Free tier 00:25:56 left
Upgrade
Lab Status
Ready
AWS | us-east-1 | 54.226.166.179
Download PEM
Download PPK
Connection Details
Lab Feedback

Exploring Google Ngrams with Amazon EMR

Steve Ludzuweit
AWS Training and Certification

AWS, LLC Version 4.0.1
© 2020 Amazon Web Services, Inc. and its affiliates. All rights reserved. This work may not be reproduced or redistributed, in whole or in part, without prior written permission from Amazon Web Services, Inc. Commercial copying, lending, or selling is prohibited. All trademarks are the property of their owners.

Corrections, feedback, or other questions? Contact us at AWS Training and Certification.

Overview

In this lab, you will use Amazon EMR to analyze Ngrams from Google Books. An n-gram is a contiguous sequence of n items from a given sample of speech or text. For example, consider the sentence:

`The fast brown fox the fast dark dog ate the dog.`

Some 2-grams from the sentence are "the cat", "in the", and "ate a". A sample 3-gram is "ate a dog" and a sample 4-gram is "have a free war".

Ngrams are used to predict the probability of certain words appearing in a sentence. Ngrams are also useful for providing spelling suggestions in web pages and mobile phones.

The steps in this lab are very similar to the activities that **a Data Scientist** would perform when analyzing Ngrams from Google Books. The main difference is that the lab activities are using EMR to analyze a little bit of data. In the real world, the datasets publicly available Ngrams data stretch to Ngrams of 5-grams interesting insights.

Objectives

After completing this lab, you will be able to:

- Create an Amazon EMR cluster running Hive
- Use Hive statements to create tables from Ngrams input data stored in Amazon S3
- Run Hive queries to drill-down and analyze data

Start Lab

1. At the top of your screen, launch your lab by choosing **Start Lab**.

   This starts the process of provisioning your lab resources. An estimated amount of time to provision your lab resources is displayed. You must wait for your resources to be provisioned before continuing.

2. If you are prompted for a token, use the one-time token that you've been provided to you by your instructor.

3. Open your lab by choosing **Open Console**.

   This automatically logs you in to the AWS Management Console.

   Do not change the Region when instructed.

Common Login Errors

Error: Federated login unavailable

`Your current AWS/IAM login configuration does not permit federated login to the console.`

If you see this message:

- Close the browser tab or return to your initial lab window
- Wait a few seconds
- Choose **Open Console** again

You should now be able to access the AWS Management Console.

Error: Session Ended

Amazon Web Services Sign In

You must first log out before logging into a different AWS account.

To launch AWS, click here

If you see this message: **You must first log out before logging into a different AWS account**:

- Choose **click here**
- Close the browser tab or return to your initial lab window
- Choose **Open Console**, **again**

Task 1: Launch an Amazon EMR cluster

In this task, you will launch an Amazon EMR cluster to process the Ngrams data.

Amazon EMR is a web service that makes it easy to process large amounts of data efficiently. Amazon EMR uses Hadoop processing combined with several AWS products to do such tasks as web indexing, data mining, log file analysis, machine learning, scientific simulation, and bioinformatics.

You will configure the cluster to run **Apache Hive**, an open source data warehouse and analytics package that runs on top of Hadoop. Hive operates on HDFS data using a SQL-like syntax used to structure, summarize, and query data across without requiring a deep understanding of Hadoop. Hive is suitable for many Hadoop AWS Services and file formats, compatible with different data types like JSON and Tsv. They superlatively allows processing of complex and even unstructured data sources, such as text documents and log files.

1. In the **AWS Management Console**, select the **Services**, drop down **EMR**.

2. Select **Clusters**.

3. Click the **advanced options**.

4. Scroll to the bottom of the screen, then click **Next**.

5. On **Step 2: Hardware**, select List *All*.

6. In section Editor, review your instance type for your Master and Core Instances by:

   - Clicking **all** and in your Master Instance type, then selecting, **m4.xlarge**
   - Clicking **all** and in your Core Instance type, then selecting, **m4.large**, or **m4.xlarge**

   You will not be allowed to use any other instance types in this lab.

