表结构数据特征

04

表结构数据

以字段或记录作为数据的引用、操作及计算的基本单位的数据

字段：整列数

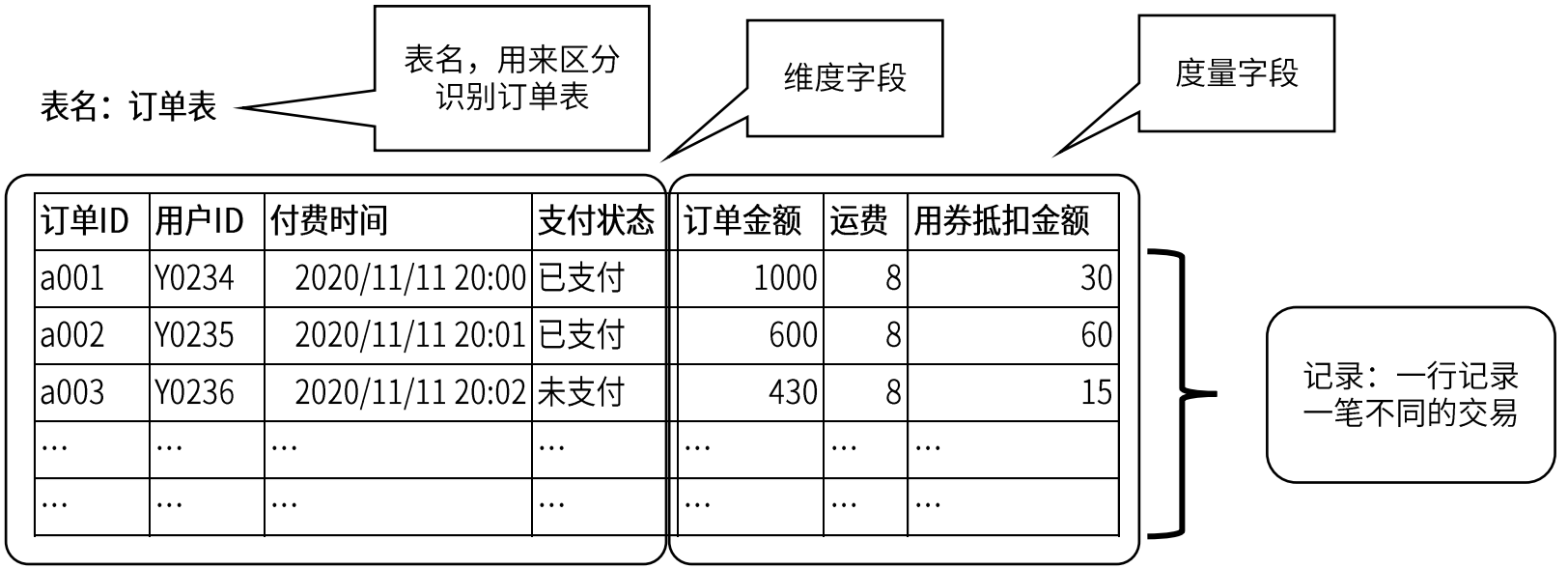
记录：整行数

维度：业务角度

度量: 业务行为结果

维度字段：文本型

度量字段: 数值型



订单ID	用户ID	付费时间	支付状态	订单金额	运费	用券抵扣金额
a001	Y0234	2020/11/11 20:00	已支付	1000	8	30
a002	Y0235	2020/11/11 20:01	已支付	600	8	60
a003	Y0236	2020/11/11 20:02	未支付	430	8	15
…	…	…	…	…	…	…
…	…	…	…	…	…	…

事实表及维度表

维度表：只包含维度信息的表

事实表： 既包含维度信息又包含度量信息的表

事实表：销售表
订单号
销售日期
客户ID
销售金额
……

事实表：采购表
进货单号
进货日期
进货金额
……
……

事实表：库存表
库存编码
产品编码
库存日期
库存金额
……

维度表：产品表
产品编号
产品名称
品牌编号
……
……

维度表：品牌表
品牌编号
品牌名称
……
……
……

维度表：客户表
客户编号
客户名称
公司规模
所属行业
……

表结构数据特征-1

以字段或记录作为数据的引用、操作及计算的基本单位的数据

第一行为标题行

第二行以后称为记录

字段名不能重名

一个字段只能有一种数据类型

订单ID	订单年份	订单月份	产品名称	城市	订单金额
a01	2015	11	Touring Tire Tube	黑龙江	5
a02	2015	11	Short-Sleeve Classic Jersey S	内蒙古	54
a03	2015	11	Touring-1000 Blue 46	吉林	2384
a04	2015	7	Mountain-200 Silver 38	辽宁	2320
…	…	…	…	…	…

表结构数据特征-2

所有字段记录行数相同

- 方形结构
- 记录行数相同
- 存在空值
- 处理缺失值

订单ID	用户ID	付费时间	支付状态	订单金额	运费	用券抵扣金额
a001	Y0234	2020/11/11 20:00	已支付	1000	8	30
a002	Y0235	2020/11/11 20:01	已支付	600	8	60
a003	Y0236	2020/11/11 20:02	未支付	430	8	15
	Y0237	2020/11/11 20:02	已支付	680	8	43

缺少一行信息，不是表结构数据

订单ID	用户ID	付费时间	支付状态	订单金额	运费	用券抵扣金额
a001	Y0234	2020/11/11 20:00	已支付	1000	8	30
a002	Y0235	2020/11/11 20:01	已支付	600	8	60
a003	Y0236	2020/11/11 20:02	未支付	430	8	15
a004	Y0237	2020/11/11 20:02	已支付	680	8	43

订单表：数据表都满足不同字段记录行数相同的特征

处理缺失值-1

根据数据类型以及生成信息重要程度的不同，使用不同方法处理缺失值

文本型字段

影响不大：以选择不进行处理，或者也可以用其他没有实际业务含义的文本字符对缺失值进行替换

影响大：招业务人员进行确认后替换，或者与业务人员核实后删除

在张三负责的商机信息中的销售阶段为“null”值，需要找销售人员张三进行核实，并用正确的有效信息替换“null”值

商机编号	客户ID	销售人员	产品编号	销售阶段	赢单率	建立商机日	预计签约日	预计成交金额（$M）
A1550	C1550	赵大	F-009	投入	0.15	2019/1/8	2020/1/7	0.70
A1919	C1919	王二	H-010	null	1	2019/1/4	2020/3/31	1.09
A1364	C1364	张三	H-011	意向	1	2019/1/1	2020/3/26	0.70
…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…

