Saving Time and Money—Why Open-Source BI Makes Sense

Barry Klawans

Barry Klawans is chief technology officer at Jaspersoft.
bklawans@jaspersoft.com

For over a decade, business intelligence (BI) has been sold as an investment—typically a six-to-seven-figure investment that promised to make organizations smarter and more competitive. Admittedly, the approach is valid. Many enterprises have achieved handsome returns on investments in data warehousing and BI, but the "high investment/high return" scenario priced BI out of the reach of most organizations.

Since time is money, I'll get to the point. I've been developing enterprise software for more than two decades; the majority of that time has been spent in the data management space. Here's my recommendation on BI software: don't buy it—at least not until you see a return.

Time Is Money

Have you considered how much it costs to simply *evaluate* enterprise software? The typical enterprise sales cycle is about nine months long. This makes the evaluation process itself a huge investment for both buyer and vendor. Much of the evaluation period is spent building the business case to justify the large, up-front investment in proprietary software licenses.

As the evaluation team builds the business case, they inevitably add requirements to the project—before even seeing any tangible results. The idea is to extract greater returns on that large investment. This process is repeated several times, complicating and extending projects. The longer it takes between project inception and rollout, the greater the lost opportunity, the more up-front risk, and the less value you realize from the software.

The proprietary vendors, in the meantime, dedicate sales, technical, and executive resources to each opportunity. (If you have a project and a budget, you're an "opportunity.") They pitch, build relationships, demonstrate products, coordinate reference calls, and eventually agree to a proof of concept or pilot program.

Now add it all up. That nine-month sales cycle represents a large chunk of the salaries paid to all the employees on your evaluation team, plus the six-digit salaries of the account executives, sales engineers, and other support staff from each vendor. The next time a software vendor pays you a visit or you assemble an evaluation team, look around the room and add it up. Note that you and other buyers like you are paying a premium for all that time.

Money Is Money

Then there are the real, hard numbers put into contracts. Add the software license fees, implementation costs, and 18–20 percent annual maintenance fees, and you have a big investment that must show a big return. Once again the focus is on time—all the costs are borne up-front before you will see any reward. How long will it take to show a return on such a large investment?

This dynamic is one of the key reasons for the growth and popularity of the open-source movement. While open-source methods are evangelized on many fronts, including security, flexibility, and competitive advantage, organizations adopt open-source technologies primarily because of the price/performance ratio. With open-source BI, you can adjust your spending as you go, depending on

