Event Recap: Google Cloud Next – August 2023

Bradley Shimmin

Chief Analyst, AI platforms, analytics and data management

askananalyst@omdia.com

OMDIA

Brought to you by Informa Tech

Contents

Seven top trends

01

Dancing a Duet

Google begins rolling out Duet AI across its portfolio

02

Embeddings afoot

New vector database tools target semantic search and LLMs

03

Fresh from the garden

The Google Vertex AI Model Garden picks up several new LLMs

04

Solid foundations

Google beefs up its flagship LLM with languages and context

05

Notebooks get real

Google's free Colab notebook finally reaches the enterprise

06

Big data & big models

New LLMs and new tools now native to Google BigQuery

07

LLMOps in overdrive

New measures for honesty, helpfulness, and harmlessness abound

Show overview

Show overview: “The old way is getting old.” – Sundar Pichai

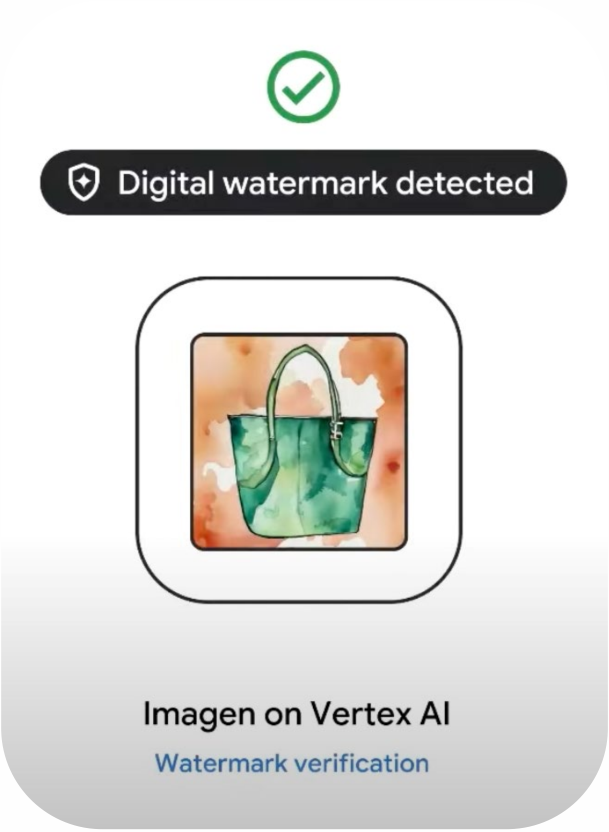
Google and Alphabet CEO Sundar Pichai on stage

Source: Google Cloud Next 2023 keynote

- To open Google Cloud Next 2023 show, CEO of Google and Alphabet, Sundar Pichai, took to the stage and straight away emphasized the fact that Google Cloud has been taking an AI-first approach to creating its products for the past seven years.
- This statement is crucial, as it takes direct aim at recent market criticism that despite inventing the concept of transformer models, the very technology that powers today’s wildly popular large language models (LLMs), the company has lost its position as a first mover in pioneering AI.
- As a proof point, Mr. Pichai noted that during Google Cloud Next 2023 the company was rolling several generative AI (GenAI) solutions that had only been announced a few months ago at Google I/O 2023, led by Duet AI – an augmentative GenAI-powered collaborator assistant that Google intends to interweave across its entire Google Cloud portfolio.
- Furthering this desire to be associated with innovation, Mr. Pichai went so far as to claim that 70% of all GenAI Unicorns (startups worth more than $1B dollars) were Google Cloud customers. That influential list includes A21 Labs and Anthropic, both of which figure heavily in Google’s emerging GenAI ecosystem.
- This focus extended offstage as well. In meeting with several Google product leaders, Omdia noted this abiding sense of acceleration with one individual happily stating that it feels currently as though Google feels once again like a start up in diving headlong into GenAI.

