Hadoop 存储技术研究进展

韩洪勇 曹源 顾龙雨 吴杰
(山东科技大学电气信息系 山东济南 250000)

摘要：当前的存储技术虽然已经取得了巨大的发展，一块小小的 U 盘就已经能够存储 128G，但是随着智能设备的发展，ipv4 都已经不够用了，存储技术也已经越来越显得捉襟见肘。需求上来了，硬件跟不上也要解决呀。硬件上需要有突破,软件上也需要有相应的跟进,目前的面对大数据的解决方案是通过分布式的存储技术来破除硬件上的限制,当然这样做也有助于对提升对数据计算能力。本文将从现在存储技术解决方案进行讲解。

关键词：Hadoop；数据获取；数据清洗；数据存储

1 数据获取

Google 一天产生的数据都是按照 PB 计算的。现代的数据具有 4v 的特点：1.volumn（体量大）、2 .variaty （样式多） 3 .Velocity（速度快）、 4 .Valueless（价值密度低）。不管数据有多大 数据采取什么结构、来源如何，只要能够带来价值,都会想办法进行处理，好比一块一块小石头也能筑起长城一样。

1.1 数据采集方法

1.1.1 系统日志采集方法

很多互联网企业都有自己的海量数据采集工具，多用于系统日志采集，如 Hadoop 的 Chukwa，Cloudera 的 Flume，Facebook 的 Scribe 等，这些工具均采用分布式架构，能满足每秒数百 MB 的日志数据采集和传输需求。

1.1.2 网络数据采集方法

网络数据采集是指通过网络爬虫或网站公开 API 等方式从网站上获取数据信息。该方法可以将非结构化数据从网页中抽取出来，将其存储为统一的本地数据文件，并以结构化的方式存储。它支持图片、音频、视频等文件或附件的采集，附件与正文可以自动关联。

1.1.3 数据库采集系统

对于企业生产经营数据或学科研究数据等保密性要求较高的数据，可以通过与企业或研究机构合作，使用特定系统接口等相关方式采集数据。

2. 数据清洗

数据获取之后还需要进行数据清洗工作,也即是剔除"脏数据"。构建业务模型，在确定特征向量以后，都需要准备特征数据在线下进行训练、验证和测试。同样，部署发布离线场景模型，也需要每天定时跑 P 加工模型特征表。而这一切要做的事，都离不开数据清洗，业内话来说，也就是 ETL 处理(抽取 Extract、转换 Transform、加载 Load)，三大法宝。大数据挖掘更多时间都在于清洗数据。

2.1 清洗对象

现在所做的数据清洗主要集中在以下四个方面:（1）检测并消除数据异常；（2）检测并消除近似的记录；（3）数据的集成；（4）特定领域的数据清洗。

2.2 清洗方法

数据清洗方法包含：手工实现、编程实现、针对特定领域问题和无关数据剔除四类。

3. 数据存储

数据经过清洗后就需要进行存储，但是当前除了少数互联网科技公司有内部自定义的数据库存储外。在数据存储领域传统的关系型数据库(RDBMS)还是占据了主流。RDBMS 设计的主要思想是 OLAP。

联机事务处理 OLTP（on-line transaction processing）、联机分析处理 OLAP（On-Line Analytical Processing）。OLTP 是传统的关系型数据库的主要应用，主要是基本的、日常的事务处理，例如银行交易。OLAP 是数据仓库系统的主要应用，支持复杂的分析操作，侧重决策支持，并且提供直观易懂的查询结果。可以看出 OLAP 已经越来越不能满足当代数据存储的要求了。一方面是单个服务器存储容量的进步远远小于数据的增长。另外一方面是 CPU 等硬件对数据的运算能力也已经赶不上数据的增长速度。

我们以商业智能 FineBI 来分析。其提供了常见的 OLAP 多维分析操作，对于用户，可以对已有的表样切换维度来进行数据钻取分析。同时支持对数据的排序与过滤功能，按照自身需求对数据分析处理。

3.1 HDFS 存储原理

庞大的数据进去到 HDFS 系统前需要先进行切块(block)操作。HDFS 中的文件在物理上是分块存储（block）。

HDFS 的块比磁盘的块大(磁盘的块一般为 512 字节),其目的是为了最小化寻址开销。如果块设置得足够大，从磁盘传输数据的时间会明显大于定位这个块开始位置所需的时间。因而，传输一个由多个块组成的文件的时间取决于磁盘传输速率。但是很多情况下 HDFS 使用默认 128MB 的块设置。块的大小：10ms100100M/s = 100M，然而真正实际开发中要把 block 设置的远大于 128MB，比如存储文件是 1TB 时,一般把 Block 大小设置成 512MB.但是也不能任意设置的太大，比如 200GB 一个，因为在 MapReduce 的 map 任务中通常一次只处理一个块中数据（切片大小默认等于 block 大小），如果设置太大，因为任务数太少（少于集群中的节点数量），那么作业的运行速度就会慢很多，此外比如故障等原因也会拖慢速度。

通常 datanode 从磁盘上读取块，但是对于频繁访问的数据块，datanode 会将其缓存到 dataNode 节点的内存中，以堆外块缓存的形式（off-heap block cache ）存在。默认情况下，一个块只缓存到一个 datanode 内存中（加入复本是 3 个，但是也只在一个 datanode 内存中缓存块）。这样的话，计算框架，比如 MR 或者 Spark 就可以在缓存块的节点上运行计算任务，可以大大提高读操作的性能，进而提高任务的效率。

结论及展望

以上经过原理层面的分析，完成了对海量数据的存储与处理。未来是否还要依靠这种方式，还要看下一代量子计算机的性能。可能我们现在每天朝思暮想地进行调优，对量子计算机来说也是很轻松就能做到的事情。

参考文献

[1] （美）怀特 著.《Hadoop 权威指南》[M].2.清华大学出版社, 2010.

[2] 陆嘉恒.Hadoop 实战（第 2 版）[M].3. 机械工业出版社华章公司, 2012.

[3] 蔡斌 陈湘萍.Hadoop 技术内幕：深入解析 Hadoop Common 和 HDFS 架构设计与实现原理[M].1-1.机械工业出版社, 2013.

作者简介：韩洪勇，男，(1999.08.13-)，山东青岛人，现于山东科技大学攻读学士学位，目前主要从事于计算机科学与技术的专业研究。