中软广发 广州

**数据同步怎么同步？使用的工具、技术、脚本是哪些？**

全量同步或者增量同步；使用kettle、sqoop或者写存储过程；

全量抽取：将目标表的全部诗句删除，再将源系统的数据全部插入目标表，但数据量大 的时候，性能太差

增量同步：只抽取新增的或修改的数据。但源表有更新的字段，在目标表中不存在，则 不需要更新，以时间戳取增量，对源表删除的数据无能为力。

方法一：***merge***（不是所有数据库都通用）若数据量大时merge删除比delete好

方法二：先将目标表的记录在***增量范围之内的删除***，然后再将增量插入

方法三：使用***游标***，逐一判断每一条增量纪录是需要更新还是插入

方法四：将增量记录***逐一插入***目标表，如果违反唯一索引则更新

**如果源表的记录删除了怎么办？**

源表的记录一般不会被删除，通常后面会加一个删除标识字段，用来标记这条记录是否失效；如果源表有数据删除了，可以查看源表的日志，看哪些记录被删除了，在目标表相应记录上加上删除标识，但是代价大，成本高。

**存储过程的结构？**

业务角度：1.处理逻辑；2.异常处理；3.日志，把运行过程记录到日志里面；

语法角度：1.过程名称、参数；2.包含在begin/and之间的代码3.存储过程的调用

**存储过程怎么往日志表里去写？实现的原理？写的内容？怎么实现呢？**

日志是用来记录错误和追溯问题、记录程序运行的情况，对程序进行修改和优化，提高性能；

1.先创一个日志表来记录日志；2.创建存储过程，来往日志表中插入日志信息；3.在其他存储过程执行的时候，调用日志的存储过程来记录具体的信息；

**创建自定义函数？**

Create or raplace function 函数名（参数1 数据类型，参数2 [in|out|in out]数据类型……）

Return 返回的数据类型

Is|as

PL/SQL 函数体；*--里面必须要有一个return子句*

**存储过程和函数的区别？**

1. 返回值的区别：函数有1个返回值，存储过程是通过参数返回，可以有多个或者没有；
2. 调用的区别：函数可以在查询语句中直接调用，而存储过程必须单独调用
3. 函数一般是用来计算并返回一个计算结果，存储过程一般是用来完成特定的数据操作（如修改、插入数据库表或执行DDL语句）

**异常处理or存储过程遇到过的报错**

PL/SQL中用exception关键字开始异常处理，异常发生时，进入异常处理部分，具体的异常与若干个when子句中知名的异常名称匹配，匹配成功就进入对应的异常处理部分，如果对应不成功，则进入others进行处理。

PL/SQL中预定义异常：--***存储过程遇到过的报错***

试图向唯一索引列插入重复值

试图进行非法游标操作

试图将字符串转换为数字

select into 中没有返回任何记录

select into 中返回多于1条记录

试图用0作为除数

试图打开一个已经打开的游标

**ETL掌握的技能，或者掌握了什么技术？**

数据库oracle、mysql、hive

ETL工具kettle、DataStage、infa、sqoop、datax

数据转换是ETL的核心环节，主要目标是将抽取到的各种数据进行数据的清洗、格式 的转换、缺失值填补

**ETL单词：*数据抽取extract-数据转换transform-数据加载load***

***Sqoop*抽到哪里？**

Sqoop主要用于在Hadoop和关系型数据库或大型机之间传输数据，可以将数据从关系型数据库管理系统导入（import）到Hadoop分布式文件系统中，或者将Hadoop中的数据转换导出（export）到关系数据库管理系统。

**Hive里面的Metastore是拿来干嘛的？**

客户端链接Metastore服务器，Metastore再去链接Mysql数据库才存取元数据。有了Metastore服务，就可以有多个客户端同时链接，而且这些客户端不需要知道Mysql数据库的用户名和密码，只需要链接Metastore服务即可

