数据仓库基础知识讲解

讲师：曾力
天善社区： http://www.flybi.net/
天善学院： http://school.flybi.net/
天善官网： www.tianshansoft.com/

天善智能
TIANSHAN SOFT
专注商业智能和数据库性能优化

- **数据仓库基础知识讲解 主要授课内容**

1.数据仓库设计基本思想
2.ETL设计基本思想
3.数据建模:PowerDesigner
4.课程表结构介绍
5.存储过程基础讲解

- 数据仓库设计基本思想

数据仓库的历史发展。

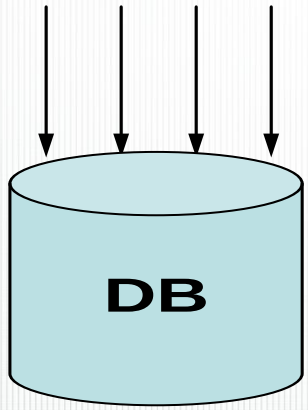
数据仓库的特点
(面向主题，随时间变化而变化，数据集成，信息相对稳定)

数据仓库的技术要求
(数据库模型设计、ETL设计,存储管理,报表设计)

常见数据仓库体系的架构
相关数据仓库的概念
(元数据、商业智能、数据挖掘、联机处理分析(OLAP)、
维度、度量、聚合汇总、钻取、旋转、切片、切块）。
雪花模型、星型模型两种常用的设计理念。

联机事务处理系统

联机事务处理系统（On-line Transaction Processing）OLTP系统：也称为生产系统，它是事件驱动、面向需求的，比如银行的储蓄系统就是一个典型的OLTP系统。OLTP在使用过程中积累了大量的数据。关系数据库概念提出之后，联机事务处理一直是数据库应用的主流。



OLTP系统

OLTP的特点：

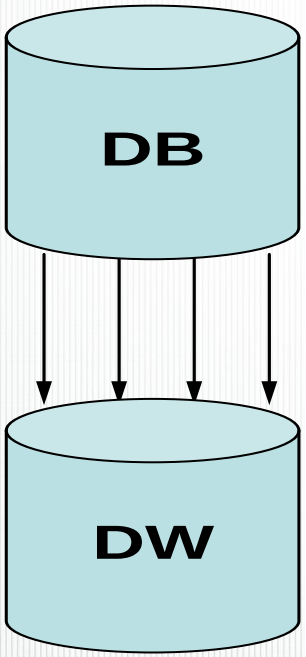
·对响应时间要求非常高;

·用户数量非常庞大,主要是操作人员;

·数据库的各种操作基于索引进行。

联机分析处理系统

联机分析处理系统（On-line Analytical Processing）OLAP系统：是基于数据仓库的信息分析处理过程，是数据仓库的用户接口部分，它是数据驱动、面向分析的。OLAP系统是跨部门、面向主题的 。



OLAP系统

OLAP的特点：

·基础数据来源于生产系统的操作数据;

·对系统的相应时间合理;

·用户数量相对较小，其用户主要是业务决策人员与管理人员。

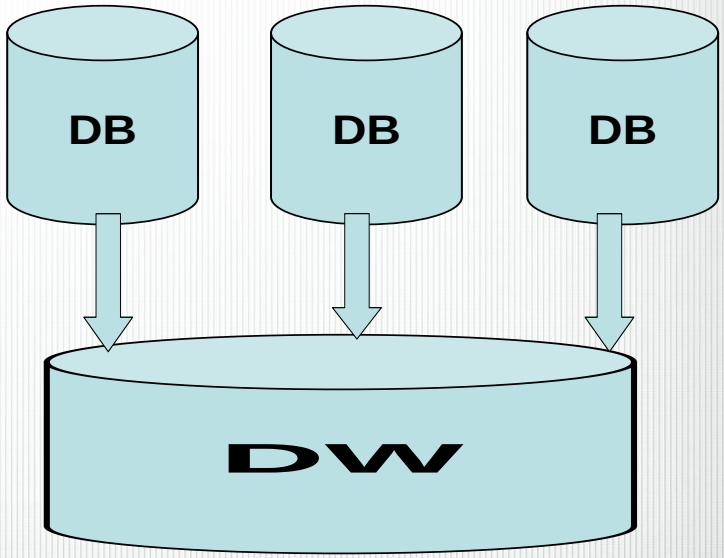
建立数据仓库的基本条件

建立数据仓库的基本条件：

第一：该行业有较为成熟的联机事务处理系统，它为数据仓库提供业务分析的客观条件和数据来源；

第二：该行业面临市场竞争的压力，需要通过数据分析服务业务，它为数据仓库的建立提供外在的动力；

第三：该行业为数据密集型行业，比如金融，供应链等；



数据仓库的特点 --面向主题

数据仓库的概念由被誉为“数据仓库之父”的WilliamH.Inmon博士提出的：**数据仓库是一个面向主题的、集成的、随时间变化的、信息相对、稳定的数据集合，**它用于对企业管理和决策提供支持。

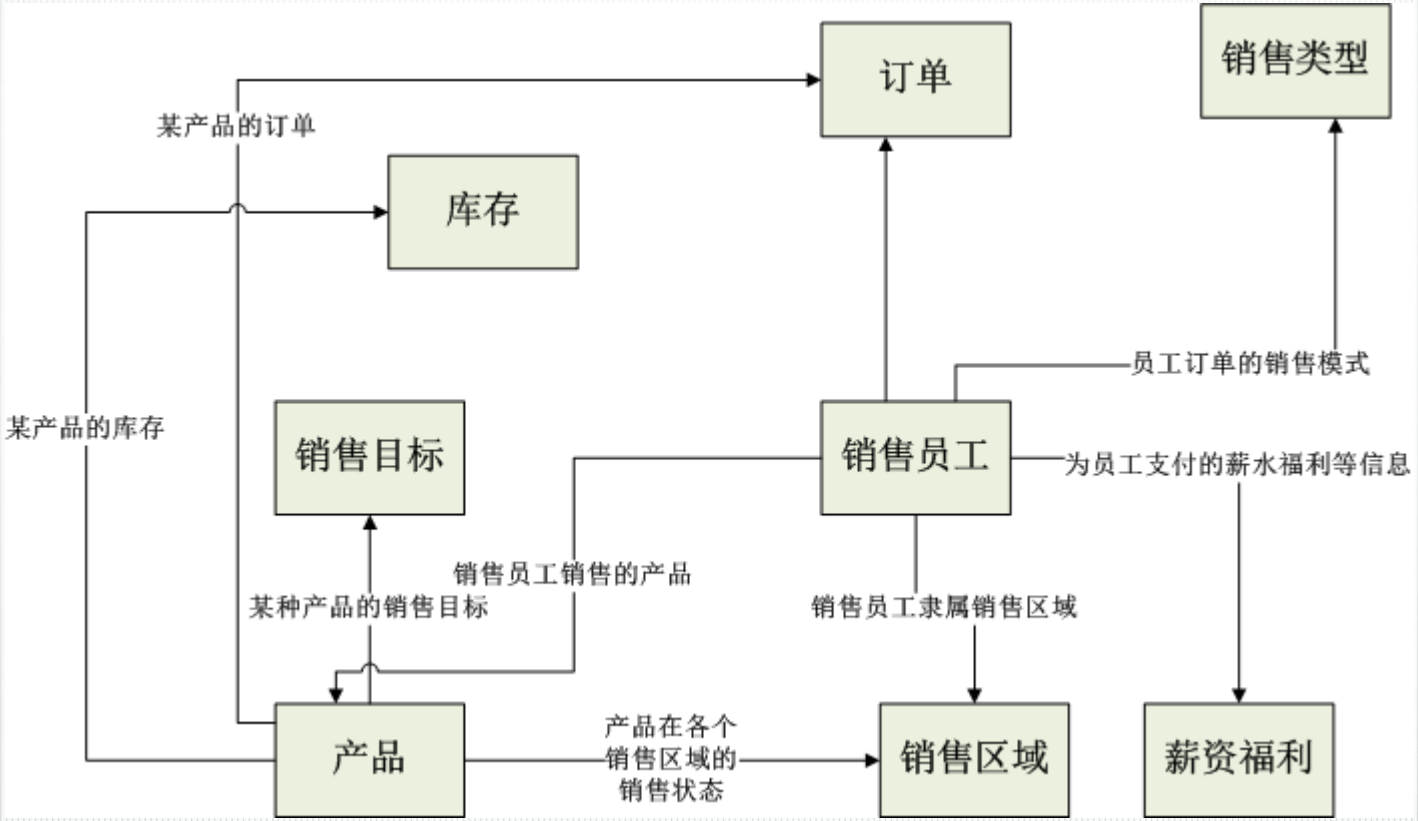
所谓主题：是指用户使用**数据仓库进行决策时所关心的重点方面，**如：客户、产品、账务、事件、服务使用、资源、客户服务、地域等；所谓面向主题，**是指数据仓库内的信息是按主题进行组织的，而不是像业务支撑系统那样是按照业务功能进行组织的；分析和明确企业所涵盖的业务范围，并且对企业业务进行高度概括性的描述，把密切相关业务对象进行归类，**它没有统一的标准，主要根据设计者的经验。不同的行业会有不同的主题域划分方式。

数据集成

数据集成：是指数据仓库中的信息不是从各个业务系统中简单抽取出来的，而是经过一系列加工、整理和汇总的过程，必须消除源数据中的不一致性，因此数据仓库中的信息是关于整个企业的一致的全局信息；

各个业务系统可能由不同的厂家独立承建，它们的数据模型设计、编码规则等都是不同的，这些数据加载到数据仓库之后，需要进行一个加工转换的过程。BOSS系统中，那地市的编码来说，CRM系统是的编码为1、2等，而BILLING系统可能根据长途区号来编码：451、452等，那么在数据仓库中，需要将各个业务系统中相同含义的数据通过规则映射为同一个编码。

数据仓库的特点 --面向主题



数据仓库的特点 --数据集成

CRM系统	
地市代码	地市名称
1	哈尔滨
2	齐齐哈尔
3	大庆
4	黑河
5	大兴安岭

Billing系统	
地市代码	地市名称
451	哈尔滨
452	齐齐哈尔
455	大庆
459	黑河
455	大兴安岭

映射规则1

映射规则2

数据仓库	
地市代码	地市名称
01	哈尔滨
02	齐齐哈尔
03	大庆
04	黑河
05	大兴安岭

随时间变化而变化

- 随时间变化：是指数据仓库内的信息并不只是反映企业当前的信息，而是记录了从过去某一时点到当前各个阶段的信息。通过这些信息，可以对企业的发展历程和未来趋势做出定量分析和预测；业务系统只记录当前的最新状态，数据仓库中可以反映一个用户的状态变化过程以及分析变化的原因。

某个用户的用户状态变化过程

数据仓库的特点 --数据相对稳定

信息相对稳定：是指一旦某个数据进入数据仓库以后，一般很少进行修改，更多的是对信息进行查询操作，通常只需要进行定期的加载和刷新。

数据仓库中几乎很少对历史数据进行修改，6月2日用户单停，那么在天这个粒度上的的数据就是本日最终停单的这个状态；而对于业务系统中，它总是最新的状态，所以业务系统的中的数据总是不断变化的。

(若出现业务审核情况造成的数据有效性加载，或者业务在审核过后指标才会出现，这种情况通常需要和客户进行确认，指标到底是计算在制单日还是审核日，我们应该尽量避免对进入数据仓库的数据进行UPDATE操作,同时再这种设计的时候,我们也要想好可能的回退方案。)

数据仓库构建思想

构造数据仓库有两种方式：**一是自上而下，一是自下而上。**

Bill Inmon先生推崇“自上而下”的方式，**即一个企业建立唯一的数据中心，就像一个数据的仓库，其中数据是经过整合、经过清洗、去掉脏数据的、标准的，能够提供统一的视图。**要建立这样的数据仓库，并不从它需要支持那些应用入手，而是要从整个企业的环境入手，分析其中的概念，应该有什么样的数据，达成概念完整性；（会考虑到很全面的设计）

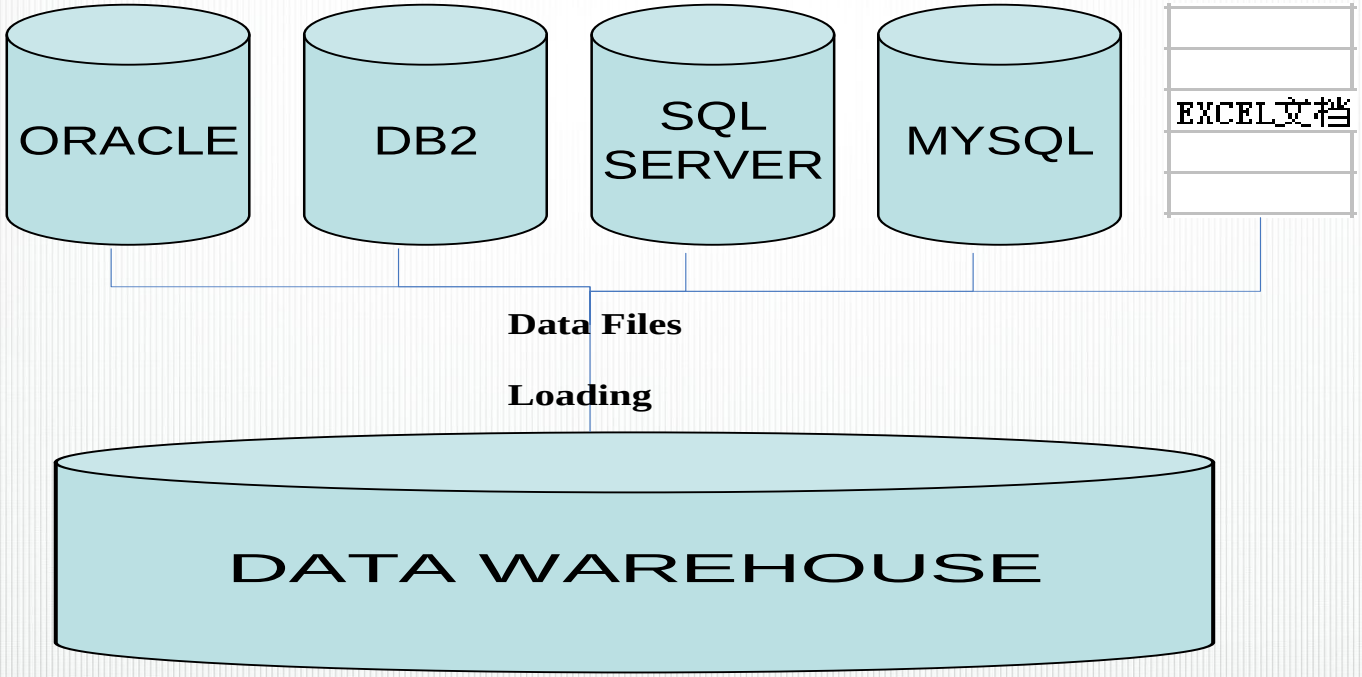
Ralph Kimbal先生推崇“自下而上”的方式，**他认为建设数据仓库应该按照实际的应用需求，加载需要的数据，不需要的数据不必要加载到数据仓库当中。**这种方式建设周期较短，客户能够很快看到结果。（针对客户的需求，需求要什么就做什么）

二者都要达到同一个目标：企业级数据仓库

实际上在建设数据仓库的时候，一般都参照这两种方式结合使用没有硬性规定

ETL(Extract/Transformation/Load)

用户从数据源抽取出所需的数据，经过数据清洗、转换,最终按照预先定义好的数据仓库模型，将数据加载到数据仓库中去；ETL是数据仓库系统中最重要的概念之一，ETL在一个数据仓库系统项目中要花一半以上的时间。



ETL(Extract/Transformation/Load)

ETL调度目标

数据来源：数据库、数据库文件、文本文件、程序生成
系统数目：单个系统/多个系统(过多的系统可以考虑接口实现)
数据库的类型：同种数据库/多种数据库

ETL调度参数设计

调度优先级/调度次序/中断标志/回滚标志/成功标志/调度开始结束时间等

ETL调度日志管理

文件记录/数据库记录
作业名称/作业执行开始-结束时间/作业执行结果/异常信息捕获/作业编号等

ETL调度JOB设计

数据文本文件加载/SQL在程序中调用/存储过程/ETL工具的WORKFLOW)

ETL(Extract/Transformation/Load)

ETL调度策略设计

全量数据加载

用户信息类数据, 状态会更新发生变化的数据

增量数据加载

流水分批**调度设计**

数据抽取一般在生产系统比较闲暇的时候进行，凌晨时候比较多，而且按照要分析数据的周期，还分为按日、按月数据；

由于涉及到的业务系统的数据量庞大，需要分批进行抽取，以及抽取数据后面的一系列处理过程；

调度并发设计

JOB并发设计、并发冲突设计、异常处理方式、成功/错误退出方式

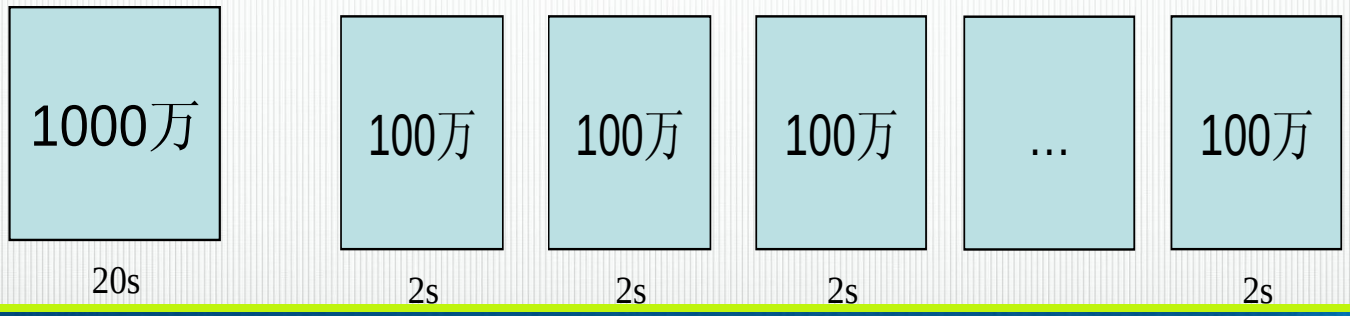
存储管理和模型设计

数据仓库的真正关键是数据的存储和管理。数据仓一般遇到的几个问题：

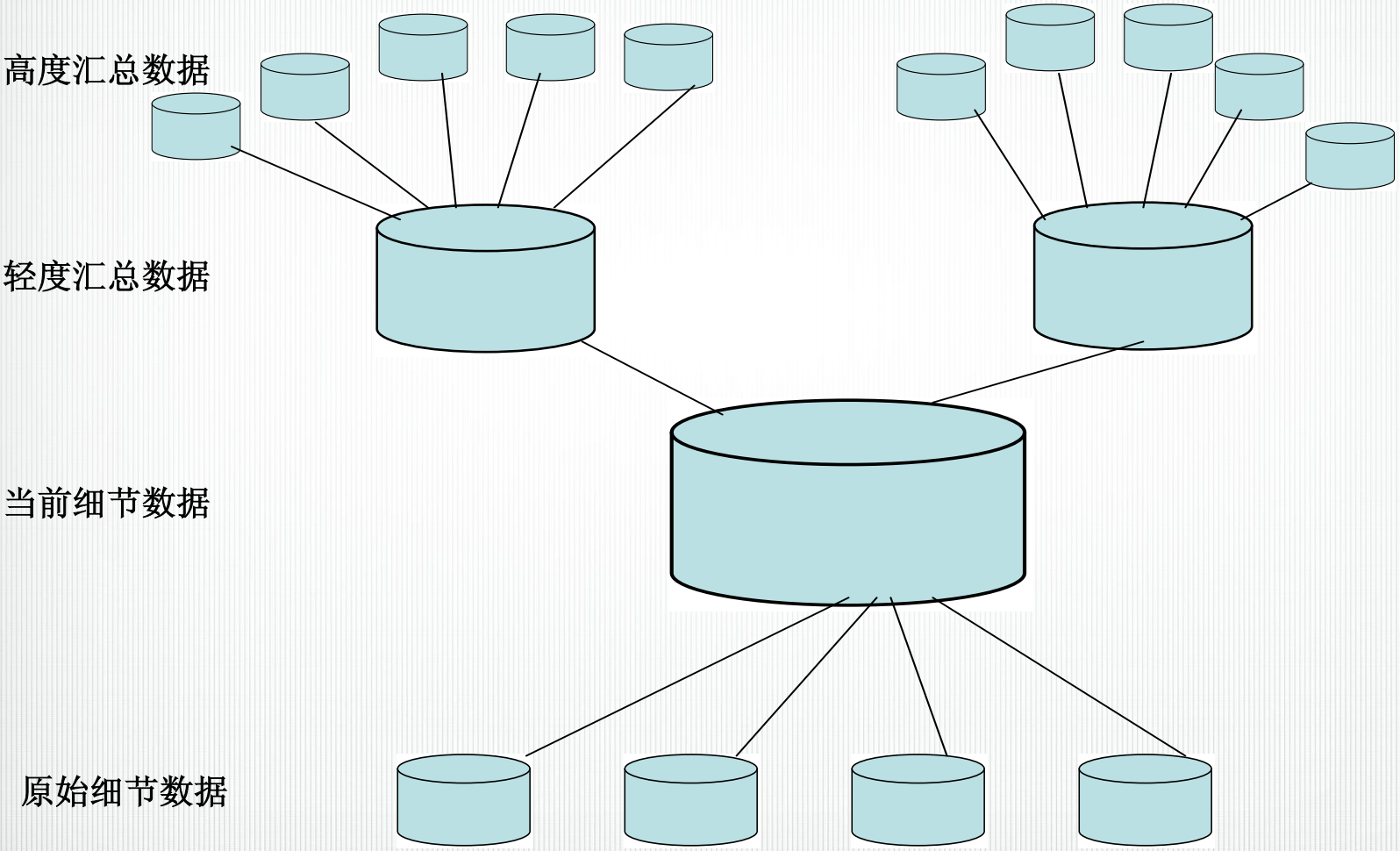
1. 大数据量的存储和管理 数据库的设计，安装、集成、根据数据抽取的详细要求设计数据库应用方案等；

