

BUSINESS
INTELLIGENCE

● R

DATA MINING ?

The best approach for effective loss prevention

Contents

Executive Summary	3
The Problem.....	4
The Technical Challenge.....	4
Data Model.....	4
What to Ask?	5
Visualisation.....	6
Everything Changes	6
Business Intelligence Systems – Main Elements.....	7
IntelliQ approach	8
Data Transformation and Data Model.....	8
Analysis Tools	11
The Conventional Business Intelligence Approach	12
Conclusions	13
About IntelliQ	14

Executive Summary

Many large organisations use business intelligence systems to monitor a wide range of key business performance indicators at corporate, divisional, departmental and business unit levels. These systems rely on pre-defined extracts and summaries of data, and are limited when in-depth, dynamic analysis of very large quantities of detailed transaction data is required.

Successful businesses continuously look for areas where they could improve performance by monitoring, reviewing and analysing the information extracted from the data produced through their day to day operations. Most modern businesses have data warehouses and reporting systems which provide the basis for monitoring their regular activity answering questions about sales, customers, marketing, finance, personnel, etc. However, for applications such as retail loss prevention, fraud detection, financial risk analysis, and marketing analytics the ability to combine extensive domain knowledge with the ability to dynamically analyse and “mine” very large quantities of raw data is a fundamental requirement.

This kind of deeper, investigative, “*train of thought*” type analysis relies on the ability to identify seemingly random patterns from hundreds of millions of individual pieces of data. Traditional business intelligence and reporting tools are designed to identify “exceptions” based on a set of rules. This approach is effective if you know what you are looking for. However, when trying to identify new trends, spending patterns, or in the case of identifying for example fraudulent refund abuse, it’s what you don’t already know that has the most value.

The objective of this White Paper is to evaluate the relative effectiveness of using conventional business intelligence techniques versus a content specific data mining technology to support comprehensive loss prevention within a large retail environment.

IntelliQ’s RetailKey LP solution is the UK market leader in the provision of retail loss prevention software. By combining over 15 years experience together with its “*best of breed*” data mining technology, RetailKeyLP provides loss prevention teams with a very powerful, easy to use business tool that not only helps drive loss out of the business but delivers a consistent return on investment year on year. Some of the UK’s largest retailers use RetailKey LP including John Lewis Partnership, Argos, Homebase, B&Q, Carphone Warehouse, New Look and many more.

The Problem

It is useful at this point to draw the distinction between most conventional business intelligence systems, and data mining. The former provide query and reporting against selected, and often highly summarised and pre-aggregated data that enable managers to monitor business performance (e.g. give me a breakdown of last month's sales by region). They provide facilities to drill down to more detail (e.g. give me a breakdown of last month's sales by salesman in the poorest performing region), but are not designed, as is data mining, for rapid, ad-hoc analyses of large amounts of complete, detailed transaction data to seek new trends or patterns to provide new insight.

Whereas most business intelligence systems concentrate on monitoring "business as usual" and highlight only those exceptions which can be expected from time to time, data mining in its true form is invaluable for aiding the detection of non-obvious, unexpected or abnormal events. Managers will also find that while their query and reporting system reports the symptoms of some problems, they do not have the ability to uncover their underlying causes.

The Technical Challenge

Data Model

Although most large companies have invested in corporate business intelligence systems, there are some challenging technical issues to be overcome if software solutions are to be effective for organisations that transact with their customers in large volumes (retailers, financial services, telecoms, etc.). First there are the sheer volumes involved - millions or even billions of sales, thousands of products, hundreds of thousands of credit/debit cards, hundreds of stores and thousands of cashiers. More than that, a key technical issue is the huge number of possible relationships between all these entities.

Database size is not the only factor that determines the speed with which a query can be resolved. Performance will also be strongly affected by the number and types of "joins" between related data – for example between sales, stores and products.

This is best illustrated with an example based on the kinds of volumes found in a large retailer. Suppose the retailer knows from its business performance reports that it is losing a percentage of its profits, and suspects that some part of this is through fraudulent activity. It is interested in analysing refunds for fraud activity (notice we are already making a pre-determination that this is where the problem lies). In a given year there are one million refunds through the 500 stores. A simple report with a breakdown of refunds by store requires only a summary table of 500 rows of data - not difficult. But this is too coarse an analysis - stores are different sizes, and fraud may be "networked" across several stores. So we need to find the patterns that exist in the

linkages between not only the one million refunds, but also the 5,000 cashiers and 1,000,000 credit cards. That's a total of 5 million billion (5×10^{15}) possible combinations! That is before we have even considered whether the 250 million sales transactions need to be examined too.

