NoSQL

Paulina Seroka & Michał Jaworowski

Zbiór danych

Temat: **US Baby Names**

Zbiór imion nadawanych dzieciom
w latach 1910-2014 w
Stanach Zjednoczonych

Liczba rekordów: **5 647 246**

PRZYKLADOWY REKORD

```
{
  "_id" : ObjectId("58ebb1538fdb91a54781be10"),
  "Id" : 1,
  "Name" : "Mary",
  "Year" : 1910,
  "Gender" : "F",
  "State" : "AK",
  "Count" : 14
}
```

Znaczenie kolumn:

- `Id` pole zawiera numer rekordu
- `Name` pole zawiera nadane imię dla dziecka
- `Year` pole zawiera rok narodzin dziecka
- `Gender` pole zawiera płeć dziecka
- `State` pole zawiera stan w USA narodzin dziecka
- `Count` pole zawiera liczbę nadań takiego imienia

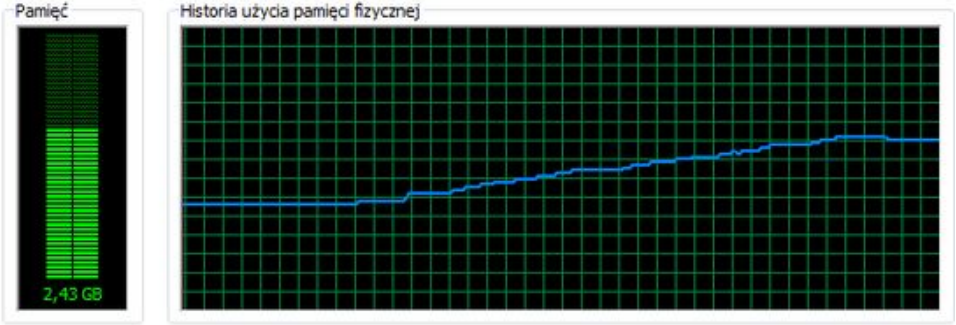
PARAMETRY KOMPUTERA TESTOWEGO

Jednostka	Parametr
System	Windows 7 64bit
Procesor	Intel(R) Core(TM) i5-2450M
Ilość rdzeni	2
Moc rdzenia	2.50GHz
Pamięć RAM	4,00 GB

Obciążenie komputera podczas importowania danych

CPU

RAM



- 1.52 GB przed uruchomieniem bazy
- 1.62 GB po uruchomieniu bazy
- 2.43 GB po imporcie danych

DYSK

Agregacje

1) Najczęściej nadawane imiona w Stanach Zjednoczonych w latach 1910-2014.

2) Znalezienie okresu największej popularności wybranych imion.

3) Średnia roczna urodzeń dziewczynek i chłopców w latach 1910-2014.

4) Top 2-6 najczęściej nadawanych imion męskich zaczynających się literą "M".

5) Zbiorcze zestawienie liczby urodzeń w każdym roku.

Agregacja 1

Najczesciej nadawane imiona

Agregacja ma na celu sprawdzenie, które imiona są najbardziej popularne.

```
db.names.aggregate(
        { $group: {
                _id: { Gender: "$Gender", Name: "$Name" },
                Number: { $sum: "$Count" }
        }},
        { $sort: { Number: -1 }
        },
        { $limit: 10 }
)
```

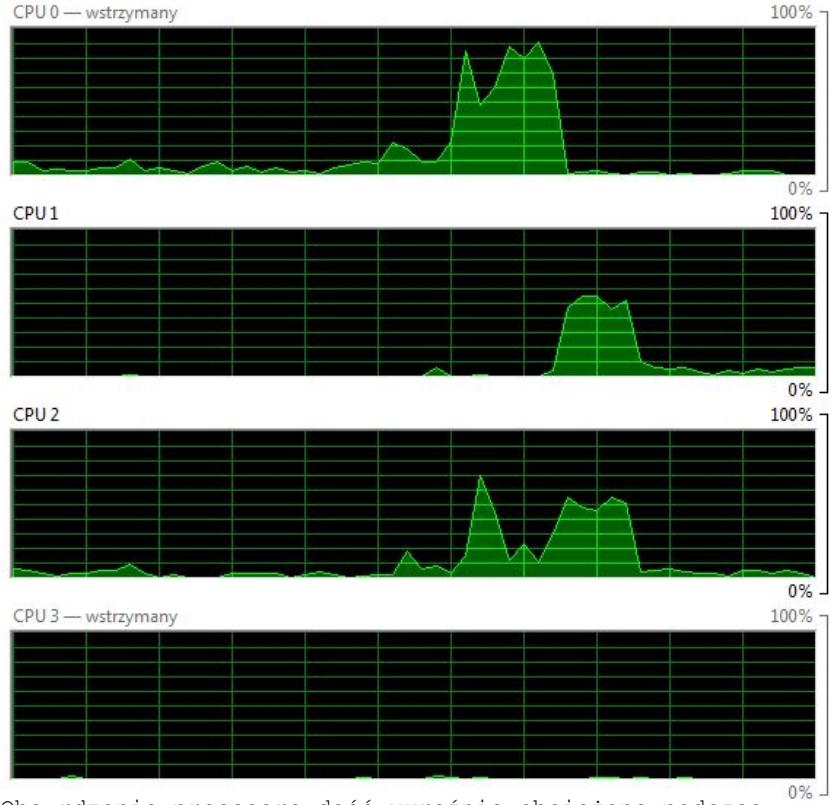
Wynik:

```
{ "_id" : { "Gender" : "M", "Name" : "James" }, "Number" : 4938965 }
{ "_id" : { "Gender" : "M", "Name" : "John" }, "Number" : 4829733 }
{ "_id" : { "Gender" : "M", "Name" : "Robert" }, "Number" : 4710600 }
{ "_id" : { "Gender" : "M", "Name" : "Michael" }, "Number" : 4295779 }
{ "_id" : { "Gender" : "M", "Name" : "William" }, "Number" : 3829026 }
{ "_id" : { "Gender" : "F", "Name" : "Mary" }, "Number" : 3730856 }
```

