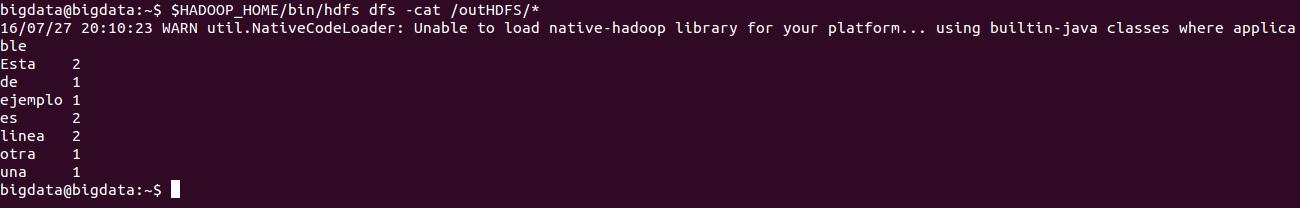
EJERCICIOS DE HADOOP

*1.* Devolver el resultado de ejecutar el siguiente comando

*$HADOOP\_HOME/bin/hdfs dfs -cat /outHDFS/\**



Nos devuelve el contenido de outHDFS

2. ¿Cuál de los siguientes funcionalidades pertenecen a un NameNode de HDFS?

a. Transferir bloques de datos de los datanodes a los clientes

b. Mantener el árbol del sistema de archivos y los metadatos de todos los ficheros y directorios

c. Controlar los procesos de Map Reduce

d. Almacenar bloques de datos

e. Ninguna de las opciones es correcta

El NameNode se encarga de gestionar la información que va a cada DataNode, así como de comprobar su estado y balancear en el caso de que alguno esté caído.

3. ¿Cuál de los siguientes funcionalidades pertenece a un DataNode de HDFS?

a. Mantener el árbol del sistema de archivos y los metadatos de todos los ficheros y directorios.

b. Controlar la ejecución de una tarea de mapeo o de reduce individual.

c. Gestionar el sistema de espacios de nombres de los archivos.

d. Almacenar y recuperar bloques cuando los clientes o el NameNode lo solicita.

e. Ninguna de las opciones es correcta.

El DataNode se encarga de almacenar devolver la información al cliente (o al NameNode si lo requiere)

4. ¿Cuál de las siguientes frases es cierta con respecto a YARN?

a. Implementa un gestor de ficheros para todos los frameworks de Hadoop.

b. Permitir acceso a los datos de HDFS a programas que no estén desarrollados en Hadoop.

c. Permitir a múltiples Namenodes con sus propios namespaces, compartir el pool de Datanodes.

d. Usar el JournalNode para decidir el NameNode activo.

e. Ninguna de las anteriores es correcta.

YARN, es un entorno de gestión de recursos y aplicaciones distribuidas dónde se pueden implementar múltiples aplicaciones de procesamiento de datos totalmente personalizadas y específicas para realizar una tarea en cuestión

5. HDFS está diseñado para

a. Ficheros grandes, acceso continuo a los datos y hardware de grandes prestaciones.

b. Ficheros pequeños, acceso continuo a los datos y hardware básico.

c. Ficheros grandes, baja latencia de acceso y hardware básico.

d. Ficheros grandes, acceso continuo a los datos y hardware básico.

e. Ninguna de las anteriores es correcta.

HDFS está preparado para manejar archivos grandes (64 Mb) y está optimizado para una escritura frente a varias lecturas. Realiza mejor su trabajo si las consultas son sobre ficheros grandes que sobre pequeños.

6. ¿Qué es común a Pig y Hive?

a. Permiten múltiples y aleatorias escrituras y lecturas.

b. Traducen lenguajes de alto nivel a trabajos de Map Reduce.

c. Todas operan con estructura de datos JSON.

d. Todas son lenguajes de flujos de datos.

e. Ninguna de las anteriores es correcta.

7. ¿Qué es Flume?

a. Un sistema de archivos distribuido.

b. Una plataforma de ejecución de tareas de MapReduce.

c. Un lenguaje de programación que traduce queries de alto nivel en tareas de map reduce.

d. Un servicio para mover grandes cantidades de datos en un cluster una vez los datos se han generado.

e. Ninguna de las anteriores.

**Apache Flume** es un servicio distribuido, fiable, y altamente disponible para recopilar, agregar, y mover eficientemente grandes cantidades de datos. Tiene una arquitectura sencilla y flexible basada en flujos de datos en streaming. Es robusto y tolerante a fallos, con mecanismos de fiabilidad configurables y muchos mecanismos de conmutación por error (failover) y recuperación. Utiliza un modelo de datos sencillo y extensible que permite la creación de aplicaciones analíticas en línea. (Wikipedia)

8. ¿En cuál de los siguientes escenarios usarías Hadoop?

a. Analizar los signos vitales de un bebé en tiempo real.

b. Obtener las tendencias de acciones bursátiles cada minuto.

c. Procesar un sensor meteorológico para predecir la trayectoria de un huracán.

d. Procesar billones de mensajes de email para ejecutar análisis de texto.

e. Ninguna de las anteriores.

Las opciones a y b están descartadas porque Hadoop no está pensado para obtener resultados en tiempo real. La opción c dependería de la cantidad de los datos. Si es un único sensor y no da una cantidad elevada de datos o no hace falta distribuir la información en distintos nodos, no tendría sentido utilizar Hadoop

9. ¿Cuál de las siguientes frases es cierta?

a. Hadoop es una nueva tecnología diseñada para reemplazar las bases de datos relacionales.

b. Hadoop incluye componentes open source y closed source.

c. Hadoop se puede usar para bigdata, DSS y OLTP.

d. Todas las anteriores son correctas.

e. Ninguna de las anteriores es correcta.

Hadoop está pensado principalmente para big data. Tano DSS (Sistema de soporte a decisiones) como OLTP (Procesamiento de transacciones en linea) también puede ser utilizado con Hadoop.

EJERCICIOS DE HDFS

A. WORD COUNT

Partiendo de las letras de las siguientes canciones

NINE INCH NAILS LYRICS "March Of The Pigs"

step right up march push

crawl right up on your knees

please greed feed (no time to hesitate)

I want a little bit I want a piece of it I think he's losing it

I want to watch it come down

don't like the look of it don't like the taste of it don't like the smell of it

I want to watch it come down

all the pigs are all lined up

I give you all that you want

take the skin and peel it back

now doesn't that make you feel better?

shove it up inside surprise! lies

stains like the blood on your teeth

bite chew suck away the tender parts

I want to break it up I want to smash it up I want to fuck it up

I want to watch it come down

maybe afraid of it let's discredit it let's pick away at it

I want to watch it come down

now doesn't that make you feel better?

the pigs have won tonight

now they can all sleep soundly

and everything is all right

PINK FLOYD "Pigs (Three Different Ones)"

Big man, pig man, ha ha, charade you are

You well heeled big wheel, ha ha, charade you are

And when your hand is on your heart

You're nearly a good laugh

Almost a joker With your head down in the pig bin

Saying "keep on digging" Pig stain on your fat chin

What do you hope to find?

When you're down in the pig mine

You're nearly a laugh You're nearly a laugh

But you're really a cry.

Bus stop rat bag, ha ha, charade you are

You fucked up old hag, ha ha, charade you are

You radiate cold shafts of broken glass

You're nearly a good laugh Almost worth a quick grin

You like the feel of steel You're hot stuff with a hat pin

And good fun with a hand gun

You're nearly a laugh You're nearly a laugh

But you're really a cry.

Hey you Whitehouse, ha ha, charade you are

You house proud town mouse, ha ha, charade you are

You're trying to keep our feelings off the street

You're nearly a real treat All tight lips and cold feet

And do you feel abused? .....!.....!.....!.....!

You gotta stem the evil tide

And keep it all on the inside

Mary you're nearly a treat Mary you're nearly a treat

But you're really a cry.

DAVE MATTHEWS BAND "Pig"

Isn't it strange

How we move our lives for another day

Like skipping a beat

What if a great wave should wash us all away

Just thinking out loud

Don't mean to dwell on this dying thing

But looking at blood

It's alive right now

Deep and sweet within

Pouring through our veins

Intoxicate moving wine to tears

Drinking it deep

Then an evening spent dancing

It's you and me

This love will open our world

From the dark side we can see a glow of something bright

There's much more than we see here

Don't burn the day away

Is this not enough

This blessed sip of life

Is it not enough

Staring down at the ground

Oh then complain and pray more from above

Greedy little pig

Stop just watch your world trickle away

Oh it's your problem now

It'll all be dead and gone in a few short years

Just love will open our eyes

Just love will put the hope in our minds

Much more than we could ever know

Don't burn the day away

Come sister my brother

Shake up your bones shake up your feet

I'm saying open up

And let the rain come pouring in

Wash out this tired notion

That the best is yet to come

But while you're dancing on the ground

Don't think of when you're gone

Love love what more is there

We need the light of love in here

Don't beat your head

Dry your eyes

Let the love in there

There are bad times

But that's ok

Just look for love in it

Don't burn the day away

Look

Here are we

On this starry night staring into space

And I must say

I feel as small as dust

Lying down here

What point could there be troubling

Head down wondering what will become of me

Why concern we cannot see

But no reason to abandon it

Time is short but that's all right

Maybe I'll go in the middle of the night

Take your hands from your eyes, my love

Everything must end some time

Don't burn the day away

Come sister my brother

Shake up your bones shake up your feet

I'm saying open up

And let the rain come flooding in

Wash out this tired notion

That the best is yet to come

But while you're dancing on the ground

Don't think of when you're gone

Love love what more is there

We need the light of love in here

Don't beat your head

Dry your eyes

Let the love in there

There are bad times

But that's ok

Just look for love in it

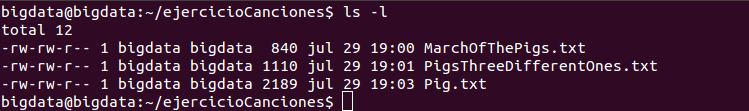
1. Crear ficheros de texto en la máquina ubuntu.

Creamos la carpeta en local donde vamos a guardar los ficheros de las letras:

$ mkdir home/bigdata/ejercicioCanciones

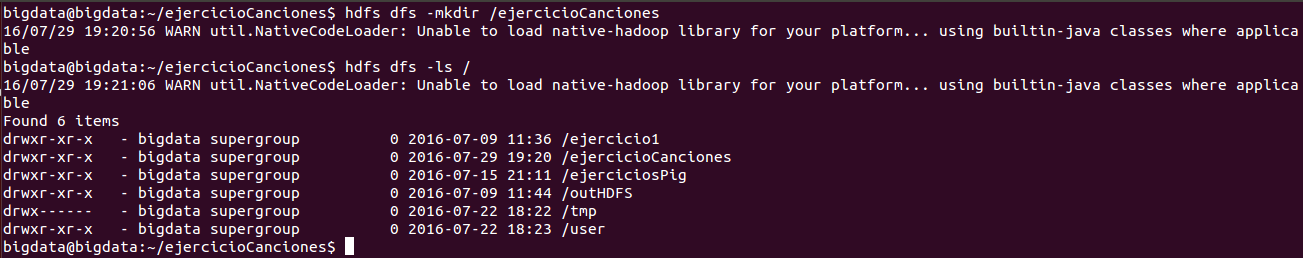


Creamos los ficheros con las canciones. Un fichero por canción:



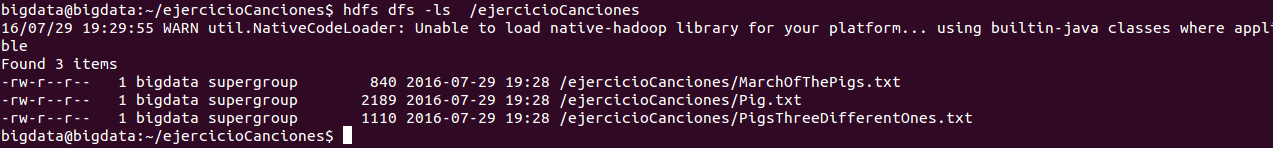
2. Subir los ficheros a HDFS

Creamos una carpeta en hdfs para subir los ficheros que hemos creado y poder tratarlos:



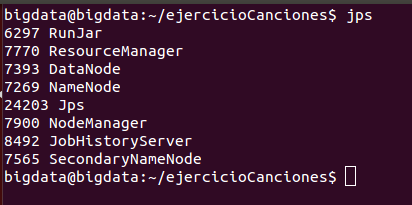
Subo los ficheros locales a la carpeta creada en hdfs



