(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 104767813 B
(45)授权公告日 2018.06.08

(21)申请号 201510162567.3

(22)申请日 2015.04.08

(65)同一申请的已公布的文献号
申请公布号 CN 104767813 A

(43)申请公布日 2015.07.08

(73)专利权人 江苏国盾科技实业有限责任公司
地址 211106 江苏省南京市江宁开发区胜太路科技创业中心

(72)发明人 何颖飞

(74)专利代理机构 南京正联知识产权代理有限公司 32243
代理人 王素琴

(51)Int.Cl.
H04L 29/08(2006.01)
G06F 17/30(2006.01)

(56)对比文件
US 2013/0227558 A1,2013.08.29,
CN 103561061 A,2014.02.05,
CN 104065716 A,2014.09.24,
CN 104320460 A,2015.01.28,
梁瑜.基于Hadoop平台的医保数据挖掘.《中国优秀硕士学位论文全文数据库》.2014,(第7期),正文第1-69页.
高贵升.基于OpenStack的计算云的研究与实现.《中国优秀硕士学位论文全文数据库》.2013,(第3期),正文第1-66页,图3-2、4-3、5-1.
张仲妹.云计算环境下的资源监控应用研究.《中国优秀硕士学位论文全文数据库》.2013,(第8期),正文第1-51页.

审查员 王田园

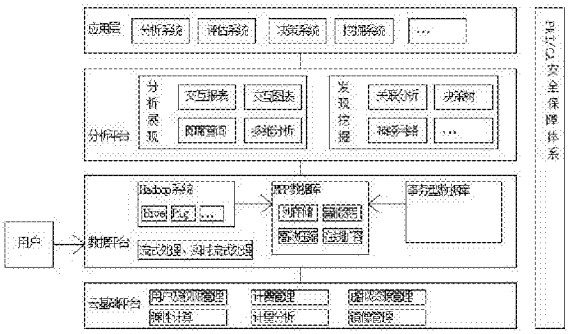
权利要求书1页 说明书9页 附图1页

(54)发明名称

基于openstack的公众行大数据服务平台

(57)摘要

本发明公开了一种基于openstack的公众行大数据服务平台，包括云基础平台、数据平台、分析平台和应用层，用户将数据传输到数据平台，分析平台通过云基础平台对数据平台的数据进行分析挖掘，应用层通过云基础平台对分析平台的数据进行应用。本发明可广泛应用于政府，企业，及各种社会组织，项目的成功实施将为我国推动“智慧城市”战略起到积极作用。



CN 104767813 B

1.一种基于openstack的公众行大数据服务平台，其特征在于：包括云基础平台、数据平台、分析平台和应用层，用户将数据传输到数据平台，分析平台通过云基础平台对数据平台的数据进行分析挖掘，应用层通过云基础平台对分析平台的数据进行应用；

所述云基础平台基于openstack的horizon模块实现云基础平台的管理系统，所述管理系统包括用户及权限管理、计费管理、虚拟资源管理、弹性计算、计量分析和镜像管理；

所述数据平台主体采用Hadoop系统、MPP数据库和事务型数据库相结合的混搭数据架构，用户将数据传输到Hadoop系统、MPP数据库或者事务型数据库，事务型数据库数据可导入MPP数据库，Hadoop系统数据可导入MPP数据库，Hadoop系统配合流式处理和实时流式处理将采集的数据进行初步处理存储入Hadoop大数据数据库中；

所述分析平台包括分析展现模块和发现挖掘模块，分析展现模块实现交互报表、交互图表、即席查询和多维分析功能，发现挖掘模块实现关联分析、决策树和神经网络功能；

所述应用层包括分析系统、评估系统、决策系统和挖掘系统的大数据应用；

所述Hadoop系统承担海量结构化数据、半结构化数据和非结构化数据分布式计算和非关系型处理，以及低价值密度结构化数据、半结构化数据和非结构化数据的存储管理，所述MPP数据库为分布式并行数据库集群，采用列存大规模分布式并行数据库集群承担复杂查询、统计和分析等OLAP分析应用的数据仓库功能，MPP数据库存储管理高价值密度的结构化数据，在系统中构建业务查询、统计专题和其他专题，实现列存储、智能索引、高效压缩和在线扩容功能，所述事务型数据库采用OLTP数据库来承担上层业务应用数据存储管理功能，用于相应的在线系统后台数据库；

在事务型数据库导入MPP数据库，Hadoop系统数据导入MPP数据库过程中，涉及到以下几种转换：数据清洗、非结构化数据结构化转换和低价值密度数据向高价值密度数据转换。

