NEC Advanced Analytics Platform V1.5
マネージドサービス
サイジングガイド

第 1.1 版

日本電気株式会社

改版履歴

版	作成日	変更内容
1	2021/09/24	「NEC Advanced Analytics Platform V1.3 マネージドサービスサイジングガイド」(AAPF V1.3 サイジングガイド)をベースに、AAPF V1.5 のサイジング資料として改版し、初版作成。
1.1	2021/12/08	RAPID 機械学習 マッチング版のサイジング手順を追加。

目次

1　本文書について

本文書は、利用者向けに事業者が提供する「NEC Advanced Analytics Platform V1.5 マネージドサービス」（以下、AAPF マネージドサービス）のサイジング手順を記載した文書です。

本文書は、秘密保持対象ドキュメントとして、事業者の許可なくコピーおよびその配布、ホームページへの掲載を禁じます。

また、事業者は本利用ガイドについて随時変更することができるものとします。

本文書で取り扱う AAPF マネージドサービスで利用するソフトウェアは、NEC Advanced Analytics Platform V1.5（以下 AAPF）です。AAPF 上の操作、また AAPF に関する制限事項や留意事項については、AAPF マニュアルをご参照ください。

本文書および AAPF マニュアルに記載の手順は、テナント利用者の環境や利用状況により、記載通りに利用できない可能性があります。環境に応じて手順を変更の上、ご利用ください。

1.1　本文書の対象読者

本文書の対象読者は、AAPF マネージドサービスの利用申請(契約申し込み)を行うにあたって、契約プランの検討を行う方を想定しています。以下の知識をお持ちであることを前提としています。

・使用する予定の分析エンジンの概要を把握していること。(使用する予定の分析エンジンのマニュアルの「Getting Started」を既にお読みであるか、同等の知識を有していること。)

・データ分析に関する一般的な知識を有していること。

1.2　本文書の位置付け

本文書は、AAPF マネージドサービスにて提供するプラン（エントリープラン、エントリープラスプラン、スタンダードプラン）を検討する方法として、各リソース(CPU コア数、メモリー、ストレージ容量)観点で記載しています。本文書により、分析環境を利用するために必要な各リソースを見積もり、用途に適したプランを選択することを支援します。CPU コア数、メモリー、ストレージ容量以外の個別のご要件がある場合は、必要なリソースの見積もりに関わらずスタンダードプランのご契約でないと実現できない場合がありますのでご注意ください。各契約プランの仕様の詳細については、「NEC Advanced Analytics Platform マネージドサービス（エントリー、エントリープラスプラン）サービス仕様書」、「NEC Advanced Analytics Platform マネージドサービス（スタンダードプラン）サービス仕様書」を参照してください。

1.3　本文書の改版

本文書の見直しは、AAPF マネージドサービスの提供内容の変更に伴い実施します。

1.4　用語集

本書で使用する各種用語は下表のとおりです。

用語	説明
1DCNN	1 次元畳み込みニューラルネットワーク（1-Dimension Convolutional Neural Network）の略。 ニューラルネットワークの手法であり、画像解析に用いられる CNN を、時系列数値データ解析用にカスタマイズした手法。RAPID 機械学習(時系列数値解析版)で使用できるアルゴリズムの一つ。
1DDCN	1-Dimension Deep Convolutional Neural Network の略。 1D CNN よりもさらに深いネットワークでの高精度な学習を可能にする。RAPID 機械学習(時系列数値解析版)で使用できるアルゴリズムの一つ。
1DOCN	1-Dimension OneClass Neural Network の略。 正常データのみによる学習を可能にする。RAPID 機械学習(時系列数値解析版)で使用できるアルゴリズムの一つ。
文クラスタリング	文書中の 1 文（程度）に相当するテキストの集合を入力として、類似したテキストを同じグループ（クラスタと呼ぶ）にまとめ上げ、クラスタの集合を生成する機能です。テキスト分析エンジンで使用できる機能の一つ。
文書判別	機械学習の手法を使って、未知の文書を決められたカテゴリに分類する機能です。テキスト分析エンジンで使用できる機能の一つ。
文数	テキスト分析における分析対象の文の数。本書では 1 文あたり 128 バイトを想定しています。
文書数	テキスト分析における分析対象の文書(テキストファイル)の数。本書では 1 文書あたり 1480 バイトを想定しています。
二値展開	カテゴリ変数(性別、職業など一般に数や量で測れない変数)を値として持つ属性の各属性値を 0 または 1 の二値に変換すること。
AACluster	AAPF マネージドサービスが提供する仮想分析環境(コンテナ)。 AAPF の Web UI 上で、ユーザー自身でイメージと性能タイプを選択して作成します。
性能タイプ	AAPF 上で AACluster を作成するときに指定する、AACluster に割り当てるリソース(CPU コア数、メモリー容量)のタイプ。

その他、上表に無い用語の説明については「AAPF マネージドサービス サービス仕様書」も合わせて参照してください。

2 はじめに

本ガイドでは、AAPF マネージドサービス で提供する分析エンジンを対象として、分析エンジンの利用形態に分けてそれぞれのサイジング手順を記載しています。

- 「異種混合学習」(SAMPO/FAB),「異種混合学習」(sklearn-fab)を利用して分析する …3 章を参照
- 「RAPID 機械学習 時系列数値解析」を利用して分析する …0 章を参照
- 「テキスト分析」を利用して分析する …6 章を参照
- 上記 4 つの分析エンジンを複数利用する …7 章を参照

