1. SparkContext哪一端创建的的?

Driver端

2. DAG是在哪一端被构建的?

Driver端

3. RDD是在哪一端生成的?

Driver端

4. 调用RDD的算子（Transformation和Action）是在哪一端调用的

Driver端

5. RDD在调用Transformation和Action时需要传入函数，函数是在哪一端声明【定义】和传入的?

Driver端

6. RDD在调用Transformation和Action时需要传入函数，请问传入的函数是在哪一端执行了函数的业务逻辑?

Executor

7. Task是在哪一端生成的呢?

Driver端

8. DAG是在哪一端构建好的并被切分成一到多个Stage的

Driver端

9. DAG是哪个类完成的切分Stage的功能?

DAGScheduler

10. DAGScheduler将切分好的Task以什么样的形式给TaskScheduler

TaskSet

11. 自定义的分区器这个类是在哪一端实例化的?

Driver端

12. 分区器中的getParitition方法在哪一端调用的呢?

Executor

13. 广播变量是在哪一端调用的方法进行广播的?

Driver端

14. 要广播的数据应该在哪一端先创建好再广播呢?

Driver端

15. 广播变量以后能修改吗?

不能

16. 广播变量广播到Executor后，一个Executor进程中有几份广播变量的数据

一个

17. 累加器事先在哪一端创建的?

Driver端

18. 累加器事先在哪一端累加的

Executor

19. shuffle算子是否一定会触发shuffle

不会，需要看情况。如果现有数据已经按照一定规则和分区进行过划分，将要做的操作还是一样的分区规则和分区数量，则不需要再次shuffle了。

20. RDD为何高效?

RDD是不可变的+lazy。转化操作，行为操作。

RDD是粗度。[每次操作 都作用于所以集合] 对于RDD的写是粗粒度的 RDD的读 操作 可以是粗粒度的也可以是细粒度的： 可以读其中的一条记录。

注：资料来源于网络。