Intellyx EA Communiqué

Jason Bloomberg, April 5, 2016

Communicating the Data Lake vs. Data Warehouse Story

Business stakeholders are increasingly calling upon enterprise architects to facilitate the massive and complex technology-driven transformations today's enterprises face. As a result, EAs often find themselves in a 'connect the dots' role, interpreting a complex technology story for a broad audience.

This EA Communique focuses specifically on the transition enterprises face as they move from *data warehouses* to *data lakes*. To a non-technical stakeholder, these two technologies appear to have similar purposes, and given the proliferation of buzzwords in today's overheated tech marketplace, there are bound to be skeptics who come to the conclusion that data lakes are nothing more than data warehouses with new price tags.

Walk over to the part of the cube farm where the data architects hang out, however, and you'll get a very different story – typically one that centers on a multitude of technical minutiae and a laundry list of odd-sounding open source product names.

Fortunately, the EA can help with both sides of this challenge: communicating the true value of data lakes as well as their differences from data warehouses to business users, while interacting with the data specialists to ensure their efforts align with ever-shifting business priorities.

Beyond the Horseless Carriage

Whenever a new technology comes along that promises to supplant an older one, people tend to think of the modern alternative in the context of the familiar – a phenomenon known as the *horseless carriage syndrome*.

The starting point for data lakes, predictably, is to think of them as a modern alternative to data warehouses – but this perspective shortchanges the power of the data lake. Once the business gets past this misconception, only then will they be able to leverage data lakes to their fullest extent.

Understanding the differences between the two, therefore, is part of the EAs challenge. The larger challenge, however, is understanding all the purposes that data lakes can serve that nobody ever envisioned for data warehouses.

A great place to start is the white paper *Will the Data Lake Drown the Data Warehouse*? by Mark Madsen of Third Nature and sponsored by SnapLogic. While Madsen didn't write this paper specifically for EAs, it can serve as a valuable resource for informing those conversations EAs should have with both business and technical users about the differences between data warehouses and data lakes.

Drowning the Data Warehouse

Madsen first places the data warehouse into the context of its history as a reporting platform for mainframe applications. Data warehouses centralized data and resolved issues of conflicting mainframe workloads, thus separating data access from transaction processing.

Data warehouses were quite successful for a number of years, but as new data requirements gradually emerged, people attempted to force new workloads into the older data warehouse architecture – an approach that rarely worked well enough to meet business needs.

Over time, data warehouses gradually became less of a vitally important business tool and more of a "data junkyard" where data go to die. Clearly, positioning a data lake as simply a replacement for a data warehouse won't address this perception. Organizations can put a wider variety of data structures into a lake, but that doesn't mean that we're simply building a better junkyard.

It's essential, therefore, for the EA to help dispel the "junkyard" context for data warehouses when discussing data lakes. Like junkyards, data warehouses are the end of the data lifecycle, but data in a data lake may serve other purposes well beyond the lake itself.

In other words, data lakes are one step on the path that data take as they provide value to the organization, more so than an endpoint simply for the collection of data. While data warehouses serve as the final resting place for data, for data lakes, the endpoint may be another application entirely.

In fact, data lakes can serve as a place to work on the data in ways that data warehouses never could. Where data warehouses separate data processing and storage with traditional extract, transform, and load (ETL) tooling, data lakes provide greater flexibility as to the timing and nature of transformations or other processing, including the standardization, cleaning, and aggregation of data. (See my earlier article on the timing of processing in data lakes.)

DATA LAKES CAN SERVE AS A PLACE TO WORK ON THE DATA IN WAYS THAT DATA WAREHOUSES NEVER COULD.

The fact that data lakes can process data is especially important considering the variety of data types and structures that lakes can accept. Perhaps the first thing a business user will learn about data lakes, in fact, is that you can put all manner of different kinds of information in them, from tabular data to text-based data to rich media like audio or video. The challenge, however, is understanding how this variety of data won't just end up as an unmanageable mishmash.

The answer is to understand that a data lake is a *platform* – a platform in the sense that other applications can run on top of it. A data lake might support log management and analysis apps, focused on recognizing root causes of problems in the operational environment. Or the same data lake might drive a natural language analysis application, understanding customer communication or other text-based, natural language tasks.

The list of potential uses for a data lake is endless – a marked contrast from the limited uses of a data warehouse.

The Intellyx Take

Part of the data lake story is support for big data, of course – but from the business perspective, big data challenges are typically separate from the everyday business data tasks business users have been dealing with for years.

As a result, the EA's challenge isn't simply helping the business understand and leverage big data-centric solutions, but also to incorporate modern data lake functionality into aspects of the business that people don't think of as big data – in other words, 'everyday' data.

Furthermore, data lakes are inherently more complicated than data warehouses, by virtue of their versatility as well as their storage and processing capabilities. Supporting this level of complexity, in fact, is the primary architectural challenge facing data lakes.

By their very nature, data lakes don't stand alone. They are a central part of a diverse and highly scalable ecosystem of applications, tools, and other infrastructure. Orchestrating such components while maintaining the flexibility and scalability so essential to the data lake value proposition is a critical capability, and one that SnapLogic brings to the table.

DATA LAKES ARE A CENTRAL PART OF A DIVERSE AND HIGHLY SCALABLE ECOSYSTEM OF APPLICATIONS, TOOLS, AND OTHER INFRASTRUCTURE.

SnapLogic is an Intellyx client. At the time of writing, no other organizations mentioned in this article are Intellyx clients. Intellyx retains full editorial control over the content of this article.