11

**商** **业** **智** **能** **在** **电** **子** **商** **务** **中** **的** **实** **践** **与** **应** **用**



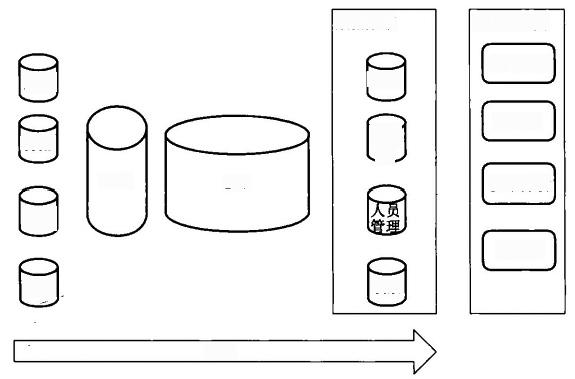
12

**第三章** **商业智能在电子商务的解决方案**

电子商务行业和传统的制造业、零售业无论从商业模式、组织架构等各个方面 都有着很大的区别。早期，很多BI 咨询公司或者实施公司把原先在传统行业的解 决方案直接搬到电子商务行业，企图复制在传统行业的成功，最后发现，不管是项 目实施的成功率还是项目后期的效果都非常的不理想，从而让整个电子商务行业 对于此行业是否适合利用BI 产生了很大的质疑。本章将结合焦点科技实施 BI 以 来的经验，系统讨论适用于电子商务行业的 BI 整体解决方案。

**3.1** **系统架构**

传统行业的BI 解决方案中核心的数据仓库架构设计如图3.1所示：



OLAP 工具

灵活报表

|CRM

**仪表盘**

进销

在

多维分析

销售

立方体

库存

ETL

人事

物流

数据集市

绩效

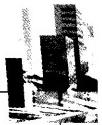
ODS

DW

**图3** **.** **1** **传统行业数据仓库架构示意图**

这个数据仓库架构移植到电子商务行业就会产生严重的“水土不服”现象，通 过总结，发现是下面几种情况限制了整个架构发挥正常的作用。

(1)电子商务行业数据源多种多样，如各种关系型数据库、社交产品产生的数 据、网络日志数据等。单一的 ETL 工具很难很好的处理所有数据源。



13

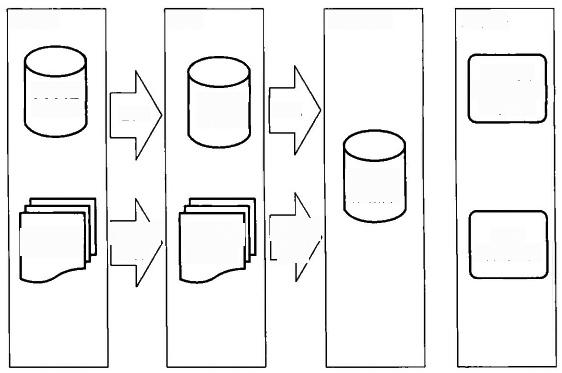
**第** **三** **章** **商** **业** **智** **能** **在** **电** **子** **商** **务** **的** **解** **决** **方** **案**

(2)电子商务数据量巨大。“大数据”这个词的流行就是因为目前各类网站产 生了巨大的网络日志，并且各个公司在处理这么大规模的数据时都遇到了很大的 问题。传统的关系型数据库目前来看已经无法满足处理如此大规模数据的要求。

(3)电子商务对于数据分析的实时性要求极高，有很大部分的数据只在短时 间内才能体现数据的价值，处理数据速度的快慢就显得尤为重要，传统的离线 ETL 处理方式已经无法满足这个需要。

(4)用户量巨大，电子商务的BI 应用很大一部分是面向网站的用户群体的， 这个群体数目巨大。传统的BI 展现工具设计的时候是为了服务企业的中高层用 户，移植到电子商务行业，无论从用户体验还是交互效率都存在巨大差距。

以上四个问题就是传统的BI 解决方案在电子商务行业屡屡失败的根本原因， 根据焦点科技两年以来的BI 实施经验来看，这些问题都是可以解决的。数据仓库 架构的设计本质上都是和传统行业的架构是一样的，只是其中的一些解决方法和 技术有所不同，具体架构设计如图3.2所示：



数据仓库

BI 工具

(QLikVlew.

cognos)

关系型

数据库

分布式)

计算

**网站**

(分布式)

分布式文 件系统

关系型 数据库!

*分布式)*

计算

关系型 数据库

网络日志

**数据源**

展现层

ODS

ETL

ETL

**图3** **.2** **电子商务行业数据仓库架构示意图**

电子商务企业实施 BI不能简单的照搬传统行业的 BI 实施经验，要在理解数 据仓库架构本质的基础上根据行业的特点进行不同技术的采用和架构的升级。本 章将详细论述在电子商务行业实施 BI 的成功经验。

**3.2** **数据仓库及模型**

与传统商业模式不同，对于电子商务行业来说，WEB 用户与网站信息提供者

**商** **业** **智** **能** **在** **电** **子** **商** **务** **中** **的** **实** **践** **与** **应** **用**

不能进行直接的信息交流和意见反馈。而经营者非常希望了解到用户喜欢什么样 的信息，不同用户对信息的不同需求等这些作为网站发展决策的重要依据。因此 建立数据仓库成为企业管理决策支持系统的基础。数据仓库通过有效地组织和存 储数据，把其内部隐藏的信息转化为商业价值，可以帮助企业从海量数据中找到真 正有用的决策信息，为提升企业效益提供服务，解决决策者们迫切关心的问题。

**3.2.1** **数据仓库的定义**

数据仓库并不是为了存储数据，而是为更好地利用企业内所有可能收集到的 数据进行决策支持。数据仓库的概念是 W.H.Inmon 在其《数据仓库》一书中提出 的，他指出“数据仓库是面向主题的、集成的、具有时间特征的、稳定的数据集合，用 以支持经营管理中的决策制定过程”[5]。简单地讲，数据仓库就是企业内部一个专 门的、大型统一的数据存储系统，支持快速、灵活、有效的分析型数据查询。

数据仓库的基本功能包括数据抽取、数据筛选和清洗、数据加载、建立数据集

市、完成数据仓库的查询、决策分析和数据挖掘等操作。

**3.2.2** **数据仓库的架构**

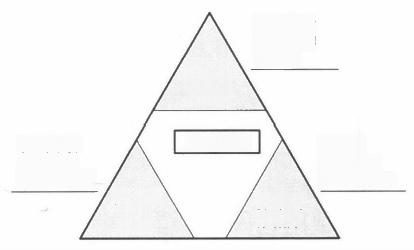
**1)数据仓库的设计思路**

对于传统 BI 来说，数据仓库主要使用关系型数据库来支持离线分析和复杂查 询。而随着电子商务行业的发展，网站需要立即对用户的网站行为进行在线分析， 比如显示某个到访顾客的所有历史来访记录，同时实时跟踪显示某个访客在一个 店铺正在访问的页面等信息，而关系型数据库的查询性能难以满足这样的需求。 内存数据库(如 HBase 等)的数据处理速度比传统数据库的数据处理速度要快很 多， 一般都在10倍以上。因此在线分析，实时 ETL 和多用户在线等对于数据查询 和处理效率要求较高的需求可以采用内存数据库来完成。

同时，电子商务行业会面临越 来越多的比如网站日志、用户行为 这样的大数据的处理，还要进行复

**杂的数据挖掘，还有半结构化数据** **的处理，因此在数据仓库系统中引** **入分布式数据库成为必然的趋势。** **比如，以** **Hadoop[6]为代表的海量数** **据处理开源工具无疑是吸引人的。**

**综上所述，对于数据仓库的设** **计思路如图3.3所示，我们可以结**



离线分析 复杂查询 传统BI

关系型

数据库

在线分析

实时ETL

多用户在线

内存数

据库

Hbase

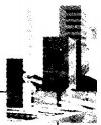
大数据

数据挖掘 半结构数据

分布式 数据库

数据仓库

**图3** **.** **3** **数据仓库的设计思路**

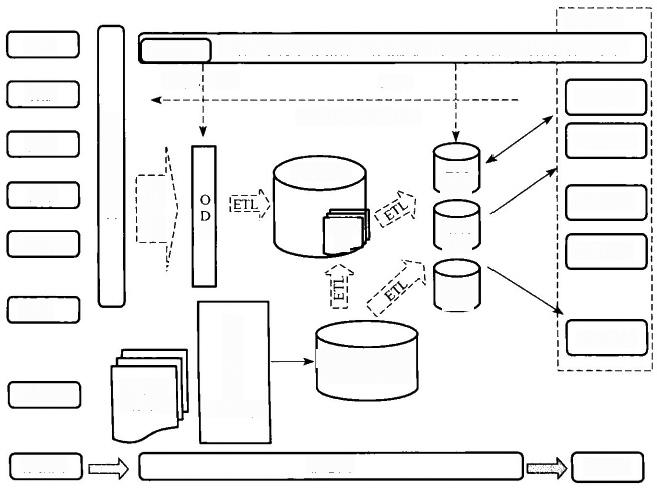
**第** **三** **章** **商** **业** **智** **能** **在** **电** **子** **商** **务** **的** **解** **决** **方** **案**

15

合使用关系型数据库、内存数据库和分布式数据库以满足电子商务行业对于商业 智能在数据处理、数据存储等方面的要求。

**2)企业级数据仓库架构**

数据仓库的目的是构建面向分析的集成化数据环境，为企业提供决策支持(De- cision Support)。数据仓库的数据来源于外部，并且开放给外部应用。数据仓库按照 数据流向可以分为三层结构，包括数据层、信息层和分析层(如图3.4所示)。



BI Server

外部数据 **BI管理系统(元数据管理、数据质量监控、权限管理、预警值管理等等)**

录入、 审批

数据挖掘

**信息引用、数据反写**

企业分析

数据仓库

客户分析

数据支持

其他

在线分析

分布式数据仓库

(Hive+HBase)

网站日志

(Hadoop)

数据层

DM

集市

销售 集市

客服 集市

登录

竞争对手

销售

客服

搜索

分布式 文件系统[

标 准 数 据 接 口

分布式存储

元数 据

信息层

分析层

數 据 抽 取

J2EE

S

**图3** **.** **4** **企** **业** **级** **数** **据** **仓** **库** **架** **构**

(1)数据层

数据层主要是数据源和标准数据接口。

数据源：对于电子商务行业来说，数据源包括网站日志、用户网站行为(如登 录、搜索等)、竞争对手数据以及其他记录网站运营(如客服、销售等)的数据等。

标准数据接口：标准数据接口主要是为了标准化数据，将源数据按照BI 的数 据类型和存储结构进行标准化，从而利于BI 应用。BI 用于决策分析的数据大多都 存在于企业的其他业务系统中，为此其他系统都需要提供数据接口给 BI 系统，支 持其源数据。通常出于对源业务系统的安全和对BI 的稳定的考虑，我们对数据接 口有如下基本要求：

· 通过数据接口，BI 抽取数据所施加给业务系统的压力要减到最小，要充分 评估数据接口对业务系统的安全性。



商 业 智 能 在 电 子 商 务 中 的 实 践 与 应 用

16

· 数据接口可以是数据库的形式，也可以是文件的形式，常用的文件形式有 XML 、TXT 等格式，不管哪种格式，都要求数据接口有实际的数据落地，并且和业 务系统不同一个库，甚至不同一台服务器。

· 数据接口要求稳定，不随业务系统变化而随意变化模型，数据接口的标准和 一切数据模型由BI系统实施人员决定；数据接口数据最好采用推送的方式到达，完 全受控于业务系统代码，而不是 BI系统人员开发代码到业务系统任意进行抓取。

· 数据接口有一定的更新周期要求，通常是一天更新一次。

· 数据接口可以采用和源业务系统完全一样的模型，也可以是中间表， 一般 建议企业级 BI 系统实施的时候数据接口保持为业务系统表结构。

(2)信息层

信息层包括 ODS、数据仓库、数据集市、ETL、元数据。

ODS: 操作数据存储(Operation Data Storage)主要用途是将多个数据源的数据 集成到一个临时缓冲区中供数据仓库使用。ODS 是数据源和目标数据仓库之间的 缓冲库，能有效减轻数据源的压力和 ETL 的压力，其表结构和数据接口表结构尽量 保持一致。比如：我们从源数据向数据仓库中加载事实表数据时，这时候我们需要进 行聚合操作，如果没有 ODS层，那么所有聚合操作的压力是在源系统形成的，这就会 给客户源系统带来很大的压力，这是在项目实施过程中经常遇到的一个问题。

