

## Spark by {Examples} (https://sparkbyexamples.com/)

### PySpark Tutorial

[PySpark Tutorial For Beginners](#)

(https://sparkbyexamples.com/pyspark-tutorial/)

[PySpark – Features](#)

(https://sparkbyexamples.com/apache-hive-tutorial/#features)

[PySpark – Advantages](#)

(https://sparkbyexamples.com/apache-hive-tutorial/#advantages)

[PySpark – Modules & Packages](#)

(https://sparkbyexamples.com/apache-hbase-tutorial/#modules-packages)

[PySpark – Cluster Managers](#)

(https://sparkbyexamples.com/apache-kafka-tutorials-with-examples/#pyspark)

[PySpark – Install on Windows](#)

(https://sparkbyexamples.com/apache-kafka-tutorials-with-examples/#pyspark-installation)

[PySpark – Web/Application UI](#)

(https://sparkbyexamples.com/spark/spark-web-ui-understanding/)

[PySpark – SparkSession](#)

(https://sparkbyexamples.com/spark/pyspark-what-is-sparksession/)

[PySpark – RDD](#)

(https://sparkbyexamples.com/pyspark-rdd)

[PySpark – Parallelize](#)

(https://sparkbyexamples.com/pyspark/pyspark-parallelize-create-rdd/)

[PySpark – repartition\(\) vs coalesce\(\)](#)

(https://sparkbyexamples.com/pyspark/pyspark-repartition-vs-coalesce/)

[PySpark – Broadcast Variables](#)

(https://sparkbyexamples.com/pyspark/pyspark-broadcast-variables/)

[PySpark](#) (https://sparkbyexamples.com/pyspark-tutorial/)

[Hive](#) (https://sparkbyexamples.com/apache-hive-tutorial/)

[HBase](#) (https://sparkbyexamples.com/apache-hbase-tutorial/)

[Kafka](#) (https://sparkbyexamples.com/apache-kafka-tutorials-with-examples/)

[FAQ's](#) (https://sparkbyexamples.com/spark-questions/)

[More](#) (https://sparkbyexamples.com/)

**Data Science with Internship -  
25,000+ Certified  
Datamites - Data Science Course...**

## PySpark Read JSON file into DataFrame

 [NNK](#)

(https://sparkbyexamples.com/author/admin/)

 [PySpark](#)

(https://sparkbyexamples.com/category/pyspark/)

**PySpark SQL**

provides `read.json("path")` to read a single line or multiline (multiple lines)

JSON file into PySpark DataFrame

and `write.json("path")` to save or

write to JSON file, In this tutorial, you

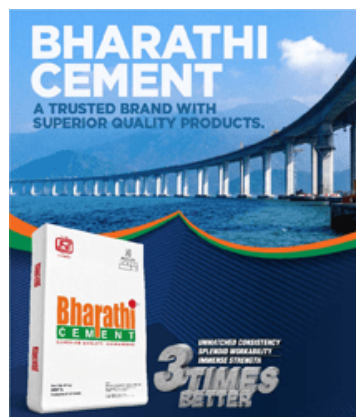
will learn how to read a single file,

multiple files, all files from a directory

into DataFrame and writing DataFrame

back to JSON file using Python

example.



**Related:**

- [PySpark Parse JSON from String Column | TEXT File](#)



[PySpark – Accumulator  
\(https://sparkbyexamples.com/pyspark/pyspark-accumulator-with-example/\)](https://sparkbyexamples.com/pyspark/pyspark-accumulator-with-example/)

---

## PySpark DataFrame

[PySpark – Create a DataFrame  
\(https://sparkbyexamples.com/pyspark/different-ways-to-create-dataframe-in-pyspark/\)](https://sparkbyexamples.com/pyspark/different-ways-to-create-dataframe-in-pyspark/)

---

[PySpark – Create an empty DataFrame  
\(https://sparkbyexamples.com/pyspark/pyspark-create-an-empty-dataframe/\)](https://sparkbyexamples.com/pyspark/pyspark-create-an-empty-dataframe/)

---

[PySpark – Convert RDD to DataFrame  
\(https://sparkbyexamples.com/pyspark/convert-pyspark-rdd-to-dataframe/\)](https://sparkbyexamples.com/pyspark/convert-pyspark-rdd-to-dataframe/)

---

[PySpark – Convert DataFrame to Pandas  
\(https://sparkbyexamples.com/pyspark/convert-pyspark-dataframe-to-pandas/\)](https://sparkbyexamples.com/pyspark/convert-pyspark-dataframe-to-pandas/)

---

[PySpark – show\(\)  
\(https://sparkbyexamples.com/pyspark/pyspark-show-display-dataframe-contents-in-table/\)](https://sparkbyexamples.com/pyspark/pyspark-show-display-dataframe-contents-in-table/)

