Análisis de la confianza con IA de NetApp

NetApp Solutions

NetApp
April 25, 2024

Tabla de contenidos

Análisis de la confianza con IA de NetApp

TR-4910: Análisis de sentimiento en las comunicaciones de los clientes con IA de NetApp

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

Este informe técnico proporciona directrices de diseño para que los clientes realicen análisis de opinión en un centro de soporte global de nivel empresarial usando las tecnologías de gestión de datos de NetApp con un marco de software de NVIDIA mediante la transferencia de aprendizaje e IA conversacional. Esta solución es aplicable a cualquier sector que desee obtener información de los clientes a partir de archivos de voz o texto grabados que representen registros de chat, correos electrónicos y otras comunicaciones de texto o audio. Hemos implementado una canalización integral para demostrar el reconocimiento automático de voz, el análisis de sensibilidad en tiempo real y el modelo de procesamiento de lenguaje natural y aprendizaje profundo: Funcionalidades de nueva formación en un clúster informático acelerado por GPU con almacenamiento all-flash conectado al cloud de NetApp. Se pueden formar y optimizar modelos lingüísticos masivos de vanguardia para realizar inferencia rápidamente con el centro de soporte global y crear una experiencia de cliente excepcional y evaluaciones de rendimiento de empleados objetivas y a largo plazo.

El análisis del sentimiento es un campo de estudio dentro del procesamiento del lenguaje natural (NLP) por el cual se extraen sentimientos positivos, negativos o neutrales del texto. Los sistemas de IA conversacionales han aumentado hasta alcanzar un nivel de integración casi global a medida que cada vez más personas acuden para interactuar con ellos. El análisis de confianza tiene una variedad de casos de uso, desde determinar el rendimiento de los empleados del centro de soporte en conversaciones con los llamantes y proporcionar respuestas automatizadas adecuadas al bot conversacional para predecir el precio de las acciones de una empresa basándose en las interacciones entre representantes de la empresa y la audiencia en llamadas trimestrales a ganancias. Además, el análisis de confianza puede utilizarse para determinar la visión del cliente sobre los productos, servicios o asistencia proporcionados por la Marca.

Esta solución integral utiliza modelos NLP para realizar un análisis de confianza de alto nivel que permita marcos analíticos del centro de soporte. Las grabaciones de audio se procesan en texto escrito y el sentimiento se extrae de cada frase de la conversación. Los resultados, agregados en un panel, se pueden elaborar para analizar los sentimientos de la conversación, tanto históricos como en tiempo real. Esta solución se puede generalizar a otras soluciones con modalidades de datos y necesidades de producción similares. Con los datos adecuados se pueden llevar a cabo otros casos de uso. Por ejemplo, las llamadas de beneficios de la empresa se pueden analizar para determinar la opinión utilizando la misma canalización de extremo a extremo. Otras formas de análisis de NLP, como el modelado de temas y el reconocimiento de entidades con nombre (TNM), también son posibles debido a la naturaleza flexible del ducto.

Estas implementaciones de IA fueron posibles gracias a NVIDIA RIVA, el kit de herramientas TAO de NVIDIA y el kit de herramientas DataOPS de NetApp trabajando juntos. Las herramientas de NVIDIA se utilizan para poner en marcha rápidamente soluciones de IA de gran rendimiento utilizando modelos y canalizaciones prediseñados. El kit de herramientas DataOPS de NetApp simplifica diversas tareas de gestión de datos para acelerar el desarrollo.

Valor para el cliente

Las empresas ven el valor de una herramienta de evaluación de empleados y de reacción del cliente para la conversación de texto, audio y vídeo para el análisis de opiniones. Los gerentes se benefician de la información presentada en el panel, permitiendo una evaluación de los empleados y la satisfacción del cliente basada en ambos lados de la conversación.

Además, el kit de herramientas DataOps de NetApp gestiona el versionado y la asignación de datos dentro de la infraestructura del cliente. Esto lleva a frecuentes actualizaciones de los análisis presentados en el panel de control sin crear costes rígidos de almacenamiento de datos.

Casos de uso

Debido al número de llamadas que procesa estos centros de soporte, la evaluación del rendimiento de la llamada puede llevar un tiempo considerable si se realiza manualmente. Los métodos tradicionales, como el conteo de bolsas de palabras y otros métodos, pueden lograr cierta automatización, pero estos métodos no capturan aspectos más matizados y un contexto semántico del lenguaje dinámico. Las técnicas de modelado de IA se pueden usar para realizar algunos de estos análisis más matices de forma automatizada. Además, con lo último en tecnología y herramientas de modelado preformadas publicadas por NVIDIA, AWS, Google y otras, puede ponerse en marcha una canalización integral con modelos complejos con una facilidad relativa personalizada.

Una canalización de extremo a extremo para el análisis de confianza del centro de soporte procesa archivos de audio en tiempo real a medida que los empleados converse con las personas que llaman. A continuación, estos archivos de audio se procesan para su uso en el componente voz a texto que los convierte en un formato de texto. Cada frase de la conversación recibe una etiqueta que indica el sentimiento (positivo, negativo o neutro).

El análisis de confianza puede proporcionar un aspecto esencial de las conversaciones para evaluar el rendimiento de la llamada. Estos sentimientos añaden un nivel adicional de profundidad a las interacciones entre empleados y personas que llaman. El panel de control de confianza asistido por IA proporciona a los administradores un seguimiento en tiempo real de la opinión dentro de una conversación, junto con un análisis retrospectivo de las llamadas pasadas del empleado.

