TCTiSe - Text Compressed Time Series

Версия формата A4

Версия документации от 05.11.2015

Автор: Махмудов Евгений Рейзудинович

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

Оглавление

Общие сведения, замечания, сокращения

TCTiSe (от Text Compressed Time Series в переводе "сжатые текстовые временные серии", произносится как ТиСиТайз) - формат данных общего назначения для хранения значений временных рядов. Как видно из названия, в общую идею заложен иной принцип работы с числовыми значениями. В TCTiSe значения временного ряда представляются в виде текстовой строки, которая сжимается такими алгоритмами как gzip, bzip2, lzma. Формат имеет блочную структуру, каждый блок имеет заголовок фиксированной длины и динамическую часть.

Принятые сокращения

ID - идентификационный номер

ASCII, UTF-8 – текстовые кодировки

UTC – Всемирное координатное время

Общие замечания

- Дата/время всегда используются в UTC
- Размер блоков в их описании указан с учетом идентификатора типа блока.

Общие технические замечания

- Бинарные типы данных указаны в аспекте использования исчислений 32 разрядных операционных систем
- В строках фиксированных полей блоков должны использоваться символы ASCII
- Строки фиксированной длины дополняются слева символами пробела. Например, поле имеет тип строка 5c, значение имеет длину 3c "ABC", в результате в поле записывается строка " ABC"
- Литералы в значениях хеш сумм md5 должны быть строчными, например "83a36c"

- Если строковое значение поля не определено, то его нужно заменить символами пробела. Например, если поле задано как строка 5c, то значение должно состоять из пяти символов пробелов "   "
- В качестве строкового разделителя значений временного ряда должен использоваться символ переноса строки \n соответствующий байту 0x0A

Типы сжатия

Формат TCTiSe предполагает использование современных, эффективных, свободных от патентных ограничений и доступных для широкого использования алгоритмов сжатия. Исходя из критерия степени сжатия, времени упаковки и распаковки, такими алгоритмами были выбраны gzip(deflate), bzip2, LZMA. Особенность каждого из алгоритмов заключается в существенных вариациях этих критериев в зависимости от предоставляемого фрагмента текста. Поэтому не существует однозначного суждения о том, какой алгоритм имеет комплексно лучший показатель. Тем не менее, алгоритм bzip2 чаще всего имеет лучший показатель степени сжатия, gzip работает быстрее всех, lzma занимает промежуточное положение.

Исходя из многочисленных экспериментов на временных, рядах для достижения лучшего показателя степени сжатия рекомендуется использовать алгоритм bzip2.

Список совместимых типов

Типы значений используются из языка программирования C версии 99.

Таблица 1. Список соответствия типов данных в C и строки параметров.

Тип	**Строка параметра**	**Размер, байт**
char	b	1
unsigned char	B	1
short	h	2
unsigned short	H	2
int	i	4
unsigned int	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

Оптимизация блочного фрагмента временного ряда

Перед упаковкой фрагмента временного ряда в блок DATA он должен пройти оптимизацию с целью увеличения степени сжатия. Первое значение является опорным и записывается в исходном виде, каждое следующее значение должно записываться как разность текущего и предыдущего значения фрагмента ряда.

Например, если фрагмент данных подлежащий упаковке в блок DATA выглядит следующим образом:

256, 259, 261, 264, 265, 266, 265, 264, 261, 259

, то после оптимизации должен выглядеть следующим образом.

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

Хеш идентификатор блока DATA

Поле "*Hash ID*" это индикатор, представляющий собой последние 6 символов шестнадцатеричного представления MD5 хэш суммы строковых значений полей блока DATA. Идентификатор предназначен для выявления блоков DATA с идентичными параметрами, в случае если совпадает основное сочетание названий станции канала и сети, но отличаются другие параметры блока.

Для корректного расчета поля необходимо представить в виде строки параметры блока в следующей последовательности*: format version*, *byte order*, *station*, *channel*, *network*, *type of sampling*, *value of sampling*, *compression method*, *type of value*.

Например, если имеется следующий набор параметров:

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

, то получится следующая строка 'A4> KLY SHZ SN512bi', MD5 сумма которой будет равна ' 934044e6b2f4f370efe94c22f4844b42', 6 последних символов этой суммы и будут составлять поле “*Hash Id*” равным '844b42'.