---

[PySpark – StructType & StructField  
\(https://sparkbyexamples.com/pyspark/pyspark-structtype-and-structfield/\)](https://sparkbyexamples.com/pyspark/pyspark-structtype-and-structfield/)

---

[PySpark – Row Class  
\(https://sparkbyexamples.com/pyspark/pyspark-row-using-rdd-dataframe/\)](https://sparkbyexamples.com/pyspark/pyspark-row-using-rdd-dataframe/)

---

[PySpark – Column Class  
\(https://sparkbyexamples.com/pyspark/pyspark-column-functions/\)](https://sparkbyexamples.com/pyspark/pyspark-column-functions/)

---

[PySpark – select\(\)  
\(https://sparkbyexamples.com/pyspark/select-columns-from-pyspark-dataframe/\)](https://sparkbyexamples.com/pyspark/select-columns-from-pyspark-dataframe/)

---

[PySpark – collect\(\)  
\(https://sparkbyexamples.com/pyspark/pyspark-collect/\)](https://sparkbyexamples.com/pyspark/pyspark-collect/)

---

[PySpark – withColumn\(\)  
\(https://sparkbyexamples.com/pyspark/pyspark-withcolumn/\)](https://sparkbyexamples.com/pyspark/pyspark-withcolumn/)

---

[\(https://sparkbyexamples.com/pyspark/pyspark-parse-json-from-string-column-text-file/\)](https://sparkbyexamples.com/pyspark/pyspark-parse-json-from-string-column-text-file/)

- [Convert JSON Column to Struct, Map or Multiple Columns in PySpark  
\(https://sparkbyexamples.com/spark/spark-from\\_json-convert-json-column-to-struct-map-or-multiple-columns/\)](https://sparkbyexamples.com/spark/spark-from_json-convert-json-column-to-struct-map-or-multiple-columns/)
- [Most used PySpark JSON Functions with Examples  
\(https://sparkbyexamples.com/pyspark/pyspark-json-functions-with-examples/\)](https://sparkbyexamples.com/pyspark/pyspark-json-functions-with-examples/)

**Note:** PySpark API out of the box supports to read JSON files and many more file formats into PySpark DataFrame.

### Table of contents:

- [PySpark Read JSON file into DataFrame](#)
- [Read JSON file from multiline](#)
- [Read multiple files at a time](#)
- [Read all files in a directory](#)
- [Read file with a user-specified schema](#)
- [Read file using PySpark SQL](#)
- [Options while reading JSON file](#)
  - [nullValues](#)
  - [dateFormat](#)
- [PySpark Write DataFrame to JSON file](#)
  - [Using options](#)
  - [Saving Mode](#)

## PySpark Read JSON file into DataFrame

Using `read.json("path")` or `read.format("json").load("path")` you can read a JSON file into a PySpark DataFrame, these methods take a file path as an argument.

[PySpark – withColumnRenamed\(\).](#)  
(<https://sparkbyexamples.com/pyspark/pyspark-rename-dataframe-column/>).

[PySpark – where\(\) & filter\(\).](#)  
(<https://sparkbyexamples.com/pyspark/pyspark-where-filter/>).

[PySpark – drop\(\) & dropDuplicates\(\).](#)  
(<https://sparkbyexamples.com/pyspark/pyspark-distinct-to-drop-duplicates/>).

[PySpark – orderBy\(\) and sort\(\).](#)  
(<https://sparkbyexamples.com/pyspark/pyspark-orderby-and-sort-explained/>).

[PySpark – groupBy\(\).](#)  
(<https://sparkbyexamples.com/pyspark/pyspark-groupby-explained-with-example/>).

[PySpark – join\(\).](#)  
(<https://sparkbyexamples.com/pyspark/pyspark-join-explained-with-examples/>).

[PySpark – union\(\) & unionAll\(\).](#)  
(<https://sparkbyexamples.com/pyspark/pyspark-union-and-unionall/>).

[PySpark – unionByName\(\).](#)  
(<https://sparkbyexamples.com/spark/spark-merge-two-dataframes-with-different-columns/>).

[PySpark – UDF \(User Defined Function\).](#)  
(<https://sparkbyexamples.com/pyspark/pyspark-udf-user-defined-function/>).

[PySpark – map\(\).](#)  
(<https://sparkbyexamples.com/pyspark/pyspark-map-transformation/>).

[PySpark – flatMap\(\).](#)  
(<https://sparkbyexamples.com/pyspark/pyspark-flatmap-transformation/>).

[pyspark – foreach\(\).](#)  
(<https://sparkbyexamples.com/pyspark/pyspark-loop-iterate-through-rows-in-dataframe/#use-foreach-loop-through-dataframe>).

