ontotext

Do you know that The Great Lakes contain 20 percent of the world's surface fresh water and are home to some 150 species of fish? Let's imagine for a second that The Great Lakes were data lakes. Imagine how many and how big fish, anglers-data analysts would catch if they know their species, locations and baits.

Data lakes – huge storage repositories of both structured and unstructured data in their native format – have been a trend in recent years. Data lakes differ from data warehouses, for example, in several crucial data management aspects.

In addition, data lakes managed under a semantic graph database (also called an RDF triplestore) help organizations optimize data, costs and resources by creating highly interlinked data and mastering huge sets of heterogeneous data. Thus, Linked Data and Linked Open Data keep fishermen constantly updated on the best locations to throw bait and build bridges invisible to other anglers.

The Origins of The Data Lake

Still, what's the data lake buzz all about?

In order to differentiate data lakes from data warehouses, let's first dig into the origins of the 'data lake' collocation. Pentaho CTO James Dixon is credited with coining the term. In a 2010 blog post Dixon wrote:

> If you think of a datamart as a store of bottled water – cleansed and packaged and structured for easy consumption – the data lake is a large body of water in a more natural state. The contents of the data lake stream in from a source to fill the lake, and various users of the lake can come to examine, dive in, or take samples.

Lakes vs. Warehouses

Diving or swimming in a data lake is nothing like just rummaging through a warehouse full of boxes of stuff. It's keeping empty boxes ready for use in case you want to put more stuff in.

Copyright: Ontotext New York | Sofia | Basel
Reach out to us: info@ontotext.com

Data lakes and traditional enterprise data warehouses differ in the way they approach data storage and management. First, warehouses contain structured data designed for specific purposes, while data lakes have all the data, for all time, allowing for any data to be used in the future.

Next, data warehouses have mostly quantitative metrics data while data lakes incorporate all types of data regardless of source, including all new sources of gathering information such as mobile, social or IoT. Also, warehouses may not have all the source data because they are built to serve a case. It's not so with data lakes.

Data lakes, being repositories for raw data, have all data in their native formats and can be accessed and used at any time. Share on X

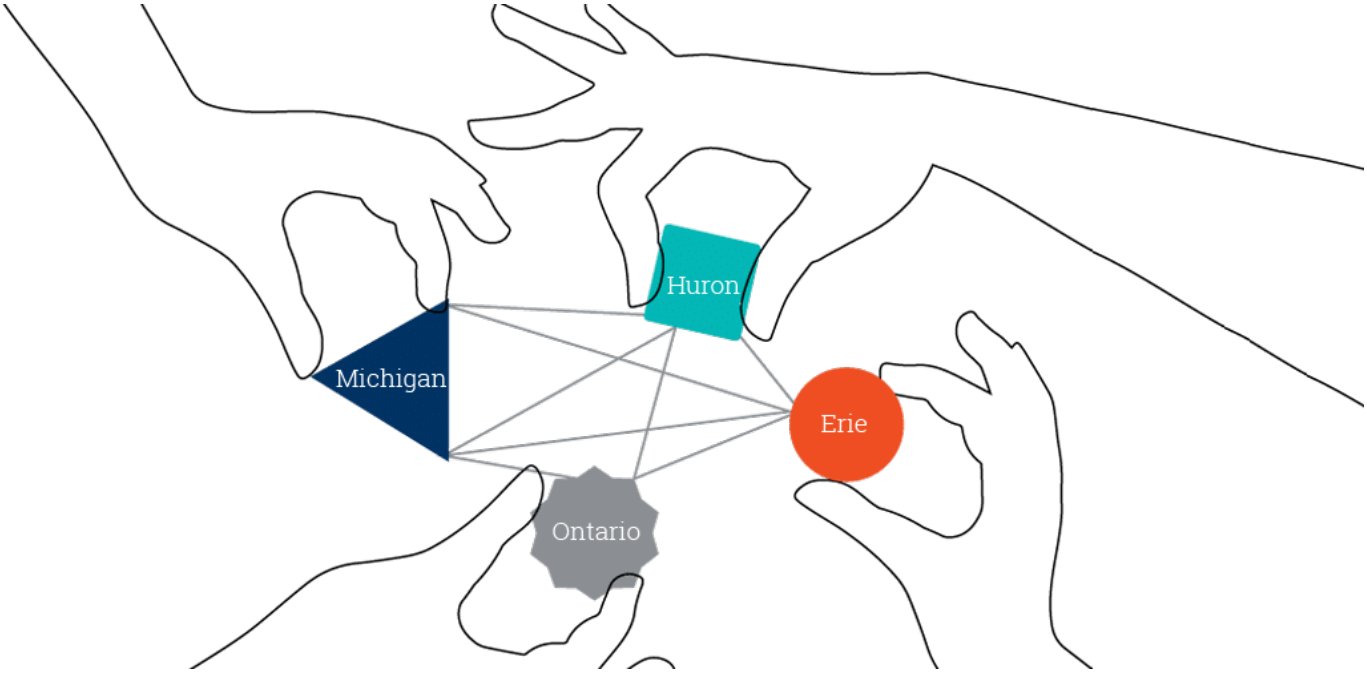
This leads to another difference between the warehouse and data lake management approaches: data lakes are more flexible with changes and are highly agile when configured and reconfigured, compared to traditional warehouse structured data.

