Rethinking Hadoop for Modern Analytics

The Joy of ASAP—Analytics by a Single Access Point

December 2019

ThoughtPoint 4 of a 5-part Series by
Dr. Barry Devlin, 9sight Consulting
barry@9sight.com

Your business may have a multiSIDEd problem—an explosion of analytical silos driven by Hadoop and other technologies. You need a single access point solution, ASAP.

Are there too many SIDEs to your business? Please excuse the new acronym, but I need a shorthand to talk about a very challenging multiSIDEd disease that has been spreading rapidly in digitally transforming business over the past few years. It's not a new disease, having existed since the early days of computing. But the growth of Hadoop has provided a fertile ground for its widespread proliferation.

A SIDE is a *standalone insight delivery ecosystem*, often called an analytical silo. Insight delivery is, of course, what a business needs to support its decision-making and action-taking processes. It begins with data discovery and collection, generation of useful information, leads to a set of tools and applications enabling businesspeople, analysts and data scientists to explore and analyze the information, and ends with the human and organizational processes to socialize decisions and ensure action. In short, an ecosystem of interdependent people, processes, information, and applications.

Business decision making is supported on all SIDEs by a surfeit of analytical silos: standalone insight delivery ecosystems.

A SIDE is not necessarily a bad thing. Within its original scope, it can deliver results quickly, with a high degree of agility as business needs change, and often offers its users a common language and context for collaboration, decisions and action. However, one person's SIDE is another's silo. The danger is that almost every such ecosystem emerges and develops in a standalone manner—among a bounded subset of people in the business, driven by specific goals and processes, based on readily available information, and built on particular tools and technology. When SIDEs proliferate, especially in the case of a Hadoop-based data lake, a chaotic multiSIDEd environment emerges with inconsistent information, misaligned insights, and conflicting decisions across the organization, raising serious governance issues for IT to tackle.

Chaotic multiSIDEd decision support environments have emerged as Hadoop data lakes have proliferated.

So, are there too many SIDEs to your business? If you are like most medium and large enterprises, the answer is a resounding "yes." Count the number of data warehouses, data marts, data lakes, business intelligence (BI) tools, analytics and artificial intelligence (AI) systems you have. I'll guess a dozen or more. And add all the spreadsheet-based systems in use. Not every example may be a SIDE, but if it is siloed in terms of users, data, or tools, it probably is. The more analytical silos you find, the higher the risk of chaos.

Sponsored by teradata.
www.teradata.com

On the wrong SIDE of history

The plague of SIDEs is a historical fact. Why has it spread so widely? The individual causes are not too difficult to understand, but their complex interactions have made this multiSIDEd problem difficult to eradicate. These causes include:

1. *Businesspeople want instant satisfaction:* Businesspeople have always wanted to just "get things done," but with digital transformation the pressure for immediacy is immense. The fastest way to instant answers is to commission a standalone solution for your specific case. Arguing against instant satisfaction is a losing strategy.

2. *There's comfort in the familiar:* Having a solution that you understand and works for your needs leads immediately to trying to expand it when new needs arise. SIDEs are thus very sticky; their users seldom want to move to another solution.

3. *Technology just won't stand still:* New tools are certainly good news for addressing long-standing or intractable problems or finding new opportunities. Unfortunately, they usually come with specific prerequisites or unique ecosystems. The biggest and most important offender is the extended Hadoop ecosystem (as discussed in previous articles in this series) which has driven an explosion of SIDEs in yet another set of disparate data stores and computing environments.

Businesspeople want it all and want it now. And they are very comfortable when they get it. Don't mess with their SIDEs!

4. *Consistency with agility is a big ask:* IT has long tried to drive cross-enterprise data consistency with efforts such as data warehousing to combat multiSIDEd proliferation. It's a laudable goal. But such projects tend to be slow to deliver and even slower to change with the business. Trying to fix the problem with yet another data warehouse or data lake exacerbates the problem.

5. *Existing data architectures are largely monolithic:* The traditional data warehouse architecture was (and is) a powerful concept. At its inception in the mid-1980s, the only technology capable of supporting its aim for data consistency combined with ease of access was relational database (RDB) technology. There's more to IT life today than RDBs, although they will play a key role in solving the multiSIDEd challenge.

Hadoop has driven an enormous growth in the number and variety of analytic solutions, leading to a multiSIDEd problem for governance.

In a 2018 global Teradata survey[1], nearly three quarters of respondents with analytics systems said that analytics environment complexity is a problem. Multiple studies, anecdotes, and personal experience confirm the growing challenge of analytical silos. As analytics and artificial intelligence opportunities have proliferated, a plethora of SIDEs have been developed, most commonly via the extended Hadoop ecosystem. At a recent count, I found over thirty different data storage environments, twenty-plus access methods, and more than fifteen streaming systems in the Hadoop project space. Together, they enable the development of an almost innumerable variety of analytical silos even within a single business function, never mind across the organization as a whole.

