2. pielikums

Intelektuālā īpašuma “Datu ezers” izsoles noteikumiem



**Objekta apraksts**

**IZSOLE:** Pētījumu rezultātā radītā intelektuālā īpašuma komercializācija Latvijas Universitātē (LU) notiek saskaņā ar Zinātniskās darbības likuma 39.5 pantu.

LU izsludina rakstisku intelektuālā īpašuma, kuru veido vienas zinātības “Datos balstīta infrastruktūra (Datu ezers), kas nodrošina datu un saskarnes risinājumus datu reģistrācijai, apstrādei un analīzei. Infrastruktūra ir noformēta kā sagatave, ko var uzstādīt uz Microsoft Azure mākoņpakalpojuma vides konkrētu projektu realizācijai”, izsoli.

**IZGUDROJUMA PĀRSKATS:**

**Izsoles objekts:** sastāv no zinātības, kas izstrādāta projekta Nr. KC-PI-2017/102 “Uz genoma un veselības datiem balstītas vēža prognozēšanas infrastruktūras izveide” (LU reģistrācijas numurs ESS2018/280) ietvaros. Intelektuālais īpašums paredzēts datos balstītas infrastruktūras izmantošanai.

Zinātības aprakstā ir iekļautas jaunas un nepublicētas metodes, scenāriji un shēmas mākoņveida datu ezeram. Zinātība satur detalizētus aprakstus, shēmas, rasējumus un programmu kodus datos balstītai infrastruktūrai.

Detalizētāku informāciju par izgudrojumu un sākotnējo komercializācijas stratēģiju var sniegt projekta vadītāja: Signe Bāliņa, signe.balina@lu.lv.

* Cenas diapazons: atbilstoši pretendenta piedāvājumam;
* Atslēgas vārdi: mākoņpakalpojumi, datu ezers, datos balstīta infrastruktūra;
* Pētījumu veica: Latvijas Universitātes Inovāciju centrs;
* Kontakta telefons: +371 29252365 (Signe Bāliņa);
* Kontakta e-pasts: signe.balina@lu.lv.

**KOPSAVILKUMS:**

Ir izveidota datos balstīta infrastruktūra (Datu ezers), kas nodrošina datu un saskarnes risinājumus datu reģistrācijai, apstrādei un analīzei. Infrastruktūra ir noformēta kā sagatave, ko var uzstādīt uz Microsoft Azure mākoņpakalpojuma vides konkrētu projektu realizācijai. Kā arī tiek nodrošināta infrastruktūras konfigurēšanas un lietošanas dokumentācija.

**IZGUDROJUMA PRODUKTS:** datos balstīta infrastruktūras vēža prognozēšanai apraksts

**IĪ galveno komponenšu īss apraksts**

Datu ezera risinājuma arhitektūra var tikt pielietota gan pētījumos, gan precīzijas medicīnā, un atbalsta šādus piecus lietojuma gadījumus:

* Datu ieguve;
* Datu apstrāde un anonimizēšana;
* Datu analīze;
* Datu nodrošināšana pētniecībai un precīzijas medicīnai.

Datu ieguve tiek īstenota, integrējot dažāda veida datus no dažādiem datu avotiem. Lai iegūtu sākotnējos un papildu pacientu datus, ir izstrādāts adaptera modelis veselības datu iegūšanai. Arhitektūra ļauj arī izstrādāt papildu adapteru modeļus, lai nodrošinātu datu iegūšanu no šobrīd neidentificētiem avotiem. Infrastruktūra ļauj piemērot kā manuālu datu ieguvi (augšupielādējot datu failus datu ezera infrastruktūrā), kā arī automātisku vai pusautomātisku datu ieguvi izmantojot infrastruktūrā iestrādātos REST servisus un “Loģiskās aplikācijas” komponentes dažāda līmeņa automatizācijas realizācijai. Iegūto datu uzglabāšanai tiek izmantota Azure Blob Storage komponente.

Datu apstrāde ietver dažādus uzdevumus: datu tīrīšanu, savstarpēju savienošanu, anonimizāciju un uzglabāšanu dažādos datu saglabāšanas, apstrādes un analīzes scenārijos. Atkarībā no projekta prasībām, var tikt piemēroti visi vai daļa no datu apstrādes uzdevumiem. Uzdevumu realizācijai tiek piedāvātas dažādas komponentes: REST servisu un Loģisko aplikāciju kombinācija, SQL datu bāze, Azure DataBricks. Lēmumu par konkrētu kompoenšu izmantošanu pieņem infrastruktūras izmantotāji, atbilstoši savām kompetencēm un konkrētām datu apstrādes vajadzībām.