Conventional data warehouse and business intelligence systems deal with these challenges by firstly, disregarding data which does not have apparent or immediate value, and secondly by summarising the data into pre-calculated totals, averages and other calculations. This reduces the data volumes the system has to deal with and eliminates much of the expensive database processing involved in both managing the data and in processing user queries. The consequence of this is a loss of information, and only a limited ability to explore underlying detail. In the retail example above, the refunds data will often not even be included in a data warehouse which has been primarily built for sales reporting. Even if it were available, no amount of clever systems analysis and design can pre-determine all the reports and queries and hope to expose all the irregular transactions that may be hidden there.

This situation is common to several business areas - retail loss prevention, market analysis, and outside of retail, in areas such as credit risk analysis, credit card fraud, airline flight analysis, consumer behaviour, network traffic analysis, and crime detection.

What to Ask?

The second main characteristic of these types of problems is that only so much can be discovered by asking singular questions, such as "What is the total value of refunds in May at Store X?". New insight may only be found through data mining techniques which are capable of trawling the data to highlight patterns, trends, and significant linkages or relationships in the data. Business managers have a model of how their business should be operating, and they will have a number of measures which they monitor to keep them apprised of the state of the business (revenue, profit, sales productivity, etc). Additionally they will ask some questions to drill into areas of concern (*"give me a breakdown of refund value by store for last month"*), and use conventional business intelligence systems and tools to do this.

Yet when business managers need to investigate fuzzier issues and ask more general questions like "How are we losing so much through shrinkage?" the problem is more difficult because they do not know enough about the problem to ask more specific questions. They need to be able to frame more general questions, and have the data itself provide the insights. This is the purpose of *data mining*.

These two approaches to business intelligence might be termed, respectively, *model-based* and *data-driven*.

The technical challenge of data mining is that it needs detailed transaction data to be effective, which means large volumes, and may need to generate and run many queries to test potentially many millions of possible relationships between the objects represented in the data (for example sales transactions, credit cards, stores, people and products). Attempting this with conventional business intelligence software will invariably lead to extremely long investigation times, and very limited results.

Visualisation

The enormous number of relationships between the objects of interest (i.e. sales, cashiers, credit cards, stores, products) also creates a major visualisation challenge - how to highlight the key relationships that indicate where problems lie? And how to present them to the investigator or analyst so they can easily interpret patterns and identify unusual activity?

Everything Changes

Finally nothing stands still. External economic conditions change, new risks emerge, fraudsters devise new ways of making their ill-gotten gains, business processes break down; there are innumerable new ways in which the business may be affected that had not made themselves known before. So management systems which rely only on over-selective or summarised views of business based on sets of pre-determined rules and expected behaviours will lack sensitivity to such changes.

Business Intelligence Systems – Main Elements

At the most fundamental level all business intelligence (BI) systems have three essential parts.

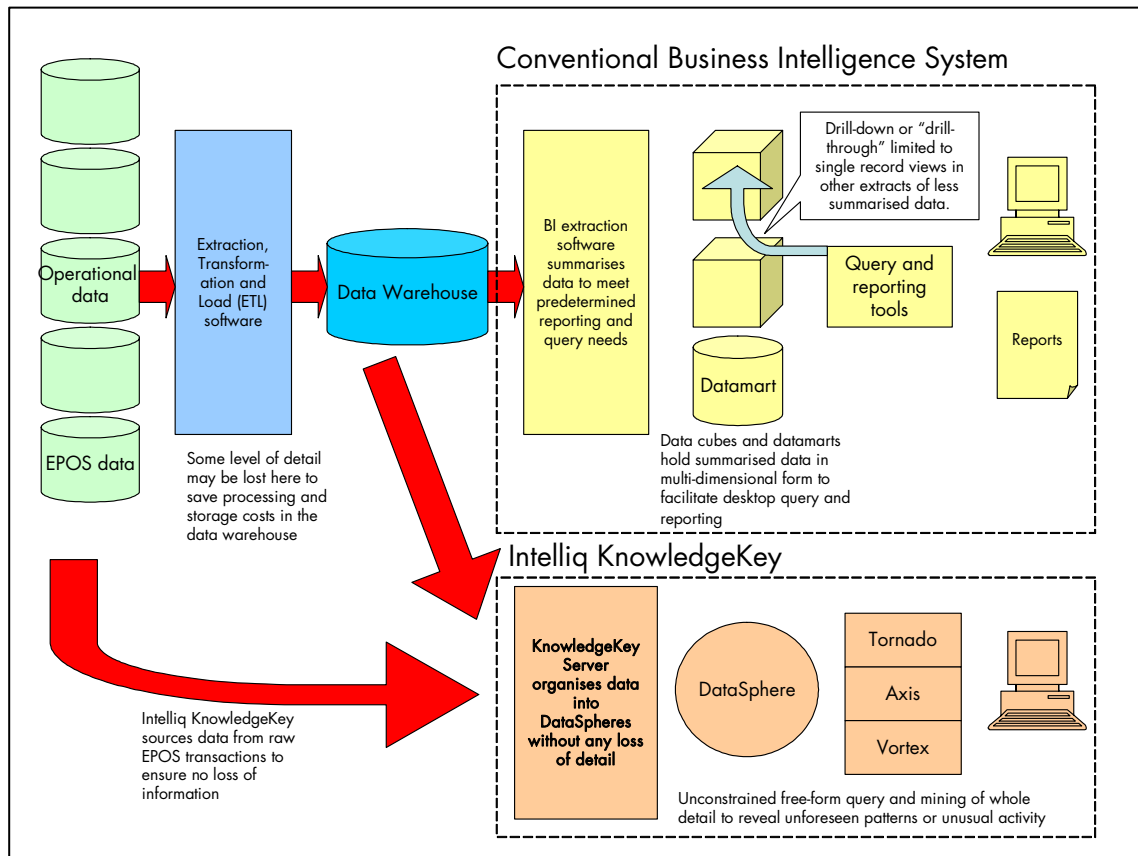


Figure 1: Data Warehouse and Business Intelligence Systems

The first is *extraction, transformation and loading* (ETL) software which extracts data from operational databases, then structures and organises the data into a form (or *data model*) suitable for efficient analyses.

The second is a set of databases and files (which may variously be *data warehouses*, *data marts* or *data cubes*) which store and make available the data in its analysable form.

The third is a set of end user tools that enable users to design, run and view reports, queries and analyses.

It is important to understand, for all the reasons above, that the design of all these three elements for data mining (or *data driven analysis*) has to be very different to the design of conventional business intelligence systems which are intended for *model-based* query and reporting.

IntelliQ approach

IntelliQ provides data mining systems, so its transformation software, data model and user tools are all purposefully designed to provide an unfettered analysis of very large volumes of detailed transaction data.

Data Transformation and Data Model

IntelliQ's technology is designed primarily to ensure that no detail is lost in the transformation of data from its sources to the query database or data warehouse. This is not a one-size fits all operation; in retail, EPOS systems are all different and produce different data content, structures and formats according to the retailer's own business situation; likewise in telecoms companies, billing systems will vary and in each sector large companies will have different operational data and systems to meet their business needs.

IntelliQ's transformation process has to be flexible to cope with this, and calls upon a rich library of software tools each designed for a specific task. With these tools, an automated data transformation process can be quickly built to meet the customer's specific requirements. Usually IntelliQ will use the original raw data (such as EPOS transactions) to create the transformed database. This is because existing data warehouses and data marts often will not store all the original detail - in the case of retail loss prevention, for example, the analysis of refund and void transactions may be critical, yet this data is often discarded for normal data warehousing or BI purposes.

The IntelliQ data model further contrasts with typical data warehouse models, in that the former ensures that it retains equal information about all types of events. Most data warehouses tend to have a bias toward certain types of events – typically sales – because of their focus on monitoring the business. Each record of an event in IntelliQ (examples of “events” may be item sale, sale transaction, credit card tender) includes all the detail related to that event – producing a very flat, and highly efficient data structure for analytic processing. In a “classical” data warehouse the model is highly normalised usually around the sales transaction. This is more efficient for disk space usage (now a relatively cheap commodity), but much less efficient for large queries, and can require skilled database analysts to ensure queries are structured carefully even to obtain correct results. See following diagram:

The Data Sphere™, once created, is self-managing. IntelliQ data middleware (“Nexus”) provides an automated data warehouse manager, load scheduler, segment manager, and other tools which means that, once created, an IntelliQ data warehouse is self-managing. In other data warehouses (and “data marts” – smaller, satellite data warehouses for departmental use), not only is a large amount of scarce database expertise required to build it, but a database administration team is also required to manage regular updates, archive older data and ensure the data is kept organised to required performance levels. A Data Sphere requires little or no resource for these tasks, and places no additional burden on the database management staff.

