The database buyer’s guide

Practical questions every business should ask when choosing open source and proprietary databases

Contents

The expanding range of data management options from open source, proprietary and IoT-focused databases offers unique opportunities for businesses to collect, manage and use data in solutions that are more tailored to their exact needs.

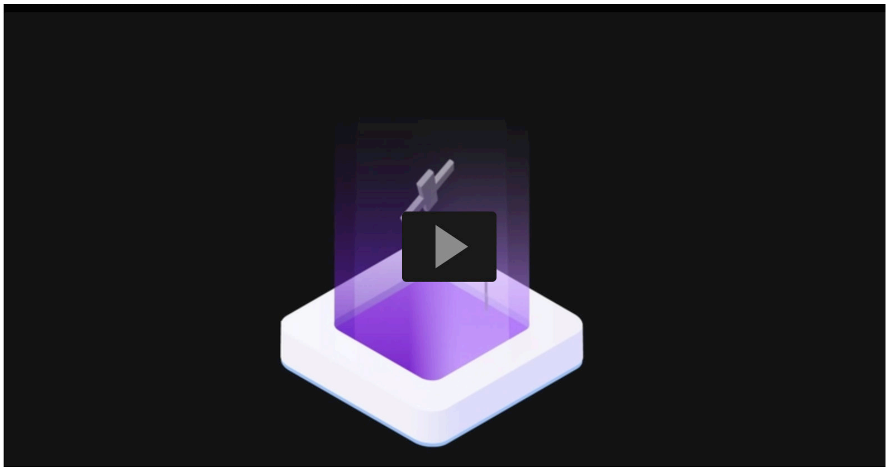
74% of organizations already use two or more database types.[1] Yet these options are not always easy to navigate or integrate. Challenges arise due to siloed data spread across multiple clouds, on-premises deployments, and trouble supporting the variety of databases and workloads that are chosen. It is similarly difficult to select a proprietary database capable of handling the increasing demand for data for AI applications and incorporating AI for greater database optimization. Comparing the multitude of open source options adds another layer of complexity, and the unique considerations of an IoT database may come as a surprise to those looking deeper into the category.

If you are trying to navigate the database landscape, you may find it helpful to ask yourself the questions below and pose them to potential database vendors as you try to find the right database for your unique needs.

How will I break down silos between multiple databases?

With multiple databases in an information architecture, it's important to ensure data does not remain siloed. All data must be available so that insights can be holistically informed.

Data virtualization is the best option; it ensures all data is accessible at a single point without needing to move or combine that data. Virtualization is considerably more efficient than extract, transform, load (ETL) processes with regard to both costs and employees' time. According to a Forrester Total Economic Impact report, using the right data and AI platform with data virtualization could reduce ETL requests by 25 to 65%.[2]



Watch the video to learn more about data virtualization with IBM Db2 on Cloud Pak for Data (2:03) →

Are different support teams needed for each database type?

Using a different support team for each database can quickly become frustrating. At best, multiple phone calls are needed and at worst your support request can turn into blame game with you as the mediator.

A simpler option is to choose a single vendor that offers multi-vendor support. In this way, after one phone call, a single team will work together to solve any issue you might experience across your architecture, even if that includes multiple open source and proprietary databases. Multi-vendor support has been observed to deliver reductions in maintenance and support spending (up to 25%), time spent on hardware support tasks (up to 20%), and time spent on vendor relationship management (up to 20%).[3]

How can I ensure data quality and discoverability across multiple databases?

To improve data quality and discoverability, seek out containerized databases that run on top of data and AI platforms featuring built-in governance. The common code and data virtualization capabilities found in the leading data and AI platforms allow governance to be applied consistently to data at a single access point, no matter which database it belongs to.

Data cleansing, metadata, user access and data lineage should all be considered because each plays a distinct role in delivering trusted data. And trusted data can be transformative in helping data scientists and other analysts spend less time searching for data and more time using it to drive insights. Trusted data also makes it easier to comply with any new data-related regulations. A recent study demonstrated that data virtualization and governance benefits such as these could lead to savings of USD 932,569 to USD 2,424,681.[4]

Do strong user group communities exist?

