# Analyse des sentiments avec NetApp ai

NetApp Solutions

NetApp
April 24, 2024

This PDF was generated from https://docs.netapp.com/fr-fr/netapp-solutions/ai/ai-sent-support-center-analytics.html on April 24, 2024. Always check docs.netapp.com for the latest.

# Sommaire

Analyse des sentiments avec NetApp ai . . . 1
Tr-4910 : analyse du sentiment issue des communications avec le client sur NetApp ai . . . 1
Cas d’utilisation . . . 2
Architecture . . . 5
Considérations relatives à la conception . . . 10
Déploiement de l’analyse des sentiments du centre de support . . . 12
Résultats de la validation . . . 14
Vidéos et démonstrations . . . 15
Conclusion . . . 17
Où trouver des informations complémentaires . . . 18

# Analyse des sentiments avec NetApp ai

## Tr-4910 : analyse du sentiment issue des communications avec le client sur NetApp ai

Rick Huang, Satish Thyagarajan et David Arnette, NetApp Diego Sosa-Coba, SFL Scientific

Ce rapport technique fournit des conseils de conception afin que les clients puissent réaliser une analyse des sentiments dans un centre de support global de niveau entreprise en utilisant les technologies de gestion des données NetApp avec un framework logiciel NVIDIA, par le biais de l'apprentissage par transfert et de la discussion sur l'IA. Cette solution s'applique à tout secteur souhaitant obtenir des informations sur les clients à partir de fichiers vocaux ou texte enregistrés représentant des journaux de conversation, des e-mails et d'autres communications audio ou texte. Nous avons mis en œuvre un pipeline de bout en bout afin de démontrer la reconnaissance vocale automatique, l'analyse des sentiments en temps réel et le réentraînement du modèle de traitement du langage naturel du deep learning. Nous avons également mis en place des fonctionnalités de recyclage sur un cluster de calcul accéléré par processeur graphique grâce au stockage 100 % Flash de NetApp connecté au cloud. Des modèles linguistiques de pointe peuvent être entraînés et optimisés pour inférence rapidement avec le centre de support global. Ils assurent ainsi une expérience client exceptionnelle et permettent d'évaluer de manière objective les performances des employés à long terme.

L'analyse des sentiments est un domaine d'étude dans le traitement du langage naturel (NLP) par lequel des sentiments positifs, négatifs ou neutres sont extraits du texte. Les systèmes d'IA conversationnels ont été étendus à un niveau d'intégration quasi global, car de plus en plus de gens viennent à les interagir avec eux. L'analyse des sentiments présente de nombreux cas d'utilisation, allant de la détermination de la performance des employés du centre de support lors de conversations avec les appelants à la fourniture de réponses de chatbot automatisées appropriées à la prévision du prix des actions d'une entreprise en fonction des interactions entre les représentants des entreprises et le public lors d'appels trimestriels sur les bénéfices. En outre, l'analyse de sentiment peut être utilisée pour déterminer l'opinion du client sur les produits, les services ou l'assistance fournis par la marque.

Cette solution de bout en bout utilise des modèles NLP pour effectuer une analyse de confiance de haut niveau qui permet de créer des cadres analytiques de centre de support. Les enregistrements audio sont traités en texte écrit et le sentiment est extrait de chaque phrase de la conversation. Les résultats, regroupés dans un tableau de bord, peuvent être conçus pour analyser les sentiments de conversation, historiquement et en temps réel. Cette solution peut être généralisée à d'autres solutions avec des modalités de données et des besoins de sortie similaires. Grâce aux données appropriées, d'autres cas d'utilisation peuvent être mis en œuvre. Par exemple, les appels sur les bénéfices de l'entreprise peuvent être analysés pour obtenir un sentiment à l'aide du même pipeline de bout en bout. D'autres formes d'analyses du PNL, telles que la modélisation de sujets et la reconnaissance d'entités nommées (NER), sont également possibles en raison de la nature souple du pipeline.

Ces implémentations d'IA ont été rendues possibles par NVIDIA RIVA, le kit NVIDIA TAO et le kit NetApp DataOps. Les outils NVIDIA sont utilisés pour déployer rapidement des solutions d'IA haute performance à l'aide de modèles et de pipelines prédéfinis. Le kit NetApp DataOps simplifie les tâches de gestion des données pour accélérer le développement.

### En valeur pour le client

Les entreprises voient les avantages d'un outil d'évaluation des employés et de réaction des clients pour les conversations texte, audio et vidéo pour l'analyse des sentiments. Les responsables bénéficient des informations présentées dans le tableau de bord, ce qui permet d'évaluer les employés et la satisfaction des clients en fonction des deux côtés de la conversation.

Par ailleurs, le kit NetApp DataOps gère les versions et l'allocation des données au sein de l'infrastructure du client. Cela entraîne des mises à jour fréquentes des analyses présentées au tableau de bord sans générer des coûts de stockage de données qui ne sont pas unwieldy.

## Cas d'utilisation

En raison du nombre d'appels traités par ces centres d'assistance, l'évaluation du rendement des appels peut prendre beaucoup de temps si elle est effectuée manuellement. Les méthodes traditionnelles, telles que le comptage des sacs de mots et d'autres méthodes, peuvent atteindre une certaine automatisation, mais ces méthodes ne prennent pas en compte les aspects plus nuancés et le contexte sémantique du langage dynamique. Les techniques de modélisation d'IA peuvent être utilisées pour effectuer certaines de ces analyses plus nuancées de manière automatisée. De plus, avec les outils de modélisation pré-entraînés et publiés par NVIDIA, AWS, Google et autres, un pipeline de bout en bout avec des modèles complexes peut maintenant être mis en place et personnalisé avec une facilité relative.

Un pipeline de bout en bout pour l'analyse des sentiments des centres de support permet d'analyser les fichiers audio en temps réel lorsque les employés convertisse avec les appelants. Ensuite, ces fichiers audio sont traités pour être utilisés dans le composant parole-texte qui les convertit au format texte. Chaque phrase de la conversation reçoit un libellé indiquant le sentiment (positif, négatif ou neutre).

L'analyse des sentiments peut fournir un aspect essentiel des conversations pour évaluer la performance des appels. Ces sentiments ajoutent une profondeur supplémentaire aux interactions entre les employés et les appelants. Le tableau de bord des sentiments assistés par l'IA fournit aux responsables un suivi en temps réel des sentiments au sein d'une conversation, ainsi qu'une analyse rétrospective des appels passés de l'employé.

