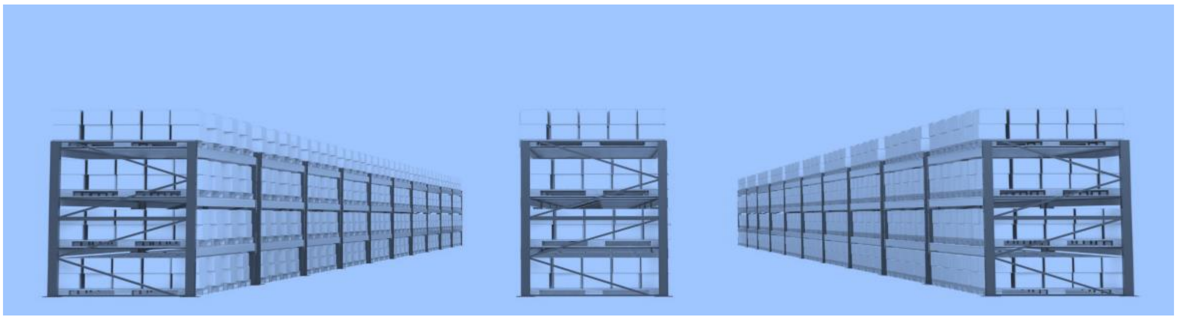
quocirca
Business and IT Analysis



Optimising the data warehouse

Dealing with large volumes of mixed data to give better business insights

October 2013

Data warehouses are struggling to keep pace with the growth of data volumes and the different types of information that organisations need to deal with. Extract, transform and load (ETL) activities on datasets can cause major issues for organisations, with growing ETL time windows impacting the capability to carry out meaningful analysis. New approaches are needed to ensure that an organisation gains the insights it needs from its data.

Clive Longbottom
Quocirca Ltd
Tel : +44 118 948 3360
Email: Clive.Longbottom@Quocirca.com

Rob Bamforth
Quocirca Ltd
Tel: +44 126 726 8138
Email: Rob.Bamforth@Quocirca.com

pentaho

quocirca

Optimising the data warehouse

Dealing with large volumes of mixed data to give better business insights

The data warehouse is struggling to keep pace with an organisation's data needs. Extraction, transformation and loading (ETL) tasks are being stretched beyond their capabilities, and organisations wedded to maintaining a standard ETL, relational data approach are finding that they have to cut corners in order to get some level of results from their analysis and reporting. However, a different approach can add a lot more value – and make the data warehouse work again.

Old approaches to data warehousing are struggling

Relational data needs to be supported by non-relational information to ensure that decisions are made against as much information as possible. The use of extended value chains across organisations means that other sources of data are being accessed. ETL windows are shrinking – yet ETL tasks are trying to extend at the same time and encompass new emerging technologies. Something has to change to make business intelligence workable.

Standard ETL is no longer the answer

The use of standard ETL approaches no longer meets the needs of the organisation. Many now have to cut back on the amount of data being moved to the data warehouse in order to fit in with the time windows available to them. Decisions are therefore being based on a sub-set of the available data – and do not include the non-relational information that now makes up a sizeable proportion of an organisation's electronic assets.

Mapping multiple, diverse data sources is an increasing problem

If only two or three different data sources are used in a data warehouse, mapping the various data schemas is relatively easy. However, in today's environment there can be many different data sources. Maintaining the schemas can be a major issue – and a change in any one of the sources can lead to major issues in the data warehouse itself. A more automated means of maintaining schemas needs to be found through the use of suitable systems.

Mixing different data types (relational and non-relational) is important

Basing corporate decisions only on data held within relational systems is dangerous. With an increasing amount of information held outside of a standard database, these sources have to be included in the mix to ensure that decisions are made against as much of the available information as possible. The use of binary large objects (BLObs) is not the best way to do this.

Bringing in Hadoop can help solve many of the issues

The use of a specialised non-relational store, such as Hadoop, can provide a platform where a mixed environment of relational and non-relational data can be brought together in a meaningful manner. The use of MapReduce can help bring the size of the data warehouse back under control – and can make the end analysis and reporting of the available information far faster.

Hadoop is not particularly user friendly

Although Hadoop has many strengths, it is not a tool for the novice. To ensure that Hadoop is implemented correctly needs deep domain expertise – and such skills are difficult to find. Some systems are now coming to market where the complexities of Hadoop are being effectively hidden from the user, making it far easier to implement and manage.

