S&P Global
Market Intelligence

Technology & Business Insight

Tech Trend in Focus

October 2023

Tech Trend in Focus: Vectors and Databases

A Match Made in Generative AI Heaven

James Curtis, **Senior Research Analyst, Data, AI & Analytics**

Gregory Zwakman, **Research Director, Market and Competitive Intelligence**

Vector-enabled databases are an emerging class of databases that are designed to store and process vectors. Vectors (or vector embeddings) are numerical representations of data — primarily for unstructured data — that are stored as a data type and searchable within the database. Vector-specific database vendors are building databases that are optimized specifically for and around the vector data model, whereas existing database vendors are adding vector data support, which is one of many data models that are supported within that database. Why the rush to embrace vectors? The answer lies in the fact that when vectors are stored in a database, it opens up a host of use-case scenarios for generative AI, from augmenting large language models (LLMs) to providing semantic search.

This report discusses vectors and specifically vector databases and the needs they fill, as well as the drivers, market direction, financial forecast and trends driving vector adoption. Several new and existing database vendors will be explored, as will the expected impact vector databases will have within the market and with enterprise clients.

To learn more or to request a demo, visit spglobal.com/marketintelligence.

Table of contents

Definition

The interest in vector-enabled databases is extraordinary, although not entirely surprising, given the broader interest in generative AI, and specifically LLMs. Vector-specific databases or existing databases with added vector search support can store, index and process high-dimensional vectors — numerical representations of unstructured data such as text, images, audio and video. To get the data in its correct form, however, requires a process called vectorization, which uses a specific type of LLM called an embedding model that converts the data into vectors. The resulting output — referred to as vector embeddings — consists of potentially tens to thousands of dimensions based on the complexity of the data object.

Because vector data is represented numerically, it is the ideal structure to be processed efficiently on modern CPUs and GPUs. After the vectorization processes, these vector embeddings are stored and indexed in the database. Indexing plays a crucial role in vector databases because it correlates heavily to search performance and accuracy.

Querying a vector database is functionally differently than traditional databases. Whereas traditional databases search on an exact match, vector-enabled databases search on similarity — that is, looking for vectors that are “nearest to” other vectors by leveraging approximate nearest neighbor search algorithms. Similarity search is quite powerful because it provides semantic searching capabilities, such as searching for the meaning within a document or text. Similarity search can also be used for image and audio searching by finding the items that are most similar, which could drive, for instance, recommendation applications.

The Take

The rise of vector-enabled databases — whether that be specific database systems optimized for vector processing or existing database systems with added vector search capabilities — is sudden and remarkable. Enterprise adoption, however, will depend largely on generative AI applications and intended purpose, as vector-enabled databases are used to augment LLMs to provide context and act as long-term memory (examples of two commonly implemented use cases). But vector-enabled databases are also expected to push the boundaries of search capabilities with their ability to perform fast and highly accurate semantic search, which is expected to drive a host of applications, from product recommendation systems to personalized search (retail, e-commerce, etc.) to fraud detection to reverse image search. While interest remains high for generative AI and, subsequently, for vector-enabled databases, the forecasted market for vector-enabled databases is still a nascent market, estimated at $269 million. The trend however is expected to creep up marginally until 2025, where much more market growth is forecasted. By 2028, we project revenue to be slightly north of $1.7 billion, with a compound annual growth rate (CAGR) of 45% from 2023 to 2028. With these forecasts, the underlining expectation is that vector-enabled databases will be the largest database adoption driver that we have witnessed in more than a decade.

The need

While enterprises are actively looking to adopt AI and build generative AI applications, enterprises are also looking deeper at their existing data and processing platforms as enablers for these applications. Generative AI applications require not only significant amounts of data, specifically unstructured data, but also need to extract meaning from the data. Current data platform systems, while capable of handling unstructured data, may require significant manual "feature engineering" to handle certain data types and be able to drive AI applications. This approach is not sustainable in the long run, thus a different approach is needed.

Vector-enabled databases fill that need and systems are optimized for managing, storing and processing vectors, or more specifically vector embeddings. It is due to precisely this concept that vector-enabled databases are able to address at least two critical needs for enterprises.

