Database preservation through conversion technology

Damir Bulic (Spectral Core Ltd Company, Dublin, Ireland)

Damir Bulic has 30 years of experience building software in a wide variety of languages, frameworks, and environments. He is the owner of Spectral Core, a company specializing in database migrations with customers in more than 90 countries. For the past 20 years, his focus has been on database tooling. His interest is in removing vendor lock-in, SQL parsing, analysis, and transformation.

Bulic's talk continued the topic of best current practice. His company helps other companies in copying data from one place to another. They support around 40 database formats and big data, migration to the cloud or data lakes. The conversion software is called "Full Convert". It supports, among many other formats, transformation into SIARD from 40 different database management systems.

The enterprise level product is called "Omni Loader" and is a distributed migration cluster. This is an ideal solution for migrating databases on premises to the cloud. It handles hundreds of terabytes similarly to "Full Convert" on a single computer for the basic use case.

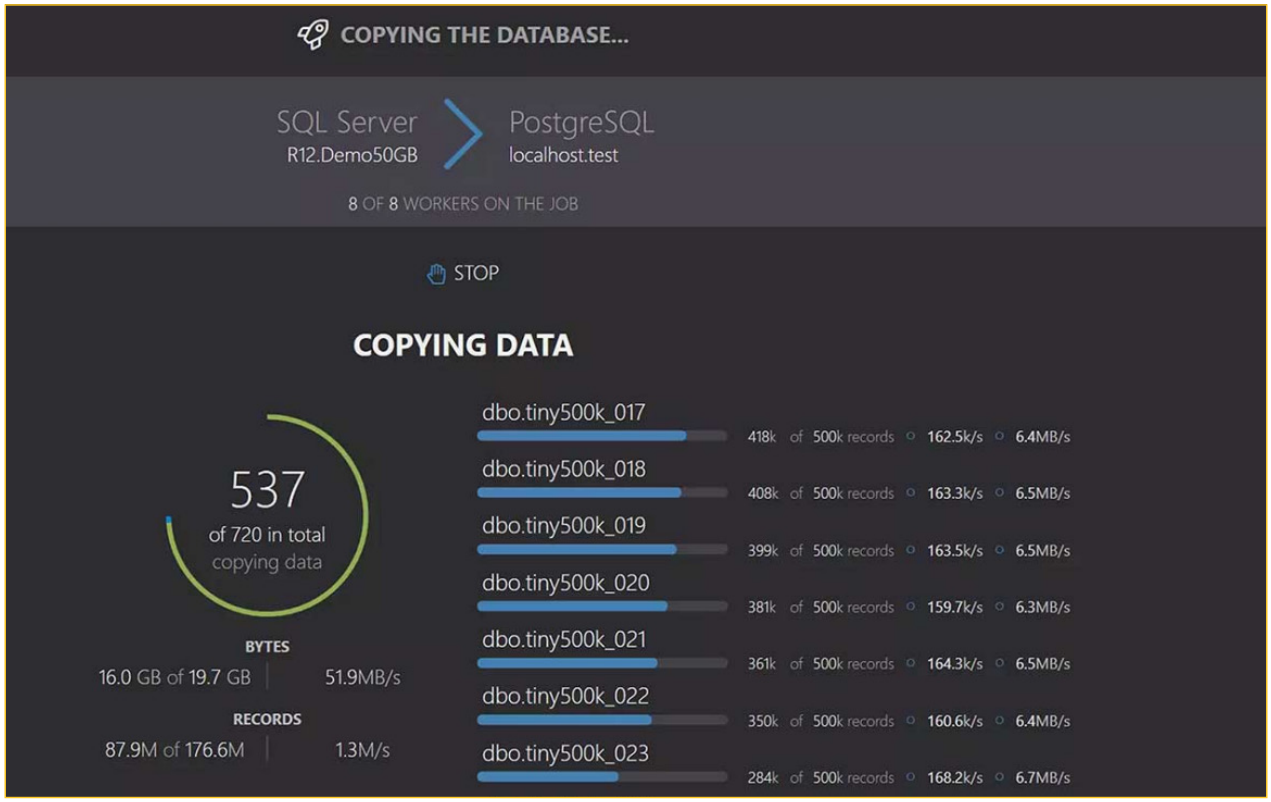

Figure 15: Spectral Core Full Convert software at work

Another Spectral Core product to mention is “SQL Tran” which allows for data definition language (DDL) translations of very complex schemas. SQL Tran extracts complex data from storage containers and copies it somewhere where you have full control. This feature is also important to prepare for obsolescence. Also, there is “Documenter”, a tool for database schema documentation. Spectral Core feels their software for working with SIARD is very effective.

Bulic then shifted to a critical review of SIARD. According to him, SIARD is useful now but will not be useful in five years’ time, as databases are getting larger and more complex. He named the following challenges for SIARD:

- It is XML and zipped
- A single file that cannot be serialised
- Hierarchical text format
- Verbose and clunky
- Useful only for small datasets
- Cannot handle much of what we see in the real world today. Datasets are growing at fast pace and SIARD will be less applicable in future.
- Full convert can handle up to 16 connections, but ZIP writing cannot be paralleled.
- SIARD cannot be supported in Omni Loader (most powerful database migration software they sell).

Looking to the future, Bulic predicted the end of the concept of vertical scaling, and Moore’s law (observation that the number of transistors in a microchip doubled about every two years, see references) being no longer applicable. According to him, horizontal scaling is the answer, meaning a division of calculation tasks in separate workloads that are executed on multiple machines and reunited for the result. The cloud is all about horizontal on-demand scaling and this trend will continue. Workflows are moving away from personal computers.

A new possible format Bulic proposed for archival data should have a separate structure for data and schema, highly compressed chunks of columnar data (10 times better compressed), and extensible data types. A perfect archival format should allow for searches and should not have to be all in one file. It could in fact be distributed across the world. It could probably be based on SQLite as this is well used everywhere. SQLite can support billions of records.

Questions and discussion

- Faria noted that the SIARD standard is already developing in some of the areas mentioned, for example how it could be segmented into different files. SIARD supports several Terabytes currently. He saw an importance to view and understand the tradeoff between performance and suitability for long-term preservation. There are benefits in having a self-documenting file without reliance on a technology stack. Performance may decrease but the benefits are around interoperability, simplicity, and the ability to use it in different contexts.

- Rechert once again got back to the purpose of preservation. Sometimes you have logic encoded into the front end. With complex databases and systems, you need to spend some effort to reconstruct it so that future users get the same results when they run the same queries etc. Every approach has its merits, but he still reminds us to think about what is needed 60 years from now. Bulic noted that there needs to be some sort of ISO format that is readable and efficient for the future.
- Appuswamy compared the discussions on Bulic's SIARD successor to the format used in DNA storage – it needs to be self-describing and compressed. He suggested the term "binary SIARD". Faria noted we have a binary SIARD once it is zipped. There may be room for improvement, Appuswamy thought.
- Incremental backup is very common in databases, Appuswamy noted. How do we handle this with emulation and other solutions?