数据仓库技术及其在银行业的应用

NCR Teradata 数据仓库事业部

目前，国内各个商业银行正面临着前所未有的激烈的市场竞争，而与此同时，随着中国加入WTO，金融自由化、国际化的速度也正在逐渐加快。不久的将来，国内各商业银行除了彼此之间相互竞争以外，还将迎接来自许多世界级外资银行的挑战。

利用先进的数据仓库技术建立集中的、包含详细交易数据的商业智能解决方案，已经成为各大银行对内加强经营管理和决策支持，对外更好地了解客户需求，开发新产品或服务，利用现有渠道对客户进行交叉销售，增加赢利能力，并在特定的业务领域提供差异化服务的重要手段。

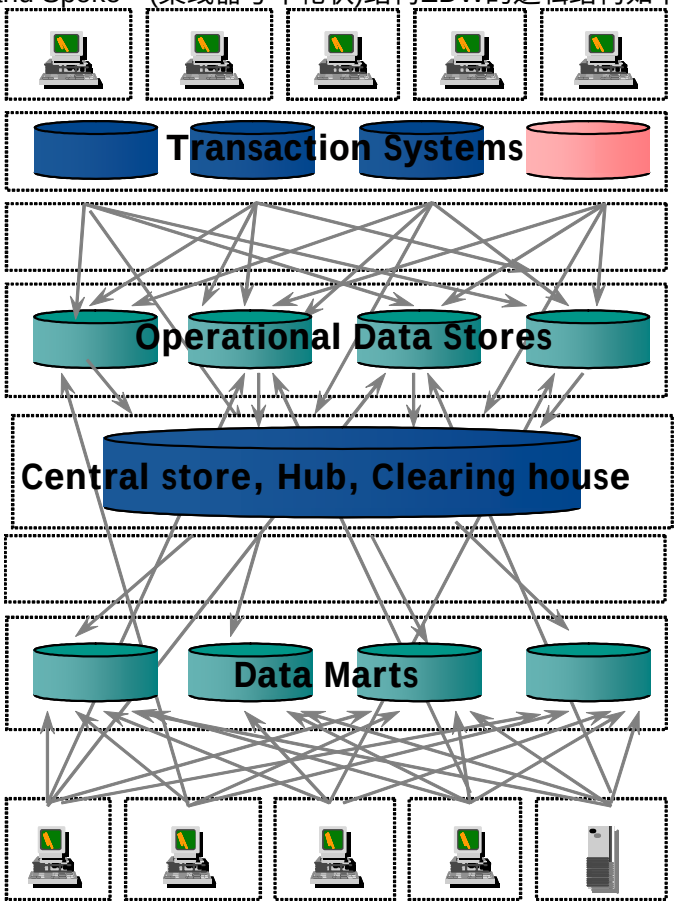
与前几年不同的是，大家目前都在谈论企业级数据仓库(Enterprise Data Warehouse)，对于数据集市的定位也基本形成共识，那就是数据集市应该从属于企业级数据仓库。所谓EDW，基本的要求是整个企业能够共享统一的数据存储模型，为各级业务人员提供一致的信息视图。实施时可以先按照需求的轻重缓急选择部分业务主题，然后逐步扩展到涵盖全部业务。

本文对业界常见的两种EDW架构作了分析，并探讨了银行业数据仓库的应用体系。

一、 两种主要的企业级数据仓库体系架构

1.1 集线器与车轮状结构的企业级数据仓库(Hub and Spoke)

“ Hub and Spoke ” (集线器与车轮状)结构EDW的逻辑结构如下图所示。



之所以把这种结构称为“ Hub and Spoke ”，是因为中央数据库汇集了来自各业务处理系统的数据，同时也负责向各从属数据集市提供信息，看上去象一个 Hub (集线器)一样。而业务人员在进行数据分析与信息访问时将根据需要连接到不同的数据集市，这种交叉复杂的连接看上去就象 Spoke(车轮辐条)一样。由于这样的关系，著名评估机构 Gartner Group 把这种结构的数据仓库形象地称为“ Hub and Spoke Data Warehouse ”。

