Accelerating Business-Centric Data Preparation in the Big Data World

The "Big Data" revolution has led to unimagined changes in the volume, variety and velocity of information available for use in business decision-making. But it also has vastly complicated the work lives of business analysts who are expected to turn mountains of data into actionable insights.

These professionals require a new approach that goes beyond the current data-preparation tools used by IT departments.

THE CHALLENGE OF DATA PREPARATION

Business analysts have always faced difficulties in preparing data for analysis. There almost always is a diversity of sources, both internal (such as sales, manufacturing and finance) and external (such as third-party providers, public sources and the Internet). Data also comes from a variety of locations, and in different formats (such as Excel, JSON and XML).

To date, analysts have spent the vast majority of their time preparing data and much less time doing the actual analysis. And the Big Data explosion has only exacerbated the problem of getting the right data in the right form for analysis. A recent InformationWeek study on Big Data reported that 59% of respondents said data quality problems are the biggest barrier to successful analytics or BI initiatives.

The ever-increasing volume and variety of data mean an increase in the time it takes to prepare data for analysis. As a result, data preparation, more than ever, is a significant bottleneck choking off informed and timely decision-making.

A recent article in the Harvard Business Review reported research on best practices for how businesses should use data. The article reported that companies with a culture of "evidence-based decision making" consistently see improvements in their business performance. One of the hallmarks of such companies, the research found, was that they "ensure that all decision makers have performance data at their fingertips every day."

That means data preparation cannot be a month-long effort exclusively assigned to IT professionals. It must be a task that business users can accomplish by themselves quickly and easily.

The goal of this new approach to preparing data must be two-fold: to reverse the ratio of data preparation to data analysis so that more time is spent on analysis; and to put the focus of data-prep tools on the business analyst, not the IT staff.

ENVISIONING NEW APPROACHES

To help break through the data-preparation bottleneck, it is important first to define the key elements of a solution. From the standpoint of data analysts, they are:

- Support for real-time, ad-hoc queries that can quickly be modified
- Support for existing tools and processes (mainly Excel)
- Interfaces expressed in business language as opposed to arcane schemas
- Processes that put the emphasis on the content of data, not the structure

A business-focused solution to data preparation also must recognize the unstructured nature of much of the data used, as well as the fact that business processes today are more fluid, less formalized and much faster paced than in the past.

Several advanced technologies are emerging to accelerate and simplify the process of data preparation from the perspective of the business analyst. Many of these techniques are well established in other technology-intensive applications and only now are being employed in data preparation. Let's look at a few of them and see how they can help.

As a common point of reference, we'll apply these techniques to a common task of business data analysts – answering an executive's request to compare sales of selected products across geographies, time periods and distribution channels. The answers being sought must draw upon multiple data sets.

One technique, for example, is latent semantic indexing, an indexing and retrieval algorithm widely used in search. It is used to look for hidden patterns between words and ideas in the context of unstructured data from multiple sources, as a way to identify commonalities across data sets. This allows a search engine to identify a word or phrase across numerous sources, even accommodating misspellings or other variations.

In the case of our analyst tasked with comparing sales numbers, one of the first tasks is to resolve commonly occurring data-preparation nuisances such as a single customer referred to by multiple names, or the same product with several part numbers.

Latent semantic indexing accelerates this task by finding the commonalities among all data sets – e.g., variations on a customer's name, or different spellings of a city – without requiring an analyst to consume valuable time manually sorting through columns and rows.

Another helpful technique is statistical graph analysis, which is used in many applications requiring the recognition of patterns in large amounts of data. It is used widely in social-media applications such as Facebook and LinkedIn, which are looking for commonalities among people, interests and preferences.

Statistical graph analysis and latent semantic indexing are often used together. While indexing extracts the key concepts referred to among multiple data sets, (e.g, a city, a part number, a reseller), graph analysis establishes linkages between concepts across data sets. In the case of our typical business analyst, this would pull together all the relevant information about a single customer. The benefit to the analyst, again, would be less effort spent on data prep and accelerated time to analysis.

A third algorithmic technique with relevance to data preparation is reinforcement learning. This "learns" a user's preferences by tracking "Yes" and "No" responses to suggestions. Again, this is similar to techniques used by social media and e-commerce sites such as Facebook, Flipboard and Amazon, which seek to reinforce a user's "Yes" responses and avoid presenting suggestions that will garner a "No" response.