An optimised data warehouse requires a new approach

New technologies are available that can make a data warehouse far more effective. However, it has to be simple to implement and run, as well as easy for end users to interact with and use directly to gain the insights required. Vendors are now providing systems that hide the complexities of the underlying technology and speed the integration and aggregation of the multiple different relational and non-relational data sources.

Conclusions

Existing approaches to data warehousing are increasingly unfit for purpose. The creation of ever-larger warehouses or data marts counts against any meaningful analysis being carried out, and the financial performance of organisations will suffer as they struggle to compete in markets where others have adopted a more modern approach. Data warehouses have to be brought up to date: a different approach is required in how multiple relational and non-relational data sources are dealt with so that the amount of data needing to be analysed and reported against in the warehouse is better controlled and managed. However, the technologies chosen must also be easy for IT staff to implement and manage – the complexities of the underlying technologies must be hidden to provide a system that is easy for all to use.

Background

According to figures released by Oracle in 2012, the amount of stored data is growing at 40% per annum and will hit 45 zettabytes (ZB – or 45,000,000,000,000,000,000,000 bytes) by 2020. Of course, this will not all be held in one database, but will be spread across millions of databases and file stores held within private and public datacentres.

Along with existing data created in enterprise applications such as customer relationship management (CRM) and enterprise resource management (ERP) systems, systems that used to be separate from the main IT environment, such as production lines, security systems and other sensor and monitoring systems, are standardising towards a TCP/IP backbone. This will result in more data that will be available to be gathered alongside this 'standard' data – and that will need to be suitably analysed.

Market pressures will continue to stress how organisations need to deal with data. Gone are the days when a rear-view mirror look at how the business has performed was good enough. Such business reporting has been superseded by a need for business intelligence and from there for a need to look to the future through the use of business analytics.

Data volumes are also not being helped from the increasing load of legal compliance that is required from organisations. In many cases, there are legal requirements to store data for a minimum of 3 years – and, in areas such as healthcare, this may extend to many decades. Many organisations now store data on a 'just in case' basis, refusing to delete anything in case it may be needed as part of a legal disclosure demand sometime in the future. This 'store everything forever' approach makes data analytics more difficult as out-of-date and old, unused data still has to be included in many analyses.

On top of this is the need to ensure that data is highly available, and that data volumes are backed up and maintained so that on the failure of any part of the existing compute platform, the business can be back up and working as rapidly as possible. Many organisations are now finding that their backup windows – the time required for a full image of a data volume to be copied – is now exceeding the available time. Although data backup vendors are addressing this issue through the use of technologies such as snapshots and incremental backups, full copies of data are still required at some stage.

This report looks at how large data volumes can be optimised in a manner that allows an organisation to get the most from its data assets, enabling business analytics to be used to provide deep insights into future possibilities and so to better compete in the markets.

Defining 'big data'

2012 and 2013 have been the years of big data. However, different vendors and commentators have defined the term to fit in with what they already provide, and this has led to a degree of confusion in the market.

From Quocirca's point of view, big data is a case of being able to deal with the various "V"s:

- **Volume** – the amount of data that has to be dealt with
- **Variety** – the different types of data being dealt with, such as that already held within a formal data store; office and other documents held within a standard file server; video, image and sound-based files; data held outside of an organisation's control, such as that held in social networks and other web-based systems.
- **Velocity** – both upstream and downstream. On the upstream side, just how fast does incoming data need to be dealt with? For example, sensor data or data coming from firewalls or other network devices can be extremely rapid and will need to be captured and dealt with in a different manner to data that is more transactional. On the downstream side, how quickly does the information receiver need the output from the

system? As machine-to-machine (M2M) automation increases, the need for real-time data results grows. Even where it is machine-to-person (M2P), expectations have changed from batch jobs where results were provided several hours to several days after a report was initiated to an expectation of results being available within seconds to minutes.

- **Veracity** – just how accurate are the results? This can be a mix based on how 'clean' the data is in the first place (garbage in, garbage out), on how many available data sources were included and also on how in-depth the analysis of the data was.
- **Value** – what will the user and/or the organisation gain from the analysis anyway? If the level of value to be gained is small compared against the cost of analysis, is it worth carrying out the analysis in the first place, or leaving resources available for higher value tasks?

