# **Benchmarking Java Streams**

### Processamento de Dados com Streams de JAVA

Afonso Silva, a70387 Octávio Maia, a71369

# Conteúdo

١.	Contextualização e objetivos	2
II.	Testes	5
1.	Cálculo dos valores de transações registadas	6
2.	Extração dos primeiros e últimos 20% de transações realizadas	10
3.	Esforço de eliminação de duplicados	14
4.	Comparação entre a aplicação de método estático, BiFunction e Lambda	17
5.	Comparação da ordenação através de um TreeSet e do método sorted	20
6.	Catálogo de transações por Mês, Dia e Hora	22
7.	Comparação do cálculo da soma com ou sem partições	25
8.	Determinação da maior transação entre uma determinada hora	29
9.	Cálculo do total facturado numa determinada semana do ano	32
10	. Cálculo do total de IVA associado a cada mês	35
11	. Comparação de performance entre JDK8 e JDK9	38

# Parte I. Contextualização e objetivos

#### Introdução

#### Execução

Todos os testes deste projeto foram desenvolvidos no mesmo programa, que se encontra hospedado no GitHub:

https://github.com/OctavioMaia/PDSJ TP2

O executável pode ser descarregado diretamente aqui e é corrido desta forma:

```
$ java -jar bjs.jar T
```

onde T é o número do teste que se pretende executar.

O resultado do teste será um tabela CSV que contém na primeira linha os nomes dos **indicadores** executados e na primeira coluna os nomes dos **inputs** indicados. Este CSV pode ser facilmente incluído numa folha de cálculo para melhor análise estatística do mesmo, ou numa table em formato de documento.

#### Implementação

De forma a uniformizar e facilitar a implementação de todos os testes requisitados, optou-se por implementar uma **interface** comun a todos os testes. Deste modo cada teste corresponderá a uma **classe** que implementa esta mesma interface.

Assume-se também que cada teste é composto por um conjunto de indicadores, que mais não são o conjunto dos métodos que o teste deve executar. Para além disso, cada teste pode receber ou não, um determinado **input** (por exemplo o conjunto de todas as transações realizadas, ou o número de números aleatórios a gerar).

```
public interface Test {

/**

* Returns a textual description of the input of this test.

* If no input is given, then it should return an empty Optional.

*

* @return textual description of the input if it exists.

*/

default Optional < String > input() {

    return Optional .empty();
}

/**

* Returns all the indicators that this test should run.

* Each indicator is an association between its textual description

* and the function (Supplier) that it should run.

*

* @return Amap, where the keys are the textual description of the

* indicator, and the values are the supplier to run.

*/
```

```
Map<String, Supplier <?>> indicators();
 st Calculates the times needed to run each one of the indicators.
   @return a map, where the keys are the textual description of the
           indicator, and the values are theirs results in seconds.
default Map<String, Double> results() {
    return indicators().entrySet().stream().collect(Collectors.toMap(
            Map. Entry::getKey,
            entry -> measure(entry.getValue()).getKey()
    ));
 * Measures the required time to run the given supplier.
 *\ Before\ the\ supplier\ is\ ran\,,\ it\ does\ a\ 5\ turns\ warmup\ and
 * cleans the memory.
   @author FMM
 * @param supplier supplier to run
 * @param <R> type of the result of the supplier
   @return An entry composed by the time in seconds that the
           supplier took to run, and the result of that supplier.
static <R> SimpleEntry<Double, R> measure(Supplier<R> supplier) {
    for (int i = 0; i < 5; i++) supplier.get();
    System.gc();
    Crono.start();
   R result = supplier.get();
    Double time = Crono.stop();
    return new SimpleEntry <> (time, result);
```

Com a utilização desta **interface** torna-se então bastante intuitiva a criação de novos testes. A classe principal do programa fica então responsável por ler o número do teste escolhido pelo utilizador e por corrê-lo com o **input** especificado.

Parte II.

**T**estes

# 1. Cálculo dos valores de transações registadas

#### Observações

```
public double sumArray() {
    double[] values = new double[this.transactions.size()];
    int i = 0;

    for (TransCaixa transaction : this.transactions) {
        values[i++] = transaction.getValor();
    }

    double sum = 0.0;

    for (i = 0; i < values.length; i++) {
        sum += values[i];
    }

    return sum;
}</pre>
```

Listing 1.1: Cálculo da soma dos valores das transações atraves de um array do tipo double

```
public double sumDoubleStream() {
    DoubleStream values = this.transactions.stream()
        .mapToDouble(TransCaixa::getValor);
    return values.sum();
}

public double sumDoubleStreamP() {
    DoubleStream values = this.transactions.parallelStream()
        .mapToDouble(TransCaixa::getValor);
    return values.sum();
}
```

Listing 1.2: Cálculo da soma dos valores das transações através de uma DoubleStream

Listing 1.3: Cálculo da soma dos valores das transações através de Stream<Double>

