HADOOP DATA LAKES CAN BE SHARK-INFESTED WATERS

Steve Swartzlander, Lead Architect 8/18/2017

Many of today's enterprises are embarking on "big data" projects to realize the benefits of various types of analytics such as predictive modeling, machine learning, natural-language processing, and others. A key part of such efforts includes gathering volumes of data into a large distributed repository often called a "data lake" which is frequently hosted in a Hadoop cluster.

Such projects come with hazards lurking below the surface. Building a small Hadoop cluster in a lab-like environment for testing and experimentation is rather simple. Building one for a large enterprise that must be secure, stable, and highly available is another matter, particularly if one is in a highly-regulated industry that deals with sensitive data.

It is necessary to accept a higher-than-usual level of uncertainty when building out a Hadoop environment. Change is rapid across the ecosystem. New Hadoop-based projects and components are constantly being introduced. Most of them are open source. A production Hadoop cluster will often include key components that are still Apache incubator projects at a 0.x release level.

Not all answers will be found in a vendor's documentation. That's true of most things in IT, of course, but it is magnified with Hadoop. Online communities are critical, as is a certain willingness to dive into code, poke through obscure files, and experiment.

Security can be a particularly interesting challenge with Hadoop. Its authentication mechanisms use Kerberos, which can be implemented in a couple major ways. Engaging the enterprise's IT security team early is important. They may be unfamiliar with the details of Kerberos and need to gain new skills themselves to be effective partners in the project.

A less-technical challenge are the third-party vendors. Word will get out that the enterprise is starting a big-data effort, and vendors will swarm. They will have both software and appliances with a big-data angle that promise to add value. Be wary of these until the intrinsic capabilities of the base components are well-understood relative to project requirements. A particular area of concern is compatibility. Each third-party product added will complicate administration, may introduce constraints on upgrading to new versions of Hadoop components, and may also have limitations in regard to operation with secured clusters.

The bottom line is that going alone into a data lake project is inadvisable. Find a partner who's been there. Consult with those who have already navigated through the barriers. Build something small first, and learn as much as possible before committing to an enterprise-class implementation. Expect to iterate through architectural options before finding a combination that works well for your environment.

A data lake project can be a very rewarding and value-creating effort. Go in with eyes open, a willingness to adapt, a desire to learn, and some expert assistance, and the benefits can be significant.