What can we do if our existing Data Warehouse is not delivering?

RESOLVING POOR PERFORMANCE IN MATURE DATA WAREHOUSES AND ESTABLISHED ANALYTICS PLATFORMS

SUMMARY

A national hospitality chain engaged with Infuse Data Solutions to deliver a scheme of work to improve the performance of their mature Data Warehouse. The SQL based Data Warehouse that had been continuously developed for over 15 years was struggling to meet the needs and expectations of the business. ETL (Extract, Transform, Load) routines were taking until the middle of the business day to complete, with regular failures and key reports were not being refreshed until the end of the business day. Infuse consultants embarked on a 3-month challenge to drastically improve this situation.

THE SITUATION

The Data Warehouse has been developed over the course of many years using best practice principles by a number of different data warehouse and ETL developers. As the business and business requirements for data and analytics grew and evolved, as did the scale and scope of the data warehouse.

The data warehouse itself consists of dozens of Fact, Dimension and Reporting tables that support the business' business intelligence reports and analytics.

The ETL routines consisted of Stored Procedures orchestrated by SQL Server Integration Services. These ETL routines have changed over time to accommodate new requirements at the same time as the data volumes have increased resulting in the degradation of the performance of these routines.

AT A GLANCE

Business Problems

- Essential activity reports provided too late or not at all.
- Operational decisions made without insight.
- Wastage and diminished profitability based on poor decisions.
- Low confidence in reporting across the organisation.

Technical Challenges

- Improve run time of ETL routines
- Reduce time taken to update SSAS cubes
- Improve performance of key business reports
- Minimise failures

Environment

- Microsoft SQL Server
- Stored Procedures & User-defined functions
- SQL Views
- SQL Server Integration Services

What can we do if our existing Data Warehouse is not delivering?

THE SOLUTION

REFACTORING QUERIES AND PROCEDURES

By benchmarking the performance of units of code within the Stored Procedures and queries in the ETL we were able to identify the long running areas of code that may be able to be re-factored to achieve better performance. This might include examining the query path to determine whether the tables are being accessed in the most efficient way. Often re-phrasing the SQL statement can achieve performance gains by using a more efficient approach to a achieve the same outcome.

REVIEW INDEX STRATEGY

A well-designed index strategy can make considerable difference to query performance. Conversely an over indexed database can have a detrimental impact on performance. Indexing key columns in star-schema is standard but many data warehouse ETL routines will have complex logic in lower layers of the data warehouse. By measuring the usage and effectiveness of existing indexes, then dropping unnecessary indexes we often experience performance improvements to Insert and Update statements. When the logic requires the return a small set of data, index optimalisation can be a powerful tool in the kitbag.

DATA ARCHIVING

As history builds up over time, it is usually only once data volumes become an issue that an archiving strategy is considered, by which time performance degradation is already taking place. By properly designing data files and filegroups, partition schemes and functions and partitioning tables and indexes archiving data can be a simple task that requires minimal manual intervention and have a multitude of benefits including speeding up processing and querying data in the data warehouse as well as considerable cost savings over time.

SYSTEM RESOURCES

Monitoring system resources such as CPU and memory during peak and idle times can give a good indication whether performance can be improved by increasing the resources available to the applications. Data warehouse workloads read large amounts of data into available cache in memory. If there is no available space in cache then this increases the amount of disk i/o which has a significant impact on performance. By tracking the usage of memory and CPU during different times, we are able accurately determine the optimum resource requirements for the application.

BENEFITS

- Using the techniques and approach we've detailed here, along with others, we were able to achieve a considerable improvement to this company's ETL and data warehouse performance and efficiency.
- ETL time reduced from 6 hours to 2 hours.
  Data warehouse now updated before reports need to be run.
- Key reports now run in minutes rather than hours.
  Information in the hands of the users sooner and any errors can be rectified and reports re-run in a timely fashion.
- Cost savings. Archived data now moved to cold storage, freeing up space in expensive high availability hot storage.