Show overview: “The old way is getting old.” – Sundar Pichai

Addressing copyright protections in GenAI output



Source: Google Cloud Next 2023 keynote

Throughout the keynote and across the entire event, Google emphasized the fact that it remains a key player in the AI marketplace and how its investments in AI have made the Google Cloud platform a future-forward investment for enterprises looking to successfully ride the current wave of interest in generative AI. Some of the more impactful announcements made by Google follow:

- Google called attention to the differentiating nature of its in-house tensor processing units (TPUs), further bolstered by its continuing partnership with GPU-powerhouse NVIDIA. This will be covered separately by Omdia’s Data Center and Applied Intelligence research groups.
- In terms of software (the focus of this report), Google accelerated and further solidified its generative AI capabilities, for example, rolling out its AI-assisted user experience, Google Duet AI.
- The vendor also introduced a comprehensive set of operational tools supporting the development of LLM-based solutions, focusing on responsible AI concerns such as model grounding through semantic search, data availability/quality capabilities, and even digital watermarking for images.
- Importantly, Google enhanced its catalogue of natively supported LLMs with several updates to its own models as well as the addition of Meta’s newly released Llama family of open source models.

Top stories

Dancing a duet

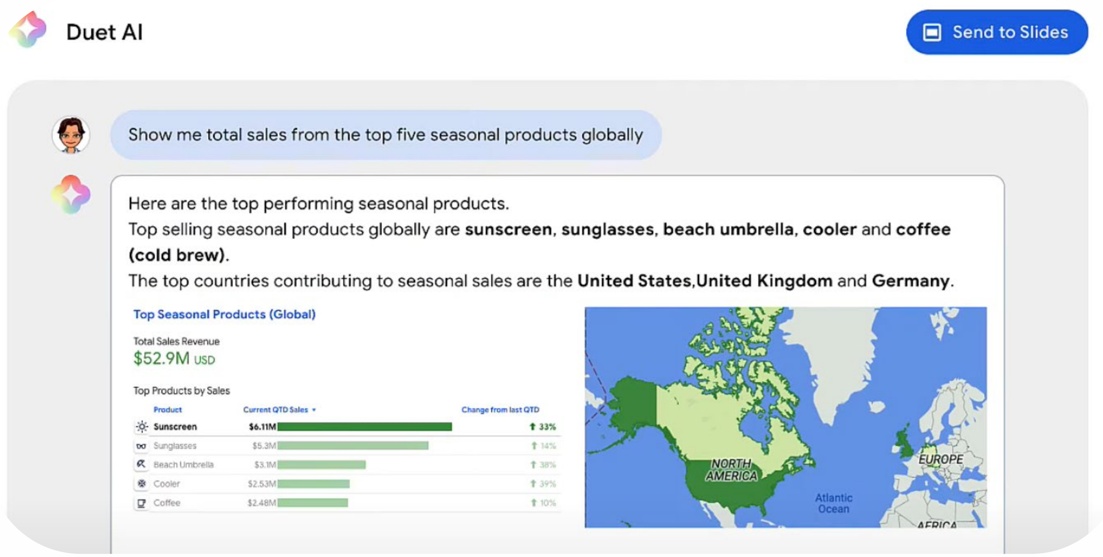
Announcement

- A few short months after introducing Google Duet AI, an always-on collaborator integrated across Google Cloud services, the company announced the general availability of Duet AI within Google Cloud Workspaces. The vendor also announced a definitive release plan and several new capabilities for Duet AI within Google Cloud itself, expanding the scope of this AI augmented copilot for developers, systems operators, data professionals, and even cybersecurity practitioners.

Assessment

- Much more than a mere GenAI collaborator, Duet AI for Google Cloud will be tightly interwoven within several tools like Google Cloud console, Cloud Workstations, and Cloud Shell Editor. For software developers, this means they'll be able to gain assistance with difficult tasks like code refactoring without having to switch context. This ability to generate code in context is key as it allows the collaborator to generate code that's specific to the project at hand, drawing from the language, libraries, data models, etc. that are in use at that time in that workspace.