Формирование значения дискретизации

Значение дискретизации закодировано в блоке в виде двух полей “*Mantissa of sampling*” и “*Power of sampling*” выражающих компьютерное представление вещественных чисел вида $M \cdot 10^p$, Где *M* это мантисса и *p* это степень.

Если значение мантиссы положительно, то оно является частотой дискретизации, если отрицательное, то выражает количество миллисекунд между отчетами.

Значение мантиссы не должно быть кратко 10, излишняя степень должна быть перенесена в поле “*Power of sampling*” (см. таблицу 2).

Таблица 2. Примеры возможных комбинаций значений.

Значение дискретизации	Мантисса	Порядок
100 Гц	1	2
500 мс	-5	2
7,8125 мс	-78125	-4
44.1 кГц	441	2
1мс	-1	0
0,5 Гц	5	-1

Описание и структура блоков

Блок DATA

Основной блок содержит в себе набор параметров для корректной интерпретации значений временного ряда, контрольные метки времени и номера блока, упакованные значения оптимизированного временного ряда. Размер фиксированной части блока 69 байт.

Имя поля	Тип данных	Возможные значения	Описание
ID	строка 10c	“TCTISEDATA”	Строка с фиксированным значением. Идентификатор блока.
Format version	строка 2c	“A4”	Версия формата TCTiSe, первый символ буквенный является мажорной версией, второй символ цифровой – минорной версией.
Hash ID	строка 6c		Последние 6 символов MD5 хэш суммы параметров блока (Подробнее см.раздел “Хеш идентификатор блока DATA”)
Byte order	строка 1c	“<”, “>”	Порядок следования байтов бинарных данных, little-endian соответствует "<", big-endian соответствует ">". Рекомендуется использовать big-endian порядок.
Station	строка 7c		Название станции регистрации, например " KLY"
Channel	строка 7c		Название канала регистрации, например " WINDS"
Network	строка 5c		Название сети регистрации, например " N1"
ID global	uint		Последовательный номер блока от момента начала аппаратной регистрации
ID channel	uint		Последовательный номер блока для канала от момента начала аппаратной регистрации

Datetime	double		Отметка даты и времени начала блока, выраженная дробным числом секунд с начала 1970 года, является возвращаемым значением функции C time()
Mantissa of sampling	int		Мантисса значения дискретизации. Если значение положительно, то оно является частотой дискретизации, если отрицательно, то выражает количество миллисекунд между отчетами.
Power of sampling	char		Порядок десятичной степени значения дискретизации.
Compression method	строка 1c	“b”, “g”, “l”	Метод сжатия, например алгоритму bzip2 соответствует “b” (Подробнее см. раздел «Типы сжатия»)
Type of value	строка 1c	см.раздел список совместимых типов	Тип упаковываемого значения, т.е. Как интерпретировать текстовые распакованные значения в типах данных C
Number of values	uint		Количество значений в строке данных
Data length	uint		Длина запакованной строки с данными
Data	строка		Строка переменной длины, содержащая запакованный оптимизированный временной ряд

Блок CUST

Блок-расширение, структура определяется свободно, при соблюдении базовых полей. Блок может быть использован для хранения второстепенных регистрируемых данных, являться уведомлением о происходящих сбоях и т.п. В поле “*Length*” используется big-endian (“>”) порядок байтов.

Имя поля	Тип данных	Значение	Описание
ID	строка 10c	“TCTISECUST”	Строка с фиксированным значением. Идентификатор блока
Extension id	строка 32c		Уникальный идентификатор, определяющий тип CUSTOM блока, должен использоваться в качестве ключа при регистрации в пользовательском приложении для извлечения структуры данных из блока. Рекомендуется использовать md5 хеш сумму строки подробно описывающей предназначение блока
Length	uint		Размер содержания расширения в байтах

Список зарегистрированных блоков расширений

Пользовательские блоки могут быть официально зарегистрированы и внесены в документацию. Для этого нужно обратиться на официальный сайт в раздел «Контакты».

bedf076edfc306dd3f4bb3995a8ce2a7

Расширение с идентификационным описанием “Text message”, предназначено для хранения текстовой информации. Структурно весь объем блока представляет собой строку в кодировке UTF-8.