**はじめての自然言語処理（NLP：Natural Language Processing）ハンズオン**

アマゾン ウェブ サービス ジャパン  
シニアエバンジェリスト　亀田 治伸

はじめに：

このハンズオンではAWSが提供している、自然言語処理関連の4つのサービスを用いて、日本語の処理を体験するハンズオンです。以下のサービスを用います。

・Amazon Translate：高速で高品質なニューラル機械翻訳サービス

・Amazon Polly: 深層学習技術を使用したテキスト読み上げサービス

・Amazon Transcribe：音声をテキストに変換する自動音声認識サービス

・Amazon Comprehend：テキストから洞察を見つけるサービス

・Amazon Elasticsearch Service：Elasticsearch,Kibanaのマネージドサービス

・AWS Cloud9：クラウドベースの統合開発環境（IDE）

・AWS Lambda: サーバレスコンピューティングサービス

作業は特に指示されない限り、東京リージョンで行いますが、途中手順によってオハイオリージョンを使いますので、文中の記載に従ってください。

1. IAMロールの設定

ハンズオンで作成するLambda関数が他のサービスの実行に必要となる権限を作成します。

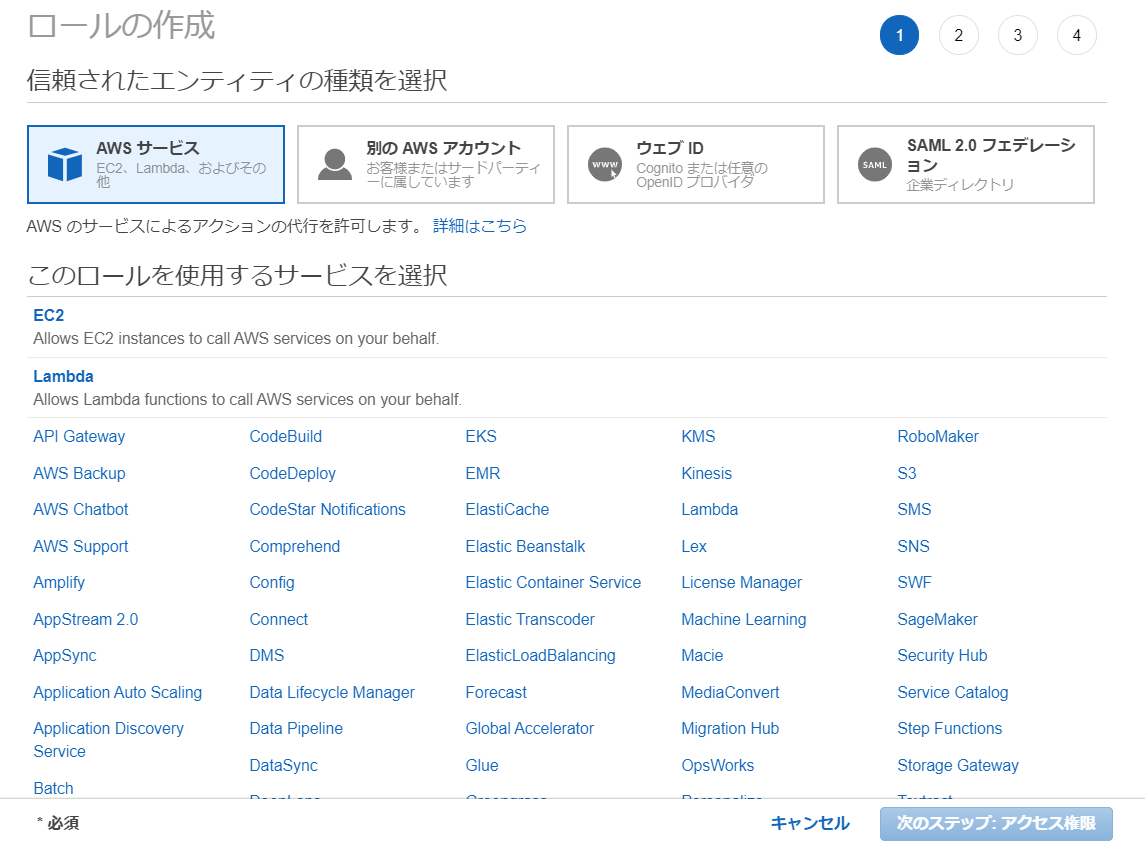
* 1. IAMのトップ画面にいき、「ロール」をクリックします。



* 1. 「ロールの作成」を押します



* 1. 「Lambda」を選択し、「次のステップ：アクセス権限」を押します



1-4.検索項目に[translate]と入力し、出てきた「TranslateFullAccess」を選びます。**右下のボタンはまだ押しません。**



* 1. 同様に「Polly」「Transcribe」「Comprehend」も検索しFullAccessにチェックを入れ最後に、「次のステップ：タグ」を押します。



* 1. 何も入力せず「次のステップ：確認」を押します。



1-6.　ロール名に「YYYYMMDDnlphandson」と入力します。（YYYYMMDDは本日の日付）



1-7.　4つのサービスへのFullAccessがついていることを確認し、「ロールの作成」を押します。以下のような緑で成功が表示されたら作業完了です。皆さんが作成したロールの名前が表示されているはずです。



1. Amazon Translate

2-1.　何か適当な英語のニュースを探してコピペしておきます。あまり長い文字だと作業が不必要に大変になるので、だいたい150文字以内にしましょう。

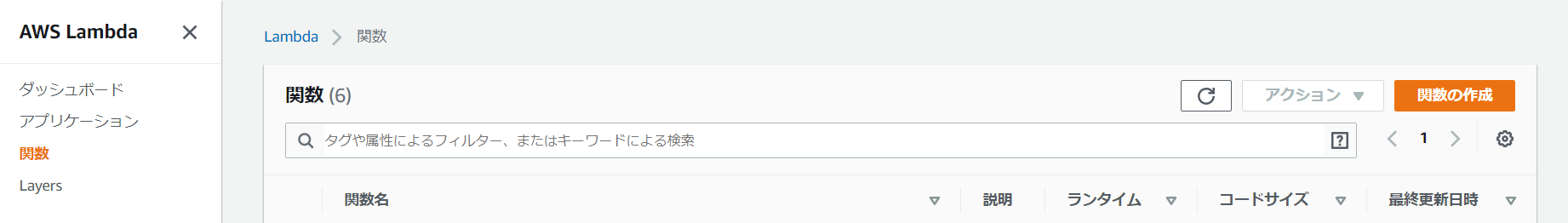
◆参考URL：

https://alpha.japantimes.co.jp/article/top\_news/202008/48249/

◆参考文章：

Terunofuji caps dramatic comeback with second title Former ozeki Terunofuji completed a stunning comeback to the top division Aug. 2 by clinching the July Grand Sumo Tournament with a final-day victory over sekiwake Mitakeumi

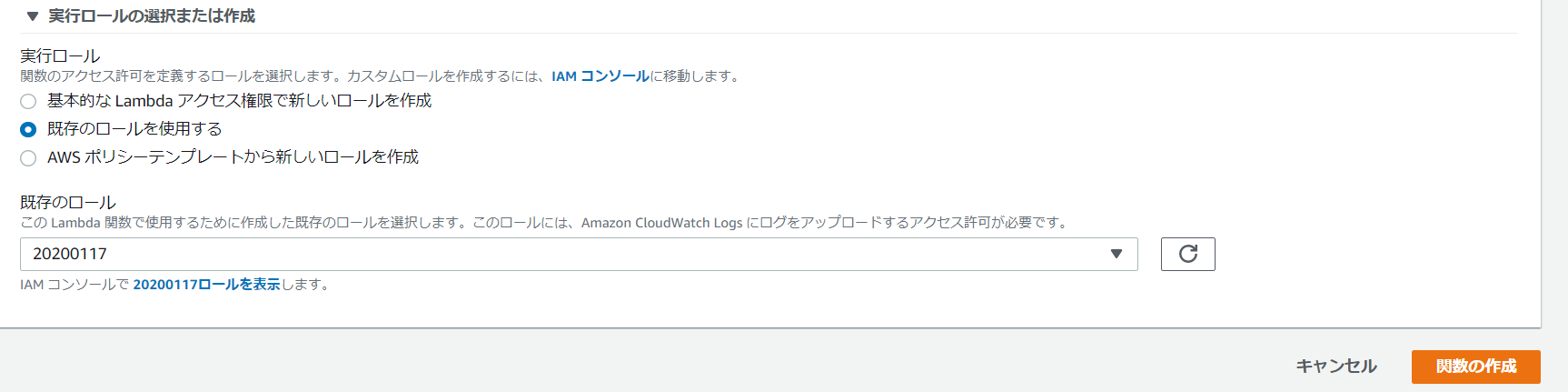
2-2.AWS Lambdaの管理者画面にいき、画面右上の「関数の作成」を押します。



2-3.関数名に「YYYYMMDDtranslate」(YYYYMMDDは本日の日付)を入力し、ランタイムは「Python3.8」を選択します。



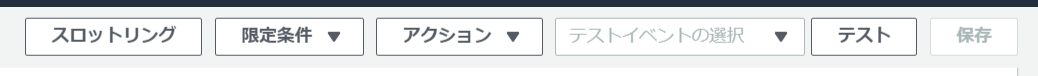
2-4.「実行ロールの選択または作成」を伸ばして、「既存のロールを選択する」を選び先ほど作成したIAMロールを選択します。選択がおわったら「関数の作成」を押します。



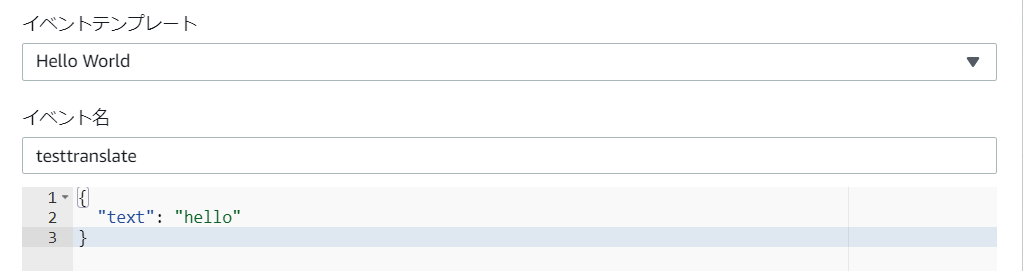
2-5.サンプルで挿入されているコードをすべて消し「translatelambda.txt」の内容をコピペします。その後右上の「保存」を押します。



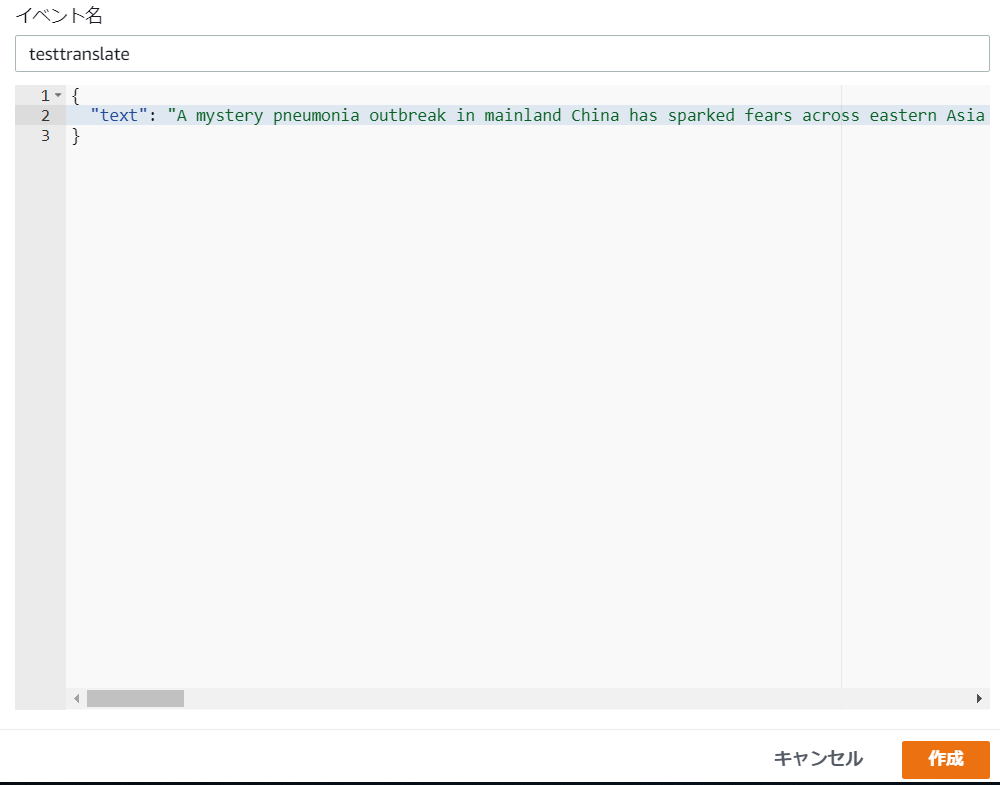
2-6.　右上の「テスト」を押します。



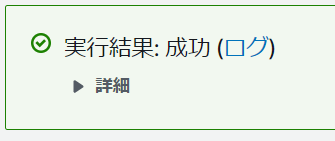
2-7．イベント名に適当な名前を入れあらかじめ挿入されている文字列を削除し、「translatelambdatest.txt」の中身に置き換えます。



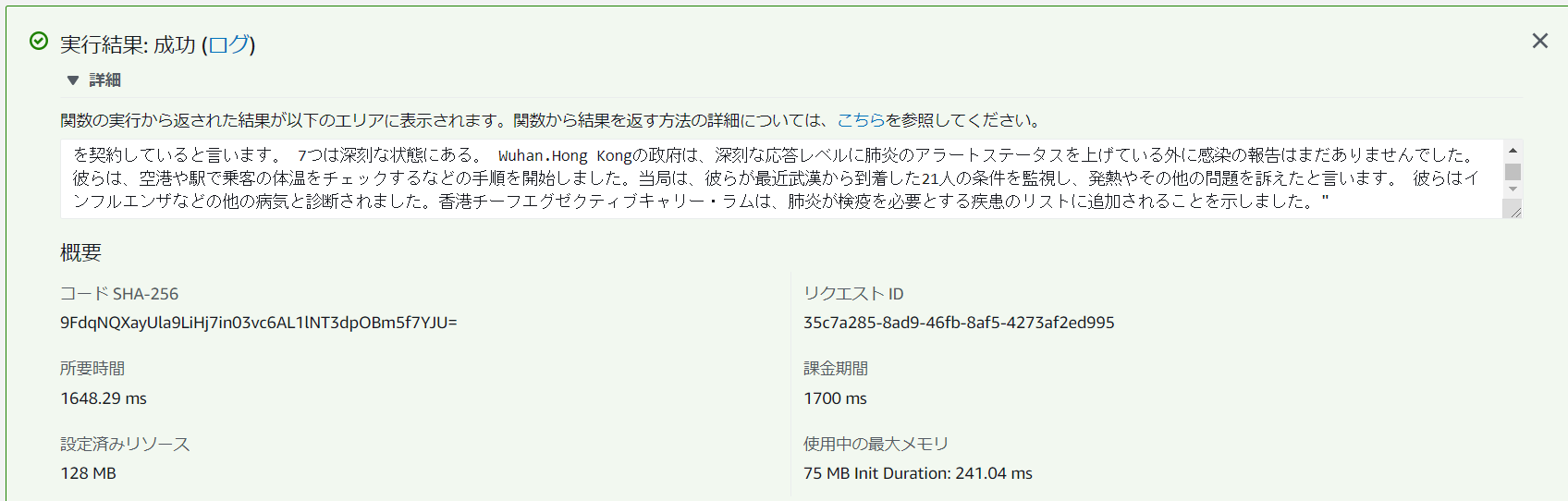
2-8.　[hello]の部分をコピペしておいた英語の記事に入れ替えます。この際**改行は全て抜いて**ください。その後右下の「作成」を押します。