处理缺失值-2

根据数据类型以及生成信息重要程度的不同，使用不同方法处理缺失值

文本型字段

影响不大：不进行处理，或者替换

影响大：替换，或者删除

数值型字段

综合考虑该数值型字段所代表的度量意义以及针对该数值型字段进行汇总计算的方式来最终决定对缺失值的具体处理方法

在张三负责的商机信息中的销售阶段为“null”值，需要找销售人员张三进行核实，并用正确的有效信息替换“null”值

商机编号	客户ID	销售人员	产品编号	销售阶段	赢单率	建立商机日	预计签约日	预计成交金额（$M）
A1550	C1550	赵大	F-009	投入	0.15	2019/1/8	2020/1/7	0.70
A1919	C1919	王二	H-010	null	1	2019/1/4	2020/3/31	1.09
A1364	C1364	张三	H-011	意向	1	2019/1/1	2020/3/26	0.70
…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…

表结构数据特征-3

一个表中有且只有一个主键

物理意义

- 单字段主键：由一个字段构成的主键
- 多字段联合主键：由多个字段构成的主键
- 非空不重复
- 定位记录行、字段名+主键值定位具体数值
- 多以“xxID”、“xxNo”、“xx编号”等名称命名

业务意义

表的业务记录单位。在一个数据表中的所有非主键字段都要围绕主键展开

确定主键的方法

如果直接对数据库中的数据表进行操作，可以通过SQL语句确认数据表的主键字段

如果间接在其他数据分析平台使用表结构数据（数据源是数据库中的数据表，使用时将数据源数据导入到其他平台使用），可以找原数据库中数据表的设计者咨询主键信息，或是直接查看数据表设计者留下的设计资料来对主键字段进行确认

通过对数据表的业务意义进行分析，推测主键字段后在用物理手段确认推测的主键字段中的记录值是否能够满足“非空”、“不重复”的要求来对主键字段进行确认

产品编号字段为主键，该表以产品为单位生成记录，一行记录代表一个不同产品，所有其他非主键字段都要用来描述及扩展产品信息

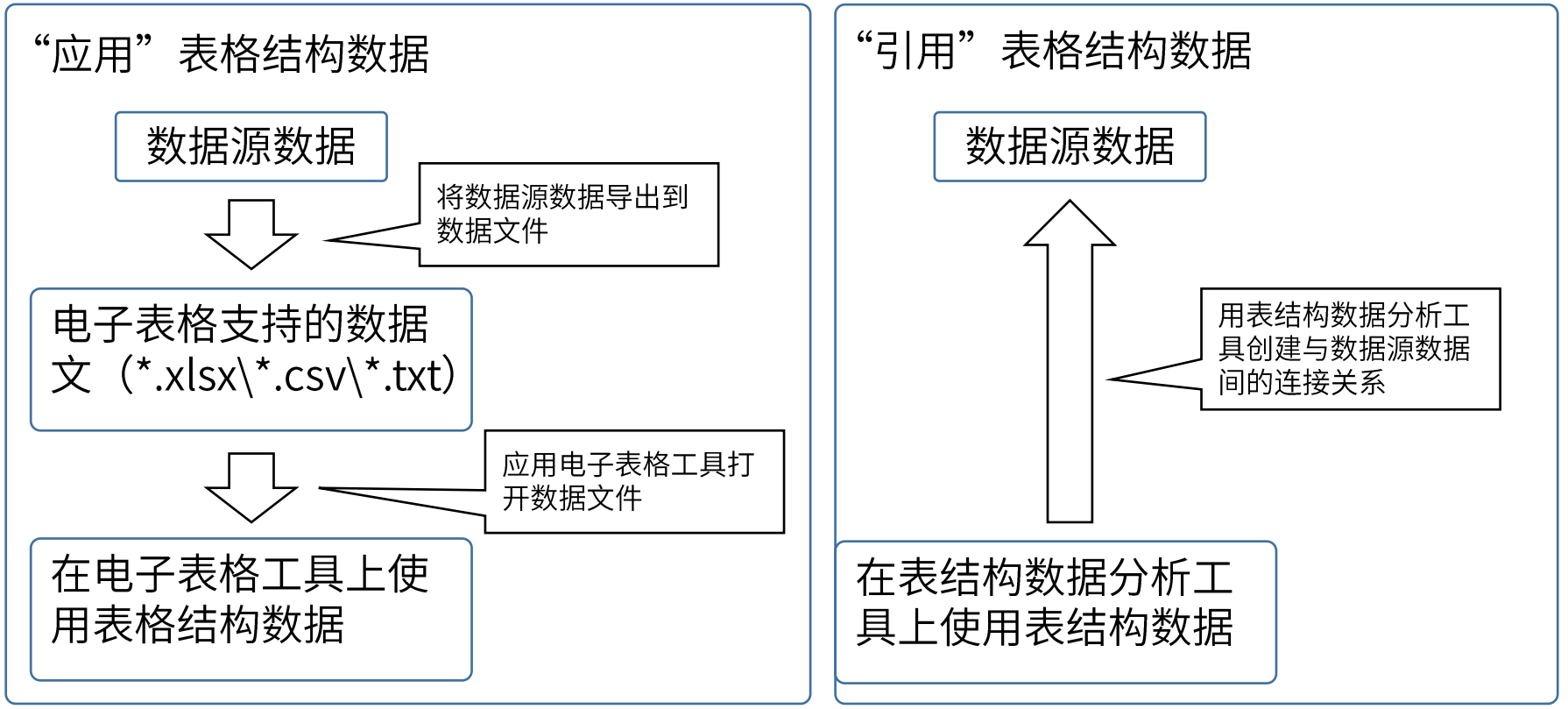
产品编号	产品名称	品牌编号	…
1101001	1101001-啤酒xx	00001	…
1101002	1101002-啤酒xx	00001	…
1201001	1201001-38xx	00002	…
…	…	…	…

表结构数据的获取方法

05

“应用”与“引用”

