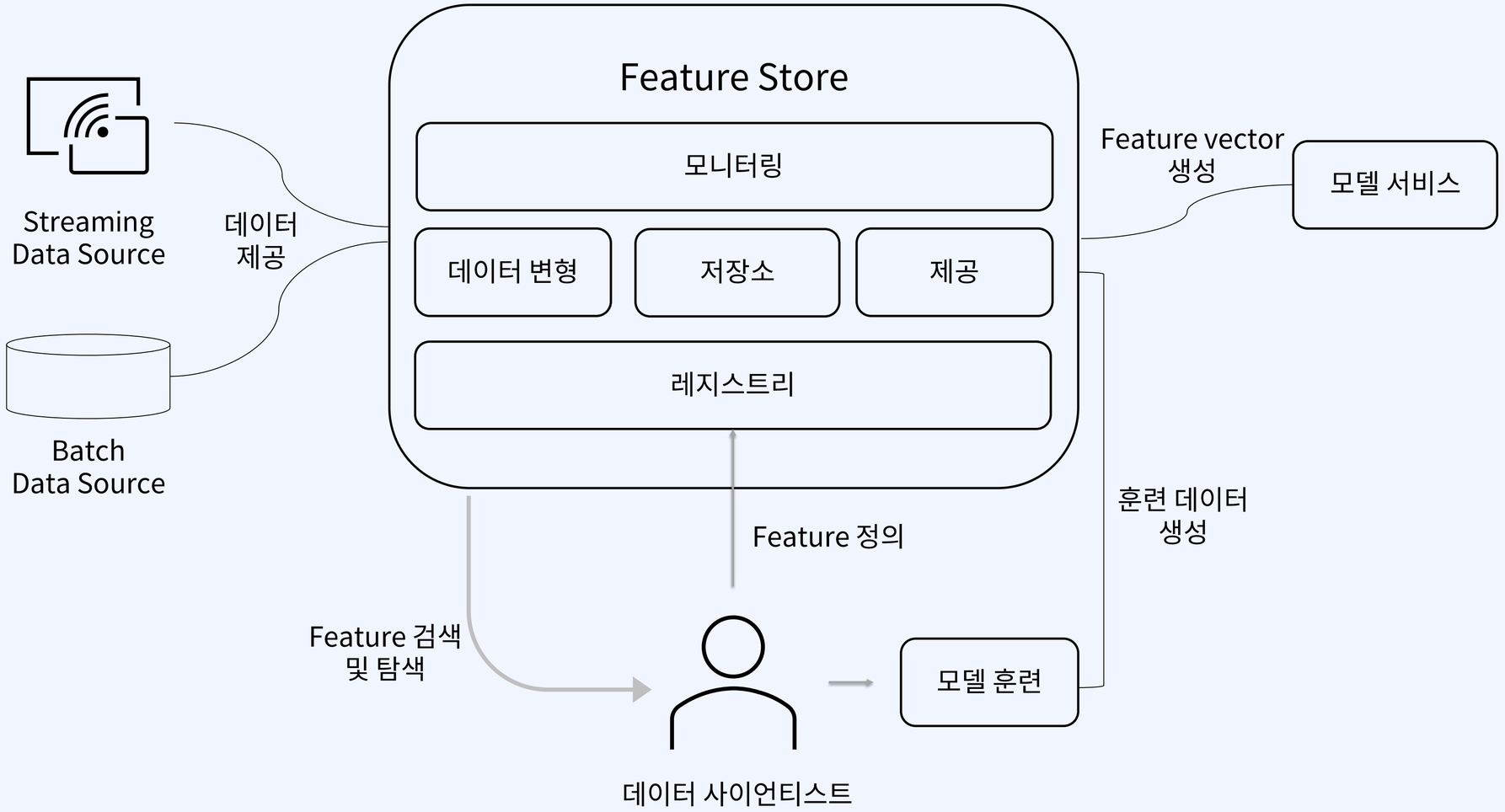
Feature Store 란?

Feature Store 는 모델 훈련과 배포 과정의 중복을 줄이기 위해 생겨난 관리형 플랫폼 입니다.



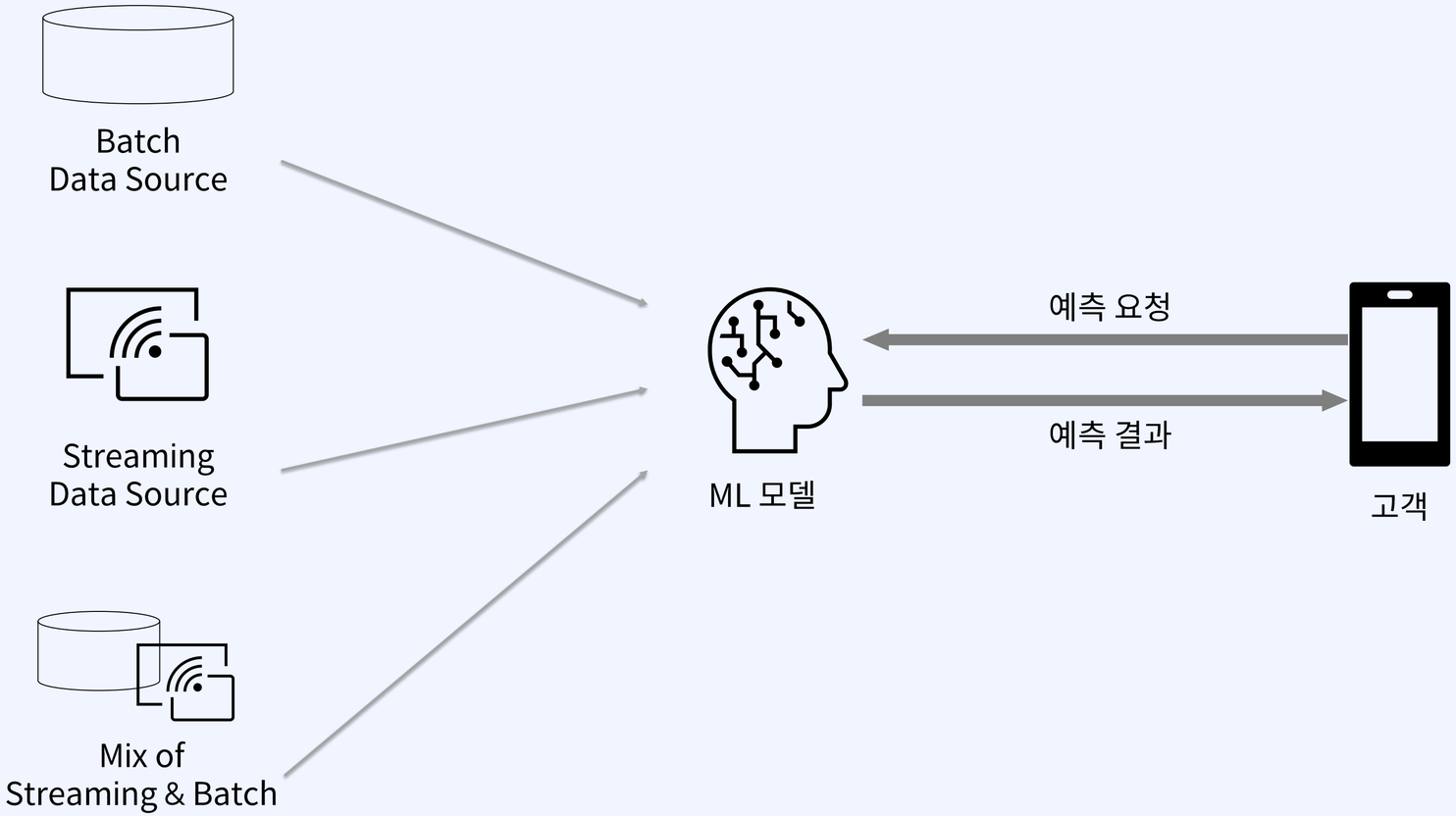
Feature Store 구성 모습

Feature Store 는 활용 모습에 따라 다양한 구성 요소를 가질 수 있습니다.



