# REAL TIME DATA WAREHOUSING AND ON LINE ANALYTICAL PROCESSING AT ABERDEEN TEST CENTER

# REAL TIME DATA WAREHOUSING AND ON LINE ANALYTICAL PROCESSING AT ABERDEEN TEST CENTER

**Michael J. Reil**
**SFA Inc. / Aberdeen Test Center**

**T. George Bartlett**
**Aberdeen Test Center**

**Kevin Henry**
**Sverdrup Technology Inc. / Aberdeen Test Center**

## ABSTRACT

This paper is a follow on to a paper presented at the 2005 International Telemetry Conference by Dr. Samuel Harley et. al., titled *Data, Information, and Knowledge Management.* This paper will describe new techniques and provide further detail into the inner workings of the VISION (Versatile Information System – Integrated, Online) Engineering Performance Data Mart.

## KEY WORDS

VISION, HDF5, data warehouse, data mart, On Line Analytical Processing (OLAP)

## INTRODUCTION

VISION embodies a comprehensive, holistic, top down approach to information collection, management, and ultimate transformation into knowledge. The Aberdeen Test Center has developed VISION to meet the increasing information demands that arise from the testing of complex military combat systems. One of the data marts managed under VISION is the Engineering Performance Data Mart, which was designed to manage developmental and operational engineering test data. This data mart will be the topic of this paper. Figure 1 illustrates the overall VISION data management scheme, with the Engineering Performance Data Mart circled.

