项目一 认识Hive



项目1 认识Hive.pptx
项目2 Hive环境搭建.pptx
项目3 Hive基本语法.pptx
项目4 Hive数据定义.pptx
项目5 数据的导入与导出.pptx
项目6 Hive数据查询.pptx
项目7 Hive函数.pptx
项目8 Hive优化.pptx
项目9 综合案例.pptx

《Hive数据仓库应用教程》

目 录 / CONTENTS

1 任务1-1 认识数据仓库

2 任务1-2 认识Hive数据仓库

3 项目小结

任务 1-1 认识数据仓库

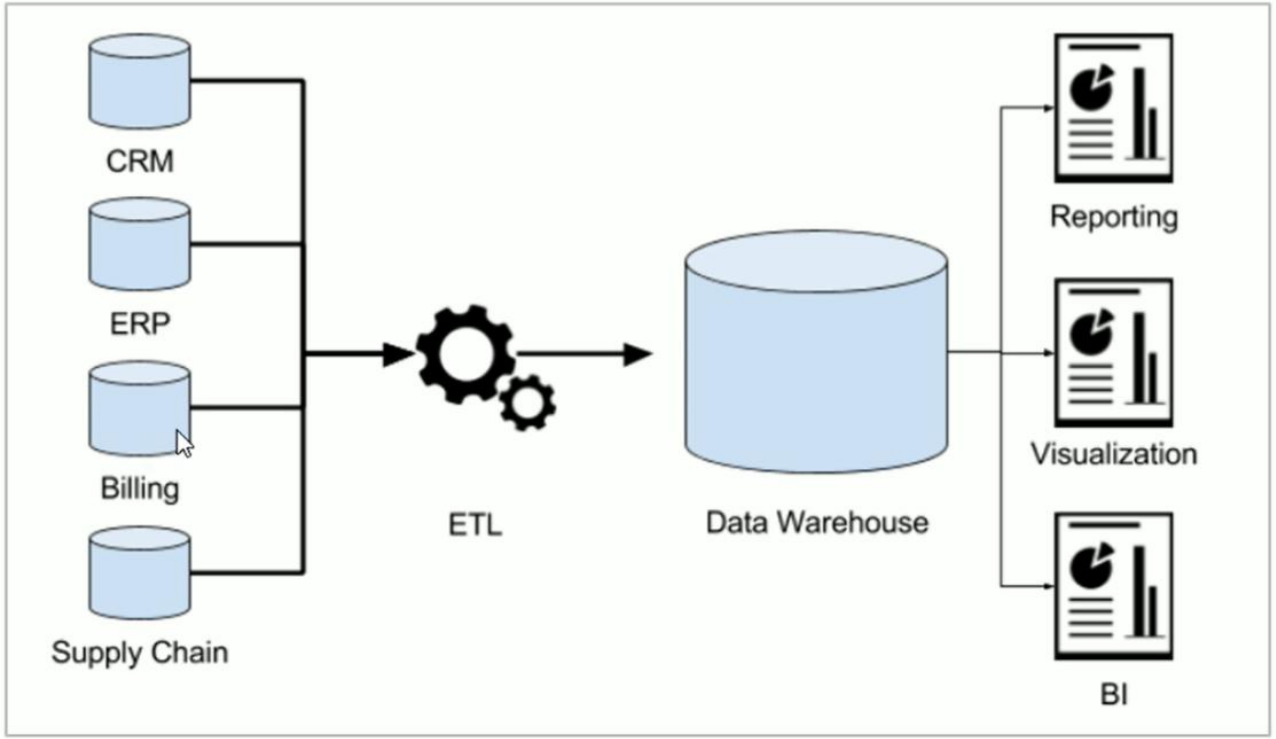
1.1.1 数据仓库的概念

- **数据仓库**（英语：Data Warehouse，简称**数仓**、**DW**），是一个用于存储、分析、报告的数据系统。
- 数据仓库的目的是构建**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（Decision Support）。



1.1.1 数据仓库的概念

- 数据仓库本身并不“生产”任何数据，其数据来源于不同外部系统；
- 同时数据仓库自身也不需要“消费”任何的数据，其结果开放给各个外部应用使用；
- 这也是为什么叫“仓库”，而不叫“工厂”的原因。



案例

数据仓库为何而来

先下结论：为了分析数据而来，分析结果给企业决策提供支撑。

企业中，信息总是用作两个目的：

（1）操作型记录的保存、（2）分析型决策的制定。

下面以中国人寿保险公司（chinalife）发展为例，阐述数据仓库为何而来？

1.1.2 数据仓库为何而来

（1） 操作型记录的保存

- 中国人寿保险（集团）公司下辖多条业务线，包括：人寿险、财险、车险，养老险等。各业务线的业务正常运营需要记录维护包括客户、保单、收付费、核保、理赔等信息。
- **联机事务处理系统（OLTP）**正好可以满足上述业务需求开展，其主要任务是执行联机事务处理。其基本特征是前台接收的用户数据可以立即传送到后台进行处理，并在很短的时间内给出处理结果。