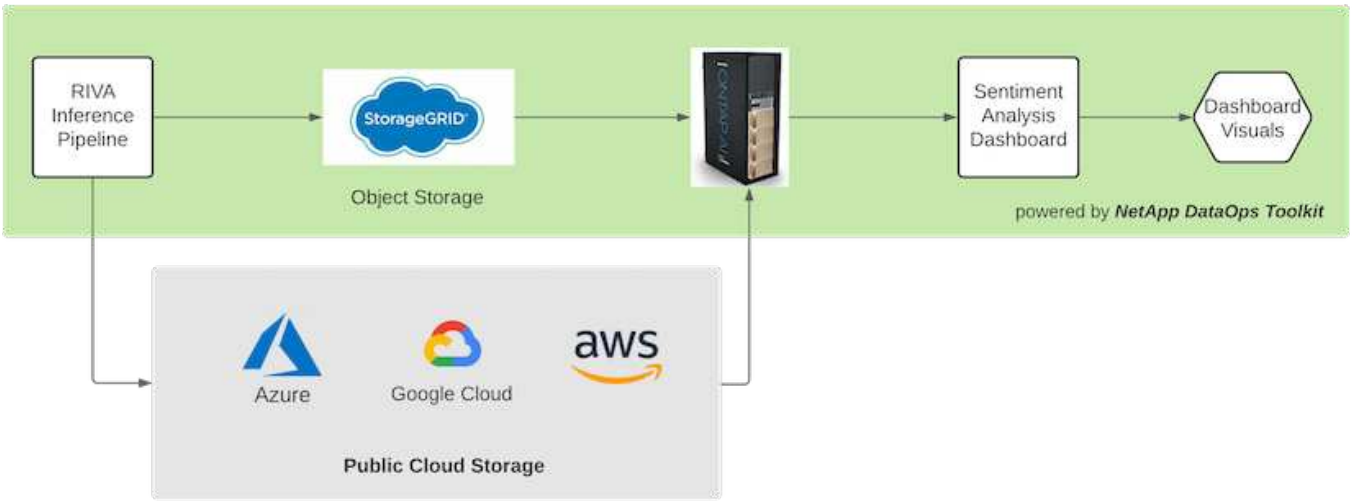
Existen herramientas prediseñadas que se pueden combinar de manera potente para crear rápidamente una canalización de IA completa para solucionar este problema. En este caso, la biblioteca NVIDIA RIVA se puede utilizar para realizar las dos tareas en serie: Transcripción de audio y análisis de sentimiento. El primero es un algoritmo de procesamiento de señales de aprendizaje supervisado y el segundo es un algoritmo de clasificación NLP de aprendizaje supervisado. Estos algoritmos listos para usar pueden ajustarse para cualquier caso de uso relevante con datos relevantes del negocio mediante el kit de herramientas TAO de NVIDIA. Esto lleva a que se estén creando soluciones más precisas y potentes por una fracción del coste y los recursos. Los clientes pueden incorporar la "Maxine de NVIDIA" Marco de trabajo para aplicaciones de videoconferencia aceleradas por GPU en el diseño del centro de soporte.

Los siguientes casos de uso son el núcleo de esta solución. En ambos casos de uso, se utiliza TAO Toolkit para realizar un ajuste preciso del modelo y RIVA para su implementación.

- Voz a texto
- Análisis de la confianza

Para analizar las interacciones del centro de soporte entre empleados y clientes, cada conversación del cliente en forma de llamadas de audio se puede realizar a través de la canalización para extraer sentimientos a nivel de frase. Esos sentimientos pueden ser verificados por un ser humano para justificar los sentimientos o ajustarlos según sea necesario. A continuación, los datos etiquetados se pasan al paso de ajuste fino para mejorar las predicciones de sentimientos. Si ya existen datos de confianza etiquetados, se puede acelerar el ajuste preciso del modelo. En cualquier caso, el ducto es generalizable a otras soluciones que requieren la ingestión de audio y la clasificación de frases.

Los resultados de opiniones de la IA se cargan en una base de datos cloud externa o en un sistema de almacenamiento gestionado por la empresa. Los resultados de la confianza se transfieren desde esta base de datos más grande al almacenamiento local para su uso dentro del panel de control que muestra el análisis de opinión de los administradores. La funcionalidad principal del panel es la de interactuar con el empleado del servicio de atención al cliente en tiempo real. Los gerentes pueden evaluar y proporcionar comentarios a los empleados durante sus llamadas con actualizaciones en vivo de la opinión de cada frase, así como una revisión histórica del desempeño pasado del empleado o de las reacciones del cliente.



La "Kit de herramientas de operaciones de datos de NetApp" Puede seguir gestionando sistemas de almacenamiento de datos incluso después de que la canalización de inferencia RIVA genere etiquetas de sentimiento. Estos resultados de la IA pueden cargarse en un sistema de almacenamiento de datos gestionado por el kit de herramientas DataOPS de NetApp. Los sistemas de almacenamiento de datos deben ser capaces de gestionar cientos de inserciones y seleccionar cada minuto. El sistema de almacenamiento del dispositivo local consulta el mayor almacenamiento de datos en tiempo real para su extracción. También es posible consultar a la instancia de almacenamiento de datos de mayor tamaño los datos históricos para mejorar aún más la experiencia de la consola. El kit de herramientas DataOPS de NetApp facilita ambos usos mediante el clonado rápido de los datos y la distribución entre todas las consolas que los utilizan.

Público objetivo

El público objetivo de la solución incluye los siguientes grupos:

- Gerentes de empleados
- Ingenieros/científicos de datos
- Administradores DE TECNOLOGÍA (en las instalaciones, en cloud o híbrida)

El seguimiento de los sentimientos a lo largo de las conversaciones es una herramienta valiosa para evaluar el rendimiento de los empleados. Mediante el panel de instrumentos de inteligencia artificial, los administradores pueden ver cómo los empleados y las personas que llaman cambian sus sentimientos en tiempo real, lo que permite evaluaciones en directo y sesiones de orientación. Además, las empresas pueden obtener información valiosa de los clientes que participan en conversaciones vocales, bots de chat de texto y videoconferencias. Dichos análisis de clientes utilizan las funcionalidades de procesamiento multimodal a escala con modernos modelos de IA y flujos de trabajo de vanguardia.

En cuanto a los datos, el centro de soporte procesa diariamente un gran número de archivos de audio. El kit de herramientas DataOPS de NetApp facilita esta tarea de gestión de datos tanto para la sintonización periódica de modelos como para consolas de análisis de opiniones.

Los administradores DE TECNOLOGÍA también se benefician del kit de herramientas de NetApp DataOps ya que permite mover datos con rapidez entre entornos de implementación y de producción. Los entornos y servidores de NVIDIA también deben gestionarse y distribuirse para permitir la inferencia en tiempo real.

Arquitectura

La arquitectura de esta solución de centro de soporte se basa en las herramientas prediseñadas de NVIDIA y el kit de herramientas Data OPS de NetApp. Las herramientas de NVIDIA se utilizan para poner en marcha rápidamente soluciones de IA de alto rendimiento utilizando modelos y canalizaciones prediseñados. El kit de herramientas DataOPS de NetApp simplifica diversas tareas de gestión de datos para acelerar el desarrollo.

Tecnología de soluciones

"RIVA DE NVIDIA" Es un SDK acelerado por GPU para crear aplicaciones de IA conversacionales multimodales que ofrecen rendimiento en tiempo real en las GPU. El kit de herramientas NVIDIA Train, Adapt y Optimize (TAO) proporciona un método más rápido y sencillo para acelerar la formación y crear rápidamente modelos de IA específicos para dominios con un alto rendimiento y precisión.

