Hive 数据仓库应用教程

第1讲 认识Hive —— 强义原稿(jiǎng yì yuán gǎo)

授课教师：李阳源

适用对象：专科大学学生 | 语言：中文（口语化） | 难词标注拼音

幻灯片 1 —— 项目一 认识Hive（封面）

同学们好！

欢迎来到《Hive数据仓库(shù jù cāng kù)应用教程》这门课。

我是你们的老师李阳源。

今天是我们的第一节课，叫做'认识Hive'。

什么意思呢？就是先了解一下Hive到底是什么东西。

大家看一下屏幕上的这张封面(fēng miàn)。

> **【幻灯片图片说明】**封面图片：左侧显示'项目一 认识Hive'大标题，右侧有一只手拿着手机，手机上方显示各种数据图标——比如柱状图(zhù zhuàng tú)、车辆、房子、网络连接等图案，代表大数据和物联网(wù lián wǎng)的应用场景。下方列出了本课程全部9个项目的名称。

我们这门课一共有9个项目，从认识Hive开始，到环境搭建(dā jiàn)、基本语法，一直到最后的综合案例(zōng hé àn lì)。

今天我们先从最基础的开始——先认识一下数据仓库是什么，然后再认识Hive。

幻灯片 2 —— 本节目录

好，大家看一下今天的课程安排(ān pái)：

我们今天主要讲三个部分：
第一部分：任务1-1，认识数据仓库。
第二部分：任务1-2，认识Hive数据仓库。
第三部分：项目小结(xiǎo jié)。

我们先从第一部分开始，了解一下什么是数据仓库。

幻灯片 3 —— 任务 1-1 认识数据仓库（标题页）

好，现在我们正式进入第一个任务——认识数据仓库。
在学Hive之前，我们必须先搞清楚什么是数据仓库。
因为Hive本身就是一种数据仓库工具。
如果你不知道数据仓库是什么，那后面学Hive就很难理解了。

幻灯片 4 —— 1.1.1 数据仓库的概念

那什么是数据仓库呢?

数据仓库，英语叫**Data Warehouse**，简称(jiǎn chēng)'数仓'或者'**DW**'。
它是一个专门用来存储(cún chǔ)、分析、报告数据的系统。

大家注意，数据仓库的目的是什么?
就是构建(gòu jiàn)一个面向分析的、集成化(jí chéng huà집적화, 통합화)的数据环境。

分析完了之后呢，把结果提供给企业(qǐ yè)，帮助企业做决策(jué cè)。

英语叫**Decision Support**，就是'决策支持'的意思。

举个简单的例子：

你开了一家网店，每天有很多订单数据、用户数据、商品数据。

你想知道哪个商品卖得最好、哪个月销量(xiāo liàng)最高。

这时候你就需要把这些数据都放到一个地方去分析——这个地方就是数据仓库。

【幻灯片图片说明】图片描述：从左到右依次展示三个部分。左边是Data Source（数据源），显示为黑色服务器和数据库图标，代表各种原始数据来源；中间是Data Warehouse（数据仓库），显示为圆柱形数据存储图标；右边是Analysis（分析），显示为一个电脑屏幕，上面有各种图表——饼图 **(bǐngtú) - 원형 차트 (파이 차트)**、柱状图 **柱状图 (zhùzhuàngtú) - 막대 그래프**、折线图 **折线图 (zhéxiàntú) - 꺾은선 그래프 (선 그래프)**，还有人物图标，代表分析人员在查看分析结果。整体流程是：数据从数据源流入数据仓库，再从数据仓库输出到分析环节。

幻灯片 5 —— 1.1.1 数据仓库的概念（续）

这里有一个很重要的特点，大家一定要记住：

数据仓库本身并不'生产'任何数据！

它的数据来源于不同的外部系统。

同时，数据仓库自身也不需要'消费(xiāo fèi)'任何的数据。

它的结果是开放给各个外部应用(yìng yòng)使用的。

所以说，这也是为什么叫'仓库(cāng kù)'，而不叫'工厂(gōng chǎng)'的原因。

工厂是生产东西的，而仓库只是存放和管理东西的。

数据仓库只是把别人的数据收集过来，整理好，给别人用。

【幻灯片图片说明】图片描述：左侧是四个数据源系统——CRM（客户关系管理）、ERP（企业资源规划）、Billing（计费系统）、Supply Chain（供应链）。中间有一个齿轮图标，标注ETL（抽取、转换、加载），代表数据处理过程。右边是一个大的圆柱形Data Warehouse（数据仓库）。从数据仓库再向右输出到三个应用：Reporting（报表）、Visualization（可视化）、BI（商业智能）。

幻灯片 6 —— 1.1.1 数据仓库的概念（案例引入）

案例：数据仓库为何而来?

