

White Paper



Data and applications

The bottom line is that we are surprised that the major vendors are not doing more to introduce the sort of automation discussed. This is particularly true as Silwood Technology, a company we have been tracking for some time, does this sort of thing

[Philip Howard](#)

Executive summary

We hear increasingly about “the data-driven enterprise”. That is, where information about the business, about customers, about products, about suppliers, about competitors and even about the world around us is regarded as more and more critical to decision making by executives from line of business managers to chief executives.

This is all well and good but it represents a paradigm shift. Thirty five years ago, before the advent of applications such as manufacturing resource planning (MRP) systems, IT was largely data-driven, principally because there was nothing else. However, as packaged application suites became more popular this shifted towards a focus on applications and this accelerated still further with the introduction of enterprise resource planning (ERP) and, more recently, customer relationship management (CRM) systems. This has been the state of affairs for approximately three decades. However, the move towards the data-driven enterprise is reversing this focus. Moreover, whereas data in the seventies was largely an IT issue, it is now a business issue with IT providing the platform and tools that allow the business to use that data.

Unfortunately, this raises an issue: how do you move from an application-focused environment to one that emphasises data? Implementations of technologies that support big data or the Internet of Things are providing much of the impetus for the move to the data-driven enterprise but, as inherently new systems, they do not suffer from the legacy constraints that apply to existing transactional systems. This is not to say that they do not have issues of their own but in this paper our focus is on how you transition from an application-centric to a data-centric environment.

Of course, there are political issues here. We are not going to discuss these but concentrate on what needs to be done from a technical point of view.

The vendor's dilemma

Vendors are in one of three positions. Companies like SAP and Oracle, which are major suppliers of application software, need to transition the information held within their applications so that this data is easily available in a self-service manner with low latency. In SAP's case, in particular, this explains the company's focus on HANA, its in-memory database technology. These companies want and need to provide the necessary analytic platforms for the data used by their transactional applications

Secondly, there are suppliers of ERP applications software, like Epicor and Infor, which do not have the database and other infrastructure technology to support near real-time analytics of the information managed by their applications and they therefore need to rely on third party data warehouse and analytic providers.

Finally, there are all the other database and data integration vendors. In the former case, IBM and Teradata, for example, want the opportunity to provide the platform of choice for analytics. On the other hand, data integration companies, which again includes IBM but also Informatica, Syncsort and so forth, want to be able to provide the glue that links transactional data to the data warehousing platforms that may be used to exploit this data.

All of these companies also want to support implementations of master data management and data governance initiatives that ensure that the information you are providing to the business is accurate, complete, timely and secure.

The key issue, however, is not the provision of low latency analytic platforms nor the ease and speed with which you can transform and load data into these environments, because the former already exist and the latter is similarly available (or, in some cases, you can host the transactional and analytic data on the same platform). No, the key issue is: how do you know what data you need for the analysis you want to do?

In case this last question sounds trivial, consider that in the average SAP ERP implementation there are tens of, if not hundreds of, thousands of tables; of which a large number will be heavily customised and a significant percentage not used at all. Moreover, the naming conventions used by SAP (and the company is not alone in this) are obscure. If you have over a thousand different tables—not uncommon—which pertain to “sales”, how do you know which ones have data in that you want to analyse?

Now, to a significant extent this used not to be a problem. Primarily, because there was no urgency at getting to the data. This is no longer true. In the data-driven enterprise the data is wanted now. This is why there is so much emphasis on low latency, near real-time processing and in-memory technologies. Nowadays, assuming the data quality is good you can, in principal, load up and go. The proviso is that you know what to load; and it is this proviso that is at issue: understanding the data in application environments is potentially the new bottleneck and the dilemma for vendors is how to go about ‘understanding’ the data quickly enough and easily enough to prevent this bottleneck from developing.

The vendor's answer

The basic requirement is to identify those tables that contain the information you want to analyse and to move (or transfer/re-define in the case of environments where transaction processing and analytics will take place on the same platform) that data to your analytic environment. You want to do this as efficiently as possible, as quickly as possible, and at lowest possible cost. In practice, this means automating as much of the process as possible.

There are traditional methods of attempting to understand your data. One is to use data modelling. This is fine for relatively small-scale deployments but once you get into thousands of tables, let alone orders of magnitude more tables than that, it becomes impossible to get any conceptual overview of what is going on—it's not that the tools can't do the job, it's just that the complexity is more than the users can handle. A further problem with data modelling is that this captures the database schema. However, it is typically the case that there are relationships between data elements that are defined at the application level (this is especially true of SAP) rather than explicitly within the database and these may not be identifiable when using data modelling. In order to resolve this particular problem you can deploy data profiling and discovery tools that have the ability to identify implicit, as opposed to explicit, relationships within the data. However, while these tools clarify the relationships that exist, they will also increase the number, so the end result is going to be even more incomprehensible than it was in the first place. Moreover, if you don't know what tables you need to profile then you are going to waste a lot of time and resource profiling random tables that you actually don't care about. Further, many data profiling tools run out of steam if they have to profile too many tables.

