



AN OVERVIEW TO COLLATION AND CHARACTER SETS

-THIYAGHU CK



ABOUT US

- Founded by a group of enthusiastic and dedicated individuals
- Hands on experience in MySQL versions 3.x, 4.0, 4.1, 5.0, & 5.1
- Experience in working with large set ups like Sify, Yahoo, Genpact, Cricinfo, BSNL, & TCS
- Have 7 Certified MySQL DBAs and a cluster certified MySQL professional with an average of 6 years of experience and 4 DBAs with an average of 3 years of experience
- Managed MySQL database driving Platforms with database volume ranging from 10GB to 6 TB



OUR SERVICES

- Design and architecture services
- Query tuning and performance optimization
- Migration from any database to MySQL
- Monitoring and remote maintenance of MySQL databases
- Design, Implementation and Maintenance of MySQL high availability solutions



AN OVERVIEW TO COLLATION AND CHARACTER SETS



COLLATION AND CHARACTER SETS

- What is Charset?

A *character set* is a set of symbols and encodings. Or a collection of signs.

A collection of signs ...

³*#⊗Ⓜ∕!%&?≠”♣☾∩∴∟≡
 {}-€∂ ← ↓ → ↓ ϕ ~+

1-9

1	2	3
4	5	6
7	8	9

The German alphabet

AaÄäBbCcDdEeFfGgHhIijjKkLlMmNnO
 oÖöPpQqRrSsßTtUuÜüVvWwXxYyZz

UNICODE

The Greek alphabet

ΑαΒβΓγΔδΕεΖζΗηΘθΙιΚκΛλΜμΝν
 ΞξΟοΠπΡρΣσΤτΥυΦφΧχΨψΩω

A-Z

ABCDEFGHIJKLMNOPQRSTUVWXYZ

Roman numbers

I V X L C D M A

ISO-8859-15

NUL	SOH	STX	ETX	EOT	ENQ	ACK	BEL	BS	HT	LF	VT	FF	CR	SO	SI
DLE	DC1	DC2	DC3	DC4	NAK	SYN	ETB	CAN	EM	SUB	ESC	FS	GS	RS	US
SP	!	"	#	\$	%	&	'	()	*	+	,	-	.	/
0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
P	Q	R	S	T	U	V	W	X	Y	Z	[\]	^	_
`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
p	q	r	s	t	u	v	w	x	y	z	{		}	~	DEL
PAD	HOP	BPH	NBH	IND	NEL	SSA	ESA	HTS	HTJ	VTS	PLD	PLU	RI	SS2	SS3
DCS	PU1	PU2	STS	CCH	MW	SPA	EPA	SOS	SGCI	SCI	CSI	ST	OSC	PM	APC
NBSP	ı	¢	£	€	¥	Š	š	š	®	®	«	»	SHY	®	-
°	±	²	³	Ž	μ	¶	·	ž	ž	»	Œ	œ	Ÿ	ı	ı
À	Á	Â	Ã	Ä	Å	Æ	Ç	È	É	Ê	Ë	Ì	Í	Î	Ï
Ð	Ñ	Ò	Ó	Ô	Õ	Ö	×	Ø	Ù	Ú	Û	Ü	Ý	Þ	ß
à	á	â	ã	ä	å	æ	ç	è	é	ê	ë	ì	í	î	ï
ð	ñ	ò	ó	ô	õ	ö	÷	ø	ù	ú	û	ü	ý	þ	ÿ



COLLATION(1/2)

- What is Collation?

A collation is a set of rules for comparing characters in a character set.

- How it encode?

Suppose that we have an alphabet with four letters: "A", "B", "a", "b".
We give each letter a number: "A" = 0, "B" = 1, "a" = 2, "b" = 3.

0=>Encoding of A

1=>Encoding of B

2=>Encoding of a

3=>Encoding of b



COLLATION(2/2)

- How it compare?

If we need to compare two strings, say “A” and “B”. It just compare 0 and 1.

