


数据仓库基本知识\_直到世界的尽头

原创 张小凡vip 2017-10-31 17:35:04 18712 已收藏 83 版权

分类专栏： 数据分析 文章标签： 数据仓库



**spark on k8s**  
spark on k8s的最新实践和更新，云计算和大数据的美妙结合，因为目前网上资料较少，官网的步骤并不详细，具体步...  
张小凡vip

¥9.90  
订阅博主

数据仓库是什么

根据统计，每个企业的数据量每2~3年时间就会成倍增长，这些数据蕴含着巨大的商业价值，而企业所关注的通常只占在总数据量的2%~4%左右。因此，企业仍然没有最大化地利用已存在的数据资源，以至于浪费了更多的时间和资金，也失去制定关键商业决策的最佳契机。

于是，企业如何通过各种技术手段，并把数据转换为信息、知识避免各种无知状态和瞎猜行为，已经成了提高其核心竞争力的主要瓶颈。

数据仓库是把数据转换为信息、知识的一种主要技术手段。

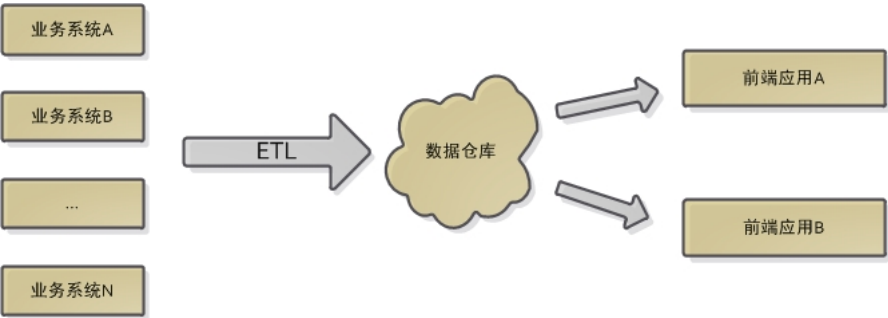
数据仓库是面向分析的存储系统

数据仓库，是企业所有级别的决策制定过程，提供所有类型数据支持的数据集合。这些数据集合出于分析性报告和决策支持目的而创建，用于支持研究管理决策。

一是为调查研究作数据支撑，二是为实现需要业务智能的企业，提供指导业务流程改进、监视时间、成本、量以及控制。

数据仓库是一个过程而不是一个项目；数据仓库是一个环境，而不是一件产品。

数据仓库提供用户用于决策支持的当前和历史数据，这些数据在传统的操作型数据库中很难或不能得到。



目标和DEMO

将联机事务处理(OLTP)经年累月所累积的大量数据资料，透过数据仓库理论所特有的资料储存架构，做数据的清理保存，提供给各种分析方法使用，如联机分析处理(OLAP)、数据挖掘(Data Mining)，并进而创建 决策支持系统(DSS)、主管资讯系统(EIS)、研究支持系统，帮助决策者研究者快速有效的自大量资料中，分析出有价值的资讯，能够快速回应外在环境变动，帮助建构商业智能(BI)，挖掘内部数据价值，产生更多高质量的内容。

数据仓库给组织带来了巨大的变化。数据仓库的建立给企业带来了一些新的工作流程，其他的流程也因此而改变。

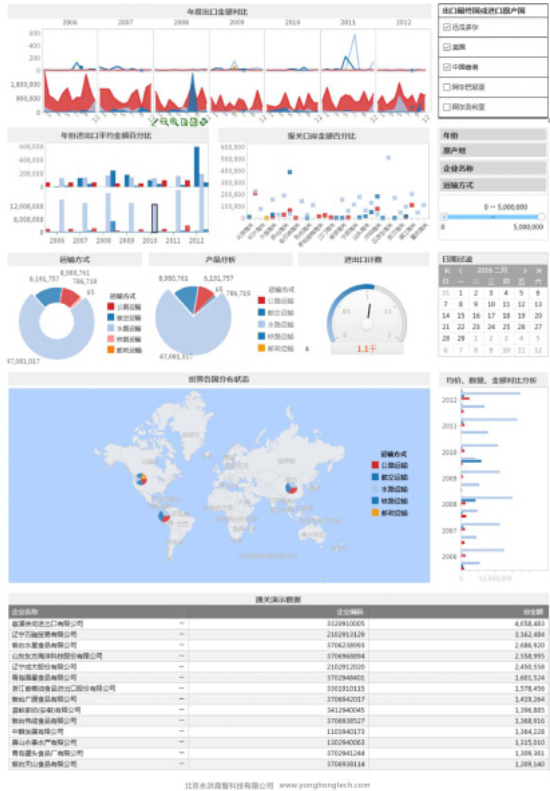
数据仓库为企业带来了一些“以数据为基础的知识”，它们主要应用于对市场战略的评价，和为企业发现新的市场商机，同时，也用来控制库存、检查生产方法和定义客户群。

通过数据仓库，可以建立企业的数据模型，这对于企业的生产与销售、成本控制与收支分配有着重要的意义，极大的节约了企业的成本，提高了经济效益，同时，用数据仓库可以分析企业人力资源与基础数据之间的关系，可以用于返回分析，保障人力资源的最大化利用，亦可以进行人力资源绩效评估，使得企业管理更加科学合理。数据仓库将企业的数据按照特定的方式组织，从而产生新的商业知识

业智能产品巨头形成了挑战。

国内BI产品起步较晚，知名的敏捷型商业智能产品有PowerBI, 永洪科技的Z-Suite, SmartBI, FineBI商业智能软件等，其中永洪科技的Z-Data Mart是一款热内存计算的数据集市产品。

国内的德昂信息也是一家数据集市产品的系统集成商。



## 阿里数加

<https://data.aliyun.com/>

点赞 13

评论<sup>4</sup>

分享

★ 已收藏<sup>83</sup>

手机看

⌘ 打赏

...

订阅博主

[https://blog.csdn.net/zzq900503/article/details/78405958?ops\\_request\\_misc=%257B%2522request%255Fid%2522%253A%2522160093681119...](https://blog.csdn.net/zzq900503/article/details/78405958?ops_request_misc=%257B%2522request%255Fid%2522%253A%2522160093681119...) 2/13

## 数据仓库的特点

数据仓库是一个面向主题的（ Subject Oriented ）、集成的（ Integrate ）、相对稳定的（ Non-Volatile ）、反映历史变化（ Time Variant ）的数据集合

### 1、面向主题

操作型数据库的数据组织面向事务处理任务，各个业务系统之间各自分离，而数据仓库中的数据是按照一定的主题域进行组织的。

### 2、集成的

数据仓库中的数据是在对原有分散的数据库数据抽取、清理的基础上经过系统加工、汇总和整理得到的，必须消除源数据中的不一致性，以保证数据仓库内的信息是关于整个企业的一致的全局信息。

### 3、相对稳定的

数据仓库的数据主要供企业决策分析之用，所涉及的数据操作主要是数据查询，一旦某个数据进入数据仓库以后，一般情况下将被长期保留，也就是数据仓库中一般有大量的查询操作，但修改和删除操作很少，通常只需要定期的加载、刷新。

### 4、反映历史变化

数据仓库中的数据通常包含历史信息，系统记录了企业从过去某一时点(如开始应用数据仓库的时点)到目前的各个阶段的信息，通过这些信息，可以对企业的发展历程和未来趋势做出定量分析和预测。

### 5、效率足够高

数据仓库的分析数据一般分为日、周、月、季、年等，可以看出，日为周期的数据要求的效率最高，要求24小时甚至12小时内，客户能看到昨天的数据分析。由于有的企业每日的数据量很大，设计不好的数据仓库经常会出问题，延迟1-3日才能给出数据，显然不行的。

## 数据仓库技术

数据仓库技术是为了有效的把操作形数据集成到统一的环境中以提供决策型数据访问的各种技术和模块的总称。所做的一切都是为了让用户更快更方便查询所需要的信息，提供决策支持。

从功能结构划分，数据仓库系统至少应该包含数据获取（ Data Acquisition ）、数据存储（ Data Storage ）、数据访问（ Data Access ）三个关键部分。

## 数据获取

### 数据源

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RDBMS中的各种业务处理数据和各类文档数据。外部信息包括各类法律法规、市场信息和竞争对手的信息等等；

### 元数据

对业务数据本身及其运行环境的描述与定义的数据，称之为元数据（ metadata ）。元数据是描述数据的数据。元数据的典型表现为对象的描述，即对数据库、表、列、列属性（类型、格式、约束等）以及主键/外部键关联等等的描述。特别是现行应用的异构性与分布性越来越普遍的情况下，统一的元数据就愈发重要了。“信息孤岛”曾经是很多企业对其应用现状的一种抱怨和概括，而合理的元数据则会有效地描绘出信息的关联性。而元数据对于ETL的集中表现为：定义数据源的位置及数据源的属性、确定从源数据到目标数据的对应规则、确定相关的业务逻辑、在数据实际加载前的其他必要的准备工作，等等，它一般贯穿整个数据仓库项目，而ETL的所有过程必须最大化地参照元数据，这样才能快速实现ETL。

## 数据转换工具

1)数据转换工具要能从各种不同的数据源中读取数据。2)支持平面文件、索引文件、和legacyDBMS。3)能以不同类型数据源为输入整合数据。4)具有规范的数据访问接口5)最好具有从数据字典中读取数据的能力6)工具生成的代码必须是在开发环境中可维护的7)能只抽取满足指定条件的数据，和源数据的指定部分8)能在抽取中进行数据类型转换和字符集转换9)能在抽取的过程中计算生成衍生字段10)能让数据仓库管理系统自动调用以定期进行数据抽取工作，或能将结果生成平面文件11)必须对软件供应商的生命力和产品支持能力进行仔细评估 主要数据抽取工具供应商：Prismsolutions. Carleton'sPASSPORT. InformationB

点赞 13

评论 4

分享

已收藏 83

手机看

打赏

...

订阅博主

ETL分别代表：提取extraction、转换transformation、加载load。

其中提取过程表示操作型数据库搜集指定数据，转换过程表示将数据转化为指定格式并进行数据清洗保证数据质量，加载过程表示将转换过后满足指定格式的数据加载进数据仓库。

数据仓库会周期不断地从源数据库提取清洗好了的数据，因此也被称为"目标系统"；

实现ETL，首先要实现ETL转换的过程。体现为以下几个方面：

- 1、空值处理：可捕获字段空值，进行加载或替换为其他含义数据，并可根据字段空值实现分流加载到不同目标库。
- 2、规范化数据格式：可实现字段格式约束定义，对于数据源中时间、数值、字符等数据，可自定义加载格式。
- 3、拆分数据：依据业务需求对字段可进行分解。例，主叫号 861082585313-8148，可进行区域码和电话号码分解。
- 4、验证数据正确性：可利用Lookup及拆分功能进行数据验证。例如，主叫号 861082585313-8148，进行区域码和电话号码分解后，可利用Lookup返回主叫网关或交换机记载的主叫地区，进行数据验证。
- 5、数据替换：对于因业务因素，可实现无效数据、缺失数据的替换。
- 6、Lookup：查获丢失数据 Lookup实现子查询，并返回用其他手段获取的缺失字段，保证字段完整性。
- 7、建立ETL过程的主外键约束：对无依赖性的非法数据，可替换或导出到错误数据文件中，保证主键唯一记录的加载。

根据以往数据仓库项目的经验，在一个数据仓库项目中，ETL设计和实施的工作量一般要占总项目工作量的40%-60%，而且数据仓库项目一般会存在二次需求的问题，客户在项目的实施过程中或者使用过程中会提出新的业务需求，而任何前端业务模型的改变都会涉及到ETL设计，因此ETL工具的选择对于整个数据仓库项目的成功是非常重要的。

## 选型

ETL工具的典型代表有:Informatica powercenter、Datastage、Oracle OWB(oracle warehouse builder)、ODI、微软DTS、Beeload、Kettle、Talend、DataSprider、Spark、等等.....