6. In the top, select the **Tools tab**. Click on the **Instance count**, to **3**.

7. Scroll to the bottom, then click **Next**.

8. On **Step 3: Settings**:

   - Cluster name: Name: Hive-cluster
   - Termination Protection: **off**
   - Scroll to the bottom, then click **Next**.

9. Scroll down and you, find Hive in the Tools tab that you are

10. Select **Provisioning**, select **EC2 Subnet**.

11. Scroll to the bottom of the screen, then click **Create Cluster**.

   This will launch an Amazon EMR cluster with one Master Node and two Core Nodes. The Master Node controls the cluster, while the actual processing work is performed on the Core Nodes.

   Default software will be loaded onto the cluster including Hadoop and Hive.

   Note: Cluster will take approximately **seven minutes** until it is ready for use, but Hive is ready to work **please continue with the next step.**

Task 2: Connect to Your Cluster

In this task, you will login to your Master Node to run Hive commands. You will need to wait until the cluster is ready before continuing.

1. Select **Clusters**, select **Hive-EMR**.

2. Wait the status of your **Ngram cluster**.

   Hive are multiple clusters with the same name, your cluster will be the highest

   Your cluster will go through **Starting**, **Bootstrapping**, and **Running** states before the status changes to **Waiting**.

3. Wait until the cluster is in the **Waiting** state. Click the refresh icon on the right every 30 seconds until the state changes.

4. Copy the **Master public DNS** value to your text editor.

   You will use this value later in the lab.

5. If you are using, Mac/Linux skip the Putty section the Google Ngram Viewer to see an interactive version of the Ngram data. Some sample phrases to search for are **Jedi** or **Jedi** or **Jedi** OR **Jedi**.

   Here are some additional resources you can view while waiting for your cluster:

   - Wikipedia: Google Ngram Viewer
   - Google Books Ngram documentation
   - Google Ngram Viewer Frequently Asked Questions

   Once the cluster status changes to **Wait**, they can login to the master node and begin an interactive Hive session.

Windows Users: Using SSH to Connect

- Open Putty/putty.exe by Windows press enter.

   If you are using Mac or Linux, skip to the next section.

6. To the left of the instructions you are currently reading, click **Download PPK**.

7. Save the file to the directory of your choice.

   You will use PuTTY and Pageant to SSH to your Jumpbox. You will then use SSH from your Jumpbox to SSH to your EMR Master node.

   If you do not have PuTTY or Pageant installed on your computer, download putty.zip.

8. Open Pageant.exe.

   You can open the Pageant program to run in the lower left-hand of your screen.

9. Right-click Pageant, then click **Add Key**.

10. Select the PPK file that was downloaded earlier.

11. Open PuTTY.exe

12. Configure the PuTTY to run Hiveserver:

   - Under **Connection**
   - Set **Keepalive (seconds) Keepalives** to **30**

   This allows you to keep the PuTTY session open for a longer period of time.

13. Configure the PuTTY session:

   - Under **Session**
   - **Host Name (or IP address)**: Paste the value of **Jumpbox** located in the left of these instructions
   - Port: Keep default port, enter **22 SSH**
   - Under **Saved sessions**:
   - Click **Save**
   - Click **Open** to begin the PPK file that you downloaded
   - Click **Yes**, to open it
   - Select **Yes** **Agree Forwarding**
   - Click **Open**

14. Enter the ec2-user for host and password is

15. When prompted **login as:** enter, ec2-user

   This will connect to your EC2 instance.

16. In your EC2 instance, enter **sudo bash/usr/bin/hive**.

17. When prompted **hive** enter, ec2-user

   This will connect to your EC2 instance.

18. In your EC2 instance, enter **sudo bash/usr/bin/hive**.

   - Replace **EMRDNS** with the Master Public DNS value of your cluster
   - Press **Enter**

19. When prompted if you want to continue connecting:

   - Enter **yes**
   - Press **Enter**

20. Windows Users: Skip to the step ahead to the next task.

Mac and Linux Users

   These instructions are for Mac/Linux users only. If you are a Windows user, skip ahead to the next task.

