



基于电信大数据的数据建模平台研究

沈雷明¹, 别志铭²

(1. 中国移动通信集团上海有限公司 上海 200060;

2. 珠海世纪鼎利通信科技股份有限公司 珠海 519085)

摘要: 基于电信大数据的 Hadoopdesigner 系统, 是针对电信大数据的特点, 根据元数据的设计思想开发的数据建模平台, 融合了数据仓库的很多建模思想和理论, 为电信大数据处理提供了一种新的方法, 也为不同种类的数据提供了统一、标准、可视化的模型, 大大降低了统一、快速满足不同业务的不同需求的难度, 提高了数据处理效率。

关键词: 数据建模; 大数据; 云计算; Hadoop; 电信

doi: 10.3969/j.issn.1000-0801.2014.06.022

Research on Data Modeling Platform Based on Big Data of Telecom

Shen Leiming¹, Bie Zhiming²

(1. China Mobile Group Shanghai Co., Ltd., Shanghai 200060, China;

2. DingLi Communications Co., Ltd., Zhuhai 519085, China)

Abstract: The Hadoopdesigner system based on big data of telecom was designed. According to the characteristics of telecom data, the system adopted the data modeling platform which was designed based on the idea of metadata. It also combined with many ideas and theories of the data warehouse modeling. It provided a new method for the treatment of big data of telecom, also it provided the unified, standard, and visual model for different kinds of data, and greatly reduced the difficulty of meeting fleetly different requirments of services by a unified way, improved processing efficiency.

Key words: data modeling, big data, cloud computing, Hadoop, telecom

1 电信行业大数据的需求

伴随电信行业 3G/4G 等业务的增加, 数据量出现爆炸式增长, 数据的规模已经超过 TB, 达到 PB 级别。海量数据导致电信运营商的数据处理和存储压力急剧变大, 而传统的数据库和数据仓库越来越无法满足电信行业的需求。许多电信运营商开始关注 Hadoop 技术, 但是电信运营商自身组织结构复杂, 产品和业务种类繁多, 并且 Hadoop 技术也在快速发展。如何能统一、快速地满足不同业务的不同需求? 这个难题摆在所有人的面前。

2 电信领域大数据的特点

(1) 电信用户数据量大

电信行业是个垄断行业, 每个国家通常只有 3 家或 4 家电信运营商, 这样就导致了运营商拥有超过一亿的用户。

(2) 用户产生的数据量大

电信用户每天都要接打电话, 通过 3G/4G 使用互联网, 即使用户只是把手机连接到电信的网络中, 都会产生类似位置更新等数据。每个用户每时每刻都在产生数据, 一亿用户汇聚起来, 数据量非常庞大。

(3) 用户产生的数据多样性

电信用户打电话、发短信、使用互联网等,还有客户的位置、在线状态等,每种数据结构都不一样,内容也不一样,其中有结构化的数据,也有非结构化的数据,非常复杂。

3 Hadoopdesigner 系统设计

本文介绍了 Hadoopdesigner 系统,它是基于 Hadoop 的数据建模平台。该系统根据电信业务系统的特点,提取出类似 counter、KPI(key performance indicator,关键绩效指标)等这样的专业电信行业模型,是为电信行业的大数据建模量身打造的。Hadoopdesigner 系统总体结构如图 1 所示。

其总体结构是由模型处理、模型对象管理、智能服务等模块构成。

3.1 模型处理

模型处理包括的功能有模型建立、模型智能评估和模型运行监控。每个功能模块的具体介绍如下。

(1) 模型建立

建模人员拿到应用的需求后,把需求中的具体业务分解成系统里面的实体对象。整个建模中,一般需要经历 4 个过程。

- 业务模型建模。它主要解决业务层面的分解和程序化。
- 专业领域建模。它主要是对业务模型进行抽象处理,生成专业的领域概念模型。
- 逻辑建模。它主要是将专业领域模型的概念实体以及实体之间的关系进行逻辑化。
- 物理建模。它主要解决如逻辑模型对不同实体的物