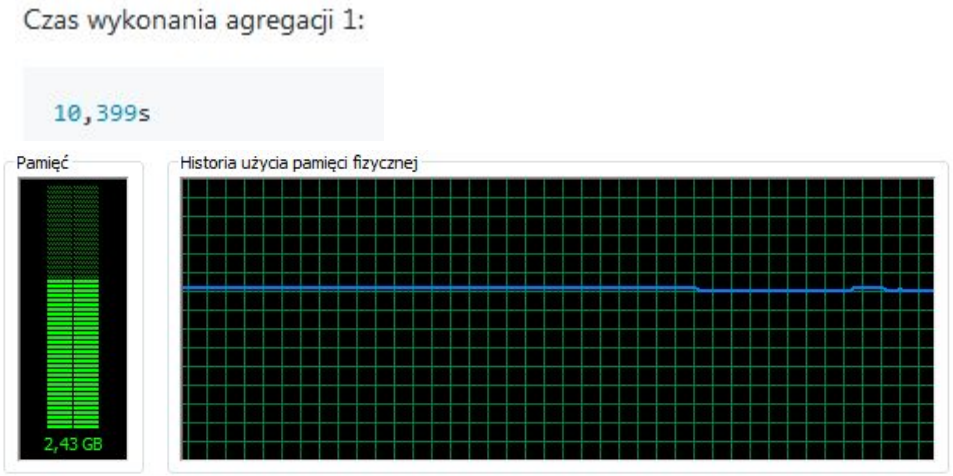
- **$group** - wymaga pola *_id*, w którym wyznaczamy po jakich polach grupujemy, pole *Number* korzysta z funkcji agregacji $sum
- **$sum** - sumuje liczbę nadanych tych samych imion poszczególnym płciom po polu *Name*
- **$sort** - sortuje malejąco względem pola *Number*
- **$limit** - ogranicza liczbę rekordów wynikowych

Agregacja 1

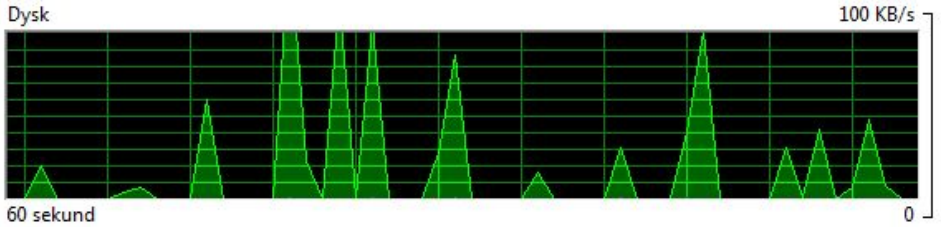
Obciążenie komputera podczas wykonania agregacji



Oba rdzenie procesora dość wyraźnie obciążone podczas wykonywania agregacji (CPU 0 i CPU 2), a także zaangażowany 3-ci wątek (CPU 1).



2.43 GB stabilnie, bez zmian podczas wykonywania agregacji.



Praktycznie bezczynny dysk podczas wykonywania agregacji (100 kb/s to najmniejsza podziałka wykresu obciążenia dysku).

AGREGECJA 2

OKRES NAJWIEKSZEJ POPULARNOSCI WYBRANYCH IMION

Agregacja ma na celu zbadanie tendencji w nadawaniu wybranych imion na przestrzeni lat.

Kiedy wybrane imię było najbardziej popularne i dlaczego?

Jakie wydarzenia w historii Stanów Zjednoczonych mają wpływ na decyzję o wyborze imienia dla dziecka?

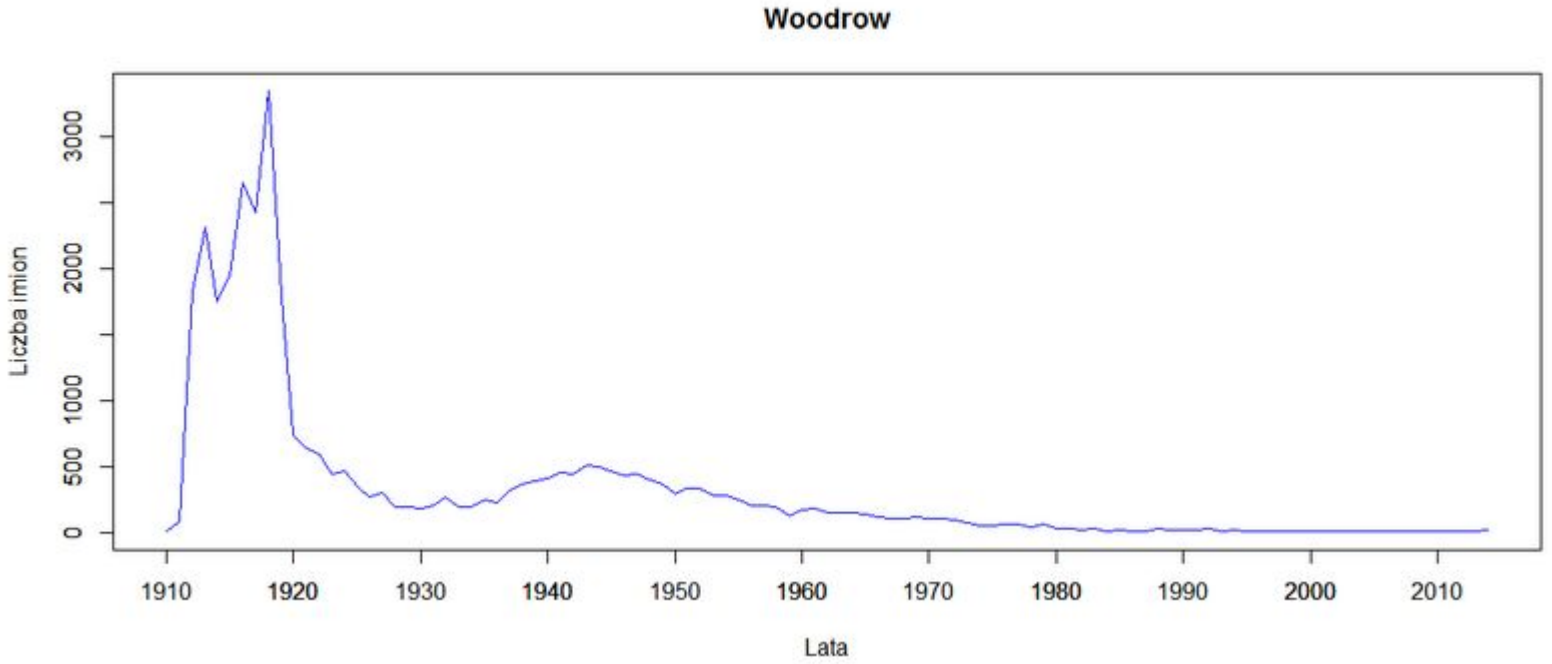
WOODROW

```
db.names.aggregate(
  { $match: { Name: "Woodrow" } },
  { $group: {
    _id: { Year: "$Year" },
    Number: { $sum: "$Count" }
  }},
  { $sort: { "_id.Year" : 1}},
  { $out : "agr1" }
)
```

