Como resultado de la práctica has de entregar un archivo en formato *"pdf"* con breves explicaciones y capturas de pantalla de los pasos principales que has dado para realizar los ejercicios.

Realizaremos dos apartados del curso del enlace de abajo en AWS Skillbuilder:

https://skillbuilder.aws/learning-plan/J38YWQY59M/data-analytics-learning-plan-includes-labs-espaol-de-espaa/TDYZZ22A7S

Data Analytics Learning Plan (includes Labs) (Español de España)

CONTENIDO

APARTADO A

INTRODUCCIÓN

Realiza el laboratorio:

Laboratorio de AWS Builder

Analyze Big Data with Hadoop (Español de España)

4.0 (3) | Básico | 1h | Español (España)

En él se analizan datos de Amazon CloudFront, servicio de AWS que genera registros de acceso que muestran todos los datos solicitados por los usuarios, con el siguiente formato (se explica el contenido de cada campo en el laboratorio):

```
**********

**** SAMPLE LOG DATA ****

**********

2017-07-05 20:05:47 SEA4 4261 10.0.0.15 eabcd12345678.cloudfront.net /test-image-2.jpeg Mozilla/5.0%20(MacOS;%20U;%20Windows%20NT%205.1;%20en-US;%20rv:1.9.0.9)%20Gecko/2009040821%20Chrome/3.0.9
```

No necesitas usar AWS Academy, el curso incluye su propio laboratorio.

CONTENIDO

APARTADO B

INTRODUCCIÓN

Realiza el laboratorio:

Big Data

Realiza el apartado:

Laboratorio de AWS Builder

Exploring Google Ngrams with Amazon EMR and Hive (Español de España)

4.5 (2) | Avanzado | 1h 15m | Español (España)

Se trabajan con datos de Google Ngram:

Google Ngram Viewer: Albert Einstein,Sherlock Holmes,Frankenstein

El *bucket* de Amazon S3 `s3://datasets.elasticmapreduce/ngrams/books/20090715/eng-1M/1gram/` contiene un conjunto de datos que es parte de un proyecto de análisis de n-gramas en libros en inglés. Este proyecto se utiliza principalmente para el análisis lingüístico y de texto a gran escala. A continuación, te doy una idea más clara de lo que puedes encontrar en ese *bucket*:

Este *bucket* en particular contiene archivos relacionados con **n-gramas de un solo término (1-gram)** de libros en inglés, recopilados en grandes cantidades. En particular, este conjunto de datos contiene un millón de palabras o términos (de ahí el nombre eng-1M).

Realiza este laboratorio creando el clúster EMR desde AWS ACADEMY (en estes caso no uses el entorno que te propone el curso)

- Una vez creado el clúster EMR en AWS ACADEMY, conéctate a él vía **`ssh`** como ya hicimos en otras prácticas y comienza el laboratorio desde el apartado:

---

11. Escribe `hive` en el terminal y pulsa `enter`. Se te redirigirá a la aplicación de Hive.

Además de las practicas que se proponen en el laboratorio responde a las siguientes preguntas:

1.- ¿Qué contiene el *bucket* s3://datasets.elasticmapreduce/ngrams/books/20090715/eng-1M/1gram/?

¿Cuánto ocupa el archivo que contiene?

2.- ¿Cuántos registros contiene la tabla **`ngrams`** que creaste en HIVE?

¿Desde qué año hasta qué año abarca la información que contiene?

3.- En la creación de la tabla **`normalized`** ¿qué significa la expresión **`REGEXP "^[A-Za-z+\'-]{3,}$"`**?

¿ Cuántos registros contiene la tabla **`normalized`**?