August 2003

DB2 Information Management Software

The Real Data Warehousing Story with DB2® Universal Database™ and DB2 Information Integrator™

IBM Software Group
Toronto Laboratory

Contents

Highlights

Introduction

It has recently come to light that there are some growing misunderstandings and confusion surrounding IBM's position on DB2 Information Integrator and its role in the data warehouse. This is not unusual when dramatically new innovations are brought to market. Also being questioned is the ability of DB2 UDB to handle the workloads associated with a data warehouse. This document addresses these misconceptions and clears up the confusion. To be clear, IBM's position is that:

- IBM does not promote the concept of "virtualized data warehouses" or claim that federated technology should be the basis of a data warehouse solution. IBM proposes that federation is a powerful tool for extending the data warehouse, not replacing it.
- DB2 UDB has extensive features and capabilities that make it an extremely powerful OLTP and decision support platform.
- DB2 UDB's multi-platform support and IBM's experience and skills in data warehousing can help ensure that the right data warehousing environment for a customer is proposed every time.

DB2 Information Integrator and Federated Technology

The virtualized or federated approach to data warehousing has certain flaws and inadequacies.

Industry experts have long agreed that the virtualized or federated approach to data warehousing has certain flaws and inadequacies. With such an approach, data usually resides in multiple, disparate systems and there is a federated layer that sits on top of them that gives the data warehouse the appearance of a single system. Issues that come up when attempting to use this approach include data cleanliness, semantic integrity, and unpredictable performance.

Our federation technology does not replace data warehousing.

IBM has clearly and consistently stated that our federation technology does not replace data warehousing, for precisely the reasons that many of the industry experts describe. IBM agrees that federating dirty or inconsistent data produces dangerous and nonsensical results. Nor can federation create an historical view of data as is typical in a warehouse. It is equally clear that a federated query is heavily dependent on network performance and availability.

IBM has carefully described the circumstances under which federation can be used to extend the data warehouse in two IBM White Papers, "*Information Integration - Extending the Data Warehouse*" and "*Information Integration - Distributed access and data consolidation*". These white papers are readily available on the IBM web site at http://ibm.com/software/data/pubs/papers as of April 2003. How federation fits into an existing data warehouse is also discussed in a DB2 Magazine article entitled "*Solving the Data Warehouse Puzzle*" which can be found at http://www.db2mag.com/db_area/archives/2003/q2/devlin.shtml.

Highlights

Federation is not a substitute for a data warehouse.

To summarize IBM's message: In a data warehouse environment, DB2 Information Integrator can enable access to certain data outside the warehouse where there is a need for real-time information, or where, for technical, legal, or other reasons, it is not reasonable or possible to create another copy of that data in the warehouse. Federation is not, however, a substitute for a data warehouse. Said differently, it is impossible or infeasible to assume that all data ever needed in a user query would always reside in the data warehouse. While some DBAs may wish to put every piece of corporate data into the data warehouse at all times, there are a number of reasons why this never happens. Consequently, clients can use programmer labor for these truly ad hoc queries, they can tell the end user to forget the business need, or they can use IBM's DB2 Information Integrator within certain constraints.

META Group analyst Doug Laney agrees with this philosophy. In a METAbit article released on August 21st, 2003 Doug had the following to say about EII (Enterprise Information Integration):

> "We have long-advocated EII merely as a means for low-volume federated access to heterogeneous data sources. Similarly, EII vendors (e.g. IBM, Metamatrix, Enosys) have been very careful not to characterize their solutions as a substitute for physical data warehouses (DW). Yet, some in the industry still have trouble appreciating that EII is an enabler for a valuable new style of data integration (i.e. "virtual" or "federated" integration)--one that can extend a DW but certainly does not intend to replace it."[1]

"… EII is an enabler for a valuable new style of data integration … one that can extend a DW but certainly does not intend to replace it."

IBM is working with customers and our field technical organization to clarify best practices on when and how to use federated queries with a data warehouse. We understand deeply that federated queries cannot side step data cleansing, semantic transformations, and other context changes to data between production systems and the data warehouse. If such transformations are extensive, federated queries may not be appropriate and the end user must look to their programming staff for help. DB2 Information Integrator can do "light transformations" but does not supersede the need for an extract-transform-load (ETL) process.

