Ramin B. Samadov
Baku State University, Baku, Azerbaijan
ramin.samedov@gmail.com

DOI: 10.25045/jpit.v08.i2.08

ALGORITHM OF CREATING BACKUP OF A DISTRIBUTED DATABASE IN CLOUD STORAGE

The paper presents the modern technologies in storage and backup systems. Traditional methods of encryption are examined. The shortcomings of traditional backup methods are shown. A backup copy for creating algorithm of a distributed database in cloud storage is developed. The results of the experiment are presented.

Keywords: *cloud computing, distributed databases, backup, encryption.*

Introduction

The information stored in IT systems is at daily risk of hacking and various attacks. Information can be lost for various reasons, such as software errors (software), user errors while working with the system, hardware and media malfunctions, and malicious attacks on information. Protection from all threats by one solution does not exist, it requires constant monitoring and constant work as the risk of losing valuable information always exists [1].

Global statistics show that the leading causes of data loss are the failure of physical and hardware (44%), system user errors (32%), more critical errors of administrators; with the highest level of access to storage systems in organizations, 14% of data loss occur due to software failures, 7% of cases are the result of virus attacks on computer systems, and the smallest percentage (3%) is due to natural disasters.

Failures are the causes of the termination of the functionality and loss of valuable data, thereby questioning the existence and operation of the entire system as a whole. The only way to store valuable information safely is to make backup copies from time to time.

In the process of a data backup system integration, many organizations face a complex set of challenges in assessing their current needs and essential data volumes in future, selecting a system that will meet all security, speed and reliability requirements, and the potential for subsequent scaling and many other conditions. The determination of the optimal solution, especially considering the wide variety of existing schemes for implementing the integration of storage and backup systems, as well as the fairly large dynamics of price changes and the emergence of new technologies in the IT market, became a difficult task.

About data backup

A backup is a duplicate instance of existing electronic information stored in a backup location. In case of a failure of the main system or if it is essential, a backup must be used to restore the system entirely [2].

Backup - the process of creating backups of systems. Archiving is a subset of many types of backups. The demand for archiving is not always available. More often, the need for this appears in the organization when it grows, and this applies not only to regular files, but also to financial databases and data, mail databases, i.e. those documents that are usually stored in archives (letters, orders, accounting documents, etc.) in the paper version.

The frequency of archiving is different. Organizations can perform it once a year and store data on external media somewhere in a room or in a banking cell. The archiving process is very simple, and the main thing is not only to perform archiving, but also to periodically check the state of the archive after it is backed up. For instance, it is necessary not only to copy new data, but also to restore at least some of them from old archives, while not forgetting to monitor the media so that it can be always read the archive made earlier. There may be a situation when it becomes necessary to use an archive copy of data that was performed more than five years ago, and the equipment, in which the data stored, was out of order broken down and has not been released for a long time. Obviously, this situation should be avoided, keeping the archives constantly in an actual and working condition [3].

Redundancy of systems is the need for back-up of not separate files but the entire system, which can consist of several components, for example, a distributed system consisting of special software, a database and file data. It is better to restore systems entirely, rather than in parts. First of all a specific software for backup is selected, of course, the price, functionality and convenience are taken into account, which play significant role. Moreover, backup of the system is essential so that it can guarantee to be restored, even in the case of a complete server failure.

The main purpose of the backup is to keep the latest backups for quick recovery in case of any need. If archives are stored for as many years as an organization exists (sometimes even longer), then it is necessary to introduce the concept of "depth of storage" for backups, i.e. the time after which the backup is obsolete, and the data on them must be overwritten; that is, a fresh copy of the data to be made. Backup copies are divided into two parts: backup copies of the information system with the entire structure and directly only the system data. Of course, backup copies of the information system with the entire structure are a very important part of the security policy. For example, organizations need to backup the mail server, Oracle databases, CRM systems.

The repository is the place where the latest and up-to-date data backups are stored and maintained.

Distributed databases

Distributed processing systems, or distributed systems (DS) that operate in computer networks, are among the most promising and rapidly developing areas of computer science. They took such a place because of their significant advantages compared to isolated systems operating on the basis of individual computers. The most successful and often used type of DS is a distributed database (DDB), which is the integration of stand-alone local databases, geographically distributed and connected through a computer network. All local databases are assumed to be initially comprehensive and coherent.

The nodes interact with each other by messaging. The means of interaction between the user and the DDB are transactions. A transaction is a sequence of operations (sub-transactions) on the DDB that redirects it from one consistent state into another consistent state. Each sub-order before starting its work must enforce a resource in each node. Two transactions conflict only if they work with the same shared resource and at least one of the operations they perform is a record. The order of execution of actions of two transactions is significant only in the case when they conflict. The execution of each individual transaction preserves the integrity of the DDB. Consequently, several sequentially executed transactions, at the same time, preserve the integrity of the DDB.