车险

保险期间：1年

产品简介：透明才安心，车程皆坦途

购买途径：网上购买

产品提供：中国人寿财产保险股份有限公司



居家无忧•家庭财产保险

保险期间：1年

产品简介：全面守护爱家，每天仅需0.54元起

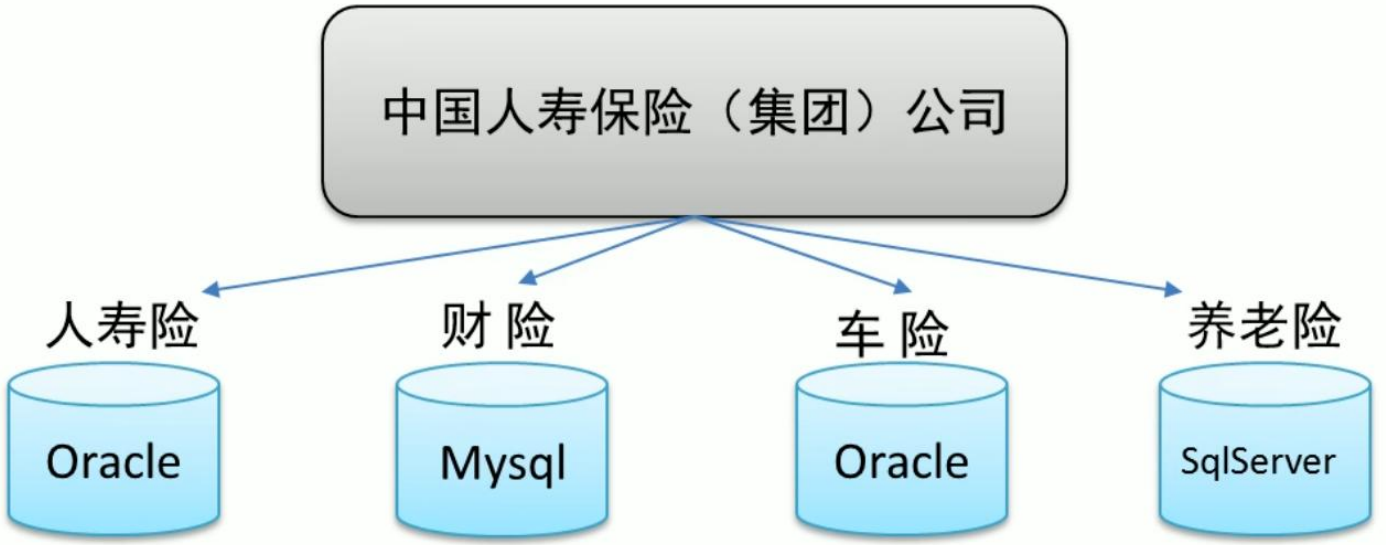
购买途径：网上购买

产品提供：中国人寿财产保险股份有限公司

1.1.2 数据仓库为何而来

（1） 操作型记录的保存

- 关系型数据库（RDBMS）是OLTP典型应用，比如：Oracle、MySQL、SQL Server等。



ID	0001
name	Allen
category	国寿盛世
amount	8（千）

ID	241568
name	Allen
category	居家无忧
amount	98（元）

ID	456621
name	Allen
category	车损险
amount	19.6（万）

ID	200145
name	Allen
category	幸福一生
amount	4200（元）

(2) 分析型决策的制定

- 随着集团业务的持续运营，**业务数据将会越来越多**。由此也产生出许多运营相关的困惑：
  - 能够确定哪些险种正在恶化或已成为不良险种？
  - 能够用有效的方式制定新增和续保的政策吗？
  - 理赔过程有欺诈的可能吗？
  - 现在得到的报表是否只是某条业务线的？集团整体层面数据如何？

1.1.2 数据仓库为何而来

(2) 分析型决策的制定

- 为了能够正确认识这些问题，制定相关的解决措施，瞎拍桌子是肯定不行的。
- 最稳妥办法就是：基于业务数据开展数据分析，基于分析的结果给决策提供支撑。也就是所谓的数据驱动决策的制定。



1.1.2 数据仓库为何而来

(2) 分析型决策的制定

- 为了能够正确认识这些问题，制定相关的解决措施，瞎拍桌子是肯定不行的。
- 最稳妥办法就是：基于业务数据开展数据分析，基于分析的结果给决策提供支撑。也就是所谓的数据驱动决策的制定。



OLTP环境开展分析可行吗?

可以，但是没必要

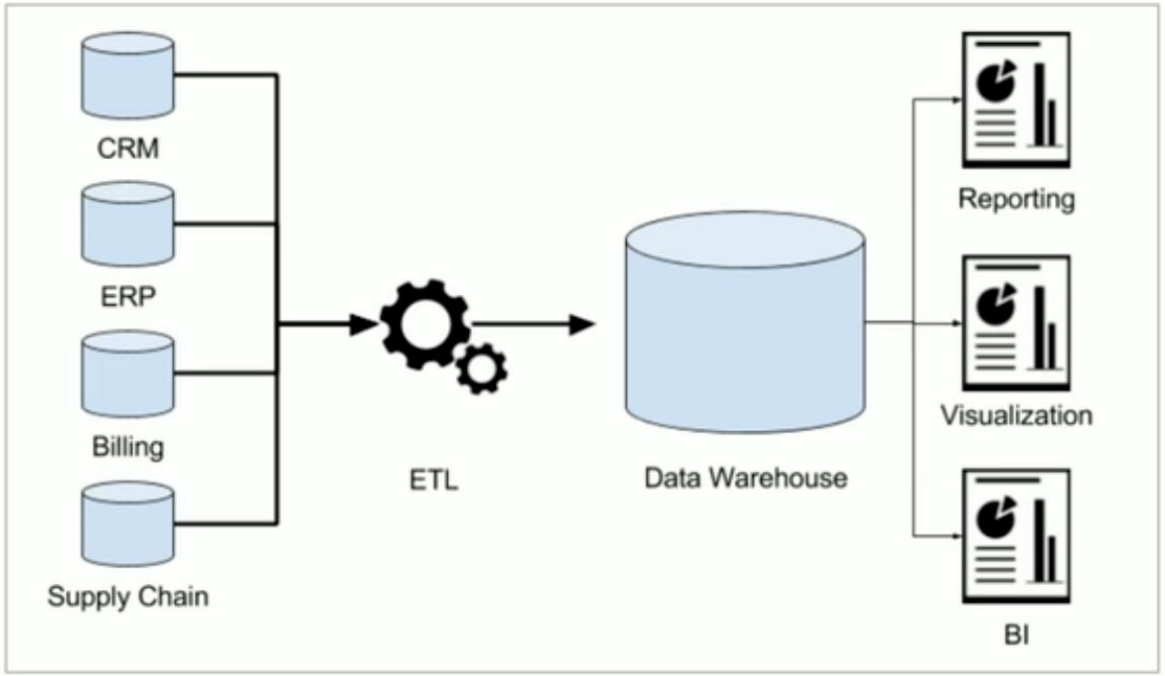
- OLTP系统的核心是面向业务，支持业务、支持事务。所有的业务操作可以分为读、写两种操作，一般来说读的压力明显大于写的压力。如果在OLTP环境直接开展分析，有以下问题需要考虑：
  - 数据分析也是对数据进行读取操作，会让读取压力倍增；
  - OLTP仅存储数周或数月的数据；
  - 数据分散在不同系统不同表中，字段类型属性不统一；

OLTP**环境开展分析可行吗**?

- 当分析所涉及数据规模较小的时候，在业务低峰期时可以在OLTP系统上开展直接分析。但是为了更好的进行各种规模的数据分析，同时也不影响OLTP系统运行，此时需要构建一个集成统一的数据分析平台。
- 该平台的目的很简单：**面向分析，支持分析**，并且和OLTP系统解耦合。
- 基于这种需求，数据仓库的雏形开始在企业中出现了。

