



(51) International Patent Classification:

G06F 9/50 (2006.01) G06F 13/16 (2006.01)
G06F 9/54 (2006.01) G06N 20/00 (2019.01)

(21) International Application Number:

PCT/US2024/023706

(22) International Filing Date:

09 April 2024 (09.04.2024)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

63/459,076 13 April 2023 (13.04.2023) US

(71) Applicant: **VISA INTERNATIONAL SERVICE ASSOCIATION** [US/US]; P.O. Box 8999, San Francisco, California 94128 (US).

(72) Inventors: **XIAO, Xuyang**; Visa International Service Association, P.O. Box 8999, San Francisco, California 94128 (US). **SHI, Ningyu**; Visa International Service Association, P.O. Box 8999, San Francisco, California 94128 (US).

tion, P.O. Box 8999, San Francisco, California 94128 (US). **GENTY, Victor, James**; Visa International Service Association, P.O. Box 8999, San Francisco, California 94128 (US). **ZHOU, Zhimin**; Visa International Service Association, P.O. Box 8999, San Francisco, California 94128 (US).

(74) Agent: **PREPELKA, Nathan, J.** et al.; The Webb Law Firm, One Gateway Center, 420 Ft. Duquesne Blvd., Suite 1200, Pittsburgh, Pennsylvania 15222 (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, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN, MU, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH,

(54) Title: SYSTEM, METHOD, AND COMPUTER PROGRAM PRODUCT FOR PROCESSING INFERENCE REQUESTS TO EXECUTE A MACHINE-LEARNING MODEL

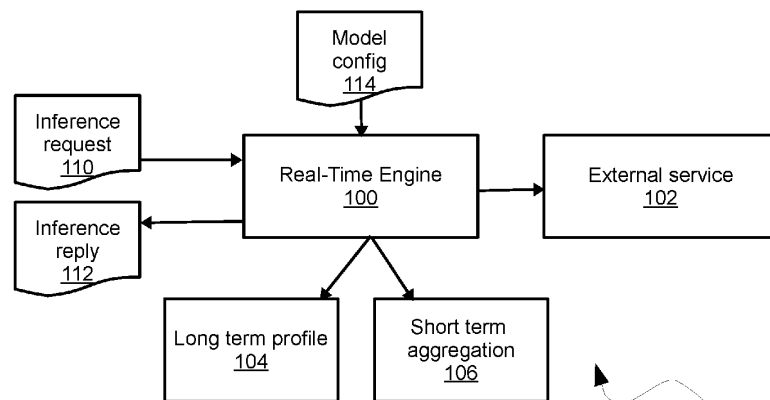


FIG. 1

(57) Abstract: Provided is a system, method, and computer program product for processing inference requests to execute a machine-learning model. The system includes at least one processor programmed or configured to receive an inference request at a request handler having a first programming language capable of multithreading, communicate a first message based on the inference request from the request handler to at least one object associated with a machine-learning model, the at least one object having a second programming language different than the first programming language, determine whether to call at least one external service based on the first message and the at least one object, communicate a second message from the at least one object to an asynchronous function having the first programming language based on determining whether to call the at least one external service, and call the at least one external service via the asynchronous function while the at least one object associated with the machine-learning model executes.

[Continued on next page]

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, CV, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, 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, ME, 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).

Declarations under Rule 4.17:

- *as to applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii))*
- *as to the applicant's entitlement to claim the priority of the earlier application (Rule 4.17(iii))*

Published:

- *with international search report (Art. 21(3))*