[PySpark – sample\(\) vs sampleBy\(\).](#)

## Data Science Course Bangalore Certified

Datamites - Data Science Courses in I

6-month/400 learning hours, 120-hour training, capstone & client projects



WEBSITE



Unlike [reading a CSV](#)  
(<https://sparkbyexamples.com/pyspark/pyspark-read-csv-file-into-dataframe/>),

By default JSON data source infers schema from an input file.

[zipcodes.json](#) (<https://github.com/spark-examples/pyspark-examples/blob/master/resources/zipcodes.json>) file used here can be downloaded from GitHub project.

```
# Read JSON file into dataframe
df = spark.read.json("resources/zipcodes.json")
df.printSchema()
df.show()
```

When you use `format("json")` method, you can also specify the Data sources by their fully qualified name as below.

```
# Read JSON file into dataframe
df = spark.read.format('org.apache.spark.sql.json')
df.load("resources/zipcodes.json")
```

## Read JSON file from multiline

PySpark JSON data source provides multiple options to read files in different options, use multiline option to read JSON files scattered across multiple lines. By default multiline option, is set to false.

[\(https://sparkbyexamples.com/pyspark/pyspark-sampling-example/\)](https://sparkbyexamples.com/pyspark/pyspark-sampling-example/)

[PySpark – fillna\(\) & fill\(\).  
\(https://sparkbyexamples.com/pyspark/pyspark-fillna-fill-replace-null-values/\)](https://sparkbyexamples.com/pyspark/pyspark-fillna-fill-replace-null-values/)

[PySpark – pivot\(\)\\_\(Row to Column\).  
\(https://sparkbyexamples.com/pyspark/pyspark-pivot-and-unpivot-dataframe/\)](https://sparkbyexamples.com/pyspark/pyspark-pivot-and-unpivot-dataframe/)

[PySpark – partitionBy\(\).  
\(https://sparkbyexamples.com/pyspark/pyspark-partitionby-example/\)](https://sparkbyexamples.com/pyspark/pyspark-partitionby-example/)

[PySpark – ArrayType Column \(Array\).  
\(https://sparkbyexamples.com/pyspark/pyspark-arraytype-column-with-examples/\)](https://sparkbyexamples.com/pyspark/pyspark-arraytype-column-with-examples/)

[PySpark – MapType \(Map/Dict\).  
\(https://sparkbyexamples.com/pyspark/pyspark-maptype-dict-examples/\)](https://sparkbyexamples.com/pyspark/pyspark-maptype-dict-examples/)

## PySpark SQL Functions

[PySpark – Aggregate Functions  
\(https://sparkbyexamples.com/pyspark/pyspark-aggregate-functions/\)](https://sparkbyexamples.com/pyspark/pyspark-aggregate-functions/)

[PySpark – Window Functions  
\(https://sparkbyexamples.com/pyspark/pyspark-window-functions/\)](https://sparkbyexamples.com/pyspark/pyspark-window-functions/)

[PySpark – Date and Timestamp Functions  
\(https://sparkbyexamples.com/pyspark/pyspark-sql-date-and-timestamp-functions/\)](https://sparkbyexamples.com/pyspark/pyspark-sql-date-and-timestamp-functions/)

[PySpark – JSON Functions  
\(https://sparkbyexamples.com/pyspark/pyspark-json-functions-with-examples/\)](https://sparkbyexamples.com/pyspark/pyspark-json-functions-with-examples/)

## PySpark Datasources

[PySpark – Read & Write CSV File  
\(https://sparkbyexamples.com/pyspark/pyspark-read-csv-file-into-dataframe/\)](https://sparkbyexamples.com/pyspark/pyspark-read-csv-file-into-dataframe/)

[PySpark – Read & Write Parquet File  
\(https://sparkbyexamples.com/pyspark/pyspark-read-write-parquet-file/\)](https://sparkbyexamples.com/pyspark/pyspark-read-write-parquet-file/)

Below is the input file we going to read, this same file is also available at [Github](https://github.com/spark-examples/pyspark-examples/blob/master/resources/multiline-zipcode.json) (<https://github.com/spark-examples/pyspark-examples/blob/master/resources/multiline-zipcode.json>).

### Looking for jobs after 12th

HCL

```
[{
  "RecordNumber": 2,
  "Zipcode": 704,
  "ZipCodeType": "STANDARD",
  "City": "PASEO COSTA DEL SUR",
  "State": "PR"
},
{
  "RecordNumber": 10,
  "Zipcode": 709,
  "ZipCodeType": "STANDARD",
  "City": "BDA SAN LUIS",
  "State": "PR"
}]
```

Using

```
read.option("multiline","true")
```

```
# Read multiline json file
multiline_df = spark.read.option(
    "multiline", "true").json("resources/multiline-zipcode.json")
multiline_df.show()
```

## Reading multiple files at a time


Using the `read.json()` method you can also read multiple JSON files from different paths, just pass all file names with fully qualified paths by separating comma, for example



[pyspark/pyspark-read-and-write-parquet-file/](#)

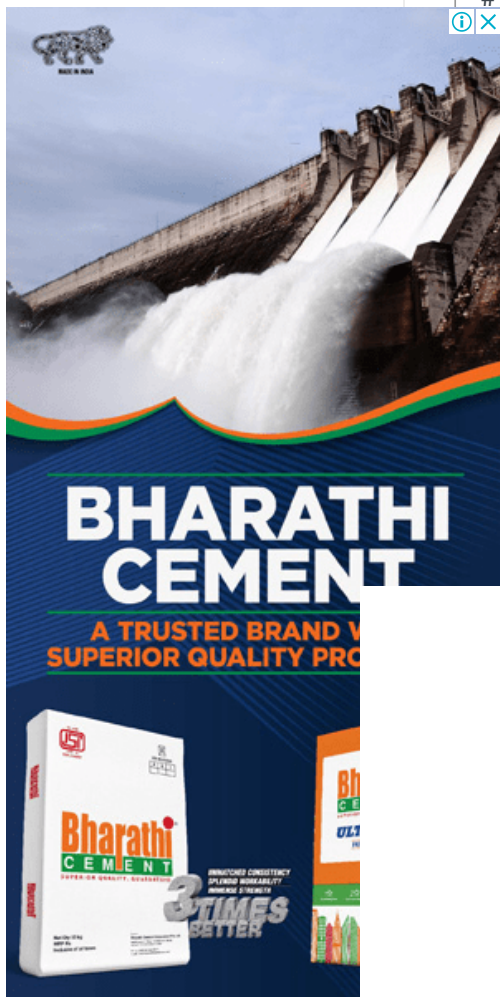
[PySpark – Read & Write JSON file](#)  
(<https://sparkbyexamples.com/pyspark/pyspark-read-json-file-into-dataframe/>)

**Laptop Bag Offer by Skechers**  
Skechers - WHC Road, Dharampeth, Nagpur

Get a Free Laptop Bag! Shop for ₹6999  
Store T&C\*

  
Nagpur

 STORE INFO 



## PySpark Built-In Functions

[PySpark – when\(\)](#)  
(<https://sparkbyexamples.com/pyspark/pyspark-when-otherwise/>)

[PySpark – expr\(\)](#)  
(<https://sparkbyexamples.com/pyspark/pyspark-sql-expr-expression-function/>)

```
# Read multiple files
df2 = spark.read.json(
    ['resources/zipcode1.json',
df2.show()
```

## Reading all files in a directory

can read all JSON files from a directory into DataFrame just by using directory as a path to the `read()` method.

```
# Read all JSON files from a folder
df3 = spark.read.json("resources/zipcodes")
df3.show()
```

## Reading files with a user-specified custom schema

PySpark Schema defines the structure of the data, in other words, it is the structure of the DataFrame. PySpark provides `StructType` & `StructField` classes to programmatically specify the structure to the DataFrame.

If you know the schema of the file and do not want to use the default `inferSchema` option, use `schema` option to specify user-defined custom column names and data types.

Use the [PySpark StructType class to create a custom schema](#) (<https://sparkbyexamples.com/pyspark/pyspark-structtype-and-structfield/>), below we initiate this class and use `add` a method to add columns to it by providing the column name, data type and nullable option.

[PySpark – lit\(\)](#)  
(<https://sparkbyexamples.com/pyspark/pyspark-lit-add-literal-constant/>).

[PySpark – split\(\)](#)  
(<https://sparkbyexamples.com/pyspark/pyspark-convert-string-to-array-column/>).

[PySpark – concat\\_ws\(\)](#)  
(<https://sparkbyexamples.com/pyspark/pyspark-convert-array-column-to-string-column/>).

[Pyspark – substring\(\)](#)  
(<https://sparkbyexamples.com/pyspark/pyspark-substring-from-a-column/>).

[PySpark – translate\(\)](#)  
(<https://sparkbyexamples.com/pyspark/pyspark-replace-column-values/#translate-replace-character-by-character>).

[PySpark – regexp\\_replace\(\)](#)  
([https://sparkbyexamples.com/pyspark/pyspark-replace-column-values/#regexp\\_replace-replace-string-columns](https://sparkbyexamples.com/pyspark/pyspark-replace-column-values/#regexp_replace-replace-string-columns)).

[PySpark – overlay\(\)](#)  
(<https://sparkbyexamples.com/pyspark/pyspark-replace-column-values/#overlay-function>).

