Project of Russian Science Foundation № 21-71-20078

Analytical processing of large arrays of heterogeneous data in the interests of state assessment, decision support and incident investigation to ensure cybersecurity of critical infrastructures

Description of work performed and scientific results obtained in 2021

1. The concept of analytical processing of large arrays of heterogeneous data on cyber security events in the interests of assessing the state, supporting decision-making and investigating computer incidents in critical infrastructures (CI) has been developed. The concept describes: principles of organizing analytical processing of large arrays of heterogeneous data; goals and tasks of advanced monitoring and security management systems for CI, as well as the requirements imposed on them; features of the application of supercomputer computing technology for the analysis of cyber security data; mechanisms to ensure status assessment, decision support and investigation of cyber security incidents; description of the CI of the categories "Internet of Things" and "cyber-physical system"; generalized architecture of the system being developed. The main tasks solved by the system were formulated: prompt detection of attacks and violations of the security policy; identification of security incidents and their prioritization; automatic response to security incidents; maintaining a knowledge base on security incidents; audit and investigation of security incidents; assessment of threats for individual resources of the CI. Their solution in the supercomputer computing environment involves the use of CUDA graphics accelerators, TensorFlow, TyTorch and Theano libraries, MPI technology, and an actor model. The main analytical components of the system are determined: detection of attacks; detection of anomalies; security assessments; risk analysis; decision making; visualization; investigation of computer incidents. The functioning of the system is carried out in the modes of training and operation.

2. A general approach and requirements for the components of real-time attack detection based on simulation and graph-oriented modeling have been developed. The analysis of the current state of research has made it possible to single out detection methods based on graphs, Bayesian networks, Markov models, Petri nets, simulation modeling and big data technologies for detecting multi-step attacks. The requirements for detection components are considered from the point of view of possible attack effects and the process of their detection, the protected CI, the structure and volume of processed data, the variability and scalability of the approach, as well as component testing. Detection is carried out in a mode close to real-time, with the ability to promptly identify known types of security incidents within the allotted time frame. The structure of the intrusion detection components assumes the possibility of embedding private intrusion detection modules in the configuration process. All built-in and included detection modules operate in parallel, as a result of which their functions can be distributed within the computational cluster of a supercomputer.

3. A general approach and requirements for the components of real-time detection of anomalous activity and violations of security criteria and policies based on analytical processing of large arrays of heterogeneous data on cyber security events have been developed. The problem of detecting in real time anomalous activity and violations of security criteria and policies is associated with solving problems caused by large volumes of analyzed data and their dimensions, and high speed of generation of new data streams. These tasks give rise to scientific and practical challenges associated with asynchronous data generation, the identification of dynamic relationships between events, the heterogeneity of the used formats and schemes for describing events, as well as the shift of the concept in the data. The key requirements for the components of detecting anomalous activity and violations of security criteria and policies are related to ensuring scalability, efficiency, completeness and reliability of the decision, as well as adaptability of procedures for processing large amounts of data on security events. In the structural diagram of the components, two levels can be distinguished: scheduling of tasks related to event processing and processing of security events. The second level is represented by a variety of software modules that provide several modes of operation: anomaly detection modules, model verification modules, as well as modules that implement additional training of models, which will allow one not to retrain all previously trained machine learning models, but to expand and refine the composition of types of violations of criteria and security policies and abnormal activity.

4. A general approach and requirements for the components of the operational assessment of the security of information, telecommunications and other critical resources based on analytical processing of large arrays of heterogeneous data have been developed. The approach includes the stages of development of requirements for security assessment, specification of security assessment, development of a security assessment project, security assessment and formation of a conclusion on security assessment. It is based on the development of a model-methodological apparatus for assessing of information, telecommunications and other critical resources based on the joint use of attack graphs and service dependency graphs. At the same time, the MPI technology and the supercomputer's distributed memory can be used in the supercomputer computing environment. The selected functional requirements determine the need for a comprehensive security assessment, supporting the adoption of protection measures, fixing criteria, procedures, tools for operational security assessment and forms of presentation of its results, specifying the conditions and limits of application of the security assessment component. The qualimetric requirements include: unbiasedness, consistency, efficiency and sufficiency of the assessment of the security of the CI resources.

