Data Backups in the Clouds

Aljaž Zrnec

University of Ljubljana, Faculty of Computer and Information Science, Trzaska 25, 1000 Ljubljana, Slovenia
E-mail: aljaz.zrnec@fri.uni-lj.si

Abstract. In the paper we present a concept of making backups in the cloud. We discuss the current practice of making backup copies of data enabling backups to be stored in a separate location and analyze the possibility of making backups in the cloud. We mainly focus on the economic and performance aspects of using cloud computing for making backup copies of data.

1 INTRODUCTION

Cloud computing [1] allows users to tap into a virtually unlimited pool of computational resources and data storage capacities via Internet. Compared to the traditional computer conception with users having full control over computing resources, cloud-computing users have little or no access to and control over the cloud-computing infrastructure. So they have to interact with computational and data resources of the cloud using appropriate APIs (API - Application Programming Interface) which have to be provided by a cloud-computing provider [2]. Despite these restrictions, cloud users receive several significant benefits such as ease of scalability, reliability, independent and dynamic adaptation of the necessary resources and paying only for the resources actually used.

The advantages offered by cloud computing have enabled many service providers to offer different kinds of services over the Web. One of them which is especially interesting is the possibility of making backup copies of data in the cloud. Therefore, in Section 2 we describe the differences between the traditional backup and the data backup in the cloud. In Section 3 we give an example of making backup in the Oracle database, first without and then by using the cloud. We also compare the cost and the speed of both cases. In conclusion, we present our findings and plans for our further work.

2 TRADITIONAL BACKUP AND BACKUP IN THE CLOUD

2.1 Traditional tape-based backup of data

Good practices in the field of database disaster recovery require backup storage of business critical data on a remote location, outside the business system. A business system usually provides writing backups on magnetic tapes and sending them to a remote location which is an expensive and complex process requiring special hardware, properly trained staff and procedures (regulations) ensuring backups to be regularly produced, and protected and the information contained in them to be obtained and used in case of an accident. Besides being today customary for business systems to use outsourcing for transporting and protecting their backup data, they still have to take care for their data integrity in their backups and above procedures.

2.2 Cloud backup

As an alternative to the today's traditional backup of data, there are many providers of cloud services becoming available, enabling data backup in the clouds.

Backup in the cloud, or the so-called online backup of data, is a way of making backups, where data from a particular database are sent via a public or private network to a data server located at a remote location. The data server is managed by a provider of cloud services, charging the customer for using the data storage service based on the required disk space, bandwidth or number of users of this service.

The system for backing up data in the cloud is based on a special application located at the user of the cloud service. The application is launched at a frequency (daily, weekly, etc.) defined in the contract on the use of the service for backing up data in the cloud (SLA - Service Level Agreement). For instance, if a user (customer) has a contract for making daily backups, then the application gathers, compresses, encrypts and sends data to a data-service provider server every 24 hours. To reduce the bandwidth used for data transfer between the user and the cloud-service provider, we may use an incremental backup. According to the contract on the use of the cloud service (SLA), the incremental backup is performed in intervals only for changing data in the original database. Since the data are transferred via the Internet, the bandwidth is usually

relatively very limited. Besides, cloud-service providers may also limit the bandwidth to prevent individual users from a disproportionate use of resources in the cloud.

Based on our analysis of the cloud-service usage performed with the Oracle database and Amazon service for backing up data in the cloud - Amazon Simple Storage Service (hereinafter referred to as Amazon S3) [3], we found that the Amazon S3 service limits the data throughput in the context of individual sessions at 2.5 to 3.5 Mb/s.

2.3 Advantages of the cloud backup

The main advantages of sending backup data over the Internet to the cloud are: flexibility of the cloud with regard to the performance needs, large amount of the available storage space, and the costs being accounted only for an actual use of resources. Also, the use of the cloud significantly simplifies the user's own computational infrastructure as there is no longer any need for own storage management (e.g. working with tapes, sending tapes to a remote location, etc.) and special hardware for making backups on tapes. Section 3.2 presents the advantages offered by the Oracle database for making backup copies of data in the cloud.

An important issue of data transferring to or from the cloud is the limited bandwidth of the Internet which prevents fast transfers of a large amount of data (the problem of making a full backup). The Amazon-service provider has solved this issue by offering a special service allowing movement of a full backup to or from the cloud and transportation with a portable hard drive. For instance after a disaster in a local operational database, the Amazon S3 service provider sends us a full backup of data on a portable hard drive by using express mail. In this way, the data storage in the cloud is comparable to the conventional data backup, especially when a remote location for storing backups is part of a business strategy that includes both local backups and backups at a remote location.

3 BACKUP IN ORACLE RDBMS

3.1 Test environment

In our analysis we used the Oracle database 11gR2 installed on a computer server with a 2.66 GHz Xeon processor, hard drive with 10.000 rotations per minute, 8 GB RAM and MS Windows Server 2008 x64 operating system. To back up data in the cloud, we selected the Amazon S3 service from the Amazon-service provider.

Amazon S3 is an essential Amazon service for storing data in the cloud. The service provides a simple web interface to store and transfer any amount of data to or from the Amazon cloud. The Amazon S3 main advantages are scalability which means that it acts independently of the number of users, reliability and

speed. The price of the service usage which is defined in SLA is also very important for being based exclusively on the cloud resources that are actually used. The Amazon S3 service can be used for both classical textual and numeric data storage as well as for serving a multimedia content in real time.

The Oracle Cloud Backup Module (hereinafter referred to as CBM) enables an Oracle database to be connected to the Amazon S3 service and to send its backups to the cloud. The module is compatible with all the Oracle database versions 9iR2 and above. It requires a network connection to the Internet for its operation. CBM is a member of the Oracle Secure Backup tools used creating backup copies of data on traditional tapes or in the cloud. CBM can also be used when the database is running within the Amazon Elastic Cloud (hereinafter referred to as Amazon EC2) where the database is located on a virtual machine (server) inside the cloud [4]. In this case, the CBM benefits from the higher internal-network bandwidth and the lower cost of use by eliminating the costs for transferring data into and out of Amazon S3. CBM is implemented using the Oracle Recovery Manager (hereinafter referred to as ORM) enabling an easy integration with external libraries for making backup copies of data. In this way, database administrators can continue to use the existing tools for making backups. The above described modules and backup processes in the cloud are shown in Fig. 1.
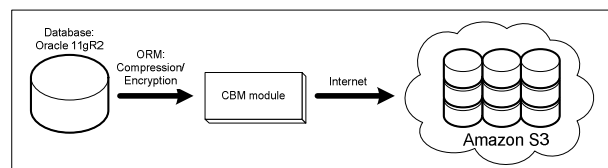

Figure 1. Oracle database backup in the cloud

3.2 Advantages of using Oracle RDBMS for Cloud backup

The Oracle cloud-backup functionality provides many advantages over the traditional tape-based offsite backups:

- *Continuous accessibility*: Backups stored in the cloud are always accessible, much in the same way as the local disk backups are. As such, in case of disaster, there is no need to call anyone and to ship or load tapes before a restore can be performed. Instead of this, administrators can initiate restore operations using their standard tools for disaster recovery (Enterprise Manager, script, etc.), just as if the offsite backup were stored locally. In this way, restore activities can be performed faster and down time can be reduced from several days to a few hours or even just a few minutes. In case of large databases, when shipping a portable disk from the Amazon cloud service-provider, a restore takes no

longer than it would take to have a tape recalled from the remote location.