2-9.　右のブラウザのスクロールバーを一番上に持っていき、「テスト」を押してください。以下が表示されれば成功です。



2-10.　「詳細」を伸ばすと翻訳された日本語がでてきます。



この日本語文字列は次のStepでも利用しますのでどこかで、テキストファイルなどにコピーをしておいてください。

1. Amazon Polly

先ほどはLambdaでTranslateを操作しましたが、実は、Translate、Polly、Transcribe、Comprehendは全てマネージメントコンソールからテストすることができるようになっています。このハンズオンではこの後何度もPollyを使用しますので、マネージメントコンソールでの手順を紹介します。

3-1．Pollyの管理者画面にアクセスします。



3-2．Translateで翻訳した文字列をコピペします。文字列はなんでもよいので、繰り返し作業を行う方は、適当な日本語を150字以下程度の文字列でコピペします。



3-3.　男性か女性の音声を選び、「ダウンロードMP3」を押すと、しばらく待てばmp3ファイルが生成されます。人名や地名など難読文字のカスタマイズもこの画面から可能です。

3-4.　ダウンロードされたmp3を再生し、正しく変換されていることを確認します。

このmp3は次のStepで使用しますので、わかりやすいところに保存しておいてください。

1. Transcribe

先ほど生成したmp3から文字起こしを行う環境を作ります。

4-1.　S3バケットを東京リージョンで作るため、S3の管理者画面に行きます。



4-2.　「バケットを作成する」を押してYYYYMMDDName（YYYYMMDDは本日の日付。Nameはご自身のアルファベットの名前）を入れます。リージョンが東京となっていることを確認し、「次へ」を押します。



4-3.　すべてデフォルトで作成しますので、もう2回「次へ」を押し、最後に「バケットを作成」を押します。



4-4.　先ほど作成したバケットをクリックし「アップロード」を押します。



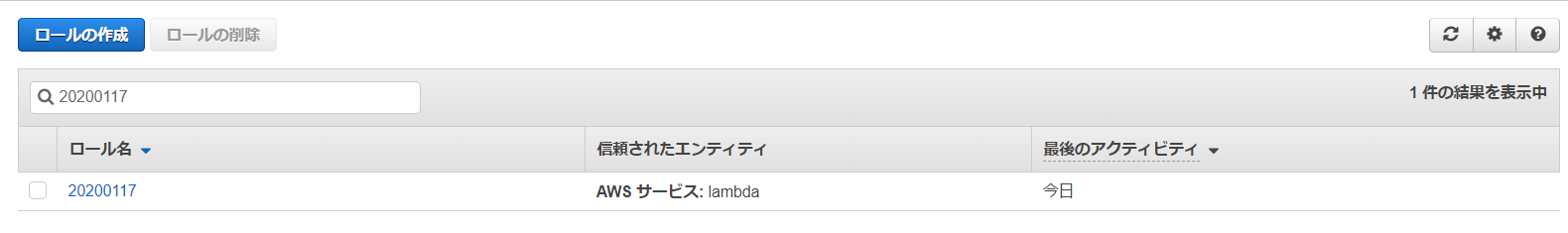
4-5.　先ほど作成したmp3をアップロードします。



4-6.　デフォルトのままアップロードしますので、「次へ」を3回、「アップロード」を1回押してください。完了すると以下のような画面になります。



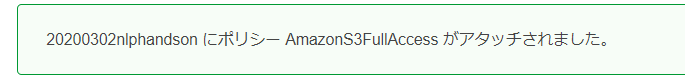
4-6-1．先ほど作成したIAMロールは、Translate、Polly、Transcribe、Comprehendへの権限が付与されていますが、このままではS3へのアクセスが許可されていません。IAMロールの画面に戻り、S３へのアクセス権限を付与します。先ほど作成したロールをIAMロール管理画面から特定し、名前をクリックします。



4-6-2.　「ポリシーをアタッチします」をおします



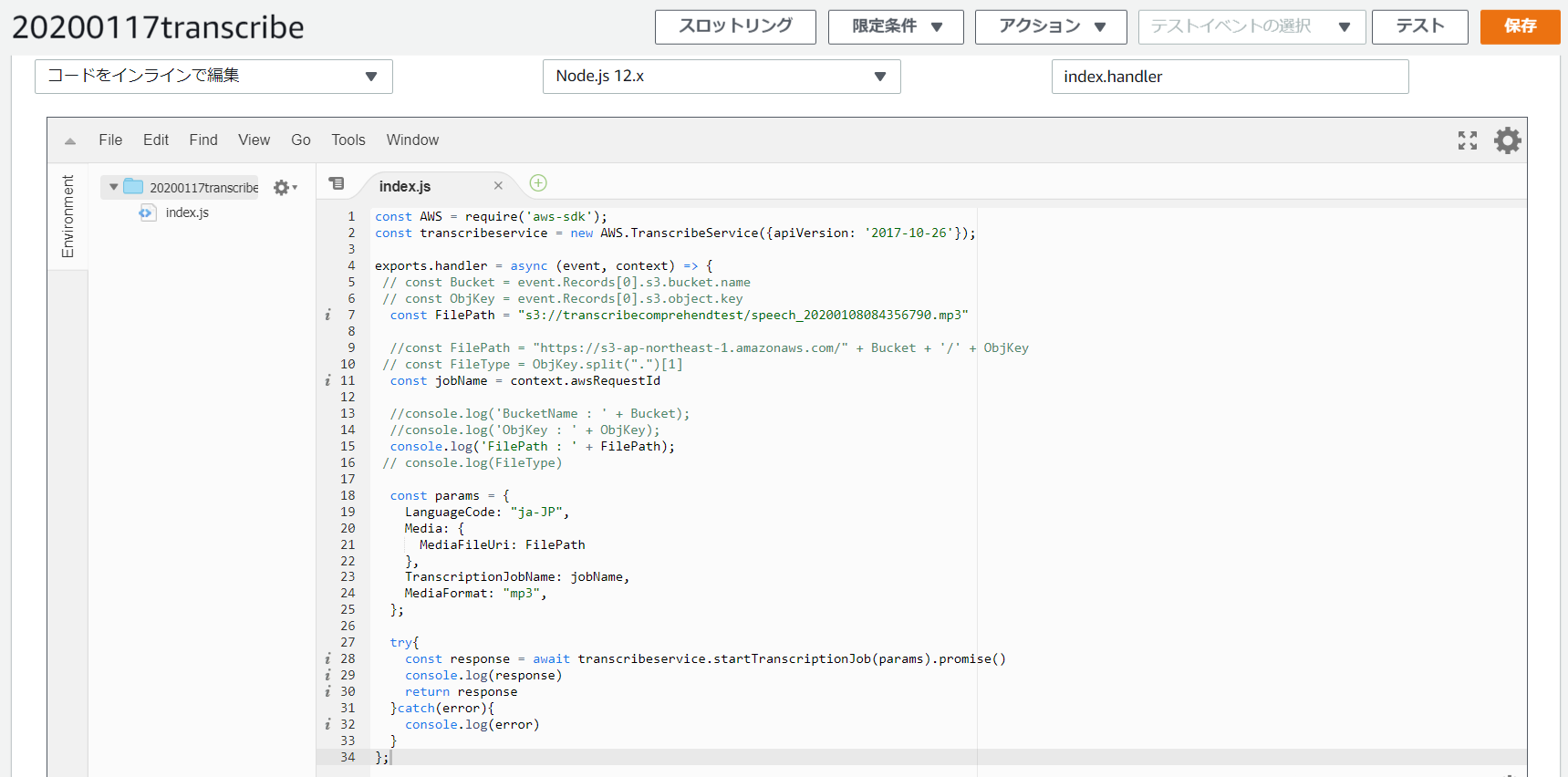
4-6-3.　「AmazonS3FullAccess」を選び「ポリシーのアタッチ」を押します。



4-7.　別のタブでLambdaの管理者画面を開き、Translateと同じ要領でYYYYMMDDtranscribeという関数を作ります。今度はランタイムにNode12.x

を選択します。**先ほど作成したロールを指定することを忘れずに行って下さい。**

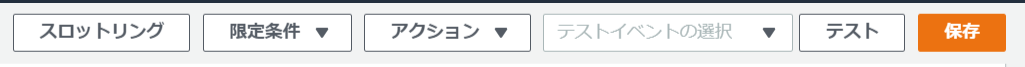
4-8.　「transcribelambda.txt」の中身をコピペします。



4-9.　 [const FilePath]の部分を、[先ほど作成したバケット名/アップしたmp3ファイル名]に変換します。以下のような文字列になるはずです。

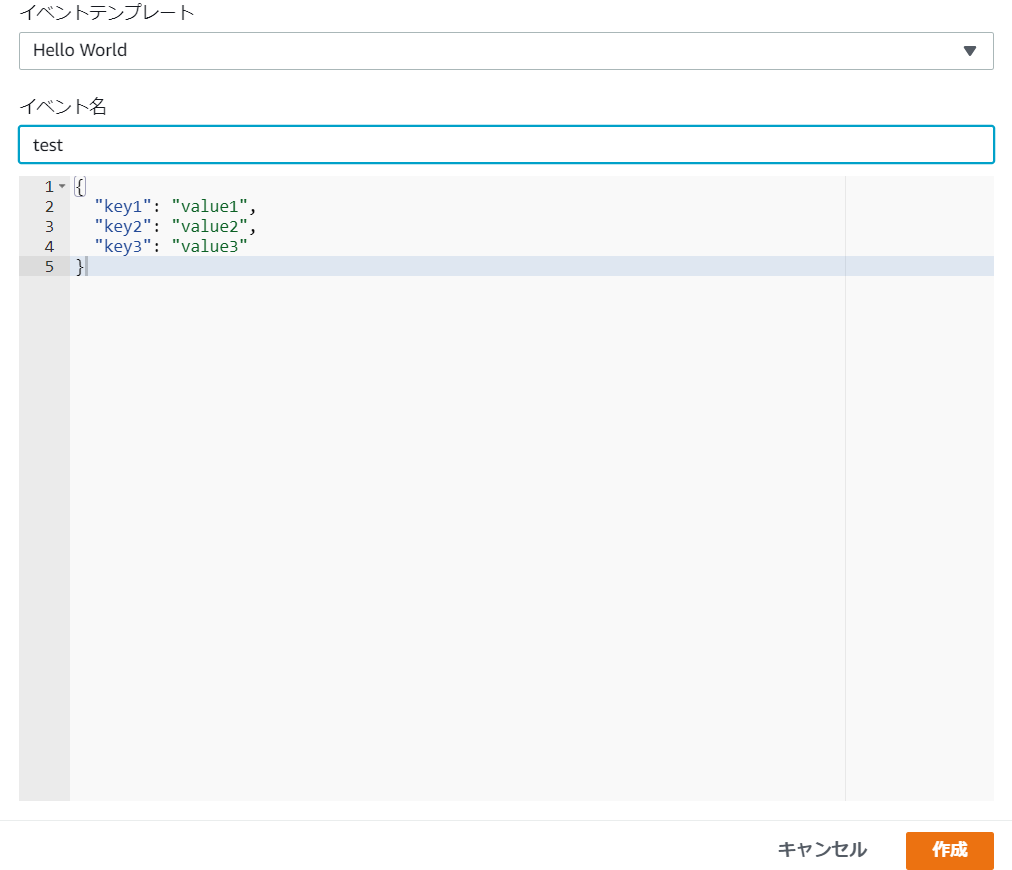
“s3://20200117kameda/speech\_20200117061325085.mp3”

4-10. 右上の「保存」を押します。



4-11.　「テスト」を押します。イベント名に適当な名前を入れて、「作成」を押します。

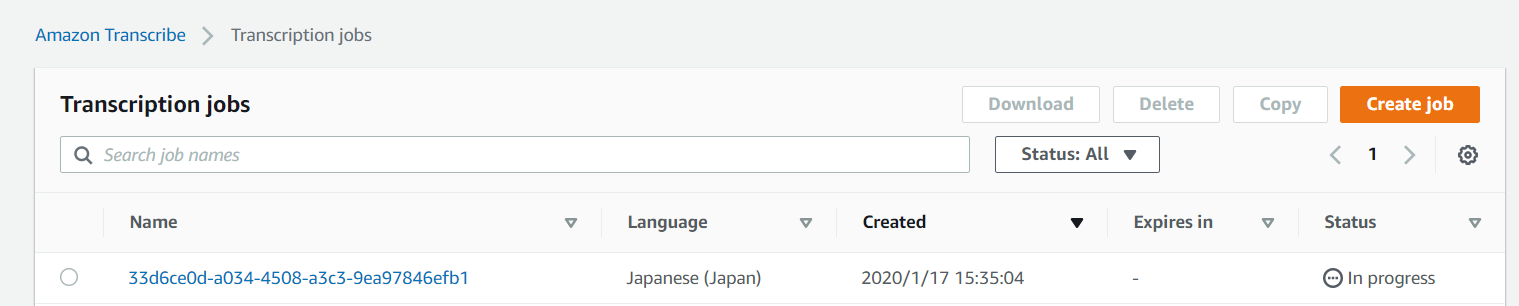
＃この関数では外部インプトは必要ありませんので、文字列はデフォルトのままダミーのテストイベントを作成する必要があります。



