Sentiment analysis with NetApp AI

NetApp Solutions

NetApp
July 31, 2024

Table of Contents

Sentiment analysis with NetApp AI

TR-4910: Sentiment Analysis from Customer Communications with NetApp AI

Rick Huang, Sathish Thyagarajan, and David Arnette, NetApp
Diego Sosa-Coba, SFL Scientific

This technical report provides design guidance for customers to perform sentiment analysis in an enterprise-level global support center by using NetApp data management technologies with an NVIDIA software framework using transfer learning and conversational AI. This solution is applicable to any industry wanting to gain customer insights from recorded speech or text files representing chat logs, emails, and other text or audio communications. We implemented an end-to-end pipeline to demonstrate automatic speech recognition, real-time sentiment analysis, and deep-learning natural-language- processing model- retraining capabilities on a GPU-accelerated compute cluster with NetApp cloud-connected all flash storage. Massive, state-of-the-art language models can be trained and optimized to perform inference rapidly with the global support center to create an exceptional customer experience and objective, long-term employee performance evaluations.

Sentiment analysis is a field of study within Natural Language Processing (NLP) by which positive, negative, or neutral sentiments are extracted from text. Conversational AI systems have risen to a near global level of integration as more and more people come to interact with them. Sentiment analysis has a variety of use cases, from determining support center employee performance in conversations with callers and providing appropriate automated chatbot responses to predicting a firm's stock price based on the interactions between firm representatives and the audience at quarterly earnings calls. Furthermore, sentiment analysis can be used to determine the customer's view on the products, services, or support provided by the brand.

This end-to-end solution uses NLP models to perform high level sentiment analysis that enables support-center analytical frameworks. Audio recordings are processed into written text, and sentiment is extracted from each sentence in the conversation. Results, aggregated into a dashboard, can be crafted to analyze conversation sentiments, both historically and in real-time. This solution can be generalized to other solutions with similar data modalities and output needs. With the appropriate data, other use cases can be accomplished. For example, company earnings calls can be analyzed for sentiment using the same end-to-end pipeline. Other forms of NLP analyses, such as topic modeling and named entity recognition (NER), are also possible due to the flexible nature of the pipeline.

These AI implementations were made possible by NVIDIA RIVA, the NVIDIA TAO Toolkit, and the NetApp DataOps Toolkit working together. NVIDIA's tools are used to rapidly deploy highly performant AI solutions using prebuilt models and pipelines. The NetApp DataOps Toolkit simplifies various data management tasks to speed up development.

Customer value

Businesses see value from an employee-assessment and customer-reaction tool for text, audio, and video conversation for sentiment analysis. Managers benefit from the information presented in the dashboard, allowing for an assessment of the employees and customer satisfaction based on both sides of the conversation.

Additionally, the NetApp DataOps Toolkit manages the versioning and allocation of data within the customer's infrastructure. This leads to frequent updates of the analytics presented within the dashboard without creating unwieldy data storage costs.

Use cases

Due to the number of calls that these support centers process, assessment of call performance could take significant time if performed manually. Traditional methods, like bag-of-words counting and other methods, can achieve some automation, but these methods do not capture more nuanced aspects and semantic context of dynamic language. AI modeling techniques can be used to perform some of these more nuanced analyses in an automated manner. Furthermore, with the current state of the art, pretrained modeling tools published by NVIDIA, AWS, Google, and others, an end-to-end pipeline with complex models can be now stood up and customized with relative ease.

An end-to-end pipeline for support center sentiment analysis ingests audio files in real time as employees converse with callers. Then, these audio files are processed for use in the speech-to-text component which converts them into a text format. Each sentence in the conversation receives a label indicating the sentiment (positive, negative, or neutral).

Sentiment analysis can provide an essential aspect of the conversations for assessment of call performance. These sentiments add an additional level of depth to the interactions between employees and callers. The AI-assisted sentiment dashboard provides managers with a real-time tracking of sentiment within a conversation, along with a retrospective analysis of the employee's past calls.

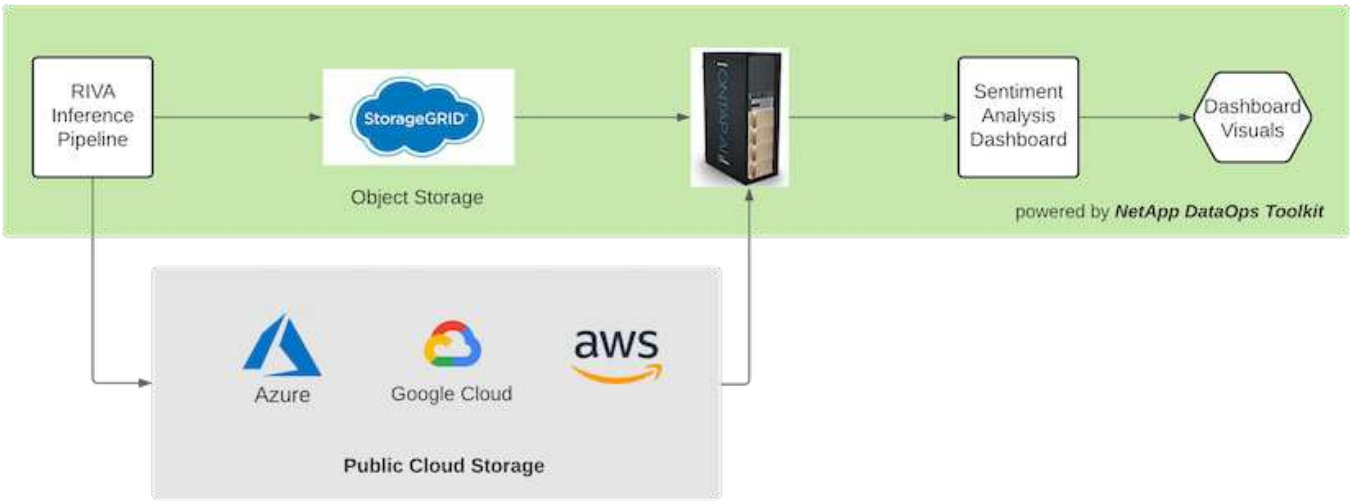
There are prebuilt tools that can be combined in powerful ways to quickly create an end-to-end AI pipeline to solve this problem. In this case, the NVIDIA RIVA library can be used to perform the two in-series tasks: audio transcription and sentiment analysis. The first is a supervised learning signal processing algorithm and the second is a supervised learning NLP classification algorithm. These out-of-the-box algorithms can be fine-tuned for any relevant use case with business-relevant data using the NVIDIA TAO Toolkit. This leads to more accurate and powerful solutions being built for only a fraction of the cost and resources. Customers can incorporate the NVIDIA Maxine framework for GPU-accelerated video conferencing applications in their support center design.

