데이터 엔지니어 초급 4일차

데이터 웨어하우스 설계

Park Suhyuk
Data Ingestion Team Leader


psyoblade


psyoblade

hive 2.3.2
spark 2.4.5
fluentd 3.8.0
sqoop 1.7.4
docker-ce 19.03.13
ubuntu 18.04 LTS

NCSOFT®

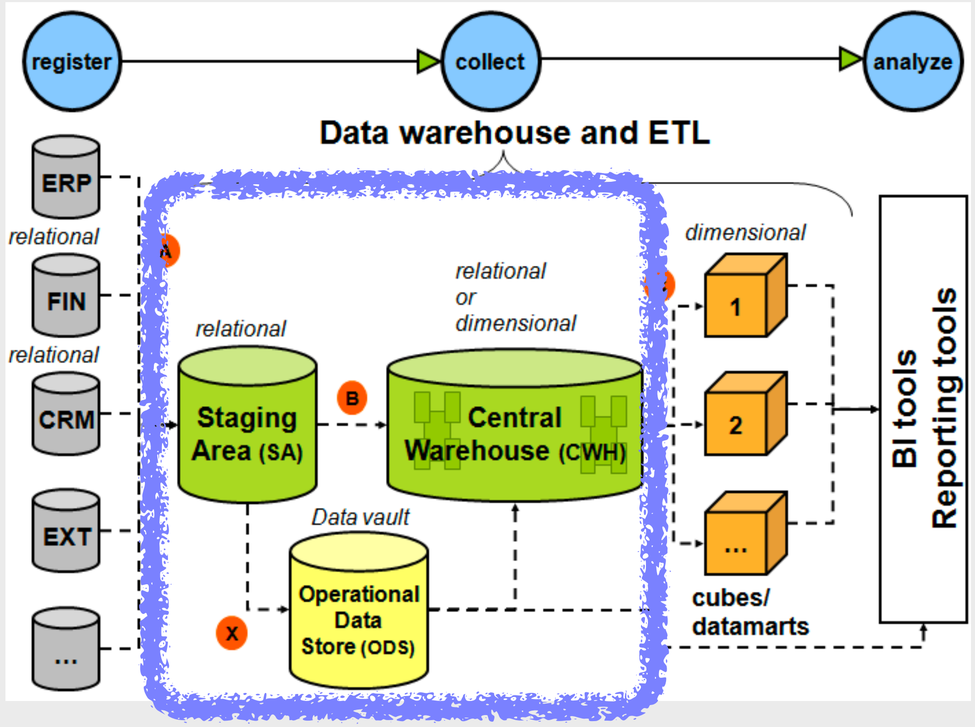
목차

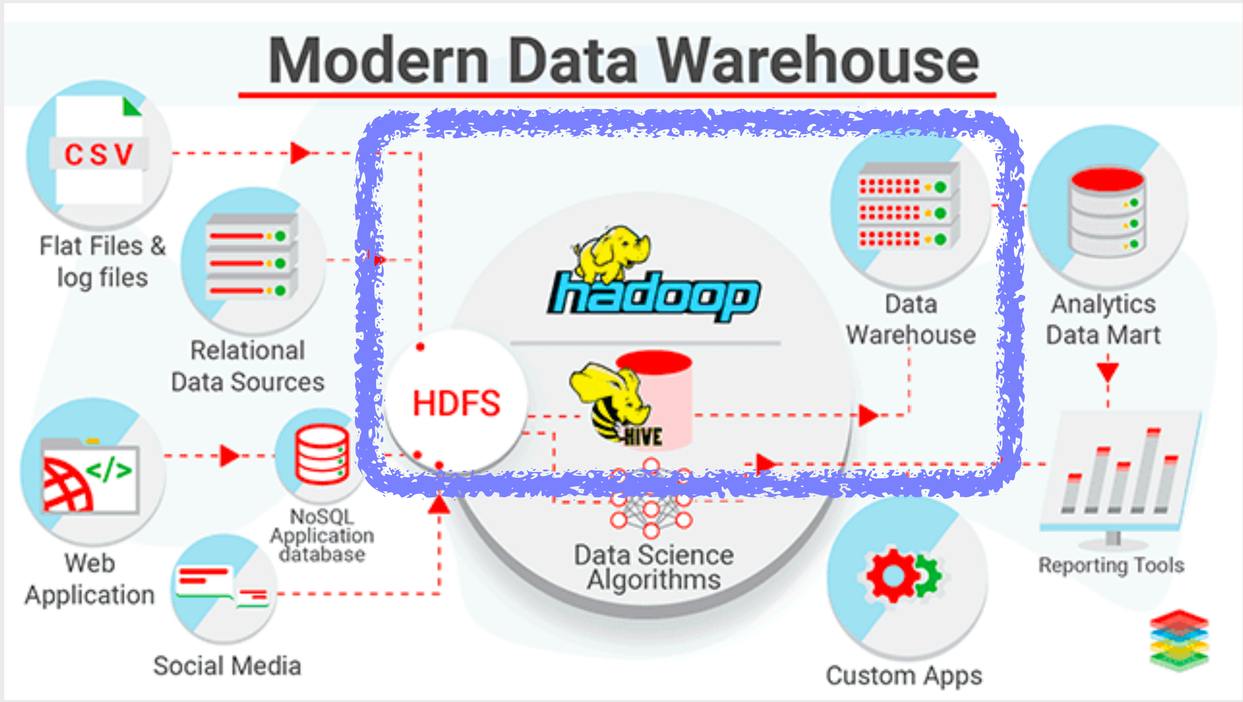
데이터 웨어하우스

데이터 웨어하우스

DataWarehouse on'Database' and 'Hadoop'

전통적인 관계형 DB는 별도의 저장소를 통해 저장하는 반면, 하둡의 경우 수십 ~ 수백대의 노드의 분산 저장소에 저장하기 때문에, 대용량 데이터의 저장 및 처리가 가능한 반면 분산된 노드에 분산되어 저장되기 때문에 데이터베이스의 정규화를 통한 조인의 최적화와 같은 기법은 Disk 및 Network I/O 비용이 너무 크기 때문에, 테이블의 비정규화를 통해 조인과 같은 연산을 최소화 하고 데이터의 압축 및 컬럼 단위 색인을 통한 최적화가 필요합니다