Applied to data preparation, reinforcement learning can simplify the work of data analysts by taking their feedback on exceptions and reinforcing the correct decisions in subsequent sessions. This saves data analysts from having to make repeated and time-consuming changes or corrections. For example, this would produce a faster, simpler way of ensuring that transactions involving a reseller

called Thomas Jones Network Solutions are grouped together, even when it may appear in some data sets as Jones Networks, or Thomas Jones, or Jones Networks Inc.

This may seem mundane, but the impact of getting the right data in the right place can have huge consequences. Consider the case of global financial services companies, who need to make absolutely certain that when millions of dollars are transferred from one account to another, it is going to the right place. These firms use what's called "legal entity identifiers" (LEIs), and it is essential that their lists of LEIs are constantly updated and accurate.

Tools like latent semantic indexing and reinforcement learning can dramatically reduce the time and increase the accuracy of this essential task.

IMPLEMENTING NEW TECHNOLOGIES

The advanced algorithmic approaches cited above can save data analyst's hours and even days of routine work. But how do analysts tap the power of these techniques? Whom do they ask for such tools?

There are two general approaches. One is to rely on an enterprise's IT department. Programming languages such as R, scripting tools and other data-science resources are widely used by IT professionals. And as noted earlier, the people who build and operate websites for search, social media and other applications have the tools and knowledge to implement techniques such as latent semantic indexing and statistical graph analysis.

The biggest downside to asking IT to apply these techniques, however, is that these tend to be fragile, complex solutions that are not optimized for the massive data sets characteristic of Big Data applications. Moreover, the IT-centric approach perpetuates an environment in which every request for improved data preparation turns into a costly, time-consuming IT project.

An alternative approach is to give data analysts tools they can use themselves by acquiring next-generation, analyst-centric data preparation software now coming to market. This software takes an algorithmic approach to data preparation and gives analysts tools that are expressly designed to look at problems from their perspective, using their language.

THE ROI FROM BETTER DATA PREPARATION

Most enterprises already have invested millions of dollars in data-management tools. So, what's the justification for spending even more on analyst-centric data-preparation software?

The first thing to remember is that most of the data-preparation tools already in use are for IT professionals. So, the benefits we're talking about will come from tools designed for *business* users.

A second important point is to understand that the typical data-preparation process used by IT departments – known as ETL – is antithetical to the needs of business data analysts. ETL is a complex, highly technical, batch-oriented process. What business users need is a process that's flexible, open-ended and responsive to fast-changing business needs.

Tools specifically designed for business users will improve performance in several ways:

- Fewer demands for large-scale IT projects, which typically consume 12 to 18 months and can involves scores of people
- Greater productivity of business analysts, who will spend less time on data preparation and more time on the tasks they're paid to do
- Better informed decision-making, and faster time-to-decision, by senior management
- Better use of data that's already being collected; sometimes, the ability to use certain data for the first time in a meaningful way
- The potential for significant revenue increases and/or cost savings through the ability to, for the first time, identify data-directed opportunities for improvement

SUMMARY

It's hard to argue with the Harvard study that identifies "evidence-based decision making" as a valuable business practice. Who doesn't agree that real facts are a sound basis for action?

But in many enterprises, this is easier said than done because of the landfill being created by the continuous and accelerating collection of so much more data. Many businesses are not capitalizing on the data they have to better inform their decision making, because it's just too hard and time-consuming to get all the data into a consistent, reliable form that can be used for ad-hoc, business-focused analysis.

New technologies and tools are now becoming available that make it possible for business users to gain much quicker access to diverse sources of data. This will make the promises of Big Data more realistic to achieve.

ABOUT PAXATA

Paxata is the first Adaptive Data Preparation™ platform built for the business analyst. Now everyone has the ability to rapidly turn all raw data into ready data for analytics—in minutes, not months – accelerating the time to right insights and action.

Our customers, Pax Pros, now have the freedom to connect, explore, transform, and combine data on their own or work with peers in a shared, transparent environment as they shape data for analytics. Paxata's seamless connection with BI tools like Tableau Software, QlikTech and Excel gives business people total flexibility to use the visualization and discovery solutions they prefer to use. Paxata is a cloud-based, self-service solution powered by breakthrough technologies including semantic algorithms, distributed computing techniques and a highly interactive visual experience.

Paxata dramatically reduces the most painful and manual steps of any analytic exercise, empowering analysts at market-leading companies like Dannon, Box and UBS to drive greater value for the business. In partnership with Cloudera, Tableau Software and QlikTech, Paxata unlocks even greater potential from Big Data and Business Intelligence investments. Founded in May 2012, Paxata is headquartered in Redwood City, California. Visit www.paxata.com, follow @paxata_news and watch www.youtube.com/paxataTV.