ROMANIA

Reimbursable Advisory Services Agreement on
Romania Capacity Building for Statistics (P167217)

OUTPUT No. 4.2 a

Recommendations and guidance provided to Recipient in running of the integrated IT system with its components, hardware (including tablets) and software (licenses), including data storage, protection and security of data for running the activities developed under the RAS (GIS, PHC2021, GAC2020, inter-census periods, SICCA)

September 2022

Revised version December 2022

THE WORLD BANK
IBRD • IDA | WORLD BANK GROUP

Project co-financed from the European Social Fund through the Operational Programme for Administrative Capacity 2014-2020

Disclaimer

This report is a product of the staff of the World Bank. The findings, interpretation, and conclusions expressed in this paper do not necessarily reflect the views of the Executive Directors of the World Bank or the governments they represent. The World Bank does not guarantee the accuracy of the data included in this work and does not assume responsibility for any errors, omissions, or discrepancies in the information, or liability with respect to the use of or failure to use the information, methods, processes, or conclusions set forth. The boundaries, colors, denominations, and other information shown on any map in this work do not imply any judgment on the part of the World Bank concerning the legal status of any territory or the endorsement or acceptance of such boundaries.
This report does not necessarily represent the position of the European Union or the Romanian Government.

This report has been delivered in September 2022 and revised in December 2022, under the Reimbursable Advisory Services Agreement on Romania Capacity Building for Statistics (P167217) signed between the Romanian National Institute of Statistics and the International Bank for Reconstruction and Development on September 17, 2019. It is part of Output 4.2 under the above-mentioned agreement.

Table of Contents

List of Acronyms

APIA	Agency for Payments and Intervention in Agriculture
ATU	Administrative Territorial Units
CAPI	Computer-assisted personal interview
CASS	Computer Assisted Survey System
CAWI	Internet-computer-assisted web interview
CNP	Personal Identification Number (Personal Numeric Code)
DTS	Territorial Statistics Directorates
ESS	European Statistical System
ESS-MH	ESS Metadata Handler
EU	European Commission
EU INSPIRE	Infrastructure for Spatial Information in the European Community
GA	Grid Application
GAC	General Agriculture Census
GAMSO	Generic Activity Model for Statistical Organizations
GSBPM	Generic Statistical Business Process Model
GEOLOC	Geographical Location Software Application
GIS	Geographic Information System
GPS	Global Positioning System
IIS	Internet Information Services
INS	National Institute of Statistics
IT	Information Technology
PHC2021	Population and Housing Census 2021
RAS	Reimbursable Advisory Services
SDC	Statistical Disclosure Control
SICCA	Solution for the Architecture and Computation of the Costs of Products and Activities of INS
STS	Special Telecommunication Service
VPN	Virtual Private Network

List of figures

List of Tables

Introduction

This report covers ***recommendations and guidance provided to the Recipient in running the integrated IT system with its components, hardware (including tablets) and software (licenses), including data storage, protection and security of data for running the activities developed under the RAS (GIS, PHC2021, GAC2020, inter-census periods, SICCA).*** This deliverable is part of Output 4.2 under the Reimbursable Advisory Services (RAS) on *Romania Capacity Building for Statistics (P167217)*. The project is implemented by the National Institute of Statistics with support from the World Bank.

The report presents an overview of the implemented IT integrated system with its components developed under the RAS: the GAC, the PHC, the SICCA, and the system for the intercensus periods. The design of the integrated IT system of statistics and actual implementation reflect the capabilities and production structure in view of satisfying the country's statistical data needs for public institutions, private sector and other organizations, citizens at large, in term of quality and reliability as set by the standards of the domain.

The integrated IT hardware and software system, developed under this project, has enabled the implementation of the two key censuses (GAC and PHC) and has prepared the set-up for the intercensus activities and for the management of activities and costs (SICCA) and it will ensure continuing and sustainable use after the completion of the project, as follows:

- the IT solution (Survey Solutions and R), consisting of data preparation, data production and data dissemination components, for the collection and processing of GAC2020 data will be used for statistical research that must be carried out every 3 years during the intercensual period in the field of agricultural statistics (structural surveys in agriculture);

- the IT solution (Survey Solutions and R/Python), consisting of data preparation, data production and data dissemination components, for the implementation of PHC2021 will be used in the regular process of producing continuous statistics on the resident population (by taking the modules related to extracting data from administrative sources and integrating them with the database for population statistics), as well as by continuing to collect data for buildings and dwellings, including geo-spatial coordinates (GIS);

- the IT solution (Survey Solutions and R) for data collection using multimodal methods (CAPI and CAWI), data validation and processing, editing, imputation and production of tables, protection and data security, for the statistics for the intercensual period will be used for any data collection activities and included in the statistical process to produce statistics with an annual frequency for census-like estimates. The smart electronic questionnaires developed within the project, as well as the knowledge gained through the training sessions and workshops will provide a solid knowledge and skills base regarding the future design of any such surveys.

- the GIS platform, initially used in the process of data collection and processing for PHC2021, will subsequently be integrated into the regular statistical production process for resident population, buildings and housing, statistics that will incorporate geo-spatial coordinates (GIS);

- the methodology for evaluating and ensuring the quality and coverage of the collected data, as well as for the protection / security of the collected data developed within the project for GAC2020 and PHC2021 will be applied (with the appropriate adaptations) for the current statistical research conducted by the INS.

- All solutions are built around the World Bank's open-source Computer Assisted Survey System (CASS) Survey Solutions as well as through R (Shiny) (and Python by case) applications and scripts. The source codes for all these tools had been shared with INS, so it is capable to modify theses scripts and adjust them to future needs.

- the SICCA (Integrated IT Solution for the Architecture and Calculation of Costs of Products and Activities) application developed in compliance with the statistical standards of the Generic Model of Activities for Statistical Organizations (GAMSO) and the Generic Model of Statistical Processes (GSBPM), will be the daily used tool for each employee from the INS and DTS for monitoring and evaluating the activities carried out in order to determine the costs of human resources, time and financial resources for the production of statistical products, but also for each activity or for the groups of activities carried out;

By considering all the above functionalities and their future use, the report summarizes the technical assistance provided to run the integrated IT system designed and implemented through project, respectively:

- IT infrastructure
    - *Output 4.1b - Recommendations to develop the technical documentation in view of organizing the procurement of an integrated IT system including, hardware (including tablets) and software (licenses), including data storage, protection and security of data for running the activities developed under the RAS (GIS, PHC2021, GAC2020, inter-census periods, SICCA)*
- PHC implementation
    - *Output 4.1c - Documented plan for the integrated system for PHC2021 implementation (details how the IT infrastructure implementation for PHC2021 will be carried out)*
    - *Output 7c - Recommendations and best practices for implementing a data management system for PHC2021 geo-spatial data and the actual sectorization*
    - *Output 13 - Recommendations and guidance on documented plan for the integrated system for the PHC2021 implementation (including user management, ID encryption, disaster recovery, maintain system availability in case of breakdown).*
- GAC implementation
    - *Output 14 - Recommendation and best practices for implementing a functional IT system for GAC2020 (Reports on (i) logistics related to data collection according to the optimal combination identified; and (ii) piloting of the GC IT application with report and assistance, and advice on best practices and guidance for the maintenance processes and actions of IT system during the go-live production) including the report on two (2) five-day (5) workshops for statisticians on methodologies and tools for transforming questionnaires in intelligent statistical e-questionnaires"*

    - *Output 7b - Recommendations, best practices and guidance in developing a methodology on customizing an IT solution to geo-reference the agricultural holdings in grid-like statistical units GAC2020.*
- SICCA implementation
    - *Output 11 - Recommendation, advice and best practices for the development and, customization of an IT Integrated Solution for the Architecture and Computation of the Costs of Products and Activities NIS – SICCA.*
- Intercensus surveys implementation
    - *Output 4.2b - Recommendations and guidance on the documented (go-live production) plan for the integrated system for inter-census periods.*

and includes few recommendations (aspects) which will be carried out by INS specialists, their partners (e.g. STS) and vendors of IT infrastructure, in maintaining and exploiting the system in an optimal way/condition, respectively the security of INS hosted IT environment and the STS hosted IT environment for data collection, whenever applicable, and the functional performance and security testing processes to be carried out. The data protection and security, applied for PHC collected data, are detailed in *Oputput 10c: Set of draft statistical census tools using multi-modal methods to promote data protection and security* and complemented *with Output 10b: Technical assistance and best practice recommendations for SDC, data confidentiality, ways to secure micro-data and aggregate data.*

The use of integrated IT system in optimal conditions requires skills on specific domains of INS staff which can be acquired through consistent implementation of multiannual training plan proposed in *Output 12 – Development of a multi-annual training needs plan (2021-2024) for NIS including STDs based on training needs* and by case with the support of specialized services partnered (e.g., STS) or outsourced as the future situations occur.

1. The INS Integrated IT system

The Integrated IT System implemented by the National Institute of Statistics (INS) envisages to sustain the statistical production process for the next 10-year period and comprises the following functional components, including the underlying application and infrastructure elements, respectively:

1. **Data storage, data processing and dissemination for the General Agriculture Census 2020 (GAC2020), including:**
    a. Solution for geo-referencing of agricultural holdings in statistical units of grid type, based on the location, so that the association between the used agricultural area and the corresponding grid code can be realized, according to the provisions of the regulation on integrated statistics at the farm level; the solution will approach different holdings that benefit from subsidies and are included in the APIA list (for which there are GIS coordinates from the APIA source) and the small holdings that are not included in the APIA list (for which the GIS coordinates must be collected).
    b. Processing of data from administrative sources and their conversion into GAC2020 database, as well as pre-filling this information into individual questionnaires. The individual questionnaires were pre-filled according to the information identified in the administrative sources (APIA, sectorization lists based on Agriculture Register, etc.).
    c. Software applications for assuring the data confidentiality and for applying the statistical disclosure control (SDC) techniques.
    d. The entire system includes, also, the CAPI data collection based on questionnaires developed with Survey Solution (data collection servers were hosted at STS).
2. **Data storage, data processing and dissemination for the Population and Housing Census 2021 (PHC2021), including:**
    a. Data processing from administrative sources and their conversion into PHC2021 database, and pre-filling of the individual questionnaires (the individual questionnaires will be pre-filled according to the information identified in the administrative sources).
    b. Solution for geo-referencing the dwellings/buildings meant to be occupied by households, collective units and buildings (including for those having other destinations) in which at least one person has his/her usual residence at the reference date of PHC2021.
    c. Sectorization
    d. Software applications for assuring the data confidentiality and for applying the statistical disclosure control (SDC) techniques.
    e. CAPI data collection based on questionnaires developed with Survey Solution (hosted by STS).
    f. CAWI data collection based on questionnaires developed with Survey Solution (hosted by STS).
3. **Solutions for data processing during Inter-censual periods to produce on annual basis data /indicators similarly to census one (for buildings, dwellings, and persons).**
4. **Sample Surveys (Survey Solutions) - CAPI and CAWI hosted at INS IT infrastructure.**
5. **Integrated IT Solution for the Architecture and Calculation of Costs of Products and Activities (SICCA).**

6. **Geographic Information System (GEOLOC) – provided by a specialized vendor and not part of this project.**

2. Hardware Infrastructure

The description of all components of hardware infrastructure are available in the report of *Output 4.1b - Recommendations to develop the technical documentation in view of organizing the procurement of an integrated IT system including, hardware (including tablets) and software (licenses), including data storage, protection and security of data for running the activities developed under the RAS (GIS, PHC2021, GAC2020, inter-census periods, SICCA).*

The IT system integrates a structure of hardware able to respond to actual and future needs of the statistical production process considering the following requirements:

- processing power - capacity of processes calculation by event or survey
- number of simultaneous users
- response time
- data storage capacity
- data transfer capacity
- operating system and licenses

The system is based on three clusters with identical nodes:

- Virtualization - Data processing (CPU & GPU) – 2 nodes
- Virtualization - Windows (Survey Solutions web – IIS, GEOLOC - GIS Server & Portal) - 2 nodes
- Virtualization – Linux – 4 nodes

The storage needs are satisfied by an external storage device. Servers connect to storage via redundant 16 Gbps Host Based Adapters (HBAs) and to network via redundant 10Gbps Network Interface Cards (NICs). The integrated system connects to INS's LAN via Optical Fiber GBICs on the Integrated Security Appliance and to the STS hosted environment via Site-to-Site VPN terminated, also, on the Integrated Security Appliance, confirmed by STS.

