

数据分析人员需要掌握sql到什么程度?

数据分析师最常用的分析工具是 Excel、SQL、Python。那么问题就来了，这些分析工具具体掌握哪些内容呢？

一、SQL 需要掌握的知识点

很多初学者会被一些资料误导，以为 SQL 要掌握到数据库管理员的水平，而去学习 SQL 所有的知识点，最后不仅太难学不会，还浪费时间。

其实不同的职位对 SQL 的要求完全不一样，例如数据库管理员职位偏重于数据库的日常维护和管理,对 SQL 要求比较高；开发工程师偏重于用 SQL 完成开发工作，需要会在编程语句中使用 SQL；数据分析师偏重于用 SQL 查询数据。

所以，作为数据分析师，SQL 主要掌握查询数据就可以了。具体来说，SQL 的查询需要掌握以下知识点，涉及到的 SQL 语法不需要会，因为这部分重点是告诉你 SQL 需要掌握什么。

1) 简单查询

掌握基本的查询语句，也就是从数据库中获取数据。例如，从学生表中获取全部数据。对应的 SQL 知识点是查询语句

(select) 。

学生表
(student)

学号	姓名	出生日期	性别
0001	猴子	1989-01-01	男
0002	猴子	1990-12-21	女
0003	马云	1991-12-21	男
0004	王思聪	1990-05-20	男

从学生表中查询出2列

```
select 姓名,性别  
from student;
```

查询结果
→

姓名	性别
猴子	男
猴子	女
马云	男
王思聪	男



会根据指定的查询条件从数据库中获取符合条件的数据，例如，从学生表中获取姓名是「猴子」的学生，这里姓名=「猴子」就是查询条件。对应的 SQL 知识点是条件语句 (where) 。

查询条件

学号	姓名	出生日期	性别
0001	猴子	1989-01-01	男
0002	猴子	1990-12-21	女
0003	马云	1991-12-21	男
0004	王思聪	1990-05-20	男

选取'姓名'列里
值为'猴子'的行

```
select 姓名,学号  
from student  
where 姓名 = '猴子';
```

查询结果

学号	姓名
0001	猴子
0002	猴子



当查询条件多的时候，需要用到运算符来表示复杂的查询条件，例如，从学生表中获取姓名是「猴子」、性别是「男」的学生，这里姓名=「猴子」、性别是「男」就是多个查询条件。对应的 SQL 知识点是掌握以下运算符：

2) 汇总分析

实际工作中，并不是简单从数据库中查询出数据就可以了，还需要分析数据。SQL 常用的分析功能包括汇总、分组、排序。

掌握下面 5 个常用的汇总函数：

汇总函数

count	•求某列的行数
sum	•对某列数据求和
avg	•求某列数据的平均值
max	•求某列数据的最大值
min	•求某列数据的最小值



例如，从学生表中查询有多少学生，就可以用汇总函数 count 分析出表中有多少行。

学生表
(student)

学号	姓名	出生日期	性别
0001	猴子	1989-01-01	男
0002	猴子	1990-12-21	女
0003	马云	1991-12-21	男
0004	王思聪	1990-05-20	男

