Data engineering

Contents

- problems
- goal
- role
- skill
- data pipeline
- tool

4 petabytes / day

60 petabytes / day

12 petabytes / day
3조 개 이벤트 / day

1.5 billion people are active on Facebook **daily**

Europe has more than 307 million people on Facebook

There are five new Facebook profiles created every second

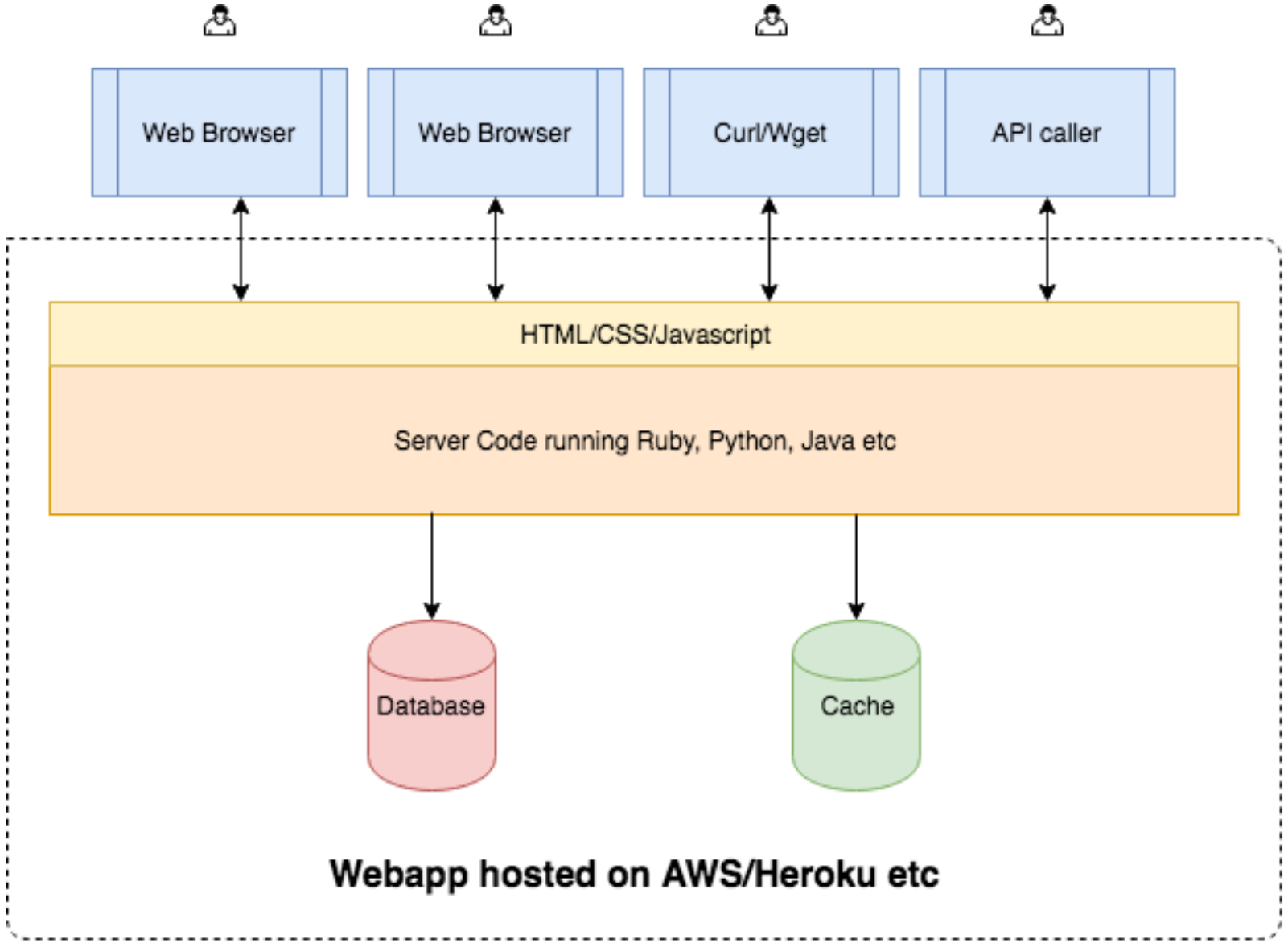
More than 300 million photos get uploaded per day

Every minute there are 510,000 comments posted and 293,000 statuses updated

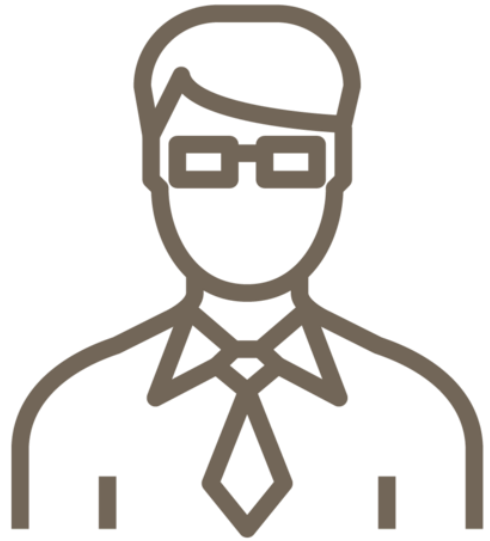
There are 600 million Instagrammers; 400 million who are active every day

