GREEN.DAT.AI – PODATKOVNI PROSTORI ZA IZVEDBO SISTEMOV UMETNE INTELIGENCE

Domen Mongus[1], Aljaž Žel[1], Mitja Žalik[1], Gregor Horvat[1], Dino Vlahek[1], Matej Brumen[1]
[1] Fakulteta za elektrotehniko, računalništvo in informatiko, Univerze v Mariboru, Koroška cesta 46, 2000 Maribor
domen.mongus@um.si, aljaz.zel@um.si, mitja.zalik@um.si, g.horvat@um.si, dino.vlahek1@um.si, matej.brumen@um.si

Navkljub splošnim trendom digitalizacije se številne zasebne in javne organizacije danes še vedno soočajo s težavami pri implementaciji in integraciji sistemov umetne inteligence v svoje poslovne procese. Pomemben razlog za to je gotovo tudi neustreznost obstoječih podatkovnih zbirk za neposredno izvedbo strojnega učenja. Podatkovni strežniki, distribucijski sistemi in podatkovna jezera danes namreč še vedno nastopajo večinoma izolirano, v obliki tako imenovanih podatkovnih silosov. Nasprotno pa je natančnost in splošna učinkovitost metod umetne inteligence pogosto neposredno odvisna od številnih kontekstnih informacij, ki jih je običajno mogoče pridobiti zgolj s povezovanjem heterogenih in pogosto tudi fizično ločenih podatkovnih virov in tokov. Pri njihovem povezovanju pa pogosto naletimo tudi na pomanjkanje metapodatkovnih standardov, splošno slabo kakovost metapodatkov ter semantično neusklajenost kar dodatno oteži in podraži njihovo harmonizacijo in transformacije v podatkovne formate, ki so primerni za strojno učenje.

Rešitev za to obljubljajo tako imenovani podatkovni prostori, ki s svojo idejo o federalizaciji podatkovnih virov predstavljajo jedro nove podatkovne strategije Evropske Unije in temelj za vzpostavitev enotnega Evropskega podatkovnega trga. Podatkovni prostor je vmesna programska infrastruktura, ki prek storitev centralnega podatkovnega kataloga posreduje dostop do distribuiranih podatkovnih skladišč. Ponudniki in uporabniki podatkov so v podatkovni prostor vključeni prek namenskih vtičnikov, ki omogočajo samodejno iskanje in odkrivanje razpoložljivih virov podatkov in sklepanje dogovora o njihovem prenosu. Ob uporabi standardiziranih podatkovnih modelov pa lahko ti vtičniki omogočajo tudi harmonizacijo podatkov s samodejno preslikavo v interne podatkovne formate uporabnika. Tako lahko vsi ponudniki podatkov delujejo neodvisno od ostalih in so izključno sami odgovorni za kakovost, varnost in zasebnost podatkov, ki jih zagotavljajo.

V okviru projekta GREEN.DAT.AI smo razvili referenčno arhitekturo podatkovnega prostora, ki poleg samih podatkov omogoča tudi izmenjavo modelov strojnega učenja in je kot tak prilagojen na izvedbo velikih sistemov umetne inteligence. V svoji zasnovi je skladen s specifikacijami mednarodnih iniciativ GAIA-X in IDSA, pri čemer temelji na naslednjih plasteh:

- Pilotna plast vsebuje vse vire podatkov, podatkovne modele in algoritme/modele strojnega učenja, ki so na voljo za skupno rabo ali izmenjavo med sodelujočimi v podatkovnem prostoru. Ta plast tako zagotavlja interoperabilnosti med pilotnimi okolji in v njih nameščenimi napravami na robu, ki sicer hranijo in obdelujejo dejanske podatke.
- Interoperabilnostna plast vsebuje vtičnike, ki so potrebni za sklepanje dogovorov in dejansko izmenjavo podatkov. Ti so nameščeni v pilotnih okoljih in so lahko generični ali specifični, odvisno od značilnosti podatkov in modelov, ki jih posamezno okolje želi objaviti ali prejeti. Podatke in modele so pri tem zmožni opisati, podati definicijo njihovega podatkovnega modela in vhodne ter izhodne parametre ponujenih modelov.
- Pogajalska plast vključuje komponente, ki se uporabljajo za zagotavljanje zaupanja, upravljanje identitet ter izvedbo ostalih funkcionalnosti, ki so potrebne za delovanja podatkovnega

prostora, vključno s skupno rabo podatkov, ponudbo storitev, orkestracijo in sklepanjem pametnih pogodb.

- Poslovna plast vključuje vse storitve, ki so potrebne za trgovanje s podatki in modeli umetne inteligence. Natančneje, ta plast predstavlja tržnico, ki uporablja katalog za promocijo orodij, storitev in informacij, potrebnih za lažjo identifikacijo podatkov in modelov, omejitev pri njihovi rabi in drugih pogojev, ki lahko vplivajo na omejitve dostopa.

Referenčno arhitekturo implementiramo v šest pilotnih okolij, ki vključujejo trgovanje z energijo vetrnih elektrarn, optimizacijo polnjenja električnih avtomobilov, distribucijo električnih koles glede na previdene potrebe, optimizacijo gnojenje in zalivanja kmetijskih površin ter odkrivanje bančnih prevar. Pri tem pa vključuje številne metode umetne inteligence, poleg tradicionalnih metod ansambelskega in globokega učenja še generiranje sintetičnih podatkov, učenje značilnic, gručenje, učenje s prenosom znanja, zaznavanje dogodkov in orkestracija cevovodov za posnemanje obnašanja.

Ključne besede: podatkovni prostori; platforme strojnega učenja; referenčna arhitektura;

GREEN.DAT.AI – DATA SPACES FOR THE IMPLEMENTATION OF ARTIFICIAL INTELLIGENCE SYSTEMS

Numerous public and private organisations are still struggling with the integration of artificial intelligence systems into their business processes, in part also due to the data silos in which they maintain their data. The solution for this is promised by the so-called data spaces, which offer federalisation of data sources. As such, it lies at the core of the new data strategy of the European Union and is considered as a foundation for the establishment of a single European data market. In the GREEN.DAT.AI project, we developed a reference architecture of the data space, which, in addition to ensuring a safe and secure exchange of data, also enables the exchange of machine learning models and, as such, is adapted to the implementation of large artificial intelligence systems. In its design, it complies with the specifications of the international initiatives GAIA-X and IDSA and prescribes four layers, including (i) a pilot layer where data is generated and processed; (ii) the interoperability layer, where data space connectors are deployed for ensure efficient data exchange, (iii) negotiation layer, where agreements of data exchange are being made, and (iv) business layer with services that promote data exchange, including data discovery services, data models, and actual marketplace.

Keywords: data spaces; machine learning platforms; reference architecture**;**