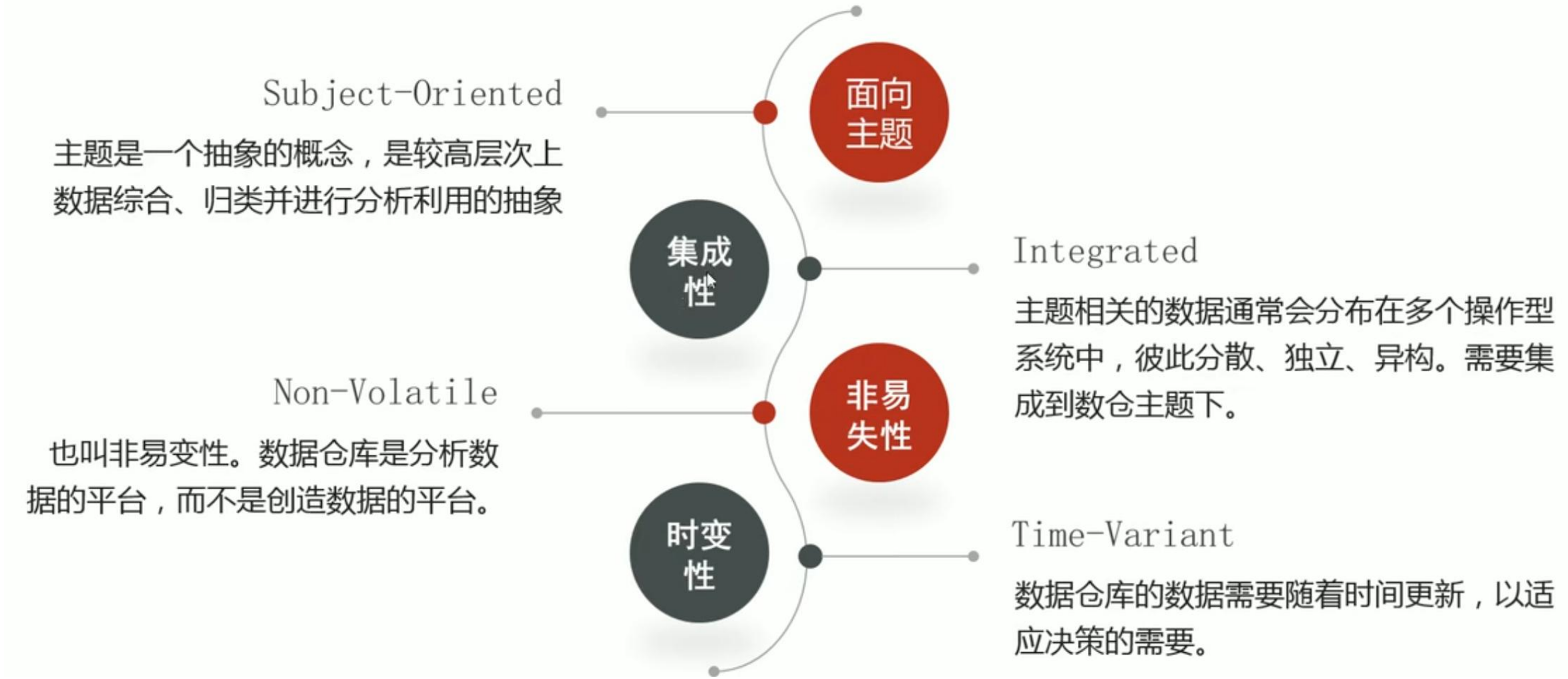
数仓体系图

- **数据仓库**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（Decision Support）。
- 数据仓库本身并不“生产”任何数据，其数据来源于不同外部系统；
- 同时数据仓库自身也不需要“消费”任何的数据，其结果开放给各个外部应用使用；



1.1.3 数据仓库主要特征

主要特征



面向主体性(Subject-Oriented)

- 数据库中，最大的特点是面向应用进行数据的组织，各个业务系统可能是相互分离的。
- 而数据仓库则是面向主题的。主题是一个抽象的概念，是较高层次上企业信息系统中的 **数据综合、归类** 并进行分析利用的抽象。在逻辑意义上，它是对应企业中某一宏观分析领域所涉及的分析对象。
- 操作型处理（传统数据）对数据的划分并不适用于决策分析。而基于主题组织的数据则不同，它们被划分为各自独立的领域，每个领域有各自的逻辑内涵但互不交叉，在 **抽象层次上对数据进行完整、一致和准确的描述**。

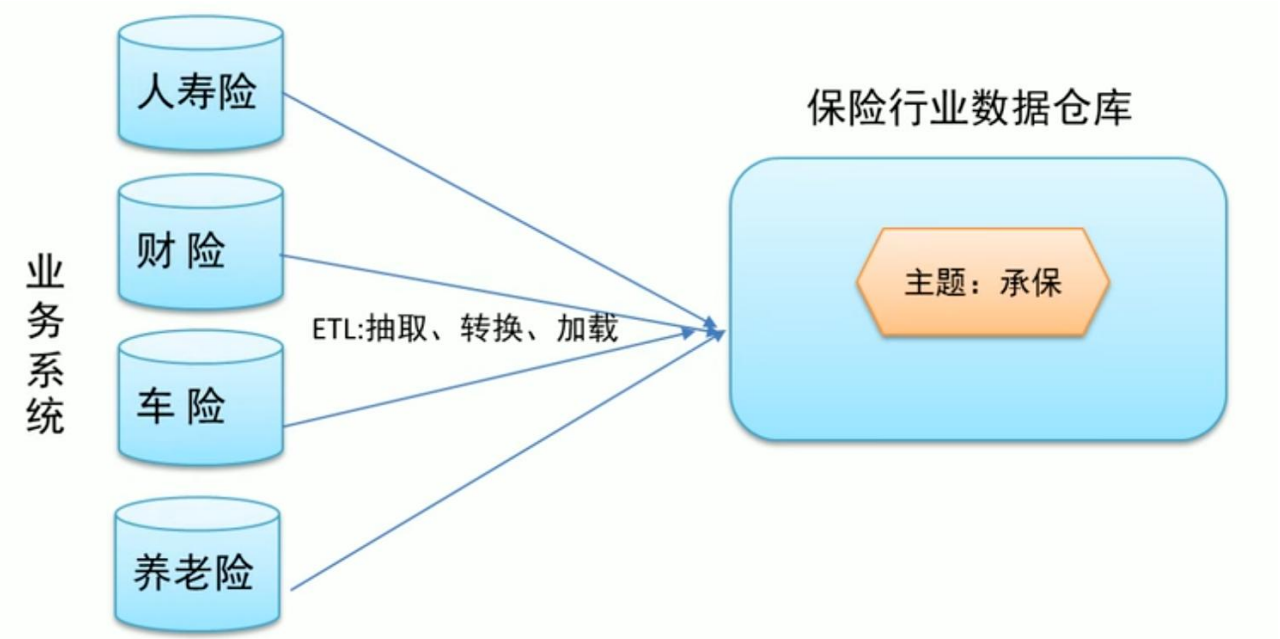


集成性(Integrated)

- 确定主题之后，就需要获取和主题相关的数据。当下企业中主题相关的数据通常会分布在多个操作型系统中，彼此分散、独立、异构。
- 因此在数据进入数据仓库之前，必然要经过统一与综合，对数据进行抽取、清理、转换和汇总，这一步是数据仓库建设中最关键、最复杂的一步，所要完成的工作有：
  - 要统一源数据中所有矛盾之处，如字段的同名异义、异名同义、单位不统一、字长不一致，等等。
  - 进行数据综合和计算。数据仓库中的数据综合工作可以在从原有数据库抽取数据时生成，但许多是在数据仓库内部生成的，即进入数据仓库以后进行综合生成的。

集成性(Integrated)