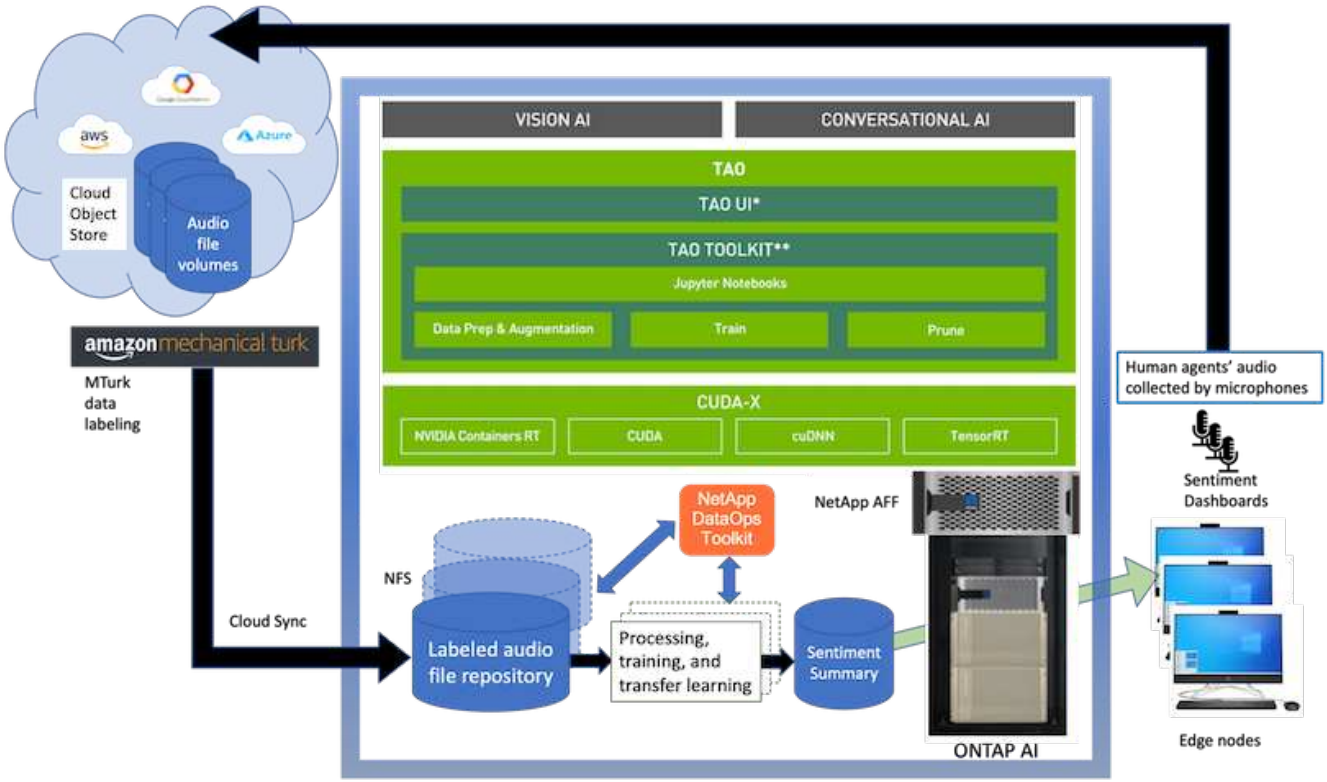
El kit de herramientas NetApp DataOPS es una biblioteca Python que facilita a los desarrolladores, científicos de datos, ingenieros de DevOps e ingenieros de datos la tarea de realizar varias tareas de gestión de datos. Esto incluye el aprovisionamiento casi instantáneo de un nuevo volumen de datos o espacio de trabajo JupyterLab, el clonado casi instantáneo de un volumen de datos o un espacio de trabajo JupyterLab y la creación casi instantánea de copias Snapshot de un volumen de datos o de un espacio de trabajo JupyterLab para su seguimiento y línea de base.

Diagrama arquitectónico

El siguiente diagrama muestra la arquitectura de la solución. Existen tres categorías de entorno principales: El cloud, el núcleo y el perímetro. Cada una de las categorías puede estar geográficamente dispersa. Por ejemplo, la nube contiene almacenes de objetos con archivos de audio en bloques de diferentes regiones, mientras que el núcleo puede contener centros de datos vinculados a través de una red de alta velocidad o Copia y sincronización de NetApp BlueXP. Los nodos EDGE denotan las plataformas de trabajo diario del agente humano individual, en las que hay disponibles micrófonos y herramientas de panel interactivas para visualizar la opinión y recopilar datos de audio de conversaciones con los clientes.

En los centros de datos acelerados por GPU, las empresas pueden utilizar NVIDIA "RIVA" Marco para crear aplicaciones de IA conversacionales, con las que el "Kit de herramientas Tao" Se conecta para la finetunización de modelos y el reciclaje mediante técnicas de aprendizaje en L de transferencia. Estas aplicaciones y flujos de trabajo de computación funcionan con la "Kit de herramientas de operaciones de datos de NetApp", Activación de las mejores capacidades de administración de datos que ONTAP tiene para ofrecer. El kit de herramientas permite a los equipos de datos corporativos realizar rápidamente prototipos de sus modelos con datos estructurados y no estructurados asociados a través de copias Snapshot y clones para llevar a cabo seguimientos, crear versiones, realizar pruebas A/B, proporcionando así seguridad, gobernabilidad, y cumplimiento de las normativas. Consulte la sección "Diseño del almacenamiento" para obtener más detalles.

Esta solución muestra los pasos detallados del procesamiento de archivos de audio, el entrenamiento de modelos NLP, el aprendizaje de transferencias y la administración de datos. La canalización integral resultante genera un resumen de sentimiento que se muestra en tiempo real en los paneles de control de los agentes de apoyo humano.



Requisitos de hardware

En la siguiente tabla se enumeran los componentes de hardware necesarios para implementar la solución. Los componentes de hardware que se usan en cualquier implementación particular de la solución pueden variar en función de las necesidades del cliente.

Pruebas de latencia de respuesta	Tiempo (milisegundos)
Procesamiento de datos	10
Inferencia	10

Estas pruebas de tiempo de respuesta se realizaron en más de 50,000 archivos de audio en 560 conversaciones. Cada archivo de audio tenía un tamaño de ~100 KB como MP3 y de ~1 MB cuando se convirtió a WAV. El paso de procesamiento de datos convierte MP3s en archivos WAV. Los pasos de inferencia convierten los archivos de audio en texto y extraen un sentimiento del texto. Estos pasos son todos independientes entre sí y pueden ser paralelizados para acelerar el proceso.

Teniendo en cuenta la latencia de transferencia de datos entre almacenes, los administradores deben poder ver actualizaciones del análisis de opinión en tiempo real en un segundo del final de la frase.

Hardware NVIDIA RIVA

Hardware subyacente	Requisitos
SO	Linux x86_64
Memoria GPU (ASR)	Modelos de streaming: Aprox. 5600 MB modelos sin secuencias: Aprox. 3100 MB
Memoria GPU (NLP)	~500 MB por modelo BERT

Hardware del kit de herramientas TAO de NVIDIA

Hardware subyacente	Requisitos
RAM del sistema	32 GB
RAM DE GPU	32 GB
CPU	8 núcleos
GPU	NVIDIA (A100, V100 y RTX 30x0)
SSD	100 GB

Sistema de almacenamiento flash

ONTAP 9 de NetApp

ONTAP 9.9, la última generación del software de gestión del almacenamiento de NetApp, permite a las empresas modernizar su infraestructura y realizar la transición a un centro de datos preparado para el cloud. ONTAP ofrece las mejores capacidades de gestión de datos y permite la gestión y protección de los datos con un solo conjunto de herramientas, sin importar dónde residan. También puede mover los datos libremente a donde sea necesario: El perímetro, el núcleo o el cloud. ONTAP 9.9 incluye numerosas funciones que simplifican la gestión de datos, aceleran y protegen los datos esenciales y permiten disfrutar de funcionalidades de infraestructura de nueva generación en arquitecturas de cloud híbrido.

