BUSINESS ANALYTICS

Building the Right Analytical Ecosystem Architecture Takes Shrewd Planning

Coauthored By:
Dan Woods, CITO Research
Ron Bodkin, Teradata Corporation

One of the problems that open source and cloud computing platforms have created is a massive expansion in the number of choices enterprise buyers have to create the perfect analytics environment. For those of us tasked with helping the rest of the business make better use of data, this wealth of choices is mostly a blessing, if you face it with the right attitude.

But, as Timo Elliot sagely pointed out in this cartoon, it can also be a curse.

The goal should be to create an analytics ecosystem where every step forward both solves urgent problems but also expands the collection of reusable data and componentry so that solving the 10th problem is much easier that solving the first. This seems simple enough, but not easy to achieve in an environment that offers so many choices.

The good news is that we have reached a point where the order of battle is clear. We'd like to share what we have learned and what the industry as a whole has learned about charting a path forward toward creating an analytical ecosystem that unlocks the power of data.

The Golden Mean Applied to Technology

Often when a new technology hits the market, its evangelist's feel compelled to make exaggerated claims. This happened when Hadoop, a legitimately new and powerful technology took the stage, followed eventually by other powerful systems such as Spark and Kafka, all based on open source.

The exaggerated claim was that the 25 years of investment spent making RDBMS technology work to serve business was going to be superseded overnight. Everything would be in Hadoop, the one repository to rule them all.

Since then, we have all come to understand that Hadoop has its strong points (storing any type of affordably data at massive scale, performing advanced machine learning, executing simple SQL queries, processing ETL and other data pipelines) but that the RDBMS and MPP systems also are still solving crucial problems such as

- Optimizing execution of complex queries
- Handling massive concurrent workloads
- Supporting huge volumes of modest sized tables
- Providing a mature ecosystem for data labs, security, and governance
- Providing mature integration with enterprise systems and analytics platforms

The best analytical ecosystem has an architecture that uses both commercial and open source alternatives. Like the ancient Greeks, we must seek a golden mean in our application of technology. The number of companies that believe Hadoop will outright replace their existing data infrastructure is down to 3 percent. The use of Hadoop is growing, but research suggests that it does not possess the full requirements of an enterprise-wide, analytical ecosystem.

According to TDWI, "78% of companies do not have Hadoop in their data warehouse environment yet. This indicates that many companies are not sure how to best leverage Hadoop to complement their data warehouses (17% have deployed Hadoop as a complement to the data warehouse).

The same lurch is happening in cloud adoption. All public cloud, all private cloud, all virtualization, all on-premise are all losing strategies. The reality is that for the vast majority of companies, a hybrid cloud approach makes the most sense. Certain applications and workloads are best handled on premise or in a private cloud, while other workloads benefit from the scalability and elasticity of the public cloud. As with open source and commercial software, it's not either-or: it's both-and. This is just as much true for analytics as for other kinds of applications. Often business departments are seduced by the lure of simplicity of standing up an isolated data mart in the cloud. But the long term cost of adding an incompatible, inconsistent, isolated data set is rarely considered.

The Messy Details of Achieving the Right Balance

The challenge we face over and over in our work helping companies understand how to create the best analytical ecosystem is understanding the needs of that particular business. Our team is expert at understanding all available technology for analytics and data science. We can tell you when Spark beats Flink and when Splunk is a better choice than Kafka. We can tell you how to use Teradata in conjunction with Presto or Hive and what workloads should go where. But that knowledge useful after we understand what you really want to get done.

The unfortunate truth is that Hadoop and other open source technologies are still not very well integrated with existing systems, and it can take a lot of retooling on the purchaser's end to align the product to be more than a data repository.

The unfortunate truth is that Hadoop and other open source technologies are still not very well integrated with existing systems

But when integrated as part of an analytical ecosystem, Hadoop can drive new use cases and success stories. For instance, Ancestry.com initially used Hadoop and integrated it into its legacy database platforms, which existed before big data came to the stage. The company at the time had a database of about 11 million customers, which in bits and bytes converted to about four petabytes of information, with much of it from genetic testing. The company used Hadoop to store the data and in-house MPP solutions to make sense of it. They were able to use the MPP solutions to link all of their legacy records with new ones and allow the company to take up an even bigger data endeavor — record matching for the company's new DNA sequencing genealogy offering.

So it can be done. But creating the appropriate ecosystem architecture requires quite a bit of forethought regarding not just a company's data but the resources it has to manage it, both financially and through employees. Here are a few steps companies can follow when plotting out their data environment.

Make the Appropriate Investment

Just because something is open source doesn't mean it's necessarily free. While any proprietary data warehouse will have up-front setup fees, there is a total cost of ownership over time to consider. Most open source companies, regardless of their industry, have to sell solutions on top of their initial product to make it more useful and to stay in business.

Traditionally, creating data lakes in Hadoop required a lot of custom coding in Java, Scala or other low level languages. The emergence of higher level data lake frameworks like Think Big's Kylo (to be open source in February 2017) or the commercial Podium Data have made this dramatically easier.

Conversely, open source engines for querying data are significantly inferior to best in class commercial data warehouse technologies. The ability to perform complex large multi-way joins and manage thousands of concurrent sessions are key areas of differentiation. This makes it valuable to use warehouses to support production data products like BI analytics on integrated data, dashboards and often complex ETL that requires joins or window functions. The use of a data lake to partially prepare data then integrate it in a data warehouse for mass consumption is the best choice for many companies. A data warehouse is designed to thousands of queries on data integrated from across the business. It's a foundation that serves everyone, as opposed to

Within a data ecosystem, there must be a step that ensures data quality. Sometimes errors a data mart focused on a single analytic or business unit. The value of integrated data that can answer many types of questions is not to be underestimated.

Within a data ecosystem, there must be a step that ensures data quality

Make Sure You Have Good Data

Errors occur because of a bad entry, but other times they occur during a system migration. Finding these errors does not occur when the data is stored, but rather in the RDBMS, so selecting one that is up to the task is vital. The solution a company chooses should be reliable and adaptable but also scalable, because the amount of data a company collects is ticking upward all the time.

Let Your Data Talk

The ultimate goal of all of this is an ecosystem that can work efficiently in real-time to give accurate answers to a company's questions. There needs to be transparency so it's easy for a data administrator to know the system is performing well.

An management tool can monitor hardware and the state of the data throughout the ecosystem and then take automated action. If a company needs results from the latest data in one system but it hasn't been uploaded to another part of the system, it can reroute users to right place where the data is currently. Finally, a visualization tool, or advanced machine learning techniques, can provide access and insights anytime anywhere so users can get consistent and real-time insights from the data.

Getting the Right Environment Gets Results

Not all data ecosystems are created equal. And building one is hard. While it may be tempting to go only open source or to jump on the latest public cloud offering, it's better to pick an optimized analytical ecosystem that combines the benefits of open source innovation with proven commercial capabilities and that can work in a hybrid cloud world. The right solution emerges quickly when everyone in the room understand the goals of the business and the power of all of the technology components available and how those components work together.

With such a common understanding, the massive amount of technology choices stops creating confusion, but instead sparks excitement about how much better a business can become by putting data and analytics to work.

This paper was created by CITO Research and sponsored by Teradata

CITO RESEARCH TERADATA.

Ron Bodkin

Founder & President, Think Big, a Teradata Company

Ron Bodkin is Founder & President, Think Big, a Teradata company. He founded Think Big to help companies realize measurable value from Big Data. Think Big is the first and leading provider of independent consulting and integration services specifically focused on Big Data solutions, with expertise on all facets of data science and data engineering.

Dan Woods

Chief Analyst and Founder, CITO Research

I do research to understand and explain how technology makes people more effective in achieving their goals. I write about data science, cloud computing, and IT management in articles, books, and on CITO Research, as well as in my column on Forbes.com.