数据架构及数据建模

2012. 09. 14

SACC2012

EN·CORE

Contents

1 数据发展战略

04 与数据相关的技术

- BI : Business Intelligence
- **Data Governance**
- Data Integration
- **Data Quality**
- Data Warehouse
- ETL
- **MDM (Master Data Management)**
- BPM (Business Process Management)
- Content Management
- CRM (Customer Relationship Management)
- CDI (Customer Data Integration)
- Customer Intelligence
- Data Acquisition, Replication
- Data Analysis
- **Data Architecture**
- **Data Management**
- Data Marts
- Data Migraiton
- Data Mining
- Data Modeling
- Data Profiling
- Data Visualization
- Database Marketing
- Database Application Performance
- Databases
- DW Design, Methodoloty
- EA (Enterprise Architecture)
- EAI (Enterprise Application Integration)
- **Metadata Management**
- Operational Data Store
- RTE (Real-Time Enterprise)
- SOA (Service-Oriented Architecture)

EN·CORE

05 数据趋势

Quantity

IDC estimates that the world will reach a zettabyte of data (1,000 exabytes or 1 million pedabytes) in 2010

Mearian, Lucas. "A zettabyte by 2010: Corporate data grows fiftyfold in three years." Computerworld, March 6, 2007.

Quality

Process failure and information scrap and rework caused by defective information costs the United States alone $1.5 trillion or more.

Gartner, Inc press release. "'Dirty Data' is a Business Problem, Not an IT Problem, Says Gartner," March 2, 2007.

Best-practice data quality programs are not a one-shot measure (clean up and move on)... To achieve such results, successful programs identify the organizational processes behind data quality. Much like regular IT housekeeping, from virus scanning or performance monitoring to data backup, the data quality program becomes part of daily IT routine.

English, Larry. "Plain English about Information Quality: Information Quality Tipping Point." DM Review, July 2007.

Over the next two years, more than 25 percent of critical data in Fortune 1000 companies will continue to be flawed, that is, the information will be inaccurate, incomplete or duplicated…

"Organizing for Data Quality." Research note from Gartner Inc., June 1, 2007.

06 数据趋势

EN·CORE

Governance

Data governance (DG) refers to the overall management of the availability, usability, integrity, and security of the data employed in an enterprise. A sound data governance program includes a governing body or council, a defined set of procedures, and a plan to execute those procedures.

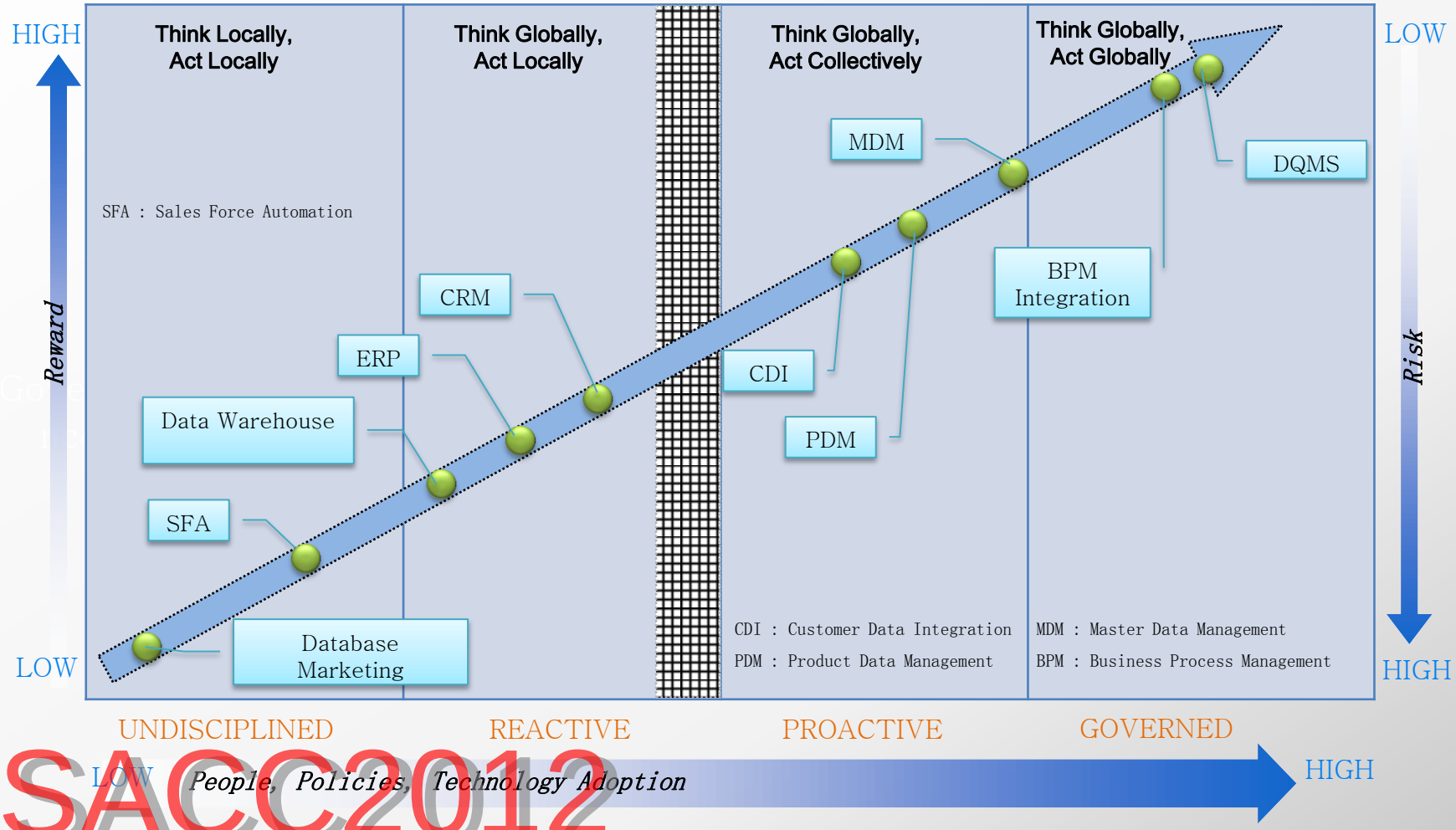
Many companies have difficulty keeping data consistent, synchronised and in a high quality state, Data needs to be managed in a robust way, so Data governance is needed.

Data governance processes can be automated using data services built using workflow and deployed on a data Management platform

Enterprise data governance requires systematic implementation of common processes via re-usable services on a data management platform.

Mike Ferguson "Accelerating Enterprise Data Governance" Intelligent Business Strategies. December 2007

07 Data Governance Maturity Model



'The Data Governance Maturity Model', 2007, DataFlux

2 数据架构

Enterprise Architecture Framework

详细内容

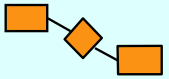
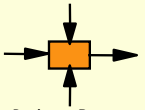
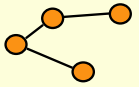
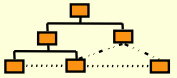
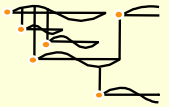
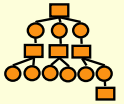
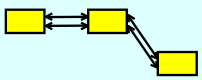
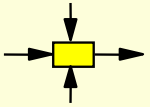
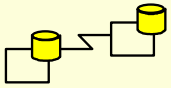
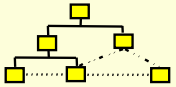
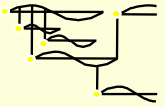
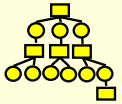
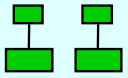
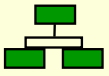
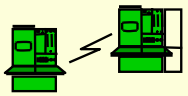
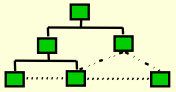
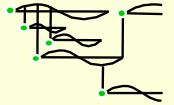
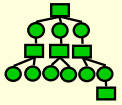
依据6W原则设计

继承关系

概念

逻辑

物理

	DATA *What*	FUNCTION *How*	NETWORK *Where*	PEOPLE *Who*	TIME *When*	MOTIVATION *Why*
SCOPE **CONTEXTUAL** ***Planner***	List of Things Important to the Business  ENTITY = Class of Business Thing	List of Processes the Business Performs  Function = Class of Business Process	List of Locations in which the Business Operates  Node = Major Business Location	List of Organizations Important to the Business  People = Major Organizations	List of Events Significant to the Business  Time = Major Business Event	List of Business Goals/Strat  Ends/Means=Major Bus.Goal/ Critical Success Factor
ENTERPRISE MODEL **CONCEPTUAL** ***Owner***	e.g. Semantic Model  Ent = Business Entity Reln = Business Relationship	e.g. Business Process Model  Proc. = Business Process I/O = Business Resources	e.g. Business Logistics System  Node = Business Location Link = Business Linkage	e.g. Work Flow Model  People = Organization Unit Work = Work Product	e.g. Master Schedule  Time = Business Event Cycle = Business Cycle	e.g. Business Plan  End = Business Objective Means = Business Strategy
SYSTEM MODEL **LOGICAL** ***Designer***	e.g. Logical Data Model  Ent = Data Entity Reln = Data Relationship	e.g. Application Architecture  Proc .= Application Function I/O = User Views	e.g. Distributed System Architecture  Node = I/S Function (Processor, Storage, etc) Link = Line Characteristics	e.g. Human Interface Architecture  People = Role Work = Deliverable	e.g. Processing Structure  Time = System Event Cycle = Processing Cycle	e.g. Business Rule Model  End = Structural Assertion Means =Action Assertion
TECHNOLOGY MODEL **PHYSICAL** ***Builder***	e.g. Physical Data Model  Ent = Segment/Table/etc. Reln = Pointer/Key/etc.	e.g. System Design  Proc.= Computer Function I/O = Data Elements/Sets	e.g. Technology Architecture  Node = Hardware/System Software Link = Line Specifications	e.g. Presentation Architecture  People = User Work = Screen Format	e.g. Control Structure  Time = Execute Cycle = Component Cycle	e.g. Rule Design  End = Condition Means = Action
DETAILED REPRESENTATIONS **OUT OF CONTEXT** ***Contractor***	e.g. Data Definition  Ent = Field Reln = Address	e.g. Program  Proc.= Language Stmt I/O = Control Block	e.g. Network Architecture  Node = Addresses Link = Protocols	e.g. Security Architecture  People = Identity Work = Job	e.g. Timing Definition  Time = Interrupt Cycle = Machine Cycle	e.g. Rule Specification  End = Sub-condition Means = Step
FUNCTIONING ENTERPRISE	**e.g. DATA**	**e.g. FUNCTION**	**e.g. NETWORK**	**e.g. ORGANIZATION**	**e.g. SCHEDULE**	**e.g. STRATEGY**

EN·CORE

数据架构流程

环境分析	数据库的使用结构，模型CASE TOOL, 环境问题、当前需求、 未来需求
架构定义	数据、数据结构、数据管理流程定义
架构原则定义	确保具有较好的扩张性、运维性、标准化、一致性、整合性，制定管理体系
参考模型选定	先进模型 (ERP, 同行业案例)
AS-IS架构设计	•逆向建模→ AS IS物理模型→标准化→AS IS逻辑化→AS IS概念化 •数据质量对象指定→质量指数制定→BR编写→数据质量测定
TO-BE架构设计	•TO BE概念模型→TO BE逻辑模型→TO BE物理模型 •错误分析→改善计划制定→改善→应用→目标修订
迁移计划设计	新一代系统构建计划或模型重构计划制定
管理系统构建	构建元数据管理系统，及数据质量管理体系(包括使用质量管理工具)
管理方案设计	设立数据架构管理部门，制定流程

