

How do you get more from your Data Warehouse?

A White Paper by Bloor Research
Author : Philip Howard
Publish date : November 2007

The need for data warehousing is growing ever more acute and poses a number of problems for data warehouse providers and users. This paper is an active guide on how to approach the current issues you may have with your existing data warehouse implementation

Philip Howard

This document is sponsored by

DATAUPIA™

The need for data warehousing is growing ever more acute as companies seek to leverage their information assets to better support their business. However, what this means in practice, and in very simple terms, is that you have more users querying more data that it is necessary to deliver in a shorter time frame (and often in real or near real time).

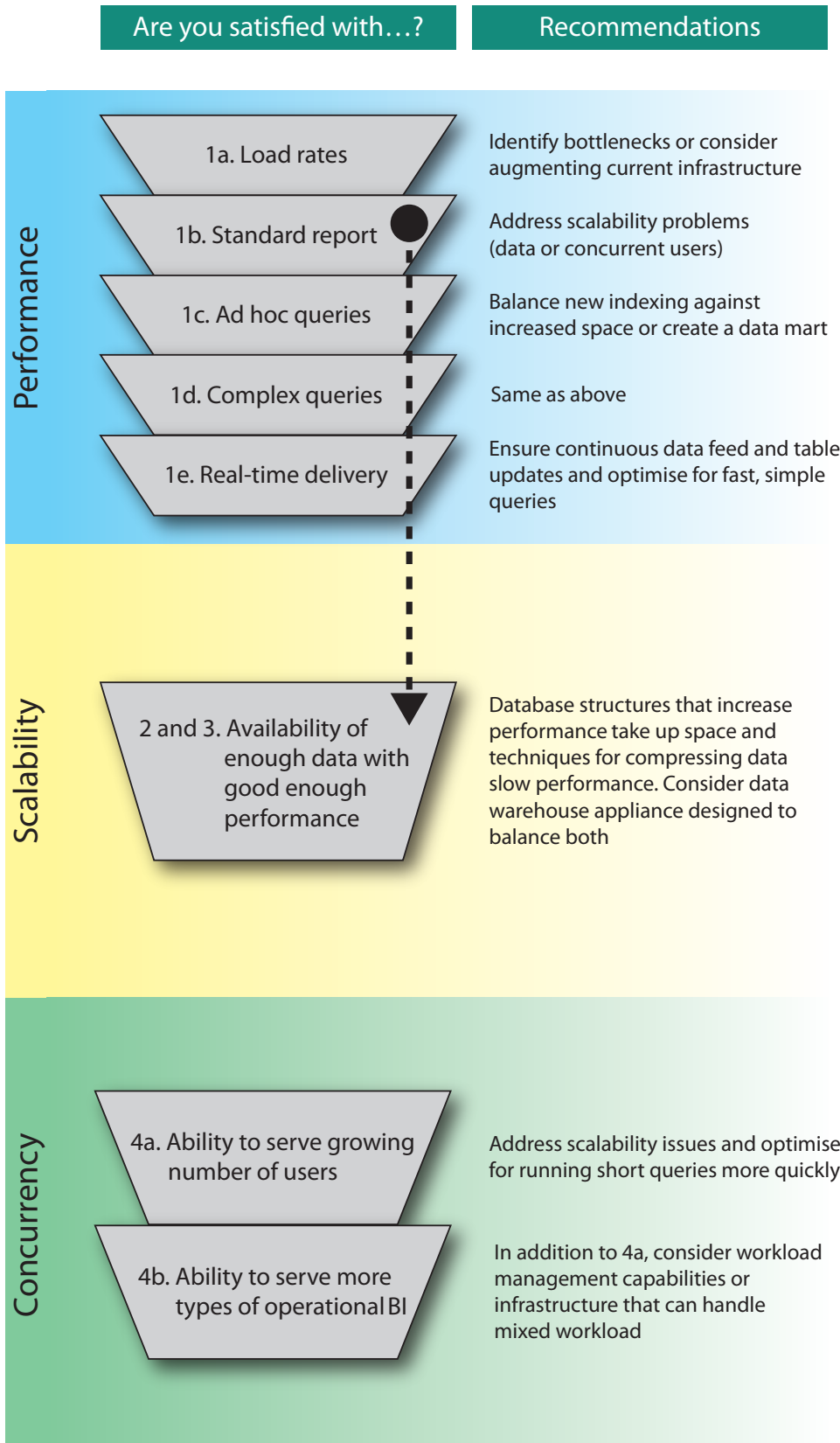
This poses a number of problems for data warehouse providers and users. The biggest is the fact that while the use of dual core and quad core processors means that Moore's law (that processing capacity doubles every 18 months) continues to be met, the same is not true of disk speeds, where improvements are more of the order of 5 or 10%. Worse, it is estimated that storage requirements are doubling every year. How then, is this circle to be squared? How can organisations meet their users' demands for the extended access that they are demanding? This paper sets out to explore the possibilities.

