A Temporal-Spatial Data Fusion Architecture For Monitoring Complex Systems

Kevin McCarty[1], Milos Manic[1], Shane Cherry[2] and Miles McQueen[2]
[1]University of Idaho, Idaho Falls, Idaho USA,
[2]Idaho National Laboratory, Idaho Falls, Idaho USA,
kmccarty@ieee.org, misko@uidaho.edu, Shane.Cherry@inl.gov, Miles.McQueen@inl.gov

Abstract – **Non-homogenous systems arise from the need to incorporate a variety of disparate systems into a cohesive functioning whole and may comprise many crucial elements of an industrialized, modern society. As a result they must be constantly monitored to ensure efficient functioning and avoid expensive breakdowns. In particular, inter-connected computer-based systems must increasingly be aware of cyber and physical threats that are dynamic and evolutionary in nature. However, difficulties arise in trying to ascertain threats and problems among the diverse sources of information generated by these systems. Finally, there is the question of how best to present this data to a human operator. Human systems require not just analysis, but presentation which encourages timely, proactive or corrective decisions. This paper presents a software architecture to solve these problems based upon data fusion using temporal-spatial relationships. As phase one of a three phase project, a prototype implementation of this architecture demonstrates application of this technique for a cohesive system. Test results showed the system capable of real-time fusion of physical, cyber and process data elements as well as analysis, display and interpretation of threats.**

I. Introduction

Modern complex systems, such as a nuclear power plant, consist of a multitude of subsystems, each of which has a specific, often critical, role [1]. Each subsystem attempts to address a particular need of the overall plant. Ensuring component and overall plant health is extremely important for both economic and other reasons:

1. Plant equipment tends to be very expensive to repair and replace.
2. Any plant that is not running at capacity incurs both fixed costs and lost opportunity costs that can run into the many millions of dollars.
3. Due to the hazardous nature of nuclear fuel and waste products, accidents can have devastating economic and environmental consequences.

While failures in these and other large, complex systems tend to be rare, they can also be catastrophic. Costs to repair or replace a typical steam generator can run over a billion dollars, not to include the revenue lost while the system is nonfunctional [2, 3]. Major accidents, such as Chernobyl (1986), France (1992), and Japan (1999) can have far-reaching consequences [4]. More recently, the prospect of terrorism adds an entirely new dimension and impetus for plant safety and security [5].

In order to prevent accidents, equipment malfunctions and acts of sabotage, plant operators rely on a myriad of equipment sensors, networks and various layers of security [1, 5, 6]. Plant subsystems are monitored through use of sensors or other means of observation such as physical inspections. Computer networks employ IDS, encryption and other forms of network security to detect and deter attackers. The physical assets are protected as well by fences, doors and other barriers to entry which employ locks, card swipe systems, fingerprint/retinal scans or other means to allow access to the select few while denying access to others.

Despite that fact that gathering data from plant sensors has been occurring for decades, preventable accidents and malfunctions still happen and security risks continue to be a major concern [1, 2, and 5]. Data repositories are often of limited use even though they are often quite extensive [7]. Some key subsystems, while closely monitored, are still susceptible to malicious outside influence that can be difficult to trace and determine. Even with increased awareness of security, better sensors and computers, databases and analysis tools, there is still a significant gap between available data and the ability to translate that data into actionable knowledge [7, 8].

Recent research into Resilient Control Systems (RCS) undertaken at the Idaho National Laboratory demonstrates the need for determining proper operation of a facility by considering and evaluating all possible threats and measures [6]. The goal of RCS Design is to be able to combine disparate system data for improved state awareness of a larger system.

Better state awareness allows for a more proactive, rather than reactive application of resources for system operation and maintenance. Implementing RCS design requires a much more cohesive approach to data collection along with better ways to combine and evaluate results across different yet interrelated systems. This is necessary due to the greater complexity and sophistication of the modern plant as well as increased threat levels and means of attack.

Traditional techniques to address this problem typically involve collecting the data in some sort of data repository. The repository may consist of a database, such as an ISAM or relational database, or just a set of text or binary files [7]. If the data is retrievable, it is displayed in a number of formats which can range from stacks of paper to an advanced software dashboard.