Duet AI for data exploration



Show me total sales from the top five seasonal products globally

Here are the top performing seasonal products.
Top selling seasonal products globally are **sunscreen**, **sunglasses**, **beach umbrella**, **cooler** and **coffee (cold brew)**.
The top countries contributing to seasonal sales are the **United States**,**United Kingdom** and **Germany**.

Top Seasonal Products (Global)

Total Sales Revenue
$52.9M USD

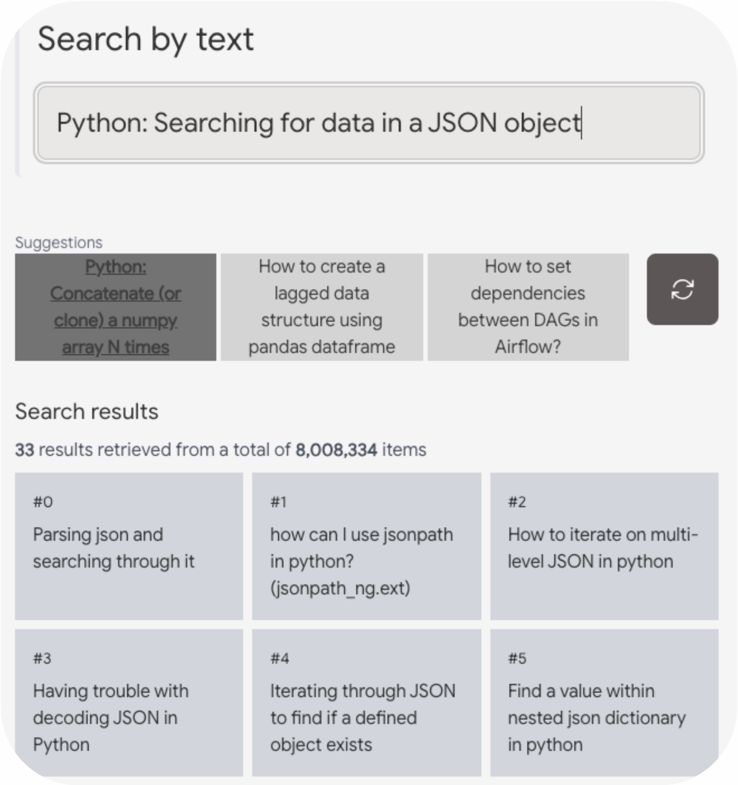
Top Products by Sales

Product	Current QTD Sales ▾	Change from last QTD
Sunscreen	**$6.11M**	**↑ 33%**
Sunglasses	$5.3M	↑ 14%
Beach Umbrella	$3.1M	↑ 38%
Cooler	$2.53M	↑ 39%
Coffee	$2.48M	↑ 10%

Notes: Using Duet AI to generate visualizations with natural language in Google Looker
Source: Google

Embeddings afoot

Google Vertex AI Embeddings API + Matching Engine demo



Search by text

Python: Searching for data in a JSON object

Suggestions

- Python: Concatenate (or clone) a numpy array N times
- How to create a lagged data structure using pandas dataframe
- How to set dependencies between DAGs in Airflow?

Search results

33 results retrieved from a total of **8,008,334** items

#0 Parsing json and searching through it

#1 how can I use jsonpath in python? (jsonpath_ng.ext)

#2 How to iterate on multi-level JSON in python

#3 Having trouble with decoding JSON in Python

#4 Iterating through JSON to find if a defined object exists

#5 Find a value within nested json dictionary in python

Notes: Google demo of *semantic search on 8 million Stack Overflow questions, returning relevant results in milliseconds.*
Source: Google

Announcement

- Following what is now a well-worn path, Google announced several points of support for vector stores/databases within Google Cloud. The company announced, for example, a new feature of AlloyDB called AlloyDB AI, which features pgvector-compatible vector search that's purported to be up to 10 times faster than regular vector-enabled PostgreSQL.

Assessment

- Certainly, every vendor with a database by now has announced or will soon announce vector store capabilities, particularly in support of retrieval augmented generation (RAG) use cases. Google is no different, announcing several vector store options at the show. AlloyDB AI is important because it seeks to unite vectors, models, and data, all within one familiar database.
- Related, Google announced a migration tool built specifically to bring Oracle MySQL users over to AlloyDB. Given Oracle's recent work to blend analytics and operational data with Oracle MySQL HeatWave, there's a battle brewing here between the two vendors.

Fresh from the garden

Announcement

- Google announced several new additions to Vertex AI Model Garden, adding Meta's popular Llama 2 and very recently released Code Llama. The company also added support for Technology Innovation Institute's Falcon LLM. And the company pre-announced support for Anthropic's Claude 2.

Assessment

- Like its main competitors (AWS, Microsoft, IBM, and Salesforce), the addition of native support for not just first party (e.g., Google PaLM 2), but also for open source and third party models like Llama 2 and Falcon will serve as literal jet fuel for the company's AI development platform, Vertex AI. As early adopters have already discovered, implementing GenAI solutions is an engineering problem that often involves highly iterative and deeply comparative looks at the way different models behave opposite one another with different prompt engineering and fine-tuning methodologies.

Google's first-party foundation models

- *PaLM for Text*
- *PaLM for Chat*
- *Imagen for text-to-image*
- *Codey for code completion*
- *Chirp for speech-to-text*

Solid foundations

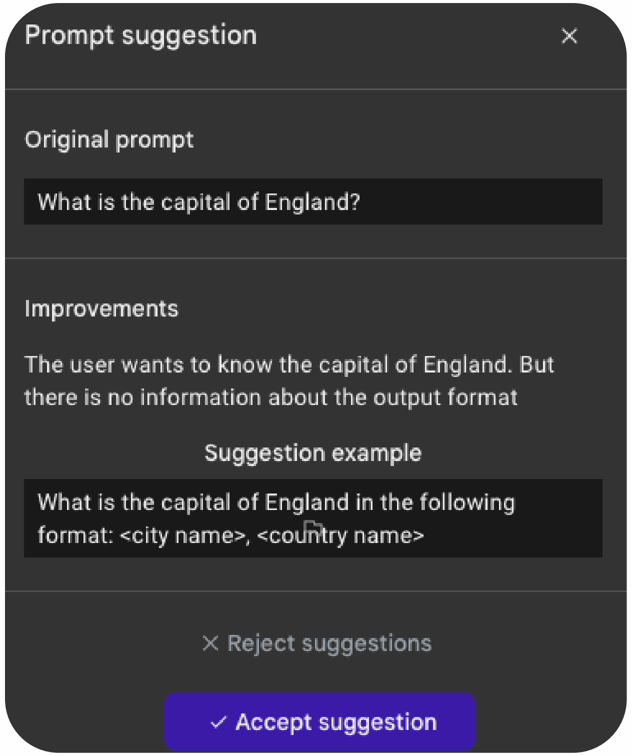
Announcement

- During the show, Google announced updates to its core LLM, Google PaLM 2, adding releasing support for 38 languages and extending the model's context window to a whopping 32,000-tokens, which will enable the model to ingest much larger documents and software. The company also announced several updates to its code generation model, Codey, as well as its image generation model, Imagen.

Assessment

- It's important to note that Google is not sitting on its DeepMind technical laurels in plying its first-party, flagship GenAI models. Beyond expanding the basics for PaLM 2 (e.g., context window expansion), the vendor announced that it's HIPAA-compliant Med-PaLM 2 model will be available in preview. And it introduced what appears to be one of the first instances of digital watermarking to Imagen. Powered by Google DeepMind's SynthID technology, Imagen can now provide companies with assurance that AI-generated artwork is authentic and free of potential copyright concerns.

