Обзор Hadoop 3.0, HDFS 3.0 (высшая математика included)

Драль Алексей, study@bigdatateam.org
CEO at BigData Team, https://bigdatateam.org
https://www.facebook.com/bigdatateam

BIGDATA TEAM

Hadoop 3.0.0

Release	Date	Released?
3.0.0-alpha1	2016-09-03	✓
3.0.0-alpha2	2017-01-25	✓
3.0.0-alpha3	2017-05-26	✓
3.0.0-alpha4	2017-07-07	✓
3.0.0-beta1	2017-10-03	✓
3.0.0 GA	2017-12-13	✓

https://dandelion.eu/datagems/SpazioDati/telecom-sms-call-internet-mi

Hadoop 3.0, что новенького?

- Улучшена интеграция с облаками (Microsoft, Alibaba, AWS, ...)

Hadoop 3.0, что новенького?

- Улучшена интеграция с облаками (Microsoft, Alibaba, AWS, ...)
- YARN:
  - ~~Spot Instances~~ Opportunistic Containers

Hadoop 3.0, что новенького?

- Улучшена интеграция с облаками (Microsoft, Alibaba, AWS, ...)
- YARN:
  - ~~Spot Instances~~ Opportunistic Containers
  - GPU, лицензии и другие ресурсы

Hadoop 3.0, что новенького?

- Улучшена интеграция с облаками (Microsoft, Alibaba, AWS, ...)
- YARN:
  - ~~Spot Instances~~ Opportunistic Containers
  - GPU, лицензии и другие ресурсы
  - 30% ускорение для shuffle-нагруженных MR-задач

Hadoop 3.0, что новенького?

- Улучшена интеграция с облаками (Microsoft, Alibaba, AWS, ...)
- YARN:
  - ~~Spot Instances~~ Opportunistic Containers
  - GPU, лицензии и другие ресурсы
  - 30% ускорение для shuffle-нагруженных MR-задач
- HDFS:
  - возможность иметь 2+ StandBy Namenode

Hadoop 3.0, что новенького?

- Улучшена интеграция с облаками (Microsoft, Alibaba, AWS, ...)
- YARN:
  - ~~Spot Instances~~ Opportunistic Containers
  - GPU, лицензии и другие ресурсы
  - 30% ускорение для shuffle-нагруженных MR-задач
- HDFS:
  - возможность иметь 2+ StandBy Namenode
  - intra- datanode balancer

Hadoop 3.0, что новенького?

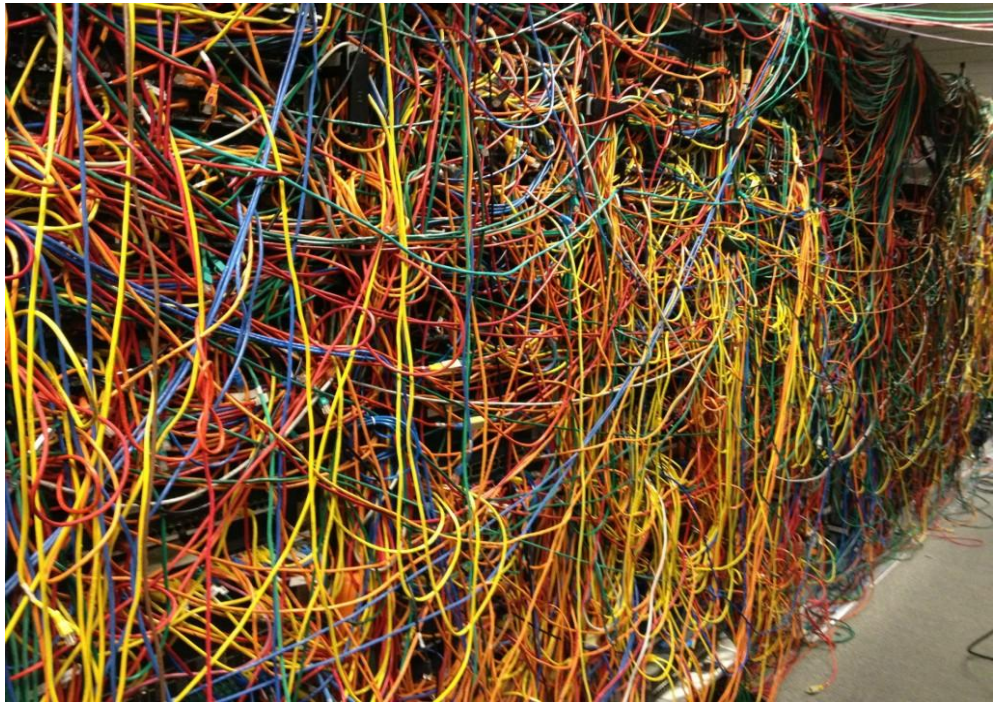
- Улучшена интеграция с облаками (Microsoft, Alibaba, AWS, ...)
- YARN:
  - ~~Spot Instances~~ Opportunistic Containers
  - GPU, лицензии и другие ресурсы
  - 30% ускорение для shuffle-нагруженных MR-задач
- HDFS:
  - возможность иметь 2+ StandBy Namenode
  - intra- datanode balancer
  - HDFS Federation → HDFS Router-Based Federation

Hadoop 3.0, что новенького?

