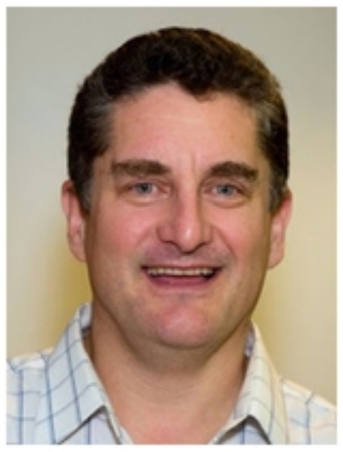
The ascendency of data lakes

The premise, the promise, the potential of new method for managing big data.



BY SEAN MARTIN

The data lake concept occupies a central place of prominence in contemporary big data initiatives. The past two years have unveiled numerous headlines, vendor solutions (including repackaging of former solutions) and enterprise use cases for the utility of this centralized approach for accumulating, analyzing and actuating big data.

The fervor for this method of managing big data is based on a simple premise that promises value for organizations regardless of size or vertical industry. Data lakes provide a singular repository for storing all data – unstructured, semi-structured and structured – in their native formats, granting access and insight to all without lengthy IT preparation.

Moreover, the data lake movement is largely spurred by adoption rates for Hadoop. As Hadoop's presence increases, its function as an integration hub for all data delivers more credence and traction to the notion of data lakes. The data lake concept may be relatively new, but the association

Big data is the principal driver of data lakes.

of Hadoop and big data is nearly as ubiquitous as big data itself.

The combination of these two factors, Hadoop's deployment as a data lake and the storage and access benefits this method produces, is largely responsible for the widespread attention data lakes have garnered. A recent post from Gartner reveals that data lake interest is "becoming quite widespread." *Forbes* indicates that "one phrase in particular has become popular for the massing of data into Hadoop, the 'Data Lake.'"

Most of all, the intrigue behind the data lake phenomenon pertains to the potential of these centralized repositories. In a world in which organizations are confronted with new and differing technologies, tools and platforms daily, data lakes offer something of an oasis: a one-stop hub for all aspects of big data, from initial ingestion to analytics-based action, that makes big data more manageable and demonstrable of its value.

DATA LAKE DRIVERS

Big data is the principal driver of data lakes. Organizations realize the business value that collecting large quantities of data engenders; they understand that exploiting this opportunity will give them an advantage over competitors who do

not. The most immediate advantages of this architecture involve costs for storage and physical infrastructure. Data lakes enable organizations to store massive amounts of data at reduced costs that were not previously available. Additionally, this architecture is extremely scalable and suited for daily ingestion of petabytes.

Alternative methods of storing such data present greater upfront costs than open source Hadoop does. Data lakes also enable organizations to simplify their infrastructure; their comprehensive nature decreases the needs for silos and data marts. Consequently, there is less physical infrastructure, which translates to cost benefits associated with managing and maintaining a single repository instead of multiple ones.

Another driver for data lakes is the increased availability and accessibility they deliver. This advantage is best measured in temporal terms. Data lakes dispel the lengthy data preparation processes that typify the involvement of IT departments with other options for managing big data. Instead, users across the enterprise can access data from the same place with a degree of immediacy that is vital to the speed at which big data is absorbed.

That accessibility correlates to an availability of data that is unparalleled with traditional database life cycles. Organizations can encompass data from different sources (with varying schema and structure, or lack thereof) that utilize multiple technologies (cloud, social, mobile, etc.). Additionally, they can do so to suit the needs of individual business units and across vertical industries, if need be.

Nonetheless, the driver that is likely to make data lakes mainstream is the perception of open source technologies. Hadoop's salience is directly related to the burgeoning familiarity, acceptance, and penetration of open source technologies. Granted, adoption rates for Hadoop reflect many of the foregoing drivers for data lakes. However, its ubiquity is also linked to a greater ease to attain upper-level management support for the data lake concept, since many executives already associate big data with Hadoop.