User group communities are essential for all types of databases. The members of these groups continue to push the boundaries of what's possible using databases and provide use cases that can serve as a blueprint to other businesses. In addition, their expertise can be leveraged if your users encounter a sticking point in their own architecture or just need some advice. Proprietary and IoT databases often have local meetups, so make sure to look for a vibrant community.

For open source databases the community is of even greater importance. Since the community itself contributes to the creation of the database and subsequent updates, it's impossible to overstate the importance of having a widespread, intelligent and well-organized community.

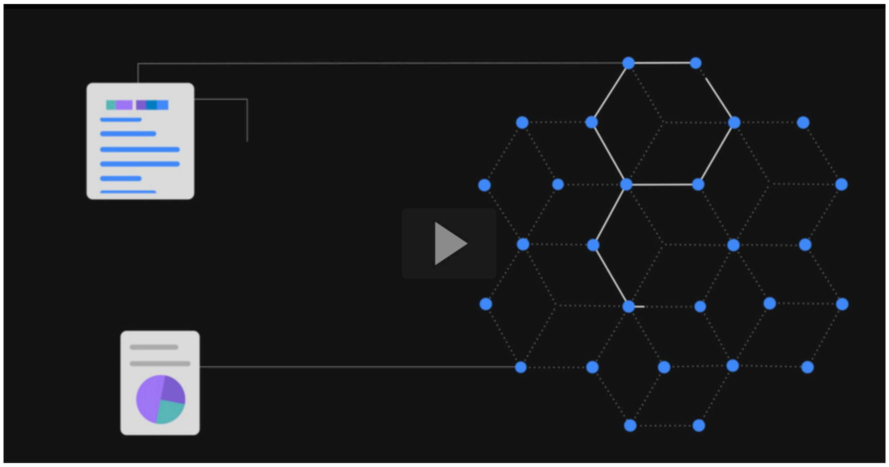
2. Assessing proprietary databases

How is the database optimized?

The way databases are optimized can have a considerable impact on their speed of insight and resulting database administrator (DBA) workload. Automated optimization is essential so that DBAs can focus on higher-value activities. Newer databases improve on this automation by incorporating machine learning (ML).

ML optimization capabilities monitor SQL performance over time and create optimized models for specific SQL statements. Therefore, more efficient access path cost estimates are created, leading to faster query execution and lower resource consumption. In some cases, up to ten times faster queries have been reported.[5]

Another aspect of optimizing the database is utilizing adaptive workload management; this is technology that automatically allocates data resources to handle a variety of workloads. The result is a reduction in time required for configuration and tuning, leading to overall database performance improvements of up to 30%.[6]



Watch the video to learn more about Machine Learning (ML) SQL optimization with IBM Db2 (2:50) →

Is the database locked into one deployment model?

The days when a database could be on-premises only or cloud only have long since passed. Each deployment location has unique advantages; on-premises options are valued for greater control while the cloud is popular for rapid flexibility and scalability. Now multicloud and multivendor deployments are gaining traction as well with 98% of surveyed organizations planning to use multiple hybrid clouds within three years[7] and 81% of respondents indicating they work with two or more cloud providers.[8]

Your best option is to use a containerized database built to run on top of a multicloud container development platform such as Red Hat® OpenShift®, which was highest ranked in the Q3 2020 Forrester Wave.[9] This configuration allows the database to run anywhere the platform does: on-premises, on one or multiple clouds, and even on clouds provided by a different vendor. Containerization also enables application portability, rapid deployment and scaling as well as easy management and configuration.

What level of developer support is offered?

It’s important to remember that database functionality doesn’t only apply to DBAs; developers must also be supported. Look for databases that natively support popular languages and libraries such as Python, JSON, GO, Ruby, PHP, Java, Node.js, Sequelize and Jupyter Notebooks. Developers will likely be familiar with one or more of these, reducing the amount of time needed for additional training.