5. A general approach and requirements for the components of operational analysis and information security risk management based on analytical processing of large arrays of heterogeneous data on cyber security events in the interests of assessing the state, supporting decision-making and investigating incidents have been developed. The purpose of these components is to determine the level of cyber security risks, to balance the cost of potential negative consequences and the cost of protection measures, and to develop recommendations for handling the identified risks. In the context of large amounts of data, the components of analysis and risk management must promptly respond to the changing situation, recalculate and compare integral assessments with the criterion. Achieving the formulated requirements for these components is possible using a set of methods, including the formation of an attack model in the form of a graph in conditions of processing big data and taking into account unknown vulnerabilities, processing large attack graphs in order to calculate security metrics, processing attack graphs with cycles, the formation of objective integral metrics and an explanation of their meaning based on the ontological approach, parallel processing of big data, in particular in the formation and processing of graph models and ontologies implemented for execution on a high-performance cluster with support for horizontal scaling of computing resources within this component.

6. A general approach and requirements for the components of operational visualization of large arrays of heterogeneous data on cyber security events in the interests of assessing the state, supporting decision-making and investigating incidents have been developed. The main problems of operational visualization of large arrays of heterogeneous data were identified. They include the choice of a visualization model, ensuring efficiency, perception of large data, processing heterogeneous data and multidimensionality of data. The generated list of requirements for operational visualization includes the ability to aggregate data by aggregating objects and measurements, supporting the conversion of heterogeneous data to a quantitative or categorical form, as well as defining the data structure, ensuring efficiency through the possibility of horizontal scaling of computing power, including visualization models that are capable of displaying aggregated objects and human-machine interaction. The general approach to visualization is presented in the form of a standard visualization pipeline adapted for visualization of large arrays of heterogeneous data, in the form of a diagram: data analysis - aggregation - markup - rendering. On the basis of this general approach, the architecture of the visualization component is proposed with support for the scaling of modules due to methods of parallel data processing.

7. A general approach and requirements for the components of decision-making on the protection of information, telecommunications and other critical resources based on analytical processing of large arrays of heterogeneous data on cyber security events have been developed. The main limitation in decision making is the large number of decision options and their characteristics that need to be analyzed. Therefore, to effectively solve the problem, it is necessary to adapt the applied model apparatus to the conditions and limitations of modern systems of parallel processing, including the use of supercomputer technologies and cluster computing. In the structural diagram of the components, the levels of task planning and decision making are highlighted. The second level is represented by a variety of software modules that provide static and dynamic operation of the component. This includes the following modules: processing event data and security data from external sources; inventory of assets; formation and updating of the attack model; formation of the attacker's model; formation of a model of countermeasures, integration with the model of attacks; interaction with components of security assessment and analysis, detection of attacks and anomalies; decision making and calculation of indicators of choice of response measures; determining the time available for making a decision; interaction with the supercomputer center; interactions with rendering components.

8. A general approach and requirements for the components of investigating computer incidents based on analytical processing of large arrays of heterogeneous cybersecurity data have been developed. The general approach to incident investigation includes the following procedures: indexing, hashing and digitally signing input data for ease of retrieval, ensuring integrity and authenticity; data about attack scenarios and their context comes from real-time attack detection components; anomaly data comes from components of real-time detection of anomalous activity and violations of security criteria and policies; the priority of detected attacks and anomalies for incident investigation is determined on the basis of data on the criticality of CI elements provided by the operational analysis and information security risk management components; reports of stored security incidents, including their context, and related incidents are provided as an output. In addition, data is generated for operational visualization components in order to display them for analytics.

The research results were published in 12 articles indexed by WoS and Scopus, and 14 articles and abstracts indexed by the RSCI. One monograph has been published. During the implementation of the project, 9 certificates of state registration of computer programs were prepared. The team members participated in the testing of the results at 12 Russian and international conferences and seminars. Also, advanced training courses in the field of cyber forensics were prepared and conducted for employees of the Investigative Committee from different regions of Russia [https://tass.ru/obschestvo/13166641/amp]. To simulate the actions of violators and collect data on security events, a prototype of an information system was developed and a Hackathon (an ethical hacking competition) was organized for students of St. Petersburg universities. The event was attended by 29 teams, more than 70 students [https://spark.ru/user/139694/blog/82190/uchyonie-spb-fits-ran-testiruyut-algoritmi-obnaruzheniya-atak].