数据仓库2012年4月16日

* 1. 解释名称：
* 数据仓库：数据仓库是一个按面向决策主题的方式组织的、集成的、相对稳定、随时间变化的，主要用于决策支持的数据集合。
* 元数据：元数据又称中介数据、中继数据，为描述数据的数据，主要是描述数据属性的信息，用来支持如指示存储位置、历史数据、资源查找、文件记录等功能。建立元数据是为了对数据库环境进行监控。
* ODS：操作型数据存储，是用于支持企业日常的全局应用的数据集合。

是企业信息系统中的一打雷重要的信息系统，这类信息系统以数据基础，通过数据统计、分析、挖掘、展现等手段为各层决策人员提供决策支持服务。

* ETL：是英文 Extract-Transform-Load 的缩写，用来描述将数据从来源端经过抽取（extract）、转换（transform）、加载（load）至目的端的过程。ETL是构建数据仓库的重要一环，用户从数据源抽取出所需的数据，经过数据清洗,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。
* 物化视图：在外观上像视图，在实质上，它是普通数据表，但是具有附加的逻辑，可能可以实现一些有效的计算功能，并提供不同刷新机制。它是用于预先计算并保存表连接或聚集等耗时较多的操作的结果，这样，在执行查询时，就可以避免进行这些耗时的操作，从而快速的得到结果。详见PPT5.4
* 数据模型：用于指导数据的设计、存储，并作为数据设计结果的检验标准的模型。
* 决策支持系统：DSS是企业信息系统中的一大类重要的信息系统，这类系统以数据为基础，通过数据统计、分析、挖掘、展现等手段为各层次决策人员 提供决策支持服务。
* OLAP：在线分析处理；用于支持复杂的分析操作；侧重对决策人员和高层管理人员的决策支持；快速、灵活地进行大数据量的复杂查询处理，以直观易懂的形式将查询结果提供给决策人员，使用户准确掌握企业的经营状况，了解市场需求，制定正确方案，增加效益。
* OLTP：联机事务处理系统；也称为面向交易的处理系统，其基本特征是顾客的原始数据可以立即传送到计算中心进行处理，并在很短的时间内给出处理结果。这样做的最大优点是可以即时地处理输入的数据，及时地回答。
* 关系数据库，是建立在关系模型基础上的数据库，借助于集合代数等数学概念和方法来处理数据库中的数据。
  1. 将如下概念串起来，描述概念之间的关系，每个概念不必单独解释。

企业信息系统架构、基础业务系统、数据仓库、数据挖掘、机器学习、RDBMS、HADOOP、Oracle、大数据、DB2、关系型数据库、多维分析引擎、数据集成、数据模型、元数据、决策支持系统。

早期的**企业信息系统架构**是单个数据库服务于所有目的。**基础业务系统**作为产生数据的系统，产生大量数据，信息型应用使用这些数据。这样的系统架构存在着很大的性能冲突，无法进行**数据挖掘**工作，即无法将操作型数据有效的转化为有用的信息。为了解决以上问题，引发了数据组织方法的变化，引入了合理架构的**数据仓库**。常见的数据仓库体系结构从下至上依次是：操作性数据库，操作数据存储，数据仓库，数据市场，应用交互。其中，操作性数据库以关系数据库为主，管理关系型数据库的软件成为**数据库管理系统**即**RDBMS**。常见的RDBMS有**Oracle**,**DB2**,SQLserver等。

为了对数据仓库环境进行监控，需要建立数据简历和**元数据**，利用**元数据**掌握**数据仓库**环境的情况。**数据仓库**是**数据集成**和高质量数据的最好来源。将数据从业务性应用集成到数据仓库时需要大量的数据处理。但是随着**大数据**时代的到来，传统的平台已无法处理如此庞大的数据，**Hadoop**生态圈基本上都是为了处理超过单机尺度的数据处理而诞生的。

