A3 HyperMegaLogLog Pro Max++

Реализация

Генерация потока (`RandomStreamGen`)

Реализован генератор строк:

- длина строки: от 1 до 30 символов;
- алфавит: `a-z` , `A-Z` , `0-9` , `-` ;
- поток формируется с повторениями через параметр `reuse_prob` .

Моменты времени t задаются как префиксы потока с шагом 5%:

$$t \in \{0.05n, 0.10n, \dots, 1.00n\}.$$

Генерация хеш-функции (`HashFuncGen`)

Используется схема:

1. Базовый хеш строки: FNV-1a 64-bit.
2. Линейное преобразование:

$$h(x) = (a \cdot x + b) \bmod 2^{32},$$

где a, b выбираются случайно, a делается нечётным.

HyperLogLog

Для каждого элемента:

1. Вычисляется 32-битный хеш.
2. Первые B бит идут в индекс регистра.
3. По оставшимся битам вычисляется ρ (позиция первого `1`).
4. Регистр обновляется максимумом.

Оценка:

$$\hat{N} = \alpha_m \cdot m^2 \left(\sum_{j=1}^{m} 2^{-M_j} \right)^{-1},$$

где $m = 2^B$, M_j — значение j-го регистра.

Также применены стандартные поправки:

- `small-range` (linear counting),
- `large-range` для 32-битного пространства.

Параметры эксперимента

- число потоков: `NUM_STREAMS = 30` ,
- длина потока: `STREAM_SIZE = 200000` ,
- шаг по времени: `STEP_PERCENT = 5` ,
- доля повторов: `REUSE_PROB = 0.78` ,
- параметр HLL: `B = 12` ,
- число регистров: $m = 2^{12} = 4096$.

Теоретические ориентиры:

$$\frac{1.04}{\sqrt{m}} = \frac{1.04}{64} = 0.01625 \ (1.625\%),$$

$$\frac{1.3}{\sqrt{m}} = \frac{1.3}{64} = 0.0203125 \ (2.031\%).$$

Методика оценки

На каждом шаге t:

1. Считалось точное значение F_0^t через `unordered_set` .
2. Считалась оценка HLL N_t.

Далее по 30 потокам вычислялись:

- $E(F_0^t)$,
- $E(N_t)$,
- σ_t (стандартное отклонение оценки),
- относительный сдвиг:

$$bias_{rel}(t) = \frac{E(N_t) - E(F_0^t)}{E(F_0^t)},$$

- относительная RMSE:

$$rmse_{rel}(t) = \frac{\sqrt{E\left[(N_t - F_0^t)^2\right]}}{E(F_0^t)},$$

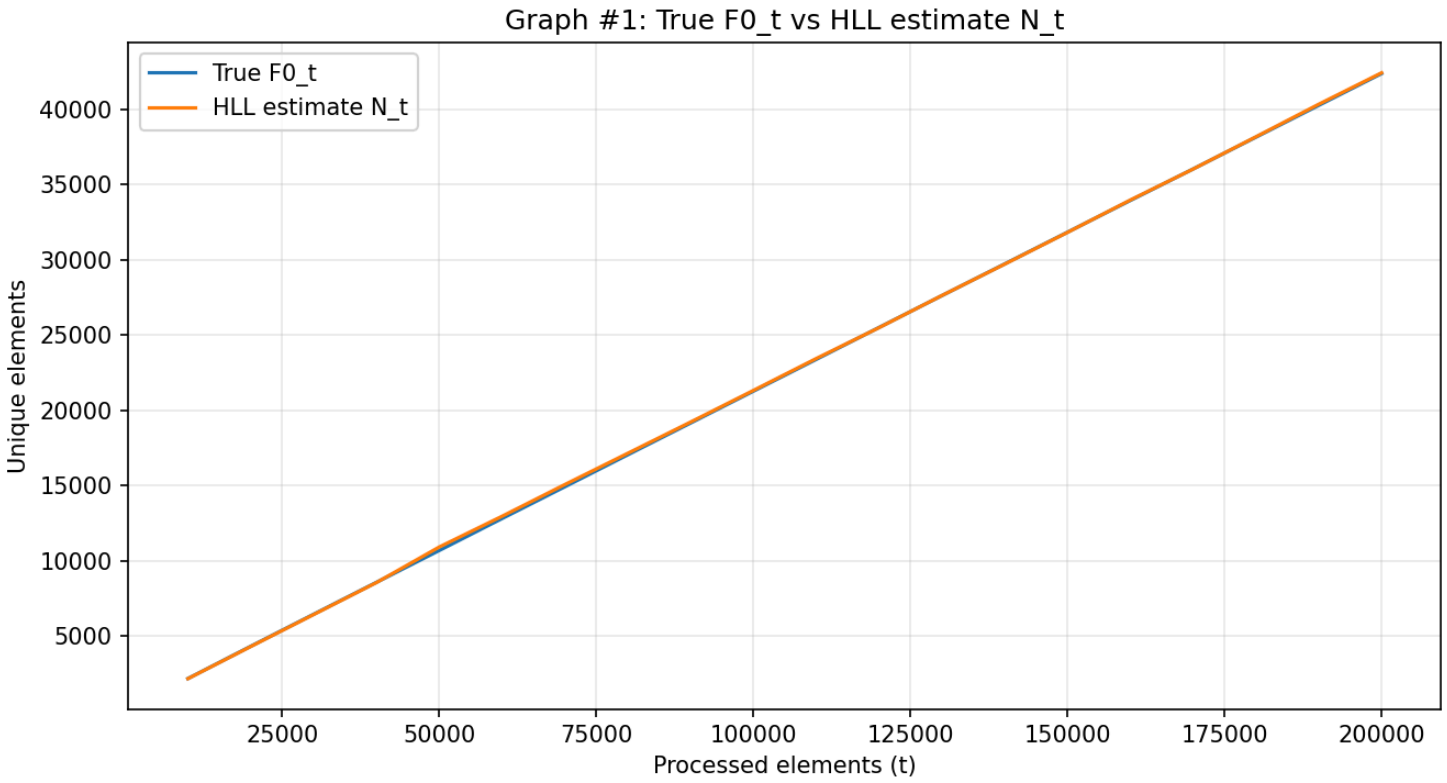
- относительная вариативность:

$$cv_t = \frac{\sigma_t}{E(F_0^t)}.$$

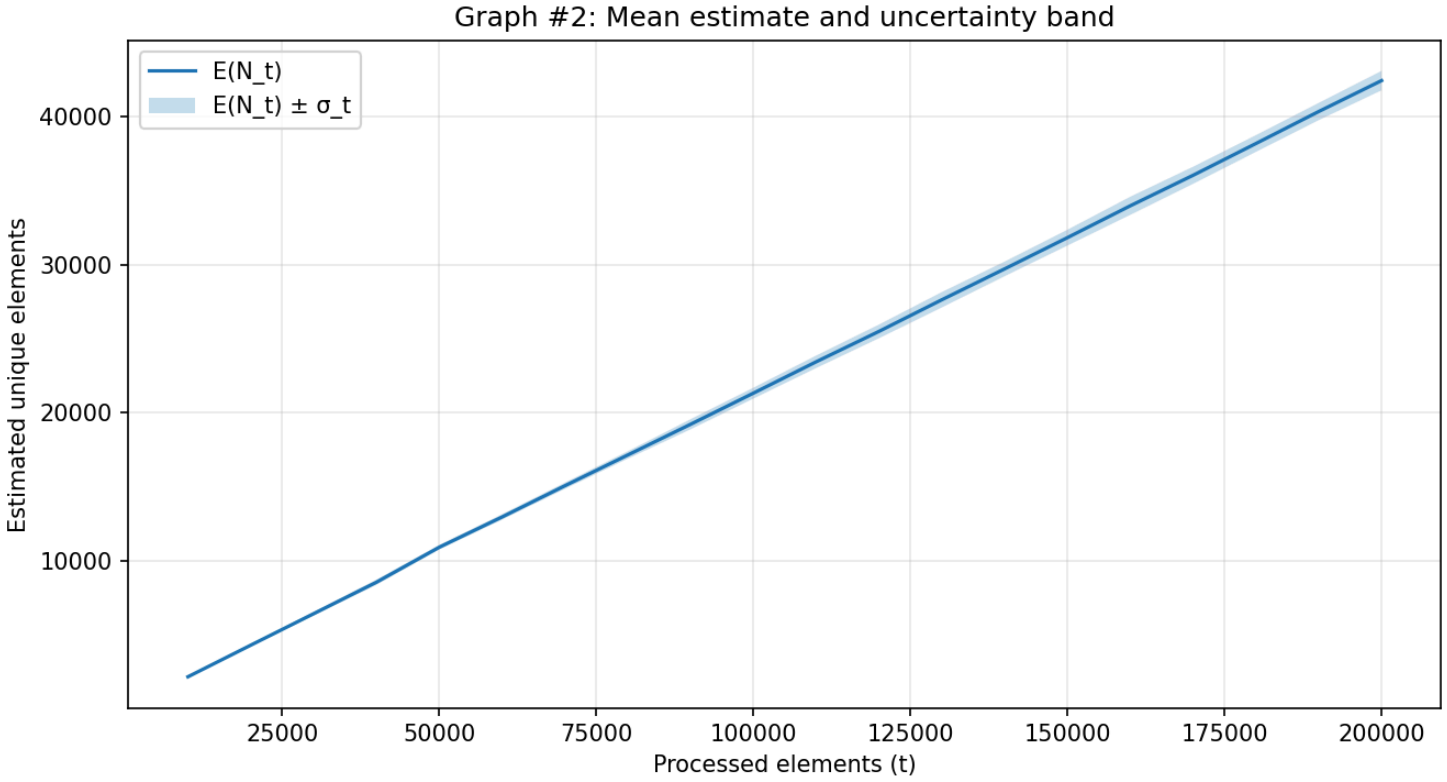
Результаты

Построены графики:

1. Сравнение F_0^t и N_t - кривые почти совпадают на всём диапазоне.



2. $E(N_t)$ и полоса неопределённости $E(N_t) \pm \sigma_t$ — полоса узкая и стабильная.



Сводные метрики:

- число шагов: `20`,
- средний $|bias_{rel}|$: `0.003753` (0.375%),
- максимальный $|bias_{rel}|$: `0.021526` (2.153%),
- средний $rmse_{rel}$: `0.014800` (1.480%),
- максимальный $rmse_{rel}$: `0.024115` (2.411%),
- средний cv: `0.016929` (1.693%),
- максимальный cv: `0.022297` (2.230%).

Доли шагов, удовлетворяющих границам:

- $cv_t \le 1.04/\sqrt{m}$: `40%`,
- $cv_t \le 1.3/\sqrt{m}$: `95%`,
- $rmse_{rel}(t) \le 1.3/\sqrt{m}$: `95%`.

Дополнительно:

- $rmse_{rel}(t) \le 1.04/\sqrt{m}$ выполняется для `90%` шагов.

Худший шаг по RMSE: $t = 50000$,

- $bias_{rel} \approx +2.15\%$,
- $rmse_{rel} \approx 2.41\%$.
  Это единичный выброс на остальных шагах качество существенно лучше.

Сравнение с теорией

1. **Точность.**
   В среднем ошибка соответствует ожидаемому уровню для $B = 12$:
   средний $rmse_{rel} = 1.48\%$ близок к теоретическим 1.625%.
2. **Стабильность.**
   Средний $cv = 1.693\%$ находится около теоретического порядка.
   Большинство шагов укладывается в ослабленную границу $1.3/\sqrt{m}$.
3. **Влияние констант.**
   Точность определяется главным образом B (т.е. числом регистров $m = 2^B$): при увеличении B ожидается снижение дисперсии и RMSE ценой роста памяти.