Advances in database technology have lead many organizations to create a data warehouse, which is often a large repository consisting of relational databases working in combination with multi-dimensional databases [8, 9]. Data in these databases can be “mined” for information

978-1-4244-7561-2/10/$26.00 ©2010 IEEE

either automatically using advanced data mining techniques or manually using reports or spreadsheets [10].

There are significant problems, however, in trying to store, mine and interpret data which comes from many different sources and appears to have little by way of "relateability" to other data.

This paper presents a software architecture and methodology for use in a nuclear plant or other large organization, designed to improve its capability to monitor key subsystems and employ both preventative and reactive techniques to minimize plant disruption and downtime. Section II presents a problem statement. Section III describes the architecture in detail. Section IV presents a prototypical implementation of the architecture. Section V presents conclusions and future work.

II. PROBLEM STATEMENT

A typical nuclear plant will employ a vast array of sensors, computers and other equipment that generate data. Because of the considerable cost of modern plant maintenance and protection, a resilient design is essential [6].

This design must incorporate data that leads to proactive rather reactive control and account for both mechanical and human threats. Consider a plant that gathers data from three major subsystems.

The first is physical security, which records time and activity around various access points throughout the plant. Physical security sensors record access times, physical breaches and the like.

The second is cyber security, which consists of an Intrusion Detection System and network monitoring software of the plant network. Cyber systems record network traffic along with analysis to identify potential security problems and threats.

The third is a chemical mixer, which much precisely maintain both the chemical consistency of a solution and then pump to resulting mixture to a receiving tank. The mixer has sensors which record volume, composition and flow rates. All the major subsystems form an integral part of plant function as shown in Fig. 1.

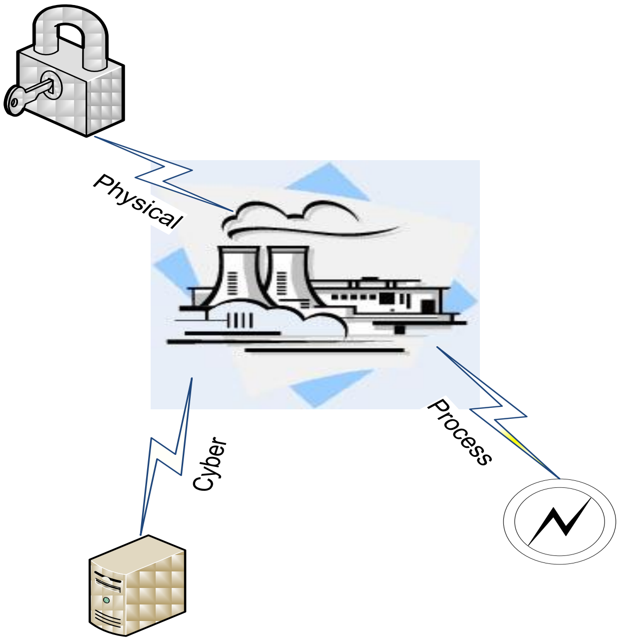


Fig. 1. Nuclear plant implementing methods of security

Aside from the large amounts of data generated that must be stored and analyzed, there are a number of other issues that prevent that data from providing useful information:

1. Different sensors, e.g. a volume sensor and card swipe, often do not share a common means of transmission; so any robust data collection system must be able to support a wide variety of input mechanisms.
2. Data formats are usually tailored to a specific device and generally incompatible with other subsystems.
3. Sensor data is susceptible to noise, incompleteness or nonstandard metrics.
4. Even in the event data can be gathered and correlated properly, there exists limited capability for relating events occurring between disparate systems such as an access control system and a proximate subsystem process. Take, for example, an intruder who forces open a door in order to gain access to a pump and tamper with a valve. Consider also, a cyber attack where regulator inputs are spoofed so that the unit wrongly calculates unit pressure. In either case, there significant damage may occur but if the break-in, cyber attacks and resulting damage are looked at in isolation there is limited capability for a plant manager to determine causality between them.

Relational Database Management Systems (RDBMS) provide useful tools for storing and retrieving datasets but require a "relationship" in order to merge disparate datasets. Multi-Dimensional Databases (MDD) suffer from the same issues, even when employing advanced schemas such as the constellation [7]. Binding different factual elements requires introducing informational ties, even if those ties are artificial.

