无处不在的大数据

科学数据

物联网数据

交通数据

金融数据

零售数据

社交网络数据

“21世纪核心的竞争是数据的竞争，谁拥有更多的数据，谁就拥有未来”

如何对大数据进行存储和分析那？

系统的瓶颈

存储容量，读写速率，计算效率...

Google提出的大数据技术

Mapreduce，bigtable，gfs

带来的革命性变化

成本降低，能用PC机，就不用大型机和高端存储

软件容错硬件故障视为常态，通过软件保证可靠性

简化并行分布式计算，无需控制节点同步和数据交换

模仿Google大数据技术的开源实现——Hadoop

Hadoop是什么？

开源的apache

分布式存储和分布式计算平台

包括两个核心组成部分

HDFS：分布式文件存储系统，存储海量数据

MapReduce：并行处理框架，实现任务分解和调度

Hadoop可以用来做什么？

搭建大型数据仓库，PB级数据的存储、处理、分析、统计等业务

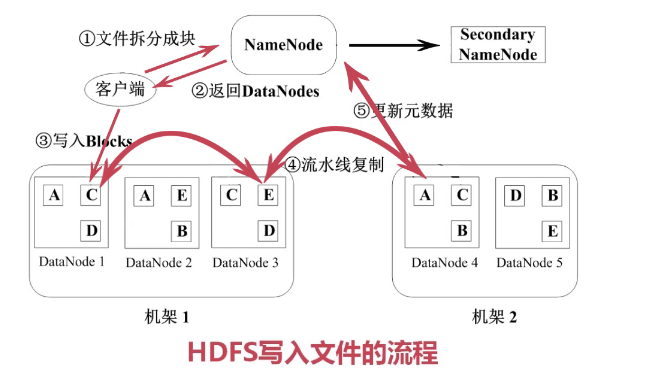
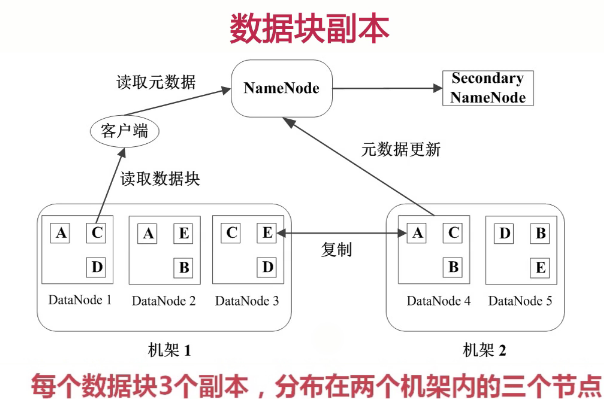
Hadoop的优势：高扩展、低成本、成熟的生态圈

Namendoe是管理节点，存放文件的元数据

文件与数据块的映射表

数据块与数据节点的映射表

Datanode是HDFS的工作节点，用来存放数据块



HDFS的特点

数据冗余，硬件容错

流式的数据访问

存储大文件

适合数据的批量读写，吞吐量高

不适合做交互式的应用，低延迟很难得到满足

适合一次写入多次读取，顺序读写

不支持多用户的并发写相同的文件