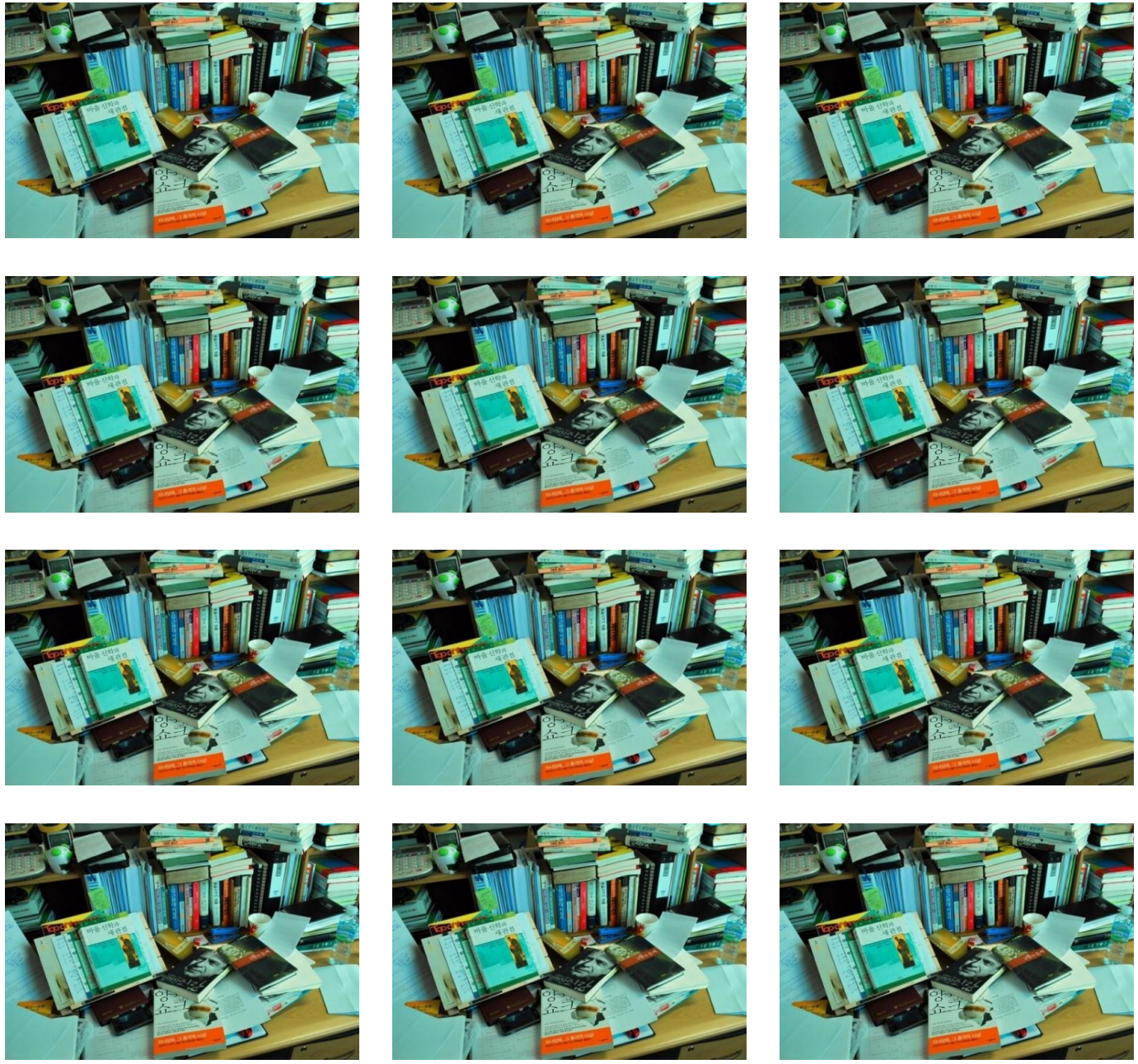
Each day 95 million photos and videos are shared on Instagram

100 million people use the Instagram “stories” feature daily

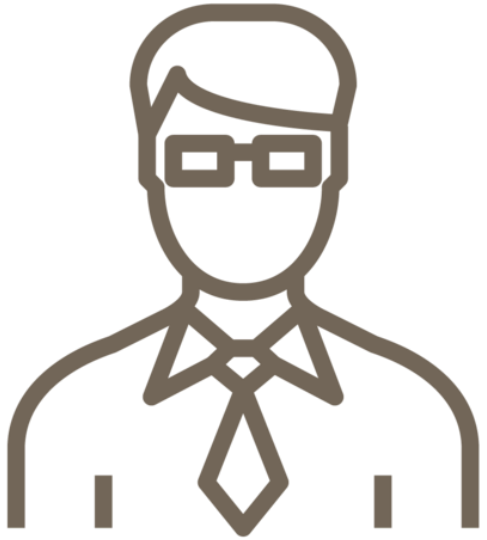

Web Browser
Web Browser
Curl/Wget
API caller
HTML/CSS/Javascript
Server Code running Ruby, Python, Java etc
Database
Cache
Webapp hosted on AWS/Heroku etc

한달전에 준 보고서
좀 찾아주세요





가지고 있는 책 들에서
어떤 단어가 가장 많이
등장하나요?



