White Paper

Bloor

Considerations for maximising analytic performance

A White Paper by Bloor Research
Author : Philip Howard
Publish date : September 2013

DB2 with BLU Acceleration should not only provide better performance in the first place but also provide consistent performance, with a corresponding requirement for less hardware and less cost.

Philip Howard

Introduction

When it comes to running business analytics, there are three key non-functional requirements that must be met: fast performance, usability and affordability. Bloor Research was asked by IBM to compare the performance capabilities of the leading business analytic platforms. Specifically, we were asked to evaluate how the combined capabilities of business analytic tools and the underlying database management system can affect the overall performance of your analytic applications, reports and dashboards.

For this paper, we took a look at four solutions:

- IBM Cognos BI running on DB2 with BLU Acceleration
- SAP Business Objects running on SAP HANA
- Oracle Business Intelligence Enterprise Edition (OBIEE) running on Oracle Exadata
- Microsoft Business Intelligence running on Microsoft SQL Server 2012.

To simplify the discussion, we split this paper into two sections:

- A comparison of the business analytics tools including any unique capabilities they deliver by integrating with the underlying database management system
- A comparison of database management systems.

BI tool comparison

Overall, we do not believe that the features of any the business intelligence products provide significant benefit in performance terms when compared to the overall effect of the database in question. This does not mean that there are not particular features of particular products that will provide a performance boost but that those benefits will be relatively small compared to the underlying database. Even where products are tied specifically to a particular database, it is the database that typically provides the features that enhance performance rather than anything that is intrinsic to the BI tool itself. Nevertheless, there are a few facts that are worth discussing.

Overall, at the business intelligence level, OBIEE and Microsoft BI have been specifically developed to take advantage of the respective databases that they run on. And while SAP Business Objects nominally claims to be an open system it is clear, as we shall see, that it is increasingly optimised for use with SAP HANA. IBM Cognos, on the other hand, has no specific features that are available in conjunction with DB2, with or without BLU Acceleration, and the product continues to be targeted across all platforms.

IBM Cognos Business Intelligence

With respect to Cognos Business Intelligence (BI) there are two things to note. The first is that this is the most open of the products in this space. This is because it uses native SQL syntax rather than being specialised for a particular platform. It will detect the version of the database in use and adjust as appropriate. The second is the feature known as Dynamic Cubes. This extends the in-memory caching capabilities of the previously introduced Dynamic Query feature, applying this caching to members, data, expressions, results and aggregates. This is an optimisation feature in the business intelligence product and outside the database per se. It is not yet available to run with SAP HANA although it will be in due course.

SAP Business Objects

SAP Business Objects is increasingly focused on SAP HANA rather than as a general-purpose product. Given the attention that SAP is paying to HANA this is not altogether surprising. However, as an example, HANA doesn't support common table expressions (CTEs) so you cannot use these functions of SAP Business Objects when run in conjunction with HANA. Of course, the same would apply if you wanted to run IBM Cognos BI in conjunction with HANA (you would have to turn CTEs off) but the upshot is that Business Objects is de-featured by being run on HANA rather than the other way round.

The one major feature of Business Objects that HANA enables is Analytic Views. An Analytic View is an OLAP cube but, unlike in other databases where these are materialised, in HANA they are calculated on the fly and the Analytic View is, in effect, the plan of how to transform the data rather than the data itself. The point here is that HANA is fast enough to calculate these on the fly rather than needing to store materialised views. This is a good thing. However, we have heard claims that the use of Analytic Views speeds up query processing more generally. We do not see how this can be the case. If the view is not materialised you cannot reuse it – you have to recalculate the cube every time you want to use it. This implies that you will get consistent performance (all other things being equal) but not any more than that.

Oracle Business Intelligence Enterprise Edition and Microsoft Business Intelligence

Neither of these products has any particular features that encourage high performance outside of their close integration with their respective databases. The only relevant comment that can be made here is that Microsoft targets its Business Intelligence solution at "start-ups, small businesses and medium businesses", which suggests that the company is not targeting large scale deployments which, in turn, suggests that either it does not scale or it will not perform beyond these environments.

Interim conclusion

In summary, as far as business intelligence itself is concerned, the various products are not going to make a huge difference to performance over and above what the database gives you. The only exception is Cognos Dynamic Cubes although we still believe that the database will always be much more significant when it comes to performance.

Database comparison

In this section we will discuss each vendor's particular database capabilities that accelerate analytic performance. There are lots of database technologies that are well-known: massive parallelism, materialised views, indexes, partitions, optimisers and so forth, so we focused on some of the newer capabilities such as in-memory, columnar processing, compression and data skipping—techniques used to find data more quickly, as these are the key new frontiers of database performance.