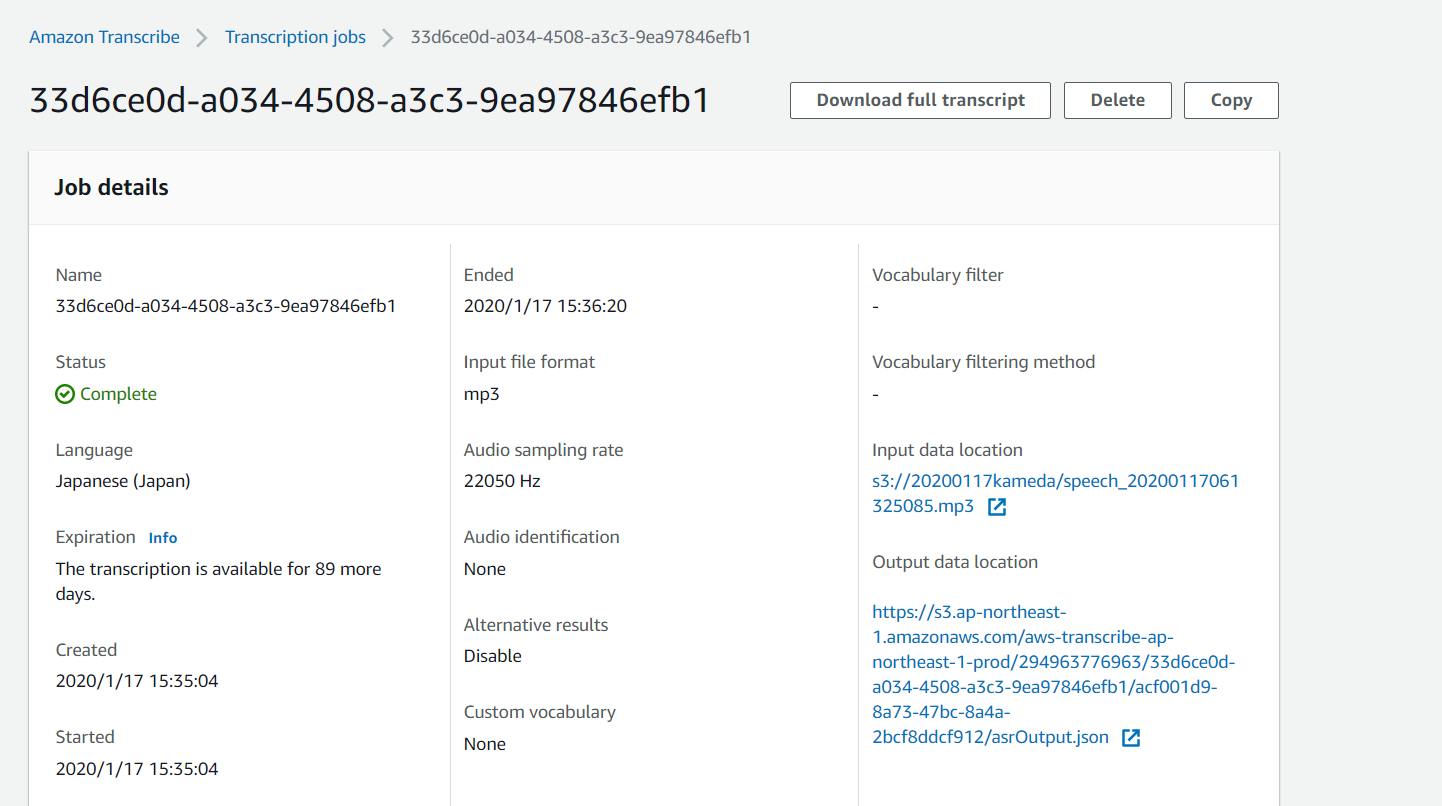
4-12.ブラウザのスクロールバーを一番上に持っていき「テスト」を押します。以下のような表示[ IN\_PROGRESS]が戻れば成功です。



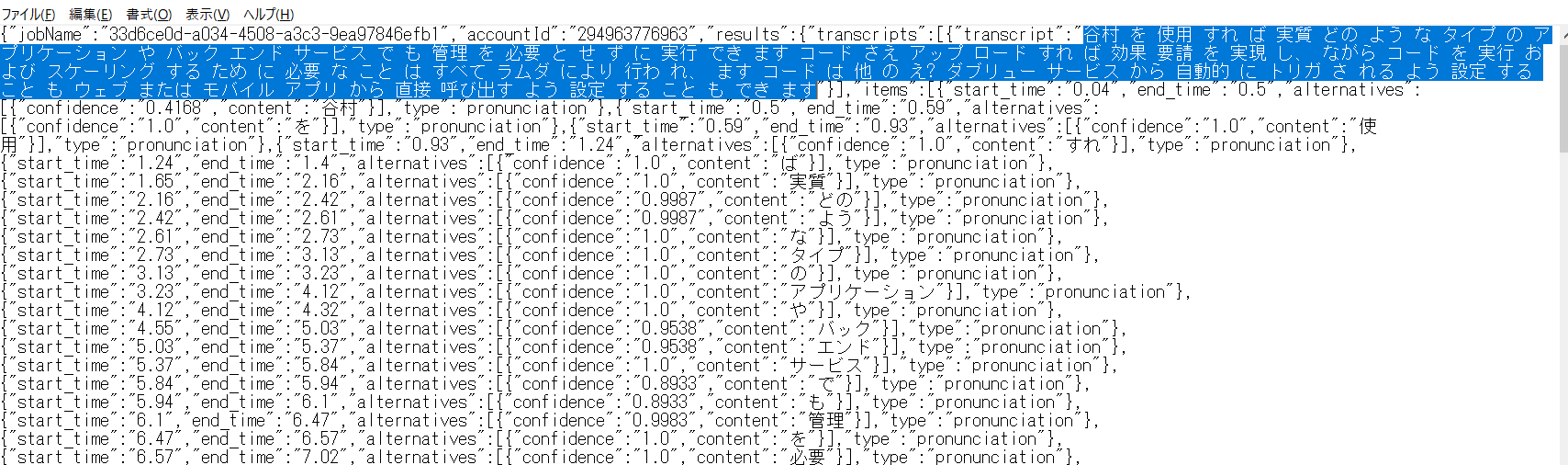
4-13.　Transcribeの管理者画面に行きます。プロジェクトが開始されていますので、[Complete]になるまでしばらくまちます。



4-14.　[Complete]になったら、Nameをクリックします。



4-15.　[Download full transcript]を押すとjsonがダウンロードされますので、開くと文字お起しされた文字列が格納されています。青い部分を次のStepで使いますので、コピペしておきます。



1. Comprehend

先ほどのTranscribeが書き起こしたmp3の内容をComprehendで分析します。

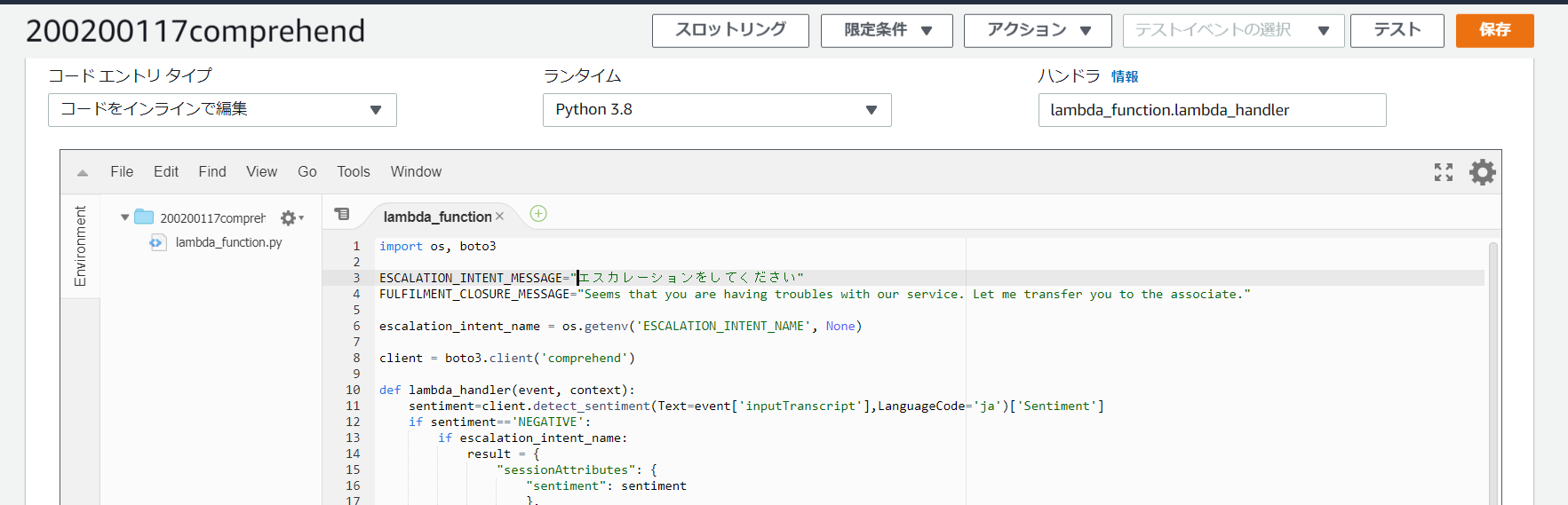
5-1.　**オレゴンリージョンで関数を作ります。画面上のリージョンをオレゴンに切り替えます。**

Lambda関数の作成画面に再度いき、YYYYMMDDComprehendで関数を作成します。

ランタイムはPython3.8を指定し、先ほどと同様に今日作成したIAMロールを設定することを忘れずに行います。（IAMロールは一度作成すれば、全AWSリージョンで利用可能な設定です）

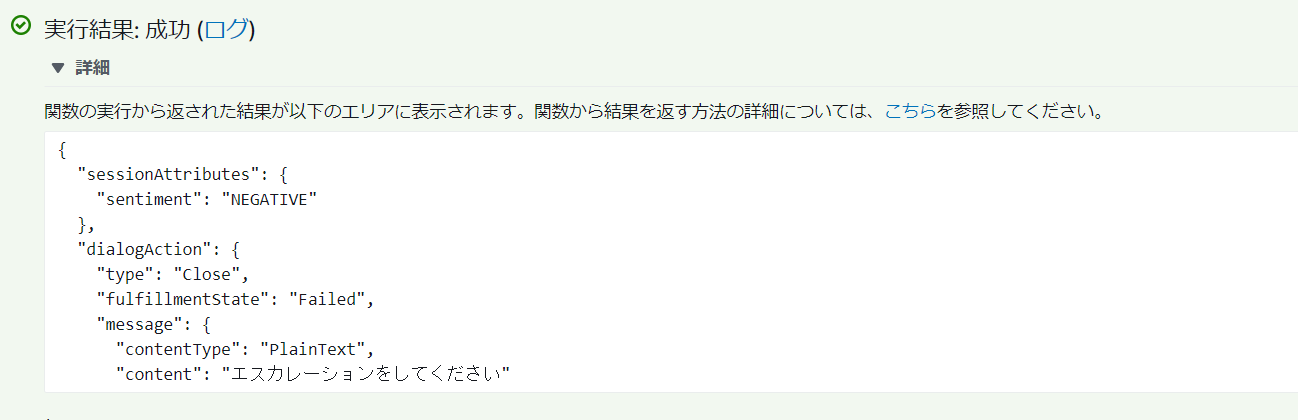
5-2.　[comprehendlambda.txt]の中身をコピペします。

このサンプルコードは、ComprehendがNegativeなメッセージであると判断した場合、"エスカレーションをしてください"というメッセージが出力されます。



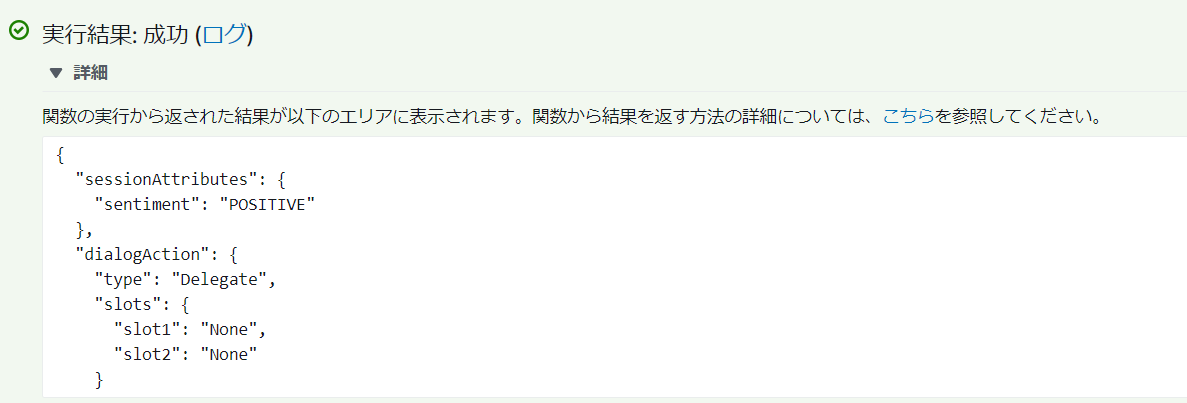
5-3.　画面右上の「保存」を押します。その後「テスト」を押し、イベント名に適当な名前を付けた後、[comprehendlambdatest.txt]をコピペします。

5-4．テスト文字列には「あなたを攻撃します」と記載されていますが、一度そのままテストを実行します。以下のような画面が出力されます。

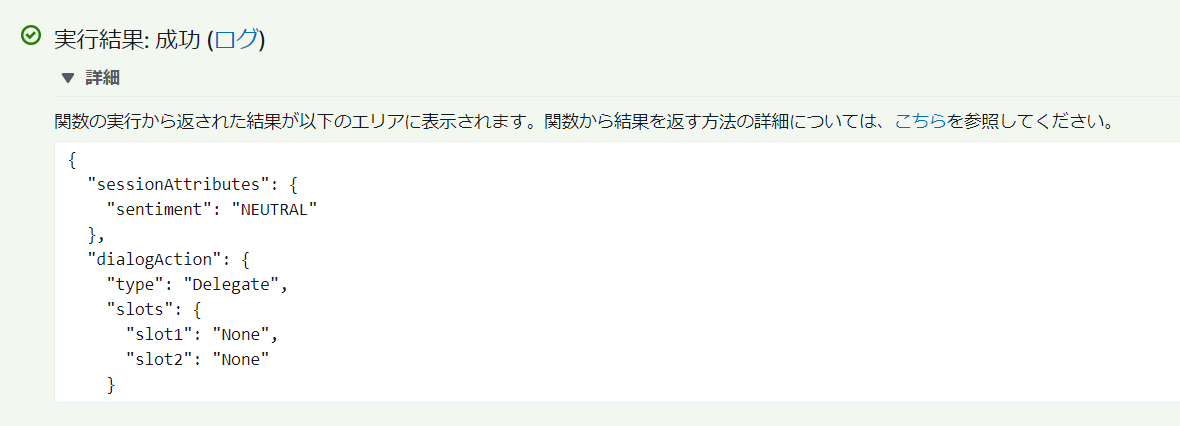


5-5.　先ほどのテストで使った文字列を「あなたが好きです」に変えて再度テストを実行してください。修正するためには、以下の箇所から「テストイベントの設定」を選択することで、先ほど設定したテストを上書きすることができます。