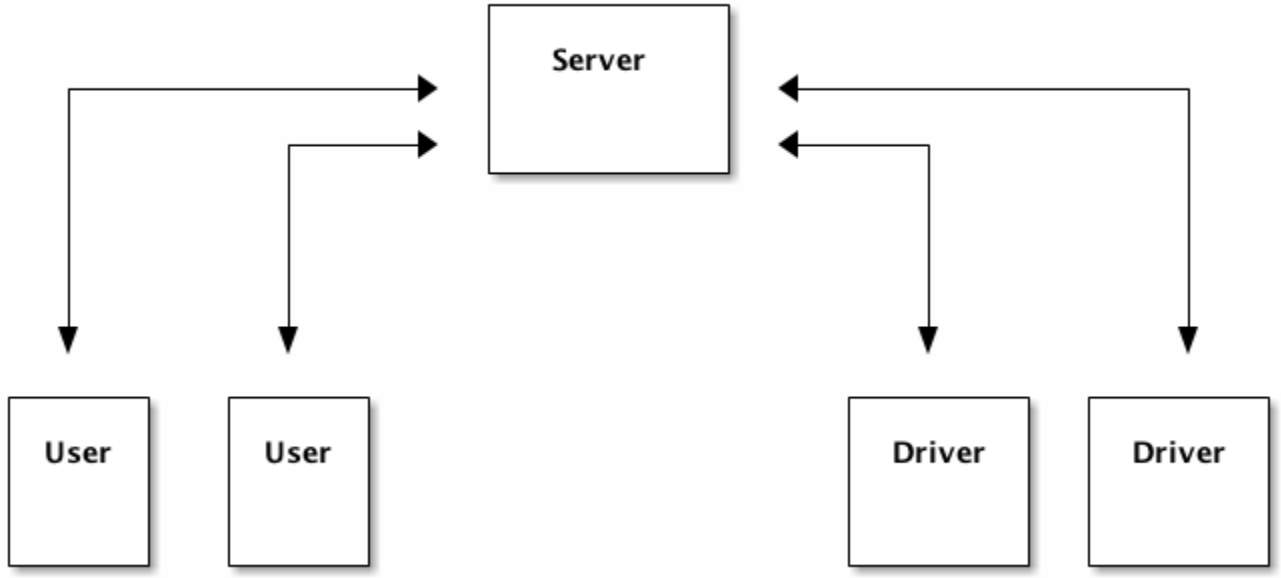
Goals of a Data Engineer

- data pipeline에 대한 이해 및 개발
- 분석가 및 데이터 과학자를 위한 데이터 및 테이블 관리
- 서비스/제품 연계를 염두하여 설계

Uber의 예를 들어봅시다

Uber와 같은 서비스에서 필요한 것

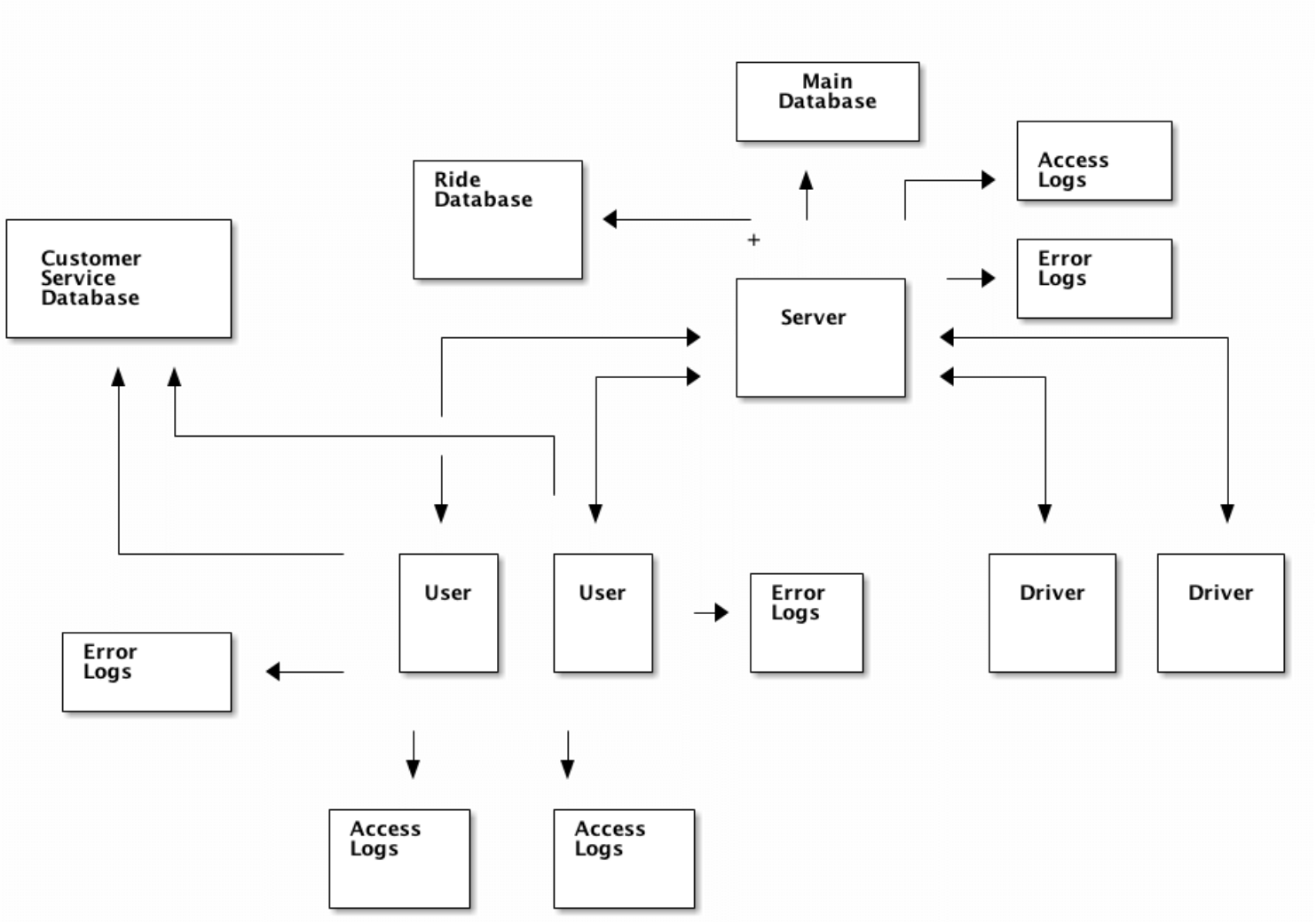
- 유저를 위한 모바일 앱
- 드라이버를 위한 모바일 앱
- 요청을 전달 및 처리하고 결제정보 업데이트 같은 세부 정보 처리 서버



