9sight Consulting

RETHINKING HADOOP FOR MODERN ANALYTICS

Hadoop—Spreadsheets on Steroids

OCTOBER 2019

THOUGHTPOINT 1 OF A 5-PART SERIES BY
DR. BARRY DEVLIN, 9SIGHT CONSULTING
BARRY@9SIGHT.COM

Hadoop offers powerful, valuable analytical tools to business and data scientists, but its negative impacts on data governance and systems management must be mitigated.

Once upon a time, Harvard Business School students Dan Bricklin and Bob Frankston created VisiCalc[1]. It was 1978 and spreadsheets soon became the "killer app" of the PC revolution. Spreadsheets dramatically empowered businesspeople, removing the drudgery of paper, pencils, erasers and calculators. They unleashed a huge wave of innovation in the use of data in business—from planning to auditing and beyond.

Subsequently, a huge wave of frustration swept over IT departments trying to manage and curate the business data released through data warehousing and BI projects. As Wayne Eckerson lamented[2]: *"Spreadsheets run amok in most organizations. They proliferate like poisonous vines, slowly strangling organizations by depriving them of a single consistent set of information and metrics..."* Spreadsheets enable base data to be changed, derived data to be miscalculated, and inconsistent results to be distributed widely—all without due data governance—in a fully decentralized and distributed computing environment.

Three decades after VisiCalc's debut, Doug Cutting's yellow elephant was anointed an Apache top-level project after a few years in gestation. By 2008, Hadoop was starting to do for data analysts (later renamed data scientists) what spreadsheets had done for businesspeople. It released a surge of innovation, this time in the analysis of "big data." And for professionals in data management and governance, it posed a greater challenge than spreadsheets. Hence my adage: Hadoop resembles spreadsheets on steroids.

> Hadoop poses a greater challenge to data management than spreadsheets ever did.

The dangers to data quality of spreadsheets are now widely understood (although still poorly addressed). Meanwhile, the strengths and weaknesses of Hadoop in enterprise computing are little discussed. Factors include the widely accepted but ill-defined concept of the data lake, introduced in 2010 by James Dixon[3], and the "Cambrian explosion of [Hadoop-]related projects" as Doug Cutting described it in a 2015 article[4].

This series of ThoughtPoints explores Hadoop's strengths and weaknesses, and what we should do about them as we enter the third decade of the 21st century. But first, what is Hadoop today and how is it used?

SPONSORED BY

teradata.

www.teradata.com

The blind men and the elephant

Hadoop always seems to evoke elephants. The parable says that depending on which part of the elephant you touch, you come to a different conclusion about what it is. Today, Hadoop is not just an elephant, but a whole menagerie of inter-related but largely independent projects named for exotic beasts—Pig and Giraph, Impala and Kudu—and clever memes, such as Zookeeper and Hive, Cassandra and Goblin.

In its earliest incarnation, more than a decade ago, Hadoop consisted of HDFS, MapReduce, and some system management software, all developed in the open source paradigm and delivered in a few tightly linked projects. Today, when we say *Hadoop*, we are referring to an assortment of more than eighty separate projects. In reality, this is a complex, extended, and deeply interdependent but independently developed ecosystem of mostly open source software to collect, prepare, process and deliver data for analytical purposes. In this series, therefore, *Hadoop* refers to this extended ecosystem.

> Hadoop is an extended, heavily interdependent ecosystem of data-manipulation software.

Having defined what Hadoop is, we must now discuss how to implement and use it. How do you eat an elephant? In very small chunks. But keeping the whole beast in mind!

The good, the bad, and the downright ugly

Hadoop's inherent goodness

Since its birth, Hadoop has enabled and driven the growth of an analytics environment, particularly of big data, that would otherwise have been prohibitively expensive or, in some cases, impossible in traditional data management settings. By defining a parallel processing environment on distributed, low-cost, commodity hardware, the Hadoop ecosystem's original designers—owners, such as Google and Yahoo, of then burgeoning big data systems—created a new, powerful set of open source intellectual property.

The data lake philosophy of allowing any type or structure of data to be stored at a user's sole discretion, combined with a mindset of enabling a wide variety of tools and analytic approaches has led Hadoop to become the destination of choice for data scientists and analytics / machine learning experts. Such freedom of choice and avoidance of pre-planning or permission-seeking from IT are especially appealing for those involved in free-flowing research into data patterns and what they might offer the business.

> For data scientists, Hadoop data lakes promise liberty without limits to play with big data.

Using full data sets rather than being limited to sampling, data scientists initially found in Hadoop a cost- and time-effective solution to the expanding set of needs and opportunities offered by social media, clickstream, Internet of Things data, and more. The environment also supports the repeated and iterative analysis required by data scientists.