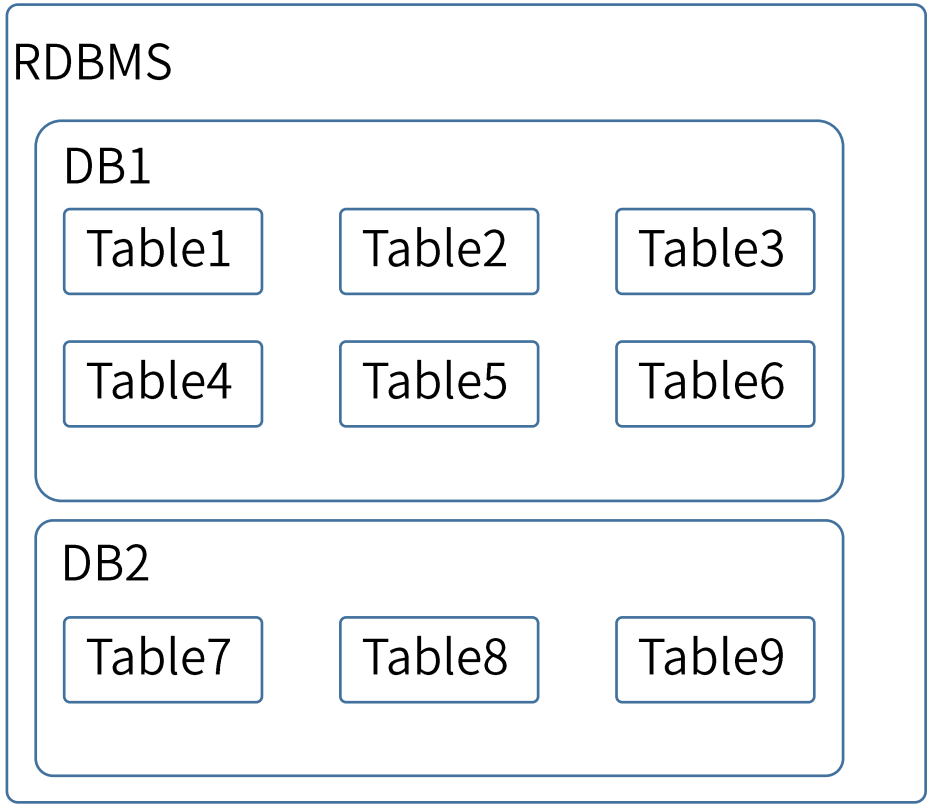
“应用”表格结构数据、“引用”表结构数据



关系型数据库管理系统

关系型数据库管理系统的主要任务是企业业务数据的存储、检索、访问与共享

- 多层级结构
- OLTP
- 可量化、结构化数据
- 提供大部分数据源
- 不善于分析



商业智能系统 - BI

用于为企业决策者快速提供完整、准确、深入的数据分析结果，帮助企业决策者实现商业洞察

强于分析

多功能模块构成

两种主要类型

多维数据集

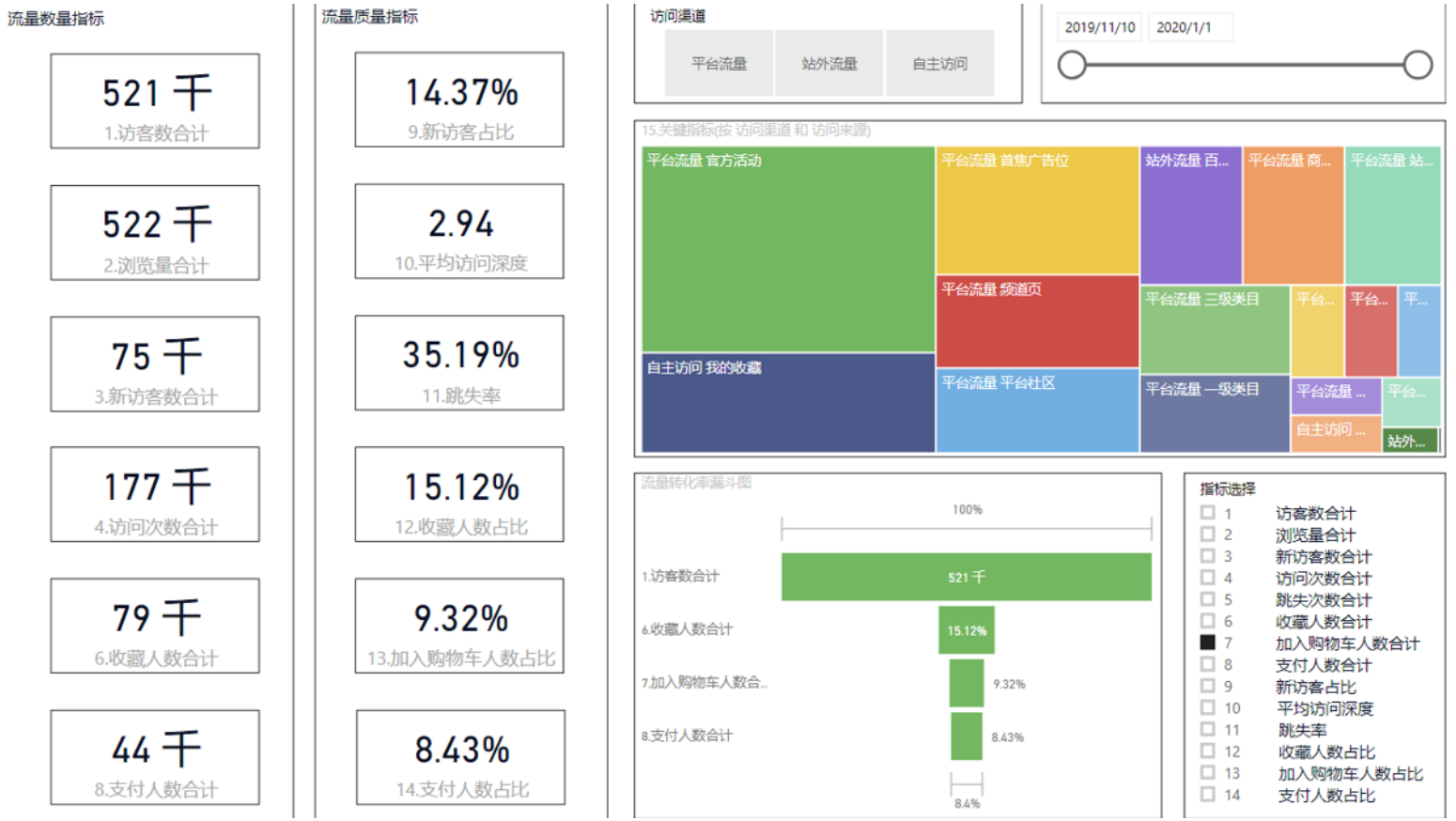
所见即所得

	企业级商业智能系统	敏捷型商业智能系统
应用范围	大型企业各相关部门	中小企业或某个大企业的业务部门
价格	高	低
数据处理加工能力	强	一般
数据分析能力	强	一般
速度	快	一般
IT技术门槛	高	低
实施及部署周期	长	短
拓展更新难度	大	小

