Amazon Elastic MapReduce

入門ガイド

API Version 2009-03-31

Amazon Elastic MapReduce: 入門ガイド

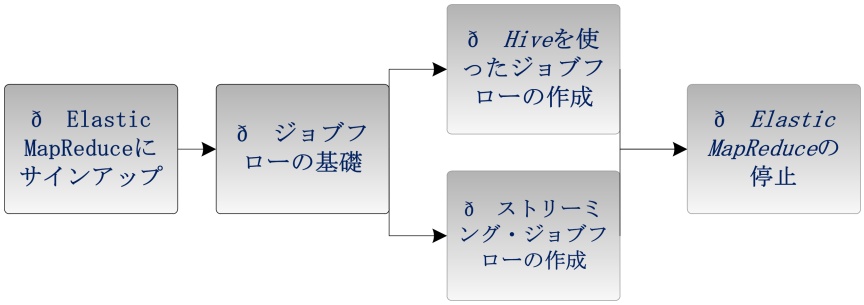
Table of Contents

Amazon Elastic MapReduce 入門

この*Amazon Elastic MapReduce 入門*は、 Elastic MapReduce の機能について、ハイレベルな概要を提供するものです。本ガイドの読了後、Elastic MapReduce の基本が理解できているはずです。これらの例は、 Elastic MapReduce コマンドライン インターフェイスをもちいて Hadoop ストリーミングや Hive ジョブフローを作成する方法、および AWS management console で Elastic MapReduce タブをもちいて、実行中のジョブフローを監視してデバッグする方法について示すものです。

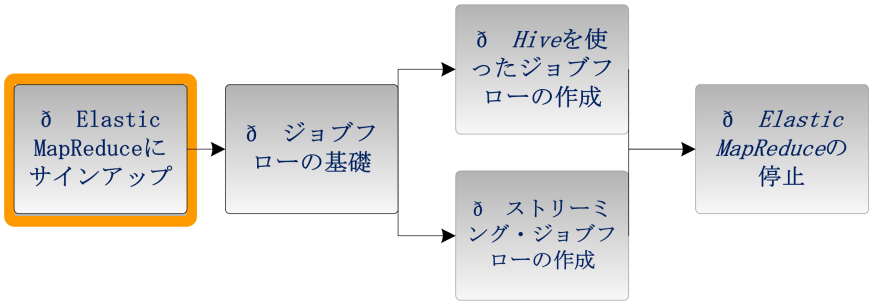
Amazon Elastic MapReduce は、大容量のデータを効率的かつ簡単に処理するウェブサービスです。Elastic MapReduce は、Hadoop 処理といくつかの AWS サービスを組み合わせて、ウェブインデックス化、データマイニング、ログファイル分析、 機械学習、科学シミュレーション、データウェアハウジングなどのタスクを行なっています。

以下の図に示されたタスクに従って、Amazon Elastic MapReduce を使用し始めることができます。



このガイドは、ジョブフローの起動と管理の手順を説明します。Elastic MapReduce を始めて使用する場合は、 Elastic MapReduceへのサインアップ (p. 2)に移動します。

Elastic MapReduceへのサインアップ



Topics

本セクションは、Elastic MapReduce.を使用する前に実行する必要のある、AWS アカウント作成とシステム設定について説明するものです。

Elastic MapReduce アカウントの取得方法

本セクションは、Elastic MapReduce アカウントへのサインアップ方法について説明するものです。このプロセスは、あらゆる Amazon Web Services、リソース、フォーラム、サポートおよび使用レポートに対するアクセスを提供する、Amazon Web Services (AWS) アカウントを作成します。Elastic MapReduce にサインアップすると、自動的に Amazon Elastic Compute Cloud (Amazon EC2) と Amazon Simple Storage Service (Amazon S3), にもサインアップします。これらは Elastic MapReduce と密接に連携しています。それらを使用しない限り、サービスに対して課金されることはありません。

Elastic MapReduce アカウントにサインアップするには

1. http://aws.amazon.com/elasticmapreduce/に進み、Elastic MapReduce へのサインアップをクリックします。
2. Eメールアドレス欄にお客様のEメールアドレスを入力します
3. AWS account にログイン:

- AWS アカウントを「既にお持ちの場合は」を選択し、お客様のパスワードを入力します。それから当社の安全なサーバーを使用してサインインをクリックします。 以下手順 4 に進みます。
- Amazon アカウントをお持ちでない場合は、新規を選択し、当社の安全なサーバーを使用し てサインインをクリックします。手順に従って、AWS アカウントを作成します。

4. Amazon Elastic MapReduce にサインアップに関する情報を確認します。利用規約に同意する場合は、サインアップの完了をクリックして、次のページの手順に従います。

Note

Amazon Elastic MapReduce のサインアップ手順の一部には、通話呼び出しを受け取り、電話のキーパッドをもちいてPINを入力することが含まれています。

Elastic MapReduce コマンドライン インターフェイスをインストールします。

Topics

Elastic MapReduce コマンドライン インターフェイス (CLI) を使用して、複数の手順からなるジョブフローを作成できます。Elastic MapReduce タブは、単一手順のジョブフローのみの作成をサポートします。本文書は、主に Elastic MapReduce CLI をもちいたジョブフローの管理方法について説明しています。AWS management console、Elastic MapReduce タブおよび Elastic MapReduce API の使用方法の詳細は、Amazon Elastic MapReduce Developer Guide および Amazon Elastic MapReduce API Reference にあります。

Ruby のインストール

Elastic MapReduce コマンドライン インターフェイスは、Ruby 1.8 を必要とします。Ruby をインストールしたら、elastic-mapreduce-ruby.zip をディレクトリに解凍してください。これで Elastic MapReduce CLI の使用準備が整います。

Ruby をインストールするには

1. Ruby 1.8 をダウンロードしてインストール:

   - Linux および Unix ユーザーは、Ruby を http://www.ruby-lang.org/en/news/2010/06/23/ruby-1-8-7-p299-released/ からダウンロードし、以下のコマンドを入力して Ruby をインストールできます:

     ```
     $ sudo apt-get install ruby-full
     ```

   - Windows ユーザーは、Ruby を次のアドレスからインストールできます。http://rubyforge.org/frs/?group_id=167&release_id=28426Ruby のディレクトリが、お客様のPATH にあるようにしてください。

2. コマンドプロンプトに以下をタイプして、Ruby が実行中であることを確認してください。
   - Linux や Unix のユーザーは、コマンドラインのプロンプト画面から、以下を入力してください:

     ```
     $ ruby -v
     ```

   - Windows のユーザーは、コマンドラインのプロンプト画面から、以下を入力してください：

     ```
     C:\ruby>ruby -v
     ```

Ruby のバージョンが表示され、Ruby がインストールされたことが確認できます。

コマンドライン インターフェイスのインストール

Elastic MapReduce CLI をダウンロードするには

1. Ruby のディレクトリ で CLI のローカルディレクトリを作成します:
   - Linux や Unix のユーザーは、コマンドラインのプロンプト画面から、以下を入力してください:

     ```
     $ mkdir elastic-mapreduce-cli
     ```

   - Windows のユーザーは、コマンドラインのプロンプト画面から、以下を入力してください：

     ```
     C:\ruby>mkdir elastic-mapreduce-cli
     ```

2. Elastic MapReduce ファイルをダウンロード:
   a. http://aws.amazon.com/developertools/2264 を閲覧してください。AWSにログインしない場合は、プロンプト画面で AWS アカウント証明書を入力します。
   b. ダウンロードをクリックします。
   c. 新規に作成したディレクトリにファイルを保存します。