The environments are all connected to the same network (including VLANS where needed) as well as to the Storage Equipment.

INS users (local or from Territorial Statistics Departments connected via INS WAN) would have LAN based access to the environments although for security considerations they should connect via the Firewall / Integrated security appliance.

The high-level view and the logical diagram of the integrated IT system are presented in the figures below (see detailed description in Output 4.1b mentioned above).

Figure 1 - High-level view of IT system

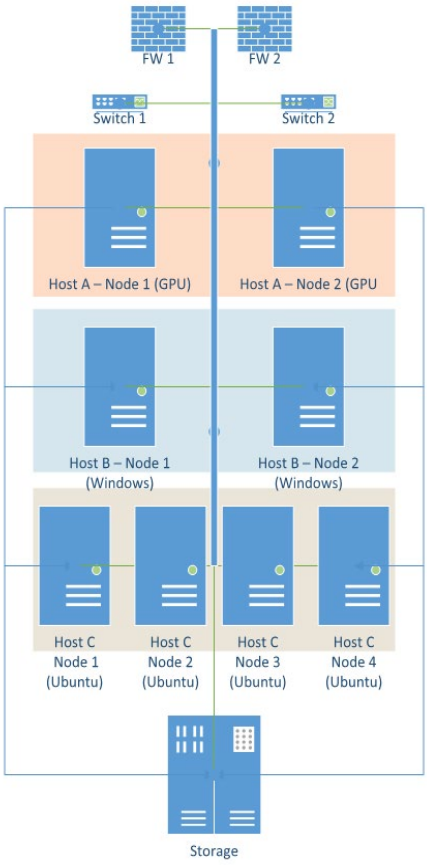
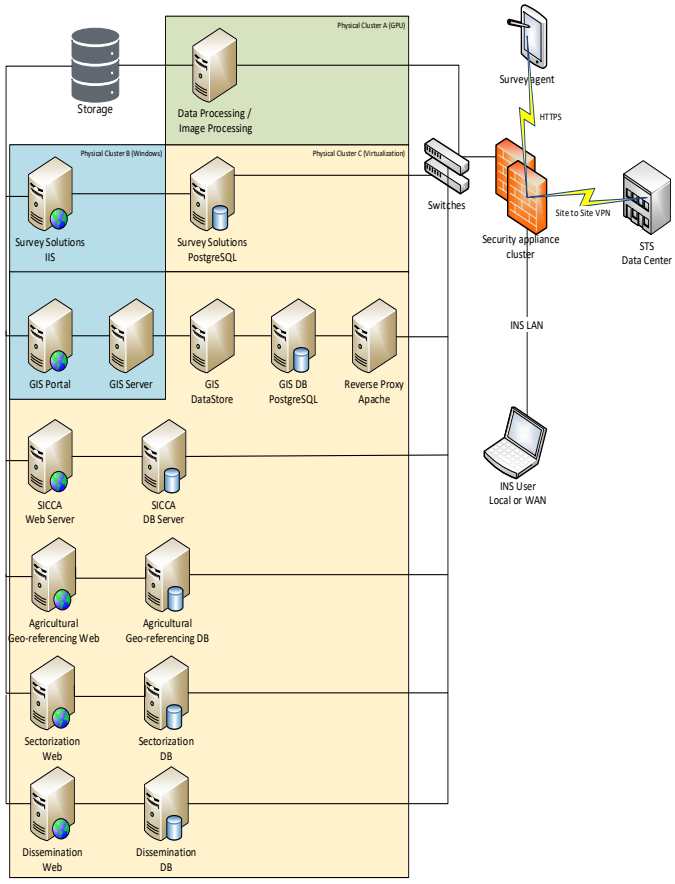


Figure 2 - Logical diagram of IT system



The detailed description of the IT system is presented in *Output 4.1b - Recommendations to develop the technical documentation in view of organizing the procurement of an integrated IT system including, hardware (including tablets) and software (licenses), including data storage, protection and security of data for running the activities developed under the RAS (GIS, PHC2021, GAC2020, inter-census periods, SICCA*.

3. Overview on IT integrated system's applications

The integrated IT system is composed of various hardware and software components (software licenses, applications, databases) that together provide the requisite information infrastructure for both data collection and data processing activities and other complementary activities that INS perform.

The applications and solutions prepared are presented further in brief. The detailed description of them is presented in corresponding output(s) mentioned for each component.

3.1. GAC 2020

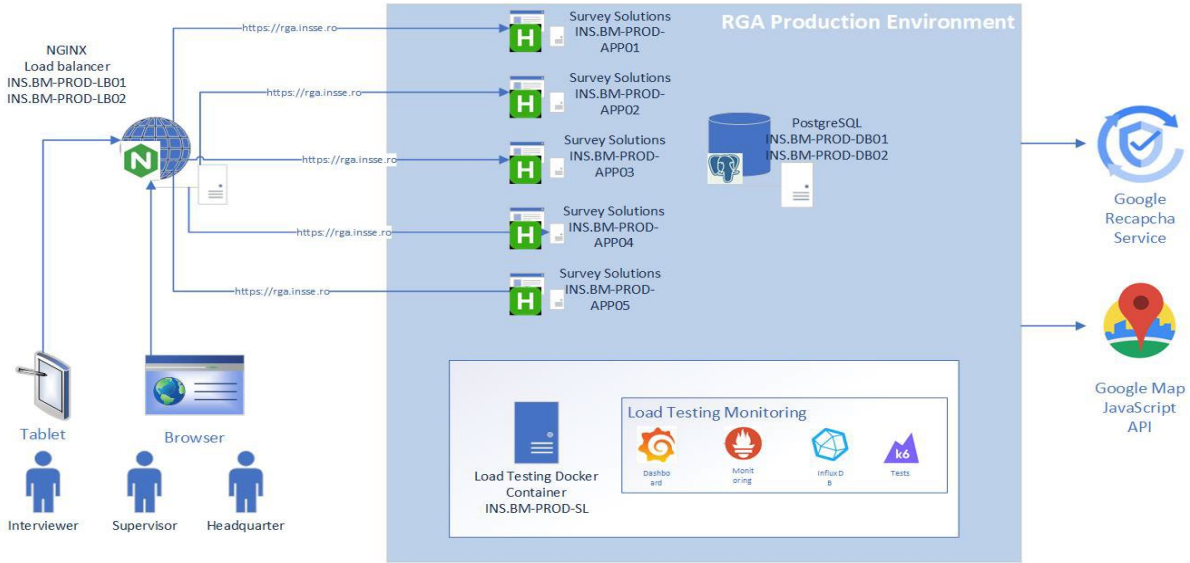
Design and implementation of GAC 2020 is subject of *Output 14 – "Recommendation and best practices for implementing a functional IT system for GAC2020 (Reports on (i) logistics related to data collection according to the optimal combination identified; and (ii) piloting of the GC IT application with report and assistance, and advice on best practices and guidance for the maintenance processes and actions of IT system during the go-live production) including the report on two (2) five-day (5) workshops for statisticians on methodologies and tools for transforming questionnaires in intelligent statistical e-questionnaires"* and of Output 7b: *Report on advisory services provided to the Recipient on the Recommendations, best practices and guidance in developing a methodology on customizing an IT solution to geo-reference the agricultural holdings in grid-like statistical units GAC2020.*

To this output, short presentations of data collection system architecture, monitoring and reporting solutions used during data collection, validation process of the collected data solutions and dissemination of GAC results, including the grid-like distribution of farms, are included below.

3.1.1. GAC data collection

The actual GAC data collection system was built on a three-tier architecture, consisting in a pair of 2 load balancers serving the web requests, 5 application server nodes and a pair of 2 database nodes. Significant metrics are captured from all nodes and sent to a monitoring node.

Figure 3 - GAC Production environment



Source: Authors

The nodes are virtual machines hosted at the STS premises. The following resources were dedicated to each node:

Table 1 - GAC IT resources assigned

Name	Provisioned Space	Guest OS	Memory Size	CPUs	IP Address
INS.BM-PROD-APP01	82.09 GB	Microsoft Windows Server 2019 (64-bit)	32 GB	8	
INS.BM-PROD-APP02	82.09 GB	Microsoft Windows Server 2019 (64-bit)	32 GB	8	
INS.BM-PROD-APP03	82.09 GB	Microsoft Windows Server 2019 (64-bit)	32 GB	8	
INS.BM-PROD-APP04	82.09 GB	Microsoft Windows Server 2019 (64-bit)	32 GB	8	
INS.BM-PROD-APP05	82.09 GB	Microsoft Windows Server 2019 (64-bit)	32 GB	8	
INS.BM-PROD-DB01	1050 GB	Ubuntu Linux 20.04 (64-bit)	128 GB	32	
INS.BM-PROD-DB02	1050 GB	Ubuntu Linux 20.04 (64-bit)	128 GB	32	
INS.BM-PROD-LB01	58.08 GB	Ubuntu Linux 20.04 (64-bit)	8 GB	4	
INS.BM-PROD-LB02	58.08 GB	Ubuntu Linux 20.04 (64-bit)	8 GB	4	
INS.VM-PROD-SL	98.09 GB	Ubuntu Linux 20.04 (64-bit)	48 GB	16	

The high availability, log management and disaster recovery were implemented and are described in Output 14 (i).

The *Annex 1 – GAC data collection information system – As-built report*, provides all details on system implemented.

3.1.2. Monitoring and reporting solutions during the GAC data collection

The monitoring and reporting process for data collection represented a daily task realized for the entire period of the GAC (May 10 – July 31, 2021) as a joint effort of INS and WB teams. The solutions provided consisted in reports for:

a) Daily monitoring
   - Monitoring of data collection
   - Monitoring of Data collection by ATU, supervisors and enumerators
   - Monitoring on UAA, cattle and pigs by ATU (county)

b) Census staff's activity
   - report for census' staff activity
   - enumerator activity and payment report
   - chief enumerator activity and payment report
   - coordinator activity and payment report

The *Annex 2 – Sample reports of daily monitoring* and *Annex 3 – Sample reports of census's staff activity* of Output 14 (i) are providing details on above reports.

