

Market Guide



Analytic Warehousing

A Market Guide by Bloor Research
Author : Philip Howard
Publish date : September 2009

In this paper we have attempted to give an overview of the main concerns in the market and those vendors that currently play in it.

Philip Howard

Executive summary

We have long been concerned that the data warehousing market is treated as if it was homogeneous, with all vendors potentially tackling all warehousing issues. This is very far from the truth: for example, some products are specifically focused on supporting analytics while others are targeted at enterprise data warehouses. Similarly, some suppliers target the low end of the market, measured in gigabytes or perhaps a few terabytes, while others concentrate on implementations with tens or hundreds of terabytes, or even petabytes. Trying to compare all of these offerings within a single paradigm makes no sense. We have therefore elected to break the market down in terms of both functionality (analytic marts and warehouses on the one hand, and enterprise data warehouses on the other) and, in the case of the former: scale (small, medium and large). We have not distinguished in scale terms across enterprise data warehouses because the two tend to go hand in hand. Each of these four sub-market examinations is available as

an individual paper in its own right, but each refers back to this one as its foundation document in which we discuss market requirements and vendors.

Despite the above we appreciate that readers will wish to see some sort of diagram illustrating the entire vendor landscape and we therefore provide this in Figure 1. The vertical axis shows numbers of installations for each vendor on a logarithmic scale (note that numbers exceeding 10,000 are not intended to be accurate) and it is organised horizontally by architecture (whether column or row-based). The size of the bubbles representing each vendor indicates their size, stability and geographic presence on a scale of 1 to 5.

Note that SAS has been excluded from this diagram as the company has failed to provide us with figures as to the number of customers using SPD Server.

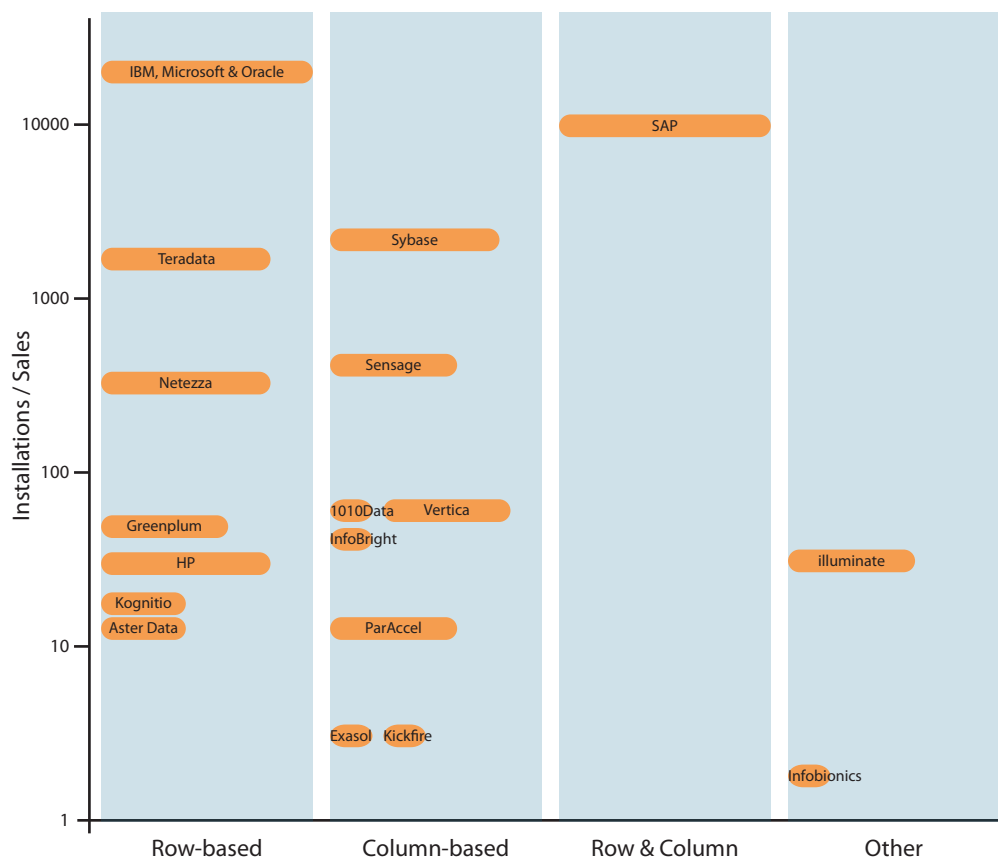


Figure 1: The vendor landscape

Introduction

You can use any database as a data warehouse or to support analytics or business intelligence. However, if there is any substantial requirement for complex or ad hoc analytics then general-purpose databases without specialised facilities will fail to give adequate performance. For this reason this report is focused exclusively on products and vendors that do have such specialised functionality and it does not include a discussion of offerings that are primarily aimed at transactional markets, nor those that do not have substantial analytic capability. In addition to a focus on analytics we are also concerned only with general-purpose products rather than those aimed at niche or hybrid markets.

Nevertheless, we still have 20 vendor's products to evaluate. However, it should be clear that they do not all compete with one another: there are multiple sub-markets within the data warehousing domain and, in addition to general requirements, we will also need to consider the particular markets that the various products are targeted at. However, as a general principle we are concerned in this report with anything involving substantial amounts of (complex and ad hoc) analytic processing as opposed to environments where query and report processing is well-defined in advance. For requirements in the latter category, standard merchant or open source databases should prove sufficient for most purposes, though there may still be advantages in moving to one of the more specialised products discussed in this paper. However, our focus is on environments that include heavy duty processing of data, which is why we have called this paper "Analytic Warehousing".

“ our focus is on environments that include heavy duty processing of data... ”

What do you want from a warehouse?

We are here using 'warehouse' as shorthand for all the categories of interest just described and in this section we will briefly outline all of the major characteristics that you might want to look for in an analytic warehousing product. Note that we do not pretend that the following list is exhaustive but it does cover all of the major issues that you might want to consider. They do not appear in any particular order.

Performance

This is arguably the most important characteristic of all. Does your warehouse give you responses to your queries fast enough to suit your business? Do they meet your service level agreements? These are not necessarily the same thing: it is the former that is fundamental. Note that if all your queries do meet that need then you don't need a warehousing product that does it faster: performance for its own sake has limited value.