Elastic MapReduce CLI をインストールするには

1. elastic-mapreduce-cli ディレクトリに移動します。
2. 圧縮ファイルを解凍します:
   - Linux や Unix のユーザーは、コマンドラインのプロンプト画面から、以下を入力してください:

     ```
     $ unzip elastic-mapreduce-ruby.zip
     ```

   - Windows ユーザーは、Windows Explorer から`elastic-mapreduce-ruby.zip`ファイルを開いてください。

証明書の設定

Elastic MapReduce の証明書ファイルは、多くのコマンドに必要な情報を提供することができます。ファイルにコマンド変数を保存し、 情報を毎回入力する手間を省くことができるので便利です。

証明書は、お客様が行なうすべてのリクエストに対して署名値を計算するために使用されます。Elastic MapReduce は、 ファイル中のお客様の証明書を自動的に検索します。`credentials.json`ファイルを編集して AWS 証明書に含めることができるので便利です。AWS キーのペアは、パスワードに似たセキュリティ証明書です。これはお客様が、実行時のインスタンスに安全に接続するために使用するものです。新しいキーペアを作成して、本ガイドと共に使用することを推奨します。

証明書ファイルを作成するには

1. `credentials.json`という名前のファイルを、`elastic-mapreduce-cli`ディレクトリに作成します。
2. 以下の行を、証明書ファイルに追加します。

```
{
"access_id": "[Your AWS Access Key ID]",
"private_key": "[Your AWS Secret Access Key]",
"keypair": "[Your key pair name]",
"key-pair-file": "[The path and name of your PEM file]",
"log_uri": "[A path to a bucket you own on Amazon S3, such as, s3n://mylog-uri/]",
"region": "[The Region of your job flow, either us-east-1, us-west-1, eu-west-1, ap-northeast-1, ap-southeast-1, or sa-east-1.]"
}
```

   リージョン名をメモします。このリージョンを使用して、Amazon EC2 のキーペアと Amazon S3 のバケットを作成します。

次のセクションでは、証明書の作成および検索方法について説明します。

AWS セキュリティ証明書

AWS は、セキュリティ証明書を使用して、お客様のデータを保護します。本セクションでは、セキュリティ証明書の閲覧方法を提示し、それらをお客様の`credentials.json`ファイルに追加できるようにします。

AWS は、アクセスキーIDおよびシークレットアクセスキーをお客様に割り当てます。アクセスキーIDをすべてのAWSサービスリクエストに含めることによって、お客様がリクエストの送信者であることを証明できます。

Note

シークレットアクセスキーは、お客様とAWSの間で秘密裏に共有されるものです。このIDを他者に公開しないでください。当社はこれを使用して、お客様が利用するAWSサービスの使用料金を請求します。このIDを、AWS に対するリクエストに含めないでください。また、問い合わせの発信元がAWS またはAmazon.com であるように見える場合でも、このIDを他者にEメール送信しないでください。Amazonのスタッフまたは関係者がシークレットアクセスキーについて尋ねることは決してありません。

AWS アクセスキーID とAWS シークレットアクセスキーの場所を特定するには

1. AWS のウェブサイトhttp://aws.amazon.comに進みます。

2. アカウントをクリックして、オプション一覧を表示します。
3. セキュリティ証明書をクリックして、お客様のAWS アカウントにログインします。お客様のアクセスキーIDが、アクセス証明書セクションに表示されます。シークレットアクセスキーは、非公開のままとなり、さらに警告が表示されます。
4. シークレットアクセスキーを表示するには、以下の図で示されているように、あなたの秘密アクセスキーエリアで表示をクリックします。

access_key変数を、アクセスキーIDの値に設定し、private_key変数を、シークレットアクセスキーの値に設定します。

Amazon EC2 キーペアを作成するには

1. AWS management console の Amazon EC2 タブに進みます。AWSにログインしない場合は、プロンプト画面で AWS アカウント証明書を入力します。
2. EC2 のダッシュボードから、credentials.json ファイルで使用したリージョンを選択し、そしてキーペア をクリックします。
3. キーペアページ上で、キーペアの作成をクリックします。
4. お客様のキーペア名を入力します。（ 例：**`mykeypair`** ）
5. 作成をクリックします。
6. 結果として生じるPEMファイルを、安全な場所に保存します。

あなたのcredentials.jsonファイルで、*keypair*パラメータを Amazon EC2 キーペア名に変更し、*key-pair-file*パラメータを PEM ファイルのロケーションと名前に変更します。

Amazon S3 バケット

変数は、Elastic MapReduce の結果とジョブフローのログファイルのために、Amazon S3 でロケーションを指定するものです。*log-uri*log-uri変数の値は、お客様がこの目的で作成する Amazon S3 バケットです。

Amazon S3 バケットを作成するには

1. https://console.aws.amazon.com/s3/home の Amazon S3 タブに進みます。AWS にログインしない場合は、求められたら AWS アカウント証明書を入力します。
2. バケットの作成をクリックします。
バケットの作成ダイアログボックスが開きます。
3. バケット名を入力します。（例：**mylog-uri**）
この名前は、グローバルに一意である必要があります。また、他のバケットで使用されるものと同じであることはできません。

   Note

   有効なバケット名の詳細については、
   http://docs.amazonwebservices.com/AmazonS3/latest/dev/BucketRestrictions.html を参照してください。

4. バケットのリージョン（region）を選択します。

あなたの Elastic MapReduce リージョンが _ _ の場合は...	Amazon S3 のリージョンを選択...
us-east-1	米国スタンダード
us-west-2	オレゴン
us-west-1	北カリフォルニア
eu-west-1	アイルランド
ap-northeast-1	東京
ap-southeast-1	シンガポール
sa-east-1	南米 – サンパウロ

5. 作成をクリックします。

   Note

   バケットの作成ウィザードでロギングを有効にする場合、バケットアクセスログのみが有効になります。Elastic MapReduce ジョブフロー ログは有効になりません。

URI**s3n://mylog-uri/**でバケットを作成しました。

バケットを作成後、それに対する適切な権限を設定してください。一般的に、お客様自身（オーナー）に読み書きのアクセス、認証されたユーザーに対しては読み込みアクセスを付与します。

Amazon S3 バケットに対する権限を設定するには

1. まだそこにいない場合は、 https://console.aws.amazon.com/s3/home の Amazon S3 タブに進んでください。AWSにログインしない場合は、プロンプト画面で AWS アカウント証明書を入力します。
2. バケットペインで、作成したばかりのバケット上で右クリックします。
3. プロパティを選択します。
4. プロパティペインで、権限タブを選択します。
5. さらに権限を追加をクリックします。
6. 権限保有者欄で、認証されたユーザーを選択します。
7. 権限保有者欄の右側で、リストを選択します。
8. 保存をクリックします。

これで、バケットを作成し、それに権限を割り当てることができました。*log-uri*変数をこのバケットのURIに、Elastic MapReduce のロケーションとして設定し、ログと結果をアップロードします。

SSH セットアップと設定

ssh または PuTTY で使用するために、SSH 証明書を設定します。

SSH 証明書を設定するには

- SSH を使用するようコンピュータを設定:
  - Linux および Unix ユーザーは、Amazon EC2 キーペアの PEM ファイル上で権限を設定します。例えば、ファイルを**mykeypair.pem**として保存した場合、コマンドは次のようになります。

    ```
    $ chmod og-rwx mykeypair.pem
    ```

  - Windows ユーザー
    a. PuTTYgen.exe を、次のアドレスからお客様のコンピューターにダウンロードします。 http://www.chiark.greenend.org.uk/~sgtatham/putty/download.html.
    b. PuTTYgen を起動します。
    c. 読み込みをクリックします。先に作成したPEMファイルを選択します。
    d. 開くをクリックします。
    e. キーのインポートが成功したことを告げるPuTTYgen 通知上でOKをクリックします。
    f. 秘密鍵の保存をクリックして、鍵をPPKフォーマットで保存します。
    g. PuTTYgen がプロンプト画面で、パスフレーズなしで鍵を保存するよう促したら、はいをクリックします。
    h. PuTTY 秘密鍵の名前を入力します。（例：**mykeypair.ppk**）
    i. 保存をクリックします。
    j. PuTTYgen アプリケーションを終了します。

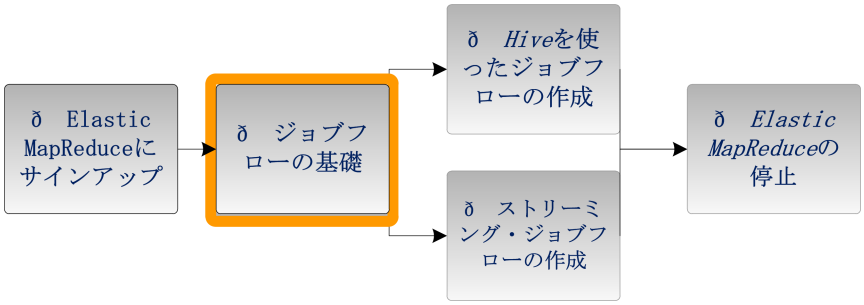
Windows ユーザーは、マスターノードにリモートで接続するには PuTTY をインストールする必要があります。

PuTTY をダウンロードするには

- Windows ユーザーのみ、PuTTY を次のアドレスからお客様のコンピューターにダウンロードしてください。 http://www.chiark.greenend.org.uk/~sgtatham/putty/download.html.Hive をもちいたサンプル ジョブフローの一部として、ssh経由で PuTTY をお客様のマスターノードに接続します。

これで、Amazon Elastic MapReduce にサインアップし、Amazon Elastic MapReduce CLI をインストールして、設定を行なうことができましたので、ジョブフローの基礎 (p. 10)に進みます。

ジョブフローの基礎



Topics

本セクションは、Elastic MapReduce コマンドラインインターフェイス (CLI) をもちいたジョブフローを作成して管理する方法について、一般的情報を提供するものです。

Elastic MapReduce は、Amazon EC2 クラスタのプロビジョニング、終了、Amazon S3 とそれの間のデータ移動、Hadoop の最適化を行います。Elastic MapReduce は、セットアップのモニタリング、Hadoop の設定、ジョブフローの実行など、サーバークラスタが必要とする、ハードウェアやネットワーキングの詳細な設定の大部分を取り除きます。

ジョブフローの作成

Elastic MapReduce CLI を使用して、あなたが実行終了するまで継続的に実行できるジョブフローを作成できます。このプロセスは、デバッグのために便利です。ステップが失敗する場合、アクティブなジョブフローに別のステップを追加することができます。シャットダウンや新しいジョブフロー開始の費用は必要ありません。

一般的に 1 つのステップには、非常に大容量のデータに対する比較的単純なオペレーションの実行が含まれます。1 つのステップは、大まかに言って、データを操作する 1 つのアルゴリズムに対応しています。1 つのジョブフローは一般的に、複数のステップから構成されます。1 つのステップの出力結果は

しばしば、次のステップの入力になります。1つまたは複数の一連のステップは、*ジョブフローと呼ばれます。*

以下のコマンドは、終了するまでリソースを消費するジョブフローを開始します。

ジョブフローを作成するには

- コマンドラインのプロンプト画面から以下のコマンドを入力してください:
  - Linux および UNIX ユーザー:

```
$ ./elastic-mapreduce --create --alive
```

  - Windows ユーザー:

```
C:\ruby\elastic-mapreduce-cli>ruby elastic-mapreduce --create --alive
```

出力結果は次のものと同様になります。

```
Created job flow JobFlowID
```

このコマンドは、単一の m1.small インスタンス上で実行されるジョブフローを起動します。*--alive* オプションは、全ステップを完了した場合でも、ジョブフローの実行が継続するよう命令を与えるものです。

新規に作成された各ジョブフローに対して、一意なジョブフローIDが割り当てられます。ジョブフローIDを使用して、ジョブフローの特定と管理を行ないます。

ジョブフローの管理

本セクションは、ジョブフローを特定して管理するいくつかの方法を紹介します。

すべての Elastic MapReduce コマンドを列挙する

*--help*パラメータを使用して、Elastic MapReduce CLI で使用可能な全コマンドを列挙できます。

Elastic MapReduce の全コマンドを列挙するには

- コマンドラインのプロンプト画面から以下のコマンドを入力してください:
  - Linux および UNIX ユーザー:

```
$ ./elastic-mapreduce --help
```

  - Windows ユーザー:

```
C:\ruby\elastic-mapreduce-cli>ruby elastic-mapreduce --help
```

各Elastic MapReduce コマンドの詳細については、Amazon Elastic MapReduce Developer Guide を参照してください。

すべてのジョブフローを列挙する

*--list*パラメータを使用して、過去2週間の全ジョブフローを列挙できます。

ジョブフローをすべて列挙するには

- コマンドラインのプロンプト画面から以下のコマンドを入力してください:
  - Linux および UNIX ユーザー:

```
$ ./elastic-mapreduce --list
```

  - Windows ユーザー:

```
C:\ruby\elastic-mapreduce-cli>ruby elastic-mapreduce --list
```

反応は以下と同様です:

```
JobFlowID     STARTING
   Development Job Flow  (requires manual termination)
```

ジョブフローのSTATESおよびジョブフローをリストアップするその他の方法の詳細については、Amazon Elastic MapReduce Developer Guide を参照してください。

特定のジョブフローについての情報を取得する

*--describe*オプションと関連ジョブフローIDを使用して、ジョブフローについての情報を取得できます。

ジョブフローについての情報を取得するには

- コマンドラインのプロンプト画面から以下のコマンドを入力してください:
  - Linux および UNIX ユーザー:

```
$ ./elastic-mapreduce --describe --jobflow [JobFlowID]
```

  - Windows ユーザー:

```
C:\ruby\elastic-mapreduce-cli>ruby elastic-mapreduce --describe --jobflow
 [JobFlowID]
```

