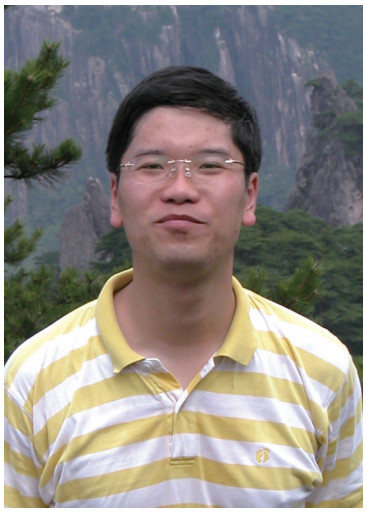
协同产品数据的活跃度管理

Activity Management of Cooperative Product Data

中航工业成都飞机工业(集团)有限责任公司　　赵　力



赵　力

高级工程师,长期从事企业信息化工作,先后参与CIMS工程、工程数据应用、企业级PLM平台建设等重大项目,现在主持整个PLM项目/三维CAPP项目等工程信息化项目,具有很强的PLM技术能力和丰富的开发经验。

在现有协同工作平台的支持下,中航工业成飞的主要核心业务(技术准备、生产准备、采购、财务等)均可以围绕协同工作平台提供的3个BOM(EBOM/PBOM/MBOM)来展开工作,其工作效率、质量、成本控制等方面均得到了巨大改善。

DOI:10.16080/j.issn1671-833x.2015.18.072

项目背景

作为我国航空工业的骨干企业,中航工业成飞多年来成功参与了多个飞机的研制任务,其中,中航工业成飞的数字化工程建设工作功不可没。中航工业成飞协同工作平台经过多年的持续建设,以及在多个飞机型号上的工程实践,得以逐步完善并实现了对中航工业成飞现行业务模式的基本覆盖,为型号飞机的成功研制提供了巨大保障。在现有协同工作平台的支持下,中航工业成飞的主要核心业务(技术准备、生产准备、采购、财务等)均可以围绕协同工作平台提供的3个BOM(EBOM/PBOM/MBOM)来展开工作,其工作效率、质量、成本控制等方面均得到了巨大改善。

随着各个型号的深入研制,参研型号的不断增加,协同工作平台中的产品数据及业务数据呈几何式爆炸性增长,对协同工作平台的运行性能和使用效能提出了更高的要求。到目前为止,纳入协同工作平台的飞机型号已经达到10个,协同工作平台内的产品数据已超过4TB,且以10G/每周的速度增长,这将对当前的协同工作平台造成极大的挑战。在近10年的协同工作平台持续建设和工程化应用过程中,曾出现过因产品数据不断增加而导致的系统使用性能下降等问题,例如2013年5月,就出现了一次因数据量剧增,导致系统性能下降,并致使系统出现周期性的宕机,严重影响型号正常研制工作的事件。

面对这些问题,一般性的处理措施就是通过提升硬件配置,以及软、硬件厂商共同参与进行性能调优来改善系统的运行性能。基于此,最近5年从IT基础设施改造、应用系统调优等方面展开了一系列的工作,取得了较好的效果。但随着系统内产品数据的进一步扩增,以支撑环境优化调整的手段开始难于奏效;放眼未来,即便是通过硬件升级改造,应用系统性能调优以及负载均衡等传统手段都将很难保证复杂系统的可靠性及持续性。

为此,业界提出一种观点:从产品数据、业务数据自身出发,分析数据的业务类别、作用范围及其使用频度,从而在系统中智能化地支持数据的管理、检索和使用。通过数据的活

跃程度来假定(推测)数据的使用可能性这一基本思想来逻辑性地缩减候选数据范围,从而提高数据的检索和使用效率。该思想目前暂时被称之为“复杂系统数据活跃度管理”。该方法能较大程度上减轻数据搜索和使用的性能,解决了系统性能的瓶颈问题。数据复杂程度如图 1 所示,近 10 年活跃数据比变化趋势如图 2 所示。