Wynik:

```
{ "_id" : { "Year" : 1910 }, "Number" : 11 }
{ "_id" : { "Year" : 1911 }, "Number" : 82 }
{ "_id" : { "Year" : 1912 }, "Number" : 1826 }
{ "_id" : { "Year" : 1913 }, "Number" : 2314 }
{ "_id" : { "Year" : 1914 }, "Number" : 1747 }
{ "_id" : { "Year" : 1915 }, "Number" : 1941 }
{ "_id" : { "Year" : 1916 }, "Number" : 2645 }
{ "_id" : { "Year" : 1917 }, "Number" : 2422 }
{ "_id" : { "Year" : 1918 }, "Number" : 3337 }
{ "_id" : { "Year" : 1919 }, "Number" : 1791 }
{ "_id" : { "Year" : 1920 }, "Number" : 733 }
...
```

- **$match** - wybiera z bazy tylko te rekordy, które zawierają w polu *Name* słowo *Woodrow*
- **$group** - wymaga pola *_id*, w którym wyznaczamy po jakich polach grupujemy, a pole *Number* korzysta z funkcji agregacji $sum, która sumuje liczbę nadanych tych samych imion po polu *Name*
- **$sort** - służy do ustawienia rekordów w kolejności rosnącej względem lat z pola *Year* (1 rosnąco, -1 malejąco)
- **$out** - wyniki tej agregacji zostają umieszczone w nowej kolekcji *agr1*



WNIOSKI

Zyskanie popularności: rok 1913 - 2314 dzieci
Ważna data: objęcie urzędu prezydenta Stanów Zjednoczonych przez Woodrowa Wilsona

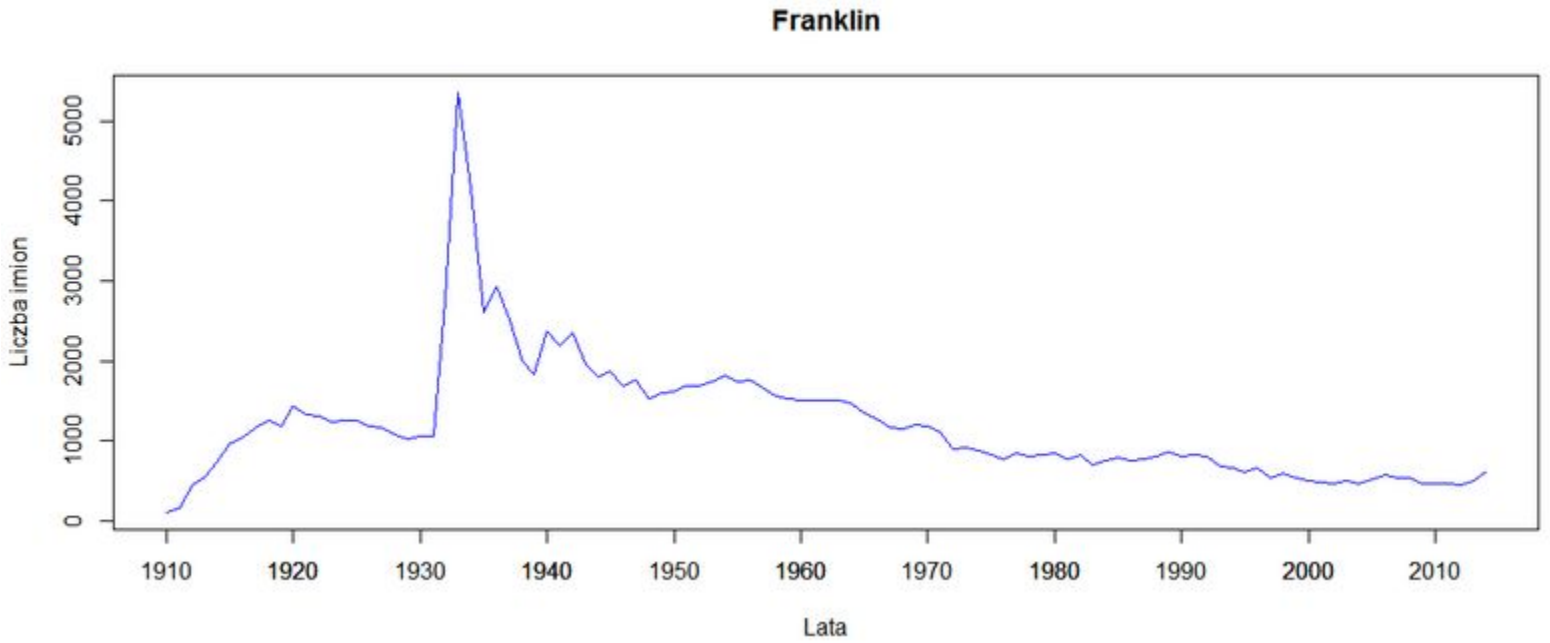
Największa popularność: rok 1918 - 3337 dzieci
Ważna data: wygłoszenie przez Woodrowa Wilsona programu pokojowego na konferencji w Wersalu

FRANKLIN

```
db.names.aggregate(
  { $match: { Name: "Franklin" } },
  { $group: {
    _id: { Year: "$Year" },
    Number: { $sum: "$Count" }
  }},
  { $sort: { "_id.Year" : 1}},
  { $out : "agr2" }
)
```

Wynik:

```
{ "_id" : { "Year" : 1910 }, "Number" : 113 }
{ "_id" : { "Year" : 1911 }, "Number" : 173 }
{ "_id" : { "Year" : 1912 }, "Number" : 456 }
{ "_id" : { "Year" : 1913 }, "Number" : 562 }
{ "_id" : { "Year" : 1914 }, "Number" : 740 }
{ "_id" : { "Year" : 1915 }, "Number" : 967 }
{ "_id" : { "Year" : 1916 }, "Number" : 1053 }
{ "_id" : { "Year" : 1917 }, "Number" : 1164 }
{ "_id" : { "Year" : 1918 }, "Number" : 1257 }
{ "_id" : { "Year" : 1919 }, "Number" : 1186 }
{ "_id" : { "Year" : 1920 }, "Number" : 1435 }
...
```



WNIOSKI

Największa popularność: rok 1933 - 5355 dzieci
Ważna data: objęcie urzędu prezydenta Stanów Zjednoczonych przez Franklina Delano Roosvelta

