6.湖南省烟草商业系统数据中心建设研究

侯杰华、邹暾、姚利军

一、背景

根据国家局2006年全国烟草行业信息化工作会议的总结讲话和2007年全国烟草行业信息化工作会议上的《突出应用，加强服务，全面推动行业数据中心建设》的讲话的精神，按照国家局对全国当前及今后一段时间的信息化工作部署安排，要求各单位认真落实“建设数据中心，突出重点工程，全面加强管理”，加快推进信息化建设步伐，建立统一标准、统一平台、统一数据库、统一网络的数据中心，有效整合信息资源，以更好地发挥信息化在深化行业改革、改造传统产业、强化基础管理、加强内部监管等方面的支撑和促进作用。

数据中心建设既是国家局领导的要求、行业改革与发展的需要，也是行业信息化建设深入发展的必然。根据国家局下发的《烟草行业数据中心建设实施意见》的要求，需要尽快建立行业两级（国家级、省级）数据中心体系。结合国家局要求及我省的实际情况，建设我省数据中心系统势在必行。

二、湖南烟草商业系统信息化建设情况

省局（公司）所辖14家市州局（公司），经过多年的信息化建设，已建成了专卖、营销、烟叶、办公自动化、财务等信息系统，并实施了应用集成，实现了数据集成与界面集成。目前，已实现了各系统的编码管理与对象管理，以及各系统间的数据交换体系，可以通过单点登录，在同一浏览器界面里访问所有应用系统。

硬件平台方面，湖南烟草系统已建立了省-市-县三级网络，各级之间采用2M的MSTP专线。省局各系统运行在IBM P670上，市州局各系统运行在IBM P650上。

软件环境方面，目前各系统均建立在IBM DB2的数据库上，采用J2EE架构，运行环境在DB2、WAS、MQ和中软R1构成的软件平台上。其中，除专卖采用省级集中外，其它各系统均采用市级集中部署方式。

在“两烟”生产、经营和专卖管理工作中，各系统经过长时间的运行，积累了大量的数据。平均每个市州局的数据总量为100GB左右，每局每天约产生30MB左右的数据。这些数据蕴藏着丰富的使用价值，是指导企业进一步发展的重要依据，应当进行充分的挖掘和利用。

三、现有系统数据利用的问题

（一）可信性问题

企业管理者为了解企业运营状况，需要从各业务系统内抽取业务数据进行经济运行分析。但直接从各业务系统中抽取数据往往出现这样的情况：从不同部门来的报表的结论不相吻合，而且相去甚远。数据缺乏可信性有如下几个原因：

数据无时间基准：各部门、各系统抽取数据的时间不一致；

数据算法上的差异：各系统、部门的统计算法不一致；

抽取的多层次问题：从数据源到为决策者提供的分析结果，经过了多次抽取，进一步恶化了上两个因素造成的后果；

外部数据问题：对外部数据的来源没有进行记录，原始数据成了数据源不明的不可靠数据；

无公共起始数据源：数据源不一致，各部门从不同的系统中抽取数据，但这些系统的数据没有数据同步或数据共享。

（二）效率问题

对于现有业务系统，如要从业务系统直接进行报表展现和数据分析处理，必须先要进行数据定位，分析很多文件和数据的布局，并对数据进行分析和合理化处理。因此，如需要进行报表展现和数据分析，就只能依靠对系统底层数据非常熟悉的IT公司开发人员。而对报表和数据分析有直接需求的用户，由于对系统底层数据不了解，无法自行直接开发报表和进行数据分析。因此，基于业务系统的报表和数据分析功能受到很大的局限。

同时，为从业务系统取得需要的数据，要写的程序数量很多，并且每个程序都需要定制，程序涉及的技术面很广，这就造成开发的工作量和难度都很大。并且，如果IT公司开发人员发生变动，新的开发人员可能对系统结构不熟悉，则更会造成开发工作的延误。