好，我们来思考(sī kǎo)一个问题：数据仓库为什么会出现?

先说结论(jié lùn)：为了分析数据而来!

分析完的结果，是给企业做决策提供支撑(zhī chēng, 떠받치다)的。

在企业里面，信息一般有两个用途(yòng tú)：

第一，操作型记录的保存——就是记录日常业务数据。

第二，分析型决策的制定——就是用数据来帮助做决定。

下面我们用中国人寿保险公司(China Life)作为例子来讲一讲。

幻灯片 7 —— 1.1.2 数据仓库为何而来——操作型记录的保存

（1）操作型记录的保存

中国人寿保险集团(jí tuán)公司，业务很多。

它有人寿险(rén shòu xiǎn 생명보험)、财险(cái xiǎn)、车险(chē xiǎn)、养老险(yǎng lǎo xiǎn)等等。

每条业务线在正常运营(yùn yíng)的时候，都需要记录很多信息：

- 比如客户信息、保单(bǎo dān)信息(**보험 증권 정보**)、收付费、核保(hé bǎo 인수심사)、理赔(lǐ péi 손해사정) 等等。

那这些数据怎么处理呢?

这时候就要用到联机事务处理系统，英语叫OLTP。

全称是On-Line Transaction Processing。

OLTP系统的特点是什么呢?

就是前台接收到用户数据，可以立即(lì jí)传送到后台进行处理。

然后在很短的时间内给出处理结果。

比如你去买保险，工作人员输入你的信息，系统马上就能处理完。

【幻灯片图片说明】图片描述：展示了两个保险产品的截图。上面是'车险'产品，保险期间1年，产品简介为'透明才安心，车程路坦途'，网上购买，由中国人寿财产保险股份(**gǔ fèn주식)**有限公司提供。下面是'居家无忧·家庭财产保险'，保险期间也是1年，每天仅需0.54元起，同样网上购买。

幻灯片 8 —— 1.1.2 数据仓库为何而来——操作型记录的保存（续）

刚才说了，OLTP系统主要用关系型数据库(guān xì xíng shù jù kù)来实现。

英语缩写(suō xiě)叫RDBMS。

常见的有Oracle、MySQL、SQL Server这些。

大家看屏幕上的图：

【幻灯片图片说明】图片描述：顶部是'中国人寿保险（集团）公司'，下方用箭头分出四个业务分支：人寿险（用Oracle数据库）、财险（用MySQL数据库）、车险（用Oracle数据库）、养老险（用SqlServer数据库）。每个分支下面各有一张数据表，记录了一个叫Allen的客户在不同险种中的信息。比如：人寿险ID是0001，产品是'国寿盛世'，金额8千；财险ID是241568，产品是'居家无忧'，金额98元；车险ID是456621，产品是'车损险'，金额19.6万；养老险ID是200145，产品是'幸福一生'，金额4200元。

大家注意看，同一个客户Allen，在四个不同的业务系统里面都有数据。

而且每个系统用的数据库还不一样!

有的用Oracle，有的用MySQL，有的用SQL Server。

这就是现实中企业的真实情况。

幻灯片 9 —— 1.1.2 数据仓库为何而来——分析型决策的制定

（2）分析型决策(jué cè)的制定(zhì dìng)

随着公司业务不断运营，数据会越来越多。

这时候就会产生(chǎn shēng)很多问题和困惑(kùn huò)：

比如说：

——哪些保险产品正在变差，已经变成不好的产品了？

——新客户和老客户续保的政策(zhèng cè)怎么定才有效？

——理赔(lǐ péi)的过程中有没有人在骗保(piàn bǎo 보험사기)？

——现在看到的报表只是某一条业务线的，整个集团的数据怎么看？

这些问题都需要通过数据分析来回答。

幻灯片 10-11 —— 1.1.2 数据驱动(qū dòng)决策

那要解决这些问题怎么办呢?