As such, data volume is just one part of a more complex big data environment. Many organisations will still have a need to deal with large amounts of data held within formal database systems – but this is not really a 'big data' issue, it is more of a 'lot of data' issue. However, even where the issue seems to be just in dealing with relational data growth, the capability to include other data sources from non-relational and file stores can lead to deeper business insights. Providing the capability for users to mix different data sources together in a timely, but accurate, way must also be considered. Quocirca strongly advises that any chosen solution for dealing with predicted relational data growth does not preclude the capability to include and manage non-relational data – including third party data sets that can add significantly to the value of analysis and reports.

The data warehouse issue

The term "data warehouse" goes back to the 1980s, when researchers from IBM created an architectural model for how data flowed from operational systems to decision support environments. This work built on concepts discussed by Bill Inmon in the 1970s.

The idea was that data from the disparate data stores would be moved into a separate data warehouse where the required analysis and reporting could be carried out without impacting the operational systems themselves.

However, each operational system would have its own data schema, and bringing them all together as a coherent data set was a problem. Therefore, a common schema (a normalised data model) needed to be designed and then the data from each operational store needed to be extracted, transformed such that it met the new schema design and then loaded into the data warehouse. During the transform step, other actions such as data cleansing and a level of data consolidation would also be carried out. This action is known as ETL. In most cases, ETL is carried out as a batch process, run on a scheduled basis to maintain data consistency and to minimise the impact on the operational systems.

However, such an approach to ETL is full of slow and unwieldy actions. Extraction of existing data across a growing number of sources can lead to issues with data quality, and non-relational data is difficult to deal with. The transform step can lead to issues with context and how metadata is used and can be slow when dealing with large datasets. Data cleansing can extend the transform step if the original data stores are not particularly well managed in the first place. The load step can lead to major time issues if intelligence has not been applied to minimise the amount of data that is being dealt with. Each step needs expertise in dealing with the data and the business rules that need to be applied to it. Storage can become an issue as data volumes grow through the need to store the original data, any staging data and the resulting data in the data warehouse itself.

Organisations need a mix of reporting capabilities, ranging from the cyclical reporting in areas such as financial and compliance reporting through to the more immediate needs of responding to market forces and completion. If a data warehouse is dependent purely on batch-based ETL and relational data, then the analysis of the data can only be as up-to-date as the last batch job. This is often suitable for cyclical reporting, but is not good for more immediate reporting needs. On top of this, as upstream data velocity increases, the volumes of data needed to be dealt with by the ETL job also increases – in many cases to the point where the time window available for ETL is no longer sufficient, and the choice then comes down to running almost continuous ETL jobs or to compromise on data integrity by not using all the data available from the data sources.

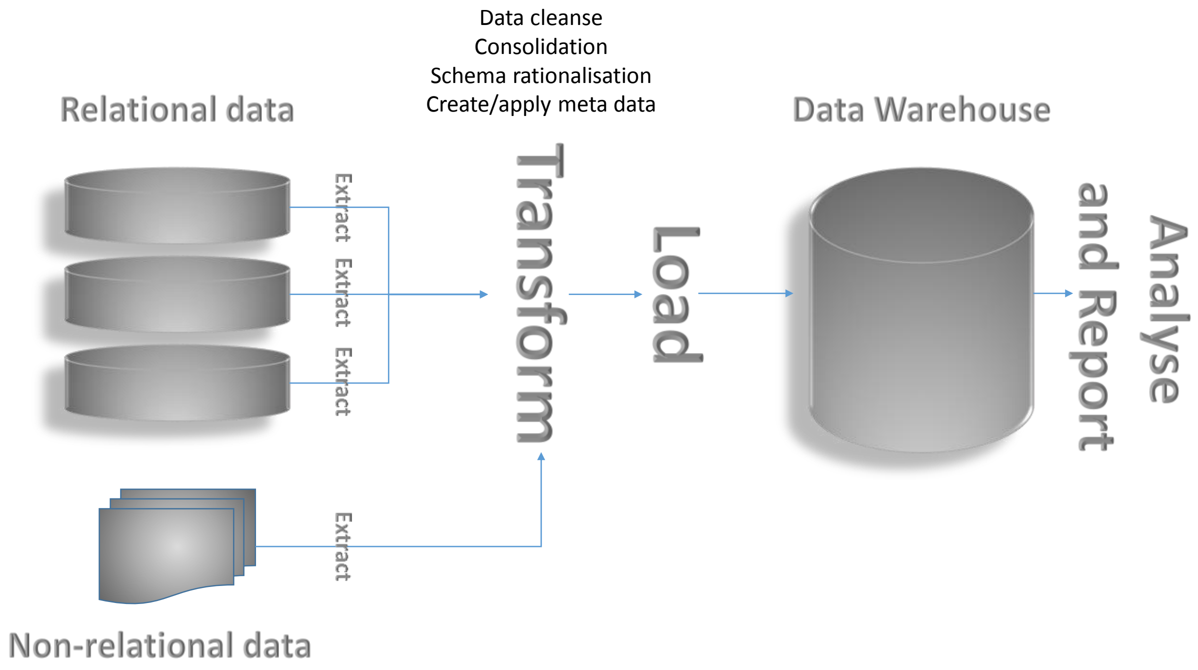


Figure 1: Traditional data warehouse approach

It has also been the case that data warehouses were dependent on the use of a 'standard' database, yet the increasing need to deal with different types of data has led to problems in dealing with less structured data sources, which tend to end up being stored within the database as a binary large object (BLOb). The use of BLObs tends to have an adverse impact on how the data can be analysed, as indexing the content generally means creating massive amounts of metadata that needs to be held referring to the BLOb itself.

It seems that existing approaches to dealing with enterprise data systems through a standard approach to data warehousing are no longer fit for purpose – so what can be done?

A new approach

Previous approaches to dealing with data warehousing were all predicated around using a standard relational database such as Oracle, IBM DB2 or Microsoft SQL Server. However, over the past couple of years, new ways of dealing with data have come to market.

Leading this is an open-source project under the Apache Foundation banner, called Hadoop. Hadoop is a platform that provides a range of functions, two of which are very useful in a data warehouse environment. In the first instance, it can act as a data store, and secondly, it can take actions on the data that provide direct benefits to both IT and the business.

Hadoop is a highly scalable system that can be implemented on a network of standard commodity x86 servers resulting in a Hadoop cluster. It is built on a filing system commonly referred to as HDFS (Hadoop distributed file system) and utilises a technology developed by Google called MapReduce.

MapReduce works on a Hadoop cluster to carry out a combination of two actions. The first is the "map" function, which performs a filter and sort to create sets of data (such as customer ID numbers or first line of addresses) and then uses a "reduce" function, which carries out actions like counting the frequency of the same items across the mapped data.

Through the use of MapReduce, the amount of information that needs to be analysed can be reduced massively by keeping redundant and dead information from the main data warehouse store, enabling analysis to be carried out far faster. Indeed, with data volumes being brought down to such small levels, the whole of the data warehouse can be loaded into memory, so giving a massive uplift in how rapidly insights can be achieved into an organisation's main data.