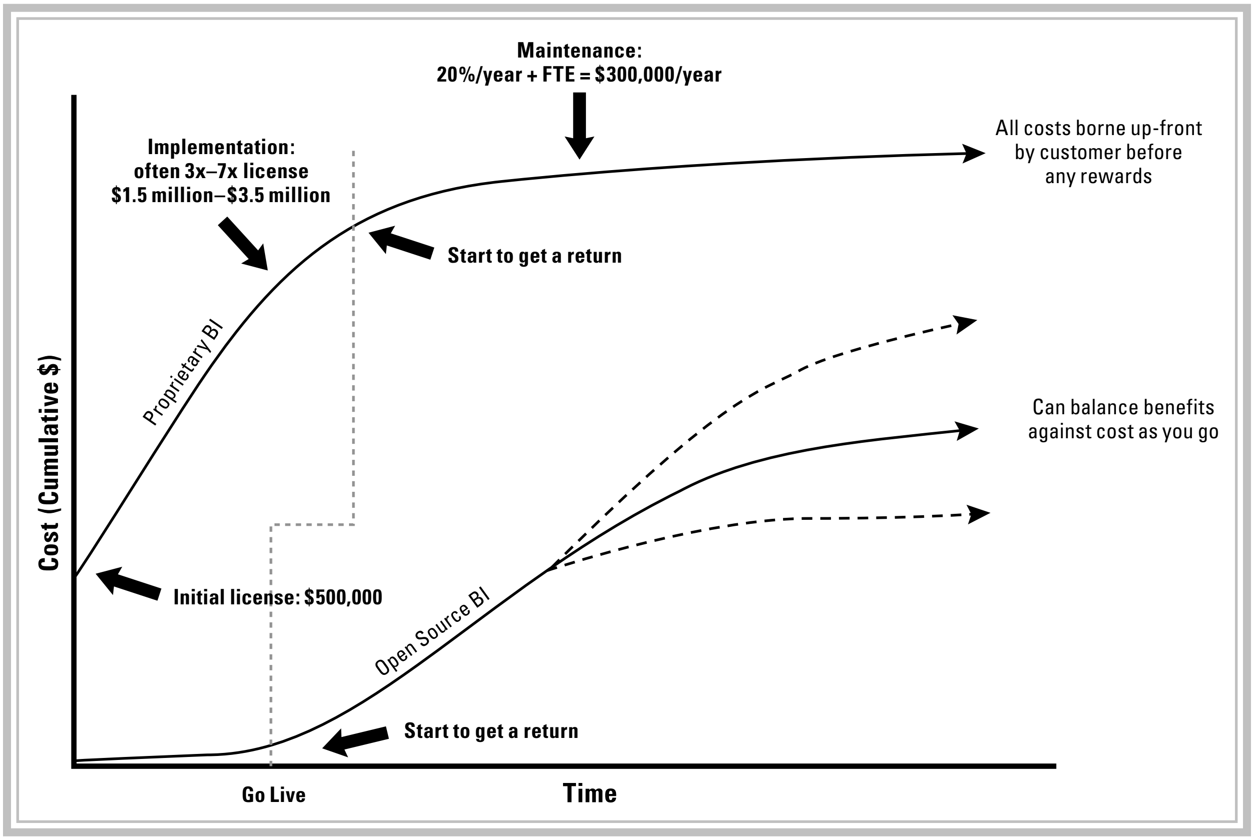


Figure 1. Open source solutions allow faster returns on smaller investments

the perceived returns. Almost all of your money is spent customizing the solution to meet your needs, not on a generic system that needs to be customized just to work. The bottom line—it's all about time and money.

Opening Up Business Intelligence

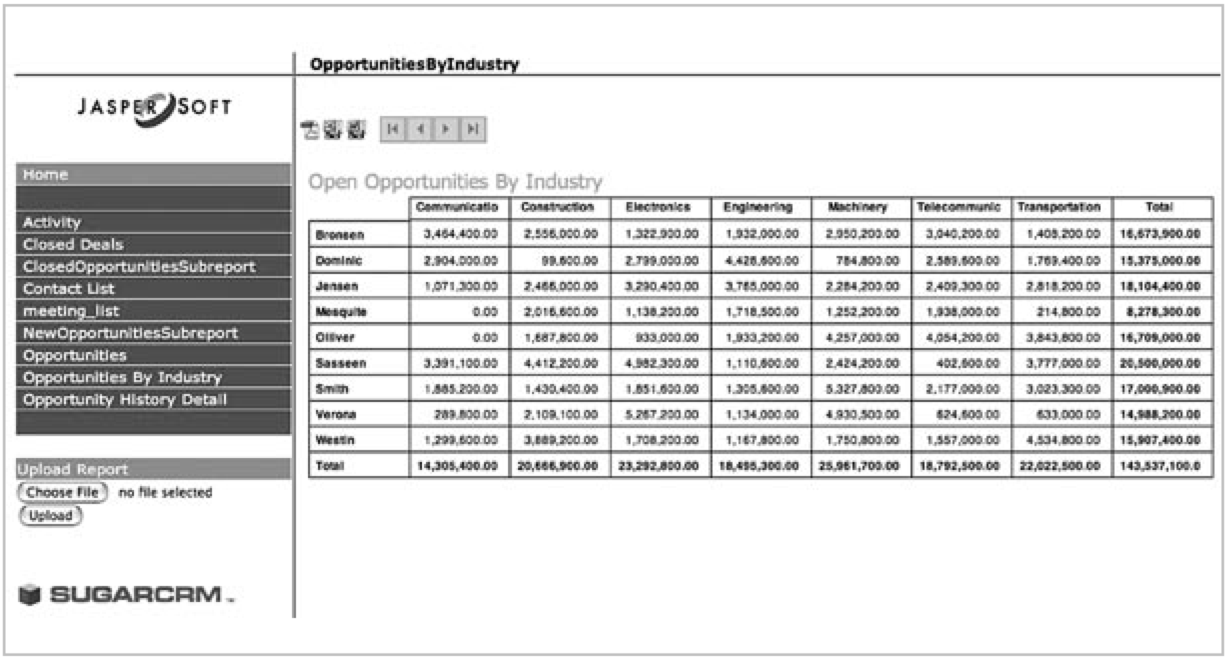
While Linux, Apache, MySQL, and other open-source products are routinely deployed in the enterprise, open-source BI solutions are just now able to stand up to their commercial competitors. The major attraction of open source is to save time and money, but there's an extra dimension to open-source BI: It allows iteration and evolution in a way that proprietary BI does not because the evaluation period is fundamentally different.

