# Towards Analytical Data Management for Numerical Simulations

Ramon G. Costa, Fábio Porto, Bruno Schulze
{ramongc, fporto, schulze}@lncc.br
National Laboratory for Scientific Computing - RJ, Brazil

**Abstract.** Numerical simulation of natural phenomena is being fostered by recent advances in powerful high processing computing platforms. Scientists in various areas, such as human cardiovascular system, model a phenomenon being studied through a set of mathematical equations. As scientists strive to obtain a more realistic simulation, a huge amount of data is produced. Unfortunately, there has been little work on supporting numerical simulation data management, which leaves simulation scientists with huge standard text files and complex analytical programs that eventually extract some meaningful information to validate scientific hypotheses. Moreover, some analytical queries cannot, as well, be represented using none of the scientific query languages. In this context, this paper tries to bridge this gap by raising some issues involved in numerical simulation data analysis. A representation for numerical simulation data is presented that considers a multidimensional model, for dimensional variables, and their corresponding physical quantities. A cloud service to interface with the numerical simulation data manager is proposed and its integration with the Neblina cloud middleware is explored.

## 1 Introduction

Many scientific areas are taking advantage of development in high processing computing to model natural phenomena through in-silico simulation. The process involved in modeling phenomena starts by observing phenomenon data, and modeling the process through a set of mathematical differential equations that expresses the variation of selected physical quantities on time-space. Next, the scientist may choose an appropriate numerical method that would solve the equations and compute for each reference point the values for selected physical quantities. Using state-of-art cluster platforms computer scientists strive to obtain the most possible realistic simulations to compute weather forecasts, just to name an application.

In this paper we propose a novel strategy for scientific simulation data management based on database approach. We focus on the support to the analysis of simulation results, taking a sample of a simulation output and loading it onto a cloud data service modeled using a multi-dimensional array representation for space and time to manage space-time multi-scale dependent data. We illustrate our discussion with the simulation of the human cardiovascular system developed at LNCC, INCT-MACC [2].

## 2 Literature review

Recent studies have demonstrated the need for a more efficient storage, indexing and processing strategies for scientific data [4], [8] and [10].

In Ogasawara [4], scientific workflows are modeled as data intensive applications. Parameter sweep experiments evaluate data represented as a set of parameter configuration values. Moreover, typical workflow activities are identified according to their data consumption and production rate and mapped to algebraic operators, such as: Map, Reduce, Filter, and JoinQuery.

The storage of numerical simulation data has been investigated by [10]. Multidimensional scientific data are modeled using an array data model. The authors propose different types of array storage models according to array sparsity.

Another important initiative in support for scientific data management is SciDB [6], a database management system for scientific applications. It offers a multidimensional model based on multiarray representation. Its functionality includes: data versioning, uniform distribution of data across the nodes of a cluster, and two query language interfaces: AFL and AQL languages.

Considering the analysis of simulation data, one question is to determine the past and future cone of information, as presented by Sowa [7]. Similar is the study of causality in databases [3], in which data that contributes to a given result is considered to cause, with a certain responsibility, such result.

The Magellan project [5] explores the adoption of Cloud Computing in scientific applications. It evaluates some recent technologies in support for HPC, such as: virtualization, MapReduce, Eucalyptus, and Hadoop. In addition, an important issue that the project aims to evaluate is to confront the performance of scientific applications running on local clusters versus a cloud environment.

## 3 Challenges

As discussed in section 1, we are interested in specifying a data management service in support to numerical simulation results analysis. This section investigates some possible representations for numeric simulation data and how to extend database functionality to support scientific data handling under it.

### 3.1 Data representation

Simulation data can be interpreted as composed by two sets of variables: dimensional and physical quantities. Typically, the dimensional variables include space and time. The space dimension refers to a mesh, which represents the topology of the physical domain as a composition of simple geometric objects. A mesh is represented by: a set of points, referring to the vertices of the geometric objects; the set of edges linking the points and the faces of the model. Observe, yet, that simulations may adopt different scales throughout the domain. Furthermore, given a physical quantity, its value in a reference coordinate may change through scales. A given simulation may be composed of data in different scales, according to the precision requirements in different parts of the physical domain.