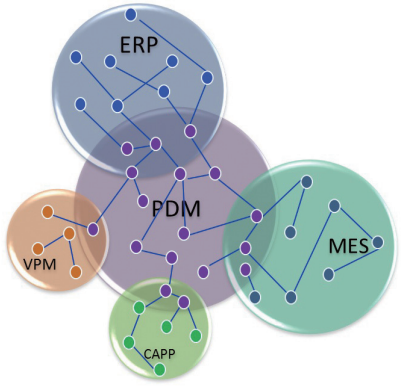


图1 数据复杂程度示意图

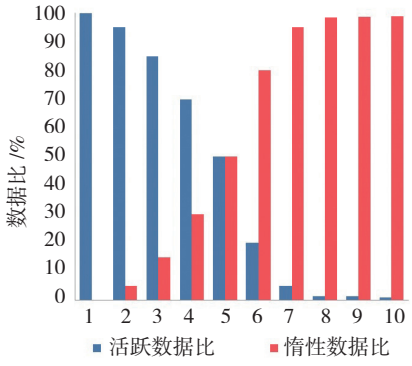


图2 近10年活跃数据比变化趋势

技术背景

1 大数据时代的来临

进入 2012 年以来,大数据(Big Data)一词越来越多地被提及与使用,人们用它来描述和定义信息爆炸时代产生的海量数据。随着数字化工程在制造企业的深化应用,企业中的各种数据正在迅速膨胀并变大,它决定着企业的未来发展,虽然现在企业可能并没有意识到数据爆炸性增长带来的问题隐患,但是随着时间的推移,它已经在逐渐影响人们的工作模式和工作效率。最终,人们将越来越多地意识到大数据对企业的重要性。大数据时代对人类的数据驾驭能力提出了新的挑战,也为人们获得更为深刻、全面的洞察能力提供了前所未有的空间与潜力。全球知名咨询公司麦肯锡在研究报告中指出,数据已经渗透到每一个行业和业务职能领域,逐渐成为重要的生产因素;而人们对于海量数据的运用将预示着新一波生产率增长和消费者盈余浪潮的到来[1]。

据全球最具权威的 IT 研究与咨询公司 Gartner 统计,今后每 5 年数据量将增长 10 倍以上,且其中 85% 将来源于期间产生的新数据类型。2012 年,大数据就带动了全球 280 亿美元的 IT 支出,至 2016 年,这一数字将达到 2320 亿美元。全球 IT 巨头都已意识到了大数据时代的来临,也意识到了大数据的重要意义。包括 EMC、惠普、IBM、微软在内的全球 IT 巨头纷纷通过收购“大数据”相关厂商来实现技术整合,亦可见其对大数据的重视。大数据发展趋势如图 3 所示[1]。

大数据主要有如下 3 个特征[1]。(1)数据类型繁多:包括网络日志、音频、视频、图片、地理位置信息等,多类型的数据对数据的处理能力提出了更高的要求。(2)数据价值密度相对较低:如随着物联网的广泛应用,信息感知无处不在,信息海量,但价值密度较低,如何通过强大的机器算法更迅速地完成数据的价值“提纯”,是大数据时代亟待解决的难题。(3)处理速度快、时效性要求高:这是大数据区分于传统数据挖掘最显著的特征。

对于实施 PLM 的企业来说,随着产品数量的增加及应用的深入,大数据的问题也日益显现,随之而来的是系统架构的复杂性、应用的复杂性、实施的难度等都成为制造企业所无法回避的问题。重视大数据问题,探求适合本企业的大数据管理解决方案,包括系统架构、数据挖掘和分析等,提前为大数据时代做好准备,将是未来制造企业所面临的前所未有的一大机遇[1]。

因此,对企业而言,大数据是对现有数据管理模式的挑战,同时也是一种全新机遇,对海量数据的分析、运用将成为未来企业竞争和增长的基础[1]。

2 活跃度概念的提出

大数据时代面临着这样的问题:如何从海量规模、多样性和快速流量的数据集中抽取有用的信息。

(1)数据挖掘技术[2]。

随着信息技术的迅速发展,数据库的规模不断扩大,从而产生了大量的数据。为了给决策者提供一个统一的全局视角,在许多领域建立了数据仓库,但大量的数据往往使人们无法辨别隐藏在其中的能对决策提供支持的信息,而传统的查询、报表工具无法满足挖掘这些信息的需求。因此,需要一种新的数据分析技术处理大量数据,并从中抽取有价值

