**大数据的数据抽取和集成需求**

在信息研究领域，信息抽取技术是一项必不可少的关键技术。面对如此海量的信息空间，如何更快更准确的抽取出用户感兴趣的内容是一个迫切需要解决的 问题，也是信息挖掘技术的一个重要研究方向。信息抽取不同于信息检索等信息 处理技术，它需要对文本进行命名实体的识别，并抽取出实体之间的关系，再加 上中文文本中词语的灵活多变、构词复杂且没有明显的标志，因此对中文命名实 体的识别及关系的抽取就显得更加困难。

目前，信息抽取的主要方法有两种，一种是基于知识库算法，这种方法需要建立一些规则，虽然这种方法的准确率较高，但是这种规则的确定是比较困难的，对编写者有较高的要求，且移植性不高；另一种是基于统计的机器学习算法，这种算法采用不同的模型，并利用人工标注的训练集进行学习，对于新的数据集则采用模型算出其相关的概率，并以此来得到最终的结果。这种方法代价较小，性能较高，便于移植，所以是当前研究的热点。

随着网络信息量的增大，对海量文本数据的信息抽取也变得复杂起来。如何利用海量文本数据来对实体关系进行更为准确的抽取是本文所要研究的一个重要问题。而这种大数据量的计算对算法的性能要求较高，采用何种策略来应对这种繁重的计算任务也是一个重要的问题。

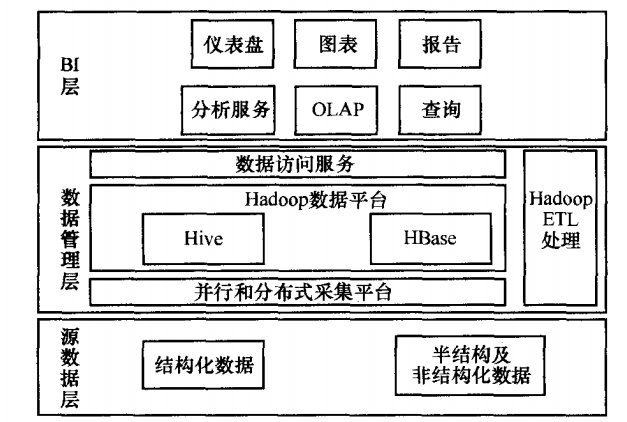
大数据通常是由来源、主体或格式不同的数据合并而成，例如来自不同地区的调查数据，来自不同市场的金融数据，来自不同实验室的基因数据等。这种基于多个数据集的建模十分常见，了解不同子样本间的异质性(heterogeneity or difference)和同质性(homogeneity or similarity)是大数据分析的两个重要目标。但它的建模比较特殊，一方面，由于不同来源的数据存在差异，各不同数据源的同一变量的系数显著性和估计值可能存在差异，传统的处理方法是简单合并所有样本，建立统一模型，但是这种方法过于笼统，忽略了数据问的异质性(heterogeneity)；另一方面，也不能分开各自建立模型，因为这样会忽略各个数据集之间的关联性。数据集成分析方法同时兼顾这两方面，通过目标函数综合不同地区的数据，从统计角度考虑数据的异质性和同质性，以多个变量为研究目标，充分考虑了不同地区问相互影响，同时求解多个模型。数据集成分析方法把不同来源、格式、特点性质的数据集中起来，相对于单一数据集模型，整合了更多的原始信息，能解决因地域、时间等因素造成的样本差异而引起的建模不稳定，在模型解释性和预测方面都具有显著优势。

传统数据仓库在这个数据爆炸的时代已经无法满足用户对数据应用的需求了，当然，传统数据仓库架构也仍在不断发展演化，这一点不容置疑，但也因其存在如上文所述的缺陷而在面对海量异构数据时表现的不够令人满意。现在有相当一部分的企业都拥有自己的数据仓库用来数据分析和数据挖掘，对于采用传统数据仓库的企业而言，大数据时代带来的机会就是使之能够利用过去无法通过传统数据仓库架构利用的数据，而Hadoop如今则为企业提供了运用新数据来源，使分析更加智慧的能力。因此，越来越多的企业开始构建自己的大数据平台来应对这个数据爆炸时代的数据应用需求，但是，同时也发现了一个问题：企业花费大量人力物力构建的传统数据仓库却成为了孤岛。于是，越来越多的企业开始考虑是否有一种通用模式可以将传统数据仓库和大数据平台整合在一起，建立一个统一的数据存储和数据处理架构，使得原来的传统数据仓库物尽所用，同时也能应对大数据的冲击?这成为一个急待解决的问题，该方案在已有传统数据仓库的基础上提供Hadoop的支持，弥补传统数据仓库在海量数据处理、存储等方面的不足，也可以依靠Hadoop的横向扩展能力突破单节点的传统数据仓库在存储和计算能力上的瓶颈，同时也重用企业费时费力构建好的传统数据仓库。