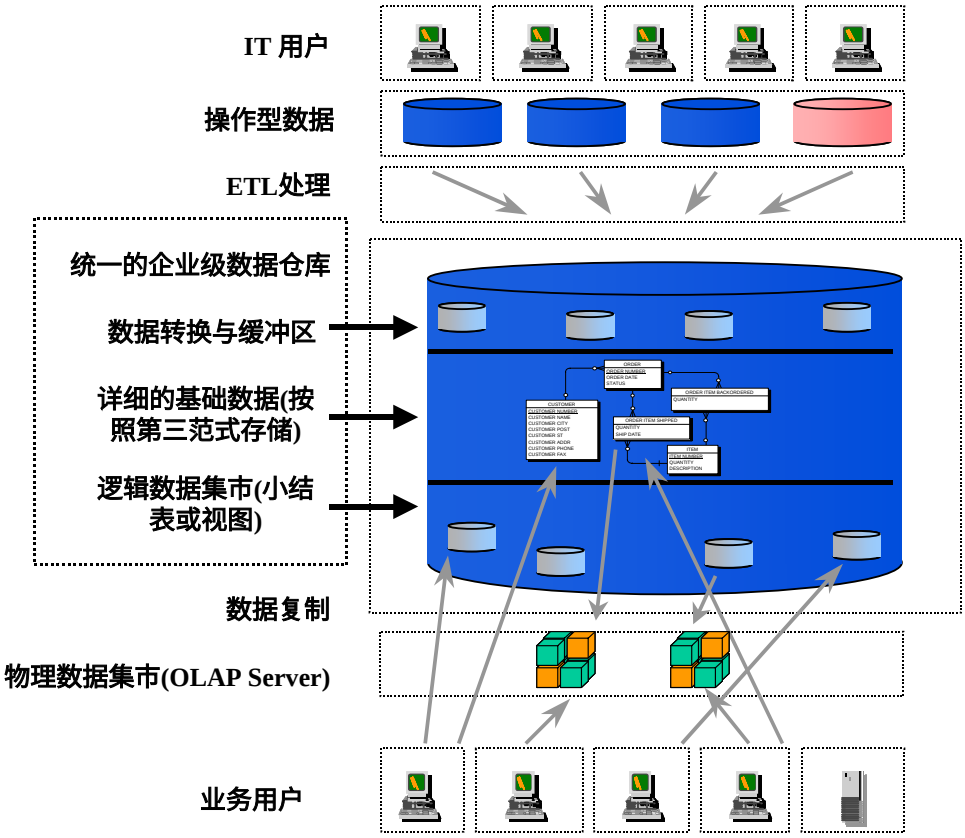
“ Hub and Spoke ”结构解决了企业内统一数据存储模型的问题，但从实际使用的角度来看仍有比较严重的缺陷。主要体现在两方面：一是业务人员对信息的访问非常不方便，很难进行跨数据集市或跨部门的信息分析。数据集市的存储模型需要根据预先定义的分析需求进行规划和设计，业务人员根据分工到指定的数据集市上去访问相关信息。如果需求发生变化，就需要对数据集市重新规划。这显然不能满足日益变化的市场需求。中央数据库只是起统一数据存储和刷新数据集市的作用，一般不提供信息访问。另一个问题是每个数据集市都需要相应的软硬件投入，当数据集市增加时，系统整体投资迅速增加，同时管理的复杂性也随之增加。这些都意味着巨大的整体拥有成本 TCO(Total Cost of Ownership)。

为什么不直接访问中央数据仓库而非要设计一个数据集市层呢？主要原因在于当中央数据库保存越来越多的数据、并发用户越来越多时，一般的数据库引擎无法承担这样的负载，只好把它们分解到不同的数据集市。

对于“ Hub and Spoke ”结构的数据仓库，Gartner Group 曾发表过不少研究报告加以分析。Gartner Group 认为，“ 数据仓库的 Hub and Spoke 结构，回避了 DBMS 技术中的弱点，无法提供适当的业务价值来平衡投资成本的显著增加(The "hub and spoke" trend in data warehouse topology circumvents weaknesses in DBMS technology and fails to deliver benefits to justify notable cost increases) ”，“ 之所以产生这种趋势，是由于对大多数 DBMS 产品而言，支持复杂的数据模型和并发查询负载都是极大的挑战(This recent trend is due to the challenge for most DBMS products to support complex data models and concurrent query workloads) ”。“ 概要来说，企业不应轻信有关这种拓扑结构的过份宣传，而忽视按照支持当前与今后的并发用户访问环境的需求，对 DBMS 产品进行全面、综合的评估与选择(Bottom Line… Enterprises should not let the hype of yet another alternative topology take away from a comprehensive DBMS evaluation and selection that will provide support for current and future user concurrency requirements) ”。

