

# ornl

ORNL/TM-13227

RECEIVED

JUN 21 1996

OSTI

**OAK RIDGE  
NATIONAL  
LABORATORY**

**LOCKHEED MARTIN** 

## Visualization for the Large Scale Data Analysis Project

R. E. Flanery, Jr.  
J. M. Donato

MANAGED AND OPERATED BY  
LOCKHEED MARTIN ENERGY RESEARCH CORPORATION  
FOR THE UNITED STATES  
DEPARTMENT OF ENERGY

ORNL-27 (3-96)

# MASTER

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED *AF*

This report has been reproduced directly from the best available copy.

Available to DOE and DOE contractors from the Office of Scientific and Technical Information, P. O. Box 62, Oak Ridge, TN 37831; prices available from (423) 576-8401, FTS 626-8401.

Available to the public from the National Technical Information Service, U.S. Department of Commerce, 5285 Port Royal Road, Springfield, VA 22161.

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government of any agency thereof.

**DISCLAIMER**

**Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.**

Computer Science and Mathematics Division

Mathematical Sciences Section

**VISUALIZATION  
FOR THE  
LARGE SCALE DATA ANALYSIS  
PROJECT**

R. E. Flanery Jr.<sup>†</sup> and J. M. Donato<sup>‡</sup>

Mathematical Sciences Section  
Oak Ridge National Laboratory  
P.O. Box 2008, Bldg. 6010  
Oak Ridge, TN 37831-6414

<sup>†</sup> Director, Advanced Visualization Research Center.  
E-mail: flaneryrejr@ornl.gov.

<sup>‡</sup> Staff Researcher. E-mail: donatojm@ornl.gov.

Date Published: April 1996

Research was supported by the Laboratory Directed Research and Development Program of the Office of Energy Research, U.S. Department of Energy.

Prepared by the  
Oak Ridge National Laboratory  
Oak Ridge, Tennessee 37831  
managed by  
Lockheed Martin Energy Research Corp.  
for the  
U.S. DEPARTMENT OF ENERGY  
under Contract No. DE-AC05-96OR22464

## List of Figures

1	AVS Network for Filter Pre-Processing . . . . .	5
2	Snapshot of the Pre-Processing Phase . . . . .	6
3	AVS Network for Post-Processor . . . . .	7
4	One View of Post-Processing Output . . . . .	8
5	An Alternate View of Post-Processing Output . . . . .	9
6	View of Raw Data Represented by Chosen Data Point . . . . .	10

VISUALIZATION  
FOR THE  
LARGE SCALE DATA ANALYSIS  
PROJECT

R. E. Flanery Jr.<sup>†</sup> and J. M. Donato<sup>‡</sup>

**Abstract**

In this paper we overview the visualization approach used as part of the Large Scale Data Analysis project. This project used the AVS5 software to create interface tools for the public domain database management system POSTGRES. This work utilized hardware and software available through the AVRC<sup>1</sup>. This work is part of ongoing research on data analysis tools for large (terabyte) data sets. Statistical analysis and browsing tools for the data sets were implemented as part of an AVS5 software system network which interfaces to the POSTGRES database management system. These networks were implemented to provide uniform graphical interfaces for both the feature extraction stage and the post processing stage of the entire system. The network of tools for these interfaces, rather than the statistical tools themselves, are the focus of this paper.

---

<sup>1</sup>Advanced Visualization Research Center at Oak Ridge National Laboratory.

<sup>†</sup>Director, Advanced Visualization Research Center. E-mail: flaneryrejr@ornl.gov.

<sup>‡</sup>Staff Researcher. E-mail: donatojm@ornl.gov.

## 1. Introduction

Data sets in the gigabyte size range are difficult, if not impossible, to analyze with current methods. At Oak Ridge National Laboratory in Tennessee, we are concerned with large data sets from a variety of applications. These applications include:

CHAMMP<sup>2</sup> (Computer Hardware, Advanced Mathematics, and Model Physics) program sponsored by DOE, and

ARM<sup>3</sup> (Atmospheric Radiation Measurement) program sponsored by DOE.

The CHAMMP project, which simulates atmospheric global climate models, will eventually be producing data for analysis at the rate of about 1TB (terabyte) per day. The ARM archive for atmospheric water vapor records is maintained at ORNL and currently contains in excess of 350 GB of data distributed in over 50,000 files.

Researchers need tools to analyze these large data sets in a straightforward fashion. The tools must be easy to use. They must present the data in a compact form, yet be able to take advantage of the supercomputers and heterogeneous networks of computers on which the data to be analyzed is created and stored [3].