- Улучшена интеграция с облаками (Microsoft, Alibaba, AWS, ...)
- YARN:
  - ~~Spot Instances~~ Opportunistic Containers
  - GPU, лицензии и другие ресурсы
  - 30% ускорение для shuffle-нагруженных MR-задач
- HDFS:
  - возможность иметь 2+ StandBy Namenode
  - intra- datanode balancer
  - HDFS Federation → HDFS Router-Based Federation
  - HDFS Erasure Coding (избыточное кодирование в HDFS)

BIGDATA TEAM

Распределенные вычисления



Fail-Recovery + Fair-Loss Link + Asynchronous

HDFS до версии 3.0

Условия:

- 100 вычислительных узлов
- до 5% узлов вышли единовременно из строя

Политика реплицирования и гарантии доступности данных:

- 1 реплика: доступность данных - 95%

BIGDATA TEAM

HDFS до версии 3.0

Условия:

- 100 вычислительных узлов
- до 5% узлов вышли единовременно из строя

Политика реплицирования и гарантии доступности данных:

- 1 реплика: доступность данных - 95%
- 2 реплики: доступность данных - 99.75%

HDFS до версии 3.0

Условия:

- 100 вычислительных узлов
- до 5% узлов вышли единовременно из строя

Политика реплицирования и гарантии доступности данных:

- 1 реплика: доступность данных - 95%
- 2 реплики: доступность данных - 99.75% (overhead - 100%)

HDFS до версии 3.0

Условия:

- 100 вычислительных узлов
- до 5% узлов вышли единовременно из строя

Политика реплицирования и гарантии доступности данных:

- 1 реплика: доступность данных - 95%
- 2 реплики: доступность данных - 99.75% (overhead - 100%)
- 3 реплики: доступность данных - 99.9875% (overhead - 200%)

BIGDATA TEAM

Немного математики

HDFS Erasure Coding

Erasure coding policies. To accommodate heterogeneous workloads, we allow files and directories in an HDFS cluster to have different replication and erasure coding policies. The erasure coding policy encapsulates how to encode/decode a file. Each policy is defined by the following pieces of information:

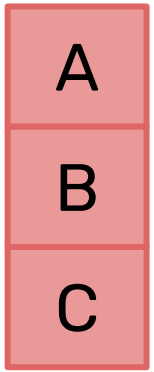
- The EC schema: This includes the numbers of data and parity blocks in an EC group (e.g., 6+3), as well as the codec algorithm (e.g., Reed-Solomon, XOR).
- The size of a striping cell. This determines the granularity of striped reads and writes, including buffer sizes and encoding work.

Policies are named codec-num data blocks-num parity blocks-cell size. Currently, six built-in policies are supported: RS-3-2-1024k, RS-6-3-1024k, RS-10-4-1024k, RS-LEGACY-6-3-1024k, XOR-2-1-1024k and REPLICATION.

