大数据框架

Hadoop

什么是hadoop?

Hadoop是一个分布式系统基础架构，可以使用户在不了解分布式底层细节的情况下开发分布式程序。充分利用集群的威力来进行高速运算和存储。

Hadoop的组成是怎样的？

分布式文件系统HDFS、MapReduce引擎、数据厂库Hive分布式数据库Hbase。

Hadoop有什么优点？

可靠、高效、可伸缩地进行分布式大数据处理；

高可靠、高扩展、高效、高容错、低成本；

备注：hadoop带有java语言编写的框架，也可以使用其他编程语言；

Hadoop在大数据中应用广泛得益于其在数据提取、变形和加载方面的优势。其分布式架构，将大数据处理引擎尽可能的靠近存储。

SPARK

什么是Spark?

Spark是专为大规模数据处理而设计的快速计算引擎，是一种开源的类似于Hadoop MapReduce的通用并行框架。

Spark是一种与Hadoop类似的开源集群计算环境，其在某些工作负载方面表现更加优越，Spark启用了内存分布数据集，除了提供交互式查询外，还可以优化迭代工作负载，因此能更好的适用于数据挖掘和机器学习等需要迭代的MapReduce算法。