Erste Schritte mit AWS

Analysieren von Big Data

Erste Schritte mit AWS: Analysieren von Big Data

Table of Contents

Übersicht

Der Begriff *Big Data* ist mittlerweile so geläufig, dass eine klare Definition schwer möglich ist. Ein vorherrschendes Thema in jedem Kontext ist, dass Big Data schwierige Daten sind: Es ist schwierig, sie in herkömmlichen Datenbanken zu speichern, auf Standard-Servern zu verarbeiten und mit typischen Anwendungen zu analysieren. Selbst "kleinere" Datenmengen können eine Komplexität aufweisen, die eine neue Herangehensweise erforderlich macht. Da Sie weitere Datenquellen und -typen untersuchen, müssen Sie auch Tools und Techniken identifizieren, mit denen Sie sie effizient verwalten und echten Nutzen ziehen können.

In diesem Handbuch werden zwei Verwendungsmöglichkeiten von Amazon Web Services für die Verarbeitung von Big Data erläutert. *Erste Schritte: Stimmungsanalyse* (p. 5) zeigt, wie Sie mithilfe von Hadoop Twitter-Daten auswerten können. *Erste Schritte: Analysieren von Webserver-Protokollen* (p. 18) veranschaulicht, wie Sie Protokolle von Apache-Webservern mithilfe von Hive abfragen können.

Wichtige AWS-Services für Big Data

Mit Amazon Web Services zahlen Sie nur für die Ressourcen, die Sie wirklich nutzen. Anstatt Cluster von physischen Servern und Speichergeräten zu unterhalten, um für einen möglichen Einsatz bereitzustehen, können Sie Ressourcen dann erstellen, wenn Sie benötigt werden. AWS unterstützt außerdem bekannte Tools wie Hadoop und vereinfacht die Bereitstellung, Konfiguration und Überwachung von Clustern für die Ausführung solcher Tools.

Die folgende Tabelle zeigt, wie Sie mit Amazon Web Services Big Data verwalten können.

Herausforderungen	Amazon Web Services	Vorteile
Datensätze können sehr umfangreich sein. Die Speicherung kann teuer werden und der Verlust sowie die Beschädigung von Daten können weitreichende Folgen haben.	Amazon Simple Storage Service (Amazon S3)	Amazon S3 kann große Datenmengen speichern. Um Ihre Anforderungen zu erfüllen, kann die Kapazität noch gesteigert werden. Es ist hoch redundant und sicher und bietet Schutz gegen Datenverlust und nicht autorisierte Nutzung. Außerdem hat Amazon S3 einen absichtlich kleinen Funktionssatz, um die Kosten gering zu halten.

Herausforderungen	Amazon Web Services	Vorteile
Das Verwalten eines Clusters von physischen Servern zur Datenverarbeitung ist teuer und zeitraubend.	Amazon Elastic Compute Cloud (Amazon EC2)	Wenn Sie eine Anwendung auf einem virtuellen Amazon EC2-Server ausführen, zahlen Sie für ihn nur solange die Anwendung ausgeführt wird. Sie können die Anzahl von Servern auch innerhalb von Minuten – nicht Stunden oder Tagen – erhöhen, um die verarbeitungstechnischen Anforderungen Ihrer Anwendung zu erfüllen.
Es kann eine Herausforderung darstellen, Hadoop und andere quelloffene Big-Data-Tools zu konfigurieren, zu überwachen und zu bedienen.	Amazon EMR	Amazon EMR erledigt die Konfigurierung, Überwachung und Verwaltung des Clusters. Außerdem integriert Amazon EMR Open-Source-Tools in andere AWS-Services, um die Verarbeitung großer Datenmengen in der Cloud zu vereinfachen. Somit können Sie sich auf die Datenanalyse und die Wertgewinnung konzentrieren.

Zwei Beispiele sollen verdeutlichen, wie AWS Ihnen bei der Arbeit mit Big Data helfen kann.

Einrichten

Bevor Sie AWS zum ersten Mal verwenden, führen Sie die Schritte in diesem Abschnitt aus:

- Schritt 1: Registrieren für Amazon Web Services (AWS) (p. 3)
- Schritt 2: Erstellen eines Schlüsselpaars (p. 3)

Diese Schritte müssen Sie nur einmal ausführen. Sie können Ihr Konto und Schlüsselpaar in beiden Tutorials dieses Handbuchs, *Erste Schritte: Stimmungsanalyse* (p. ?) und *Erste Schritte: Analysieren von Webserver-Protokollen* (p. ?), verwenden.

Schritt 1: Registrieren für Amazon Web Services (AWS)

Wenn Sie ein AWS-Konto erstellen, wird es automatisch für alle AWS-Services registriert. Berechnet werden Ihnen nur die Services, die Sie nutzen.

Wenn Sie bereits ein AWS-Konto haben, wechseln Sie zum nächsten Schritt. Wenn Sie kein AWS-Konto haben, führen Sie die folgenden Schritte aus, um ein Konto zu erstellen.

Registrieren Sie sich für ein AWS-Konto wie folgt:

1. Wechseln Sie zu http://aws.amazon.com/ und klicken Sie dann auf Anmelden.
2. Folgen Sie den Anweisungen auf dem Bildschirm.

   Der Anmeldeprozess beinhaltet auch einen Telefonanruf und die Eingabe einer PIN über die Telefontastatur.

Schritt 2: Erstellen eines Schlüsselpaars

Sie müssen ein Schlüsselpaar erstellen, um Verbindungen mit Amazon EC2-Instances herzustellen. Aus Sicherheitsgründen verwenden EC2-Instances ein Schlüsselpaar aus einem privaten und einem öffentlichen Schlüssel anstelle eines Benutzernamens und Kennworts für die Authentifizierung von Verbindungsanforderungen. Die öffentliche Hälfte des Schlüsselpaars ist in der Instance eingebettet. Somit können Sie den privaten Schlüssel für die sichere Anmeldung ohne Kennwort verwenden.

In diesem Schritt erstellen Sie ein Schlüsselpaar mithilfe der AWS Management Console. Zu einem späteren Zeitpunkt verwenden Sie dieses Schlüsselpaar zum Herstellen einer Verbindung mit den in den Tutorials verwendeten Amazon EC2-Instances.

To generate a key pair

1. Open the Amazon EC2 console at https://console.aws.amazon.com/ec2/.
2. In the left navigation pane, under Network and Security, click Key Pairs.
3. Click Create Key Pair.
4. Type **`mykeypair`** in the new Key Pair Name box and then click Create.
5. Download the private key file, which is named `mykeypair.pem`, and keep it in a safe place. You will need it to access any instances that you launch with this key pair.

   Important

   If you lose the key pair, you cannot connect to your Amazon EC2 instances.

   For more information about key pairs, see Getting an SSH Key Pair in the *Amazon Elastic Compute Cloud User Guide*.

Erste Schritte: Stimmungsanalyse

Topics

Die *Stimmungsanalyse* (Sentiment Analysis) bezieht sich auf verschiedene Methoden der Untersuchung und Verarbeitung von Daten zur Bestimmung subjektiver Reaktionen, meist Grundstimmungen oder Meinungen einer Gruppe zu einem bestimmten Thema. Die Stimmungsanalyse kann eingesetzt werden, um z. B. die allgemeine positive Einstellung eines Blogs oder Dokuments zu messen oder um Wählereinstellungen hinsichtlich eines politischen Kandidaten zu erfassen.

Stimmungsdaten entstammen häufig Social Media-Diensten und ähnlichen benutzergenerierten Inhalten wie Bewertungen, Kommentaren und Diskussionsgruppen. Die Datensätze nehmen also tendenziell soweit zu, dass sie als "Big Data" betrachtet werden können.

Angenommen, Ihre Firma hat vor Kurzem ein neues Produkt freigegeben und Sie möchten feststellen, wie es von den Kunden aufgenommen wird. Sie wissen, dass Social Media dazu beitragen können, eine umfangreiche Stichprobe der öffentlichen Meinung zu erfassen. Sie haben jedoch keine Zeit, um jede Erwähnung zu überwachen. Sie brauchen eine bessere Methode, um die Gesamtstimmung zu ermitteln.

Amazon EMR integriert Open-Source-Frameworks zur Datenverarbeitung in die vollständige Suite von Amazon Web Services. Die daraus resultierende Architektur ist skalierbar und effizient. Sie eignet sich ideal zum Analysieren großer Mengen von Stimmungsdaten, wie z. B. Tweets, in einem bestimmten Zeitraum.

In diesem Tutorial starten Sie einen AWS CloudFormation-Stapel, der ein Skript zum Erfassen von Tweets bereitstellt. Sie speichern die Tweets in Amazon S3 und passen eine Mapper-Datei an, um sie mit Amazon EMR zu verwenden. Anschließend erstellen Sie einen Amazon EMR-Cluster, der ein Python Natural Language Toolkit verwendet, das mit einem Hadoop-Streaming-Auftrag implementiert wird, um die Daten zu klassifizieren. Schließlich untersuchen Sie die Ausgabedateien und werten die Gesamtstimmung des Tweets aus.

Normalerweise wird für dieses Tutorial weniger als eine Stunde benötigt. Sie zahlen nur für die Ressourcen, die Sie tatsächlich nutzen. Dieses Tutorial enthält einen Schritt für die Bereinigung, um sicherzustellen, dass Ihnen keine zusätzlichen Kosten entstehen. Sie können sich auch unter dem Thema Preise (p. 29) informieren.

Important

Bevor Sie beginnen, sollten Sie sicherstellen, dass Sie die in Einrichten (p. 3) beschriebenen Schritte ausgeführt haben.

Klicken Sie auf Weiter, um das Tutorial zu starten.

Schritt 1: Erstellen eines Twitter-Entwicklerkontos

Damit Sie Tweets für die Analyse erfassen können, müssen Sie ein Konto auf der Entwickler-Website von Twitter sowie Anmeldeinformationen für die Verwendung der Twitter-API erstellen.

