基于数据仓库的高校数据统计服务平台研究

龙新征
北京大学计算中心

提纲

- 背景
- 需求分析
- 平台设计
- 平台实现
- 平台部署

背景

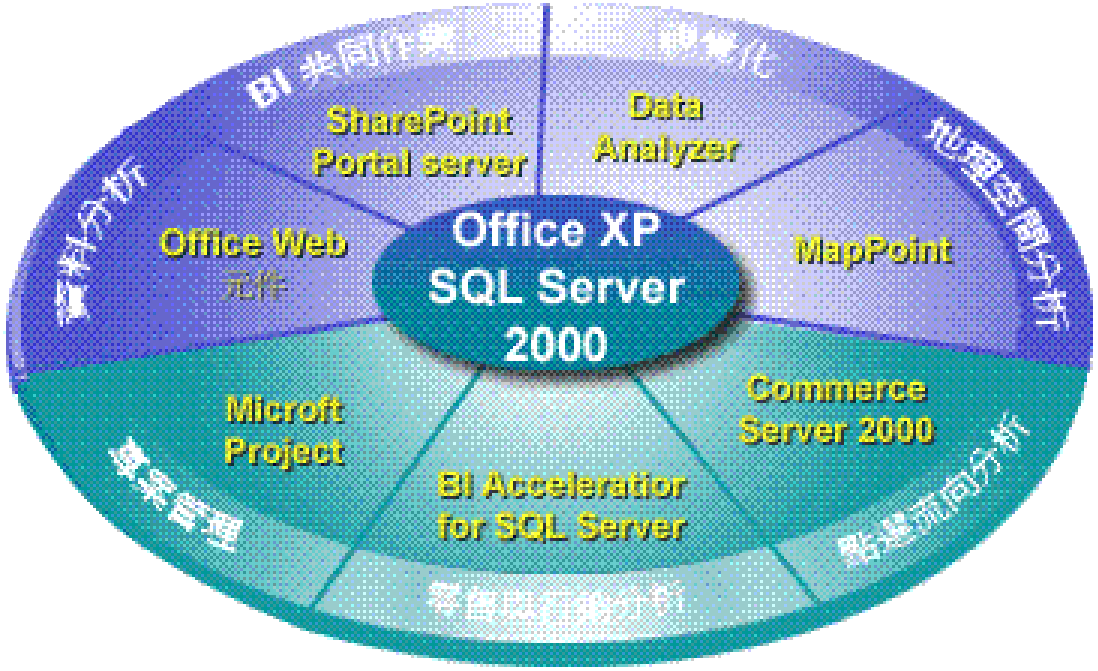
- 各高校建成的信息管理系统越来越多，海量数据背后隐藏着许多重要信息，是学校正常运转的核心资源，以灵活便捷的方式对数据进行统计、分析，进而为高校管理与决策提供支持的需求日益强烈

背景

- 商业智能(**BI, business intelligence**)
- **Gartner Group, Howard Dresner, 1996**：一类由数据仓库(或数据集市)、查询报表、数据分析、数据挖掘、数据备份和恢复等部分组成，以帮助企业决策为目的的技术及其应用。

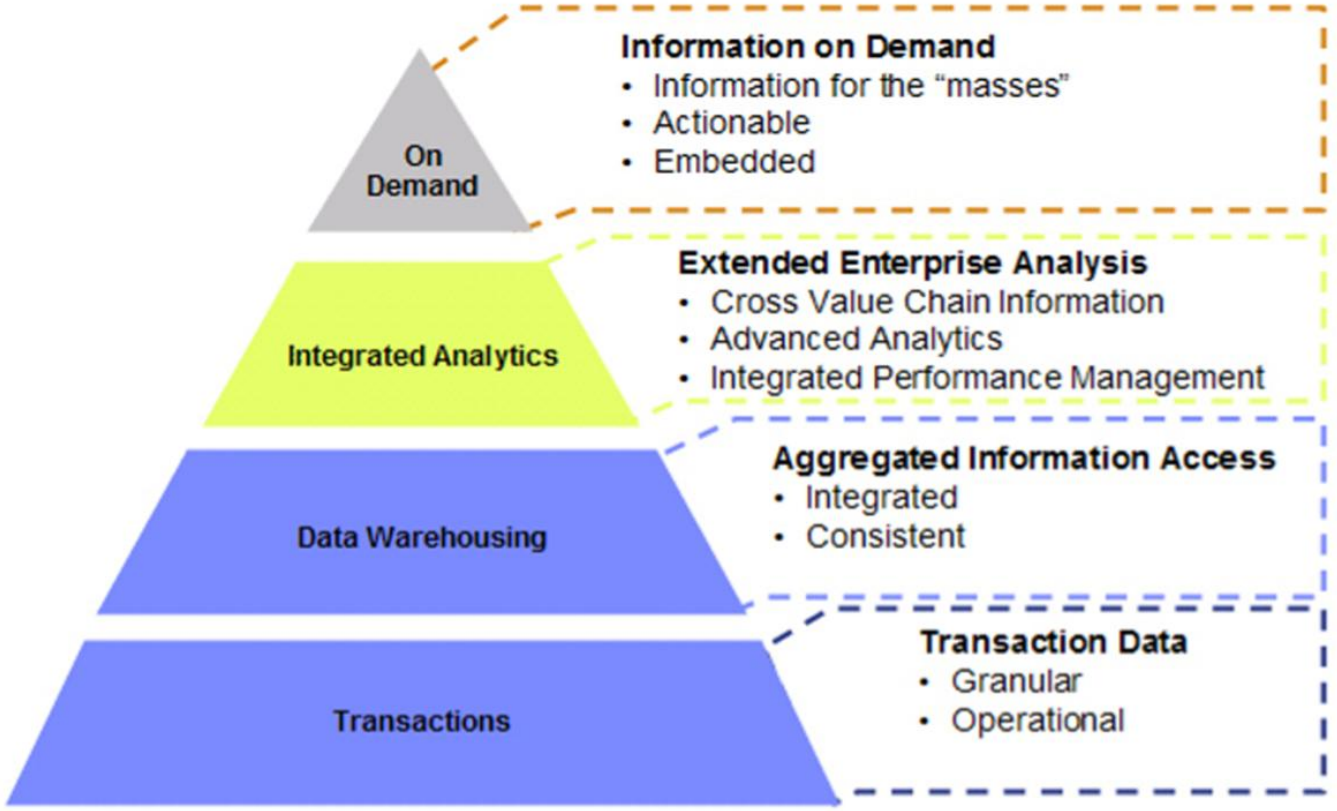
背景

- 微软的**BI**体系框架



背景

IBM的BI体系框架



背景

- 建立**BI**系统的基本步骤包括：
  - 确认和解读数据源；
  - 进行数据采集和存储管理；
  - 构建模型并在此基础上分析数据

背景

- 商业智能的基础是数据仓库(DW，Data Warehouse)
- 数据仓库是一个面向主题的、集成的、相对稳定的、反映历史变化的数据集合。
  - 数据仓库为有效地为BI系统提供了全局一致的数据环境，也为历史数据综合数据的处理提出了一种行之有效的解决方法。

背景

- 国外:**BI**应用已经进入了数据分析阶段，有些已经积累了高端的数据挖掘经验；
- 国内:**BI**的应用则还停留在数据整合的初级阶段，应用的主要领域集中在电信、保险、销售等行业，
- 国内高校的情况不容乐观：
  - 数据分散在不同的源系统中，数据的规范性和共享性还存在很大问题；
  - 数据统计和分析基本上是基于单个系统中的操作型数据进行的，既不能反映出不同系统之间的数据关联，又缺乏对数据的全局把握；还会因为操作型数据的动态性和分散性影响统计结果的准确性，也无法对历史数据进行统计和分析。

背景

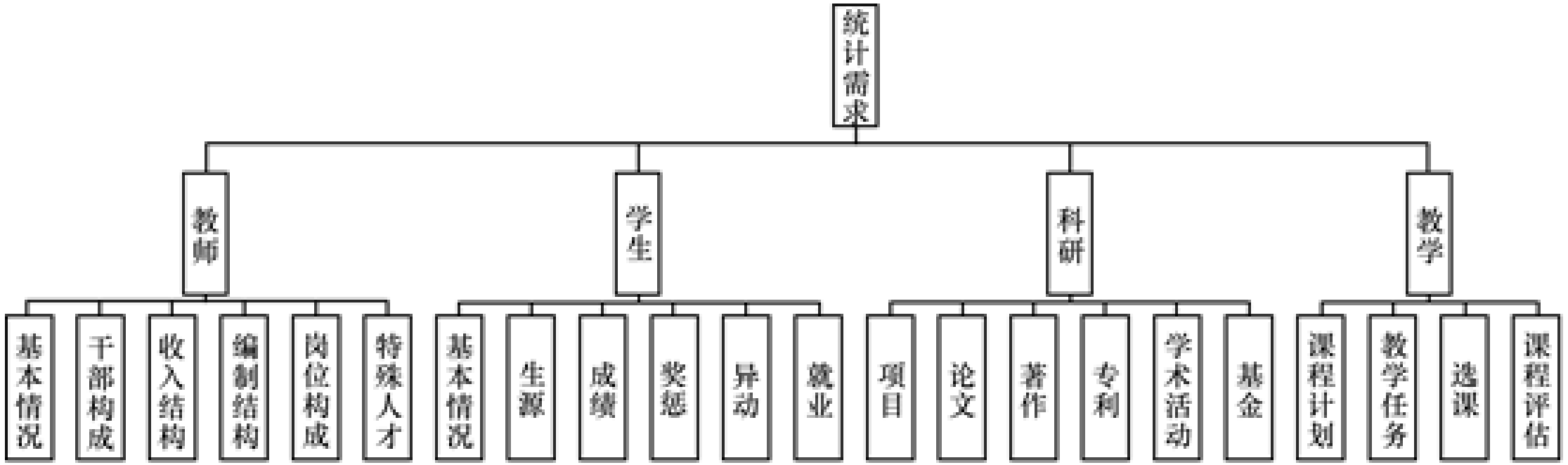
- 要建立高校的**BI**应用，满足为高校管理与决策提供支持的需求，首先必须打破不同应用系统之间的“藩篱”，建立全局一致的数据仓库，将操作型数据转换为静态的、稳定的、规范化的、能够反映历史的分析型数据，然后在此基础上搭建统一的数据统计服务平台。

需求分析

- 数据统计需求按用途不同分为2类：
  - 一是以年报表或者季度报表的形式上报上级部门的统计数据，有着固定的报表格式、复杂的报表内容、专门的统计口径，有的报表甚至还有严格的填报流程；
  - 二是部门日常所需要统计数据，往往和某种类型具体业务相关，和第一类需求相比，统计数据的格式和内容比较简单，没有复杂的填报流程，但时间粒度要求更细，要以月报表、周报表乃至日报表的形式提供统计结果，并且要求提供数据钻取的功能。

需求分析

- 这些需要统计的数据涉及高校人事、学生、科研、教学等各个领域，每个领域下面又细分为了很多不同的细类，具体如图1所示。

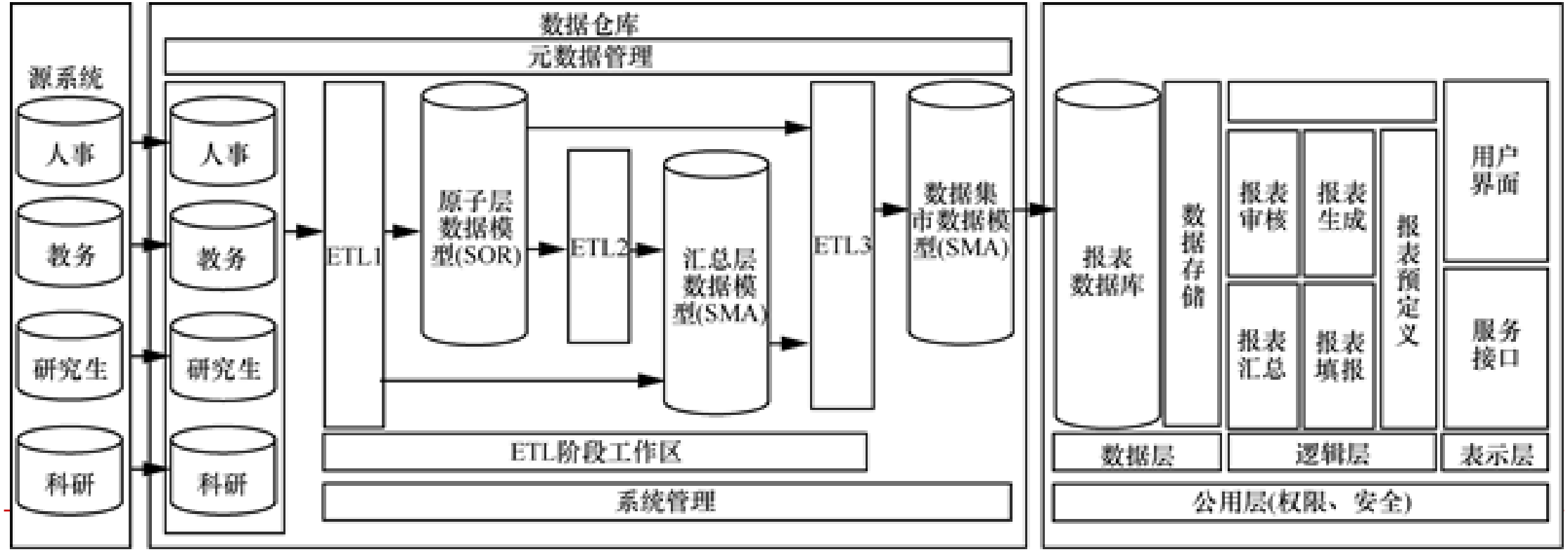


平台设计

- 总体架构
- 数据仓库建模
- 统计数据模型

总体架构

- 高校数据统计服务平台由源系统、数据仓库和统计平台三部分组成。



总体架构

- **源系统**：高校业务涉及的数据源比较广泛，主要有人事系统、教务系统、研究生系统、科研系统等，再加上校园网之外的一些其他外部数据源，构成了数据统计服务平台的数据基础，由于系统业务职能和具体需求不同，在实现时会选用不同的数据库，数据结构也可能存在较大差异，从而导致数据间有较大的异构性和不一致性。

总体架构

- **数据仓库**：数据仓库全面接收源系统数据，**ETL**进程对数据进行规范化、验证、清洗，并最终装载进入数据集市，通过数据集市支持系统进行数据查询、分析；整个数据仓库包含四大层次：
  - **复制层(SSA, system-of-records-staging-area)**
  - **原子层(SOR, system-of-record)**
  - **汇总层(SMA, summary-area)**
  - **集市层(DM, data mart)**

总体架构

- **复制层(SSA, system-of-records-staging-area)**：直接复制源系统的数据，尽量保持业务数据的原貌；与源系统数据唯一不同的是，复制层中的数据在源系统数据的基础上加入了时间戳的信息，形成了多个版本的历史数据信息；
- **原子层(SOR, system-of-record)**：基于模型开发的一套符合**3NF**范式规则的表结构，它存储了数据仓库内最细层次的数据，并按照不同的主题域对数据分类存储；根据目前部分需求，将全校数据在原子层中按人事、学生、教学、科研四大主题存储；原子层是整个数据仓库的核心和基础，在设计过程中应具有足够的灵活性，以能应对添加更多的数据源、支持更多的分析需求，同时能够支持进一步的升级和更新；

总体架构

- 汇总层(SMA, summary-area)：汇总层是原子层和集市层的中间过渡，由于原子层的数据是高度规范化数据，因此要完成一个查询需要大量的关联工作，同时集市层中的数据粒度往往要比原子层高很多，对要生成集市层中的汇总数据需要进行大量的汇总工作，因此，汇总层根据需求把原子层数据进行适度的反范（例如，设计宽表结构将人员信息、干部信息等多个表的数据合并起来）和汇总（例如，一些常用的人头汇总、机构汇总等）；从而提高数据仓库查询的性能。

总体架构

- **集市层(DM, data mart)**：集市层保存的数据是供用户直接访问的；可以将集市层理解成最终用户直接最终想要看的数据；集市层主要是各类粒度的事实数据，通过提供不同粒度的数据，适应不同的数据访问需求；集市层中的数据以**2**种不同类型存储：一类以星型模型建设，便于部门日常的灵活查询和统计，另一类按宽表以及重新组织的适应固定报表的表结构存储，便于高校的年统和季度统计工作。

总体架构

- **统计平台**：高校数据统计服务平台采用B/S架构的3层体系结构，即：数据操作层、逻辑层、表示层。
  - **数据操作层**
  - **逻辑层**
  - **表示层**

总体架构

- **数据操作层：**充分考虑系统的高可用性，数据统计服务平台与数据仓库所使用的数据库互相独立，由此保证数据统计服务平台对数据进行加工处理时不会影响数据仓库中的数据；数据存取模块实现对数据统计服务平台数据的访问。

总体架构

- **逻辑层**：分为报表预定义、报表查询、报表生成、报表填报、报表审核及报表汇总等模块；每个模块分别实现不同的功能；在统计平台中，不同身份的用户其功能权限和数据权限是不一样的：报表预定义是给系统管理员用的；报表生成、报表填报是给院系管理人员使用的，只能查看和操作本院系的数据；报表审核、报表汇总是给学校相关部门的管理人员用的，可以操作全校数据；功能权限和数据权限通过公用层与身份认证服务平台对接，统一进行管理

总体架构

- **表示层**：提供交互界面给用户使用，此外还提供一些服务接口供其他系统调用

数据仓库建模

- 目前较为流行的数据仓库的建模方法较多，常用的有**Inmon**所提倡的范式建模法和**Kimball**所提倡的维度建模法。

数据仓库建模

- 维度建模法针对各个维做了大量的预处理，通过这些预处理能够极大地提升数据仓库的处理能力，相对于范式建模法来说，在性能上占据了明显的优势；同时维度建模非常直观，紧紧围绕着业务模型，可以直观地反映出业务模型中的业务问题。不需要经过特别的抽象处理即可以完成维度建模。因此高校数据统计服务平台的数据仓库采取维度建模的方式构建。
- 维度建模法采用事实表—维表的方式来构建数据仓库，数据集市、事实表存储实际的数据，维表存储事实表中对象的属性，事实表和维表的关联关系常用的是“星型模型”。

数据仓库建模

- 维度建模的步骤
- **结合具体需求确定分析主题**，结合高校主要业务定义了一个公共维度主题和人事、学生、教学、科研4个业务主题：公共维度包含时间维、地理维、国标及校标，时间维和地理维在不同的应用场景可以使用视图形式转换为具体的分析维度，国标和校标主要用来解决在数据集成过程中的一致性问题；人事主题核心内容是教师的基本情况，具体分析主体有收入、岗位、职称以及杰出人才等；学生主题核心内容是在校生基本情况，具体分析主题有招生、成绩、奖惩、异动、就业等；科研主题主要分析全校师生科研成果完成情况，根据实际业务可以纳入所有科研成果，如项目、论文、著作、专利、学术活动等；教学主题以教学活动相关内容为主，如课程计划、教学任务、选课、教学工作量等。

数据仓库建模

- **确定分析粒度**，通俗地说就是分析对象的详细程度。为了满足分析的可扩展性及需求的多样性，以最小粒度来设计数据模型总是能达到最好的分析效果，如：记录每个学生的明细情况、记录每项科研成果的详细情况。

