List Price	Enterprise Price
$495	Please contact us

AN ECKERSON GROUP TOP TIER REPORT

Vector Databases

Leading Products You Should Know

By Sumit Pal and Wayne Eckerson

November 2024

Eckerson Group

ABOUT THE AUTHORS

Sumit Pal was a VP analyst in data management and analytics for Gartner, where he advised CTOs, CDOs, CDAOs, enterprise architects, and data architects on data strategy, data architectures, implementation, and choosing tools, frameworks, and vendors for building data platforms and end-to-end data-driven systems. Sumit has advised organizations and CDOs with guidance, insights, and analysis on best practices to solve data challenges.

Sumit has more than 30 years of experience in the software and data industry spanning companies from start-ups to enterprises. His book is SQL on Big Data: Technology, Architecture, and Innovation (Apress, 2016), and he has developed a massive open online course on big data. Sumit has also hiked to Mt. Everest Base Camp.

Wayne W. Eckerson is an international thought leader in the data and analytics field since the early 1990s. He is a noted speaker, sought-after consultant, and widely read author. Eckerson has conducted groundbreaking research studies, chaired numerous conferences, and consulted with organizations around the world during his 25 years in the field. Eckerson has also written two books: The Secrets of Analytical Leaders: Insights from Information Insiders (2012) and Performance Dashboards: Measuring, Monitoring, and Managing Your Business (2005/2010). He has degrees from Williams College and Wesleyan University.

ABOUT ECKERSON GROUP

Eckerson Group is a boutique consultancy that helps organizations get more value from data. We specialize in developing actionable data strategies and tailored data solutions for commercial and non-profit organizations. Our consultants each have 25+ years of hands-on experience and specialize in data architecture, data governance, data literacy, self-service, data products, and operating models. Our clients say we are hard-working, insightful, and humble. It stems from our love of data and desire to help organizations turn data into insights and action. **Explore what Eckerson Group can do for you!**

TABLE OF CONTENTS

Executive Summary 4

Introduction 6

Market Trends 9

- Vendor Comparison Chart 13
- Pinecone 14
- Qdrant 16
- Weaviate 17
- Chroma 19
- Zilliz 20

Conclusion 23

About Eckerson Group 24

EXECUTIVE SUMMARY

The generative AI boom has elevated the vector database from a niche position to a star player. A vector database stores and queries embeddings of data from unstructured data sources like text files, images, audio & video. Embeddings are mathematical representations of words, sentences, documents, as numerical vectors in a high-dimensional space that captures semantic meaning and relationships between these linguistic entities.

While companies have long used vector databases to recognize patterns and recommend actions, now they are using them to search documents as part of GenAI initiatives. Vector databases feed relevant content to language models (LMs), enabling companies to enrich prompts, fine-tune models, and govern outputs. Vector databases are becoming an increasingly important component of the generative AI stack. They support retrieval augmented generation, which supplement LM prompts with domain-specific content to make responses more accurate.

This report evaluates, scores, and compares the leading vector database products. The aim is to help zero in on the right vector database for your organization across use cases and domains. The report outlines the strengths and weaknesses of each product; and compares products along different dimensions in a visual quadrant chart. Most importantly, it publishes the worksheet that our analysts used to evaluate and score vendor products. You can modify the sheet and create your own scoring and ranking.

Introduction

INTRODUCTION

The decade of vector databases is here. It catapulted into fame with the advent of generative AI (GenAI) and the use of retrieval augmented generation (RAG), an architectural technique that organizations are using to increase the accuracy of GenAI output. Though vector models and databases have been there for over two decades, their timing with the arrival of LLMs and their synergy with foundation models have given them a prized spot in emerging GenAI architectures.

Vector databases promise faster, better, cheaper, vector search capabilities that leverage a variety of novel indices to search through millions of primarily unstructured data points. Vector databases support low-latency user queries based on semantic similarity rather than content accuracy.

Vectors are data points converted to fixed-length one-dimensional numerical representation in floating point numbers called an embedding. These numbers encode relationships in the content being represented across multiple dimensions. Vector databases have specialized novel indices to accelerate searching on these embeddings with the main goal to reduce latency of the searches at scale. Similarity search via embeddings is becoming a compelling use case across multiple applications from personalization, and recommendation to RAG applications.

The market is growing explosively and evolving quickly, though canonical reference architecture and best practices are yet to emerge. Incumbent database players (existing relational and non-relational) have joined the race too and have quickly added capabilities to store, search, retrieve, and update vectors

This Report

5
VENDORS

This report profiles five pure-play vector databases: Chroma, Pinecone, Qdrant, Weaviate, and Zilliz. This is not an exhaustive list. There are about a dozen vendors we could have profiled. However, these five vendors are good representations of this emerging category.

Each profile in this report examines the product's heritage, focus, differentiators, and challenges. The report also contains a quadrant chart that plots each vendor on two axes, "company scale" and "functionality", with the lower left area representing emerging players and the upper right area representing established players.

The heart of the report is the companion worksheet that we compiled for each product. Each worksheet contains our synthesized notes for 40+ evaluation criteria divided into four categories: 1) company background 2) company scale 3) product background and 4) product functionality. We scored the items in the second and fourth categories to create the quadrant chart described above. Feel free to add your own notes to the worksheet and score the products as you see fit.

Qualifications

In the beginning of 2024, we reached out to more than a dozen vendors that we felt represented this category. We reached out to both native vector database vendors as well as non-native vector database vendors (MongoDB, Datastax, Databricks, OpenText, Rockset). Databricks did not respond to our invitation. Rockset was acquired by OpenAI - so we had to drop it. Native vendors like Zilliz did not respond to our invite - though we have included it in our report as we felt it is a critical player in the market. We dropped the idea of including incumbent non-native vendors as it would not be a fair comparison to native vendors especially in terms of market and customer size and enterprise capabilities.

Methodology

Rather than sending out a worksheet to vendors and getting "Yes/No" responses to questions, we scheduled multiple briefings with each vendor to drill into the details behind our evaluation criteria. We captured the details in note form, which we subsequently validated and scrubbed for legibility. This approach required us to adapt our taxonomy of criteria as we learned about new features and attributes we didn't think about initially. Since our research spanned nine months, vendors have likely added or changed features since we documented their capabilities. So please validate this research with your own.

Scoring

To create the quadrant chart, we scored the product worksheets on two dimensions: company scale and functionality. For company scale, we used normalized counts. That is, for each section in the company scale section, we normalized the counts for each item.

We judged functionality subjectively on a 4-point scale with 4= Excellent, 3 = Good, 2 = Average, and 1= Poor or non-existent. To generate scores, we compared written descriptions of vendor functionality in each of the 19 categories in the worksheet. In some cases, we reviewed vendor videos to validate functionality before entering a score. We sent the worksheets (but not the scores) to vendors to validate our inputs.

Market Trends

MARKET TRENDS

Vendors

The vendors profiled in this report are startups. Most were founded just a few years ago. All have under $10 million in annual recurring revenue. Some of these are closed-source like Pinecone, others, like Zilliz are a closed-source but fully-managed commercial solution built atop Milvus while others are source-available. Deployment models vary from self-hosted (on-premise) to managed/cloud-native with a client-server architecture. One of the vendors, Chroma, is written in Python and uses an OLAP engine called Clickhouse written in C++, which provides a differentiation with embedded mode. Others are developed in languages like Rust and Golang.

Use Cases

Vector databases have revolutionized the way the mathematical representation of data is stored and accessed at scale. It is being leveraged across a range of use cases from semantic and multi-modal (image and video) retrieval and searches, Q&A, and automated categorization, fraud detection, anomaly detection, data deduplication across medical, media, gaming, entertainment, retail and legal domains.

Pricing

Each of the vendors had the following 4 pricing models.

a. Basic starting, community model, enterprise model

b. Cloud-based deployment with FinOps

c. Vector Search Costs

d. Costs to Scale Elastically

Chroma Cloud

Consumption-based pricing that scales on data, network, and computation.

