In PySpark, structType() and StructField() are functions used to define the schema for structured data, specifically when working with DataFrame and Dataset API. These functions are essential when you want to create a structured representation of your data, similar to defining a table schema in a relational database.

**StructType():**

StructType() is a constructor function in PySpark that creates a new instance of a StructType object. It is used to define the schema for a structured data type, which is essentially a collection of fields, where each field has a name and a data type. This function allows you to define the structure of a DataFrame or Dataset.

**Syntax:**

from pyspark.sql.types import StructType

schema = StructType([

StructField("field\_name1", data\_type1, nullable1),

StructField("field\_name2", data\_type2, nullable2),

# Add more fields as needed

])

Breakdown of the syntax:

field\_name: The name of the field/column in the DataFrame.

data\_type: The data type of the field, which can be one of the PySpark data types (StringType, IntegerType, FloatType, DoubleType, BooleanType, etc.).

nullable: A boolean value indicating whether the field can contain null values (True) or not (False).

**StructField():**

StructField() is a constructor function in PySpark that creates a new instance of a StructField object. It represents a single field within a StructType schema. This function is used when you need to define a field with a specific name and data type.

**Syntax:**

from pyspark.sql.types import StructField, StringType, IntegerType

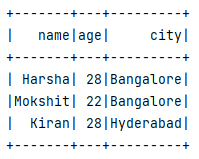
field1 = StructField("field\_name", StringType(), nullable=False)

field2 = StructField("another\_field", IntegerType(), nullable=True)

1. Example of Creating a DataFrame with a specified schema:



OutPut:



1. Nested StructType: we'll define a schema with a nested structure. We'll create a DataFrame representing information about employees, where each employee has personal details and a list of projects they are working on.



Output:

A screen shot of a computer code

Description automatically generated

1. StructType as a column in DataFrame: In this scenario, we'll create a DataFrame where one of the columns contains a structType. We'll represent information about books, where each book has a title, author, and publication details stored as a struct.



Output:

A screenshot of a computer code

Description automatically generated