

MINISTRY OF SCIENCE AND EDUCATION OF  
THE REPUBLIC OF KAZAKHSTAN

Non-Profit Joint Stock Company  
ALMATY UNIVERSITY OF POWER ENGINEERING AND  
TELECOMMUNICATIONS

Department Telecommunication systems and networks

«Admitted»

Head of the Department Baykenov A.S.  
c.t.s., professor

(Surname and initials, degree, rank)

«    » 20 y.  
(sign)

DIPLOMA PROJECT

Theme: Infocommunication infrastructure for  
analytical systems of a telecommunications provider

Specialty: 5B071900 – Radio engineering electronics and telecommunications

Implemented by: Zamakhov A.V. ICTe-14-9  
(Student's surname and initials) group

Scientific Supervisor: Panchenko S.V., M.S., senior lecturer  
(Surname and initials, degree, rank)  
(sign) «25» 05 2018 y.

Reviewer: Vassin V.V., M.S., CTO KVINT LLP  
(Surname and initials, degree, rank)  
(sign) «25» 05 2018 y.

Advisors:  
of Economy section: Tuzelbaev B.I. associate professor  
(Surname and initials, degree, rank)  
(sign) «24» 05 2018 y.

of Life activity safety section: senior lecturer Begimbetova A.S., Ph D  
(Surname and initials, degree, rank)  
(sign) «24» 05 2018 y.

of Computer Science section: Panchenko S.V., M.S., senior lecturer  
(Surname and initials, degree, rank)  
(sign) «25» 05 2018 y.

Standards compliance controller: Panchenko S.V., M.S., senior lecturer  
(Surname and initials, degree, rank)  
(sign) «25» 05 2018 y.

Almaty 2018 y.

**MINISTRY OF SCIENCE AND EDUCATION OF THE REPUBLIC OF  
KAZAKHSTAN**

**Non-Profit Joint Stock Company  
ALMATY UNIVERSITY OF POWER ENGINEERING AND  
TELECOMMUNICATIONS**

Institute of Space Engineering and Telecommunications (ISET)

Specialty: 5B071900 – Radio engineering electronics and telecommunications

Department: Telecommunication systems and networks

**ASSIGNMENT**

**For diploma project implementation**

Student: Alexandr Vladislavovich Zamakhov

(name, patronymic and surname)

Theme: Infocommunication infrastructure for  
analytical systems of a telecommunications provider

Approved by Rector order № 155 of « 23 » 10 2017 y.

Deadline of completed project: « 25 » 05 2018 y.

Initial data for project, required parameters of designing result, object initial data:

The governmental program „Digital Kazakhstan” for 2017-2020y  
Elasticsearch searching engine technical guide  
CMDBuild system technical guide  
The complex project "National network of digital television  
and radio broadcasting of the Republic of Kazakhstan

List of questions for development in diploma project or brief content:

- 1) Analysis of the data analytics problem in  
telecommunication provider
- 2) Determination of the required search engine for the  
analytical system
- 3) Comparison and selection of the information analytical system
- 4) Identification of incoming data sources and organization  
of interaction with them
- 5) Description of the software part of the system infrastructure
- 6) Calculation of incoming data flow to the analytical system
- 7) Calculation of the hardware infrastructure for the  
analytical system






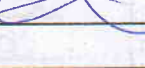

List of illustrations (with exact specifying of mandatory drawing):

- 1) Kibana interface and toolbars scheme
- 2) Scheme of report generation using Kibana
- 3) Logstash system operation diagram
- 4) Elasticsearch search engine operation scheme
- 5) Logstash and Elasticsearch interaction scheme
- 6) Splunk analytical system operation diagram
- 7) Diagram of the software infrastructure of designed system
- 8) Scheme of the calculated hardware infrastructure of the analytical system

Recommended main references:

- 1) The governmental program "Digital Kazakhstan" for 2017-2020
- 2) Administrator's Guide for using Kibana software 2018
- 3) Description of the Elasticsearch Python client API 2018
- 4) CMDBuild API docs or examples // free access by [cmdbuild.org](http://cmdbuild.org)
- 5) Zabbix API documentation 3.0
- 6) Elasticsearch: Hardware Definitive Guide [2.x] Free access by [elastic.co/guide](http://elastic.co/guide)

Project adviser with corresponding sections specifying:

Section	Advisor	Dates	Sign
Economy	Tuzelbaev B.I.	12.04 - 24.05	
Life safety	Beginbetova A.S.	01.04 - 24.05	
Computer science	Panchenko S.V.	25.05.2018	
Standards compliance	Panchenko S.V.	25.05.2018	
Technical part	Panchenko S.V.	25.05.2018	

## of diploma project implementation

[illegible]

Assignment issue date « 12 » 01 2018 y.

Head of Department: \_\_\_\_\_ Baykenov A.S  
(sign) (Surname and initials)

Scientific Supervisor:  Panchenko S.V.  
(sign) (Surname and initials)

Assignment submitted for implementation: A. Zamakhov (Sign) Zamakhov A.V. (Surname and initials)

## **Андатпа**

Бұл дипломдық жұмыста телекоммуникациялық провайдер деректерін талдауға арналған инфокоммуникациялық инфрақұрылым әзірлеуі көрсетілген. Ашық бастапқы кодпен қажетті бағдарламалық шешімдер іріктеуі орындалды. Қажетті есептеу қуаттарын қамтамасыз ету мақсатында аспаптық инфрақұрылым таңдалған. Серверлік жабдыққа болжамдалған жүктеме және қажетті дисктік кеңістік көлемі есептеуі жүргізілген. Бұл жұмыста сонымен қатар талдауға жұмсалатын еңбек күшімен шартталған жобаның экономикалық тиімділігінің сипаттамасы және желі мониторингіне негізделген есептеулер жүргізілді. Өміртіршілік қауіпсіздігі бөлімінде жұмыс ғимаратының табиғи және жасанды жарықтандыру есептеулері көрсетілген, сонымен қатар қажетті кондициялау жүйесі таңдалған.

## **Аннотация**

В данной дипломной работе представлена разработка инфокоммуникационной инфраструктуры для аналитики данных телекоммуникационного провайдера. Осуществлен подбор необходимых программных решений с открытым исходным кодом. Подобрана аппаратная инфраструктура с целью обеспечения необходимых вычислительных мощностей. Проведен расчет предполагаемой нагрузки на серверное оборудование и объема необходимого дискового пространства. В данной работе также представлено описание экономической эффективности проекта, которая обуславливается сокращением трудозатрат на аналитику и формирование отчетности на основе данных мониторинга сети. В разделе безопасности жизнедеятельности представлен расчет естественного и искусственного освещения рабочего помещения, а также подобрана необходимая система кондиционирования.

## **Abstract**

In this diploma project presented the design of an infocommunication infrastructure for data analytics of a telecommunications provider. Was carried out the selection of the necessary software solutions with open source. The hardware infrastructure was selected to provide the necessary computing power. The estimated load on server hardware and the amount of disk space required is calculated.. This paper also presents a description of the economic efficiency of the project, which is caused by a reduction in labor costs for analytics and reporting based on network monitoring data. In the life safety section, the calculation of natural and artificial illumination of the workplace is presented, and the necessary conditioning system is selected.

## Content

Introduction .....	8
1 Analysis of the problem .....	11
2 Toolset for the system design.....	15
2.1 Lucene searching engine .....	15
2.2 JSON format.....	16
2.3 Elasticsearch search engine.....	16
2.4 Kibana as a tool for visualization and analysis of data .....	22
2.5 Logstash as a tool for log collecting .....	25
2.6 Splunk analyzing system.....	26
2.7 Comparing of ELK stack and Splunk .....	33
2.8 Data sources for analytical system.....	37
2.9 General description of the designed infrastructure .....	45
3 Calculation part .....	49
3.1 Assessment of the required capacity of the directions for the transmission of state control data from the RTS to the regional and Republican Center .....	49
3.2 Definition of parameters for the server part of the infrastructure of the information analytical system .....	55
3.3 Design of a fault-tolerant system. ....	59
4 Life safety .....	63
4.1. Analysis of working conditions .....	63
4.2 Calculation of heat input due to temperature difference.....	65
4.3 Heat input from solar radiation through glazing.....	66
4.4 Heat input from people .....	67
4.5 Heat supply from lighting devices and office equipment .....	67
4.6. The overall heat balance and the choice of the split-system air conditioner ..	68
4.7 Initial data for calculation and selection of the lighting system .....	69
4.8 Calculation of natural light .....	70
4.9 Calculation of artificial lighting .....	72
5 Estimation of economic efficiency of the project .....	75
5.1 Capital expenditures for the implementation of the investment project.....	76

5.2 Description of labor costs associated with the work of the technical department, control service, the department of analytics and the central apparatus without the use of information and analysis system .....	79
5.3 Calculation of operating costs of a telecommunications provider using an information and analytical system .....	86
5.4 Calculation of economic efficiency .....	88
5.5 Conclusion about economic efficiency .....	89
Conclusion .....	91
List of Abbreviations .....	93
List of references .....	94
Appendix A Listing of the Zabbix interaction module .....	96
Appendix B Listing of the CMDBuild interaction module.....	98
Appendix C Reference from antiplagiarism checking	
Appendix D Electronic version of the DP and demonstration video-materials (CD-R)	
Appendix E Handout materials	



## **Introduction**

Currently, any modern project in any sphere of activity cannot be imagined without content. It is an information that is the basis of any project and it does not matter whether what is the source which is providing this data. The question of search, systematization and analysis is now quite acute. Every day this issue is becoming more relevant due to the growth in the amount of data from many sources.

Thus, companies now faces the problem of collecting, searching and organizing data. And the requirements for the tool are becoming more complex and broad. In some cases, already standard tools are indispensable. The more complex the application and the more complex the content structure, more special methods of processing the result or special types of search are required or the amount and format of the data is special, these all causes the need for own search and analytical system[1].

After all, these tools allow to correspond to the modern level in the question of the speed of reaction to emerging events, their processing and making a certain decision.

In modern society, the center of economic development is transferred from the material spheres of production (energy and raw materials basis) to the information field. Progressive movement, including in the field of economics, is determined today and will be determined in the next decade by the improvement of information technologies. The information society is the current stage of the social evolution of mankind.

The society, based on the information economy, already by its structure avoids the majority of socio-economic and environmental problems situationally gravitating over us today, and its potential is supposed to expand exponentially in all key parameters ("knowledge-generate knowledge") [1].

The most important manifestation of the qualitative technological breakthrough that led to the emergence of the information society, and at the same time one of its essential features is the emergence and rapid dissemination of data analysis technologies. This dramatically reduces the importance of financial resources from the point of view of the competitiveness of societies and corporations: if before they were the main source of power, now they turn into its consequence.

Companies work with large data sets, often unstructured and unrelated. Manual verification of such data can be very time-consuming. That is why the direction of intelligent analytical systems is now developing at a rapid pace. There are tools and technologies, combining which it is possible to create systems capable of analyzing the flow of data coming in the course of work of the world's largest corporations. It is possible to use this kind of analytics in small enterprises to correctly allocate resources depending on the incoming information. Since at the initial stage of development of the company is very limited in means, the necessary information can become a cornerstone for all activities. Now the proverb "The one



who owns the information - owns the world" is more relevant than ever [2].

Based on the definition, the enterprise information systems includes such basic elements as information flows (information, database data, document stores, etc. and related software), technical means, human resources, management.

Sources, where information comes from, can be many and it is very important to structure and systematize this data.

Solving problems in the field of telecommunications - the standard area of application of analytical methods. It is with the birth of telephony that the development of many theoretical and probabilistic methods (the queueing theory, reliability, random processes of a special kind) is connected.

Systems, which at the moment can quickly and in an easy understand end-user form, handle a large data flow relatively little. In this diploma project, the main examples of such solutions are considered, and the most appropriate option is chosen, taking into account the specifics of the incoming data in the telecommunications field.

The ultimate analytical system will allow making important decisions at the level of telecom industry executives and not only. As it is known, the task of the company's management is to make decisions. But to make decisions it is necessary not only to have information (reports) from different departments, but also to understand how this information is interrelated. And this is only half the case, since it concerns only internal information. However, each company works to meet certain market needs and depends on its situation (the situation in the market, which is determined by a variety of complex economic indicators). In addition, no firm exists separately from society or the state, and, therefore, it is also necessary to take into account socio-political factors and trends. If the enterprise has sufficiently large sizes, then these tasks are not only complex enough to be performed solely by the head of the enterprise, but require knowledge of special methods, tools and software.

One of the main tasks of the operator companies is to get an idea of how subscribers use the network and what data comes in the course of monitoring the equipment. Such awareness will allows to identify potential products and offers, effectively use investments in resources and see trends before competitors, which in the end helps to maintain the advantage. Analytical systems are an effective tool to reduce downtime of telecommunications equipment and as a result of disruptions in broadcasting. After all, improving the quality of service is considered one of the most important tasks facing the provider of telecommunications services.

Analysis of the data of the telecommunications operator is the key to knowing how well the network suits subscribers and potential consumers, and how to minimize the number of problem areas associated with the incorrect operation of telecommunications equipment. These factors are the most important in planning the activities of the telecommunications provider and further development plans. By analyzing the use of systems, it is possible to better serve subscribers and enhance the company's image among potential customers. Also, the analyst in this case can also be used to determine with sufficient accuracy how subscribers evaluate the

quality of the services provided. So the intellectual analysis of the incoming feedback, including the determination of the general mood of the left commentary on the service provided, is a sought-after technology for companies operating in the telecommunications field [3].

The use of analytical information in the field of telecommunications can help reduce the outflow of customers, increase the loyalty of users and minimize the risks of disruption of broadcasting due to inaccurate information about the operation of equipment.

Using analytical systems it is possible to take data from multiple sources, combine and match them to get a complete picture of all the events on the network. This allows you to maintain a high level of awareness in real time about the status of network usage. The platform of intellectual analytics allows to reveal territories where subscribers are more inclined to leaving and initiate targeted marketing campaigns aimed at their retention. Notifications and automation of a large part of the reporting, if necessary, can also provide accurate results of relevant users and units. Analytical systems also allow you to detect problems on a visual, freely configurable panel, both at the level of the entire network, and easily obtain detailed data at the level of one node or set of nodes. Such panels allow you to configure a graphical display of the situation on the network and provide the most accurate and necessary statistics for all persons interested in improving the quality of the services provided.

If we consider specific examples of the use of analytics in the telecommunications field, then the most appropriate subject for study are broadcasters. They have a large flow of analyzed data, both coming in during monitoring of equipment, and from end users.

Analytical systems help to increase customer satisfaction by increasing the availability of all TV and radio broadcasting equipment, providing significant analysis capabilities for components that control network performance and are early indicators of problems or failures[3].

First of all, it is necessary to understand the sources of data, whether these data are sufficient for decision-making, and then to estimate the amount of incoming data to calculate the infrastructure of the analytical system.

## **1 Analysis of the problem**

At the moment, the solution of the problem of processing a large array of data coming in the process of monitoring various technological equipment is one of the main tasks facing various telecommunications providers in Kazakhstan.

Now the idea of constant improvement of the quality of services is widely spread and it is the equipment monitoring system that provides the opportunity to react quickly to events occurring in the telecommunications network. Round-the-clock control over monitoring systems is conducted and on the basis of the obtained data the team of specialists solves problems of various nature, arising during the broadcast. Eliminating downtime in broadcasting is an extremely important task, since it is on how stable the work of telecommunication equipment depends on the most important indicator in this industry, namely the network availability coefficient. This is an indicator that shows the possibility that the equipment will be working in certain time.

On the basis of this and similar parameters that problems of a republican nature are solved, such as the strategy for the development of entire industries. From the efficiency of the equipment depends not only the activities of the provider itself, but also a number of other organizations that are partners in this case. Performance indicators can be a guarantee of the quality of the services provided during negotiations. Increased customer loyalty, as well as the nature of the relationship between the company customer and the supplier of telecommunications services becomes more trustworthy.

Because of the fact that network reliability indicators are critical and have a national scale, they are continuously monitored by government supervisory structures. Therefore, the development of strategic directions in the work of the telecommunications provider should be based on some analysis. Since at the moment the amount of incoming data has reached the scale when manual processing becomes impossible, it becomes obvious that the development of automated analytical systems is required. This is the basis of assistance is assigned in the process of determining the priority directions in the development of the industry.

Broad and effective application of software and hardware solutions has become one of the factors of the company's survival and success in the face of intense competition. Automated information systems have become widespread.

The problem of analyzing the initial information for decision making was so serious that a separate type of information systems appeared - information and analytical systems (IAS).

Information-analytical systems (IAS) are designed on the basis of real-time data to help in making managerial decisions.

The main purpose of the IAS is dynamic presentation and multivariate analysis of historical and current data, analysis of trends, modeling and forecasting the results of various managerial decisions.

The result of applying IAS tools is, on the one hand, procedural analytical reports oriented to the needs of users of various categories, on the other - means for interactive analysis of information and rapid construction of reports by non-programmers using familiar notions of the subject domain.

In modern conditions, the efficiency of the activities of most economic entities is largely determined by the efficiently organized information support of activities. Virtually any organization, not just telecommunications companies, has computer resources, which allows them to actively use them in the process of processing incoming data, as well as in making a decision. Only a few currently use information and analytical systems to make effective use of available and incoming information. The existence of a coherent information management system can eliminate the probabilistic nature of the management decisions made, duplication of information and its loss, and as a result, to improve management effectiveness.

Consider what are the advantages of implementing this technology and how it affects the effectiveness of entrepreneurial structures.

If we take into account Kazakhstan companies, in which information and analytical systems are actively used, there are many elements left in them that are aimed exclusively at working with databases[2].

The process of making managerial decisions involves the need to process a huge amount of information, a significant excess of information on the physiological capabilities of the human brain on the perception and processing of information led to the need for the use of technical means. In commercial and educational organizations that solve complex problems of allocating significant resources, the price of damage from the choice of not the best solutions is exceptionally high. In such situations that the only effective means of minimizing errors in decision making is the use of special methods, technologies and software tools for information processing, which include information and analytical systems.

In general, information in the management of economic systems is understood as a set of data used to solve economic and, in particular, management tasks. The importance of information in management is indisputable, but traditional systems use information that generalizes the state of the management entity and for the aggregate period of time. Because of what the necessary information cannot always be provided on time, and their generalization leads to some relativity, proximity of values to real indicators.

Creation of information and analytical systems allows to significantly increase the number of processed data and more quickly provide the necessary information, the requirements for which in modern management systems are changing (for example, the use of quantitative management methods).

Of course, automation of managerial decision-making requires more information that has not been recorded and stored in a traditional management system. However, the additional costs of collecting information are justified by more accurate and operational decisions.

The function of collecting and storing information with concomitant revision in information and analytical systems is performed by information storages.



Due to the large volume and complexity, data analysis has two directions - operational analysis of data (information), the abbreviation of the name - OLAP - is widespread. The main objective of OLAP-analysis is to quickly extract the necessary analytics for the justification or decision-making information.

Information-analytical systems are a superstructure over existing information applications in the enterprise and do not require their replacement.

As further programs, which have already been adopted by the government of the RK, imply a significant growth and dynamic development of the telecommunications industry, it means that the tools for information processing must become as modern as possible.

The telecommunications industry is the sphere of activity where the incoming data are very important. To assess the scale of the incoming data stream, it is necessary to consider their sources. So, in the course of the development of the strategy of "Digital Kazakhstan" in the country, more and more attention is paid to the development of data collection and analysis systems. Since 2011, measures have been taken to switch to digital television and radio broadcasting technologies, and work has been carried out to modernize the satellite network with the transition to the DVB-S2 / MPEG-4 digital standard. The introduced satellite TV and radio broadcasting network for the first time provided an opportunity for the residents of the country to receive domestic (freely available) and foreign channels in any geographical location in Kazakhstan. As a result, based on the results of the first half of 2016, the total number of subscribers of the national satellite TV broadcasting network "OTAU TV" amounted to about 1.17 million connections. In parallel, since 2012 the project "Introduction and development of digital terrestrial television and radio broadcasting in the Republic of Kazakhstan" is being implemented. To date, 336 radio and television stations have been introduced, which provide 72% coverage of the country's population with digital terrestrial television broadcasting. According to the state program "Digital Kazakhstan" for 2017-2020, the number of newly installed radio and television stations should be 127, and in 2018 - 227. In total, according to the frequency-territorial plan of the Republic of Kazakhstan, it is planned to complete the construction and reconstruction of 827 radio television stations, which will provide 95% coverage of the population of the Republic of Kazakhstan with digital terrestrial broadcasting by 2019. As a result, 10 to 30 TV channels have been broadcast in the cities of Astana, Almaty, regional centers, and up to 15 TV channels in cities and towns below the regional center [1].

The growth of the Kazakh segment of satellite TV is expected due to 50% of the rural population, which is still without access to cable TV. [2] The expansion of national TV market players to the regions of the country begins by merging with local operators. The demand for mobile Internet is growing due to the ever more frequent use by the population of a mobile device instead of a PC. At the same time, the fixed data transmission market is still relevant in the business environment.

Let's cover the main medium-term trends, which are visible in 2018 on the telecom market in Kazakhstan. The first is the growth of traffic in networks - the

revenues from Internet access remain the flagship of the telecom (they exceed the revenues for mobile communications and for the first time in history exceeded KZT 20 billion a month - the weighted average tenge rate in the US dollar is 326) - the majority of smartphone users (44% ), based on the research of 4Service, consumes more than 6 GB of Internet monthly, and the share of such users has grown by 14%, compared to last year. The second important trend is the continued growth in the share of smartphones. According to various estimates, it ranges from 55 to 60%. If we rely on the research of the mentioned company 4Service, today 83% of all subscribers of mobile operators use the Internet on their smartphones. On the other hand, the average monthly expenses of the vast majority of mobile users (47%) are in the range of 1,000 to 2,000 tenge. At the same time, there was a tendency to reduce communication costs. Finally, another trend - the system one - and it is connected with the said transformation. Operators are trying to find new sources of revenue and reducing costs[3].

Considering such a large scale of work, we can conclude that the development of analytical systems for processing the entire incoming data array and the formation of conclusions based on them is one of the most promising tasks of the entire telecommunications industry. Analysis and structuring of information is a priority area of development in the field of data management in the coming years. Information coming from various sources is one of the most important resources of our time. It is their careful processing that will help create new competitive advantages, as well as improve the level of services provided. And it is the building of the infrastructure of the latest information and analytical systems that is the most necessary method for solving the problems associated with the increasing flow of data.

## **2 Toolset for the system design**

### **2.1 Lucene searching engine**

The basis for the information-analytical system is primarily the search engine. And the right choice is the most important factor at the stage of establishing the basic infrastructure for the system.

For our system, we need an engine with which it is possible to integrate with other programs. One option for implementation is Lucene.

With the help of this search engine it is possible to perform a number of operations, usual for information and analytical systems. These functions in most cases are: performing searches based on compiled queries, indexing, and organizing data storage. Also, based on Lucene, a simultaneous search is performed on the whole set of input requests, together with optimization.

There are two ways to speed up the indexing process. The first is the optimization of individual segments of the index, while the other, almost comparable in time, is re-indexing [4].

The tasks of language support when integrating the analyzer are also important. Initially, Lucene has the opportunity to analyze not only English, but also Russian. But when using Russian in comparison with other solutions, such as Sphinx, Lucene has a rather low speed of indexing. Also significant drawbacks of this search engine is the lack of a full API and various difficulties with working with databases. The implementation of a distributed file system is also possible, as well as clustering, processing and storing indexes, but this task requires third-party solutions. As will be discussed later, Elasticsearch will be used to implement this function [4].

Now consider the main advantage of Lucene over other solutions of search engines. Namely, a simple implementation and low requirements for the format of index files.

To organize a full search, using several documents at once, based on certain keywords for the query, uses the Apache Lucene library. This library is written using the Java programming language, while there are separate ports to other languages and platforms written in C #, C ++ according to[4].

Apache Lucene is an open source software solution that allows to complement the work of individual, self-developed modules. This allows to significantly expand the functionality of the search engine, according to the tasks that will face the information-analytical system.

In order to search, the main data source is used, namely the index. The index is the combined repository of many documents. In this case, the main component of them are given data fields. The main type of data that is received for processing is a string. However, you can configure and additional support for numeric data types, with different retention periods. To create an index, you need a set of documents with marked data fields.

## **2.2 JSON format**

While working on the infrastructure of any information and analytical system, it is necessary to pay attention to such an important aspect as the choice of the format for data exchange. Since most software solutions that will be selected to build an analytical system will somehow be connected to the web interface, the most logical format will be JavaScript Object Notation (JSON). Although JSON can not be called a strict subset, the syntax for this data exchange format has much in common with JavaScript data structures.

This format can be called the most preferable, because it can represent a variety of different types of data. This can be numbers, elements of Boolean algebra, values written down as strings, different ordered arrays of values, and what will most often occur in information-analytical systems, namely objects. Objects in this case will be the pair between the key and the value. In this case, a key can mean a string, but the value can be a large range of data types. So for example, a value in this case can be understood as various primitives and ordered sequences of values and logical key-value pairs. It is also possible to support more complex elements, such as regular expressions of various formats, proprietary functions, and so on. Objects that are associated with dates in some way go through the serialization procedure, which means they will be represented as a string that contains information about the date in the ISO format. This is what helps to prevent data loss. In JSON, it is also possible to represent different types of data by converting a value to a string of the format that you want to deserialize in the future.[10]

In JSON, there are a number of features that are available in XML. Such an example is the ability to store data in a hierarchical form. At the same time, there are quite a lot of software tools between these two formats.

At the same time, JSON also has a text format, which, although it uses different conventions that are used in C-like programming languages, is nevertheless independent of the implementation language. It is this quality that we take as the basis of its choice when building an infrastructure for the exchange and processing of data.

The data can be presented in two main structures. The first is, as already mentioned in the form of an object. That is, in the key format and the corresponding value. In different languages, this idea is implemented a little differently, but the principle in all of them remains the same. The second type is a list of ordered values. In most programming languages, this concept is known as an array of data, a list or different sequences.

And, of course, another reason for choosing this format is JSON support in almost all programming languages.

## **2.3 Elasticsearch search engine**

In the thesis will be considered two search engines, two software solutions of the same problem. Between them, a comparison will be made for individual functionalities, but first, need to consider each decision separately. The first tool for compiling search queries based on keywords will be Elasticsearch. It can perform a



wide range of functions, especially if the stack of technologies offered by Elastic is immediately applied, but it was originally designed for text processing, and the result is also in the form of text according to the query, together with the statistical result[5].

