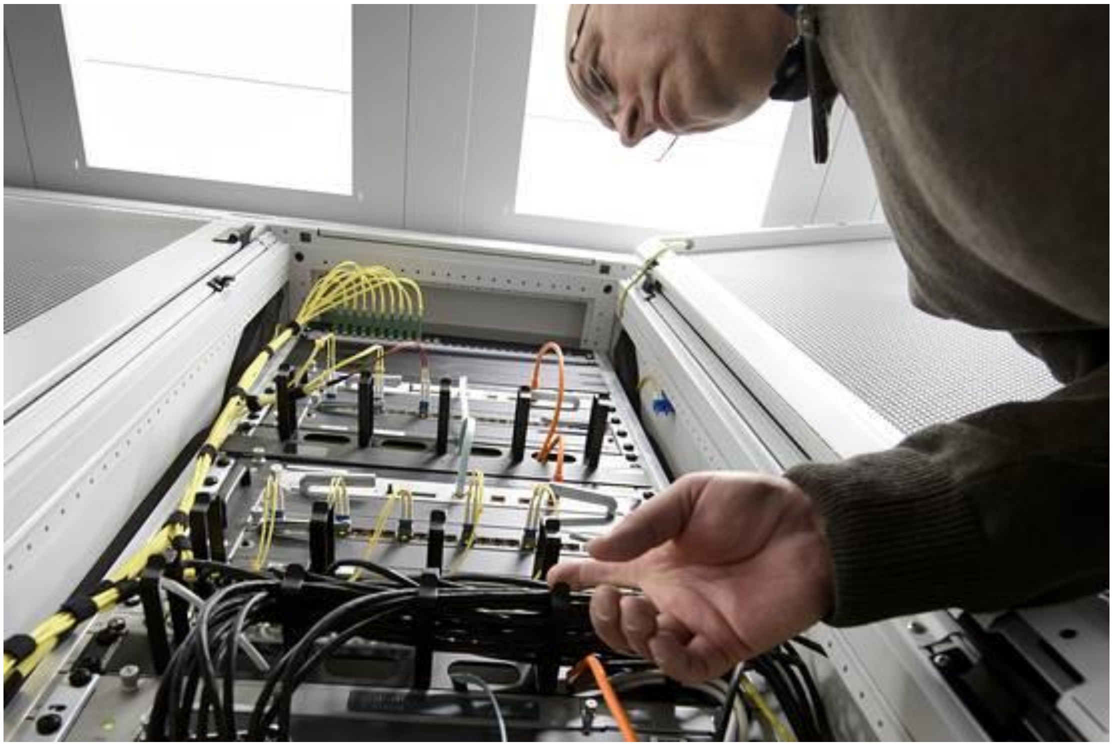
THE WALL STREET JOURNAL.
WSJ.com

June 3, 2015, 6:18 PM ET

The Shift to a New Data Architecture

By Thomas H. Davenport



Computer servers at the Deltalis Swiss Mountain Data Center in Switzerland
Fabrice Coffrini/Agence France-Presse/Getty Images

I am not the first to say this—various IT vendors have made similar pronouncements—but I am confident that we are moving to a new data and technology architecture. This will be profound for everyone in the industry—vendors, their customers, IT people, data analysts, and business users. Of course, like every major architectural transition, it won't happen overnight, and many companies will put it off as long as possible. And it won't be entirely new, but rather a hybrid of existing and new technologies.

That's why I am neither surprised nor depressed by Gartner's recent pronouncement that the adoption of Hadoop—the poster child for this new architecture, but certainly not all of it—is "steady but slow." It's hard to imagine that it would be otherwise. I thought the fact that 26% of companies in Gartner's survey sample already had Hadoop projects underway, and 46% intended to have them in place within a couple of years, was a pretty fast transformation.