商业智能系统 - BI

用于为企业决策者快速提供完整、准确、深入的数据分析结果，帮助企业决策者实现商业洞察

- 强于分析
- 多功能模块构成
- 两种主要类型
- 多维数据集
- 所见即所得



ETL功能

将数据从数据源端经过抽取（Extract）、清洗转换（Transform）之后加载（Load）到数据仓库

E – 抽取

创建与不同数据源间的连接关系，对这些数据源中的数据进行“引用”

T – 清洗转换

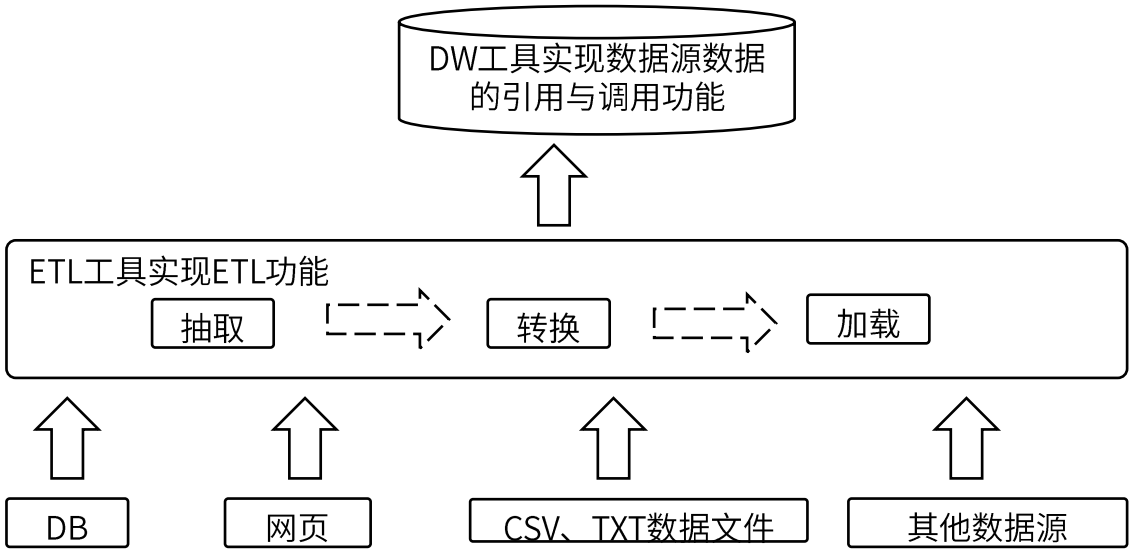
- 清洗的主要任务是筛选过滤不完整、错误及重复的数据记录
- 对“粒度”不一致的数据进行转换
- 对业务规则不一致的数据进行转换

