

MySQL日本語処理完全解説

住商情報システム株式会社
プラットフォームソリューション事業部門
IT基盤ソリューション事業部
オープンソースシステム部
玉川 修一

今日の内容

- MySQL概要
- 日本語処理対策
 - 問題点
 - 原因
 - 解決策

MySQL概要

MySQLとは

- オープンソースのデータベース

- スウェーデン発
- オープンソースのRDBMS
- 1日あたりの平均ダウンロード数5万
- 全世界で1000万インストール
(2006年4月現在)
- 普及度は世界No.1のOSS RDBMS

MySQLバージョン

- 4.0以下と4.1以上では日本語の取り扱いが異なる

- 5.1(beta)
- 5.0(GA)
- 4.1(GA)

- 4.0(old)
- 3.23(old)



GA: Generally Available

日本語処理の留意点

- いろいろあります

1. 文字化け対策基礎
2. クライアントライブラリ
3. (株)問題
4. Unicode変換ルール問題
5. ラウンドトリップ変換問題
6. Javaのキャラクターセット
7. 日本語メタデータ
8. 日本語全文検索

1.文字化け対策基礎

“MySQLで文字化けしてしまいます！”

- FAQ
- 特に ver.4.1のリリース以降
- 他のマルチバイトユーザーも経験

文字化けの例

- 4.1をインストールはしてみたものの...

```
mysql> CREATE TABLE t1(a CHAR(1))
      > DEFAULT CHARACTER SET = sjis;
Query OK, 0 rows affected (0.09 sec)
```

```
mysql> INSERT INTO t1 VALUES('あ');
Query OK, 1 row affected, 1 warning (0.05 sec)
```

```
mysql> SELECT * FROM t1;
```

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

```
| a |
```

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

```
| ? |
```

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

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



?

MySQLで日本語使える？

- はい、ちゃんと使えます。

- ほとんどの場合が「キャラクターセット」の設定ミスが原因
- 正しい設定を行えば、簡単に解決できる場合が多数

問題を理解する為に

- まずは基礎から解説

- キャラクターセットとは？
- 主なキャラクターセット
- MySQL 4.0以下 と MySQL 4.1以上の違い

MySQLのキャラクターセットとは？

- 文字集合 + 文字エンコーディング

- 文字集合
 - どんな文字が使えるか
- 文字エンコーディング
 - どうやって文字を表現するかというルール
 - sjis エンコーディングで「あ」という文字は「0x82A0」と表現
 - ujis エンコーディングで「あ」という文字は「0xA4A2」と表現
- MySQLのキャラクターセット
 - 文字集合と文字エンコーディングの組み合わせ
- 文字コード
 - あるキャラクターセットにおける文字の値
 - sjisキャラクターセットで、「あ」の文字コードは「0x82A0」

現在使える主なキャラクターセット

- 日本語が使えるのは6種類

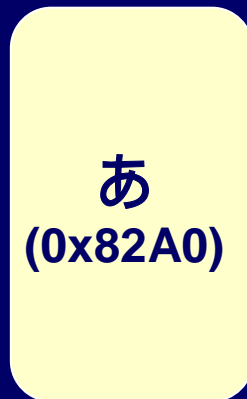
sjis	<ul style="list-style-type: none">■ 主にWindows向け■ ASCII文字、ひらがな、カタカナ、JIS第一、第二水準漢字等が使用可能
ujis	<ul style="list-style-type: none">■ 主にUnix/Linux向け■ sjisの文字集合に加えて、補助漢字等が使用可能
cp932	<ul style="list-style-type: none">■ sjisの文字集合に加え、NEC特殊文字等の外字が使用可能■ 文字コード体系はsjisとほぼ同等
eucjpms	<ul style="list-style-type: none">■ ujisの文字集合に加え、NEC特殊文字等の外字が使用可能■ 文字コード体系はujisとほぼ同等
utf8	<ul style="list-style-type: none">■ 各国言語が使用可能な国際文字集合■ Ver. 4.1以降で使用可能
ucs2	<ul style="list-style-type: none">■ 各国言語が使用可能な国際文字集合■ Ver. 4.1以降で使用可能
latin1	<ul style="list-style-type: none">■ MySQLのデフォルトキャラクターセット■ ASCII文字、アクセント付アルファベットのみ使用可能