Overall, SAP HANA has been designed as an in-memory database while DB2 with BLU Acceleration, although it has significant in-memory capabilities, assumes that there will never be enough memory for a complete environment. Oracle and Microsoft, on the other hand, despite any claims to the contrary, both use memory in a traditional manner for caching. SAP HANA is completely columnar (in the sense that it never uses row storage) while DB2 with BLU Acceleration supports both row and column-based tables. SAP, Oracle and IBM provide improved compression via their columnar support. They are, of course, all databases, but their architectures are quite different.

IBM DB2 with BLU Acceleration

BLU Acceleration effectively does the following, which was not available previously:

- It introduces the use of 'actionable' compression. The big advantage of this is that for many query processes you do not need to de-compress the data in order to query it. This is because the techniques now used for compression (known as frequency-based dictionary compression) preserve order. This means that predicate evaluation, joins, grouping and count queries can all be performed without de-compressing the data. It should further be noted that the rate of compression achieved by DB2 with BLU Acceleration is significantly improved using these techniques, when compared to previous versions of DB2.

- Data may be held in columnar format. This is not an all or nothing choice. Where it makes sense to use columns you can use columns and where it makes sense to use rows you can use rows combined with conventional indexes. We do not believe that we need to reiterate the value of using a columnar approach for appropriate query types. Note that column-organised tables do not have secondary structures, such as materialised query tables, thereby eliminating any need for synchronisation.

- It provides in-memory processing through what it calls "dynamic in-memory" capabilities that allow columns of data to reside in memory for query purposes. Unlike SAP HANA, which has been designed purely to provide in-memory capability, DB2 with BLU Acceleration has been designed on the premise that your data warehouse will always exceed the size of the memory that is available. Of course, you can always pay a fortune to have perhaps hundreds of terabytes of memory but that is an impractical proposition for most organisations.

- It advances the ZoneMap technology gained when IBM acquired Netezza. It calls this data skipping and the way that it works is that the database holds metadata about where data is stored and this allows you to 'skip' those areas of the database where the data that the query requires is not stored. This can very significantly speed up relevant queries.

- It uses parallel vector processing. This is cross-core parallelism: you can parallelise across the cores within a single CPU as well as across sockets—and parallelises Multiple Data elements per Single Instruction cycle in each core (SIMD). IBM claims one of their strengths is the efficiency of the parallelism made possible by some deep engineering to reduce "memory access latency", in other words the time it takes to get data from RAM into the CPU where it can be processed. Most systems have a fair bit of memory access latency, but it gets disruptive across sockets. BLU Acceleration is doing something special to keep those latencies low.

There are some additional points we should make. Firstly, DB2 with BLU Acceleration is not available for use with the PureData family of products as yet. Secondly, it is only available on single server systems and not on large shared nothing platforms although, having said that, we understand that IBM has run proofs of concept using DB2 with BLU Acceleration with warehouses well in excess of 100TB. Finally, it is worth mentioning that partitions are not supported with DB2 with BLU Acceleration at present, although we understand them to be on IBM's roadmap.

Database comparison

Microsoft SQL Server

The key feature of Microsoft SQL Server 2012 Parallel Data Warehouse is what is known as xVelocity. This provides memory-optimised Columnstore indexes. This means that only the columns needed to satisfy a query need to be read. However, while this will improve performance where appropriate it will also worsen the performance of some other queries. You would like to have the optimizer work this out but as Microsoft states on its website, "the cost models the optimizer uses are approximate and sometimes the optimizer chooses to use the Columnstore index for a table when it would have been better to use a row store (B-tree or heap) to access the table." It then suggests using a work around.

xVelocity also enables superior compression rates: either dictionary-based or using run-length encoding, depending on the type of data. However, we do not believe that this will give as good a rate of compression as IBM and data will need to be de-compressed in order to be queried.

Oracle Exadata

Oracle claims that Exadata is column-based. It is not. It uses columns to enable better rates of compression than it could get otherwise. Moreover, the data needs to be de-compressed when it is queried: the data is compressed in the storage server and, with only one or two exceptions, it needs to be decompressed after it is moved from there for query processing. It also claims to have in-memory capabilities. It does not. Its flash cache is exactly that: a cache (which is also located on the Storage Server and not the Database Server. What Exadata does have is Smart Scans.

Support for large table scans is what Exadata, as an add-on to the Oracle database, was specifically designed to address. Put simply, it puts processing close to the disk to provide an MPP-like, shared nothing approach. Data is streamed off disk and pre-processed locally before the results are passed to the database itself for final processing. This results in faster table scans and reduced traffic across the network between the Storage Server and the database. The major weakness of this technique is that Smart Scans, which operate at a lower level than the database, are turned off if there is a write to the data page in question. In other words it will be best suited to static environments or situations where it is acceptable to update data on a periodic basis. If you are operating in real-time or near real-time mode then Smart Scans will need to be reserved only for that part of your queries that access historic data. A further issue with Smart Scans is that when they are used the data is decompressed in the Storage Server so that uncompressed, rather than compressed, data is passed to the database layer.