With proprietary BI, considerable time is spent selling and buying based on anticipated results. With open-source products, the evaluation is tangible, productive, fast, and free. If you have a specific question, you can download a reporting solution, build the report, get the answer you need (or immediately evaluate another solution that better meets your needs), and you move on.

Can you do this with proprietary software? Yes and no. Proprietary BI vendors may give you permission to install and evaluate their software for free, but getting to that point may take much longer, and they are likely to expand the context of the problem to demonstrate greater long-term value. That's not bad, but getting fast answers for free is better. With open-source BI tools and 24/7 user communities, nimble teams can solve discrete problems one after another, building knowledge and executive support along the way.

Embed, Manage, and Analyze

I made two claims that need explanation. First, I argued that you shouldn't pay for BI software until you see the value, and second, that you should iterate and evolve with open source. Here's how to do both.



Open Opportunities By Industry

	Communicatio	Construction	Electronics	Engineering	Machinery	Telecommunic	Transportation	Total
Bronsen	3,464,400.00	2,556,000.00	1,322,900.00	1,932,000.00	2,950,200.00	3,040,200.00	1,408,200.00	16,673,900.00
Dominic	2,904,000.00	99,600.00	2,799,000.00	4,428,600.00	784,800.00	2,589,600.00	1,769,400.00	15,375,000.00
Jensen	1,071,300.00	2,466,000.00	3,290,400.00	3,765,000.00	2,284,200.00	2,409,300.00	2,818,200.00	18,104,400.00
Mesquite	0.00	2,016,600.00	1,138,200.00	1,718,500.00	1,252,200.00	1,938,000.00	214,800.00	8,278,300.00
Olliver	0.00	1,687,800.00	933,000.00	1,933,200.00	4,257,000.00	4,054,200.00	3,843,800.00	16,709,000.00
Sasseen	3,391,100.00	4,412,200.00	4,982,300.00	1,110,600.00	2,424,200.00	402,600.00	3,777,000.00	20,500,000.00
Smith	1,885,200.00	1,430,400.00	1,851,600.00	1,305,600.00	5,327,800.00	2,177,000.00	3,023,300.00	17,000,900.00
Verona	289,800.00	2,109,100.00	5,267,200.00	1,134,000.00	4,930,500.00	624,600.00	633,000.00	14,988,200.00
Westin	1,299,600.00	3,889,200.00	1,708,200.00	1,167,800.00	1,750,800.00	1,557,000.00	4,534,800.00	15,907,400.00
Total	14,305,400.00	20,666,900.00	23,292,800.00	18,495,300.00	25,961,700.00	18,792,500.00	22,022,500.00	143,537,100.0

Figure 2. *Open source has moved up the stack to integrate applications, reporting, and analytics*

Before you do anything else, take a good look at your report requests backlog. If it overwhelms you, just take a slice of it—perhaps 10 requests, or work with reports that you already created with a tool you're familiar with. Categorize each request into one or more buckets: embed, manage, and analyze.