2. 针对决策支持查询的优化 分区表、索引、簇集索引、MQT、SQL优化等方式来；

3. 支持多维分析的查询方式 是否有相关报表软件及查询方式的优化；



数据仓库简单的结构



数据仓库层次描述

	STAGE层	ODS层	MID层	MRT层
作用	提供业务系统数据文件的临时存储。	提供业务系统细节数据的长期沉淀；为未来分析类需求的扩展提供历史数据支撑；支撑中间汇总层数据生成	支撑DM层数据生成；方便应用需求处理，提高性能；支撑专题分析和数据挖掘	面向分析类应用所构建的数据存储；为报表、KPI、OLAP和指标体系等应用提供数据支撑
数据模型	与业务系统一致	3NF,与企业级数据模型一致	介于DM与DW之间，反范式设计，增加数据冗余	多维模型
数据存储粒度	存储业务系统数据的原始粒度	存储详单、客户资料等细节数据的原始粒度；经过转换处理后的数据	对用户等数据的轻度加工	中度、高度汇总数据
数据周期	临时性	长期保留，详单类可考虑6个月左右	长期保留	原则上保留所有数据

数据仓库的概念--元数据

按照传统的定义，元数据（Metadata）是关于数据的数据。在数据仓库系统中，元数据可以帮助数据仓库管理员和数据仓库的开发人员非常方便地找到他们所关心的数据；元数据是描述数据仓库内数据的结构和建立方法的数据，可将其按用途的不同分为两类：**技术元数据和业务元数据。**

技术元数据是存储关于数据仓库系统技术细节的数据，是用于开发和管理数据仓库

使用的数据，主要包括以下信息：

数据仓库结构的描述，包括仓库模式、视图、维、层次结构和导出数据的定义，以及数据集市的位置和内容；

业务系统、数据仓库和数据集市的体系结构和模式；

汇总用的算法，包括度量和维定义算法，数据粒度、主题领域、聚集、汇总、预定义的查询与报告；

由操作环境到数据仓库环境的映射，包括源数据和它们的内容、数据分割、数据取、清理、转换规则和数据刷新规则、安全（用户授权和存取控制）。

数据仓库的概念--联机处理分析(OLAP)

简写为OLAP,随着数据库技术的发展和应用，数据库存储的数据量从20世纪80年代的兆（M）字节及千兆（G）字节过渡到现在的兆兆（T）字节和千兆兆（P）字节，同时，用户的查询需求也越来越复杂，涉及的已不仅是查询或操纵一张关系表中的一条或几条记录，而且要对多张表中千万条记录的数据进行数据分析和信息综合，关系数据库系统已不能全部满足这一要求。在国外，不少软件厂商采取了发展其前端产品来弥补关系数据库管理系统支持的不足，力图统一分散的公共应用逻辑，在短时间内响应非数据处理专业人员的复杂查询要求.

数据仓库与OLAP的关系是互补的，现代OLAP系统一般以数据仓库作为基础，即从数据仓库中抽取详细数据的一个子集并经过必要的聚集存储到OLAP存储器中供前端分析工具读取。典型的OLAP系统体系结构如下图所示：

OLAP系统按照其存储器的数据存储格式可以分为关系OLAP（RelationalOLAP，简称ROLAP）、多维OLAP（MultidimensionalOLAP，简称MOLAP）和混合型OLAP（HybridOLAP，简称HOLAP）三种类型。

数据仓库的概念--联机处理分析(OLAP)

ROLAP

ROLAP将分析用的多维数据存储在关系数据库中并根据应用的需要有选择的定义一批实视图作为表也存储在关系数据库中。不必要将每一个**SQL**查询都作为实视图保存，只定义那些应用
频率比较高、计算工作量比较大的查询作为实视图。对每个针对**OLAP**服务器的查询，优先利用已经计算好的实视图来生成查询结果以提高查询效率。同时用作**ROLAP**存储器的**RDBMS**也针对**OLAP**作相应的优化，比如并行存储、并行查询、并行数据管理、基于成本的查询优化、位图索引、**SQL**的**OLAP**扩展**(cube,rollup)**等等。

MOLAP

MOLAP将**OLAP**分析所用到的多维数据物理上存储为多维数组的形式，形成“立方体”的结构。
维的属性值被映射成多维数组的下标值或下标的范围，而总结数据作为多维数组的值存储在数组的单元中。由于**MOLAP**采用了新的存储结构，从物理层实现起，因此又称为物理**OLAP**（**PhysicalOLAP**）；而**ROLAP**主要通过一些软件工具或中间软件实现，物理层仍采用关系数据
库的存储结构，因此称为虚拟**OLAP**（**VirtualOLAP**）。

HOLAP

由于**MOLAP**和**ROLAP**有着各自的优点和缺点（如下表所示）,且它们的结构迥然不同，这给分析人员设计**OLAP**结构提出了难题。为此一个新的**OLAP**结构——混合型**OLAP**（**HOLAP**）被提出，它能把**MOLAP**和**ROLAP**两种结构的优点结合起来。迄今为止，对**HOLAP**还没有一个正式的
定义。但很明显，**HOLAP**结构不应该是**MOLAP**与**ROLAP**结构的简单组合，而是这两种结构

数据仓库的概念--维度

管理人员往往希望从不同的角度来审视业务情况，比如从时间、地域、产品、客户等来看收入、利润、支出等业务统计数字。每一个分析的角度可以叫做一个维，因此，我们把多角度分析方式称为多维分析。以前，每一个分析的角度需要制作一张报表。在线多维分析工具的主要功能，是根据用户常用的多种分、析角度，事先计算好一些辅助结构，以便在查询时能尽快访问到所要的汇总数字，并快速地从一维转变到另一维，将不同角度的信息以数字、直方图、饼图、曲线等等方式展现在用户面前。

