10TH INTERNATIONAL COMMAND AND CONTROL RESEARCH AND TECHNOLOGY SYMPOSIUM

The Future of C2

Title of Paper:
The Data Warehouse in Service Oriented Architectures and Network Centric Warfare

Topic: C4ISR Architecture

Author: Jack Lenahan
POC: Jack Lenahan
Organization: Office of the Chief Engineer
Space and NAVAL Warfare Systems Command
Charleston, S.C.
Address: P.O. Box 190022
N. Charleston, South Carolina: 29419
Phone: 843-218-6060
Email: John.Lenahan@Navy.mil

Abstract
Since Network Centric Warfare (NCW) theory stresses shared understanding, command dispersal, and improved situational awareness does it not follow then, that data availability, mining, and superior analytics must be available at all policy and command levels to support superior decision making? Analyzing the anticipated massive amount of GIG data will almost certainly require data warehouses and federated data warehouses. The central question being addressed here is: Will a new Data Warehouse Paradigm be required for Network Centric Warfare Service Oriented Architectures (SOA)? This research attempts to answer this question by analyzing Service Oriented Architecture (SOA) based "Virtual Data Warehouses", Corporate Information Factories, and SOA based federated data warehouses.
The research concludes that "Composeable Data Warehouse Services" offer the best methodology for supporting decision making at all levels of dispersed command. "On Demand - Composeable Data Warehouse Capabilities", based upon web services, should be implemented and registered on the GIG for testing and deployment if successful. These new paradigms will require that adaptive and agile Extract, Transform, and Load (ETL) services, dynamic report creation services, composeable mining engines, robust Meta data tagging for discovery and analysis, and more sophisticated analytics services be developed to fully exploit the vast amounts of Global Information GRID data which is expected to accumulate.

Introduction

Since NCW theory stresses command dispersal, or moving decisions to more levels of the command chain, then does it not follow that data availability and superior analytics must be available at all levels to support superior decision making? Data architects must address how to adapt the traditional data warehouse to the needs of the diverse NCW user communities. NCW will provide an enormous amount of data to the GIG. Analyzing this massive amount of data will almost certainly require data warehouses and federated warehouses. A view of aggregating local and global data into a data warehouse for mining and analytics by means of a federated warehouse is shown by the following graphic.

