数据库内核月报 - 2020 / 08

当期文章 Database • 案例分析 • UTF8与GBK数据库字符集

问题背景

现有数据库A与数据库B，数据库A服务端由GBK编码，数据库B服务端由UTF8编码，需要完成数据库A至数据库B的数据导入，测试中发现A库数据插入B数据库时的部分数据进行查询时存在编码转换报错。

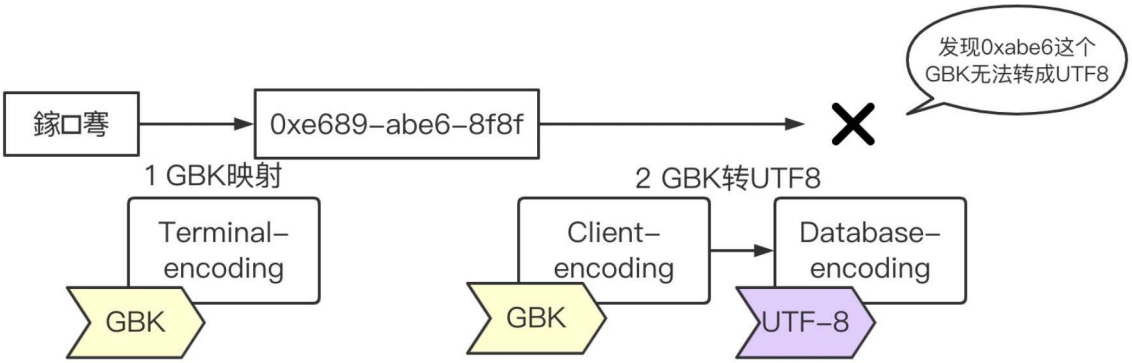
问题分析

角色分析

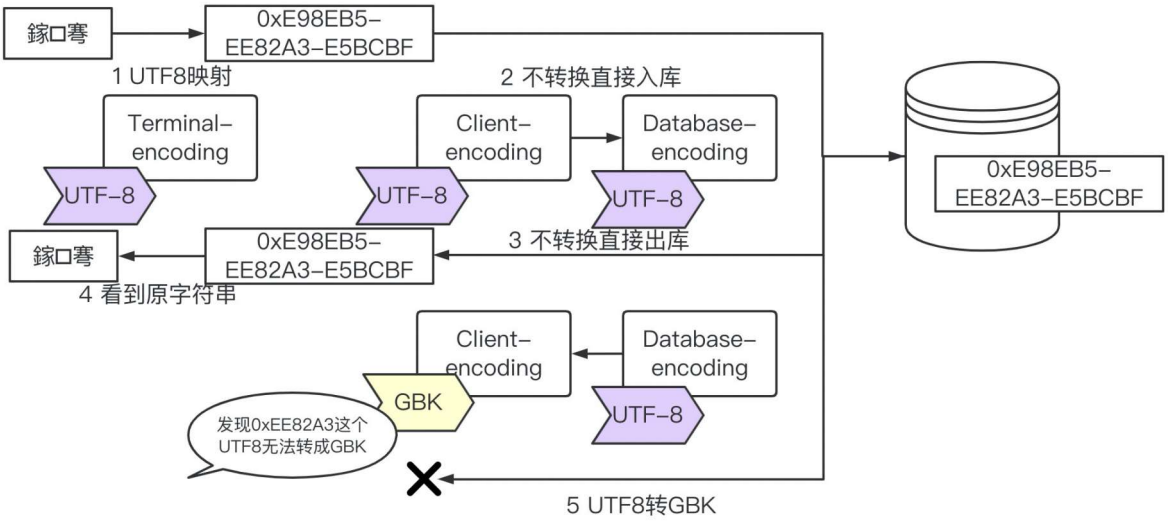
首先阐述影响字符编码的几个要素 • Terminal-encoding(用户客户端编码，Iterm编码，终端编码)：该编码格式不参与编码转换，其负责将一个字符串映射成字符编码。例如‘鎵’这个字，如果被作为GBK解析，会解析成 0xE689; 如果被当成UTF8解析，会被解析成 0xE98EB5 。这些二进制编码用户感知不到，而被数据库储存。 • Client-encoding(数据库客户端编码，client_coding)：该编码格式是数据库识别该编码格式的一个参考。由于字符串被Server-encoding解析进入数据库后以二进制编码的形式存储，由Client-encoding唯一标识数据库中的二进制编码原本是什么编码格式。 ⚠如果数据库只有二进制编码是毫无意义的，因为数据库只储存了类似于0xe689abe68f8f这样的二进制，如果没有编码格式甚至不知道其可以被解析为几个字符，解析的规则由字符集指定。 • Database-encoding(数据库服务器编码) 对数据库B来说是 UTF-8模式且不支持GBK格式。含义就是所有的二进制编码如果不是UTF8编码，会被转义成UTF8编码入库；同理如果读出时client_encoding不是UTF8会被转义成其他二进制编码出库。