MySQL 4.0以下の処理

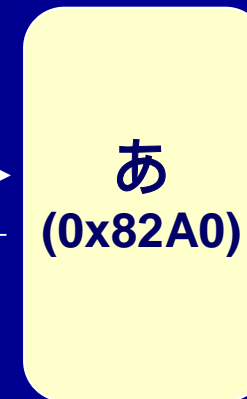
- 文字コードの変換はなし

- クライアント → サーバー
 - 文字コードをそのまま格納
- サーバー → クライアント
 - 文字コードをそのまま表示

クライアント



サーバー



MySQL 4.0以下の処理

- latin1でも日本語OK

- キャラクターセットにかかわらず、文字コードをそのまま送受信
- キャラクターセットがデフォルトのlatin1でも日本語の利用は(見かけ上*)可能だった
- キャラクターセットを変換する機能はない

*文字列長の認識や、ソート順序に影響有り

MySQL 4.1以上の処理

- 機能が強化され高度になりました

- テーブル(カラム)キャラクターセットの概念
- 文字コード変換機能 + Unicodeサポート
- 文字コードの自動変換機能
- キャラクターセット変数の追加

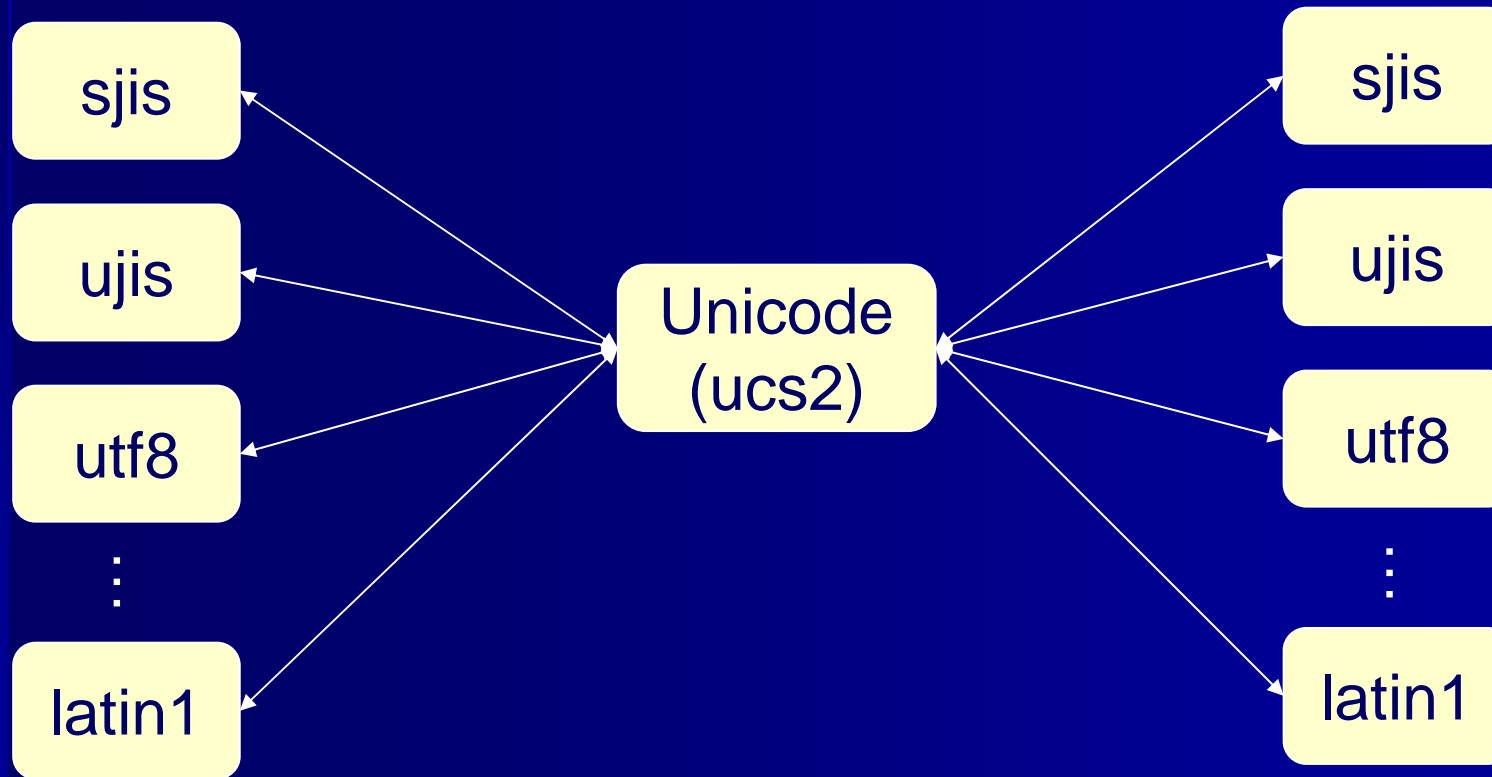
MySQL 4.1以上の処理

- テーブルキャラクターセットが指定できます

```
CREATE TABLE `table1` (  
  `column1` char(8) default NULL  
