CESTE

Escuela Internacional de Negocios

Zaragoza (España)

¿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

CESTE

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

CESTE

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

CESTE

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 componenentes
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

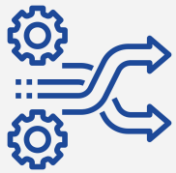
1. Motor unificado de analítica para procesar datos a gran escala
2. Motor de procesamiento distribuido
3. Permite desarrollo en Java, Scala, Python y R.
4. PySpark es la interfaz en Python de Apache Spark

Estructura de datos en Spark | Evolución en busca de mejoras

RDDs	DataFrames	Datasets
• Estructura fundamental de datos de Spark • Colección distribuida • Inmutables	• Datos distrubidos de forma tabular • Tienen schema • Se puede usar SQL sintax	• 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

CESTE

www.ceste.es