图 16-11 直观地表示了贷款分析模型所能实现的所有的分析角度(维度)和层次(粒度)：

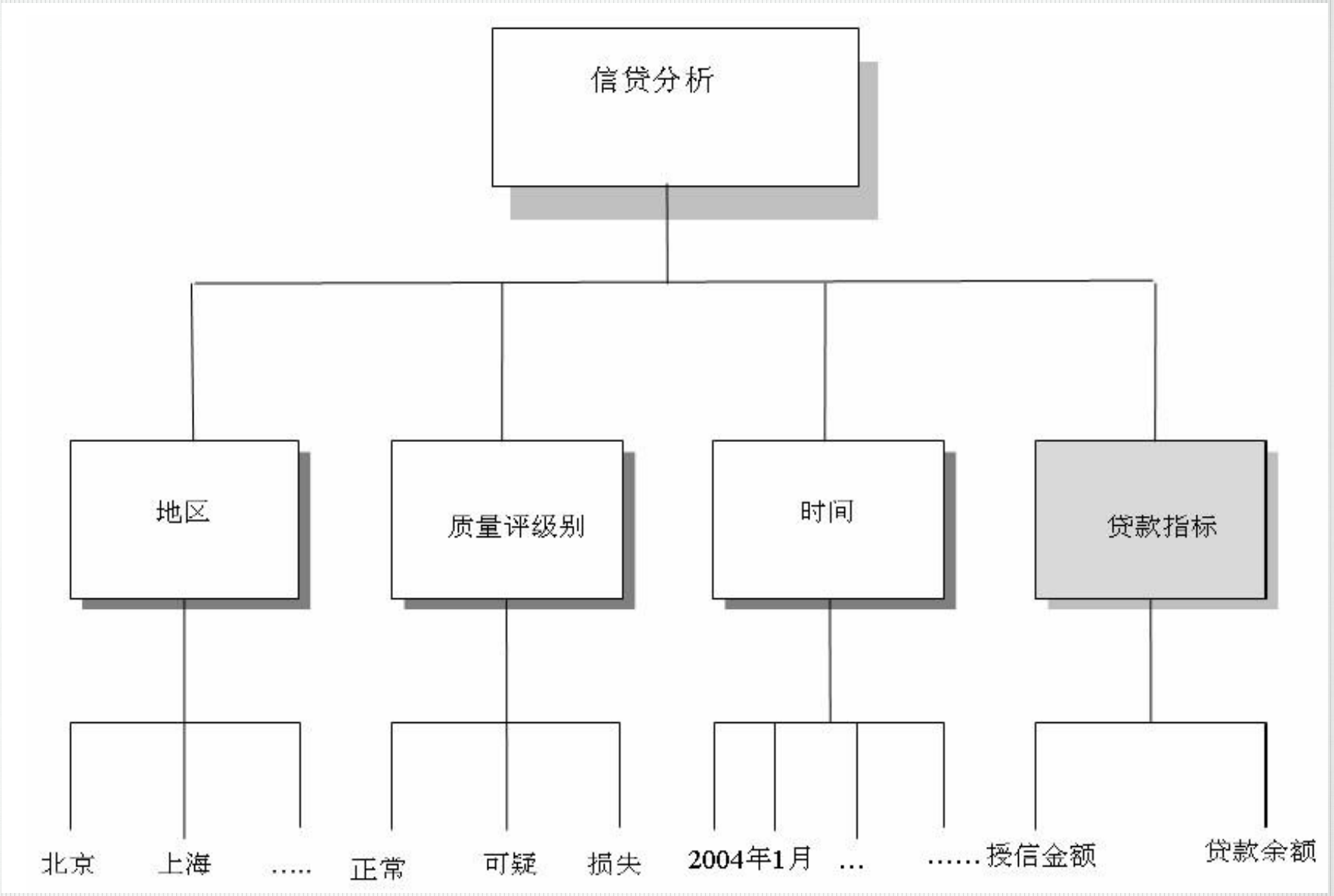
维度 →

粒度 ↓

时间	贷款银行	区域	贷款质量
年	商业银行总行	省	正常/不良
季度	省级分行	市	五级分类
月	市分行		

度量指标(事实)：授信金额、贷款余额

图 16-11 贷款分析的角度和层次



数据仓库的概念--切片/切块/钻取/旋转/转轴

切片和切块(Slice and Dice)

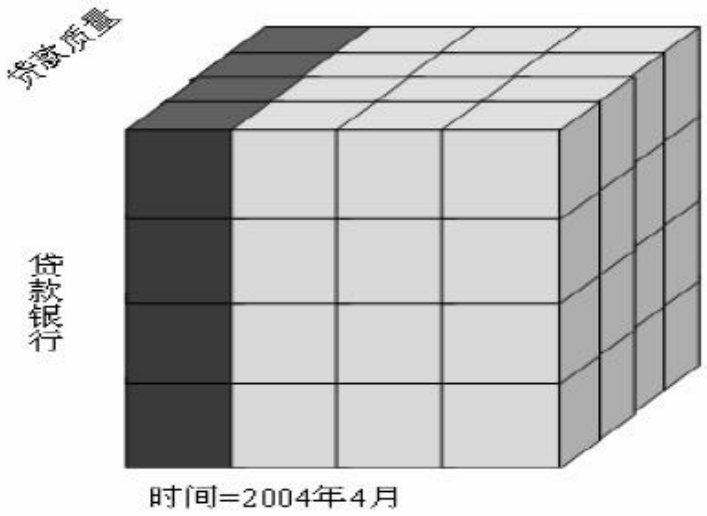
在多维数据结构中，按二维进行切片，按三维进行切块，可得到所需要的数据。每次都是沿其中一维进行分割称为分片，每次沿多维进行的分片称为分块。

钻取(Drill)

钻取包含向下钻取(Drill-down)和向上钻取(Drill-up)， 钻取的深度与维所划分的层次相对应。

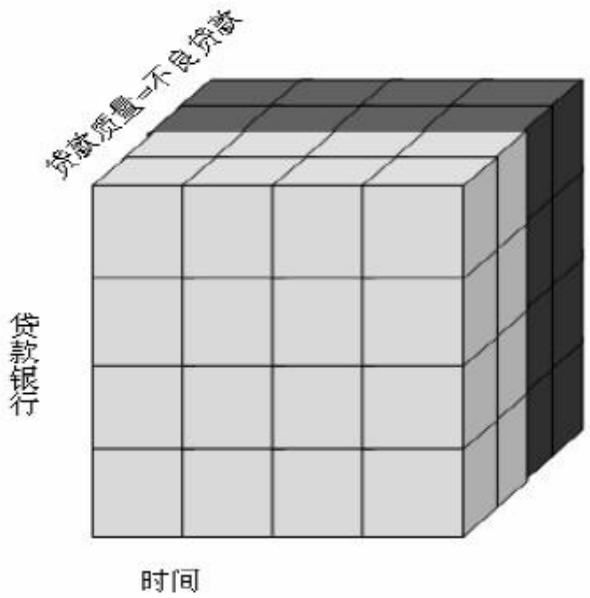
旋转(Rotate)/转轴(Pivot)