Feature 추출과 제공의 어려움

Feature 들은 매우 다양한 데이터 원천으로부터 추출된다.
이로 인해 생기는 어려움들은 무엇일까?



Feature 추출과 제공의 어려움

1) 데이터 원천의 종류에 따라 지원되는 데이터 변형이 다르다.

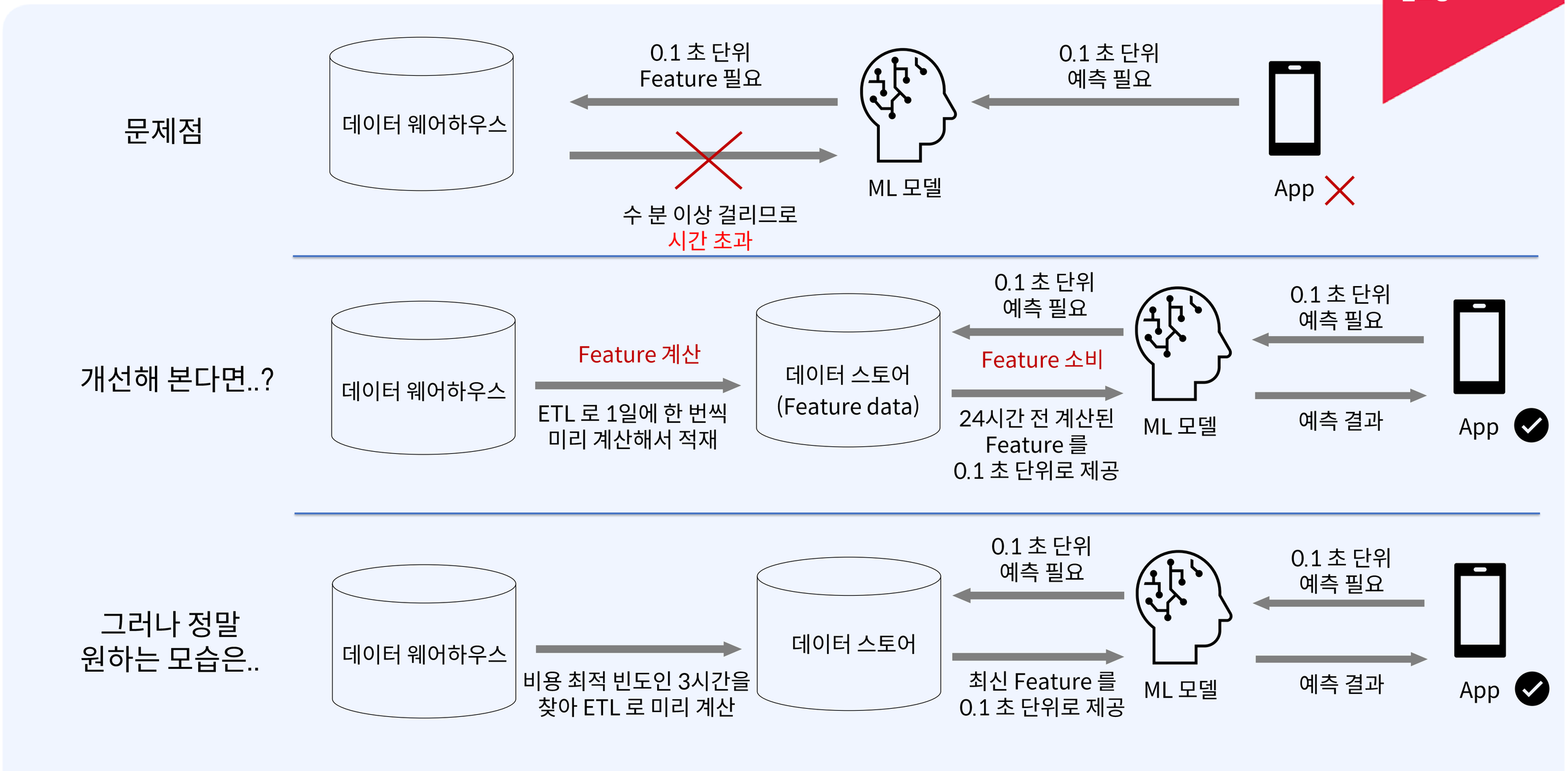
	데이터 웨어하우스 (예: Snowflake)	트랜잭션 데이터 (예: MySQL)	실시간 데이터 (예: Kafka)	예측 요청 데이터
데이터 양	전체 이력	최근 이력	실시간	현재
데이터 빈도	일/시간 단위	0.1~1초 단위	0.1~1초 단위	현재

