Information Integration Software Solutions
March 2003

DB2® Information Management Software

Information integration

Distributed access and data consolidation

By Dr. Barry Devlin
IBM Software

Table of Contents

Introduction

Business today increasingly demands a unified view of information. Massive amounts of data are already stored throughout the business, but in a fragmented and disintegrated manner. In order to be efficient and responsive, business users need the ability to use this data transparently, wherever and however it resides, without concern for its timeliness, consistency or security. They require real, useful information. They demand that the IT infrastructure take care of all the details of how the data can be integrated into the information they need.

This demand for integration of information is not easily satisfied today. Some of the required functionality does not exist. Other enabling technology is still emerging. Available functionality is spread over many products in different technology categories.

Businesses typically use the traditional integration approaches of extract, transform and load (ETL) and replication. These approaches, here called data placement or consolidation, integrate information by physically consolidating it locally in advance of using it. An emerging approach, known as distributed access or enterprise information integration (EII) enables users to obtain direct access to data in its original locations.

The purpose of this white paper is to provide high-level guidance on what functionality exists in these areas, what products are available, where they play and how they work together. This is done in the context of IBM's overall vision of information integration, which is outlined in the first section of this paper.

Second, the paper describes and contrasts the two major approaches to integrating information, distributed access and data placement or consolidation, showing where one method may be preferred over the other and where they can work together.

Finally, it shows how current EII, ETL and replication products fit within the information integration framework. IBM's and partner products in these areas are introduced and positioned relative to one another. Examples show where a combination of EII, ETL and replication tools is required to solve the business needs.

IBM’s vision of integrated information

Today, any but the simplest of business tasks requires the use of information from the variety of data sources that businesses have built over many years. These sources may be local or remote, on the intranet, extranet or Internet. The data may be stored in any of a variety of formats—relational or non-relational databases, flat files, unstructured content stores, etc. The data may be current or point-in-time copies. Often, the users need not only to read from these sources but to write to them as well.

This complex, and indeed, continuously changing, information environment presents significant challenges to business users and applications as well as to the IT people who must maintain and manage it. Put simply, IBM’s vision of information integration is to significantly reduce or even eliminate these issues.

The underlying principle of information integration is that users should be able to see all of the data they use as if it resided in a single source. Information integration technology shields the requester from all the complexities associated with accessing data in diverse locations, semantics, formats and access methods. Using a standards-based language such as structured query language (SQL) or extensible markup language (XML) through XQuery, for example, or a standard Web services or content API, information integration middleware enables users, or applications acting on their behalf, to access information transparently, without concern for its physical implementation.

Alternative methods of integrating information

The goal of providing an integrated view of information can be achieved in two ways, either separately or in combination. The primary methods of information integration are:

1. Providing distributed access to data through data access or federation
2. Moving the data to a more efficient or accessible location—data consolidation or placement.

Together, these sets of functionality form the heart of what is required to integrate information. In the simplest terms, federation takes a query in one location and distributes the appropriate parts of it to act upon the data wherever and in whatever form it resides. Distributed access corresponds to the enterprise information integration (EII) category of technology. Data placement, on the other hand, brings together data from a variety of locations into one place, in advance, so that a user query does not always need to be distributed. This approach corresponds to extract, transform and load (ETL) and replication functionality.

Both approaches require extensive and largely common supporting functionality. Neither distributed access nor data placement could exist without mapping and transformation functionality, which ensure data integrity. Furthermore, depending on the business requirement, the same data may need to be consolidated in some cases and federated in others. Therefore, a common set of transformation and mapping functionality is required in both cases to maintain consistency across the data used by the business.

Mapping provides the ability to understand the relationships between different pieces of data, and the ability to determine, for example, that the columns in a virtual table actually correspond to the attributes of a different data source. In a more complex example, mapping relates "customer" in a query or application to a combination of data, some residing locally in IBM DB2® Universal Database, some residing remotely in Oracle and a third part in an IBM Lotus® Notes® database.

Transformation is the functionality that actually converts and combines the data related through a mapping between different representations. For example, a simple transformation could convert a number to a string, while more complex transformations include converting hierarchical to relational data, and even conversions based on elaborate business rules.

Caching provides a temporary data store that can improve the performance of federation by transparently storing a local copy of a result set. From the viewpoint of data placement, it is simply a store containing a possibly transformed copy of some remote source, which might need to be managed. The cache is thus another key link between the two modes of integration, allowing the strength of one mode to support the weakness of the other.

Mapping, transformation and caching depend on a detailed description of the environment in which they operate. Such description includes business meaning, relationships, location, technical format and so on. In short, metadata. Such metadata must be both comprehensive and consistent, and be useful from the discovery and definition of an integration project right through to the operation of a federated query. A comprehensive and logically consistent set of metadata, whether materialized in a single physical store or distributed across multiple stores, is the fundamental basis of any vision of information integration.

Distributed access and data consolidation

The functionality required for full query federation is, of course, far more complex than that. The vast majority of data does not reside in relational databases. Such data is accessed through a variety of languages or even applications, which seldom provide the rich set functions available in SQL. As a result, federation must simulate the required functionality, participate in query optimization, push down sub-queries to the source, mediate between the different access mechanisms and manage security and transactional integrity as needed. Federation is not confined to read-only access; many business tasks also mandate distributed write access. The federated query in its most complete form must handle all of these requirements.

The data consolidation or ETL/replication approach. Data consolidation or placement is the traditional approach to integrating information and, in contrast to federation, moves the data to the query. It has always been considered less complex than federation, as data consolidation creates a second, local copy of the data, pre-processed as required, thus reducing the need for extensive data manipulation and remote access within the user query. Data consolidation, because it operates off the critical time path of the user's query, also allows for substantial and complex transformation of the data to address issues of cleanliness, semantic and temporal consistency and so on. It therefore exhibits varying levels of complexity. At its simplest, it is a manually initiated database unload followed by a load of the target system. At its most complex, it may involve the automated, real-time, multi-way synchronization of databases on a number of remote systems. In most cases today, it is somewhere in between.

A key consideration for data consolidation is the maximum latency that can be tolerated when transferring the data from source to target. Typically, business needs specify how up-to-date a copy of the data must be. In data warehouses for example, the frequency required might be daily or weekly and the latency of data consolidation can easily extend to many hours. At the other extreme, the need for almost real-time data, such as in stock market systems, requires minimum latency in data consolidation.

Two of the most important factors determining the minimum latency possible in data consolidation are the complexity of the transformations required and the volumes of data to be transferred. These factors lead to two complementary approaches to consolidating data. ETL is optimized for larger data volumes and is often associated with more complex transformations, while replication emphasizes the transfer of individual data records and is often restricted to simpler transformations.

Comparing federation and data consolidation

Federation and data consolidation are actually similar concepts. Both involve requesting and receiving data that originally resides outside the physical confines of the database with which the users interact. The key difference is in the timing of these requests from and transfers of data to that database. With federation, both occur after the user issues his or her request. In data consolidation, both occur beforehand and, in practice, the request occurs once during transfer definition while the transfer may subsequently occur many times.

From the end users' point-of-view, or that of the applications acting on their behalf, federation and data consolidation act in opposite ways. Federation integrates the required information on the fly, directly from its original sources, acting only after the end user decides what n th

It is sometimes suggested that federation can be employed to provide users direct, unplanned access to any data, anywhere because a federated query allows data to be combined on the fly. This is a dangerous myth, because such unrestrained access can lead to significant problems for both the users and the IT systems. Federation actually demands even more rigorous analysis, modeling, control and planning than data consolidation to avoid significant performance and semantic problems for users if they try to combine data that is inconsistent in meaning and structure. Consolidated data stores such as data warehouses and operational data stores constructed through data placement avoid such issues and will thus continue to play an important role.

With these considerations in mind, it’s easy to discern when federation or data consolidation is the appropriate method of information integration.

Federation, or distributed query, is the appropriate method when:

- *Real-time or near real-time access to rapidly changing data is required.* Making copies[1] of rapidly changing data can be costly, and there will always be some latency in the process. Through federation, the original data is accessed directly and joined in the query. However, the performance, security, availability and privacy aspects of accessing the original data must be considered.
- D*irect immediate write access to the original data is required.* Working on a data copy is generally not advisable when there is a need to insert or update data as data integrity issues between the original data and the copy can occur. Even if a two-way data consolidation tool is available, complex two-phase locking schemes are required. However, writing directly to the database of another application is generally prohibited. Therefore, it is generally recommended to call the owning application via an API through the federated query to request any updates.
- *It is technically difficult to use copies of the source data.* When users require access to widely heterogeneous data and content, it may be difficult to bring all the structured and unstructured data together in a single local copy. Or, when the source data has a very specialized structure, or has dependencies on other data sources, it may not be possible to sensibly query a local copy of the data. In such cases, accessing the original source is recommended.