2.如权利要求1所述的基于openstack的公众行大数据服务平台，其特征在于：所述多维分析具有钻取分析、数据排序、图表显示、MDX语言查询、聚合运算、行列转换、隐藏空行/列功能。

3.如权利要求2所述的基于openstack的公众行大数据服务平台，其特征在于：还包括PKI/CA安全保障体系，PKI/CA安全保障体系提供安全基础设施。

基于openstack的公众行大数据服务平台

技术领域

[0001] 本发明涉及系统处理技术领域，特别涉及一种基于openstack的公众行大数据服务平台。

背景技术

[0002] 随着“宽带中国”、“三网融合”、“智慧城市”等信息化建设在江苏的全面推进，公众信息需求和信息消费能力不断提升，信息公共服务设施不断完善，江苏已经积累并将继续产生庞大的数据，为产业发展提供了丰富的数据资源。

[0003] 相比较较成熟的大数据企业，其多在某方面或某几方面比较专长，但其更多地专注于企业的精确化业务和科学化管理，对于政府的精确化业务如何与企业的精确化业务进行真正意义上的对接服务，怎样让政府职能部门通过为企业创造更加“实惠”的电子商务环境，怎样让企业心甘情愿接受政府职能部门的“服务”，这是个瓶颈问题。另外，各行各业的大数据企业，要么只关注于技术，比如Cloudera，拓尔斯，启明星辰，等等；要么关注技术与部分业务的结合，比如百度大数据，只是将技术跟用户搜索行为数据，公共数据结合；阿里大数据，只是将技术跟电商数据，信用数据结合；而腾讯大数据，只是将关系数据，社交数据跟大数据技术相结合。

发明内容

[0004] 本发明需要解决的技术问题是现有的大数据服务平台都是比较专业的服务平台，没有面向民生的处理各类繁杂数据以及数据应用的服务平台。

[0005] 为解决上述技术问题，本发明提供了一种基于openstack的公众行大数据服务平台，包括云基础平台、数据平台、分析平台和应用层，用户将数据传输到数据平台，分析平台通过云基础平台对数据平台的数据进行分析挖掘，应用层通过云基础平台对分析平台的数据进行应用；

[0006] 所述云基础平台基于openstack的horizon模块实现云基础平台的管理系统，所述管理系统包括用户及权限管理、计费管理、虚拟资源管理、弹性计算、计量分析和镜像管理；

[0007] 所述数据平台主体采用Hadoop系统、MPP数据库和事务型数据库相结合的混搭数据架构，用户将数据传输到Hadoop系统、MPP数据库或者事务型数据库，事务型数据库数据可导入MPP数据库，Hadoop系统数据可导入MPP数据库，Hadoop系统配合流式处理和实时流式处理将采集的数据进行初步处理存储入Hadoop大数据数据库中；

[0008] 所述分析平台包括分析展现模块和发现挖掘模块，分析展现模块实现交互报表、交互图表、即席查询和多维分析功能，发现挖掘模块实现关联分析、决策树和神经网络功能；

[0009] 所述应用层包括分析系统、评估系统、决策系统和挖掘系统的大数据应用。

[0010] 其中，所述Hadoop系统承担海量结构化数据、半结构化数据和非结构化数据分布式计算和非关系型处理，以及低价值密度结构化数据、半结构化数据和非结构化数据的存

储管理，所述MPP数据库为分布式并行数据库集群，采用列存大规模分布式并行数据库集群承担复杂查询、统计和分析等OLAP分析应用的数据仓库功能，MPP数据库存储管理高价值密度的结构化数据，在系统中构建业务查询、统计专题和其他专题，实现列存储、智能索引、高效压缩和在线扩容功能，所述事务型数据库采用OLTP数据库来承担上层业务应用数据存储管理功能，用于相应的在线系统后台数据库。

[0011] 在事务型数据库导入MPP数据库，Hadoop系统数据导入MPP数据库过程中，涉及到以下几种转换：数据清洗、非结构化数据结构化转换和低价值密度数据向高价值密度数据转换。

[0012] 更进一步的，所述多维分析具有钻取分析、数据排序、图表显示、MDX语言查询、聚合运算、行列转换、隐藏空行/列功能。

[0013] 作为本发明的进一步改进，还包括PKI/CA安全保障体系，PKI/CA安全保障体系提供安全基础设施。

[0014] 本发明选用开源的云操作系统openstack来构建云平台。OpenStack由三部分组成，分别是Nova、Swift、Glance，OpenStack可以单独提供其中的一部分，也可以将这三部分组合起来，搭建一个通用的云平台。本发明将重点研发基于openstack的自动化云平台部署技术、虚拟化资源优化管理技术以及云平台管理系统的开发与实现，为大数据应用提供一个强大计算平台。

