Andy Dilkes is a senior partner in LakeFront Data and CTO responsible for the Reference Data Factory data management suite

Data Management

Acquisition, Integration & Distribution

Myths and Lessons Learned

The management of reference and enterprise data in the financial services enterprise has proven to be a very challenging initiative. For the most part, the difficulties experienced by these initiatives can be boiled down to a handful of missteps. This article will describe a view of the "seven deadly sins" that can easily mark the demise of an enterprise data project. These projects are typically high exposure and large budget initiatives leading to constant monitoring, review and scrutiny from top level executives. In the end however, the success of the program will ultimately be measured by the rate of adoption across the enterprise.

Reference data projects initially achieve high marks and are deemed successful through the first set of deliverables or the first phase of the project. This is due to the initial development phase being engineered under the direction of the "first client", offering a clear set of requirements and therefore providing a simple measuring stick for success. The challenge, and ultimately the true measure of success, lies in the ability of the program to quickly and seamlessly adopt new data sources and data consumers into the program.

Given detailed requirements in terms of data content, data representation and success criteria we can assume with reasonable certainty that most large scale projects will be able to declare success through the first client adoption of the program. While this is a positive and necessary milestone, extreme care must be exercised to ensure that the solution architecture, infrastructure, capabilities and representations are engineered in a forward thinking and adaptive model. Too often program development falls into the trap of building to suit the 'first client' and loses sight of the ultimate service oriented goal.

The following lessons learned are based on beliefs, approaches and methodologies that have resulted in non-adaptive solutions that reduce the rate of success in consumer adoption. The result is a constant re-engineering and maintenance cost, significant delivery delays and ultimately program failure.

♦It's Just an ETL

The first deadly sin is the assumption that a reference data solution is simply an ETL from a vendor or in-house data source into a database table. It is easy to assume that data can be loaded into a database table through any reasonably mature ETL tool and voila… done. The problem does not need to be viewed as complex (as in some of the proprietary solutions), but experience makes it clear that a simple table loading infrastructure will not solve the problem.

The data delivered from vendor sources is typically a proprietary reflection of the vendor operational implementation. Data is either represented through their data model (snapshot) or in a transactional representation of an event that took place within their operational environment. ETL tools and other simple loading mechanisms lend themselves well to acquiring and persisting data delivered in the form of a snapshot. But ETL tools lend themselves very poorly to managing representations that depend on content interpretation, current state and processing sequence to ensure integrity. Given the complexity of some of the data sources to be processed, it's clear that some level of transactional capability is required. To this point, even snapshot type of sources would be better viewed in a transactional model that supports some degree of auditing, delta determination and re-processing capability. An ETL infrastructure may get you through the "first client" deliverable but will likely fail to provide scaling going forward.

♦It's Just Data

Vendor provided data and even internal data sources can be very complex. In general they are a representation of an operational model that serves as the basis of a complex data management application. Each vendor and data source will have their own unique characteristics regarding content, granularity, delivery format, delivery technology, process sequencing and

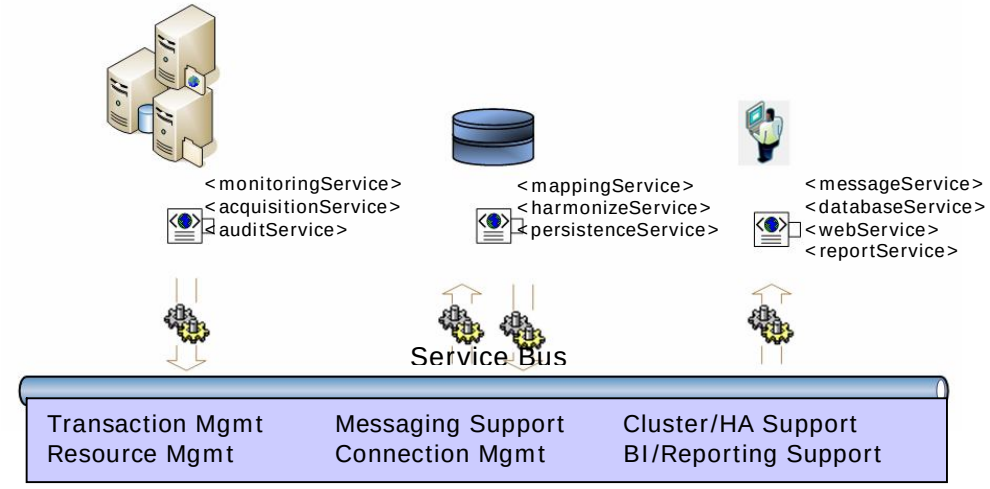
data dependencies. The documentation delivered with vendor products describes the technical content or even the data content at an attribute level. The available documentation typically falls short of describing the 'bigger picture' in terms of data granularity, dependency and cross product integration.

Acquiring the knowledge to effectively harvest vendor data products requires a great deal of time and 'trial and error' investigation. Data misrepresentation will compromise the integrity of the program and lead to expensive reengineering costs. A reference data project should engage a subject matter expert with specific and extensive experience in the data products that will be included in the program. Even if viewed as an overhead expense, this will pay for itself many times over.

♦ **Build It…. They Will Come**

No they won't! Of the challenges facing an enterprise reference data program, promoting adoption across the enterprise is second to none. The business case for a centralized reference data program is often clear and the financial and operational benefits of a successful implementation are substantial. In the end however, just like any other business, your success will be based on adopting and retaining customers.

