The Information Difference

Data Warehouse Landscape - Q4 2009

The Information Difference Landscape is a high level assessment of the main and most innovative vendors in a market at a point in time. The diagram shows three dimensions. The size of the bubble is an indication of the customer base of the vendor i.e. the number of corporations it has sold to, adjusted for deal size. The larger the bubble, the broader the customer base, though it is by no means to scale. The technology dimension position is derived from a weighted set of scores based on four factors: customer satisfaction as measured by a survey of reference customers, analyst impression of the technology, maturity of the technology and breadth of technology in terms of its coverage against our functionality model. The market strength position is derived from a weighted set of scores based on five factors: data warehouse revenues, growth, financial strength, breadth of partner network and geographic coverage.

The data warehouse market has been established for at least two decades, yet has seen a resurgence of activity in the last few years as a series of new vendors have entered the market. Traditionally data warehouses were custom-built applications on top of traditional relational database platforms (mainly from Oracle, IBM and Microsoft). The major relational databases were initially optimised mainly for transaction processing, which has lots of concurrent updates and a relatively small amount of data accessed by any one user. By contrast data warehouse access is typically read-only, and usually involves access to lots of data in a single query. Consequently, early data warehouse pioneers often found performance issues with large data warehouses.

Over time the major database vendors responded by adding specific functionality for data warehousing (such as specialist indexes that are suited to read-only processing) while Teradata in particular established the market for an “appliance”, a specialist database complete with hardware, tuned specifically to data warehouse loads, and able to take advantage of massively parallel processing (MPP). It is important to understand that the data warehouse market is itself connected to the market for business intelligence tools, and also to the data integration market, whose technologies are frequently used to gather data from multiple sources and feed data into a data warehouse. The traditional relational database vendors have made significant acquisitions in recent years, with Oracle, IBM and Microsoft now offering a complete suite of data integration, data warehousing platform and business intelligence tools.

However, while this consolidation was going on, the market also diverged. Columnar databases, pioneered by Sybase IQ, used a different technology that was well-suited to read-only processing, and allowed significant data compression, boosting performance; others (such as ParAccel and Vertica) have since followed this route. Software applications to handle the creation of maintenance of data warehouses, such as Kalido Dynamic Information Warehouse and SAP BW also appeared; Wherescape is a newer example of this. For relational deployments there has been considerable discussion of alternative schema designs, with third normal form, “star” and “snowflake” schemas all having different pros and cons and gaining adherents. Some deployments involve a multi-stage approach, with an “operational data store” in conjunction with a data warehouse.

In the last five years the choice of platform offerings has further diversified. Netezza achieved considerable success as an appliance, with its approach of using co-processor hardware to boost performance. A series of other MPP-based appliance vendors also emerged, a pioneer being Kognitio, some using row-based databases and some columnar approaches. The most recent hardware-related developments have been offerings that take advantage of specialist chips (Kickfire with its own specialist chip technology, and Ingres with its Vectorwise feature). For the most demanding applications and where price is a secondary consideration, appliances using solid state memory are beginning to appear: Teradata already has an example of this. The largest vendors have also embraced the appliance route: around half of IBM's new customers choose their appliance offering, while Oracle has its Exadata offering, recently updated to Exadata V2.

Another feature of the market has been the dramatic increase in the volumes of data that companies need to manage. As recently as 2005 the largest data warehouses were around 100 terabytes (TB) in size; now there are several examples at the petabyte level. Such volumes clearly represent significant operational challenges, and it has meant that the data warehouse appliance market has become somewhat segmented. Differing approaches have targeted varying subsets of the market.

At the low end, Kickfire aims firmly at providing low-cost data marts at the few TB range, as do Infobright, Exasol and Illuminate with its "correlation database". Moving up from this into slightly larger workloads are products such as Vertica and ParAccel. Asterdata has, in our view, a particularly elegant MPP-based architecture, and has some quite large deployments. HP also entered the high-end market with its Neoview offering. At the rarefied end of the market, there are a few petabyte sized data warehouses deployed (Teradata, Netezza, Greenplum, Oracle, IBM).