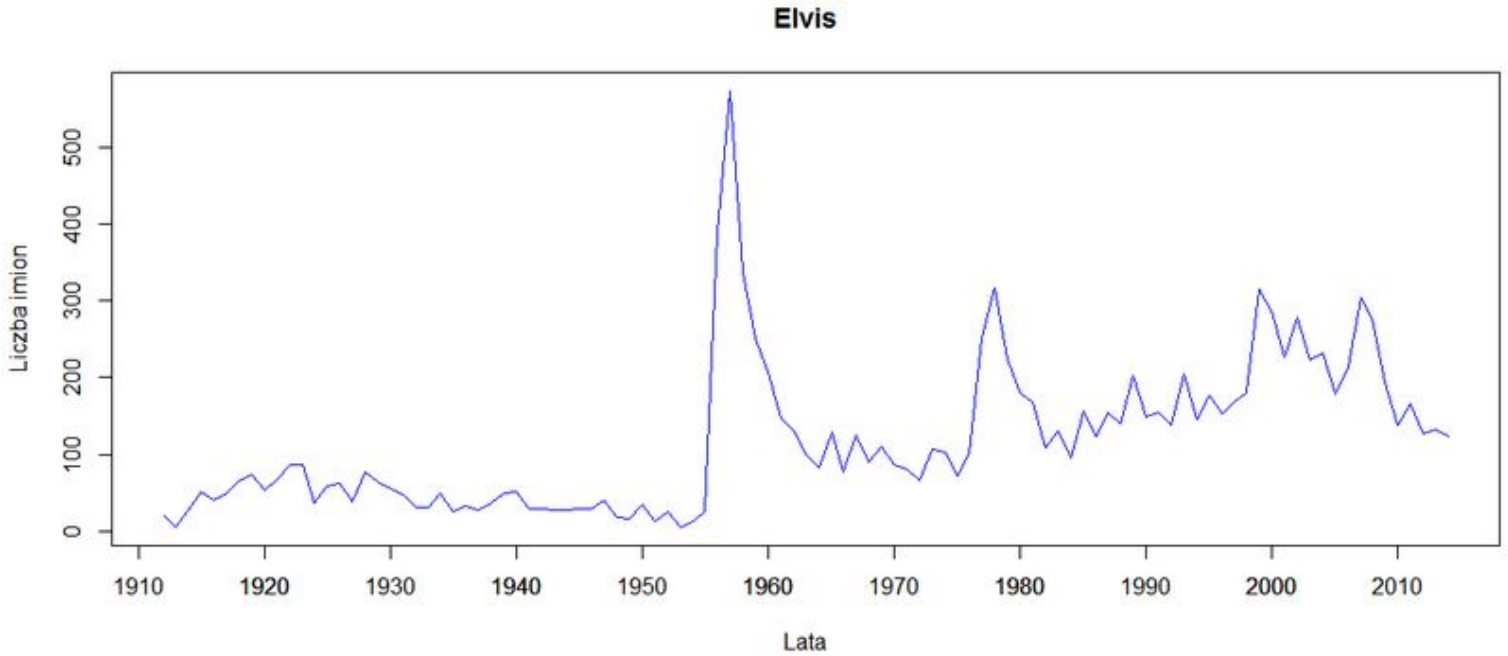
Skoki popularności: rok 1936 - 2933, rok 1940 - 2375
Ważna data: ubieganie się o drugą i trzecią kadencję, obie wygrane

ELVIS

```
db.names.aggregate(
  { $match: { Name: "Elvis" } },
  { $group: {
    _id: { Year: "$Year" },
    Number: { $sum: "$Count" }
  }},
  { $sort: { "_id.Year" : 1}},
  { $out : "agr3" }
)
```

Wynik:

```
{ "_id" : { "Year" : 1912 }, "Number" : 20 }
{ "_id" : { "Year" : 1913 }, "Number" : 6 }
{ "_id" : { "Year" : 1914 }, "Number" : 28 }
{ "_id" : { "Year" : 1915 }, "Number" : 51 }
{ "_id" : { "Year" : 1916 }, "Number" : 41 }
{ "_id" : { "Year" : 1917 }, "Number" : 50 }
{ "_id" : { "Year" : 1918 }, "Number" : 67 }
{ "_id" : { "Year" : 1919 }, "Number" : 73 }
{ "_id" : { "Year" : 1920 }, "Number" : 53 }
{ "_id" : { "Year" : 1921 }, "Number" : 69 }
{ "_id" : { "Year" : 1922 }, "Number" : 86 }
...
```



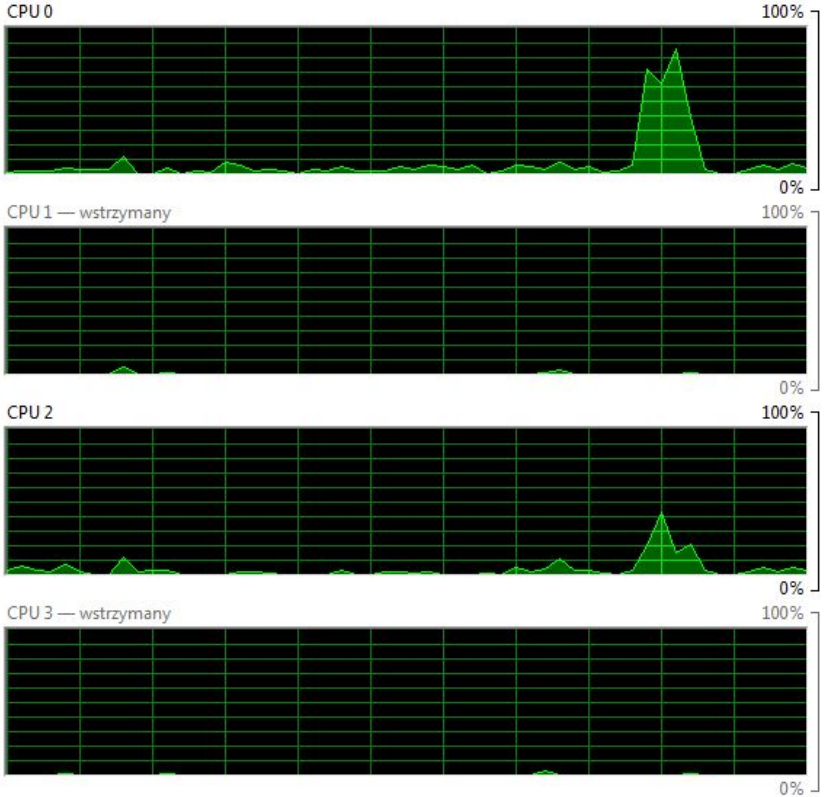
WNIOSKI

Zyskanie popularności: rok 1956 - 389, rok 1957 - 574 dzieci
Ważna data: muzyka Elvisa Presleya zaczęła być rozpoznawana na całym świecie

