ark SQL COALESCE on DataFrame – mples

Updated on May 8, 2020 by Vithal S

will know the importance of coalesce function if you are from or Data Warehouse background. Coalesce function is one of videly used function in SQL. You can use the coalesce function turn non-null values. In this article, we will check how to use k SQL coalesce on an Apache Spark DataFrame with an nple.

```
mp = testDF.withColumn('newColumn', coalesce(testDF['id'],testDF['number']))
mp.show()
+------+---------+
|number|newColumn|
+------+---------+
|     1|        1|
|     2|        2|
|     3|        3|
|  null|        4|
+------+---------+
```

k SQL COALESCE on DataFrame

coalesce is a non-aggregate regular function in Spark SQL. The esce gives the first non-null value among the given columns ull if all columns are null.

SQL COALESCE on DataFrame Examples

can apply the COALESCE function on DataFrame column values
ou can write your own expression to test conditions.

wing example demonstrates the usage of COALESCE function
e DataFrame columns and create new column. We have used
ark to demonstrate the Spark coalesce function.

Data

vill be using following DataFrame to test Spark SQL COALESCE
tion.

```
DF = sqlContext.createDataFrame([(1,1), (2,2),
,3), (4,None)], ["id", "number"])
```

```
-+------+
d|number|
-+------+
1|     1|
2|     2|
1|     3|
4|  null|
-+------+
```

Required Pyspark Functions

first step, you need to import required functions such as
Column, WHERE, etc.

xample, execute the following command on the pyspark
mand line interface or add it in your Python script.

RECENT POSTS

What are SELECT INTO
Alternatives in Snowflake?

What are Different Methods to
Create Snowflake Tables?

How to Duplicate or Clone SQL
Tables – Methods

can use the coalesce function either on *DataFrame* or in *kSQL query* if you are working on tables.

COALESCE Function on DataFrame

is one of the commonly used method to get non null values.

wing example create new column which contains all non null s.

```
eate new column with non Null values
= testDF.withColumn('newColumn',
esce(testDF['id'], testDF['number']))

eck the content of new df
show()


-+------+---------+
d|number|newColumn|
-+------+---------+
1|     1|        1|
2|     2|        2|
1|     3|        3|
4|  null|        4|
-+------+---------+
```

ou can see, newColumn contains non-null values. Optionally, can [rename this column and drop existing columns](#).

ESCE Function in Spark SQL Queries

xample, consider below example which use coalesce in
es. Note that, we have registered Spark DataFrame as a temp
using registerTempTable method.

```
ntext.sql("select *, coalesce(id, number) as
lumn from temp_table").show()
```

```
-+------+---------+
d|number|newColumn|
-+------+---------+
1|     1|        1|
2|     2|        2|
1|     3|        3|
4|  null|        4|
-+------+---------+
```

ed Articles,

Spark SQL Create Temporary Tables, Syntax and Examples
Spark SQL update DataFrame Column

this helps 🙂

of Selena hez	How to Export Spark-SQL…	Spark SQL Analytic…	Spark CASE
iew Mentor	dwgeek.com	dwgeek.com	dwgeek.c

ace park…	Spark SQL Date and…	Spark SQL EXPLAIN…	How t Spark
	dwgeek.com	dwgeek.com	dwgeek.c

DWGEEK.COM

CATEGORIES

Apache Spark

Azure Synapse

BigData

Data Mining

Data Warehouse

General

Greenplum

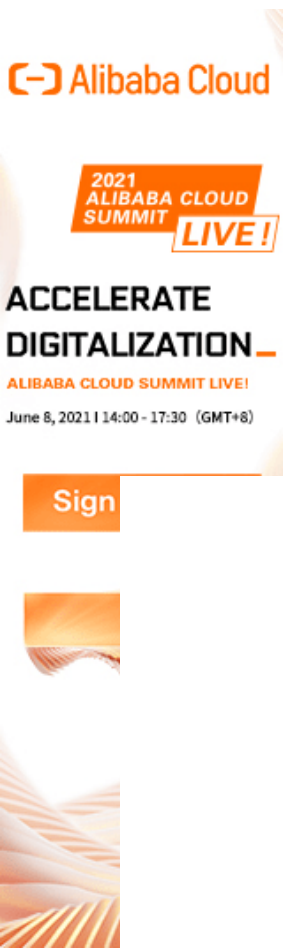Netezza

Redshift

Snowflake

Uncategorized

Vertica

Master Big Data Analytics

E&ICT Academy, IIT Guwahati

Learn Hadoop, Spark with Machine Learning , Python for Data Analytics, Statistic

intellipaa

DWGEEK.COM

THE ONE BEHIND DWGEEK.COM

I'm Vithal, a techie by profession, passionate blogger, frequent traveler, Beer lover and many more.. I write about Big Data, Data Warehouse technologies, Databases, and other general software related stuffs.

More..