1. 数据的存储和分析

1990年一个普通硬盘可以存储1370MB，传输速率4.4MB/S，需要大约5分钟读完所有数据.

几十年后的今天，TB级硬盘到处可见，但是传输速率约100MB/S左右。那么读完整个1TB硬盘数据需要2.5小时。那么该如何解决?

假设我们有100个硬盘，在每个硬盘存储1%，并行读取大约2分钟读完所有数据。

1. 与关系型数据库的对比
2. 磁盘寻址时间提高速度远远小于数据传输速率提高速度。关系型数据库使用B树结构进行数据的更新查询操作，对于最大到GB的数据量，一般相对数据量较小，效果很好。但是大数据量时，B树使用排序/合并方式重建数据库以更新数据的效率远远低于MapReduce。
3. 关系型数据使用的是结构化数据，在数据库阶段按具体类型处理数据。MapReduce数据模式半结构化(xml，Excel)或非结构化（图像，日志）。所以MapReduce更适合做日志记录的分析
4. MapReduce适合一次写入、多次读取数据的应用，而关系型数据库更适合持续更新的数据集
5. Hadoop的简史

Hadoop是Apache Lucene的创始人Doug Cutting创建。最初Doug Cutting在想做网络搜索引擎。

2002年号称支持10亿的网络搜索引擎(Nutch)面世了。但是在运用发现不好用，且硬件投资50万美元。

2003年谷歌公布了自己产品的架构“谷歌分布式文件系统”GFS论文。

2004年开始开源实现(NDFS),同年谷歌公布了MapReduce论文。

2005年Nutch团队也开源实现了Mapreduce.

2006年2月，Nutch正式将NDFS和Mapreduce移出重名为Hadoop.同年Cutting加入雅虎

2008年1月，Hadoop成为Apache顶级项目，

2008年4月，在910个节点的集群上运行的hadoop用时209s对1TB数据排序

2009年5月，hadoop再次打破记录用时62s对1TB数据排序，（据说谷歌公布他们的Mapreduce和GFS用时68s完成1TB的数据排序）