PySpark Tutorial

PySpark Tutorial For Beginners (https://sparkbyexamples.com/pyspark-tutorial/)

PySpark – Features (https://sparkbyexamples.com/pyspark-tutorial/#features)

PySpark – Advantages (https://sparkbyexamples.com/pyspark-tutorial/#advantages)

PySpark – Modules & Packages (https://sparkbyexamples.com/pyspark-tutorial/#modules-packages)

PySpark – Cluster Managers (https://sparkbyexamples.com/pyspark-tutorial/#cluster-manager)

PySpark – Install on Windows (https://sparkbyexamples.com/pyspark-tutorial/#pyspark-installation)

PySpark – Web/Application UI (https://sparkbyexamples.com/spark/spark-web-ui-understanding/)

PySpark – SparkSession (https://sparkbyexamples.com/pyspark/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/)

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

PySpark Read and Write Parquet File

Prabha (https://sparkbyexamples.com/author/prabha/) - PySpark (https://sparkbyexamples.com/category/pyspark/)

Pyspark SQL provides methods to read Parquet file into DataFrame and write DataFrame to Parquet files, `parquet()` function from `DataFrameReader` and `DataFrameWriter` are used to read from and write/create a Parquet file respectively. Parquet files maintain the schema along with the data hence it is used to process a structured file.

In this article, I will explain how to read from and write a parquet file and also will explain how to partition the data and retrieve the partitioned data with the help of SQL.

PySpark – Accumulator (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/)

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/)

PySpark – Convert 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/)

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

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

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

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

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

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

Below are the simple statements on how to write and read parquet files in PySpark which I will explain in detail later sections.

```
df.write.parquet("/tmp/out/peop
parDF1=spark.read.parquet("/tem
```

Before, I explain in detail, first let's understand What is Parquet file and its advantages over CSV, JSON and other text file formats.

What is Parquet File?

Apache Parquet (https://parquet.apache.org/) file is a columnar storage format available to any project in the Hadoop ecosystem, regardless of the choice of data processing framework, data model, or programming language.

Advantages:

While querying columnar storage, it skips the nonrelevant data very quickly, making faster query execution. As a result aggregation queries consume less time compared to row-oriented databases.

It is able to support advanced nested data structures (https://sparkbyexamples.com/pyspark/pyspark-structtype-and-structfield/).

Parquet supports efficient compression options and encoding schemes.

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()

Pyspark 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. Pyspark by default supports Parquet in its library hence we don't need to add any dependency libraries.

Apache Parquet Pyspark Example

Since we don't have the parquet file, let's work with writing parquet from a DataFrame. First, create a Pyspark DataFrame from a list of data using `spark.createDataFrame()` method.

```
data =[("James ","","Smith","36
              ("Michael ","Rose
              ("Robert ","","Wi
              ("Maria ","Anne",
              ("Jen","Mary","Br
columns=["firstname","middlenam
df=spark.createDataFrame(data,c
```

above example, it creates a DataFrame with columns firstname, middlename, lastname, dob, gender, salary.

Pyspark Write DataFrame to Parquet file format

Now let's create a parquet file from PySpark DataFrame by calling the `parquet()` function of `DataFrameWriter` class. When you write a DataFrame to parquet file, it

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

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

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

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

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

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

PySpark SQL Functions

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

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/)

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

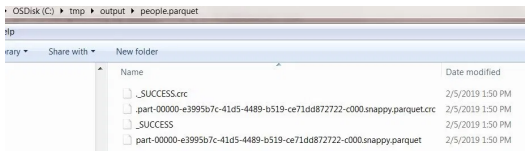
PySpark Datasources

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

PySpark – Read & Write Parquet File (https://sparkbyexamples.com/

automatically preserves column names and their data types. Each part file Pyspark creates has the .parquet file extension. Below is the example,

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



Pyspark Read Parquet file into DataFrame

Pyspark provides a `parquet()` method in `DataFrameReader` class to read the parquet file into dataframe. Below is an example of a reading parquet file to data frame.

```
parDF=spark.read.parquet("/tmp/
```

Append or Overwrite an existing Parquet file

Using append save mode, you can append a dataframe to an existing parquet file. Incase to overwrite use overwrite save mode.

