

Hadoop

Форматы хранения данных

Типы хранения данных

- * txt (обычный текстовый файл)
- * SequenceFile (двоичный типа key-value)
- * Avro (двоичный от Дуга Кашинга)
- * Protobuf (двоичный от Google)
- * ORC (колоночный)

Текстовый формат txt

* Плюсы:

* простой

* понятный

* хорошо сжимается

* легко обрабатывать на любом ЯП

Текстовый формат txt

- * Минусы:

- * неэффективное использование диска

- * нет схемы

- * нет сериализации (надо делать)

- * Кодировки

SequenceFile

- * Самый первый двоичный формат
- * Храним все в виде ключ - значение
- * Поддержка компрессии
 - * уровня блока
 - * уровня файла

SequenceFile

структура

- * Header
- * Record
 - * Record length
 - * Key length
 - * Key
 - * (Compressed?) Value
 - * A sync-marker every few k bytes or so.

SequenceFile

структура header

- * version - SEQ4 или SEQ6
- * keyClassName - класс для ключа
- * valueClassName - класс для значения
- * compression - флаг компресии
- * blockCompression - флаг блочной компресии
- * compressor class - кодак для компресии
- * metadata - метаданные
- * sync - маркер

Avro

- * Двоичный
- * Есть схема
- * Есть стандартная сериализация
- * Хорошо сжимается
- * Высокая производительность
- * Поддержка большинства ЯП
- * Самоописательный

Avro schema

```
{  
  "namespace": "ru.mail.avro",  
  "name": "SvdUid",  
  "type": "record",  
  "fields": [  
    {  
      "name": "uid_type",  
      "type": "string"  
    },  
    {  
      "name": "uid",  
      "type": "string"  
    }  
  ]  
}
```


Анго типы данных

- * `null`: пусто
- * `boolean`: двоичное
- * `int`: 32-bit целое
- * `long`: 64-bit целое
- * `float`: 32-bit число с плавающей точкой
- * `double`: 64-bit число с плавающей точкой
- * `bytes`: массив байт
- * `string`: строка в юникоде

Avro типы данных

- * records

- * enum

- * arrays

- * maps

- * unions

- * fixed

Avro enum

```
{  
  "doc": "User type",  
  "name": "user_type",  
  "type": {  
    "name": "EXPRESS_USER_TYPES",  
    "type": "enum",  
    "symbols": [  
      "_1POSITIVE",  
      "_2NEGATIVE"  
    ],  
    1  
  },  
  "order": "ascending"  
}
```


Avro array

```
{
```

```
  "doc": "Features",
```

```
  "name": "features",
```

```
  "type": {"type": "array", "items": "double"},
```

```
  "order": "ignore"
```

```
}
```


Avro map

```
{  
  "name": "step",  
  "type": {  
    "type": "map",  
    "values": "string"  
  }  
}
```


Avro union

```
{
```

```
  "name": "photo_big",
```

```
  "type": ["string", "null"]
```

```
}
```


Avro schema plugin

<plugin>

<groupId>org.apache.avro</groupId>

<artifactId>avro-maven-plugin</artifactId>

<version>\${org.apache.avro.cdh.version}</version>

<executions>

<execution>

<phase>generate-sources</phase>

<goals>

<goal>schema</goal>

</goals>

Avro schema plugin

<configuration>

<sourceDirectory>\${project.basedir}/src/main/avro/</sourceDirectory>

<outputDirectory>\${project.basedir}/target/generated-sources/java/</outputDirectory>

