

ORACLE®

A low-angle photograph of the Atomium structure in Brussels, Belgium, against a clear blue sky. The structure's spherical nodes are highly reflective, mirroring the sky and surrounding environment. A small flag is visible on top of one of the spheres. The overall image has a semi-transparent blue overlay with the word 'MySQL' in a large, light blue font.

Regular Expressions with full Unicode support

The ins and outs of the new regular expression functions and the ICU library

Martin Hansson
Software Development
MySQL Optimizer Team

Safe Harbor Statement

The following is intended to outline our general product direction. It is intended for information purposes only, and may not be incorporated into any contract. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions. The development, release, and timing of any features or functionality described for Oracle's products remains at the sole discretion of Oracle.

What Happened?

Old regexp library (Henry Spencer)

- Does not support Unicode
- Limited Features
- No resource control
- Only Boolean Search

<https://mysqlserverteam.com/new-regular-expression-functions-in-mysql-8-0/>

Not some niche feature

Feature Requests for Extracting Substring:

Bug#79428 No way to extract a substring matching a regex

Bug#29781 Adding in Pattern Replace (RegExp) for MySQL Engine

Bug#16357 add in functions to do regular expression replacements in a select query

Bug#9105 Regular expression support for Search & Replace

51 “affects me” total

CTE had 59 “affects me”

New Regular Expression Functions

REGEXP_INSTR

REGEXP_LIKE

REGEXP_REPLACE

REGEXP_SUBSTR

Program Agenda

- Security
- ICU library
- Unicode
- Working with Unicode in Regular Expressions

Two Security Concerns



Memory



Runtime

Security

Cap on runtime

```
mysql> SELECT regexp_instr(  
          'AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAC',  
          '(A+)+B' );
```

ERROR 3699 (HY000): Timeout exceeded in regular
expression match.

Security

Cap on Memory

```
mysql> SELECT regexp_instr(  
    '', '((((((( {120}) {11}) {11}) {11}) {80}) {11}) {4}' );
```

ERROR 3699 (HY000): Timeout exceeded in regular expression match.

```
mysql> SET GLOBAL regexp_stack_limit = 239;
```

```
mysql> SELECT regexp_instr(  
    '', '((((((( {120}) {11}) {11}) {11}) {80}) {11}) {4}' );
```

ERROR 3698 (HY000): Overflow in the regular expression backtrack stack.

Program Agenda

- Security
- ICU library
- Unicode
- Working with Unicode in Regular Expressions

ICU library



Building ICU

Need three libraries

- i18n library
 - Regular expressions
 - Character sets
- Common library
- Data Library

Program Agenda

- Security
- ICU library
- Unicode
- Working with Unicode in Regular Expressions

UTF-32

ab🍺d

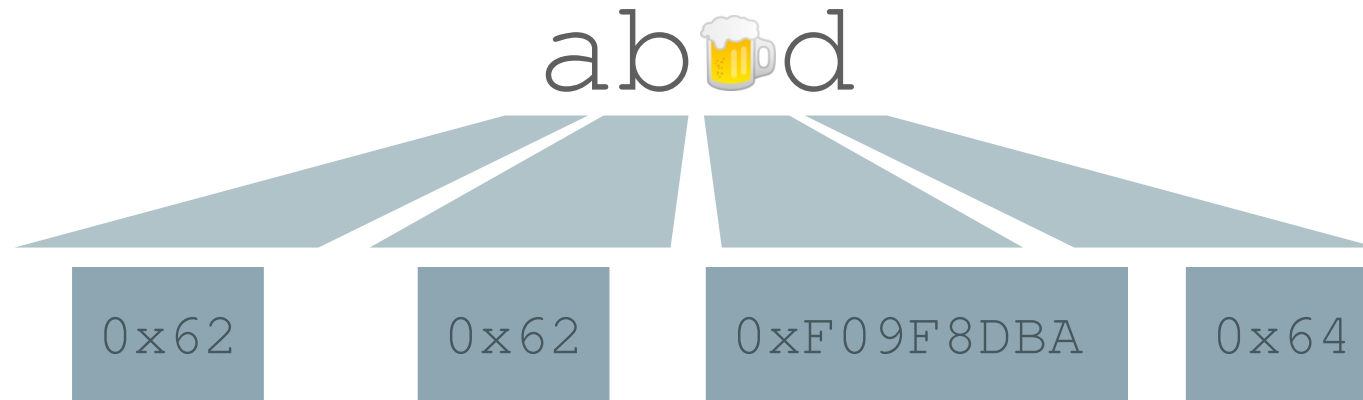
0x00000061

0x00000062

0x0001f37a

0x00000064

UTF-8



UTF-16

ab🍺d

0x0062

0x0062

0x3CD87ADF

0x0064

Under the Hood

- Count codepoints
- Convert to UTF-16
- Use the C API
- Convert back *if needed*