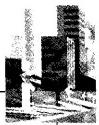
通常 ODS 可以包括下面三层。

· 映射层：把所有要用到的数据源表的所有字段映射到本地数据库，实现数 据的落地，对于外部数据也可以通过系统管理到这一层，实现所有数据的集中。

· 数据预处理层：在这一层可以进行数据整合、数据筛选和增加关联表等预 处理，所有的预处理都是为了简化和提升 ETL 过程。

· 数据清洗：对于有质量问题的数据进行数据清洗，清洗规则以业务部门定 义为准。

数据仓库：如图3.3所示，我们建议数据仓库采用关系型数据库、内存数据库 和分布式数据库的综合结构。关系型数据库用以支持传统 BI 数据需求。云技术 的出现打开了大数据处理的新篇章，采用分布式数据库来支持BI 大数据的存储与 计算，可以显著的提升数据处理和数据搜索的性能。在 Hadoop 分布式系统架构 中，提供了一个数据仓库工具 Hive,Hive 本身建立在 Hadoop 的体系结构上，能 很好地处理不变的大规模数据(例如网络日志)上的批量任务；但是，Hive 不提供 在线事务处理和实时查询，这一切可以由Hadoop 的另一个成员 HBase 来完成， HBase 为查询而生的，它通过组织起节点内所有机器的内存，提供一个超大的内存 Hash 表，它需要组织自己的数据结构，包括磁盘和内存中的，表在 HBase 中是物 理表，而不是逻辑表，搜索引擎使用它来存储索引，以满足查询的实时性需求。



**第三章** **商业智能在电子商务的解决方案**

数据仓库是根据实际的业务主题逐步进行设计的， 一般开始阶段都会选择核 心业务进行设计。

数据集市(Data Mart,简称 DM): 数据集市是面向特定的主题进行设计，根据主 题业务的实际需要可以对仓库数据进行筛选、过滤、汇总、表合并等操作。数据集市 的设计需要选择合适的集市粒度，但没必要保留最明细的数据，根据实际分析需要我 们选择最为合适的粒度，只有这样才能最大限度提高性能和整体设计的合理性。

ETL:ETL 占据整个 BI实施最少50%的工作量。现在有大量的工具可以用 来开发ETL 过程，也可以直接基于存储过程进行开发，各有优缺点。前者易于维 护和管理，但是要对工具非常熟悉才能确保有比较好的性能；后者必须对所有开发 人员进行培训，培训开发中的设计标准，这样才能避免因为过多存储过程代码而最 终无法管理的局面，如果管理恰当可以带来较高的性能。

元数据：元数据(Meta Date),其实应该叫做解释性数据，即数据的数据。主要 记录数据仓库中模型的定义、各层级间的映射关系、监控数据仓库的数据状态及 ETL 的任务运行状态。 一般会通过元数据资料库(Metadata Repository)来统一地 存储和管理元数据，其主要目的是使数据仓库的设计、部署、操作和管理能达成协 同和一致。

(3)分析层

17

分析层主要是数据的应用，如数据挖掘，报表分析，数据支持等。

**3.2.3** **数据仓库模型**

数据模型是对现实事物的反映和抽象，它可以帮助我们更加清晰地了解客观 世界。在分析完业务需求之后，我们就需要建立数据仓库模型以反映业务逻辑。 正确和完备的数据模型，是BI 进行一切数据分析的基础，是决定数据仓库项目成 功与否的重要因素。

采用第三范式(3NF) 建模对事务处理来说很有好处，但是对于数据仓库来说， 要求模型能够提供较好的查询性能，因此数据仓库建模通常使用维度建模[。接 下来，将介绍数据仓库维度建模的方法。

**1)数据仓库建模的原则**

电子商务行业的信息系统具有业务复杂、数据量大、系统庞大等特点。电子商 务行业的数据仓库模型设计应该遵循以下原则：

· 满足不同用户的需求

电子商务行业的BI用户包括“买家”、“卖家”以及平台供应商自己，平台供应 商内部还可以分为不同的业务部门(比如客服、销售等),或者分为高层领导、中层 领导和普通员工。不同层次、不同级别、不同岗位、不同立场的用户关注的信息不



**商** **业** **智** **能** **在** **电** **子** **商** **务** **中** **的** **实** **践** **与** **应** **用**

同。数据仓库必须能够支持不同用户的信息需求。

· 兼顾查询效率与数据粒度的需求

过细的数据粒度会占用大量的存储空间，并且会降低查询速度。但是高度汇 总的数据却不能保证数据分析的灵活性。数据仓库模型的设计需要同时兼顾查询 效率与数据粒度。电子商务行业每天将会产生大量的日志数据，比如用户访问日 志，如果都将这些数据保留在数据仓库中，将会消耗大量的存储空间。但是，对于 数据分析来说，当前一个时间段内(比如最近一年内)的明细日志数据是有分析价 值的，但是在这之前的日志数据可能我们只关心用户在每一天的点击量，那么可以 将这些历史数据按天汇总后再存储，这样将会降低存储并且方便查询。

· 支持用户需求变化

用户的需求会随着业务或者市场的变化而变化。电子商务行业正处于高速发 展阶段，同时也带来了激烈的竞争，用户的需求变化也会越来越频繁。数据仓库模 型必须能够支持用户不断变化的需求。

· 避免业务运营系统性能影响

电子商务平台的用户可能来自不同的国家，在任何时候都会有用户访问平台， 业务运营系统可能在24小时内都处于忙碌状态。数据仓库每天都需要抽取业务 系统的数据，在设计数据仓库模型时应该考虑尽量减少和缩短对于业务系统的访 问时间和资源占用，从而降低对业务系统性能的影响。

18

· 提供可扩展性

数据模型的可扩展性决定了数据仓库对新的需求的适应能力，建模既要考虑 眼前的信息需求，也要考虑未来的需求。比如对于电子商务行业的竞争对手数据 的分析，现在只有A 企业是我们的竞争对手， 一个月后新加入的B 企业可能也成 了我们的竞争对手，我们可以把 A 、B的数据都放在同一个表中，但这样显然不方 便进行数据分析(比如数据对比)。如果再增加竞争对手C、D、E 、…,这个表将会 变得越来越庞大，更加不利于进行数据分析。在这种情况下，针对不同的竞争对手 建立不同的模型，会更加方便数据存储和数据分析。

**2)星形模型和雪花形模型**

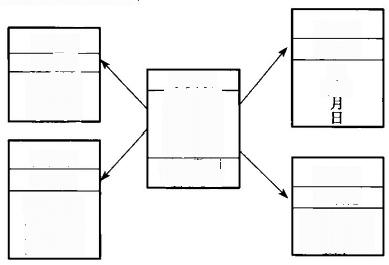
目前常用的数据仓库数据模型为多维数据模型，这种模型主要以星形模式和 雪花形模式在关系数据库系统中存在。

(1)星形模型

星形模型是一种多维的数据关系，它由一个事实表(Fact Table)和一组维表 (Dimension Table)组成，如图3.5所示。每个维表都有一个维作为主键，所有这些 维则组合成事实表的主键，换言之，事实表主键的每个元素都是维表的外键。事实 表的非主属性称为事实(Fact),它们一般都是数值或其他可以进行计算的数据；而



**第** **三** **章** **商** **业** **智** **能** **在** **电** **子** **商** **务** **的** **解** **决** **方** **案**



**维大都是文字、时间等类型的数据。**

时间维

部门维

部门键

事实表

产品维

产品键

产品名称

产品描述

产品售价

产品质量

**图3** **.5** **星型模型**

照

销 售 数 额

销得金额

地域维 地域键

鬣



时间键

年

季度

使用星形模式主要有两方面的原因：

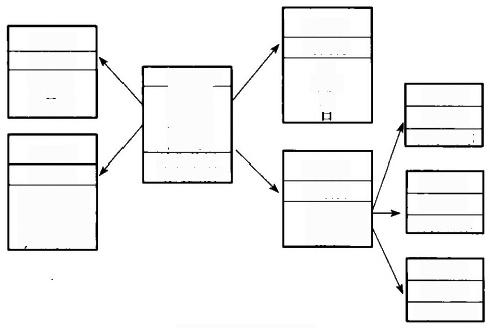
· 提高查询的效率。采用星形模式设计的数据仓库的优点是由于数据的组 织已经过预处理，主要数据都在庞大的事实表中，所以只要扫描事实表就可以进行 查询，而不必把多个庞大的表连接起来，查询访问效率较高。同时由于维表一般都 很小，甚至可以放在高速缓存中，所以与事实表作连接时其速度较快。

· 便于用户理解。对于非计算机专业的用户而言，星形模式比较直观，通过 分析星形模式，很容易组合出各种查询。

19

(2)雪花形模型

在实际应用中，随着事实表和维表的增加和变化，星形模型会产生多种衍生模 型，包括星系模型、星座模型、二级维表和雪花模型。雪花模型是对星形模型维表 的进一步层次化，将某些维表扩展成事实表，这样既可以应付不同级别用户的查 询，又可以将源数据通过层次间的联系向上综合，最大限度地减少数据存储量，因 而提高了查询功能。如图3.6所示，将地域维表分解为国家、省份、城市等维表。



时间维

时间键

年

季度

昌

露

雪花模型

事实表 时间键 地域键

部门键

产品键 销售数额

销售金额

图3 . 6

国家维 国家键 国家名称

省份维

省份键

省份名称

城市维 城市键 城市名称

部门维

部门键

总公司

分 ，A: 司

分 公 可

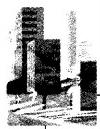
代理处

产品维

产品键

名 称 普济 质量

品品品品 产产产产



20

**商业智能在电子商务中的实践与应用**

雪花模型的维度表是基于范式理论的，因此是介于第三范式和星形模式之间 的一种设计模式，通常是部分数据组织采用第三范式的规范结构，部分数据组织采 用星形模式的事实表和维表结构。在某些情况下，雪花模式的形成是由于星形模 式在组织数据时，为减少维表层次和处理多对多关系而对数据表进行规范化处理 后形成的。

雪花模型的优点是：在一定程度上减少了存储空间；规范化的结构更容易更新 和维护。同样雪花模式也存在不少缺点：雪花模式比较复杂，用户不容易理解；浏 览内容相对困难；额外的连接将使查询性能下降。

在数据仓库中，通常不推荐“雪花化”。因为在数据仓库中，查询性能相对 OLAP 系统来说更加被重视，而雪花模式会降低数据仓库系统的性能。

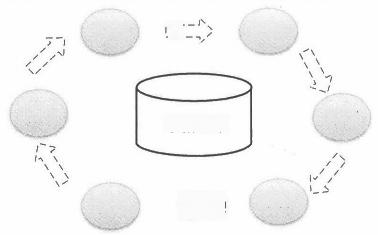
**3)数据仓库的模型设计**

数据仓库的设计以支持企业进行有效的BI 实施为目的，企业 BI 对内、外部的 用户提供统一的分析平台和有价值的信息与知识。以中国制造网为例，BI 系统提 供从市场分析、产品开发与推广、销售、客服和客户等全程分析与决策支持。我们 建议以星形结构模型建立数据仓库。设计一个数据仓库主要要考虑如下方面：

(1)主题设计

数据仓库是根据实际的业务主题逐步进行设计的，数据仓库的开发也有主题 性， 一般开始阶段都会选择核心业务进行设计。根据上文提到的业务点，进行数据

仓库的主题划分，如图3.7所示：



产品

数据仓库

客服

市场

客户

销售

推广

==3

*C=.*

**图3.7** **数据仓库主题划分**

· 客户主题

客户主题下主要存储客户的详细信息，记录客户的基本属性以及注册信息等 数据。客户信息数据体现最新的客户状态，当客户资料发生变化时，旧的客户信息 将被新的客户信息更新掉。客户信息数据中保留每一个生命周期的客户的最新 数据。

· 市场主题

市场主题下主要存储竞争对手数据，如竞争对手客户信息、竞争对手的产品目



**第三章** **商业智能在电子商务的解决方案**

录设置情况等。

· 产品主题

产品主题下存储客户的产品信息，包括产品的基本属性，产品信息数据体现最 新的产品状态，当产品信息发生变化时，旧的产品信息将被新的产品信息更新掉。 产品信息永久保存。

· 推广主题

推广主题下存储业务推广产生的数据信息，比如展会信息等。

· 销售主题

销售主题下记录产品(这里的产品指网站提供的服务)的销售情况，销售人员 的业绩等信息，主要存储了销售人员详细信息、销售合同信息、销售合同产品信息、 销售订单信息、服务信息表、销售回款信息等数据。

· 客服主题

客服主题下主要存储了客服客户服务信息、客服工作记录等数据。

(2)事实设计

事实表是历史数据的事实，不具有在业务上重复出现的性质，比如销售数据， 事实表更多的是体现主题特征。事实表是星形结构的核心，记录主题的主干内容。 比如销售订单数据，主要内容有日期、订单ID、客户、订单状态、订单金额等。订单 ID为主键，订单金额为事实表的主干信息，其他字段为维度。

21

(3)维度设计

维度表是存放分析角度数据的表，维度数据具有高重现性，比如产品数据维度 表经常为不同主题所共用；根据以往经验每个企业维度都不会太多，每个主题虽然 都可能有自己特征的维度，但是公共维度占大部分。常用的公共维度有时间维度 (年、月、日、周等)、产品数据维度(产品ID、公 司ID、产品名称、产品状态等)、客户 信息维度(公司ID、公司名称、公司所在地区、公司联系人、公司主营行业)等。

(4)代理关键字

维表关键字只能用代理关键字。使用代理关键字可以避免不同数据源数据的 统一性问题，因为不同数据源的数据可能缺乏统一的关键字编码；代理关键字可以 处理缓慢变化维，因为代理关键字可以描述历史变化信息，区分同一维度值ID 在 属性变化前后的区别。例如员工维表，某员工变换部门，从销售一部换到销售二 部，那么在表中就会记录下该人员的两条信息，该员工属于销售一部的记录作为历 史记录，该员工属于销售二部的记录作为当前记录。