图3 大数据发展趋势

的潜在知识，数据挖掘（Data Mining）技术由此应运而生，数据挖掘技术也正是伴随着数据仓库技术的发展而逐步完善起来的。但是并非所有的信息发现任务都被视为数据挖掘，例如，使用数据库管理系统查找个别的记录，或通过因特网的搜索引擎查找特定的 Web 页面，则是信息检索（Information Retrieval）领域的任务。

数据挖掘以数据库、人工智能、数理统计、可视化 4 大支柱技术为基础。描述或说明一个算法设计分为 3 个部分：输入、输出和处理过程。数据挖掘算法的输入是数据库，算法的输出是要发现的知识或模式，算法的处理过程则涉及具体的搜索方法。从算法的输入、输出和处理过程 3 个角度，可以确定数据挖掘主要涉及 3 个方面：挖掘对象、挖掘任务、挖掘方法。挖掘对象包括若干种数据库或数据源，例如关系数据库、面向对象数据库、空间数据库、时态数据库、文本数据库、多媒体数据库、历史数据库，以及万维网（WEB）等。挖掘方法可以粗分为：统计方法、机器学习方法、神经网络方法和数据库方法。统计方法可细分为：回归分析、判别分析等。机器学习可细分为：遗传算法等。神经网络方法可细分为：前向神经网络、自组织神经网络等。数据库方法主要是多维数据分析方法等。

数据挖掘是指从数据集合中自动抽取隐藏在数据中的那些有用信息的非平凡过程，这些信息的表现形式为：规则、概念、规律及模式等。它可帮助决策者分析历史数据及当前数据，并从中发现隐藏的关系和模式，进而预测未来可能发生的行为。数据挖掘的过程也叫知识发现（Knowledge Discovery in Database，KDD）的过程，它是一门涉及面很广的交叉性新兴学科，涉及到数据库、人工智能、数理统计、可视化、并行计算等领域。数据挖掘是一种新的信息处理技术，其主要特点是对数据库中的大量数据进行抽取、转换、分析和其他模型化处理，并从中提取辅助决策的关键性数据。数据挖掘是知识发现（KDD）过程中的一个特定步骤，它用专门算法从数据中抽取模式（patterns），并不是用规范的数据库查询语言（如 SQL）进行查询，而是对查询的内容进行模式的总结和内在规律的搜索。传统的查询和报表处理只是得到事件发生的结果，并没有深入研究发生的原因，而数据挖掘则主要了解发生的原因，并且以一定的置信度对未来进行预测，用来为决策行为提供有利的支持 [2]。

活跃度数据管理模型从根本上即为数据挖掘的一种算法，与传统算法相比融入了产品研发数据的多维度属性，从而获取有用信息供用户使用 [2]。

（2）结合实际情况的活跃度概念提出。

协同工作平台经过长期的应用后，数据量越来越大，数据库中有效数据的比例越来越低，系统检索性能也逐渐加大，结合数据的实际应用情况，客户提出采用数据活跃度的概念提取系统中的有效信息，纳入到活跃数据库管理，提升系统的使用效率。

现状概述

1 系统架构

协同工作平台采用 J2EE 标准的 3 层架构：客户层 / 服务器层 / 数据库层。

（1）客户层主要的应用为：协同工作平台数据的创建 / 查询 / 统计 / 流程签审，该工作主要在 IE 上进行操作；CATIA 设计数据的提交，该工作设计人员在 CATIA 集成设计环境中完成设计，通过集成模块 WGM 将设计数据提交到协同工作平台。

（2）服务器层的主要服务为：Apache 服务，主要应用于用户认证，负载均衡的配置等；Tomcat 服务，编译 JSP 源页面，接收用户的访问请求并返回用户访问结果；WindchillDS 服务，轻量级目录访问协议（LDAP），用户管理用户认证信息、用户群组信息；Windchill 服务，用户执行用户访问后的具体业务逻辑，并通过对数据库的访问进行数据的读写操作。

（3）数据库层的主要应用为：外部电子仓库，存储用户上载的业务数据的电子文档，是协同工作平台最主要的数据存储方式；Oracle 数据库，存储业务数据的基本信息，提供用户的读写访问。