开源的工具具有eclipse的etl插件:CloverETL和Octopus

在购买现成的工具之外，还有自己从头开发ETL程序的。

ETL工作看起来并不复杂，特别是在数据量小、没有什么转换逻辑的时候，自己开发似乎非常节省成本。的确，主流的ETL工具价格不菲，动辄几十万；而从头开发无非就是费点人力而已，可以控制。至于性能，人大多是相信自己的，认为自己开发出来的东西知根知底，至少这些程序可以完全由自己控制。

就目前自主开发的ETL程序而言，有人用c语言编写，有人用存储过程，还有人用各种语言混杂开发，程序之间各自独立。这很危险，虽然能够让开发者过足编码的瘾，却根本不存在架构。

有位银行的朋友，他们几年前上的数据仓库系统，就是集成商自己用c语言专门为他们的项目开发。单从性能上看似似乎还不赖，然而一两年下来，项目组成员风雨飘零，早已物是人非，只有那套程序还在那里；而且，按照国内目前的软件工程惯例，程序注释和文档是不全或者是不一致的，这样的程序已经对日常业务造成很大阻碍。最近，他们已经开始考虑使用ETL工具重新改造了。

## 扩展阅读

数据仓库项目应该如何选择ETL工具：ETL or E-LT? <http://blog.csdn.net/mengdebin/article/details/41151533>

ETL构建企业级数据仓库五步法

<http://blog.csdn.net/xcbdsu/article/details/6637775>

## 数据存储

### 数据集市 ( Data Marts )

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

( subjectarea )。在数据仓库的实施过程中往往可以从一个

点赞 13

评论 4

分享

已收藏 83

手机看

打赏

...

订阅博主

## 数据仓库管理

安全和特权管理；跟踪数据的更新；数据质量检查；管理和更新元数据；审计和报告数据仓库的使用和状态；删除数据；复制、分割和分发数据；备份和恢复；存储管理。

## 选型

在大数据时代，数据仓库的重要性更胜以往。Hadoop平台下的Hive，Spark平台下的Spark SQL都是各自生态圈内应用最热门的配套工具，而它们的本质就是开源分布式数据仓库。

在国内最优秀的互联网公司里(如阿里、腾讯)，很多数据引擎是架构在数据仓库之上的(如数据分析引擎、数据挖掘引擎、推荐引擎、可视化引擎等等)。不少员工认为，开发成本应更多集中在数据仓库层，不断加大数据建设的投入。因为一旦规范、标准、高性能的数据仓库建立好了，在之上进行数据分析、数据挖掘、跑推荐算法等都是轻松惬意的事情。反之如果业务数据没梳理好，各种脏乱数据会搞得人焦头烂额，苦不堪言。

## 数据访问

数据仓库通常需要提供具有直接访问数据仓库功能的前端应用，这些应用也被称为BI(商务智能)应用

有数据查询和报表工具

应用开发工具

经理信息系统（EIS）工具

联机分析处理（OLAP）工具

数据仓库建设好以后，用户就可以编写SQL语句对其进行访问并对其中数据进行分析。但每次查询都要编写SQL语句的话，未免太麻烦，而且对维度建模数据进行分析的SQL代码套路比较固定。

于是，便有了OLAP工具，它专用于维度建模数据的分析。而BI工具则是能够将OLAP的结果以图表的方式展现出来，它和OLAP通常出现在一起。(注：本文所指的OLAP工具均指代这两者。)

这种情况下，OLAP不允许访问中心数据库。一方面中心数据库是采取规范化建模的，而OLAP只支持对维度建模数据的分析；另一方面规范化数据仓库的中心数据库本身就不允许上层开发人员访问。而在维度建模数据仓库中，OLAP/BI工具和数据仓库的关系则是这样的：