Pinecone

a. Starter (Free); Standard- Storage ($0.33/GB/month), Read Units ($8.25/1M RUs), Write Units ($2.00/1M WUs); Enterprise- Starting at $0.144/hr

b. Pinecone serverless

c. $70/month

d. Write cost scales linearly by the number of writes within an index, query scales sub-linearly compared to the total size of records within an index or namespace

Qdrant

a. Community-Free ($0), Enterprise- Price on Request

b. Qdrant Cloud- Billed hourly ($25/hr/node)

c. From $8.54/month on AWS us-east (1 vector search on 1 dimension, 1 GB of RAM)

d. Vertical scaling (from 2GB to 64GB multiplied with 1GB price), Horizontal scaling (adding more nodes/sharding)

Weaviate

a. 14 days free trial, Enterprise Cloud / BYOC (handled by sales team)

b. Weaviate Cloud, Serverless ($25/mo)

c. $25/mo

d. Multiplied by $0.29 per 5,000 data objects > 60K

Architecture

Native vector database vendors follow the client-server architecture as it is the most battle-hardened model adopted by other databases. Chroma supports an embedded/ serverless hosting model. Storage, Indexing, Separation of Storage and Compute, Scalability, Sharding, Replication, and Multi-tenancy, Consistency are the main architectural dimensions that native vector databases try to optimize. Ziliz/Milvus, Pinecone (serverless) support separation of storage and compute & Separation of query and insertions. Only Ziliz and Weaviate support tunable Consistency, while all the vendors support a multi-replication model and partitioning with namespaces & logical groups. Ziliz supports dynamic sharding.

Integration

Integration has to be considered across different dimensions. Integration with existing data and application ecosystem for the organization. Integration with ever-evolving AI systems, tools, libraries, and frameworks. Using a native vector database will result in fragmented systems with the use of multiple specialty systems. It could lead to duplicated data, data

movement, additional infrastructure, deployment, additional licensing costs, limited query language power, and poor enterprise features as compared to incumbent databases, data lakehouse,s and data warehouses with vector capabilities. While Weaviate and Pinecone have strong integration with other tools in the modern data stack, others are catching up. However, almost all of them are very competitive in integrations with the AI stack.

Indexing

Vector indexing is about organizing the vector embeddings to optimize the retrieval. It leverages different classes of algorithms that arrange high-dimensional vectors so that similar vectors are grouped together for blazing fast access. There exist a plethora of different indexing techniques like Inverted File (IVF) and its variants to Hierarchical Navigable Small World (HNSW) , a sophisticated approach that uses graph-like structure as the primary ones.

Index building mainly requires vector and matrix operations and hence is computation- and memory-intensive. Picking a vector index is an open-ended question - the key being the right index depends on application requirements. Remember vector search performs approximate nearest neighbor search, meaning there is a loss of recall once the index size hits a certain threshold.

