b) 数据预处理

现实世界中数据大体上都是不完整，不一致的脏数据，无法直接进行信息的抽取和挖掘，或挖掘结果差强人意。为了提高信息抽取和数据挖掘的质量产生了数据预处理技术。数据预处理技术在对所收集数据进行分类或分组之前使用，可以大大提高分类、分组和挖掘的质量，降低实际运行所需要的时间。一般来说，数据预处理有多种方法：数据清洗，数据集成，数据转换，数据归约等。

1 数据清洗

数据清洗的主要原理是，利用有关技术如数理统计、数据挖掘或预定义的清理规则将脏数据转化为满足数据质量要求的数据。主要处理的脏数据类型有三种：

1. 残缺数据

这一类数据主要是一些应该有的信息缺失，如供应商的名称、分公司的名称、客户的区域信息缺失、业务系统中主表与明细表不能匹配等。对于这一类数据过滤出来，按缺失的内容分别写入不同Excel文件向客户提交，要求在规定的时间内补全。补全后才写入数据仓库。

1. 错误数据

这一类错误产生的原因是业务系统不够健全，在接收输入后没有进行判断直接写入后台数据库造成的，比如数值数据输成全角数字字符、字符串数据后面有一个回车操作、日期格式不正确、日期越界等。这一类数据也要分类，对于类似于全角字符、数据前后有不可见字符的问题，只能通过写SQL语句的方式找出来，然后要求客户在业务系统修正之后抽取。日期格式不正确的或者是日期越界的这一类错误会导致ETL运行失败，这一类错误需要去业务系统数据库用SQL的方式挑出来，交给业务主管部门要求限期修正，修正之后再抽取。

1. 重复数据

对于这一类数据——特别是维表中会出现这种情况——将重复数据记录的所有字段导出来，让客户确认并整理。

一般来说，数据清理是将数据库精简以除去重复记录，并使剩余部分转换成标准可接收格式的过程。数据清理标准模型是将数据输入到数据清理处理器，通过一系列步骤“ 清理”数据，然后以期望的格式输出清理过的数据（如上图所示）。数据清理从数据的准确性、完整性、一致性、惟一性、适时性、有效性几个方面来处理数据的丢失值、越界值、不一致代码、重复数据等问题。

数据清理一般针对具体应用，因而难以归纳统一的方法和步骤，但是根据数据不同可以给出相应的数据清理方法。

1. 解决不完整数据（ 即值缺失）的方法

大多数情况下，缺失的值必须手工填入（ 即手工清理）。当然，某些缺失值可以从本数据源或其它数据源推导出来，这就可以用平均值、最大值、最小值或更为复杂的概率估计代替缺失的值，从而达到清理的目的。

1. 错误值的检测及解决方法

用统计分析的方法识别可能的错误值或异常值，如偏差分析、识别不遵守分布或回归方程的值，也可以用简单规则库（ 常识性规则、业务特定规则等）检查数据值，或使用不同属性间的约束、外部的数据来检测和清理数据。

1. 重复记录的检测及消除方法

数据库中属性值相同的记录被认为是重复记录，通过判断记录间的属性值是否相等来检测记录是否相等，相等的记录合并为一条记录（即合并/清除)。合并或者清除是消重的基本方法。

1. 不一致性（ 数据源内部及数据源之间）的检测及解决方法

从多数据源集成的数据可能有语义冲突，可定义完整性约束用于检测不一致性，也可通过分析数据发现联系，从而使得数据保持一致。目前开发的数据清理工具大致可分为三类：数据迁移工具允许指定简单的转换规则，如：将字符串gender替换成sex。sex公司的PrismWarehouse是一个流行的工具，就属于这类。

数据清洗工具使用领域特有的知识（ 如，邮政地址）对数据作清洗。它们通常采用语法分析和模糊匹配技术完成对多数据源数据的清理。某些工具可以指明源的“ 相对清洁程度”。工具Integrity和Trillum属于这一类。数据审计工具可以通过扫描数据发现规律和联系。因此，这类工具可以看作是数据挖掘工具的变形。

2 数据集成

数据集成通过应用间的数据交换从而达到集成，主要解决数据的分布性和异构性的问题，其前提是被集成应用必须公开数据结构，即必须公开表结构，表间关系，编码的含义。在企业数据集成领域，已经有了很多成熟的框架可以利用。通常采用联邦式、基于中间件模型和数据仓库等方法来构造集成的系统，这些技术在不同的着重点和应用上解决数据共享和为企业提供决策支持。在这里将对这几种数据集成模型做一个基本的分析。

1. 联邦数据库系统

联邦数据库系统( FDBS)由半自治数据库系统构成，相互之间分享数据，联盟各数据源之间相互提供访问接口，同时联盟数据库系统可以是集中数据库系统或分布式数据库系统及其他联邦式系统。在这种模式下又分为紧耦合和松耦合两种情况，紧耦合提供统一的访问模式，一般是静态的，在增加数据源上比较困难；而松耦合则不提供统一的接口，但可以通过统一的语言访问数据源，其中核心的是必须解决所有数据源语义上的问题。

1. 中间件模式

中间件模式通过统一的全局数据模型来访问异构的数据库、遗留系统、Web 资源等。中间件位于异构数据源系统(数据层) 和应用程序(应用层) 之间，向下协调各数据源系统，向上为访问集成数据的应用提供统一数据模式和数据访问的通用接口。各数据源的应用仍然完成它们的任务，中间件系统则主要集中为异构数据源提供一个高层次检索服务。

