(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 102685221 B
(45)授权公告日 2014.12.03

(21)申请号 201210130726.8

(22)申请日 2012.04.29

(73)专利权人 华北电力大学(保定)
地址 071003 河北省保定市永华北大街619号

(72)发明人 王德文 宋亚奇 肖磊 肖凯

(74)专利代理机构 石家庄冀科专利商标事务所有限公司 13108
代理人 李羡民 高锡明

(51)Int.Cl.
H04L 29/08(2006.01)
H04L 12/26(2006.01)

(56)对比文件
CN 101800440 A,2010.08.11, 全文.
CN 101917067 A,2010.12.15, 全文.
CN 102368634 A,2012.03.07, 全文.
梁竹靓,石超.基于CORBA技术的分布式电力监控系统的设计.《电力系统保护与控制》.2008,第36卷(第17期),67-70,93.
韩如月,李俊刚,宋小会,魏勇,狄军峰.输变电设备状态监测系统设计.《高压电器》.2012,第48卷(第1期),58-63,69.

审查员 唐文森

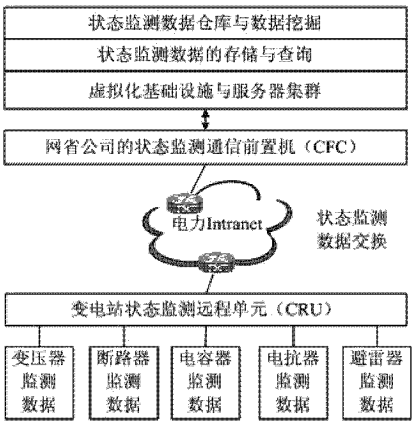
权利要求书2页 说明书7页 附图2页

(54)发明名称

一种状态监测数据的分布式存储与并行挖掘方法

(57)摘要

一种状态监测数据的分布式存储与并行挖掘方法,它通过Web服务描述语言定义变电站状态监测远程单元与状态监测通信前置机的功能服务模型,通过简单对象访问协议进行电力广域网环境下电力设备状态监测数据交换;它将大规模状态监测数据冗余存储在分布式文件系统中,对状态监测数据文件创建索引表,并插入到大规模结构化数据表中,根据查询请求完成状态监测数据查询;通过提取、转换与加载生成基础数据与多维度的分析性数据建立数据仓库,并通过MapReduce任务分解与结果汇总,实现关联规则、分类和聚类数据挖掘算法的并行执行。本发明能有效地对智能电网环境下海量电力设备状态监测信息进行分布式数据交换、冗余存储与快速并行处理。



CN 102685221 B

1. 一种状态监测数据的分布式存储与并行挖掘方法，其特征是，通过 Web 服务描述语言定义变电站状态监测远程单元与状态监测通信前置机的功能服务模型，通过简单对象访问协议进行电力广域网环境下电力设备状态监测数据交换；将大规模状态监测数据冗余存储在分布式文件系统中，对状态监测数据文件创建索引表，并插入到大规模结构化数据表中，根据查询请求完成状态监测数据查询；通过提取、转换与加载生成基础数据与多维度的分析性数据建立数据仓库，并通过映射与化简并行编程模型进行任务分解与结果汇总，实现关联规则、分类和聚类数据挖掘算法的并行执行；

状态监测数据仓库建立与数据挖掘并行化处理步骤如下：

a. 状态监测数据的抽取：对于现有的长期存储在关系型数据库中的电力设备历史状态数据，经过数据净化、转换、标准化后，以文件的形式存储于 HDFS 的数据结点上；

b. 状态监测数据仓库的建立：首先采用 Hive 查询语言 HiveQL 创建表，表的定义、字段以及间隔符信息均存储于元数据库中，然后加载 HDFS 数据文件到表以构造数据文件目录；根据变电站、设备类型、监测类型与时间状态监测主题组织成分区，按照列属性将数据组织成数据桶；

c. 状态监测数据分析：客户端发起状态监测数据分析请求，根据请求命令的内容查询元数据库中对应的表模式，若满足则进入数据文件目录查询相应的表，通过 HiveQL 找到状态监测量字段，获取满足条件的状态量值，进行聚类、求和、汇总、报表生成操作，最后，将操作生成的查询分析计划存储在 HDFS 数据仓库中，并将状态数据分析结果返回给客户端；

d. 状态监测数据挖掘的并行化：将包括关联规则、分类和聚类的算法运行分发给作业进程管理下的各个任务进程共同完成；设置并行化引擎实例，通过映射与化简并行编程模型 MapReduce 将学习过程中的大规模数据集运算分割为若干训练子集分配给多个映射节点 Mapper，在映射节点上分别执行各种操作得到中间结果，最后通过化简节点 Reducer 将结果合并，实现算法的并行执行。