5-6.　テストを実行すると分析結果が「Positive」に代わり、エスカレーションがされなくなりました。

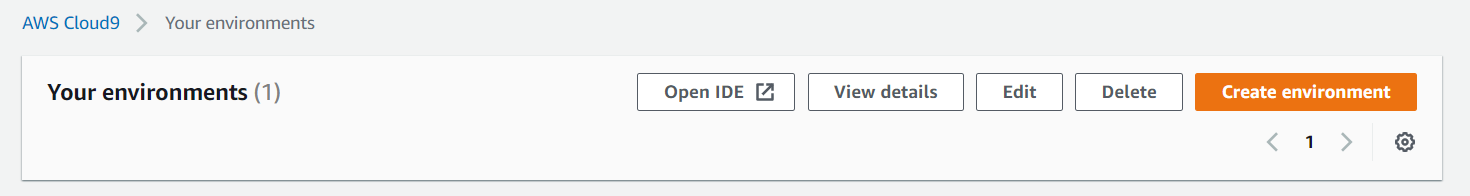


5-7.　では先ほどTranscribeで文字お起しした文字列に入れ替えて再度テストをしてみてください。ほとんどのニュース記事の場合、例え内容が悲惨な事件などであっても、その論調は感情を持たせず客観的に記載されていますので、[Neutral]と出力されるケースが多いようです。

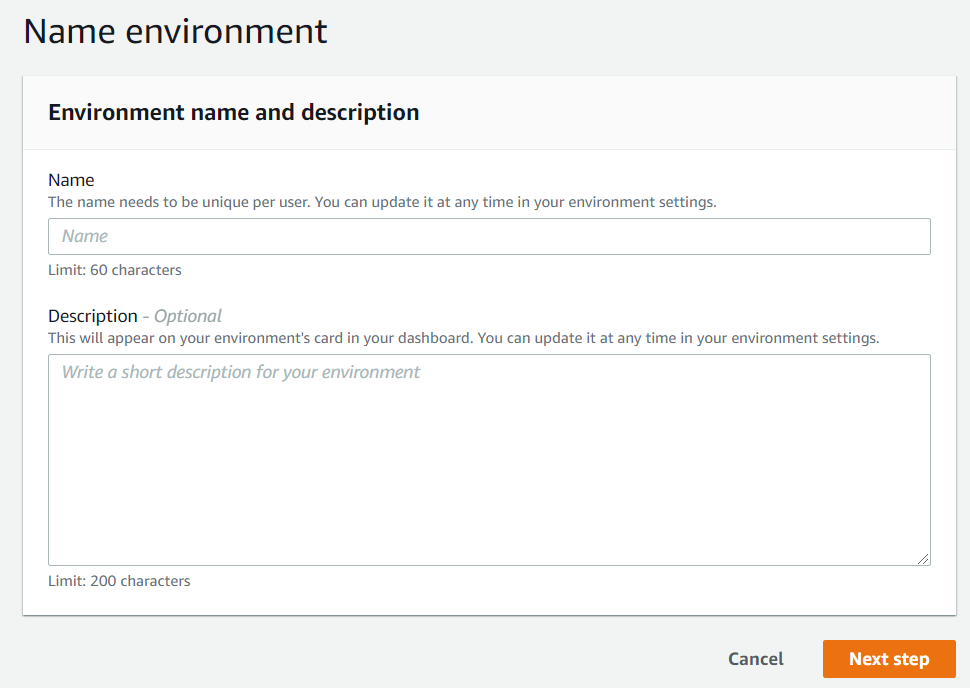


6.　TranscribeとComprehendを連携させ、音声の分析結果をElasticsearch serviceのkibanaで可視化する。**以下の作業はすべてオレゴンで行います。Cloud9は起動した後、個別のIAMロール設定が可能ですが、この手順ではデフォルト権限設定を利用するため、Comprehend、Transcribe、Elasticsearch Service、S3、Cloud9へのフル権限を持っているIAMユーザーで管理者画面にログインして作業を行ってください。**

6-1.　Cloud9を起動するため、Cloud9の管理者画面にいきます。

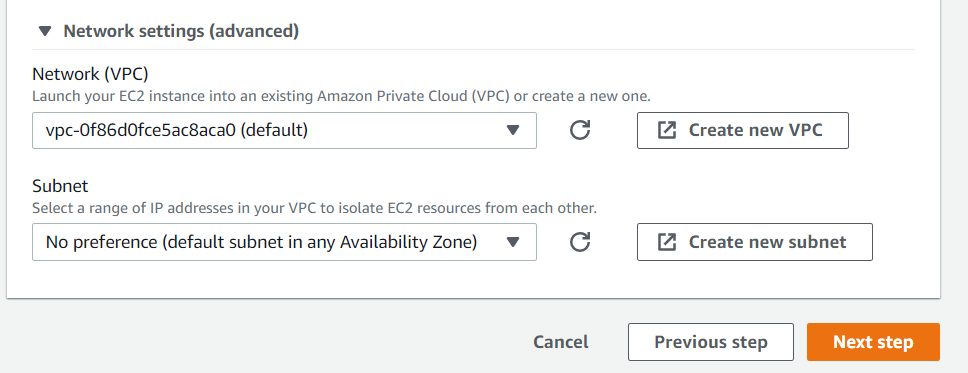


6-2.　「Create Environment」を押します。[Name]にはYYYYMMDDhandson（YYYYMMDDは本日の日付）と入力し、[Next Step]を押します。

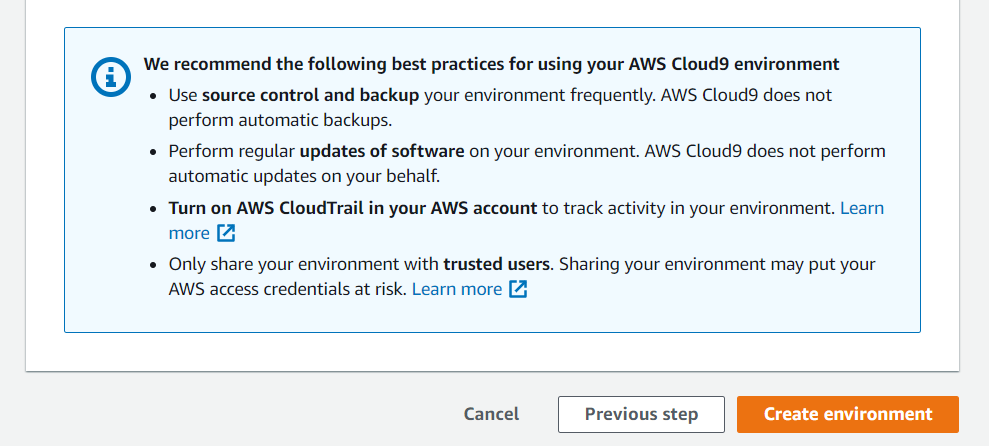


6-3.　画面下の[Network Settings]を以下の様に伸ばし、パブリックサブネットを持つVPCとパブリックサブネットを選択します。不明な場合は、チューターに聞いて下さい。別タブでVPCの管理者画面を開くと確認しやすくなります。

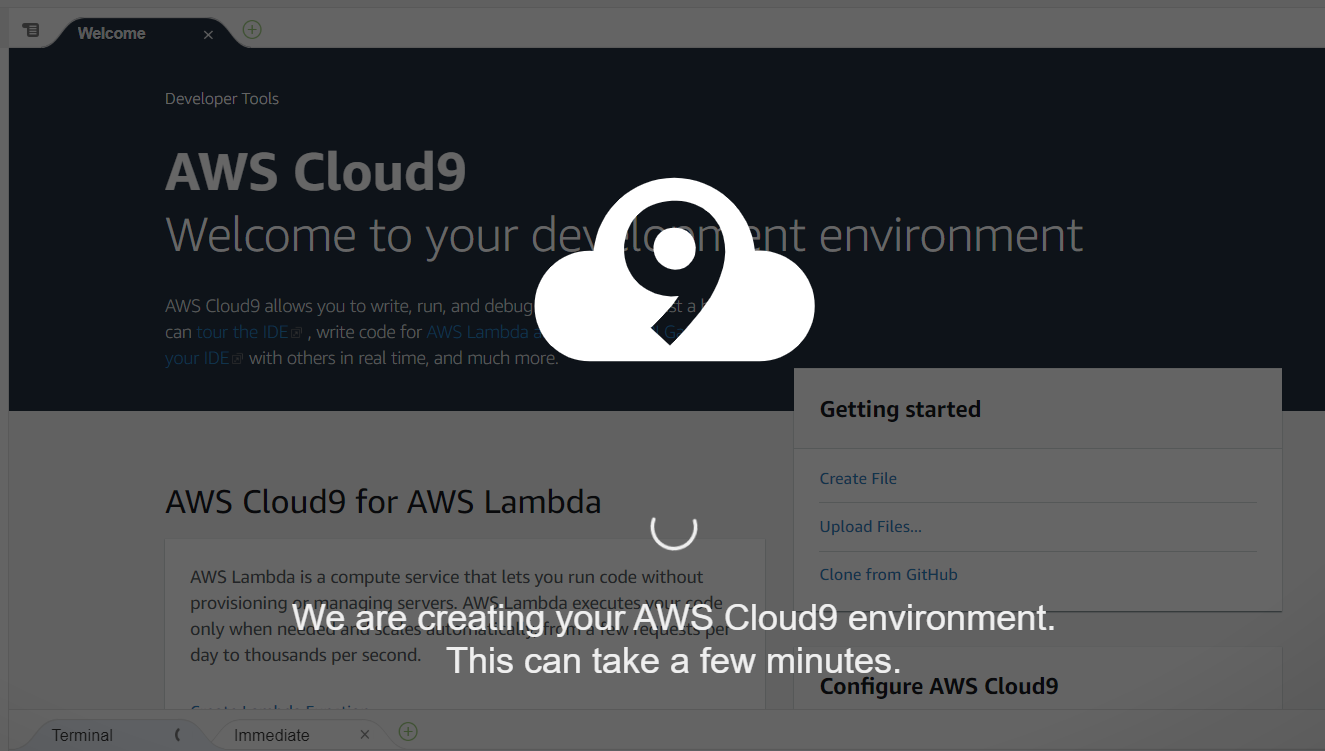
VPC周りの設定に慣れてらっしゃらない方は、以下の画面のようにdefaultを選択してください。



6-4.　「Next Step」を押します。



6-5.　「Create environment」を押すと以下の画面に切り替わります。



6-6.　起動は数分間かかりますが、完了するとコンソールが使えるようになります。

6-7．オレゴンにS3バケットを作成するために、以下のコマンドをタイプして下さい。

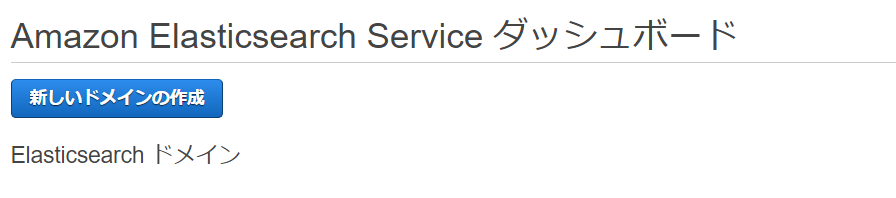
aws s3 mb s3:*//****<yyyymmdd>transcribe<name>*** *--region us-west-2*

(YYYYMMDDは本日の日付、nameはお名前)

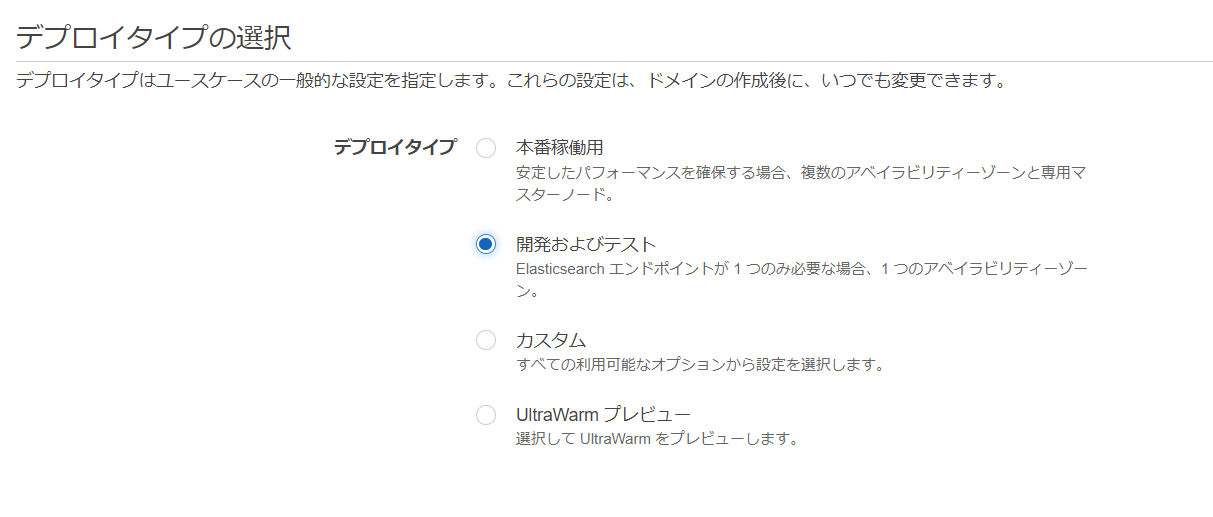


6-8.　念のためS3の管理者画面でS3バケットが正しく作られているか確認します。表示されない場合、再読み込みをしてみましょう。

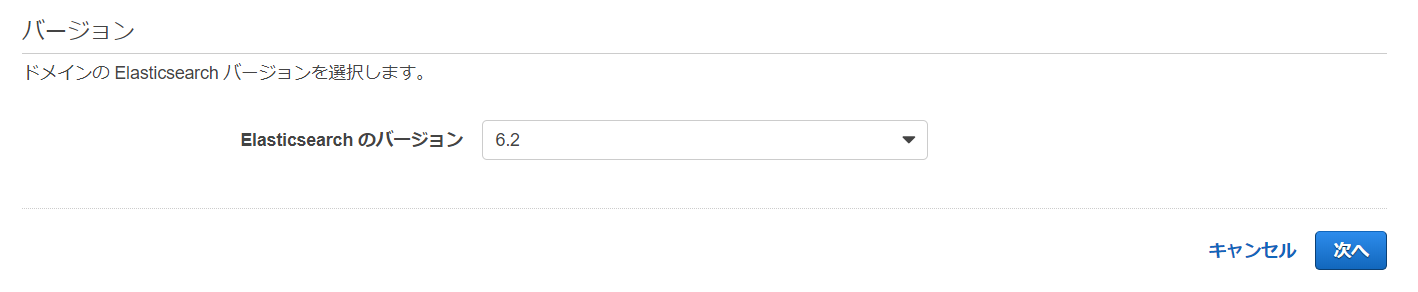
6-9．　オレゴンにElasticsearch serviceを起動するためElasticsearch serviceの管理者画面に遷移します。