理化及性能等一些具体的技术问题。

其中逻辑模型和物理模型由 Hadoopdesigner 系统根据业务模型和专业领域的模型自动生成。建模人员只需要对业务模型和专业领域的模型进行设计。

(2) 模型智能评估

常用的数据建模有范式建模法、维度建模法、实体建模法等。对于电信的信令分析应用来说,一般采用维度建模法。针对各个维度作大量的预处理,如按照维度进行预先的统计、分类等。通过这些预处理,能够极大地提升系统的处理能力。Hadoopdesigner 系统就是通过可视化的界面,辅助建模人员使用维度建模法实现应用,并且对设计的模型进行专业的智能评估和分析,给出分析报告。

(3) 模型运行监控

业务和数据建模是个复杂的过程,对于复杂的应用,不可能一下子就能设计出非常合适的模型,需要对模型运行过程进行监控,及时地发现问题和解决问题,以完成对模型的持续优化。

3.2 模型对象管理

建模人员在业务模型建模和专业领域建模后,会从复杂的应用中提取出实体、事件、说明、关系等抽象的对象,再通过 Hadoopdesigner 系统的可视化界面,建模人员把抽象对象配置到系统中,就完成了整个应用建模过程的第一个阶段。剩下的任务就是根据模型运行的情况,调整、优化模型。目前 Hadoopdesigner 系统支持的常用模型有以下几个类型。

(1) 数据源模型

目前数据的来源主要是通过 FTP 传输过来的文件。文件格式一般是 csv。从应用角度出发,数据可分为 CS、PS

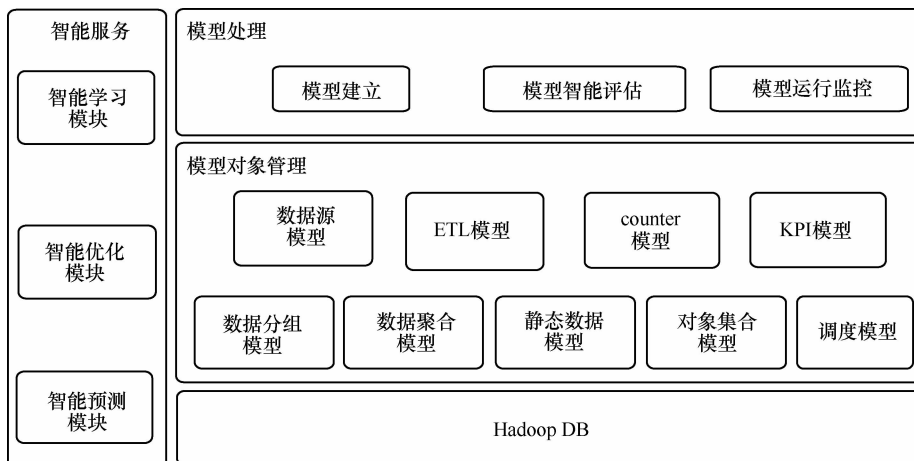


图 1 Hadoopdesigner 系统总体结构



等,它们的数据字段信息等是不同的。因此需要根据不同的应用描述不同数据来源的信息,统一数据读取接口。Hadoopdesigner 系统提供一些常用的数据源定义模板,在导入后,直接修改便可以使用。

(2)ETL 模型

定义好数据的来源后,数据就会源源不断地通过入云程序把数据导入 Hadoop 中。在数据真正入云前,需要定义数据 ETL(extract, transform and load)的规则和模型,对数据进行清洗、转换和集成以及提供数据入云的速率、数据错误率等数据质量报告等。对于常用 ETL 规则, Hadoopdesigner 系统也提供一些模板,方便用户直接使用。

