Database archiving

By Heiko Müller, University of Edinburgh

1. Introduction

In a computational context, data archiving refers to the storage of electronic documents, data sets, multimedia files, and so on, for a defined period of time. Database archiving is usually seen as a subset of data archiving. Database archiving focuses on archiving data that are maintained under the control of a database management system and structured under a database schema, e.g., a relational database. The primary goal of database archiving is to maintain access to data in case it is later requested for some particular purpose such as a Freedom of Information (FoI) request. In fact, compliance with government regulations on data preservation are the main driver for the majority of current data archiving efforts.

The term database archiving, however, can be problematic as it is used differently by different communities. The most common definition of database archiving comes from the business and government communities as is defined as '*the process of removing selected records from operational databases that are not expected to be referenced again and storing them in an archive data store where they can be retrieved if needed*'.1 Following this definition, database archiving requires the active selection and appraisal of data records to ensure that those no longer deemed necessary for daily operational or reference purposes are moved to a separate data store for longer term retention. The data is maintained in the archive for as long as required based on legal or institutional requirements. If and when the data is no longer required, it is either transferred or destroyed.2

When archiving scientific and reference data, database archiving is frequently regarded as maintaining a collection of database snapshots over time (see for example Buneman *et al* (2002)). This form of database archiving involves making off-line copies of the data and managing these copies efficiently. In addition to being able to reconstruct the database as it was at a certain point in time, database archives enables tracking and querying of the history of objects, for example: "How did the functional annotation of a protein change over time and are these changes reflected by changes to the amino acid sequence?"

2. Short-term Benefits and Long-term Value

There are a number of business and research related benefits associated with database archiving. Some of these benefits include:

- **Verification of scientific findings:** Archives of scientific data allow retrieval of historic database states. These database snapshots in turn can be used to verify scientific findings that were derived based on the data.
- **Querying database history:** Managing database snapshots in a single archive enables tracking and querying the history of database objects. This capability to answer temporal queries is especially valuable for reference databases like the CIA World Factbook, as it can answer queries like "How did the energy consumption in China change over the past 15 years?"
- **Cost savings related to reduced hardware requirements:** The bigger a production database, the more infrastructure, i.e., server and disk capacity, it requires. Regularly scheduled database archiving frees up disk capacity, thus saving money on hardware upgrades. Maintaining active

data on-line and selecting the most appropriate storage medium for archived data further ensures cost-effective balance of storage media throughout the information life cycle.

- **Improved application performance:** Overloaded databases degrade performance. Database archiving ensures that production databases are maintained at a manageable size to improve performance and availability of critical systems. Removing rarely used data from production databases frees processing power and improves application performance.
- **Improved database administration:** Large databases require more administration efforts for upgrade, migration, and backup, potentially leading to long outages for critical applications.
- **Meeting legal and regulatory requirements:** Several compliance regulations mandate enterprises to store all related data for several years before deleting it. Archiving is a viable approach to storing such data and meeting compliance requirements.
- **Customer service issues:** Maintaining historic information in an archive enables the resolution of customer-related queries that may span several month or years.

3. e-Science Perspective

"*Entries [in the EMBL Nucleotide Sequence Database] are subject to changes, but only the most recent versions have been preserved. However, it has become necessary to see entries exactly as they were in the past, especially when references are made to specific versions of sequences from third party annotation entries, or from journal articles. Additionally, patent attorneys are interested in seeing the exact content of an entry at a given date in the past.*"

— Leinonen R, Nardone F, Oyewole O, Redaschi N, Stoehr P. "The EMBL sequence version archive" in *Bioinformatics*, Vol. 19, No. 14, (2003).

4. Issues to be Considered

- A comprehensive enterprise database archiving solution must provide the capability to archive data from a variety of sources, for example, relational database management systems from different vendors, or XML documents.
- When moving and removing records in a database management system, the implemented archiving solution has to ensure that existing integrity constraints are satisfied in both the production database and the archive.
- Data in the archive has to be accessible within a reasonable time-frame without requiring extensive manual manipulation. When storing data on tapes, for example, it can take weeks to browse and scan data which could result in a considerable time and cost penalty.
- Government regulations on data retention usually define penalties for altering or deleting important business data. Thus, the electronic storage media used for database archiving must preserve data records in a non-rewritable, non-erasable format.
- When archiving scientific data, the scale of the data is quickly becoming a limiting factor. For example, the Sloan Digital Sky Survey contains around 100 terabytes of data and petabytes of data will not be unusual in the future. Copying a petabyte of data, however, would take approximately 3 years with current technologies.[3]

5. Additional Resources

- Buneman, P. *et al*. "Archiving scientific data" in *Proceedings of the 2002 ACM SIGMOD international Conference on Management of Data*, Madison, Wisconsin, June 03 - 06, (2002).
- Forrester Consulting "Why database archiving should be part of your enterprise DBMS strategy" .
- Lee, J. "Database Archiving: A Critical Component of Information Lifecycle Management" in Database Journal (2004).
- Mullins, C. S. "Database Archiving for Long-Term Data Retention" , (2006).
- Robb, D. "Preventing Database Bloat With Archiving" .

- Szalay, A. "Preserving digital data for the future of eScience" in Science News (2008).

1 *Cf*. Mullins (2006)

2 *Ibid*. see the DCC Curation Lifecycle Model for the range of activities that may be involved in database archiving
3 Szalay (2008)

Digital Curation Centre
Appleton Tower, 11 Crichton Street, Edinburgh, EH8 9LE | t. +44 (0)131 651 1239

The DCC is funded by Joint Information Systems Committee