**比较传统面向数据仓库和大数据集成的各自目的和特点**

数据仓库是在企业管理和决策中面向主题的、集成的、与时间相关的、不可修改的数据集合，与其它数据库应用不同的是，数据仓库更像一种过程，对分布在企业内部各处的业务数据的整合、加工和分析的过程。建设数据仓库的最终目的是为企业提供决策支持。传统关系型数据仓库存在的一些不足，最为明显的有：一、现今企业传统数据仓库和关系型数据库擅长处理结构化数据，并且可以存储大量的数据，但传统数据仓库通常采用小机加盘阵高性能一体机建设，成本非常高，很大程度上限制了其扩展能力，即遇到性能瓶颈时，存储能力与计算能力的扩展是个很大的难题；二、传统数据仓库“擅长处理结构化数据”限制了可处理的数据种类，同时这种缺点还影响到传统数据仓库在面对海量异构数据时的表现，这通常意味着不少有价值的数据源是传统数据仓库无法涉及的。

从现有数据仓库建设理论和经验人手，引入部分大数据技术，特别是实现非结构化数据的收集、存储和处理是一种比较可行的方法。例如，将Hadoop技术应用于对数据的采集、ETL、存储、处理，开发提供给传统的数据仓库BI工具，其架构如图所示。



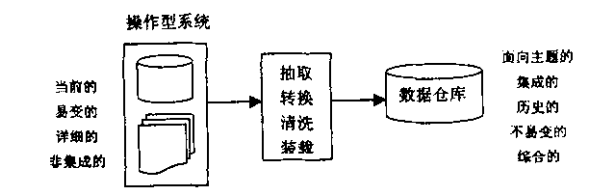
在这个架构中，主要改变了传统数据仓库单节点数据处理和存储的方式，利用了Hadoop强大的数据处理能力，将各类数据处理成结构化数据，向上提供给传统BI工具，对数据进行分析和结果展示。在这个基础架构之上。可以 根据数据处理速度及分析响应能力，逐层进行细化及分解，优化组合 MPP数据库、内存数据库等各类技术，从而满足BI层分析展示的需要同。另外，还可以在数据管理层利用传统数据仓库和Hadoop共同合作，由传统数据仓库工具对结构化数据进行处理，由Hadoop对更大规模的非结构化数据进行预处理，并将两者处理后的数据存储至结构化数据库中，以便于BI层进行分析和展示。 除了技术层面上数据仓库与大数据的融合之外，非常重要的一点是传统数据仓库在具体应用理论、方法和实施上的成功经验，如基于数据驱动的螺旋式开发方法、调研及需求设计、ETL、数据建模、元数据管理唧等各个方面同样具有很多的可借鉴性与融合性。

大数据需求的产生背景与数据仓库类似，人们希望利用新技术处理越来越多的数据、挖掘更大的数据价值。因此从需求角度来说．无论是数据库、数据仓库还是大数据都是解决不同需求、处理不同级别数据量的技术，它们之间并无冲突．所以短期内并不会出现由谁取代谁的结果，而应该是针对不同需求和现状进行技术选择，各种技术相互补充、相互协作。

**大数据集成的主要技术**

**ETL**

ETL，即数据抽取、转换、清洗、装载的过程，是构建数据仓 库最重要的步骤之一。用户从数据源抽取出所需的数据，经过 数据清洗，最终按照预先定义好的数据仓库模型，将数据加载到 数据仓库中去。如图所示。



