

The AI Database

By: Alexei Balaganski

Date: May 6, 2024

As I was writing about Oracle's new SQL Firewall some time ago, I had no idea it will be published on the same day the company officially announced the general availability of its flagship database product, Oracle Database 23ai. What a coincidence! And what a twist with the new name!

I have mixed feelings about the name, by the way. On the one hand, it appears to be the most unnecessary, marketing-driven change ever. I can vividly imagine thousands of DBAs, developers, consultants, journalists, and other IT professionals rolling their eyes and scratching their heads. Now they must update all their documentation and writings to reflect it, and that time could have been spent on more productive things.

On the other hand, in contrast to many other vendors throwing "AI" into their product names, Oracle does have quite a lot to show for it. Artificial intelligence was, of course, a substantial part of the database for quite some time already. The concept of the Autonomous Database was introduced back in 2017. Machine learning algorithms have been a part of the database core for years as well. However, in 2024, the only kind of AI everyone is talking about is Generative AI. And so, among over 300 major new features, the new release introduces several new ones crucial for not just enabling GenAI capabilities for business applications, but implementing them in a universal, globally scalable, secure and, last but not least, compliant manner.

The most notable addition is AI Vector Search, a set of capabilities to enable native support for generating, storing, indexing, and querying vector data directly in the database. This is, of course, a crucial requirement for implementing retrieval-augmented generation (RAG) to enhance the accuracy of large language model responses with additional information from external sources (such as an organization's own sensitive information that it does not want to share with LLM operators in any other way).

Now Oracle Database 23ai can natively support storing vector embeddings for unstructured content in the same table as existing relational data. This allows for creating complex queries across them, combining traditional SQL with semantic search, as well as geospatial information, graphs, and so on – which is impossible with a standalone vector database. In contrast, Oracle's converged database approach keeps all data in a single location, where it is uniformly protected by layers of data security controls, including the SQL Firewall.

When running on the Oracle Exadata platform, its underlying smart storage technology will even natively accelerate vector search operations to run AI applications at a massive scale. Curiously, even if you are not yet ready to migrate your application to the 23ai release, it is still possible to replicate existing data from other sources using the GoldenGate 23ai service and let Oracle Cloud handle the AI operations for you.

Another major feature introduced in the new release is JSON Relational Duality, a technology that aims to solve the decades-long debate between the fans of the relational and the document data models. Until now, the developers were forced to make an early design decision between the efficiency and consistency of SQL and the simplicity and flexibility of NoSQL, and to resort to additional middleware layers to address the shortcomings of both approaches.

With JSON Relational Duality Views (what a mouthful of a name!), developers no longer need to make this choice. It is now possible to store data in relational format but access it in the form of JSON documents. A view can be declared across multiple tables with a structure described using the familiar GraphQL syntax. The database then takes care of all the rest, including automated table updates when documents are modified, lock-free concurrency control to support stateless operations, and making the data available across a range of APIs, from SQL to REST or even MongoDB.

While the idea might sound somewhat trivial in hindsight, implementing it in a scalable, reliable, and standardized way has required years of research and development. But now the capability is officially available, and developers are encouraged to try it – even the free edition of the database includes the feature, along with the similarly improved Property Graph Views.

I have no intention to mention every of the 300 features introduced in the new release, but one thing that is especially close to me both as an analyst covering data security and compliance solutions and as a citizen of the European Union is Oracle's Globally Distributed Database. With the introduction of built-in RAFT-based replication, the new release brings the concept of database sharding to a new level of scalability and performance. A global, hyperscale database that is distributed and replicated across multiple geographical locations in real-time is now a reality – and the data within it can be transparently localized according to complex rules.

For example, all information linked to EU citizens will be stored only in datacenters located in Germany, enforcing the EU data sovereignty regulations, while the data related to Indian citizens will be only kept within the borders of India. And yet, for business applications, the entire customer base will appear as a single table, enabling efficient but still compliant transactions and analytics. With new privacy regulations being introduced constantly, adapting existing applications becomes a matter of just adding new sharding rules to the database – no need to change anything in the business logic.

All these new capabilities are now officially available in the Oracle Cloud, both in their public regions and in the Cloud@Customer private cloud, as well as in Microsoft Azure as a part of Oracle's and Microsoft's joint Oracle Database@Azure offering. The on-prem availability is yet to be announced.

Copyright

©2024 KuppingerCole Analysts AG all rights reserved. Reproduction and distribution of this publication in any form is forbidden unless prior written permission. All conclusions, recommendations, and predictions in this document represent KuppingerCole's initial view. Through gathering more information and performing deep analysis, positions presented in this document will be subject to refinements or even major changes. KuppingerCole disclaims all warranties as to the completeness, accuracy and/or adequacy of this information. Even if KuppingerCole research documents may discuss legal issues related to information security and technology, KuppingerCole does not provide any legal services or advice and its publications shall not be used as such. KuppingerCole shall have no liability for errors or inadequacies in the information contained in this document. Any opinion expressed may be subject to change without notice. All product and company names are trademarks™ or registered® trademarks of their respective holders. Use of them does not imply any affiliation with or endorsement by them.

KuppingerCole Analysts support IT professionals with outstanding expertise in defining IT strategies and in relevant decision-making processes. As a leading analyst company, KuppingerCole provides first-hand vendor-neutral information. Our services allow you to feel comfortable and secure in taking decisions essential to your business.

KuppingerCole, founded in 2004, is a global, independent analyst organization headquartered in Europe. We specialize in providing vendor-neutral advice, expertise, thought leadership, and practical relevance in Cybersecurity, Digital Identity & IAM (Identity and Access Management), Cloud Risk and Security, and Artificial Intelligence, as well as for all technologies fostering Digital Transformation. We support companies, corporate users, integrators, and software manufacturers in meeting both tactical and strategic challenges and making better decisions for the success of their business. Maintaining a balance between immediate implementation and long-term viability is at the heart of our philosophy.

For further information, please contact clients@kuppingercole.com.