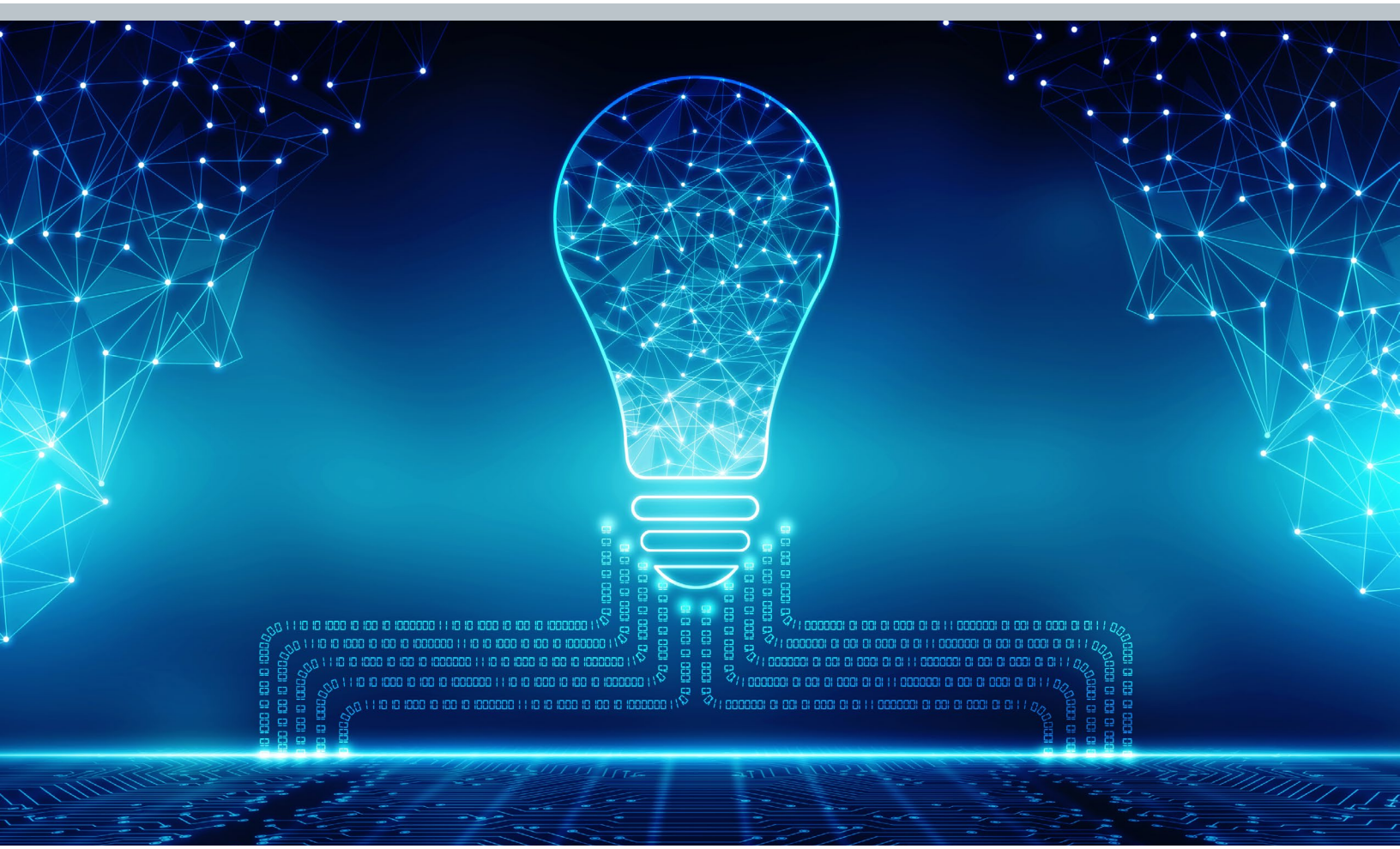
PAPER

Redmond MAGAZINE



EVOLUTION AND CONVERGENCE: HYBRID DATA

By Don Jones

actian™
activate your data

www.actian.com

Modern businesses rely heavily on data to make strategic decisions, run the business, and pivot based on market trends and demands. But, to achieve this, modern businesses presently have to deal with an overwhelming array of different databases, each specialized to a single purpose: transactional, analytical, and so on. Further, each database typically requires an immense amount of planning, architecture, and tuning in order to deliver the information that the business needs in order to act. The sheer quantity and specialization of that planning, architecture, and tuning often prevents the very agility that the underlying data was supposed to deliver in the first place.