[0015] HDFS是基于Google的Bigtable储存原理的一种开源分布式文件系统实现，它有着高容错性，能够提供高吞吐量数据访问，适合那些有着超大数据集应用程序，本发明采用HDFS分布式文件系统搭建在云平台的基础上为大数据应用提供储存，利用自主知识产权的基于hadoop中小文件优化和倒排索引算法实现高效的大数据检索，建立本体库实现不同类型的数据进行统一处理和存储，同时采用MapReduce为大数据应用提供计算框架。

[0016] 本发明基于开源云操作系统Openstack技术、国产顶尖数据库分析系统gbase 8a应用分布式文件系统HDFS和MapReduce作为大数据的存储和计算框架，通过拥有自主知识产权的基于hadoop中小文件优化和倒排索引算法实现高效的大数据检索，采用本体库对不同类型的数据进行统一处理和存储，优化了向量机分类模型以及关联规则的名词提取模式，提高计算效率和分析精度。本项目的安全方案基于PKI技术，以SAAS模式为客户提供大数据服务。

[0017] 本发明可广泛应用于政府，企业，及各种社会组织，项目的成功实施将为我国推动“智慧城市”战略起到积极作用。

附图说明

[0018] 图1是本发明的框架图。

具体实施方式

[0019] 下面详细说明本发明的优选技术方案。

[0020] 本发明的基于openstack的公众行大数据服务平台，包括云基础平台、数据平台、分析平台和应用层，用户将数据传输到数据平台，分析平台通过云基础平台对数据平台的数据进行分析挖掘，应用层通过云基础平台对分析平台的数据进行应用；

[0021] 所述云基础平台基于openstack的horizon模块实现云基础平台的管理系统，所述管理系统包括用户及权限管理、计费管理、虚拟资源管理、弹性计算、计量分析和镜像管理；

[0022] 所述数据平台主体采用Hadoop系统、MPP数据库和事务型数据库相结合的混搭数据架构，用户将数据传输到Hadoop系统、MPP数据库或者事务型数据库，事务型数据库数据可导入MPP数据库，Hadoop系统数据可导入MPP数据库，Hadoop系统配合流式处理和实时流式处理将采集的数据进行初步处理存储入Hadoop大数据数据库中；

[0023] 所述分析平台包括分析展现模块和发现挖掘模块，分析展现模块实现交互报表、交互图表、即席查询和多维分析功能，发现挖掘模块实现关联分析、决策树和神经网络功能；

[0024] 所述应用层包括分析系统、评估系统、决策系统和挖掘系统的大数据应用；

[0025] 还包括PKI/CA安全保障体系，PKI/CA安全保障体系提供安全基础设施。

[0026] 其中，所述Hadoop系统承担海量结构化数据、半结构化数据和非结构化数据分布式计算和非关系型处理，以及低价值密度结构化数据、半结构化数据和非结构化数据的存储管理，所述MPP数据库为分布式并行数据库集群，采用列存大规模分布式并行数据库集群承担复杂查询、统计和分析等OLAP分析应用的数据仓库功能，MPP数据库存储管理高价值密度的结构化数据，在系统中构建业务查询、统计专题和其他专题，实现列存储、智能索引、高效压缩和在线扩容功能，所述事务型数据库采用OLTP数据库来承担上层业务应用数据存储管理功能，用于相应的在线系统后台数据库。

[0027] 在事务型数据库导入MPP数据库，Hadoop系统数据导入MPP数据库过程中，涉及到以下几种转换：数据清洗、非结构化数据结构化转换和低价值密度数据向高价值密度数据转换。

[0028] 所述多维分析具有钻取分析、数据排序、图表显示、MDX语言查询、聚合运算、行列转换、隐藏空行/列功能。

[0029] 从数据结构化程度看，由于政府、企事业单位、民生相关数据存在多种数据，包括结构化数据、半结构化和非结构化数据。结构化数据主要为工商、税务、财政等产生的结构化数据，这些数据可以直接保存在相应的MPP数据仓库中。另外，政府、企事业单位、民生系统中同时存在相应的半结构化数据，主要包括大量的word、pdf、xml、html等文本类数据。这些数据可以通过相应的MPP数据仓库带有全文检索功能进行相应的分析挖掘。另外，系统中存在大量非结构化数据，如监测音、视频、图片等非结构化信息，这些信息可以通过实时流处理系统将这些信息导入Hadoop系统中。