(3)counter 模型

counter 模型是系统定义单一维度的统计汇总模型。它是系统里面非常重要的模型,是维度建模法的基础。系统可以通过定义的 counter 模型,对需要统计分类的维度进行预处理和优化。counter 模型根据数据所处的位置分类,可以分为两类:一是云外 counter 模型,主要定义在数据入云的同时,对单个维度的汇总统计;二是云内 counter 模型,主要是定义在 Hadoop 里面,对单个维度的汇总统计。通过两种类型的 counter 模型,可以充分利用系统的基础和存储能力,大大提高以后应用汇总统计的速度和能力。比如定义通话中的主叫成功次数、主叫失败次数等,都可以定义为 counter 模型。不同的行业可以定义不同的 counter 库。针对电信的信令分析行业,笔者建立了 counter 模型库,提供了常用的几百个 counter 对象。

(4)KPI 模型

counter 模型相当于单个实体,KPI 模型就是通过各种运算符等关系,连接多个不同实体,组成一个新的 KPI 模型。比如计算通话总次数这个 KPI,就可以通过定义的通话成功 counter 模型加上定义的通话失败 counter 模型,构成通话总次数的 KPI 模型。在实际的应用中,KPI 模型会经常被引用,并且不同的 KPI 可以直接相互地引用。例如需要定义通话成功率的 KPI 模型,就可以通过定义的通话成功 counter 模型加上定义的通话总次数的 KPI 对象,构成通话成功率的 KPI 模型。针对电信的信令分析行业的常用 KPI 模型,Hadoopdesigner 系统也定义了上百个 KPI 模型,方便用户组建模的时候调用。

(5)数据聚合模型

在对数据进行汇总统计的时候,需要指定对某个维度进行聚合。通过数据聚合的模型,可以很简单地从上面定

义的数据源模型的维度里,任意挑选一个维度进行处理,然后再从下拉框中选择一个对这个维度聚合的算法。

聚合的算法常有两类:一是对时间的聚合算法,对某个时间维度进行不同时间段的聚合,如 15 min、30 min、1 h 等聚合;二是通过特定的算法,把维度从细粒度变成粗粒度,如从信令中基站控制器的维度汇总到移动交换中心的维度。数据聚合模型需要根据具体的行业需求,开发出不同聚合算法的组件。

(6)数据分组模型

上述模型定义完成后,通常需要对数据进行分组和分类统计汇总等。数据分组模型就是用来实现数据分组和分类的模型。系统会根据数据源的模型,取出相关的分组维度,提供单选和多选、排序等方式方便用户选择需要的维度。

(7)静态数据模型

在实际应用中,静态数据的使用是不可避免的。因此对静态数据的管理和建模也是很重要的。在信令分析应用中,如手机终端的静态信息会经常被使用。但是静态数据不一定是固定不变的数据,只是该数据变化和更新较少,实际上还是需要静态数据进行增删改等操作。静态数据模型按关联的位置可以分为两种:一是在数据的分析阶段关联的静态数据,对于 Hadoop 来说,就是在 MapReduce 的 map 阶段关联的静态数据,如分析某个特定型号的手机的通话次数,需要通过 TAC 号来关联终端型号的静态数据;二是在汇总阶段的关联静态表,对于 Hadoop 来说,即是 MapReduce 的 reduce 阶段关联的静态数据,如分析所有不同型号的手机的通话次数,需要按 TAC 汇总后,再通过 TAC 号关联终端型号的静态数据,当然这个例子也可以在 map 阶段汇总,但是考虑到性能内存等问题,最好是在汇总后进行关联。

(8)对象集合模型

Hadoopdesigner 系统支持按星型模式或雪花模式来设计。上述模型都是周围的“小星星”或“小雪花”,那么结构的中心实体就是对象集合模型,是用户最关心的基本实体和查询活动的中心,为 Hadoop 的活动提供定量数据。它用来集合和引用上面各种模型和对象,组成一个实际的应用。为了方便用户的建模,Hadoopdesigner 系统扩展对象集合模型面向对象的设计,主要表现在以下两个方面。

- 支持多个对象的继承关系。如 A 对象集合模型使用了通话成功次数的 counter 模型,B 对象集合模

型使用了通话失败次数的 counter 模型,当C对象集合模型同时需要使用通话成功和失败次数的 counter 模型时,就可以定义C从A和B继承过来。

- 重载的功能。如A对象集合模型使用了通话成功次数的 counter 模型,定义数据源模型是来自BSSAP对象,B对象集合模型也是使用了通话成功次数的 counter 模型,但是数据源来自RANAP对象。只需要定义B对象集合模型从A对象集合模型继承,同时把数据源重新定义覆盖即可。

(9)调度模型

在海量数据的分析中,做一次汇总统计是一个漫长的过程,快则半小时,慢则需要几个小时。因此在很多的实际应用中,都需要定时来调度运行应用。Hadoopdesigner 系统调度的最大特点在于能实现智能调度。它根据批量处理的多个模型的优先级别运行模型。若系统不忙时,优先等级比较低的模型可以运行;若系统忙时,根据系统的负荷选择运行等级高的模型。

3.3 智能服务

设计一个模型不难,维护和优化好这个模型就显得更为困难。因此 Hadoopdesigner 系统专门增加了智能学习模块、智能优化模块、智能预测模块,方便建模人员维护和优化好模型,分别简单介绍如下。

(1)智能学习模块

在开启智能监控属性时,系统会跟踪模型的执行计划的每个模块的属性,如运行时磁盘的 I/O、CPU 和内存的占用情况、运行的时间等,用于优化模块,找出最优的模型以及给出影响了整个模型运行速度的对象。

(2)智能优化模块

根据学习到的模型的每个具体参数和系统里面配置的优化参数阈值,决定是否启动优化。例如需要经常使用的、重要的数据,如果模型第1次运行需要多次从磁盘读取同一数据,那么模型第2次运行时会把数据读取到内存中,以后多次从内存中访问。系统自动检测当前的硬件和软件状态,给出状态报告,发送给管理员。

(3)智能预测模块

主要是根据性能、调度、对象集合模型等做预测。例如一个应用统计一天内打电话、发短信超过100次的用户信息,系统会根据这个应用自动预测出骚扰电话、广告识别、重大的突发事件等应用(地震、火灾等)。当把应用预测出后,用户下次做类似的应用时,不用重新开发和设计,只在

预测结果上修改即可。

4 Hadoopdesigner 系统的主要特点

Hadoopdesigner 系统的主要特点如下。

- 该系统是基于电信大数据开发的数据建模工具,专门为电信行业的海量数据处理提供服务的一套工具。
- 对于不同的电信业务需求,采用统一的、标准化模型,整个处理过程是可视化的,并且整个处理过程也是闭环的。
- 根据电信行业的业务特点,不再需要专业的设计和开发人员响应不同的需要,只需要一个了解电信业务知识的建模人员就可以完成,大大降低了难度。
- 该系统采用可视化配置界面,方便用户把复杂的问题简单化、模型化和可视化,提高建模人员处理问题的效率。
- 该系统增加了智能服务的功能,方便建模人员根据具体的软硬件情况以及应用需求,维护和优化自己的模型,提升模型运行的速度。

5 Hadoopdesigner 系统在电信行业的应用

在实际应用过程中,发现了 Hadoopdesigner 系统的一些不足,笔者总结了一些经验。大数据的建模既涉及业务知识,也涉及具体的技术,既需丰富的电信行业经验,同时也需要一定的信息技术。要实现数据模型,最重要的是需要一个非常适用的方法论,指导对业务进行抽象、处理,生成各个阶段的模型。

该系统已经在三四个省级运营商公司部署。某省运营商公司的信令共享平台上的本系统已正常运行一年多的时间。其中 Hadoop 的集群规模为400台PC服务器,存储容量为4PB,数据增量为12TB/日,记录数约为216亿条/日。目前建模平台应用的场景主要有以下两种。

- 固定的需求。建模人员通过平台配置好模型,周期性或定时运行任务,用户通过界面查询结果或把结果导出报表查看。
- 突发型的临时需求。接到需求后,快速定制好模型,直接运行任务后导出报表。

用户通过使用建模平台,实实在在地解决了问题,极大地改善了用户体验,获得用户好评。