1.2 集中式企业级数据仓库

第二种企业级数据仓库的架构是集中式的，其逻辑结构如下图所示：



与前面讨论的“ Hub and Spoke ”结构相比，主要的差别在于：

1、 数据集市分成物理与逻辑两种，物理数据集市设立在中央数据仓库之外，具有专门的软硬件设备。一般都使用 OLAP 服务器，按照特定需求组建多维立方体来提供多维信息分析。逻辑数据集市设立在中央数据仓库之内，由在基础数据之上形成的小结表或者逻辑视图组成。业务人员既可以访问多维立方体，也可以访问中央数据仓库内的小结表或者逻辑视图。这些分析主要针对预先定义的业务需求，并且粒度比较粗。基于 OLAP 服务器的数据集市比基于 RDBMS 的数据集市要容易维护得多，当然规模也相对较小。

2、 中央数据库采用符合数据库范式理论(一般为第三范式)的存储模型来保存基础数据，从而为整个企业提供一致的信息视图。上面说明的数据集市主要针对粒度较粗、预先定义的分析需求，对于动态的业务查询、粒度较细的或者针对基础数据的分析需求则由中央数据库提供。因此业务人员可以直接访问到最基础的详细数据，特别是高级业务分析师(Power User)，将更频繁地基于详细数据进行分析，以便挖掘出内在的、隐含的业务规则，帮助企业主管更好地进行业务决策。

3、 在中央数据库中还设立了一个数据转换与缓冲区(Data Staging Area)，作为 ETL 处理的一部分。由于在很多数据仓库的 ETL 处理流程中，需要对源数据作一些比较复杂的转换与清洗工作，如果仅借助于 ETL 工具实现这种转换与清洗，由于没有数据库的支撑(ETL 工具均在数据库之外运行)，经常会产生比较严重的性能问题。于是在一些系统中增加一个 ODS(Operational Data Store)层来进行数据的整理，但这就像设立基于 RDBMS 的数据集市一样，将大大增加整体投资和管理复杂性。理想的方法是如上图所示，在中央数据库中设置一部分存储空间来作为数据转换与缓冲区，借助数据仓库引擎强大的复杂查询处理能力，通过

SQL 实现数据的转换与清洗。这种实现方法简单、快速、并且不容易出错，当然对中央数据仓库引擎的处理能力就提出了更高的要求。

这种集中式的数据仓库结构解决了“ Hub and Spoke ”结构中存在的诸多问题，是一种比较理想的企业级数据仓库系统架构，能够为企业带来真正的业务价值与回报。但由于把详细数据分析、部分的数据转换与清洗等复杂处理均集中在中央数据仓库，从而对作为数据仓库引擎的 RDBMS 和相应的服务器带来了极大的挑战。选择这种数据仓库基础平台的基本要求是：

1、 线性扩展能力。原始数据对任何一个数据仓库来说，都是最主要的负载之一。随着数据量的增长，系统性能会逐渐下降。为了维持合理的业务查询响应时间，要求数据仓库引擎和相应的数据库服务器具有优良的线性扩展能力。一些系统的扩展能力非常有限，当数据量增长到一定规模时（比如 TB 级以上）已经很难满足日常的业务分析要求，不得不把数据分离到多个小规模的数据集市，形成所谓的“ Hub and Spoke ”结构。

2、 并行处理能力。许多业务查询与分析都是动态(Ad-hoc Query)的，数据库传统的索引技术对动态分析和模糊查询的帮助不大。系统只有具有非常好的并行处理能力，才能满足复杂的、动态的分析需求，并且承担比较复杂的数据转换与清洗工作。

3、 简单的系统管理。对于大型的数据仓库应用系统而言，如何能有效而简单地进行系统管理是非常重要的。特别是当数据量不断扩大时，如果没有一种有效而且简单的系统管理措施，那么系统的运行费用将会很高。