靠拍脑袋(pāi nǎo dài)做决定? 那肯定不行!

PPT上说的'瞎拍(xiā pāi)桌子'就是这个意思——乱猜(luàn cāi:되는대로 추측하다)是不行的。

最靠谱(kào pǔ)的办法是什么?

就是基于(jī yú)业务数据来做数据分析，然后用分析结果来支持决策。

这就叫做'数据驱动决策'，英语叫Data-Driven Decision Making。

> **【幻灯片图片说明】**图片描述：右侧展示了'Data-Driven Decision Making'（数据驱动决策）的流程图，分5个步骤：1. Know Your Mission（明确你的使命）；2. Identify Data Sources（确定数据来源）；3. Clean and Organize Data（清理和整理数据）；4. Perform Statistical Analysis（进行统计分析）；5. Draw Conclusions（得出结论）。

好，这里PPT上又提出了一个很关键的问题：

在哪里进行数据分析?

用数据库可以吗?

我们接着往下看。

幻灯片 12 —— OLTP环境开展(kāi zhǎn)分析可行(kě xíng)吗?

答案是：可以，但是没必要!

为什么这么说呢?

OLTP系统的核心(hé xīn)是面向业务的，支持日常业务操作。
所有的操作分两种：读操作和写操作。
一般来说，读的压力比写的压力大很多。

如果你还要在OLTP系统上面做数据分析，会有什么问题呢?

第一，数据分析也是读数据的操作，会让系统读的压力翻倍(fān bèi:곱절로 늘어나다.)!
就好比一条公路本来就很堵了，你再往上面加车，那不是更堵了吗?

第二，OLTP系统只保存最近几周或者几个月的数据。
做分析经常需要看很长时间的历史数据，OLTP里面没有。

第三，数据分散(fēn sàn)在不同系统的不同表里面。
字段类型、名称都不统一(tǒng yī)，很难放在一起分析。

幻灯片 13 —— OLTP环境开展分析可行吗？（续）

如果数据量比较小，在业务低峰期(dī fēng qī)的时候，勉强可以在OLTP上直接分析。

但是为了更好地做各种规模(guī mó)的数据分析，同时又不影响OLTP系统的正常运行：

我们需要构建一个集成统一的数据分析平台。

这个平台的目的很简单：

面向分析，支持分析，并且和OLTP系统解耦合(jiě ǒu hé:decoupling)。

'解耦合'就是把它们分开，互不影响的意思。

基于这种需求(xū qiú)，数据仓库的雏形(chú xíng:초기모델, 시작단계의 형태)就开始出现了！

幻灯片 14 —— 1.1.3 数据仓库主要特征(tè zhēng)——数仓体系图

好，我们来看一下数据仓库的整体体系图。

再强调(qiáng diào)一遍((yí biàn한번) 数据仓库的核心特点：

——目的是构建面向分析的集成化数据环境。

——分析结果为企业提供决策支持。

——本身不'生产'任何数据，数据来源于外部系统。

——自身也不'消费'数据，结果开放给外部应用使用。

【幻灯片图片说明】左侧四个数据源（CRM、ERP、Billing、Supply Chain）通过ETL流程进入中间的Data Warehouse，再输出到右侧三个应用（Reporting、Visualization、BI）。

幻灯片 15 —— 1.1.3 数据仓库主要特征——四大特征总览(zǒng lǎn)

数据仓库有四个最重要的特征，大家一定要记住：

第一，面向主题(zhǔ tí)——Subject-Oriented

主题是一个抽象(chōu xiàng)的概念，就是把数据按照分析的角度(jiǎodù)来分类整理。

第二，集成性(jí chéng xìng)——Integrated

把不同系统里面分散的数据集中到一起来。

第三，非易失性(fēi yì shī xìng)——Non-Volatile

数据仓库是用来分析的，不是用来改数据的。数据进来以后就稳定不变。

第四，时变性(shí biàn xìng)——Time-Variant

数据仓库的数据需要随着时间不断更新，来满足决策的需要。

幻灯片 16 —— 面向主体性 Subject-Oriented

我们来详细(xiáng xì)说说第一个特征——面向主题。

在普通数据库里面，数据是面向应用来组织(zǔ zhī)的。

每个业务系统可能是相互分离(fēn lí)的，各管各的。

但数据仓库不一样，它是面向主题的。

什么叫主题呢?

就是把企业里面某个大的分析领域相关的数据综合(zōng hé)在一起、归类(guī lèi:분류하다. 범주로 나누다)在一起。

举个例子:

在保险行业(háng yè)，我们可以有'客户'这个主题、'保险理赔(보험금 청구)'这个主题、'承保(chéng bǎo:보험계약체결)'这个主题。

每个主题包含和这个话题相关的所有数据。

主题之间互不交叉(jiāo chā)，但每个主题内部的数据是完整的、一致的。

幻灯片 17 —— 集成性 Integrated

好，第二个特征——集成性。

确定了主题之后，就要去获取(huò qǔ)相关的数据了。
但问题是，企业里面的数据通常分布在很多不同的系统里面。
它们彼此(bǐ cǐ)分散、独立、而且结构不同（异构 yì gòu）。

所以在数据进入数据仓库之前，必须经过统一和综合的处理。
具体来说就是对数据进行：
抽取(chōu qǔ :extraction)——从各个系统把数据拿出来；
清理(qīng lǐ:cleaning)——去掉脏数据(zāng shùjù)、错误数据；
转换(zhuǎn huàn:transformaton)——统一格式和标准；
汇总(huì zǒng summarization집계)——把数据合并到一起。

比如说，A系统里面客户性别用'男/女'表示，B系统用'M/F'表示。
这种情况就要统一成一种格式，才能放到数据仓库里面。

幻灯片 18 —— 集成性 Integrated（图解）

【幻灯片图片说明】图片描述：左侧是四个业务系统数据库——人寿险、财险、车险、养老险，属于'业务系统'。中间标注'ETL：抽取、转换、加载'过程，用箭头汇聚到右侧的'保险行业数据仓库'中，仓库内部标注了'主题：承保'，表示所有数据都集成到承保这个主题下面。

大家看这张图，就是我们刚才说的那个意思。
来自四个不同业务系统的数据，通过ETL处理以后，统一集成到数据仓库里面。
虽然每个系统的数据命名可能不同、格式也不同，
但放到数据仓库之前，需要先把这些不一致的地方去掉。

幻灯片 19 —— 非易失性 Non-Volatile

好，第三个特征——非易失性。这个词有点难，我解释一下。

数据仓库是分析数据的平台，而不是创造数据的平台。
我们是通过数据仓库去分析数据中的规律(guī lǜ)，
而不是去修改(xiū gǎi)里面的规律。
所以数据进入数据仓库后，就稳定(wěn ding 안정적)不变了。

普通数据库呢，每天都在不断地增删改查(zēng shān gǎi chá)。
因为要支持日常业务，必须随时更新。

但数据仓库不需要每一笔业务都实时更新。
它是每隔一段时间，把一批新的数据批量(pī liàng)导入进来就可以了。

简单说就是：数据仓库里面主要是查询(chá xún)操作，
修改和删除(shān chú)操作非常少。

幻灯片 20 —— 时变性 Time-Variant

最后一个特征——时变性。

数据仓库里面保存的是各种历史数据。
这些数据可能和某个日期(rì qī)、星期、月份、季度(jì dù)或者年份有关。

虽然用户不能随便修改数据仓库的数据，
但数据仓库的数据并不是永远(yǒng yuǎn)不变的。
因为业务在变化，分析的结果只能反映(fǎn yìng)过去的情况。
所以数据仓库需要随着时间定期(dìng qī)更新。

有三个具体表现(biǎo xiàn)：
第一，数据仓库保存数据的时间范围比普通数据库长得多。
第二，普通数据库存的是当前数据，数据仓库存的是历史数据。
第三，数据仓库的数据是按照时间顺序(shùn xù)追加(zhuī jiā 추가)的，都带有时间标记。