通过旋转可以得到不同视角的数据。



时间	授信银行	发放行级别	贷款	贷款质量	贷款金额
2004年04月	工行吉林吉林分行	地市行	正常	正常贷款	273684.00
			关注	正常贷款	114924.00
			次级	不良贷款	80930.80
			可疑	不良贷款	73231.00
	工行吉林四平分行	地市行	正常	正常贷款	81309.00
	工行吉林辽源分行	地市行	正常	正常贷款	82573.48
			可疑	不良贷款	14450.00
			损失	不良贷款	17261.00
	农行吉林分行	地市行	正常	正常贷款	95106.50
			损失	不良贷款	42574.00
	农行吉林吉林分行	地市行	正常	正常贷款	11260.00
	农行吉林辽源分行	地市行	正常	正常贷款	17080.00
			关注	正常贷款	10120.00
	中行吉林吉林分行	地市行	正常	正常贷款	57391.93
			关注	正常贷款	90815.89
			次级	不良贷款	18600.00
			可疑	不良贷款	64980.44
			损失	不良贷款	7771.00
	中行吉林四平分行	地市行	关注	正常贷款	16865.00
	中行吉林辽源分行	地市行	关注	正常贷款	26494.81
			损失	不良贷款	11094.87
	建行吉林吉林分行	地市行	正常	正常贷款	50000.00
			关注	正常贷款	20000.00
	建行吉林四平分行	地市行	正常	正常贷款	1000.00
			关注	正常贷款	126900.00
			次级	不良贷款	54300.00
	交行吉林吉林分行	地市行	正常	正常贷款	15750.00
	光大吉林分行	省行	正常	正常贷款	66570.00
			关注	正常贷款	42600.00
			可疑	不良贷款	48003.44
	合计				1633641.15

图 16-12 切片一 2004 年 4 月份所有贷款情况



			时间	2004年04月
贷款质量	授信银行	发放行级别	贷款	贷款金额
不良贷款	工行吉林吉林分行	地市行	次级	80930.80
			可疑	73231.00
	工行吉林辽源分行	地市行	可疑	14450.00
			损失	17261.00
	农行吉林分行	地市行	损失	42574.00
	中行吉林吉林分行	地市行	次级	18600.00
			可疑	64980.44
			损失	7771.00
	中行吉林辽源分行	地市行	损失	11094.87
	建行吉林四平分行	地市行	次级	54300.00
	光大吉林分行	省行	可疑	48003.44
	合计			433196.55

图 16-13 切片二 所有不良贷款情况

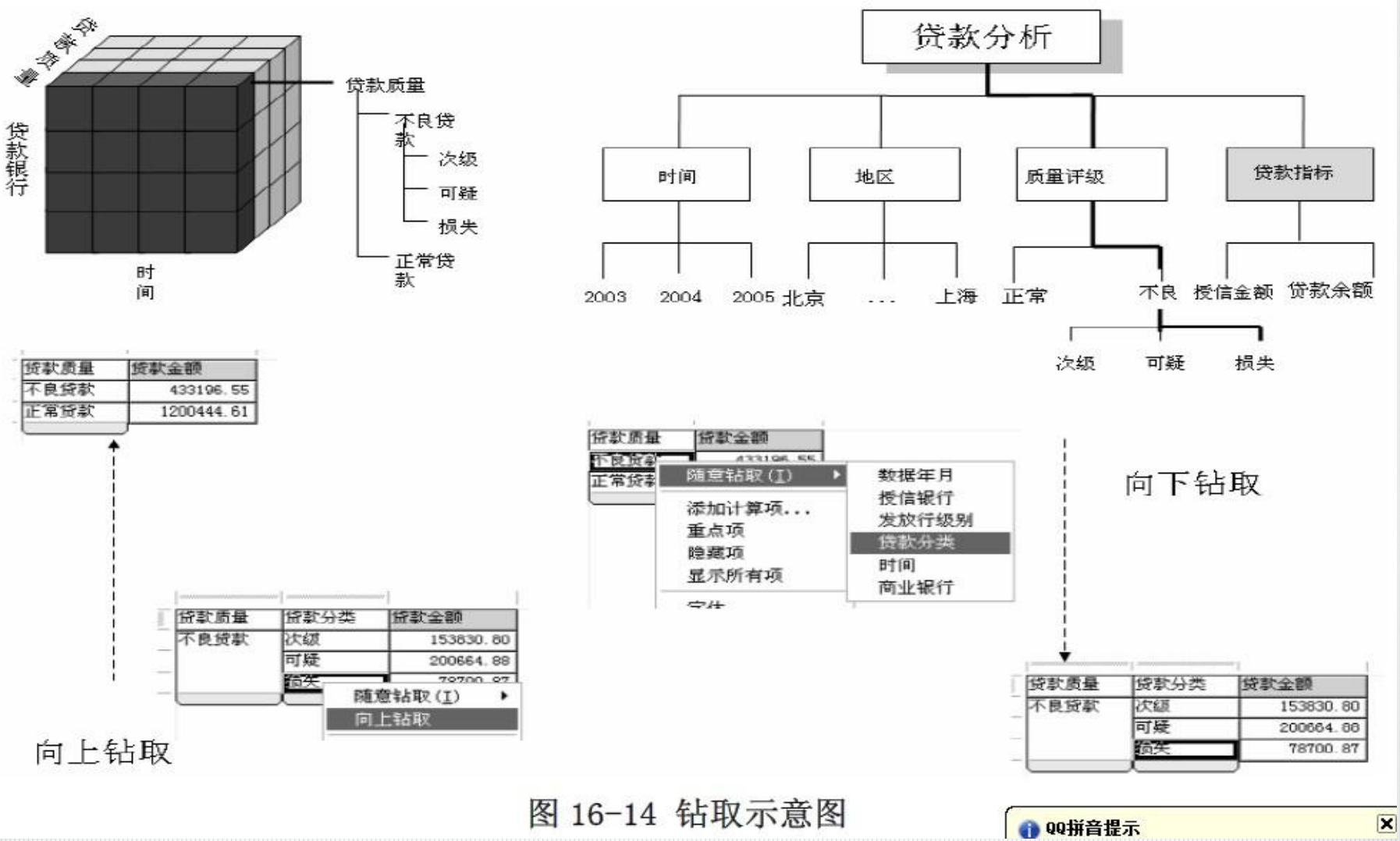
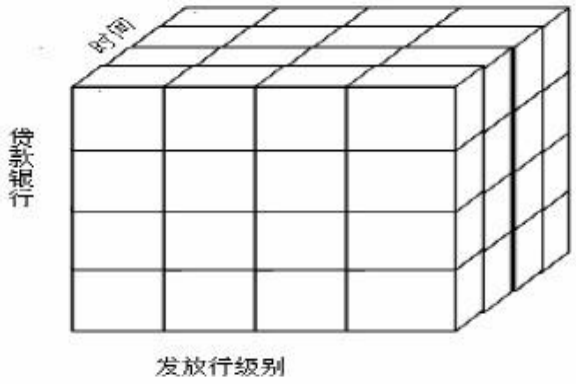
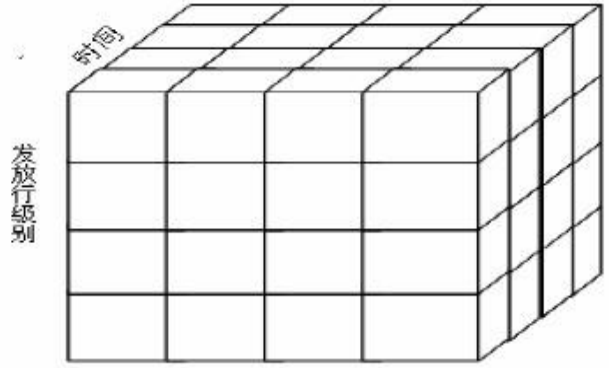