创建数据仓库最重要的一个步骤是将数据从各种操作型数据库系统中抽取出来，排除数据中的数据缺陷，完成一系列转换，最后将数据加载到数据仓库。这个过程非常繁杂，是整个数

据仓库建设过程中工作量最大的一部分。下面将详细介绍它们的设计细节。

（l）数据抽取 并不是源数据库的所有细节数据对于数据仓库的主题域都是有用的，必须根据已确定主题的需要，从原有操作型数据库中抽取相关数据到数据仓库。一般在设计数据

抽取时要考虑以下几个方面：

源数据库和目标数据库各自的数据库格式是否一致？

从源数据库中要访问哪些文件和表？

从源数据库中可以提取哪些字段，抽取记录的条件是什么？

目标数据库中的表结构是什么？

（2）数据转换 数据仓库中的数据往往来自一个或多个异构的数据库系统，这些数据源之间往往存在着不一致的问题，如不一致的字段长度、不一致的赋值等。数据不一致会严重影

响数据仓库的数据质量。数据转换就是处理这些不一致性的过程。数据转换一般包括两个方面的内容。一方面是数据名称及格式的统一，如统一的命名、统一的数据格式、统一的计量单位等等。另一方面，数据仓库中存在着源数据库中可能不存在的数据，因此需要创建新的数据逻辑视图并进行以下的转换：把一个字段的各个部分隔成两个或多个字段。把一个记录的两个或多个字段组合成一个字段。把来自多个记录的字段结合成一个记录。

（3）数据清洗 数据质量是决定信息价值的关键因素。高质量的信息导致高质量的决策，而低质量的信息将导致低劣的决策。数据仓库是分析决策的基础，所以数据仓库中数据的

准确性是非常重要的。传统数据库中的数据大都存在错误，这是活生生的事实。尽管数据仓库项目能把注意中心放在数据质量问题和引导未来的进一步改进上，但数据仓库开发者却并不能改变数据仓库的基础——历史数据的质量。因此，有必要对进入数据仓库的数据进行全面检查并使他们尽可能无差错。这一过程就称作数据清洗。数据清洗应该处理许多可能存在的错误类型。这些类型包括数据源中丢失数据和有错误数据，还包括两个或多个数据源

里的不一致数据和冲突数据，所有这些都必须处理。有时如果系统不能自动处理，还必须借助于手工操作来完成。

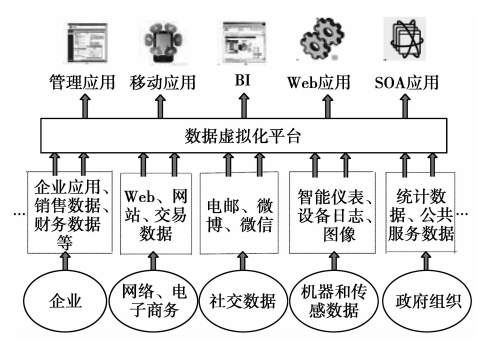
（4）数据装载 这一步所进行的工作是运行以上的处理步骤，将数据装入到数据仓库。主要的工作包括确定数据装入的次序、载入初始数据等。

2. 3 ETL 工具选择原则目前已有众多商用数据仓库产品走向市场。IBM、OracIe、Sybase、CA、NCR、SAS、Microsoft 等公司已相继推出了各自的数据仓库解决方案，它们的 ETL 工具也各有其优势和不足。在选择 ETL 工具时必须遵守以下原则：支持多种数据源，如 DBMS、电子表格、平面文件。支持多种平台，支持多种数据库。具有规范的数据访问接口。具有灵活的可编程性和调用外部程序的功能。工具生成的代码必须是在开发环境中可维护的。能只抽取满足指定条件的数据和源数据的指定部分。能在抽取过程中进行数据类型转换和字符集转换并能计算生成衍生的字段。具有直观的视图、灵活的配置，能自动调用以定期实现管理工作。

数据同步

数据虚拟化是针对异构、多源、多所有者的数据集，通过对数据资源的逻辑虚拟化，实现数据的集成管理并提供统一的访问接口，以便为各种数据消费需求提供跨数据源整合的数据服务。数据消费者不用关心数据从哪些数据源来，如何进行集成，以及数据的存储位置与方式、访问接口等细节，数据虚拟化将这些技术细节对用户应用隐藏，通过一个逻辑抽象层集成管理、整合各个数据源。数据的清洗、转换与加载在逻辑抽象层完成，实现用户以完全透明的方式访问所有的数据源。同时，数据处理整合的周期变得更短更灵活，并且能够确保数据的统一访问、建模、部署、优化和管理，逻辑上就像统一的一个数据资源，用户只需通过统一的接口进行访问即可数据虚拟化。

如图所示，来自于企业、网络、感知、社交等等对象产生或提供的数据源，如企业销售数据、网站交易数据、统计数据等，具有数据多样化、异构化、分布式、不同开放性等特点，数据虚拟化系统提供跨数据源、跨平台的数据集成、管理与整合服务，为各种应用需求提供所需的数据服务。



通过数据虚拟化，对数据消费者而言，不需提取和存储大量的异构数据集，只需查询请求已发布的数据服务，然后通过对应的 ＡＰＩｓ 查询获取所需数据资源，从而大大简化用户对各个分散数据源信息的访问。数据虚拟化在数据质量、缓存、查询处理等方面也有了较大的改善。通常在数据集成中引起数据质量问题的根源是对源数据的多次复制转移，而数据