0!=1

- What is the collation here?

Only one rule in this case: “compare the encodings.”

- Issue here?

Lowercase and Uppercase letters are equivalent? . No, so the rules should be (1) Treat the lowercase letters “a” and “b” as equivalent to “A” and “B”; (2) Then compare the encodings. We call this a *case-insensitive* collation



SHOW COLLATION AND CHARSET

```
mysql> show character set;
```

Charset	Description	Default collation	Maxlen
big5	Big5 Traditional Chinese	big5_chinese_ci	2
dec8	DEC West European	dec8_swedish_ci	1
cp850	DOS West European	cp850_general_ci	1
hp8	HP West European	hp8_english_ci	1
koi8r	KOI8-R Relcom Russian	koi8r_general_ci	1
latin1	cp1252 West European	latin1_swedish_ci	1
latin2	ISO 8859-2 Central European	latin2_general_ci	1

```
mysql> SHOW COLLATION LIKE 'latin1%';
```

Collation	Charset	Id	Default	Compiled	Sortlen
latin1_german1_ci	latin1	5		Yes	1
latin1_swedish_ci	latin1	8	Yes	Yes	1
latin1_danish_ci	latin1	15		Yes	1
latin1_german2_ci	latin1	31		Yes	2
latin1_bin	latin1	47		Yes	1
latin1_general_ci	latin1	48		Yes	1
latin1_general_cs	latin1	49		Yes	1
latin1_spanish_ci	latin1	94		Yes	1

```
8 rows in set (0.00 sec)
```



COLLATION CHARACTERISTICS

- One character sets may have many collation.
- Each character set has one collation that is the *default collation*. For example, the default collation for latin1 is latin1_swedish_ci.
- There is a convention for collation names. They start with the name of the character set associated.
 - Latin1 should have the collation name start with latin_XXXXX
 - Utf8 should have the collation name start with utf8_XXXXX
- If the collation ends with _ci which is the Case Insensitive and the collation ends with _cs is Case Sensitive and _bin is binary.



4 LEVELS(1/4)

- Character sets and collations at four levels: **server, database, table, and column.**

I. Server Character Set and Collation:

- MySQL variable `character_set_server / collation_server`
- The encoding that the server is using internally.
- Default `character_set_server` is *latin1*.
- Default `collation_server` is *latin1_swedish_ci*.
- Values can be changed at runtime.



4 LEVELS(2/4)

II. Database Character Set and Collation:

- MySQL variable `character_set_database` / `collation_database`
- If the charset and collation is not defined then the default value is `character_set_server`
- Syntax: `CREATE DATABASE db_name CHARACTER SET latin1 COLLATE latin1_swedish_ci;`
- Character set and collation have to match to each other.

```
mysql> CREATE DATABASE coll_test CHARACTER SET latin1 COLLATE
utf8_general_ci;
ERROR 1253 (42000): COLLATION 'utf8_general_ci' is not valid for
CHARACTER SET 'latin1'
```



4 LEVELS(3/4)

III. Table Character Set and Collation:

- No such variables.
- Syntax: `CREATE TABLE t1 (...)CHARACTER SET latin1 COLLATE latin1_danish_ci;`
- If charset or the collate is missing, it will take the default



4 LEVELS(4/4)

IV. Column Character Set and Collation:

- No such variables.
- Syntax: CREATE TABLE t(..., v varchar(100) CHARSET mychar COLLATE mycoll, ...) ...;
- Example:

```
CREATE TABLE t1( col1 VARCHAR(5) CHARACTER SET latin1  
COLLATE latin1_german1_ci);
```

```
ALTER TABLE t1 MODIFY col1 VARCHAR(5) CHARACTER SET  
latin1 COLLATE latin1_swedish_ci;
```



LATIN1

- *Latin alphabet No. 1*, is part of the ISO/IEC 8859 series of ASCII-based standard character encodings, first edition published in 1987.
- It is informally referred to as **Latin-1**
- This character-encoding scheme is used throughout The Americas, Western Europe, Oceania, and much of Africa. It is also commonly used in most standard Romanization of East-Asian languages.
- Modern languages with complete coverage of their alphabet
Afrikaans, Albanian, Breton, Catalan, Danish, English (UK and US), Faroese, Galician, German, Icelandic, Irish (new orthography),....., Occitan, Portuguese, Rhaeto-Romanic, Scottish Gaelic, Spanish, Swahili, Swedish, Walloon, Basque.