One of the issues raised with respect to Oracle Exadata is workload management. Oracle assigns resources to queries on a first-come, first-served basis but if the load is high then later arriving queries get downgraded, receive fewer resources and therefore run more slowly. Furthermore, you can't increase resources to a query if the load subsequently decreases, meaning that resources are now idle.

To avoid inconsistent response times and minimise idle resources, Oracle recommends queuing incoming queries so that it may more smoothly allocate the resources required. Unfortunately, in a high concurrency environment with many incoming queries, such as enterprise data warehousing, query queuing consumes a large amount of redo log space making it difficult or impossible for some workloads to use this technique to improve performance on a given system. Query queuing appears to be useful only in cases of limited concurrency and duration.

In cases of workload management of mixed workloads, such as running OLTP and BI simultaneously on the same system, beware of workload isolation issues. For example, analytic queries have a tendency to overrun the system, with large amounts of Storage Server CPU being consumed to de-compress analytic query data, thereby slowing performance for OLTP operations. If this results in OLTP performance falling below SLA requirements then query performance will need to be throttled back. Ultimately, although Oracle offers controls to govern I/O consumed by a workload, there is no ability to control the Storage Server CPU consumed by a workload, making it difficult to isolate and maintain SLAs in a mixed workload environment.

The other major thing that Exadata brings to the table is hardware. While extra hardware may improve performance you have to pay for this: literally.

Database comparison

SAP HANA

SAP HANA is an in-memory, columnar database system that runs only on x86 platforms. Unlike all of the other vendors discussed in this paper, the whole database has been designed to be held in memory. The implication of this is that SAP HANA will not be a practical proposition for very large data warehouses (for cost reasons if for no other) for which SAP Sybase IQ would be a more practical proposition. As we have discussed, this purely in-memory approach has some advantages such as the ability to support analytic views that do not require views to be materialised.

Column-stores enable a better compression rate, which reduces storage requirements, but data has to be de-compressed during query processing except that SAP claims to be able to do equality matching without de-compressing.

As part of our research, we reviewed a report of free memory usage by SAP HANA during query evaluation. Data is loaded, in columns, into memory, but memory usage then spikes as you de-compress the data. The software also creates intermediate objects during processing which further uses up memory. It is interesting to note that the version of SAP HANA under test was the first version of the software that provided a facility to remove data from memory. This strongly suggests that SAP needed to do this in order to reduce the pressure on memory. In particular, if you run out of memory then your query will fail. In theory, memory usage should be ameliorated by the fact that HANA only loads relevant column data, although the extent to which this will help will vary depending on how many columns need to be read. The bottom line is, and we have heard this from a variety of sources, that SAP HANA, at least at present, is not only in-memory but memory-bound.

Conclusion

In conclusion, neither SAP HANA nor Microsoft SQL Server, with or without their respective business intelligence products, are in the same ballpark as Oracle and IBM when it comes to improving the performance of your analytic applications. The reasons are different but nonetheless true.

In the case of Microsoft, the company has never been a player in the high-end data warehousing market and it has never been seen as a potential host of data mining or complex analytics. Nor has the introduction of the Parallel Data Warehouse done anything to change this perception. As its website for Microsoft BI clearly implies, Microsoft SQL Server together with Microsoft BI are clearly targeted at small and medium-sized organisations, which are Microsoft's traditional markets.

As for SAP HANA, we can only conclude that the product is not yet ready to compete on analytic performance except in very limited environments. It is likely to be fine for specific workloads with limited amounts of data but, as a general-purpose database, we believe that it needs several more years of development before it will be able to stand up against the likes of Oracle and IBM. Let us be clear: Oracle and IBM have been developing databases for over 30 years and they have vast experience. Is it any wonder that SAP, with no history of successfully developing databases, cannot develop a brand new database to the same level within just a handful of years?

So, at least as far as performance is concerned the choice is between the old rivals: Oracle and IBM. While the performance figures we have seen favour IBM, the margins are not huge: DB2 with BLU Acceleration is not outperforming Oracle by orders of magnitude but offers more like twice the performance. That's still pretty significant. But to be honest, it is the inconsistency of performance that worries us most about Oracle Exadata, with queries getting downgraded and queued queries getting timed out. Given that many analytic environments today are mission critical, performance inconsistency can have serious financial repercussions as you have to add hardware (and therefore expense) in order to maintain service level agreements.

DB2 with BLU Acceleration should not only provide better performance in the first place but also provide consistent performance, with a corresponding requirement for less hardware and less cost.

Further Information

Further information about this subject is available from http://www.BloorResearch.com/update/2179

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