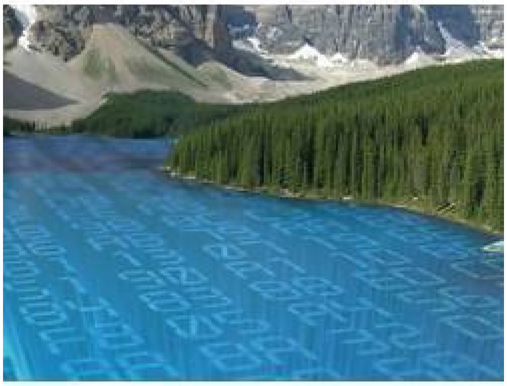
Communications of the ACM article – The Data Lake Concept Is Maturing



The data lake allows raw data, structured and unstructured and in multiple formats, to be stored in one conceptual depository.

Credit: SolutionsReview.com

John O'Brien, CEO of Radiant Advisors, a Boulder, CO-based technology research consultancy, has a succinct view of the current state of databases:

"The concept of database is gone, in my opinion."

O'Brien's sentiment should not be interpreted to mean the reams of data being created by computers and devices across the world will no longer have coherence once stored. Rather, the latest concept in data organization, the data lake—in which raw data, structured and unstructured, of many formats, is stored in one conceptual depository—has emerged.

In contrast to the rigidly pre-formatted data warehouse, the data in the lake is accessed and its schema created upon use ("schema on read") by those with varied purposes, from real-time operational optimization to predictive analytics. This approach, experts say, was an obvious answer to meeting the needs of the new era of big data analytics, in which loosely coupled data is expected to yield valuable new insights.

“I would say the isolated database is gone, but that was already long gone, anyway,” said Jans Aasman, CEO of Franz, Inc., the Oakland, CA-based developer of the AllegroGraph, a semantic graph database employed in a pioneering healthcare data lake at Montefiore Medical Center in New York City. Aasman said the hodgepodge of database technology in large enterprises is a result of the tendency to “solve the problem at hand, and you never have the time to think about the long-term implications of how it integrates with the rest of the things you have.

“What I see now is that companies are aware of this fact. Every big bank, every big pharmaceutical firm, now the hospital world, now knows the way of single unconnected databases is untenable,” he said.

Lakes, Not Swamps

In 2014, Gartner analysts Andrew White and Nick Heudecker released The Data Lake Fallacy: All Water and Little Substance, a report that checked the hype around the data lake. Among the risks White and Heudecker delineated were the inability to determine data quality or the lineage of findings by previous users of the same data; security and access control; and semantic consistency and performance.

The report served as a watershed of sorts.

“Gartner was not alone,” Radiant’s O’Brien said. “A lot of industry analysts, peers I have, were slamming the data lake concept. What we were finding in our clients was data lakes were alive and well; they needed work. The problem was nobody had a good definition of it. Nobody had best practices. It’s easy in hindsight to come up with best practices; it’s much harder when you’re trying to figure out something new.”

In the past year, however, research O’Brien did for both a white paper and market survey has led him to conclude, “I do believe there is a good definition of data lake. There is a component of an overall architecture and strategy. I think the needle has moved a healthy degree in the last year, from confusion to informed.”

The technical architecture of the data lake has, indeed, coalesced into a handful of foundational platforms, from storage and file management to governance to database architecture itself. Some of the emerging building blocks include:

Storage and Processing Building Blocks

While it may not be exclusively synonymous with the data lake, the Apache Hadoop Distributed File System (HDFS) is one of the dominant data lake storage platforms. The introduction of Hadoop YARN (Yet Another Resource Negotiator) in 2012 revolutionized the HDFS ecosystem, adding capabilities for real-time and near-real-time processing to the MapReduce batch-oriented architecture. Ron Bodkin, founder and president of Mountain View, CA-based big data consultancy Think Big, now a part of Teradata, said nearly all his company’s deployments use YARN.

