Spark和MapReduce之间的区别？各自优缺点?

可回答：1）spark和maprecude的对比；2）mapreduce与spark优劣好处

问过的一些公司：阿里云(2022.10)，银联(2022.10)，携程(2022.09)，vivo(2022.09)，滴滴(2022.09)(2020.09)，网易云音乐(2022.09)，快手(2022.08)，字节(2022.08)x2(2022.05)(2020.09)(2020.06)(2019.11)x4，快手(2022.08)，星环科技(2022.07)，海康威视(2022.06)，字节日常实习(2022.03)，思科cisco(2021.11)，腾讯PCG(2021.10)，腾讯云(2021.10)，阿里(2021.10)，蔚来(2021.09)，重庆富民银行(2021.09)，网易杭研院(2021.09)，网易严选(2021.08)，小米(2021.08)(2020.09)(2019.09)，华为精英计划(2021.07)，触宝(2021.07)，有道(2021.03)，作业帮社招(2020.09)，58(2020.09)，一点资讯(2020.08)，多益(2020.08)，360实习(2020.04)，阿里菜鸟(2020.04)，腾讯互娱(2020.03)，蘑菇街实习(2020.03)x2，阿里淘系(2019.11)，美团大众点评(2019.10)，微众银行(2019.09)，网易有道(2019.08)，招商银行信用卡中心(2019.04)，光大银行(2019.03)，头条(2018.11)

参考答案：

1、Spark处理数据是基于内存的，而MapReduce是基于磁盘处理数据的

MapReduce是将中间结果保存到磁盘中，减少了内存占用，牺牲了计算性能。

Spark是将计算的中间结果保存到内存中，可以反复利用，提高了处理数据的性能。

2、Spark在处理数据时构建了DAG有向无环图，减少了shuffle和数据落地磁盘的次数

Spark计算比MapReduce快的根本原因在于DAG计算模型。一般而言，DAG相比MapReduce在大多数情况下可以减少shuffle次数。Spark的DAGScheduler相当于一个改进版的MapReduce，如果计算不涉及与其他节点进行数据交换，Spark可以在内存中一次性完成这些操作，也就是中间结果无须落盘，减少了磁盘IO的操作。但是，如果计算过程中涉及数据交换，Spark也是会把shuffle的数据写磁盘的。

3、Spark比MapReduce快

有一个误区，Spark是基于内存的计算，所以快，这不是主要原因，要对数据做计算，必然得加载到内存，Hadoop也是如此，只不过Spark支持将需要反复用到的数据Cache到内存中，减少数据加载耗时，所以Spark跑机器学习算法比较在行（需要对数据进行反复迭代）。

4、Spark是粗粒度资源申请，而MapReduce是细粒度资源申请

粗粒度申请资源指的是在提交资源时，Spark会提前向资源管理器（YARN，Mess）将资源申请完毕，如果申请不到资源就等待，如果申请到就运行task任务，而不需要task再去申请资源。

MapReduce是细粒度申请资源，提交任务，task自己申请资源自己运行程序，自己释放资源，虽然资源能够充分利用，但是这样任务运行的很慢。

5、MapReduce的Task的执行单元是进程，Spark的Task执行单元是线程

进程的创建销毁的开销较大，线程开销较小。

6、Spark优缺点

优点：

1）Spark把中间数据放到内存中，迭代运算效率高。

Spark支持DAG图的分布式并行计算的编程框架，减少了迭代过程中数据的落地，提高了处理效率。

2）Spark 容错性高

Spark 引进了弹性分布式数据集 RDD (Resilient DistributedDataset) 的抽象，它是分布在一组节点中的只读对象集合，这些集合是弹性的，如果数据集一部分丢失，则可以根据“血统”（即允许基于数据衍生过程）对它们进行重建。另外在RDD 计算时可以通过 CheckPoint 来实现容错。

3）Spark更加通用

Spark提供的数据集操作类型分为：Transformations和Actions两大类。Transformations包括Map、Filter、FlatMap、Sample、GroupByKey、ReduceByKey、Union、Join、Cogroup、MapValues、Sort等多种操作类型，同时还提供Count, Actions包括Collect、Reduce、Lookup和Save等操作。

缺点：

1）内存问题

JVM的内存overhead太大，1G的数据通常需要消耗5G的内存。

2）性能问题

由于大量数据抄被缓存在RAM中，Java回收垃圾缓慢的情况严重，导致Spark性能不稳定。

7、MapReduce优缺点

优点：

1）MapReduce 易于编程

它简单的实现一些接口，就可以完成一个分布式程序，这个分布式程序可以分布到大量廉价的 PC 机器上运行。也就是说你写一个分布式程序，跟写一个简单的串行程序是一模一样的。就是因为这个特点使得 MapReduce 编程变得非常流行。

2）良好的扩展性

当你的计算资源不能得到满足的时候，你可以通过简单的增加机器来扩展它的计算能力。

3）高容错性

MapReduce 设计的初衷就是使程序能够部署在廉价的 PC 机器上，这就要求它具有很高的容错性。比如其中一台机器挂了，它可以把上面的计算任务转移到另外一个节点上运行， 不至于这个任务运行失败，而且这个过程不需要人工参与，而完全是由Hadoop内部完成的。

4）适合 PB 级以上海量数据的离线处理

可以实现上千台服务器集群并发工作，提供数据处理能力。

缺点：

1）不擅长实时计算

MapReduce无法像MySQL一样，在毫秒或者秒级内返回结果。

2）不擅长流式计算

流式计算的输入数据是动态的，而MapReduce的输入数据集是静态的，不能动态变化。这是因为 MapReduce 自身的设计特点决定了数据源必须是静态的。

3）不擅长 DAG（有向无环图）计算

多个应用程序存在依赖关系，后一个应用程序的输入为前一个的输出。在这种情况下，MapReduce并不是不能做，而是使用后，每个MapReduce作业的输出结果都会写入到磁盘， 会造成大量的磁盘 IO，导致性能非常的低下。