上記各章で見積もった必要な CPU コア数、メモリー容量、ストレージ容量を基に、「8 章 契約プラン、オプションの選択」にて、AAPF マネージドサービスが提供する契約プランやオプションの選択・検討を行います。

AAPF 上で「テキスト分析 with Deep Learning」 (TDL)の利用をご検討で、その性能や必要メモリー量の見積もりが必要な場合は、AAPF お問い合わせ窓口(aapf-contact@aar.jp.nec.com)までご相談ください。

3　「異種混合学習」を利用するシステムのサイジング

本サービスにおいて、「異種混合学習」の分析エンジンを 1 ユーザーが利用する場合の 1 ユーザーに必要な各種スペックの参考値を見積もる方法を示します。

3.1　前提条件

以降で算出する見積もり値は、異種混合学習で利用する分析対象データの特性によって結果が大きく異なります。より確からしい見積もりを出すためには、本番運用と同等の特性を持つサンプルデータを用いるなどして実測することを推奨します。

次の表は、本サービス上で幾つかのサンプルデータに基づく 1 回あたりの学習処理(モデルの作成)を基に実測値をまとめたものですので、実測が難しい場合の見積もりの参考に留めてご利用ください。

表 3.1　(参考) 1 回あたりの学習処理における実測値

ケース	サンプル数×属性数	入力データサイズ	**メモリー使用量**	**ストレージ使用量**	**実行時間**
(A)	200,000	0.4MB	**580MB**	**4.4MB**	**0.004 時間**
(B)	550,000	2.4MB	**981MB**	**12.8MB**	**0.021 時間**
(C)	900,000	1.8MB	**887MB**	**17.5MB**	**0.021 時間**

- 属性数は、サンプルデータに含まれる説明変数の候補となる数を示します。
- 二値展開を行う属性数が多い場合、およびその属性の中の値の種類数が多い場合、メモリー使用量が極端に大きくなる場合があります。
- 表 3.1 のメモリー使用量、ストレージ使用量、および実行時間は、並列度 1(使用 CPU コア数 1)での実測値です。
- 入力データサイズは、サンプル数、および属性データの特性によって決まります。
- 予測処理(予測結果の取得)でのメモリー使用量、ストレージ使用量、および実行時間は、学習処理の時よりは小さくなります。
- 表 3.1 の実測値は SAMPO/FAB での実測値ですが、sklearn-fab においても SAMPO/FAB の数値を使用して見積もってください。

3.2　サイジング手順

以下に各種スペックの参考値を見積もる手順を示します。

3.2.1　CPU コア数とメモリー

■ 並列度

異種混合学習エンジンを用いて作成するモデルの数とその作成期間、実測値もしくは 表 3.1 を参考に算出した 1 回の学習実行時間を基に、学習処理の並列度を以下で求めます。

実行可能サイクル数 =ceil (全モデルの作成に許容可能な時間[h] /
（1 回の学習実行時間[h] × ランダムリスタート数))

並列度 = ceil (全モデル作成数 / 実行可能サイクル数)

- ランダムリスタート数は、異種混合学習エンジンが持つ初期解のランダム性を考慮して試行する学習の回数を示します。求める精度に応じて最大 30 程度の範囲で決定します。
- ceil (...) は算出した値の小数点以下を切り上げます。

■ CPU コア数

上記の並列度を基に、必要な CPU コア数を以下で求めます。

CPU コア数 = ceil (並列度 × ((並列度 - 1) × 性能劣化係数 + 1) × 安全係数)

- AAPF マネージドサービスでの性能劣化係数は 0.04 とします。動作実績に応じて調整してください。
- 安全係数は標準で 2.0 とします。要件に応じて調整してください。

■ メモリー

実測値もしくは 表 3.1 を参考に算出した 1 回の学習処理でのメモリー使用量を基に、必要なメモリー容量を以下で求めます。

メモリー容量[GB] = 1 回の学習処理でのメモリー使用量[GB] × 並列度

- 表 3.1 を参考に見積もる場合、式中のメモリー使用量は 1GB=1,024MB で換算します。
- 1 つの AACluster が使用するメモリー使用量は全体に占める割合が僅かなため考慮しません。

3.2.2 ストレージ容量

1 回の学習への入力データサイズ、および実測値もしくは 表 3.1 を参考に算出した 1 回の学習処理でのストレージ使用量を基に、必要なデータストレージ容量を以下で求めます。

データストレージ容量[GB] =
（1 回の学習への入力データサイズ[GB] + 1 回の学習処理でのストレージ使用量[GB]）×
（全モデル作成数 × ランダムリスタート数 × 学習処理結果の保存世代数）

- 表 3.1 を参考に見積もる場合、式中のストレージ使用量は 1GB=1,024MB で換算します。
- 異種混合学習では、学習処理ごとにファイルシステムに結果を保存します。要件に応じて保存する世代数を決定します。

4 「RAPID 機械学習 マッチング」を利用するシステムのサイジング

本サービスにおいて、「RAPID 機械学習 マッチング」の分析エンジンを１ユーザーが利用する場合の１ユーザーに必要な各種スペックの参考値を見積もる方法を示します。

4.1 前提条件

以降で算出する見積もり値は、利用する分析対象データのサイズによって結果が大きく異なります。より確からしい見積もりを出すためには、本番運用と同等のサイズを持つサンプルデータを用いるなどして実測することを推奨します。

