**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**

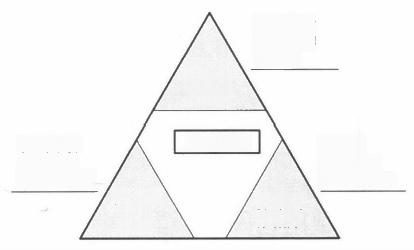
**1)数据仓库的设计思路**

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

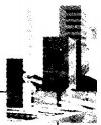
大数据

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**

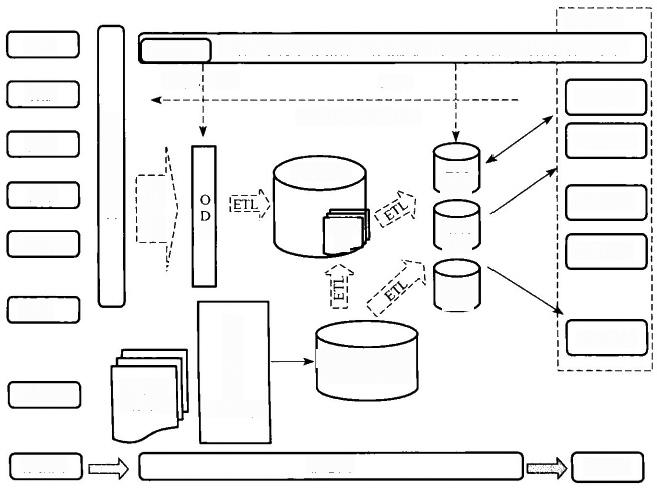
**第** **三** **章** **商** **业** **智** **能** **在** **电** **子** **商** **务** **的** **解** **决** **方** **案**

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。

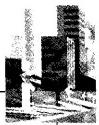
通常 ODS 可以包括下面三层。

· 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**

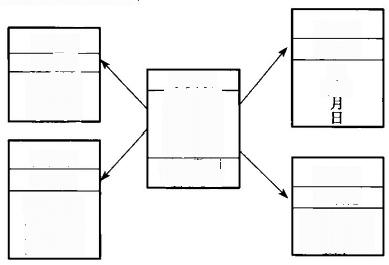
目前常用的数据仓库数据模型为多维数据模型，这种模型主要以星形模式和 雪花形模式在关系数据库系统中存在。

(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：

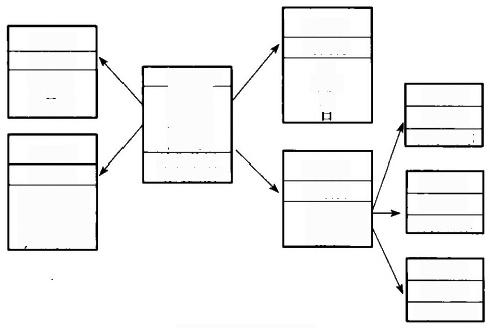
· 提高查询的效率。采用星形模式设计的数据仓库的优点是由于数据的组 织已经过预处理，主要数据都在庞大的事实表中，所以只要扫描事实表就可以进行 查询，而不必把多个庞大的表连接起来，查询访问效率较高。同时由于维表一般都 很小，甚至可以放在高速缓存中，所以与事实表作连接时其速度较快。

· 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

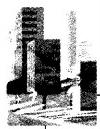
代理处

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。

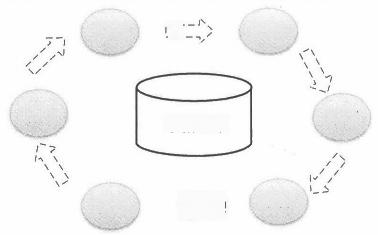
**3)数据仓库的模型设计**

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。