지원되는 변형

매우 큰 크기의 Batch 단위 변형	✓			
작은 크기의 Batch 단위 변형	✓	✓		
시계열 변형	✓	✓	✓	
간단한 변형	✓	✓	✓	✓

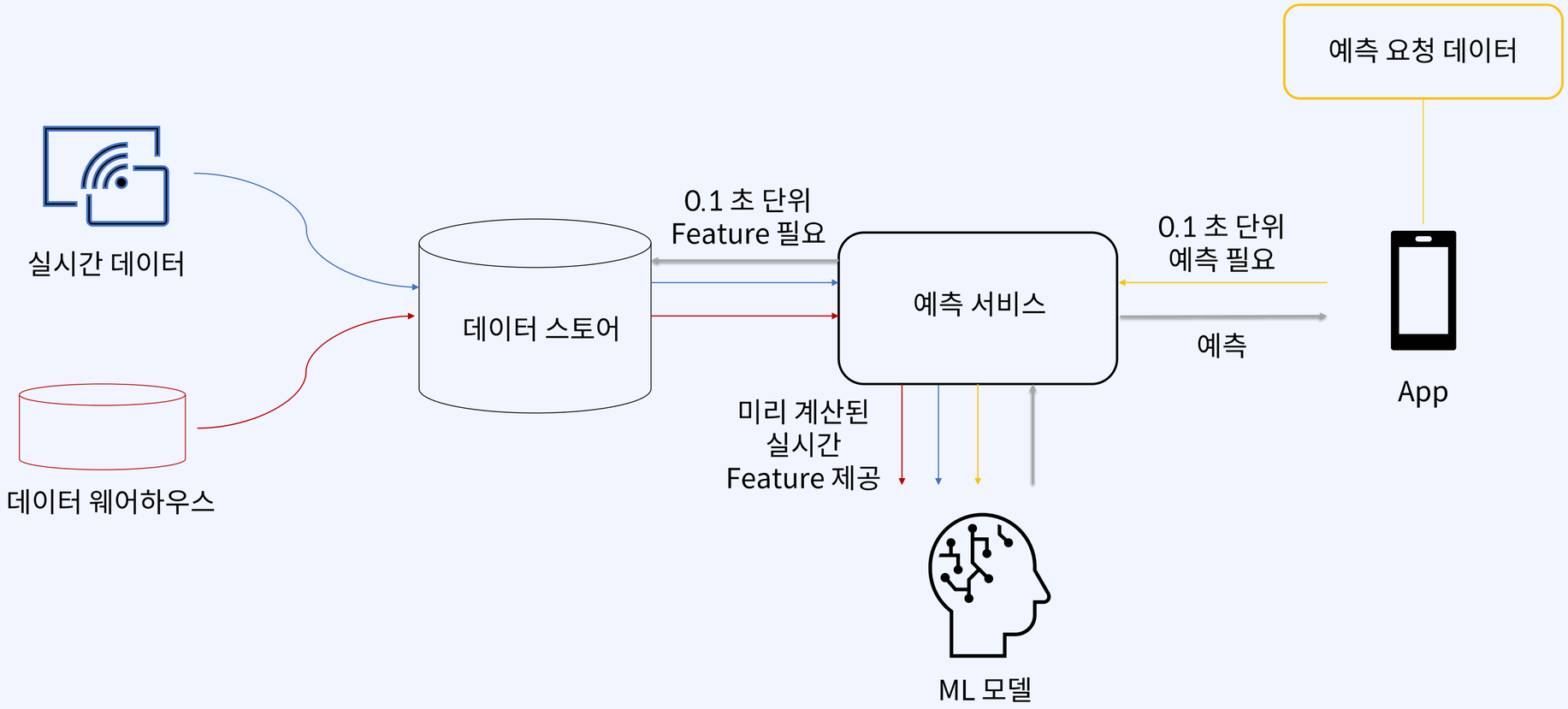
Feature 추출과 제공의 어려움

2) App 이 필요한 시간 안에 Feature 를 제공하지 못한다.



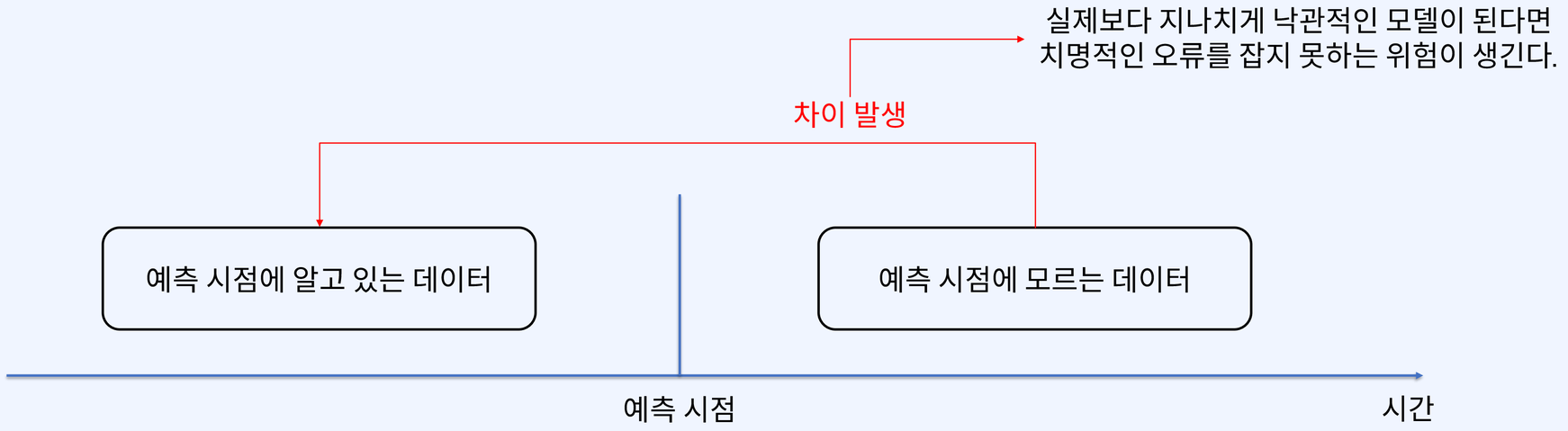
Feature 추출과 제공의 어려움

3) Batch, 실시간 데이터, 예측 요청 데이터를 한 번에 처리하는 것이 매우 어렵다.