RAPID 機械学習マッチング版の入力データは、テキスト列と数値列、カテゴリ値列が混在するデータ行の連続で、テキスト列のエントリーには可変サイズのテキスト文字列が含まれます。入力データのサイズは、データの列数、行数以外に、テキスト文字列の最大値によって決まります。

以下の表は、本サービス上でいくつかのサンプルデータに基づく学習コマンド実行時の実測値をまとめたものですので、実測が難しい場合の見積もりの参考に留めてご利用ください。

表 4.1 学習コマンド実行時の実測値（マッチング機能）

	データ条件				測定項目		
	テキスト文字数	テキスト列数	数値列数	データ行数	入力データサイズ	メモリー使用量	ストレージ使用量
1	1,000	0	50	10,000	3.9MB	94MB	116KB
2	1,000	1	50	10,000	64MB	498MB	176KB
3	1,000	10	50	10,000	601MB	4,201MB	724KB
4	1,000	1	10	10,000	61MB	488MB	156KB
5	1,000	1	100	10,000	68MB	513MB	200KB
6	1,000	1	50	100,000	635MB	1,566MB	176KB

表 4.2 学習コマンド実行時の実測値（フィルタリング機能）

	データ条件				測定項目		
	テキスト文字数	テキスト列数	数値列数	データ行数	入力データサイズ	メモリー使用量	ストレージ使用量
1	1,000	0	50	10,000	2.4MB	84MB	12KB
2	1,000	1	50	10,000	33MB	148MB	44KB
3	1,000	10	50	10,000	304MB	702MB	300KB
4	1,000	1	10	10,000	31MB	144MB	40KB

5	1,000	1	100	10,000	35MB	151MB	48KB
6	1,000	1	50	100,000	325MB	734MB	44KB
7	1,000	5	50	100,000	1,566MB	2,774MB	160KB

・表 4.1、表 4.2 の測定は、マッチング機能とフィルタリング機能のそれぞれで行い、パラメータは全てデフォルト値を用いています。
・テキスト列の 1 項目のテキストデータは 1000 文字の日本語文字列を使用しています。

- 各表の実測値は下記の対応で測定項目の変化の基準としてください。
  ・①、②、③：テキスト列を増やした場合の測定項目の変化
  ・②、④、⑤：数値列を増やした場合の測定項目の変化
  ・②、⑥：データ行数を増やした場合の測定項目の変化

4.2　サイジング手順

以下に各種スペックの参考値を見積もる手順を示します。

4.2.1　メモリー容量

RAPID 機械学習マッチング版のメモリー使用量は、一般に学習処理において最も多くのメモリー容量を必要とするため、学習処理におけるメモリー使用量を基準に見積もります。学習に使用するメモリー量は、入力データのサイズに大きく依存するため、見積もる際は入力データのサイズを基に考えます。表 4.3、表 4.4 のメモリー使用量を基準とし、実際に用いる入力データの行数、列数に応じて増加・減少すると考えて計算します。

表 4.3　学習コマンド実行時の実測値（マッチング機能）

	データ条件				測定項目	
No	テキスト文字数	テキスト列数	数値列数	データ行数	入力データサイズ	**メモリー使用量**
1	1,000	0	50	10,000	3.9MB	**94MB**
2	1,000	1	50	10,000	64MB	**498MB**
3	1,000	10	50	10,000	601MB	**4,201MB**
4	1,000	1	10	10,000	61MB	**488MB**
5	1,000	1	100	10,000	68MB	**513MB**
6	1,000	1	50	100,000	635MB	**1,566MB**

表 4.4　学習コマンド実行時の実測値（フィルタリング機能）

	データ条件				測定項目	
	テキスト文字数	テキスト列数	数値列数	データ行数	入力データサイズ	**メモリー使用量**
1	1,000	0	50	10,000	2.4MB	**84MB**
2	1,000	1	50	10,000	33MB	**148MB**
3	1,000	10	50	10,000	304MB	**702MB**
4	1,000	1	10	10,000	31MB	**144MB**
5	1,000	1	100	10,000	35MB	**151MB**
6	1,000	1	50	100,000	325MB	**734MB**
7	1,000	5	50	100,000	1,566MB	**2,774MB**

- 表 4.3、表 4.4 は、学習モードが分類の場合のメモリー使用量の実測値ですが、回帰の場合もほぼ同様のメモリー使用量となります。

4.2.1　ストレージ容量

RAPID 機械学習　マッチング版では、前処理、学習処理、予測処理を実行するため、入力データのサイ

ズ、前処理実行によるストレージ消費の増加量、学習処理実行によるストレージ消費の増加量、予測処理実行によるストレージ消費の増加量の合計を基準に見積もります。

表 4.5　学習コマンド実行時の実測値（マッチング機能）

	データ条件				測定項目	
No	テキスト文字数	テキスト列数	数値列数	データ行数	入力データサイズ	**ストレージ使用量**
1	1,000	0	50	10,000	3.9MB	**116KB**
2	1,000	1	50	10,000	64MB	**176KB**
3	1,000	10	50	10,000	601MB	**724KB**
4	1,000	1	10	10,000	61MB	**156KB**
5	1,000	1	100	10,000	68MB	**200KB**
6	1,000	1	50	100,000	635MB	**176KB**