A distributed database backup creation

In order to create a backup of a distributed database, it is necessary to consider the status in which the reservation is wanted to be made. There are two methods for creating a backup for a distributed database [4]:

The first method is the creation of a backup of the distributed database at the time the database is extinguished. But, the big shortcoming of this method is that the database must be completely turned off. This method is not suitable for systems operating in 24x7 mode;

The second method is to create a backup copy of the distributed database at its working state. Creating a backup copy of the database takes a considerable amount of time, and the data itself is constantly changing. Therefore, the process of creating a consistent database backup is getting more complex. For these purposes, the RMAN utility for Oracle databases is usually used.

The problem of data hacking, backup encryption

While working with confidential data, there is always the risk of losing this data. In the case of the critical data loss, the very existence of the organization may be at risk. To prevent an unauthorized access to data, it is recommended to encrypt the data. It is also proposed to encrypt each backup copy of the system. For encryption, you can use the PGP utility. PGP is software with the library of procedures and functions, through which it is possible to perform encryption of messages, files and digital signatures. At the same time, it is possible to encrypt any information in electronic form, including transparent encryption of data on different storage devices, for example, hard disks. The PGP client creates a key pair: public and private keys. When creating a key pair, their owner is indicated with the name and email address of the mail, the type of the key, its length and the key completion time. Concurrently, digital signatures are encrypted and verified with a public key. A private key is required to decrypt and create a digital signature. Using PGP encryption, it is possible to use hashing, data compression, symmetric key encryption, and finally, public key encryption. Each step can be implemented using any of several supported algorithms. Symmetric encryption can be performed using one of the seven symmetric algorithms on a session key (AES, CAST5, 3DES, IDEA, Twofish, Blowfish, Camellia). Cryptographic persistent pseudo-random number generators are required to be used in order to generate a session key. The session key is eventually encrypted with the public key of the recipient using the algorithms RSA or Elgamal, depending on the type of key that will be used by the recipient. All public keys must match the username or email address specified when creating the key [5].

When working with confidential data there is always the risk of losing this data. In the event of the loss of critical data, the very existence of the organization may be in question. To prevent unauthorized access to data, it is recommended to encrypt the data. It is also proposed to encrypt each backup copy of the system. For encryption, the PGP utility can be used.

Problems of a backup copy of the DDB creation

The rapid increase in the amount of stored data can be observed every day, along with this, the complexity of their protection increases. All data are stored in a conventional data store. A typical data store has the risk of losing data by crashing physical cumulative disks or because of errors made by administrators serving the storage system. However, in case of loss of a disk with data backup copy, there is a risk of recovering this data from an attacker. Organizations often do not have a complete copy of the storage system. Once the data loss occurs, all data are lost.

To solve the abovementioned problem, an algorithm has been developed, by the help of which the data is backed up to the cloud data storage. The Google Drive is used as a cloud storage. According to the functionality descriptions, the Google Drive cloud data storage has a large staff of professional administrators who monitor the operation of the storage system. The resilience factor of the entire cloud storage system is much higher than the local storage system.

The fault tolerance of the entire system increases, while using cloud storage. Moreover, the issue of confidentiality of information is in question. Data encryption must be used for solving this issue before sending the backup to the cloud data store. Thus, in case of data loss in the cloud data warehouse, attackers will not be able to use them as the readability of the encrypted data is impossible.

Algorithm for creating the backup copy of distributed database

Step 1.The unit of time is denoted by the symbol “B”. “B” will take values from 0 to “P”.

Step 2. The distributed database is denoted by symbol “C”. At each moment of time “B”, the state of the DDB will constantly change.

Step 3. It is required to create a backup copy of “C”. For this purpose, each new backup copy of “C” will be sent to the repository of the backup copies “C”. The repository is denoted by “P”, and in the unit of time “B” from 0 to “P” backup copies of the DDB will go into the repository.

Step 4. Using encryption libraries, we will create public and private encryption keys. The private encryption key must be transferred to other systems that must decrypt the data encrypted by this private key.

Step 5. The backup copies "C", received in step 3, will be encrypted with the key created in step 4. While encrypting the encryption library, it is necessary to enter the key phrase, by which the files will be encrypted.

Step 6. Encrypted files received after the Step 5 will be denoted by status 1, and unencrypted files will be denoted by status 0.

Step 7. All the files from the repository “P” with the status 1 will be transferred to the clouds. The clouds will be denoted by “0”, and all backups with status 1 will be sent to the cloud in the unit of time from 0 to “P”.

The Fig.1 graphically shows the work of the algorithm for creating a backup copy of the DDU in cloud technologies.

