BUILDING A DATA LAKE FOR THE ENTERPRISE

Best Practices Series

Data lakes are forming as a response to today's big data challenges, offering a cost-effective way to maintain and manage immense data resources that hold both current and future potential to the enterprise. However, enterprises need to build these environments with great care and consideration, as these potentially critical business resources could quickly lose their way with loose governance, insecure protocols, and redundant data.

In their book, *BI and Analytics on a Data Lake: The Definitive Guide*, Sameer Nori and Jim Scott identify the role of a data lake, serving as a repository for raw data "which was previously too expensive to store and process." It can then be moved to a data warehouse, or simply serve as "as an online archive for infrequently accessed data." The benefits of a data lake are far-reaching, from boosting big data analytics capabilities to simply serving as a holding tank for raw data that can be refined at some future point by more advanced systems. The best part is data lakes are agnostic to the format of the data being supported—data can come in any format, from any vendor, for any purpose.

In a survey of 385 IT and data managers, Unisphere Research found that the data lake is increasingly recognized as both a viable and compelling component within a data strategy, with companies large and small continuing to move toward adoption ("Data Lake Adoption and Maturity Survey Findings Report," Unisphere Research/Radiant Advisors, October 2015). At the time of the survey, more than 32% of enterprises already had an approved budget to launch a data lake initiative and another 15% had already submitted a budget for adoption. On top of that, 35% were researching and considering a data lake approach. In addition, the Unisphere survey found clear early use cases exist for the data lake. However, governance and security are still top of mind as key challenges and success factors for the data lake.

Besides the mechanical considerations of data storage, there is a need to integrate complex datasets to build more intelligent analytic frameworks for more targeted and intelligent decisions. The data lake helps move big data analytics beyond handling large files of flat data. Advanced analytics, sitting on a data lake, will enable the creation of a robust set of business capabilities, such as predictive and prescriptive analytics.

For example, a key area in which data lakes are proving their potential is the healthcare sector. A semantic data lake for healthcare is underway at Montefiore Medical Center, which involves a sophisticated machine learning project that is slated to go live for patient care in the summer of 2017. The data lake supports predictive analytics to clinicians, starting with the ability to flag patients entering the health system at high risk of experiencing a serious crisis event within 48 hours. The system also generates customized checklists of intervention tasks sent to clinicians that may help to avert or lessen the impact of the crisis.

The following are best practices for making the most of data lakes in the enterprise:

THINK ABOUT THE BUSINESS

As with any technology endeavor, a data lake that is created simply for the purpose of having a data lake will accomplish little. The Unisphere survey finds 70% of data managers see data lakes as a way to boost data discovery, data science, and big data projects—functions critical to today's analytics-driven enterprises. A majority of respondents, 58%, also view data lakes as key to enabling real-time analytics or operationalized insights—again, functions that are critical to businesses that seek to succeed in the 21st century. Thus, alignment with business objectives—better understanding customer needs or creating new revenue streams—is vital, and the data lake needs to accomplish purposes that data warehouses may not be capable of addressing.

THINK EXPERTISE

As with any major data environment, it's critical to have competent professionals. Hadoop skills are still notoriously hard to come by, as is expertise in data governance, data architecture, and data security. Managing data lakes requires renaissance professionals with skills in a broad range of areas. Needed are people who have strong backgrounds in data management and analytics. Data scientists or high-level data analysts are also required to help business users get the most out of their data experiences.

THINK ABOUT WHAT DATA REALLY NEEDS TO BE CAPTURED AND STORED

In today's on-demand environment, many third-party data sources are available almost instantaneously from their original sources, such as Google or Twitter Analytics, or other sites with indexed information that are made available for analysis. These sources are readily available to enterprise decision makers, and it may not be necessary to capture and store the data in a separate corporate location. Data that may need to be stored and maintained is that which is transitory—such as readings from sensors. Before making decisions that will require storage capacity and security considerations, determine what can be best maintained on the web and what needs to be captured. Another consideration is the fact that established data is often very difficult to move—data already existing in transactional systems, such as mainframe platforms, may need to stay where it is.

THINK LONG-TERM, THINK ARCHITECTURALLY

Enterprise architecture and future infrastructure requirements are part of the equation in looking forward over the next 3, 5, or 10 years. While data lakes can be quite fluid and requirements and technologies are changing quickly beneath data managers' feet, a data lake still must be part of a long-term, comprehensive strategy. Since the data lake needs to map to the business, it's important to plan how and what kind of data will be maintained in the repository, who will have access to it, and how it will be governed. Achieving a robust infrastructure capable of scaling for large capacity is also important. Technology strategies such as faster in-memory computing and the use of cloud computing should also be part of the architectural equation. There are many different types of data-level solutions that can be considered and designed, including Hadoop, Spark, NoSQL databases, and data as a service. It will also be necessary to eventually be able to accommodate next-generation approaches, such as real-time data streaming, machine learning, artificial intelligence, and 3D interfaces that will become an essential part of business processes. Also important are the significant technical distinctions between lakes and warehouses. While data warehouse data is highly structured, data lakes store data in its original format—structured, unstructured, and semi-structured. Because data warehouse data is loaded through extract, transform, and load processes, it is more expensive per byte than data lakes, which do not thoroughly vet incoming data. In addition, data warehouses are more likely supported by commercial vendors, while data lakes tend to be home-grown affairs.