表 4.6　学習コマンド実行時の実測値（フィルタリング機能）

	データ条件				測定項目	
No	テキスト文字数	テキスト列数	数値列数	データ行数	入力データサイズ	**ストレージ使用量**
1	1,000	0	50	10,000	2.4MB	**12KB**
2	1,000	1	50	10,000	33MB	**44KB**
3	1,000	10	50	10,000	304MB	**300KB**
4	1,000	1	10	10,000	31MB	**40KB**
5	1,000	1	100	10,000	35MB	**48KB**
6	1,000	1	50	100,000	325MB	**44KB**
7	1,000	5	50	100,000	1,566MB	**160KB**

- 表 4.5、表 4.6 は、学習モードが分類の場合のメモリー使用量の実測値ですが、回帰の場合もほぼ同様のメモリー使用量となります。
- ストレージ容量としては、表 4.5、表 4.6 のストレージ使用量に入力データ（※1）を加えてストレージ容量を見積もってください。さらに学習・予測処理を行う前に前処理としてデータの加工を行う場合、入力データサイズをさらに加えて見積もってください（※2）。

  ※1：表 1、表 2 ストレージ使用量には入力データが使用するストレージ容量を含みません。

  ※2：前処理の実行で入力データサイズとほぼ同等のストレージを消費します。

5　「RAPID 機械学習 時系列数値解析」を利用するシステムのサイジング

本サービスにおいて、「RAPID 機械学習 時系列数値解析」の分析エンジンを１ユーザーが利用する場合の１ユーザーに必要な各種スペックの参考値を見積もる方法を示します。

5.1　前提条件

以降で算出する見積もり値は、利用する分析対象データのサイズや用いるアルゴリズムによって結果が大きく異なります。より確からしい見積もりを出すためには、本番運用と同等のサイズを持つサンプルデータを用いるなどして実測することを推奨します。

RAPID 機械学習　時系列数値解析に入力する時系列数値データ（以下、入力データ）のサイズは、サンプル数（行数）とセンサ数（列数）によって決まります。アルゴリズムについては、利用開始時に決まっていない場合、最大値によって見積もってください。

以下の表は、本サービス上でいくつかのサンプルデータに基づく 1 回当たりの処理を基に実測値をまとめたものですので、実測が難しい場合の見積もりの参考に留めてご利用ください。

表 5.1 学習処理１回当たりの実測値(1DDCN アルゴリズム)

No	入力データ行数	入力データ列数	入力データサイズ	メモリー使用量	ストレージ使用量	実行時間
1	10,000	10	0.26MB	**0.30MB**	**4,312KB**	0.197 時間
2	10,000	50	1.05MB	**0.46MB**	**4,348KB**	0.206 時間
3	10,000	100	2.06MB	**0.50MB**	**4,392KB**	0.214 時間
4	100,000	10	2.69MB	**0.37MB**	**4,312KB**	2.500 時間
5	100,000	50	10.69MB	**0.67KB**	**4,348KB**	2.620 時間
6	100,000	100	20.69MB	**0.77MB**	**4,352KB**	2.777 時間

表 5.2 学習処理１回当たりの実測値(1DOCN アルゴリズム)

No	入力データ行数	入力データ列数	入力データサイズ	メモリー使用量	ストレージ使用量	実行時間
1	10,000	10	0.26MB	**0.13MB**	**1,244KB**	0.031 時間
2	10,000	50	1.05MB	**0.56MB**	**1,304KB**	0.043 時間
3	10,000	100	2.06MB	**1.01MB**	**1,396KB**	0.058 時間
4	100,000	10	2.69MB	**0.81MB**	**1,244KB**	0.450 時間
5	100,000	50	10.69MB	**2.67MB**	**1,304KB**	0.629 時間
6	100,000	100	20.69MB	**9.58MB**	**1,396KB**	0.851 時間

表 5.3 予測処理１回当たりの実測値(1DDCN アルゴリズム)

No	入力データ行数	入力データ列数	入力データサイズ	メモリー使用量	ストレージ使用量	実行時間
1	10,000	10	0.26MB	0.04MB	**1,404KB**	0.009 時間
2	10,000	50	1.05MB	0.08MB	**1,596KB**	0.010 時間

3	10,000	100	2.06MB	0.15MB	**1,792KB**	0.012 時間
4	100,000	10	2.69MB	0.14MB	**19,502KB**	0.117 時間
5	100,000	50	10.69MB	0.44MB	**18,092KB**	0.125 時間
6	100,000	100	20.69MB	1.05MB	**18,096KB**	0.137 時間

表 5.4 予測処理 1 回当たりの実測値(1DOCN アルゴリズム)

No	入力データ行数	入力データ列数	入力データサイズ	メモリー使用量	**ストレージ使用量**	実行時間
1	10,000	10	0.26MB	0.03MB	**488KB**	0.012 時間
2	10,000	50	1.05MB	0.08MB	**492KB**	0.013 時間
3	10,000	100	2.06MB	0.15MB	**492KB**	0.014 時間
4	100,000	10	2.69MB	0.03MB	**7,756KB**	0.198 時間
5	100,000	50	10.69MB	0.47MB	**7,732KB**	0.200 時間
6	100,000	100	20.69MB	1.08MB	**7,788KB**	0.271 時間

上記の測定条件は以下の通りです。

- ストレージ使用量は、各処理の実行によるストレージ消費の増加量を測定したものです。
- 実行時間は、8CPU コアにおける測定結果です。
- 各アルゴリズムの測定条件は以下の通りです。
  - 1DDCN(1-Dimension Deep Convolutional Neural Network)アルゴリズム
    - 学習モード: 分類
    - 学習・予測コンフィグのパラメータ
      - WINDOW_SIZE: 256（ウィンドウサイズ（一度に抜き出して入力するデータの行数））
      - TRAIN_EPOCH: 10（学習の反復回数（epoch 数））
      - その他のパラメータはデフォルト値
  - 1DOCN(1-Dimension OneClass Neural Network)アルゴリズム
    - 学習モード: 分類
    - 学習・予測コンフィグのパラメータ
      - WINDOW_SIZE: 256
      - 1DOCN_PREDICT_WINDOW_SIZE: 128（判断区間のサイズ）
      - TRAIN_EPOCH: 10
      - その他のパラメータはデフォルト値

5.2 サイジング手順

以下に各種スペックの参考値を見積もる手順を示します。

5.2.1 CPU コア数(参考)

必要な CPU コア数については、「表 5.1、表 5.2」の実行時間を参考にしてください。

CPU コア数を増やしても実行時間にほとんど差が無い場合があります。一方で、複数の学習や予測処理を同時に実行したい場合は、一般にコア数が多い方が、少ない場合に比べて全体の実行時間が短縮されます。

5.2.2　メモリー容量

RAPID 機械学習　時系列数値解析では、一般に学習処理において最も多くのメモリー容量を必要とするため、学習処理におけるメモリー使用量を基準に見積もります。学習に使用するメモリー量は、アルゴリズムごとで入力データのサイズに大きく依存するため、見積もる際は入力データのサイズを基に考えます。表 5.1、表 5.2 のメモリー使用量を基準とし、実際に用いる入力データの行数、列数に比例して増加・減少すると考えて計算します。例えば、列数が同じで行数が 2 倍の場合は、メモリー使用量は 2 倍と考えます。同様に、行数が同じで列数が 2 倍の場合も、メモリー使用量は 2 倍と考えます。

その他の分析条件に対する考え方は以下の通りです。

- 表 5.1、表 5.2 は、学習モードが「分類」の場合のメモリー使用量の実測値ですが、回帰の場合もほぼ同等のメモリー使用量となります。
- 1DCNN アルゴリズムを用いる場合のメモリー使用量は、一般に 1DDCN アルゴリズムと同等かそれ以下となるため、1DDCN の場合を基準にメモリー使用量を見積もります。

5.2.1　ストレージ容量

RAPID 機械学習　時系列数値解析では、少なくとも学習処理と予測処理を実行するため、入力データのサイズ、学習処理実行によるストレージ消費の増加量、予測処理実行によるストレージ消費の増加量、の合計を基準に見積もります。ストレージ使用量は、アルゴリズムごとで入力データのサイズに大きく依存するため、見積もる際は入力データのサイズを基に考えます。表 5.1～表 5.4 のストレージ使用量を基準とし、実際に用いる入力データの行数、列数に照らし合わせて見積もります。学習処理によるストレージ使用量は、表 5.1、表 5.2 のストレージ使用量を基準とし、実際に用いる入力データの行数・列数を基にしますが、単純な比例関係にはならないため、表 5.1、表 5.2 の中で最も近く、かつ大きめの入力データのサイズで見積もります。例えば、9,000 行・30 列の入力データの場合は、10,000 行・50 列の入力データにおけるストレージ使用量として見積もります。入力データサイズが表 5.1、表 5.2 のサイズを超える場合、表の最大値で見積もってください。予測処理によるストレージ使用量は、表 5.3、表 5.4 のストレージ使用量を基準とし、実際に用いる入力データの行数に比例して増加・減少すると考えて計算します。例えば、行数が 2 倍の場合は、ストレージ使用量は 2 倍と考えます。予測処理においては、入力データの列数はストレージ使用量に大きくは影響しません。

その他の分析条件に対する考え方は以下の通りです。

- 表 5.1、表 5.2 は、学習モードが「分類」の場合における学習処理のストレージ使用量の実測値ですが、学習モードが「回帰」の場合も、ストレージ使用量はほぼ同等となります。

- 表 5.3、表 5.4 は、学習モードが「分類」の場合における予測処理のストレージ使用量の実測値ですが、学習モードが「回帰」の場合、ストレージ使用量は「分類」の時より小さくなります。
- 1DCNN アルゴリズムを用いる場合のストレージ使用量は、一般に 1DDCN アルゴリズムと同等かそれ以下となるため、1DDCN の場合を基準にストレージ使用量を見積もります。
- 上記で算出されるストレージ使用量は、必要となる最小のストレージ使用量となります。例えば、前処理を実行する場合、上記で算出されたストレージ使用量に加えて、出力する前処理済ファイルのストレージを格納するストレージも必要となります。

6 「テキスト分析」を利用するシステムのサイジング

本サービスにおいて、「テキスト分析」の分析エンジンを 1 ユーザーが利用する場合の 1 ユーザーに必要な各種スペックの参考値を見積もる方法を示します。

6.1 前提条件

以降で算出する見積もり値は、テキスト分析で利用する分析機能・分析対象データの特性によって結果が大きく異なります。より確からしい見積もりを出すためには、本番運用と同等の特性を持つサンプルデータを用いるなどして実測することを推奨します。