2 总体思路

考虑到系统中型号数据量虽然繁多复杂，但针对不同时期或时间段所关注的数据具有一定的单一性，即可以区分出数据的活跃度。通过针对活跃数据的管理，能降低系统的负载压力。

总体解决思路为：基于型号、型号生命周期、数据类型、关注程度等多个维度的动态数据活跃度定义，见图 4。

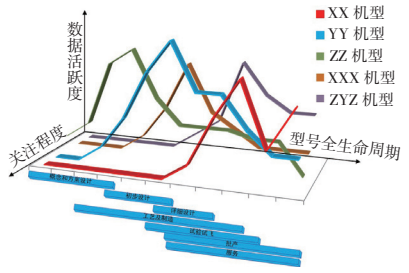


图4 多维度的动态数据活跃度定义

基于协同工作平台的动态活跃数据管理系统架构分为应用层、数据分析挖掘层、基础数据库层，未来还可以根据业务需要，发展基于 ERP、MES 等多系统的综合动态活跃数据管理平台，见图 5。

应用层：基于已定义并提出的活跃数据库，优化检索查询、文件管理、报表管理等整体应用性能。

数据分析挖掘层：通过数据活跃度定义、分析，综合多维度数据活性定义，挖掘、提取、清理活跃数据，使活跃数据库能够动态更新，是动态

活跃数据管理系统的“发送机”。

基础数据层：包括完整数据库、动态活跃数据库、全文检索库等基础应用数据层。

3 活跃度数据管理的数学模型

产品数据的活跃度是指在一定期间内该数据被访问的频度。在实际应用过程中，往往还需要考虑该数据所处的上下文环境以及数据本身所具备的业务价值等，因此广义的数据活跃度是综合考虑数据重要性、价值和访问频度的一个衡量指标。

活跃度数据管理就是要求基于活跃值，提出一个综合的衡量计算模型，能准确、有效地定义数据的活跃程度。这是典型的多属性决策算法，而本算法的最大难点在于属性的不确定性，且各属性属于不同的度量空间。

3.1 活跃度参数集定义

$$A_P = \{\frac{1}{I_P}|\frac{1}{T_P}|P_P|U_P|\frac{1}{O_P}|V_d\} \quad 。$$

式中，I_p 为产品重要度；T_p 为业务对象重要度标志；P_p 为数据的成熟度；U_p 为用户重要核心程度；O_p 为操作关键度标志；V_d 为同一业务对象在持续一段时间内的数据平均访问次数。

根据实际情况，活跃度参数初始定义如表 1 所示。

3.2 活跃度参数集权重定义

考虑到不同参数对活跃度的影响程度会有所不同，因此为每个属性设定了相应的权重，以体现各参数对活跃度的影响程度。活跃度参数集权重集定义如下：

$$A_W = \{I_{W_p}|T_{W_p}|P_{W_p}|U_{W_p}|O_{W_p}\} \quad 。$$

3.3 数学模型定义

基于上述因素，采用加权平均算法，得出活跃度数学模型：

$$A = \sum_1^{V_d}(\frac{\sum_1^5(A_{JP} \times A_{W_t})}{\delta})_j = \sum_1^{V_d}((\frac{1}{I_P} \times I_{W_p} + \frac{1}{T_P} \times T_{W_p} \times P_P \times P_{W_p} + U_P \times U_{W_p} + \frac{1}{O_P} \times O_{W_p})/\delta)_j \quad 。$$

3.4 活跃数据基准系数 δ

由于系统中历史数据的访问次数是无法评估的，因此，为了判断数据的活跃度，设定被访问次数超过 5 次以上的数据才是活跃数据，从而给出活跃数据的基准系数：δ=5。基准系数根据实际情况定义，可根据实际业务情况进行调整。