https://hadoop.apache.org/docs/r3.0.0/.../hadoop-hdfs/HDFSErasureCoding.html

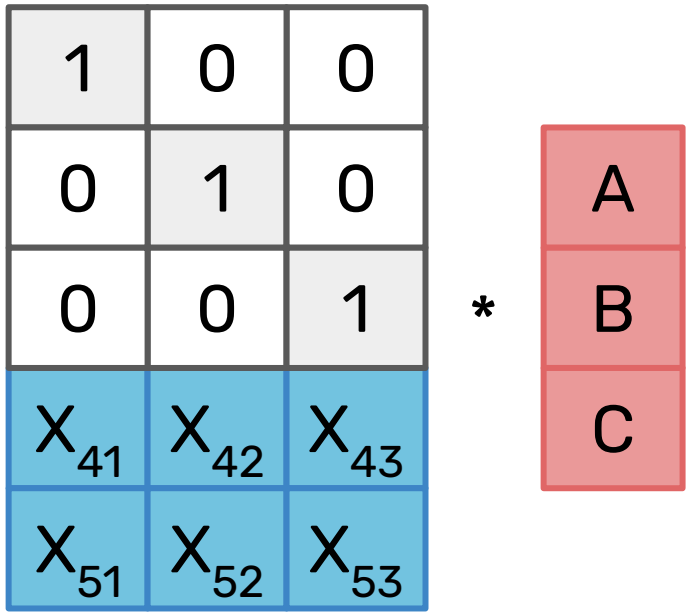
BIGDATA TEAM

Политика хранения из HDFS 3.0



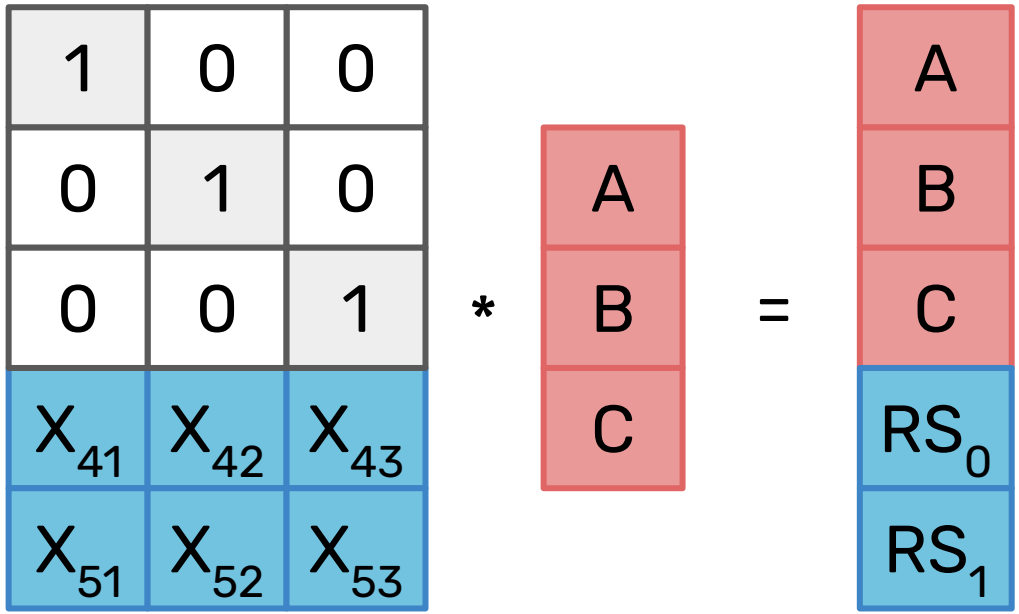
BIGDATA TEAM

Политика хранения из HDFS 3.0



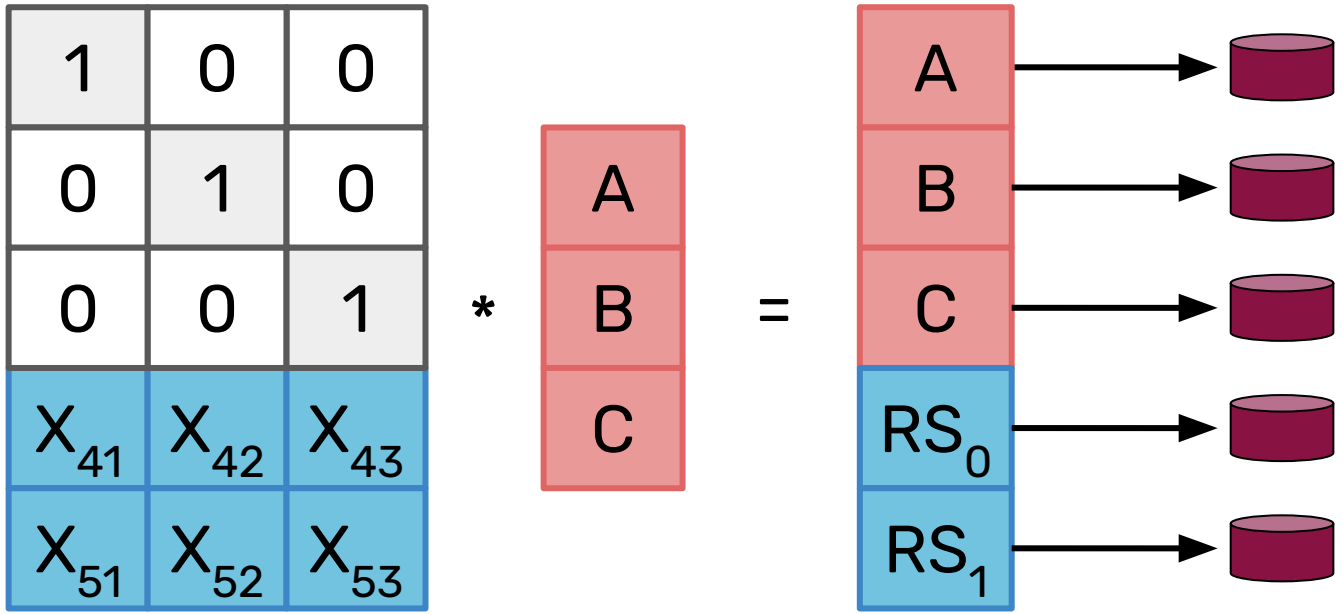
BIGDATA TEAM

Политика хранения из HDFS 3.0



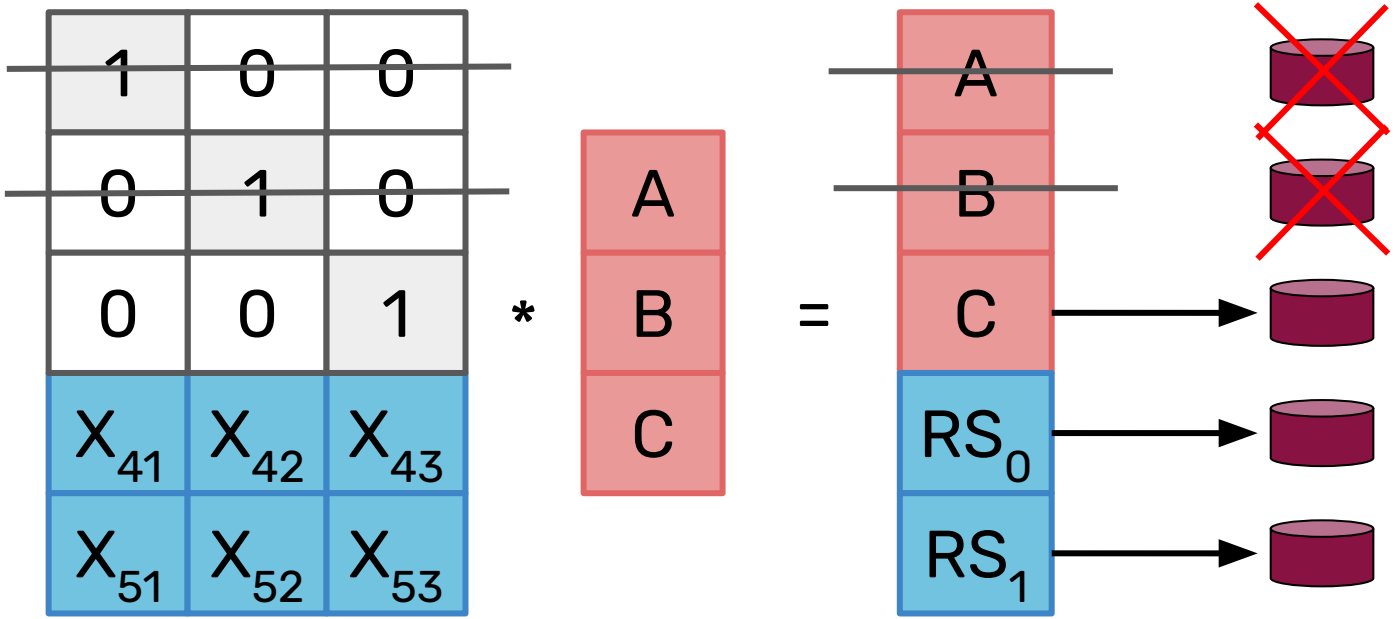