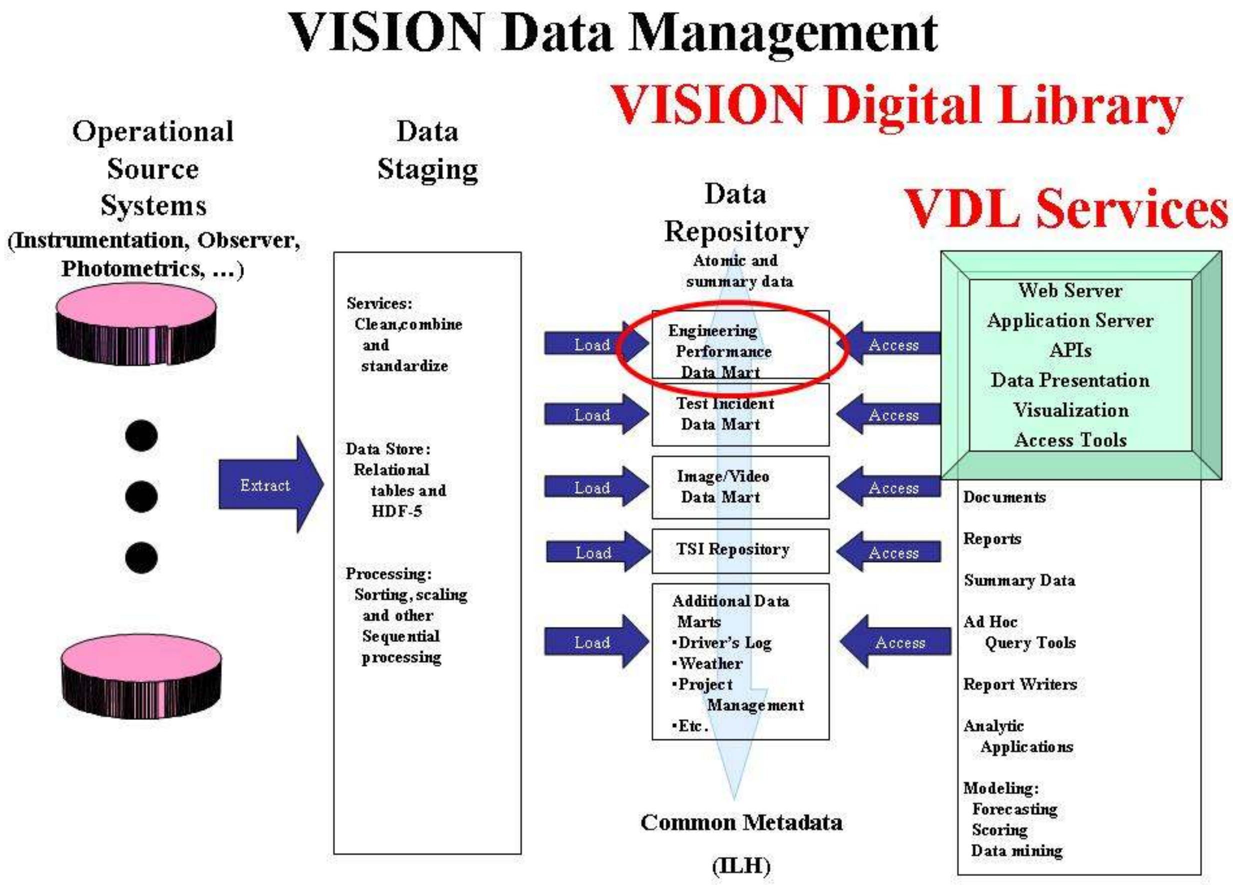


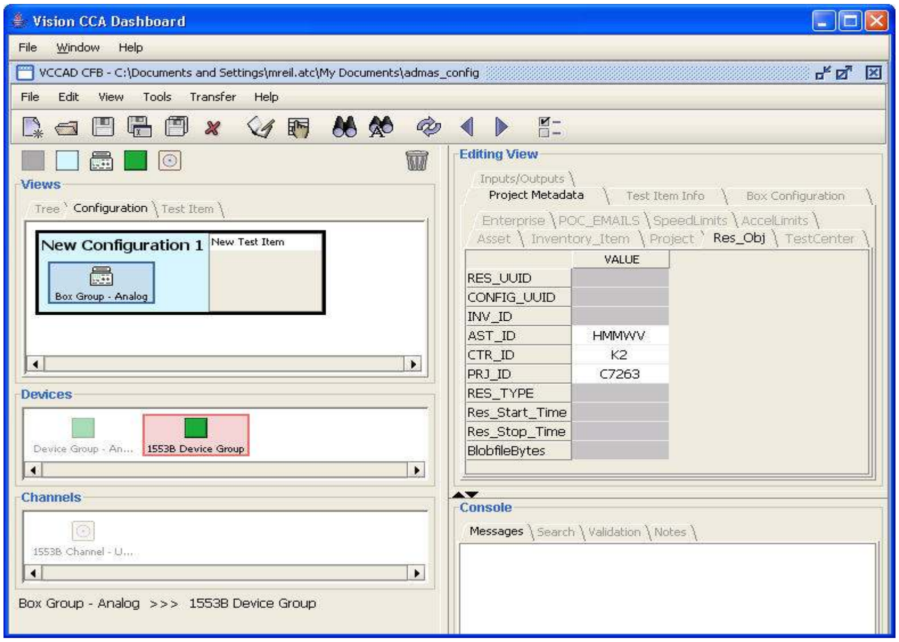
**Figure 1 VISION Data Management**

## PRE TEST SETUP

To gain the maximum ROI (Return on Investment) from test data, it must be tagged with all applicable metadata. Data consumers, which include analysts, modelers, simulation event planners and as yet unknown future users, will be far disconnected from the actual test that produced the data. Their only reliable means of placing the data into context is via the associated metadata. The VISION instrumentation suite of Advanced Distributed Modular Acquisition System (ADMAS) enforces this requirement by means of an XML based configuration file. The configuration file serves multiple purposes:

- Program the ADMAS for data collection.
- Physically couple the metadata with the data (if you have one, you have the other).
- Control the post-processing of the data for entry into the data mart.

A ‘Configuration File Builder’ application has been developed to assist the test engineer with this task. This application pulls what metadata it can from the project administrative databases, making it easier for the test engineer to enter correct values. There are a few required pieces of metadata, including Project ID, System ID, Test Item ID and Test Location. Figure 2 shows a screenshot of the ‘Configuration File Builder’ application.


Vision CCA Dashboard
File Window Help
VCCAD CFB - C:\Documents and Settings\mreil.atc\My Documents\admas_config
File Edit View Tools Transfer Help
Views
Tree Configuration Test Item
New Configuration 1
New Test Item
Box Group - Analog
Devices
Device Group - An...
1553B Device Group
Channels
1553B Channel - U...
Box Group - Analog >>> 1553B Device Group
Editing View
Inputs/Outputs
Project Metadata
Test Item Info
Box Configuration
Enterprise POC_EMAILS SpeedLimits AccelLimits
Asset Inventory_Item Project Res_Obj TestCenter
Console
Messages Search Validation Notes

| | VALUE |
|---|---|
| RES_UUID | |
| CONFIG_UUID | |
| INV_ID | |
| AST_ID | HMMWV |
| CTR_ID | K2 |
| PRJ_ID | C7263 |
| RES_TYPE | |
| Res_Start_Time | |
| Res_Stop_Time | |
| BlobfileBytes | |

**Figure 2 Configuration File Builder**

## RAW DATA FILE COLLECTION

During the test event, the ADMAS records applicable ‘Run Time’ metadata such as date/time, GPS position, operator’s name etc. The pre-test entered metadata, the engineering data itself, and any run time collected metadata are stored in a single file (a BLOb – Binary Large Object) by the ADMAS. The format of the raw data is optimized for the real-time requirements of the ADMAS. The metadata however, is in the well known XML format at the head of every ADMAS BLOb file. Figure 3 shows an ADMAS and the layout of all ADMAS BLOb files.