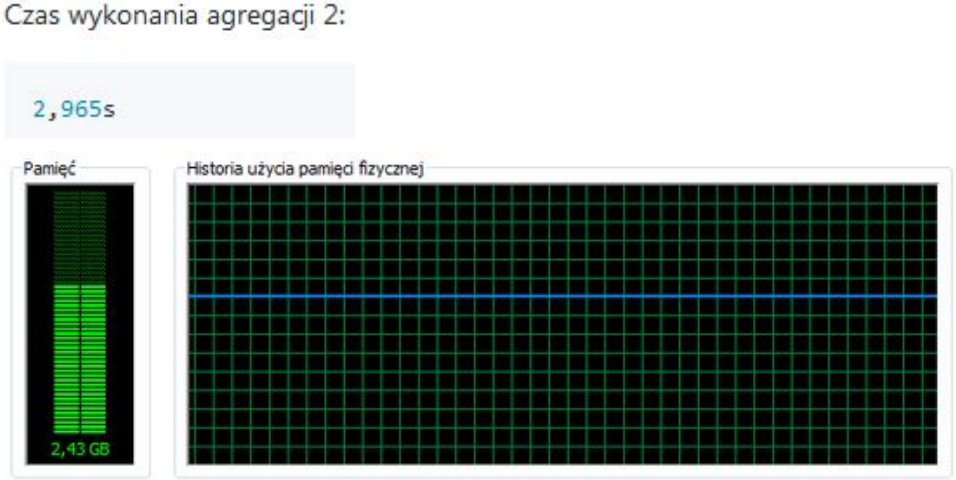
Skoki popularności: rok 1977 - 255, rok 1978 - 317 dzieci
Ważna data: śmierć „Króla Rock and Rolla” w roku 1977

Agregacja 2

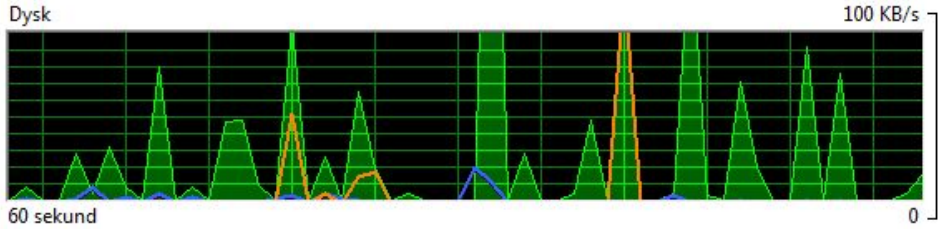
Obciążenie komputera podczas wykonania agregacji



Obciążony głównie rdzeń pierwszy (CPU 0), drugi zdecydowanie mniej. Brak zaangażowania innych wątków do wykonania tej agregacji.



2.43 GB stabilnie przez cały proces. Brak zaangażowania pamięci RAM do tej agregacji.



Delikatne skoki obciążenia dysku podczas wykonywania agregacji. Największy skok podczas eksportowania wyniku do nowej kolekcji.

Agregacja 3

Srednia liczba urodzen

Agregacja ma sprawdzić jaka jest średnia urodzeń dziewczynek i chłopców w każdym roku.

Podobno statystycznie na 100 mężczyzn przypada 108 kobiet.
Postanowiliśmy sprawdzić czy znajduje to odzwierciedlenie w naszym zbiorze.

Płeć	Teoria	Nasz zbiór
Dziewczynki	108	?
Chłopcy	100	?

```
db.names.aggregate(
  { $group: {
    _id: { Gender: "$Gender", Year: "$Year" },
    Suma: { $sum: "$Count" }
  }},
  { $group: {
    _id: { Gender: "$_id.Gender" },
    Average: { $avg: "$Suma" }
  }},
  { $sort: { "Average" : -1}}
)
```

Wynik:

```
{ "_id" : { "Gender" : "M"}, "Average" : 1477269.0571428572 }
{ "_id" : { "Gender" : "F"}, "Average" : 1369238.8095238095 }
```

- **$group** - pierwsze grupowanie wymaga pola *_id*, grupuje względem pól *Gender* oraz *Year*, a pole *Suma* korzysta z funkcji agregacji $sum, która sumuje liczbę nadanych kobiecych i męskich imion w poszczególnych latach po polu *Count*
- **$group** - drugie grupowanie również wymaga pola *_id*, na bazie wyniku poprzedniego grupowania grupuje względem pola *_id.Gender* i za pomocą funkcji agregacji $avg wylicza średnią ze zliczonych sum dla poszczególnych płci
- **$sort** - opiera się o wcześniej utworzone pole *Average* i sortuje malejąco względem tego pola

WNIOSKI

Założeniem agregacji było sprawdzenie czy teoria o rodzeniu się większej ilości kobiet niż mężczyzn jest prawdziwa (100 mężczyzn do 108 kobiet).

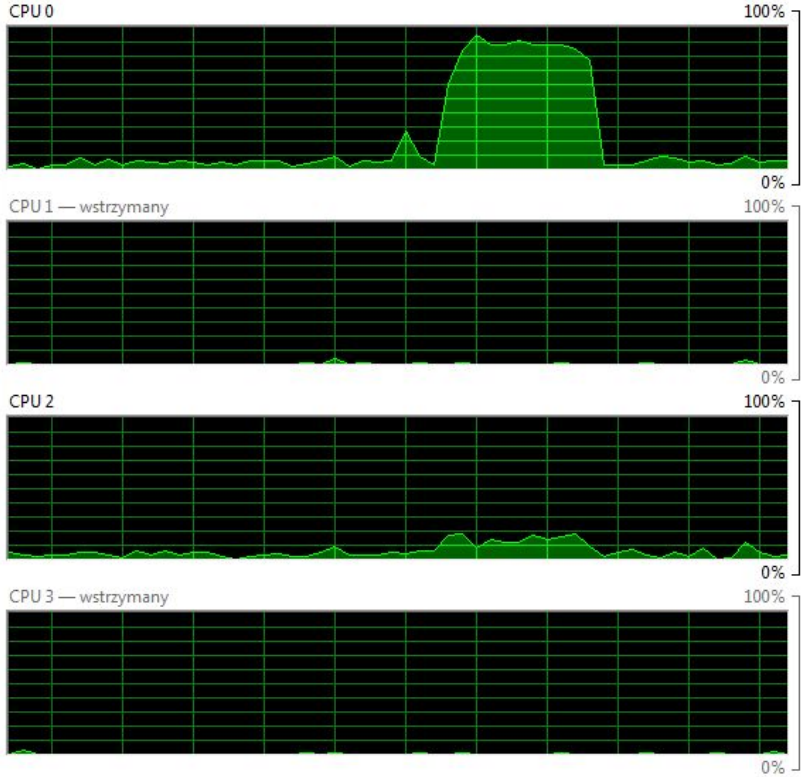
W wyniku otrzymaliśmy ~1 477 269 mężczyzn do ~1 369 239 kobiet rocznie.