In addition, the line between "transactional" and "analytics" is blurring for many businesses. Businesses simply need the information and data that they need, and they need it quickly; the technological distinctions between different "kinds" of databases isn't a distinction that the business itself cares about. Yet, businesses are "stuck" with different kinds of databases, simply due to the way these database technologies have evolved to this point.

Enterprises find themselves in a hybrid data situation, with data lakes and swamps creating disparate pools of structured and unstructured information that needs to be organized and analyzed for insights. Being in a hybrid data environment essentially creates data silos, with each different database contributing something to the business, but with each of them operating independently. To use all that data efficiently, people need to learn where to look for what they need, and the databases need to anticipate those needs, none of which benefits the business in the end.

Dealing with hybrid data requires a new class of databases that take advantage of modern hardware capabilities. This new, modern database can offer the agility

DEALING WITH HYBRID DATA REQUIRES A NEW CLASS OF DATABASES THAT TAKE ADVANTAGE OF MODERN HARDWARE CAPABILITIES.

Cover Image: Blackboard/Shutterstock.com

businesses need, with less human overhead, while maintaining a standards-friendly approach to data management and analysis.

FROM SPECIALIZATION TO GENERAL PURPOSE, AND BACK AGAIN

We all know that the computer industry is, ultimately, an industry of cycles and seasons. The first computers were extremely specialized, used for little more than planning rocket launches. So, too, the first databases were extremely specialized, often catering specifically to online transaction processing.

THERE'S A LESSON TO TAKE FROM COMPUTER HARDWARE. MODERN HARDWARE IS BOTH GENERAL-PURPOSE AND SPECIALIZED.

Over time, more general-purpose computers were developed. As is often the case in a "one size fits all" approach, these general-purpose computers compromised toward breadth of function, sometimes paying a cost of lower performance. Again, databases followed suit, developing "one size fits all" databases that were suitable for a broad variety of tasks, perhaps without the higher performance of their more specialized predecessors.

That cycle continued, as new breeds of specialized computers were invented to deal with performance-intensive tasks like CGI. Databases also specialized, with dozens of variants emerging to deal with specific types of data storage, analysis, transformation, and so on. But this specialization created its own problems, forcing companies to adopt dozens of open-source projects, database platforms, query languages, and more in order to deal with all of their different business needs.

Critically, each specialized database often needed to anticipate the business' query needs, with the database more or less custom-designed to support whatever the businesses needed to do with their data. Index design, query tuning, even table structure all became exceedingly mission-critical tasks requiring immense human input. All that custom-tuning also made it difficult or impossible to anticipate new business needs. Simply asking for a new type of dashboard could require the creation of a whole new database, custom-designed and -tuned for that specific task. For many businesses, it's all a bit much. Entire teams of data analysts, scientists, and specialists lovingly nurture each unique, special database, and the business accepts the commensurate loss in agility and increased overhead.

But there's a lesson to take from computer hardware. Modern hardware is both general-purpose and specialized. Processor cores are tuned for specific

tasks, such as high- or low-power tasks, or for mathematically intensive graphics processing. Unlike the last round of specialized computing, though, modern hardware lets users and developers take a "hands-off" approach. They can treat the hardware as general purpose, and just throw workloads at it. The hardware and software are smart enough to recognize specific workload patterns, and engage its various specialized subsystems as needed. It's all the benefit of a high-performing, highly specialized system, with all the ease of a general-purpose system. That's where databases need to go next.

Imagine a data store capable of holding many different kinds of data structures: row-based data, column-based structures, and so on. A single interface accepts queries, and dynamically figures out the best engine and structure to run it through, just as modern computer hardware engages its own specialized subsystems. It's a hybrid database approach.

Just the idea is compelling. Modern processor design has allowed developers to spend less time worrying about which hardware needs to do which bit of computer, enabling those developers to spend more of their brainpower creating amazing software. So, too, can hybrid databases enable businesses to spend less brainpower designing and hand-tuning specialized databases, and instead spend that brainpower creating true innovation for the business.