Performance & Scalability

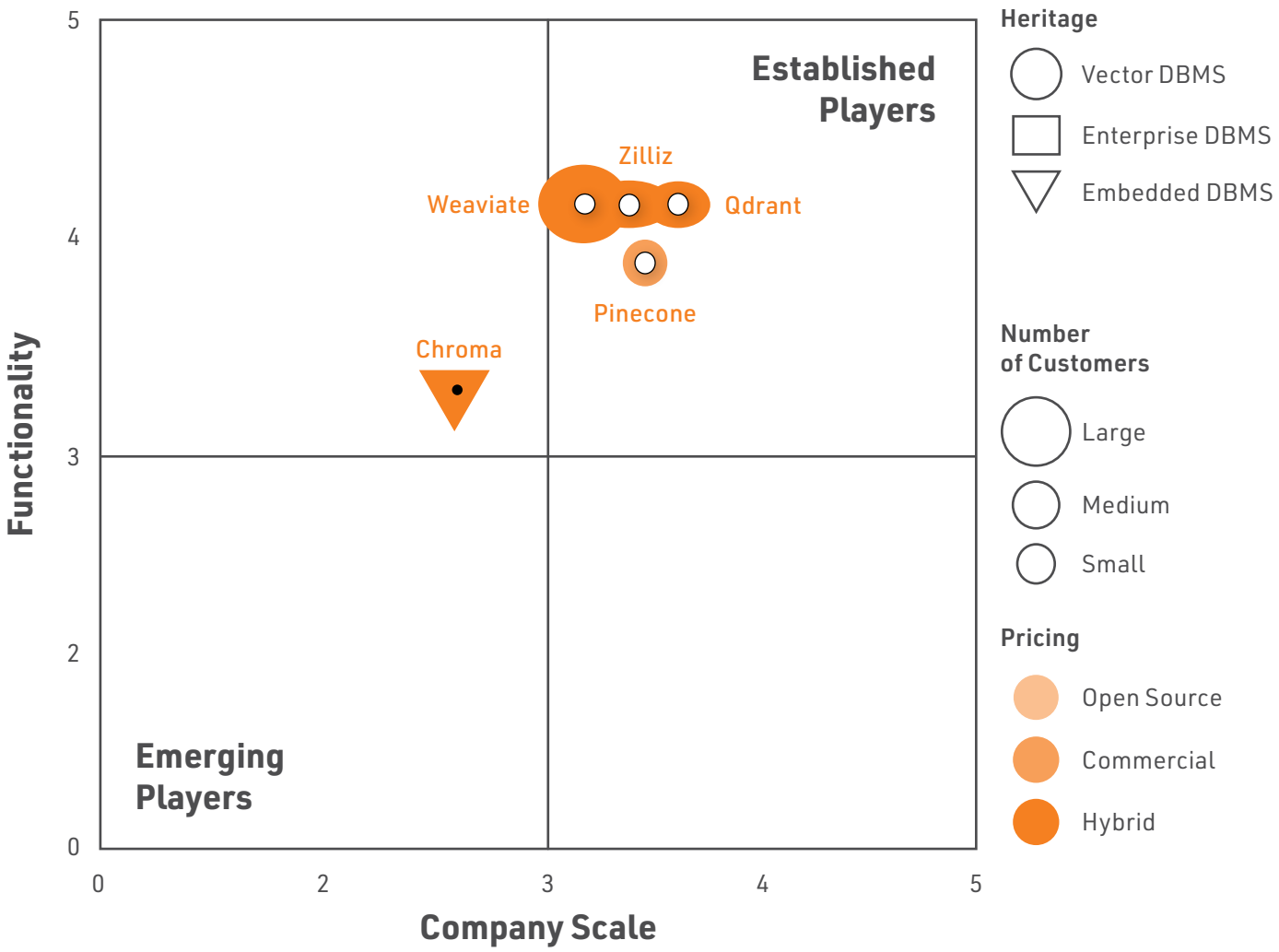
The performance of vector databases centers around Throughput, Latency, Relevancy, and Scalability. The performance of vector databases can be compared only if they achieve the same precision. When evaluating benchmark results, understand the precision threshold. Native vector databases benefit from SOTA and are optimized and focused on vector operations and are often developed in modern programming languages like Go and Rust For many customers choice of vector databases boils down to price performance. Non-native databases offer high scale but performance can vary widely, depending on functionality, indexing, and hardware acceleration. These do not necessarily store or index optimally for vector operations and for million-scale vector search performance could suffer.

GenAI

As the RAG (Retrieval Augment Generation) pattern of working with proprietary data to enhance LLM responses to use queries gains traction, vector models and vector databases will become pervasive and imperative to GenAI solutions. Vector databases are continually evolving and innovating to build better indexes, support CRUD features,

Vendor Profiles

VENDOR COMPARISON CHART



PINECONE

Focus

Founded in 2019, Pinecone's vision is to help customers build enterprise-grade AI with a developer-first experience. Recognizing the gap in traditional databases for efficiently managing high dimensional vectors, Pinecone decided to create a scalable, managed vector database service to revolutionize the handling of vector data for ML/AI applications.

Pinecone offers a cloud-native vector database that simplifies infrastructure management, allowing users to focus on developing AI solutions without worrying about maintenance or monitoring. Marquee customers include Microsoft, Notion, Blackstone, CVS Health, Sixfold, Plaid, Klarna, Gong, Accenture, Disco, Hubspot, Shopify, ClickUp, GoDaddy, Zapier, CourseHero, ZoomInfo, Frontier Medicines, Route, expel, BambooHR, Klue, Chipper across multiple verticals.

A typical customer leverages Pinecone to solve accuracy issues with GenAI responses and searches at scale. Businesses that manage and query large volumes of high-dimensional vector data essential for AI applications turn to Pinecone. Primary use cases center around AI assistant, multi-modal AI search and classification.

Differentiators

Pinecone delivers high-performance vector search with fast, relevant, and fresh results delivered at scale at minimal cost. It offers a free developer experience with resource provisioning, and it supports scalability with pod-based and serverless indexes. Pinecone remains highly performant during concurrent updates/queries and inserts. Vendors in the vector database space often compare their performance benchmarks to Pinecone. It also supports fast concurrent writes without significant performance bottlenecks, role-based access control, and end-to-end encryption.