THINK SECURE

One of the most pressing concerns about data lakes is that they may end up as "data swamps." The worry is that they may end up storing untrustworthy data, sensitive corporate data that may be subject to unintentional exposure, or even data that violates corporate policies or ethics, such as personally identifiable information or pornography. Accordingly, the Unisphere survey finds that respondents report the most vexing challenges are metadata management issues (71%), security (67%), and governance (71%). Rules and policies that outline access and permissions need to be established, just as in a standard database environment. The same safeguards and processes employed to protect traditional data environments—such as data encryption—need to be extended to data lakes. Regulations and mandates—such as HIPAA or Sarbanes-Oxley—will apply to data lakes as well. Data held within data lakes is also subject to legal discovery or proceedings, in the same manner as all other electronic data. In addition, disaster recovery and business continuity should be part of a security strategy. In the event of system failure, what datasets get priority in terms of recovery? Which datasets should be replicated in real time, immediately available to end users without so much as a hiccup? As many data lakes may serve as repositories for historical data, they may not require priority, real-time replication.

THINK SELF-SERVICE

All across the data and IT space, there is a growing emphasis on enabling end users to access IT services as they need them. A data lake that delivers value would enable business users to make inquiries with little or no help from the IT department. Framed within permissions, users need to be able to see all the datasets available for analysis or reporting. Ideally, users should be able to, at the click of a button, view an on-demand catalog of datasets and be able to run analysis or reports at any time. They should also be able to identify potential new data sources as well, and understand the trustworthiness of the data. Through a robust self-service approach, the data lake can realize significant value to the business.

—Joe McKendrick

Transform Data Lakes Into Knowledge

DATA LAKE CHALLENGES

Data Lakes address the Total Cost of Ownership (TCO) of data warehouses and the onslaught of Big Data. Unlike data warehouses, Data Lakes try to minimize pre-processing of data and fish out what is needed later. Extract, Transform, Load (ETL) and data integration is delayed until you need to do queries or analytics—so-called late binding. That all sounds great, but tossing everything into the lake in native formats has a number of challenges that need to be addressed.

Meaningful analytics on complex data in a Data Lake requires:

- Intelligent metadata management for curation, provenance and known quality of the data
- Semantics for consistent and uniform taxonomies between silos
- Semantics for cross-linking or data integration—otherwise, silos stay silos
- Entity-level access control/governance

WHY SEMANTICS

Data Warehouses, Data Marts and Data Lakes all lack critical data semantics and are bound by the rigidity of the ubiquitous Relational Database schema. However, a Data Lake based on Semantic Models addresses both problems.

At the core of a Semantic Data Lake model we find two W3C standards: the URI and RDF. The URI is the well-known Uniform Resource Identifier, and RDF stands for data based on the Resource Description Framework. Instead of using strings like "Diabetes Mellitus," Diabetes Mellitus NOS," or "糖尿病" to represent the disease, it is represented by a URI such as:

<http://linkedlifedata.com/resource/umls/id/C0011860>

Using "C0011860" in the above URI example is the Unified Medical Language System (UMLS) designator for Diabetes Mellitus. The beginning of the term "http://linkedlifedata.com/resource/umls/id/" typically denotes the organization defining that URI. This way, the same term <http://linkedlifedata.com/resource/umls/id/C0011860> is unlikely to be used by other organizations for different diseases.

In essence, we no longer use character strings that are ambiguous and without data semantics to represent a concept like Diabetes Mellitus. Instead, EVERY concept within a semantic database is represented by a unique (and universal) URI, and the same URIs within a semantic database or among different databases are designed to represent the same concept. The linkage between data elements is thus automatically established.

COGNITIVE COMPUTING WITH SEMANTIC DATA LAKES

Cognitive Computing in the context of databases and Data Lakes is the use of Artificial Intelligence to do very intelligent inferencing and analytics. Cognitive Computing in general combines structured and unstructured enterprise data that is cross-linked through the use of semantics and uniform terminologies. Once you have a standard vocabulary based on semantics, you can apply algorithms through a database or Data Lake that can understand and leverage the terminology. This is only possible by using a semantic graph database, which also links data and creates self-describing data. By applying machine learning to semantically linked data, you have the beginning of Artificial Intelligence.

The foundation for Cognitive Computing and AI lies in the multitude of facets of semantic technology that provision these applications. The genesis is uniform terminologies underpinned by vocabularies and taxonomies. Semantic Graph databases provide the environment in which semantic statements are used to draw inferences; ontologies enrich the contextualized understanding of data linked together. The ability to learn over time and operate autonomously is the crux of AI, and depends entirely on semantic technologies.

If you really want to solve complex Artificial Intelligence problems, you need a data system that goes beyond just data. You have to create a system that can link to anything outside your own predefined parameters—and that can learn from previous experiences, too. That is where Cognitive Computing comes into the picture, and that is why you need to pay very close attention to how all of these data sets link together. Semantic computing is adaptable to those changes, if it is set on the right path to begin with, and that is what differentiates it from other forms of Big Data analytics.

ALLEGROGRAPH—SEMANTIC GRAPH DATABASE

Unlike traditional relational databases, AllegroGraph provides the unique ability to link data, without manual user intervention, coding, or the database being explicitly pre-structured. AllegroGraph processes data with contextual and conceptual intelligence to resolve queries and help the clients to build predictive analytics, which help them to make better, real-time decisions.

THE SEMANTIC DATA LAKE

A Semantic Data Lake is incredibly agile. The architecture quickly adapts to changing business needs, as well as to the frequent addition of new and continually changing data sets. No schemas, lengthy data preparation, or curating is required before analytics work can begin. Data is ingested once and is then usable by any and all analytic applications. Best of all, analysis isn't impeded by the limitations of pre-selected data sets or pre-formulated questions, which frees users to follow the data trail wherever it may lead them.

FRANZ ALLEGROGRAPH
http://franz.com/