White Paper

Bloor

IBM Informix in hybrid workload environments

A White Paper by Bloor Research
Author : Philip Howard
Publish date : August 2012

... for hybrid environments Informix has a number of unique capabilities that cannot be matched by either conventional data warehousing vendors or traditional data warehouses

Philip Howard

Executive summary

Informix, one of the databases offered by IBM, has a long and illustrious history. In particular, it is well known as a platform for supporting applications developed by third parties and ISVs, primarily for use in transactional environments. This is because Informix is a robust offering that combines high performance and functionality with minimal administrative requirements. However, this paper does not discuss these features of the product in any detail. Instead, we are here concerned with the use of Informix either in a hybrid transactional/analytic environment or for pure data warehousing applications, especially those involving 'big data' analytics based on sensor and other machine-generated data.

In this paper we will start with a brief discussion of various types of analytics and data warehousing, and then consider the sort of facilities that are required to support such environments, along with a consideration of the features of IBM Informix that lend themselves to such deployments. We will conclude with a discursion into areas where Informix has special abilities, not generally available in other database products, which lend themselves to specific types of analytic applications.

Analytics

To start, we need to distinguish between analytics and analytic applications. The difference between these two is that the latter provides a packaged approach whereas the former is more tool-based. As a result, analytic applications tend to be more user-friendly and more suitable for use by business users as opposed to those that are more IT savvy. Given its history of supporting third party vendors, IBM Informix is especially targeted at analytic applications although, of course, it supports business intelligence and data mining tools such as IBM Cognos and SPSS as well as products from other suppliers. It will often be the case that analytic applications have ad hoc query capabilities built-in to complement the various pre-designed queries, reports and other functions that form the mainstay of the application.

Leaving aside enterprise data warehouses, whose characteristics are well known, there are three types of environment we need to discuss:

1. Hybrid environments. Analytic applications are normally thought of as separate and distinct from transactional applications. However, that doesn't have to be the case. For example, it could easily make sense to combine a CRM solution with marketing optimisation or campaign management, a purchase ledger application with supply chain optimisation, or sales order processing with customer lifetime profitability analysis. Unfortunately, while this may make sense from a logical (and business) perspective it requires a database platform that has been optimised to support such hybrid environments and the truth is that most databases can't do this within a single environment. The big advantage, of course, is that you only require a single system with a single hardware implementation if you can support a hybrid environment, which potentially means reduced up-front costs. You also don't need to move data from one environment to the other and there is no need to synchronise the data. Furthermore, a single set of administrative requirements means lower on-going costs, but it is this that represents the catch for most database products because the majority of relational databases that are suitable for both transactional and analytic operations on a separate basis require very different tuning and administrative requirements to optimise performance for both environments in conjunction. Thus it becomes impractical to implement such a hybrid system. However, as we shall see, IBM Informix is an exception to this general rule.

2. Analytic applications. Of course, the advantages of hybrid transactional/analytic environments do not necessarily mean that this is always the best approach. You may, for example, want to perform analytics against transactional data that has been derived from multiple transactional systems either because your analytics needs to combine data from different types of source system (as an example, perhaps combining customer data and service data) or because you are bringing together data from geographically dispersed systems. Analytic applications may be deployed in one of two ways:

   a) As a single application, typically for a department within a large organisation. An example would be revenue assurance with telecommunications. These are sometimes called edge applications.

   b) In conjunction with other analytic applications, either to support the whole of a small or medium-sized business or a department requiring multiple related applications. An environment such as this would be regarded as a data mart in the case of a department or perhaps as a data warehouse for a small or medium-sized company. Unless all of the analytic applications involved form part of a suite it is likely that an environment such as this will require business intelligence or data mining tools in addition to the analytic applications.

Analytics

3. Breakthrough applications. These represent a different type of hybrid environment. However, whereas traditional hybrid environments combine transactional data with the analysis of that data, so-called breakthrough applications combine sensor-based or machine-generated data into applications that have both operational and analytical implications. An example is smart metering, which we will discuss later. Another common example is in motor insurance, where companies are increasingly installing telemetry into vehicles to monitor the quality and speed of driving, which can lead to a bonus for the driver if he or she drives well. Breakthrough applications often include a real-time or near real-time element (in the example quoted both in the collection of the data and in the ability to inform emergency services if an accident is detected [sudden deceleration followed by lack of movement]) and may also require geospatial data and/or time stamps (where it is important to know the time at which something happened).