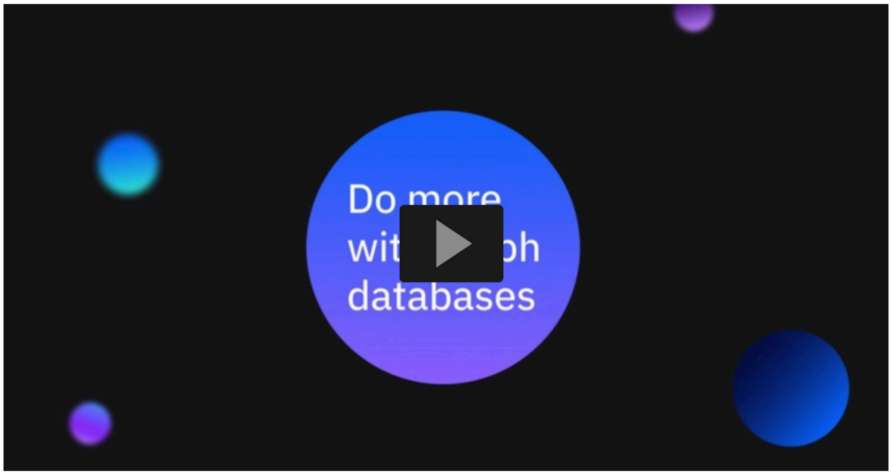
In addition, perpetually free developer trials should be available with full functionality; this minimizes the impacts of time-restricted trials that only provide access to some features. The trials should also be accompanied with product tours, tutorials, and other education to help developers get up to speed and deliver ROI faster or simply get them through a coding task they might need a bit of help on.

Does the database have multi-model functionality such as graph and SQL?

Although no database should be expected to cover the entirety of an organization’s data management needs, there are opportunities to combine some functionality in order to save costs and deliver better insight.

Consider the inclusion of graph database functionality within a standard operational database. Combining both with multi-model functionality means an organization no longer needs to pay for a standalone graph database, saving clients a minimum of USD 200,000 annually.[10] In addition, graph can run directly against relational data while SQL analytics can run directly against graph data.

This functionality is particularly important in graph-heavy industries such as healthcare and finance. In addition to graph, consider a database that accesses JSON and XML data with no migration or duplication; similar multi-model advantages can be achieved with them, too.



Watch the video to learn more about graph database Analytics functionality with IBM Db2 (2:38) →

Can the database be containerized for fast time-to-value?

An important aspect to evaluate in databases is their time-to-value, and more specifically how quickly the database can be deployed. Containers bundle applications with software libraries, making applications very portable and quickly deployable. Containers are fast, simple to use and compatible with automation tools as well, so the database can start delivering value right away. Containerizing the database also enables mutlicloud compatibility.

Does the database integrate with a data and AI platform?

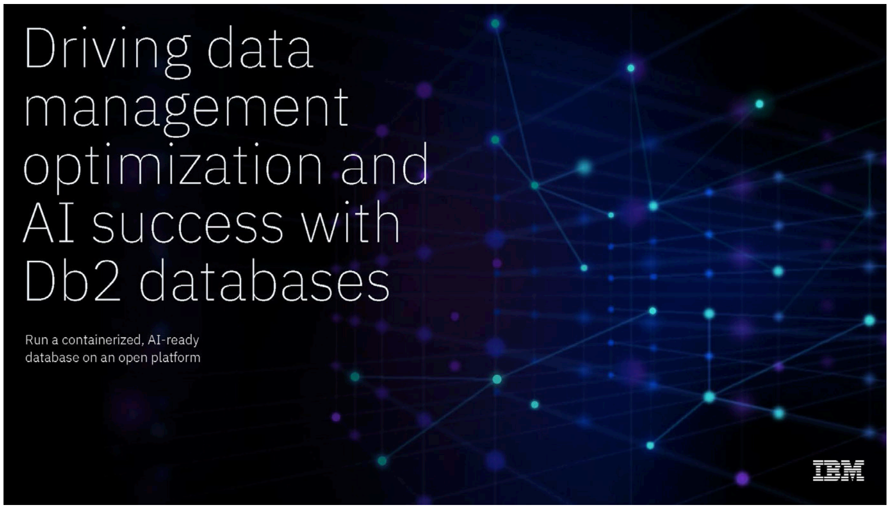
Integrating the database with an AI platform makes it easier and faster to move AI applications into production. Data is a key component of AI applications, and utilizing a database that is integrated to a data platform enables people in multiple job roles to leverage the data to develop, test and deploy AI applications.