***3.3*** ***ETL***

数据由各种系统数据库、文件、网页等存储方式进入数据仓库需要经过抽取



**商** **业** **智** **能** **在** **电** **子** **商** **务** **中** **的** **实** **践** **与** **应** **用**

(Extraction) 、转换(Transform) 和装载(Load) 这样一个过程，简称 ETL, 这个过程 非常具有挑战性，但是只要能认识ETL 的本质和核心问题，还是能把 ETL 简单 完成。

焦点科技的 ETL 完全基于数据库存储过程进行开发，在这个过程中非常重视 完整的 ETL 设计思想、方案、架构及技术标准，下面我们从各个角度来介绍焦点科 技BI 人员对ETL 的设计及开发的认识，从中大家或许可以领略一点技术的魅力。

**3.3.1** **ETL** **目标和基本要求**

简单点理解 ETL 的目标就是保证数据仓库、数据集市数据的及时性和准确 性，从而保证我们各种分析需求的数据要求。基本要求包括：

(1)数据完全准确。数据完全准确是 ETL 的第一要求，数据不准的系统没人 敢用，不准的数据甚至会带来巨大危害，毕竟高层领导的分析和战略决策都需要 BI 数据来支撑。

(2)更新及时。及时更新数据会给 BI 带来很好的用户体验，但不太现实，因 为 ETL 运算量过于庞大，所有 ETL 都实施或者频繁更新难于实现，但是对于实 时性要求比较高的分析，提高更新频率是有必要的；对不同需求有不同的更新策 略就非常有必要了。

22

(3)系统稳定。区别一般系统稳定性这里有两层含义，除了服务不能停之外 还要求 ETL 在更新数据的时候不会经常中断，如果中断也能快速恢复。

(4)对其他系统影响小。虽然每天要抽取的数据很多，但是任何系统都不能 容忍对自身影响太大的请求，快速从业务系统获取数据是 ETL 第一步要解决的 问题。

当然如果能做到如下两条，那更为完美，焦点科技这方面已经达到。

(5)代码简单并且能够标准化。

(6)维护简单。

**3.3.2** **ETL 的开发**

**1)企业级** **ETL开发基本思路**

· ETL 包的分解：企业级 ETL 开发首先要做的事情是 BI 项目和企业业务 主题规划，然后按不同主题来划分ETL 包，最后得到分主题的设计ETL 包，包之 间数据可以单向依赖，不能互相依赖，这样我们在后面就可以按照依赖关系进行 调度。

· 分包分项目开发 ETL: 根据项目进度我们安排包的开发，先上项目先开 发，后上后开发，如果项目为同一主题，我们可以共用一个包，这样在数据统一性方

**第** **三** **章** **商** **业** **智** **能** **在** **电** **子** **商** **务** **的** **解** **决** **方** **案**

面会更好。

· ETL 更新顺序：根据表之间的数据传递关系，建立起表之间的依赖关系， 根据表之间的依赖关系建立起包之间的依赖关系，如表 A 有数据来源于表 B,那么 我们认为表 A依赖表 B, 对包同理。原则上表和表、包和包之间不互相依赖。

· ETL 中断纠正机制：ETI, 中断是商业智能系统常见的问题，建立起简单有 效的 ETL 中断处理机制是必不可少的工作之一。

· ETL 统一调度：确定更新时间、更新频率、更新对象。

**2)ETL 的挑战及应对思路**

· ETL 技术架构：技术架构只有合不合适，没有说哪个最好，前面我们在系 统架构方面基本了解到现在我们针对电子商务行业采用了 ODS 、DW 、DM 这样 三层架构，不管什么工具都要符合这样三层架构思想进行设计，这样我们在架构层 面就有明确的目标定位，相当于统一了思想，从而为后面的整体设计，分步、分工实 施提供了可能。系统架构也是我们后面制定技术标准的大框架。

· ETL 更新机制：ETL 开发除了数据质量之外，最具挑战的就是数据更新机 制的设计，ODS 、DW 、DM的每个表都面临数据怎么更新的问题。只有最合适，没 有说哪个方法最优，不管哪个层级的表数据更新策略可以简单分为：

I. 全量更新，每次把表数据清空，重新生成，对 ETL 来讲小表我们可以每次 全量更新，而对于每天增量数据较多的表则不能全量更新，只能利用增量 更新办法。

23

Ⅱ.增量更新，其中增量更新分完全增量更新和相对增量更新，完全增量更新 是指加载到目标表的数据和目标表原数据既不出现重合也不出现缝隙， 这是最理想的一种更新方式，但是因为增量判断机制复杂，增量方式众多 (新增、更新、删除),每种增量方式的增量算法完全不一样，开发难度巨 大，特别是数据源如果不稳定，那将出现致命性错误，甚至导致数据仓库 必须重新初始化这样大的事件；相比较相对增量更新可以允许数据有一 定重合度，通过关键字等判断自动删除重合部分再简单加载新数据，不用 面临一个指标更新要几个算法应付各种更新机制的问题，这里只需要一 种更新算法就可以，设计简单，性能也不会有太大牺牲，如果大数据引入 分区技术，每次通过移除分区来删除重复数据那性能几乎没有损失。

· 异常中断纠正机制：很多企业在开发 BI 之初，面临 ETL 中断之后是重新 跑数据还是哪里断了哪里重启这样一个问题。 一般来讲重新跑风险小，但是时间 上要允许，而断点处开始执行则面临人工干预的问题，即使再有经验的人操作这个 风险也始终存在。事实上 ETL 开发到一定程度可以用点时间来解决该问题，现在 焦点科技通过合理设计在效果上实现了融合两者优点的方法：



**商** **业** **智** **能** **在** **电** **子** **商** **务** **中** **的** **实** **践** **与** **应** **用**

24

I. 建立表之间依赖关系。

Ⅱ.任何报错只影响对该 ETL 对象表有依赖关系的表，受影响的表不执行， 节省时间，因为跑了也是白跑，不受影响的表继续往下执行。

Ⅲ.跑完之后系统自我检查是否有报错，有报错再次重跑，重跑过程中系统自 动调度已发生错误的 ETL 进行重跑，其他成功的表不再重跑，依赖于该 ETL 的表其 ETL 都需要重跑。

IN. 如果再次重跑还是失败那么通过邮件预警到相关人员，检查错误排除错 误直到解决后再次重跑。

· 大数据的 ETL 技术：解决大数据，是电子商务行业必须面临的问题，现在 主流的办法还是分布式计算，其中 Hadoop 最为常用，当然 EMC 的 Greenplum 更 为成熟，其费用也不菲，有条件的建议大家应用后者。

· 数据质量：我们可以认为BI 产生的数据和业务部门需要的数据不一致问 题都是数据质量问题，所以在开发 ETL 过程中处处是风险：

I. 首先是需求问题，BI需求最大难点是难于统一各个独立业务部门的指标 定义，除非自上而下强行统一，但一般来讲 BI 实施之初不可能有这种力 度，比较可行的办法还是在数据仓库放基础数据，面向不同业务部门设计 独立的集市来满足其需求，当BI 在各个业务部门应用到一定程度，他们

会自动统一这些定义和需求，否则他们将无法沟通，当然 BI 团队可以加 快这个进程。

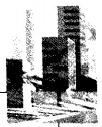
Ⅱ.其次就是架构问题，在数据仓库层面要放基础数据，基础数据不要有冗 余， 一个字段在多个表出现将会导致数据难于统一，这是最大忌讳，为了 提升仓库到集市的数据统一性，我们可以在仓库层生成结构性的中间表， 让大部分基础运算都统一到这层，这样为仓库进集市的数据统一性又降 低了难度。集市设计最好是分主题，在同一主题下，数据定义要求完全一 致，表尽量共用设计，这样不管是哪个项目在同一体系内数据都是一致 的，另外就是不同定义的数据哪怕在不同主题间也要区别命名。

Ⅲ.最直接的数据质量问题还是 ETL 开发和测试，ETL 测试不仅要看数据总 量，必要时候还得交叉维度和层级进行数据测试，简单测试指标包括汇 总、计数等。

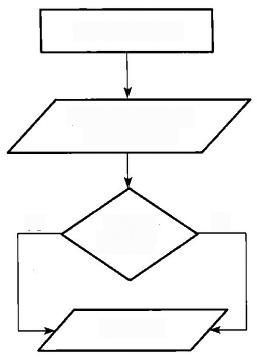
IN. 系统维护要能自动预警数据质量问题，这样可以让相关人员在第一时间 发现问题、解决问题。

**3)电子商务行业可采用的** **ETL** **开发**

一般的设计数据仓库都可分为 ODS 、DW 、DM三层，下面我们来看一下这三 层我们是怎么做的。



**ODS** **全量加载：**



清空ODS 目标表

将数据源的数据全 量插入目标表

插入是

否成功?

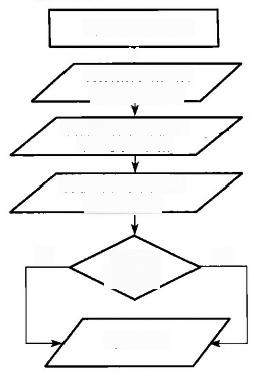
写入日志

否

是

第 三 章 商 业 智 能 在 电 子 商 务 的 解 决 方 案

(1)ODS 层全量加载、增量加载 ETL 开发流程(图3.8)



ODS 增量加载：

清空ODS临时表

王

将数据源增量数 据插入临时表

删除目标表与临时 表重复的数据

将临时表数据全量

插入目标表

是 插入是 否成功?

写入日志

否

**图3.8** **ODS\_ETL 开发流程图**

· 全量加载将清空原有数据，重新载入最新的源数据。

· 增量加载每次只更新历史数据中有更新的记录，并且载入历史数据中没有 的新数据。

25

(2)DW 层 ETL 开发

· 事实表先更新，再更新维表。

· DW 层事实表全量加载 ETL 开发：与 ODS 层全量加载 ETL 开发方法类 似，在此不再重回复。

· DW层事实表增量加载 ETL 开发：与 ODS 层增量加载ETL 开发方法基 本一致，但是需要把相关维表字段关联修改为维表的代理关键字。

· DW层维表增量加载 ETL 开发如图3.9所示。

(3)DM层

· DM层全量加载 ETL 开发：与ODS 层全量加载 ETL 开发方法类似，在此 不再重复。

· DM 层 增 量 加 载 ETL 开 发 如 图 3 . 1 0 所 示 。

**3.3.3** **ETL** **的测试**

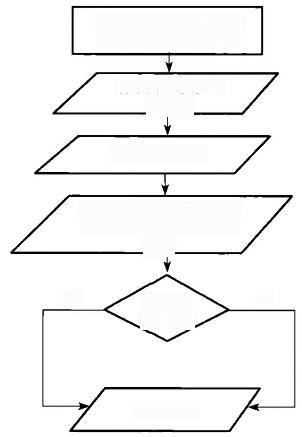
ETL 数据测试采用黑盒测试[8],包括 ETL 常规检查和业务逻辑测试。

ETL 常规检查包括①ETL 脚本是否有运行错误；②ETL 编码是否符合 ETL 技术规范。

ETL 业务逻辑测试是 ETL 测试的重点，主要包括：①业务逻辑测试：指标设

**商** **业** **智** **能** **在** **电** **子** **商** **务** **中** **的** **实** **践** **与** **应** **用**

DW 维表增量加载：



将增量数据按照历史数 据的ROWNUM 排序

将增量数据插入

临时表

清空DW 维表

将ODS 源数据表与临时 表关联，将数据插入维

表中

插入是

否成功?

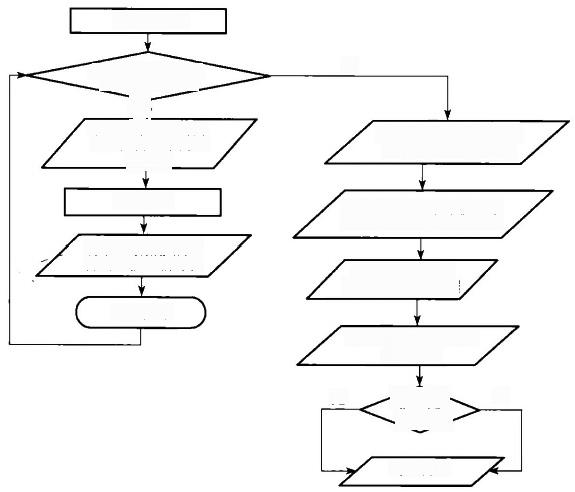
写入日志

是

否

**图3** **.9** **DW\_ 维表增量加载** **ETL 开发流程图**

26



否

取出数据表中时间截范 围内的不同业务时间

将临时表1的数据按照维 度字段汇总到临时表2中

删除目标表与临

时表2重复的数据

将临时表2.的数据插

入目标表

插 入 是

香成功?

写入日志

清空临时表

是否还有指标取百 不同的DW 表 ?