Input	(1) sum Array	(2) sum DoubleStream	(3) sum DoubleStreamP	(4) sum Stream	(5) sum StreamP
1000000 transactions	0,010215	0,010794	0,009364	0,026481	0,015645
2000000	0,021934	0,021800	0,009912	0,053940	0,026226
transactions 4000000	0,044380	0,047799	0,026682	0,113063	0,051981
transactions 8000000 transactions	0,078939	0,089719	0,046436	0,214286	0,126348

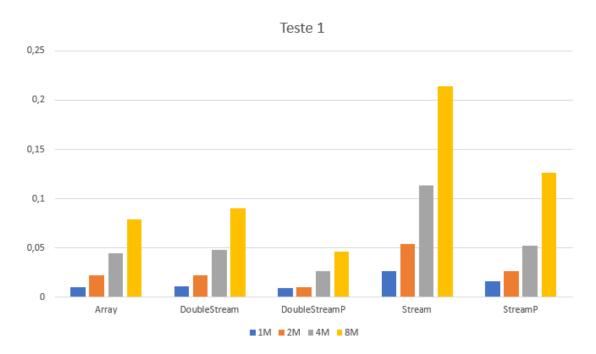


Figura 1.1.: Representação gráfica destes resultados

Após uma breve observação dos gráficos, podemos afirmar que a estrutura de dados mais adequada é de facto a DoubleStreamP que implementa streams paralelas. Em contraste, a pior estrutura em nível de performance é a Stream, sendo até 5 vezes mais lenta que a DoubleStreamP.

# 2. Extração dos primeiros e últimos 20% de transações realizadas

#### Observações

```
public SimpleEntry<List<TransCaixa>, List<TransCaixa>> byDateList() {
   int nelems = 20 * this.transactions.size() / 100;

   List<TransCaixa> sorted = new ArrayList<>(this.transactions);
   sorted.sort(Comparator.comparing(TransCaixa::getData));

   List<TransCaixa> first = sorted.subList(0, nelems);
   List<TransCaixa> last = sorted.subList(
        sorted.size() - 1 - nelems, sorted.size() - 1);

   return new SimpleEntry<>(first, last);
}
```

```
public SimpleEntry<List<TransCaixa>, List<TransCaixa>> byDateSet() {
   int nelems = 20 * this.transactions.size() / 100;

   TreeSet<TransCaixa> sorted = new TreeSet<>(
        // Com este comparador garante-se que a lista fica ordenada
        // e nao se removem os elementos iguais
        (t1, t2) -> t1.getData().isBefore(t2.getData()) ? -1 : 1
);
   sorted.addAll(this.transactions);

   List<TransCaixa> first = new ArrayList<>(sorted)
        .subList(0, nelems);
   List<TransCaixa> last = new ArrayList<>(sorted.descendingSet())
        .subList(0, nelems);

   return new SimpleEntry<>(first, last);
}
```

```
. limit(nelems)
. collect(Collectors.toList());
return new SimpleEntry<>(first , last);
}
```

	(1)	(2)	(3)	(4)
Input	byDateList	byDateSet	byDateStream	byDateStreamP
1000000	1,029137	1,939311	1,861916	2,239986
transactions				
2000000	2,331822	3,706627	$3,\!858658$	4,378531
transactions				
4000000	5,745454	9,411794	$9,\!281572$	$10,\!556521$
transactions				
8000000	12,635500	21,085741	20,579760	22,002980
transactions				

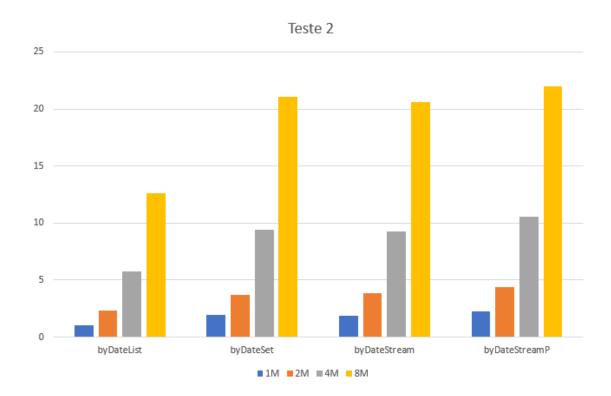


Figura 2.1.: Representação gráfica destes resultados

Após uma breve análise do gráfico podemos concluir que a implementação usando *List* é a mais rápida por uma margem substancial (quase 2x mais rápida que as outras alternativas implementadas). Curiosamente a implementação paralela utilizando Streams, *byDateStreamP* é mais lenta que a sua implementação sequencial, *byDateStream*. Podemos concluir que para os dataset fornecido a paralelização não é benéfica, antes pelo contrário. Por fim, a implementação usando Sets e Streams sequenciais têm performance quase idênticas, sendo a sua diferença em tempos de execução na casa dos 5%.