Uber의 예를 들어봅시다

필요한 데이터 저장소

- 메인 DB(사용자, 드라이버 정보)
- 서버 분석 로그 / 서버 엑세스 로그
  (요청 당 한 줄 씩 추가)
- 서버 에러 로그(서버 상의 에러)
- 앱 분석 로그
- 앱 이벤트 로그
  (버튼클릭, 결제 정보 업데이트)
- 앱 에러로그(앱 상의 에러)
- 고객 서비스 DB
  (고객 서비스 에이전트에서의 고객과
  상호작용 정보: 이메일, 음성녹음)



Uber의 Data Engineer들은

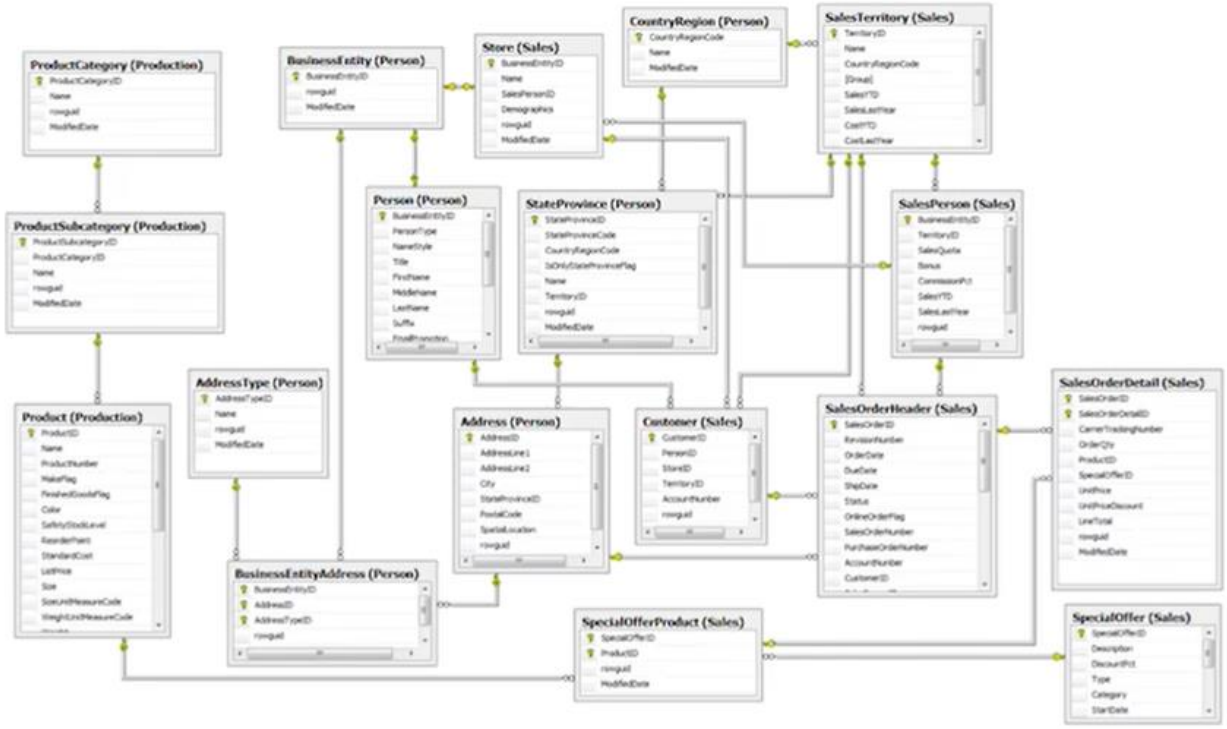
- 모바일 앱 / 서버 로그 실시간 수집
- 구문 분석한 후 사용자에게 연결할 파이프라인 설계 및 구축
- 구문 분석한 로그 DB 저장 및 쿼리 시스템 구축
- 로드 벨런싱 통한 분산 병렬 처리 시스템 구축
- 오토 스케일링 시스템 인프라 작업