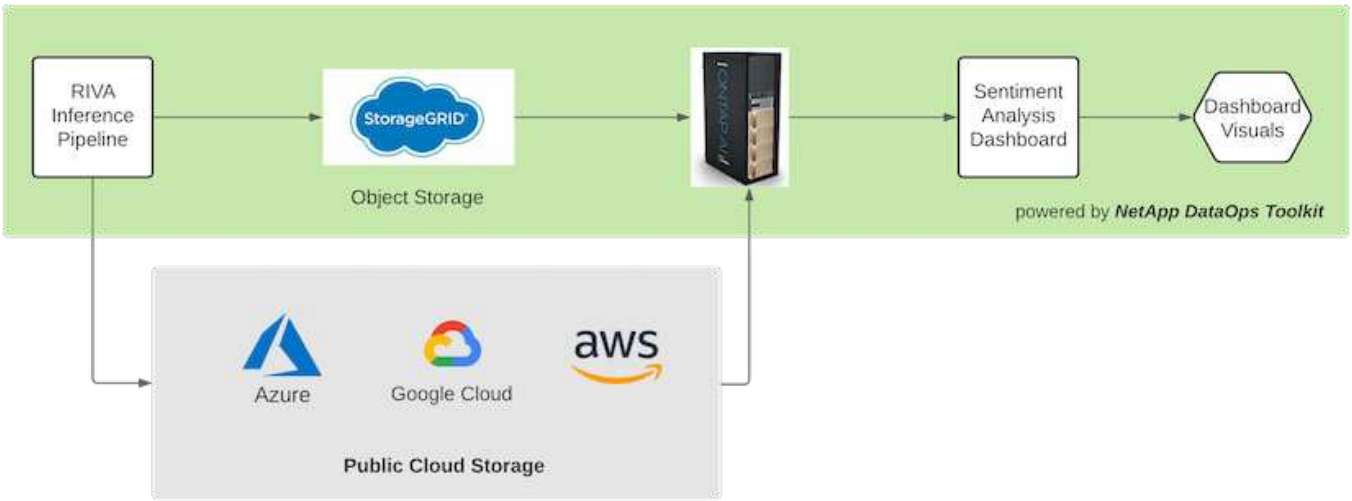
Des outils prédéfinis peuvent être combinés de façon puissante afin de créer rapidement un pipeline d'IA de bout en bout pour résoudre ce problème. Dans ce cas, la bibliothèque NVIDIA RIVA peut être utilisée pour effectuer les deux tâches en série : transcription audio et analyse de sentiment. Le premier est un algorithme de traitement du signal d'apprentissage supervisé et le second est un algorithme de classification NLP d'apprentissage supervisé. Ces algorithmes prêts à l'emploi peuvent être ajustés pour toute utilisation pertinente avec les données pertinentes de l'entreprise à l'aide du kit NVIDIA TAO. Cela entraîne la création de solutions plus précises et plus puissantes pour un coût et des ressources réduits. Les clients peuvent intégrer le "NVIDIA Maxine" Infrastructure pour les applications de visioconférence avec accélération graphique dans leur conception de centre de support.

Les utilisations suivantes sont au cœur de cette solution. Dans les deux cas, TAO Toolkit permet de régler avec précision les modèles et RIVA pour le déploiement de modèles.

- Parole-à-texte
- Analyse des sentiments

Pour analyser les interactions entre les employés et les clients du centre de support, chaque conversation client sous forme d'appels audio peut être exécutée dans le pipeline pour extraire des sentiments de phrase. Ces sentiments peuvent ensuite être vérifiés par un humain pour justifier les sentiments ou les ajuster selon les besoins. Les données étiquetées sont ensuite transmises à l'étape de réglage fin pour améliorer les prédictions de sentiment. Si des données de sentiment étiquetées existent déjà, le réglage fin du modèle peut être accéléré. Dans les deux cas, le pipeline est généralisable à d'autres solutions qui exigent l'ingestion de l'audio et la classification des phrases.

Les résultats des sentiments d'IA sont soit téléchargés vers une base de données cloud externe, soit vers un système de stockage géré par l'entreprise. Les sorties de sentiment sont transférées de cette base de données plus grande vers le stockage local pour une utilisation dans le tableau de bord qui affiche l'analyse de sentiment pour les gestionnaires. La principale fonctionnalité du tableau de bord consiste à interagir avec l'employé du service client en temps réel. Les responsables peuvent évaluer et fournir des commentaires sur les employés lors de leurs appels avec des mises à jour en direct du sentiment de chaque phrase, ainsi qu'un examen historique des performances passées de l'employé ou des réactions des clients.




Le "Kit NetApp DataOps" Peut continuer à gérer les systèmes de stockage des données même après que le pipeline d'inférence RIVA génère des étiquettes de sentiments. Les résultats de cette IA peuvent être téléchargés vers un système de stockage des données géré par le kit NetApp DataOps. Les systèmes de stockage des données doivent être capables de gérer des centaines d'insertions et de sélectionner chaque minute. Le système de stockage de périphériques local interroge le stockage de données plus important en temps réel pour l'extraction. Le tableau de bord peut également être interrogé sur ses données historiques afin d'optimiser l'expérience qu'il fournit. Le kit NetApp DataOps facilite ces deux utilisations en clonant rapidement les données et en les distribuant dans tous les tableaux de bord.

## Public visé

La solution cible plusieurs groupes d'utilisateurs :

- Responsables des employés
- Ingénieurs/data Scientists
- Administrateurs IT (sur site, dans le cloud ou hybride)

Le suivi des sentiments tout au long des conversations est un outil précieux pour évaluer les performances des employés. Grâce au tableau de bord ai, les responsables peuvent voir comment les employés et les appelants changent leurs sentiments en temps réel, ce qui permet des évaluations en direct et des sessions de guidage. De plus, les entreprises peuvent bénéficier de précieux avis de la part de clients engagés dans des conversations vocales, des chatbots de texte et des visioconférences. Cette analytique client exploite les fonctionnalités de traitement multimodal à grande échelle avec des modèles et des workflows d'IA modernes et à la pointe de la technologie.

En ce qui concerne les données, un grand nombre de fichiers audio sont traités quotidiennement par le centre d'assistance. Le kit NetApp DataOps facilite le traitement des données pour affiner régulièrement les modèles et les tableaux de bord d'analyse de sentiment.

Les administrateurs INFORMATIQUES profitent également du kit NetApp DataOps car il leur permet de déplacer rapidement des données entre les environnements de déploiement et de production. Les environnements et les serveurs NVIDIA doivent également être gérés et distribués pour permettre l'inférence en temps réel.

# Architecture

L'architecture de cette solution de centre de support s'articule autour des outils préfabriqués de NVIDIA et du kit NetApp DataOps. Les outils NVIDIA sont utilisés pour déployer rapidement des solutions d'IA haute performance à l'aide de modèles et de pipelines prédéfinis. Le kit NetApp DataOps simplifie les tâches de gestion des données pour accélérer le développement.

## Technologie de la solution

"NVIDIA RIVA" Est un kit de développement accéléré par processeur graphique pour concevoir des applications d'IA multimodales qui fournissent des performances en temps réel sur les GPU. Le kit NVIDIA train, Adapt, and Optimize (TAO) offre un moyen plus rapide et plus simple d'accélérer l'entraînement et de créer rapidement des modèles d'IA spécifiques à un domaine, précis et performants.