The following use cases are at the core of this solution. Both use cases use the TAO Toolkit for model fine-tuning and RIVA for model deployment.

- Speech-to-text
- Sentiment analysis

To analyze support center interactions between employees and customers, each customer conversation in the form of audio calls can be run through the pipeline to extract sentence-level sentiments. Those sentiments can then be verified by a human to justify the sentiments or adjust them as needed. The labeled data is then passed onto the fine-tuning step to improve sentiment predictions. If labeled sentiment data already exists, then model fine-tuning can be expedited. In either case, the pipeline is generalizable to other solutions that require the ingestion of audio and the classification of sentences.

AI sentiment outputs are either uploaded to an external cloud database or to a company- managed storage system. The sentiment outputs are transferred from this larger database into local storage for use within the dashboard that displays the sentiment analysis for managers. The dashboard's primary functionality is to interface with the customer service employee in real time. Managers can assess and provide feedback on employees during their calls with live updates of the sentiment of each sentence, as well as an historic review of the employee's past performance or customer reactions.



The NetApp DataOps Toolkit can continue to manage data storage systems even after the RIVA inference pipeline generates sentiment labels. Those AI results can be uploaded to a data storage system managed by the NetApp DataOps Toolkit. The data storage systems must be capable of managing hundreds of inserts and selects every minute. The local device storage system queries the larger data storage in real-time for extraction. The larger data storage instance can also be queried for historical data to further enhance the dashboard experience. The NetApp DataOps Toolkit facilitates both these uses by rapidly cloning data and distributing it across all the dashboards that use it.

Target Audience

The target audience for the solution includes the following groups:

- Employee managers
- Data engineers/data scientists
- IT administrators (on-premises, cloud, or hybrid)

Tracking sentiments throughout conversations is a valuable tool for assessing employee performance. Using the AI-dashboard, managers can see how employees and callers change their feelings in real time, allowing for live assessments and guidance sessions. Moreover, businesses can gain valuable customer insights from customers engaged in vocal conversations, text chatbots, and video conferencing. Such customer analytics uses the capabilities of multimodal processing at scale with modern, state-of-the-art AI models and workflows.

On the data side, a large number of audio files are processed daily by the support center. The NetApp DataOps Toolkit facilitates this data handling task for both the periodic fine-tuning of models and sentiment analysis dashboards.

IT administrators also benefit from the NetApp DataOps Toolkit as it allows them to move data quickly between deployment and production environments. The NVIDIA environments and servers must also be managed and distributed to allow for real time inference.

Architecture

The architecture of this support center solution revolves around NVIDIA's prebuilt tools and the NetApp DataOps Toolkit. NVIDIA's tools are used to rapidly deploy high-performance AI-solutions using prebuilt models and pipelines. The NetApp DataOps Toolkit simplifies various data management tasks to speed up development.

Solution technology

NVIDIA RIVA is a GPU-accelerated SDK for building multimodal conversational AI applications that deliver real-time performance on GPUs. The NVIDIA Train, Adapt, and Optimize (TAO) Toolkit provides a faster, easier way to accelerate training and quickly create highly accurate and performant, domain-specific AI models.

The NetApp DataOps Toolkit is a Python library that makes it simple for developers, data scientists, DevOps engineers, and data engineers to perform various data management tasks. This includes near-instantaneous provisioning of a new data volume or JupyterLab workspace, near-instantaneous cloning of a data volume or JupyterLab workspace, and near-instantaneous snapshotting of a data volume or JupyterLab workspace for traceability and baselining.