First, the need to get unstructured data into a usable format for processing and enable access to it. Vectors are ideally suited to handle high-dimensional data for which enterprises generate significant quantities. High-dimensional data — also known as unstructured data — consists of text documents, images, audio and video files, among others. While traditional databases have improved in their ability to store and process unstructured data, most database systems are more suited for structured data that is often expressed as a single value, such as integers and text strings that are stored as tables or documents.

The second is the need to glean meaning from the data, to be able to search and query the data for business-critical decision-making. In traditional database systems, searching is based on performing an exact match against the data because the system is searching on one-dimensional data. Conversely, vectors consist of multidimensional data, and therein lies its value. The multidimensional capability of a vector means that it can provide meaning in the data, or semantics. When a semantic search is done on a vector-enabled database, it is not looking for an exact match, but rather is executing a similarity search based on a given query prompt. The search could be to find similar products, or it could be based on finding text that is contextually similar.

Drivers

November 30, 2022, will go down as a technology watershed moment because it marked the debut of ChatGPT, an LLM developed by OpenAI that is capable of generating a response based on a user prompt. Not only did ChatGPT elicit great excitement, like from developers as a tool to augment coding, but it also stirred great concern from other groups, for instance, at educational institutions where students began using ChatGPT to automatically generate school essays. While the technology may be used for nefarious purposes, what ChatGPT proved was that there were real, tangible benefits for generative AI, specifically within the question-and-answer framework. The possibilities seemed endless and piqued the interest of enterprises for generative AI.

LLM usage adoption is a leading driver for the adoption of vector-enabled databases as they have proven to mitigate many of the shortcomings users experience when using LLMs.

- **Provide knowledge (context)**. Known as retrieval augmented generation (RAG), this approach uses a vector-enabled database as a type of external knowledgebase and provides similar information and context that the LLM is not trained on. For instance, RAG enables organizations to "augment" their own propriety data with that of the LLM to ensure an accurately generated response that is more relevant to the organization.
- **Provide memory**. This approach maintains a "memory" of a chat history within a vector-enabled database that can then be used to provide relevant information specific to the existing session. In essence, the history provides context to the current session and helps to address topic changes that may have occurred since the last session that can be utilized when generating responses. Providing memory also has the added benefit of getting around the problem of token limitation by using only the most relevant content, because some LLMs have fixed-sized input limits.
- **Acting as a cache**. This approach leverages a vector-enabled database that functions as a cache for previously run queries and generated responses. If there is a new query that is semantically similar to a previous query and it has been cached, then the response can be retrieved. For enterprises, the benefit is a potentially faster response time with reduced cost because there is no need to query the LLM.

All querying of vector-enabled databases employs semantic search to some extent, and this includes scenarios that augment LLMs. However, another significant driver of vector-enabled database adoption, which does not involve augmenting LLMs, centers on purely sematic search for a variety of generative AI applications. For instance, semantic search applications can be used for product recommendations, fraud detection and reverse image search.

Trends

Vector databases continue to evolve in a number of areas.

- **Pure-play vector database versus incumbent database with vector-added support**: As enterprises weigh their need for a vector-enabled database, it will become apparent early on whether that means deploying a pure-play vector database or going with an existing database with vector-added support. It is not a new debate, especially in the database space, where graph and time-series databases are just two examples of functionality offered as stand-alone databases or included as part of an existing database. Historically, however, specialty databases are favored when it comes to driving the largest, most demanding workloads. But the tradeoff should not be dismissed; it requires adding yet another database system to an already complex database environment, which ultimately takes resources, time and additional expenses to manage.

- **Technical differentiators**: Vector-specific databases or databases with vector management support added can vary widely. Compared with other data models such as the relational data model, vectors require upfront processing to convert unstructured data to a vector, also known as vectorization, to produce vector embeddings. Vectorization, however, requires vector embedding models that vary in complexity and capability. Database support for embedding models (including API support for access) may likewise differ among vendors. When and why to use one embedding model over another can have a significant impact that is often dependent upon each use case. While vectorization support may be one of many technical differentiators, there are others, such as indexing, scalability, partitioning, concurrency and deployment model.

- **Managed and self-managed in the cloud or on-premises**: Cloud is quickly becoming the favored deployment model for running database workloads. It is not to say that on-premises is completely dead. It is not; it has its place under specific circumstances. The cadre of pure-play vector database vendors are gravitating toward the cloud, offering either fully managed services or self-hosted offerings, where enterprises deploy the vector database to their VPN. Of the emerging group of vector database vendors, none are delivering an on-premises offering. Some vector-specific databases are available as open source and enterprises could certainly deploy these offerings on-premises if they chose. Incumbent database vendors that are adding vector support may offer on-premises options, particularly if it is part of their existing deployment suite. But the trend appears to be that vector-enabled databases favor a cloud model over on-premises.

Market

The vector-enabled database market is multifaceted in that there are pure-play vector database vendors, incumbent database vendors that have added vector support to their existing databases, and a collection of libraries, engines and APIs (much of which is open source) that augment vector search capabilities. Below are an initial list of players.

- **Amazon Web Services Inc.**: Vector Engine for Amazon OpenSearch, a new cloud service, enables vector search without requiring management of a vector database infrastructure. Amazon Aurora PostgreSQL and Amazon RDS for PostgreSQL both support the pgvector (a PostgreSQL extension) that enables storing of vector embeddings and performing similarity searches.
- **Activeloop**: Deep Lake is billed as a "data lake for AI." Deep Lake stores a litany of data types, including vectors which enables similarity searching.
- **Alibaba Group Holding Ltd.**: AnalyticDB for PostgreSQL that includes a vector engine. Hologres is an interactive data warehouse service that incorporates Proxima, a library extension, to enable vector support.
- **Chroma Inc.**: The eponymous offering is referred to as a vector "embeddings database" that is currently only available as open source, with a hosted version on the horizon. Chroma was developed to directly ease the AI application developer experience.
- **ClickHouse Inc.**: The ClickHouse database is open source and primarily geared toward real-time analytic workloads, but has been updated to enable vector search. The company provides a hosted and server-side database offering with support.
- **Couchbase Inc.**: Couchbase Server is a multi-model NoSQL database, and the company has added the vector data model to its managed Capella database service to enable developer application productivity, which the company calls Capella IQ.
- **Databricks Inc.**: Known for its Lakehouse Platform, Databricks has added vector search to its Lakehouse AI capabilities. Vector search is closely integrated with the company's Unity Catalog offering, which automatically converts data and queries to vectors.
- **DataStax Inc.**: Known for its commercial offering atop Apache Cassandra, the company has added the ability to store and process vectors (or vector embeddings) inside Astra DB, the company's database-as-a-service offering, which also integrates with the company's data streaming offering.
- **Dremio**: Billing itself as delivering an open-source data lakehouse, the company is jumping into the generative AI space with a forthcoming offering called Vector Lakehouse Capabilities that enables storing and searching vectors.
- **Elasticsearch BV**: The company added vector search to its distributed, RESTful, JSON-based search engine that is built atop Apache Lucene.
- **Google Cloud**: Google LLC's Vertex AI Search enables vector search capabilities that can drive generative AI applications based on a corpus of data. AlloyDB for PostgreSQL and Cloud SQL for PostgreSQL both support the pgvector (a PostgreSQL extension) that enables storage for vector embeddings and similarity searches.
- **Kinetica DB Inc.**: Analytics database vendor Kinetica promotes tight integration with LLMs (ChatGPT, for instance) where it can perform "vectorization queries" in parallel because the data is stored in vectors and is processed with an architecture that incorporates GPUs and CPUs.

- **KX**: KDB.AI is the company's vector database offering, which it promotes as a hybrid database that integrates time-series and real-time data. ok

- **LanceDB (Eto Labs Inc.)**: Its database is an open-source vector database. A cloud service is under development and currently available by request only.

- **Marqo**: Based in Australia, the company offers a cloud-based vector database service with an eye on a providing a full-service offering in which customers will not be required to bring third-party tooling (i.e., foundation models), as the these are available from a single API.

- **Metal Technologies Inc.:** The company positions itself as providing vector embeddings as a service, with a focus primarily on the developer experience, not needing to build out infrastructure, storage or tooling to manage the vector embeddings. Free, developer and business tiers are offered.

- **Microsoft Corp.:** Azure SQL database can store and search on vectors. Vector search (public preview) was recently added to Azure Cognitive Search as well as to Azure Cosmos DB for MongoDB vCore. Azure Data Explorer, the company's analytics service, is able to store and search on vectors because of its dynamic data type.

- **MongoDB Inc.**: Known for its document-model NoSQL database catering to developers, the company has added vector search to its managed cloud database service, Atlas.