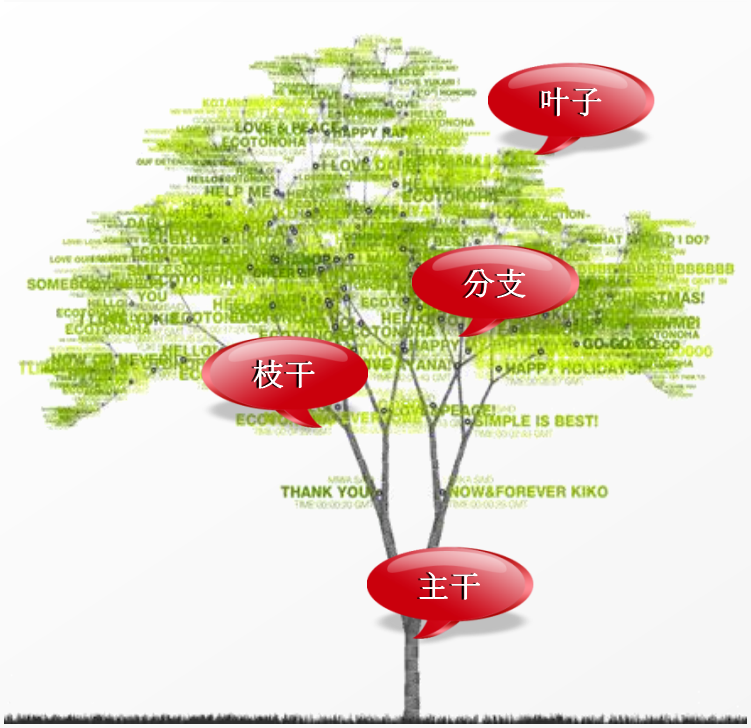
EN·CORE

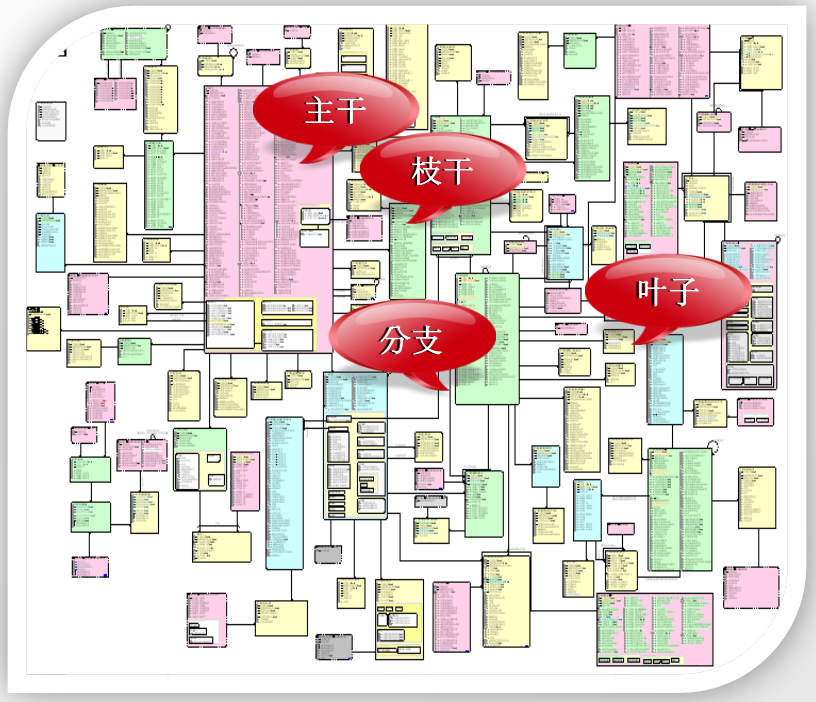
10 Enterprise Architecture Framework

	数据	数据架构	数据管理流程
EDA (继承关系)		主题域 继承关系模型	数据管理政策
DA (概念)	数据标准 (词汇,术语,域)	概念模型 数据集成架构	标准管理 需求管理
Modeler (逻辑)	模型数据 (模型申请，变更)	逻辑模型 物理模型	数据模型管理 数据流管理
DBA (物理)	管理数据 (数据流、备份、验证、使用)	数据库	数据库管理
User (运维)	业务数据 (实际数据)	用户视图	数据应用管理

3 数据模型

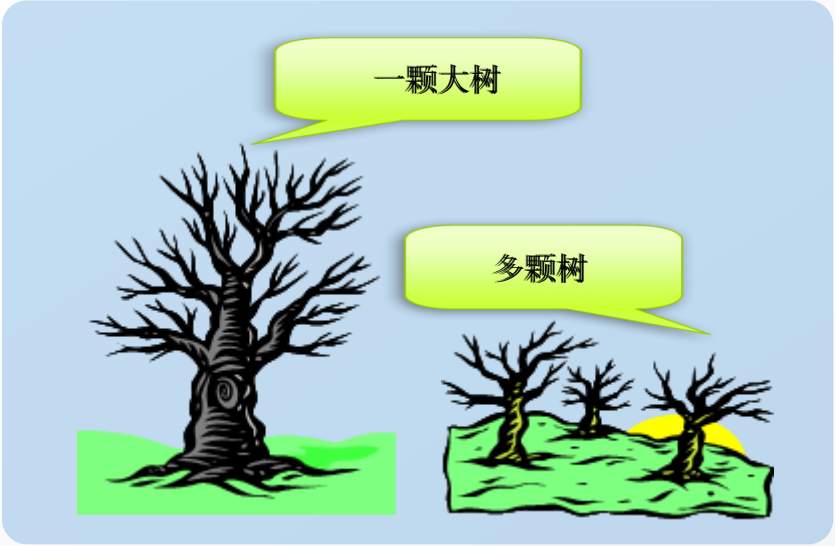
1.1 数据模型与树的比较





主干	是根本，最上层集合，是中心，一定要稳固，个数少。
枝干	与主干相连，业务上的最上层集合，存在多层，是分支的父母。
分支	与枝干相连，个数较多，存在多处，上面张着叶子。
叶子	处于最底层，描述的是业务处理中的行为，分支类型不同深度不同。

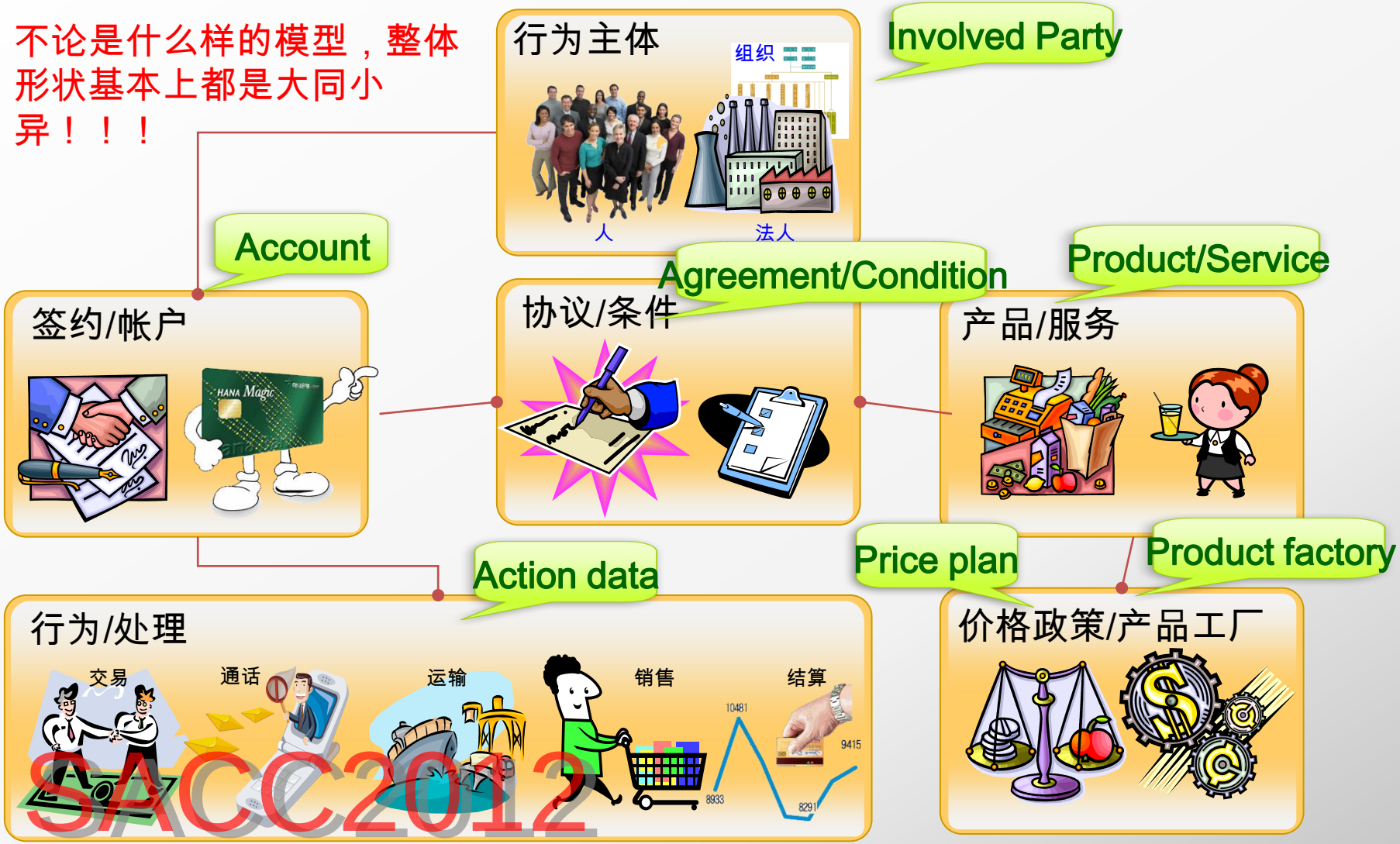
数据模型与树的比较



就像树的种类不同，枝干及外形不同一样，行业不同数据模型的骨骼形状也不尽相同。

1.3 数据模型的骨骼

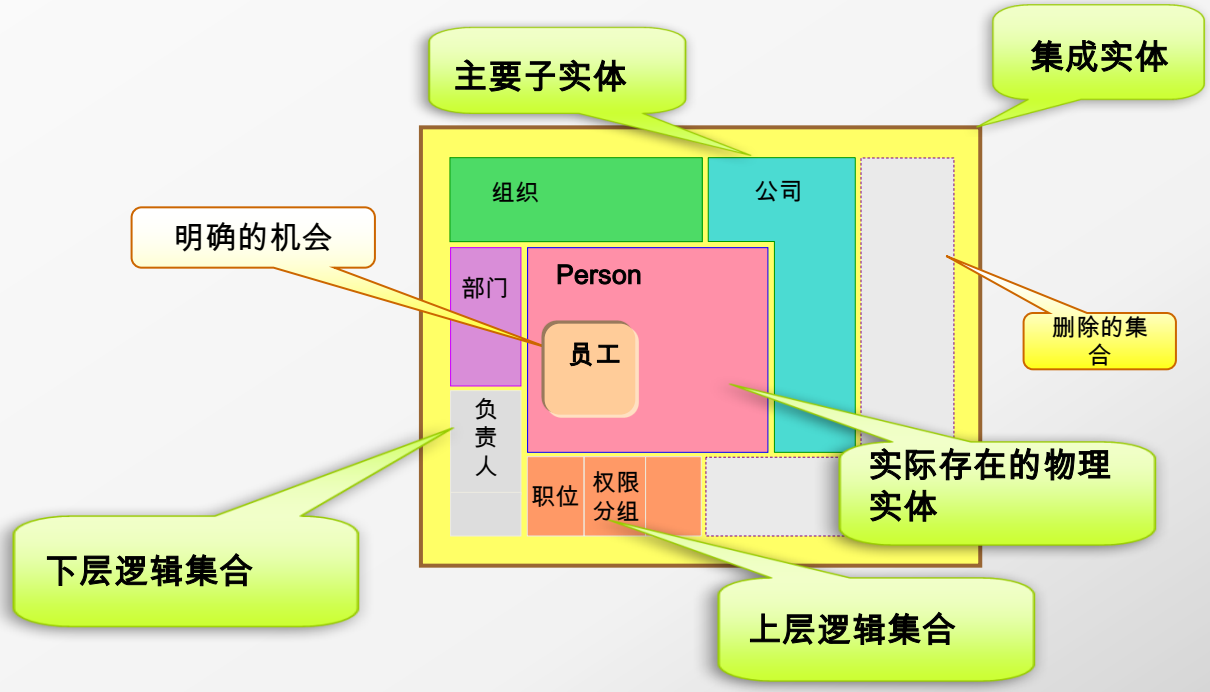
不论是什么样的模型，整体形状基本上都是大同小异！！！



1.4 集成实体的定义原则

集成实体定义的原则

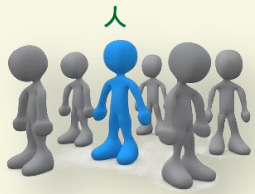
- 优先定义子实体
- 优先定义实际存在的物理实体
- 优先定义比较明确的实体
- 以实体存在的集合为基准，决定其之上和之下的逻辑集合



1.5 物理个体与逻辑个体

物理(原始)实体

可以看得到或摸得到的真实存在或正式定义的实体，不能定义为其它实体的原始实体



逻辑(概念) 实体

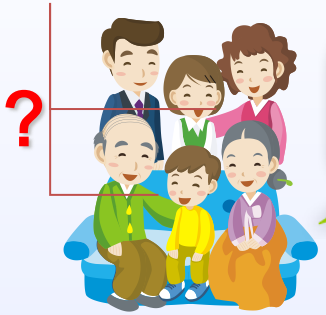
不能看到或碰到的，只是在概念上存在的实体（从原始实体派生出来的实体）

- 目前在大部分的系统里只对眼前看到的物理集合做了定义
- 如果出现特殊情况的话，会导致很严重的问题
- 虽然物理集合只有固定的一个，但逻辑集合却可以根据需要随时定义
- 属性必须属于经过正确正规定义好的集合

16 逻辑实体存在的理由

必须在正确的位置定义信息

家庭成员数	住址类型	派出所名称	派出所编号

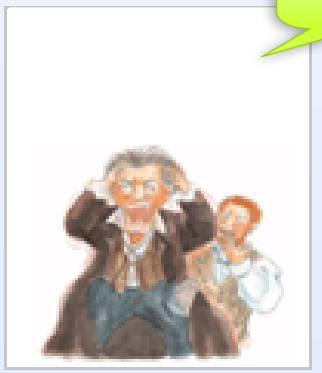


如不存在家庭集合

限额管理?

1人?

2人?



