企业大数据管理解决方案

梁 钢
(北京华胜天成科技股份有限公司，北京 100192)

摘 要：大数据的应用方兴未艾，根据国内企业的应用场景，给出了企业大数据管理解决方案。此方案还为数据的进一步处理打下了基础。

关键词：大数据；企业大数据管理
中图分类号：TP302.1 **文献标识码**：B **文章编号**：1674-7720(2013)24-0007-04

Big data dolution for enterprise

Liang Gang
(Beijing Teamsun Technology Co., Ltd., Beijing 100192, China)

Abstract: We provide a solution for big data management, which is useful to Enterprise. The design of system satisfies user's requirement.

Key words: big data; enterprise big data management

IT 行业一直在不断地努力，以最佳方案满足日益增长的各种需求。继云计算之后，大数据又成为业界关注的热点。云计算更多地体现在它的商业模式与服务模式上，而大数据则更关注数据的处理，而这些纷杂的数据则是关系社会、企业乃至个人生活的核心关键，可以说数字时代数据为要。

1 大数据参考架构

通常人们认为大数据具有 4V 特点，即：Variety(多样性)、Volume(大容量或海量)、Velocity(快速)和 Value(价值)。至于大数据的严格定义，则是人者见人、智者见智，莫衷一是[1]。

根据调研与实践，本文给出了相关的参考架构，如图 1 所示。

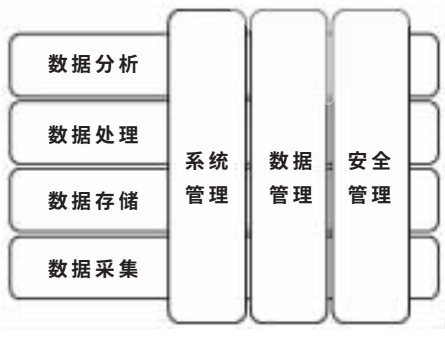

图 1 大数据参考架构

可以将大数据的参考层次分为 4 个：

(1) 数据采集。主要涉及对数据源的采集，包括各种结构化与非结构化数据、静态数据与动态实时数据等。

(2) 数据存储。主要涉及对数据的存储，包括分布式存储、海量存储、虚拟存储等。

(3) 数据处理。主要涉及对数据的转换、传输、分发等。

(4) 数据分析。主要涉及对数据的清洗、比对、挖掘、钻取等。

同时，按照数据平台管理、数据维护、安全保护等维度，存在着贯穿各层的管理机制，即：

(1) 系统管理。对构建的系统平台进行管理与维护。

(2) 数据管理。按照数据生命周期对数据进行管理。

(3) 安全管理。对数据隐私、数据安全、访问安全、系统安全等方面进行管理。

2 企业大数据解决方案

由于大数据的应用很多，本文更加关注企业所处的混杂数据的应用场景，基于上面给出的参考架构，给出相应的解决方案。

2.1 应用场景

企业的数据是企业的核心资料，企业信息化的核心问题就是数据的应用的效率与效果。目前企业的数据主要包括：财务类数据、管理类数据、业务类数据等，这些数据可以是结构化数据和非结构化数据。从容量上看，

随着信息化应用的不断提高，可以达到 GB 或 TB 级，对于一些行业，甚至有可能达到 PB 级。

2.2 解决方案

本文提出的企业大数据解决方案是从业务连续性的角度来考虑用户数据的问题。参考了业界流行的 ISO20000、ISO27000、BCP/DRP、SOA 等相关标准和技术，从安全、服务的范畴来管理数据、保护数据、使用数据。

方案主要解决企业用户的结构化与非结构化数据的存储、管理，为企业相关应用提供基础数据，为企业的业务连续性保驾护航。

2.2.1 技术特点

方案主要融合了信息安全技术、数据管理技术、数据同步复制技术、数据库技术、商务智能技术等，区别于现有的数据备份产品、数据复制产品、数据管理产品，更关注数据在复制之后能够被快速使用与恢复，以延续业务的连续性。