另外，随着报表和数据分析的应用，会有很多新的报表及数据需求提出。由于基于业务系统直接开发报表和数据分析功能的滞后性，无法快速实现领导与管理人员的思路，等到实现以后，现实情况可能已经发生变化，生成的报表和数据分析结果已经失去意义。

（三）从数据转化为信息的能力

业务系统中的数据一般都是明细数据，而分析系统需要的信息一般需要汇总或者经过一定的运算，在现有系统数据中无法直接检索得到。由于目前的业务系统为OLTP（联机事务处理）系统，为保证数据的正确性，实施了数据锁等保护机制，如进行OLAP（联机分析处理）工作，会产生大量数据查询请求，在用户数较少时，对业务系统性能影响尚不明显，一旦用户数较多，则会导致业务系统性能明显下降。

由于目前的业务系统只保有单一业务数据，一旦需要进行经济运行分析，往往需要到各系统中寻找需要的数据，并且很多数据分布在上、下级不同单位及部门之间。这就需要用户在各个系统之间频繁切换，手工进行数据汇总和计算，并且还需要自行解决各系统由于统计口径不一致导致的数据不一致问题。这样的经济运行分析结果，往往不能全面、综合、动态地反映全省生产经营状况，不能为生产经营辅助决策提供有效的依据。

（四）应用需求迫切

目前，省局领导需要了解全省经济运行状况和企业最新情况；省局各业务部门需要了解本部门所辖系统的综合情况以及各市州局的具体情况；市州局领导需要了解本单位的经济运行状况、企业最新情况，以及与兄弟单位的比较情况；市州局各部门负责人及工作人员如客户经理等，需要了解本业务系统的明细、综合分析、需求预测；各级领导与人员对数据展现、分析和报表均有不同需求。

（五）问题的解决方案

把这些对业务数据的应用需求综合起来，可以归结为四个问题：数据的一致性问题、数据的完备性问题、数据展现与分析的手段问题、数据展现与分析的效率问题。这几个问题在各业务系统中无法得到有效的解决，而这正是数据中心要解决的问题。

因此，湖南烟草需要搭建一套以数据仓库技术为基础的数据中心，使管理层及业务人员能够真正从各个观察角度，分析监控全省的经营情况，为省、市州局（公司）领导、各业务部门提供经营决策依据，了解经济运行情况，掌握企业动态，为各业务人员提供统计分析与即时查询，做好预测，更好地开展业务工作。

三、数据中心简介

（一）、什么是数据中心

按照国家局的定义，数据中心是以信息资源标准为基础、信息安全为保障的数据交换服务平台、数据加工存储平台和数据分析应用平台。数据中心主要由数据仓库、数据展现、数据分析及数据挖掘的工具，以及支撑它们的硬件平台组成，其核心部分是数据仓库。

（二）数据中心与应用系统的区别

1、数据仓库与应用系统数据库的区别

数据仓库是面向主题的、集成的、相对稳定的、反映历史变化的的数据集合，用以支持经营管理中的决策制定过程。数据仓库的主要作用是实现数据的集中，并按主题进行存储与管理。数据仓库是为了建立一种体系化的数据存储环境，将分析决策所需要的大量数据从传统的操作环境中分离出来，使分散、不一致的操作数据转换成集成、统一的信息。

数据仓库中的数据面向主题，与传统数据库面向应用相对应。主题是一个在较高层次上将数据归类的标准，每一个主题对应一个宏观的分析领域；数据仓库的集成特性是指在数据进入数据仓库之前，必须经过数据加工和集成，这是建立数据仓库的关键步骤，首先要统一原始数据中的矛盾之处，还要将原始数据结构做一个从面向应用向面向主题的转变；数据仓库的稳定性是指数据仓库反映的是历史数据的内容，而不是日常事务处理产生的数据，数据经加工和集成进入数据仓库后是极少或根本不修改的；数据仓库是不同时间的数据集合，它要求数据仓库中的数据保存时限能满足进行决策分析的需要，而且数据仓库中的数据都要标明该数据的历史时期。

