

一种支持大规模数据的多维可视化分析框架

游进国¹, 杨卓萃¹, 胡建华¹, 奚建清²

(1. 昆明理工大学信息工程与自动化学院, 昆明 650051; 2. 华南理工大学软件学院, 广州 510641)

摘要: 以 Hadoop 为代表的可扩展大规模数据库难以进行多维可视化分析。为此, 设计基于 B/S 架构的可视化分析框架 Bizard。数据模型通过封装底层数据接口以支持业界多维数据访问协议 XMLA, 从而在展现层易于接入支持 XMLA 的传统分析工具, 同时采用视图物化技术提高分析性能, 利用互联网技术丰富用户分析体验。实验结果表明, 该框架能在高达千万条记录级的数据上进行多维可视化分析。

关键词: 数据仓库; 可视化分析; Hadoop 软件; 大规模数据; XMLA 协议

Multi-dimensional Visualized Analysis Framework for Supporting Large Scale Data

YOU Jin-guo¹, YANG Zhuo-luo¹, HU Jian-hua¹, XI Jian-qing²

(1. School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650051, China;

2. School of Software Engineering, South China University of Technology, Guangzhou 510641, China)

【Abstract】 To address the multi-dimensional visualization analysis problem of large scale data storages with high scalability represented by Hadoop, this paper designs and implements a visualization analysis framework, called Bizard. Bizard's data model encapsulates the underlying data access interface and provides XMLA protocol for the presenting layer, which makes it convenient to use conventional analytical tools. Meanwhile, Bizard uses materialized view technologies to improve query performance and the RIA technology to enrich the user analysis experience. Experimental results show that the framework can process multi-dimensional visualization analysis on large data sets with the number of rows up to tens of millions.

【Key words】 data warehouse; visualized analysis; Hadoop software; large scale data; XMLA protocol

DOI: 10.3969/j.issn.1000-3428.2011.19.007

1 概述

随着 Web2.0 技术的迅猛发展, 生活中的各个领域, 包括通信、在线交易、社会网络、航空影像等数据密集型应用产生了大量数据。据 IDC 报告, 2006 年全球数据存储需求为 1 610 亿 GB, 预计 2010 年将达到 9 880 亿 GB^[1]。如何有效地存储和分析如此巨大的数据为学术界和工业界都带来了巨大的挑战。传统的并行模型和机制很难解决数据扩展性和容错性问题, 以 Hadoop^[2]为代表的 NoSQL^[3]大规模数据处理技术主要基于 MapReduce^[4]并行编程模式, 较好地满足了现有的数据存储和计算需求。

大规模数据必然会驱动数据可视化和商业智能分析需求, 以人们易于理解和洞察的可视化形式展示出来, 从而转变为可决策的信息。IDC 对已经实施商业智能解决方案的客户进行调研后, 得出结论: 可视化是必须的, 并且 80% 的商业智能的客户发现需要有效的可视化技术和工具^[5]。然而数据可视化, 尤其是大规模数据可视化的研究并不多。NoSQL 在许多方面与传统关系型数据库有差别, 如 Schema-Free 方式, 强调高可扩展性, 在一定程度上放松事务性要求。在性能方面, 由于数据巨大, 在线实时计算和查询难以支撑。目前, Jasper、Pentaho 等开源社区正在做 NoSQL 数据的图表展示方面的尝试, 但主要是针对关系型 SQL 查询分析的, 难以支持复杂的查询以及 OLAP 多维分析。而商业产品 MicroStrategy、Business Objects、ArcPlan、Oracle BIEE、Cognos PowerPlay、Dundas 还侧重对传统关系数据库的支持。

本文以多维数据为分析模型, 以预计算和视图物化技术

为优化手段, 以富互联网应用技术(Rich Internet Application, RIA)为界面交互方式, 着重设计了 Web 框架, 简称 Bizard, 以支持在大规模数据上通过透视表和透视图等可视化方式进行上卷、下钻、转轴、切片、切块的复杂分析操作, 进行商业智能应用。Bizard 封装 Hive/Hadoop 的接口为 XMLA 接口协议, 在浏览器端通过丰富的可视化组件和多维分析方式提高了用户体验。

2 Bizard 架构

Bizard 采用 MVC 模式, 系统从下往上主要分为数据层、应用层和展现层。

数据层包括传统关系型数据源和 NoSQL 型数据源, 其中, 传统关系型数据源有支持 XMLA 多维数据访问协议的 Microsoft SQL Server Analysis Services(SSAS)和开源的 Mondrian 服务器, NoSQL 方式有 Hive/Hadoop 以及自主设计的基于 Hadoop 的并行分布式大型数据仓库 HDW^[6]等。本文为应用层的 OLAP 模型设计了 XMLA 的数据源驱动, 将

基金项目: 云南省教育厅科学研究基金资助项目(09C0109); 云南省应用基础研究基金资助项目(2010ZC030); 广东省科技计划基金资助项目“基础软件与应用构建创新平台”(2006B80407001); 广东省国际科技合作基金资助项目(2009B050700008)

作者简介: 游进国(1977—), 男, 讲师、博士, 主研方向: 数据仓库, 并行计算; 杨卓萃, 硕士研究生; 胡建华, 副教授; 奚建清, 教授、博士生导师

收稿日期: 2011-04-12 **E-mail:** jgyou@126.com

NoSQL 数据库进行封装, 以支持 XMLA 访问接口。

应用层分为 OLAP 模型、控制器和分析视图。OLAP 模型基于 XMLA 协议设计, 提供了对 XMLA 协议的解析和响应。不同的数据源(SQL 方式和 NoSQL 方式的)被应用层的 OLAP 模型封装为多维数据模型。控制器负责请求的分发和处理, 将请求发给相应的 OLAP 模型的接口进行处理, 返回结果集交给分析视图。分析视图可为多种可视化组件, 分为透视表、透视图等组件。可视化组件可以注册为观察者, 任意一个可视化组件的分析操作, 都可能引起 OLAP 模型的变化, 并通过系统控制器通知其他作为观察者的可视化组件做相应变化。分析视图通过 HTTP 协议向客户端发送特定结构的 XML 和标准的 HTML。

展现层支持众多符合 W3C 标准的浏览器(如 Internet Explorer、Firefox 或 Opera)。Bizard 系统提供给用户的 Web 页面采用 JQuery 向应用服务器发送请求, 并通过 JQuery 接口从应用服务器端接收 XML 数据。可视化组件的客户端脚本(如透视表客户端)通过对 XML 的解析、渲染, 呈现 HTML。JQuery 接口还将处理过后的数据传递给 Flex、透视图等 RIA 图形组件进行展现。

Bizard 系统总体架构如图 1 所示。

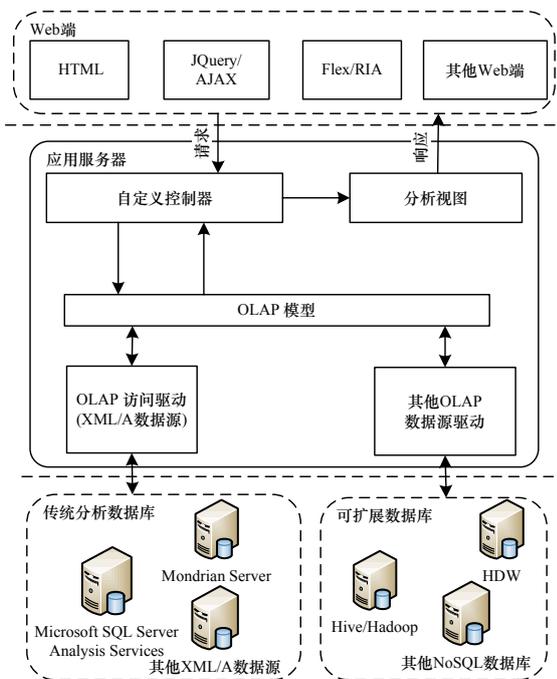


图 1 Bizard 系统架构

3 关键技术

3.1 OLAP 模型引擎

XMLA 提供了多维数据的基于 Web 服务的统一访问接口协议, 包括元数据发现接口 Discover 和命令执行接口 Execute。底层非 XMLA 数据源访问被 OLAP 模型引擎封装为 Discover 接口和 Execute 接口。

如图 2 所示, 元数据信息被保存在 Repository 库中, 该库可被保存在 Hadoop 的名字节点上或以 XML 形式保存在其他能被访问到的机器上。通过 Discover 接口获得多维模型以及星型架构的定义。OLAP 模型接收到浏览器端生成的 MDX 语句后, 由 Execute 接口处理, 其对 MDX 进行解析、计算后, 生成了 2 种中间结果形式。对于 Hive/Hadoop, 生成 SQL 语句, 交由 Hive 的 JDBC 接口执行; 对于 HDW, 解析为点查

询和范围查询的接口形式, 直接由 HDW 调用 MapReduce 执行。这 2 种形式可以分别视为对应 ROLAP 形式和 MOLAP 形式。

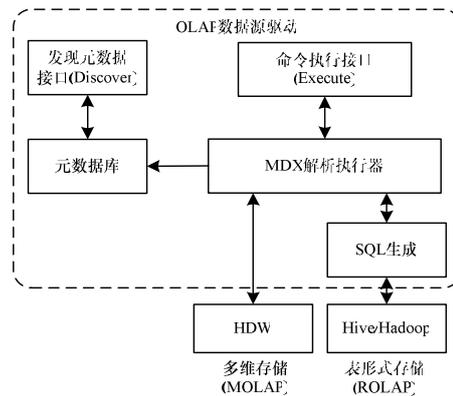


图 2 OLAP 模型引擎

Hive 的 HQL 语法对 From 子句不支持“From A, B”形式, 但支持“From A Join B”形式, 因此在生成 SQL 语句时, 需要注意 Hive 的方言。

3.2 海量数据分析的支持

NoSQL 面向的是多达千万条记录级别及以上的海量数据, 在线实时计算很难满足在线查询分析的需求。目前, Hive 更多的应用是离线计算。对此, 通过建立多维数据模型, 对一些维度属性组合的 GroupBy 进行预计算和事先物化。查询时先搜索物化视图是否可以回答或者是能否从物化视图上派生出来。如果可以, 则由物化视图给出结果; 否则, 从事实表和维表进行连接和聚合查询, 并将结果作为物化视图保存。

对于 HDW, 其采用了有效的预计算算法, 即封闭立方体算法^[6], 不仅较大地压缩数据, 并且还保留上卷、下钻语义关系。基于 Hadoop 的分布式文件系统 HDFS 和并行计算框架 MapReduce 模块, HDW 实现了封闭立方体的分布式存储和并行查询计算算法, 达到了较高的扩展性。

3.3 基于 RIA 技术的可视化分析

Bizard 不仅能够展现已有的支持 XMLA 协议的可视化分析工具, 如 JPivot、Excel/OWC 等, 并且基于 Flex 实现了多种分析图呈现形式, 采用 JQuery 实现了多维模型的树形展示以及图表的拖拉。此外, JQuery 在浏览器端负责向服务器异步发送请求和接收 XML 数据, 从而完成可视化组件的渲染和呈现。相应地, RIA 上的事件也会有消息调用的形式发给 JQuery 接口, 由该接口向 Web 服务器发送请求, 完成数据与展示同步。借助 RIA 强大的富客户端展示特性, 给用户呈现出多样的可交互的分析效果。

4 实现结果与分析

系统是基于 JEE 5 平台进行开发的, 开发语言为 Java 1.6, 集成开发环境为 Eclipse 3.5。运行环境如下: 客户端为安装浏览器的 PC 机器, 应用服务器安装 JRE 和 Tomcat 6.0 Web 服务器, 数据服务器为 Hadoop 集群, 配置有 Hadoop 0.20.0, Hive 0.6.0。

可视化组件的实现如下:

(1)透视表: 基于开源组件 JPivot 1.8.0 开发, 以利用其上卷、下钻、切片、切块及排序等分析操作, 自定义单元格颜色、图标, 可导出分析报表到 Excel 等功能。

(2)透视图: 采用 Flash Builder 4.0 平台, 基于 Flex 3.5 (下转第 31 页)