[1] The term “copy” is used loosely here and in the following discussion. It may or may not imply an exact copy; for ease of use, performance or other reasons, the data may have been substantially transformed during the data consolidation process.

- *The cost of copying the data exceeds that of accessing it remotely.*
  The performance impacts and network costs associated with querying remote data sets must be compared with the network, storage and maintenance costs of storing multiple copies of data. In some cases, there will be a clear preference for a federation-based approach, such as when:
  – Data volumes of the original sources are too large to justify copying it
  – Data is too seldom used to justify copying it
  – A very small or unpredictable percentage of the data is ever used
  – Data is accessed from many remote and distributed locations, which would imply multiple copies.
- *It is illegal or forbidden to make copies of the source data.*
  Creating a local copy of source data that is controlled by another organization or that resides on the Internet may be impractical, due to security, privacy or licensing restrictions, but federated access with discrete queries may be allowed.
- *The users' needs are not known in advance.*
  Allowing users immediate, ad hoc access to necessary data is an obvious argument in favor of federation. However, this can be a risky strategy, because of the potential for users to create "queries from hell" that negatively impact both source system and network performance and give poor response times. In addition, because of the current level of semantic inconsistencies across data stores within organizations, there is a high risk that such queries would return answers that are little more than nonsense!

In many ways, the arguments for data consolidation are the exact opposite of those for federation; however, there are some additional specific indications:

- *Read-only access to reasonably stable data is required.*
  Creating regular copies from the data source can hide the ongoing churn of information in cases where users don't want to see every change that occurs in the source data.
- *Users need historical or trending data.*
  Historical and trending data is seldom available in operational data sources, but can be built up over time through the data consolidation process. This is a very common data warehousing requirement.

- *Data access performance or availability are overriding requirements.* Users routinely want quick data access, necessitating that a local, pre-processed copy of the data be made available. As seen in data warehousing environments, these queries can be very complex or a multidimensional view of historical or trending data is needed. As a result, consolidation is a fundamental technology in data warehousing. However, as described in the IBM white paper *Information integration—Extending the data warehouse*, federated query can be effectively used for specific needs, such as accessing real-time data.
- *Users' needs are repeatable and can be predicted in advance.* When users' queries are well-defined, repeated and require access to only a known subset of the source data, it makes sense to create a copy for local access and use. This is particularly true when a specific set of users needs a view of the data that differs substantially from the way the data is stored in the source. In this case, data consolidation can create a specialized copy of the data which performs well and is more easily understood by the users.
- T*ransformations or joins needed are complex or long-running.* In cases where joins or transformations are complex or long-running, it is inadvisable to have them run as part of a user query due to potentially poor performance and high costs. In such cases, creating a copy of the data through data consolidation makes more sense.

Successfully combining federation and data consolidation approaches

There are times when it makes sense to combine the strengths of federation and data consolidation.

The first case is where federated query can leverage data consolidation functionality under the covers. There are situations in which a federated query, even in the simplest case described previously, simply will not work. Network performance or availability issues, may prevent the query from running. Allowing an application direct access to a remote data set may be politically or technically impractical. Or, the underlying data may be of a structure that would unacceptably slow the performance of such a query.

In such cases, federation can use data consolidation to create or manage cached data in support of such needs. Today, system designers can explicitly specify and define these data caches in advance. In the future, information integration technology could automatically configure them as required. Clearly, this approach limits the federated query in some ways, as the cached data is not fully current and write access may be limited if the cache does not fully support two-way synchronization. However, in other respects, using data consolidation underneath federation does expand the solution set that federation can provide.

Conversely, how can data consolidation benefit from federation? One of the challenges faced by any data consolidation tool is the number and variety of sources and targets with which it must exchange data. In general, therefore, data consolidation tools are often optimized for a subset of the available sources and targets. Placing federation behind data consolidation expands this set of sources and targets as well as allowing pre-joining of data from multiple sources. The trade-off here is clearly performance, in exchange for the number of sources or targets that can be accessed, and the approach may be restricted to cases where limited quantities of data are involved.

Federation and data consolidation tools

IBM and its Business Partners deliver a variety of tools providing federation and consolidation functionality. However, it should be recognized that no single product delivers all of the functions envisaged above. Some products address one area of functionality, others focus on another, and some types of function are still emergent.