In summary, the IntelliQ transformation software and data warehouse has a number of technical features which stand it apart from other business intelligence systems for analysing transaction data, including:

- In order to achieve fast analysis times (processing a large number of automated queries against large volumes of data), data mining requires an uncomplicated, “flat” data model. IntelliQ’s transformation software “denormalises” the data into a single table (rather like a very large spreadsheet) which is better for the bulk processing required by data mining. The data model, however, continues to mirror a transactional view of the business retaining the concept of a transaction (which may be of many types such as sale or refund) which is composed of many transaction items (e.g. sale item);
- The Nexus middleware includes an intelligent SQL engine which optimises query generation to produce the most efficient SQL. Together with the IntelliQ data model, this ensures highly efficient and rapid query processing. This provides the ability to ask complex questions of the large volumes of detailed data in a “Train of Thought Analysis™” fashion – meaning users can ask questions interactively following a train of enquiry without the constraints imposed by summary or selective data;
- Automatic addition of sequencing to all records to preserve the chronology of transaction items. This is important for ensuring the detection of relationships or patterns in data where the order of events may be the significant factor;
- Use of Aggregate Fields for the pre-calculation of commonly used totals (transaction totals, refund totals, etc);
- Ability to “promote” data items from an item level to transaction level, where such data is commonly used to seek linkages (for example credit card number may be recorded at item level in operational EPOS data, whereas for analysis purposes its relevance for the vast majority of cases is at the transaction level).

This capability is part of an integrated package – designed from the outset to provide a wholly viable solution to analysing large volumes of transaction data. That means the rapid query performance does not come at a cost of highly skilled, constant database tuning and maintenance, but instead once in production, manages itself with scheduled and automated update and maintenance processes.

Analysis Tools

Whereas many BI systems can do a good job at reporting against the values in data (e.g. total sales this month) IntelliQ have developed software to look for significant yet non-obvious relationships and patterns in the data (e.g. is there a pattern between refunds, credit cards and cashiers?), and then to present these patterns in highly visual ways for fast and easy interpretation.

IntelliQ's Vortex product (see Figure 2: IntelliQ Vortex) uses a visualisation which presents a link analysis in such a way that the user can readily see the significant linkages between objects and events, despite the very many possible combinations. It also has sophisticated filtering and zoom capabilities to clarify unusual patterns still more. For example, it is used by retailers to detect unusual patterns of activity between cashiers, credit cards and refunds when looking for fraud.

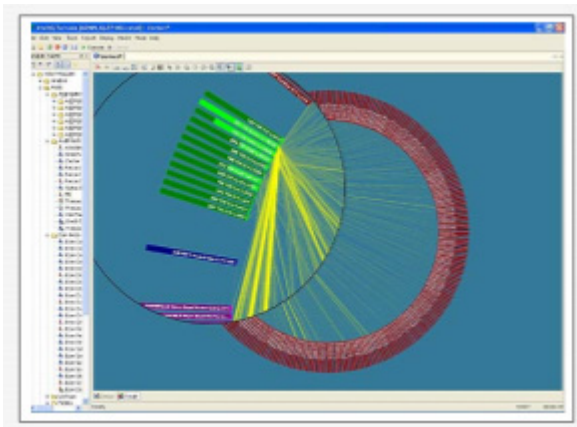


Figure 2: IntelliQ Vortex

As well as the powerful data mining and visualisation capability of Vortex, IntelliQ's product set includes end user tools similar to those of other BI products – a full function query tool (Tornado) and a graphical charting tool (Axis). In addition to the usual query capabilities, Tornado includes some features provided as a result of IntelliQ's close collaboration with its customers.

An example of this is its "Blueprint" functionality which allows the user to link a series of reports where each is to be executed depending upon conditions met by the results of former reports. These can be scheduled to run automatically, so in effect a complete and sophisticated analysis of conditions can be affected at the end of a business day. It includes an ability to issue alert emails when certain conditions are met. In many systems offering this option the email alert function loses credibility with users because their simplistic rules or conditions mean alerts too frequently are sent out unnecessarily. In Tornado, the Blueprint functionality is such that these emails are only generated when there are highly probable reasons for alerting users. Again this is strengthened by IntelliQ's innate ability to work with detailed data – allowing "alert" conditions to far more robust than those based on a more summarised or selective view.

The Conventional Business Intelligence Approach

As previously discussed, conventional Business Intelligence systems are generally predicated on a "model-based" approach. The data for reporting and query is selected and organised to mirror a model of how the business should be operating. They focus on monitoring straightforward measures of business performance, such as sales volumes, stock value, customer numbers, number of visits, etc., and even when they highlight exceptions, they are only against these predefined measures.

