10.10.7.2 The gb18030 Character Set

In MySQL, the `gb18030` character set corresponds to the "Chinese National Standard GB 18030-2005: Information technology—Chinese coded character set", which is the official character set of the People's Republic of China (PRC).

Characteristics of the MySQL gb18030 Character Set

- Supports all code points defined by the GB 18030-2005 standard. Unassigned code points in the ranges (GB+8431A439, GB+90308130) and (GB+E3329A36, GB+EF39EF39) are treated as `'?'` (0x3F). Conversion of unassigned code points return `'?'`.

- Supports UPPER and LOWER conversion for all GB18030 code points. Case folding defined by Unicode is also supported (based on `CaseFolding-6.3.0.txt`).

- Supports Conversion of data to and from other character sets.

- Supports SQL statements such as `SET NAMES`.

- Supports comparison between `gb18030` strings, and between `gb18030` strings and strings of other character sets. There is a conversion if strings have different character sets. Comparisons that include or ignore trailing spaces are also supported.

- The private use area (U+E000, U+F8FF) in Unicode is mapped to `gb18030`.

- There is no mapping between (U+D800, U+DFFF) and GB18030. Attempted conversion of code points in this range returns `'?'`.

- If an incoming sequence is illegal, an error or warning is returned. If an illegal sequence is used in `CONVERT()`, an error is returned. Otherwise, a warning is returned.

- For consistency with `utf8` and `utf8mb4`, UPPER is not supported for ligatures.

- Searches for ligatures also match uppercase ligatures when using the `gb18030_unicode_520_ci` collation.

- If a character has more than one uppercase character, the chosen uppercase character is the one whose lowercase is the character itself.

- The minimum multibyte length is 1 and the maximum is 4. The character set determines the length of a sequence using the first 1 or 2 bytes.

Supported Collations

- `gb18030_bin`: A binary collation.

- `gb18030_chinese_ci`: The default collation, which supports Pinyin. Sorting of non-Chinese characters is based on the order of the original sort key. The original sort key is `GB(UPPER(ch))` if `UPPER(ch)` exists. Otherwise, the original sort key is `GB(ch)`. Chinese characters are sorted according to the Pinyin collation defined in the Unicode Common Locale Data Repository (CLDR 24). Non-Chinese characters are sorted before Chinese characters with the exception of `GB+FE39FE39`, which is the code point maximum.

- `gb18030_unicode_520_ci`: A Unicode collation. Use this collation if you need to ensure that ligatures are sorted correctly.

10.10.8 The Binary Character Set