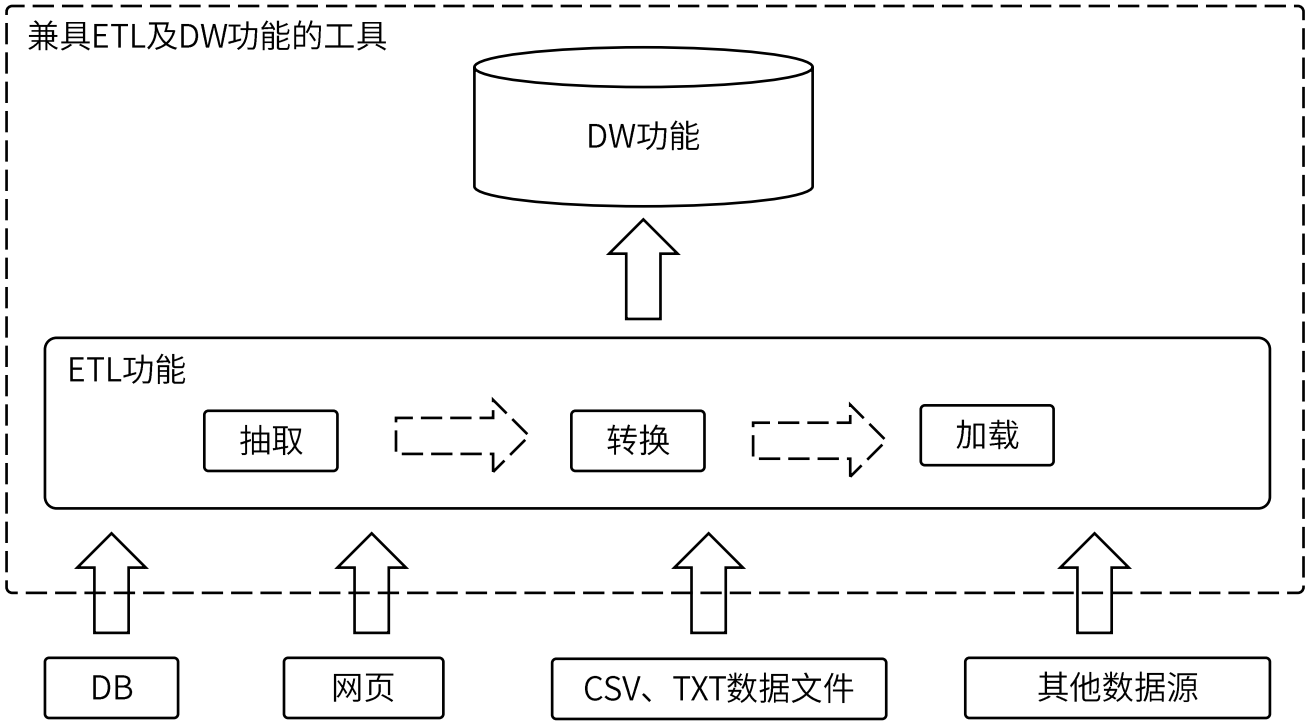
L – 加载

将抽取出来的数据经过清洗与转换后加载到数据仓库中进行存储与使用

数据仓库 - DW

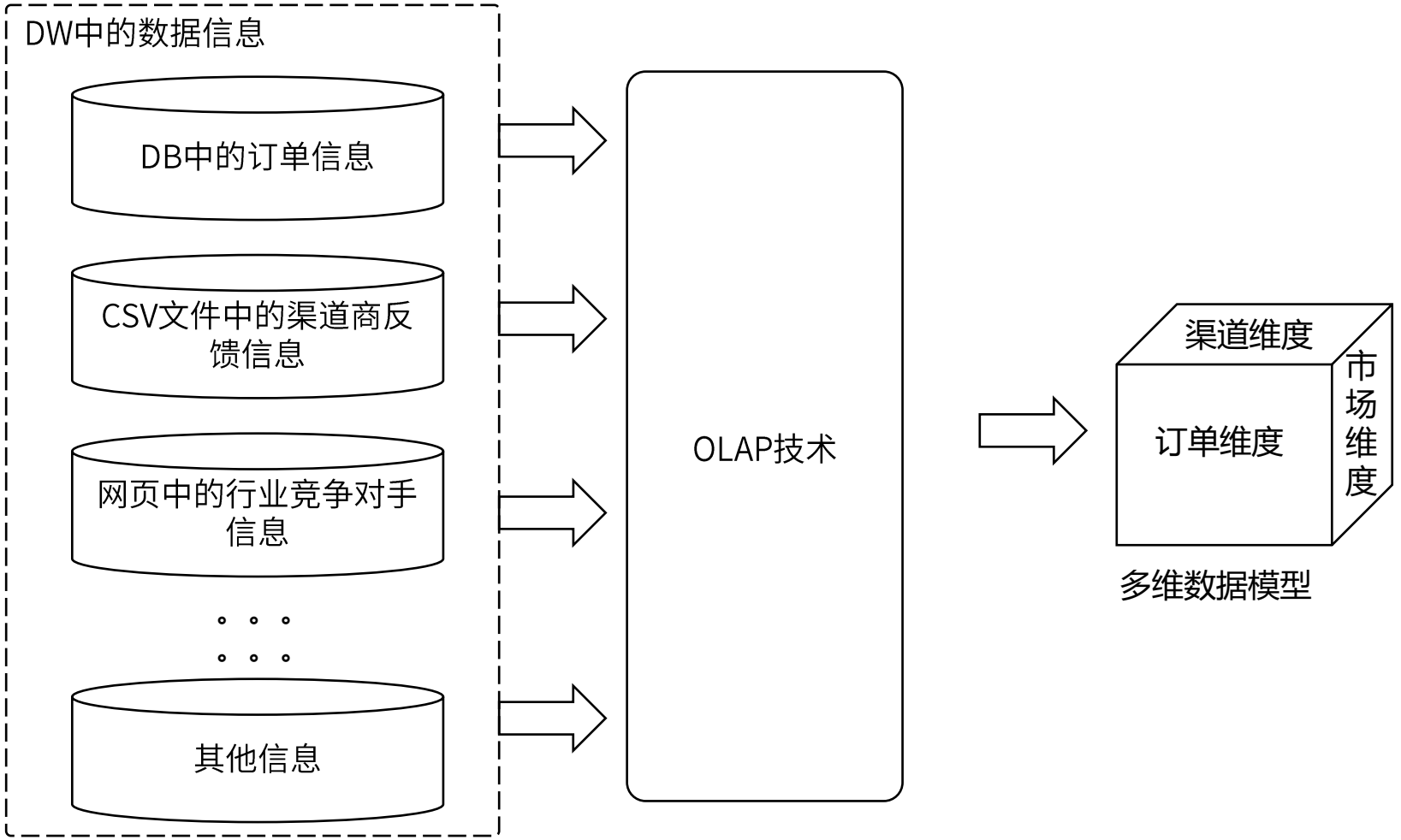
用来存储分析所需要的不同数据源上的所有相关数据信息





OLAP

连接信息孤岛、创建多维数据模型

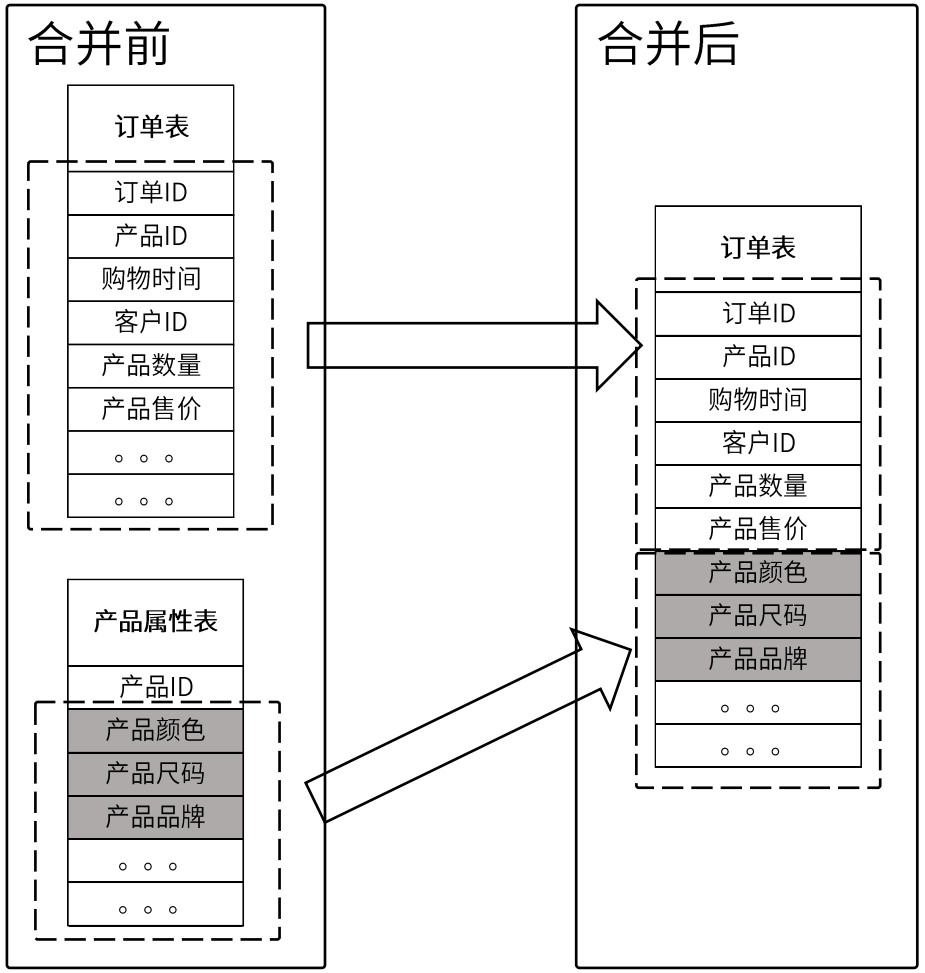


表结构数据的使用方法

06

表结构数据的横向合并

将不同表中的字段信息合并到同一个表中使用



表结构数据的横向合并

将不同表中的字段信息合并到同一个表中使用

通过公共字段匹配	拥有相同记录值的字段
左表与右表	连接命令左侧的表为左表，右侧的表为右表
连接方向	决定表的主附关系，主要使用“左连接”\“右连接”\“内连接”
对应关系	决定连接结果行数是对应项乘积的结果
E-R图	多表连接的鸟瞰图

