www.triangleinformationmanagement.com @triangleinfor

01256 860056

Common Data Warehouse Issues

It takes forever to load

After the initial project to deliver the data warehouse has finished, the data volumes increase over time. This often leads to ever increasing overnight load times, with the common problem that people cannot run reports until well into the working day because the warehouse is still building.

This is always a symptom of one or more of the following design flaws:

The initial warehouse was created as a full refresh every day…

The quickest, simplest and therefore cheapest data warehouse designs just truncate and reload all or most of the data daily. Fast enough for the first few months, but a year or two later the 5-minute load times have increased to several hours. This is a form of technical debt, a shortcut which needs eventually to be addressed by implementing incremental data loading.

Be aware that if you are truncating your data every day you will be causing massive fragmentation of your database files and disks and this may lead to further performance issues.

The indexing is non-existent or ill conceived

Despite improvements in database tools, efficient indexing is still easy to get wrong and can cause performance issues. Whilst an index can improve performance of a query, the same index will also create an overhead for an insert.

Often we will find a data warehouse with a huge number of indexes which have been automatically created by a tuning wizard leading to poor load times as all the indexes get updated.

This is often compounded by a poor data warehouse design and reporting tool. In a correctly designed data warehouse utilising star schemas the indexing strategy is straightforward to implement and a good reporting tool will be able to identify the correct columns to join and group by as required. If you have many indexes on each table, the chances are you are degrading your load times. Conversely if you have no indexes you will be sub optimal on query response times.

The data types are inconsistent or incorrect…

Often overlooked, the data types and character sets chosen in a data warehouse can have a negative effect on performance and quality. It is common to find warehouses where the data types for a single attribute vary wildly from table to table – the same attribute being stored as a number, varchar or date in different tables. This often happens when the ETL tool can create tables automatically based upon a few rows of data.

Getting the data types wrong can mean at best longer load times and at worst data corruption.
Don't expect your database to tell you it's truncating your data as it converts between a NVARCHAR(250) and VARCHAR(250) or corrupting it between character sets that are not fully compatible.

The fastest and safest way to move data is to match the data types and character set of the source system. A conversion will slow down your load times and could have unintended data corruption. If you do need to convert data types always do it explicitly with a function, do not rely on implicit conversions in the database.

The design approach is confused or inappropriate…

The largest recommendation we regularly make is to reconsider the appropriate data warehouse design with respect to best practise and then fully commit to one approach or the other.

There are two clear alternate approaches to building a data warehouse. The CIF Corporate Information Factory approach recommended by Bill Inmon and the requirements driven, Dimensional Methodology, Star Schema based approach recommended by Ralph Kimball. Both recommend the creation of conformed star schemas for reporting from (Inmon has adopted Kimball's star schemas in his latest books) - apart from this, any similarity ends.

Inmon CIF Approach

The Inmon CIF approach is typically what businesses think of when they hear the term data warehousing and has led to the perception that data warehousing is enormously expensive, complex and prone to fail. Called the top down or data driven approach as it includes all information from business systems on the assumption it may be useful combined into in a single 3rd normal form or normalised enterprise data warehouse.

In this methodology, disposable data marts (conformed star schemas) are created only after the complete enterprise data warehouse has been created.

If you are truly looking to build a permanent data warehouse based upon the business (and not the systems which may come and go) then this is a worthwhile endeavour, however the

creation of the enterprise data warehouse is typically the stumbling block for many attempting this as it adds a huge amount of time and complexity to the project.

However very often what was intended to be an enterprise data warehouse simply becomes another staging area, maybe renamed as an integration layer and the true benefits of thisapproach are never realised, but the added complexity and overhead remain. We regularly see

'Enterprise Data Warehouses' which are not normalised, source data from only one source system, add no real benefit and may be an unnecessary overhead. Often these types of data warehouse are the result of consultancies wanting to sell a larger and more complex project than was required.

Ralph Kimball's Dimensional Approach

Migrating to Ralph Kimball's Dimensional approach can help streamline and simplify a failing data warehouse. Removing the need for a 3rd Normal form enterprise data warehouse and its bottom-up requirements driven approach, means that you can cut out the dead wood, reduced load times and concentrate on the business need.

In Kimball's approach, we only need a staging area in which to perform any necessary staging, integration and data quality, and a star schema area containing de-normalised data in dimensions and facts. We do not have an enterprise data warehouse.

Inmon Vs Kimball

The Inmon approach is the Rolls Royce solution and is best suited to global companies who must integrate multiple different ERP systems and are often required to add new, yet unknown data sources following acquisitions or migrations from one ERP to another. For these companies, it makes sense to invest in a source independent enterprise database based upon their business.

For most organisations resources and time are an issue, so the enterprise data warehouse is a luxury. In most situations, we can easily integrate data from different source systems directly into the star schemas if required, without the need for an enterprise database

The warehouse is carrying too much technical debt…

As the months and years go by it is common for quick fixes and additions to creep into the warehouse that address an urgent business need. It may have been intended later to return to the issues and create a better solution but they are often forgotten.

When you get a build-up of these quick fixes it is a form of technical debt. The fixes may degrade the purity of the design, break rules or just complicate the design making support harder and load times slower. We have seen warehouse designs hacked to the extent data

loops through the staging, presentation layers back to staging again or combines data from several layers in a single query. This makes problem solving and even scheduling overly complex.

At some point the accumulation of technical debt may lead to the demise of the data warehouse as it becomes more and more complex. Any data warehouse that cannot be left scheduled to run on its own or which requires daily intervention is probably suffering from technical debt.

Best practise these days would be to set aside one day in 5 and all free time to proactively work on reducing the technical debt. It will make your life easier!

Are you suffering from one or more of these issues?

If one or more of the above issues are familiar in your organisation and you feel your existing data warehouse is not meeting your business requirements or expectations, ask us for a review and go forward plan.