



西安交通大学管理学院  
THE SCHOOL OF MANAGEMENT  
XI'AN JIAOTONG UNIVERSITY

# 6 从数据库到数据分析 [了解]

刘跃文 博士，副教授

西安交通大学管理学院  
信息管理与电子商务系

[liuyuewen@mail.xjtu.edu.cn](mailto:liuyuewen@mail.xjtu.edu.cn)

V1, 2018-12-20

# 提纲

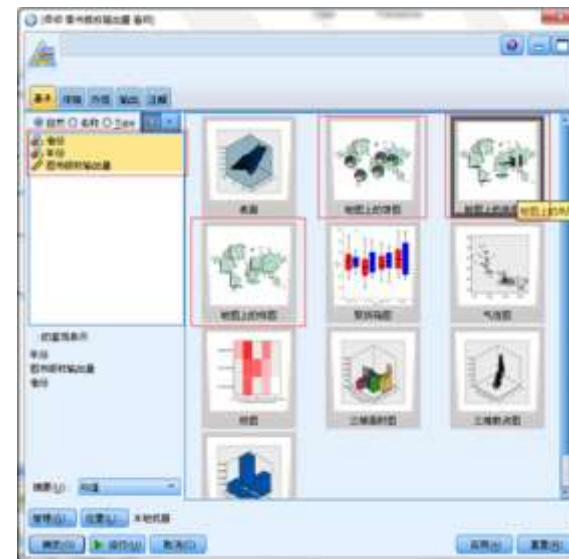
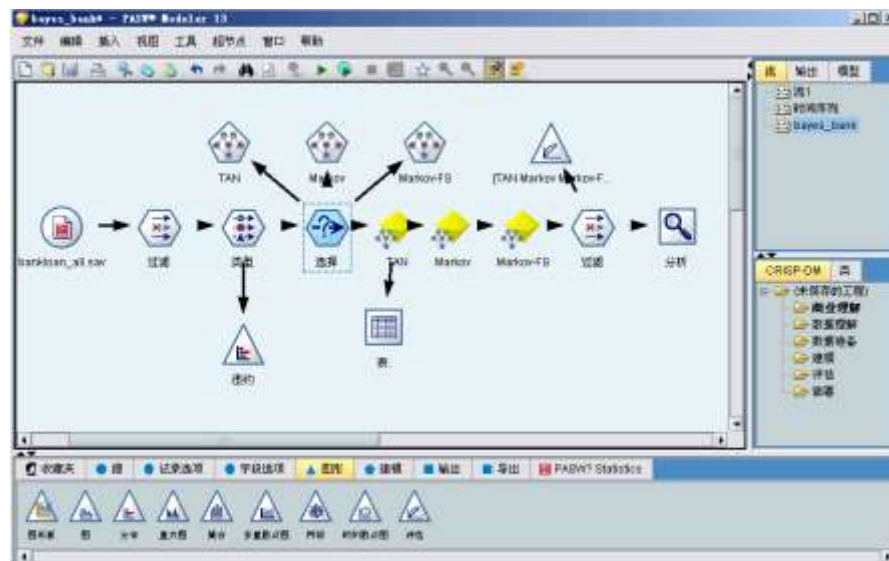
- SQL语句的特点
- SQL语句是大部分数据分析软件的基础
- E-R模型为数据分析提供理论指导
- 数据分析还要学什么

# 1 SQL语句的特点

- 面向过程的语言 C
- 详细规定每一个流程及动作
- 面向对象的语言 C++, Java
- 对对象的实例进行操作
- 面向数据的语言 SQL
- 对数据表进行操作
- 思维模式的核心：数据表
- 大幅度减少循环操作

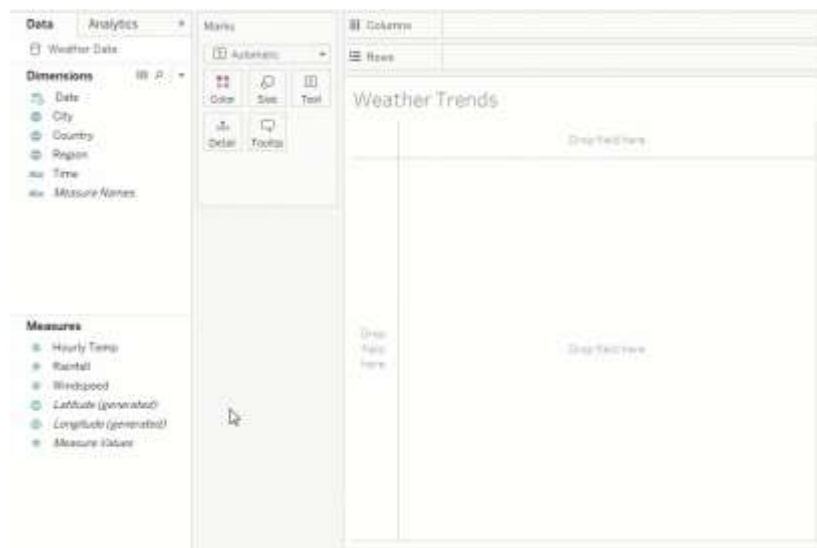
## 2 SQL语句是大部分数据分析软件的基础

- IBM SPSS Modeler
  - Clementine → SPSS → IBM
  - IBM SPSS Modeler



# 软件说明

- Tableau <https://www.tableau.com/>
  - 在校学生可以申请desktop版的许可证，免费试用1年



# 以Modeler为例：



Excel



选择



排序



追加



汇总



样本



过滤器



字段重排



合并



填充



导出



类型

## 常用的节点



Excel

- | 品名   | 上月末库存    |          | 本月购进   |          | 本月原材料加工费 |      |
|------|----------|----------|--------|----------|----------|------|
|      | 数量       | 金额       | 数量     | 金额       | 数量       | 单价   |
| 材料01 | 100.00   | 120.00   | 41.00  | 748.00   | 181.00   | 6.16 |
| 材料02 | 101.00   | 121.00   | 28.00  | 437.00   | 129.00   | 6.33 |
| 材料03 | 102.00   | 122.00   | 32.00  | 426.00   | 134.00   | 6.09 |
| 材料04 | 103.00   | 123.00   | 35.00  | 492.00   | 138.00   | 6.46 |
| 材料05 | 104.00   | 124.00   | -      | -        | 104.00   | 1.19 |
| 材料06 | 105.00   | 125.00   | -      | -        | 105.00   | 1.19 |
| 材料07 | 106.00   | 126.00   | -      | -        | 106.00   | 1.19 |
| 材料08 | 107.00   | 127.00   | -      | -        | 107.00   | 1.19 |
| 材料09 | 108.00   | 128.00   | 17.00  | 306.00   | 125.00   | 3.47 |
| 材料10 | 109.00   | 129.00   | -      | -        | 109.00   | 1.18 |
| 材料11 | 110.00   | 130.00   | -      | -        | 110.00   | 1.18 |
| 材料12 | 111.00   | 131.00   | -      | -        | 111.00   | 1.18 |
| 材料13 | 112.00   | 132.00   | -      | -        | 112.00   | 1.18 |
| 材料14 | 113.00   | 133.00   | -      | -        | 113.00   | 1.18 |
| 材料15 | 114.00   | 134.00   | -      | -        | 114.00   | 1.18 |
| 材料16 | 115.00   | 135.00   | -      | -        | 115.00   | 1.17 |
| 合计   | 1,720.00 | 2,040.00 | 153.00 | 2,409.00 | 1,873.00 | 2.38 |

[illegible]