This new architecture involves not only Hadoop, but an entire series of new technologies. What they have in common is that many are open source, accommodate a wide variety of data types and structures, run on commodity hardware, are somewhat challenging to manage, and are amazingly cheap to get started on proofs-of-concept. They don't have all the features of traditional IT architectures, but they are gaining the necessary ones rapidly. Some organizations call them "data lakes," others refer to them by the technologies they include, such as Hadoop. In short, they are a classic disruptive innovation in the Clay Christensen sense. However, it's not disruptive enough to send existing architectures to the dustbin; the new one is just added to the mix.

I've spoken with a variety of companies over the past few weeks on this issue. Many of the conversations were facilitated by Informatica, for whom I spoke at their recent "Big Data Ready Summit." Informatica and other proprietary software vendors obviously have a big stake in this transition. They've primarily been in the world of the previous architecture, which involved relational databases, enterprise data warehouses, and ETL (extract, transform, and load) software to get the data in. But Informatica and other proprietary software vendors are producing new offerings to extend and leverage the existing architecture and insulate business users from the complexity of the new one. As Mike Vaughan, Senior Vice President of Data Strategy and Architecture at BB&T Corporation, a super regional bank, put it:

We are evolving our data environments to deal with increasing volumes and sources of structured and unstructured data and provide additional analytic capabilities to the business. There is still a role for commercial products; traditional warehouse and ETL tools still have a place. They serve as a great foundation and springboard into the new environment. Where we have made investments, we will leverage them.

Mr. Vaughan and the other data executives I interviewed are investing considerable energy and resources into these new data environments. The details of them vary a bit, but it's clear that there will be multiple places to store and analyze data in the future, rather than just one. Data may first be stored in a data lake so that it can be explored, cleaned, and prepared. If it can be structured in a relational format (basically rows and columns) and needs to be used frequently and kept highly secure, it may go into a data warehouse. If it stops being used frequently, it may go back to a HDFS (Hadoop Distributed File System)-based archive.

This movement of data places a premium on the ability to keep track of data at all times. The companies doing this work, then, are almost all focused on creating better metadata and data lineage. Those firms that are banks are being strongly encouraged to do so by regulators, but they all think it makes sense.

The new architecture will also rise or fall on its ability to exploit the existing skills of employees. If people have to learn Pig, Hive, Python and the like to use data in these new environments, the economic and competitive benefits will be limited. At the Dutch telecom KPN, the company is using data platforms and tools to enable the use of SQL-based queries. Thomas Reichel, the company's Lead Data Architect, told me:

We are using a SQL abstraction layer, and it's working very well to preserve our data analysts' investment in SQL skills. It took only a couple of days to get them going, and we were able to develop new applications worth €180,000 in two weeks. Our goal is to make data science a universal skill in the company without requiring the exotic data science skills.

Many of the early adopters of these new architectures are determined to prevent business users and analysts from having to know the details of where data is residing at any moment, and how it can be accessed and analyzed. A senior technologist at a large bank (not BB&T) said:

We are trying to create a layer of abstraction by which you can invoke a variety of open-source tools and languages without knowing what is actually doing the work behind the scenes. We hope to make this approach just part of the analytic process—you won't have to know R [an open-source statistics language] to do analytics, you won't have to know SQL to do queries. We don't even frame it as a transition because people really love their existing tools. We will keep those around in the background, but we hope not to use them as much in the future.

So this architectural revolution won't be televised, but it will be revolutionary. It will bring the power of open-source tools and big data to large, established firms, but none of the existing capabilities will go away. It will combine the cost and processing power advantages of Hadoop and open-source tools with the ease of access, governance and security of data management and data warehousing. Just as many organizations still have mainframes in their data centers, the new architecture won't involve throwing much away. But it will bring in some amazing new capabilities.

Thomas H. Davenport is a Distinguished Professor at Babson College, a Research Fellow at the Center for Digital Business, Director of Research at the International Institute for Analytics, and a Senior Advisor to Deloitte Analytics.