The work presented in this paper is part of ongoing research on data analysis tools for large (terabyte) data sets, [2], entitled "Analysis of Large Scientific Data Sets." The project utilizes statistical techniques to extract useful or interesting information from data sets too large for normal browsing techniques. AVS5 [1] is used to build the networks for the statistical filters which extract the information. These statistical filters are encapsulated as AVS5 modules. This allows us to interactively update the statistical parameters through the graphical interface. This stage is called the "feature extraction stage." A typical choice of filter parameters could produce filtered information of only a very small (e.g. 0.1%) percentage of the original large data set size. This resulting filtered information is stored and accessed via the POSTGRES DBMS [4, 5].

In this paper, we will focus on the graphical interface tools between AVS5 and POSTGRES.

AVS5 is the Application Visualization System by Advanced Visual Systems, Inc. which provides an interactive 3D visualization system and a sophisticated graphical user interface. Further information about AVS5 can be obtained via <http://www.avs.com/>.

POSTGRES is a public domain Data Base Management System. For more information refer to <http://s2k-ftp.CS.Berkeley.EDU:8000/postgres/>.

The goals of the "Analysis of Large Scientific Data Sets" project were:

- To automate the process of extracting useful or interesting information from the raw data sets.
- To store the resulting data using an appropriate data model for later retrieval and analysis.
- To provide useful tools for browsing the extracted data which are too large and unwieldy to manipulate or browse by current means.
- To implement a Graphical User Interface for the above.

In light of these goals, the following subgoals were also deemed important:

---

<sup>2</sup><http://www.esd.ornl.gov/programs/chammp/chammp.html>

<sup>3</sup><http://www-armarchive.ornl.gov/>

- To have a graphical user interface that even non-experts in statistical analysis techniques, AVS, or POSTGRES could use.
- That information be presented in a compact form, yet allow details to be easily accessed.
- That the tools utilize a client-server structure to allow these large data sets to reside on a large mainframe or supercomputer, yet allow the users to perform analysis and visualization at their workstations.

## 2. Overview of the System

The system interface consists of two main phases:

- Feature Extraction, and
- Visualization and Analysis.

In the feature extraction phase, the user supplies the system with specifications necessary to create a filtered representation of the large data set. As a minimum, the user specifies the following information:

- the input file containing the raw scientific data sets;
- the feature extractor to use in filtering the data.

The system then begins to read the raw data in intervals, applying the feature extractor to chunks of data, and writing a representation of the filtered data to the database. During this process histogram and box-plots of the raw data sections are displayed allowing the user to tailor parameters to better reduce or filter the resulting data representation.

In the visualization and analysis phase, the power of scientific visualization is used to discover unusual or interesting features of the original large data set in an easily comprehensible yet compact format. Through the visual interface the user displays one of the filtered representations of the original data. The system generates the necessary query language commands to the database system to load the appropriate filtered data.

By clicking a button in the user interface, the user may select different views of the filtered data representation. For example, the user might view "statistical value" versus "event time" using a color schema to depict variations in the "time between intervals." This technique gives a visually compact representation of three dimensional relationships in a two dimensional display where color variations represent the third dimension of data information.

Other views of the data are easily selected by a click of an interface button.

## 3. Technical Issues

The technical approach to the overall system just described contained a number of interesting, and sometimes difficult, implementation issues.

We use AVS5 to create the Graphical User Interface (GUI) as well as to create many of the analysis tools for the process. AVS5 was chosen because it provides a useful GUI, two and three dimensional renderers, and it is easily portable as well as user-extendible.

The database management system chosen was POSTGRES. POSTGRES is a public domain, easily portable, system. It allows back-end functions to better utilize the CPUs of the host machine, and provides object oriented extensions that allow useful Large Objects to be stored with the data.

However, interfacing these two systems presented the interesting challenge.

For example, during the visualization and analysis phase, the user selects some point of interest. The selection is then used to determine and load the corresponding raw data. Unfortunately, AVS5 did not easily allow access to the point location information. Hence a separate module implementing an interactive two-dimensional browser and graphic display was created to provide this crucial information via the user interface.

Due to the size of these scientific data sets, they are typically stored on large mainframes or supercomputers. It is crucial to minimize the amount of data down-loaded and processed on the client workstation. Hence, another aspect of this project was utilizing POSTGRES' back-end function capability to do as much processing on the back-end server machine as possible.

#### 4. Example System Usage

In this section we present and describe a number of typical system screens as seen from the user's viewpoint. Due to the dark background of the original color pictures from AVS, the screen pictures shown in this paper are in reversed video grey scale color.