- *High level of reliability*: As the storage clouds are disk-based, they are inherently more reliable than tapes. Moreover, cloud vendors typically keep multiple redundant copies of data to offer a higher level of availability and scalability.
- *Unlimited increase in space and no upfront capital expenditure*: The cloud provides a virtually unlimited capacity to store data with no upfront capital expenditure. This means that the cloud dynamically adjusts the size of the storage space to hold the required backup data and users pay only for the space actually used.
- *Reduced usage of backup tapes and reduced offsite storage costs*: The clouds reduce or even eliminate the need for the tape-based backup. This can lead to significant savings in buying the tape-backup hardware and software and also savings in tape storage costs at remote locations.
- *Easy provisioning of the test and development environments*: Cloud backups are accessible from anywhere via the Internet. They can be used to quickly clone databases to create custom test or development environments. For instance, a cloud backup stored in Amazon S3 can be cloned to a virtual machine (virtual server) running in Amazon EC2 by using a simple script.

3.3   Data-security assurance

In shared, publicly accessible environments, such as the storage cloud, data security and privacy are particularly important. Therefore, when sending data to the cloud, the Oracle CBM module uses a special functionality of the ORM component, which uses an encryption to assure data security and privacy. In this way, data are protected twice. One level of protection against an unauthorized access is already ensured by the cloud-service provider, the second level is guaranteed by the above-mentioned encryption of the backup data before being sent into the cloud. This reduces the risk of theft or an unauthorized access to the data during transport as well as when storing data in the cloud.

3.4   Data compression

Since the CBM module integrates into Oracle RDBMS, it can independently identify and skip unused space (blocks) in the database before the backup is made and sent into the cloud. At the same time, the ORM component offers many possibilities for data compression directly impacting the speed of making a backup copy. Relatively slow Internet connections have the greatest impact on the speed of the cloud-backup creation. Thus, in our analysis of the impact of data compression on the speed of the data backup creation

we compared making a backup copy of compressed and uncompressed data.

3.5   Cloud-backup performance

As already mentioned in Section 2.2, Amazon S3 may throttle the throughput of an individual session from 2.5 Mb/s to 3.5 Mb/s to prevent individual users from consuming disproportional amounts of cloud resources. However, it is possible to use a correct combination of parallelism and data compression for the backup to attain a data throughput of 43 Mb/s to 55 Mb/s, which strongly impacts the backup-creation speed.

To examine the performance of backup data in the cloud with the Oracle database, we made several measurements in which we determined the time needed for creation of a backup. Firstly, we measured the duration of the backup creation in the database that was located on our test server, and secondly on a virtual server located in Amazon EC2 [5]. For each database we observed the impact of data compression on time taken for the backup creation. Our measurements were performed for both the complete and incremental backups. The size of a full backup was 250 GB and the size of an incremental backup was 10% of the data changes made to the database. The results of the measurements are summarized in Table 1.

Database location	Data throughput		Full backup time		Incremental backup time	
	Compression		Compression		Compression	
	No	Yes	No	Yes	No	Yes
Test server	10 MB/s	43 MB/s	< 6 h	> 2 h	< 1 h	> 30 min
Virtual server in Amazon EC2	35 MB/s	55 MB/s	< 2 h	> 1 h	< 20 min	> 10 min

Table 1. Oracle cloud-backup performance

Some of the observations drawn from these results are:
- The time for a full or incremental backup of the test database (without using data compression) located on a local test server is three times longer than that of the database located on a virtual server in Amazon EC2.
- The time to back up the test database located on a test server (using data compression) is twice (full backup) or thrice the time (incremental backup) of the database located on a virtual server in Amazon EC2.
- The time for a full backup of the test database located on a local test server is three times longer when no compression is used and in case of an incremental backup, it is two times longer when no compression is used.
- The time for a full or incremental backup of the test database located on a virtual server in Amazon EC2 is two times longer when no compression is used.

It can be concluded that both the throughput of the network (in this case Internet) and the level of data compression have a great impact on the speed of making backups in the cloud. Also to be noted is that the Oracle database 11g versions and the above use of advanced compression mechanisms are significantly faster and more efficient in terms of CPU overhead than the pre-11g compression.