The notion of dark data, and the realization that elucidating such data improves big data's ROI, also contributes to the ascendency of data lakes. Positioning an organization's entire data assets into a single place provides the first step in attaining insight, and then value, from them comprehensively. With the majority of the world's newly generated data involving unstructured and semi-structured forms, data lakes are poised as the optimal environment to parse and utilize such data in accordance with structured data for a holistic overview of data assets.

COMPARATIVE ADVANTAGES

A comparison between data lakes and traditional repository methods for big data illustrate a number of pivotal advantages and disadvantages – for both. Data lakes are arguably displacing data warehousing as the de facto means of storing data and facilitating analytics. Multiple facets of data warehouses render them unsuitable for the quantities and varieties of big data that are required to truly profit from this technology. The most readily apparent are storage costs, which are exorbitant compared to those for Hadoop. The increase in sources and types of big data merely exacerbates the storage issue, and makes the warehouse approach particularly unwieldy.

This fact is compounded by the time consumption of warehousing and the traditional BI it was designed to support. The business is constantly waiting for IT to model, prepare and transform data before any analysis and reporting is performed, which decreases the value of the velocity at which big data is ingested and consumed. Therefore, warehousing is incongruent with the current self-service movement within data management, which seeks to empower the business and give it more control over its data.

COMPARATIVE DISADVANTAGES

Data lakes rectify the cost concerns for storage and the rapidity of access associated with warehousing time-sensitive big data. However, these benefits become disadvantageous without critical aspects of data management that require

more than just depositing data into Hadoop or NoSQL stores; failing to implement them frequently results in these points of chaos:

Lack of context and meaning: Large data volumes, disparate data types and big data sources are collected in data lakes without any sort of context or readily discernible meaning. Without those conventional, lengthy preparation processes facilitated by IT, end users (or data scientists) are left to implement them as best they can, oftentimes without formal training in this critical prerequisite. The result is an obfuscation of data's meaning and makes data discovery extremely difficult.

Inconsistent data: The jumbled data in data lakes lack semantic and metadata consistency, creating further ambiguity about data's meaning, purpose and relation to other data. Subsequently, there are considerable deleterious effects for …

Data governance: The unrestrained approach of unmanaged data lakes considerably worsens some of the hallmarks of data governance including role-based access to data, security concerns, and transparent data lineage and traceability.

Another serious problem that implementers of early data lakes struggle to address is the scarcity of the data scientist and big data manipulation or even big data programming skills that are usually needed to extract value for or even obtain clean access to the data residing in the data lake. As inflexible and cumbersome as they are, data warehouses can draw on an army of DBAs, armed with a host of mature data wrangling technologies and will generally produce reliable reports on a regular schedule. In many cases data lakes can rapidly resemble a "Wild West" for data.

MAXIMIZING DATA LAKE UTILITY

The data lake concept fulfills its promise via smart data lakes that leverage semantic models and graphs to eliminate the aforementioned points of disorder while adding additional advantages such as delivering drastically improved business end-user self-service capability. Semantic models (based on ontologies) provide concise descriptions of data and are visually represented in a semantic graph. These ontologies clarify data and enhance context by denoting just what the data mean, regardless of source, structure, type or schema. The visual representation of data in a graph illustrates their relationships to one another, providing further context and the foundation for application and analytics usage. These definitions and relationships are digestible for the business and other end users, which expedites their access to and deployment of big data.

Utility solutions:

Role-based access: Semantic technologies also maintain the necessary governance and security policies for long-term sustainability of data lakes. Organizations can implement role-based access to data in accordance with governance protocols by specifying who can and cannot view data elements as expressed by triples. Such access is one of the primary means of engendering order and structure to data lakes based on enterprise-wide policies. Thus, even though the data is in one place, restrictions and permissions to their use are as enforceable as if the data were siloed according to governance mandates, providing internal security for disparate use cases of the same repository.