However, this is a white paper with a difference. It is not intended to be merely a discussion of data warehousing but an active guide on how to approach the current issues you may have with your existing data warehouse implementation. In particular, it uses a dialogue-based approach along the lines of "if you have this problem then". (See the diagram in the next section which traces the decision points in this dialogue.) Note that we are assuming that you already have a data warehouse as green field implementations would require a different set of questions. For the same reason we will not be discussing, at least in any depth, what you might do if you decide to completely replace your existing database.

Leading on from this last point, we will also not be discussing what you might do if you are using a data warehouse based on Informix, RedBrick or some other database that is no longer treated by its vendor as a primary data warehousing platform. While you might add a data mart from, say, Netezza, or you might upgrade your hardware, ultimately this will be no more than a stopgap: if you are having concurrency, scalability or performance issues now, with a database that is unlikely to be significantly improved in the foreseeable future, then eventually you will have to replace that system.

What we will be discussing is those environments where you have a viable potential future with your existing supplier and we will consider what you might do, depending on the issues that you face. Note that while this paper has been sponsored by Dataupia it is not limited to discussions of when that company's product might be suitable for your environment.

The following diagram summarizes the questions and recommendations that will set you on the right track for getting more out of your data warehouse.



There are four basic technical questions (with sub-questions) that need to be asked, and these are concerned with performance (how fast?), scalability (how much data?), concurrency (how many users?) and how do these compound each other (do you have both a performance and scalability problem?). Relevant non-technical questions include manageability and costs (initial and ongoing) and, in particular: are you prepared to consider a replacement system? Many companies (according to one recent survey, 75%) are not. As has been previously noted, a completely different set of questions would be needed if we wished to discuss replacement, so the following questions and answers assume that you are at least interested in retaining your investment in your existing system.

Question 1 – is your data warehouse performance good enough?

If the answer is yes go to question 2. If not:

Question 1a: can you load data into your warehouse fast enough?

Question 1b: are standard reports too slow?

Question 1c: are ad hoc, unpredictable queries (even simple ones) too slow?

Question 1d: are complex queries too slow?

Question 1e: can you deliver the results of queries (using real-time data) in (near) real-time?

If the answer to Question 1a is “yes” go to the next paragraph, otherwise if you simply can’t load fast enough to support the timeliness of queries then you will need to identify the bottleneck. If it is in the hardware you may be able to upgrade it, if it is in the data loading software you may be able to replace it, but if it is in the database you may have a problem. Given the growth in data storage requirements most vendors have plans to allow greater loader speeds. If these are adequate and have a reasonable timescale, then that may be satisfactory. If not, then the only alternative to replacing your existing system will be to augment it with an appliance from a vendor such as Dataupia. This should have minimal impact on your existing systems, as you can run existing applications without change, and the company may well be able to offer faster loading speeds (currently at around 2Tb per hour) than your current system. Note that Dataupia only currently runs in conjunction with Oracle, SQL Server and DB2.

If the answer to Question 1b is “yes” you haven’t got a performance issue you’ve got a scalability issue: go to Question 2. However, before you go there one possible remedy is to offload part of your query processing. For example, we know of one site where a data warehouse appliance is used as a front-end to a Teradata warehouse: the appliance being used to calculate aggregates and feed the results into the main warehouse, thus reducing the load on the Teradata system so that improved query performance can be attained. Such appliances are available from Netezza, DATAlegro, Greenplum, Dataupia and others.

If the answer to Question 1c or 1d is “no”, then go to the next paragraph. If the answer is “yes” then you need to bear in mind that the reason why your standard reports and queries run satisfactorily is because there are indexes defined against the relevant data to make them run faster, but for unpredictable and complex queries such indexes are often not available. If there are only a limited number of such queries that you ever run then it may be possible to define relevant indexes and other database artefacts (such as materialised views) that will help to resolve these queries. However, bear in mind that these will increase the size of the storage requirement, which impacts on scalability and that this will have knock-on effects on performance; and they will also impact on load performance. If this makes scalability into an issue or if there are too many such queries to realistically build indexes for, then you should consider running these query types on a dedicated data mart that has been based on technology that has been designed to support

these sorts of queries. Typical providers would be data warehouse appliance vendors such as Netezza, Greenplum, DATAlegro, Dataupia or column-based products such as those provided by Vertica, ParAccel or Sybase IQ. Note that these products are not limited to use as data marts (see Question 1b), even when used in conjunction with existing data warehouse products.