- 下图说明了保险公司综合数据的简单处理过程，其中数据仓库中与“承保”主题有关的数据来自于多个不同的操作型系统。这些系统内部数据的命名可能不同，数据格式也可能不同。把不同来源的数据存储到数据仓库之前，需要去除这些不一致。



非易失性，非异变形(Non-Volatitle)

- 数据仓库是分析数据的平台，而不是创造数据的平台。我们是通过数仓去分析数据中的规律，而不是去创造修改其中的规律。因此数据进入数据仓库后，它便稳定且不会改变。
- 操作型数据库主要服务于日常的业务操作，使得数据库需要不断地对数据实时更新，以便迅速获得当前最新数据，不至于影响正常的业务运作。在数据仓库中只要保存过去的业务数据，不需要每一笔业务都实时更新数据仓库，而是根据商业需要每隔一段时间把一批较新的数据导入数据仓库。
- 数据仓库的数据反映的是一段相当长的时间内历史数据的内容，是不同时点的数据库快照的集合，以及基于这些快照进行统计、综合和重组的导出数据。
- 数据仓库的用户对数据的操作大多是数据查询或比较复杂的挖掘，一旦数据进入数据仓库以后，一般情况下被较长时间保留。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。

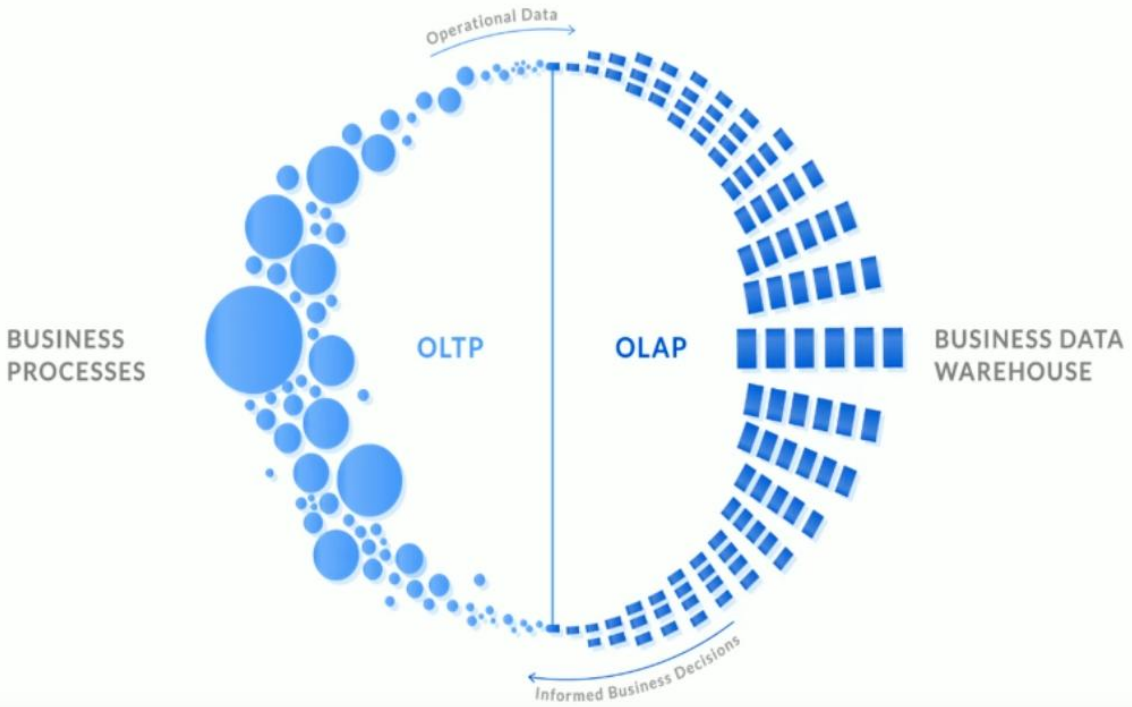
时变性(Time-Variant)

- 数据仓库包含各种粒度的历史数据，数据可能与某个特定日期、星期、月份、季度或者年份有关。
- 虽然数据仓库的用户不能修改数据，但并不是说数据仓库的数据是永远不变的。分析的结果只能反映过去的情况，当业务变化后，挖掘出的模式会失去时效性。因此数据仓库的数据需要随着时间更新，以适应决策的需要。从这个角度讲，数据仓库建设是一个项目，更是一个过程。
- 数据仓库的数据随时间的变化表现在以下几个方面。
  - 数据仓库的数据时限一般要远远长于操作型数据的数据时限。
  - 操作型系统存储的是当前数据，而数据仓库中的数据是历史数据。
  - 数据仓库中的数据是按照时间顺序追加的，它们都带有时间属性。

1.1.4 OLTP, OLAP

概念

- 联机**事务**处理 OLTP（On-Line Transaction Processing）。
- 联机**分析**处理 OLAP（On-Line Analytical Processing）。



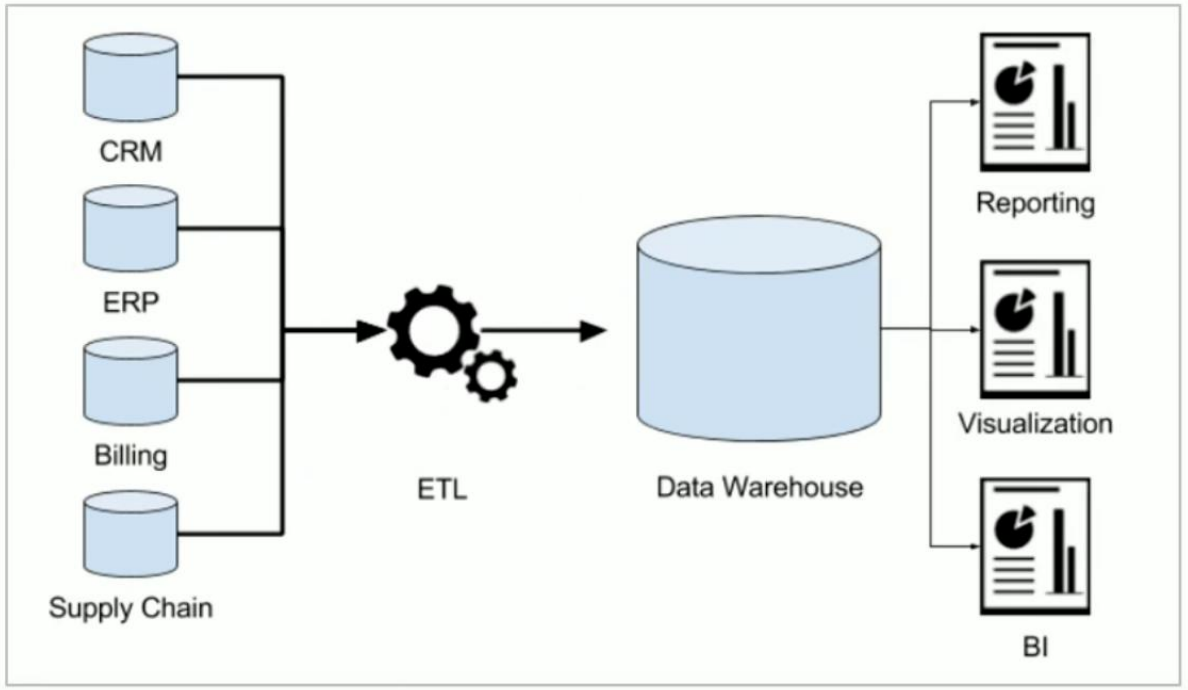
概念

- 操作型处理，叫联机事务处理OLTP（On-Line **Transaction** Processing），主要目标是做数据处理，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。
- 用户较为关心操作的响应时间、数据的安全性、完整性和并发支持的用户数等问题。
- 传统的**关系型数据库系统（RDBMS）作为数据管理的主要手段，主要用于操作型处理**。



