

## Scientific Data Management Using Ad-hoc Queries over Simulation Data

### Goal

The goal of the DataFoundry query infrastructure is to help scientists understand and explore the large data sets being generated by complex scientific simulation codes.

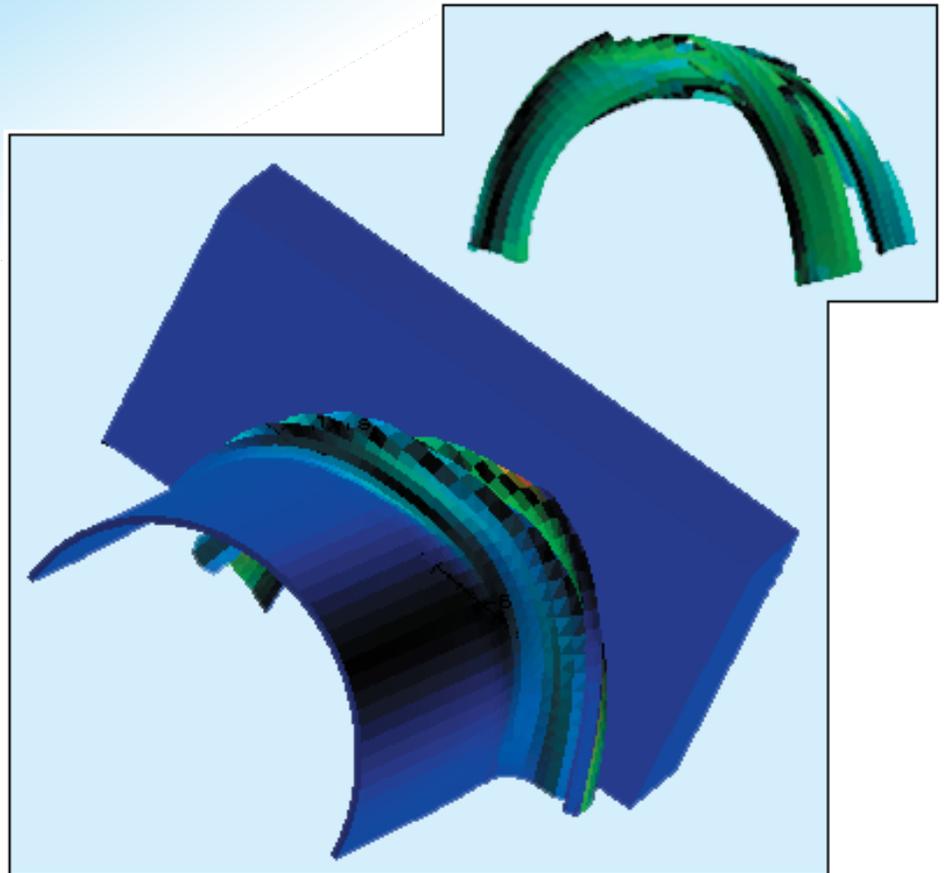
### Technology

DataFoundry extends the approximate query research being performed in the database community. It uses a multi-resolution index to access mathematical and statistical models of the original data set and provide approximate query capabilities over large-scale simulation data.

### Applications

The DataFoundry query infrastructure can be used to help analyze any mesh-based data sets. Currently, we are evaluating our system using astrophysics data generated by the Djehuty project.

Traditionally, scientists have analyzed simulations by using visualization tools to display and explore the data sets these complex codes have generated. As computing power increases, these simulations are able to model physical reality with increasing precision. Unfortunately, this increased precision usually results in a corresponding increase in the data set size. Currently, large-scale simulations produce more data than can be effectively visualized. There are ongoing efforts to address this problem by modifying visualization tools to provide a multi-resolution view of the data, and to provide alternative representations of information (such as iso-contour graphs). DataFoundry's query infrastructure is complementary to these efforts.



*Figure 1. The figure on the left is a pseudo-color plot on pressure of a can partially crushed against a wall. The figure on top is the subset of that image containing only those zones in which the pressure value is between 20% and 50% of the maximum value.*

DataFoundry's query infrastructure provides an environment in which scientists use queries to identify the subset of the data that is currently of interest to them, and visualize only that information. Ideally, the scientist would be able to ask a question against the original data set and immediately receive a completely correct answer. The query results would be presented as a subset of the data and the scientist would be able to interact with it through the same visualization tool as they normally use.