Furthermore, as the ecosystem has evolved through open source development, the infrastructure has matured into a full-function, parallel processing environment (with version 2 in 2013), adding streaming techniques, and most recently support for cloud-like object storage. Application functions, such as data mining, machine learning, and artificial intelligence, have also been made available—often first—in the Hadoop environment, providing data scientists with leading-edge solutions to their demanding needs.

Small wonder that businesses have come to see Hadoop-based data lakes as the best thing since sliced and diced spreadsheets, believing they offered innovative analytics and timely solutions at reasonable costs. Sadly, however, that turned out not to be the case.

Innovation in business analytics has been spurred by Hadoop's speed of evolution.

When good solutions go bad

There is little argument that the open source development approach offers one of the fastest and move innovative way of delivering new function. In the rapidly emerging and evolving analytics environment of the past decade and more, such speed, flexibility and innovation have been highly valued characteristics. However, open source creates its own problems, especially in a market as diverse and complex as analytics.

The first drawback is in the sheer number of projects that Hadoop has spawned, either directly or indirectly. Providing a coherent roadmap to this environment is near impossible. Identifying which projects offer which function, overlaps or gaps in functionality, or inter-dependencies or conflicts between them is challenging as new Hadoop projects are kicked off frequently. Add to that the difficulty in figuring out which projects have lost momentum with their often-voluntary development teams and what to do if the function for which you chose a particular project cannot be found elsewhere. The innovation that was so desirable in the early stages of market evolution can become less attractive as the market matures. Systems management in these circumstances is deeply challenging.

The Hadoop ecosystem is an unmanageable jungle of symbiotic projects that are difficult to profitably commercialize.

More recently, a second problem has emerged. The companies that launched Hadoop distributions (or distros)—designed in part to tackle the above systems management problems—have struggled to create a profitable business model around largely "free" software. The emergence of cloud solutions has also impacted the Hadoop market. The recent withdrawals, collapses, sell-offs and consolidations of many of the largest players has shaken confidence in the Hadoop ecosystem and suggests that some contraction in the number and variety of projects may be imminent.

Businesses that bought into the innovative promises of Hadoop, expecting to benefit from new tools, such as Graph analytics or machine learning, met another challenge. Moving the insights to production often involved a return to relational techniques that were often poorly supported in the relational tools available in Hadoop. Achieving their business goals turned out to be more difficult than anticipated.

Ugly is as ugly does

Data governance and management are often painted as the ugly stepsisters of business progress. Correct and accurate results matter, as do the actions and processes needed to achieve them. But their dependence on high quality data in decision making is poorly appreciated by business. Making the case for investment in such quality data has too often and incorrectly been left to IT, the same department that is frequently blamed for standing in the way of business action. The business success of spreadsheets and their serious impact on data management have contributed to the ugliness of the business-IT gap.

Hadoop has further widened this traditional gap while simultaneously obscuring the necessary collaboration between business and IT roles in data governance. On one hand, Hadoop has led to the creation of enormous data lakes, often with minimal IT involvement, with their subsequent and rapid degradation to data swamps and failed projects[5].

Hadoop has widened and deepened the business-IT gap in data management and governance.

On the other hand, data scientists need significantly better data skills than spreadsheet users but focus more on data manipulation than data management. Data management can, of course, be applied retrospectively to data lakes, but it is often too little, too late.

It is in the impact on data management and governance that the idea of "spreadsheets on steroids" applies most strongly. Hadoop offers a set of tools with high business value and expectations, in a highly distributed environment for multiple users, with little oversight or control of data quality. More worrying, these users are more technically skilled—and thus potentially more capable of impacting data quality as they accumulate enormous quantities of data from often poorly described, external sources with little coordination. In the worst cases, this can lead to different departments buying the same data multiple times and using it in different ways to prove competing propositions.

With Hadoop, data scientists may acquire programming skills but get limited support for good data management.

Business beyond steroids

My comparison of Hadoop to spreadsheets on steroids dates back many years, but the metaphor has become increasingly appropriate as the scope and importance of the extended Hadoop ecosystem has since expanded. The data governance and systems management challenges encountered are stronger than ever. However, recent developments in the data management landscape offer some hope that these issues can and should be addressed now.

The enormous increase in popularity and power of cloud offerings, with their much vaunted and valued elasticity in both data storage and processing power, as well as their outsourcing of systems management, has led to a new questioning of the appropriateness of an on-premises implementation strategy for big data. The fact that the cloud is the main source for much of the data Hadoop handles adds further weight to the argument. Hadoop's cost advantage versus traditional data processing and storage solutions has been turned against it by the cloud vendors. As a result, some analysts are predicting the imminent demise[6] of Hadoop. Although I believe this analysis to be over-simplistic[7], it seems likely that we may have reached peak-Hadoop as evidenced by recent significant changes in the Hadoop vendor space.