https://stackoverflow.com/questions/25474114/data-warehouse-and-or-database

지표 생성을 위한 모델링

지표 모델링

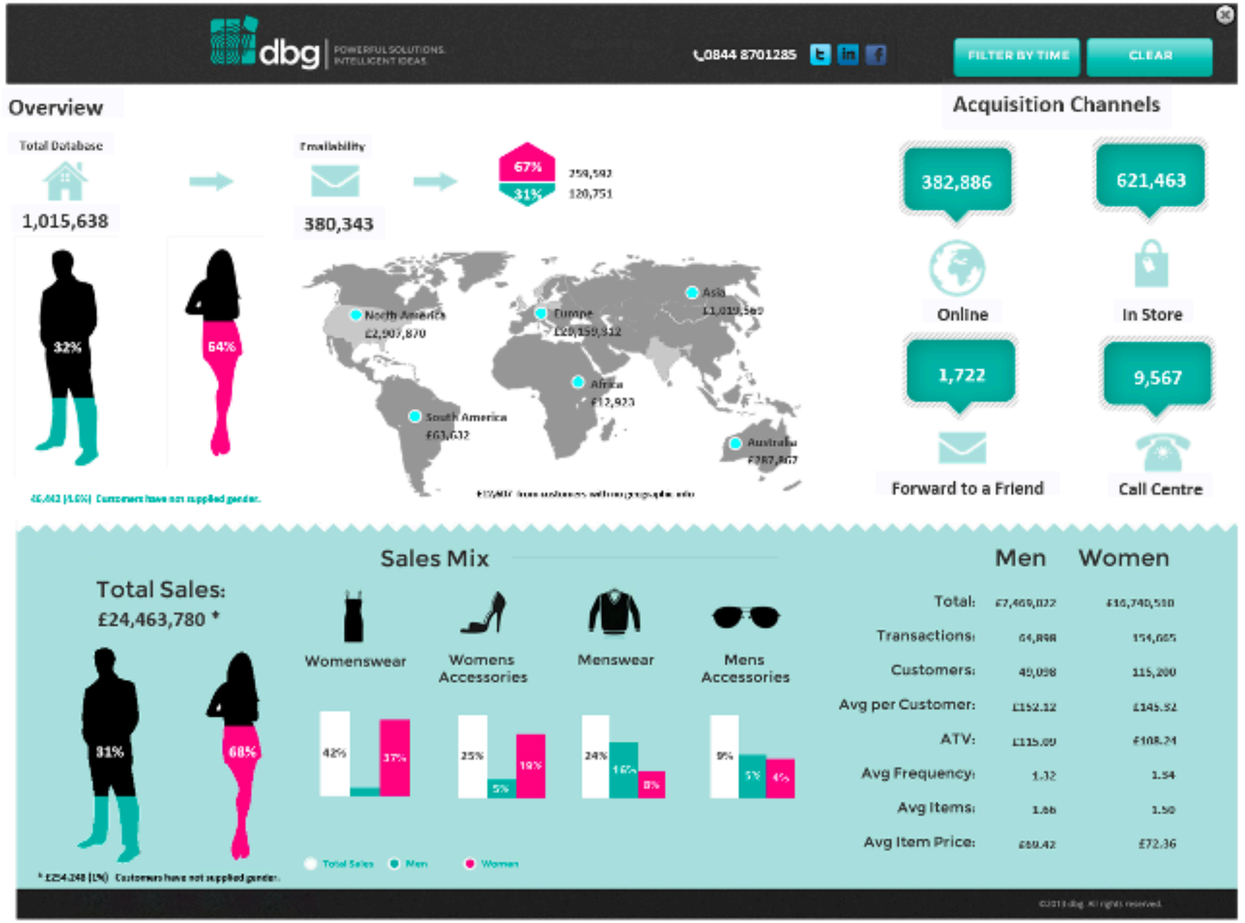
Summary Index Modeling

요약 지표 생성을 위한 몇 가지 예제를 통해 어떤 형태로 데이터를 구조화 해야 효과적인 지에 대해 설계합니다. 예를 들어 유저 당 매출, 신규 유저수 혹은 연령대 별 매출 등의 지표를 효과적으로 생성하기 위해서는 적절한 모델링이 필요하며 크게 2가지 유형의 데이터 테이블이 필요합니다.

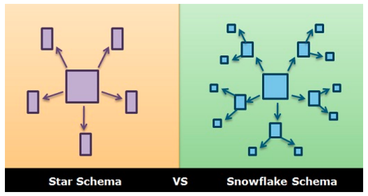
- Dimension (차원) : 고객 혹은 상품 등의 상태 (나이, 성별, 지역, 등급)
- Fact (사실) : 고객의 행위 혹은 서비스를 사용하면서 발생한 모든 사건 (서비스 가입/접속, 상품 구매 등의 이벤트)

