

## Working with HIVE data

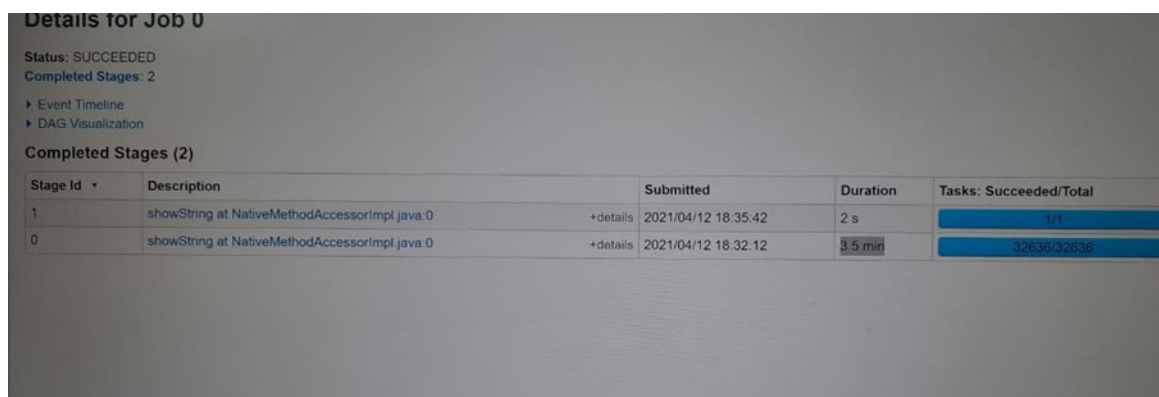
### Read HIVE data

Reading hive table is different from reading raw files. While reading the hive table Spark gets the schema information from hive metastore, so spark do not have to read the input files to infer the schema. However, reading the hive table also can make spark perform worst if the hive table is not well optimized.

Let us go through some examples

**Scenario 1:** - I have a partitioned hive table for sales information. This table is partition by date, and we have 190 days of data. Size of the tables is small, only 500MB.

While reading the table and doing a simple count (\*) in spark I had noticed that it had spawned 32 thousand task. This is too many tasks for such a small table.



**Details for Job 0**

Status: SUCCEEDED  
Completed Stages: 2

▶ Event Timeline  
▶ DAG Visualization

**Completed Stages (2)**

Stage Id	Description	Submitted	Duration	Tasks: Succeeded/Total
1	showString at NativeMethodAccessorImpl.java:0	2021/04/12 18:35:42	2 s	1/1
0	showString at NativeMethodAccessorImpl.java:0	2021/04/12 18:32:12	3.5 min	32636/32636

*Running count on poorly design hive table*

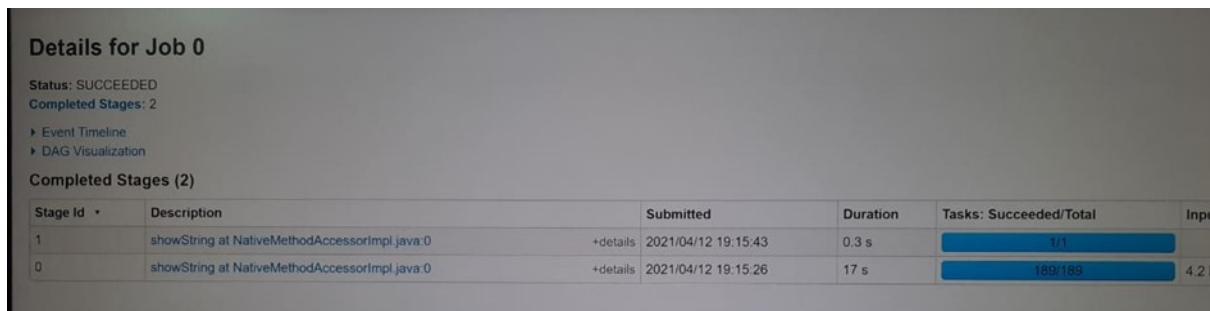
It took 3.5 minutes for the count to finish.

Then I started digging into tables. I realized that this hive table is poorly

designed and has lots of small files. Almost every partition has more than 100 files in KBs. In entire 190 partitions it had 32636 files. This is the reason spark had spawned so many tasks.

I had merged all the small files and kept 1 file per partition and also used compressions(snappy).

Now I ran the same count job and this time spark ran only 189 tasks and finish the count in 17 sec. Due to compression, the file size also reduced to 4.2 MB from 549 MB.



**Details for Job 0**

Status: SUCCEEDED  
Completed Stages: 2

- Event Timeline
- DAG Visualization

Completed Stages (2)

Stage Id	Description	Submitted	Duration	Tasks: Succeeded/Total	Input
1	showString at NativeMethodAccessorImpl.java:0	2021/04/12 19:15:43	0.3 s	1/1	
0	showString at NativeMethodAccessorImpl.java:0	2021/04/12 19:15:26	17 s	189/189	4.2 MB

*Running count on optimized hive table*