- **Neo4j Inc.**: Graph NoSQL database vendor, Neo4j, announced the inclusion of vector search to its core database as an enabler for generative AI applications and integration with LLMs.

- **Oracle Corp**.: The company recently announced vector search support as part of the Oracle 23c release. Oracle's strategy is that vector search should and can be augmented with additional business data (non-vectorized data) to enhance and drive use cases that involve generative AI.

- **Pinecone Systems Inc.**: A pure-play, proprietary vector database vendor with an optimized vector data model. Offered exclusively as a managed cloud service on all three cloud hyperscalers.

- **Qdrant**: Pronounced as "quadrant," this vendor offers its vector database, open-sourced under an Apache 2.0 license, but the company also offers a managed cloud and enterprise (managed on-premises) option.

- **Redis Ltd**.: Known as an in-memory NoSQL key-value database that is often used as a caching layer to other systems, Redis can also be used as a primary datastore and has expanded to accommodate multiple other data models, including vectors, for vector search.

- **Rockset**: Categorized as a real-time analytics database that can land and query data near immediately, the company has added the ability to store vector embeddings as well as enable vector search.

- **SingleStore Inc.**: The company offers a hybrid database capable of both transactional and analytics workloads that centers on the company's Universal Storage model, which makes vectors and other data types all queryable by SQL.

- **Snowflake Inc.**: The company's "data cloud" strategy is a data clearing house for all types of data, including vectors, to enable a variety of different workloads — not just analytic or operational workloads, but also vector search.

- **Supabase Inc.**: Positioned as an alternative to Firebase, and as such, one of Supabase's services is a vector database environment built atop PostgreSQL, which also leverages the pgvector PostgreSQL extension to enable storage for vectors and vector search.
- **Tencent Holdings Ltd.**: Chinese technology and media company Tencent announced a forthcoming cloud-based vector database.
- **TileDB Inc.**: The company's database is array-based with a claim that it can take any data — structured or unstructured — and convert it to high-dimensional arrays for searching. Vectors are one-dimensional arrays, so a transition to support vectors and vector search was an easy move for the company.
- **Timescale Inc.**: The company's time-series database is built atop PostgreSQL and integrates the pgvector PostgreSQL extension for storing and enabling vector search.
- **Weaviate B.V.**: A pure-play vendor with a vector database that is available as open source, managed cloud and hybrid cloud, where Weaviate manages the database via its data plane from a customer virtual private cloud.
- **YugabyteDB:** Known for its globally distributed SQL database that is built on PostgreSQL, the company is leveraging the pgvector PostgreSQL extension to enable vector search.
- **Zilliz**: A pure-play vendor with a vector database that is built atop the open-source vector database Milvus, of which Zilliz is the original developer. The company provides a managed cloud service for Milvus.

Competitive differentiation in this space is driven by several factors, including:

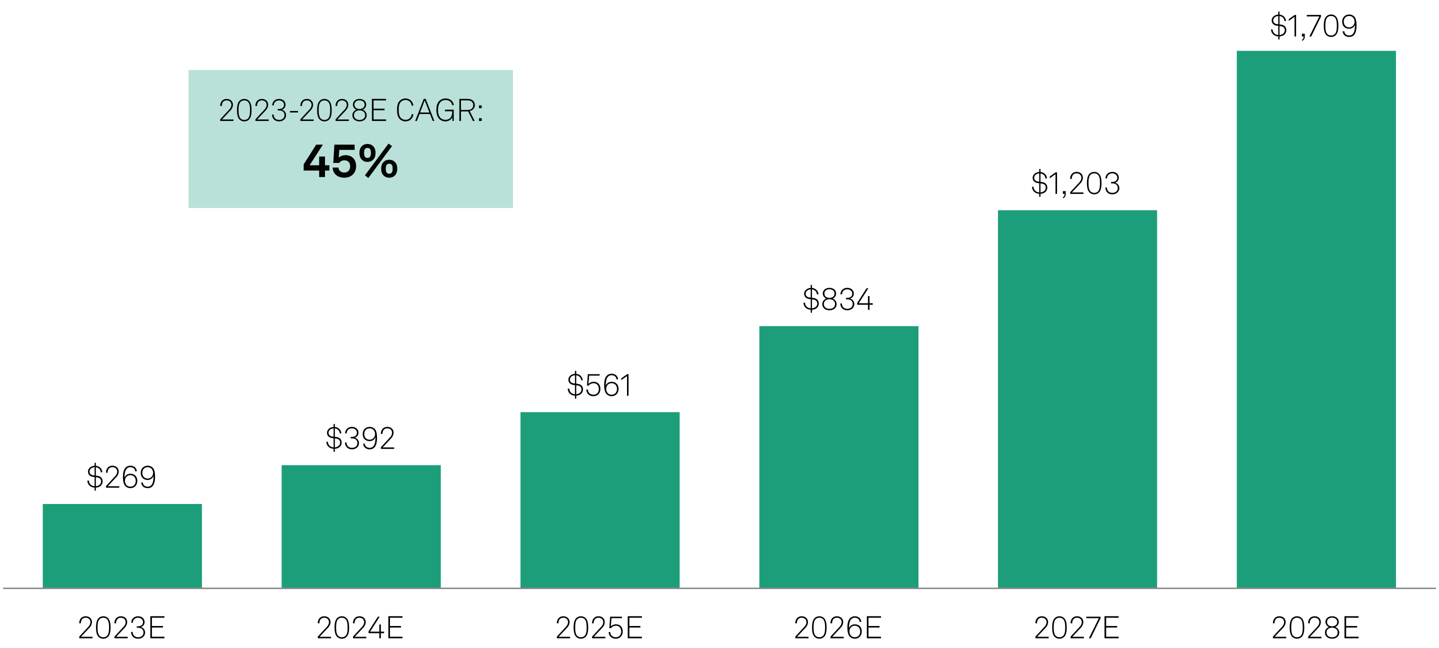
- Database scalability
- Concurrency
- Indexing type
- Embedding models supported
- Search operations for vector distance
- Query language support
- Database partitioning
- Deployment location
- Pricing
- Support services
- Security certifications3

Market size and forecast

The estimates and forecasts for the vector-enabled database market were generated by a mix of proprietary market intelligence, qualitative analyst input and custom statistical analyses, with end user survey results supporting our findings.

The S&P Global Market Intelligence team generated a revenue estimate for each vendor participating in the vector database market, a bottom-up methodology. The majority of vendors included in the analysis are privately held and typically do not disclose revenue, and the revenue streams we track at public companies do not showcase vector-specific revenue. Therefore, we rely primarily on proprietary market intelligence to create revenue models for each vendor in our market sizing. The analysts who make up our segment and the quantitative teams who drive the forecasts regularly hold briefings with vendors, end users and other market participants. It is the intelligence gleaned from these interactions that generates our current estimates and form the basis for our forecasts. Our team of quantitative analysts, using a modeling discipline rooted in equity research, utilizes this data to create proprietary revenue models for each vendor.

Figure 1: Estimated global vector database market annual revenue, 2023-2028 ($M)



Source: S&P Global Market Intelligence estimates, 2023.

With several vector-specific database vendors releasing products this year, as well as a of host of incumbent database vendors just starting to release vector search support, our estimate for the current size of the market points to a nascent segment, currently forecasted at $269 million for 2023 (see Figure 1). We expect another modest year in 2024 before trending significantly upward starting in 2025, where much more substantial growth is expected and where CAGR is forecasted at nearly 45% through to 2028.

Implications

- **Uncovering deeper use case applications**. The adoption of LLMs is already having an impact on vector-enabled databases that provide context and long-term memory. But vector-enabled databases are also used to improve search, particularly semantic search. Many of the same areas will be impacted — e-commerce, biosciences, financial services, healthcare, technology and so forth — but the depth can be much greater. In healthcare, for instance, patient records, symptom detection and medical history could all be enhanced with vector search applications.
- **Weighing performance with accuracy or vice versa**. Performance and accuracy will begin to take on greater meaning as enterprises begin to adopt vector-enabled database systems. But favoring one over the other is a matter of preference and need, driven largely by enterprise use application. While vendors generally promote high-performing, accurate results, the reality often comes down to which search metrics are used and how the vectors are indexed.
- **Knowing LLM characteristics**. Not all LLMs are created equal, which should not come as a surprise to enterprises. LLMs vary on multiple levels, such as number of parameters, performance, training type data, age of the data and so on — and this is for good reason. When the use case is for a vector database system is to augment an LLM for context or long-term memory, the linked LLM characteristics need to come into play to match the intended scenario.
- **Considering database system differentiation**. One database to rule them all sounds good in theory, but is not in reality. This applies to vector-enabled database as well. With these systems, there are numerous variables that come into play depending on the intended purpose. For instance, does the application cater to an image-centric use case or a text-centric application? Embedding models are designed to produce a certain type of vector embedding with specific characteristics, such as for image search or text. Distance metrics and associated LLMs are also factors to consider that could drive the requirements for a purpose-driven vector-enabled database system. All of these are likely to weigh on enterprises. It is also possible that enterprises may need to deploy multiple vector-enabled database systems within their environment, based on the apparent strength of each.

