Title (en)

SYSTEM AND METHOD FOR PREDICTING AN OVERALL SIMILARITY SCORE BETWEEN TWO PRIMARY ENTITIES OF A DATA LAKE

Title (de)

SYSTEM UND VERFAHREN ZUR VORHERSAGE EINER GESAMTÄHNLICHKEITSBEWERTUNG ZWISCHEN ZWEI PRIMÄRENTITÄTEN EINES DATENLACKS

Title (fr)

SYSTÈME ET PROCÉDÉ POUR PRÉDIRE UN SCORE GLOBAL DE SIMILARITÉ ENTRE DEUX ENTITÉS PRIMAIRES D'UN LAC DE DONNÉES

Publication

EP 4272089 A1 20231108 (FR)

Application

EP 21851988 A 20211231

Priority

- FR 2014297 A 20201231
- IB 2021062515 W 20211231

Abstract (en)

[origin: WO2022144848A1] One of the aims of said invention is to provide an objective and reproducible tool for quantifying redundancy in a data lake. To achieve this, the inventors propose training a machine learning model, using existing data lakes, to predict an overall similarity score which is representative of the similarity between two data entities of a data lake. In practice, instead of comparing each of the data fields of the data entities, the invention proposes determining the overall similarity score from intermediate similarity scores which are calculated for random samples of the data entities.

IPC 8 full level

G06F 16/35 (2019.01)

CPC (source: EP)

G06F 16/355 (2019.01)

Designated contracting state (EPC)

AL AT BE BG CH CY CZ DE DK EE ES FI FR GB GR HR HU IE IS IT LI LT LU LV MC MK MT NL NO PL PT RO RS SE SI SK SM TR

Designated extension state (EPC)

BA ME

Designated validation state (EPC)

KH MA MD TN

DOCDB simple family (publication)

WO 2022144848 A1 20220707; EP 4272089 A1 20231108; EP 4272090 A1 20231108; WO 2022144852 A1 20220707

DOCDB simple family (application)

IB 2021062515 W 20211231; EP 21851988 A 20211231; EP 21854759 A 20211231; IB 2021062519 W 20211231