Copia y sincronización de NetApp BlueXP

"Copia y sincronización de BlueXP" Es un servicio de NetApp que ofrece una sincronización de datos rápida y segura que permite transferir archivos entre recursos compartidos de archivos NFS o SMB en las instalaciones a cualquiera de los siguientes destinos:

- StorageGRID de NetApp
- ONTAP S3 de NetApp
- Cloud Volumes Service de NetApp
- Azure NetApp Files
- Simple Storage Service (Amazon S3)
- Sistema de archivos Elastic de Amazon (Amazon EFS)
- Azure Blob
- Google Cloud Storage
- Almacenamiento de objetos en cloud de IBM

BlueXP Copy and Sync mueve los archivos a donde los necesitas de una forma rápida y segura. Una vez transferidos los datos, estarán completamente disponibles para su uso tanto en el origen como en el destino. BlueXP Copy and Sync sincroniza los datos de forma continua en función de tu programación predefinida, moviendo solo los deltas, por lo que se reducen al mínimo el tiempo y el dinero que se invierten en la replicación de datos. BlueXP Copy and Sync es una herramienta de software como servicio (SaaS) fácil de configurar y usar. Las transferencias de datos activadas por BlueXP Copy and Sync se llevan a cabo por agentes de datos. Puedes poner en marcha agentes de datos de BlueXP Copy y Sync en AWS, Azure, Google Cloud Platform o en las instalaciones.

StorageGRID de NetApp

La suite de almacenamiento de objetos definida por software StorageGRID admite una amplia gama de casos de uso en entornos multicloud públicos, privados e híbridos sin problemas. Con innovaciones líderes del sector, StorageGRID de NetApp almacena, protege y preserva datos no estructurados para usos múltiples, incluida la gestión automatizada del ciclo de vida durante largos periodos de tiempo. Para obtener más información, consulte "StorageGRID de NetApp" sitio.

Requisitos de software

En la siguiente tabla se enumeran los componentes de software necesarios para implementar esta solución. Los componentes que se usan en cualquier implementación particular de la solución pueden variar en función de las necesidades del cliente.

Máquina host	Requisitos
RIVA (anteriormente JARVIS)	1.4.0
TAO Toolkit (antes Transfer Learning Toolkit)	3.0
ONTAP	9.9.1
SO DGX	5.1
DOTK	2.0.0

Software NVIDIA RIVA

De NetApp	Requisitos
Docker	>19.02 (con nvidia-docker instalado)>=19.03 si no se utiliza DGX
Controlador NVIDIA	Más de 465.19.01 418.40+, 440.33+, 450.51+ y 460.27+ para las GPU del centro de datos
So del contenedor	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
Servidor de inferencia Triton	2.9.0

Software NVIDIA TAO Toolkit

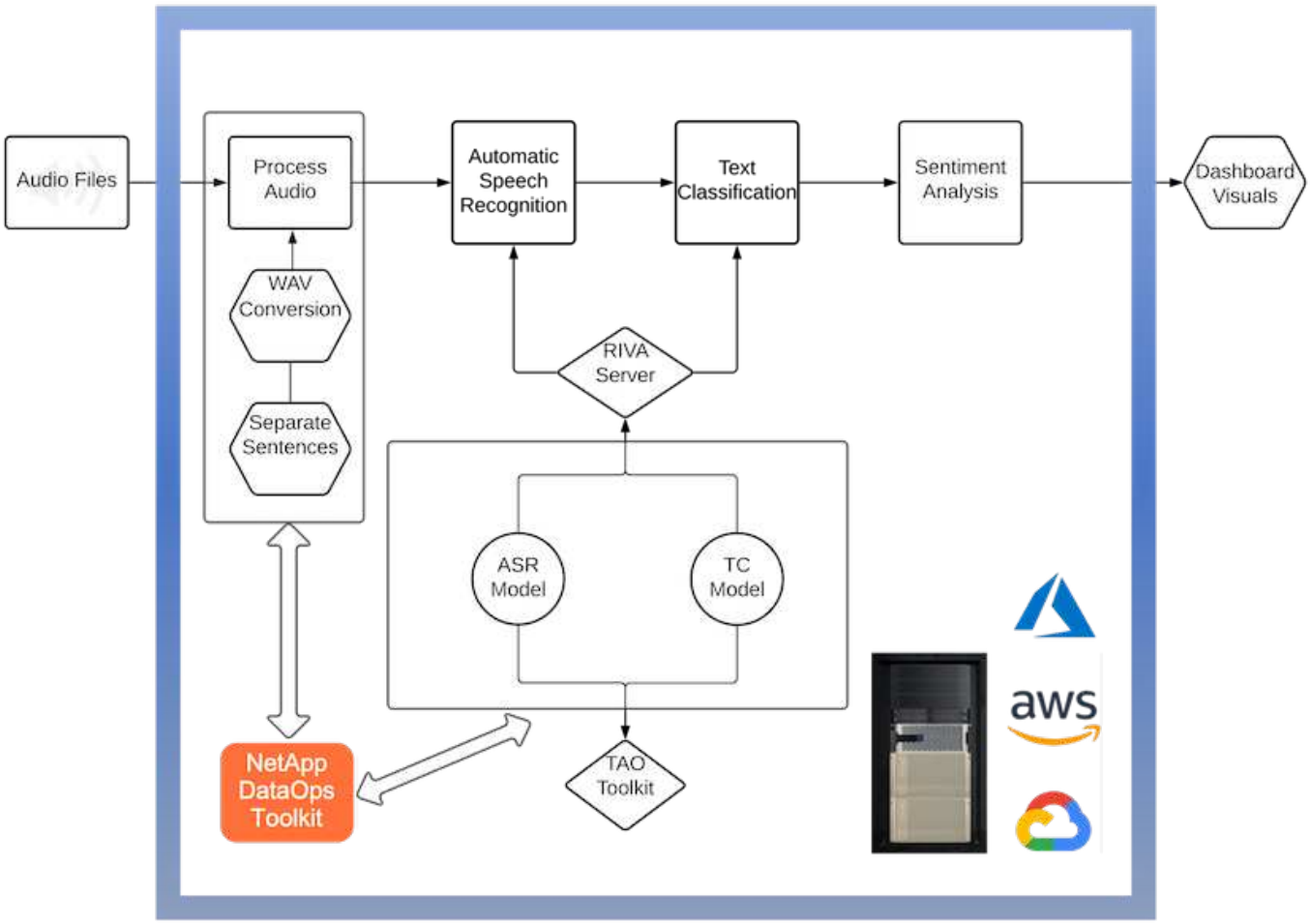
De NetApp	Requisitos
Sistema operativo Ubuntu 18.04 LTS	18.04
python	>=3.6.9
docker-ce	>19.03.5
docker-API	1.40

