The Right Vantage Point Offers Advanced SQL Views

JANUARY 2020

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

With Hadoop suffering from a quintet of mid-life crises, the start of a new decade is an appropriate time to take a different Vantage point that places Advanced SQL at the heart of digital transformation efforts.

Hadoop is, according to some industry observers, dying or already dead. I disagree. However, it is—at best—experiencing a serious mid-life crisis. Or, to switch metaphors in midstream, the Hadoop train is no longer an express, is on a track to an unknown destination, and running out of steam. Those already aboard must choose whether to continue the journey or change trains at the next station.

The Multiple Crises of Being Hadoop

Throughout this ThoughtPoint series, *"Rethinking Hadoop for Modern Analytics,"* I have explored the challenges faced now by Hadoop at the end of its first decade.

Hadoop is a complex, deeply interdependent but independently developed ecosystem of mostly open source software addressing data use for analytics.

First is a ***crisis of identity***. What exactly is Hadoop? I've coined the phrase *extended Hadoop ecosystem* to reflect the reality that an originally tightly bounded set of a few open-source, data-centric components has grown willy-nilly to more than eighty separate projects, 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. It's difficult to market something so amorphous and ever-changing.

Second is a ***crisis of confidence***. Industry analysts have begun to discount Hadoop. Its last appearance in Gartner's Data Management Hype Cycle[1] was in 2017, deep in the trough of disillusionment, labelled "obsolete before plateau." Vendors of Hadoop "Distros" have gone broke, pulled out, or merged even as they slipped Hadoop down several levels in their marketing blurb to focus instead on platforms or broader infrastructure plays.

The third is a ***crisis of deployment***. Systems management of the extended Hadoop ecosystem is notorious in its difficulty. With so many projects to choose from, looking for a specific function can be challenging. Even more so is knowing how it integrates with other Hadoop projects as they evolve, and whether it will continue to be supported and grown. Even discovering if or when development has been abandoned is a challenge.

SPONSORED BY
teradata.
www.teradata.com

Fourth is a ***crisis of cloudiness***. The significant, ongoing growth in cloud implementations of big data projects, combined with the fact that most such data is externally sourced, has also contributed to Hadoop's problems. Native feature-as-a-service and serverless approaches, as well as object stores have become more attractive than Hadoop offerings.

Finally, and most importantly, Hadoop has engendered a ***crisis of data governance***. This crisis should not be blamed entirely on Hadoop. The seeds were sown with the rise of spreadsheets, when IT lost control of consistency and quality of data dispersed throughout the organization. Data governance demands a set of beliefs and skills seldom found among spreadsheet users or, indeed, Hadoop developers or data scientists, due to their diverse backgrounds. And in thrall to quick wins promised by Hadoop-fueled analytics, and often failing to link project failures to data quality issues, business has too often ignored or underfunded necessary data quality efforts.

Hadoop is experiencing five crises of identity, confidence, deployment, cloudiness, and data governance, all of which put it in a vulnerable spot.

"Reports of My Death Are Greatly Exaggerated"

Despite this quintet of crises, Mark Twain's widely misquoted riposte[2] is appropriate. Hadoop will continue to live on in many instances where businesses have made significant investments with real returns or even where there is a reluctance to admit a lack of obvious success. There also continue to be use cases and business or infrastructure needs where Hadoop is the most appropriate answer.

Nonetheless, this dawning decade is the time to reevaluate Hadoop's role and reposition its uses and strengths. At the core of a digital business, data quality is imperative and software governance trumps *ad hoc* innovation. Hadoop's crises confirm that it cannot be at the heart of digital transformation. We need a new Vantage point.

Finding the Right Vantage Point

One of the strengths of the extended Hadoop ecosystem is the speed of innovation that comes from open source software development. Applied in the wrong place, it can also be its biggest weakness.

We have experienced a decade of analytics innovation, driven at least in part by the extended Hadoop ecosystem. Of course, digital business demands that this innovation continue. However, it is built on an increasingly unstable foundation of ill-defined and poorly managed data accumulating in data lakes, also known as data swamps for this very reason. We must therefore put a new and dedicated focus on creating a core of well-governed, quality data that also supports speedy and successful innovation. We need the best of both worlds: well-governed data open to innovative analytical use.

The innovation in analytics demanded by digital transformation needs a firm foundation of quality data and well-governed IT systems.

In most organizations, these worlds are seen as completely antithetical to one another. The conflict is often characterized by the data warehouse / BI community declaring spreadsheets and similar tools an uncontrollable plague, while the business berates the data warehouse for being slow, stultifying and preventing them from running the business as needed. These perceptions and conflicts are based on a false dichotomy.