) ENGINE=MyISAM  
DEFAULT CHARSET=sjis
```

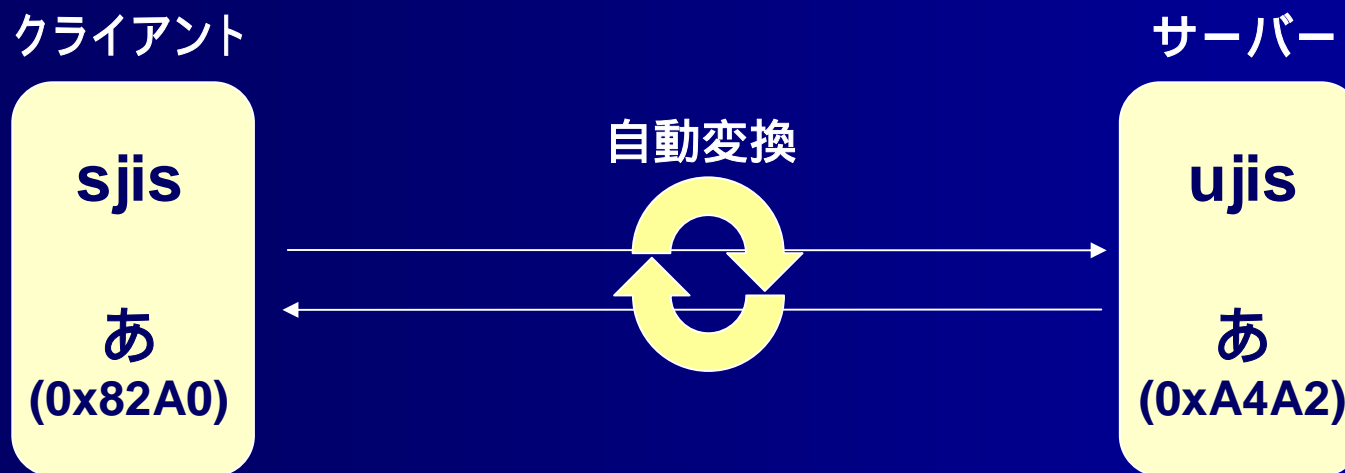
MySQL 4.1の処理

- Unicodeを介して変換できます



MySQL 4.1の処理

- 自動変換が行われます



MySQL 4.1の処理

- 変数が増えました

クライアントプログラム起動時の
--default-character-set
オプションが影響

character_set_client	クライアントから入力された文字の キャラクタセット	クライアント
character_set_connection	テーブルアクセスのないサーバ処理の キャラクタセット	
character_set_results	返された結果を表示するキャラクタセット	
character_set_server	CREATE DATABASE のデフォルト キャラクタセット	サーバー
character_set_database	CREATE TABLE のデフォルトキャラクタセット	

mysqld起動時の
--default-character-set
オプションが影響

文字化け発生のメカニズム

- こんなイメージです (4.1 <)