[0030] 从数据价值密度上看，政府、企事业单位、民生系统中包含两类价值密度数据，即低价值密度数据和高价值密度数据。其中的高价值密度数据可以直接存入相应的MPP数据库中，进行相应的数据分析、挖掘查询。对于价值密度较低的数据类型，可以通过流处理系统，将这些数据导入MPP数据库中。MPP数据库可以方便的进行数据查询、统计，系统将Hadoop数据导入相应的MPP数据库中，便于进行标准化相关的分析、挖掘查询。

[0031] 以下为数据平台相关产品的优势和关键技术：

[0032] 1)hadoop技术

[0033] Hadoop 大数据处理平台是业界知名的开源平台，以分散存储和并行计算为基础的大数据平台，利用低成本的通用计算设备(PC)组成大型集群，构建下一代具备高性能的

海量数据分布式计算服务平台。Hadoop 符合 GNU 相关规范，属于完全开放源代码的体系架构，不仅属于完全免费模式，而且更是便于二次开发和平台定制。

[0034] 基于 Hadoop 的大数据平台主要优势如下：

[0035] 处理性能强大：采用 HDFS 分散存储和 MR 并行计算技术，通过分发数据，hadoop 可以在数据所在的节点上并行地处理它们，这使得处理非常的快速。

[0036] 投资成本低：采用开源免费模式，基于 Hadoop 大数据平台的 HDFS 分散存储、HIVE 数据仓库、HBASE 数据挖掘等组件均免费，可以节省大量的项目建设投资。

[0037] 应用广泛：Hadoop 是目前应用最广泛云计算平台，在电信运营商、互联网厂商、银行相关企业等行业获得广泛应用。Yahoo 已经建设了规模为 45000 个节点的 Hadoop 平台，淘宝也已经建设了超过 1100 个节点的 Hadoop 云平台。

[0038] 2)MPP分布式数据库

[0039] GBase 8a MPP Cluster适合替代现有关系数据结构下的大数据处理，具有较高的效率。本项目的存储平台MPP 数据库采用分布式数据库集群 GBase 8a MPP Cluster。GBase 8a MPP Cluster 是一款基于 Shared Nothing + MPP 扁平架构的分布式并行数据库集群，具备高性能、高可用、高扩展特性，可以为超大规模数据管理提供高性价比的通用计算平台。GBase 8a MPP Cluster采用列存储、MPP 集群、分布式文件系统、全文检索、高效数据压缩、智能索引等新技术，达到与国际上领先的 NewSQL 数据库产品的技术同步。GBase 8a MPP Cluster 的数据处理规模达到 PB 级以上，实现对结构化、半结构化和非结构化数据进行统一处理。GBase 8a MPP Cluster产品为本项目提供了新型数据库产品基础及分布式集群架构，其全文检索系统为本项目奠定了半结构化数据查询分析高性能的基础。实时备份和恢复系统与数据抽取转化加载管理系统作为成熟的外围工具，使本项目更加完备、易用，有益于促进实现本项目的产业化推广。

[0040] 通过研究大规模并行处理、分布式存储、分布式文件系统、Shared Nothing 集群、全文检索、高效数据压缩、智能索引等技术，实现了集群模式下对结构化、半结构化和非结构化数据进行统一处理的列存数据库产品。该产品具备高性能、高可用、高扩展特性，在技术上达到与国际领先的 NewSQL 数据库产品的技术同步，主要技术性能指标与国外产品相当。

[0041] 系统在现有结构化 RDBMS 的基础上，扩展非结构化数据引擎，并在产品内部集成分布式文件管理模块，对于不同类型的数据文件提供统一的存储管理。实现列存储数据库集群存储结构化数据，半结构化和非结构化数据存储在分布式文件系统上。该产品实现数据处理规模达到 PB 级以上。

[0042] 分布式并行数据库集群是在 GBase 8a 列存储数据库基础上开发的，基于现代云计算理念和 shared nothing 架构的并行数据库集群，可支持 TB 到 PB 级别结构化数据存储查询，高性能、高可用、高扩展的分布式、并行的数据库系统。以其独特的扁平架构，高可用性和动态扩展能力，为超大型数据管理提供一个高性价比的通用平台。

[0043] 分布式并行数据库集群主要具备以下特征和优势：

[0044] 真正的列存储RDBMS体系架构；

[0045] 支持海量数据存储、查询；

[0046] 数据分布的灵活性：基于策略的数据加载模式；

[0047] 数据加载高效:装载数据速度大于2TB/小时;

[0048] 压缩优势:缺省为轻量级数据压缩,加载后数据不膨胀,启动高级压缩后可以达到1:10以上的综合压缩比,压缩状态下查询性能不下降;