BIGDATA TEAM

Политика хранения из HDFS 3.0



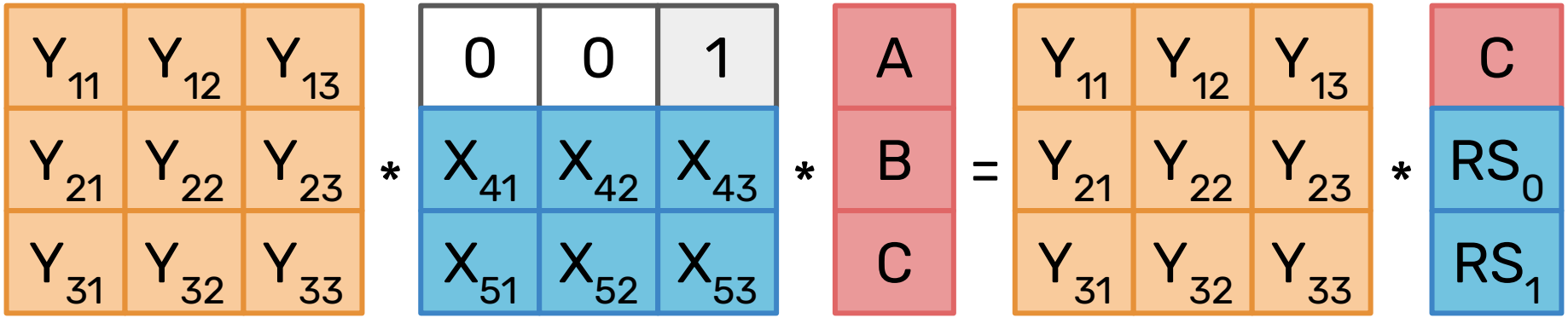
BIGDATA TEAM

Политика хранения из HDFS 3.0



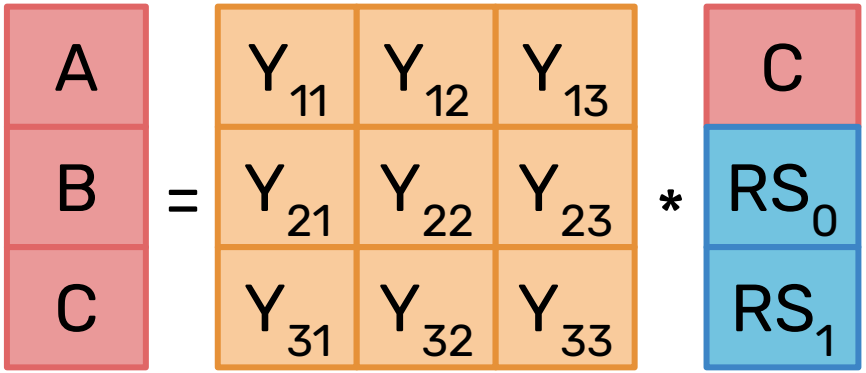
BIGDATA TEAM

Политика хранения из HDFS 3.0



BIGDATA TEAM

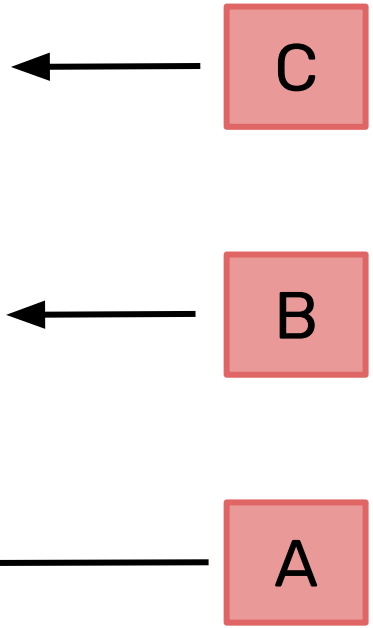
Политика хранения из HDFS 3.0



BIGDATA TEAM