De NetApp	Requisitos
kit de herramientas de nvidia-container	>1.3.0-1
nvidia-container-runtime	3.4.0-1
nvidia-docker2	2.5.0-1
controlador nvidia	>455
python-pip	>21.06
nvidia-pyindex	Última versión

Detalles de casos de uso

Esta solución se aplica a los siguientes casos de uso:

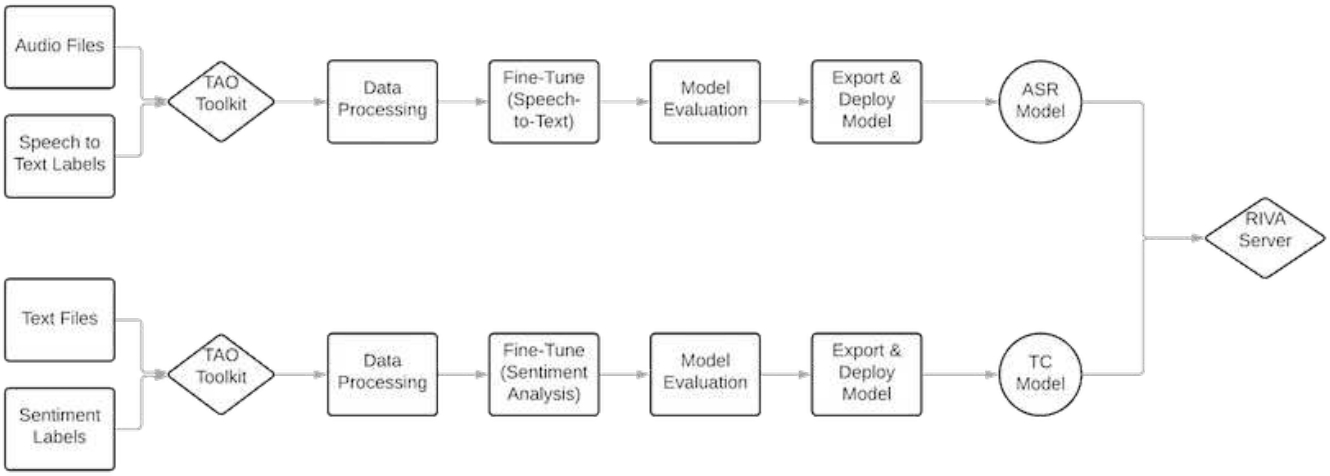
- Voz a texto
- Análisis de la confianza



El caso práctico de voz a texto comienza ingiriendo archivos de audio para los centros de soporte. Este audio se procesa entonces para ajustarse a la estructura requerida por RIVA. Si los archivos de audio aún no se han dividido en sus unidades de análisis, esto se debe hacer antes de pasar el audio a RIVA. Una vez procesado el archivo de audio, se pasa al servidor RIVA como una llamada API. El servidor emplea uno de los muchos modelos que aloja y devuelve una respuesta. Este mensaje de voz a texto (parte del reconocimiento

automático de voz) devuelve una representación de texto del audio. A partir de ahí, la canalización pasa a la parte del análisis de confianza.

Para el análisis de confianza, la salida de texto del reconocimiento automático de voz sirve como entrada a la clasificación de texto. Text Classification es el componente NVIDIA para clasificar el texto en cualquier número de categorías. Las categorías de sentimiento varían de positivo a negativo para las conversaciones del centro de apoyo. El rendimiento de los modelos se puede evaluar utilizando un conjunto de holdout para determinar el éxito del paso de ajuste fino.



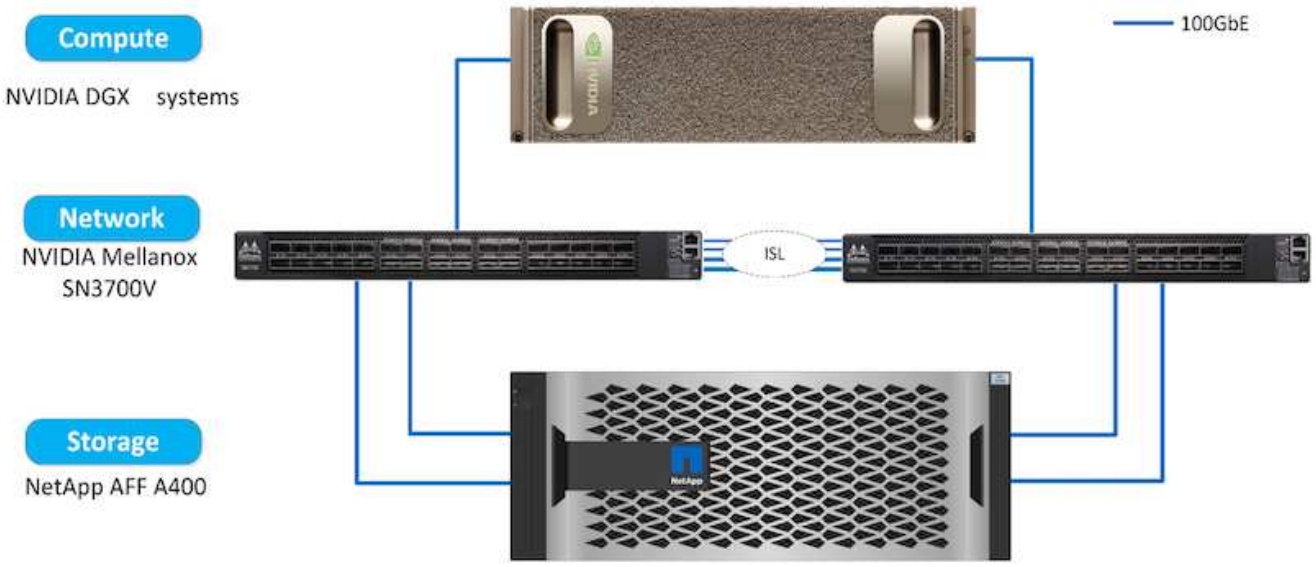
Una canalización similar se utiliza tanto para el análisis de voz a texto como para el análisis de sentimiento dentro del TAO Toolkit. La principal diferencia es el uso de etiquetas que se requieren para la afinación de los modelos. La canalización del kit de herramientas TAO comienza con el procesamiento de los archivos de datos. A continuación, los modelos preformados (procedentes de "Catálogo de NVIDIA NGC") se ajustan con precisión mediante los datos del centro de soporte. Los modelos ajustados con precisión se evalúan en función de sus mediciones de rendimiento correspondientes y, si tienen un rendimiento superior al de los modelos preformados, se implementan en EL servidor RIVA.

Consideraciones de diseño

En esta sección se describen las consideraciones de diseño para los distintos componentes de esta solución.