To determine how to categorize the requests, use the following criteria:

- **Embed**: the data often comes from a single application, such as a CRM, ERP, or other business application. Often the best way to present this information is in the application itself—embed the report and everyone benefits.
- **Manage**: the user wants a report to show up in her inbox on a regular schedule, requires a repository of canned reports, or needs a report to be centrally shared but restricted by role.
- **Analyze**: the data comes from more than one application, for which the SQL would crush the server. Tip-offs for when a report falls into the *analyze* category include the word "by," as in, "I want to see our sales 'by' state" or "deal size and discount rate 'by' brand." Another way to determine whether this is the proper category is by asking your business users to tell you when report data is exported into spreadsheets, indicating that the layout or formatting may need improvement. If data leaves a report and ends up in a spreadsheet, it's usually because additional analysis is needed.

A single request can belong in more than one bucket. Do your best, but don't worry too much about how to categorize each request. The point is to learn and quickly build on that knowledge for your users' benefit. Allow iteration and evolution!

I suggest you start with a low-effort, low-cost, big win. That's the "embed" list. Very good open-source reporting libraries are available that allow you to include reporting systems inside existing applications. You typically need someone with Java skills to put the libraries in the right place. As for creating the reports themselves, the requirements vary. You can use WYSIWYG editors or APIs. Depending on the maturity of the tool, you'll need basic SQL skills for the query, plus scripting (such as Groovy or JavaScript) or Java skills for more sophisticated layouts.

Create a few reports and embed them in your applications. Re-creating existing reports can accelerate the learning curve. It also allows your users to express their opinions. Are the reports you create using the open-source tool better? In what way? Are there additional features in the open-source tool you can use to improve your reports? Could they be created more easily or more cheaply using existing tools? What else is needed to make the report better? What would that take?

While a single report from a single data source isn't what people think about when we say "business intelligence," this approach goes a long way to achieving its goal: getting the right data to the right people in the right context so they can make better, faster business decisions. The effort is minimal, the cost is minimal, and the payoff is huge.

Managed reporting is more complex, and therefore may cost more for labor, support subscriptions, or other services. Compared to commercial solutions, however, open-source products are a fraction of the cost, and you pay only for what you need, when you need it. Open-source managed reporting solutions compare well with commercial solutions. Prioritize the "manage" list, and then build a repository for just one or two business units. For example, focus on Sales, Marketing, or Finance first. Let the business users know your reports use open-source code. Get their feedback and build support from users and executives within that business unit.

Most complex is "analyze." For commercial vendors, this phase requires big budget, big vision, and lots of time. With open source, you need small budget and a few burning questions that need to be answered. For both commercial and open-source solutions, you need DBA, data warehousing, ETL, and programming skills, but you don't necessarily require full-time, on-staff employees. Some tasks are one-time tasks, such as implementing the initial cube or star schema. For commercial vendors, the scope generally gets so large that you also need dedicated program/project management. With open source, you can start small, so you may not need as much administrative overhead.

After that, the keys to success with the "analyze" category are the same for commercial and open-source solutions. Remember that your transactional data store is not suitable for analysis. You need a separate data mart (or data warehouse). Analysis is conducted from data cubes or star schemas. For open-source solutions, we'll focus on star schemas, because they can be built on a good open-source relational database.

Know the questions that must be answered—these come from your report requests backlog. Your task is not to "build a data mart" or "roll out BI," but to solve business problems. If you have the right questions, you can anticipate what kinds of answers you need. Remember the criteria for the "analyze" bucket: "I want to see deal size and discount rate 'by' product." "Deal size" and "discount rate" are facts and "product" is a dimension. (See Figure 3.)