Nevertheless, this approach has some levels of automation involved, in that the process of reverse engineering your database schema is automated and that data profiling will automatically detect potential relationships, though you will need a person to determine if they are real relationships or not. Data profiling can also detect empty tables in the sense that these are tables in which all fields are null. On the other hand, it won't help you to select particular tables that have data in that you care about.

The other common approach used is to call in an Oracle or SAP specialist for some consulting. He or she has the advantage, over the tools just discussed, of knowing which of the more than a thousand sales-related tables actually are likely to have meaningful information contained in them. On the other hand, they won't know how you have customised your environment and they will have to spend time learning this on the job, which is expensive. This is entirely a manual process other than the fact that the consultant should have in-built knowledge of the environment, which they can directly access.

As far as today's vendors are concerned that's it: the foregoing, separately or together, makes up the answer. SAP, for example, is increasingly marketing its EIM (enterprise information management) products in conjunction with SAP HANA and that makes sense. However, it remains a process-centric approach—both consulting and data modelling/profiling are essentially process-based in the sense that they are about 'how' the data is used, where what you really want to know is: 'what' is relevant.

The potential answer

What would a better way look like? Let's suppose you want to analyse customer data: who bought what, where, how often, in which combinations, what propensity did they have to buy one product after buying another, and so on?

The first and most obvious thing to do would be to filter out all the tables in your application that have nothing to do with customers and/or products. It shouldn't be beyond the wit of man to do this automatically once you enter your filter/search terms. Next, you could filter out all tables with no values in them. Or, you could do this first. In any case, this should also be automatic.

This is fine if you are happy working with tables, but a lot of business people aren't: tables are more of an IT thing. So, instead you might want to start by considering applications. For example, you might start with the sales order application and perhaps some related applications and then have this hypothetical tool tell you what tables are used by those applications, followed if necessary, by filtering out empty tables.

Whichever approach you take you are now in a position to model the remaining tables in way that does not overwhelm you: a few tens of tables perhaps, rather than tens of thousands. So there is still going to be an element of data modelling but it is not going to be extensive since all the initial work has been automated.

Of course there is a caveat: you can't go doing this against your production system, both for performance and security reasons. In other words you need a product or tool that will read the relevant metadata and then extract it into its own repository, where you can explore it. Ideally, such a repository would be synchronised with the source application so that if the metadata in the latter is changed then this is reflected automatically in the former.

We have to say that all of this sounds obvious, and it does not actually sound very difficult to do, which makes it even more surprising that none of the major vendors are offering such a facility at present.

Conclusion

While all the leading vendors have recognised the trend away from process-centric computing to data-centric computing, caused in part by the fact that processes are being automated out of existence and that “data is the new oil” they continue to be laggards when it comes to automating IT processes away. While self-service BI is all the rage, database administration is, for example, far more complicated than it needs to be (though some vendors are more blameworthy than others) and we believe that there is far more that can be done to automate data integration processes and, in the context of this paper, more automation is needed in understanding application environments.

The bottom line is that we are surprised that the major vendors are not doing more to introduce the sort of automation discussed. This is particularly true as there is a supplier—Silwood Technology, a company we have been tracking for some time—that does this sort of thing. While the company already has some partnerships with leading vendors (but not typically those mentioned) we are surprised that they are not more extensive. There are clear advantages to the sort of approach described in this paper and we have encouraged Silwood (with their product Safyr) to actively seek partnering relationships with other major vendors. We think this is a sensible approach that will be well received by end-user communities and offer commercial advantage to the vendors.

The point is that firing up a warehouse today is relatively easy and quick, and so is analysing the data once it’s in there. If vendors can provide their customers with the ability to know what to load in hours or days, when it currently takes weeks or months, then everyone wins!

Further Information

Further information about this subject is available from <http://www.BloorResearch.com/update/2204>

Bloor Research overview

Bloor Research is one of Europe's leading IT research, analysis and consultancy organisations. 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 over two decades 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.

About the author

Philip Howard
Research Director - Data 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 walking Benji the dog.

Copyright & disclaimer

This document is copyright © 2014 Bloor Research. No part of this publication may be reproduced by any method whatsoever without the prior consent of Bloor Research.

Due to the nature of this material, numerous hardware and software products have been mentioned by name. In the majority, if not all, of the cases, these product names are claimed as trademarks by the companies that manufacture the products. It is not Bloor Research's intent to claim these names or trademarks as our own. Likewise, company logos, graphics or screen shots have been reproduced with the consent of the owner and are subject to that owner's copyright.

Whilst every care has been taken in the preparation of this document to ensure that the information is correct, the publishers cannot accept responsibility for any errors or omissions.



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

Tel: +44 (0)207 043 9750
Fax: +44 (0)207 043 9748
Web: www.BloorResearch.com
email: info@BloorResearch.com