9sight Consulting

RETHINKING HADOOP FOR MODERN ANALYTICS

AI and Analytics—All Gold Taps but No Plumbing

NOVEMBER 2019

THOUGHTPOINT 3 OF A 5-PART SERIES BY
DR. BARRY DEVLIN, 9SIGHT CONSULTING
BARRY@9SIGHT.COM

Hadoop data lakes have delivered many gold tap applications but the plumbing infrastructure is not fit for purpose. Businesses should consider migrating such plumbing to a future-proof, relational environment, such as Teradata Vantage.

From customer churn to climate crisis, there is no challenge that AI and analytics cannot address. Major analyst firms publish predictions of a trillion-dollar impact AI will have on the world economy in the next decade. The business pages of the much-maligned mainstream media paint pretty pictures of algorithm-enhanced enterprises in the near future. New and exciting gold-standard applications will reduce costs and drive profits for digital businesses. Or so the stories go.

Beyond obviously extravagant claims—and business executives can smell them a mile off—many of these apps can become reality. They, and their BI precursors, are the gold taps of the title, the business wins that inspire big changes and successes in many enterprises. However, the question arises: Will water ever flow from these faucets? That demands some seriously unsexy and equally costly plumbing behind the marble tiles.

New AI- and analytics-based apps are vital for digital business, but the provision of quality data is often overlooked.

Selling (and buying) applications vs. infrastructure

Multiple analogies describe the dilemma. Gold taps vs. pipework, phones vs. network, automobiles vs. highways. The challenge is always the same. What end-users value is what is visible to them, what delivers results. They seldom want to think about the hard work and expensive infrastructure needed to make the applications work.

The parable of the COO who saw the light

Among my imaginary friends at Trucoeur, introduced in my previous ThoughtPoint[1], is the inimitable COO, Celine Dejavu, who courageously admits she fell for the promise of AI although she had, according to herself, "seen it all before". It was she who championed the data lake project to reduce operating costs for Trucoeur and its customers. The business case was indisputable. Advanced analytics and AI tools could predict when certain trucks were likely to break down, based on a combination of operating data that was already available and data about driving conditions and history of usage that could be easily obtained. These gold taps offered significant RoI for Trucoeur and its customers.

SPONSORED BY

teradata.

www.teradata.com

As Dejavu said later, "I was so impressed by the demonstrations of what machine learning applications could predict about component failures. The dashboards and graphs were excellent. I could immediately see how our operations staff could use them to make decisions and, indeed, how maintenance actions could be scheduled automatically. The business case was obvious.

"The vendors freely admitted that the extensive data set used was a mix of real trucking data, enhanced with readily available public data. I knew we had a lot of such data already and were planning to collect even more. The project scope and size seemed well-defined and I immediately made the decision to give the go-ahead."

Vendors of analytics and AI apps may underemphasize the difficulty in obtaining the best source data for high quality predictions.

Dejavu had just been dazzled by her reflection in the gold taps and overlooked what was needed behind them. It has been a common experience for businesspeople since the earliest days of business intelligence (BI) and data warehousing. It's easier to demo the BI application and to impress the business than to discuss data sourcing. In fact, many BI, analytics and, now, AI vendors have long emphasized business value and ease of use in their sales pitches and glossed over the data sourcing with some glib assurances that "our tool connects easily to every common database and file store."

The plumbing is far less shiny, but Dejavu should have focused on what it took to ensure that the data flows freely and cleanly to the users and to enable any cleansing, consolidation and reconciliation required. Or she should have involved a senior data expert who could have posed the right questions and examined the underlying assumptions about data availability, cleanliness and consistency at Trucoeur.

Is gold-plated plumbing the answer?

Few of us, in real life, would consider gold-plating our pipework, especially the majority of it that is hidden behind walls and ceilings. However, when it comes to data governance and management, putting some additional thought into and making some additional investment in our infrastructure is a good use of time and money. Just as highways and water systems have been run down through years of neglect in many nations, numerous enterprises have also cut corners on infrastructure, outsourced it to the cloud, or looked to open source solutions, all in the name of reducing expense.

Prior cost-cutting of data management tools and outsourcing of staff has led to unforeseen dangers in digital transformation.

Not only are we seeing the cumulative impact of years of cost-cutting in data management infrastructure, but the effect is coming at the worst possible moment, with AI and analytics making extraordinary demands on digitally transforming businesses. In the past, the strongest data management focus was reserved for production, administrative and financial data (I call it process-mediated data) used to run and manage the business. However, recent developments in our ability to analyze and act on externally sourced data, such as social media and IoT (Internet of Things) data, have demonstrated that poor governance of such data can lead to serious ethical and societal wrongs, as Cathy O'Neil describes in her excellent "*Weapons of Math Destruction*[2]."

How then should you think about gold-plating our data governance and management infrastructure? Does that imply the wholesale replacement of the existing infrastructure? Or is it an add-on? The answers depend on whether your enterprise is one that has retained much of its enterprise data warehouse (EDW)—a well-structured and maintained data preparation and storage environment—or has moved wholesale to the data lake.