At its core, Elasticsearch in the infrastructure of the system is almost completely standalone database server, which is written in the Java programming language. Its main functionality includes: the adoption and storage of data in a complex format, previously optimized in the form required for search. Another advantage of this system is that the work with it is facilitated by the fact that its main protocol is implemented using JSON, a description of which is also given in this thesis. Scaling of the system based on Elasticsearch is implied initially after the installation by clustering.

The core for Elasticsearch is the Lucene search engine, which was previously considered. Elasticsearch in this case is an infrastructure built on the basis of Java libraries Lucene. The main functional that is used in this case is the actual algorithms for implementing partial text negotiation, and also for storage in the form of indices available for search queries. Elasticsearch has its own functional API.

As mentioned earlier, the use of Lucene alone is inappropriate, since it was originally designed to integrate with other software products. Elasticsearch provides an intuitive interface for working with both API and search queries.

It is possible to distinguish several key features that visually reflect the essential advantages of Elasticsearch, over a small Lucene functional:

- 1) Simplified work with the API
- 2) Interaction with a number of programming languages, not just the Java family.
- 3) Scalability by clustering and replication
- 4) Optimal values initially for working with complex classes of the search engine Lucene

There are a number of tasks for which Elasticsearch will be used in the process of building an infrastructure for data analytics:

- 1) Search for a large amount of data for the most accurate match with the compiled search query.
- 2) Search based on specific words in the query
- 3) Implementing autocomplete based on sample words, and the earlier issued query results.
- 4) Storage of a large number of not fully structured data in the form of a JSON format.
- 5) Creation of unique records, the distinctive character of which will be regulated by a number of conditions [5].

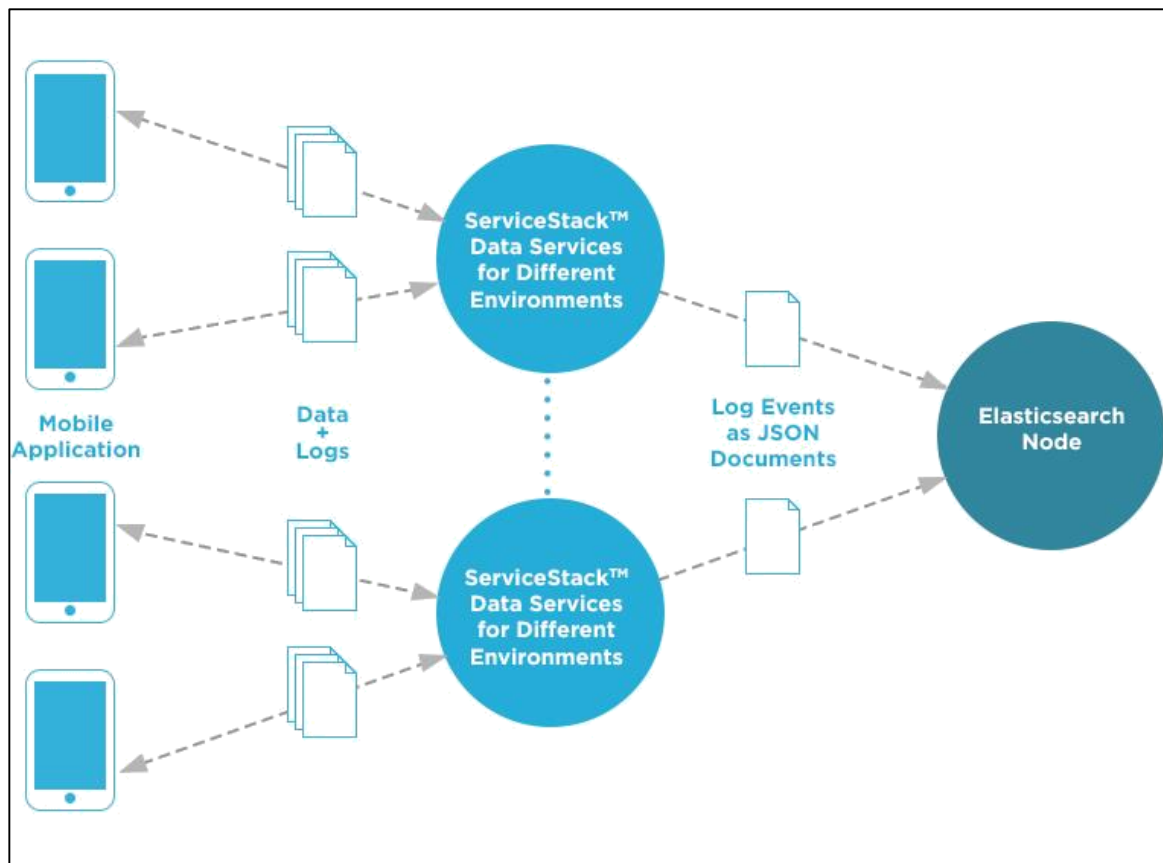


Figure 2.1 – Infrastructure of Elasticsearch

Although Elasticsearch is usually used to provide approximate results for queries, there are, however, opportunities for accurate comparison and statistical operations. In this case, the main function of this search engine is still to provide the most approximate results, but not exactly the same. This property separates the search procedure in Elasticsearch from traditional solutions used in databases.

That is why during the construction of the infrastructure for the analytical system, Elasticsearch was chosen as a tool for processing data and displaying them in the form of a statistical set of information [5].

**2.3.1 Elasticsearch client.** Another functional feature of Elasticsearch is the existence of a low-level client. Its main purpose is to provide a basis for all programming code in Python, which is related to the Elasticsearch functionality. That is why the main principles of Elasticsearch client are independence from other software products and wide opportunities for expansion [8].

To use the client library, which can be considered high-level it is necessary to use functionality of Elasticsearch-dsl. This is a kind of more python-oriented library, which is oriented towards interaction with product, named Elasticsearch.py.

The main purpose of developing a client in Python is to provide the most possible flexible solutions for interacting with the REST API. Therefore, you can make the assumption that using some APIs with Python will be difficult. To solve this problem, some software solutions were developed, in the form of an even higher-level library. All these actions are aimed at ensuring the simplest use of

Elasticsearch for narrowly focused tasks.

It is possible to consider the key features of the Elasticsearch client such as:

**Continuous connection.** Elasticsearch.py uses a persistent connection to provide communication between individual pools. For this, a single connection is used for each individually configured node. To ensure this interaction, it is possible to use two different implementations of the HTTP protocol. The transport interaction layer creates one instance for each preconfigured connection class for each network node. In this case, if one node is inaccessible, it goes into a timeout state. It is assigned a certain class, and the restoration of the data stream occurs only after a time-out period has elapsed. By default, all nodes on the network do not have an exact hierarchy until they are moved to the pool. After the move, various techniques are used to balance the resulting load between the nodes [8].

At this stage, there are opportunities for using the API. So, passing the arguments for the Elasticsearch class sets up a policy for the interaction of network nodes. In order to take advantage of third-party functionality that is not initially supported, there must be a skill in creating a subclass for the required component. This subclass is passed as a parameter. In turn, it will already be used by default.

**Automatic reconnection.** After a failure occurs due to a connection problem, the node goes into a fault state. At the same time, it will be held in the state for a time-out for a predetermined amount of time. After this time expires, a second attempt will be made to connect to the other node. In this case, the timeout duration will accumulate with the growth of unsuccessful connection attempts. This makes it possible to reduce the probability of distributing the load to the node that, according to observations, is in an inoperative state. In this case, if no node in the network is available, the order of connection will be based on the principle of the least time-out.

**Sniffing.** To ensure that a single node can be tested, as well as whole joins, the sniffing function is used. Also, SSL is configured for this purpose. The client thus allows you to connect to the very cluster of Elasticsearch, after having passed the stages of different kind of verification.

**Logging of processes.** In order to apply registration processes from python, two loggers are used: the product itself and elasticsearch.trace. They differ in the objects of application. For example, the first is used to register standard user actions. The other, in turn, is used to register individual requests to the server. This process is carried out by means of commands written in curl. In this case, they can be used with json, launched from the command line.

**2.3.2 Elasticsearch API.** The main principle of functioning in the API Elasticsearch is the mapping of the original REST API as close as possible. This implies the difference between the arguments according to the obligation. From this we can conclude that there is a significant difference between calls with different types of arguments. However, the recommended type is keyword. This is due to the issues of consistency and system security of functioning

Elasticsearch is client of low level. But along with this, it is possible to

directly display information which received from Python directly to the endpoints. In this case, there are a number of attributes that provide access to different clients in API.

For a full-fledged configuration, you can set your own class for your own tasks. In doing so, it will be used to provide the connection class parameter. In order to understand the basic capabilities of the API, such as creating and deleting, retrieving information, and so on, you need to consider the basic parameters by which they are invoked. And also the functionality that they cover after the call.

Create parameters. In some cases it has the meaning of indexing. Because of the main functionality which described as adding JSON document, indexing and making it searchable [8].

Table 2.1 – Description of the Create method parameters

Name	Description
index	it is possible to specify each index by the name
doc_type	It is possible to work with several types of documents.
id	the authentication of each document by the certain number
body	The operating document itself
parent	The specification of the parent document is also occurring by the number of it
pipeline	The processing is based on the principle of pipeline, and this parameters shows the preprocessed id number.
refresh	It is depend on the option. In case of true this parameter affectes shard and makes the processed operation visble for search. In the case of wait for makes the visibility of operation. And in the case of false, which is default value it makes no refresh.
routing	In the process of routing also there is the specification by the number
timeout	Each operation has its own timeout and this parameter explicates it.
timestamp	Not only the operation has timestamps but documents too. This parameter explicates timestamp for it.

Delete parameters. Delete a typed JSON document from a specific index based on its id.

It is possible to see the main parameters which are used during the work with the function Delete.

It includes the similar functions as with create, but also some of them are different [8].



Table 2.2 – Description of the Delete method parameters

Name	Description
index	it is possible to specify each index by the name
doc_type	It is possible to work with several types of documents.
id	the authentication of each document by the certain number
body	The operating document itself
parent	The specification of the parent document is also occurring by the number of it
pipeline	The processing is based on the principle of pipeline, and this parameters shows the preprocessed id number.
refresh	It is depend on the option. In case of true this parameter affects shard and makes the processed operation visible for search. In the case of wait for makes the visibility of operation. And in the case of false, which is default value it makes no refresh.
routing	In the process of routing also there is the specification by the number
timeout	Each operation has its own timeout and this parameter explicates it.

Exists parameters. Returns a Boolean value indicating whether or not given document exists in Elasticsearch . Get parameters. Get a typed JSON document from the index based on its id. Get parameters are the same with the Exists.[8]

Table 2.3 – Description of the Exists and Get methods parameters

Name	Description
index	it is possible to specify each index by the name
doc_type	It is possible to work with several types of documents.
id	the authentication of each document by the certain number
_source	There are two values: True and False. From it depends will the the source field return or not. And another option is a list of fields that is possible to return.
_source_exclude	The parameter that will shows the list which is possible to exclude from returned field.
_source_include	The parameter that will shows the list which is possible to include from returned field.
parent	The specification of the parent document is also occurring by the number of it
preference	With the help of this option it is possible to make the specification of shards or nodes. It describes the necessary operation that should be executed. By default the value will be random
realtime	There are to modes of the operation. The first one is the processing in real-time and this parameter is calls for it. The second is operating in search mode.

Table 2.3 continuation

Name	Description
refresh	It is depend on the option. In case of true this parameter affects shard and makes the executed operation visible for search. In the case of wait for makes the visibility of operation. And in the case of false, which is default value it makes no refresh.
routing	In the process of routing also there is the specification by the number.

## 2.4 Kibana as a tool for visualization and analysis of data

Above, the search engine Elasticsearch was used, which is used to obtain results based on search queries. Now, when building an infrastructure, the task is to visualize these results. There is a need to search, interact with data and view them in a convenient graphical form. To solve these problems, a software product, also with open source, is used for visualization and analysis called Kibana. With the help of this tool it is possible to visualize data in the form of various diagrams, maps, dashboards and maps.

There are four main sections in the interface of this product, with which the visualization management process is implemented, namely: Discover, Dashboard, Settings, Visualize.

Each section has its own functions, aimed at the overall development of the infrastructure and for the implementation of the process of continuous improvement.

Kibana simplifies the visualization and analysis of large data samples. The Kibana interface consists from follow sections: Discover, Visualize, Dashboard, Settings [6].

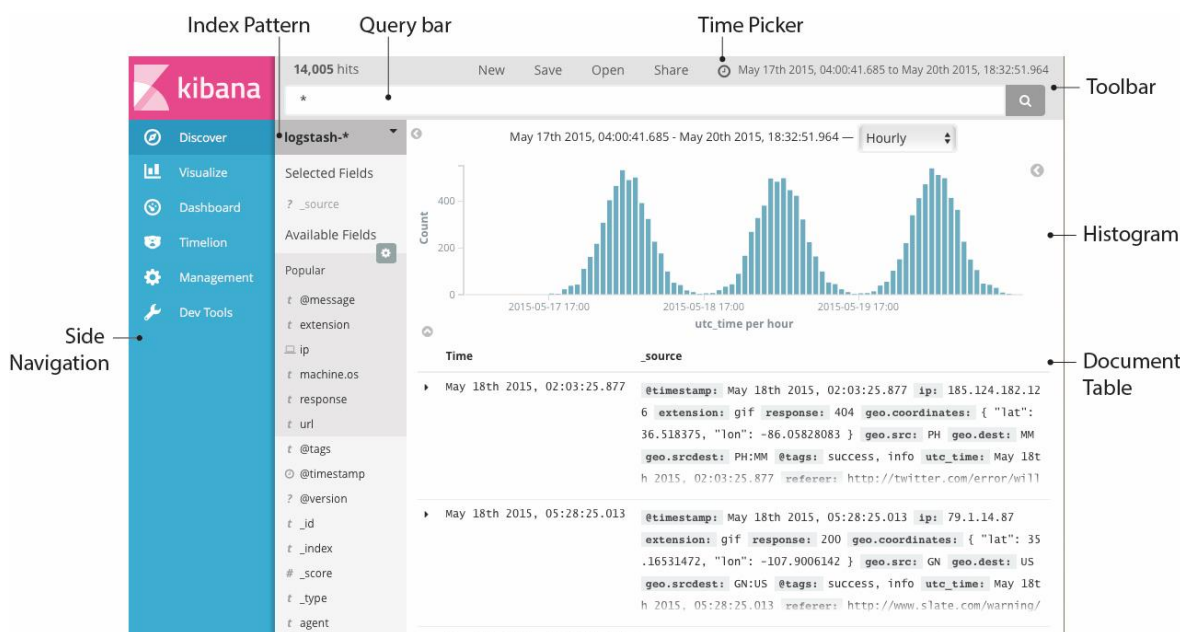


Figure 2.2 – Interface of Kibana Discover

Consideration of the functionality of the main sections must begin with Kibana Discover. Initially, according to the preliminary settings, the last logs of the ELK stack will be displayed in this section. Also, there are opportunities for filtering and compiling search queries based on the input log. At the same time throughout the section there is a possibility to select the time range of interest.

The elements of this section are:

- 1) Search bar. It is located in the main navigation menu, and is used to search for specified fields and whole lines
- 2) The time filter, which is used to filter the results based on both relative and absolute time scores.
- 3) Field selector, to select fields for visualization.
- 4) Histograms that initially display all information received from logs and have the ability to further fine-tune
- 5) View log log itself, to view the selected messages, as well as visualize the results obtained during the filtering. If the setup procedure was not carried out in advance, all of the log messages will be displayed.

Since Kibana is primarily a data visualization tool, the "Visualization" section also deserves a separate review. It allows you to create, modify and view your own visualization data. This software solution provides various types of visualization, such as vertical bar, pie charts, maps, tables, etc. It is also possible to share the results visualization with other users [6].

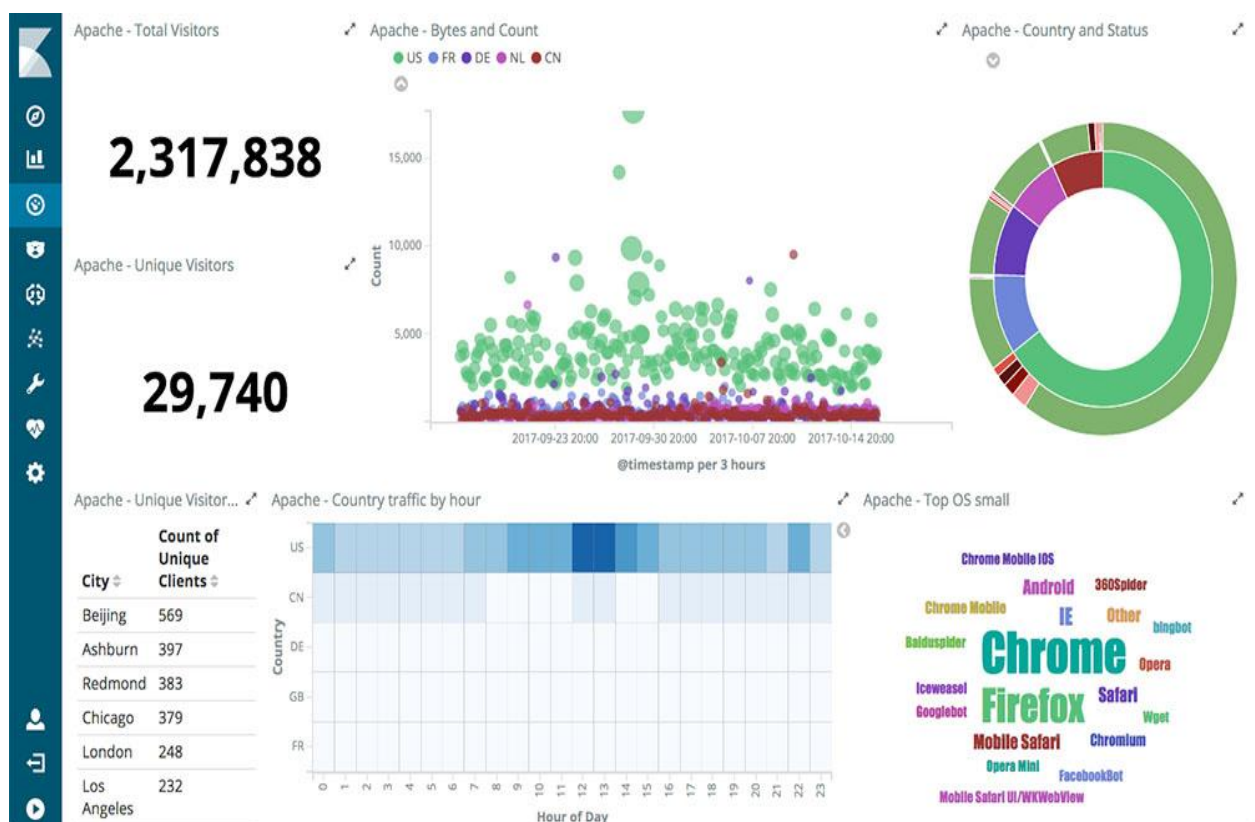


Figure 2.3 – Interface of the controlling panel of Kibana

Thus, it will be possible to flexibly configure the visualization of incoming

data, calls and problems, depending on the needs and the level of admission to the information of each individual user of the developed information and analytical system.

On the Kibana toolbar page, it is possible to create the dashboard, which will represent the united information, obtained during the analysis. Using the toolbar, the process of combining several visualizations into one page, and then filtering them based on a specific search query, is performed [6].

2.4.1 Design of reports in Elasticsearch. Automating the processes of reporting is one of the most important tasks facing any enterprise in every field of activity. This is due to the need for monitoring and tracking the results and incoming information. Each company, especially in the telecommunications industry, where the information from the reports can depend on the functioning of critical equipment, is engaged in the process of improving and accelerating the formation of reporting. It is for these purposes that the ELK stack has the ability to generate reporting documents based on the behavior of users in Kibana. All of them are formed as separate documents in Elasticsearch.

At the same time, it is possible to share access to information, and therefore, to reporting by using X-Pack, with the use of modules responsible for security. In this case, the application of the Watcher software solution is possible under the condition of additional permissions for the Kibana application [11].

In order to generate reports, you must distribute roles in the system. In this case, the `report_user` role is used, which directly interacts with the reporting system. The second role is a user who interacts with Kibana, and has the ability to configure their own reports. Names for this user are set by `kibana_user`.

At the same time, Kibana has its own administrative capabilities for assigning roles. This is done through the management interface by users or through the API. The API in this case has quite flexible features.

When developing the infrastructure of the system, it is necessary to understand the algorithm of reporting in Elasticsearch. In a phased manner, it looks like this:

- 1) For proper reporting, you must install an additional License plug-in. It's pretty easy to find in the installation directory of the Elasticsearch itself.

- 2) Since Kibana plays a key role in the formation of reports, it is the changes that take place there that are reflected in the documentation, an additional Reporting application is needed

- 3) All changes in the settings should be reflected in the configuration file Kibana - `kibana.yml`. Therefore, we separately set the identity for the `report.encryptionKey` for authentication.

- 4) The Kibana application is launched directly.

- 5) In order to verify that the process of installing the reporting modules was correctly performed in the browser on the address `localhost:5601`, check the presence of a button in Kibana "Create report" [11].

Forming the reports is available in several types such as PDF, XML.

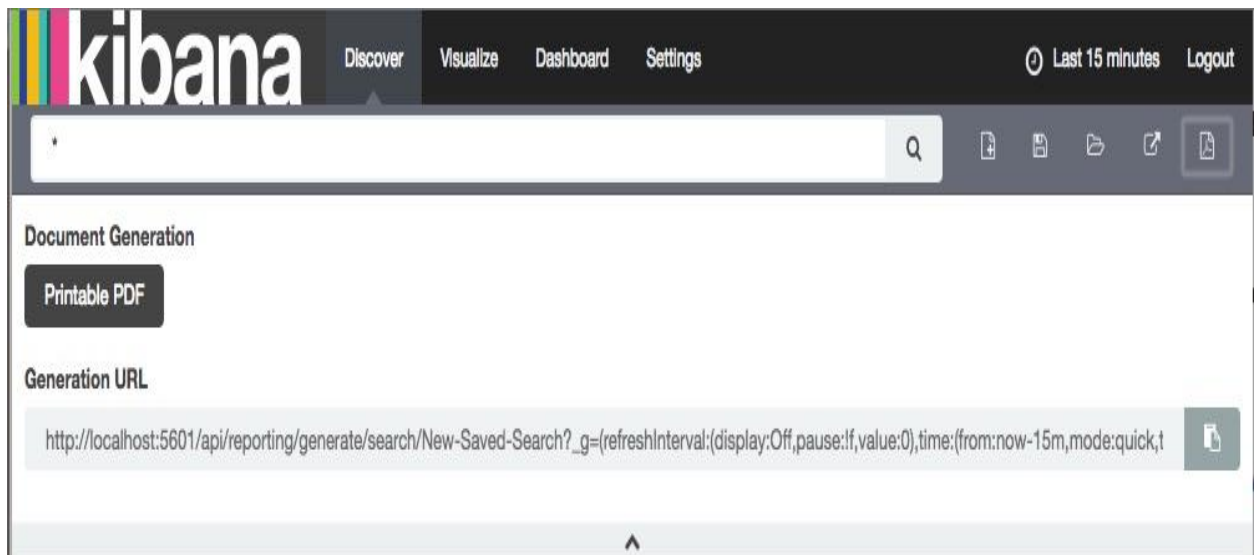


Figure 2.4 – Result of the reporting configuration

## 2.5 Logstash as a tool for log collecting

Logstash (or LS) is a powerful tool for collecting, organizing and analyzing logs, which helps to get a general idea of the environment, and also to detect server problems in time. One way to increase the efficiency of the installation Logstash uses filters to collect application logs and structure data, so that data can be easily accessed and analyzed [12].

Grok parses text patterns with regular expressions, and then assigns an identifier to them.

Logstash filters include a sequence of grok templates that find and analyze different log messages, and then assign them identifiers, so that the logs are structured.

Input, filter and output blocks can be any number. It all depends on your needs and capabilities of iron.

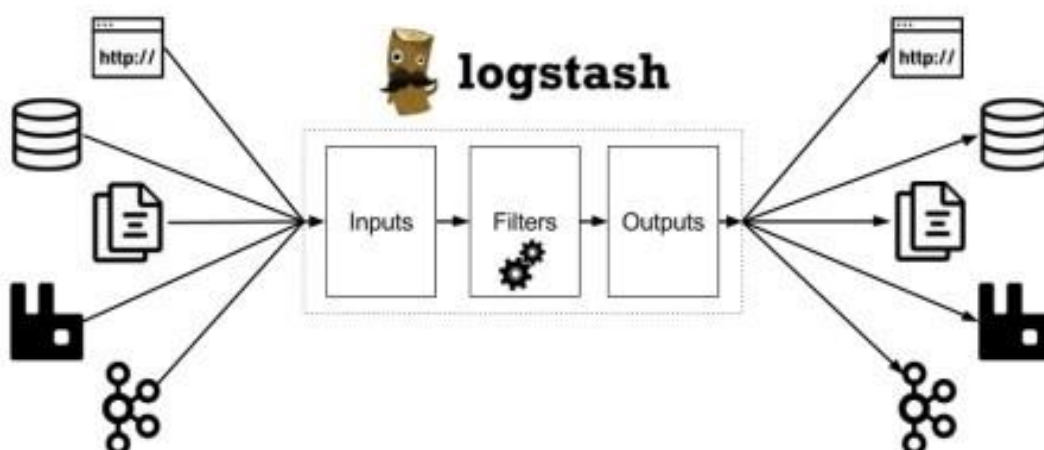


Figure 2.5 – Infrastructure of Logstash

Empty lines and lines starting with # - Logstash ignore. So commenting the configuration files will not cause any problems.

### 1. INPUT

This method is the input point for logs. It defines the channels on which the logs will go to Logstash.

### 2. FILTER

In this block, basic manipulations with logs are configured. This can be a breakdown by key = value, and the removal of unnecessary parameters, and the replacement of existing values, and the use of geoIP or DNS queries for IP addresses or host names.

### 3. OUTPUT

The name of this block / method speaks for itself - it specifies the settings for outgoing messages. Similar to the previous blocks, you can specify any number of outgoing subblocks here [12].