17 逻辑实体存在的理由

逻辑个体 物理个体的

以物理集合为基准，
导出逻辑集合。

SACC2012

EN·CORE

4 逆向建模

10 逆向建模的定义

基于现有系统中的元数据重新设计数据模型

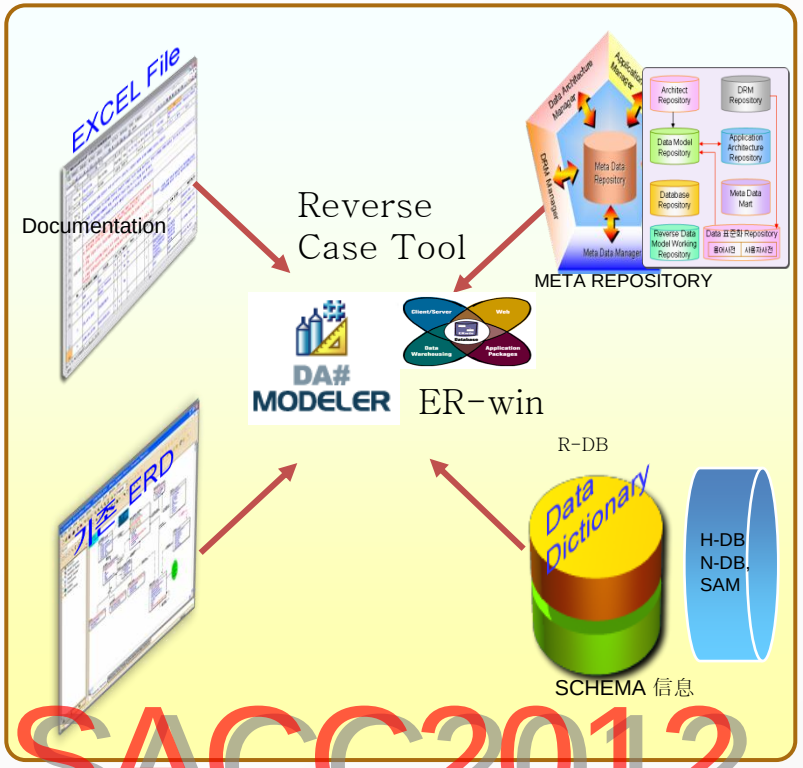
- 当前系统缺乏数据模型，模型中存在很多问题，模型错综复杂难以管理时，利用系统元数据信息重新设计数据模型
- 利用建模工具从DBMS Dictionary或相关文档中收集元数据信息，通过对其中包含的表和列的分析，准确寻找不同实体之间的相互关系。

逆向建模的流程

20 逆向建模的流程—1

元数据收集及分析

收集



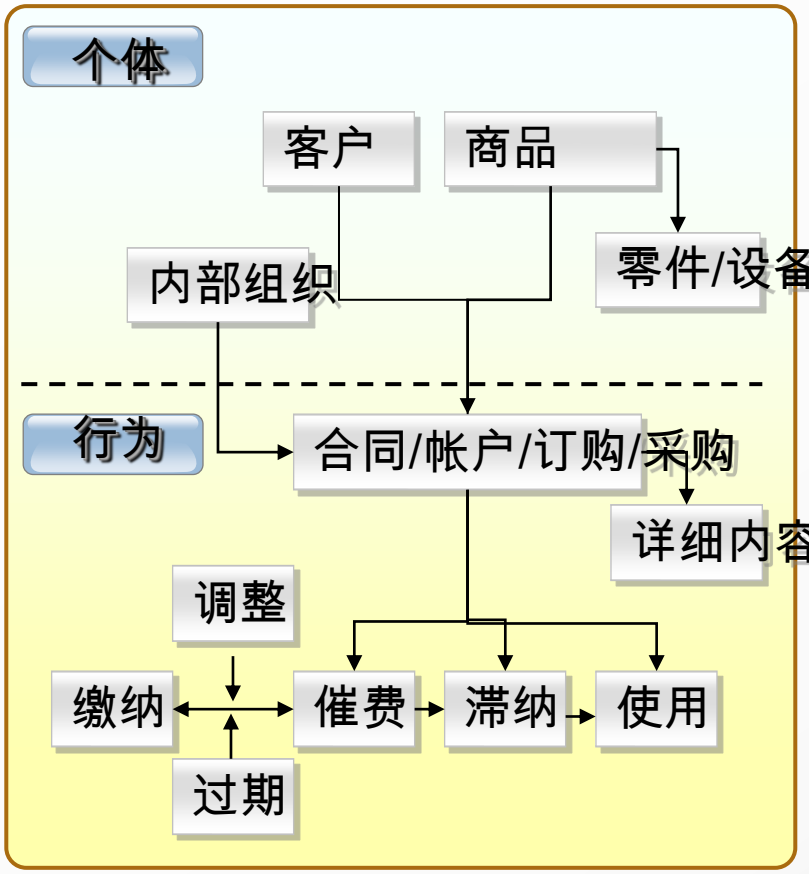
分析结果

表英文名称	表中文名称	表说明	使用与否
tbl_paymentType	付款方式	付款时使用的方式，现金，刷卡等方式。	使用
payment	付款	付款信息	使用
collect	会费	会费管理表(年会费)	使用
collect_wait	预付款	预付款信息	使用
contract	合同	合同信息	使用
contract_hist	合同履历	合同履历表	使用
contract_joiner	合同认证	合同的认证管理	使用
request	缴费	缴费信息管理	使用
staff	MA用户	MA用户管理	使用
cart	购物框	购物框	使用
BranchClub	加盟俱乐部	加盟俱乐部	预定删除
BranchDiv	加盟店地址	加盟店地址	预定删除
tbl_findBranch_bk	会员信息	会员信息管理	备份

21 逆向建模的流程—2

表主题域分类

标准域



分类结果

分类	分类名称	表中文名称
个体	客户	客户,公司，工厂…
	产品/服务	产品，检验服务，费用,打折，注册条件…
	零件/设备	零件，终端，网络设备…
	内部组织	员工，代理商，部门，分公司…
	事件	事故，诉讼
行为	合同/帐户/订购/采购	客户帐户,注册，订购，采购合同…
	(合同/帐户/订购/采购)详细	服务合同详细，订购详细，采购详细
	使用	通话详单，交易详单…
	滞纳	电话费， 打折详细…
	催费	催款， 催费详细，销售…
	缴纳	转账手续费,
	调整	客户投诉，退货，交换
	延迟	未收，过期帐户…

逆向建模的流程—3/6

关系查找

关系定义	继承父实体的识别符(键)→通过识别符查找与之有关系的实体
关系查找	从对象实体的属性中查找从父实体中继承来的属性，并利用它来找其父实体。
关系结束	• 绝对关系：父实体的识别符也是子实体的识别符-先有父实体后有子实体 • 相对关系：父实体的识别符是子实体的一般属性–子实体的辅助属性
关系维持方法	PK: 完整性– 识别符的角色 , FK: 参考完整性
关系难以确认的理由	• ERD不完整，为提升DBMS执行速度而未指定FK • 辅助识别符滥用，关系模糊
关系查找的方法	真正意义上的主语

逆向建模的流程—3/6

绝对关系，继承关系(identification relationship)

- 两个实体之间存在关系是指子实体继承父实体的识别符，并将其作为关系属性来使用。

朋友

姓名	年龄	出生年月
金强	5	2007.01.04
张智创	9	2003.07.23

电话号码
017-234-4567
010-3033-1234

地址
北京市…
上海市…

识别关系

朋友电话

姓名	电话号码
金强	017-234-4567
金强	010-3033-1234
张智创	010-3033-1234

逆向建模的流程—3/6

绝对关系，继承关系(identification relationship)