The user starts AVS, selects the "Network Editor", then "Network Tools", and then selects "Read Network" in order to supply the name of the AVS network to load. It is at this stage that the user can choose either to pre-process (filter) data from a large data set or to post-process (visualize filtered) meta-data from a previous filtering run.

In this case, the user has chosen to pre-process using the change-point filter. The user specifies any desired system characteristics (such as the name of the output database to create or to which to add data). The screen setup for this filter process is shown in Figure 1. This figure shows the actual AVS network for the filter along with parameters available for user specification. Here the user has specified that the variable to monitor is the "Total water vapor along LOS" where LOS stands for "line-of-sight." The time variable is chosen to be the "Time offset from base time." There are also default ranges of variables prearranged within the filter code. These ranges can be changed by the user.

Once the user is satisfied with the setup parameters and files, the user selects "Extract Feature and Monitor" in order to begin pre-processing the data with the filter (in this case, the change-point filter) and to be able to watch this filtering process. Figure 2 is a snapshot during the pre-processing phase. The screen is updated as the data are processed. The user can watch the box-plots and histogram data being displayed in order to select new variables ranges of interest for the variables. This is particularly useful when the user wants to insure a particular amount of data filtering. The pre-processing stage may be rerun once the user is satisfied with the variables and ranges selected and the level of data filtering achieved. The pre-processing may also be run for a different set of variables, ranges or filtering process. The filters generate meta-data about the raw data which is stored in the POSTGRES database.

Once the pre-processing is completed, the meta-data is available to the the user via the AVS network interface immediately or for use at a later time.

In the post-processing phase, the goal is to actually utilize the now existent filter generated meta-data to examine relationships or trends. The user is also able to selectively examine the original raw data for features within the meta-data.

To utilize a post-processing module of the system, the user returns to the AVS Network Editor screen and again selects "Network Tools" and then "Read Network." This time the user specifies one of the post-processing systems. Figure 3 shows the AVS network for the situation where the user has selected the "newer changept postgres" post-processor. The database, class names, and variable names are specified as part of the module "New\_Read\_Postgres.DB."

The user selects the button "new\_choose\_point" to begin the post-processing. In Figure 4 the output from the module is depicted (normally in color, but for this paper in grey scale). Different colors/shades represent levels of values for the second variable. Here the user is viewing "Statistical" value versus "Event Time" where the colors/shadings relate to the values for the

third variable "Time Between Events." Not all of the information may be able to displayed in a useful fashion on just one single page, hence there is a slider that allows the user to browse (move through) the information in the display window. There is also a "Swap Axes" button that allows the user to swap which variable values will be used for the x-axis. By selecting the button, we switch the plot to view "Statistical" value versus "Time Between Intervals" where the colors/shadings relate to the values for the variable "Event Time." This view is shown in Figure 5.

The user may examine the raw data that generated a particular point simply by clicking on that point. For example, in Figure 5 there is a very unusual point in the upper right hand corner of the window. We click on this point and are shown the view in Figure 6. When the user clicked on the point, the module generated appropriate POSTGRES commands to query the database for location information of the raw data. This location information is used to access and view the corresponding original raw data which may have resided in a different database or in a large file of some format.

This is the power point of the system, in that the user may peruse a small percentage of data and yet still be able to examine interesting or unusual events from the raw data itself without having to download the entire large data set (which may not even be physically possible).