レスポンスは以下と同様です：

```
{
  "JobFlows": [
    {
      "Name": "Development Job Flow (requires manual termination)",
      "LogUri": "s3n:\/\/YourBucket\/FileName\/",
      "ExecutionStatusDetail": {
        "StartDateTime": null,
        "EndDateTime": null,
        "LastStateChangeReason": "Starting instances",
        "CreationDateTime": DateTimeStamp,
        "State": "STARTING",
        "ReadyDateTime": null
      },
      "Steps": [],
      "Instances": {
        "MasterInstanceId": null,
        "Ec2KeyName": "KeyName",
        "NormalizedInstanceHours": 0,
        "InstanceCount": 5,
        "Placement": {
          "AvailabilityZone": "us-east-1a"
        },
        "SlaveInstanceType": "m1.small",
        "HadoopVersion": "0.20",
        "MasterPublicDnsName": null,
        "KeepJobFlowAliveWhenNoSteps": true,
        "InstanceGroups": [
          {
            "StartDateTime": null,
            "SpotPrice": null,
            "Name": "Master Instance Group",
            "InstanceRole": "MASTER",
            "EndDateTime": null,
            "LastStateChangeReason": "",
            "CreationDateTime": DateTimeStamp,
            "LaunchGroup": null,
            "InstanceGroupId": "InstanceGroupID",
            "State": "PROVISIONING",
            "Market": "ON_DEMAND",
            "ReadyDateTime": null,
            "InstanceType": "m1.small",
            "InstanceRunningCount": 0,
            "InstanceRequestCount": 1
          },
          {
            "StartDateTime": null,
            "SpotPrice": null,
            "Name": "Task Instance Group",
            "InstanceRole": "TASK",
            "EndDateTime": null,
            "LastStateChangeReason": "",
            "CreationDateTime": DateTimeStamp,
            "LaunchGroup": null,
            "InstanceGroupId": "InstanceGroupID",
            "State": "PROVISIONING",
            "Market": "ON_DEMAND",
            "ReadyDateTime": null,
            "InstanceType": "m1.small",
```

```
            "InstanceRunningCount": 0,
            "InstanceRequestCount": 2
          },
          {
            "StartDateTime": null,
            "SpotPrice": null,
            "Name": "Core Instance Group",
            "InstanceRole": "CORE",
            "EndDateTime": null,
            "LastStateChangeReason": "",
            "CreationDateTime": DateTimeStamp,
            "LaunchGroup": null,
            "InstanceGroupId": "InstanceGroupID",
            "State": "PROVISIONING",
            "Market": "ON_DEMAND",
            "ReadyDateTime": null,
            "InstanceType": "m1.small",
            "InstanceRunningCount": 0,
            "InstanceRequestCount": 2
          }
        ],
        "MasterInstanceType": "m1.small"
      },
      "BootstrapActions": [],
      "JobFlowId": "JobFlowID"
    }
  ]
}
```

ジョブフロー変数名と値の詳細については、Amazon Elastic MapReduce Developer Guide および Amazon Elastic MapReduce API Reference を参照してください。

ジョブフローのデバッグ

Elastic MapReduce のデバッグ機能を使用するには、`credentials.json`ファイルで Amazon S3 バケットロケーションを指定する必要があります。証明書の設定 (p. 5)ステップの一部として作成したファイルで、*`log_uri`*変数を指定しました。

AWS Management Console の Elastic MapReduce タブを使用するか、Amazon S3 タブから直接閲覧することによって、Elastic MapReduce ログファイルにアクセスします。

Note

ログファイルへの書き込みが停止するタイミングと、Amazon S3 でそれが可能となるタイミングの間には、5 分間の遅延が発生します。

また Hadoop デバッギングも利用可能であり、ジョブフローの課題や問題を特定することができます。Hadoop デバッギングを有効にして設定する方法については、Amazon Elastic MapReduce Developer Guide を参照してください。

ストリーミング ジョブフローにステップを追加する

*`RunJobFlow`*変数*`KeepJobFlowAliveWhenNoSteps`*が*`True`*に設定されている場合、ジョブフローにステップを追加できます。この値は、ジョブフローの完了に成功した後も、Amazon EC2 クラスタを稼動させ続けます。*`KeepJobFlowAliveWhenNoSteps`*の既定の設定は*`True`*であり、*`--describe`*

*--jobflow [**JobFlowID**]*コマンドを使用して検証できます。ジョブフローIDを識別するには、前の特定のジョブフローについての情報を取得する (p. 12)セクションをご参照ください。

既定の変数値を使用してステップをジョブフローに追加するには

- コマンドラインのプロンプト画面から以下のコマンドを入力してください:
  - Linux および UNIX ユーザー:

```
$ ./elastic-mapreduce -j JobFlowID --stream
```

  - Windows ユーザー:

```
C:\ruby\elastic-mapreduce-cli>ruby elastic-mapreduce -j JobFlowID --stream
```

*--stream*コマンドは、既定の変数を使用してストリーミングステップを追加します。AWS Management Console 内 *Hadoop* ストリーミングは、任意の実行可能プログラムまたはスクリプトを Hadoop mapper または reducer として使用するジョブフローを、作成して実行できるようにする Hadoop の機能です。CLI または AWS Management Console のElastic MapReduce タブで、追加したばかりのステップを閲覧できます。

AWS management console でジョブフローを閲覧するには

1. https://console.aws.amazon.com/elasticmapreduce/homeの Elastic MapReduce タブに進みます。AWSにログインしない場合は、プロンプト画面で AWS アカウント証明書を入力します。
2. 更新をクリックします。
3. 追加されたステップのあるジョブフローをクリックします。
4. ウィンドウ下部の詳細ペインで、ステップタブをクリックします。

追加されたステップについての情報が、ステップタブに表示されます。

ジョブフローの終了

ジョブフローでの作業が完了したら、それを終了し、AWS リソースの使用に伴う課金がこれ以上行なわれないようにします。

ジョブフローを終了するには

- コマンドラインのプロンプト画面から以下のコマンドを入力してください:
  - Linux および UNIX ユーザー:

```
$ ./elastic-mapreduce --terminate JobFlowID
```

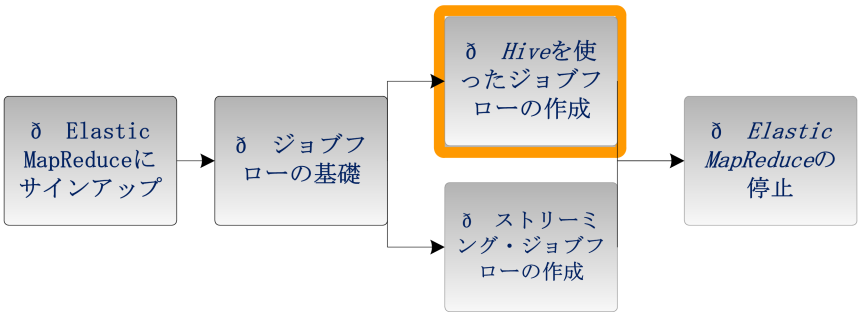
  - Windows ユーザー:

```
C:\ruby\elastic-mapreduce-cli>ruby elastic-mapreduce --terminate JobFlowID
```

おめでとうございます！Elastic MapReduce インスタンスの作成と終了に成功しました。また利用可能ないくつかのオプションについて学習しました。

これで、ジョブフローの作成、デバッグ、終了方法について知ることができました。ストリーミングジョブフローを作成する (p. 17)へ進んでください。

ストリーミング ジョブフローを作成する



本例は、Hadoop ストリーミングを使用して、データセットの単語発生回数をカウントする方法について表示するものです。特定のエラーについてたくさんのログを検索したい場合、または各ユーザー名に対してなされたブログ投稿数を知りたい場合、このタイプのジョブフローが適切となります。Hadoop ストリーミングは、Python、Ruby や PHP などの言語で記述された MapReduce プログラムの実行を可能にします。

単語の発生回数をカウントするには、入力データと出力データのワードカウントペアを通じて反復する mapper 関数を必要とします。以下の例が示すように、Python で mapper 関数を作成することができます。

```
#!/usr/bin/python

 import sys
 import re

 def main(argv):
   line = sys.stdin.readline()
   pattern = re.compile("[a-zA-Z][a-zA-Z0-9]*")
   try:
     while line:
       for word in  pattern.findall(line):
         print  "LongValueSum:" + word.lower() + "\t" + "1"
       line =  sys.stdin.readline()
```

```
  except "end of file":
    return None
if __name__ == "__main__":
main(sys.argv)
```

Amazon Elastic MapReduce で Hadoop ストリーミングジョブを実行するには、この mapper 関数が Amazon S3 にアップロードされている必要があります。

Python スクリプトを、お客様自身の Amazon S3 ロケーションに保存できます。参考として、本例は Amazon S3 のロケーション `s3://elasticmapreduce/samples/wordcount/wordSplitter.py` に保存されています。