HARDWARE PLAYS A KEY ROLE

Databases have traditionally evolved along with their hardware platforms, since data is one of the most intense workloads a computer will ever deal with. The shift toward hybrid data is no different: by leveraging powerful, single-instruction, streamed number-crunching features of modern processor cores, hybrid solutions can make decisions and analyze massive data sets incredibly fast. Originally designed for high-performance computing and compute-intense graphics modeling, these processor features also enable a new generation of faster, specialized data analysis on general-purpose hardware.

This is where a new approach to databases, called vectorized, can come into play. Using principles like vector processing, and single-instruction, multiple data (SIMD) capabilities of newer CPU hardware, vectorized query execution can process the same operation on multiple chunks of data simultaneously, and exploit data-level parallelism.

For years, most databases of all kinds have leveraged basic hardware advances, from multi-processor systems to hyper-threading and multi-core processors. Even so, many "modern" database platforms still utilize a relatively small percentage of the computing power available to them. This "spare power" is where vector databases can realize equivalent or better performance in a self-tuning, self-architecting model. By

leveraging every trick the hardware knows how to perform, a hybrid database can deliver performance without a massive capital investment in planning and hand-tuning.

"HANDS-OFF" APPROACH

The real key to this new approach is an elimination of the specialized planning, architecting, and tuning so common to legacy data platforms. No longer do you need a specialized data team to construct task-specific schemas, design dashboard-specific data shards, or analyze and create indexes to support specific queries. No longer do you need a new database, with its own extract/transform/load tasks, to support every new query or database that the business dreams up. Instead, you just feed your queries to the hybrid database and—like its hybrid-processor hardware cousin—the database figures out what to do.

Ad-hoc queries, long the bane of database performance, simply run, and they run fast. Internally, the hybrid database figures out the base data structures to use, the best database engine to run, and it "just works." Hybrid database systems can even learn over time, automatically and continually optimizing themselves for the most common workloads they see, while still executing one-off queries with ease. The business value is striking and immediately obvious: new queries don't require massive new planning processes. Existing queries don't require constant tuning as their data sets grow and the exact business need changes. Data just works. The business is free to be agile, to ask whatever questions it needs, to build whatever dashboards it needs, and to focus on the business, not on the database structures and tuning.

THE BUSINESS VALUE IS STRIKING AND IMMEDIATELY OBVIOUS: NEW QUERIES DON'T REQUIRE MASSIVE NEW PLANNING PROCESSES.

Even maintaining this new hybrid database system should be as "hands-off" as possible. When an update is available, you shouldn't have to worry that the database will be offline for days on end, or that when it comes back up all of your previous careful fine-tuning will stop working, resulting in a slower system than before the update. Updates should be invisible, and regardless of changes to the internal system structures, the database performance should continue to be the same, or better, than what you're used to.

STANDARDS-FRIENDLY

Of course, no new database platform is going to be embraced if it requires businesses to rip-and-replace everything they already know and have. A new, hybrid database system must support your existing standards and connections.

SQL should still be the query language of choice, immediately enabling connectivity to existing business systems, and creating familiarity for developers, business analysts, and data scientists throughout your organization. Core connectivity to common enterprise database platforms, like Hadoop and Spark, are also a must, helping to preserve your existing investments, reports, dashboards, and extensibility to Machine Learning and AI.

Choice should also be standard, meaning a hybrid database system needs to be able to run on Windows or Linux, or in the cloud platform of your choice. This helps a hybrid database fit into your business, rather than demanding your business change to accommodate a new solution.

CONCLUSION

We've been taught by years—decades, even—of experience that hand-designing and hand-tuning databases is the only way to get the information as quickly as businesses need. But it's time to face the fact that that baseline assumption is simply no longer true. The database, in hybrid form, has evolved.

Hybrid data represents the simple facts on the ground for most modern businesses: we have a huge amount of data stored in an enormous, and growing, array of forms and schemas. A single, combined database, that can handle both operational and analytics needs quickly and efficiently, is the simple convergence of our hybrid into a single data source, a convergence which was perhaps inevitable. What's needed to happen, to support that convergence, is the end of hand-designed schemas, hand-tuned queries, and discrete "silos" of data that each require their own language, their own tuning tricks, their own architectures, and so on.

Instead, a truly modern database simply makes it all "just work." Queries go in, and the database itself figures out the best way to execute that query, right then. It learns over time, ensuring that the most common query workloads are as optimized as possible. And it does it all without that massive amount of human investment previously needed, freeing up human ingenuity and innovation to truly benefit the business.

Find out more:
www.actian.com

actian™
activate your data