



FEATURE



Database Archiving for Tomorrow

[By Trevor Eddolls, senior consultant for NEON Enterprise Software]

Organizations are generating and keeping more data now than at any time in history. Most companies understand the importance of archiving their data, but compliance with new regulations means that a database archive solution must not only work today, but must still be effective 30 years from now.

There are many reasons to archive. Databases are getting bigger — Gartner has quoted growth rates of 125%. In addition, the type of data that can be stored in a database has increased. Originally, databases stored characters and numbers and dates and times, but DB2 9.1, for example, happily stores unstructured data such as images, video, and XML documents.

Large databases perform less well than smaller ones. They require more CPU, and back-ups and REORGs take longer. This is the reason organizations began archiving data. They later found that they could use this archived data for data mining activities — such as finding which marketing strategies had worked best in the past or for providing ideas for seasonal variations in sales.

Perhaps the most compelling reason to archive is compliance with regulations, such as Sarbanes-Oxley (SOX), HIPAA (Health Insurance Portability and Accountability Act), and BASEL II, plus what is estimated to be over 150 state and federal laws. These regulations dictate the length of time that electronically stored information (ESI) needs to be retained. Indeed, the regulations — and they do depend on the industry — have increased data retention periods from 20 to 30 years or more. Compliance with these regulations will drive demands for data archiving solutions.

So what do we actually mean by database archiving? It has been defined as the process of removing selected data records from operational databases that are not expected to be frequently referenced again and storing them in an archive data store, where they can be securely retained and retrieved as needed, and then discarded at the end of their legal life.

So the first part of any archive strategy must be to store data that is not needed to complete a business transaction (operational) or that is not needed for reporting or other queries (reference). Secondly, it needs to archive related records associated with a business object — which will involve taking data from different tables in the databases. It should not try to archive at the file or row level because of the way business data can be spread about. Thirdly, this archive process needs to be policy driven and automated.

Not only does the data have to be in the archive for possibly decades in order to comply with regulations, it also has to be accessible to authorized people and must be retrievable using standard SQL queries. In addition to access and retrieval characteristics, it's important to be able to produce reports about the data using standard techniques.

The next important archive characteristic is compliance with Section 802 of the Sarbanes-Oxley Act and rule 240.17a-4 of the Securities and Exchange Act (1934). These regulations affect the authenticity of the archive. Companies face severe penalties if they alter or delete their archived data. So, for compliance reasons, the archived records must be stored in a format that is both non-rewritable and non-erasable.

If data from a database archive is restored, then it needs to go back into the same columns and tables in which it originally existed. Information about these tables and columns is called metadata, so for an archive to be successful, it must also store the metadata along with the data. Over time the database may be modified as new versions are brought out, or, with company acquisitions and mergers, the database in use may change. This is why archiving the metadata is so important. No matter what happens, the archive data will remain accessible in its original format. In terms of compliance, recent amendments to the Federal Rules of Civil Procedure (FRCP) affect the discovery of electronically stored information. Rule 34b states that, "A party who produces documents for inspection shall produce them... as they are kept in the usual course of business..." So, basically, this means that the archive data has to be independent of the originating database.



FEATURE

Once data has reached the end of its “legal life” and is no longer required to be retained, the archive solution should have a policy for the automatic deletion of that data from the archive.

In the event that litigation occurs or is pending, data is placed in a litigation hold. That means it cannot be deleted or changed for any reason. Having decided what information might be available and needed in the court case, the next stage is to be able to locate that data in the archive. This is where e-discovery can be used. It is important that the archive stores data in a way that allows e-discovery tools to work fairly quickly. There have been cases where huge fines have been imposed because electronic documents have not been produced in a timely fashion (For example, Serra Chevrolet v. General Motors). Once litigation is over, the data may still have a long legal life ahead of it before it can be deleted.

It almost goes without saying that if archives are to store data for up to 30 years that they

will be very big. Figures in petabytes (10¹⁵ bytes) have been suggested. The analyst firm Enterprise Strategy Group concluded that between 2005 and 2010 the required digital archive capacity will increase by more than a factor of 10 — from 2500 petabytes in 2005 to 27,000 petabytes in 2010.



Sophisticated archiving systems will prevent data from being altered or deleted, while at the same time allowing it to be accessed and retrieved. The archive data is stored on a storage area network (SAN) using encapsulated archive data objects (EADO), which allow access and retrieval of data from the archive, while also maintaining the authenticity of the data and preventing it from being overwritten or deleted. This ensures that users are compliant with the growing list of regulations today and

tomorrow, and also for many decades into the future.

About the Author

Trevor Eddolls is a Senior Consultant for NEON Enterprise Software, a Sugar Land, TX-based technology leader in enterprise data availability software and services. Trevor has over 25 years of experience in all aspects of IT, and for many years he was editor of Xephon’s (www.xephonusa.com) *Update* journals. You can read his weekly blog at mainframeupdate.blogspot.com. He can be contacted by email at trevor@itech-ed.com. Visit www.neonesoft.com.

ON THE NET

NEON Enterprise Software
www.neonesoft.com

Trevor Eddolls’ Blog
mainframeupdate.blogspot.com

EmploymentCrossing is the largest collection of active jobs in the world.

We continuously monitor the hiring needs of more than 250,000 employers, including virtually every corporation and organization in the United States. We do not charge employers to post their jobs and we aggressively contact and investigate thousands of employers each day to learn of new positions. No one works harder than EmploymentCrossing.

Let EmploymentCrossing go to work for you.