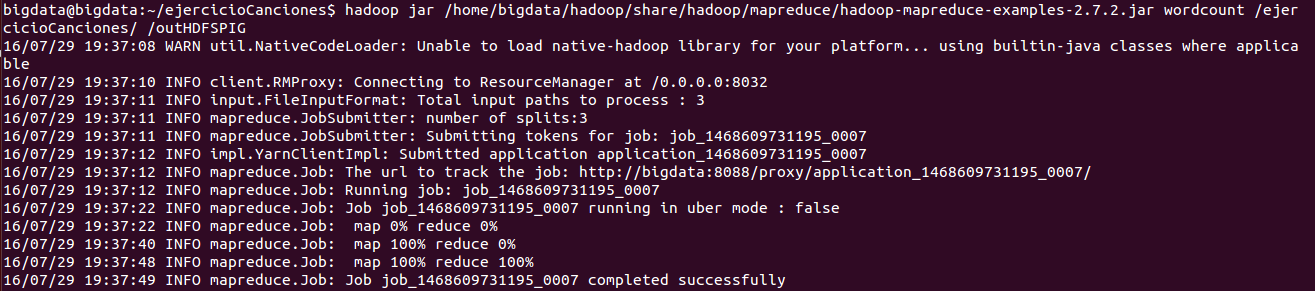


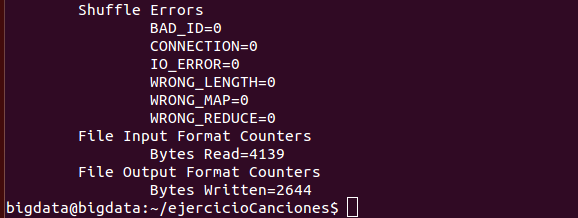
3. Ejecutar un word count y resolver las preguntas del formulario.

Comprobamos que los demonios estén levantados:

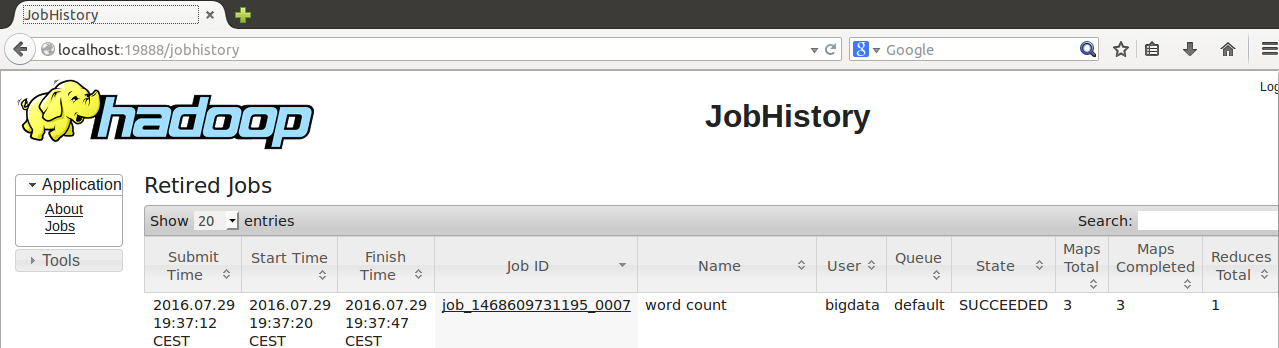


Ejecutamos el Jar “wordCount”





1. Acceder al jobhistory e indicar los valores de "Status" y "Map Total" del job con el nombre "word count"



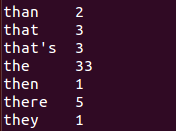
Vemos que el status es “SUCCEEDED” y Maps Total tiene un valor de 3

1. Palabra que se repite más veces

Vemos la salida del programa

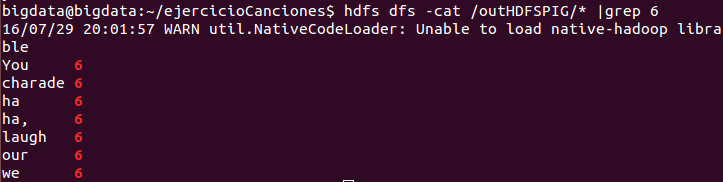


Nos devuelve el listado de todas las palabras junto con las veces que se repite cada una de ellas, ordenadas alfabéticamente. Si buscamos la que tiene un mayor número de repeticiones, nos encontramos que “the” se repite 33 veces