[0049] 可扩展性:单个集群可达到128个节点,支持PB级的数据库存储;

[0050] 并发特性:读写没有互斥, 支持MVCC, 支持边数据入库边查询,支持2000以上并发用户;

[0051] 并行特性:充分利用现代多核CPU资源;

[0052] 集群架构:扁平架构, 无单点故障(SPOF),无master瓶颈,具有高扩展性;

[0053] 集群调度占用资源少:网络带宽需求小;

[0054] 可靠性:支持全量,增量备份/恢复;

[0055] 易用性:不用特殊索引,调优,物化视图等;

[0056] 易于维护:支持集群在线扩展;

[0057] 高可用:支持数据冗余,自动故障探测和管理,自动fail over,自动同步;

[0058] 高效率:提供智能索引为统计分析查询提供高效率;

[0059] 安全,监控能力:支持用户权限管理,提供图形化管理诊断工具。

[0060] 分布式并行数据库集群主要实现以下技术特性:

[0061] Shared Nothing+MPP架构

[0062] 远程数据加载:远程数据加载工具Remote Loader 实现了数据库的高速数据加载功能,DBA在进行数据仓库维护的时候,不再需要将源数据库中的数据导出形成数据文件,再通过手工的方式进行数据导入,现在只需要进行简单的配置,远程高速数据加载工具就会自动完成数据的导入加载功能。Remote Loader支持多种数据源,包括Oracle、DB2、Sybase IQ、SQL Server等,能够实现绝大多数数据库系统的数据加载。

[0063] 主要突破一下技术难点:

[0064] 数据分片存储:海量数据存储的一大特征是事实表超大,有些事实表的记录数能够达到几百亿条,这就要求数据存储空间非常大,有些能达到几十TB,传统的数据库无法对超大的数据表进行存储,更无法实现数据分析功能。分布式并行数据库集群可以将超大的数据表进行数据和索引分片进行存储,数据分片方法可采用(range、round robin、hash)等不同方法,适用于不同的场景,数据与其相关的索引的分片存储在一个节点中。这项技术带来的优势是能够实现更大数据规模的存储,同时能够发挥多节点并行计算的优势,提升数据查询分析性能。数据的存储支持水平进行分区,分区条件可为range、key、hash。

[0065] 大表关联:大表关联用于对存在于源节点服务器和目的节点服务器中的表进行查询,其中源节点服务器接收协调节点服务器发送的查询命令,该查询命令中包含从原始查询语句中筛选出的用于查询所述源节点服务器中的表的查询条件,根据查询命令执行查询操作,并将查询结果数据写入临时文件,将临时文件传输到目的节点服务器,目的节点服务器将临时文件中的查询结果数据加载到临时表中,对临时表及存在于目的节点服务器中的表进行连接查询。应用本申请实施例进行跨节点数据查询时,通过对源节点中的表进行筛选,并将筛选结果发送到目的节点进行连接查询,由此提高了跨节点数据查询的性能;并且,通过利用外部数据传输工具及批量导入工具,提高了数据传输性能和数据导入性能,以及通过可配置的方式,提供了跨节点数据查询的可扩展性。

[0066] 列存数据引擎:集群中的列存数据引擎技术包括结构化数据和半结构化数据引擎,采用读写分离、结构化数据和非(半)结构化数据引擎同步机制,保障海量数据的快速读写性能。

[0067] 高效的透明压缩:集群具备压缩功能,并实现透明性,压缩比能够达到 1:10 甚至更优。数据库允许用户根据需要设置配置文件,选择是否进行压缩。在启用压缩的情况下分析型关系数据库根据数据的不同特性以及不同的分布状况,自动采用相应的压缩算法。

[0068] 智能索引:集群具备智能索引技术,将智能索引建立在数据包上(粗粒度索引。智能索引中包含了描述数据间相互依赖关系的高级信息,有效的解决复杂的多表连接和子查询,能够准确识别数据包的需要,最大限度地减少磁盘 I/O。并且智能索引所占空间小,大约是数据的百分之一。

[0069] 分布式查询、调度:本项目的分布式查询、调度包括:简单查询、两表JOIN查询、两表JOIN 查询计划和Limit查询。

[0070] 高可用机制:本项目为保证系统的高可用性,提供多副本机制,例如Tp表的复本Tp`可根据应用和资源需求提供副本数量的自动化管理。提供副本管理工具副本数量和位置的调整。Tr表会部署在各个节点上。数据进行分片,Tp表被分片成Tp1、Tp1、Tp3,被存储在不同节点上。