Gold plating the data warehouse

The key component of a gold-plated infrastructure to support modern analytics and AI is a modern EDW. And although it may be stretching the analogy beyond its elastic limit, the gold-plating should be on the inside of the pipes; this is about effect rather than appearance. From its original conception in the mid-1980s, the EDW[3] has been first and foremost about data quality and consistency. The approach was to consolidate data from disparate operational systems into a central store based on relational database (RDB) technology. Although minuscule by today's measures, the volume of data involved was at the limit of what could then be comfortably handled, and although many developers have tried to include *all* data there, this is increasingly impossible as data volumes have grown.

A modern EDW is the key component of a high-quality infrastructure to support analytics and AI.

A modern EDW must be more flexible to handle today's volume, velocity and variety of data but retains a relational core with the ability to store, manage and access data in a distributed, multi-structured environment. Teradata Vantage is a prime example of this approach. Rather than suggesting that all data should be stored in—or even pass through—the EDW, only a subset, called core business information (CBI), belongs there. CBI is central to the very existence of the enterprise and its correctness and consistency is vital to the success of all operational and analytical work.

Teradata Vantage is a prime example of a modern EDW with a relational core and the ability to store, manage and access data in a distributed, multi-structured environment.

So, if your enterprise is one that has retained significant EDW infrastructure, gold-plating it makes a lot of sense. By making it the prime location for CBI and CSI (context-setting information, as described in ThoughtPoint 2 of this series[1]) and extending its reach to access non-relational data stores, the EDW becomes the primary support environment for all data governance and management in pursuit of digital transformation—true gold-plated plumbing.

Can a data lake be gold plated?

If your enterprise is one of those who have abandoned traditional RDBs in favor of a Hadoop-based data lake, gold plating may prove difficult, depending on the level of data governance and management embedded in your data lake. There are two key considerations.

First is how chaotic is the existing data lake storage. If it consists of thousands (or hundreds of thousands) of files, loaded as needed by multiple users, seldom if ever deleted, containing multiple copies or versions of the same data, and so on, gold plating the plumbing will likely be costly and time-consuming. Emerging metadata management / data catalog products for data lakes can offer a layer of limited governance and management on top of this collection of data but fail to address its underlying lack of structure.

Second is the extent to which your data lake contains well-structured relational Hadoop databases. Like most things in Hadoop, there are multiple approaches. Some projects have their own RDBs. Others offer SQL access to HDFS, object stores, or NoSQL stores. Some focus on transactional processing (OLTP) while others specialize in columnar format or even in-memory store (OLAP) use cases. Although good data management and governance function could be developed in such systems, implementation often focuses on specific application function—essentially in support of particular gold taps.

Hadoop-based data lakes require a modern, extended relational EDW core to maintain the data quality and consistency needed for successful AI and analytics apps.

The bottom line is that gold-plating a data lake is seldom recommended. Rather a strategy of establishing (or re-establishing) CBI and CSI in a modern enterprise data warehouse with migration of generic data management function should be pursued.

A Hadoop migration strategy

Businesspeople want and need gold taps. Who can blame them for wanting the best possible applications to report on and analyze data and make AI-based predictions? What they don't need is to worry about the plumbing behind the taps—how well it performs, if it delivers consistent and reliable data, how easy it is to maintain. For the business, these qualities should be a given.

Traditionally, the plumbing consisted of the EDW, including all its data storage, management and preparation infrastructure, based largely on a relational foundation. The data warehousing industry has invested three decades of effort ensuring this infrastructure meets a range of quality, timeliness, consistency, and maintainability needs. In effect, vendors of RDBs, such as Teradata, have been internally gold plating their offerings.

As data volumes, velocity and variety grew and analytics and AI needs increased, a Hadoop-based data lake approach gained credence in the past ten years. Strongly driven by specific business-led big-data, analytical and, more recently, AI projects—gold taps—Hadoop open-source projects have been slow to address data governance and management requirements. As feared by data warehouse professionals, many data lakes have silted up and become data swamps. The plumbing has not been delivered to spec and is not fit for gold plating.

As a result, many enterprises that have pursued a singular data lake strategy to store and make available all data should now consider migrating significant portions of that infrastructure—those that create and manage core business information and context-setting information—to a more robust, performant and maintainable environment based on modern relational database technology, such as Teradata Vantage.

> Enterprises that built a singular Hadoop-based data lake strategy should consider migrating key sections to a modern, extended relational environment, such as Teradata Vantage, to deliver a more robust, performant, maintainable environment.

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

[2] Cathy O'Neil, *"Weapons of Math Destruction"*, Crown Books, 2016, https://weaponsofmathdestructionbook.com/

[3] Barry Devlin, "An architecture for a business and information system", IBM Systems Journal, February 1988, http://bit.ly/EBIS88