<imports>

<import>\${project.basedir}/src/main/avro/session.avsc</import>

/imports>

</configuration>

</execution>

</executions>

</plugin>

Avro internals

```
dm — a.pilipenko@rbhp74:~ — ssh -A a.pilipenko@rbhp74-ext.rbdev.mail.ru — 117x28
...a.pilipenko@rbhp74-ext.rbdev.mail.ru  ...lipenko@rbhp74:tmp/mining — -bash ...  ~/projects/dm — -bash  ~/projects/dm — -bash  ...p — a.pilipenko@rbhp74:~ — -bash ... +
a.pilipenko@rbhp74:~$ head titles.avro
Objavro.schema?{"type":"record","name":"KeyValuePair","namespace":"org.apache.avro.mapreduce","doc":"A key/value pair
","fields":[{"name":"key","type":{"type":"record","name":"User","namespace":"ru.mail.avro","fields":[{"name":"vid","t
ype":"string"}, {"name":"okid","type":["null","string"],"default":null}, {"name":"email","type":["null","string"],"defa
ult":null}, {"name":"vkid","type":["null","string"],"default":null}, {"name":"category","type":["null","int"],"doc":"Ca
tegories are positive integers (1, 2, 3) by convention","default":null}, {"name":"start","type":["null","long"],"doc":
"Unix time in seconds SINCE which the user is valid","default":null}, {"name":"end","type":["null","long"],"doc":"Unix
time in seconds UNTIL which the user is valid","default":null}, {"name":"mmid","type":["null","string"],"doc":"ID fro
m Moy mir <https://my.mail.ru/>","default":null}]},"doc":"The key"}, {"name":"value","type":{"type":"record","name":"S
parseFeatureVector","namespace":"ru.mail.avro","fields":[{"name":"features","type":{"type":"array","items":{"type":"r
ecord","name":"Feature","fields":[{"name":"feature_id","type":"string"}, {"name":"value","type":["double","null"]}, {"n
ame":"timestamps","type":[{"type":"array","items":"int"},"null"]}]}}]},"doc":"The value"}}}avro.codenull"[?2???({?R
?C?0???100003157415171209130kG1TYKcdh3bUNBdSwhgwXw==203476925&2159496962638519427"t:ofisn??Q???t:a??Q???$t:obrazovani
i(esli??Q???
t:znak??Q???t:kitel??Q???t:stranic??Q???t:raspolozen??Q???t:forum??Q???t:kompan??Q???t:#NUM#??Q???
t:form?
?Q???
t:voen??Q???t:nagrudn??Q???t:raspolaga??Q???
t:mvd??Q???t:sotrudnik??Q???t:znack??Q???1000080c362985255343248109040
t:#URL_SHORT#0 K?t:novost0 K?
t:pocht0 K?
t:igr0 K?t:poisk0 K?t:internet0 K?10001175t:tv?/?b???t:#URL_SHORT#/?/?b???t:zavtr?/?b???t:moskv?/?b???t:nedel?/?
b???t:teleperedac?/?b???t:kanal?/?b???t:novostj?6?i?t:vkluclj?6?i??
t:poln?/?b???t:peredac?/?b???t:teleprogramm?
/?b???t:programm?/?b???100040118
t:tehnicesk0 K?t:podderzk0 K?t:klient-bank0 K?t:sistem0 K?t:internet0 K?100048c020TvbxiudH2ZrUNBdSwhgwXw==428627
228(10889783474718379551
t:#URL_SHORT#q?t:novostq?
```


