Big Data Information Security Maintenance

Conference Paper · September 2014
DOI: 10.1145/2659651.2659655

4 authors, including:

- Natalia G. Miloslavskaya
  National Research Nuclear University MEPHI
  116 PUBLICATIONS  35 CITATIONS
  SEE PROFILE

- Alexander Tolstoy
  National Research Nuclear University MEPHI
  91 PUBLICATIONS  31 CITATIONS
  SEE PROFILE

- Sergey V. Zapechnikov
  National Research Nuclear University MEPHI
  22 PUBLICATIONS  4 CITATIONS
  SEE PROFILE

All in-text references underlined in blue are linked to publications on ResearchGate, letting you access and read them immediately.
Big Data Information Security Maintenance

Natalia Miloslavskaya, Mikhail Senatorov, Alexander Tolstoy, Sergey Zapechnikov
Information Security of Banking Systems Department
National Research Nuclear University MEPhI (Moscow Engineering Physics Institute)
31 Kashirskoe shosse, 115409, Moscow, Russia
+7-495-7885699*9084
NGMiloslavskaya@mephi.ru

ABSTRACT
The need to protect big data, particularly those relating to information security maintenance (ISM) of an enterprise’s IT infrastructure (ITI), and their processing is shown. Related worldwide experience of addressing big data ISM issues is summarized. An attempt to formulate a big data ISM problem statement is undertaken. An infrastructure for big data ISM is proposed. The importance of big data visualization is discussed.

Categories and Subject Descriptors
K.6.5 [Security and Protection]
General Terms
Information Security, Big Data

Keywords
Big Data; Information Security; Secure Infrastructure.

1. INTRODUCTION
Modern enterprises of different scope and size collect and store a huge amount of data about the current state of their IT infrastructure (ITI). These data need to be processed correctly and promptly to identify and visualize possible information security (IS) threats, vulnerabilities to be eliminated and at least IS incidents occurred, to optimize ITI monitoring strategy and resources, to calculate current and forecast further IS risks and so on. The data are generated from the different types of information considered in a particular context. This information comes not just from the separate domain controllers, proxy servers, DNS servers, information protection tools (IPT). It also describes current configuration of network devices, generalized characteristics of network traffic, application and network services functioning, activity and specific actions of individual end-users, as well as contains e-mails, web-based content, digitized audio and video, the data of business processes, enterprise’s internal documents and analytical data for many years of its existence.

Volumes and heterogeneity of data and related activity for further scrupulous monitoring and analysis are very high. A problem of structured, consolidated and visual presentation of data to make timely and informed decisions in the field of IS management for all enterprise’s ITI assets is one of the biggest ITI challenges nowadays. Ever-increasing volumes of data on IS events, ITI assets, their vulnerabilities, users, IS threats, IS risks and related information, as well as the need for more rapid obtaining of systematic and analyzed in a certain way “raw” heterogeneous information for faster understanding of the current situation in IS for ensuring ITI are well-known problems of big data.

“Big Data IT” term combines a set of IT existing in the area with three “V”: 1) volume (very large volumes of data), 2) velocity (very high data transfer rate) and 3) variety (weak structured data, which is primarily understand as data structure irregularity and difficulty of extracting homogeneous data from a stream and identifying some correlations) later added by veracity, variability, value and visibility. Common terminology in the field of big data has not yet been developed. Our interpretation of the big data concept is the datasets of such size and structure that exceed the capabilities of traditional programming tools (databases, software, etc.) for data collection, storage and processing in a reasonable time and a-fortiori exceed the capacity of their perception by a human. Data can be structured, semi-structured and unstructured that makes it impossible to manage and process them effectively in a traditional way. There are two types of big data processing [1]:

1) batch processing in pseudo real or soft real-time, when data already stored in the non-volatile memory are processed (only the stored data are processed) and probability and time characteristics of data conversion process are mainly determined by the requirements of the applied problems;

2) stream processing in hard real-time, when collected data without storing to non-volatile media are processed (only the processing operations results are stored) and probability and time characteristics of data conversion process are mainly determined by incoming data rate, since the appearance of the queues at the processing nodes leads to irreversible loss of data.

