What is the difference between two dates in terms of days, months, and years in PySpark?

Authored by

stats writer

June 24, 2024

The difference between two dates in terms of days, months, and years in PySpark refers to the number of days, months, and years that have elapsed between two given dates. This can be calculated using the PySpark function "datediff", which calculates the difference between two dates in terms of days, or the "months_between" function, which calculates the difference in terms of months. The result of these functions can vary depending on the leap years and the length of the months in a given year. Additionally, the PySpark function "year" can be used to calculate the difference in terms of years. Overall, the difference between two dates in PySpark is a useful tool for analyzing time-based data and understanding the time intervals between events.

Using PySpark SQL functions `datediff()`, `months_between()`, you can calculate the difference between two dates in days, months, and years. Let's see this by using a DataFrame example. You can also use these to calculate age.

Get Differences Between Dates in Days

The `datediff()` is a PySpark SQL function that is used to calculate the difference in days between two provided dates. datediff() is commonly used in SQL queries or DataFrame operations to compute the duration between two timestamps or date values. In the example below, I will calculate the differences between the date column and the current date. To get the current date, I will use the `current_date()` function.

```
# Imports
from pyspark.sql import SparkSession
from pyspark.sql.functions import col, current_date, datediff

# Create SparkSession
spark = SparkSession.builder.appName('SparkByExamples.com').getOrCreate()

# Create DataFrame
data =
df=spark.createDataFrame(data=data,schema=)

# Calculate the difference between two dates
df2 = df.select(
col("date"),
current_date().alias("current_date"),
datediff(current_date(),col("date")).alias("datediff")
)
df2.show()
```

This code snippet selects the "`date`" column from the DataFrame and computes the difference in days between the current date and the values in the "date" column. Finally, it renames the resulting columns appropriately and assigns the resulting DataFrame to the variable `df2`.

```
# Output:
+----------+------------+--------+
| date|current_date|datediff|
+----------+------------+--------+
|2019-07-01| 2021-02-26| 606|
|2019-06-24| 2021-02-26| 613|
|2019-08-24| 2021-02-26| 552|
+----------+------------+--------+
```

Get Differences Between Dates in Months

Use the PySpark SQL `months_between()` function to get the number of months between two dates. The below code snippet calculates month differences between the date column from the DataFrame and the current date, including differences in days, and months.

```
# Imports
from pyspark.sql.functions import col, current_date, datediff, months_between,
round
```

```
# Calculate the difference between two dates in months
df3 = df.withColumn("datesDiff", datediff(current_date(), col("date")))
.withColumn("monthsDiff", months_between(current_date(), col("date")))
.withColumn("monthsDiff_round", round(months_between(current_date(), col("date")), 2))

df3.show()
```

This yields the below output. This adds new columns named `datediff`, `monthdDiff`, and `monthsDiff_round` to the DataFrame. The `monthsDiff_round` represents the difference in months between the current date and the dates in the "date" column, rounded to two decimal places.

```
# Output:
+---+----------+---------+-----------+---------------+
| id| date|datesDiff| monthsDiff|monthsDiff_round|
```

```
+---+----------+---------+-----------+---------------+
| 1|2019-07-01| 1730|56.80645161| 56.81|
| 2|2019-06-24| 1737|57.06451613| 57.06|
| 3|2019-08-24| 1676|55.06451613| 55.06|
+---+----------+---------+-----------+---------------+
```

Get Differences Between Dates in Years

To calculate the difference between two dates in years using PySpark, you can utilize the `months_between()` function to get the difference in months and then convert it into years.

```
# Imports
from pyspark.sql.functions import col, current_date, datediff, months_between,
round, lit
```

# Calculate the difference between two dates in years
df4 = df.withColumn("datesDiff", datediff(current_date(), col("date")))
.withColumn("yearsDiff", months_between(current_date(), col("date")) / lit(12))
.withColumn("yearsDiff_round", round(months_between(current_date(), col("date")) / lit(12), 2))
df4.show()

This yields the below output.

```
# Output:
+---+----------+---------+-----------------+---------------+
| id| date|datesDiff| yearsDiff|yearsDiff_round|
+---+----------+---------+-----------------+---------------+
| 1|2019-07-01| 1730|4.733870967500001| 4.73|
| 2|2019-06-24| 1737|4.755376344166667| 4.76|
| 3|2019-08-24| 1676| 4.5887096775| 4.59|
+---+----------+---------+-----------------+---------------+
```

Alternatively, you can also use the `datediff()` function to get the difference in days and then convert it into years.

```
# Using datediff()
result = df.withColumn("years_diff", expr("datediff(end_date, start_date) /
365.25"))
```

Calculating Differences when Dates are in Custom Format

Let's see another example of the difference between two dates when dates are not in DateType format `yyyy-MM-dd`. When dates are not in DateType format, all date functions return null. Hence, you need to first convert the input date to DateType using `to_date()` function and then calculate the differences

```
# Imports
from pyspark.sql.functions import current_date,to_date

# Create DF with custom date formats
data2 =
df2=spark.createDataFrame(data=data2,schema=)

# Convert date and calculate
df2.select(
months_between(to_date(col("date"),"MM-dd-yyyy"),current_date()).alias("monthsdiff")
).show()
```

PySpark SQL Example to get the Difference Between Dates

Let's use the SQL example to calculate the difference between two dates in years. Similarly, you can calculate the days and months between two dates. Here, the spark.sql() is used to run the SQL queries by using the functions explained above.

```
# SQL Example
spark.sql("select  round(months_between('2019-07-01',current_date())/12,2)  as
years_diff")
.show()
```

Conclusion:

In this tutorial, you have learned how to calculate days, months, and years between two dates using PySpark Date and Time functions datediff(), months_between(). You can find more information about these functions at the following blog

Happy Learning !!

Related Articles: