InDetail

Bloor

Netezza Data Virtualizer powered by Composite Software

A InDetail Paper by Bloor Research
Author : Philip Howard
Publish date : June 2010

We have long been fans of both Netezza and Composite Software and it is good to see an alliance between two best-of-breed vendors in complementary spaces

Philip Howard

Executive summary

It is often the case in large enterprises that multiple data marts are deployed, along with a central enterprise data warehouse acting as a system of record. While these data marts serve the requirements of the departments or divisions that implement them, a frequent requirement is to merge data across these implementations. While this can be done by replicating the data across systems, this is an expensive and inflexible option. A preferred solution is to use federated technology that will support queries across systems. This is what Netezza Data Virtualizer does. It also supports other use cases but this will be the primary one for most companies.

Fast facts

Powered by the Composite Information Server, the Netezza Data Virtualizer provides a method for linking multiple Netezza appliances into a single, federated environment. It is available directly from Netezza. A full licence version of Composite Information Server is available directly from Composite Software, which will allow federation between Netezza appliances and both third party data warehouses and operational systems.

It is assumed in this paper that readers are familiar with Netezza and its technology does not therefore require any discussion here. We thus limit ourselves to a brief consideration of Composite's technology, together with details of how it interrelates to Netezza, plus a discussion of potential uses of this combined technology. More detailed descriptions of both Composite Information Server and Netezza are available separately from Bloor Research.

Key findings

In the opinion of Bloor Research the following represent the key facts of which prospective users should be aware with respect to Netezza Data Virtualizer:

- Netezza is the leading new generation data warehousing provider and Composite Software is the leading provider of data federation capabilities. Thus this is an alliance that makes obvious sense.
- While some other companies may offer lightweight data federation capabilities as an extension to their BI or ETL offerings, we believe that Composite has the richest product set for supporting federated queries on the market.
- With NPS 5.0 and beyond, Composite Software supports federated environments that include Netezza via a Netezza-specific extension of Composite's industry-standard JDBC interface. With this release a native interface (which understands Netezza DDL) has been released with additional specific optimisations designed to support Netezza implementations. Earlier Netezza installations will only be able to leverage the generic optimisation capabilities of Composite Information Server.
- New Netezza-specific optimisations include data-ship-enabled joins for large resultant data sets across multiple appliances, SQL pass-through when querying a single appliance, and a number of select and insert performance improvements to leverage advanced Netezza capabilities.
- Netezza Data Virtualizer is available directly from Netezza, which provides first line support for the product.
- In addition to supporting federated query functions there is also a Composite Active Cluster option that supports high availability active-active clustered Netezza environments.
- For users wishing to extend the use of Composite Information Server to third party data warehousing or transactional systems, or to support migration from these environments to Netezza, then the product's capabilities are available directly from Composite Software.

The bottom line

In all too many organisations, data marts and analytic appliances stand as islands within their environments. Or, where they are not islands, data is copied across databases in order to allow queries that span the data in these islands. The first approach is inflexible, the second is expensive. The ability to run federated queries overcomes this issue in a way that is both agile and relatively inexpensive. Of course, you could use Composite Information Server with Netezza prior to this release but the software was not optimised for Netezza. This release not only highlights that this capability is available but optimises federated queries for a Netezza environment. This both extends Netezza's offering and provides a competitive advantage in the sense that Composite Information Server has not been similarly optimised for most of Netezza's competitors. We therefore believe that this release will be good for both companies.

Use cases

There are a number of use cases for which Netezza Data Virtualizer will make sense. These include (but are not limited to):

1. Federating across data marts. For example, suppose that your organisation is structured into separate divisions, each with its own data mart or warehouse. You could then use Netezza Data Virtualizer to run queries about, say, customers, across these different data sources, without having to copy the data into a further warehouse. A similar requirement might occur where data marts are vertically oriented (for example, customer, product, and supplier analytics) and you want to run queries across these vertical boundaries.

   Federating across aged data. This is illustrated in Figure 1. Here you are storing information in different environments, depending on the age of the data, but you may still want to run queries across these different environments.

2. In Figure 1, the implication is that you are using Netezza both as an operational data store and for data warehousing, with Netezza Mantra as your compliance data vault. This could be further extended to front-office transactional data sources using the full licence Composite Information Server product available directly from Composite Software. Similar considerations would apply where not all parts of the data warehousing environment are provided by Netezza.

3. Migration. Another use is to licence the full version of the Composite Information Server where you are intending to migrate from a third party data warehousing environment to Netezza. By implementing Netezza while the previous incumbent is still in place you could use Composite Information Server to build virtual queries that duplicate application calls to your existing database and then use this facility to gracefully migrate from the existing environment to the new Netezza one.

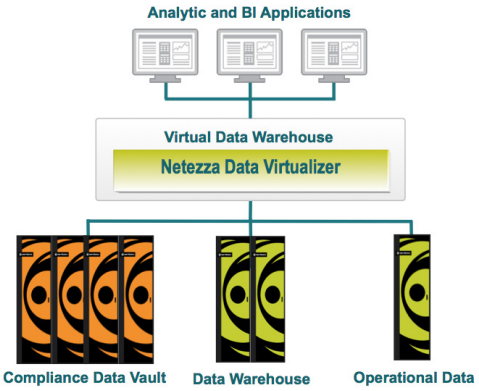

Figure 1: Federated queries against aging data

4. Not all use cases involve queries or, at least, not just queries. An optional add-on is the Netezza Data Virtualizer Active Cluster option. Active Cluster allows you to deploy multiple instances of Netezza Data Virtualizer across a cluster for scalability, load balancing and high availability purposes. It provides a peer-to-peer environment with automatic synchronisation across instances and it also supports the caching of shared metadata. Thus you could deploy a cluster of Netezza appliances deployed in an active-active fashion. This sort of high availability requirement is becoming more and more important, given the mission critical nature of many analytic applications. Moreover, there are also potential extensions to this scenario: for example, suppose you have a requirement for large batch updates but still need to be able to provide query capabilities: you can use the Active Cluster option to enable each appliance to be updated separately while the other is supporting relevant queries. This sort of continuous availability is becoming more and more of a requirement in many organisations.

Product availability

Netezza Data Virtualizer is based around the latest version of Composite Information Server (version 5.1.2), and will run with Netezza NPS (5.0 or later), TwinFin, Skimmer and future products. That said, the two companies have a number of existing joint customers, demonstrating that Composite Information Server will run with earlier Netezza appliances, though this integration will be generic and will lack the more advanced features, which we will discuss in due course, associated with this release.

Composite Information Server is a pure Java application and it supports 64-bit native JVM and runs on Windows 2000 and later, Sun OS and Solaris, Linux (Red Hat and SUSE), AIX and HP-UX. Clients may be based on Windows 2000 or Windows XP. The product has strong support for web services, whose capabilities are built-in (thus it is not dependent on a third party environment such as JBoss) and there is support for SOAP over JMS (publisher and subscriber) on both Sonic ESB and TIBCO BW, as well as REST.

More generally, data sources supported include DB2 on zOS via a native driver and any ODBC (64bit) and JDBC compliant data sources (Teradata, DB2, Oracle, Informix, Sybase, SQL Server, MySQL and so on); there is also an ADO.NET driver and the company resells Data Direct (Progress) Shadow RTE software to connect to legacy mainframe sources such as VSAM, CICS, Adabas, IMS/DB, IMS/TM and Natural. Microsoft Excel, flat files and XML data sources are all supported and in the case of XML the product supports XQuery, so you can use this to run queries against the XML if you want to. Note too that Composite Information Server can be used as a data source by other instances of Composite Information Server: this is particularly useful if you have multiple instances deployed across multiple geographies.

Product details

In the following sections we will briefly detail the major features of Composite Information Server itself and the specifics of the interaction between this and Netezza appliances. As previously noted the discussions of the individual products are not intended to represent detailed product evaluations.