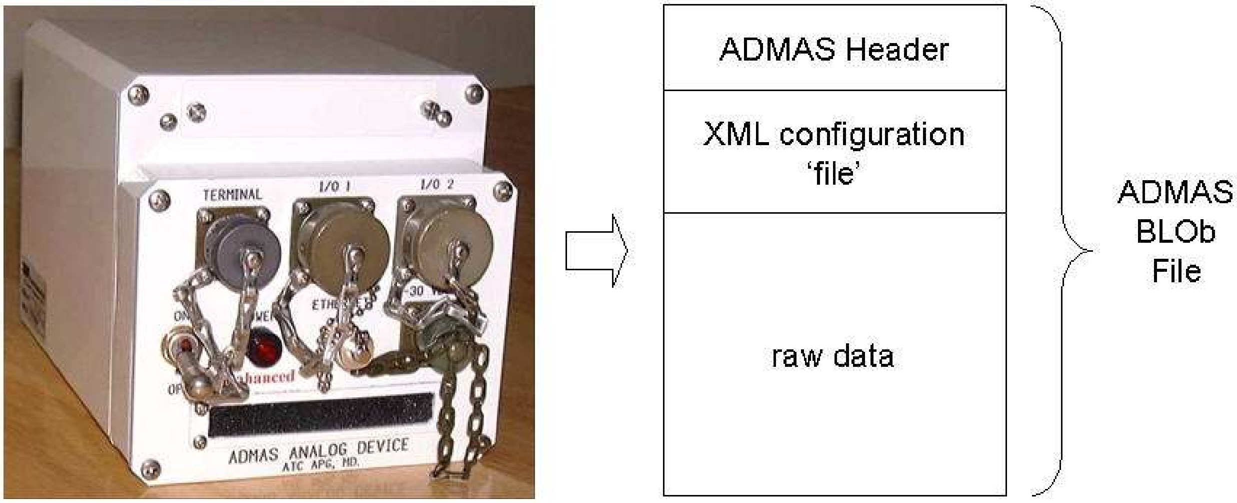


**Figure 3 ADMAS and BLOb file layout**

## DATA HARVESTING

ADMAS' record BLOb files on PC card flash drives. BLOb files are harvested from the ADMAS via either wireless LAN, hard wire LAN or by physically removing the flash cards from the ADMAS and inserting them into a networked PC. The raw BLOb files are transferred via SSL to the data mart for processing. The method of transport is not really important, and can be customized to suit each range's unique networking requirements. The only real requirement is that the files somehow make it to the data mart. Figure 4 shows the generic process.

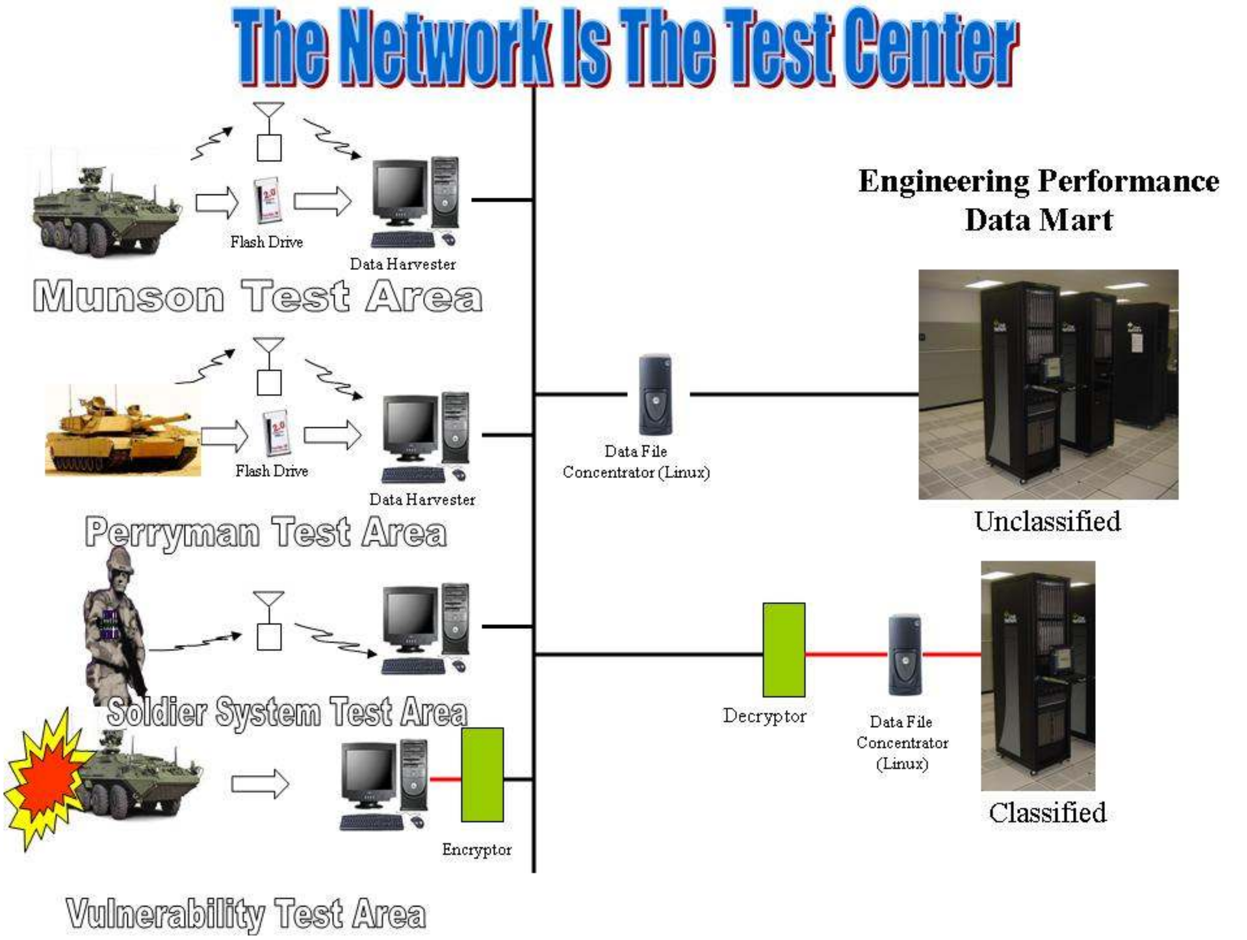


**Figure 4 Data Harvesting**

## DATA TRANSFORMATION AND HDF5

Unattended, triggered processes run on the data mart to convert the raw data to processed data, which is stored in HDF5 files. HDF5 (http://hdf.ncsa.uiuc.edu/HDF5/) is developed and maintained by NCSA, and is ideally suited for storing large scientific data sets. Generally, one HDF5 file is produced for each raw data file. The processing instructions in the raw file's XML configuration section control what additional processing steps are done, such as validation, reporting etc.

The processes are triggered by the arrival of new raw data files using 'AntFlow' (http://onionnetworks.com/products/antflow/), which is an open source framework for workflow automation based on hot folders. Ant (http://ant.apache.org/) is a java based open source framework for creating tasks. These tasks are run by AntFlow when new files appear in the hot folders. The tasks run the java applications, which do the converting, validating and reporting. A typical data flow process is shown in figure 5. This follows the 'Extract, Transform and Load' data warehousing technique (http://en.wikipedia.org/wiki/Extract,_transform,_load).