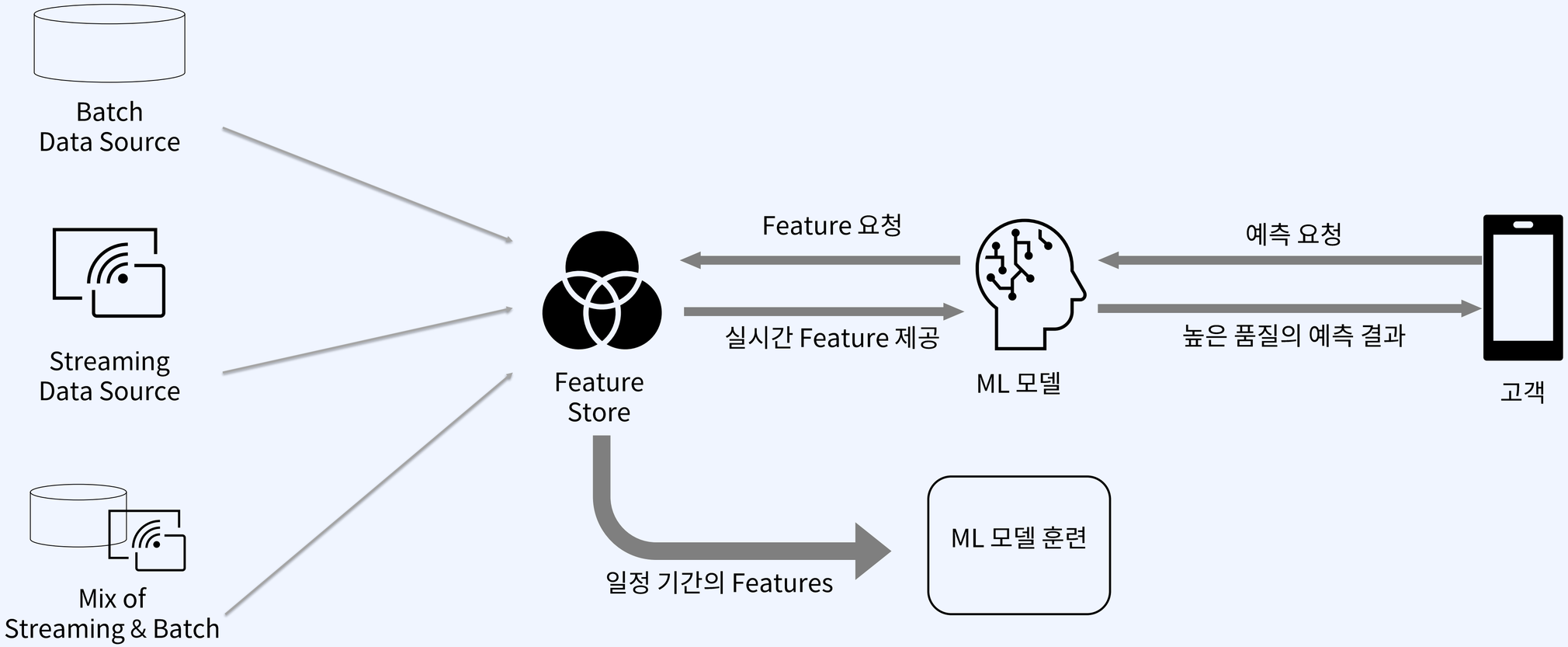
Feature 추출과 제공의 어려움

4) 데이터 누수에 의해 훈련/제공 데이터의 차이가 발생한다



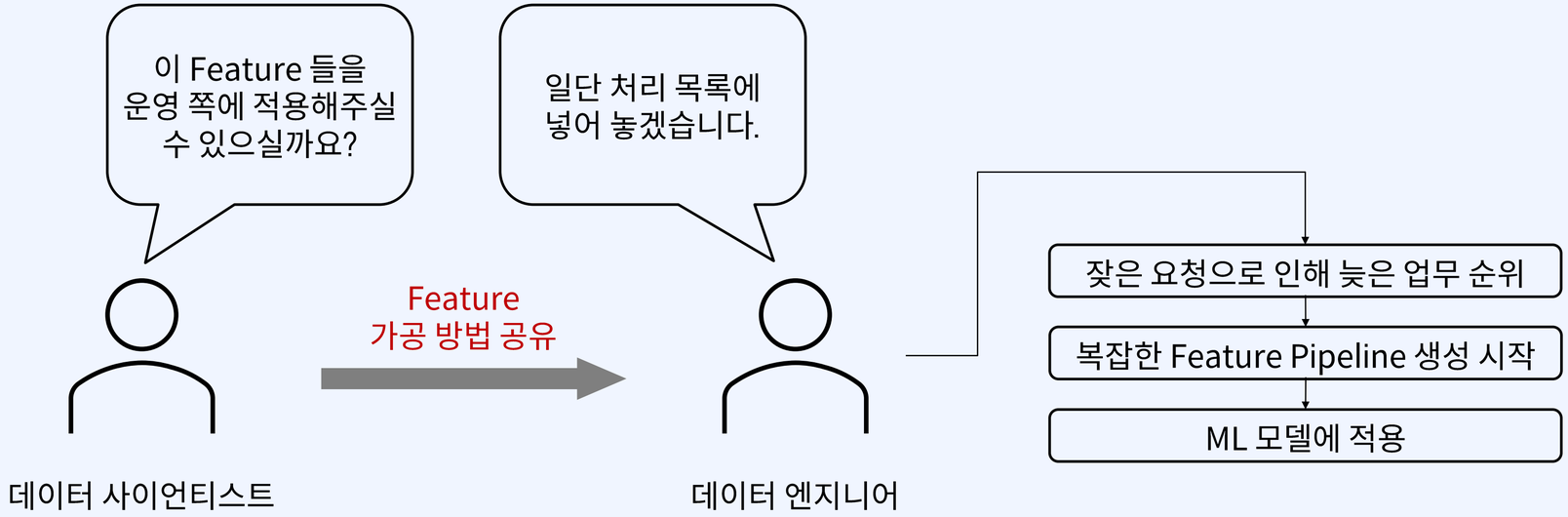
Feature 추출과 제공의 어려움

해결책) Feature Store 를 통해 필요한 모든 데이터로부터 feature 를 추출하고 제공한다



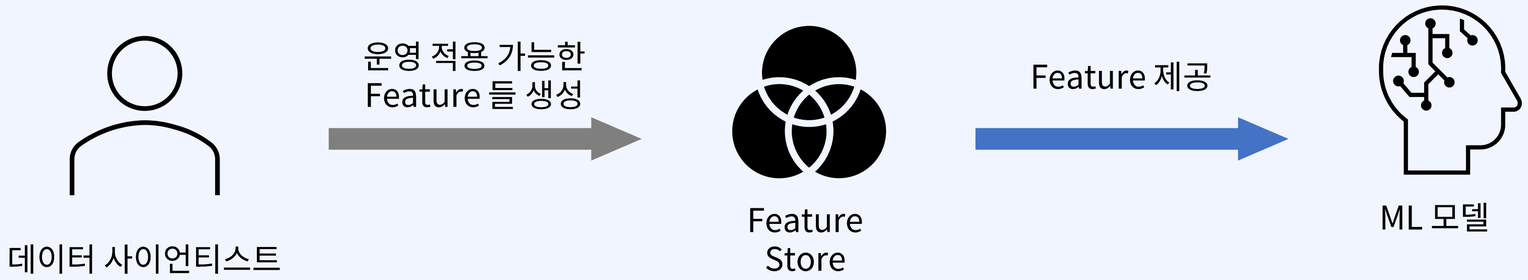
Data Pipeline 에 익숙하지 않은 ML Team

Feature 데이터 생성을 위해서 대부분의 경우 데이터 엔지니어를 필요로 한다.