Furthermore, queries that join data from DB2 with SQL Server or Oracle across a network have some growing up to do. IBM's optimization technology does an excellent job of choosing the best query plan for each database accessed, moving the least amount of data at any given time. Nevertheless, this is not a panacea. In not having some critical database statistics or contextual knowledge, the optimizer may not choose the ideal query plan for a federated query. However, in a large percentage of cases, the federated query can be effective, providing value to the end user and simplifying the DBA's workload.

The federated query can be effective, providing value to the end user and simplifying the DBA's workload.

Highlights

Initially, we recommend DBAs explore federated queries for static reporting used weekly, monthly or yearly. This allows reporting to occur on occasionally used data, thereby reducing the need to have all the data in the data warehouse – a potential cost savings in servers and storage. Similarly, federated queries work well in the extract-transform-load process feeding the data warehouse. We've also found that federated queries can be effective in bringing together single instance data from multiple systems, such as a unified customer record gathered from an ODS and a content repository. IBM does *not* recommend using true ad hoc federated queries joining the data warehouse with production databases interactively. A DBA must guide interactive ad hoc queries in the initial stages to achieve success. Some best practices performance guidelines exist today with much more to come as we explore this innovative strategy with our clients.

A Powerful OLTP and Decision Support Platform

It is simply irrational to claim that a feature that helps data warehousing will automatically hurt OLTP.

Some niche vendors claim that no platform can do multiple things well and if one tries then it must be optimal for nothing. It is simply irrational to claim that a feature that helps data warehousing will automatically hurt OLTP, and vice-versa. By some estimates, 70-80% of RDBMS development investment is for common functionality favorable to any application: reliability, availability, serviceability, administration, international languages, SQL, APIs, etc. Said differently, good performance in OLTP does not guarantee bad performance in data warehouses, except where development budgets are severely limited. IBM does not have this problem.

These kinds of "functional specialist" comparisons may apply to surgeons, airline pilots, or stock traders, but not to software. Software can excel at many tasks with enough time and investment. Otherwise, we would see operating systems designed only for OLTP or data warehousing. Such custom operating systems faded in the 1980s because software can be made to excel at many workloads, and in some cases simultaneous workloads. We ask, is UNIX® incapable of handling OLTP and data warehousing? Our customers know the answer.

Highlights

DB2 UDB has features that make it flexible and scalable enough to handle the operational and analytic needs of almost any user.

IBM was the first to publish a TPC-H result at the 10 TB scale factor.

DB2 UDB has many features that make it an extremely powerful OLTP and decision support platform, features that make it flexible and scalable enough to handle the operational and analytic needs of almost any user. Proof of this can be seen in the industry benchmarks that DB2 UDB participates in. For example, DB2 UDB recently held the #1 position for performance in the Transaction Processing Performance Council TPC-C benchmark which demonstrates that DB2 UDB can perform OLTP type workloads very well[2]. Likewise, IBM was the first to publish a TPC-H result at the 10 TB scale factor[3]. TPC-H is a decision support environment and DB2 is one of only two vendors that have published at this large a scale factor, demonstrating that DB2 UDB is able to handle complex decision workloads. IBM also recently published #1 results at the 100 GB[4] and 300 GB[5] scale factors showing DB2 UDB's ability to not only handle large warehouses but also the smaller scale factors as well (see http://www.tpc.org for details).

DB2 UDB has shown that it can effectively handle "real world" workloads.

More importantly, in some people's opinion, DB2 UDB has shown that it can effectively handle "real world" workloads by publishing leadership results in various vendor benchmarks such as SAP, PeopleSoft, Baan, J.D. Edwards, i2, and Siebel.

DB2 UDB is a flexible engine with the ability to excel at both OLTP and warehousing.

