Home (/) About (https://sparkbyexamples.com/about-sparkbyexamples/) | { One stop for all Spark Examples }

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

Spark (https://sparkbyexamples.com/)

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/)

Spark Read and Write Apache Parquet

NNK (https://sparkbyexamples.com/author/admin/) - Apache Spark (https://sparkbyexamples.com/category/spark/)

How to write and read Parquet files in Spark
APACHE Spark
Parquet
https://sparkbyexamples.com
One stop for all Spark Examples

(https://sparkbyexamples.com/spark-and-parquet-example/spark-parquet-example/)

Example of Spark read & write parquet file

In this tutorial, we will learn what is Apache Parquet?, It's advantages and how to read from and write Spark DataFrame to Parquet file format using Scala example. The example provided here is also available at Github

Spark RDD – Read text file (https://sparkbyexamples.com/apache-spark-rdd/spark-read-multiple-text-files-into-a-single-rdd/)

Spark RDD – Read CSV (https://sparkbyexamples.com/apache-spark-rdd/spark-load-csv-file-into-rdd/)

Spark RDD – Create RDD (https://sparkbyexamples.com/apache-spark-rdd/different-ways-to-create-spark-rdd/)

Spark RDD – Create Empty RDD (https://sparkbyexamples.com/apache-spark-rdd/spark-how-to-create-an-empty-rdd/)

Spark RDD – Transformations (https://sparkbyexamples.com/apache-spark-rdd/spark-rdd-transformations/)

Spark RDD – Actions (https://sparkbyexamples.com/apache-spark-rdd/spark-rdd-actions/)

Spark RDD – Pair Functions (https://sparkbyexamples.com/apache-spark-rdd/spark-pair-rdd-functions/)

Spark RDD – Repartition and Coalesce (https://sparkbyexamples.com/spark/spark-repartition-vs-coalesce/)

Spark RDD – Shuffle Partitions (https://sparkbyexamples.com/spark/spark-shuffle-partitions/)

Spark RDD – Cache vs Persist (https://sparkbyexamples.com/spark/spark-difference-between-cache-and-persist/)

Spark RDD – Persistance Storage Levels (https://sparkbyexamples.com/spark/spark-persistence-storage-levels/)

(https://github.com/sparkbyexamples/spark-examples/blob/master/spark-sql-examples/src/main/scala/com/sparkbyexamples/spark/dataframe/ParquetExample.scala) repository for reference.

Apache Parquet Introduction

Apache Parquet (https://parquet.apache.org/) is a columnar file format that provides optimizations to speed up queries and is a far more efficient file format than CSV or JSON, supported by many data processing systems.

It is compatible with most of the data processing frameworks in the Hadoop (https://en.wikipedia.org/wiki/Hadoop) echo systems. It provides efficient data

[Spark RDD – Broadcast Variables (https://sparkbyexamples.com/spark/spark-broadcast-variables/)](https://sparkbyexamples.com/spark/spark-broadcast-variables/)

[Spark RDD – Accumulator Variables (https://sparkbyexamples.com/spark/spark-accumulators/)](https://sparkbyexamples.com/spark/spark-accumulators/)

[Spark RDD – Convert RDD to DataFrame (https://sparkbyexamples.com/apache-spark-rdd/convert-spark-rdd-to-dataframe-dataset/)](https://sparkbyexamples.com/apache-spark-rdd/convert-spark-rdd-to-dataframe-dataset/)

Spark SQL Tutorial

[Spark SQL – Create DataFrame (https://sparkbyexamples.com/spark/different-ways-to-create-a-spark-dataframe/)](https://sparkbyexamples.com/spark/different-ways-to-create-a-spark-dataframe/)

[Spark SQL – Select Columns (https://sparkbyexamples.com/spark/spark-select-columns-from-dataframe/)](https://sparkbyexamples.com/spark/spark-select-columns-from-dataframe/)

[Spark SQL – Add and Update Column (withColumn) (https://sparkbyexamples.com/spark/spark-dataframe-withcolumn/)](https://sparkbyexamples.com/spark/spark-dataframe-withcolumn/)

[Spark SQL – Rename Nested Column (https://sparkbyexamples.com/spark/rename-a-column-on-spark-dataframes/)](https://sparkbyexamples.com/spark/rename-a-column-on-spark-dataframes/)

