DATA ANALYSIS

BY NIMO BEEREN, MARIA POGODAEVA, HENRIQUE DIAS, PANAGIOTIS BANOS, ÇAĞLA SÖZEN, GABRIELA SLAVOVA

TU/e EINDHOVEN UNIVERSITY OF TECHNOLOGY

GHTORRENT PROJECT DATASET

GitHub is an online code archive and collaboration platform with 65+ Million users and 200+ Million repositories. All the information regarding users, repository commits, coding languages, every bit of information that surrounds a repository such as comments, pull requests, issue tickets etc. have been stored and made available through the GHTorrent Project database.

ATTRIBUTES:
ID
LOGIN
CREATED_AT
TYPE
FAKE
DELETED
LONG
LAT
COUNTRY_CODE
STATE
CITY
LOCATION

PREPROCESSING

- Record flagged as **fake**? → Delete record
- Create **country** column using **country_code**

# Records: 32.430.223 → 24.562.103

DEPENDENCIES

HARD

country_code → country

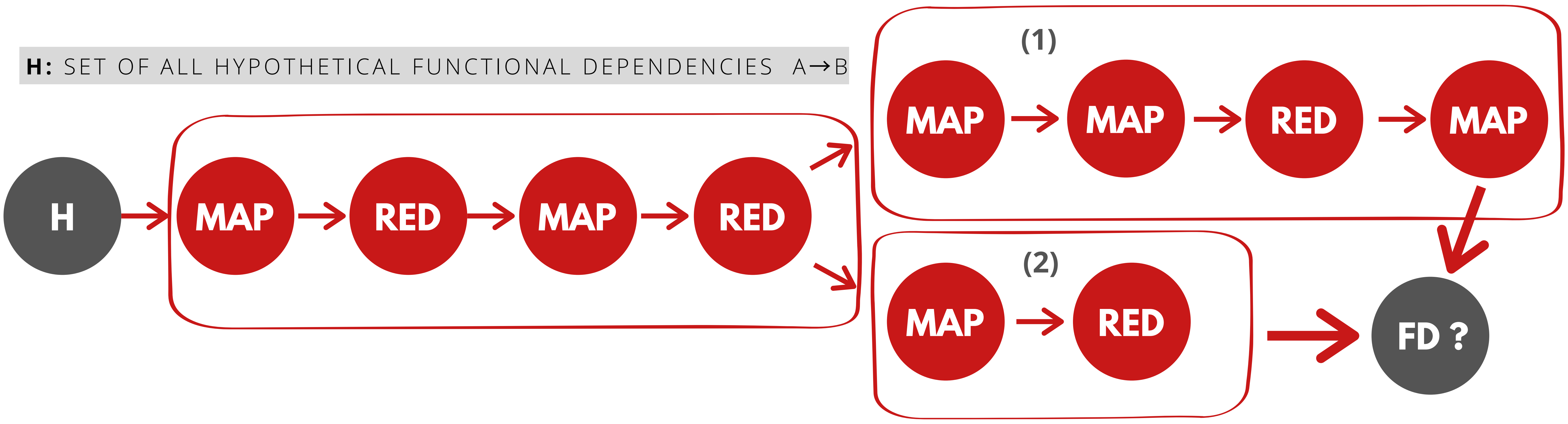
SOFT

state→ country
city → country
city → state
company → country

DELTA

city, country_code → long, lat

DISTRIBUTION OF COMPUTATION


H: SET OF ALL HYPOTHETICAL FUNCTIONAL DEPENDENCIES A→B

H → MAP → RED → MAP → RED

(1) MAP → MAP → RED → MAP

(2) MAP → RED

FD ?

COMMON OPERATIONS

1. **Map** each record to a pair *<(a, b), 1>*
2. **Reduce** all pairs **by key** to *<(a b), sum>*
3. **Map** to *<(a, (b, sum)>* pairs
4. **Reduce** each pair of pairs **by key** then aggregate non unique *< a,(b, sum) >*

(2) δ-FDs

5. **Map** each pair of records to *TRUE* iff their difference* is within a range *δ*, *FALSE* otherwise
6. **Reduce** each pair of booleans by *B1∧B2*

*Absolute/Edit distance for numerical/string data

(1) HARD / SOFT FDs

5. **Map** each pair *<(a,b), sum>* to *<t_a, P_a>* where *t_a* is the total number of records and *P_a* is the probability of randomly selecting two records with the same right-hand value, given that they have that particular left-hand value
6. **Map** each pair *<t_a, P_a>* to *<t_a,(t_a*P_a)>*.
7. **Reduce** each pair of pairs *<t1,(t1*P1)>*, *<t2,(t2*P2)>*.to *<t1 + t2, t1*P1 + t2*P2>*
8. **Map** each pair *<T, t*P>* to the value *t*P/T*

*Dataset is available at: https://drive.google.com/drive/folders/16xy-5SntyM0atfJNQh3Ognw2XZcrdZHz
**Video available at: https://drive.google.com/file/d/1eFp5FyAIZSvtWxqEX71ozvfmtmv_pBIm/view

Video avaiable at: