the Availability Digest

Don’t Wait for the Other Shoe to Drop

February 2007

Redundant systems are great for protecting against a failure. But once a failure occurs, fix it fast before a second failure turns a problem into a disaster.

Such a delay cost a small company dearly. What should have been a weekend ordeal turned into a two-month disaster.

A Smooth Running Credit and Collection Service

The subject of our story is a company that has provided credit and collection services to the food industry for over fifty years. It maintains a large, up-to-date credit database for companies in the food industry and provides this credit information to its subscribers. Credit information includes lines of credit, returned checks, slow payment, credit litigation, and other delinquency information, as well as trade references.

Upon a request from a subscriber, the company’s credit investigators further research a specific company to provide more detailed current credit information on that company.

Changes in credit status are reported via a bulletin which gives subscribers early warning of pending credit problems. The company encourages its subscribers to contribute to the bulletin so as to extend its reach beyond publicly available information. The bulletin is augmented by news feeds, articles, and white papers concerning the food industry.

A subscriber may at any time submit delinquent accounts to the company for collection. The company’s collection professionals immediately review and determine the best approach to ensure collection. This action may range from collection efforts directly by the company to referral for legal action. The requesting subscriber is given frequent status reports on its collection accounts. Collection services are provided on a contingency basis, so the subscriber pays only a portion of what is collected.

Going Online

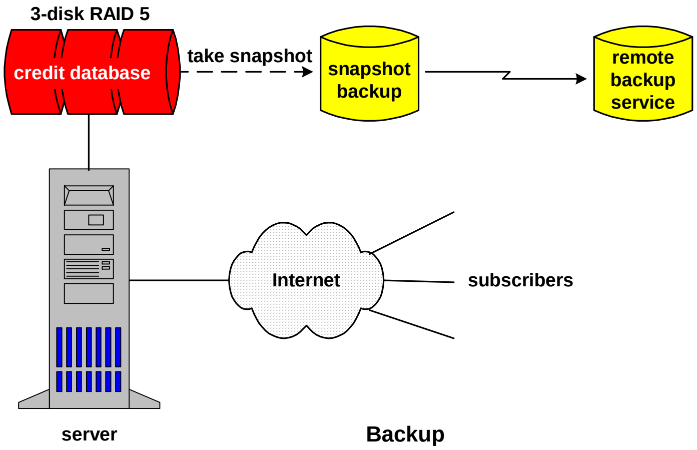
After decades of service to its subscribers, the company decided to offer its services online. The Web-based system maintains the credit information database, which is available to its subscribers. It also holds the credit bulletin and its associated news feeds, articles, and white papers.

Subscribers can submit delinquent accounts and receive status reports on collection activity via the Web. With several years of successful Web-based service behind it, the company has become substantially an online operation. The timely provision of its services is now very dependent upon the continuing operation of its Web services.

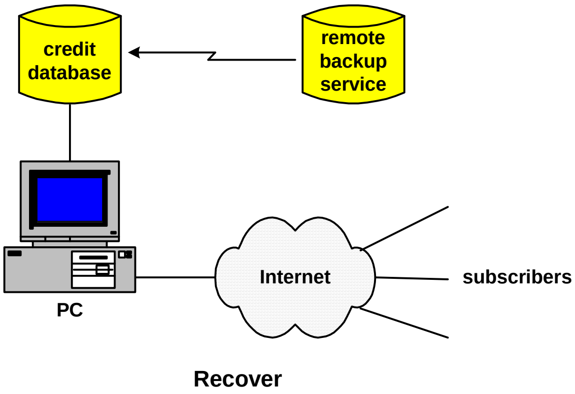
The company chose a Dell PowerEdge Server running the Windows 2000 Server operating system. The database manager is SQL Server. To guarantee database persistence, data is stored on a three-disk RAID 5 array. Thus, the data is still available should a disk in the RAID array fail. The system hosts several custom applications created by the company to support its services.

The Backup and Recovery Strategy

To ensure continuity of service in the event of a failure of the server, the database was backed up nightly to the eSureIT remote backup service provided by Intronics. A backup PC, available on the company's premises, could be put into service and could substitute for the primary server should that server fail. It was recognized that the capacity of this PC backup could not handle peak loads, but it was felt that it would get the company through the downtime of the failed primary server.



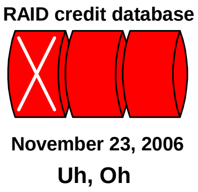
The database backup procedure started with taking a snapshot of the RAID database and storing it on a 300 gigabyte local disk. That snapshot was then compressed, encrypted, and sent to the remote backup service for safe storage. Recovery entailed loading the remote copy of the database onto the local backup PC, bringing up its applications, and connecting it to the network of workstations and the Internet.



As with any good business continuity plan, the recovery plan was tested by transferring the snapshot of the database from the local backup disk to the backup PC, bringing up the applications on that server, and testing that everything worked correctly. Recovery time was demonstrated to be about four hours, a period deemed acceptable by the company.

With what was believed to be a tested backup and recovery plan in place, the company was protected from a primary server failure, right? Wrong, as we shall see.

Problem #1: A Drive Failure



On November 23, 2006, drive 0 in the RAID array dropped offline. As expected, this created no interruption in service.

The company has a next-day service contract which covers the cost of failed parts. The service vendor was contacted, and a service person arrived onsite to replace the failed disk. It was expected that all that was needed to be done was to insert the new drive and to rebuild it from the data on the other surviving drives. The full redundancy of the server would then be restored.

Problem #2: A Rebuild Failure

However, after multiple attempts, drive 0 would not rebuild. The company was informed that its only option was to rebuild the server. This would require taking down the server, restoring the RAID configuration, formatting the drives (which would delete the database), installing the operating system, installing the applications, restoring the database, and thoroughly testing the repaired server before returning it to service.

The company's IT department developed a plan to carefully migrate the failed server's functions to a temporary server with the same RAID configuration, to repair and rebuild the primary server, and to restore it to service. This would avoid the limited-capacity problem associated with the use of the backup PC.

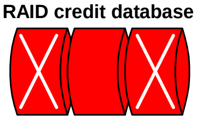
To do this migration required Microsoft Certified System Engineer services, which were estimated to cost $1,600.

Problem #3: Unconcerned Management

Unfortunately, the company's management didn't see this migration as a priority item. After all, the system was still running fine. What was the chance of a second disk failure? Based on management's experience to date, disk drives were highly reliable; and the probability of a second disk failure was essentially nil. Besides, it would be in a better position to cover this cost in the coming year. The result was no action.

Problem #4: Murphy's Law

If an event is highly unlikely, it still means that it will occur someday in the future. And that someday might be tomorrow.



December 14, 2006

Oops!

In the company's case, that someday was Thursday, December 14th. On that fateful day, a second drive failed; and the primary server went down.

Now the IT folks were faced with a failed server and with no time to arrange for a loaner server as they had before. Their only option was to fail over to the backup PC and to take the consequences of reduced capacity.

The timing couldn't have been worse. The failure occurred just before the Christmas and New Year's holidays. Not only did this mean potential long hours for the IT staff during this time of family commitments, but its usual support contacts would be charging higher rates to be available over the holiday weekends.

So much for management saving money. The company was now only interested in how long it would take to get back into operation since its only recourse during this downtime was inefficient manual operations. IT's best estimate was six days, including working over the weekend.

Problem #5: A Reluctant PC Backup

Though the "recovery plan" had been practiced in the past, the IT staff now realized that only a portion of the recovery plan had been tested. It had recovered from the backup snapshot disk on the primary server (which was now unavailable), but it had never attempted recovery from the remote backup service. Furthermore, none of its recovery plans had been documented.

It turned out that recovering from the remote service to the same RAID configuration that created the backup was straightforward. However, recovering to a different configuration – in this case, a PC with a single disk – just didn't work. Googling the "internal consistency error" that was being reported yielded lots of discussion but no solutions.

Finally, after a long weekend (which included a 5 am to 11 pm shift on Saturday), the staff was able to get the backup PC working; and service was restored with full functionality the next Monday morning. To no one's surprise, the PC's performance matched its prediction. The PC made a terrible server, but that was all they had.