在维度建模数据仓库中，OLAP不但可以从数据仓库中直接取数进行分析，还能对架构在其上的数据集集群做同样工作。

数据挖掘工具。

信息发布系统

把数据仓库中的数据或其他相关的数据发送给不同的地点或用户。基于Web的信息发布系统是对付多用户访问的最有效方法。

## 数据可视化选型

你想知道的经典图表全在这

<https://zhuanlan.zhihu.com/p/24168144>

R语言

<http://www.cnblogs.com/muchen/p/5332359.html>

pentaho

FineBI

PowerBI

<http://www.cnblogs.com/muchen/p/5389960.html>

<http://www.cnblogs.com/muchen/p/5391101.html>

点赞 13

评论 4

分享

已收藏 83

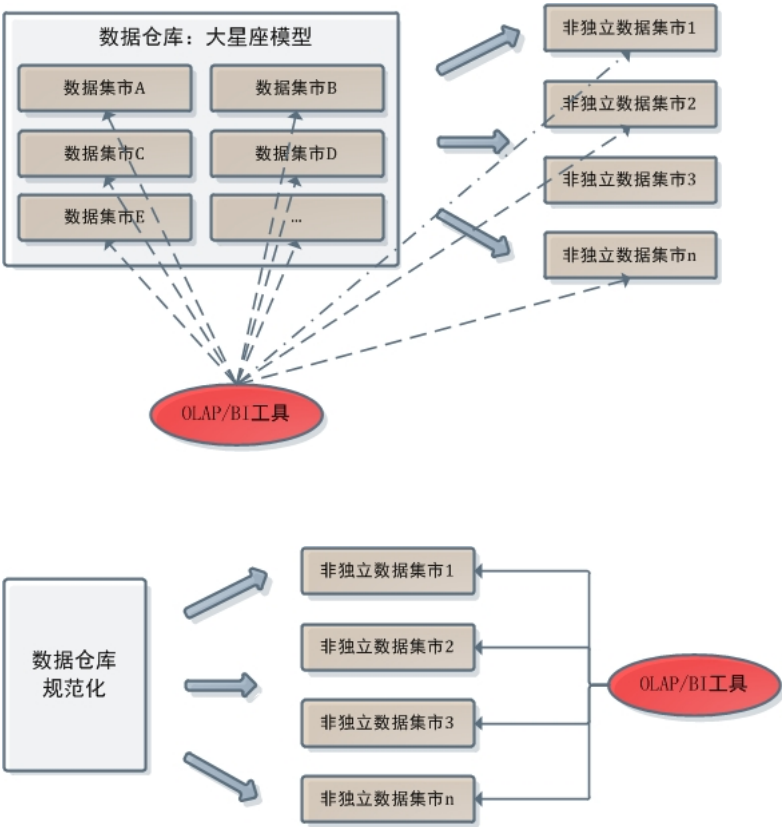
手机看

打赏

...

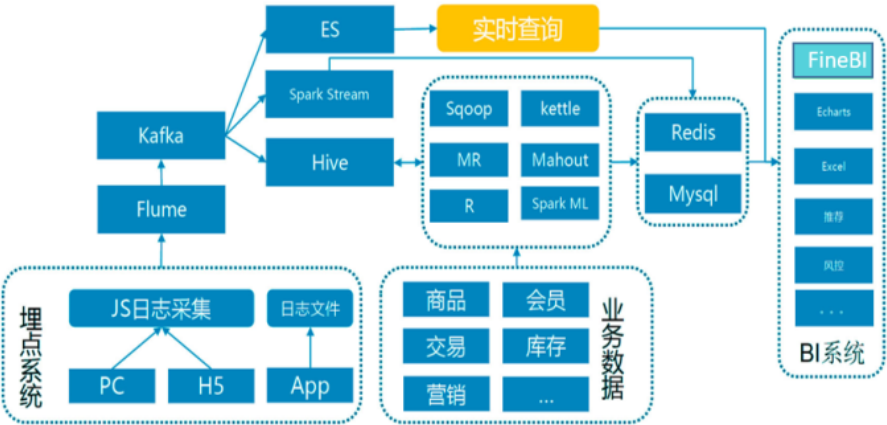
订阅博主





案例

facebook的ppt上了解到的是他们在hive上做大数据量的分析，计算结果放到oracle上做BI展示和计算 hadoop MR or hive上ETL计算完的结果表，同步到oracle中，连接传统BI工具，呈现报表，阿里、腾讯、盛大都是这样的



与传统数据库的对比

企业的数据处理大致分为两类：

一类是操作型处理，也称为联机事务处理，它是针对具体业务在数据库联机的日常操作，通常对少数记录进行查询、修改。

另一类是分析型处理，一般针对某些主题的历史数据进行分析，支持管理决策。

数据仓库：数据仓库系统的主要应用主要是OLAP（On-Line Analytical Processing），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。

举个最常见的例子，拿电商行业来说好了。基本每家电商公司都会经历，从只需要业务数据库到要数据仓库的阶段。

电商早期启动非常容易，入行门槛低。找个外包团队，做了一个可以下单的网页前端 + 几台服务器 + 一个MySQL，就能开门迎客了。这好比手工作坊时期。

第二阶段，流量来了，客户和订单都多起来了，普通查询已经有压力了，这个时候就需要升级架构变成多台服务器和多个业务数据库（量大+分库分表），这个阶段的业务数字和指标还可以勉强从业务数据库里查询。初步进入工业化。

