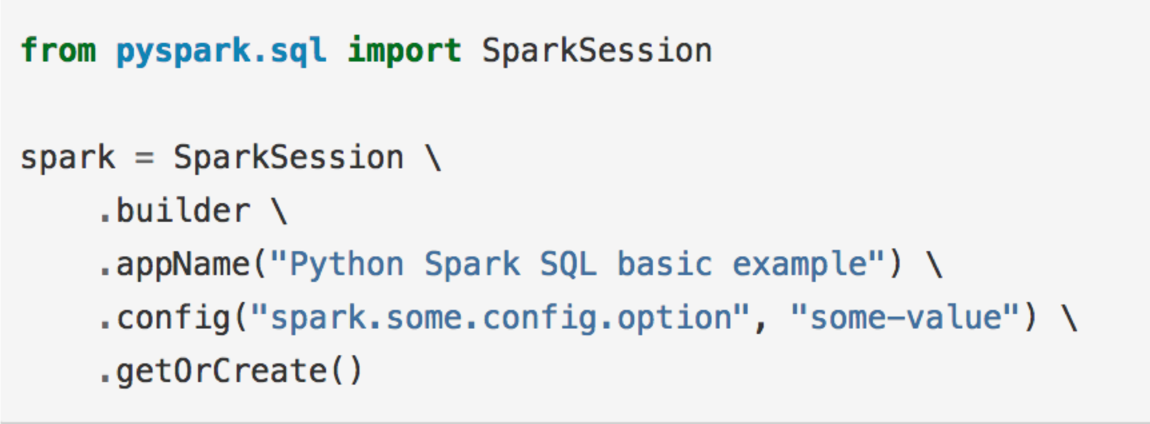
# DataFrames

Create DataFrames

# Create DataFrames

To work with DataFrames you must first create a SparkSession.



```
from pyspark.sql import SparkSession

spark = SparkSession \
    .builder \
    .appName("Python Spark SQL basic example") \
    .config("spark.some.config.option", "some-value") \
    .getOrCreate()
```

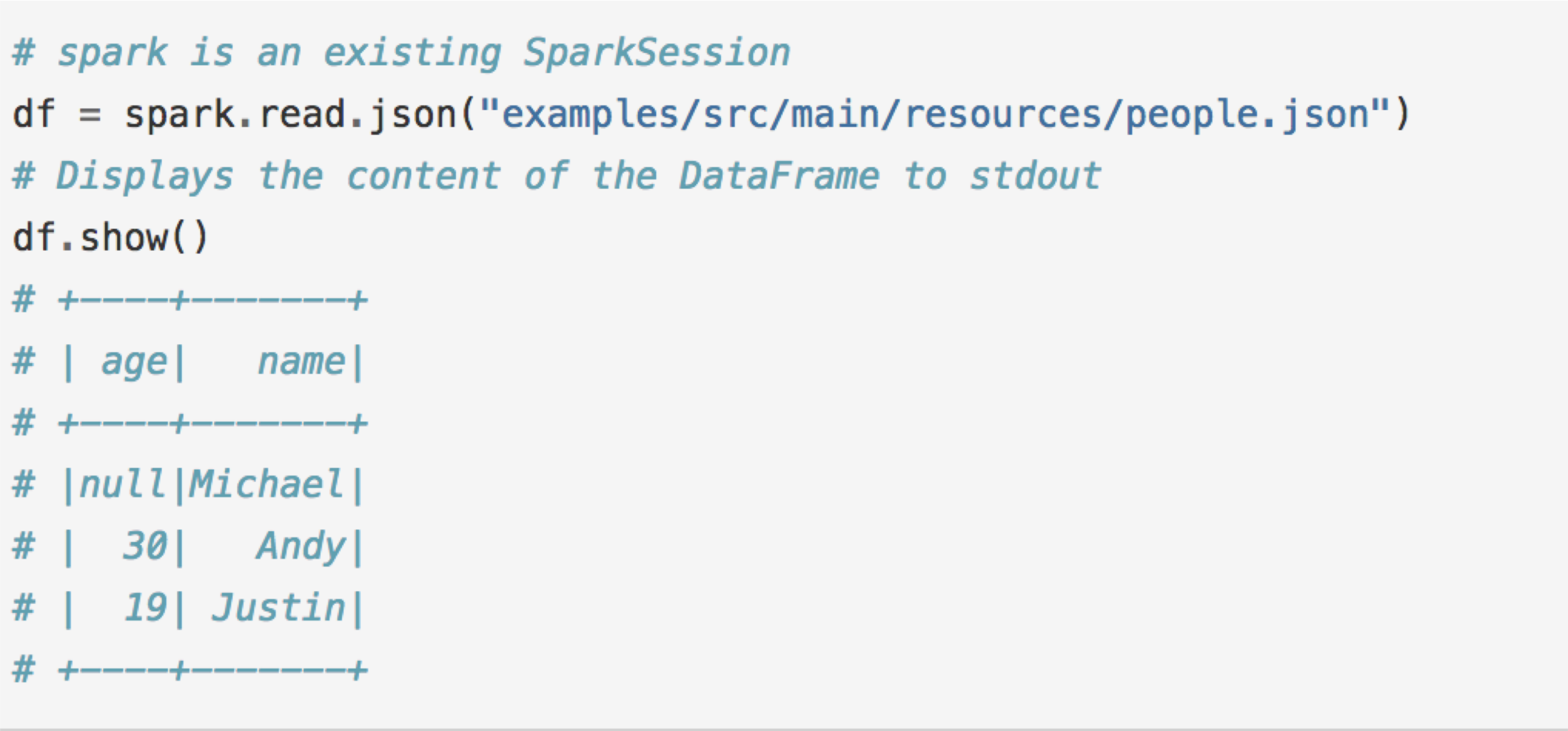
(notebook : SparkDataFrame/ CreateDataFrames.ipynb)

# Creating DataFrames

There are two ways to create DataFrames

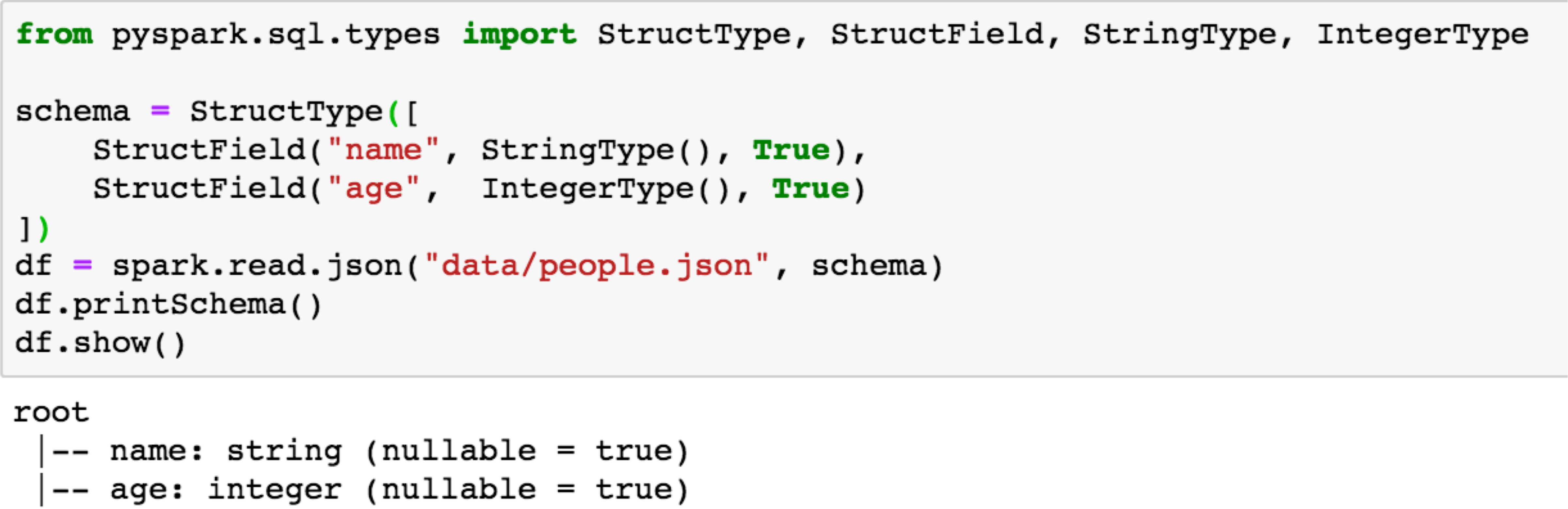
- ➢From an external file using either an inferred schema or one that you provide
- ➢From an exiting RDD

# Create a DataFrame from an outside source with an inferred schema



```
# spark is an existing SparkSession
df = spark.read.json("examples/src/main/resources/people.json")
# Displays the content of the DataFrame to stdout
df.show()
# +----+-------+
# | age|   name|
# +----+-------+
# |null|Michael|
# |  30|   Andy|
# |  19| Justin|
# +----+-------+
```