2. 根据权利要求 1 所述状态监测数据的分布式存储与并行挖掘方法，其特征是，状态监测数据交换的具体方法为：

a. 状态监测服务接口的定义：采用 Web 服务描述语言 WSDL，为变电站状态监测远程单元 CRU 定义抽象服务接口 CRUServiceInterface 以及控制命令操作，为状态监测通信前置机 CFC 定义抽象服务接口 CFCServiceInterface 以及上传心跳信息、配置信息、状态监测数据操作；

b. 状态监测服务消息与参数的描述：为上述操作定义请求消息与响应消息，并约束输入与返回参数的数据类型，其中心跳信息请求消息输入参数包括标识符、工作状态与网络状态，配置信息请求消息输入参数包括标识符、数据上传周期与配置参数，状态监测数据请求消息输入参数包括监测数据代码、监测数据值与告警状态；

c. 状态监测信息传输方式与消息格式的定义：将 CRUServiceInterface 与 CFCServiceInterface 服务接口绑定为简单对象访问协议 SOAP，传输方式采用超文本传送协议 HTTP，并采用 document/literal 作为状态监测信息的编码方式；

d. 访问端点的部署绑定： 为 CRUServiceInterface 与 CFCServiceInterface 指定特定网络地址来定义访问端点，通过该地址访问所提供的状态监测服务；

e. 状态监测数据的交换过程如下：

① CRU 处于堵塞状态，周期性主动唤醒后，发起调用远程对象 CFC 的状态监测数据服务的远程过程调用 RPC 请求；

② 状态监测数据服务的 RPC 请求被封装成一个采用结构化描述语言 XML 编码的 SOAP 请求消息，发送到 CFC 的 SOAP 服务器上；

③ CFC 的 SOAP 服务器解码收到的 SOAP 请求消息，对变压器、断路器与容性设备的状态监测数据进行业务逻辑处理，判断是否存在缓存的尚未发出的配置与控制命令，再将处理结果封装成 SOAP 响应消息；

④ CRU 获得状态监测数据服务的响应消息后，判断是否执行配置与控制命令。

3. 根据权利要求 2 所述状态监测数据的分布式存储与并行挖掘方法，其特征是，状态监测数据的存储与查询的具体步骤如下：

a. 从 CFC 收集的状态监测数据以文件形式组织，直接将数据以二进制的形式存放到文件里，不包含任何的冗余数据，将数据转化为便于查询的结构化形式，读取状态监测数据文件，逐行扫描每个状态监测数据记录；

b. 将文件扫描检测和索引创建分布在不同节点上，设置主节点服务器对状态文件检测和索引创建，检测是否产生新的状态监测文件，将新的状态监测文件名整合成索引创建请求，并分发给子节点处理，如果子节点失效，转移到其他子节点上，子节点部分检测请求是否到来以及是否为合理，每当子节点接收到一个状态监测数据文件索引创建任务，将从 Hadoop 分布式文件系统 HDFS 中读取的状态监测文件数据读入内存中，并记录该文件的名称；

c. 对文件中的每个状态监测数据记录逐行扫描，提取出对查询有效字段，添加到列表中，根据这些常用字段建立索引表；

d. 将状态监测数据文件产生的索引表插入到分布式列存储的 Hadoop 结构化数据表 HBase 中，接受并处理用户的状态数据查询请求，并检测该请求是否合理，查询遍历索引表；

e. 索引表中行键为查询字段，偏移量为状态监测数据记录在状态数据文件中的位置，即文件名加偏移量，查询将通过文件名和偏移量来获取数据，一张表的行键按照字节序顺序排序，对于指定查询条件，拼接成合理的查询字节序，通过直接定位到行键或者行键的上一个行键，快速获得满足条件的状态监测索引数据，读取后续的数据，获得满足条件的状态监测数据位置信息；当行键不满足时，则查询索引完毕；

f. 根据所获得所有满足条件的状态监测位置信息集合，从状态数据文件中读出所有的状态监测数据记录，将查询结果返回给客户端。

一种状态监测数据的分布式存储与并行挖掘方法

技术领域

[0001] 本发明涉及一种智能电网海量状态监测数据的分布式存储与并行挖掘方法，属数据处理技术领域。

背景技术

[0002] 随着大规模波动式能源发电与高渗透率分布式电源的大量接入、负荷特性的日趋复杂，电网规模越来越大，电网安全、稳定运行所面临的压力也越来越大。从智能电网的发展策略和建设进展可以看出，尽管各国智能电网的功能特性、关键技术和建设重点不尽相同， 但是实现电网信息化，即全面整合电网稳态、动态、暂态运行信息，建设基于全景数据的分析与计算平台，为智能电网各类业务应用提供支持和服务，使电力企业的管理模式从分散化到集中化进行转变，则是各国智能电网的基本特征之一。

[0003] 伴随着特高压电网的建设、可再生能源和分布式能源的不断接入，电网规模将急剧增大。随着传感测量、物联网以及通信等技术的不断发展，电网数据的采样频率将明显提高、采集范围将极大扩展、电网运行数据规模将急速增长。电力设备状态监测装置所采集的实时数据将积累出海量的时间序列历史数据。智能电网的状态监测数据具有广域、全景、海量、实时、准确可靠的特征，远远超出了传统电网状态监测的范畴，它不仅涵盖一次系统设备，还囊括了二次系统设备；不仅包括实时在线状态数据，还应包括设备基本信息、试验数据、运行数据、缺陷数据、巡检记录、带电测试数据等离线信息，面对这些海量的、分布式的、异构的、复杂的状态数据，常规的数据存储与管理方法会遇到极大的困难，现有的数据分析与处理能力不足以支撑智能电网状态信息的分析优化与辅助决策。仅以绝缘子泄漏电流监测为例，假设 10ms 采集一次数据，一个杆塔在一个月内就达到了 2.5 亿条，对于关系数据库来说，在一张 2.5 亿条记录的表里面进行 SQL 查询，效率是极其低下乃至不可忍受的。

[0004] 目前，一般以 Oracle、Sybase 等标准商用数据库与数据仓库存储历史数据，这种体系结构仍然保持了传统的数据库管理系统的特点，存储的是相对静止的数据，而对于存储变化快、连续、海量的时序数据的管理能力是非常有限的。虽然可以采用实时库和历史库相结合的方式，在标准商用数据库平台上外挂实时数据库，用来管理内存实时数据，历史数据文件是以存档文件的形式存在。由于实时数据库大多由厂商自行开发，并且采用各自的专用接口、互不兼容，给系统的二次开发、异构系统的集成、数据共享与管理造成了极大困难。

[0005] 研究人员采用数据流、并行计算、分布式计算以及网格计算等技术对电网运行数据的高效查询、高性能的分析与挖掘进行了大量研究工作。目前，数据流的处理算法与降载策略还没有解决，应用理论体系尚不成熟完善，数据流管理系统仍停留在原型系统的研发阶段，例如 Stanford 大学的 STREAM 项目、UC Berkeley 大学的 Telegraph CQ 项目以及 Aurora 项目等。网格计算曾一度被认为是提升电力系统分析与计算能力的有效技术，但是网格计算主要侧重于聚合分布的松散耦合资源、强调资源共享，适用于计算密集型的应用、难以自动扩展，网格的构建大多为完成某一个特定的任务需要，或者支持挑战性的应用，通

常被用来解决计算敏感型的科研、数学、学术问题，对企业应用的支持不够，限制了其在电网企业的大规模应用。

[0006] 云计算是一种新兴的计算模型，具备可靠性高、数据处理量巨大、灵活可扩展以及设备利用率高等优势，正成为信息领域研究的热点，给上述问题的解决带来了机遇。

发明内容

[0007] 本发明的目的在于克服现有技术的不足、提供一种状态监测数据的分布式存储与并行挖掘方法，实现智能电网环境下海量电力设备状态监测信息的分布式数据交换、冗余存储管理、快速查询与处理。

[0008] 本发明所称问题是以下述技术方案实现的：

[0009] 一种状态监测数据的分布式存储与并行挖掘方法，它通过 Web 服务描述语言定义变电站状态监测远程单元与状态监测通信前置机的功能服务模型，通过简单对象访问协议进行电力广域网环境下电力设备状态监测数据交换；它将大规模状态监测数据冗余存储在分布式文件系统中，对状态监测数据文件创建索引表，并插入到大规模结构化数据表中，根据查询请求完成状态监测数据查询；通过提取、转换与加载生成基础数据与多维度的分析性数据建立数据仓库，并通过映射与化简并行编程模型进行任务分解与结果汇总，实现关联规则、分类和聚类数据挖掘算法的并行执行。

[0010] 上述状态监测数据的分布式存储与并行挖掘方法，状态监测数据交换的具体方法为：

[0011] a. 状态监测服务接口的定义：采用 Web 服务描述语言（WSDL），为变电站状态监测远程单元（CRU）定义抽象服务接口 CRUServiceInterface 以及控制命令操作，为状态监测通信前置机（CFC）定义抽象服务接口 CFCServiceInterface 以及上传心跳信息、配置信息、状态监测数据操作；

[0012] b. 状态监测服务消息与参数的描述：为上述操作定义请求消息与响应消息，并约束输入与返回参数的数据类型，其中心跳信息请求消息输入参数包括标识符、工作状态与网络状态，配置信息请求消息输入参数包括标识符、数据上传周期与配置参数，状态监测数据请求消息输入参数包括监测数据代码、监测数据值与告警状态；

[0013] c. 状态监测信息传输方式与消息格式的定义：将 CRUServiceInterface 与 CFCServiceInterface 服务接口绑定为简单对象访问协议（SOAP），传输方式采用超文本传送协议（HTTP），并采用 document/literal 作为状态监测信息的编码方式；

[0014] d. 访问端点的部署绑定：为 CRUServiceInterface 与 CFCServiceInterface 指定特定网络地址来定义访问端点，通过该地址访问所提供的状态监测服务；

[0015] e. 状态监测数据的交换过程如下：

[0016] ① CRU 处于堵塞状态，周期性主动唤醒后，发起调用远程对象 CFC 的状态监测数据服务的远程过程调用（RPC）请求；

[0017] ② 状态监测数据服务的 RPC 请求被封装成一个采用结构化描述语言（XML）编码的 SOAP 请求消息，发送到 CFC 的 SOAP 服务器上；

[0018] ③ CFC 的 SOAP 服务器解码收到的 SOAP 请求消息，对变压器、断路器与容性设备的状态监测数据进行业务逻辑处理，判断是否存在缓存的尚未发出的配置与控制命令，再

将处理结果封装成 SOAP 响应消息；

[0019] ④ CRU 获得状态监测数据服务的响应消息后，判断是否执行配置与控制命令。

[0020] 上述状态监测数据的分布式存储与并行挖掘方法，状态监测数据的存储与查询的具体步骤如下：

[0021] a. 从 CFC 收集的状态监测数据以文件形式组织，直接将数据以二进制的形式存放到文件里，不包含任何的冗余数据，将数据转化为便于查询的结构化形式，读取状态监测数据文件，逐行扫描每个状态监测数据记录；

[0022] b. 将文件扫描检测和索引创建分布在不同节点上，设置主节点服务器对状态文件检测和索引创建，检测是否产生新的状态监测文件，将新的状态监测文件名整合成索引创建请求，并分发给子节点处理，如果子节点失效，转移到其他子节点上，子节点部分检测请求是否到来以及是否为合理，每当子节点接收到一个状态监测数据文件索引创建任务，将从 Hadoop 分布式文件系统（HDFS）中读取的状态监测文件数据读入内存中，并记录该文件的名称；

[0023] c. 对文件中的每个状态监测数据记录逐行扫描，提取出对查询有效字段，添加到列表中，根据这些常用字段建立索引表；

[0024] d. 将状态监测数据文件产生的索引表插入到分布式列存储的 Hadoop 结构化数据表（HBase）中，接受并处理用户的状态数据查询请求，并检测该请求是否合理，查询遍历索引表；

[0025] e. 索引表中行键为查询字段，偏移量为状态监测数据记录在状态数据文件中的位置，即文件名加偏移量，查询将通过文件名和偏移量来获取数据，一张表的行键按照字节序顺序排序，对于指定查询条件，拼接成合理的查询字节序，通过直接定位到行键或者行键的上一个行键，快速获得满足条件的状态监测索引数据，读取后续的数据，获得满足条件的状态监测数据位置信息；当行键不满足时，则查询索引完毕；

[0026] f. 根据所获得所有满足条件的状态监测位置信息集合，从状态数据文件中读出所有的状态监测数据记录，将查询结果返回给客户端。

[0027] 上述状态监测数据的分布式存储与并行挖掘方法，状态监测数据仓库建立与数据挖掘并行化的技术方案如下：

[0028] a. 状态监测数据的抽取：对于现有的长期存储在关系型数据库中的电力设备历史状态数据，经过数据净化、转换、标准化后，以文件的形式存储于 HDFS 的数据结点上；

[0029] b. 状态监测数据仓库的建立：首先采用 Hive 查询语言（HiveQL）创建表，表的定义、字段以及间隔符信息均存储于元数据库中，然后加载 HDFS 数据文件到表以构造数据文件目录；根据变电站、设备类型、监测类型与时间状态监测主题组织成分区，按照列属性将数据组织成数据桶；

[0030] c. 状态监测数据分析：客户端发起状态监测数据分析请求，根据请求命令的内容查询元数据库中对应的表模式，若满足则进入数据文件目录查询相应的表，通过 HiveQL 找到状态监测量字段，获取满足条件的状态量值，进行聚类、求和、汇总、报表生成操作，最后，将操作生成的查询分析计划存储在 HDFS 数据仓库中，并将状态数据分析结果返回给客户端；

[0031] d. 状态监测数据挖掘的并行化：将包括关联规则、分类和聚类的算法运行分发给

作业进程(JobTracker, 部署在主节点)管理下的各个任务进程(TaskTracker,部署在从节点)共同完成;设置并行化引擎实例,通过映射与化简并行编程模型(MapReduce)将学习过程中的大规模数据集运算分割为若干训练子集分配给多个映射节点(Mapper),在 Mapper 节点上分别执行各种操作得到中间结果,最后通过化简节点(Reducer)将结果合并,实现算法的并行执行。

[0032] 本发明采用 WSDL 对 CRU 与 CFC 的状态监测服务进行建模,可以摆脱硬件平台与软件工具的限制,确保了系统的可移植和互操作。采用 SOAP 作为分布式环境中交换数据的简单协议,使得 CRU 与 CFC 完全可以跨越防火墙在电力 Intranet 上进行状态监测数据交换。

[0033] 大规模的廉价服务器集群技术可以直接利用闲置的服务器搭建,且不要求服务器类型相同,大幅降低建设成本。虚拟化技术通过对服务器、存储设备与网络设备等硬件资源进行虚拟化,可以屏蔽各个电力网省公司和直属单位千差万别的硬件资源,以虚拟机为单位进行统一的自动化管理,一方面可以提高资源利用率,另一方面可以简化管理与维护工作。

[0034] HDFS、HBase 以及 HiveQL 等海量分布式数据存储与管理技术可以保障智能电网海量状态监测数据的可靠存储、高效管理与快速查询。MapReduce 并行编程模型以及并行数据挖掘可以为设备状态检修提供高性能并行处理能力。

[0035] 本发明能有效地对智能电网环境下海量电力设备状态监测信息进行分布式数据交换、冗余存储与快速并行处理。

附图说明

[0036] 下面结合附图对本发明作进一步详述。

[0037] 图 1 是智能电网状态监测数据处理系统结构图

[0038] 图 2 是状态监测数据的存储与查询流程图;

[0039] 图 3 是状态监测数据仓库的建立流程图;

[0040] 图 4 是状态监测数据挖掘的并行化流程图。

[0041] 图中及文中各符号为:CFC、状态监测通信前置机;CRU、变电站状态监测远程单元; WSDL 、Web 服务描述语言; HDFS、Hadoop 分布式文件系统;HBase、Hadoop 结构化数据表;HiveQL、Hive 查询语言;JobTracker、作业进程;TaskTracker、任务进程;Mapper、映射节点;Reducer、化简节点;MapReduce、映射与化简并行编程模型;CRUServiceInterface、CRU 服务接口;CFCServiceInterface、CFC 服务接口;HTTP、超文本传送协议;RPC、远程过程调用;SOAP、简单对象访问协议;XML、结构化描述语言。

具体实施方式

[0042] 本发明公开的一种状态监测数据的分布式存储与并行挖掘方法,包括状态监测数据交换、状态监测数据存储与查询、状态监测数据仓库与数据挖掘,所述状态监测数据交换中,建立 Web 服务描述语言定义变电站状态监测远程单元与网省公司状态监测通信前置机的功能服务模型,通过简单对象访问协议进行电力广域网环境下变压器、断路器与容性设备等状态监测数据交换;所述状态监测数据存储与查询中,将大规模状态监测数据冗余存储在分布式文件系统中,通过对状态监测数据文件创建索引表,插入到大规模结构化数据

表中，根据查询请求完成状态监测数据查询。所述状态监测数据仓库与数据挖掘中，通过提取、转换与加载生成基础数据与多维度的分析性数据建立数据仓库，并通过映射与化简并行编程模型将任务分解与结果汇总，实现关联规则、分类和聚类等数据挖掘算法的并行执行。本发明能有效地对智能电网环境下海量电力设备状态监测信息进行分布式数据交换、冗余存储与快速并行处理。

[0043] (1)状态监测分布式数据交换

[0044] 变电站设备状态监测的分布式数据交换由变电站状态监测远程单元(CRU)与网省公司状态监测通信前置机(CFC)构成，采用 Web 服务描述语言 (WSDL) 定义 CAG 与 CAC 的状态监测服务接口、状态监测服务消息与参数、状态监测信息传输方式与消息格式，建立状态监测数据交换的服务模型，通过简单对象访问协议 (SOAP) 实现变压器、断路器与容性设备等状态监测数据的远程传输，具体方法如下：

[0045] 1)状态监测服务接口的定义。为 CRU 与 CFC 分别定义抽象服务接口 CRUServiceInterface 与 CFCServiceInterface。CRU 与 CFC 之间的数据交换分为主动上传与命令下发两类过程。CRU 平时处于堵塞状态，周期性主动唤醒，向 CFC 上传数据，为 CFC 定义上传心跳信息、配置信息与状态监测数据等操作，供 CRU 来调用。另外，CFC 还需要主动唤醒 CRU，来下发控制命令，为 CRU 定义控制命令等操作，供 CFC 来调用。

[0046] 2)状态监测服务消息与参数的描述。为上述操作定义请求消息与响应消息，请求消息类似于函数的输入参数，而响应消息类似于函数的返回值，并约束输入与返回参数的数据类型，其中心跳信息请求消息输入参数包括标识符、工作状态与网络状态等，配置信息请求消息输入参数包括标识符、数据上传周期与配置参数等，状态监测数据请求消息输入参数包括监测数据代码、监测数据值与告警状态等；

[0047] 3)状态监测信息传输方式与消息格式的定义。将 CRUServiceInterface 与 CFCServiceInterface 服务接口绑定为简单对象访问协议(SOAP)，传输方式采用超文本传送协议(HTTP)，并采用 document/literal 作为状态监测信息的编码方式。