There is really only one way to test for performance and that is to run your own queries against your own data and measure the results. Be sure that this proof of concept includes a representative mix of both queries and load processes running concurrently. Note that TPC-H benchmarks published by the various vendors may provide a useful indication of performance but that they do not do more than that: the test schema for TPC-H will not represent your requirements and, as many vendors do not bother with TPC-H, they do not represent a particularly useful comparative tool. Moreover, new hardware and software releases mean that results continually leapfrog one another so, at best, a benchmark result is a snapshot of a particular point in time.

Apart from query performance there is also the question of load performance (and, for that matter, backup and export performance). Can you load data fast enough into your warehouse? If you produce six terabytes of data to load every day, is the database fast enough to load this within whatever batch window you have available? Note that databases that do not use indexes (see later) potentially have an advantage here, because they do not need to update indexes as well as load the data itself, so the key metric is not how fast you load the data but how long it takes before you can query it, taking into account the construction of all necessary indexes, summary tables, materialised views and the like. An interesting corollary to the issue of batch load speeds is that

the better the database supports real-time loading then the less necessary batch load speeds become, because you can trickle feed the data into the warehouse on a continuous basis. Thus six terabytes per day represents a heavy requirement if your batch load window is only one hour, but if you trickle feed the data over 24 hours then you only need to load at 250 Gb per hour. However, this doesn't mean that you don't need efficient bulk loading capability because you have to load the warehouse in the first place.

Scalability

There are multiple aspects to scalability. The first is whether you can store all the data you need to, both now and for the future. You need to consider the raw data that you want to store, how much additional space you need for indexes, materialised views and other constructs and then how much more capacity you require for disk mirroring to support backup, high availability and so forth. Next you have to apply compression, which may apply only to data or to data and indexes and, in any case, will vary by data type and by vendor (in both cases, potentially significantly). Put that all together and you know how much disk capacity you need today and you can make some sort of forecast as to how much you will need tomorrow. However, you will have to be careful to compare apples with apples when you are talking with vendors: what do the figures they are quoting refer to: raw data, with indexes, with or without mirroring, compressed? You may get quotes of warehouse sizes that actually refer to distributed environments in which there are multiple instances of the database. Also bear in mind that many suppliers today are quoting their ability to support petabyte sized warehouses. In most cases this is purely theoretical.

The second issue about scalability is memory (or, perhaps more accurately, data processing). Systems that make extensive use of in-memory capability in order to get performance may require very large amounts of memory, which may mean a much greater number of processors. Determining the balancing act between in-memory and off-disk processing on the one hand, and suitable hardware provisioning (and cost) on the other, may be a complex exercise.

Thirdly, can you support all the users you need to, both now and in the future, which is essentially a concurrency issue? Note that

What do you want from a warehouse?

users here may not actually be people but may be business processes or operational functions that have queries or look-ups embedded within them. Note that there is a big difference between the number of users that can be supported based on x number of queries per day, of y average duration per user; and the number of simultaneous queries that can be supported. Both answers need to be able to meet your business needs.

Finally, you could also argue that scalability also applies to loading, exporting and backing up data. As you grow, your need to perform these functions will grow too, and your solution needs to scale in line with that growth.

Architecture

A priori it does not matter what sort of architecture a product has providing it meets your performance, scalability and functional requirements. Arguments about architectures may be of interest to theorists but are irrelevant when it comes to deciding what is best for your business.

That said, there is a major rift within the warehousing community between those vendors that offer a row-based approach and those that use columns, though there are a couple of vendors that use neither of these approaches and one (in the future two) vendors that use both.

Columns are good for whole table scans, reducing the amount of data that needs to be read (you only read the columns you are actually interested in—though if the number of columns is large this may not make much difference), removing or reducing the number of indexes you need thereby reducing tuning requirements, improving bulk load speed (you don't have indexes to update, though this advantage may dissipate if you have separate read and write stores), and can generally offer better compression thus reducing storage requirements.

On the other hand rows are better, amongst other things, for single (typically real-time) inserts and updates, when you are looking up a single row or when a query involves only a few rows. Rows are also amenable to the use of things like join indexes, though some column-based products do have these.

Analytic queries typically involve large volumes of data and therefore columns are well suited to these environments. However, if there is a substantial amount of real-time activity or look-ups, for example if the warehouse is being used to host a master data management system or to support operational BI, then a row-based approach may be indicated. In particular, transactional capability and ACID compliance are important here.

As may be imagined, row-based vendors want to get over the whole table scan issue and to reduce the amount of data that needs to be read, while column-based vendors want to provide a way around their disadvantages. To date we would say that the former (at least some of them) have been more successful than the latter. There are several row-based suppliers that have addressed the challenges outlined on their side, either by using specialist techniques whilst eschewing indexes or, at the other extreme, by implementing such things as join indexes or parallel join capability. Conversely, in the columnar camp the only issue that has really been addressed (and then by some vendors only) is the issue of real-time updates. Note that when we say that row-based vendors have been "more successful" we mean that literally: this still doesn't mean that compression is as good in row-based systems as in the best column-based ones, for example.

For what it is worth, of the vendors in this report eight use a column-based architecture (and a ninth is in the offing), ten use a row-based approach, one uses both (a second will be joining them) and two use something else; four of the vendors make extensive use of in-memory processing (and this number is increasing); two suppliers use or will use a shared everything approach, two offer a shared disk approach, one offers a choice between shared disk and shared nothing, and sixteen prefer shared nothing; five offer SMP (symmetric multi-processing) and all the others prefer MPP (massively parallel processing) except two that use an asymmetric approach with an SMP system on the front-end and MPP close to the disks—two of the SMP vendors will be offering an MPP architecture in due course. In addition, two vendors require some specialist hardware: a SQL chip and an FPGA (field programmable gate array) respectively. Some suppliers recommend blades, some do not. Most vendors use direct attached storage but not all, some support storage area networks (some better than others). What does all this tell you? Not a lot.

What do you want from a warehouse?

Compression

Compression can make a big difference to the amount of storage you require and therefore the footprint, price and running costs of your warehouse. Compression can also improve performance, because you can read more data in a single I/O and, in an MPP architecture where you moving data from node to node (for example, when doing joins or performing complex operations of various types), then the use of compression can reduce network bottlenecks. However, compression is an area where architecture makes a difference. In theory, columnar databases are easier to compress than row-based ones, because you can apply optimal algorithms by datatype. However, this is dependent on the implementation of compression and some companies are better at this than others. Typically, compression rates offered are in the three to five times range but some vendors can get better much better rates than this. If a vendor uses indexes then it will be useful if they can offer index compression, which not all vendors do. Similarly, the ability to compress backup data and temporary tables will be an advantage.

Indexes

Some vendors offer lots of indexes. These take up more disk space and slow down load speeds (because you have to update the index as well as the data). They also require administration and tuning even though much of this process may be automated. On the other hand, indexes add capability (for example, text indexes) and speed up performance against those tables (or multiple tables if using join indexes) that are suitably indexed; but this pre-supposes that you know what queries you are going to ask, which is often not the case, especially when it comes to complex analytics.

There are various approaches to overcoming the issue of querying non-indexed data: one is to index everything, but you need a special kind of (correlation or associative) database to do this; another is to use what can be best described as an anti-index whereby the database is told where the data is not, thereby reducing the required amount of I/O; a third would be to build indexes on the fly; and a fourth is to use a column-based approach where each column acts as an index. Many column-based databases, in particular, do not support indexes

at all and the same is essentially true for a number of the newer row-based products on the market. However, in the latter case, being row-based, it is typical that these vendors will include look-up indexes or their equivalent.

It is also worth briefly commenting on optimisers. While a significant number of vendors have quite advanced specialised capabilities in this area there are also a number that don't. In particular, for analytics a major concern is with the execution of correlated sub-queries, which are sub-queries that use values from the outer query in their WHERE clauses. A number of companies have only limited facilities for handling these in an efficient manner.

Mixed query workloads

There is always a trade off between shorter and longer running processes in any environment, which is complicated further by the need to meet SLAs. Facilities such as scheduling, query governing and prioritisation have historically been provided to meet this balance of requirements and, for analytic data marts and traditional data warehouses, they are probably more or less sufficient. However, for the more advanced enterprise data warehouses that we see today, which include support for master data management and operational BI as well as traditional analytics, BI and data mining, the balance to be struck is much more complex and requires additional capability. In particular, you need the ability to set a floor or ceiling on the resources to be allocated to a query. For example, there is no point assigning parallel processing capabilities to a look-up that is only going to read a single row. Moreover, you need to be able to do things like assigning memory and threads dynamically, taking account of the current workload. You will also need transactional support with full ACID capabilities for real-time updates and inserts in a master data management or similar environment. Relevant interfaces to be able to monitor workload and performance will be useful. In general, mixed query workload management is in its early days and we can expect improvements in this area over the coming years: capability in this area is typically a reflection of the maturity of the underlying product.

What do you want from a warehouse?

Key areas to consider are ease of use in creating and updating mixed workload policies, the granularity of these, and the ability to pre-define, as well as dynamically re-prioritise, resources. The last of these is fundamental for mission-critical environments where new workloads may mean that adjustments to CPU and disk allocations are necessary to meet SLAs.

Encryption

Encryption is important not just for data protection purposes per se, but also in environments where you are hosting data for multiple customers or providing warehousing services to multiple user communities. In any of these cases it is important to be able to ensure that one customer cannot access another's data. Of course, part of providing this security involves user privileges, access control, passwords and so forth but encryption (with different keys/algorithms used for each customer) is an important back-up provision so that even if the wrong data is accessed it cannot be read.

It is preferable to be able to encrypt data by column rather than at the database or table level. The main reason for this is because it is expensive, in performance terms, to encrypt data, because you have to decrypt it during the query process. Therefore it will be most beneficial to limit encryption to the data that actually needs encryption, and this can be most efficiently done on a column basis. A number of row-based database vendors provide column-based encryption, so this is not a differentiator for column-based approaches.

In-database mining

Historically, data mining has been achieved by the extraction of data from the data warehouse into the mining tool, where the data is then processed. This is slow because of the volumes of data to be extracted. One way to compensate for this is to sample the data rather than work with it in its entirety but this may diminish the accuracy of the results and may miss outliers. The solution is to process the mining within the database itself, and the leading data mining vendors, as well as some open source suppliers, are working to provide this sort of functionality. At present there are only a few warehousing vendors that have this sort of capability, which provides these suppliers with definite advantages over their competitors.

Note that those vendors that combine SQL with MapReduce arguably also provide a form of in-database mining.

In addition to embedding general-purpose data mining capabilities within the database it is also useful to embed other specialised analytic functions (spatial analytics is a good example), either by the vendor doing this directly or by providing user-defined functions, or something similar, that will allow relevant partners to embed this sort of functionality.

Unstructured data

There is a growing demand to be able to analyse unstructured data, either on its own or in conjunction with structured analytics or data mining. For example, text mining, XML analytics, video and audio analysis, and so on are all growing areas of interest. The number of vendors that support even one of these is limited but there are suppliers with text indexing and/or with partners that have built functionality for things like face recognition.

Perhaps the most significant recent warehousing development that supports unstructured data is MapReduce. This is described in the accompanying box. Note that the architecture is essentially similar to a number of warehousing products.

MapReduce is a framework for computing certain kinds of distributable problems across multiple nodes. It consists of two steps:

The Map step, where a master node takes the query, sub-divides it, and distributes the sub-queries to worker nodes, which may themselves further sub-divide the sub-queries. Each worker node then processes that sub-query and passes the answer back to the master.

The Reduce step, where the master node combines the results passed to them by the worker nodes to formulate the final answer.

The advantage of MapReduce is that it allows distributed processing of both the map and reduction operations. Provided that mapping operations are independent, all maps can be performed in parallel. Similarly, you can perform reduce in parallel provided that all outputs of the map operation that share the same key are presented to the same reducer, at the same time.

As a framework, MapReduce supports multiple programming approaches, including Python, Perl, Java, C, C++, R, C# and others.

What do you want from a warehouse?

While most vendors do not yet support MapReduce at all there are four main approaches to MapReduce amongst those that do: some vendors support something similar, which isn't actually MapReduce; some support it but haven't done anything with it; some have integrated it directly with SQL so that you can combine the two environments in a single query, in much the same way that some vendors allow you to combine XML functions into SQL; and one company has indirectly linked SQL and MapReduce. This last is implemented via Hadoop, which is an open source Java framework that supports MapReduce. Here, you pass the results of an SQL query to Hadoop or vice versa rather than attempting a close coupling. The argument for the last of these approaches is that MapReduce and SQL developers are not usually the same people, at least at the present time.