Avro internals

```
dm — a.pilipenko@rbhp74:~ — ssh -A a.pilipenko@rbhp74-ext.rbdev.mail.ru — 117x28
...a.pilipenko@rbhp74-ext.rbdev.mail.ru  ...lipenko@rbhp74:/tmp/mining — -bash ...  ~/projects/dm — -bash  ~/projects/dm — -bash  ...p — a.pilipenko@rbhp74:~ — -bash ... +
a.pilipenko@rbhp74:~$ hadoop fs -text /data/dm/datasets/2017-12-13/titles/part-r-00044.avro | head
{"key":{"vid":"100003157","okid":{"string":"41517120913"},"email":{"string":"kG1TYKcdh3bUNBdSwHgwXw=="},"vkid":{"string":"203476925"},"category":{"int":0},"start":null,"end":null,"mmid":{"string":"2159496962638519427"}}, "value":{"features":[{"feature_id":"t:ofisn","value":{"double":0.165},"timestamps":null}, {"feature_id":"t:a","value":{"double":0.165},"timestamps":null}, {"feature_id":"t:obrazovani(esli","value":{"double":0.165},"timestamps":null}, {"feature_id":"t:znak","value":{"double":0.165},"timestamps":null}, {"feature_id":"t:kitel","value":{"double":0.165},"timestamps":null}, {"feature_id":"t:stranic","value":{"double":0.165},"timestamps":null}, {"feature_id":"t:raspolozen","value":{"double":0.165},"timestamps":null}, {"feature_id":"t:forum","value":{"double":0.165},"timestamps":null}, {"feature_id":"t:komp an","value":{"double":0.165},"timestamps":null}, {"feature_id":"t:#NUM#","value":{"double":0.165},"timestamps":null}, {"feature_id":"t:form","value":{"double":0.165},"timestamps":null}, {"feature_id":"t:voen","value":{"double":0.165},"timestamps":null}, {"feature_id":"t:nagrudn","value":{"double":0.165},"timestamps":null}, {"feature_id":"t:raspolaga","value":{"double":0.165},"timestamps":null}, {"feature_id":"t:mvd","value":{"double":0.165},"timestamps":null}, {"feature_id":"t:sotrudnik","value":{"double":0.165},"timestamps":null}, {"feature_id":"t:znack","value":{"double":0.165},"timestamps":null}]]}}
{"key":{"vid":"1000080c","okid":{"string":"362985255343"},"email":null,"vkid":{"string":"248109040"},"category":{"int":0},"start":null,"end":null,"mmid":null}, "value":{"features":[{"feature_id":"t:#URL_SHORT#","value":{"double":8.333333333333334E-4},"timestamps":null}, {"feature_id":"t:novost","value":{"double":8.333333333333334E-4},"timestamps":null}, {"feature_id":"t:poct","value":{"double":8.333333333333334E-4},"timestamps":null}, {"feature_id":"t:igr","value":{"double":8.333333333333334E-4},"timestamps":null}, {"feature_id":"t:poisk","value":{"double":8.333333333333334E-4},"timestamps":null}, {"feature_id":"t:internet","value":{"double":8.333333333333334E-4},"timestamps":null}]]}}
{"key":{"vid":"10001175f","okid":null,"email":null,"vkid":null,"category":{"int":0},"start":null,"end":null,"mmid":null}, "value":{"features":[{"feature_id":"t:tv","value":{"double":0.30916666666666665},"timestamps":null}, {"feature_id":"t:#URL_SHORT#","value":{"double":0.30916666666666665},"timestamps":null}, {"feature_id":"t:zavtr","value":{"double":0.30916666666666665},"timestamps":null}, {"feature_id":"t:moskv","value":{"double":0.30916666666666665},"timestamps":null}, {"feature_id":"t:nedel","value":{"double":0.30916666666666665},"timestamps":null}, {"feature_id":"t:teleperedac","value":{"double":0.30916666666666665},"timestamps":null}, {"feature_id":"t:kanal","value":{"double":0.30916666666666665},"timestamps":null}, {"feature_id":"t:novost","value":{"double":0.30333333333333334},"timestamps":null}, {"feature_id":"t:vklu","value":{"double":0.30333333333333334},"timestamps":null}, {"feature_id":"t:poln","value":{"double":0.30916666666666665},"timestamps":null}]]}}
```


Protobuf

(Protocol Buffers)

- * От Google
- * Нужен компилятор (надо скачать *)
- * Есть схема
- * Нет схемы при сериализации *

Protobuf schema

```
syntax="proto2";  
package ru.mail.proto;  
option java_package = "ru.mail.proto";  
option java_outer_classname = "GeoProto";  
message LatLonMsg {  
    optional double lat = 1;  
    optional double lon = 2;  
}
```


Protobuf

типы данных

- * double
- * float
- * int32
- * int64
- * uint32
- * uint64
- * sint32
- * sint64
- * fixed32
- * fixed64
- * sfixed32
- * sfixed64
- * bool
- * string
- * bytes

Protobuf

Серилизация

- * Magic (несколько байт как маркер)
- * Тип сообщения (int - id в репозитории)
- * Длина сообщения (int - в байтах)
- * Тело сообщения

Protobuf recap

- * Хорош для сериализации маленьких сообщений
- * Неплохо сжимает
- * Неплохая производительность
- * Нет стандартного механизма описания
- * как следствие - надо писать свой InputFormat

ORC

- * Колоночный
- * Простая интеграция с Hive, Spark
- * Оптимизация сериализации комплексных типов
- * Можно сплитить без полного сканирования
- * Можно эффективно сливать файлы
- * Регулируемые параметры потребления памяти для чтения и записи

ORC

ТИПЫ ДАННЫХ

* Integer

- * boolean (1 bit)
- * tinyint (8 bit)
- * smallint (16 bit)
- * int (32 bit)
- * bigint (64 bit)