Erstellen Sie ein Twitter-Entwicklerkonto wie folgt:

1. Wechseln Sie zu https://dev.twitter.com/user/login und melden Sie sich mit Ihrem Benutzernamen und Kennwort für Twitter an. Wenn Sie noch kein Twitter-Konto haben, klicken Sie auf den Link Sign up, der unter dem Feld Username angezeigt wird.
2. Wenn Sie bereits die Entwickler-Website von Twitter verwendet haben, um Anmeldeinformationen zu erstellen und Anwendungen zu registrieren, wechseln Sie zum nächsten Schritt.

   Wenn Sie die Entwickler-Website von Twitter noch nicht verwendet haben, werden Sie aufgefordert, die Website zu ermächtigen, Ihr Konto zu verwenden. Klicken Sie auf Authorize app, um fortzufahren.
3. Wechseln Sie zur Twitter-Anwendungsseite unter https://dev.twitter.com/apps und klicken Sie auf Create a new application.
4. Folgen Sie den Anweisungen auf dem Bildschirm. Unter Name, Description und Website der Anwendung können Sie beliebige Angaben machen, da Sie ja lediglich Anmeldeinformationen für dieses Tutorial und nicht eine echte Anwendung erstellen.
5. Auf der Detailseite zu Ihrer neuen Anwendung werden Werte für Consumer key und Consumer secret angezeigt. Notieren Sie diese Werte. Sie benötigen sie zu einem späteren Zeitpunkt in diesem Tutorial. Wenn gewünscht, können Sie Ihre Anmeldeinformationen in einer Textdatei speichern.
6. Klicken Sie auf der Detailseite der Anwendung unten auf Create my access token. Notieren Sie die Werte für Access token und Access token secret oder fügen Sie sie der Textdatei hinzu, die Sie im vorhergehenden Schritt erstellt haben.

Wenn Sie die Anmeldeinformationen für Ihr Twitter-Entwicklerkonto zu einem anderen Zeitpunkt abrufen müssen, können Sie unter https://dev.twitter.com/apps die Anwendung auswählen, die Sie für die Zwecke dieses Tutorials erstellt haben.

Schritt 2: Erstellen eines Amazon S3-Buckets für die Amazon EMR-Dateien

Amazon EMR-Aufträge verwenden normalerweise Amazon S3-Buckets für Eingabe- und Ausgabedatendateien sowie für Mapper- und Reducer-Dateien, die nicht von Open-Source-Tools bereitgestellt werden. Für die Zwecke dieses Tutorials erstellen Sie einen eigenen Amazon S3-Bucket, in dem Sie die Datendateien und einen benutzerdefinierten Mapper speichern.

Erstellen Sie einen Amazon S3 Bucket mithilfe der Konsole wie folgt:

1. Melden Sie sich bei der AWS Management Console an und öffnen Sie die Amazon S3-Konsole unter https://console.aws.amazon.com/s3/home.
2. Klicken Sie auf Create Bucket.
3. Geben Sie einen Namen für Ihren Bucket ein, beispielsweise **`mysentimentjob`**.

   Note

   Um die Anforderungen von Hadoop zu erfüllen, sind die mit Amazon EMR verwendeten Namen von Amazon S3-Buckets auf Kleinbuchstaben, Zahlen, Punkte (.) und Bindestriche (-) beschränkt.

4. Übernehmen Sie die Einstellung US Standard für Region und klicken Sie auf Create.
5. Klicken Sie in der Liste All Buckets auf den Namen Ihres neuen Buckets.
6. Klicken Sie auf Create Folder und geben Sie dann **`input`** ein. Drücken Sie die Eingabetaste oder klicken Sie auf das Häkchen,
7. Wiederholen Sie diesen Schritt, um einen weiteren Ordner mit der Bezeichnung **`Mapper`** auf der Ebene des input-Ordners zu erstellen.
8. Für die Zwecke dieses Tutorials (um sicherzustellen, dass alle Services die Ordner verwenden können), sollten Sie die Ordner veröffentlichen. Aktivieren Sie die Kontrollkästchen neben Ihren Ordnern. Klicken Sie auf Actions und dann auf Make Public. Klicken Sie auf OK, um die Veröffentlichung der Ordner zu bestätigen.

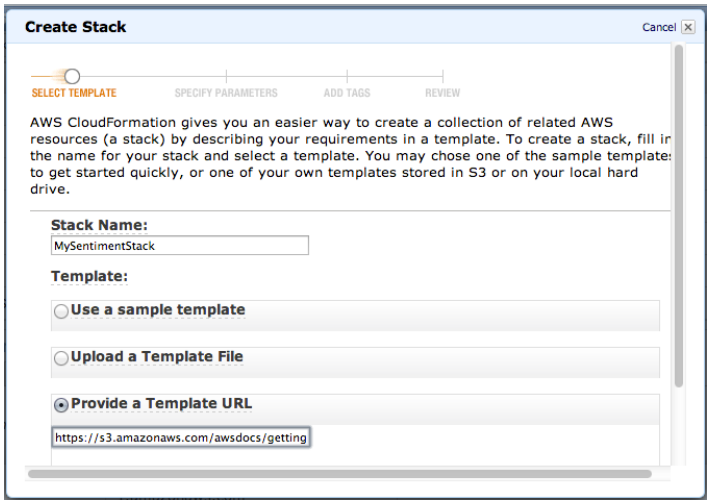
Notieren Sie die Namen Ihrer Buckets und Ordner. Sie benötigen sie zu einem späteren Zeitpunkt.

Schritt 3: Erfassen und Speichern von Stimmungsdaten

In diesem Schritt starten Sie eine Instance mithilfe einer AWS CloudFormation-Vorlage. Anschließend verwenden Sie die Tools in der Instance, um Daten per Twitter-API zu erfassen. Außerdem speichern Sie die gesammelten Daten mithilfe eines Befehlzeilen-Tools in dem von Ihnen erstellten Amazon S3-Bucket.

Starten Sie den AWS CloudFormation-Stapel wie folgt:

1. Öffnen Sie die AWS CloudFormation-Konsole unter https://console.aws.amazon.com/cloudformation.
2. Stellen Sie sicher, dass US East (N. Virginia) in der Regionenauswahl in der oberen Navigationsleiste ausgewählt ist.
3. Klicken Sie auf Create Stack.
4. Geben Sie im Feld Stack Name einen beliebigen Namen zum Kennzeichnen Ihres Stapels ein, z. B. **`MySentimentStack`**.
5. Wählen Sie unter Template die Option Provide a Template URL aus. Geben Sie **`https://s3.amazonaws.com/awsdocs/gettingstarted/latest/sentiment/sentimentGSG.template`** in das Feld ein (oder kopieren Sie die URL von dieser Seite und fügen Sie sie in das Feld ein). Klicken Sie auf Continue.



6. Geben Sie auf der Seite Specify Parameters Ihre Anmeldeinformationen für AWS und Twitter ein. Der Name für Key Pair muss mit dem Schlüsselpaar übereinstimmen, das Sie in der Region "US-East region" in Schritt 2: Erstellen eines Schlüsselpaars (p. 3) erstellt haben.

   Um die besten Ergebnisse zu erzielen, kopieren Sie die Twitter-Anmeldeinformationen von der Twitter-Entwickler-Website oder der Textdatei, in der sie gespeichert sind, und fügen Sie sie ein.

   Note

   Die Reihenfolge der Felder für die Twitter-Anmeldeinformationen auf der Seite Specify Parameters entspricht möglicherweise nicht der Anzeigereihenfolge auf der Twitter-Entwickler-Website. Stellen Sie sicher, dass Sie die richtigen Werte in die jeweiligen Felder einfügen.

7. Aktivieren Sie das Kontrollkästchen, um zu bestätigen, dass die Vorlage IAM-Ressourcen erstellen darf, und klicken Sie auf Continue. Klicken Sie auf der Seite Add Tags erneut auf Continue.
8. Überprüfen Sie die Einstellungen und stellen Sie sicher, dass Ihre Twitter-Anmeldeinformationen richtig sind. Sie können die Einstellungen ändern, indem Sie auf den Bearbeitungslink für einen bestimmten Schritt klicken.
9. Klicken Sie auf Continue, um den Stapel zu starten. Ein Bestätigungsfenster wird geöffnet. Klicken Sie auf Close.
10. Das Bestätigungsfenster wird geschlossen und Sie kehren zur AWS CloudFormation-Konsole zurück. Ihr neuer AWS CloudFormation-Stapel wird in der Liste mit dem Status CREATE_IN_PROGRESS angezeigt.

    Note

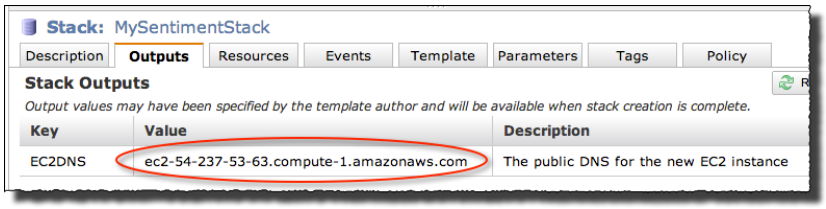
    Es dauert einige Minuten, bis Ihr Stapel gestartet ist. Klicken Sie auf der Seite Stacks auf Refresh, um zu sehen, ob der Stapel erfolgreich erstellt wurde.

Weitere Informationen zu AWS CloudFormation erhalten Sie unter Walkthrough: Updating a Stack.

Erfassen Sie Tweets mithilfe des AWS CloudFormation-Stapels wie folgt:

Wenn Ihr Stapel den Status CREATE_COMPLETE hat, kann er verwendet werden.

1. Klicken Sie im unteren Bereich auf die Registerkarte Outputs, um die IP-Adresse der Amazon EC2-Instance abzurufen, die von AWS CloudFormation erstellt wurde.