Diseño informático y de red

En función de las restricciones en la seguridad de los datos, todos los datos deben permanecer dentro de la infraestructura del cliente o en un entorno seguro.



Diseño del almacenamiento

El kit de herramientas Data OPS de NetApp sirve como servicio principal para gestionar los sistemas de almacenamiento. El kit de herramientas DataOps es una biblioteca Python que facilita a los desarrolladores, científicos de datos, ingenieros de DevOps e ingenieros de datos la tarea de gestión de datos realizar varias tareas de gestión de datos, como el aprovisionamiento casi instantáneo de un nuevo volumen de datos o un espacio de trabajo JJupyterLab, el clonado casi instantáneo de un volumen de datos o el espacio de trabajo JuppyterLab. Y una copia de Snapshot casi instantánea de un volumen de datos o un espacio de trabajo JupyterLab para poder seguir su seguimiento o crear una línea de base. Esta biblioteca de Python puede funcionar como una utilidad de línea de comandos o una biblioteca de funciones que se pueden importar a cualquier programa de Python o a cualquier ordenador portátil Jupyter.

Mejores prácticas de RIVA

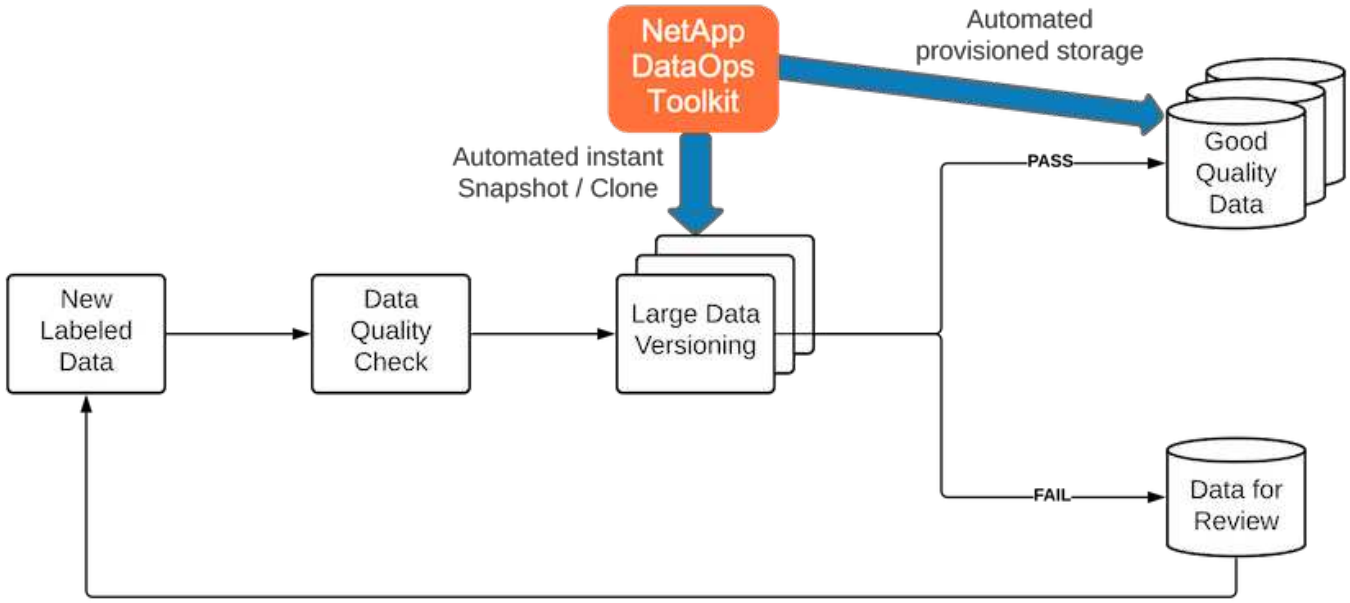
NVIDIA proporciona varias funciones generales "mejores prácticas de datos" Para utilizar RIVA:

- **Utilice formatos de audio sin pérdidas si es posible.** el uso de códecs con pérdida como MP3 puede reducir la calidad.
- **Aumentar los datos de entrenamiento.** la adición de ruido de fondo a los datos de entrenamiento de audio puede inicialmente disminuir la precisión y, sin embargo, aumentar la solidez.
- **Limite el tamaño del vocabulario si utiliza texto raspado.** muchas fuentes en línea contienen tipopos o pronombres auxiliares y palabras poco comunes. La eliminación de estos elementos puede mejorar el modelo de idioma.
- **Utilice una frecuencia de muestreo mínima de 16 kHz si es posible.** sin embargo, trate de no resampling, ya que al hacerlo se reduce la calidad de audio.

Además de estas mejores prácticas, los clientes deben priorizar la recopilación de un conjunto de datos de ejemplo representativo con etiquetas precisas para cada paso de la canalización. En otras palabras, el conjunto de datos de ejemplo debería reflejar proporcionalmente las características especificadas ejemplificadas en un conjunto de datos de destino. De forma similar, los anotadores del conjunto de datos tienen la responsabilidad de equilibrar la precisión y la velocidad del etiquetado, de modo que se maximice la calidad y la cantidad de los datos. Por ejemplo, esta solución de centro de soporte requiere archivos de audio, etiquetas de texto etiquetadas y etiquetas de sentimiento. La naturaleza secuencial de esta solución significa que los errores desde el principio de la canalización se propagan hasta el final Si los archivos de audio son de

mala calidad, las transcripciones de texto y las etiquetas de sentimiento también serán.

Esta propagación de errores también se aplica a los modelos entrenados en estos datos. Si las predicciones de sentimiento son 100% exactas pero el modelo de voz a texto tiene un rendimiento deficiente, entonces la canalización final está limitada por las transcripciones iniciales de audio a texto. Es esencial que los desarrolladores consideren el rendimiento de cada modelo individualmente y como un componente de una canalización mayor. En este caso en particular, el objetivo final es desarrollar una canalización que pueda predecir con precisión el sentimiento. Por lo tanto, la métrica general sobre la cual evaluar la canalización es la precisión de los sentimientos, que afecta directamente la transcripción del habla al texto.



El kit de herramientas DataOPS de NetApp complementa la canalización de comprobación de la calidad de los datos mediante su tecnología de clonado de datos casi instantánea. Cada archivo etiquetado debe evaluarse y compararse con los archivos etiquetados existentes. La distribución de estas comprobaciones de calidad a través de distintos sistemas de almacenamiento de datos garantiza una ejecución rápida y eficaz de estas comprobaciones.