# 行和列

列，字段，变量



| ID         | FriendID   | ActionTime      | ActionTime | Flag |
|------------|------------|-----------------|------------|------|
| 2010006067 | 2020699652 | 2012 年 7 月 14 日 | 13:32:55   | 1    |
| 2010006067 | 2020699652 | 2012 年 8 月 8 日  | 23:41:19   | 1    |
| 2010006067 | 2020699652 | 2012 年 8 月 16 日 | 16:12:10   | 1    |
| 2010006067 | 2020699652 | 2012 年 8 月 16 日 | 16:42:23   | 1    |
| 2010006067 | 2020699652 | 2012 年 8 月 16 日 | 16:42:54   | 1    |
| 2010006067 | 2020699652 | 2012 年 8 月 16 日 | 16:43:30   | 1    |
| 2010006067 | 2020699652 | 2012 年 8 月 16 日 | 17:03:16   | 1    |
| 2010006067 | 2020699652 | 2012 年 8 月 17 日 | 12:33:59   | 1    |
| 2010006067 | 2020699652 | 2012 年 8 月 19 日 | 13:09:15   | 1    |
| 2010006067 | 2020699652 | 2012 年 8 月 19 日 | 14:51:21   | 1    |
| 2010006067 | 2020699652 | 2012 年 8 月 19 日 | 15:21:46   | 1    |
| 2010006067 | 2020699652 | 2012 年 8 月 20 日 | 4:15:47    | 1    |
| 2010006067 | 2020699652 | 2012 年 8 月 20 日 | 13:19:16   | 1    |
| 2010006067 | 2020699652 | 2012 年 8 月 20 日 | 15:05:35   | 1    |
| 2010006067 | 2020699652 | 2012 年 8 月 21 日 | 2:52:34    | 2    |

行，记录，  
一条数据



# 行的处理

- 筛选：条件选择、随机选择（样本） where, top



选择



样本

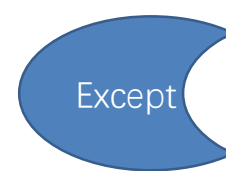
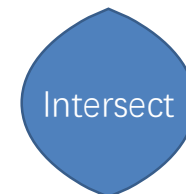
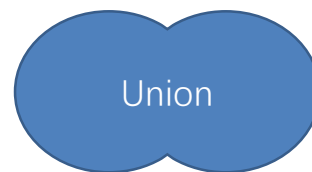
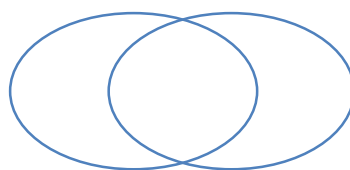
- 排序：单一条件排序、综合条件排序 order by, asc, desc



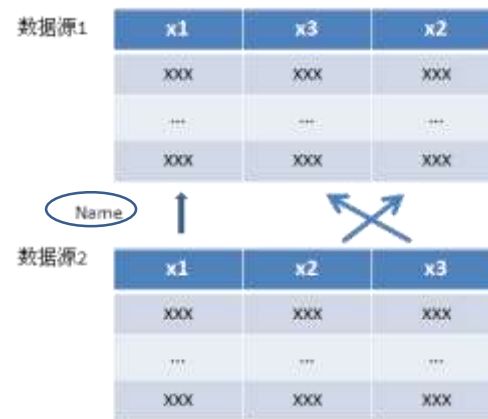
排序

# 集合查询

- 追加



- 数据纵向比较（集合比较）
- 并集 Union
- 交集 Intersect
- 补集 Except



# 去重操作 distinct

- 优点：可对字符串型数据进行汇总操作
- 几种用法：
  - (1) 每组取第一条
  - (2) 每组合并



| ID | 备注  | 频次 |
|----|-----|----|
| 1  | 王大锤 | 17 |
| 2  | 张晓丽 | 8  |

| ID | 备注           |
|----|--------------|
| 1  | 王大锤；大锤；锤哥；锤子 |
| 2  | 张晓丽；丽姐；张总    |



| ID | 备注  | 频次 |
|----|-----|----|
| 1  | 王大锤 | 17 |
| 1  | 大锤  | 8  |
| 1  | 锤哥  | 5  |
| 1  | 锤子  | 1  |
| 2  | 张晓丽 | 18 |
| 2  | 丽姐  | 6  |
| 2  | 张总  | 3  |



