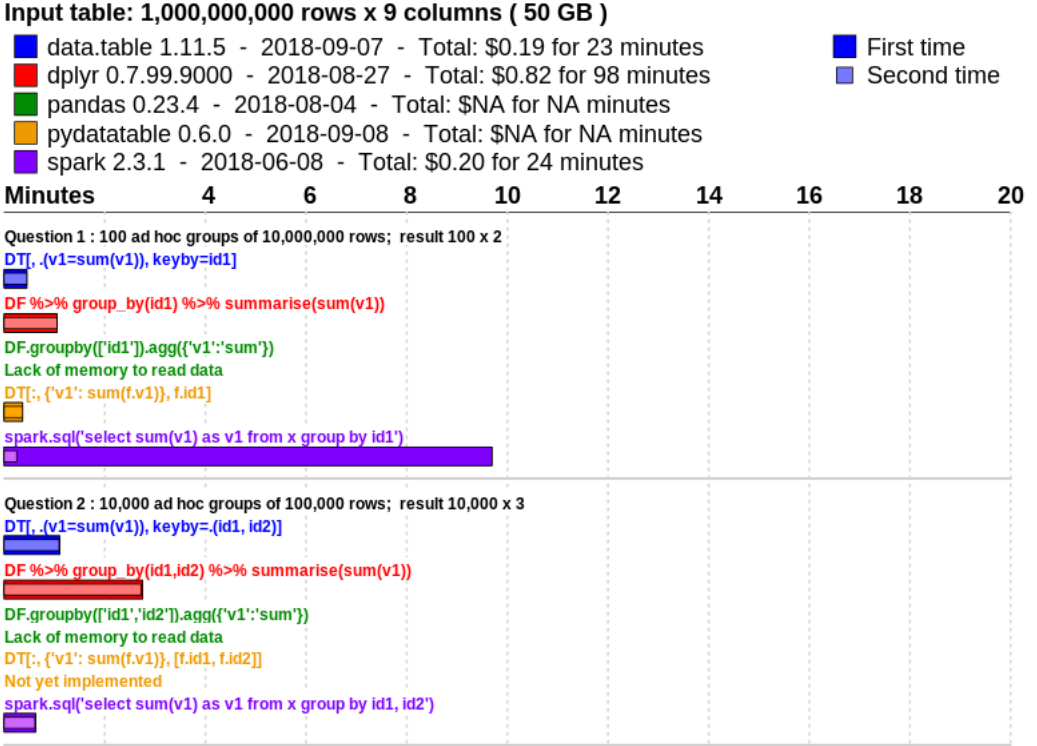
Input table: 1,000,000,000 rows x 9 columns (50 GB)
data.table 1.11.5 - 2018-09-07 - Total: $0.19 for 23 minutes
dplyr 0.7.99.9000 - 2018-08-27 - Total: $0.82 for 98 minutes
pandas 0.23.4 - 2018-08-04 - Total: $NA for NA minutes
pydatatable 0.6.0 - 2018-09-08 - Total: $NA for NA minutes
spark 2.3.1 - 2018-06-08 - Total: $0.20 for 24 minutes
First time
Second time
Minutes
4
6
8
10
12
14
16
18
20
Question 1 : 100 ad hoc groups of 10,000,000 rows; result 100 x 2
DT[, .(v1=sum(v1)), keyby=id1]
DF %>% group_by(id1) %>% summarise(sum(v1))
DF.groupby(['id1']).agg({'v1':'sum'})
Lack of memory to read data
DT[:, {'v1': sum(f.v1)}, f.id1]
spark.sql('select sum(v1) as v1 from x group by id1')
Question 2 : 10,000 ad hoc groups of 100,000 rows; result 10,000 x 3
DT[, .(v1=sum(v1)), keyby=.(id1, id2)]
DF %>% group_by(id1,id2) %>% summarise(sum(v1))
DF.groupby(['id1','id2']).agg({'v1':'sum'})
Lack of memory to read data
DT[:, {'v1': sum(f.v1)}, [f.id1, f.id2]]
Not yet implemented
spark.sql('select sum(v1) as v1 from x group by id1, id2')