### 当前几种常见的数据处理分析技术

大数据的分布式处理技术与存储形式、业务数据类型等相关，针对大数据处理的主要计算模型有**MapReduce分布式计算框架**、**分布式内存计算系统**、**分布式流计算系统**等。

#### MapReduce分布式计算框架

批处理的分布式计算框架，适合对各种结构化、非结构化数据的处理。

大致流程：从HDFS中获取数据，将其按照大小分片，进行分布式处理，最终输出结果。

该框架对数据来源的限制：

1、 数据已经存在在HDFS当中。

2、 数据间是少关联的。各个任务执行器在执行负责的数据时，无需考虑对其他数据的影响，数据之间应尽可能是无联系、不会影响的。

基于Map/Reduce这种单级的数据处理模型进行，因此，如果数据间的关联系较大，需要进行数据的多级交互处理（某个阶段的处理数据依赖于上一个阶段），需要进行多次map/reduce，并且map/reduce每次执行都需要遍历整个数据集。

开源框架参考：Hadoop

#### 分布式内存计算系统

基于内存的分布式批处理引擎，通俗说就是把计算机集群分布，增加可用内存量，形成所谓的计算机集群，并且在集群间引入特定的通信机制。

系统的特点：

内存效率较其他两种系统要快很多（提高10倍左右），可有效减少数据读写和移动的开销，提高大数据处理性能；

有通用性；

有良好的可扩展性；

开源框架参考：spark

#### **分布式流计算系统**

流式数据的应用大多是监控型的，即持续运行在连续数据流上的连续查询。

流式数据的特点：无限性；时序性（数据流的时序性有两方面含义。一方面指元组按序到达，不能在整个数据流上实现随机访问。另一方面指其元组当中通常包含直接或间接的时间标签，如时间戳，序列号。）；大小和规模不可预知。

分布式流计算系统适用场景（系统的特点）：

1.可以用来用来处理源源不断的消息，并将处理之后的结果保存到持久化介质中。

2.处理组件可分布式，处理延迟极低。

3.易于扩展。

4.信息处理的完整性较高；

5.适用于多种语言开发；

开源框架参考：storm