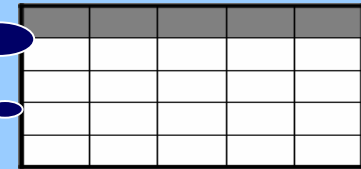
クライアント
(Windowsマシン)

「あ」をInsert
(0x82A0)

クライアントキャラクタセット変数
character_set_client = latin1
character_set_connection = latin1
character_set_result = latin1

入力された文は
'0x82'と'0xA0'の
2文字だな

送られてきた文字は
latin1の'0x82'と'0xA0'
か...。テーブルがsjis
なので、sjisに変換だ！



default character set = sjis

文字化け発生

あ -> ??

今から送る文字は
latin1の'0x82'と'0xA0'
だぞ！

文字化け発生の条件 (4.1 <)

- ポイントは変換と文字集合

- 文字コードの変換が発生
- Unicodeとのマッピングが定義されていない
 - 文字集合の範囲外の文字を使用
 - 変換元と変換先の文字集合の違い

対策 (4.1 <)

- ポイントは変換の回避

- 文字コードの変換を回避
 - クライアント/サーバー間でキャラクターセットを統一
- 変換元キャラクターセットと変換先キャラクターセットの互換性を理解して使用

対策(設定例)

- この様に設定すればOK

mysqlクライアントを使う限りではmy.cnf
ファイルで設定可能

```
[mysqld]
```

```
default-character-set = sjis
```

```
[mysql]
```

```
default-character-set = sjis
```

2.クライアントライブラリ

問題はクライアントライブラリ

- キャラクターセットは常にlatin1

- my.cnfでは設定済み
- mysqlクライアントでは問題なし
- PHP等のプログラムでは文字化け
- 問題はバイナリ配布のクライアントライブラリ
- キャラクターセットはlatin1でコンパイル
- my.cnfファイルの設定は影響しない

対策

- 方法は3通り

- ソースからコンパイル
`./configure --with-charset=character_set_name`
- 接続直後にSET NAMES
*character_set_name*の実行
- mysqld起動オプション
`--skip-character-set-client-handshake`
(MySQL 4.1.15, 5.0.13 ~)

対策

- この様に設定すればOK(例)

my.cnf ファイル

```
[mysqld]
```

```
default-character-set = sjis
```

```
skip-character-set-client-handshake
```

```
[mysql]
```

```
default-character-set =sjis
```

3.(株)問題

文字化けの例

- (株)が文字化け！

```
mysql> CREATE TABLE `table1` (`column1` char(8))
-> ENGINE=MyISAM
-> DEFAULT CHARSET=utf8;
Query OK, 0 rows affected (0.12 sec)
```

```
mysql> insert into table1 values('(株)');
Query OK, 1 row affected, 1 warning (0.00 sec)
```

```
mysql> select * from table1;
+-----+
| column1 |
+-----+
| ?      |
+-----+
```

