Spark架构

Spark Core（RDD+DAG）

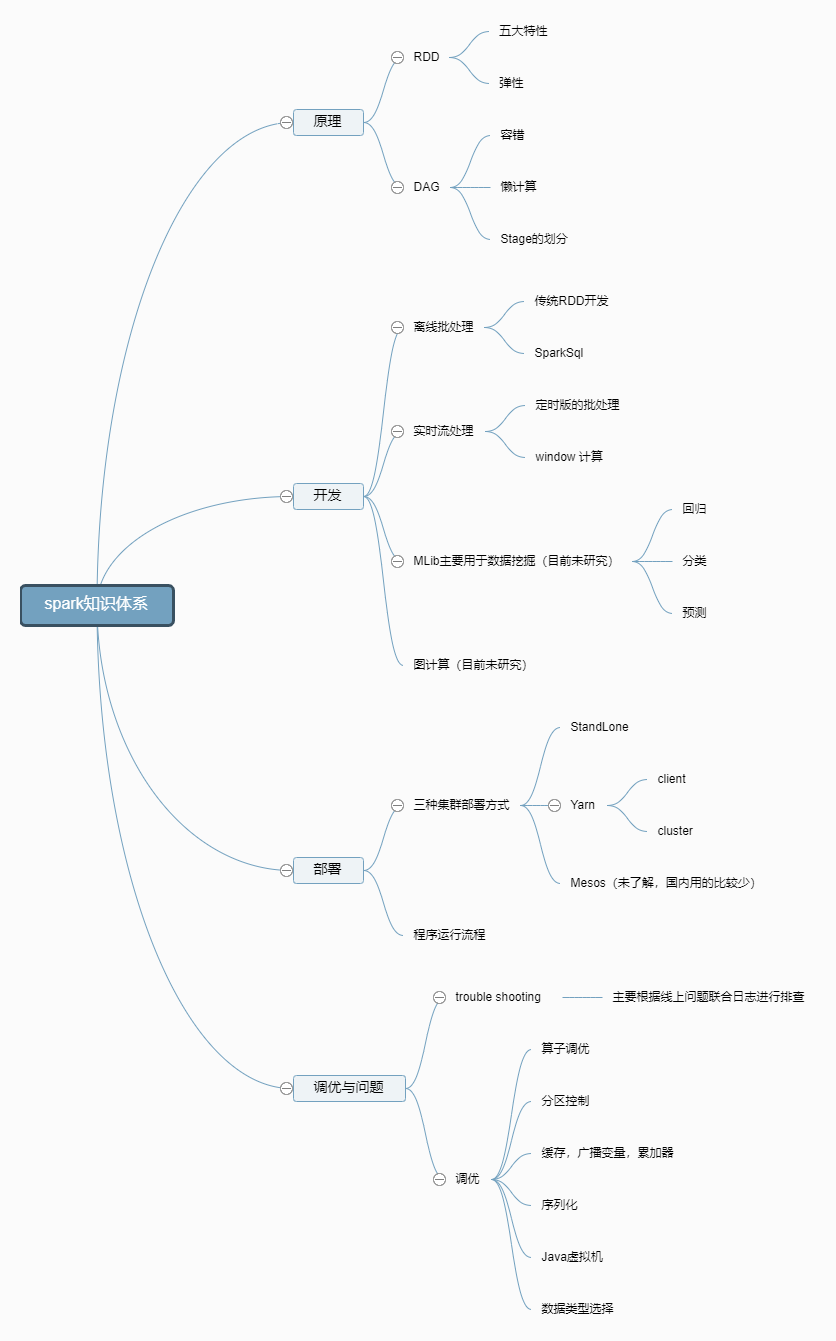
Spark Streaming

MLib

GraphX

sparkSql

Spark知识体系：



Spark知识体系总结：

Spark是一个分布式容错的计算框架，提供流处理和批处理的甚至机器学习等工具。框架的核心也就是spark core，其中rdd，dag，分区，shuffle是比较重要的，几乎所有的工具都跟他们有关。值得一提的是他这种移动计算不移动数据的设计思想也是值得佩服。不管是spark sql还是saprk straming 最终都是使用的rdd，在我理解而言，他们都是相对于rdd而言的高级api。在高级api的帮助下，基本编码的门槛得到降低，紧接着的问题就是调优，不管是官方，还是面试调优也是会被经常到。如果是我，我觉得不管用什么方法去调优，都要知道，影响spark性能的IO和内存占主要因素，而且这个也是可以通过程序员努力到达一些效果，另外的CPU，集群规模，怎么说呢，目前很多企业成本还是有限的。知道了这些因素，调优的点基本也就很明显了，内存有关的有哪些，java虚拟机，内存分配，数据结构，序列化（当然序列化也跟IO有关），跟IO有关的，分区文件，本地聚合算子，rdd本身具有就近计算，缓存，checkpoint，广播变量，BlockManager。

官方也正在做努力，比如tungsten（对内存，CPU缓存，以及spark代码生成都有很好的优化吧）,分布式内存文件系统。SparkStreaming本质还是那些算子，他们更像是定时去执行一些任务而实现流式处理，所以有准实时和实时的概念。当然也针对一些需求提出来窗口计算，窗口计算是什么呢：比如我要用每个一小时统计过去两个小时的点击量。这个时候就要用到窗口函数。

至于Mlib和GraphX有时间研究的话在继续吧！