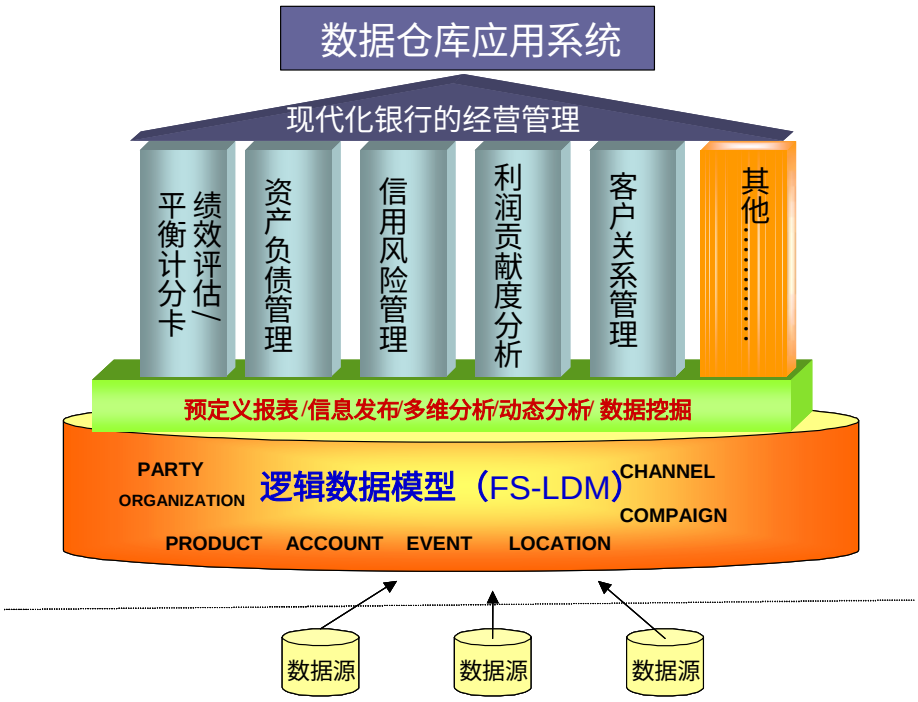
二、 数据仓库技术在银行业的应用

数据仓库体系结构属于基础设施的建设，只有稳固的数据仓库基础设施才能支撑灵活多样的数据仓库应用。对于银行业来说，数据仓库的应用面非常广，基本上涵盖了银行经营管理与业务运作的各个方面。

现在国内几大商业银行都在着手调研、准备或者尝试实施基于数据仓库技术的各种解决方案。比如，中国工商银行进行了以个人客户关系管理（PCRM）和业绩价值管理（PVMS）为主题的应用试点，中国银行则全面规划了信用卡系统，其中很重要的一个子系统就是基于数据仓库技术的销售和客户服务系统，中国农业银行正在广东分行进行经营分析系统的建设，中国民生银行也全面启动了客户信息管理（CIM）和企业级数据仓库的建设。

银行通过逐步建立企业级数据仓库，可以对全行业务数据进行集中存储和统一管理，科学合理地对信息进行详细分类，及时准确收集信息和分析信息，确保管理层随时掌握银行的经营风险、运营情况和经营目标。在引入详细交易数据以后，可以通过各种数据的关联分析，衡量各类客户需求、满意度、赢利能力、潜在价值、信用度和风险度等指标，帮助银行识别不同的客户群体，确定目标市场，为实施差别化服务、产品合理定价的策略提供技术支持。

根据国内外银行使用数据仓库的经验，银行业数据仓库应用的体系和分类大体如下图所示：



2.1 平衡计分卡与绩效评估

“平衡计分卡”将绩效评估指标分成四个重要的层面：财务层面（Financial）、客户层面（Customer）、流程层面（Processes）及员工学习与成长层面（Learning and Growth），然后从全行观点、各业务部门观点、地理区域观点、地区分行/部门观点及至个别员工等不同级别，由上而下由粗而细区分为不同等级的绩效衡量指标，这样既可以提供银行经营所需的信息，又不会使信息过于庞杂而失去效用，更重要的是可以促进策略与远景目标的实现。