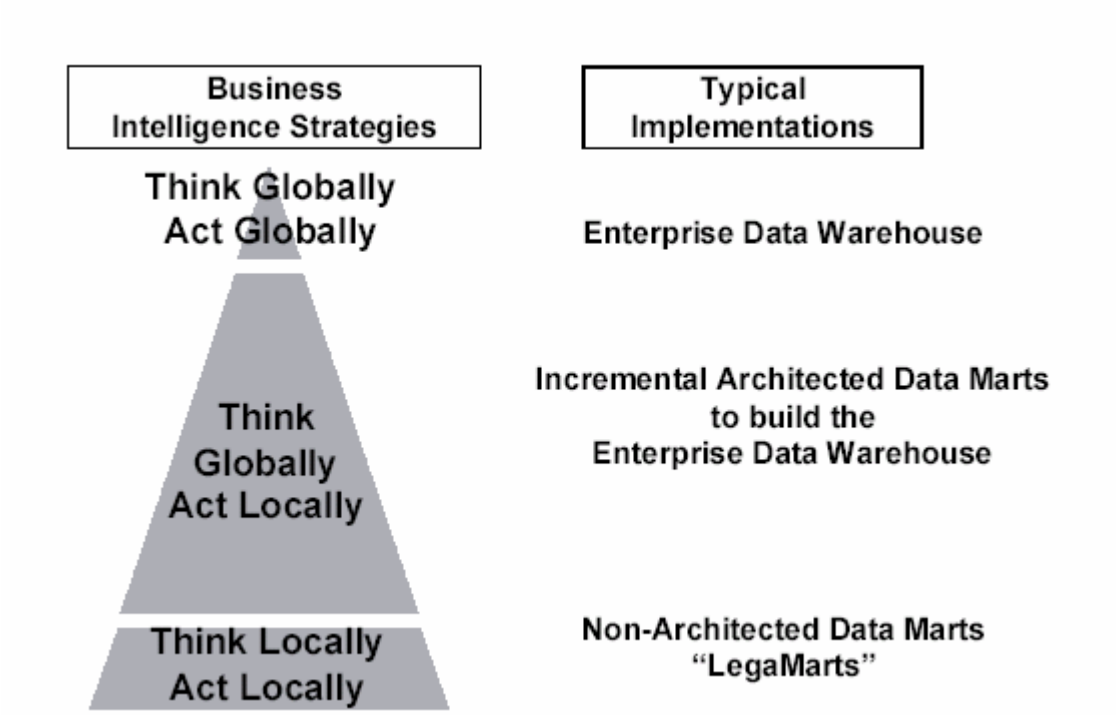

Figure 1 – Depiction of Relationship to Global Activity Planning[1] and Analysis versus Local Activity Planning and Analysis with Legacy Data Stores (LegaMarts)

There must be a methodology which will enable the analysis of the vast amounts of C4ISR data at a global level in order to support many of the global effects based operations strategies required for policy and NCW implementation. There must also be a methodology available at the mid levels and lower levels of command in order to support the NCW goals of dispersed command and superior decision making. Novel data discovery and correlation can be achieved by analyzing data at both the local and global levels. This is one of the great promises of NCW theory. In order to support that promise in a timely fashion, we must have new mechanisms to exploit the data. The traditional warehouse implementation paradigm takes many months or years to bring to fruition. NCW needs to dramatically reduce this warehouse implementation and analysis time to days.

Proposal

An SOA orchestrated web services approach can be used to implement both composeable standalone NCW warehouses and composeable federated global data warehouses. These types of data warehouses amount to "virtual data warehouses". They would exist only on the GIG as composed orchestration sets only for as long as required. If my recommendation of a GRID is accepted for NCW, then the computational complexity of a federated warehouse will be manageable. Web services dedicated to each layer of warehouse management may be a time saving device for dynamic warehouse construction, mining and performing analytics on the expected massive amounts of GIG data. A quote from related research offered by John Medicke of IBM[2] follows. **"**Reacting with speed to changing business conditions, occurring either internally or in the marketplace, requires insight and agility. More than ever, businesses need to leverage their business intelligence systems to fuel this responsiveness. However, many legacy data warehouse environments are holding organizations back with their crusty structures and stale data. There is good news. Business Performance Management solutions offer a new approach to corporate informational alignment. Business performance management derives performance from the business process, coalescing the operational and the analytical environments. The business process management environment, as the superstructure of operational activity, is the perfect channel for the fluid execution of performance management. Business measures are streamlined into the analytical environment and actionable knowledge is turned into action by driving business process operations."

To repeat from the quote above: "crusty structures and stale data". Well stated and accurate. Warehouses are usually designed with a particular financial return on investment goal in mind. This drives the data structures to represent corporate assets in such as manner as to facilitate mining of "anticipated data or qualitative results" already suspected of being of value. Thus the stale data comment. But in the case of NCW, we wish to discover new and unexpected things from all the posted data. It seems contrary to NCW theory to wait a year until the warehouse is built in order to exploit the potentially rich SOA NCW data population sets. I would like to transition now to a few basic definitions. This is unfortunately needed due to the lack of understanding of the differences in data storage products.

Data warehouse terminology

A data warehouse has several attributes that distinguish it from a relational database.

1. A data warehouse is by definition **"READ ONLY".** Data mining will yield meaningless results if the underlying data content is permitted to change unexpectedly. Citing Inmon's paper ("Inmon is the Father of the Data Warehouse"), "a (data) warehouse is a subject-oriented, integrated, time-variant, and non-volatile collection of data in support of management's decision-making process."
    - Subject-oriented: Data that gives information about a particular subject instead of about a company's on-going operations.
    - Integrated: Data that is gathered into the data warehouse from a variety of sources and merged into a coherent whole.

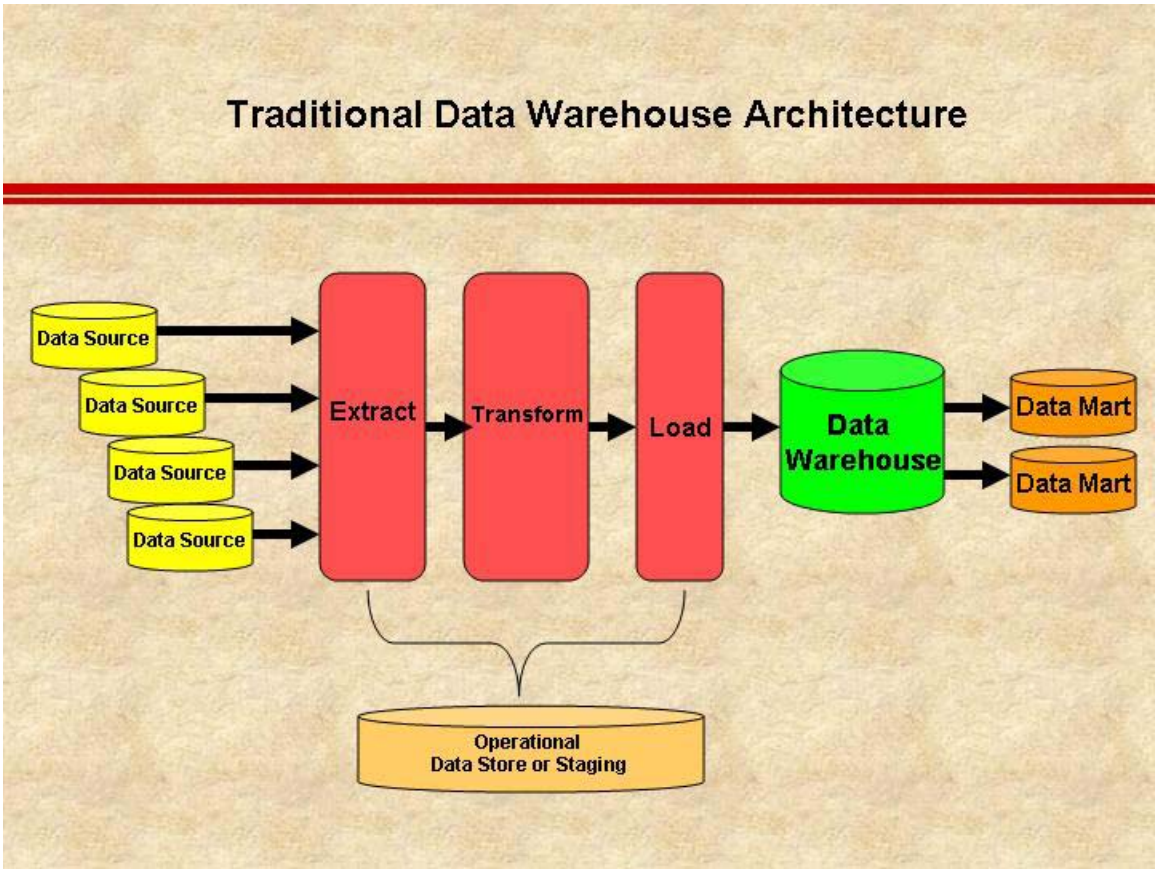

Figure 2 – Traditional Data Warehouse Architecture depicting the ETL Sequence

- Time-variant: All data in the data warehouse is identified with a particular time period. (Thus it cannot change after initial load and is therefore read only)

2. Data warehouses are deliberately de-normalized for improving the reporting and general access speeds. Thus, they do not follow E.F. Codd's model in terms of normalization.
3. Warehouses get their data from multiple sources and usually have a serious data format reconciliation problem. Thus, one of the first data warehouse design tasks is data standardization of the same data but in different formats from multiple sources. For example, a ball bearing part number may exist as follows: 123-456-789 in source A. But source B represents the same part number as 987-654-321, while source C represents it as a "smart" number in the format xxzz-123-456-789-yynnbc-aabb. In the previous example the smart number contains information relevant to the storage of ball bearings by division (xxzz) and a supplier id (yynnbc-aabb). In order to load all the data concerning this ball bearing from the 3 sources, a formal warehouse process called Extract Transform & Load must be designed and executed to consolidate the part number formats

and prepare for the storage of only one reportable format. ETL phases are typically long in execution.