High availability

Data warehouses and marts are increasingly mission critical, not just when they are supporting functions like master data management but also for operational BI, business processes that need to query the warehouse, and so on. This means that simply having backup and recovery, or even no single point of failure, is not enough. You also require automated failover and the ability to update or patch the software without having to take the whole system down so, for example, you would like the ability to take down an individual node within a cluster for these purposes as well simply because you want to add or remove a node from the cluster. For the same reason you would also want back-ups to be online and not require any planned outage. Even better, one would like support (available from some vendors) for online restoration and recovery, while queries are still running.

Where recovery is not online it is important to minimise downtime. It is not enough, for example, to have automated failover to a replica if a server fails. Traditional approaches take a long time to restore the copies/replicas by requiring time-consuming copies of all the data. A better approach is to look at the 'delta' of data that has changed, which results in copying much less data in order to restore the mirrored copies, thus reducing downtime.

With respect to disks, most suppliers use direct attached storage (DAS) rather than a storage area network (SAN), because the former offers

better price/performance. However, a SAN offers more advanced abilities when it comes to high availability. Typically you therefore have a choice between these approaches, whereas what you would really like is a hybrid approach that aims to offer a maximised combination of performance and availability. Vendors are now starting to introduce such capability.

One other requirement for disks, appropriate to some environments, is support for worm drives or other read-only storage that is required for data that may be needed for evidentiary purposes.

Implementation

The introduction of appliance technology a few years ago has led everyone to recognise the importance of being able to install and implement a warehouse in the easiest fashion possible in the shortest amount of time. As a result, even vendors that do not offer appliances per se have been busy making their implementation process as simple as possible.

However, implementation isn't just about switching on the hardware, installing the software and loading the data. A major concern is how you structure the warehouse, particularly for enterprise data warehouses rather than analytic databases. For these you need a (customisable) data model. The leading vendors tend to directly provide these for a number of vertical sectors, while other companies partner with third party suppliers of these models. An alternative is to use a product from a vendor such as Kalido or BIReady that will help to rapidly construct the logical structure of the warehouse.

Integration

Data models are not the only things you may require from third parties. It is also likely that you will want data integration, data quality, business intelligence and perhaps data mining, data federation or master data management applications, amongst others. In some cases these may be provided by the same vendor but, in any case, the degree of integration between the warehouse product and the third party software should be checked. This is because different products may work in different ways. For example, MicroStrategy makes extensive use of temporary tables in determining query results. If you have a warehousing product that does not intrinsically use temporary

What do you want from a warehouse?

tables very much or does not use them well then additional integration work will be required to make that warehouse perform efficiently with MicroStrategy. The same might apply, in different ways, to integrations with any third party product. Thus it will be helpful if integration has been certified or where there are existing customers who can testify to the soundness of the integration. Note that two companies announcing that they have a 'partnership' does not necessarily mean anything other than that they have a marketing pact.

Administration

In order to get maximum performance it is common with some warehousing products that you have lots of indexes, materialised views and other constructs. Historically, the tuning needed to maintain these was onerous and represented an advantage when compared to index-less products from competitive vendors. However, autonomic functions in the leading databases means that this is no longer much of an issue. Nevertheless, this doesn't mean that the issue has gone away: scheduling, mixed query workload management and other functions still require administration regardless of the supplier. Perhaps the one area where there remains a divide is with respect to aggregates. The management of pre-aggregation is arguably the most time-consuming administration function if you ignore tuning, so products that are fast enough to allow you to calculate aggregates on the fly have a significant advantage over those that don't.

It is also worth noting that there may be system and operational administration implications as opposed to purely database concerns. Things like scaling (adding or removing nodes), replication restoration, adding new network interface controllers and so on, should be supported from the same management console as database administration functions.

Delivery

The most typical ways that warehousing products are delivered are as an appliance, as software or pre-installed. In addition, there are service providers who will host a warehouse or mart for you and they may even provide data mining, forensics or other services. An example of the latter is KPMG, which will host short-term (six weeks to eighteen months) marts on demand. In addition, there are some vendors of warehousing products offering this service (DaaS: data warehousing as a service) directly.

Another way to implement short term data marts is via either private or public clouds and a number of vendors support these, either directly or indirectly. Also, there is the possibility of supporting virtualisation, which could be effective when used in conjunction with a private cloud, though there will typically be a performance downside.

Cost

While there is little point in discussing the importance of licensing and maintenance costs, bear in mind also that physical running costs (footprint, power, cooling and so on), ongoing administrative requirements, and consulting and other services (whether provided by the vendor or a third party) all contribute to the cost of a system. The rapidity with which you can deploy a solution may also have relevance in your decision making process.

Sub-markets

The points made in the previous section apply to some environments more than others and we need to be clear what those different environments are. We can broadly categorise them as follows:

Data Marts: typically a subject-specific database for intensive querying and analytics—if complex analytics and unpredictable queries (including data mining) predominate, then these are sometimes referred to as analytic data marts. Where this element is less important then the need may well be fulfilled by a standard database, in which case it falls outside the scope of this report.

- **Edge appliances:** this is a special purpose data mart, where an application has been built on top of the warehouse or appliance—a classic example would be for real-time pricing in the telco sector. Note that some of the vendors, excluded from this report for not being general-purpose, offer products in this area. There is nothing in particular to distinguish an edge appliance from a data mart except for packaging and the fact that there may be multiple vendors involved.
- There is also a software equivalent of this which we might describe as for embedded analytic databases. The main requirement for these, apart from performance, is the same as for any other embedded database, which is that it should have a small footprint and not require administration. The companies targeting this market tend to be those with lower scalability.
- **Analytic Data Warehouse (ADW):** this is what we have been used to calling an enterprise data warehouse but that name has now been usurped (see next). It supports reporting, BI/OLAP and a substantial amount of analytics and data mining.
- **Enterprise Data Warehouse (EDW),** sometimes described as EDW 2.0, this is a superset of ADW that includes support for master data management, operational BI and other functions. This changes the workload on the system (lots of short running and look-up queries as well as longer running analytics).