* Floating point

* float

* double

* String types

- * string
- * char
- * varchar

* Binary blobs

* binary

* Date/time

* timestamp

* date

* Compound types

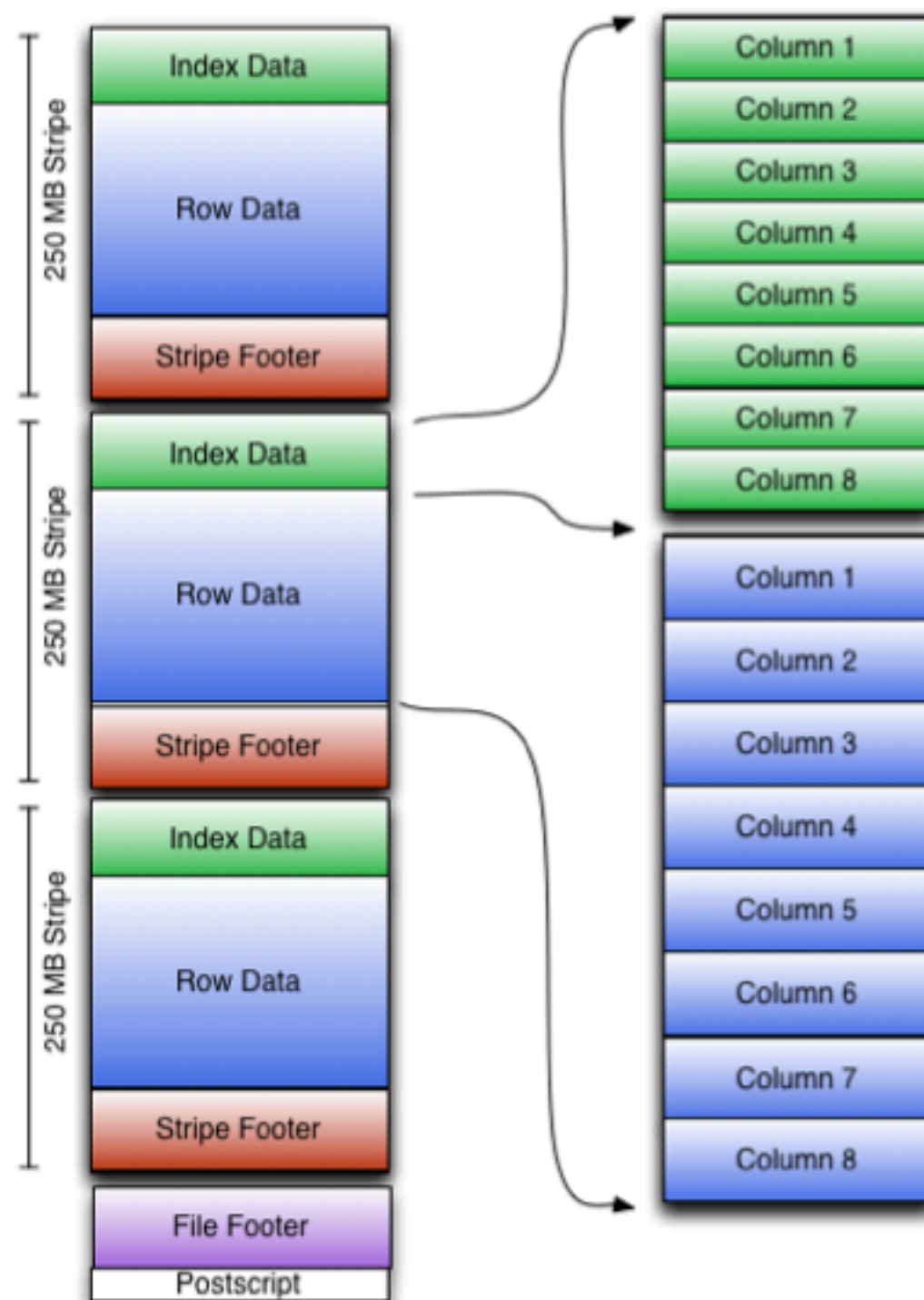
* struct

* list

* map

* union

ORC структура



ORC postscript

```
message PostScript {
```

```
// the length of the footer section in bytes
```

```
optional uint64 footerLength = 1;
```

```
// the kind of generic compression used
```

```
optional CompressionKind compression = 2;
```

```
// the maximum size of each compression chunk
```

```
optional uint64 compressionBlockSize = 3;
```

```
// the version of the writer
```

```
repeated uint32 version = 4 [packed = true];
```

```
// the length of the metadata section in bytes
```

```
optional uint64 metadataLength = 5;
```

```
// the fixed string "ORC"
```

```
optional string magic = 8000;
```

```
}
```

```
enum CompressionKind {
```

```
NONE = 0;
```

```
ZLIB = 1;
```

```
SNAPPY = 2;
```

```
LZO = 3;
```

```
LZ4 = 4;
```

