Máster en Big Data

Tecnologías de Almacenamiento

6. Hands-On: Desarrollo MapReduce

Àlex Balló Vergés
2018-2019

Índice

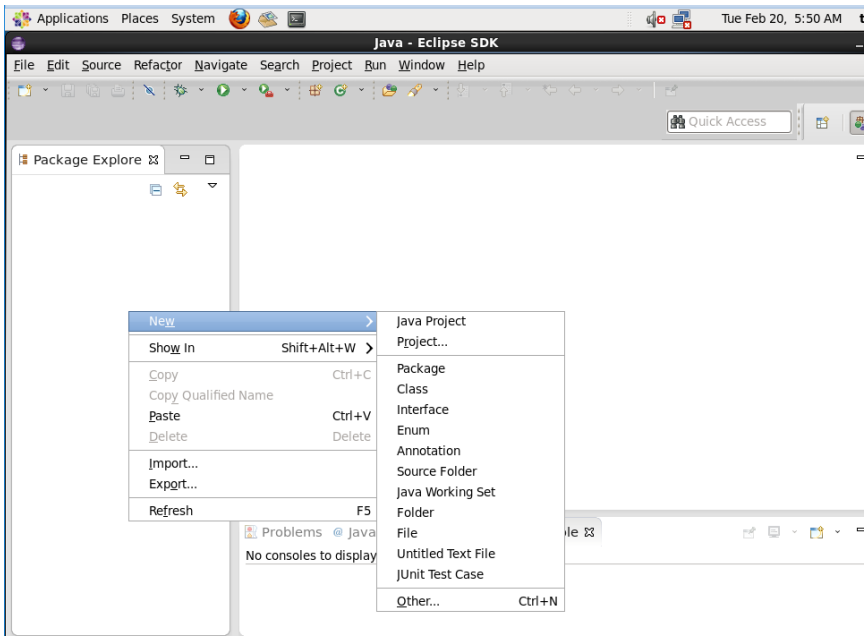
1. Introducción

El objetivo de este Hands-On es poner en práctica conceptos avanzados en el desarrollo de Jobs de MapReduce

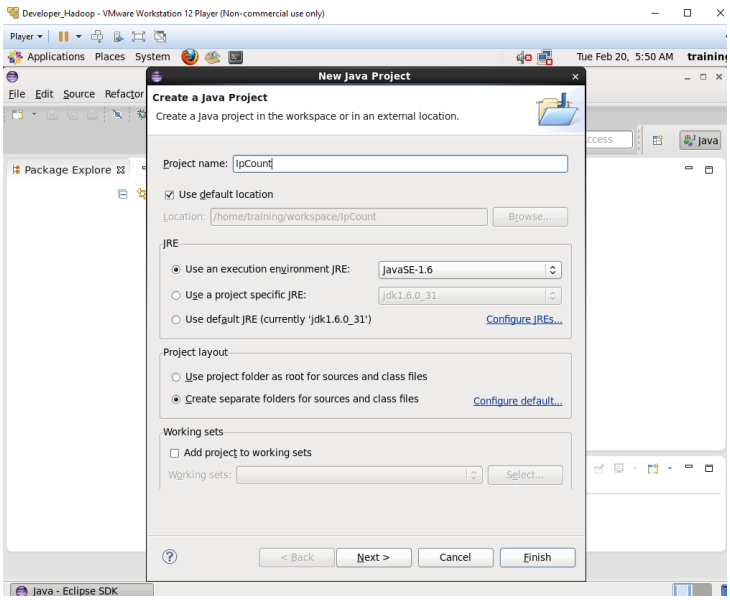
2. Entorno de desarrollo

Para realizar el desarrollo lo haremos mediante el IDE Eclipse de la máquina virtual importada en ejercicios anteriores.

