



(51) International Patent Classification:
G06F 16/215 (2019.01)

(21) International Application Number:
PCT/US2021/019572

(22) International Filing Date:
25 February 2021 (25.02.2021)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
62/981,646 26 February 2020 (26.02.2020) US
17/006,504 28 August 2020 (28.08.2020) US

(71) Applicant: AB INITIO TECHNOLOGY LLC [US/US];
201 Spring Street, Lexington, Massachusetts 02421 (US).

(72) Inventors: JOYCE, John; 503 Walnut Street, Newton, Massachusetts 02460 (US). ISMAN, Marshall A.; 11 Valley Spring Road, Newton, Massachusetts 02458 (US). MELBOUCI, Sandrick; 10409 Easterday RD, Myersville, Maryland 21773 (US).

(74) Agent: DEVRIES, Gretchen A. et al.; Fish & Richardson P.C., P.O. Box 1022, Minneapolis, Minnesota 55440-1022 (US).

(81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, IT, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW,

(54) Title: GENERATING RULES FOR DATA PROCESSING VALUES OF DATA FIELDS FROM SEMANTIC LABELS OF THE DATA FIELDS

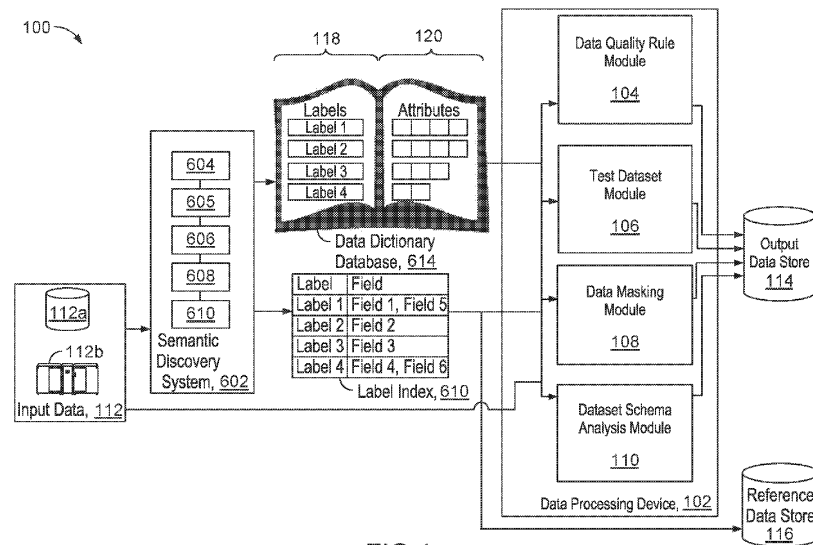


FIG. 1

(57) Abstract: Methods and systems are configured to determine a semantic meaning for data and generate data processing rules based on the semantic meaning of the data. The semantic meaning includes syntactical or contextual meaning for the data that is determined, for example, by profiling, by the data processing system, values stored in a field included in data records of one or more datasets; applying, by the data processing system, one or more classifiers to the profiled values; identifying, based on applying the one or more classifiers, one or more attributes indicative of a logical or syntactical characteristic for the values of the field, with each of the one or more attributes having a respective confidence level that is based on an output of each of the one or more classifiers. The attributes are associated with the fields and are used for generating data processing rules and processing the data.

[Continued on next page]

SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN,
TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(84) Designated States (*unless otherwise indicated, for every kind of regional protection available*): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

- *with international search report (Art. 21(3))*
- *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))*