Tarea de Hadoop

$ docker start -ia hadoop-pseudo

➜ ~ sudo -u hdfs hadoop fs -mkdir /experimentos

➜ ~ sudo -u hdfs hadoop fs -ls /

➜ ~ sudo -u hdfs hadoop fs -mkdir /user/jared27

➜ ~ sudo -u hdfs hadoop fs -rm -r /user/jared27

➜ ~ sudo -u hdfs hadoop fs -mkdir /etl

➜ ~ sudo -u hdfs hadoop fs -mkdir /data

➜ ~ sudo -u hdfs hadoop fs -mkdir /app

➜ ~ sudo -u hdfs hadoop fs -mkdir /user/itam

➜ ~ sudo -u hdfs hadoop fs -mkdir /user/itam/datasets

➜ ~ sudo -u hdfs hadoop fs -mkdir /user/itam/experimentos

➜ ~ sudo -u hdfs hadoop fs -mkdir /user/itam/datasets

➜ ~ sudo -u hdfs hadoop fs -mkdir /user/itam/datasets/ufos

➜ ~ sudo -u hdfs hadoop fs -mkdir /user/itam/datasets/gdelt

➜ ~ sudo -u hdfs hadoop fs -copyFromLocal /home/itam/data/UFO-Dic-2014.tsv /user/itam/experimentos

➜ ~ sudo -u hdfs hadoop fs -copyFromLocal /home/itam/data/UFO-Nov-2014.tsv /user/itam/experimentos

➜ ~ sudo -u hdfs hadoop fs -cat /user/itam/experimentos/UFO-Dic-2014.tsv | head

➜ ~ sudo -u hdfs hadoop fs -cat /user/itam/experimentos/UFO-Dic-2014.tsv | wc -l

➜ data sed -i 's@Date / Time@Date\_Time@head' data/UFO-Dic-2014.tsv

sed: -e expresión #1, carácter 25: opción desconocida para `s'

➜ data sed -ei 's@Date / Time@Date\_Time@head' data/UFO-Dic-2014.tsv

sed: -e expresión #1, carácter 1: se esperaba \ después de «a», «c» o «i»

➜ data sed -e -i 's@Date / Time@Date\_Time@head' data/UFO-Dic-2014.tsv

sed: -e expresión #1, carácter 1: orden desconocida: «-»

➜ data sed -e -i 's@Date / Time@Date\_Time@head' /UFO-Dic-2014.tsv

sed: -e expresión #1, carácter 1: orden desconocida: «-»

➜ data sed -e -i 's@Date / Time@Date\_Time@head' UFO-Dic-2014.tsv

sed: -e expresión #1, carácter 1: orden desconocida: «-»

➜ data sed -e -i 's@Date / Time@Date\_Time@g' UFO-Dic-2014.tsv

sed: -e expresión #1, carácter 1: orden desconocida: «-»

➜ data sudo sed -e -i 's@Date / Time@Date\_Time@g' UFO-Dic-2014.tsv

sed: -e expresión #1, carácter 1: orden desconocida: «-»

➜ data sudo sed -i 's@Date / Time@Date\_Time@g' UFO-Dic-2014.tsv

➜ data cat UFO-Dic-2014.tsv | head

Date\_Time City State Shape Duration Summary Posted

12/31/14 23:59 Wailuku HI Circle 20 minutes Exactly one year to the date more red orange floating objects. 1/07/2015

12/31/14 23:59 St. Thomas (U. S. Virgin Islands) Triangle 5 minutes Red triangle seen over us Virgin Islands on New Year's Eve. 1/09/2015

12/31/14 23:30 San Jose del Cabo (Baja California)(Mexico) Triangle 10 minutes Two triangular lights over San Jose del Cabo New Year's eve. I was not drinking. 1/07/2015

12/31/14 23:30 Boynton Beach FL Light 1 minute Orange light moves west, flickers, then takes off full speed out of sight. 1/07/2015

