The Secret to Data Lake Success: A Data First Strategy

by Rick Sokolosky

JUN 2, 2015

Good data warehousing practices teach us to start with business requirements, get only the data we need, build it all into a highly-engineered dimensional model that links everything to everything, and lock down the data once it is created. After doing that for 20 years now, we're pretty good at it. There is only one problem. It doesn't come close to delivering access with the speed and agility that businesses really need from their data. Data warehouse managers live in constant fear of the next berating over failure to deliver, the time necessary to deliver, and the costs to keep the lights on, let alone add new information. Where did we go wrong?

We forgot about the data.

Need for Speed – and Flexibility

There are two main complaints leveled at the data warehouse, outside of cost:

1. By the time the answers are available, the questions are no longer important

2. The business cannot answer the questions themselves

So why is the data warehouse failing to deliver on these requirements? The organization spent a lot of time and money to create a "slice and dice" environment that should give the business what they need. Unfortunately, in today's environment, accounting for every question in one model is impossible. New data sources are emerging at a breakneck pace. New questions are sprouting up even faster. A highly engineered environment that only takes the data it needs a upfront is going to have difficulty adapting to rapidly changing requirements.

Faced with the complexity of data loading and transformation processes, as well as a highly intricate data model, the average data warehouse change takes nine months and costs over $1 million to complete. When you build a complex system with one purpose in mind, don't expect agility.

Start with the Data

We can't keep trying the same thing over and over again; businesses have had enough. Meanwhile, semi-pro data warlords are springing up all over the enterprise, creating their own fiefdoms with un-regulated and inefficient processes to get at the data they need. At least the data warehouse kept the house in order, even if it missed the mark on discovery analytics. The answer is to start with the source

data itself, making it an asset that should be available to those who need it. This is the whole idea behind the Operational Data Store (ODS) and more recently, the Data Lake. Source data should be centrally managed and made available across the organization.

Rigid approaches and IT "soup to nuts" data and analytics ownership have blocked the information-creating process enough to provoke isolated data fiefdoms in many organizations. Business requirements should drive what sources are taken in their entirety into the environment. For even better results, that source data should be lightly integrated and standardized into easily understood subject areas for greater usability. *Data First,* made possible by new developments in big data technologies and approaches, is the answer to these needs.

*Data First***: The Principles**

Flying in the face of what we learned as best data practices for data warehousing, *Data First* operates on these principles:

1. **Data storage is cheap and no longer a constraint** – Data can now be reproduced as many times as needed, in the necessary forms, at an acceptable ROI.

2. **Centralized data efforts are focused on subject areas** – Put time and effort into making the data usable across a host of purposes in collections of like data vs. creating it for a specific purpose.

3. **Encourage data discovery** – By making the data available, the "warlords" can discover new answers with total control while contributing to the overall ecosystem.

4. **The businesses should have control of their analytics, not the underlying enterprise assets** – It is not the business-led analytics that cause issues, it is the redundant and inconsistent acquiring of source data that creates problems. The business has the right to expect a comprehensive data asset that enables their analytics and consequently, their decisions.

Filling in the Gaps

By incorporating *Data First* into the process of creating information, businesses can create the analytics they need at the pace they require. The data assets are available to the organization, and data discovery helps feed better function-specific repositories, such as data marts. That speed and flexibility are the beauty of *Data First* – the missing piece to finally let us fill the gaps left by the data warehouse.

Rich Sokolosky is partner, life sciences practice leader for [NewVantage Partners](#), a provider of data management and analytics-driven strategic consulting services to Fortune 1000 firms. The company specializes in big data strategy consulting, thought-leadership, execution, and business value realization.

For further information, contact Randy Bean, Managing Partner, NewVantage Partners at rbean@newvantage.com.