MapReduce计算模型在大数据处理方面获得了巨大的成功，但如今实现该计算模型的绝大多数分布式系统（如Hadoop）基于“磁盘→磁盘”的非循环数据流处理，使得这些系统在机器学习、交互式数据挖掘等需要重复使用数据集的应用场景效率低。基于此，加州伯克利大学AMP实验室开发了类MapReduce的通用并行计算框架Spark，该框架提出了一种叫“弹性的分布式数据集（Resilient Distributed Datasets，简称RDD）”的粗粒度数据集抽象，RDD能有效地利用内存资源，高效地实现迭代式计算。

Shark是基于Spark的数据仓库，既支持SQL查询处理，也支持机器学习函数。Shark是与Hive兼容的，用户可以无需修改Hive的查询以及数据就能在shark上执行。

性能方面，shark执行SQL查询的速度比Hive快100倍，执行机器学习算法比Hadoop快100倍。

Shark为什么比Hive有如此高的性能提升，下面有请今天主讲人沈昌干做本次seminar的演讲。