[Spark SQL – Drop column (https://sparkbyexamples.com/spark/spark-drop-column-from-dataframe-dataset/)](https://sparkbyexamples.com/spark/spark-drop-column-from-dataframe-dataset/)

[Spark SQL – Where | Filter (https://sparkbyexamples.com/spark/spark-dataframe-where-filter/)](https://sparkbyexamples.com/spark/spark-dataframe-where-filter/)

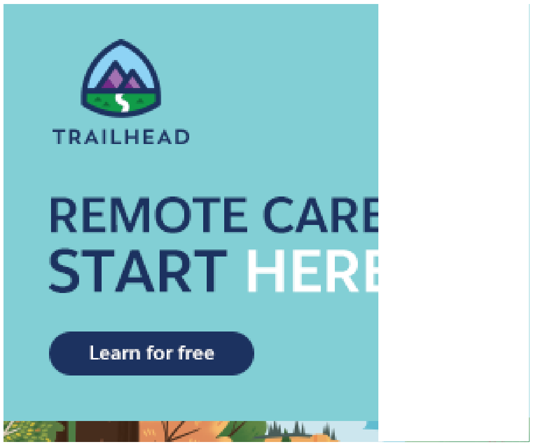
[Spark SQL – When Otherwise (https://sparkbyexamples.com/spark/spark-case-when-otherwise-example/)](https://sparkbyexamples.com/spark/spark-case-when-otherwise-example/)

compression and encoding schemes with enhanced performance to handle complex data in bulk.

Spark SQL provides support for both reading and writing Parquet files that automatically capture the schema of the original data, It also reduces data storage by 75% on average. Below are some advantages of storing data in a parquet format. Spark by default supports Parquet in its library hence we don’t need to add any dependency libraries.

Apache Parquet Advantages:

Below are some of the advantages of using Apache Parquet. combining these benefits with Spark improves performance and gives the ability to work with structure files.



- **Reduces IO operations.**
- **Fetches specific columns that you need to access.**
- **It consumes less space.**
- **Support type-specific encoding.**

Apache Parquet Spark Example

[Spark SQL – Collect data to Driver (https://sparkbyexamples.com/spark/spark-dataframe-collect/)](https://sparkbyexamples.com/spark/spark-dataframe-collect/)

[Spark SQL – Distinct (https://sparkbyexamples.com/spark/spark-remove-duplicate-rows/)](https://sparkbyexamples.com/spark/spark-remove-duplicate-rows/)

[Spark SQL- Pivot Table DataFrame (https://sparkbyexamples.com/spark/how-to-pivot-table-and-unpivot-a-spark-dataframe/)](https://sparkbyexamples.com/spark/how-to-pivot-table-and-unpivot-a-spark-dataframe/)

[Spark SQL – Data Types (https://sparkbyexamples.com/spark/spark-sql-dataframe-data-types/)](https://sparkbyexamples.com/spark/spark-sql-dataframe-data-types/)

[Spark SQL – StructType | StructField (https://sparkbyexamples.com/spark/spark-sql-structtype-on-dataframe/)](https://sparkbyexamples.com/spark/spark-sql-structtype-on-dataframe/)

[Spark SQL – Schema (https://sparkbyexamples.com/spark/spark-schema-explained-with-examples/)](https://sparkbyexamples.com/spark/spark-schema-explained-with-examples/)

[Spark SQL – Groupby (https://sparkbyexamples.com/spark/using-groupby-on-dataframe/)](https://sparkbyexamples.com/spark/using-groupby-on-dataframe/)

[Spark SQL – Sort DataFrame (https://sparkbyexamples.com/spark/spark-how-to-sort-dataframe-column-explained/)](https://sparkbyexamples.com/spark/spark-how-to-sort-dataframe-column-explained/)

[Spark SQL – Join Types (https://sparkbyexamples.com/spark/spark-sql-dataframe-join/)](https://sparkbyexamples.com/spark/spark-sql-dataframe-join/)

[Spark SQL – Union and UnionAll (https://sparkbyexamples.com/spark/spark-dataframe-union-and-union-all/)](https://sparkbyexamples.com/spark/spark-dataframe-union-and-union-all/)