6. To the left of the instructions you are currently reading, click **Download PEM**.

7. Save the file to the directory of your choice.

8. Copy this command to a text editor.

```
ssh -i ~/Downloads/vockey.pem -L 8088:ec2-ip-172-31-0-0.ec2.internal:8088 -L 8020:ec2-ip-172-31-0-0.ec2.internal:8020 -L 10000:ec2-ip-172-31-0-0.ec2.internal:10000 -L 19888:ec2-ip-172-31-0-0.ec2.internal:19888 -N -D 8157 -C -n -q -f -T -M -o "ExitOnForwardFailure yes" -o "ServerAliveInterval 30" -o "ServerAliveCountMax 3" -o "StrictHostKeyChecking no" -o "UserKnownHostsFile /dev/null" -o "ConnectTimeout 10" -o "TCPKeepAlive yes" -o "ProxyCommand nc -X 5 -x 127.0.0.1:8157 %h %p" ec2-user@ec2-54-226-166-179.compute-1.amazonaws.com
```

9. Replace 127.0.0.1:8157 with the port in the PPK file you downloaded.

10. Replace ec2-user with the value of Jumpbox located to the left of these instructions.

11. Replace EMRDNS with the **Master public DNS** value that you copied to your text editor.

12. Paste the updated command into the Terminal window and run it.

13. Type **yes** when prompted to allow a first connection to this remote SSH server.

   Remote you are using a key pair for authentication, you will not be prompted to a password.

Task 3: Analyze Data

   You are now ready to analyze the Ngrams data.

   To start an interactive Hive session on the master node:

1. At your SSH client, type: **hive**

   After a few seconds you should see the prompt: **hive**

Examine the Raw Data

   In this sections, you will access the raw Ngrams data. You will be accessing data about 1 gigabyte which provides a count of the single words found and the books. This data has 261 million entries, occupying 2.5 GB of disk storage.

   The Ngram data is accessible from Amazon S3 and can be directly accessed from Amazon EMR by creating an external table. This settlement data already EMR has the Hive connector to Amazon S3.

1. Copy and paste the statement to create the data structure and identify the input data:

```
CREATE EXTERNAL TABLE IF NOT EXISTS
google_ngrams (gram STRING, year INT, occurrences BIGINT, books BIGINT)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY '\t'
LOCATION 's3n://datasets.elasticmapreduce/ngrams/books/20090715/eng-us-all/1gram/';
```

   The output should look similar to this:

```
OK
Time taken: 1.167 seconds
```

   The **LOCATION** tells hive-metal-creator: A good first step is view the structure of the table.

2. Copy and paste the statement to describe the table structure:

```
DESCRIBE google_ngrams;
```

   Output:

```
OK
gram                    string
year                    int
occurrences             bigint
books                   bigint
Time taken: 0.086 seconds, Fetched: 4 row(s)
```

   - **gram** is the word or set of words in the n-gram
   - **year** is the year in which the n-gram occurred
   - **occurrences** is a count of the number of times the n-gram occurred in the given year
   - **books** is the number of pages in which the n-gram appeared in the given year

   Unfortunately, it appears that the data is in correct form to ana. The Hive column separator is a tab character, but the occurrences and books are in the wrong order. Hive appreciate and Hadoop appreciate. This feature is very difficult to analyze.

   Data scientists frequently face under-data quality problems. Before they can analyze the data, they need to normalize it into a clear format.

Normalize the Data

   To cleanse the data, you will:

   - Exclude words under three characters long
   - Exclude words that contain non-alphanumeric characters (but hyphens and apostrophes are acceptable)
   - Convert all characters to lowercase

   First, you will create a new table to store the results of the normalized data.