概念

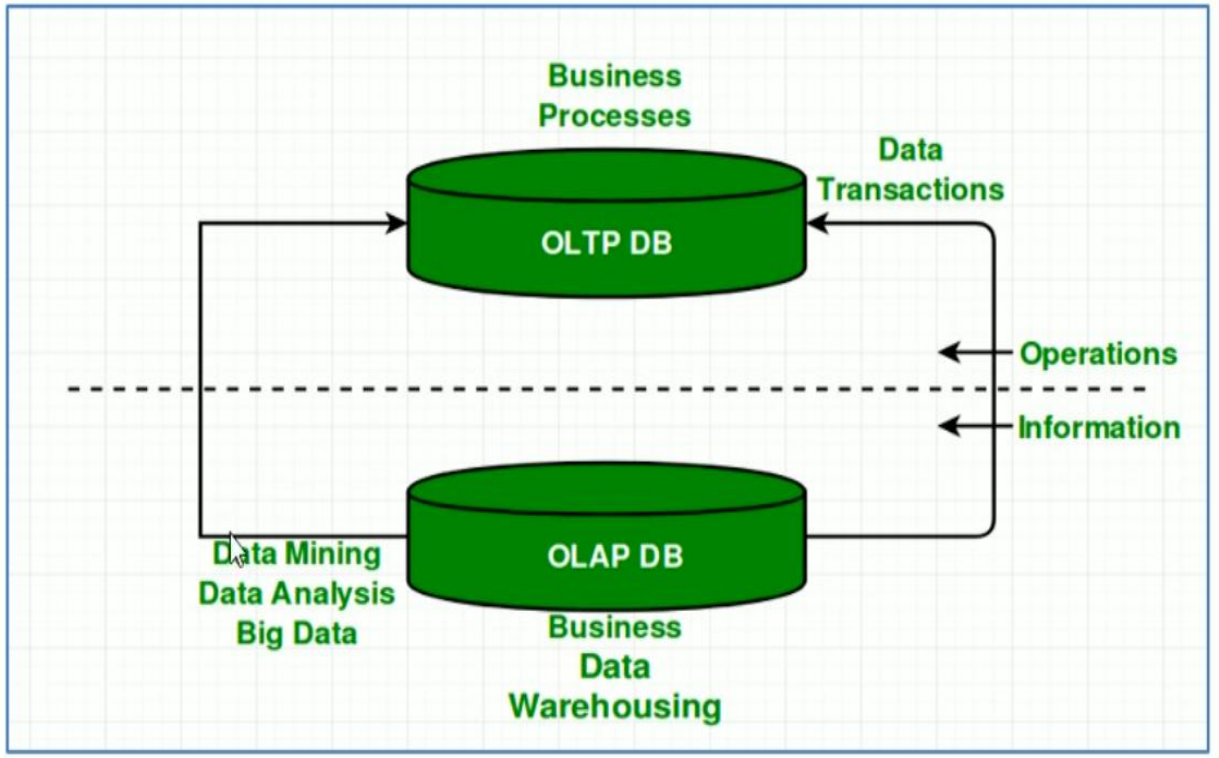
- 分析型处理，叫联机分析处理OLAP（On-Line **Analytical** Processing），主要目标是做数据分析。
- 一般针对某些主题的历史数据进行复杂的多维分析，支持管理决策。
- **数据仓库是OLAP系统的一个典型示例**，主要用于数据分析。



1.1.4 OLTP, OLAP

对比

	OLTP	OLAP
数据源	仅包含当前运行日常业务数据	整合来自多个来源的数据，包括OLTP和外部来源
目的	面向应用，面向业务，支撑事务	面向主题，面向分析，支撑分析决策
焦点	当下	主要面向过去、面向历史 实时数仓除外
任务	读写操作	大量读而很少写操作
响应时间	毫秒	秒、分钟、小时或者天 取决于数据量和查询复杂性
数据量	小数据，MB,GB	大数据，TP，PB

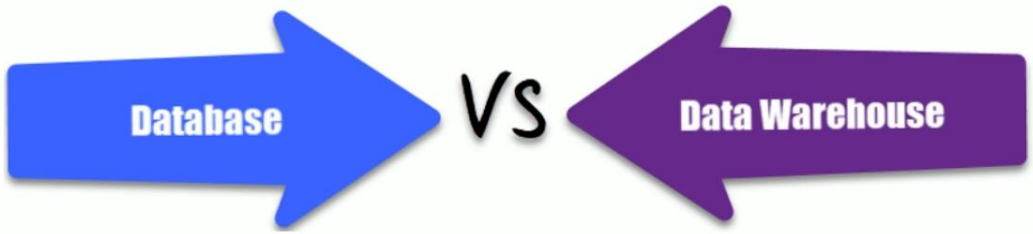


区别

- 数据库与数据仓库的区别实际讲的是OLTP与OLAP的区别。
- OLTP系统的典型应用就是RDBMS，也就是我们俗称的数据库，当然这里要特别强调此数据库表示的是关系型数据库，Nosql数据库并不在讨论范围内。
- OLAP系统的典型应用就是DW，也就是我们俗称的数据仓库。

区别

- 数据仓库不是大型的数据库，虽然数据仓库存储数据规模大。
- 数据仓库的出现，并不是要取代数据库。
- 数据库是面向事务的设计，数据仓库是面向主题设计的。
- 数据库一般存储业务数据，数据仓库存储的一般是历史数据。
- 数据库是为捕获数据而设计，数据仓库是为分析数据而设计。



区别

- 数据仓库（Data Warehouse）是面向整个集团组织的数据，数据集市（Data Mart）是面向单个部门使用的。
- 可以认为数据集市是数据仓库的子集，也有人把**数据集市叫做小型数据仓库**。数据集市通常只涉及一个主题领域，例如市场营销或销售。因为它们较小且更具体，所以它们通常更易于管理和维护，并具有更灵活的结构。