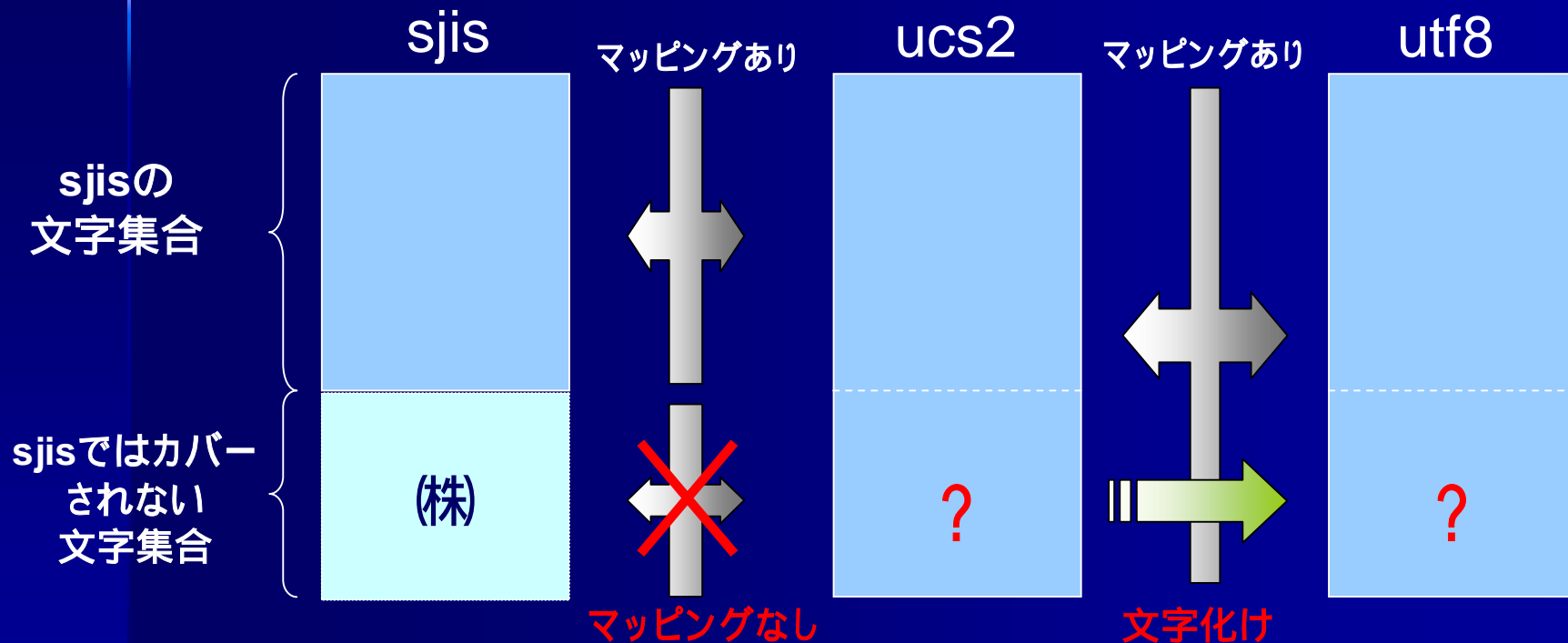
「(株)」問題

- 機種依存文字の文字化け

- Windows機種依存文字が化けする問題
 - NEC特殊文字
 - IBM拡張文字
 - NEC選定IBM拡張文字
- WindowsのシフトJISとMySQLのsjisの文字集合の違いが原因

イメージ図

- 変換マッピングルールがないのです



“シフトJIS” と “sjis”

- 実は違うんです

- MySQLの”sjis”はIANAの”Shift_JIS”
- Windowsの”シフトJIS”はWindows Code Page 932 (通称cp932)
- IANAではWindows31-Jに相当
- cp932/Windows31-Jの文字集合をカバーするキャラクターセットが必要

解決策は後ほど...

4.Unicode変換ルール問題

Unicode変換ルール問題

- 変換ルールも異なる”シフトJIS”

- この世には異なるシフトJISが存在
- それぞれ一部の文字についてUnicodeへの変換ルールが異なる
- 変換が異なる文字の例

\ ~ -

Unicode変換ルール問題

- Unicodeに変換すると違う文字に！

sjis

0x8191

(全角)

その他

0x8191

(全角)



ucs2

0x00A2

¢ (半角)

0xFFE0

(全角)

変換の結果
全角が半角に！

変換後も
全角のまま

解決策は後ほど...