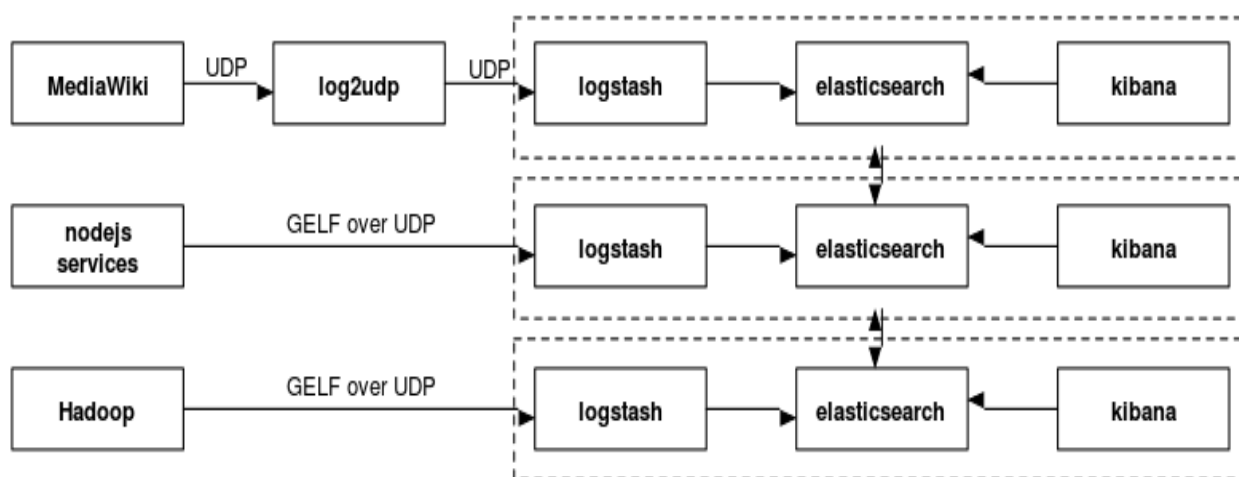


Figure 2.6 – Ways of interaction of Logstash and Elasticsearch

Yes, Logstash works great with Elasticsearch, but the sources and recipients of data can be a huge set of services, from message queues to TCP sockets.

But the process of describing this tool skipped one element - the user interface. Whatever you say, analyzing logs from the command line is not the most productive activity. To fix this, the next component of the stack ELK, Kibana, is discussed.

## 2.6 Splunk analyzing system

Splunk is an information-analytical system that performs distribution functions, as well as real-time log analysis. A brief description of the basic principles of functioning is possible in the following way: the infrastructure consists of the Splunk server, which deals with what indexes, and also performs primary analytics of the log flow, as well as machines that generate the source data and pass these logs to the server. The very same server infrastructure in its case can be a whole cluster of distributed physical machines. The information between them is divided according to the technical capabilities for processing, using the MapReduce



technology. In order to transfer from the machines that generate information to the server, Splunk has a product forwarder. It deals with the fact that it sends log changes in real time using SNMP or NFS / SMB management technologies. At the same time, a script is also possible for manually sending data to Splunk through a stack of TCP / IP technologies. Since in most cases Splunk integration is done with Windows, it is also possible to process data coming from Windows Events and the registry [13].

Logs that arrive for analysis are information broken down into separate lines. The end result obtained during indexing is a field where a direct match is established. An example can be a string name-value. To send requests to fields, a special programming language is used that is oriented to Splunk. With its help, it is possible to filter, sort, merge, format individual tables, access vocabularies and, based on these data, perform visualization. SPL allows you to work with separate lines, and combine and group multi-line objects up to one line.

In the Splunk system, all processed logs are available for queries. This means that the concept of archiving is missing. In this case, the servers through which the information passes must display the full amount of data that they locally store and process [13].

The main interface Splunk is a web service. This is where the tools for creating their own dashboards are located. It is on their basis that it is possible to format applications for Splunk for their own tasks. Also, Splunk allows you to use ready-made analysis configurations that are publicly available.

2.6.1 Splunk features. Using a lot of preconfigured and also customizable input ways, the Splunk system use basically all types of sources that can appear during the work of information-analytical system. Data which is usually file-base should be sent through senders that are directly located with the close interaction with sources of data, at the same moment DevOps, and another different kinds of data, such IoT are possible to be accessed by event collector such as API or also possible TCP or UDP port. It is also possible to extract information using sources which are based on API-based structures by using some modular inputs. Usual sources of information about IT or also important themes such as data and information security applications also possible to embedded and analyzed [13].

Real Time Architecture: Splunk collects, searches, monitors and analyzes data in real time on a variety of large enough (hundreds of TB of data per day) and all this - one system.

Splunk can provide real-time data collection from thousands of heterogeneous sources - and this can be either a physical or virtual host, or a cloud. Also Splunk supports searching not only in real time, but also over the entire time interval, the data for which were collected. That is, it is possible to search, monitor, alert, report and analyze for any time (historical data and real-time data in one solution). Finally, Splunk provides speed in obtaining results and interactivity of search queries on large amounts of data.

**Universal Machine Data Platform:** Splunk is a universal platform for machine data that provides comprehensive data collection, processing and analysis. Thus, we can index any machine data with a timestamp regardless of the structure and format. Splunk is able to combine the machine data + user data+ business data, which makes it extremely versatile

**Schema on the Fly:** Splunk searches for time, that is, you do not need to know the data structure in advance to generate a query. You can choose a time interval, enter a couple of keywords and quickly get acquainted with the data. There are no hard limits on columns, tables, and so on. This greatly increases the flexibility of the system. Also any request can be stopped, put on a pause or show intermediate results [13].

**Agile Reporting & Analytics:** Splunk provides the ability to build analytics, reports and their visualization. In addition to the target data, the system can also access external directories, for example, in SQL databases. Also I would like to say that Splunk is quite an open system and it is possible to add your module, although the visualization capabilities are quite diverse.

**Scales from Desktop to Enterprise:** Splunk uses MapReduce technology, which provides load balancing and horizontal scalability of the system, that is, we can start from one server for Splunk, and with the increase in data - quickly add a couple of new servers and distribute the load. Also, thanks to MapReduce technology, Splunk can quickly process really large amounts of data without requiring an outstanding hardware.

**Fast Time to Value:** Splunk allows you to quickly get the result from use. Implementation takes hours or days, not weeks and months. The same with scaling and operation [13].

Metrics are obtained in time and also compressed, in some cases stored and processed for the further extraction numerical data which is much more effective than regular solution as logs. It is recognized as a data with the first class priority, suitable for purpose of scaling and another kinds of performance. Using metric data increases the speed by 20% compared to previous versions.

The creation of scheme is occurring automatically during in the process of search and also can work with data that comes in real time. As the data is usually obtained in an unprocessed format, it is not necessary to buy expensive ETL or another products which are related to the normalization of data , and also there is no need to create the schema manually by adding some new kinds of data performing new search processes.

When entering different data which is appears in real time, the processes of extracting and also the normalizing of timestamps starts to be necessary to troubleshoot ( to answer the question what and when something gone wrong) research and deep research of transaction flow from one end to another. The time of any event can be determined by the Splunk system, sometimes in the untypical and also unconventional formats. Data which has no timestamps, will get it after the analyzing of context.

Behavior and activities analytics view through the historical context using the same interface as above. It is possible to use not only simple boolean queries in the field searches, but also statistical searches and subroutines. And then visualize the results, see the templates and compare all the necessary information. It is possible to save queries and plan to execute at same specified intervals to start the element which called dashboard.

Sending e-mail, different activities such as order on the site and calling via VoIP technologies as result create several events in different IT elements. It is possible to use this information for each transaction to identify the most problematic areas.

Increase and decrease allows to identify main trends and anomalies . Dynamically installed in the dashboard everywhere including charts until unprocessed events or define user views and decrease the appearing noise.

Optimize the execution of queries for the bigger data set. Using the data sampling allows to process events with the bigger speed, helping to analyze large sets with the speed of their appearing.

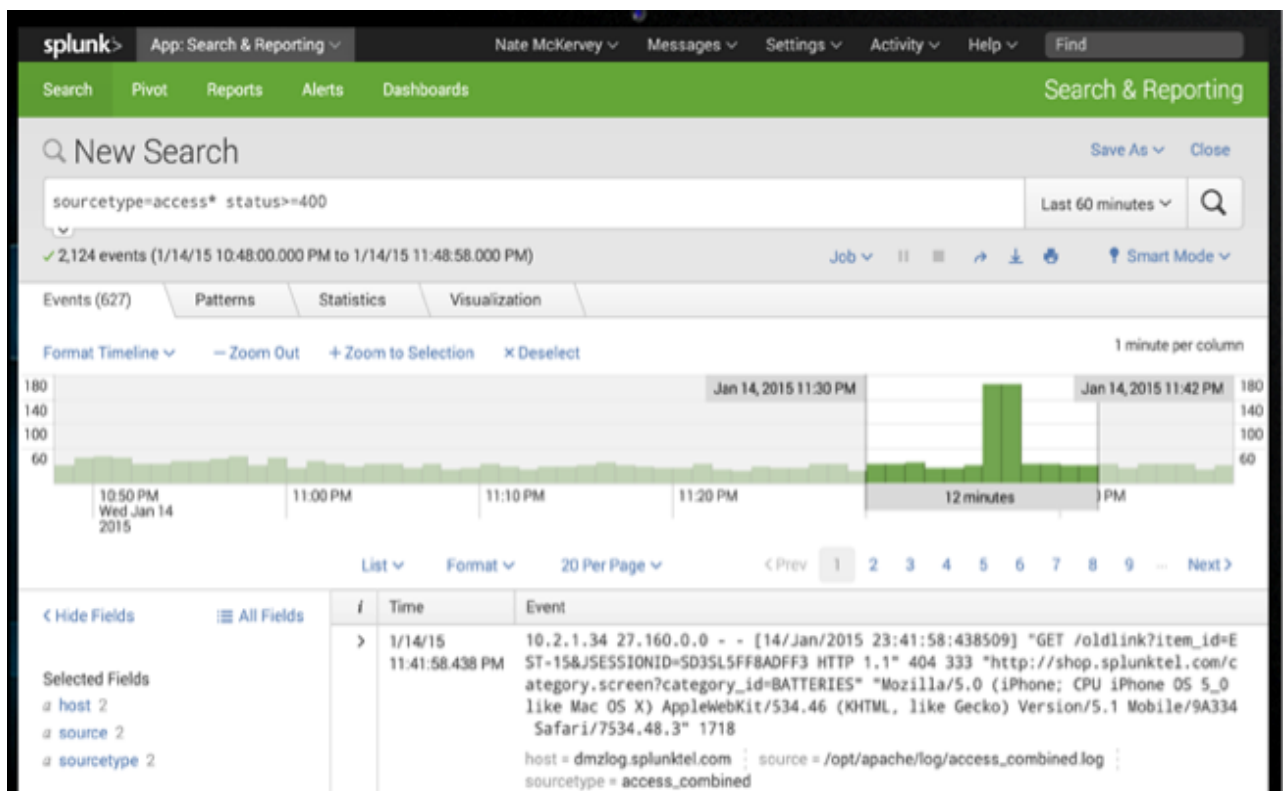


Figure 2.7 – Real time search in Splunk

To avoid various downtime in the processing of information in various enterprises, and primarily the telecommunications industry, it is possible to create own models of analytics and also use models on machine learning Splunk. In this system, it is possible to create own models based on some built-in solutions in the field of machine learning, which means enhanced API, the ability to segregate information and management based on learning systems and some pre-configured

algorithms that improve the efficiency of product use. It is also possible to create algorithms based on popular Python libraries, which are open source.

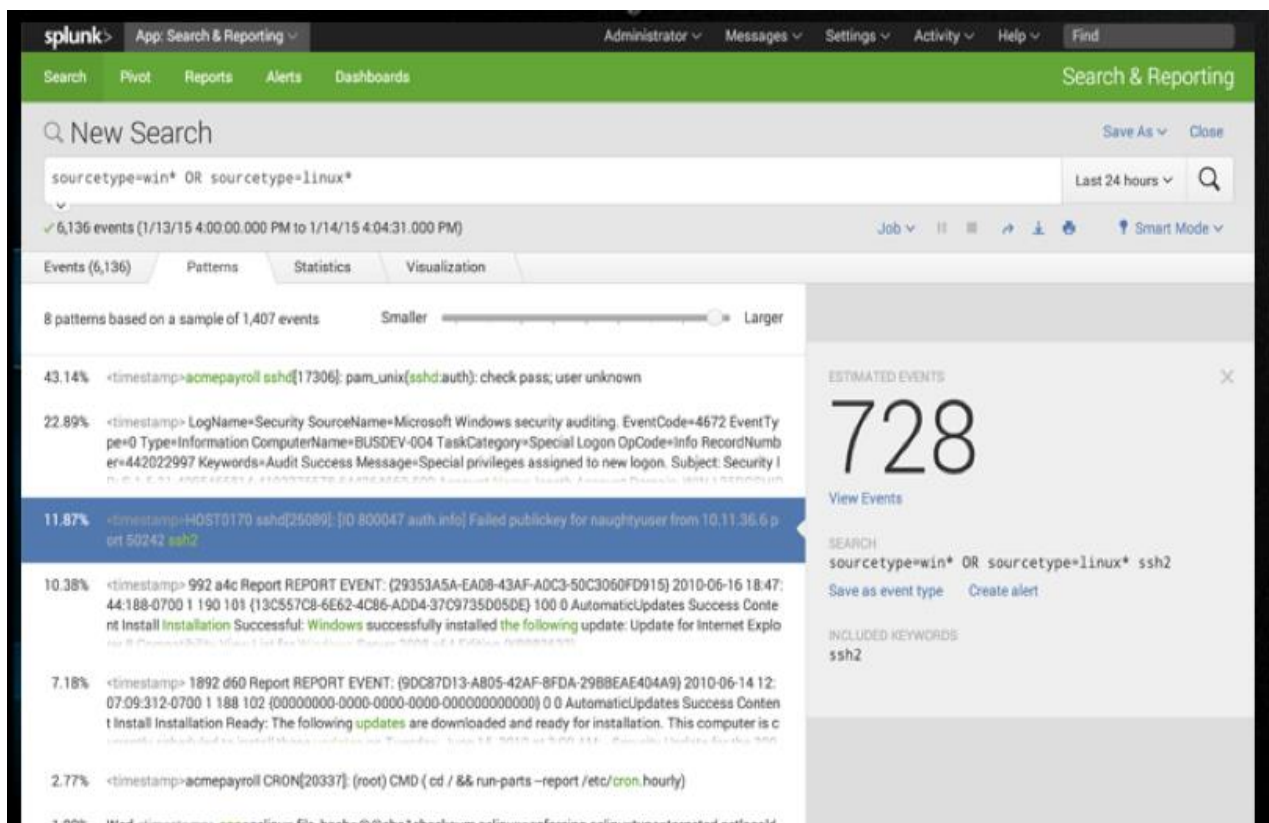


Figure 2.8 – Correlation between coming events

Correlation of possible events is a purposeful search for relationships between individual events in a data stream that initially seem unrelated, provided that the data can come from different sources. For example, it is possible to track several events displayed as a series as a single transaction in order to be able to measure its duration or status at the current time. It is also possible to automate some correlations results in order to be able to create various kinds of warnings in case of deviations from specified indicators or to track various business indicators that are most important in some areas of activity. The considered information-analytical system Splunk is able to ascertain the correlations of various events based on the time and geographical location of the source of information, as well as the sequence of receipt, search, association and various transactions [13].

This solution also provides the ability to automatically determine some patterns in incoming machine data, as well as templates, regardless of where the information came from or what type of data they have. In order to identify different trends, outbursts, or details of information, it is possible to use an increase or decrease in scale using a visualized timeline through the entire information flow.

To develop some structure in the data set, it is possible to use defined data sets as a models, different tables and some kinds of search inquiries. These structures can be used as a building element for analysis and reporting. Data models show the relationship between different data without the need to process a real flow

of information. Structured mapping of complex data flows is possible using tables. Various considerations make it possible to increase and expand the possibilities of applying existing data and events, due to the fact that they interact with external resources.

The user of the system is given the ability to create a tabular representation of data, to which it is possible to provide multiple access to a wide range of users.

Research tables can improve, filter, refine and collect data using some elements of the interface, without using any SPL. After preparing tables, it often becomes necessary to share structured data with a number of other users, and it is for these tasks that you can use Pivot including and to create highly narrowed reports and monitoring dashboards [13].

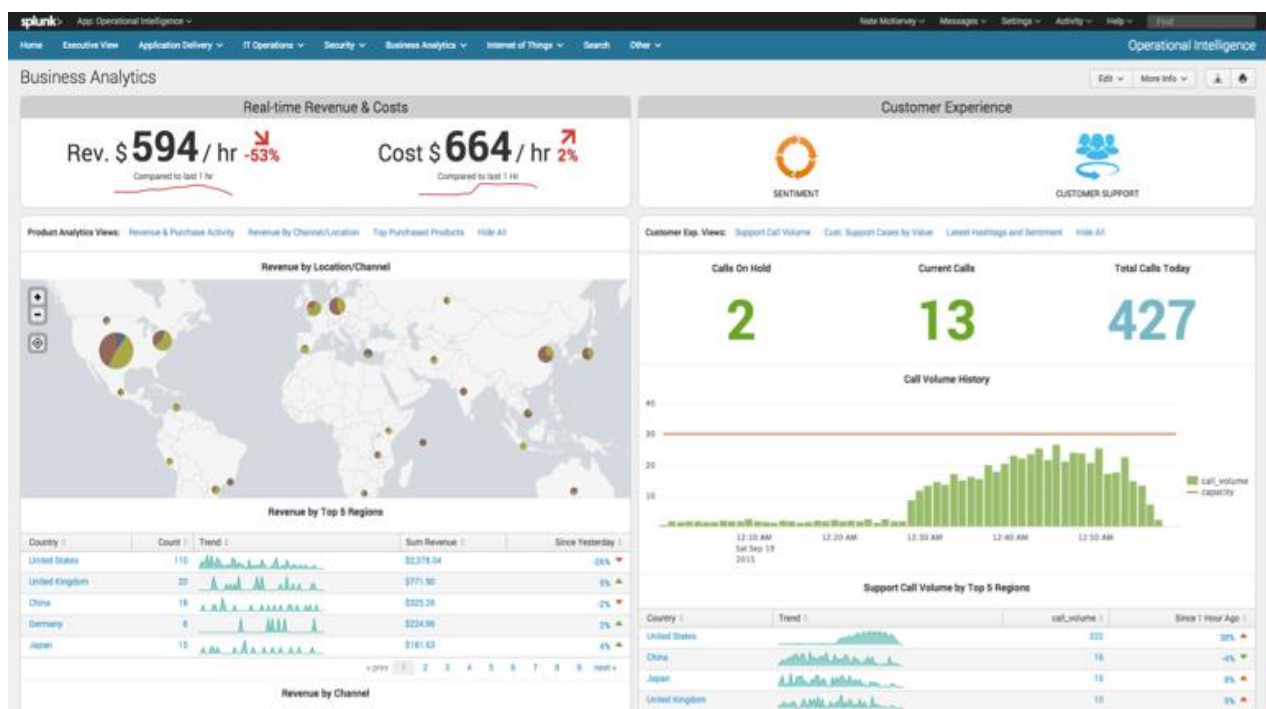


Figure 2.9 – Splunk Dashboard interface

One of the main tasks now, which are facing the developer of information-analytical systems is to make the processing results understandable and effective. For this, it is possible to select the necessary tools from various diagrams and various kinds of visualizations. Intuitive diagrams and visualizations with interactive capabilities help on the basis of complex data to identify problems, both existing and potential, as well as opportunities for improvement. In the system Splunk there is a set for visualization, which provides ready-made solutions for some tasks that are in the process of analytics.

Combining several charts and reports occurs on the toolbar, where it is also possible to select some items for reuse. It also creates and configures for its own need information panels that are used in various areas, such as: management analysis, security, audits, business processes, operating groups and software developers.

Panels can be created to use different libraries for joint and complementary

functioning.

Customizable internal processes allow you to navigate to other panels, forms, as well as various related external websites. It remains possible to go to the basic data through the corresponding results in the dashboard.

Users can easily view users through mashup files with various web applications, and also configure integration between security consoles and programs for managing enterprise resources. Monitoring panels can not only be viewed, but also edited on mobile devices, since graphics and time frames do not use Flash in their operation [13].

Analysis and retrieval of information allow solving management problems and uncover some unused capabilities, but at the same time constant monitoring of emerging events, conditions and critical indicators for the company in terms of efficiency, allow to ensure a continuous flow of all transactions arising in the course of economic activity.

In order to inform responsible teams and management, it is possible to use pre-configured searches and algorithms to create dashboards in real time and further visualize the data.



Figure 2.10 – Events monitoring in Splunk

A variety of critical events and the emergence of conditions for their appearance can be signaled by customizable warnings. For example, the function for user warnings allows you to call any script or algorithm of actions, which can be sending letters to a certain circle of people or starting the correction procedures. In the course of the activity of the telecommunications provider, it is possible that different types of events arise, which cause the need to send alerts. They can be set



to some specific threshold value, in templates for a particular type of equipment and to emerging trends or methods of fraud.

Data security is now one of the key factors in the selection of information and analytical systems, since information must necessarily be stored in the technical infrastructure. Splunk in this case meets the standards of information security, allows for safe processing and analysis of data, as well as provide control by auditing companies. This also makes it possible to integrate with the already implemented enterprise solutions.

To ensure the disguise and confidentiality of data received from various sources, encryption access to the streams and TCP / SSL protocols are used. Users' access is protected using HTTPS or SSH, which limits access to the command line.

In order to monitor the actions performed by users, which tools can be used to solve specific tasks, which dashboards can be used to retrieve data-all this allows you to implement access control elements and system audit, based on the distribution of the level in the hierarchy of the company.

It is possible, however, to create your own access policies for the entire enterprise based on the differences between user classes.

In order to monitor, track and record what actions can be performed by users, which tools to use and what data can be accessed, access control and audit elements are used. It is possible to set up an organization policy mapping and a class of the user who is requesting information [13].

For single sign-on to popular data identification solutions, you can configure Splunk integration with SAML. There are also integration possibilities with other software solutions such as Okta, CA SiteMinder, PingFederate, OneLogin and others.

## **2.7 Comparing of ELK stack and Splunk**

The systems Splunk and ELK essentially use different applied methods to solve one global problem. In the process of choosing one or another software solution, people rely on the structure of the organization and on the amount of time that they are willing to devote to the analysis of journals. The system Splunk allows you to search for information based on a large amount coming data extract the desired result. ELK may require more time for organizing, plan and perform the structuring of the processing at the very beginning of the project, but close to the end of the project result justifies time which was spent [14].

Splunk bases on three structure components, the first of which is the forwarder. It passes the incoming data stream to the remote indexers. The second element is an indexer which performs the tasks of storing, then indexing and responding to incoming search queries. The last component of the processing of information is the search head which executes the functions of the main web interface of system, where the actual merging occurs, and in the case of a large scale of analysis and distribution by servers. In Splunk, you can also integrate most functions into separate applications with the help of SDK. The default options which includes the security operation, operational monitoring, and behavioral

analysis of users. Since Splunk is a software product with the fixed price in it, the usage billing is generated by indexing of the steam volume of the analyzed data.

ELK is a complete solution, which unlike Splunk is an open source product, and consists of three main software solutions - Elasticsearch , Logstash, Kibana. All these products are Elastic's own development, and they are provided with full technical support from the company [15].

Elasticsearch itself is a search engine that bases on the search engine called Lucene , while in its essence is the base of this NoSQL. Kibana is a solution that facilitates data analysis through visualization and customized monitoring panel, but also is a kind of graphical instrument panel. The processing and transport unit is represented by Logstash. The main function of it is to fully fill Elasticsearch with data and integration with other areas such as Kafka, Graphite, Nagios, etc. is also possible.

Both systems, such as Splunk and ELK, can be used as the monitor applications, business intelligence, security system, and primarily for the process of monitoring and analyzing infrastructure in various IT operations.

The process of comparing these two systems was performed on the basis of key functions, and provided solutions from different developers.

Loading data. The solution from Splunk in this case is relatively simple. Immediately after installation, the component forwarders get the initial, preliminary configuration in order to select from a large number of data sources, which can be both files and directories, as well as various events that occur on the network, Windows sources, as well as data coming from logs applications.

In the ELK stack, everything is a bit more complicated. To send data from the initial source of information to the final destination, Logstash the kind of software that is used. However, for this you need to make a setting, so that each search field has been identified before it is sent directly to the Elasticsearch. This can be difficult if the system user does has no experience of work with any scripting programming languages, but in this case there is strong support in the community.

Visualization. In order to edit some visual instruments and add updated components to the monitoring dashboard, Splunk Web has its own web interface that includes flexible solutions to the controlling function. It is also possible to create custom management elements, which can also be configured differently depending on the hierarchy for different classes of users. In this case, everyone can make their own settings, if necessary. The solution from Splunk also can provide the visualizations on remote mobile devices along with application in web and visualization components, which in turn can be relatively easily configured using XML format [16].

The ELK stack, as well as the Splunk, implements a system of various kinds of visualizations, including linear diagrams, tables, as well as their visual representation in the monitoring dashboard in web. All these graphic elements are based on the Kibana visualization tool of ELK stack. The search filter in most cases is placed depending on the situation in different ways. For example, if a query was used, the filter will be automatically applied to all elements of the management

panel. In Splunk this is organized in almost the same way, but there is still an optional configuration based on XML. At the same time, Kibana has no ability to provide user management, but if ELK fee-based support is involved, then this option is still implemented.

Search capabilities. For any log management platform, the ability of search processing is the most critical parameter for evaluation. In both systems, you can search in a separate search field. However, the query syntax in this case is significantly different. So, in Kibana search queries are built on the basis of Lucene query syntax, and in Splunk the same operation is performed in a separate programming language developed by the company, which is called the Splunk Search Processing Language. For those who already own scripting programming languages to master the syntax of Lucene will not be a big problem, but the situation with Splunk is not so optimistic. To compose searches there, you need to spend time learning this programming language.