表 6.1～表 6.4 は、本サービス上で文クラスタリング、文書判別のモデル作成のための学習、文書判別の 3 つの機能についてサンプルデータに基づく処理の実測値を元にまとめたものです。

文書判別については、文書判別(学習処理)のケース(E)の 28,000 文書で判別モデルを作った場合の実測値になります。また、文書判別は、出力ファイルとして csv ファイルと asf ファイルを出力することができますのでそれぞれのパターンで計測しています。

本結果については、実測が難しい場合の見積もりの参考に留めてご利用ください。

表 6.1 文クラスタリングにおける実測値

ケース	文数	入力データサイズ	メモリー使用量	ストレージ使用量	実行時間
(A)	5,000	0.8MB	42MB	5.9MB	0.002 時間
(B)	10,000	1.6MB	59MB	12.9MB	0.010 時間
(C)	15,000	2.3MB	81MB	21.2MB	0.024 時間
(D)	20,000	2.9MB	107MB	30.7MB	0.044 時間
(E)	30,000	3.8MB	174MB	53.2MB	0.102 時間
(F)	40,000	5.1MB	258MB	80.6MB	0.184 時間

表 6.2 文書判別(学習処理)における実測値

ケース	文書数	入力データサイズ	メモリー使用量	ストレージ使用量	実行時間
(A)	3,500	5.3MB	2,930MB	78.1MB	0.021 時間
(B)	7,000	10.3MB	3,620MB	111MB	0.034 時間
(C)	14,000	20.7MB	4,980MB	178MB	0.059 時間
(D)	21,000	31.1MB	6,340MB	244MB	0.084 時間
(E)	28,000	41.5MB	7,770MB	311MB	0.109 時間

表 6.3 文書判別(判別処理-csv 出力)における実測値

ケース	文書数	入力データサイズ	メモリー使用量	ストレージ使用量	実行時間
(A)	11,000	16.6MB	191MB	45.0MB	0.024 時間

(B)	23,000	33.4MB	**192MB**	**94.1MB**	**0.051 時間**
(C)	46,000	67.6MB	**195MB**	**188MB**	**0.102 時間**
(D)	69,000	102.8MB	**197MB**	**282MB**	**0.153 時間**
(E)	92,000	137.9MB	**200MB**	**376MB**	**0.204 時間**

表 6.4　文書判別(判別処理-asf 出力)における実測値

ケース	文書数	入力データサイズ	メモリー使用量	ストレージ使用量	実行時間
(A)	11,000	16.6MB	**188MB**	**1,311MB**	**0.030 時間**
(B)	23,000	33.4MB	**190MB**	**2,742MB**	**0.064 時間**
(C)	46,000	67.6MB	**195MB**	**5,484MB**	**0.127 時間**
(D)	69,000	102.8MB	**200MB**	**8,225MB**	**0.191 時間**
(E)	92,000	137.9MB	**205MB**	**10,967MB**	**0.255 時間**

- 入力データサイズは、文数(文書数)、および文(文書数)の文字数によって決まります。

6.2　文クラスタリング　サイジング手順

以下に文クラスタリングの処理を行う場合の各種スペックの参考値を見積もる手順を示します。

6.2.1 実行時間と CPU コア数(参考)

文クラスタリング処理にかかる実行時間は、表 6.1 の文数、入力データサイズ(文の平均サイズが約 128 バイト)の場合、以下の近似式で求めることができます。

実行時間[h] = ({0.000000428 × 文数 × 文数} - {0.000578 × 文数}) / 3,600

上記は、AACluster の性能タイプの CPU コア数が 8 の場合の実行時間になります。文クラスタリングの実行時間は、文数の 2 乗オーダで増える部分があるため、CPU コア数を増加させても、それに比例した実行時間の短縮はできません。1 万文で 0.01 時間、10 万文で 1.2 時間、20 万文で 4.7 時間、30 万文で 11 時間、50 万文で 30 時間程度の実行時間がかかることが想定されます。クラスタリング対象の文集合の性質によっても実行時間は異なるため、初期の検証時には、文クラスタリング対象の文を最大でも 20 万文程度に絞り込むことをお勧めします。

なお、上記の実行時間を維持した上で、文クラスタリング処理を並列に複数動かす場合は、CPU コア数を増加させることを検討してください。

6.2.2 メモリー

文クラスタリング処理に必要なメモリー量は、表 6.1 の文数、入力データサイズ(文の平均サイズが約 128 バイト)の場合、以下の近似式で求めることができます。

メモリー容量[GB] = (1 回の処理でのメモリー使用量[GB])
= ({0.0000936 × 文数 × 文数} + {2.12 × 文数} + 30,000) / (1,024 * 1,024)

必要なメモリー量は、1 万文で 0.059GB、10 万文で 1.2GB、20 万文で 4.1GB、30 万文で 8.9G、50 万文で 24GB 程度になることが想定されます。

複数並列で動かす場合は、並列度に応じて、メモリー量を増加させることを検討してください。

6.2.3 ストレージ容量

文クラスタリング処理に必要なデータストレージ容量は、表 6.1 の文数、入力データサイズ(文の平均サイズが約 128 バイト)の場合、以下の近似式で求めることができます。

データストレージ容量[GB] = ({0.000000024 × 文数 × 文数} + {0.00105 × 文数}) / 1,024

必要なデータストレージ容量は、1 万文で 0.012GB、10 万文で 0.33GB、20 万文で 1.1GB、30 万文で 2.4G、50 万文で 6.2GB 程度になることが想定されます。数百 GB 規模のストレージ容量があれば、ストレージ容量が大きな問題になることは少ないと考えられます。