5.ラウンドトリップ問題

ラウンドトリップ問題

- ディレクトリパスが変!?

```
C:\WINDOWS\system32\cmd.exe - mysql -u root -p --port=5015

mysql> show variables like '%dir';
+-----+-----+
| Variable_name | Value                               |
+-----+-----+
| basedir       | c:\mysql-5015\                     |
| character_sets_dir | c:\mysql-5015\share\charsets\     |
| datadir       | c:\mysql-5015\data\               |
| innodb_data_home_dir |                                     |
| innodb_log_arch_dir |                                     |
| innodb_log_group_home_dir | .\                                 |
| slave_load_tmpdir | C:\WINDOWS\TEMP\                  |
| tmpdir        |                                     |
+-----+-----+

8 rows in set (0.42 sec)

mysql>
```

ラウンドトリップ問題

- こんなイメージです



解決策

- cp932で3つまとめて全て解決

- 新キャラクターセットcp932を実装
- 3つの問題すべて解決
 - (株)問題
 - Unicode変換ルール問題
 - ラウンドトリップ問題
- ちなみにこれ、SCSがやりました

cp932とは

- sjisの拡張
- Windows機種依存文字をカバー
 - (株)問題を解決
- Windows方式のUnicode変換
 - Unicode変換ルール問題を解決
 - ラウンドトリップ問題を解決

sjisとcp932

- 文字集合が違います

%&¥ abc 123
アイ
あいう
アイウ
亜井宇

sjis

jisx0201

jisx0208

cp932

jisx0201

jisx0208

NEC特殊文字

IBM拡張文字

NEC選定IBM拡張文字

(株) 平城

續襲鎧

續襲鎧

sjisとcp932

- 変換ルールも違います

cp932

0x8191



ucs2

0xFFE0

0x5C

¥



0x005C

¥

0x815F

\



0xFF3C

\

eucjpms

- ujisで(株)を使いたい場合に

- ujisの拡張
- cp932互換
- ujisでWindows機種依存文字を格納する場合に使用
- 5.0以上で使用可能

詳細は日本MySQLパートナー会の記事参照
「cp932 eucjpms」でGoogle!

対処方法

- sjis/ujisよりもcp932/eucjpmsを

my.cnf ファイル

```
[mysqld]
```

```
default-character-set = cp932
```

```
skip-character-set-client-handshake
```

```
[mysql]
```

```
default-character-set = cp932
```

cp932: 4.1,5.0

eucjpms: 5.0

6.Javaの注意点

JavaとMySQLのキャラクターセット

- Connector/Jでマッピング

- キャラクターセットの名称はJavaとMySQLで異なる

- Connector/Jでマッピング

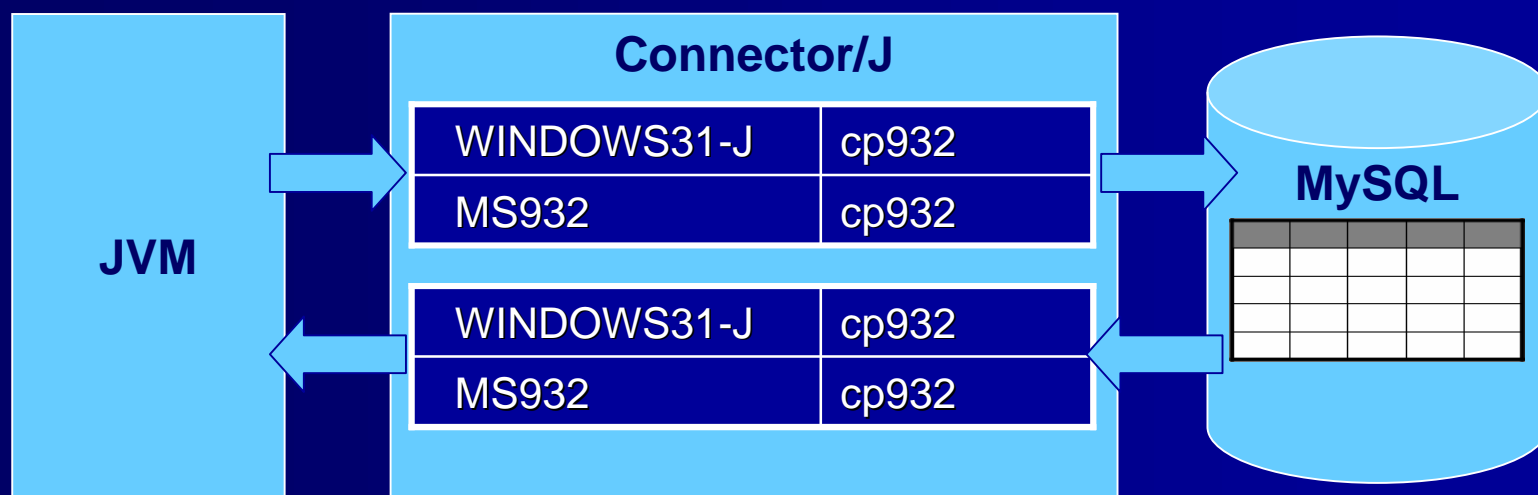
例)

