Bloor

Spotlight

Spotlight Paper by Bloor
Author **Philip Howard**
Publish date **March 2015**

Self-service data preparation

It is the automation and self-service capabilities that are provided by data preparation platforms that makes them significant.

Author **Philip Howard**

Introduction

> **What is now emerging as a new market segment are tools that enable data preparation in a self-service manner.**

Data preparation is the meat in the sandwich between data sources on the one hand and business intelligence, (predictive) analytics and visualisation on the other. For some time there has been an increasing trend towards self-service business intelligence and analytics, with user departments licensing tools from companies such as Tableau and Qlik so that business analysts can undertake analysis without needing resort to IT. This self-service approach is fine when it is limited to information that is internal to the relevant department. However, it is increasingly the case that users want to analyse multiple sources of data, often of different types and in different formats, which derive from a variety of internal and external sources.

Such data, as we shall discuss, needs "preparation" in order to put it into a form that is suitable for analysis. Historically this has meant relying on IT, which is the antithesis of self-service. What is now emerging as a new market segment are tools that enable data preparation in a self-service manner. These are the subject of this paper. We will discuss what such products offer, why they are important and what sort of features to look for.

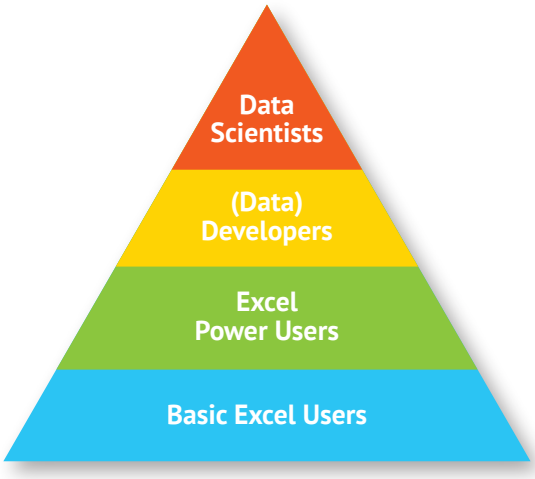
Why prepare data?

> Typical business users spend 70 to 80% of their time simply searching for the right data.

Data needs to be prepared before it can be analysed. It may need to be collated from multiple sources, it may require transformation into a consistent format, you may want to cleanse or de-duplicate the data and you may wish to enrich or complete the data.

There are, broadly speaking, three classes of product that support the preparation of data. The first of these are historic products designed to be used with, and often built into, data mining products. That is, they are essentially designed for relational data and were never designed to handle social media, text, sensor data, and so on. We are not discussing products of this type in this paper. These are not self-service offerings.

Figure 1: Potential users for self-service data preparation



The two types of product that we are considering support all sorts of structured, semi-structured and unstructured data and they are targeted primarily at one of two constituencies: either at business users or at data scientists. Both classes of product tend to be suitable for developers that want, for example, to embed analytics into operational applications. In so far as business users are concerned, different vendors are currently aiming at different parts of the stack illustrated in **Figure 1**, in particular to the extent that they support basic Excel users as opposed to power users. Whatever the target the primary intent (there may be secondary benefits, especially for IT) is to provide self-service capabilities for their respective audiences. We need to discuss these use cases separately. However, before we do that, and in order to avoid any confusion, we should make it clear that platforms of the sort described in this paper do not provide long-term data storage, they are simply places where you can manipulate data prior to analysis.

The business user

Business users want self-service. They want to be able to access data, manipulate it and analyse it without having to go through IT. Or, to be more accurate, it is not IT per se that is the issue but the time it typically takes for IT to deliver the relevant information. In today's business world a delay of a few weeks or even a few days may be too long.