确定为活跃的产品数据及该产品数据所关联的业务对象将转存入活跃数据库的各索引区，提供业务应用。

根据时间的推移，数据的活跃度将不断地进行重新计算演化。其逻辑框图如图 6 所示。

4 整体处理流程

产品数据的活跃度管理过程见图 7。

数据活跃度的管理从用户访问出发，采集用户的日常访问信息（日常操作 + 用户的关键操作），作为数

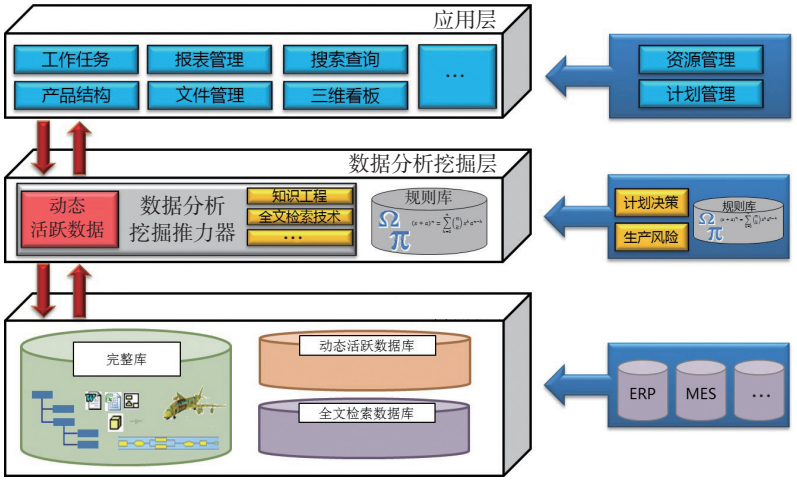


图5 基于动态活跃数据管理的3层系统架构

表1 活跃度参数及其对应的值

参数	业务对象	预设值	权重值 /%	备注
I_p	XX 产品	3	20	该值为初始值，可在管理功能中调整，取值范围为：1– 重要，2– 一般，3– 不重要
	XXX 产品	1	20	该值为初始值，可在管理功能中调整，取值范围为：1– 重要，2– 一般，3– 不重要
	⋮		⋮	
T_p	文档	1	20	该值为初始值，可在管理功能中调整
	零部件	2	20	该值为初始值，可在管理功能中调整
	⋮			该值为初始值，可在管理功能中调整
P_p	编制中	1		该值为初始值，可在管理功能中调整
	审核中	2		该值为初始值，可在管理功能中调整
	已发布	3		该值为初始值，可在管理功能中调整
	⋮			该值为初始值，可在管理功能中调整
U_p	设计师	2		该值为初始值，可在管理功能中调整
	审核者	1		该值为初始值，可在管理功能中调整
	⋮			该值为初始值，可在管理功能中调整
O_p	下载			该值为初始值，可在管理功能中调整
	查看			该值为初始值，可在管理功能中调整