EII

EII tools provide distributed access to remote data and correspond to the distributed access or federated query functionality described earlier. This tools category is becoming increasingly important as distributed e-business demands extensive and immediate access to dispersed data for both read-only and read/write purposes.

IBM DB2 Information Integrator, along with its predecessor products—IBM DB2 DataJoiner® and IBM DB2 Relational Connect—provide EII functionality. DB2 Information Integrator allows definition of and querying across an integrated view of diverse and distributed data on a variety of databases and file formats, including:

- DB2, Informix, Oracle, Sybase, Microsoft SQL Server, and Teradata databases as well as ODBC sources
- XML, Microsoft Excel, OLE DB and flat files
- Web services, message queues and application data sources
- IBM Lotus® Extended Search sources (content repositories, LDAP directories, Web, email databases, syndicated content, etc).

Caching functionality is an important requirement in EII. Caching typically implies very low or no latency, automated transfer of data with minimal transformation between sources and target cache. In the longer term, however, this concept requires the ability to replicate data in both directions and to resolve conflicts between source databases and the target cache. A Materialized Query Table (MQT) is a basic form of cache that is already available, where the results of queries are stored for later reuse. DB2 Information Integrator can use MQTs, but setup and management of these caches still requires manual intervention.

DB2 Information Integrator for Content, as well as its predecessor offering, IBM Enterprise Information Portal, also offer EII functionality, however tailored more toward content integration and based on the IBM Content Manager programming model. It provides search and access across the Content Manager products and other content repositories, Lotus databases, relational databases and wide-ranging content available with Lotus Extended Search. It includes a sophisticated information mining capability that uses Web crawling and text-mining algorithms to provide structure to unstructured content. And DB2 Information Integrator for Content provides an advanced workflow application to enable businesses to increase productivity, reduce production times and improve communication and collaboration.

Extract, transform and load (ETL)

ETL tools are characterized by their ability to transfer large volumes of data efficiently, often with sophisticated transformations, between source and target. They are normally used to populate data warehouses and data marts, where low latency is typically not a strong requirement and complex transformations are required to make the data suitable for end users.

DB2 Warehouse Manager provides ETL function primarily among members of the DB2 family. It uses the full range of SQL function, Web services, plus a rich set of pre-built transformations to provide transformation between source and target and adopts an agent approach in order to distribute processing over multiple nodes. IBM DB2 Warehouse Manager also provides an example of the ability of an ETL tool to use federation under the covers, by linking to DB2 Information Integrator as one of its sources of data.

Support for high-volume ETL from a wide variety of heterogeneous sources is provided through a Business Partner product, Ascential DataStage from Ascential Software. There are versions of the product that run on the IBM 390® platform and on the Microsoft® Windows NT® and Unix® platforms. DataStage provides extensive support for data transformations as well as recent add-ons to provide data quality and cleansing support.

Both products are traditionally focused on the data warehousing market; however, data warehousing is being driven inexorably towards more real-time and distributed data needs. As a result, these tools can be considered as a part of the broader scope of integrating information, providing broad data consolidation function. In particular, their extensive transformation and bulk data handling capabilities are key components in any wide scale information integration solution.

Replication

Replication tools aim to provide a low latency transfer of data from source to target, through the capture of source database changes, the transfer and application of those changes to the target system. With a focus on low latency, transformations are often limited in complexity. Replication is often used to populate operational data stores (ODS) or to periodically distribute data between a central operational database and multiple, remote subsets of that database that operate autonomously between updates.

IBM DB2 DataPropagator™ provides replication functionality between members of the DB2 family while DB2 Information Integrator provides replication among mixed relational databases such as DB2, Informix, Oracle, Sybase and Microsoft SQL Server. Both DB2 DataPropagator and DB2 Information Integrator support data transformation through the use of the full set of SQL functionality, Web services and stored procedures.

ETL tool vendors today are increasingly offering replication functionality, often under the term "changed data capture". As in the case of the ETL products, replication tools can also provide a subset of information integration data consolidation functionality. Here the emphasis is on the ability to address immediate or near real-time data transfer needs rather than complex transformation or bulk data handling.

Combining tool sets to solve business problems

Having explored the conceptual functions of information integration and introduced today's federation and data consolidation tools, it's appropriate to look at a few examples that illustrate how they work together to solve business problems.

