writeANGLE Distinguishing Content

Breaking the ETL Bottleneck: Accelerating Data Preparation for Agile BI

Exclusive to BI Journal

By Prakash Nanduri and Nenshad Bardoliwalla, Paxata

Executive Summary

For over 30 years, the process of preparing data for business analysis has been painful for IT and data analysts alike. And it has become increasingly complex as business users utilize BI tools for analysis and visualization. With the widespread adoption of cloud computing and the Big Data revolution that is spawning new data sources with unprecedented volume, variety and velocity, IT and data analysts are at a breaking point.

It is clear that traditional ETL processes are not equipped to meet the fast turnaround requirements of today's data analysts, who are being asked to glean insights from ever-growing data sets at shorter and shorter intervals. While conventional ETL clearly has its place, IT has conceded that far more data is being used than can be handled in the 12- to 18-month project lifecycles they typically manage.

A new detente is emerging between business analysts and IT: there is agreement that new approaches and tools are needed to address the ad-hoc and increasingly agile requirements of today's data analytics.

The Challenge of Data Preparation

Business-focused data analysts have always faced challenges in assembling, cleaning and organizing the data they need to perform their jobs downstream of the trusted data that comes from IT. Even if the only data business analysts used came out of IT, it would still come from multiple systems within an enterprise (e.g., sales, marketing, manufacturing, distribution or finance), none of which share common schemas. And increasingly, data is coming from personal systems (e.g., Excel and Access) and from proprietary systems (such as SAP or Oracle). Data also can come from premium third-party sources (such as Dun & Bradstreet and Nielsen), and public sources like the Census.

In addition, data can exist in multiple formats (such as Excel files, flat files, relational databases, and XML/JSON), and can be organized in different ways – for example, the sales team might categorize Japan as part of Asia-Pacific, while Finance might treat Japan as an entity all its own.

But of course internal sources of data are no longer the only place business analysts are going. The explosion of cloud computing and Web services has opened up vast new sources of information that may be useful in business analysis. The problem with this data is that while it may be attractive to use, it rarely meets the standards of IT departments for quality, security, and consistency of format.

Compounding the "data variety" challenge is the persistent issue of data quality and standardization. It's highly unlikely that data coming from third-party sources has the same level of quality as data held by IT, nor will it map to corporate-defined policies. For example, corporate data might use nine-digit ZIP codes while the data from a partner might not include ZIP codes at all. Or, the marketing team might refer to products using a SKU while manufacturing will use a BOM number..

Traditional methods for preparing data for analysis – which combine highly sophisticated Data Integration, Data Enrichment, Data Quality and Master Data Management processes – continue to be needed, but business analysts still have to do a lot of work downstream from IT before they can get the data ready for their ad-hoc analytics.

It is not surprising that many data analysts spend as much as 80% of their time just preparing the data they need and only 20% doing the actual analysis that they're passionate about and are paid to do.

Consequently, the data preparation phase has turned into a significant bottleneck. While there are many productivity-enhancing BI tools currently available, far too often the data needed for analysis isn't in a format that can be readily exploited for timely insights.

What's wrong with the tools that enterprises already use for data preparation? Why aren't they sufficient to solve this problem?

The short answer is that most of the data-preparation tools in use today have been designed for IT professionals, who are very good at understanding the *structure* of data, but who are not business-oriented people familiar with the *content* of data.

This is not an indictment of IT staffs. Rather, it is recognition that IT professionals and business analysts have different goals and areas of expertise. (Even language can highlight the differences. The term "data warehouse" shows the scale at which IT staffs view their task, while the business analyst may be investigating a single product or market.)

To date, the primary tool that business analysts rely on is Excel, where they copy and paste data from all their different sources, write complex formulas to try to make the data fit together, and have no ability to repeat the process once it's finally done.

Another aspect of the IT/business gap is the veracity of the data. IT is usually regarded as the guardian of data security and integrity. But once a business analyst has imported data and starts analyzing it, IT has no visibility into how data is being used. This can lead to arguments over whose data is being used, whether it's the right data, and whether it's accurate.

The Pitfalls of ETL

Enterprise IT professionals have implemented numerous tools over the years to improve enterprise information management, particularly in the areas of data integration, data quality, and master data management. In addition, ETL has become a broadly accepted process for moving data from multiple sources into a standard, uniform environment.

There are, however, numerous aspects of ETL and other central IT practices that run counter to the needs and interests of business analysts.

Let's just look at one aspect of these IT practices – master data management. Master data management is, of course, a very valuable practice with an important objective – ensuring that there is an authoritative master set of data free from duplications, errors and other problems. But the goals of master data management (uniformity and centralization) are almost antithetical to the needs and work environment of the business analyst, who wants to slice data from several perspectives, perform what-if scenarios, and generally have free rein to pursue whatever might be of interest.

But there are other aspects of ETL and centralized data hygiene practices that are counter-intuitive to the world of business data analysis:

- ETL is complex and highly technical. It is clearly the province of IT professionals who have vastly different backgrounds, training and vocabulary from business analysts.

- ETL is primarily a batch process dealing with large volumes of data. It is not suitable for real-time, ad-hoc queries seeking a quick answer to an open-ended, multi-faceted question.