Key features

There are three key features of IBM Informix that lend themselves to hybrid and analytic environments, which we will discuss in turn. However before doing so it is worth emphasising that all of these features are part of a single platform that can implemented within a single instance of the database.

Flexible Grid

For those requiring high availability and/or scalability IBM offers the IBM Informix Flexible Grid together with High Availability Data Replication, Remote Secondary Standby Database Servers and Shared Disk Secondary Servers. While the last three of these do what they say on the tin the Flexible Grid may need some explanation. The Flexible Grid supports the definition of a multi-node heterogeneous cluster that makes it possible to run an application on any node within the grid. This means that you can have a geographically dispersed environment with different commodity hardware (and operating systems) implemented in different locations, according to need, and yet have the whole environment centrally controlled. Not only are DML operations replicated, but so are DDL operations. This means that when a table is created on one node, it is automatically created on all nodes within the Flexible Grid. In addition, it is possible to automatically start replication of the table as soon as it is created, which greatly simplifies the deployment of a replication solution. A major feature of the Flexible Grid is that it supports continuous availability. That is, operations can continue regardless of whether downtime is planned or unplanned. For example, Game Show Network uses the Informix Flexible Grid and has had no unplanned (or planned) downtime for two years. According to Susan Marciano, Vice President of Technical Operations "the Flexible Grid feature of Informix enables us to perform rolling upgrades without any outages, so players can go on playing with no interruption and no impact on our revenue. That's worth its weight in gold." Other use cases occur wherever 24x7 operations are critical (call centres in, for example, the insurance sector) and, especially, where costs are a major factor, since the Flexible Grid runs on commodity hardware. Minimal administration is a further major benefit.

Warehouse Accelerator

IBM offers various Warehouse Editions of Informix. These all include the Informix Warehouse Accelerator to support analytics. This is an extension to the normal database used for transactional purposes. It enables query processing in-memory and provides a column-based approach to avoid any requirement for indexes, temporary space, summary tables or partitions. In other words it is entirely suitable for supporting analytic applications because the lack of these features means that administration is both minimised and consistent across transactional and analytic environments. Moreover, the Warehouse Accelerator can be implemented on the same system as the relevant transactional environment. When this is the case you use Smart Analytics Studio, which is a graphical development tool, to define the data (and its schema) that you want query and the Warehouse Accelerator will automatically offload this data, which is now stored separately from the OLTP environment. It is processed in its own memory space so that there is no conflict with the operational aspects of the environment and transactional performance will not be impacted. Note that there is no need to change your existing business intelligence tool(s).

There are a number of other features worth mentioning:

The Accelerator uses vector processing. This is a form of processing that takes advantage of modern day CPU characteristics, which is orders of magnitude faster for computationally intensive tasks, which analytics frequently are.

- The database optimiser has been specifically optimised to support both transactional and analytic workloads where a hybrid environment is being deployed. It is also worth noting that the optimiser knows what data is in the data mart and what is not. The optimiser determines whether the query can be satisfied by the Accelerator and, if so, it routes the query there. If not, it will choose to execute the query within Informix. Now, if a query saves the result into a temporary table as part of the Select statement, as is often done by certain BI tools, then the Accelerator can speed up that portion of the query.

Key features

- In-database analytic capabilities are available from Fuzzy Logix, which has ported its library of analytic and statistical capabilities to run with the IBM Informix database.

- Informix itself uses the same deep compression technology as DB2, which provides benefits both in terms of storage capacity and performance. However, the Warehouse Accelerator uses a different approach: in the process of loading data into the Warehouse Accelerator, the software takes a snapshot of the data in the Informix database and uses its own proprietary encoding method (approximate Huffman encoding). This has the advantage that it allows predicate evaluation without having to decompress the data.