数据仓库最根本的特点是物理地存放数据，而且这些数据并不是最新的、专有的，而是来源于其它业务系统数据库的。数据仓库的建立并不是要取代业务系统数据库，它要建立在一个较全面和完善的信息应用的基础上，用于支持高层决策分析，而事务处理数据库在企业的信息环境中承担的是日常操作性的任务。

2、应用系统数据与数据中心数据的区别：

应用系统数据（操作型数据）	数据中心数据（分析型数据）
面向应用	面向主题
详细的	概要的
在访问瞬间是准确的（当前值）	代表过去的数据，快照（历史数据—）
为日常工作服务	为管理者服务
可修改	不可修改
重复运行操作	启发式运行操作
处理需求预先可知	处理需求事先不知道
对性能要求高	对性能要求宽松
一次访问一个单元	一次访问一个集合
事务处理驱动	分析处理驱动
需要进行更新控制	不需要进行更新控制
高可用性	宽松的可用性要求
整体管理	以子集管理
非冗余性	总是存在冗余
静态结构，可变的内容	结构灵活
一次处理数据量小	一次处理数据量大
支持日常操作	支持管理需求
访问频繁	访问频率较低
细节性数据	汇总或运算后数据

3、应用系统与数据中心用户对需求的区别

应用系统用户对系统的需求是明确的。传统的需求由需求驱动，为建立系统，首先必须理解需求，然后进入设计和开发阶段。

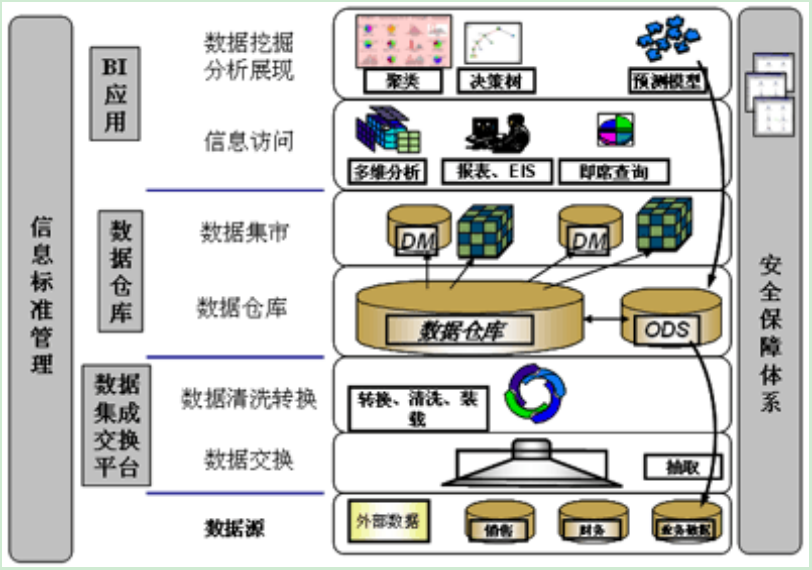
数据中心的用户首先是个商务人员和管理人员，主要工作是定义和发现在企业决策中使用的信息，他是在发现模式下工作，只有看到报表和数据后，才能决定是否有必要进行分析、怎样分析、需要什么数据、什么分析手段。

因此，传统的系统开发生命周期不适用于数据中心的分析领域。在传统的系统开发生命周期中，首先假设在设计之初，需求是已知的，至少是可发现的，而在数据中心的用户眼中，要到数据中心开发生命周期最后才发现真正的需求，要从现有需求开始，将新的需求考虑在内几乎是完全不可能的事情。

数据中心的需求由数据开始，得到数据后，将数据集成，然后检验数据存在什么偏差，之后，针对数据写程序，分析程序的结果，最后系统需求才得到理解。一旦系统需求得到理解，就需要对系统的设计进行调整，然后针对不同的数据集开始新的开发周期。