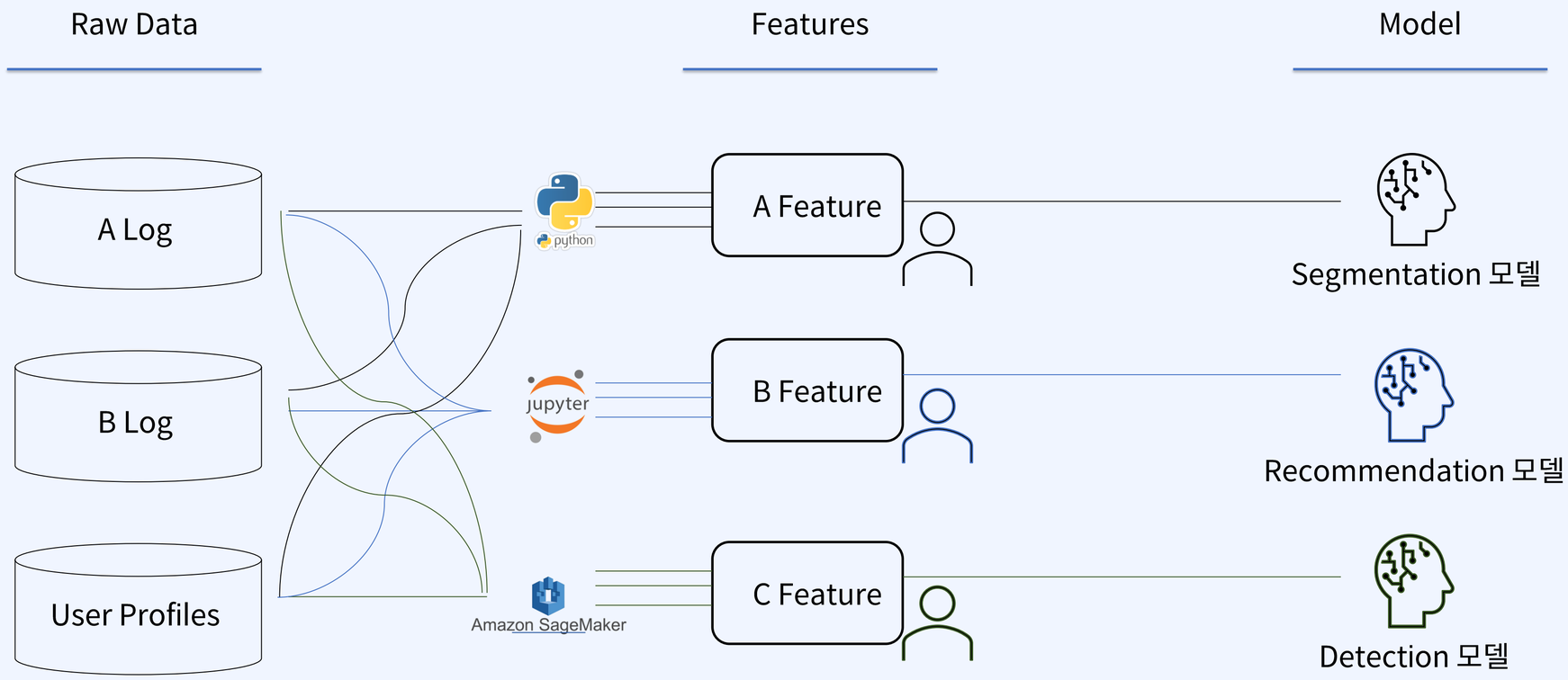
Data Pipeline 에 익숙하지 않은 ML Team

해결책) 데이터 사이언티스트가 직접 Feature Store 를 통해 운영에 필요한 Feature 들을 배포하게 된다



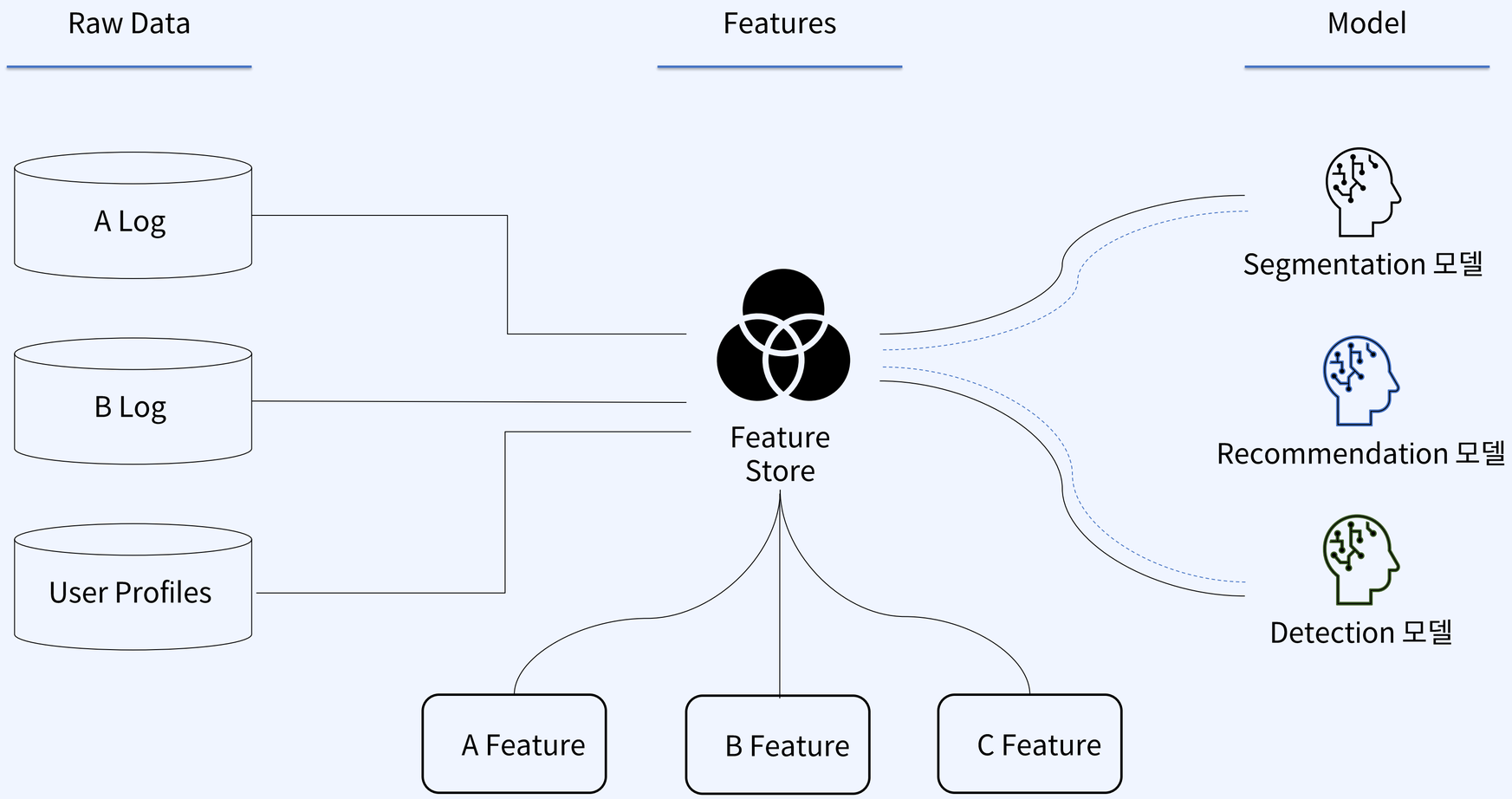
표준화되지 않은 데이터로 데이터 관리 어려움

다양한 ML모델 각각 원천 데이터를 이용해 모델 생성을 하게 되면
여러 분석 시스템에서 데이터를 중복 사용하여 관리가 어렵게 된다.



표준화되지 않은 데이터로 데이터 관리 어려움

해결책) Feature Store 를 사용하여 Feature 들을 데이터 자산의 하나로 관리 가능해진다.



데이터 문제 발생 시 운영 중인 모델 중지

데이터에 이슈가 생기면 운영에 배포된 모델 서비스가 중지되어 버린다.

- 데이터 원천으로부터의 연결이 끊기게 된다.
- 실행 가능한 Feature 들의 분포가 급격히 바뀐다.
- 데이터 품질 신뢰도가 떨어진다.

→ Feature Store 를 통해 급격한 데이터 분포 변화를 막고, 데이터 품질 신뢰를 얻을 수 있다

정리 Feature Store 를 통해 여러 문제점들을 해결할 방안이 마련될 수 있다.

문제 상황

1. Feature 의 추출과 제공의 어려움으로 인해…
 1) 데이터 원천의 종류에 따라 지원되는 데이터 변형이 다르다
 2) App 이 필요한 시간 안에 Feature 를 제공하지 못한다.
 3) Batch, 실시간 데이터, 예측 요청 데이터를 한 번에 처리하는 것이 매우 어렵다.
 4) 데이터 누수에 의해 훈련/제공 데이터의 차이가 발생한다

2. Feature 데이터 생성을 위해서 대부분의 경우 데이터 엔지니어를 필요로 한다

3. 다양한 ML모델 각각 원천 데이터를 이용해 모델 생성을 하게 되면 여러 분석 시스템에서 데이터를 중복 사용하여 관리가 어렵게 된다.

4. 데이터에 이슈가 생기면 운영에 배포된 모델 서비스가 중지되어 버린다.