The backup speed is also accelerated by the ORM module thus enabling us to use multiple parallel transmission channels to fully utilize the network. The highest performance in our tests was achieved with 64 simultaneous channels. The Oracle database 11g allows multiple channels to back up a single file in parallel, meaning that the parallelism is increased beyond the number of the data files to be backed up.

3.6 Cloud-backup cost assessment

Our assessment of the cost of backing-up data in the cloud [6] was based on the cost of using the Amazon cloud services (Amazon S3 and Amazon EC2). The cost of the Amazon S3 service includes the price of a 325 GB cloud storage (see description of the cloud backup scenario below) and the price of the Amazon S3 usage. The cost of the Amazon EC2 service includes the price of the virtual-server image set up in Amazon EC2 and the price of the data transfer into the cloud. We estimated the total costs for the period of one month, under assumption that at the beginning we transfer 250 GB of data to the cloud (full backup copy of the database) and then make incremental backup of database three times a month (weakly). The Internet connection speed was limited to 10 Mbit/s. The size of the incremental backup was 25 GB. We used a special tool on the Amazon S3 web page to calculate the necessary cost of using the cloud storage. Our conclusion was that using a portable hard drive to transfer the full backup to the cloud is not reasonable as the cost of such transfer amounts to 235$ compared to the transfer of the same amount of data over a 10 Mbit Internet connection, at the cost of 25$; it lasts three days and six hours. Besides, the transport of the portable hard drive to the Amazon computer center also lasts from three to four days. The total cost of the backing-up data in the cloud (using the Amazon S3 service) is presented in Table 2.

Amazon S3 service			
Storage price (the price of the first 1 TB/month is 0,14$/GB):			
Amount of data	Storage time (days)	Calculation: 0,00452$/GB day * Storage time * Amount of data	Cost
full backup - 250 GB	31 days	0,00452*31*250	35,00$
1st incremental backup	24 days	0,00452*24*25	2,71$
2nd incremental backup	17 days	0,00452*17*25	1,92$
3rd incremental backup	10 days	0,00452*10*25	1,13$
The cost of data transfer to the cloud:			
1 x 250 GB		1 x 25$	25,00$
3 x 25 GB		3 x 2,5$	7,50$
TOTAL:			73,26$

Table 2. Cloud backup cost estimation

When using the Amazon EC2 service, with the database located on a virtual server in the cloud, the cost of the backup is the same, since the data are also transferred to the Amazon S3 service. The advantage of using a virtual server is mainly in the speed of backing up the data because of the data being transferred within the cloud (the Amazon-service provider), where we are not limited to a throughput of the Internet. So in our case, a full backup would be made in less than three days. Also to be mentioned is the cost of the virtual server usage. It is 0.62$/h for an Extra Large, High-memory instance of a server running Windows Server 2008 x64 operating system. The cost for one month is therefore 461.28$. Compared to the cost of the physical server, this is almost 10-times less. For example, the cost of our test server with an equivalent operating system is 4200$.

3.7 Traditional backup speed and cost estimation

In our measurement of the time of a traditional backup we added the time for making backup on tape and the time used in transportation of the tape to a remote location. The time measured was for a full and incremental backup without and with data compression, similarly as described in Section 3.5. We used a tape drive for recording data on a tape, courier service for transporting tapes to a remote location and a service for tape storage in the vault at a selected remote location. The selected tape drive was able to record all the data (full or incremental copy) on a single cartridge. The times taken in producing full and incremental backups are shown in Table 3.

	Full backup time		Incremental backup time	
	Compression		Compression	
	No	Yes	No	Yes
	< 1:15 h	> 15 min	< 10 min	> 1:42 min
Transport to a remote location and storage in the vault	< 3 h	< 3 h	< 3 h	< 3 h
Total time	< 4:14 h	> 3:15 h	< 3:10 h	> 3:1:4 h

Table 3. Traditional backup performance

From the full or incremental backup times given above, it can be concluded that:

- The minimum time required for a full backup is 3 hours and 15 minutes. Compared to the time required for a full backup in the cloud (2 hours), that is by 62.5% more.
- The minimum time required for incremental backups is approximately 3 hours and 2 minutes, which is six-times more compared to the time needed for incremental backup in the cloud (30 minutes).