In this paper we are focused specifically on environments that make extensive use of analytics. As it happens, there is not a lot of difference between the requirements for an ADW and a data mart, so from here on we will assume that these are concatenated and our emphasis is therefore on analytic data marts and warehouses on the one hand and EDWs on the other. Note that this is not necessarily a difference of scale: it is entirely possible, indeed it is quite common, that a medium sized company struggling with a home-grown MySQL-based warehouse with just a few hundred of gigabytes of data will be looking for functionality that is entirely similar to that which is required by a large enterprise in a full enterprise data warehouse. While they perhaps will not need to support MDM, they are likely to require look-ups and have frequent queries running against small numbers of rows. So they will require at least a modicum of mixed query workload capability even if not in the same detail as some larger enterprises.

Of course, the other determining factor is scale. If a vendor doesn't scale below 10Tb and you have 500Gb then it is unlikely to be suitable for you on cost grounds. Similarly, if you need hundreds of terabytes and a supplier is limited to supporting 20Tb then that is probably not much use either.

In this report we have therefore broken down both our analysis by whether a product is suitable for use as an analytic data mart or warehouse as opposed to an enterprise data warehouse. We have further split the first of these categories into three scalability buckets: up to 5Tb, 5–50Tb and 50Tb+.

Vendors

1010 Data

Product:	1010data version 5.
Architecture:	column-based, shared-nothing grid.
Offering:	SaaS only (service can be in customer's data centre).
Customers:	85.
Geography:	North America.
Scalability:	100Gb to 10Tb raw data, should scale (linearly) much larger.
Target sectors:	primarily financial services, followed by retail and consumer packaged goods; also healthcare & pharmaceuticals; in theory anyone with big data
Functionality:	primarily analytic data mart, supports look-up indexes and has mixed query workload capability so is suitable for EDW.
Status:	successful in financial markets, growing elsewhere.
Strengths:	typical benefits of columnar approach, support for time series, very fast load speeds, SaaS.
Weaknesses:	little known outside financial sector, doesn't support real-time updates
Comment:	a leading contender for smaller analytic requirements where a SaaS-based solution is preferred. May be suitable for larger environments but yet to be proved.

Aster Data

Product:	Aster nCluster Database and nCluster Cloud Edition, version 3.0.1.
Architecture:	row-based, massively parallel, shared-nothing with Queen, Worker and Loader server groups.
Offering:	software only, can be provided as appliance or via cloud.
Customers:	double digits.
Geography:	North America and UK (Europe).
Scalability:	100Gb to (claimed) petabytes of raw data, largest customer 270Tb.
Target sectors:	particularly social networking and other online (Web 2.0), also finance, insurance and communications.
Functionality:	primarily for analytics requiring MapReduce.
Status:	new boy, market leader for MapReduce support.
Strengths:	in-database MapReduce closely integrated with SQL, has transactional capabilities.
Weaknesses:	likely to require more disk space than columnar approach, no support for change data capture, no support for encryption.
Comment:	product of choice for integrated SQL/MapReduce functions. Proven ability for large-scale deployments.

EXASOL

Product:	EXASolution version 3.0.
Architecture:	column-based, in-memory hybrid, shared-nothing massively parallel cluster.
Offering:	software or as appliance.
Customers:	5.
Geography:	Germany, Austria and Switzerland plus Japan (where Exasol is resold by Hitachi Systems and Services).
Scalability:	100Gb to 50Tb.
Target sectors:	any.
Functionality:	analytic data marts and warehouses in general; supports look-up (and join) indexes so suitable for EDWs.
Status:	relatively small but only home-grown German vendor (apart from SAP).
Strengths:	all advantages of columnar approach plus in-memory giving performance boost, automatic creation and maintenance for join indexes, specialised version of Linux provides extra cluster capability for high availability, German supplier in home markets, full ACID capabilities for transactions.
Weaknesses:	no support for encryption.
Comment:	a leading solution for smaller environments. May be suitable for larger deployments but yet to be proved. German origin provides significant benefit in German-speaking countries.

Vendors

Greenplum

Product:	Greenplum Database version 3.3.
Architecture:	open source SMP version available for free download, otherwise shared-nothing, row-based MPP.
Offering:	software, private or public cloud.
Customers:	70+.
Geography:	offices in Australia, South-East Asia and UK as well as USA.
Scalability:	claims from low Tb up to petabytes, sweet spot probably up to 50Tb.
Target sectors:	all.
Functionality:	Data marts and some EDWs.
Status:	a year ago was probably second to Netezza as next generation leading vendor but now being overhauled.
Strengths:	open source option, supports MapReduce in conjunction with SQL, strong cloud support.
Weaknesses:	mixed query workload capability could be more extensive, no data encryption, row-based approach will require additional disk storage and administration.
Comment:	good all-round offering but seems to be losing market momentum.

HP

Product:	Neoview version 2.4.
Architecture:	row-based, massively parallel shared-nothing.
Offering:	appliance.
Customers:	25–35 customers (50–60 installed).
Geography:	worldwide.
Scalability:	10–144Tb (3Tb minimum).
Target sectors:	any, especially financial services, retail, manufacturing, health and life sciences, communications/media/entertainment, public sector.
Functionality:	primarily targeted at EDW.
Status:	major vendor but yet to make major impact in data warehousing terms.
Strengths:	extensive support for mixed query workloads, high availability, advanced features for predicate processing.
Weaknesses:	no support for encryption, no support for data compression, which will mean greater storage requirements, priced to compete with high-end data warehouses and therefore relatively expensive compared to data mart vendors.
Comment:	so far, has failed as much of an impression on the market as we (and HP) would like.

IBM

Product:	InfoSphere Warehouse and InfoSphere Balanced Warehouse, both based on DB2 version 9.7.
Architecture:	row-based in-memory hybrid, shared nothing massively parallel.
Offering:	software, may be pre-installed; cloud support.
Customers:	lots.
Geography:	worldwide.
Scalability:	could be less than 1Tb, have customers with hundreds of terabytes, potentially support petabytes.
Target sectors:	all.
Functionality:	primarily EDW.
Status:	one of the market leaders: behind Oracle in terms of absolute numbers but ahead of it in terms of large scale implementations.
Strengths:	transaction processing support, extensive compression (not just data), XML analytics (only vendor), in-database mining, support for MapReduce.
Weaknesses:	will take up more disk space than some column-based competitors, probably can't compete for analytic data marts (though future release of IBM Smart Analytic System may change this), ongoing database tuning for non-integrated offerings.
Comment:	competitive throughout and a leading vendor for larger-scale systems. Making a determined effort to counter the claims of new entrants to the market, with more to come.

