DATA PREPARATION:
An Essential Element for Your BI Strategy

By Rich Julius
Product Manager, Progress Easyl
Abstract

The increased availability of user-friendly, self-service business intelligence (BI) visualization tools underscores a need for better data integration and data that is coherent and available for analysis. This requires data preparation. Data that is properly prepared offers BI users greatly improved data access, a unified view of disparate data sources such as CRM or ERP systems, better data quality and integrity, while also improving collaboration between BI stakeholders. This paper examines how data preparation works, as well as the benefits and challenges of different data preparation techniques.
Introduction

The Data Analytics sector continues to experience impressive growth. A.T. Kearney forecasts that worldwide spending on “Big Data” will grow at a rate of 30% (CAGR) from now until 2018, when the market will be $114 billion.\(^1\) Cloud-based Business Intelligence (BI), a subset of this larger market, is expected to grow at 31%.\(^2\) The reason for this growth is clear: Businesses that engage in BI do better. Bain’s “Big Data: The Organizational Challenge” report reveals that companies that use analytics are twice as likely to have top quartile financial performance than those that do not, and are five times more likely to make decisions “much faster than the competition.”\(^3\)

The way that companies do analytics is changing. Collaboration is considered a key success factor in organizations that embrace BI. An A.T. Kearney/Carnegie Mellon University study showed that organizations that do analytics well (“the leaders”) foster a culture of cross-functional collaboration and co-creation of BI assets.\(^4\) A new breed of self-service BI visualization tools make this trend possible. Individuals and small teams can now take on ambitious BI projects with minimal support from large-scale, centralized IT groups or data analytics specialists.

Clean, high-quality data as a basic input to the BI workflow is required for efficient and accurate BI visualization. Analysts need access to report-optimized data sets that are easy to blend, no matter how diverse their origins.

Traditionally, the preparation of data for use in BI has been a cumbersome, time-intensive process. This is changing with the availability of new data preparation tools that create report-optimized data sets. This paper explores data preparation, what it is, and why it’s needed.

The goal of this paper is to increase your understanding of data preparation, the alternatives to doing it manually, and the value of self-serve data preparation tools.

---

1. Beyond Big: The Analytically Powered Organization
2. Business Intelligence, RedWood Capital, April 2014
3. Big Data: The Organizational Challenge
What is Data Preparation?

Data preparation is the work required to give BI tools a view of the data that is relevant to running a business, setting strategy, and making important decisions.

The reality of today’s organizations is that the data needed for BI invariably resides in more than one place. There’s data in customer relationship management (CRM) systems, enterprise resource planning (ERP) suites, custom-built databases, file repositories, financial systems, and so forth. Typically, data is stored in different formats, schemas, and conventions.

Data preparation includes unifying disparate sources of data. The data must ultimately conform to a common schema in order to work properly with BI tools. Figure 1 shows a simple example of blended customer data from the CRM and ERP systems. In this case, the CRM and ERP systems have different field names and values for key customer attributes. In the CRM, Central Co. is listed as “England” with revenue of $643,132. In the ERP, it’s called “Central Company,” located in “United Kingdom” with revenue of £432,114. It’s the same record expressed in incompatible ways.

Data preparation, shown in the bottom square, unifies these two data sources into a common schema and set of data values. Now, the BI tool can do analytics on a consistent data set.

---

**Figure 1 – Example of data preparation to unify disparate data from CRM and ERP systems**

### ERP

<table>
<thead>
<tr>
<th>CUST</th>
<th>CTRY</th>
<th>REV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acme, Ltd.</td>
<td>Great Britain</td>
<td>$ 64,228.90</td>
</tr>
<tr>
<td>Big Corp.</td>
<td>United States</td>
<td>$ 354,254.00</td>
</tr>
<tr>
<td>Central Company</td>
<td>United Kingdom</td>
<td>£ 423,113.16</td>
</tr>
</tbody>
</table>

### CRM

<table>
<thead>
<tr>
<th>Customer</th>
<th>Country</th>
<th>Revenue</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acme</td>
<td>UK</td>
<td>£ 42,345.00</td>
</tr>
<tr>
<td>Big, Inc.</td>
<td>USA</td>
<td>$ 354,254.00</td>
</tr>
<tr>
<td>Central Co.</td>
<td>England</td>
<td>$ 643,132.00</td>
</tr>
</tbody>
</table>