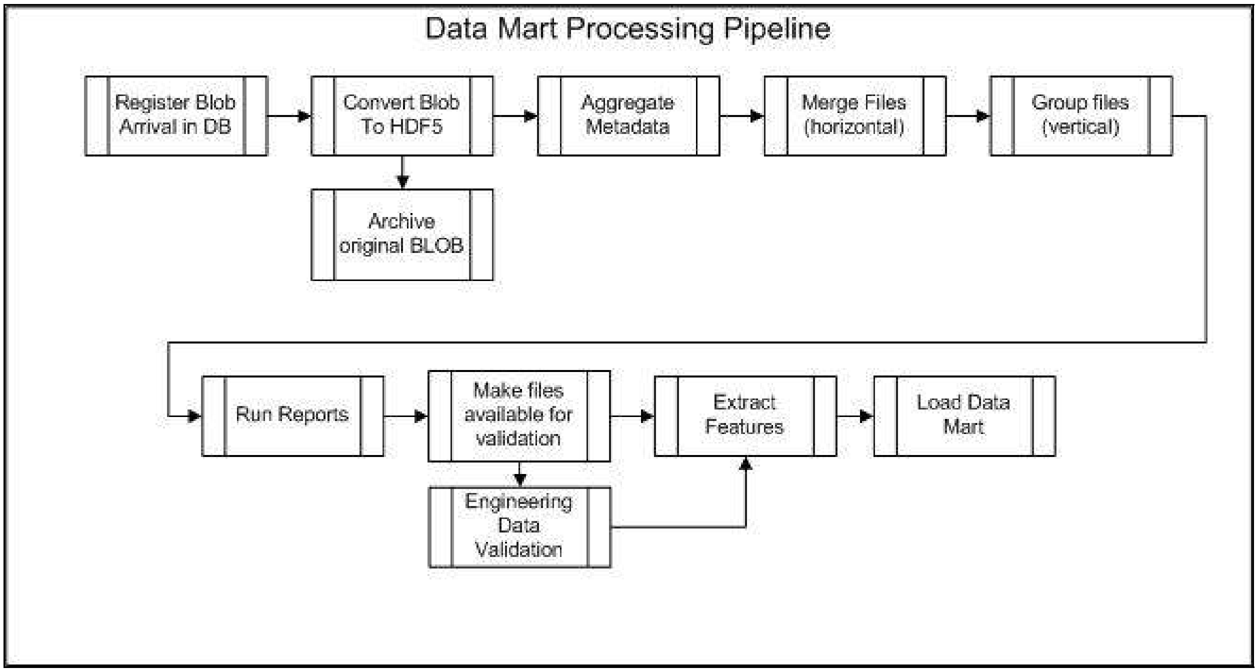


**Figure 5**

## DATA MART POPULATION

The Engineering Performance Data Mart consists of both a relational database and a collection of HDF5 files. All of the metadata and none of the data are stored in the relational database. The data is kept in the HDF5 files, with a pointer from the relational database to the individual HDF5 files. This technique keeps the database small, and therefore queries against it run very fast. Loading of the data mart is done via a java application that reads the metadata from the HDF5 files, connects to the relational database, and issues the appropriate SQL "Insert …"statements to populate the tables. The HDF5 files are then moved to the distributed storage pool, where the OLAP tools can find them.

## DATA MART PHYSICAL MAKEUP

The Engineering Performance Data Mart is comprised of two Linux servers, a Linux cluster, and a large distributed storage pool. The data mart components consist of the following:

- Hardware
    - One Opteron dual processor CentOS-4 database server
    - One Opteron dual processor CentOS-4 application server
    - Distributed File System
        - One Opteron dual processor CentOS-4 Lustre metadata server
        - Six Opteron dual processor CentOS-4 Lustre object storage servers with 6.4 terabyte storage each (38.4 TB total).
    - One Linux cluster comprised of one head node and 32 slave nodes, each Opteron dual processors. Gigabit Ethernet interconnect.
- Software
    - In house developed java classes
    - PostgreSQL 8.1 (http://www.postgresql.org/)
    - Apache Tomcat 5.5 application server (http://tomcat.apache.org/)
    - Lustre 1.4.6 distributed file system (http://www.lustre.org/).

Note that the software stack of the data mart is made up of open source software and the hardware stack is using 64 bit processors. The cluster is used to populate and query the data mart using distributed parallel processing. The Lustre file system stores the HDF5 files.