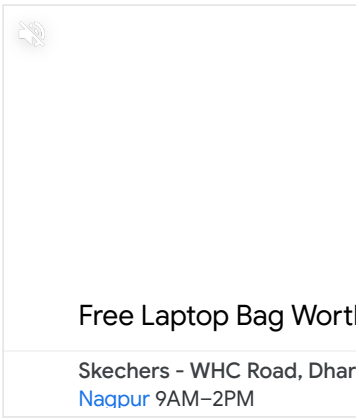
```
df.write.mode('append').parquet
df.write.mode('overwrite').parq
```

Executing SQL queries DataFrame

Pyspark Sql provides to create temporary views on parquet files for executing sql queries. These views are available until your program exists.

pyspark/pyspark-read-and-write-parquet-file/)

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




```
parqDF.createOrReplaceTempView(
parkSQL = spark.sql("select * f
```

Creating a table on rquet file

/ let's walk through executing SQL ries on parquet file. In order to cute sql queries, create a temporary / or table directly on the parquet file ead of creating from DataFrame.

```
spark.sql("CREATE TEMPORARY VIE
spark.sql("SELECT * FROM PERSON
```

e, we created a temporary view SON from "people.parquet" file. gives the following results.

```
---------+----------+--------+
Firstname|middlename|lastname|
---------+----------+--------+
  Robert |          |Williams|
   Maria |      Anne|   Jones|
 Michael |      Rose|        |
   James |          |   Smith|
     Jen|      Mary|   Brown|
---------+----------+--------+
```

eate Parquet partition e

en we execute a particular query on PERSON table, it scan's through all the rows and returns the results back. This is similar to the traditional database query execution. In PySpark, we can improve query execution in an optimized way by doing partitions on the data using pyspark `partitionBy()` (https://sparkbyexamples.com/pyspark/pyspark-partitionby-example/) method. Following is the example of partitionBy().

PySpark Built-In Functions

PySpark – when() (https://sparkbyexamples.com/pyspark/pyspark-when-otherwise/)

PySpark – expr() (https://sparkbyexamples.com/pyspark/pyspark-sql-expr-expression-function/)

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)

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/)

PySpark – to_date() (https://sparkbyexamples.com/pyspark/pyspark-to_date-convert-timestamp-to-date/)

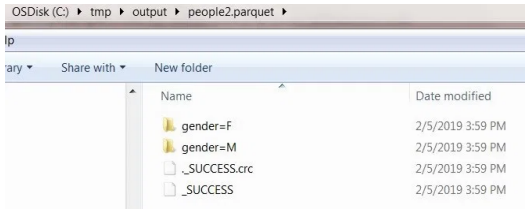
PySpark – date_format() (https://sparkbyexamples.com/pyspark/pyspark-date_format-convert-date-to-string-format/)

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())

```
df.write.partitionBy("gender","
```

When you check the people2.parquet file, it has two partitions “gender” followed by “salary” inside.



Retrieving from a partitioned Parquet file

The example below explains of reading partitioned parquet file into DataFrame with gender=M.

```
parDF2=spark.read.parquet("/tmp
parDF2.show(truncate=False)
```

Output for the above example is shown below.

```
+---------+----------+--------+
|firstname|middlename|lastname|
+---------+----------+--------+
|Robert   |          |Williams|
|Michael  |Rose      |        |
|James    |          |Smith   |
+---------+----------+--------+
```

Creating a table on Partitioned Parquet file

Here, I am creating a table on partitioned parquet file and executing a query that executes faster than the table without partition, hence improving the performance.

PySpark – explode() (https://sparkbyexamples.com/pyspark/pyspark-explode-nested-array-into-rows/)

PySpark – array_contains() (https://sparkbyexamples.com/pyspark/pyspark-arraytype-column-with-examples/#array_contains)

PySpark – array() (https://sparkbyexamples.com/pyspark/pyspark-arraytype-column-with-examples/#array)

PySpark – collect_list() (https://sparkbyexamples.com/pyspark/pyspark-aggregate-functions/#collect-list)

PySpark – collect_set() (https://sparkbyexamples.com/pyspark/pyspark-aggregate-functions/#collect-set)

PySpark – create_map() (https://sparkbyexamples.com/pyspark/pyspark-convert-dataframe-columns-to-maptype-dict/)

PySpark – map_keys() (https://sparkbyexamples.com/pyspark/pyspark-maptype-dict-examples/#map_keys)

PySpark – map_values() (https://sparkbyexamples.com/pyspark/pyspark-maptype-dict-examples/#map_values)

PySpark – struct() (https://sparkbyexamples.com/pyspark/pyspark-structtype-and-structfield/#update-struct-function)