是

取出数据表中时间

戳范围内的不同业

务时间

循环处理

将增量数据按照业 务时间插入临时表

结束循环

是

否

**图3** **.** **10** **DM 层增量加载** **ETL 开发流程图**



**第三章** **商业智能在电子商务的解决方案**

计是否符合业务逻辑；②数据量测试：目标表数据量是否与数据源数据量一致； ③唯一性检查：检查 ETL 重复调度是否会出现重复数据；④性能测试：测试 ETL 的效率是否在可接受的范围内。

ETL 的测试流程包括：理解业务逻辑，开发测试用例，编写测试文档，通知开 发人员修改 BUG,BUG 验证等。

**3.4** **多维分析**

OLAP 的目标是满足决策支持或者满足在多维环境下特定的查询和报表需 求，它的技术核心是“维”这个概念。多维数据分析工具的集合构成了OLAP 。存 储在各种数据源中的数据通过ETL 存储到数据仓库中，对数据仓库中的数据进行 多维建模，便于利用数据挖掘技术和工具发现隐含的信息，还可以利用前端展现工 具将数据分析结果展现出来，以供用户查阅。

**3.4.1** **多维数据分析技术**

如何理解多维数据中的维呢?维是我们分析问题和观察事物的角度，从不同 的角度对事物进行分析观察可以得到不同的结果，也能够让我们能够更加全面和 真实的了解事物的本质。比如公司销售业务数据中，记录了公司产品(会员)详细 情况，则我们可从以下几个方面来对销售数据进行分析：

27

(1)从产品的角度，可以按产品的类别、品牌来查看产品的销售情况。

(2)从客户的角度，可以按客户的类别、地区等来查看产品的购买情况。

(3)从销售代表的角度，可以按销售代表的部门、级别等来查看产品销售 业绩。

(4)从时间的角度，可以按年度、季度、月份等来观察产品销售的变动情况。

其中产品、客户、销售代表、时间分别是四个不同的维度，每个维度都从不同方面体 现了销售数据的特征，而每个维度又可按粒度的不同划分成多个层次，称为维度成员。

多维分析中另一个重要的概念是数据指标，简称指标，指标代表了数据中的可 度量的属性，在上面的销售数据中有两个重要的指标是销售数量和销售金额。

了解了多维数据的维的概念之后，便可对数据进行多维分析操作，常见的多维 分析操作主要有：钻取(上钻和下钻)、切片、切块、旋转。焦点科技目前使用的 QlikView 软件即可实现上述多维分析操作。

钻取是改变维的层次，变换分析的粒度。钻取包括上钻和下钻，上钻是在某一 维上将低层次的细节数据概括到高层次的汇总数据的过程，减少了分析的维数；下 钻则是相反，它从汇总数据深入到细节数据进行观察和增加新维。比如：将各销售

**商** **业** **智** **能** **在** **电** **子** **商** **务** **中** **的** **实** **践** **与** **应** **用**



28

人员的销售业绩汇总为各个销售部门的销售业绩，整个销售部的销售情况可以细 分为各销售部的销售情况，进一步细分到各个销售人员的销售业绩。在多维分析 中，如果在某一维度上限定了一个值，则称为对原有分析的一个切片；如果对多个 维度进行限定，每个维度限定为一组取值范围，则称为对原有分析的一个切块。如 果变换维度的顺序和方向，或交换两个维度的位置，则称为旋转。

**3.4.2** **多维数据分析的实现**

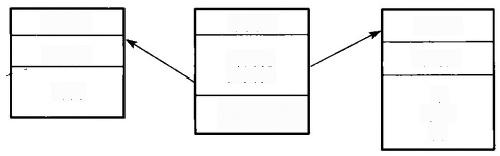
要进行多维数据分析，首先需要提取维度和指标信息并且确立维度与指标之 间的关系，然后根据统计规则统计出指标的结果，最后将统计结果展现出来。这三 个步骤分别对应于数据仓库建模、ETL 开发、前端展现的BI 项目开发过程。下面 以焦点科技销售BI 系统下的销售业绩统计分析为例(为了便于描述，对数据进行 了简化),详细阐述多维数据分析的实现。

**1)提取维度和指标信息**

数据仓库中销售业绩数据主要包括部门人员信息和销售数据。部门人员信息 为：销售人员、所在部门。销售数据为：日期、销售人员、销售额。销售提成为销售 额的3%。进行销售业绩统计分析的维度有：时间维、人员维，数据指标有销售额、 销售提成。

**2)实现主题扩展和关联**

由于各个主题间维度在定义时都是相互独立的，其间并无隐含的关联信息，但 是为了数据一致性和关联性，我们将会建立各个分析主题维度的关系信息。比如 销售的提成主题和部门的业绩主题就是上下层的关系，我们只要在销售提成主题 上加入部门维度，既可以实现提成主题分析也可以向上钻成部门业绩的多维分析 主题，如图3.11所示：



事实表

时间键

人员键

第转器藏

时间维

时间键

年

季度

月

人员维

人员键

部门

**图3** **.** **11** **销售业绩数据模型**

**3)执行数据整合**

建立好数据模型之后，利用 ETL 将数据仓库中的数据按照客户需求进行相应 的统计汇总得到多维分析数据。最终形成QlikView 报表，实现钻取、分区、分块、 旋转等操作。销售业绩统计的结果可以生成如下的 QlikView 报表。

**第** **三** **章** **商** **业** **智** **能** **在** **电** **子** **商** **务** **的** **解** **决** **方** **案**

29

图3.12展现了各部门各销售代表每月的销售业绩情况，可以在图3.12 的基 础上进行上钻得到各部门各销售代表的年度销售情况(如图3.14),以供各部门主 管或相关人员查看和分析。在图3.13基础上再进行上钻，即可得到各销售部的年 度销售情况。相反的，如果在图3.13基础上，进行下钻，可以得到如图3.12所示 的明细数据。在图3.12中想要查看各部门各季度的销售情况，则可以将时间维度 与人员维度进行旋转(交换顺序),结果如图3.15所示。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 销售情况明细表 部门 |  | | | | |
| 销售代表 | 单度 | 月份 | 销售超 | 销售提成 |
| 二  销售一部 | 李明 | =  一季度 | 一月 | 1000 | 30 |
| 二月 | 1200 | 36 |
| 三月 | 1400 | 42 |
| =  二季度 | 四 月 | 1100 | 33 |
| 五月 | 1500 | 45 |
| 天月 | 1700 | 51 |
| 二  三季度 | 七月 | 1100 | 33 |
| 八月 | 1100 | 33 |
| 九月 | 1000 | 30 |
| =  四季度 | 十月 | 1600 | 48 |
| 十一月 | 1300 | 39 |
| 十二月 | 1500 | 45 |
| 王丽 | 日  一季度 | 一月 | 1200 | 36 |
| 二月 | 1700 | 51 |
| 三月 | 1100 | 33 |
| 二季度 | 四 月 | 1100 | 33 |
| 五月 | 1000 | 30 |
| 六月 | 1100 | 33 |
| 三季度 | 七月 | 1000 | 30 |
| 八月 | 1600 | 48 |
| 九月 | 1300 | 39 |
| 四季度 | 十月 | 1500 | 45 |
| 十一月 | 1400 | 42 |
| 十二月 | 1500 | 45 |
| 销售二部 | 刘阳 | 季度 | 一月 | 1000 | 30 |
| 二月 | 1200 | 36 |
| 三月 | 1000 | 30 |
| 二季度 | 西月 | 1100 | 33 |
| 五月 | 1500 | 45 |
| 天月 | 1500 | 45 |
| 三季度 | 七月 | 1100 | 33 |
| 八月 | 1100 | 33 |
| 九月 | 1100 | 33 |
| 四季度 | 千月 | 1500 | 45 |
| 十一月 | 1500 | 45 |
| 十二月 | 1000 | 30 |
| 张三 | 季度 | 一月 | 1400 | 42 |
| 月 | 1500 | 45 |
| 三月 | 1400 | 42 |
| 西  二季度 | 月 | 1600 | 48 |
| 五月 | 1100 | 33 |
| 六月 | 900 | 27 |
| 三季度 | 七月 | 1100 | 33 |
| 八月 | 1100 | 33 |
| 九月 | 1000 | 30 |
| 四季度 | 十月 | 1100 | 33 |
| 月 | 1500 | 45 |
| 十二月 | 1500 | 45 |
| 合计 | | | | 60800 | 1824 |
|  | | | | | |

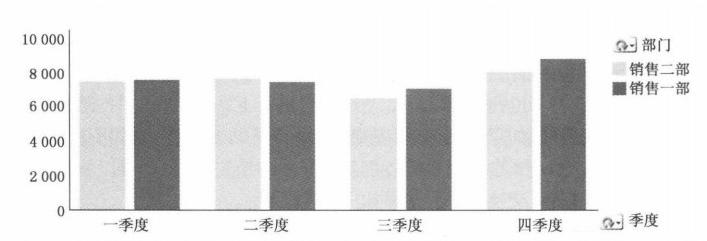
**图3.12** **销售业绩分析报表\_明细表**

QlikView 报表中可以便捷地使用图形分析来展现多维数据，如图3.13所示， 利用柱形图展现出各季度各部门的销售情况，便于对比分析。并且在柱形图上仍 然可以实现钻取、分区、分块等多维分析操作。比如，将季度下钻到月份，如图3.16 所示。



30

**商** **业** **智** **能** **在** **电** **子** **商** **务** **中** **的** **实** **践** **与** **应** **用**



**图3.13** **销售业绩分析报表\_柱形图**

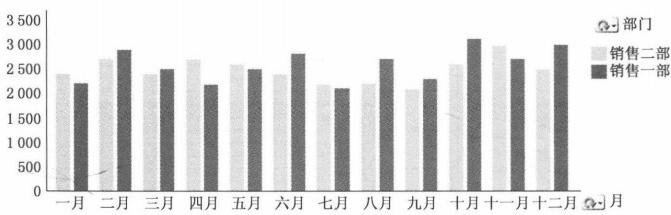
|  |  |  |  |
| --- | --- | --- | --- |
| 销售情况明细表 | | | |
| 部门 -] | 销售代表 季度 月份 | ]销售颇 | 销售提成 |
| 销售一部 | 牵明 | 15500 | 465 |
| 王商 | 15500 | 465 |
| =  销售二部 | 刘阳 | 14600 | 438 |
| 张三 | 15200 | 456 |
| 合计 | | 60800 | 1824 |

**图3.14** **销售业绩分析报表\_上钻**

销售情况明细表

|  |  |  |  |
| --- | --- | --- | --- |
| 部门 ~ | 年度 月份 辆 要 代 表 | 销售额 | 销售提成 |
| 2  销售一部 | 一季度 进 | 7600 | 228 |
| 三垂度 | 7500 | 225 |
| 三季度 生 | 7100 | 213 |
| 西季度 | 8800 | 264 |
| 销售二部 | 季度 | 7500 | 225 |
| 季度 生 | 7700 | 231 |
| 三季度 + | 6500 | 195 |
| 四季度 注 | 8100 | 243 |
| 合计 | | 60800 | 1824 |

**图3.15** **销售业绩分析报表\_旋转**



**图3.16** **销售业绩分析报表\_柱形图下钻**

**3.4.3** **多维分析在商业智能中的应用**

**1)覆盖各层用户需求**

BI 在电子商务行业中的用户有买家、卖家、网站内部中高层管理人员以及一 线操作员。多维数据分析可以从不同的客户角度进行数据分析，满足各个层次、各

31

**第** **三** **章** **商** **业** **智** **能** **在** **电** **子** **商** **务** **的** **解** **决** **方** **案**

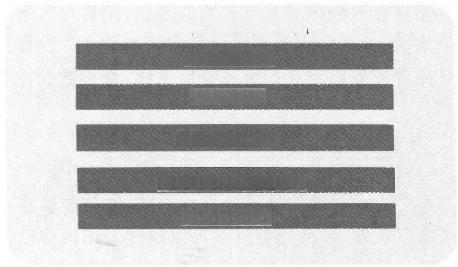
个级别用户的统计分析需求。比如，销售业绩分析中，底层销售代表需要知道一定 的时间周期内的自己的销售额和自己的提成情况；而部门主管们更加关注的是本 部门的月、季度或者年的计划及绩效。

**2)支持数据挖掘**

多维分析的结果可以用于支持数据挖掘，我们可以基于现有分析结果，挖掘出 更有价值的隐含在数据中的信息。比如，基于销售数据的多维分析，进行简单的数 据挖掘就可以发现出每个季节或月份比较畅销的产品类型、不同的用户对于产品 的喜好等，从而可以根据这些信息制订合适的产品销售计划。

**3.5** **数据挖掘**

数据挖掘是从海量数据中挖掘潜在的信息与知识，挖掘结果的应用往往可以 带来产品的创新以及业务的创新。B2B 、B2C 等电子商务平台在每天的运营过程 中，信息系统都会产生大量数据信息，数据挖掘技术能够把历史积累的大量数据进 行分析以及模型化的挖掘和处理，从中发现隐藏的规律或模式，提取辅助商业决策 的关键性数据，为决策提供支持。MIC 挖掘价值体现见图3.17所示。



MIC挖掘价值体现

**产品深度支持**

**精确化销售**

客户精细化服务

**精细化推广与市场培养**

**客户深度认识**

**图3.17** **MIC** **挖掘价值体现**

**3.5.1** **价值体现**

**1)产品深度支持**

新产品设计：对于竞争激烈的互联网企业，好的产品就是企业的生命。随着信 息化的发展，如何能够设计出一个用户满意的产品，已经不再完全凭借个人的直觉 和行业经验，更多的还是需要数据挖掘和分析的支持，例如产品的结构设计、产品 如何定价，以及该产品适用人群分析等。

