White Paper

Bloor

Enabling in-database advanced analytics with Sybase IQ

Whether it is to better understand customer behaviour, optimise your supply chain or to recognise and prevent fraud, it should be clear that there is significant value to be derived from advanced analytics and in-database analytics as a discipline

Philip Howard

Introduction

There is a great deal of discussion right now about advanced analytics, predictive analytics and in-database analytics. What are these? How does advanced analytics (which we take to include, but not be limited to, predictive analytics) differ from normal analytics? Why should you care? If you are interested in what it can offer, why might you want to implement inside a database (data warehouse) rather than conventionally? Moreover, if that is the case, what should you do about it and what sorts of facilities and capabilities should you be looking for from vendors? These are the questions we are going to seek to answer in this paper, with specific reference to the capabilities provided by Sybase IQ. However, before we discuss the advantages (or otherwise) of enabling advanced analytics through in-database processing, we need to understand analytics within a business context.

What is analytics?

Unfortunately there are no clearly agreed definitions of analytics, beyond tautologies such as "the science of analysis". So we need to start not from the perspective of what the enabling technology is, but from what it is intended to do; and what we are trying to achieve, as a business, is to make better business decisions based, at least in part, on quantitative data. However, this is simply a definition of business intelligence taken in its broadest terms and we need to be able to distinguish advanced (and typically interactive) analytics from conventional business intelligence.

In our view there are at least one, and often two, major functional differences between advanced analytics and business intelligence, which is primarily focused on reports and summarisations. In the first case, analytics specifically looks to discover and understand patterns of behaviour and activity through deep data exploration and ad hoc analysis. Business intelligence, on the other hand, simply presents what has happened around what is known: if there are patterns or trends within that history then it is up to you as the viewer to identify them; all that business intelligence will do is to structure and present the information in ways that may help you to extrapolate into the future. Secondly, those identified patterns are commonly, though not always, used to predict the future and, thereby, improve business performance. Most business intelligence tools have no such capability. Furthermore, it should be readily appreciated that understanding complex patterns of activity and (potentially) projecting them into the future, is significantly more complex than reporting on, for example, sales of products by store, which is the typical bailiwick of business intelligence solutions. Advanced analytics comes into play when you want to look at sales of products by store, correlate that with the profile of customers, mix that with the impact of promotions, add in seasonal patterns, and geographic location considerations—and then explore and model new business scenarios to exploit this information. Thus you are now dealing with many diverse data sources, large volumes of data, and you require ad hoc interactive analysis.

However, we need to go a step further. Just because a query is more complex than can be typically handled by a conventional business intelligence solution (for example, queries involving multi-way joins, whole table scans and correlated sub-queries) does not necessarily mean that those queries fit within the domain of advanced analytics, even though they are often referred to as such. Specifically, the difference between advanced analytics and other types of query processing is that the former involves a two stage process rather than a single operation. When conducting advanced analytics the first stage is to discover patterns or correlations in behaviour and the second stage is used to compare a real-world event, person or circumstance to the discovered pattern or model to see whether it fits the expected behaviour or is anomalous. This second stage is referred to as "scoring". Thus we build a model that aims to predict whether credit card holders will default on their debt or if a particular customer will churn at the end of his contract or if we can up-sell or cross-sell him additional products.

Thus we come to a definition:

> "Analytics is used to a) discover behavioural characteristics and b) allows actual events and objects to be compared with those characteristics, for either discovery or predictive purposes."

Note that we have made no distinction between analytic tools such as data mining or statistical modeling on the one hand and analytic applications on the other. The latter employ the techniques of the former even if that is hidden from the user.

Why use analytics?

Analytics is employed in a wide range of environments and can either be supplied as applications or through the use of statistical or data mining tools, depending on whether you want a more packaged or do-it-yourself approach. Typically, we can classify the types of environments where analytics may be useful under four headings.

Security

There are three main security use cases:

- Fraud detection/prevention: while forensic analysis can discover fraud after the fact, prevention is much to be preferred. Thus, you can use analytics for credit card fraud prevention, or for detecting social security or benefit fraud. Similarly, on-line poker sites use analytics to prevent fraud (or, at least, to nip it in the bud) as do insurance companies for detecting false claims. This is a particular area of focus for Sybase, especially when Sybase IQ is used in conjunction with Fuzzy Logix (see later).

- Security event management: if your infrastructure is threatened by hackers, viruses, SQL injection attacks or other menaces it is important to recognise what sort of attack this is in order to react appropriately. This requires recognition of the pattern of activity that this threat represents, such as a low and slow attack. Further, analytics for its own sake is not enough: for business purposes what you discover needs to be actionable. For example, for malicious network attacks you may know what is happening but it takes a long time (using a traditional approach) to pull in data from other relevant sources, do cross analysis, validate your findings and determine the right action to take. If you are under attack you need to be able to respond, if not instantly, then at least within minutes—and the same thing applies to the other use cases that follow, though this sort of low latency may not be required in all environments.

- Law enforcement and security agencies: while these bodies do not normally discuss the use of technology within their organisations the use of analytics to understand criminal or subversive behaviour is obvious.

Risk

There are both general-purpose and specific risk elements. In the general case are considerations such as: if you want to apply for a loan, a mortgage or credit card, how likely are you to default? Of course, risk management is much broader than this: it applies in capital markets, in insurance, and whenever a customer is allowed a credit limit, but whatever the environment is, you will want to analyse your potential risk. Note that in the banking and insurance sectors there may be associated compliance issues associated such as Basel II and Solvency II. Again, this is an area of special focus for Sybase.

Verticals

There are specific requirements associated with particular vertical markets, even if we leave risk management aside. For example, in capital markets, analytic models are used to predict the movement of share and commodity prices as well as foreign exchange movements, in order to support investment decisions. With its long history in serving financial markets this is a further area of specialisation for Sybase.

Another example is traffic analysis, primarily used in the telecommunications sector, where you want to analyse your call network or caller roaming patterns in order to optimise traffic patterns (avoid potentially un-secure paths, avoid paths that are known to have malicious activity, reduce the number of network hops in call roaming, maximise roaming patterns to meet service level agreements and so on). Similar considerations apply to other providers of networks, regardless of whether these are transport, IT or utility networks.

You can also use analytics across various scientific disciplines, ranging from biology to archaeology.

Why use analytics?

General-purpose

Finally, there is a range of analytic requirements that span a number of user environments. These include:

- Customer behaviour: for (a real) example, you are a telecommunications company and you want to establish how teenage girls use their mobile phones as opposed to the way that teenage boys use them, so that you can more effectively market to each of these segments. There are lots of other examples where you want to understand customers better, including market basket analysis, loyalty card optimisation, customer segmentation, and so on.

- Supplier behaviour and spend analytics: both of these, which are slightly different (one with the emphasis on particular suppliers and the other on what you actually sourcing), can be analysed in a variety of ways to optimise your supply chain and cut costs.

- Influencer analytics: associated with customer intelligence, this is an emerging field in which you want to analyse who influences whom (using social media sites, Twitter and so on) on the basis that if you can influence the influencer then you have a greater reach into your chosen market.

General-purpose

Characteristics of analytics

The main characteristic of analytic functions or applications, especially in the development phase, is that, leaving aside scoring, you need large volumes of data, often from many diverse data sources, to work with. Producing statistically significant models requires processing a lot of data. Fortunately, there is a lot of it around. However, just as we have already seen that there is a distinction between scoring and other aspects of analytics there are, in fact, multiple aspects to consider. These include:

1. Data exploration: this is the initial process of understanding the data; first of all knowing what it represents and then performing some simple statistical calculations such as means, medians, standard deviations and so forth. Some simple classifications by age group or location may also be conducted at this stage. Products like Sybase IQ have appropriate functions built into them for making these calculations within the database.

2. Data cleansing: data mining and other tools are less tolerant of poor quality data than the data warehouse in more general terms. That said, most such products have workarounds for handling things like null fields. On the other hand, having to use these workarounds means that the subsequent analysis will be less accurate: if there is no value in a field then you cannot draw conclusions based on that value.

3. Data transformations: depending on the algorithms you plan to use (and you may use several, comparing their results to see which is most efficient or useful) you may need to convert the data into a particular format. For example, some techniques require a range of values between zero and one, so you would have to map your existing values to that range. Similarly, you may want to combine data elements to work off calculated fields. Again, Sybase IQ has suitable functions within the database.

These three processes take up the majority of the time spent by data analysts and statisticians. Yet these are essentially non-productive tasks in the sense that they only prepare the way: it is the process of building predictive or other models, in step 4, that actually creates value for the business, because it helps it to better understand customers, suppliers or other players/factors, and because it supports the scoring process (step 5). Two further steps are reporting and model maintenance but these are relatively less time consuming than the first four or five steps.

Traditional analytic processing

Before discussing the benefits of in-database analytics it is important to understand how analytics have traditionally been processed, in order to see where deficiencies have occurred in the past or, in many cases, where they still do. This is actually a two-fold discussion because we need to discuss scoring separately from discovery.

Scoring

The reason why we need to discuss scoring separately from discovery is because discovery has never previously been an in-database process: until now the data was always extracted to an analytic server for processing. On the other hand, when it comes to scoring, a number of database vendors (though by no means all) have offered in-database scoring as an option for some years, though the truth is that these facilities were not widely used, as they were not supported by the leading predictive analytics and data mining vendors. This too is changing.

In most cases the way that scoring has worked historically is that data is extracted from the data warehouse to an analytic server for scoring purposes and then the final scores are commonly bulk loaded back to the data warehouse (or other environment). In these types of cases in-database scoring allows you to now handle more complex (or custom) models with lots of variables, transformation, methods, and so on, in order to get accurate and reliable results, in addition to performance gains. In addition, while scoring is definitely not a difficult task it can be time consuming. There are certainly examples where traditional scoring takes several hours or a day to complete. With in-database scoring the processing time can come down to a few seconds or minutes, so there are substantial performance benefits to be gained.

Discovery

The discovery and model building phase of analytics has been supported without benefit of in-database capabilities for decades so it is hardly fair to say that that is not feasible also. However, it does suffer from drawbacks.

The way that analytics historically has worked is that an application, say a data mining tool, is running on an application server somewhere external to the data warehouse. When a model is being investigated the tool extracts data from the data warehouse and processes it. However, as we have already discussed, in order to build an accurate model you need to process a lot of data. This means transferring that same lot of data across your network with all the implications that that has on performance. In addition, the application server itself is a potential performance bottleneck and that situation will be exacerbated if you have multiple application servers handling multiple data types, which is often the case. As a result of these performance limitations it is commonplace to sample the data instead of processing the entire data set. This means that there is less network traffic and less processing strain. Unfortunately, it also means significant additional work in the data preparation phase (steps 1 to 3 above), because you have to ensure that the sample is representative of the data set as a whole. For example, in any large set of data it is likely that there will be outliers (that is, values outside the normal expected range). This creates a problem, not only when you use an algorithm that requires values in a range of zero to one (because it skews the results) but also because, if this is a valid result and not a data quality issue, then such outliers need to be properly represented in the sample you take, not least because outlier analysis is often a fruitful form of investigation. The other problem with sampling is that it is more error prone: it is very likely that you will miss important outliers or anomalies.

So, using traditional methods you either process all the data, which is more accurate but requires significantly more processing power and bandwidth, or you sample the data, which is less accurate and requires more work.
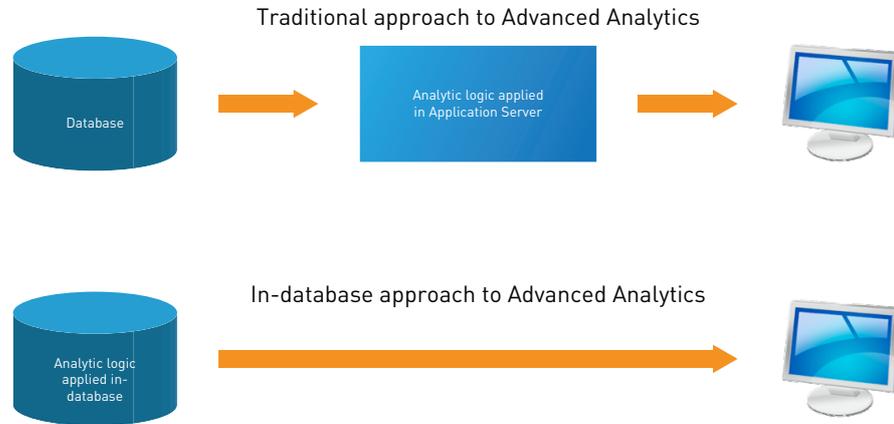
In-database analytics

Traditional approach to Advanced Analytics

Database → Analytic logic applied in Application Server →

In-database approach to Advanced Analytics

Analytic logic applied in-database →

Figure 1: Traditional vs in-database analytics

As might be imagined, in-database analytics means performing analytic functions in the database without having to extract the data to an application server. This is illustrated in Figure 1.

This approach eliminates any network issues and leverages the parallel capabilities of the data warehouse. This, in turn, means that you do not have to sample the data but rather can analyse complete data sets, thus eliminating some of the preparation work you would otherwise have to do, while simultaneously increasing accuracy and, of course, improving performance. Plus you can address diverse data types derived from many more data sources. As an example of the performance gains you can get from using this approach, Figure 2 represents the performance difference achieved in a customer proof-of-concept when running the same complex data mining process either using a traditional approach or in-database. Note that the same database software and the same hardware were used in both cases.

In addition, there has historically been a 'people' dimension because of the lack of common ground between IT and business analysts or data miners/statisticians. The warehouse team is concerned with meeting system-level agreements, maintaining data and application security and keeping the systems running. The analytic team is concerned with delivering high-value insights, responding quickly to business requests and manipulating large volumes of data in a highly iterative fashion for model development. Bringing these functions

into the warehouse has helped to bridge this divide between IT and business analysts, which is critical to obtaining accurate results, improved quality of modelling outcomes, and improved time to results in a reliable manner.
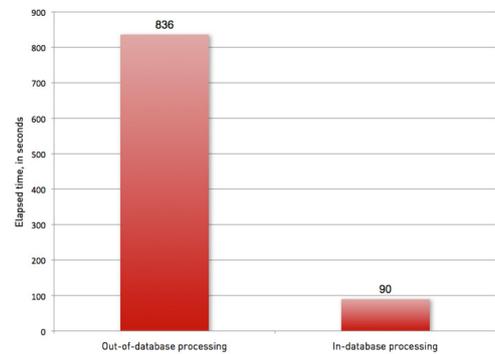

Figure 2: Marketing model calculation performance

Beyond these considerations it is important to understand exactly what in-database processing on complete data sets versus aggregate/sample data sets can bring to the table and we can do this by looking at the first five steps described above:

1. Data exploration: the various calculations required at this stage are not, in themselves, difficult or complex. However, they will typically require whole table scans unless there are specific mechanisms in place to avoid this. Traditional relational databases will typically be at a disadvantage for this purpose precisely because they will have to read the entirety of each table or make use of a plethora of indexes. It will be preferable to use a data warehouse technology

In-database analytics

that avoids either of these requirements: Sybase IQ's column-based approach being an example of such an approach.