Program Agenda

- Security
- ICU library
- Unicode
- Working with Unicode in Regular Expressions

Case folding

Simple case sensitivity

```
mysql> SELECT regexp_like( 'a', '(?i)A' ); # mode modifier
1
```

```
mysql> SELECT regexp_like( 'a', 'A', 'i' ); # match_parameter
1
```

```
mysql> SELECT regexp_like(
    'a' COLLATE utf8mb4_0900_as_cs, 'A' ); # collation
0
```

Case folding

Simple case sensitivity

```
mysql> SELECT regexp_like( 'Abc', 'abC', 'c' );
```

```
→ 0
```

```
mysql> SELECT regexp_like( 'Abc', 'abC', 'i' );
```

```
→ 1
```

Case folding

Case-mapping process

A → a

B → b

C → c

Case folding

Full Case Folding

ß → ss

```
mysql> SELECT regexp_like( 'ß', '^ss$', 'c' );
```

→ 0

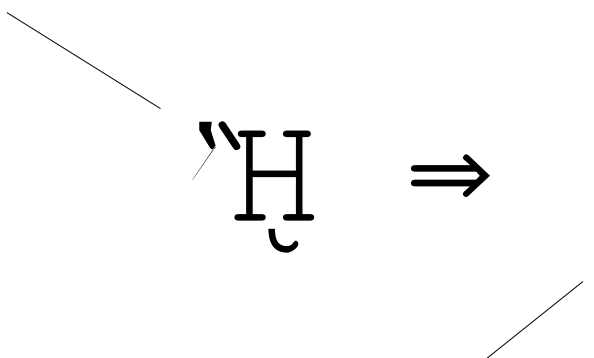
```
mysql> SELECT regexp_like( 'ß', '^ss$', 'i' );
```

→ 1

Case folding

Full Case Folding

U+1F9B GREEK CAPITAL LETTER ETA WITH DASIA AND VARIA AND
PROSGEGRAMMENI



“Η ⇒ ἥ ι

The diagram illustrates the case folding process. A line connects the text 'U+1F9B GREEK CAPITAL LETTER ETA WITH DASIA AND VARIA AND PROSGEGRAMMENI' to the character '“Η'. Another line connects the character 'ἥ ι' to the text 'U+1F23 U+03B9 GREEK SMALL LETTER ETA WITH DASIA AND VARIA'. A double arrow points from '“Η' to 'ἥ ι', indicating the folding operation.

U+1F23 U+03B9 GREEK SMALL LETTER ETA WITH DASIA AND VARIA

Case folding

Has to Look Like a String in order to Match

```
mysql> SELECT regexp_like( 'ß', '^ss$' );  
→ 1
```

```
mysql> SELECT regexp_like( 'ß', '^s+$' );  
→ 0
```

```
mysql> SELECT regexp_like( 'ß', '^s{2}$' );  
→ 0
```

Case folding

Can't start Match Within Expanded Character

```
mysql> SELECT regexp_like( 'ß', 's$' );  
→ 0
```

```
mysql> SELECT regexp_like( 'ß', '^s' );  
→ 0
```

Case folding

Collations

```
mysql> select 'ß' collate utf8mb4_de_pb_0900_ai_ci = 'ss'\G
***** 1. row
'ß' collate utf8mb4_de_pb_0900_ai_ci = 'ss': 1
```

```
mysql> select 'ß' collate utf8mb4_de_pb_0900_as_cs = 'ss'\G
***** 1. row
'ß' collate utf8mb4_de_pb_0900_as_cs = 'ss': 0
```

Case folding

Language Dependent Case Folding

```
mysql> SELECT regexp_like( 'I', 'i' );  
→ 1
```

```
mysql> SELECT regexp_like( 'İ', 'i' );  
→ 0
```

```
mysql> SELECT regexp_like( 'I', 'ı' );  
→ 0
```

Beware of Conversion!

```
mysql> set names latin1;
mysql> create table t1 ( a char ( 10 ) );
mysql> insert into t1 values ( 'å' );
mysql> select a from t1\G
***** 1. row
a: å
mysql> select regexp_like( a, 'å' ) from t1\G
***** 1. row
regexp_like( a, 'å' ): 1
```

Beware of Conversion!

Use Hex Codes!

```
mysql> select hex( a ) from t1;
```

```
+-----+
```

```
| hex( a ) |
```

```
+-----+
```

```
| C383C2A5 |
```

```
+-----+
```

wait, what?!

Beware of Conversion!