图 16-14 钻取示意图



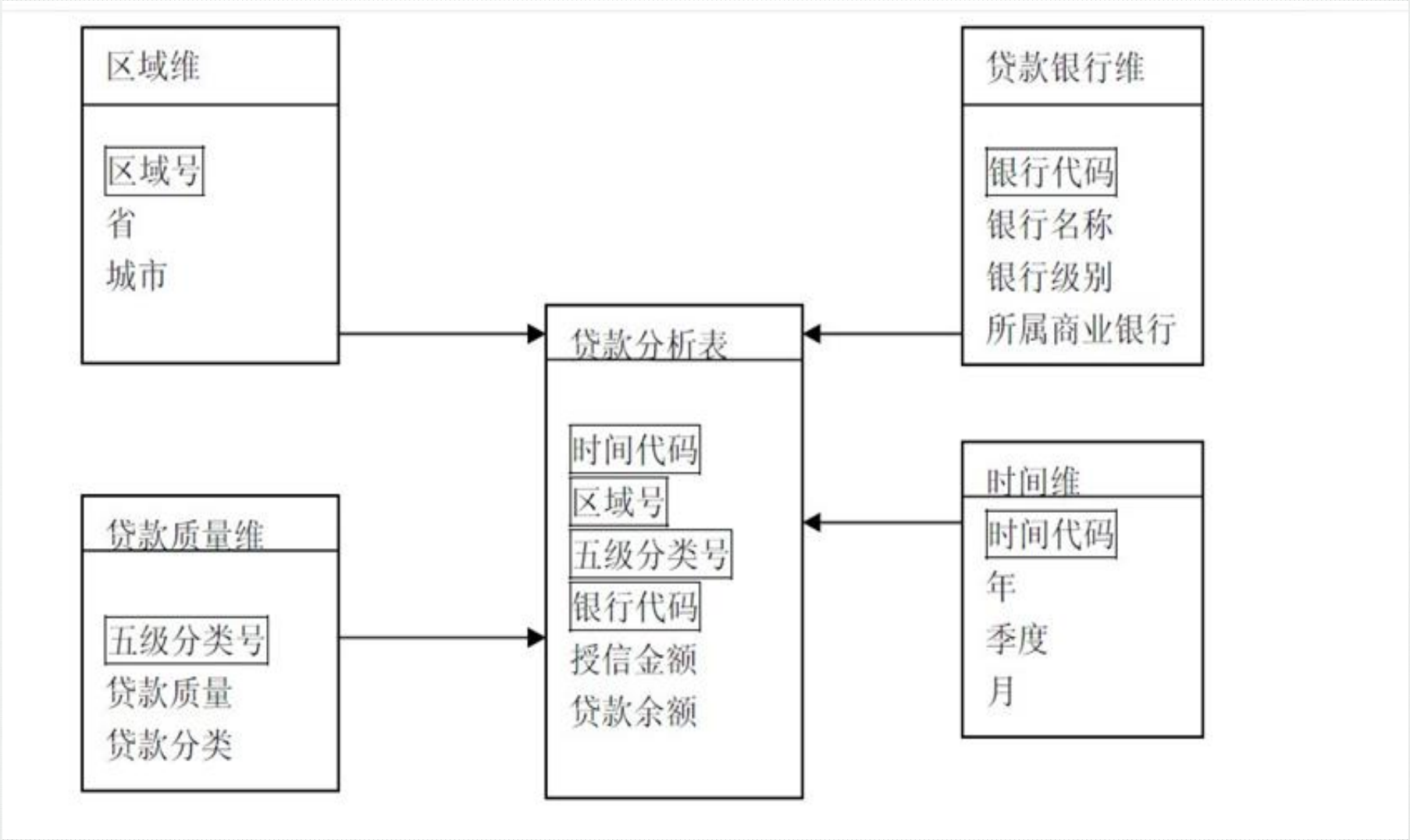
	时间	2004年04月
授信银行	发放行级别	贷款金额
工行吉林吉林分行	地市行	542769.80
工行吉林四平分行	地市行	81309.00
工行吉林辽源分行	地市行	114284.48
农行吉林分行	地市行	137680.50
农行吉林吉林分行	地市行	11260.00
农行吉林辽源分行	地市行	27200.00
中行吉林吉林分行	地市行	239559.28
中行吉林四平分行	地市行	16865.00
中行吉林辽源分行	地市行	37589.68
建行吉林吉林分行	地市行	70000.00
建行吉林四平分行	地市行	182200.00
交行吉林吉林分行	地市行	15750.00
光大吉林分行	省行	157173.44
合计		1633641.16

旋转



	时间	2004年04月
发放行级别	授信银行	贷款金额
地市行	工行吉林吉林分行	542769.80
	工行吉林四平分行	81309.00
	工行吉林辽源分行	114284.48
	农行吉林分行	137680.50
	农行吉林吉林分行	11260.00
	农行吉林辽源分行	27200.00
	中行吉林吉林分行	239559.28
	中行吉林四平分行	16865.00
	中行吉林辽源分行	37589.68
	建行吉林吉林分行	70000.00
	建行吉林四平分行	182200.00
	交行吉林吉林分行	15750.00
	合计	1476467.72
省行	光大吉林分行	157173.44
	合计	157173.44