Big data IT should be understood primarily as IT, aimed at processing very large-scale arrays of semi-structured data in real-time. Big data IT fundamentally differs from traditional IT so that it becomes data-centric or data-driven. Processing various requests (orders, etc.) device or medium (computer, cluster, Cloud) is put at the center of the data processing for traditional IT. And big data are considered primarily as continuous flowing substance, processing mechanisms for which must be built in the streams themselves. Wherein a downstream rate for data incoming for processing and a rate of results delivery should be no lower that the stream rate, as otherwise this would lead to an infinite growth or queues or useless storage of infinitely increasing volumes of raw data. All technologies supporting big data IT were developed without taking into account the ISM requirements to them. IS is a quality (property) of information to maintain its confidentiality, integrity, availability, authenticity, accountability, non-repudiation and reliability. ISM is a complicated process divided into many sub-processes of maintaining the secure (protected) state of information characterized by confidentiality, integrity, availability, etc. Information protection implying “protection of information” involves all the activities aimed at IS maintaining, namely preventing diversion of protected
information, unauthorized and unintended impacts on the protected information. IPT are the tools that implement information protection in a particular environment – a separate computer, network or whole enterprise’s IT. Urgency of an ISM problem for big data IT is defined by the contradiction between the increasing needs in the processing such data arising in various spheres of human activity, and lack of capacity to ensure the confidentiality, integrity, availability and other properties of data to be processed and, in particular, of personal data, and ISM for data processing facilities and ITI’s resources, including system and application software for the data centers processing big data. Therefore the given issue is particularly relevant as it is very far from full resolving.

2. RELATED WORK

Big data IT is now probably the most popular and debated topic in the field of IT. The former group head of security services for Barclays, who now works for Splunk, S.Gailey said [2]: “Security is now a big data problem because the data that has a security context is huge. It’s not just a collection of security tools producing data, it’s your whole organization. If you’re going to ignore some of that data, or if you can’t analyze it, then you are not doing security properly.” A wrong from many scientists and practitioners point of view opinion that the tasks listed above can be solved using well-known approaches and familiar IPT is expressed in some publications. The radical restructuring of IT applied to big data processing causes the need for the fundamentally different approaches to ISM with an appropriate methodological base being in its infancy nowadays. The validity of such an opinion is due to the fact that the computer, network technologies are built on the concept of defending a single physical entity (such as a server or database) and not distributed, large-scale computing environment supporting big data IT.

The intensive studies of general ISM issues for big data IT are conducted all over the world.

In 2012, the U.S. government has announced a strategic “Big Data Initiative”, which involves six federal ministries and departments of the USA.

The University of Berkeley (California) is the main Initiative’s executor.

The special research and technology center for big data (called The Science and Technology Center for Big Data) was created at the Massachusetts Institute of Technology in the Laboratory of Computer Science and Artificial Intelligence.

Research in the field of methodology of big data cryptographic protection is carried out by a large group of applied cryptography in the Laboratory of Security, the Computer Science Department of the Stanford University (USA).

The project “Big Data Public Private Forum” under Horizon 2020 strategic initiatives is implemented under the auspices of the European Commission. Scientific papers on the subjects are published by the international research team, including scientists from Germany, the Netherlands, France, Switzerland, and Denmark.

Projects in the field of big data ISM are actively implemented by the leading manufacturers of IT products and systems. In particular, the results of IBM research are implemented in a series of InfoSphere Guardium products to protect structured and unstructured data in real-time and offline mode.

Not all individual tasks receive the same attention because of the ISM problem breadth and diversity. Main directions of research in this area are currently focused as follows.

1. Development of methods of ensuring confidentiality of information processed in an untrusted environment by performing operations over the encrypted data. Level of the problem solutions achieved currently can theoretically perform operations over the encrypted data, described by arbitrary Boolean functions with some restrictions [3], but the complexity of the corresponding algorithms and mathematical apparatus used do not yet provide (not even close) a level of performance suitable for practical applications, and especially for applying in the real-time systems.

2. Development of big data access control mechanisms. Traditional access control methods, based on a discretionary, mandatory and role models, are not applicable when dealing with big data. Here’s leading research area is the creation of new access control models based on the attribute model [4].

3. Development of methods and means of remote integrity control for data arrays processed in the environment being external to the data owner and therefore untrusted. Due to the difficulties of complete control procedures implementation, the methods of probabilistic integrity monitoring have a leading role here. However, all known methods are designed only for static arrays control. So-called locality-sensitive hashing functions and locally decodable codes [5], allowing to monitor portions of a large array, are developed. But their application to rapidly changing datasets requires more research.

4. Development of data provenance methods for tracing and recording the origins of data and its movement between databases for rapidly changing datasets [6].

5. Development of data-processing methods using proxy re-encryption schemes (with a limited set of operations) and verification of the computing results. The most significant achievement in the area is by far the open source program compiler written in C language, created by a research team from Microsoft and several universities [7].

6. Development of methods for secure collection, processing and storage of IS events evidence submitted in electronic form (so-called digital evidence) when processing the big IS-related data in order to avoid tampering and to support ISM [8].

7. Study of the problem of bilateral and multilateral secure computing for different initial conditions and settings (IS here means a property to maintain confidentiality and authenticity of not only the input and output data, but also all the intermediate results of computations performed) [9].

Majority of methods developed were designed to test the possibility to implement the stated functionality in principle. Currently available solutions are applicable generally to data processing in pseudo real-time (soft real-time). Methods and algorithms proposed to date are mainly designed to provide either some IS aspects for permanently stored datasets or their separate elements or to protect processing big data tools and corresponding infrastructure in the framework of which they are processed. These methods focus primarily on the application for batch processing. Appropriate methods and algorithms focused on stream processing, are practically absent at present.

3. BIG DATA INFORMATION SECURITY PROBLEM STATEMENT

The ISM issues for big data IT can be divided into four categories according to a number of authoritative sources (in particular, the documents of the Big Data Working Group of the Cloud Security Alliance) [http://cloudsecurityalliance.org]:

- providing IS for the big data infrastructure;
• providing the logical integrity and IS monitoring in real-time;
• personal privacy ensuring;
• rational data management.

The processing object itself within big data IT is “sensitive” because it contains inter alia important information about the enterprise’s ISM level. When big data IT is used for ISM purposes, the result is greatly enhanced visibility into enterprise’s ITI and creation of trusted ITI being able to distinguish suspicious activities from normal. Large volumes of ISM-related information must be mapped out, broken down, classified, normalized and correlated to identify the IS risks to the enterprise’s ITI. After processing in a specific context these data become meaningful information, requiring its “hiding” from the intruders. Such data are the most important protection object. With more meaningful information located in and moving between more places in ITI than ever before, there is also a vastly increased number of points and ways to hack it, resulting in theft, fraud, reputational loss, etc. Therefore, big data processing processes themselves as well as all their inputs, intermediate results of computations performed and outputs require ISM.

The key point of ISM that is rationale for big data IT, like any other technology, is a risk-based approach. A complete-cycle risk assessment for big data IT should be fulfilled. Only a few risks associated with big data IT can be highlighted here [10]:

• as any new technology big data IT is not well understood and it will introduce new vulnerabilities to enterprises’ ITI;
• its implementations typically include an open source code, often having unrecognized (until sometime) back doors and default credentials (proprietary software also has them, but it is harder to find them without having an open source code); if an intruder will find some vulnerability in it earlier than others, he/she can use it for his/her own profit, until the vulnerability will be eliminated;
• not reviewed attacks against cluster’s nodes and inadequately hardened servers;
• insufficient control over user’s authentication and access to data from the multiple locations;
• not fulfilled regulatory requirements, with access to logs and audit trails problematic;
• malicious data input and inadequate data validation.

Of course this list should be significantly broadened, for example, by the IS risks associated with virtualized and distributed nature of big data IT itself. Thus the current lack of a systematic approach to ISM for big data processing in real-time constrains this IT usage.