What typically happens today is that users go and grab the data anyway, load into it Excel or a business intelligence tool of some sort and go and do their own analyses. What's wrong with that? Several things. Firstly, the data users are accessing may be subject to compliance restrictions that business users may not be aware of, which means that the organisation, by endorsing such an approach, could be in breach of regulations and subject to financial penalties. Secondly, if multiple users and departments are *"doing their own thing"* then this can lead to a proliferation of business intelligence tools that is costly to license and maintain. Thirdly, there are governance considerations, there may be duplicate or missing information that can skew the results of any analyses, rendering these misleading. Finally, and arguably most importantly, this do it yourself approach can lead to a lot of wasted time. Estimates vary but according to Forbes magazine typical business users spend 70 to 80% of their time simply searching for the right data.

The basic problem is that traditional tools for accessing, manipulating, governing and transforming data were designed for developers and other professionals within the IT department.

While such tools are getting easier to use they were never designed for business users. What have now started to emerge are products that have these capabilities but which are specifically targeted at end users who understand their business data but are not otherwise technical. These offerings are intended not just to make preparing the data simpler and easier but also to significantly reduce the amount of time spent searching for data.

Data scientists

Historically, when we only used to analyse relational data, it used to be estimated that 80% of the time spent in data mining was spent in preparing the data prior to statistical analysis or performing other analyses. And we are talking about 80% of processes that took months overall to complete. In fact these figures have come down in recent years as various data warehousing vendors have implemented in-database capabilities that automate at least parts of this process. However, that doesn't help in the current age of big data and won't help in the coming age of the Internet of Things. This is because those data warehouses are designed for processing structured data and not text (for example) and, secondly, because the scale of the data to be prepared has to be considered: most data warehouses do not have the capacity, or you don't want to pay for the capacity, to do this preparation in-database. So you need a pre-processing capability for preparing the data that is (typically) external to the data warehouse.

Thus for data scientists there is a similar issue as for business users except that the sorts of analyses that data scientists will be performing will typically be more sophisticated and complex, and will require more advanced tooling.

IT implications

There are a couple of important corollaries for IT departments. The first is that developers who want to build some analytics into operational processes can also make use of the sort of capabilities that can be provided for end users. In this respect most developers are in much the same boat as business users: they may be expert coders but they don't tend to have detailed knowledge about things like data governance or compliance. Even where developers are familiar with the data and data models, self-service data preparation can be a significant boon.

The second consequence of deploying a tool for preparing data is that you can use such products to monitor how they are being used because such products have auditing and data built into them, even though these features may be hidden from users. However, these capabilities are accessible by IT who can use them to discover what data users are accessing, how they are manipulating it and what sorts of analyses they are performing. This has two benefits: firstly, it means that IT can proactively advise the business if they are, for example, not complying with PII (personally identifiable information) regulations. Secondly, it means that when it comes to creating a production version of a particular analysis – because it has been successful and the business wants to make it repeatable – then the logic underlying that analysis has been captured by the platform, which will make the production process much easier and faster.

Preparing data

The foregoing discussion probably feels like putting the cart before the horse. What actually is involved in data preparation?

The end goal of data preparation is to have consistent, accurate, complete data, in a standardised format, ready for analysis. So that you can just throw the resulting dataset at the relevant algorithm, process, business intelligence tool or application and you can expect it to run successfully and return useful results. The key question is what would prevent that happening?

There are multiple answers to that question. Here are a few:

- **Empty numeric fields.** Many statistical algorithms work on the basis that a field has a value between zero and one. You cannot assign a blank field with a value of zero (say) and expect to get reliable results.
- **Missing records.** Suppose that you are analysing time-series records: what do you do if some of the readings are missing? This can happen when, for example, the battery in a sensor dies, or when atmospheric conditions disrupt signals.
- **Missing values.** Suppose that you are running an analysis by city. You may have postal codes but missing city fields. It's easy enough to determine the latter from the post codes but it's a process that needs to be completed. The same applies where you may need something like SIC codes, which can be obtained from third party sources.
- **Duplicate records.** Duplication in name and address data is a well-known phenomenon and there are well-known techniques and tools for de-duplication. Less well known is that sensor data is duplicated at rates between 8% and 16%, depending on the type of sensor, so these also need de-duplication. If you are analysing social media the same applies: how do you handle re-tweets, for example?
- **Invalid data.** For example, if your algorithm is expecting a numeric value in a particular field and the actual data is alphanumeric then your analysis will not run properly.
- **Disparate formats.** When combining data from multiple sources the data will often be in different formats (web logs, CRM data, social media, text, sensor data et al) yet your analytic processes will require a consistent format. You will therefore need to transform the data into a consistent format.

