incorta

THE DEATH OF THE STAR SCHEMA THREE KEY INNOVATIONS DRIVING ITS RAPID DEMISE

There are better ways to analyze complex data—ways that eliminate the need for star schemas and their resulting maintenance nightmare.

Intelligent Solutions, Inc.

Intelligent Solutions, Inc.
Solve your business puzzles with Intelligent Solutions

The Death of the Star Schema

Three Key Innovations Driving Its Rapid Demise

Claudia Imhoff, Ph.D.

Sponsored By
incorta.

Table of Contents

Introduction

Relational databases in the 1980s were typically designed using the Codd-Date rules for data normalization. It was the most efficient way to store the data used for OLTP operations. As Business Intelligence (BI) and multidimensional analysis became popular, the relational databases began to have performance issues when multiple joins were requested to resolve a query. The implementation of star schema designs was a brilliant way to get around the performance issues and ensure that multidimensional queries could be resolved in a reasonable amount of time.

The star schema is a physical instantiation of a multi-join process – a denormalization of 3rd Normal Form (3NF) data for performance reasons. Dimensions surround the resulting measures and, as long as business users described their needs up front and in enough detail, the dimensions behaved rationally; all was good, and the query did not require anything other than the dimensions offered.

Unfortunately, the analytic process is frequently unpredictable. Business users think up an unimaginable number of ways to query the data, and the data itself often changes in unpredictable ways. These result in the constant need for new dimensions, new and mostly redundant star schemas, maintenance difficulties in handling slowly changing dimensions, and other significant problems. These, in turn, cause the analytical environment to become overly complex and difficult to maintain, with long lead times for new capabilities – all resulting in an unsatisfactory environment for both the users and those maintaining it.

The data warehouse – still the foundation for analytics usage – requires the extraction of data from analytically hostile environments (e.g., operational applications and external sources of data) followed by the integration and quality processing of that data into single versions used in analytics. To add the star schema design and its physical creation on top of an already complex ETL process has been the bane of most data warehouse developers.

There IS a Better Way!

In the 30 years or so since BI and data warehousing were introduced, there have been massive advances in database technologies to support analytical environments. Two of these advances have caused significant increases in performance: columnar storage of data and the leveraging of in-memory capabilities. Fortunately, these innovations have improved performance in two ways. First, columnar storage has more effective compression, yielding fewer disk seeks, faster scan times, and a more efficient use of CPU for data in memory. Second, the combination of columnar storage with in-memory caching eliminated a significant bottleneck by creating agile and more flexible data warehouse environments.

There's a third innovation, however, that is delivering significant query performance gains. This innovation revisits the query algorithm and eliminates some of the heavy lifting inside the data warehouse that is the root cause of engine slowdown. Incorta's Direct Data Mapping™ is an example of innovation in this space and presents a glimpse of future data warehousing. It can reduce response times for complex queries from hours to seconds.

With the advances of columnar database technology, in-memory technology, and Direct Data Mapping™, physical star schemas are no longer needed. Yes, that's right. These three in combination massively increase performance in data warehouses, and remove the need to create physical joins of analytical data ahead of time—resulting in tremendous agility and flexibility while reducing time to value.

Benefits

The benefits of being free from the constraints of star schemas and their corresponding indexes – these rigid and inflexible database designs – are significant:

- A fast, flexible, virtual database design means business users can ask impromptu questions (with virtually unlimited dimensionality) that lead to far more insight and increased satisfaction with and usage of analytical environments.
- Business users can perform hundreds of queries on more complex data or drill down into very detailed operational data and be assured of acceptable response times – something they are not able to do with traditional data warehouse technologies.
- Removing the need for physical star schemas means reduced time, effort, and money spent on: (i) star schema design sessions, (ii) creation of the ETL processes that transform the data into star schema compliant designs, and (iii) the massive maintenance needed for these designs. With the recovered time formerly spent on the creation and maintenance of star schemas, developers can focus time on backlogged business user requests.
- Significant benefit comes from reducing the amount of substantial data storage needed due to data redundancy in star schemas and the creation of their indexes.
- Without the need for physical star schemas, developers can increase the availability and volumes of more complex source data for analytics.

While the ETL process does not go away it is certainly simplified by the elimination of unneeded processes used in creating star schemas and flattening or reshaping data. Data is extracted and then directly loaded into a modern data warehouse using data processing technologies and Python scripts like those available within Incorta. After these processes, business users can begin to ask whatever questions they want without being told the design won't support their query or having to pay severe performance penalties.

Getting Started

Many organizations have “legacy” data warehouse implementations based on physical star schema or 3NF database designs. To migrate to a modern environment, perform the following steps:

1. Evaluate your ETL processes to determine where the star schema bottlenecks are, decide which star schemas are particularly burdensome in terms of their creation and/or maintenance, and target these for migration to the new environment.
2. Group these troublesome star schemas into the business problems they solve. At first, you may choose to deal with just the star schema data (aggregates, measurements, dimensional data) for resolving a specific business problem or set of business problems.
3. Create a migration plan to move the set of star schema data used to resolve one business problem at a time to a new, modern environment. This gives you a quick win just by presenting the star schema data in a far more usable format – one that is modern, more performant, and far more flexible.
4. Next, begin analyzing the detailed data that went into the star schema data. This data can add even more flexibility and agility to the overall analytics environment. And again, there is minimal additional overhead in terms of the ETL process.
5. Finally, expand your data acquisition horizons to data that you may have thought was just beyond the reach of your development capabilities. In the new environment – where data volume, query performance, and time to delivery are not such big problems – you may find that the overall analytics environment can be greatly and easily enriched.

Keep in mind that when you reduce the burden of star schema design, creation, and maintenance, you can use that time to start reducing the overall backlog of analytical requests. You may also want to reexamine the need for cubing technology, since it too may no longer be needed.

For those of you fortunate enough to have a “green field” situation where very little or no analytical capabilities exist, you can quickly move to this new world of star schema-free analytic environments. You will still need a good understanding of your business users’ problems and the data they will need to solve these problems. You will also need to determine how much data integration and quality processing will be required to bring together the sets of needed data. But you have the very real advantage of never needing to design, produce, or maintain star schemas.

Summary

With innovative technologies, a rethinking of data warehouse design and processes is necessary. But don't throw the baby out with the bath water. Yes, you still need a data warehouse – it is your repository of data used for analytical activities. Yes, you still need ETL or some form of integration and data quality processing, just not as much as you needed in the past – there is no need to transform data into star schema or 3NF compliant designs.

Yes, you need to perform maintenance on the data stored in the data warehouse, but you are dealing with less data than in traditional data warehouses and with simpler data schemas. And finally, yes, you can now solve many of the problems you faced in the past by bringing better, faster, and more flexible decision-making to your organization.

More from the Sponsor

Typically, complex RDBMS queries are slow because data needs to be moved around – either in-memory or on disk – to generate the desired result set. Incorta's Direct Data Mapping™ completely circumvents this data movement and hence is able to return result sets within sub-seconds, no matter how large the data set or how complex the query. This approach dramatically speeds up insights by completely eliminating the need for star schemas and by paring down the unnecessary parts of ETL that are costly, cumbersome, and brittle. With Incorta, IT can be ultra-responsive and deliver new reports within minutes as opposed to weeks, and business users can feed their data curiosity by conducting real-time conversations with their data to make more accurate and timely decisions.

incorta.

Direct Data Mapping

Intelligent Solutions, Inc.

Solve your business puzzles with Intelligent Solutions