（三）数据中心系统架构

数据中心系统架构图



数据中心按照功能分为以下几个部分：

1、元数据（Meta Data）：元数据是数据仓库的核心，是关于数据的数据，是关于数据和信息资源的描述信息。它通过对数据的内容，质量，条件和其他特征进行描述和说明，帮助人们有效地定位、评论、比较、获取和使用相关数据。

元数据有多种不同的形式，一种形式是业务元数据，另一种是技术元数据。业务元数据就是对业务人员有用或有价值的元数据，技术元数据就是对技术人员有用或有价值的元数据。

元数据在数据仓库中比在传统操作型环境中更重要。由于数据仓库是在一种启发式、迭代式的开发生命周期上运作的，为了更加有效利用数据仓库，用户应该能够对准确和实时的元数据进行访问。没有一个好的元数据来源来运作的话，决策分析分析人员的工作就非常困难。

2、源数据（Source Data）：指分布在不同的应用系统中，存储在不同的平台和不同的数据库中的大量的数据信息，是数据仓库的物质基础。

3、数据变换工具（ETL）：为了优化数据仓库的分析性能，源数据必须经过变换以最适宜的方式进入数据仓库。在数据仓库的体系架构中，ETL的主要作用在于其屏蔽了复杂的业务逻辑从而为各种基于数据仓库的分析和应用提供了统一的数据接口，这也可以说是构建数据仓库最重要的意义所在。由于ETL 在数据仓库和业务系统之间搭建了一座桥梁，确保新的业务数据能源源不断进入数据仓库，同时用户的分析和应用也能反应出最新的业务动态，虽然ETL 在数据仓库架构中技术含量并不算高，但其涉及到大量的业务逻辑和异构环境，因此在一般的数据仓库项目中ETL部分往往是牵扯精力最多的。

4、数据仓库（DW）：数据仓库一般按三层进行设计：操作数据存储（ODS）、数据仓库（DW）、数据集市（DM）。数据仓库的真正关键是数据的存储和管理。数据仓库的组织管理方式决定了它有别于传统数据库，同时也决定了其对外部数据的表现形式。要决定采用什么产品和技术来建立数据仓库的核心，则需要从数据仓库的技术特点着手分析。针对现有各业务系统的数据，进行抽取、清理，并有效集成，按照主题进行组织。 数据仓库中的数据存在着不同的综合级别，一般称之为"粒度"。粒度越大，表示细节程度越低，综合程度越高。

数据仓库DW(Data Warehouse) 既是一种结构和方法，又是一种技术。各种信息从不同信息源提取出来，然后将其转换成公共的数据模型并和仓库中已有的数据集成，当用户向仓库查询时，需要的信息已准备就绪，数据冲突、表达不一致等问题已经得到解决，这样，决策查询更容易、更有效。

1）ODS

操作型数据存储(ODS)是各种详细数据的集合，用于满足企业综合的、集成的以及部分操作型的处理需求,是一种混合性的结构。它是从业务系统过渡到数据仓库核心层的操作数据的模型，其设计接近于业务系统，其目标是对数据仓库核心层尽量屏蔽不同业务的差异性并降低ETL设计、处理和调度的复杂度，提高ETL性能。ODS一般只存放相对较短时间的历史数　据。

2）DW

DW层也称数据仓库数据库，统一存放所有业务系统的历史数据。系统通过ETL系统采集所有的历史数据，并定期自动把新增加的业务数据装载到DW。基于DW可生成面向不同应用主题的数据集市（Data Mart）。

3）DM

DM层是根据分析主题的需求建立的数据高度聚集的集市层，一般按照星型模型组织，建立对应的事实表及维表。DM层一般是面向部门级应用领域的分析主题，存储的是部门级的数据。