Last but surely not least, data lakes allow for a faster pace in getting actionable insights because raw heterogeneous data in native format can be used for various types of Big Data analytics and predictive models whenever needed. Unlike data warehouses which keep transformed and structured data for business professionals mostly.

Linking Data Lakes

Copyright: Ontotext New York | Sofia | Basel
Reach out to us: info@ontotext.com

ontotext



The Great Lakes Waterway of natural channels and artificially built canals allows ships to navigate through the lakes Superior, Michigan, Huron, Erie and Ontario. Though all five lakes are interconnected, water transport needed civil engineering works to pass through the Niagara Falls for example.

Out of the wildlife and civil engineering and into data lakes, we find huge repositories of structured, semi-structured and unstructured data from various sources, kept in native format, and for all time. So how can one navigate and search for insights in such lakes?

The idea of data lakes revolves around having a vast repository of all enterprise data in one place, waiting to be accessed and crunched equally by all business departments and applications, without the need to specially prepare for it. Therefore, tagging and linking the raw data via metadata is essential to identifying relationships out of huge heterogeneous items.

Linked Data, with an RDF database, enables organizations to quickly access their critical actionable information. Share on X

The graph database, where linked data is stored in a knowledge graph structure, allows

Copyright: Ontotext New York | Sofia | Basel
Reach out to us: info@ontotext.com

businesses to reuse data in future applications. By attributing semantic relations to the concepts in raw disparate data, organizations build the bridges to creating data-driven commercial decisions whenever the business environment calls for them.

Building a way to navigate through all the data keeps the lakes fresh and clean and swarming with fish. It also prevents them from becoming data swamps where data is unusable for any operational value.

Enhancing Optimization by Using Data Lakes

The use of data lakes helps organizations optimize their data, costs and resources. Data is being enhanced with the collecting, hosting and analyzing flexible and easily scalable raw heterogeneous datasets. The costs for deploying and maintaining data lakes are lower than those for using traditional enterprise data warehouse solutions, experts agree.

Data lake deployment also optimizes resources by minimizing the labor costs for development and data clean-up until the organization decides how the relevant data it has access to at any time serves its business purposes.

In its 2014 report Technology Forecast: Rethinking integration, PwC said:

> Every industry has a potential data lake use case. A data lake can be a way to gain more visibility or put an end to data silos. Many companies see data lakes as an opportunity to capture a 360-degree view of their customers or to analyze social media trends.

So, data lakes have the potential to lead organizations to untapped streams of data analytics and new streams of revenues. By using linked data in the data lakes, enterprises build bridges to extracting more powerful and more relevant insights from their Big Data analytics.

Copyright: Ontotext New York | Sofia | Basel
Reach out to us: info@ontotext.com

Building Linked Data Bridges To Fish In Data Lakes

Author: Milena Yankova

ontotext

Copyright: Ontotext New York | Sofia | Basel

Reach out to us: info@ontotext.com