结构化大数据解决方案

1 引言

现在，各行各业都已经认识到“大数据”处理的价值，“大数据”已经成为一个炙手可热的名称。可是，在 2000 年，我们开始研究“大数据”的时候，它还只是学术小圈子里的术语。经过十多年的研究和开发，我们先后推出了多个大数据处理的解决方案。它们涉及到“大数据”的存储、访问和分析等问题的解决。

从内部组织结构来分，“大数据”可以分为非结构化大数据和结构化大数据。所谓“非结构化大数据”就是数据内部元素之间没有固定的，一致的组织关系的数据。例如各种视频、文字、图片和表格的组合体。而“结构化大数据”内部具有固定的，一致的组织关系和结构。非结构化大数据的处理技术已经得到了足够多的关注。但是，如何处理海量的结构化大数据则仍然是一个问题。现在，针对结构化大数据，我们推出新一代的解决方案。

2 结构化大数据处理的挑战

结构化大数据通常存储在数据库中，以数据表的方式存储、访问、查询和管理。当数据表的容量增长到一定程度后，就将面临以下问题：

1）数据更新速度无法满足应用要求。通常数据表都有索引、外键等内部组织结构。当表的内容增大后，数据更新时对这些内部结构的维护开销也急剧增大，从而使得数据更新速度急剧下降，以致无法满足应用需求。

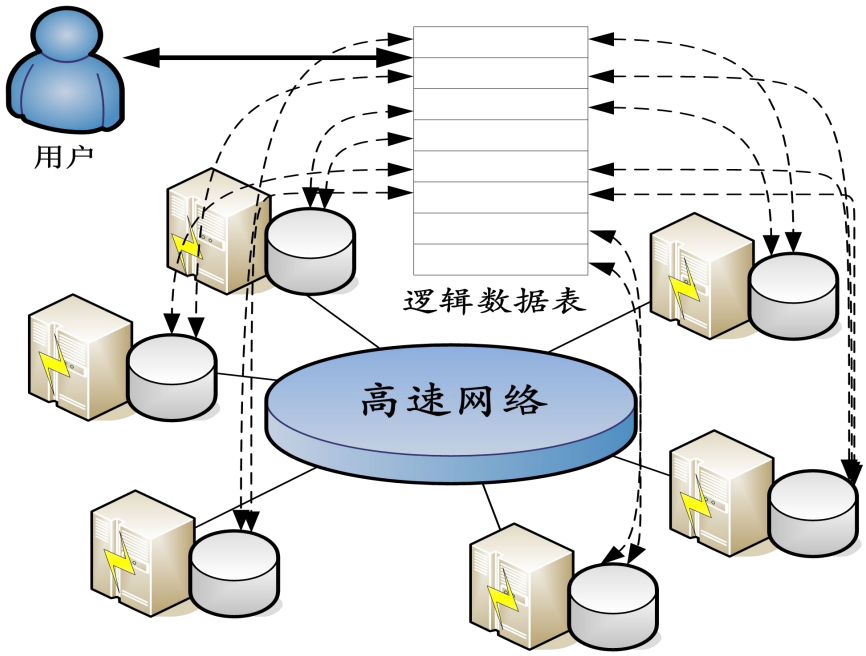
2）数据查询速度无法满足应用要求。当数据表内容增大都一定程度后，一台服务器已经无法将表数据大量载入内存，从而使得数据分析严重依赖磁盘数据读取。而磁盘读取的效率远远低于内存访问，从而使得数据查询速度下降数个数量级。因此，随着表数据增大，查询速度将急剧下降，以致无法满足应用要求。

3）数据表越来越不可靠。累计的数据就是财富，而将这些财富放在单个存储设备上，将变得越来越无法让人放心。存储设备一旦出现故障，宝贵的财富也就化为乌有。因此，越来越需要高可靠的数据存储方案。

针对上述三个问题，我们推出数字有机体结构化大数据处理解决方案，力图全面解决上述三个问题。

3 数字有机体结构化大数据处理解决方案

数字有机体结构化大数据处理的解决方案以数字有机体工作库为软件核心，依托高速网络互连的服务器系统，解决结构化大数据查询的难题。其逻辑结构如图所示。



数字有机体工作库系统将一个逻辑数据表按照一定的规则自动分解到多台服务器的自有存储系统中。当处理涉及到该逻辑数据表的查询时，数字有机体工作库自动将查询分解到各台服务器上协同执行，通过整合这些服务器的处理能力，加速查询处理。和 MPP(Massively Parallel Processing)数据库不同，数字有机体工作库在结构化大数据处理上具有以下功能：

1）数字有机体工作库支持同一个库中不同的表采用不同的分片策略，且支持某些表不进行分片处理。这使得它可以适应各种复杂的应用环境。对小表的查询不会因为无意义的分片而降低性能。

2）在进行小表和大表连接时，系统采用特殊的链接处理机制，处理效率更高，查询需要的时间更短。

3）系统支持多种分片方式。当采用范围和列表方式时，系统可以根据查询条件过滤不需要访问的分区，避免盲目的全表扫描，大幅度减少处理开销，从而可以支持更大的查询并发度。

4）系统没有单一入口或者 master 节点，元数据维护更加简单，可以支持更大规模的系统。

5）可以直接更新各个分片数据表，从而消除数据插入的瓶颈，解决表数据增大后插入速度降低的问题。

6）系统支持分片多复制策略。即每个分片都可以有多个拷贝。分片复制间的同步支持实时日志同步方式或者异步日志同步方式。提升数据的可靠性，使得数据表不会因为某个存储设备故障而丢失。

7）系统支持多个表间的诱导分片机制。当多个表间采用诱导分片时，可以极大的提升诱导连接的查询处理速度。

数字有机体工作库系统的上述功能使得它有别于现有的数据库集群、MPP 或者列式数据库，可以解决结构化大数据处理的上述难题。

4 方案特点

1）可以利用各种高性价比的资源搭建系统，无需专用的共享存储设备，无需专用的高速网络。这使得系统具有良好的性价比。并且可以充分利用现有的各种资源，实现对已有投资的保护。

2）系统具有良好的扩展性，可以根据需要增加服务器数量，通过调整数据分片即可重构系统，从而能够按需部署系统。这是许多大数据一体机无法提供的能力。

3）高查询处理速度。通过充分整合各台服务器的处理能力，可以获得极大的查询加速比，从而缩短查询处理时间。在一个测试实例中，它将一个需要数分钟才能完成的查询缩短到几秒的水平。

4）数据表很大时仍然有较高的数据插入速度。通过数据表分片和分片表直接插入技术，即使数据表记录数很多，数据插入速度仍然可以满足应用需求，从而解决大数据表插入速度无法满足需求的问题。

5）支持分片多复制，有效地提升了数据表的可用性。即使某些服务器或者存储设备故障，数据表仍然可以访问。

5 客户价值

1）以前无法想象的大数据实时分析现在可以进行了。在没有数字有机体工作库时，当面对一个庞大的数据表时，无法进行和其他表之间的联合统计分析。为此，不得不采取预先统计的方式，这极大地限制了数据分析人员的想象空间，从而使得数据挖掘缺乏灵活性，无法满足各种实时数据分析的要求。现在，利用数字有机体工作库提供的大数据处理机制，这些都变成可以完成的事情。

2）数据无忧了。以前，总担心大数据的存储设备损坏。如有每天进行离线备份，又需要花费大量的时间，而且离线存储介质的保管也是一个问题。要想建立在线的灾备系统又需要大量的投入。现在，数字有机体工作库可以解决这个问题。它支持在线本地备份和在线异地备份等，这可以有效地提升数据可用性。

3）数据业务的可靠性提升了。数字有机体工作库系统中的每台服务器都可以作为数据业务服务器，即使某些服务器出现故障，系统仍然可以继续提供服务。

4）可以在线保存更多的数据了。数字有机体系工作库统支持海量数据的存储，并且支持海量数据的查询，这使得可以保存更多的在线数据，从这些数据中发掘出更多的商业价值。