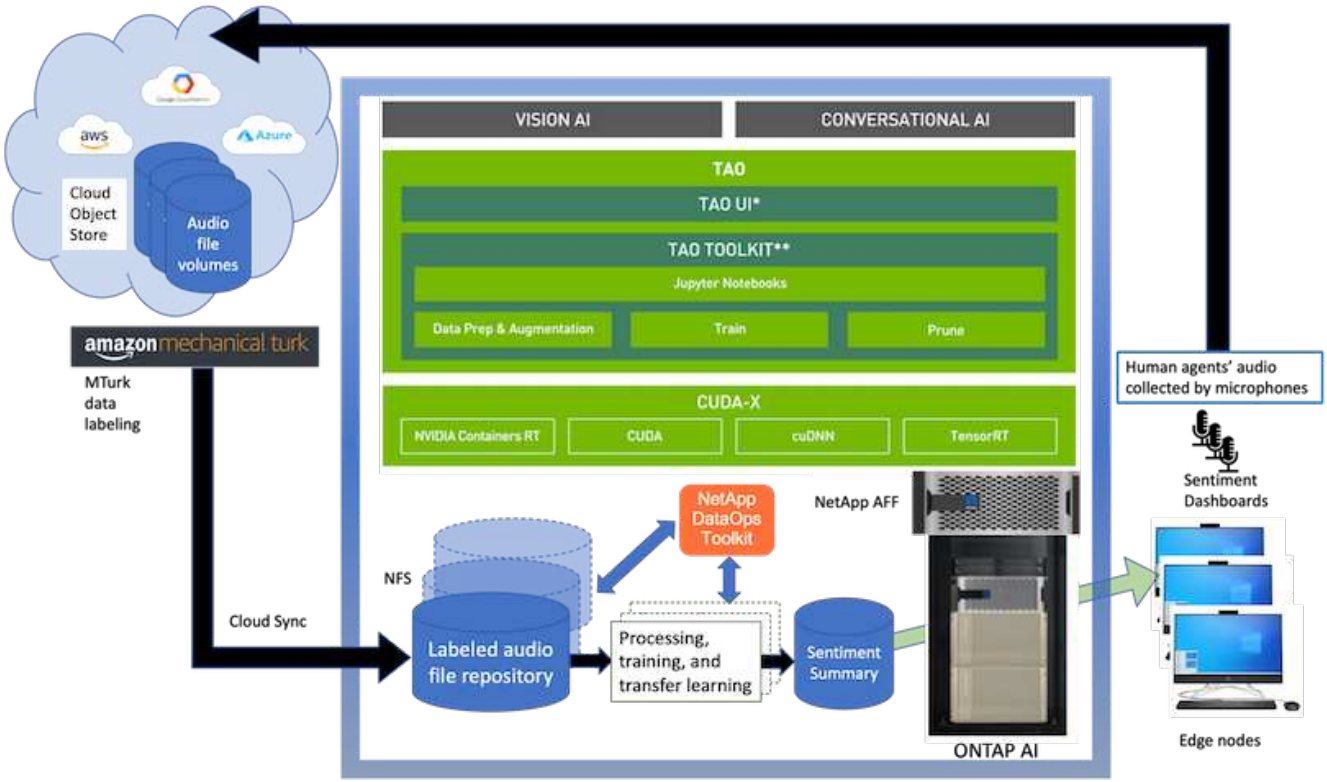
Le kit NetApp DataOps est une bibliothèque Python qui simplifie la tâche des développeurs, des data Scientists, des ingénieurs DevOps et des ingénieurs en gestion des données. Cela inclut le provisionnement quasi instantané d'un nouveau volume de données ou d'un nouvel espace de travail JupyterLab, le clonage quasi instantané d'un volume de données ou d'un espace de travail JupyterLab, ainsi que le clonage quasi instantané d'un volume de données ou d'un espace de travail JupyterLab à des fins de traçabilité et de base.

## Diagramme architectural

Le schéma suivant illustre l'architecture de la solution. Il existe trois catégories d'environnement principales : le cloud, le cœur et la périphérie. Chacune des catégories peut être géographiquement dispersée. Par exemple, le cloud contient des magasins d'objets avec des fichiers audio dans des compartiments de différentes régions, tandis que le cœur peut contenir des data centers reliés par un réseau haut débit ou par copie et synchronisation NetApp BlueXP. Les nœuds périphériques indiquent les plates-formes de travail quotidiennes de chaque agent humain, où des microphones et des outils de tableau de bord interactifs sont disponibles pour visualiser le sentiment et collecter des données audio à partir de conversations avec les clients.

Dans les data centers accélérés par GPU, les entreprises peuvent utiliser NVIDIA "RIVA" Framework pour créer des applications d'IA conversationnelles, vers lesquelles le "Tao Toolkit" Se connecte pour le finetuning de modèles et le recyclage à l'aide de techniques d'apprentissage par transfert de L. Ces workflows et applications de calcul sont optimisés par le "Kit NetApp DataOps", Fournir les meilleures fonctionnalités de gestion des données qu'ONTAP offre. Ce kit permet aux équipes de gestion de données d'entreprise de créer rapidement des prototypes de modèles avec des données structurées et non structurées associées, à l'aide de snapshots et de clones pour la traçabilité, la gestion des versions, les tests A/B, ce qui assure la sécurité et la gouvernance et la conformité réglementaire. Voir la section "Design du stockage" pour en savoir plus.

Cette solution présente les étapes détaillées du traitement des fichiers audio, de l'entraînement des modèles NLP, de l'apprentissage par transfert et de la gestion des données. Le pipeline de bout en bout qui en résulte génère un résumé des sentiments qui s'affiche en temps réel sur les tableaux de bord des agents du support humain.




### Configuration matérielle requise

Le tableau suivant répertorie les composants matériels requis pour implémenter la solution. Ils peuvent varier selon la mise en œuvre de la solution et les besoins du client.

| Tests de latence de réponse | Temps (millisecondes) |
|---|---|
| Le traitement de données | 10 |
| Inférence | 10 |

Ces tests de temps de réponse ont été exécutés sur plus de 50,000 fichiers audio pour 560 conversations. Chaque fichier audio a une taille de ~100 Ko en MP3 et de ~1 Mo lorsqu'il est converti en fichier WAV. L'étape de traitement des données convertit des fichiers MP3 en fichiers WAV. Les étapes d'inférence convertissent les fichiers audio en texte et extraient un sentiment du texte. Ces étapes sont toutes indépendantes les unes des autres et peuvent être parallélisées pour accélérer le processus.

En tenant compte de la latence du transfert des données entre les magasins, les responsables devraient être en mesure de voir les mises à jour de l'analyse des sentiments en temps réel dans une seconde de la fin de la phrase.

#### Matériel NVIDIA RIVA

| Sous-jacent | De formation |
|---|---|
| OS | Linux x86_64 |
| Mémoire GPU (ASR) | Modèles de diffusion en continu : ~5600 Mo modèles sans diffusion en continu : environ 3100 Mo |
| Mémoire GPU (NLP) | Environ 500 Mo par modèle BERT |

#### Matériel du kit d'outils NVIDIA TAO

| Sous-jacent | De formation |
|---|---|
| RAM système | 32 GO |
| MÉMOIRE VIVE DU GPU | 32 GO |
| CPU | 8 cœurs |
| GPU | NVIDIA (A100, V100 et RTX 30x0) |
| SSD | 100 GO |

### Système de stockage Flash

#### NetApp ONTAP 9

ONTAP 9.9, la dernière génération de logiciel de gestion du stockage de NetApp, permet aux entreprises de moderniser l'infrastructure et de passer à un data Center prêt pour le cloud. Avec des capacités de gestion des données à la pointe du secteur, ONTAP permet de gérer et de protéger les données avec un seul ensemble d'outils, quel que soit leur emplacement. Vous pouvez aussi déplacer vos données librement partout où elles sont nécessaires : la périphérie, le cœur ou le cloud. ONTAP 9.9 comprend de nombreuses fonctionnalités qui simplifient la gestion des données, accélèrent et protègent les données stratégiques, et permettent d'utiliser des fonctionnalités d'infrastructure nouvelle génération dans toutes les architectures de cloud hybride.

#### Copie et synchronisation NetApp BlueXP

"Copie et synchronisation BlueXP" Est un service NetApp permettant une synchronisation rapide et sécurisée des données, qui vous permet de transférer des fichiers entre des partages de fichiers NFS ou SMB sur site vers l'une des cibles suivantes :

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

BlueXP Copy and Sync déplace les fichiers où vous le souhaitez, rapidement et de manière sécurisée. Une fois vos données transférées, elles peuvent être utilisées à la fois sur la source et sur la cible. BlueXP Copy and Sync synchronise en continu les données en fonction de votre planification prédéfinie, en déplaçant uniquement les données modifiées. Le temps et l'argent consacrés à la réplication des données sont ainsi réduits. BlueXP Copy and Sync est un outil SaaS (logiciel en tant que service) simple à configurer et à utiliser. Les transferts de données déclenchés par la copie et la synchronisation BlueXP sont effectués par les courtiers de données. Vous pouvez déployer des courtiers de données BlueXP Copy and Sync dans AWS, Azure, Google Cloud Platform ou sur site.