场景分析

场景a 终端字符集为GBK，数据库client_encoding为GBK，database_encoding 为 UTF8 该场景下，鎵□骞 被还原出了正确的原编码，然而这个编码被当成GBK去转义UTF8，发现 0xabe6这个编码（原场景中的□ ）无法作为一个GBK去转UTF8 导致转义失败。

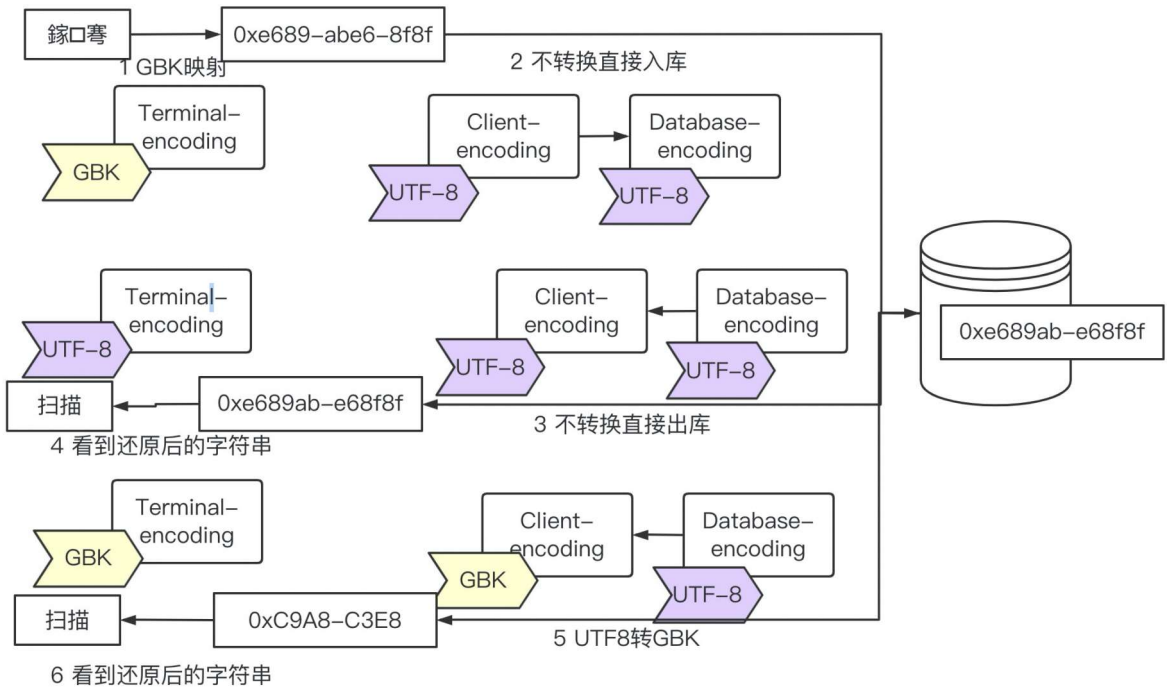


场景b 终端字符集为UTF8，数据库client_encoding为UTF8，database_encoding 为 UTF8 该场景下，镓口骞 被按照UTF编码的格式还原出了UTF8编码并入库。如果仍旧按照这个格式读出，可以得到原字符；如果按照GBK的格式转码，发现 0xEE82A3 这个编码没有GBK对应的字符。