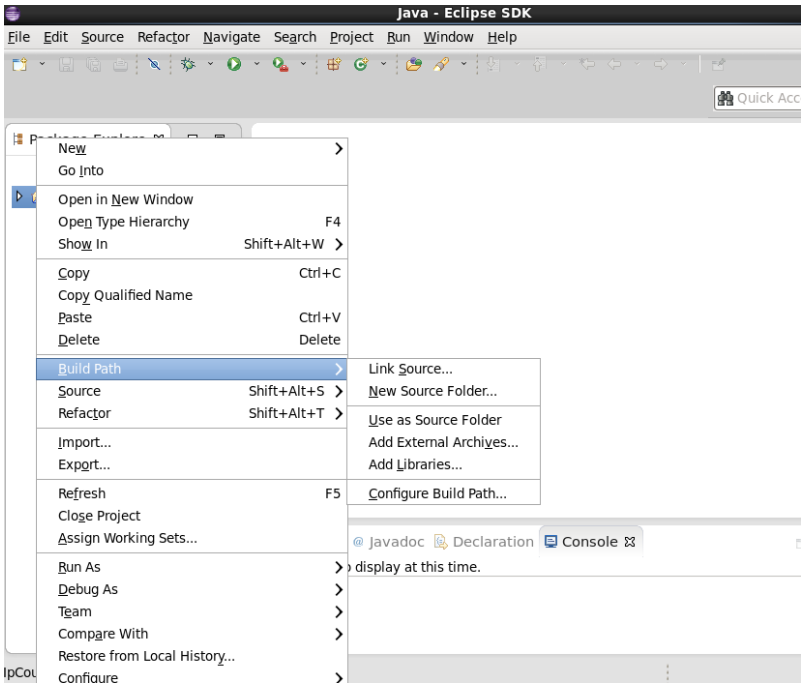
Para crear un nuevo proyecto, haremos click derecho sobre el package explorer New → Java Project



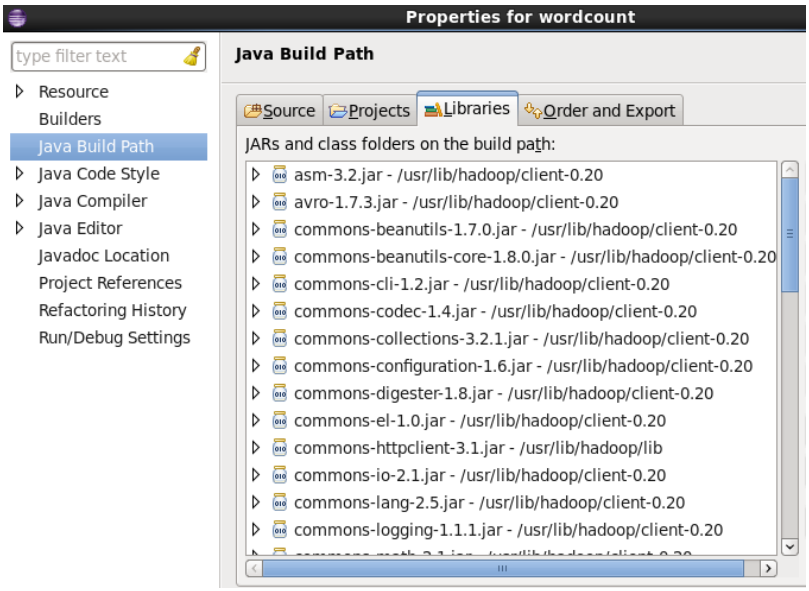
Introducimos el nombre del proyecto y click en Finish



Importamos manualmente las lilbrerías necesarias haciendo click derecho sobre el proyecto que acabamos de crear y seleccionamos Build Path → Configure Build Path



En la pestaña de libraries, seleccionamos Add Extertnal Jars e importamos todo el contenido de la carpeta /usr/lib/hadoop/client-0.20/



3. Tool Runner y parámetros

Desarrollar y ejecutar el siguiente MapReduce:

Aprovechando el ejercicio del Hands-On anterior (**AvarageWordLength)** realizar las siguientes modificaciones:

- La clase driver use ToolRunner

- Modificar el Mapper para referenciar una variable booleana llamada caseSensitive. Si esta variable es true, el mapper no diferenciara entre mayúsculas ni minúsculas, si es false, hará una conversión de todas las letras a minúscula.