[Spark SQL – map() vs mapPartitions() (https://sparkbyexamples.com/spark/spark-map-vs-mappartitions-transformation/)](https://sparkbyexamples.com/spark/spark-map-vs-mappartitions-transformation/)

Before we go over the Apache parquet with the Spark example, first, let's [Create a Spark DataFrame (https://sparkbyexamples.com/spark/different-ways-to-create-a-spark-dataframe/)](https://sparkbyexamples.com/spark/different-ways-to-create-a-spark-dataframe/) from `Seq` object. Note that [toDF() (https://sparkbyexamples.com/spark/different-ways-to-create-a-spark-dataframe/#from-collection)](https://sparkbyexamples.com/spark/different-ways-to-create-a-spark-dataframe/#from-collection) function on sequence object is available only when you import implicits using `spark.sqlContext.implicits._`. This complete spark parquet example is available at [Github (https://github.com/sparkbyexamples/spark-examples/blob/master/spark-sql-examples/src/main/scala/com/sparkbyexamples/spark/dataframe/ParquetExample.scala)](https://github.com/sparkbyexamples/spark-examples/blob/master/spark-sql-examples/src/main/scala/com/sparkbyexamples/spark/dataframe/ParquetExample.scala) repository for reference.

```
val data = Seq(("James ","","Smi
               ("Michael ","Rose"
               ("Robert ","","Wil
               ("Maria ","Anne","
               ("Jen","Mary","Bro

val columns = Seq("firstname","m

import spark.sqlContext.implicit
val df = data.toDF(columns:_*)
```

The above example creates a data frame with columns “firstname”, “middlename”, “lastname”, “dob”, “gender”, “salary”

Spark Write DataFrame to Parquet file format

Using `parquet()` function of `DataFrameWriter` class, we can write Spark DataFrame to the Parquet file. As mentioned earlier Spark doesn't need any additional packages or libraries to use Parquet as it by default provides with Spark. easy isn't it? so we don't

[Spark SQL – foreach() vs foreachPartition() (https://sparkbyexamples.com/spark/spark-foreachpartition-vs-foreach-explained/)](https://sparkbyexamples.com/spark/spark-foreachpartition-vs-foreach-explained/)

[Spark SQL – map() vs flatMap() (https://sparkbyexamples.com/spark/spark-map-vs-flatmap-with-examples/)](https://sparkbyexamples.com/spark/spark-map-vs-flatmap-with-examples/)

[Spark SQL – Persist and Cache (https://sparkbyexamples.com/spark/spark-dataframe-cache-and-persist-explained/)](https://sparkbyexamples.com/spark/spark-dataframe-cache-and-persist-explained/)

[Spark SQL – UDF (User Defined Functions) (https://sparkbyexamples.com/spark/spark-sql-udf/)](https://sparkbyexamples.com/spark/spark-sql-udf/)

[Spark SQL – Array (ArrayType) Column (https://sparkbyexamples.com/spark/spark-array-arraytype-dataframe-column/)](https://sparkbyexamples.com/spark/spark-array-arraytype-dataframe-column/)

[Spark SQL – Map (MapType) column (https://sparkbyexamples.com/spark/spark-dataframe-map-maptype-column/)](https://sparkbyexamples.com/spark/spark-dataframe-map-maptype-column/)

[Spark SQL – Flatten Nested Struct Column (https://sparkbyexamples.com/spark/spark-flatten-nested-struct-column/)](https://sparkbyexamples.com/spark/spark-flatten-nested-struct-column/)

[Spark SQL – Flatten Nested Array Column (https://sparkbyexamples.com/spark/spark-flatten-nested-array-column-to-single-column/)](https://sparkbyexamples.com/spark/spark-flatten-nested-array-column-to-single-column/)

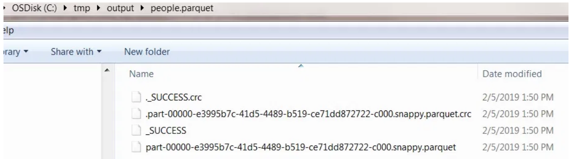
[Spark SQL – Explode Array & Map Columns (https://sparkbyexamples.com/spark/explode-spark-array-and-map-dataframe-column/)](https://sparkbyexamples.com/spark/explode-spark-array-and-map-dataframe-column/)

[Spark SQL – Sampling (https://sparkbyexamples.com/spark/spark-sampling-with-examples/)](https://sparkbyexamples.com/spark/spark-sampling-with-examples/)

have to worry about version and compatibility issues. In this example, we are writing DataFrame to “people.parquet” file.

```
df.write.parquet("/tmp/output/pe
```

Writing Spark DataFrame to Parquet format preserves the column names and data types, and all columns are automatically converted to be nullable for compatibility reasons. Notice that all part files Spark creates has parquet extension.