2. Stellen Sie eine Verbindung mit der Instance per SSH und mithilfe des Benutzernamens `ec2-user` her. Weitere Informationen über das Verbinden mit einer Instance und das Konfigurieren von SSH-Anmeldeinformationen und -Tools erhalten Sie unter Connecting to Your Linux/UNIX Instances Using SSH oder Connecting to Linux/UNIX Instances from Windows Using PuTTY. (Lassen Sie die Abschnitte mit der Beschreibung der Dateiübertragung außer Acht.)
3. Geben Sie im SSH-Fenster den folgenden Befehl ein:

```
cd sentiment
```

4. Die Instance wurde mit Tweepy vorkonfiguriert, einem Open-Source-Paket für die Verwendung mit der Twitter-API. Python-Skripts zum Ausführen von Tweepy werden im Verzeichnis `sentiment` angezeigt. Geben Sie den folgenden Befehl ein, um zu überprüfen, ob sie vorhanden sind:

```
ls
```

   Sie sollten Dateien mit der Bezeichnung `collector.py` und `twaiter.py` sowie `twitterparams.py` sehen.
5. Geben Sie zum Erfassen von Tweets den folgenden Befehl ein, wobei *`term1`* für Ihren Suchbegriff steht.

```
python collector.py term1
```

   Einen Suchbegriff aus mehreren Wörtern müssen Sie in Anführungszeichen setzen. Beispiele:

```
python collector.py kindle
python collector.py "kindle fire"
```

   Beim Erfassungsskript wird die Groß- und Kleinschreibung nicht beachtet.
6. Drücken Sie die Eingabetaste, um das Erfassungsskript auszuführen. Im SSH-Fenster sollte die Meldung "Collecting tweets. Please wait" angezeigt werden.

   Mit dem Skript werden 500 Tweets erfasst, was einige Minuten dauern kann. Wenn Sie nach einem Thema suchen, das auf Twitter derzeit nicht populär ist (oder wenn Sie das Skript so bearbeitet haben, dass mehr als 500 Tweets erfasst werden), dauert die Ausführung des Skripts länger. Sie können es jederzeit mit der Tastenkombination STRG+C unterbrechen.

   Wenn die Ausführung des Skripts beendet ist, wird im SSH-Fenster die Meldung "Finished collecting tweets" angezeigt.

   Note

   Wenn Ihre SSH-Verbindung während der Ausführung des Skripts unterbrochen wird, stellen Sie erneut eine Verbindung mit der Instance her und führen Sie das Skript mit nohup aus (z. B. `nohup python collector.py > /dev/null &`).

Speichern Sie die erfassten Tweets in Amazon S3 wie folgt:

Ihr Stimmungsanalysestapel wurde mit s3cmd, einem Befehlszeilen-Tool für Amazon S3 vorkonfiguriert. Sie verwenden s3cmd zum Speichern der Tweets in dem Bucket, den Sie zuvor erstellt haben.

1. Geben Sie im SSH-Fenster den folgenden Befehl ein: (Das aktuelle Verzeichnis sollte noch `sentiment` sein. Wenn dies nicht der Fall ist, navigieren Sie mit `cd` zum Verzeichnis `sentiment`.)

   ```
   ls
   ```

   Sie sollten eine Datei mit dem Namen `tweets.`*`date-time`*`.txt` sehen, wobei *`date`* das Datum und *`time`* die Uhrzeit der Ausführung des Skripts wiedergeben. Diese Datei enthält die ID-Nummern und den vollständigen Text der Tweets, die mit Ihren Suchbegriffen übereinstimmen.
2. Geben Sie den folgenden Befehl ein, um die Twitter-Daten nach Amazon S3 zu kopieren. Hierbei ist *`tweet-file`* die Datei, die Sie im vorherigen Schritt bestimmt haben, und *`your-bucket`* ist der Name des zuvor erstellten Amazon S3-Buckets.

   ```
   s3cmd put tweet-file s3://your-bucket/input/
   ```

   Beispiel:

   ```
   s3cmd put tweets.Nov12-1227.txt s3://mysentimentjob/input/
   ```

   Important

   Vergessen Sie nicht den abschließenden Schrägstrich, um anzugeben, dass "input" ein Ordner ist. Andernfalls erstellt Amazon S3 ein Objekt mit dem Namen `input` in Ihrem S3-Basis-Bucket.

3. Geben Sie den folgenden Befehl ein, um zu überprüfen, ob die Datei auf Amazon S3 hochgeladen wurde:

   ```
   s3cmd ls s3://your-bucket/input/
   ```

   Sie können auch mit der Amazon S3-Konsole unter https://console.aws.amazon.com/s3/ den Inhalt Ihrer Buckets und Ordner anzeigen.

Schritt 4: Anpassen des Amazon EMR-Mappers

Wenn Sie eigene Hadoop-Streaming-Programme erstellen möchten, müssen Sie ausführbare Mapper- und Reducer-Dateien schreiben. Weitere Informationen erhalten Sie unter Process Data with a Streaming Cluster im *Entwicklerhandbuch für Amazon Elastic MapReduce*. Für dieses Tutorial steht ein Mapper-Skript zur Verfügung, das Sie anpassen können, um es mit Ihrem Twitter-Suchbegriff zu verwenden.

Passen Sie den Mapper wie folgt an:

1. Öffnen Sie auf Ihrem lokalen System einen Text-Editor Ihrer Wahl. Kopieren Sie das folgende Skript und fügen Sie es in eine neue Datei ein.

   ```
   #!/usr/bin/python
   ```

```
import cPickle as pickle
import nltk.classify.util
from nltk.classify import NaiveBayesClassifier
from nltk.tokenize import word_tokenize
import sys

sys.stderr.write("Started mapper.\n");


def word_feats(words):
    return dict([(word, True) for word in words])


def subj(subjLine):
    subjgen = subjLine.lower()
    # Replace term1 with your subject term
    subj1 = "term1"
    if subjgen.find(subj1) != -1:
        subject = subj1
        return subject
    else:
        subject = "No match"
        return subject


def main(argv):
    classifier = pickle.load(open("classifier.p", "rb"))
    for line in sys.stdin:
        tolk_posset = word_tokenize(line.rstrip())
        d = word_feats(tolk_posset)
        subjectFull = subj(line)
        if subjectFull == "No match":
            print "LongValueSum:" + " " + subjectFull + ": " + "\t" + "1"
        else:
            print "LongValueSum:" + " " + subjectFull + ": " + classifi
er.classify(d) + "\t" + "1"


if __name__ == "__main__":
    main(sys.argv)
```

2. Bearbeiten Sie die folgende Skriptzeile:

```
subj1 = "term1"
```

Ersetzen Sie *term1* durch den Suchbegriff, den Sie in Schritt 3: Erfassen und Speichern von Stimmungsdaten (p. 7) verwendet haben. Beispiel:

```
subj1 = "kindle"
```

Important

Stellen Sie sicher, dass Sie die Formatierung der Datei nicht ändern. Falsche Einrückungen bewirken, dass das Hadoop-Streaming-Programm fehlschlägt.

Speichern Sie die bearbeitete Datei. Vielleicht möchten Sie das Skript generell durchsehen, um ein Gefühl für die Funktionsweise des Mappers zu bekommen.

Note

In Ihren eigenen Mappern möchten Sie die Konfiguration wahrscheinlich vollständig automatisieren. Das manuelle Bearbeiten in diesem Tutorial dient nur zur Veranschaulichung. Weitere Informationen zum Erstellen von Amazon EMR-Arbeitsschritten und Bootstrap-Aktionen erhalten Sie unter Create Bootstrap Actions to Install Additional Software und Steps im Entwicklerhandbuch für Amazon Elastic MapReduce.

3. Wechseln Sie zur Amazon S3-Konsole unter https://console.aws.amazon.com/s3/ und suchen Sie den Ordner `mapper`, den Sie in Schritt 2: Erstellen eines Amazon S3-Buckets für die Amazon EMR-Dateien (p. 6) erstellt haben.
4. Klicken Sie auf Upload und folgen Sie den Anweisungen auf dem Bildschirm, um die angepasste Mapper-Datei hochzuladen.
5. Veröffentlichen Sie die Mapper-Datei, indem Sie erst die Datei und anschließend die Optionen Actions und Make Public auswählen.

Schritt 5: Erstellen eines Amazon EMR-Clusters

Important

In diesem Tutorial werden die Änderungen berücksichtigt, die im November 2013 an der Amazon EMR-Konsole vorgenommen wurden. Wenn die Ansichten Ihrer Konsole von den Abbildungen in diesem Handbuch abweichen, wechseln Sie zu einer neueren Version. Klicken Sie hierfür auf den Link, der oben in der Konsole angezeigt wird:

Mit Amazon EMR können Sie einen Cluster mit Software, Bootstrap-Aktionen und Arbeitsschritten konfigurieren. In diesem Tutorial führen Sie ein Hadoop-Streaming-Programm aus. Wenn Sie einen Cluster mit einem Hadoop-Streaming-Programm in Amazon EMR konfigurieren, geben Sie einen Mapper und einen Reducer sowie alle zugehörigen Dateien an. Die folgende Liste enthält eine Zusammenfassung der in diesem Tutorial verwendeten Dateien.

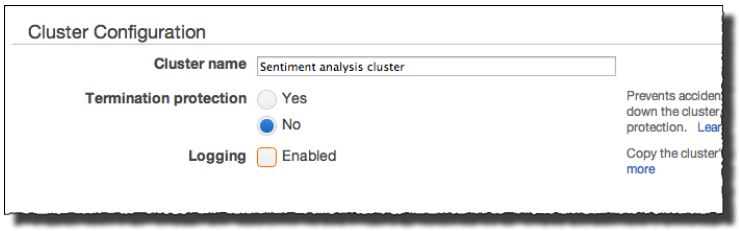
- Für den Mapper verwenden Sie die Datei, die Sie im vorhergehenden Schritt angepasst haben.
- Für die Reducer-Methode verwenden Sie das vordefinierte Hadoop-Paket `aggregate`. Weitere Informationen über das Aggregate-Paket erhalten Sie in der Dokumentation zu Hadoop.
- Üblicherweise ist für die Stimmungsanalyse eine Form der Verarbeitung natürlicher Sprache erforderlich. In diesem Tutorial verwenden Sie das Natural Language Toolkit (NLTK), eine gängige Python-Plattform. Sie installieren das Python-NLTK-Modul mithilfe einer Bootstrap-Aktion von Amazon EMR. Mit Bootstrap-Aktionen wird benutzerdefinierte Software in die von Amazon EMR bereitgestellten und konfigurierten Instances geladen. Weitere Informationen erhalten Sie unter Create Bootstrap Actions im *Entwicklerhandbuch zu Amazon Elastic MapReduce*.
- Neben dem NLTK-Module verwenden Sie eine Klassifikator-Datei für die natürliche Sprache, die in einem Amazon S3-Bucket bereitgestellt wird.
- Für die Eingabedaten und Ausgabedateien des Auftrags verwenden Sie den Amazon S3-Bucket, den Sie erstellt haben (in dem nun die erfassten Tweets enthalten sind).