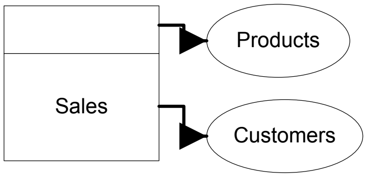


Fig. 2. Typical Entity Relationship linking sales to products and customers

Furthermore, in order for the data that is stored to be mined effectively, it must be formatted in such a way so as to lead to mining models that are understandable and lead to successful classifications and predictions. This often requires Extraction, Transform and Load (ETL) functions to clean, discretize and transform data from a source input to final repository.

Finally, once all that is accomplished, a human interface must be devised that presents the knowledge in a usable format. The information must be accurate, timely and actionable or suggestive enough that an operator can quickly isolate problems and take corrective measures and achieves the stated goals for a resilient design.

III. Data Fusion Architecture

Data Fusion is defined as integrating disparate data elements into a coherent and usable framework for storage, analysis and display. By "fusing" different data elements, relations are made possible and the advantages of RDBMS and MDD can be brought to bear. In the test nuclear facility that requires combining physical, cyber and process into a unified data model that can be analyzed and interrogated. This is accomplished via a series of steps:

1. The various data sources (physical, cyber and process sensors) export data to a central location.
2. The data is time-stamped and cross-referenced to obtain sufficient temporal-spatial characteristics for data mining.
3. Data is then cleaned and transformed and stored in a temporary data repository.
4. Processes then migrate and transform the data from the repository to a permanent data warehouse.
5. Analysis, data mining is performed on the new data.
6. Existing rules are applied and new information obtained from the analysis is applied to the rules engine.
7. Any appropriate alerts are generated.
8. Display data is created and made available for retrieval.

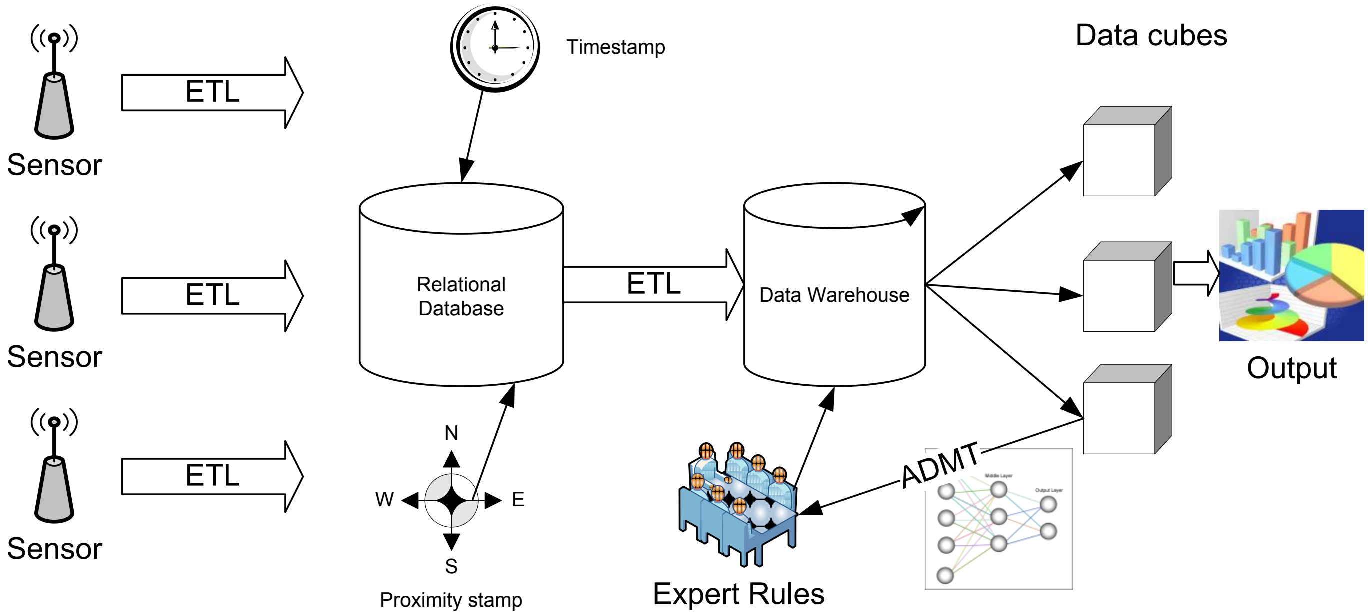


Fig. 3. Data Fusion Architecture

Fig. 3 shows what an implementation of such an architecture might look like. In addition, it is highly desirable that all data transactions be as secure as possible to avoid corruption and tampering. For that various techniques are available such as SHA (Secure Hash Algorithm), DES (Data Encryption Standard) over TCP.

