TCTiSe - Text Compressed Time Series

Format version A4

Documentation version date 05.11.2015

Author: Makhmudov Evgeniy Reyzudinovich

email: john_16@list.ru, mer@emsd.ru

web site: http://emsd.ru/larmnow/tctise/en/main.html,

Python API: http://bitbucket.org/john_16/tctise

Content

General information, comments, reduce

TCTiSe (Text Compressed Time Series) - general-purpose data format for storing the values of the time series. As the name implies, a general idea laid a principle of working with numerical values. In TCTiSe values of time series are presented in the form of a text string, which is compressed algorithms such as gzip, bzip2, lzma. The format is the block structure, each block has a fixed length header and a dynamic part.

Abbreviations

ID — identification number

ASCII, UTF-8 - text encoding

UTC - Universal coordinate time

General remarks

Date / time always used in UTC.

Size of the blocks in their description is quoted with the identifier of the block type.

General technical notes

- Binary data types specified in the aspect of using estimates of 32 bit operating systems
- Strings in fixed field should be used ASCII characters
- Fixed-length strings are complemented by a space character to the left . For example, the field is string of 5 chars has a length 3 chars with value "ABC", resulting string should be " ABC"
- Literals in the values of the hash md5 sums must be lowercase, e.g. "83a36c"
- If the string value of the field is not defined, it must be replaced by a space character. For example, if a column is defined as a line of 5 chars , the value shall consist of five space characters "   "
- As a string delimiter of time series values to be used a carriage return \n corresponding byte 0x0A

Compression types

TCTiSe format involves to use modern, efficient , free of patent restrictions and available for widespread compression algorithms. Based on the criterion of the degree of compression, time packing and unpacking , were chosen algorithms gzip (deflate), bzip2, LZMA. Features each of the algorithms is significant and variations of these criteria , depending on the portion of the text provided . Therefore, there is no unambiguous judgments about which algorithm is the best indicator of the complex . However, bzip2 algorithm often has the best degree of compression , gzip is faster all , lzma occupies an intermediate position .

On the basis of numerous experiments on the time series to achieve the best degree of compression algorithm is recommended to use bzip2.

List of compatibility of types

Value types are used from the C programming language version 99.

Table 1. List of relevant data types in C and string of parameters.

Type	String of parameter	Size in bites
char	b	1
unsigned char	B	1
short	h	2
unsigned short	H	2
integer	i	4
unsigned integer	I	4
long	l	4
unsigned long	L	4
long long int	q	8
unsigned long long int	Q	8
float	f	4
double	d	8

Optimizing fragment of the time series block

Before packaging in the time-series fragment of block DATA, he must go optimization to increase the compression ratio. The first value is the reference and written in its original form, each next value should be recorded as the difference between the current and previous value of fragment. For example, if a piece of data to be packaged in a DATA chunk is as follows:

256 259 261 264 265 266 265 264 261 259

after optimization should be as follows.

256 3 2 3 1 1 -1 -1 -2 -3

Hash identifier of block DATA

Field "Hash ID" is an indicator , which is the last 6 characters hexadecimal representation of the MD5 hash sum of string values fields of block DATA. The identifier is used to identify blocks DATA with identical parameters, if the same combination of names of the main channel and the network station , but differ in other parameters of the block.
For correct calculation of the field, parameters of block must be submitted as a string in order : format version, byte order, station, channel, network, type of sampling, value of sampling, compression method, type of value.

For example, if you have the following set of parameters :

Format version	A4
Byte order	>
Station	KLY
Channel	SHZ
Network	SN5
Mantissa of sampling	1

Power of sampling	2
Compression method	b
Type of value	i

You have the following string 'A4> KLY SHZ SN512bi', MD5 sum which is equal to '934044e6b2f4f370efe94c22f4844b42', the last 6 characters of this amount will be field "Hash Id" equal '844b42 '.

Formation of sampling value

Sampling value encoded in the block in the form of two fields "Mantissa of sampling" and "Power of sampling" expressing the computer representation of real numbers of the form M · 10^p, Where M is the mantissa and p is the degree.

If the mantissa is positive, it is the sampling rate, if negative, it represents the number of milliseconds between the reports.

The mantissa must not be a multiple of 10 , an excessive degree should be moved in the "Power of sampling" (see Table 2) .

Table 2. Examples of possible combinations of values.

Sampling value	Mantissa	Power
100 Hz	1	2
500 ms	-5	2
7.8125 ms	-78125	-4
44.1 kHz	441	2
1 ms	-1	0
0.5 Hz	5	-1

Blocks description

Block DATA

The main block contains a set of parameters for the correct interpretation of time series values, checking of the timestamp and block number, packed optimized values of the time series. Size of fixed portion of block is 69 byte.

Field name	Type of data	Possible values	Description
ID	string, 10 chars	“TCTISEDATA”	String value with the fixed length which mean Block identificator.
Format version	string, 2 chars	“A4”	Version of TCTiSe format. First char is letter with major version, second char is digit with minor version.
Hash ID	string, 6 chars		Last 6 chars of MD5 hash sum of block parameters. (For details see chapter “Hash identifier of block DATA”)
Byte order	string, 1char	“>”,“<”	Byte order of binary data, little-endian is “<”, big-endian is “>”
Station	string, 7 chars		Name of station, for example “ KLY”
Channel	string, 7 chars		Name of channel, for example “ SHZ”
Network	string, 5 chars		Name of network, for example “ N1”
ID global	uint		Block number from the moment of start registration
ID channel	uint		Block number of channel from the moment of start registration
Datetime	double		Datetime stamp of beginning data in block, is a count of seconds from 1970 year. Is return value of C functions time()
Mantissa of sampling	int		Mantissa of sampling value. If value is positive it is a sampling frequency, if

			negative it is a count of milliseconds between samples
Power of sampling	char		Power degree of ten in sampling value
Compression method	string, 1char	“b”, “g”, “l”	Compression method, for example char b is bzip2 algorithm (See details in chapter “Compression types”)
Type of values	string, 1 char	See details in chapter “List of compatibility of types”	Type of packed value. How interpretate unpaked string values in C data types.
Number of values	uint		Number of values in packed string
Data length	uint		Length of packed string
Data	string		String with unfixed size, included packed optimizied time series

Block CUST

Block-extension, structure is determined freely, while respecting the basic fields. Block can be used for secondary storage of recorded data, notification of failures occurring, etc. In the "Length" field uses big-endian (">") byte order.

Field name	**Type of data**	**Possible values**	**Description**
ID	string, 10 chars	“TCTISECUST”	String value with the fixed length which mean Block identificator.
Extension id	string, 32 chars		A unique identifier that identifies the type CUSTOM block should be used as a key when registering the user application to retrieve the data structure of the block. It is recommended to use the md5 hash of a string which widely describe purpose of the block

Length	uint		Content length of a block in bytes

List of registered blocks of extension

Custom blocks can be officially registered and entered into the documentation. To do this, visit the official site in the "Contacts" section.

bedf076edfc306dd3f4bb3995a8ce2a7

Extension with identification description "Text message", intended for storing text information. Structurally, the entire volume of the block is a string in UTF-8.