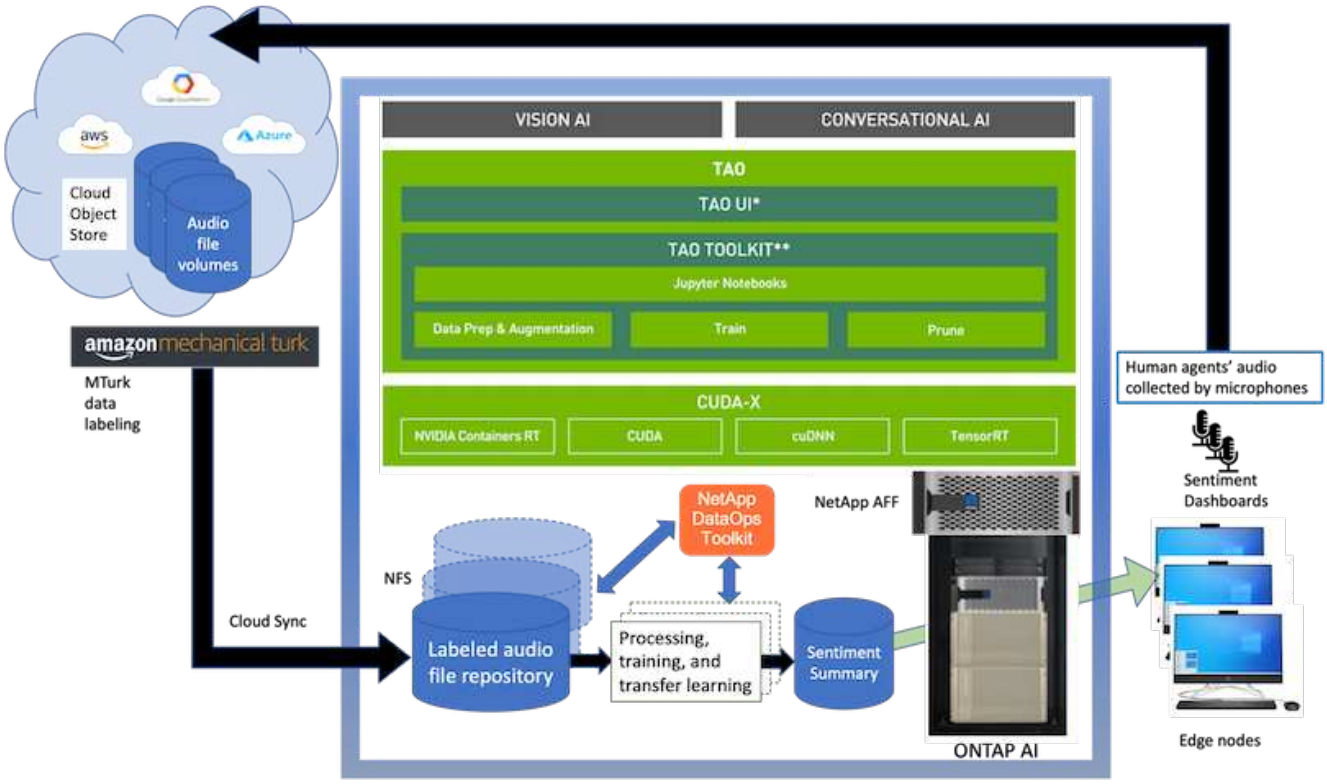
Architectural Diagram

The following diagram shows the solution architecture. There are three main environment categories: the cloud, the core, and the edge. Each of the categories can be geographically dispersed. For example, the cloud contains object stores with audio files in buckets in different regions, whereas the core might contain datacenters linked via a high-speed network or NetApp BlueXP Copy and Sync. The edge nodes denote the individual human agent's daily working platforms, where interactive dashboard tools and microphones are available to visualize sentiment and collect audio data from conversations with customers.

In GPU-accelerated datacenters, businesses can use the NVIDIA RIVA framework to build conversational AI applications, to which the Tao Toolkit connects for model finetuning and retraining using transfer L-learning

techniques. These compute applications and workflows are powered by the NetApp DataOps Toolkit, enabling the best data management capabilities ONTAP has to offer. The toolkit allows corporate data teams to rapidly prototype their models with associated structured and unstructured data via snapshots and clones for traceability, versioning, A/B testing, thus providing security, governance, and regulatory compliance. See the section "Storage Design" for more details.

This solution demonstrates the audio file processing, NLP model training, transfer learning, and data management detail steps. The resulting end-to-end pipeline generates a sentiment summary that displays in real-time on human support agents' dashboards.



Hardware requirements

The following table lists the hardware components that are required to implement the solution. The hardware components that are used in any particular implementation of the solution might vary based on customer requirements.

Response latency tests	Time (milliseconds)
Data processing	10
Inferencing	10

These response-time tests were run on 50,000+ audio files across 560 conversations. Each audio file was ~100KB in size as an MP3 and ~1 MB when converted to WAV. The data processing step converts MP3s into WAV files. The inference steps convert the audio files into text and extract a sentiment from the text. These steps are all independent of one another and can be parallelized to speed up the process.

Taking into account the latency of transferring data between stores, managers should be able to see updates to the real time sentiment analysis within a second of the end of the sentence.

NVIDIA RIVA hardware

Hardware	Requirements
OS	Linux x86_64
GPU memory (ASR)	Streaming models: ~5600 MB Non-streaming models: ~3100 MB
GPU memory (NLP)	~500MB per BERT model

NVIDIA TAO Toolkit hardware

Hardware	Requirements
System RAM	32GB
GPU RAM	32GB
CPU	8 core
GPU	NVIDIA (A100, V100 and RTX 30x0)
SSD	100GB

Flash storage system

NetApp ONTAP 9

ONTAP 9.9, the latest generation of storage management software from NetApp, enables businesses to modernize infrastructure and transition to a cloud-ready data center. Leveraging industry-leading data management capabilities, ONTAP enables the management and protection of data with a single set of tools, regardless of where that data resides. You can also move data freely to wherever it is needed: the edge, the core, or the cloud. ONTAP 9.9 includes numerous features that simplify data management, accelerate, and protect critical data, and enable next generation infrastructure capabilities across hybrid cloud architectures.

NetApp BlueXP Copy and Sync

BlueXP Copy and Sync is a NetApp service for rapid and secure data synchronization that allows you to transfer files between on-premises NFS or SMB file shares to any of the following targets:

- NetApp StorageGRID
- NetApp ONTAP S3
- NetApp Cloud Volumes Service
- Azure NetApp Files
- Amazon Simple Storage Service (Amazon S3)
- Amazon Elastic File System (Amazon EFS)
- Azure Blob
- Google Cloud Storage
- IBM Cloud Object Storage

BlueXP Copy and Sync moves the files where you need them quickly and securely. After your data is transferred, it is fully available for use on both the source and the target. BlueXP Copy and Sync continuously

synchronizes the data, based on your predefined schedule, moving only the deltas, so that time and money spent on data replication is minimized. BlueXP Copy and Sync is a software as a service (SaaS) tool that is simple to set up and use. Data transfers that are triggered by BlueXP Copy and Sync are carried out by data brokers. You can deploy BlueXP Copy and Sync data brokers in AWS, Azure, Google Cloud Platform, or on-premises.

NetApp StorageGRID

The StorageGRID software-defined object storage suite supports a wide range of use cases across public, private, and hybrid multi-cloud environments seamlessly. With industry leading innovations, NetApp StorageGRID stores, secures, protect, and preserves unstructured data for multi-purpose use including automated lifecycle management for long periods of time. For more information, see the NetApp StorageGRID site.

Software requirements

The following table lists the software components that are required to implement this solution. The software components that are used in any particular implementation of the solution might vary based on customer requirements.