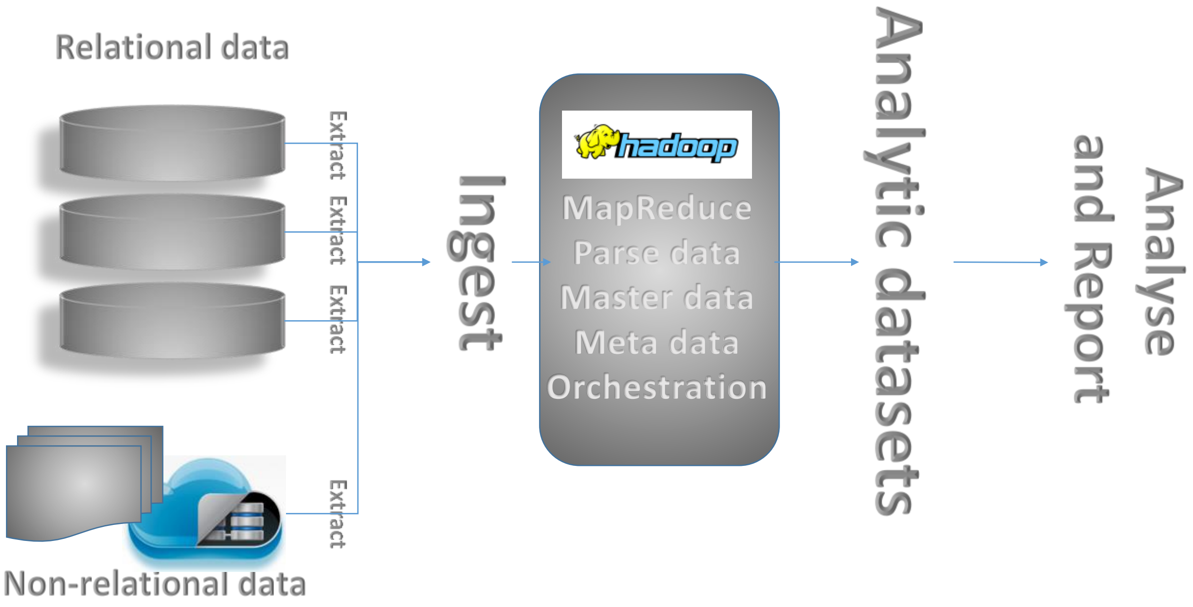


Figure 2: Data warehouse with Hadoop

However, Hadoop is not a tool for the unwary. It requires a high degree of knowledge in order to be able to make the most of it – knowledge that remains scarce at the moment. Actions need to be scripted or coded and, without the right skills, mistakes can be made that can either make Hadoop work inefficiently or can result in data errors.

However, some vendors are now coming to market with Hadoop-based systems that hide the complexities of Hadoop from both end users and technical staff. The use of visual front ends can provide a means of defining how different data sets need to be mapped, enriched and modelled, while interactive analytical front ends enable end users to carry out their own analysis and reporting activities.

As an example of how a Hadoop-based approach can help an organisation, consider the need for a retail organisation to run a report on where in the country are the customers who spend the most on their premium range of food items. The company will have a database of goods purchased, many of which will be from identifiable shoppers based on their use of loyalty cards. The purchase database will generally not be the same database as the loyalty card database. Historically, the total purchase database and the total loyalty card database would be extracted separately, transformed to meet a different data schema and then loaded into a new single database in the data warehouse which could then be analysed to get what the company requires.

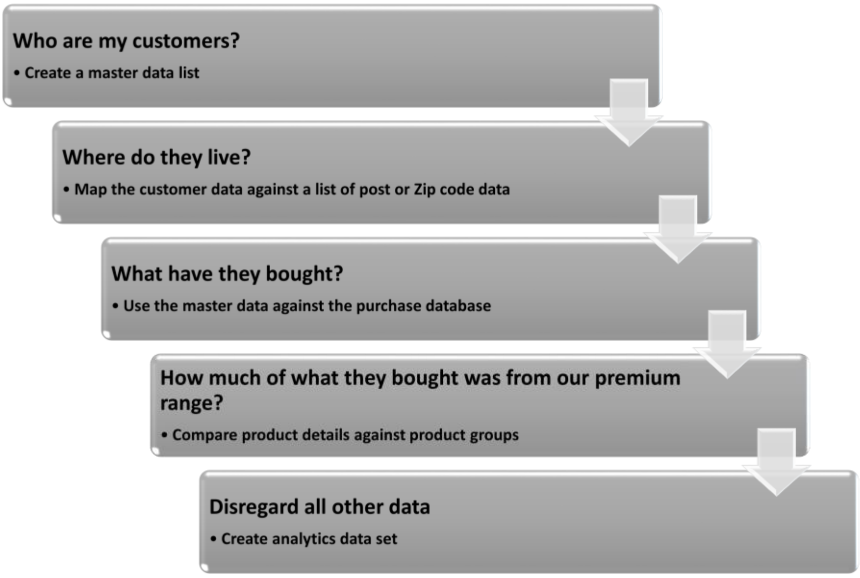

Who are my customers?
• Create a master data list
Where do they live?
• Map the customer data against a list of post or Zip code data
What have they bought?
• Use the master data against the purchase database
How much of what they bought was from our premium range?
• Compare product details against product groups
Disregard all other data
• Create analytics data set

Figure 3: Simple analytics requirement against different data sets

In the new data warehouse, the structured (relational) data and the unstructured (non-relational) data are 'ingested' into a Hadoop cluster. Using master data constructs (for example, customer name or inventory object), the raw data can then be parsed and reduced to create analytic datasets that become the 'active' part of the data warehouse. The analysis and reporting is carried out against these datasets, rather than against the overall aggregated total data of the multiple source databases.

Data sets from outside can be included through simple data connectors that understand external data provider formats. Indeed, even web-based search data can be ingested and parsed in a contextually aware manner to add considerable value to the resulting analytic datasets.