[PySpark – to\\_timestamp\(\)](#)  
([https://sparkbyexamples.com/spark/pyspark-to\\_timestamp-convert-string-to-timestamp-type/](https://sparkbyexamples.com/spark/pyspark-to_timestamp-convert-string-to-timestamp-type/)).

[PySpark – to\\_date\(\)](#)  
([https://sparkbyexamples.com/pyspark/pyspark-to\\_date-convert-timestamp-to-date/](https://sparkbyexamples.com/pyspark/pyspark-to_date-convert-timestamp-to-date/)).

[PySpark – date\\_format\(\)](#)  
([https://sparkbyexamples.com/pyspark/pyspark-date\\_format-convert-date-to-string-format/](https://sparkbyexamples.com/pyspark/pyspark-date_format-convert-date-to-string-format/)).

[PySpark – datediff\(\)](#)  
(<https://sparkbyexamples.com/pyspark/pyspark-difference-between-two-dates-days-months-years/#datediff>).

[PySpark – months\\_between\(\)](#)  
([https://sparkbyexamples.com/pyspark/pyspark-difference-between-two-dates-days-months-years/#months\\_between\(\)](https://sparkbyexamples.com/pyspark/pyspark-difference-between-two-dates-days-months-years/#months_between())).

```
# Define custom schema
schema = StructType([
    StructField("RecordNumber", IntegerType(), nullable=False),
    StructField("Zipcode", IntegerType(), nullable=False),
    StructField("ZipCodeType", IntegerType(), nullable=False),
    StructField("City", StringType(), nullable=False),
    StructField("State", StringType(), nullable=False),
    StructField("LocationType", IntegerType(), nullable=False),
    StructField("Lat", DoubleType(), nullable=False),
    StructField("Long", DoubleType(), nullable=False),
    StructField("Xaxis", IntegerType(), nullable=False),
    StructField("Yaxis", DoubleType(), nullable=False),
    StructField("Zaxis", DoubleType(), nullable=False),
    StructField("WorldRegion", IntegerType(), nullable=False),
    StructField("Country", StringType(), nullable=False),
    StructField("LocationText", StringType(), nullable=False),
    StructField("Location", StringType(), nullable=False),
    StructField("Decommisioned", BooleanType(), nullable=False),
    StructField("TaxReturnsFiled", IntegerType(), nullable=False),
    StructField("EstimatedPopulation", IntegerType(), nullable=False),
    StructField("TotalWages", IntegerType(), nullable=False),
    StructField("Notes", StringType(), nullable=False)
])

df_with_schema = spark.read.schema(schema).json("resources/zipcode")
df_with_schema.printSchema()
df_with_schema.show()
```

## Read JSON file using PySpark SQL

PySpark SQL also provides a way to read a JSON file by creating a temporary view directly from the reading file using `spark.sqlContext.sql("load JSON to temporary view")`

```
spark.sql("CREATE OR REPLACE TEMPORARY VIEW zipcodes "
          "(path 'resources/zipcode.json') AS "
          "SELECT * FROM json_tuple(path, 'resources/zipcode.json')")
spark.sql("select * from zipcodes")
```

## Options while reading JSON file



[PySpark – explode\(\)](https://sparkbyexamples.com/py-spark/pyspark-explode-nested-array-into-rows/)  
(<https://sparkbyexamples.com/py-spark/pyspark-explode-nested-array-into-rows/>).

[PySpark – array\\_contains\(\)](https://sparkbyexamples.com/py-spark/pyspark-arraytype-column-with-examples/#array_contains)  
([https://sparkbyexamples.com/py-spark/pyspark-arraytype-column-with-examples/#array\\_contains](https://sparkbyexamples.com/py-spark/pyspark-arraytype-column-with-examples/#array_contains)).

[PySpark – array\(\)](https://sparkbyexamples.com/py-spark/pyspark-arraytype-column-with-examples/#array)  
(<https://sparkbyexamples.com/py-spark/pyspark-arraytype-column-with-examples/#array>).

[PySpark – collect\\_list\(\)](https://sparkbyexamples.com/py-spark/pyspark-aggregate-functions/#collect-list)  
(<https://sparkbyexamples.com/py-spark/pyspark-aggregate-functions/#collect-list>).

[PySpark – collect\\_set\(\)](https://sparkbyexamples.com/py-spark/pyspark-aggregate-functions/#collect-set)  
(<https://sparkbyexamples.com/py-spark/pyspark-aggregate-functions/#collect-set>).

[PySpark – create\\_map\(\)](https://sparkbyexamples.com/py-spark/pyspark-convert-dataframe-columns-to-maptype-dict/)  
(<https://sparkbyexamples.com/py-spark/pyspark-convert-dataframe-columns-to-maptype-dict/>).