旧产品优化：每一产品都有其生命周期， 一个过去很好的产品现在可能已经过 时了，所以对于老产品的效果分析与挖掘是必不可少的，怎样正确的对当前产品的



**商** **业** **智** **能** **在** **电** **子** **商** **务** **中** **的** **实** **践** **与** **应** **用**

效果作出客观评价，同时对产品将来的发展趋势作出准确的预测都需要数据挖掘 技术的支持。

**2)客户深度认识**

服务客户：俗话说知己知彼，百战不殆。电子商务对于用户来讲提供的也是一 种服务，如何能够让用户对我们的服务认可，首先就需要对客户有所了解，这样我 们才能够有针对性地为客户提供服务。例如客户细分、流失客户挖掘等都是通过 数据挖掘的技术使我们对客户能有一个更深入地了解。

客户风险控制：对客户深度认识除了能够为客户更好的服务以外，还可以控制 由客户带来的风险，尤其是互联网这样一个特殊的行业，欺诈等现象时有发生，影 响非常严重，可是少量的异常客户隐藏在海量的客户当中我们很难区分，但有了数 据挖掘技术的支持，我们可以对客户有更深度地认识，就可以把异常客户区分出 来，并对异常客户进行监控，从而对客户可能带来的风险进行控制。

**3)精确化营销与服务**

精确化营销与服务是基于对客户深度认识的基础上，针对于不同的客户采用 不同的营销策略和服务策略，提高营销的成功率和服务效果。要做到精确化单靠 人工去完成成本非常大，或者说根本无法实现，但数据挖掘技术可以精确地发现用 户的需求并且挖掘出最合适的产品与营销服务策略，销售与客服人员可以通过数 据挖掘结果的数据支持，自动的实现精确化的营销与服务。例如现在大型网站都 非常流行的个性化推荐服务以及客户维系挽留系统，都是数据挖掘技术在精确化 营销与服务的应用。

32

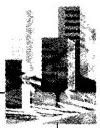
**4)精确化推广与市场培养**

精确化推广：网络推广是电子商务网站得到迅猛发展的重要手段之一，网络推 广最大的难题就是推广成本与推广效果的平衡问题，如何能够在有限的推广成本 前提下取得最好的推广效果，数据挖掘技术可以通过公司内部以往推广的一些历 史数据挖掘出最有效的推广方式，提供辅助精确化推广的决策支持。

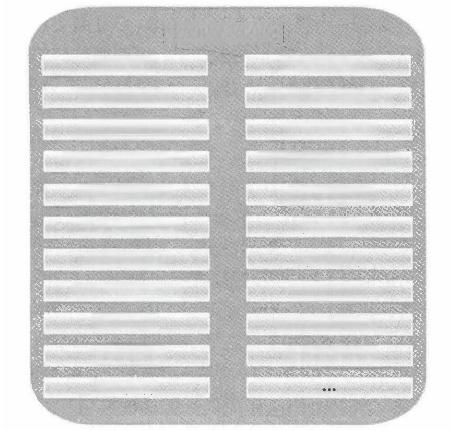
精确化市场培养：市场是公司赖以生存的外部环境，数据挖掘技术能够对目前 市场情况作出准确评估，并对市场将来的可能的变化趋势作出预测，使决策层根据 市场分析结果作出正确的市场决策，抓住市场机会，规避市场风险。

**3.5.2** **主要应用**

上文中我们在讨论数据挖掘价值体现的时候实际上已经提及了一些数据挖掘 的应用，例如客户细分、客户流失预警等。就是通过这些方方面面的数据挖掘应 用，数据挖掘的价值才能在电子商务行业中体现出来。经过总结，数据挖掘在电子 商务行业主要有如图3.18所示的应用。



**第** **三** **章** **商** **业** **智** **能** **在** **电** **子** **商** **务** **的** **解** **决** **方** **案**



数据挖掘应用

客户流失预警

产品搜索效果评估

关联产品推荐

搜索利益最大化挖掘

点击流挖掘 市场评估 市场预测

风险预警

买家个性化推荐 销售工作预警

客户分群 潜在客户挖掘

客户价值提升 客户维系挽留设计

客户价值评估 客户忠诚度评估 客户需求预测 客户满意度评估

产品设计

产品推荐

产品定价

**图3.18** **数据挖掘的应用**

**3.5.3** **应用实现要求**

**1)完善的开发规划**

33

应用之间的关系要求：数据挖掘的多项应用并不是独立的，绝大部分应用需要 联合起来才能发挥价值，例如客户维系挽留系统一般就需要以客户流失预警系统 为前提，客户细分是大部分其他应用的基础，这种依赖关系存在但是又不绝对，所 以导致不同应用之间的关系非常复杂。必须制订完善的开发计划。

应用的价值要求：对于一般的应用开发计划肯定是价值大、难度低的先开发， 但是数据挖掘的应用有时候价值的大小很难直接评估，例如客户细分从短期内看 价值不大，但是它是很多挖掘应用的基础，长期来看价值非常大。因此应用的价值 和长远的开发计划关系非常密切，只有详细的开发计划出来了，才能准确地评判什 么样的应用应该先做，什么样的应用应该后做。

数据要求：数据挖掘一般对历史数据都有要求，虽然现在的电子商务都有自己 的数据仓库，储存了大量的历史数据，但是具体到某一个实际的应用，我们还是经 常会发现数据不足的问题，原因就是历史数据往往是业务系统实现现有业务而产 生的，不一定会考虑到后续的数据分析与挖掘，缺少一些关键的分析指标，因此经 常为了能够收集到这些数据而将一些很重要的数据挖掘应用延后，如果我们有完 整的数据挖掘应用开发规划，就可以根据计划提前准备必要的数据。

**2)与时俱进的挖掘技术**

虽然多数的数据挖掘应用只用到了最基础的挖掘算法，基础的挖掘算法非常

**商** **业** **智** **能** **在** **电** **子** **商** **务** **中** **的** **实** **践** **与** **应** **用**

有用，但是挖掘技术毕竟是一门难度非常高的技术， 一旦需要用到比较高端的挖掘 技术时才来学习是来不及的，因此在平时的工作当中，保持挖掘团队技术的先进性 是非常必要的，这些技术是数据挖掘应用能够稳定实现的有力保证。

**3.5.4** **常见问题**

**1)执行效率**

复杂算法的执行效率：电子商务行业是业内公认的海量数据的行业之一，在这 里有很多效果非常好的算法因为效率的问题而搁浅，特别是有些实际的数据挖掘 应用不能用抽样的方式解决效率问题。因此在电子商务行业有时候并不是结果最 准确的挖掘算法才是最好的。 一定要找到成本与效果的平衡点，这样的挖掘算法 才是最好。

业务系统实时要求：很多挖掘算法不是把挖掘结果分析出来就算结束，这些结 果还要提供给业务系统，供业务系统实时处理事务时使用。这类需求要求实时性 要求非常高， 一种比较实用的方法是在BI 系统中离线训练挖掘模型，等模型规则 评估通过以后，将规则封装到业务系统中进行实时处理。另一种方法就是尽量提 升挖掘算法的效率，BI 实时处理数据并实时同步到业务系统当中。

**2)数据质量**

34

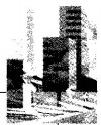
数据缺失：电子商务行业大部分用户是虚拟的，这就导致了用户产生的数据经 常是部分信息丢失，例如用户如果不登录就在网站上进行浏览操作，那么这些操作 的主体信息就会缺失，间接地与主体相关的信息也会缺失。

数据失真：电子商务行业并不是完全需要实名制，因此很多信息用户自愿填 写，填写只要符合规则就可以，对填写的真实性一般是不做验证的，这就导致了大 量数据失真。

做电子商务行业的数据挖掘的时候，由于数据的缺失与失真导致数据的预处 理一般要比其他行业工作量大很多，挖掘的难度也大很多。在这一点上在做电子 商务行业数据挖掘项目的时候需要充分评估项目的难度与实现周期，提前做好 准备。

**3)云计算**

近年来云计算越来越流行，特别是发展迅速，数据量暴增的电子商务行业，云 计算的应用取得了长足的进步；相对于传统的数据挖掘技术，能不能更有效的和云 计算相结合，开发出基于云计算技术的数据挖掘应用，是电子商务行业未来数据挖 掘新的发展方向。焦点科技目前做的云计算平台下的个性化推荐系统就是基于云 计算平台的数据挖掘应用。



35

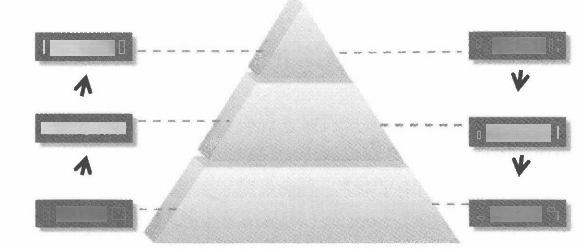
**第三章** **商业智能在电子商务的解决方案**

**3.6** **BI** **在电子商务领域的应用**

**3.6.1** **电子商务运营决策分析**

随着整个电子商务行业格局的不断调整，B2C 、B2B 、C2C网站层出不穷，我国 的电子商务运营企业，正面临激烈的竞争和前所未有的挑战。如何提高对市场的 快速反应能力，改善企业管理水平，提高经营效率，达到建设行业领先商务平台的 战略目标，成为我国电子商务行业当前面临的一个重要课题。

商业智能作为一种企业辅助决策的解决方案，在传统制造业取得了显著的效 果。电子商务行业和传统的制造业有着很大的不同，电子商务行业不仅要辅助公 司的中高层做决策，还要和网站的运营部门结合起来，做网站的分析。如图3.19 所示，电子商务决策分析系统为分析层、决策层提供决策分析平台，为操作层以及 各业务系统提供执行优化的智能帮助。



指标和报表

分析层

分析4策咯

操作层

(ERP、SCM、CRM、MES、HR、FINANCE)

**对数据进行分析**

战咯决策

执行优化

数据积累

决策层

**图3.19** **电子商务决策分析系统**

**1)企业级** **KPI 决策系统**

满足企业信息整合，由粗放型经营和管理向集约型转变。对于企业决策者也就 是公司的高层人员而言，必须对企业运作中的各种信息和数据进行监控、分析、管理， 并据此进行判断、决策，进而采取行动。随着各种数据处理和分析应用软件的涌现和 升级，企业发现需要处理和分析的信息和数据越来越多，越来越复杂，在处理和分析 信息中所投入的时间成本更大。而使用仪表盘能可视化地呈现信息，且有助于判断、 监控并支持决策，从而有效地提升信息系统的实时信息处理能力，这一特性使得“仪 表盘”成为缓解和解决上述问题的重要途径，并日益赢得使用者的青睐。

**2)部门级绩效考核系统**

目前很多企业在绩效考核中面临的问题是这些企业大多建设实施了 ERP 系

**商** **业** **智** **能** **在** **电** **子** **商** **务** **中** **的** **实** **践** **与** **应** **用**



36

统，实现了业务处理的自动化，但在管理决策上很多企业并未真正实现管理信息自 动化。这样企业绩效管理的处理过程是，ERP 系统进行业务处理产生了大量业务 数据，管理部门在此基础上把数据从 ERP 中倒出，利用数据库处理或 EXCEL 手 工进行数据整理和分析，按照企业的绩效管理的考核方法，如 KPI 的各种指标等 进行计算和分析，最后得出绩效考核的“数量”结果。

**3)业务监控及分析**

对于电子商务网站，主营业务就是网站，了解网站运营情况，对运营进行 系统有效的分析至关重要。运用BI 对运营数据进行分析，从而提升网站整体 运营效果，监控并分析网站运营状况，找出存在的问题，提出改进优化意见，实 施改进措施，进一步提升网站整体性能，获得更高的知名度，留住老用户，吸引 新用户。

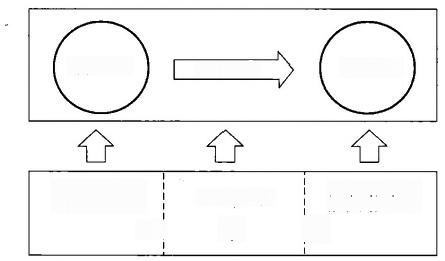
**4)分析型工作流程**

目前市场成形的大型分析型系统主要有 ERP 系统、SCM 系统、CRM 系统等， 这些系统都给公司运营管理带来了巨大的价值。因此要提高公司在电子商务领域 的核心竞争力，分析型流程定制是公司运营发展的必然趋势。分析型工作流程纳 入到运营系统后，会把BI 和业务紧密结合，使公司运营系统化、体系化，同时也会 提高运营系统的响应效率与质量，从而提升操作人员的工作效率与工作质量，规范

化运营体系保障业务运营的正常进行。

**3.6.2** **对外-支撑网站用户分析**

BI在支撑网站上有其自身的优势，无论是完整的成熟仓库架构还是先进的数 据挖掘技术，都是其他系统无法取代的。网站作为服务客户的一个平台，网站运营 无非是以下几个方面：用户对象、服务工具、服务方式，而BI 在电子商务服务中的 这几个方面都有应用(图3.20),下面是其具体应用框架。