PySpark – countDistinct() (https://sparkbyexamples.com/pyspark/pyspark-count-distinct-from-dataframe/)

PySpark – sum(), avg() (https://sparkbyexamples.com/pyspark/pyspark-dataframe-groupby-and-sort-by-descending-order/)

PySpark – row_number() (https://sparkbyexamples.com/pyspark/pyspark-window-functions/#row_number)

PySpark – rank() (https://sparkbyexamples.com/pyspark/pyspark-window-functions/#rank)

```
spark.sql("CREATE TEMPORARY VIE
spark.sql("SELECT * FROM PERSON
```

Below is the output .

```
+---------+----------+--------+
|firstname|middlename|lastname|
+---------+----------+--------+
|   Maria |      Anne|   Jones|
|      Jen|      Mary|   Brown|
+---------+----------+--------+
```

Complete Example of PySpark read and write Parquet file

```
import pyspark
from pyspark.sql import SparkSe
spark=SparkSession.builder.appN
data =[("James ","","Smith","36
             ("Michael ","Rose
             ("Robert ","","Wi
             ("Maria ","Anne",
             ("Jen","Mary","Br
columns=["firstname","middlenam
df=spark.createDataFrame(data,c
df.write.mode("overwrite").parq
parDF1=spark.read.parquet("/tmp
parDF1.createOrReplaceTempView(
parDF1.printSchema()
parDF1.show(truncate=False)

parkSQL = spark.sql("select * f
parkSQL.show(truncate=False)

spark.sql("CREATE TEMPORARY VIE
spark.sql("SELECT * FROM PERSON

df.write.partitionBy("gender","

parDF2=spark.read.parquet("/tmp
parDF2.show(truncate=False)

spark.sql("CREATE TEMPORARY VIE
spark.sql("SELECT * FROM PERSON
```

PySpark – dense_rank() (https://sparkbyexamples.com/pyspark/pyspark-window-functions/#dense_rank)

PySpark – percent_rank() (https://sparkbyexamples.com/pyspark/pyspark-window-functions/#percent_rank)

PySpark – 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)

PySpark – 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)

PySpark – 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)

Conclusion:

We have learned how to write a Parquet file from a PySpark DataFrame and reading parquet file to DataFrame and created view/tables to execute SQL queries. Also explained how to do partitions on parquet files to improve performance (https://sparkbyexamples.com/spark/spark-performance-tuning/).

Hope you liked it and, do comment in the comment section.

Share this:

(https://sparkbyexamples.com/pyspark/pyspark-read-and-write-parquet-file/?share=facebook&nb=1)

(https://sparkbyexamples.com/pyspark/pyspark-read-and-write-parquet-file/?share=reddit&nb=1)

(https://sparkbyexamples.com/pyspark/pyspark-read-and-write-parquet-file/?share=pinterest&nb=1)

(https://sparkbyexamples.com/pyspark/pyspark-read-and-write-parquet-file/?share=tumblr&nb=1)

(https://sparkbyexamples.com/pyspark/pyspark-read-and-write-parquet-file/?share=pocket&nb=1)

(https://sparkbyexamples.com/pyspark/pyspark-read-and-write-parquet-file/?share=linkedin&nb=1)

(https://sparkbyexamples.com/pyspark/pyspark-read-and-write-parquet-file/?share=twitter&nb=1)

TAGS: PARQUET (HTTPS://SPARKBYEXAMPLES.COM/TAG/PARQUET/), PARQUET PERFORMANCE IMPROVEMENT (HTTPS://SPARKBYEXAMPLES.COM/TAG/PARQUET-PERFORMANCE-IMPROVEMENT/), READ PARQUET TO DATAFRAME (HTTPS://SPARKBYEXAMPLES.COM/TAG/READ-PARQUET-TO-DATAFRAME/), WRITE DATAFRAME TO PARQUET (HTTPS://SPARKBYEXAMPLES.COM/TAG/WRITE-DATAFRAME-TO-PARQUET/)

THIS POST HAS 0 COMMENTS

Raymond

25 APR 2021 REPLY

Great post/site

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/)

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/)

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/)

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/)

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

(https://www.facebook.com/sparkbyexamples) (https://www.linkedin.com/in/n-k-b860a8193/) (https://twitter.com/sparkbyexamples) (https://github.com/spark-examples/)

Copyright sparkbyexamples.com