Contemporary state of the confidential information leakages through the cloud data repositories

A V Kalach\textsuperscript{1,2}, E A Timofeeva\textsuperscript{3}, D A Bakhteeva\textsuperscript{4}, D D Dayneko\textsuperscript{4} and VA Spirin\textsuperscript{1}

\textsuperscript{1}Voronezh institute of the Russian Federal Penitentiary Service, 1a Irkutskaya, Voronezh, Russia
\textsuperscript{2}Voronezh State Technical University, 14 Moskovsky pr-t, Voronezh, Russia
\textsuperscript{3}Research Institute of the Federal Penitentiary Service of Russia, 15 Narvskaya, Moscow, Russia

E-mail: AVKalach@gmail.com

Abstract. Information concerning the leakage of the confidential data held in the cloud repositories of Amazon, Mongo DB, in the file hosting of Google Drive kind, as well as cloud servers for back-up in the period of 2016-2018. Certain conclusions made concerning necessity for the development of the systems enabling information security with the use of the cloud technologies at the expense of cyber-resilience enhancement.

1. Introduction

More and more significance at the information resources market is pertains to the cloud technologies. State organizations and commercial companies more and more feel usability of the cloud services in the virtual media and they are ready to transfer their data bases there. Along with the development as of the cloud technologies as the information medium as a whole the cases of inadvertent data compromising happened occasionally. But at present the threat of data loss becomes more tangible while the number of incidents is registered more and more often. Majority of the leakages of the confidential information from the open repositories is a consequence of the incorrect grant access to the users or invalid configuration management of the systems [1 – 8].

Specialists of the analytical center InfoWatch studied the cases of leakages of the corporative data bases through the cloud services and other file repositories. The subject of investigations was various public reports on the leakages of the confidential data through the cloud repositories (Amazon, Mongo DB, etc.), folders in the file hosting (Google Drive, etc.), as well as the back-up cloud servers for the period of 2016-2018. In our work no incidents were considered connected with the information leakage through web-servers and post-servers. The emphasis was placed for the problems of the personal and other data leakage from the objects of data virtualization (cloud repositories and services) [9].

2. Confidential information leakage through the cloud data repositories

Basing on the public reports in media and other open information resources InfoWatch analytical center recorded an increase of the confidential data through the cloud servers and repositories with a Network access in 2018 by almost 1.5 times as compared with the previous year. This trend is already kept for several years. For example, the leakage increased by 4.4 times as compared with 2016 year [9].

The largest number of the compromised records was fixed in 2017 and it was a record. More than
1.7 billion of records was lost from the unlocked servers (and this is about 13% of all the data lost in that year). Most of the lost data owned to the company dealing with network marketing – River City Media. As a result of incorrect back-up an error occurred and thus a great data base with a level of 1,34 billion records was compromised. In 2018 data losses from the servers with open access were about 1.3 billion of records, and in the major incident it was lost 400 million of the records. Note, that about 40% of the data leakage occurred at high-technological companies.

For the period of 2017 - 2018, the largest share of the leakages connected with the incorrect maintenance and management of the cloud servers as well as the other errors during the work concerning data virtualization appropriate to the high-technology IT-segment – manufacturers of IT-production, IT-services, social networks and so on [9].

Highly developed companies owing a good technological potential are prone to the use of the visualization instruments being in trend, in particular, the use of external data repositories. But, unfortunately, the staff of these companies does not take into account some nuances in the work with the cloud services and consequently, the cases of the data leakage become more often.

Thus, as a result about 90% of all the data that were lost from the unprotected servers in 2018 occurred just in high-tech industry (table 1).

**Table 1.** Sectoral distribution of the leakages from the open cloud information resources in 2017-2018.

| Sectors of leakages                     | 2017 г. | 2018 г. |
|----------------------------------------|---------|---------|
| State organizations                    | 14,3%   | 8,5%    |
| Medicine                               | 8,2%    | 11,4%   |
| High technologies                      | 32,6%   | 40%     |
| Education                              | 4,1%    | 7,2%    |
| Retail                                 | 10,2%   | 7,2%    |
| Industry and transport                 | 12,2%   | 7,2%    |
| Banks, finances and insurance          | 12,2%   | 1,4%    |
| Other                                  | 6,2%    | 17,1%   |

During the same period of the year, servers of the medical and educational institutions also became the «victims» – the number of leakages considerably increased. But at the same time the loss of data from the financial sector, industry and state sector drops rather noticeably. Personal data leak in four out of five cases. Majority of the leakage, namely more than 80%, is personal data. 9.2% of the cases related to the data compromising from the open repositories is the leakage of payment information as well as the commercial secrets and industrial know-how (table 2) [9].