- ETL is procedural, involving a set of programming-oriented tasks. Business users need a declarative, simple-to-use, results-oriented function.

- ETL is a centralized, structured process oriented toward uniformity. Business analysis is a decentralized, creative process oriented toward diversity.

- ETL is often conducted in isolation from other relevant issues such as data quality, entity resolution and master data management.

None of this is to say that ETL is bad or outdated. It is still very valuable to IT staffs and to many tasks related to data management, especially for slowly evolving data sources like ERP or CRM systems. However, ETL is just not relevant or helpful in meeting the needs of the business analyst trying to answer a question in a day or week.

For example, let's say that a large consumer-goods manufacturer is experiencing problems during the rollout of a new product. It appears that availability in some regions is inadequate to meet demand. In the affected areas, advertising and promotions are being run for an item that isn't yet on retailers' shelves. A multi-function team has been assembled to resolve the problem.

The ability of the team to come up with a solution will depend, at least initially, on having access to accurate, real-time data and insights as quickly as possible. But the relevant data are in different systems and formats such as real-time sales data, lists of distributors and retailers, advertising schedules, production figures.

The team will be looking to its data analysts to provide the right information right away. In such a situation, analysts can't afford to spend days assembling and cleaning sources of information before any analysis can be performed.

It's a classic example of the bottleneck described above: The data is there, the expectations are high, but the time needed to prepare the data for analysis can negate the value of whatever insights are eventually made.

Looking for a Better Way

What can be done about the bottleneck in data preparation?

The first thing to realize is what *can't* be expected, at least in the short term. It's unrealistic to expect (nor is it desirable for) IT departments to dispense with valuable practices that ensure data integrity, security and reliability. Nor can we expect business analysts to quit using their favored tools, mainly Excel.

What does the business analyst most need in a data-preparation solution? There are several key elements:

- Support for real-time, ad-hoc queries. Today's executives want to make decisions quickly. Every request for supporting data can't be an 18-month IT project.

- Support for existing tools and processes. Excel is the tool of choice among the vast majority of business analysts, but self-service BI tools like QlikView and Tableau are also finding acceptance.

- Interfaces that are intuitive and expressed in business language. Business analysts need to be able to push icons that say "sort by region" and "rank by profitability" instead of writing convoluted macros and instructions.

- Processes that put the emphasis on the content of the data, not the structure. You don't hear business analysts talking to their bosses about schemas and disaggregation; they talk about what the data means to their business.

Here are a few techniques that can lead to a solution.

1. Apply new technologies from other fields to the task of improving data preparation.

One technique, for example, is latent semantic indexing, which is widely used in search. This is an indexing and retrieval method that helps identify relationships between words and ideas in unstructured texts. It tackles two of the most difficult problems in queries: multiple words with similar meanings and words with more than one meaning.

Certain problems in data preparation cited earlier – such as single customers referred to by multiple names, or the same product with several different part numbers – can be addressed using latent semantic indexing.

Another helpful technique is statistical graph analysis, which is used in many applications requiring the recognition of patterns in large amounts of data, among them plotting all the connections between participants in a social network.

Statistical graph analysis is highly useful in settings where there is too much data for a single human being to identify and analyze all the possibilities – a situation that is becoming well recognized in today's "big data" environment. This technique can considerably accelerate the data preparation process when the question being posed is of the needle-in-a-haystack variety.

2. Learn from past queries and data-preparation experiences so every new query doesn't start from scratch.

Many business-oriented data queries draw from the same or similar sources time after time. For example, a data analyst for the VP of product management might

frequently use product lists, sales figures and price lists, while an analyst in distribution will often draw upon lists of wholesalers and retailers along with production schedules and sales forecasts.

For these analysts, a technology known as semantic typing can be very helpful. It can detect the types of data being worked with and suggest other relevant sources of data, so that an analyst might simply have to say "yes" to the suggestion that other data sources be imported, rather than requesting each separately.

In addition, semantic typing can automatically make corrections to simple errors such as misspellings or incorrect addresses. The cumulative impact on the daily life of a data analyst could be significant in terms of time spent on routine tasks such as data downloads and cleaning.

Machine learning – specifically, a sub-field known as reinforcement learning – also can simplify the work of data analysts by taking a user's feedback on exceptions and reinforce the correct decisions in subsequent sessions.

3. Hide the complexities inherent in data preparation from the business analyst.

Smartphones and tablets are conditioning us, as consumers, to expect that the complexities of many requests are made invisible behind the simple tap of an icon. For example, a GPS navigation device hides dozens of turn-by-turn driving directions, while an on-line banking app has eliminated the need for deposit slips and even ATMs.

Similar improvements should be made in the process of preparing data for business analysis by removing work steps that resemble programming and replacing them with simplified actions that require no more than tapping an icon.

The result of these and other steps should (and can) be a vast improvement for the life of the business analyst. Instead of slow, cumbersome, batch-oriented processes, data preparation can be a quick, intuitive and iterative process that leverages what IT staffs and business analysts each do best.

Data preparation – which today is a significant bottleneck compromising actionable intelligence – should soon become a barely visible interlude on the way to data-enabled insight and action.

###