#### NetApp StorageGRID

La suite de stockage objet Software-defined de StorageGRID prend en charge de nombreux cas d'utilisation de manière transparente dans les environnements multiclouds publics, privés et hybrides. Grâce à des innovations de pointe, NetApp StorageGRID stocke, sécurise, protège et préserve les données non structurées à une utilisation polyvalente, y compris la gestion automatisée du cycle de vie sur de longues périodes. Pour plus d'informations, reportez-vous à la section "NetApp StorageGRID" le site.

### Configuration logicielle requise

Les composants logiciels requis pour implémenter cette solution sont répertoriés dans le tableau suivant. Ils peuvent varier selon la mise en œuvre de la solution et les besoins du client.

| Machine hôte | De formation |
|---|---|
| RIVA (anciennement JARVIS) | 1.4.0 |
| TAO Toolkit (anciennement Transfer Learning Toolkit) | 3.0 |
| ONTAP | 9.9.1 |
| SYSTÈME D'EXPLOITATION DGX | 5.1 |
| DOTK | 2.0.0 |

#### Logiciel NVIDIA RIVA

| Logiciel | De formation |
|---|---|
| Docker | Pour 19.02 (avec nvidia-docker installé)>=19.03 si vous n'utilisez pas DGX |
| Pilote NVIDIA | Plus de 465.19.01 418.40+, 440.33+, 450.51+ et 460.27+ pour les processeurs graphiques du data Center |
| OS de conteneur | Ubuntu 20.04 |
| CUDA | 11.3.0 |
| CcuBLAS | 11.5.1.101 |
| Distance | 8.2.0.41 |
| NCCL | 2.9.6 |
| TensorRT | 7.2.3.4 |
| Serveur d'inférence Triton | 2.9.0 |

#### Logiciel NVIDIA TAO Toolkit

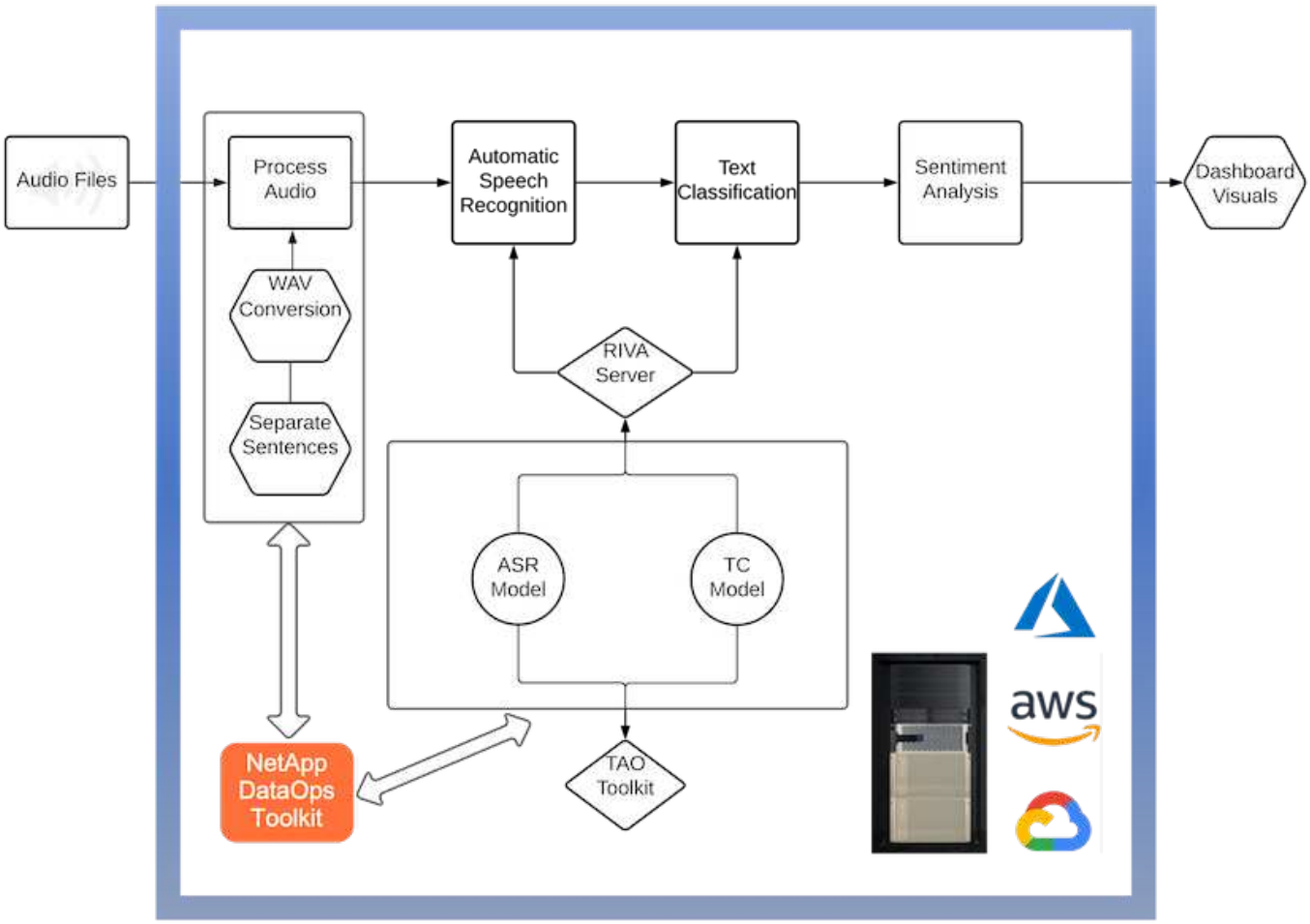
| Logiciel | De formation |
|---|---|
| Ubuntu 18.04 LTS | 18.04 |
| python | >=3.6.9 |
| docker-ce | Pour 19.03.5 |
| docker-API | 1.40 |

| Logiciel | De formation |
|---|---|
| kit-conteneur-nvidia | Pour 1.3.0-1 |
| exécution-conteneur-nvidia | 3.4.0-1 |
| nvidia-docker2 | 2.5.0-1 |
| pilote nvidia | Pour 455 |
| python-pip | Pour 21.06 |
| nvidia-pyindex | Dernière version |

### Détails du cas d'utilisation

Cette solution s'applique aux cas d'utilisation suivants :