このジョブフローのサンプル入力データは`s3://elasticmapreduce/samples/wordcount/input` にあります。

本例は、*`aggregate`*と呼ばれる埋め込み reducer を使用します。この reducer は、wordSplitter mapper 関数によって出力される単語数を加算するものです。単語の接頭辞に ロング型のデータを使用することが知られています。

ストリーミング ジョブフローを実行するには

- コマンドラインのプロンプト画面から以下のコマンドを入力してください:
  - Linux および UNIX ユーザー:

```
 $ ./elastic-mapreduce --create --stream \
     --mapper  s3://elasticmapreduce/samples/wordcount/wordSplitter.py \
     --input   s3://elasticmapreduce/samples/wordcount/input \
    --output  [A path to a bucket you own on Amazon S3, such as, s3n://my-
bucket] \
     --reducer aggregate
```

  - Windows ユーザー:

```
C:\ruby\elastic-mapreduce-cli>ruby elastic-mapreduce --create --stream \
     --mapper  s3://elasticmapreduce/samples/wordcount/wordSplitter.py \
     --input   s3://elasticmapreduce/samples/wordcount/input \
    --output  [A path to a bucket you own on Amazon S3, such as, s3n://my-
bucket] \
     --reducer aggregate
```

出力結果は次のようになります :

```
Created job flow JobFlowID
```

本例は、実行に数分を要する場合があります。特定のジョブフローについての情報を取得する (p. 12) の手順で説明されたように、または AWS management console の Elastic MapReduce タブから、このジョブフローを監視できます。

ストリーミング ジョブフローを表示するには

1. https://console.aws.amazon.com/elasticmapreduce/homeの Elastic MapReduce タブに進みます。AWSにログインしない場合は、プロンプト画面で AWS アカウント証明書を入力します。
2. 更新をクリックします。
3. Hadoop ストリーミング ジョブフローをクリックします。Hadoop ストリーミング ジョブフローは、*STATE*でリストアップされます。
4. デバッグをクリックします。

   ジョブフロー*STATE*が`COMPLETED`である場合、Elastic MapReduce ログファイルへのリンクが表示されます。
5. ジョブフローが完了していない場合は、閉じるをクリックして、 1 分間待機してから再度手順 4 を試みてください。

   Note

   [アクション]欄には、ジョブの閲覧へのリンクがあります。リンクをクリックすると、警告が表示されます。このジョブフローの作成時にデバッギングを有効にしなかったため、[ジョブ]、[タスク]および[タスク試行]は利用できません。こうしたその他の結果を作成するには、Hadoop デバッギングを有効にして設定する必要があります。

6. Elastic MapReduce ログファイルを閲覧したら、閉じるをクリックします。

`credentials.json`ファイルで指定した Amazon S3 バケットには、その他のElastic MapReduce ログファイルも含まれています。

これらのログに関する情報については、Amazon Elastic MapReduce Developer Guide を参照してください。

Tip

Hadoop ストリーミング ジョブフローを実行する度に、新しい`--output`ロケーションまたは機能しなくなるジョブフローを指定する必要があります。新しいバケットを作成するだけでなく、既存のバケット内でフォルダを指定することもできます。

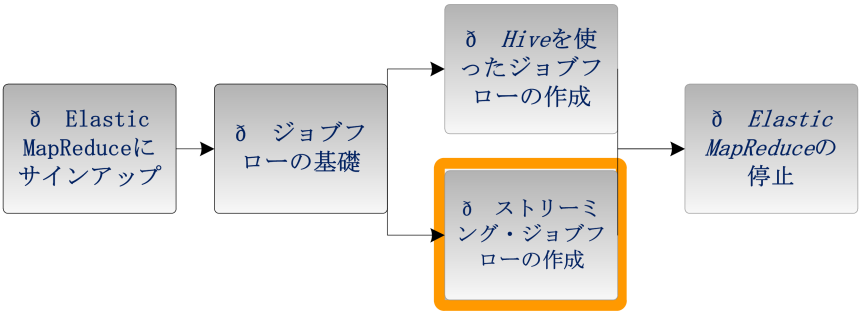
ジョブフローの結果を表示するには

1. https://console.aws.amazon.com/s3/home の Amazon S3 タブに進みます。AWSにログインしない場合は、プロンプト画面で AWS アカウント証明書を入力します。
2. `--output`で参照した Amazon S3 バケットに移動します。

お客様のジョブフローの結果は、テキストファイルの形式で保存されます。結果のファイルには、データセットの単語発生回数と共に、見つかった全単語の一覧が含まれています。

Hadoop ストリーミング ジョブフローが完了したので、Hive を用いたジョブフローの作成 (p. 20)へ移動します。

Hive を用いたジョブフローの作成



Topics

このサンプル Hive スクリプトは、広告インプレッションとクリックログデータを組み合わせ、ターゲット型オンライン広告の成功度を評価するものです。このスクリプトは、ログデータの 2 つのセットを組み合わせ、情報を Hive クラスタに配置し、結果を特定のディレクトリに出力します。以下のスクリプトは、2009-04-13 8:00 と 2009-04-13 9:00 の間に発生し、Mozilla ブラウザから twitter.com に照会されたすべてのインプレッションを処理します。

このビジネスの問題に関する詳細な説明は、次のチュートリアルにあります。*Hive および Amazon Elastic MapReduce をもちいたコンテクスト型広告*
http://developer.amazonwebservices.com/connect/entry!default.jspa?categoryID=269&externalID=2855

Hive は、データをまとめ、Hadoop ファイルに保存されている大規模なデータセットにクエリ問い合わせや分析を行なうためのツールです。SQL を基にした Hive QL と呼ばれる単純なクエリ言語が提供されています。Hive によって、従来の map/reduce プログラマーが、さらに洗練された分析を行なうために、カスタム mapper やレデューサーに入力を行なうことができます。

Hive スクリプトの作成

参考用に、サンプルスクリプトが Amazon S3 の `s3://elasticmapreduce/samples/hive-ads` に保存されています。このスクリプトをお客様自身の Amazon S3 ロケーションに保存して、適宜 Hive コマンドを変更することもできます。

このジョブフローのサンプルデータは
s3://elasticmapreduce/samples/hive-ads/libs/twitter-impressions.qにあります。

コメントされたスクリプトは次のとおり:

- 広告インプレッションのデータを読み込むために、カスタム SerDe が使用されます。

```
ADD JAR ${SAMPLE}/libs/jsonserde.jar ;
```

- 外部テーブルが作成され、広告インプレッションのデータの編成方法について、Hive に指示を与えます。

```
CREATE EXTERNAL TABLE impressions (
   requestBeginTime string, adId string, impressionId string, referrer string,

    userAgent string, userCookie string, ip string
  )
  PARTITIONED BY (dt string)
  ROW FORMAT
    serde 'com.amazon.elasticmapreduce.JsonSerde'
    with serdeproperties ( 'paths'='requestBeginTime, adId, impressionId,
    referrer, userAgent, userCookie, ip' )
  LOCATION '${SAMPLE}/tables/impressions' ;
```

- 単一のパーティションテーブルが作成され、時間を基にしてパーティションを設定されます。

```
ALTER TABLE impressions ADD PARTITION (dt='2009-04-13-08-05');
```

- 一時的なテーブルがジョブフローのローカルHDFSパーティションに作成され、広告のインプレッションとクリックに関する中間データを保存します。

```
CREATE TABLE tmp_impressions (
   requestBeginTime string, adId string, impressionId string, referrer string,

    userAgent string, userCookie string, ip string
  )
  STORED AS SEQUENCEFILE ;
```

- 特定期間の広告インプレッションテーブルのデータは、パーティションが設定されたテーブルに挿入されます。

```
INSERT OVERWRITE TABLE tmp_impressions
    SELECT
      from_unixtime(cast((cast(i.requestBeginTime as bigint) / 1000) as int))
 requestBeginTime,
      i.adId, i.impressionId, i.referrer, i.userAgent, i.userCookie, i.ip
    FROM
      impressions i
    WHERE
      i.dt   '{DAY}-${HOUR}-00' and i.dt < '{NEXT_DAY}-${NEXT_HOUR}-00'
;
```

- 特定のインプレッションデータは、Amazon S3 の出力テーブルに保存されます。

```
CREATE EXTERNAL TABLE output_impressions (
   requestBeginTime string, adId string, impressionId string, referrer string,

    userAgent string, userCookie string, ip string
    )
    PARTITIONED BY (day string, hour string)
    ROW FORMAT DELIMITED
    FIELDS TERMINATED BY '\t'
    LINES TERMINATED BY '\n'
    STORED AS TEXTFILE
    LOCATION '${OUTPUT}/impressions'
  ;
```

- 出力テーブルには、特定の期間 Mozilla ブラウザから twitter.com に照会された、すべての広告インプレッションが追加されます。

```
INSERT OVERWRITE TABLE output_impressions PARTITION (day='${DAY}',
hour='${HOUR}')
  SELECT
    i.requestBeginTime, i.adId, i.impressionId, i.referrer, i.userAgent,
i.userCookie, i.ip
  FROM
    tmp_impressions i
  WHERE
    i.referrer = 'twitter.com' and i.userAgent like '%Mozilla%'
    ;
```

Hive を用いてジョブフローを起動する

Hive をもちいてジョブフローを実行するには、CLI をもちいて Elastic MapReduce ジョブフローを作成し、ジョブフローのマスターノードにログインして、Hive スクリプトを起動します。

Hive を使用してジョブフローを作成するには

- コマンドラインのプロンプト画面から以下のコマンドを入力してください:
  - Linux および UNIX ユーザー:

```
$ ./elastic-mapreduce --create --alive --name "Hive Job Flow" --hive-inter
active
```

  - Windows ユーザー:

```
C:\ruby\elastic-mapreduce-cli>ruby elastic-mapreduce --create --alive --
name "Hive Job Flow" --hive-interactive
```

出力結果は次のようになります :

```
Created job flow JobFlowID
```

このジョブフローは、*STARTING*から*WAITING*の状態に移行するのに数分間を要します。特定のジョブフローについての情報を取得する (p. 12)の手順で説明されたように、または AWS management console の Elastic MapReduce タブから、このジョブフローを監視できます。

CLIをもちいて、アクティブなジョブフローをすべてリストアップするには

- コマンドラインのプロンプト画面から以下のコマンドを入力してください:
  - Linux および UNIX ユーザー:

    ```
    $ ./elastic-mapreduce --list --active
    ```

  - Windows ユーザー:

    ```
    C:\ruby\elastic-mapreduce-cli>ruby elastic-mapreduce --list --active
    ```

アクティブなジョブフローは最初、次のようになります :

```
JobFlowID      STARTING
     Hive Job Flow
   PENDING        Setup Hive
```

ジョブフローが Hive スクリプトを受け付ける準備が整うと、それは次のようになります : :

```
JobFlowID      WAITING        ec2-184-72-128-177.compute-1.amazonaws.com
     Hive Job Flow
   COMPLETED      Setup Hive
```

マスターノードに接続するには、そのマスターノードに対するDNSと、ルートログインが必要です。DNSは、アクティブなジョブフローの出力結果に表示されます。この例では、DNS は `ec2-184-72-128-177.compute-1.amazonaws.com`です。ルートログインまたはユーザー名はhadoopです。

ジョブフローが*WAITING*の状態にあるときは、SSH を使用して、マスターノードに接続してください。

マスターノードに接続するには

1. コマンドラインのプロンプト画面から以下のコマンドを入力してください:
   - Linux および UNIX のユーザー:

     ```
     & ./elastic-mapreduce --ssh --jobflow JobFlowID
     ```

     サンプルのジョブフローの*job flow ID*を使用します。
   - Windows ユーザー:
     a. PuTTYの開始

b. カテゴリリストのセッションを選択します。ホスト名欄に**hadoop@*DNS***を入力します。入力は、*hadoop@ec2-184-72-128-177.compute-1.amazonaws.com* のようになります。
c. カテゴリリストで、接続、SSHの順に展開し、Authを選択します。オプション制御SSH認証ペインが表示されます。
d. 認証用秘密鍵ファイルを取得する目的でファイルを開くをクリックし、先に生成しておいた秘密鍵ファイルを選択します。本ガイドに従っていれば、ファイル名は**mykeypair.ppk** となっているはずです。
e. [OK]をクリックすると、
f. 開くをクリックして、お客様のマスターノードに接続します。
g. PuTTYセキュリティ警告ポップアップYes をクリックします。

マスターノードへの接続に成功すると、出力結果は以下のようになります：

```
Using username "hadoop".
Authenticating with public key "imported-openssh-key"
Linux domU-12-31-39-01-5C-F8 2.6.21.7-2.fc8xen #1 SMP Fri Feb 15 12:39:36
EST 2008 i686
------------------------------------------------------------------------
------

Welcome to Amazon Elastic MapReduce running Hadoop and Debian/Lenny.

Hadoop is installed in /home/hadoop. Log files are in /mnt/var/log/hadoop.
 Check
/mnt/var/log/hadoop/steps for diagnosing step failures.

The Hadoop UI can be accessed via the following commands:

  JobTracker    lynx http://localhost:9100/
  NameNode      lynx http://localhost:9101/

------------------------------------------------------------------------
------
```

2. 以下のコマンドで、サンプルの Hive スクリプトを実行します。

```
hadoop@domU-12-31-39-07-D2-14:~$ hive \
    -d SAMPLE=s3://elasticmapreduce/samples/hive-ads \
    -d DAY=2009-04-13 -d HOUR=08 \
    -d NEXT_DAY=2009-04-13 -d NEXT_HOUR=09 \
    -d OUTPUT=[A path to a bucket and a folder you own on Amazon S3, such
as, s3://my-bucket/folder] \
    -f s3://elasticmapreduce/samples/hive-ads/libs/twitter-impressions.q
```

Hive スクリプトがジョブフローに追加されます。出力結果は以下のようになります：

```
10/08/20 14:57:34 WARN conf.Configuration: DEPRECATED: hadoop-site.xml found
 in the classpath.
Usage of hadoop-site.xml is deprecated. Instead use core-site.xml, mapred-
site.xml and hdfs-site.xml to
override properties of core-default.xml, mapred-default.xml and hdfs-de
fault.xml respectively
```

```
Hive history file=/mnt/var/lib/hive/tmp/history/hive_job_log_ha
doop_201008201457_1658787617.txt
Testing s3://elasticmapreduce/samples/hive-ads/libs/jsonserde.jar
converting to local s3://elasticmapreduce/samples/hive-ads/libs/jsonserde.jar
Added /mnt/var/lib/hive/downloaded_resources/s3_elasticmapreduce_samples_hive-
ads_libs_jsonserde.jar
to class path
Found class for com.amazon.elasticmapreduce.JsonSerde
OK
Time taken: 11.531 seconds

...


Starting Job = job_201008201445_0003, Tracking URL = http://domU-12-31-39-
01-5C-F8.compute-1.internal:
9100/jobdetails.jsp?jobid=job_201008201445_0003
Kill Command = /home/hadoop/.versions/0.20/bin/../bin/hadoop job  -
Dmapred.job.tracker=
domU-12-31-39-01-5C-F8.compute-1.internal:9001 -kill job_201008201445_0003
2010-08-20 14:59:07,714 Stage-2 map = 0%,  reduce = 0%
2010-08-20 14:59:22,254 Stage-2 map = 100%,  reduce = 0%
2010-08-20 14:59:31,450 Stage-2 map = 100%,  reduce = 33%
2010-08-20 14:59:37,608 Stage-2 map = 100%,  reduce = 100%
Ended Job = job_201008201445_0003
Loading data to table output_impressions partition {day=2009-04-13, hour=08}
30 Rows loaded to output_impressions
OK
Time taken: 64.647 seconds
```