Given the value of the program from an overall corporate perspective, consumers are compelled to engage. However as an established operating entity, they are likely to display an understandable reluctance to change, especially change that will introduce significant resource consumption or risk to their operation. The lack of willingness to participate will quickly reduce adoption unless you can provide a seamless integration that minimizes operational impact. In order to secure consumers, the program has to display an adaptive capability and have virtually no expectation of changing the consumer's operational model. The adaptive data enterprise has been able to satisfy these requests. A capabilities model that includes technology, format, content and source independence will go a long way toward impeding any barriers to adoption.



♦ **Build Infrastructure as You Go**

A major misstep that many reference data initiatives fall into is the development of infrastructure or capabilities as part of the data engineering aspect of the program. The engineering of data representations, by its nature is a vertical development effort to acquire, manage and represent data from different sources and technologies and deliver the data to downstream consumers. This aspect of the development process is ongoing and will continue as long as you are bringing on new data sources and consumers into the program. In an adaptable environment, this is not a code development exercise but rather an exercise to configure and deploy services that manage new data sources or consumer within an established infrastructure. If you are developing, testing or even enhancing this capabilities model throughout the process of consumer take-on, your chances of successful integration drop dramatically. It is critical that this infrastructure is in place and deemed production ready so that the focus can be on integrating new data sources or consumers. The program should be about data representation, not about developing complex capabilities on the fly.

A critical consideration in the engineering methodology is the 'buy vs. build' argument. The in-house development of infrastructure designed to handle messaging, connectivity, resources and failover is difficult, resource intensive and prone to failure. If you are considering a commercial proprietary solution you must also consider the infrastructure on which it is based to ensure extensibility, adaptability and operability.

♦ It's All About the Golden Copy

Most reference data initiatives start off with the notion of a 'Golden Copy'. It's the idea of the golden copy that seems to provide the conceptual basis for a centralized reference and market data repository. The concept of the 'Golden Copy' implies "this is it"; the final consolidation of all data sources ready for distribution. There was a time when this approach may have made sense, when the profile of data consumers was alike. Today's consumer base is diversified with vastly different requirements across all dimensions of reference data. Consumers require personal subscriptions into the universe of financial instruments with the capability of defining their own level of granularity, preferred sources and change indicators in addition to delivery content, format and technology. In essence, the consumer requires their own personal 'view' that delivers definable content under definable circumstances through a definable format and technology.

The deployment of a centralized reference data model handcuffs the project's ability to provide a multi-dimensional view of the data. Rigid data models imply that a representation of a financial instrument has been pre-determined with no consideration for a consumer specific requirement. Given the diversity of the potential consumer base, a project would be better served with a centralized repository of facts from which a consumer specific view, defined both horizontally and vertically can be derived. The idea of a one time centralized derivation debilitates a fluid and adaptable data representation model.

In general, a golden view of the data must be maintained at the consumer level. A single golden copy will fail to meet the needs a diverse and ever growing consumer base.

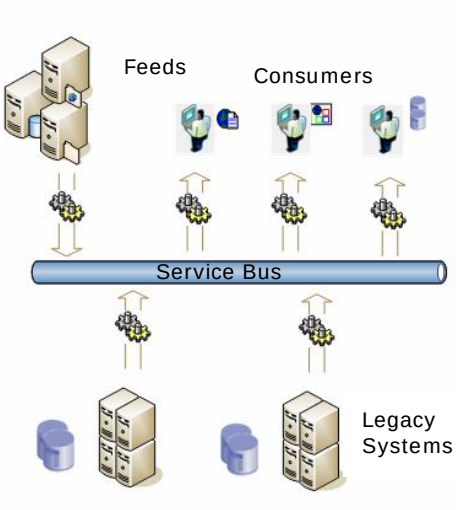
♦ Ignore Your Legacy

The investment in existing (legacy) systems should not be summarily dismissed. Maintaining and upgrading legacy systems is one of the more difficult challenges CIOs face today. Despite their obsolescence, legacy systems continue to provide a competitive advantage through supporting unique business processes and containing invaluable knowledge and historical data. However challenging, harvesting legacy system data often expedites time to market, minimizes cost and reduces risk.

An adaptive data management platform should include an integration capability that extends beyond the acquisition of vendor data from flat files. Integrating diverse data sources including legacy systems is critical to providing a central standardized service implementation that can adapt to a consumer business unit without modification. A service infrastructure should display the ability integrate existing systems as a service component within the framework. In this infrastructure the exposure of legacy data becomes standardized from a consumer point of view. All data, irrespective of the ultimate source is accessed by the consumer through the same interfaces.

♦ It's All About Technology

Developing and/or implementing technology capabilities that acquire, manage and distribute reference



throughout the enterprise are obvious and necessary mandate for a centralized reference data initiative. The current buzz around reference data management programs (sometimes referred to as Enterprise Data Management), covers not only the technical capabilities model but also organizational management model. Both these models must interact at virtually every step of the process.

The recent push toward management of reference data with a focus on governance and stewardship is an inevitable evolution given the complexity of vendor relationships, the ever growing demand for this data throughout the enterprise and the focus on mitigating operational risk. The development and application of governance policies is a critical aspect of a successful reference data program.

www.lakefrontdata.com

(416) 364 8575