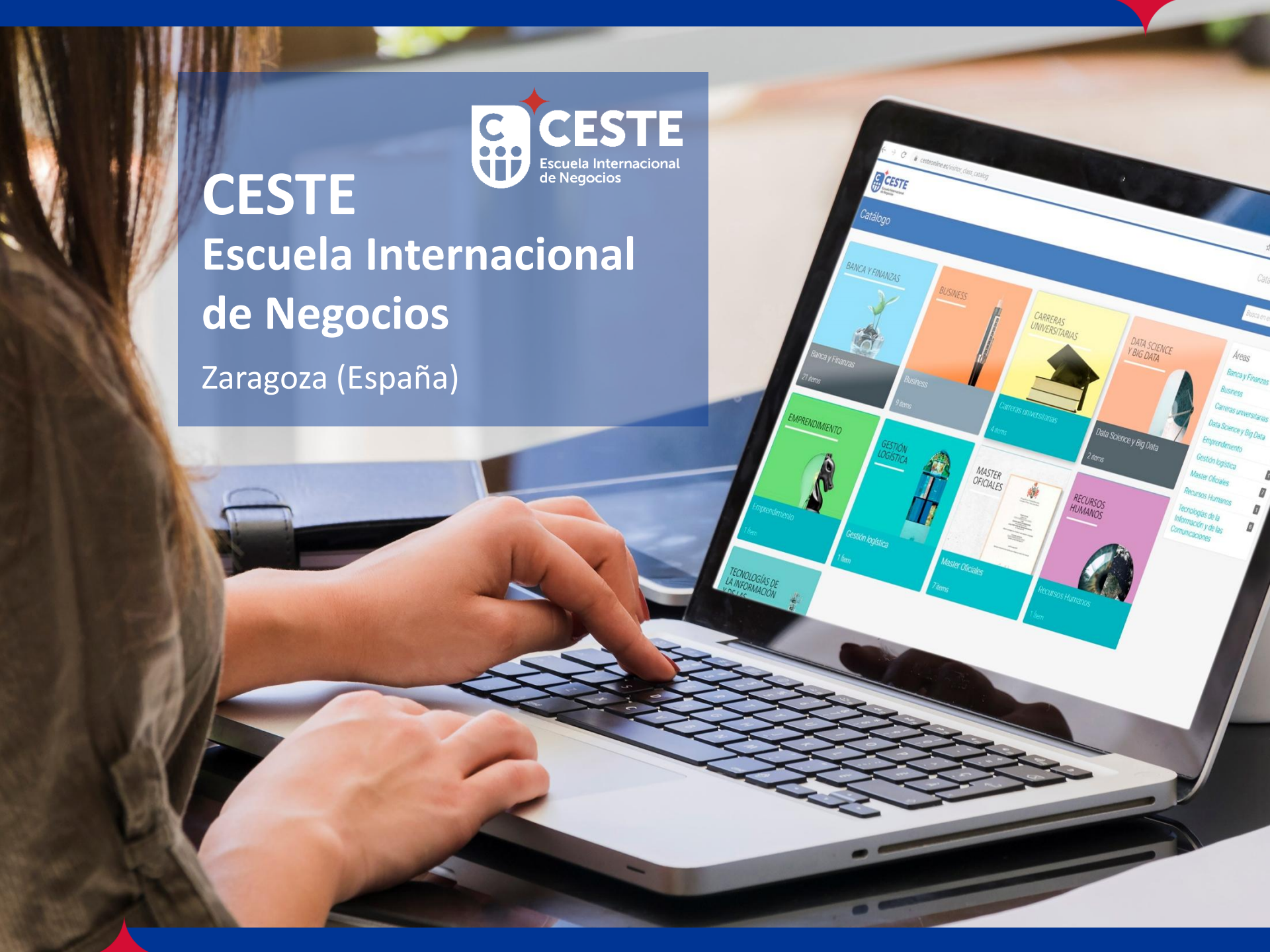

CESTE

Escuela Internacional de Negocios

Zaragoza (España)

Concepto	Qué es	Cuándo se crea	Nivel
Job	Acción completa *count, save, collect*	Se ejecuta una acción	Puede contener varios stages
Stage	Operaciones que pueden ejecutarse sin mover datos	Cada operación que requiere shuffle *groupby, join, repartition*	Contiene varios task
Task	Unidad indivisible de ejecución	Por partición de los datos	Se ejecutan en pararlelos los executors

Ejemplo 1 | Un solo Job, un solo Stage

```
df = spark.range(0, 10)
df.count()
```

- **Jobs:** 1 -> count es acción
- **Stages:** 1 -> No hay suffle (solo lectura)
- **Tasks:** El número de particiones del DF (el que haya por defecto)

Ejemplo 2 | Un solo Job, varios Stages

```
df = spark.range(0, 10_000).repartition(4)
df.groupBy().sum().collect()
```

- **Jobs:** 1 -> collect es acción
- **Stages:** 2 -> groupby require shuffle
- **Tasks:** 8 -> Cada stage tiene 4 Task

Ejemplo 3 | Varios Jobs

```
df = spark.range(0, 100)

df.count()           # Primer Job
df.filter("id < 50").count()  # Segundo Job
```

- **Jobs:** 2 -> cada acción dispara un nuevo Job
- **Stages:** 2 (1 por cada job)
- **Tasks:** El número de particiones del DF (el que haya por defecto)

Ejemplo 4 | Join con shuffle

```
df1 = spark.range(0, 100).withColumnRenamed("id", "key")
df2 = spark.range(50, 150).withColumnRenamed("id", "key")

df1.join(df2, "key").count()
```

- **Jobs:** 1 -> count es acción
- **Stages:** 2 -> join implica shuffle
- **Tasks:** El número de particiones del DF (el que haya por defecto)

www.ceste.es