区别

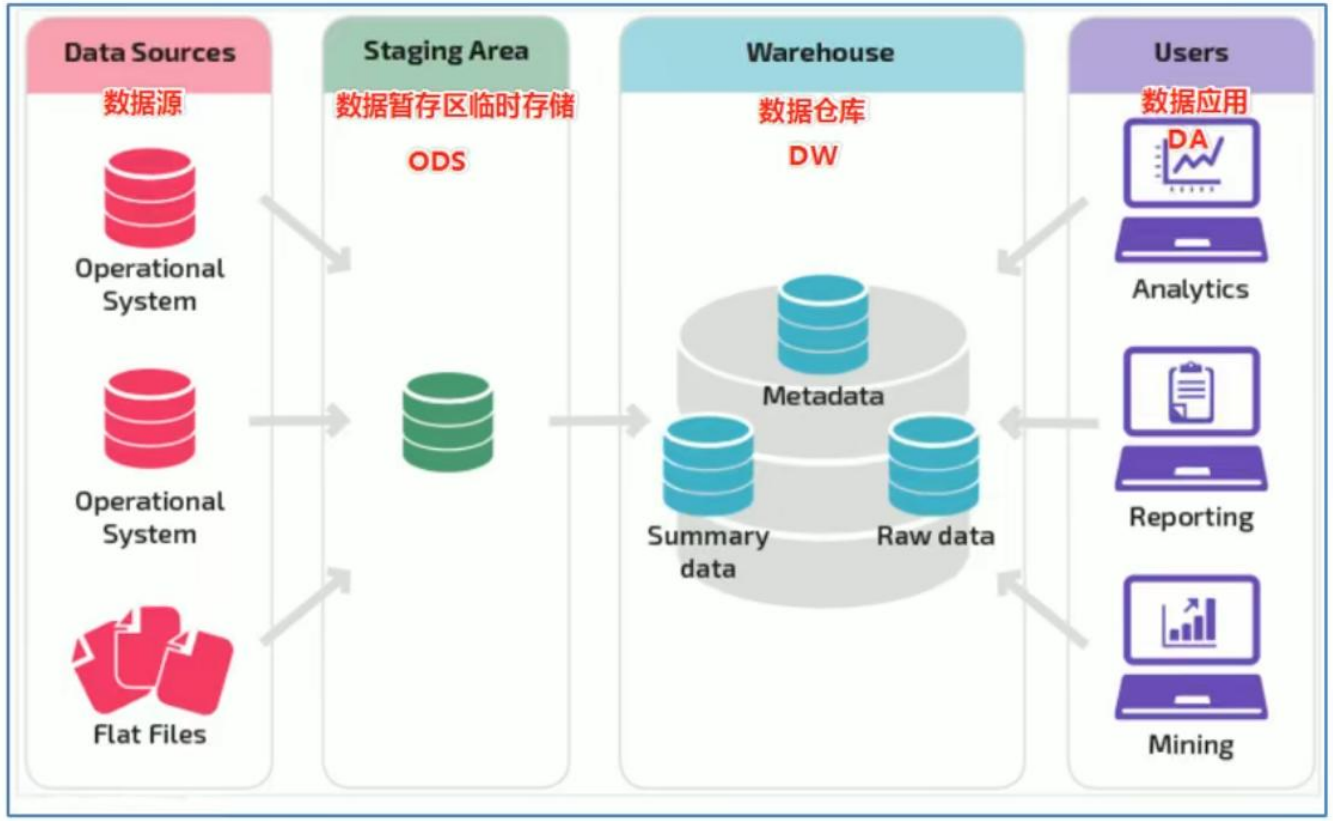
- 下图中，各种操作型系统数据和包括文件在内的等其他数据作为数据源，经过ETL(抽取转换加载）填充到数据仓库中；数据仓库中有不同主题数据，数据集市则根据部门特点面向指定主题，比如Purchasing（采购）、Sales（销售）、Inventory（库存）；
- 用户可以直接基于主题数据开展各种应用：数据分析、数据报表、数据挖掘。



1.1.7 数据仓库分层架构

分层思想和标准

- 数据仓库的特点是本身不生产数据，也不最终消费数据。按照数据流入流出数仓的过程进行分层就显得水到渠成。
- 每个企业根据自己的业务需求可以分成不同的层次。但是最基础的分层思想，理论上分为三个层：操作型数据层（ODS）、数据仓库层（DW）和数据应用层（DA）。
- 企业在实际运用中可以基于这个基础分层之上添加新的层次，来满足不同的业务需求。



(Operational Data Store)
(Data Warehouse)
(Data Application)

阿里巴巴数仓3层架构

- 为了更好的理解数据仓库分层的思想以及每层的功能意义，下面结合阿里巴巴提供的数仓分层架构图进行分析。
- 阿里数仓是非常经典的3层架构，从下往上依次是：ODS、DW、DA。
- 通过元数据管理和数据质量监控来把控整个数仓中数据的流转过程、血缘依赖关系和生命周期。
- 当下我们不做深入探讨，只做宏观了解掌握。



阿里巴巴数仓3层架构：ODS层(Operation Data Store Layer)

- **操作型数据层**，也称之为源数据层、数据引入层、数据暂存层、临时缓存层。
- 此层存放未经过处理的原始数据至数据仓库系统，结构上与源系统保持一致，是数据仓库的数据准备区。
- 主要完成基础数据引入到数仓的职责，和数据源系统进行解耦合，同时记录基础数据的历史变化。



阿里巴巴数仓3层架构: : DW层(Data Warehouse Layer)

- **数据仓库层**，由ODS层数据加工而成。主要完成数据加工与整合，建立一致性的维度，构建可复用的面向分析和统计的明细事实表，以及汇总公共粒度的指标。内部具体划分如下:
  - 公共维度层（DIM）：基于维度建模理念思想，建立整个企业一致性维度。
  - 公共汇总粒度事实层（DWS、DWB）：以分析的主题对象作为建模驱动，基于上层的应用和产品的指标需求，构建公共粒度的汇总指标事实表，以宽表化手段物理化模型。
  - 明细粒度事实层（DWD）：将明细事实表的某些重要维度属性字段做适当冗余，即宽表化处理。



阿里巴巴数仓3层架构: DA层(Data Application Layer)

- **数据应用层**，面向最终用户，面向业务定制提供给产品和数据分析使用的数据。
- 包括前端报表、分析图表、KPI、仪表盘、OLAP专题、数据挖掘等分析。



1.1.7 数据仓库分层架构

思考

1．为什么要对数据仓库进行分层？

2．分层的好处是什么？



好处

分层的主要原因是在管理数据的时候，能对数据有一个更加清晰的掌控，详细来讲，主要有下面几个原因：

- **清晰数据结构**

每一个数据分层都有它的作用域，在使用表的时候能更方便地定位和理解。

- **数据血缘追踪**

简单来说，我们最终给业务呈现的是一个能直接使用业务表，但是它的来源有很多，如果有一张来源表出问题了，我们希望能够快速准确地定位到问题，并清楚它的危害范围。

- **减少重复开发**

规范数据分层，开发一些通用的中间层数据，能够减少极大的重复计算。

1.1.7 数据仓库分层架构

好处

- 把复杂问题简单化
  将一个复杂的任务分解成多个步骤来完成，每一层只处理单一的步骤，比较简单和容易理解。而且便于维护数据的准确性，当数据出现问题之后，可以不用修复所有的数据，只需要从有问题的步骤开始修复。
- 屏蔽原始数据的异常
  屏蔽业务的影响，不必改一次业务就需要重新接入数据