Pokazuje to, że jeśli ta teoria jest prawdziwa, to nie znajduje ona pokrycia w Stanach Zjednoczonych, gdyż proporcje wyszły odwrotne, około 108 urodzonych chłopców przypada na 100 urodzonych dziewczynek.

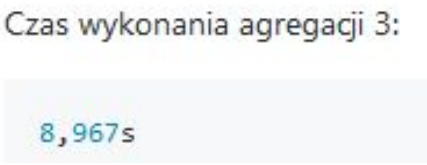
Płeć	Teoria	Nasz zbiór
Dziewczynki	108	100
Chłopcy	100	108

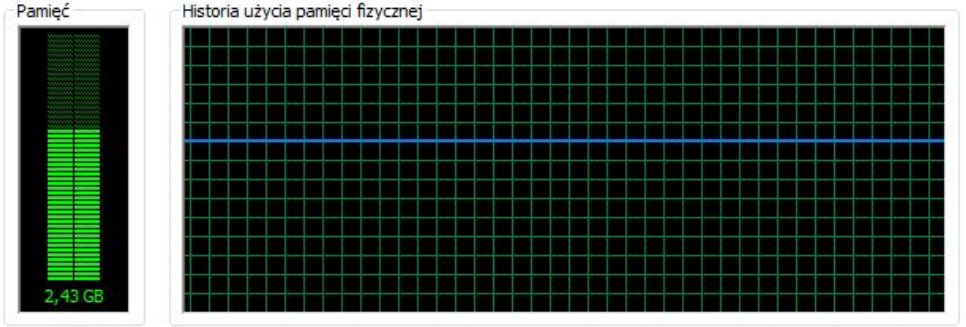
Agregacja 3

Obciążenie komputera podczas wykonania agregacji

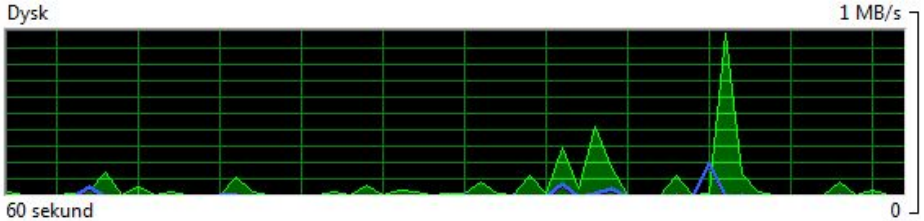


Obciążony prawie całkowicie rdzeń pierwszy (CPU 0), drugi jedynie lekko asystuje. Brak zaangażowania innych wątków do wykonania tej agregacji.





2.43 GB stabilnie przez cały proces. Brak zaangażowania pamięci RAM do tej agregacji.



Delikatne skoki obciążenia dysku podczas wykonywania agregacji. Największy skok podczas zwracania wyniku.

Agregacja 4

Top 2-6 najczesciej nadawanych imion meskich zaczynajacych sie na litere "M"

```
db.names.aggregate(
  { $match: {
    $and: [{
      Name: { $regex: new RegExp(/^M/) },
      Gender: "M"
    }]
  } },
  { $group: {
    _id: { Name: "$Name"},
    Suma: { $sum: "$Count" }
  }},
  { $sort: { Suma : -1}},
  { $limit: 6 },
  { $skip: 1 }
)
```

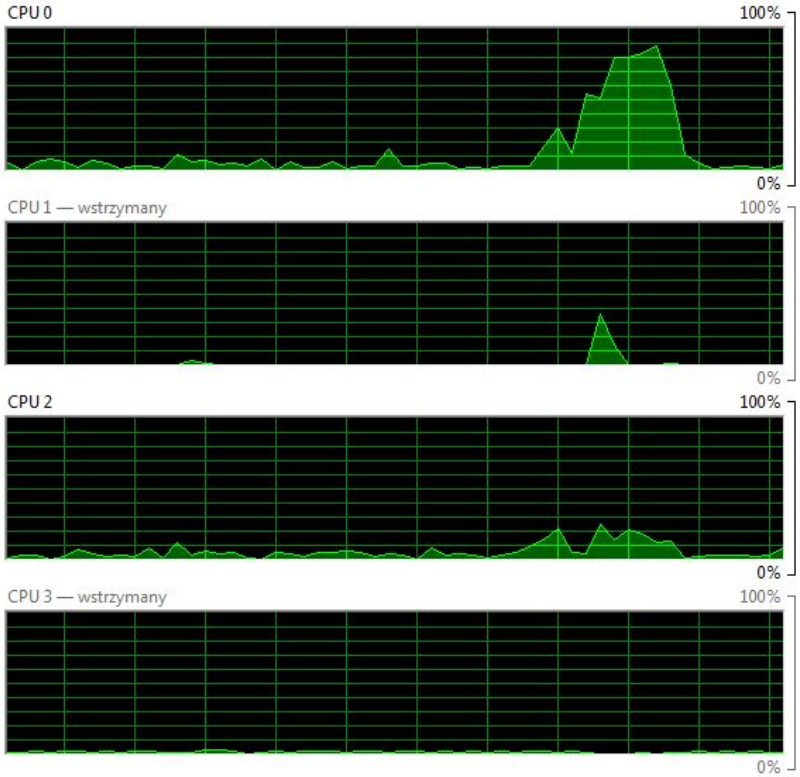
Wynik:

```
{ "_id" : { "Name" : "Matthew"}, "Suma" : 1548424 }
{ "_id" : { "Name" : "Mark"}, "Suma" : 1339737 }
{ "_id" : { "Name" : "Martin"}, "Suma" : 290519 }
{ "_id" : { "Name" : "Marvin"}, "Suma" : 242835 }
{ "_id" : { "Name" : "Melvin"}, "Suma" : 234118 }
```

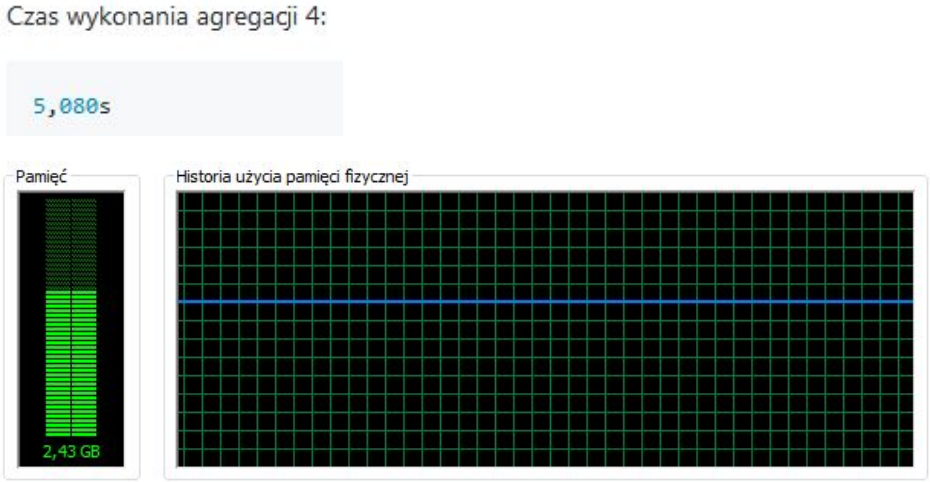
- **$match** - wybiera z bazy tylko te rekordy, których pole *Name* zaczyna się od litery “*M*” oraz płeć jest “*M*”
- **$group** - grupowanie wymaga pola *_id*, na bazie wyniku poprzedniego ograniczania grupuje względem pola *Name* i za pomocą funkcji agregacji $sum zlicza ilość nadanych imion, zapisując pod nowe pole *Suma*
- **$sort** - opiera się o wcześniej utworzone pole *Suma* i sortuje malejąco względem tego pola
- **$limit** - ogranicza liczbę zwracanych rekordów do 6
- **$skip** - pomija pierwszy rekord, nieistotny z punktu widzenia założenia wyszukiwania

Agregacja 4

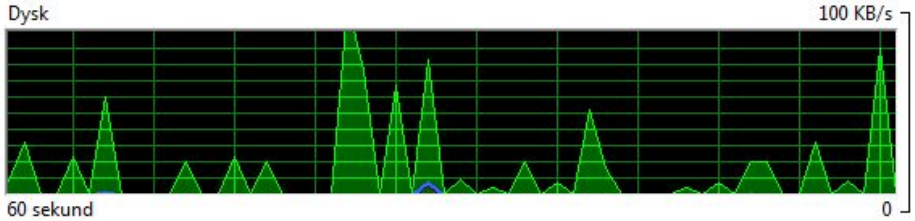
Obciążenie komputera podczas wykonania agregacji



Rdzeń pierwszy (CPU 0) zdecydowanie bardziej obciążony od rdzenia drugiego (CPU 2) podczas wykonywania agregacji, dodatkowo lekko zaangażowany 3-ci wątek (CPU 1).



2.43 GB stabilnie, bez zmian podczas wykonywania agregacji.



Praktycznie bezczynny dysk podczas wykonywania agregacji (delikatne skoki niezwiązane z agregacją).

Agregacja 5

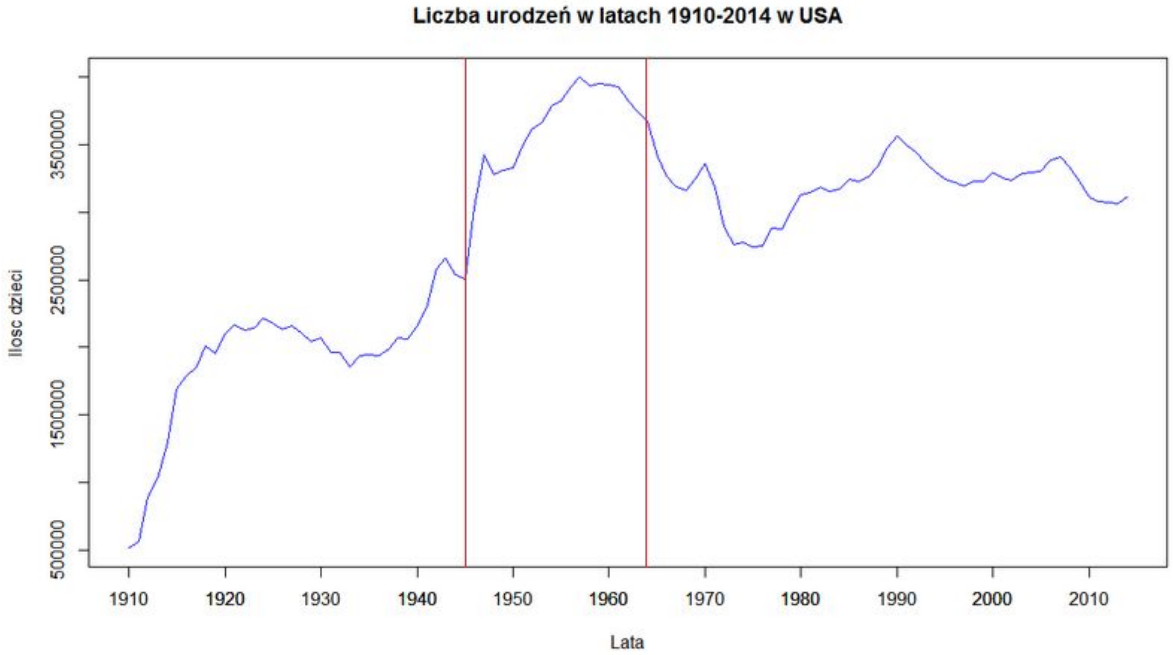
Liczba urodzeN w latach 1910-2014

Agregacja ma na celu sprawdzenie liczby urodzonych dzieci w każdym roku (dziewczynek i chłopców).

```
db.names.aggregate(
  { $group: {
    _id: {
      Year: "$Year"},
      Number: {
        $sum: "$Count"}
    }},
  { $sort: { "_id.Year": 1 } },
  { $out: "agr5" }
)
```

Wynik:

```
{ "_id" : { "Year" : 1910 }, "Number" : 516318 }
{ "_id" : { "Year" : 1911 }, "Number" : 565810 }
{ "_id" : { "Year" : 1912 }, "Number" : 887984 }
{ "_id" : { "Year" : 1913 }, "Number" : 1028553 }
{ "_id" : { "Year" : 1914 }, "Number" : 1293322 }
{ "_id" : { "Year" : 1915 }, "Number" : 1690022 }
{ "_id" : { "Year" : 1916 }, "Number" : 1786510 }
{ "_id" : { "Year" : 1917 }, "Number" : 1855696 }
{ "_id" : { "Year" : 1918 }, "Number" : 2013381 }
{ "_id" : { "Year" : 1919 }, "Number" : 1954834 }
{ "_id" : { "Year" : 1920 }, "Number" : 2101157 }
...
```



