**数据库的范式**

第一范式：一张表，确定主键，以及需要记录的每个字段无法再分。

第二范式：首先是1NF,当一张表里的部分字段不完全依赖主键，通过非主键来确定。比如分数表：学号、系名、宿舍楼、课程、分数。系名、宿舍楼可以通过学校确定。则可以分成两张表。

第三范式：首先是 2NF，另外非主键列必须直接依赖于主键，不能存在传递依赖，一个订单表【Order】（OrderID，OrderDate，CustomerID，CustomerName，CustomerAddr，CustomerCity）主键是（OrderID）。 其中 OrderDate，CustomerID，CustomerName，CustomerAddr，CustomerCity 等非主键列都完全依赖于主键（OrderID），所以符合 2NF。不过问题是 CustomerName，CustomerAddr，CustomerCity 直接依赖的是 CustomerID（非主键列），而不是直接依赖于主键，它是通过传递才依赖于主键，所以不符合 3NF。 通过拆分【Order】为【Order】（OrderID，OrderDate，CustomerID）和【Customer】（CustomerID，CustomerName，CustomerAddr，CustomerCity）从而达到 3NF。

第二范式（2NF）和第三范式（3NF）的概念很容易混淆，区分它们的关键点在于，2NF：非主键列是否完全依赖于主键，还是依赖于主键的一部分；3NF：非主键列是直接依赖于主键，还是直接依赖于非主键列。第一范式确定表的主键，需要掌握的所有字段。第二范式将只依赖于主键一部分的字段剔除，第三范式把依赖于非主键字段的字段剔除。比如台账库，剔除指标属性是符合第二范式，保留报表iD剔除报表属性是为了符合第三范式。

**sql 查询**

**Mysql 表分大小写**

**Oracle 表名不分大小写**

**Sprk连接在DBI中用hive语句查询，hive函数可以用在sparklyr里面**

多表查询的内交只需要写where a.字段 =b.字段 AND 。。。。即可。

oracle里日期可以直接用符合日期格式的字符串来表示。不用转化为日期。

sql 取反用NOT

like \_表示一个字符，%表示多个字符，字符、数字都可以

SELECT INTO 在SQLsever里可用,

--(1)

SELECT IDENTITY(int, 1,1) AS ID\_Num

INTO NewTable

FROM OldTable

--(2)

SELECT ID\_Num = IDENTITY(int, 1, 1)

INTO NewTable

FROM OldTable

--(3)

select identity(int,1,1) kk,\* into #t from T

为所有T表的数据自动加自增序列，并放入临时表#t

**备份数据表**

ORACLE里可以用

create table NEW\_TEMPTABLE as select \* from TABLE1

-- 使用insert into select 拷贝数据（注意红色部分，可以自动生成id序列值）

insert into test2(id,testname,createtime,falg)

select seq\_test.nextval,t1.testname,t1.createtime,t1.falg from test1 t1;

-- 使用 create table select 创建被拷贝数据（注意要删除test2表先）

create table test2 as select t1.id,t1.testname,t1.createtime,t1.falg from test1 t1;

mysql 里用

create table dust select \* from student;//用于复制前未创建新表dust的情况下

Create table new\_table\_name (Select \* from old\_table\_name);

insert into dust select \* from student;//已经创建了新表dust的情况下

**union用处**

当在查询两种情况，比如按照佣金字段统计是否领取佣金的人数时候，不能用groupby因为会统计每个佣金的人数。可以用分开where查询，最后用union对接过进行合并。当然也可以decode函数计算新列，作为子查询。

**Sql的判断符**

not in （） 注意不能有空

in（） 与 =ANY() 是一样的。还有>any() 比返回的最小值要大 <any()比返回的最大值要小 >all() 比最大值都大 <all()比最小值要小。

**子查询中尽量加别名**

复杂查询 先列号查询格式，然后根据查询字段确定需要什么表，然后找到表与表之间的关联关系

R 与dplyr

int代表整数

dbl代表浮点数或者实数

chr代表字符向量或者字符串

dttm代表日期-时间

lgl代表逻辑向量，只含TRUE和FALSE

fctr代表因子，R用它来代表含固定可能值的分类变量

date代表日期

n\_distinct(...,na.rm=FALSE)