Rethinking Hadoop for Modern Analytics

Since its initial rise in 2008, Hadoop and its yellow elephant logo have become ubiquitous in the data world and are largely synonymous with the concept and implementation of big data. Both terms—Hadoop and big data—have been so used and abused in marketing by vendors and consultants alike that their real meanings have been confused and obscured.

In the past year or two, Hadoop has fallen somewhat out of favor, experiencing a set of midlife crises: of identity, confidence, deployment, cloudiness, and data governance. This series of five ThoughtPoints, published from October 2019 to January 2002, explores Hadoop's strengths and weaknesses, and what we should do about them as we enter a new decade when analytics has become a central aspect of digital transformation.

We conclude that Hadoop is not yet dead, but that in significant aspects of its current use, enterprises would be well advised to revisit relational technology as a foundation for improved data and systems management, and as a single access point for analytics distributed among multiple technologies and across a hybrid on-premises and cloud delivery environment. Teradata Vantage™ with Advanced SQL is offered as a sound foundation for such an approach.

Contents

SPONSORED BY

teradata.

www.teradata.com

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?

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.

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.

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

Relational is the New Black—Uniting Data and Context

OCTOBER 2019

THOUGHTPOINT 2 OF A 5-PART SERIES BY
DR. BARRY DEVLIN, 9SIGHT CONSULTING
BARRY@9SIGHT.COM

Hadoop's early premise that big data stores should be largely schemaless and interpreted only at time of use is faulty and must be revisited to ensure valid and appropriate use of the entire information assets of a digital business.

Unicorns. And unstructured data. What do they have in common? Since the beginning of the new millennium, the IT industry has been enthralled by both, yet neither exists. Let's (reluctantly) forget unicorns and instead show that *unstructured data* is also a mythical beast. Check out this data fragment: {06Harrym601gngr35m119052018}. It may seem unstructured, but readers can likely see a name and maybe a date in there; it's clearly structured. Mix up the letters and numbers, but the result is still not fully unstructured data because each character has a detectable binary structure. Data must have structure; otherwise it would just be noise. *Unstructured data* is an oxymoron.

Such pattern recognition in data shows that structure—or ***schema***—is the basis of meaning. Here's a second fragment: {99, Meghan, f, 507, dkbr, 38, m2, 19052018}, with added structure via CSV formatting. Royalists can now identify the context and maybe guess more fields[8]. Context is key to meaning. As seen in my book, *Business unIntelligence*[9], context is the difference between data that computers crunch and information that humans use. Without a viable data schema, finding context and meaning in data is well-nigh impossible, putting insight discovery, decision making, and action taking in digital business at high risk or error and ultimate failure.

A data schema is more than a structure; it is the key to understanding the meaning of its content, especially when used in a digital business.

Schema-on-read—what's that all about?

In the first decade of this century, the processing challenges of volume, velocity and variety of externally sourced big data drove a new, structure-lite view of data storage and management, seen in both the Hadoop and NoSQL approaches. By 2010, a new breed of data professionals declared this an exciting, novel concept: ***schema-on-read***. It proposed that big data should, in preference, be stored in whatever format in which it arrived and that all definition and interpretation of its structure, context and meaning should be postponed until someone needed to use it for some business purpose.

The rationale was that upfront structuring of incoming data was too onerous or in some cases even impossible because:

1. The data volumes were too large and its arrival velocity too fast to allow the structuring and processing required to load it into traditional, relational databases

2. Data structures were variable and rapidly changeable over time, making relational databases largely incompatible with such data and, more important, that it was extremely costly to adjust fixed table schemas to changes in incoming data structure

3. Maintaining the data in its raw form and original silos would ensure nothing of interest was lost, and allow the maximum flexibility in analytics and other business uses

Like the proverbial road to hell, this one is also paved with good intentions. There is some truth—to varying degrees—in each of the above arguments. Technological evolution has weakened some of the original rationale. Since the schema-on-read approach was formulated, many organizations have built data lakes designed to avoid the problems and benefit from the opportunities listed above. Many have fallen foul of the hidden traps that occur when meaning, context and structure are overlooked as the following story demonstrates.