Beachten Sie, dass die in diesem Tutorial verwendeten Dateien nur der Veranschaulichung dienen. Wenn Sie eigene Stimmungsanalysen durchführen, müssen Sie eigene Mapper schreiben und Stimmungsanalysemodelle erstellen, die auf Ihre Bedürfnisse zugeschnitten sind. Weitere Informationen zum Erstellen

eines Stimmungsanalysemodells erhalten Sie unter Learning to Classify Text im Buch *Natural Language Processing with Python*, das auf der NLTK-Website kostenlos zur Verfügung steht.

Erstellen Sie einen Amazon EMR-Cluster mithilfe der Konsole wie folgt:

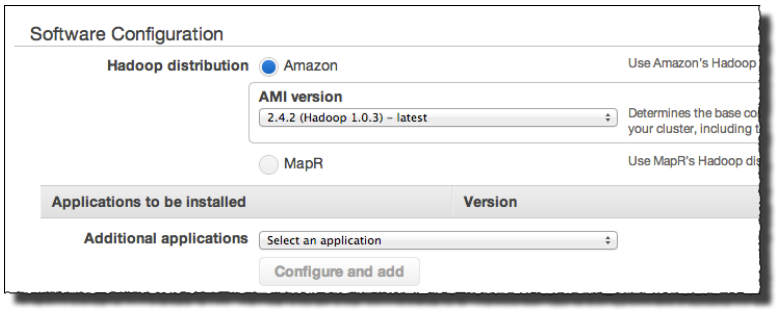
1. Öffnen Sie die Amazon EMR-Konsole unter https://console.aws.amazon.com/elasticmapreduce/.
2. Klicken Sie auf Create cluster.
3. Geben Sie im Abschnitt Cluster Configuration unter Cluster name einen Cluster-Namen ein oder verwenden Sie den Standardwert **`My cluster`**. Legen Sie Termination protection auf No fest und deaktivieren Sie das Kontrollkästchen Logging enabled.



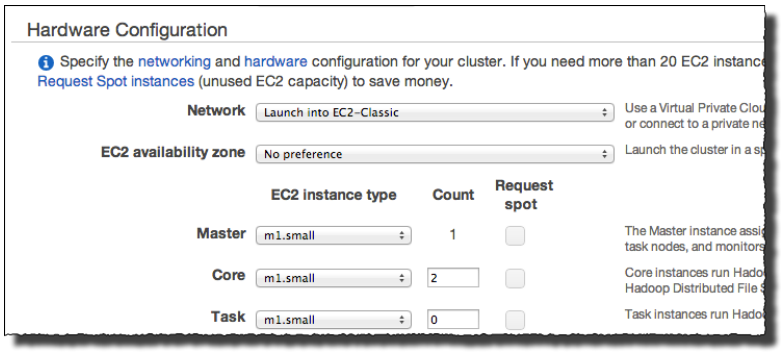
Note

In einer Produktionsumgebung können Protokollierung und Debugging sinnvolle Tools sein, um fehlerhafte oder unwirtschaftliche Schritte oder Anwendungen in Amazon EMR zu analysieren. Weitere Informationen zur Verwendung von Protokollierung und Debugging in Amazon EMR erhalten Sie unter Troubleshooting im *Entwicklerhandbuch für Amazon Elastic MapReduce*.

4. Übernehmen Sie im Abschnitt Software Configuration die Standardeinstellung für Hadoop distribution: Amazon und letzte AMI version. Klicken Sie unter Applications to be installed auf jedes "X", um Hive und Pig aus der Liste zu entfernen.



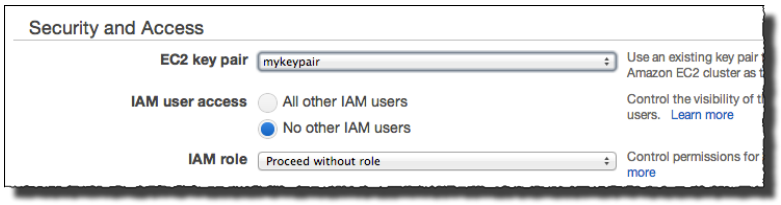
5. Behalten Sie die Standardeinstellungen im Abschnitt Hardware Configuration bei. Die standardmäßigen Instance-Typen, ein m1.small-Master-Knoten und zwei m1.small-Core-Knoten tragen dazu bei, die Kosten für dieses Tutorial gering zu halten.



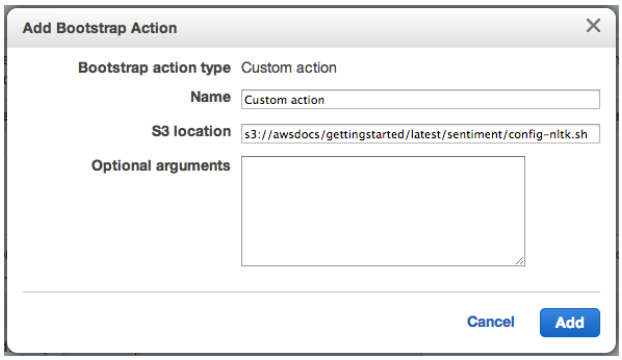
Note

Beim Analysieren von Daten in einer echten Anwendung können Sie die Größe oder Anzahl dieser Knoten erhöhen, um die Verarbeitungskapazität zu steigern und die Rechenzeit zu optimieren. Sie können auch Spot-Instances verwenden, um die Amazon EC2-Kosten weiter zu senken. Weitere Informationen zu Spot-Instances erhalten Sie unter Lowering Costs with Spot Instances im *Entwicklerhandbuch für Amazon Elastic MapReduce*.

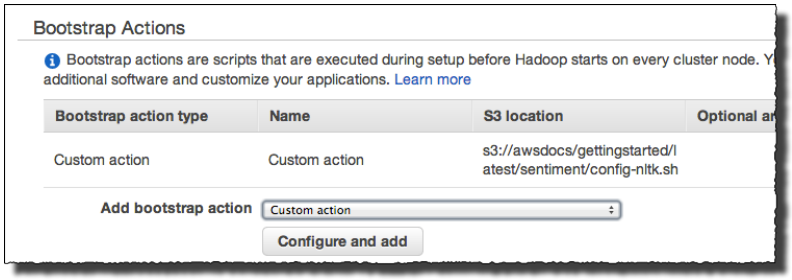
6. Wählen Sie im Abschnitt Security and Access unter EC2 key pair das EC2-Schlüsselpaar aus, das Sie zuvor erstellt haben. Behalten Sie die IAM-Standardeinstellungen bei.



7. Wählen Sie im Abschnitt Bootstrap Actions in der Liste Add bootstrap action die Option Custom action aus. Sie fügen eine benutzerdefinierte Aktion hinzu, mit der das Natural Language Toolkit im Cluster installiert und konfiguriert wird.
8. Geben Sie im Popup-Fenster Add Bootstrap Action unter Name einen Namen für die Aktion ein oder übernehmen Sie die Einstellung Custom action. Geben Sie im Feld Amazon S3 Location die Zeichenfolge **`s3://awsdocs/gettingstarted/latest/sentiment/config-nltk.sh`** ein (oder kopieren Sie die URL von dieser Seite und fügen Sie sie ein). Klicken Sie anschließend auf Add. (Bei Bedarf können Sie das Shell-Skript auch herunterladen und durchsehen.)



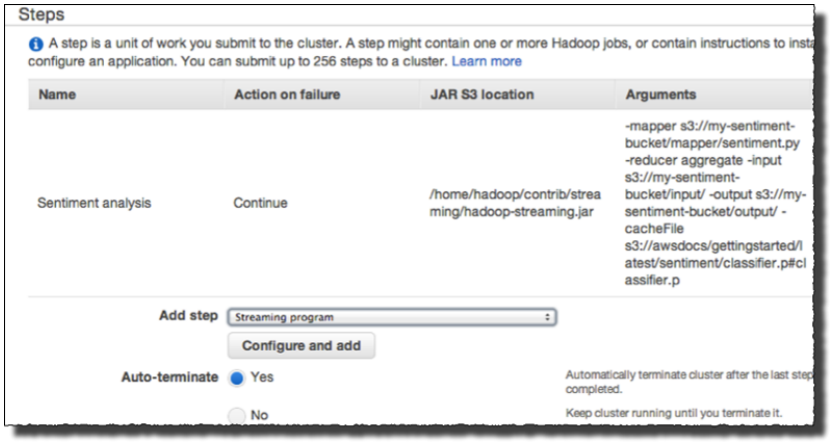
Im Bereich Bootstrap Actions müsste nun die hinzugefügte benutzerdefinierte Aktion angezeigt werden.



9. Im Abschnitt Steps legen Sie den Hadoop-Streaming-Auftrag fest. Wählen Sie in der Liste Add step die Option Streaming program aus und klicken Sie anschließend auf Configure and add.
10. Konfigurieren Sie im Popup-Fenster Add Step den Auftrag wie nachfolgend aufgeführt und ersetzen Sie dabei *`Ihr-Bucket`* durch den Namen des zuvor erstellten Amazon S3-Buckets:

Name	Sentiment analysis
Mapper	s3://*Ihr-Bucket*/mapper/sentiment.py
Reducer	Aggregate
S3-Eingabespeicherort	s3://*Ihr-Bucket*/input
S3-Ausgabespeicherort	s3://*Ihr-Bucket*/output *(stellen Sie sicher, dass dieser Ordner nicht bereits vorhanden ist)*
Argumente	-cacheFile s3://awsdocs/gettingstarted/latest/sentiment/classifier.p#classifier.p
Aktion bei Fehler	Continue

Klicken Sie auf Add. Im Abschnitt Steps müssten nun die Parameter für das Streaming-Programm angezeigt werden.