If the answer to Question 1e is "yes" go to Question 2. On the other hand, if you are having a problem with real-time queries then you need to bear in mind that there are three aspects to supporting such environments. The first is that you need to be able to trickle feed data into the data warehouse (though "trickle" may be a misnomer, in some environments it can represent a flood). This is not hard, at least in theory, as there are a number of providers of such a capability. Secondly, you need to update relevant records in the warehouse and, thirdly, you have to be able to run a lot of small queries (often just look-up queries) in a short space of time. It is the last two elements of this that are the most awkward. For updates and look-up you effectively need something akin to OLTP (on-line transaction processing capability) so the merchant databases (and Teradata and Sybase IQ have suitable functionality as well) will be most suitable; and the same applies to the mixed workload support that will be required to balance these requirements with the traditional functions of a data warehouse. Hopefully, your existing supplier will be working on upgrades to its workload management capabilities (IBM has just done exactly this with its recent DB2 9.5 release) so that this balance can be satisfied in that way. If not, then it may be best to consider splitting the functions of the warehouse so that the existing system caters for real-time enquiries and a separate data mart (see Questions 1c and d) is set up to support traditional warehousing requirements. In theory it would also be possible to do this the other way round, using HP NeoView (which has advanced workload management) for the real-time functions. If neither of these approaches appeals, then augmenting your existing system by adding a complementary technology such as Dataupia's could help.

Question 2 – can you store enough data and still get good enough performance?

This is the scalability issue, if the answer to question 2 is "yes" then go to Question 4. Otherwise, before we discuss scalability we must clarify it. The fact is that you can see all sorts of figures bandied about for data warehouse sizes but this is not always (or even often) about comparing apples with apples. Suppose that you have a 100Tb data warehouse. How much of that data is never queried or is not even able to be queried (because it consists of photographs, for example)? If that's 10% of your warehouse then you effectively have only a 90Tb warehouse. Now, how many indexes, materialised views and other constructs have you defined to speed up the query process? In a typical, uncompressed merchant database implementation the size of the raw database will multiply by anything between 2 and 5 times to incorporate these structures. So, if we assume 3x then that means that our 100Tb warehouse actually only stores 30Tb of data. Conversely, a data warehouse appliance or column-based database that doesn't use indexes will typically give you 100Tb of raw data for 100Tb of storage.

This situation is complicated by compression. In Oracle 11g and DB2 9.1 (and later) both Oracle and IBM offer compression, with rates of around 2 or 3 times (more sometimes), depending on the type of data to be compressed. However, bear in mind that there is an overhead involved in decompressing the data. If a lot of data is to be read then you can more than make that up by having to read less data and the compression will therefore provide improved performance. However, for small tables it will better not to compress them. In most warehouses this is not an issue. However, what will be an issue is whether you compress indexes: if the indexes are not big enough to merit compression then that very significant part of the total storage requirement will not be affected by compression. The other thing to bear in mind is that data warehouse appliance vendors are increasingly introducing compression as well, while the column-based suppliers can do compression significantly better than anybody else (because it is easier to compress against a single datatype).

Anyway, if the answer to Question 2 is no, then the first thing that you might consider, if you are an Oracle or DB2 user, is to upgrade to the latest versions of those databases and then implement compression. However, this is probably only a short-term measure. Given that data is doubling ever year, and with increased compliance requirements and simply for good business reasons, it is likely that any benefit derived from compression will be eaten up very quickly. Moreover, it is unlikely that further developments in compression algorithms will do more than provide incremental enhancements in storage saving and/or performance in the future.

For Microsoft users or for Oracle or DB2 users for whom compression is insufficient (or has already been used up) then there remain two alternatives, assuming that you have exhausted hardware options or, in the case of Oracle, the deployment of RAC (real application clusters). The first is that you could adopt a federated warehouse strategy, deploying one or more further instances of your existing database, in effect splitting the warehouse into multiple subsets; secondly, you could extend the warehouse through a product such as the Dataupia Satori Server. The latter option is likely to be much easier to implement and administer.

Question 3 – do you have both a performance and a scalability issue?

If the answer to this question is “no” then go to Question 4; conversely, if it is “yes” then we would recommend that you resolve the scalability issue first, as outlined in the response to Question 2. The reason for this is that it may be scalability that is causing the performance problem in the first place. It is perfectly possible from a hardware perspective to hang hundreds of terabytes of data onto any database that you like—MySQL, for example—but try to do any significant work on it and it will either perform like a dog or not at all. Where is the problem for this poor performance? The answer is in the lack of scalability of the database itself. So, the first thing to address is this issue: compress the data if you can or otherwise look at the likes of Dataupia. Only if this fails to resolve all of your performance issues should you go back to Question 1. Note that once you go beyond compressing the data you are likely to be running a proof of concept so you do not need to decide to augment your existing system with Dataupia and then consider whether to add an appliance-based data mart: you can more or less parallelise these decisions.