Schema-on-read, seen in Hadoop data lakes, was conceived to ease the challenges of the three Vs of big data.

The parable of the truckloads of data

Trucoeur is an imaginary French truck manufacturer that sells vehicles across the EU. A key competitive goal is to reduce operating costs for its customers. Unplanned downtime and maintenance are expensive; having a truck off the road can cost more than €1,000 per day, excluding parts and labor. When big data emerged in the early 2010s, Trucoeur saw an opportunity to move from scheduled to preventative maintenance by tracking and analyzing dozens of data points in near real-time from onboard mechanical sensors, historical warranty, and parts inventory information, as well as third party data sources—such as weather, geolocation, vehicle usage, and traffic patterns. Predicting high-risk part failures would allow maintenance to be planned around truck schedules, locations, parts availability and more. Savings of more than 25% were anticipated.

The plan was to gather the necessary data in its raw form from over two dozen different feeds into a Hadoop data lake and allow data scientists to access and analyze it to create models of failure modes and predicted timing based on sensor and other externally sourced data. The results would be merged with internal warranty and inventory data. Maintenance plans—what to repair or replace where and when to minimize truck, and even driver, downtime—would then be sent to fleet operators.

The business goals were excellent. The technology budget—based on commodity hardware and open source software—looked very affordable. The project team was staffed and appropriately skilled with Hadoop programmers, experienced Unix systems administrators, and a mix of experienced and newly minted data scientists who knew R and understood and could model truck maintenance.

With hardware and software installed, the data center began to hum quietly as the first data was easily ingested to a schema-on-read model. What could possibly go wrong?

Schema-on-read offers welcome agility in the building and initial loading of a data lake.

In short, data could—and did—go wrong. Very wrong. As the fourth and fifth feeds were connected, alarm bells began to ring, albeit quietly, but ring, nonetheless. The first traffic data from the UK was subtly different from the mainland data already loaded and began to throw the models off track. Of course, the miles vs. kilometers difference was known

and accounted for, but traffic-based predictions became unreliable. The problem was traced to another factor that was well-known but not included in the earlier data: the UK drives on the opposite side of the road and UK trucks have the steering wheel on the right, variables that had to be retrofitted to all other data and models.

"No problem," said the data scientists. "Adding new fields is easy in schema-on-read." So, they did. And when the next data problem arose, they added yet more fields. Sometimes they had to exclude or reinterpret existing fields in specific cases. Soon, data in different files and stores was becoming incompatible in subtle but challenging ways.

Then the new variant of the PQ-Plus truck was released. The engineers had added some newly requested sensor data. No problem with schema-on-read. What turned out to be more of a problem was that the engineers had also subtly redesigned some of the existing sensor data outputs for speed and efficiency. That took some time and luck to discover when the number of unpredicted truck breakdowns began to creep up again.

As additional sources are ingested, schema-on-read may lead to the degradation of a data lake into a data swamp.

While schema-on-read is good at addressing the three problems and opportunities listed at the top of page 2, it also brings its own set of challenges, potentially turning a data lake into a data swamp. Let's take a look at the opposite of schema-on-read.

Schema-on-write—what's right with this?

Proponents of schema-on-read contrast it to the traditional "schema-on-write" approach. This latter term was seldom if ever seen prior to the emergence of schema-on-read, because it was almost universally accepted that data should be well-structured by design. The weight of expert opinion was strongly in favor of designing data storage according to the relational model introduced by Dr. E.F. Codd in 1970[10] and Dr. Peter Chen's 1976 seminal paper on entity-relationship modelling[11].