Today, there is a solution. Let's call it ASAP—Analytics by a Single Access Point—and most organizations do indeed need it ASAP. To make the acronym work, I'm using *analytics* here in the broadest sense to cover all types of BI, AI, spreadsheets, etc. Let's explore the solution now.

ASAP—Analytics by a Single Access Point can solve the multiSIDEd problem.

The future is ASAP

I have already hinted at the core of ASAP: the relational database. In ThoughtPoint 2 of this series, *"Relational is the New Black—Uniting Data and Context*[2]*,"* I discussed the concept of the extended relational environment, exemplified by Teradata Vantage^{TM}. The extensions comprise the technology needed to better support the volume, velocity and variety of externally sourced data, storage and processing for non-relational data formats, direct access to data stored locally and in remote, distributed data stores, and separation of compute and storage independently on premises and/or in the cloud.

However, the key to ASAP lies in one further feature of the system: the support for direct access to all data stores via SQL, R, Python and more. By embedding R and Python support, as well as extensive analytics functions in SQL, businesspeople, analysts, data scientists, and others can continue to use the languages they already know and love. SQL is the most common language for BI, while R and Python are the most popular analytics environments. The "magic happens" in the vertical bar between data stores and analytic engines in Figure 1.

The secret to ASAP is to retain as much of the user-facing aspects of existing SIDEs as possible.

To the right, analytic engines, languages, and tools represent the paths by which information is made available to businesspeople, analysts and data scientists. These are the user facing components of the pervasive SIDEs. These are the components that keep their users coming back for more data, more insights, more functionality. These are the sticky components of SIDEs, and anybody who wants to tackle the multiSIDEd challenge must recognize that, in general, users will cling to them like drowning men to a life raft.

To the left lie all the various data stores with their various strengths and weaknesses. There will be times when moving data from one to another may make sense or even be possible. However, the sizes and skill investments in the different stores will make such migrations a costly exercise, to be taken only when absolutely necessary. As a result, we should assume that there will always exist a variety, and even a changing variety, of data stores. The magic needed to solve the multiSIDEd challenge ***must*** occur in the vertical "translation" bar between these two sides.

Figure 1: Teradata Vantage overview

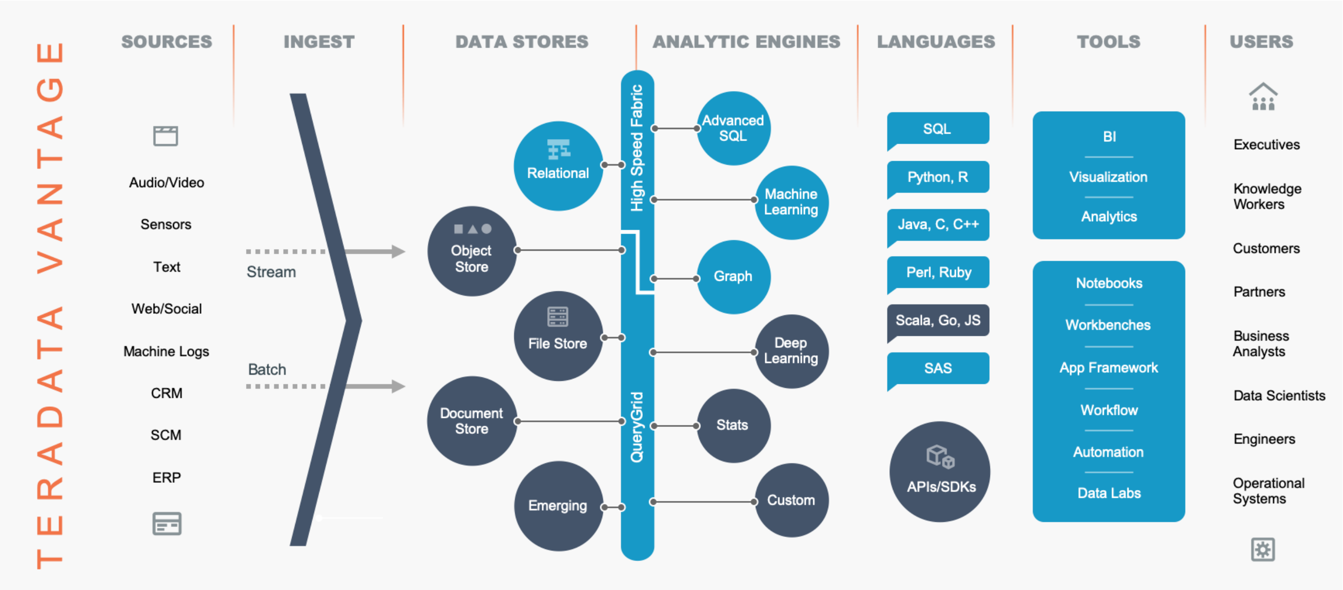

Figure 1 identifies two tools—High Speed Fabric and Query Grid—that compose this bar. However, their identities are less important, representing a point-in-time view of a moving technical implementation of a set of function often called ***data virtualization***. Sadly, that designation carries a lot of marketing baggage. In *Business unintelligence*[3], I called it ***reification***—defined as the process of making something abstract real. *Abstract* here is the business understanding of information and insights, as well as the applications and tools that represent them in ways that users prefer. *Real* is the actual data as stored in its various forms and stores shown. Reification translates one to the other, in both directions. A request for information in whatever language or tool preferred by the user is translated into the required languages of the underlying stores and routed to them in real time. The data results are combined and returned in the user's preferred language and tool.