Prompt engineering augmented support



Prompt suggestion

Original prompt

What is the capital of England?

Improvements

The user wants to know the capital of England. But there is no information about the output format

Suggestion example

What is the capital of England in the following format: <city name>, <country name>

Notes: Automated prompt engineering suggestion in Google MakerSuite.
Source: Google

Notebooks get real

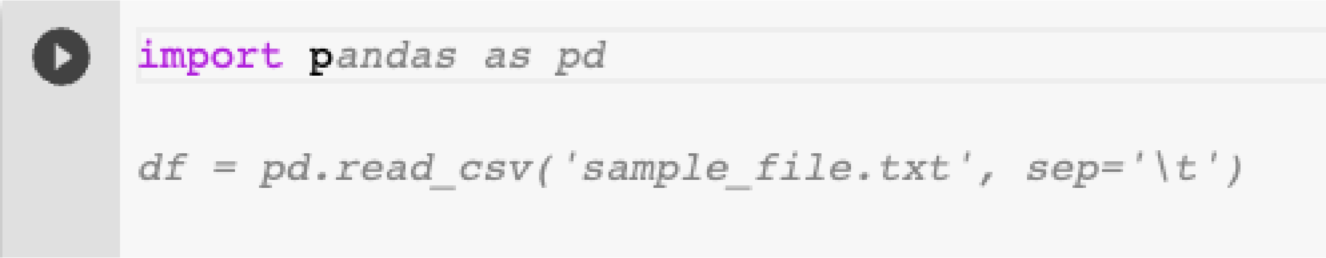
Announcement

- Google announced Google Colab Enterprise, a new offering based on its well-regarded but prosumer-oriented data science notebook experience, Google Colab. Available in preview today with general availability planned for later in September, this new version of Colab will integrate with Vertex AI and other Google Cloud services.

Assessment

- A move that has been coming for some time, the introduction of Colab Enterprise comes as no surprise. However, it is tremendously important to the vendor, as Colab can now serve as a direct, consistent path between enterprise students/enthusiasts and enterprise practitioners.
- This is also a big deal because it enables Google to directly integrate Colab within its numerous AI, data, and analytics solutions. Already, for example, the company announced that BigQuery Studio would work directly with Colab Enterprise, extending enterprise security and compliance support to Colab notebooks.

Duet AI within Google Colab

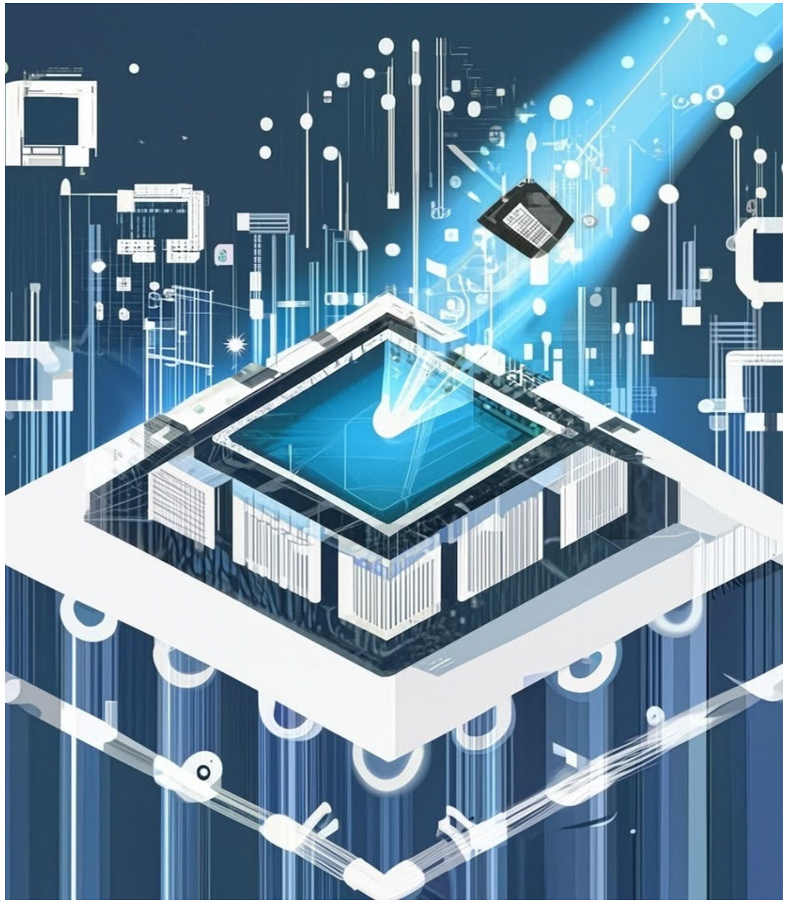


```
import pandas as pd

df = pd.read_csv('sample_file.txt', sep='\t')
```