Vendors

illuminate Solutions

Product:	iLuminate version 4.0.
Architecture:	correlation, shared everything.
Offering:	software.
Customers:	52 (39 on maintenance).
Geography:	North, South and Central America and Europe (Spain); actively recruiting partners.
Scalability:	Up to 10Tb, theoretically can scale to hundreds of terabytes.
Target sectors:	any.
Functionality:	data marts.
Status:	actually has more implementations than many better known competitors, primarily in Spanish speaking countries.
Strengths:	active in geographies (Spanish speaking) not addressed by many other vendors, theoretically should be very fast.
Weaknesses:	does not have full transactional support, have to explain technology.
Comment:	major contender within Spanish-speaking market, particularly for smaller systems, but lacks some of the features of its more established competitors.

Infobionics

Product:	Infobionics Knowledge Server.
Architecture:	cellular architecture leveraging associative technology.
Offering:	software.
Customers:	few.
Geography:	US only.
Scalability:	limited.
Target sectors:	government, scientific, research, anywhere where data relationships are unknown or variable.
Functionality:	data marts.
Status:	new.
Strengths:	dynamic data model, XML support, supports semantically rich data.
Weaknesses:	have to sell the technology before the product, no track record.
Comment:	interesting technology, particularly of interest to the scientific and research communities—is not included in detailed comparative papers as the company failed to respond to our requests for further information.

Infobright

Product:	Infobright Enterprise and Community Editions, version 3.2.
Architecture:	column-based but stored in 'data packs', SMP today; shared-everything multi-server planned end 2009.
Offering:	software (open source and licensed enterprise editions), cloud support.
Customers:	70+.
Geography:	North America and Europe (Poland—where underlying technology was developed, plus UK office), have partners worldwide.
Scalability:	200Gb to 50Tb (assuming average 10:1 compression), will increase once multi-server version released.
Target sectors:	Web 2.0 users, online companies, mid-tier enterprises, emerging businesses, ISVs and SaaS providers (as OEMs).
Functionality:	primarily data marts and embedded analytic databases, perhaps EDW for SMEs.
Status:	less well-known than some of its competitors but clearly making headway.
Strengths:	open source option, all advantages of columns but should be even faster because of architecture, high compression rate, supports look-up indexes, looks like MySQL to developers.
Weaknesses:	single server only at present (limits load speeds in particular), no encryption, limited mixed query workload capability.
Comment:	advanced columnar technology makes the company a serious contender for smaller analytic environments. It remains to be seen whether it can translate this with larger systems when its multi-server implementation is released.

Vendors

Kickfire

Product:	Kickfire Analytic Appliance, version 1.1.
Architecture:	column-based SMP with SQL chip configured as co-processor.
Offering:	appliance.
Customers:	less than 10.
Geography:	North America; active in Europe but no office.
Scalability:	focuses on 500Gb to 3Tb but will scale to 10Tb, can add further external storage.
Target sectors:	high tech data marts, online analytics, financial services and MySQL users.
Functionality:	Oracle data marts and my MySQL EDWs.
Status:	relatively new.
Strengths:	all advantages of columns, MySQL compatibility with migration wizard, full mixed query workload capability, can define indexes, full ACID compliance for transactions, SQL chip potentially offers price/performance advantages with low power consumption.
Weaknesses:	no support for encryption.
Comment:	will appeal to low-end analytic users, especially where they are existing MySQL users.

Kognitio

Product:	Kognitio WX ₂ analytical database, version 6.1.7.
Architecture:	row-based, MPP shared-nothing; makes extensive use of memory; commodity blades recommended for optimal performance.
Offering:	software, appliance or DaaS (data warehousing as a service).
Customers:	21 (14 installed).
Geography:	offices in UK and USA, active in Europe.
Scalability:	targets 1 to 100Tb but can scale both lower and higher.
Target sectors:	any.
Functionality:	analytic data marts and EDW.
Status:	long established (previously WhiteCross), not solely a warehousing company has steady revenue stream from other parts of business.
Strengths:	in-memory techniques improve performance and concurrency, support for mixed storage architectures.
Weaknesses:	no support for encryption, no data compression.
Comment:	serious competitor across the board. Its DaaS offering is particularly interesting.

Microsoft

Product:	SQL Server 2008 and SQL Server Project "Madison".
Architecture:	row-based, SMP but project Madison (based on technology acquired from DATAlegro) will be MPP shared-nothing. This is due during the first half of 2010. The larger environment is/will be based on a hub and spoke environment with Madison at the hub and spokes being either Madison or SQL Server.
Offering:	software, cloud capability.
Customers:	lots.
Geography:	worldwide.
Scalability:	SQL Server 2008 Enterprise and Fast Track Data Warehouse focus on 1Tb to 32Tb but have customers with as much as 55Tb. This should be extended significantly (to hundreds of terabytes) when Madison is available.
Target sectors:	all.
Functionality:	EDW and data marts but performance may be an issue for analytic data marts where requirements are complex and/or ad hoc—applies to SQL Server rather than Madison.
Status:	a market leader.
Strengths:	transaction processing support, lots of BI capability, user-defined aggregates, table-valued functions that are MapReduce-like, SQL Server can run on up to 256 cores on single SMP system.

Vendors

- Weaknesses:** supports encryption but not by column. SQL Server, as opposed to Madison, will likely have greater storage requirements than columnar approaches because of need for indexes and other constructs, with ongoing database tuning. Similarly, SQL Server is unlikely to be able to compete for analytic data marts where requirements are ad hoc and/or complex (this will change when Madison becomes available).
- Comment:** obviously competitive for smaller and medium-sized deployments. This will be extended significantly once Madison becomes available.

Netezza

- Product:** Netezza TwinFin, version 5.0.
- Architecture:** row-based, shared nothing, asymmetric MPP (SMP front-end) with snippet processors. Uses standard IBM (potentially others) blade hardware with FPGA-based sidecars (mezzanine cards) for snippet processing; note that this replaces previous proprietary hardware.
- Offering:** appliance, available in cloud via partners.
- Customers:** 200+ (450+ installations).
- Geography:** North America, EMEA and South-East Asia.
- Scalability:** 2Tb+, will scale into hundreds of terabytes, theoretical maximum 1.3Pb.
- Target sectors:** all.
- Functionality:** data marts and EDWs.
- Status:** leading vendor amongst the new generation of suppliers.
- Strengths:** very fast, minimal administration, special spatial analytic offering, in-database data mining, MapReduce support shortly, competitive pricing that is independent of disk storage, ACID compliance, compressed data cache to support operational workloads.
- Weaknesses:** currently does not have optimal capability for mixed query workload management but is better than many and has significant enhancements planned.
- Comment:** the leader or a leader in all analytic markets. Is making significant strides in adding mixed query workload capability that will make it a serious contender for EDWs.