Datu analīzes var tikt veikta izmantojot specializētās komponentes datu ezera infrastruktūras ietvaros, tādas kā Machine Learning Studio, Azure Batch un Microsoft Genomics API (sekvenēto genoma datu sekundārajai analīzei). Tāpat datu analīzēvar tikt izmantotas Hadoop / Spark tehnoloģijas, kas darbojas Azure HDInsight klasteros un Azure Stream Analytics un Data Lake Analytics. Kā arī datu analīze var tikt realizēta sadarbībā ar augstas veiktspējas skaitļošanas (HPC) pakalpojumiem.

Datu sagatavošana pētniecībai un precīzijas medicīnai var tikt nodrošināta pēc infrastruktūrā reģistrēto datu apstrādes un analīzes. Iegūtais rezultāts var tikt kopīgots, nodrošinot piekļuvi ārējiem sadarbības partneriem. Šīs funkcionalitātes nodrošināšanai tiek izmantota Azure Blob Storage komponentes kopīgošanas funkcija.

Datu ezera infrastruktūra ietver šādu galveno funkcionalitāti:

* autentifikācija un autorizācija,
* lietotāju pārvaldība, nodrošinot atbalstu vairākām lietotāju lomām,
* izmantojot SQL komponenti datu apstrādei var tikt nodrošinātas šādas funckionālās iespējas:
* reģistrēto pētniecības projekta datu reģistrāciju, parametru koriģēšanu un projekta deaktivēšanu,
* reģistrēto dokumenta veidu pārvaldība - to reģistrāciju, parametru koriģēšanu un deaktivēšanu (dokumenta veids nodrošina iespēju glabāt informāciju par dokumenta formātu, struktūru un transformācijām, ļaujot no ienākošā dokumenta iegūt atsevišķas datu vienības),
* reģistrēto dokumenta validācija, izmantojot sistēmā reģistrētos dokumentu veidus,
* reģistrēto dokumentu datu transformāciju veikšana, izmantojot dokumenta veida transformāciju, iegūstot atsevišķas datu vienības un reģistrējot iegūtās datu vienības normalizētā datu krātuvē,
* atļauju reģistrācija parauga datu izmantošanai vienā vai vairākos pētniecības projektos, esošo atļauju parametru korekcija un esošo atļauju slēgšana,
* specializēta datu krātuve normalizētu parauga datu glabāšanai, nodrošinot iespēju papildināt parauga datus ar jaunām datu vienībām, veikt esošu datu vienību vērtību korekciju, kā arī datu vienību statusa maiņu (dzēšanu),
* kā alternatīva SQL komponentei datu apstrādei ir DataBricks komponente, kas ļauj:
* var tikt veidotas augstas veiktspējas darba plūsmas izmantojot dažādas valodas: python, scala, R,
* piekļūt pie uzkrātiem datiem Azure Blob Storage komponentē,
* veikt nepieciešamo datu apstrādi,
* apstrādāto datu novietošana atpakaļ Azure Blob Storage komponentē;
* audu parauga ģenētiskās sekvencēšanas rezultāta datnes reģistrācija, sekvencēšanas rezultāta datnes sekundārā analīze, izmantojot Microsoft Genomics servisus,
* atbalsta rīks datu importam no Excel failu formāta, veicot nepieciešamās darbības dokumentu veidņu ģenerēšanai, kā arī dokumentu (individuālo ierakstu) ģenerēšanai
* paplašināts ģenētiskā materiāla sekundārās analīzes atbalsta risinājums, izmantojot Azure Databricks un Cromwell,
* Azure Machine Learning Studio integrācija datu ezerā,
* pētniecības datu, kas tiek glabāti ārējā datu avotā un ir pieejami caur REST servisiem (piemēram, HPC datu krātuve), uzraudzība un transportēšana,
* pētniecības datu, kas tiek glabāti ārējā datu avotā un ir sasaistīti ar Galaxy serveri, uzraudzība un aprakstīšana,
* datu kopu raksturojošo datu (metadatu) ģenerēšana,
* datu filtrēšanas pieprasījumu apstrāde.

**IZGUDROJUMA MĒRĶIS UN BŪTĪBA:**

Izgudrojuma mērķis ir būtiski uzlabot vēža prognozēšanas iespēju.

**INTELEKTUĀLĀ ĪPAŠUMA AIZSARDZĪBA:**

Zinātība ir rakstiski fiksēta kā apraksts, aprakstā ir iekļautas jaunas un nepublicētas metodes, scenāriji un shēmas mākoņveida vēža diagnosticēšanai. Zinātība satur detalizētus aprakstus, shēmas, rasējumus un programmu kodus.

**PAPILDU INFORMĀCIJA:**

* Produkts – datos balstīta infrastruktūra vēža diagnostikai
* Viena produkta masu izgatavošanas izmaksas ir sākot no 500 eiro.

**INTELEKTUĀLĀ ĪPAŠUMA DOKUMENTĀCIJA** kurā ietilpst:

Intelektuālā īpašuma izstrādes apraksts, tehniskā dokumentācija, shēmas, rasējumi, programmas kodi.