中间件模式是比较流行的数据集成方法，它通过在中间层提供一个统一的数据逻辑视图来隐藏底层的数据细节，使得用户可以把集成数据源看为一个统一的整体。这种模型下的关键问题是如何构造这个逻辑视图并使得不同数据源之间能映射到这个中间层。

1. 数据仓库模式

数据仓库是在企业管理和决策中面向主题的、集成的、与时间相关的和不可修改的数据集合。其中，数据被归类为广义的、功能上独立的、没有重叠的主题。这几种方法在一定程度上解决了应用之间的数据共享和互通的问题，但也存在以下的异同:联邦数据库系统主要面向多个数据库系统的集成，其中数据源有可能要映射到每一个数据模式，当集成的系统很大时，对实际开发将带来巨大的困难。

数据仓库技术则在另外一个层面上表达数据之间的共享，它主要是为了针对企业某个应用领域提出的一种数据集成方法，也就是我们在上面所提到的面向主题并为企业提供数据挖掘和决策支持的系统。

1. 数据转换

由于数据量的不断增加，原来数据构架的不合理可能导致数据不能满足各方面的要求。由数据库的更换，数据结构的更换，从而需要数据本身的转换。常用的数据转换方式及其适用场景有如下五种：

1. 对数转换

将原始数据自然对数值作为分析数据，如果原始数据中有零，可以在底数中加上一个小数值。此种转换的适用情况有：

1. 部分正偏态资料。
2. 等比资料。
3. 各组数值和均值比值相差不大的资料。
4. 平方根转换

此种转换的适用情况有：

1. 服从泊松分布的资料。
2. 轻度偏态资料。
3. 样本的方差和均数呈正相关的资料。
4. 变量的所有个案为百分数，并且取值在0%～20%或者80%～100%的资料。
5. 平方根反正弦转换

此种转换的适用情况有：变量所有个案为百分数，并且取值广泛的资料。

1. 平方转换

此种转换的适用情况有：

1. 方差和均数的平方呈反比。
2. 资料呈左偏。
3. 倒数变换

此种转换的适用情况有：和平方转换相反，需要方差和均数的平方呈正比，但是，倒数转换需要资料中没有接近或者小于零的数据。

4 数据规约

数据归约是指在尽可能保持数据原貌的前提下，最大限度地精简数据量（完成该任务的必要前提是理解挖掘任务和熟悉数据本身内容）。数据归约技术可以用来得到数据集的归约表示，它虽然小，但仍大致保持原数据的完整性。这样，在归约后的数据集上挖掘将更有效，并产生相同(或几乎相同)的分析结果。数据归约主要有两个途径：属性选择和数据采样，分别针对原始数据集中的属性和记录。数据规约主要分为三类：

1. 特征归约

特征归约是从原有的特征中删除不重要或不相关的特征，或者通过对特征进行重组来减少特征的个数。其原则是在保留、甚至提高原有判别能力的

同时减少特征向量的维度。特征归约算法的输入是一组特征，输出是它的一个子集。在领域知识缺乏的情况下进行特征归约时一般包括3个步骤:

1. 搜索过程:在特征空间中搜索特征子集，每个子集称为一个状态由选中的特征构成。
2. 评估过程:输入一个状态，通过评估函数或预先设定的阈值输出一个评估值搜索算法的目的是使评估值达到最优。
3. 分类过程:使用最终的特征集完成最后的算法。

特征归约处理的效果：

1. 更少的数据，提高挖掘效率
2. 更高的数据挖掘处理精度
3. 简单的数据挖掘处理结果
4. 更少的特征。
5. 样本归约

样本都是已知的，通常数目很大，质量或高或低，或者有或者没有关于实际问题的先验知识。样本归约就是从数据集中选出一个有代表性的样本的子集。子集大小的确定要考虑计算成本、存储要求、估计量的精度以及其它一些与算法和数据特性有关的因素。

初始数据集中最大和最关键的维度数就是样本的数目，也就是数据表中的记录数。数据挖掘处理的初始数据集描述了一个极大的总体，对数据的分析只基于样本的一个子集。获得数据的子集后，用它来提供整个数据集的一些信息，这个子集通常叫做估计量，它的质量依赖于所选子集中的元素。取样过程总会造成取样误差，取样误差对所有的方法和策略来讲都是固有的、不可避免的，当子集的规模变大时，取样误差一般会降低。一个完整的数据集在理论上是不存在取样误差的。与针对整个数据集的数据挖掘比较起来，样本归约具有以下一个或多个优点:减少成本、速度更快、范围更广，有时甚至能获得更高的精度。

1. 特征值归约

特征值归约是特征值离散化技术，它将连续型特征的值离散化，使之成为少量的区间，每个区间映射到一个离散符号。这种技术的好处在于简化了数据描述，并易于理解数据和最终的挖掘结果。

特征值归约可以是有参的，也可以是无参的。有参方法使用一个模型来评估数据，只需存放参数，而不需要存放实际数据;有参的特征值归约有以下两种:

1. 回归:线性回归和多元回归
2. 对数线性模型:近似离散多维概率分布。

无参的特征值归约有3种:

1. 直方图:采用分箱近似数据分布，其中V-最优和MaxDiff直方图是最精确和最实用的
2. 聚类:将数据元组视为对象，将对象划分为群或聚类，使得在一个聚类中的对象“类似”而与其他聚类中的对象“不类似”在数据归约时用数据的聚类代替实际数据;
3. 选样:用数据的较小随机样本表示大的数据集，如简单选择n个样本(类似样本归约)、聚类选样和分层选样等。