6-10.　「新しいドメインの作成」を押します。

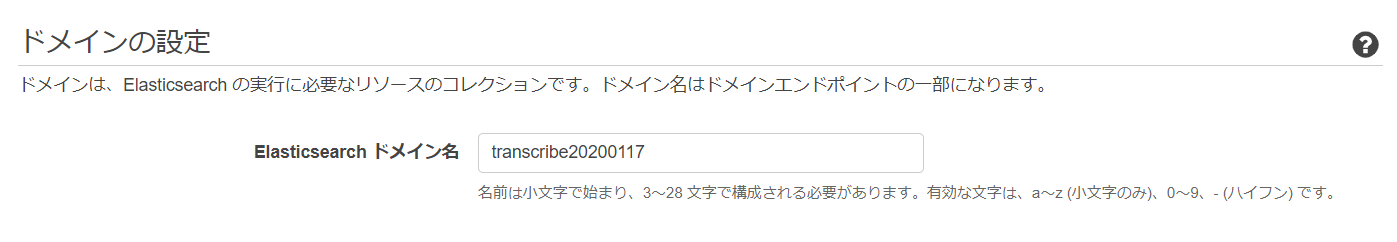


6-11.「開発およびテスト」を選びます。



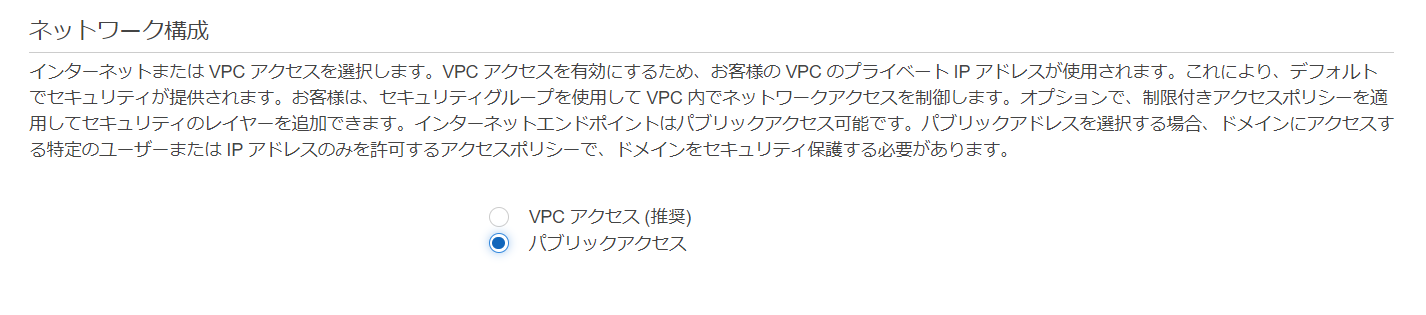
6-12.「6.2」を選んで「次へ」を押します。

6-13. ドメイン名に[transcribetestYYYYMMDD](YYYYMMDDは今日の日付)と入力します。



6-14．「次へ」を押します。

6-15. ネットワーク構成で「パブリックアクセス」を選びます。

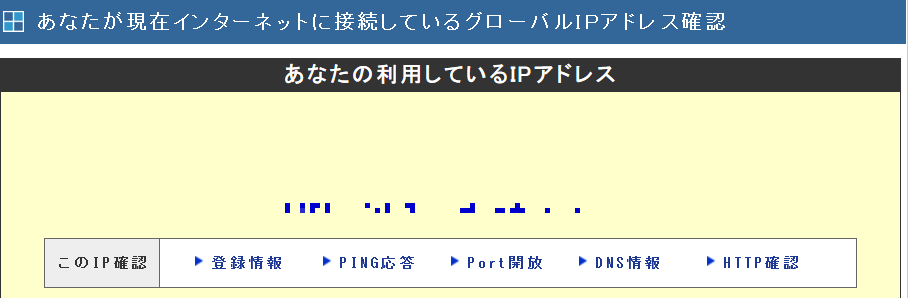


6-16.　「アクセスポリシー」を以下のように特定ソースIPv4アドレス(※1)から許可、特定IAM ARN(※2)から許可に限定するよう作成します。



※1:IPv4アドレスは以下サイト等からアクセス元IPアドレスを確認して入力して下さい。

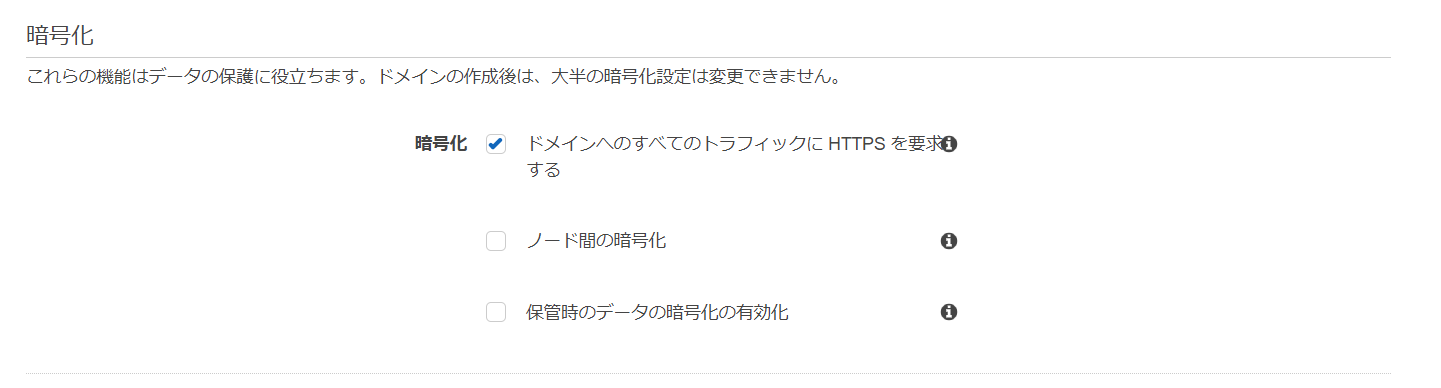
<https://www.cman.jp/network/support/go_access.cgi>



※2:IAM ARNはIAMサービスを開き、現在利用しているAdmin相当のIAMユーザもしくはIAMロールのARNを開いてコピー、ペーストします。



6-17.　暗号化はそのまま設定します。



6-18.　「次へ」を押します。最後に確認画面が表示されるので「確認」を押します。

6-19.　現在作成中です。作成が完了したらエンドポイント及びKibanaのURLが表示されますので、しばらくまち、コピーしておきます。（別のタブでCloud9の画面に戻って次のStepに進んでも問題ありません）



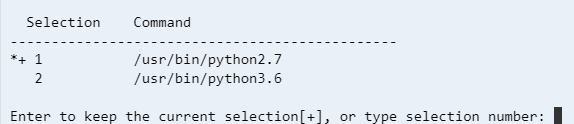
6-20．以下2つのコマンドをCloud9で実行します。コピペすると正常に動作しないケースがありますので。手で入力します。

python –-version

pip –-version

ハンズオンを行う時期によって出力が異なりますが、以下のようにPythonはバージョン３、pipがバージョン2を指している場合、以下のコマンドを実行して、pipの向き先を変更します。

sudo update-alternatives --config python



2を選びます。これでPython3環境の整備が終わりました。

6-21．　必要なライブラリをインストールするため、以下の3つのコマンドを実行します。赤字で「Permission denied」と表示された場合、先頭にsudoをつけて実行してください。

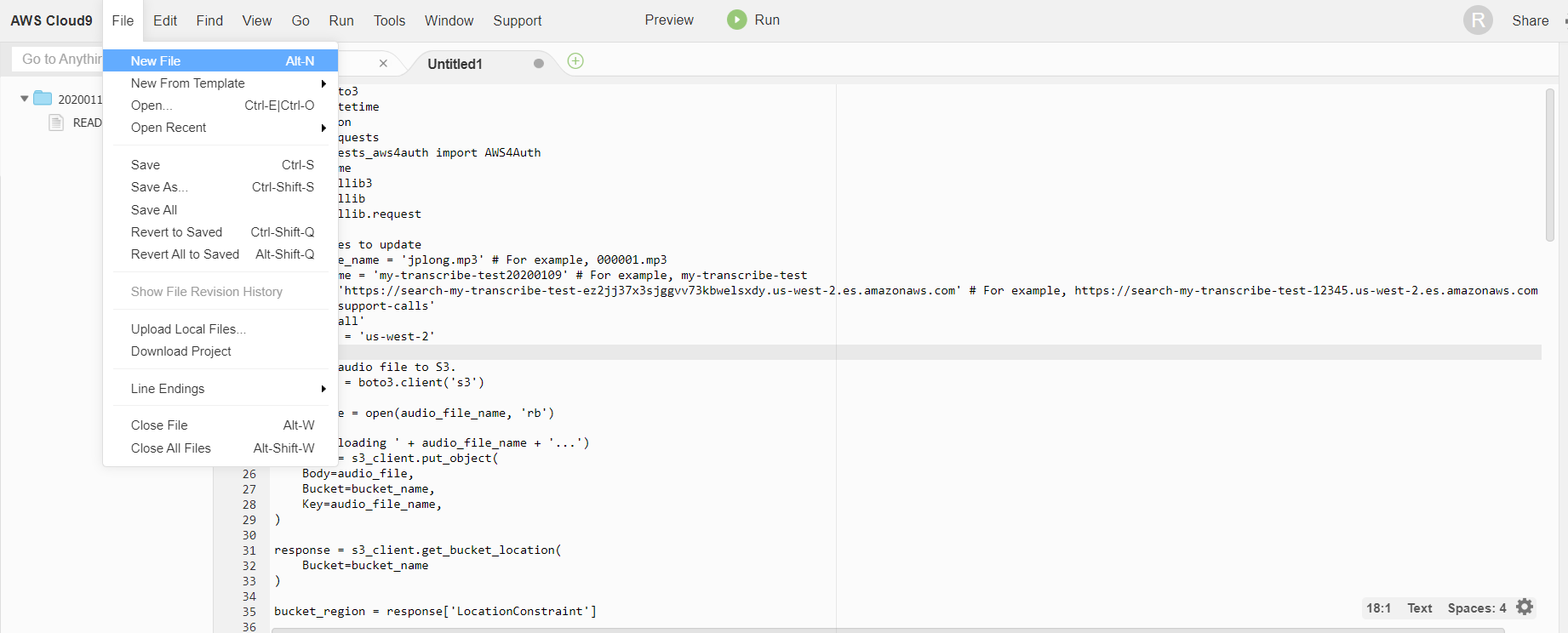
pip install boto3

pip install requests

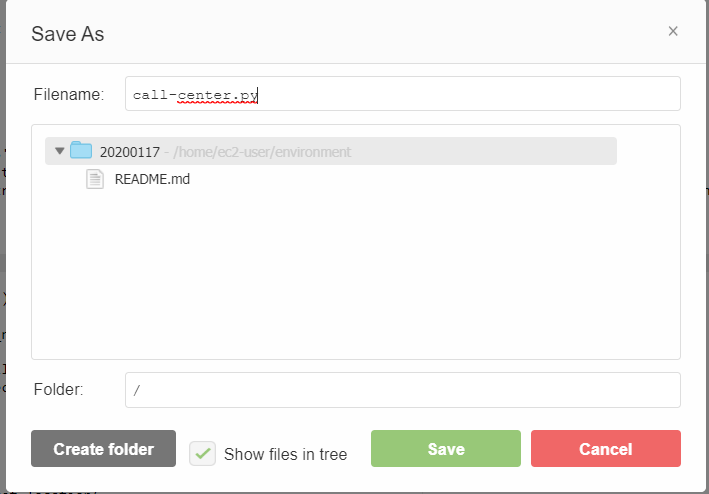
pip install requests\_aws4auth

（エラーとなる場合は、sudo pip install requests\_aws4authとしてください）

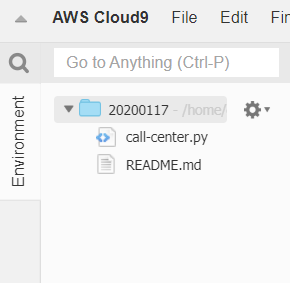
6-22.　必要なライブラリのインストールが完了しましたので、File→New Fileと選んで、新しく現れたタブにcall-center.pyの中身をコピペします。



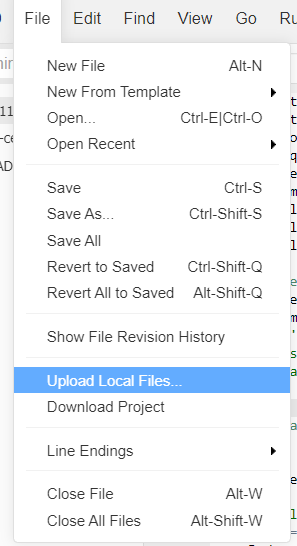
6-23.　File→Saveと押して、Filenameに[call-center.py]と入力し「Save」を押します。



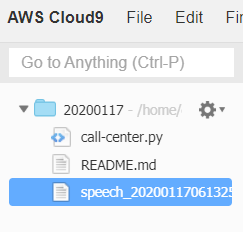
6-24. 以下のフォルダ構成となります。



6-25.　File→Upload Local Filesを選び、先ほど作成したmp3をアップロードします。



6-26.　Upload画面は自動で閉じませんので、右上のバツを押して閉じます。以下のフォルダ構成となります。



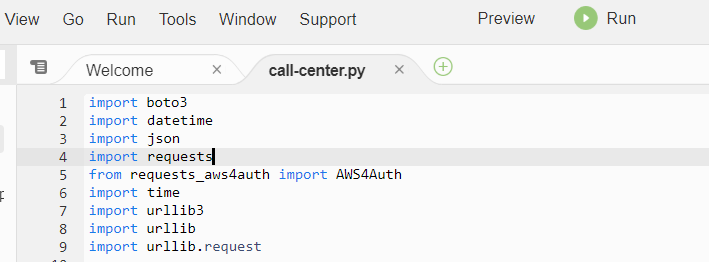
6-27.　ソースコードを修正します。12行目、13行目、14行目を修正します。