These are just a few examples but the message should be clear: the data needs preparing before it can be analysed.

Terminology

As is not unusual in new IT areas there are a host of different terms used in and around data preparation and it is worthwhile to put some meat onto the bones of this varied terminology. Common vocabulary includes:

- **Data shaping** (some vendors call this data preparation): essentially the art (or science) of preparing the data so that null fields, for example, are completed with something useful, that the data is enriched in appropriate ways, and so on. Data shaping is essentially a task that a business analyst or educated end user can perform, prior to doing analysis. Note that that this is not the same as data shaping in the context of Microsoft ActiveX Data Objects, which is something entirely different.

> **Ease of use needs to be explored further. Business users are not going to be running complex statistical analyses, while data scientists and statisticians will be doing so. Ultimately, platforms need to cater to both constituencies.**

- **Data inferencing:** essentially a data profiling capability, ensuring that data in a column matches the metadata that is supposed to describe it and to what extent. Also used as a pre-cursor for recommending join strategies across multiple datasets.
- **Data munging:** the process of transforming a set of data, typically by means of some sort of algorithm or through the use of coding mechanisms, into a format suitable for analysis.
- **Data wrangling:** this is effectively the combination of data shaping and data munging. Like data munging this is typically a task that is more appropriate for data scientists (sometimes called data wranglers in this context) than business users.
- **Data unification:** there are vendors in the data preparation space that are particularly focused on bringing together many (large) datasets from multiple sources prior to data shaping or wrangling, and in providing a collaborative environment for this unification, as opposed to focusing on wrangling processes directly. To this extent they may be seen as complementary to other tools that are focused more specifically on wrangling. In due course, we would expect these distinctions to blur.
- **Data blending:** a term that is virtually synonymous with data unification though there is not quite the same emphasis on bringing together "many" datasets. Data harmonisation is also similar, with an emphasis on consistency in the resulting dataset.
- **Data curation:** this is primarily about data governance and compliance. However, in the self-service world of data preparation business users don't want to know or do anything about governance but, on the other hand, they don't want to get into trouble from doing something they shouldn't. Thus data curation is effectively self-service data governance (and compliance), where the governance is built-in and not something you see. A corollary to having data curation is that IT can monitor who is accessing what data, how they are manipulating it, and what they are using it for.
- **Data refinery:** a data refinery is something that potentially encompasses all of the above. While a number of vendors are referring to their products as data refineries in practice they are not complete yet, even if refinement is the end goal.

In practice there are products focused on end users, there are products targeted at data scientists, there are some whose strengths are more in the area of unification and there are others that emphasise wrangling. Some vendors are addressing or will address the data refinery market. At the present time the market is disparate and still emerging and that is why we are being cautious about our terminology. At some point in the future we would expect these various trends to merge.

None of the preceding is new. We are talking about data quality, data profiling, data enrichment and data transformation. Why, therefore, do you need a platform? Why can't you rely on existing tooling? And if you already have data governance processes in place why can't you rely on that?

Let us take the last question first. Suppose that you already have data quality processes in place. The first point is that this may not be applicable to, say, sensor data, simply because you haven't implemented that yet or because you don't need it for operational purposes. Secondly, this won't apply to things like social media data that derive from external sources or even, for that matter, to internally held information such as documents or emails.