2.2 资产负债管理

在一个效率市场中经营的商业银行必然会面临大幅度的价格波动，这将对银行的（净）收益和资产、负债以及一些资本的价值产生巨大的影响。如果严重的话，还可能会使银行面临很大的偿付风险。另外，随着当今世界对银行经营的监管力度的不断增强，各监管机构和国际公约都相继对如何控制银行经营中的各项风险作出了具体的规定，甚至还对一些关键性的指标（如“资本充足率”等）给出了建议的算法和具体数值，这些都使得资产负债管理在银行中的地位越来越重要。

“资产负债管理”应用模块的主要任务就是帮助银行科学考核和管理银行自身资产、负债以及由于经营活动而产生的市场风险、外汇与流动性风险，寻找建立在合理风险回报基础上的资本分配方法，从而使银行能够很好地控制经营风险并提高利差的收益回报，在流动性、安全性、盈利性的经营原则中寻找到一个最佳的平衡点。

2.3 信用风险管理

商业银行经营的最终目标是为了获取最大的利润，而贷款业务是银行最主要的利润来源之一，因此它的质量和收益对银行的兴衰成败有着至关重要的影响。有效降低信用风险，提高贷款质量，是银行取得利润最大化的关键因素。

“ 信用风险管理 ” 应用模块通过对全行信贷数据的分析，准确识别、计量和控制信用风险并实现风险的组合分析和相关分析，从而确定合理的贷款结构和适当的利差，制定规范、科学、有效的贷款政策

2.4 利润贡献度分析

利润贡献度分析的主要目标是帮助银行了解其利润贡献度构成因子的分布状况，使行领导能够很容易地从不同的角度进行绩效评估，制定相应的经营策略，并进一步完善分行及业务部门的自身分析和流程规划。

传统的利润贡献度分析是从总帐系统出发，通过分摊的方式来进行计算。这种分析太过粗糙，一些国际先进银行已经利用基于行为的成本稀量方法（ABC：Activity Based Costing），在详细交易数据的基础上用净利息收入、其他收入、直接费用、间接费用和风险准备等五大因子来计算每个帐户的利润贡献度，并进一步归整到每一位客户、每一项金融产品、每一种渠道和每一个机构对银行的利润贡献度。

通过实施这样的 “ 利润贡献度分析 ” 应用模块，可以帮助银行建立精确的、全行一致利润贡献度评估方法论，从而得到关于客户、产品、分行、部门利润贡献的准确信息及影响因素。获取的这些分析结果在银行市场、财务、规划、风险和产品管理等部门的决策支持中占有不可或缺的重要地位。

2.5 客户关系管理

“ 客户关系管理 ” 应用模块通过分析数据仓库中各种数据信息以及相互之间的关联，从多个方面衡量各类客户的忠诚度、满意度、赢利能力、潜在价值、信用度、风险度等关键性指标和需求差异性，为银行制定正确的市场行销策略提供科学的决策支持。

客户关系管理是一个动态处理过程，它不仅包括一系列产品和服务，还要求银行要能够完整地认识整个客户生命周期，并提供与客户沟通的统一平台，提高员工与客户接触的效率和客户反馈率。

实施 “ 客户关系管理 ”，可以帮助银行建立全行的客户单一视图，提供清楚准确的客户轮廓，帮助各业务部门了解、分析客户，更好地进行客户细分，提高客户行销和服务水平，在适当的时间，通过适当的渠道，为客户提供适当的产品和服务，从而增强市场的综合竞争能力。最终实现以产品为中心的经营模式向以客户为中心的经营模式的转变。

三、 小结

数据仓库的实施是一个长期的过程，在基础设施建立完成后，随着应用的逐步开展和深入，其投资回报也逐步增加。在一次数据仓库用户大会上，香港东亚银行的CTO在结束其演讲时深有体会地谈到，东亚银行花了两至三年的时间来完善其数据仓库的基础设施，现在终于得到可观的回报。许多同行到东亚交流数据仓库使用经验时，常常羡慕其先进的客户关系管理系统及其产生的效益，却往往忽略了东亚银行在前期建置数据仓库基础设施时所耗费的时间、心血和投资。

对于国内各大商业银行而言，我们同样需要一定的时间来建立数据仓库基础设施，并在建置的过程中逐步完善数据质量。这种打基础的过程是无法省略的。更为重要的是，在建立数据仓库的过程当中，我们还可以培养一批既懂数据仓库技术、又精通银行业务的高级分析人才，这对于更好发挥数据仓库价值是非常重要的。