方案为用户数据的进一步加工处理打下了基础，有助于用户整合数据、整合应用、数据加工、商务智能、决策分析等。

主要特点：

(1)支持多种数据库的不同版本，也支持多种异构数据库之间的同步，如 Oracle、SQL Server、MySQL、Sybase、DB2、AS400 等可以同步到 Oracle 数据库或其他数据库上。

(2)支持一对一、一对多、多对一、多对多等异构数据库同步方式。

(3)比较强的数据加工能力，可以选择数据源的不同字段，也可以对数据源做相应的转换、逻辑判断、映射等处理，还可以设置在数据同步时做异常数据检查等。

(4)比较强的传输能力，内置数据传输平台，满足复杂网络情况下的数据可靠传输，支持广域网下的数据同步，支持跨网段的数据同步，支持物理隔离情况下的数据同步。

(5)易用性。提供中文工具，方便可视化操作和监控。

2.2.2 技术原理

统一支持结构化数据和非结构化数据的同步及相应加工。提供可视化工具配置结构化数据和非结构化数据的同步与加工。

(1)非结构化数据文件既可以通过系统内置的传输平台同步到备份方的文件夹下，也可以将备份方文件夹下的数据文件映射到数据库上。

对于非结构化的文件备份，可以在数据源方部署一个节点，负责监控和发送文件，通过可视化配置的数据推送服务，选择要发送的文件夹、文件、接收节点、接收文件夹等信息，通过定时等调度策略将文件发送到备份方。当然要发送的文件(或文件夹)、备份方的文件(或文件夹)可以来自于接口表或接口文件，通过接口表(或接口文件)实现文件的备份。

能实现非结构文件到结构化数据的映射，可视化配置非结构文件到异构系统的映射服务，可视化定义文件分类处理服务，根据文件的不同分类调用相应非结构文件到异构系统的映射服务。

(2)结构化数据方面支持 Oracle、MS SQL Server、IBM DB2、AS 400、Sybase ASE、Sybase IQ、MS Access、MySQL、PostgreSQL、Intersystems Cache、Informix、Gupta SQL Base、dBase III, IV or 5、Firebird SQL、MaxDB (SAP DB)、Hypersonic、Generic database、SAP R/3 System、CA Ingres、Borland Interbase、KingbaseES 等不同版本的数据库作为源或者目标。

其技术原理如图 2 所示，核心主要包括数据源层、数据管理服务器组层、数据镜像服务器组层、数据存储层这几部分。数据源可以是不同业务系统的数据库，也可以是文件系统；ReiKing 引擎实现了将异构的数据源(数据库或非结构化的文件等)备份到相应的镜像服务器的数据库或文件系统上，ReiKing 引擎部署在服务器上，一台机器可以部署一个或多个 ReiKing 引擎，根据同步业务负载情况通过扩展引擎数或者机器数实现性能和可靠性的扩展；数据镜像服务器组的数据库服务器接收来自于 ReiKing 引擎的数据，并通过数据库服务器保存到结构化数据存储；数据存储层可以通过数据库服务器保存结构化数据，也可以通过 ReiKing 引擎直接保存要同步的文件等信息。

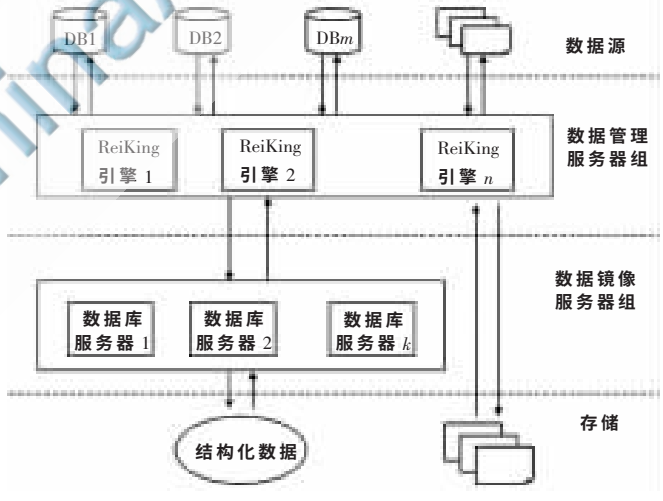

图 2 技术原理

可以生成数据备份引擎，通过业务逻辑策略的定制，一个引擎可以完成一个或者多个数据源的备份，也可以多个引擎完成一个数据源的备份。一台机器可以部署一个引擎，也可以一台机器部署多个引擎，随着业务应用及信息系统不断扩展，方便通过增加引擎等手段的拓展。

引擎之间可以互为备用，示意图如图 3 所示。

有一个或多个引擎组成运行环境，引擎可以分组处理，也可以互为备份。

图 3 示意图

机器 C 运行管理服务器、管理工具，保存统一的规则和定制相互备份的运行服务器的使用规则；机器 A、机器 B 保存各自的使用规则，平时独立运行，各自做相应的处理服务，当任何一台机器出现故障时，另一台机器根据设定规则，启动出现故障的机器上运行引擎，并做相应的调动运行处理。

经过数据同步、交换、集中等整合处理后的数据还可以做数据治理等深加工，包括数据质量的管理、主数据的管理、数据的监控、数据审核等，方便数据分析、数据决策等数据应用；同时，还可以实现数据的共享和交换，配置出共享的数据服务，通过安全的授权和权限鉴定，方便实现数据的安全共享，减少对数据库的直接访问，保证使用者只能访问经过授权的数据。此外还可以实现数据生命周期管理，经过备份的数据可以再被归档到不同的库中，需要时可以按需恢复使用备份和归档的数据。

2.2.3 基于流计算的数据加工服务

数据管理提供了基于流计算的数据加工服务，对来自数据库、文件、队列、应用系统等异构系统的数据，在数据流中做加工处理，包括数据转换/清洗、数据复制、差错数据发现、数据传输、数据比对、数据装载、流程处理、数据路由等处理，如图 4 所示。主要功能如下：

(1)异构数据源或目标

① 数据库：Oracle、SQL Server、DB2、Sybase ASE、Sybase IQ、Informix、My SQL、Access 等。

② 数据文件：文本文件(分隔符、定长等)、Excel 文件、DBF 文件、睿智文件、XML 等。

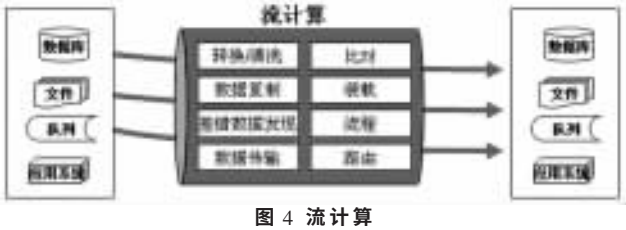

图 4 流计算

③XML：XML 文件或内存中的 XML。

④内存表：由其他系统或消息队列传来的数据可以存在内存表中作为输入，经过整合处理后也可以放到内存表中，提供给被调用方，还可以放到消息队列中，由消息中间件传输处理。

⑤数据库表、消息队列内容、文件、XML 之间的相互转换。

⑥支持异构的字符集，数据源或目标可以是中文、西文等字符集。

⑦支持数据库 SQL。支持 SQL 语句调用、支持 SQL 函数调用、支持 SQL 存储过程调用。

⑧支持结构体，方便自定义类型、自定义结构、结构体成员的抽取。

(2)支持实时、增量、批量、全量的抽取。抽取条件可以是静态语句、动态 SQL 规则、来自变量、来自变量组等。

(3)数据转换处理

①格式转换，包括字段拆分/合并、不同格式间转换。

②静/动态字段，包括系统时间、动态序列号、给定值。

③比对、翻译转换处理，包括基于规则表的翻译、给定规则翻译、给定数据的比对处理等。

④数学运算，不同的数据对象之间作数学运算。

⑤聚类处理，根据一个或几个字段做聚类操作。

⑥身份证转换等转换处理。

(4)数据路由

①采用“一次抽取，按条件路由”的机制。

②支持一对多的数据推送方式。

③减轻对数据源如数据库的压力。

④提高处理的性能。

⑤路由条件可以是动态的，也可以是组合的。

(5)支持数据比对装载处理。支持和目标内容做比对操作，并根据比对结果做相应的增加、覆盖、删除等处理。

(6)支持缓慢变化维、日志、比对、数据回写等增量抽取策略。

(7)缓慢变化维处理

①提供缓慢变化维模版和向导，方便缓慢变化维的设计。

②可以保留最新值、保留上次数据值，也可以保留给定时间范围或最近的数据，还可以保留所有的历史数据值。

2.2.4 安全

数据安全处理主要包括系统认证安全、传输安全、安全授权和鉴定[2]。

(1)系统安全认证

系统安全实现提供运行时鉴定，ReiKing 引擎运行时

验证运行机器和 Key，只有都匹配时才能执行，保证 ReiKing 运行的加工规则只能在 ReiKing 环境下运行。ReiKing 提供安全连接认证机制，每个节点都有不同的密钥用于实现建立连接时的加密处理和安全的认证。

(2)传输安全

提供可靠的安全传输机制，保证了数据传输中的数据的一致性、完整性。除了网络传输的重送和数据冗余校验机制外，还提供了数据稽核机制，对传输的数据量、文件数量、实体完整性和非空字段进行稽核。

对于涉密数据，还提供了安全加密传输机制，可以根据密钥对所需数据进行加密后传送。

(3)安全授权和鉴定

提供安全授权管理，满足不同用户安全权限的需求。比如有的用户只有浏览的权限而没有编辑的权限，有的用户只能编辑自己的对象而不能访问别人的对象，有的用户只有设计的权限而没有运行任务的权限，而管理员拥有全部的权限。

提供分级安全管理功能，实现了如下安全管理：

①系统提供管理员(包含超级管理员、部门管理员、组管理员)、开发者、使用者等多种权限级别的用户管理，可以由上级管理员授权下级管理员权限，满足总公司和下属企业两级权限管控的管理需要，如系统管理员只能设置本单位及下属单位的用户。

②分项授权，对运行服务器、数据库连接、服务、流程、整合服务等分项授权，权限包括执行权、编辑权、读取权，满足系统级、数据库级、软件功能级、记录级和字段级等多级别的安全控制需要。

③通过用户管理和权限管理，系统对数据实现分级管理，本单位的用户或系统管理员只能对本单位或下属单位的数据进行维护，不可调整上级单位的数据。

④系统提供较完善的日志管理，能详细记录各用户(含系统管理员)在系统中的操作情况。

⑤身份和权限的鉴定，操作者在做开发管理，或者数据服务使用者在使用服务时，都会根据该用户的授权做相应的身份鉴定和权限鉴定。

数据服务使用安全，当应用程序通过 Web Service 方式、API 方式、事件等方式使用数据服务时，其访问情况将由安全授权来决定。

根据国内企业的大数据应用特点，本文提出了相应的大数据管理解决方案。实践证明，该方案能够较好地解决国内企业各种数据源的数据的存储、处理等问题，并为解决业务连续性问题打下了基础。可以说这是一种性价比很高、易于操作的方案。

参考文献

[1] RAJARAMAN A，ULLMAN J D. 大数据：互联网大规模数据挖掘与分布式处理[M]. 王斌，译. 北京：人民邮电出版社，2012.

[2] 梁钢，茅秋吟. 云计算 IaaS 平台的信息安全和运维服务设计[J]. 电子技术应用，2013，39(7)：63-64，96.

(收稿日期：2013-09-23)

作者简介：

梁钢，男，1969 年生，硕士，高级工程师，主要研究方向：云计算、大数据、容灾、安全方面解决方案的研究与推广。