This new architecture does not require any change to existing enterprise applications – it still utilises relational data from the databases underpinning the applications. However, it adds an easier and more granular capability to deal with new data types – and provides a direct means of gaining greater control over how the data warehouse is managed and how it provides value to the business.

Business benefits

The use of an optimised data warehouse approach does not just create an elegant technology platform. The use of Hadoop can help through consolidation of data at a highly granular level so that less data then needs to be analysed and reported against, leading to faster reporting capabilities. By gaining faster insights into corporate data, better decisions can be made, with better capabilities to respond to changes in the market.

The inclusion of different data types is also important. For example, few organisations have the capability or the money needed to maintain a competitive market database when the likes of Dun and Bradstreet are around, or a full database of consumer demographics when companies like Equifax are around to fully concentrate on this. Likewise, unless your organisation needs multi-layered maps with, for example, detailed plans of where utilities run, then it makes far more sense to use the likes of Google or Bing Maps rather than PBBI MapInfo. The use of an inclusive solution based on Hadoop means that organisations can utilise data sets from anywhere. Integration should be able to be carried out into the chosen system through simple connectors, rather than hard-wired coded connections. Through this means, organisations will gain a highly flexible, future-proof, data-agnostic system that will provide business benefits now and in the future.

However, this only works where the tools are made available to end users so that they can use them themselves. Predicating a system on the capabilities of users to code or create scripts to gain access to the data that they need to analyse, or to script the report itself, will just lead to users not using the systems provided. Wherever possible, simple visual front ends need to be provided that are intuitive and shield end users from complexity. At the same time, the simplified mixing of data by end users must be managed at the source to ensure accuracy and timeliness of data supporting decisions based on governed data.

Conclusions

Data growth is not going to slow down. Indeed, it is highly likely that the amount of data an organisation will have to deal with will continue to grow, and that the different data types and sources will also grow exponentially.

Trying to deal with this by just throwing more resources at an existing data warehouse will not work. ETL windows will continue to grow; with organisations having to make decisions on how much of the underlying data is actually used in any analysis so as to meet the requirements of the business and the capabilities of the technology platform.

The importance of different data types continues to grow as well. Building a data warehouse strategy on purely a relational database basis is doomed to failure – less structured, non-relational data such as office documents, voice and video will also need to be included in the mix in an intelligent and meaningful manner.

Batch ETL will remain an important part of an organisation’s analysis and reporting mix for some time, but this needs to be bolstered with a more inclusive system that allows for greater data inclusivity and ease of use for end users.

Dealing with such a mix of needs requires a rethink of how a data warehouse works. Using a system that enables early data consolidation against a scalable commodity platform means that the initial extract and transformation stages can be carried out more intelligently, leading to the resulting analysis and reporting being carried out in a much faster timeframe.

Any chosen system should hide the underlying technology complexity from the users and the IT people implementing and running the system. Hadoop is a rapidly changing system, and an organisation that is dependent on deep skills within their workforce to keep up to speed with what is happening in the area will be disadvantaged through the possibility of being held to ransom by those who feel that they have the power through knowledge of how such technology works. A well-chosen system will provide all the benefits of Hadoop without the technical complexities being a problem. Visual front ends and automated processes will make it easier for businesses to reap the benefits of a more modern data warehouse.

Travian Games Case Study

Travian Games is a German company that creates browser-based and massively multiplayer online games (MMOGs). With over 120 million users across more than 50 countries using 42 languages, Travian Games has to deal with massive amounts of data.

Business issue

Travian Games uses a 'freemium' model for its games. Users can try out a free version and Travian Games then tries to move the user up from this free version to a paid version through the offer of advanced features. To be effective, the offer must be compelling, must be 'of the moment' and create a sense of urgency in the player to move to the paid version.

Travian Games uses advanced analysis of player data in order to continuously improve the game experience and identify the triggers for moving a player from a free to a paid version. Player behaviour needs to be monitored and reported on a continuous basis, leading to massive amounts of data needing to be analysed.

As players number have grown rapidly, Travian Games has struggled to keep pace with the need to analyse the data using existing approaches of proprietary tools across different databases of information.

Solution thought process

Travian Games needed a system that would pull together all corporate data, including not only the player data, but also the sales and customer data into a single data warehouse. The chosen system would have to be self-contained: it would need to be able to carry out the various aspects of data integration, analysis and reporting without the need for 'bolt-on' extras. It would also need to be easy to use so that business workers and those working on the continuous development of the games could run ad hoc reports as they saw fit.

