



## The Open Source Data Management and Analytics Software for Scientific Research

---

### SciDB Overview

#### The Need

Science and industry are growing increasingly data-intensive. Analyses of terabytes and even petabytes of data are becoming routine. Operating at this scale in an efficient manner involves rethinking systems from the ground up. Building new systems from scratch for each new peta-scale project is grossly inefficient and unnecessary.

Representatives from the scientific, industrial, and computer science communities dealing with extremely large databases met at the 1<sup>st</sup> XLDB workshop in October 2007, and a followup meeting in March 2008. These sessions highlighted the rise of large-scale analytics, the convergence of analytic needs between science and industry, and frustration with slow progress in this area due to repeated re-invention.

The consensus from the meetings was that the communities should join forces to build a next-generation open source data management system for data-intensive scientific analytics.

#### Open Source

An open source product is required for widespread adoption in the academic science world, where fear of “data lockup” is pervasive. Open source is also viewed positively in the industrial world as a key tool to manage the cost of large data systems. In order to build the required product using distributed development resources, the system architecture must be modular and extensible. This design in turn encourages contributions from the entire community, including both public and private sector participants. The result is not only a superior product to what any single contributor could create, but also a product that can support a diverse array of applications through user-defined functions and procedures while leveraging a single, coherent core.

#### A New Data Management System

SciDB will not be a traditional database. Instead, SciDB will be optimized for data management of big data and for big analytics. To distinguish it from traditional DBMSs, we're calling it a DMAS, a Data Management and Analytics Software System. It will not be optimized for online transaction processing (OLTP) and will only minimally support transactions at all. It will not need to provide strict atomicity, consistency, isolation, and durability (ACID) constraints. It will not have a rigidly-defined, difficult-to-modify schema.

Instead, SciDB will be built around analytics. Storage will be write-once, read-many. Bulk loads, rather than single-row inserts, will be the primary input method. “Load-free” access to minimally-structured data will be provided.

The standard relational model is often inefficient for the types of data used for complex analytics. Time series and spatial grids may be represented in relations, but only at a severe cost in both space and processing time. SciDB will be organized around multidimensional array storage, a generalization of relational tables that can provide orders of magnitude better performance.



## The Open Source Data Management and Analytics Software for Scientific Research

---

### Data-Intensive Scientific Analytics

In order to support the vast collections of data being obtained by new instruments or new simulations, SciDB will need to be scalable up to petabytes and beyond. This scale necessitates the use of more than one machine; SciDB will run on incrementally scalable clusters or clouds of industry standard hardware. The system will also be scalable down to megabytes to enable researchers to use the same interface on a laptop as on a 10,000-node cloud. Computation must scale equally with the storage. Functions and procedures will execute in parallel, as close to the data being operated on as possible.

Operating on a large number of industry standard nodes requires that reliability be engineered into the system from the very first release. SciDB will be designed to continue operating in the face of node failure, without even restarting a long-running operation in progress. Scalability also means that expensive human administrative costs cannot increase even linearly with the size of the data. The system will accordingly be designed for automated operations with minimal administrative overhead.

Complex analytics will be simplified with SciDB because arrays and vectors are first-class objects with built-in optimized operations. Spatial operators and time-series analysis will be easy to express. Interfaces to common scientific tools like R and eventually MATLAB and IDL, as well as programming languages like C++ and Python, will be provided.

Many features important to science that have been developed in the research community but have not been incorporated into commercial databases will be standard with SciDB, including versioning, provenance tracking, and support for uncertain data with error bars. By building these features into the system, rather than patching them on with external tools, the accuracy and consistency of the data and the resulting analyses will be ensured.

### The Users

SciDB will primarily be designed to meet the needs of data-intensive scientific analytics in the public and private sectors. User communities expected to benefit include sciences such as astronomy, biology, geoscience (geology, oceanography, atmospheric science, environmental science), medicine, and physics; science-based industries such as remote sensing, resource extraction (oil, gas, minerals), medical imaging, and pharmaceuticals; and other organizations with vast amounts of data and complex analytical needs such as Internet, telecommunications, and financial services.

### The Organization and Founding Community

Open source projects must have a vibrant support base and achieve a critical mass of user adoption in order to survive and thrive. A new nonprofit foundation, SciDB.org, has been created to lead development and evangelize the product, providing the necessary focus to seed this community. Founding design participants include world-class thought leaders from a wide variety of disciplines.