In assessing the cost of a traditional tape-based backup, we assumed that we needed a backup tape drive, appropriate number of tape cartridges, courier service for transporting the tapes to a remote location and service for tape storage in the vault at a remote location. The cost was similar to the one in Section 3.6 evaluated for the period of one month and under assumption on that at the beginning we had to make a full backup of the database, and then three times a month (weakly) we made an incremental backup of the database. The total cost of making tape-based backups at a remote location is presented in Table 4.

Traditional tape based backup			
	Unit cost	Number of units	Cost
Tape drive	2750$	1	2750$
Transport of the tapes with a courier service	40$	4	160$
Lease of the vault (1 year)	240$	1	240$
Cartridge (400 GB) for a full backup	45$	1	45$
Cartridge (400 GB) for an incremental backup	45$	3	135$
TOTAL			3330$

Table 4. Traditional backup cost estimation

The cost of a tape-based backup using the above scenario would be 3330$ in the first month because of the need of buying a tape drive and hiring a vault for at least a year. In the months to follow (in a period of one year), the cost would decrease because of the price of the tape drive and the price of the vault. Thus, the reduced cost would be 340$. Irrespective of the above, making such backup copies is still 4.6 times more expensive than providing data backups in the cloud.

4 CONCLUSION

Business systems can use very different scenarios for making their backups. Judging from the obtained results, the decision of using a cloud should be well considered. First, business systems should determine their requirements about the speed of making backups and the costs associated with it. Only then they can find the turning point on which to decide whether it makes more sense to use a traditional way of backing-up data or to use a cloud.

Besides solving the issues of performance and costs discussed in this paper, business systems should answer a number of other important questions involved with making backups in the cloud [7]: Can they trust the data to a cloud-service provider [8]? What will happen to the data if a service provider stops working? Can data be transferred "to another cloud" of another cloud-computing provider? Are data really rejected after becoming unusable and are they in the cloud safe from a theft? To sum up, there are still many questions needing to be answered in the context of further research before the use of cloud computing can really flourish.

REFERENCES

[1] B. HAYES, Cloud Computing, *Communications of the ACM*, Vol. 51, No. 7, pp. 9-11, 2008
[2] A. REED, S. G. BENNET, Silver Clouds, Dark Linings: A Concise Guide to Cloud Computing, *Prentice hall*, ISBN-13: 978-0-131-38869-7, 2010
[3] Amazon Simple Storage Service (Amazon S3), http://aws.amazon.com/s3/
[4] Amazon Elastic Compute Cloud (EC2), http://www.amazon.com/ec2/
[5] G. J. POPEK, R. P. GOLDBERG, Formal requirements for virtualizable third generation architectures, *Communications of ACM*, Vol. 17, No. 7, pp. 412–421, 1974
[6] N. STINCHCOMBE, Cloud computing in the spotlight, Vol. 6, No. 6, pp. 30-33, 2009
[7] S. VRHOVEC, R. RUPNIK, A model for resistance management in IT projects and programs, *Electrotechnical review*, Article in reviews
[8] D. SVANTESSON, R. CLARKE, Privacy and consumer risks in cloud computing, *Computer Law & Security Review*, Vol. 26, No. 4, pp. 391-397, 2010

Aljaz Zrnec graduated in 1999 and received his master degree in 2002 from the Faculty of Computer and Information Science, University of Ljubljana. In 2006 he received his PhD degree in method engineering. He works in the Laboratory for data technologies as a lecturer and assistant in the field of databases. His research work involves database management systems and cloud computing. He is the author or coauthor of numerous professional and scientific papers.