How can I use PySpark to read data from a Hive table in SQL?

Authored by

stats writer

June 24, 2024

PySpark is a powerful tool for processing large datasets in a distributed computing environment. It also has the capability to interact with Hive, a data warehouse system built on top of Hadoop. This allows users to seamlessly read data from a Hive table in SQL using PySpark. By establishing a connection to the Hive database, users can use PySpark commands to query and manipulate data in the Hive table. This enables efficient analysis of big data sets without the need for data transfer or data duplication. Additionally, PySpark's integration with Hive allows for seamless integration with other data analysis tools and frameworks.

How to read or query a Hive table into PySpark DataFrame? PySpark SQL supports reading a Hive table to DataFrame in two ways: the `SparkSesseion.read.table()` method and the `SparkSession.sql()` statement.

To read a Hive table, you need to create a SparkSession with enableHiveSupport(). This method is available at `pyspark.sql.SparkSession.builder.enableHiveSupport()` which is used to enable Hive support, including connectivity to a persistent Hive metastore, support for Hive SerDes, and Hive user-defined functions.

Steps to Read Hive Table into PySpark DataFrame

1. Create Spark Session with Hive Enabled

In order to read the hive table into pySpark DataFrame first, you need to create a SparkSession with Hive support enabled. In case you wanted to read from remove hive cluster refer to How to connect Remote Hive Cluster from Spark.

```
from os.path import abspath
from pyspark.sql import SparkSession

#enableHiveSupport() -> enables sparkSession to connect with Hive
warehouse_location = abspath('spark-warehouse')
spark = SparkSession
.builder
.appName(arabpsychology.com)
.config("spark.sql.warehouse.dir", warehouse_location)
.enableHiveSupport()
.getOrCreate()
```

PySpark reads the data from the default Hive warehouse location which is `/user/hive/warehouse` when you use a Hive cluster. But on local, it reads from the current directory. You can change this behavior using the `spark.sql.warehouse.dir` configuration while

creating a `SparkSession` .

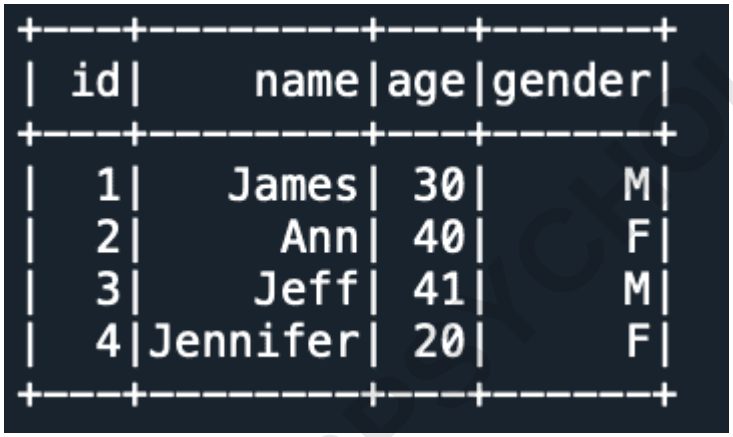
2. PySpark Read or Query Hive Table into DataFrame

In my previous article, I saved a Hive table from PySpark DataFrame, which created Hive files at the default location, which is inside the spark-warehouse directory within the current directory.

Let's read the Hive table into PySpark DataFrame.

```
# Read Hive table
df = spark.sql("select * from emp.employee")
df.show()
```

Yields the below output.



```
+---+--------+---+------+
| id|    name|age|gender|
+---+--------+---+------+
|  1|   James| 30|     M|
|  2|     Ann| 40|     F|
|  3|    Jeff| 41|     M|
|  4|Jennifer| 20|     F|
+---+--------+---+------+
```

3. Using spark.read.table()

Alternatively, you can also read by using spark.read.table() method. here, spark.read is an object of the class DataFrameReader.

```
# Read Hive table
df = spark.read.table("employee")
df.show()
```

4. PySpark Read Hive Table from Remote Hive

```
from os.path import abspath
from pyspark.sql import SparkSession
```

#enableHiveSupport() -> enables sparkSession to connect with Hive
warehouse_location = abspath('spark-warehouse')
spark = SparkSession
.builder
.appName(arabpsychology.com)
.config("spark.sql.warehouse.dir", "/hive/warehouse/dir")
.config("hive.metastore.uris", "thrift://remote-host:9083")
.enableHiveSupport()
.getOrCreate()

# or Use the below approach
# Change using conf
spark.sparkContext().conf().set("spark.sql.warehouse.dir", "/user/hive/warehouse");
spark.sparkContext().conf().set("hive.metastore.uris", "thrift://localhost:9083");

5. Conclusion

In this article, you have learned how to read the Hive table into Spark DataFrame by creating SparkSession with `enableHiveSupport()` and using the dependencies required to connect to the Hive. Also, learned how to read by using `SparkSesseion.read.table()` method and the `SparkSession.sql()`.

You can find the complete working example at GitHub PySpark Hive Example

Related Articles