THE DATA WAREHOUSE IS DEAD.
LONG LIVE THE DATA WAREHOUSE!
Abstract

This paper explores the idea that modern Business Intelligence and Business Analysis tools render the concept of the data warehouse obsolete through the use of on-the-fly, in-memory data manipulation, and discusses the possibility that the data warehouse no longer has a place in delivering high quality information to users.

Simpler Times

Once upon a time, life was simple; or at least, the basic approach to Business Intelligence was fairly easy to describe... A process of collecting information from systems, building a repository of consistent data, and bolting on one or more reporting and visualisation tools which presented information to users.

Times have changed. I attended a presentation recently in which the speaker made the point that the capabilities of our computer hardware now far outstrip the functionality offered by the software we use. In simple terms, if we exploited every last drop of power at our disposal, we would be able to do far more and far more quickly; but it takes time to develop, test and bring to market operating systems and software applications. The hardware on which the latest product began development will almost certainly be several years old by the time the software reaches release.

Additionally, the requirement to maintain compatibility with older hardware and software means we’re always running a compromise. The price we pay for being able to retain our old documents and still use them with our new software is to not exploit the full potential of our computers.

What’s the problem?

One area in which the lag effect described above has had a particular impact is in the Business Intelligence and Business Analysis field; the quantity, complexity and richness of the data we hold is growing at an incredible, almost exponential rate. Why store a name when we can have a picture too? Why hold a picture when a video is better?
Additionally, the growing desire to interrogate less structured data such as that held within social media environments like Facebook and Twitter, has led us to a situation where we’re being held back by the capabilities of the platforms at our disposal for extracting, manipulating and presenting the data.

So why not ditch the platforms? Why restrict ourselves to the ‘traditional’ approach which we’re all so used to, when we have all of this latent power available? What we need is a new breed of products designed to take advantage of some of the plentiful, cheap, hardware at our disposal; a new type of product which will make accessing our information and answering our questions quicker, easier, and more interactive than ever before. If these products were made accessible to ‘average’ users, and less reliant on expensive IT departments to implement and maintain, we’d have reached Business Intelligence utopia, right?

**Enter the data discovery tool**

In the last couple of years, we’ve seen a wide range of new products which set out to address some of these needs. Data discovery tools such as QlikView, Spotfire and Tableau, which now sit in the marketplace alongside the ‘traditional’ Business Intelligence products from players like IBM, Oracle and Microsoft have grown significantly in capability, market share and rate of development. These products focus on rapid deployment and ease of use; they take advantage of hardware, provide a rapid development route and highly interactive interface, and are designed to be end-user focussed and code-free to simplify development and maintenance. Job done, you might think; Business Intelligence utopia achieved.

Sadly, it’s not quite time for tea and medals just yet. Whilst the new generation of Business Intelligence tools certainly make analysis and data interrogation accessible to a wider range of organisations, and from the market’s perspective are driving developments in the products offered by the rest of the vendors, we haven’t quite re-written the rule book.

**Changing approach**

For many years we’ve been using ETL (extract, transform and load) processes to capture the information we want from our source systems and populate a data repository for later use. In doing so, we accept the fact that we’re not using ‘real-time’ (or even ‘near-real-time’ in many cases) data, as it takes time to pull all the
information from all of the systems together. It’s such a common way of doing things we even gave it a catchy name – data warehousing.

Having a data-warehouse wasn’t enough on its own, however. As data volumes grow, the data warehouse simultaneously increases in size and complexity. The original objective, which was to consolidate sources of data into one high performance common repository, consistent in content and accuracy, begins to get lost. As the data warehouse becomes bigger, it becomes slower. Answering questions and queries takes longer, and the concept and value offered by the data warehouse is undermined. A new method of storing and retrieving vast amounts of information becomes needed.

Thankfully, OLAP (OnLine Analytical Processing) is there to rescue us from the data-quagmire; by taking the data from the data warehouse and manipulating it into a new ‘cube’ form specifically designed for fast reporting and analysis, OLAP provides a method through which we can interrogate millions or tens of millions of data elements in the blink of an eye. The power of colossal amounts of data can be distilled into a series of data structures which hold all the answers we need at our fingertips. All we need to do is ask the questions.

The new breed of products mentioned above has begun to change the way we think about Business Intelligence, but has it changed our approach? The vendors of such data discovery tools make much of their products’ abilities to interrogate live data on-the-fly, and of their abilities to deliver the power of a data warehouse directly to a user’s desktop without database skills, code, or reporting expertise.

The data warehouse is dead!