A public sector entity in Peru is a user of the Warehouse Accelerator. It had previously used Informix for OLTP purposes and had a separate data warehouse. However, in the case of analytics for things such as tax fraud the warehouse was too slow, so the organisation looked at replacement warehouses as well as the Informix Warehouse Accelerator and ultimately chose the latter. For some queries the organisation has reported that the environment is now as much as 100 times faster than previously. Moreover, there is no ETL to worry about, no transformations, no cube building or aggregations. Some Warehouse Accelerator customers have reported even greater performance gains. For example, one US retail organisation has reported performance gains of more than 750 times. This is probably exceptional but other reports of 36 times (US government agency) and 80 times (European government agency) mean that one to two orders of magnitude would seem like a good rule of thumb, although of course it depends on your existing environment.

Some Warehouse Accelerator customers have reported even greater performance gains. For example, one US retail organisation has reported performance gains of more than 750 times.

Time Series

There are many applications that require an understanding of time. For example, in capital markets you need to know when a trade was initiated and when it was completed. In smart metering (which we consider in more detail below) and other sensor-based environments you will be taking measurements on a regular basis and you need to know what value was recorded and at what time. Bearing in mind that measurements are taken every 15 minutes, say, it would be wasteful to record a time stamp for every single measurement: it is more efficient to simply store the start date and time, and record what the time interval between measurements is. This is what is known as a time series and it is a native capability of IBM Informix. It saves disk space and will provide better performance characteristics when querying the data. Moreover, we know of no other transactional (relational) database that has native time series capability.

We know of no other transactional (relational) database that has native time series capability

Requirements

If you want to develop or run analytic applications then you need to think about the requirements of the platform that will support that application and you also need to think about the data warehousing environment in general, since you will need to support ad hoc enquiries, for example, for which performance will be a major consideration. We will look at each of these requirements in this section and consider how IBM Informix meets these. Note that even if you are simply considering a general-purpose data warehouse rather than developing or implementing one or more analytic applications, then many of these requirements will be 'nice to haves' even if they are not actually requirements.

Requirement 1: fire and forget

End users do not know about or want to know about the database that underpins their application. They certainly do not want any involvement with database administration. It is thus essential that any database underpinning an application is invisible to the customer and remains that way. This is as true for transactional environments as it is for analytic applications. However, there is an additional consideration when it comes to analytics. This is that, in order to get good performance (see requirement 3) for query-based applications you need, at least in conventional environments, to create indexes, materialised views and other such database constructs in order to achieve that performance. (Or you can throw a lot of hardware at the problem but this would countermand requirement 5—see below). Now, this is perfectly feasible. However, it is not flexible (see requirement 6) in the sense that this may well not support adequately performing ad hoc queries against data that you did not index in the first place. Moreover, every time you add functionality to your application you will need to change the supported indexes. Worse, different workloads may mean that different indexes, materialised tables and so forth will be differently suitable for different customers. Moreover, these workloads may change over time. What this will mean is that the database will need to be tuned on an ongoing basis in order to maintain performance. The customer does not want to do this and does not, in any case, know how. For all of these reasons a traditional relational database will not be suitable for embedding as an analytic engine, precisely because these all require exactly this sort of tuning and, moreover, different tuning regimes for transactional environments, which makes hybrid implementations difficult if not impossible. The opposite is true of the Warehouse Accelerator (see above) because it is column-based, requires no indexes and runs in its own memory space.

Requirement 2: ease of implementation

There are two aspects to this, the first being that the database should be easy to install in the first instance and, where applicable, that the resulting application, with its underlying database, is similarly easy to install and implement at the customer site. In particular, there should be no requirement for the end user to configure any of the database elements during the implementation process. The fact that IBM Informix is already widely used by ISVs and other third party providers strongly suggests that the product meets this requirement.

Requirement 3: high performance

While high performance is important for all environments this is arguably even more critical for analytic implementations than for transactional ones. The reason is twofold. In the first place, it is not easy to predict which particular analyses users are going to want to run at any particular time, so you have to be able to cater for multiple instances of the most onerous and complex queries and offer performance to match. In a transactional environment, on the other hand, you know (roughly) the volume of throughput that has to be catered. Secondly, the number of transactions will grow as the business grows and this should be reasonably predictable. However, when it comes to analytics there is not only more and more data available to analyse there are also new types of data that it might be appropriate to include in queries. While the former can be estimated it may be very difficult to guess at the performance requirements associated with new sources of information. While, to a certain extent, this will be resolved through scalability (see requirement 9) you do not want to be constantly expanding your system and any solution should have sufficient headroom in terms of performance so that scaling is an infrequent requirement.

