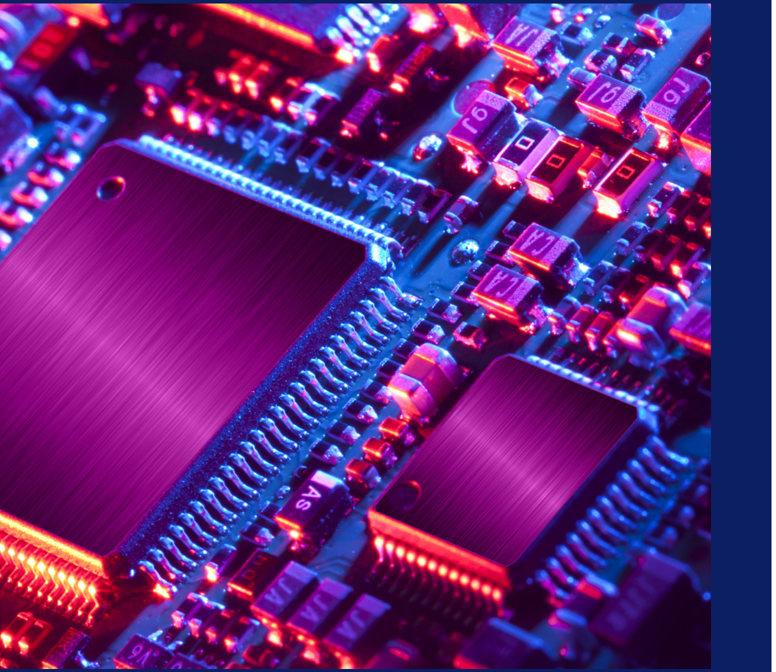
MAKING BIG DATA EASY WITH DATA VIRTUALIZATION

Author: **Rick F. van der Lans**



INTENDA

It is no longer special for organizations to own multiple really big data systems. Organizations increasingly collect and analyze massive amounts of data to improve business and decision-making processes.

Unfortunately, developing big data systems is not always easy and can be risk-prone. The technology to store the data can be complex, use unfamiliar database concepts, use a different and proprietary language or API, and it can support complex, non-flat data structures.

An approach taken by organizations to simplify working with big data technology is to copy the data from the big data system to a more traditional SQL-based system. While this simplifies data access, such an approach has several drawbacks:

• The process of copying big data to a SQL database can be too time-consuming. Extracting millions of records from one system and loading it into another, even when almost no transformations need to take place, can take a long time.

• If the big data system is massive, the SQL database will be massive as well. So, the organization ends up with two big databases. And if the SQL databases is accessed by a wide range of queries, many indexes are required to speed up processing. All those indexes lead to extra storage. In fact, the SQL database can end up being factors bigger than the big data system itself.

• The SQL database needs to be managed by DBAs, which costs time.

• There will be license costs for this SQL database.

• The more copies of data that are stored, the more challenging it is to comply with GDPR rules on personally identifiable data, especially with regard to supporting the customers' "right to be forgotten."

• When data consumers have access to the SQL databases, they have access to a non-up-to-date copy of the big data, so real-time access to data is not an option, in other words, the data latency is high.

• Some big data systems are classified as big, not because of the volume of data stored, but because of the massive amounts of streaming message that need to be analyzed real-time. When data in a streaming environment needs to be stored before it can continue, it's not streaming anymore. There is a delay.

All these aspects stand in the way of big data projects and especially the aspect of making big data available to data consumers. Data virtualization can simplify the use of data stored in big data systems by operating as a data abstraction layer between the data consumers and big data systems. Data consumers access the big data systems through the data virtualization platform.

Unfortunately, developing big data systems is not always easy and can be risk-prone.

The latter can transform the database concepts and proprietary language or API of the big data systems to more well-known interfaces such as SQL. It can also transform complex, non-flat data structures into flat, relational data structures. In other words, it can create a SQL-based view on big data systems.

The advantages of this approach are:

- No time is wasted on copying data.

- No additional databases need to be developed and managed.

- There are no license costs for an extra database. Evidently there are license costs for the data virtualization platform.

- There are fewer problems with GDPR compliance, as no or fewer copies of the data are stored.

- Because the big data system is accessed directly, data consumers have access to real-time data.

- Streaming data is not supported by all data virtualization platforms, but it is supported by fraXses. In other words, big data streaming applications can be supported and do not require intermediate storage of data.

In addition, products such as fraXses can deploy some of the big data technologies internally.

For example, for specific queries, fraXses creates, optimizes and enhances filter / join / aggregation pushdown query plans that are pushed into Apache Spark to execute queries using parallel processing. Data virtualization platforms also support advanced forms of data security and can extend or replace the underlying data security system.

Some big data experts are not really aware what data virtualization can do for them. Therefore, I recommend them to study this technology. It may reduce the risks of big data projects and speed up the exploitation of big data within organizations.

fraXses supports the streaming of data. You do not need an intermediate storage of data.