Reification (or data virtualization) offers a translation layer between multiple user-preferred tools and multiple IT-preferred data stores.

At one time, this might have been seen as magic. The technology is available today to perform the task reliably and with the required performance as a key part of Teradata Vantage. This is the function that solves the multiSIDEd problem. This is what I mean by Analytics by a Single Access Point, ASAP.

The parable of the new broom

The Hadoop-based data lake debacle at the mythical truck company, Trucoeur, cost my imaginary friend, Celine Dejavu, her job as COO. It seems it was partly my fault. Soon after the publication of her story ThoughtPoint 3, *"AI and Analytics—All Gold Taps but No Plumbing*[4]*,"* she was replaced by Jacques Noveauhomme, who—as his name suggests—was determined to be the new broom that sweeps clean.

Technically, the cleanup of the truck maintenance scheduling application—a true SIDE—required dealing with the plumbing that collected the data required from its multiple sources and ensured its quality and consistency. Some of the key data stores were migrated from Hadoop to Teradata and a robust delivery system was implemented behind the database and remaining Hadoop stores. A plan is in place to move some of the data lake storage to a cloud-based object store. This is a complex migration, but it turned out to be the easy part.

The users of the scheduling application were not impressed with the first version of the plan which would have required them to rewrite their application in a new language and understand how to access and use data from multiple and changing locations. A new plan was quickly written based on the reification function in Teradata Vantage, allowing users to continue with their original application with minimal rewrites.

Looking forward, Trucoeur—like many digital businesses—is planning a range of AI-based applications that will involve new tools and novel data stores. Some will be brand new; others will be a reworking of existing Hadoop-based applications. New opportunities will certainly emerge as businesspeople and data scientists take advantage to emerging data sources. All will be driven by often urgent business needs and each will risk delivering yet another analytical silo, yet another SIDE. However, ensuring data quality and consistency will remain a function of the extended relational environment now placed firmly at the core of the data management environment. And the concept of analytics by a single access point will be placed at the core of their thinking ASAP. New broom sweeping clean.

Businesspeople will continue to have urgent needs, leading to demand for new SIDEs. ASAP thinking and Teradata Vantage can prevent the emergence of new chaos.

This is the fourth article in a series of five ThoughtPoints on "Rethinking Hadoop for Modern Analytics." The complete series of articles is:

1. *Hadoop—Spreadsheets on Steroids http://bit.ly/2N59ZCO*
2. *Relational is the New Black—Uniting Data and Context http://bit.ly/2CSpV6t*
3. *AI and Analytics—All Gold Taps but No Plumbing http://bit.ly/2DCKXqe*
4. *The Joy of ASAP—Analytics by a Single Access Point http://bit.ly/2S2vjga*
5. *The Right Vantage Point Offers Advanced SQL Views http://bit.ly/2TZ1Epr*

An omnibus edition of all five articles is also available at http://bit.ly/36lWy95

*Dr. Barry Devlin is among the foremost authorities on business insight and one of the founders of data warehousing, having published the first architectural paper on the topic in 1988. With over 30 years of IT experience, including 20 years with IBM as a Distinguished Engineer, he is a widely respected analyst, consultant, lecturer and author of the seminal book, "Data Warehouse—from Architecture to Implementation" and numerous White Papers. His book, **"Business unIntelligence—Insight and Innovation Beyond Analytics and Big Data"** was published in October 2013.*

Barry is founder and principal of 9sight Consulting. He specializes in the human, organizational and technological implications of deep business insight solutions combining all aspects of internally and externally sourced information, analytics, and artificial intelligence. A regular contributor to Twitter (@BarryDevlin), TDWI Upside, and more, Barry is based in Bristol, UK, and operates worldwide.

[1] Teradata Press Release, "Global Survey: Analytic Insights Remain Trapped in Complexity and Bottlenecks", October 2018, https://www.teradata.co.uk/Press-Releases/2018/Global-Survey-Analytic-Insights-Remain-Trapp

[2] Barry Devlin, "Relational is the New Black—Uniting Data and Context", November 2019, http://bit.ly/2CSpV6t

[3] Barry Devlin, *"Business unIntelligence"*, 2013, Technics Publications, New Jersey, http://bit.ly/BunI-TP2

[4] Barry Devlin, "AI and Analytics—All Gold Taps but No Plumbing", November 2019, http://bit.ly/2DCKXqe