服务工具 服务方式 用户对象

异常用户挖掘

网站功能优化 数据增值服务

个性化服务

系

规范用户入口

BI

统

**图3** **.20** **BI 网站支撑系统框架**



**商** **业** **智** **能** **在** **电** **子** **商** **务** **的** **解** **决** **方** **案**

**第** **三** **章**

**1)个性化服务**

电子商务竞争日趋激烈，网站的服务质量对于网站的发展至关重要。网站的 用户千差万别，采用统一的服务不可能满足所有用户的需求，如何对每一个用户采 取有针对性的个性化服务就是BI 的应用领域之一。

个性化服务主要可以区分为：问候、用户定制、实时消息和特色推荐。但是无 论哪一种个性化服务都必须面对服务时机的选择、服务对象的识别、用户偏好的总 结以及服务与对象匹配几个环节，只是有时候某些情况下某些环节相对简单比较 容易实现。

个性化服务做得好，可以大大缩减用户浏览网站所花费的时间，提高用户的效 率，增加客户对站点的好感，提高用户黏性，对增加网站用户规模、提升网站知名度 都有好处。

**2)网站异常用户挖掘**

电子商务网站欺诈现象一直是用户非常关注的一个问题，很多电子商务企业 甚至把欺诈问题看成是关乎网站生死的大问题来抓，但是在互联网这种特殊的环 境模式下欺诈问题一直很难解决。

电子商务的欺诈现象，主要形式有诈骗、钓鱼以及一些虚假注册、重复注册等。 这些异常用户隐藏在海量的用户群体当中，利用传统的统计方法和人工识别方法 很难区分，BI 可以利用数据挖掘技术，根据用户的历史行为数据的挖掘分析，以及 用户基本信息的文本挖掘等手段，将异常用户与其他的正常用户进行区分，并将分 析结果提供给相应的负责人进行处理，大大提高异常用户的甄别能力。

37

网站欺诈用户的减少，为其他用户的正常交易提供一个安全的环境，使用户在 网站上交易更放心，无疑会大大提高用户对网站的认同感，对于网站的发展有着非 常积极的作用。

**3)网站效果分析及优化**

任何一个网站都不可能很完美，网站都是随着功能的不断改进和完善而不断 进步的，那么功能好不好，应该如何完善就成了网站时时刻刻需要面对的问题；同 时网站作为电子商务企业获利的工具，网站现有资源是否合理利用也是企业面临 的另外一个问题。

BI 通过数据挖掘与分析，剖析用户使用网站功能的每一个细节，通过流量分 析，与用户访问路径的挖掘，合理评估网站功能的好坏，并对网站功能可能存在的 问题进行定位，为产品人员合理的优化和改进相应的功能提供参考依据；BI 通过 网站资源的流量监控，对网站资源的价值进行评估，并发现新的有潜在价值的优秀 资源，为运营部门对资源产品的合理定价提供参考依据。

网站功能的改进和优化可以保证网站能够与时俱进，确保网站能够有持

**商** **业** **智** **能** **在** **电** **子** **商** **务** **中** **的** **实** **践** **与** **应** **用**



38

续的竞争力，使网站能够不断地做大做强；网站资源的合理利用则是网站持续 盈利能力的保证，用户认可并且具有强大盈利能力的网站才是真正意义上好 的网站。

**4)数据分析增值与服务**

电子商务网站多年来沉淀了大量的数据资源，这些数据资源就好比有价值的 矿石，除了利用这些资源优化网站间接的为企业创造价值以外，能不能运用这些丰 富的数据资源，提炼出对用户有价值的信息，提供给需要的客户，从而直接为企业 创造收益成为了很多企业的目标。

电子商务网站可分析的有价值信息很多，根据电子商务行业的特点这些信息 可以分为下面几类。

· 可以帮助用户更加了解市场的市场分析：如用户所处行业的供求变化趋 势，哪些产品热门等。

· 可以帮助用户更加了解他的客户的客户分析：例如用户来源于哪里，他们 是如何找到我的，哪些客户需要重点关注等。

· 可以帮助用户提高他在网站上推广效果的效果评估及优化：例如用户购买 推广产品的推广效果，哪些推广方式可以帮助用户提高推广效果等。

这些分析既可以帮助用户发现市场机会规避市场风险，又能帮助用户提高推

广效果，核心就是在为用户创造价值。用户能够通过我们的数据分析增值与服务 创造价值，那么他们自然愿意花费这些价值中的一部分来购买这些服务，因此数据 分析增值与服务所能为公司创造的价值的多少完全取决于增值服务为客户带来的 价值的多少，为客户带来的价值越高，那么增值服务本身的价值就越大。

从上面的分析我们不难发现，网站作为电子商务企业服务于用户的一个工具， 核心还是服务于用户，因此 BI 对网站的支撑实际上就是对网站用户的支撑，用户 的需求在变化，那么 BI 对网站的支撑也要变化，以用户的价值为导向的 BI 的价值 才能够得以体现。

**3.7** **软硬件方案**

商业智能在电子商务行业中实施面临着较多的挑战和问题，如何选用合适的 软硬件产品来支持企业的 BI实施是一个重要的问题。

首先是处理海量数据的问题，随着业务的发展，电子商务企业每天都会产生巨 大的数据量，比如网站用户信息、网站访问日志等记录，这些数据少则数十万数百 万，多则上亿；如何选择合理的软件硬件方案来高效准确地处理这些数据是商业智 能成功实施的一个关键。商业智能在电子商务行业中会面对不同层次的用户，从



**第** **三** **章** **商** **业** **智** **能** **在** **电** **子** **商** **务** **的** **解** **决** **方** **案**

底层的操作人员到公司高层领导，还包括网站用户。不同层次和类型的用户会产 生大量的差异化的需求，用户对实时性、数据粒度、用户终端等各方面的需求各不 相同，因此BI 实施过程中需要选择合适的软件和硬件，才能保证数据顺利的从业 务系统、数据仓库、数据集市一直到最后的报表展现都能满足不同用户的需求。

电子商务行业的 BI 涉及的关键技术有数据仓库及 ETL 技术、OLAP 技术和数 据挖掘技术，市场上已经存在着许多厂商的BI产品以及一些开源的项目来支持这些 技术(表3.1)。BI产品种类繁多，每家厂商几乎都会宣称自己的产品是如何的好，在 构建企业自身的 BI系统时应该根据企业自身的需求考虑软件硬件产品的选择。

**表3.1** **主流** **BI 工具**

|  |  |  |  |
| --- | --- | --- | --- |
| 数据库 | ETL工具 | O LAP工具 | 数据挖掘工具 |
| DB2 | DataStage | BO | SAS EM |
| Oracle | PowerCenter | Cognos | Clementine |
| SQL server | SSIS | QilikVicw | Weka |
| Teradata  Greenplum | Datalntegrator |  |  |

BI产品主要涉及数据库、ETL 工具、OLAP 工具及数据挖掘工具，有的厂商有 很全的产品线，从数据库、ETL 到前端展现及数据挖掘都有涉及，能够提供全面的 解决方案，比如Cognos 和 Oracle;用户可以购买厂商的整体解决方案实现 BI。很 多BI 产品都是很优秀的，但是对于企业来说要考虑各方面问题，好的产品不一定 会适合自身。

39

数据库方面 DB2 、Oracle都是很好的产品，Teradata 和 EMC 公司的 Green- plum 在大数据处理方面有着很强的能力。根据 Gantner 在2012年2月发布的数 据仓库魔力象限[]显示 Teradata在数据仓库方面处于领导者地位， Teradata 在实 际的应用中也显示出其强大的数据处理能力，另外 Oracle 、IBM也处于领导者象 限，EMC 公司的 Greenplum 凭借其云计算方面的能力也进入领导者象限(图 3.21)。ETL 工具方面比较主流的产品有 DataStage 和 PowerCenter, 这两者占据 了国内主要市场份额，DataStage 在被 IBM 收购之后显得后劲十足。分布式部署 云计算是解决大数据最好的方法，Greenplum 在云计算方面表现出色，但是这是需 要付费的，当前主流的 Hadoop 也很流行，可以让企业以较低的硬件投入实现分布 式，当然这对开发人员的要求更高。

主流的 OLAP 工具有 BO 、Cognos 以及QlikView 等，要评价工具好不好除了 看数据准不准这类开发问题，更重要的就是要在考虑功能强大、开发简单之外，速 度、美观、简单必不可少；很多厂商在性能上有自己解决办法，但大部分对开发者设

labilily to execute H

 **商** **业** **智** **能** **在** **电** **子** **商** **务** **中** **的** **实** **践** **与** **应** **用**

challengers leaders

|  |  |
| --- | --- |
| 1010data | Teradat  看  \*Oracle  IBM  ●  ·EMC/Greenplum  Sybase an SAP Company  Microsoft |
| ParAccel Kognitio  SAND Technology  Infobright  \*Actian  ●Exaso | Vertica |

niche players visionaries

—completeness of vision I—

As of February 2012

**图3.21** **Gantner 数据仓库魔力象限**

计能力要求较高，比如Cognos 产品就要求编程实现，这样就需要开发者对产品非 常熟悉才可行。Gartner 前端报表工具最具魔力象限连续几年纳入了 QlikView 这 个产品，这个产品代表了基于内存的报表分析技术，美观简单，用户体验极佳。

数据挖掘产品领域，有 SAS 、SPSS 等两大厂家，而像 IBM 、Teradata 也都有

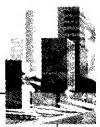
40

自己的挖掘工具。相对而言，SAS EM功能更全面，性能更加优秀，算法也更全面，

具有图形化界面，可视化操作。SPSS Clementine 开发更简便，功能上集成度更高， 其中 SAS 需要编程，对开发者的要求更高一些。做好数据挖掘仅仅有工具是不够 的，更需要数据挖掘人员对数据挖掘的理解和认识。

**3.8** **实施策略及风险控制**

客观地说，懂得安装数据库、建表、写存储过程，再会用展现工具开发报表基本就 算 BI入门了，但要实施企业级的BI项目，只满足于这几点是远远不够的。如果实施 策略不当，很可能造成投入巨资硬性推动后产生不了价值，或者东西很好并且有价 值，因为方法不得当最后还是得不到客户认可，甚至被否定。所以这里我们来讲一下 BI的实施策略，有代表性的恰恰是数据仓库支持OLAP 理念的两位创始人Bill Inmon 和 Ralph Kimball,Bill Inmon倡导自上而下实施策略(图3.22),而 Ralph Kim- ball 则提出不同看法，认为数据仓库应该实施自下而上的实施策略。但经过那么多 年的BI实施，对于一位有经验的BI实施人员我们认为可以采用第三种策略，那就是 整体规划、分步实施，这种方式可以融合两种策略的优点，又可以避免其中的短处。

**第** **三** **章** **商** **业** **智** **能** **在** **电** **子** **商** **务** **的** **解** **决** **方** **案**

41



**VS**

Ralph Kimball

Bil Inmon

自上而下

自下而上

优点

●投入小、周期短 ●见效快、风险小 ●业务理解高

缺点

●数据孤岛

●整体设计有挑战

优点

●数据仓库有整体设计

●数据一致性高

缺点

●投入大、周期长

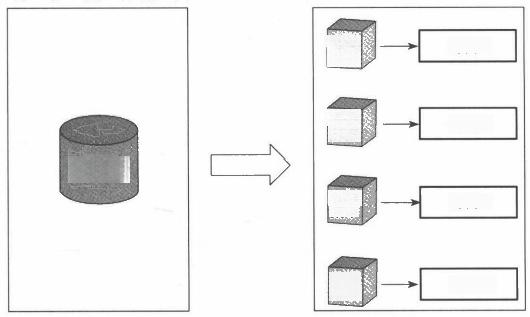
●业务理解低

●见效慢，风险大

**图3.22** **传统的两大** **BI实施策略：自上而下、自下而上**

**3.8.1** **自上而下策略**

自上而下实施策略(图3.23)最开始由 Bill Inmon 提出，其核心思想还是突出 前期的设计进而达到开发一步到位，实施的时候先统一规划业务，整体设计数据仓 库，在数据比较稳定后再面向各个部门及各个层级人员整体实施企业级 BI应用。



第一步：建立数据仓库 第二步：每个部门建立一个集市

数据

集市

数据

集市

数据金库

部门3

部门…

数据

集市

数 据

集市

部门1

部门2

**图3** **.** **23** **自** **上** **而** **下** **BI 实** **施** **策** **略**

这种策略优点是：

(1)数据仓库模型统一设计，结构简单易于管理及扩展，冗余数据少。

(2)虽然有数据集市，但是数据集市是基于数据仓库的基础上进行设计，数据 来源于数据集市，可以说是数据仓库派生出来面向特定主题的一个子集。

(3)数据仓库和数据集市在设计上属于两个不同层，只要数据仓库质量比较 高，增加主题应用会变得相对简单。



**商** **业** **智** **能** **在** **电** **子** **商** **务** **中** **的** **实** **践** **与** **应** **用**

其缺点是：

(1)初期就大规模实施数据仓库，投资大。

(2)需要较长时间才能使数据仓库开发到比较完善的程度，从而导致项目周 期长，见效慢。

(3)业务需求过于追求全面，导致项目难度大大提升，普遍情况是在 BI 实施 之前大部分用户是很难想象 BI 有什么功能，能对他们有什么帮助，更难提出合理 意见，这种时候唯一的办法就是引入高端咨询人员进行需求规划及设计。可想而 知对于塑造差异化竞争力的今天来说，企业采用自上而下的实施策略存在巨大的 风险。

