数 据 仓 库 概 念 的 简 单 理 解

一个典型的企业数据仓库系统通常包含数据源、 数据存储与管理、 OLAP服务器以及前端工具与应用四个部分。如下图所示：

数据源：

是数据仓库系统的基础，是整个系统的数据源泉。通常包括企业内部信息和外部信息。 内部信息包括存放于企业操作型数据库中 （通常存放在 RDBMS中）的各种业务数据和办公自动化 （OA）系统包含的各类文档数据。 外部信息包括各类法律法规、市场信息、竞争对手的信息以及各类外部统计数据及各类文档等；

数据的存储与管理：

是整个数据仓库系统的核心。 在现有各业务系统的基础上， 对数据进行抽取、清理，并有效集成， 按照主题进行重新组织， 最终确定数据仓库的物理存储结构，同时组织存储数据仓库元数据（具体包括数据仓库的数据字典、记录系统定义、数据转换规则、数据加载频率以及业务规则等信息）。按照数据的覆盖范围，数据仓库存储可以分为企业级数据仓库和部门级数据仓库 （通常称为“ 数据集市 ”，Data Mart ）。数据仓库的管理包括数据的安全、归档、备份、维护、恢复等工作。这些功能与目前的 DBMS基本一致。

OLAP服务器：

对分析需要的数据按照多维数据模型进行再次重组，以支持用户多角度、多层次的分析，发现数据趋势。其具体实现可以分为： ROLAP、MOLAP和 HOLAP。ROLAP基本数据和聚合数据均存放在 RDBMS之中；MOLAP基本数据和聚合数据均存放于多维数据库中；而 HOLAP是 ROLAP与 MOLAP的综合，基本数据存放于 RDBMS之中，聚合数据存放于多维数据库中。

前端工具与应用：

前端工具主要包括各种数据分析工具、报表工具、查询工具、数据挖掘工具以及各种基于数据仓库或数据集市开发的应用。其中数据分析工具主要针对OLAP服务器，报表工具、数据挖掘工具既针对数据仓库，同时也针对 OLAP服务器。？

集线器与车轮状结构的企业级数据仓库 ？

这种结构也称为“ Hub and Spoke ”，这是因为中央数据库汇集了来自各业务处理系统的数据，同时也负责向各从属数据集市提供信息，看上去像一个 Hub (集线器) ;而业务人员在进行数据分析与信息访问时将根据需要连接到不同的数据集市，这种交叉复杂的连接看上去就像 Spoke(车轮辐条) 一样。 “ Hub and Spoke ”结构解决了企业内统一数据存储模型的问题，但从实际使用的角度来看仍有比较严重的缺陷： 一是业务人员对信息的访问非常不方便， 很难进行跨数据集市或跨部门的信息分析；另一个问题是每个数据集市都需要相应的软硬件投入，当数据集市增加时， 系统整体投资迅速增加， 同时管理的复杂性也随之增加。这些都意味着巨大的整体拥有成本 TCO(Total Cost of Ownership) 。

为什么不直接访问中央数据仓库而非要设计一个数据集市层呢？主要原因在于当中央数据库保存越来越多的数据、 并发用户越来越多时， 一般的数据库引擎无法承担这样的负载，只好把它们分解到不同的数据集市。对于“ Hub and Spoke ”结构的数据仓库， Gartner Group 也认为， “ 数据仓库的 Hub and Spoke 结构，回避了 DBMS 技术中的弱点，无法提供适当的业务价值来平衡投资成本的显着增加 ”，“ 之所以产生这种趋势，是由于对大多数 DBMS 产品而言，支持复杂的数据模型和并发查询负载都是极大的挑战 ”。 ？

集中式企业级数据仓库

第二种企业级数据仓库的架构是集中式的，这解决了“ Hub and Spoke ”结构中存在的诸多问题， 是一种比较理想的企业级数据仓库系统架构， 能够为企业带来真正的业务价值与回报。 但由于把详细数据分析、 部分的数据转换与清洗等复杂处理均集中在中央数据仓库， 从而给作为数据仓库引擎的 RDBMS 和相应的服务器带来了极大的挑战。选择这种数据仓库基础平台的基本要求是： ？