Problem #6: One-Day Service is Six-Day Service Plus

In the meantime, efforts were frantically underway to bring up the primary server. The service vendor was contacted to come out and rebuild the primary server. According to the company's service contract, the service provider was committed to next-day service. However, in spite of that commitment, resource overbooking and holiday schedules delayed the service call for six precious days.

These same problems continued to plague the recovery effort. As it turned out, it would take a total of 21 days to recover the primary server.

Problem #7: Loss of Network Services

A problem that had not been anticipated was that the company's Domain Controller, which provides necessary network services, ran on the primary server. It therefore had crashed with the primary system and was no longer available. Normally, the Domain Controller is decommissioned, which is a controlled shutdown. Then a new domain controller can be put into place. However, the server crash precluded a soft decommission.

The Domain Controller maintains domain-specific data such as logons, share permissions, scripts, computer accounts, and IP addresses. This data is kept in a Global Catalog along with policies describing the network.

The loss of the Domain Controller created significant instability for the backup PC. For instance, workstations couldn't log on since security policies and personal settings were unavailable. After a few days, workstation IP addresses, which had been leased from the Dynamic Host Configuration Protocol (DHCP) server also hosted by the primary server, began to expire, disabling the workstations. These parameters had to be laboriously recreated manually on every company workstation.

Without the Domain Controller, the network was slowly dying. The company's Exchange Server and Web Server started running unbearably slow. PCs were crashing every morning when they requested updates from the now-dead antivirus server which had been running on the primary server. To correct this, antivirus protection had to be disabled. This left the company's workstations without protection from viruses, spam, spyware, and malicious scripts. Users could not find their documents and would get many errors when they attempted to log on or log off.

To add insult to injury, a RAID drive on the company's Exchange Server failed. Fortunately, this was replaced and restored without incident.

When the primary server was finally recovered, the fact that the Domain Controller had not been properly decommissioned caused further problems. The Domain Controller had to be manually decommissioned and its links broken so that it could be rebuilt. All of the prior server entries had to be removed manually from the Global Catalog. This was akin to going through a telephone book, looking for instances of a given item, and deleting them.

Furthermore, the new server could not rejoin the network until a manual search of the policies found the policy that was preventing the server from rejoining.

Wait! It's Not Over

All of these problems contributed to the 21 days required to restore the system. But the problems weren't over yet.

On the first day that the primary server was again in operation, the IT staff was overwhelmed with user complaints. Settings were incorrect. Scripts were pointing to the old server. It took several days to work out these problems.

And the IT staff is still not done. As of this writing in early February, there is still a raft of tasks that must be performed before recovery is complete:

- SQL maintenance and backup jobs need to be reinstalled.
- All onsite and remote backups need to be reinstalled and reconfigured.
- The antivirus server needs to be reinstalled and configured.
- Some printers and document scanners are still offline.
- Workstations still crash.
- Many scripts are still pointing to the wrong places.
- User roaming profiles are not yet working.
- The Exchange server is still exhibiting problems.

As the company's IT Manager said, "Overall, without taking into consideration all of the configuration and troubleshooting that still needs to be done, the time from failure to restore took thirty days and cost well over what it would have cost to properly plan and execute this service."

Lessons Learned

As is often the case, a chain of problems led to the eventual disaster that befell this company. If any link in the chain had been broken, the company's problems may not have happened or would have been far less serious.

Many lessons from this experience are obvious from hindsight, which, unfortunately, is always 20/20. These include:

- Educate managers about the availability aspects of the system, and impress on them the importance of timely maintenance. The middle of a crisis is no time to do this.

- When a failure of a spare component occurs, it must be a top priority to fix it rapidly. This would have precluded the unanticipated crash of the company's primary server.

- Recovery plans must be thoroughly tested. Testing just what seems to be a critical part of the plan isn't enough. In this case, the IT staff was comfortable with the fact that it could recover from the primary's database snapshot and assumed that recovering from the remote backup would be no different.

- Business operations ought to be tested with the backup system to ensure that there are no unanticipated problems with the backup configuration when it is in actual service. This would have exposed the network problems associated with the Domain Controller and other services that were running on the primary server.

- The recovery plan must be well documented. When the pressure is on to recover is no time to be trying to figure out what to do.

All of these tasks have nontrivial costs associated with them. It is up to management to decide if these costs are less than the costs of a multiday disastrous failure.