**数据仓库**中的数据服务于决策分析的目的，也就是在数据仓库中建立起DSS（**决策支持系统**）应用。**数据模型**是**数据仓库**体系结构化环境的核心，指导数据的设计存储及检验标准。

以**关系型数据库**为基础的**数据仓库**不能很好地支持分析型应用逻辑，功能不够灵活，所以需要进一步研究归纳出通用的**数据模型**，实现**多维分析引擎**。

1. 简答题：
2. 请说明数据粒度和粒度设计的概念，主要考虑因素？粒度分析为什么要与技术环境评估同时进行？

答：数据粒度是描述数据环境中各种数据的细节程度或综合程度的高低的指标。数据细节程度越高，粒度级就越低，粒度越小；数据细节程度越低，粒度级就越高，粒度越大

粒度设计是数据仓库中一个重要的设计问题，影响整个平台的架构。因为它影响着存放在数据仓库中的数据量的大小，同时影响数据仓库所能回答的查询类型。粒度的大小需要数据仓库在设计时在数据量大小与查询的详细程度之间作出权衡。

考虑的因素：用户对数据的需求，要接受分析的类型，可接受数据的最低粒度，能够存储的数据量，信息占用数据仓库空间的大小。

粒度分析为什么要与技术环境评估同时进行？

技术环境评估指标包括：

 管理大数据量数据的能力

 进行灵活数据存取的能力

 根据数据模型重组数据的能力

 透明的数据发送和接收能力

 周期性成批装载数据的能力

 作业管理能力

这些指标都与数据有关，粒度是描述数据环境中各种数据的细节程度或综合程度的高低的指标，数据细节程度越高，粒度级就越低，粒度越小；数据细节程度越低，粒度级就越高，粒度越大。粒度级设计是数据仓库中一个重要的设计问题，影响整个平台的架构。

技术环境准备中需要进行软硬件配置的工作，主要考虑以下问题：预期在数据仓库上分析处理的数据量；进出数据仓库的数据通信量。而数据量和数据通信量都是在粒度设计中需要考虑的因素

1. 孤岛式和蜘蛛网信息系统架构？

蜘蛛网式的系统架构产生原因：1.组织架构与业务的发展，信息系统林立，业务交叉；2.在不同思维数据层上，存在大量的没有合理规划与控制的数据抽取程序；

信息孤岛的系统架构产生原因：在一个单位的各个部门之间由于种种原因造成部门与部门之间完全孤立，各种信息（如财务信息、各种计划信息等）无法或者无法顺畅地在部门与部门之间流动。

这两种架构存在的问题：1.数据的可信度低；2.生产效率低，开发新信息型应用的效率低；3.从数据到信息转换困难。

解决这些问题的常见方案：从数据应用架构的方法论角度作出调整。提出体系结构化的（成体系的）数据仓库

3. 请简述决策支持应用开发周期的特点，并说明与面向基础业务数据库应用系统的开发周期的主要区别。

* 决策支持系统开发周期的特点：不确定性 支持决策系统要完成的任务在系统开发前是不明确的，即需求不明确
* 开发周期的区别：

开发方法不同，通常的业务系统在开发之初开发人员需要与最终用户访谈获得需求，这些需求转变为系统设计和数据模型设计，

DSS则由于用户的需求不明确，这种变化的需求用户是参与整个系统的开发过程。

1. 有如下数据仓库数据模式设计活动，请将这些活动分别归入概念模型设计、逻辑模型设计和物理模型设计中。
2. 确定集成范围
3. 模式分解
4. 画出ERD
5. 确定数据存储位置
6. 编写建库脚本
7. 建立企业全局数据视图
8. 确定每个主题包含的实体集以及每个实体集的属性集
9. 确定粒度层次
10. 确定索引结构
11. 确定存储分配
12. 分析确定数据来源