12行目：アップロードしたmp3ファイル名

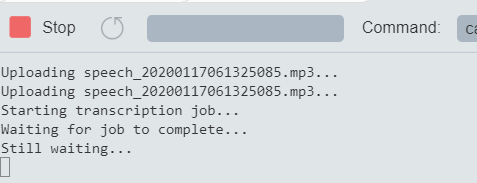
13行目：オレゴンに作成したS3バケット名

14行目：Elasticsearch serviceのエンドポイント

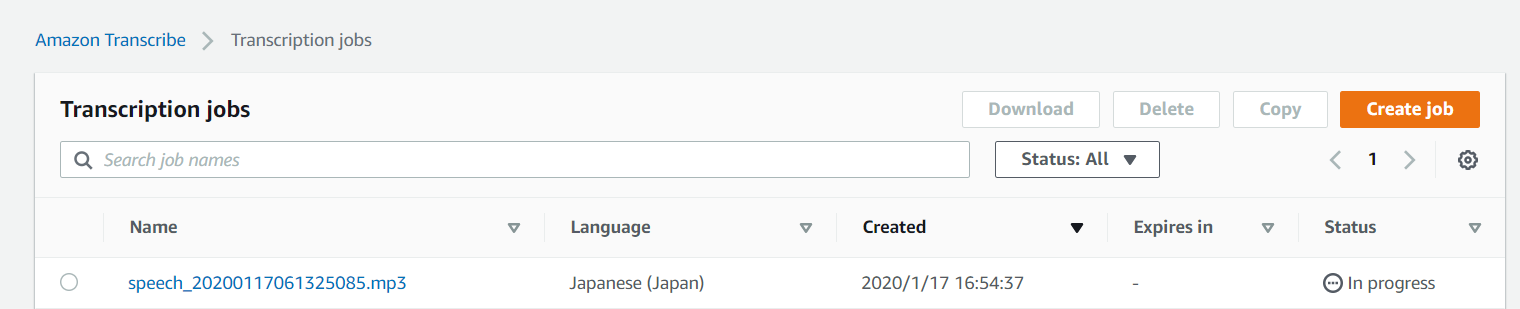
6-28.　ファイルをFile→Saveしたらいよいよ実行です。画面上のRunボタンを押します。



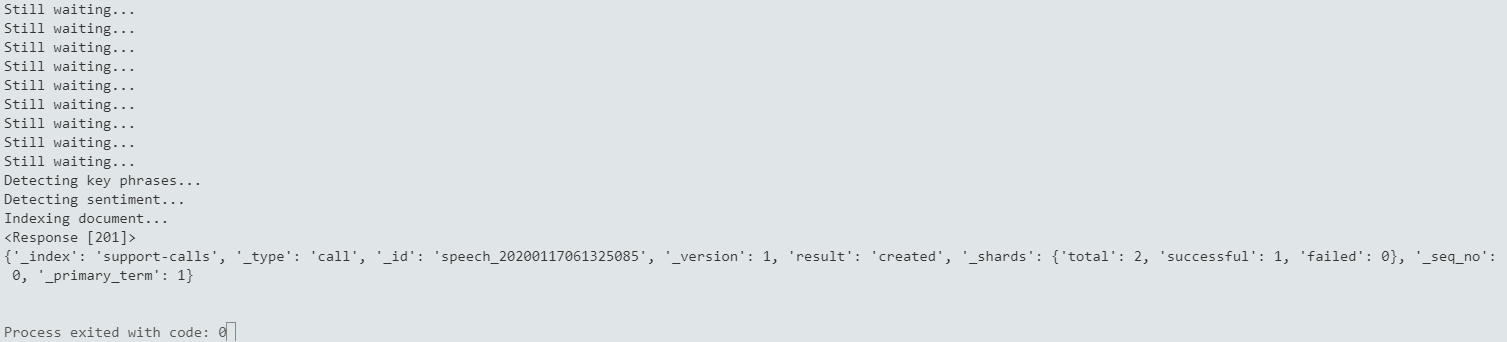
6-29.　以下の画面が表示されれば正しく動作しています。



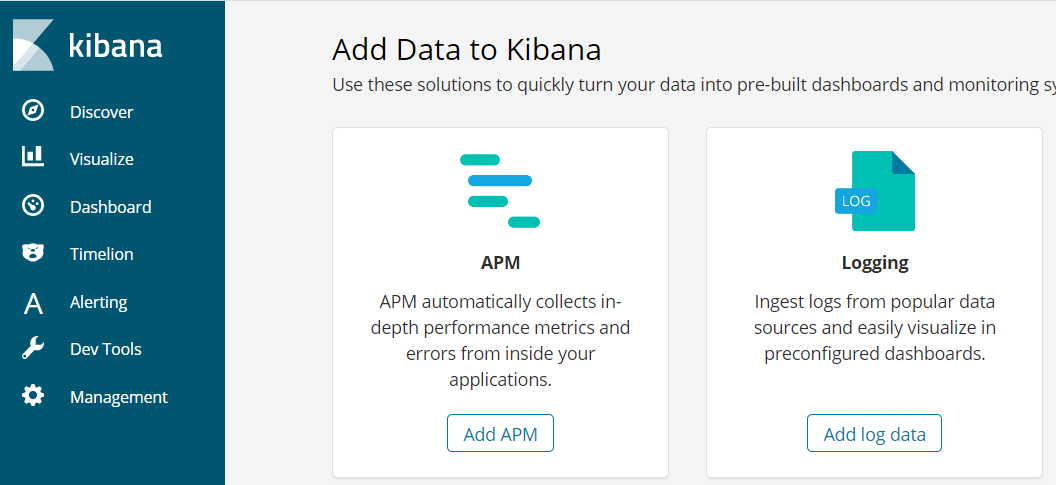
6-30.　オレゴンリージョンのTranscribeの画面に行きます。文字起こしが開始されています。



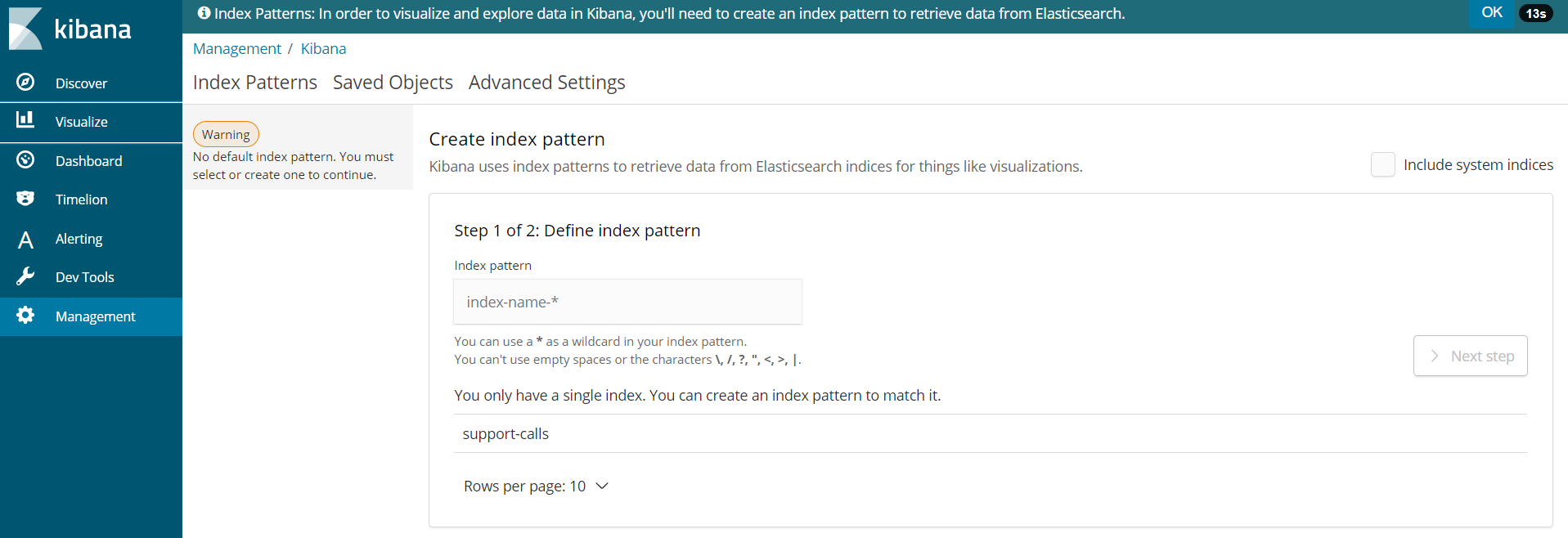
z6-31.　以下のような表示がされれば実行は完了ですTranscribeが文字起こしを行い、その後ComprehendがKeyword分析を行いElasticsearch Serviceにデータが投入されています。