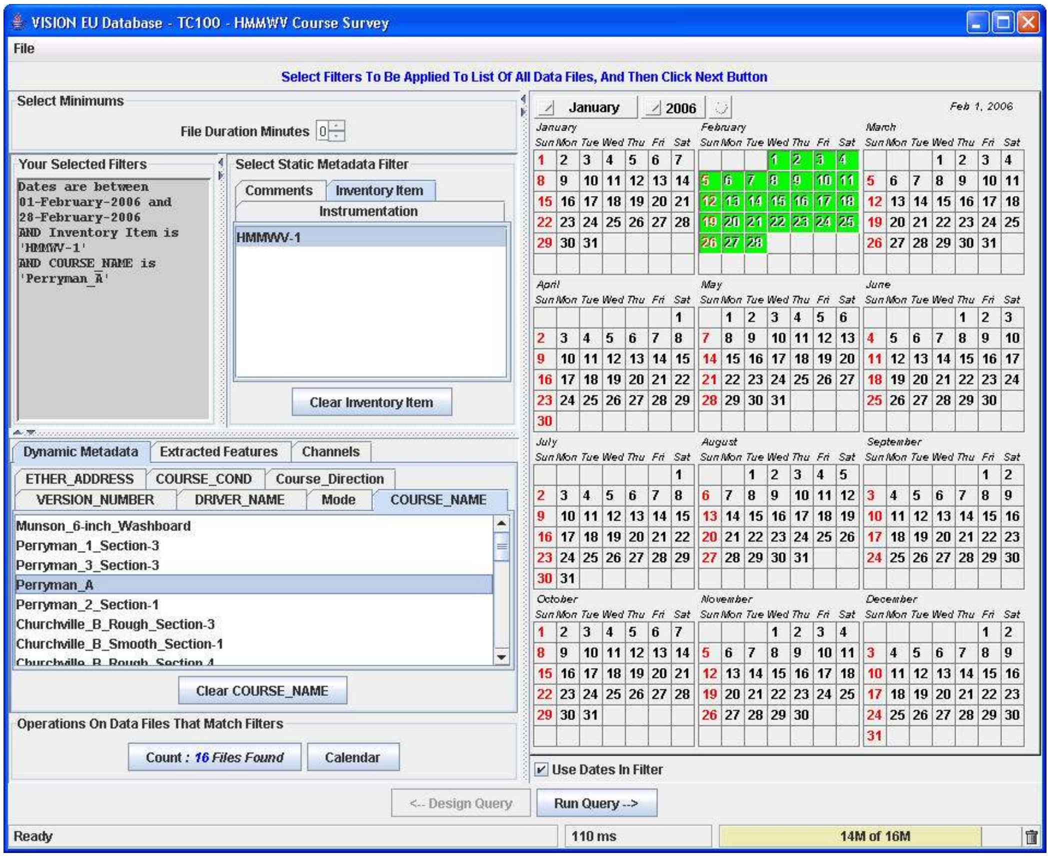
## ON LINE ANALYTICAL PROCESSING TOOLS

A web based OLAP toolbox has been developed to allow querying of the data mart. Access to the Engineering Performance Data Mart – which is a web based application – is via the VISION Digital Library (VDL) - as is access to all VISION data marts. This single sign-on allows the project coordinators to control and track access to the data marts. By logging into VDL and navigating to the project of interest, the user would click on the data mart URL (assuming permission), thus launching the OLAP toolbox in his browser. Tools have been developed which allow the user to view aggregate composites, or to drill down to individual data files and parameters. The user builds queries to submit to the data mart via the java GUI, by selecting metadata values that are presented as pick lists. The more metadata that was collected before and during the test, the better the OLAP user will be able to define his/her query. Some of the things that users can do include:

- View all of the collected metadata tags, and all of the different values used for these tags
- View how many files are available, and what time period they cover, for any combination of metadata
- View a GPS map of the location of the item under test for any combination of metadata
- Apply metadata filters to the query results
- View time series traces of all parameters
- Create custom plots of parameters from a pick list

- Save plots as JPEG files on users computer
- Download raw files, HDF5 files or Excel spreadsheets of selected data.
- Create composite histograms, time in limits, threshold crossing or trend reports across multiple files.
- Download Google Earth (http://earth.google.com/) files for displaying GPS locations during the entire test, or just select portions of the test.

Using the OLAP GUI, and the power of the Linux cluster, all of these tasks can be initiated by the user, run on the cluster, and the results sent back to the user in seconds. The vast majority of queries return to the user in under 5 seconds. To test the performance of the data mart, we ran a composite histogram query of an acceleration parameter across all files in an 8 month duration test (38000 data files, with 5.5 billion samples). The histogram plot was displayed in less than 2 minutes. Screenshots of the OLAP GUI are shown in figures 6 through 9.