**Table 2.** Distribution of the leakages in accordance to the data types for the period of 2017-2018.

| Type of the data     | 2017   | 2018   |
|----------------------|--------|--------|
| Personal data        | 77,8%  | 81,6%  |
| Commercial secrets   | 8,9%   | 9,2%   |
| Payment information  | 6,7%   | 9,2%   |
| State secret         | 6,7%   | -      |

A quarter of the leakages happens from the servers of Amazon S3. As for the distribution of the compromised servers by the types of data some considerable changes occurred. The leader in data compromising in 2018 just as in the previous years remained the cloud repository Amazon S3. It accounted for more than a quarter of leaks (table 3).

A tendency to the leakage of the confidential information in 2018 was observed for the server MongoDB, as well as such platforms as Elasticsearch and Apache, file hosting Google Drive. By-turn, such services as Rsync and GitHub reduces the number of cases connected with data leakage under the information back-up by 3 and 7 times, respectively. In every second case the victim was the US
Company.

Table 3. Data on the compromised repositories, 2017-2018.

| Repositories     | 2017  | 2018  |
|------------------|-------|-------|
| Amazon S3        | 28.6% | 25.7% |
| Apache           | 2%    | 4.3%  |
| FTP              | 2%    | 1.4%  |
| GitHub           | 10.2% | 1.4%  |
| Microsoft SQL    | 8.2%  | 1.4%  |
| Mongo DB         | 6.1%  | 15.7% |
| rsync            | 8.2%  | 2.8%  |
| Other            | 34.7% | 34.3% |
| Google Drive     | -     | 5.7%  |
| Elasticsearch    | -     | 7.1%  |

As for the distribution of the leakage by the countries, here strong changes are observed. In 2018, just as before, the leadership remains for the USA, but the data leakage reduced from 75.5% to 47.1%. But in Canada the number of leakages increased almost by 3 times, while in India – by 2 times (table 4) [9].

Table 4. Leakage distribution by the countries in 2017-2018.

| Countries       | 2017  | 2018  |
|-----------------|-------|-------|
| Canada          | 2.0%  | 5.9%  |
| China           | 2.0%  | 1.5%  |
| Great Britain   | 6.1%  | 4.4%  |
| India           | 4.1%  | 8.8%  |
| Sweden          | 2.0%  | -     |
| USA             | 75.5% | 47.1% |
| Other           | 8.2%  | 19.1% |
| Finland         | -     | 2.9%  |
| France          | -     | 2.9%  |
| Russia          | -     | 2.9%  |
| Brazil          | -     | 4.4%  |

Data leakage from the unprotected cloud data repositories is rampant all over the world. A number of the large Companies such as American Express, Honda, Nokia, Sky Brazil, and several state structures became the victims of such incidents. Potential source of the data leakage is every repository where thousands and even tens thousands of the errors in configuration are detected.

Only one error in the work of corporations can become a reason of data leakage from the unprotected server. A lot of companies loose large data bases since cybercriminals perform monitoring of the cloud resources.

Analysis of the data on the leakage of confidential information from the cloud data repositories enables to make a conclusion on the appropriateness of increase of their cyber-resilience. Cyber-resilience is a particular case of the one-sided information conflict in cyber-space when the attacking part uses different strategies in order to gain its purpose while the defending party applies the strategies providing stable performance of the system for the control of the objects protecting from the attacking impacts. One should note that a specific feature of the cyber contention is that at least two (or more) control subsystems each other tend to expand their control impact by the shared common resource (of the global communication space) [10]. As a result, it is possible to consider hypothesis on the behavior of the information systems in the conditions of the information conflict that describes a set of the possible attacking and protecting strategies as one of the starting points. At the stage of formalization of
the cyber-resilience process a system of hypotheses generates corresponding variety of the models and algorithms of cyber-resilience.