Methodology

S&P Global Market Intelligence provides essential insight into key trends driving digital transformation across the entire technology, media and telecommunications (TMT) landscape. By offering a combination of expert analyst insight and differentiated data, our TMT Research group enables the industry with the information and perspectives they require to make more effective decisions.

Reports such as this offer a holistic perspective on key trends and themes driving the technology, media and telecommunications space over the coming year. These markets evolve quickly, so S&P Global Market Intelligence offers a wide range of research services that provide critical marketplace updates on an ongoing basis. These reports, datasets and perspectives are published frequently, in numerous short- and long-form factors, available on S&P Capital IQ Pro and 451Research.com. Forward-looking M&A analysis and perspectives on strategic acquisitions and the liquidity environment for TMT companies are also updated regularly via these platforms, backed by industry-leading databases such as the 451 Research M&A KnowledgeBase.

Our research is organized into channels that align with the prevailing key issues driving digital transformation across TMT. These channels are: Applied Infrastructure & DevOps; Cloud & Managed Services Transformation; Cloud Native; Customer Experience & Commerce; Data, AI & Analytics; Datacenter Services & Infrastructure; ESG; Fintech; Global Media; Global Mobile; Global Multichannel and Broadband; Information Security; Internet of Things; and Workforce Productivity & Collaboration.

For more information about S&P Global Market Intelligence, please go to: Vital TMT Insight | S&P Global Market Intelligence (spglobal.com).

About the authors

James Curtis

Senior Research Analyst, Data, AI & Analytics

James is a senior research analyst at S&P Global Market Intelligence leading the database and data platforms vertical within the Data, AI, and Analytics channel. Previously, he covered business intelligence, analytics and reporting, along with machine learning-driven and data science.

Before joining S&P Global Market Intelligence in 2015, James held several senior roles in technology, marketing, and communications. He served as VP of a large BPO firm where he oversaw marketing for analytic solutions. He held senior technical marketing roles at Netezza and later at IBM with responsibility for data warehousing, analytics, and big data products. He has managed global programs at HPE and worked as a case editor at Harvard Business School.

Gregory Zwakman

Research Director of Market and Competitive Intelligence

As Research Director of Market and Competitive Intelligence, Greg oversees the development and maintenance of the Market Monitor & Forecast product line. The Market Monitor provides market sizing, share and growth forecasts for emerging enterprise technology markets including IoT; data, AI and analytics; information security; cloud computing; datacenters; and hosting/service providers. Greg also serves as an analyst catering to the buy side and regularly participates in custom consulting projects focused on total and segment addressable market (TAM/SAM), ROI analysis, custom market share and size analysis, and more for a range of clients.

About S&P Global Market Intelligence

At S&P Global Market Intelligence, we understand the importance of accurate, deep and insightful information. Our team of experts delivers unrivaled insights and leading data and technology solutions, partnering with customers to expand their perspective, operate with confidence, and make decisions with conviction.

S&P Global Market Intelligence is a division of S&P Global (NYSE: SPGI). S&P Global is the world's foremost provider of credit ratings, benchmarks, analytics and workflow solutions in the global capital, commodity and automotive markets. With every one of our offerings, we help many of the world's leading organizations navigate the economic landscape so they can plan for tomorrow, today. For more information, visit www.spglobal.com/marketintelligence.

CONTACTS

Americas: +1 800 447 2273
Japan: +81 3 6262 1887
Asia Pacific: +60 4 291 3600
Europe, Middle East, Africa: +44 (0) 134 432 8300

www.spglobal.com/marketintelligence
www.spglobal.com/en/enterprise/about/contact-us.html