Use Hex Codes!

```
mysql> select hex( a ) from t1;
```

```
+-----+
```

```
| hex( a ) |
```

```
+-----+
```

```
| C383C2A5 |
```

```
+-----+
```

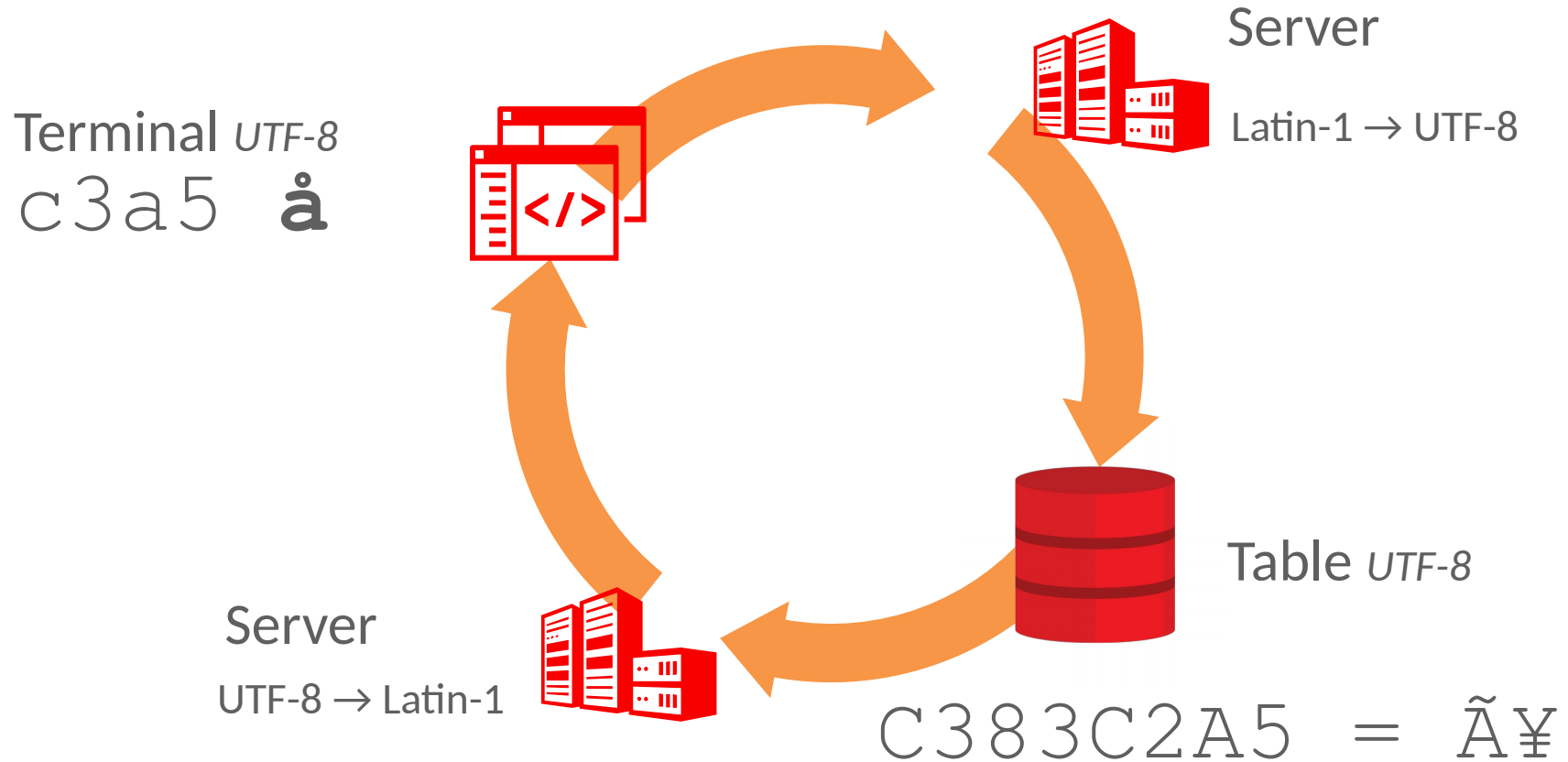
wait, what?!

å is encoded as:

Latin-1: 0x e5

UTF-8: 0x c3 a5

Conversion flow



Power Tip

Use Hex Codes and Character set Introducers!

```
mysql> set global character_set_client = utf8mb4;
```

```
mysql> select _utf8mb4 0xc3a5, _latin1 0xe5;
```

+-----+-----+	
_utf8mb4 0xc3a5	_latin1 0xe5
+-----+-----+	
å	å
+-----+-----+	

```
mysql> set global character_set_client = latin1;
```

```
mysql> select _utf8mb4 0xc3a5, _latin1 0xe5;
```

+-----+-----+	
_utf8mb4 0xc3a5	_latin1 0xe5
+-----+-----+	
å	å
+-----+-----+	

Questions?

ORACLE®