Skill

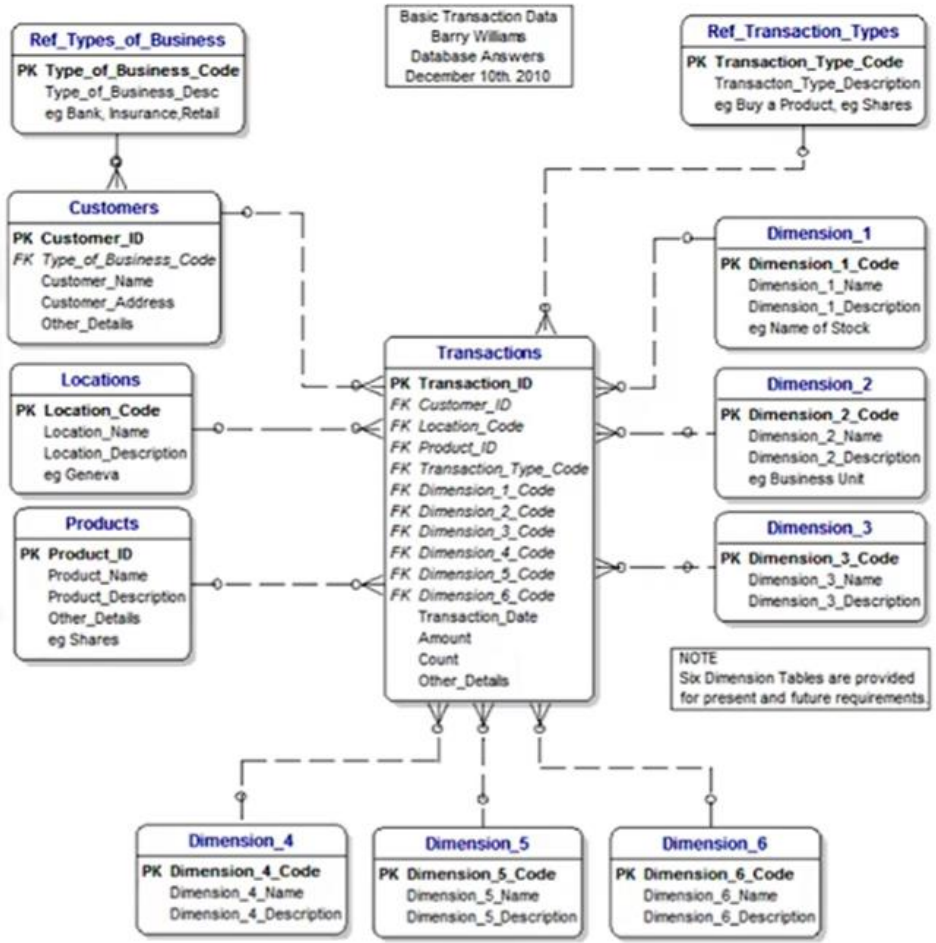
- hard skill
  - data modeling
  - Automation
  - ETL development(reliable data pipeline development)
  - Combining data sources
  - Architecting distributed systems
  - Architecting data stores
  - infra-structure
- soft skill
  - Product understanding
  - Collaborating with data science teams
  - building the right solutions for them