[PySpark – map\\_keys\(\)](https://sparkbyexamples.com/py-spark/pyspark-maptype-dict-examples/#map_keys)  
([https://sparkbyexamples.com/py-spark/pyspark-maptype-dict-examples/#map\\_keys](https://sparkbyexamples.com/py-spark/pyspark-maptype-dict-examples/#map_keys)).

[PySpark – map\\_values\(\)](https://sparkbyexamples.com/py-spark/pyspark-maptype-dict-examples/#map_values)  
([https://sparkbyexamples.com/py-spark/pyspark-maptype-dict-examples/#map\\_values](https://sparkbyexamples.com/py-spark/pyspark-maptype-dict-examples/#map_values)).

[PySpark – struct\(\)](https://sparkbyexamples.com/py-spark/pyspark-structtype-and-structfield/#update-struct-function)  
(<https://sparkbyexamples.com/py-spark/pyspark-structtype-and-structfield/#update-struct-function>).

[PySpark – countDistinct\(\)](https://sparkbyexamples.com/py-spark/pyspark-count-distinct-from-dataframe/)  
(<https://sparkbyexamples.com/py-spark/pyspark-count-distinct-from-dataframe/>).

[PySpark – sum\(\).avg\(\)](https://sparkbyexamples.com/py-spark/pyspark-dataframe-groupby-and-sort-by-descending-order/)  
(<https://sparkbyexamples.com/py-spark/pyspark-dataframe-groupby-and-sort-by-descending-order/>).

[PySpark – row\\_number\(\)](https://sparkbyexamples.com/py-spark/pyspark-window-functions/#row_number)  
([https://sparkbyexamples.com/py-spark/pyspark-window-functions/#row\\_number](https://sparkbyexamples.com/py-spark/pyspark-window-functions/#row_number)).

[PySpark – rank\(\)](https://sparkbyexamples.com/py-spark/pyspark-window-functions/#rank)  
(<https://sparkbyexamples.com/py-spark/pyspark-window-functions/#rank>).

## nullValues

Using nullValues option you can specify the string in a JSON to consider as null. For example, if you want to consider a date column with a value “1900-01-01” set null on DataFrame.

## dateFormat

dateFormat option is used to set the format of the input DateType and TimestampType columns. Supports all [java.text.SimpleDateFormat](https://docs.oracle.com/javase/10/docs/api/java/time/format/DateTimeFormatter.html) (<https://docs.oracle.com/javase/10/docs/api/java/time/format/DateTimeFormatter.html>) formats.

**Note:** Besides the above options, PySpark JSON dataset also supports many other options.

## Applying DataFrame transformations

Once you have [create PySpark DataFrame](https://sparkbyexamples.com/pyspark/different-ways-to-create-dataframe-in-pyspark/) (<https://sparkbyexamples.com/pyspark/different-ways-to-create-dataframe-in-pyspark/>) from the JSON file, you can apply all transformation and actions DataFrame support. Please refer to the link for more details.

## Write PySpark

### DataFrame to JSON file

Use the PySpark DataFrameWriter object “write” method on DataFrame to write a JSON file.

```
df2.write.json("/tmp/spark_outp
```

## PySpark Options while writing JSON files

56% OFF

Setu Lean Lite

Checkout the latest products

Setu Nutrition

Setu Nutrition



[PySpark – dense\\_rank\(\).  
\(https://sparkbyexamples.com/pyspark/pyspark-window-functions/#dense\\_rank\)](https://sparkbyexamples.com/pyspark/pyspark-window-functions/#dense_rank)

[PySpark – percent\\_rank\(\).  
\(https://sparkbyexamples.com/pyspark/pyspark-window-functions/#percent\\_rank\)](https://sparkbyexamples.com/pyspark/pyspark-window-functions/#percent_rank)

[PySpark – typedLit\(\).  
\(https://sparkbyexamples.com/pyspark/pyspark-lit-add-literal-constant/#typedlit\)](https://sparkbyexamples.com/pyspark/pyspark-lit-add-literal-constant/#typedlit)

[PySpark – from\\_json\(\).  
\(https://sparkbyexamples.com/pyspark/pyspark-json-functions-with-examples/#from\\_json\)](https://sparkbyexamples.com/pyspark/pyspark-json-functions-with-examples/#from_json)

[PySpark – to\\_json\(\).  
\(https://sparkbyexamples.com/pyspark/pyspark-json-functions-with-examples/#to\\_json\)](https://sparkbyexamples.com/pyspark/pyspark-json-functions-with-examples/#to_json)

[PySpark – json\\_tuple\(\).  
\(https://sparkbyexamples.com/pyspark/pyspark-json-functions-with-examples/#json\\_tuple\)](https://sparkbyexamples.com/pyspark/pyspark-json-functions-with-examples/#json_tuple)

[PySpark – get\\_json\\_object\(\).  
\(https://sparkbyexamples.com/pyspark/pyspark-json-functions-with-examples/#get\\_json\\_object\)](https://sparkbyexamples.com/pyspark/pyspark-json-functions-with-examples/#get_json_object)

[PySpark – schema\\_of\\_json\(\).  
\(https://sparkbyexamples.com/pyspark/pyspark-json-functions-with-examples/#schema\\_of\\_json\)](https://sparkbyexamples.com/pyspark/pyspark-json-functions-with-examples/#schema_of_json)

While writing a JSON file you can use several options.