Another difference is that Splunk software provides the capability for dynamic data exploration. It allows users to extract the results as a field for searching with formatting so that it is possible to carry out the search process and unconfigurable fields. At the same time, the elasticsearch query fields must be preconfigured and also additionally configured to be aggregated by the properties of the log file [14, 15].

The difference between the syntaxes of these two systems lies in the fact that the Splunk programming language (SPL) supports the so-called search pipeline. In it can be united by a single chain all the commands through a special connecting sign. This means that it is possible to organize a search in such a way that one execution of one command will be the input data for the next one. The syntax for Lucene queries is actually much simpler. It can output results based on the query without additional conversion.

Support for the community. Both systems, like Splunk and ELK, have a large number of supporters and active users, which are integrated into communities and forums. That the work with the product is very manageable, ELK also has its own, carefully written and extensive documentation. However, Elastic itself offers free training courses to work with the system around the world. However, Splunk also has well-developed documentation, together with strong community support, but at the same time it still offers professional services to support users. Training materials from Splunk are also easily accessible, and technical support is relevant at an unlimited time.

Curve of learning. The learning curve for ELK products looks flat, based on the company's policy. Elastic offers paid courses, but they are already highly specialized, but due to the high popularity of the platform and the open source code on the Internet, there are enough free training materials. For Splunk, the learning curve is quite moderate, especially if there is a need for specialized information. Although the company offers a trial period, which is only a month, at the same time, paid courses have a relatively high cost.

User management. In this case, ELK Stack provides protection as a separate, paid tool, which is based on different roles of users. Splunk and ELK with paid support provide user management right after purchasing a paid version of the product, along with the already-audited users [16].

Pricing. As already mentioned above, Splunk is a paid software with a clearly marked price. At the same time, if several sources of incoming data are integrated into the platform, the cost will increase in direct proportion to the size of the data stream, namely the traffic that is processed by the system.

ELK stack, which is an open source software, is provided for use on a free basis, but in this respect, not everything is so unambiguous. The final cost of using the products also includes the price of the equipment necessary for the proper functioning of the system and the price for the maintenance of the platform itself. In order to reduce the final cost of using ELK, it is necessary to develop separate modules, plug-ins, functions and tools.

Vendor Lock-In. A high price tag for Splunk products however has substantial grounds. After all, after this software solution is purchased, the user provides a modified and complete product. Of course, binding a user to only one supplier is not the most optimal solution, but this provider provides all the necessary solutions from and to. Although ELK is an open source product, it can not provide all the needs and capabilities that are necessary for the comfortable work of users, right after it is installed. Many functional requirements require further development and additional financial investments, for example, the simplest technical task of notifying users of an emerging event is not provided initially [14, 15].

One of the views on how one can consider which of the two software solutions is preferable is to proceed from which company provides the operating system solution. That is, whether Microsoft or Linux is used. So, when using Microsoft, Splunk is the preferred solution for the information-analytical system. However, if the choice is given to Linux, the optimal choice is the ELK stack. While making this choice, it is necessary to rely on the tasks that the enterprise wants to solve. Since usually the telecommunications providers of the republican scale the bulk of servers and the very operation of the system is built on the Linux operating system, the choice becomes obvious.

Both solutions are constantly improving and improving the functionality, and therefore the gap between the product with open source ELK and the proprietary solution from Splunk is constantly decreasing.

If we consider the product provided by companies separately, then we need to rely on the services we offer. So, if a mature product is provided, it will be most optimal to use Splunk, since the development budget will allow it to be done and the solution itself is more elaborate. However, if a dynamic change in the services is planned, ELK will preferably be used [16].

A set of possibilities. At the same time, the opportunities that are provided by both these solutions allow solving almost all the tasks that are put in the course of the work of large companies. Both are dynamically configurable and easily configurable. At the same time, almost all the necessary functionality is provided

for this category, namely: visualization of the incoming data stream, initial analysis, customized query-based search capabilities, and reporting automation [14, 15, 16].

Ease of use. What Splunk is that ELK is pretty easy to install and start to use, especially considering the set of tools that is provided after the initial configuration of the system. However, it immediately becomes clear that the Splunk solution is intuitively clear right after the system starts, which at the same time can not be said about ELK. In this case, the functions for managing users are quite complicated, in comparison with Splunk. At the same time, in order to facilitate the setting and lower the entry threshold for use in ELK, there is a separate software offer - AWS.

And the last parameter on which in the given degree work the comparison of these two systems is carried out is the API and the possibilities for expansion. In this regard, Splunk provides a carefully documented API for 200 endpoints for access. At the same time, ELK developed a mechanism for searching and analyzing information based on distributed systems with support for JSON and RESTful API. Also, additional integration of various user solutions is possible, which are developed in such programming languages as Python, Java, .NET and some others.

## **2.8 Data sources for analytical system**

2.8.1 Enterprise service bus. To organize the interaction between the sources of incoming data and the analytical system, a separate software solution is employed, which is an Enterprise Service Bus. This product is interconnected with data sources and their processing at all stages of information processing. That is, its main task is to simplify as much as possible the request of individual services, by the method of interconnection of the end user and the provider of the required service. In this case, the main functions of this system is the conversion of messages, since regardless of the type of incoming or requested information, it will in any case pass through the Service bus. The second function is the establishment of the path through the entire information-analytical system - routing. At the same time, these two functions allow to solve a more global problem within the framework of constructing an analytical system, namely, to create interaction between loosely coupled sections of the system. All this is implemented in the SOA architecture. At the same time, analysts consider the ESB to be a combination of all the functions of already existing solutions for establishing an intermediate interaction [17].

All the implementation of various web-service is due to the formation of a simple protocol to access objects written in the language of the description of Web services.

To enable the integration of existing and developed monitoring subsystems through unified interfaces, it is proposed to introduce an enterprise service bus into an already existing infrastructure.

Enterprise Service Bus (ESB) plays the role of middleware software that provides a centralized and unified event-driven message exchange between different information systems on the principles of service-oriented architecture.

The basic principle of the service bus - concentration messaging between different systems through a single point, wherein, if necessary, provided transactional control, data conversion, safety messages. All settings processing and reporting is also assumed to be concentrated at a single point, and shaped in terms of service, thus the replacement of any information system that is connected to the bus, no need to migrate to other systems.

Advantages of implementing ESB:

Abstraction and unification of interfaces of solutions

Offset of activities in the field of integration of corporate systems in the direction of configuration.

Excluding the central component / broker from the architecture - the transition to a distributed architecture.

Simplified integration and disconnection of loosely coupled systems.

Simplify the procedure for developing and upgrading systems, reducing system downtime to zero.

Potential problems of ESB implementation:

Reducing the speed of interaction of systems, in particular, for those that are already integrated.

With basic implementation, an additional single point of failure.

Complicating the configuration and maintenance procedures [17].

2.8.2 CMDBuild. At the moment, some telecommunication operators already have the initial developments regarding the construction of systems for recording and processing events. In particular, they apply CMDBuild. With it, the telecommunications operator solves the tasks of configuring databases using an open source product. The system was developed with the aim of realizing the idea of creating problem tickets for each event that occurred and obtaining initial statistical data on the state of the network [18].

In this case, CMDBuild has the necessary functionality to manage some elements of the hardware. This includes both conventional computers and peripheral devices, as well as software, various reporting, services, and so on.

Virtually on the basis of CMDBuild implemented a number of tasks that the telecommunications operator is facing in the course of its activities. So it provides an opportunity to create a user interface. There is an additional module for compiling and processing reports. Although it requires quite a lot of labor for configuration, the functionality is still ensured at a sufficient level. Similarly with other information-analytical systems, CMDBuild has the ability to create information panels. However, in the telecommunications provider, which is considered in this diploma project, this system is used mainly to create an integrated document management system. Basically all the functionality of CMDBuild in this case is limited to creating problem tickets, monitoring their implementation and creating static reports based on events that occurred in Zabbix. However, the system has the ability to interact with external systems through a web service. This consists in displaying the incoming data, as well as the already processed amount of



information on the maps and mnemonic diagrams. But as it was said before, this functional is not applied in practice.

The main reasons for the narrowing of the CMDBuild functional in real use are the configuration difficulties. Since the capabilities of this system are wide enough, and all of them are reflected in the initial stage of work, this creates additional difficulties for the end user, who may not have sufficient qualifications. This creates a paradox. The system, which was originally aimed at speeding up the processes of user interaction with data and making the reporting phase more convenient, in fact slowed it down. This is due to the incomplete understanding of the functioning of the system and the tasks that faced the telecommunications operator [18].

At the moment, a large amount of data passes through CMDBuild. Basically this happens at the stage of reporting, so the system being developed should have the ability to interact with this software solution. For this, the possibility of establishing an interrelation with the corporate bus of the enterprise was used. In Python, a separate module is written for bus communication (Appendix B).

2.8.2.1 CMDBuild API. Since to create a module that will interact between the corporate bus and directly CMDBuild you need knowledge of the API, then consideration of its basic concepts is necessary.

CMDBuild allows to create an interface that makes it possible to single out individual objects and see their relationship. It is possible to perform additional input of information, both reference and dynamically updated for all realized objects. As an example, you can consider the interaction of two objects: the server and IP addresses. So you can set the relationship between these objects as 1 (server) to a set of K (IP addresses). In this case, the next step will be the transfer of information regarding the IP addresses of this server [19].

With all this, when this stage is implemented it will provide an opportunity to determine the location of the server on the network, and thus to receive the necessary background information at any time, without resorting to the use of manual method of interaction with the system.

In order to understand how this will work, for example, with the monitoring system that is present at any telecommunications provider, the following example can be considered. So for the monitoring system built on Zabbix it will be possible to instruct in the automatic mode to apply a preconfigured template with the necessary list of parameters for monitoring.

As the next stage of the system, the process of obtaining information from CMDBuild is considered. In order to solve these tasks, the previously mentioned SOAP and also the REST interface are applied.

These methods are the most preferable, and this is due to the probability of updating the database model itself. After this kind of update, getting information through SQL queries will be problematic.

Although CMDBuild also has enough extensive documentation for the functioning of the application, the problem lies elsewhere. There is no precise

description of the possibilities, and all training on the use of the system is based on consideration of individual examples that arise during the functioning.

The procedure for using REST CMDBuild is as follows:

Initially, it is need to authenticate. Next is the process of getting tokens, to identify the current session, during subsequent REST calls. In this case, the authentication token itself is of relatively great importance [19].

An example of the use of this tool is available in Appendix B of this project.

**2.8.3 Zabbix monitoring system.** At present, telecommunications providers in Kazakhstan are in the process of introducing a large part of radio-television stations (RTS) into the monitoring system. This is due to the transition of Kazakhstan to the format of digital television. At present, more than 200 RTS are integrated into the monitoring system. The total number of digital stations that will be built and implemented in monitoring is 827. The monitoring system is a critical element in the functioning and monitoring of the network. With the advent of digital television, the possibilities for monitoring and recording events have expanded. Simultaneously with this transition, continuous broadcasting is carried out by 1218 analog RTS. Their monitoring is carried out manually, and in the future it is planned to completely abandon this type of control [1].

The entire monitoring system of the telecommunications provider in question is built on the Zabbix system. Its main advantages are flexible possibilities for expanding the scale of event registration. The basic principles of functioning is the interrogation of devices located on the RTS via MIB. After that, the obtained values are compared with the values that are preset as threshold values. If this limit is exceeded, a message is generated in the Zabbix monitoring system.

Zabbix provides opportunities for building a distributed monitoring system. There is another significant advantage of the system, which is that the system is an open source solution. This means that in the future, to integrate Zabbix with the corporate bus of the enterprise and with the information-analytical system, it is possible to write separate modules for the interaction between the elements [7].

Zabbix is used to monitor a wide range of network parameters. This concept includes separate devices, power supply systems, cooling and device integrity. At the same time, there are flexible possibilities for creating notifications depending on the events occurring. This makes it possible to minimize the response time to accidents occurring and to timely resolve problems related to the cessation of broadcasting. It is the implementation of uninterrupted broadcasting and the provision of services is the most important task facing the telecommunications provider.

There is support for both pollers and trappers. At the same time, the formation of statistics and reporting is included in the functional of the system. However, all these operations require additional work on configuration and configuration.

It is also possible to display the arising events on the mnemonic diagrams. For example, if a certain event occurs in some remote object, it will be quickly fixed

and displayed on the map. The monitoring operator will be able to respond to the problem that has occurred according to the job description and to eliminate the accident as soon as possible.

The Zabbix system is already an implemented solution in the network monitoring process and has a fairly large range of tasks to be solved:

Data collection:

- Continuous monitoring of parameters of equipment availability and performance
- Opportunities for wide use of various protocols, including SNMP, IPMI, JMX and others.
- Ability to perform custom checks.
- Since the threshold values for the occurrence of an alarm event for each device are different, the need to establish user intervals also takes place.

Flexible options for setting threshold values:

- Alarm thresholds are used in the monitoring system, which are called triggers. However, in the course of their work they refer to the values that are in the database.

Fixing of emergency events is important during the functioning of the telecommunications provider's systems, so the ability to configure notifications of responsible persons, interact with their different types, recipients is an important task.

However, notifications should be informative and have as much information as possible about the event, so that the operator can transfer a problematic application to technicians in the immediate vicinity of the radio and television station.

In the future, it is planned to expand the decision-making capabilities based on the data coming from the Zabbix system by implementing the elements of the expert system [7].

Zabbix allows you to create a high-quality storage of information coming in during monitoring. Also, in the process of fixing information, there may be a need for clearing the history, if an unsatisfactory threshold value is established on some of the triggers and the number of operations has reached a large scale.

Zabbix allows you to create and use templates for repetitive devices on the network, which increases the system's ability to scale with the least amount of effort.

**2.8.3.1 Zabbix API.** In order to most effectively and flexibly work with Zabbix all possible methods of API application are used. Zabbix API allows for flexible interaction and configuration of the system itself. Also with it you can access individual data items for a certain period. The API in this case performs a number of functions necessary to create new modules and applications that interact with Zabbix.

In this case Zabbix API allows creating new software solutions and modules for interaction with other systems. This facilitates the process of integration

between Zabbix and the corporate bus of the enterprise, and hence with the information and analytical system [20].

Since a lot of attention in this infrastructure is paid to the process automation process, which are routine for the telecommunications provider, the availability of the Zabbix API also positively affects this factor.

Zabbix API is presented in the form of a web interface and functions on the basis of the web. The structure is a whole series of methods that are applied depending on the problem being solved. All requests and responses between the client and API used use JSON encoding.

If you look more closely at the structure of the Zabbix API, you can say that it is a series of methods combined into separate API groups. In this case, each method is designed to perform only one specific task. An example is the `host.create` method. Based on the name, we can conclude that it is used to create new nodes in the monitoring network. In this case, these groups formed the name "class".

Once the web interface is installed, it is possible to interact with the API using an HTTP request. In this case, all requests are sent to a certain file that is in the same folder as the web interface.

The order of execution of requests looks like a certain action plan. Immediately after the installation of the web interface is done, it is possible to use HTTP requests in order to use the API functionality. This is done by using HTTP POST. The file itself is called `api_jsonrpc.php` [20].

In the process of infrastructure development for the information-analytical system, the Zabbix API is used at the stage of developing a separate module for the interaction of the monitoring system of the telecommunications operator with the corporate bus of the enterprise. The efficiency of such a module allows creating an additional and important data flow concerning the process of functioning of the entire network. In this case, the data comes in a structured form and it is possible to further segment them according to the source of the initial data and the information contained in the messages.

**2.8.4 Requirements for the event registration and management system.** During consideration of the existing monitoring system, it becomes clear that it solves a wide range of tasks. At the same time, it can contain a whole list of additional modules, such as performance management, monitoring of the level of services provided, direct technological monitoring of the network, provision of a storage system, and monitoring of the state of hardware and software [21].

Sending of messages about the happened events and the control over their decision can also be the same modules. The peculiarities of these processes can be considered by the example of two structures that could not function without an IT infrastructure. They are banks and a telecommunications provider.

For example, the IT service of the bank provides its users with a full range of services, from access to the Internet, postal services, to specialized financial instruments. At the same time, the number of these narrowly directed banking systems, interconnected, is relatively small. From this it follows that the first place

is working with service-resource models, which should reflect the current state of controlled systems. In doing so, they help identify non-functioning elements, as well as their level of impact on the provision of services to end users. In this case, in the event of a malfunction, the degree of responsibility of specific employees for the elimination of malfunctions must be prescribed in advance.

Telecommunications providers have only one service for numerous customers. However, the fact is that the number of subscribers is incommensurably higher. Therefore, it becomes obvious that the performance of the recording and event processing systems is critical, namely: the number of events that the system can process per unit of time.

At the same time, the system can be multifunctional, have ample opportunities for additional settings, but the most important is the opportunity to use it without having specialized skills. To do this, you need to determine who will use it and with what methods. Regardless of the end-user (monitoring operator, technician on the RTS or administrator), it is necessary to ensure that the messages about the events will be communicated to the responsible persons and be resolved in a timely manner [21].

At the moment, there is a need in the telecommunications operator under consideration to introduce an additional solution as a part of the general information and analytical system in the form of a system for recording and managing events. ERMS is an easy-to-use interface for the end user. It does not require high qualification for use, but it requires considerable effort in developing the system.

In doing so, you can formulate a set of recommendations that need to be followed during the concept development process, as well as a description of the functional of the future system.

The first thing to begin with is the definition of the type of events being processed. As already highlighted, in this case there are two main data sources at the moment. Namely: data coming from Zabbix during monitoring and data coming from CMDBuild.

Allocation of the main types of data in the future will give an understanding of what may be the failures in the operation of the infrastructure, and also determine their possible impact on the functioning and speed of the internal business processes of the telecommunications provider [9].

At the moment, CMDBuild receives a lot of unnecessary information in the form of statistical data. This is connected with the processes of implementation of the monitoring system, and as a result, in the initial stages, a large number of triggers are possible. At the same time, only the most important events should fall into our statistical sample, which must necessarily be processed and accepted for execution. The importance of this is due to the impact on the final broadcasting of the telecommunications provider and also on the quality of the services provided.

In order to identify really important events that have a significant impact, it is necessary to conduct an initial survey of the monitoring engineers of the telecommunications provider. They can certainly determine which problems are typical and how they can be built on a solution algorithm. Since the possible

number of emergent accidents is colossal, it will be irrational to waste time compiling the entire list of possible situations. It is possible to obtain statistics on typical failures and how they affect services. At the same time, the entire procedure will help, using the most accurate formulations, to minimize the time of detection and elimination of negative events.

Also an important task is to define methods and means of collecting information during monitoring. You need to determine the capabilities of these devices and the ability to scale. It is possible, such a situation that individual devices can fix only a limited range of emerging events. The telecommunications provider in this case has a list of parameters necessary for monitoring and reporting. These documents describe the parameters and their threshold values, which must necessarily be present in the monitoring system, and as a consequence, be processed by the information and analytical system. In this case, the devices that carry out the monitoring process must have the opportunity for modernization in connection with the constantly growing flow of incoming data.

An important task is also the definition of the resources by which the process of transferring events to a single event processing and management center will be implemented. In this example, such centers can be, as well as the regional directorate of television and radio broadcasting, and larger communication control centers.

It is also necessary to analyze already existing configuration management processes. After all, some systems, such as CMDBuild, have the ability to solve similar problems in their functionality. However, as preliminary research showed, the available resources are not involved in this task and most of the information processing is done manually.

In order to set priorities for services, it is necessary to identify those devices that have the highest influence on the broadcasting process and the provision of other services.

When developing this system for registering events, it is necessary to conduct constant testing of individual elements of the model. It is also important to check the logic and simulate the occurrence of certain situations and monitor their solution. After all, in some cases it is enough to react once to the incident on time and prevent negative consequences in order to recoup the development and implementation of the entire system as a whole. Given the state importance of the services provided, this becomes even more urgent [21].

When designing a system, there is also the question of whether to use a ready-made solution or to develop the system yourself. After analyzing, it was found that there are four alternative solutions to this problem. The first option is to develop your own web application. Moreover, it has a number of advantages, including the ability to install on an existing infrastructure, the lack of a constant subscription fee, a simple end-user interface, the presence of local technical support. The second option is the SaaS service. It has the necessary functionality to solve problems, but it does not have local technical support and its placement on its own infrastructure is not considered possible. The third option is CMDB solutions



already considered earlier. Its main disadvantage is the complexity of the configuration and the use of the end user. And the last possible method is a corporate solution from major software developers like Microsoft and IBM. The main stopping factor is the price of this decision and the lack of local technical support.

According to the comparison, the best option would be to develop your own ERMS web application, since this is the only software product that can later be transferred to the service of the engineer of the telecommunications provider itself.

After carrying out the procedure for selecting a system for recording and managing events, it is necessary to designate the logic of work, while optimizing for the already existing monitoring system. Only with this condition the developer will have complete certainty in the tasks and reasons for doing this work. If you do the opposite, it's possible that in the project documentation the system will work, and in fact there will be a mismatch with real problems. Such situation already arose with the telecommunications operator during the implementation of the system on CMDBuild, and this experience should be taken into account.

Successful implementation also depends on the continued operation of this system. To do this, it is necessary to work on the regulation and training of employees.

If we consider the implementation of regulations, it should be noted that the process of managing events should be closely related to the already existing processes of change management. Any changes you make must be reflected in the monitoring settings and the existing additional services. This means that otherwise there will be a need to develop separate regulations for each change in the composition and settings of the entire network.

## **2.9 General description of the designed infrastructure**

So, after all the necessary tools and technologies required for building an information and analytical system were considered, it is possible to compile a structural block diagram of all the elements.

There will be two main sections: data sources and the analytical system itself. Also in the scheme there is an element of notification of users about the happened accidents, but it does not belong to the block of the analytical system. There is also the opportunity to integrate the expert system into the work of the analytical system in the future, however, in order to create the initial knowledge base, it is necessary to spend considerable time in consultations with the employees of the telecommunications provider. After all, they know the order of troubleshooting the network, and consequently, they will be able to make scenarios that will be shown to monitoring operators in the future. The diagram also shows the RTS, where the equipment sends data to the Zabbix monitoring system.

Also, as one of the further steps for the development of the data collection system (sources), it is possible to develop a system for registering and managing events based on the ERMS software solution. This will create a unified reporting,

which will pass the analysis procedure in an already functioning information and analytical system.

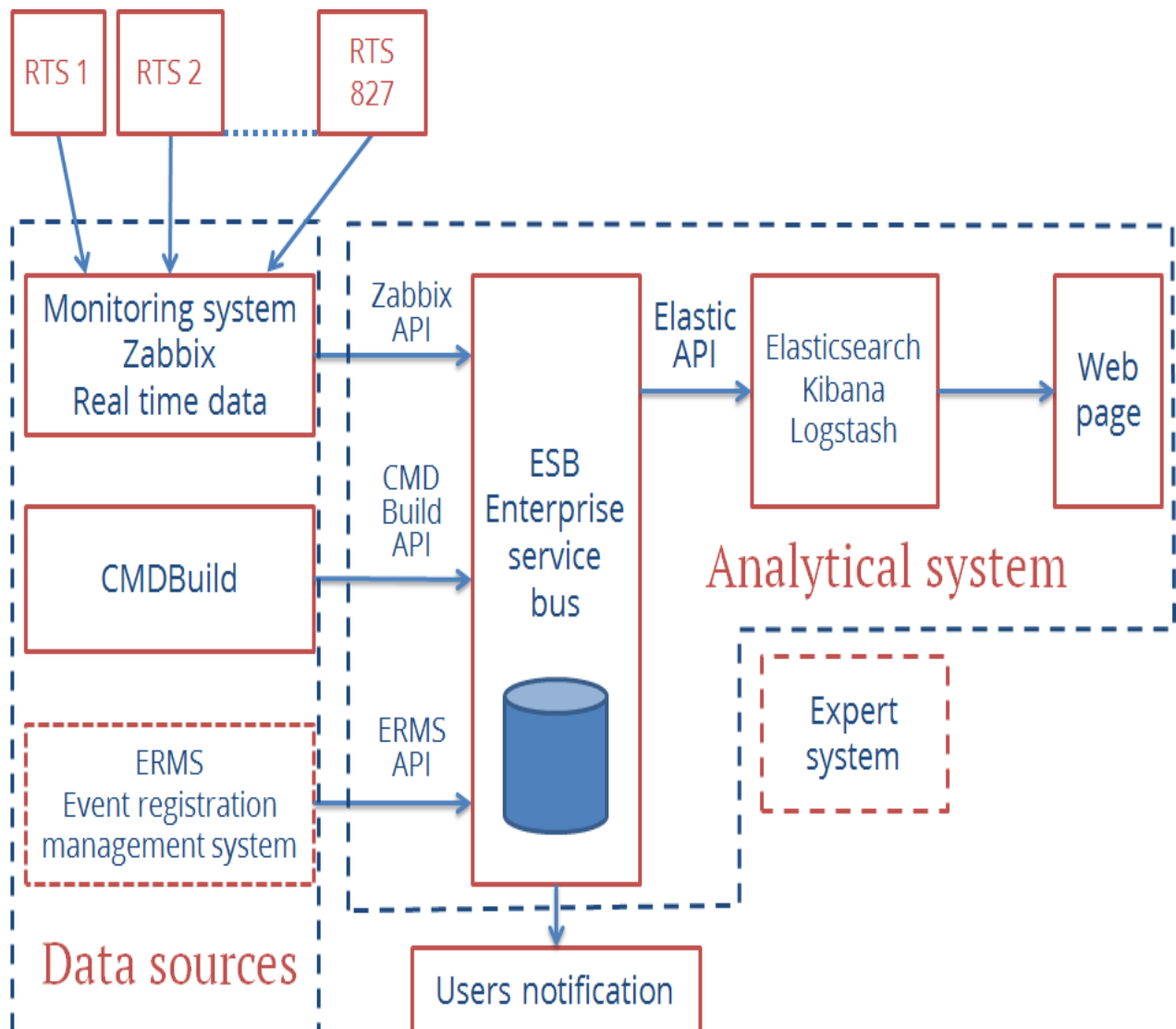


Figure 2.11 – Infrastructure of the designed system

As mentioned earlier, in this graduation paper two modules have been developed for interaction with the enterprise service bus. These modules allow to interact with the monitoring system, as well as the existing system for registering events, built on the basis of CMDBuild. This module has the functionality to compile statistical reporting on events occurring in the network.

The initial setup of the ELK stack was also performed to more clearly display information coming from data sources. There is also a web page for the supervisor of the system with a personal cabinet. There is an opportunity for the cabinet to interact with the ELK stack.