As to the second question we posed, you could rely on existing tooling. However, many organisations won't have tools that address the requirements of sensors, log data, social media et al. Most data governance in most organisations has been focused on governing relational data and not the sorts of data that characterise big data or the Internet of Things. Even if you do have such tools it is likely that they are separate and distinct, with multiple different user interfaces, even if they are provided by a single vendor. Further, traditional tools were designed to be used by developers not business users and will not typically be suitable for self-service business use.

The platform

Almost all the vendors in this space describe their products as being some form of platform. However, they don't necessarily mean the same thing by this term. Specifically, pure play vendors tend to have built a complete platform to support the various facilities provided, whereas more established suppliers may leverage existing capabilities from within their traditional portfolios. In addition, while most companies in this space do not provide any form of business intelligence themselves (beyond simple things like count), this is not always true. While all the vendors expect to integrate with products such as Tableau some suppliers provide things such as storyboards that extend beyond the sort of collaborative projects that are typical in this space.

Whether we call it a platform or not, the question is why you should need a comprehensive suite of capabilities within a holistic environment? The answer is because you want a single place for doing all of this preparation, against whatever type of data, with a single user interface. You also want it to be able to scale easily and inexpensively, which is why many solutions are built on top of NoSQL databases of some sort: either Hadoop or one of its derivatives, or a triple store.

Automation

The key to providing self-service is automation. Platforms in this space typically provide pre-built algorithms to automate various tasks as well as machine learning capabilities (some more than others at this point in time). In both cases these functions are typically based on semantics or some form of artificial intelligence. There are a number of specific capabilities with respect to these that are worth highlighting.

The first pertinent (semantic) capability is the ability to recognise that what appear to be different things (they have different names) are actually the same. A simple form of this would be recognising that a table called "customer-name" is the same as a table called "customer_ID". However, naming conventions (or lack thereof) are such that this is not always a realistic proposition, so what data preparation platforms can do is to recognise that the two tables have very similar content and they can therefore suggest to the user that these are tables of the same type. Note that you need a recommendation here rather than the software just deciding something: a supplier table might look similar to a customer table, for example, and you don't want those two equated.

The second point is that, based on a recognition of the type of data you are working with, the software can recommend other types of data that might be suitable to enrich your data. For instance, if you are working with location-based data, then the platform might suggest including additional geocoding data or population demographics.

And thirdly, there are also quality aspects to this. An example would be that the software might recognise that while Paris, France and Paris, Texas are both valid locations Paris, Brazil is not.

More broadly typical capabilities that might be automated include:

1. Recognising how you had processed data coming in from various previous sources in the past and then predicting how you are going to want to process the data coming in from a new source. The software can then present you with a list of options so that you can simply select your preferred approach and the platform will do the rest for you. Similar facilities will be provided when enriching data from third party sources or for assigning default values.
2. Allowing you to work on a subset of the data, prepare it appropriately and then run the relevant transformations that you have defined against the whole dataset. The platform will present you with any exceptions that might not have been represented in your original subset of the data.
3. Providing automated facilities for things like column splitting, sorting, dropping or aggregation as well as upper/lower case conversion, plus more complex processes such as matching. All such processes are food for the machine learning (**see 1**) that will make data preparation easier in future.
4. Reasoning capabilities that will suggest column types to you (that is, what is the appropriate datatype for this column) based on a histogram of values for that column. As is common with data profiling, mismatched values will be highlighted.
5. Workflow so that you can assign tasks that cannot be automated. Again, machine learning can capture such details to improve performance in the future.
6. In a collaborative environment where there are multiple users, the platform can capture information about what other users are doing (crowdsourcing) and make recommendations on that basis: which are the most popular transformations, what are the most used entities, what are the most valuable attributes? This, in effect, is capturing the *"wisdom of the organisation"*. The platform should also directly support a collaborative environment
7. Once you have 6 in place you can also use this to inform IT so that it can get visibility into what the business is doing with its data. This potentially means that IT departments can start to be proactive rather than reactive with respect to the provision of information (which will be a cultural shift).