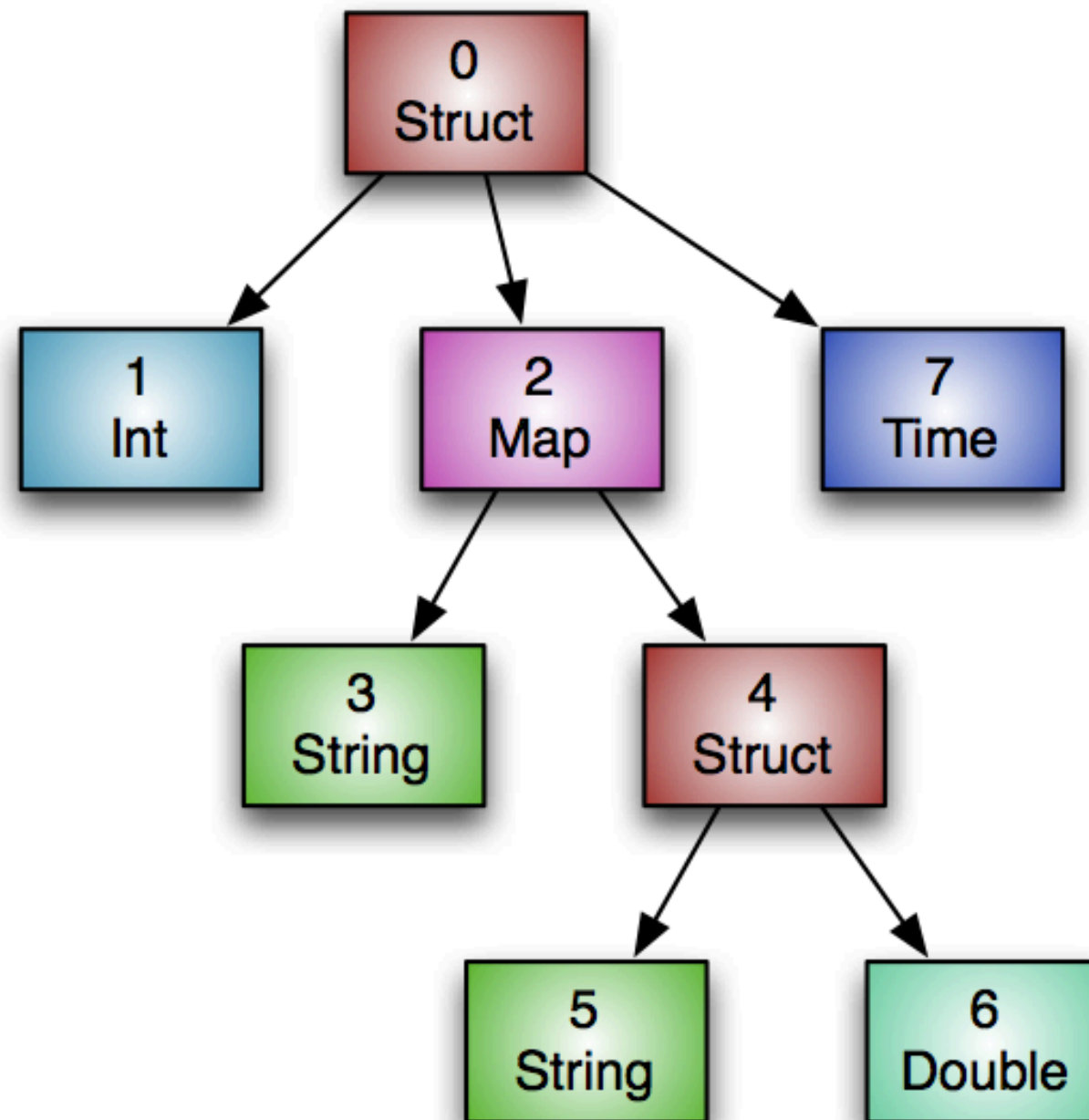
```
ZSTD = 5;
```

```
}
```


ORC file footer

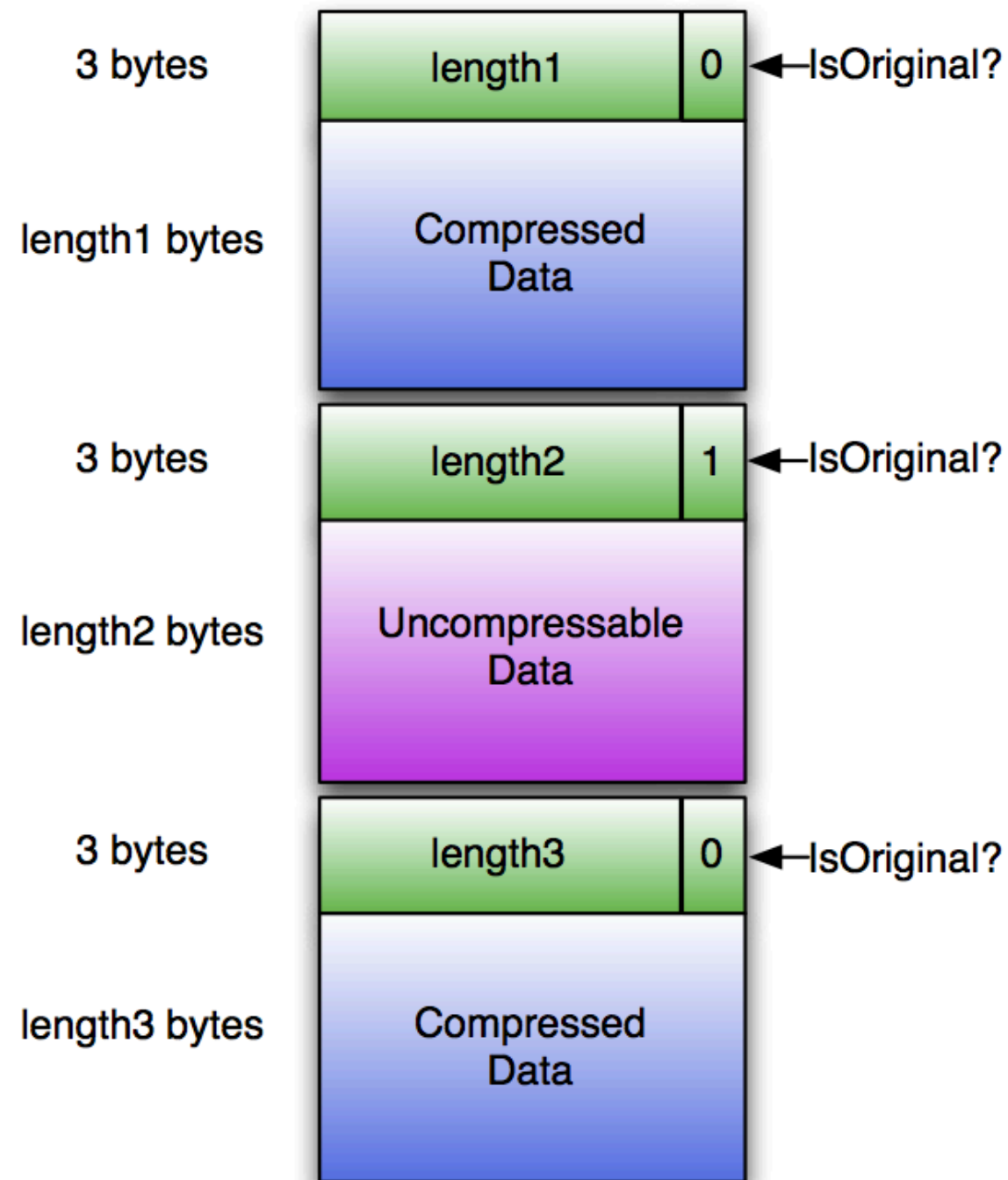
- * Список stripe-ов
- * Количество строк в каждом страйпе
- * Типы данных всех полей
- * Индексы по полям (min, max, count, sum)

ORC структура строки



ORC сжатие

- * Пишется блоками (256к)
- * Проверка размера (жать/не жать)
- * Поддержка ZLIB, SNAPPY



ORC

индексы

- * File level (общая статистика по всем полям в файле)
- * Stripe level (статистика по каждому полю страйпа)
- * Row level (статистика по каждому полю по 10K строк в страйпе)

ORC

ИНДЕКСЫ

- * `orc.create.index=true`
- * `orc.row.index.stride=10000`
- * `min+max+sum`
- * `orc.bloom.filter.columns="field1,feild2"`
- * `orc.bloom.filter.fpp=0.05`

ORC benchmark

