Towards Passive, Migration-Free, Standardized, Long-Term Database Archival

Raja Appuswamy
EURECOM, Biot, France

"How would you archive databases for the next 60 years such that they incur no migration cost, and they remain usable in 2080?" This was an open challenge raised by digital preservation experts from the Landesarchiv of Baden-Württemberg [12], who, similar to other memory institutions (archives, museums, libraries, etc.), have faced several challenges in archiving culturally significant, historic data stored in digital databases since early 1960s.

On the hardware front, all current media technologies suffer from media decay and have a limited lifetime of few decades at best. Further, current media technologies tightly couple the storage medium, with the technology to read data off the medium. This leads to media obsolescence, where data stored in an older medium is no longer readable by new readers. Memory institutions, in contrast, are tasked with preserving databases for much longer duration requiring expensive, periodic data migration. On the software front, database engines store data in proprietary file formats that evolve rapidly. This leads to format obsolescence, as data archived in an engine's native format becomes unreadable by even a newer version of the same engine. In addition to archiving data, it is also necessary to archive the application logic expressed in stored procedures, SQL queries, and views, as they provide the context in which data is accessed. Unfortunately, even SQL statement archiving is not standardized due to deviations from ANSI/ISO SQL caused by vendor-specific extensions [15].

These challenges are not specific to memory institutions; the growth of data fueled by AI and BI, coupled with new compliance requirements, has created the perfect breeding ground for long-term archival challenges in today's data-driven enterprises. It is time to revisit the archival hardware–software stack, especially given recent advances in the development of storage technologies based on novel media. One such medium that has received a lot of attention recently is synthetic DNA [2, 4, 7, 9, 14].

DNA possesses several advantages over current media. First, its theoretical density is at least eight orders of magnitude higher than contemporary magnetic media [5]. Second, DNA is very durable and can last millennia at ambient temperature [6]. Third, DNA decouples the medium (biological molecules) from read technology (sequencing). Thus, data stored in DNA can be left untouched without migration as it does not suffer from media decay or obsolescence.

Designing read and write pipelines for DNA-based database archival opens up several opportunities for data management research. For instance, in project OligoArchive [13], we are developing database-aware encoding [2], decoding [10], and bootstrap [1] techniques for reliably archiving databases on DNA [11]. In addition to these challenges, several others, like designing indexed access paths over DNA [14], exploiting rather than masking DNA errors by using it as an approximate storage medium [8], or enabling near-molecule computation by executing search [3] or query [2] operations using biochemical reactions over DNA, remain open for further research.

While DNA solves issues at the medium level, it does not solve format obsolescence issues. Digital preservation experts have standardized text-based, archival file formats (SIARD [15]) and rely on exporting data out of a database into a software independent archival file to overcome format obsolescence. However, text-based archival leads to data bloat, and tooling support is limited to a few relational databases. In contrast, the rise of open-source binary file formats like Arrow and Parquet is a step towards eliminating format obsolescence. But more work is required to understand their conformance to SQL and their ability to archive non-relational data and contextual application logic.

With archival problems affecting memory institutions and enterprises alike, it is time for database vendors and researchers to tackle the Landesarchiv challenge with solutions and standards that can make low-cost, migration-free data archival feasible.

REFERENCES

[1] R. Appuswamy and V. Joguin. Universal layout emulation for long-term database archival. In *CIDR*, 2021.

[2] R. Appuswamy, K. Lebrigand, P. Barbry, M. Antonini, O. Madderson, P. Freemont, J. MacDonald, and T. Heinis. OligoArchive: Using DNA in the DBMS storage hierarchy. In *CIDR*, 2019.

[3] C. Bee, Y.-J. Chen, D. Ward, X. Liu, G. Seelig, K. Strauss, and L. Ceze. Content-based similarity search in large-scale dna data storage systems. *bioRxiv*, 2020.

[4] G. M. Church, Y. Gao, and S. Kosuri. Next-Generation Digital Information Storage in DNA. *Science*, 337(6102), 2012.

[5] S. R. Corporation. 2018 semiconductor synthetic biology roadmap. `https://www.src.org/program/grc/semisynbio/ssb-roadmap-2018-1st-edition_e1004.pdf`, 2018.

[6] D. Coudy, M. Colotte, A. Luis, S. Tuffet, and J. Bonnet. Long term conservation of dna at ambient temperature. implications for dna data storage. *PLOS ONE*, 16(11), 11 2021.

[7] Y. Erlich and D. Zielinski. DNA Fountain enables a robust and efficient storage architecture. *Science*, 355(6328), 2017.

[8] G. Franzese, Y. Yan, G. Serra, I. Onofrio, R. Appuswamy, and P. Michiardi. Generative dna: Representation learning for dna-based approximate image storage. In *VCIP*, 2021.

[9] N. Goldman, P. Bertone, S. Chen, C. Dessimoz, E. M. LeProust, B. Sipos, and E. Birney. Toward Practical High-capacity Low-maintenance Storage of Digital Information in Synthesised DNA. *Nature*, 494, 2013.

[10] E. Marinelli and R. Appuswamy. Onejoin: Cross-architecture, scalable edit similarity join for dna data storage using oneapi. In *ADMS*, 2021.

[11] E. Marinelli, E. Ghabach, T. Bolbroe, O. Sella, T. Heinis, and R. Appuswamy. Dna4dna: Preserving culturally significant digital data with synthetic dna. In *iPRES*, 2021.

[12] K. Naumann. `https://www.landesarchiv-bw.de/de/aktuelles/termine/72973`.

[13] OligoArchive. Oligoarchive: Eu fet initiative on intelligent dna data storage. `https://oligoarchive.eu/`.

[14] L. Organick, S. D. Ang, Y.-J. Chen, R. Lopez, S. Yekhanin, K. Makarychev, M. Z. Racz, G. Kamath, P. Gopalan, B. Nguyen, C. N. Takahashi, S. Newman, H.-Y. Parker, C. Rashtchian, K. Stewart, G. Gupta, R. Carlson, J. Mulligan, D. Carmean, G. Seelig, L. Ceze, and K. Strauss. Random access in large-scale DNA data storage. *Nature Methods*, 11(5), 2014.

[15] SFA. SIARD Suite. `https://www.bar.admin.ch/bar/en/home/archiving/tools/siard-suite.html`.