场景c 终端字符集为GBK，数据库client_encoding为UTF8，database_encoding 为 UTF8 该场景下，镓口骞 被还原出了正确的原编码，并被当成UTF8去入库。这种情况下，

不管是UTF8去读，还是GBK去读，都可以读出正确的字符串。



问题小节

• 在一个合规的流程中，Terminal_encoding 及 Client_encoding 应该是完全一致的。这两这其实是一体的两面，分别代表了一个字符串应该被如何编码，和一个编码如何被解析成字符串。这就对应了场景a和场景b的1234 。场景a问题是由于鎵口褰 这个字符串本身不能被GBK编码。 • 由于鎵口褰 这个字符串本身就是‘扫描’错误解析下的产物，场景c 通过这种不合规的实验还原出了原字符。

问题原因

出现这些问题的根本原因是A库中的“GBK”范围大于B库中设置的GBK。A所谓“GBK”编码的字符集实际上是GB18030。

编码背景资料

GB2312、GBK与GB18030

• GB 2312 或 GB 2312-80 是中国国家标准简体中文字符集，全称《信息交换用汉字编码字符集·基本集》，又称 GB 0，由中国国家标准总局发布，1981 年 5 月 1 日实施。GB 2312 编码通行于中国大陆；新加坡等地也采用此编码。中国大陆几乎所有的中文系统和国际化的软件都支持 GB 2312。GB 2312 标准共收录 6763 个汉字，GB 2312 对任意一个图形字符都采用两个字节表示 • GBK 即汉字内码扩展规范，K 为汉语拼音 Kuo Zhan（扩展）中“扩”字的声母。英文全称 Chinese Internal Code Specification。GBK 共收入

21886 个汉字和图形符号，包括：GB 2312 中的全部汉字、非汉字符号。BIG5 中的全部汉字。与 ISO 10646 相应的国家标准 GB 13000 中的其它 CJK 汉字，以上合计 20902 个汉字。其它汉字、部首、符号，共计 984 个。 • GB 18030，全称：国家标准 GB 18030-2005《信息技术中文编码字符集》，是中华人民共和国现时最新的内码字集，是 GB 18030-2000《信息技术信息交换用汉字编码字符集基本集的扩充》的修订版。GB 18030 与 GB 2312-1980 和 GBK 兼容，共收录汉字70244个。与 UTF-8 相同，采用多字节编码，每个字可以由 1 个、2 个或 4 个字节组成。编码空间庞大，最多可定义 161 万个字符。支持中国国内少数民族的文字，不需要动用造字区。汉字收录范围包含繁体汉字以及日韩汉字。GB 18030 编码是一二四字节变长编码。 • 国家标准GB18030-2000《信息交换用汉字编码字符集基本集的补充》是我国继GB2312-1980和GB13000-1993之后最重要的汉字编码标准，是我国计算机系统必须遵循的基础性标准之一。GB18030-2000编码标准是由信息产业部和国家质量技术监督局在2000年 3月17日联合发布的，并且将作为一项国家标准在2001年的1月正式强制执行。GB18030-2005《信息技术中文编码字符集》是我国制订的以汉字为主并包含多种我国少数民族文字（如藏、蒙古、傣、彝、朝鲜、维吾尔文等）的超大型中文编码字符集强制性标准，其中收入汉字70000余个。

编码小节

• GB2312 -> GBK -> GB18030 是逐渐扩充的集合，其向下兼容 • 我国现有的汉字编码字符集标准是 GB18030

解决方案

在导入与导出数据时，如果A库是“GBK“或类“GBK”字符集传输或存储数据，B库设置客户端字符集为“GB18030”。

阿里云RDS-数据库内核组

欢迎在github上star AliSQL

阅读： -