Are tools like natural language querying available to make insight discovery easier?

Everyone from data scientists to business users can benefit from a simpler route to insights. The best databases use the latest technologies to enable this simplicity at all levels of the business. One way to do so is through natural language querying.

With natural language querying, users can input a question similar to how they would with a search engine like Google or Bing and receive visual insights based on underlying data. This gives business users the opportunity to gather insights more quickly than they might have otherwise, encouraging them to be more data-driven without needing to involve data scientists.

Of course, data scientists will benefit too, not just from a reduction in business user queries but by using natural language querying to obtain preliminary results or quickly test theories. These more agile checks will help inform the most profitable avenues of work without full-scale query or model building, driving greater efficiency.



Driving data management optimization and AI success with Db2 databases

Run a containerized, AI-ready database on an open platform

IBM

Read the ebook: Driving data management optimization and AI success with Db2 databases (20 pages) →

Are APIs available to help with integration?

Databases and data management are a foundational component of most applications. For that reason, it is important to consider a database’s support for application programming interfaces (APIs). For example, the availability of REST APIs enables developers to integrate their applications to data from everywhere in the organization.

Does the database have advanced authorization, encryption and key management?

Look for databases operating under Key Management Interoperability Protocol 1.1 that integrate with centralized key managers. The database should also have the option of being hosted in data centers around the world to meet global regulations. And the database should be in line with the latest NIST and FIPS standards, with advanced authorization, encryption, and comprehensive security controls.

Can the database support mission critical workloads with geographically dispersed clusters?

Geographically dispersed clusters can help mitigate or eliminate planned and unplanned outages, so you should consider this option to support heightened availability. Both change-queue-based and change data capture replication should be available as well to provide your organization choice in how you manage your HADR strategy. Because databases support mission-critical workloads, minimizing down time is paramount and the failover time should be measured in seconds.

Is the database capable of quick, independent scaling of storage and compute?

Deployment and scaling should not only be possible, but quick as well. Getting a database or additional clusters up and running in a matter of hours through traditional methods or minutes using a containerized approach will help streamline architecture growth. Techniques like compression and data skipping can also help lower the need for scaling by conserving space. And there should be independent scaling of storage and compute. Independent scaling helps tailor costs and capacity to a business's exact needs better than tiered plans can.

3. Assessing open source databases

Can open source meet enterprise scalability, availability and security standards?

Open source databases position businesses to capitalize cost-effectively on the vast amounts of data generated in today's world. The code is built and maintained by a large community and offered for free. With so many developers involved, high levels of innovation can be expected and bugs are found and fixed quickly. A wide variety of tools, plug-ins, and code is also available over the internet.

Yet the enterprise often has needs for additional functionality, security, and governance not provided by the open source community—and that's where vendors can play a role. Vendors can utilize and augment the code for enhanced security and governance and release it under an open source license. Thus, vendors are able to offer the quality, flexibility and scalability of open source and add the additional capabilities enterprises expect, such as a SQL-on-Hadoop engine.

When should MongoDB be considered?

MongoDB is a JSON document database built on scale-out architecture. It offers high-volume data storage, scalability and caching for real-time analytics, so it should be considered when developers need to build scalable applications using agile methodologies.

In addition, MongoDB's ability to ingest data without defining a schema provides considerable flexibility. Its architecture is based on collections and documents, allowing hierarchical relationships, storing arrays and representing other more complex structures more easily.

Read this white paper to learn more about MongoDB →

When should EDB be considered?