Composite Information Server

To put it simply, when you are using Composite Information Server you will build database views that you can run queries against using a standard business intelligence tool. These views are exactly like the views you would construct in any database environment. The only difference is that these views span multiple data sources (Netezza specifically in the case of Netezza Data Virtualizer) rather than being limited to a single database. As such, these are not instantiated in the database but are, in fact, virtual views. This is why Composite sometimes refers to its technology as a data virtualisation platform, though this is not to be confused with virtualisation as provided by the likes of VMware.

Of course, there is a lot of technology that underlies Composite's capabilities but this is hidden as far as the user is concerned and we will not discuss it here. What is relevant to the user is illustrated in Figure 2.

The actual aspects of the Composite environment that the user will interact with are as follows:

- Composite Discovery is used during the development and high-level design phases to discover and model key entities and relationships.

- Composite Studio and Composite Designer provide two different approaches to development. Both tools are graphical with the former providing a traditional (E-R diagrams and the like) database-centric environment for relational developers, and the latter providing an Eclipse-based environment for web services and Java developers. There is support for XML and stored procedures can be written in Java, XQuery or Composite Script, as required. Both products automatically generate relevant code.

- Composite Manager does what its name suggests, providing an administrative interface for access security, metadata management, source code control and so on.

- Composite Application Data Services provides pre-built objects for leading ERP suites such as SAP, Siebel (Oracle), Oracle E-Business, Salesforce.com and PeopleSoft

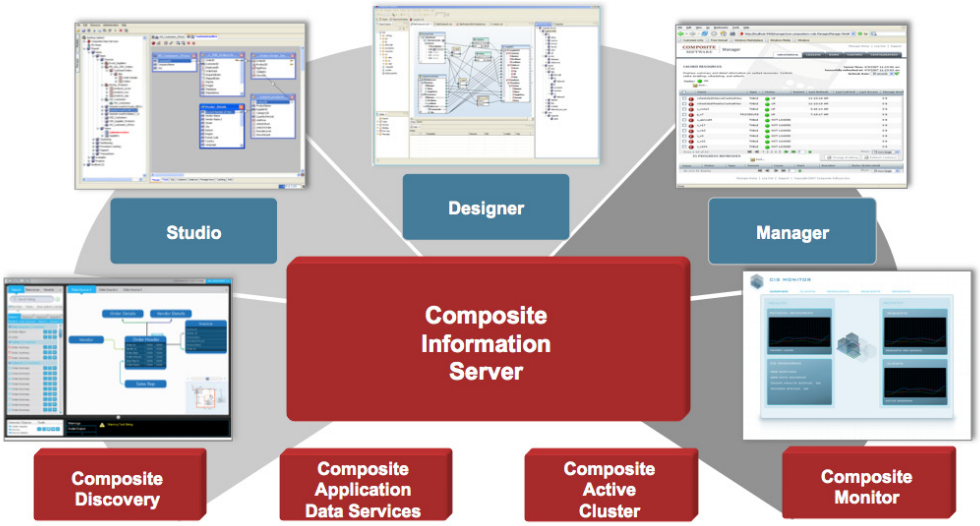

Figure 2: Composite Data Virtualisation Platform

Product details

(Oracle) environments. These integrate with the relevant environments at a deep level, abstracting common business objects such as orders, invoices, customers, suppliers, and so on. This module also provides SQL to MDX translators for SAP BW and Oracle Essbase.

- Composite Active Cluster, which we have already discussed.

- Composite Monitor, which not only provides monitoring capabilities for the run-time environment but also supports SNMP and links to third party system monitoring tools.

- Composite Information Server itself provides the run-time query engine that will perform the various queries that have been defined.

More generically, there are two major foci within Composite Information Server. The first is ease of use and productivity for developers, as provided by the Studio and Designer tools, and the second is on performance. The latter point requires further discussion.

In order to maximise performance in a federated environment the key is to move as little data around the network as possible. At its simplest this means that if you have to join two tables residing in different places, then you move the smaller to the larger not vice versa. However, there is more to it than this. To begin with you may not have all the information you need to determine the best stratagem for performing a particular query, so you need to have good methods for estimating data sizes. In addition, you cannot simply parse SQL as is, because it is likely to be inefficient: you need to collapse all the SQL layers before you can determine the best query plan. Moreover, this has to be a global plan that spans all of your data resources so that the query knows when to use push-down optimisation and when not to, for example.