↓ 输入

```
select count(姓名)  
from student;
```

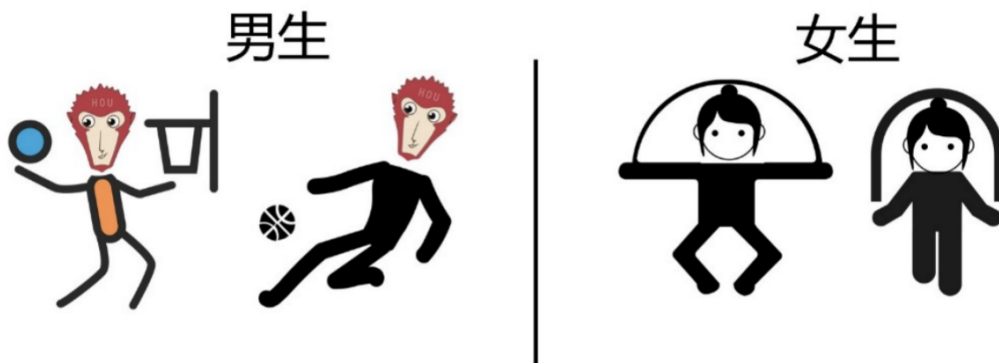
↓ 输出

count(姓名)
4



分组是把数据按某个条件分为几组，然后分析每一组数据。例如从学生表中查询每个性别的学生人数，这就需把学生按「性别」分组（也就是男生一组，女生一组），然后汇总每一组的人数。对应的 SQL 知识点是分组语句（group by），同时还需要掌握 having 语句，它可以对分组结果指定查询条件。

分组



SQL分组：group by



排序是对数据按照从大到小，或者从小到大的顺序排序。例如，在上学期期间每次做操都要按大小个子排队，这里的排队，就是按个子大小来排序。在 sql 中要会使用排序语句（order by）对查询结果排序。

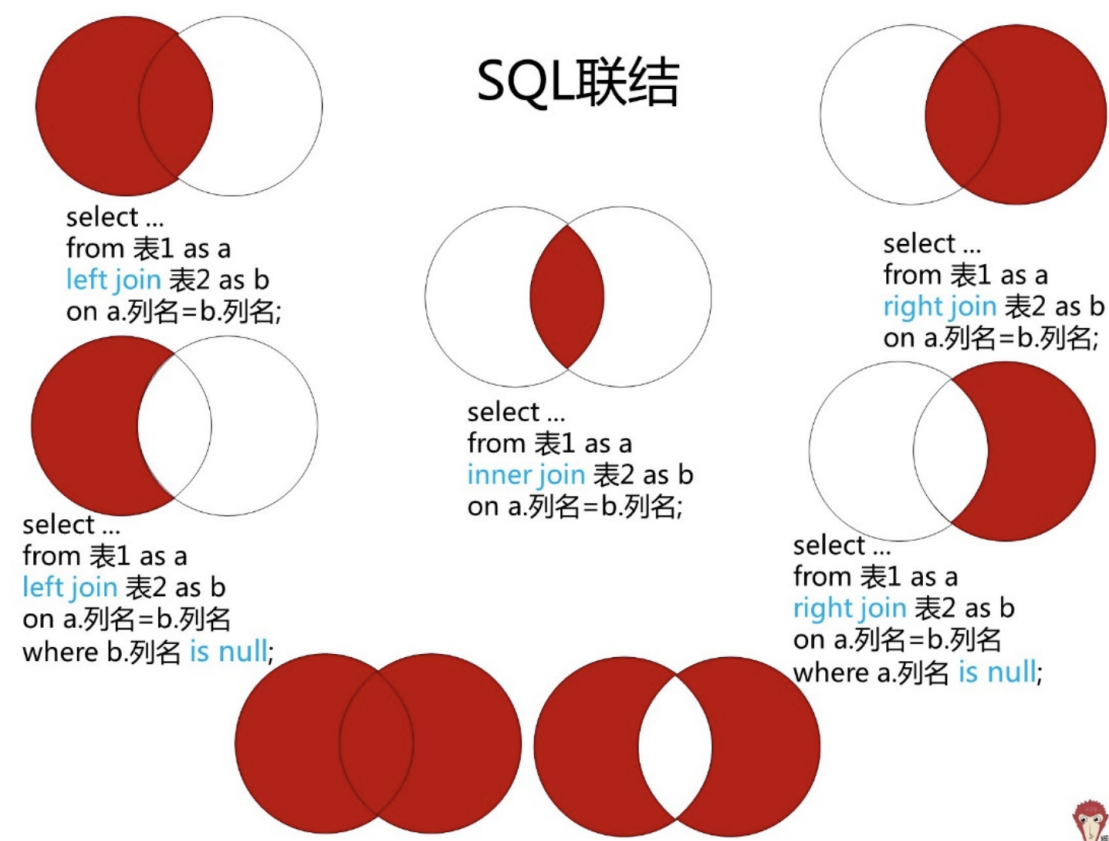
3) 复杂查询

实际工作中，业务分析比较复杂，往往需要很多 SQL 语句来完成复杂查询，这时候就需要用子查询来实现。子查询知识点需要掌握 3 种子查询类型：子查询、标量子查询、关联子查询。

