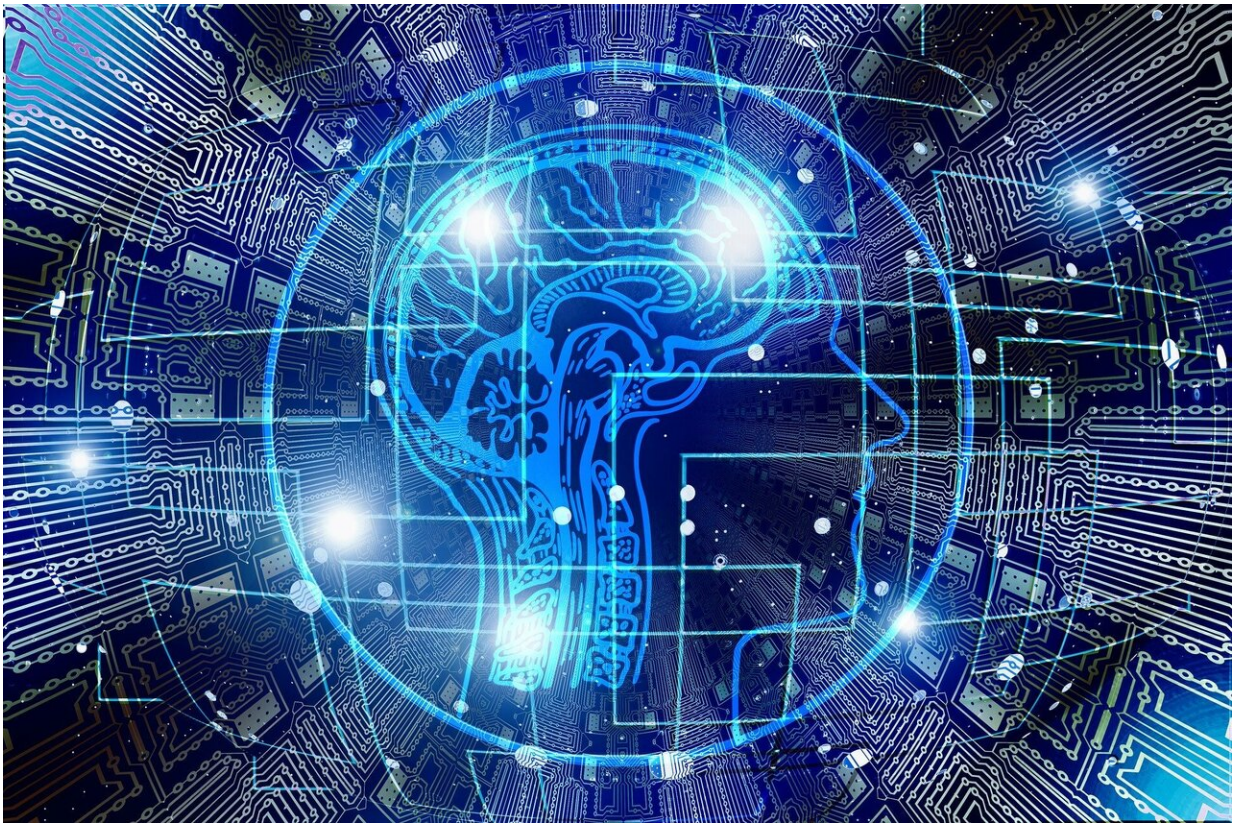
# Data management system developed to bridge the gap between databases and data science

June 9 2020



Credit: Pixabay/CC0 Public Domain

Relational databases are used to store information or data in such a way that it preserves relations between the data. This property makes it a useful tool for data scientists. There is, however, a gap between the

relational database research community and data scientists. This leads to inefficient use of databases in data science. Ph.D. student Mark Raasveldt tried to bridge the gap between the relational databases and data science. Ph.D. defense 9 June 2020.

## Integration with analytical tools

Most data scientists use analytical tools, such as R, Python and C/C++, for their research. These tools are difficult to integrate with current database systems, resulting in slow and cumbersome data analysis. "Data scientists have opted to reinvent database systems by developing a zoo of data management alternatives that perform similar tasks to classical database management systems, but have many of the problems that were solved in the database field decades ago," says Raasveldt.

"The database research community has made tremendous strides in developing powerful database engines that allow for efficient analytical query processing." Raasveldt tried to combine these innovations in the database science with the analytical tools that are mostly used by data scientists. "We investigate how we can facilitate efficient and painless integration of analytical tools and relational database management systems," says Raasveldt.

## Large datasets

Another issue with the use of standard database systems in computer science is the size of the data that is handled. Most database systems are not optimized for large data sets and large-scale data analysis using remote servers. To optimize the database systems, there are three methods that can be considered.

"We focus our investigation on the three primary methods for database-

client integration: client-server connections, in-database processing and embedding the database inside the client application," Raasveldt explains. For every method, he studied the implementations in existing database systems and he evaluated how efficient they are for the large datasets and workloads that are common in data science.

## DuckDB

Raasveldts final result was a new data management system, called DuckDB, that was purpose-built for efficient and painless integration with R and Python (and other analytical tools). This management system is meant to be used as a mature database system that is not only used for research purposes.

"In DuckDB, we take all the lessons that we have learned investigating database-client integrations and create an easy-to-use and highly efficient embedded database." Raasveldt will continue his work as a postdoc at the CWI, where he will work on further developing DuckDB.

**More information:** DuckDB: www.duckdb.org

Provided by Leiden University

Citation: Data management system developed to bridge the gap between databases and data science (2020, June 9) retrieved 13 May 2024 from https://techxplore.com/news/2020-06-bridge-gap-databases-science.html