Hadoop是适合于海量数据的分布式存储和分布式计算的框架

三大核心：

HDFS(Hadoop Distributed File Sytem hadoop分布式文件系统)：分布式存储

MapReduce:用于分布式计算（批处理 离线计算）

Yarn：资源管理系统

分布式存储

把一个整体文件拆分成多个模块，存储在不同的主机上，通过对不同主机的上文件块（block）在内存中的映射

来对这些文件进行管理。

分布式计算

把一个比较复杂的计算任务，分解成许多个简单的小的计算任务，在不同的机器上运行，得到结果，最后将这些

机器上的结果进行汇总，得出最终结果

**1. Hadoop核心件组有哪些? 广义hadoop指什么?**

l 核心组件有：Hdfs、Yarn、MapReduce;

l 广义上指一个生态圈，泛指大数据技术相关的开源组件或产品，如hdfs、yarn、hbase、hive、spark、pig、zookeeper、kafka、flume、phoenix、sqoop。

**2. Spark与hadoop之间有什么联系**

l Spark也是一个生态圈，发展非很快，在计算方面比mapreduce要快很多倍，供了一个简单而丰富的编程模型，支持多种应用，包括ETL、机器学习、数据流处理、图形计算；

l Hadoop和Spark在某些方面有重叠，但组件之间都可以很好的协同作战。