Figure 1: AVS Network for Filter Pre-Processing

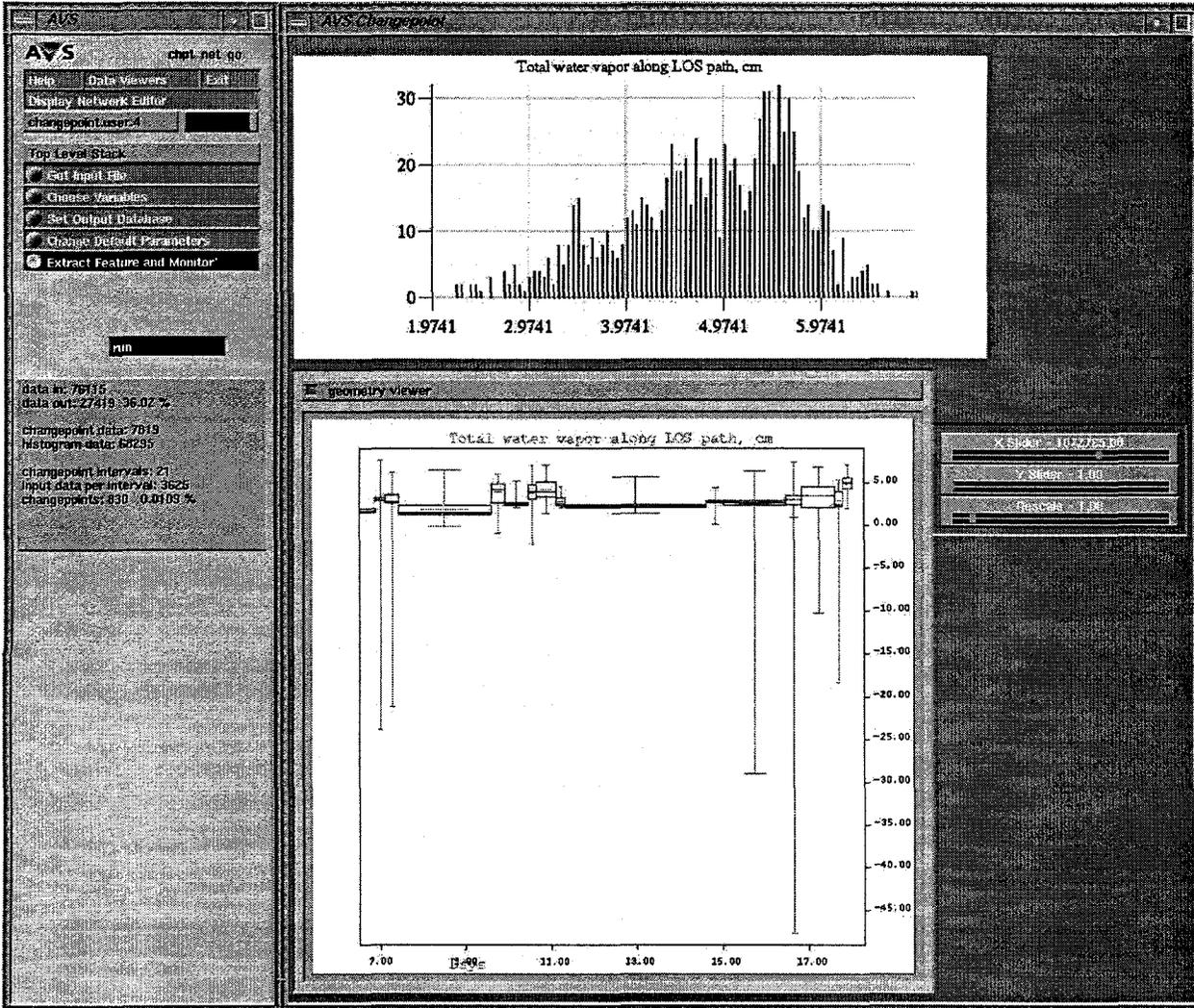


Figure 2: Snapshot of the Pre-Processing Phase



Figure 3: AVS Network for Post-Processor

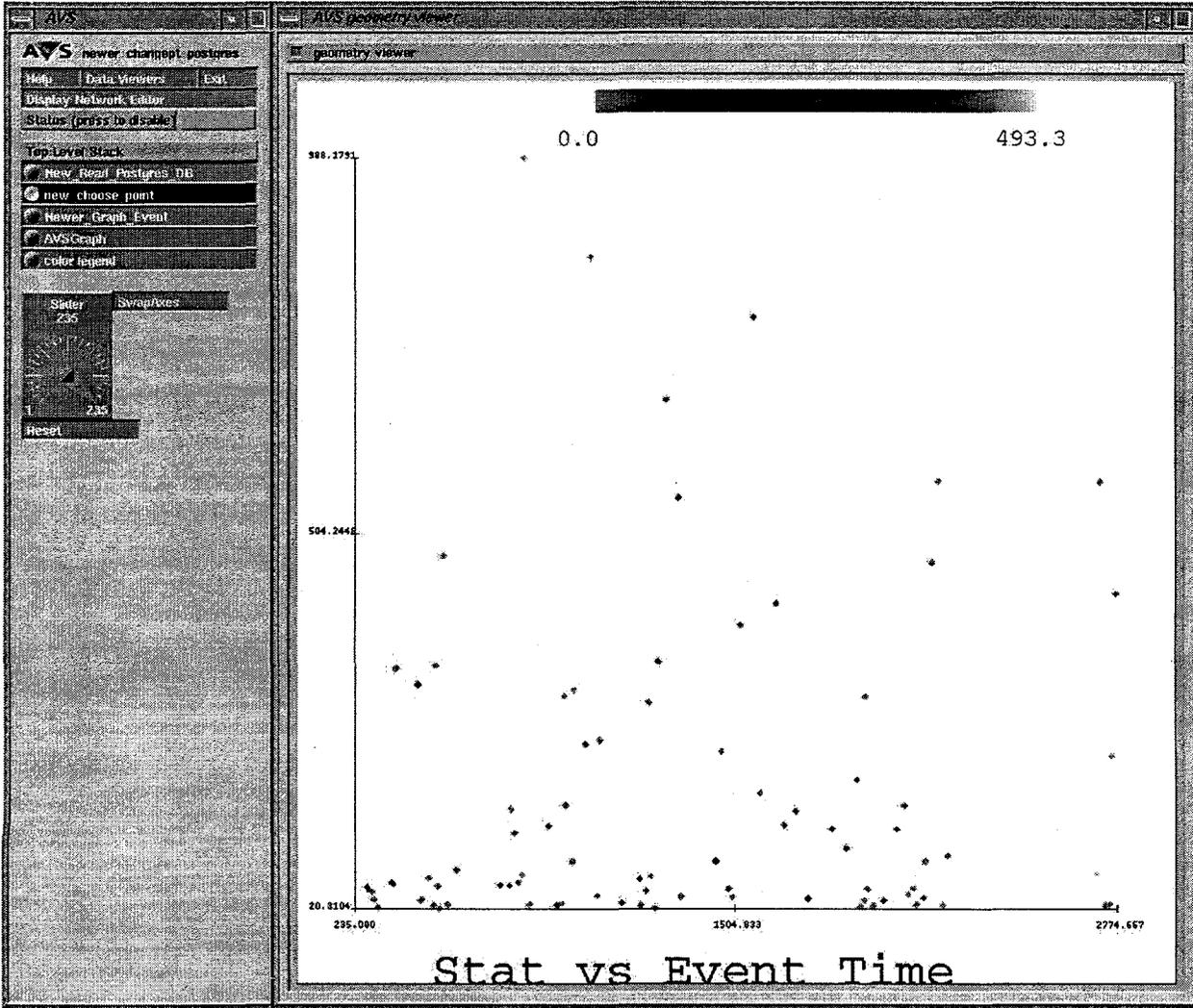


Figure 4: One View of Post-Processing Output

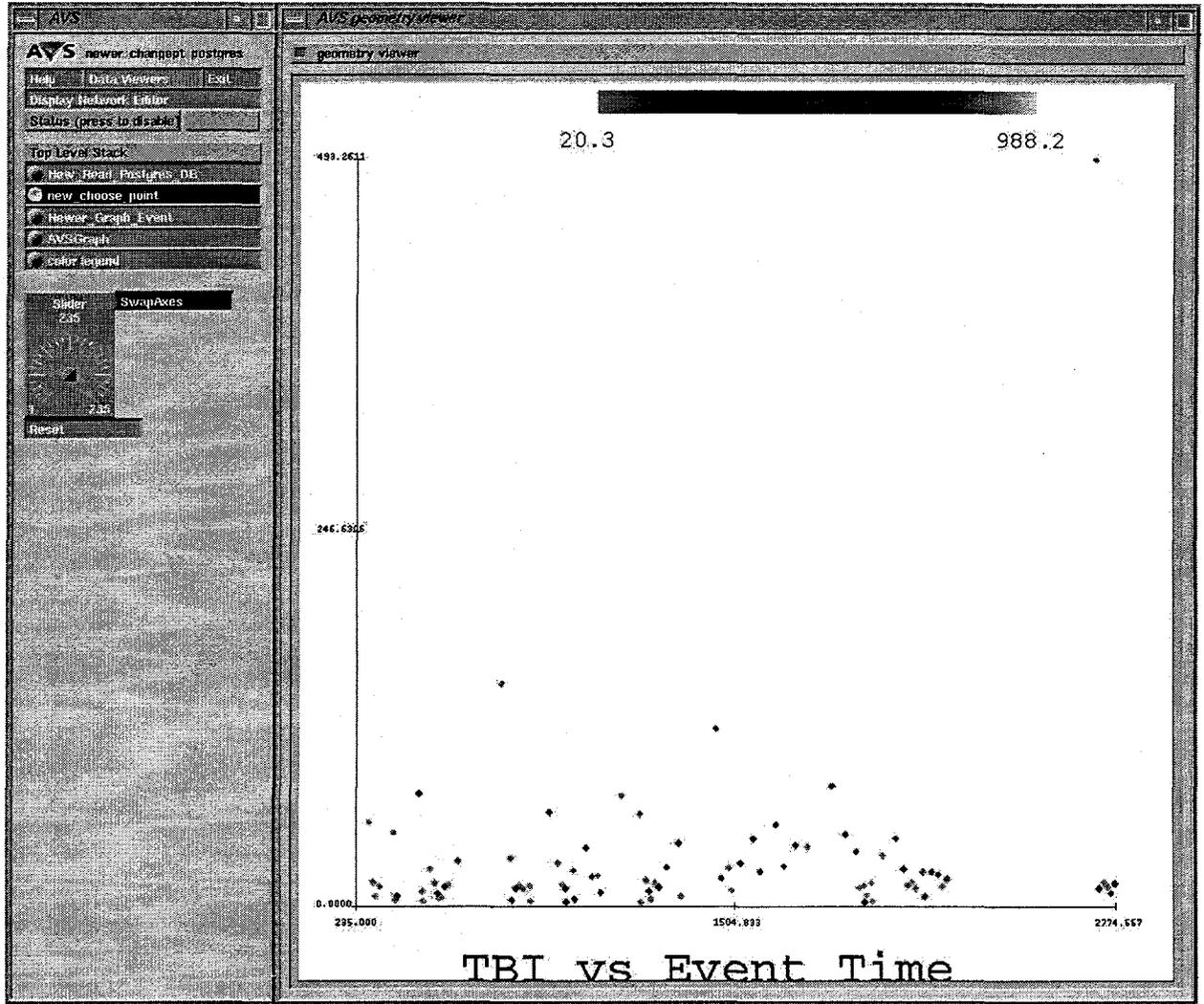


Figure 5: An Alternate View of Post-Processing Output

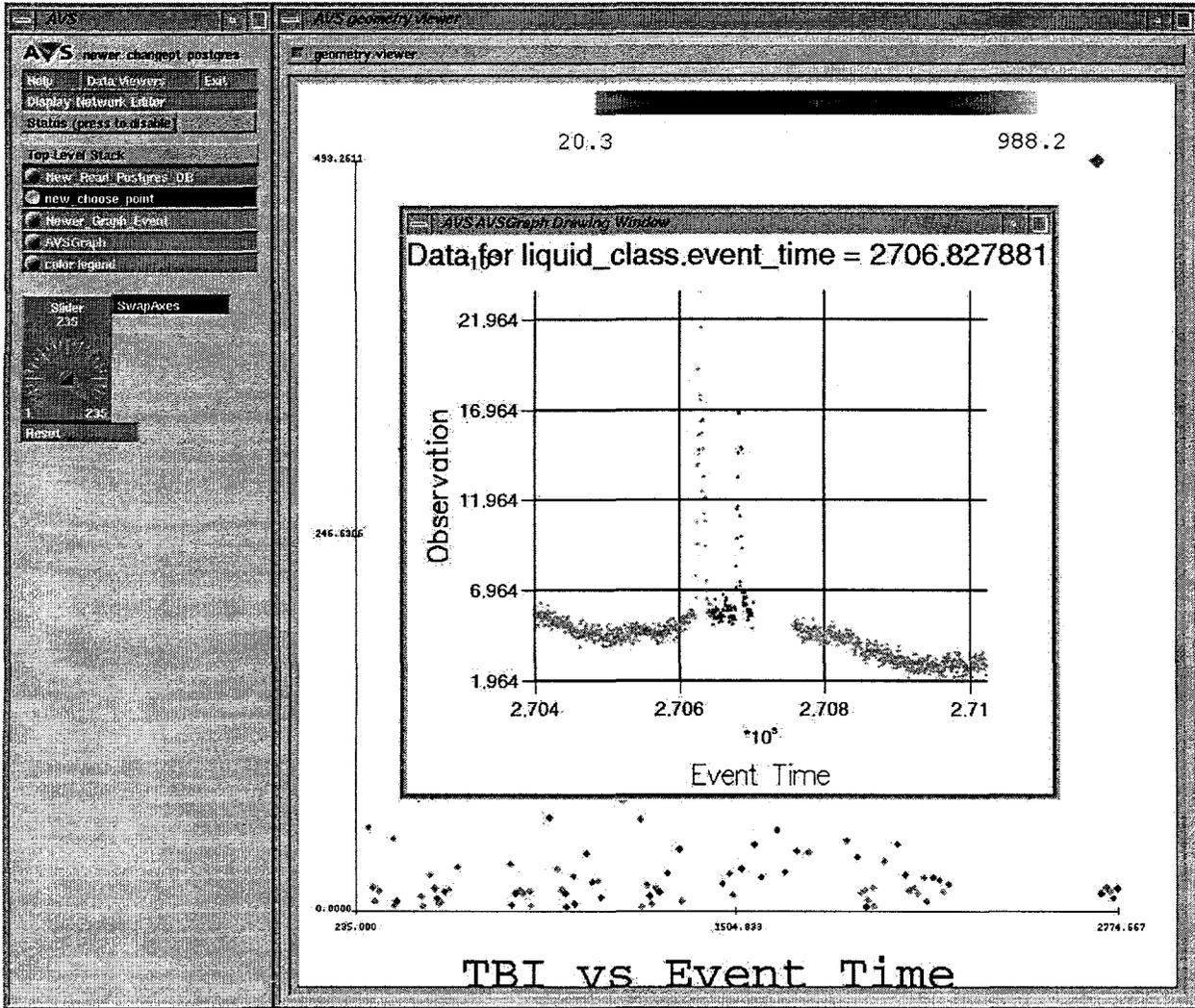


Figure 6: View of Raw Data Represented by Chosen Data Point

## 5. Summary and Future Goals

In this paper we described the visualization approach used as part of the Large Scale Data Analysis project. Statistical