information management

Why Most Business Intelligence Tools Fail the 'Hadoop Test'

by SARAH GERWECK
FEB 3, 2016 6:30am ET

Traditionally, Business Intelligence leverages only some of the most basic statistical techniques available. BI is still largely using 17th-century statistical techniques: counts, sums, averages and extrema. At most, we might use techniques that were used by Gauss and Galton in the 19th century (e.g., standard deviations and quantiles).

In traditional BI, when we're slicing and dicing, we take data that's defined over some complex dimensional space and project it down onto a smaller dimensional space that's easy to understand. Like virtually everything in statistics, we're doing this mostly through regression and clustering.

For example, the reason graphs are so important in BI is that the human brain is extremely good at doing regression and clustering by looking at things like scatterplots—but we're pretty bad at doing it numerically. Where BI does use numbers directly, it's usually in situations where some kind of regression or clustering is already mostly built into the numbers.

For example, if you want to quickly understand a business, you might look first at their quarterly revenue and their growth against the previous quarter or year. The quarterly receipts put that company into some kind of cluster of size, and deltas are a simple form of time-series regression. We're pretty much always trying to look for trends and correlations (regression) or reduce large groups into representative units (clustering).

Where BI falls short

There are now well-known statistical techniques for dimensional reduction (e.g., principal component analysis - PCA) but there is a huge potential for even simpler techniques like automatically pointing out dimensions that correlate well with your KPIs.

On the operational side of things, where we're often looking at individual items more than aggregate behavior, there are also plenty of statistical techniques. One opportunity is using Bayesian techniques to identify maximum-likelihood tops and bottoms. You can put this feature into a broader statistical context of noise reduction and significance testing: we want to know whether something that looks unusual really is.

This also includes things like outlier detection, which is immensely useful to users. Operational and row analysis might also make effective use of things like clustering and similarity analysis, but that quickly starts to move from BI to data mining.

The last area where BI falls short is that it's too naively Cartesian in its data domains. BI is traditionally looking only at either simple enumerations (e.g., gender, ZIP code) or numerical axes (age, income, impressions).

Hadoop naturally holds semi-structured data, such as maps, which offer ways for developers to store these traditional forms of data without doing the work to canonicalize them. This is very powerful in terms of allowing access to more data and lowering barriers to entry, but you end up with the same visualizations, tables, etc.

The most obvious area of truly new data is in graph analytics. Many businesses have access to information about the social interactions of their customers, or information on the marketing path their customers have taken, and traditional BI has very little ability to aggregate and visualize that information. Traditional BI is also simplistic in its treatment of things like time series and geography that contain intrinsic structures that are not always part of the raw data.

How to integrate more stats into BI

One apparent challenge to utilizing these techniques is that the audiences for BI mostly won't understand them—but that's actually okay: Google's users mostly don't understand PageRank either. The solution is to reach the business user with these techniques by presenting the results as recommendations. You don't tell a user "our PCA analysis says x": you say "you're looking at the CTR for a campaign: we think you might be interested the CTR of this campaign is especially well determined by household income and age." Ideally you do this without any paperclips tapping on the screen.

Once you have those recommendations, you offer a set links that take a user directly to visualizations of the options you presented. You want the user to see the connections you're identifying in no more than a couple seconds. The user is still driving, and trying one of our recommendations about where to go next costs almost nothing. The statistics are there to aid the user in exploring their data: they're not the output of the analysis.

This notion that the user is driving is how BI remains different from a data mining product, as the notion that the statistics aren't the output is how BI remains separate from statistical tools. They're all using many of the same fundamental technical tools, but the type of user, the amount of input, and the type of output are all fairly different. Again, like Google, the UX is optimized to make the complex appear simple.

Whereas in the dimensional-reduction side of things the user doesn't have to trust your statistics at all, things are a little different on the row-level, operational side of things. Here there is a spectrum of UI options. The simplest might be something like "sort by Bayesian CTR," which would still show users the true clicks and impressions of each cell or campaign without letting them even see the maximum-likelihood rate you calculated. You can step further away from BI's traditional boundaries by exposing your statistical measures or even by allowing them to set up alerts simply based on something like "statistically unusual."

The areas of time series, geospatial analysis and graphs are a bit more wide open. There are a lot of untapped and well-founded techniques here, but how to make them fit well into BI is another question. At one end of the spectrum, simply exposing new measures and let users make of them what they will. (E.g., we could do something like a Influence score by customer just by finding some eigenvectors).

When it comes to geospatial data, cartograms are very exciting. We could introduce automatic dimensions like distance from coast, red/blue/swing state, etc. This area of automatically generated dimensions offers a lot of rich features that both make a great impression and offer good jumping-off points for serious analysis. Imagine automatically correcting currency metrics for inflation and an integration with GeoIP databases.

To truly realize the promise of these kinds of data will require a lot of thought and experimentation to understand how to visualize it to users and allow exploration to integrate human intelligence.

Going Further

Traditionally, BI was a naturally conservative form of data manipulation: the costs of making a mistake with a business's data are high, and the rewards were not as big as they should have been. As the reward increases, the appetite for risk (adopting new methods and software) increases. Consumer Internet companies have been mavens in using modern statistical methods to run their businesses, and that data driven decision making philosophy is gaining steam in the enterprise.

The key to finding the right balance is to stay true to BI's heart: the human brings the intelligence and the computer acts like a professional clerk. Our job as that clerk is to carry out the intentions of the boss in a way that makes things fast and frictionless—but also to use our experience with the data to point out things the boss might find useful. Hadoop is helping us make it fast to access unprecedented libraries of data, the next inflection point for BI is to act as a guide while users explore their data.

AtScale's OLAP server uses Hadoop cluster idle time to generate statistics and identify relationships, then surface them to today's data explorer. The goal is to bring automated advanced guiding analytics to the Hadoop ecosystem.

The future of BI is evolution, not revolution. We get from here to there by marrying foundational BI concepts with modern statistics and visualization. BI principles are sound: we don't need to reinvent, we need to support their mature ecosystem, and build on top of them. Hadoop emerged as *the* data platform. Enabling more data means our 300-year-old math isn't strong enough, we need more modern math and algorithms.

(About the author: Sarah Gerweck is chief architect at AtScale)