With star schemas, facts are the core. The fact table will be "tall"—that is, it will have many rows, but each row is small. Fact tables contain the most detailed information, and the actual size depends on the facts you are modeling. If you are modeling sales information, you will have one row in your fact table for each order. Put another way, the fact table holds numerical data for analysis, plus indices. Dimensions are smaller, "wide," de-normalized tables that hold structured information about the facts. Dimensions are the "by" tables.

Design the fact and dimension tables before starting so you know where you're going. Acknowledge that you will not get it right the first time, however; and know that users will change their requirements just when you think you have it right. Business intelligence is iterative, and that's one of the key reasons you don't want to invest huge amounts of time or money in a big project until you deliver real value.

Here are a few tips on maintaining the database:

- You can use a commercial or open-source database for your data mart.
- Tune your database correctly. Create indices on the dimensions used for searching and on all the columns used for the primary key/foreign key relations.
- Don't worry about the impact of indices on inserts—you won't be doing any.
- Take advantage of any data warehousing extensions your RDBMS provider has created. For example, if using MySQL, use the MERGE engine to build fact tables from subtables, preferably striped across multiple disks.
- Structure queries to search on dimensions first, then pull in only the fact entries needed. For example, in the star schema in Figure 3, you want to identify the product first and then only pull in the sales facts related to that product.
- Data-mart tables are *big*—avoid full table scans.

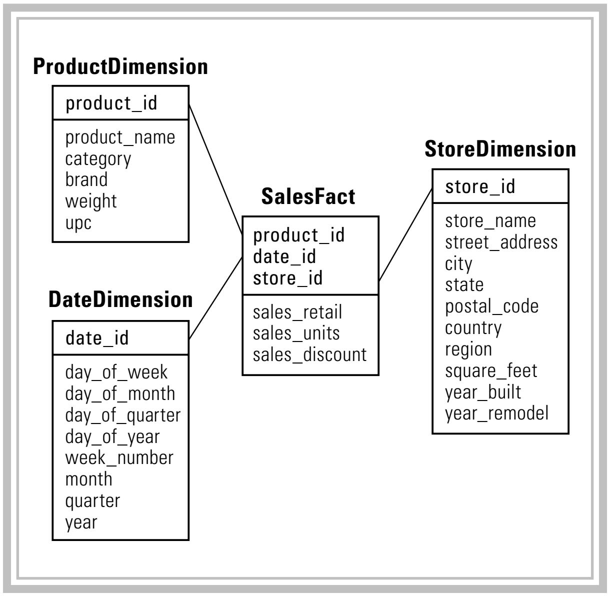


Figure 3. *Typical star schema for analysis*

- Run “ANALYZE TABLE…” or its equivalent after updating the database so the RDBMS updates its performance optimization plans.
- Learn how to use the tools your RDBMS vendor gives you to understand how the database is going to perform query execution.

Finally, be prepared to iterate the ETL process many times to move the data from production systems to the data mart. As you set up the ETL process, you will raise new questions that in turn influence which facts and dimensions must be included in the data mart.

Countering the Downsides to Open Source

Despite all the advantages, open-source technology is not risk-free. The major hurdles are skill set, support, product maturity, and the glut of products to choose from. Fortunately, the open-source movement has matured so you can clear these hurdles. Here’s how.

Smaller organizations priced out of proprietary BI solutions might think that open-source BI is also beyond their reach. This is because they don’t have programmers on staff to evaluate and deploy open-source solutions. For example, a $100,000 company may have only a handful of people in IT, and they are dedicated to maintaining e-mail servers and backup systems, security management, and provisioning employees. They have neither the time nor the required skills to set up and maintain a data mart.