3.1.3. Validation process of the GAC collected data

The purpose of the validation process was to support the transmission of the results to the Eurostat, in the predefined format of the Eurofarm file (the same for all EU member states)

and the dissemination of the results to other users (national / international): publications (predefined tables) and tables produced following ad-hoc requests.
The tools for the validation of the data collected consisted in:

- R script on the automatic acceptance and rejection of the questionnaire collected during GAC
- the control tables (the 31 tables developed by INS and WB team) used for data processing-validation and to support the comparison with other data sources (e.g.: GAC2010, ASA 2016, APIA);

The application for GAC2020 Control Tables is functional on the INS IT environment ('*Data Processing Server*' - link **http://xx.x.xxx.xx/GAC_Dashboard**.

The *Annex 4 – Script for automatic acceptance / rejection of questionnaires* and *Annex 6 – Sample of data control tables (TC no.0)* of Output 14(i) provide details on validation process.

3.1.4. Dissemination of the GAC results

For dissemination, the INS and WB analyzed the possibility of a solution as an application - database interface to allow the INS staff to browse the GAC data, filter it based on territorial levels, and group it based on several layers. As a result, the WB team proposed and supported the development of a Shiny[1] web application based on the R programming language, allowing fast data integration, wrangling and reproducibility (see https://shiny.rstudio.com/).

The application is to quickly integrate the raw data (at record level/agriculture holding level) provided by GAC, to summarize key variables at the ATU level, to group the variables and allow fast exports to deliver in the case of data requests from INS or external data users.

The application for GAC data dissemination is deployed on the INS *'Data Processing Server'* - link **http://xx.x.xxx.xx/rga-diseminare/**, and the underlying source code has been shared with INS in *Annex 7 – Script for GAC data dissemination* of Output 14(i).

3.1.5. Geo-referencing the agricultural holdings in grid-like statistical units

The geo-referencing of the agricultural holding in gridded statistical units is the process of association between the location (GPS) of the agricultural holding and the corresponding grid code, according to the specific regulations[2] relevant for integrated farm statistics, including GAC2020 and EU regulations[3].

The GA (Grid Application) allows to visualize large census data sets as well as corresponding spatial analytics either as points or as grid data sets. It is intended to support the work of statisticians with basic GIS knowledge.
The application is available on the INS '*Data Processing server*' (link: **xx.x.xxx.xx/rga-grid/)** and the source code, written in R has been shared with INS.

The annexes of the *Output 7b: Recommendations, best practices andguidance in developing a methodology on customizing an IT solution to geo-reference the agricultural holdings in grid-like statistical units GAC2020*, provide the complete user manual for the application as well as the R script for data matching and corrections of GPS coordinates.

---

[1] Shiny is an R package that makes it easy to build interactive web apps straight from R.

[2] D2.8.I.2 Data Specification on Geographical Grid Systems – Technical Guidelines; INSPIRE, 2014; Directive 2007/2/EC of theEuropean Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE)

[3] EU Regulation 2018/1098

3.2. PHC 2021

The design of PHC 2021 data collection system was subject of *Output 4.1c – Documented plan for the integrated system for PHC2021 implementation (details how the IT infrastructure implementation for PHC2021 will be carried out).* Based on it, the implementation for actual PHC data collection system was carried out and the as-built system is described in *Output 13 – Recommendations and guidance on documented plan for the integrated system for the PHC2021 implementation (including user management, ID encryption, disaster recovery, maintain system availability in case of breakdown)* and includes, also, the dissemination solutions (hyper-cubes, grid-like distribution of PHC data and the tabular dissemination of PHC data). The IT solutions dedicated to the PHC is completed by the sectorization solution described in report of *Output 7c - Recommendations and best practices for implementing a data management system for PHC2021 geo-spatial data and the actual sectorization.*

The integrated IT system of PHC comprises the following IT solutions:

- Data collection:
  - Survey Solutions for CAWI and CAPI data collection process and monitoring solution
  - Self-registration solution for CAWI stage of PHC
- Data processing and validation
- Data editing and production for dissemination: hypercubes, grid distribution and tabular dissemination of PHC data
- (Sectorization for data collection)

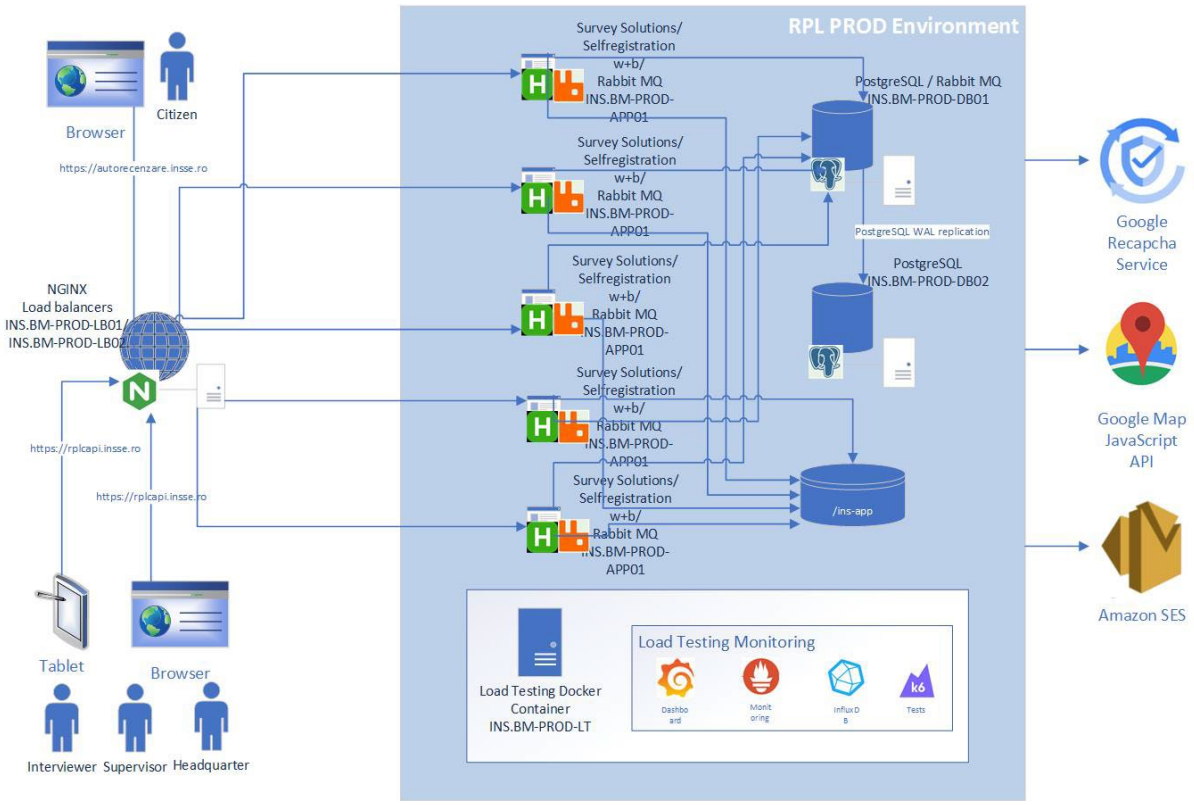
This output only provides a short description of the solution and any details can be found in corresponding subject report.

3.2.1. PHC data collection

The data collection system includes the Survey Solutions application for the CAWI and CAPI data collection process as well as the self-registration solution for CAWI stage of PHC (see PHC production environment figure).

The actual PHC data collection system was built on a three-tier architecture, consisting in a pair of 2 load balancers serving the web requests, 5 application server nodes and a pair of 2 database nodes. Significant metrics are captured from all nodes and sent to a monitoring node.

Figure 4 - PHC Production environment



Source: Authors

The nodes are virtual machines hosted in STS premises. The following resources were dedicated to each node:

Table 2 - PHC IT resources assigned

Name	Provisioned Space	Guest OS	Memory Size	vCPUs	IP Address
INS.BM-PROD-APP01	1000 GB	Microsoft Windows Server 2019 (64-bit)	96 GB	24	
INS.BM-PROD-APP02	82.09 GB	Microsoft Windows Server 2019 (64-bit)	96 GB	24	
INS.BM-PROD-APP03	82.09 GB	Microsoft Windows Server 2019 (64-bit)	96 GB	24	
INS.BM-PROD-APP04	82.09 GB	Microsoft Windows Server 2019 (64-bit)	96 GB	24	
INS.BM-PROD-APP05	82.09 GB	Microsoft Windows Server 2019 (64-bit)	96 GB	24	
INS.BM-PROD-DB01	1050 GB	Ubuntu Linux 20.04 (64-bit)	256 GB	48	
INS.BM-PROD-DB02	1050 GB	Ubuntu Linux 20.04 (64-bit)	256 GB	48	
INS.BM-PROD-LB01	58.08 GB	Ubuntu Linux 20.04 (64-bit)	32 GB	16	
INS.BM-PROD-LB02	58.08 GB	Ubuntu Linux 20.04 (64-bit)	32 GB	16	
INS.VM-PROD-SL	98.09 GB	Ubuntu Linux 20.04 (64-bit)	48 GB	16	

The detailed description of PHC Data collection information system is part of Output 13, as an as-built report.

The PHC data collection monitoring, processing and validation is performed using scripts on R statistical software language deployed on INS IT environment, the ***'Data Processing Server'*** *(actual IP xx.x.xxx.xx)*.

3.2.2. Monitoring and reporting solutions during the PHC data collection

The monitoring and reporting process for data collection represented a daily task realized for the entire period of the PHC (March 15 – July 31, 2022) as a joint effort of INS and WB teams. The solutions provided consisted in reports for:

- Monitoring of data collection during CAWI phase (March 15 - May 27), and CAPI phase (May 31-July 31); and
- Monitoring of Data collection by ATU, supervisors and enumerators.

The analysis of the paradata can use a special purpose application available for INS staff (no credentials are required) on ***Data Processing Server*** path **phc__censustools.**

3.2.3. PHC Data processing and validation

The PHC data processing and validation is performed using scripts R statistical programing language deployed on INS IT environment.

The purpose of the validation process was to prepare the PHC final data and to support the dissemination of results, including the transmission of them to the Eurostat.

The tools for the validation of the data collected consisted in R script on the automatic acceptance of the questionnaire collected during PHC. The post enumeration validation was prepared by the team of INS together with the SDC treatment of final data using the method(s) selected from the ones described in Output 10b: *Technical assistance and best practice recommendations for SDC, data confidentiality, ways to secure micro-data and aggregate data.*

The details of R scripts are subject of Output 13 and are to be presented in report of it.

3.2.4. Data editing and production for dissemination: of PHC results

The PHC data dissemination comprises three specific solutions.

The PHC data hypercubes, mandatory to be published on Eurostat portal follow the guidelines provided in this respect by this organization, will be retrieved from the repository of PHC data organized on INS IT environment (the INS team is preparing the IT environment and install the necessary applications).

The grid distribution of PHC data allows the dissemination of 20 indicators (13+7) on grids (INSPIRE) of 1 sqKm for entire Romania's territory. The geo-referencing of the PHC2021 data in gridded statistical units is the process of association between the location (GPS) of the collected variables and the corresponding grid code, according to the EU Regulation no. 1799/2018. By that is produced and sent to Eurostat a one two-dimensional table that cross-tabulates the set of grid cells against the 13 (+7) selection of categories from the census topic breakdowns specified in Annex I of the Regulation. The disseminated PHC data on grid is prepared by using an R/Shiny statistical software.