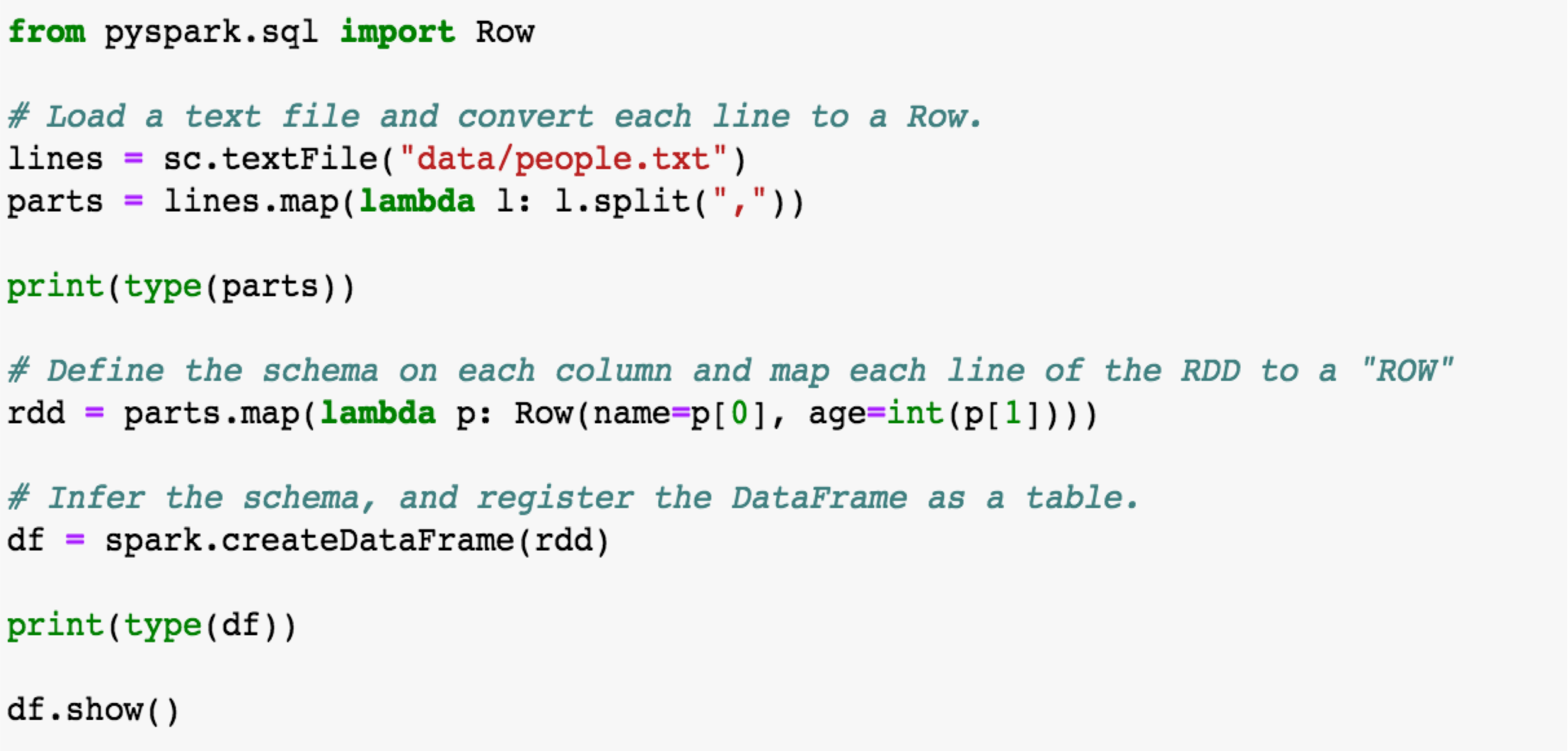
# Define the schema, then read into a DF



```python
from pyspark.sql.types import StructType, StructField, StringType, IntegerType

schema = StructType([
    StructField("name", StringType(), True),
    StructField("age",  IntegerType(), True)
])
df = spark.read.json("data/people.json", schema)
df.printSchema()
df.show()
```