Improving call center performance

The IBM white paper *Information integration—Extending the data warehouse* discusses the advantages to be gained in a call center environment by allowing queries to be federated across the warehouse, which provides historical information for the agents, and the operational systems, which record the current status. In this example, federation allows access to up-to-date account information when required from the operational environment. However, in such an environment, query performance is difficult to guarantee.

Over time, it may become clear that certain items of information from the operational systems are requested with some regularity. For example, it may be that some of the biggest customers regularly query payment status. While this is data that is not deemed worth storing in the warehouse in near rea-ltime, the creation of an ODS updated every few minutes may allow rapid access to this data. DB2 DataPropagator can be used to regularly update the ODS with data from operational systems based on DB2. Meanwhile, DB2 Information Integrator replication functionality can do the same from other relational operational systems, which become yet another source of data for the federated query.

Expanding data warehouse sources

DB2 Warehouse Manager is typically used to provide ETL functionality in fully DB2 warehouse environments. DB2 Information Integrator expands its scope to handle native extraction of information from non-IBM relational databases.

Supporting a distributed network of operational systems

A system model that is often seen in retail and other distributed environments involves a central master operational system and a number of intermittently connected satellite systems running in remote locations. In this mode of operation, the central system sends updates to the satellite systems at some time, which later return updated data to the central database. For example, in the retail environment, the central system sends price and promotion updates to all the stores early in the morning before they open and the stores send point-of-sale information to the central system at the close of business for the day. The characteristic feature of this model is that although updates flow in both directions, there is no possibility of conflicts because the flow is in different directions at different times and/or the updates affect different parts of the databases.

Suppose now that the central database runs on DB2 and the satellites are Oracle implementations. DB2 Information Integrator captures the changes in the central DB2 price and promotion database applies them to the Oracle database price and promotion tables. The point-of-sale systems run during the day using these base tables. At the close of business, these sales are summarized in various ways and these changes are captured by the replication capability of DB2 Information Integrator and applied to the central system.

In this case, replication and federation work together to synchronize this distributed environment. In addition, DB2 Information Integrator can also be used to facilitate business analysts tracking product sales. Here, the analysts have access only to summary information on the central system but if they need to drill down to detailed information, DB2 Information Integrator can take their queries and seamlessly route them to the Oracle databases where the detailed information resides.

Conclusion

The vision for information integration: in tomorrow's world, business users will neither know nor care about data, its format or its location. Rather, the information they need will be at their fingertips in a form they can immediately understand and use. The complexity of systems and stores, content and format, integrity and performance will all be managed by information integration middleware. Not only will this completely insulate business users from underlying system complexity; it will simplify and at least partially eliminate this complexity. It will automate the management of this environment, shielding IT from excessive intricacy.

Such a breakthrough will not be achieved overnight. However, today's information integration tooling does provide a good starting point. The DB2 Information Integrator federation function creates the appearance of a single, integrated repository of information for the business user or requesting application. In short, it hides original location and format behind the facade of a standard query interface.

But we know that it is neither possible nor desirable to always leave the data in its original format or location. Usability, simplicity and performance all dictate that, sometimes, data must be moved or copied, mirrored or transformed from its original form, either temporarily or permanently. Today's ETL and replication tools provide the basis for such functionality.

ETL products such as DB2 Warehouse Manager and Ascential DataStage emphasize unidirectional, bulk data transfer with complex transformations. Replication products such as DB2 DataPropagator and DB2 Information Integrator support the efficient transfer of database changes or transactions with minimal transformation. Database caching technology, such as materialized query tables in DB2 Universal Database and DB2 Information Integrator, demonstrates yet another aspect of this need to create local copies of remote data.

Individually and collectively, as we have shown, today's EII, ETL and replication tools can be effectively used in a variety of information integration scenarios. Each tool provides useful and necessary function that is a component of many information integration implementations.

None of these tools alone provides a complete realization of the information integration vision; however, each is a building block. As we extend and assemble them, we will establish, over time, IBM's vision of a truly integrated information environment.

For more information

Please contact your IBM marketing representative or an IBM Business Partner, or call 1-800 IBM CALL within the U.S. Also, visit our Web site at:
ibm.com/software/data/integration

IBM United States
Software Group
Route 100
Somers, NY 10589
U.S.A.

Produced in the United States of America
03-03

References in this publication to IBM products or services do not imply that IBM intends to make them available in all countries in which IBM operates.

♻

Printed in the United States on recycled paper containing 10% recovered post-consumer fiber.

e business software