해결 방안

1. Feature Store 를 통해 필요한 모든 데이터로부터 feature 를 추출하고 제공한다

2. 데이터 사이언티스트가 직접 Feature Store 를 통해 운영에 필요한 Feature 들을 배포하게 된다

3. Feature Store 를 사용하여 Feature 들을 데이터 자산의 하나로 관리 가능해진다.

4. Feature Store 를 통해 급격한 데이터 분포 변화를 막고,
데이터 품질 신뢰를 얻을 수 있다

Part 04. [Chapter 2-1] Feature Store 기초

3 Feature Store 활용 사례 알아보기

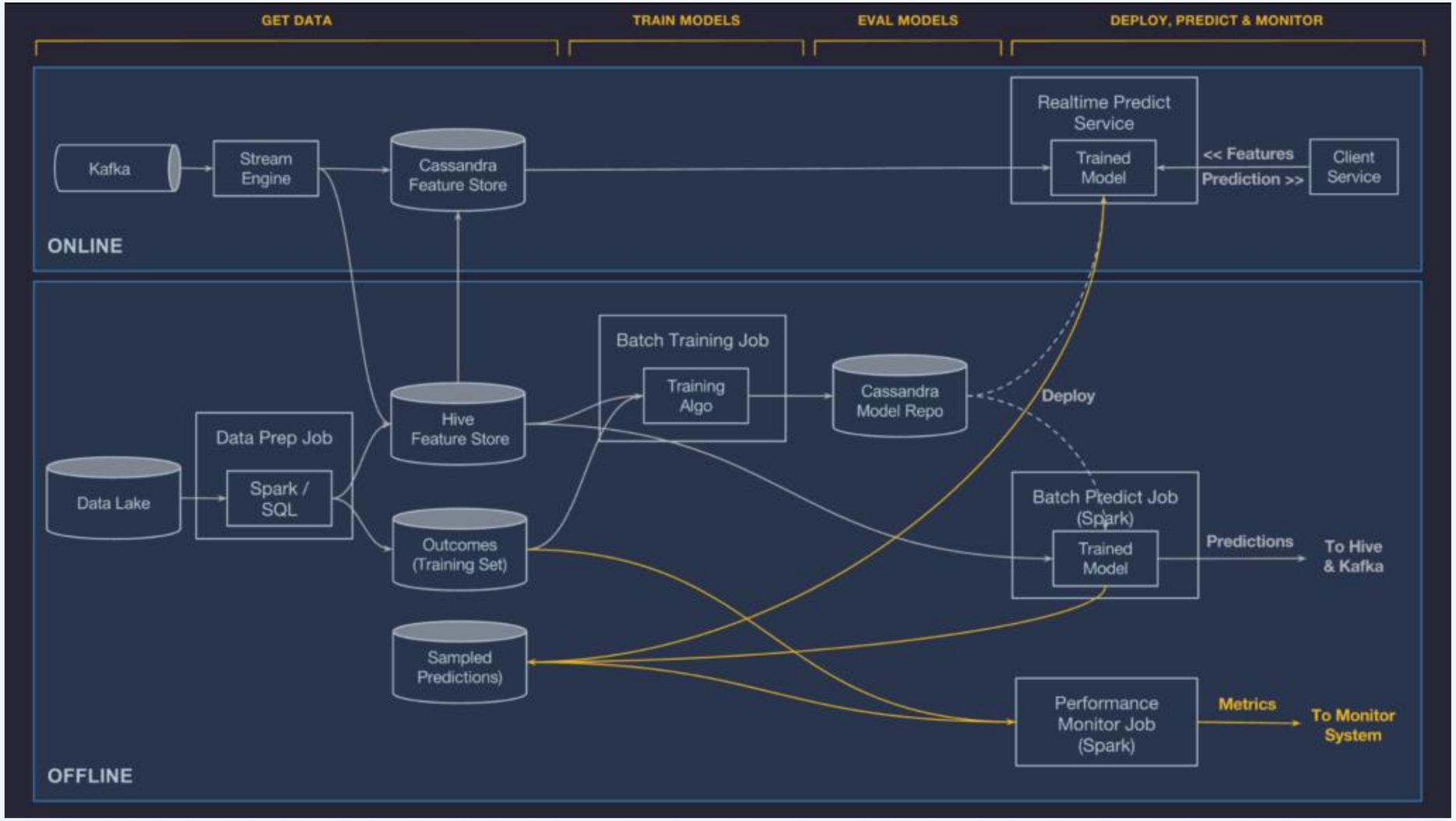
Feature Stores For ML

국내에서의 적은 관심과 다르게 해외 기업에는 많은 활용 사례가 존재한다.
Feature Stores for ML(https://www.featurestore.org/) 에서 다양한 정보를 얻을 수 있다.

이미지 출처 : Feature Store for ML

사례 1) Uber

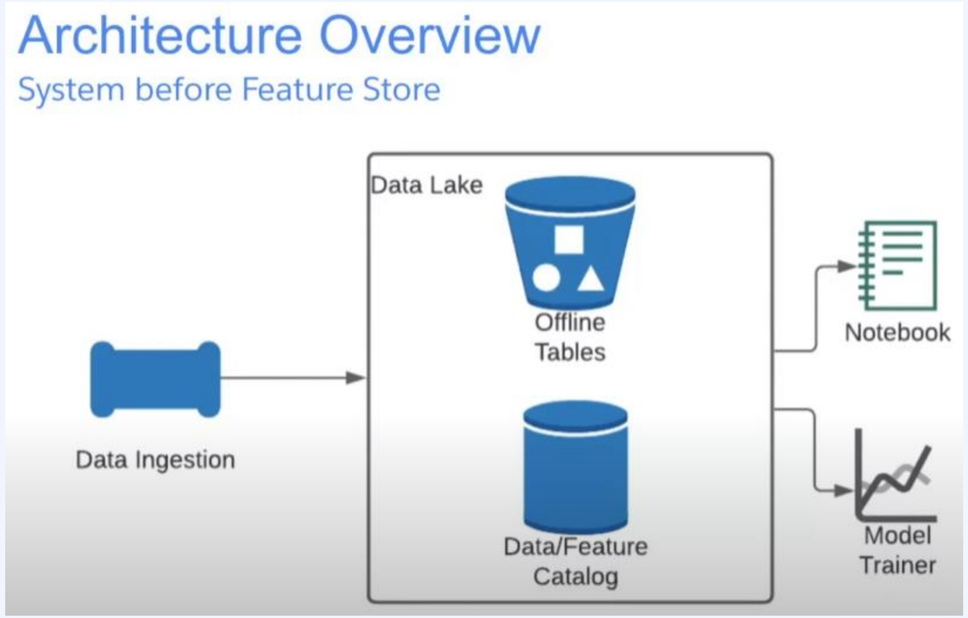
Uber Eats 의 배달 시간 예측 ML 모델 서비스

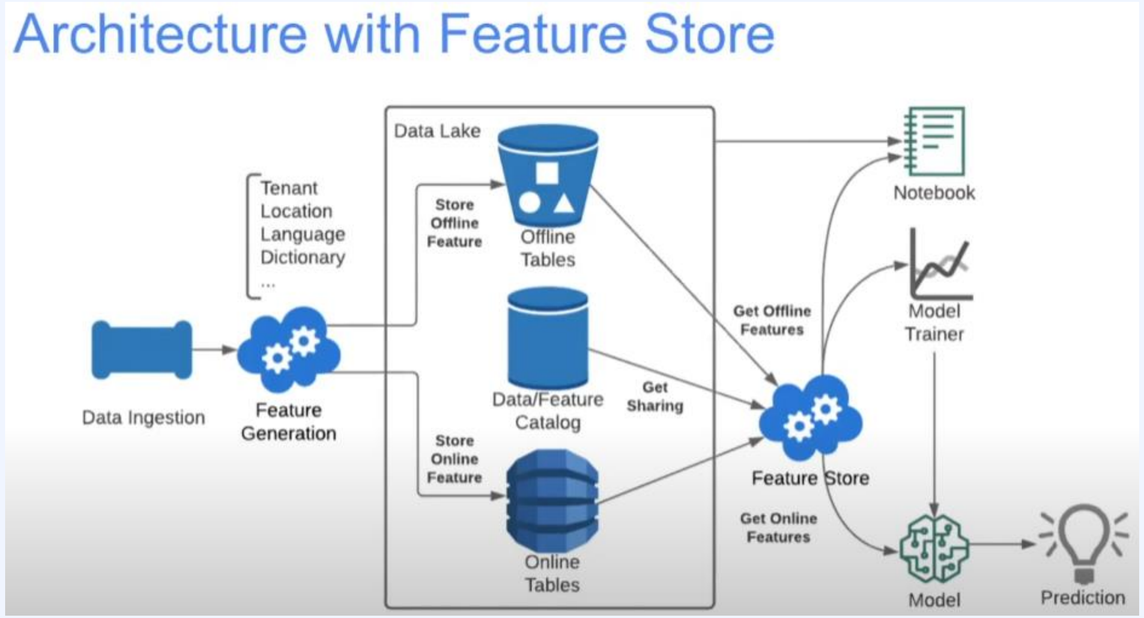


이미지 출처 : Uber blog (https://eng.uber.com/michelangelo-machine-learning-platform/)

