컴퓨터 캡스톤 디자인 1

THE DATA CIVILIZER SYSTEM

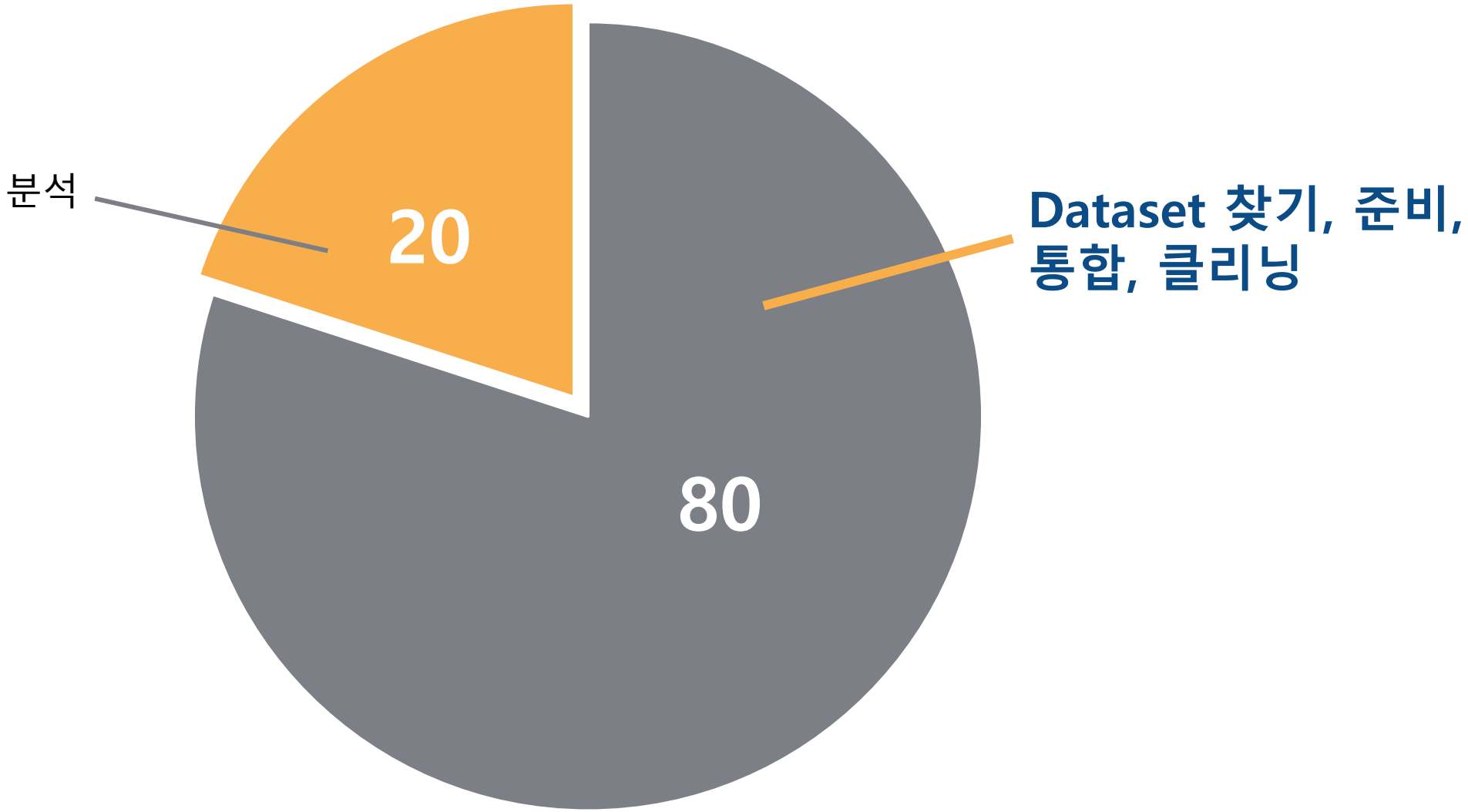
소프트웨어학부
[팀명] 강은지 박서연 석예림

Contents

01

DATA CIVILIZER 개요

기존 문제점



The Data Civilizer System

: 해결 방법

검색

: 수 많은 테이블에서

관심 dataset 신속하게 검색

연결

: 관련 dataset연결

Cleaning

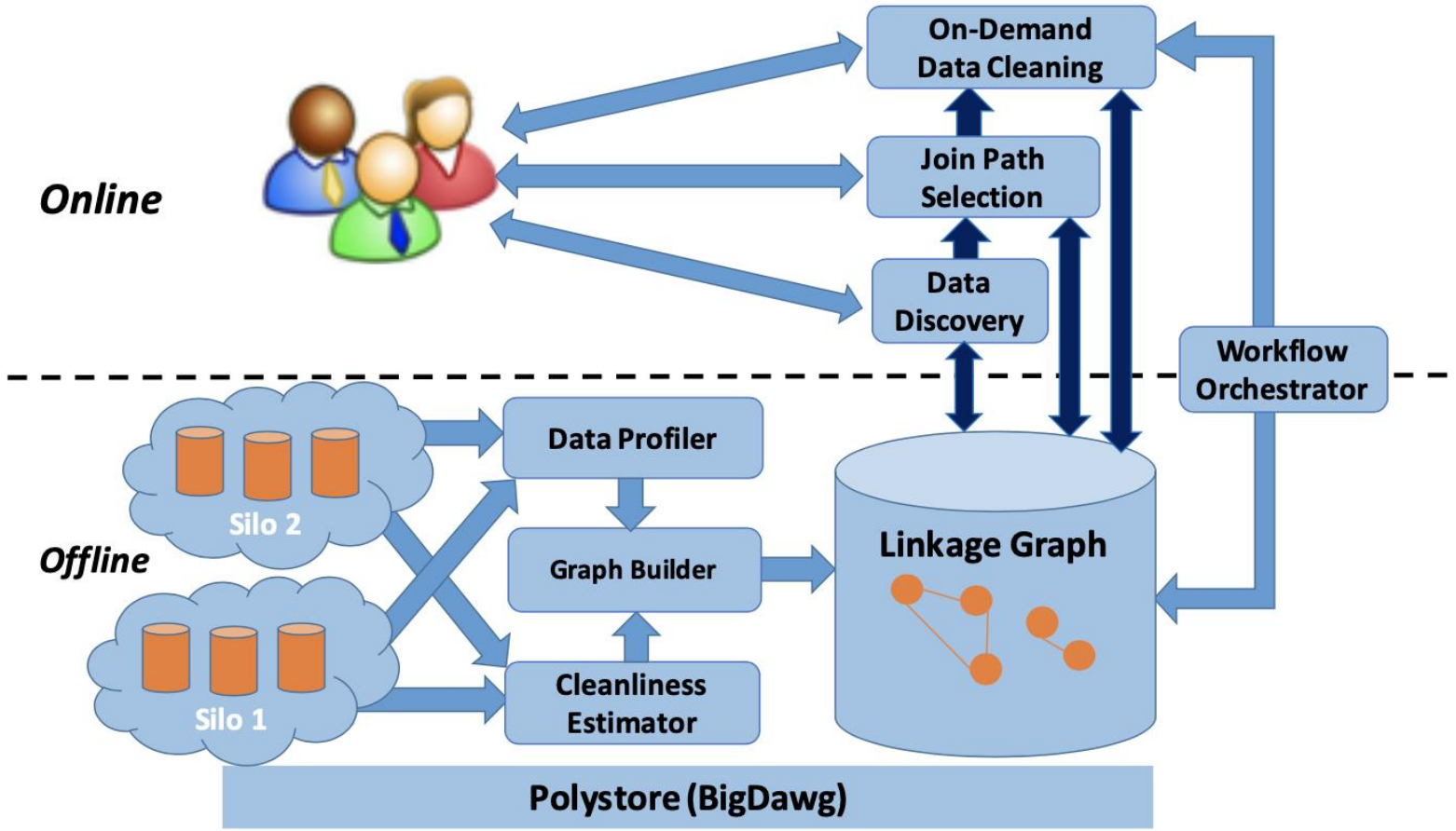
: 원하는 데이터 cleaning

계산

: dataset들을 호스팅하는

분산된 데이터 저장소의 응답

The Data Civilizer System



02

Cleanliness

Polystore Query Processing

- **BigDAWG polystore 사용**
  ※ Big Data Analytics Working Group
  : 여러 저장 시스템 query 처리 위해
- 가정: **사용자가 검색 실행**