11. Legen Sie unterhalb der Schrittparameter Auto-terminate auf Yes fest.
12. Überprüfen Sie die Cluster-Einstellungen. Wenn alles richtig ist, klicken Sie auf Create cluster.

Es wird eine Zusammenfassung des neuen Clusters angezeigt, dessen Status Starting ist. Es dauert einige Minuten, bis Amazon EMR die Amazon EC2-Instances für den Cluster bereitstellt.

Schritt 6: Untersuchen der Ausgabe der Stimmungsanalyse

Wenn der Status des Clusters in der Amazon EMR-Konsole Waiting: Waiting after step completed ist, können Sie die Ergebnisse untersuchen.

Untersuchen Sie Ergebnisse des Streaming-Programms wie folgt:

1. Wechseln Sie zur Amazon S3-Konsole unter https://console.aws.amazon.com/s3/home und suchen Sie den Bucket, den Sie in Schritt 2: Erstellen eines Amazon S3-Buckets für die Amazon EMR-Dateien (p. 6) erstellt haben. Sie müssten einen neuen `output`-Ordner in Ihrem Bucket sehen. Sie müssen eventuell auf den Aktualisierungspfeil in der rechten oberen Ecke klicken, damit das neue Bucket angezeigt wird.

2. Die Auftragsausgabe ist in mehrere Dateien aufgeteilt: eine leere Statusdatei mit der Bezeichnung `_SUCCESS` sowie mehrere `part-xxxxx`-Dateien. Die `part-xxxxx`-Dateien enthalten vom Hadoop-Streaming-Programm generierte Stimmungsmesswerte.
3. Laden Sie eine Ausgabedatei herunter, indem Sie sie in der Liste auswählen, auf Actions klicken und dann Download auswählen. Klicken Sie mit der rechten Maustaste auf den Link im Popup-Fenster, um die Datei herunterzuladen.

   Wiederholen Sie diesen Schritt für jede Ausgabedatei.
4. Öffnen Sie die Dateien in einem Text-Editor. Sie enthalten die Gesamtanzahl von positiven und negativen Tweets für Ihren Suchbegriff sowie die Gesamtanzahl von Tweets, die nicht mit einem der positiven oder negativen Begriffe im Klassifikator übereinstimmen (meist, weil das Schlagwort in einem anderen Feld statt im tatsächlichen Text des Tweets stand).

   Beispiel:

```
kindle: negative      13
kindle: positive     479
No match:              8
```

   In diesem Beispiel ist die Stimmung überwältigend positiv. In den meisten Fällen liegen die Summen für positive und negative Werte näher bei einander. Für Ihre eigenen Stimmungsanalysen sollten Sie Daten über mehrere Zeiträume erfassen und vergleichen und möglicherweise unterschiedliche Suchbegriffe verwenden, um möglichst genaue Messwerte zu erhalten.

Schritt 7: Bereinigen

Um zu verhindern, dass Ihrem Konto Zusatzkosten entstehen, sollten Sie die in diesem Tutorial verwendeten Ressourcen beenden.

Löschen Sie den AWS CloudFormation-Stapel wie folgt:

1. Wechseln Sie zur AWS CloudFormation-Konsole unter https://console.aws.amazon.com/cloudformation.
2. Wählen Sie im Abschnitt AWS CloudFormation Stacks Ihren Stapel für die Stimmungsanalyse aus.
3. Klicken Sie auf die Schaltfläche Delete Stack. Sie können auch mit der rechten Maustaste auf den ausgewählten Stapel klicken und dann auf Delete Stack klicken.
4. Klicken Sie im angezeigten Bestätigungsdialogfeld auf Yes, Delete.

Note

Einmal begonnen, können Sie die Löschung des Stapels nicht abbrechen. Der Stapel durchläuft den Prozess bis zum Status DELETE_IN_PROGRESS. Nachdem der Stapel gelöscht wurde, hat er den Status DELETE_COMPLETE.

Da Sie ein Hadoop-Streaming-Programm ausgeführt und so konfiguriert haben, dass es nach Ausführung der Schritte in diesem Programm automatisch beendet wird, müsste der Cluster nach Abschluss der Verarbeitung automatisch beendet worden sein.

Stellen Sie wie folgt sicher, dass der Amazon EMR-Cluster beendet wurde:

1. Wenn die Cluster-Liste nicht bereits anzeigt wird, klicken Sie oben in der Amazon Elastic MapReduce-Konsole im Menü Elastic MapReduce auf Cluster List.
2. Vergewissern Sie sich, dass in der Cluster-Liste Status für Ihren Cluster mit Terminated angegeben ist.

Beenden Sie einen Amazon EMR-Cluster wie folgt:

1. Wenn die Cluster-Liste nicht bereits anzeigt wird, klicken Sie oben in der Amazon Elastic MapReduce-Konsole im Menü Elastic MapReduce auf Cluster List.
2. Wählen Sie in der Cluster-Liste das Feld links vom Cluster-Namen aus und klicken Sie anschließend auf Terminate. Klicken Sie im Bestätigungsfenster, das angezeigt wird, auf Terminate.

Der nächste Schritt ist optional. Hierbei wird das zuvor erstellte Schlüsselpaar gelöscht. Schlüsselpaare werden Ihnen nicht berechnet. Wenn Sie beabsichtigen, sich eingehender mit Amazon EMR zu befassen oder das andere Tutorial in diesem Handbuch abzuschließen, sollten Sie das Schlüsselpaar behalten.

Löschen Sie ein Schlüsselpaar wie folgt:

1. Wählen Sie im Navigationsbereich der Amazon EC2-Konsole Key Pairs aus.
2. Wählen Sie im Inhaltsbereich das von Ihnen erstellte Schlüsselpaar aus und klicken Sie anschließend auf Delete.

Der nächste Schritt ist optional. Hierbei werden zwei Sicherheitsgruppen gelöscht, die von Amazon EMR für Sie beim Starten des Clusters erstellt wurden. Sicherheitsgruppen werden Ihnen nicht berechnet. Wenn Sie beabsichtigen, sich eingehender mit Amazon EMR zu befassen, sollten Sie sie behalten.

Löschen Sie Amazon EMR-Sicherheitsgruppen wie folgt:

1. Klicken Sie im Navigationsbereich der Amazon EC2-Konsole auf Security Groups.
2. Klicken Sie im Inhaltsbereich auf die Sicherheitsgruppe ElasticMapReduce-slave.
3. Klicken Sie im Detailbereich der Sicherheitsgruppe "ElasticMapReduce-slave" auf die Registerkarte Inbound. Löschen Sie alle Aktionen mit Verweis auf ElasticMapReduce. Klicken Sie auf Apply Rule Changes.
4. Klicken Sie im Inhaltsbereich auf ElasticMapReduce-slave und anschließend auf Delete. Klicken Sie zur Bestätigung auf Yes, Delete. (Diese Gruppe muss gelöscht werden, bevor Sie die Gruppe "ElasticMapReduce-master" löschen können.)
5. Klicken Sie im Inhaltsbereich auf ElasticMapReduce-master und anschließend auf Delete. Klicken Sie zur Bestätigung auf Yes, Delete.

Sie haben das Tutorial zur Stimmungsanalyse abgeschlossen. Lesen Sie auch die anderen Themen dieses Handbuchs, um weitere Informationen über Amazon Elastic MapReduce zu erhalten.

Erste Schritte: Analysieren von Webserver-Protokollen

Topics

Angenommen, Sie hosten eine bekannte E-Commerce-Website. Um besser auf Ihre Kunden eingehen zu können, möchten Sie die Apache-Web-Protokolle analysieren, da Sie erfahren möchten, auf welche Art und Weise Ihre Website gefunden wird. Insbesondere möchten Sie herausfinden, welche Ihrer Online-Werbekampagnen am erfolgreichsten Datenverkehr zu Ihrem Online-Shop lenkt.

Die Webserver-Protokolle sind für den Import in eine MySQL-Datenbank jedoch zu umfangreich und verfügen nicht über ein relationales Format. Sie benötigen ein anderes Verfahren, um sie zu analysieren.

Amazon EMR integriert Open-Source-Anwendungen wie Hadoop und Hive in Amazon Web Services und stellt somit eine skalierbare und effiziente Architektur zum Analysieren großer Datenmengen, wie z. B. Apache-Web-Protokollen, bereit.

Im folgenden Tutorial importieren wir Daten von Amazon S3 und erstellen einen Amazon EMR-Cluster über die AWS Management Console. Anschließend stellen wir eine Verbindung mit dem Master-Knoten des Clusters her, in dem wir Hive ausführen, um die Apache-Protokolle mithilfe einer vereinfachten SQL-Syntax abzufragen.

Normalerweise wird für dieses Tutorial weniger als eine Stunde benötigt. Sie zahlen nur für die Ressourcen, die Sie tatsächlich nutzen. Dieses Tutorial enthält einen Schritt für die Bereinigung, um sicherzustellen, dass Ihnen keine zusätzlichen Kosten entstehen. Sie können sich auch unter dem Thema Preise (p. 29) informieren.

Important

Bevor Sie beginnen, sollten Sie sicherstellen, dass Sie die in Einrichten (p. 3) beschriebenen Schritte ausgeführt haben.

Klicken Sie auf Weiter, um das Tutorial zu starten.