The every-day monitoring of these standard measures of business performance can be done against data which is summarised against a limited number of dimensions (e.g. sales by product, sales by store, sales by date, etc). In addition BI systems provide drill-down and multi-dimensional capabilities to investigate to some extent what lies beneath the highest level summaries - e.g. to query sales by product within a store.

But in order to deliver reasonable desk-top performance, business intelligence systems generally access data that is either pre-summarised to meet those reporting and query needs which can be pre-determined, or pre-filtered to eliminate data which was not recognised as having high utility during data warehouse requirements analysis.

To keep the data volumes within the capabilities of these systems, much of the detailed data will either be discarded or rolled up into summarised totals, especially where such detail appears to have little utility for day-to-day reporting. Indeed the challenge for the analyst designing business intelligence applications is to ensure that the data made available to users is the minimum necessary for their identifiable requirements. This reduces the difficulties in delivering reasonable query performance to end users, and reduces the costs and complexities of managing the large, complex databases of most data warehouses. This is doubly important because any single user may require several views (stored in structures such as cubes and data marts) of the data for different reports or queries, thereby duplicating data and increasing complexity. So any data which cannot be specifically identified and justified for inclusion is usually discarded as unnecessary for purpose.

The effect of this is that reports and queries, even when they are designed to report on the target problem, may not be alerting management to dysfunctional activity in their business, since the vital clues of unusual behaviour may be masked. This is exacerbated when these tell-tale signatures of such activity (as is the case for retail fraud) are constantly changing. For example, a report of overall levels of refunds by store may not be sensitive to dishonest refund activity by a single cashier. (In fact refunds data may not even be available in the data warehouse). If however we could analyse the fine details of the transactions to detect unusual behaviour patterns - perhaps an unusual number of refunds to one credit card - then the problem would be quickly highlighted and provide the evidence to resolve the issue.

Conclusions

In summary IntelliQ differs from, and complements, conventional business intelligence products by providing the ability to freely explore detailed data to seek new patterns in data, while other products are designed for repetitive query and reporting and can only highlight exceptions against pre-determined criteria.

The primary benefits of IntelliQ's data mining approach are:

- RetailKey LP has been designed for business users. Loss prevention analysts can quickly and easily build complex report and queries without the need to be experienced data analysts or IT specialists.
- It allows the discovery of previously undetected patterns in data which can reveal significant issues or opportunities for the business. This contrasts with query and reporting systems which at best can only report variances from definable norms. We could say that where query and reporting might provide an answer if you know the question, true data mining reveals new insights even when you don't know the question!
- It preserves all the valuable detail in business transactions such as EPOS data where much is usually discarded for business intelligence systems. Examples are void transactions, which for routine business reporting, are seen to have little value compared to the cost of processing and storing them in cubes or data marts. IntelliQ's intelligent and efficient use of this data is invaluable in the detection of business process abnormalities and fraud.
- It is very efficient and effective at finding the relationships between multiple large data entities (for example sales and credit cards), where other business intelligence systems struggle with this, being better suited where data only needs to be grouped by a smaller ranges of categories (for example sales by store).
- Its less complicated architecture designed specifically for analysis of large volumes of transaction data, its integrated data warehouse creation toolset, and reduced reliance on the creation of multiple summarised views, cubes or data marts means that time to implement is considerably less than the normal experience with business intelligence systems. ROI is typically faster as a result. In addition it is more adaptable, and requires little or no IT resource to be diverted for data warehouse maintenance or end user support.

In this paper we have discussed the differences between IntelliQ's data mining approach and conventional BI query and reporting, and used the example of retail loss prevention to illustrate those differences. Most of IntelliQ's work to date has been in helping major retailers to identify and reduce losses, but it is finding that its expertise and its product's ability to discover previously hidden information have substantial value elsewhere. Examples of these include benefit claims analysis in social services, credit card usage in financial services and market analysis in retail.

About IntelliQ

Founded in 1991, IntelliQ is a UK based company specialising in the development of industry specific data mining solutions. IntelliQ has become the UK's leading supplier of loss prevention software by combining its considerable retail domain expertise together with its world class data mining technology to deliver a proven business focused solution that delivers significant year on year financial returns to its customers.

In addition to retail, IntelliQ software is used by a range of other industries including financial services, market analytics and government agencies.

Contact:

IntelliQ Ltd.
Bellerive House
3 Muirfield Crescent
London E14 9SZ
United Kingdom

Phone: +44 20 7517 1000
Email: info@intelliq.com
www.intelliq.com