사용자가 특정 칼럼 집합을 포함하는 복합적인 테이블에 관심이 있는 경우 관련 데이터 셋을 결합하는 과정 필요

PK-FK 마이닝 기술 사용

관심있는 테이블을 계산하기 위해 가장 적합한 Join Path 선택

Join Path Selection

기존 : 비용 중심적으로 선택 -> 데이터 청결 문제 발생

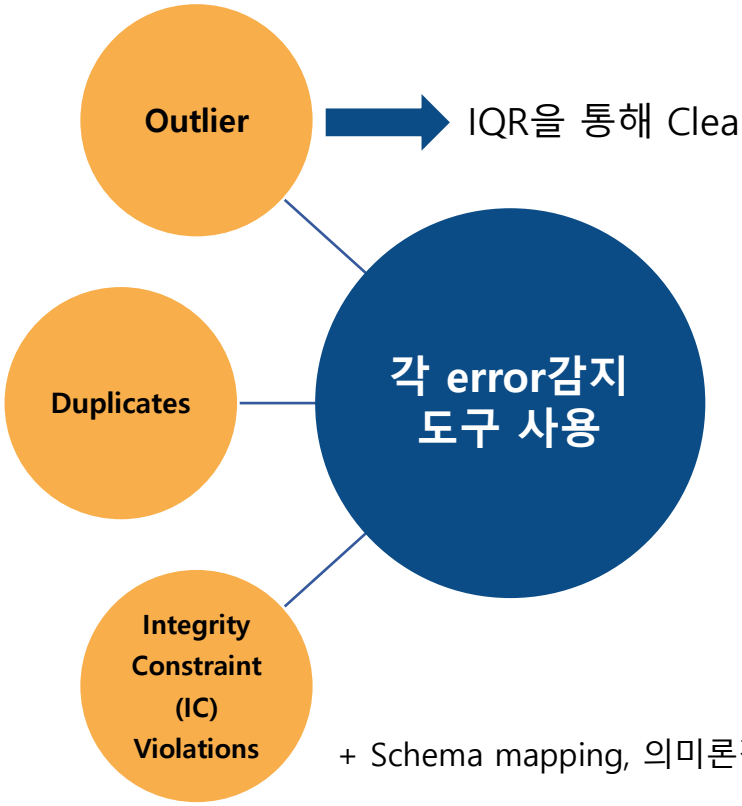
- Highest quality result

- 쿼리를 실행한 테이블과 연관된 데이터를 cleaning

Cleanliness Model

-> 복합적인 오류 검출

- Data error 유형



+ Schema mapping, 의미론적으로 동일한 테이블 병합, 병합된 테이블 안에 IC violation 감지

03

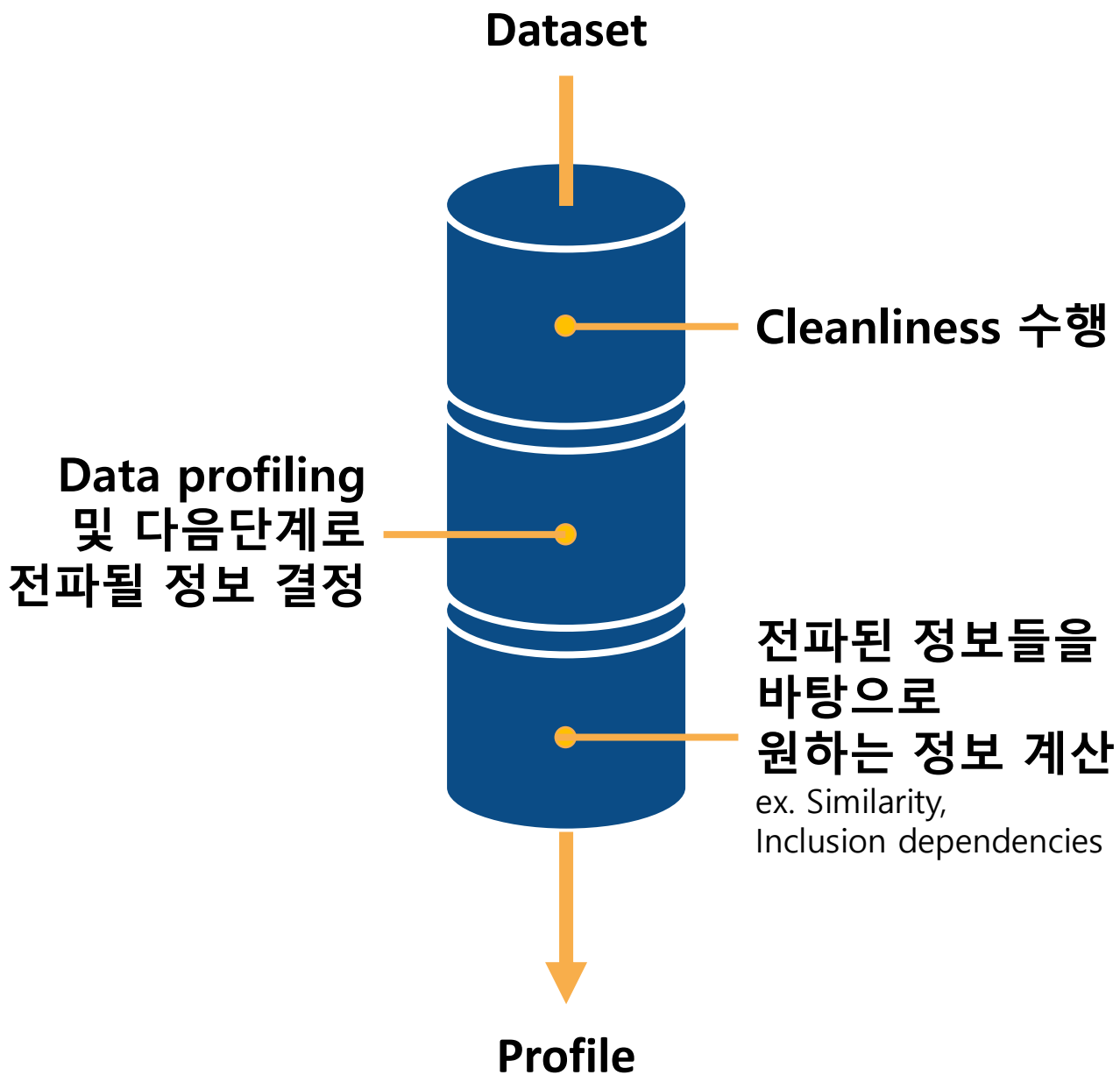