### 3.2 Using SciDB for storing simulation data

Our first effort to represent numerical simulation data uses SciDB. A user specifies multidimensional structures by providing the range values for each dimension and a list of attribute values to compose a cell. In addition, a versioning mechanism keeps historical values for each attribute. In this context, the following mapping strategy has been defined: for each set of physical quantities corresponding to a phenomenon being simulated in a given scale; define the set of $\Delta$ dimensions; specify the list of physical quantities $\Pi$ to be computed; create an array having the dimensions as $\Delta$ and attributes as $\Pi$. Using the AQL [6], we would define the following schema for the 3D model of an artery.

```
CREATE ARRAY Geometry3D <velocity: point3D, pression: double,
   displacement: point3D> [ simulations=0:*,1,0, t=0:500,500,0,
   x=1:7000,1000,0, y=1:7000,1000,0, z=1:36000,1000,0]
```

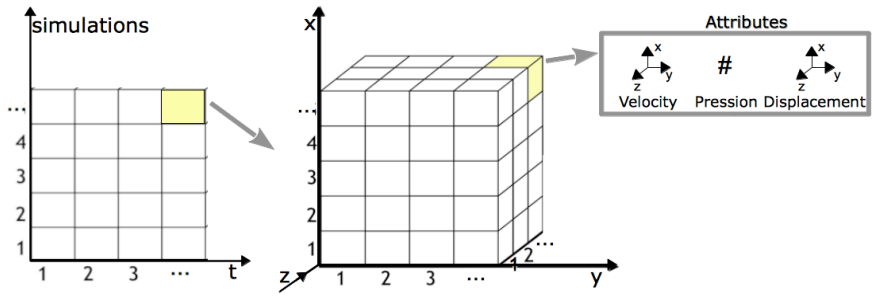


**Fig. 1.** 3D graphical representation of a multidimensional array for the artery model

Through the AQL query language, we can use the following schema to represent 1D (i.e different scale) of the human cardiovascular system:

```
CREATE ARRAY Geometry1D <velocity: double, pression: double,
flow: double> [ simulations=0:*,1,0, t=0:500,500,0 ]
```

A 1D model represents each cross-section summarized in a single point. Thus, each point contains the values for the physical quantities on each time step. Additionally, It's refers to the different simulations over the same mesh. Observe that the array data model adopted by SciDB enables the direct representation of multidimensional objects produced in numerical simulations.

### 3.3 Analyzing numerical simulation data

In order to support the analysis of numerical simulation output, algebraic operators must be provided, eg.: drill down through scales - given a reference coordinate in scale $s_i$, return the corresponding set of points in scale $s_j$; AQL and

AFL languages do not have sufficient mechanisms to support many analytical queries. The challenge is to create new operations and functions to bridge this gap, as well as, to coupling them to algebraic operators. An example is to obtain the values for pression where the highest values are achieved:

```
aggregate(Geometry3D, max(pression));
```

### 3.4 Analytical data management service

From an architectural point of view, we expect to develop a service to interface with the solvers - producing simulation data - and with scientists - submitting analytical queries to the system. The Simulation Data Management Service (SDMS) is responsible for providing such an interface as a cloud service, and should manage the storage and retrieval of the simulation data making it transparent to scientific applications. The availability of the SDMS as a cloud service fosters the collaborative use of simulation data among scientists of a same research project and among different projects.

An important aspect of the cloud service approach is the possibility to explore elasticity [9]. Indeed, depending on the requested analysis, a huge amount of data may be retrieved from the data storage device. In such a scenario, the system may allocate extra memory for processing and freeing it as the computation ends. Such a policy is not only a sign of altruism in a collaborative environment, but may reduce the costs involved in supporting the computing platform. Finally, the SDMS should offer to scientific applications an API for accessing its services using traditional programing languages, such as C++, Java an Scripting languages such as Python.