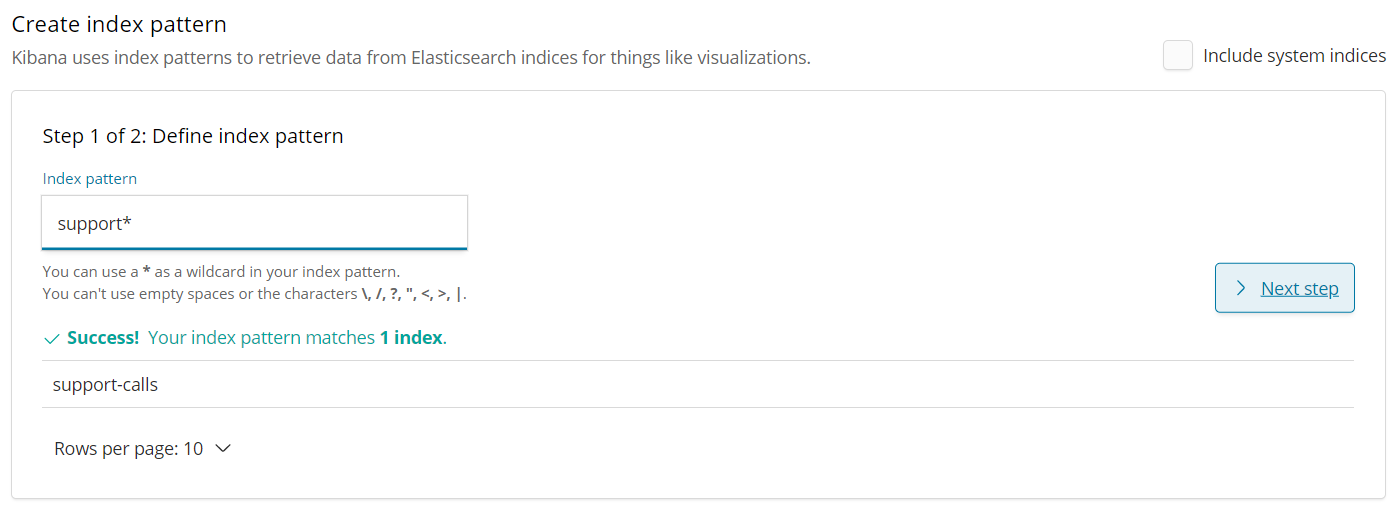
6-31.　先ほどコピーしたKibanaのURLにブラウザでアクセスします。



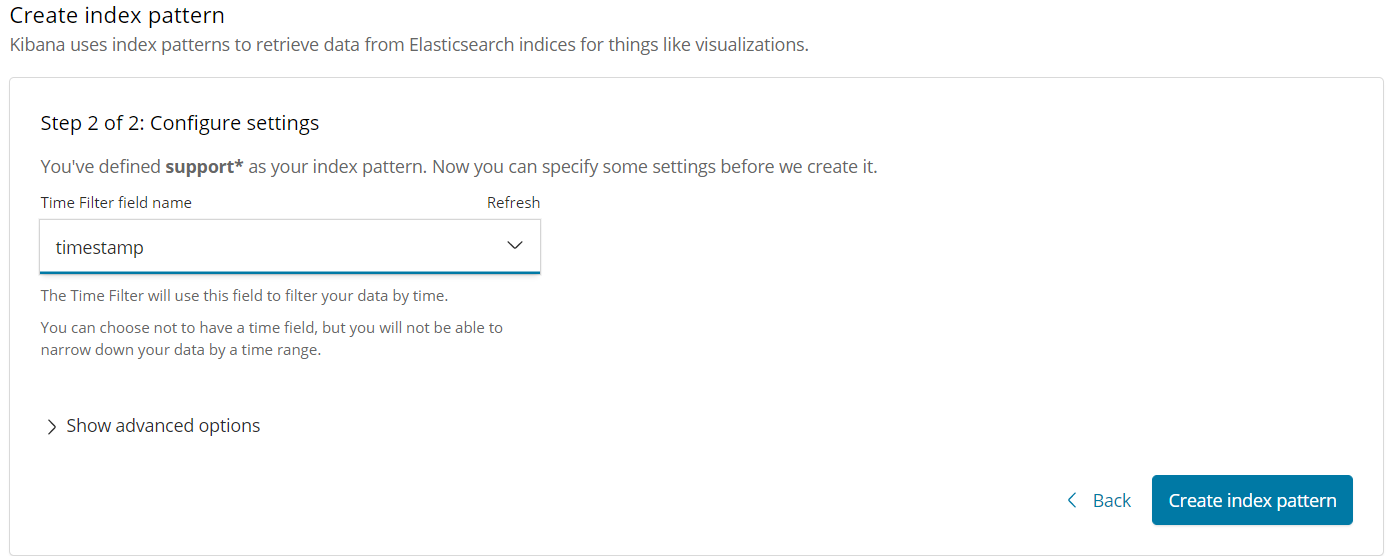
6-32.　「Visualize」を押します。



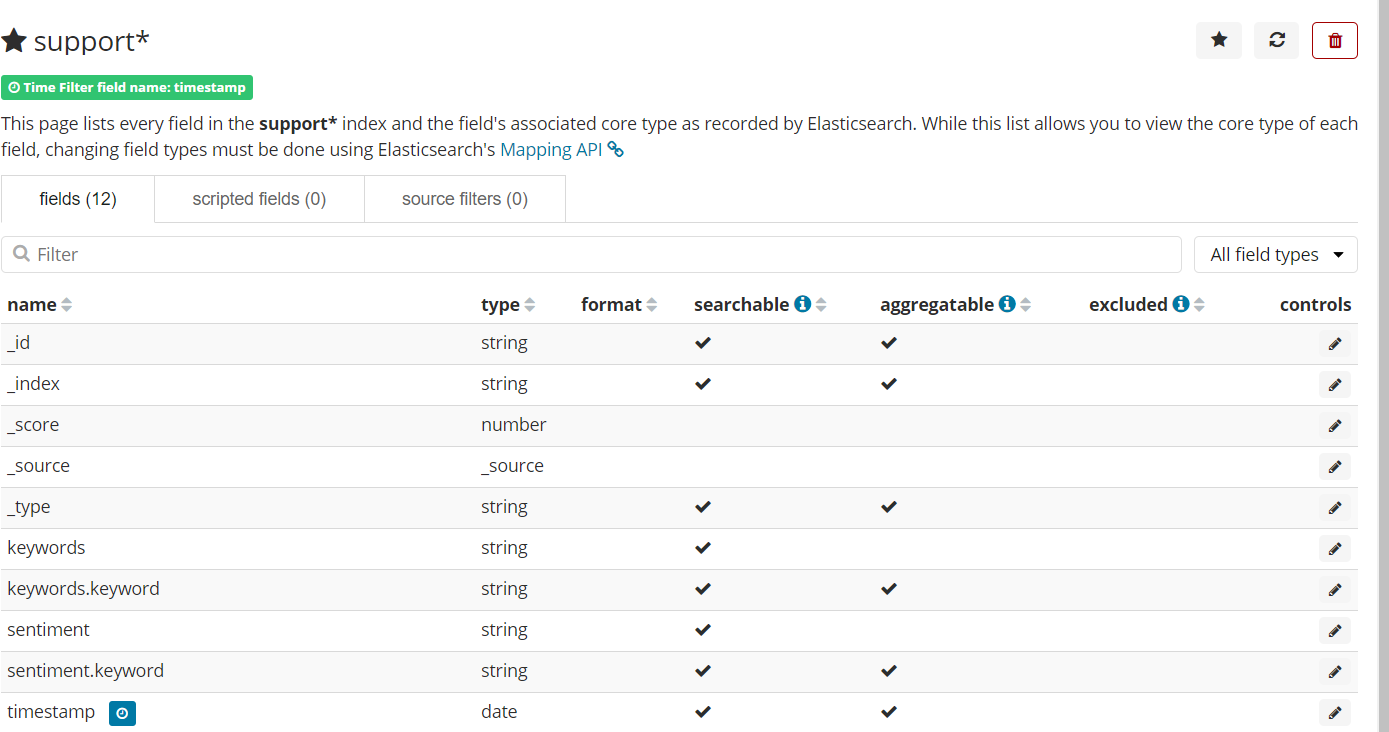
6-32.　[Index Pattern]に[support\*]と入力し[Next Step]を押します。



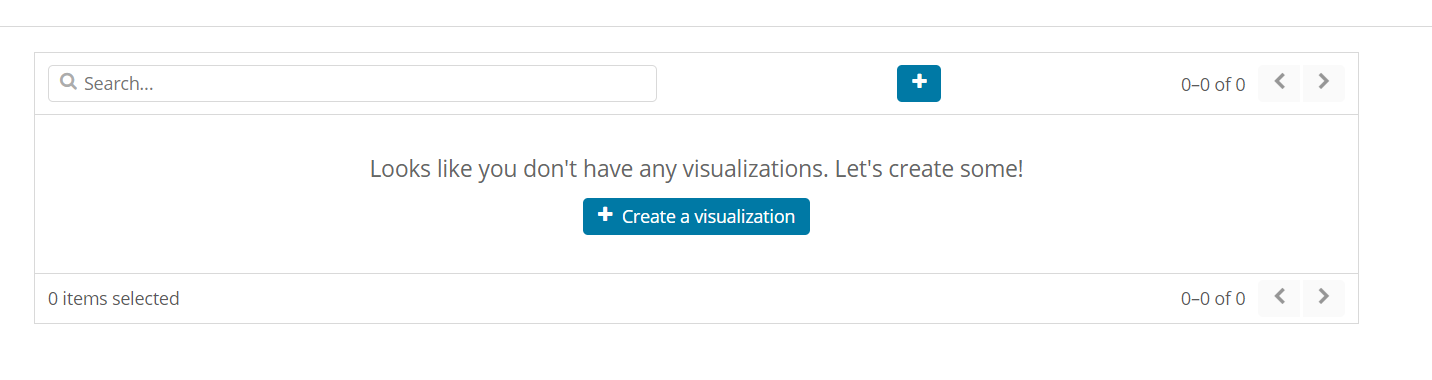
6-33. [Time Filter field name]にtimestampを選び[Create Index Pattern]を押します。