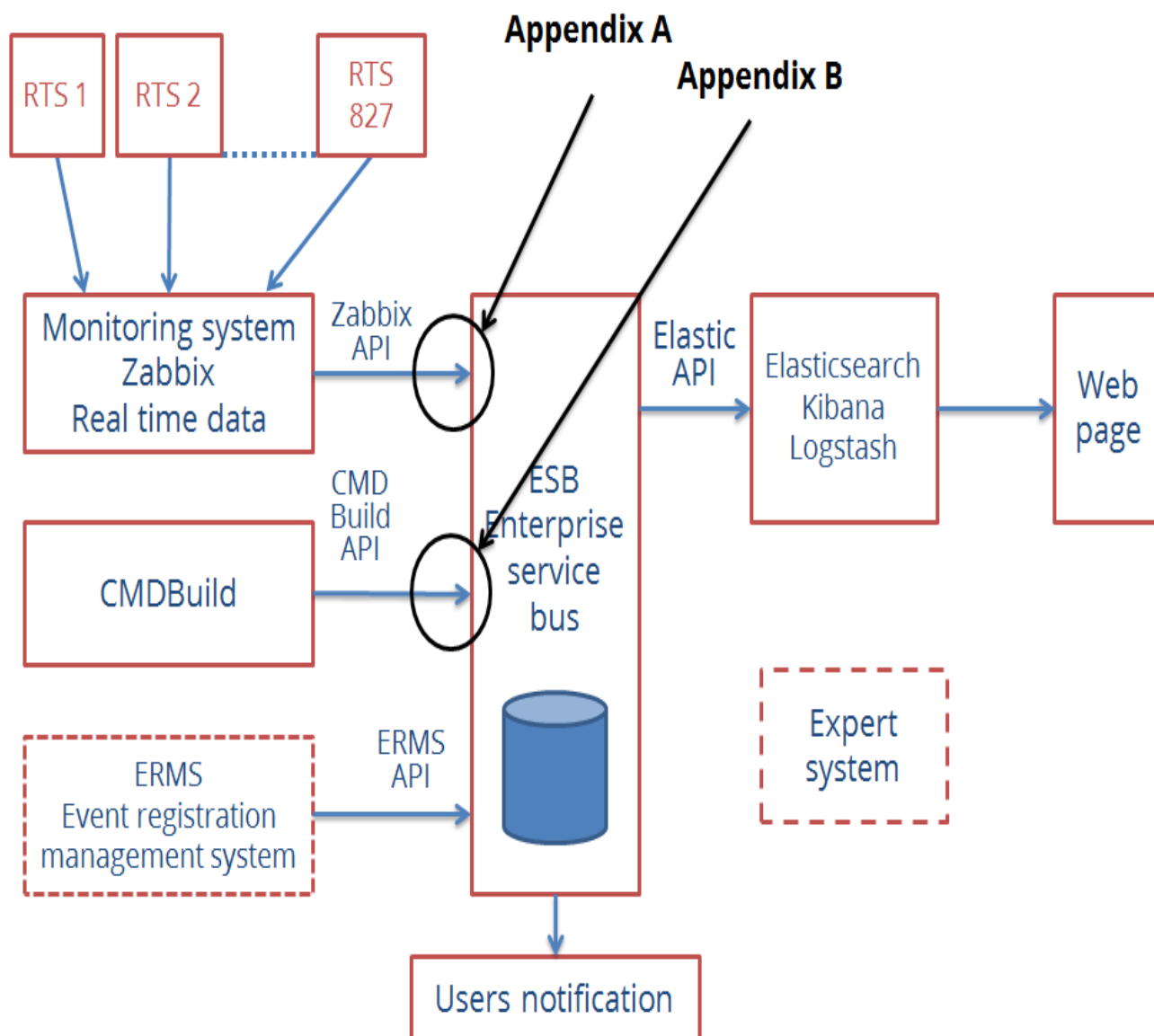


Figure 2.12 – Matching the parts of paper with the infrastructure

The operation of the system based on Elasticsearch is shown below. The mnemoscheme is a description of the process of initial data collection, processing and loading into the Elasticsearch database.

In this case, the data sources are several resources and, consequently, a common information collector is needed, where it will be accumulated and modified to a unified form. All this will allow the processing of the information-analytical system. In the role of collector in the case under consideration is the enterprise bus database.

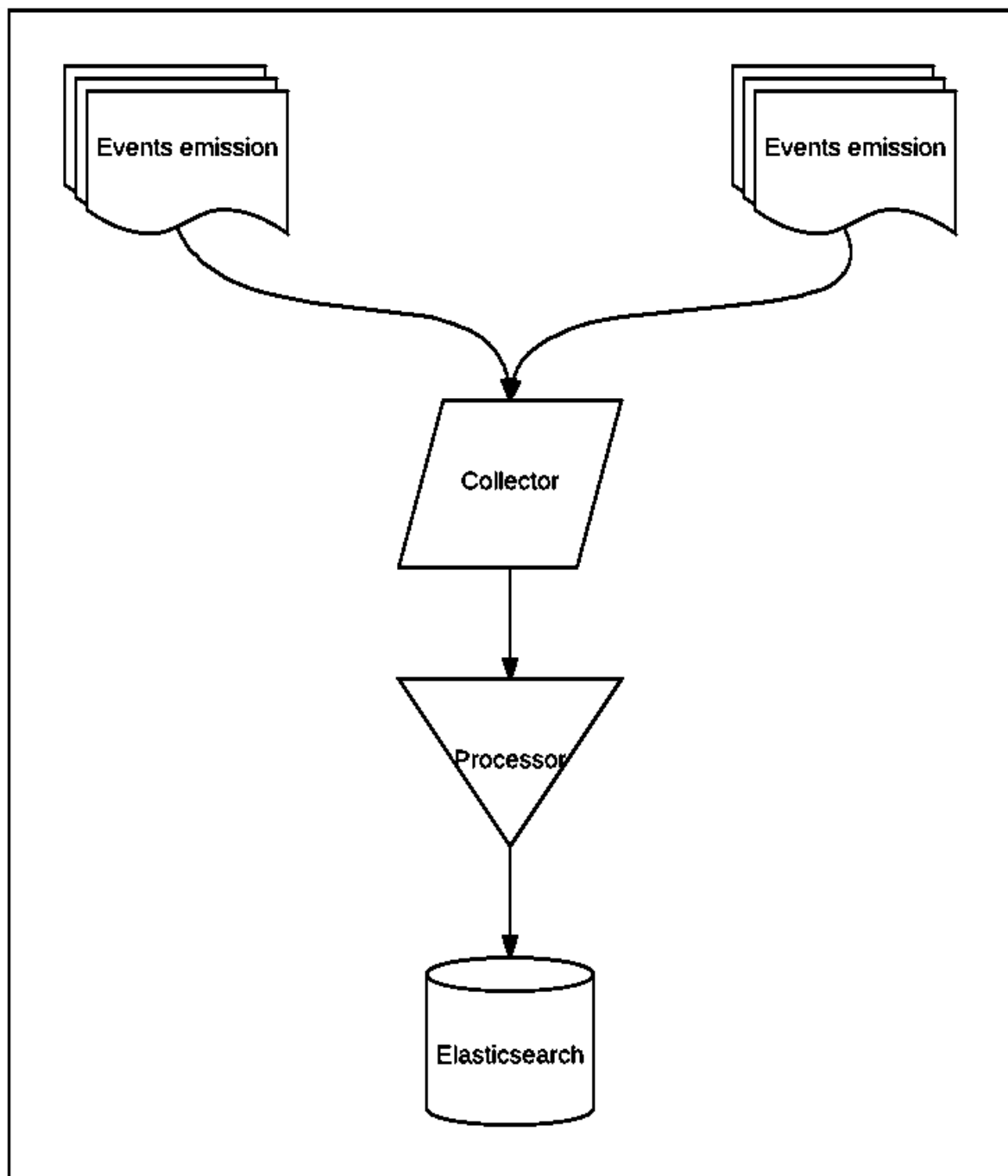


Figure 2.13 – Process of work of Elasticsearch

Since the process of integrating the system with a telecommunications provider takes quite a long time, the decision to organize data collection from students and employees of the NJSC "Almaty university of power engineering and telecommunications" was taken for the test of functioning of all elements. Data with reviews passed the indexing process in the information and analysis system and became available for search and segmentation. This allowed to bring the original statistics into the cabinet of the supervisor of the system and make this setting interactive. However, as part of the implementation based on the telecommunications provider, this will be only part of the analyzed information, since most of the data will come from the equipment.

All these actions will allow integration of the information and analytical system into the existing infrastructure of the telecommunications provider in the shortest possible time. This is due to the fact that the number of connected stations in the network of the provider is growing at a significant pace. So at the time the infrastructure was created for the analytical system, the number of RTS was 350. By mid-2018, the number of functioning stations, and therefore connected to the monitoring system, would be 450 RTS [1].

After all the initial software development activities have been carried out, it is necessary to prepare the hardware infrastructure for the analytical system, which will be discussed in the next chapter.

### **3 Calculation part**

When calculating the hardware need for a successful operation of the informational analytical system, it is necessary to take into account a number of factors. Initially, it is necessary to estimate the number of data sources, as well as the amount of information that they transmit. Once the amount of incoming data is determined, it is necessary to select the necessary server part that will be able to support the processing of all incoming data.

#### **3.1 Assessment of the required capacity of the directions for the transmission of state control data from the RTS to the regional and Republican Center**

There several steps which we need to cover while calculation the volume of data, which continuously coming to the analytical system of telecommunication provider. The main source is the data which comes from the equipment according to the monitoring MIBs.

Table 3.1 of section 3 of this project provides an approximate estimate of the volume of MIB - all significant data of the status of a typical RTS (without taking into account headings, addresses and other data for "packaging" information data in packages). The total number of controlled (logical and digital) parameters of the RTS is about 496. The total volume of this data for one RTS is about 1058 bytes [1].

The SNMP protocol assumes the use of the IP / UDP protocol at the transport level (that is, the transmitted data at the transport level is not acknowledged). Reliability and reliability of the transmission of RTS state data is determined by the probability of providing the channel and the reliability of the transmission in the VSAT satellite network. The minimum packet length is approximately 50 bytes, the maximum is 1500 bytes. Equipment controllers of various types can provide MIB data transfer:

- in the form of a table containing all the essential parameters (in this case, one request is made for the transmission of all parameters and the packing of data into transport packages is optimized). The average MIB of one hardware unit is about 110 bytes. The number of requests and corresponding IP packets to transfer

the total amount of MIB data to a typical RTS will be equal to the number of blocks in with SNMP agents in the RTS;

- consistently in time (each request requires a separate request, the data of each parameter is "packed" into a packet of a minimum length of 50 bytes).

Table 3.1 - Approximate values of the MIB volume - all significant data of the status of a typical RTS [1]

Name of the equipment	Number of sufficient parameters	Average volume of data about parameters, bytes	Number of block on RTS	Overall data volume about state, bytes
Satellite redundancy reservation switch $2 \times 82 \times 8$	20	40	1	40
Satellite Receiver	11	22	8	176
Remultiplexer	112	224	2	448
Modulator DVB-T2 with the equipment of the system of single time	28	56	2	112
Amplifying Broadcasting Route	3	6	2	12
AFD of broadcasting	2	4	1	4
Control Receiver DVB-T2 (STC-1000)	15	30	1	30
VSAT terminal	15	30	1	30
RTS Management and control block	8	16	1	16
Power supply and life support equipment	16	32	1 complex.	32
Sensors of security and other alarm systems	20	40	1 complex.	40

Table 3.1 continuation

Name of the equipment	Number of sufficient parameters	Average volume of data about parameters, bytes	Number of block on RTS	Overall data volume about state, bytes
Traffic Flow Analyzer DVMS	15	30	1	30
The device of visual monitoring Telescreen	11	12	4	88
TOTAL				1058

In the RTS MCB can be performed preliminary processing, generalization and storage of MIB data (general table of the RTS MIB is formed from the MIB tables of the equipment blocks,).

When "packing" the data on the status of the RTS into IP packets, taking into account the headers, addresses of equipment blocks, timestamps and other service data, we assume that the length of the MIB data packet is 1.3-2 times the total volume of the data to be transmitted [1].

Table 3.2 provides estimates of the volume of MIB packet data for a typical RTS, depending on the processing options and the generalization of these data to the RTS.

Table 3.2 - Estimates of MIB packet data volume of a typical RTS, depending on processing options

Variant of formation and processing of data MIB	Number of IP packets	Average packet length, bytes	Volume of packet data, bytes
Sequential query	496	50	24800
Table query for each block	11	200	2200
Request for the RTS table as a	1	1500	1500

When SNMP protocol is used, in addition to MIB data (all significant data of the state of the equipment blocks of a typical RTS), the signals from the generalized state "Trap" are generated by the equipment units. Different variants of signal formation "Trap" can be used for different equipment blocks [13].

These signals may contain [2]:



- only data on the absence or presence of an accident of the unit without indicating the cause of the accident;
- data on the accident indicating the signs and causes of the accident;
- data on the output of the MIB parameters for permissible threshold values, etc.

In accordance with the data of section 3, we assume that the length of the IP packet with the signal "Trap" for each equipment block is 100 bytes and every 4 seconds each equipment unit generates its "Trap" signal.

If the Trace is not pre-processed on the RTS, then the number of Trap signals, equal to the number of blocks with SNMP agents, must be transmitted every 4 seconds. There are 14 such units in the standard RTS:

- switchboard for reservation of satellite reception paths (1 unit);
- Receiver-remultiplexer PVR-7100 (2 units);
- modulator DVB-T2 (2 units);
- controller STC-1000 (1 unit);
- the controller of the power supply system (1 unit);
- controller of the life support system (1 unit);
- VSAT terminal (1 unit);
- visual monitoring Telescreen (4 units);
- DVMS transport stream analyzer (1 unit)

Thus, the total volume to be sent to the "Trap" data packets of the generalized state directly from the equipment blocks is 1100 bytes every 4 seconds.

If the RTS pre-processes the equipment's MIB signals and generates one common "Trap" signal of the generalized RTS state, then every 4 seconds it is necessary to transmit one "Trap" signal with an average volume of 300 bytes.

Let's consider possible variants of data transmission MIB and "Trap" from RTS to NCC. These states can be transferred on the initiative of the RTS at random times, or at the initiative of the NCC in a rigid time cycle, with the timely allocation of each RTA of the relevant time windows. We will evaluate each of the options for the network of data collection on the status of "Trap".

3.3.1 Transmission of signals "Trap" from RTS at random times. We assume that all blocks of equipment of all RTS networks without mutual synchronization with each other form "Trap" signals every 4 seconds. Signaling is performed using multiple access to one or more system channels.

The total transmission speed in the system channel without taking protective intervals (less than 5% of its bandwidth) for optimized QPSK mode (with 9/10 encoding) is about  $c_{chan} = 450$  kbit/s [1].

Required bit rate of a single signal "Trap":

$$c_{trap} = \frac{R * 8}{t} \quad (3.8)$$

where  $R = 100$  bytes – the length of the IP packet with the Trap signal for each equipment block;

$t = 4$  sec. – time cycle.

$$c_{trap} = (100 * 8)/4 = 200 \text{ bit/s.}$$

The average service intensity for a system of 827 RTSs of 14 blocks in each is calculated by the formula:

$$E_{RTS} = \frac{n_{RTS} * n_{message} * c_{trap}}{c_{chan} * 1000} \quad (3.9)$$

where  $n_{message} = 14$  – summarized number of messages coming from one RTS.

$$E_{RTS} = \frac{827 * 14 * 200}{450 * 1000} = 5.14 \text{ Erl}$$

When generating the "Trap" signal of the generalized RTS state, the required transmission rate of one "Trap" signal is calculated using the formula (3.8):

$$c_{trap} = (300 * 8)/4 = 600 \text{ bit/s.}$$

The average service intensity for a system of 827 RTSs of 14 blocks each according to the formula (3.9):

$$E_{RTS} = \frac{827 * 14 * 600}{450 * 1000} = 15,43 \text{ Erl}$$

Table 3.6 gives estimates of the required bandwidth for Trap signaling, depending on the given probability of failure in providing a channel (the probability of bringing the signal to the CCC within 4 seconds) for a random access system without a queue for maintenance.

Table 3.3 - Required bandwidth for the transmission of "Trap" signals depending on the given probability of failure in providing a channel

Required service intensity (individual block signals), Erl	5,14			
Probability of service	0,9999	0,9990	0,995	0,992
Number of channels with transmission speed of 450 kbit/s	16	14	12	12

Table 3.3 continuation

Required total bit rate kbit / s	7200	6300	5400	5400
Required service intensity (generalized RTS signal), Erl	15,43			
Probability of service	0,999	0,995	0,99	0,95
Number of channels with a transfer rate of 450 kbit / s	3 2	30	26	26
Required total transmission rate, kbit / s	14400	13500	11700	11700

Thus, the total throughput of system channels in the network with random access in the transmission of signals "Trap" of the generalized state of the RTS for the probability of bringing 0.999 (the loss in each cycle of the signal from no more than one RTS of 827) is 14.4 Mbit / s.

3.1.2 Transmission of signals "Trap" from RTS in a rigid (deterministic) time cycle. When operating in a rigid (deterministic) time cycle with the allocation of each RTS to the respective time windows, the required throughput is defined as the product of the rate of transmission of one signal by the total number of blocks with SNMP agents in the broadcast system.

For the variant with the transmission of individual block signals, the total speed can be calculated by the formula:

$$c_{sum} = c_{trap} * n_{block} * n_{RTS} \quad (3.10)$$

$$c_{sum} = 200 * 14 * 827 = 2315 \text{ kbit/s}$$

It can be seen that with the use of only six system channels with a capacity of 450 kbit / s guaranteed (with a probability close to one), data on the generalized state of the equipment of all RTS networks during 4 seconds will be provided. At the same time, the decrease in the required bandwidth, compared to a network with random access will be significant.

For the variant with the transfer of generalized signals "Trap" RTS when working in a hard time cycle with the allocation of each RTC of the corresponding time windows, the required capacity is also calculated using formula (3.10):

$$c_{sum} = 500 * 1 * 827 = 414 \text{ kbit/s}$$

Thus, in the network of data collection of the generalized state of RTS "Trap" equipment, it is advisable to organize a rigid time cycle ensuring guaranteed bringing the signals of the generalized state of each RTS to the CCC within 4

seconds.

Unlike signals "Trap", for the formation of which no requests of NCC are required, for transfer to the Center of MIB signals from NCC to RTS should come relevant requests [2]:

- Get-request-MIB request by block name;
- Get Next-request-MIB query for sequential viewing of the hardware block table.

Set - transfer of control commands to the state change of the equipment block. The RTS equipment generates Get-response packets - responses to the Get-request, Get Next-request and Set commands. The MIB data request is required by the NCC to solve two classes of tasks:

- remote control of RTS equipment to eliminate the current failures that occurred on it (when receiving a corresponding signal from Trace from the RTS), or carrying out planned measures to change the operating modes of the equipment;
- Replenishment of NCC databases with the discreteness necessary to predict failures based on the correlation analysis of data on the status of the RTS network.

The bandwidth requirements for the transmission of MIB data for each of the task classes are formulated in different ways.

### **3.2 Definition of parameters for the server part of the infrastructure of the information analytical system**

For the functioning of the system, an additional installation of servers is required, which will collect data for subsequent analytics. In this case, the main features are: not the criticality of instantaneous recording and the large amount of stored data. In this case, the query forms are aggregate samples with a search of a huge number of available records. And this means that the use of RAM in an effective mode in this case will be difficult.

All this leads to the fact that when choosing a server part, you need to rely on large sizes of hard drives, average processor performance and average amount of memory.

Given the total size of incoming data, it is possible to calculate the maximum amount of disk space that will be used in three years of system operation. This period was chosen based on the standards of the telecommunications provider regarding the time of data storage.

$$V_{data} = c_{sum} * t \quad (3.11)$$

$$V_{data} = 414 * 1000 * 60 * 60 * 24 * 365 * 3 = 38 * 10^{12} \text{ bytes}$$

Based on the received value, it is possible to determine that the system requires 38 terabytes of disk space.

This means that it is necessary to use a server with 12 bays to install hard drives.

According to the volume of incoming data and the recommendations of the

software solution providers (Elasticsearch), two servers, each with 8 cores and a processing frequency of 2.4 GHz, are necessary for the system functioning.

In this case, the value of RAM for each of them will be a total of 16 GB.

These parameters will cover the needs of the server part when using the transmission of individual block signals in a strict time cycle, which means that the transmission of generalized Trap signals will also be possible.

3.2.1 Basic principles for the server part of the infrastructure. As the amount of data grows, you need to think about hardware. In order to deploy the system under the management of Elasticsearch, you need to follow the recommendations from both the software solution provider and the basic principles for organizing the system. For this, it is necessary to take into account a number of factors, such as: reservation, provision of sufficient computing power, and more. Recommendations in this case are not strict plans, and not always require a huge number of servers for the full operation of the system. But at the same time, these recommendations give starting points in the construction of infrastructure, which in most cases are described in the technical manual for Elasticsearch, based on the production experience of the developers [22].

At the moment it is possible to build a system based on any number and size of machines. The size and processing power in this case can truly amaze: hundreds of gigabytes of RAM, dozens of cores of the best processors. You can also make thousands of small servers with a fairly low processing power. It is necessary to determine which approach will be the most appropriate solution in the case of building an information and analytical system.

In short, it is necessary to give preference to medium and large machines. This is due to the fact that managing a system consisting of 1000 small servers is extremely difficult, and the cost of operation will be commensurate with the cost of running the system itself.

In this case, and huge machines can not be given preference. Huge resources do not guarantee an efficient and balanced work, but they certainly guarantee a high price for such a decision. In this case, there are great difficulties when migrating servers [22].

3.2.2 Definition of the RAM size for the server part When working with information and analytical systems, the first question that needs to be closed is the issue with sufficient memory. This is due to the fact that the sorting and aggregation process is very demanding for resources, which means that there should be enough space for the system to work. In this case, if the amount of data is relatively small, the entire available stock can be redirected to the cache of the operating system file system. As already discussed in the review of the search engine Lucen, the data structure is a disk format. That's why Elasticsearch actively interacts with the operating system's cache. In this case, a server with 64 GB of RAM will be an ideal option for building an infrastructure on it, but more common hardware solutions at 32 GB and 16 GB. Since the amount of incoming data is relatively small, the most suitable option is a solution with 16 GB. After all, when using a server with less

than 8 GB, the system starts to function counterproductively. This leads to the need for the operation of many small servers. However, too much memory can cause a number of problems [23].

3.2.3 Definition of the number and characteristics of CPU for the server part. The processing of complex, filtered, queries, a constant indexing process, percolation, and in this case also access to non-Latin encodings, since the system should work with Cyrillic - all this has a serious impact on the CPU. This is the reason that the choice of processor should be approached with special care. However, for this it is necessary to plunge deeply into the specification of available hardware solutions to understand how they will behave with Java [23].

For example, the Xeon E5 v4 processor provides almost 60 percent performance gain than the third version, when working with Java. Xeon D also allows for horizontal scaling of the system, immediately after the intensive indexing that occurs is distributed among the remaining nodes. However, the importance of the work of each node, especially if we take into account their small number, is difficult to overestimate. Thus, determining the right technical solution in terms of the choice of processor is also important for horizontal growth of the system.

If we consider how the Elasticsearch distributes the load to the CPU, it is possible to see that the entire load is distributed over several pools. Each pool has its own purpose. For example, this is a pool for standard operations, such as searching, indexing, for obtaining some operations, for performing mass operations.

It is possible to configure several threads in each pool, including configuring the type and length of the queue that occurs during processing. The number of threads, by default never exceeds 32, but if there are more available cores, the actual value will be displayed [23].

The default settings are quite stable and effectively load the system, so it is not recommended to change them. It is possible to adapt the queue size, because in case of filling the queue, superfluous requests and operations will be rejected.

Most operations associated with the operation of the Elasticsearch are fairly simple, so the requirements put to the processor are also reduced. Thus, the thin configuration of the processor settings is less important than other resources, such as memory. It is enough to choose a modern processor with several cores. Usually, processors with 4 and 8 cores are used. In our case, three servers with 8 Xeon E5 cores are used, since it is comparable to the newer versions for computational resources, but the possibilities for increasing the number of streams are wider. Also for the processing of http requests, processors of the same single-chip model with 6 cores are used, since this type of operations will be significantly less than search.

In the matter of choosing between more modern processors with a high frequency of work, it is better to choose those with more cores, so when the information and analytical system is working, the issue of parallelizing the load is more important.

There are several parameters and cases in which the appearance of a system violation is obvious:

1) The load on the processor is more than 90%. This is too high a value for one node, and forces you to apply load sharing technologies to other nodes.

2) The average value of the load, which can mean problems related to I / O and inconsistencies when switching context. In some cases, this may be due to errors in Java or the kernel itself.

3) The current status of the pools, namely the value of the rejected flows, as a result of the raw data. The increase in the size of the queue in the long term will have an extremely negative impact on the operation of the entire system.

4) The amount of data, records and search queries per second are also important quantitative indicators of the system performance.

3.2.4 Definition of the type of memory for the processed data storage. Disks for the operation of clusters of the information-analytical system are extremely important, especially for clusters with indexing. For a system where requests are processed, and the data is very important, since they are related to the functioning of the provider's telecommunications network, the storage system plays an important role. Since disk space is the slowest subsystem in the entire infrastructure, that's why clusters can very quickly fill all available space, and this leads to a bottleneck in the entire system [22].

SSD drives in terms of performance, recording speed is noticeably superior to the older technology used in HDD drives. The key factor, however, in this case is the price. Therefore, in order to save on servers where the data indexing process is applied, SSDs will be installed to provide the necessary speed, and in data warehouses where processing is almost not performed, but large-scale storage of data on incidents with a prescription of up to three years is required, HDD is used.

3.2.5 Definition of the interconnection between the servers. In this case, whatever the bandwidth, the more the better. Low latency ensures that the data exchange process will not be hampered by the physical parameters of the transmission medium. In most cases, 1 GB ethernet cable is used in information systems, and it is possible to use a cable with a bandwidth of up to 10 GB for improved infrastructure operation [23].