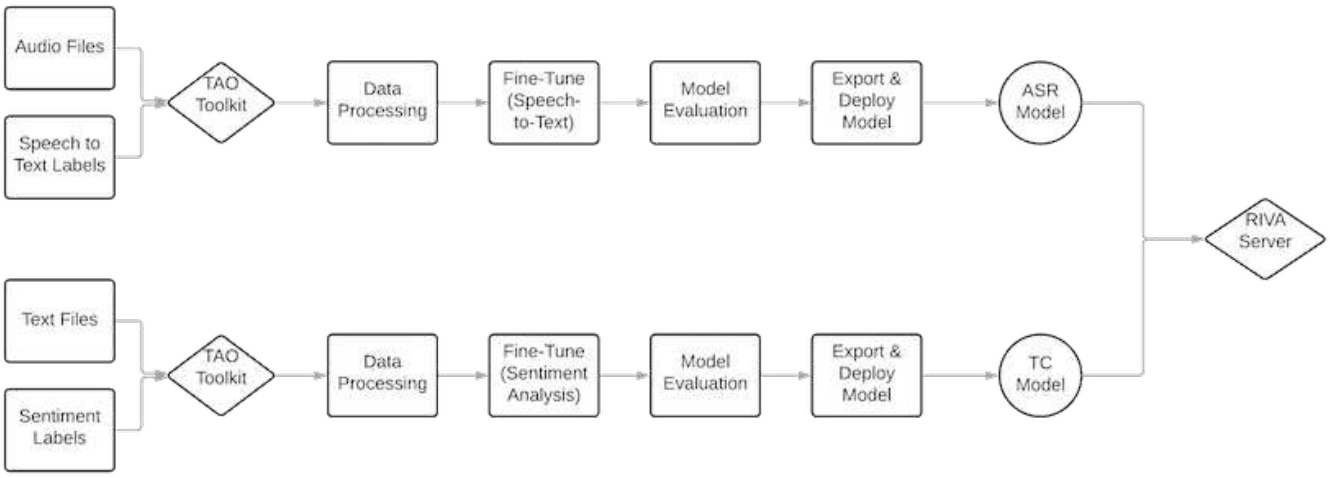
- Parole-à-texte
- Analyse des sentiments




L'utilisation de texte à texte commence par l'acquisition de fichiers audio pour les centres de support. Ce son est ensuite traité pour s'adapter à la structure requise par RIVA. Si les fichiers audio n'ont pas déjà été divisés en unités d'analyse, cela doit être fait avant de transmettre l'audio à RIVA. Une fois le fichier audio traité, il est transmis au serveur RIVA en tant qu'appel API. Le serveur utilise l'un des nombreux modèles qu'il héberge et renvoie une réponse. Ce message vocal au texte (fait partie de la reconnaissance vocale automatique) renvoie

une représentation textuelle de l'audio. À partir de là, le pipeline passe à la partie analyse de sentiment.

Pour l'analyse des sentiments, la sortie du texte de la reconnaissance vocale automatique sert d'entrée à la classification de texte. La classification de texte est le composant NVIDIA permettant de classer le texte dans n'importe quelle catégorie. Les catégories de sentiments varient de positif à négatif pour les conversations du centre de support. Les performances des modèles peuvent être évaluées à l'aide d'un jeu de retenue pour déterminer la réussite de l'étape de réglage précis.



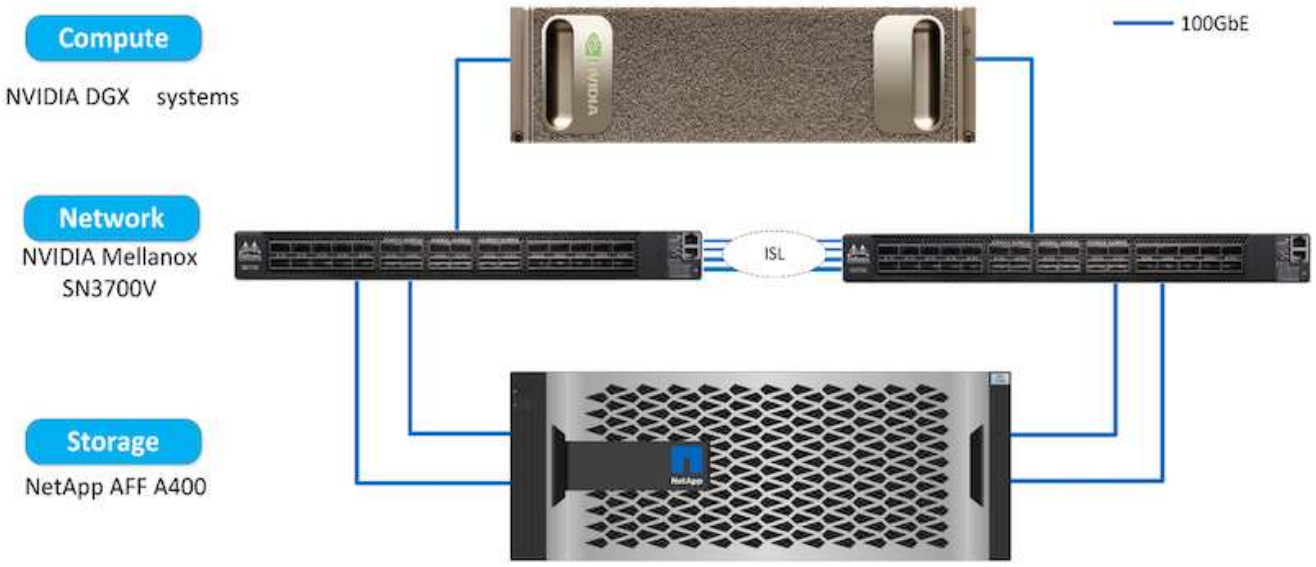

Un pipeline similaire est utilisé à la fois pour l'analyse de la parole au texte et de l'opinion dans la boîte à outils TAO. La différence majeure réside dans l'utilisation d'étiquettes nécessaires au réglage précis des modèles. Le pipeline TAT Toolkit commence par le traitement des fichiers de données. Puis les modèles pré-entraînés (provenant du "Catalogue NVIDIA NGC") sont affinée à l'aide des données du centre de support. Les modèles optimisés sont évalués en fonction des metrics de performance correspondants. S'ils sont plus performants que les modèles pré-entraînés, ils sont déployés sur le serveur RIVA.

# Considérations relatives à la conception

Cette section décrit les différents composants de cette solution à prendre en compte lors de sa conception.

## Conception du réseau et des ressources de calcul

En fonction des restrictions liées à la sécurité des données, toutes les données doivent rester dans l'infrastructure du client ou dans un environnement sécurisé.




## Conception du stockage

Le kit NetApp DataOps sert de service principal pour la gestion des systèmes de stockage. Le kit DataOps est une bibliothèque Python qui facilite le clonage quasi instantané d'un volume de données ou d'un espace de travail JupyterLab par exemple, des développeurs, des data Scientists, et des ingénieurs DevOps, Et le snapshots quasi-instantané d'un volume de données ou d'un espace de travail JupyterLab à des fins de traçabilité ou de base. Cette bibliothèque Python peut fonctionner comme un utilitaire de ligne de commande ou une bibliothèque de fonctions pouvant être importées dans n'importe quel programme Python ou ordinateur portable Jupyter.