The main problem of ISM for big data IT can be correctly formulated as follows – to create a complex consisting of models and methods allowing to ensure protection of big data processing in a hard real-time at the processing nodes and data centers. As the initial condition of a problem to be addressed it is assumed that the node’s input receives a high speed stream of the semi-structured data (including possibly formed by mixing several separate streams). Data centers carry out continuous processing of an incoming stream at a rate not lower than the downstream speed, without loss of any part of the stream. Stream processing can consist of an arbitrary number of separate elementary operations, each entry of which is supplied by either a source stream or a stream resulting from an already completed elementary operation. Each elementary operation generates as an output a stream that is either transferred to the input of another elementary operation, or is an output stream of a data center. Thus, the data center’s functions are described by a graph whose vertices are the elementary operations and ribs are the streams between them. The specific content of elementary operations performed and configuration of the streams between them are determined by application tasks that the data center performs. The output stream does not necessarily coincide with the intensity of the input streams (most often it is smaller since it contains only the meaningful information obtained as a result of the input data analytical processing). Under these conditions it is requires primarily to provide IS (possibly within defined quantitative indicators for individual requirements) for passing through the data center’s information stream, including both traditional basic IS aspects (data confidentiality, integrity, availability and authenticity), and specific new requirements arising due to the data stream’s characteristics (large volume, weak structured, high intensity) and processing procedures. Such requirements, in particular, may be the requirements to control access to the separate elements or data structures coming in a stream; tracking of personal data in a stream, or the relationships between anonymous data allowing to identify reliably the persons to whom they relate.

The second group of requirements relates to an algorithmic component of processing: it is required to provide confidence in the data center’s components that implement some elementary operations, and if a sufficient level of confidence in them is not possible, to verify the operations implementation results and to correct all random errors or intentionally introduced distortions.

Finally, the third group of requirements relates to ISM for the data center’s infrastructure. The most common current approach to big data processing, regardless of the relevance of requirements to processing in real-time, is creation of a computational tools cluster and software implementation of massively parallel processing of the elements of the very large-scale arrays. In the absence of the consumer’s own ITI he/she can use the services of a cloud computing provider. In this case, the cloud computing services are provided according to the “infrastructure as a service” (IaaS) model, as the other models ("platform as a service" and “software as a service”) are obviously irrational or even impossible in the presence of real-time requirements.

Solvable problem in this formulation is one of the key problems leading to a satisfactory solution of ISM issues for big data IT. Scientific novelty of the formulated problem is determined by the following factors:

1) previously known solutions of tasks to ensure the IS processing of very large-scale arrays of semi-structured data were obtained for the pseudo real-time processing conditions; all such solutions are mainly focused on MapReduce methodology; and there are only a few, not related to a single cycle, works devoted to the analysis of IS in real-time and discussing some differences between them;

2) create a complex of protected object’s (big data IT) models taking into account, in contrast to the previously known, the operation of ISM tools and mechanisms as the elements of the queuing systems network representing the data center’s processing stream;

3) create the formal IS threats and intruder models, taking into account not only functional, but also computing capabilities of the intruder, allowing to formulate the correct statements on the properties of the security algorithms and to receive their evidence, in contrast to the previously used qualitative or semi-formal models;

4) obtain solutions for a complex of ISM issues being implemented as the processing algorithms for an incoming stream and the internal streams between the data center’s elementary operations, as well as the processes of interaction between the data center’s resources, with the data owner and
users, providing performing protection functions in hard real-time.

4. **SECURE INFRASTRUCTURE FOR BIG DATA**

Data centers having certain computational power, storage, and applications necessary to support an enterprise business are spaces specifically designed to accommodate its big data. Special case of applying the requirements listed above is to address the problems of ensuring own large-scale ITI’s IS, where such a large and comprehensive structured arrays of IS events-related data are formed that they can be classified as big data. To solve this problem is possible only in the development of a single secure infrastructure (SI) that allows all IS-related information to be collected, indexed, normalized, analyzed and shared in a secure way.