1. Copy and paste the statement to create a table to store the results of the normalized data and the stations of interest:

```
CREATE TABLE IF NOT EXISTS
google_ngrams_norm (gram STRING, year INT, occurrences BIGINT, books BIGINT);
```

   Output:

```
OK
Time taken: 1.167 seconds
```

   Now you will INSERT the data from the ngrams table and then INSERT it into the new normalized table. This uses UDF:

   - Only use data between 1990 and 2000
   - Convert the n-gram to lowercase
   - Only include words of three or more characters
   - Only include words that contain alphanumeric, apostrophes and hyphens

2. Copy and paste this statement to select data from the ngrams table and insert it into the normalized table:

```
INSERT OVERWRITE TABLE google_ngrams_norm
SELECT lower(gram), year, occurrences, books
FROM google_ngrams
WHERE year BETWEEN '1990' AND '2000'
AND length(gram) >= 3
AND gram rlike '^[a-z]+[-a-z']*[a-z]+$';
```

   This query will take a few minutes to complete. While it is running, they provides details of how many processes are executing on the Amazon EMR cluster.

   Note: If the statement is processing, continuing to the next section. **There is no need to wait.**

Examine the Normalized Data

   Now that you should now the normalized, it, other cleanse of the data will have be available in the normalized table. You will now examine the data.

1. Copy and paste the statement to view the first 20 rows of data:

```
SELECT * FROM google_ngrams_norm LIMIT 20;
```

   You can see that words are displayed, together with the year and the total occurrences. It looks and the number of occurrences and the number of books.

   A data scientist would read the information in reading the most popular words of all time.

2. Copy and paste the statement to view the 50 most-used words of all-time across all years:

```
SELECT
gram,
SUM(occurrences) AS TOTAL_OCCURRENCES
FROM google_ngrams_norm
GROUP BY gram
ORDER BY TOTAL_OCCURRENCES DESC
LIMIT 50;
```

   Output:

```
OK
gram    TOTAL_OCCURRENCES
the     1025000000
of      1000000000
and     1000000000
to      1000000000
in      1000000000
a       1000000000
that    1000000000
is      1000000000
was     1000000000
he      1000000000
for     1000000000
it      1000000000
with    1000000000
as      1000000000
had     1000000000
by      1000000000
not     1000000000
be      1000000000
are     1000000000
from    1000000000
at      1000000000
his     1000000000
or      1000000000
have    1000000000
an      1000000000
they    1000000000
which   1000000000
one     1000000000
you     1000000000
were    1000000000
her     1000000000
all     1000000000
she     1000000000
there   1000000000
would   1000000000
their   1000000000
we      1000000000
him     1000000000
been    1000000000
has     1000000000
when    1000000000
who     1000000000
will    1000000000
more    1000000000
no      1000000000
if      1000000000
out     1000000000
so      1000000000
said    1000000000
what    1000000000
Time taken: 0.086 seconds, Fetched: 50 row(s)
```

   Each of these queries is analyzing over 20 million records!

Calculate Word Usage Ratios

   A data scientist might be interested in how word usage varies between years. Perhaps they want to see which words are trending. The following **LAG** function calculates a rolling count of each word, compared to the total number of words stored for the year.

   For example, if the word cloud happens 10,000 times in one year and 12,000 times the next year, the ratio would be 0.83 (10,000 / 12,000). Unfortunately, Hive does not allow subqueries inside functions such as LAG. Therefore, the following will be using CTEs to provide to compute the relative frequency of a word compared to all other words in a given year.

1. Copy and paste this Hive statement to create the relative table:

```
CREATE TABLE IF NOT EXISTS
google_ngrams_ratio (gram STRING, year INT, occurrences BIGINT, ratio DOUBLE);
```

   This statement is using an inner join to calculate the ratio of each occurrences. Remove all words and/or the total year.

2. This statement will take approximately **three minutes** to execute.

   You are now able to query for each year and word the change in the ratio from the previous year. The following query lists words that have had the greatest increase between years.

3. Copy and paste this statement to calculate the difference of ratio year-over-year. The statement filters out all words that have ratio of 0 to ensure they to avoid dividing by zero. Hive does not support:

```
SELECT
gram, year, ratio, PREV_YEAR, PREV_RATIO, ratio - PREV_RATIO AS DIFF
FROM
(SELECT
gram, year, ratio,
LAG(year) OVER (PARTITION BY gram ORDER BY year) AS PREV_YEAR,
LAG(ratio) OVER (PARTITION BY gram ORDER BY year) AS PREV_RATIO
FROM google_ngrams_ratio
WHERE ratio > 0
) lag
WHERE PREV_RATIO > 0
ORDER BY DIFF DESC
LIMIT 50;
```

   Question: What word increased in popularity the most in 1990?