A fast and fault-tolerant network is absolutely essential for the operation of a multi-node system. However, it is necessary to avoid infrastructures that include several data processing centers. In this case, even the proximity of these centers can not correct the situation. It is also desirable not to use data centers that are far apart from each other. All this has a very significant effect on the system throughput.

Initially, the nodes on which Elasticsearch operates function a delay of no more than 150 ms, which can not be reliably ensured in the case of a large distance between data centers.

At the same time, most data centers claim that the communication channels between them are reliable and have a low delay, but during an emergency situation it will be a big problem to transfer a constant stream of data to other servers. In this case, the costs of eliminating the emergency situation are incomparably greater than the savings from using the data centers. In the case of a telecommunications



provider, data security is much more important than imaginary savings in the short term.

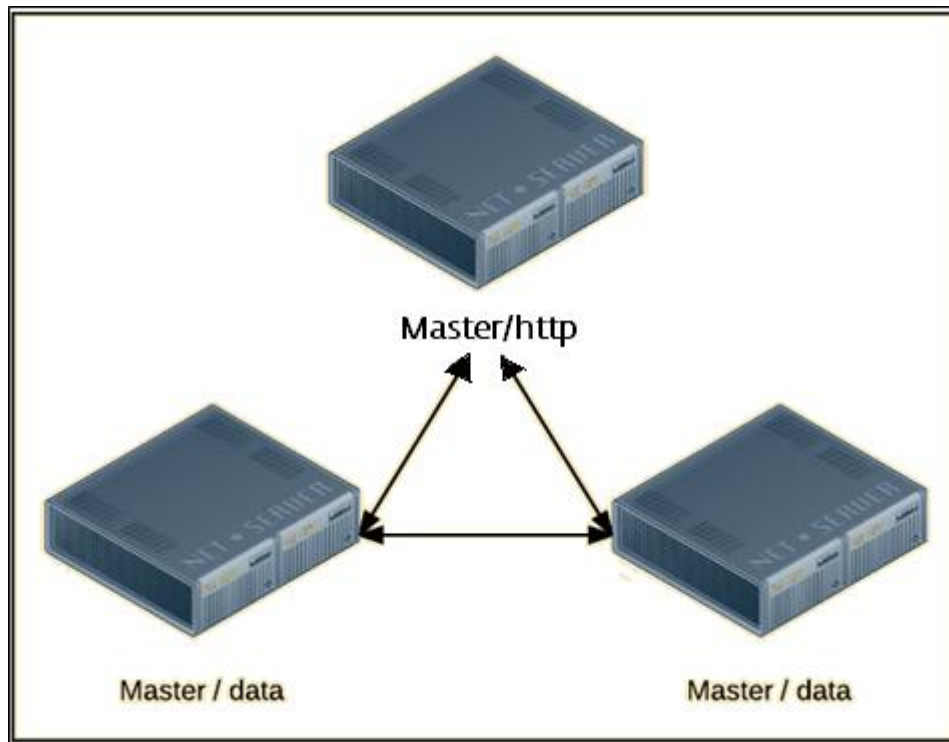
### **3.3 Design of a fault-tolerant system.**

In our case, the system is deployed on several nodes, and therefore it is necessary to further work on the issue of fault tolerance. At the same time, the main task is the development of such a system, in which in the case of the loss of the entire data center, it is possible to guarantee data processing without interrupting the maintenance. In this situation is quite realistic, because the reasons for which the data center can suddenly stop working is a large number.

In most cases, considering the issue of fault tolerance, one of the options may be the loss of already indexed data. Since all data is somehow accumulated in a single data store, where all the major hard disks are located, with information for the last three years of the system. Consequently, the loss of application logs, which are usually stored in the ELK stack, is not critical for the entire infrastructure of the system.

When it is possible for some time to suspend the processing of data, but do not lose the basic data, the fault-tolerant system can be deployed on two nodes. For this, the exact same configuration is used as in the basic case. In doing so, it allows you to significantly distribute the incoming load [23].

However, with the option of notifying about events or if you need to search for events without interrupting the system, everything becomes more complicated. And there is a need for a third server, with the same configuration as in the previous cases. Since the data stream in the case of a telecommunications provider is mostly cut off by local RTSs and only an important part of the messages arrives at the central system, it is sufficient to establish the minimum allowed three nodes in the network. In this case, two of them will be active master nodes.



Picture 3.1 – Basic design of the nodes

Description of the hardware infrastructure for the operation of the information and analytical system.[23]

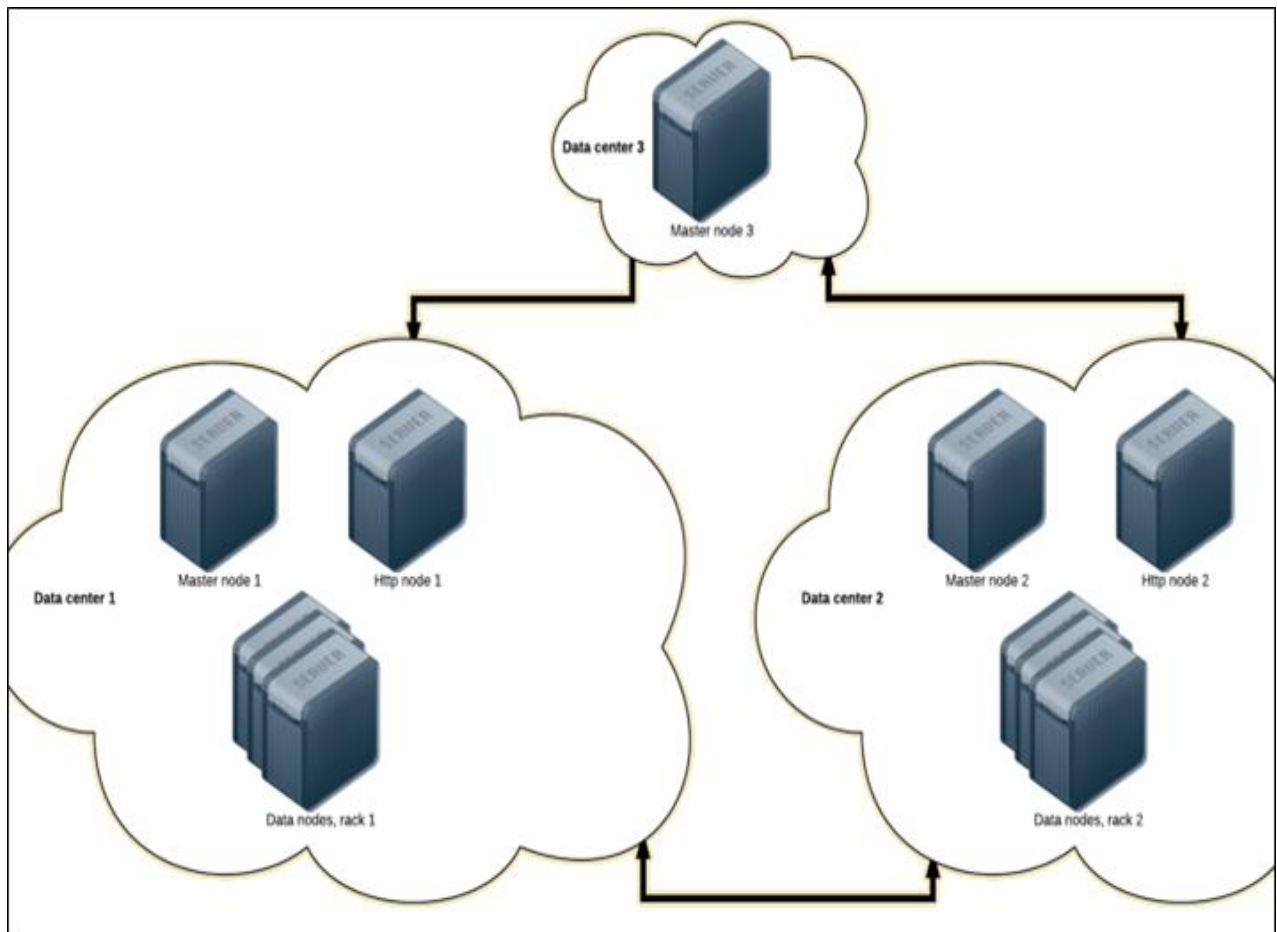
The system will be operated by using the following 4 elements:

- 1) Receiving wizards nodes that perform the main job of managing the entire cluster;
- 2) HTTP servers that process user requests;
- 3) Nodes of data, where all the processed information is available, available for searching for a time period of three years;
- 4) Coordination nodes that perform balancing functions, and are optional with such a small infrastructure.

The minimum requirements for the system are three separate places for placing devices, to minimize losses in the event of an emergency. Namely, two places will be needed to organize the launch of half of the cluster and one more to run the main copy node for fault-tolerant operation.

Three main nodes are used, because according to the basic fault tolerance recommendations, the number of master nodes must be odd. This is necessary to prevent a situation with a complete loss of the data center.

Two http nodes also make up the system's infrastructure. Everyone has a main data center. At the same time, all the load must be evenly balanced for the maximum possible efficiency of the system.



Picture 3.2 – Design of the server part of infrastructure

Elasticsearch provides an opportunity to organize a system, called shard allocation awareness. This function allows you to identify problem fragments and replicas in individual zones. The ability to single out data in one data center in one data zone will allow to limit the problematic part and prevent the entire site from stopping working.[23]

Elasticsearch has made significant changes in the field of event processing. All this became possible thanks to well-organized interaction of three software solutions like Elasticsearch-Logstash-Kibana. In the case of the described infrastructure, it is also a hot data warehouse, which allows you to search for available information.

At the same time, the infrastructure of this system is more or less unified, regardless of hardware.

Heterogeneous events are forced into a single queue. In the process of forming queues, you need to achieve two main goals: you need to make sure that data processing is not a bottleneck for the entire company and you need to make sure that the event is not lost in the case of the system.

The information processing tool also performs functions for normalizing events. This is necessary for data analytics, and since the system uses only uniform information, it brings some of the unification into the operation of the entire infrastructure.

ELK also sends events to a single hot storage, where it is possible to organize a search for the events that happened in the network and to form further reporting.

So now we can describe which servers that we need for the data processing in the master node in the table below. [24]

Table 3.4 – The main parameters of the master node server

Lenovo ThinkServer RD550 (70CX000EEA)	
Processor type	Intel Xeon E5-2630v3
Processor frequency	2400 MHz
Processor core	Haswell
Number of processor cores	8
Volume of RAM	8 Gb
Frequency of memory	2133 MHz
Form factor	1U
Power consumption	750 W
RAID controller	720ix
Volume of hard disk (HDD)	none
Weight	13.5 kg

Table 3.5 – The main parameters of http node server

Lenovo ThinkServer RD550 (70CX0014EA)	
Processor type	Intel Xeon E5-2630v3
Processor frequency	2400 MHz
Processor core	Haswell
Number of processor cores	6
Volume of RAM	8 Gb

*Table 3.5 continuation*

Lenovo ThinkServer RD550 (70CX0014EA)	
Frequency of memory	2133 MHz
Form factor	1U
Power consumption	750 W
RAID controller	720ix
Volume of hard disk (HDD)	none
Weight	13.5 kg

## **4 Life safety**

### **4.1. Analysis of working conditions**

The functioning and support of the working capacity of the analytical system is supported by system engineers. It is necessary to create comfortable and safe working conditions. The workplace is a room located in the city of Almaty, which has an area of 112 m<sup>2</sup>, volume - 448 m<sup>3</sup>. Also in this room there is a window with an area of 22,9 m<sup>2</sup>.

Sufficient illumination of workplaces, technical serviceability of the equipment used, provision of fire safety, as well as the creation of a normal microclimate in the workplace - the conditions necessary for the organization of a comfortable workplace that meets the safety standards.

The room is located in the city of Almaty. This region is characterized by a variety of weather conditions in different periods of the year. Geographical position, the relief of the surrounding area are factors that make the change in weather conditions lightning fast. In this regard, the workplace must create conditions for air comfort. These conditions are formed depending on the systems of aspiration, heating systems, ventilation and air conditioning. Proper management of these systems will create satisfactory working conditions.

The amount of the working equipment in the room is 16. The air conditioning is located near the window. Fire extinguisher is carbonic acid, because of the work character with big amount of electronic devices. It is located near the exit from the room. Working hours are from 9a.m to 6p.m.

Simultaneously in the room are working 2 engineers.

Modern air conditioning systems allow for various manipulations with air. Changing the parameters of cooling, heating, cleaning, setting the humidity is made by easy adjustment. Such systems are very intelligent. The user does not need to pay attention to monitoring these parameters. The system independently monitors

these parameters and also independently supports them within the necessary limits.

In summer, heat input through external structures (walls, ceiling) is usually positive. The calculation is complicated by the fact that the air temperature varies greatly during the day, and the solar radiation further heats the external surface of the building. In winter, heat is lost through external structures. The temperature fluctuations in winter are less, and the heating of surfaces by solar radiation is negligible.

Heat input (or loss of heat) due to the temperature difference depends not only on external conditions, but also on the temperature inside the room.

Calculation of heat receipts due to heat transfer is carried out in accordance with the construction standards SNiP 11-3-79 [26].

Calculation of heat quantity

The amount of heat  $Q_{lim}$ , transferred by heat transfer through a fence (wall) with area  $S$ , having heat transfer coefficient  $k$ , is calculated by the formula:

$$Q_{lim} = S * k * (T - t) * Y \quad (4.1)$$

Here,  $T$  is the calculated outside temperature,  $t$  is the calculated internal temperature, and  $Y$  is the correction factor, the value of which is selected according to SNiP 2.04.05-91 [26].

The calculated outside temperatures depend on the region, and the internal temperatures are chosen taking into account the comfort or technological requirements, depending on the purpose of the room.

This formula is simplified and does not take into account a number of factors. In order to take into account the direction with respect to the sides of the world, solar radiation, heating the walls, etc., it is necessary to introduce corrections into this formula. They are constituent parts of the coefficient  $Y$ .

The absorption of solar radiation by the fence depends on the following factors:

- Wall colors: the heat absorption coefficient reaches 0.9 for the dark color of the outer walls and only 0.5 for the light walls.

- Thermal characteristics of walls: the more massive the wall, the greater the delay in the flow of heat into the room. The thermal load when the massive wall is heated is distributed for a longer time. If the walls are thin and light, then the thermal loads increase and change rapidly when the external conditions change. This requires more expensive and powerful air conditioning.

Heat input from solar radiation through glazed openings

The heat of solar radiation can significantly increase the heat input into the building (for example, in a store with display windows). The room is transferred to 90% of the sun's heat, and only a small part is reflected by the glass. The most intense heat radiation comes in summer, in clear weather.

Heat input of radiation is taken into account in the heat balance of the building only for summer and transition times, when the outside temperature exceeds +10 degrees.

What influences the arrival of heat radiation?

The heat input of solar radiation depends on the following factors:

- Kinds and structures of fencing materials
- The surface states (for example, less radiation passes through dirty glass)
- Angle, under which the sun's rays fall to the surface
- Orientations of the room to the sides of the world (heat losses from radiation through windows facing north are not taken into account at all)

The calculated value of heat input from radiation is taken to be the greater of two values:

1. Heat coming through the glazed surface of that wall, which is most advantageously located relative to the heat input or having the maximum light surface
2. 70% of the heat coming through the glazed surfaces of two perpendicular walls of the room.

If it is necessary to reduce heat losses from solar radiation, it is recommended to take the following measures:

- Orient the rooms to the north
- to make a minimum amount of light apertures

Use protection from sunlight: double glazing, whitewashing of glass, curtains, blinds, etc.

When using complex solar protection, the heat input from radiation can be reduced by almost half, and the power of the required refrigeration unit will decrease by 10-15%.

In rooms for various purposes, the thermal loads that occur outside the room (external ones) act mainly; and also the heat loads arising inside buildings (internal).

External heat loads are represented by the following components:

- 1) Heat input or heat loss, which are determined by the temperature difference inside and outside the building.
- 2) Temperature difference is determined by the fact that in summer the temperature is positive in the building, because Warm air comes from outside into the building, and in winter everything happens in the opposite way and the temperature has a negative value;
- 3) The load also manifests itself in the form of perceived heat through the glazed parts of the building from the sun;
- 4) Heat input from infiltration.

## **4.2 Calculation of heat input due to temperature difference**

In summer, heat input through external structures (walls, ceiling) is usually positive. The calculation is complicated by the fact that the air temperature varies greatly during the day, and the solar radiation further heats the external surface of the building. In winter, heat is lost through external structures. The temperature fluctuations in winter are less, and the heating of surfaces by solar radiation is negligible.

Heat input (or loss of heat) due to the temperature difference depends not



only on external conditions, but also on the temperature inside the room.

Calculation of heat receipts due to heat transfer is carried out in accordance with the construction standards SNIP 11-3-79 [26].

Depending on the time of year and the time of day, external heat loads can be positive.

Heat losses and heat loss as a result of the temperature difference are determined by the formula:

$$Q_{lim}^s = 0, kW$$

$$Q_{limSummer} = V_{room} * X_0(t_{ocalc} - t_{incalc}), W \quad (4.1)$$

where  $V_{room}$  – room volume,  $m^3$ :

$$V_{room} = 14 * 8 * 4 = 448 m^3$$

$X_0$  – specific thermal characteristic,  $W / m^3 \text{ } ^\circ C$ :

$$X_0 = 0,42 W/m^3 * C$$

$t_{ocalc}$  – outdoor temperature. For the cold period - the average temperature of the coldest month at 13 o'clock, for a warm period - the average temperature of the hottest month at 13 o'clock.

$$t_{ocalc} = -26 \text{ } ^\circ C$$

$t_{incalc}$  – internal temperature, is selected taking into account the comfort conditions or technological requirements imposed on production processes.

$$t_{incalc} = 18 \text{ } ^\circ C$$

$$Q_{limsummer} = 448 * 0.42 * (18 - (-26)) = 448 * 0.42 * 44 = 8,279 W$$

### 4.3 Heat input from solar radiation through glazing

Excess heat of solar radiation, depending on the type of glass, is absorbed by the room environment to almost 90%, the rest is reflected. The maximum thermal load is achieved at the maximum radiation level, which has direct and scattered components. The intensity of radiation depends on the width of the terrain, the time of the year and the time of day.

Heat input from solar radiation through glazing is determined by the formula:

$$Q_{sr} = m * F(q^I + q^{II}) * \beta * K^1 * K^2 \quad (4.2)$$

where  $q^I, q^{II}$  – thermal fluxes from direct and diffuse solar radiation,  $W / m^2$ ;

$F_{Io}, F_{IIo}$  – areas of the light opening, irradiated and non-irradiated by direct solar radiation,  $m^2$ ;

$\beta_{c.t.}$  – coefficient of heat transmission

$$\beta_{c.t.} = 0,15.$$

$m = 1$  – number of windows

$F = 22,9 m^2$  - area of the one window

K<sub>1</sub> – glare factor of glazing;  
 K<sub>1</sub>= 0.6;  
 K<sub>2</sub> – coefficient of glazing contamination;  
 K<sub>2</sub> = 0,95.

$$Q_{sr} \text{ at SE} = 1 * 22,9 * (73 + 77) * 0.15 * 0.6 * 0.95 = 293,7 \text{ W}$$

$$Q_{sr} \text{ at SW} = 1 * 22,9 * (214 + 79) * 0.15 * 0.6 * 0.95 = 573,7 \text{ W}$$

$$Q_{sr} = 293,7 + 573,7 = 867,4 \text{ W}$$

#### 4.4 Heat input from people

Internal loads in residential, office or service areas are mainly composed of heat:

- heat that is released by people;
- heat that is released from lighting equipment and electrical appliances;
- heat that stands out from computers, interactive whiteboards, etc.;

Heat input from people depends on the intensity of the work performed and the parameters of the ambient air. The heat emitted by a person is composed of a tangible (apparent), that is, a room transferred to the air by convection and radiation, and the latent heat expended on the evaporation of moisture from the surface of the skin and from the lungs.

In summer at 24 ° C, one man gives an apparent heat of 67 W, and at 18 ° C - 89 W. The woman allocates 85% of the norm of heat emissions of an adult male. Then the allocation of apparent heat in the room will be:

$$Q_p = n_m * Q_{act} + n_w * 0,85 * Q_{act} \quad (4.3)$$

$$Q_{p \text{ at } 24^\circ \text{C}} = 1 * 67 + 1 * 0.85 * 67 = 124 \text{ W}$$

$$Q_{p \text{ at } 18^\circ \text{C}} = 1 * 89 + 1 * 0.85 * 89 = 165 \text{ W}$$

#### 4.5 Heat supply from lighting devices and office equipment

Heat supply from lighting devices, office equipment and equipment is calculated as follows. Heat input from lamps is determined by the formula:

$$Q_{light} = \eta * N_{light} * F_{floor} \quad (4.4)$$

where  $\eta$  – coefficient of transition of electrical energy into thermal energy

(for incandescent lamps)  $\eta = 0,55$ ;

$N_{light}$  – installed lamp power ( $N = 60 \text{ W/m}^2$ );

$F_{floor}$  – floor area =  $14 * 8 = 40 \text{ m}^2$

$$Q_{\text{light}}=0,55*60*112=3696 \text{ W}$$

The heat emitted by the production equipment is determined by the formula:

$$Q_{\text{eq}}=\eta*n*P_1$$

$\eta$  = Efficiency = 0.75

$n$  = number of the equipment = 16

$P_1$  = power of the equipment

$$Q_{\text{equip}}=\eta*n*P_1=0.75*16*0.5*1000=6000 \text{ W}$$

#### **4.6. The overall heat balance and the choice of the split-system air conditioner**

Based on the calculations performed, we will compose the heat balance in the room:

$$Q_{\text{all}}= Q_{\text{equip}}+Q_{\text{light}}+Q_p+Q_{\text{sr}}-Q_{\text{return}} \quad (4.5)$$

$$\text{At } 24^\circ\text{C } Q_{\text{all}}= 6000+3696+124 +867.4 -0= 10687 \text{ W}$$

$$\text{At } 18^\circ\text{C } Q_{\text{all}} = 6000+3696+165 +867.4 -8,279 =2408 \text{ W}$$

The most suitable air conditioner is the Pioneer KFF36UW / KON36UW with the following specifications:

On the area up to: 120 m<sup>2</sup>.

Operating modes: heating / cooling

Compressor type: Normal On / Off

Min. Noise level: 41 dBA

Air purification: Standard

Cooling capacity: 11.6 kW.

Heating power: 11.7 kW.

Electricity consumption: 3.5 kW.

Electricity consumption cold: 3.77 kW.

Dimensions of the indoor unit: 660x1280x205 mm

Dimensions of the external unit: 857x903x354 mm

Weight of indoor unit: 33 kg.

After the performing this calculation we obtained the results of the  $Q_{\text{all}}=10687 \text{ W}$  for summer period and  $Q_{\text{all}}= 2408 \text{ W}$  for the winter. Both of these parameters are average results for this type of calculation. The comfortable temperature is very important for the workers which are basically spend most of their work time in one room. So the right calculation of the heat balance will increase the productivity of the work and will affect to efficiency of the company.

In order to properly design systems that provide normalized parameters of the microclimate in the room, it is necessary to compose the heat balance of the

building (building).

The heat balance is composed of heat loss and heat accumulation in the room. Heat receipts include:

- 1) Heat released by people;
- 2) From solar radiation;
- 3) From sources of artificial lighting;
- 4) From heated equipment and products;
- 5) From equipment using mechanical and electrical energy;
- 6) Heat release from cooling products and material;
- 7) Heat release during condensation of water vapor.

Warmths from solar radiation through the windows, called in SNIIP the term "translucent openings", are determined only for a warm period in the event that in the calculation room there are windows or transparent glazed doors.[26]

Thermal radiation from the sun, which depends on the latitude of the terrain, the orientation of the opening and the estimated hour of the day, can flow through the windows into the room directly with direct sunlight (direct radiation) and by reflection from the surrounding surfaces (scattered radiation). It should be noted that some of the incoming heat flux is absorbed by dust in the atmosphere. At the same time, some of the heat is reflected from the glass. The structure of overlappings also plays an important role in calculating the absorption of heat flux.

These factors cause the fact that, the value of the heat flow is reduced. This may be influenced by a number of different factors, such as: the level of atmospheric pollution and the environment, and the principles of the structure of windows.

Heat, which still gets into the working room itself, can not be transferred completely to heating the air. Since part of this heat flux will be absorbed by the floor and ceiling, as well as the inner walls of the room.

To determine what part will be absorbed, it is necessary to take into account various factors like: the area of internal partitions, the material and the time when solar radiation enters the room.

#### **4.7 Initial data for calculation and selection of the lighting system**

The calculation of the artificial lighting in the working rooms has a big impact on the productivity of work and on visual performance. The absence of the normal light conditions may cause the reduced physical and moral condition of workers and also increase the occupational traumatism.

The conditions of artificial lighting in industrial enterprises have a great influence on the visual performance, physical and moral condition of people, and, consequently, on labor productivity, product quality and occupational traumatism.

Premises with a permanent stay of people should have, as a rule, artificial lighting. When designing new premises, when reconstructing old ones, when designing the natural lighting of rooms, it is necessary to determine the area of the light apertures that provide the normalized value of KEO.

Calculation of artificial lighting is to solve the following tasks: selection of

the lighting system, type of light source, location of luminaires, performing lighting calculations and determining the power of the lighting installation.

For designing of premises it is necessary to define the quantity of light apertures which would correspond to sanitary norms.

We will calculate a system of general lighting for a room with a length = 14m, width = 8 m, height = 4 m, with whitewashed ceiling and light walls with shutters. The working surface will be located at a height  $h_{\text{work}} = 0,7$  m. Normalized illumination, according to SNiP RK 2.04-02-2011, 200 lux.[26]

Calculation of lighting is performed using the light flux utilization method, which is designed to calculate the overall illumination of horizontal surfaces in the absence of large shading objects. The height of the beginning of the window  $h_{\text{b.window}} = 0,8$  m. The discharge of visual work IV, b. The distance to a nearby building  $P = 12$  m, the window height  $h_{\text{window}} = 3.5$  m, the window length  $L = 7.6\text{m}$ , respectively, the window area  $S = 22,9 \text{ m}^2$ .

#### 4.8 Calculation of natural light

Calculation of natural light is to determine the area of light apertures.