表结构数据的横向合并

将不同表中的字段信息合并到同一个表中使用

通过公共字段匹配

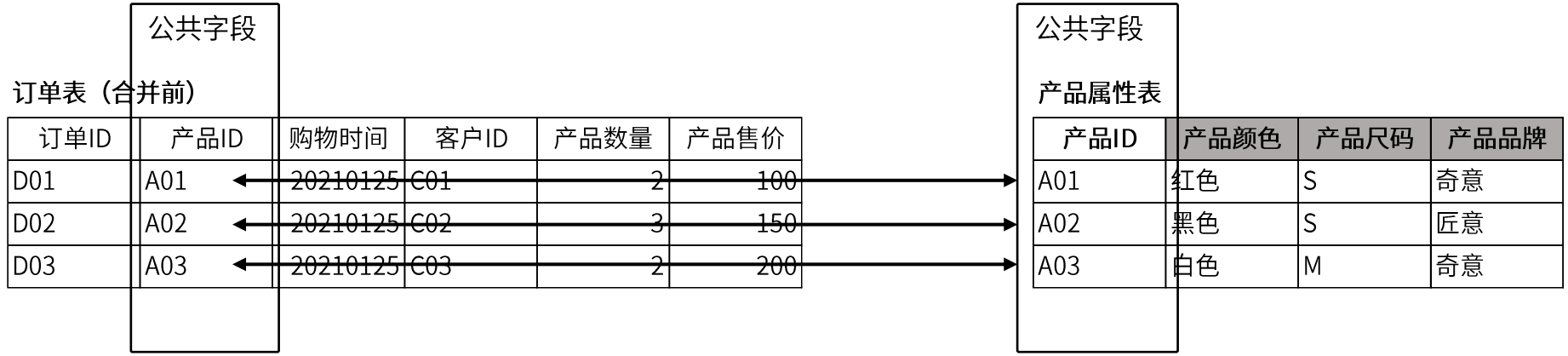
拥有相同记录值的字段

左表与右表

连接方向

对应关系

E-R图



订单表（合并前）

订单ID	产品ID	购物时间	客户ID	产品数量	产品售价
D01	A01	20210125	C01	2	100
D02	A02	20210125	C02	3	150
D03	A03	20210125	C03	2	200

产品属性表

产品ID	产品颜色	产品尺码	产品品牌
A01	红色	S	奇意
A02	黑色	S	匠意
A03	白色	M	奇意

订单表（合并后）

订单ID	产品ID	购物时间	客户ID	产品数量	产品售价	产品颜色	产品尺码	产品品牌
D01	A01	20210125	C01	2	100	红色	S	奇意
D02	A02	20210125	C02	3	150	黑色	S	匠意
D03	A03	20210125	C03	2	200	白色	M	奇意

表结构数据的横向合并

将不同表中的字段信息合并到同一个表中使用

左表与右表

连接命令左侧的表为左表，右侧的表为右表

连接方向

决定表的主附关系，主要使用“左连接”\“右连接”\“内连接”

销售人员表（左表）

销售人员ID	销售人员姓名
S01	赵大
S02	王二
S03	张三

订单表（右表）

订单ID	销售人员ID	订单金额
A01	S01	10000
A02	S02	15000

内连接合并结果

销售人员ID	销售人员姓名	订单ID	销售人员ID	订单金额
S01	赵大	A01	S01	10000
S02	王二	A02	S02	15000

左连接合并结果

销售人员ID	销售人员姓名	订单ID	销售人员ID	订单金额
S01	赵大	A01	S01	10000
S02	王二	A02	S02	15000
S03	张三	null	null	null

右链接合并结果

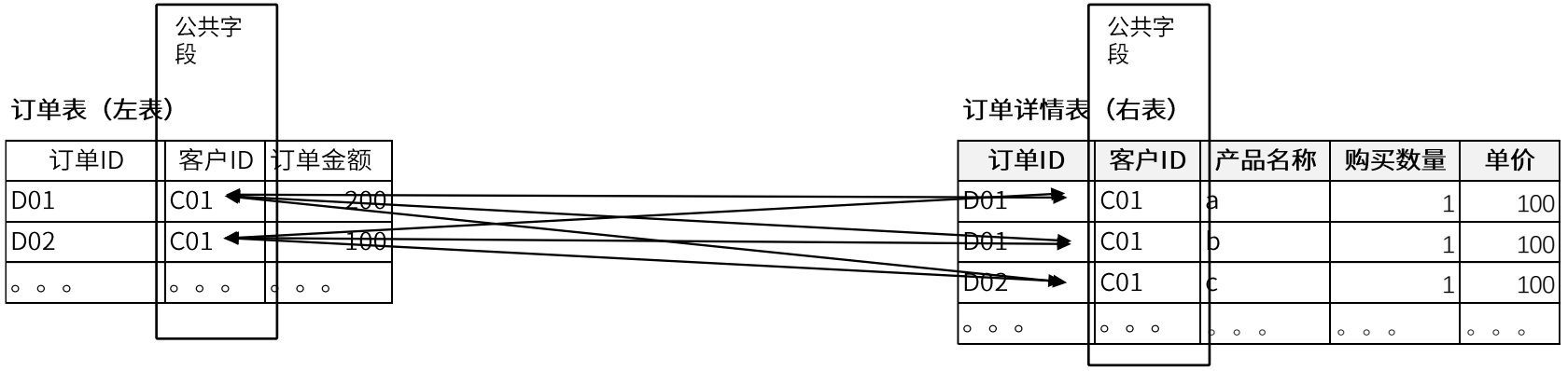
销售人员ID	销售人员姓名	订单ID	销售人员ID	订单金额
S01	赵大	A01	S01	10000
S02	王二	A02	S02	15000