幻灯片 21-22 —— 1.1.4 OLTP 和 OLAP 概念

好，接下来我们来讲两个非常重要的概念：OLTP和OLAP。

OLTP —— 联机事务(shì wù)处理

全称：On-Line Transaction Processing
主要目标是做数据处理。
针对具体业务，在数据库上进行日常操作，通常是对少量记录进行查询和修改。
用户关心的是：响应(xiǎng yìng)速度快不快？数据安全不安全？
传统的关系型数据库系统（RDBMS）就是典型的OLTP应用。
比如Oracle、MySQL、SQL Server、MongoDB等等。

幻灯片 23 —— OLAP 概念

OLAP —— 联机分析处理

全称：On-Line Analytical Processing

主要目标是做数据分析。

一般针对某些主题的历史数据，进行复杂的多维(duō wéi)分析。

目的是支持管理层做决策。

数据仓库就是OLAP系统的一个典型例子！

简单总结一下：

OLTP——侧重(cè zhòng)'事务'处理，对应的是数据库。

OLAP——侧重'分析'处理，对应的是数据仓库。

幻灯片 24 —— OLTP 与 OLAP 对比

我们来做一个对比(duì bǐ)：

数据源方面：

OLTP——只包含当前运行的日常业务数据。

OLAP——整合(zhěng hé 종합)来自多个来源的数据。

目的方面：

OLTP——面向应用，面向业务，支撑事务。

OLAP——面向主题，面向分析，支撑分析决策。

关注焦点(jiāo diǎn)：

OLTP——关注当下。

OLAP——主要关注过去和历史，也有实时数仓(shí shí shù cāng)。

操作类型：

OLTP——读写操作都有。

OLAP——大量读，很少写。

响应时间：

OLTP——毫秒(háo miǎo)级别，非常快。

OLAP——秒、分钟、小时甚至天，取决于数据量和查询复杂度。

数据量：

OLTP——小数据，MB到GB。

OLAP——大数据，TB到PB级别。

> 【幻灯片图片说明】图片描述：右侧展示了OLTP DB和OLAP DB的关系图。上方是OLTP DB（绿色），处理Business Processes和Data Transactions；下方是OLAP DB（绿色），进行Data Mining、Data Analysis、Big Data等操作。两者之间通过Operations和Information双向箭头连接，中间有一条虚线(xū xiàn)分隔。

幻灯片 25-26 —— 1.1.5 数据仓库与数据库的区别

数据库和数据仓库的区别，其实就是OLTP和OLAP的区别。

OLTP的典型应用就是RDBMS，也就是我们平时说的'数据库'。
注意这里说的是关系型数据库，NoSQL不在讨论范围。

OLAP的典型应用就是DW，也就是我们说的'数据仓库'。

【여기에서 다음 슬라이드로 전환】
大家一定要记住这几点区别：

第一，数据仓库不是大型的数据库！虽然数据仓库存的数据很多。
第二，数据仓库并不是要取代(qǔ dài)数据库，它们各有各的用处。
第三，数据库是面向事务设计的，数据仓库是面向主题设计的。
第四，数据库存的是业务数据，数据仓库存的是历史数据。
第五，数据库是为了捕获(bǔ huò)数据而设计的，数据仓库是为了分析数据而设计的。

【幻灯片图片说明】图片描述：底部有两个相对的箭头，左边蓝色箭头写着'Database'，右边紫色箭头写着'Data Warehouse'，中间一个大大的'VS'，表示两者的对比关系。

幻灯片 27-28 —— 1.1.6 数据仓库与数据集市(jí shì)

好，再讲一个概念——数据集市(Data Mart)。

数据仓库是面向整个集团(jí tuán)组织(zǔ zhī)的数据。
而数据集市是面向单个部门(bù mén)使用的。

简单说，数据集市就是数据仓库的一个子集(zǐ jí)。
也有人叫它'小型数据仓库'。

[여기서 다음 슬라이드 전환]
举个例子：
一个大公司的数据仓库里面有所有部门的数据。
但销售部门只需要和销售相关的数据，这部分就叫销售数据集市。
采购(cǎi gòu 조달,구매)部门只需要采购相关的数据，这部分就叫采购数据集市。