Name	Date modified
._SUCCESS.crc	2/5/2019 1:50 PM
.part-00000-e3995b7c-41d5-4489-b519-ce71dd872722-c000.snappy.parquet.crc	2/5/2019 1:50 PM
_SUCCESS	2/5/2019 1:50 PM
part-00000-e3995b7c-41d5-4489-b519-ce71dd872722-c000.snappy.parquet	2/5/2019 1:50 PM

Spark Read Parquet file into DataFrame

Similar to write, DataFrameReader provides parquet() function (spark.read.parquet) to read the parquet files and creates a Spark DataFrame. In this example snippet, we are reading data from an apache parquet file we have written before.

[Spark SQL – Partitioning (https://sparkbyexamples.com/spark/spark-partitioning-understanding/)]

Spark SQL Functions

[Spark SQL String Functions (https://sparkbyexamples.com/spark/usage-of-spark-sql-string-functions/)]

[Spark SQL Date and Timestamp Functions (https://sparkbyexamples.com/spark/spark-sql-date-and-time-functions/)]

[Spark SQL Array Functions (https://sparkbyexamples.com/spark/spark-sql-array-functions/)]

[Spark SQL Map Functions (https://sparkbyexamples.com/spark/spark-sql-map-functions/)]

[Spark SQL Sort Functions (https://sparkbyexamples.com/spark/spark-sql-sort-functions/)]

[Spark SQL Aggregate Functions (https://sparkbyexamples.com/spark/spark-sql-aggregate-functions/)]

[Spark SQL Window Functions (https://sparkbyexamples.com/spark/spark-sql-window-functions/)]

[Spark SQL JSON Functions (https://sparkbyexamples.com/spark/spark-most-used-json-functions-with-examples/)]

Spark Data Source API

[Spark – Read & Write CSV file (https://sparkbyexamples.com/spark/spark-read-csv-file-into-dataframe/)]

[Spark – Read and Write JSON file (https://sparkbyexamples.com/

```
val parqDF = spark.read.parquet(
```

printing schema of DataFrame returns columns with the same names and data types.

Append to existing Parquet file

Spark provides the capability to append DataFrame to existing parquet files using “append” save mode. In case, if you want to overwrite use “overwrite” save mode.

```
df.write.mode('append').parquet(
```

Using SQL queries on Parquet

We can also create a temporary view on Parquet files and then use it in Spark SQL statements. This temporary table would be available until the SparkContext present.

```
parqDF.createOrReplaceTempView("
val parkSQL = spark.sql("select
```

Above predicate on spark parquet file does the file scan which is performance bottleneck like table scan on a traditional database. We should use partitioning in order to improve performance.

spark/spark-read-and-write-json-file/)

Spark – Read & Write Parquet file (https://sparkbyexamples.com/spark/spark-read-write-dataframe-parquet-example/)

Spark – Read & Write XML file (https://sparkbyexamples.com/spark/spark-read-write-xml/)

Spark – Read & Write Avro files (https://sparkbyexamples.com/spark/read-write-avro-file-spark-dataframe/)

Spark – Read & Write Avro files (Spark version 2.3.x or earlier) (https://sparkbyexamples.com/spark/using-avro-data-files-from-spark-sql-2-3-x/)

Spark – Read & Write HBase using “hbase-spark” Connector (https://sparkbyexamples.com/spark/spark-read-write-using-hbase-spark-connector/)

Spark – Read & Write from HBase using Hortonworks (https://sparkbyexamples.com/spark/create-spark-dataframe-from-hbase-using-hortonworks/)

Spark – Read & Write ORC file (https://sparkbyexamples.com/spark/spark-read-orc-file-into-dataframe/)

