spark access first n rows - take vs limit

Asked 7 years ago Modified 1 year, 1 month ago Viewed 169k times

54

I want to access the first 100 rows of a spark data frame and write the result back to a CSV file.

Why is `take(100)` basically instant, whereas

```
df.limit(100)
      .repartition(1)
      .write
      .mode(SaveMode.Overwrite)
      .option("header", true)
      .option("delimiter", ";")
      .csv("myPath")
```

takes forever. I do not want to obtain the first 100 records per partition but just any 100 records.

Why is `take()` so much faster than `limit()`?

apache-spark apache-spark-sql limit

Share Edit Follow

edited Feb 7, 2021 at 17:11
Marioanzas
1,895 ● 2 ● 18 ● 36

asked Oct 19, 2017 at 14:31
Georg Heiler
17.6k ● 40 ● 172 ● 311

1 So, you can use `take(100)` indeed; what is the question? – desertnaut Oct 20, 2017 at 10:39

2 Why take is so much faster than limit. – Georg Heiler Oct 20, 2017 at 11:10

1 Indeed I could, but so far have not seen a way to create a df of the local array to use Sparks nice CSV handling capabilities. Limit should just provide this. – Georg Heiler Oct 20, 2017 at 11:13

6 Answers

Sorted by: Highest score (default)

68

Although it still is answered, I want to share what I learned.

```
myDataFrame.take(10)
```

-> **results in an Array of Rows.** This is an action and performs collecting the data (like collect does).

```
myDataFrame.limit(10)
```

-> **results in a new Dataframe.** This is a transformation and does not perform collecting the data.

I do not have an explanation why then limit takes longer, but this may have been answered above. This is just a basic answer to what the difference is between take and limit.

Share Edit Follow

edited Oct 15, 2020 at 9:23

pfnuesel
15.2k ● 14 ● 60 ● 73

answered Jun 26, 2019 at 12:09

Kaspatoo
1,353 ● 2 ● 13 ● 28

The difference between action and transformation is correct, but that does not explain why limit should take longer than take (once the plan executes). – Arjen P. De Vries Nov 11, 2020 at 8:16

26

This is because predicate pushdown is currently not supported in Spark, see this very good answer.

Actually, take(n) should take a really long time as well. I just tested it, however, and get the same results as you do - take is almost instantaneous irregardless of database size, while limit takes a lot of time.

Share Edit Follow

edited Mar 16, 2018 at 9:45

answered Mar 16, 2018 at 9:14

Thomas
5,064 ● 5 ● 44 ● 74

Collect only works in spark dataframes. When I collect first 100 rows it is instant and data resides in memory as a regular list. Collect in sparks sense is then no longer possible. – Georg Heiler Mar 16, 2018 at 9:35

You are right of course, I forgot take returns a list. I just tested it, and get the same results - I expected both take and limit to be slow. – Thomas Mar 16, 2018 at 9:47

stackoverflow.com/questions/35869884/... <- This question however explicitely states that others have problems with `take()` as well - which version of pyspark are you using? – Thomas Mar 16, 2018 at 9:48

1 Spark scala 2.2 – Georg Heiler Mar 16, 2018 at 11:32

-1

limit work for me:

```
limitDF= df.limit(5)
```

Even better approach is limit your data with filter

```
filterDF= df.filter("name = 'jitendra'").limit(5)
```

Share Edit Follow

answered Oct 4, 2023 at 5:43

Jitendra Bansiwal

113 ● 1 ● 4

-1

Limit() will not work in partition, so it will take more time to execute

Share Edit Follow

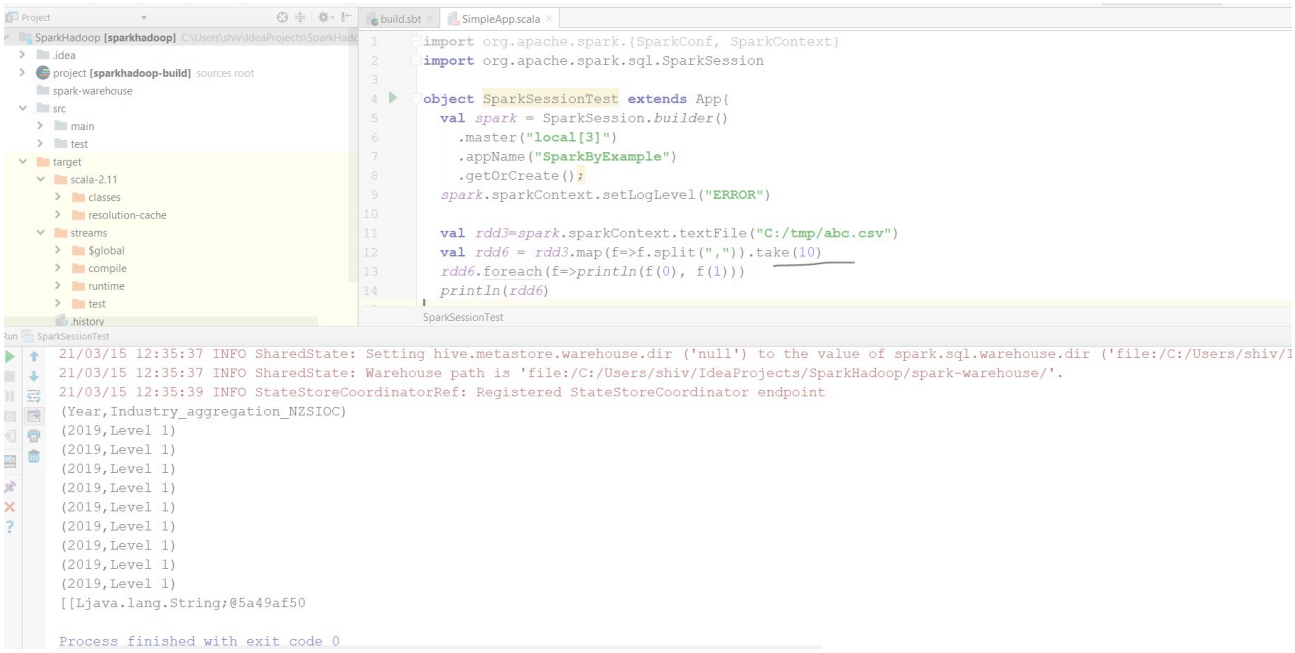
answered Feb 18, 2022 at 14:12

Ratheesh

19 ● 5

-3

You can use take(n) to limit the data. Adding the complete code with output in the screenshot.



Share Edit Follow

answered Mar 15, 2021 at 7:08

Shyam Gupta

503 ● 4 ● 8

-9

`.take()` could be the answer, but I used a simple head command like below

```
df.head(3)
```

`.take()` did not work for me.

Share Edit Follow

edited Jan 7, 2021 at 17:59

10 Rep
2,270 ● 7 ● 20 ● 33

answered Jan 7, 2021 at 12:35

V

Vignesh M21
1

1 Unfortunately, this is not an answer, it's more a feeling ... – Gilles Bodart Oct 18, 2022 at 12:59