data modeling

How Do We Turn This

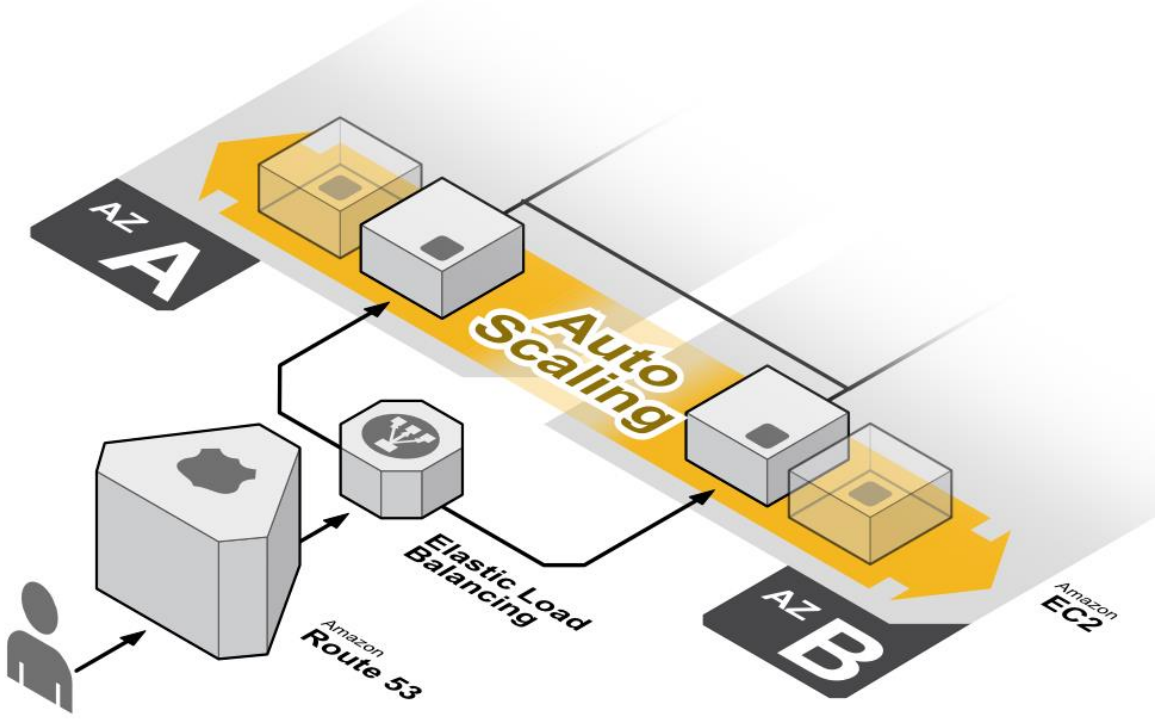


Into This

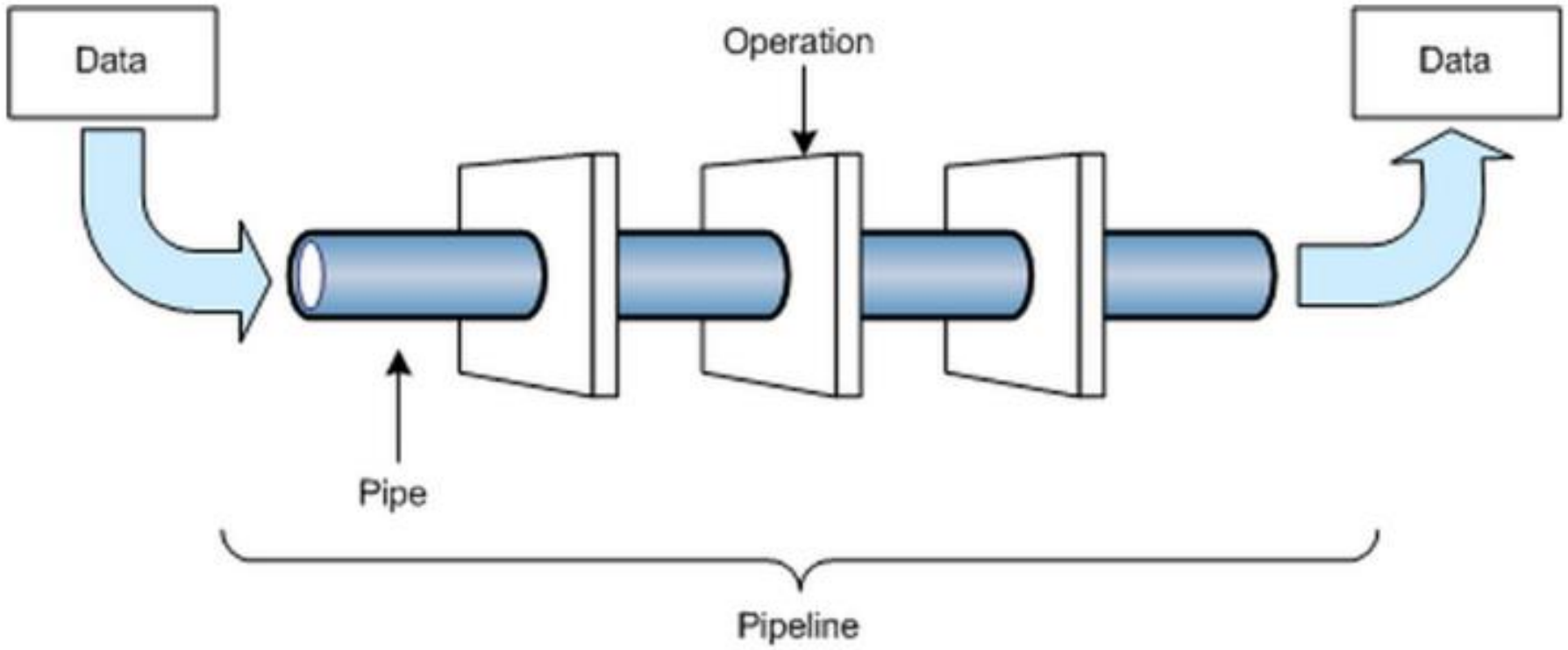


Automation

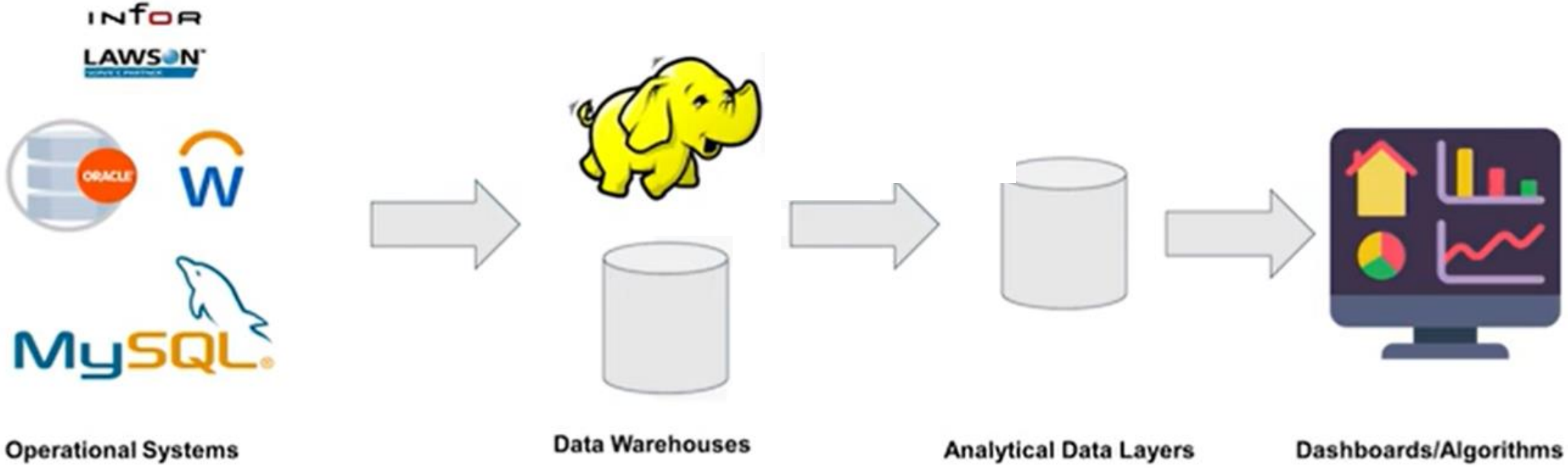
- workflow managing 자동화
- Dependencies managing 자동화
- 장애 시 fail-back / fail-over 자동화
- Auto Scaling



