«

JSON Files

Scala Java **Python** R Sql

Spark SQL can automatically infer the schema of a JSON dataset and load it as a DataFrame. This conversion can be done using `SparkSession.read.json` on a JSON file.

Note that the file that is offered as *a json file* is not a typical JSON file. Each line must contain a separate, self-contained valid JSON object. For more information, please see JSON Lines text format, also called newline-delimited JSON.

For a regular multi-line JSON file, set the `multiLine` parameter to `True`.

```
# spark is from the previous example.
sc = spark.sparkContext

# A JSON dataset is pointed to by path.
# The path can be either a single text file or a directory storing text files
path = "examples/src/main/resources/people.json"
peopleDF = spark.read.json(path)

# The inferred schema can be visualized using the printSchema() method
peopleDF.printSchema()
# root
#  |-- age: long (nullable = true)
#  |-- name: string (nullable = true)

# Creates a temporary view using the DataFrame
peopleDF.createOrReplaceTempView("people")

# SQL statements can be run by using the sql methods provided by spark
teenagerNamesDF = spark.sql("SELECT name FROM people WHERE age BETWEEN 13 AND 19")
teenagerNamesDF.show()
# +------+
# |  name|
# +------+
# |Justin|
# +------+

# Alternatively, a DataFrame can be created for a JSON dataset represented by
# an RDD[String] storing one JSON object per string
jsonStrings = ['{"name":"Yin","address":{"city":"Columbus","state":"Ohio"}}']
otherPeopleRDD = sc.parallelize(jsonStrings)
otherPeople = spark.read.json(otherPeopleRDD)
otherPeople.show()
# +---------------+----+
# |        address|name|
# +---------------+----+
# |[Columbus,Ohio]| Yin|
# +---------------+----+
```

Find full example code at "examples/src/main/python/sql/datasource.py" in the Spark repo.