概念：1/3/6/7：确定集成范围、画出ERD、建立企业全局数据视图、确定每个主题包含的实体集以及每个实体集的属性集

逻辑：2/8/11模式分解、确定粒度层次、分析确定数据来源

物理：确定数据存储位置、编写建库脚本、确定索引结构、确定存储分配

1. 请简述多维数据模型的目的，并说明在ROLAP架构中，多维数据模型与RDBMS中的关系型数据之间的关系

多维数据模型是为了满足用户从多角度多层次进行数据查询和分析的需要而建立起来的基于事实和维的数据库模型，其基本的应用是为了实现OLAP(Online Analytical Processing)。

当然，通过多维数据模型的数据展示、查询和获取就是其作用的展现，但其真的作用的实现在于，通过数据仓库可以根据不同的数据需求建立起各类多维模型，并组成数据集市开放给不同的用户群体使用，也就是根据需求定制的各类数据商品摆放在数据集市中供不同的数据消费者进行采购。

多维数据模型的功能目标：多维数据模型的功能是描述多维数据，有效支持多维分析操作。

**多维数据模型应该达到以下目标**

.逻辑上表示主题中的实体和实体间的关系；

.逻辑上表示主题各属性的定义域，即维的定义域；

.逻辑上应该能支持一定程度上的自动聚集逻辑，实现数据的动态计算，视图间的动态计算；

.逻辑上应该能实现有效的维护机制；

.能支持多维分析；

.支持分布式逻辑。

多维数据模型MDDM是一个三元组(F, DS, AIS)其中

 F为基事实表

 DS为定义于基事实表上的维集合

 AIS是定义于基事实表和DS上的聚集项集合。

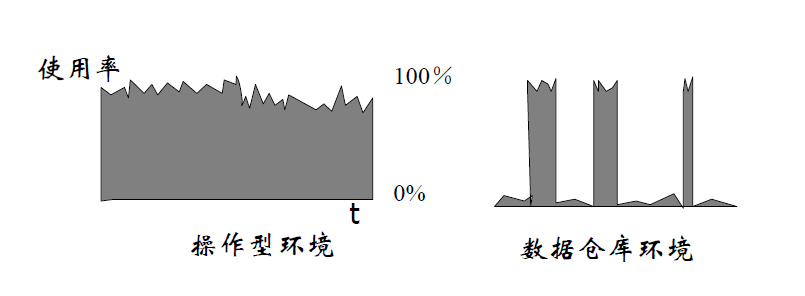
维表的主键用于连接事实表，作为事实表的外键，实现数据的参照完整性。所有维表将其他的一个或多个字段组成维表的自然键(natural key),一般来自源系统的键值。关系型数据：关系型数据是以关系数学模型来表示的数据。多维数据模型中以二维表（事实表）的形式来描述数据。

基于关系型数据库OLAP，或称为ROLAP以关系数据库为核心。维表和事实表都是用二维关系表的方式存放的。通过表来连接来获取事实。

ROLAP表示基于关系数据库的OLAP实现（Relational OLAP）。以关系数据库为核心,以关系型结构进行多维数据的表示和存储。ROLAP将多维数据库的多维结构划分为两类表:一类是事实表,用来存储数据和维关键字;另一类是维表,即对每个维至少使用一个表来存放维的层次、成员类别等维的描述信息。维表和事实表通过主关键字和外关键字联系在一起,形成了"星型模式"。对于层次复杂的维,为避免冗余数据占用过大的存储空间,可以使用多个表来描述,这种星型模式的扩展称为"雪花模式"。特点是将细节数据保留在关系型数据库的事实表中，聚合后的数据也保存在关系型的数据库中。这种方式查询效率最低，不推荐使用。