Notes: Here, the user need only type “import p” to initiate a Duet AI completion recommendation.
Source: Google

Big data and big models



Source: Google Duet AI

Announcement

- Beyond introducing AlloyDB AI, Google made several database-related announcements, bringing Duet AI directly into Google Cloud Spanner (AlloyDB and Cloud SQL will follow shortly). The vendor also added new opportunities for customers to meld transactional and analytical data via Cloud Spanner Data Boost, which lets users analyze Cloud Spanner data in BigQuery and other tools without impacting performance.

Assessment

- The importance of fielding GenAI augmentative capabilities within and across Google's wide portfolio of data and analytics tools cannot be overstated. Omdia foresees a rapid shift toward GenAI-assisted code generation, product navigation, and management tasks. With Duet AI in Cloud Spanner, for example, data professionals can generate code to structure, modify, or query data in a purely declarative manner, using only natural language.
- It's also interesting to see Google put Duet AI to use well beyond basic augmentation to also encapsulate automation. The company, for instance, is putting Duet AI to work in automating complex tasks like migrating Oracle customers to Google Cloud PostgreSQL (now in GA) and eventually AlloyDB.

LLMOps in overdrive

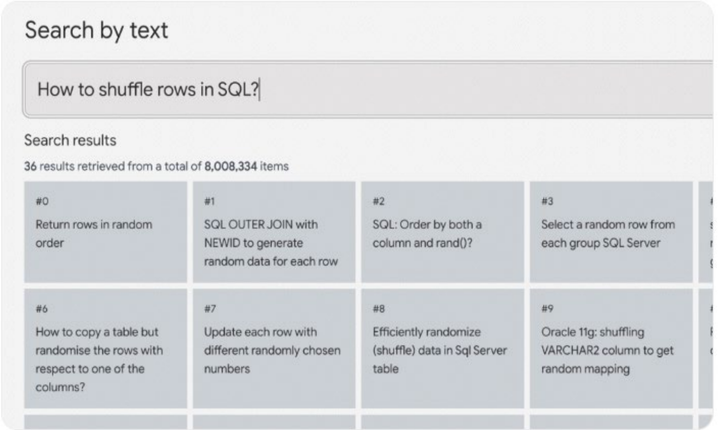
Announcement

- Lastly, Google announced a wide array of innovations, all taking aim at operationalizing LLMs (referred to as LLMOps). For example, the company introduced Automatic Side by Side and Automatic Metrics features within Vertex AI, which lets practitioners test and compare several LLMs against a ground truth reference dataset. Google also updated Vertex AI Feature Store, building it directly into BigQuery.

Assessment

- It is important to note that these together with Google's early noted announcements surrounding vector embeddings and API-level grounding for PaLM (in private preview) will work across all models within the company's Model Garden through Google Vertex AI. This includes both text and image embeddings. Going further, when these embeddings are coupled with Google's Matching Engine (Google's vector search service), customers can ground their model output against business data in tabular format. This is crucial for Google because it opens up the ability for customers to work synchronously between 3rd party, open source, and first party models, all using the same responsible AI tooling.