[0071] 3)流式处理:在流数据不断变化的运动过程中实时地进行分析,捕捉到可能对用户有用的信息,并把结果发送出去,如 SystemS、Storm、S4 等,其中大数据最被最广泛使用的是 Storm。

[0072] Storm 是 Twitter 开源的分布式实时计算系统, Storm 通过简单的 API 使开发者可以可靠地处理无界持续的流数据,进行实时计算,开发语言为 Clojure 和 Java,非 JVM 语言可以通过 stdin/stdout 以 JSON 格式协议与 Storm 进行通信。Storm 的应用场景很多:实时分析、在线机器学习、持续计算、分布式 RPC、ETL 处理等。

[0073] Storm 被广泛应用于实时分析,在线机器学习,持续计算、分布式远程调用等领域。其优势如下:

[0074] 实时分析用户的属性,并反馈给搜索引擎。最初,用户属性分析是通过每天定时运行的MR job 来完成的。为了满足实时性的要求,希望能够实时分析用户的行为日志,将最新的用户属性反馈给搜索引擎,能够为用户展现最贴近其当前需求的结果。

[0075] 实时分析系统监控网站性能。利用 HTML5 提供的 performance 标准获得可用的指标,并记录日志。Storm 集群实时分析日志和入库。使用 DRPC 聚合成报表,通过历史数据对比等判断规则,触发预警事件。

如果,业务场景中需要低延迟的响应,希望在秒级或者毫秒级完成分析、并得到响应,而且希望能够随着数据量的增大而拓展。使用 Storm 就再合适不过了。

[0076] 除了低延迟,Storm 的 Topology 灵活的编程方式和分布式协调也会给本项目带来方便。

[0077] 属性分析的项目,需要处理大量的数据。使用传统的 MapReduce 处理是个不错的选择。但是,处理过程中有个步骤需要根据分析结果,采集网页上的数据进行下一步的处理。这对于MapReduce 来说就不太适用了。但是,Storm 的 Topology 就能完美解决这个问题。

[0078])事务型数据库:事务型数据库是建立在关系数据库模型基础上的数据库,借助于集合代数等概念和方法来处理数据库中的数据,同时也是一个被组织成一组拥有正式描述性的表格,该形式的表格作用的实质是装载着数据项的特殊收集体,这些表格中的数据能以许多不同的方式被存取或重新召集而不需要重新组织数据库表格。事务型数据库的定义造成元数据的一张表格或造成表格、列、范围和约束的正式描述。每个表格(有时被称为一个关系)包含用列表示的一个或更多的数据种类。每行包含一个唯一的数据实体,这些数据是被列定义的种类。当创造一个事务型数据库的时候,用户定义数据列的可能值的范围和可能应用于那个数据值的进一步约束。而SQL语言是标准用户和应用程序到关系数据库的接口。其优势是容易扩充,且在最初的数据库创造之后,一个新的数据种类能被添加而不需要修改所有的现有应用软件。

[0079] 事务型数据库技术较为成熟,该项目中采用较为先进的开源数据库或国产事务型数据库。

[0080] 事务型数据库在对性能要求比较高时,可以采用内存数据库。

[0081] 由于平台中存取的数据有一定的保密性要求,在搭建数据平台选用数据过程中,该项目应着重强调事务型数据库的安全性。

[0082] 3、分析平台

[0083] 1)分析展现

[0084] 大数据可视化基于南大通用的GBase BI实现交互式图表、即席查询、多维分析等可视化技术。展现平台灵活易用,可以通过拖拽式形成各类型展现图形,支持多种图形显示。

[0085] 即席查询

[0086] 简单快速灵活的数据查询工具。用户通过简单拖拽树状数据源模型树,任意组合查询内容及查询条件,选择多种函数的运算,就可图文并茂展现查询结果,并可快速生成报表,打印图和报表。即席查询支持复杂的查询条件基本覆盖SQL功能,用户可以根据业务需要,任意组合查询条件。操作简单,通过拖拽可以使用户快速的查询出所需数据。表格的格式设置,根据需求选择设置数据、单位、文字、表格等格式。数据范围预警,用户设置预警数据范围,表格自动以不同颜色区分显示预警数据。图形数据实时转换,圈选表格数据,将数据转换成统计图形,更直观的表现数据。多种自定义函数,可以对数据进行总和、平均、最小值、最大值、计数、标准差、方差、上期、本期、同期、同比、同比增长率、同比发展率、环比、环比增长率、环比发展率、Top N等多种函数运算。数据和图形打印,图表与表格数据可以同时打印。报表输出,导出各种格式,如:Excel、PDF、CSV。多优先级数据排序,可以设置多列按优先级排序。