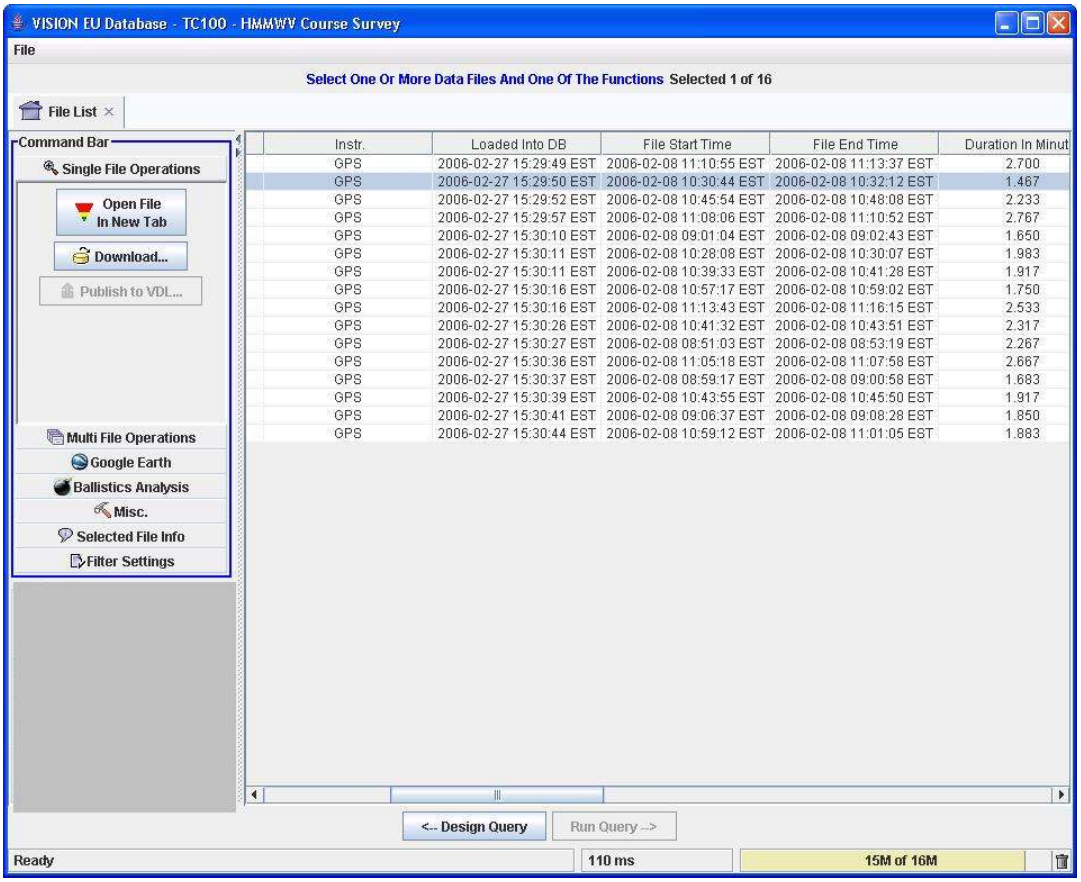

VISION EU Database - TC100 - HMMWV Course Survey
File
Select Filters To Be Applied To List Of All Data Files, And Then Click Next Button
Select Minimums
File Duration Minutes 0
Your Selected Filters
Select Static Metadata Filter
Comments
Inventory Item
Instrumentation
Clear Inventory Item
Dynamic Metadata
Extracted Features
Channels
ETHER_ADDRESS
COURSE_COND
Course_Direction
VERSION_NUMBER
DRIVER_NAME
Mode
COURSE_NAME
Clear COURSE_NAME
Operations On Data Files That Match Filters
Count : 16 Files Found
Calendar
January
2006
Feb 1, 2006
January Sun Mon Tue Wed Thu Fri Sat 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31
February Sun Mon Tue Wed Thu Fri Sat 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28
March Sun Mon Tue Wed Thu Fri Sat 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31
April Sun Mon Tue Wed Thu Fri Sat 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
May Sun Mon Tue Wed Thu Fri Sat 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31
June Sun Mon Tue Wed Thu Fri Sat 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
July Sun Mon Tue Wed Thu Fri Sat 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31
August Sun Mon Tue Wed Thu Fri Sat 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31
September Sun Mon Tue Wed Thu Fri Sat 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
October Sun Mon Tue Wed Thu Fri Sat 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31
November Sun Mon Tue Wed Thu Fri Sat 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
December Sun Mon Tue Wed Thu Fri Sat 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31
☑ Use Dates In Filter
<-- Design Query
Run Query -->
Ready
110 ms
14M of 16M

```
Dates are between
01-February-2006 and
28-February-2006
AND Inventory Item is
'HMMWV-1'
AND COURSE_NAME is
'Perryman_A'
```

HMMWV-1

Munson_6-inch_Washboard
Perryman_1_Section-3
Perryman_3_Section-3
Perryman_A
Perryman_2_Section-1
Churchville_B_Rough_Section-3
Churchville_B_Smooth_Section-1
Churchville_B_Rough_Section-4