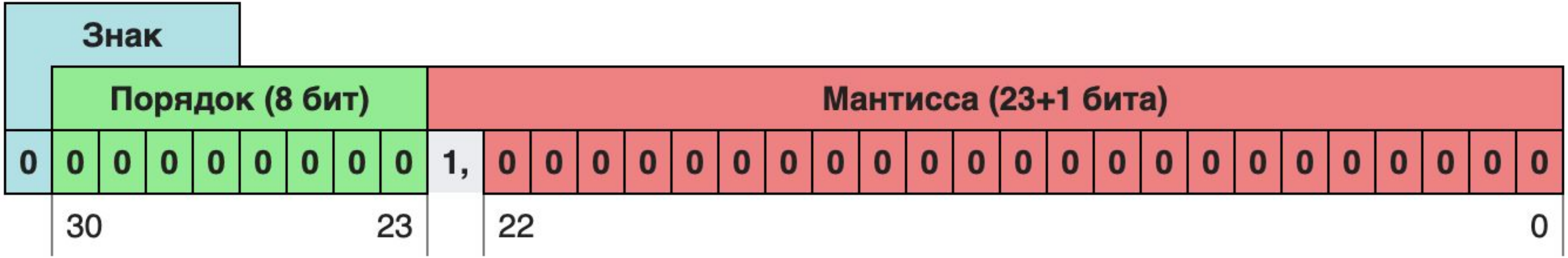
Представление целых чисел

Число	Бинарное представление
2^n-1	011...111
...	...
2	000...010
1	000...001
0	000...000
-0	100...000
-1	100...001
-2	100...010
...	...
$-(2^n-1)$	111...111

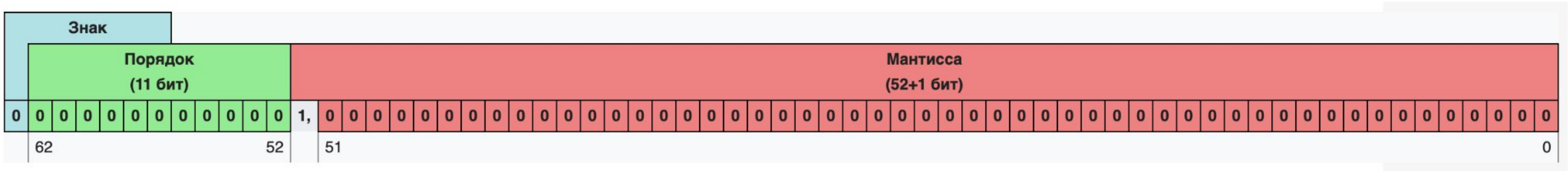


Представление вещественных чисел

- Одинарная точность (float, 4 байта)



- Двойная точность (double, 8 байт)