4. Extraction is defined as the process of retrieving the data to be loaded from the identified data sources in native legacy formats.
5. Transformation is defined as the process of converting data from multiple sources in different formats into a single format for each field in the warehouse, prior to its being loaded.
6. The operational data store is the place that extracted data is sent to and the place where the conversions are performed.
7. The data warehouse is the read only relational engine where the transformed data is loaded into or stored according to the star or snowflake schemas designed to support reporting dimensions.
8. Data marts are read only department or organizational level "mini-warehouses". They are fed by the main warehouse itself. They exist to reduce the load on the central warehouse and to target data mining performance improvements due too the reduced number of rows required for departmental (not corporate level) analysis.
9. A Federated[3] Data Warehouse Architecture is an overall system architecture that accommodates multiple DW/data mart (DM) systems, operational data stores (ODS), amorphous reporting systems, analytical applications (AAs), etc. As the Internet is a network of networks, a federated DW architecture is an architecture of architectures. It provides a framework for the integration, to the greatest extent possible, of disparate DW, DM and analytical application systems.

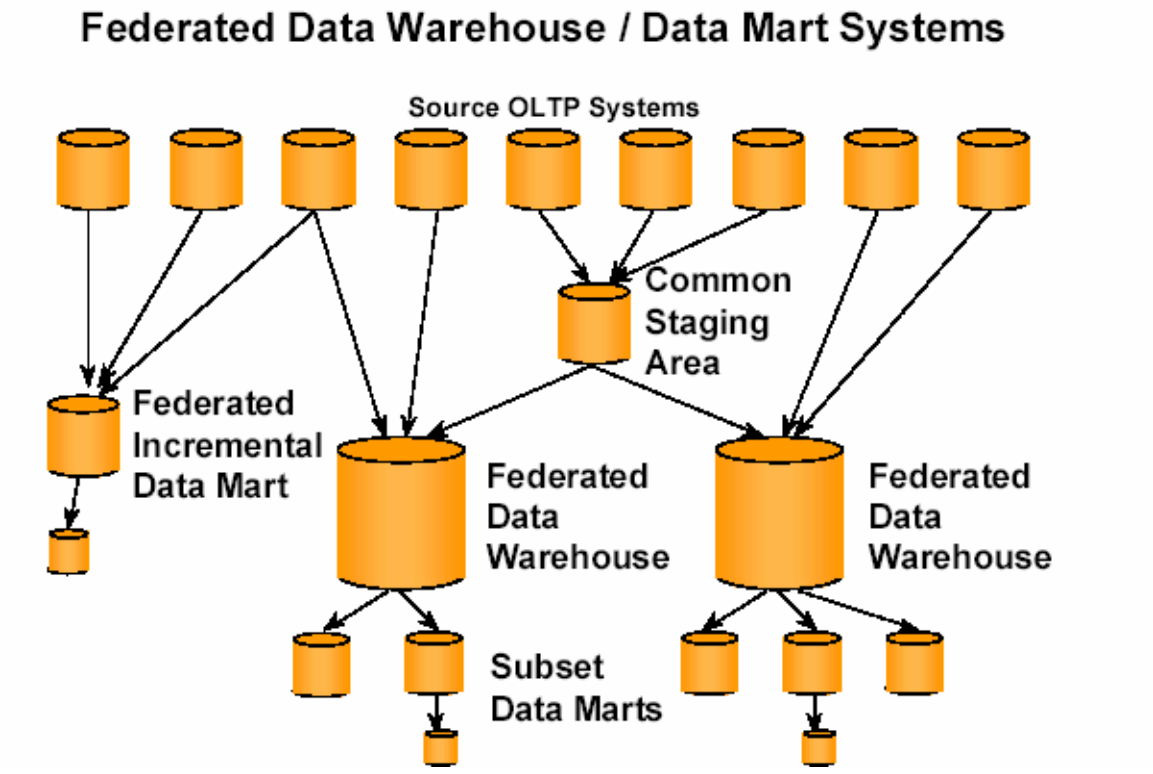

Figure 3 – Federated Data Warehouse Model

10. The Corporate Information Factory (CIF)[4] – (GIF = Government Information Factory) - The CIF is a blueprint of a data warehouse architecture - The CIF was developed by Bill Inmon and Claudia Imhoff as a framework for the information processing of an organization at a high architectural level. It specifies which components of information processing there will be and where they will be placed. As information systems develop over time, the CIF-framework acts as a guideline telling what to add to the information systems architecture and where this addition should be made. The CIF has transformed over the last decade from a DW blueprint to a framework describing a complete Business Intelligence suite, including Web Environment, Portals, Analytical Applications, etc.
11. From the same source (Jens Körner) "The GIF is the counterpart of the CIF in the public sector. Inmon's GIF is an information systems blueprint for government agencies; namely for federal, state, and local agencies, and takes into consideration the known ERP and DW needs for:
    - Operational and informational processing
    - Multidimensional processing and reporting
    - Managing very large amounts of data
    - High availability
    - Data mining and exploration, and so forth…
    - In addition, the GIF takes into account the need for:
        i. Interagency passage of data
        ii. Integrated electronic security
        iii. Predictive security (the ability to use data to anticipate threats before they occur)
        iv. Reconciliation of data
        v. Addressing the challenges of stovepipe systems
    - GIF vs. CIF distinctions
        i. (i) No concept of profits in the Public Sector
        ii. (ii) Multiple objectives in the public sector vs. single objective in the private sector