Requirements

Requirement 4: high availability

High availability (see the previous discussion on the IBM Informix Flexible Grid) is always a potential requirement whenever an application is deemed to be mission critical. The question, of course, is whether analytic applications are regarded as mission critical, and the answer is that it depends on the application and the user. For example, if you have a real-time requirement for say, security event monitoring, then high availability is likely to be essential. On the other hand, if you are using analytics to support some sort of customer intelligence application then whether or not you will require high availability will depend on how important that customer intelligence is perceived to be. The conclusion therefore must be that in some environments it is a must have while in others it is a nice to have. Given that in the latter category there will always be some customers that would like the high availability option then it effectively means that this must be offered by the database provider so that you can at least offer the option.

Requirement 5: low cost

The requirement for low costs, subject to meeting all our other rules, is obvious. Moreover, this applies not just to the license fee to the software provider but also to the running costs, footprint and hardware requirements that the system has. Note that hardware requirements include consequential needs such as cooling. It will be clearly be advantageous if the database will run on low cost commodity hardware that is as green (in terms of heat output and power requirements) as possible.

One specific aspect of database technology that relates to technology cost is compression (which also improves performance), because this will directly affect the amount of disk space that you will require and, in the case of IBM Informix, how much data can be fitted into memory. Compression rates are variable, depending on the type of data to be compressed. Typical rates provided by the Warehouse Accelerator are around 3:1 but IBM reports that it has seen figures as good as 5/6:1. The product uses what IBM calls deep compression, which uses a form of tokenisation and, of course, it is much easier to compress columnar data efficiently (because each column has the same datatype).

Requirement 6: flexibility

Do you want to offer your customer an environment in which he can only query what you have pre-prepared for him or do you want to allow him to make ad hoc or train-of-thought enquiries that go beyond the prescriptions of any particular application? While you will clearly want to offer as comprehensive a set of out-of-the-box analytic functions as possible, nobody can predict what every possible customer will think is worth enquiring about. We therefore believe that to offer as much flexibility as possible will be appealing. However, this means that the analytic database must support such flexibility.

In theory, you can ask any question of any database. However, in practice, complex queries (see requirement 8), run against large datasets that have not been pre-designed, will often take a long time to run or will time out unless you have the right sort of underlying architecture. In particular, databases that depend on indexes, materialised views and pre-calculated aggregates (amongst other things) will slow down to the point of being unusable if ad hoc queries are required against data for which these things have not been pre-defined. To support more flexible analytic processes you will therefore require a database that does not require these constructs, such as the Warehouse Accelerator.

Requirement 7: loading

There are two circumstances in which the loading capability of the database will be relevant: either because you have large amounts of data to be loaded or because you need to load data in real-time or near real-time. In some circumstances it may be that you have both issues to contend with. On the other hand, in some environments, loading may not be a major issue at all. Where it is a concern, you will need a product that supports a high ingestion rate and (near) real-time capabilities or both. It is worth noting that the latter supports the former: if you need to load 12Tb of data per day it requires much less capability by way of ingestion if you do this on a real-time basis at 500Gb per hour than if you do it on a batch basis with a batch window of two hours (if you have that much time available) that requires a load rate of 6Tb per hour.

Requirements

In terms of raw loading capacity this is simply a question of the size of the pipe into the database, bearing in mind any parallelism that is provided. On the other hand, being able to support real-time loading requires support for the ability to micro-batch data (say, batches of one minute) or explicit trickle feeding mechanisms such as change data capture or streaming capability. In most instances the first of these will be fine but if you want genuinely real-time capability then you will need the latter unless (and we have never seen this) you can micro-batch at the second or sub-second level. Of course, IBM is a market leader in data integration technology but the IBM Informix Warehouse Editions include a built-in ETL (extract, transform and load) capability so that you do not have to license separate technology for this purpose.