ジョブフロー手順が完了します。

ssh または PuTTY の終了

- をタイプして、エンターを押します。

ジョブフローを終了するには

- コマンドラインのプロンプト画面から以下のコマンドを入力してください:
  - Linux および UNIX ユーザー:

```
$ ./elastic-mapreduce --terminate JobFlowID
```

  - Windows ユーザー:

```
C:\ruby\elastic-mapreduce-cli>ruby elastic-mapreduce --terminate JobFlowID
```

ジョブフローの結果を閲覧するには

1. https://console.aws.amazon.com/s3/home の Amazon S3 タブに進みます。AWSにログインしない場合は、プロンプト画面で AWS アカウント証明書を入力します。
2. Amazon S3 バケットと、*-d OUTPUT*の一部として Hive スクリプトで参照したパスに移動します。本サンプルの結果は、フォルダのテキストファイルにあります。
   `\impressions\day=2009-04-13\hour=08`

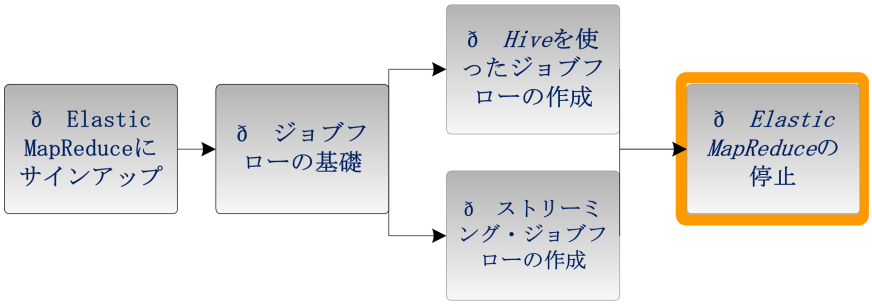
お客様のジョブフローの結果は、テキストファイルの形式で保存されます。

`credentials.json`ファイルで指定した Amazon S3 バケットには、その他のElastic MapReduce ログファイルも含まれています。

これらのログに関する情報については、Amazon Elastic MapReduce Developer Guide を参照してください。

これで Hive をもちいたジョブフローが完了したので、リソースをクリーンアップし、不必要な課金が発生しないようにする方法を学んでください。これを行なうには、Elastic Map Reduce の停止 (p. 27)に移動します。

Elastic Map Reduce の停止



本ガイドで説明されるElastic MapReduce のサンプルを完了しています。

残りのサービスに対して課金されないようにするために、望まないジョブフローやファイルを、Elastic MapReduce と Amazon S3 サービスから削除してください。

Elastic MapReduce のジョブフローを停止する

アクティブなジョブフローを列挙して、そして必要なくなったものを終了することによって、Elastic MapReduce のリソースを使用していないことを確認できます。

アクティブなジョブフローをすべて列挙するには

- コマンドラインのプロンプト画面から以下のコマンドを入力してください:
  - Linux および UNIX ユーザー:

```
$ ./elastic-mapreduce --list --active
```

  - Windows ユーザー:

```
C:\ruby\elastic-mapreduce-cli>ruby elastic-mapreduce --list --active
```

ジョブフローIDを使用して、終了したい各ジョブフローを識別してください。

ジョブフローを終了するには

- コマンドラインのプロンプト画面から以下のコマンドを入力してください:
  - Linux や Unix のユーザーは、コマンドラインのプロンプト画面から、以下を入力してください:

```
$ ./elastic-mapreduce --terminate [job flow ID]
```

  - Windows のユーザーは、コマンドラインのプロンプト画面から、以下を入力してください：

```
C:\ruby\elastic-mapreduce-cli>ruby elastic-mapreduce  --terminate [job
flow ID]
```

すべてのジョブフローを終了すると、関連するすべての Amazon EC2 インスタンスが削除されます。

ログファイルの削除

証明書の設定 (p. 5)手順の一部として、*log-uri*を指定することによって、お客様の全ジョブフローは Elastic MapReduce ログを生成し、それらを Amazon S3 に保存しました。

Elastic MapReduce ログファイルをこれ以上必要としない場合は、Amazon S3 ストレージに対して課金されないように、ファイルを削除してください。

Amazon S3 のファイルを削除する

1. https://console.aws.amazon.com/s3/home の Amazon S3 タブに進みます。AWSにログインしない場合は、プロンプト画面で AWS アカウント証明書を入力します。
2. バケットペインでバケット名をクリックすることによって、*log-uri*として指定されたバケットとフォルダに移動します。それから、オブジェクトとフォルダペインでフォルダをクリックします。
3. アクションをクリックして、削除を選択し、フォルダとそのコンテンツすべてを削除します。

本チュートリアルの一部として使用していたサービスに対して、現在は課金されていません。

おめでとうございます！正常に起動し、接続し、ジョブフローを終了しました。Amazon Elastic MapReduce の詳細およびどのように続けるかについては、ここからどこへ進むべきですか？ (p. 29)をご覧ください。

ここからどこへ進むべきですか？

Topics

Amazon Elastic MapReduce は、本ガイドでは網羅しきれないほど多くの機能を提供する贅沢なサービスです。これには例えば、Hadoop ロギング& Pig、Custom JAR ジョブフロー& Bootstrap Action&、および仮想プライベートネットワーキングがあります。本セクションは、その他のリソースへのリンクを提供し、Elastic MapReduce の理解を深める手助けをします。

Elastic MapReduce にアクセスする別の方法

本ガイドには、Elastic MapReduce をもちいてジョブフローを起動 & 終了する方法が示されています。コマンドライン インターフェイス経由で Elastic MapReduce の使用を継続するか、その他のインターフェイスの 1 つを試みることができます。

コマンドライン インターフェイスの使用を継続する

Elastic MapReduce コマンドライン インターフェイスについてさらに学ぶには、Amazon Elastic MapReduce Developer Guide を参照してください。CLI は、すべてのElastic MapReduce 関数を完全にサポートしています。プログラミングを行なったり、Elastic MapReduce ライブラリを使用したり必要はありません。

Console の使用

AWS Management Console の Elastic MapReduce 管理タブには、デバッグ出力を監視する以外にも、数多くの機能が含まれています。management console における Elastic MapReduce の使用方法について学ぶには、Amazon Elastic MapReduce Developer Guide へ進んでください。また console には、お客様をサポートするためのヘルプがあります。

Web Service API に対して直接コード化を行なう

Elastic MapReduce Query API に対して直接コードを記述したい場合は、Amazon Elastic MapReduce Developer Guide へ進んでください。このガイドは、API リクエストの作成と認証方法、および API をもちいた Elastic MapReduce の使用方法について説明するものです。すべての API アクションに関する完全な説明については、Amazon Elastic MapReduce API Reference に進んでください。

Elastic MapReduce についてさらに学ぶ

本セクションは、Elastic MapReduce のその他の機能をリストアップし、詳細な情報の場所について示すものです。Elastic MapReduce についてのその他の情報は、AWS ウェブサイトのElastic MapReduce 記事&チュートリアルエリアにもあります。