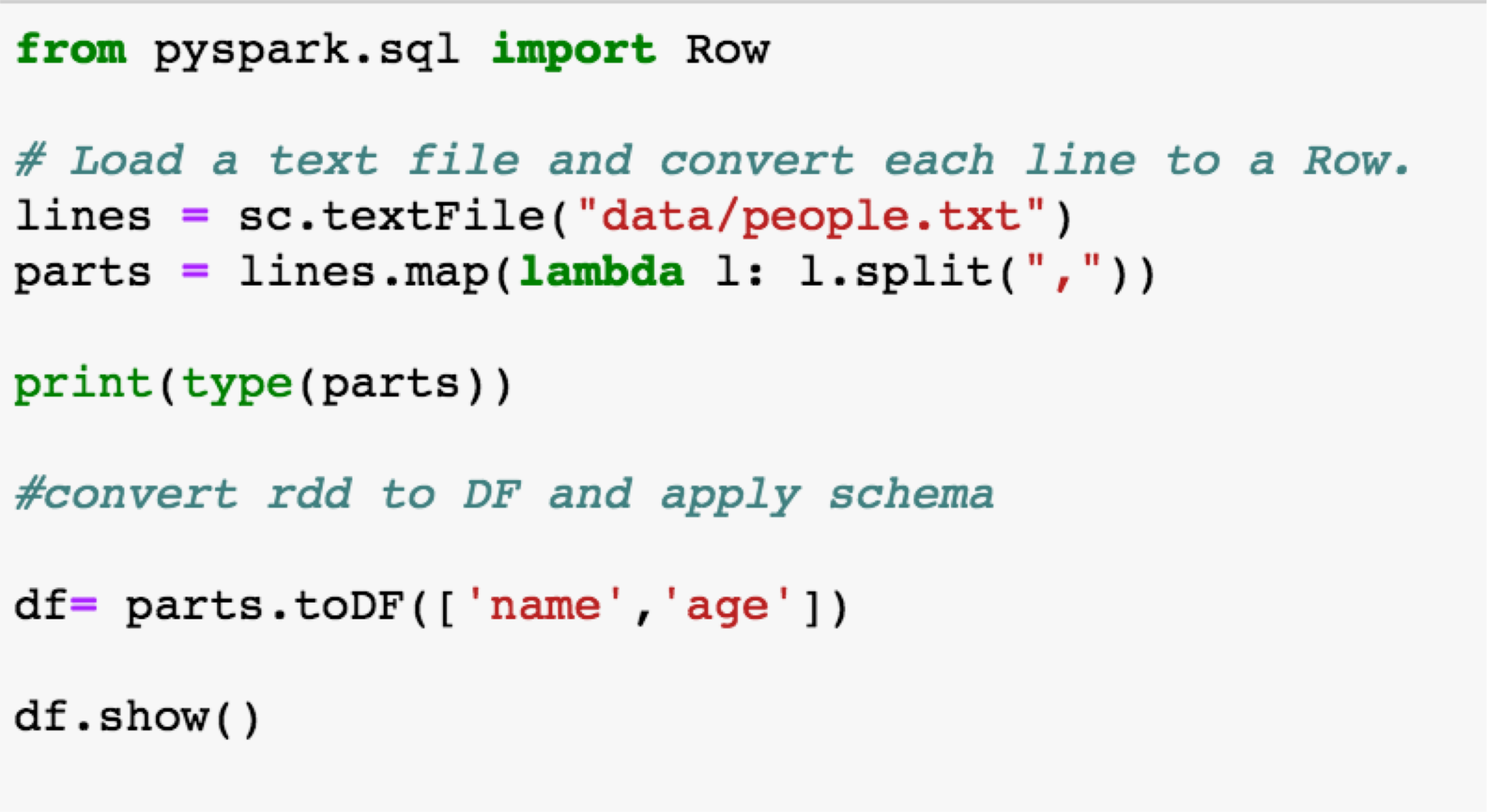
```
root
 |-- name: string (nullable = true)
 |-- age: integer (nullable = true)
```

# Read an RDD, then map to a "ROW" (df), apply a schema



```python
from pyspark.sql import Row

# Load a text file and convert each line to a Row.
lines = sc.textFile("data/people.txt")
parts = lines.map(lambda l: l.split(","))

print(type(parts))

# Define the schema on each column and map each line of the RDD to a "ROW"
rdd = parts.map(lambda p: Row(name=p[0], age=int(p[1])))

# Infer the schema, and register the DataFrame as a table.
df = spark.createDataFrame(rdd)

print(type(df))

df.show()
```

# Read data into a RDD, convert to DF using toDF() and schema



```python
from pyspark.sql import Row

# Load a text file and convert each line to a Row.
lines = sc.textFile("data/people.txt")
parts = lines.map(lambda l: l.split(","))

print(type(parts))

#convert rdd to DF and apply schema

df= parts.toDF(['name','age'])

df.show()
```