Implementar el análisis de confianza del centro de soporte

La implementación de la solución consta de los siguientes componentes:

1. Kit de herramientas de operaciones de datos de NetApp
2. Configuración de NGC
3. Servidor NVIDIA RIVA
4. Kit de herramientas TAO de NVIDIA
5. Exportar modelos TAO a RIVA

Para realizar la implementación, lleve a cabo los siguientes pasos:

Kit de herramientas Data OPS de NetApp: Compatibilidad con el análisis de confianza del centro de soporte

Para utilizar la "Kit de herramientas de operaciones de datos de NetApp", lleve a cabo los siguientes pasos:

1. Instalación del kit de herramientas de PIP.

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

2. Configurar la gestión de datos

```
netapp_dataops_cli.py config
```

Configuración de NGC: Respaldar el análisis de confianza del centro

Para configurar "NVIDIA NGC", lleve a cabo los siguientes pasos:

1. Descargar el NGC.

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

2. Agregue su directorio actual a la ruta de acceso.

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

3. Debe configurar la CLI de NGC para su uso con el fin de poder ejecutar los comandos. Introduzca el siguiente comando, incluida su clave de API cuando se le solicite.

```
ngc config set
```

Para sistemas operativos que no están basados en Linux, visite "aquí".

NVIDIA RIVA Server: Análisis de opinión del centro de soporte

Para configurar "RIVA DE NVIDIA", lleve a cabo los siguientes pasos:

1. Descargar los archivos RIVA de NGC.

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

2. Inicialice LA configuración DE RIVA (`riva_init.sh`).
3. Inicie EL servidor RIVA (`riva_start.sh`).
4. Inicie EL cliente RIVA (`riva_start_client.sh`).
5. En EL cliente RIVA, instale la biblioteca de procesamiento de audio ("FFMPEG")

```
apt-get install ffmpeg
```

6. Inicie el "Juppyter" servidor.
7. Ejecute el portátil de canalización de inferencia DE RIVA.

Kit de herramientas TAO de NVIDIA: Análisis de opinión del centro de soporte

Para configurar NVIDIA TAO Toolkit, lleve a cabo los siguientes pasos:

1. Prepare y active una "entorno virtual" Para TAO Toolkit.
2. Instale el "paquetes requeridos".
3. Tire manualmente de la imagen utilizada durante el entrenamiento y ajuste preciso.

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

4. Inicie el "Juppyter" servidor.
5. Ejecute el cuaderno TAO de afinación fina.

Exportar modelos TAO a RIVA: Apoyar el análisis de confianza del centro

Para usar "Modelos TAO Toolkit EN RIVA", lleve a cabo los siguientes pasos:

1. Guarde los modelos en el cuaderno TAO de sintonización fina.
2. Copie los modelos TAO entrenados en el directorio del modelo RIVA.
3. Inicie EL servidor RIVA (`riva_start.sh`).

Obstáculos para la implementación

Estas son algunas cosas que debe tener en cuenta a medida que desarrolla su propia solución:

- El kit de herramientas Data OPS de NetApp se instala primero para garantizar que el sistema de almacenamiento de datos se ejecute de forma óptima.
- NVIDIA NGC debe instalarse antes de cualquier otra cosa porque autentica la descarga de imágenes y modelos.

- RIVA se debe instalar antes que TAO Toolkit. LA instalación DE RIVA configura el demonio docker para extraer imágenes según sea necesario.
- El DGX y el docker deben tener acceso a Internet para descargar los modelos.

Resultados de validación

Tal como se ha mencionado en la sección anterior, los errores se propagan por la canalización siempre que haya dos o más modelos de aprendizaje automático ejecutándose de forma secuencial. Para esta solución, el sentimiento de la frase es el factor más importante a la hora de medir el nivel de riesgo de las acciones de la empresa. El modelo de voz a texto, aunque esencial para la canalización, sirve como unidad de preprocesamiento antes de que se puedan predecir los sentimientos. Lo que realmente importa es la diferencia en el sentimiento entre las frases de la verdad en el terreno y las oraciones predichas. Esto sirve como proxy para la velocidad de error de la palabra (WER). La precisión de voz a texto es importante, pero el WER no se utiliza directamente en la métrica de canalización final.

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

Estas métricas de sentimiento se pueden calcular para la puntuación F1, la recuperación y la precisión de cada frase. Los resultados se pueden agregar y mostrar dentro de una matriz de confusión, junto con los intervalos de confianza de cada métrica.

La ventaja de usar el aprendizaje de transferencia es un aumento en el rendimiento de los modelos por una fracción de los requisitos de datos, el tiempo de entrenamiento y el coste. Los modelos ajustados también deben compararse con sus versiones de referencia para garantizar que el aprendizaje de la transferencia mejore el rendimiento en lugar de deteriorarlo. En otras palabras, el modelo ajustado debería tener un mejor rendimiento en los datos del centro de soporte que el modelo preentrenado.

Evaluación de la canalización

Caso de prueba	Detalles
Número de prueba	Métrica de sentimiento de canalización
Probar los requisitos previos	Modelos ajustados para modelos de análisis de voz a texto y de sentimiento
Resultado esperado	La métrica de sentimiento del modelo ajustado funciona mejor que el modelo preentrenado original.

Métrica de sentimiento de canalización

1. Calcule la métrica de sentimiento para el modelo de referencia.
2. Calcule la métrica de sentimiento para el modelo ajustado.
3. Calcular la diferencia entre estas métricas.
4. Calcule la media de las diferencias entre todas las frases.

Vídeos y demostraciones

Existen dos portátiles que contienen la canalización de análisis de confianza: ""“Support-Center-Model-Transfer-Learning-and-Fine-Tuning.ipynb”" y.. ""“Support-Center-sentimiento-Analysis-Pipeline.ipynb”". Juntos, estos portátiles muestran cómo desarrollar una canalización para la incorporación de datos de centro de soporte y extraer sentimientos de cada frase mediante modelos de aprendizaje profundo de última generación ajustados a los datos del usuario.

Centro de apoyo - Análisis de confianza Pipeline.ipynb

Este cuaderno contiene la canalización RIVA de inferencia para la incorporación de audio, la conversión a texto y la extracción de sentimientos para su uso en un panel externo. El conjunto de datos se descarga y procesa automáticamente si aún no se ha hecho. La primera sección del bloc de notas es la voz a texto que controla la conversión de archivos de audio a texto. A continuación se muestra la sección Análisis de sentimientos que extrae sentimientos para cada frase de texto y muestra dichos resultados en un formato similar al panel propuesto.

> Este portátil debe ejecutarse antes del entrenamiento del modelo y del ajuste preciso porque el conjunto de datos MP3 debe descargarse y convertirse al formato correcto.

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

Centro de soporte - formación de modelos y afinación.ipynb

El entorno virtual del kit de herramientas de TAO debe configurarse antes de ejecutar el cuaderno (consulte la sección del kit de herramientas de TAO en la descripción general de comandos para obtener instrucciones de instalación).