The dynamic tables dissemination with PHC results is an application written in R using the Shiny framework used for the dissemination of PHC data aggregated/grouped by variables and at ATU, county, region, and country level.

The details of the underlying R scripts are subject of Output 13 and are presented there and run on *'Data Processing server'*.

3.2.5. Sectorization for PHC

Sectorization in the context of a population and housing census refers to the process of creating segments of the geographic census area with a similar number of census units inside each of these segments. For the PHC2021 this means that the whole census area (= geographic territory of Romania) must be segmented into geographic areas with a population of dwelling units (e.g. 100; 200; 600; ± 20%). The solution for sectorization was developed as a script in R and with the support of some customized C++ functions and was deployed on the INS IT environment (Analysis Server). The script can be accessed/run as a standalone RStudio project through the RStudio Server available on INS environment '*Data Processing Server'* path: ***arhiva2/ins_sectorization***

The script code and the detailed method of using to produce segments are presented in report of *Output 7c - Recommendations and best practices for implementing a data management system for PHC2021 geo-spatial data and the actual sectorization.*

3.3. Inter-census

During the intercensus periods, INS will regular be updating the information necessary to carry out census data on the population and housing, on an annual basis, mainly by collecting data from respondents through Survey Solutions using CAWI and/or CAPI modes, and from administrative sources by updating the registers. It is envisaged a system that allows obtaining up-to-date information on the size of the population and its characteristics at the time of the draw or at a specified reference moment, respectively to have a structured list of units, each of which contains a number of attributes, a regular update mechanism, and a component for producing outputs.

The use of registers to produce data for population and housing is based on an algorithm for regular updating during intercensus. This will use variables and dessagregations data from 22 administrative sources (estimated as potential to use) and with a specific timeline (date and frequency – monthly and/or biannually, biannually, annually) depending on the statistical event and updated attributes.

The algorithm is developed on R Statistical Software, R Shiny and is run on '*Data Processing Server'*. The installation of Survey Solutions data collection system for the intercensus periods on the INS IT environment was performed by INS experts.

The register for population is subject of *Output 4.2b - Recommendations and guidance on the documented (go-live production) plan for the integrated system for inter-census periods* and are presented in the report of it.

3.4. SICCA

Design, installation, configuration and optimization of SICCA is subject of "Recommendation, advice and best practices for the development and, customization of an IT

Integrated Solution for the Architecture and Computation of the Costs of Products and Activities NIS – SICCA".

The purpose of the SICCA is to support the cost management of the INS and it relies on three objectives: (i) the calculation of the direct and indirect costs for all statistical products at the level of generic activities, component/subcomponent and projects, standardized according to GSBPM / GAMSO; (ii) the optimization of the statistical production costs and the determination of the costs; and (iii) the improvement of the INS budget based on evidence / records.

The SICCA is designed and developed to cope with several technical requirements. The application is accessible simultaneously to all 2,000 INS employees (500 employees at the central headquarter and 1,500 employees in the territorial statistical departments, located throughout the entire country). It is a web application accessed only with an internet browser, without requiring additional software like plug-ins or similar, with an interface accessible through the most widely used internet browsers: Microsoft Edge, Internet Explorer, Firefox, Chrome. It is installed in the INS LAN and users connect via the INS WAN.

All the various software components of the application were designed and implemented using open-source technologies. The technology stack consists of:

- Operating system – any major OS is supported, including Linux, OSX and Windows;
- Programming languages – PHP (backend), Javascript/HTML/CSS (frontend);
- Database server – MySQL/MariaDB/PostgreSQL;
- Web Server – Apache/NGINX;

Details on the functionalities of the application are available the annexes provided with the Output 11 - *Recommendation, advice and best practices for the development and, customization of an IT Integrated Solution for the Architecture and Computation of the Costs of Products and Activities NIS – SICCA*, respectively: *Annex 1 - Preliminary (initial) Business and Technical Requirements; Annex 2 - Web administrator manual; Annex 3 - User Manual; Annex 4 – Administrator (backend) Manual; Annex 5 – Database and Nomenclatures; Annex 6 – Source code of application.*

4. Security and Testing considerations of IT environment

4.1. Security

In terms of physical security of IT environment, the INS has its own procedures and decides accordingly for their application. These are completed with the logical security actions (e.g., VPN solution, systems configurations, etc., already in place) to ensure the robustness security of the IT environment.

Physical security of IT infrastructure hosted at STS and dedicated to data collection (and used in the current project) is covered by organizational STS procedures and protocols. Physical access to this environment is only allowed for authorized STS personnel.

Logical access is allowed only to persons which have eligible clearance level, and it is possible only through two factor authentication VPN.

The web exposed applications are subject to security assessment conducted by an STS dedicated team. In the process of data collection for performed censuses, the assessment report was presented by STS team to INS. The recommendations regarding applications functionality and applications configuration are applied accordingly.

4.2. Testing

Functional testing is subject to INS team and its purpose is to assure that all the designed functionalities are covered up by effective checks.

In the data collection process for censuses the performance testing was supported with WB team experts' activities in order to ensure that planned capacity and agreed responses times are met. The ***Annex 1*** describes detailed performance test planning, execution, and results.

The security testing for Internet exposed components of the system is subject of STS experts' activities and are covered up on existing STS – INS protocol and security reports.

5. Annexes

Annex 1 – *PHC Test planning, execution, and results*

Competence makes a difference!
Project selected under the Administrative Capacity Operational Program, co-financed by European Union from the European Social Fund

Annex 1 – PHC Test planning, execution and results

Assumptions

In order to estimate the peak number of concurrent users that the system can handle, we targeted 16000000 users (more than 70% of estimated population) in 60 days with a 20-hour window for registration or interview completion. Based on pilot findings we will develop an estimation model in order to find out the peak number of concurrent users that system should be able to serve.

Pilot findings

In order to calculate the peak number of concurrent users, we use information gathered from pilot. Here are the most important findings from pilot:

Table 1 – Average interview duration

TEAM MEMBER	3/10/2021	3/11/2021	3/12/2021	3/13/2021	3/14/2021	3/15/2021	3/16/2021	AVERAGE
All teams	1 hour, 51 minutes	2 hours, 30 minutes	3 hours, 46 minutes	1 hour, 19 minutes	1 hour, 35 minutes	51 minutes	21 minutes	
supervisor	1 hour, 51 minutes	2 hours, 30 minutes	3 hours, 46 minutes	1 hour, 19 minutes	1 hour, 35 minutes	51 minutes	21 minutes	**1 hour, 35 minutes**

Table 2 – Number of completed interviews

TEAM MEMBER	3/10/2021	3/11/2021	3/12/2021	3/13/2021	3/14/2021	3/15/2021	3/16/2021	AVERAGE	TOTAL
All teams	710	562	479	485	676	931	822	666.43	4,665
supervisor	710	562	479	485	676	931	822	666.43	4,665

You might notice that average completion time of interviews is 95 minutes and total number of completed interviews for the 7 days of the pilot is 4665. We can assume the average duration of the registration will be about 5 minutes, so we can use 100 minutes as overall time to complete both self-registration and interview.

In order to find the peak number of users in a day, we normalized the distribution of users along the 7 days, then along 20 hours. In the real model we have multiple peaks usually on the middle of the weeks and multiple times during a day. By assuming a single peak, we get sure that no other peaks will be higher than this one.

Table 3 – Model to estimate highest peak of distribution by days - pilot

Day	Visitors
1	383
2	736
3	1020
4	1020
5	736
6	383
7	143

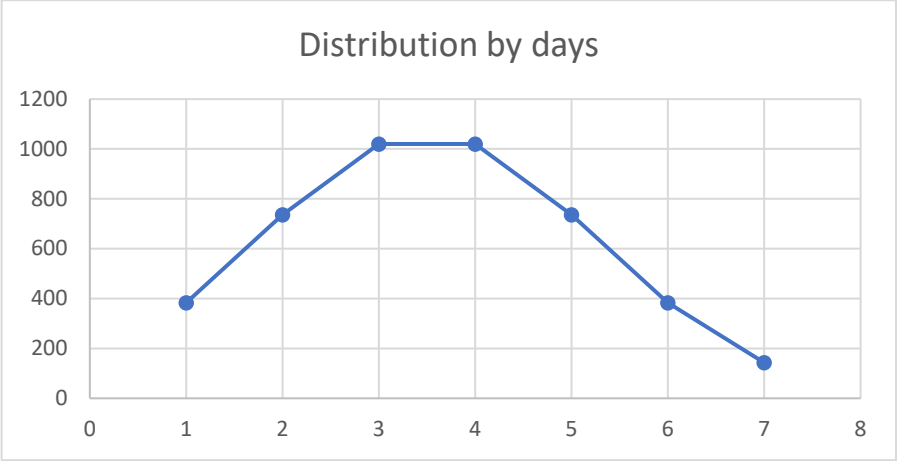
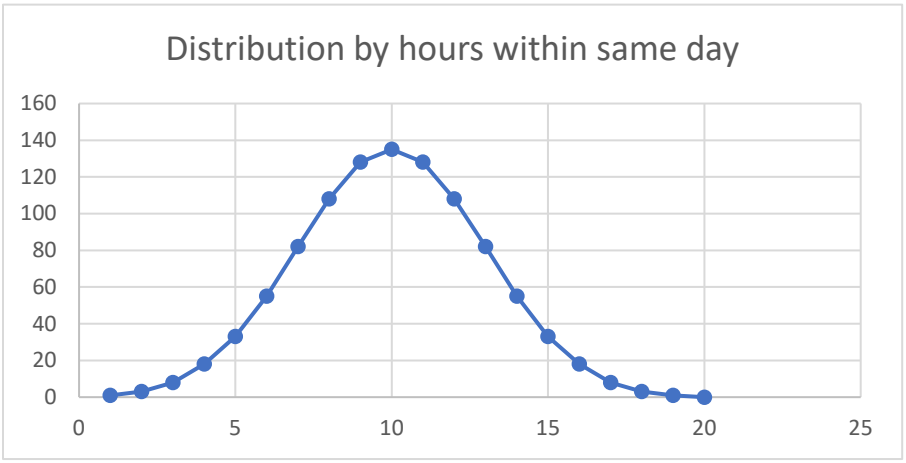


Table 4 – Model to estimate highest peak of distribution by hours - pilot

Hour	Visitors
1	1
2	3
3	8
4	18
5	33
6	55
7	82
8	108
9	128
10	135
11	128
12	108
13	82
14	55
15	33
16	18
17	8
18	3
19	1
20	0