To support all of this, Composite makes use of multiple algorithms for estimating the costs of a query as well as rules for determining push-down and other functions. Thus the product uses a hybrid rule and cost-based optimiser. There are also some 'exotic' optimisations such as the ability to flip a join over unions to a union over joins, which can be useful in certain scenarios. Needless to say, more optimisations tend to appear with each new release. In addition, the product supports hints so that developers can give, say, an order of magnitude indication when creating views, which will help the optimise to work most efficiently.

Complementary functions

Work to optimise Netezza Data Virtualizer has primarily been on Composite's part, though with design engineering and testing assistance from Netezza. The emphasis, as with Composite more generally, has been on performance. Notable features include:

- Support for distributed semi-joins (so-called data-ship joins) in order to allow the creation of temporary tables.

- Use of the Netezza bulk loading feature to improve performance.

- The ability to use Netezza to store cached information (requiring the ability to understand Netezza DDL).

- Support for pass-through where there are OLAP functions that are Netezza-specific.

On the commercial side, Netezza Data Virtualizer is directly sold and marketed by Netezza and is not available from Composite. Conversely, if you want to use Composite Information Server to federate with beyond Netezza's, then this capability is only available from Composite directly. First level support in the former case is provided directly by Netezza.

The vendors

Composite Software

Composite Software is a privately held company based in California that was founded in 2002. At that time there were a number of start-ups focusing on data federation. Composite, however, has outlived almost all of its pure-play competitors and its Composite Information Server (CIS) is now the leading product in this sector.

Apart from technical considerations, a major reason behind this success has been that the company, from the outset, has focused on a dual approach to marketing with both a strong direct sales force and a concentration on partnerships. In particular, in 2004 Cognos (now IBM) not only signed an OEM agreement (CIS is embedded in Cognos) with Composite but also took a 10% share in the company. Further, CIS is also embedded within BMC's Atrium product and the company also has strong partnerships with a number of other vendors, of which its partnership with Netezza is the latest example. In total, this indirect channel represents approximately half of Composite's revenues with the remainder coming from its direct customers. The product is aimed primarily at Global 2000 accounts with typical deployments at a departmental level, though many of the company's customers have multiple deployments; for example, both Pfizer and Wachovia have around 20 such deployments. The company, like Netezza, has a worldwide presence.

Composite web site: www.compositesw.com

Netezza Corporation

Netezza Corporation is named after the Urdu (the official language of Pakistan, though it is also spoken in parts of India) word netēzza, meaning 'results'. It is pronounced as 'net-teaser' and the symbol on the 'ē' is known as a macron.

The company was founded in 2000 and is listed on NASDAQ. Its first generally available product was shipped in January 2003. The company uses a direct sales force model together with a distributor programme and also leverages systems integrators. It has partnerships with all of the leading data integration and business intelligence vendors and the company also supports an active developer community through the Netezza Developer Network.

Since its founding Netezza has shown rapid organic growth, demonstrating clear leadership across the latest generation of data warehousing vendors, with more than 400 installations to date. It has also acquired spatial analytic technology from IISi and acquired both NuTech Solutions and Tizor, where the former provides specialised data mining and analytic technologies and the latter, re-branded as Netezza Mantra, provides compliance and audit solutions not just for Netezza implementations but also for enterprise data more broadly.

Netezza web site: www.netezza.com

Summary

We have long been fans of both Netezza and Composite Software and it is good to see an alliance between two best-of-breed vendors in complementary spaces. We believe that Netezza Data Virtualizer adds substantial value for users of more than one Netezza appliance that want to be able to meet one or more of the use cases previously described. It will also open up additional possibilities for Composite Software where users want to extend their federated capabilities beyond those of Netezza per se, whether for operational or migration purposes.

Further Information

Further information about this subject is available from http://www.BloorResearch.com/update/2034

Copyright & disclaimer

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