Other options

available nullValue, dateFormat

## PySpark Saving modes

PySpark DataFrameWriter also has a method mode() to specify SaveMode; the argument to this method either takes overwrite, append, ignore, errorifexists.

overwrite – mode is used to overwrite the existing file

append – To add the data to the existing file

ignore – Ignores write operation when the file already exists

errorifexists or error – This is a default option when the file already exists, it returns an error

```
df2.write.mode('Overwrite').json('data')
```

## Source code for reference

This example is also available at [GitHub PySpark Example Project \(https://github.com/spark-examples/pyspark-examples/blob/master/pyspark-read-json.py\)](https://github.com/spark-examples/pyspark-examples/blob/master/pyspark-read-json.py) for reference.



```

from pyspark.sql import SparkSession
from pyspark.sql.types import StructType, StructField, StringType, IntegerType, DoubleType

spark = SparkSession.builder \
    .master("local[1]") \
    .appName("SparkByExamples.com") \
    .getOrCreate()

# Read JSON file into dataframe
df = spark.read.json("resources/zipcode.json")
df.printSchema()
df.show()

# Read multiline json file
multiline_df = spark.read.options(headerLines=1) \
    .json("resources/multiline.json")
multiline_df.show()

# Read multiple files
df2 = spark.read.json(
    ["resources/zipcode2.json",
     "resources/zipcode3.json"])
df2.show()

# Read All JSON files from a directory
df3 = spark.read.json("resources/zipcodes")
df3.show()

# Define custom schema
schema = StructType([
    StructField("RecordNumber", IntegerType, True),
    StructField("Zipcode", IntegerType, True),
    StructField("ZipCodeType", IntegerType, True),
    StructField("City", StringType, True),
    StructField("State", StringType, True),
    StructField("LocationType", IntegerType, True),
    StructField("Lat", DoubleType, True),
    StructField("Long", DoubleType, True),
    StructField("Xaxis", IntegerType, True),
    StructField("Yaxis", DoubleType, True),
    StructField("Zaxis", DoubleType, True),
    StructField("WorldRegion", StringType, True),
    StructField("Country", StringType, True),
    StructField("LocationText", StringType, True),
    StructField("Location", StringType, True),
    StructField("Decommisioned", IntegerType, True),
    StructField("TaxReturnsFiled", IntegerType, True),
    StructField("EstimatedPopulation", IntegerType, True),
    StructField("TotalWages", IntegerType, True),
    StructField("Notes", StringType, True)
])

df_with_schema = spark.read.schema(schema) \
    .json("resources/zipcode.json")
df_with_schema.printSchema()
df_with_schema.show()

```

```
# Create a table from Parquet File
spark.sql("CREATE OR REPLACE TABLE zipcodes
          " (path 'resources/zipcodes.parquet')
spark.sql("select * from zipcodes")

# PySpark write Parquet File
df2.write.mode('Overwrite').json("resources/zipcodes.json")
```

## Conclusion:

In this tutorial, you have learned how to read a JSON file with single line record and multiline record into PySpark DataFrame, and also learned reading single and multiple files at a time and writing JSON file back to DataFrame using different save options.

## References:

- [Databricks read JSON](https://docs.databricks.com/data/data-sources/read-json.html)  
(<https://docs.databricks.com/data/data-sources/read-json.html>)
- [Spark json datasource](https://spark.apache.org/docs/latest/sql-data-sources-json.html)  
(<https://spark.apache.org/docs/latest/sql-data-sources-json.html>)
- [jsonlines.org](http://jsonlines.org/) (<http://jsonlines.org/>)
- [json.org](https://www.json.org/json-en.html) (<https://www.json.org/json-en.html>)
- [Spark JsonFileFormat scala class](https://github.com/apache/spark/blob/master/sql/core/src/main/scala/org/apache/spark/sql/execution/datasources/json/JsonFileFormat.scala)  
(<https://github.com/apache/spark/blob/master/sql/core/src/main/scala/org/apache/spark/sql/execution/datasources/json/JsonFileFormat.scala>)

Happy Learning !!