- 将两个实体之间的关系创建成一个关系实体，并继承这两个实体中的识别符。

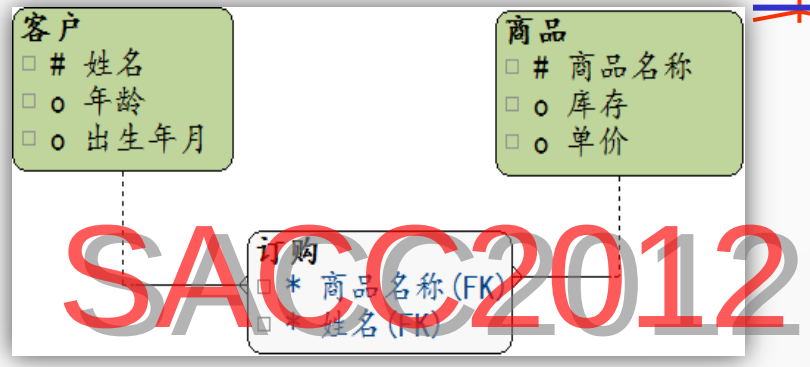
客户

姓名	年龄	出生年龄
金强	1	2007.01.04
张智创	4	2003.07.23

订购

商品

商品	库存	单价
电话	1	20000
TV	4	30000



客户	商品
金强	电话
金强	TV
张智创	电话

逆向建模的流程—3/6

相对关系, 非继承关系(None identification relationship)

- 两个实体之间存在相对关系是指一个实体继承另一个实体的识别符属性，但将其作为一般属性来使用。

朋友

姓名	年龄	出生年月
金强	1	2007.01.04
张智创	4	2003.07.23

地址编码
001
002

地址名称
北京市
上海市

参考继承

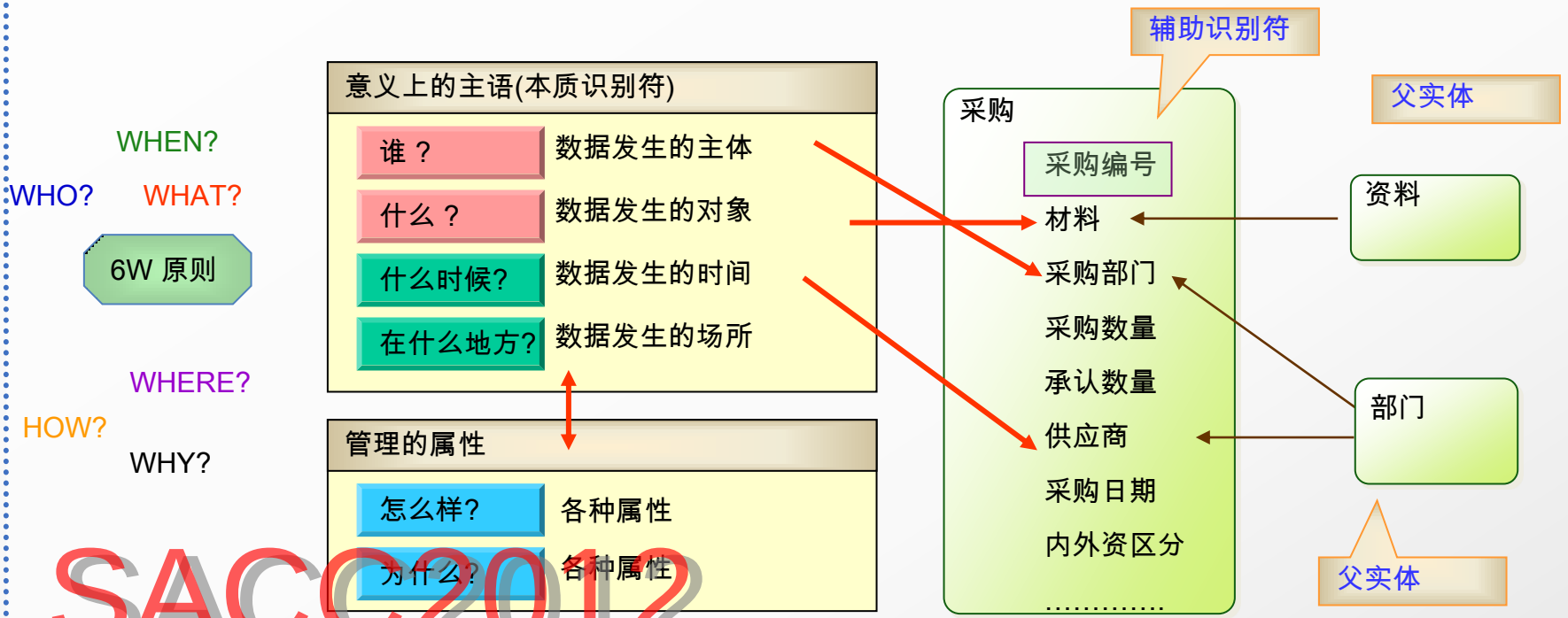
地址

地址编码	地址名称
001	北京市
002	上海市

26 逆向建模的流程—4

意义上的重要

- 个体实体位于最上层，只拥有相对关系，主要以此为基准去查找绝对关系。
- 行为实体是查找绝对关系父实体的对象->需要使用关联识别符寻找父实体。
- 利用形成数据的4个原因来验证其正确性。



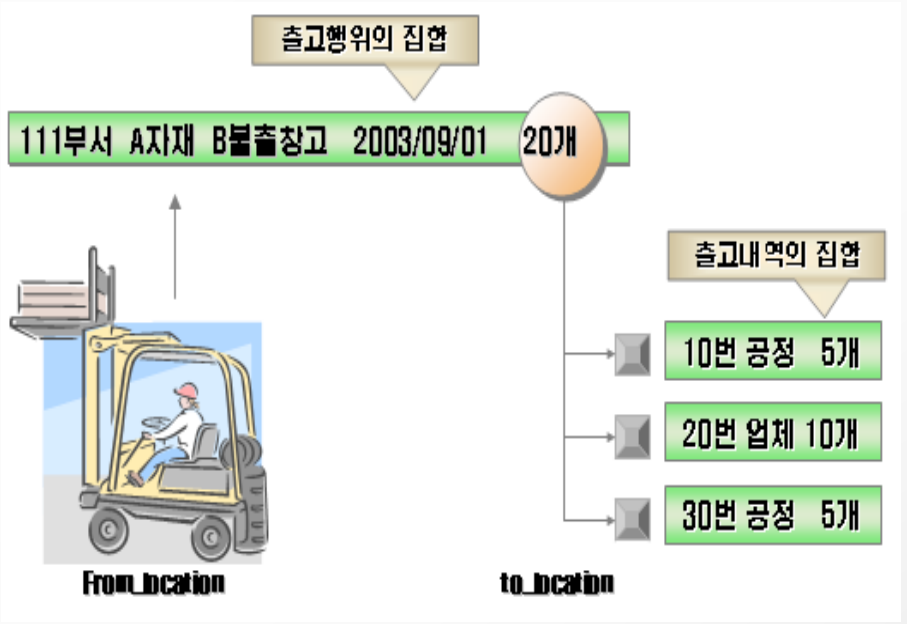
27 逆向建模的流程—5

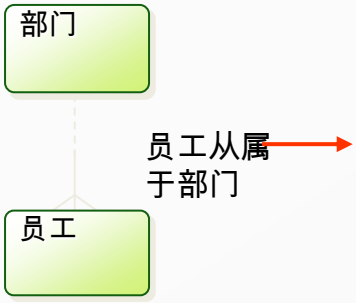
关系确认

材料出库详细
#出库编号
出库区分
出库日期
材料编码
出库部门
使用部门
出库数量
申请日期
出库仓库编号
供货商编号
受理人

?