[0087] 交互式图表

[0088] 基于web的交互式图表的绘制工具。支持多模型、多数据源,图形间的联动、钻取,多种函数运算,丰富的图表种类、多种颜色、格式、文字、坐标轴、图例、背景等属性的设置,简单拖拽制作的过程,用户可以制作在ICTD这个工作舞台上,随心所欲的绘制出任意风格的图表,得到想要的分析结果。

[0089] 支持与Ad hoc相同的数据过滤和查询条件,不包括明细查询,所需数据和关键KPI指标采用最适合的图形或者格式显示,使用ICTD可以制作出非常专业的图形化数据分析报

告。支持丰富的图形模板，迎合分析不同种类的数据需求。支持树状数据源，支持连接不同的数据源。支持多种自定义函数，包括同比、环比、同比增长率、环比增长率等。支持图形的数据联动及钻取，动态的分析数据。支持图形的风格类型转换，包括图表风格及图表类型两种转换。具有预警设置，适宜预警提醒的仪表盘及温度计等图表进行预警设置。多轴展示，图表中可同时显示多个不同数量级的度量值。多图表混合布局，数据展现形式更丰富灵活。

[0090] 多维分析

[0091] 数据深度挖掘的分析工具，通过上卷、下钻、钻取明细等操作，实现对数据的动态深入分析。以此来为决策者提供依据。多维分析技术具有钻取分析、数据排序、图表显示、MDX语言查询、聚合运算、行列转换、隐藏空行/列等功能。

[0092] 2)发现挖掘

[0093] 本课题所研究的大数据管理分析平台在对海量多源异构数据进行高效存储的基础上，通过大数据分析挖掘技术对数据的潜在的价值进行分析、挖掘和展示。平台基于南大通用商业智能系统GBase BI完成数据的分析展示，并进一步融合了数据挖掘的建模分析和挖掘算法等关键技术。

[0094] 大数据管理分析平台通过数据挖掘实现对平台所存储数据的潜在价值的深度挖掘，通过数据对象化技术建立模型，并使用平台强大的计算能力对数据的模型进行更新维护，保证平台能够具有对多源异构数据的良好分析挖掘性能。

[0095] 大数据管理分析平台针对所存储的海量数据，采用数据分类、估计、预测、相关性分组、聚类与描述，实现对结构化数据、半结构数据和非结构化数据的自我认知过程。平台采用面向对象的三螺旋模型(虚虚关联，虚实关联，和实实关联)建立动态的数据关联形态，通过简单关联、时序关联和因果关联规则，找出海量多源异构数据中隐藏的关联关系网，实现数据的统一描述。

[0096] 数据关联分析和数据挖掘主要由数据存储过程和分布式MapReduce计算框架实现，数据关联的思想采用“面向对象”的数据组织，具体通过三螺旋模型(虚虚关联，虚实关联，和实实关联)建立动态的数据关联形态，依托于服务总线，实现与存储层的数据交互工作。

[0097] 关联规则是一种简单，实用的分析规则，它描述了一个事物中某些属性同时出现的规律和模式，是数据挖掘中最成熟的主要技术之一。关联规则在数据挖掘领域应用很广泛，适合于在大型数据集中发现数据之间的有意义关系。大多数关联规则挖掘算法能够无遗漏发现隐藏在所挖掘数据中的所有关联关系，但是，并不是所有通过关联得到的属性之间的关系都有实际应用价值，要对这些规则进行有效的评价，筛选有意义的关联规则。

[0098] 4.应用层

[0099] 应用层完成相关分析系统、评估系统、决策系统、挖掘系统等相关大数据应用。

[0100] 系统可以在金融、电力、智能交通、电子政务、司法系统等发挥相应的分析、挖掘、决策支持能力。具体可以根据相应系统数据和业务进行。

[0101] 5.PKI/CA

[0102] PKI技术提供安全基础设施。PKI技术采用证书管理公钥，通过CA把用户的公钥和用户的其他标识信息绑定在一起，使用户可以在多种应用环境下方便的使用加密和数字签名技术，从而保证网上数据的机密性、完整性和有效性，同时采用静态职责分离RBAC的访问

控制机制，强化应用对大数据平台访问的私密性和安全性。

[0103] 在研发过程中，采用开源加自研（在开源数操作系统openstack的基础上，利用hdfs,mapreduce, PKI等技术，先研究确认国际主流的各种数值和分析算法，实现基于大数据的搜索，关联规则，向量机分类模型等数据挖掘算法，最终完成基于openstack的大数据服务平台）的技术路线。

