



EBook Gratis

APRENDIZAJE spark-dataframe

Free unaffiliated eBook created from
Stack Overflow contributors.

#spark-
dataframe

Tabla de contenido

Acerca de.....	1
Capítulo 1: Comenzando con el marco de datos de chispa.....	2
Observaciones.....	2
Examples.....	2
Instalación o configuración.....	2
Cargando datos en un DataFrame.....	2
Creditos.....	4

Acerca de

You can share this PDF with anyone you feel could benefit from it, downloaded the latest version from: [spark-dataframe](#)

It is an unofficial and free spark-dataframe ebook created for educational purposes. All the content is extracted from [Stack Overflow Documentation](#), which is written by many hardworking individuals at Stack Overflow. It is neither affiliated with Stack Overflow nor official spark-dataframe.

The content is released under Creative Commons BY-SA, and the list of contributors to each chapter are provided in the credits section at the end of this book. Images may be copyright of their respective owners unless otherwise specified. All trademarks and registered trademarks are the property of their respective company owners.

Use the content presented in this book at your own risk; it is not guaranteed to be correct nor accurate, please send your feedback and corrections to info@zzzprojects.com

Capítulo 1: Comenzando con el marco de datos de chispa

Observaciones

Esta sección proporciona una descripción general de qué es el marco de datos de chispa y por qué un desarrollador puede querer usarlo.

También debe mencionar cualquier tema grande dentro del marco de datos de chispa y vincular a los temas relacionados. Dado que la Documentación para el marco de datos de chispa es nueva, es posible que deba crear versiones iniciales de esos temas relacionados.

Examples

Instalación o configuración

Instrucciones detalladas sobre cómo configurar o instalar el marco de datos de chispa.

Cargando datos en un DataFrame

En Spark (scala) podemos obtener nuestros datos en un DataFrame de varias maneras diferentes, cada una para diferentes casos de uso.

Crear DataFrame desde CSV

La forma más fácil de cargar datos en un DataFrame es cargarlos desde un archivo CSV. Un ejemplo de esto (tomado de la [documentación oficial](#)) es:

```
import org.apache.spark.sql.SQLContext

val sqlContext = new SQLContext(sc)
val df = sqlContext.read
  .format("com.databricks.spark.csv")
  .option("header", "true") // Use first line of all files as header
  .option("inferSchema", "true") // Automatically infer data types
  .load("cars.csv")
```

Crear DataFrame desde RDD implícitamente

Muy a menudo en las aplicaciones de chispas tenemos datos en un RDD, pero necesitamos convertir esto en un DataFrame. La forma más fácil de hacer esto es usar la función RDD `.toDF()`, que determinará implícitamente los tipos de datos para nuestro DataFrame:

```
val data = List(
  ("John", "Smith", 30),
  ("Jane", "Doe", 25)
)
```

```
val rdd = sc.parallelize(data)

val df = rdd.toDF("firstname", "surname", "age")
```

Crear DataFrame desde RDD explícitamente

En algunos casos, el uso del enfoque `.toDF()` no es la mejor idea, ya que necesitamos definir explícitamente el esquema de nuestro DataFrame. Esto se puede lograr utilizando un `StructType` que contenga un `Array of StructField`.

```
import org.apache.spark.sql.types._
import org.apache.spark.sql.Row

val data = List(
  Array("John", "Smith", 30),
  Array("Jane", "Doe", 25)
)

val rdd = sc.parallelize(data)

val schema = StructType(
  Array(
    StructField("firstname", StringType, true),
    StructField("surname", StringType, false),
    StructField("age", IntegerType, true)
  )
)

val rowRDD = rdd.map(arr => Row(arr :_*))

val df = sqlContext.createDataFrame(rowRDD, schema)
```

Lea Comenzando con el marco de datos de chispa en línea: <https://riptutorial.com/es/spark-dataframe/topic/8988/comenzando-con-el-marco-de-datos-de-chispa>

Creditos

S. No	Capítulos	Contributors
1	Comenzando con el marco de datos de chispa	Ben Horsburgh, Community