SPARK

### Spark生态系统

* Shark：Shark基本上就是在Spark的框架基础上提供和Hive一样的HiveQL命令接口，为了最大程度的保持和Hive的[兼容性](http://baike.baidu.com/view/80015.htm" \t "_blank)，Shark使用了Hive的API来实现query Parsing和 Logic Plan generation，最后的PhysicalPlan execution阶段用Spark代替[Hadoop](http://baike.baidu.com/view/908354.htm" \t "_blank)[MapReduce](http://baike.baidu.com/view/2902.htm)。通过配置Shark参数，Shark可以自动在内存中缓存特定的RDD，实现数据重用，进而加快特定数据集的检索。同时，Shark通过UDF用户自定义函数实现特定的数据分析学习算法，使得SQL数据查询和运算分析能结合在一起，最大化RDD的重复使用。
* SparkR：SparkR是一个为R提供了轻量级的Spark前端的R包。 SparkR提供了一个分布式的data frame数据结构，解决了 R中的data frame只能在单机中使用的瓶颈，它和R中的data frame 一样支持许多操作，比如select,filter,aggregate等等。（类似dplyr包中的功能）这很好的解决了R的大数据级瓶颈问题。 SparkR也支持分布式的机器学习算法，比如使用MLib机器学习库。[1]  SparkR为Spark引入了R语言社区的活力，吸引了大量的数据科学家开始在Spark平台上直接开始数据分析之旅。