Challenges

Pinecone is not open source and does not participate in ANN Benchmarks. Its IMI index is a proprietary index type and creates storage overheads and computationally intensive. Pinecone lacks row-level security, database backups, bulk operations, and complete ACID compliance. It limits metadata to 40KB per vector, necessitating additional queries to the main datasource for metadata retrieval.

Pinecone utilizes APIs for sending and receiving data from its indexes. This approach can be a bottleneck in data-intensive applications as it runs into synchronization issues with the primary datasource and can cause indexes to be unsynchronized from the source after periods of high data-intensity workload.

QDRANT

Focus

Founded in 2021, Qdrants's vision & focus is in developing a high-performance, open-source vector database for AI applications. Its vision stems from making advanced AI and machine learning technologies accessible and efficient for developers and businesses. It was designed from ground up to provide extensive filtering support making it a great product for semantic-based matching and faceted search applications.

Qdrant is well suited for tasks that involve managing and searching through large volumes of high-dimensional data, like vectors. Use cases include semantic search, recommendation systems,, personalization, fraud detection, natural language processing, advertising, e-commerce, medical diagnostics, media and games, document search and more. Marquee customers include Bosch, CBInsights, Discord, J&J, Dailymotion, Disney and many more.

Qdrant is Apache 2.0 license open source and supports capabilities to be deployed locally as a container. Qdrant Cloud is the fully managed option available on AWS, Google Cloud, and Azure, while Qdrant hybrid cloud version runs in a Kubernetes cluster from any cloud provider, on-premises infrastructure, or edge locations.. Qdrant also supports a private cloud experience with maximum control and security.

Differentiators

Unlike other solutions, Qdrant does not have a limitation on the number of vector dimensions. Qdrant also puts no limitation on metadata size. Qdrant supports immediate search for new vectors with background index updates, and the most differentiating features for dissimilarity and diversity search are on the roadmap. Qdrant also has a query planner that optimizes the query execution based on indexes, complexity, and cardinality parameters.

Challenges

As a newer tool, Qdrant is playing catch-up with features and capabilities though the gap is rapidly reducing with each release. Qdrant only supports a customized HNSW indexing algorithm. Qdrant also does not support strong transaction guarantees, which allows it to perform operations with low overhead. Qdrant does not guarantee atomic distributed updates and incurs increased latency to see the results of writes. Qdrant does not automatically rebalance shards if your cluster size or replication factor changes. Qdrant cannot scale down cloud clusters.

WEAVIATE

Focus

Founded in 2019, Weaviate's vision is rooted in deriving insights from data using AI and ML. Weaviate is dedicated to creating open-source, scalable, AI-first infrastructure, for high quality AI and semantic applications. The goal is to make it easier for developers and enterprises to leverage AI and ML models for semantic search and data understanding. Marquee customers include - Instabase, NetApp, Stack Overflow, Writesonic, Red Bull, Shippo.

The ideal prospect for Weaviate, are organizations whose pain points are related to managing, searching, and deriving insights from large volumes of unstructured data. Weaviate has raised $70 million with over 1000 commercial customers, and more than 1M open source downloads a month. Use cases include – recommendations, personalization and RAG-based applications.

Differentiators

Some of the key differentiators of Weaviate are:

- Developer-friendly AI-native platform that simplifies development and maintenance of AI applications.
- API-driven and easy to integrate with the ML ecosystem.
- Easy embedding creation, integration of ML models, hybrid search.
- Built for scale and cost optimization with capability to store vectors at multi-billion scale with millisecond latency.
- Native multi tenancy, tenant offloading, filtering, and vector compression to maximize resource efficiency.
- It is cloud, model, and deployment agnostic, and can self-host on-prem.
- Extensive integration with the LLM ecosystem and built-in integration for embedding, generative, and re-ranker ML models.

Challenges

Weaviate has been developed in Golang, and it requires K8S for scalability which means it requires infrastructure for large data sets. Cost implications for Weaviate's fully managed offering over the long term are unknown. With static sharding as data outgrows server capacity, a time-consuming and complex process of re-sharding is required. Weaviate implements proprietary HSNW indexing.

