作业 7

171850532 司富元

1. 简述为什么会有 Spark
   - Hadoop 计算框架对非批大数据问题的支持局限性
   - Spark 拥有大数据处理一站式解决平台的完善生态系统
     - 包括 Spark SQL，Spark Streaming，MLlib，GraphX 等组件
   - Spark 本身性能卓越
     - 内存计算思想显著提高计算性能：提出基于内存的弹性分布式数据集（RDD）
     - 集诸多计算模式之大成：流式计算、迭代计算、图计算、内存计算

2. 对比 Hadoop 和 Spark
   - Spark 迭代运算效率高
     - Spark 中间数据存于内存，而 MapReduce 计算结果存于 HDFS
       - 磁盘数据运算速度快于 Hadoop 2x-10x
       - 内存数据运算速度快 100x
     - Spark 支持 DAG 图的分布式并行计算，减少数据落地
   - Spark 容错性高
     - 引入 RDD 抽象，可通过 CheckPoint 实现容错
   - Spark 更加通用
     - Spark 提供相当多的数据集操作，而 MapReduce 只提供两种有限操作
     - Spark 中用户可命名、物化、控制中间结果的存储、分区，而 MapReduce 中各处理节点只能通过 Shuffle 通信
     - Spark 支持 Java、Python、Scala
   - Spark 只是分布式计算，Hadoop 分布式计算和存储
   - Spark 擅长迭代工作，可用于机器学习。而 Hadoop 迭代工作效率不理想
   - Spark 生态完善，拥有 Spark SQL，Spark Streaming，MLlib，GraphX 等组件，支持常见 Use case

3. 简述 Spark 的技术特点
   - 适用于需要多次操作特定数据集的 use case

    - 反复操作越多，受益越大
- 不适用异步细粒度更新状态（增量修改）的 use case
    - 如 web 服务存储或增量的 web 爬虫和索引
- 适用于要求实时统计分析但数据量不是很大的 use case
- 总之，Spark 是一中基于内存的迭代是分布式计算框架
    - 适用于迭代、关系查询、流式查询等计算密集 use case