从上面我们知道，自上而下的实施策略要发挥其优点避免风险必须具备如下 条件：

(1)有足够的财力。

(2)直接打造企业级数据仓库，需要高层领导支持，解决跨部门协调资源的 困难。

(3)业务形态较为常见，有强大的成熟的业务咨询团队和实施团队，有能力化 解整体设计及开发过程中的任何业务及技术问题。

(4)业务系统较为统一，数据仓库数据质量有足够保证。

42

另外 Bill Inmon 提出自上而下的实施策略同时要求数据集市能够上下钻取， 这种功能刚好数据立方体(Cube) 工具能够实现，最著名的有 Hyperion 公司的 Essbase, 还有就是 Cognos 的 Powerplay, 这些数据立方体工具笔者估计一开始应 该是为了实现Planning 应用而开发的，我们知道 Hyperion 和 Cognos 的 Planning 软件都是很知名的，Planning 软件在时间和组织架构上非常需要往上汇总、往下分 摊，后面这种上下钻取的功能因为刚好满足BI 的基本需求，过去十年被 BI 厂商广 泛采纳和推广，并取得了不错成绩；但是从解决方案和工具来看，现在技术和工具 只能说广义上的Bill Inmon的上下钻取，现在上下钻取在技术和架构上有如下三 大流派：

(1)Cube 数据集市支撑的物理 OLAP。

(2)关系型数据库加前端缓存支撑的逻辑型 OLAP。

(3)关系型数据库加内存支撑的逻辑型OLAP。

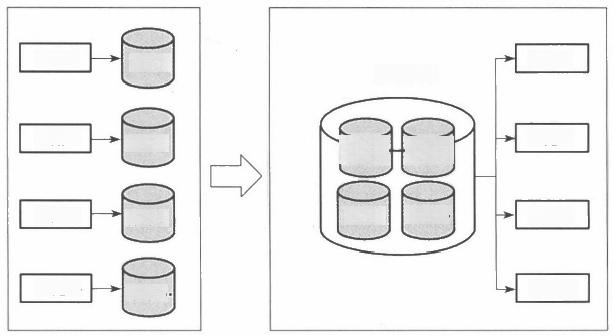
从上面的划分来看，当前数据模型设计和前端展现工具之间、软件和硬件之间 互相支撑，这也是一种趋势，特别是现在云计算的成熟也必将促使BI 在未来3～5 年里其技术及工具方面发生巨大变化；从这个层面笔者再次提醒我们开发者要真 正理解 BI的本质和问题，了解主流的工具和技术，只有正确认识前者，合理利用后 者才能做好企业级 BI实施，特别是具有大数据的电子商务行业更是如此。

**第** **三** **章** **商** **业** **智** **能** **在** **电** **子** **商** **务** **的** **解** **决** **方** **案**

43

**3.8.2** **自下而上策略**

**虽然都是数据仓库理念的创始人，但不同于** **Bill Inmon,Ralph Kimball 提倡** 自**下而上的数据仓库实施策略(图3.24),Ralph Kimball 认为“不同数据集市合起** **来就是我们要的数据仓库”,这种策略的优缺点也同样明显。**



第 一步：按主题建立集市 第二步：分主题集市合并为数据仓库

主题1

数据仓库

主题2

集市2+

主题3

主题…

集市

集市3

主题…

集市…

主题1

集市2

集市3

集市1

主题2

主题3

集市…

**图3.24** **自下而上** **BI 实施策略**

这种策略优点是：

(1)直接设计数据集市进行OLAP 分析，让BI 快速产生价值，项目见效快。

(2)不断叠加的数据仓库建设策略能保证数据仓库设计更接近实际需要，避 免仓库特别是模型的不合理设计，从而降低风险。

(3)另外在自下而上的策略里面Kimball提到维度建模的概念，从而产生了当 前最为流行的数据仓库按照维度和事实的设计，特别是Kimball 对维度建模的若 干建议对数据仓库设计具有非常高的指导价值。

其缺点是：

虽然前期集市设计速度快、风险小，但数据集市设计更多是在部门级或者比较 单一的业务领域，这种设计难于推广到其他业务部门和领域，所以前期只注重数据 集市设计是难于形成企业级数据仓库，具体来说有以下三个问题难于统一：

(1)数据粒度：数据仓库一般尽量保持必要的明细，而数据集市则是看当前业 务需要。

(2)数据范围：数据集市只选择当前需要的数据，哪怕是同一个表在不用应用

的时候也会在不同程度上进行过滤。

(3)数据定义：这个是问题最大的，业务部门的指标选择和定义往往有差别，

 **商** **业** **智** **能** **在** **电** **子** **商** **务** **中** **的** **实** **践** **与** **应** **用**

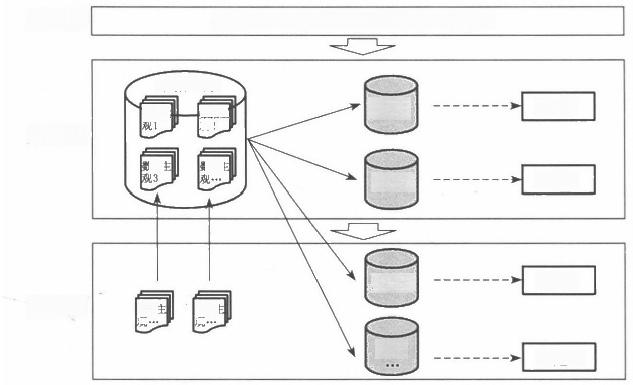
所以不同业务部门的需求在集市设计上必然会有区别。

这里不得不说的是 Bill Inmon 和 Ralph Kimball 提出的策略都没有绝对意义 上的自上而下或者自下而上，但在实际工作中，BI 项目组、IT 领导往往都喜欢绝对 的自上而下，能够获取到公司最大的资源支持，结果也容易导致投入大，但是见效 小而慢，甚至失败；而业务部门在久久看不到效果的时候又总是喜欢绝对的自下而 上，每个部门自成一派，最终企业级 BI 迟迟建立不起来。为此我们还是要从本意 上理解两种策略，特别是吸取其长处。

从上面我们知道自上而下、自下而上两种策略都有可取之处，也各有致命弊 端，显然实际开发及应用过程中，谁都不会简单使用其中一种，都会两者兼备，这里 最为提倡的是整体规划，逐步实施，下面进行详细介绍。

**3.8.3** **整体规划、逐步实施**

简单概述不难发现 Bill Inmon 重视数据仓库的整体设计，从而保证系统的完 整性，而 Ralph Kimball 重视快速推进，快速出效果，重视项目的实用性。应该说两 者意图都值得肯定，所以我们提倡两者兼备的整体规划、逐步实施，如图3.25 所示。



数据分解(把数据分解成相对独立的主题)

数据仓库

主题1

数据主

第二步

主题2

集市3 主题3

数据主

观

主题…

第三步

第一步

集市2

集市1

数据主

观2

集市

数据主 观 ·

数据主

据主

**图3.25** **整体规划、分步实施**

这里我们建议把 BI 实施分成三步走。

第一步：把具有相互独立性的业务数据分成不同的主题，特别是要明确核心业 务主题及维度信息表。

第二步：选定1～2个分析主题，对核心业务进行数据仓库设计并且实现数据 入库；维表设计和开发尽可能一步到位，满足主题分析的集市数据要求； 一个分析

**第** **三** **章** **商** **业** **智** **能** **在** **电** **子** **商** **务** **的** **解** **决** **方** **案**

主题一个集市。

第三步：根据用户使用情况及规划可以不断增加分析主题，扩展主题集市，如 果数据仓库数据不够，那么需要不断补入数据， 一般尽可能同一主题数据一起 补入。

分步实施在推进整体规划上需要遵循下面的原则。

(1)基本原则：分步实施虽然要满足整体的规划，体现其整体性，但更要重视 每一期项目的实效性，项目实施过程中可以根据实际情况对规划及计划进行合理 调整。

(2)制订实施计划：基于整体规划和主题划分，对业务主题进行重要性分级及 差距分析，从而形成分步实施计划。

(3)基本任务：针对不同的业务需求建立独立的集市，以满足不同的数据分析 需求及性能需求；对数据仓库基本数据进行逐步完善，重视模型的完善和数据质量 管理。

(4)初期目标：BI 初期开发偏向于系统的基础性建设及快速出效果。在初期 一般都会把数据仓库建设到70%左右，完成公共维度和基本事实数据的入库；BI 实施效果得到体现，可以让企业对于 BI价值有直观认识，便于后面工作开展。

(5)中远期目标：中远期重视数据质量管理和 BI 整体应用体系。数据质量问 题对 BI应用是致命的，长期来讲需要建立起数据质量管理体系；还要成体系的推 进 BI应用，让BI 支持的分析点互相支撑，形成统一的企业级决策分析平台。

45

优点：

(1)整体规划、分步实施很好地融合了自上而下及自下而上BI 建设思路的各 自优点。

(2)注重数据仓库的整体设计，70%左右的前期设计可以建立起数据仓库的 基本框架。

(3)分步实施的业务应用可以让 BI不断得到回报，从而让企业更有信心。

(4)同步推进的业务应用可以不断检验数据仓库模型的正确性，从而避免在 仓库模型建设上出现大的反复。

缺点及应对措施：

这种策略的不足就是怎么来保证项目的整体规划能力和分步实施效果。

(1)业务规划能力：BI 应用规划不得不面临不同主题间的横向区别，同一主题 业务对于不同级别用户的需求也是有区别的，这就要求 BI 规划人员既要懂业务又 要懂BI 应用，且要有较高的 BI行业咨询能力。

(2)技术方案能力：因为数据仓库的完整性，BI 分步实施是对技术方案的一个 挑战，如果没有领先的技术方案是难以保证项目的独立性和设计的整体性，所以团



**商** **业** **智** **能** **在** **电** **子** **商** **务** **中** **的** **实** **践** **与** **应** **用**

队必须要有领先的 BI 技术方案。

(3)项目实施能力：分步实施重视不断出效果，如果前期效果连续失败，那整体

规划还是一句空谈，因此保证项目实施效果，特别是初期效果尤其关键，所以要求 项目组有较高的项目实施能力；另外可以选择对报表分析需求较大的核心业务部 门开始，并获得业务部门的支持，这样成功可能性要大。

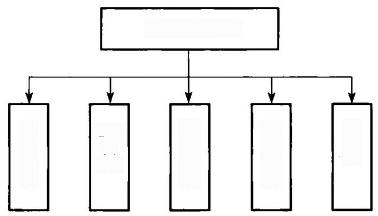
**3.9** **BI 项目风险控制**

BI实施风险包含技术、业务、实施、推广等方面，但BI 最大风险还是来自于很多 人认为BI 可有可无，有则应该更好，无也没关系的这样一个定位。事实上BI的应用 是不能将就的，必须实实在在有价值，确实好用才行，所以其难度和风险非同一般。

**3.9.1** **组织风险**

不管什么事情最终都是人在落实，所以组织的合理性是控制风险的第一步，要快 速实现企业级 BI应用需要非常合理的划分组织功能，让每一块的人做自己专业的 事，并且每一块的要求门槛要尽量低。如图3.26 所示的企业级 BI组织具有较大 优势：

46



維 护

数据挖捌

报表分析

咨询团队

BI 团队负责人

DW 及 ETL

**图3.26** **BI 团队组织结构**

在团队负责人下面设立咨询团队、DW 及 ETL 团队、报表分析团队、数据挖掘 团队及 BI 维护团队，咨询团队的设立是我们的一大特点，并且为项目快速出效果 起了决定作用。

(1)咨询团队：是该组织结构里面最为重要的团队，要求团队成员精通 BI 技 术及业务，并且有很强的项目规划及推广能力，软实力上乘；能准确对公司BI 业务 进行规划，对规划点能有序的安排实施，随时把控项目存在风险，能根据用户实际 情况快速合理地做出项目调整，做到快速推进项目，并取得实际应用成果；这个团 队要求有基本的咨询素质、很强的做事欲望、丰富的BI 实施经验、上乘的协调沟通 能力、熟悉咨询方法论且有较高的文档录入及表达能力；这个团队是所有项目的龙



**第三章** **商业智能在电子商务的解决方案**

头，事实证明咨询团队建立得好则BI 项目的成功率就高，反之失败可能性就大。

(2)DW 及 ETL 团队：团队任务是负责数据仓库、ETL 技术架构及标准定制， 负责BI 项目的 ETL 开发以及负责 ETL 及 DW 技术研究。在电子商务行业里， DW 面临的大数据处理问题是该团队区别一般行业的重头所在，在 DW 相对较为 稳定之后，建议把 DW团队独立出来，专门从事大数据存储及运算的工作，而项目 中的ETL 和数据集市设计则由 ETL 团队负责设计及开发。