analysis and graphical browsing tools for the data sets were implemented as networks within the AVS5 software system. These networks interfaced directly to the POSTGRES database management system to display the statistical metadata and to gain access to the the original raw data. These networks provide uniform graphical interfaces for both the feature extraction stage and the post processing stage of the entire system. A separate module was also created to implement a two-dimensional browser for the large information data set. We utilized the POSTGRES back-end capabilities to provide as much processing on the server machine as possible. This frees resources on the user's workstation (such as an SGI machine) to allow for the rapid rendering and perusal of the metadata and data.

In the statistical analysis and visual representation of large scientific data sets this project has already made a number of successful steps. However, this system is still an initial prototype. There are a number of features still to be designed and implemented. In some cases, we would like to completely revamp certain mechanisms. In particular, we would like to replace AVS5 by the more recent AVS/EXPRESS system in order to gain a number of interface and module implementation features.

We would also like to implement a PVM (Parallel Virtual Machine) or MPI (Message Passing Interface) version of this system to generically provide parallel processing of the filtering and analysis tools. This would allow the system to be used on parallel and heterogeneous distributed machines, including the currently available parallel supercomputers, such as the Intel Paragon.

Most of all, we would like to extend the system, both in terms of feature extraction and visual capabilities, to more complex data sets, such as three-dimensional data, and to multivariate statistical analyses.

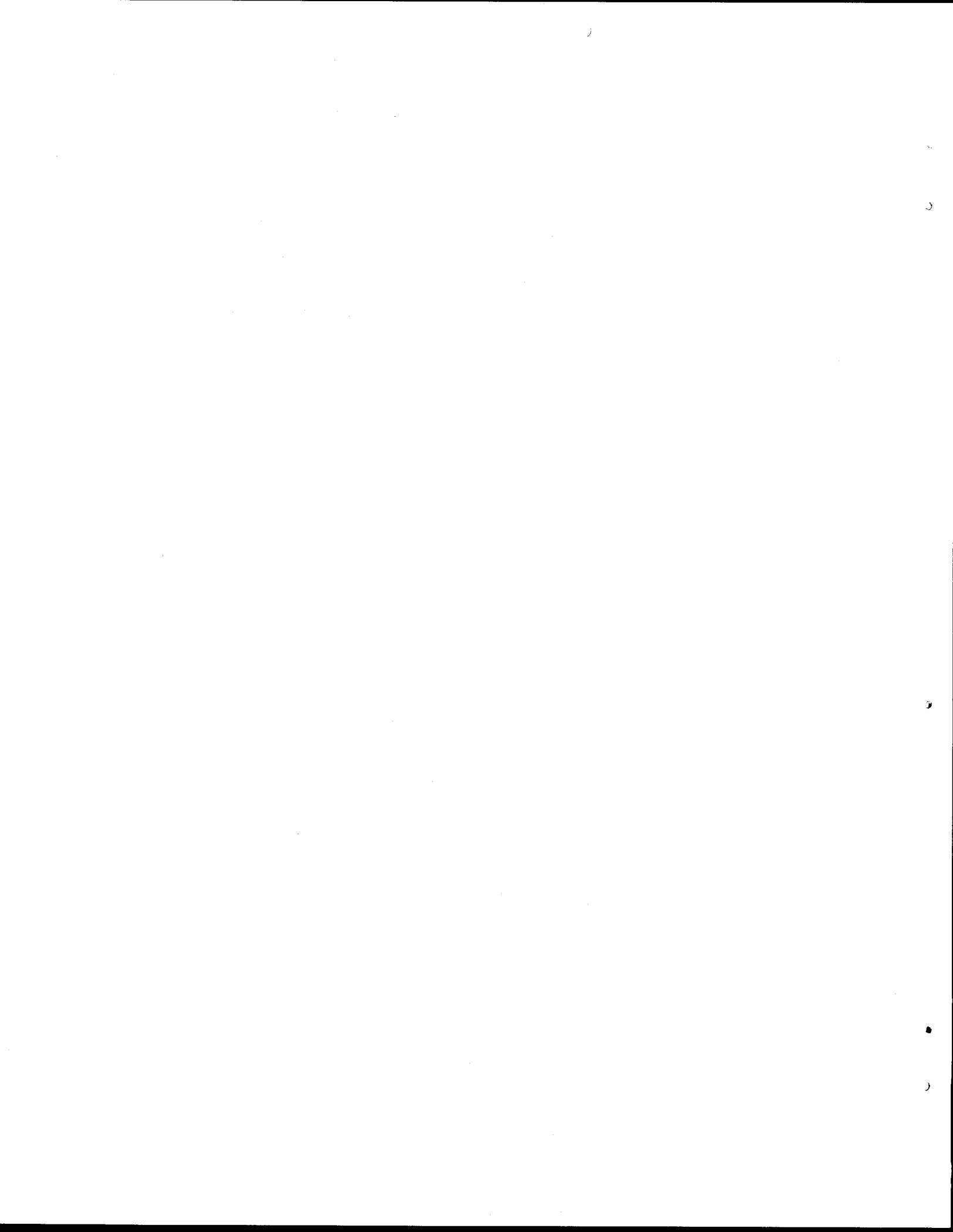
## Acknowledgments

This project was initially funded as a Laboratory Director's Research and Development (LDRD) project under the direction of Darryl J. Downing. Members of the statistical analysis team for this project are Darryl J. Downing, V.V. Fedorov, W.F. Lawkins, M.D. Morris, and G. Ostrouchov.

We also take this moment to remember Toby Mitchell.

## 6. References

- [1] Advanced Visual Systems, Inc. *AVS Developer's Guide*, 1992, Waltham, MA.
- [2] D.J. Downing, W.J. Lawkins, M.D. Morris, and G. Ostrouchov, *FEaTureS - Feature Extraction for Long Time Series: User's Guide*, Oak Ridge National Lab, technical report.
- [3] P. Kochevar, Z. Ahmed, J. Shade, and C. Sharp, *A Simple Visualization Management System: Bridging the Gap Between Visualization and Data Management*, San Diego Supercomputer Center, in *Proceedings Visualization '93*, edited by G.M. Nielson, April 30, 1993.
- [4] M. Stonebraker and G. Kemnitz, *The POSTGRES Next Generation DBMS*, EECS Department, University of California, Berkeley.
- [5] M. Stonebraker and L.A. Rowe, *The Design of POSTGRES*, EECS Department, University of California, Berkeley.



**INTERNAL DISTRIBUTION**

- |                      |                                |
|----------------------|--------------------------------|
| 1-2. T. S. Darland   | 25. C. E. Oliver               |
| 3. E. F. D'Azevedo   | 26. G. Ostrouchov              |
| 4-8. J. M. Donato    | 27-31. S. A. Raby              |
| 9. D. J. Downing     | 32. B. D. Semeraro             |
| 10. G. A. Geist      | 33-37. R. F. Sincovec          |
| 11. N. W. Grady      | 38. D. R. Tufano               |
| 12. V. V. Fedorov    | 39. Central Research Library   |
| 13-17. R. E. Flanery | 40. ORNL Patent Office         |
| 18. W. F. Lawkins    | 41. K-25 Appl Tech Library     |
| 19-23. M. R. Leuze   | 42. Y-12 Technical Library     |
| 24. M. D. Morris     | 43. Lab Records Dept - RC      |
|                      | 44-45. Laboratory Records Dept |

**EXTERNAL DISTRIBUTION**

46. Dr. Dan Hitchcock ER-31, Mathematical, Information, & Computational Sciences Div. Office of Computational & Technology Research Office of Energy Research, U.S. Department of Energy Washington, DC 20585
47. Office of Assistant Manager for Energy Research and Development, Department of Energy, Oak Ridge Operations Office, P.O. Box 2001 Oak Ridge, TN 37831-8600
- 48-49. Office of Scientific & Technical Information, P.O. Box 62, Oak Ridge, TN 37831