Reading: User-Defined Schema (UDS) for DSL and SQL

Skills Network

Estimated time needed: 10 minutes

How to Define and Enforce a User-Defined Schema in PySpark?

In this reading, you will learn how to define and enforce a user-defined schema in PySpark.

Spark provides a structured data processing framework that can define and enforce schemas for various data sources, including CSV files. Let's look at the steps to define and use a user-defined schema for a CSV file in PySpark:

Step 1:

Import the required libraries.

```
1. 1
```

```
1. from pyspark.sql.types import StructType, IntegerType, FloatType, StringType, StructField
```

Copied!

Step 2:

Define the schema.

Understanding the data before defining a schema is an important step.

Let's take a look at the step-by-step approach to understanding the data and defining an appropriate schema for a given input file:

1. **Explore the data:** Understand the different data types present in each column.

2. **Column data types:** Determine the appropriate data types for each column based on your observed values.

3. **Define the schema:** Use the 'StructType' class in Spark and create a 'StructField' for each column, mentioning the column name, data type, and other properties.

Example:

```
1. 1
2. 2
3. 3
4. 4
5. 5
6. 6
7. 7
```

```
1. schema = StructType([
```

```
2.      StructField("Emp_Id", StringType(), False),
3.      StructField("Emp_Name", StringType(), False),
4.      StructField("Department", StringType(), False),
5.      StructField("Salary", IntegerType(), False),
6.      StructField("Phone", IntegerType(), True),
7. ])
```

Copied!

'False' indicates null values are **NOT** allowed for the column.

The schema defined above can be utilized for the below CSV file data:

Filename: employee.csv

```
1. 1
2. 2
3. 3
4. 4
```

```
1. emp_id,emp_name,dept,salary,phone
2. A101,jhon,computer science,1000,+1 (701) 846 958
3. A102,Peter,Electronics,2000,
4. A103,Micheal,IT,2500,
```

Copied!

Step 3: Read the input file with user-defined schema.

```
1. 1
2. 2
3. 3
4. 4
5. 5
6. 6
7. 7
8. 8
9. 9
```

```
1. #create a dataframe on top a csv file
2. df = (spark.read
3.    .format("csv")
4.    .schema(schema)
5.    .option("header", "true")
6.    .load("employee.csv")
7. )
8. # display the dataframe content
9. df.show()
```

Copied!

Step 4: Use the `printSchema()` method in Spark to display the schema of a DataFrame and ensure that the schema is applied correctly to the data.

```
1. 1
```

```
1. df.printSchema()
```

Copied!

Through the preceding four steps, you've acquired the ability to establish a schema for a CSV file. Additionally, you've employed this user-defined schema (UDF) to read the CSV file, exhibit its contents, and showcase the schema itself.

Author(s)

- Raghul Ramesh

Changelog

Date	Version	Changed by	Change Description
2023-09-07	0.2	Rahul Rawat	QA pass with edits
2023-09-06	0.1	Gagandeep Singh	Initial version created