表结构数据的横向合并

将不同表中的字段信息合并到同一个表中使用

对应关系

决定连接结果行数是对应项乘积的结果



左连接

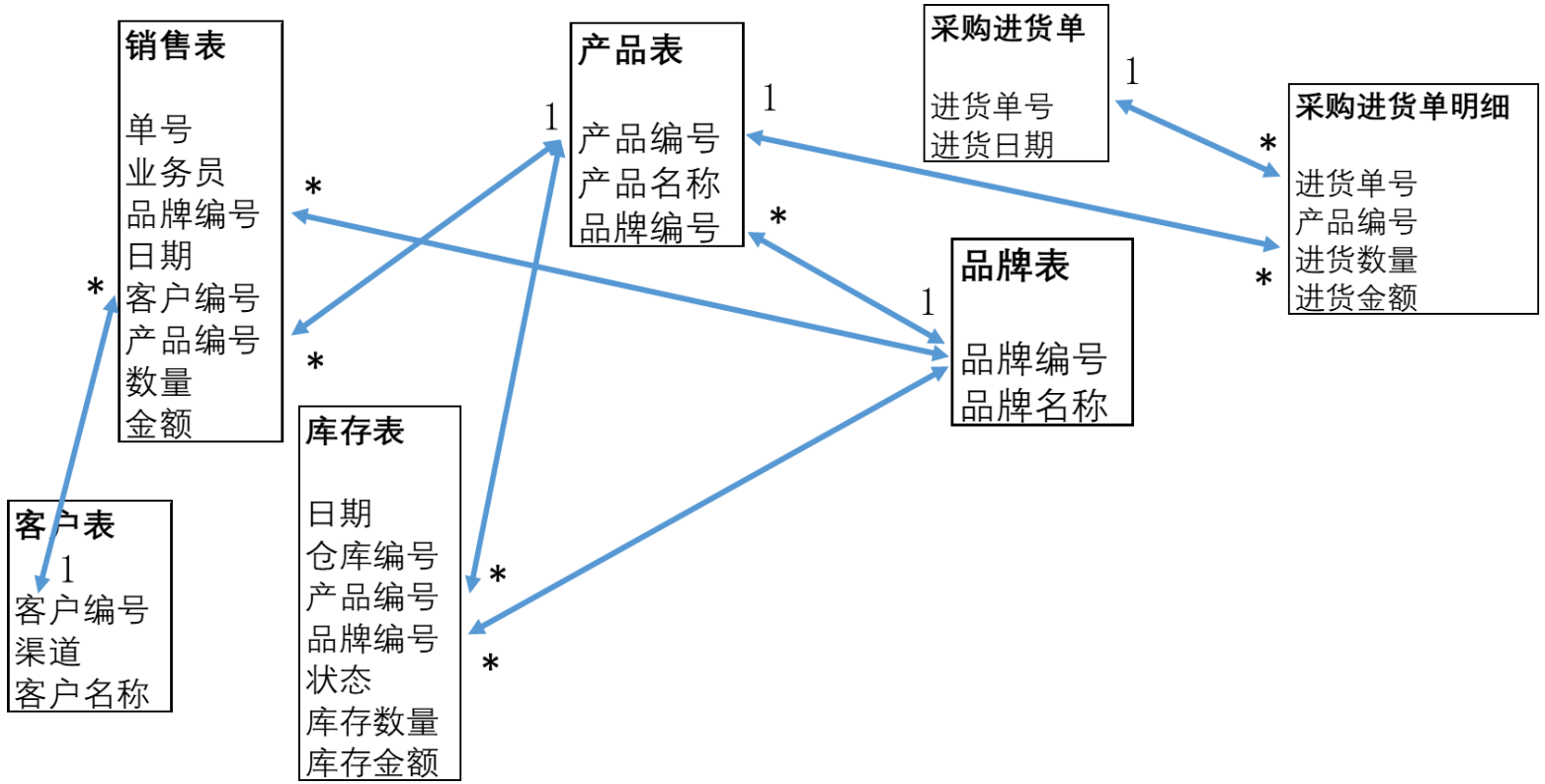
订单ID	客户ID	订单金额	订单ID	客户ID	产品名称	购买数量	单价
D01	C01	200	D01	C01	a	1	100
D01	C01	200	D01	C01	b	1	100
D01	C01	200	D02	C01	c	1	100
D02	C01	100	D01	C01	a	1	100
D02	C01	100	D01	C01	b	1	100
D02	C01	100	D02	C01	c	1	100
...	...	...	...	...	...	...	...

表结构数据的横向合并

将不同表中的字段信息合并到同一个表中使用

E-R图

多表连接的鸟瞰图



表结构数据的纵向合并

多表中记录信息合并到同一个表中进行使用的合并方式称为纵向合并

字段个数相同

相同位置字段的数据类型相同

去重合并与全合并

第一季度销售表

订单ID	销售人员ID	订单金额
D01	C01	200
D02	C02	100
...	...	...

第二季度销售表

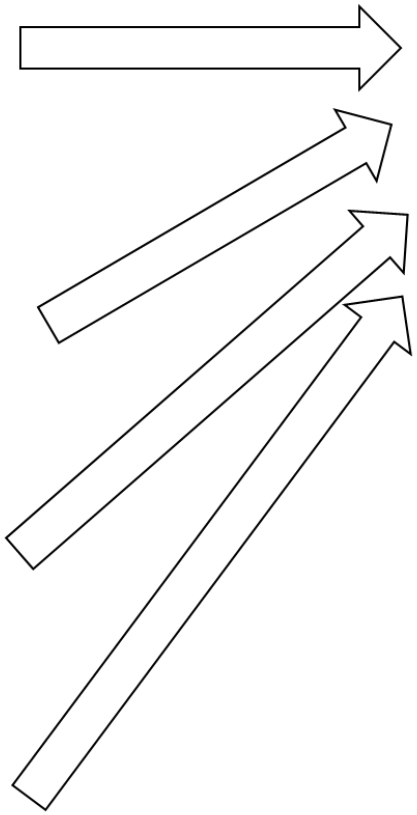
订单ID	销售人员ID	订单金额
D03	C01	300
D04	C02	400
...	...	...

第三季度销售表

订单ID	销售人员ID	订单金额
D05	C01	100
D06	C02	300
...	...	...