To solve this problem, talk to your preferred value-added reseller (VAR) or systems integrator. There are over 30,000 regional VARs in North America alone. Low barriers to entry and community support make open-source solutions very attractive to the smaller VARs, which increases your ability to get a great implementation at a great price. If you don’t have a preferred VAR, check the open-source provider’s “partners” Web page, or post a comment to a community site. You’ll be sure to get a response. The good news is that the money you spend with a VAR or systems integrator will be on customizing the solution to your needs, not license fees.

Many open-source projects are part-time and poorly supported or abandoned. This is a huge risk when building mission-critical systems and applications. Over the last five years, commercial vendors have sponsored many top-tier open-source solutions. These vendors offer pay-as-you-go support and services, receive and post bug fixes, receive and implement new features, and publish a public roadmap developed with community input.

Many commercial and open-source solutions can handle up to a terabyte of data. If you have more than a terabyte in a single store, you may need to invest in a commercial product that specializes in very large data sets.

Organizations such as O’Reilly’s CodeZoo, Open BRR, SpikeSource, and freshmeat offer information and services that make it easier to evaluate and implement open-source solutions. Technology-industry analysts such as Gartner, Forrester, and the 451group also track popular open-source solutions alongside their commercial counterparts.

Finally, open source should not be confused with free or not-for-profit. For a comparison, review the Free Software Foundation and Open Source Initiative (OSI) Web sites at www.fsf.org and www.osi.org. You’ll find essays that explain the differences between the free software and open-source movements, and learn important information about the GPL (GNU public license), LGPL (lesser GNU public license), and much more.

Conclusion

Business intelligence has been a “high-investment/high-return” solution for a long time. However, its price kept BI out of reach for most organizations. A new methodology is available with open source that brings BI to the masses.

Open source provides tremendous freedom. The software is freely available, and with open source, you can quickly prototype your desired solution. This allows you to approach executive management sooner with a value proposition that applies to your organization.

You also have the freedom to iterate and evolve. Rather than assembling large teams to catalog and warehouse all corporate data, nimble teams can solve discrete business problems one after another, building knowledge and delivering value along the way with little to no "investment" needed. ■

For More Information

Open-Source Resources		
Business Readiness Rating (BRR)	Allows community ratings of open-source software in an open and standardized way.	www.openbrr.org
Free Software Foundation (FSF)	Promotes the development and use of free software. Provides information that compares free software with open source.	www.fsf.org
freshmeat.net	Resources and information on Linux downloads and Linux training.	www.freshmeat.net
JasperForge.org	Development portal for open-source business intelligence solutions.	www.jasperforge.org
O'Reilly CodeZoo	Provides metrics and information to help developers find open-source components that are actively maintained. Ratings are provided by O'Reilly.	www.codezoo.com
Open Source Initiative (OSI)	Non-profit corporation dedicated to managing and promoting the Open Source Definition.	www.opensource.org
SourceForge.net	Provides free hosting to open-source software development projects. Has the largest repository of open source code and applications, but can be difficult to assess project maturity.	sourceforge.net
SpikeSource	For-profit organization that distributes, integrates, manages, and supports open-source solutions.	www.spikesource.com

Open-Source Solutions for Business Intelligence			
	Organization Name	**License Type**	**Web Site**
Reporting	JasperReports	LGPL	www.jaspersoft.com
	JfreeCharts	LGPL	www.jfree.org
	BIRT	EPL	www.eclipse.org/birt
Analytics	Mondrian	CPL	mondrian.sourceforge.net
	Jpivot	CPL	jpivot.sourceforge.net
	OpenI	MPL	openi.sourceforge.net
ETL	Clover	LGPL	cloveretl.berlios.de
	Octopus	LGPL	www.enhydra.org
	Kettle	LGPL	www.kettle.be
Database	MySQL	GPL, commercial	www.mysql.com
Integrated Packages	JasperIntelligence	GPL, commercial	www.jaspersoft.com