The combination of TPC-C, TPC-H and application benchmarks are proof-points that DB2 UDB is not an overly generalized, less-than-capable database, but a flexible engine with the ability to excel at both OLTP and warehousing. This is a critical differentiation for DB2 UDB and a key part of its value proposition to customers. The unification of transactional and warehousing capabilities in a single database engine is also the basis for real-time analytics, a concept which – clever marketing notwithstanding -- flies in the face of the specialized warehousing-only DBMS approach.

Platform Flexibility and the Power to Choose

Flexibility enables customers to leverage their existing investments and prevent future lock-in.

IBM has a reference architecture that is used to guide our customers to their strategic warehousing solution.

To the critics who say that DB2 UDB's platform flexibility is also its weakness, we would like to point out that flexibility enables customers to leverage their existing investments and prevent future lock-in. IBM does recognize though that a large class of customers want to approach warehousing as a "solution". Thus, IBM has a reference architecture that is used to guide our customers to their strategic warehousing solution and offers pre-tested solutions of DB2 UDB, hardware, and services. These solutions are complemented by partner software where needed. A great example of this is the recently announced DB2 ICE (Integrated Cluster Environment) that is a complete solution stack of DB2 UDB on Linux with IBM eServer xSeries™ hardware. Optional components include IBM's WebSphere® Application Server™ and Tivoli® systems management software. DB2 ICE won the best of its class solution at the recent LinuxWorld Expo (see http://ibm.com/db2/linux/ice for more details on DB2 ICE).

Highlights

DB2 Information Integrator opens up several new choices for DBAs and programmers to solve real business problems.

IBM does not recommend virtual data warehouses for any customer using any technology in the market today.

Conclusion

IBM's DB2 Information Integrator adds value to the existing data warehouse in applications wide and varied such as Business Activity Management (BAM), real time data warehousing, and simple reporting across multiple databases. DB2 Information Integrator opens up several new choices for DBAs and programmers to solve real business problems. IBM does not recommend virtual data warehouses for any customer using any technology in the market today. Furthermore, we continue to invest in data transformations and business integration technology that will give DB2 Information Integrator access to more types of data with more and more sophisticated run-time transformations. As customers explore this newfound functionality, product enhancements and best practices will enable new ways to exploit the data warehouse in the overall enterprise architecture. This is the next step for extending data warehouses in the 21st century.

REFERENCES

1. http://www.metagroup.com/cgi-bin/inetcgi/jsp/displayArticle.do?oid=42390

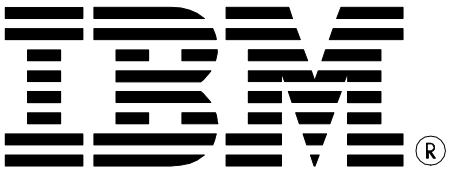
2. IBM DB2 UDB 8.1 on IBM eServer pSeries p690 running IBM AIX 5L V5.2 (160 x IBM Power4 1300MHz); Metrics: 62,214.7QphH@10000GB*, 243.00 $ US per QphH@10000GB; Available: 05/15/2003; Date Submitted: 12/23/2003

3. IBM DB2 UDB 8.1 on IBM eServer pSeries 690 Turbo Model 7040-681 running IBM AIX 5L V5.2 (32 x Power4+ 1700 MHz); Metrics: 763,898 tpmC, $8.31/tpmC; Available: 11/08/2003; Date Submitted: 06/30/2003

4. IBM DB2 UDB 8.1 on IBM eServer 325 running Suse Linux Enterprise Server 8 (16 x AMD Opteron 2 GHz); Metrics: 12,216.10 QphH@100GB*, 71.00 $ US per QphH@100GB; Available: 11/08/2003; Date Submitted: 07/29/2003

5. IBM DB2 UDB 8.1 on IBM eServer 325 running Suse Linux Enterprise Server 8 (16 x AMD Opteron 2 GHz); Metrics: 13,194.90 QphH@300GB*, 65.00 $ US per QphH@300GB; Available: 11/08/2003; Date Submitted: 07/29/2003

* The TPC believes that comparisons of results published with different scale factors are misleading and discourages such comparisons.



IBM Canada
8200 Warden Avenue
Markham, ON
L6G 1C7
Canada

Printed in United States of America
08-2003

.