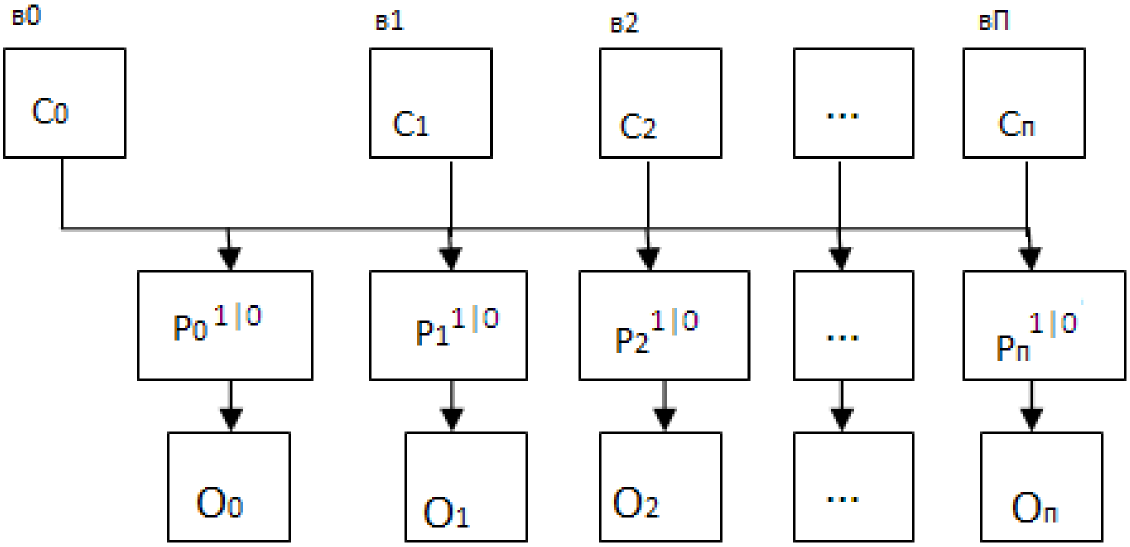

Fig.1. The scheme of the algorithm for creating a backup copy of an DDB in cloud storage

Description of technologies and instructions for the experiment of the algorithm

Following software tools will be used for the experiment on the basis of the algorithm for the creation of a backup copy of the DDB in cloud storage: Oracle DDB version 11g, Windows 7 operating system, Kleopatra PGP encryption library and Google Drive cloud technologies.

For a backup copy of the database creation and sending the created backup to the repository, the Oracle RMAN Recovery Manager utility must used, namely:

```
RMAN> backup database plus archive log.
```

Files that are a backup copy of the running Oracle database will be created based on the results of the abovementioned command. Subsequently, the encrypting of the received DDB backup copies must be enabled by the use of the public key of the encryption library. With the purpose of doing this, the appropriate files must be selected and a key phrase to be entered in the Kleopatra software. After selecting the encryption method and starting the process, the files are encrypted using a key and a passphrase. The received encrypted files are sent to the Google Drive cloud technologies.

Results of experiments

It is essential to use several databases with different data volumes for carrying out experiments. The first experiment requires a database with a data volume of 12G; for the second experiment - a database with a data volume of 29G; the third experiment - a database with a data volume of 41G. For all three DDBs, backups are created in two ways.

The first method is a way to create a backup by copying backup data to local storage disks, and the second method is based on the algorithm described in this article. Table 1 presents the findings of three experiments.

Table 1.

The results of the experiments of the algorithm for creating a backup copy of DDB in cloud storage

Duration (seconds)	Experiment number	Operation type
196	1	Creating a backup copy on local disks
196	1.a	Creating a backup copy in repository
42	1.b	Encryption of a backup
135	1.c	Sending the encrypted data to the cloud
272	2	Creating a backup copy on local disks
272	2.a	Creating a backup copy in repository
71	2.b	Encryption of a backup
220	2.c	Sending the encrypted data to the cloud
375	3	Creating a backup copy on local disks
375	3.a	Creating a backup copy in repository
123	3.b	Encryption of a backup
310	3.c	Sending the encrypted data to the Cloud

As can be seen from the Table 1, more time is required to create a backup copy of the DDB in cloud storage. But, due to the use of cloud storage, the fault tolerance of the backup database storage system is increasing. In case of failure or fire of the entire server infrastructure, there will be no loss of backup copies of the DDB. Because of this, it will be possible to restore the working capacity of the whole system.

Conclusion

Thus, the creation of a backup copy of the DDB in cloud storage increases the resilience of the entire system due to more secure storage of backup copies. The algorithm is developed, the scheme of its operation is shown, and the practical application of the algorithm is described in the paper.

References

1. Kazakov V.G., Fedosin S.A. Technologies and algorithms of backup copy. Russian competitive selection of review-analytical articles on the priority area “Information and Telecommunication Systems”, 2008, 49 p. Www.ict.edu.ru/ft/005653/62330e1-st17.pdf
2. Jansen W. A. Cloud hooks: Security and privacy issues in cloud computing / Proceedings of the 44th Hawaii International Conference on System Sciences, 44th Hawaii International Conference System Sciences (HICSS), January 04–07, 2011, pp.1–10. doi>10.1109/HICSS.2011.103
3. Rahumed A. A secure cloud backup system with assured deletion and version control / Proceedings of the 40th International Conference Parallel Processing Workshops (ICPPW), 13-16 Sept. 2011,Taipei City, Taiwan, pp.160–167. DOI: 10.1109/ICPPW.2011.17
4. Li Q., Xu H. Research on the backup mechanism of oracle database / Proceedings of the International Conference Environmental Science and Information Application Technology (ESIAT), 4–5 July 2009, pp.423–426.DOI: 10,1109 / ESIAT.2009.294
5. Zimmermann P. R. The official PGP user's guide, MIT press, 1995, 21 p.