12/31/14 23:30 Smyrna GA Disk 15 minutes Very bright orange/red lights on disc shaped craft heading due south over Atlanta on New Years Eve 1/07/2015

12/31/14 23:15 Onancock VA Light 2-3 seconds Fast bright light disapears with in a second 1/07/2015

12/31/14 22:45 Bear Lake UT Sphere 5 minutes Dark orange light that was stationary in the sky and turned from 1 light to 3 lights in a crescent shape. Stayed in one place for ab 1/07/2015

12/31/14 22:45 Somersworth NH Fireball 10 minutes Two orange fireballs seen moving silently across the night sky in Somersworth, NH. 1/07/2015

12/31/14 22:15 Vancouver WA Light 15 minutes Orange orbs seen in southeast sky over Vancouver WA on New Years Eve at about 10:15pm 1/07/2015

➜ data cd..

zsh: command not found: cd..

➜ data cd ..

➜ ~ kite-dataset csv-schema data/UFO-Dic-2014.tsv --class UFO -o ufos.avsc --delimiter "\t"

OpenJDK 64-Bit Server VM warning: INFO: os::commit\_memory(0x00000000f0b80000, 41418752, 0) failed; error='No se pudo asignar memoria' (errno=12)

#

# There is insufficient memory for the Java Runtime Environment to continue.

# Native memory allocation (malloc) failed to allocate 41418752 bytes for committing reserved memory.

# An error report file with more information is saved as:

# /home/itam/hs\_err\_pid6897.log

➜ ~ kite-dataset csv-schema data/UFO-Dic-2014.tsv --class UFO -o ufos.avsc --delimiter "\t"

OpenJDK 64-Bit Server VM warning: INFO: os::commit\_memory(0x00000000f0b80000, 41418752, 0) failed; error='No se pudo asignar memoria' (errno=12)

#

# There is insufficient memory for the Java Runtime Environment to continue.

# Native memory allocation (malloc) failed to allocate 41418752 bytes for committing reserved memory.

# An error report file with more information is saved as:

# /home/itam/hs\_err\_pid6985.log

➜ ~ kite-dataset csv-schema data/UFO-Dic-2014.tsv --class UFO -o ufos.avsc --delimiter "\t"

OpenJDK 64-Bit Server VM warning: INFO: os::commit\_memory(0x00000000f0b80000, 41418752, 0) failed; error='No se pudo asignar memoria' (errno=12)

#

# There is insufficient memory for the Java Runtime Environment to continue.

# Native memory allocation (malloc) failed to allocate 41418752 bytes for committing reserved memory.

# An error report file with more information is saved as:

# /home/itam/hs\_err\_pid7077.log

➜ ~ kite-dataset csv-schema data/UFO-Dic-2014.tsv --class UFO -o ufos.avsc --delimiter "\t"

OpenJDK 64-Bit Server VM warning: INFO: os::commit\_memory(0x00000000f0b80000, 41418752, 0) failed; error='No se pudo asignar memoria' (errno=12)

#

# There is insufficient memory for the Java Runtime Environment to continue.

# Native memory allocation (malloc) failed to allocate 41418752 bytes for committing reserved memory.

# An error report file with more information is saved as:

# /home/itam/hs\_err\_pid7172.log

➜ ~ kite-dataset csv-schema /data/UFO-Dic-2014.tsv --class UFO -o ufos.avsc --delimiter "\t"

OpenJDK 64-Bit Server VM warning: INFO: os::commit\_memory(0x00000000f0b80000, 41418752, 0) failed; error='No se pudo asignar memoria' (errno=12)

#

# There is insufficient memory for the Java Runtime Environment to continue.

# Native memory allocation (malloc) failed to allocate 41418752 bytes for committing reserved memory.

# An error report file with more information is saved as:

# /home/itam/hs\_err\_pid7260.log

➜ ~

➜ ~ kite-dataset csv-schema data/UFO-Nov-2014.tsv --class UFO -o ufos.avsc --delimiter "\t"