Other features you would expect from a platform are that it should be highly visual and easy to use and that it should provide auditing functions and traceability. While some vendors in this space have started out by targeting data scientists the trend is towards making data preparation platforms suitable for business analysts and others who want a self-service approach. There may still be deeper functions (perhaps supported by a platform specific language) that are aimed at data scientists but general functions should be easy to use by all. We should add that some suppliers have taken the opposite approach and have started with a focus on business users. They will no doubt broaden their capabilities to support data scientists in the future.

Ease of use needs to be explored further. Business users are not going to be running complex statistical analyses, while data scientists and statisticians will be doing so. Ultimately, platforms need to cater to both constituencies as both sets of people (as well as application developers wishing to embed analytics into applications or services) have the same sorts of requirement for data preparation.

Conclusion

It is the automation and self-service capabilities that are provided by data preparation platforms that makes them significant. Whether those capabilities are based upon semantics or upon some other technology is really neither here nor there. Being able to recognise repeated actions, similarities and synergies and make recommendations based on these, which lead to automated actions, is a major benefit. It is features such as these that will drive down data preparation times (both in absolute and percentage terms) and will enable collaboration and self-service by users.

Bloor Research believes that platforms of the type under discussion represent a major step forward from what was previously available and, as a result, we expect this market to grow and thrive over the next few years. While there is a general divide between breadth and depth in terms of the offerings provided by suppliers, over time we expect this division to vanish.

About the author

PHILIP HOWARD
Research Director / Information Management

Philip started in the computer industry way back in 1973 and has variously worked as a systems analyst, programmer and salesperson, as well as in marketing and product management, for a variety of companies including GEC Marconi, GPT, Philips Data Systems, Raytheon and NCR.

After a quarter of a century of not being his own boss Philip set up his own company in 1992 and his first client was Bloor Research (then ButlerBloor), with Philip working for the company as an associate analyst. His relationship with Bloor Research has continued since that time and he is now Research Director focused on Data Management.

Data management refers to the management, movement, governance and storage of data and involves diverse technologies that include (but are not limited to) databases and data warehousing, data integration (including ETL, data migration and data federation), data quality, master data management, metadata management and log and event management. Philip also tracks spreadsheet management and complex event processing.

In addition to the numerous reports Philip has written on behalf of Bloor Research, Philip also contributes regularly to ***IT-Director.com*** and ***IT-Analysis.com*** and was previously editor of both ***Application Development News*** and ***Operating System News*** on behalf of Cambridge Market Intelligence (CMI). He has also contributed to various magazines and written a number of reports published by companies such as CMI and The Financial Times. Philip speaks regularly at conferences and other events throughout Europe and North America.

Away from work, Philip's primary leisure activities are canal boats, skiing, playing Bridge (at which he is a Life Master), dining out and foreign travel.

Bloor overview

Bloor Research is one of Europe's leading IT research, analysis and consultancy organisations, and in 2014 celebrated its 25th anniversary. We explain how to bring greater Agility to corporate IT systems through the effective governance, management and leverage of Information. We have built a reputation for 'telling the right story' with independent, intelligent, well-articulated communications content and publications on all aspects of the ICT industry. We believe the objective of telling the right story is to:

- Describe the technology in context to its business value and the other systems and processes it interacts with.
- Understand how new and innovative technologies fit in with existing ICT investments.
- Look at the whole market and explain all the solutions available and how they can be more effectively evaluated.
- Filter 'noise' and make it easier to find the additional information or news that supports both investment and implementation.
- Ensure all our content is available through the most appropriate channel.

Founded in 1989, we have spent 25 years distributing research and analysis to IT user and vendor organisations throughout the world via online subscriptions, tailored research services, events and consultancy projects. We are committed to turning our knowledge into business value for you.

Copyright and disclaimer

2nd Floor
145–157 St John Street
LONDON EC1V 4PY
United Kingdom

Tel: **+44 (0)20 7043 9750**
Web: **www.Bloor.eu**
email: **info@Bloor.eu**