Being pure vector databases, it comes with the challenges of competing with other enterprise scale vector enabled databases that can store a variety of data types and has limited SQL support. Weaviate? Is still working on enterprise features such as CRUD, ACID, DR, RBAC..

CHROMA

Focus

Founded in 2021, Chroma is a late-comer in the already crowded vector database market. Its focus is to make it easy for developers to quickly get a vector store up and running. It's often used to build search and recommendation applications in digital media, e-commerce, and content discovery.

Differentiators

Chroma's aim is to make AI systems robust, safe, and accessible to machine learning newcomers. It seeks to balance open-source accessibility with the viability of commercial offerings, hoping to democratize AI tools and capabilities. Chroma addresses the pains related to managing large volumes of data for AI applications, particularly those involving embeddings for machine learning and NLP tasks.

Chroma is designed specifically for handling high-dimensional color vectors. Efficient indexing and retrieval of color data, supports complex color-based query operations.

Chroma is a Python/TypeScript wrapper on top of Clickhouse with an open-source vector index HNSW. Most vector database are self-hosted (on-premises) or a managed cloud service, or both However, Chroma is also available in an embedded mode, where it can be tightly coupled with the application as a serverless deployment.

Challenges

Chroma only supports the HNSW indexing approach and does not support its own storage layer. Chroma does not yet participate in ann-benchmarks and has not been benchmarked independently.

ZILLIZ

Focus

Zilliz is the commercial parent entity of open-source Milvus founded in 2019. Its goal is to develop a high-performance, open-source vector database for AI applications. Zilliz is designed to empower developers and data scientists in building the next generation of AI and search applications. It offers a robust platform for scalable, efficient, and accurate vector search and analytics, supporting a wide array of AI-driven applications.

Zilliz aims to empower applications with AI capabilities by making advanced data analytics and search accessible and efficient. Zilliz's vision focuses on providing efficient data processing technologies to enable real-time analytics and similarity searches. Ideal prospects for Zilliz are customers that want to handle vector data efficiently, address scalability challenges, support real-time data processing, and manage and search large-scale high dimensional vector data..

Differentiators

Zilliz achieves high scalability through a combination of proxies, load balancers, message brokers, Kafka and Kubernetes. Zilliz supports a large number of indexing approaches from Flat, Annoy, IVF, HNSW and DiskANN (only one amongst the current players)

Other differentiators include:

- It supports a variety of vector data types and search algorithms.
- Ability to scale in a cost-effective manner to a billion vectors, by supporting bigger-than-memory datasets through indexes optimized for SSD storage and hardware acceleration.
- It's one of a few vector databases to support four levels of consistency.
- It separates computing and storage and different computing tasks like querying, data wrangling, and indexing.
- It supports temporality with time travel capabilities and working with Parquet files.

- It cloud complies with SOC 2 Type II and ISO/IEC 27001 certifications and supports granular RBAC.

Challenges

Zilliz uses Go, a garbage collected language and Kubernetes for scalability. This requires a good amount of infrastructure skills and resources when data gets large. This makes the databases complex and resource intensive. Its APIs aren't as developer friendly as other products and it still lacks geo-spatial support. Zilliz can be an overkill for smaller to medium sized datasets.

Conclusion

and improve performance and operational efficiency.

CONCLUSION

If your organization or business unit is looking for a native vector database, consider the ones profiled in this report. These represent the pure play vector databases that focus on scalability, low latency, high throughput, speed and options of deployment, infrastructure management and continuously innovating set of capabilities. Although they may not have all the bells and whistles of incumbent databases incorporating vector capabilities, they are quickly catching up and innovating at a faster pace. Incumbent players have a lesser footprint and infrastructural requirements but can fall short in terms of latency and high throughput with concurrency as they were built for a different set of use cases and data types.

ABOUT ECKERSON GROUP

Eckerson Group is a boutique consultancy that helps organizations get more value from data. We specialize in developing actionable data strategies and tailored data solutions for commercial and non-profit organizations. Our consultants each have 25+ years of hands-on experience and specialize in data architecture, data governance, data literacy, self-service, data products, and operating models. Our clients say we are hard-working, insightful, and humble. It stems from our love of data and desire to help organizations turn data into insights and action. **Explore what Eckerson Group can do for you!**

Eckerson Group