### Prepared for BI

<table>
<thead>
<tr>
<th>Customer</th>
<th>Country</th>
<th>Revenue</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acme, Ltd.</td>
<td>UK</td>
<td>$ 64,228.90</td>
</tr>
<tr>
<td>Big, Inc.</td>
<td>USA</td>
<td>$ 354,254.00</td>
</tr>
<tr>
<td>Central Co.</td>
<td>UK</td>
<td>£ 643,132.00</td>
</tr>
</tbody>
</table>
In some cases, data sources are so different that it’s difficult to do simple field mapping and normalizations. Data may reside in a variety of cloud systems that use different APIs for data access, or it may be found in on-premise databases, or NoSQL “big data” repositories, or even distributed across the enterprise in a confederation of Excel files. Data preparation tools handle the standardization and blending from these heterogeneous systems into report-optimized data sets.

Data preparation can also involve enriching data. For example, it is possible to match business names with the government’s Standard Industrial Classification (SIC) codes from public records. Adding the SIC code to the company name data set enriches it and allows for great analytical flexibility. Similarly, preparing a list of company names ensures consistent quality. Cross-checking company names with an index such as Hoovers corrects improper or out-of-date names. “Johnson and Johnson” becomes “Johnson & Johnson,” “Warner Brothers” becomes “Warner Bros.” and so forth.

Data preparation is usually necessary for any kind of iterative BI process. If two teams are going to work jointly on analytics, there should be data preparation steps built into the project workflow that defines how the data is to be prepared.

Data preparation can also leverage the intelligence of data analysis techniques such as machine learning, natural language, and semantic analysis. This is becoming more relevant with new data inputs from unstructured sources such as social media or online commentary. For instance, a website might contain 100,000 comments, each of which contains words such as “great,” “terrible,” “love,” or “hate.” To make sense of this data for effective BI, the preparation process might involve assigning comments with the word “love” a rating of 10 points while comments with the word “hate” would get a 1. From this preparation process, the unstructured comment data set becomes a quantitative table that is easily analyzed.

Why is Data Prep Needed?

Data preparation has always existed in some form, but today it is highly relevant. When data existed in isolated system-of-record silos and data was cleaned and analyzed by dedicated teams of experts, self-serve data preparation for analytics was less common. Today, with data mined from both on-premise databases and cloud sources, the situation is quite different.
The new generation of self-service BI visualization tools has had an impact on data preparation from the perspectives of both IT and business. For IT, new APIs and data management tools make it simpler to pull data out of multiple sources, but the readiness of the data cannot be taken for granted. Just because it’s possible to create a rapid mash-up of data sets for analysis does not mean that the data is of high quality or adequately prepared for coherent analytics.

Business managers and knowledge workers have a higher level of expectation for data analytics today. They expect speed and flexibility, and the availability of self-service data prep tools has taken a great deal of the burden off the shoulders of centralized technical staff. While this development is generally good from a business perspective, knowledge workers need data preparation tools that are easy to use or the process becomes too complex for them.

Pressure and a sense of urgency surround issues of data preparation. More businesses today depend on rapid analysis of heterogeneous data sets for decision-making at all levels. For example, many companies estimate the “Lifetime Value of a Customer” as a way to plan marketing and customer support.

Figure 2 shows four different data inputs to calculate how much a customer is worth over his or her complete lifetime. This analytical model looks at past purchases from the General Ledger database, estimates pending in CRM, orders pending in ERP, and response patterns to digital marketing – all of which taken together can predict how much a customer is likely to spend in the future. The challenge is to prepare a unified data set to support the BI process.

---

**Figure 2 – Data inputs to estimate the lifetime value of a customer**

For IT, new APIs and data management tools make it simpler to pull data out of multiple sources, but the readiness of the data cannot be taken for granted.

Business managers and knowledge workers have a higher level of expectation for data analytics today. They expect speed and flexibility, and the availability of self-service data prep tools has taken a great deal of the burden off the shoulders of centralized technical staff.

The challenge is to prepare a unified data set to support the BI process.
Benefits of Data Prep

Data preparation confers numerous benefits on businesses that utilize BI. Done right, data prep ensures that BI analysis and data visualization will be as complete as possible. Prep also helps with accuracy and integrity of the BI result. In the example of the lifetime value of a customer, if the data sets are not well correlated, then the results will be flawed, skewing business planning based on those results.