Spark – Read Binary File (https://sparkbyexamples.com/spark/spark-read-binary-file-into-dataframe/)

Spark Streaming & Kafka

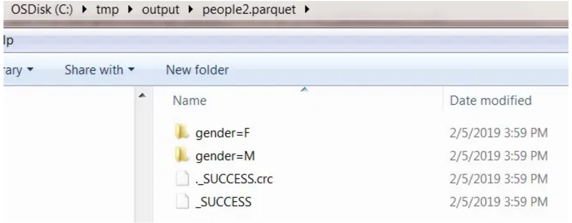
Spark Streaming – OutputModes (https://sparkbyexamples.com/spark/spark-streaming-outputmode/)

Spark parquet partition – Improving performance

Partitioning is a feature of many databases and data processing frameworks and it is key to make jobs work at scale. We can do a parquet file partition using spark `partitionBy()` function.

```
df.write.partitionBy("gender","s
        .parquet("/tmp/output/pe
```

Parquet Partition creates a folder hierarchy for each spark partition; we have mentioned the first partition as gender followed by salary hence, it creates a salary folder inside the gender folder.



Name	Date modified
gender=F	2/5/2019 3:59 PM
gender=M	2/5/2019 3:59 PM
._SUCCESS.crc	2/5/2019 3:59 PM
_SUCCESS	2/5/2019 3:59 PM

This is an example of how to write a Spark DataFrame by preserving the partitioning on gender and salary columns.

```
val parqDF = spark.read.parquet(
parqDF.createOrReplaceTempView("
val df = spark.sql("select * fro
```

The execution of this query is significantly faster than the query without partition (https://sparkbyexamples.com/spark/spark-performance-tuning/). It filters the data first on gender and then applies filters on salary.

Spark Streaming – Reading Files From Directory (https://sparkbyexamples.com/spark/spark-streaming-read-json-files-from-directory/)

Spark Streaming – Reading Data From TCP Socket (https://sparkbyexamples.com/spark/spark-streaming-from-tcp-socket/)

Spark Streaming – Processing Kafka Messages in JSON Format (https://sparkbyexamples.com/spark/spark-streaming-with-kafka/)

Spark Streaming – Processing Kafka messages in AVRO Format (https://sparkbyexamples.com/spark/spark-streaming-consume-and-produce-kafka-messages-in-avro-format/)

Spark SQL Batch – Consume & Produce Kafka Message (https://sparkbyexamples.com/spark/spark-batch-processing-produce-consume-kafka-topic/)

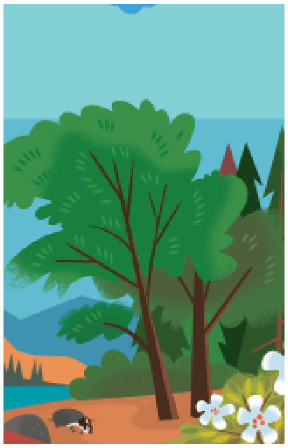
Spark Read a specific Parquet partition

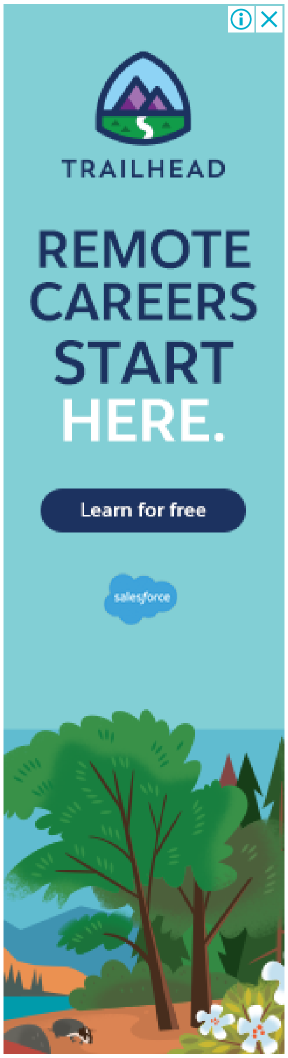
```
val parqDF = spark.read.parquet(
```

This code snippet retrieves the data from the gender partition value “M”.