이 두가지 데이터를 통해 아래와 같은 요약 지표 생성이 가능합니다

- 지역 별 가입자 수
- 연령대 별 매출, 혹은 이벤트 참여도
- 고객 등급별 접속 횟수

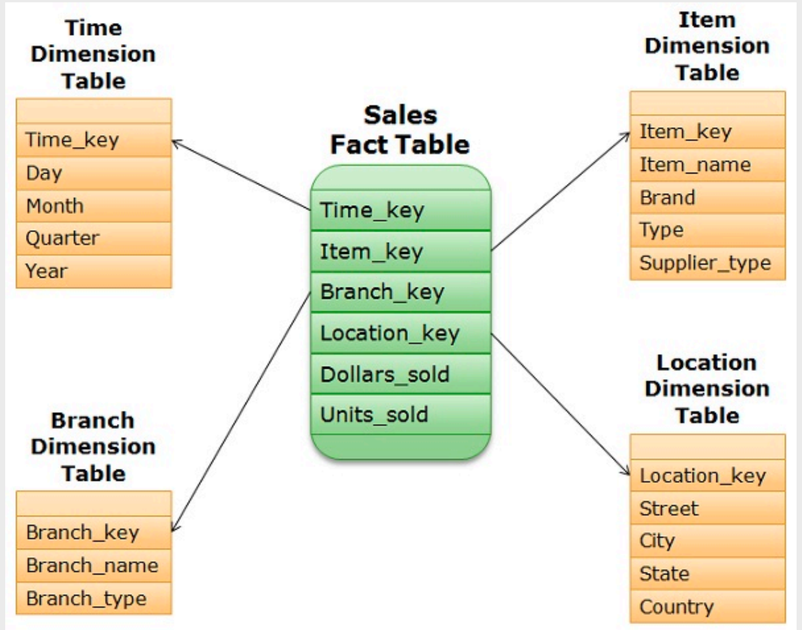


데이터 모델링

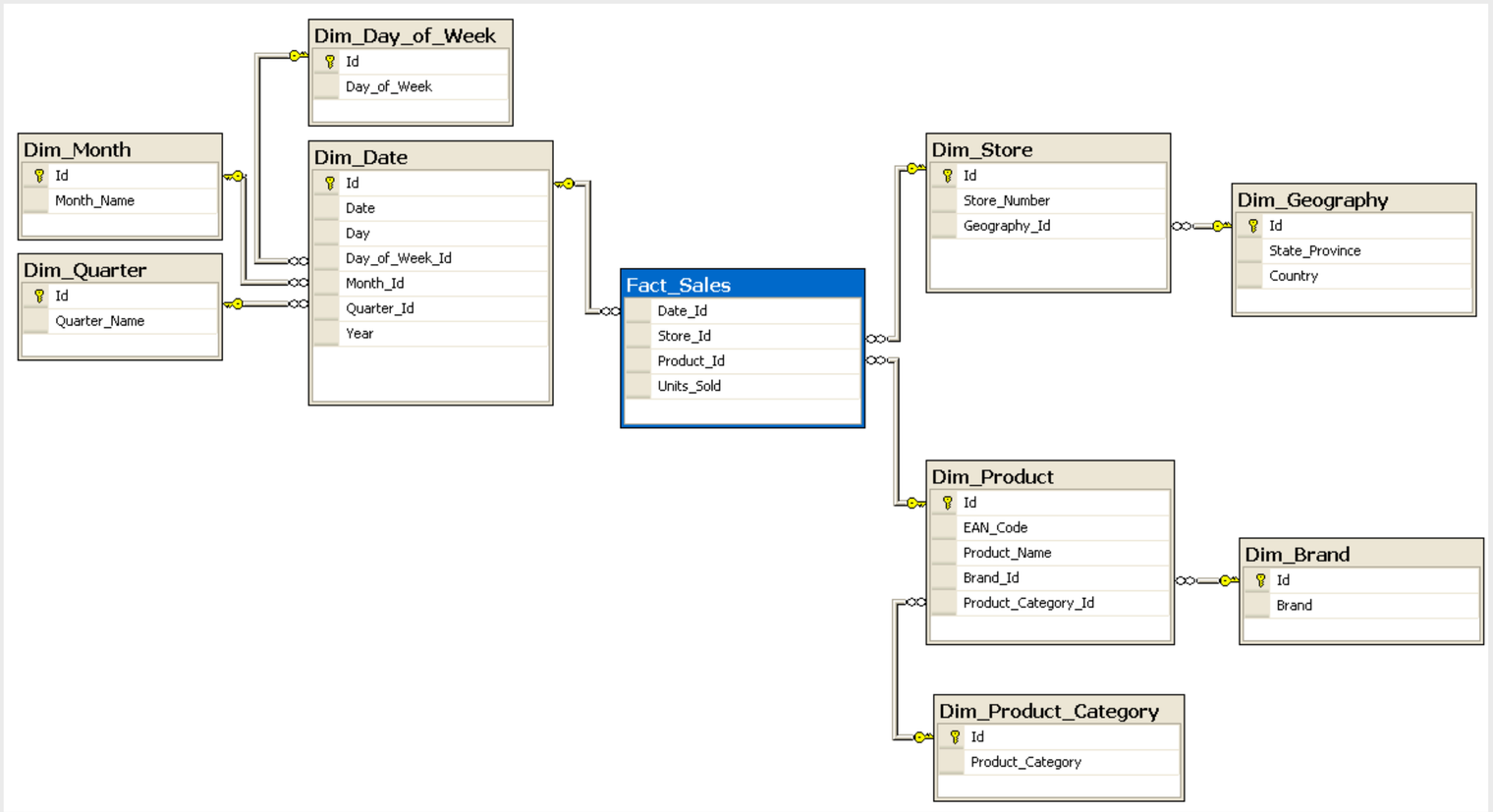


Differences Between Star and Snowflake Schema

Star Schema (스타 스키마)



Snowflake Schema (눈송이 스키마)



지표 설계

Differences Between Star and Snowflake Schema

가장 단순한 데이터 웨어하우스 스키마이며, 팩트 테이블의 PK 와 차원 테이블의 관계로 이루어진 스키마이며, 모양의 이름을 따서 붙인 이름입니다. 무엇보다 스타 스키마는 이해하기 쉬우며, 테이블 설계 및 애플리케이션 사용에 있어 간결하고 명확하여 서비스 유지보수 및 운영에 용이합니다. 반면에 눈송이 스키마의 경우 정규화를 통한 중복 데이터를 최소화 하고, 저장공간을 절약할 수 있으나 더 많은 조인이 필요하여 저장과 읽기에 성능에 악영향을 미칠 수 있으므로 스타 스키마에 비해 널리 사용되지는 않습니다. 특히 NoSQL 과 같은 분산 저장소의 경우 조회 성능에 가장 나쁜 구조입니다.

