
*White Paper presented by*

# Analytic Data Solutions, LLC

**Title:** Information Based Design - Next Generation Data Warehouses for Healthcare Providers

**Author:** By John Majarwitz and Kishore Nair

**Date Written:** 9/29/2010

**Date Posted (to website):** 12/19/2011

**Brief Description:**

This paper outlines an incremental strategy for developing a data warehouse solution that is based upon getting early benefit to end users while being able to build upon the work that has been done. The key principles to delivering incremental value while avoiding throw away work over time are based on planning the work broadly while executing the work modularly.

4800 Spring Park Road
Box #18
Jacksonville, FL 32207
Office: 866.611.8595
Fax: 877.893.1410
www.metaanalytix.com

# Information Based Design - Next Generation Data Warehouses for Healthcare Providers

**By John Majarwitz and Kishore Nair**

## Overview

Very large companies can afford big investments and long lead times to build out new enterprise data warehouse environments. These can be several year projects that cost tens of millions of dollars, or more. Success (based on full achievement of the initial vision) is far from guaranteed. Not uncommonly, after several years, planning for the "next generation" data warehouse begins to take root and the process begins anew …

Operating in this manner is not an option for smaller companies and many larger companies are also taking a more modular approach to next generation warehousing in lieu of the "big bang" approach. Companies that provide health care services fall in this category as they are in an increasingly competitive market with additional regulatory and industry pressures that make it imperative that investments in decision support and analytics lead to successful solutions that produce business value, both short term and long term. This paper outlines an incremental strategy for developing a data warehouse solution that is based upon getting early benefit to end users while being able to build upon the work that has been done. The key principles to delivering incremental value while avoiding throw away work over time are based on planning the work broadly while executing the work modularly.

For many health care providers, taking this approach will produce decision support and analytics solutions that are created and supported at a low initial cost while also delivering increasing value over time. While the approach described in this paper is not unique to health care providers, the examples and patterns used will be from this industry sector. The goal of this paper is to help motivate the health care provider sector to take advantage of best practices to create greater decision support and analytical capabilities over time.

## What is a Data Warehouse Environment?

Before getting into more detail about planning broadly and executing modularly, we will first establish a definition of the data warehouse environment to serve as a reference architecture. *Figure 1* below depicts the major functions and data flow through a data warehouse environment.

4800 Spring Park Road
Box #18
Jacksonville, FL 32207
Office: 866.611.8595
Fax: 877.893.1410
www.metaanalytix.com

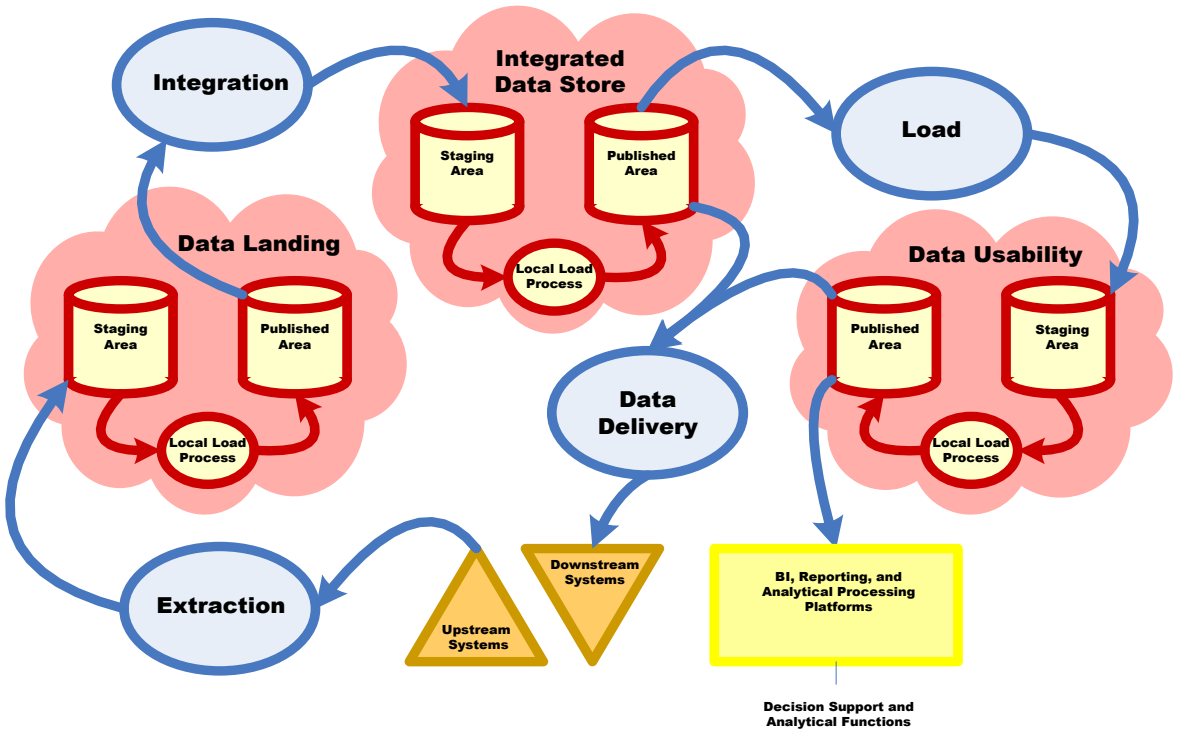


*Figure 1: Logical Reference Architecture for a Data Warehouse Environment*

In Figure 1, the **red zones** indicate the major stores of data within the data warehouse environment and the **blue zones** indicate the flow of data into and out of the data stores. It is important to view this as a logical architecture and not make assumptions about the physical implementation of servers and databases to fulfill the architecture at this point. This is because the way that the logical architecture is mapped to a physical systems architecture will depend upon a variety of technical factors including the size, scale, desired availability and current technical platforms of the organization. This can result in a systems architecture that places all of the data stores and processing flows within a single server; or it can place each data store on a separate server along with a separate server for processing the flow of data; or it can be any one of several other system topologies. The mapping of the logical architecture to a systems architecture is a topic outside the scope of this paper and will be covered by another white paper. The important point to recognize for purposes of planning broadly and executing modularly is that the logical reference architecture should be observed in order to provide the basic properties of separation of concerns and tight coupling that will provide sustainability over time. These properties are discussed in more detail below.

## The Red Zones

Each of the red zones fulfills a major function in the environment. Common to each zone is the concept of a staging area and a publishing area. The staging area is used to collect input and perform work in

4800 Spring Park Road
Box #18
Jacksonville, FL 32207
Office: 866.611.8595
Fax: 877.893.1410
www.metaanalytix.com

progress. The publishing area is used to make data available for downstream consumption. In order to reliably connect the red zones together it is important that there is a well established protocol for what data can be consumed downstream and when this can occur. Separation of staging and publishing establishes this definition. Once again, it is important not to interpret staging and publishing as being separate databases, or even necessarily implemented as databases in all of the zones as this will be determined by the mapping to the physical systems architecture. Each of the three red zones is defined below:

**Data Landing Zone** – This is the point of interface from external sources of data into the data warehouse environment. Source data needs to be collected and organized in terms of dependencies before it can be processed into a more integrated model. This is also the place where data cleansing and standardization can be performed as well as validation of data elements for quality assurance. The data is usually structured according to the source system structure at this point. So, it can be thought of as a collection of source models or "islands" of data that are not designed as a whole. In this zone, the staging area does not necessarily mean that data is physically transported from the source system and stored into the data warehouse environment (although this is often times the case). In some cases, it is possible to source and process the data as a function that connects to the source system data store and extracts and processes the data into the data landing zone's publishing area.

**Integrated Data Store** – This is the point where the various sources are integrated into a common model that is organized around the enterprise business data model instead of the data model of each source system. For example, a common data structure to maintain patient contact information would exist in this zone whereas there may be several sources of patient contact information, each with a slightly different structure. Creating a well formed integrated data model is the single most important piece of the data warehouse environment because this will serve as the foundation for how easy or hard it will be for the downstream decision support and analytical functions to be performed. Fortunately, there is a strong existing body of business models and data models within the health care provider sector that can be leveraged in creating this integrated data model.

4800 Spring Park Road
Box #18
Jacksonville, FL 32207
Office: 866.611.8595
Fax: 877.893.1410
www.metaanalytix.com

**"There can be multiple data stores within the Data Usability Zone each specializing in a subject area of data needed to fulfill certain business functions such as finance, clinical care, or research. These are often referred to as data marts."**

**Data Usability Zone** – This is the point of interface to end users. The data is available for access by business intelligence and reporting platforms in order to fulfill decision support and analytical functions. Data can also be extracted for delivery to external targets such as regulatory bodies, business partners, vendors, suppliers, etc. In some system architectures, the Data Usability Zone is combined with the Integrated Data Layer meaning that users directly access the published version of the integrated data model. This is an acceptable approach for smaller companies to start with but the distinction between the data integration processing being completed before data is published to end users needs to be rigorously adhered to in order to maintain integrity. As companies become larger, this distinction becomes even more critical, and, eventually, the two zones are usually separated within the systems architecture. There can be multiple data stores within the Data Usability Zone each specializing in a subject area of data needed to fulfill certain business functions such as finance, clinical care, or research. These are often referred to as data marts. Later in the paper we will cover the options for publishing data into data marts.

## The Blue Zones

The blue zones represent the extract, transform, and load (ETL) processes that are responsible for moving data from one point to another and converting the data into the desired structure. It is important to distinguish between duplication of data and replication of data in the architecture. A replica of data is a controlled copy usually created for ease of access, scalability, or performance reasons. When replicating, the ETL processes are carefully designed to always keep the controlled copy in synch with its defined source, which is the system of record. Replication of data is an appropriate tool when applied for the right reasons. Duplication of data is an independently managed copy, which, from that point forward is not supported through a controlled ETL process. Within a data warehouse environment, duplication is never an appropriate tool. Preventing duplicate data can be a tricky issue, however. For example two distinct source systems may provide overlapping data that, when assembled in the data warehouse environment, is duplicated, but not easily detected and managed. Each of the blue zones is described below:

**Extraction** – This addresses sourcing data from upstream systems which are typically the transactional systems that support the operations of the business. This data is incorporated into the Data Landing Zone. The major emphasis is on acquiring the necessary source system data. There can be processing to clean up, validate, and standardize data elements.

4800 Spring Park Road
Box #18
Jacksonville, FL 32207
Office: 866.611.8595
Fax: 877.893.1410
www.metaanalytix.com

**Integration** – This addresses significant transformation to combine the data from separate source systems into a common, integrated data model. This is typically the most complex and performance sensitive processing within the data warehouse environment.

**Loading** – This addresses the movement of the data from the Integrated Data Source to the Data Usability Zone. Since the data is already integrated into a common model the processing involved is fairly straightforward. However, there may be conversion of the model from a relatively normalized design (in the Integrated Data Source) to a dimensional design (in the Data Usability Zone) since dimensional models are often preferred for end users because they are a simpler form to query whereas normalized models are sometimes preferred for data integration because they are easier for maintaining data integrity.

**"You do not need to have all of the data in the data warehouse for some of the data to be of use for decision support and analytics. Taking advantage of this observation allows for a project to be executed modularly."**

**Delivery** – This addresses being able to provide data extracts to downstream targets that are external from the EDW. The source can be the Integrated Data Store or the Data Usability Zone.

## The Integrated Data Model

Now that we have base-lined to the logical reference architecture we have a better understanding of the processing and storage framework that comprises the data warehouse environment. But this is not sufficient to give us insight into how to approach creating this environment because we do not yet have a good enough understanding of some key principles that are important to the integrated data model. As stated above, creating a well formed integrated data model is the single most important piece of the data warehouse environment and any methodology for developing a data warehouse environment needs to take this into account. Being able to plan the work globally and execute modularly is tied to these basic two observations regarding the integrated data model:

1) You do not need to have all of the data in the data warehouse for some of the data to be of use for decision support and analytics. Taking advantage of this observation allows for a project to be executed modularly.
2) As additional data is sourced in the data warehouse environment, it will relate to existing data already in the environment. Taking advantage of this observation results in a better data model, providing consistent quality, high performance, and low cost. This requires one to plan globally.

*Figures 2, 3* and *4* below help to illustrate and expand on these observations. In Figure 2, the integrated data model is shown as having two categories of data: core data and extended data. Core data is the fundamental data that describes the business. This data is used prevalently for many decision support

4800 Spring Park Road
Box #18
Jacksonville, FL 32207
Office: 866.611.8595
Fax: 877.893.1410
www.metaanalytix.com

and analytic functions. For the health care provider sector, the basic business entities that describe the business are part of the core data. These include customer/patient, product/service, appointments, diagnosis, procedure, claims, invoice, etc. These entities are described by the attributes that are commonly needed. Extended data is more specialized data that deeply describes a specific area of the business. This can include additional attributes that are highly specialized within the core entities. Or it can include separate entities with their own specialized attributes. Examples within the health care provider sector can include: detailed clinical data for specific conditions or areas of analysis such as genomics research or cardiovascular study; detailed data for departmental operations such as ambulatory, emergency, radiology, etc; general ledger and other financial data for detailed financial analysis; and many other data sets.

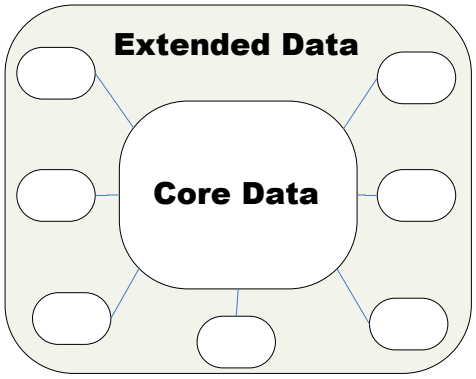


*Figure 2: Core Data and Extended Data Comprising an Integrated Data Model*

The core data and extended data do not need to be collected all at once. Instead, there are usually some convenient groupings that can drive the work program in a more modular fashion. These groupings are an ontology of data entities and elements that describe a subject area. As shown in *Figure 3*, a grouping can be thought of as a "slice" of the integrated data model. This is usually organized from the point of view of a primary business function and one or more source systems that operationally support the function. Thus by sourcing data from these systems and loading it into the data warehouse environment we can have sufficient data to provide decision support and analytics for that business function. Within the healthcare provider sector, some example groupings can be: patient/visit/ treatment data; pharmaceutical related data; supplier/logistics/facilities management data; academic/training/human resource data; general ledger/accounts payable/accounts receivable data; and many other potential groupings.

4800 Spring Park Road
Box #18
Jacksonville, FL 32207
Office: 866.611.8595
Fax: 877.893.1410
www.metaanalytix.com

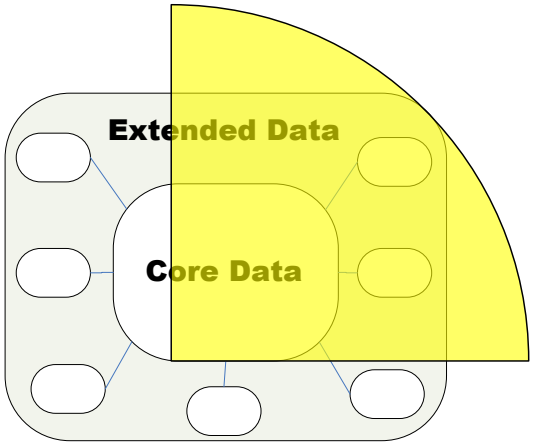


*Figure 3: A "slice" of the integrated data model that describes a primary business function*

A second slice is shown in *Figure 4* containing the grouping of data for another primary business function. Once again this set of data can be sourced and loaded into the data warehouse environment in a modular fashion. The important thing to notice, however, is that there is an overlap within the core data with the first slice. Some of the same data that is needed to describe the first business function (yellow slice) is also needed to describe the second business function (blue slice).

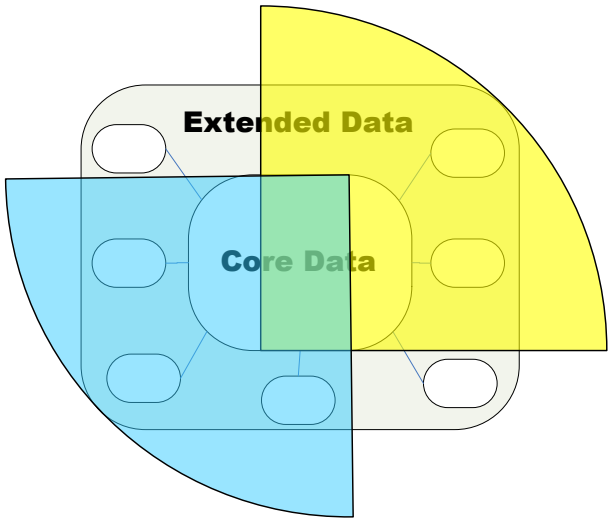


*Figure 4: Data Groupings for Two Different Business Functions, Showing Their Overlap*

4800 Spring Park Road
Box #18
Jacksonville, FL 32207
Office: 866.611.8595
Fax: 877.893.1410
www.metaanalytix.com

This overlap is shown as the green area rectangle, and, of course, is to be expected. But the way that it is handled is crucial to the sustained success of the data warehouse environment. The integrated data model needs to keep a singular data model, meaning that there is only one definition for a business entity and attributes are in one and only one place. Therefore, the areas of overlap from the sources for the second business function need to be integrated into the same data model that was delivered for the first business function. It is the strict observance of this practice that results in an overall integrated data model. By planning broadly across a number of business functions and the corresponding source systems, the work program can be modularized and prioritized thereby allowing for incremental delivery of value based on the needs of the business.

For example, a core data grouping of "patient" might be relevant to both business functions of "improving quality of care" and "reducing cost of care". Depending on the needs of the business, the data maybe further modularized as shown:

**Improving Quality of Care**

a) Reducing medication errors
b) Reducing patient wait times
c) Reducing central line infection rates

**Reducing Cost of Care**

a) Patient's average length of stay
b) Smoking cessation counseling
c) Average daily patient census

So, even though the core data definition of "patient" remains the same throughout the enterprise, different attributes of it may be used to group it along the needs of the business. Between the two business functions there is a lot of common data including patient visits, diagnoses, treatments, costs, age, etc. Yet, as indicated by the lists above, there is additional data that Quality of Care may be interested in using that isn't needed by Cost of Care and vice-versa. It is important that the business function that is sourced first, establishes the common data into an integrated model so that it is available to the second business function which then only has to add the additional data from sources not yet extracted.

## Putting It All Together

Now that we have a definition of the logical reference architecture and the integrated data model, it is time to discuss how to organize and manage the workstream for delivering a data warehouse environment. The basic "recipe" for the first work-stream is:

4800 Spring Park Road
Box #18
Jacksonville, FL 32207
Office: 866.611.8595
Fax: 877.893.1410
www.metaanalytix.com

1) Identify the primary business functions that you first want to be able to support. Prioritize these functions in rank order.
2) Identify the sources that are required for the first business function. These sources will probably contribute both core and extended data.
3) Source all the data that the sources provide, even if there are data elements that are not needed by the business function being targeted. It is very low cost to move these extra data elements into the Data Landing zone of the data warehouse environment. This is less costly than having to go back at a later point in time and re-evaluate a source system to gather additional data elements when they are needed by a business function.

> **"Critical to long term success is being able to organize work-streams based on the business data that supports business functions"**

4) Design the initial integrated data model taking into account the data needed for the business function. Pay special attention to the design of the core data model. Focus on the data elements required by the business function being supported. Leave behind the extra data elements in the Data Landing zone that have been sourced but do not have a business requirement held against them. While sourcing extra data is low cost, designing extra data into the integrated data model can be time consuming so this is best left as a deferred activity once the business justification is evident. The exception to this rule is if it is straightforward for how to incorporate the extra business data elements into the design. In some situations this may be the case, and, if so, the advantage of integrating this data is worth it because little additional effort will be spent.
5) Load from the Integrated Data Store into the Data Usability layer. This topic will be covered in more detail in the section on data provisioning later in this paper.

For the remaining work-streams, the steps are:

1) Select the next business function. Re-evaluate and re-prioritize as needed.
2) Same as step 2 above. Keep in mind that, if a source from any of the prior iterations has already been extracted, its data will already be in the Data Landing zone or the Integrated Data Store.
3) Same as step 3 above.
4) Extend the design of the integrated data model by incorporating the additional data. As mentioned earlier, this is easier said than done, and careful data analysis and design techniques are needed to properly evaluate the additional data elements.
5) Load from the Integrated Data into the Data Usability layer. This includes providing additional data into any other already existing data marts in addition to supplying the data for the new business function.

4800 Spring Park Road
Box #18
Jacksonville, FL 32207
Office: 866.611.8595
Fax: 877.893.1410
www.metaanalytix.com

Critical to long term success is being able to organize work-streams based on the business data that supports business functions. While business functions are fairly distinct, data is shared between them. Thus an issue with data quality generally has an impact to more than one business function. The data warehouse environment builds upon itself in the sense that, as more and more data is added, there is increasing potential for more sophisticated decision support and analytical business functions to be added, though it is also important to bear in mind that, in the beginning, core data will predominate over extended data. While both core business data and extended business data are needed to perform advanced business functions, you can performs basic business functions without requiring extended data.

## Data Usability Layer Revisited

Now that we have a firm understanding of how to organize and manage the work- stream, let's wrap up by taking a closer look at the alternatives for creating data marts within the Data Usability zone. While there is a single integrated data model within the Integrated Data Store, there does not have to be a single model accessed by all business domains in the Data Usability zone. Various business domains, such as Finance, Clinical Care, or Marketing may have a dedicated data mart created for their use with the subset of data from the integrated data model that is needed for the decision support and analytical business functions within the domain.

Typically, these data marts will organize the data in a dimensional model or star schema. This may be a different structure from the integrated data model which might be designed based on a more normalized relational data model. While this means that the data is structured differently in the data marts, it is important to note that there is no additional data or information created when the change in structure takes place. The conversion to a dimensional model is done to make access easier for end users whereas the normalized model tends to make it easier to manage as an integrated model. It is generally a straightforward process to create the dimensional model.

Even though there may be multiple data marts in the Data Usability zone, they all are produced as subsets of the integrated data model. Each data mart does not have its own distinct data model, although it might appear that way. In fact, if the data marts are all compared, much of the same shared core data will be evident in each of them. Yet each data mart can be individually managed in terms of when, and under what conditions, it gets updated. As mentioned earlier the data is replicated in the data marts, meaning that it is controlled and managed from the integrated data store which is the source of record. At no point is there duplication of data, which is a copy that is not managed and is uncontrolled.

4800 Spring Park Road
Box #18
Jacksonville, FL 32207
Office: 866.611.8595
Fax: 877.893.1410
www.metaanalytix.com

There is no hard and fast rule that requires separate data marts to be built for each business domain. A single mart can be created containing all dimensional models needed across all business domains. The decision factors for determining how to organize the data marts depend largely upon accessibility, performance, availability, security, and various other requirements. Security and data access control play a particularly significant role in this decision for the provider healthcare sector since privacy is essential and heavily regulated. This requires certain data to be de-identified, obfuscated, or completely filtered when used by many business functions. This typically results in separate data marts for each business domain so that data security and access can be better managed and controlled. For example, medical cost analysis requires access to clinical data of patients but it does not require patient identifying information, which is prohibited by regulation. Thus a data mart containing de-identified patient information with detailed clinical data may be an attractive approach for more general access, while a data mart containing patient identifying data without detailed clinical data (and tightly restricted access) may also be appropriate.

**"There is no hard and fast rule that requires separate data marts to be built for each business domain"**

## Conclusion

By planning broadly and executing modularly, it is possible to deliver near term and sustainable business value for decision support and analytic functions in the health care provider sector. Doing so requires careful management of the integrated data model and data integration processes. Leveraging data models, architectural and design patterns, and ETL best practices are an important enabler to get started quickly, deliver early results, and remain on the right track.

**About the Authors:**

**John Majarwitz** has worked professionally in information systems and technology areas for over 25 years, including as a Chief Architect at AT&T Bell Laboratories, and, currently, as Vice President of Solutions and Technology at Meridian Technologies, Inc, an information technology consulting firm. John also operates End To End Information Solutions, which is dedicated to technical publication and information dissemination, and holds B.S. and M.S. degrees in engineering from Yale and Carnegie Mellon Universities.

**Kishore Nair** is an award winning technologist and entrepreneur. Kishore has a wealth of knowledge in technology, business intelligence, data management and entrepreneurship. In the past, he has worked at Fortune 100 banking and healthcare companies, leading several multi-million dollar technology initiatives and informatics initiatives. He is currently the CIO at Meta Analytix. He has a Master's Degree from Birla Institute of Technology and Sciences, Pilani, India.

4800 Spring Park Road
Box #18
Jacksonville, FL 32207
Office: 866.611.8595
Fax: 877.893.1410
www.metaanalytix.com

**About Meta Analytix:**

**Meta Analytix is a comprehensive informatics solutions and consulting firm with a focus on Healthcare providers. Our solutions, IATROMATIX, INTEGRA (the first BI appliance for Healthcare) and Knowledge On Demand (SaaS) provide affordable business intelligence solutions to organizations of all sizes. To learn more, please visit us at http://www.metaanalytix.com**