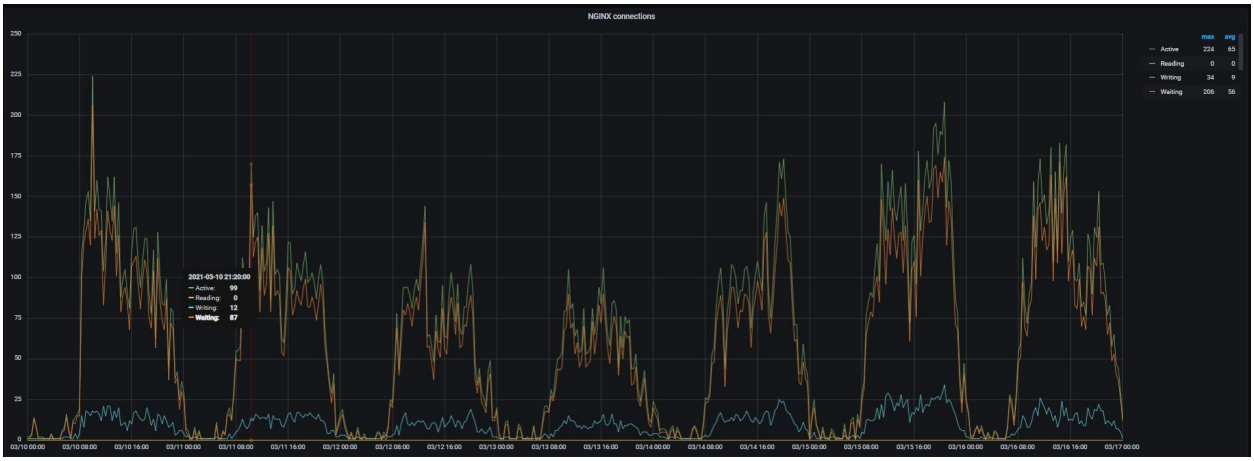


By dividing the number of maximum number of users per day by number of interviews that can be done in an hour multiplied by active time interval in hours, we get the number of peak concurrent users:

Table 5 – Peak number concurrent users

Peak number of concurrent users / registration	**225**
Total number*	4665
Number of days	7
Max number per day	1020
Average completion time in minutes*	100
Active time interval in hours*	20
Max number by hour	135

This calculation provided a number of 225 maximum concurrent users during the pilot. By comparing this estimated number with real data, we validate our estimation model to be used for production. The real number of maximum concurrent users should be determined by the maximum number of connections recorded by NGINX proxies, which in our case is 224:



Note that the concurrent connections on load balancers includes also headquarter users, observers and supervisors, so in reality the real number of concurrent users which are involved in self-registration and interview completion are actually lower.

Production environment calculations

Using the estimation models from the pilot we can calculate peak number of concurrent users for production stage:

Table 6 – Model to estimate highest peak of distribution by days - production

Day	Visitors
1	65658
2	74524
3	84213
4	94739
5	106108
6	118315
7	131341
8	145154
9	159709
10	174944
11	190782
12	207131
13	223885
14	240920
15	258102
16	275283
17	292305
18	309004
19	325208
20	340744
21	355439
22	369124
23	381635
24	392821
25	402542
26	410673
27	417112
28	421772
29	424593
30	425538
31	424593
32	421772
33	417112
34	410673
35	402542
36	392821
37	381635
38	369124
39	355439
40	340744
41	325208
42	309004
43	292305
44	275283
45	258102
46	240920
47	223885
48	207131
49	190782
50	174944
51	159709
52	145154
53	131341
54	118315
55	106108
56	94739
57	84213
58	74524
59	65658
60	57590

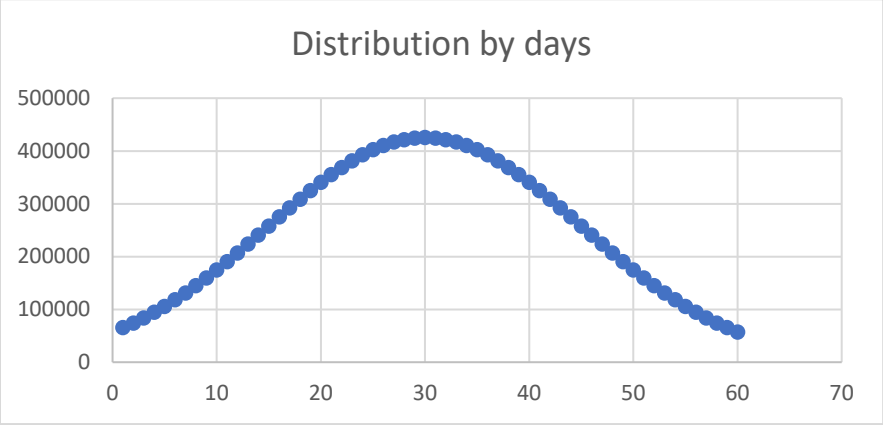
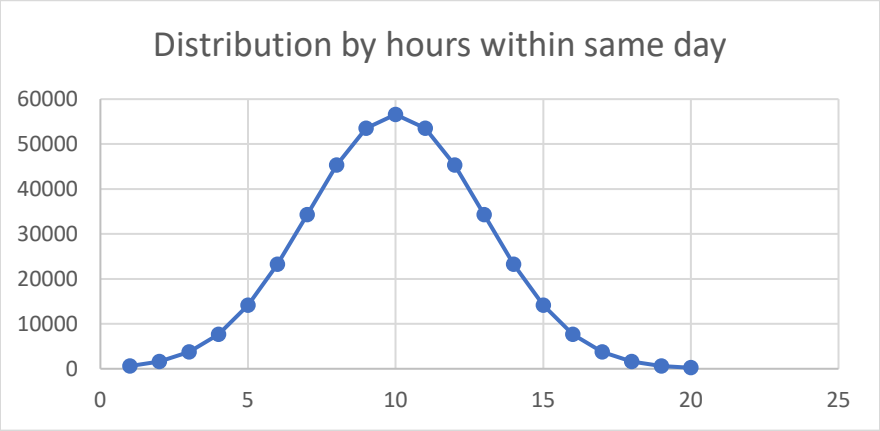


Table 7 - Model to estimate highest peak of distribution by hours - production

Hour	Visitors
1	628
2	1616
3	3719
4	7658
5	14110
6	23264
7	34322
8	45312
9	53530
10	56588
11	53530
12	45312
13	34322
14	23264
15	14110
16	7658
17	3719
18	1616
19	628
20	218



As a result we can determine:

Table 8 – Peak number concurrent users for production environment

Number of peak concurrent users / registration	**94313**
Total number	16000000
Number of days	60
Max number per day	425538
Average completion time in minutes*	100
Active time interval in hours*	20
Max number by hour	56588

Testing scenario – stress test

Testing scenario will assume that a person will register to https://autorecenzare.insse.ro, as a single person with random address (taken from the list of available county/locality/satul/street/streetNo database) with 1 household consisting on 1 person. Then the person will complete the interview by answering to 60 questions with fixed answers.

For stress test, we will consider a 5 minute ± 30 seconds period for self registration and 60 minutes ± 6 minues for completing the interview. Between those periods it will be a 5 minute ± 30 seconds of pause (to simulate waiting for the interview link e-mail).

Four machines will be used to generate total of 100000 users that will self register then complete the interviews. Two phases will be used for user generation behavior for each of the 4 load generators machines:

- Ramp-Up from 0 to 25000 during 1 hour
- Stay at 25000 users for another 1 hour

Then, in order to properly close the interviews, there will be a ramp-down period with variable duration, calculated by considering the maximum time needed to complete the rest of the interviews. This is not relevant for performance testing but is needed for smooth test execution, meaning that all open interviews should be completed.

Test scripts will report also performance metrics related to application behavior:

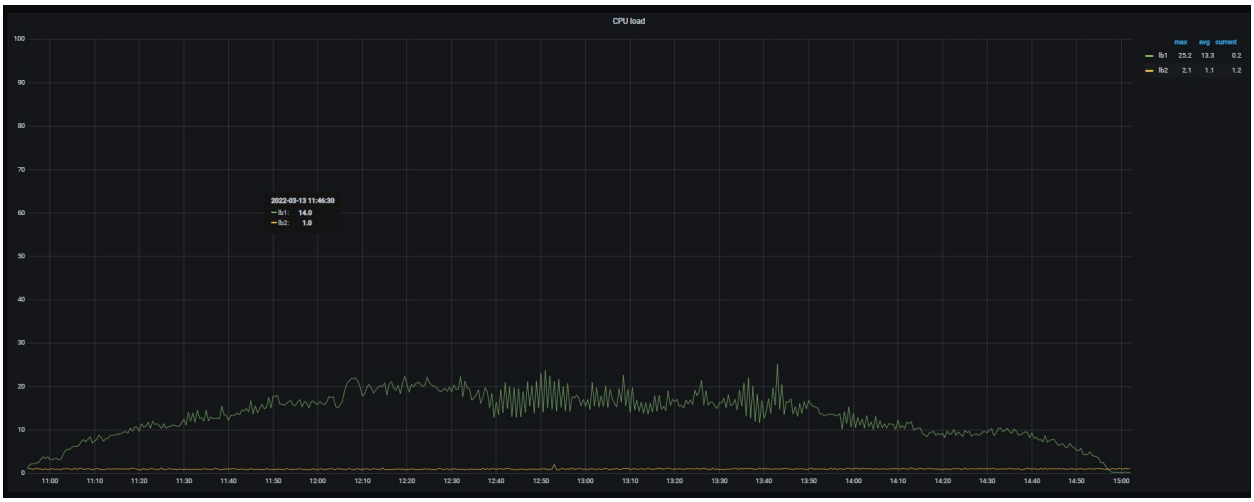
- Request which takes longer that 5 seconds – threshold 99% of responses to be under 5 seconds
- Interviews with questions not answered – threshold 99% of answers to be processed without errors

Test results

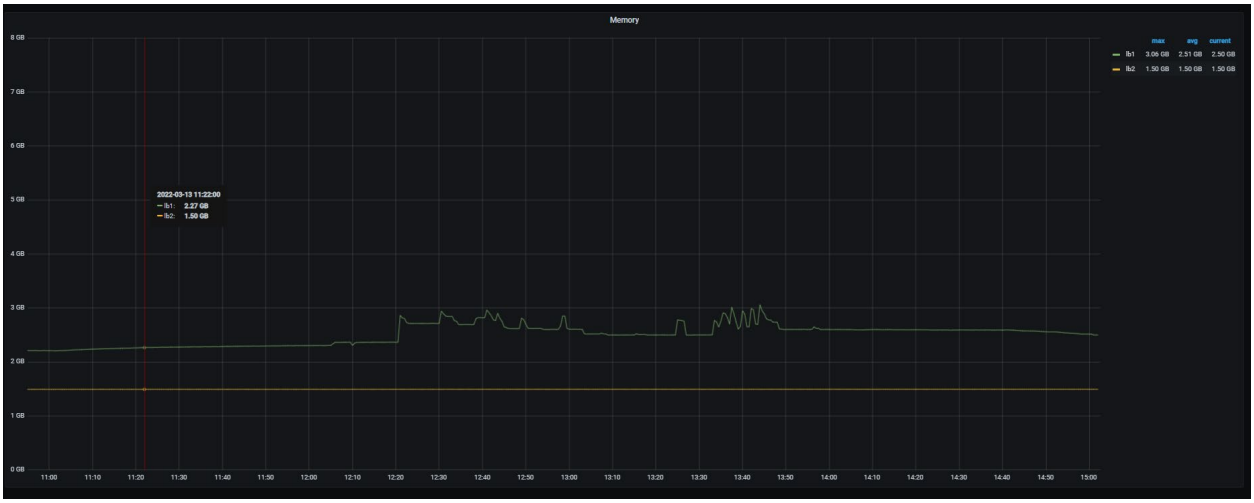
Below you can find the results of the performance stress test and the most important - metrics related to system components. This is the overall monitoring panel:

Load balancers

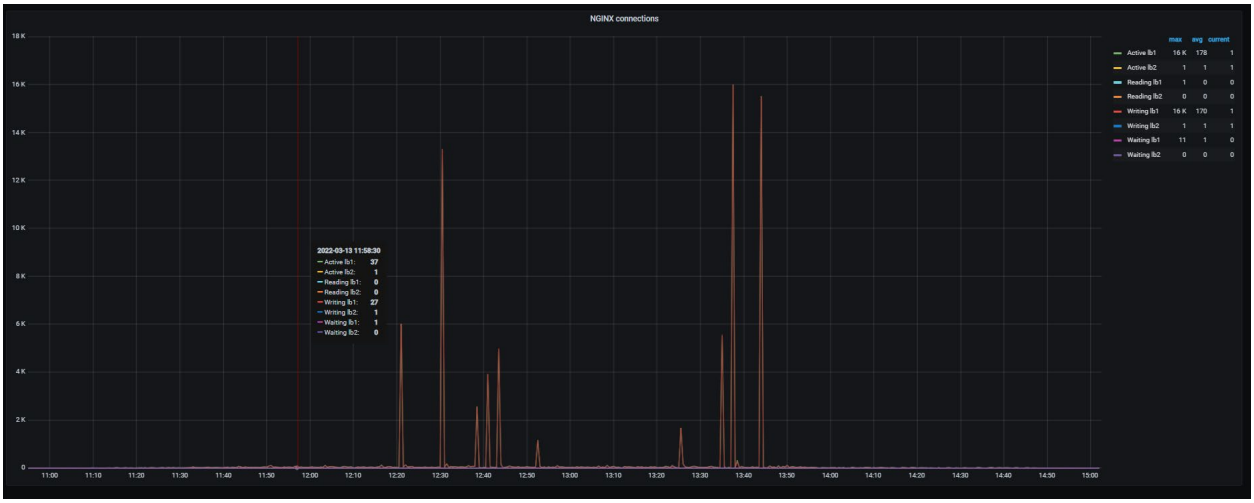
CPU



Memory

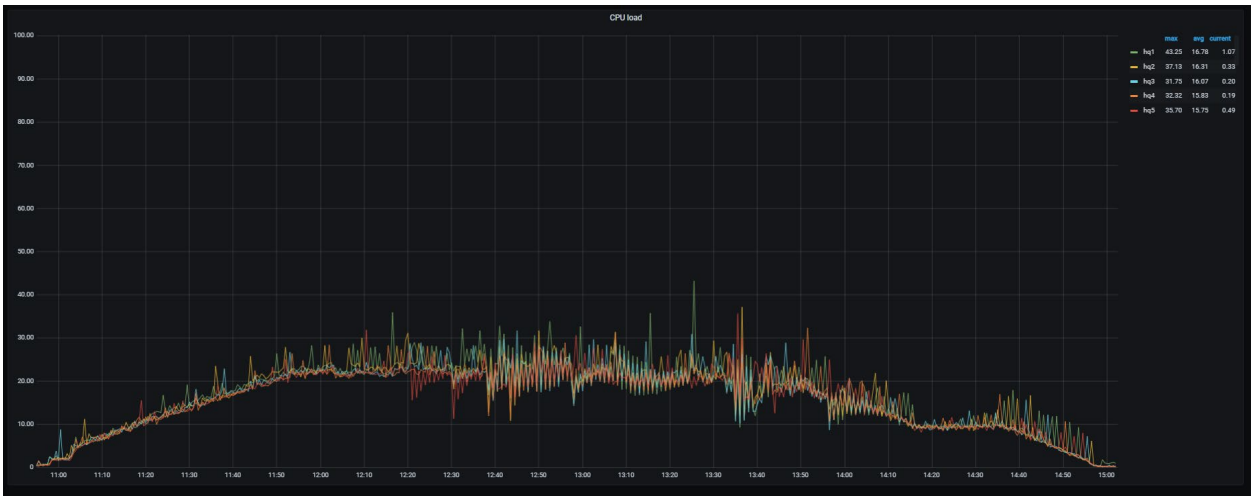


Concurrent connections



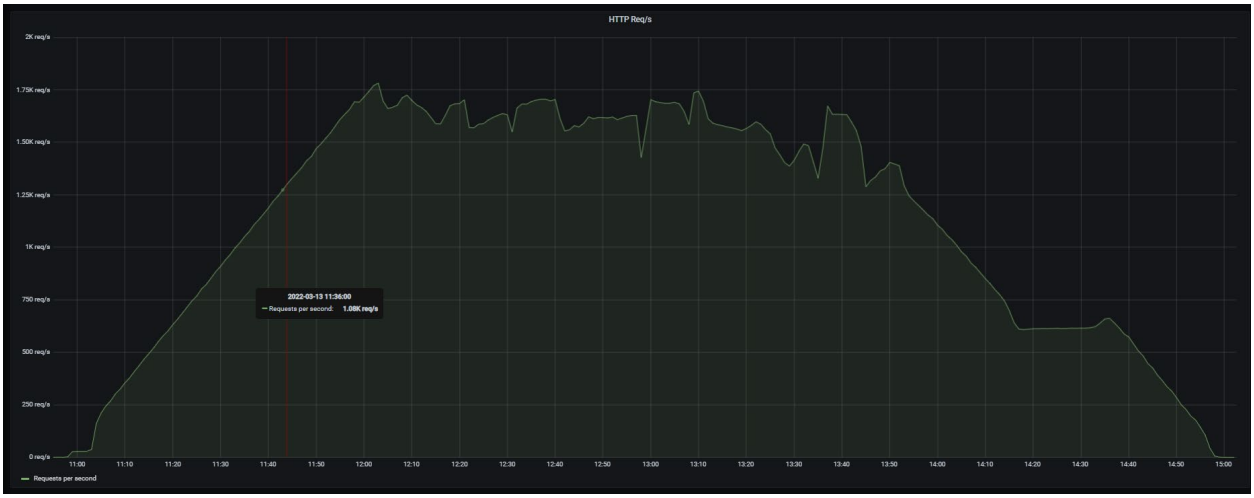
Application servers

CPU



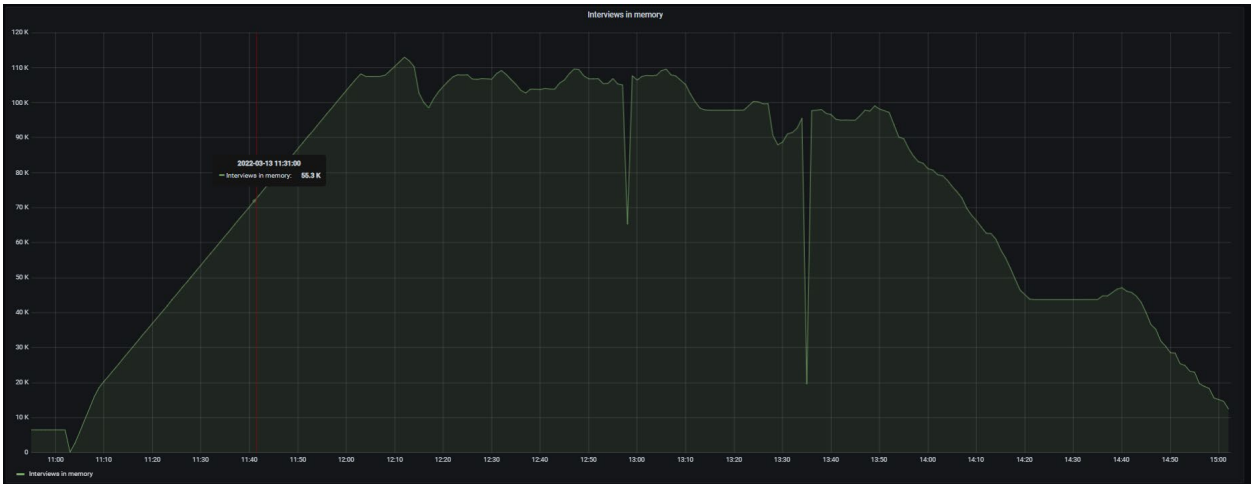
Memory

HTTP req/s

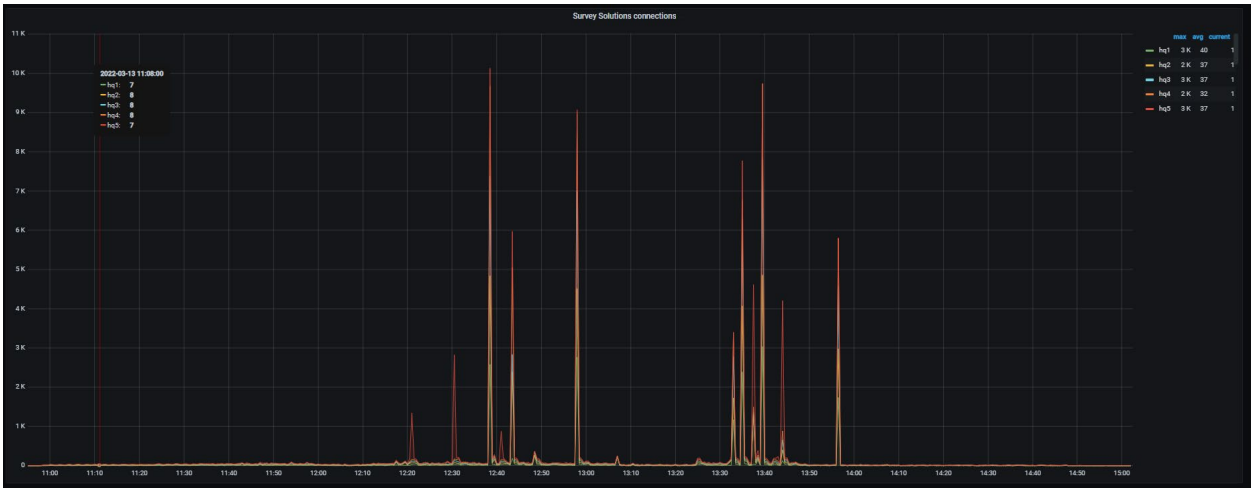


Survey Solutions

Interviews in memory

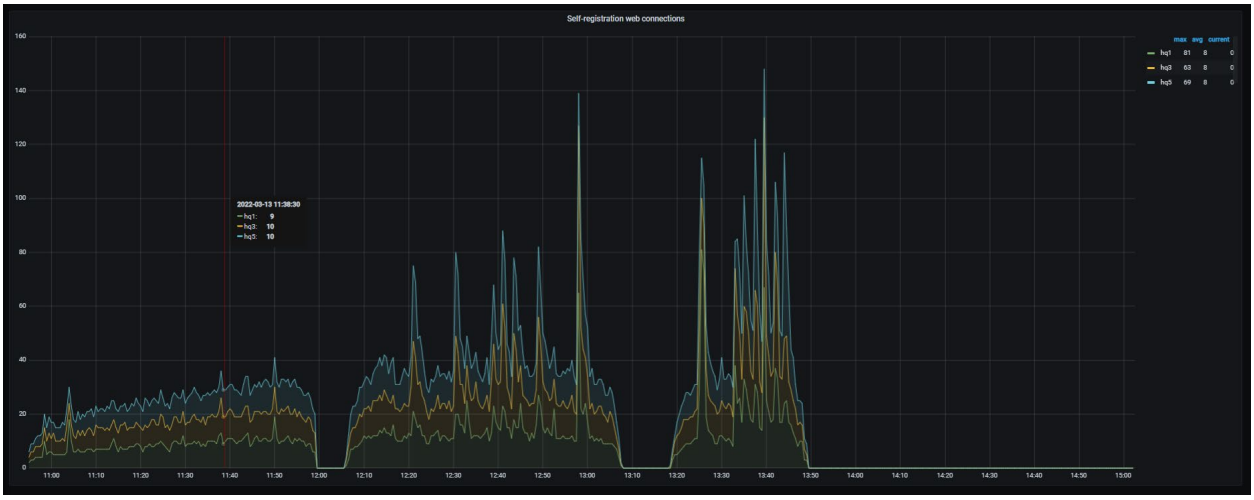


Concurrent connections



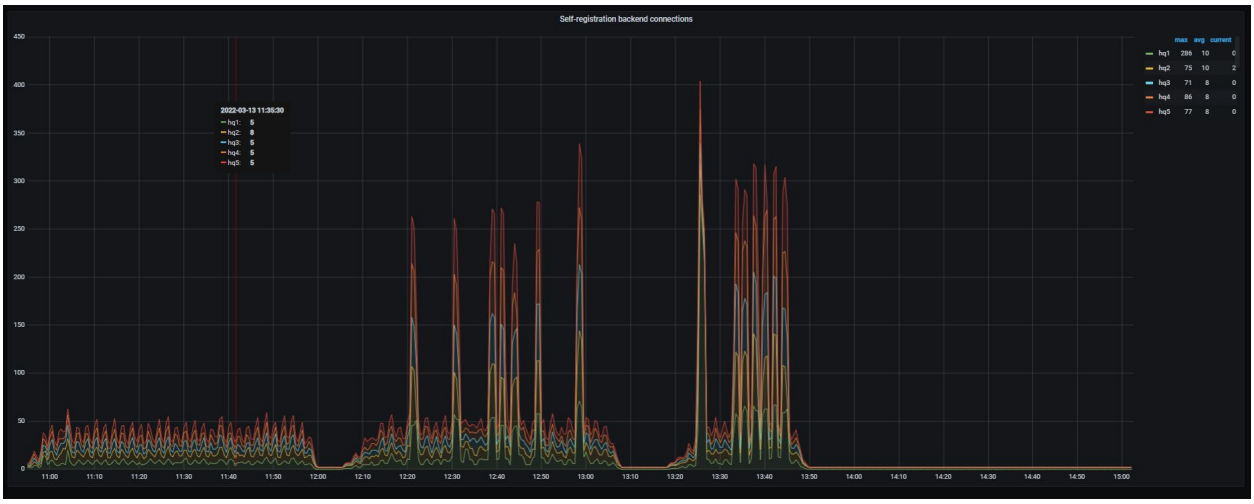
Selfregistration web

Concurrent connections



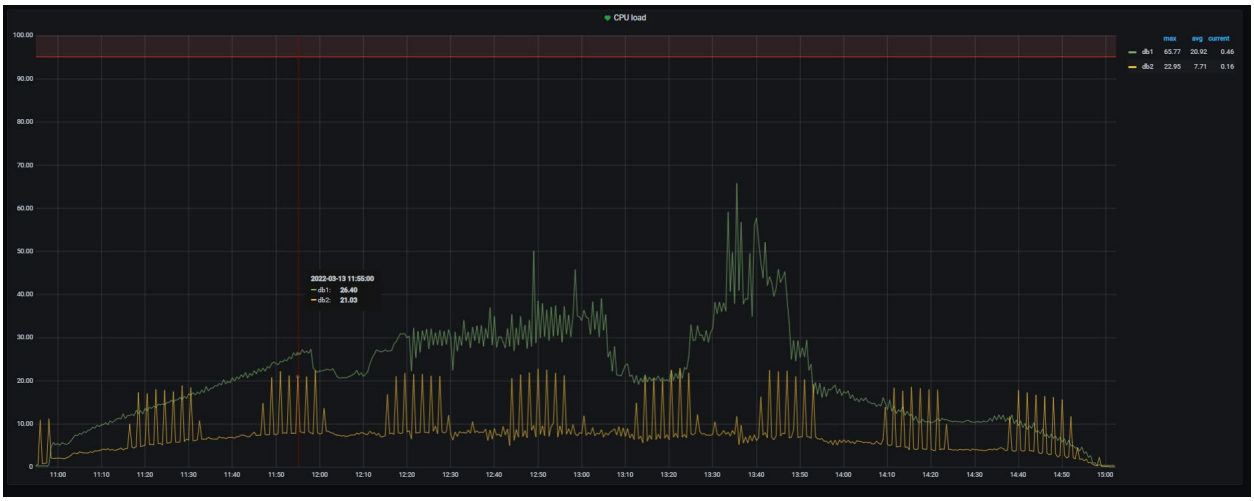
Selfregistration backend

Concurrent connections

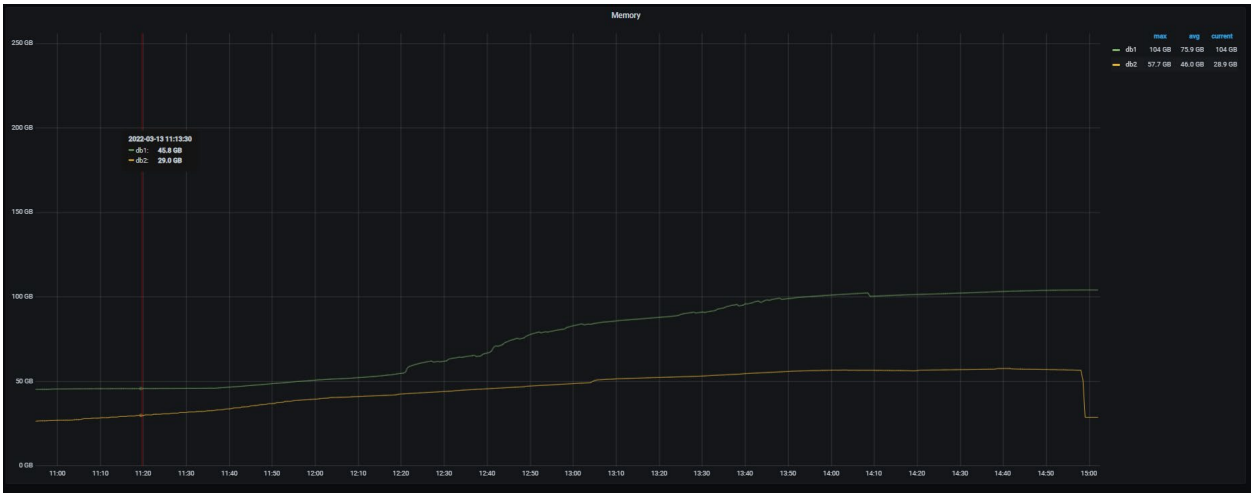


Database servers

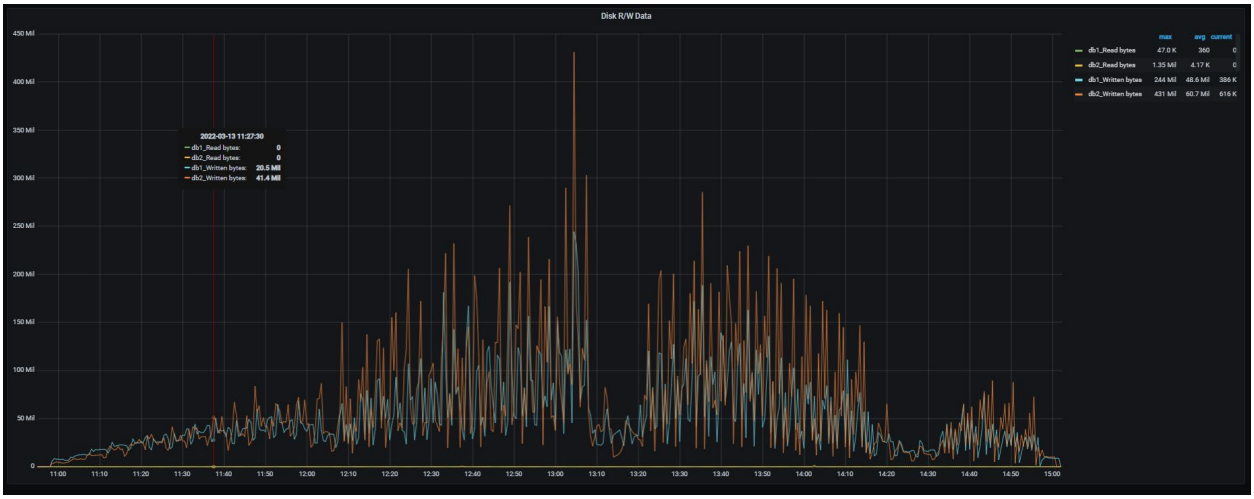
CPU



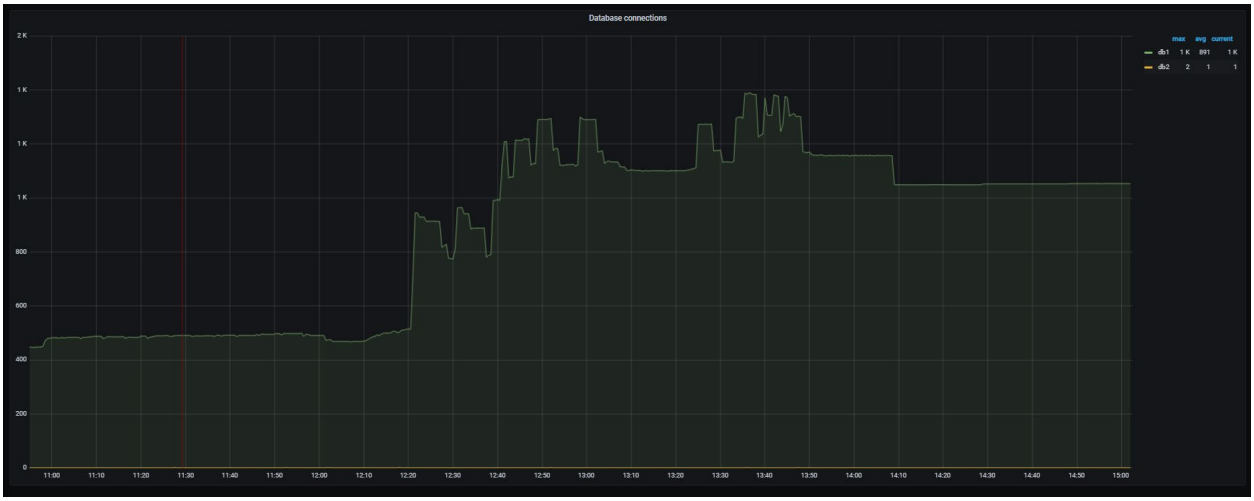
Memory



Disk R/W



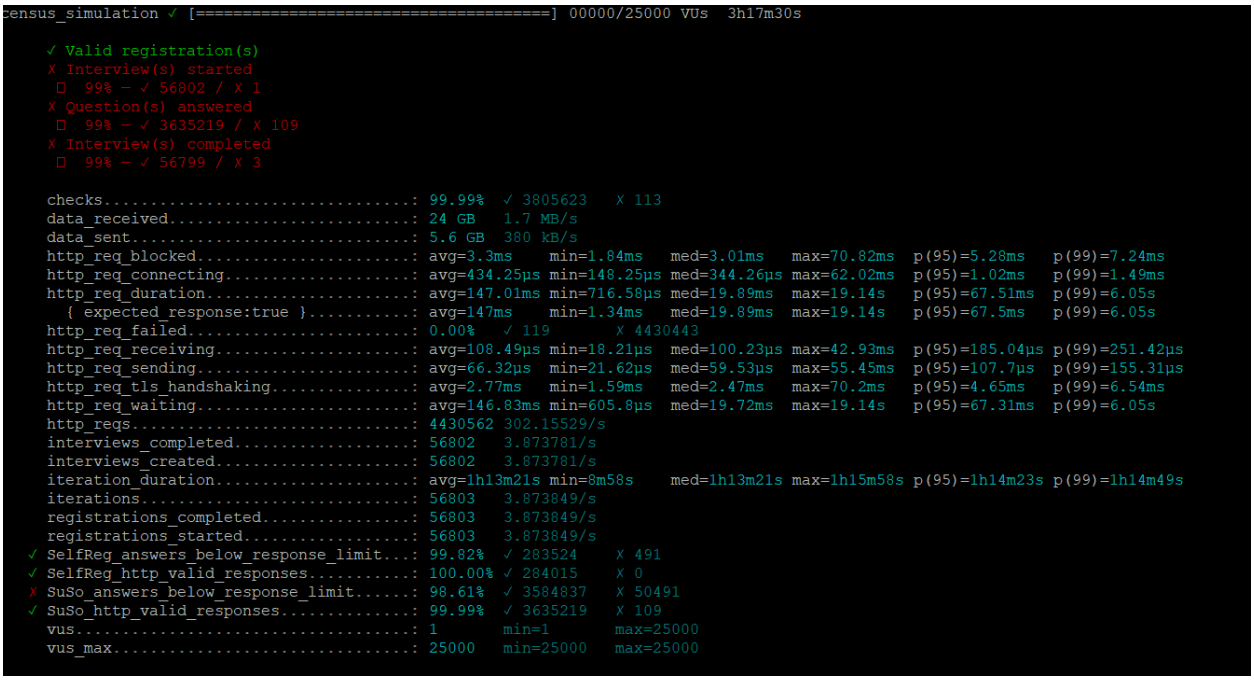
Database connections



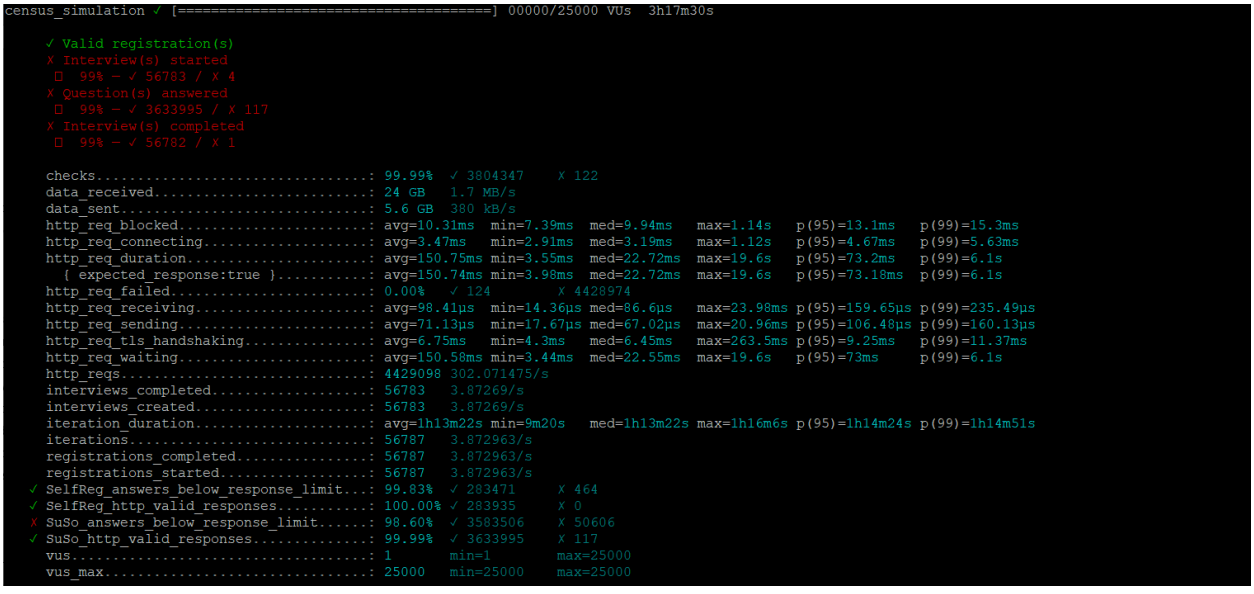
Load scripts results

Self-registration and interview completion load script

First load generator:



Second load generator:



Third load generator:

```
census_simulation ✓ [======================================] 00000/25000 VUs  3h17m30s

    ✓ Valid registration(s)
    ✗ Interview(s) started
      □  99% — ✓ 56775 / ✗ 2
    ✗ Question(s) answered
      □  99% — ✓ 3633457 / ✗ 143
    ✗ Interview(s) completed
      □  99% — ✓ 56773 / ✗ 2

    checks.............................: 99.99%  ✓ 3803782    ✗ 147
    data_received......................: 24 GB   1.7 MB/s
    data_sent..........................: 5.6 GB  380 kB/s
    http_req_blocked...................: avg=10.45ms  min=7.36ms  med=10.07ms  max=1.14s    p(95)=13.16ms  p(99)=15.34ms
    http_req_connecting................: avg=3.48ms   min=2.9ms   med=3.2ms    max=1.12s    p(95)=4.68ms   p(99)=5.54ms
    http_req_duration..................: avg=151.27ms min=3.63ms  med=22.72ms  max=19.61s   p(95)=73.53ms  p(99)=6.11s
      { expected_response:true }.......: avg=151.26ms min=3.98ms  med=22.72ms  max=19.61s   p(95)=73.52ms  p(99)=6.11s
    http_req_failed....................: 0.00%   ✓ 151        ✗ 4428311