사례 2) Salesforce

많은 수의 App 들의 ML 모델 기반 서비스

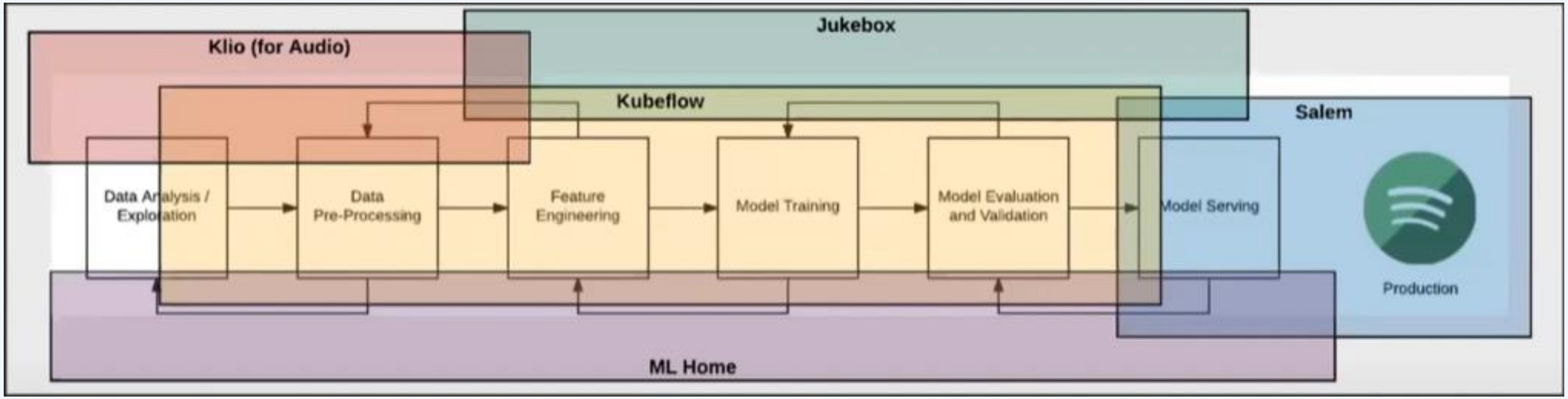




자료 출처 : Global Feature Store Meetup #7 (https://www.youtube.com/watch?v=n_V3RYFg5Zo/)

사례 3) Spotify

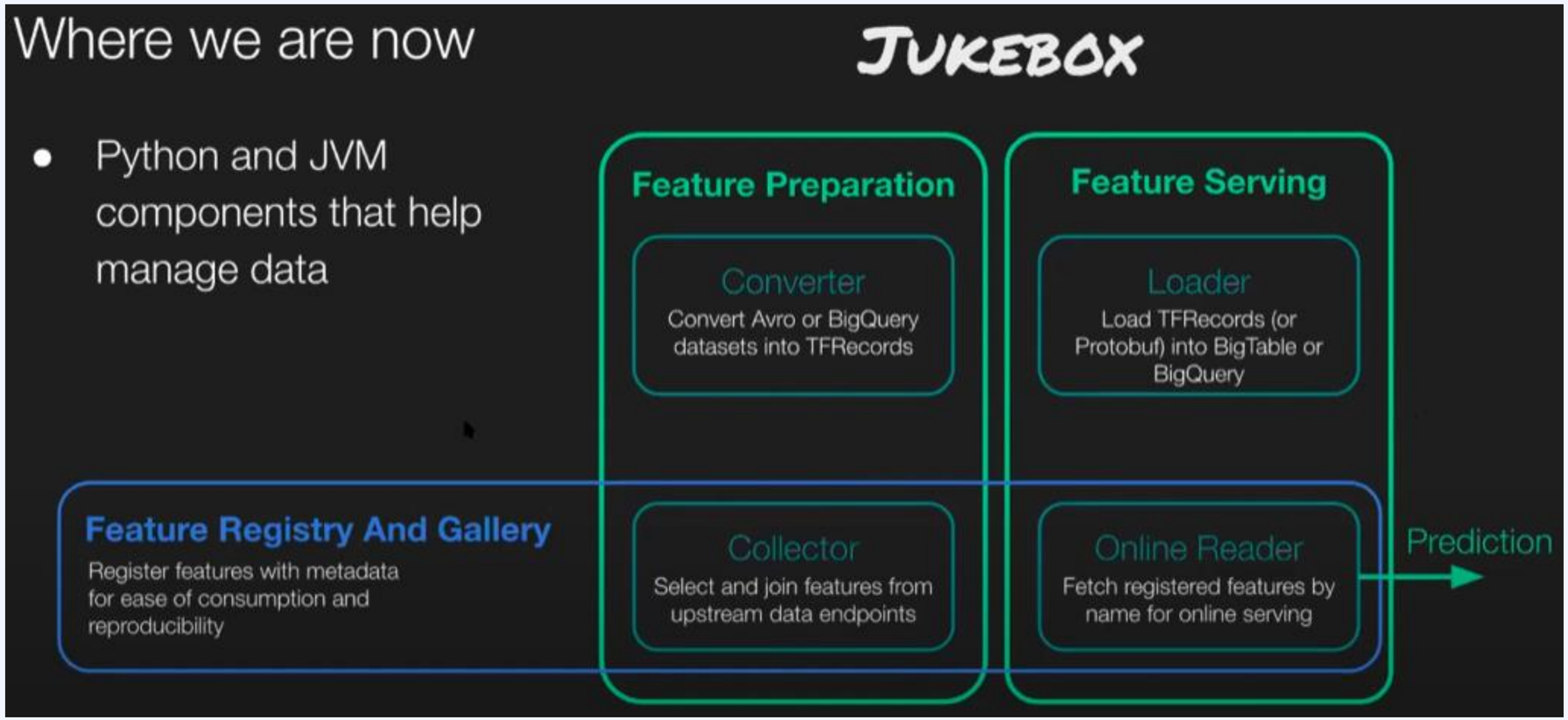
여러 제품들의 다양한 ML Lifecycle



자료 출처 : Global Feature Store Meetup #6 (https://youtu.be/2RRcOO_Nvvs)