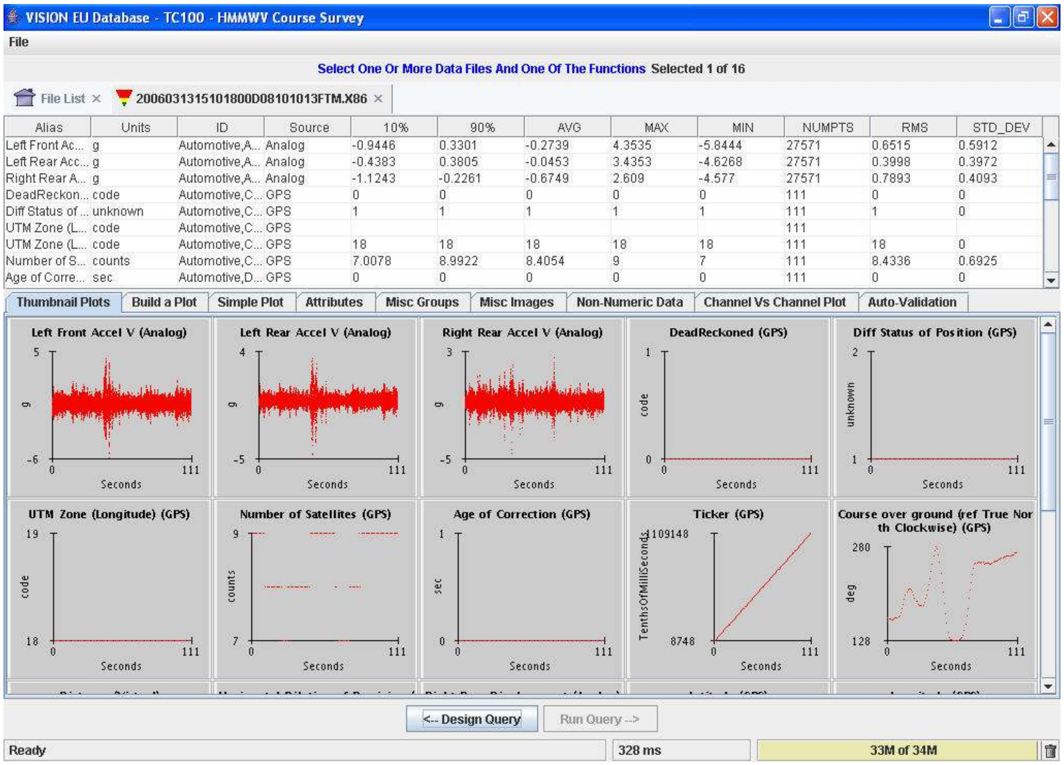
**Figure 6 – Metadata filter page**



Select One Or More Data Files And One Of The Functions Selected 1 of 16

| Instr. | Loaded Into DB | File Start Time | File End Time | Duration In Minut |
|---|---|---|---|---|
| GPS | 2006-02-27 15:29:49 EST | 2006-02-08 11:10:55 EST | 2006-02-08 11:13:37 EST | 2.700 |
| GPS | 2006-02-27 15:29:50 EST | 2006-02-08 10:30:44 EST | 2006-02-08 10:32:12 EST | 1.467 |
| GPS | 2006-02-27 15:29:52 EST | 2006-02-08 10:45:54 EST | 2006-02-08 10:48:08 EST | 2.233 |
| GPS | 2006-02-27 15:29:57 EST | 2006-02-08 11:08:06 EST | 2006-02-08 11:10:52 EST | 2.767 |
| GPS | 2006-02-27 15:30:10 EST | 2006-02-08 09:01:04 EST | 2006-02-08 09:02:43 EST | 1.650 |
| GPS | 2006-02-27 15:30:11 EST | 2006-02-08 10:28:08 EST | 2006-02-08 10:30:07 EST | 1.983 |
| GPS | 2006-02-27 15:30:11 EST | 2006-02-08 10:39:33 EST | 2006-02-08 10:41:28 EST | 1.917 |
| GPS | 2006-02-27 15:30:16 EST | 2006-02-08 10:57:17 EST | 2006-02-08 10:59:02 EST | 1.750 |
| GPS | 2006-02-27 15:30:16 EST | 2006-02-08 11:13:43 EST | 2006-02-08 11:16:15 EST | 2.533 |
| GPS | 2006-02-27 15:30:26 EST | 2006-02-08 10:41:32 EST | 2006-02-08 10:43:51 EST | 2.317 |
| GPS | 2006-02-27 15:30:27 EST | 2006-02-08 08:51:03 EST | 2006-02-08 08:53:19 EST | 2.267 |
| GPS | 2006-02-27 15:30:36 EST | 2006-02-08 11:05:18 EST | 2006-02-08 11:07:58 EST | 2.667 |
| GPS | 2006-02-27 15:30:37 EST | 2006-02-08 08:59:17 EST | 2006-02-08 09:00:58 EST | 1.683 |
| GPS | 2006-02-27 15:30:39 EST | 2006-02-08 10:43:55 EST | 2006-02-08 10:45:50 EST | 1.917 |
| GPS | 2006-02-27 15:30:41 EST | 2006-02-08 09:06:37 EST | 2006-02-08 09:08:28 EST | 1.850 |
| GPS | 2006-02-27 15:30:44 EST | 2006-02-08 10:59:12 EST | 2006-02-08 11:01:05 EST | 1.883 |

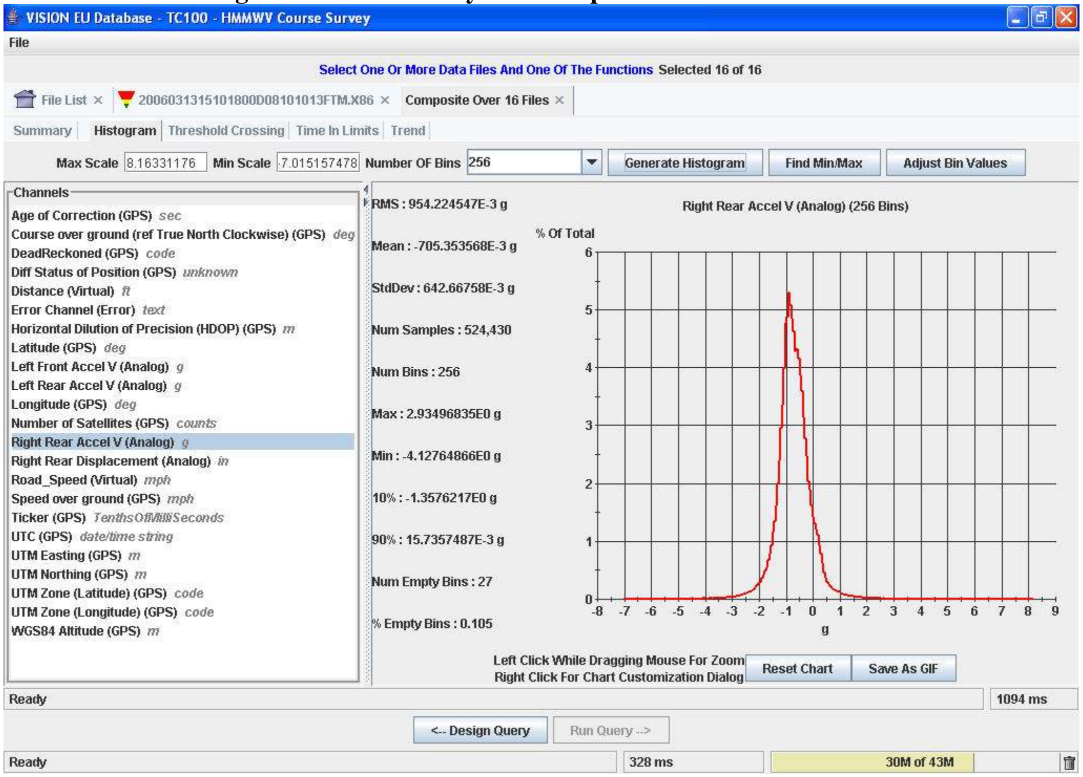
**Figure 7 – List of files returned from query**