Semantic search for grounding LLMs



Search by text

How to shuffle rows in SQL?

Search results

36 results retrieved from a total of 8,008,334 items

#0
Return rows in random order

#1
SQL OUTER JOIN with NEWID to generate random data for each row

#2
SQL: Order by both a column and rand()?

#3
Select a random row from each group SQL Server

#6
How to copy a table but randomise the rows with respect to one of the columns?

#7
Update each row with different randomly chosen numbers

#8
Efficiently randomize (shuffle) data in Sql Server table

#9
Oracle 11g: shuffling VARCHAR2 column to get random mapping

Notes: A demonstration showing search results across 8 million Stack Overflow questions
Source: Google

Key takeaway

Key takeaway from Google Cloud Next 2023

- Google is currently operating in "startup" mode, accelerating time to market for key GenAI capabilities like Duet AI in Google Cloud, AlloyDB AI, Grounding in Vertex AI, PaLM 2 enhancements, et al.
- At a high level, this will help the vendor overcome market perception that it been too conservative relative to rivals Microsoft and OpenAI.
- In looking at the announcement in detail, however, Google is not trying to recapture market momentum.
- Rather, the vendor is assembling the infrastructure necessary to operationalize GenAI solutions in the enterprise at scale with full control over Responsible AI H3 concerns (model helpfulness, harmlessness, and honesty).
- Taking this a step further, Google intends to help customers do so no matter which models they select, be those first-party, third-party, or open source models. With more than 100 models available in Vertex AI Model Garden, the company is well on its way in this regard.

Bradley Shimmin

Chief Analyst, AI Platforms, analytics and data management

Appendix

Appendix

Further reading

AWS extends generative AI platform capabilities with autonomous agents, (August 2023)

Technology Analysis: Responsible LLM Tools and Practices, (August 2023)

Google Cloud Next 2023 Keynote

Google Cloud Next 2023 Keynote on Duet AI in GCP analytics tools

Google Vertex AI Model Garden

Author

Bradley Shimmin, Chief Analyst, AI platforms, analytics and data management

askananalyst@omdia.com

Omdia Consulting

We hope that this analysis will help you make informed and imaginative business decisions. If you have further requirements, Omdia’s consulting team may be able to help you. For more information about Omdia’s consulting capabilities, please contact us directly at consulting@omdia.com.

Citation Policy

Request external citation and usage of Omdia research and data via citations@omdia.com.

Disclaimer

The Omdia research, data and information referenced herein (the “Omdia Materials”) are the copyrighted property of Informa Tech and its subsidiaries or affiliates (together “Informa Tech”) or its third party data providers and represent data, research, opinions, or viewpoints published by Informa Tech, and are not representations of fact.

The Omdia Materials reflect information and opinions from the original publication date and not from the date of this document. The information and opinions expressed in the Omdia Materials are subject to change without notice and Informa Tech does not have any duty or responsibility to update the Omdia Materials or this publication as a result.

Omdia Materials are delivered on an “as-is” and “as-available” basis. No representation or warranty, express or implied, is made as to the fairness, accuracy, completeness, or correctness of the information, opinions, and conclusions contained in Omdia Materials.

To the maximum extent permitted by law, Informa Tech and its affiliates, officers, directors, employees, agents, and third party data providers disclaim any liability (including, without limitation, any liability arising from fault or negligence) as to the accuracy or completeness or use of the Omdia Materials. Informa Tech will not, under any circumstance whatsoever, be liable for any trading, investment, commercial, or other decisions based on or made in reliance of the Omdia Materials.

Get in touch

Americas	Europe, Middle East & Africa	Asia Pacific
customersuccess@omdia.com	customersuccess@omdia.com	customersuccess@omdia.com
08:00 – 18:00 GMT -5	8:00 – 18:00 GMT	08:00 – 18:00 GMT + 8