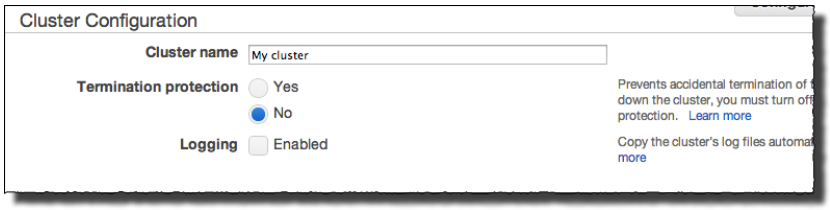
Schritt 1: Erstellen eines Clusters mithilfe der Konsole

Important

In diesem Tutorial werden die Änderungen berücksichtigt, die im November 2013 an der Amazon EMR-Konsole vorgenommen wurden. Wenn die Ansichten Ihrer Konsole von den Abbildungen in diesem Handbuch abweichen, wechseln Sie zu einer neueren Version. Klicken Sie hierfür auf den Link, der oben in der Konsole angezeigt wird:

Erstellen Sie einen Cluster mithilfe der Konsole wie folgt:

1. Melden Sie sich bei AWS Management Console an und öffnen Sie die Amazon Elastic MapReduce-Konsole unter https://console.aws.amazon.com/elasticmapreduce/.
2. Klicken Sie auf Create Cluster.
3. Geben Sie im Abschnitt Cluster Configuration unter Cluster name einen Cluster-Namen ein oder verwenden Sie den Standardwert **`My cluster`**. Legen Sie Termination protection auf No fest und deaktivieren Sie das Kontrollkästchen Logging.

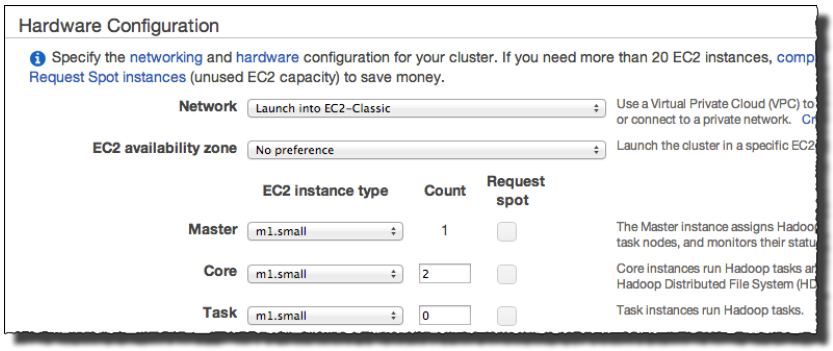


Note

In einer Produktionsumgebung können Protokollierung und Debugging sinnvolle Tools sein, um fehlerhafte oder unwirtschaftliche Schritte oder Programme in Amazon EMR zu analysieren. Weitere Informationen zur Verwendung von Protokollierung und Debugging in Amazon EMR erhalten Sie unter Troubleshooting im *Entwicklerhandbuch für Amazon Elastic MapReduce*.

4. Übernehmen Sie im Abschnitt Software Configuration die Standardeinstellung für Hadoop distribution: Amazon und letzte AMI version. Behalten Sie unter Applications to be installed die standardmäßigen Hive-Einstellungen bei. Klicken Sie auf "X", um Pig aus der Liste zu entfernen.

5. Behalten Sie die Standardeinstellungen im Abschnitt Hardware Configuration bei. Die standardmäßigen Instance-Typen, ein m1.small-Master-Knoten und zwei m1.small-Core-Knoten tragen dazu bei, die Kosten für dieses Tutorial gering zu halten.

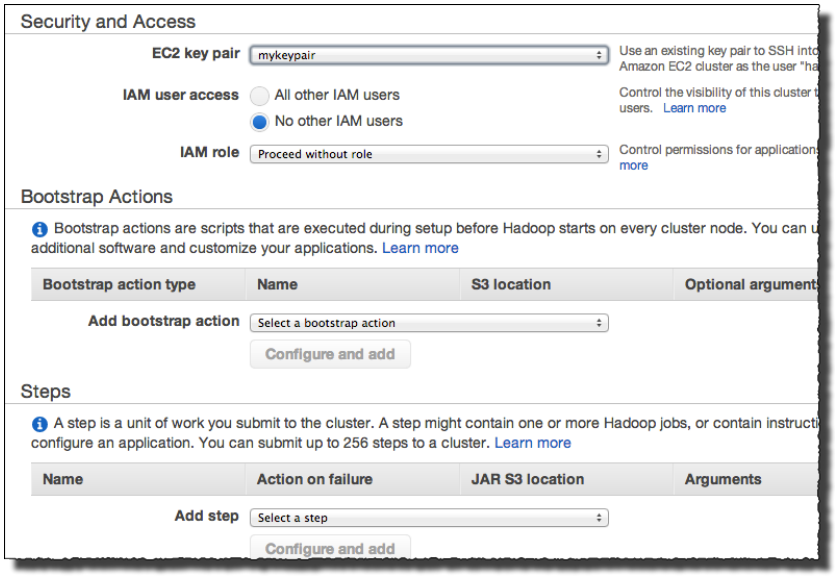


Note

Beim Analysieren von Daten in einer echten Anwendung können Sie die Größe oder Anzahl dieser Knoten erhöhen, um die Verarbeitungskapazität zu steigern und die Rechenzeit zu verkürzen. Sie können auch Spot-Instances verwenden, um die Amazon EC2-Kosten weiter zu senken. Weitere Informationen zu Spot-Instances erhalten Sie unter Lowering Costs with Spot Instances im *Entwicklerhandbuch für Amazon Elastic MapReduce*.

6. Wählen Sie im Abschnitt Security and Access unter EC2 key pair das EC2-Schlüsselpaar aus, das Sie im vorhergehenden Schritt erstellt haben. Behalten Sie die IAM-Standardeinstellungen bei.

   Übernehmen Sie die Standardeinstellungen für Bootstrap Actions und Steps. Mit Bootstrap-Aktionen und Schritten können Sie Anwendungen anpassen und konfigurieren. In diesem Tutorial wird Hive verwendet, das bereits im Amazon Machine Image (AMI, Amazon-Computerabbild) installiert ist. Eine zusätzliche Konfiguration ist also nicht erforderlich.



7. Überprüfen Sie die Einstellungen. Wenn alles richtig ist, klicken Sie auf Create cluster.

   Es wird eine Zusammenfassung des neuen Clusters angezeigt, dessen Status STARTING ist. Es dauert einige Minuten, bis Amazon EMR die Amazon EC2-Instances für den Cluster bereitstellt.

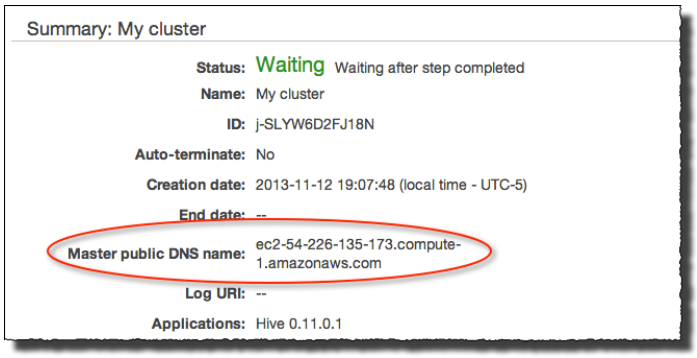
Schritt 2: Verbinden mit dem Master-Knoten

Wenn der Cluster in der Amazon EMR-Konsole den Status WAITING hat, können Sie eine Verbindung mit dem Master-Knoten herstellen. Zuerst müssen Sie den DNS-Namen des Master-Knotens abrufen und die Verbindungs-Tools sowie Ihre Anmeldeinformationen konfigurieren.

Ermitteln Sie den DNS-Namen für den Master-Knoten wie folgt:

- Wenn Sie aktuell die Seite "Cluster Details" nicht anzeigen, wählen Sie zunächst den Cluster auf der Seite "Cluster List" aus.

  Auf der Seite "Cluster Details" wird der öffentliche DNS-Name unter Master public DNS name angezeigt. Notieren Sie den DNS-Namen, da Sie ihn im nächsten Schritt benötigen.



Sie können mit Secure Shell (SSH) eine Terminal-Verbindung mit dem Master-Knoten öffnen. Bei den meisten Linux-, Unix- und Mac OS-Installationen wird eine SSH-Anwendung standardmäßig installiert. Windows-Benutzer können eine Verbindung mit dem Master-Knoten mithilfe der Anwendung PuTTY herstellen. Später in diesem Thema erhalten Sie plattformspezifische Anweisungen, um eine Windows-Anwendung zum Öffnen einer SSH-Verbindung zu konfigurieren.

Sie müssen zunächst Ihre Anmeldeinformationen konfigurieren. Andernfalls gibt SSH eine Fehlermeldung zurück, dass Ihre Datei mit dem privaten Schlüssel nicht geschützt ist, und lehnt den Schlüssel ab. Diesen Schritt müssen Sie nur bei der ersten Verwendung des privaten Schlüssels zum Herstellen der Verbindung ausführen.

Konfigurieren Sie Anmeldeinformationen unter Linux/Unix/Mac OS X wie folgt:

1. Öffnen Sie ein Terminal-Fenster. Auf den meisten Computern mit Mac OS X finden Sie das Terminal unter "Anwendungen/Dienstprogramme/Terminal". In vielen Linux-Distributionen lautet der Pfad "Applications/Accessories/Terminal".
2. Legen Sie die Berechtigungen für die PEM-Datei Ihres Amazon EC2-Schlüsselpaars so fest, dass nur der Besitzer berechtigt ist, auf den Schlüssel zuzugreifen. Wenn Sie beispielsweise die Datei als `mykeypair.pem` in Ihrem Home-Verzeichnis gespeichert haben, können Sie folgenden Befehl verwenden:

```
chmod og-rwx ~/mykeypair.pem
```

Stellen Sie eine Verbindung mit dem Master-Knoten mithilfe von Linux/Unix/Mac OS X wie folgt her:

1. Geben Sie im Terminal-Fenster den folgenden Befehl ein. Hierbei gibt der Wert des Parameters `-i` den Speicherort der Datei für den privaten Schlüssel aus Schritt 2: Erstellen eines Schlüsselpaars (p. 3) an. In diesem Beispiel wird angenommen, dass sich der Schlüssel in Ihrem Home-Verzeichnis befindet.

   ```
   ssh hadoop@master-public-dns-name \
   -i ~/mykeypair.pem
   ```

2. Es wird die Warnung angezeigt, dass die Authentizität des Hosts nicht überprüft werden konnte. Geben Sie **yes** ein, um mit dem Herstellen der Verbindung fortzufahren.

Wenn Sie einen Windows-basierten Computer verwenden, müssen Sie einen SSH-Client installieren, um eine Verbindung mit dem Master-Knoten herstellen zu können. In diesem Tutorial wird PuTTY verwendet. Wenn Sie PuTTY bereits installiert und ein Schlüsselpaar konfiguriert haben, können Sie diese Vorgehensweise überspringen.

To install and configure PuTTY on Windows

1. Download PuTTYgen.exe and PuTTY.exe to your computer from http://www.chiark.greenend.org.uk/~sgtatham/putty/download.html.
2. Launch PuTTYgen.
3. Click Load. Select the PEM file you created earlier. You may have to change the search parameters from file of type “PuTTY Private Key Files (*.ppk) to “All Files (*.*)”.
4. Click Open.
5. On the PuTTYgen Notice telling you the key was successfully imported, click OK.
6. To save the key in the PPK format, click Save private key.
7. When PuTTYgen prompts you to save the key without a pass phrase, click Yes.
8. Enter a name for your PuTTY private key, such as mykeypair.ppk.

Stellen Sie eine Verbindung mit dem Master-Knoten mithilfe von Windows/PuTTY wie folgt her:

1. Starten Sie PuTTY.
2. Klicken Sie in der Liste Category auf Session. Geben Sie im Feld Host Name den Eintrag "hadoop@*DNS*" ein. Die Eingabe wird ungefähr wie folgt angezeigt: hadoop@ec2-184-72-128-177.compute-1.amazonaws.com.
3. Erweitern Sie in der Liste Category die Optionen Connection und SSH und wählen Sie anschießend Auth aus.
4. Klicken Sie im Bereich Options controlling SSH authentication auf Browse for Private key file for authentication und wählen Sie anschließend die zuvor generierte Datei mit dem privaten Schlüssel aus. Falls Sie den Anweisungen dieses Handbuchs nachgehen, lautet der Dateiname `mykeypair.ppk`.
5. Klicken Sie auf Open.
6. Klicken Sie zum Öffnen des Master-Knotens auf Open.
7. Klicken Sie im Fenster "PuTTY Security Alert" auf Yes.

Note

Weitere Informationen zum Installieren von PuTTY und zum Herstellen einer Verbindung mit einer EC2-Instance mithilfe von PuTTY erhalten Sie unter Connecting to Linux/UNIX Instances from Windows Using PuTTY im *Amazon Elastic Compute Cloud User Guide*.

Wenn Sie die Verbindung mit dem Master-Knoten über SSH erfolgreich hergestellt haben, werden eine Willkommensnachricht und eine Eingabeaufforderung ähnlich der Folgenden angezeigt:

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

Welcome to Amazon EMR 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/

--------------------------------------------------------------------------------
hadoop@ip-10-245-190-34:~$
```

Schritt 3: Starten und Konfigurieren von Hive

Apache Hive ist eine Data Warehouse-Anwendung, mit der Sie Daten von Amazon EMR-Clustern mit einer SQL-ähnlichen Sprache abfragen können. Da Hive beim Erstellen des Clusters unter Applications to be installed aufgeführt war, steht es zur Verwendung im Master-Knoten bereit.

Um mit Hive Protokolldateien eines Webservers interaktiv abrufen zu können, müssen Sie noch einige zusätzliche Bibliotheken laden. Die Java-Archivdatei `hive_contrib.jar` im Master-Knoten enthält diese zusätzlichen Bibliotheken. Wenn Sie diese Bibliotheken laden, werden sie von Hive mit dem MapReduce-Auftrag gebündelt, der gestartet wird, um die Abfragen zu verarbeiten.

Weitere Informationen zu Hive erhalten Sie unter http://hive.apache.org/.

Starten und konfigurieren Sie Hive im Master-Knoten wie folgt:

1. Geben Sie an der Befehlszeile des Master-Knotens `hive` ein und drücken Sie anschließend die Eingabetaste.
2. Geben Sie an der Eingabeaufforderung `hive>` den folgenden Befehl ein und drücken Sie anschließend die Eingabetaste.

   ```
   hive> add jar /home/hadoop/hive/lib/hive_contrib.jar;
   ```

   Warten Sie auf die Bestätigungsmeldung ähnlich der folgenden:

```
Added /home/hadoop/hive/lib/hive_contrib.jar to class path
Added resource: /home/hadoop/hive/lib/hive_contrib.jar
```

Schritt 4: Erstellen der Hive-Tabelle und Laden von Daten in HDFS

Damit Hive mit Daten interagieren kann, müssen die Daten vom aktuellen Format (im Fall von Apache-Web-Protokollen eine Textdatei) in ein Format übersetzt werden, das als Datenbanktabelle dargestellt werden kann. Hive verwendet für diese Übersetzung einen Serializer/Deserializer (SerDe). Es gibt SerDes für eine Reihe von Datenformaten. Weitere Informationen zum Schreiben von benutzerdefinierten SerDes erhalten Sie unter Apache Hive Developer Guide.

Der in diesem Beispiel verwendete SerDe analysiert die Protokolldateidaten anhand von regulären Ausdrücken. Er stammt von der Open-Source-Community zu Hive. Sie finden ihn unter https://github.com/apache/hive/blob/trunk/contrib/src/java/org/apache/hadoop/hive/contrib/serde2/RegexSerDe.java. (Dieser Link dient nur zur Referenz. Für die Zwecke dieses Tutorials müssen Sie den SerDe nicht herunterladen.).

Mit diesem SerDe können die Protokolldateien als Tabelle definiert werden, die mit SQL-ähnlichen Anweisungen später in diesem Tutorial abgefragt wird.

Übersetzen Sie die Apache-Protokolldateidaten in eine Hive-Tabelle wie folgt:

- Kopieren Sie den folgenden mehrzeiligen Befehl. Fügen Sie den Befehl an der `hive`-Eingabeaufforderung ein und drücken Sie anschließend die Eingabetaste.

```
CREATE TABLE serde_regex(
  host STRING,
  identity STRING,
  user STRING,
  time STRING,
  request STRING,
  status STRING,
  size STRING,
  referer STRING,
  agent STRING)
ROW FORMAT SERDE 'org.apache.hadoop.hive.contrib.serde2.RegexSerDe'
WITH SERDEPROPERTIES (
  "input.regex" = "([^ ]*) ([^ ]*) ([^ ]*) (-|\\[[^\\]]*\\]) ([^
\"]*|\"[^\"]*\") (-|[0-9]*) (-|[0-9]*)(?: ([^ \"]*|\"[^\"]*\") ([^
\"]*|\"[^\"]*\"))?",
  "output.format.string" = "%1$s %2$s %3$s %4$s %5$s %6$s %7$s %8$s %9$s"
)
LOCATION 's3://elasticmapreduce/samples/pig-apache/input/';
```

Der LOCATION-Parameter im Befehl gibt den Speicherort für eine Reihe von Apache-Beispielprotokolldateien in Amazon S3 an. Ersetzen Sie die oben angegebene Amazon S3-URL durch den Speicherort Ihrer Protokolldateien in Amazon S3, um Protokolldateien Ihres Apache-Webservers zu analysieren. Um

die Anforderungen von Hadoop zu erfüllen, dürfen Namen von mit Amazon EMR verwendeten Amazon S3-Buckets nur Kleinbuchstaben, Zahlen, Punkte (.) und Bindestriche (-) enthalten.

Nachdem Sie den oben angegebenen Befehl ausgeführt haben, sollten Sie eine Bestätigung wie folgende erhalten:

```
Found class for org.apache.hadoop.hive.contrib.serde2.RegexSerDe
OK
Time taken: 12.56 seconds
hive>
```

Sobald Hive die Daten geladen hat, verbleiben sie im HDFS-Speicher (Hadoop Distributed File System), solange der Amazon EMR-Cluster ausgeführt wird, selbst wenn die Hive-Sitzung beendet und die SSH-Verbindung geschlossen wird.

Schritt 5: Abfragen von Hive

Sie können nun beginnen, die Apache-Protokolldateidaten abzufragen. Nachfolgend einige Beispielabfragen, die Sie ausführen können:

Zählen der Anzahl von Zeilen in den Protokolldateien des Apache-Webservers

```
select count(1) from serde_regex;
```

Zurückgeben aller Felder aus einer Zeile von Protokolldateidaten

```
select * from serde_regex limit 1;
```

Zählen der Anzahl von Anforderungen vom Host mit der IP-Adresse 192.168.1.198

```
select count(1) from serde_regex where host="192.168.1.198";
```

Um Abfrageergebnisse zurückzugeben, übersetzt Hive die Abfrage in einen Hadoop MapReduce-Auftrag und führt ihn im Amazon EMR-Cluster aus. Während der Ausführung des Hadoop-Auftrags werden Statusmeldungen angezeigt.

Hive SQL ist ein Teilsatz von SQL. Wenn Sie SQL kennen, können Sie problemlos Hive-Abfragen erstellen. Weitere Informationen zur Abfragesyntax erhalten Sie in der Hive-Sprachreferenz.

Schritt 6: Bereinigen

Um zu verhindern, dass Ihrem Konto Zusatzkosten entstehen, sollten Sie den Cluster nach Abschluss dieses Tutorials beenden. Da Sie den Cluster interaktiv verwendet haben, muss er manuell beendet werden.

Trennen Sie die Verbindung mit Hive und SSH wie folgt:

1. Drücken Sie im SSH-Fenster oder -Client "STRG+C", um Hive zu beenden.
2. Geben Sie an der SSH-Eingabeaufforderung `exit` ein und drücken Sie anschließend die Eingabetaste. Anschließend können Sie das Terminal- oder PuTTY-Fenster schließen.