Question 4 – can you support all of your user demands (and how do you know)?

If you can genuinely answer “yes” to this question (in other words you have checked) then go to the next section. Before discussing what to do if you answered “no” we need to make it clear here what we mean by “users” since we do not necessarily mean people. In particular, there is a trend towards embedding query capabilities into business processes and applications, so-called operational BI. Here the application, process or service is the user and information is often required in real-time. We therefore need to ask some subsidiary questions, the first of which is:

Question 4a: is this because of growth in conventional users (in other words you are deploying BI more widely across the enterprise)?

If the answer to Question 4a is “no” then go to Question 4b; on the other hand, if the answer is yes then you almost certainly have a performance problem: you are simply not processing queries fast enough to support the increased number of users. If you also have a scalability problem go back to Question 3 but, if not, go back to Question 1.

Question 4b: and/or is it because of growth in operational BI?

If the answer to Question 4b is “no” then go to the next section; otherwise you either have a conventional performance and/or scalability issue (go back to Questions 1 or 3, as appropriate) or you have an issue with the workload management capabilities of your existing system. In the latter case, this means that the database is struggling to meet the combined requirements of the many short queries: go to Question 1e.

Supplementary Issues

There are, of course, a number of supplementary issues that you may wish to consider. For example, you may be concerned about floor space in your data centre, or power and cooling requirements. Or you may have issues with the management and administration of your existing system. While we are not going to discuss these in detail it is worth pointing out that both complementary appliances (Dataupia) and stand-alone warehouse appliances (Netezza, Greenplum et al) tend to perform better on these scores than conventional systems, sometimes by a substantial margin. Otherwise, compression will help, at least in the short term, by reducing disk requirements (and, thereby, costs) and replacement hardware may also be a possibility, at least in terms of saving floor space.

There is no quick and easy answer to what you should do with your data warehouse. For some companies a replacement product may be the best way forward, in which case all bets are off. However, this paper does not delve into such areas but considers the potential options facing you on the assumption that you wish to retain your investment in your existing systems. Put simply, the options then are to upgrade your existing system (either through new releases of the software and/or new hardware), to augment your existing system through the use of complementary technology provided by a company such as Dataupia, or to off-load part of your processing either into a data mart or for a particular purpose such as calculating aggregates. In this last case you may wish to consider a data warehouse appliance but you might also be able to accomplish the same thing using a further instance of your main database; though this is more likely in the case of data marts than for a special purpose like calculating aggregates on the fly.

The options that will be most suitable for your organisation will, of course, depend on your circumstances. We have attempted here to highlight the main issues that you are likely to face and what are the most likely approaches available that may be able to fix these.

Bloor Research has spent the last decade developing what is recognised as Europe's leading independent IT research organisation. With its core research activities underpinning a range of services, from research and consulting to events and publishing, Bloor Research is committed to turning knowledge into client value across all of its products and engagements. Our objectives are:

- Save clients' time by providing comparison and analysis that is clear and succinct.
- Update clients' expertise, enabling them to have a clear understanding of IT issues and facts and validate existing technology strategies.
- Bring an independent perspective, minimising the inherent risks of product selection and decision-making.
- Communicate our visionary perspective of the future of IT.

Founded in 1989, Bloor Research is one of the world's leading IT research, analysis and consultancy organisations—distributing research and analysis to IT user and vendor organisations throughout the world via online subscriptions, tailored research services and consultancy projects.



Philip Howard
Research Director - Data

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 what is now P3ST (Wordsmiths) Ltd 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. His practice area encompasses anything to do with data and content and he has five further analysts working with him in this area. While maintaining an overview of the whole space Philip himself specialises in databases, data management, data integration, data quality, data federation, master data management, data governance and data warehousing. He also has an interest in event stream/complex event processing.

In addition to the numerous reports Philip has written on behalf of Bloor Research, Philip also contributes regularly to www.IT-Director.com and www.IT-Analysis.com and was previously the 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 published a number of reports published by companies such as CMI and The Financial Times.

Away from work, Philip's primary leisure activities are canal boats, skiing, playing Bridge (at which he is a Life Master) and walking the dog.

Copyright & disclaimer

This document is copyright © 2007 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.



Suite 4, Town Hall,
86 Watling Street East
TOWCESTER,
Northamptonshire,
NN12 6BS, United Kingdom

Tel: +44 (0)870 345 9911
Fax: +44 (0)870 345 9922
Web: www.bloor-research.com
email: info@bloor-research.com