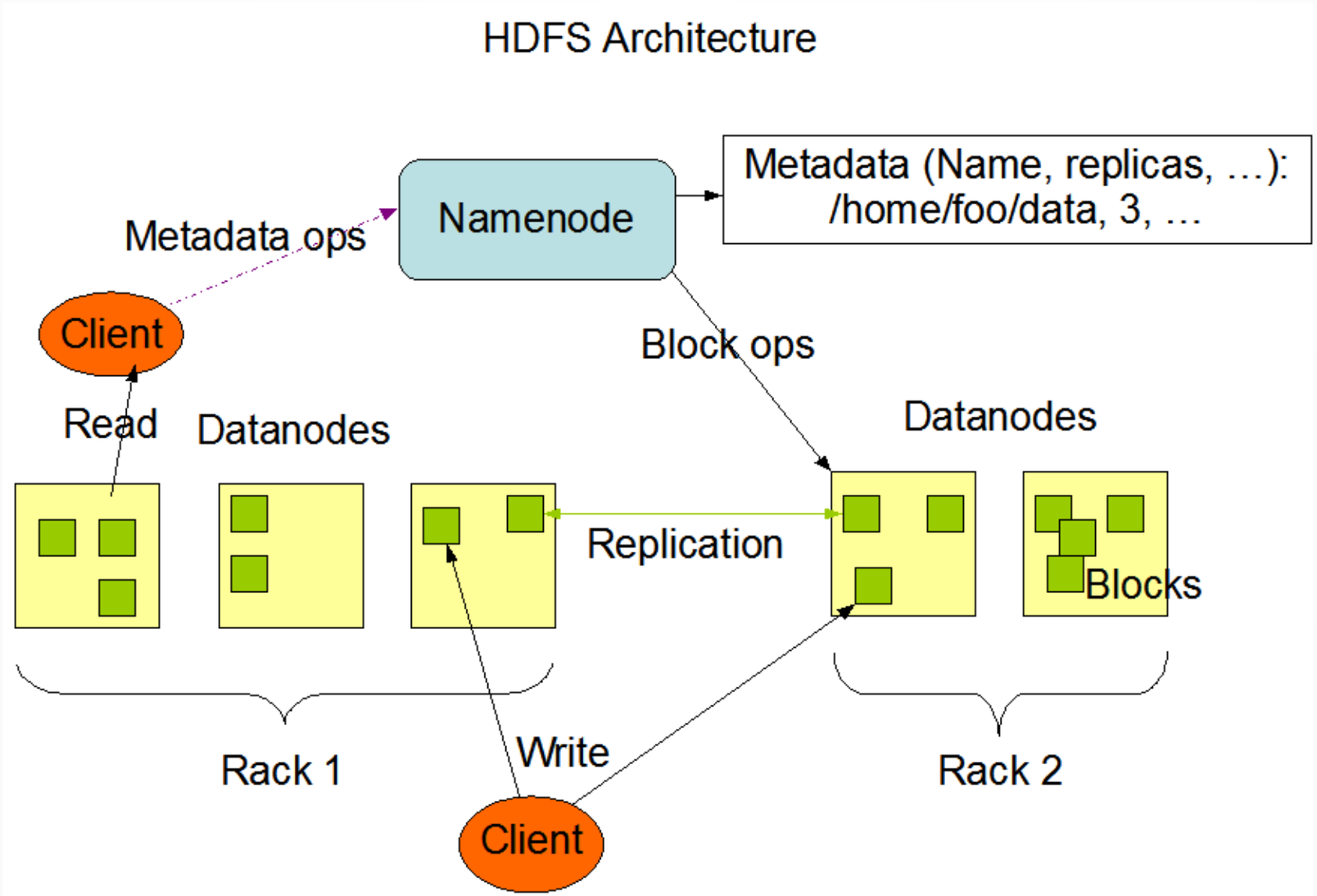
CURSO DE PROGRAMACION SCALA
Sesión 14

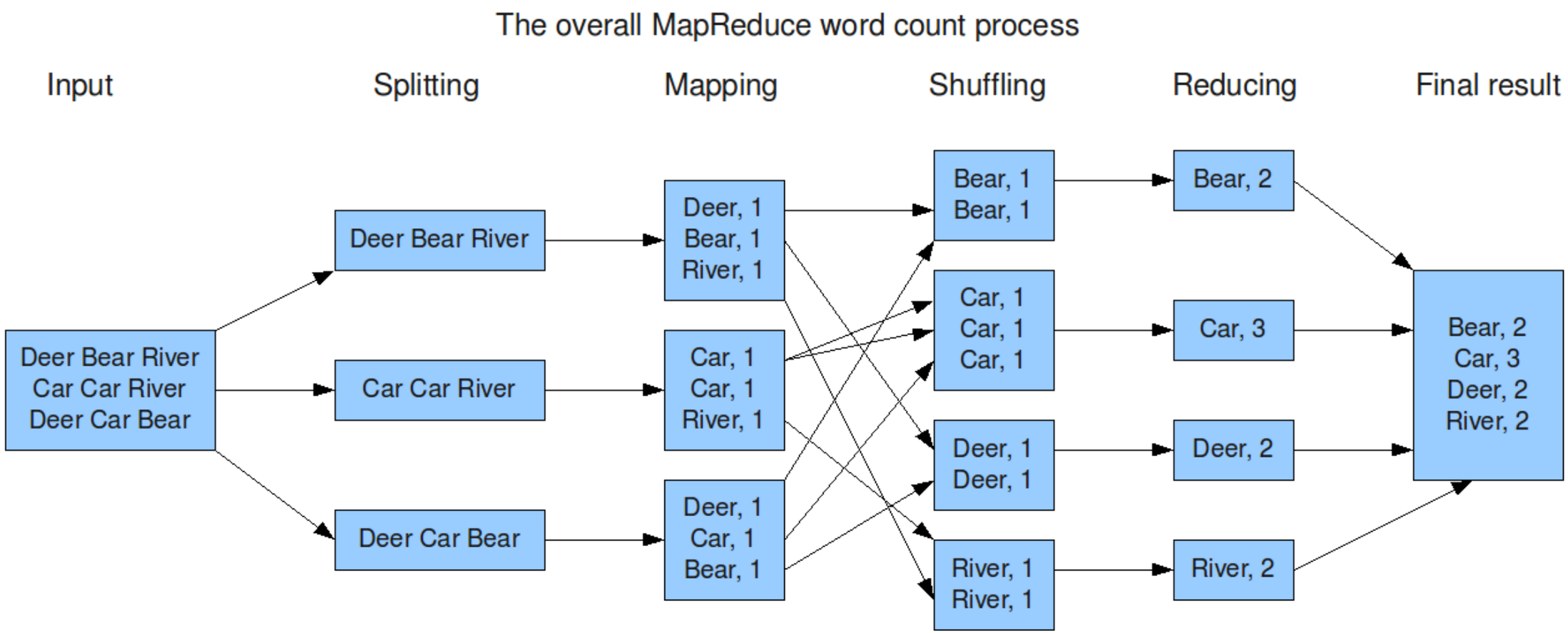
Sergio Couto Catoira

Índice

- Hadoop
- HDFS
- Spark (http://spark.apache.org/downloads.html)
  - Arquitectura
  - APIs
  - Otros Conceptos
- Ejercicio sobre RDD
- Ejercicio sobre Dataframe
  - UDF

