|  |
| --- |
| Máster en Big Data |
| Tecnologías de Almacenamiento |
| 1. Hands-On: Desarrollo MapReduce Avanzado |

Índice

[1. Introducción 3](#_Toc507233944)

[2. Entorno de desarrollo 3](#_Toc507233945)

[3. Tool Runner y parámetros 5](#_Toc507233946)

[4. Combiner 5](#_Toc507233947)

[5. Partitioner 6](#_Toc507233948)

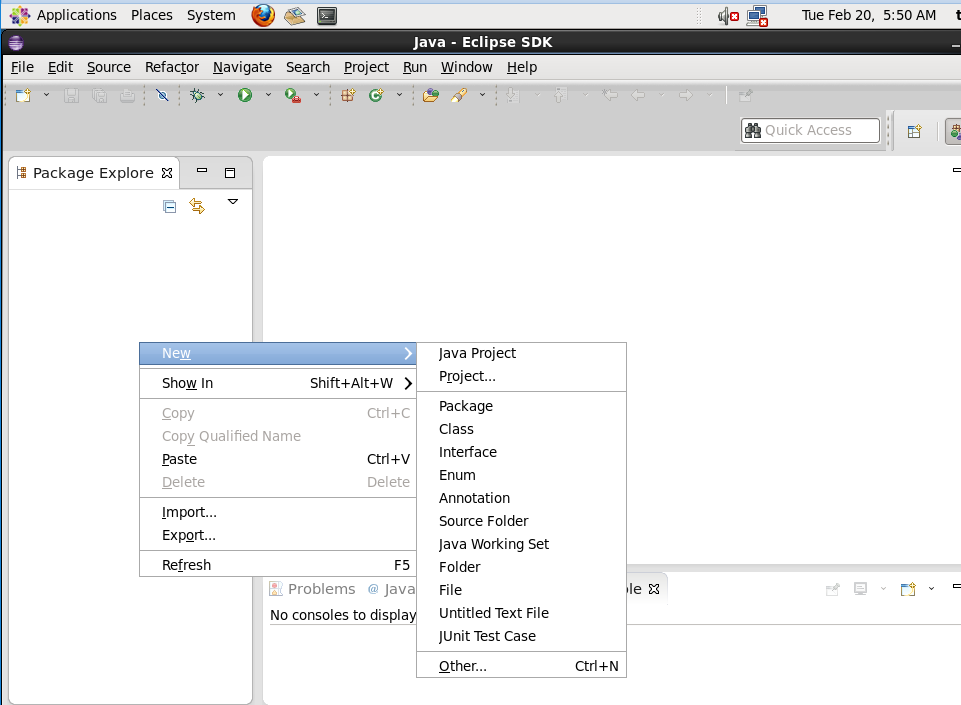
# Introducción

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

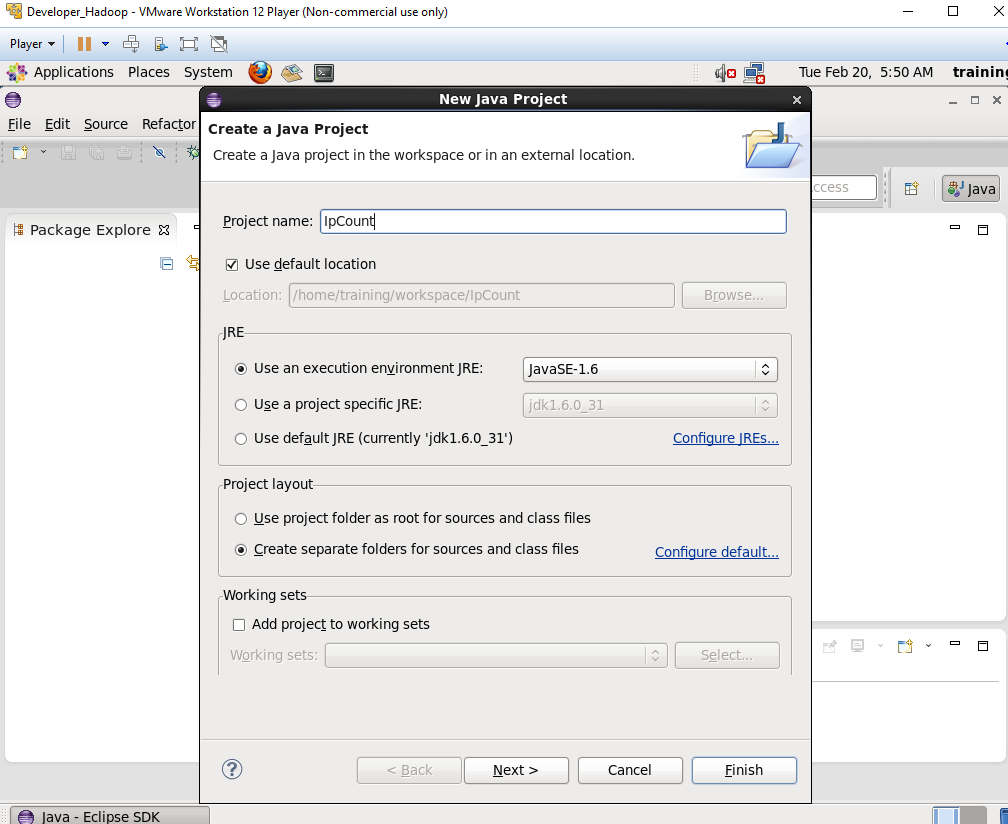
# Entorno de desarrollo

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