Schema-on-write demands that you model and structure your data and storage before gathering data. Data modelling is, in simplistic terms, the process of refining the rather messy reality of real-world information into something that is suitable for the neat and tidy—and definitely naïve—mindset of a digital computer. Modelling is only a process of rationalization and documentation. To be useful in a computer, it must lead to a schema for the data that implements the model and instantiates the metadata—or, as I prefer to call it, *context-setting information*—that defines its meaning.

This leads us right back to Harry and Megan and the question of how best to build data structures that incorporate context and meaning, preferably in a form that is easily understood by people and performs well for reading, writing, and computation by computers. We already have such an approach: the relational model as instantiated in relational database management systems (RDBMSs).

Schema-on-write—just classical modelling and database design—has long been central to data management in setting context and defining meaning for business.

The RDBMS is a tried and tested technology with forty years of experience embedded. It was true that it did not handle the volumes, velocity and variety of big data well when schema-on-read was gestating. However, relational technology has improved and expanded in scope since then in modern RDBMS environments, such as Teradata Vantage. Furthermore, the challenges and opportunities of big data have also evolved in the interim. Relational is the new black—not just fashionable, but stylish, hardwearing and suitable for all weathers.

Extended relational solves modern big data challenges

When schema-on-read was devised as a solution to the three Vs of big data almost a decade ago, for most businesses, big data was a largely separate and distinct area of data processing, isolated from the traditional day-to-day computing that ran their operations and decision-making support activities. In today's digital business, the distinction has completely disappeared. Big data and traditional data—both externally and internally sourced—are intermixed and used together in multiple business processes. It is no longer realistic to treat them as independent processing environments.

A combined strategy and convergent architecture are required. This is not to say that a single technology base can answer all requirements. Rather, one technology must be chosen as the core—the *primum inter pares*—of the diverse set of technologies required to support all modern information and data management needs. The mandatory and obvious business requirement for pervasive context and omnipresent meaning points clearly toward schema-on-write, instantiated in relational database technology, as the only viable approach to storing and managing the core information of the business. I describe this strategy in detail in the IDEAL (conceptual) and REAL (logical) architectures of *Business unIntelligence*[9].

An extended relational environment, such as Teradata Vantage, offers the core data storage and connectivity to meet the complex data needs of digital business today.

An extended relational environment, such as Teradata Vantage, supports this strategic approach by providing:

- Full support in the relational model for a significant range of volumes and velocities of data ingestion and storage
- The ability to easily change existing database schemas to support data variety and later changes to defined schemas
- Ingestion, storage and management of data in non-relational formats, such as CSV, JSON, XML and more within the RDBMS
- Direct access via SQL, R, Python and a wide array of analytics functions to all data stored in the RDBMS and to remote, distributed data stores, including Hadoop and object stores, such as Amazon S3 and Azure Blob
- Separation of compute and storage, and implementation of both independently on premises and/or in the cloud

Taken together, these features favor a schema-on-write approach to data management, while not precluding the use of schema-on-read where needed and appropriate.

Integrating data and context—done or redone

Digital business is a "big data" world where an enormous percentage of data comes into the enterprise from external—and often poorly constructed and managed—sources. It is vital that data scientists and businesspeople can use it correctly and validly in decision making and action taking. To do so, the structure, context and meaning of this data must be made and kept fully clear from its earliest arrival in the enterprise until the last moment it is used in the digital business value chain. Schema-on-write based on the relational model and exemplified by Teradata Vantage is the most appropriate approach to achieving this goal.

Teradata Vantage combines all required function to give business the ability to use data correctly and validly in decision making and action taking.

While some long-standing data management professionals may see this as no more than a return to old wisdom, the reality is much more. The extended relational approach differs from traditional data warehousing by allowing data to reside outside the RDBMS, while—in contrast to data lakes—mandating that such diverse data is governed according to best data management principles from the relational environment.

The extended relational approach differs from data warehousing by allowing data to reside outside the RDBMS but governed to best data management principles.

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, 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.

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 under-emphasize 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.

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*[12]."

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

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[13] 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) 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.

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.