# 列的处理 Select

- 重命名、不显示某些列



过滤器

- 调整顺序



字段重排

| ID         | FriendID   | ActionTime      | ActionTime | Flag |
|------------|------------|-----------------|------------|------|
| 2010006067 | 2020699652 | 2012 年 7 月 14 日 | 13:32:55   | 1    |
| 2010006067 | 2020699652 | 2012 年 8 月 8 日  | 23:41:19   | 1    |
| 2010006067 | 2020699652 | 2012 年 8 月 16 日 | 16:12:10   | 1    |
| 2010006067 | 2020699652 | 2012 年 8 月 16 日 | 16:42:23   | 1    |
| 2010006067 | 2020699652 | 2012 年 8 月 16 日 | 16:42:54   | 1    |
| 2010006067 | 2020699652 | 2012 年 8 月 16 日 | 16:43:30   | 1    |
| 2010006067 | 2020699652 | 2012 年 8 月 16 日 | 17:03:16   | 1    |
| 2010006067 | 2020699652 | 2012 年 8 月 17 日 | 12:33:59   | 1    |
| 2010006067 | 2020699652 | 2012 年 8 月 19 日 | 13:09:15   | 1    |
| 2010006067 | 2020699652 | 2012 年 8 月 19 日 | 14:51:21   | 1    |
| 2010006067 | 2020699652 | 2012 年 8 月 19 日 | 15:21:46   | 1    |
| 2010006067 | 2020699652 | 2012 年 8 月 20 日 | 4:15:47    | 1    |
| 2010006067 | 2020699652 | 2012 年 8 月 20 日 | 13:19:16   | 1    |
| 2010006067 | 2020699652 | 2012 年 8 月 20 日 | 15:05:35   | 1    |
| 2010006067 | 2020699652 | 2012 年 8 月 21 日 | 2:52:34    | 2    |

# 字段计算

- 创建新的字段 ... as



- 修改原有字段 Update



# 连接查询

- 数据的连接条件：
  - 连接条件也是一种“条件”
  - 一般用法：表A.ID=表B.ID



- 数据的横向合并的连接方式：
  - 内连接 (inner join)
  - 全外连接 (full outer join)
  - 局部外连接 (partial outer join)
  - 反连接 (anti-join)

| 数据源1 |     |     |       | 数据源2 |     |     |
|------|-----|-----|-------|------|-----|-----|
| 编号   | x1  | x2  | Order | 编号   | x3  | x4  |
| 1    | xxx | xxx | ←     | 1    | xxx | xxx |
| 2    | xxx | xxx | ←     | 2    | xxx | xxx |
| 3    | xxx | xxx | ←     | 4    | xxx | xxx |
| 4    | xxx | xxx | ←     | 3    | xxx | xxx |
| ...  | ... | ... |       | ...  | ... | ... |

| 数据源1 |     |     |      | 数据源2 |     |     |
|------|-----|-----|------|------|-----|-----|
| 编号   | x1  | x2  | Keys | 编号   | x3  | x4  |
| 1    | xxx | xxx | ←    | 1    | xxx | xxx |
| 2    | xxx | xxx | ←    | 2    | xxx | xxx |
| 3    | xxx | xxx | ↗    | 4    | xxx | xxx |
| 4    | xxx | xxx | ↘    | 3    | xxx | xxx |
| ...  | ... | ... |      | ...  | ... | ... |

# 连接的种类

表A

| A1 | A2 | A3 |
|----|----|----|
| 1  | a  | M  |
| 2  | b  | N  |
| 3  | c  | O  |
| 4  | d  | P  |
| 5  | e  | Q  |

表B

| A1 | A4 |
|----|----|
| 3  | X  |
| 5  | Y  |
| 7  | Z  |

表A × B

| A1 | A2 | A3 | A1 | A4 |
|----|----|----|----|----|
| 1  | a  | M  | 3  | X  |
| 1  | a  | M  | 5  | Y  |
| 1  | a  | M  | 7  | Z  |
| 2  | b  | N  | 3  | X  |
| 2  | b  | N  | 5  | Y  |
| 2  | b  | N  | 7  | Z  |
| 3  | c  | O  | 3  | X  |
| 3  | c  | O  | 5  | Y  |
| 3  | c  | O  | 7  | Z  |
| 4  | d  | P  | 3  | X  |
| 4  | d  | P  | 5  | Y  |
| 4  | d  | P  | 7  | Z  |
| 5  | e  | Q  | 3  | X  |
| 5  | e  | Q  | 5  | Y  |
| 5  | e  | Q  | 7  | Z  |

关键字(A1)连接

| A1 | A2 | A3 | A1 | A4 |
|----|----|----|----|----|
| 3  | c  | O  | 3  | X  |
| 5  | e  | Q  | 5  | Y  |

内连接

| A1 | A2 | A3 | A4 |
|----|----|----|----|
| 3  | c  | O  | X  |
| 5  | e  | Q  | Y  |

不同的局部外连接

| A1 | A2 | A3 | A4 |
|----|----|----|----|
| 1  | a  | M  |    |
| 2  | b  | N  |    |
| 3  | c  | O  | X  |
| 4  | d  | P  |    |
| 5  | e  | Q  | Y  |

| A1 | A2 | A3 | A4 |
|----|----|----|----|
| 3  | c  | O  | X  |
| 5  | e  | Q  | Y  |
| 7  |    |    | Z  |

全外连接

| A1 | A2 | A3 | A4 |
|----|----|----|----|
| 1  | a  | M  |    |
| 2  | b  | N  |    |
| 3  | c  | O  | X  |
| 4  | d  | P  |    |
| 5  | e  | Q  | Y  |
| 7  |    |    | Z  |

# 分组统计 Group by

- 行：按什么字段汇总（关键字段），就剩下这些字段的不重复记录数
- 列：除了关键字段，只剩下统计字段



| 姓名 | 性别 | 身高  |
|----|----|-----|
| 王  | 男  | 170 |
| 张  | 女  | 160 |
| 李  | 女  | 162 |
| 赵  | 男  | 175 |
| 王  | 男  | 171 |



| 性别 | 个数 | 平均身高 |
|----|----|------|
| 男  | 3  | 172  |
| 女  | 2  | 161  |



### 3 E-R模型为数据分析提供理论指导

- 从表格中抽取对象、关系
- 逐个分析对象、关系



# 动态数据的基本构成

- 动态数据的基本构成：
  - 记叙文6要素：人物、时间、地点、事件起因、经过、结果
  - 动态数据5要素：{ID、时间、地点、事件、属性}
  - 例：学生卡号，时间，POS机号，消费类型，金额
- 5个要素中，4+个维度，1+个度量
  - 普遍存在的度量：数据条数
- 所有详细的账目都是动态数据



# 实体分析

- 分类型变量
  - 人、地、事、物、时间、组织……
- 数值型变量
  - 金额、次数

# 关系分析

|    | 时间 | 人 | 地 | 事 | 属性 |
|----|----|---|---|---|----|
| 时间 |    |   |   |   |    |
| 人  |    |   |   |   |    |
| 地  |    |   |   |   |    |
| 事  |    |   |   |   |    |
| 属性 |    |   |   |   |    |



- 以销售数据为例：
- 人-事：什么人擅长卖什么货物
- 事-地：什么地点消费什么货物的数量较大
- 事-时间：货物销量的时间周期（淡旺季分析）
- 人-属性：销售人员的销售额分析
- .....



- 关联规则：
- 时间-时间：在什么时候购买的，还会在什么时候购买
- 地-地：什么地方购买了，什么地方也会购买
- 事-事：购买什么的，还会购买什么
- .....



## 4 数据分析学什么

- **数据分析的流程**
  - 数据采集
  - 数据探索（理解所掌握的数据）
    - 描述性统计、趋势分析、分布分析、关联分析、数据质量评估等
  - 分析报告
    - 图表选择、规律识别、异常分析、报告撰写
- **数据处理（数据体系建设）**
  - 指标层面：抽取、构建、降维、重要性评估等
  - 数据层面：数据连接、数据选择、数据清洗等
- **建模与预测**
  - 数据挖掘与建模：模型选择、结果比较、结果解释、部署应用



# 数据挖掘的实现方式

|        | LV 1（无需编程）                | LV2（R/Python）                  | LV3（大数据）                      | LV3（专业）                                  |
|--------|---------------------------|--------------------------------|-------------------------------|------------------------------------------|
| 数据采集   | 八爪鱼                       | R/Python, ...                  |                               |                                          |
| 数据管理   | 文本+Excel                  | DB                             | Hadoop/Spark<br>分布式架构         | 取决于所选专业<br>DB                            |
| 数据探索   | Excel+Tableau<br>+Modeler | DB, Tableau, R/Python          | MR/Scala ...                  | 取决于所选专业<br>R/Python                      |
| 可视化分析  | Tableau+Excel+PPT         | Tableau+Excel+PPT,<br>R/Python | Tableau+Excel+PPT<br>R/Python | 取决于所选专业<br>R/Python                      |
| 数据处理   | Modeler                   | DB, R/Python                   | MR/Scala ...                  | 取决于所选专业<br>R/Python                      |
| 数据挖掘建模 | Modeler                   | R/Python                       | Scala, R/Python               | 取决于所选专业<br>R/Python                      |
| 理论要求   | 基础的统计知识<br>数据挖掘模型原理       | 计算机基础<br>统计模型                  | 分布式基础知识                       | 文本/图像/视频/声音/网络/基因领域知识及理论<br>高级模型：DL/增强学习 |

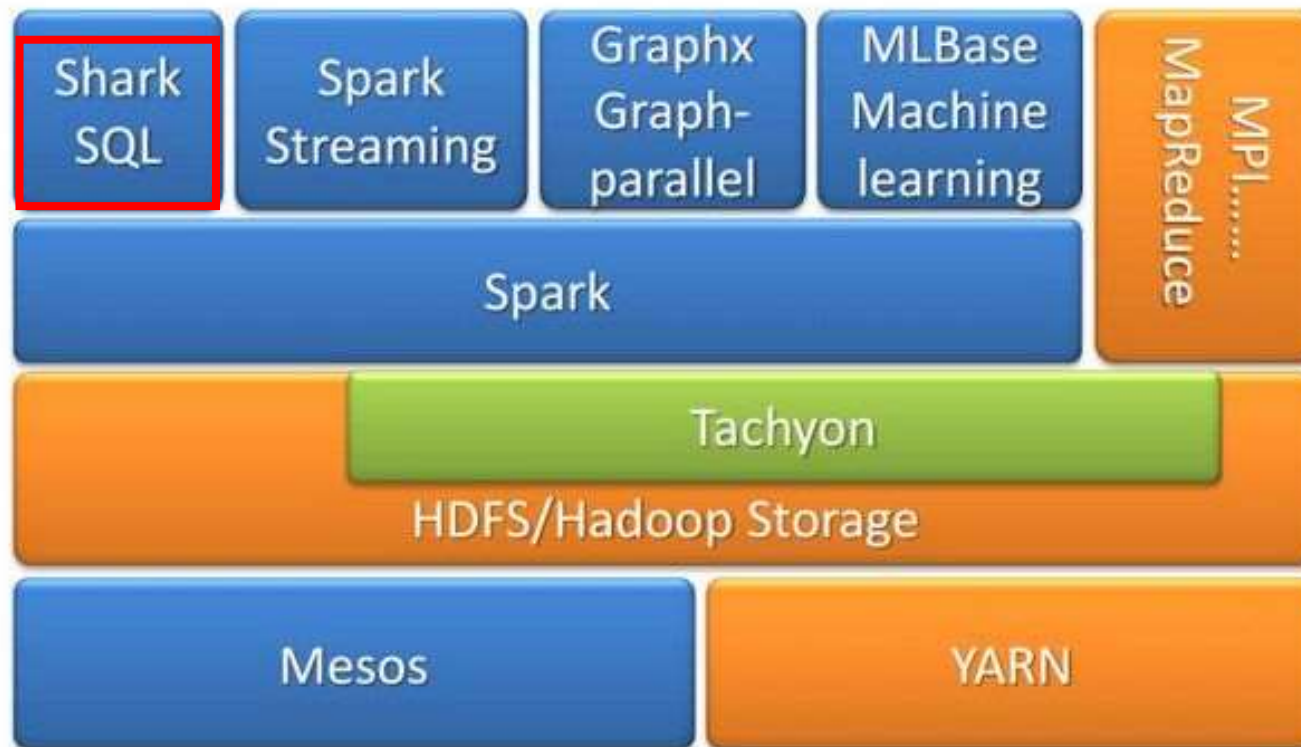
# Hadoop生态体系

- Hadoop对SQL的支持：



# Spark生态体系

- Spark对SQL的支持（Shark是Hive on Spark）



# 数据库与工作/生活的关系：

- 回到开篇问题：将来和数据库打交道的可能性有多大呢？





谢谢！

liuyuewen@xjtu.edu.cn