To address this confusion, in *Business unIntelligence*[3], I characterized two modes of analytics and decision making: center-out and edge-on. The former focuses on the provision and use of well-governed data ***to*** the business while the latter emphasizes innovative analytics and exploration ***by*** the business. Both are required. Their characteristics are:

Characteristic	*Center-out*	*Edge-on*
Data provenance	A correct, centrally controlled "single version of the truth" exists	Multiple and possibly conflicting versions of truth can exist
Data flow	From central store to users	Directly from user to user
Data manipulation by users	Basic data is read-only; users control derived data	Users have full control over all data
Process focus	Reporting and *ad hoc* performance analysis	Creative exploration and innovation
Typical tools	BI reporting and query tools	Spreadsheets and analytic tools
Data quality	Can be closely controlled and managed	Open to rapid degradation
Work approach	Hierarchical and standardized	Emergent prototyping and innovation

A data warehouse is the prime example of the center-out approach, while Hadoop is much closer to edge-on. It should be clear from the contrasting characteristics that neither approach on its own is enough to meet the needs of a digital business. Both are needed, but must be applied in the right places. Specifically, where data governance is a primary concern, a center-out approach is mandatory and data warehousing principles and tools must be applied; where innovation and exploration is sought, an edge-on approach can be applied and Hadoop, spreadsheets and similar tools can be utilized.

Creating a dual-focus center-out and edge-on environment requires a carefully crafted combination of centralized governance of core business information and virtualized access to data in disparate locations through the diverse types of function needed by the business. To the businessperson, this gives the appearance that all data is in a single store and the confidence that it can be accessed through any tool they choose.

A careful combination of centralized core business information governance and seamless access to disparate and distributed data are key to digital business implementation.

As discussed in *"The Joy of ASAP—Analytics by a Single Access Point*[4]*,"* Teradata has been developing such function over several years and their Vantage™ platform provides a firm foundation on which to build such a dual-focus system. At its core, Vantage consists of a full-function, parallel-processing, relational database with strong reliability, scalability, and data integrity features that have evolved over four decades. This is complemented by two key access components—High Speed Fabric and QueryGrid— mediating access to the core relational database as well as an expanding set of other data storage systems, including object stores, file stores, document stores and more.

Getting a Modern SQL View

With the emergence of digital business, the traditional relational database model has had to evolve to support data governance and quality needs far beyond those driven by traditional business operational and informational processes. In this multi-decade evolution, Teradata led the way with its focused support for all aspects of informational work, including parallel processing, columnar storage, high-speed ingestion, specialized analytical functions, and more.

Wth the Vantage platform, Teradata defined the Advanced SQL engine, featuring:

1. *4D Analytics:* integrating when (Time Series and Temporal) and where (Geospatial) analytics on relational data
2. *Support for multiple data types and structures:* from relational data to multi-structured data such as web logs, XML, JSON, and CSV
3. *Hybrid row/column data store:* mix and match rows and columns to create the optimal structure for specific data and query patterns
4. *In-memory technology:* fast access to the most frequently used data and rapid answers to complex questions with Intelligent Memory and In-Memory Optimization
5. *External system access:* orchestration of access to disparate external analytic engines and file systems, so users can focus on business value rather than data integration
6. *Workload management and data resilience:* real time monitoring and management of a mixed workload environment and fallback protection in case of problems or errors

Taken together, these features and more in the pipeline are the basis for building the integrated environment often called a *logical data warehouse*[5] that offers the best of distributed warehouse-based governance, supporting Hadoop features where needed.

In Conclusion…

One aim of this series of ThoughtPoints has been to document where Hadoop—or, more precisely, the extended Hadoop ecosystem—currently stands in the marketplace and in existing implementations. While not as endangered as claimed by some observers, it is clear that we have passed "peak Hadoop" and that the ecosystem is facing challenges on multiple fronts.

The second objective of the series was to explore what options exist for current and future Hadoop customers and what will drive their choices. We observe that data quality and IT governance are becoming ever more challenging and are certain that these challenges can be addressed only by revisiting the foundational platform choices for data collection, storage, processing and use. Given its longstanding history of reliability and integrity, the relational platform, extended with modern features to extend its reach and versatility, is a clear winner for fulfilling the needs of center-out control and governance.

Hadoop will live on in existing, successful implementations as well as remaining a useful environment for exploration and innovation in edge-on analytics. Successful digital transformation will increasingly depend on developing an environment that combines and integrates this approach with center-out governance needs. Teradata Vantage with the Advanced SQL engine provides the ideal foundation for such a combined, well-integrated center-out and edge-on modern analytics environment demanded by a digital business.

Teradata Vantage with Advanced SQL provides the ideal foundation for the combined, well-integrated center-out and edge-on modern analytics environment demanded by a digital business.

This is the fifth and final 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] Help Net Security, *"Gartner reveals the 2017 Hype Cycle for Data Management"*, October 2018, https://www.helpnetsecurity.com/2017/10/02/hype-cycle-data-management/

[2] http://www.thisdayinquotes.com/2010/06/reports-of-my-death-are-greatly.html

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

[4] Barry Devlin, *"The Joy of ASAP—Analytics by a Single Access Point"*, December 2019, http://bit.ly/2S2vjga

[5] Gartner, "Understanding the Logical Data Warehouse: The Emerging Practice", 2012, https://www.gartner.com/en/documents/2057915/understanding-the-logical-data-warehouse-the-emerging-pr

SPONSORED BY

www.teradata.com

teradata.