**上传文件到Hadoop的命令？用什么命令来传？**

1. 使用Hadoop命令向文件夹中放入数据

hadoop fs ***-put*** /mydir/dept.txt hdfs://hadoop-master:9000/test/dept/

1. 使用FTP命令链接远程机器，把文件上传到远程机器上面，再使用sqoop/load命令从 本地加载到hive表中。

湖南龙通科技 上海

1. 自我介绍
2. 为什么想要做这个，本专业不是学这个的
3. 最近的项目介绍
4. Oracle里面有哪些对象？
5. 表（Table）；
6. 索引（Index）；
7. 簇（Cluster）；
8. 视图；
9. 同义词Synonym；
10. 序列（Sequence）；
11. 过程（Procedure）和函数（Function）；
12. 触发器Trigger；
13. 约束Constraint；
14. Oracle里面的表空间类型（系统表空间、临时表空间、自定义表空间）
15. Oracle里面表空间高水位现象是怎么产生的

1.操作表时使用delete删除了大量数据。
 2.在插入时使用了/append nologging/语句，append关键字会从为表分配段中的随 机位置插入，水位线会不断增高。
 3.Sql load 时默认使用truncate 自带了reuse storage参数，导致truncate以 后水位线不会降低。

1. 使用truncate来解决高水位现象这个你自己有试验过吗？

1.直接truncate table drop storage
 2.建立一张维护表定期move并重建索引或者shrink space。
 3.表数据建表时按照日期建立了备份表，保留一定天数数据
 4.Rename表名，重建表，重建索引，将数据导入重建表，drop原表，然后rename重 建表为原表

1. 什么时候用trancate什么时候用delete

当表被TRUNCATE 后，这个表和索引所占用的空间会恢复到初始大小，

 DELETE操作不会减少表或索引所占用的空间。

 drop语句将表所占用的空间全释放掉。

1、在速度上，一般来说，drop> truncate > delete。

2、在使用drop和truncate时一定要注意，虽然可以恢复，但为了减少麻烦，还是 要慎重。

3、如果想删除部分数据用delete，注意带上where子句，回滚段要足够大；

 如果想删除表，用drop；

 如果想保留表而将所有数据删除，如果和事务无关，用truncate即可；

 如果和事务有关，或者想触发trigger，还是用delete；

 如果是整理表内部的碎片，可以用truncate跟上reuse stroage，再重新导入/ 插入数据

1. 关联条件和过滤条件有什么区别

内关联无区别。

外关联，无论on后面的条件是否为真都返回主表的数据，where条件不为真的全部过 滤掉，没有外关联的含义了

1. 把一个条件写在关联后面和写在过滤后面有什么区别，写在on后面和写在where后面 有什么区别
2. Count（\*）count（1）count（列）有什么区别

从执行结果来说:

count(1)和 count(\*)之间没有区别，因为 count(\*)count(1)都不会去过滤空值， 但 count(列名)会过滤空值。

从执行效率来说:

(1)如果列为主键，count(列名)效率优于 count(1) ;如果列不为主键，count(1)效率优于 count(列名) ;

(2)如果表中存在主键，count(主键列名)效率最优;

(3)如果表中只有一列，则 count(\*)效率最优;如果表有多列，且不存在主键，则 count(1) 效率优于 count(\*);

1. 索引有什么类型

1，普通索引：普通索引是最基本的索引，它没有任何限制，值可以为空；仅加速查询。
 2，唯一索引：唯一索引与普通索引类似，不同的就是：索引列的值必须唯一，但允许 有空值。如果是组合索引，则列值的组合必须唯一。
 3，主键索引：主键索引是一种特殊的唯一索引，一个表只能有一个主键，不允许有空 值。
 4，组合索引：组合索引指在多个字段上创建的索引，只有在查询条件中使用了创建索 引时的第一个字段，索引才会被使用。使用组合索引时遵循最左前缀集合。
 5，全文索引：全文索引主要用来查找文本中的关键字，而不是直接与索引中的值相比 较。fulltext索引跟其它索引大不相同，它更像是一个搜索引擎，而不是简单的where 语句的参数匹配。