第四季度销售表

订单ID	销售人员ID	订单金额
D07	C01	600
D08	C02	900
...	...	...



全年销售表

订单ID	销售人员ID	订单金额
D01	C01	200
D02	C02	100
D03	C01	300
D04	C02	400
D05	C01	100
D06	C02	300
D07	C01	600
D08	C02	900
...	...	...

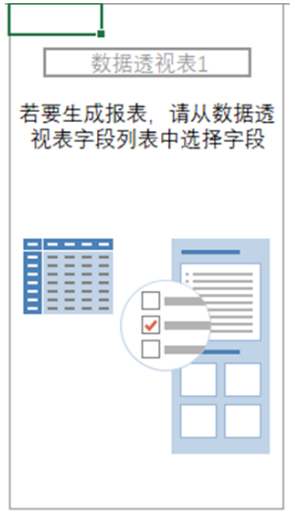
表结构数据的汇总

数据透视 -- 对零散数据进行汇总分析

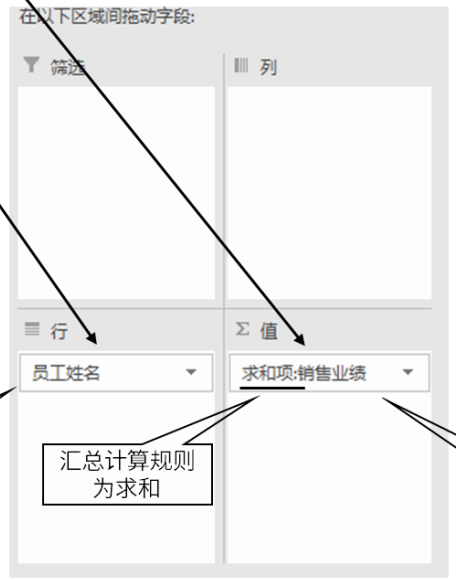
订单ID	员工姓名	销售业绩	销售提成
a001	邹*倩	75894	3794.7
a002	马*雄	92522	4626.1
a003	王*可	15457	772.85
a004	邹*倩	20698	1034.9
a005	刘*珊	17616	880.8
a006	张*纯	22371	1118.55
a007	谢*佳	52560	2628
a008	何*敏	9213	460.65
a009	蕾*拉	51142	2557.1

数据源：销售业绩表

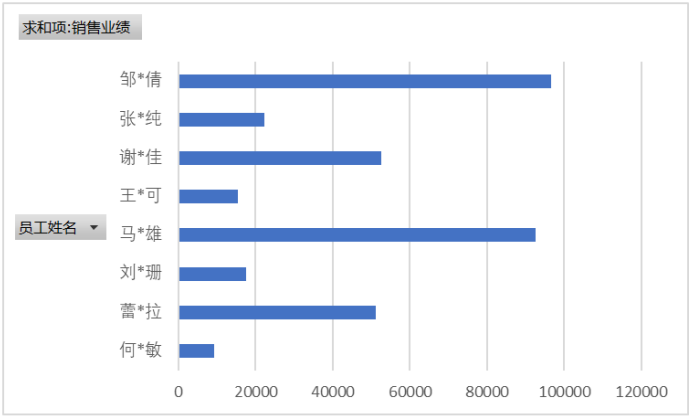
维度字段：员工姓名
度量字段：销售业绩
汇总计算规则：求和



数据透视表模板



设置维度、度量及汇总计算规则



数据透视图

行标签	求和项:销售业绩
何*敏	9213
蕾*拉	51142
刘*珊	17616
马*雄	92522
王*可	15457
谢*佳	52560
张*纯	22371
邹*倩	96592
总计	357473

数据透视表

表结构数据的汇总

数据透视 -- 对零散数据进行汇总分析

维度	业务观测角度
度量	业务行为结果
汇总计算规则	衡量业务行为结果好坏程度的测量仪

表结构数据的汇总

维度筛选度量、度量被维度筛选

订单ID	员工姓名	销售业绩	销售提成
a001	邹*倩	75894	3794.7
a002	马*雄	92522	4626.1
a003	王*可	15457	772.85
a004	邹*倩	20698	1034.9
a005	刘*珊	17616	880.8
a006	张*纯	22371	1118.55
a007	谢*佳	52560	2628
a008	何*敏	9213	460.65
a009	蕾*拉	51142	2557.1

行标签	求和项:销售业绩
何*敏	9213
蕾*拉	51142
刘*珊	17616
马*雄	92522
王*可	15457
谢*佳	52560
张*纯	22371
邹*倩	96592
总计	**357473**

表结构数据的汇总

汇总计算规则

合计规则：将相同维度值下对应的多个度量值相加在一起、一般用SUM函数代表合计规则

计数规则：对相同维度值下的度量个数进行计数、COUNT非空计数、DISTINCTCOUNT去重计数

平均规则：用合计规则的结果除以计数规则的结果（平均=合计/计数）、一般用AVERAGE函数表示

最大值规则：求相同维度之下最大的度量值、一般用MAX函数表示

最小值规则：求相同维度之下最小的度量值、一般用MIN函数表示

表结构数据的汇总

汇总计算规则

数据源：销售业绩表

订单ID	员工姓名	销售业绩	销售提成
a001	邹*倩	75894	3794.7
a002	马*雄	92522	4626.1
a003	王*可	15457	772.85
a004	邹*倩	20698	1034.9
a005	刘*珊	17616	880.8
a006	张*纯	22371	1118.55
a007	谢*佳	52560	2628
a008	何*敏	9213	460.65
a009	蕾*拉	51142	2557.1

	合计规则	计数规则	平均规则	最大值规则	最小值规则
行标签	求和项:销售业绩	计数项:销售业绩2	平均值项:销售业绩2	最大值项:销售业绩2	最小值项:销售业绩2
何*敏	9213	1	9213	9213	9213
蕾*拉	51142	1	51142	51142	51142
刘*珊	17616	1	17616	17616	17616
马*雄	92522	1	92522	92522	92522
王*可	15457	1	15457	15457	15457
谢*佳	52560	1	52560	52560	52560
张*纯	22371	1	22371	22371	22371
邹*倩	96592	2	48296	75894	20698
总计	357473	9	39719.22222	92522	9213

数据分析的业务意义

数据分析是连接零散数据与人类认知间的桥梁

零散数据

数据分析

数据透视分析

数据挖掘分析

人类认知