Spark Option: inferSchema vs header = true

Asked 1 year, 8 months ago Active 1 year, 8 months ago Viewed 12k times

8

4

Reference to [pyspark: Difference performance for spark.read.format("csv") vs spark.read.csv]

I thought I needed `.options("inferSchema" , "true")` and `.option("header", "true")` to print my headers but apparently I could still print my csv with headers.

What is the difference between header and schema? I don't really understand the meaning of "inferSchema: automatically infers column types. It requires one extra pass over the data and is false by default".

csv apache-spark header apache-spark-sql schema

Share Improve this question Follow

edited Jul 8 '19 at 10:32
Shaido
22.2k 17 54 64

asked Jul 8 '19 at 1:20
user1342124
331 1 3 11

1 Answer

Active | Oldest | Votes

23

The header and schema are separate things.

Header:

If the csv file have a header (column names in the first row) then set `header=true` . This will use the first row in the csv file as the dataframe's column names. Setting `header=false` (default option) will result in a dataframe with default column names: `_c0` , `_c1` , `_c2` , etc.

Setting this to true or false should be based on your input file.

Schema:

The schema refered to here are the column types. A column can be of type String, Double, Long, etc. Using `inferSchema=false` (default option) will give a dataframe where all columns are strings (`StringType`). Depending on what you want to do, strings may not work. For example, if you want to add numbers from different columns, then those columns should be of some numeric type (strings won't work).

By setting `inferSchema=true` , Spark will automatically go through the csv file and infer the schema of each column. This requires an extra pass over the file which will result in reading a file with `inferSchema` set to true being slower. But in return the dataframe will most likely have a correct schema given its input.

As an alternative to reading a csv with `inferSchema` you can provide the schema while reading. This have the

types. In addition, for csv files without a header row, column names can be given automatically. To provde schema see e.g.: Provide schema while reading csv file as a dataframe

Share Improve this answer Follow

answered Jul 8 '19 at 10:30

Shaido
22.2k 17 54 64