2. Data quality: the same applies to data quality as to data exploration, albeit that there may also be a manual effort involved—for example, deciding how to cope with outliers.

3. Data transformation: because of their parallel engines, data warehouses such as Sybase IQ are now favoured for transformation purposes in general (hence the trend away from ETL—extract, transform and load—and towards ELT, because the warehouse can process these mappings more effectively). There may also be issues around whole table scans as already discussed.

4. Model creation: this is where the complex statistical and other processes and algorithms come into play and the extent to which any particular warehouse will provide advantages will be product specific. For example, if the warehouse does not support linear regression and that's what you need to do then you are back to the situation of having to extract data as in traditional scenarios. It will therefore be advantageous to have as many data and application-processing functions as possible supported natively by the warehouse so that you can do as much work as possible within that environment. In the case of Sybase IQ the following (extensive) facilities are provided out-of-the-box:

    a. Aggregate functions, which include window functions to support the calculation of such things as moving averages and time series analysis. The latter include specific financial functions to support capital markets.

    b. Analytical functions supporting ANSI, SQL, and OLAP extensions amongst others.

    c. Data type conversion, data and time, HTTP (to support web services), numeric (mathematical), string and system functions.

    d. Text analysis capabilities that can be combined with the foregoing.

In addition, there are C/C++ function libraries available from various Sybase partners, notably Fuzzy Logix, which will run directly within Sybase IQ. These contain hundreds of pre-built mathematical, statistical and other functions as well as advanced capabilities such as cubic spline interpolation (used for calculating treasury yields as well as for fraud analysis and customer churn) and Monte Carlo simulations. The capabilities provided by these libraries will provide high performance solutions in targeted use cases and markets such as risk management, fraud detection and prevention and other data intensive applications in various verticals.

5. Scoring: as we have already discussed, this is not a particularly onerous task. That said, by no means all products support this capability and it is a function that one would clearly like to have, for performance reasons.

Conclusion

Whether it is to better understand customer behaviour, optimise your supply chain or to recognise and prevent fraud, it should be clear that there is significant value to be derived from advanced analytics and in-database analytics as a discipline. While we have discussed this during the course of this paper our primary interest is in the additional benefits that accrue from running analytics in the database as opposed to not doing so. There are, essentially, three such benefits: increased accuracy, from read only to interactive analysis, thereby allowing more complex analysis; improved performance particularly on large data volumes with response times often in seconds to minutes; and reduced development time for advanced applications (from months to less than a week in many cases). It is a debatable point as to which of these is most important but we would highlight the very fact that complex analytics is enabled in ways which are otherwise impossible and the fact that shorter development time means lower resource utilisation and reduced costs—put that together with depth of analysis, accuracy, and performance enhancements and it's like having your cake and eating it too.

In so far as Sybase IQ is concerned, Sybase has embedded an extensive array of capabilities within its database, in order to provide the sort of in-database analytics under discussion. Moreover, it has done so well ahead of the vast majority of its competitors within the data warehousing market and with a generally more extensive set of in-built functions: the company is therefore well placed to capitalise on these capabilities as demand for advanced analytics continues to heat up.

Further Information

Further information about this subject is available from
http://www.BloorResearch.com/update/2048

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

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 what is now P3ST (Wordsmiths) Ltd 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. His practice area encompasses anything to do with data and content and he has five further analysts working with him in this area. While maintaining an overview of the whole space Philip himself specialises in databases, data management, data integration, data quality, data federation, master data management, data governance and data warehousing. He also has an interest in event stream/complex event processing.

In addition to the numerous reports Philip has written on behalf of Bloor Research, Philip also contributes regularly to www.IT-Director.com and www.IT-Analysis.com and was previously the 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 published a number of reports published by companies such as CMI and The Financial Times.

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

Copyright & disclaimer

Bloor