Host machine	Requirements
RIVA (formerly JARVIS)	1.4.0
TAO Toolkit (formerly Transfer Learning Toolkit)	3.0
ONTAP	9.9.1
DGX OS	5.1
DOTK	2.0.0

NVIDIA RIVA Software

Software	Requirements
Docker	>19.02 (with nvidia-docker installed)>=19.03 if not using DGX
NVIDIA Driver	465.19.01+ 418.40+, 440.33+, 450.51+, 460.27+ for Data Center GPUs
Container OS	Ubuntu 20.04
CUDA	11.3.0
cuBLAS	11.5.1.101
cuDNN	8.2.0.41
NCCL	2.9.6
TensorRT	7.2.3.4
Triton Inference Server	2.9.0

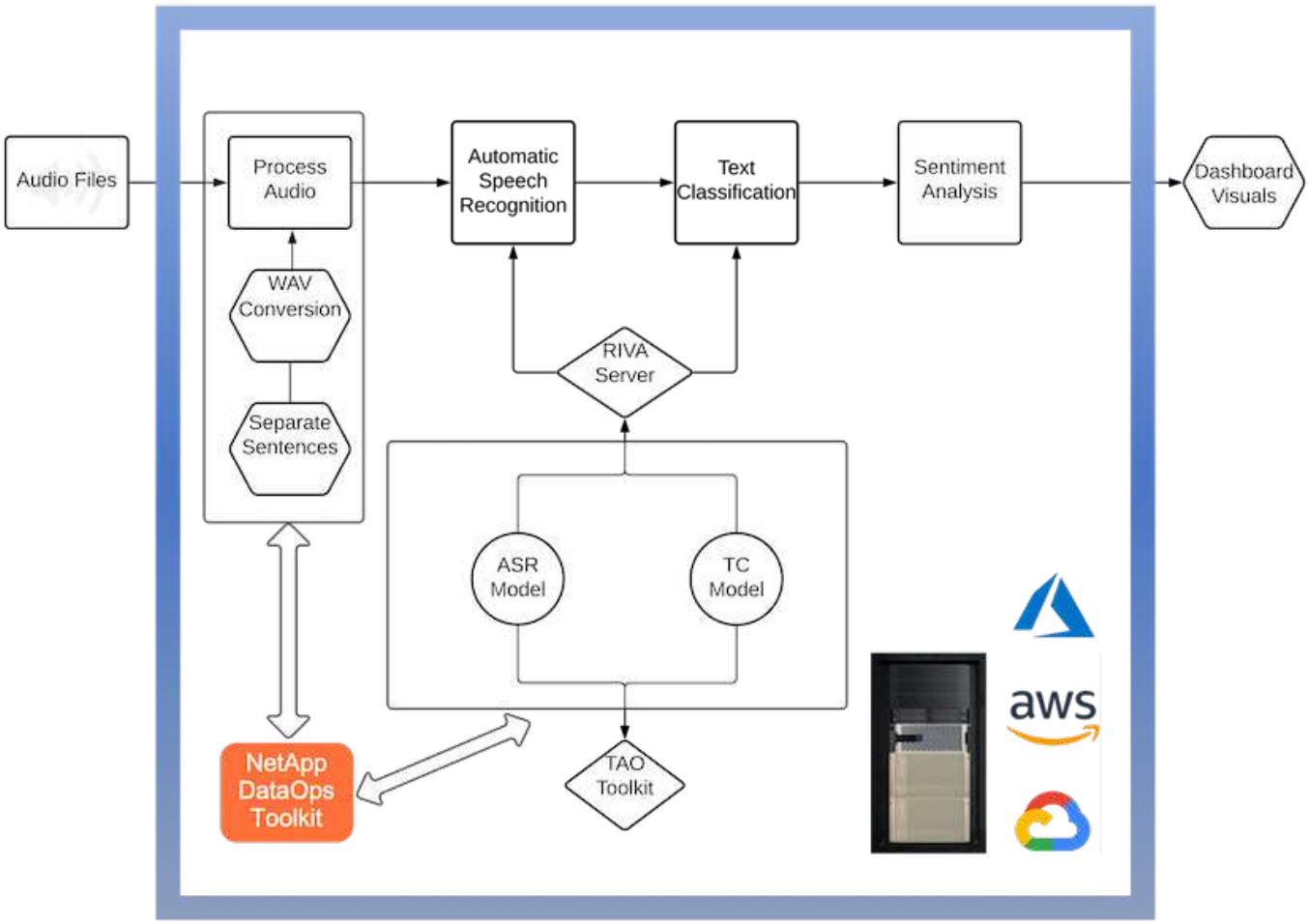
NVIDIA TAO Toolkit software

Software	Requirements
Ubuntu 18.04 LTS	18.04
python	>=3.6.9
docker-ce	>19.03.5
docker-API	1.40
nvidia-container-toolkit	>1.3.0-1
nvidia-container-runtime	3.4.0-1
nvidia-docker2	2.5.0-1
nvidia-driver	>455
python-pip	>21.06
nvidia-pyindex	Latest version

Use case details

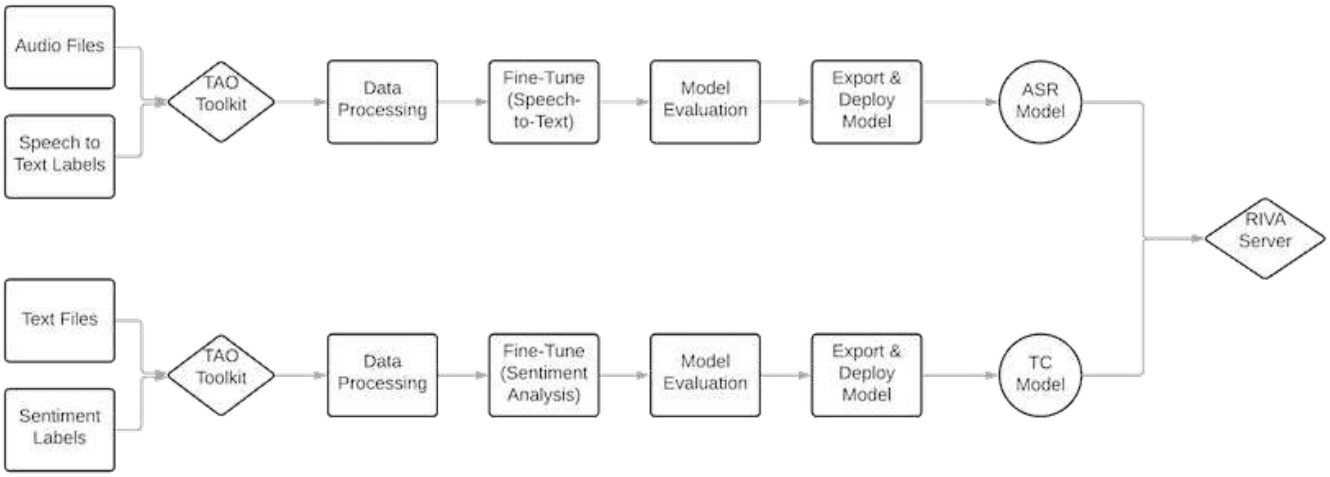
This solution applies to the following use cases:

- Speech-to-text
- Sentiment analysis



The speech-to-text use case begins by ingesting audio files for the support centers. This audio is then processed to fit the structure required by RIVA. If the audio files have not already been split into their units of analysis, then this must be done before passing the audio to RIVA. After the audio file is processed, it is passed to the RIVA server as an API call. The server employs one of the many models it is hosting and returns a response. This speech-to-text (part of Automatic Speech Recognition) returns a text representation of the audio. From there, the pipeline switches over to the sentiment analysis portion.

For sentiment analysis, the text output from the Automatic Speech Recognition serves as the input to the Text Classification. Text Classification is the NVIDIA component for classifying text to any number of categories. The sentiment categories range from positive to negative for the support center conversations. The performance of the models can be assessed using a holdout set to determine the success of the fine-tuning step.



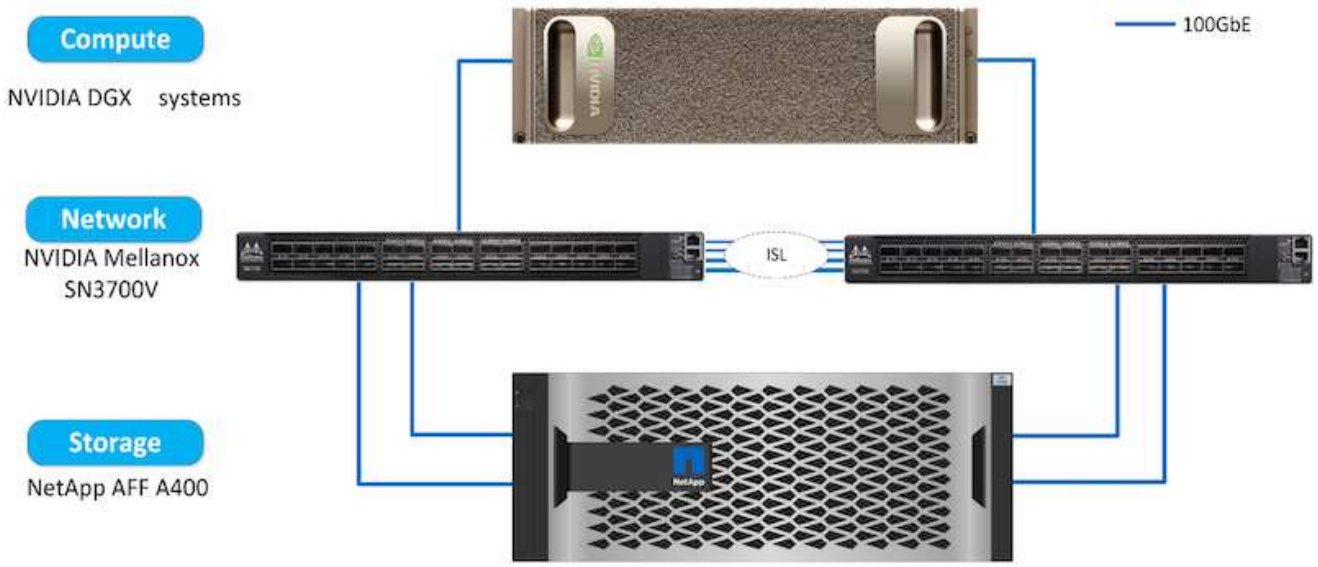
A similar pipeline is used for both the speech-to-text and sentiment analysis within the TAO Toolkit. The major difference is the use of labels which are required for the fine-tuning of the models. The TAO Toolkit pipeline begins with the processing of the data files. Then the pretrained models (coming from the NVIDIA NGC Catalog) are fine-tuned using the support center data. The fine-tuned models are evaluated based on their corresponding performance metrics and, if they are more performant than the pretrained models, are deployed to the RIVA server.

Design considerations

This section describes the design considerations for the different components of this solution.

Network and compute design

Depending on the restrictions on data security, all data must remain within the customer's infrastructure or a secure environment.



Storage design

The NetApp DataOps Toolkit serves as the primary service for managing storage systems. The DataOps Toolkit is a Python library that makes it simple for developers, data scientists, DevOps engineers, and data engineers to perform various data management tasks, such as near-instantaneous provisioning of a new data volume or JupyterLab workspace, near-instantaneous cloning of a data volume or JupyterLab workspace, and near-instantaneous snapshotting of a data volume or JupyterLab workspace for traceability or baselining. This Python library can function as either a command line utility or a library of functions that can be imported into any Python program or Jupyter Notebook.

