When a disruptive technology comes along, it completely changes the way things are done, offering such a clear and substantial improvement that it sweeps away established methods. A disruptive technology has now emerged in computing, one that fundamentally changes how organizations use data.

The technology, known as the Data Lake, eliminates the rigid, narrow structures that constrain the data and have long diminished our ability to gain knowledge and insight. And it abolishes the cumbersome, time-intensive process of molding the data to fit those structures. Without these limitations, business and government can collect, manage, and analyze data substantially faster and at far less cost than even the most advanced current techniques. Organizations can also use all of the available data—not just a limited portion—to aid decision making and drive mission and business success. With these and other disruptive qualities, the Data Lake technology developed by Booz Allen Hamilton is rapidly changing the data analytics landscape across government and business.

**BREAKING FROM THE STATUS QUO**

Current computing methods require information to be tailored to individual data structures before analysis can begin. This is typically a labor-intensive task in which analysts spend the bulk of their time simply formatting the data. In addition, only a limited amount of information can be placed into any one of the structures, forcing us to set aside potentially valuable data. Just as problematic, we have to frame the kinds of questions we want to ask of the data in advance. So to gain insight from data to make decisions, we have to guess beforehand at what kind of insight we are looking for—a catch-22 that makes it hard for us to move beyond our preconceptions.

It can be very costly if the guesses are wrong. We have to tear down the data structures, guess again, and then rebuild the structures, hoping this time we are right. This iterative building and rebuilding is such a labor-intensive process that, as a practical matter, organizations typically make decisions using only a relatively small portion of the data available to them.
But what if you did not need the rigid data structures, or the building and rebuilding? What if you could simply choose the data and ask the questions as you went along, quickly and easily refining your queries as you gained greater insight? These capabilities are now possible with the Data Lake technology.

FROM SCHEMA-ON-WRITE TO SCHEMA-ON-READ

The Data Lake’s transformative shift from “schema-on-write” to “schema-on-read” eliminates the need for rigid data structures. With schema-on-write, which underlies conventional computing methods, we must first decide how we want to organize the data for analysis—for example, the row and column headings of a database. The choices we make circumscribe the data that can be examined and define the kinds of questions we can ask. We are already narrowing our possibilities before we have even begun using the data.

Once conventional structures have been built, the extract, transform, and load (ETL) process pulls information from a data source and transforms or molds it to fit the strict parameters that have been set. This way, the schema (basic organization) is imposed on the data as it is written into the structure. Everything is now locked in place. It is difficult to change the structure, the range of usable data, or the types of questions we can ask.

Schema-on-read operates on an entirely different model. Information is pulled from a data source (“read”) and is ingested into the Data Lake, a highly flexible storage medium. Instead of rigid data structures and the ETL process, the Data Lake uses the key/value store, an innovative approach founded on schema-on-read. Individual pieces of data are given identifying tags so that the data can later be located and sorted for analysis. These metadata tags essentially constitute the schema of the Data Lake. But, since the tags do not dictate how the data can be used—they just tell us where it can be found—the schema is a light one and imposes no limitations. When we are ready to conduct an analysis, we use the tags to choose any or all of the data and ask whatever questions we wish. We do not need to guess which portion of the data we might need, unlike when using the conventional approach. The Data Lake scales to an organization’s growing repository of data and makes it all available for any inquiry.

There are no rigid data structures with schema-on-read, so there is also no need for an extensive data preparation process. The Data Lake easily accepts all types of data including unstructured, which is particularly difficult to prepare using the conventional approach. Information is also far more secure since the Data Lake controls the visibility of the data to a granular level.

INSIDE THE NEW TECHNOLOGY

The shift from schema-on-write to schema-on-read arises from the confluence of three major factors: two highly advanced software programs (Accumulo and Hadoop) and the extraordinary power of the cloud. Accumulo provides the Data Lake’s key/value store, while Hadoop makes it possible to process the data on thousands of networked computers at once in a cloud environment. While Accumulo and Hadoop are publicly available and each is currently used independently in a number of experimental applications, the Data Lake is the only technology that brings them together to create the revolutionary schema-on-read.

This powerful new technology applies the flexibility of schema-on-read to an almost unlimited amount of data, made possible by the third major factor, cloud computing. Conventional techniques simply use the cloud to make the rigid structures fatter, but the Data Lake uses cloud computing to its full potential, opening up the vast and diverse data that defines our age.

The Data Lake as a disruptive technology is steadily displacing the conventional schema-
on-write approach across business and government. Organizations are discovering an immediate value: collecting, managing, and analyzing large amounts of data is no longer an expensive proposition. Schema-on-read enables high-powered analytics, providing business with opportunities for top-line growth and helping government agencies achieve their mission goals in new and better ways.

FOR MORE INFORMATION

Michael Delurey, EngD
Delurey_Mike@bah.com
703-902-6858

Steven Escaravage
Escaravage_Steven@bah.com
202-346-9516