Geophysical Research Abstracts
Vol. 16, EGU2014-16290, 2014
EGU General Assembly 2014

Unified Access Architecture for Large-Scale Scientific Datasets

Risav Karna
Jacobs University Bremen, CS Department, Bremen, Germany (r.karna@jacobs-university.de, 4915778685442)

Data-intensive sciences have to deploy diverse large scale database technologies for data analytics as scientists have now been dealing with much larger volume than ever before. While array databases have bridged many gaps between the needs of data-intensive research fields and DBMS technologies (Zhang 2011), invocation of other big data tools accompanying these databases is still manual and separate the database management's interface. We identify this as an architectural challenge that will increasingly complicate the user's work flow owing to the growing number of useful but isolated and niche database tools. Such use of data analysis tools in effect leaves the burden on the user's end to synchronize the results from other data manipulation analysis tools with the database management system.

To this end, we propose a unified access interface for using big data tools within large scale scientific array database using the database queries themselves to embed foreign routines belonging to the big data tools.

Such an invocation of foreign data manipulation routines inside a query into a database can be made possible through a user-defined function (UDF). UDFs that allow such levels of freedom as to call modules from another language and interface back and forth between the query body and the side-loaded functions would be needed for this purpose. For the purpose of this research we attempt coupling of four widely used tools Hadoop (hadoop1), Matlab (matlab1), R (r1) and ScaLAPACK (scalapack1) with UDF feature of rasdaman (Baumann 98), an array-based data manager, for investigating this concept. The native array data model used by an array-based data manager provides compact data storage and high performance operations on ordered data such as spatial data, temporal data, and matrix-based data for linear algebra operations (scidbusr1). Performances issues arising due to coupling of tools with different paradigms, niche functionalities, separate processes and output data formats have been anticipated and considered during the design of the unified architecture. The research focuses on the feasibility of the designed coupling mechanism and the evaluation of the efficiency and benefits of our proposed unified access architecture.

Zhang 2011: Zhang, Ying and Kersten, Martin and Ivanova, Milena and Nes, Niels, SciQL: Bridging the Gap Between Science and Relational DBMS, Proceedings of the 15th Symposium on International Database Engineering Applications, 2011.

Baumann 98: Baumann, P., Dehmel, A., Furtado, P., Ritsch, R., Widmann, N., "The Multidimensional Database System RasDaMan", SIGMOD 1998, Proceedings ACM SIGMOD International Conference on Management of Data, June 2-4, 1998, Seattle, Washington, 1998.

hadoop1: hadoop.apache.org, "Hadoop", http://hadoop.apache.org/, [Online; accessed 12-Jan-2014].

scalapack1: netlib.org/scalapack, "ScaLAPACK", http://www.netlib.org/scalapack,[Online; accessed 12-Jan-2014].

r1: r-project.org, "R", http://www.r-project.org/,[Online; accessed 12-Jan-2014].

matlab1: mathworks.com, "Matlab Documentation", http://www.mathworks.de/de/help/matlab/,[Online; accessed 12-Jan-2014].

scidbusr1: scidb.org, "SciDB User's Guide", http://scidb.org/HTMLmanual/13.6/scidb_ug,[Online; accessed 01-Dec-2013].