Operationally, data preparation is a recommended best practice because it gives participants in the BI process confidence in the quality of the end results. Data prep can also speed up the BI cycle, assuming it is performed with the right tools, and help knowledge workers avoid review cycles and corrections to BI visualization.

A related benefit of consistent data preparation is that it enables a generally easier, more flexible BI capability. For instance, if the preparation workflow is in place to estimate the lifetime value of a customer, it would be relatively simple to extend the process and estimate the lifetime value of a lost customer. Without starting over again, participants already know that high quality, prepped data is available for analysis.

Continued >>
How Data Prep Can Be Performed

The three most common methods of data preparation are manual manipulation in spreadsheets, large data warehouse projects, or dedicated data preparation technology solutions. The following section reviews how each method works and its pros and cons from the IT and business perspectives.

Manual “Brute Force” Data Preparation

Manual data preparation is arguably the most common – and most inconvenient – approach to data preparation. Almost every analyst has done some version of it at one point or another. Typically, data from each source is exported in a spreadsheet format. Then, someone has to go through each table manually and ensure that the data is properly prepared for analysis. It takes a huge amount of time and effort. Figure 3 shows the workflow, which involves multiple manual steps, including joining tables, emailing prepared data sets for review, and so forth.

The manual approach to data preparation does offer some benefits, though. It is inherently flexible. Knowledge workers can adapt the workflow to any number of changes in BI needs or data sources. The method has virtually no up-front cost or project overhead. The big disadvantages to manual data preparation are that it’s slow, error prone, and requires a considerable investment in costly analyst time, which translates to lost productivity.

The big disadvantages to manual data preparation are that it’s slow, error prone, and requires a considerable investment in costly analyst time, which translates to lost productivity.
Data Preparation through Large IT Projects

At the other end of the data preparation spectrum, large-scale IT projects can automate the handling of data before it goes into BI. These IT projects most commonly involve creating “data marts” or “data warehouses.” In this approach, software developers and database administrators collaborate on a customized solution that extracts data from selected databases. The solution then puts the data through a series of transformations and corrections as specified by the particular BI use case.

The data warehouse method of data preparation offers a high level of control and certainty over the process. With automated extraction and integration, the IT project approach eliminates much of the guesswork and review required with manual data processing. It’s also fast—once you have it built. A properly configured data preparation process can execute its tasks in a tiny fraction of the time required by manual prep. This comes at a cost, though. Setting up the solution itself can cost hundreds of thousands of dollars and take months. Plus, a dedicated team is needed to maintain it and make time-consuming changes as line-of-business decision-makers pose new questions and demand additional metrics.

Technology Solutions

A new generation of data preparation tools has come onto the market that puts some of the power of the large data prep IT projects into the hands of knowledge workers. These tools enable IT, analysts, and line-of-business users to extract data from databases, execute joins, run filters, and implement rules-based data corrections before submitting the prepared data set to BI tools for analysis.

Technology solutions give knowledge workers the ability to do data preparation quickly and accurately without requiring a large IT investment. They are quick to implement, flexible, and reliable. They support a collaborative approach between data analysts, line-of-business decision-makers, and systems administrators.

Unlike monumental data warehouse initiatives, they can be used to quickly answer key business questions, and these questions can be built into a library of reports and analytics. Like any new workplace technology, data preparation tools require new skill sets and a learning curve. However, the business benefits of having an economical and rapid data preparation capability more than justifies the solution.
Conclusion

BI visualization is an increasingly critical element of business management. As more businesses rely on data analysis to make decisions, knowledge workers face pressure to prepare data more quickly and accurately. BI works when analytics tools have access to high integrity, accurate data. This requires thorough preparation of data, including joining tables, correcting values, correlating field names, and structuring natural language data sets.

The manual approach is ill-suited for this new reality; it is simply too slow and error-prone. Undertaking large IT projects to set up data preparation workflows is not always practical. The costs are too high and the development process too slow and inflexible to work in today’s rapid-cycling BI environment.

A new class of data preparation tools, which give end users the ability to conduct automated, sophisticated data prep, offer the best solution. With these technology solutions for data prep, knowledge workers can quickly turn complex heterogeneous data into analytics-ready data sets, flexibly reacting to changes in BI requirements in real time and enabling informed, data-driven business decisions.

To get more information about today’s leading-edge data preparation tools, click below

Learn more