    http_req_receiving.................: avg=108.51µs min=14.68µs med=96.92µs  max=314.82ms p(95)=170.7µs  p(99)=254.65µs
    http_req_sending...................: avg=77.4µs   min=18.37µs med=74.01µs  max=5.57ms   p(95)=110.89µs p(99)=167µs
    http_req_tls_handshaking...........: avg=6.87ms   min=4.34ms  med=6.54ms   max=282.68ms p(95)=9.29ms   p(99)=11.43ms
    http_req_waiting...................: avg=151.08ms min=3.51ms  med=22.54ms  max=19.61s   p(95)=73.31ms  p(99)=6.11s
    http_reqs..........................: 4428462 302.323947/s
    interviews_completed...............: 56775   3.875938/s
    interviews_created.................: 56775   3.875938/s
    iteration_duration.................: avg=1h13m22s min=8m46s   med=1h13m22s max=1h16m12s p(95)=1h14m25s p(99)=1h14m51s
    iterations.........................: 56777   3.876074/s
    registrations_completed............: 56777   3.876074/s
    registrations_started..............: 56777   3.876074/s
  ✓ SelfReg_answers_below_response_limit...: 99.84%  ✓ 283447     ✗ 438
  ✓ SelfReg_http_valid_responses.....: 100.00% ✓ 283885     ✗ 0
  ✗ SuSo_answers_below_response_limit.....: 98.59%  ✓ 3582578    ✗ 51022
  ✓ SuSo_http_valid_responses........: 99.99%  ✓ 3633457    ✗ 143
    vus................................: 0       min=0        max=25000
    vus_max............................: 25000   min=25000    max=25000
```

Fourth load generator:

```
census_simulation ✓ [======================================] 00000/25000 VUs  3h17m30s

    ✓ Valid registration(s)
    ✓ Interview(s) started
    ✗ Question(s) answered
      □  99% — ✓ 3634038 / ✗ 138
    ✗ Interview(s) completed
      □  99% — ✓ 56783 / ✗ 1

    checks.............................: 99.99%  ✓ 3804389    ✗ 139
    data_received......................: 24 GB   1.7 MB/s
    data_sent..........................: 5.6 GB  380 kB/s
    http_req_blocked...................: avg=10.31ms  min=7.37ms   med=9.94ms   max=1.21s   p(95)=13.09ms  p(99)=15.28ms
    http_req_connecting................: avg=3.47ms   min=2.91ms   med=3.18ms   max=1.03s   p(95)=4.66ms   p(99)=5.62ms