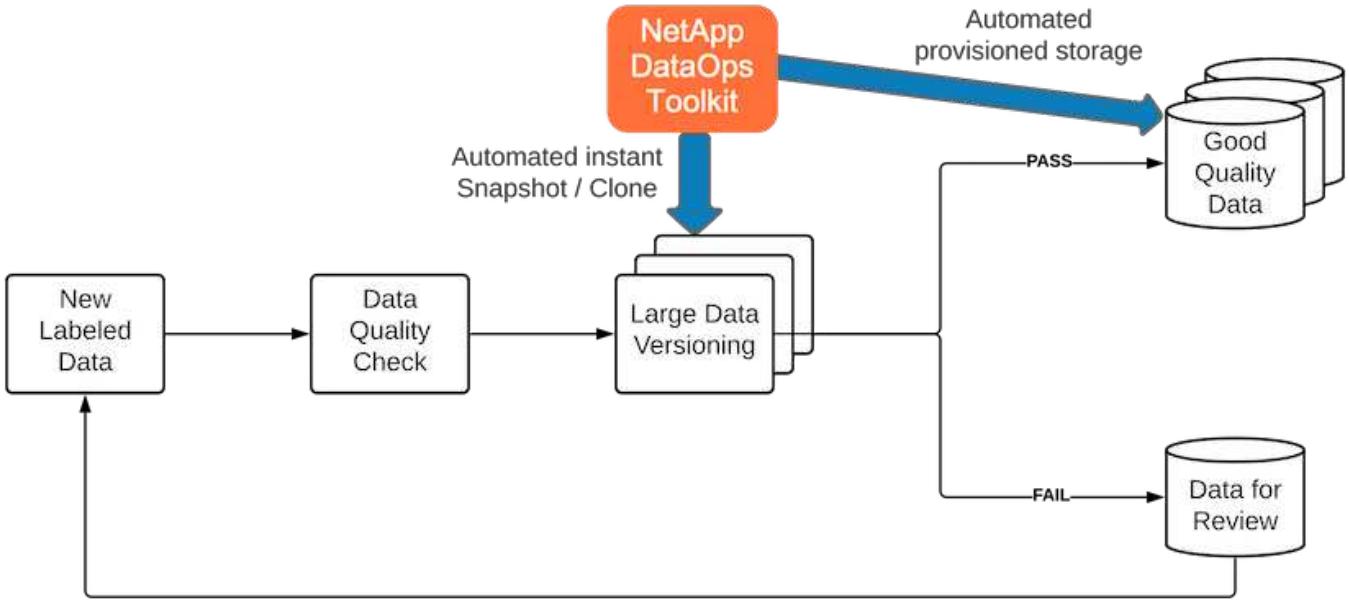
## Meilleures pratiques RIVA

NVIDIA présente plusieurs avantages généraux "de meilleures pratiques sur les données" Pour utiliser RIVA :

- **Utilisez des formats audio sans perte si possible.** l'utilisation de codecs avec perte tels que MP3 peut réduire la qualité.
- **Augmenter les données de formation.** l'ajout de bruit de fond aux données de formation audio peut d'abord diminuer la précision tout en augmentant la robustesse.
- **Limiter la taille du vocabulaire en cas d'utilisation de texte gratté.** de nombreuses sources en ligne contiennent des fautes de frappe ou des pronoms auxiliaires et des mots peu communs. La suppression de ces éléments peut améliorer le modèle de langue.
- **Utilisez un taux d'échantillonnage minimum de 16 kHz si possible.** toutefois, essayez de ne pas rééchantillonner, car cela diminue la qualité audio.

Outre ces meilleures pratiques, il est nécessaire que les clients privilégient la collecte d'un dataset d'échantillon représentatif, avec des étiquettes précises pour chaque étape du pipeline. En d'autres termes, le dataset exemple doit refléter de manière proportionnelle les caractéristiques spécifiées présentées dans un dataset cible. De la même façon, les annotations du dataset ont la responsabilité d'équilibrer la précision et la vitesse d'étiquetage, de sorte que la qualité et la quantité des données soient optimisées. Par exemple, cette solution de centre de support nécessite des fichiers audio, du texte étiqueté et des étiquettes de sentiment. La nature séquentielle de cette solution signifie que les erreurs du début du pipeline sont propagées jusqu'à la fin Si les fichiers audio sont de mauvaise qualité, les transcriptions de texte et les libellés de sentiment seront aussi.

Cette propagation d'erreur s'applique également aux modèles entraînés sur ces données. Si les prévisions de sentiment sont exactes à 100 % mais que le modèle de parole à texte fonctionne mal, alors le pipeline final est limité par les transcriptions audio à texte initiales. Il est essentiel que les développeurs considèrent les performances de chaque modèle individuellement et comme un composant d'un pipeline plus vaste. Dans ce cas particulier, l'objectif final est de développer un pipeline capable de prédire avec précision le sentiment. Par conséquent, la mesure globale sur laquelle évaluer le pipeline est la précision des sentiments, que la transcription de la parole vers le texte affecte directement.




Le kit NetApp DataOps complète le pipeline de contrôle de la qualité des données grâce à la technologie de clonage quasi-instantané. Chaque fichier étiqueté doit être évalué et comparé aux fichiers étiquetés existants. La distribution de ces contrôles de qualité sur différents systèmes de stockage des données permet de réaliser ces vérifications rapidement et efficacement.

# Déploiement de l'analyse des sentiments du centre de support

Le déploiement de la solution implique les composants suivants :

1. Kit NetApp DataOps
2. Configuration NGC
3. Serveur NVIDIA RIVA
4. Kit d'outils NVIDIA TAO
5. Exporter des modèles TAO vers RIVA

Pour effectuer le déploiement, procédez comme suit :

## Kit NetApp DataOps : analyse des sentiments des centres de support

Pour utiliser le "Kit NetApp DataOps", effectuez les opérations suivantes :

1. PIP installer la boîte à outils.

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

2. Configuration de la gestion des données

```
netapp_dataops_cli.py config
```

## Configuration NGC : analyse du sentiment du centre de support

Pour configurer "NVIDIA NGC", effectuez les opérations suivantes :

1. Télécharger le contrôleur NGC.

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

2. Ajoutez votre répertoire actuel au chemin d'accès.

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

3. Vous devez configurer l'interface de ligne de commandes NGC pour que vous puissiez exécuter les commandes. Entrez la commande suivante, y compris votre clé API lorsque vous y êtes invité.

```
ngc config set
```

Pour les systèmes d'exploitation qui ne sont pas basés sur Linux, rendez-vous sur "ici".

## Serveur NVIDIA RIVA : analyse des sentiments des centres de support

Pour configurer "NVIDIA RIVA", effectuez les opérations suivantes :

1. Téléchargez les fichiers RIVA de NGC.

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

2. Initialiser LA configuration RIVA (`riva_init.sh`).
3. Démarrez LE serveur RIVA (`riva_start.sh`).
4. Démarrez le client RIVA (`riva_start_client.sh`).

5. Dans le client RIVA, installez la bibliothèque de traitement audio ( "FFMPEG")

```
apt-get install ffmpeg
```

6. Démarrez le "Jupyter" serveur.
7. Exécutez le bloc-notes RIVA Inférence Pipeline.

### Kit NVIDIA TAO : analyse des sentiments du centre de support

Pour configurer la boîte à outils NVIDIA TAO, procédez comme suit :

1. Préparez et activez un "virtualisé" Pour la boîte à outils TAO.
2. Installer le "packages requis".
3. Tirez manuellement l'image utilisée pendant l'entraînement et le réglage précis.

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

4. Démarrez le "Jupyter" serveur.
5. Exécutez l'ordinateur portable TAT Fine Tuning.

### Exporter des modèles TAO vers RIVA : prendre en charge l'analyse des sentiments des centres de support

À utiliser "Modèles TAO Toolkit à RIVA", effectuez les opérations suivantes :

1. Enregistrez des modèles dans le portable TAT Fine Tuning.
2. Copier des modèles entraînés TAO dans le répertoire de modèle RIVA.
3. Démarrez LE serveur RIVA (`riva_start.sh`).

### Les obstacles au déploiement

Voici quelques points à garder à l'esprit lors du développement de votre propre solution :

- Le kit NetApp DataOps est installé en premier pour assurer le fonctionnement optimal du système de stockage des données.
- NVIDIA NGC doit être installé avant toute autre chose, car il authentifie le téléchargement des images et des modèles.
- RIVA doit être installé avant la boîte à outils TAO. L'installation DE RIVA configure le démon docker pour extraire les images si nécessaire.
- DGX et docker doivent avoir un accès à Internet pour télécharger les modèles.