Data profiling

- Graph를 build하기 위해 먼저 데이터를 통계 및 요약을 하는 과정 -

Data Profiling

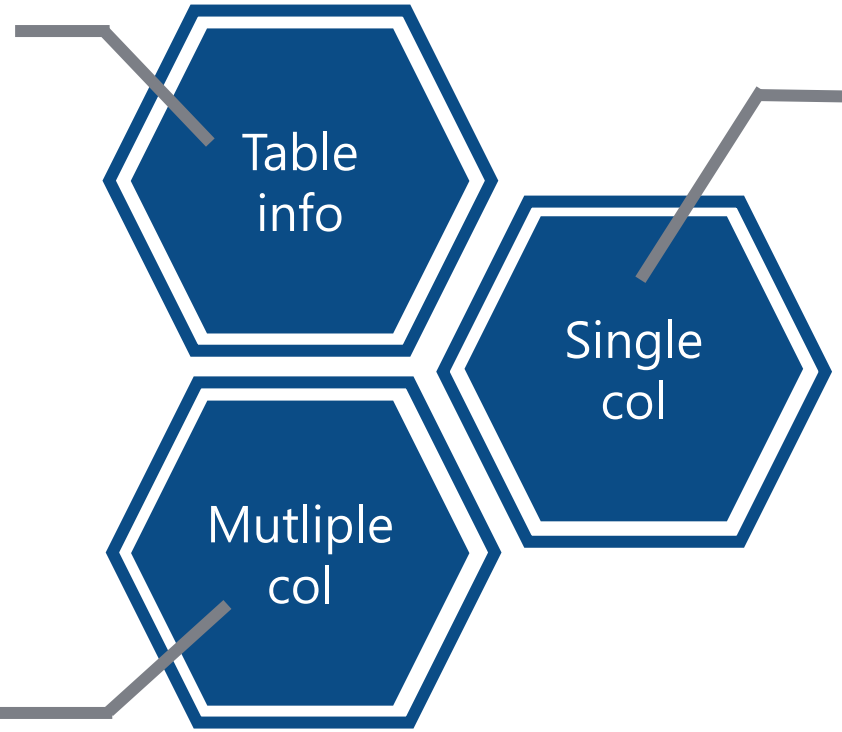
Dataset **Paging** 작업을 통한
증분 프로파일링 진행

Reducing memory pressure



Data Profiling

number of variables
missing cells
duplicate rows
column types



number range
sum, mean, var, min, max
histogram
distinct count
unique
quartile, IQR
Kurtosis, Skewness
Common Values

Pearsons's
Spearman's
Kendall's
Phik's
Carmer's

Numerical Value Similarity

01. Column 유사성 계산

profile로 계산된 데이터 빈도수를 바탕으로

$$\frac{\text{(겹치는 데이터에서 작은 빈도수)}}{\text{(전체 데이터의 합집합)}}$$

02. Inclusion Dependency(IND)

$$\frac{\text{(칼럼간 교집합)}}{\text{(FK테이블의 칼럼집합_중복제거)}} = 1$$

=> PK에서 FK로의 IND 성립

Text Value Similarity

GloVe
(Global Vectors)

- 단어 동시 등장 여부(확률) 보존
- GloVe로 임베딩된 단어 벡터 내적 = 동시 등장 확률의 로그값
- 카운트 기반의 LSA와 예측기반의 Word2Vec의 단점 보완