To further describe the CIF concept, I have included the following quotation from Dr. Claudia Imhoff:

"A Corporate Information Factory is a content delivery mechanism. One half deals with "getting data in" and consists of the operational systems, the data warehouse and/or operational data store, and the complex process of data acquisition. Much has been written about these components, especially the extract, transform and load (ETL) part of data acquisition. The ultimate deliverable for this part of the CIF is a repository of integrated, enterprise-wide data for either strategic (data warehouse) or tactical (operational data store) decision making.

The other half of the CIF deserves some more attention. It is summarized as "getting information out". The ultimate deliverable for this half of the CIF is an easily used and understood environment in which to perform analyses and make decisions. Most of the

highly touted business intelligence (BI) benefits are derived from getting information out - data consistency, accessibility to critical data, improved decision making, etc."

CIF as "Information ECOSYSTEM" Discussion

The following "information ecosystem"[5] discussion demonstrates that other pertinent research has occurred into the area of agile or adaptive data warehouse systems. This very relevant research complements the points that I am attempting to make. According to the author (J.M. Firestone) the CIF and GIF can be compared to nature's ecosystem as follows:

"W. H. Inmon's vision of the IT future is an information ecosystem…"

"With different components, each serving a community directly while working in concert with other components to produce a cohesive, balanced information environment. Like nature's ecosystem, an information ecosystem must be adaptable, changing as the inhabitants and participants within its aegis change. Over time, the balance between different components and their relationship to each other changes as well, as the environment changes. Sometimes the effect will appear on seemingly unrelated parts (sometimes disastrously!). Adaptability, change, and balance, are the hallmarks of the components of a healthy information ecosystem."

Further, "the corporate information factory (CIF) is the physical embodiment of the notion of an information ecosystem."

In other words, the CIF "is an architecture for the information ecosystem, consisting of the following architectural components:

- An applications environment
- An integration and transformation layer (I & T layer)
- A data warehouse with current and historical detailed data
- A data mart(s)
- An operational data store (ODS)
- An Internet and Intranet
- A metadata repository"

Composing an SOA XML Model Discussion

If we start with NCW theory, then all services in the SOA should be composeable. Can we compose a data warehouse? The answer appears to be a resounding "yes" given the proper set of web services registered in UDDIs, detailed meta tagging, a set of data sources, and a set of orchestration tools and a GRID for distribution of compute intensive transformations. The XML & Web Services Based Model shown in the below graphic offers an approach to composeable data warehouses that deserves some attention.