1. 索引有什么优点和缺点

*索引的优点如下：*

通过创建唯一索引可以保证数据库表中每一行数据的唯一性。

可以给所有的 MySQL 列类型设置索引。

可以大大加快数据的查询速度，这是使用索引最主要的原因。

在实现数据的参考完整性方面可以加速表与表之间的连接。

在使用分组和排序子句进行数据查询时也可以显著减少查询中分组和排序的时间

*缺点主要如下：*

创建和维护索引组要耗费时间，并且随着数据量的增加所耗费的时间也会增加。

索引需要占磁盘空间，除了数据表占数据空间以外，每一个索引还要占一定的物理空间。如果有大量的索引，索引文件可能比数据文件更快达到最大文件尺寸。

当对表中的数据进行增加、删除和修改的时候，索引也要动态维护，这样就降低了数据的维护速度。

使用索引时，需要综合考虑索引的优点和缺点。

索引可以提高查询速度，但是会影响插入记录的速度。因为，向有索引的表中插入记录时，数据库系统会按照索引进行排序，这样就降低了插入记录的速度，插入大量记录时的速度影响会更加明显。这种情况下，最好的办法是先删除表中的索引，然后插入数据，插入完成后，再创建索引。

1. Linux系统，查看的命令是什么

1、列出文件清单命令，*ls*。ls命令能够列出当前目录下的所有内容。ls 命令的执行方式为：# ls [-选项] [文件名或者目录名]

2、*cat*命令，功能：在标准输出上显示文件。语法：cat [-vTEuAte] 文件；

3、more命令，功能：在终端屏幕按屏显示文本文件。语法： more [-pcdls] 文件；

4、less命令，less命令的功能几乎和more命令一样，也是用来按页显示文件，不同之处在于less命令在显示文件时允许用户既可以向前又可以向后翻阅文件。

5、head命令，功能：显示指定文件的前若干行。缺省设置为显示10行。语法：head [-n] 文件；

6、tail命令，功能：显示指定文件的末尾若干行。缺省设置为显示10行。语法：tail [+ / - num ] [参数] 文件；

7、grep、fgrep和egrep命令，功能：这组命令以指定模式搜索文件，并通知用户在什么文件中搜索到与指定的模式匹配的字符串，并打印出所有包含该字符串的文本行，在该文本行的最前面是该行所在的文件名。grep命令一次只能搜索一个指定的模式；egrep命令检索扩展的正则表达式（包括表达式组和可选项）；fgrep命令检索固定字符串，它不识别正则表达式，是快速搜索命令。语法：grep [-EFbcihlnvxef] [查找模式] [文件名1，文件名2，……]，egrep [选项] [查找模式] [文件名1，文件名2，……]，fgrep [选项] [查找模式] [文件名1，文件名2，……]。

1. Linux里面查看文件某一行数据怎么查看

1、tail date.log 输出文件末尾的内容，默认10行

 tail -20 date.log 输出最后20行的内容

 tail -n -20 date.log 输出倒数第20行到文件末尾的内容

 tail -n +20 date.log 输出第20行到文件末尾的内容

 tail -f date.log 实时监控文件内容增加，默认10行

2、head date.log 输出文件开头的内容，默认10行

 head -15 date.log 输出开头15行的内容

 head -n +15 date.log 输出开头到第15行的内容

 head -n -15 date.log 输出开头到倒数第15行的内容

3、sed -n "开始行，结束行p" 文件名

 sed -n '70,75p' date.log 输出第70行到第75行的内容

 sed -n '6p;260,400p; ' 文件名 输出第6行 和 260到400行

 sed -n 5p 文件名 输出第5行

tail 和 head 加上 -n参数后 都代表输出到指定行数，tail 是指定行数到结尾，head是开头到指定行数

+数字 代表整数第几行，

-数字 代表倒数第几行。

1. 在职还是离职
2. 离职原因
3. 现在薪资状况
4. 期望薪资