Achieving the above requires a number of software components (and applicable hardware).

1. A relational database.
2. A data warehouse
3. Data mining tools
4. Presentation tools
5. Software for data input and extraction

6. APIs between the sensors, databases and presentation software.

Advanced Data Mining Techniques (ADMTs), combined an expert system provide the necessary sophistication and analysis of the data. ADMTs are automated algorithms used for classifying data and making predictions based upon recognizable patterns and lie at the heart of this architecture. ADMTs have been used successfully in Intrusion Detection Systems (IDS), image and character recognition, adversarial game-play and a host of other applications.

At the heart of each ADMT is an algorithm that looks for patterns in the data that can be described or classified. This description or classification becomes knowledge which can then be turned into rules for analysis or prediction [7].

There are a number of different types of ADMT. One such ADMT is the Artificial Neural Network (ANN). Each neuron of an ANN separates data elements into one of two groups. This is done by running data inputs through a neuron's processing unit combined with a series of weights applied to each input. For a set of *n* inputs *x* and weights *w* the output signal *o* is given by the following:

$$o = f\left(\sum_{i=1}^{n} w_i x_i\right) \quad (1)$$

By combining the output of neurons into an ANN, it is able to perform more sophisticated classifications, such as the one demonstrated in Fig. 4.

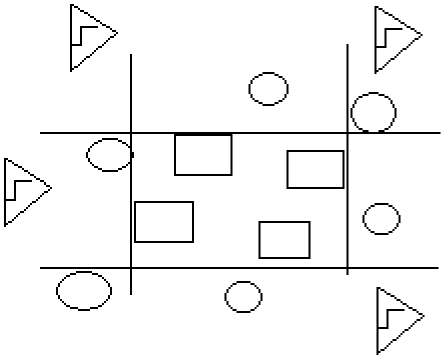


Fig. 4. Neurons in a Neural Network classify squares and circles

There are a number of algorithms in the ADMT family including decision trees, Support Vector Machines, k-means/medoids clustering, fuzzy c-means and others. Each brings certain advantages and disadvantages to data classification and rule generation and used together can provide an even more comprehensive overview of the data.

An example in the practical use of ADMTs is in credit card fraud detection. A person's card usage history forms a "pattern" that can be compared against recent usage. Significant deviation from the expected pattern is flagged as potential fraud and investigated. Because of the huge volume of transaction data, the ADMT is able to explore and analyze in a manner impractical and too time-consuming for a human.

ADMTs are automated and designed to run against large datasets that would typically overwhelm human interpretation, but are also capable of working with humans to develop extensive knowledge bases and rule sets. ADMTs search for information that is often hidden from traditional search/query techniques and attempt to "learn" from data behavior. This new knowledge can then be incorporated into additional rules that can govern an expert system [7].

ADMTs are particularly valuable because, unlike a traditional static expert system, they can adapt more readily as new knowledge becomes available. Threats today are very sophisticated and can evolve rapidly. Software that must identify these threats must be able to respond in kind. ADMTs ability to adapt quickly provides the final key component of the fusion architecture.

A research grant in partnership with the Idaho National Laboratory was initiated to implement a data fusion architecture with the express goal of improving human system interaction between plant operators and the systems they monitor. The intent is to improve plant up-time as well as threat detection and response.

Because the data fusion architecture involves a number of sophisticated components, work was divided into three phases. The first phase, described in this paper, involved developing input models, threat scenarios, test procedures and building a prototype architecture to demonstrate the overall concept.