    http_req_duration..................: avg=150.99ms min=3.66ms   med=22.72ms  max=19.49s  p(95)=73.33ms  p(99)=6.08s
      { expected_response:true }.......: avg=150.98ms min=4.05ms   med=22.72ms  max=19.49s  p(95)=73.31ms  p(99)=6.08s
    http_req_failed....................: 0.00%   ✓ 141        ✗ 4429011
    http_req_receiving.................: avg=98.53µs  min=14.36µs  med=86.71µs  max=21.78ms p(95)=160.1µs  p(99)=238.12µs
    http_req_sending...................: avg=71.19µs  min=16.87µs  med=67.05µs  max=10.27ms p(95)=106.57µs p(99)=161.26µs
    http_req_tls_handshaking...........: avg=6.75ms   min=4.33ms   med=6.45ms   max=248.1ms p(95)=9.24ms   p(99)=11.36ms
    http_req_waiting...................: avg=150.82ms min=3.52ms   med=22.55ms  max=19.49s  p(95)=73.14ms  p(99)=6.08s
    http_reqs..........................: 4429152 302.295271/s
    interviews_completed...............: 56784   3.87558/s
    interviews_created.................: 56784   3.87558/s
    iteration_duration.................: avg=1h13m22s min=1h10m58s med=1h13m22s max=1h16m1s p(95)=1h14m24s p(99)=1h14m50s
    iterations.........................: 56784   3.87558/s
    registrations_completed............: 56784   3.87558/s
    registrations_started..............: 56784   3.87558/s
  ✓ SelfReg_answers_below_response_limit...: 99.84%  ✓ 283489     ✗ 431
  ✓ SelfReg_http_valid_responses.....: 100.00% ✓ 283920     ✗ 0
  ✗ SuSo_answers_below_response_limit.....: 98.60%  ✓ 3583307    ✗ 50869
  ✓ SuSo_http_valid_responses........: 99.99%  ✓ 3634038    ✗ 138
    vus................................: 1       min=1        max=25000
    vus_max............................: 25000   min=25000    max=25000
```

Test conclusions

Even the loadscripts reports some threshold reached, test results are reflecting that the system is prepared to handle the required load.
Since the system is upgradable, suggestion is to increase resources on application server memory (from 64G to 96G and amount of cores from 16 to 32. This will improve response time and system overall feasibility.