Business Is Intelligent $^{SM}$

PICAROS SOLUTIONS
www.picarossolutions.com

# Enterprise Data Infrastructure

## A Strategic Approach

Reporting and analytic needs of a business have been met in many ways as we all know and as the industry will summarize, typically ends up in two broad categories:

- A plethora of datamarts built independently of each other and potentially when the situation starts to become unmanageable or complex enough, consolidation seems to happen and a warehouse materializes
- A warehouse would have been conceived and started and reporting marts produced out of it

In many cases, the efforts would have been embarked on with no real enterprise focus and the changing needs of the business, where each project was self contained in content and any additional need had to be met via an elaborate mechanism. The ability to leverage previous projects was limited, or at best scattered. However, as the organizational demand for information grows, it is inevitable that short-term solutions would become very expensive to deliver and sustain. This would bring to bear the problems that IT comes to deal with eventually.

Most organizations do not have a clean slate to start thinking at an enterprise scale, and those that are not “big enough” might make the mistake of starting in the same path and fall into the same pitfalls as others. A strategic way of approaching the data needs for the business (regardless of size in my opinion) has become a necessity if an organization is to remain competitive and successful. This would enable an organization to capture timely, accurate and consistent information eliminating the need to for multiple one time datasets to deal with individual business requests, thereby making it simpler to consolidate information, simplification of support, increased response time to business requests with lower incremental effort, as well as better business decision making.

This strategy implies a paradigm shift where solution design while focusing on meeting the immediate business need, should ensure that the long term interests of the organization is incorporated into the design. This strategy is easier said than done and all the things we all know and talk about have to happen:

- Management support for the vision – sell it if necessary
- Adjustment to funding strategies – marginal increase to incorporate a small amount of forced scope creep
- Aligning the IT teams to this view of the world and why we are building a little more

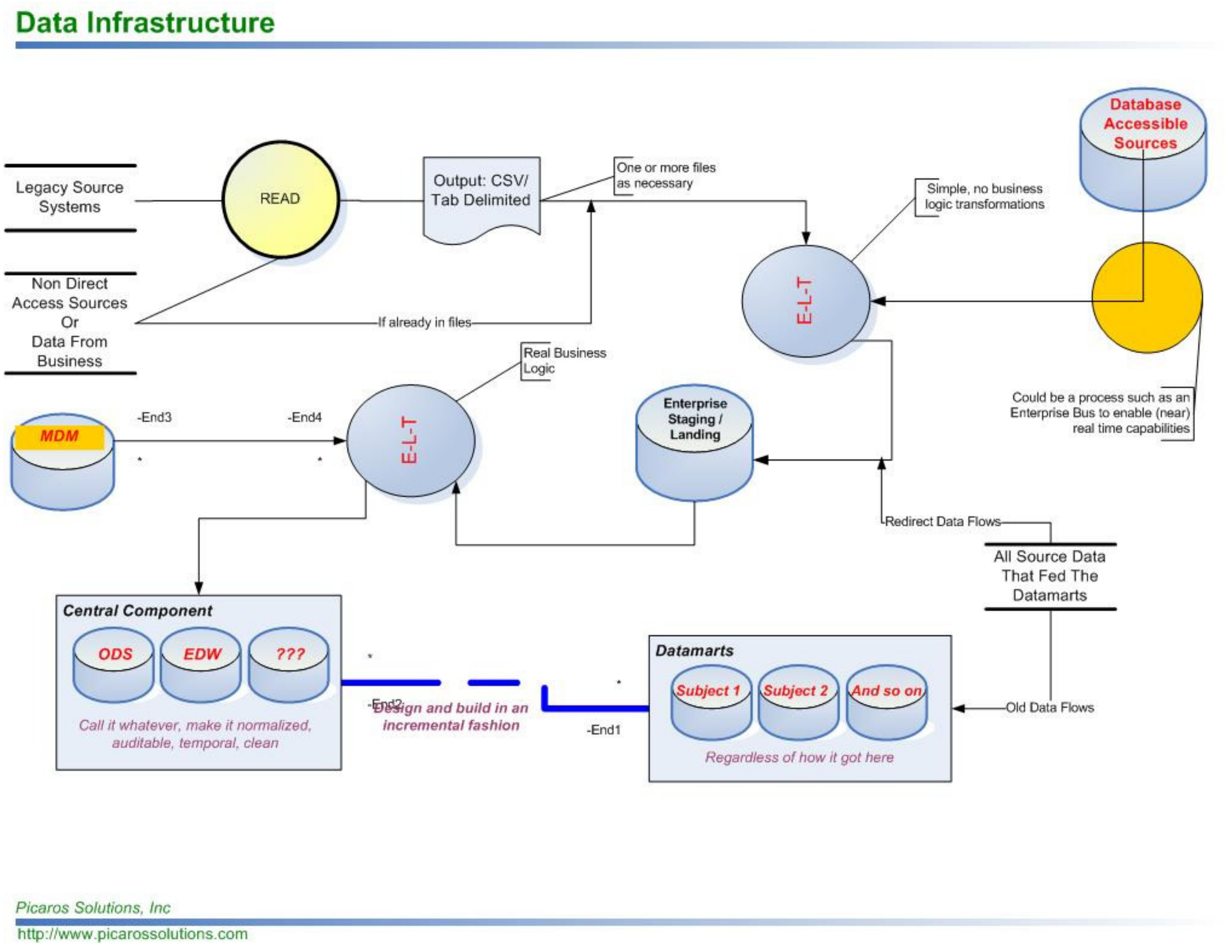
If done right, this strategy will foster an organic growth of information at the enterprise scale to aid in establishing the foundation for an enterprise data infrastructure, albeit within the boundaries of immediate business necessitated project and the funding therein.

So, what will this buy us in the long run:

- A scalable, extensible, reusable, environment
- Industry standards and best practices can be more easily incorporated (smaller bites)
- No special funding to scare business in investing more as it is a “marginal increase” to project funding that was or is being approved
- Projects delivered are leveraged for next project resulting in sizable gains in effort

- The rhetorical but useful concept of  a single version of truth, to ensure alignment between strategic and tactical objectives
- And so on

**Fig1: High Level Architecture**


Data Infrastructure
Database Accessible Sources
Legacy Source Systems
READ
Output: CSV/ Tab Delimited
One or more files as necessary
Simple, no business logic transformations
Non Direct Access Sources Or Data From Business
If already in files
E-L-T
Real Business Logic
Could be a process such as an Enterprise Bus to enable (near) real time capabilities
MDM
-End3
-End4
E-L-T
Enterprise Staging / Landing
Redirect Data Flows
All Source Data That Fed The Datamarts
Central Component
ODS
EDW
???
Call it whatever, make it normalized, auditable, temporal, clean
-End2
Design and build in an incremental fashion
-End1
Datamarts
Subject 1
Subject 2
And so on
Regardless of how it got here
Old Data Flows
Picaros Solutions, Inc
http://www.picarossolutions.com

The philosophy behind this architecture is simple, but a few salient aspects of this are as below:

**Enterprise Staging/Landing**:

- This is considered the entry point of significance for this architecture
- Intention is to let all kinds of data flow into this space with little to no transformation. Unless a source system allows for special datatypes that the physical implementation of this architecture will not support, the data comes in as-is. Examples of data here:
  - Source system data at the lowest grain that will ever be needed for reporting/analysis. There would appear to be a definite overhead here if the immediate business need does not warrant this grain, but that can be mitigated with careful analysis and engaging the business from a short term and strategic need perspective
  - Business provided data in spreadsheets, Access databases or csv files
  - Textual extracts coming from third party sources
- Periodicity of data or the grain of the temporal aspects can be controlled here, again a meaningful analysis will enable the decision around capturing every wrinkle of data captured in the source system or capture data from sources once or a few defined times of the day
- If there are existing data feeds from different sources creating datamarts, these need to be redirected to the staging area till the needs can be recreated using the overall strategy of sourcing granular data from sources

- Can be leveraged to do data analysis for quality and completeness in case source systems do not present a viable option
- Careful planning of archival process to ensure the area does not become unmanageable

**EDW/ODS/???**:

A name is not as significant in my opinion as what the objective/outcome is;

- This is considered the central/core aspect of this architecture
  - There are other methodologies – those that advocate an ODS or a warehouse or both. Personally, other than rhetoric I do not see a significant value in so many different schemas. Understanding the short term and long term objectives of the organization's reporting and analytic needs, it should be fairly a defined process to pick a strategy.
- The primary design objective here is a "normalized" design integrating an organization's breadth of data from multiple sources
  - The "normalized" aspect comes with a caveat: controlled redundancy and controlled violation of the rules of normalization as conventionally accepted and the primary purpose of this is to tackle performance issues and/or to make ELT development easy
- The design goal might sound esoteric at some level, however is simplistic in premise: for any subject area being designed (customer will probably come first) it should be done comprehensively whether each aspect of it is developed for immediate needs
  - Why customer, if an MDM solution is in place. Chances are the MDM does not exist, and if it does, it is another source of data for this layer
  - While it is probably not possible to create a full design due to time, budget and other constraints, the design should incorporate necessary hooks to expand it seamlessly as the architecture grows, in such a way that the incremental growth does not affect what has already been built
    - Spawn more dependant entities to capture hitherto unknown details about the customer (for instance), rather than alter existing structures
- The Staging layer is the primary source of all data with the exclusion of an MDM layer if it exists – else create one to capture and maintain all reference and lookup information that will eventually help in slicing/dicing for analysis purposes
- This layer should incorporate all of the following objectives:
  - Lowest grain of data to satisfy all reporting and analytic needs
  - Integrity of the data
  - Auditing capabilities to ensure whatever is stored is valid, traceable and verifiable
  - History: as-of analysis and reporting
  - Data quality processes

**Analytics and Reporting Datamarts**:

Then comes the layer(s) that will actually be exposed to the users who will be presented all the information being gathered in a simple, friendly manner depending on the need. Our strategy here is to slowly and effectively retire datamarts that were built directly with specialized feeds over time to remove disconnects in the data coming from the disparate data sources and the integrated one. Basically drive towards a situation where the single source of truth can actually be leveraged effectively.

In conclusion, the core nature of this architecture should be a key component in a reporting/analytic infrastructure. However, there are challenges that we need to confront. How should the development and deployment happen to ensure success? How or what needs to happen to gain business and management support? Unfortunately, there is no silver bullet. An organization's internal workings will determine what the road map for this looks like, and how to incorporate this within the decision making process in the organization.