虚拟化恰恰从根本上解决了这个问题。数据虚拟化通过逻辑抽象层实现源数据的共享，而且不需关心数据源的位置，避免了数据多次复制、转移、加载过程导致的数据不一致等问题，从而提高了数据质量，降低数据出错的风险。

在数据集成方面，郭树盛等提出一种基于数据虚拟化的新型ＴＥＬ方法，利用数据虚拟化创建虚拟表，在数据抽取和加载之前完成虚拟转换任务，避免应用临时数据存储区来暂存抽取过来的多源数据，减少数据缓存压力。此外文中还针对传统 ＥＴＬ中存在的数据临时存储、查询响应慢问题，提出基于数据虚拟化的 ＳｅａＢａｓｅ架构是一个关系云数据库而非一组数据库的联邦）。由于ＳＱＬ查询中可能包含不同应用的部分相同数据，利用数据虚拟化在 ＳｅａＢａｓｅ中缓存元数据而非源数据，一方面减少查询的响应时间，一方面也优化了数据存储。所以，数据虚拟化在数据集成方面带来了很多优势，为数据集成提供了一个新的方向。

数据虚拟化在未来将有广泛的应用。对于企业而言，通过数据虚拟化在数据仓库、应用程序、文本数据等数据源之上建立整合所有系统信息的数据层，可以减少数据的存储和维护成本，避免企业内部及外部数据孤岛的存在。对于一些难以用数字解释和认知的对象，如人类复杂多变的表情不能用单纯的数据进行准确表达，将数据虚拟化应用于反映认知对象信息的海量数据分析，有助于快速进行认知和决策。另外，通过数据虚拟化可以将多源数据进行整合，并以数据服务的方式发布到外部，这样就会催生出大量新的数据服务和应用。

DaaS

DDAS的出现离不开云计算带来的大数据时代优势，在现代社会，数据作为一种传递信息的服务形式能够为他人活动提供各种便利，比如上网查询资料、资源共享、数据收集与整合等，数据传递的各类信息都对人们的生活产生深刻的影响，这种影响力得到贯彻的途径即数据服务。 可以说，DAAS 是大数据时代发展到一定阶段必然会出现的产物。根据互联网统计数据，截止到 2012 年，网络数据量已经从TB级别跃升到 PB、EB 乃至 ZB 级别， 期间数据增长的速度不断加快，发展到 2012年，人均产生数据量高达200GB 以上，超越古代历史上所有文献历史资料的总量。尤其是从 2012 年-2013 年，数据的增长和产生达到了前所未有的速度与规模，由此可见，大数据时代潮流来临的必然，数据作为互联网时代的宝贵财富，既可以为决策提供参考依据，也可以转化为各种各样的财富。DAAS 本质是数据服务，也可以进一步理解为桌面云服务，这个概念最早由京华科讯提出，意为通过云计算理念将桌面作为服务形式提供给用户，可以说是 SaaS（软件即服务）内容的一个重要构成部分， 通过桌面虚拟化技术和 IAAS 架构将构建在此基础上的桌面服务分发给用户。这种服务的优势在于安全性能更佳、硬件成本和管理成本降低，能够实现快速部署，支持功能强大，管理简单，容灾能力强，业务连续能力佳，用户所需花费更低。

云计算的出现为高校整合教育资源提供了崭新的途径与方法， 云计算技术参与构建教学资源网络平台将会优化广大师生的网络学习环境和资源共享环境。 将云计算应用于高校资源库的建设有利于为教学模式创新提供强有力支持， 提升教学资源利用率和利用水平，为现代教育信息化理论的丰富和实践提供宝贵经验。作为以服务为核心思想的大数据时代，采用云服务成为发展主流，其显著的优势使得能够在高校信息化进程建设方面提供切实保

障。 比如云端数据的储藏，不仅有效节省了空间资源，还通过建立权限不一的云端服务平台实现资源的高度整合与共享。 云端提供的软件服务将无数软件置于服务器， 解决了各种标准和协议的网络准入，并且可实现多人网络协作使用，这种数据服务促使桌面云服务有了更好的发展空间与市场。新时期云计算、云服务的出现满足了广大用户多元化、 多层次的网络需求， 个人定制服务成为主流， 这种更加人性化的设置和服务无疑将会促使资源网络的构建变得更加普遍和强势， 用户通过个人账号即可享受高质量的数据

服务。