谁？
什么？
什么时候?
什么地方?





但是需从业务的角度确认数据的所属

所有权如随着受理人部门的改变而改变，则受理人将成为主体，否则出库部门将成为主题。

逆向建模的流程—6

逻辑化

86之前客户

- 删除未使用表

客户

临时客户

- 合并充分实体

客户基本信息

客户详细信息

- 垂直分割合并

个人客户

企业客户

团体客户

- 水平分割合并

个人客户
. 客户id . 客户姓名
. 身份证 .兴趣爱好

企业客户
.客户id .客户名
.营业执照编号
. 法人代表姓名

客户
. 客户ID
. 客户名

个人客户
. 身份证号码
. 兴趣爱好

企业客户
.营业执照编号
.法人代表姓名

- 属性合并

客户

个人

企业 人员 法人

团体

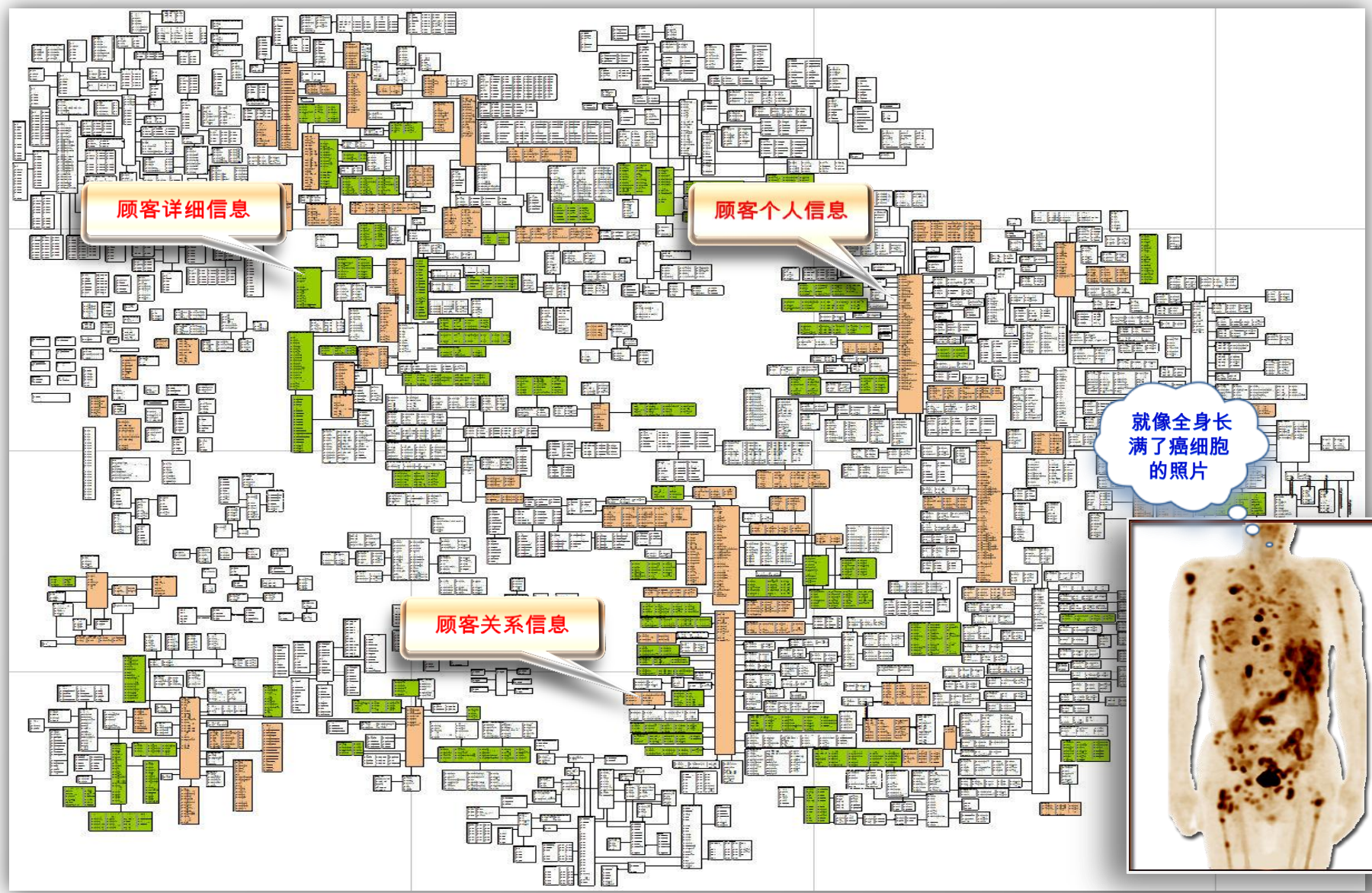
- 实体详细化

客户

联系方式(p)
. 固定电话
. 手机
. 家庭住址