SI can be created and implemented in two possible ways – as a built-in secure data center within ITI or stand-alone secure data center. In both cases it should be open, flexible and scalable with well-defined input and output data formats, high performance and high availability as everyday requirements. And to be truly effective it should have multilayered defense, be constantly monitored and well maintained including operation management, performance and availability monitoring, compliance management, security information and event management (SIEM), etc.

A good centralized system, monitoring information itself, its states, access to its, IS events related to it and its processing processes, is needed. Moving from a reactive to a proactive strategy is the best approach of SI implementation.

And of course SI should be integrated with the existing security tools and processes in place – initially in parallel to the existing connections (for example it concerns SIEM tools) and finally completely retooling for big data IT.

The basic foundation of SI design that seeks to improve its scalability, performance, flexibility, resiliency and maintenance is the traditional layered approach with the following layers: service delivery, software, platform, infrastructure supply, operations and management layers. The layers themselves and their interconnections should be protected by the corresponding measures and tools.

Virtualized networks should be used. It is a good solution for carrying big data to use VLANs between the data centers and virtual devices as the internal network for virtual hosts implementing virtual switches.

The individual virtual servers should be protected in accordance with the standards recognized by an enterprise. It is a good practice to uninstall unneeded services (for example FTP) and to have a timely software and operating systems patch management process in place, backup services and encryption.

It is vital to separate an enterprise’ users regular traffic from the big data traffic characterizing IS of its ITI. According to their main mission the firewalls as IPTs need to examine every flowing through them packet for every session. Their usage in SI is only possible by ensuring that only trusted big data traffic is moving through the encrypted network tunnels and eliminating firewalls in between the data centers’ elements. And encryption, particularly when dealing with big data analytics, should be used also.

All the processes, services, procedures, documentation (including different IS policies in the first place such as the data center IS policy), software, personnel matters, including, for example, issues of IS training for the data center’s personnel and much more should be worked out for developing, implementing, analysis and improving SI for big enterprise’s IS-related data, as for any other infrastructure.

Here are only the main ideas. The reason is that the topic of secure data center reference model designing protected from the different points of view – organizational, technical, hardware, software – deserves a separate and careful consideration. Before designing it a few models should be worked out to formulate main requirements to data center ISM:

1) a typical data center model the from the operational and business points of view;
2) a formal IS threats to the typical data center model;
3) an IS intruder violating data center IS model.

5. **BIG DATA VISUALIZATION**

IPT can register millions of IS events of different origins and consequences in the intranets of large enterprises during one day only. The amount of work required to identify the truly important data from the viewpoint of IS events and to obtain information on IS incidents can be extremely large. Big data visualization, increasing possibility of its perception by a human, improves the quality of IS event detection and IS incident management intellectual processes, during which the relevant information is collected, converted, stored, retrieved and used. This visualization promotes the formulation of hypotheses and accumulation of new knowledge about the IS event types, causes, relationships and consequences, which can then be used to improve the enterprise’s ITI IS in general.

Unlike imaging physically existing phenomena or data (e.g., the human body, territories, hurricane, etc.), IS-related information visualization should represent some abstract data, characterizing ITI’s protection state. For example, how to present the low-level network information or information obtained in a long time observation of a separate network device to detect IS events as already accomplished attacks or only developing on administrator’s eyes attempts of attacks against his/her enterprise’s ITI. Two main typical problems of any visualization arise here:

1) complexity meaning the ability to visualize various forms of reporting;
2) scalability meaning the ability to visualize large amounts of data in terms of algorithmic complexity and also the ability to display large amounts of information in understandable for a human way.