The complete code can be downloaded from GitHub (https://github.com/sparkbyexamples/spark-examples/blob/master/spark-sql-examples/src/main/scala/com/sparkbyexamples/spark/dataframe/ParquetExample.scala)

Complete Spark Parquet Example





```
package com.sparkbyexamples.spar

import org.apache.spark.sql.Spar

object ParquetExample {

  def main(args:Array[String]):U

    val spark: SparkSession = Sp
      .master("local[1]")
      .appName("SparkByExamples.
      .getOrCreate()

    val data = Seq(("James ","",
                 ("Michael ","Ro
                 ("Robert ","","
                 ("Maria ","Anne
                 ("Jen","Mary","

    val columns = Seq("firstname
    import spark.sqlContext.impl
    val df = data.toDF(columns:_
    df.show()
    df.printSchema()
    df.write
      .parquet("/tmp/output/peop
    val parqDF = spark.read.parq
    parqDF.createOrReplaceTempVi
    spark.sql("select * from Par
    val parkSQL = spark.sql("sel
    parkSQL.show()
    parkSQL.printSchema()
    df.write
      .partitionBy("gender","sal
      .parquet("/tmp/output/peop
    val parqDF2 = spark.read.par
    parqDF2.createOrReplaceTempV
    val df3 = spark.sql("select
    df3.explain()
    df3.printSchema()
    df3.show()
    val parqDF3 = spark.read
      .parquet("/tmp/output/peop
    parqDF3.show()
  }
}
```

Conclusion:

You have learned how to read a write an apache parquet data files in Spark and also learned how to improve the performance (https://sparkbyexamples.com/spark/spark-performance-tuning/) by using partition and filtering data with a partition key and finally appending to and overwriting existing parquet files.

Happy Learning !! 🙂

Share this:

(https://sparkbyexamples.com/spark/spark-read-write-dataframe-parquet-example/?share=facebook&nb=1)

(https://sparkbyexamples.com/spark/spark-read-write-dataframe-parquet-example/?share=reddit&nb=1)

(https://sparkbyexamples.com/spark/spark-read-write-dataframe-parquet-example/?share=pinterest&nb=1)

3

(https://sparkbyexamples.com/spark/spark-read-write-dataframe-parquet-example/?share=tumblr&nb=1)

(https://sparkbyexamples.com/spark/spark-read-write-dataframe-parquet-example/?share=pocket&nb=1)

(https://sparkbyexamples.com/spark/spark-read-write-dataframe-parquet-example/?share=linkedin&nb=1)

(https://sparkbyexamples.com/spark/spark-read-write-dataframe-parquet-example/?share=twitter&nb=1)

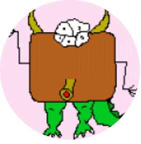
TAGS: APACHE PARQUET (HTTPS://SPARKBYEXAMPLES.COM/TAG/APACHE-PARQUET/), APACHE PARQUET SPARK (HTTPS://SPARKBYEXAMPLES.COM/TAG/APACHE-PARQUET-SPARK/), SPARK READ PARQUET (HTTPS://SPARKBYEXAMPLES.COM/TAG/SPARK-READ-PARQUET/), SPARK WRITE PARQUET (HTTPS://SPARKBYEXAMPLES.COM/TAG/SPARK-WRITE-PARQUET/)

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

NNK (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/)

THIS POST HAS 5 COMMENTS



Anonymous

3 MAR 2021 REPLY

Great Work.

NNK 3 MAR 2021 REPLY

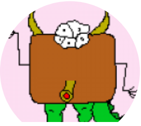
Glad you like it.

Kass 2 JAN 2021 REPLY

Nice work, thanks! Perhaps good to mention that partitioning is supported in various data formats (csv, json etc) and not just parquet.

NNK 2 JAN 2021 REPLY

Hi Kass, Thanks for your feedback. Certainly, I will update the article with your suggestion.



Pickard 14 NOV 2019 REPLY

Great tutorial, thank you!

Leave a Reply

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/)

Recent Posts

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/)

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 Kafka (https://sparkbyexamples.com/category/kafka/)

Apache HBase (https://sparkbyexamples.com/category/hbase/)

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 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/)

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://twitter.com/sparkbyexamples)

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

(https://www.linkedin.com/in/n-nk-b860a8193/)

(https://github.com/spark-examples/)