IV. Implementation

The architecture implements and extends previous research and design work in Resilient Control Systems performed at the Idaho National Laboratory [6]. The prototype implementation consists of a combination of Commercial, Off-The-Shelf (COTS) software, along with customized third party libraries and number of internal modules. A Microsoft SQL Server 2008 RDBMS served as the platform for the relational database and data warehouse. SQL Server also provides a suite of fairly sophisticated data mining tools for training models and testing results. The presentation layer and dashboard was built using software components from Dundas Software and Infragistics Software along with custom software programmed in Microsoft Visual Studio.

In a large and diverse environment such as a nuclear plant, it is not reasonable to expect input devices to all share a common interface, data format or even protocol [8]. The architecture must accommodate this by being robust enough to support a wide variety of input types as shown in Fig. 5.

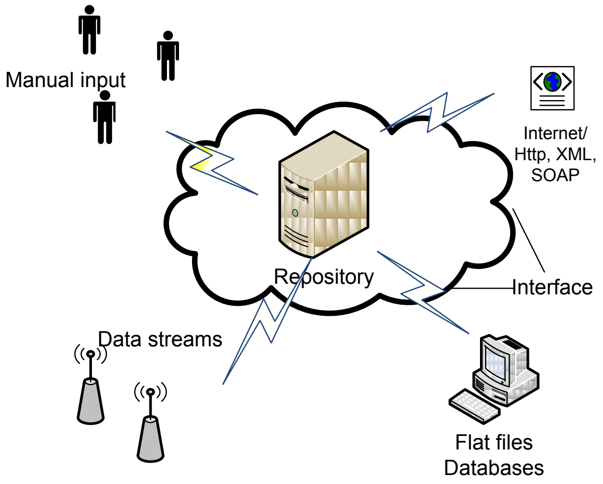


Fig. 5. Input from various sources to data repository

Sensor data is simulated using custom software created in Visual Studio, PHP and Matlab and consists of flat files, excel files and direct input via software simulator. A series of relational databases house the data and a callable API was built to interface the simulation software with the RDBMS. For data mining, a basic rules engine was constructed along with a knowledge base. Stored procedures handle the ETL process and provide timestamp information to "fuse" the various components.

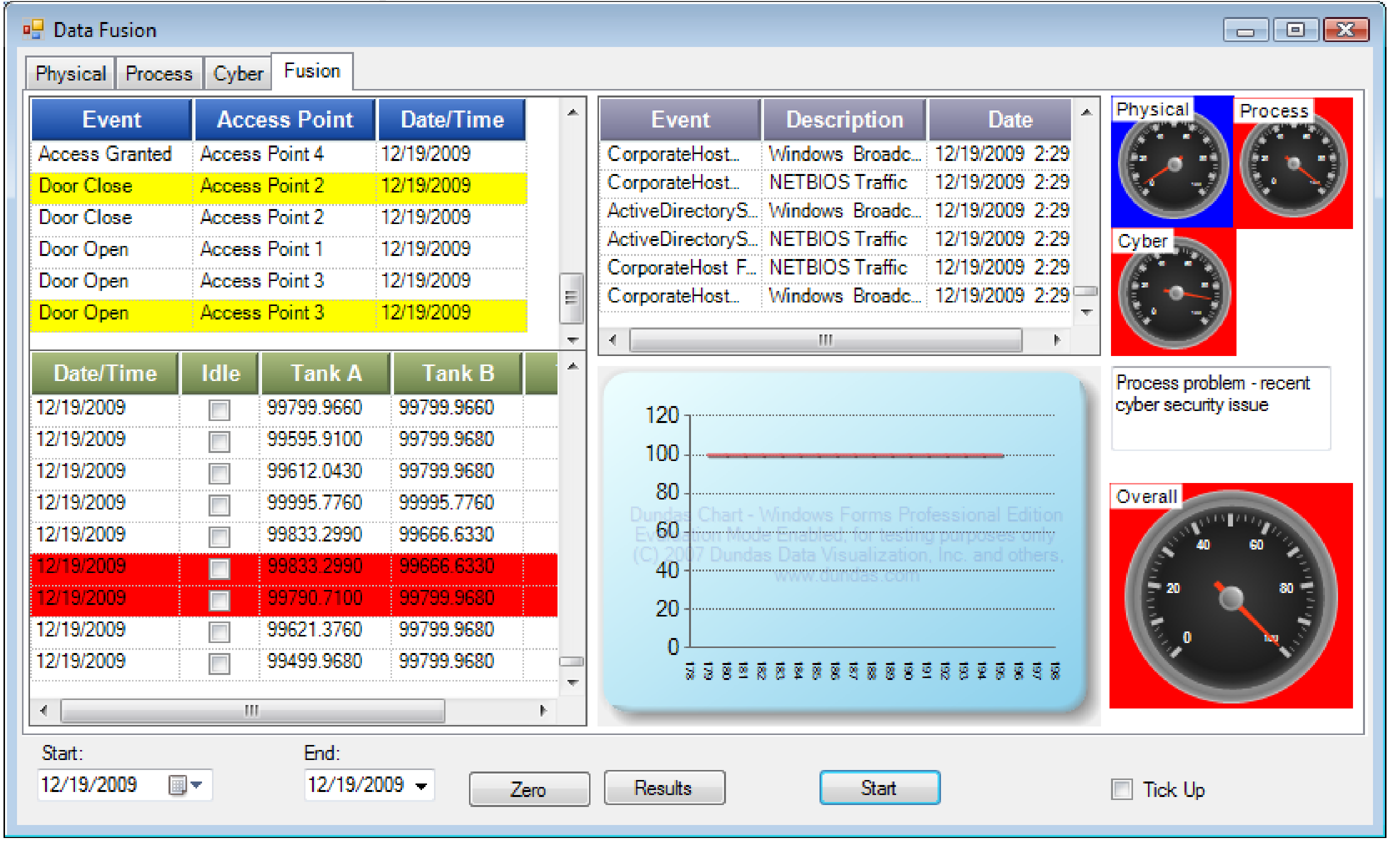
In order to determine overall system status as well as each subsystem, an algorithm and lookup table was used to determine the seriousness of each event, and assigning it a value. Event values are cumulative, but degrade with the passage of time. Alerts become a function of the combined event values over some time *t*. Hence for *n* subsystem events *e*, at any given point in time t_0 alerts are calculated as follows:

$$a_k = \sum_i^n \sum_j^m e_{ij} f(t_0 - t_j) \qquad (2)$$

"Relateability" *R* of a subsystem event s_a is also a function of the time proximity of an event in one subsystem to events occurring in other subsystems:

$$R_{s_a} = \sum_i^n \sum_j^m e_{ij} g(t_0 - t_j),\ i \neq a \qquad (3)$$

This relationship forms the basis of the data fusion and provides the rules engine and knowledge base with enough information to determine problem areas and possible actions to take.


Data Fusion
Physical
Process
Cyber
Fusion
Physical
Process
Cyber
Overall
120
100
80
60
40
20
0
Dundas Chart - Windows Forms Professional Edition
Evaluation Mode Enabled, for testing purposes only
(C) 2007 Dundas Data Visualization, Inc. and others.
www.dundas.com
Zero
Results
Start
Tick Up

Event	Access Point	Date/Time
Access Granted	Access Point 4	12/19/2009
Door Close	Access Point 2	12/19/2009
Door Close	Access Point 2	12/19/2009
Door Open	Access Point 1	12/19/2009
Door Open	Access Point 3	12/19/2009
Door Open	Access Point 3	12/19/2009

Event	Description	Date
CorporateHost...	Windows Broadc...	12/19/2009 2:29
CorporateHost...	NETBIOS Traffic	12/19/2009 2:29
ActiveDirectoryS...	Windows Broadc...	12/19/2009 2:29
ActiveDirectoryS...	NETBIOS Traffic	12/19/2009 2:29
CorporateHost F...	NETBIOS Traffic	12/19/2009 2:29
CorporateHost...	Windows Broadc...	12/19/2009 2:29

Date/Time	Idle	Tank A	Tank B
12/19/2009	□	99799.9660	99799.9660
12/19/2009	□	99595.9100	99799.9680
12/19/2009	□	99612.0430	99799.9680
12/19/2009	□	99995.7760	99995.7760
12/19/2009	□	99833.2990	99666.6330
12/19/2009	□	99833.2990	99666.6330
12/19/2009	□	99790.7100	99799.9680
12/19/2009	□	99621.3760	99799.9680
12/19/2009	□	99499.9680	99799.9680