(MySQL) cp932 = (Java) MS932

(MySQL) cp932 = (Java) Windows-31J

cp932使用時の問題点

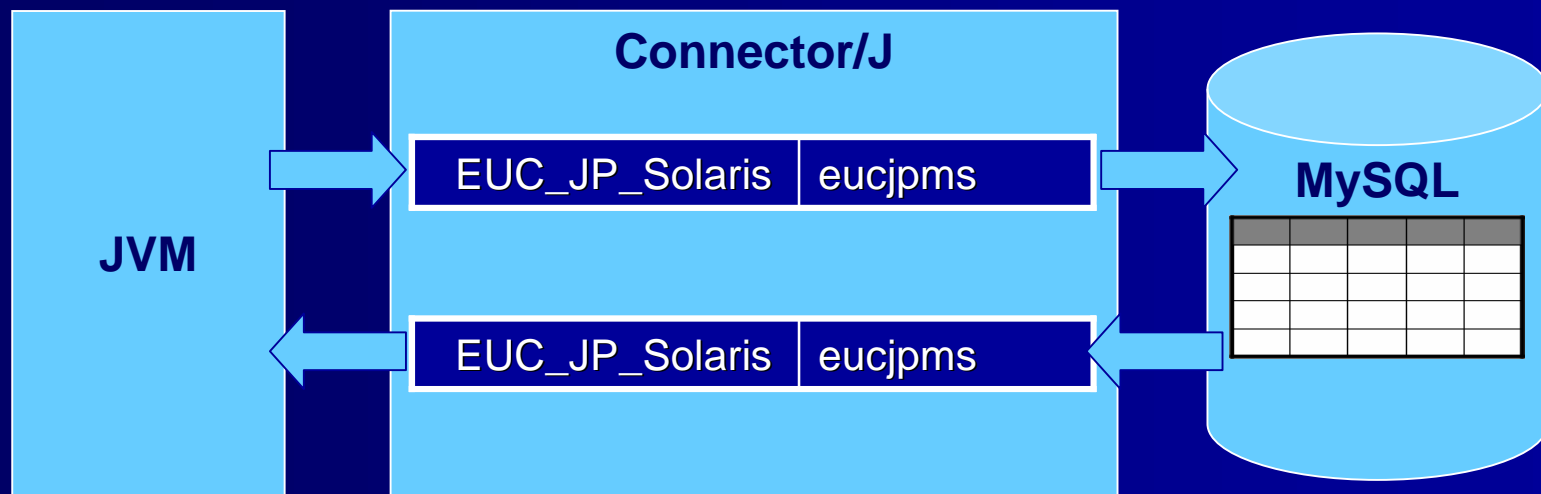
- Connector/Jがcp932に完全対応していなかった
- Windows-31J(MS932)を使用しても(株)を表示できない!?
- INSERTは出来るがSELECTはエラーに



eucjpms使用時の問題点

- eucjpmsとJavaのキャラクターセットは非完全互換

- eucjpmsは互換性のないEUC_JPにマッピング
- (株)が文字化け



EUC_JP_Solaris

- better solution (best solution)

- eucjpmsと完全互換のわけではない
- 互換性のない文字
 - |
- このような特殊記号のみ影響但し、(株)は使用可能
- 実はこれもSCSが実装しました

対策

- Javaでのまとめ

- cp932にはWINDOWS-31J / MS932を使用
- eucjpmsにはEUC_JP_Solarisを使用
 - 互換性のない文字について留意すること