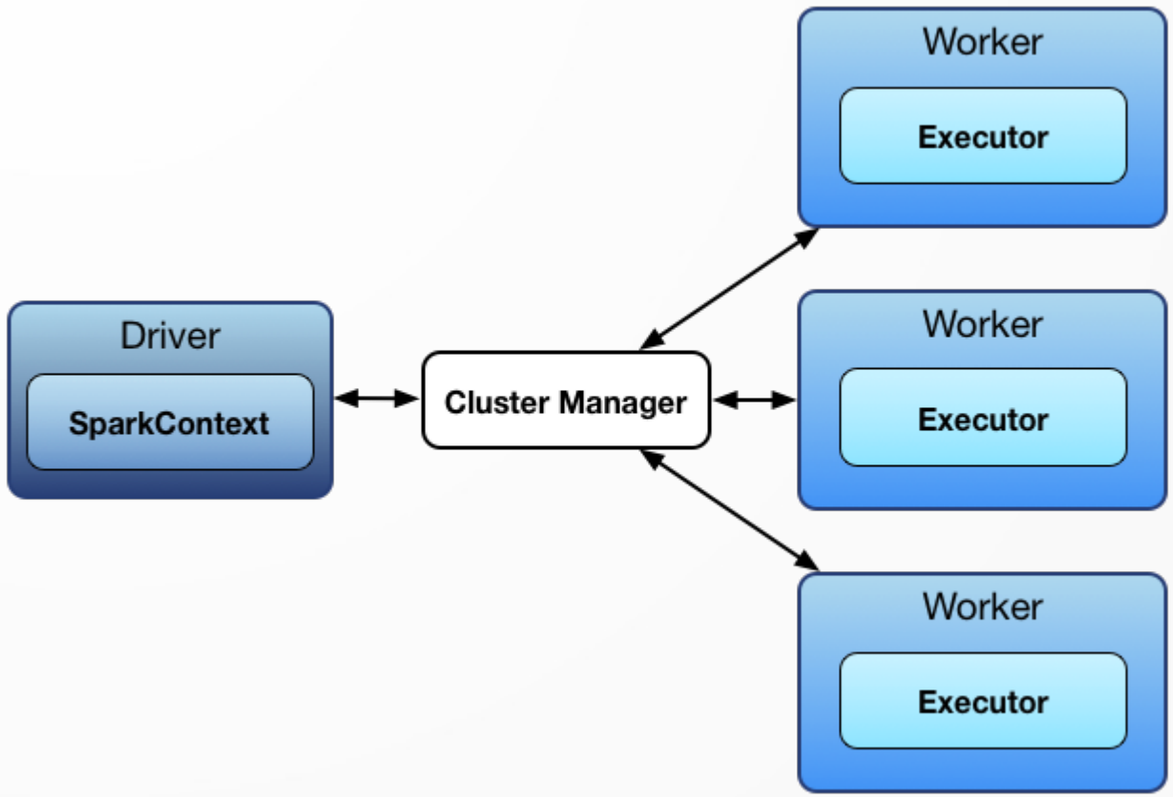

HDFS Architecture
Namenode
Metadata (Name, replicas, …):
/home/foo/data, 3, …
Metadata ops
Client
Block ops
Read
Datanodes
Datanodes
Replication
Blocks
Rack 1
Write
Rack 2
Client

MapReduce


The overall MapReduce word count process
Input
Splitting
Mapping
Shuffling
Reducing
Final result
Deer Bear River
Car Car River
Deer Car Bear
Deer Bear River
Car Car River
Deer Car Bear
Deer, 1
Bear, 1
River, 1
Car, 1
Car, 1
River, 1
Deer, 1
Car, 1
Bear, 1
Bear, 1
Bear, 1
Car, 1
Car, 1
Car, 1
Deer, 1
Deer, 1
River, 1
River, 1
Bear, 2
Car, 3
Deer, 2
River, 2
Bear, 2
Car, 3
Deer, 2
River, 2

Spark - Arquitectura

- Driver: JVM propia
- Worker: Instancias
  - Cada uno al menos un ejecutor
- Executor: JVM Propia (a partir de Spark 2)
  - Ejecuta las tareas con los datos de su nodo
- Task: Cada executor puede



Spark - API

- RDD (Resilient distributed dataset)
  - Sin esquema
- DataFrame
  - Con esquema (similar a una tabla)
  - Datos organizados en columnas con nombre
- Dataset (a partir de Spark 2)
  - Unifica API con Dataframe

Spark - Conceptos

- Laziness
  - Transformaciones
  - Acciones
- Shuffle and Sort
- Spill a disco
- Spark-shell

RDD - Ejercicios

- WordCount - Genera un programa que cuenta las palabras del fichero y las genere en un fichero de salida ordenado (src/main/resources/Shakespeare.txt)
- Los pasos a ejecutar:
  - Leer fichero (sc.TextFile)
  - Splitearlo en palabras
  - Eliminar símbolos y pasar todo a mayusculas/minusculas (.replaceAll("[,.!?:;\"')(-]", ""))
  - Filtrar espacios
  - Contar palabras
  - Ordenar
  - Escribir fichero de salida

RDD - Ejercicios

- Alturas - Genera un programa que lea el fichero alturas.csv y calcula la media de altura por sexo
- Debes filtrar datos erróneos (vacíos o negativos) y corregir los que vengan mal (en metros en lugar de en centímetros)
- Métodos a tener en cuenta
  - groupByKey => Luego lo sustituiremos por aggregate
  - Map
  - Filter
  - toDouble

Dataframe - Ejercicios

- Alturas – Mismo programa que el anterior pero con dataframes
- Métodos a tener en cuenta
  - Leer df => ss.read.csv("src/main/resources/alturas.csv")
  - ToDF => Permite pasar un esquema
  - WithColumn => Crea una columna
  - myDataFrame(“columna”) => Select de una columna