источник

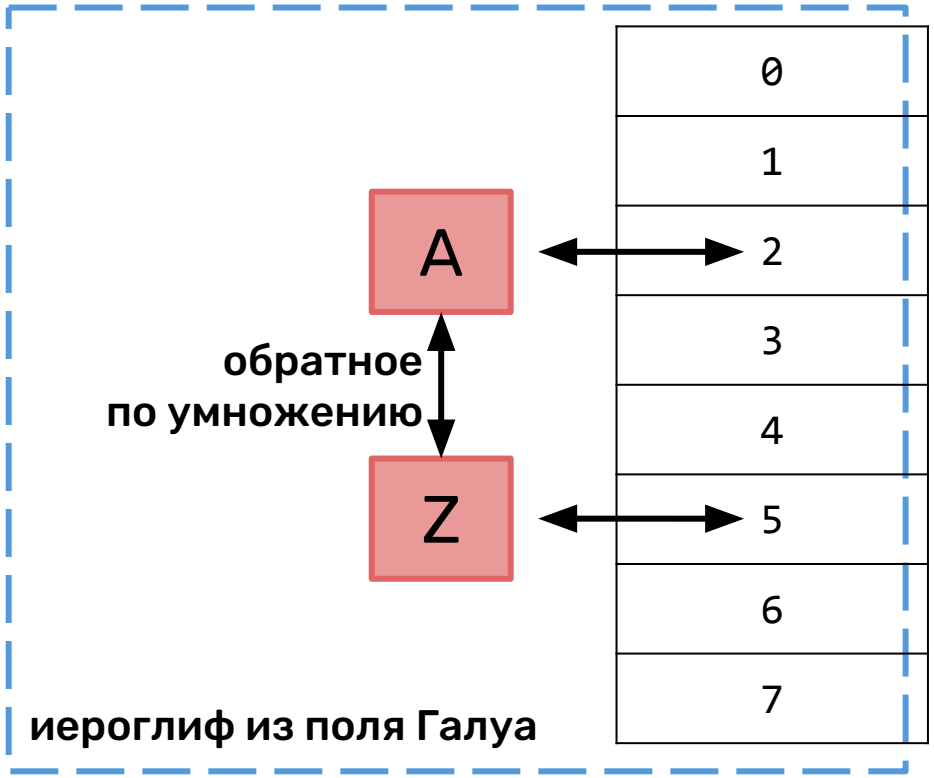
BIGDATA TEAM

Готовимся к погружению в математику

BIGDATA TEAM

Конечное поле Галуа

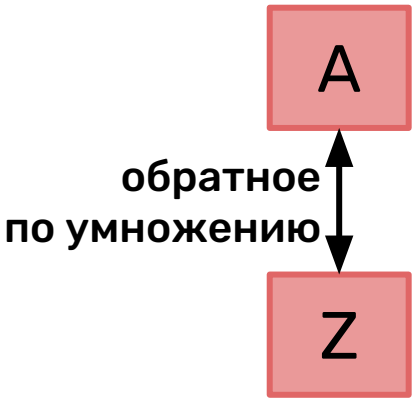
- GS(2^n) (сокращение от **G**alois **F**ield)



BIGDATA TEAM

Конечное поле Галуа

- GS(2^n) (сокращение от **G**alois **F**ield)



0	000
1	001
2	010
3	011
4	100
5	101
6	110
7	111

иероглиф из поля Галуа

соответствие бинарной системе счисления

Q&A

Операция сложения в GS(2^3)

BIGDATA TEAM

Операция сложения

A	B	C = A ⊕ B
0	0	0
0	1	1
1	0	1
1	1	0

```
  01000011
+ 01110010
----------
  00110001
```

BIGDATA TEAM

Галуа и умножение

Эвари́ст Галуа́ (фр. Évariste Galois; 1811-1832)

В 18 лет придумал поля* Галуа (1830)

А что сделал ты, когда тебе было 18?

*см. уточнения например здесь и здесь

$GF(2^3)$

- Порождающий многочлен (generator polynomial): $g(x) = x^3 + x + 1$

$GF(2^3)$

- Порождающий многочлен (generator polynomial): $g(x) = x^3 + x + 1$

1	001	**1**
x	010	**2**
x + 1	011	**3**
x^2	100	**4**
x^2 + 1	101	**5**
x^2 + x	110	**6**
x^2 + x + 1	111	**7**

BIGDATA TEAM

$GF(2^3)$

- Порождающий многочлен (generator polynomial): $g(x) = x^3 + x + 1$

1	001	**1**
x	010	**2**
x + 1	011	**3**
x^2	100	**4**
x^2 + 1	101	**5**
x^2 + x	110	**6**
x^2 + x + 1	111	**7**

умножение:

- 3 = (x + 1), 5 = (x^2 + 1)
- 3 * 5 = (x + 1) * (x^2 + 1) = x^3 + x + x^2 + 1 = ~~(x^3 + x + 1)~~ + x^2 = x^2 = 4

пример работы с коэффициентами (XOR):

- 5 + 6 = (x^2 + 1) + (x^2 + x) = ~~(x^2 + x^2)~~ + x + 1 = x + 1 = 3

BIGDATA TEAM

Таблица умножения в $GF(2^3)$

			1	**2**	**3**	**4**	**5**	**6**	**7**
1	001	**1**	1	2	3	4	5	6	7
x	010	**2**	2	4	6	3	1	7	5
x + 1	011	**3**	3	6	5	7	4	1	2
x^2	100	**4**	4	3	7	6	2	5	1
x^2 + 1	101	**5**	5	1	4	2	7	3	6
x^2 + x	110	**6**	6	7	1	5	3	2	4
x^2 + x + 1	111	**7**	7	5	2	1	6	4	3