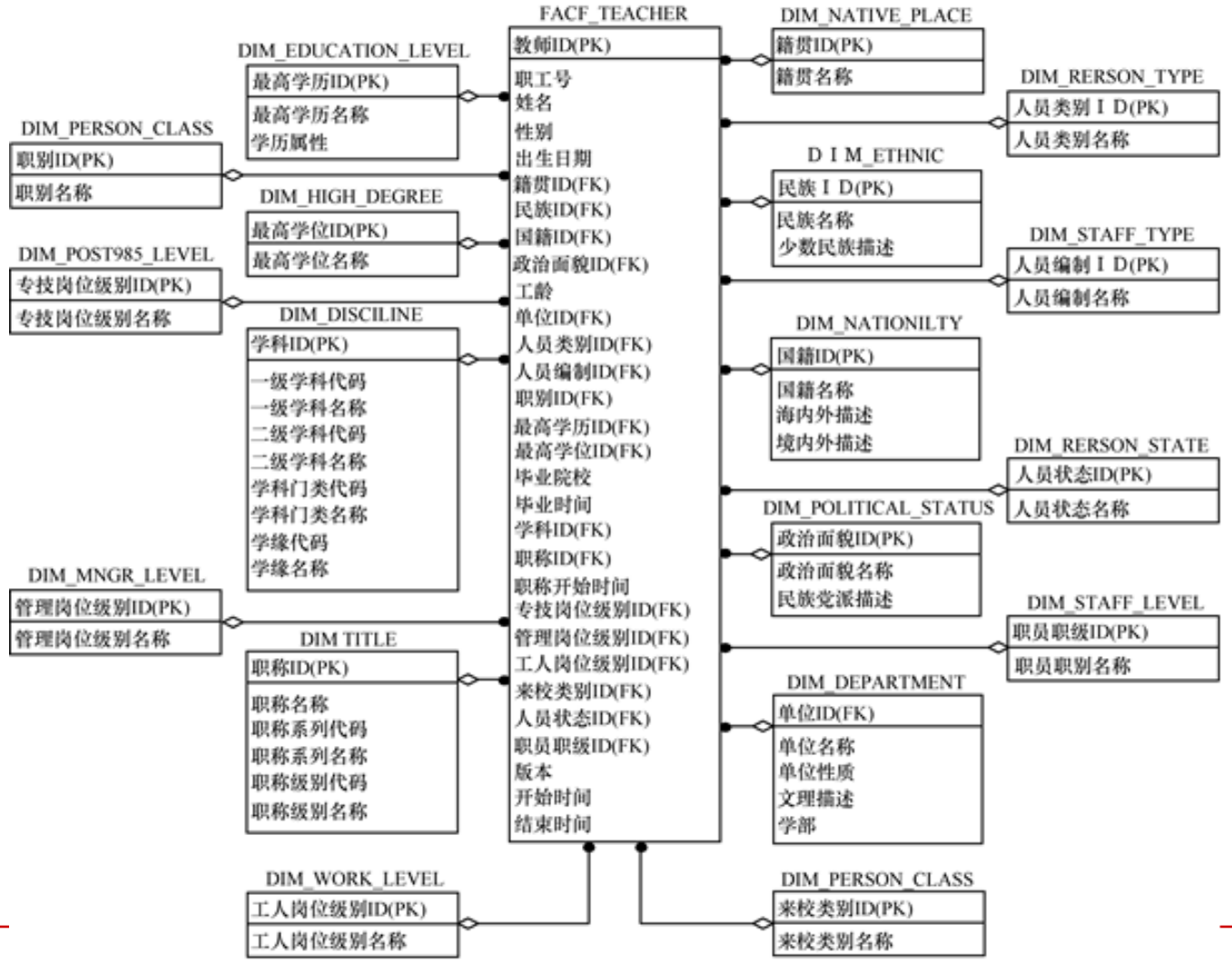
数据仓库建模

- **设计维表**，维度是统计和分析数据的角度，与统计查询的参数相对应。在选取维度时应该将实体作为一个对象，把与该对象相关的所有重要属性都提取出来作为独立维度。

数据仓库建模

- **设计事实表**，为了跟踪具有生命周期的活动数据的变化过程以保留历史信息，设计事实表时使用缓慢变化维的方法以捕获变化数据。事实表中的版本、开始时间和结束时间3个字段是实现缓慢变化的核心。版本表示同一事物历史状态的顺序，开始时间和结束时间表示在该段时间内该事物处于某一状态，每一条数据的结束时间等于新数据的开始时间，这样该事物不同时间段的状态就分布在一条时间轴上，从而可以得到任一时间点该事物的状态信息

数据仓库建模

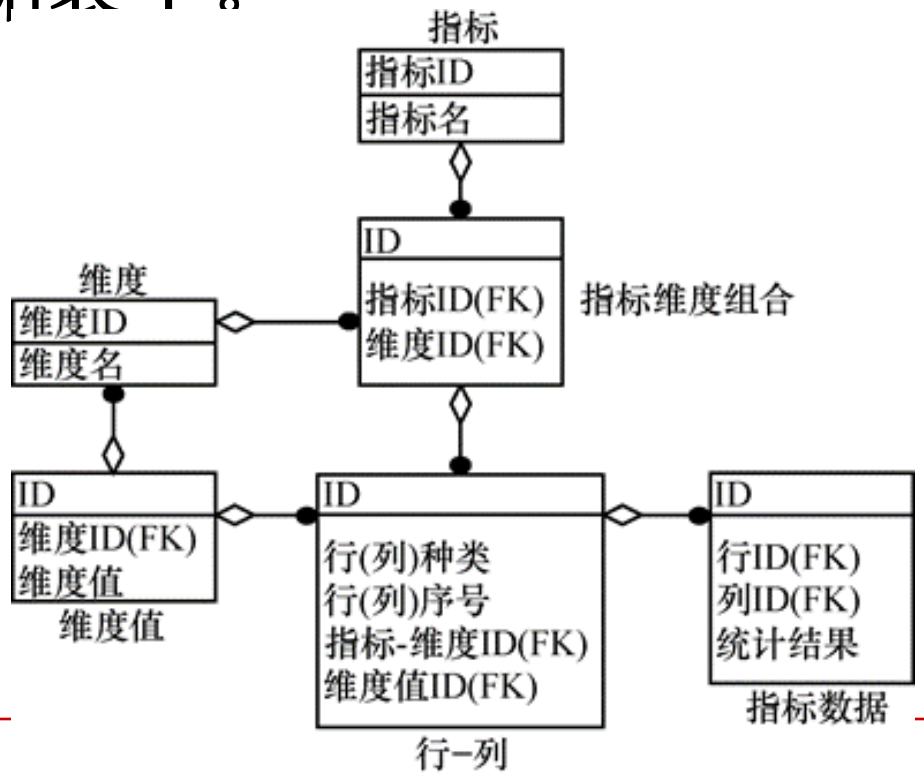

FACF_TEACHER
教师ID(PK)
职工号
姓名
性别
出生日期
籍贯ID(FK)
民族ID(FK)
国籍ID(FK)
政治面貌ID(FK)
工龄
单位ID(FK)
人员类别ID(FK)
人员编制ID(FK)
职别ID(FK)
最高学历ID(FK)
最高学位ID(FK)
毕业院校
毕业时间
学科ID(FK)
职称ID(FK)
职称开始时间
专技岗位级别ID(FK)
管理岗位级别ID(FK)
工人岗位级别ID(FK)
来校类别ID(FK)
人员状态ID(FK)
职员职级ID(FK)
版本
开始时间
结束时间
DIM_EDUCATION_LEVEL
最高学历ID(PK)
最高学历名称
学历属性
DIM_PERSON_CLASS
职别ID(PK)
职别名称
DIM_HIGH_DEGREE
最高学位ID(PK)
最高学位名称
DIM_POST985_LEVEL
专技岗位级别ID(PK)
专技岗位级别名称
DIM_DISCILINE
学科ID(PK)
一级学科代码
一级学科名称
二级学科代码
二级学科名称
学科门类代码
学科门类名称
学缘代码
学缘名称
DIM_MNGR_LEVEL
管理岗位级别ID(PK)
管理岗位级别名称
DIM TITLE
职称ID(PK)
职称名称
职称系列代码
职称系列名称
职称级别代码
职称级别名称
DIM_WORK_LEVEL
工人岗位级别ID(PK)
工人岗位级别名称
DIM_NATIVE_PLACE
籍贯ID(PK)
籍贯名称
DIM_RERSON_TYPE
人员类别ＩD(PK)
人员类别名称
D I M_ETHNIC
民族ＩD(PK)
民族名称
少数民族描述
DIM_STAFF_TYPE
人员编制ＩD(PK)
人员编制名称
DIM_NATIONILTY
国籍ID(PK)
国籍名称
海内外描述
境内外描述
DIM_RERSON_STATE
人员状态ID(PK)
人员状态名称
DIM_POLITICAL_STATUS
政治面貌ID(PK)
政治面貌名称
民族党派描述
DIM_STAFF_LEVEL
职员职级ID(PK)
职员职别名称
DIM_DEPARTMENT
单位ID(FK)
单位名称
单位性质
文理描述
学部
DIM_PERSON_CLASS
来校类别ID(PK)
来校类别名称

统计数据模型

- 确定统计相关的星型模型，即数据统计针对的是事实表中间的那些事实，涉及到哪些统计指标，统计的粒度如何。
- 确定报表中具体的每一行和每一列分别代表的统计指标，统计指标简单地说即维度取值，每个统计指标对应到维表中是某个维度取某个值，也有可能是多个维度取值的累加。
- 确定单元格的统计方法，每个单元格的统计指标应该是其对应的行、列所代表的维度取值的并集。
- 将维度转化为可执行查询的语句，去事实表中查询出相应的统计数据和事实数据，为了方便，在数据统计服务平台的报表数据库中还可以将查询到的统计结果固化，以数据库表的形式存储下来。

统计数据模型

- 将1张业务报表拆分为5张配置表，它们分别是指标表、维表、维值表、行列表、指标维度组合表；最后计算得到的结果存储在指标数据表中。



平台实现

- ETL处理
- 前台展示

ETL处理

- **ETL(extraction-transformation-loading)**负责将分散的、异构数据源中的数据抽取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。**ETL**是实施数据仓库的核心和灵魂，**ETL**规则的设计和实施约占整个数据仓库搭建工作量的**60%～80%**。