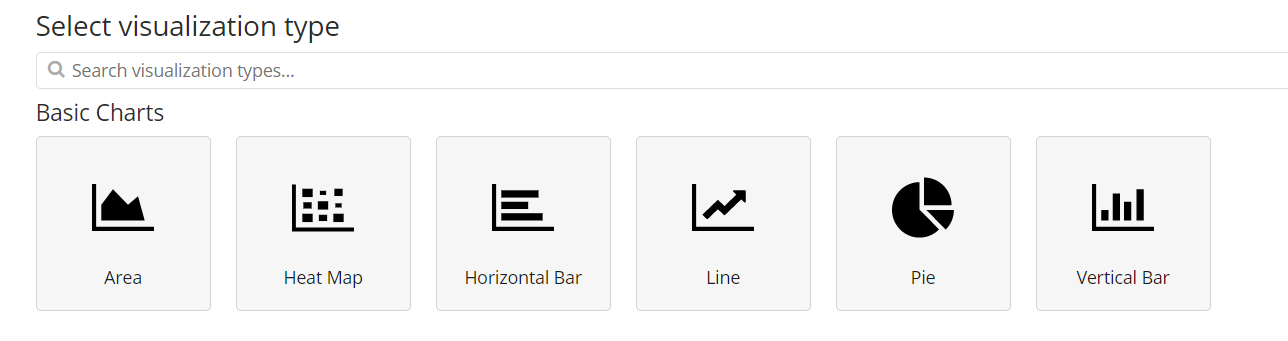
6-34.　正しくデータが投入され、設定も正しければ以下の画面が表示されます。



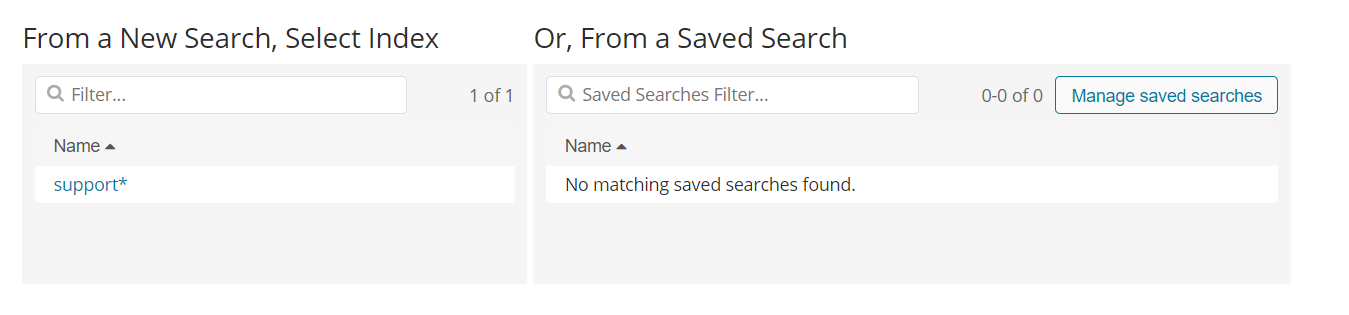
6-35.　もう一度Visualizationを押します。



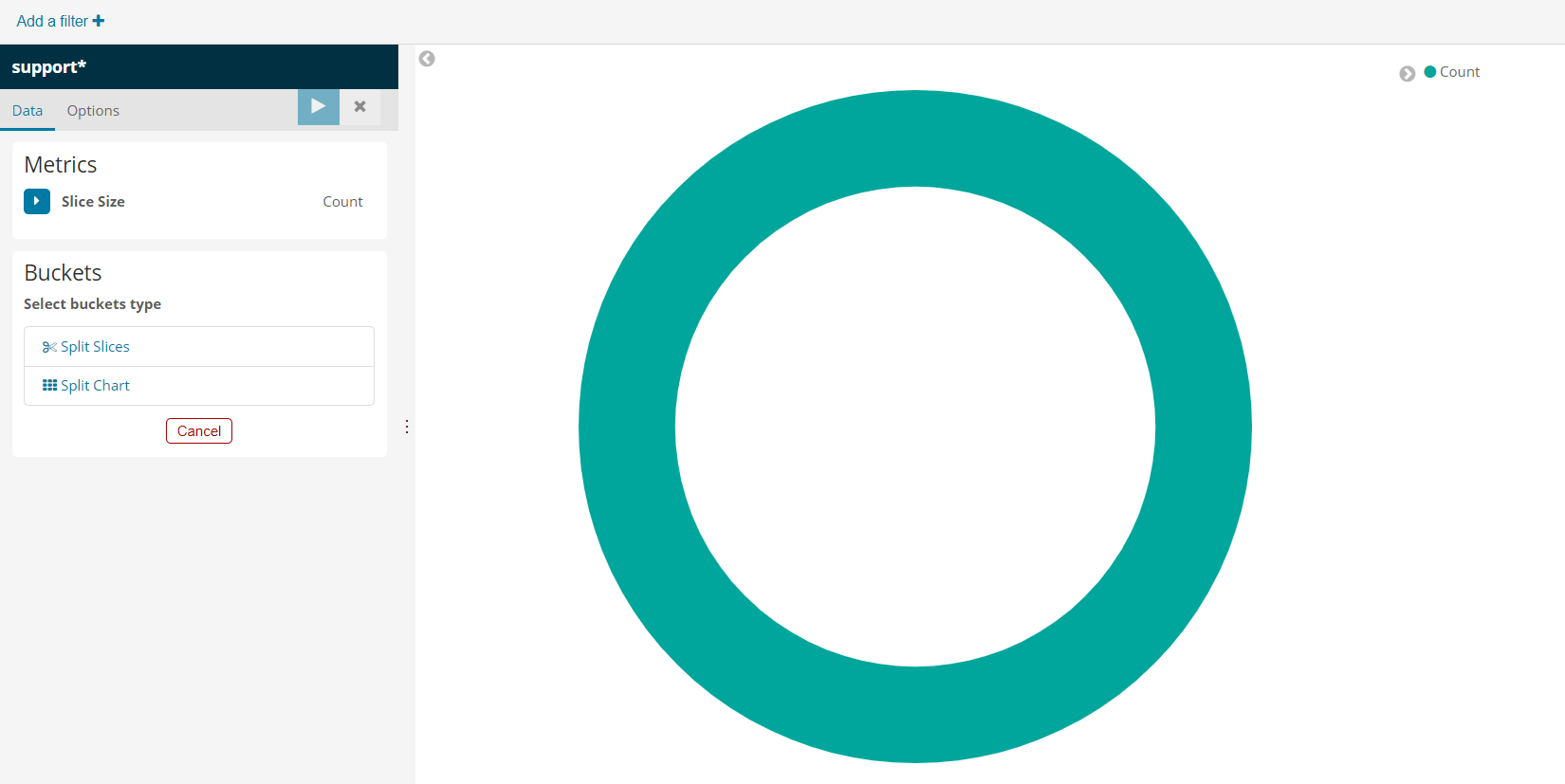
6-36.　[Create a visualization]を押します。



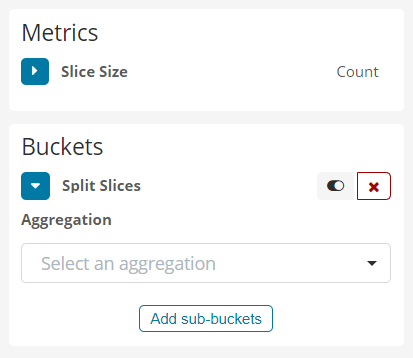
6-37.　円（Pie）グラフを選びます。



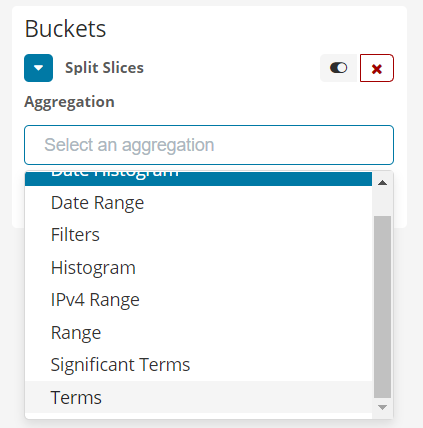
6-38.　先ほど作成した[support\*]を選びます。



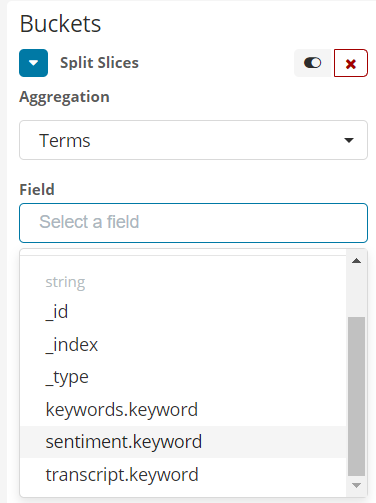
6-39.　Bucketsから[Split Slices]を選びます。



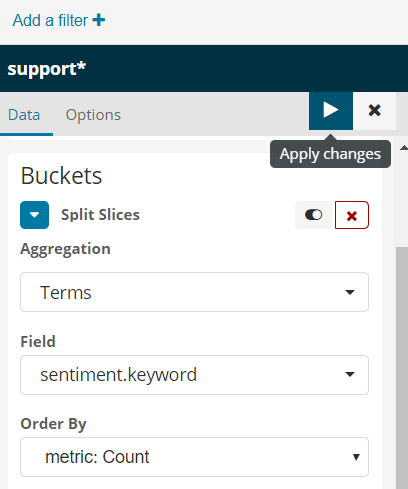
6-40.　[Aggregation]で[Terms]を選びます。



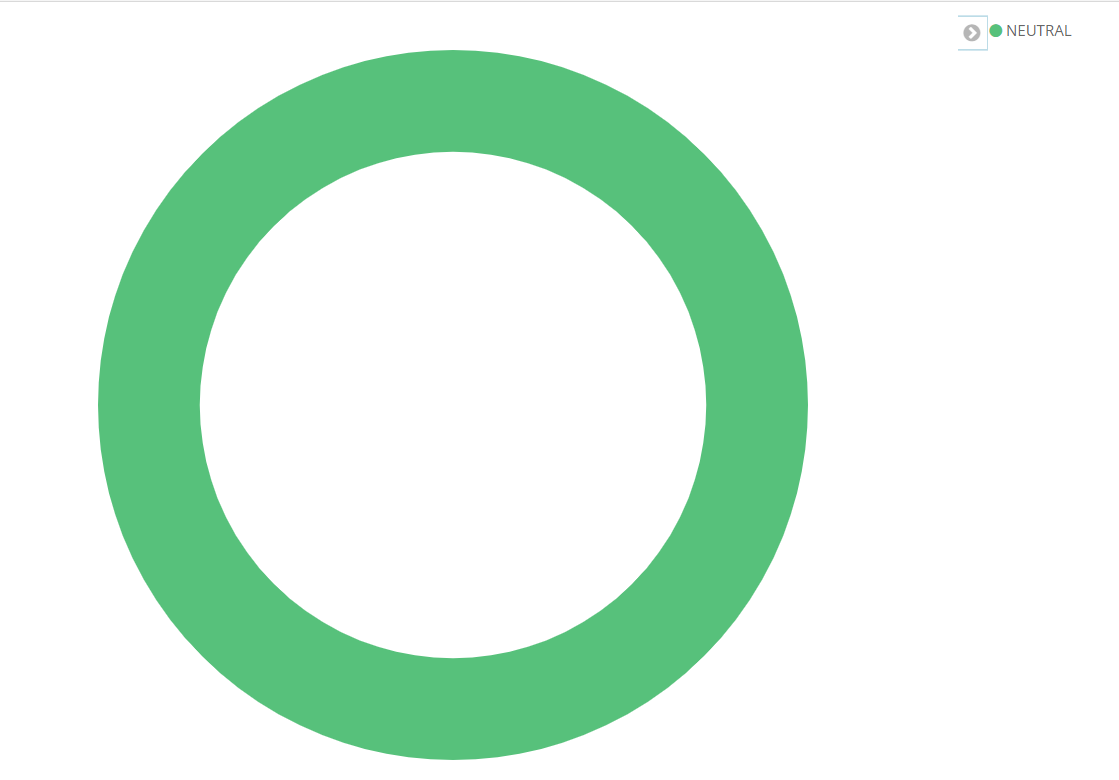
6-41. [Filed]で[sentiment.keyword]を選びます。



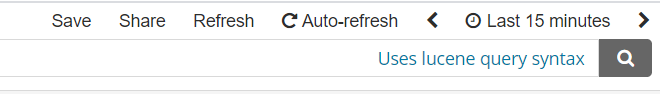
6-42.　[apply changes]を押します。



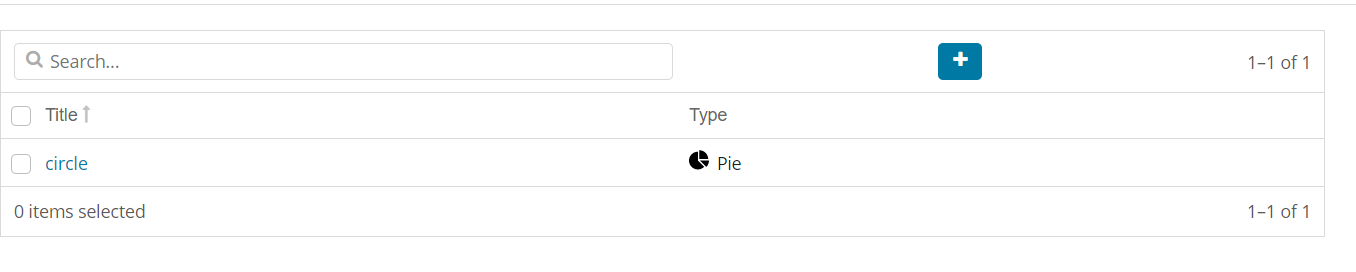
6-43.　先ほど分析した文字列の結果（Positive,Neutral,Negative）が表示されます。まだ1件しか分析結果がありませんが、複数分析を行えば増えていきます。



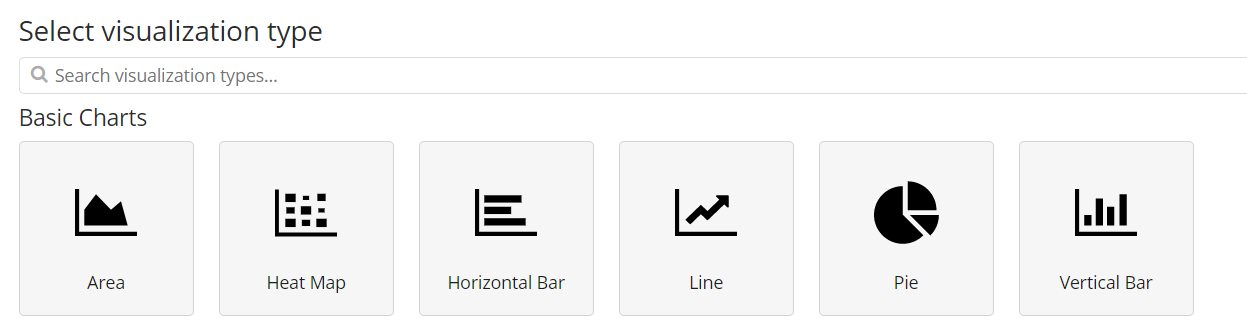
6-44. 右上の[Save]で適当な名前で保存しておきます。



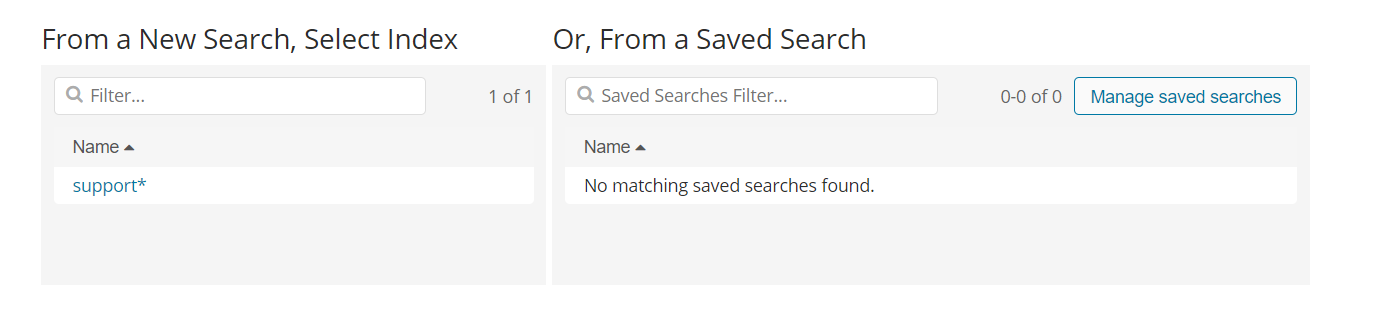
6-45.　再度[Visualize]を押します。



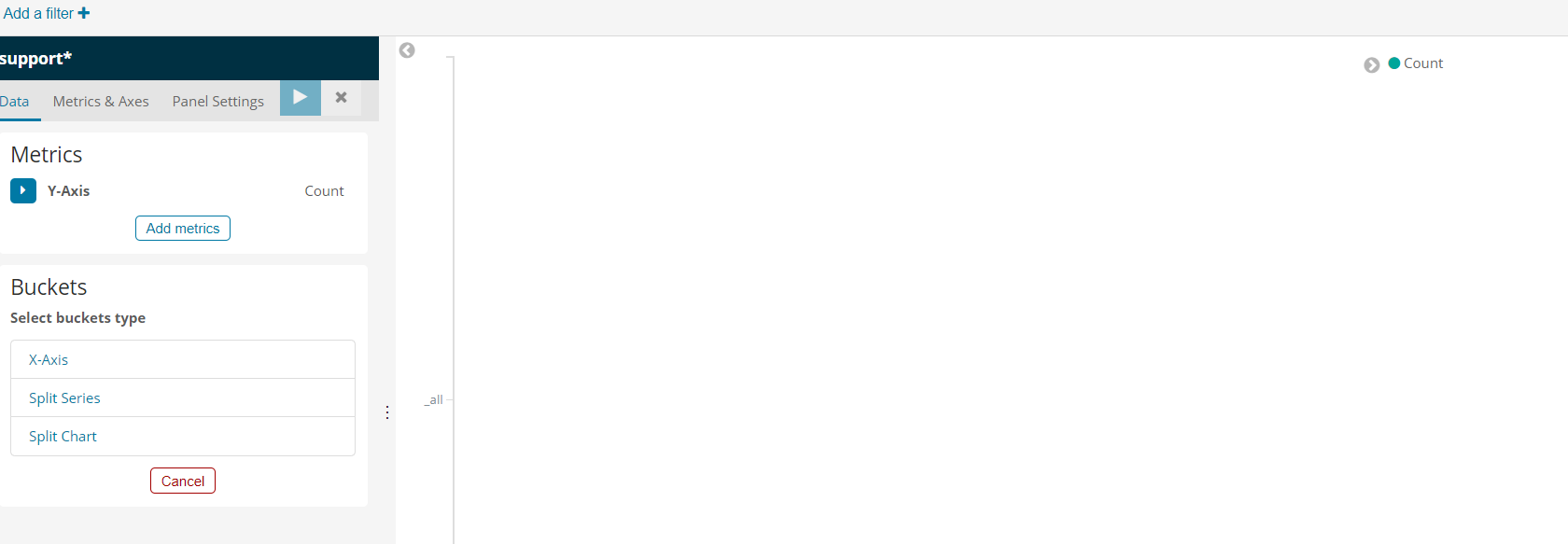
6-46.　先ほどのものが保存されていますが、もう1つ作成するので+ を押します。



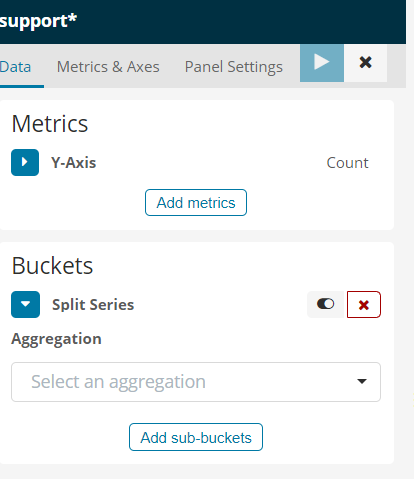
6-47.　今度はHorizontal Barを選びます。



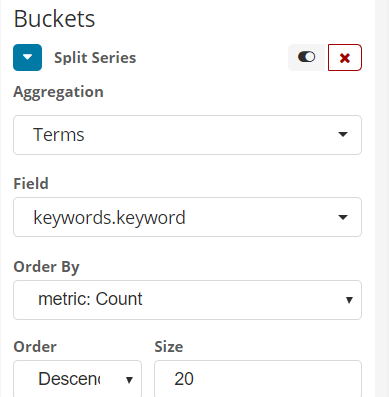
6-48.　また[support\*]を選びます。



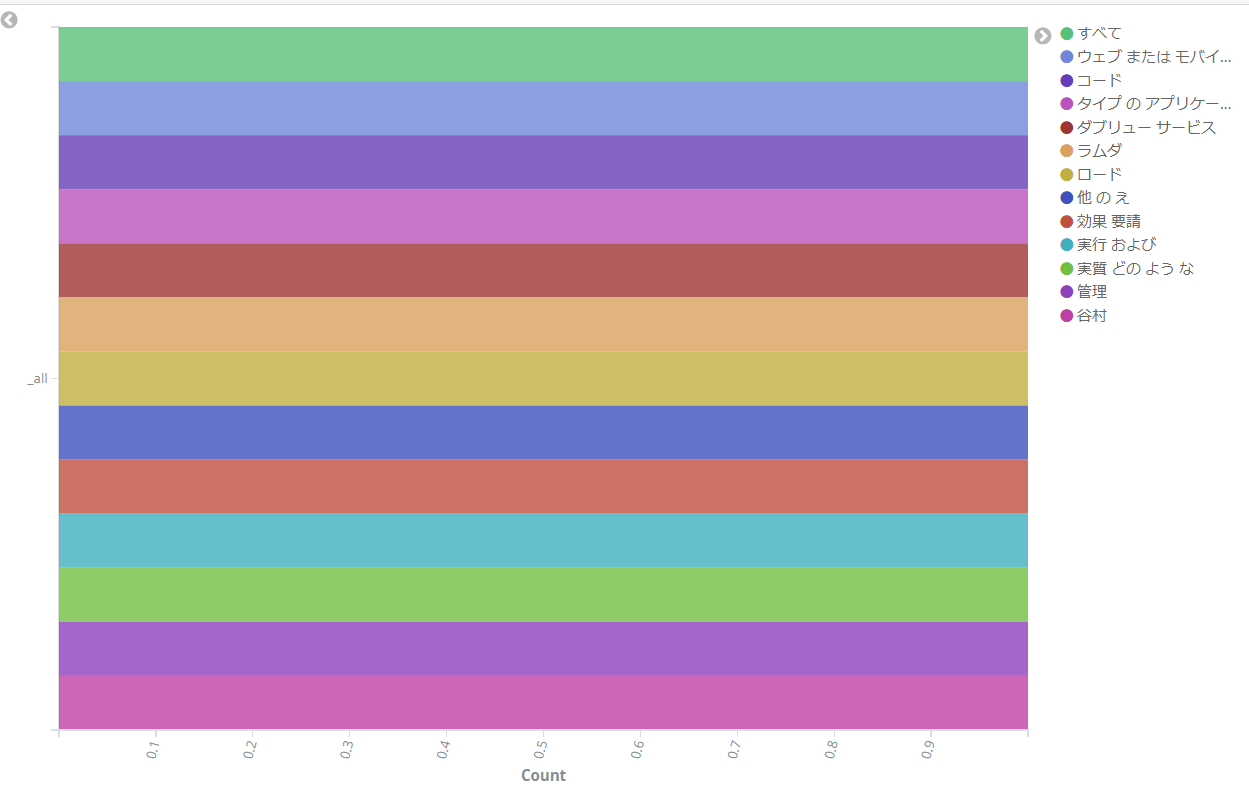
6-49.　今度は[Split Series]を選びます。



6-50.　以下の画面のように設定し、Apply changesを押します。



6-51. 以下のように、取り込んだ単語が件数毎にグラフ化されます。



まだ1件しか取り込んでいないため、全て1件ですが、複数回繰り返すうちにグラフ生成されていきます。

７．　リソースの削除

本日生成したElasticesearc Service,Cloud9,S3バケットを削除してください。また以下は残しておいても料金は発生しませんが、気になる方は合わせて削除をお願いします。

おつかれさまでした！