```
exit
```

Beenden Sie einen Amazon EMR-Cluster wie folgt:

1. Wenn Ihnen die Cluster-Liste nicht bereits anzeigt wird, klicken Sie oben in der Amazon Elastic MapReduce-Konsole auf Cluster List.
2. Wählen Sie in der Cluster-Liste das Feld links vom Cluster-Namen aus und klicken Sie anschließend auf Terminate. Klicken Sie im Bestätigungsfenster, das angezeigt wird, auf Terminate.

Der nächste Schritt ist optional. Hierbei wird das zuvor erstellte Schlüsselpaar gelöscht. Schlüsselpaare werden Ihnen nicht berechnet. Wenn Sie beabsichtigen, sich eingehender mit Amazon EMR zu befassen oder das andere Tutorial in diesem Handbuch abzuschließen, sollten Sie das Schlüsselpaar behalten.

Löschen Sie ein Schlüsselpaar wie folgt:

1. Wählen Sie im Navigationsbereich der Amazon EC2-Konsole Key Pairs aus.
2. Wählen Sie im Inhaltsbereich das von Ihnen erstellte Schlüsselpaar aus und klicken Sie anschließend auf Delete.

Der nächste Schritt ist optional. Hierbei werden zwei Sicherheitsgruppen gelöscht, die von Amazon EMR für Sie beim Starten des Clusters erstellt wurden. Sicherheitsgruppen werden Ihnen nicht berechnet. Wenn Sie beabsichtigen, sich eingehender mit Amazon EMR zu befassen, sollten Sie sie behalten.

Löschen Sie Amazon EMR-Sicherheitsgruppen wie folgt:

1. Klicken Sie im Navigationsbereich der Amazon EC2-Konsole auf Security Groups.
2. Klicken Sie im Inhaltsbereich auf die Sicherheitsgruppe ElasticMapReduce-slave.
3. Klicken Sie im Detailbereich der Sicherheitsgruppe "ElasticMapReduce-slave" auf die Registerkarte Inbound. Löschen Sie alle Aktionen mit Verweis auf ElasticMapReduce. Klicken Sie auf Apply Rule Changes.
4. Klicken Sie im Inhaltsbereich auf ElasticMapReduce-slave und anschließend auf Delete. Klicken Sie zur Bestätigung auf Yes, Delete. (Diese Gruppe muss gelöscht werden, bevor Sie die Gruppe "ElasticMapReduce-master" löschen können.)
5. Klicken Sie im Inhaltsbereich auf ElasticMapReduce-master und anschließend auf Delete. Klicken Sie zur Bestätigung auf Yes, Delete.

Varianten

Die Tutorials in diesem Handbuch stellen lediglich zwei Beispiele dafür dar, wie Sie Amazon EMR einsetzen können, um mit Big Data zu arbeiten. Diese Seite bietet eine Zusammenfassung anderer Optionen, die Sie ausloten können.

Varianten für Hadoop

- Stimmungsanalyse nach Standort

  In diesem Tutorial haben wir nur den Inhalt der Tweets analysiert. Vielleicht möchten Sie in Ihren Analysen geografische Daten erfassen, aus denen Datensätze für bestimmte Regionen oder Postleitzahlbereiche erstellt werden, um Stimmungen nach Standort zu analysieren.

  Weitere Informationen zu den geografischen Aspekten von Twitter-Daten erhalten Sie in der Twitter API-Ressourcendokumentation, z. B. die Beschreibung der Ressource reverse_geocode.
- Untersuchen von Korpora und Daten

  Das Natural Language Toolkit beinhaltet verschiedene Korpora, von ausgewählten Texten vom Project Gutenberg bis hin zu Patienteninformationsbroschüren. Die Stanford Large Network Dataset Collection umfasst verschiedene Typen von Stimmungsdaten sowie Amazon-Produktbewertungsdaten. Eine Fülle von Filmkritik-Daten steht über Cornell zur Verfügung. Sie werden vielleicht feststellen, dass der Einsatz konkreterer Daten statt allgemeiner Twitter-Daten präzisere Ergebnisse liefert.
- Erproben anderer Klassifikatormodelle

  In diesem Tutorial wurde ein naiver Bayes-Klassifikator auf die Stimmungsdaten angewendet. Das Natural Language Toolkit unterstützt einige Klassifizierungsalgorithmen, so z. B. Maxent (Maximum-Entropie) und Support Vector Machines (SVMs) über das Scikit-Learn-Modul. Je nachdem, welchen Datentyp Sie analysieren und welche Funktionen Sie für die Bewertung auswählen, könnten andere Algorithmen bessere Ergebnisse liefern. Weitere Informationen erhalten Sie unter Learning to Classify Text im Buch *Natural Language Processing with Python*.

Varianten für Hive

- Erstellen eines Skripts für Hive-Abfragen

Das interaktive Abfragen von Daten führt am direktesten zu Ergebnissen. Interaktive Abfragen können hilfreich sein, um Daten zu untersuchen und die eigene Herangehensweise weiterzuentwickeln. Sobald Sie einen Satz von Abfragen erstellt haben, den Sie regelmäßig ausführen möchten, können Sie den Prozess automatisieren, indem Sie Ihre Hive-Befehle in einem Skript speichern und es nach Amazon S3 hochladen. Weitere Informationen zum Starten eines Clusters mithilfe eines Hive-Skripts erhalten Sie unter Launch a Hive Cluster im *Entwicklerhandbuch für Amazon Elastic MapReduce*.

- Verwenden von Pig statt Hive zum Analysieren von Daten

  Amazon EMR bietet Zugriff auf viele Open-Source-Tools, so auch auf Pig, das mithilfe der Sprache "Pig Latin" MapReduce-Aufträge abstrahiert. Ein Beispiel für die Analyse von Protokolldateien mit Pig erhalten Sie unter Parsing Logs with Apache Pig and Amazon Elastic MapReduce.

- Erstellen von benutzerdefinierten Anwendungen zum Analysieren von Daten

  Wenn Sie kein Open-Source-Tool finden, das Ihren Anforderungen entspricht, können Sie eine benutzerdefinierte Hadoop-MapReduce-Anwendung schreiben und in Amazon EMR ausführen. Weitere Informationen erhalten Sie unter Run a Hadoop Application to Process Data im *Entwicklerhandbuch für Amazon Elastic MapReduce*.

  Alternativ können Sie einen Hadoop-Streaming-Auftrag erstellen, mit dem Daten aus Standardeingaben gelesen werden. Ein Beispiel erhalten Sie unter Erste Schritte: Stimmungsanalyse (p. 5) in diesem Handbuch. Weitere Informationen erhalten Sie unter Launch a Streaming Cluster im *Entwicklerhandbuch für Amazon Elastic MapReduce*.

Preise

Der AWS – Einfacher Monatsrechner hilft Ihnen bei der Schätzung Ihrer monatlichen Rechnung. Er bietet eine Kostenaufschlüsselung pro Service sowie eine Schätzung der monatlichen Gesamtkosten. Sie können den Rechner auch verwenden, um eine Kostenschätzung und -aufschlüsselung für häufige Lösungen zu erhalten.

Schätzen Sie die Kosten mit "AWS – Einfacher Monatsrechner" wie folgt:

1. Wechseln Sie zu http://calculator.s3.amazonaws.com/calc5.html.
2. Wählen Sie im Navigationsbereich einen Webservice aus, den Sie derzeit verwenden oder zukünftig verwenden möchten. Geben Sie die geschätzte monatliche Nutzung für diesen Service ein. Klicken Sie auf Add To Bill, um die Kosten der Gesamtberechnung hinzufügen. Wiederholen Sie diesen Schritt für jeden Webservice, den Sie verwenden.
3. Klicken Sie auf die Registerkarte mit der Bezeichnung Estimate of Your Monthly Bill, um die geschätzten monatlichen Gesamtkosten anzuzeigen.

Laden Sie das Whitepaper How AWS Pricing Works herunter, um weitere Informationen zu erhalten. Preisdetails sind auch für jeden Service verfügbar. Beispiel: Preise für Amazon Simple Storage Service

Related Resources

The following table lists some of the AWS resources that you'll find useful as you work with AWS.

Resource	Description
AWS Products & Services	Information about the products and services that AWS offers.
AWS Documentation	Official documentation for each AWS product, including service introductions, service features, and API reference.
AWS Discussion Forums	Community-based forums for discussing technical questions about Amazon Web Services.
AWS Support	The home page for AWS Support, including access to our Discussion Forums, technical FAQs, and AWS Support Center.
Contact Us	This form is *only* for account questions. For technical questions, use the Discussion Forums.
AWS Architecture Center	Provides the necessary guidance and best practices to build highly scalable and reliable applications in the AWS cloud. These resources help you understand the AWS platform, its services and features. They also provide architectural guidance for design and implementation of systems that run on the AWS infrastructure.
AWS Security Center	Provides information about security features and resources.
AWS Economics Center	Provides access to information, tools, and resources to compare the costs of Amazon Web Services with IT infrastructure alternatives.
AWS Technical Whitepapers	Provides technical whitepapers that cover topics such as architecture, security, and economics. These whitepapers have been written by the Amazon team, customers, and solution providers.
AWS Blogs	Provides blog posts that cover new services and updates to existing services.

Resource	Description
AWS Podcast	Provides podcasts that cover new services, existing services, and tips.

Dokumentverlauf

Dieser Dokumentverlauf entspricht der Version von *Analysieren von Big Data mit AWS in Getting Started with AWS* vom 14.11.2013.

Änderung	Beschreibung	Veröffentlichungsdatum
Tutorial zur Stimmungsanalyse hinzugefügt	Neues Tutorial hinzugefügt, das die Verwendung von Amazon EMR für Stimmungsanalysen zu Twitter-Daten veranschaulicht.	14. November 2013
Amazon EMR-Konsolenreferenz aktualisiert	Wurde aufgrund der Verbesserungen an der Amazon EMR-Konsole aktualisiert.	13. November 2013
Preise für Amazon EC2 und Amazon EMR aktualisiert	Wurde aufgrund der Senkung der Gebühren für Amazon EC2 und Amazon EMR-Service aktualisiert.	6. März 2012
Preise für Amazon S3 aktualisiert	Wurde aufgrund der Senkung der Gebühren für Amazon S3-Speicher aktualisiert.	9. Februar 2012
Neuer Inhalt	Neues Dokument erstellt	12. Dezember 2011