## Résultats de la validation

Comme mentionné dans la section précédente, les erreurs sont propagées dans l'ensemble du pipeline chaque fois qu'au moins deux modèles d'apprentissage machine

sont exécutés dans l'ordre. Pour cette solution, le sentiment de la phrase est le facteur le plus important dans la mesure du niveau de risque stock de l'entreprise. Le modèle de parole à texte, bien qu'essentiel au pipeline, sert d'unité de prétraitement avant que les sentiments ne puissent être prédits. Ce qui compte vraiment, c'est la différence de sentiment entre les phrases de vérité au sol et les phrases prévues. Cela sert de proxy pour le taux d'erreur de mot (WER). La précision de la parole au texte est importante, mais le WER n'est pas directement utilisé dans la mesure finale du pipeline.

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

Ces mesures de sentiment peuvent être calculées pour le score F1, le rappel et la précision de chaque phrase. Les résultats peuvent ensuite être agrégés et affichés dans une matrice de confusion, ainsi que les intervalles de confiance pour chaque mesure.

L'apprentissage par transfert améliore les performances des modèles pour une fraction des besoins de données, des délais d'entraînement et des coûts. Les modèles optimisés doivent également être comparés à leurs versions de base afin de garantir que l'apprentissage par transfert améliore les performances au lieu de l'associer. En d'autres termes, le modèle affinée doit optimiser les performances des données de centre de support par rapport au modèle pré-entraînés.

### Évaluation du pipeline

| Cas de test | Détails |
|---|---|
| Numéro de test | Mesure du sentiment du pipeline |
| Conditions préalables au test | Modèles optimisés pour les modèles d'analyse de la voix au texte et des sentiments |
| Résultat attendu | La mesure de sentiment du modèle affinée est plus performante que le modèle pré-pré-pré-entraînés d'origine. |

### Mesure du sentiment du pipeline

1. Calculez la mesure du sentiment pour le modèle de référence.
2. Calculez la mesure du sentiment pour le modèle affinée.
3. Calculez la différence entre ces mesures.
4. Faire la moyenne des différences entre toutes les phrases.

## Vidéos et démonstrations

Deux ordinateurs portables contiennent le pipeline d'analyse des sentiments : ""Support-Center-Model-Transfer-Learning-and-Fine-Tuning.ipynb"" et "« Support-Center-sentiment-analyse-pipeline.ipynb »". Ensemble, ces ordinateurs portables montrent comment développer un pipeline permettant d'ingérer les données du data Center et d'extraire les sentiments à partir de chaque phrase à l'aide de modèles de deep learning les plus sophistiqués sur les données de l'utilisateur.

## Centre de soutien - Pipeline d'analyse de sentiment.ipynb

Cet ordinateur portable contient le pipeline RIVA d'inférence pour l'importation de fichiers audio, la conversion en texte et l'extraction de sentiments à utiliser dans un tableau de bord externe. Le jeu de données est automatiquement téléchargé et traité si ce n'est pas déjà fait. La première section du bloc-notes est la parole en texte qui gère la conversion de fichiers audio en texte. La section analyse de sentiment est suivie de la section analyse de sentiment qui extrait des sentiments pour chaque phrase de texte et affiche ces résultats dans un format similaire au tableau de bord proposé.

> Cet ordinateur portable doit être exécuté avant l'entraînement et le réglage précis du modèle car le jeu de données MP3 doit être téléchargé et converti au format approprié.

# 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

## Centre de support - modèle de formation et affinage.ipynb

L'environnement virtuel TAO Toolkit doit être configuré avant d'exécuter l'ordinateur portable (voir la section boîte à outils TAO dans la vue d'ensemble des commandes pour les instructions d'installation).

Cet ordinateur portable s'appuie sur le kit d'outils TAO pour affiner les modèles de deep learning sur les données clients. Comme pour le précédent bloc-notes, celui-ci est divisé en deux sections pour les composants parole-à-texte et analyse de sentiment. Chaque section passe par le traitement des données,

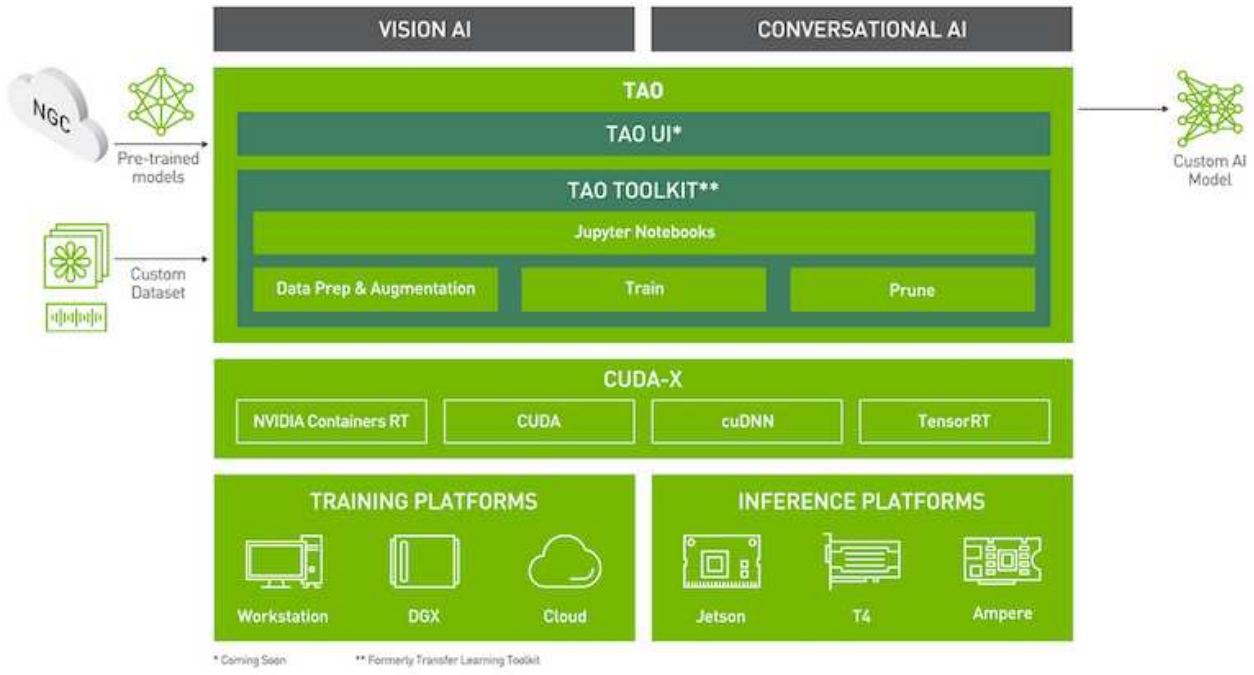
l'entraînement et le réglage des modèles, l'évaluation des résultats et l'exportation des modèles. Enfin, vous avez la fin du déploiement de vos deux modèles plus adaptés à RIVA.

## 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

L'expérience client étant de plus en plus considérée comme un champ de bataille essentiel de la concurrence, un centre de support mondial avec intelligence artificielle est devenu un composant essentiel que les entreprises de presque tous les secteurs ne peuvent se permettre de négliger. La solution proposée dans ce rapport technique a été démontrée pour soutenir la prestation de telles expériences client exceptionnelles. Le défi consiste désormais à s'assurer que les entreprises prennent des mesures pour moderniser leur infrastructure et leurs workflows d'IA.