Advanced analytics is needed for true visibility. Many potential application areas for the systems, visualizing big ITI IS-related data can be identified in order to ensure its IS:

- network attacks recognition, anomalous activity detection, detection of unauthorized (fraud) operations with information and control over the unreliable employees;
- information stream analysis, tracing of the network packets moving through the communication channels, finding the ways of computer viruses dissemination or botnets functioning; access control to all ITI’s resources and ITI’s vulnerabilities detection;
- examination of the malware code or virus detection and extraction of the network attacks’ signatures in a large volume of information from the affected systems;
- system configurations control, analysis of the IPTs’ settings effectiveness, the study of the interaction of the various IS maintenance technologies and separate IPTs;
• correlation of IS events detected and some IS events “view” for dedicated communication channels, network devices, network protocols, services, applications, etc.;
• establishing some trends (that is possible only on the basis of big data usage), modeling and development of IS maintenance rules (for example, to configure a firewall or IDS/IPS);
• allocation of priorities requiring corresponding immediate actions to improve IS management for the whole enterprise’s ITI;
• visualization of a computer crime’s evidence for further investigation, etc.

Big IS-related data visualization is possible in two ways: generally across the whole ITI and by individual IS events. In the first case it is a summary of the following information related to ITI IS: indicators of ITI IS in general and of its individual parts, including IS risks (by type, consequences, implementation likelihood) and vulnerabilities localization points; indicators of the IPT efficiency by types and locations of their use; statistics on the number of IS events and incidents by type, severity, location, etc.; status and current state of IS events and incidents; detection of priority areas requiring immediate attention, etc. Considering a separate IS event, the information describing it from the following angles must be visualized [11]:

1) who: access subject;
2) why: for example, due to an accident, involuntary or unintentional action (error), ignorance, negligence or irresponsibility, malicious actions (internal or external);
3) what: unauthorized access to the system and/or information, installing unauthorized programs (without consent of the owner), remote causing malfunction in the information system, stressful situation creation, physical intrusion or illegal act, a variety of human errors, breakdown or failure of equipment, problems in the environment, etc.;
4) how: possibility, a list of methods and tools for each of the above “what”, especially, for attacks;
5) status of the IS event: an attempt being implemented, a successful or failed hacking;
6) which vulnerability was exploited;
7) on which type of assets (basic and related): databases, applications, systems, networks and telecommunications, storage device, end-user devices, people, facilities and environment, etc.;
8) what are the consequences to the properties of information (confidentiality, integrity, availability, etc.);
9) what are the consequences for business in general: direct or indirect.

Known that log files are one of main information sources of IS events. They usually contain important information about these events in terms of IS maintenance, which can be visualized:

• user’s (access subject) identifier (ID);
• dates and times of his log-ins and log-outs, details of the key events;
• name of the host that initiated the event to be registered and/or its location;
• records of successful and rejected attempts to access the objects;
• changes in the system configuration;
• modifying the lists of access subjects and objects;
• changes in access subjects authorities and access objects (protected information resources) status;
• all privileged actions (using the supervisor account, running and stopping the system, connecting/disconnecting to the input/output devices);
• starting programs and processes that access to protected information resources;
• system utilities and applications usage;
• all established sessions;
• files that have been accessed, and the type of access;
• network addresses and protocols;
• changes in the intensity and volume of incoming/outgoing traffic (including the Internet traffic);
• printing materials containing sensitive information;
• access to critical systems and services (e.g., web server, DBMS, mail server, synchronization services, etc.);
• system warnings and failures;
• alarm raised by the access control system and IDS;
• changes or attempts to change settings and tools for system protection management;
• activation and deactivation of protection systems, such as anti-virus systems and IDS;
• appearance of new devices with uncontrolled access (wireless access points, USB-connected devices, etc.), etc.

Today there are many interesting approaches and solutions in the field of information visualization – from simple graphs, charts, and histograms of different types, plans, drawings, hierarchical structures, graphs, tables with color gradation information, geographical and topographical maps, scanning and enlargements of the selected image to the dynamically changing complex animated images. Obviously, the choice of a particular method of visualization, size, shape and color of the images will be determined by the kind of information displayed, and often in order to create a complex image, at the same information should rather visualized in several ways complementary to each other. All these approaches are applicable to visualize ITI IS big data.