Unfortunately, because of the size and complexity of the underlying simulation data, it is infeasible to query the original data directly. Furthermore, even providing relational-style indices over this data is impractical because of the number of individual objects and the wide variety of queries that

scientists wish to make. To overcome these problems, DataFoundry performs an extensive pre-processing step in which mathematical and statistical models of the data are generated and indexed. Because these are models of the data, instead of the data itself, we can achieve significant data compression while still retaining the essence of the data.

Scientists interact with these models through a graphical interface that allows them to easily define complex queries and explicitly trade response time for query accuracy. The interface connects to a query engine that uses the models to provide approximate answers to the queries. When the query completes, the results are automatically displayed using a standard visualization tool.

## Data Warehousing and Integration for Scientific Data Management

### Technology

DataFoundry combines leading-edge database, data integration, wrapper generation, and meta-data technologies to enable powerful distributed queries across a large number of independent, heterogeneous data sources.

### Application

DataFoundry is a flexible, scaleable infrastructure that is powerful enough to meet the demands of realistic scientific environments. It is relevant to application areas that require the use and sharing of large-scale, distributed, heterogeneous information and is being developed in the genomics arena to enable new computational analysis techniques for functional genomics.

Computational methodologies are proving to be a viable and cost-effective alternative to the slower and more expensive experimental methods prevalent in many scientific domains. However, these new techniques require access to large amounts of highly complex and dynamic data that may be distributed across multiple, autonomous, heterogeneous data sources.

### Difficulties in Managing Scientific Data

Data access and integration technologies play a central role in addressing the scientific challenges facing domains such as genomics. Unfortunately, the data management solutions available today do not scale to the hundreds of dynamic, scientific data sources needed to solve these complex problems. In functional genomics, for example, scientists need access to data from domains such as DNA sequence, protein structure, genetic trait,

```
FFREDLAFQK  
AREFSSEQTRAN  
SPTRRELQVWGG  
ENNSLSEAGADR  
QGTVSFNFPQITL  
WQRPLVTIR IGG  
QLKEALL DTGAD  
DTVLEEMNLP GK  
WKPKMIGGIGGF  
IKVRQYDQIPVEI
```



Figure 2. A partial sequence of the Human Immunodeficiency Virus protein (HIV Protease) is shown on the left, and the protein's 3D structure is shown on the right.

genetic linkage, physical mapping, and taxonomy, which are currently spread across hundreds of data sources each using customized interfaces and data formats.

An important characteristic of scientific data sources is that not only does the data available from a source change rapidly, but the formats in which the data are disseminated also change frequently. Managing data that is continually evolving raises difficult data management issues that current solutions such as traditional data warehouses and federated databases are too rigid and labor-intensive to address. In addition, the nomenclature between data sources is often conflicting and the data can be duplicative, erroneous, and inconsistent.

### The DataFoundry Approach

DataFoundry is currently pursuing a hybrid approach in which a mediated data warehouse is used to provide access to a consistent view of data from the core data sources while a multi-database infrastructure provides limited access to a large number of external sources.

A mediated warehouse architecture is used because it is able to

provide reliable access to a consistent view of data from multiple data sources. We have augmented the traditional architecture with a meta-data infrastructure that reduces the effort required to incorporate new sources into the warehouse and adapt to changes in previously integrated sources. This infrastructure uses a meta-data based mediator generator program to create both a collection of C++ classes that correspond to the abstract concepts being represented in the warehouse and a mediator class that transforms the data into the appropriate format and enters it into the warehouse.

### Multidisciplinary Collaboration

The DataFoundry data access and integration project is a multidisciplinary effort involving scientists from the Center for Applied Scientific Computing and the Biology and the Biotechnology Research Program.

*For additional information about the DataFoundry project, contact Terence Critchlow, (925) 423-5682. [critchlow@llnl.gov](mailto:critchlow@llnl.gov).*