Data Encoding: Experiments and Results

Aditya Avinash
Kian Win ong
Gaurav Saxena

Encoding Techniques

- Variable length encoding for integer and long
- UTF-8 encoding for Strings

These techniques are used in Google protocol buffer

Variable length integer encoding

Value Range	Number of bytes used
0 - 127 (2^7 -1)	1
128- 16383 (2^14 - 1)	2

Encoding String using UTF-8

- Variable length encoding
- Size 1 to 6 bytes. Rarely goes to 6
- ASCII characters are one byte
-Encoded value : [length as varint][each character stored as UTF-8 characters in byte stream]

Example: String : “testing”

encoded value: 07 74 65 73 74 69 6e 67

Encoding space efficiency results: (Experimental data)

Schema:

```
{{
    {
        name: String
        id      : Integer
        email : String
    }
}}
```

Data:

```
{{
    {
        name : “Aditya”
        id          : 1
        email : “aavinash”
    }
 …… 600 tuples
}}
```

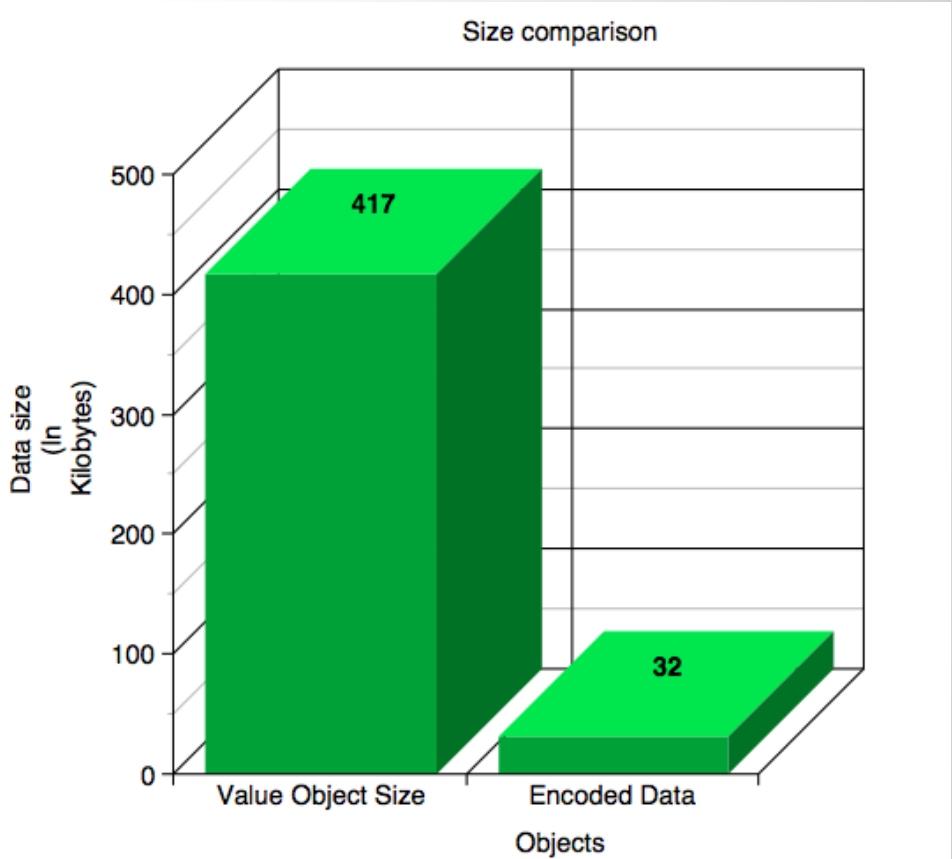
Results

Size of Value object:
417 KB

Size of encoded data :
32 KB

Encoded data is about **13 times** smaller than the value object

*encoded data verified



Results with Real data :BioHeatMap data

Schema:

{{
{article_count : integer, ...
{{
{report_item_id :.....

Data:

{{
{article_count : 929, ...
{{
{report_item_id : 37,
....

Results

Size of Value object:
555 KB

Size of encoded data :
21.8 KB

Encoded data is about **25.49 times** smaller than the value object

*encoded data could not be verified