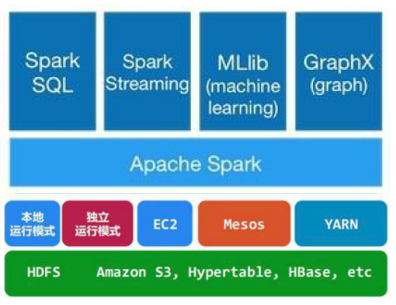
Spark2.4.3基本原理与安装

1. spark是一个实现快速通用的集群计算平台。它是由加州大学伯克利分校AMP实验室 开发的通用内存并行计算框架，用来构建大型的、低延迟的数据分析应用程序。它扩展了广泛使用的MapReduce计算

模型。高效的支撑更多计算模式，包括交互式查询和流处理。spark的一个主要特点是能够在内存中进行计算，及时依赖磁盘进行复杂的运算，Spark依然比MapReduce更加高效。

1. 中间结果输出：基于MapReduce的计算引擎通常会将中间结果输出到磁盘上，进行存储和容错。出于任务管道承接的，考虑，当一些查询翻译到MapReduce任务时，往往会产生多个Stage，而这些串联的Stage又依赖于底层文件系统（如HDFS）来存储每一个Stage的输出结果。
2. Spark是MapReduce的替代方案，而且兼容HDFS、Hive，可融入Hadoop的生态系统，以弥补MapReduce的不足。



1. 安装步骤
   1. 前置安装JDK就可以(Spark 不一定依靠Hadoop)
   2. 下载spark-2.4.3-bin-hadoop2.6.tgz
   3. 选择3个节点打通无密登录(1台做master节点 2台做worker节点)
   4. 在conf下cp slaves 和 spark-env.sh
      1. Slaves(配置worker节点)
         1. 加入两个workder节点名
      2. Spark-env.sh
         1. export SPARK\_MASTER\_HOST=node1 #主节点IP
         2. export SPARK\_MASTER\_PORT=7077 #任务提交端口
         3. export SPARK\_WORKER\_CORES=2 #每个worker使用2核
         4. export SPARK\_WORKER\_MEMORY=3g #每个worker使用3g内存
         5. export SPARK\_MASTER\_WEBUI\_PORT=8888 #修改spark监视窗口的端口默认8080
   5. 修改sbin下spark-config.sh
      1. export JAVA\_HOME=/usr/soft/jdk1.8.0\_201
   6. 把主节点上配置好的spark发布到2个worker节点上
   7. 启动spark
      1. 到主节点spark的sbin下运行 start-all.sh