Select One Or More Data Files And One Of The Functions Selected 1 of 16

| Alias | Units | ID | Source | 10% | 90% | AVG | MAX | MIN | NUMPTS | RMS | STD_DEV |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Left Front Ac... | g | Automotive,A... | Analog | -0.9446 | 0.3301 | -0.2739 | 4.3535 | -5.8444 | 27571 | 0.6515 | 0.5912 |
| Left Rear Acc... | g | Automotive,A... | Analog | -0.4383 | 0.3805 | -0.0453 | 3.4353 | -4.6268 | 27571 | 0.3998 | 0.3972 |
| Right Rear A... | g | Automotive,A... | Analog | -1.1243 | -0.2261 | -0.6749 | 2.609 | -4.577 | 27571 | 0.7893 | 0.4093 |
| DeadReckon... | code | Automotive,C... | GPS | 0 | 0 | 0 | 0 | 0 | 111 | 0 | 0 |
| Diff Status of ... | unknown | Automotive,C... | GPS | 1 | 1 | 1 | 1 | 1 | 111 | 1 | 0 |
| UTM Zone (L... | code | Automotive,C... | GPS |  |  |  |  |  | 111 |  |  |
| UTM Zone (L... | code | Automotive,C... | GPS | 18 | 18 | 18 | 18 | 18 | 111 | 18 | 0 |
| Number of S... | counts | Automotive,C... | GPS | 7.0078 | 8.9922 | 8.4054 | 9 | 7 | 111 | 8.4336 | 0.6925 |
| Age of Corre... | sec | Automotive,D... | GPS | 0 | 0 | 0 | 0 | 0 | 111 | 0 | 0 |

Figure 8 – Time history traces of parameters from one file


VISION EU Database - TC100 - HMMWV Course Survey

Select One Or More Data Files And One Of The Functions Selected 16 of 16

Max Scale 8.16331176 Min Scale -7.015157478 Number OF Bins 256

Channels

Age of Correction (GPS) *sec*
Course over ground (ref True North Clockwise) (GPS) *deg*
DeadReckoned (GPS) *code*
Diff Status of Position (GPS) *unknown*
Distance (Virtual) *ft*
Error Channel (Error) *text*
Horizontal Dilution of Precision (HDOP) (GPS) *m*
Latitude (GPS) *deg*
Left Front Accel V (Analog) *g*
Left Rear Accel V (Analog) *g*
Longitude (GPS) *deg*
Number of Satellites (GPS) *counts*
Right Rear Accel V (Analog) *g*
Right Rear Displacement (Analog) *in*
Road_Speed (Virtual) *mph*
Speed over ground (GPS) *mph*
Ticker (GPS) *TenthsOfMilliSeconds*
UTC (GPS) *date/time string*
UTM Easting (GPS) *m*
UTM Northing (GPS) *m*
UTM Zone (Latitude) (GPS) *code*
UTM Zone (Longitude) (GPS) *code*
WGS84 Altitude (GPS) *m*

RMS : 954.224547E-3 g
Mean : -705.353568E-3 g
StdDev : 642.66758E-3 g
Num Samples : 524,430
Num Bins : 256
Max : 2.93496835E0 g
Min : -4.12764866E0 g
10% : -1.3576217E0 g
90% : 15.7357487E-3 g
Num Empty Bins : 27
% Empty Bins : 0.105

Left Click While Dragging Mouse For Zoom
Right Click For Chart Customization Dialog

Ready 1094 ms

Ready 328 ms 30M of 43M

Figure 9 – Composite histogram of one parameter across 16 files

## CONCLUSION

The VISION Engineering Units Data Mart at the Aberdeen Test Center contains both developmental and operational test data for over 150 Army, Navy and Marine Corp projects, dating from 1999 to the present. The metadata totals 100 gigabytes and the HDF5 data totals 1.5 terabytes. The data mart has a flexible architecture, allowing for configuration and deployment in a number of different ways. As an example, a remote data mart was set up at “internet challenged” sites at Camp Pendleton, CA and Marine Corp Base Hawaii to support testing of the Expeditionary Fighting Vehicle (EFV) in 2004 through 2005. Because of the flexible configuration and deployment options of the system, the entire data mart was able to run on a single Dell workstation with one 2TB external network attached storage device. The use of well defined interfaces at all of the software tiers allows for additional features to be developed and plugged in with little or no rework to the infrastructure. As an example of this, in 2001 the Department of Transportation wrote a custom graphical analysis component (in java) for one of their projects that conformed to the analysis interface that we published. They gave us the component, we plugged it into the OLAP GUI, and they were able to use it immediately. This allowed them to leverage the data mart’s metadata querying capability into their own analysis tool. Through the use of open source software, notably HDF5, PostgreSQL and Linux along with in house developed cross platform software components (java based), the VISION Engineering Units Data Mart is in an excellent position to meet the unknown challenges of future generations of testing.

## ACKNOWLEDGMENTS

Dr. Samuel F. Harley – Aberdeen Test Center – VISION Technical Manager

## REFERENCES

Harley, Samuel F., Data, Information, and Knowledge Management, International Telemetry Conference, 2005