1、线性扩展能力。原始数据对任何一个数据仓库来说，都是最主要的负载之一。随着数据量的增长， 系统性能会逐渐下降。 维持合理的业务查询响应时间，要求数据仓库引擎和相应的数据库服务器具有优良的线性扩展能力。 一些系统的扩展能力非常有限，当数据量增长到一定规模时（比如 TB 级以上），就很难满足日常的业务分析要求了， 因而不得不把数据分离到多个小规模的数据集市， 形成所谓的“ Hub and Spoke ”结构。

2、并行处理能力。许多业务查询与分析都是动态的，数据库传统的索引技术对动态分析和模糊查询的帮助不大。 系统只有具有非常好的并行处理能力， 才能满足复杂的、动态的分析需求，并且承担比较复杂的数据转换与清洗工作。

3、简单的系统管理。对于大型的数据仓库应用系统而言，如何能有效而简单地进行系统管理是非常重要的。 特别是当数据量不断扩大时， 如果没有一种有效而且简单的系统管理措施，那么系统的运行费用将会很高。

数据仓库的实施是一个长期的过程，在基础设施建立完成后，随着应用的逐步开展和深入， 其投资回报也逐步增加。 在建立数据仓库过程中需要一定的时间来建立数据仓库基础设施， 并在建置的过程中逐步完善数据质量。 这个打基础的过程是无法省略的。 更为重要的是， 在建立数据仓库的过程当中， 还可以培养一批既懂数据仓库技术、 又精通该领域业务的高级分析人才， 这对于更好地发挥数据仓库价值是非常重要的

附：联机事务处理 OLTP及联机分析处理（ OLAP）？

当今的数据处理大致可以分成两大类：联机事务处理 OLTP（on-line transaction processing ）、联机分析处理 OLAP（On-Line Analytical Processing ）。 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。 下表列出了 OLTP与 OLAP之间的比较。

	OLTP	OLAP
用户	操作人员，低层管理人员	决策人员，高级管理人员
功能	日常操作处理	分析决策
DB 设计	面向应用	面向主题
数据	当前的， 最新的细节的 ， 二维的分立的	历史的， 聚集的， 多维的集成的， 统一的
存取	读/ 写数十条记录	读上百万条记录
工作单位	简单的事务	复杂的查询
用户数	上千个	上百个

DB 大小	100MB-GB	100GB-TB

?OLAP是使分析人员、管理人员或执行人员能够从多角度对信息进行快速、一致、交互地存取，从而获得对数据的更深入了解的一类软件技术。 OLAP的目标是满足决策支持或者满足在多维环境下特定的查询和报表需求，它的技术核心是"维"这个概念。

“维”是人们观察客观世界的角度，是一种高层次的类型划分。“维”一般包含着层次关系，这种层次关系有时会相当复杂。通过把一个实体的多项重要的属性定义为多个维（dimension），使用户能对不同维上的数据进行比较。因此OLAP也可以说是多维数据分析工具的集合。

OLAP的基本多维分析操作有钻取（roll up 和 drill down）、切片（slice）和切块（dice）、以及旋转（pivot）、drill across、drill through 等。

· 钻取是改变维的层次，变换分析的粒度。它包括向上钻取（roll up）和向下钻取（drill down）。roll up 是在某一维上将低层次的细节数据概括到高层次的汇总数据，或者减少维数；而 drill down 则相反，它从汇总数据深入到细节数据进行观察或增加新维。

