Be cautious about open source data mining software

It's free but "should be evaluated like any other software"

By Anh Nguyen, Computerworld UK / Published 00:25, 16 November 10

Businesses should not deploy open source software for data mining just because it is generally cheaper, an open source consultant has advised.

"Don't focus solely on cost savings," said Jos van Dongen, an associate and principal at business intelligence (BI) consultancy Deloitte Group at the Predictive Analytics World conference in London yesterday.

"[Open source] could turn out more expensive because it could require specialised people and more work. Other benefits could be more important."

To this end, van Dongen, who is also an independent consultant in open source BI software, said that companies should evaluate open source software as they would any software, open source or proprietary.

"It doesn't matter if the software is free if it takes longer to build, manage and deploy solutions to end users, or if it is unstable, or missing key features. Don't select just because it is open source," he said.

Van Dongen compared the different benefits of WEKA KnowledgeFlow, an open source tool, and proprietary software SPSS Modeler from IBM, to illustrate his point.

While WEKA is free, extendable and embeddable, and covers more than 95 percent of data mining, van Dongen recognised that SPSS had certain advantages over it.

"SPSS isn't a cheap solution but it is scriptable and it is very powerful. The types of analyses covered by SPSS are much broader than what you can do with WEKA."

"WEKA is old-school datamining, [for instance] you can't do text analytics in WEKA. Whereas with SPSS, you can run only part of a model, run different branches or easily compare different models in the same working environment."

"SPSS is a much more mature interface to work with. So if you want more intuition and power, skip WEKA, go for SPSS."

Despite this, van Dongen believes that if a business does not have any existing tools for data mining, they should make open source the default option.

For organisations in this situation, he recommended open source data mining system RapidMiner, which provides capabilities such as data integration, data analysis and reporting. RapidMiner was rated this year as the most popular data mining tool in the KDnuggets Data Mining and Analytics Software Poll.

However, van Dongen advised against businesses taking a 'rip and replace' approach to its implementation, and suggested instead that businesses plan to augment their existing software with open source.
“Look at gaps in the BI portfolio and data warehouse stack, and use open source to supplement your systems. Try to work in conjunction with existing solutions,” he said.

In addition, van Dongen said that most organisations are adopting open source in an ad-hoc fashion, on a project-by-project basis. He therefore recommended firms consider developing open source policies in order to standardise the process.

1 person liked this.

Like

Add New Comment

Type your comment here.

Showing 2 comments

Sort by Popular now

Jaganadha 5 hours ago

What about Apache Mahout. It is a good one FOSS too

Like

Frank Xavier 6 hours ago

RapidMiner indeed covers significantly more data mining and text mining functionality than Weka and SPSS and I know of several organisations that replaced SPSS Clementine by RapidMiner. However, I fully agree with Jos van Dongen that in many cases a gradual transition for the deployment of open source BI and data mining software is a good idea. Accordingly I know several organisations using solutions like SAS and RapidMiner in parallel. While they keep SAS in place for existing deployments, they approach new projects with RapidMiner.

Like
FAST FORWARD your ability to scale with confidence.
Reduce footprint and cost with 197:1 server consolidation. With the HP ProLiant DL980 G7 server powered by Intel® Xeon® processor 6500 series only £15,499 excl. VAT until 31st January 2011. Find out more.

Announcing the New M86 Secure Web Gateway 10.0
The M86 Secure Web Gateway (M86 SWG) protects organizations from new and dynamic malware using real-time analysis technology. Easy to manage and integrate, the M86 SWG requires minimal support and resources, making it a cost-effective solution.

Subscribe to Newsletter    Register for an account
RSS     Events

© 1994 - 2010 IDG Inc. | About Us | Contact Us | Privacy Policy | Advertise with us | All Topics | Webcasts | Search Computerworld UK | Site Map

Techworld    CIO UK    Macworld
PC Advisor    Marketbase    CFO World
MacVideo    Digital Arts