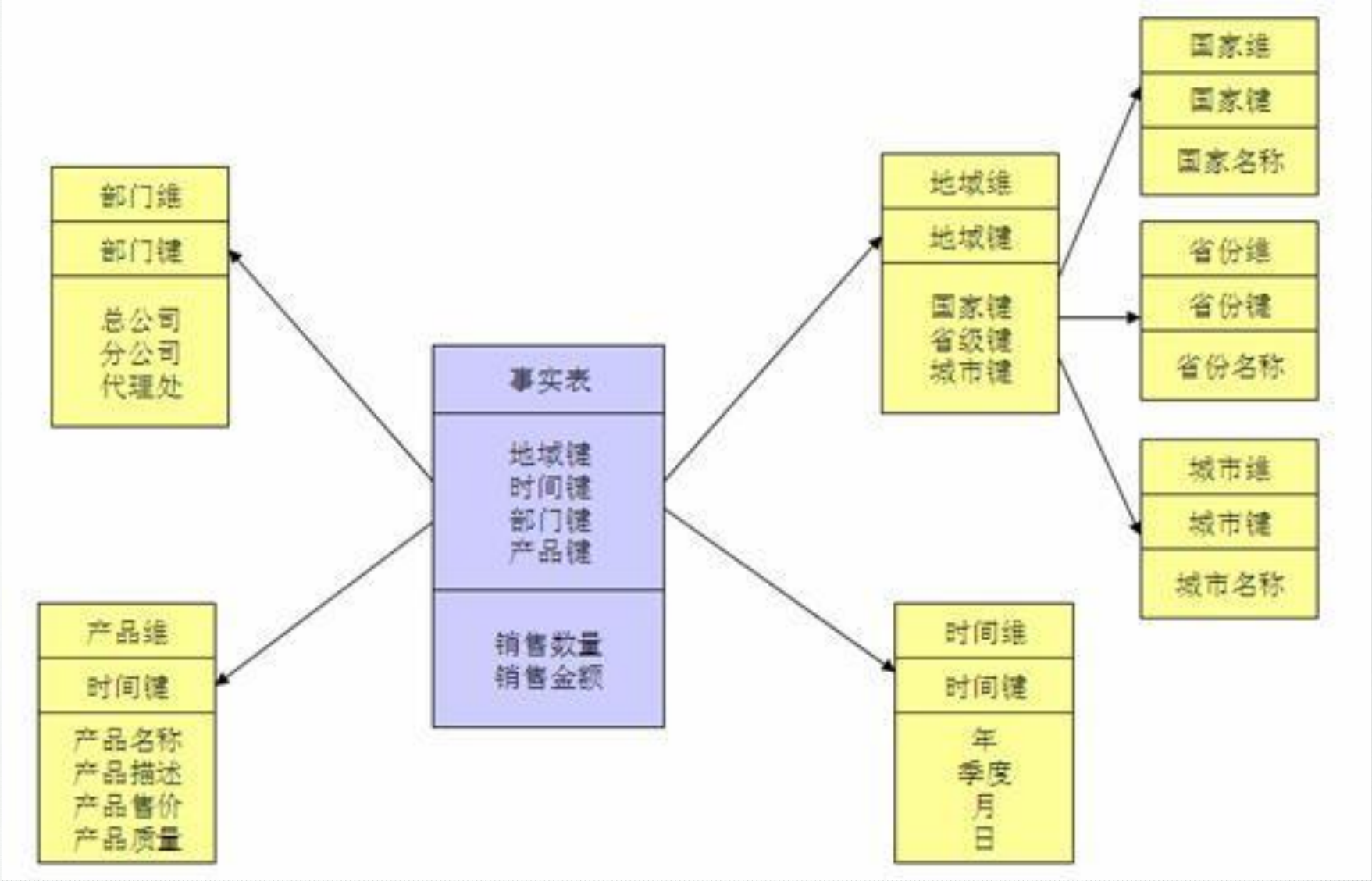
图 16-15 旋转示意图



雪花模式是对星形模式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。雪花模式的维度表是基于范式理论的，因此是界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。

雪花模式的优点是: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OLTP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。



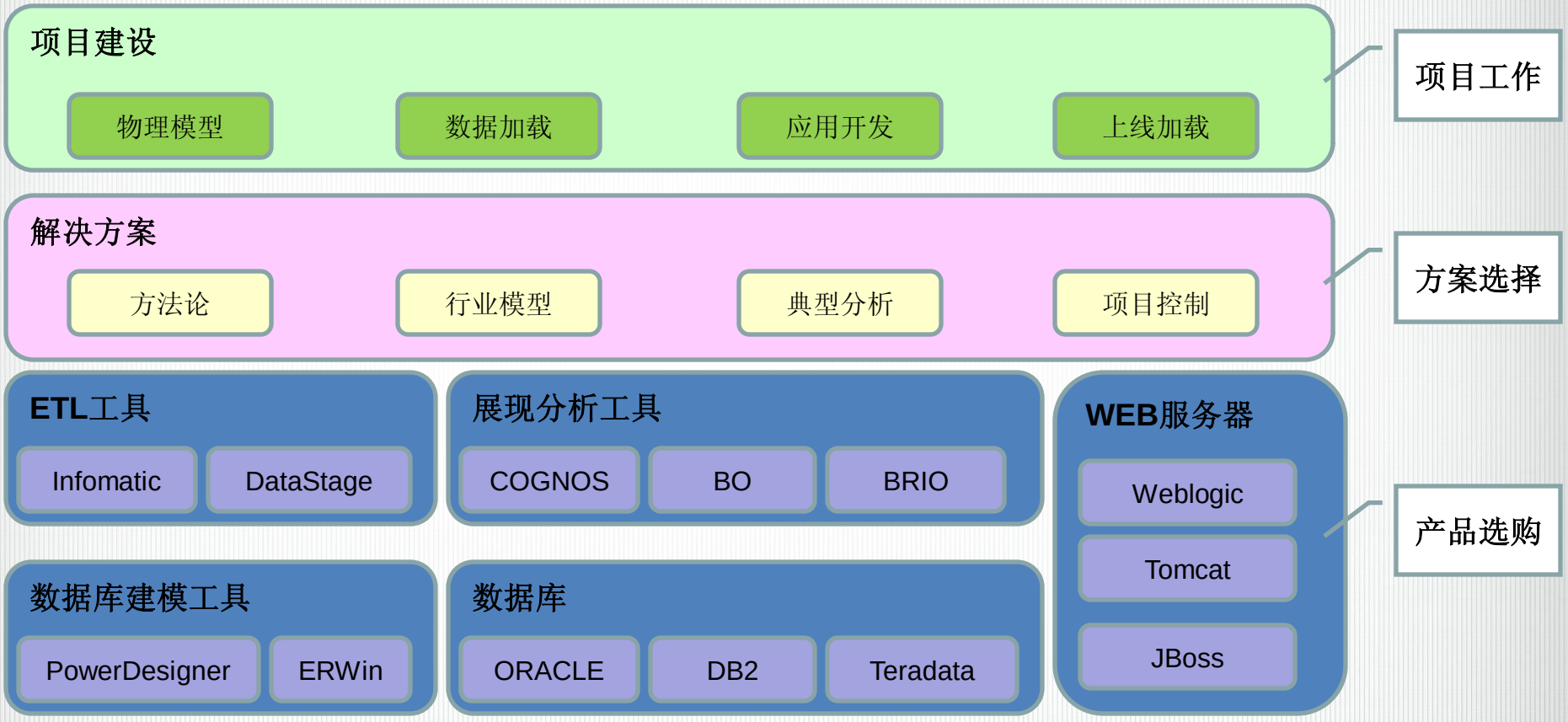
在不考虑缓慢变换的情况下
一般大多数事实表设计 如下:

id　dim1id dim2id dim3id dim4id ...measure1 measure2....

大多数维度表的设计 如下：
level1id level1name level2id levelname ..... levelnid(pk) levelname

PowerDesginer基本使用

- PowerDesginer选择物理模型数据库
- PowerDesigner创建表
- PowerDesigner创建主外键关系
- PowerDesigner创建索引
- PowerDesginer创建视图
- PowerDesginer数据库建表语句导出
- PowerDesginer反向工程


项目建设
物理模型
数据加载
应用开发
上线加载
项目工作
解决方案
方法论
行业模型
典型分析
项目控制
方案选择
ETL工具
Infomatic
DataStage
展现分析工具
COGNOS
BO
BRIO
WEB服务器
Weblogic
Tomcat
JBoss
产品选购
数据库建模工具
PowerDesigner
ERWin
数据库
ORACLE
DB2
Teradata

THANKS

天善智能

专注商业智能和数据库性能优化