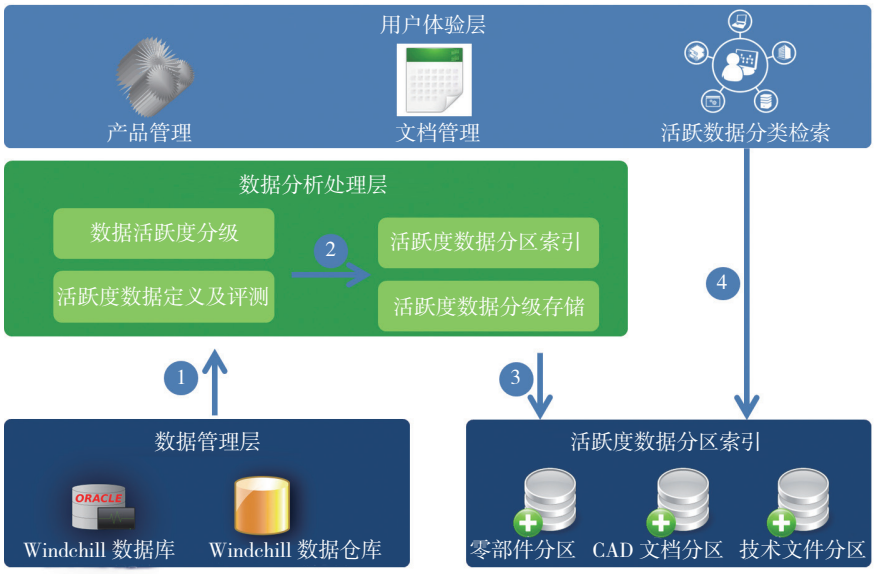


图6 逻辑框图

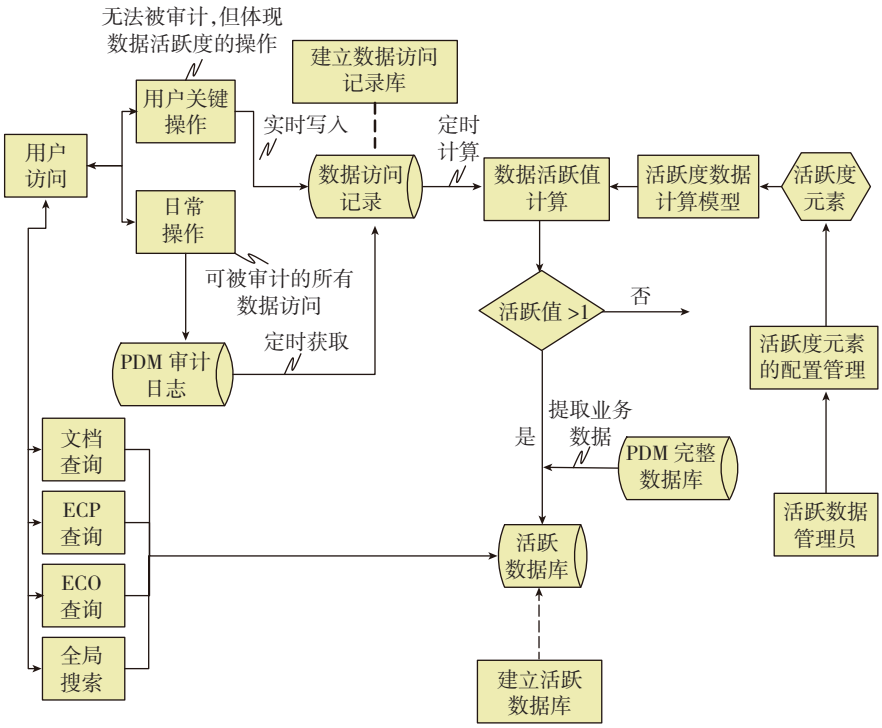


图7 处理流程图

据活跃度的基础数据源。

建立数据访问记录库，用来存储用户对业务数据对象的访问信息，并计算每次用户访问的活跃值。

通过活跃值的累加计算，某个对象的活跃值 >1 后，系统将自动将这些数据加入到活跃数据库中。

活跃数据库中记录了所有主要业务对象的基本信息，可用来作为用户快捷搜索的数据来源。

活跃数据库建立后，用户的日常查询访问，如文档查询、ECP 查询、ECO 查询、全局查询等就可以使用活跃数据库进行，以快速返回用户的查询结果。

结束语

协同产品数据的活跃度管理的实施，将作为公司在大数据管理应用的探索和研究，为未来大数据管理的深化应用提供了理论基础和实践经验。

（1）定义动态活性数据规则库，通过数据活跃度定义和数据活跃度分析，定义协同工作平台活跃数据与惰性数据；（2）通过对活性数据定义、标识，形成独立存储的活跃数据库，极大提升了数据检索、下载等应用效率；（3）通过数据活跃度的管理实施，区分出系统中活跃数据与惰性数据，并对活跃数据进行统一管理；（4）通过改变应用层对后台数据的访问层级，提升系统的操作响应能力；（5）实现活跃数据库的定期清理，大大降低了活跃数据库急剧膨胀的风险；（6）基于动态活跃数据库的应用层扩充改造，提升应用性能，扩大检索范围；基于动态活跃数据库的多维度数据展示活跃度。

未来有望在以下几个方面进行探索和深化应用：

（1）商业智能 BI 的全面深化应用，为未来移动终端的推广提供坚实的基础；（2）通过数据分析改善现有管理模式，实现向大数据环境下的全新产品数据管理模式的转变，以提高产品和服务质量；（3）以分析型型号数据为基础，优化现有产品组织模式，科学配置制造资源，构建产品研制数据监控分析模型；（4）建立各种针对产品研制的系统性算法模型库，发掘数据中存在的隐藏关系，为各级决策者提供多维的、直观的、全面的、深入的分析预测性数据，进而主动把握市场动态，采取适当的策略，获得更大的企业效益。

参 考 文 献

[1] 迈尔·舍恩博格，库克耶．大数据时代．杭州：浙江人民出版社，2013.

[2] Tan P N, Steinbach M, Kumar V, et al. Introduction to Date Mining. 北京：人民邮电出版社，2006.

（责编 宁军）