Similarly to the visualization tasks gradation in control and safety systems in nuclear power plants in crisis and situational centers, etc., that operate at different levels of authority and responsibility of users, the following gradation of ITI IS big data visualization tasks is considered:

1) information visualization for dynamic situation monitoring of the current ITI ISM state based on the most simple methods of data visualization, providing user’s (operator’s) quick response; for example, data on a single IS event or incident;
2) information visualization for the analysis of the gathered statistic data, strategic planning and ITI ISM strategy and tactics correction decision-making, which requires the use of more sophisticated methods of data analysis and visualization for the user – the analyst; for example, it is summarized data in terms of ITI protection in general and of its separate parts, including IS risks, IPT efficiency indicators by type and place of use, statistics on the number of IS incidents and events, etc.

Some generalized indicators with conditional gradation of the data observed are undoubtedly very useful for both types of users identified above. For example, highlighting areas with the green (acceptable), yellow (showing a tendency for increase in the IS risk level), orange (substantially exceeds a certain threshold level of IS risk) and red (the most critical area of an unacceptable level of risk and requiring urgent intervention) color. The main requirement is that information must be displayed in the most visual, clear, understandable way for a human allowing an unambiguous interpretation.

Modern systems used for visualization of ITI IS big data are far from perfect and needs of users – both operators and analysts of data centers. They have one major drawback – they only reflect
information concerning what has already occurred, i.e. the past, whereas it is necessary to make decisions about the future for rapid adjustment activities to maintain or even strengthen ITI IS. Therefore, it is required not to develop the classic SIEM systems (there is a sufficient number of them already on the market), that can display nice summary reports on any referral assessment ITI IS source, but the systems that can predict the impact on the most critical parameters relating to ITI ISM. Big data IT with distributed and parallel computing will complement these functions by mapping of large-scale network addresses space, network streams, simultaneous content correlation of several network packets, etc.

6. CONCLUSION
At present the main scope of big data IT is very diverse:

- fundamental and applied research in the field of elementary fundamental and applied research in the field of elementary particle physics (in particular, the international experiments with the accelerators);
- astronomy (e.g. processing data from big telescopes on our universe);
- biology (particularly human genome decoding);
- economy (in particular, macroeconomic development and financial risks forecasting);
- environmental monitoring, tracking the state of the atmosphere and weather forecasting;
- pharmaceuticals (synthesis of new chemical compounds and development of new medications);
- numerous business applications (analysis of consumers’ activity, demand, preferences, the effectiveness of advertising, etc.);
- providing big data IT as services (ready functional modules) in the implementation of other IT (in particular, search technology, deep data analytics to identify hidden patterns), the primary sources of information search and retrieval of the main content (semantics) in the extra-large arrays of documents without their direct reading by a human (particularly in news arrays, arrays of legislation, arrays of scientific publications (for example, to detect plagiarism), etc.);
- analytical processing of data about the ITI’s state to identify anomalies in the functioning of the system, IS incidents and intrusion prevention, etc.

But this wide range of applications is currently limited due to the many unresolved ISM issues for big data IT. It is obvious that the solution of at least a part of the most urgent tasks can give a strong impetus for the development of the specified scope in depth. In particular, the big data IT expansion is possible for the following important applications:

- information and analytical support to government, public and commercial organizations;
- monitoring of financial transactions, including transactions with the bank cards, in order to identify numerous fraudulent transactions;
- highly effective automated technological processes’ control;
- scientific data processing for clinical medicine, the creation of individual medications, telemedicine and remote provision of health services, etc.

Addressing the ISM issues will allow, on the one hand, to increase consumers confidence in big data IT and, on the other hand, significantly reduce the risks of unauthorized or unwanted information extraction technology by applying data mining and analytical intelligence on the Internet, in particular, can reduce the risks of collecting data by terrorist groups, will allow to counteract legalization of proceeds from crime and terrorist financing. For that purpose big data ISM problem statement is formulated, the main approaches to its solution and their specificity are outlined, as well as a secure infrastructure for big data, related to an enterprise’s ITI ISM, is briefly discussed.

The further research logically continuing related works is intended to address the stated ISM issues for the data centers processing big data.

7. REFERENCES