Unknown error: Bad header for field, should start with a character or \_ and can contain only alphanumerics and \_ 0: "Date / Time"

➜ ~ sed -i 's@Date / Time@Date\_Time@g' data/data/UFO-Nov-2014.tsv

➜ ~ kite-dataset csv-schema data/UFO-Nov-2014.tsv --class UFO -o ufos.avsc --delimiter "\t"

➜ ~ kite-dataset create dataset:hdfs:/user/itam/datasets/ufos --schema ufos.avsc

➜ ~ kite-dataset schema dataset:hdfs:/user/itam/datasets/ufos

{

"type" : "record",

"name" : "UFO",

"doc" : "Schema generated by Kite",

"fields" : [ {

"name" : "Date\_Time",

"type" : [ "null", "string" ],

"doc" : "Type inferred from '11/30/14 21:45'",

"default" : null

}, {

"name" : "City",

"type" : [ "null", "string" ],

"doc" : "Type inferred from 'Northumberland'",

"default" : null

}, {

"name" : "State",

"type" : [ "null", "string" ],

"doc" : "Type inferred from 'PA'",

"default" : null

}, {

"name" : "Shape",

"type" : [ "null", "string" ],

"doc" : "Type inferred from 'Oval'",

"default" : null

}, {

"name" : "Duration",

"type" : [ "null", "string" ],

"doc" : "Type inferred from '5 minutes'",

"default" : null

}, {

"name" : "Summary",

"type" : [ "null", "string" ],

"doc" : "Type inferred from 'Orange object that changed directions and colors.'",

"default" : null

}, {

"name" : "Posted",

"type" : [ "null", "string" ],

"doc" : "Type inferred from '12/05/2014'",

"default" : null

} ]

}

➜ ~ kite-dataset csv-import data/UFO-Nov-2014.tsv dataset:hdfs:/user/itam/datasets/ufos --delimiter "\t"

Added 529 records to "dataset:hdfs:/user/itam/datasets/ufos"

➜ ~ kite-dataset csv-import data/UFO-Dic-2014.tsv dataset:hdfs:/user/itam/datasets/ufos --delimiter "\t"

Added 338 records to "dataset:hdfs:/user/itam/datasets/ufos"

➜ ~ kite-dataset show dataset:hdfs:/user/itam/datasets/ufos

### (py)Spark

# ufo

pyspark

import csv

from io import StringIO

ufos\_nov = sc.textFile("hdfs://localhost/user/itam/experimentos/UFO-Nov-2014.tsv")

ufos\_nov.count()

ufos\_nov.take(5)

ufos\_nov.first()

ufos\_nov.map(lambda line: (line.split('\t')[2]))\

.distinct()\

.count()

def load\_tsv(archivo):

return csv.reader(StringIO(archivo[1]), delimiter='\t')

ufos\_nov = sc.textFile("hdfs://localhost/user/itam/experimentos/UFO-Nov-2014.tsv").flatMap(load\_tsv)

ufos\_nov.take(3)[2]

➜ ~ kite-dataset csv-schema data/20150508.export.CSV --class gdelt -o gdelt.avsc --delimiter "\t" --no-header

**➜ ~** kite-dataset create dataset:hdfs:/user/itam/datasets/gdelt --schema gdelt.avsc

**➜ ~** kite-dataset schema dataset:hdfs:/user/itam/datasets/gdelt

{

"type" : "record",

"name" : "gdelt",

"doc" : "Schema generated by Kite",

"fields" : [ {

"name" : "field\_0",

"type" : [ "null", "long" ],

"doc" : "Type inferred from '431551843'",

"default" : null

}, {

"name" : "field\_1",

"type" : [ "null", "long" ],

"doc" : "Type inferred from '20050510'",

"default" : null

}, {

"name" : "field\_2",

"type" : [ "null", "long" ],

"doc" : "Type inferred from '200505'",

"default" : null

}, {