



CESTE

Escuela Internacional de Negocios

Zaragoza (España)

A person's hands are shown typing on a silver laptop keyboard. The laptop screen displays the CESTE online catalog website. The website has a blue header with the CESTE logo and the word 'Catalogo'. The main content area is a grid of colored boxes representing different academic programs. The visible categories include: BANCA Y FINANZAS (Banking and Finance), BUSINESS, CARRERAS UNIVERSITARIAS (University Majors), DATA SCIENCE Y BIG DATA (Data Science and Big Data), EMPRENDIMIENTO (Entrepreneurship), GESTIÓN LOGÍSTICA (Logistics Management), MASTER OFICIALES (Official Masters), and RECURSOS HUMANOS (Human Resources). Each category box shows a small icon, the program name, and the number of items available. On the right side of the website, there is a sidebar with the heading 'Áreas' (Areas) and a list of categories: Banca y Finanzas, Business, Carreras universitarias, Data Science y Big Data, Emprendimiento, Gestión logística, Master Oficiales, Recursos Humanos, Tecnologías de la Información y las Comunicaciones, and Zaragoza.

¿Quién soy?

- Pablo Sanz Caperote
- Doble Grado en Ingeniería Informática – Matemáticas
- Profesional con más de 4 años de experiencia entorno al mundo de los datos.
- Casi 2 años trabajando con Databricks
- Varias certificaciones en Clouds (AWS, Azure, GCP)



www.linkedin.com/in/pablosanzcaperote

Introducción a Apache Spark: RDDs, DataFrames, Datasets y Spark SQL



Introducción a Apache Spark: RDDs, DataFrames, Datasets y Spark SQL

Índice | Fundamentos y diagnóstico de rendimiento

Cambiar

1. ¿Qué es spark?
2. Estructura de datos
3. Funciones en pyspark



Objetivos | Una primera toma de contacto con Spark y PySpark



1. Entender Spark y sus principales componentes
2. Uso de Notebooks en Databricks
3. Uso de dbutils
4. Ejercicios sobre Spark

¿Qué es Apache Spark? | La mejor forma de tratar un volumen grande de datos



1. Motor unificado de analítica para procesar datos a gran escala
2. Motor de procesamiento distribuido
3. Permite desarrollo en Java, Scala, Python y R.
4. PySpark es la interfaz en Python de Apache Spark

Estructura de datos en Spark | Evolución en busca de mejoras

RDDs

- Estructura fundamental de datos de Spark
- Colección distribuida
- Inmutables

DataFrames

- Datos distribuidos de forma tabular
- Tienen schema
- Se puede usar SQL syntax

Datasets

- Combina lo mejor de RDDs y DF
- Funcionan con datos estructurados y no



¿Cómo operar con los datos? | Trasformaciones y acciones

Trasformaciones



- Definen como se modifican los datos (evaluación perezosa):
 - Select()
 - Filter()
 - withColumn()
 - groupBy()
 - Join()

Acciones



- Ejecutan el procesamiento y devuelven resultados:
 - Show()
 - Count()
 - Collect()
 - Write()



DEMO

A PRACTICAR CON PYSPARK



www.ceste.es