### 3.5 Scientific computing in cloud

Regarding the SDMS, an important issue is the research and development of mechanisms that would enable its deployment in a private cloud as a Service (SaaS) [9]. In this context, we should highlight the software Neblina, presented in [1]. Neblina is a middleware developed at LNCC that offers users an interface to cloud resources. Through Neblina a cloud infrastructure, including an application, may be accessed and managed. Typical functionalities include: resources capacity provision, user management, virtualized and physical resources management interface, remote access to the resources and their monitoring.

The SDMS has been integrated into Neblina. This integration makes the cloud environment transparent to SDMS, enabling for instance the activation of its services. In Fig. 2 the architecture of the integrated environment is shown and the numbers suggest a retrieving order.

## 4 Conclusion

We investigate the requirements involved in designing a data management service in support for numerical simulation analysis. A multidimensional modeling approach represents the dimensions used in referencing each individual simulation

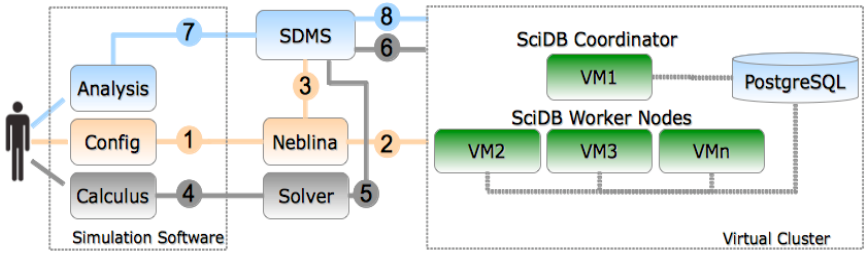


**Fig. 2.** Numerical simulation environment

point, and maps each point to its respective physical quantity values. We observe that although the multi-array model adopted by SciDB enables the implementation of the multidimensional representation, further extensions are required to fully support numerical data representation and analysis requirements. In particular, some analysis may require physical quantities to be computed over different abstractions, such as computing their values in a face or edge of a geometry object in a mesh. Moreover, supporting modeling through different scales would require a relationship between multiple representations of the same multidimensional space-time. Some proposed analytical queries can not, as well, be represented using none of the SciDB query languages. New functions and user data types would be needed to cope with those. We expect that this work will provide a better understanding concerning the needs involved in analytical data management for multidimensional numerical simulations.

## References

1. Fernandes, F.J., et al.: Neblina - espaços virtuais de trabalho para uso em aplicações científicas. In: XIXX SBRC. pp. 965–972. Campo Grande, Brazil (Jun 2011)
2. Laboratório Nacional de Computação Científica: Medicina Assistida por Computação Científica (Mar 2012), `http://macc.lncc.br/`
3. Meliou, A., et al.: Causality in databases. Data Eng. Bull. 33(3), 59–67 (2010)
4. Ogasawara, E., et al.: An algebraic approach for data-centric scientific workflows. In: 37th Intl Conference on VLDB. vol. 4, pp. 1328–1339. Seattle, USA (Aug 2011)
5. Ramakrishnan, L., et al.: Magellan: experiences from a science cloud. In: 2nd intl workshop on Scientific cloud computing. pp. 49–58. New York, USA (2011)
6. SciDB Inc.: SciDB User's Guide (2011), `http://www.scidb.org/`
7. Sowa, J.F.: Knowledge Representation: Logical, Philosophical, and Computational Foundations. Course Technology (Aug 1999)
8. Stonebraker, M., et al.: Requirements for science data bases and scidb. In: Conference on Innovative Data Systems Research. Asilomar, USA (Jan 2009)
9. Zhang, Qi, et al.: Cloud computing: state-of-the-art and research challenges. Journal of Internet Services and Applications 1(1), 7–18 (May 2010)
10. Zhang, Yi, et al.: Storing matrices on disk: Theory and practice revisited. In: 37th Intl Conference on Very Large Data Bases. Seatle, USA (Aug 2011)