切片和切块是在一部分维上选定值后，关心度量数据在剩余维上的分布。如果剩余的维只有两个，则是切片；如果有三个，则是切块。

旋转是变换维的方向，即在表格中重新安排维的放置（例如行列互换）。

OLAP有多种实现方法，根据存储数据的方式不同可以分为 ROLAP、MOLAP、HOLAP。

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"。

MOLAP表示基于多维数据组织的 OLAP实现（Multidimensional OLAP ）。以多维数据组织方式为核心 , 也就是说 ,MOLAP使用多维数组存储数据。 多维数据在存储中将形成 " 立方块（Cube）" 的结构 , 在 MOLAP中对 " 立方块 " 的" 旋转 "、" 切块 "、" 切片 " 是产生多维数据报表的主要技术。

HOLAP表示基于混合数据组织的 OLAP实现（Hybrid OLAP）。如低层是关系型的，高层是多维矩阵型的。这种方式具有更好的灵活性。

还有其他的一些实现 OLAP的方法，如提供一个专用的 SQL Server，对某些存储模式（如星型、雪片型）提供对 SQL查询的特殊支持。

OLAP工具是针对特定问题的联机数据访问与分析。 它通过多维的方式对数据进行分析、查询和报表。维是人们观察数据的特定角度。例如，一个企业在考虑产品的销售情况时， 通常从时间、 地区和产品的不同角度来深入观察产品的销售情况。这里的时间、 地区和产品就是维。 而这些维的不同组合和所考察的度量指标构成的多维数组则是 OLAP分析的基础，可形式化表示为（维 1，维 2，……，维 n，度量指标），如（地区、时间、产品、销售额）。多维分析是指对以多维形式组织起来的数据采取切片（ Slice ）、切块（ Dice）、钻取（ Drill-down 和 Roll-up ）、旋转（Pivot ）等各种分析动作，以求剖析数据，使用户能从多个角度、多侧面地观察数据库中的数据，从而深入理解包含在数据中的信息。

根据综合性数据的组织方式的不同， 目前常见的 OLAP主要有基于多维数据库的 MOLAP及基于关系数据库的 ROLAP两种。MOLAP是以多维的方式组织和存储数据， ROLAP则利用现有的关系数据库技术来模拟多维数据。在数据仓库应用中，OLAP应用一般是数据仓库应用的前端工具， 同时 OLAP工具还可以同数据挖掘工具、统计分析工具配合使用，增强决策分析功能。

附:OLAP主流产品

?Hyperion Essbase

?Oracle Express

?IBM DB2 OLAP Server

?Sybase Power dimension

?Informix Metacube

Hyperion Essbase

?以服务器为中心的分布式体系结构

?有超过 100 个的应用程序

?有 300 多个用 Essbase 作为平台的开发商

?具有几百个计算公式，支持多种计算

?用户可以自己构件复杂的查询。

?快速的响应时间，支持多用户同时读写

?有 30 多个前端工具可供选择

?支持多种财务标准

?能与 ERP或其他数据源集成

?全球用户超过 1500 家

?Oracle Express

?Oracle DW 支持 GB~TB数量级

?采用类似数组的结构，避免了连接操作，提高分析性能

?提供一组存储过程语言来支持对数据的抽取

?用户可通过 Web和电子表格使用

?灵活的数据组织方式，数据可以存放在 Express Server 内，也可直接在 RDB 上使用

?有内建的分析函数和 4GL用户自己定制查询

?全球超过 3000 家

?IBM DB2 OLAP Server

– 把 Hyperion Essbase 的 OLAP引擎和 DB2的关系数据库集成在一起。

– 与 Essbase API 完全兼容

– 数据用星型模型存放在关系数据库 DB2中

?Informix Metacube

– 采用 metacube 技术，通过 OLE和 ODBC对外开放，

– 采用中间表技术实现多维分析引擎，提高响应时间和分析能力

– 开放的体系结构可以方便地与其他数据库及前台工具进行集成

?Sybase Power dimension

– 数据垂直分割（按“ 列 ”存储）

– 采用了突破性的数据存取方法 ------bit-wise 索引技术

– 在数据压缩和并行处理方面有多到之处

– 提供有效的预连接（ Pro-Jion ）技术