[0104] 在开发过程中利用开源云操作系统openstack，开源数据挖掘软件weka，mhout，具有知识产权的搜索，MPP数据库，挖掘算法，以及满足国密要求的PKI安全技术，形成开放，安全的大数据服务平台，具有引导大数据产业发展的示范意义。

[0105] 项目完成后可以提供城市基础设施实时监测与分析、城市交通状况分析、疫情趋势分析、全面的客户数据分析、产品分析及推荐、广告优化与投放、企业动态绩效分析和风险评估等方面的服务。

[0106] 本发明的大数据服务平台面向全省政府、企业、民众提供各类资讯、社交、电子商务、智慧应用的统一服务，覆盖“智慧政务”、“智慧民生”、“大数据”、“智慧交通”、“智慧物流”、“智慧医疗”、“智慧教育”、“智慧家政”、“智慧工商”、“智慧农业”、“智慧物业”、“智慧……”等多项智慧应用。通过江苏省公共服务平台，我们将拥有全行业的各类数据，同时拥有部分政府数据的运营权。借助大数据技术，通过对这些数据的挖掘，我们就能够发现更有意义的信息，从而更好的支撑行业各企业的发展，更好的促进政府由管制型政府向服务型政府的转变。同时，借助我司具有知识产权的安全解决方案---证书应用服务平台，我们为大数据安全提供了一条非常安全的解决方案。总之，我们的解决方案无论从安全，还是商业模式上来看，都是唯一的。

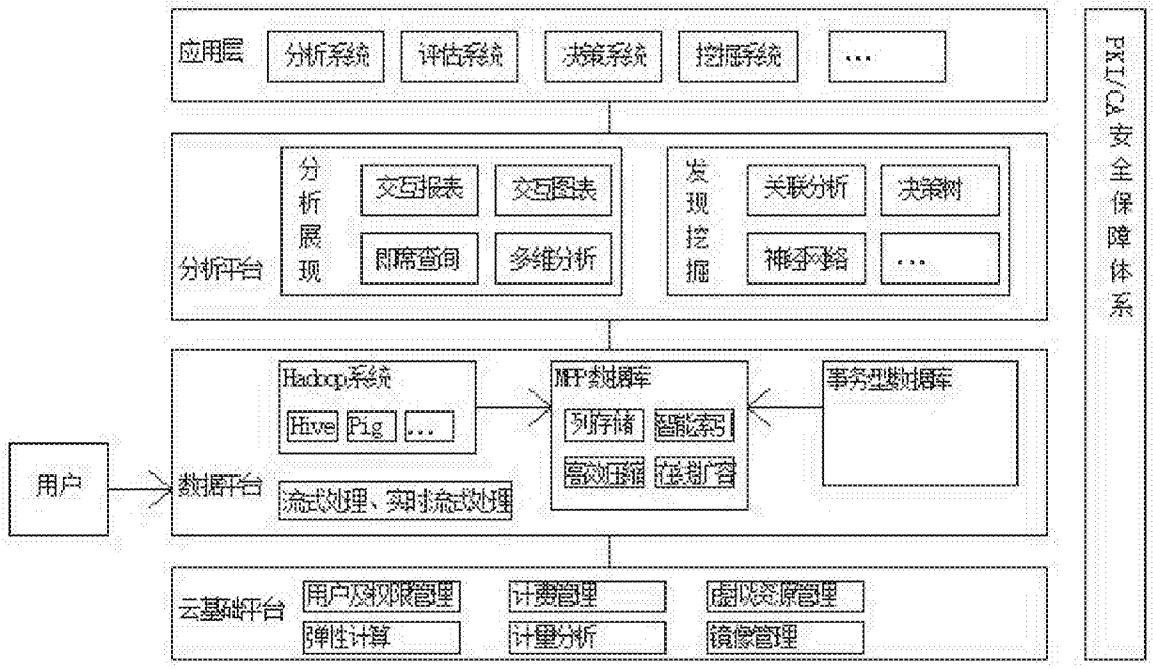

应用层
分析系统
评估系统
决策系统
挖掘系统
…
分析平台
分析展现
交互报表
交互图表
即席查询
多维分析
发现挖掘
关联分析
决策树
神经网络
…
PKI/CA 安全保障体系
用户
数据平台
Hadoop系统
Hive
Pig
…
MPP数据库
列存储
智能索引
事务型数据库
流式处理、实时和准实时处理
云基础平台
用户及权限管理
计费管理
虚拟资源管理
弹性计算
计量分析
镜像管理

图1