【幻灯片图片说明】图片描述：展示了一个完整的数据仓库架构流程图，从左到右分为五列：DATA SOURCES（数据源）→ ETL → WAREHOUSE（仓库）→ DATA MARTS（数据集市）→ USERS（用户）。数据源包括两个Operational System（操作型系统）和Flat Files（平面文件）。通过ETL进入中间的仓库后，再分别输出到三个数据集市：Purchasing（采购）、Sales（销售）、Inventory（库存）。最右侧用户通过Analytics（分析）、Reporting（报表）、Mining（挖掘）来使用这些数据。

幻灯片 29 —— 1.1.7 数据仓库分层(fēn céng)架构(jià gòu)——分层思想

好，接下来是一个很重要的知识点——数据仓库的分层架构。

数据仓库有个特点：本身不生产数据，也不最终消费数据。
所以按照数据流入流出的过程来分层，是很自然的做法。

最基础的分层思想，理论上分为三个层：
第一层：ODS层——操作型数据层 (Operational Data Store)
第二层：DW层——数据仓库层 (Data Warehouse)
第三层：DA层——数据应用层 (Data Application)

企业在实际应用中，可以在这三层的基础上再增加新的层次。

> **【幻灯片图片说明】**图片描述：右侧展示了一个分层架构图。从左到右分为四大块：Data Sources（数据源，包括Operational System和Flat Files）→ Staging Area（暂存区/ODS）→ Warehouse（数据仓库/DW，包含Metadata元数据、Summary data汇总数据、Raw data原始数据）→ Users（数据应用/DA，包括Analytics分析、Reporting报表、Mining挖掘）。

幻灯片 30 —— 阿里巴巴数仓3层架构

为了帮大家更好地理解分层的思想，我们来看一个真实的案例。
这是阿里巴巴(Ā lǐ bā bā)提供的数据仓库分层架构图。

阿里巴巴的数仓是非常经典(jīng diǎn)的3层架构，从下往上依次是：
ODS（源数据）→ DW（数据仓库）→ DA（数据应用）。

同时通过元数据(yuán shù jù)管理和数据质量(zhì liàng)监控来把控整个数据流转过程。

现在我们只做宏观(hóng guān)了解，不做深入探讨。

【幻灯片图片说明】图片描述：阿里巴巴数仓架构图。最底层是'源数据'，包括点击流数据(Click stream)、数据库数据(OLTP)、文档数据(Documents)、其它数据(Other)。中间是'数据仓库'层，经过ETL处理后包含：明细数据(DWD)、低粒度汇总加工数据(DWB)、高粒度汇总数据(DWS)、数据集市和宽表(DM)、数据应用(ST)。最上层是'数据应用'，包括KPI、账单应用、日志产品应用、报表展示、自定义查询、数据分析、数据挖掘等。右侧有'元数据管理'和'数据质量监控'两个纵向模块贯穿所有层次。

幻灯片 31 —— ODS层 (Operation Data Store Layer)

ODS层，也叫操作型数据层、源数据层、数据引入层、临时缓存(huǎn cún)层。

它的作用是什么呢?
就是把未经处理的原始(yuán shǐ)数据存放到数据仓库系统里面。
结构上和源系统保持一致，相当于数据仓库的'准备区'。

打个比方(bǐ fang)：
这就像快递(kuài dì)到了之后，先放在收发室(shōu fā shì)。
还没有拆(chāi)开整理，只是先放在那里，等后面再处理。

幻灯片 32 —— DW层 (Data Warehouse Layer)

DW层，就是数据仓库层。
这一层是由ODS层的数据加工而成的。

主要工作是：数据加工与整合(zhěng hé)。
建立一致性的维度(wéi dù)，构建可以重复(chóng fù)使用的分析表。

DW层内部还可以细分(xì fēn)：
DIM——公共维度层，建立统一的维度。
DWS/DWB——公共汇总粒度(lì dù)事实层，构建汇总指标(zhǐ biāo)。
DWD——明细(míng xì)粒度事实层，保存详细的事实数据。

现在听不懂没关系，以后我们会慢慢学。

幻灯片 33 —— DA层 (Data Application Layer)