Les meilleures implémentations d'IA dans le service client ne sont pas à remplacer des agents humains. L'IA leur permet plutôt de créer des expériences client exceptionnelles grâce à l'analyse des sentiments en temps réel, à l'escalade des conflits et à un calcul multimodal affectueux pour détecter des indices verbaux, non verbaux et faciaux avec lesquels des modèles d'IA complets peuvent formuler des recommandations à grande échelle et compléter les lacunes d'un agent humain. L'IA permet également une meilleure correspondance entre un client particulier et les agents disponibles actuellement. Grâce à l'IA, les entreprises peuvent extraire un sentiment précieux des clients concernant leurs pensées et impressions sur les produits, services et image de marque du fournisseur.

La solution peut également être utilisée pour construire des données de séries chronologiques destinées aux agents de support afin de servir de mesure objective d'évaluation des performances. Les enquêtes classiques sur la satisfaction des clients manquent souvent de réponses suffisantes. En recueillant des sentiments à long terme des employés et des clients, les employeurs peuvent prendre des décisions éclairées concernant le rendement des agents de soutien.

L'association de NetApp, des SFL Scientific, des frameworks d'orchestration open source et de NVIDIA rassemble les dernières technologies en tant que services gérés, tout en offrant une grande flexibilité pour accélérer l'adoption des technologies et améliorer le délai de mise sur le marché des nouvelles applications d'IA/ML. Ces services avancés sont fournis sur site, facilement ports pour les environnements cloud natifs, ainsi que pour les architectures de déploiement hybride.

# Où trouver des informations complémentaires

## Pour en savoir plus sur les informations données dans ce livre blanc, consultez ces documents et/ou sites web :

- Démonstrations interactives en 3D

  "www.netapp.com/ai"

- Entrez en contact directement avec un spécialiste de l'IA NetApp

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

- Description de la solution NVDIA base Command Platform avec NetApp

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

- Infographie NetApp pour l'IA 10 : bonnes raisons de choisir NetApp

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

- L'IA dans le domaine de la santé : l'apprentissage profond pour identifier les lésions de la COVID-19 dans les analyses CT pulmonaires

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

- L'IA dans le domaine de la santé : surveillance de l'utilisation du masque facial dans les établissements de santé. Livre blanc

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

- Rapport technique sur l'IA dans le domaine de la santé : imagerie diagnostique

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

- L'IA pour le Retail : NetApp transversaux à l'IA avec NVIDIA RIVA

  "https://docs.netapp.com/us-en/netapp-solutions/ai/cainvidia_executive_summary.html"

- Description de la solution NetApp ONTAP ai

  https://www.netapp.com/pdf.html?item=/media/6736-sb-3939.pdf

- Description de la solution du kit DataOps NetApp

  https://www.netapp.com/pdf.html?item=/media/21480-SB-4111-1220-NA-Data-Science-Toolkit.pdf

- Description de la solution NetApp ai Control plane

  https://www.netapp.com/pdf.html?item=/media/6737-sb-4055.pdf

- Transformation du secteur grâce à l'eBook Data Drive IA

  "https://www.netapp.com/us/media/na-337.pdf"

- Description de la solution d'IA NetApp EF-Series

  https://www.netapp.com/pdf.html?item=/media/26708-SB-4136-NetApp-AI-E-Series.pdf

- Description de la solution NetApp ai et Lenovo ThinkSystem for ai Inférence

  https://www.netapp.com/pdf.html?item=/media/25316-SB-4129.pdf

- Description des solutions NetApp pour l'IA et le ML Lenovo ThinkSystem pour les entreprises

  https://www.netapp.com/pdf.html?item=/media/25317-SB-4128.pdf

- NetApp et NVIDIA : redéfinir le contenu du possible avec la vidéo d'IA

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

**Informations sur le copyright**

Copyright © 2024 NetApp, Inc. Tous droits réservés. Imprimé aux États-Unis. Aucune partie de ce document protégé par copyright ne peut être reproduite sous quelque forme que ce soit ou selon quelque méthode que ce soit (graphique, électronique ou mécanique, notamment par photocopie, enregistrement ou stockage dans un système de récupération électronique) sans l'autorisation écrite préalable du détenteur du droit de copyright.

Les logiciels dérivés des éléments NetApp protégés par copyright sont soumis à la licence et à l'avis de non-responsabilité suivants :

CE LOGICIEL EST FOURNI PAR NETAPP « EN L'ÉTAT » ET SANS GARANTIES EXPRESSES OU TACITES, Y COMPRIS LES GARANTIES TACITES DE QUALITÉ MARCHANDE ET D'ADÉQUATION À UN USAGE PARTICULIER, QUI SONT EXCLUES PAR LES PRÉSENTES. EN AUCUN CAS NETAPP NE SERA TENU POUR RESPONSABLE DE DOMMAGES DIRECTS, INDIRECTS, ACCESSOIRES, PARTICULIERS OU EXEMPLAIRES (Y COMPRIS L'ACHAT DE BIENS ET DE SERVICES DE SUBSTITUTION, LA PERTE DE JOUISSANCE, DE DONNÉES OU DE PROFITS, OU L'INTERRUPTION D'ACTIVITÉ), QUELLES QU'EN SOIENT LA CAUSE ET LA DOCTRINE DE RESPONSABILITÉ, QU'IL S'AGISSE DE RESPONSABILITÉ CONTRACTUELLE, STRICTE OU DÉLICTUELLE (Y COMPRIS LA NÉGLIGENCE OU AUTRE) DÉCOULANT DE L'UTILISATION DE CE LOGICIEL, MÊME SI LA SOCIÉTÉ A ÉTÉ INFORMÉE DE LA POSSIBILITÉ DE TELS DOMMAGES.

NetApp se réserve le droit de modifier les produits décrits dans le présent document à tout moment et sans préavis. NetApp décline toute responsabilité découlant de l'utilisation des produits décrits dans le présent document, sauf accord explicite écrit de NetApp. L'utilisation ou l'achat de ce produit ne concède pas de licence dans le cadre de droits de brevet, de droits de marque commerciale ou de tout autre droit de propriété intellectuelle de NetApp.

Le produit décrit dans ce manuel peut être protégé par un ou plusieurs brevets américains, étrangers ou par une demande en attente.

LÉGENDE DE RESTRICTION DES DROITS : L'utilisation, la duplication ou la divulgation par le gouvernement sont sujettes aux restrictions énoncées dans le sous-paragraphe (b)(3) de la clause Rights in Technical Data-Noncommercial Items du DFARS 252.227-7013 (février 2014) et du FAR 52.227-19 (décembre 2007).

Les données contenues dans les présentes se rapportent à un produit et/ou service commercial (tel que défini par la clause FAR 2.101). Il s'agit de données propriétaires de NetApp, Inc. Toutes les données techniques et tous les logiciels fournis par NetApp en vertu du présent Accord sont à caractère commercial et ont été exclusivement développés à l'aide de fonds privés. Le gouvernement des États-Unis dispose d'une licence limitée irrévocable, non exclusive, non cessible, non transférable et mondiale. Cette licence lui permet d'utiliser uniquement les données relatives au contrat du gouvernement des États-Unis d'après lequel les données lui ont été fournies ou celles qui sont nécessaires à son exécution. Sauf dispositions contraires énoncées dans les présentes, l'utilisation, la divulgation, la reproduction, la modification, l'exécution, l'affichage des données sont interdits sans avoir obtenu le consentement écrit préalable de NetApp, Inc. Les droits de licences du Département de la Défense du gouvernement des États-Unis se limitent aux droits identifiés par la clause 252.227-7015(b) du DFARS (février 2014).

**Informations sur les marques commerciales**

NETAPP, le logo NETAPP et les marques citées sur le site http://www.netapp.com/TM sont des marques déposées ou des marques commerciales de NetApp, Inc. Les autres noms de marques et de produits sont des marques commerciales de leurs propriétaires respectifs.