Share this:



(<https://sparkbyexamples.com/pyspark/pyspark-read-json-file-into-dataframe/?share=facebook&nb=1>)



(<https://sparkbyexamples.com/pyspark/pyspark-read-json-file-into-dataframe/?share=reddit&nb=1>)



(<https://sparkbyexamples.com/pyspark/pyspark-read-json-file-into-dataframe/?share=pinterest&nb=1>)

2



(<https://sparkbyexamples.com/pyspark/pyspark-read-json-file-into-dataframe/?share=tumblr&nb=1>)



(<https://sparkbyexamples.com/pyspark/pyspark-read-json-file-into-dataframe/?share=pocket&nb=1>)



(<https://sparkbyexamples.com/pyspark/pyspark-read-json-file-into-dataframe/?share=linkedin&nb=1>)



(<https://sparkbyexamples.com/pyspark/pyspark-read-json-file-into-dataframe/?share=twitter&nb=1>)

#### TAGS: [JSON](#)

([HTTPS://SPARKBYEXAMPLES.COM/TAG/JSON/](https://sparkbyexamples.com/tag/json/)),

#### [MULTILINE](#)

([HTTPS://SPARKBYEXAMPLES.COM/TAG/MULTILINE/](https://sparkbyexamples.com/tag/multiline/)), [SCHEMA](#)

([HTTPS://SPARKBYEXAMPLES.COM/TAG/SCHEMA/](https://sparkbyexamples.com/tag/schema/)).



**NNK**

([Https://Sparkbyexamples.Com/Author/Admin/](https://sparkbyexamples.com/author/admin/))

(<https://sparkbyexamples.com/author/admin/>)

SparkByExamples.com is a Big Data and Spark examples community page, all examples are simple and easy to understand and well tested in our development environment [Read more ..](#)

(<https://sparkbyexamples.com/about-sparkbyexamples/>)

## Leave a Reply

Enter your comment here...

### Categories

Apache Hadoop

(<https://sparkbyexamples.com/category/hadoop/>)

Apache Spark

(<https://sparkbyexamples.com/category/spark/>)

Apache Spark Streaming

(<https://sparkbyexamples.com/category/spark/apache-spark-streaming/>)

Apache Kafka

(<https://sparkbyexamples.com/category/kafka/>)

Apache HBase

(<https://sparkbyexamples.com/category/hbase/>)

### Recent Posts

Spark regexp\_replace() – Replace String Value

([https://sparkbyexamples.com/spark/spark-regexp\\_replace-replace-string-value/](https://sparkbyexamples.com/spark/spark-regexp_replace-replace-string-value/))

How to Run a PySpark Script from Python?

(<https://sparkbyexamples.com/pyspark/run-pyspark-script-from-python-subprocess/>)

Spark SQL like() Using Wildcard Example

(<https://sparkbyexamples.com/spark/spark-sql-like-using-wildcard-example/>)

Spark isin() & IS NOT IN Operator Example

(<https://sparkbyexamples.com/spark/spark-isin-is-not-in-operator-example/>)

### About SparkByExamples.Com

SparkByExamples.com is a Big Data and Spark examples community page, all examples are simple and easy to understand, and well tested in our development environment [Read more ..](#) (<https://sparkbyexamples.com/about-sparkbyexamples/>)

### Follow Us



Apache Cassandra  
(<https://sparkbyexamples.com/category/cassandra/>)

Snowflake Database  
(<https://sparkbyexamples.com/category/snowflake/>)

H2O Sparkling Water  
(<https://sparkbyexamples.com/category/h2o-sparkling-water/>)

PySpark  
(<https://sparkbyexamples.com/category/pyspark/>)

Spark – Get Size/Length of Array & Map Column  
(<https://sparkbyexamples.com/spark/spark-get-size-length-of-array-map-column/>)

Spark Using Length/Size Of a DataFrame Column  
(<https://sparkbyexamples.com/spark/spark-using-length-size-of-a-dataframe-column/>)

Spark rlike() Working with Regex Matching Examples  
(<https://sparkbyexamples.com/spark/spark-rlike-regex-matching-examples/>)

Spark Check String Column Has Numeric Values  
(<https://sparkbyexamples.com/spark/spark-check-string-column-has-numeric-values/>)

Spark Check Column Data Type is Integer or String  
(<https://sparkbyexamples.com/spark/spark-check-column-data-type-is-integer-or-string/>)

(<https://www.facebook.com/sparkbyexamples/>)

(<https://www.linkedin.com/company/sparkbyexamples/>)

(<https://twitter.com/sparkbyexamples>)

(<https://www.youtube.com/channel/UC8193l93>)

(<https://github.com/sparkbyexamples>)

(<https://www.b860a.com/sparkbyexamples/>)

(<https://www.sparkbyexamples.com>)

(<https://www.sparkbyexamples.com>)