5、OLAP服务器：On-Line Analytical Processing（联机分析、OLAP）是一类软件技术，它们使用户能够以交互形式快速、一致地探查数据，用户看到的是经过转换后的原始数据的各种信息视图，它们可以反映业务的真实维数。OLAP 报告将业务数据结构、过程、算法和逻辑的复杂性集成到了它的多维数据结构中，然后向最终用户呈现容易理解的维信息视图，让他们能够以非常自然的方式探索业务数据。

OLAP具体实现可以分为：ROLAP、MOLAP和HOLAP。ROLAP基本数据和聚合数据均存放在RDBMS之中；MOLAP基本数据和聚合数据均存放于多维数据库中；HOLAP基本数据存放于RDBMS之中，聚合数据存放于多维数据库中。

OLAP服务器可以为用户提供快速的响应和交互式操作，对数据分析中经常使用的诸如求和、总计、平均、最大、最小等操作结果预先计算并存储起来，以便于决策支持系统使用；还可以让用户对数据进行多维分析，可以利用数据的下钻、上钻、旋转、切片、切块操作，从不同的角度分析数据的相关性，可以从整体来分析，也可以详细到最小的数据单元。

6、前端工具：主要包括各种报表工具、查询工具、数据分析工具、数据挖掘工具以及各种基于数据仓库或数据集市的应用开发工具。其中数据分析工具主要实现对数据仓库中的数据进行分析和综合。数据挖掘工具负责从大量的数据中发现数据的关系，找到可能忽略的信息，预测趋势和行为。报表工具主要提供用各种手段展现数据分析与数据挖掘的结果，并制作出各种形式的报表。

四、建立数据中心对应用系统的有利影响：

建立数据中心后，可从生产环境中移走大量历史数据。可使生产环境更易于纠错、易于重构、易于监控、易于索引，更具有可塑性。

建立数据中心后，可从生产环境中移走信息型处理，包括报表、显示、抽取等。可大量降低应用系统的维护与开发工作。

数据中心可实现多个数据源的集成，消除应用层的不一致性。

由于数据中心的数据是非易失性的，因此数据中心可以保留数据的历史变化情况。操作型的应用系统的数据一般是周期性更新的，但数据中心的数据通常在载入后并不进行一般意义上的更新，当应用系统中数据产生后继变化时，一个新的快照就会写入数据中心，从而在数据中心保存了数据的历史状况。

五、建设数据中心需要注意的问题

建设数据中心只能一步步地进行设计并载入数据，是进化性的，而非革命性的，因此，数据中心的建设要采用有序地反复和一步步进行的方式。

数据中心的建设是一个长期的过程，我们在开始阶段只能建立数据仓库的架构，将现有应用系统的数据进行汇聚和重构，并根据目前的应用需求建立一些主题，利用工具对其进行展现和分析。随着应用系统新的数据产生、新的应用系统的建立、用户面的扩展、数据的深入分析挖掘以及新的应用需求的提出，必然会要再次对数据中心数据结构进行优化、建立新的主题与数据集市、设计新的报表以及数据分析模块，乃至围绕数据中心对现有应用系统进行重构。因此，数据中心的建设不能一蹴而就，应该根据目前企业管理需求及战略发展方向，整体规划，分步实施，夯实基础，首先规划与建设好数据仓库，然后在此基础上逐步扩展。

参考文献

1、《数字烟草发展纲要》，国家烟草专卖局，2005.10。
2、《整合资源 提升能力，推动行业信息化建设和谐发展——张保振副局长在2007年全国烟草行业信息化工作会议上的讲话》，国家烟草专卖局，2007.4。
3、《突出应用，加强服务，全面推动行业数据中心建设—高锦主任在2007年全国烟草行业信息化工作会议上的报告》，国家烟草专卖局，2007.4。
4、《烟草行业数据中心建设实施意见》，国家烟草专卖局信息中心，2007.4
5、《数据仓库》，W.H.Inmon 著；机械工业出版社