6.3　文書判別(学習)　サイジング手順

以下に文書判別(学習)の処理を行う場合の各種スペックの参考値を見積もる手順を示します。

6.3.1 実行時間と CPU コア数(参考)

文書判別の学習処理にかかる実行時間は、表 6.2 の文書数、入力データサイズ(文書の平均サイズが約 1480 バイト)の場合、以下の近似式で求めることができます。

実行時間[h] = ({0.00130 × 文書数}　+ 30) / 3,600

上記は、AACluster の性能タイプの CPU コア数が 8 の場合の実行時間になります。文書判別の学習処理は CPU コア数を増加させても実行時間は改善されません。1,000 文書で 0.012 時間、5,000 文書で 0.026 時間、1 万文書で 0.044 時間、5 万文書で 0.19 時間程度の実行時間がかかることが想定されます。CPU コア数を増加しても実行時間の短縮はできません。文書数に対する実行時間が許容できない場合は、文書数を小さくする必要があります。

なお、上記の実行時間を維持した上で、文書判別の学習処理を並列に動かす場合は、CPU コア数を増加させることを検討してください。

6.3.2 メモリー

文書判別の学習処理に必要なメモリー量は、表 6.2 の文書数、入力データサイズ(文書の平均サイズが約 1,480 バイト)の場合、以下の近似式で求めることができます。

メモリー容量[GB] = (1 回の処理でのメモリー使用量[GB])
=　({0.204 *　文書数} + 56.1) / 1,024 + 2.2

必要なメモリー量は、1,000 文書で 2.4GB、5,000 文書で 3.2GB、1 万文書で 4.2GB、5 万文書で 12GB 程度になることが想定されます。

複数並列で動かす場合は、並列度に応じて、メモリー量を増加させることを検討してください。

6.3.3 ストレージ容量

文書判別の学習処理に必要なデータストレージ量は、表 6.2 の文書数、入力データサイズ(文書の平均サイズが約 1,480 バイト)の場合、以下の近似式で求めることができます。

データストレージ容量[GB] =　({0.00998 *　文書数} + 47.0) / 1,024

必要なデータストレージ容量は、1,000 文書で 0.054GB、5,000 文書で 0.092GB、1 万文書で 0.14GB、5 万文書で 0.52GB 程度になることが想定されます。数百 GB 規模のディスク容量があれば、ディスク容量が大きな問題になることは少ないと考えられます。

6.4　文書判別(判別)-csv 出力　サイジング手順

以下に文書判別処理で csv を出力する場合の各種スペックの参考値を見積もる手順を示します。なお、判別モデルは、表 6.2 の文書判別(学習処理)のケース(E)の約 28,000 文書で作ったもの利用した際のものです。

6.4.1 実行時間と CPU コア数

文書判別の実行時間は、表 6.3 の文書数、入力データサイズ(文書の平均サイズが約 1,480 バイト)の場合、以下の近似式で求めることができます。

実行時間[h] = ({0.00797 × 文書数}) / 3,600

上記は、AACluster の性能タイプの CPU コア数が 8 の場合の実行時間になります。1,000 文書で 0.002 時間、5,000 文書で 0.011 時間、1 万文書で 0.022 時間、5 万文書で 0.11 時間程度の実行時間がかかることが想定されます。文書数に対する実行時間が許容できない場合は、文書を分割して外部で並列に文書判別処理を呼び出す必要があります。並列の度合いに従って CPU コア数の増加を検討してください。

6.4.2 メモリー

文書判別のメモリー容量は、表 6.3 の文書数、入力データサイズ(文書の平均サイズが約 1480 バイト)の場合、以下の近似式で求めることができます。

メモリー容量[GB] = (1 回の処理でのメモリー使用量[GB])
= ({0.000114 × 文書数} + 194) / 1,024

必要なメモリー量は、文書数に大きな影響は受けずに、1,000 文書で 0.185GB、5,000 文書で 0.186GB、1 万文書で 0.186GB、5 万文書で 0.190GB 程度になることが想定されます。

6.4.3 ストレージ容量

文書判別のデータストレージ容量は、表 6.3 の文書数、入力データサイズ(文書の平均サイズが約 1480 バイト)の場合、以下の近似式で求めることができます。

データストレージ容量[GB] = (0.00429 * 文書数) / 1,024

必要なデータストレージ容量は、文書数(入力データサイズ)に比例して、1,000 文書で 0.0041GB、5,000 文書で 0.0205GB、1 万文書で 0.0409GB、5 万文書で 0.205GB 程度になることが想定されます。

6.5　文書判別(判別)-asf 出力　サイジング手順

以下に文書判別処理で asf(言語解析結果を含む JSON 形式)を出力する場合の各種スペックの参考値を見積もる手順を示します。なお、判別モデルは、表 6.2 の文書判別(学習処理)のケース(E)の約 28,000 文書で作ったもの利用することを想定しています。

6.5.1 実行時間と CPU コア数(参考)