Oracle

- Product:** HP Oracle Database Machine; you can implement Oracle Database 11g on its own but the former is the company's primary offering.
- Architecture:** Oracle Database 11g is a shared-disk system; Exadata (the storage subsystem in the HP Oracle Database Machine) is MPP based.
- Offering:** Oracle Database 11g is software, the HP Oracle Database Machine is appliance-like.
- Customers:** lots.
- Geography:** worldwide.
- Scalability:** Oracle Database 11g only from 500Gb, Database Machine from 2Tb scaling to (at least) hundreds of terabytes.
- Target sectors:** all.
- Functionality:** EDW or EDW with data marts.
- Status:** market leader.
- Strengths:** much faster scan speeds and much reduced traffic from Exadata storage to the database both result in significantly improved performance.
- Weaknesses:** Oracle's planned acquisition of Sun raises potential doubts about future of HP relationship, upgrade complications for existing users.
- Comment:** the introduction of Exadata makes Oracle a contender for analytics in a way that was not previously the case.

Vendors

ParAccel

Product:	ParAccel Analytic Database (PADB), version 2.0.
Architecture:	shared-nothing, MPP-based hybrid column/in-memory; either direct attached storage (DAS) or DAS in conjunction with SAN (for high availability).
Offering:	software or appliance (Scalable Analytic Appliance), the latter is also resold by EMC.
Customers:	10–20.
Geography:	focused on US market at present (but EMC worldwide).
Scalability:	1 to 100Tb, theoretically more.
Target sectors:	any, with traction in retail, financial services and analytic service providers.
Functionality:	analytic data mart or EDW.
Status:	rapidly growing next generation vendor.
Strengths:	all advantages of columns combined with in-memory capabilities to give additional performance boost, probably most advanced SAN implementation (for performance as well as HA) available on the market, advanced specially-designed optimiser, very fast load speeds, competitive price/performance.
Weaknesses:	limited mixed query workload capability.
Comment:	One of the leading vendors for analytic warehouses in all categories.

SAP

Product:	SAP NetWeaver BW, version 7.01 with SAP NetWeaver Accelerator.
Architecture:	Layered scalable architecture; BW is typically based on an Oracle, DB2, MaxDB or SQL Server database (in the future Teradata) and is row-based; Accelerator is column-based (used for InfoCubes only at present), has an MPP architecture and uses blade servers, indexes are held in memory.
Offering:	BW is software, Accelerator is an appliance.
Customers:	lots.
Geography:	worldwide.
Scalability:	no lower limit stated, should scale upwards to hundreds of terabytes.
Target sectors:	all.
Functionality:	EDW or EDW with associated data marts.
Status:	a market leader.
Strengths:	integration with other SAP environments, some optimisation (more to come) for use with SAP Business Objects for business intelligence, future optimisation with other Business Objects technology including data integration, data quality and metadata management; advantages of using columns limited to InfoCubes at present but will be extended to further analytics in future, plus conventional BI capabilities; good mixed query workload capabilities, support for near-line storage.
Weaknesses:	underlying layered architecture is complex but this is increasingly being hidden and the layering automated.
Comment:	BW Accelerator significantly improves analytic performance albeit that this is limited to InfoCubes at present. The planned Teradata port will further enhance its competitive position.

SAS

Product:	Scalable Performance Data (SPD) Server.
Architecture:	row-based, SMP.
Offering:	software (partners offering hosted services).
Customers:	n/a.
Geography:	worldwide.
Scalability:	up to hundreds of terabytes.
Target sectors:	any.
Functionality:	analytic data marts.
Status:	primarily sold as underpinning to analytic applications.

Vendors

Strengths:	optimised functions for analytic processing, integrated capabilities for supporting SAS data mining.
Weaknesses:	no data encryption, no focus on stand-alone sales, row-based approach will require extra disk storage and more administration, SMP approach.
Comment:	primary market is as underlying platform for SAS applications. Little or no focus on stand-alone sales.

Sensage

Product:	Event Data Warehouse, version 4.5.
Architecture:	columnar, MPP shared-nothing.
Offering:	software but available as an appliance from OEM partners such as HP and Tokyo Electron, private and public cloud options.
Customers:	400+ (500+ installations).
Geography:	US and EMEA offices, partners worldwide.
Scalability:	minimum 10Tb, sweet spot 100Tb to 2Pb, should scale further than this.
Target sectors:	communications, government, financial services, health services, insurance and retail.
Functionality:	large analytic warehouses for security, compliance and operational risk; CDR and IPDR in communications along with location intelligence; transaction fraud prevention.
Status:	leading vendor in log management and data retention markets moving into more generalised warehousing environments.
Strengths:	all advantages of using columns, proven scalability, MapReduce-like capability, time-stamped event data, transaction support for non-event data, in-database analytics, significant pre-built and in-built reporting and query capability, support for ODBC/JDBC to integrate with third party BI tools.
Weaknesses:	limited experience (and partnerships) in general-purpose warehousing market, limited mixed query workload capability.
Comment:	has significant advantages wherever event-based data is needed either on its own or in conjunction with other data.

Sybase

Product:	Sybase IQ Analytics Server, version 15.1.
Architecture:	column-based, shared-disk grid SMP with multiple read and read/write nodes.
Offering:	3 software Editions: Enterprise Edition, Small Business Edition (which is also available via Amazon Cloud), Single Application Server Edition; or as an appliance with three base editions that include the analytics server, ETL, reporting tool and management software.
Customers:	1,600+ (3,100+ installations).
Geography:	worldwide.
Scalability:	can be used at Gb level, regard sweet spot as 5 to 500Tb of raw data, theoretically scales to petabytes.
Target sectors:	by function—for advanced analytics, fraud detection, report acceleration, data aggregator analytics, where significant amounts of structured and unstructured data; also for risk management in capital markets, compliance and data/text mining.
Functionality:	analytics, report servers, data marts, and EDWs.
Status:	largest number of customers after the big merchant vendors; leading column-based vendor.
Strengths:	all the advantages of columns, significant additional indexing capabilities including text indexing, advanced encryption and security, support for in-database analytics, mixed query workload management.
Weaknesses:	use of indexes may impose additional overheads compared to other columnar approaches.
Comment:	a leading vendor for all classes of analytic warehousing.