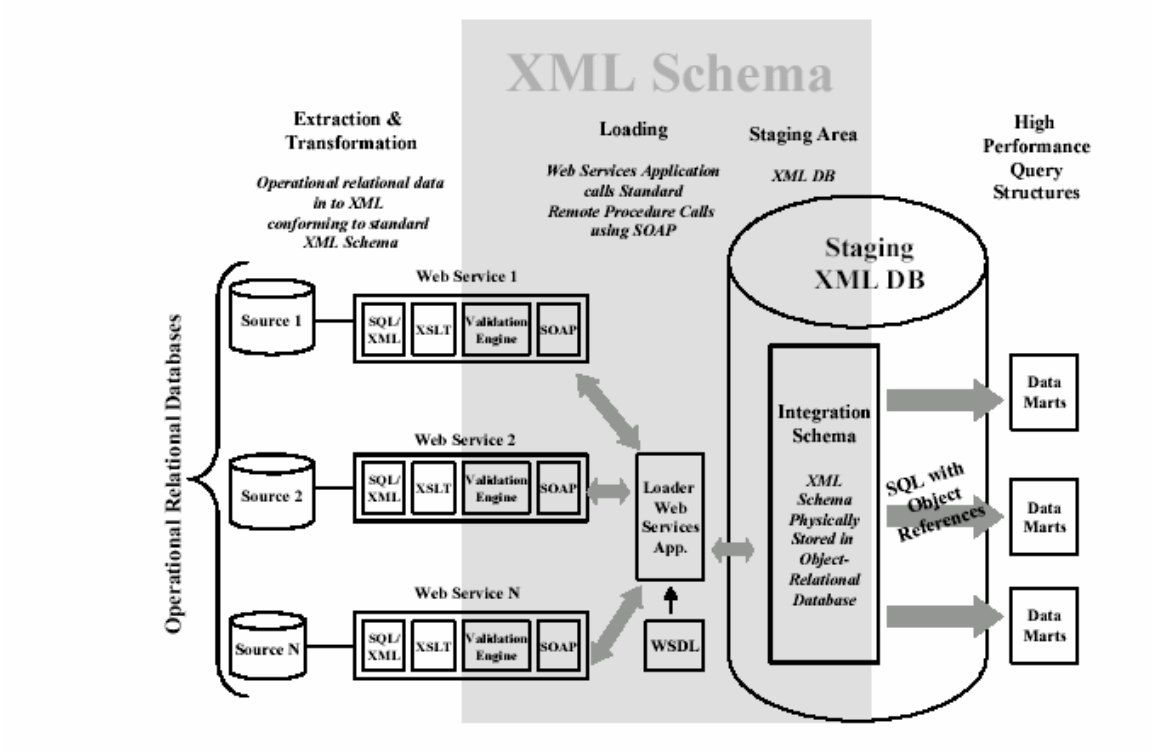

Figure 3 – Web Services[6] and XML Schema Based SOA Model for Agile Warehouse Creation

This model offers quite a bit of architectural diversity to the data architects. First, the ETL layers are componentized into web service sets. This means that the data extraction, transformation, and loading can be distributed over computers on a GRID thus reducing the time it takes to construct the data profiles and execute analytics. By defining the extract and transformation functions as services, the schema can be composed through the use of other tools and services and made available quickly as a "staged XML DB" to facilitate the loading and orchestration of analytical capabilities at any level of the command chain or GIG user communities. Thus, this model seems to support the dynamic creation of data warehouses at both the global levels and at "N local" levels. Second, the reporting from the warehouse or the marts can be supported by the use of web services also. This offers a tremendous flexibility.

Composeability discussion

Architects, who assume a priori, that they know what any given data user in any community of interest is going to require for day to day operations and decisions support, are arrogant and suffer from an "omniscience complex". The point of "publish all data" is to make data available to previously disenfranchised users. Composeability or the ability for an average user to define and construct (or have an intelligent assistant agent define and construct on his behalf) data warehouses, new analytics or data mining capabilities, and report types, supports the requirements of dispersed command and superior decision making at a greater level of granularity than has been available prior to this time. The capability to compose data warehouses or request that user specific analytic reports be

generated offers a potentially tremendous asset to the lowest levels of command. It also permits discovery of data relationships and the uncovering of novel facts pertinent to a particular user community which would not be possible if the traditional warehouse paradigm is followed. This is true since the user community would not be at the mercy of the "omniscient architects" and thus would be free to try their own compositions. By combining GRID technology with composeable data warehouses, speed and customer interest are more likely to be satisfied.

Relevant Research

An interesting related research paper[7] has been written by Dr. Claudia Imhoff concerning the use of agents in assisting the warehouse user. This approach seems very close to the idea that I am promoting, mainly that of **"on demand warehouse composition services".** In her paper, "Intelligent Solutions: Lessons from the Farm - Managing the Data Delivery Process", she discusses the use of a "request coordinator" agent in a CIF.