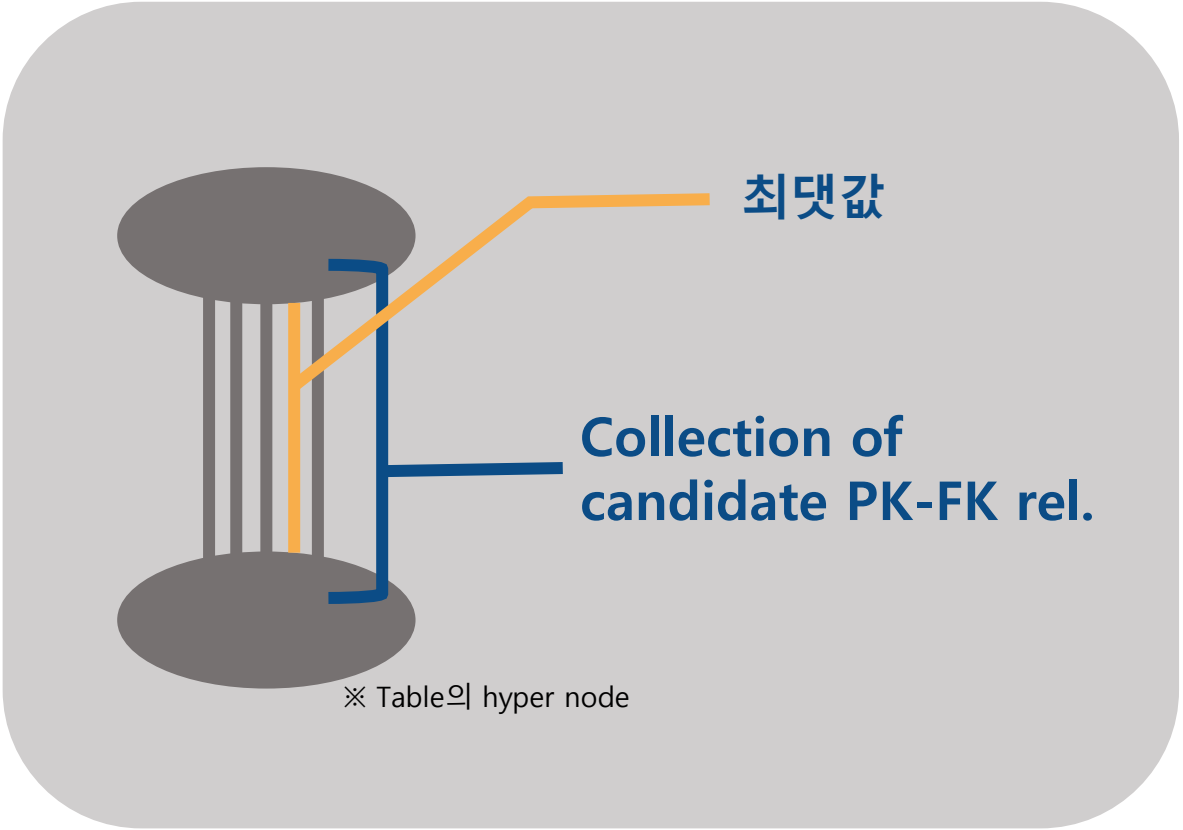
Fasttext

- 하나의 단어 안에도 여러 단어들이 존재하는것으로 간주(subword)
- 자신이 가진 데이터로 embedding하고 싶은 경우, 형태소 분석이 완료된 데이터이어야함
- word2vec보다 더 좋은 성능을 얻을 수 있으나 계산 시간은 더 소요됨

LSA
(Latent Semantic Analysis)

- 토픽 모델링
- Truncated SVD를 활용하여 문서에 숨어있는 의미를 이끌어내기 위한 방법
- 데이터가 큰 경우 차원축소 효과 및 단어와 문맥 간의 내재적인 의미 보존 가능
- 새로운 문서나 단어가 추가되면 아예 처음부터 작업을 새로 시작해야 됨

Refine Candidate PK-FK relationship



최댓값을 가지는 edge를 제외하고 모두 제거 (=refine)

※ 4 machine learning classification(refine) algorithm from Rostin et al
- 『A. Rostin: A machine learning approach to foreign key discovery』

04

Update

Update에서, profiling을 비교해서
변화가 있을 경우 graph DB 업데이트

“어떤 테이블에 대해 profile을 진행하여 기존의 값과 비교할 것인가?”

모든 테이블에 대해 매번 profile을 진행
하는 것은 시간이 많이 소요됨

DB query log 읽어와서 업데이트 된 데이터 파악

(mysql.general_log)

Data Civilizer System Structure