3 keys to keep your data lake from becoming a data swamp

Información de publicación: Olavsrud, Thor.

RESUMEN (ENGLISH)

For years, buoyed by technologies like Apache Hadoop, organizations have been seeking to build data lakes-enterprise-wide data management platforms that allow them to store all of their data in their native format. Data lakes promise to break down information silos by providing a single data repository the entire organization can use for everything from business analytics to data mining. [...]Avi Perez, CTO of business intelligence (BI) software specialist Pyramid Analytics said he sees many customers and prospects whose data lakes are deteriorating into data swamps-massive repositories of data that are completely inaccessible to end users.

TEXTO COMPLETO

For years, buoyed by technologies like Apache Hadoop, organizations have been seeking to build data lakes-enterprise-wide data management platforms that allow them to store all of their data in their native format. Data lakes promise to break down information silos by providing a single data repository the entire organization can use for everything from business analytics to data mining. Raw and ungoverned, data lakes have been pitched as a big data catch-all and cure-all.

But Avi Perez, CTO of business intelligence (BI) software specialist Pyramid Analytics said he sees many customers and prospects whose data lakes are deteriorating into data swamps-massive repositories of data that are completely inaccessible to end users.

"Databases are really expensive," Perez said. "The data lake fundamentally answers that problem. Data lakes, and all big data initiatives, come from, one, pressure in the marketplace to have one, and secondly, real-world data generators spitting up gobs of data that you need to find a way to store."

But while a number of the world's most successful companies have built businesses around their data lakes (Google is a prime example), many others are collecting data without any clear way to get value from it.

"They just collect dust," Perez said "You're just collecting junk. I think they'll get abandoned. Eventually you cut the budget for stuff that's big and expensive and not doing anything."

That's not to say the idea behind data lakes is a bad one. Perez is convinced that all companies will need one eventually. But creating a data lake that your end users can actually benefit from requires deliberation.

To avoid drowning in your own data lake, Perez recommends adopting three principles.

1. Collect less data, at least in the beginning

Perez said one of the biggest mistakes organizations make is collecting too much data, simply because they can. Consider your smartphone. If you own one, chances are you've got hundreds or more pictures stored on it.

"You end up with a billion pictures on your phone, and yet 99% of them are probably garbage that you would get rid of in a heartbeat," he said. "It's gotten so easy to take pictures with your phone, it's essentially free. And you probably think, 'One day I'll go and clean it up,' but of course no one ever does. You're collecting an enormous amount of information, but you have no way to work your way through it to use it effectively."

When you inevitably want to show someone a particular photograph, finding it can require scrolling through an enormous volume of junk.

The same thing happens with data lakes, Perez said. Storing data in Hadoop is inexpensive enough that it's often considered free. But the sheer volume of data that accumulates can make it difficult to actually access the data that could provide you with valuable insights.

"I think the way to avoid it is actually to turn the spigot way, way down," Perez said. "Work on the presumption that just because it's cheap to collect the data does not necessarily make it cheap to use it. It could actually be quite expensive. So don't collect information from everywhere and all the time. Keep it focused with a data set where you have a specific plan as to how you're going to mine it."

2. Adopt a machine learning strategy

Even with a focused data set, gleaning insight from data at scale requires automation.

"You need an automated system to clean it up," Perez said. "AI, machine learning, deep learning, whatever term you want to use, it's the magical solution for wading your way through your information. I maintain that the easiest way to get the value out of your huge 5PB data lake is to start with having a technique for how you're going to learn from it."

To start, Perez said, pick a data set you know and select a machine learning technique for going through it. You will likely have to acquire new skills to do it effectively, either through training or hiring.

"Machine learning is a black art," he said. "It's not easy to do. You need very specific skills."

3. Determine the business issue you're trying to address

Here's where everything comes full circle: You need to start with a clear vision of the business problem you're trying to solve. With an objective in mind, it should be relatively easy to zero in on the data you need to collect and the best machine learning technique for gleaning insight from that data.

For instance, imagine you're a big-box retailer. You might decide that you want to understand what kind of customers are coming into your stores. You could capture photographs of customers entering your stores and then use a convoluted neural network (CNN)-a type of deep learning neural network that excels at computer vision problems-to process the images. The CNN could determine whether any individual image is male or female, a child or an adult, a child and an adult, a young person and an old person, etc.

"Once you've got all of that done, tie it up with a business initiative and give that capability to your business users," Perez said. "It could help you determine, 'We need to market more to men because we're not getting enough men.' You really need to have a clear strategy in advance. If you don't, just the mere collection of things becomes a huge negative to the process."

Once you've built a capability with a business initiative in mind, it's often possible to iterate on that capability to provide the business with even more targeted solutions. For instance, once you can identify who's coming into your stores, you can apply that same capability to determining who's walking past your cosmetics counters.

CIO

Credit: By Thor Olavsrud

DETALLES

Materia:	Datasets; Neural networks; Initiatives; Machine learning; Big Data; Automation; Deep learning; Artificial intelligence
Término de indexación de negocios:	Asunto: Machine learning Big Data Automation Artificial intelligence
Empresa/organización:	Nombre: Google Inc; NAICS: 334310, 519290
Título:	3 keys to keep your data lake from becoming a data swamp
Autor:	Olavsrud, Thor
Título de publicación:	Computerworld Hong Kong; Newton

Año de publicación:	2017
Fecha de publicación:	Jun 15, 2017
Sección:	Feature
Editorial:	Questex, LLC
Lugar de publicación:	Newton
País de publicación:	United States, Newton
Materia de publicación:	Computers--Computer Industry
ISSN:	10234934
Tipo de fuente:	Revista especializada
Idioma de la publicación:	English
Tipo de documento:	Feature
ID del documento de ProQuest:	1933320250
URL del documento:	https://universidadviu.idm.oclc.org/login?url=https://www.proquest.com/trade-journals/3-keys-keep-your-data-lake-becoming-swamp/docview/1933320250/se-2?accountid=198016
Copyright:	Copyright Questex, LLC Jun 15, 2017
Última actualización:	2024-11-14
Base de datos:	ProQuest One Academic; ProQuest One Business

ENLACES