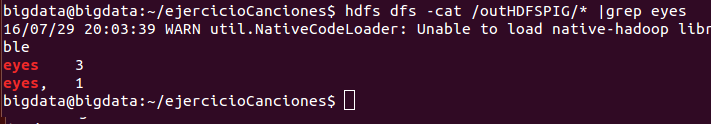
1. Palabra que se repite 6 veces

Unimos la orden anterior a un grep y sacamos las palabras que se repiten 6 veces:



1. Cuántas veces aparece la palabra eyes

Con el sistema anterior:



B. SUDOKU

4. Ejecutar el programa sudoku con los siguientes parámetros:

- Programa: sudoku

- Path fichero

En la ruta del fichero habrá que indicar la ruta en la que se encuentre un fichero que tendrá la siguiente información:

? 4 6 3 ? ? 8 2 5

? ? ? ? ? 2 4 ? ?

8 2 ? ? 6 ? ? 7 ?

7 ? ? 4 ? ? ? ? 2

3 8 ? ? ? ? 6 ? 9

? 6 ? 2 8 ? ? ? ?

? 7 ? ? ? 5 ? 3 ?

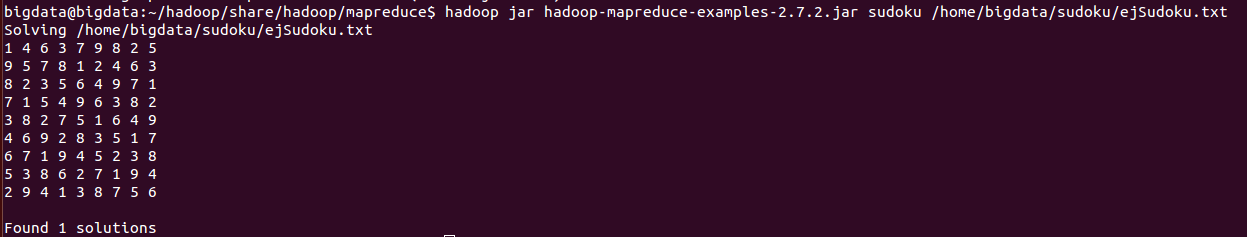
5 3 8 ? ? 7 ? ? ?

? 9 ? ? ? ? ? ? 6

Creamos un fichero, ejSudoku.txt, con los datos en la carpeta local “sudoku”. Hay que tener cuidado de que no haya ninguna línea vacía al final del documento, pues el programa daría un error.

Indicar todos los resultados del sudoku de ejemplo

Ejecuto el jar, en la carpeta mapreduce, con los parámetros que se piden:



El sudoku solo tiene una solución

C. PI

5. Ejecutar el programa "pi" con los siguientes parámetros de entrada e indicar el valor devuelto.

- Programa: pi

- Nº de mapeos: 2

- Nº de muestras por mapeo: 4





6. Ejecutar el programa "pi" con los siguientes parámetros de entrada e indicar el valor devuelto.

- Programa: pi

- Nº de mapeos: 5

- Nº de muestras por mapeo: 10





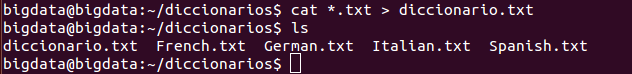
EJERCICIOS DE MAP REDUCE

1. Escribe las traducciones de la palabra PIG.

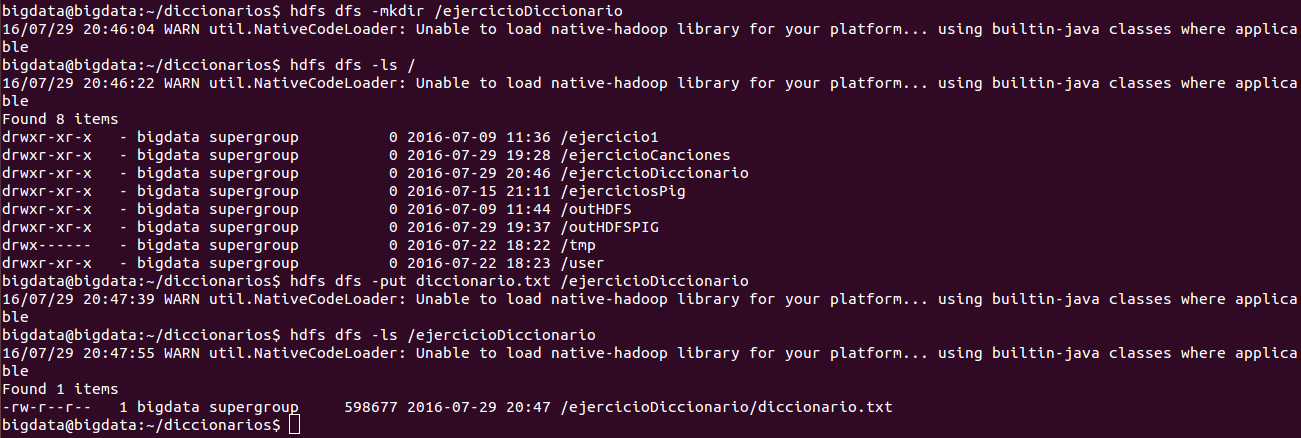
Creamos un directorio para guardar los diccionarios y descargamos en él los ficheros



Hacemos un merge de los 4 ficheros para aprovechar que hadoop trabaja con ficheros grandes:

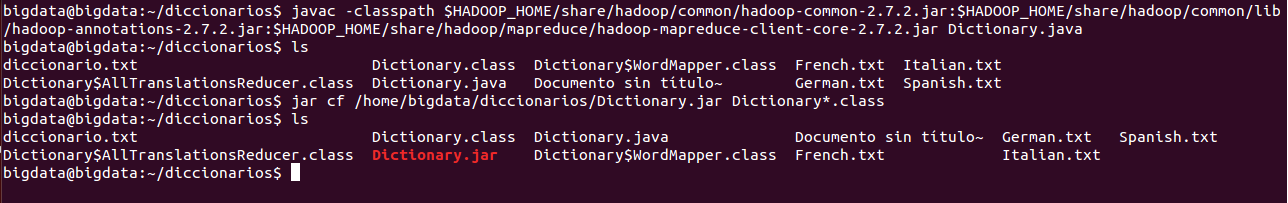


Subimos el fichero a HDFS, pero primero tenemos que crear la carpeta remota:



Ahora creamos el fichero .java. Utilizamos el código que necesita las rutas de entrada y salida para ejecutarse.

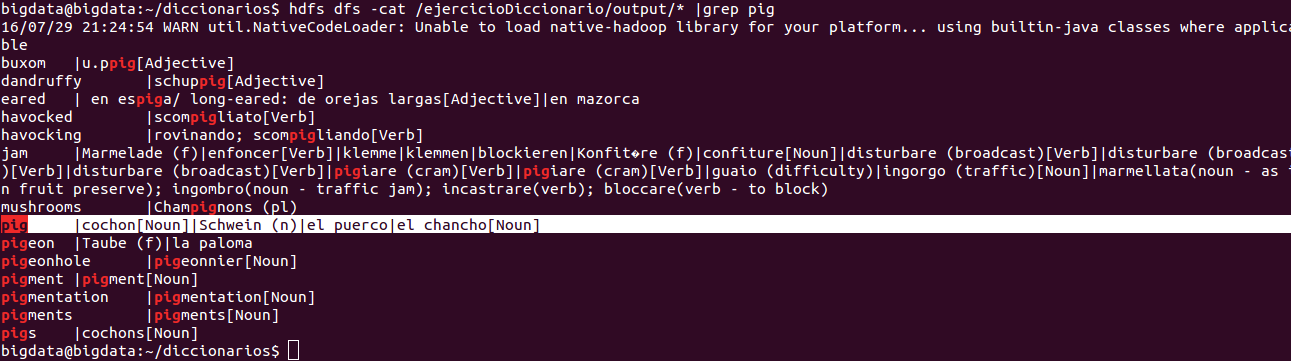
Generamos el .jar correspondiente:



Ya tenemos el archivo .jar con el código del diccionario. Ahora lo ejecutamos pasándole las rutas como parámetro:



Por último miramos el resultado de la ejecución y pedimos que nos devuelva todas las palabras con “pig”



La traducción que nos interesa es la que se marca en la consola.