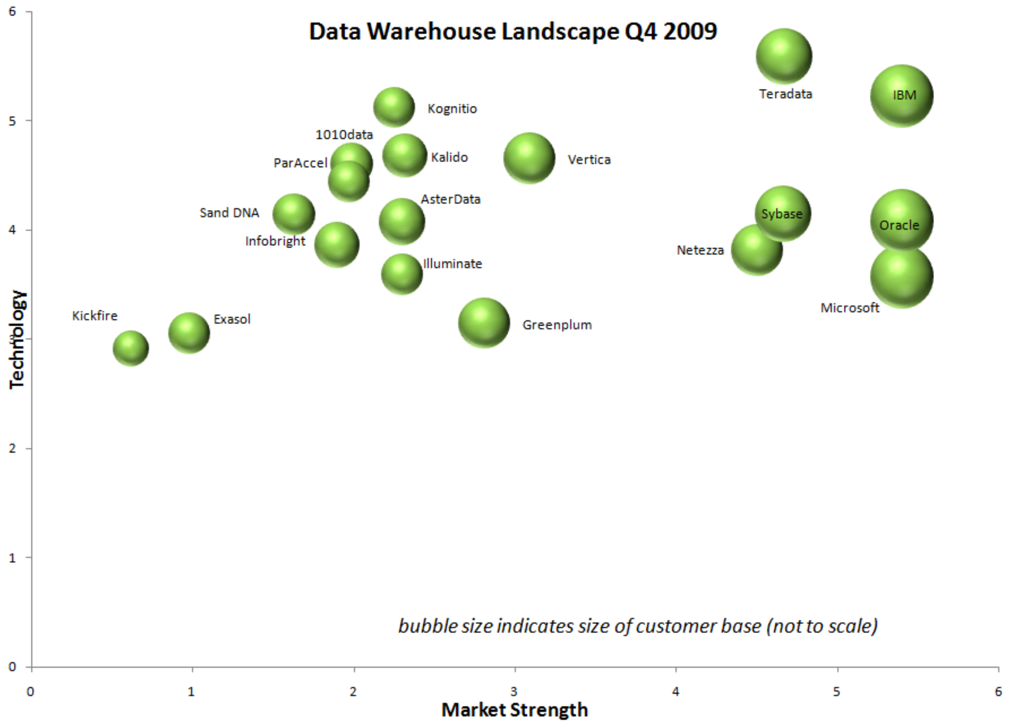
Microsoft was previously content to offer SQL Server, which was typically deployed in data warehouses of at most tens of terabytes size, but is about to enter the appliance market in 2010 through its Madison offering, based on its acquisition of DataAllegro. This product will scale up to hundreds of terabytes.

It should be considered that size in terms of data is not the only scalability dimension: the number of concurrent users that can be supported may be equally important. The type of query/analysis required also has a major impact on performance. For example, some highly computationally intensive queries can constitute a performance challenge to any SQL-based database. One approach to this is to embed support for specific libraries of algorithms in the database itself, while a more recent approach has been to adopt the distributed computing framework MapReduce, early adopters of which have been Asterdata and Greenplum. One trend in the market is the storage and efficient interpretation of different data types, such as spatial data and documents, and managing these in conjunction with traditional structured data.

One market niche has been addressed by Sand, which these days specialises in providing "near line storage", particularly suited to archiving applications where a company wants to keep large volumes of data for occasional access, and does not want to store this data on high-cost storage. A further recent trend has been to offer a data warehouse service over the web, or "in the cloud". 1010data is one company taking this approach (as does Kognitio).

With such different sub-markets it is important that end-users carefully consider the alternatives appropriate to them to match their particular need; simplistic overviews of the market, such as this Landscape, cannot capture specific customer requirements, and any technology selection process should be discussed in detail with an analyst.

As part of the research process vendors were asked to provide customer references, who were sent a survey on their satisfaction with the vendor's products (if they failed to provide references, a neutral score was assigned). Based on this survey, the data warehouse vendor with the happiest customers was Paraccel, closely followed by Vertica, Teradata, Kognitio and Sand.


Data Warehouse Landscape Q4 2009
Teradata
IBM
Kognitio
1010data
Kalido
ParAccel
Vertica
AsterData
Sand DNA
Infobright
Sybase
Oracle
Netezza
Illuminate
Microsoft
Kickfire
Exasol
Greenplum
Technology
bubble size indicates size of customer base (not to scale)
Market Strength