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:

4. *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.

5. *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.

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

6. *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.

7. *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.

8. *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[14], 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,"* I discussed the concept of the extended relational environment, exemplified by Teradata Vantage™. 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.

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

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.

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

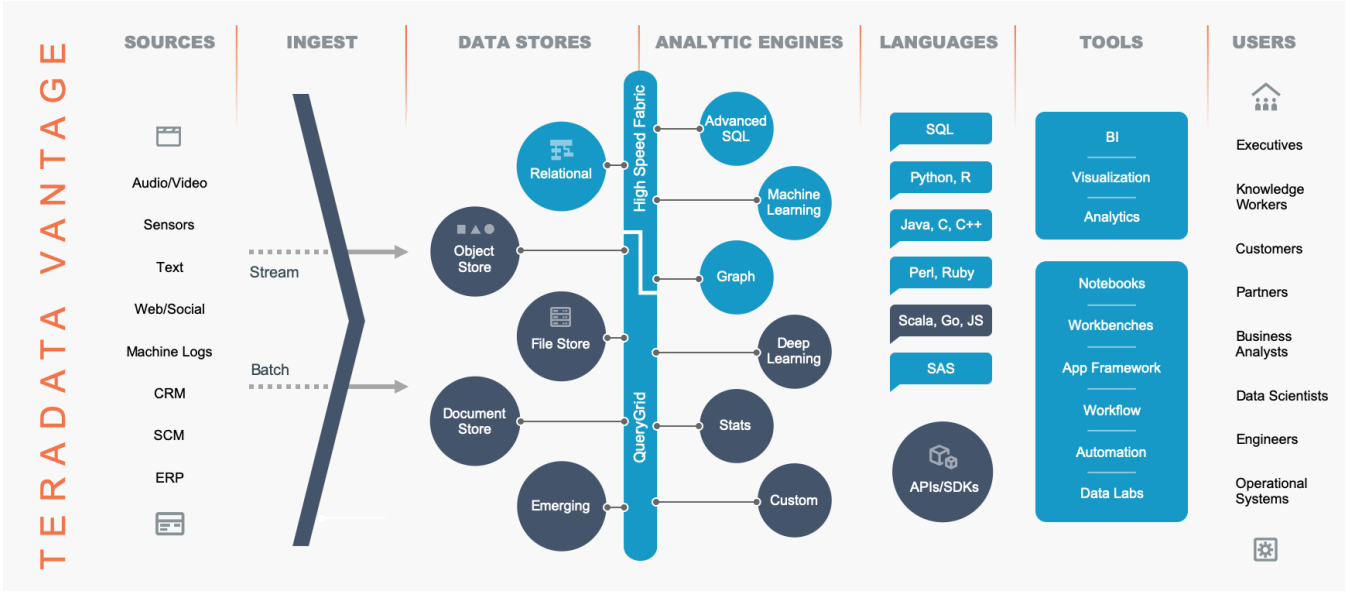

Figure 1 identifies two tools—High Speed Fabric and Query Grid—that compose this bar. However, their identities are less important, representing a point-in-time view of a moving technical implementation of a set of function often called ***data virtualization***. Sadly, that designation carries a lot of marketing baggage. In *Business unintelligence*[9], 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,"* 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.

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[15] 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.

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[16] 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*[9], 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,"* 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*[17] 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/36IWy95

*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] 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

[8] The fields I concocted in the data fragment are: Royal succession, Name, Sex, Height (feet and inches), Hair color, Age, Marriage number, Date of last marriage

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

[10] Edgar F. Codd, A Relational Model of Data for Large Shared Data Banks. 1970, *Communications of the ACM*, 13(6), pp. 377-387

[11] Peter P. Chen, The Entity-Relationship Model—Toward a Unified View of Data. March1976, *ACM Transactions on Database Systems*, 1(1), pp. 9-36, http://bit.ly/35NrQGH

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

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

[14] 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

[15] 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/

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

[17] 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