사례 3) Spotify

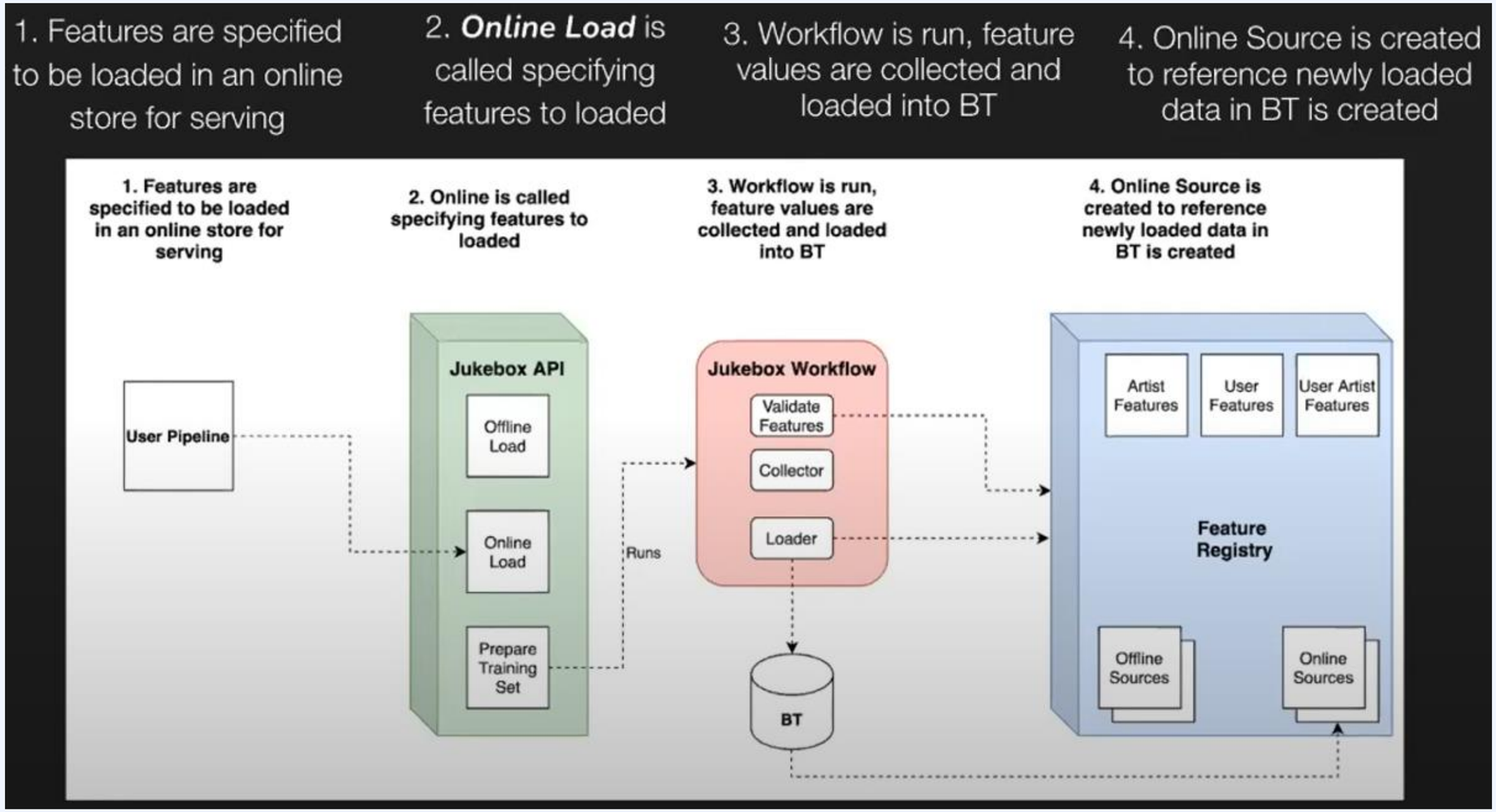
Jukebox 의 Feature Engineering



자료 출처 : Global Feature Store Meetup #6 (https://youtu.be/2RRcOO_Nvvs)

사례 3) Spotify

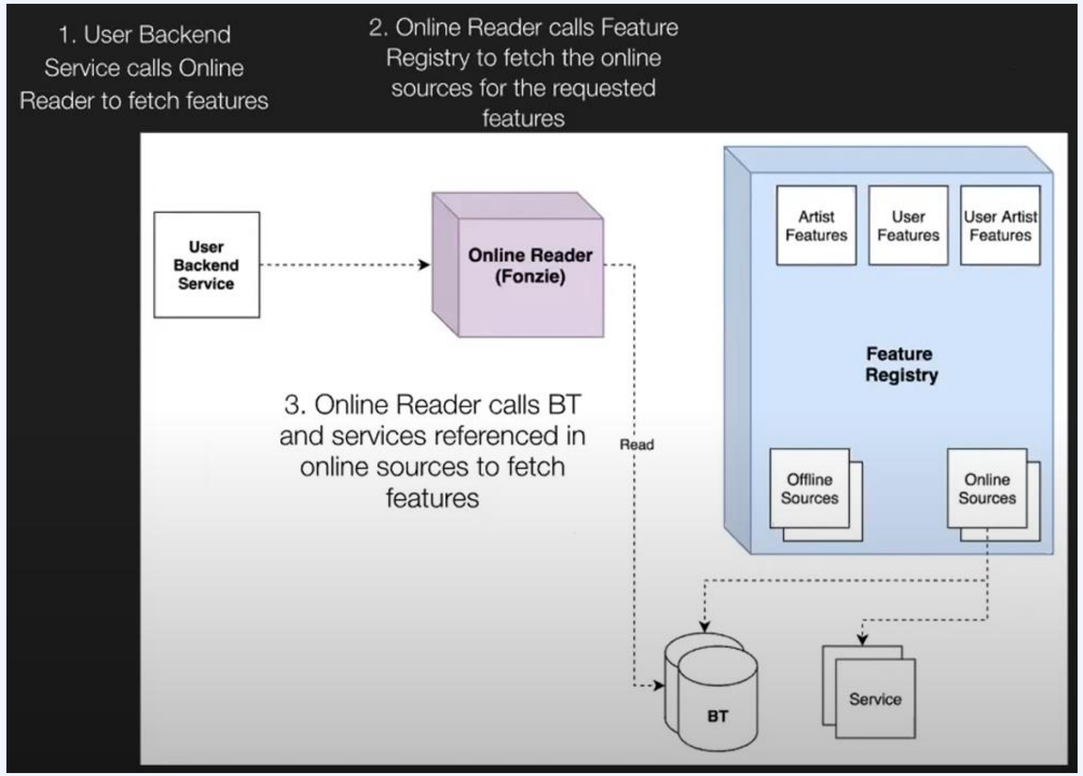
Online Load Workflow

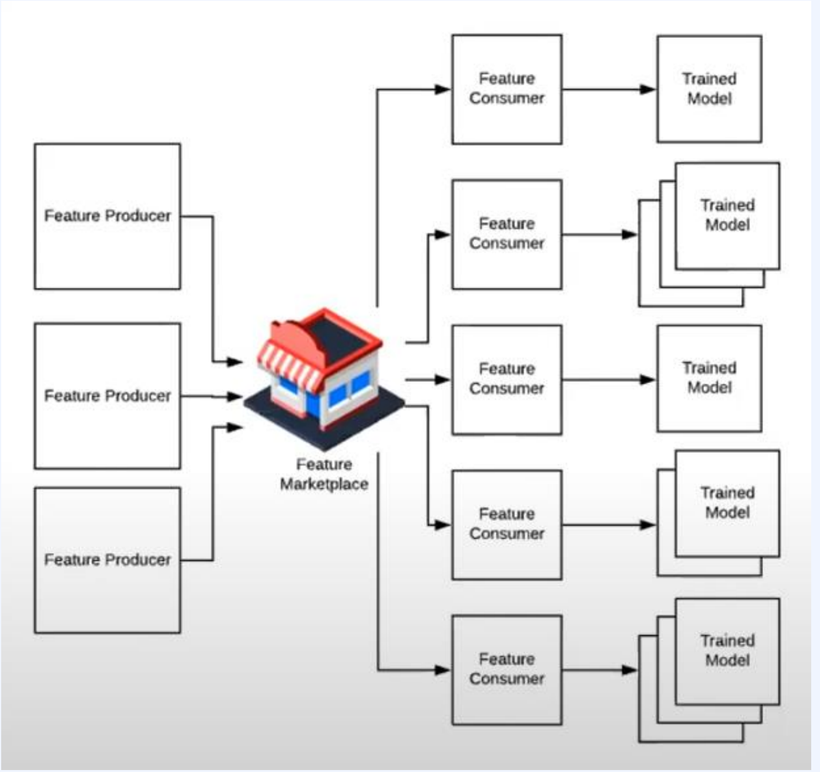


자료 출처 : Global Feature Store Meetup #6 (https://youtu.be/2RRcOO_Nvvs)

사례 3) Spotify

Online Serving 과 향후 계획





자료 출처 : Global Feature Store Meetup #6 (https://youtu.be/2RRcOO_Nvvs)