EDB PostgreSQL is a powerful, open source object-relational database system. Based on traditional RDBMS (relational database management system), PostgreSQL is best when you need a transactional, standards-compliant, ACID (atomicity, consistency, isolation and durability)-compliant, out-of-the-box database solution.

In addition, PostgreSQL integrates well with other tools and handles data integrity and complex operations with ease. It is very stable and easy to maintain with use cases ranging from e-commerce to data warehousing solutions.

Read this white paper to learn more about EDB →

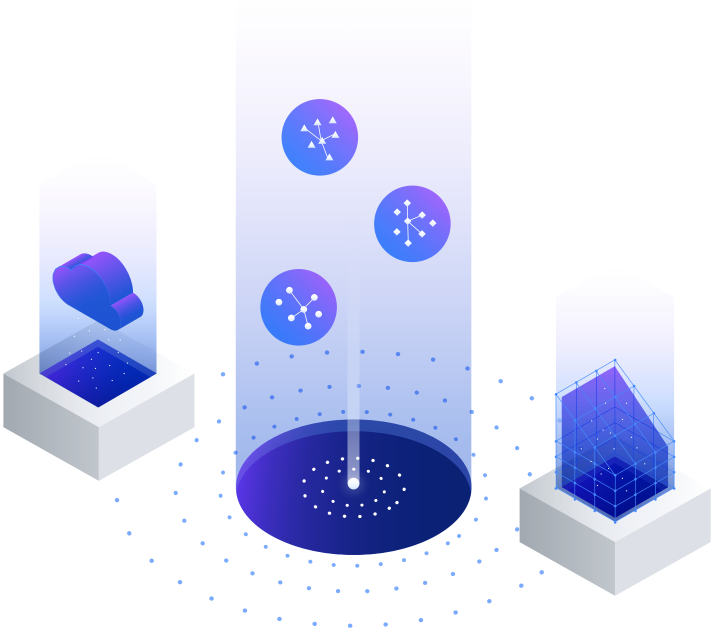
What level of expertise does the open source vendor have?

Anyone can package an open source database with a few other tools and claim to have an enterprise-grade open source solution. You should exercise care, looking into the vendor's level of expertise and history with open source. Seek out vendors that have not only sold, but contributed to multiple open source endeavors, vendors that frequently build on top of and integrate open source solutions with industry expertise so that they can design a solution that fits the unique needs of your business.

How flexible is open source with regard to deployment locations and data types?

Due to the inherent, community-based nature of open source solutions, they offer extremely high flexibility for deployments and data types. Because of how important hybrid cloud, multicloud and modernization are within the marketplace, they receive considerable attention from the open source community. The same can be said of popular development tools.

Open source solutions are also inherently transparent and can help avoid vendor lock-in. With an entire community able to view and work with the complete codebase, workload portability is second nature. This, in turn, means that a hybrid and multicloud environment is easier to create.



4. The benefits of an IoT-specific database

Is the database truly built for IoT or is IoT just an option?

A database that can handle IoT data is vastly different than a database built with IoT in mind. A good way to separate the two is to analyze whether the database is built to be embedded into IoT gateways. Embedding the database helps reduce latency by placing it closer to the source of the data, which is often crucial in IoT use cases.

Several factors determine how embeddable a database is, including a small footprint (see next question), automated administrative capabilities for easier installation and low-touch or no-touch management, and a dashboard that provides a simple way track the performance of all embedded databases through customizable alerts. Check to make sure that the IoT database you're considering has all of these components.

Does the database have a small enough footprint?

IoT database use cases often include being embedded in IoT gateways, meaning that a small footprint is crucial. While the size will vary based on individual needs and architecture, finding a solution capable of achieving a sub-100 MB footprint in optimal conditions is a great place to start.

Can the database perform time series and geospatial analysis?

Native time series and geospatial analysis capabilities are crucial for IoT databases because of common situations that generate IoT data. For example, machines sending regular or even irregular status updates will need to have that data tracked over time and data GPS locators will similarly need to be analyzed across space.

When comparing options, look for specialized compression capabilities to help with storage costs, subsecond timestamps for greater granularity, the ability to use your own coordinate system, and the length of time the vendor has offered time series and geospatial analysis as part of its database.

Read this paper to learn more about time series and geospatial analytics →

Has the database proven itself within large-scale organizations?

One of the best ways to determine an IoT database's level of performance is to analyze how it performs in some of the largest organizations. Ask your sales rep to walk you through use cases from Fortune 100 organizations. Even if you don't need as extensive an architecture, the database's ability to facilitate such a large implementation will demonstrate much about its simplicity and its usefulness in specific situations.

By asking the questions above, you'll get a better sense of the integration possible between databases, the AI-driven differentiators in proprietary databases, how open source options compare, and what considerations are most important for embedded, built-for-IoT databases.

5. More resources

By asking the questions above, you'll get a better sense of the integration possible between databases, the AI-driven differentiators in proprietary databases, how open source options compare, and what considerations are most important for embedded, built-for-IoT databases.

Dive deeper into some of the industry-leading databases across all of the categories considered here as well as a data and AI platform that can help integrate them with governance and analytics. Or schedule a free 30-minute conversation with a database expert to discuss your strategy.

Explore leading database solutions →

Discover a data and AI platform →

IBM Corporation
New Orchard Road
Armonk, NY 10504

Produced in the United States of America
July 2021

This document is current as of the initial date of publication and may be changed by IBM at any time. Not all offerings are available in every country in which IBM operates.

The performance data discussed herein is presented as derived under specific operating conditions. Actual results may vary. It is the user’s responsibility to evaluate and verify the operation of any other products or programs with IBM products and programs.

THE INFORMATION IN THIS DOCUMENT IS PROVIDED “AS IS” WITHOUT ANY WARRANTY, EXPRESS OR IMPLIED, INCLUDING WITHOUT ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND ANY WARRANTY OR CONDITION OF NON-INFRINGEMENT. IBM products are warranted according to the terms and conditions of the agreements under which they are provided.

The client is responsible for ensuring compliance with laws and regulations applicable to it. IBM does not provide legal advice or represent or warrant that its services or products will ensure that the client is in compliance with any law or regulation.

Statement of Good Security Practices: IT system security involves protecting systems and information through prevention, detection and response to improper access from within and outside your enterprise. Improper access can result in information being altered, destroyed, misappropriated or misused or can result in damage to or misuse of your systems, including for use in attacks on others. No IT system or product should be considered completely secure and no single product, service or security measure can be completely effective in preventing improper use or access. IBM systems, products and services are designed to be part of a lawful, comprehensive security approach, which will necessarily involve additional operational procedures, and may require other systems, products or services to be most effective. IBM DOES NOT WARRANT THAT ANY SYSTEMS, PRODUCTS OR SERVICES ARE IMMUNE FROM, OR WILL MAKE YOUR ENTERPRISE IMMUNE FROM, THE MALICIOUS OR ILLEGAL CONDUCT OF ANY PARTY.

YLQYMQWY

[1] ScaleGrid. 2019 Open Source Database Report: Top Databases, Public Cloud vs. On-Premise, Polyglot Persistence. June 2019.

[2] Forrester Consulting. Total Economic Impact™ of IBM Cloud Pak for Data. February 2020.

[3] Forrester Consulting. The Total Economic Impact of IBM Multivendor Support Services. 2019.

[4] Forrester Consulting. Total Economic Impact™ of IBM Cloud Pak for Data. February 2020.

[5] Based on IBM internal testing

[6] Based on IBM internal testing

[7] IBM Institute for Business Value. A blueprint for data in a multicloud world. October 2019.

[8] Laurence Goasduff. Why Organizations Choose a Multicloud Strategy. Gartner. May 2019.

[9] Forrester Research. The Forrester Wave™: Multicloud Container Development Platforms, Q3 2020.

[10] John Mark. Neo4j Enterprise Commercial Prices. January 2018.