However, Bodkin said the storage component also goes beyond HDFS in many cases. “When people are doing an on-premise deployment, HDFS is by far the most common storage layer that is used for building data lakes,” Bodkin said, adding the proprietary MapR-FS file system by vendor MapR is also popular. Among cloud-based data lakes, however, Bodkin said HDFS yields to object-storage architecture, exemplified by Amazon’s Simple Storage Service (S3), “and when people spin up a Hadoop cluster on Amazon, the HDFS layer is typically being used more as a medium-term cache for faster performance. …People typically keep long-term data in S3 and will integrate multiple clusters with downstream feeds into data warehouses.”

Bodkin also said many data lake architects are taking a hybrid approach when selecting a NoSQL database with which to work with their Hadoop clusters. MongoDB, he said, is typically used for department-level cache applications, Apache Cassandra for highly distributed interactive applications, and Apache Hbase for analytic applications which Bodkin said "can tolerate a bit more latency, having a smaller number of places where machine-learned models sit right next to your compute cluster in Hadoop."

Yet another type of data lake architecture, using graph databases, is emerging in healthcare. The industry is especially ripe for a data lake approach as better-coordinated care between unaffiliated providers is expected to improve individual patient outcomes, and as new insights from aggregated patient records are expected to improve predictive health analytics at the individual level. Additionally, public health data sources such as the National Institute of Health's MeSH database for consistent medical vocabulary, and SNOMED, the international standard for clinical terms, support Linked Open Data (LOD). These readily available resources, and the organizational complexity of modern medicine, lead many industry executives to advocate for a graph database approach, with its flexible node-property-relationship architecture, as a natural fit.

Aasman and Franz are embarking on a partnership with Intel, enterprise Hadoop developer Cloudera, Cisco, and Montefiore, to create a semantic data lake for healthcare. The project features Franz's AllegroGraph, a semantic graph database and application framework for building Semantic Web applications that can store data and metadata as triples; query these triples through various query APIs such as the World Wide Web Consortium's standard SPARQL and Prolog; and apply RDFS++ reasoning with its built-in reasoner. AllegroGraph is also compatible with the W3C's Linked Open Data which, according to Franz materials, shares both the data and the structure of the data via the W3C standard—schema information is built into the data and passed to AllegroGraph at load time, eliminating the need for building schemas.

“I think graph technology is incredibly important, just to get insight to the overall schema that ties all this data together,” Aasman said. “Whether you then put all the data in the graph is a second issue, but it’s incredibly important that you have a meta graph of how everything fits together. And the only way to do that is a graph.”

Governance Tools

One of the most vexing quandaries facing pioneering data lake proponents, as might be expected, is governance. In a lake full of raw data meant to be accessible by a wide range of users, how could the security and provenance of the data be assured? The very concept of the data lake was antithetical to the heavily curated enterprise data warehouse model at the heart of so many businesses’ data policies.

“If you want discovery to go fast and it’s highly iterative, you want to lower the barriers for governance,” O’Brien said. “If I had to request access to this other dataset and governance has a long chain, I would never do discovery. We are telling governance you have to find new policies to enable users and get out of the way, but still manage risk.”

To address that situation, developers from Hadoop platform vendor Hortonworks, Aetna, J.P. Morgan Chase, Merck, SAS, Schlumberger, and Target in 2015 committed to the Apache Atlas project, a governance technology designed to exchange metadata with other tools and processes within and outside of the Hadoop stack, enabling platform-agnostic governance controls. In addition, vendors with value-added products have emerged, such as Waterline Data, which offers technology that automatically catalogs files and fields, including business metadata and data lineage; automates cataloging, understanding, and inspecting data to accelerate the data preparation process; and supports data governance by discovering data lineage, sensitive data, and managing data assets for compliance.

Bodkin compared the emerging data lake ecosystem, in which time-consuming schema construction is eschewed, to the maturation of Agile software development: “you only invest in things that you see value in, and not speculate and say, ‘Well, we think it’ll be valuable some day, so let’s put the work in now and hope it pays off.’

“The fact you can start a data lake where you have raw data that is not well understood; you’ll almost always have some structured data from the get-go, like some kind of unique identifier and a time stamp, and there is usually a little bit more than that. There are some things that are just obviously needed. So it’s more like, going back to that concept of agile modeling, you start off with something that looks a lot more like a file system and the more you work with it, the more it starts to resemble a database.”

***Gregory Goth** is an Oakville, CT-based writer who specializes in science and technology.*