Vendors

Teradata

Product:	Teradata Database 13.
Architecture:	row-based shared-nothing running on either MPP or SMP.
Offering:	Teradata Data Mart Appliance 551, Teradata Data Warehouse Appliance 2550, Teradata Extreme Data Appliance 1550, Teradata Active Enterprise Data Warehouse 5555, Teradata Data Mart Edition (software only).
Customers:	900+ (2,000+ installations).
Geography:	worldwide.
Scalability:	for EDW typically 10 to 500Tb but proven implementations at petabytes scale, for analytic data marts typically 6 to 50Tb but have customers with as little as 250Gb.
Target sectors:	all.
Functionality:	EDW and data marts.
Status:	market leader at top end of scalability.
Strengths:	established performance characteristics, advanced mixed query workload capability, extensive indexing, in-database geospatial, porting to support SAP.
Weaknesses:	perceived (incorrectly according to Teradata) to be expensive and not interested (ditto) in data marts.
Comment:	the leading vendor in the EDW market and also competitive for purely analytic environments, especially at the high end of scalability.

Vertica

Product:	Vertica Analytic Database, version 3.5.
Architecture:	column-based, MPP shared-nothing or shared-disk (SAN).
Offering:	software only, public cloud, VMWare appliance, traditional appliance.
Customers:	85+.
Geography:	mainly North America, recently opened UK office, has clients in Italy.
Scalability:	from Thumbdrive to hundreds of terabytes, theoretically petabytes.
Target sectors:	especially financial services, telecommunications, healthcare and marketing analytics.
Functionality:	analytic data marts.
Status:	rapidly growing next-generation vendor.
Strengths:	all advantages of columns plus Flexstore provides further improved performance, support for external MapReduce environments, excellent load speeds.
Weaknesses:	limited mixed query workload capability, no data encryption (but may be encoded and/or compressed).
Comment:	the, or a leading, vendor for all classes of analytic warehouses and data marts.

Vendors

Vendors not included

There are a number of vendors that have not been included in this report that might be worth considering under particular circumstances. These include:

- MySQL and other transactional databases: where there is little analytic requirement. In the case of MySQL relatively small scale (a few hundred gigabytes) unless an alternative storage engine (from Tokutek, say) is used.
- EnterpriseDB: has plans to target the low-end (up to 5/10Tb) data warehousing market. Does not have any particular features to support analytics but does have an MPP architecture. Most likely to be useful for companies growing out of MySQL or Oracle (EnterpriseDB is Oracle compatible) with limited analytic requirement.
- Previous generation column-based products: these include Alterian, Kx Systems and SAND technology, which specialise in market campaign management, capital markets and near-line storage respectively. Note that the capital markets sector is also being targeted by Sybase and Vertica (in conjunction with Streambase), amongst others.
- Log management vendors: there are a number of these that target the data retention and event markets in particular, where they will meet Sensage, Vertica, Netezza and others in competition.
- HadoopDB and Cloudera: commercial implementations of Hadoop: neither of these is mature enough yet in our opinion.
- Dataupia: although a current warehousing vendor with solutions that back-end onto Oracle and SQL Server databases we have removed Dataupia from consideration in this report because it's financial future is currently in serious doubt.
- Calpont: a Texas-based columnar vendor that has yet to come to market. The company has been waiting in the wings for some years so the delay in releasing a product is potentially worrying.
- MonetDB and LucidDB: open source, columnar databases; the former developed at CWI in Amsterdam. No commercial implementations in either case (LucidEra has ceased trading).
- Vectornova: Mexican-based company that trades in Latin America and Europe. Open sourced, columnar database that uses vector processing. Has both a software-only and an appliance (2Tb) offering with others planned. Supports both J and R.
- Ingres, the open source vendor: has announced that it will be introducing a new data warehousing product in 2010 in conjunction with VectorWise. The latter is a spin-off from the CWI Institute in Amsterdam (where MonetDB was developed) and it has been working, using vector processing, to exploit the parallel capabilities inherent in today's chips (individual cores), which traditional databases ignore. This provides orders of magnitude performance gains at that level. In addition to vector processing the product will include both column and row-based storage.
- XtremeData: new, US-based vendor offering appliances with user capacities of 30, 60, 105 and 225Tb. MPP, shared-nothing architecture based on the PostgreSQL database. Uses FPGAs. Architecture looks very similar to Netezza though uses FPGAs in a somewhat different way.

Conclusion

The analytic warehousing market is a dynamic one. New vendors continue to appear and innovative approaches are widespread. In this paper we have attempted to give an overview of the main concerns in the market and those vendors that currently play in it. In the papers associated with this one we will delve deeper into the individual sub-markets that make up this space but as a preview, the following table shows the leading vendors in each category:

	Category leader(s)	Highly recommended	Strong contender(s)	Serious possibilities
Small-scale analytic warehouses (up to 5Tb)	ParAccel Vertica	Exasol	1010Data InfoBright Netezza Sybase	Teradata
Medium-scale analytic warehouses (5 to 50Tb)	Netezza ParAccel Vertica	Sybase	1010Data Exasol IBM InfoBright Sensage Teradata	
Large-scale analytic warehouses (over 50Tb)	Netezza	Teradata	IBM Sensage Sybase Vertica	Oracle ParAccel
Enterprise data warehouses	Teradata	IBM	Netezza	Oracle Sybase

Of course, this ignores special situations: if you want to use SQL tightly integrated with MapReduce, for example, then you would certainly want to look at Aster Data and possibly Greenplum; if you want time-stamping then you would look at Sensage, and if you want spatial analytics then you might look at Netezza, Teradata, IBM or Oracle; and so on.

Further Information

Further information about this subject is available from <http://www.BloorResearch.com/update/1052>

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 whole story' with independent, intelligent, well-articulated communications content and publications on all aspects of the ICT industry. We believe the objective of telling the whole 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

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 © 2009 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