A more important consideration, because of its implications for data governance and systems management, is the ongoing evolution of traditional relational database environments, such as Teradata Vantage™, to include additional function and support access to data and function beyond their classical boundaries. The relational paradigm, combined with the data modelling approaches that sprang from it, remains the best environment from which to monitor and manage data quality. Furthermore, with four decades of focus on reliability, availability and serviceability, relational databases offer the most stable foundation for core business data and its relationships to newer data classes and sources.

Hadoop has been pressed on one side by the growth of the cloud and on the other by a renaissance in relational databases.

These thoughts suggest three simultaneous directions of evolution for Hadoop use:

1. *Rebuild in the cloud:* Where cost and elasticity are primary drivers, components such as low cost object storage (Amazon S3 and Azure Blob) are attractive and the cloud will likely become the implementation of choice for analytics that use large and variable resources in largely standalone applications.

   There are a number of cloud offerings from major providers that allow companies to build a data warehouse/ data lake environment within the confines of one chosen cloud platform. Where business needs can be satisfied within this environment, the rebuild will need to ensure that the data governance and systems management challenges listed above are adequately addressed. As relatively recent database developments, the breadth of SQL support and the depth of reliability, serviceability and data governance functions may be limited with these newer cloud-only solutions. In addition, hybrid use cases—both data and processing—can prove difficult.

2. *Hang on with Hadoop, on-premises and into the cloud :* Companies that have invested heavily in highly specialized Hadoop applications and have the technical skills to maintain them may well stick with Hadoop as a valid, justifiable technology.

   This approach protects existing investments in Hadoop infrastructure and skills, both on-premises and in the move to the cloud clearly emerging among Hadoop vendors. However, it preserves existing systems management complexity and extends it to the cloud. Data management challenges and costs are exacerbated as data must now be managed across both environments. With added complexity, the opportunity to repeat previous mistakes should not be underestimated.

3. *Rediscover relational:* In cases where data quality and integrated operational analytics are vital, or where technical and systems management skills are more limited, migration of existing or planned Hadoop applications to a modern relational-centric environment will be the solution of choice.

   Modern, advanced relational environments, such as Teradata Vantage, have evolved in recent years from traditional products with well-established reliability, availability and scalability (RAS) characteristics and proven systems management capabilities. They have been extended in scope to handle additional data types and analytical function. Furthermore, they provide direct access to data in other stores, including cloud object stores, such as Amazon S3 and Microsoft Azure Blob storage.

   Teradata Vantage offers advanced relational and analytic features, as well as offering direct access to data on other platforms, including object stores.

   In addition to offering mature and robust analytical technology and connectivity across a hybrid on-premises/multi-cloud environment, this approach builds on the strong data governance and management, data integration, lower development costs, and workload flexibility of a mature and comprehensive advanced relational environment. While some existing workloads or data types are not yet supported, direct access to most Hadoop environments is possible.

Data quality and integration issues loom large in digital transformation projects. Data from multiple sources, both internal and external, including many of dubious quality and consistency, is central to digital business. When such data is used in decision making, assuring its governance and management is essential, especially in areas of high business impact or where ethical implications may exist. Migration of such data and projects—existing or planned—to a relational-centric environment is a vital step in addressing these issues. Option three above is therefore the approach of choice for the majority of companies struggling with on-premises Hadoop data lakes.

Digital business demands an intense focus on data quality and consistency to which the relational model is key.

The old data management adage “garbage in, garbage out” has become so important that it has entered the popular lexicon. Data governance and management experts in today’s digital-first business world need a phrase that reflects the speed of decision making and the extensive implications of getting it wrong. Perhaps “fresh in, filth out” might work.

This is the first 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.

Brand and product names mentioned in this paper are trademarks or registered trademarks of Teradata and other companies.

[1] Dan Bricklen, *"Software Arts and VisiCalc History"*, 2009, http://www.bricklin.com/history/sai.htm

[2] Wayne Eckerson, *"Taming spreadsheet jockeys"*, ADTMag, September 2002, https://adtmag.com/articles/2002/09/01/taming-spreadsheet-jockeys.aspx

[3] James Dixon, *"Pentaho, Hadoop, and Data Lakes"*, October 2010, https://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and-data-lakes

[4] Matt Asay, *"Beyond Hadoop: The streaming future of big data"*, http://www.infoworld.com/article/2900504/big-data/beyond-hadoop-streaming-future-of-big-data.html

[5] Kayla Matthews, *"The difference between a data swamp and a data lake? 5 signs"*, April 2019, https://www.information-age.com/data-swamp-data-lake-123481597/

[6] Andrew Brust, *"Cloudera and Hortonworks' merger closes; quo vadis Big Data?"*, January 2019, https://www.zdnet.com/article/cloudera-and-hortonworks-merger-closes-quo-vadis-big-data/

[7] Barry Devlin, *"The Death of Hadoop?"*, February 2019, http://bit.ly/2YbPdGb