[0048] 4)访问端点的部署绑定。指定特定网络地址来定义访问端点，通过该地址访问所提供的状态监测服务，例如 CFC 服务访问端点(CFCServicePort)的网络地址为 http://202.206.212.90/CFC_WS/CFCService.asmx，客户端将通过该地址访问 CFC 所提供的状态监测服务。

[0049] 5)状态监测数据的交换过程如下：

[0050] a)CRU 处于堵塞状态，周期性主动唤醒后，发起调用远程对象 CFC 的状态监测数据服务的远程过程调用(RPC)请求；

[0051] b)状态监测数据服务的 RPC 请求被封装成一个采用结构化描述语言(XML)编码的 SOAP 请求消息，发送到 CFC 的 SOAP 服务器上；

[0052] c)CFC 的 SOAP 服务器解码收到的 SOAP 请求消息，对变压器、断路器与容性设备等状态监测数据进行业务逻辑处理，判断是否存在缓存的尚未发出的配置与控制命令等，再将处理结果封装成 SOAP 响应消息；

[0053] d)CRU 获得状态监测数据服务的响应消息后，判断是否执行配置与控制命令。

[0054] (2)状态监测数据的存储与查询

[0055] 利用虚拟化监视器或虚拟化平台对服务器、存储设备与网络设备等硬件资源进行

虚拟化，以虚拟机为单位构建 Web 服务器集群、应用服务器集群与数据库服务器集群作为运行环境。将收集的海量状态监测数据存储在 Hadoop 分布式文件系统（HDFS）集群中，采用主 / 从架构，主节点负责检测 HDFS 是否有新的文件产生，并分发给子节点让其创建索引，子节点根据文件记录创建索引，并插入到 Hadoop 结构化数据表（HBase）中。查询客户端发送请求，在获得状态监测数据查询列表后，从 HDFS 的状态监测数据文件中读出详细的状态监测数据记录，并逐一返回客户端，如图 2 所示。

[0056] 状态监测数据的存储与查询的具体步骤如下：

[0057] 1）从 CFC 收集的状态监测数据以文件形式组织，直接将数据以二进制的形式存放到文件里，不包含任何的冗余数据，将数据转化为便于查询的结构化形式。读取状态监测数据文件，逐行扫描每个状态监测数据记录；

[0058] 2）将文件扫描检测和索引创建分布在不同节点上，设置主节点服务器对状态文件检测和索引创建，检测是否产生新的状态监测文件，将新的状态监测文件名整合成索引创建请求，并分发给子节点处理。如果子节点失效，转移到其他子节点上。子节点部分检测请求是否到来以及是否为合理，每当子节点接收到一个状态监测数据文件索引创建任务，将从 HDFS 中读取的状态监测文件数据读入内存中，并记录该文件的名称；

[0059] 3）对文件中的每个状态监测数据记录逐行扫描，提取出对查询有效字段，添加到列表中，根据这些常用字段建立索引表，例如在状态监测数据索引中，其索引字段为“变电站 id+ 监测时间 + 数据”；

[0060] 4) 将状态监测数据文件产生的索引表插入到 HBase 中，接受并处理用户的状态数据查询请求，并检测该请求是否合理，查询遍历索引表；

[0061] 5) 索引表中行键为查询字段，偏移量为状态监测数据记录在状态数据文件中的位置，即文件名加偏移量，查询将通过文件名和偏移量来获取数据，一张表的行键按照字节序顺序排序，对于指定查询条件，拼接成合理的查询字节序，通过直接定位到行键或者行键的上一个行键，快速获得满足条件的状态监测索引数据，读取后续的数据，获得满足条件的状态监测数据位置信息；当行键不满足时，则查询索引完毕；

[0062] 6) 根据所获得所有满足条件的状态监测位置信息集合，从状态数据文件中读出所有的状态监测数据记录，将查询结果返回给客户端。

[0063] （3）状态监测数据仓库与数据分析

[0064] 通过提取、转换与加载生成规范的、无冗余的基础数据，并生成多维度的分析性数据存储在分布式数据仓库中。通过映射与化简并行编程模型（MapReduce）将任务分解与结果汇总，实现电力设备状态检修中关联规则、分类和聚类等数据挖掘算法的并行化。状态监测数据仓库建立与数据挖掘并行化的技术方案如下：

[0065] 1)状态监测数据的抽取。对于现有的长期存储在关系型数据库中的电力设备历史状态数据，经过数据净化、转换、标准化后，以文件的形式存储于 HDFS 的数据结点上。

[0066] 2）状态监测数据仓库的建立。首先采用 Hive 查询语言（HiveQL）创建 Hive 表，Hive 表的定义、字段以及间隔符信息均存储于元数据库中，然后加载 HDFS 数据文件到 Hive 表以构造数据文件目录。根据变电站、设备类型、监测类型与时间等状态监测主题组织成分区，按照列属性将数据组织成数据桶。