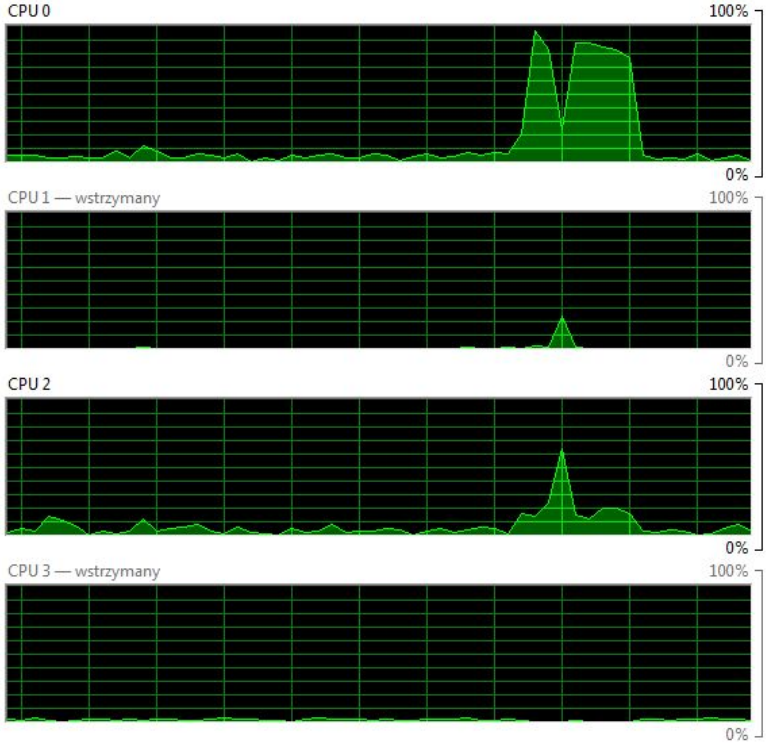
	1964;3674865
	1965;3420034
1945;2501963	1966;3268132
1946;3032350	1967;3184540
1947;3427532	1968;3160072

WNIOSKI

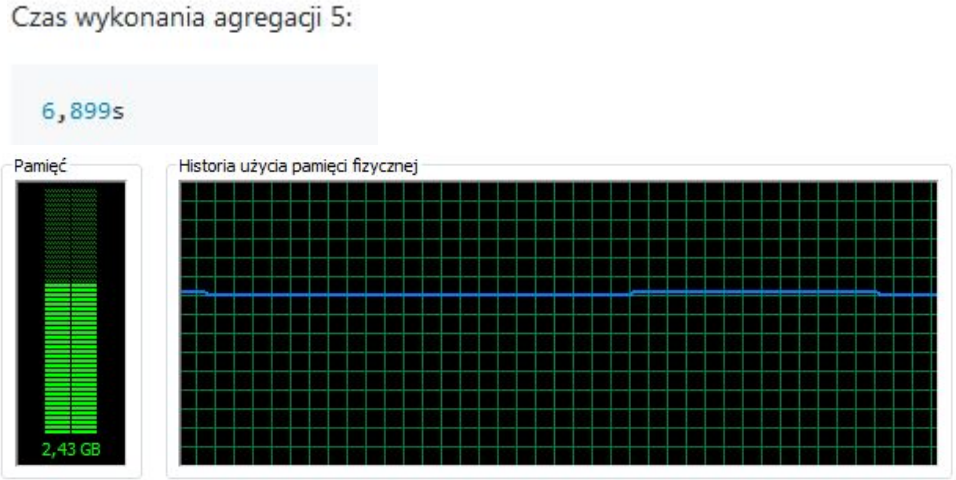
W Stanach Zjednoczonych powojenny wyż demograficzny (tzw. baby boom) datuje się na okres od 1945-1946 do 1964 roku, co znajduje odzwierciedlenie w naszym zbiorze i zostało zaznaczone na wykresie. Gwałtowny wzrost liczby urodzeń następuje w 1946 roku - rodzi się o 530387 (21%) więcej dzieci niż rok wcześniej, w kolejnym 1947 roku więcej o 395182 (13%). Od 1964 roku liczba urodzeń zaczyna miarowo spadać.

Agregacja 5

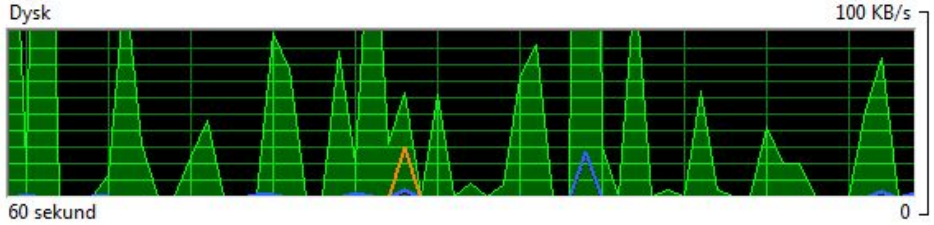
Obciążenie komputera podczas wykonania agregacji



Rdzeń pierwszy (CPU 0) zdecydowanie bardziej obciążony od rdzenia drugiego (CPU 2) podczas wykonywania agregacji, dodatkowo lekko zaangażowany 3-ci wątek (CPU 1).



2.43 GB stabilnie, bez zmian podczas wykonywania agregacji.



Praktycznie bezczynny dysk podczas wykonywania agregacji (jedyny skok podczas eksportowania wyniku do nowej kolekcji).

PODSUMOWANIE UZYTYCH OPERATORÓW I FUNKCJI

Operator	Agregacja 1	Agregacja 2	Agregacja 3	Agregacja 4	Agregacja 5
Czas	10,399	2,965	8,967	5,080	6,899
$group	V	V	V	V	V
$sort	V	V	V	V	V
$match	X	V	X	V	X
$sum	V	V	V	V	V
$avg	X	X	V	X	X
$limit	V	X	X	V	X
$out	X	V	X	X	V
&skip	X	X	X	V	X
$regex	X	X	X	V	X
$and	X	X	X	V	X

ZESTAWIENIE CZASÓW

ORAZ KOLEJNOSC WYKORZYSTANYCH OPERATORÓW I FUNKCJI

Agregacja 1	Agregacja 3	Agregacja 5	Agregacja 4	Agregacja 2
10,399s	8,967s	6,899s	5,080s	2,965s
$group	$group	$group	$match	$match
$sum	$sum	$sum	$and	$group
$sort	$group	$sort	$regex	$sum
$limit	$avg	$out	$group	$sort
	$sort		$sum	$out
			$sort	
			$limit	
			$skip	