Requirement 8: complex analytics

By complex analytics we do not necessarily mean that the questions that customers want to pose are complex but that they are complex at the database level. While there is no formal definition of what constitutes a complex query they typically involve such things as multi-way joins, whole table scans, correlated sub-queries and other functions that are either compute intensive, I/O intensive or both. Your solution has to be able to perform such queries in a timely manner and you will therefore require a database product that can cope with such a workload, also bearing in mind that these queries may be ad hoc (requirement 6) and must perform to expectations (requirement 3). While there is not space here to discuss every combination of events that may be involved in complex analytics, a good starting point is the need for whole table scans. This is a query killer. If you have to do whole table scans against large sets of data and, worse, have to do so repeatedly because of multi-way joins then your analytic performance will die. Facilities to avoid whole table scans should therefore be considered de rigueur. In fact, this is one of the major advantages of using a column-based approach to supporting analytics because using columns avoids the need for whole table scans.

Requirement 9: scalability

If there is one thing that is certain it is that the data that your customers want to analyse will grow. Even without considering 'big data', research conducted by Bloor Research indicates that the amount of data that customers are storing roughly doubles every five years. Whatever solution you choose needs to be able to easily scale as data requirements grow, which is, again, where the IBM Informix Flexible Grid fits in. However, it is not just a question of being able to store larger amounts of data. That data also needs to be loaded, so loading capability needs to be scalable also. Moreover, it is likely that more queries will be run by more users, as the value of your analytic application or platform becomes apparent to the customer, so the database will also need to be scalable in terms of the user and queries that it can support.

Breakthrough applications: smart metering

As we have noted, so-called breakthrough applications are hybrid applications that involve sensor-based or machine-generated data as opposed to transactional data, though this sort of data may also be included. For example, smart metering applications, which we will discuss in a moment, may well need to process conventional customer data as well as data retrieved from the meters. However, before we discuss this it is worth noting that IBM Informix is the only relational database, as far as we know, that natively supports time series within a transactional environment. It is true that data warehousing products often support time series analysis but the support is not native (that is, it is not embedded in the database engine per se) and doesn't apply to operational environments so IBM Informix is uniquely well-placed to capitalise on demand for time-series based breakthrough applications. That is, any environment, where it is important to know when things happened as well as what happened. The most obvious example of this is smart metering, which also requires location-based (geospatial) information.

Smart meters are increasingly being deployed by energy and utility companies and relevant applications have both operational and analytic requirements. In the former case, for example, you need to recognise outages and handle them in a timely manner, while in the latter case you need analytics for planning purposes (for example, at what times of day do we need to ramp up power production), marketing (what incentives can we provide for more efficient use of resources) and fraud detection, to quote just some examples. Typical solutions for smart metering applications require two different databases—one for operational purposes and to support analytics—and you have to move data from one to the other and ensure it is synchronised. In our view, an integrated system for supporting this sort of environment would be superior even if it didn't save you money (which it should do in the case of IBM Informix) simply because of the removal of complexity. However, as we have discussed, the products that can support hybrid operational/analytic environments are few and far between and there are even fewer that natively support both time series and geospatial data.

Conclusion

IBM Informix is widely deployed in transactional environments but is less commonly used as a traditional data warehouse. The purpose of the paper has been to examine its suitability for supporting analytic applications that are either developed in-house or by users or by third party developers. In particular, we have considered different classes of analytic applications, whether stand-alone or of hybrid type and, if the latter, whether combined with transactional or 'big' operational data.

In our opinion IBM Informix is suitable for use to support all of these environments. To be fair, it has more competition for stand-alone analytic applications: there are a number of other vendors in the data warehousing space that target this capability. However, for hybrid environments Informix has a number of unique capabilities that cannot be matched by either conventional data warehousing vendors or traditional data warehouses. In particular, applications that require time series, with or without spatial capabilities, will lend themselves especially well to the IBM Informix environment.

Further Information

Further information about this subject is available from http://www.BloorResearch.com/update/2142

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

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 his own company 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 focused on Data Management.

Data management refers to the management, movement, governance and storage of data and involves diverse technologies that include (but are not limited to) databases and data warehousing, data integration (including ETL, data migration and data federation), data quality, master data management, metadata management and log and event management. Philip also tracks spreadsheet management and complex event processing.

In addition to the numerous reports Philip has written on behalf of Bloor Research, Philip also contributes regularly to IT-Director.com and IT-Analysis.com and was previously 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 written a number of reports published by companies such as CMI and The Financial Times. Philip speaks regularly at conferences and other events throughout Europe and North America.

Away from work, Philip's primary leisure activities are canal boats, skiing, playing Bridge (at which he is a Life Master), dining out and walking Benji the dog.

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