4. Combiner

Desarrollar y ejecutar el siguiente MapReduce:

Añadir un combiner al proyecto **IpCount** realizado en el Hands-On anterior

```
job.setMapperClass(IpMapper.class);
job.setReducerClass(IpReducer.class);
job.setCombinerClass(IpReducer.class);
```

```
[training@localhost Jars]$ hadoop fs -cat /user/training/combinerIpCount/part-r-00000 | head -n 10
10.1.1.113      1
10.1.1.125      12
10.1.1.144      1
10.1.1.195      4
10.1.1.236      12
10.1.1.5        1
10.1.10.155     2
10.1.10.197     2
10.1.10.198     1
10.1.10.48      1
```

5. Partitioner

Desarrollar y ejecutar el siguiente MapReduce:

Aprovechando el proyecto original **IpCount** realizar los cambios pertinentes para escribir un Job con múltiples reducers e implementar un partitioner que redirija la salida según el mes del año hacia un reducer concreto.

Es decir, en total habrán 12 reducers (uno para cada mes del año) y el partitioner será el encargado de redirigir esa clave/valor hacia el reducer correcto.

La salida final consistirá en 12 ficheros, uno para cada mes del año, y contendrán el número de veces que se ha repetido la ip en ese mes del año.

Solución:

Input: `96.7.4.14 - - [24/Apr/2011:04:20:11 -0400] "GET /cat.jpg HTTP/1.1" 200 12433`

Output key: `96.7.4.14`

Output value: `Apr`

```
[training@localhost Jars]$ hadoop fs -ls /user/training/ipCountPartitioner
Found 14 items
-rw-r--r--   1 training supergroup          0 2019-04-08 14:55 /user/training/ipCountPartitioner/_SUCCESS
drwxr-xr-x   - training supergroup          0 2019-04-08 14:51 /user/training/ipCountPartitioner/_logs
-rw-r--r--   1 training supergroup     151256 2019-04-08 14:54 /user/training/ipCountPartitioner/part-r-00000
-rw-r--r--   1 training supergroup     452255 2019-04-08 14:54 /user/training/ipCountPartitioner/part-r-00001
-rw-r--r--   1 training supergroup    1273871 2019-04-08 14:54 /user/training/ipCountPartitioner/part-r-00002
-rw-r--r--   1 training supergroup     318451 2019-04-08 14:54 /user/training/ipCountPartitioner/part-r-00003
-rw-r--r--   1 training supergroup     423827 2019-04-08 14:54 /user/training/ipCountPartitioner/part-r-00004
-rw-r--r--   1 training supergroup     415929 2019-04-08 14:54 /user/training/ipCountPartitioner/part-r-00005
-rw-r--r--   1 training supergroup     487404 2019-04-08 14:55 /user/training/ipCountPartitioner/part-r-00006
-rw-r--r--   1 training supergroup     732348 2019-04-08 14:55 /user/training/ipCountPartitioner/part-r-00007
-rw-r--r--   1 training supergroup     356677 2019-04-08 14:55 /user/training/ipCountPartitioner/part-r-00008
-rw-r--r--   1 training supergroup     493539 2019-04-08 14:55 /user/training/ipCountPartitioner/part-r-00009
-rw-r--r--   1 training supergroup     509767 2019-04-08 14:55 /user/training/ipCountPartitioner/part-r-00010
-rw-r--r--   1 training supergroup     278156 2019-04-08 14:55 /user/training/ipCountPartitioner/part-r-00011
```

```
[training@localhost Jars]$ hadoop fs -cat /user/training/ipCountPartitioner/part-r-00001 | head -n 10
10.1.101.31     1
10.1.102.70     1
10.1.107.165    3
10.1.109.100    1
10.1.112.26     1
10.1.114.150    18
10.1.116.117    1
10.1.120.96     1
10.1.123.18     1
10.1.123.193    17
```