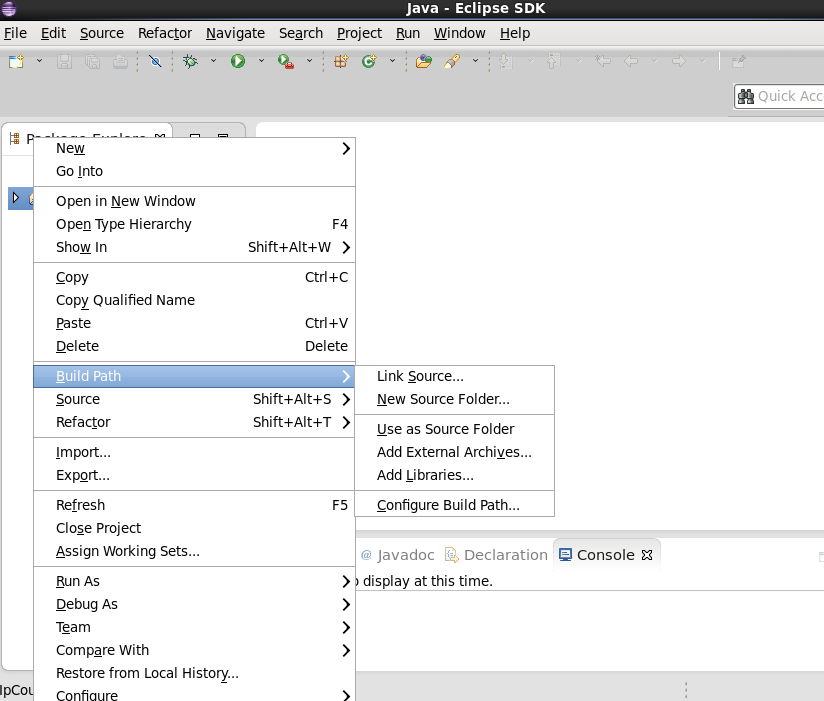
Para crear un nuevo proyecto, haremos click derecho sobre el package explorer New à Java Project



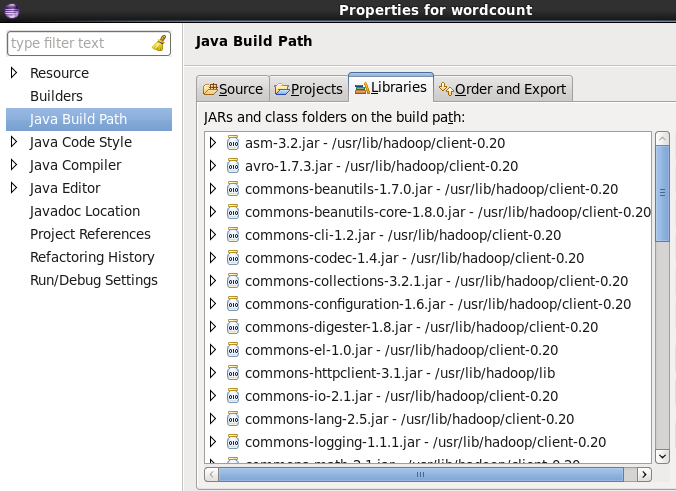
Introducimos el nombre del proyecto y click en Finish



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



# 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.

# Combiner

Desarrollar y ejecutar el siguiente MapReduce:

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

# 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:

