


Spark Connector Python Guide



On this page

- Prerequisites
- Getting Started
 - Python Spark Shell
 - Create a `SparkSession` Object
- Tutorials

SOURCE CODE:

For the source code that contains the examples below, see [introduction.py](#) .

Prerequisites

- Basic working knowledge of MongoDB and Apache Spark. Refer to the MongoDB documentation  and Spark documentation  for more details.
- Running MongoDB instance (version 2.6 or later).
- Spark 2.4.x.
- Scala 2.11.x or 2.12.x

Getting Started

Python Spark Shell

This tutorial uses the `pyspark` shell, but the code works with self-contained Python applications as well.

When starting the `pyspark` shell, you can specify:

- the `--packages` option to download the MongoDB Spark Connector package. The following package is available:
 - `mongo-spark-connector_2.11` for use with Scala 2.11.x
- the `--conf` option to configure the MongoDB Spark Connector. These settings configure the `SparkConf` object.

NOTE: When specifying the `spark.mongodb.input.uri` and `spark.mongodb.output.uri` configuration options, you must specify the database and collection appropriately. For details and other available MongoDB Spark Connector options, see the [MongoDB Spark Connector Configuration Options](#).

The following example starts the `pyspark` shell from the command line:

```
./bin/pyspark --conf "spark.mongodb.input.uri=mongodb://127.0.0.1/test.myCollection?readPreference=primary" \
--conf "spark.mongodb.output.uri=mongodb://127.0.0.1/test.myCollection" \
--packages org.mongodb.spark:mongo-spark-connector_2.11:2.4.0
```

- The `spark.mongodb.input.uri` specifies the MongoDB server address (`127.0.0.1`), the database to connect (`test`), and the collection (`myCollection`) from which to read data, and the read preference.
- The `spark.mongodb.output.uri` specifies the MongoDB server address (`127.0.0.1`), the database to connect (`test`), and the collection (`myCollection`) to which to write data. Connects to port `27017` by default.
- The `packages` option specifies the Spark Connector's Maven coordinates, in the format `groupId:artifactId:version`.

The examples in this tutorial will use this database and collection.

Create a `SparkSession` Object

NOTE: When you start `pyspark` you get a `SparkSession` object called `spark` by default. In a standalone Python application, you need to create your `SparkSession` object explicitly, as show below.

If you specified the `spark.mongodb.input.uri` and `spark.mongodb.output.uri` configuration options when you started `pyspark`, the default `SparkSession` object uses them. If you'd rather create your own `SparkSession` object from within `pyspark`, you can use `SparkSession.builder` and specify different configuration options.

from pyspark.sql
mongoDB. Documentation Search Documentation

```
my_spark = SparkSession \
    .builder \
    .appName("myApp") \
    .config("spark.mongodb.input.uri", "mongodb://127.0.0.1/test.coll") \
    .config("spark.mongodb.output.uri", "mongodb://127.0.0.1/test.coll") \
    .getOrCreate()
```

You can use a `SparkSession` object to write data to MongoDB, read data from MongoDB, create `DataFrames`, and perform SQL operations.

Tutorials

- Write to MongoDB
- Read from MongoDB
- Aggregation
- Filters and SQL