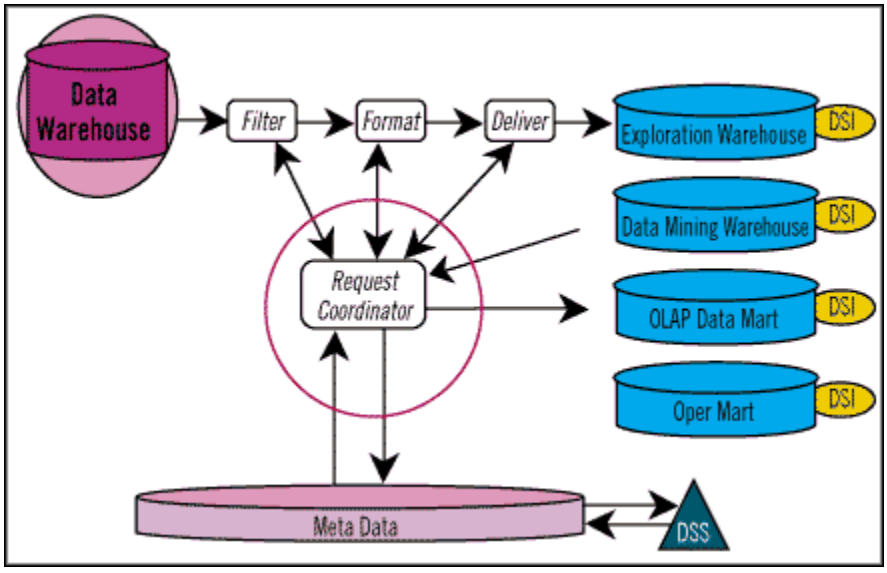

Figure 4 – CIF Model with Request manager – by Claudia Imhoff

To quote extensively from the author **(**Claudia Imhoff) "The request coordinator is like the farmer who plans his next season carefully, determining what seeds will be planted, which fields will have what crops, where efficiency of scale, market value and time to market (harvest schedule) play a role, etc. In the CIF, the request coordinator first captures the business user requests, prioritizes them and then profiles them to fully understand the request. Meta data plays an important part in this step - it is used to determine whether a new mart is warranted or an existing one can be enhanced to accommodate the request. If a data mart that can satisfy the request already exists, then the function simply gives the users access to the mart, perhaps adding a bit of new data, a new report or creating a view specifically for that set of users. If a mart does not exist, then the coordinator must begin the process of filtering the right data from the warehouse,

formatting it to the correct technological format, and delivering that data to the new mart per the requested schedule."

The quote above is provided to demonstrate that serious attention has been given to dynamic construction of warehouse components. I believe that the extension of this model to the actual dynamic creation of at least data marts is relevant and necessary to making data warehousing meaningful on the GIG.

Results

The Composeable Data Warehouse offers the best methodology for achieving superior decisions at all levels of dispersed command. Composeable Data Warehouse capabilities, based upon web services, should be implemented and registered on the GIG for testing and deployment if successful. Whether the base model is an SOA XML model, a CIF, or GIF model or a combination of all these models, should be decided upon by the services or DoD as a representative of the JOINT community.

References and Relevant Research

1. "Architectures and Approaches for Successful Data Warehouses", By Douglas Hackney, April 2002, Source: http://www.egltd.com/presents/ArchitecturesApproaches.pdf
2. "Realizing A Real-Time Enterprise With Business Performance Management", 08/02/2004 By John Medicke, Executive IT Architect; Architect, SMB and Mid-Market Architecture, IBM, Source: http://www.ebizq.net/
3. "Data Warehouse Delivery: Federated FAQs" - by Douglas Hackney Published August 15, 2000, Source: http://www.datawarehouse.com/article/?articleid=2887
4. "Bill Inmon's Corporate Information Factory", 01/23/2004, by Jens Körner - Source: http://www.wiwi.hu-berlin.de/~guenther/DW/jenskoerner.ppt.
5. "The Corporate Information Factory or the Corporate Knowledge Factory?", Joseph M. Firestone – Source: http://www.dkms.com/papers/cifckf.pdf - Also, see W. H. Inmon, Claudia Imhoff, and Ryan Sousa, "Corporate Information Factory", (New York, NY: John Wiley & Sons, 1998), Pp. 2-3
6. "XML Schema and Web Services for ETL in the Staging Area of a Scientific Data Warehouse" - Mykola Dudar1,2, Owen Eddins2, Susan M. Wilson2, Susan R. Atlas2,3 and Robert Veroff1,* - 1 Department of Computer Science, 2 Center for High Performance Computing, 3 Department of Physics and Astronomy and Center for Advanced Studies, University of New Mexico, MS C01 1190, 1 University of New Mexico, Albuquerque, NM 87131, USA
7. "Intelligent Solutions: Lessons from the Farm - Managing the Data Delivery" Process by Claudia Imhoff - *Published in DM Review in November 2004*. Printed from DMReview.com