第三个阶段，一般需要3-5年左右的时间，随着业务指数级的增长，数据量的会陡增，公司角色也开始多了起来，开始有了CEO、CMO、CIO，大家需要面临的问题越来越复杂，越来越深入。

高管们关心的问题，从最初非常粗放的：“昨天的收入是多少”、“上个月PV、UV是多少”，逐渐演化到非常精细化和具体的用户的集群分析，特定用户在某种使用场景中，例如“20~30岁女性用户在过去五年的第一季度化妆品类商品的购买行为与公司进行的促销活动方案之间的关系”。这类非常具体，且能够对公司决策起到关键性作用的问题，基本很难从业务数据库从调取出来。

原因在于：业务数据库中的数据结构是为了完成交易而设计的，不是为了而查询和分析的便利设计的。

业务数据库大多是读写优化的，即又要读（查看商品信息），也要写（产生订单，完成支付）。

因此对于大量数据的读（查询指标，一般是复杂的只读类型查询）是支持不足的。而怎么解决这个问题，此时我们就需要建立一个数据仓库了，公司也算开始进入信息化阶段了。

数据仓库的作用在于：数据结构为了分析和查询的便利；只读优化的数据库，即不需要它写入速度多么快，只要做大量数据的复杂查询的速度足够快就行了。

那么在这里前一种业务数据库（读写都优化）的是业务性数据库，后一种是分析性数据库，即数据仓库。

最后总结一下：

数据库比较流行的有：MySQL, Oracle, SqlServer等

数据仓库比较流行的有：AWS Redshift, Greenplum, Hive等。

这样把数据从业务性的数据库中提取、加工、导入分析性的数据库就是传统的ETL工作。

数据仓库的方案建设的目的，是为前端查询和分析作为基础，由于有较大的冗余，所以需要的存储也较大。

为了更好地为前端应用服务，数据仓库必须有如下几点优点，否则是失败的数据仓库方案。

- 1.效率足够高。
- 2.数据质量。
- 3.扩展性。

两类数据库的不同点：

#### 1.数据组成差别 - 数据时间范围差别

一般来讲，操作型数据库只会存放90天以内的数据，而分析型数据库存放的则是数年内的数据。这点也是将操作型数据和分析型数据进行物理分离的主要原因。

#### 2.数据组成差别 - 数据细节层次差别

操作型数据库存放的主要是细节数据，而分析型数据库中虽然既有细节数据，又有汇总数据，但对于用户来说，重点关注的是汇总数据部分。

而对于分析型数据库来说，因为汇总数据比较稳定不会发生改变，而且其计算量也比较大(因为时间跨度大)，因此它的汇总数据可考虑事先计算好，以避免重复计算。

### 3.数据组成差别 - 数据时间表示差别

操作型数据通常反映的是现实世界的当前状态；而分析型数据库既有当前状态，还有过去各时刻的快照，分析型数据库的使用者可以综合所有快照对各个历史阶段进行统计分析。

### 4.技术差别 - 查询数据总量和查询频度差别

操作型查询的数据量少而频率多，分析型查询则反过来，数据量大而频率少。要想同时实现这两种情况的配置优化是不可能的，这也是将两类数据库物理分隔的原因之一。

### 5.技术差别 - 数据更新差别

操作型数据库允许用户进行增，删，改，查；分析型数据库用户则只能进行查询。

### 6.技术差别 - 数据冗余差别

数据的意义是什么？就是减少数据冗余，避免更新异常。而如5所述，分析型数据库中没有更新操作。因此，减少数据冗余也就没那么重要了。

例如Hive是一种数据仓库，而数据仓库和分析型数据库的关系非常紧密。它只提供查询接口，不提供更新接口，这就使得消除冗余的诸多措施不需要被特别严格地执行了，可以保留冗余。

