***2015年复习材料***

**数据仓库：**面向主题的、集成的、非易失性和时变性的勇于支持管理决策的数据集合。

**DSS:**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。

------------------------------------------------------------------------------------------------------------------------------

**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、概念题：请将如下概念串起来，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

答：

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**。

---------------------------------------------------------------------------------------------------------------------------------

2、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。

具体不同，见PPT，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。

-------------------------------------------------------------------------------------------------------------------------------

3、对于企业常见的 孤岛式信息系统架构和蜘蛛网信息系统利用架构

（1）分别给出这两种架构产生的原因；1.2.5节和1.2.6节

（2）指出这两种架构中存在的主要问题；

答：1.3节：

数据的可信度低：部门之间数据不一致，不同步；

开发效率低：数据定位存在问题，数据编辑处理的问题；

从数据到信息转化困难：以现有的、分离的、缺少数据集成的平台为基础，要将数据转化成有用的信息存在很大的问题。

（3）给我解决这些问题的常见方案；

答：从数据应用构架组织的方法论角度作出调整。需要建立一个具有合理架构的数据仓库，在数据仓库中包含两类数据：原始数据（原始业务数据，操作型数据，业务型数据），导出数据（派生数据，决策支持数据）

-------------------------------------------------------------------------------------------------------------------------------

4、决策支持应用开发周期特点，并说明与面向基础业务的数据库应用系统的开发周期的主要区别。

决策支持应用开发是工作于发现模式，DSS应用开发的大致步骤见PPT3.3.6 -140页

3.3.6节：DSS应用开发与OLTP应用的区别：

DSS应用开发从数据出发；应用需求不能在开发初期明确了解；是一个不断循环的过程；

而OLTP数据库设计：有一组较为确定的应用需求，作为数据库系统设计和开发的出发点和基础；以明确的物流、数据流和数据处理流形式表示。

-------------------------------------------------------------------------------------------------------------------------------

5、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。

答：多维数据模型的目的：PPT5.1—16页：-🡪功能目标；

**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。

------------------------------------------------------------------------------------------------------------------------------

概念题：

1. **数据仓库**

数据的集合，用于管理决策，面向主题设计的，存储的一般是历史数据，有意引入冗余，采用反范式的方式来设计。是为分析数据而设计。数据仓库，是在数据库已经大量存在的情况下，为了进一步挖掘数据资源、为了决策需要而产生的，它决不是所谓的“大型数据库”。面向主题的、集成的、与时间相关且不可修改的数据集合。

1. **元数据**

元数据（Metadata），又称中介数据、中继数据，为描述数据的数据（data about data），主要是描述数据属性（property）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。

1. **ODS**

第四章：ODS是用于支持企业日常的全局应用的数据集合。

具体见 PPT4.2.1

1. **ETL**

ETL:是英文 Extract-Transform-Load 的缩写，用来描述将数据从来源端经过取（extract）、转换（transform）、加载（load）至目的端的过程。ETL一词较常用在数据仓库，但其对象并不限于数据仓库。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，经过数据清理,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。是一种数据仓库技术。

1. **物化视图**

详见PPT5.4

物化视图是包括一个查询结果的数据库对像，它是远程数据的的本地副本，或者用来生成基于数据表求和的汇总表。物化视图存储基于远程表的数据，也可以称为快照。

物化视图可以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，而从快速的得到结果。物化视图有很多方面和索引很相似：使用物化视图的目的是为了提高查询性能；物化视图对应用透明，增加和删除物化视图不会影响应用程序中SQL语句的正确性和有效性；物化视图需要占用存储空间；当基表发生变化时，物化视图也应当刷新。

物化视图可以查询表，视图和其它的物化视图。

1. **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。

1. **决策支持系统**

决策支持系统(Decision Support System ，简称DSS)是辅助决策者通过数据、模型和知识，以人机交互方式进行半结构化或非结构化决策的计算机应用系统。提供决策支持，自己不做决策。

1. **OLTP和OLAP**

OLTP:On-Line Transaction Processing联机事务处理系统,时间要求高。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。

OLAP:联机分析处理，时间要求宽松。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。

6、说明数据粒度和粒度设计的概念，在数据利用环境的数据体系建设过程中，进行粒度设计时主要需要考虑的因素有哪些？

PPT:2.5

2.5数据粒度：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。

粒度设计：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。粒度级设计影响了存放在数据库中数据量的大小，也影响数据仓库所能回答的查询类型，合理设计粒度级，达到有效存储数据，有效访问数据，有效分析细节数据等目标。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。所以设计粒度级时主要考虑的因素有：考虑存储空间、访问效率、权衡问题的覆盖面和回答问题的效率。从而设计合适的多粒度级。

-------------------------------------------------------------------------------------------------------------------------------