**Annex 2**

Graphical user interface, text, application

Description automatically generated

**Description of the Object**

**AUCTION:** The commercial use (commercialization) of intellectual property resulting from research at the University of Latvia (LU) takes place according to Article 39.5 of the Law on Scientific Activity.

The LU announces a written auction of intellectual property consisting of one know-how “Data-driven infrastructure (Data Lake) providing data and interface solutions for data registration, processing and analysis. The infrastructure is designed as a template that can be installed on a Microsoft Azure cloud service environment for the implementation of specific projects”.

**OVERVIEW OF THE INVENTION**

**Object of the auction**: consists of a know-how developed under Project No. KC-PI-2017/102 “Development of cancer forecasting infrastructure based on genome and health data” (LU registration number ESS2018/280). The intellectual property is intended for the use of data-driven infrastructure.

The description of the know-how includes novel and unpublished methods, scenarios and schemes for a cloud-based data lake. The know-how contains detailed descriptions, schemes, drawings and program codes for the data-driven infrastructure.

More detailed information on the invention and the initial commercialisation strategy can be provided by the Project Manager: Signe Bāliņa, [signe.balina@lu.lv](mailto:signe.balina@lu.lv).

* Price range: according to the applicant's offer;
* Keywords: cloud services, data lake, data-driven infrastructure;
* The study was carried out by: Innovation Center of the University of Latvia;
* Contact phone: +371 29252365 (Signe Bāliņa);
* Contact e-mail: [signe.balina@lu.lv](mailto:signe.balina@lu.lv).

**SUMMARY:**

Data-driven infrastructure (Data Lake) providing data and interface solutions for data registration, processing and analysis has been developed. The infrastructure is designed as a template that can be installed on a Microsoft Azure cloud service environment for the implementation of specific projects. Documentation on the configuration and use of the infrastructure is also provided.

**PRODUCT OF THE INVENTION:** description of data-driven infrastructure for cancer forecasting.

**Brief Description of the Main Components of the IP**

The Data Lake solution architecture can be applied in both research and precision medicine and supports the following five cases of use:

* Data extraction;
* Data processing and anonymisation;
* Data analysis;
* Providing data for research and precision medicine.

Data extraction is implemented by integrating different types of data from different data sources. An adapter model for health data extraction has been developed to extract baseline and follow-up patient data. The architecture also allows the development of additional adapter models to enable data extraction from currently unidentified sources. The infrastructure allows for manual data extraction (by uploading data files to the data lake infrastructure), as well as automatic or semi-automatic data extraction using REST services and Logical Application components embedded in the infrastructure to implement different levels of automation. The Azure Blob Storage component is used to store the extracted data.

Data processing includes various tasks: data cleaning, cross-referencing, anonymisation and storage in different scenarios of data storage, processing and analysis. Depending on the project requirements, all or part of the data processing tasks may be applied. Various components are offered for the implementation of the tasks: combination of REST services and Logical Applications, SQL database, Azure DataBricks. The decision on the use of specific components is made by the users of the infrastructure, according to their competences and specific data processing needs.

Data analysis can be performed using specialised components within the data lake infrastructure, such as Machine Learning Studio, Azure Batch and Microsoft Genomics API (for secondary analysis of sequenced genomic data). Also, Hadoop / Spark technologies running on Azure HDInsight clusters and Azure Stream Analytics and Data Lake Analytics can be used for data analysis. Data analysis can also be carried out in collaboration with High Performance Computing (HPC) services.

Data preparation for research and precision medicine can be provided after processing and analyzing the data recorded in the infrastructure. The resulting output can be shared by providing access to external partners. To provide this functionality, the sharing feature of the Azure Blob Storage component is used.

The data lake infrastructure includes the following core functionality:

* authentication and authorization,
* user management, providing support for multiple user roles,
* the following functionalities can be provided using the SQL component for data processing:
* registration of registered research project data, parameter adjustment and project deactivation,
* management of registered document types – their registration, parameter adjustment and deactivation (the document type provides the option to store information on the format, structure and transformations of the document, allowing the extraction of individual data units from the incoming document),
* validation of the registered document using the document types registered in the system,
* performing data transformations of registered documents using document type transformation, extracting individual data units and registering the extracted data units in a normalised data library,
* registration of authorisations for the use of sample data in one or more research projects, adjustment of parameters of existing authorisations and closure of existing authorisations,
* a dedicated data library for the storage of normalised sample data, allowing the addition of new data units to the sample data, the adjustment of values of existing data units and the change (deletion) of the status of data units,
* as an alternative to the SQL component for data processing is the DataBricks component that allows:
* high-performance workflows can be created using different languages: python, scala, R,
* access to data stored in the Azure Blob Storage component,
* perform the necessary data processing,
* placing the processed data back into the Azure Blob Storage component,
* registration of the genetic sequencing result file of the tissue sample, secondary analysis of the sequencing result file using Microsoft Genomics services,
* support tool for importing data from Excel, performing the necessary operations for generating document templates, as well as for generating documents (individual records)
* extended support for secondary analysis of genetic material using Azure DataBricks and Cromwell,
* integration of Azure Machine Learning Studio into the data lake,
* monitoring and transport of research data stored in an external data source and accessible via REST services (e.g. HPC data library),
* monitoring and describing research data stored in an external data source and linked to the Galaxy server,
* generation of data (metadata) characterizing the data set,
* processing data filtering requests.

**OBJECTIVE AND PURPOSE OF THE INVENTION:**

The objective of the invention is to substantially improve cancer forecasting.

**PROTECTION OF INTELLECTUAL PROPERTY:**

The know-how is expressed in writing as a description that includes novel and unpublished methods, scenarios and schemes for a cloud-based cancer diagnosis. The know-how contains detailed descriptions, schemes, drawings and program codes.

**ADDITIONAL INFORMATION:**

* Product – data-driven infrastructure for cancer diagnosis.
* The cost of mass production of a single product starts at 500 Eur.

**INTELLECTUAL PROPERTY DOCUMENTATION** comprising:

Intellectual property development description, technical documentation, schemes, drawings, program codes.