Examples(1/2)

```
mysql> CREATE TABLE t01(v varchar(100));
mysql> INSERT INTO t01(v) VALUES('a');
mysql> SELECT COLLATION(v) FROM t01;
+-----+
| COLLATION(v) |
+-----+
| latin1_swedish_ci |
+-----+
mysql> CREATE TABLE t02(v varchar(100)) CHARSET utf8;
mysql> INSERT INTO t02(v) VALUES('a');
mysql> SELECT COLLATION(v) FROM t02;
+-----+
| COLLATION(v) |
+-----+
| utf8_general_ci |
+-----+
mysql> CREATE TABLE t03(v varchar(100)) COLLATE utf8_bin;
mysql> INSERT INTO t03(v) VALUES('a');
mysql> SELECT COLLATION(v) FROM t03;
+-----+
| COLLATION(v) |
+-----+
| utf8_bin |
+-----+
```



Examples(2/2)

```
mysql> CREATE TABLE t07(v varchar(100) CHARSET utf8) COLLATE utf8_bin;
```

```
mysql> INSERT INTO t07(v) VALUES('a');
```

```
mysql> SELECT COLLATION(v) FROM t07;
```

```
+-----+
| COLLATION(v) |
+-----+
| utf8_general_ci |
```

```
+-----+
mysql> alter table t07 add column v1 varchar(10);
Records: 1 Duplicates: 0 Warnings: 0
```

```
mysql> select collation(v1) from t07;
```

```
+-----+
| collation(v1) |
+-----+
| utf8_bin      |
```

```
+-----+
mysql> select collation(v) from t07;
```

```
+-----+
| collation(v) |
+-----+
| utf8_general_ci |
```

```
+-----+
1 row in set (0.00 sec)
```



CHARACTER SET VARIABLES(1/3)

- There are 8 variables as listed below.

```
mysql> show variables like 'character_%';
```

```
+-----+-----+
| Variable_name      | Value |
+-----+-----+
| character_set_client | latin1 |
| character_set_connection | latin1 |
| character_set_database | latin1 |
| character_set_filesystem | binary |
| character_set_results | latin1 |
| character_set_server | latin1 |
| character_set_system | utf8 |
| character_sets_dir   | C:\Program Files\MySQL\MySQL Server 5.1\share\charsets\ |
+-----+-----+
```

```
8 rows in set (0.00 sec)
```



CHARACTER SET VARIABLES(2/3)

- **character_set_system**
 - The character set used by the server for storing identifiers.
 - The value is always utf8 and not able to change.
- **character_set_server**
 - The encoding that the server is using internally.
- **character_set_client**
 - The character set for statements that arrive from the client.
- **character_set_results**
 - The character set used for returning query results such as result sets or error messages to the client.
- **character_set_connection**
 - Character set for odbc/jdbc/.. transfer layer.



CHARACTER SET VARIABLES(3/3)

- **character_set_database**
 - The character set used by all the tables in the database. If not defined, the variable has the same value as `character_set_server`.
- **character_set_filesystem**
 - The file system character set.
 - Used in `LOAD DATA INFILE` and `SELECT ... INTO OUTFILE` statements and the `LOAD_FILE()` function.
 - Such file names are converted from [character_set_client](#) to [character_set_filesystem](#) before the file opening attempt occurs. The default value is binary.
- **character_sets_dir**
 - The directory where character sets are installed.
- **Note:**

`character_set_client`, `character_set_result` and `character_set_connection` should have same values.



THANK YOU