Provenance and regulatory compliance: Provenance issues are addressed due to the inherent consistency of semantic models and the ease with which it is possible to augment data sets with metadata capturing the originating context and full data lineage; the ensuing

traceability and lineage is critical for determining regulatory compliance. This method allows organizations to analyze any variety of data sources and applications—emails, online account activity, trades, etc.—to see just where and how data was used, and if it was done or should be done in accordance to regulations. The degree of meta-tagging and metadata consistency that such models provide also improves regulatory compliance by enabling semantic models to be mapped to compliance protocols in conjunction with relevant metadata attributes.

Data discovery: The combination of open data standards-based semantic models and their graphic representation also enhances the data discovery process, as end users can query the relationships and meaning of data associated with data sets to see which are appropriate for specific use cases. The application of the semantic standards ensure that the data is both immediately available for reuse and that it is self-describing through the use of standards-based tags that tie them to the associated business concept. This application of semantic technologies may provide the greatest utility to organizations via the sort of celeritous integration of complex unstructured, semi-structured and structured data sets – of any magnitude and type – according to highly specific needs of end users. Depending on the discernible attributes and context of data elements.

In life science organizations for example, clinicians and data scientists have found significant value in quickly juxtaposing the data from multiple clinical trials results through ad hoc queries that navigate across multiple data sets.

In financial services, identifying the potential for misuse of material nonpublic information can be extremely arduous. Links and relationships need to be examined by compliance officers to understanding what, how, why and when information is shared and whether it is compliant or not. Similarly difficult is tying together information that builds a comprehensive picture of counterparty risks.

2016 PREDICTIONS

Analytic expansion: Of all the ways that semantically enhanced data lakes will influence the data landscape in 2016, their impact on analytics will be the most profound. The numerous aforementioned possibilities of such data lakes coalesce into the fact that by deploying them, it is possible to place an organization's entire data assets on an RDF graph, elucidating the relationships between elements in such a way that effectively overcomes the dark data phenomenon. Innately understanding the context and meaning of data

prior to analysis profoundly affects the type, degree and nature of analytics performed, which considerably refines their results and use.

Semantics at scale: The ultimate expression of what is actually an expansion of analytical prowess is the concept of semantics at scale, in which the organization utilizing a smart data lake graph is optimized for analytics with in-memory, massively parallel processing of semantically tagged data. Such an engine, when combined with a smart data lake's RDF graph and ontological models of business meaning, incorporates all relevant enterprise data for comprehensive results at a speed which semantic technology advancements have only recently been able to produce.

Democratization of stewardship: The expedience of access and availability of data provided by data lakes is aligned with the self-service movement and the notion of the democratization of big data that in turn supports it. Data lakes will contribute to the solidification of these trends by facilitating the democracy of data stewardship. Semantic models and semantic graphs will help end users discern data and their relations to other data elements, which will enable a more pervasive form of governance than that conventionally reinforced by only a few dedicated data stewards. With increasing regulatory mandates, this enterprise-wide ubiquity of data stewardship will prove vital to organizations.

Automating IT and data science: Additionally, the alignment of smart data lakes with the self-service movement will result in automation of some of the more mundane, but highly necessary aspects of data science and the work of IT departments. Facets of integration, data discovery and data preparation that consume so much time of those working in these two departments are either expedited or unnecessary with smart data lakes, enabling these professionals to concentrate on more substantial ways to improve data-driven processes and drive more quickly to value.

Finally, the preeminence of smart data lakes themselves will be another trend that should foment in the new year. The interest in this method for managing big data deployments will continue to multiply as organizations realize that they can facilitate all of its benefits, while negating its detriments, through the utilization of user-friendly semantic technologies that belong in front offices as much as, if not more so, than in back ones.

Sean Martin *is the founder and chief technical officer of Cambridge Semantics, a provider of smart data solutions driven by semantic web technology. Prior to Cambridge Semantics, he spent 15 years with IBM Corporation where he was a founder and the technology visionary for the IBM Advanced Internet Technology group.*