ETL处理

- 数据抽取
  - 包括初始化数据装载和数据刷新：初始化数据装载主要关注的是如何建立维表、事实表，并把相应的数据放到这些数据表中，在数据仓库建模小节中已经做了详细介绍；而数据刷新关注的是当源数据发生变化时如何对数据仓库中的相应数据进行追加和更新等维护

ETL处理

- 触发器方式（又称为快照式）来实现数据刷新，具体来说就是：在**SSA**层需要抽取数据的数据表上建立了插入、修改、删除**3**个触发器（**trigger**），每当源数据库中数据表中的数据发生变化时，复制到**SSA**的数据也会相应发生改变，相应的触发器将变化的数据写入一个临时区（**buffer**）；在数据库层定义了一系列的作业（**job**）和存储过程（**procedure**）：作业规定了包括数据刷新频率和数据刷新先后次序在内的一系列任务调度策略，调用相应的存储过程从临时表中抽取需要刷新的数据，临时表中抽取过的数据被标记或删除；
- 触发器方式的好处是：数据抽取的性能高、规则简单，对于编程人员来说易于上手，特别适合北京大学数据仓库现有规模还较小的特点，是一种简单易行的好办法；但随着以后数据仓库规模的越来越大，数据表越来越多，需要编写的触发器、存储过程和作业就越来越多，可能会不利于管理

ETL处理

数据清洗

- 主要是针对源数据库中出现的二义性、重复、不完整、违反业务或逻辑规则等问题的数据进行统一的处理，下表列出了北京大学在对业务系统进行数据清洗时发现的几类最常见的问题及针对这些问题所采取的策略。

表 1　北京大学业务系统数据清洗的常见问题及策略

主要问题	表现形式	产生原因	清洗策略
完整性问题	大量空值字段的出现	源系统中对很多字段没有做非空限制	1) 交由源系统重新录入、补齐 2) 在数据仓库对应的维表中建立一个新的字段，将这些空值字段的值统一赋值
	超出字典表范围	填写这些值的时候是直接让用户填写而非下拉框选择	1) 交由源系统重新录入、补齐 2) 在数据仓库对应的维表中建立一个新的字段，将这些空值字段的值统一赋值
一致性问题	一个特定的字段在不同的表中内容不同	录入，同步的问题	1) 交由源系统重新录入、修改 2) 选取最可靠的表中的字段为确定值，
	应该成为主键的值不唯一	源系统中未建立有效的主键关系	消除错误、重复的主键

ETL处理

- 数据转换
  - 主要是为了将数据清洗后的数据转换成数据仓库所需要的数据：来源于不同源系统的同一数据字段的数据字典或者数据格式可能不一样，在数据仓库中需要给它们提供统一的数据字典和格式，对数据内容进行归一化；另一方面，数据仓库所需要的某些字段的内容可能是源系统所不具备的，而是需要根据源系统中多个字段的内容共同确定；
    - 例如，数据仓库中的人员类型“事业单位专业技术人员”实际上是根据人事表中“编制类型=事业单位”、“岗位级别=985”并且“人员类别=在职职工”等多个字段的内容共同得出的，像这样字段的形成也依赖于数据转换

ETL处理

- 考察的工具
  - ColverETL:开源ETL工具，免费版本支持的连接组件太少（Pass）
  - Kettle: 功能完善，组件齐全的处理平台
  - Talend: 功能完善，组件齐全的处理平台
  - Jitterbit: ETL工具，但是功能比较简单，维护、日志、监控等功能缺乏
  - Apatar: ETL工具，非服务器结构，适合单机版本开发小的ETL程序
  - OpenDigger：ETL工具，非图形化接口
  - Spring batch：主要用于实现调度平台，配置方法和spring工具

ETL处理

	Kettle	Talend
结构	C/S结构	非C/S结构
资料库	支持	免费版不支持
迁移和部署	不同环境可直接打开共享资料库，无需迁移	jar包方式部署
接口支持	主流数据库，文件和Webservice	数据库，文件，Webservice，MQ，Socket等
转换功能	常用ETL转换控件	常用ETL转换控件
性能	大数据量下偶尔会出现OutofMemory，需要对虚拟机进行手动调整	支持磁盘外排序，SQLLoader等功能
友好性	图形化开发	基于Eclipse插件的全图形化开发
支持	提供付费支持，免费的文档并不是很多且不详细	有在线论坛免费支持，以及开发文档，中文资源不多
成本	开源LGPL License	开源GPL License
开发难度	拖拽式开发配置	拖拽式开发配置

ETL处理

- **Kettle优势**
  - **LGPL License限制较为宽松**
  - 免费的**Repository**使得版本管理和代码迁移非常容易
  - 任务调度支持定时,时间和命令
  - 支持**Job Duplication**

ETL处理

- Talend优势
  - 接口支持非常丰富，包括：各种数据库，文件(Excel CSV Jason XML Mail等)，外围系统(SAP,CRM,FTP,SCP,JMS等)，网络(WS,Socket,RPC,RSS,SOAP等)，流(Buffer, Row)
  - 有若干高性能组件如：外排序，批量插入(如SQLLoader)
  - 结构简单，只发布jar包