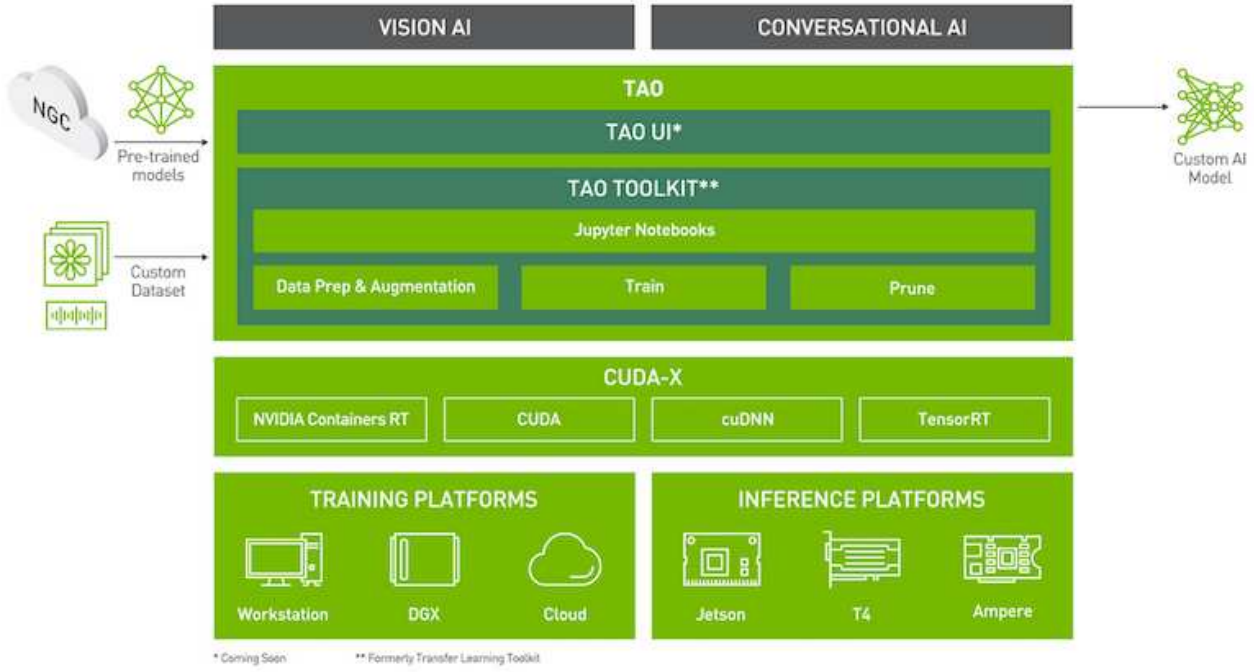
Este cuaderno confía en el kit de herramientas TAO para ajustar los modelos de aprendizaje profundo en los datos de los clientes. Al igual que en el cuaderno anterior, este se separa en dos secciones para los componentes de análisis de opinión y voz. Cada sección abarca el procesamiento de datos, la formación de modelos y el ajuste preciso, la evaluación de los resultados y la exportación de modelos. Por último, existe una sección final para implementar tanto sus modelos ajustados para su uso en 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.

Conclusión

A medida que se considera cada vez más la experiencia de los clientes como un campo de batalla competitivo clave, un centro de soporte global aumentado por IA se convierte en un componente fundamental que las empresas de casi todos los sectores no pueden permitirse el lujo de descuidar. La solución propuesta en este informe técnico ha quedado demostrada para respaldar la prestación de estas experiencias excepcionales a los clientes, y el reto consiste ahora en garantizar que las empresas tomen acciones para modernizar su infraestructura de IA y sus flujos de trabajo.

Las mejores implementaciones de IA en servicio al cliente no consisten en sustituir a los agentes humanos. Más bien, la IA puede empoderarlos para crear experiencias de cliente excepcionales mediante análisis de

sensibilidad en tiempo real, escalado de disputas y computación afectiva multimodal para detectar señales verbales, no verbales y faciales con las que los modelos de IA integrales pueden hacer recomendaciones a escala y complementar lo que podría carecer un agente humano individual. Asimismo, la IA puede proporcionar una mejor unión entre un cliente concreto con los agentes disponibles actualmente. Gracias a la IA, las empresas pueden extraer un valioso sentimiento de los clientes en cuanto a sus pensamientos e impresiones sobre los productos, servicios y la imagen de la Marca del proveedor.

La solución también se puede usar para construir datos de series temporales para que los agentes de soporte sirvan como métrica de evaluación del rendimiento objetivo. Las encuestas convencionales de satisfacción de clientes a menudo no tienen respuestas suficientes. Al recopilar la opinión de los empleados y los clientes a largo plazo, los empleadores pueden tomar decisiones informadas con respecto al desempeño de los agentes de apoyo.

La combinación de NetApp, SFL Scientific, marcos de orquestación de código abierto y NVIDIA reúne las tecnologías más recientes como servicios gestionados con una gran flexibilidad para acelerar la adopción de tecnología y mejorar el plazo de comercialización de las nuevas aplicaciones de IA/ML. Estos servicios avanzados se ofrecen en las instalaciones que se pueden transportar fácilmente para entornos nativos del cloud así como arquitecturas de puesta en marcha híbrida.

Dónde encontrar información adicional

Si quiere más información sobre el contenido de este documento, consulte los siguientes documentos o sitios web:

- Demostraciones interactivas en 3D

  "www.netapp.com/ai"

- Conecte directamente con un especialista en IA de NetApp

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

- Descripción de la solución NVDIA base Command Platform con NetApp

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

- NetApp para IA 10: Infografía buenas razones

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

- IA en la sanidad: Aprendizaje profundo para identificar las lesiones del COVID-19 en exploraciones pulmonares TC

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

- AI en la atención sanitaria: Monitorización del uso de la máscara facial en entornos sanitarios documento técnico

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

- IA en la sanidad: Informe técnico de imágenes de diagnóstico

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

- IA para minoristas: Inteligencia artificial de NetApp conversacional con NVIDIA RIVA

  "https://docs.netapp.com/us-en/netapp-solutions/ai/cainvidia_executive_summary.html"
- Resumen de la solución ONTAP AI de NetApp

  https://www.netapp.com/pdf.html?item=/media/6736-sb-3939.pdf
- Resumen de la solución del kit de herramientas de operaciones de datos de NetApp

  https://www.netapp.com/pdf.html?item=/media/21480-SB-4111-1220-NA-Data-Science-Toolkit.pdf
- Resumen de la solución del plano de control de IA de NetApp

  https://www.netapp.com/pdf.html?item=/media/6737-sb-4055.pdf
- Libro electrónico "transformando el sector con los datos impulsan la IA"

  "https://www.netapp.com/us/media/na-337.pdf"
- Resumen de la solución de IA EF-Series de NetApp

  https://www.netapp.com/pdf.html?item=/media/26708-SB-4136-NetApp-AI-E-Series.pdf
- Resumen de la solución de NetApp AI y Lenovo ThinkSystem para la inferencia de IA

  https://www.netapp.com/pdf.html?item=/media/25316-SB-4129.pdf
- Resumen de la solución NetApp AI y Lenovo ThinkSystem para IA empresarial Y ML

  https://www.netapp.com/pdf.html?item=/media/25317-SB-4128.pdf
- NetApp y NVIDIA: Redefiniendo lo que es posible con el vídeo de IA

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