Solution chosen

Travian Games was already a user of open source systems, with its games development platform being based on Linux. It evaluated several open source systems as to their capabilities in meeting its needs.

Travian Games decided that Pentaho provided what it needed. A system based on the Pentaho Business Analytics Platform was implemented:

- Pentaho Data integration (PDI) provides the integration across the multiple different relational and non-relational data sources that Travian Games needed to deal with, carrying out the ETL tasks to create the single data warehouse.
- Pentaho Anlayzer provides the ad hoc and more formal analysis of the data warehouse.
- Pentaho Reporting provides end users with clear insights into all relevant corporate data.

The overall solution is built on a Linux Mint Debian platform, with MySQL and MongoDB data sources as well as an Infobright database for the Pentaho standard interfaces.

Business benefits

Travian Games' developers now have a tool that enables them to precisely analyse game play and so drive continuous improvement in the game experience. They can also more accurately predict the trigger points for gaining upgrades from free to paid models and so have been able to drive greater revenues for Travian Games.

Business users are now able to run reports that they need directly against a complete set of corporate data and information sets. The user-friendly nature of the Pentaho solution allows ad hoc reports to be run by anyone at any time, so enabling faster decisions to be made against a more complete information set than previously.

About Pentaho

Pentaho is delivering the future of business analytics. Pentaho's open source heritage drives its continued innovation in a modern, integrated, embeddable platform built for the future of analytics, including diverse and big data requirements. Powerful business analytics are made easy with Pentaho's cost-effective suite for data access, visualization, integration, analysis and mining. For a free evaluation, download Pentaho Business Analytics at www.pentaho.com/get-started.

pentaho

REPORT NOTE:
This report has been written independently by Quocirca Ltd to provide an overview of the issues facing organisations seeking to maximise the effectiveness of today's dynamic workforce.

The report draws on Quocirca's extensive knowledge of the technology and business arenas, and provides advice on the approach that organisations should take to create a more effective and efficient environment for future growth.

About Quocirca

Quocirca is a primary research and analysis company specialising in the business impact of information technology and communications (ITC). With world-wide, native language reach, Quocirca provides in-depth insights into the views of buyers and influencers in large, mid-sized and small organisations. Its analyst team is made up of real-world practitioners with first-hand experience of ITC delivery who continuously research and track the industry and its real usage in the markets.

Through researching perceptions, Quocirca uncovers the real hurdles to technology adoption – the personal and political aspects of an organisation's environment and the pressures of the need for demonstrable business value in any implementation. This capability to uncover and report back on the end-user perceptions in the market enables Quocirca to provide advice on the realities of technology adoption, not the promises.

Quocirca research is always pragmatic, business orientated and conducted in the context of the bigger picture. ITC has the ability to transform businesses and the processes that drive them, but often fails to do so. Quocirca's mission is to help organisations improve their success rate in process enablement through better levels of understanding and the adoption of the correct technologies at the correct time.

Quocirca has a pro-active primary research programme, regularly surveying users, purchasers and resellers of ITC products and services on emerging, evolving and maturing technologies. Over time, Quocirca has built a picture of long term investment trends, providing invaluable information for the whole of the ITC community.

Quocirca works with global and local providers of ITC products and services to help them deliver on the promise that ITC holds for business. Quocirca's clients include Oracle, IBM, CA, O2, T-Mobile, HP, Xerox, Ricoh and Symantec, along with other large and medium sized vendors, service providers and more specialist firms.

Details of Quocirca's work and the services it offers can be found at http://www.quocirca.com

Disclaimer:
This report has been written independently by Quocirca Ltd. During the preparation of this report, Quocirca may have used a number of sources for the information and views provided. Although Quocirca has attempted wherever possible to validate the information received from each vendor, Quocirca cannot be held responsible for any errors in information received in this manner.

Although Quocirca has taken what steps it can to ensure that the information provided in this report is true and reflects real market conditions, Quocirca cannot take any responsibility for the ultimate reliability of the details presented. Therefore, Quocirca expressly disclaims all warranties and claims as to the validity of the data presented here, including any and all consequential losses incurred by any organisation or individual taking any action based on such data and advice.

All brand and product names are recognised and acknowledged as trademarks or service marks of their respective holders.