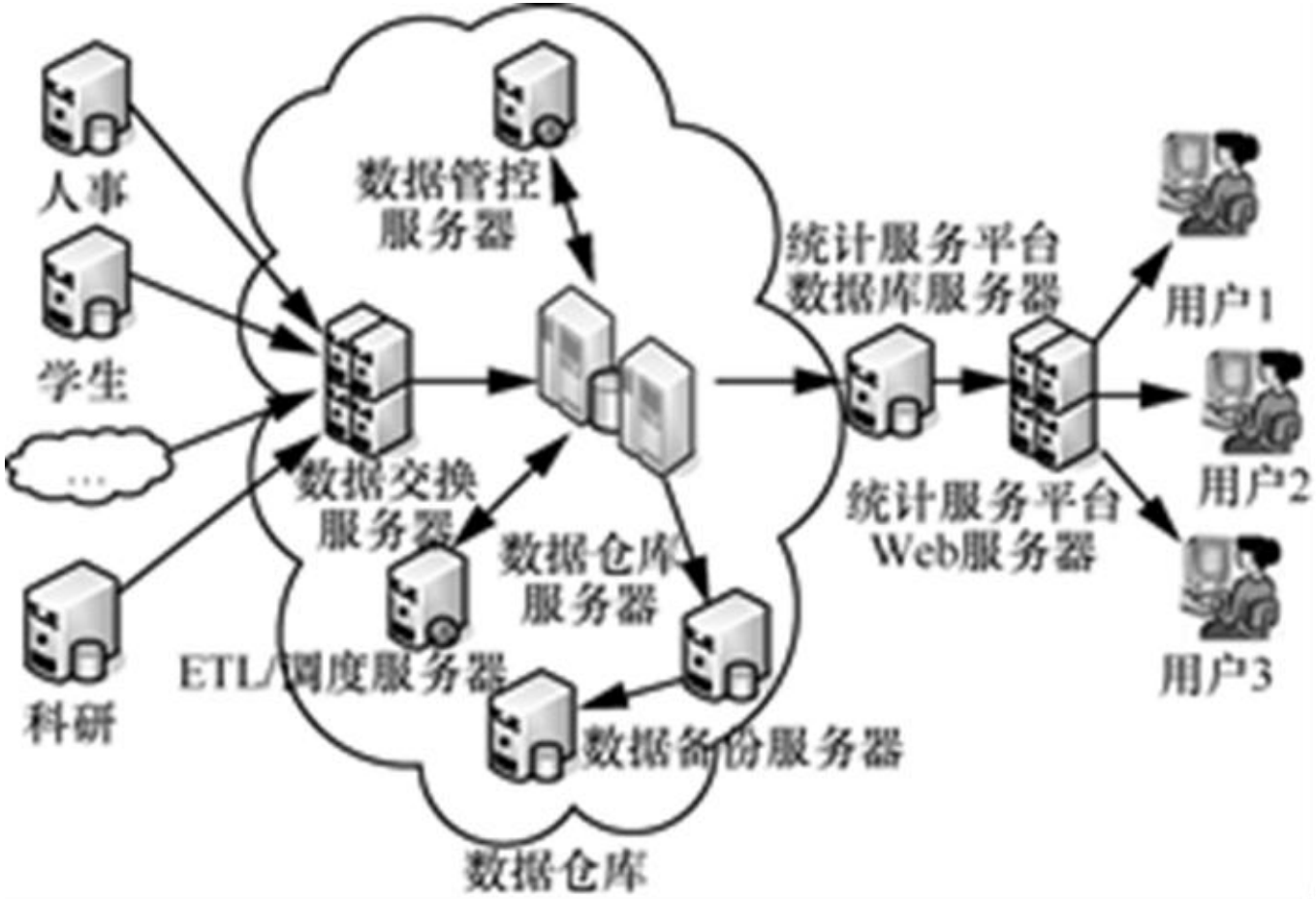
前台展示

- 常见的数据仓库的前端展示工具有**BO**、**Cognos**等，能基于**Web** 的直观界面，能提供报表、图表、仪表盘等多种展示方式。但都是商业产品，价格比较昂贵。
- **ExtJs**是一款开源的创建前端用户界面，是一个基本与后台技术无关的前端**ajax**框架，具有功能强大、编程简单的特点，数据统计服务平台的用户界面基于**ExtJS**开发。

前台展示界面


第一表：2013年上半年北京大学党员队伍状况统计表（一）——总人数
重新生成
保存数据
表内校验
表间校验
编辑批注
统计数据展示区
查询结果
导出
细节数据展示区

平台部署


人事
学生
...
科研
数据管控
服务器
数据交换
服务器
数据仓库
服务器
ETL/调度服务器
数据备份服务器
数据仓库
统计服务平台
数据库服务器
统计服务平台
Web服务器
用户1
用户2
用户3

结束语

- 提出了基于数据仓库技术的高校数据统计服务平台，通过合理的架构设计、科学的数据建模实现了对数据的集中存储、加工，以及统计数据生成、统计数据查询等功能。该系统能够有效地满足高校新形势下的业务发展需求，对于促进高校数据集约化管理水平的提升、搭建数据统筹管理和决策支持服务的长效机制框架具有十分重要的意义

谢谢！