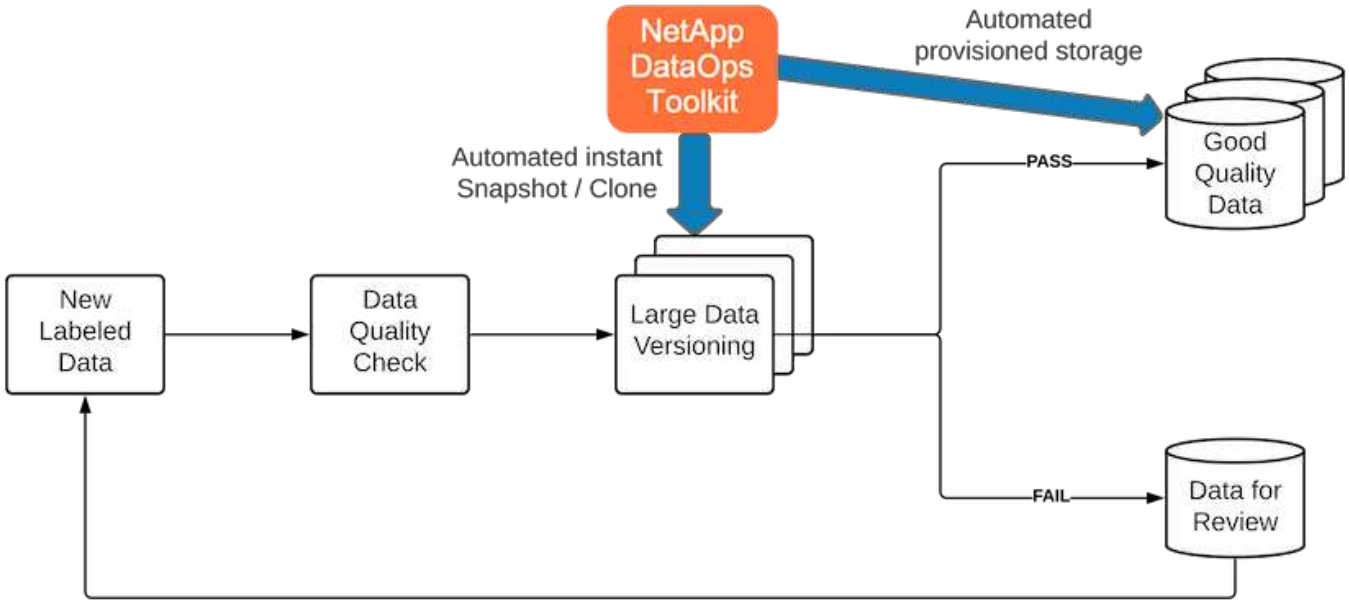
RIVA best practices

NVIDIA provides several general best data practices for using RIVA:

- **Use lossless audio formats if possible.** The use of lossy codecs such as MP3 can reduce quality.
- **Augment training data.** Adding background noise to audio training data can initially decrease accuracy and yet increase robustness.
- **Limit vocabulary size if using scraped text.** Many online sources contain typos or ancillary pronouns and uncommon words. Removing these can improve the language model.
- **Use a minimum sampling rate of 16kHz if possible.** However, try not to resample, because doing so decreases audio quality.

In addition to these best practices, customers must prioritize gathering a representative sample dataset with accurate labels for each step of the pipeline. In other words, the sample dataset should proportionally reflect specified characteristics exemplified in a target dataset. Similarly, the dataset annotators have a responsibility to balance accuracy and the speed of labeling so that the quality and quantity of the data are both maximized. For example, this support center solution requires audio files, labeled text, and sentiment labels. The sequential nature of this solution means that errors from the beginning of the pipeline are propagated all the way through to the end. If the audio files are of poor quality, the text transcriptions and sentiment labels will be as well.

This error propagation similarly applies to the models trained on this data. If the sentiment predictions are 100% accurate but the speech-to-text model performs poorly, then the final pipeline is limited by the initial audio- to- text transcriptions. It is essential that developers consider each model's performance individually and as a component of a larger pipeline. In this particular case, the end goal is to develop a pipeline that can accurately predict the sentiment. Therefore, the overall metric on which to assess the pipeline is the accuracy of the sentiments, which the speech-to-text transcription directly affects.



The NetApp DataOps Toolkit complements the data quality-checking pipeline through the use of its near-instantaneous data cloning technology. Each labeled file must be assessed and compared to the existing labeled files. Distributing these quality checks across various data storage systems ensures that these checks are executed quickly and efficiently.

Deploying support center sentiment analysis

Deploying the solution involves the following components:

1. NetApp DataOps Toolkit
2. NGC Configuration
3. NVIDIA RIVA Server
4. NVIDIA TAO Toolkit
5. Export TAO models to RIVA

To perform deployment, complete the following steps:

NetApp DataOps Toolkit: Support center sentiment analysis

To use the NetApp DataOps Toolkit, complete the following steps:

1. Pip install the toolkit.

```
python3 -m pip install netapp-dataops-traditional
```

2. Configure the data management

```
netapp_dataops_cli.py config
```

NGC configuration: Support center sentiment analysis

To set up NVIDIA NGC, complete the following steps:

1. Download the NGC.

```
wget -O ngccli_linux.zip
https://ngc.nvidia.com/downloads/ngccli_linux.zip && unzip -o
ngccli_linux.zip && chmod u+x ngc
```

2. Add your current directory to path.

```
echo "export PATH=\"\$PATH:$(pwd)\"" >> ~/.bash_profile && source
~/.bash_profile
```

3. You must configure NGC CLI for your use so that you can run the commands. Enter the following command, including your API key when prompted.

```
ngc config set
```

For operating systems that are not Linux-based, visit here.

NVIDIA RIVA server: Support center sentiment analysis

To set up NVIDIA RIVA, complete the following steps:

1. Download the RIVA files from NGC.

```
ngc registry resource download-version
nvidia/riva/riva_quickstart:1.4.0-beta
```

2. Initialize the RIVA setup (`riva_init.sh`).
3. Start the RIVA server (`riva_start.sh`).
4. Start the RIVA client (`riva_start_client.sh`).
5. Within the RIVA client, install the audio processing library (FFMPEG)

```
apt-get install ffmpeg
```

6. Start the Jupyter server.
7. Run the RIVA Inference Pipeline Notebook.

NVIDIA TAO Toolkit: Support center sentiment analysis

To set up NVIDIA TAO Toolkit, complete the following steps:

1. Prepare and activate a virtual environment for TAO Toolkit.
2. Install the required packages.
3. Manually pull the image used during training and fine-tuning.

```
docker pull nvcr.io/nvidia/tao/tao-toolkit-pyt:v3.21.08-py3
```

4. Start the Jupyter server.
5. Run the TAO Fine-Tuning Notebook.

Export TAO models to RIVA: Support center sentiment analysis

To use TAO Toolkit models in RIVA, complete the following steps:

1. Save models within the TAO Fine-Tuning Notebook.
2. Copy TAO trained models to the RIVA model directory.
3. Start the RIVA server (`riva_start.sh`).