文書判別の実行時間は、表 6.4 の文書数、入力データサイズ(文書の平均サイズが約 1,480 バイト)の場合、以下の近似式で求めることができます。

実行時間[h] = ({0.00996 × 文書数}) / 3,600

上記は、AACluster の性能タイプの CPU コア数が 8 の場合の実行時間になります。1,000 文書で 0.003 時間、5,000 文書で 0.014 時間、1 万文書で 0.028 時間、5 万文書で 0.14 時間程度の実行時間がかかることが想定されます。文書数に対する実行時間が許容できない場合は、文書を分割して外部で並列に文書判別処理を呼び出す必要があります。並列の度合いに従って CPU コア数の増加を検討してください。

6.5.2 メモリー

文書判別のメモリー容量は、表 6.4 の文書数、入力データサイズ(文書の平均サイズが約 1,480 バイト)の場合、以下の近似式で求めることができます。

メモリー容量[GB] = (1 回の処理でのメモリー使用量[GB])
= ({0.00212 × 文書数} + 190) / 1,024

必要なメモリー量は、文書数に大きな影響は受けずに、1,000 文書で 0.19GB、5,000 文書で 0.19GB、1 万文書で 0.19GB、5 万文書で 0.20GB 程度になることが想定されます。

6.5.3 ストレージ容量

文書判別のデータストレージ容量は、表 6.4 の文書数、入力データサイズ(文書の平均サイズが約 1,480 バイト)の場合、以下の近似式で求めることができます。

データストレージ容量[GB] = (0.125 * 文書数) / 1,024

必要なデータストレージ容量は、文書数(入力データサイズ)に比例して、1,000 文書で 0.12GB、

5,000 文書で 0.60GB、1 万文書で 1.2GB、5 万文書で 6.0GB 程度になることが想定されます。
asf（言語解析結果を含む JSON 形式）で出力する場合、オリジナルの入力文書の約 80 倍の容量が必要となることに注意して下さい。言語解析結果や文書判別結果を asf で保持する場合は、上記に応じたストレージ容量を確保してください。

7　「4 分析エンジン」を利用するシステムのサイジング

本サービスにおいて、複数のエンジンを利用する場合の見積もり方法を示します。

7.1　前提条件

3 章～6 章のうち、利用する分析エンジンが記載された章における前提条件を参考にしてください。

7.2　サイジング手順

以下に各種スペックの参考値を見積もる手順を示します。

■ CPU コア数

3 章～6 章のうち、利用する分析エンジンが記載された章における手順で見積もった値を用いて、利用形態に合わせて最適なスペックを選択してください。

- 複数の分析エンジンを同時に利用する場合：それぞれ見積もった値の合算値
- 複数の分析エンジンを同時に利用しない場合：それぞれ見積もった値の最大値

■ メモリー

3 章～6 章のうち、利用する分析エンジンが記載された章における手順で見積もった値を用いて、利用形態に合わせて最適なスペックを選択してください。

- 複数の分析エンジンを同時に利用する場合：それぞれ見積もった値の合算値
- 複数の分析エンジンを同時に利用しない場合：それぞれ見積もった値の最大値

■ データストレージ容量

3 章～6 章のうち、利用する分析エンジンが記載された章における手順で見積もった値の合算値を用いて、システム作成時のスペックを指定してください。

8　契約プラン、オプションの選択

3 章～7 章で算出した必要なリソース(CPU コア数、メモリー容量、ストレージ容量)を基に、AAPF マネージドサービスで提供する契約プラン（エントリー、エントリープラス、スタンダード)、およびストレージ追加オプションで提供されるリソースと照らしあわせて、契約するプランを選択します。各契約プランの詳細については、「NEC Advanced Analytics Platform V1.5 マネージドサービス（エントリー、エントリープラスプラン）サービス仕様書」および「NEC Advanced Analytics Platform V1.5 マネージドサービス（スタンダードプラン）サービス仕様書」を参照してください。

契約プラン選択時の注意事項は以下の通りです。

- スタンダードプランでしか実現できない要件がある場合は、見積もった必要なリソース量に関わらず、スタンダードプランを契約する必要があります。
- エントリープラン、エントリープラスプランでは後からメモリーを追加できないため、およびエントリープランでは後からストレージを追加できないことに注意が必要です。不安な場合は 1 ランク上のプランを選択される等ご検討ください。
- エントリープラスプランで 1 ユーザーあたりに拡張できるストレージ容量は限りがあるため、数百 GB 単位の容量が必要な場合スタンダードプランを検討する必要があります。
- スタンダードプランで提供する CPU コア数やメモリーは、AAPF システム(Web サーバやシステム管理等)が使用する分も含んでいるので、仮想分析環境で使える CPU コア数は「契約 CPU コア数－4」、メモリー容量は「契約メモリー容量[GB]－32[GB]」として見積もってください。
- エントリープラン、エントリープラスプランの CPU コア数については、最低保証のコア数です。サーバの負荷状況によっては、AACluster 作成時に指定した性能タイプの CPU コア数より多いコア数を使用される場合があり、そのため契約しているプランの CPU コア数より多いコア数が使用される場合があるため、実測する場合には注意が必要です。本書に掲載している各実測値の表中の実行時間は、各表外に掲載している CPU コア数のみを使用した場合の実測値となっています。

商標について

- Red Hat は、米国およびその他の国における Red Hat,Inc. の商標または登録商標です。
- Linux は、Linus Torvalds 氏の米国およびその他の国における商標または登録商標です。
- その他、本マニュアルに掲載された各社名、各製品名、各ロゴは、各社の商標または登録商標です。

NEC Advanced Analytics Platform V1.5

マネージドサービス

サイジングガイド

2021 年 12 月

日本電気株式会社