The total area of the windows is determined by the formula for lateral illumination:

$$S_0 = \frac{S_n * e_n * \eta_0 * K_{\text{build}} * K_p}{100 * \tau_0 * r_1} \quad (4.6)$$

Where  $S_n$  – floor area,  $\text{m}^2$ :

$$S_n = B * L = 14 * 8 = 112 \text{ m}^2$$

$e_n$  – normalized value of KEO:

$$e_n = e_{KEO} * m * c \quad (4.7)$$

$e_{KEO}$  - value of KEO according to Table 3.12 [25] for IV zone:  $e_{KEO} = 3.5$

$m$  – coefficient of light climate, is determined from Table 3.1 [25] for the orientation of light holes SE  $m=0,65$

$$e_n = 3,5 * 0,9 * 0,8 = 2,52$$

$K_s$  – safety factor according to table 3.11 [25]:  $K_s = 1,75$ ;

$\tau_0$ - common light transmittance  $\tau_0 = \tau_1 \cdot \tau_2 \cdot \tau_3 \cdot \tau_4$ ,

$\tau_1$ - coefficient of light transmittance of the material according to Table 6 [25]: for double glass  $\tau_1 = 0,8$

$\tau_2$ - coefficient that takes into account the loss of light in the light-coverings according to Table 7 [25]:  $\tau_2=0,8$

$\tau_3$ - the coefficient that takes into account the loss of light in the bearing structures, with side illumination is 0,9.

$\tau_4$ - coefficient that takes into account the loss of light in sunscreens, see table 3.6 [25]:  $\tau_4=1$

$$\text{Then } \tau_0 = 0,8 \cdot 0,8 \cdot 0,9 \cdot 1 = 0,576$$

$\eta_0$  – light characteristics of windows according to the table 3.2 [25]:

$$\frac{L}{B} = \frac{14}{8} = 1.75$$

$$h_1 = h_{ok} + h_{b.win} - h_{surf} = 3 + 0,8 - 0,7 = 3,1 \text{ m,}$$

where  $h_1$  – height from the level of the conventional working surface to the top of the window.

$$\frac{B}{h_1} = \frac{8}{3.1} = 2.58$$

$$\eta_0 = 12$$

$r_1$  – coefficient that takes into account the increase in KEO with side lighting due to light reflected from the surfaces of the room and the underlying layer adjacent to the building, see Table 3.9 [25]:

$$\frac{B}{h_1} = \frac{8}{3.1} = 2.58$$

$$\frac{H}{B} = \frac{4}{8} = 0,5$$

$$\frac{L}{B} = \frac{14}{8} = 1.75$$

$$r_1 = 4,3$$

$K_{build}$  – coefficient that takes into account the shading of windows by opposing buildings according to Table 3.8 [25]:

$$K_{build} = 1,7$$

We substitute all the values in the calculation formula:

$$S_0 = \frac{112 * 2.52 * 12 * 1.4 * 1.2}{100 * 0.576 * 4.3} \approx 22.9 \text{ m}^2$$

Since there was one-way side lighting, the area of the light apertures on one side would be 22,9 m<sup>2</sup>

Since the height of the window openings is 3 m, it follows that their length will be 22,9/3=7,6 m.

Thus, the area of the light apertures on both sides will be 22 m<sup>2</sup> (Figure 4.1).

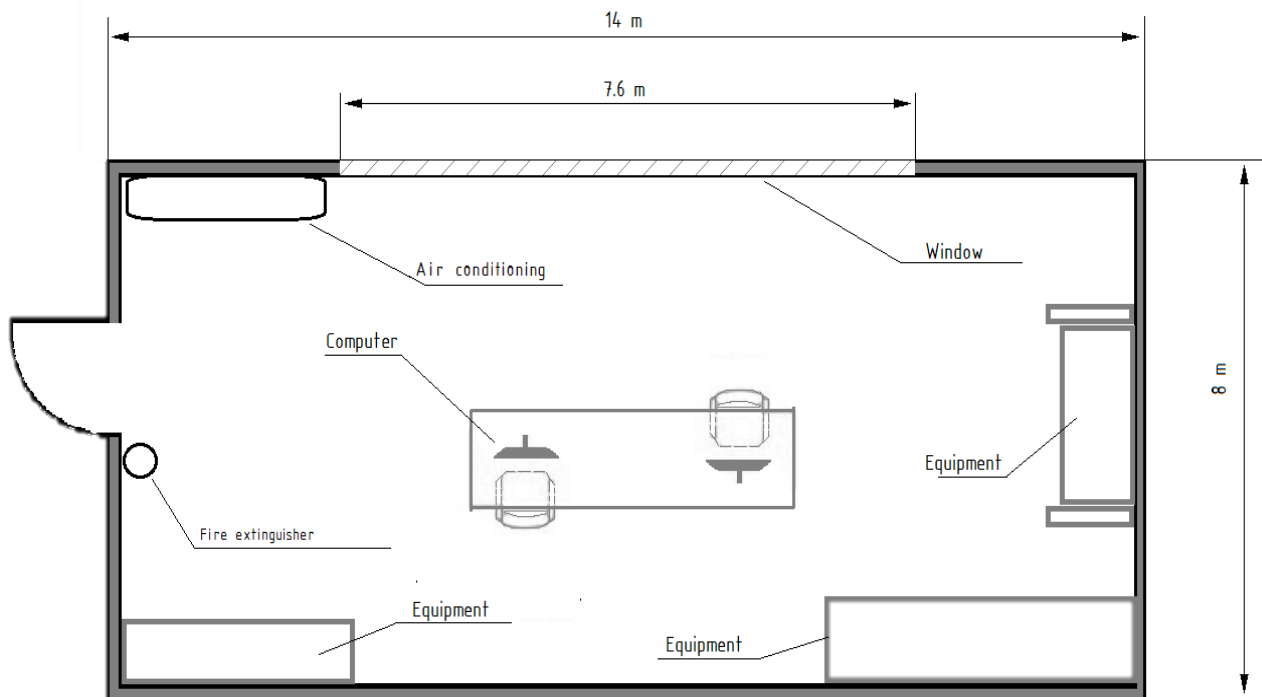


Figure 4.1 – Scheme of the room in natural light

#### 4.9 Calculation of artificial lighting

Discharge of visual work IV (b), therefore the normalized illumination according to Table 3.12 [25] is 200 lux.

We shall verify the correspondence of a given quantity and type of luminaires to a standardized value using a point method.

Determination of the calculated suspension height:

$$h_{\text{calc}} = H - (h_{\text{worksp}} + h_{\text{overhang}}), \quad (4.8)$$

$$h_{\text{calc}} = 4 - (0,7 + 0,2) = 3.1 \text{ m}$$

Distance between luminaires (Z):

$$L_A = 14/7 = 2\text{m},$$



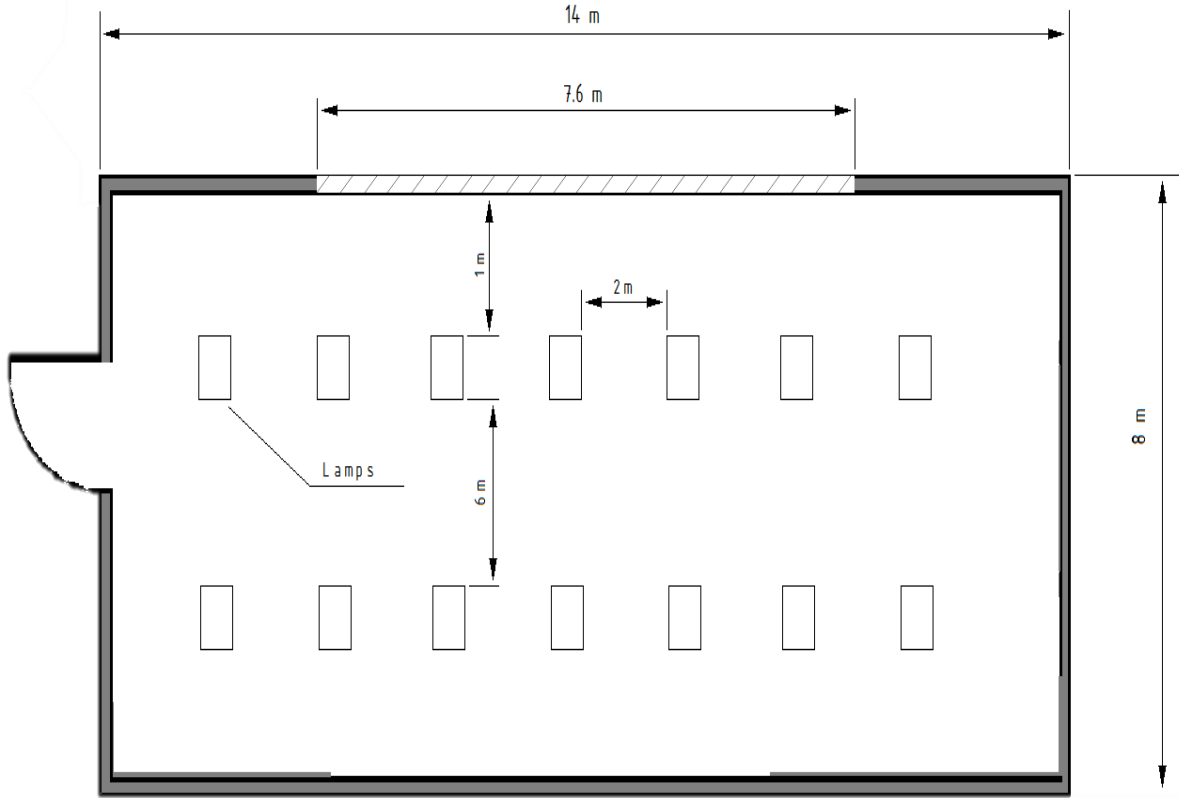


Figure 4.2 – Scheme of the room with the lamps location

We designate the control point A. To this end, we define the complete conditional lighting of all luminaires as follows:

We find the projection of the distance to the ceiling from point A to the luminaire -  $d_i$ .

Next, determine the angle between the ceiling and the line  $d_i$ . On this angle we find conditional illumination. We verify that condition:

$$E_y \geq E_{\text{norm}}, \quad (4.9)$$

where

$$E_y = F \cdot \mu \cdot \frac{\sum_{i=1}^m e_{Gi}}{1000 \cdot K_s}, \quad (4.10)$$

Coefficient of stock  $K_s = 1.5$

Coefficient considering the action of equidistant fixtures  $\mu = 1.15$

Light flow  $F = 3740 \text{ lm}$

Luminaire type: PVLM 1\*40;

$$e_{Gi} = \frac{I_{\alpha_i} \cos^3(\alpha_i)}{h_{\text{calc}}^2}, \quad (4.11)$$

$$\alpha_i = \arctg\left(\frac{d_i}{h}\right), \quad (4.12)$$

The distance from the central point to the luminaire  $d_1$  is found as:

$$d_1 = 3 \text{ m, then}$$

$$\alpha_1 = \arctg\left(\frac{3}{3.1}\right) = 44^\circ, \text{ on this value } I_{\alpha} \approx 84 \text{ cd.}$$

$$e_{G1} = \frac{44 \cdot 0.074}{3.1 \cdot 1.5} = 0.7 \text{ lx.}$$

Calculate  $E_{g2}$ :

$$d_2 = 3.6 \text{ m.}$$

$$\alpha_2 = \arctg\left(\frac{3.6}{3.1}\right) = 49^\circ, \quad I_{\alpha 2} = 70 \text{ cd.}$$

$$e_{G2} = \frac{70 \cdot 0.148}{3.1 \cdot 1.5} = 1.664 \text{ lx.}$$

Calculate  $E_{g3}$ :

$$d_3 = 5 \text{ m.}$$

$$\alpha_3 = \arctg\left(\frac{5}{3.1}\right) = 58^\circ, \quad I_{\alpha 3} = 52 \text{ cd.}$$

$$e_{G3} = \frac{52 \cdot 0.148}{4.65} = 1.61 \text{ lx.}$$

Calculate  $E_{g4}$ :

$$d_4 = 6.7 \text{ m.}$$

$$\alpha_4 = \arctg\left(\frac{6.7}{3.1}\right) = 65^\circ, \quad I_{\alpha 4} = 44 \text{ cd.}$$

$$e_{G4} = \frac{44 \cdot 0.37}{4.65} = 3.44 \text{ lx.}$$

The total conditional illumination is:

$$\sum e_g = 2 \cdot 3.44 + 4 \cdot (0.7 + 1.664 + 1.61) = 39.456 \text{ lx}$$

The total illumination is:

$$E_{AG} = \frac{\mu \cdot F_{\lambda} \cdot 2}{1000 \cdot K_s} \cdot \sum E_r = \frac{3740 \cdot 1.15}{1000 \cdot 1.5} \cdot 39.456 = 64.5 \text{ lx}$$

Illuminance in the workplace is not considered sufficient, therefore, we perform reconstruction of illumination.

We will perform the reconstruction using the coefficient of utilization

method.

Define the index of the room (i):

$$i = \frac{A \cdot B}{h_{calc} \cdot (A + B)} = \frac{14 \cdot 8}{3.1 \cdot 22} = 1.64$$

Let us determine the coefficient of use of the light flux ( $\eta$ ):

According to the table 5.12 [29]  $\eta=85\%$

Number of lamps with the necessary illumination  $E=200$  lx:

$$N = \frac{E_n \cdot S \cdot Z \cdot K_z}{F \cdot \eta}, \quad (4.13)$$

where  $Z$  – coefficient of uneven lighting, equal to 1,1-1,2;

$K_z$  – coefficient of uneven lighting, equal to

$$F = \frac{400 \cdot 1.5 \cdot 14 \cdot 8 \cdot 1.15}{14 \cdot 0.6} = 9200 \text{ mF}$$

In order to provide the necessary level of illumination in the office with the parameters of 14x8x4 meters, you need to choose a DLL lamp, whose power is 250 watts. The luminous flux of this type of lamp reaches 13,000 mF. According to the general layout of the room, the number of lamps is 14 pieces located along two lines at a distance of 6 meters from each other and 2 meters between the lamps. We choose a DLL lamp with a power of 250 W and a light flux of 13000 mF. [27, 29]

When designing a room layout, it is also necessary to determine the area of the light holes that are in the room. They provide the normal value of KEO. For a room with dimensions of  $14 \times 8 \times 4$  meters to provide a normalized value of KEO, the  $e_N$  is 0.585, for the IV visual performance, light holes with a total area of  $22 \text{ m}^2$  are required.

Also, in order to provide comfortable working conditions, the calculation of artificial lighting in the room was made. The initial calculation by the point method made it possible to find out that that light flux is insufficient and it was necessary to produce the type of lamps for more accurate ones.

The point method for calculating these parameters was applied in connection with the fact that with its help it is possible to perform calculations at the level of nominal illumination. The main disadvantage of this method of calculation is the fact that it is impossible to accurately determine the efficiency of lamps [28].

## 5 Estimation of economic efficiency of the project

The software product which is considered has ample opportunities for implementation at enterprises that have different areas of activity. At the moment, the data analyst in one form or another is used universally. This chapter will describe the effect of implementing an information and analytical system in the

infrastructure of the largest telecommunications provider in Kazakhstan. Due to the fact that the volume of incoming data from radio and TV stations will only grow, the implementation of this analytical system will eliminate the need for additional hiring of staff for data processing, reduce the human factor in information transfer and management decisions, accelerate internal business processes by reducing the time on the formation of accountability for responsible persons.

Also an important effect is the optimization of labor costs, aimed at reporting for the central apparatus of the telecommunications provider, and thus engineers who previously deal exclusively with reporting on the events occurring in the network may be transferred to more qualified work sites. As already described, at the moment there is a problem of interaction between the control department and technicians at regional radio and television stations, due to the low qualifications of the latter and local features of the terminology. The introduction of an information and analytical system will allow standardizing and unifying the process of describing emerging events [30].

Due to the fact that the response to emergent accidents in the network will be accelerated, it will allow timely to carry out the necessary preventive measures, thereby reducing the risks associated with the provision of services and services to end users. This reflects the indicator of network availability. This parameter is strategically important, since it is the part of the information security of the country. Any actions that are aimed at improving the situation with such a field of activity as the provision of television and radio services are critical not only for the leadership of the telecommunications provider itself, but also for the Ministry of Information and Communication of the Republic of Kazakhstan and the National Security Committee.

The importance of timely transmission of information to the public through television calls cannot be overemphasized. Consequently, the introduction of an information-analytical system of events occurring in the telecommunications network will produce a strong social effect. After all, in the event of an emergency, it is this method of communication that is the main one.

### **5.1 Capital expenditures for the implementation of the investment project**

In order to determine the necessary investment for the development and implementation of the information and analytical system, it is necessary to begin with.

In this case, the total amount of all necessary investments will be determined by the formula:

$$\sum K = K_{soft} + K_{equip} + K_{educat}, \quad (5.1)$$

where  $K_{soft}$  - capital investments for development of information-analytical system;  
 $K_{equip}$  - capital investments for the purchase of additional server capacities and equipment;

$K_{educat}$  - capital investments for the subsequent training of users of the system.

Before considering all the investments separately, it was necessary to determine how the software would be developed, and therefore to estimate the possible costs.

In the case of a telecommunications provider, a study was conducted on possible software solutions for the implementation of the task facing the information and analytical system. There are several possible solutions and Table 5.1 shows their comparative characteristics:

Table 5.1 - Comparison of existing solutions

Program ming solutions	Has the necessary functionality	Possible to set on own infrastructure	Usage simplicity	Absence of the continuou s paying	Existing of local technical assistance	Possible to transfer to IT department
Event Registrati on Manage ment System	+	+	+	+	+	+
SaaS service	+	-	+	-	-	-
CMDB	+	+	-	+	-	-
Commer cial solutions	+	+	-	-	+	-

As can be seen from the table, the best solution for interaction between technicians on the RTS and all other responsible services is when developing a system based on the system for registering and managing events. The main factor in favor of this solution is that in the future, the complete independence of the telecommunications provider from the software vendor is possible. Due to the fact that the product being developed is of national importance, this factor is very important.

According to the commercial offer from suppliers of system solutions, the cost of development will be:

$$K_{soft} = 21\,200\,000 \text{ (tg)}.$$

where  $K_{soft}$  - the price from the developer of system decisions from ICT service.

However, after the system has been developed, all rights to use it and use for commercial purposes are transferred directly to the telecommunications provider without time limits.

In order to support the functioning of the information and analysis system, three Lenovo ThinkServer RD550 servers (70CX000EEA) will be needed. This choice is due to the technical characteristics and the necessary processing performance. And also will use two http servers Lenovo ThinkServer RD550 (70CX0014EA).

Also, in connection with the increased amount of data related to information analysis, it will be necessary to install 8 additional Western Digital Blue 6TB 5400rpm 64MB WD60EZRZ 3.5 SATAIII. To connect all the servers to the existing infrastructure, you will need to install cable boxes and patchcords. Since a permanent system life support is required for the equipment, an uninterruptible power supply unit will be additionally installed. In connection with the fact that the servers will stand in the existing server room, the installation of additional cooling elements and the purchase of racks is not required [24].

The list of necessary equipment with purchase prices is given below in Table 5.2:

Table 5.2 - list of necessary equipment

Equipment	Price for one, tg	Quantity	Sum tg.
Lenovo ThinkServer RD550 (70CX000EEA)	830800	3	2492400
Lenovo ThinkServer RD550 (70CX0014EA).	531700	2	1063400
Western Digital Blue 6TB 5400rpm 64MB WD60EZRZ 3.5 SATAIII	46248	8	369984
Power supplier	34572	1	34572
Cablegone 15 meters	3000	1	3000
Patch cord Ethernet 15 meters	6000	1	6000
Total			3 969 356

As for the development of software, the engineers of the telecommunications provider will be involved only for consultations, then there will be no additional costs associated with training the employees to create the system. However, after the system has been put into operation, although staff members are already familiar with certain elements, it is still necessary to carry out a number of explanatory measures. Since they will be conducted by the staff of the information technology

department of the telecommunications provider, the costs will only be required for training employees of this direction. Training will be conducted directly by the developer of the information and analytical system.

Table 5.3 - Costs associated with training employees to work with the system

Service list	Cost, tg.
One training course for the whole department	300000
Brochures and manuals for working with the system (50 pcs.)	100000
Total	400000

Summarizing all the expenses listed above, it is possible to make them into a single table on the necessary investments for system development:

Table 5.4 - Total investment required for implementation

Nomenclature	Price, tg
Development of an information and analysis system by a third-party service provider	21 200 000
Cost of equipment and server capacity	3 969 356
The cost of training DIT staff to work with the system	400 000
Total	25 569 356

$$\sum K = 21\,200\,000 + 3\,969\,356 + 400\,000 = 25\,569\,356 \text{ tg.}$$

As a result, it can be said that the investments necessary to implement this system represent costs comparable to the cost of technical support for the monitoring system per year. However, the effect of its implementation is comparable to the development of a monitoring system for digital radio and television stations.

## **5.2 Description of labor costs associated with the work of the technical department, control service, the department of analytics and the central apparatus without the use of information and analysis system**

In the course of polling employees at local radio and television stations, a number of processes were clarified, which are performed by technicians at all 450 stations that have been entered at the moment. Since the total number of stations will be 827, in the future, without the introduction of software, the final labor will grow exponentially, and not linearly. This is due to the fact that human risks in the process of preventing and eliminating accidents increase exponentially in connection with the increasing burden.

Table 5.5 - Labor costs of technical personnel on RTS

Workflows of the technical service				
Name	Time per unit (hour)	Frequency per month	Number of repetitions	Total complexity of man-hours
Manual switching of inputs on transmitters	0,25	1	450	112,50
Manual switch of exciters on transmitters	0,25	1	450	112,50
Change of input parameters of receivers	0,25	0,0833	450	9,37
Formation of multiplex on RTS receivers when adding or removing TV channels	0,75	0,1667	450	56,26
Reboot of CAM-modules on RTS	0,5	2	4	4,00
Reboot VB120 on RTS	0,0833	1	1	0,08
Update of VB120 software on RTS	3	0,0833	450	112,46
Downloading logs from the RTS transmitter for incidents	0,166	1	1	0,17
Downloading logs from RTS switches in case of incidents	0,166	0,25	6	0,25
Monitoring Version of RTS Switch Configurations	0,166	0,5	14	1,16
Automating the launch of DMX services on the NMX	0,166	0,33	15	0,82
Reboot TSE800 / NetCCU / Exciter on RTS	0,25	2	88	44,00



*Table 5.5 continuation*

Name	Time per unit (hour)	Frequency per month	Number of repetitions	Total complexity of man-hours
Reloading of air conditioners on RTS	3	1	24	72,00
Reloading Hughes on RTS while hanging up	0,25	1	88	22,00
Elimination and prevention of accidents on the RTS, maintenance and additional installation of devices, registration of events in the electronic journal	8	30	80	19200
Total				19747,57

After the initial information was collected by the technical personnel directly on the RTS, it is submitted for further consideration to the control service. At the moment this process is as follows:

There is a monitoring system that captures events occurring on the network. All important events that in some way may affect the provision of television and radio broadcasting services are recorded in an electronic journal. However, in order to transfer this data to the journal, 5 people are involved in 24 hours 365 days a year. This is as follows: telephone monitoring specialists call operators at large radio and television stations and regional TV and radio broadcasting administrations and upload the information to the technical log in the form of an Excel spreadsheet. In this case, operators have difficulties in filling out the data. They are related to the fact that the qualification of individual technicians located on remote RTS may not be sufficient to accurately form the cause and describe the emergency situation that has arisen at the station. At the same time, the same device can be called differently in different regions of the country, the so-called "business" name. At the same time, there is no single standard and tools for compiling such reports.

After all these data are collected, they become the basis for the future report for the central apparatus of the telecommunications provider. Consequently, the labor costs of the control service can be presented in Table 5.6

Table 5.6 - Labor costs of the monitoring service associated with journalization network events

Controlling service workflows				
Name	Time per unit, hour	Frequency per month	Number of repetitions	Total labor intensity, man-hours
Formation of reports "Journal of TS work"	120	30	1	3600
Provision of data for the formation of reports from operators with ODRT and RTS	1	23	14	322
Total				3922

After the initial report is generated, it immediately passes to the process coordination and network quality analysts. There, according to regulatory documents regarding equipment downtime and the termination of broadcasting, a report on the availability of the network is generated. From these data, a report is compiled into TV and radio channels and the Ministry of Information and Communications of the Republic of Kazakhstan. Based on this report, tariffs for TV and radio channels and advertising, as well as reports on the information security of the country, are being formed.

The labor costs for interacting with the reporting of events occurring in the network can be described as follows: three analysts spend an average of 1.5 working weeks per month on the formation of the report. From this, you can describe the total labor costs in Table 5.7:

Table 5.7 - The labor costs of the process coordination service and network quality analysts aimed at generating reporting

Workflows for Process Coordination and Network Quality Analyzes				
Name	Time per unit, hours	Frequency per month	Number of repetitions	Total labor intensity, man-hours
Generating a network availability report	192	1	1	192
Total				192

At the same time, the importance of this parameter can be assessed on the basis of the possibility to transmit the message to the entire population of the country, through the continuous broadcasting of national channels. In the case of inaccessibility, some television services, there may be a misunderstanding of some aspects of the strategy.

Now, when all the necessary data regarding the interaction of services with reporting related to the provision of information by the controlling and responsible authority are obtained, it is possible to calculate the total amount of costs for these processes [31].

The cost is the following formula:

$$C_{\text{current}} = WF + C_{\text{soc}} + P_e, , \quad (5.2)$$

where WF is the Wage Fund;

$C_{\text{soc}}$  - social tax;

$P_e$  - the cost of electricity;

At the same time, the payroll fund is a WF of the technical service, control service and process coordination services and network quality analysts.

$$WF = WF_{\text{tech.}} + WF_{\text{contr.}} + WF_{\text{analyt.}}$$

where  $WF_{\text{tech.}}$  - The payroll of technical service;

$WF_{\text{contr}}$  - payroll of the control service;

$WF_{\text{analyt.}}$  - Wage fund of the service of process coordination and network quality analysis.

The wage fund, in turn, consists of the basic and additional wages:

$$WF = P_{\text{basic } i} + P_{\text{add } i}. \quad (5.3)$$

Where  $WF_i$  - the payroll of the i-th employee;

$P_{\text{basic } i}$  - the basic salary of the i-th employee;

$P_{\text{add } i}$  - is the additional salary of the i-th employee.