Deployment roadblocks

Here are a few things to keep in mind as you develop your own solution:

- The NetApp DataOps Toolkit is installed first to ensure that the data storage system runs optimally.
- NVIDIA NGC must be installed before anything else because it authenticates the downloading of images and models.
- RIVA must be installed before the TAO Toolkit. The RIVA installation configures the docker daemon to pull images as needed.
- DGX and docker must have internet access to download the models.

Validation results

As mentioned in the previous section, errors are propagated throughout the pipeline whenever there are two or more machine learning models running in sequence. For this solution, the sentiment of the sentence is the most important factor in measuring the firm's stock risk level. The speech-to-text model, although essential to the pipeline, serves as the preprocessing unit before the sentiments can be predicted. What really matters is the difference in sentiment between the ground truth sentences and the predicted sentences. This serves as a proxy for the word error rate (WER). The speech-to-text accuracy is important, but the WER is not directly used in the final pipeline metric.

```
PIPELINE_SENTIMENT_METRIC = MEAN(DIFF(GT_sentiment, ASR_sentiment))
```

These sentiment metrics can be calculated for the F1 Score, Recall, and Precision of each sentence. The results can be then aggregated and displayed within a confusion matrix, along with the confidence intervals for each metric.

The benefit of using transfer learning is an increase in model performance for a fraction of data requirements, training time, and cost. The fine-tuned models should also be compared to their baseline versions to ensure the transfer learning enhances the performance instead of impairing it. In other words, the fine-tuned model should perform better on the support center data than the pretrained model.

Pipeline assessment

Test case	Details
Test number	Pipeline sentiment metric
Test prerequisites	Fine-tuned models for speech-to-text and sentiment analysis models
Expected outcome	The sentiment metric of the fine-tuned model performs better than the original pretrained model.

Pipeline sentiment metric

1. Calculate the sentiment metric for the baseline model.
2. Calculate the sentiment metric for the fine-tuned model.
3. Calculate the difference between those metrics.
4. Average the differences across all sentences.

Videos and demos

There are two notebooks that contain the sentiment analysis pipeline: “Support-Center-Model-Transfer-Learning-and-Fine-Tuning.ipynb” and “Support-Center-Sentiment-Analysis-Pipeline.ipynb”. Together, these notebooks demonstrate how to develop a pipeline to ingest support center data and extract sentiments from each sentence using state-of-the-art deep learning models fine-tuned on the user’s data.

Support Center - Sentiment Analysis Pipeline.ipynb

This notebook contains the inference RIVA pipeline for ingesting audio, converting it to text, and extracting sentiments for use in an external dashboard. Dataset are automatically downloaded and processed if this has not already been done. The first section in the notebook is the Speech-to-Text which handles the conversion of audio files to text. This is followed by the Sentiment Analysis section which extracts sentiments for each text sentence and displays those results in a format similar to the proposed dashboard.

This notebook must be run before the model training and fine-tuning because the MP3 dataset must be downloaded and converted into the correct format.

Call Center - Sentiment Analysis Pipeline

This notebook demonstrates how to build a pipeline for sentiment analysis of call center conversations. The goal of this pipeline is to develop sentiment analysis for use within an external dashboard.

This tutorial will guide you through the use of NVIDIA's RIVA for automatic speech recognition and text classification. This tutorial uses NetApp cloud storage for data storage and a pre-trained RIVA model.

Channels

These are the channels on which RIVA is hosting models.

- speech: `51051`
- voice: `61051`

These channels **must** be aligned with `riva_speech_api_port` and `riva_vision_api_port` within `config.sh`

In [4]:
```
speech_channel = "localhost:51051"
voice_channel = "localhost:61051"
```

Speech-To-Text

Automatic Speech Recognition (ASR) takes as input an audio stream or audio buffer and returns one or more text transcripts, along with additional optional metadata. ASR represents a full speech recognition pipeline that is GPU accelerated with optimized performance and accuracy. ASR supports synchronous and streaming recognition modes.

For more information on NVIDIA RIVA's Automatic Speech Recognition, visit here.

Constants

Use these constants to affect different aspects of this pipeline:

- `DATA_DIR`: base folder where data is stored
- `DATASET_NAME`: name of the call center dataset
- `COMPANY_DATE`: folder name identifying the particular call center conversation

Support Center - Model Training and Fine-Tuning.ipynb

The TAO Toolkit virtual environment must be set up before executing the notebook (see the TAO Toolkit section in the Commands Overview for installation instructions).

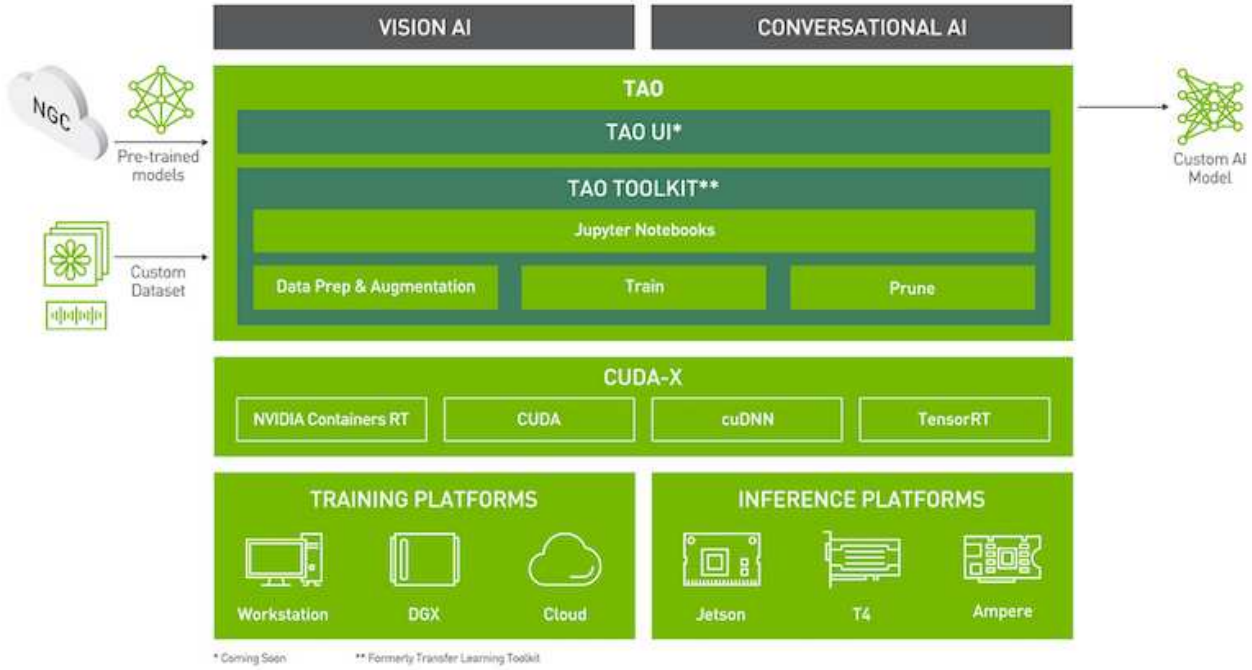
This notebook relies on the TAO Toolkit to fine-tune deep learning models on the customers data. As with the previous notebook, this one is separated into two sections for the Speech-to-Text and Sentiment Analysis components. Each section goes through data processing, model training and fine-tuning, evaluation of results, and model export. Finally, there is an end section for deploying both your fine-tuned models for use in RIVA.

Call Center - Model Transfer Learning and Fine-Tuning

TAO Toolkit is a python based AI toolkit for taking purpose-built pre-trained AI models and customizing them with your own data. Transfer learning extracts learned features from an existing neural network to a new one. Transfer learning is often used when creating a large training dataset is not feasible in order to enhance the base performance of state-of-the-art models.

For this call center solution, the speech-to-text and sentiment analysis models are fine-tuned on call center data to augment the model performance on business specific terminology.

For more information on the TAO Toolkit, please visit here.



Installing necessary dependencies

For ease of use, please install TAO Toolkit inside a python virtual environment. We recommend performing this step first and then launching the notebook from the virtual environment. Please refer to the README for these instructions.

Conclusion

As customer experience has become increasingly regarded as a key competitive battleground, an AI-augmented global support center becomes a critical component that companies in almost every industry cannot afford to neglect. The solution proposed in this technical report has been demonstrated to support the delivery of such exceptional customer experiences, and the challenge now is to ensure businesses are taking actions to modernize their AI infrastructure and workflows.

The best implementations of AI in customer service are not to replace human agents. Rather, AI can empower them to create exceptional customer experiences via real-time sentiment analysis, dispute escalation, and multimodal affective computing to detect verbal, non-verbal, and facial cues with which comprehensive AI

models can make recommendations at scale and supplement what an individual human agent might be lacking. AI can also provide a better match between a particular customer with currently available agents. Using AI, businesses can extract valuable customer sentiment regarding their thoughts and impressions of the provider's products, services, and brand image.

The solution can also be used to construct time-series data for support agents to serve as an objective performance evaluation metric. Conventional customer satisfaction surveys often lack sufficient responses. By collecting long-term employee and customer sentiment, employers can make informed decisions regarding support agents' performance.

The combination of NetApp, SFL Scientific, opens-source orchestration frameworks, and NVIDIA brings the latest technologies together as managed services with great flexibility to accelerate technology adoption and improve the time to market for new AI/ML applications. These advanced services are delivered on-premises that can be easily ported for cloud-native environment as well as hybrid deployment architectures.

Where to find additional information

To learn more about the information that is described in this document, review the following documents and/or websites:

- 3D interactive demos

  www.netapp.com/ai

- Connect directly with a NetApp AI specialist

  https://www.netapp.com/artificial-intelligence/

- NVDIA Base Command Platform with NetApp solution brief

  https://www.netapp.com/pdf.html?item=/media/32792-DS-4145-NVIDIA-Base-Command-Platform-with-NetApp.pdf

- NetApp for AI 10 Good Reasons infographic

  https://www.netapp.com/us/media/netapp-ai-10-good-reasons.pdf

- AI in Healthcare: Deep learning to identify COVID-19 lesions in lung CT scans white paper

  https://www.netapp.com/pdf.html?item=/media/31240-WP-7342.pdf

- AI in Healthcare: Monitoring face mask usage in healthcare settings white paper

  https://www.netapp.com/pdf.html?item=/media/37490-NA-611-Monitoring-face-mask-usage-in-healthcare-settings.pdf

- AI in Healthcare: Diagnostic Imaging Technical Report

  https://www.netapp.com/pdf.html?item=/media/7395-tr4811.pdf

- AI for Retail: NetApp Conversational AI using NVIDIA RIVA

  Executive Summary

- NetApp ONTAP AI solution brief

  https://www.netapp.com/pdf.html?item=/media/6736-sb-3939.pdf
- NetApp DataOps Toolkit solution brief

  https://www.netapp.com/pdf.html?item=/media/21480-SB-4111-1220-NA-Data-Science-Toolkit.pdf
- NetApp AI Control Plane solution brief

  https://www.netapp.com/pdf.html?item=/media/6737-sb-4055.pdf
- Transforming Industry with Data Drive AI eBook

  https://www.netapp.com/us/media/na-337.pdf
- NetApp EF-Series AI solution brief

  https://www.netapp.com/pdf.html?item=/media/26708-SB-4136-NetApp-AI-E-Series.pdf
- NetApp AI and Lenovo ThinkSystem for AI Inferencing solution brief

  https://www.netapp.com/pdf.html?item=/media/25316-SB-4129.pdf
- NetApp AI and Lenovo ThinkSystem for enterprise AI and ML solution brief

  https://www.netapp.com/pdf.html?item=/media/25317-SB-4128.pdf
- NetApp and NVIDIA – Redefining What is Possible with AI video

  https://www.youtube.com/watch?v=38xw65SteUc

Copyright information

Trademark information