[0067] 3）状态监测数据分析。参看图 3，客户端发起状态监测数据分析请求，根据请求

命令的内容查询元数据库中对应的表模式，若满足则进入 Hive 数据文件目录查询相应的 Hive 表，通过 HiveQL 找到状态监测量字段，获取满足条件的状态量值，进行聚类、求和、汇总、报表生成等操作。最后，将操作生成的查询分析计划存储在 HDFS 数据仓库中，并将状态数据分析结果返回给客户端。

[0068] 4）状态监测数据挖掘的并行化。将包括关联规则、分类和聚类的算法运行分发给作业进程（JobTracker，部署在主节点）管理下的各个任务进程（TaskTracker，部署在从节点）共同完成；设置并行化引擎实例，通过映射与化简并行编程模型（MapReduce）将学习过程中的大规模数据集运算分割为若干训练子集分配给多个映射节点（Mapper），在 Mapper 节点上分别执行各种操作得到中间结果，最后通过化简节点（Reducer）将结果合并，实现算法的并行执行，如图 4 所示。

[0069] 专业术语解释

[0070] （1）云计算

[0071] 一种网格计算、分布式计算、并行计算、效用计算、网络存储、虚拟化、负载均衡等传统计算机技术和网络技术发展融合的产物。云计算通过网络把多个成本相对较低的计算实体整合成一个具有强大计算能力的系统，并借助商业模式把计算能力分布到用户手中。

[0072] （2）智能电网

[0073] 智能电网，就是电网的智能化，也被称为“电网 2.0”，它是建立在集成的、高速双向通信网络的基础上，通过先进的传感和测量技术、先进的设备技术、先进的控制方法以及先进的决策支持系统技术的应用，实现电网的可靠、安全、经济、高效、环境友好和使用安全的目标，其主要特征包括自愈、激励和包括用户、抵御攻击、提供满足 21 世纪用户需求的电能质量、容许各种不同发电形式的接入、启动电力市场以及资产的优化高效运行。不同国家针对本国的能源现状制定了不同的智能电网目标。美国侧重于建设现代化电力系统，并注重需求侧管理和可再生能源的应用；欧洲侧重推广分布式发电，比如微电网组网及运行、分布式发电控制、需求侧管理等；日本将主要围绕大规模开发太阳能等新能源，确保电力系统稳定，构建智能电网；中国提出建设“坚强智能电网”，包含电力系统的发电、输电、变电、配电、用电和调度共 6 个环节，具有信息化、自动化、互动化的智能技术特征。

[0074] （2）状态监测

[0075] 状态监测包括在线监测，必要时的离线检测及试验，以及不与运行设备直接接触的所有可得到运行状态数据的手段，在线监测是指直接安装在设备本体上可实时记录表征设备运行状态特征量的测量系统及技术。

[0076] （3）状态监测远程单元（CRU）

[0077] 部署在变电站内的，能以标准方式对站内各类综合监测单元或状态监测装置进行状态监测信息获取及控制的一种装置。

[0078] （4）状态监测通信前置机（CFC）

[0079] 部署在主站系统侧的一种关口设备，能以标准方式远程连接变电站内状态监测设备，获取并校验各类状态监测信息，并可进行控制的一种计算机。

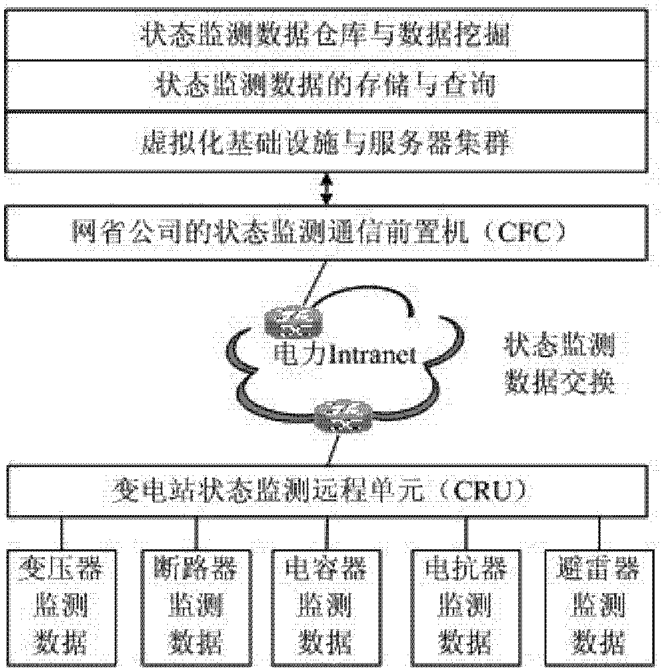

状态监测数据仓库与数据挖掘
状态监测数据的存储与查询
虚拟化基础设施与服务器集群
网省公司的状态监测通信前置机（CFC）
电力Intranet
状态监测
数据交换
变电站状态监测远程单元（CRU）
变压器
监测
数据
断路器
监测
数据
电容器
监测
数据
电抗器
监测
数据
避雷器
监测
数据