Conceptually, the in-memory model adopted by these new products allows them to do away with all of this. ETL processes are handled within the user’s client environment; in much the same way as Excel allows them to import data from other files, their new-breed Business Intelligence tools allow them to select a system, identify the data they want from it, and load it into the client. Once that’s done, they repeat the process with the various data sources they wish to interrogate, and the product itself seamlessly stitches the data together in a dynamic data warehouse type structure, establishing relationships between elements, and building its own OLAP cubes (or a proprietary alternative). With just a few hours more work, they can build an interactive dashboard with filters, charts and scorecards, and publish it to
colleagues or the wider world. Truly, Business Intelligence-in-a-box.

Of course, it’s not quite that simple. Behind the client environment, there lies a platform which facilitates all of this functionality; naturally, storing all of this data in memory requires lots of… well, memory. Servers supporting even small scale data discovery deployments routinely are provisioned with 32, 64 or 128GB of RAM; larger deployments can scale to multiple servers providing terabytes of memory.

Whilst the majority of IT systems are being virtualised to take advantage of the reduced lifetime cost offered by this approach, data discovery is one of the few technologies which does not immediately lend itself to virtualisation; the virtual infrastructure required to support a high-performance, highly-available data discovery server architecture requires a specification far higher than is commonly necessary to support all of the other server roles which it would contain, with a consequently higher price-tag. Businesses already using virtual infrastructure have to confront the need to ‘physicalise’ their Business Intelligence platform – which might prove difficult to justify to the board if you’ve just been extolling the virtues of virtualisation.

Assuming you accept these factors, is it therefore fair to assume that everything else is straightforward? In my opinion, sadly not; whilst the concept of the in-memory data discovery products is good, can one really do away with the data warehouse?

Long live the data warehouse!

Data warehouses are complex because they need to be; bringing multiple data sources together isn’t the sort of job which can be done in an afternoon. For a business with multiple business systems, millions of data rows, or thousands of users, the process of defining requirements, designing data warehouse structure, and creating ETL routines will be a lengthy one. It will require skills which can only be provided by technical experts, and an investment of time from users and stakeholders. Building the analysis (OLAP) layer on top of the data warehouse content will also be a challenge, relying on well documented requirements and an understanding and recognition of the different types of data and reports needed by different users and user groups.

The most important part of the data warehouse’s role is to ensure consistency of your reporting data, as well as ensuring that related data from disparate sources is accurately consolidated and stored. Consider the example of data pulled from a CRM system, and that from an ERP system; both customer and supplier records may have a column called ‘name’, but the two are not inherently related. I certainly don’t want to
see the suppliers’ details joined in the same dataset as customers’ details when I ask to find customer records associated with the name ‘Smith’.

Resolving these issues is straightforward when one understands the data sources and can relate them to the particular metric or measure one is interested in; recognising the ways in which data should be transformed, aggregated, or categorised in order to render it meaningful, and implementing a means of so doing, are the reasons we need data managers, database administrators, and Business Intelligence developers. They’re also the reason we need the data-warehouse; to ensure the integrity and accuracy of the reporting environment.

The in-memory approach may claim to be able to function without a data warehouse, but in the real-world businesses are imperfect, and the data held within source systems is often questionable in terms of accuracy and consistency. Effective direct access by a data discovery tool to source systems assumes that the data sources are high quality, data relationships are clear and easily determinable, and that users understand the impact and implications of directly connecting to live systems.

Imagine a situation where a user’s data-discovery ETL routine requires access to both the customers and orders tables in a large business system. Whilst for a smaller business there may only be a few thousand customers, and only a few orders per customer, if one scales up to a business the size of Amazon, with millions of customers and tens of millions of orders, the impact of an un-optimized query running against a live system could be hugely significant. Data warehouses provide a layer of insulation between source data and end user; they can be populated with data at quiet times, and the ETL routines which pull the data from them are developed, tested and optimized by experts.

Additionally, empowering users to directly access the potentially inconsistent data held in a live system can result in correspondingly inconsistent results. A less experienced user trying to find the same information as one who understands the data in more detail almost certainly will take longer, and come up with a different answer to their question. A data warehouse provides data which has been checked for accuracy, is guaranteed to be consistent, and which exposes commonly used data in a single view. Answering the question becomes much easier when there is only one path to (or as we often say ‘one version of’) the truth.

End-users reporting using information from several different systems also becomes much simplified using the data warehouse as a source. The complexities of extracting information from multiple sources, often using different methods of connection,
at different times, and in multiple formats, can be divorced from the reporting definition. The database administrator and Business Intelligence specialists will manage the process of normalising data and structuring it in a logical and consistent manner within the data warehouse, leaving users free to concentrate on the questions they need to answer and the outputs they need to generate.

Adding a new data source to a single data repository can also be done easily; a new system, with new information, will be analysed, its data-export routine created, data warehouse table structure designed, and all of these tested in isolation before a user is given access. With the ‘self-service’ data discovery model, the checks and balances associated with integration and testing of data sources can be lost, and the value and integrity of the business information users provide can suffer.

Perhaps the most important thing the data warehouse brings to an organisation is the sense of trust in the data it engenders within its users. When senior management buy-in to a data-warehouse as the primary source of business information, the data contained therein becomes an everyday part of business life; those people who are responsible for providing outputs to them and managing the process of reporting and analysis will arrive at the same answer to the same question no matter how many times or how many different people are asking it. Assumptions will shift, and the reliability and accuracy of the data will become accepted as fact; decisions made based on facts are made more quickly, with greater confidence, and better results.

**Conclusion**

It’s clearly fair to say that there has been a recent fundamental shift in the way we think about Business Intelligence. The new tools we have at our disposal are driving a new approach to the management and use of our business data. Whilst there are many reasons why a data warehouse may be a great thing, it’s impossible to ignore the new wave of products which give us new ways of working with information. Data discovery is here to stay; the approachability and ease of use that it offers is a powerful driver, and there are undoubtedly businesses for which this type of tool is a perfect fit.

The traditional approach has become more widely referred to as ‘enterprise Business Intelligence’; indeed, with the complexity of its implementation and the typical cost, it’s no surprise that larger organisations find it easier to afford. Data warehouses still have a place in this model, and will almost certainly remain a key part of this type of reporting and management information platform.
We can’t ignore the proliferation of data within business and the wider world, however. Data sources are growing in number, size, and intricacy and the result is a situation where the very structure of an enterprise Business Intelligence approach, whilst being a great strength, can also become its biggest weakness. Reacting to the pace of change can be difficult when using this model and the rapidity of development offered by the data discovery tool offers significant benefits here.

So the data warehouse isn’t dead. Perhaps its importance has been diminished somewhat, and it may no longer be the de facto solution to every Business Intelligence problem as which it was once perceived. Increasingly, the future of the data warehouse looks to be as a source of data alongside others, no longer the only source; the rapidly changing, unstructured nature of new information sources definitely requires us to think outside the data-warehouse-box.

I believe the data warehouse will still be around for a while, and its value will continue to be delivered through integration and consolidation of multiple disparate sources of more slowly changing data. In the high-speed world in which we are living, however, the data discovery tool and its fresh approach to delivering the information we crave is here to stay.

The use of both data-discovery and more traditional BI tools alongside one another is increasing too; one of the applications of the data-discovery approach is to use its rapid development cycle to prototype or test new outputs or visualisations intended to ultimately be delivered by the enterprise platform. This capitalises on perhaps one of the most important positives of the data-warehouse approach, which is that the repository and the data contained therein become independent of the user interface; a data-warehouse created using a common database engine can be accessed, and its content exploited for reports, scorecards, visualisations, or exports, using any one of a number of tools. The ability to ‘mix-’n-match’ approaches, with data discovery for power users wanting quick answers provided by one tool, and dynamic scorecards provided by enterprise BI for senior management has led to the situation where the data discovery tools’ value can be enhanced by a data warehouse, rather than replacing it.

I have previously made the point, and still firmly believe that there is no right answer to the Business Intelligence question. Every business is different, and the mechanism by which they choose to accommodate their information requirements will be correspondingly so; the only certainty is that not taking advantage of the data sources at your disposal is to not take advantage of one of the most valuable assets you hold. After all, in Sir Francis Bacon’s words, “knowledge is power”.
About the Author: Dan Burrows

Senior Consultant | Business Technology
Joined in 2008

Dan joined Waterstons from a varied career background, working as a mainframe programmer and database administrator in the insurance industry, and later as a hardware and software support engineer before a spell in management within the NHS. He specialises in business consultancy, working primarily in the healthcare and airport sectors.

Key Specialisms:

- Business Consultancy
  - IT Strategy Review
  - System Selection
  - Business Intelligence and Reporting
- Systems Analysis
  - Systems Integration
  - Requirements Analysis and Specification
  - Information and Data Quality Analysis

Key Projects:

Project management of Business Intelligence platform development for an international airport

ERP and Service Management system selection project for marine and land-based engineering firm

Comparative systems analysis for international food manufacturing business