. 公司地址
. 电子邮件

- 属性详细化

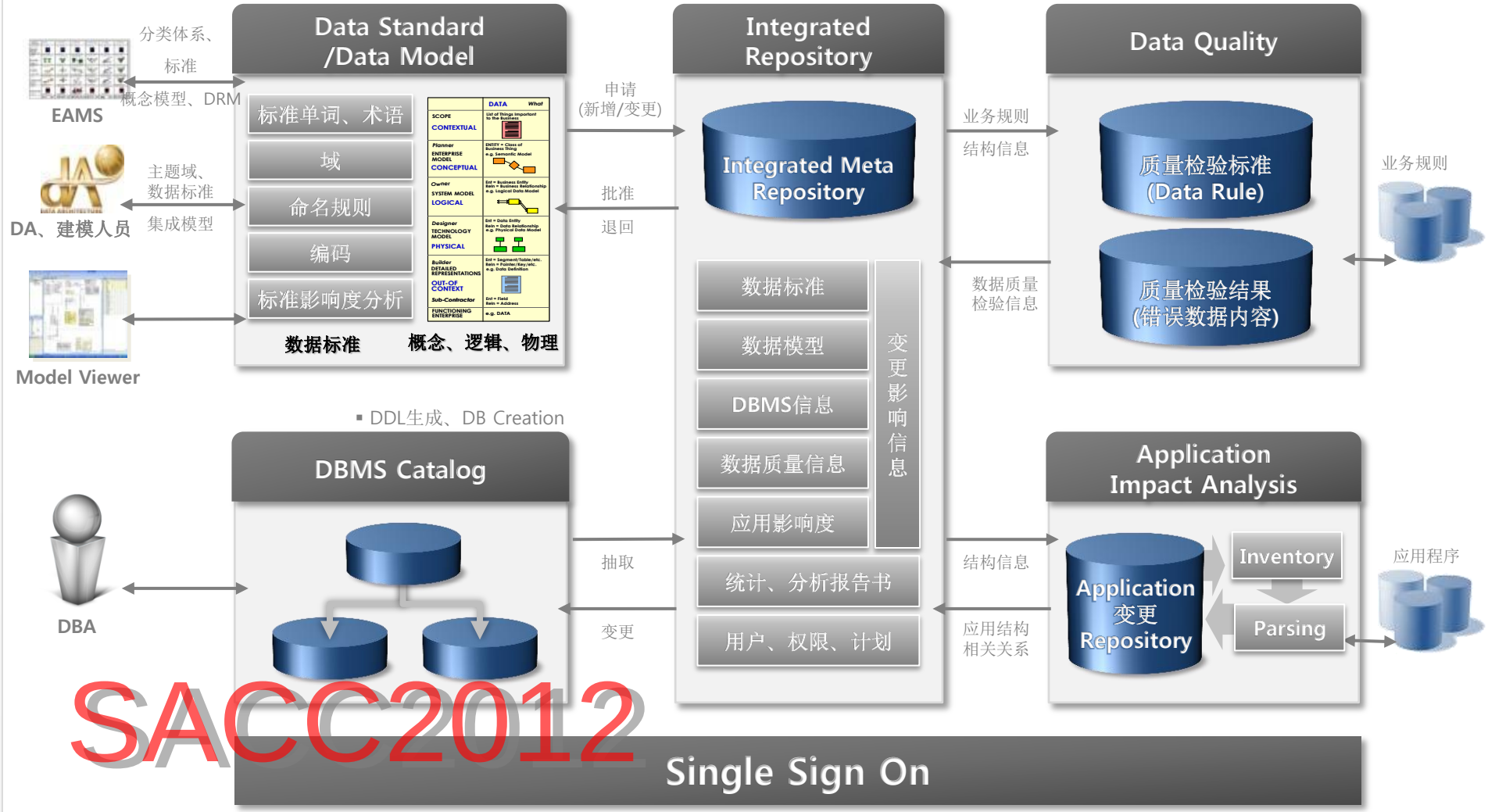
20 逆向建模的实际案例



EN·CORE

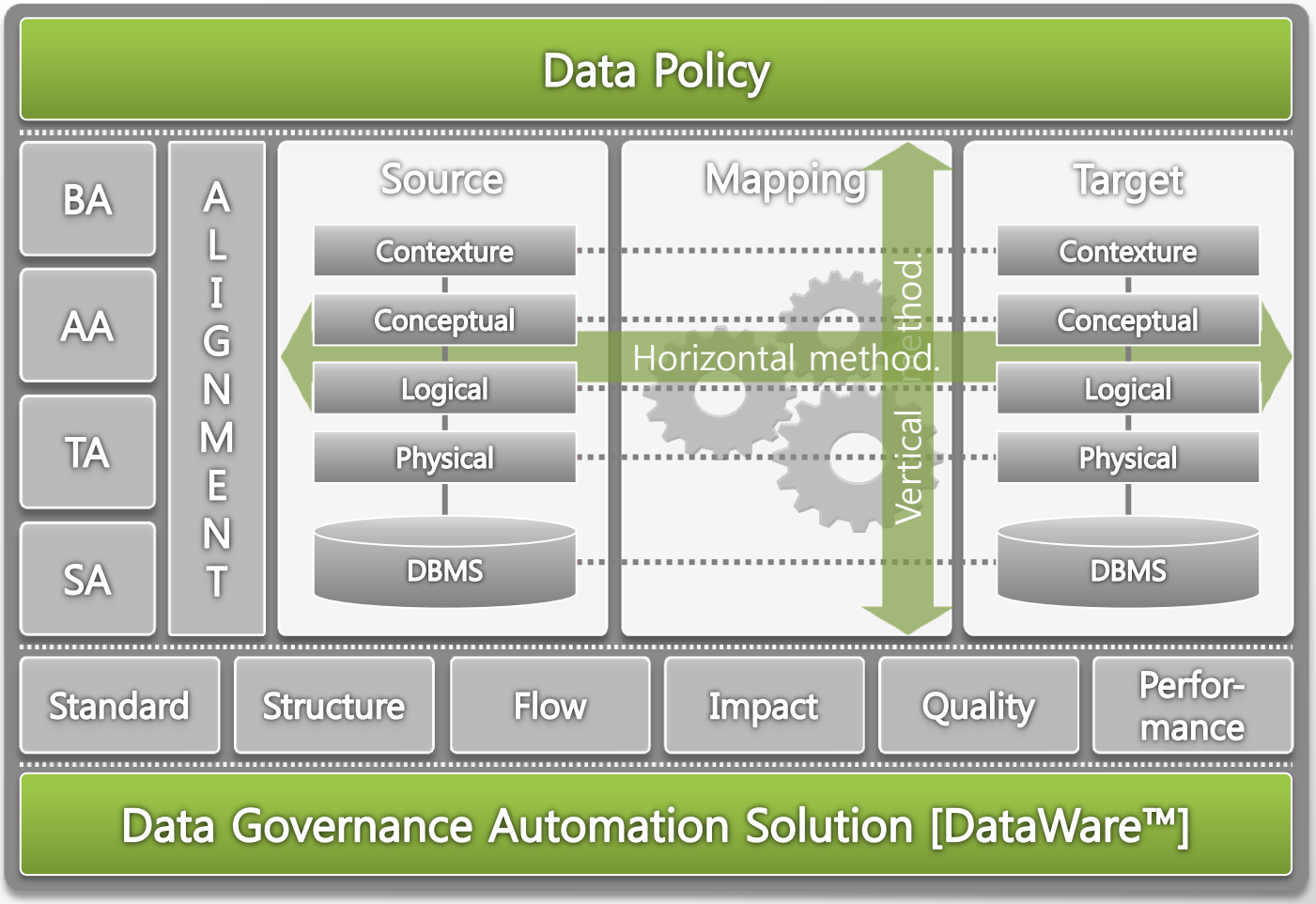
数据架构结构图

数据标准、数据结构、数据质量、变更影响度管理系统结构示例



2.1 EN-CORE Data Framework

Enterprise Data governance Framework



Policy Layer

1	DA团队设计
2	数据管理政策制定
3	数据管理体系制定

Method Layer

1	数据建模
2	逆向建模
3	数据库架构

Function Layer

1	数据标准化
2	数据架构
3	数据流程管理
4	数据质量
5	数据库性能优化

Solution Layer

EN·CORE

Thank you!

SACC2012

EN·CORE