한글명	스타 스키마 (star)	눈송이 스키마 (snowflake)
계층 구조	차원 테이블에 저장	별도의 테이블에 저장
조인 연산	단일 조인만 팩트 테이블과 수행	차원의 깊이 만큼의 조인이 발생
설계 복잡도	비교적 간단함	매우 복잡함
정규화 수준	비 정규화 된 테이블 구조	정규화 된 테이블 구조
데이터 중복	높은 수준의 데이터 중복	매우 낮은 수준의 데이터 중복

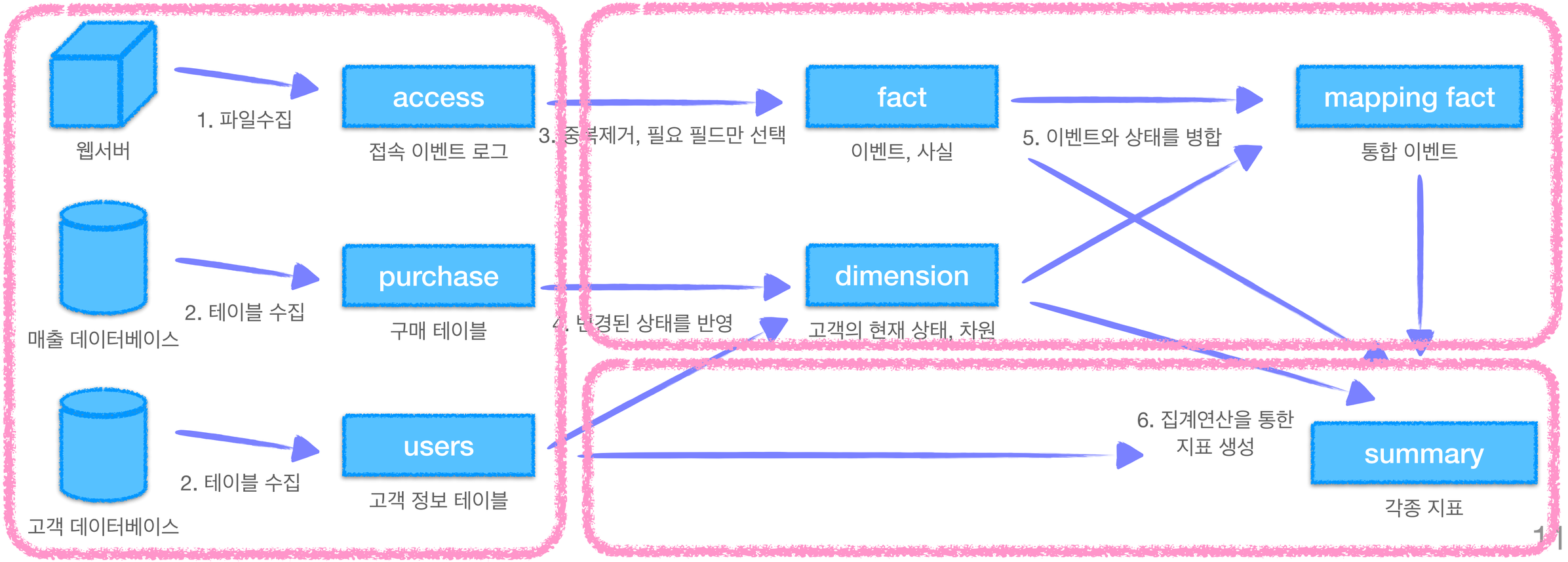
https://dbrang.tistory.com/1381

데이터 파이프라인 구성

데이터 파이프라인 구성

Data Warehouse Pipeline

고객의 행위에 해당하는 이벤트, 매출 및 고객 정보를 저장하고 있는 테이블 등이 데이터의 원천이며, 이러한 데이터를 일관된 형태 혹은 포맷으로 저장하는 단계를 입수(Ingestion, Extract)라고 하며, 분산 환경에서 전처리, 가공 및 적재를 변환(Transform) 그리고 조회를 위해 저장하는 단계를 적재(Load)라고 합니다



https://github.com/psyoblade/data-engineer-basic-training/tree/master/day4