BIGDATA TEAM

Таблица умножения в $GF(2^3)$

			1	**2**	**3**	**4**	**5**	**6**	**7**
1	001	**1**	1	2	3	4	5	6	7
x	010	**2**	2	4	6	3	1	**7**	5
x + 1	011	**3**	3	6	5	7	4	1	2
x^2	100	**4**	4	3	7	6	2	5	1
x^2 + 1	101	**5**	5	1	4	2	7	3	6
x^2 + x	110	**6**	6	7	1	5	3	2	4
x^2 + x + 1	111	**7**	7	5	2	1	6	4	3

BIGDATA TEAM

Магия степеней в GF(2^n)

Порождающий элемент (primitive element, generator): 2			степени							
			0	**1**	**2**	**3**	**4**	**5**	**6**	**7**
1	001	**1**	1	1	1	1	1	1	1	1
x	010	**2**	1	2	4	3	6	7	5	1
x + 1	011	**3**	1	3	5	4	7	2	6	1
x^2	100	**4**	1	4	6	5	2	3	7	1
x^2 + 1	101	**5**	1	5	7	6	3	4	2	1
x^2 + x	110	**6**	1	6	2	7	4	5	3	1
x^2 + x + 1	111	**7**	1	7	3	2	5	6	4	1

BIGDATA TEAM

Магия степеней в GF(2^n)

- Порождающий элемент (primitive element, generator): 2

			степени							
			0	1	2	3	4	5	6	7
1	001	1	1	1	1	1	1	1	1	1
x	010	2	1	2	4	3	6	7	5	1
x + 1	011	3	1	3	5	4	7	2	6	1
x^2	100	4	1	4	6	5	2	3	7	1
x^2 + 1	101	5	1	5	7	6	3	4	2	1
x^2 + x	110	6	1	6	2	7	4	5	3	1
x^2 + x + 1	111	7	1	7	3	2	5	6	4	1

умножение:

- 3 = 2^3, 5 = 2^6
- 3 * 5 = 2^3 * 2^6 = 2^(3+6=9) = 2^(9 mod 7) = 2^2 = 4

деление:

- 7 = 2^5
- 7 / 2 = 2^5 / 2^1 = 2^(5-1) = 2^4 = 6

возведение в степень:

- 3^5 = (2^3)^5 = 2^(15 mod 7) = 2^1 = 2

BIGDATA TEAM

Как выбирать X, чтобы получить Y?

Y_{11}	Y_{12}	Y_{13}
Y_{21}	Y_{22}	Y_{23}
Y_{31}	Y_{32}	Y_{33}

*

0	0	1
X_{41}	X_{42}	X_{43}
X_{51}	X_{52}	X_{53}

*

A
B
C

=

Y_{11}	Y_{12}	Y_{13}
Y_{21}	Y_{22}	Y_{23}
Y_{31}	Y_{32}	Y_{33}

*

C
RS_0
RS_1

BIGDATA TEAM

Как выбирать X, чтобы получить Y

Y_{11}	Y_{12}	Y_{13}
Y_{21}	Y_{22}	Y_{23}
Y_{31}	Y_{32}	Y_{33}

*

0	0	1
X_{41}	X_{42}	X_{43}
X_{51}	X_{52}	X_{53}

*

A
B
C

=

Y_{11}	Y_{12}	Y_{13}
Y_{21}	Y_{22}	Y_{23}
Y_{31}	Y_{32}	Y_{33}

*

C
RS_0
RS_1

$$V_m = \begin{bmatrix} 1 & X_1 & X_1^2 & \dots & X_1^{m-1} \\ 1 & X_2 & X_2^2 & \dots & X_2^{m-1} \\ . & . & . & . & . \\ 1 & X_m & X_m^2 & \dots & X_m^{m-1} \end{bmatrix}$$

матрица Вандермонда

$$\begin{pmatrix} \frac{1}{a_1+b_1} & \frac{1}{a_1+b_2} & \dots & \frac{1}{a_1+b_n} \\ \frac{1}{a_2+b_1} & \frac{1}{a_2+b_2} & \dots & \frac{1}{a_2+b_n} \\ \dots & \dots & \dots & \dots \\ \frac{1}{a_n+b_1} & \frac{1}{a_n+b_2} & \dots & \frac{1}{a_n+b_n} \end{pmatrix}$$

матрица Коши

$$H = \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \frac{1}{6} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \frac{1}{6} & \frac{1}{7} \\ \frac{1}{4} & \frac{1}{5} & \frac{1}{6} & \frac{1}{7} & \frac{1}{8} \\ \frac{1}{5} & \frac{1}{6} & \frac{1}{7} & \frac{1}{8} & \frac{1}{9} \end{bmatrix}.$$

матрица Гильберта

HDFS до версии 3.0

Условия:

- 100 вычислительных узлов
- до 5% узлов вышли единовременно из строя

Политика реплицирования и гарантии доступности данных:

- 1 реплика: доступность данных - 95%
- 2 реплики: доступность данных - 99.75% (overhead - 100%)
- 3 реплики: доступность данных - 99.9875% (overhead - 200%)
- RS-3-2-1024k: доступность данных - ??% (overhead - 67%)

HDFS до версии 3.0

Условия:

- 100 вычислительных узлов
- до 5% узлов вышли единовременно из строя

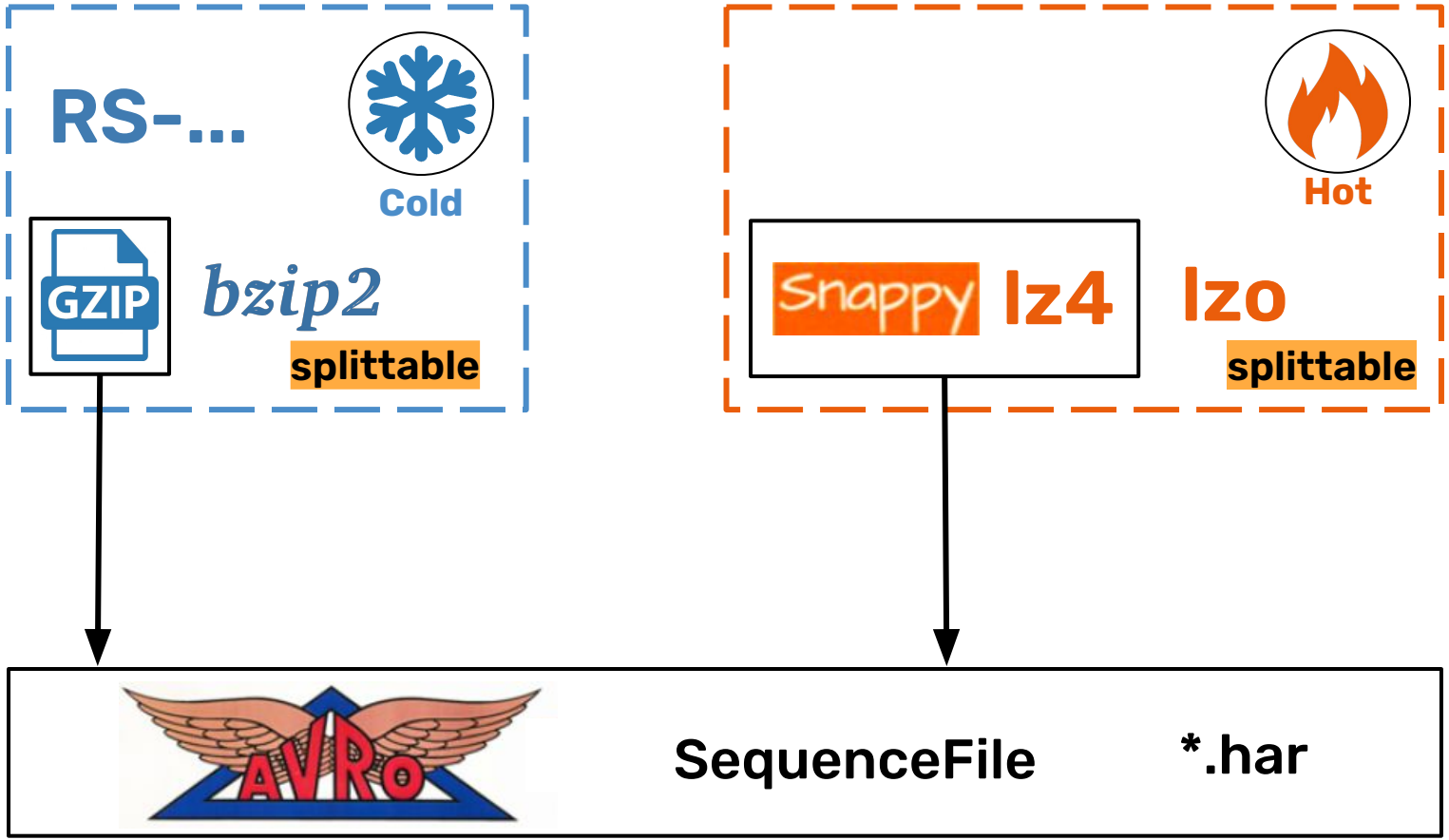
Политика реплицирования и гарантии доступности данных:

- 1 реплика: доступность данных - 95%
- 2 реплики: доступность данных - 99.75% (overhead - 100%)
- 3 реплики: доступность данных - 99.9875% (overhead - 200%)
- RS-3-2-1024k: доступность данных - 99.94% (overhead - 67%)

Q&A

Какие минусы у RS-10-4-1024k?

Стандартные подходы



Список литературы на лето

HDFS Erasure Coding in Production

- https://blog.cloudera.com/hdfs-erasure-coding-in-production/

Резюме

- Было круто, повторим?
- Hive 3+
- Spark 3+
- ...