Conceptual models of cyber-resilience for the objects of protection in the conditions of the informational conflict register a set of the possible a priori and a posteriori knowledge on the strategies of the opposing parties and act as their verbal (weakly-formalized) models (figure 1).

![Figure 1. Graphic model for the representation of multivariate state of the investigated object in the conditions of impact.](image)

Formally, the model of the problem situation can be defined with the help of the following elements representing sets of the most significant factors from the viewpoint of the person making decision (PMD) [10]:

- \( N = \{1, 2, \ldots, I_0\} \) – a set of the participants in the conflict;
- \( S \subseteq N \) – subset of the participants in the conflict joined by the common purpose and interacting with each other (coalition);
- \( R \) – hierarchy relation at the coalitions;
- \( U^S \) – a set of strategies of \( S \)-th coalition;
- \( \Theta^S \) – information that is disposed by \( S \)-th coalition;
- \( X^S \) – Cartesian product of the sets of coalition strategies – a set of situations;
- \( W^S(X^S) \) – gain function of \( S \)-th coalition;
- \( P^S \) – model representing the preferences of \( S \)-th coalition at the set of situations in the model of problem situation.

For each of the coalition at a set of situations generally it is necessary to define the binary preference relation. The purpose of coalition in the conflict is in any case the achievement of the most preferable situation. The situation is a result of choice of their own strategies by all of the participants (players). In
general, the gain of i-th player does not coincide with the gain $W^{(i)}(\cdot)$, that he can provide for himself, acting alone, since participating in coalition $S$ he can get more interest. This can be explained by the fact that:

$$W^S(\cdot) \geq \sum_{i \in S} W^{(i)}(\cdot).$$

Thus, the most general model of the conflict is defined by a system of sets:

$$\{N, \{U^S, \theta^S, P^S, W^S\}_{S \subseteq N}, R\},$$

Mechanism $H$ of the conflict’s model is that participants of the conflict from a set of $N$ have an effect on some system, and as a result they get certain gains:

$$H : U^{[1]} \times U^{[2]} \times \cdots \times U^{[k]} \times T \rightarrow (W^{[1]}, \ldots, W^{[k]}),$$

where $T$ – is a set of points of time in the development of the conflict. Each of the participants of the conflict acts in accordance with certain rules trying to achieve his purpose. It is assumed that all of the conflict’s participants get some information $\theta^{(i)}$ on the state of a system. In most cases information $\theta^{(i)}$ is concerned with the kind of functions $W^{(i)}$ for the gain of separate players or $W^S$ of their coalitions as well as the set of admissible strategies for participants and coalitions of the conflict.

3. Conclusions

Thereby, along with increase of the skill level of system administrators and users it is required to perform the revisions of the information resources, to employ the tools for the control of access, to perform monitoring of unprotected cloud information repositories with a limited access and actively apply the current models of cyber-resilience.

References

[1] Glotina I M 2015 Latent nature of economic security threats Proceedings of the III International Scientific and Practical Conference 51-6
[2] Shpeko M V 2014 Data Encryption in Cloud Services Materials of the All-Russian Youth Scientific and Practical School (Kemerovo: Kuzbass State Technical University named after Gorbachev T F) 281-2
[3] Baturina M V 2018 The future behind cyber insurance Insurance case 7 28-33
[4] Prigodin S A 2019 Legal framework for cybersecurity in the Russian Federation and development trends Compilation of scientific papers on the VII International Scientific and Practical Conference 36-41
[5] Bobrysheva G V and Pudovkina E M 2017 Cloudy safety Compilation of articles of the XVII International Scientific and Technical Conference (Penza: Autonomous non-profit scientific and methodological organization Volga House of Knowledge) 99-104
[6] Kalatch A V 2019 Systems analysis and assessment of current threats to information security Journal of the Voronezh Institute of FPSR of Russia 1 69-74
[7] Grebennikov N 2017 Cyber threats today: warned - means armed First mile (Moscow: Advertising and publishing center "Technosphere") 4(65) 76-8
[8] Poltavtseva M A 2019 A Consistent Approach to Building Secure Big Data Processing and Storage Problems of information security. Computer systems 2 29-44
[9] Digests and reviews of Analytical center InfoWatch Retrieved from: https://www.infowatch.ru/
[10] Komarovich V F and Saenko I B 2002 Computer Information Wars: Concept and Realities Information security. Confidant 4-5 84-8