1.1.8 ETL, ELT区别



思考

1．在数仓的分层架构中，ETL是什么动作？

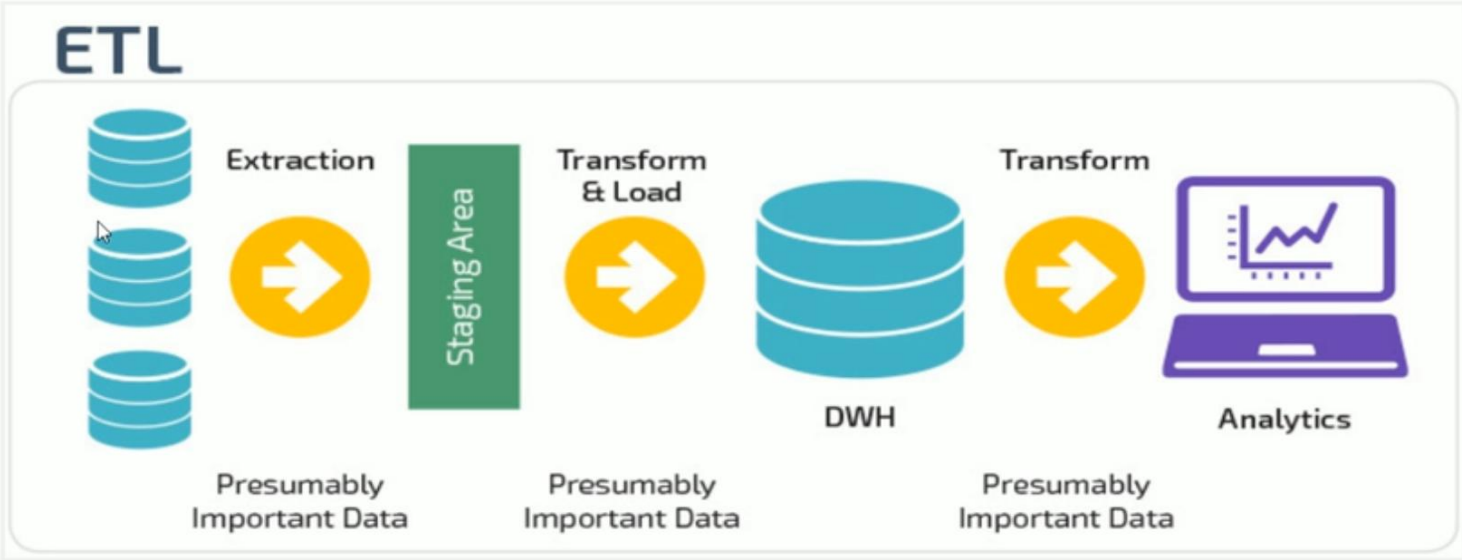
2．ETL和ELT差别在哪？

1.1.8 ETL, ELT区别

ETL 概念

- Extract, Transform, Load ETL

首先从数据源池中提取数据，这些数据源通常是事务性数据库。数据保存在临时暂存数据库中（ODS）。然后执行转换操作，将数据结构化并转换为适合目标数据仓库系统的形式。然后将结构化数据加载到仓库中，以备分析。

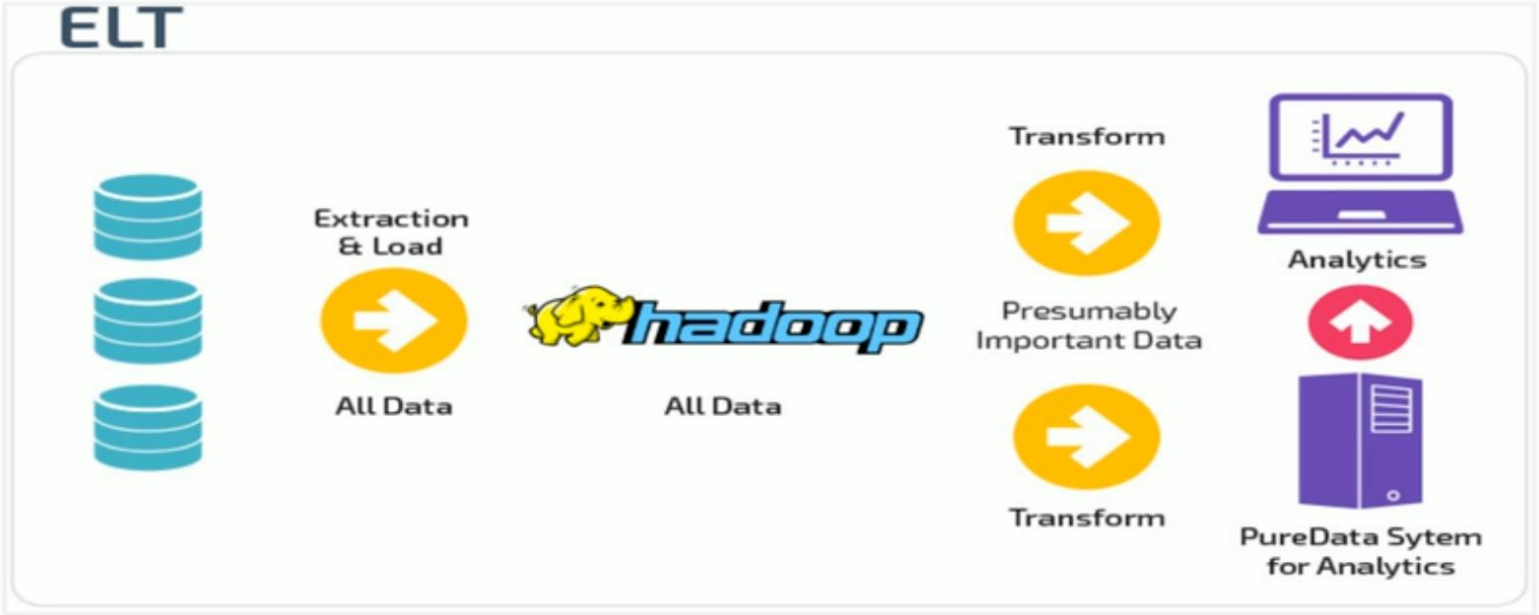


1.1.8 ETL, ELT区别

ELT 概念

- Extract, Load, Transform ELT

使用ELT，数据在从源数据池中提取后立即加载。没有专门的临时数据库（ODS），这意味着数据会立即加载到单一的集中存储库中。数据在数据仓库系统中进行转换，以便与商业智能工具（BI工具）一起使用。**大数据时代的数仓这个特点很明显。**