ETL pipeline development



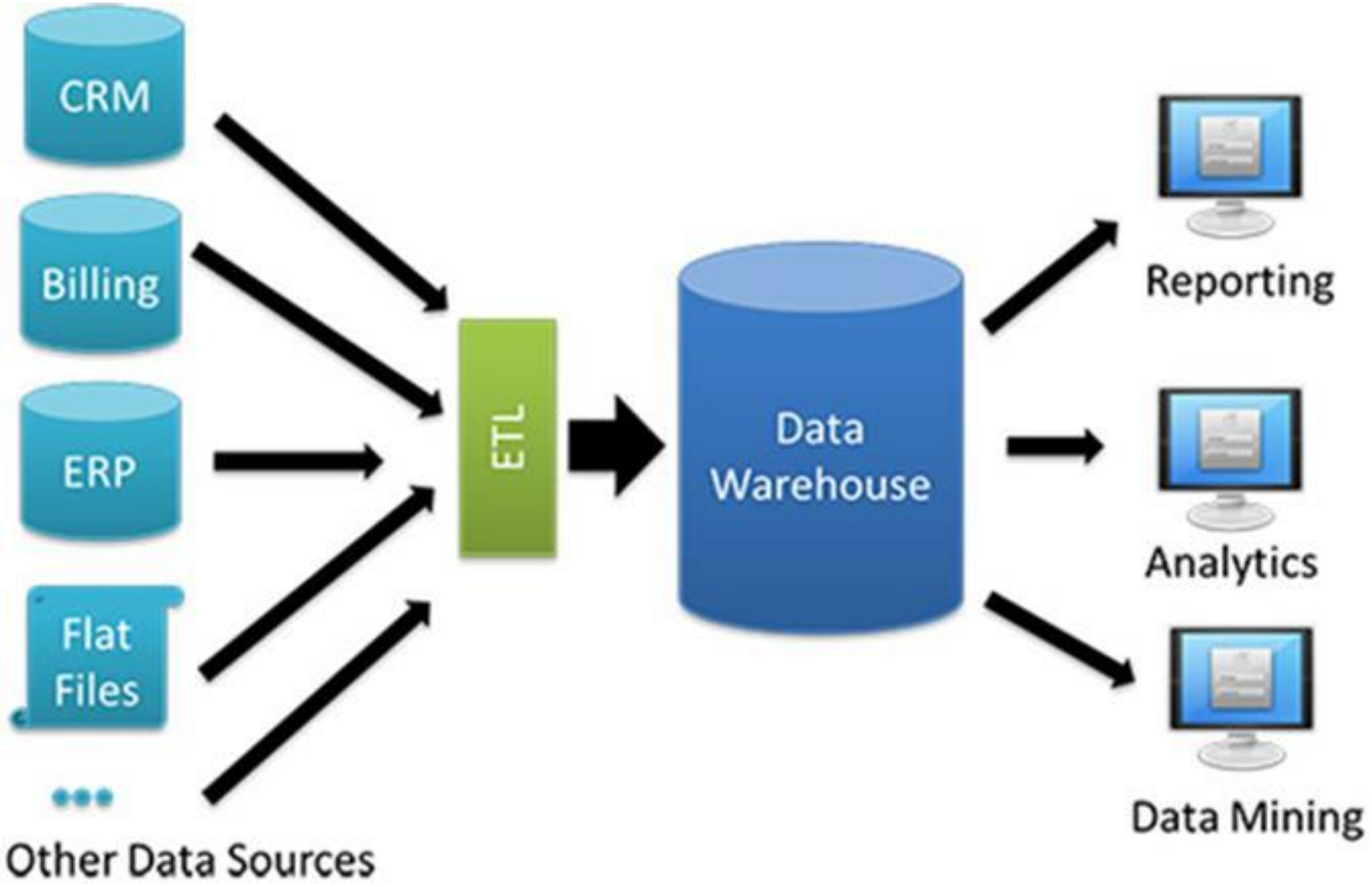
ETL pipeline development



Tools for Data Engineer

- pipeline framework(apache airflow, kubeflow pipelines)
- Distributed data Storage(hadoop, ceph)
- Distributed data processing engine(spark)
- 대쉬보드 형식의 분석 레이어

combine data sources and architecting

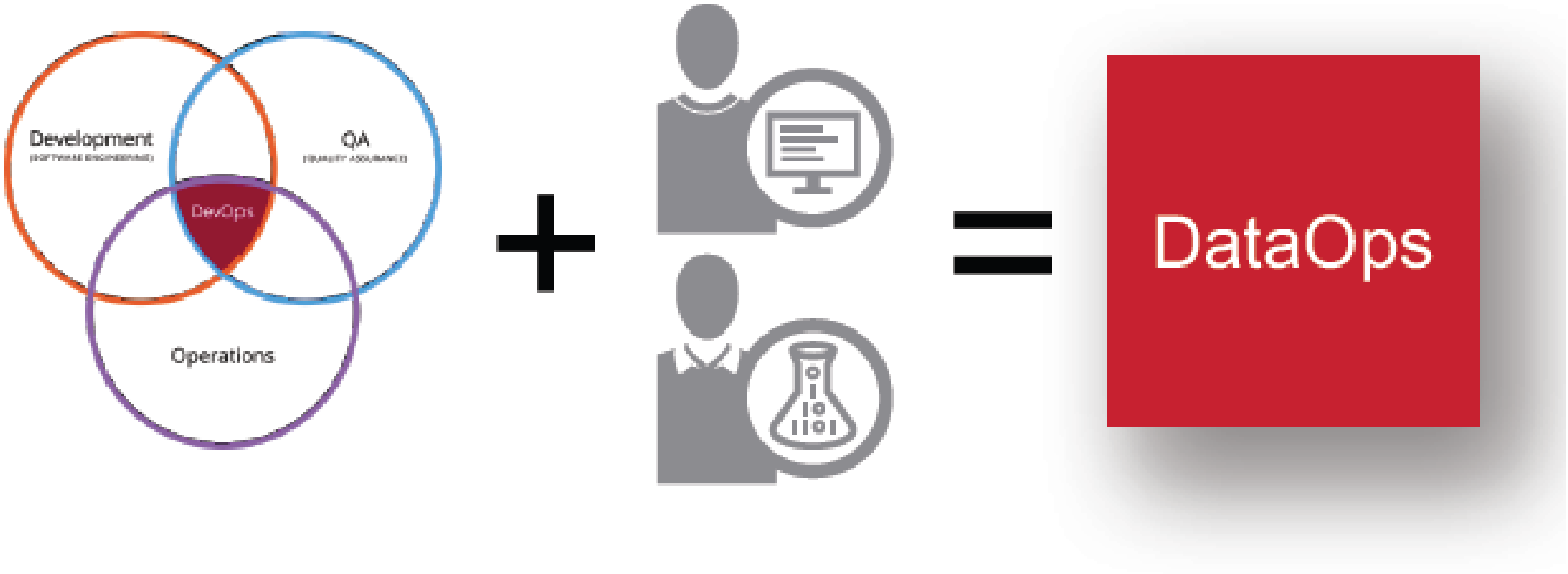


Product understanding

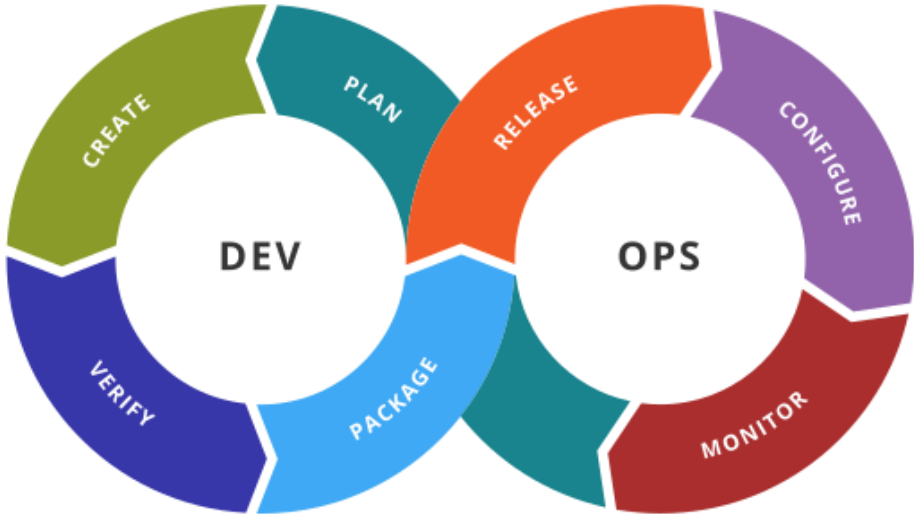
- 해결하고자 하는 문제는 무엇인가?
- 가용할 수 있는 자원은 무엇이 있는가?
- 얻을 수 있는 데이터는 어떤 것들이 있는가?
- 각 데이터는 어떤 특성이 있는가?
- 해당 도메인과 문제와 데이터에 맞는 아키텍처는 무엇인가?
- 엔드 유저에게 제공하는 솔루션의 형태는 무엇인가?

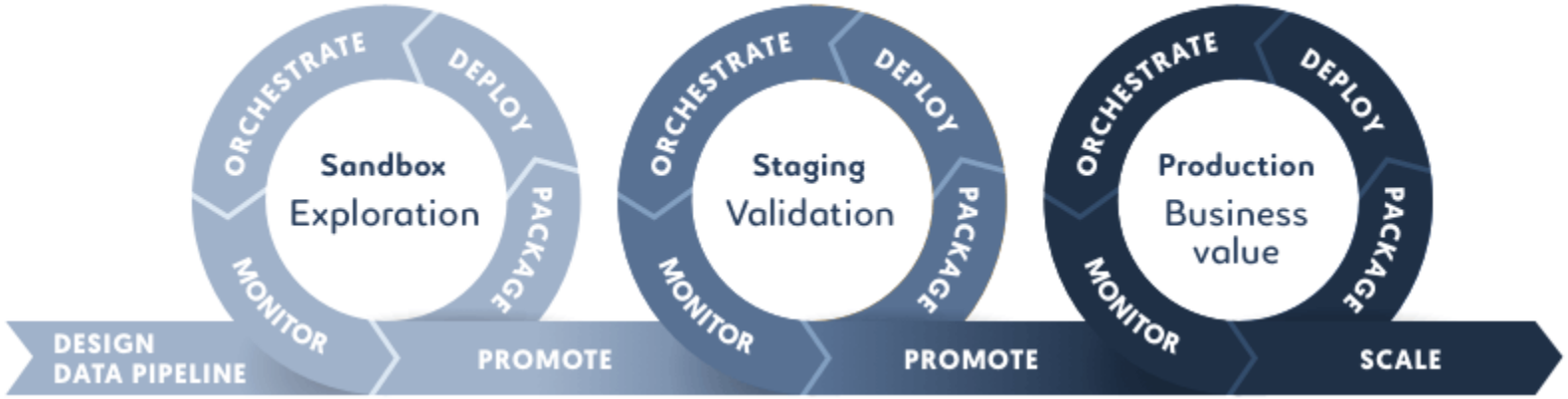
Collaborating with data science team

Collaborating with data science team



Collaborating with data science team





DataOps

data(infra) engineer / data science

출처 : Netflix Presents: A Human Friendly Approach to MLOps, Julie Pitt and Ashish Rastogi, Netflix
https://www.youtube.com/watch?v=fOSZuONmLbA