Java側キャラクターセット	MySQL側キャラクターセット
WINDOWS-31J	cp932
MS932	cp932
EUC_JP_Solaris	eucjpms
SJIS	sjis
EUC_JP	ujis
UTF-8	utf8

(Connector/J ver. 3.1.9以上)

7.日本語テーブル名問題

日本語テーブル名問題

- データベース名 = ディレクトリ名
- テーブル名 = ファイル名
- 4.1以降、テーブル名、データベース名等のメタデータはutf8で保存
- データベース名やテーブル名に日本語を使用した場合に問題あり (bug#3906)

日本語テーブル名問題

```
C:\ コマンドプロンプト - mysql -u root -p --port=5016

mysql> create table `あ`(a int);
Query OK, 0 rows affected (0.06 sec)

mysql> create table `い`(a int);
ERROR 1050 (42S01): Table '繻・ already exists
mysql> create table `う`(a int);
ERROR 1050 (42S01): Table '繻・ already
mysql> create table `え`(a int);
ERROR 1050 (42S01): Table '繻・ already
mysql> create table `お`(a int);
ERROR 1050 (42S01): Table '繻・ already
mysql>
```

C:\mysql\mysql-5.0.19\data\test

名前	サイズ	種類
繻・frm	9 KB	ファイル
繻・MYD	0 KB	ファイル
繻・MYI	1 KB	ファイル

Utf8で保存された名前を使用してファイルを作成し、シフトJISで表示している為文字化けが発生

日本語テーブル名問題

```
C:\ コマンドプロンプト - mysql -u root -p --port=5107

mysql> create table `あ`(a int);
Query OK, 0 rows affected (0.05 sec)

mysql> create table `い`(a int);
Query OK, 0 rows affected (0.05 sec)

mysql> create table `う`(a int);
Query OK, 0 rows affected (0.05 sec)

mysql> create table `え`(a int);
Query OK, 0 rows affected (0.05 sec)

mysql> create table `お`(a int);
Query OK, 0 rows affected (0.03 sec)

mysql>
```

Ver. 5.1.6.にて修正

名前	サイズ	種類
@304a.frm	9 KB	FRM ファイル
@304a.MYD	0 KB	MYD ファイル
@304a.MYI	1 KB	MYI ファイル
@3042.frm	9 KB	FRM ファイル
@3042.MYD	0 KB	MYD ファイル
@3042.MYI	1 KB	MYI ファイル

@+Unicode
コードポイントで表現

対処方法

- 日本語メタデータの使用は控えましょう
- 5.1を使用するまでは、日本語メタデータを使用しない方が安全

8.日本語全文検索

MySQLで日本語全文検索

- 機能はあるが分かち書きが必要

- MyISAMストレージエンジンでは全文検索用インデックスをサポート
- 但し日本語のような言語には未対応
- 別途分かち書きを行ってやる必要有り

対処方法

- MySQL + Senna*という選択肢

- 組み込み型の全文検索エンジン
- MySQLに組み込み可
- SQLのみで全文検索が可能に

*未来検索ブラジルが開発した、オープンソースソフトウェア

MySQL+Senna

- 詳細はこちらで

- Open Source Pavilion
 - MySQL + Sennaを紹介
- ブース
 - MySQL 5.0 + Sennaのデモ。

まとめ

まとめ

- MySQLでの日本語処理対策

課題	対策
基本的な文字化け	キャラクターセットオプションの設定
クライアントライブラリ	1. --skip-client-character-set-handshake 2. SET NAMES キャラクターセット名 3. ./configure --with-charset=キャラクターセット名
(株)問題	キャラクターセットcp932の使用
Unicode変換ルール問題	
ラウンドトリップ問題	
JAVA	1. cp932に対してMS932/Windows-31Jを使用 2. EucjpmSに対してEUC_JP_Solarisを使用
日本語メタデータ	5.1まで原則使用しない
日本語全文検索	MySQL + Sennaを使用

MySQLならSCS

- MySQLオフィシャルトレーニング
- MySQLライセンス
- システム構築
 - 詳細は www.scs.co.jp/mysql

ご清聴ありがとうございました。

ご質問は mysql@scs.co.jp まで