(3)报表分析：这里报表分析广义上包含大数据查询与导出、报表、多维分析 及仪表盘。传统的对内报表分析在电子商务行业也同样需要，没有什么太多不同， 都需要企业级仪表盘、OLAP 分析、分析型流程定制等。不同于对内的报表分析， 电子商务的 BI应用重心在于对网站用户分析支持，所以要求报表分析的功能更加 强大和灵活，网站的报表分析偏向简单及合理设计，用户体验是该工作的重大挑 战，从事这部分工作的人员必须有很强的网站用户体验背景或能力。

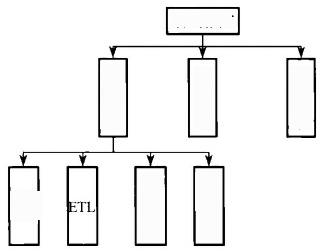
(4)数据挖掘：专门从事项目的数据挖掘技术开发，承担企业级数据挖掘模型 研究和人才培养。因为数据挖掘的特殊性，该团队非常有必要和大学、科研院所进 行合作，促进自身技术不断提升，随时掌握最前沿的挖掘技术和应用。

(5)BI 维护：不同于一般业务系统，BI 项目每天有大量数据需要处理，很容易 出错，如果团队成员没有较好的BI 基础，那么在遇到问题时候只能束手无策，所以 需要配备有BI 背景的人员进行维护才比较可行，当然这部分维护人员也可以放到 公司维护团队去。BI 项目没有数据是万万不行的，所以如何为BI 项目快速提供样

47

板数据是维护团队必须解决的问题。

BI 项目团队组织架构如图3.27所示，采 用咨询人员负责效果，项目经理负责实施的 组合方式。这样可以保证企业内实施 BI, 在 BI团队内部既要对开发负责，还要对结果负 责，让最懂BI 的人来总控有利于保证项目效 果和进度。因为咨询推广人员不用参与到每 个项目开发，所以往往一个咨询推广人员可 以同时带几个项目，这样可以最大限度发挥 咨询推广人员的高端及关键作用，有利于项 目的并行开发。



项目经理

数据挖掘

报表分析

项目经理

咨询推广

项 目 经 理

pw

**图3.27** **BI 项目团队组织架构**

**3.9.2** **技术风险控制**

到目前为止，BI 在国内已经快速发展了近十年，软件、实施、人才都已经较为 成熟，所以只要方法得当技术上不应该存在太大风险，笔者这里建议我们企业要把



**业** **智** **能** **在** **电** **子** **商** **务** **中** **的** **实** **践** **与** **应** **用**

**商**

下面几件事情做好。

(1)深度理解 Gartner 关于BI 产品的评价，慎重选择合适产品，多看看专业点 评，应该说BI 产品发展变化快，Gartner 的排名就变化快，要特别注意一些领先技 术在 BI中的应用，比如基于内存的分析、云计算、数据挖掘等。

(2)注意用户体验，很多BI 项目失败就在于其用户体验差。

(3)注意实施难度，好的产品开发几天能上手，有些产品开发复杂，甚至安装 都很困难。

(4)找个 BI 实施高手，保证项目的技术架构，这一点非常重要。最忌讳的是 业务系统的开发人员转行做BI, 因为BI 不管是在应用还是在设计上都与业务系统 有很大的区别，甚至有矛盾之处。

如果有资源，可以考虑内外结合地实施 BI, 提高 BI 项目前期的成功率。目前， 市场上已经有很多领先的公司具备较高的 BI实施水平，比如东南融通、文思创新、 华为、亚信等。

**3.9.3** **业务风险**

业务风险是BI 实施的巨大风险。我们最开始在公式化的实施 BI, 完全遵循用 户的要求，可能花了半年做了三个项目，到最后却没有用户使用，后来进行总结并 且调整项目经理之后很快重新得到用户认可，并且应用得越来越好，这里不得不说 BI业务风险巨大。如何降低业务风险，笔者认为应该注重下面几点。

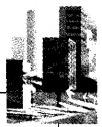
(1)普及 BI 知识：在实施 BI 初期，业务人员一般很难想象 BI 是什么,能给他 们带来什么价值，更提不出具体的需求，为此一定要提供一个直观的东西让用户知 道 BI 是什么,比如做一个 Demo, 在 Demo 里面把 BI的功能充分展现一下，拖拉旋 转、上钻下钻、图表制作等。

(2)用户决定需求：虽然用户提不出完整的BI 需求，但BI 人员要善于理解用 户的意图，合理设计其需求，并且最终由用户来决定。

(3)有用就做，不要等：如果用户提不出完整需求，可以把对用户有用的先做， 不求全，只要做出来之后用户会真正用起来就行，随着用户对 BI 的使用越来越多 会对BI 越来越熟悉，这样离提出全面完善的需求也就不远了。

(4)BI 目标是帮助用户：BI 功能很强大，但是BI 需求来自用户，BI 应用需要 用户，BI 提升完善更是需要用户，BI 不可能取代用户的工作，只可能帮助用户更好 地工作，让用户可以把更多的时间用来做更有价值的东西。

(5)业务规则需要统一：来自不同部门不同层面的需求，如果是同一个问题， 最好用相同的指标定义，并且具有相同的指标名称，不同定义的指标名称一定要区 别开来，避免部门间、上下间的信息不一致。



**第三章** **商业智能在电子商务的解决方案**

(6)选择合适的项目经理：因为BI 业务需求来源比较灵活，实施力度比较有 弹性，所以 BI 项目经理需要具备较好的沟通理解能力。

**3.9.4** **实施风险**

企业级BI实施我们建议采用整体规划、分步实施的策略，这种策略要成功实 施需要很多配套的管理办法。通过这一年多的BI 实施我们形成了完整的以效果 为导向的 BI实施方法，基本包括下面几点。

**1)QA 关键点评审及控制**

· 立项：主要看项目需求价值和实施的可行性，其中项目需求价值一定要非 常清楚，没有明确价值点和用户对象的项目不立项，最好是评审人员有丰富的分析 基础，并且了解公司业务。项目可行性一般较容易控制，不是太难。

· 需求及设计评审：和一般项目不同，BI 项 目DEMO 和实际开发很接近，难 以快速模拟，基本上需求、设计、开发一起出来，所以需求和设计放在一起评审较为 可行，这样在立项时候的项目实施范围及需求价值要求就有较高预判性，这恰好是 我们咨询推广人员的价值所在。

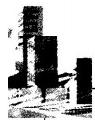
· 试运行：BI 项目最难的地方就在于数据能否准确， 一般来讲项目都需要一 个试运行阶段，在试运行阶段业务部门安排几个人员进行试用，将发现的问题反馈 给项目组进行修改， 一般来讲致命性问题(比如数据正确性问题)都要解决了才能 上线，其他不是很严重的问题(比如格式调整)可以放到下一期项目进行完善，从而 保证项目总体进度较为顺利，不会因为局部小问题导致项目周期大幅延长；另外 BI功能对很多用户来讲也是陌生的，用户还是要有初步的使用之后才能提出更多 的改进建议，所以没有致命性问题， 一般建议还是先推给用户进行使用，在使用中 进行完善。

49

· 正式上线：在试用一段时间之后我们需要给用户发调查问卷，如果用户认 为项目对他们有价值并且没有太大的问题那么就可以上线，这是以效果为导向的 项目开发流程中重要的一环。项目是否上线绝对不是项目周期说了算，而是真正 由用户的需求来决定；项目上线之后我们要持续关注用户的使用效果，监控项目的 访问情况，同时也要注意用户意见的收集；另外使用效果不好的项目，如果没有足 够的理由将不给予后期开发，如果是BI 人员问题那么需要对BI 人员进行调整，这 样能促进用户和开发团队更加重视项目的使用效果。

**2)技术方案及标准**

整体规划、分步实施最为考验的是项目的技术方案，要求方案能够有前瞻性， 这里需要一系列的解决方案；另外快速实施及出效果是分步实施的关键，这样对技 术标准也有很高的要求。因此对于BI 项目最起码必须具备如下方案和标准：

**商** **业** **智** **能** **在** **电** **子** **商** **务** **中** **的** **实** **践** **与** **应** **用**



50

· BI 系统架构方案；

· BI 系统数据接口技术标准；

· 数据模型设计技术标准；

· ETL 技术标准；

· 系统调度技术标准；

· 前端展现技术标准；

· 数据挖掘技术标准。

**3)项目效果评估及监控**

BI项目的效果评估非常有必要，可以取访问人次、访问时长、访问人数、用户 覆盖率等来作为评估的关键指标，对这些指标值及趋势进行监控，将有助于评估出 项目的价值，通过趋势的预警还便于我们及时发现问题，改进项目。

**3.10** **小结**

传统的数据仓库架构直接移植到电子商务行业会出现“水土不服”的现象，原 因在于电子商务行业的数据源多而杂，需要处理和分析的数据量巨大，对数据分析 的实时性要求较高，并且需要支持的用户量巨大。要解决这些问题，需要对数据仓

库的整体架构做出调整和优化。传统 BI 的数据仓库主要使用关系型数据库来支 持离线分析和复杂查询，难以满足大数据处理和实时查询等需求，因此可以结合使 用关系型数据库、内存数据库和分布式数据库以满足电子商务行业对于商业智能 在数据处理、数据存储等方面的要求。

数据仓库按照数据流向可以分为三层结构，包括数据层、信息层和分析层。

数据层主要是数据源和标准数据接口。信息层包括 ODS、数据仓库、数据集 市、ETL、元数据。分析层主要是数据的应用，如数据挖掘、报表分析、数据支持等。

在数据仓库架构上，笔者建议在传统数据仓库的基础上，同时采用分布式数据 仓库来支持BI 大数据的存储与计算，可以显著地提升数据处理和数据搜索的性 能。在 Hadoop 分布式系统架构中，数据仓库工具 Hive 和基于内存的数据库 HBase 可以分别解决大数据批量任务和实时性查询需求。

正确和完备的数据模型，是BI 进行一切数据分析的基础，是决定数据仓库项 目成功与否的重要因素。目前常用的数据仓库数据模型为多维数据模型，这种模 型主要以星形模式和雪花形模式在关系数据库系统中存在，两种模式各有优缺点， 但在数据仓库中，通常不推荐“雪花化”,查询性能相对OLAP 系统来说更加被重 视，而雪花模式会降低数据仓库系统的性能。

ETL 负责将分布的、异构数据源中的数据如关系数据、平面数据文件等抽取

**第三章** **商业智能在电子商务的解决方** **案**

到临时中间层后进行清洗、转换、集成，最后加载到数据仓库或数据集市中，成为联 机分析处理、数据挖掘的基础。

存储在各种数据源中的数据通过 ETL 存储到数据仓库中，对数据仓库中的数 据进行多维建模，再利用多维分析技术满足决策支持或者满足在多维环境下特定 的查询和报表需求。多维分析技术的核心是“维”的概念，维指的是分析问题和观 察事物的角度。

在数据仓库或多维分析的基础上，对数据进行更深层次的分析和知识提取，需 要应用到数据挖掘技术。数据挖掘的价值体现于发现深度隐藏的规律或模式，数 据挖掘在电子商务行业有广阔的运用空间，如精细化营销、市场评估与预测、客户 分析(客户分群、流失预警、忠诚度分析等)、个性化推荐、网站异常预警与监控等。

BI 在电子商务的应用包括对内的运营决策支持和对外的网站用户支持。电 子商务对内决策分析系统为分析层、决策层提供决策分析平台，为操作层以及各业 务系统提供执行优化的智能帮助，如企业级 KPI 决策系统、部门级绩效考核系统、 业务监控和分析以及分析型工作流程。BI 对于网站用户的支持主要体现在个性 化服务、异常用户挖掘、网站效果分析与优化和提供数据分析增值与服务。

选择合适的软硬件产品是支持企业实施 BI的关键问题。BI 产品种类繁多，每 家厂商几乎都会宣称自己的产品是如何的好，在构建企业自身的 BI 系统时应该根 据企业自身的需求考虑软件硬件产品的选择。数据库方面 DB2、Oracle 都是很好 的产品，Teradata 和 EMC 公司的 Greenplum 在大数据处理方面有着很强的能力。 根据 Gantner 在2012年2月发布的数据仓库魔力象限显示，Teradata 在数据仓库 方面处于领导者地位。主流的 OLAP 工具有 BO 、Cognos 以及 Qilikview 等， Gartner 前端报表工具最具魅力象限连续几年纳入了Qlikview 这个产品，这个产品 代表了基于内存的报表分析技术，美观简单、用户体验极佳。数据挖掘产品领域， 有 SAS 、SPSS等两大厂家，而像 IBM 、Teradata 也都有自己的挖掘工具。做好数 据挖掘仅仅有工具是不够的，更需要数据挖掘人员对数据挖掘的理解和认识。

51

BI的实施策略，有代表性的是数据仓库支持 OLAP 理念的两位创始人 Bill Inmon 和 Ralph Kimball,Bill Inmon倡导自上而下实施策略，而 Ralph Kimball 则 提出不同看法，认为数据仓库应该实施自下而上的实施策略。笔者认为可以采用 第三种策略，那就是整体规划、分步实施，这种方式可以融合两种策略的优点，又可 以避免其中的短处。

BI 实施风险包含技术、业务、实施、推广等方面，但BI 最大风险还是来自于很 多人认为 BI 可有可无，有则应该更好，无也没关系的这样一个定位。笔者总结了 BI 实施风险主要有组织风险、技术风险、业务风险、实施风险。