   How about the increase in popularity of the word internet?

2. Copy and paste the statement:

```
SELECT
gram, year, ratio, PREV_YEAR, PREV_RATIO, ratio - PREV_RATIO AS DIFF
FROM
(SELECT
gram, year, ratio,
LAG(year) OVER (PARTITION BY gram ORDER BY year) AS PREV_YEAR,
LAG(ratio) OVER (PARTITION BY gram ORDER BY year) AS PREV_RATIO
FROM google_ngrams_ratio
WHERE ratio > 0
) lag
WHERE PREV_RATIO > 0
ORDER BY DIFF DESC
LIMIT 50;
```

   The results show the statistics in use of the Internet around 1990-1995 when the World Wide Web was invented. (The Internet itself has been around a lot longer than 1990).

   Finally, it is not all of the following words from the most popular words of each length for the year 2000.

1. Copy and paste the statement to view the most-used words of each length for the year 2000:

```
SELECT
gram, LENGTH(gram) AS LENGTH, TOTAL_OCCURRENCES
FROM
(SELECT
gram,
SUM(occurrences) AS TOTAL_OCCURRENCES,
RANK() OVER (PARTITION BY LENGTH(gram) ORDER BY SUM(occurrences) DESC) AS RANK
FROM google_ngrams_norm
WHERE year = '2000'
GROUP BY gram
) rank
WHERE RANK = 1
ORDER BY LENGTH;
```

   Output:

```
OK
gram    LENGTH  TOTAL_OCCURRENCES
the     3       1025000000
and     3       1000000000
that    4       1000000000
which   5       1000000000
there   5       1000000000
would   5       1000000000
their   5       1000000000
Time taken: 0.086 seconds, Fetched: 7
```

```
Amazon Athena Query Editor
SELECT
  year, SUM(Sales) AS Sales,
  Region
FROM
  [AWS_DMS_parquet_db].[dms_parquet_table]
WHERE
  year = 2016
GROUP BY
  year, Region
ORDER BY
  year, Sales DESC
LIMIT 10;
```

These queries provide an idea of the capabilities of Amazon EMR and how to analyze massive quantities of data.

Achievements

🏆 Congratulations! You completed the lab and you are now able to:

- Create an Amazon EMR cluster using the console
- Use the Amazon EMR cluster to process data from Google Nexus Pixel data stored in Amazon S3
- Run Hive queries to drill down and analyze data

End Lab

Follow these steps to close the console, end your lab, and evaluate the experience.

16. Return to the AWS Management Console.
17. On the Navigation bar, choose **awsstudent/AccessRoleOver**, and then choose **Sign out**.
18. Choose **End lab**.
19. Confirm **OK**.
20. (Optional)
    - Select the applicable number of stars (1)
    - Type a comment
    - Choose **Submit**
      - 1 star = Very dissatisfied
      - 2 stars = Dissatisfied
      - 3 stars = Neutral
      - 4 stars = Satisfied
      - 5 stars = Very satisfied

You may close the window if you don't want to provide feedback.

Additional Resources

For more information on the topics covered in this lab, consult the following references:

Amazon Elastic MapReduce:

- http://aws.amazon.com/elasticmapreduce/
- http://docs.aws.amazon.com/ElasticMapReduce/latest/GettingStarted/

Apache Hadoop:

- http://hadoop.apache.org/
- http://wiki.apache.org/hadoop/FrontPage

Apache Hive:

- http://hive.apache.org/
- http://cwiki.apache.org/confluence/pages/viewpreviousversions.action?pageId=27838462

For more information about AWS Training and Certification, see:
http://aws.amazon.com/training/

Your feedback is welcome and appreciated.

If you would like to make any feedback, suggestions, or corrections, please provide the details in our AWS Training and Certification Contact Form.

Continue questing