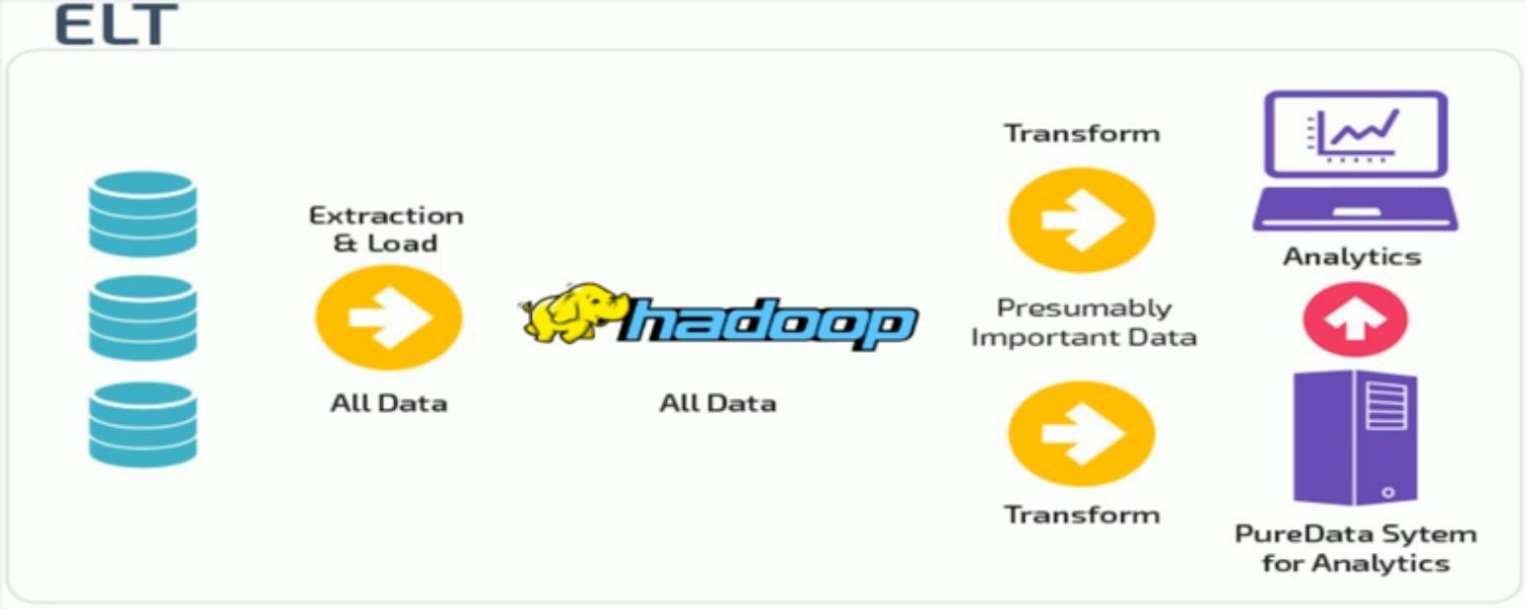


1.1.8 ETL, ELT区别

ELT 概念

- Extract, Load, Transform ELT

使用ELT，数据在从源数据池中提取后立即加载。没有专门的临时数据库（ODS），这意味着数据会立即加载到单一的集中存储库中。数据在数据仓库系统中进行转换，以便与商业智能工具（BI工具）一起使用。**大数据时代的数仓这个特点很明显。**



1.1.9 Hive数据类型测验（4选1）

问题01:

数据仓库的核心理念是什么?

A. 用于快速处理日常事务操作
B. 用于生产并创造新的业务数据
C. 构建面向分析的集成化数据环境，为决策提供支持
D. 主要用于存储和备份企业所有原始数据

1.1.9 Hive数据类型测验（4选1）

问题02:

数据仓库的数据主要来源于哪里？

A. 由数据仓库系统自己生成
B. 仅来源于企业内部的一个业务系统
C. 来源于不同的外部系统（如OLTP系统）
D. 仅来源于手工录入的报表

1.1.9 Hive数据类型测验（4选1）

问题03:

为什么企业中需要独立构建数据仓库，而不是直接在OLTP系统上进行分析?

A. OLTP系统设计目标是支持事务处理，直接分析会增加其负载，且数据分散、历史数据有限
B. OLTP系统无法存储任何业务数据
C. 数据仓库的建设成本永远低于OLTP系统
D. OLTP系统不具备数据读取功能

1.1.9 Hive数据类型测验（4选1）

问题04:

数据仓库的哪一个特征强调其数据是围绕特定分析领域（如“客户”、“销售”）组织的？

A. 集成性 (Integrated)
B. 时变性 (Time-Variant)
C. 非易失性 (Non-Volatile)
D. 面向主题性 (Subject-Oriented)

1.1.9 Hive数据类型测验（4选1）

问题05:

数据仓库的“非易失性”主要指什么？

A. 数据一旦进入数据仓库，就不能被查询
B. 数据进入数据仓库后保持稳定，一般不进行实时更新或删除，反映的是历史快照
C. 数据仓库的数据必须每分钟更新一次
D. 数据仓库只存储当前最新的业务数据

1.1.9 Hive数据类型测验（4选1）

问题06:

ODS层（操作型数据层）在数据仓库分层架构中的主要作用是什么？

A. 直接面向最终用户提供报表和图表
B. 存放未经处理的原始数据，与源系统结构一致，作为数据准备区
C. 负责复杂的数据分析和挖掘任务
D. 替代原始业务数据库进行联机事务处理

1.1.9 Hive数据类型测验（4选1）

问题07:

在阿里巴巴的三层数仓架构中，负责“建立整个企业一致性维度”的是哪一层？

A. ODS层
B. DWD层（明细粒度事实层）
C. DIM层（公共维度层）
D. DA层（数据应用层）

1.1.9 Hive数据类型测验（4选1）

问题08:

数据仓库分层的主要好处不包括以下哪一项?

A. 使数据结构更清晰，便于定位和理解
B. 便于追踪数据血缘关系，快速定位问题
C. 增加数据处理的复杂度以提升安全性
D. 通过开发通用中间层数据，减少重复计算

1.1.9 Hive数据类型测验（4选1）

问题09:

关于ETL和ELT的主要区别，以下描述正确的是？

A. ETL是先转换后加载，ELT是先加载后转换
B. ETL不需要转换数据，ELT需要复杂转换
C. ETL主要用于大数据实时处理，ELT用于小型数据库
D. ETL和ELT在数据处理顺序上没有区别

1.1.9 Hive数据类型测验（4选1）

问题10:

数据仓库与数据库（OLTP）的主要区别是什么？

A. 数据仓库用于事务处理，数据库用于分析
B. 数据仓库是面向主题设计用于分析历史数据，数据库是面向应用设计用于捕获和处**理**当前业务数据
C. 数据仓库必定比数据库的规模小
D. 数据仓库将取代数据库成为唯一的数据存储系统