Process problem - recent cyber security issue

Start: 12/19/2009

End: 12/19/2009

Fig. 6. Application Dashboard

V. Results

Data files and software simulation combined to simulate the output of a live system with input from physical, cyber and process system data. These were imported into the database and processed as described. Performance was acceptable. On a single-processor laptop, the system was able to provide real-time information, raise alerts and make actionable suggestions. Information was presented in a dashboard format, shown in Fig 6., and successfully related a pump malfunction with a recent simulated series of cyber attacks. In a more traditional setting, an operator might be completely unaware of the cyber tampering and could inadvertently damage the pump unit by adjusting it improperly. By combining the different data streams in a more coherent collection, an operator might now better respond to the real threat, a cyber attack and tampering, instead of simply responding to a pump malfunction. The Resilience Control System Design goal of combining disparate system data for improved state awareness is achieved.

VI. Conclusion and Future Work

The phase 1 prototype demonstrated basic data fusion capabilities combined with data analysis and display functionality, demonstrating a mock attack against a nuclear subsystem. Disparate data from physical sensors, cyber sensors and process data were joined along a timeframe and combined with a rules engine to produce basic threat analysis.

The prototype provides a generic interface in support of a diverse array of potential inputs, a relational staging area, ETL functions, a data warehouse and display software. The dashboard is considered acceptable to users and similar to those an operator might see in an actual facility.

Phase 2 and phase 3 work needs to be done to upgrade the prototype. These upgrades will include more advanced ADMTs, a more robust data warehouse, sophisticated rules engine and knowledge base. Security components will also be required in order to bring this design to a production system that can be used to support a live nuclear facility. Along with overall improvements in functionality and display, work must be done to train and evolve ADMTs, requiring and better models with more detailed and real-world data.

Future work also will involve implementing the data fusion architecture with additional ADMT techniques using Fuzzy type-1 and type-2 contextual rules along with improved and more flexible neural network algorithms than is provide with the software. These techniques will allow human-based reasoning and interpretation to be more closely aligned with data analysis and presentation. Being able to use human-relatable terms such as THREAT LEVEL = VERY HIGH or CAUSE = SOMEWHAT LIKELY or SOLUTION = MODERATELY APPLICABLE allows human-system interaction to occur much more understandably. Finally, fusion using spatial information needs to implemented and tested.

References

[1] W. Yu, E. Popova, E. Kee, A. Sun, R. Grantom (2005) Basic factors to forecast maintenance cost for nuclear power plants, 13th International Conference on Nuclear Engineering

[2] T. Anderson (2005) Diablo canyon power plant: steam generator repair/replacement cost/benefit analysis; Proceedings of the 2005 Crystal Ball User Conference

[3] The High Cost of Nuclear Power, Executive Summary (March 2009), http://www.uspirg.org/home/reports/report-archives/more-reports/more-reports2/the-high-cost-of-nuclear-power-why-america-should-choose-a-clean-energy-future-over-new-nuclear-reactors-new-jersey. Accessed 15 January 2010

[4] Major Nuclear Power Plant Accidents, http://www.atomicarchive.com/Reports/Japan/Accidents.shtml. Accessed December 15, 2009

[5] Backgrounder on Emergency Preparedness at Nuclear Power Plants http://www.nrc.gov/reading-rm/doc-collections/fact-sheets/emerg-plan-prep-nuc-power-bg.html. Accessed December 15, 2009

[6] C.G. Rieger, D.I. Gertman, M.A. McQueen (2009) Resilient control systems: next generation design research, 2nd Conference on Human System Interactions, pp. 632-636.

[7] J. Han, M. Kamber (2006) Data Mining Concepts and Techniques, 2nd Ed, Morgan Kaufmann Publishers.

[8] S. Cheng, B. Shen,.K. M Peng, Q Zhao, R. Xue, C. Gong (2009) Research on coordinated control in nuclear power plant; IEEE Conference on Machine Learning and Cybernetics

[9] M. Kitayama, R. Matsubara, Y. Izui (2002) Application of data mining to customer profile analysis in the power electric industry, IEEE Power Engineering Society Winter Meeting.

[10] S. Harinath, S.R. Quinn (2006) Professional SQL Server Analysis Services 2005 with MDX, Wiley Publishing