DA层，就是数据应用层。
这是最上面一层，面向最终用户(zuì zhōng yòng hù)的。

它根据业务需求(xū qiú)，提供给产品和数据分析人员使用的数据。

包括：
前端报表(bào biǎo)、分析图表、KPI指标、仪表盘(yí biǎo pán)、
OLAP专题分析、数据挖掘(wā jué)等等。

简单说，这一层就是最终让领导(lǐng dǎo)和业务人员能直接看到分析结果的地方。

幻灯片 34-36 —— 分层的好处

PPT上提了两个思考题：

1. 为什么要对数据仓库进行分层？
2. 分层的好处是什么？

主要有以下几个好处：

第一，清晰(qīng xī)数据结构。

每一层都有自己的作用，用表的时候更方便找到和理解。

第二，数据血缘(xuè yuán)追踪(zhuī zōng)。

最终给业务用的表，来源可能很多。

如果某个来源出了问题，分层以后能快速定位(dìng wèi)问题在哪里。

第三，减少重复(chóng fù)开发。

分层规范以后，中间层的数据可以多次复用(fù yòng)，减少重复计算。

第四，把复杂问题简单化。

把一个大任务分成多个小步骤(bù zhòu)，每层只做一件事，更好理解。

数据出问题时，也只需要从出问题的那一层开始修复。

第五，屏蔽(píng bì)原始数据的异常(yì cháng)。

业务系统经常变化，有了分层以后，不用每次业务变化都重新处理所有数据。

幻灯片 37-38 —— 1.1.8 ETL 与 ELT 的区别

最后我们来讲一个在数仓分层架构中很常见(cháng jiàn)的概念——ETL和ELT。

ETL 是什么?

ETL就是Extract, Transform, Load的缩写。

翻译成中文就是：抽取(chōu qǔ)、转换(zhuǎn huàn)、加载(jiā zǎi)。

具体流程是这样的：

第一步，Extract——从数据源里把数据提取(tí qǔ)出来。

第二步，Transform——对数据进行转换处理，变成统一的格式。

第三步，Load——把处理好的数据加载到数据仓库中。

在ETL中，数据先在暂存区(ODS)做转换处理，处理完了再进仓库。

【幻灯片图片说明】图片描述：ETL流程图——左侧多个数据库图标代表数据源，数据先经过Extraction（抽取），进入Staging Area（暂存区，绿色标识），然后经过Transform & Load（转换并加载）进入DWH（数据仓库），最后经过Transform输出到Analytics（分析端）。

幻灯片 39-40 —— ELT 概念

ELT 是什么?

ELT就是Extract, Load, Transform的缩写。

注意顺序变了！先Load（加载），后Transform（转换）。

具体流程是这样的：

数据从源系统提取出来后，不经过暂存区处理，直接加载到数据仓库中。

然后在数据仓库系统内部再进行转换。

为什么可以这样呢?

因为大数据时代，数据仓库系统本身的计算能力(jì suàn néng lì)非常强大。

比如Hadoop、Hive这些工具，自己就可以处理海量数据的转换。

所以就不需要在外面先转换好再放进来了。

大数据时代的数仓，这个ELT特点非常明显。

> 【幻灯片图片说明】图片描述：ELT流程图——左侧多个数据库代表数据源（标注All Data），数据经过Extraction & Load（抽取并加载）直接进入Hadoop平台（显示Hadoop大象Logo），中间标注All Data。然后在Hadoop内部进行Transform（转换），一部分输出到Analytics（分析端），另一部分输出到PureData System for Analytics（分析专用系统）。

好，ETL和ELT的主要区别就是转换(T)在什么时候做：

ETL——先在外面转换好，再加载进仓库。

ELT——先加载进仓库，再在仓库里面转换。

课堂小结

好，同学们，今天的'认识数据仓库'这部分内容就讲到这里了。

我们来回顾(huí gù)一下今天学了什么：

1. 数据仓库（Data Warehouse）是一个面向分析的数据存储平台。
2. 数据仓库为何而来？——为了分析数据，支持企业决策。
3. 四大特征(tè zhēng)：面向主题、集成性、非易失性、时变性。
4. OLTP是事务处理（对应数据库），OLAP是分析处理（对应数据仓库）。
5. 数据库和数据仓库是不同的东西，各有各的用处。
6. 数据集市是数据仓库的子集，面向特定部门。
7. 数仓分层架构：ODS → DW → DA。
8. ETL和ELT的区别：转换步骤的先后不同。

下节课我们将继续(jì xù)学习任务1-2——认识Hive数据仓库。

谢谢大家！