# 数据仓库与数据挖掘的一些基本概念

## 名词

**BI（Business Intelligence）：商业智能**

**DW（Data Warehouse）：数据仓库**

**OLTP（On-Line Transaction Processing）: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能够马上传送到计算机中心进行处理，并在非常短的时间内给出处理结果。这样做的最大长处是能够即时的处理输入的数据，及时的回答。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，详细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须要的时间。

**OLTP数据库**旨在使事务应用程序仅写入所需要的数据，以便尽快处理单个事务。

**OLAP（On-Line Analytical Processing）：联机分析处理**

OLAP是Codd于1993年提出的。

**OLAP理事会的定义**：OLAP是一种软件技术，它使分析人员可以迅速、一致、交互的从各个方面观察信息，以达到深入理解数据的目的，这些信息是从原始数据直接转换过来的，他们以用户easy理解的方式反映企业的真是情况。

**OLAP**大部分策略都是将关系型的或普通的数据进行多维数据存贮，以便于进行分析，从而达到联机分析处理的目的。这样的多维DB也被看做一个超立方体，沿着各个维方向存贮数据，他同意用户沿事物的轴线方便的分析数据，与主流业务型用户相关的分析形式一般有切片和切块以及下钻、挖掘等操作。

**DataMart：数据集市**，为了特定的应用目的或应用范围，而从数据仓库中独立出来的一部分数据，也可称为部门数据或主题数据（subject area）。在数据仓库的实施过程中往往能够从一个部门的数据集市着手，以后再用几个数据集市组成一个完整的数据仓库。须要注意的就是在实施不同的数据集市时，同一含义的字段定义一定要相容，这样在以后实施数据仓库时才不会造成大麻烦。

**Data Mining：数据挖掘。**

**ETL**：ETL各自是“Extract”，“Transform”，“Load”三个单词的首字母缩写，也就是“抽取”，“转换”，“装载”，但我们日常往往简称其为数据抽取。ETL是BI/DW（商务智能/数据仓库）的核心和灵魂，依照统一的规则集成并提高数据的价值，是负责完毕数据从数据源向目标数据仓库转换的过程，是实施数据仓库的重要步骤。

**MetaData：元数据，**元数据是描写叙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，技术元数据和商业元数据。技术元数据是数据仓库的设计和管理人员用于开发和日常管理数据仓库时用的数据。包含：数据源信息，数据转换的描写叙述，数据仓库内对象和数据结构的定义，数据清理和数据更新时用的规则，源数据到目的数据的映射，用户访问权限，数据备份历史记录，数据导入历史记录，信息公布历史记录等。

**商业元数据**从商业业务的角度描写叙述了数据仓库中的数据。包括：业务主题的描写叙述，包括的数据、查询、报表。

**元数据**为访问数据仓库提供了一个信息文件夹（information directory），这个文件夹全面描述了数据仓库中都有什么数据、这些 数据怎么得到的、和怎么访问这些数据。是数据仓库执行和维护的中心，数据仓库server利用他来存贮和更新数据，用户通过它来了解和访问数据。

## Q1：什么是数据仓库

数据仓库是一个面向主题的（Subject Oriented）、集成的（Integrate）、相对稳定的（NonVolatile）、反映历史变化（Time Variant）的数据集合，用于支持管理决策。对于数据仓库的概念我们能够从两个层次予以理解：①数据仓库用于支持决策，面向分析型数据处理，它不同于企业现有的操作型数据库；②数据仓库是对多个异构数据源的有效集成，集成后依照主题进行了重组，并包括历史数据，并且存放在数据仓库中的数据一般不在改动。企业数据仓库的建设是以现有企业业务系统和大量业务数据的积累为基础。数据仓库不是静态的概念，仅仅有把信息及时交给须要这些信息的使用者，供他们做出改善其业务经营的决策，信息才发挥作用，信息才有意义。而把这些信息加以整理、归纳和重组，并及时提供给对应的管理决策人员是数据仓库的根本任务。

## Q2：为什么要建立数据仓库

企业建立数据仓库是为了填补现有数据存储形式已经不能满足信息分析的需要。数据仓库理论中的一个核心理念就是：事务型数据和决策支持型数据的处理性能不同。

企业在他们的事务操作收集数据。在企业运作过程中：随着订货、销售记录的进行，这些事务型数据也连续产生。为了引入数据，我们须优化事务型数据库。

处理决策支持型数据时，一些问题常常会被提出：哪类客户会购买哪类产品？促销后销售额会变化多少？价格变化后或者商店地址变化后销售额又会变化多少呢？在某一段时间内，相对其他产品来说哪类产品特别easy卖呢？哪些客户添加了他们的购买额？哪些客户又削减了他们的购买额呢？

事务型数据库能够为这些问题做出解答，可是他所给的答案往往并不能让人十分满意。在运用有限的计算机资源时经常存在着竞争。在添加 新信息的时候我们需要事务性数据库是空暇的。而在解答一系列详细的有关信息分析的问题的时候，系统处理新数据的有效性又会被大大减少。还有一个问题就在于事务型数据总是在动态变化之中的。决策支持型处理需要相对稳定的数据，从而问题都能得到一致连续的解答。

数据仓库的解决方法包含：将决策支持型数据处理从事务型数据处理中分离出来。数据依照一定的周期（通常在每晚或每周末），从事务型数据库中导入决策支持型数据库—即数据仓库。数据仓库是按回答企业某方面的问题来分“主题”组织数据的，这是最有效的数据组织方式。

另外，企业日常运作的信息系统通常是由多个传统系统、不兼容数据源、数据库与应用所共同构成的复杂数据集合，各个部分之间不能彼此交流。从这个层面看：眼下执行的应用系统是用户花费了非常大精力和财力构建的、不可替代系统，特别是系统的数据。而建立数据仓库的目的就是要把这些不同来源的数据整合组织起来统一管理，从而做到数据的一致性与集成化，提供一个全面的、单一入口的解决方式。这个让我想到了SOA的理念，只是前者是数据层面的整合优化，后者是应用服务层面的整合优化。

## Q3：什么是数据挖掘

数据挖掘（data mining）就是从大量的、不全然的、有噪声的、模糊的、随机的实际应用数据中，提取隐含在当中的、人们事先不知道的、但又是潜在实用的信息和知识的过程。

数据挖掘是一门交叉学科，他把人们对数据的应用从低层次的简单查询，提升到从数据中挖掘知识，提供决策支持。在这样的需求牵引下，汇聚了不同领域的研究者，尤其是数据库技术、人工智能技术、数理统计、可视化技术、并行计算等方面的学者和project人员，投身到数据挖掘这一新兴的研究领域，形成新的技术热点。