### 7.功能差别 - 数据读者差别

操作型数据库的使用者是业务环境内的各个角色，如用户，商家，进货商等；分析型数据库则只被少量用户用来做综合性决策。

### 8.功能差别 - 数据定位差别

这里说的定位，主要是指以何种目的组织起来。操作型数据库是为了支撑具体业务的，因此也被称为"面向应用型数据库"；分析型数据库则是针对各特定业务主题域的分析任务创建的，因此也被称为"面向主题型数据库"。



## Prism and Data Warehouse both use database, what are the differences?

### • Database of Prism

- OLTP (OnLine Transaction Processing)
- Data organization for business systems
- Organized for speed of transaction processing.
- Optimized for Data Integrity
- Data is changing continually

### • Database of Data Warehouse

- OLAP (OnLine Analytical Processing)
- Data organization for analysis
- Optimized for Retrieval Speed
- Optimized to Simplify Query
- Data was composed by daily snapshot of Prism's data, then cleaned up, calculated and added a version in the format of 'yyyymmdd'



Copyright © 2003 Schick Asia, Inc. All rights reserved. 27



基础数据的架构

关键问题

一般问题 (不完全是技术或文化, 但很重要) 包括但不限于以下几点:

- 业务用户想要执行什么样的分析?
- 你现在收集的数据需要支持那些分析吗?

数据在哪儿?

数据清洗范围

数据的清洁度如何?

相似的数据有多个数据源吗?

什么样的结构最适合核心数据仓库 (例如维度或关系型)?

技术问题包括但不限于以下几点:

- 在你的网络中要流通多少数据? 它能处理吗?
- 需要多少硬盘空间?
- 硬盘存储需要多快?
- 你会使用固态还是虚拟化的存储?

2)建立数据模型和数据仓库的物理设计

3)定义数据源

4)选择数据仓库技术和平台

5)从操作型数据库中抽取、净化、和转换数据到数据仓库-ETL依照模型进行初始加载、增量加载、缓慢增长维、慢速变化维、事实表加载等数据集成

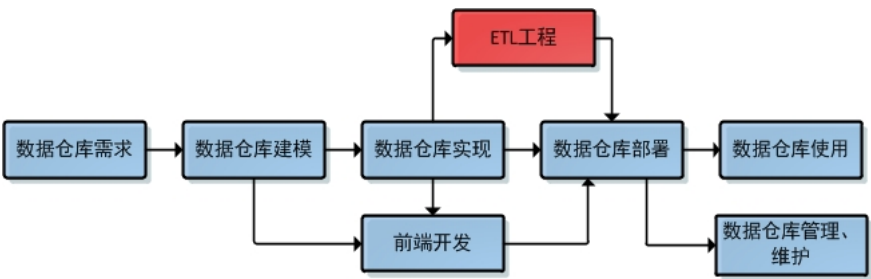
6)选择访问和报表工具

7)选择数据库连接软件

8)选择数据分析和数据展示软件

9)更新数据仓库-并根据业务需求制定相应的加载策略、刷新策略、汇总策略、维护策略。

较之数据库系统开发, 数据仓库开发只多出ETL工程部分。然而这一部分极有可能是整个数据仓库开发流程中最为耗时耗资源的一个环节。  
因为该环节要整理各大业务系统中杂乱无章的数据并协调元数据上的差别, 所以工作量很大。在很多公司都专门设有ETL工程师这样的岗位, 大的公司甚至专门聘请ETL专家。



只会Python，能找到计算机视觉的工作吗  
还是需要学习什么

09-24

大数据与数据仓库入门到精通

本课程以CDH作为大数据平台，详细介绍CDH平台各个组件在生产环境的应用及开发，并结合实际的业务场景，...

09-20

评论

优质评论可以帮助作者获得更高权重

码农\*gotobackto:

独一无二的数据库建模指南系列教程升级版 网盘地址：https://pan.baidu.com/s/1k3BkvCS5JPJhtCpQnw40JA 提取码：hwcj 1月前 回复 ...

JkYU5821:

专业做大数据获客 绝对真实有效 支持小量测试 看我昵称加喂 4月前 回复 ...

Nicky\_1218:

ETL部分，informatica跟powercenter不是一个东西么，powercenter是informatica的一部分。 2年前 回复 ...

码哥\*张小凡vip

回复：感谢提醒，已修复

点赞 13

评论 4

分享

已收藏 83

手机看

打赏

订阅博主