The basic wage in this case is determined by the formula:

$$P_{\text{basic } i} = P_{\text{av } i} \times T_i, \quad (5.4)$$

where  $P_{\text{basic } i}$  is the wages of the i-th employee (n);

$P_{\text{av } i}$  - average daily salary of the i-th employee;

$T_i$  - labor intensity of the work of the i-th personnel, (pers.d.).

In this case, the additional salary of performers, taking into account the loss of time for leaves and sickness (an average of 15% of the basic salary is taken):

$$P_{\text{add } i} = 0,15 \cdot P_{\text{basic } i}, \quad (5.5)$$

The payroll fund is:

Table 5.8 - Calculation of the wage fund

Personnel	$P_{av\ i, \text{ tg/day}}$	$T_i, \text{ h.days}$	$P_{basic\ i, \text{ tg}}$	$P_{add\ i, \text{ tg}}$	$WF_i, \text{ tg}$
Technical service (engineer on RTS and RDRT)	5000	822	4 110 000	616500	4726500
Control service (specialist of controlling service)	9400	164	1 541 000	231 150	1 772 150
Service for the coordination of processes and network quality analysis (specialist of the department of analytics)	7300	8	58 400	8760	67 160
WF, tg					6 565 810

Social tax is calculated by the formula:

$$P_{soc} = (WF - P_p) \cdot 9,5\%, \quad (5.6)$$

where  $P_p$  - pension contributions.

The PF is 10% of the FOT and is calculated by the formula:[31]

$$P_p = WF \cdot 0.1, \quad (5.7)$$

$$P_p = 6\,565\,810 \cdot 0.1 = 656\,581(\text{tg}).$$

Thus, the social tax is equal to:[31]

$$P_{social} = (6\,565\,810 - 656\,581) \cdot 9,5\% = 561\,376(\text{tg}).$$

Calculation of electricity costs

[Decision on the price kW / h from January 1, 2018]

When calculating the electricity costs for the time of report formation, only data are used regarding the functioning of the control unit on the RTS and ODRT, as well as the operation of the servers in the two communication control centers in Almaty and Astana.

The servers on the RTS are installed in a single sample. There is a group of servers on the RDRT, there is a whole cluster on the CCC. The cost of electricity

can be calculated depending on the number of working elements. Since the power consumption by servers at the telecommunications provider is round the clock, it is possible to use the following formula when determining costs:[32]

$$E_{total} = (E_{RTS} * N_{RTS} + E_{RDRT} * N_{RDRT} + E_{CCC} * N_{CCC}) * t * 30 \quad (5.8)$$

where  $E_{total}$  - the general consumption of the electric power

$E_{RTS}$  - power consumption by servers on one RTS

$E_{RDRT}$  - power consumption by servers on one HWRT

$E_{CCC}$  - power consumption by servers on one NCC

$N_{RTS}$  - number of RTS

$N_{RDRT}$  - amount of RDRT

$N_{CCC}$  - number of NCC

$t$  - time of electricity consumption (24 hours a day)

Therefore, according to the formula 3.8, the amount of electricity consumed will be:

$$E_{total} = (220 * 450 + 500 * 14 + 8000 * 2) * 24 * 30 = 122000 * 24 * 30 = 87\,840\,000 \text{ kW} * \text{hour/month}$$

The total amount of electricity costs can be calculated by the formula

$$P_e = E_{total} * Tariff * K_{use}, \quad (5.9)$$

where  $E_{total}$  is the cost of electricity;

Tariff - tg / hour;

$K_{use}$  - factor, 0.9.

$$P_e = 87\,840\,000 * 22,9 * 0,9 = 1,81 \text{ billion, tg.}$$

Accordingly, this number shows how much the largest telecommunications provider spends on electricity per month to support the operation of all the servers in Kazakhstan in the telecommunications network, which interact with each other to ensure the transfer of data.

Now we can calculate the total costs of the telecommunications provider at the moment, which in one way or another are related to the work and processing of information about the functioning of the network. By the formula 5.2:

$$P_{current} = 6\,565\,810 + 561\,376 + 1\,810\,000\,000 \text{ tg} = 1\,887\,127\,186 \text{ tg.}$$

Thus it is possible to allocate separately expenses for a payment fund and a social tax:

$$P_{WF+social\ tax} = 6\,565\,810 + 561\,376 = 7\,127\,186 \text{ tg.}$$

### 5.3 Calculation of operating costs of a telecommunications provider using an information and analytical system

After the main cost items related to the work of the telecommunications provider in the sphere of data processing and transmission were analyzed, it is possible to estimate the rapid effects of the introduction of the information and analysis system. In this case, the effects associated directly with the data analytics and the conclusions based on them are not considered, since these results are difficult to predict. However, the effect on the automation of some processes can be estimated now.

If you consider directly by services, you can start with a technical service.

Table 5.9 - Labor costs of technical personnel on RTS

Workflows of the technical service				
Name	Time per unit (hour)	Frequency per month	Number of repetitions	Total complexity of man-hours
Elimination and prevention of accidents on the RTS, maintenance and additional installation of devices, registration of events in the electronic journal	8	30	80	19200
Total				19200

As it is possible to see all the third-party, routine actions will be automated, which will enable technicians to deal exclusively with tasks to eliminate and prevent accidents.

The labor costs of the control service will be reduced to the work of one person for one hour per shift [31].

Table 5.10 - Labor costs of the controlling service of logging-related network events

Controlling service workflows				
Name	Time per unit (hour)	Frequency per month	Number of repetitions	Total complexity of man-hours
Formation of reports "Journal of TS work"	3	30	1	90
Provision of data for the formation of reports from operators with ODRT and RTS	0,2	23	14	64,4
Total				154,4

It is possible to see the main effects of automation are related to the control service. And as a result, you can send the employees' free time to develop scenarios and instructions for eliminating emergency situations on the RTS.

The work of the process coordination and network quality analysis service aimed at generating reports is also reduced to the work of one specialist for 3 hours per month.

Table 5.11 - Labor costs of the process coordination service and network quality analysts aimed at generating reporting for the dispensers

Workflows for Process Coordination and Network Quality Analyze				
Name	Time per unit (hour)	Frequency per month	Number of repetitions	Total complexity of man-hours
Generating a network availability report	3	1	1	3
Total				3

Estimating the current data, it is possible to calculate the total cost of a payroll using an information and analytical system.

Table 5.12 - Calculation of the wage fund with application

Personnel	$P_{av\ i, \text{ tg/day}}$	$T_i, \text{ h.days}$	$P_{basic\ i, \text{ tg}}$	$P_{add\ i, \text{ tg}}$	$WF_i, \text{ tg}$
Technical service (engineer on RTS and RDRT)	5000	800	4 000 000	600 000	4 600 000
Control service (specialist of controlling service)	9400	6,43	60 442	9066	69 508
Service for the coordination of processes and network quality analysis (specialist of the department of analytics)	7300	0,125	912	136	1048
WF, tg					4 670 556

From these data it can be seen that the main quick effect from the implementation of this information-analytical system is observed in the form of automation of most of the work of the control service.

If consider changes in the cost of electricity, then the changes when connecting five servers will be completely insignificant for comparison.

The social tax will be[31]

$$P_{social} = (4\ 670\ 556 - 467055) \cdot 9,5\%,$$

$$P_{social} = (6\ 565\ 810 - 656\ 581) \cdot 9,5\% = 399\ 332 \text{ (tg)}.$$

Total costs for the labor compensation fund and social tax will be:

$$P_{WF+social\ tax\ with\ inf.system} = 4\ 670\ 556 + 399\ 332 = 5\ 069\ 888 \text{ tg}.$$

#### 5.4 Calculation of economic efficiency

An indicator of the effectiveness of the implementation of the information and analytical system is the reduction of labor costs for reporting on the availability of the network. In this case, the values for 450 digital broadcast stations were considered, but in two years these stations will be 2 times larger, therefore it is possible to extrapolate the results to other stations. At the same time, the speed of report formation will be reduced from a week to several hours. As a result, the speed of response and allocation of funds to eliminate the problems that arise will be significantly reduced. The level of qualification of the employees of the control



service will be increased. And the control itself will become more thorough and visible.

$$P_{WF+social\ tax\ with\ current} = 7\ 127\ 186\ \text{tg per month}$$

$$P_{WF+social\ tax\ with\ inf.system} = 5\ 069\ 888\ \text{tg per month}$$

Annual costs for the creation of an information and analytical system

$$P_{inf.an.syst} = 25\ 569\ 356\ \text{tg.}$$

Annual savings of financial assets ( $E_{year}$ ):

$$\begin{aligned} E_{year} &= (P_{WF+social\ tax\ with\ current} - P_{WF+social\ tax\ with\ inf.system}) * 12\ months \\ &= (7\ 127\ 186 - 5\ 069\ 888) * 12 = 24\ 687\ 576\ \text{tg.} \end{aligned}$$

Economic efficiency is:

$$E_a = \frac{E_{year}}{\sum K} = \frac{24\ 687\ 576\ \text{tg.}}{25\ 569\ 356\ \text{tg.}} = 0,965$$

Payback period for this project:

$$T_p = \frac{1}{0,965} = 1,036\ \text{year} = 12,5\ \text{months}$$

The normative coefficient of economic efficiency of capital investments:

$$E_n = \frac{1}{1,036} = 0,96$$

The value of the expected economic effect in the first year from the introduction of the information and analytical system:

$$E_y = 24\ 687\ 576 - 25\ 569\ 356 \cdot 0,965 = 6730\ \text{tg}$$

## 5.5 Conclusion about economic efficiency

According to the assessment of comparative economic efficiency, it was found that when implementing an information and analytical system it is possible to optimize labor costs associated with the acceptance of information from engineers on regional RTSs. It also shows the reduction in the labor costs of the monitoring service for reporting to the central apparatus (management) and the telecommunications provider's analysis service.

The introduction of the information and analytical system will allow to increase the indicators of network availability, and as a result, to increase the availability of services for the end user. This measure will improve the quality of TV and radio broadcasting services and will provide opportunities to prevent emergencies arising on RTS and RDRT. The continuity of broadcasting is an important factor for the information policy not only of the service provider, but also of the Ministry of Information and Communications.

During the development of the system, the telecommunications provider will receive a powerful infrastructure for data analytics, not only about the status of the network, but also of all types of information. In the future such methods can be provided on a commercial basis in the form of an additional service. No other company in the country has such significant computing resources. And the information and analytical system will allow using them with maximum economic efficiency.

Comparison with the current labor costs shows that annually the telecommunications provider will save 24 million tenge, and during the commissioning of all TV and radio broadcasting stations this figure will reach almost 50 million annually, not counting the profits received from the provision of analytical services to outside organizations.

## Conclusion

During the implementation of this project, the necessary tools were selected with open source, with the possibility of developing their own modules for interacting with data sources. This choice made it possible to develop modules for interaction with the monitoring system Zabbix, as well as with the CMDBuild system. During the selection of the necessary software solutions, a sequential comparison was made between the two systems, namely Elasticsearch and Splunk. This analysis made it possible to determine the most suitable solution for the needs of the information and analytical system. This solution was Elasticsearch, as it possessed all the necessary list of characteristics. This software infrastructure has been tested on the basis of data collected through the system for collecting proposals of the Almaty University of Power Engineering and Telecommunications. This procedure allowed to debug the interaction of infrastructure elements and the correct display of the analysis results. The calculation of the required hardware for the functioning of the information-analytical system was performed. The size of the incoming data stream from the monitoring system and from the CMDBuild system was determined. This allowed us to assess the required financial investments in the project, namely, in hardware and software development and determine the economic efficiency of investments. In the life safety section, a plan of the required room has been developed, artificial and natural lighting calculations, as well as ventilation systems

At the moment, the data analyst of the telecommunications provider has become a top priority. This is due to the fact that the process of building radio and television stations is already established and will soon reach its peak. The developed infrastructure for analytical systems will become the basis on which all other software solutions that will be developed by the company will be based. Given the great potential of telecommunications providers in the field of computing resources, it is possible to assume that this analyst will be able to appear as a separate service for private companies.

To assess how the events will develop in terms of the load on the developed infrastructure, it is necessary to consider the pace of construction of radio and television stations in Kazakhstan.

The main document that sets the terms for the development of the telecommunications industry is the state program Digital Kazakhstan. The transition to the latest digital technology has been underway since 2011. In this way, measures were taken to transfer satellite TV and radio broadcasting to the DVB-S2 / MPEG-4 standard. The number of service users provided by the telecommunications provider has also grown. Their number was 1.2 million connections. It is necessary to ensure the pace of construction and technical infrastructure. As of mid-2018, about 450 stations have been put into operation and a monitoring system, and there will be 82 in all. That is why the preparation of a technical infrastructure is a priority. After all, previous models of work showed their non-competitiveness. Handling ever-increasing amounts of data manually and spending most of the time working for

skilled professionals to correct and unify reporting is an unaffordable luxury.

Preliminary testing on the basis of a pilot project in a higher educational institution allowed to identify the basic errors and eliminate them when working with a large telecommunications provider, where the data flow is incommensurably larger and more constant. However, even on the basis of the university, the infrastructure achieved high results in terms of performance and fault tolerance, which, of course, does not allow us to make a final conclusion about the reliability of the system while working with a huge data flow.

Users and customers also want to be heard by the 's management. But at the moment there was no effective mechanism for creating and processing incoming messages. There was no consolidation of data, there was no possibility for carrying out analytics and collecting statistical data. The developed infrastructure for analytical systems allows to make so that data from the equipment go in close interaction with the data received from users. It is this formula of work that will allow you to achieve your goals in the state program. Analysis and structuring of information is a priority area of development in the field of data management in the coming years. Information coming from various sources is one of the most important resources of our time. It is their careful processing that will help create new competitive advantages, as well as improve the level of services provided.

## **List of Abbreviations**

IAS – Information-analytical system  
OLAP - On-Line Analytical Processing  
API application programming interface  
JSON - JavaScript Object Notation  
XML - eXtensible Markup Language  
REST API - Representational State Transfer  
HTTP - HyperText Transfer Protocol  
ELK –Elasticsearch Logstash Kibana  
NFS - Network File System  
SNMP - Simple Network Management Protocol  
TCP - Transmission Control Protocol  
UDP - User Datagram Protocol  
ETL - Extract, Transform, Load  
SSH - Secure Shell  
SSL - Secure Sockets Layer  
ESB - Enterprise Service Bus  
MIB - Management Information Base  
IP - Internet Protocol  
UDP - User Datagram Protocol  
VSAT - Very Small Aperture Terminal  
DVB-T2 - Digital Video Broadcasting — Second Generation Terrestrial  
CCC – Communication Control Center  
RDRT – Regional Directory of Radiotransmission

## List of references

- [1] The governmental program "Digital Kazakhstan" for 2017-2020y.
- [2] Article: Analog TV is inferior to digital, or Modern trends of the Kazakhstan segment of TV viewing 23.03.2018.// free access by nomad.su (request date 08.05.2018).
- [3] Article: Results of 2017: telecom in Kazakhstan. The most important events, achievements, industry revenues in the past year, the prospects for the development of the industry.// free access by profit.kz (request date 08.05.2018).
- [4] Description of technical abilities of search Lucene// free access by: <https://lucene.apache.org>, (request date 20.02.2018).
- [5] Description of technical abilities of software Elasticsearch// free access by: [www.elastic.co](http://www.elastic.co) (request date 20.02.2018)
- [6] Administrator's Guide to using the Software Kibana 2018// free access by: [www.elastic.co/products/kibana](http://www.elastic.co/products/kibana), (request date 21.02.2018)
- [7] Description of the Zabbix monitoring system// free access by <http://www.zabbix.com/ru/> (date of request 10.05.2018)
- [8] Description of the Elasticsearch Python client API// free access by: <http://elasticsearch-py.readthedocs.io/en/master/> (date of request 10.05.2018)
- [9] Andrei Smolyaninov: Event management. Process or system?// free access by : <https://www.osp.ru> //04.04.2013
- [10] Technical description of the JSON format// free access by: [www.json.org](http://www.json.org), (date of request 10.05.2018)
- [11] Guide for the creation of reports in Elasticsearch // free access by: <https://www.elastic.co/guide/en/reporting> // (date of request 10.05.2018)
- [12] Technical guide for the Logstash programming product// free access by <https://www.elastic.co/products/logstash> //(date of request 10.05.2018)
- [13] Technical guide for the search engine Splunk// free access by <https://www.splunk.com> // (date of request 10.05.2018)
- [14] Asaf Yigal // Splunk and the ELK Stack: A Side-by-Side Comparison // free access by <https://devops.com> (date of request 10.05.2018) //June 27, 2017
- [15] Alex Zhitnitsky - Splunk vs ELK: The Log Management Tools Decision Making Guide // free access by <https://blog.takipi.com> (date of request 10.05.2018) February 23, 2016
- [16] Karun Subramanian. Splunk vs ELK//free access by <http://karunsubramanian.com> ( date of request 10.05.2018) December 11, 2017
- [17] Feraga, Matthias.How to: choosing between lightweight and traditional ESBs. 6 Jun 2011
- [18] Technical guide for the CMDBuild system// free access by [www.cmdbuild.org](http://www.cmdbuild.org) (date of request 10.05.2018)
- [19] CMDBuild API docs or examples // free access by [www.cmdbuild.org](http://www.cmdbuild.org) (date of request 10.05.2018)
- [20] Zabbix API Documentation 3.0 // free access by <https://www.zabbix.com/documentation/3.0> (date of request 08.05.2018)

- [21] Nick Morpus 8 Free and Open Source Event Registration Software December 20th, 2017 free access by <https://blog.capterra.com> (date of request 08.05.2018)
- [22] Elasticsearch: Hardware Definitive Guide [2.x] free access by <https://www.elastic.co/guide/> (date of request 08.05.2018)
- [23] Fred de Villamil. Designing the Perfect Elasticsearch Cluster: the (almost) Definitive Guide. Free access by <https://thoughts.t37.net> (date of request 05.05.2018)
- [24] Technical description and prices of servers// free access by <https://modcom.kz/> (date of request 08.05.2018)
- [25] Абдимуратов Ж. С., Мананбаева С. Е. Безопасность жизнедеятельности. Методические указания для выполнения секции «Расчет промышленного освещения» для дипломных работ всех специальностей. - Алматы: АУЭС, 2009. - 20 с.
- [26] СНИП РК 2.04-05-2002 Естественное и искусственное освещение. Государственные стандарты в области архитектуры и градостроительства.
- [27] М.В. Айзенберг Методическое указание по осветительным приборам. - М. 1983.
- [28] Никитин В.Д. Расчет освещения точечным методом - Томск: Ред.. С. М. Кирова, 1985.
- [29] Г. М. Кнорринг, Методическое указание по планированию электрического освещения. - Л.: Энергия, 1976.
- [30] Samuelson P., Nordhaus W. Economics. - М.: Williams, 2014. - P. 55. - 1360 p.
- [31] Tax code of the Republic of Kazakhstan on 1st of January 2018
- [32] Tariff plans of AlmatyEnergoSupply in force from 1 January 2016// free access by <https://www.esalmaty.kz> (date of request 07.05.2018).

## Appendix A

### Listing of the Zabbix interaction module

```

18 class ZabbixReportv2:
19     HEADERS = {"content-type": "application/json"}
20
21     AUTH_Q = {
22     }
23
24     EVENT_GET_Q = {
25     }
26
27     TRIGGER_GET_Q_OLD = {
28     }
29
30     TRIGGER_GET_Q = {
31     }
32
33     def getLastIndexOfDictArray(self, list, key, value):
34         index = -1
35         for i, dic in enumerate(list):
36             if dic[key] == value:
37                 index = i
38         return index
39
40     def getTriggerDescr(self, authkey, objectid, priority=0):
41         triggerget_q = self.TRIGGER_GET_Q
42         triggerget_q['auth'] = authkey
43         triggerget_q['params']['triggerids'] = objectid
44         triggerget_q['params']['min_severity'] = priority
45         return self.jsonRequest(triggerget_q)
46
47     def __init__(self, URL, zb_login, zb_password, host, user, password, db, ignored_duration=10):
48         self.URL = URL
49         self.login = zb_login
50         self.password = zb_password
51         self.DB_HOST = host
52         self.DB_USER = user
53         self.DB_PASS = password
54         self.DB_NAME = db
55         self.ignored_duration = int(ignored_duration)
56         self.db = MySQLdb.connect(self.DB_HOST, self.DB_USER, self.DB_PASS, self.DB_NAME,
57                                   charset='utf8', use_unicode=True)
58
59     def get_equipment_type(self, hostname):
60         patterns = Constants.EQUIPMENT_TYPES.keys()
61         for pattern in patterns:
62             regex = re.compile(pattern)
63             if regex.match(hostname):
64                 return Constants.EQUIPMENT_TYPES[pattern]
65         return ["Неизвестно", "Неизвестно"]
66
67     def jsonRequest(self, jdata):
68         jdata['id'] = random.randint(1, 999)
69         data = json.dumps(jdata)
70         response = requests.post(self.URL, data, headers=self.HEADERS)
71         #print(response)
72         if response.status_code == 500:
73             resp = {'result': []}
74         else:
75             resp = response.json()
76             if resp['id'] != jdata['id']:
77                 # print 'Request and Query ID are not equal!', resp['id'], jdata['id']
78                 exit()
79         return resp['result']
80
81     def getAuthKey(self, user, password):
82         auth_q2 = self.AUTH_Q
83         auth_q2['params']['user'] = self.login
84         auth_q2['params']['password'] = self.password
85         return self.jsonRequest(auth_q2)

```



## Appendix A continuation

```
135 def getProblemTriggers(self, lastChangeTillTime = datetime.now(), priority=4):
136     TRIGGER_GET_Q = {
137         "jsonrpc": "2.0",
138         "method": "trigger.get",
139         "params": {
140             "monitored": 1,
141             #"output": ["description", "error", "priority", "lastchange", "comments"], #"extend",
142             #"selectHosts": ["host"],
143             'min_severity': None,
144             'lastChangeTill': None,
145             "selectGroups": ["name"],
146             "expandDescription": 1,
147             #"expandData": 1,
148             "filter": {
149                 "value": 1,
150             },
151             "sortfield": "hostname",
152             "sortorder": "DESC"
153         },
154         "auth": None,
155         "id": None
156     }
157     triggerget_q = TRIGGER_GET_Q
158     authkey = self.getAuthKey(self.login, self.password)
159     triggerget_q['auth'] = authkey
160     triggerget_q['params']['lastChangeTill'] = lastChangeTillTime.timestamp()
161     triggerget_q['params']['min_severity'] = priority
162
163     return self.jsonRequest(triggerget_q)
164
165
166 def getEvents(self, start_time, end_time, priority):
167
168     authkey = self.getAuthKey(self.login, self.password)
169     eventget_q = self.EVENT_GET_Q
170     eventget_q['auth'] = authkey
171
172     eventget_q['params']['time_from'] = start_time.timestamp()
173     eventget_q['params']['time_till'] = end_time.timestamp()
174
175
176
177     enriched_event_tmp = {
178         'start_time': None,
179         'end_time': None,
180         'region': None,
181         'site': None,
182         'description': None,
183         'objectid': None,
184         'priority': None,
185         'hostname': None,
186         'acknowledges': None
187     }
188     objectids = []
189     enriched_events = []
190     events = []
```

Figure A1 – Listing of the module for the interaction between Zabbix monitoring system and enterprise service bus

## Appendix B

### Listing of the CMDBuild interaction module

```
CMDBUILD.py x
12 class CMDBuildReport:
13
14     def __init__(self, host, user, password, db, cmdbhost, cmdbuser, cmdbpass):
15         self.logger = logging.getLogger(__name__)
16         #print(__name__)
17         self.DB_HOST = host
18         self.DB_USER = user
19         self.DB_PASS = password
20         self.DB_NAME = db
21         self.db = MySQLdb.connect(self.DB_HOST, self.DB_USER, self.DB_PASS, self.DB_NAME,
22                                   charset = 'utf8', use_unicode=True)
23
24         self.CMDB_HOST = cmdbhost
25         self.CMDB_USER = cmdbuser
26         self.CMDB_PASS = cmdbpass
27
28     def cmdb_auth(self):
29         cmdbuild_url = "http://" + self.CMDB_HOST + "/cmdbuild/services/rest/v2/sessions/"
30         data = {'username': self.CMDB_USER, 'password': self.CMDB_PASS}
31         headers = {'Content-type': 'application/json', 'Accept': '*/*'}
32         r = requests.post(cmdbuild_url, data=json.dumps(data), headers=headers)
33         r1 = r.json()
34         sessionid = r1["data"]["_id"]
35         return sessionid
36
37     def cmdb_get_users(self):
38         sessionid = self.cmdb_auth()
39         cmdbuild_url = "http://" + self.CMDB_HOST + "/cmdbuild/services/rest/v2/classes/Employees/cards"
40         headers = {'Content-type': 'application/json', 'Accept': '*/*', 'CMDBuild-Authorization': sessionid}
41         r = requests.get(cmdbuild_url, headers=headers)
42         return r.json()["data"]
43
44     def cmdb_get_severities(self):
45         sessionid = self.cmdb_auth()
46         cmdbuild_url = "http://" + self.CMDB_HOST + "/cmdbuild/services/rest/v2/lookup_types/Severity/values"
47         headers = {'Content-type': 'application/json', 'Accept': '*/*', 'CMDBuild-Authorization': sessionid}
48         r = requests.get(cmdbuild_url, headers=headers)
49         return r.json()["data"]
50
51     def cmdb_get_process_status(self):
52         sessionid = self.cmdb_auth()
53         cmdbuild_url = "http://" + self.CMDB_HOST + "/cmdbuild/services/rest/v2/lookup_types/IMProcessStatus/v"
54         headers = {'Content-type': 'application/json', 'Accept': '*/*', 'CMDBuild-Authorization': sessionid}
55         r = requests.get(cmdbuild_url, headers=headers)
56         return r.json()["data"]
57
58     def cmdb_get_departments(self):
59         sessionid = self.cmdb_auth()
60         cmdbuild_url = "http://" + self.CMDB_HOST + "/cmdbuild/services/rest/v2/classes/CompanyDivision/cards"
61         headers = {'Content-type': 'application/json', 'Accept': '*/*', 'CMDBuild-Authorization': sessionid}
62         r = requests.get(cmdbuild_url, headers=headers)
63         return r.json()["data"]
```

Figure B1 – Listing of the module for the interaction between CMDBuild system and enterprise service bus