图 1

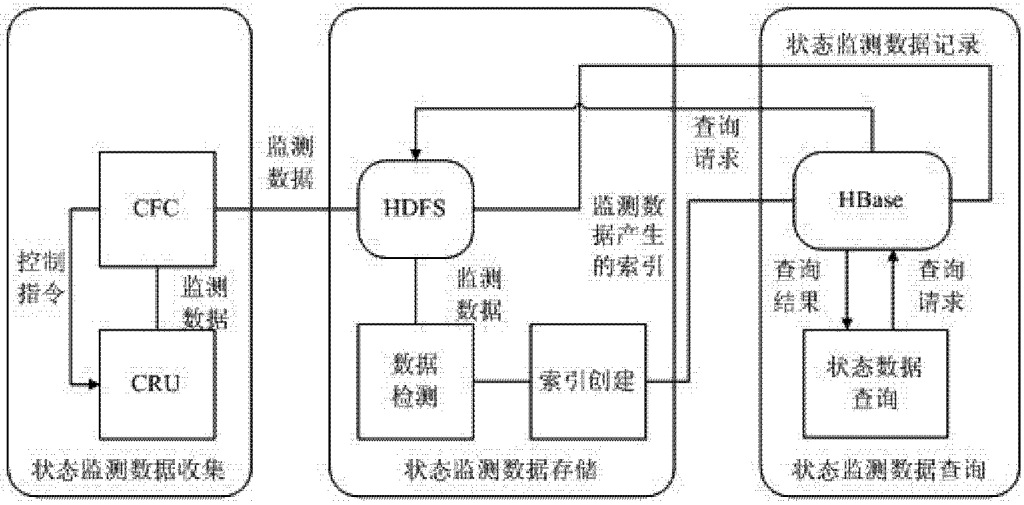

状态监测数据记录
监测
数据
查询
请求
CFC
HDFS
HBase
监测数
据产生
的索引
控制
指令
监测
数据
监测
数据
查询
结果
查询
请求
CRU
数据
检测
索引创建
状态数据
查询
状态监测数据收集
状态监测数据存储
状态监测数据查询

图 2

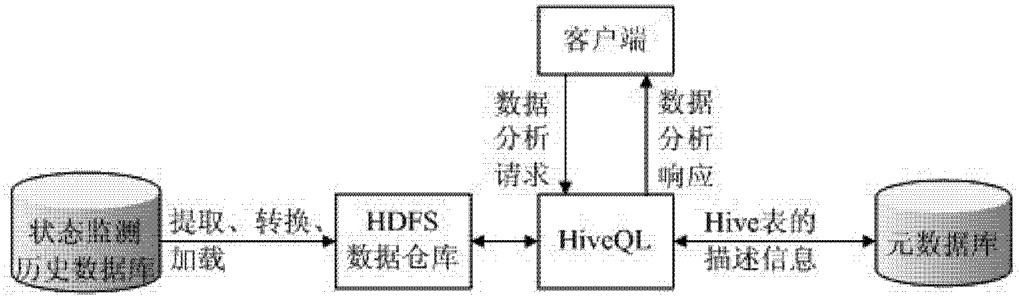

客户端
数据分析请求
数据分析响应
状态监测历史数据库
提取、转换、加载
HDFS数据仓库
HiveQL
Hive表的描述信息
元数据库

图 3

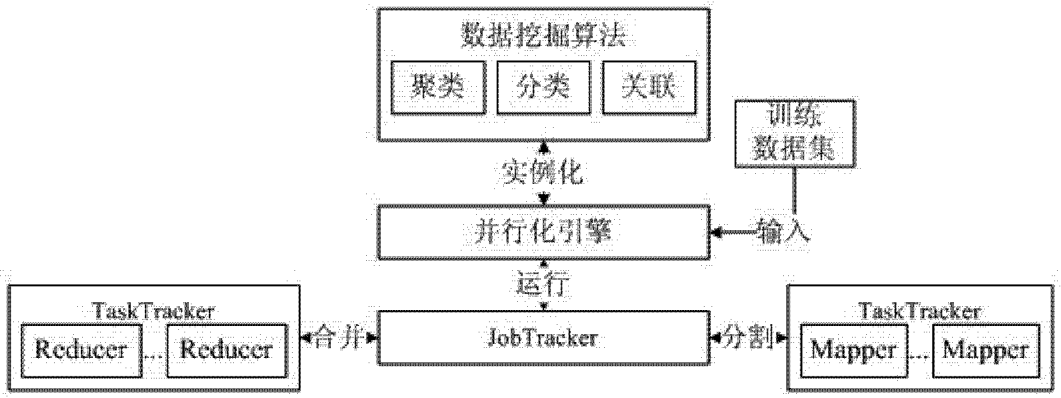

数据挖掘算法
聚类
分类
关联
训练数据集
实例化
并行化引擎
输入
运行
TaskTracker
Reducer
Reducer
合并
JobTracker
分割
TaskTracker
Mapper
Mapper

图 4