ストリーミング ジョブフロー

本ガイドで提供されるストリーミング ジョブフローの例は、Amazon Elastic MapReduce の基本機能に注目しています。Elastic MapReduce でストリーミング ジョブフローを使用することに関する詳細な情報は、以下のチュートリアルを参照してください：

- チュートリアル: Amazon Elastic MapReduce、Python、および Hadoop ストリーミングと似ている商品を見つける http://aws.amazon.com/articles/2294

Hive をもちいたジョブ f ロー

本ガイドで提供される Hive を用いたジョブフロー例は、Amazon Elastic MapReduce で Hive を使用する場合の基本機能に注目しています。Elastic MapReduce で Hive を使用することに関する詳細な情報については、以下を参照してください：

- チュートリアル: High Performance Computing のインスタンスで Apache Hive と Amazon Elastic MapReduce を使用したコンテクスト型広告 http://aws.amazon.com/articles/2855
- ビデオ: Amazon Elastic MapReduce 上の Hive 入門 http://aws.amazon.com/articles/2862

Pig をもちいたジョブフロー

Pig はオープンソースのApache ライブラリで、Hadoop 上で稼動します。ライブラリは、Pig Latin と呼ばれる言語で記述されたSQLのようなコマンドを受け付け、これらのコマンドをMapReduceジョブフローに変換します。Pig によって、SQLのようなコマンドと構文を使用してクエリを作成することができ、Java のような低水準言語を使用して MapReduce を記述する際の複雑さを回避することができます。一度に 1 つの Pig Latin コマンドを実行することもできますが、複数の Pig Latin コマンドから構成される 1 つのスクリプトを記述してタスクを達成することのほうがずっと一般的です。Elastic MapReduce は、そのようなスクリプトを Amazon S3 にアップロードする際に使用することができます。

Elastic MapReduce で Pig を使用することに関する詳細な情報については、以下を参照してください：

- チュートリアル: Apache Pig と Elastic MapReduce でログを解析 http://aws.amazon.com/articles/2729
- ビデオ: Elastic MapReduce 上の Apache Pig 入門http://-aws.amazon.com/-articles/-2735

カスタム JAR ファイルをもちいたジョブフロー

カスタム JAR ジョブフローは、Amazon S3 にアップロードされた、コンパイルされた Java プログラムを実行します。 プログラムは、起動したい Hadoop のバージョンに合わせてコンパイルするようにしてください。また、Hadoop JobClient インターフェイスをもちいて Hadoop ジョブを送信してください。

カスタム JAR ファイルで Elastic MapReduce を使用することに関する詳細な情報については、以下のチュートリアルを参照してください：

- チュートリアル: Amazon Elastic MapReduce ジョブフローの作成とデバッグの方法 http://aws.amazon.com/articles/3938

カスケーディングをもちいたジョブフロー

カスケーディングは、複雑で、自由に拡張でき、障害耐性のあるデータ処理ワークフローを Hadoop 上で定義して実行するために、APIを提供するオープンソースのプロジェクトです。

Elastic MapReduce でカスケーディングを使用することに関する詳細な情報については、以下のチュートリアルを参照してください。

- チュートリアル: Cascading Multitool http://aws.amazon.com/jobflows/2293

ブートストラップ アクション

ブートストラップ アクションは、Hadoop の開始に先立って、ジョブフローの全ノード上で実行するプログラムのことです。ブートストラップ アクションを使用すれば、以下のことが行なえます：

- ノードにソフトウェアをインストールする
- 既定の Hadoop サイト設定を修正する
- Java 変数の、Hadoop デーモンの使用方法を変更する

ジョブフローの開始時に、AWS Management Console または Elastic MapReduce コマンドライン クライアントで、ブートストラップ アクションを指定できます。あらかじめ定義されたいくつかのブート ストラップ アクションを使用できます。これには Configure Hadoop, Configure Daemons および Run-if などがあります。

ブートストラップ アクションに関する詳細な情報については、Amazon Elastic MapReduce Developer Guide または以下のチュートリアルを参照してください。

- チュートリアル: Amazon Elastic MapReduce ジョブフローの作成とデバッグの方法 http://aws.amazon.com/articles/3938

Hadoop デバッギング

Elastic MapReduce ロギングに加えて、詳細な Hadoop ログを生成するオプションもあります。Hadoop ロギングは、ジョブフロー作成時に有効になっている必要があります。また、ログを保存するには、Amazon SimpleDB にサインアップする必要があります。

Hadoop デバッギングに関する詳細な情報については、Amazon Elastic MapReduce Developer Guide を参照してください。

Hadoop の詳細を学ぶ

Apache Hadoop はオープンソースの Java ソフトウェア フレームワークであり、サーバークラスタをもちいた大規模なデータセットの処理をサポートしています。

Hadoop フレームワークの詳細な情報については、http://hadoop.apache.org/core/を参照してください。

Elastic MapReduce リソース

以下の表は、本サービスを利用する際に役立つ関連リソースをまとめたものです。

リソース	説明
Amazon Elastic MapReduce Getting Started Guide	本文書簡単な使用例を基に、サービスのクイックチュートリアルを提供します。例と手順が含まれています。
Amazon Elastic MapReduce Developer Guide	Elastic MapReduce に関する概念的な情報を提供し、Elastic MapReduce の機能の使用方法について説明します。
Amazon Elastic MapReduce API Reference	全Elastic MapReduce API の技術的説明が含まれています。
Amazon Elastic MapReduce Quick Reference Card	コマンドライン変数とそれらのオプションをすべて説明します。
Elastic MapReduce 技術上のよくある質問	この製品について開発者がよく質問する事項を網羅しています。
Elastic MapReduce リリースノート	現在のリリースについて高レベルの概要を提供します。また、新機能や修正、公となっている問題について注記を行ないます。
AWS 開発者リソースセンター	ドキュメンテーション、コード例、リリースノートをはじめとする、AWS ベースの革新的なアプリケーション開発に役立つさまざまな情報が収められた、中心的起点となるリソースセンターです。
AWS Management Console	Elastic MapReduce やその他の AWS 製品のほとんどの機能を、プログラミングなしに実行することができます。
フォーラム	Amazon Web サービスに関する技術的な質疑応答の場である開発者向けのコミュニティです。
AWSサポートセンター	当社の開発者フォーラム、技術上のよくある質問、サービスステータスページ、AWSプレミアムサポート（本プログラムに契約している場合）へのアクセスなど、AWS テクニカルサポートのホームページです。
AWS プレミアムサポート情報	AWS プレミアムサポートの情報のためにメインとなるウェブページは、マンツーマンの、対応の迅速なサポートチャネルであり、AWS インフラストラクチャ サービスで、お客様がアプリケーションを構築して実行する手助けを行なっています。
Elastic MapReduce 製品情報	Elastic MapReduce に関する情報のための主要なウェブページ

リソース	説明
お客様の AWS アカウントに関する質問のフォームお問い合わせ	このフォームは、アカウントに関する質問*専用*の窓口です。技術的な質問については、ディスカッションフォーラムをご利用ください。
利用規約	Amazon.com, Amazon.co.jp およびその関連会社における著作権・商標・その他に関する規約について詳細に説明しています。

フィードバックを提供してください

お客様のフィードバックは私どものドキュメントの改善に非常に重要です。もしよろしければ、本Elastic Map Reduce入門ガイドでのご感想・ご意見をこちらに記載をお願いいたします。

ご利用いただきどうもありがとうございます。

本ガイドについて

これは*入門ガイド*です。最終更新日 December 13, 2011.