4) 多表查询

工作里往往数据不是放在一个表里，而是不同的数据放在不同的表里，例如下图是学校数据库里的 4 张表。

这就需要会从不同的表中根据业务需求来获取数据，对应的就是掌握 SQL 多表查询的语法,包括下图的交叉联结、内联结、左联结、右联结、全联结。



4) 窗口函数

窗口函数是 SQL 的高级功能，可以解决这几类业务问题：

1. 排名问题。例如对每个班级的学生分别按成绩排名。
2. TopN 问题。例如找到每个类别下用户点击最多的 5 个商品是哪些。
3. 累计求和问题。例如医院的累计确诊人数是多少。
4. 分组比较问题。例如查找单科成绩高于该科目平均成绩的学生信息。

二、Excel 需要掌握下面的知识点：

数据清洗

数据重复【删除重复项】
数据抽取【*left, right, mid*】
数据计算【*average, sum, max, min, date, if, or, countif*】

数据分析

分组汇总【数据透视表】
描述统计分析【分析工具库】
多表关联查询【*vlookup, index*和*match*】

数据可视化

常用图表的制作
【散点图、折线图、柱状图、条形图等】



1) 数据清洗相关的功能

查找和删除重复数据，对应的是 Excel 里的删除重复项功能。

数据抽取，也就是从一堆杂乱的数据中找出想要的数 据，例如薪资「10k-20k」，需要从中截取出最低薪水「10」和最高薪水「20」，这就要用到 Excel 的字符串截取函数 *left*、*right*、*mid*) 。

数据计算函数，包括平均值(average)、求和(sum)、最大值(max)、最小值(min)等)，日期函数(date)，判断函数 (if, or, countif) 。

2) 数据分析相关的功能

描述统计分析对应 Excel 的分析工具库功能。

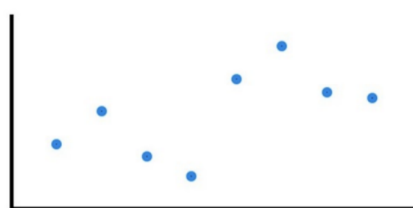
分组汇总对应 Excel 的数据透视表。

多表查询对应 Excel 的 vlookup, index 和 match 函数。

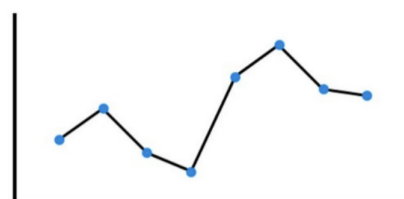
3) 数据可视化

会制作常用的图表，包括散点图、折线图、柱状图、条形图等。

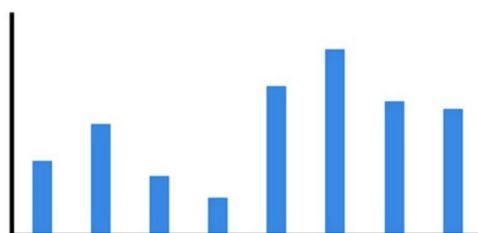
4种常用的图形



散点图



折线图



柱状图



条形图



三、Python 需要掌握的知识点

1) Python 数据分析工具

最常用的工具是 Anaconda、Jupyter Notebook 或者 JupyterLab，学会如何使用这些工具。

2) Python 基础语法

包括 4 个关键知识：数据类型、容器、函数、条件判断、循环。

快速学会
Python的
4个关键点

1) 数据 5种数据类型：字符串、数字，容器、布尔、空值None)； 容器：列表、元组、集合、字典 2) 函数 3) 条件判断 4) 循环
--

3) Python 数据分析包 (numpy, pandas)

4) Python 可视化图形包 (matplotlib)的使用，会绘制常用的图表。

5) Python 机器学习包 (scikit-learn)

6) 能够用 Python 操作数据，进行数据清洗，数据抽取，数据可视化等