PPT 5.2.2维层次结构的物理表示。

6. .如下两个图标分别是操纵系统环境与数据仓库环境在一天内的硬件资源典型使用率模式，横坐标是时间，纵坐标是使用率（0-100%）



1. 请说明为什么这两种环境的硬件资源使用率会有如此不同？（5分）
2. 请说明这种现象对企业信息系统架构设计的影响？（5分）

PPT1.7：

操作性环境和数据仓库环境的很大不同点在于硬件的使用模式上。

在操作型环境中用户的行为特点是数据的存取操作频率高而每次操作处理的时间短，但因需要进行大量的数据存取因此对硬件的使用率高。

而在数据仓库环境中一般是进行查询操作用，查询数据用于分析，这类操作频率低但进行时需要消耗大量的系统资源。

对于操作型处理来说硬件使用利用模式相对稳定，可预测。

对于数据仓库处理或应用来说它的硬件使用模式相当不稳定。

硬件的使用模式不同说明，不应将两种应用混在一起，分开以后，可以针对不同的处理，分别进行优化处理。

7 设有如下数据仓库的数据模型设计\*\*，请

（6）建立企业全局数据视图；（7）确定每个主题包含的实体集以及每个实体集的属性集；（8）确定粒度层次；（9）确定索引结构；（10）确定存储分配；（11）分析确定数据来源。（11分）

8 .假设有3年的销售数据，有如下两种存储方案：

方案1：在数据库系统中用户看的表只有一个，但是有6个分区，每半年一个物理分区。

方案2：在数据库系统中有6个独立的表，分别存储近三年的数据。

1） 这两种方案哪个是数据库系统层分区，哪种是应用层分区？各有什么优缺点。（8）

2） 数据分区对数据归档、数据清除有什么影响？（4）

1是系统层分区，2是应用层分区；系统层分区：由 DBMS实现数据分区；应用层分区：由数据仓库应用平台实现数据的分区管理。一般在系统层分区的便于访问，少写访问代码，在应用层的访问如果有涉及多个表，则需要手工编写代码加以实现。但是，因为不同分区可能在结构上会有些不同，所以，也常常需要在应用层进行分区。 如果在系统层进行分区，则分区间的数据的结构必须是一样的，如果数据的时间中跨度很长，则数据定义很有可能会发生变化。这样的话，这种方法就不可行。多数情况，有可能是两种方法相结合。

数据分区时数据归档、数据清除有什么影响？

如果数据是分区存储的，在数据继承和查询时将会非常方便和高效，因为可以利用分区单独对自己感兴趣的那部分数据进行操作，读写与管理豆浆非常的方便高效。

9. 联机事务处理与联机分析处理

OLTP 联机事务处理 也称为面向交易的处理系统，其基本特征是顾客的原始数据可以立即传送到计算中心进行处理，并在很短的时间内给出处理结果。

这样做的最大优点是可以即时地处理输入的数据，及时地回答。也称为实时系统(Real time System)。衡量联机事务处理系统的一个重要性能指标是系统性能，具体体现为实时响应时间(Response Time)，即用户在终端上送入数据之后，到计算机对这个请求给出答复所需要的时间。OLTP是由数据库引擎负责完成的。

OLTP 数据库旨在使事务应用程序仅写入所需的数据，以便尽快处理单个事务。

联机分析处理（OLAP）系统是数据仓库系统最主要的应用，专门设计用于支持复杂的分析操作，侧重对决策人员和高层管理人员的决策支持，可以根据分析人员的要求快速、灵活地进行大数据量的复杂查询处理，并且以一种直观而易懂的形式将查询结果提供给决策人员，以便他们准确掌握企业（公司）的经营状况，了解对象的需求，制定正确的方案。

OLTP是传统的关系型数据库的主要应用，主要是基本的、日常的事务处理，例如银行交易。OLAP是数据仓库系统的主要应用，支持复杂的分析操作，侧重决策支持，并且提供直观易懂的查询结果。

OLTP 系统强调数据库内存效率，强调内存各种指标的命令率，强调绑定变量，强调并发操作；

OLAP 系统则强调数据分析，强调SQL执行市场，强调磁盘I/O，强调分区等。

10概念模型设计、逻辑模型设计、物理模型设计