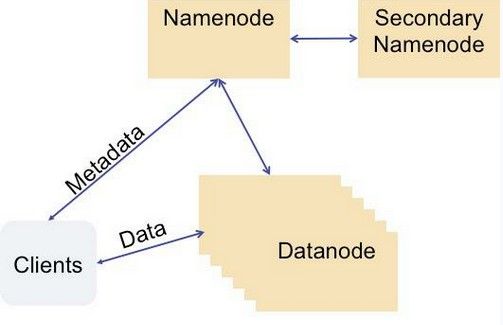
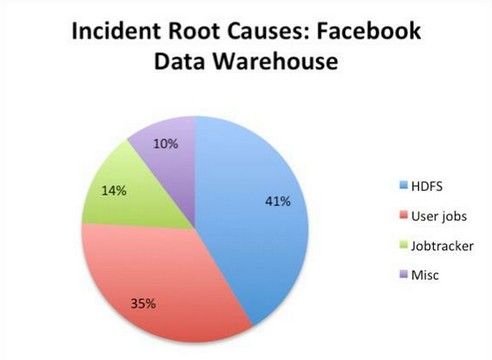
我们知道，Facebook使用Hadoop来进行大数据的处理，但Facebook又是如何保障频繁、庞大的数据请求等高压环境下不发生故障的呢？我们一起来了解一下Facebook内部的Hadoop使用情况以及其NameNode故障转移技术。  
　　Facebook Hadoop集群内目前的HDFS物理磁盘空间承载超过100PB的数据（分布在不同数据中心的100多个集群）。由于HDFS存储着Hadoop应用需要处理的数据，因此优化HDFS成为Facebook为用户提供高效、可靠服务至关重要的因素。  
　　HDFS Namenode是如何工作的？  
　　HDFS客户端通过被称之为Namenode单服务器节点执行文件系统原数据操作，同时DataNode会与其他DataNode进行通信并复制数据块以实现冗余，这样单一的DataNode损坏不会导致集群的数据丢失。  
　　但NameNode出现故障的损失确是无法容忍的。NameNode主要职责是跟踪文件如何被分割成文件块、文件块又被哪些节点存储，以及分布式文件系统的整体运行状态是否正常等。但如果NameNode节点停止运行的话将会导致数据节点无法通信，客户端无法读取和写入数据到HDFS，实际上这也将导致整个系统停止工作。



　　The HDFS Namenode is a single point of failure (SPOF)  
Facebook也深知"Namenode-as-SPOF"所带来问题的严重性，所以Facebook希望建立一套系统已破除"Namenode-as-SPOF"带来的隐患。但在了解这套系统之前，首先来看一下Facebook在使用和部署HDFS都遇到了哪些问题。  
　　Facebook数据仓库的使用  
在Facebook的数据仓库中部署着最大的HDFS集群，数据仓库的使用情况是传统的Hadoop MapReduce工作负载在大型集群中一小部分运行MapReduce批处理作业  
因为集群非常庞大，客户端和众多DataNode节点与NameNode节点传输海量的原数据，这导致NameNode的负载非常沉重。而来自CPU、内存、磁盘和网络带来的压力也使得数据仓库集群中NameNode高负载状况屡见不鲜。在使用过程中Facebook发现其数据仓库中由于HDFS引发的故障占总故障率的41%。



　　HDFS NameNode是HDFS中的重要组成部分，同时也是整个数据仓库中的重要组成部分。虽然高可用的NameNode只可以预防数据仓库10%的计划外停机，不过消除NameNode对于SPOF来说可谓是重大的胜利，因为这使得Facebook可执行预订的硬件和软件回复。事实上，Facebook预计如果解决NameNode可消除集群50%的计划停机时间。  
　　那么高可用性NameNode是什么样子的？它将如何工作？让我们来看一下高度可用性NameNode的图表。  
　　在此结构中，客户端可与Primary NameNode与Standby NameNode通信，同样众多DataNode  
也具备给Primary NameNode与Standby NameNode发送block reports的能力。实质上Facebook所研发的AvatarNode就是具备高可用NameNode的解决方案。  
　　Avatarnode：具备NameNode故障转移的解决方案  
为了解决单NameNode节点的设计缺陷，大约在两年前Facebook开始在内部使用AvatarNode工作。  
　　同时AvatarNode提供了高可用性的NameNode以及热故障切换和回滚功能，目前Facebook已经将AvatarNode贡献到了开源社区。经过无数次的测试和Bug修复，AvatarNode目前已在Facebook最大的Hadoop数据仓库中稳定运行。在这里很大程度上要感谢Facebook的工程师Dmytro Molkov。  
　　当发生故障时，AvatarNode的两个高可用NameNode节点可手动故障转移。AvatarNode将现有的NameNode代码打包并放置在Zookeeper层。  
　　AvatarNode的基本概念如下：  
    具备Primary NameNode与Standby NameNode  
    当前Master主机名保存在ZooKeeper之中  
    改进的DataNode发送block reports到Primary NameNode与Standby NameNode  
    改进的HDFS客户端将在每个事物开始之前对Zookeeper进行检查，如果失败会转移到另外的事务之中。同时如果AvatarNode故障转移出现在写入的过程中，AvatarNode的机制将允许保证完整的数据写入。  
　　或许有人会Facebook这一解决方案的名字感到好奇，这是因为Facebook的Hadoop工程师Dhruba Borthakur来到公司时正好是James Cameron阿凡达电影热映时间。（我们应该感到庆幸，如果是1998年的话或许应该叫TitanicNode了）。  
　　AvatarNode经受住了Facebook内部最苛刻的工作环境，未来Facebook将继续大幅度改善AvatarNode的可靠性和HDFS集群的管理性。并整合与一般高可用性框架的整合，还将实现无人值守、自动化与安全故障转移等特性。  
　　Facebook已将自身使用的Hadoop与AvatarNode解决方案托管到GitHub。感兴趣的朋友可下载研究。当然不止Facebook在试图解决Hadoop的缺陷，MapR和Cloudera的产品也具备相似的能力。  
　　【编辑推荐】