# 3. Esforço de eliminação de duplicados

#### Observações

```
public Integer[] uniqueArray() {
    Set<Integer> nodups = new HashSet<>();
    for (int value : this.values) {
        nodups.add(value);
    }
    return nodups.toArray(new Integer[nodups.size()]);
}
```

Listing 3.1: Eliminação dos duplicados através de um array de inteiros

```
public Integer[] uniqueList() {
    List<Integer> aux = new ArrayList <>();
    for (int value : this.values) {
        aux.add(value);
    }

    List<Integer> nodups = new ArrayList <>(new HashSet <>(aux));
    return nodups.toArray(new Integer[nodups.size()]);
}
```

Listing 3.2: Eliminação dos duplicados através de uma lista de inteiros

```
public int[] uniqueIntStream() {
    IntStream values = new Random().ints(this.values.length, 0, 9999);
    return values.distinct().toArray();
}
```

Listing 3.3: Eliminação dos duplicados através de uma stream de inteiros

Input	(1) uniqueArray	(2) uniqueList	(3) uniqueIntStream
1000000 random numbers	0,014765	0,027333	0,022026
2000000 random numbers	0,028226	0,055699	0,045615
4000000 random numbers	0,055047	0,112336	0,087037
8000000 random numbers	0,113187	0,296811	0,177688

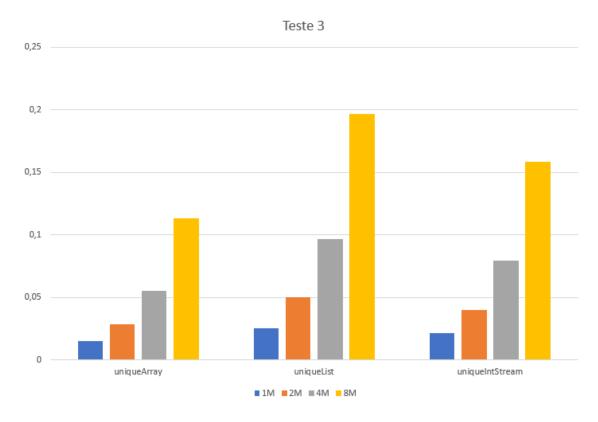


Figura 3.1.: Representação gráfica destes resultados

Após a respetiva geração de 1, 2, 4 ou 8 milhões de números aleatórios compreendidos entre 1 e 9999, procedemos à remoção dos elementos duplicados. A alternativa mais viável para tal feito demonstrou tratar-se do uniqueArray que utiliza um HashSet como estrutura auxiliar. A segunda melhor alternativa seria a uniqueStream que utiliza o operando distinct() seguindo de uma coleção para array. Em último lugar vem a uniqueList, que embora também use um HashSet como estrutura auxiliar como o uniqueArray tem uma performance muito pior, tal que o tempo de execução do uniqueArray para 8 milhões de números aleatórios é igual ao tempo de execução do uniqueList para 4 milhões.

# 4. Comparação entre a aplicação de método estático, BiFunction e Lambda

#### Observações

```
public static int div(int x, int y) {
    return x / y;
}

public int [] divSMethodStream() {
    return Arrays.stream(this.values)
        .map(x -> div(x, 2)).toArray();
}

public int [] divSMethodStreamP() {
    return Arrays.stream(this.values).parallel()
        .map(x -> div(x, 2)).toArray();
}
```

Listing 4.1: Divisão de todos os números por 2 através de um método estático

```
public int [] divBiFunStream() {
    BiFunction<Integer, Integer, Integer> f = (x, y) -> x / y;
    return Arrays.stream(this.values)
        .map(x -> f.apply(x, 2)).toArray();
}

public int [] divBiFunStreamP() {
    BiFunction<Integer, Integer, Integer> f = (x, y) -> x / y;
    return Arrays.stream(this.values).parallel()
        .map(x -> f.apply(x, 2)).toArray();
}
```

Listing 4.2: Divisão de todos os números por 2 através de uma BiFunction

```
public int[] divLambdaStream() {
    return Arrays.stream(this.values).map(x -> x / 2).toArray();
}
public int[] divLambdaStreamP() {
    return Arrays.stream(this.values).parallel().map(x -> x / 2).toArray();
}
```

Listing 4.3: Divisão de todos os números por 2 através de um Lambda

	(1) divS	(2) divS	(3) div	(4) div	(5) div	(6) div
	Method	Method	$\operatorname{BiFun}$	$\operatorname{BiFun}$	Lambda	Lambda
Input	Stream	StreamP	Stream	StreamP	Stream	StreamP
1000000 random numbers	0,004767	0,001324	0,007617	0,004647	0,004802	0,002074
2000000 random numbers	0,008992	0,003859	0,015164	0,007746	0,008965	0,002696
4000000 random numbers	0,018092	0,008287	0,030338	0,017974	0,023628	0,007737
8000000 random numbers	0,036636	0,010449	0,060451	0,032753	0,036349	0,012460

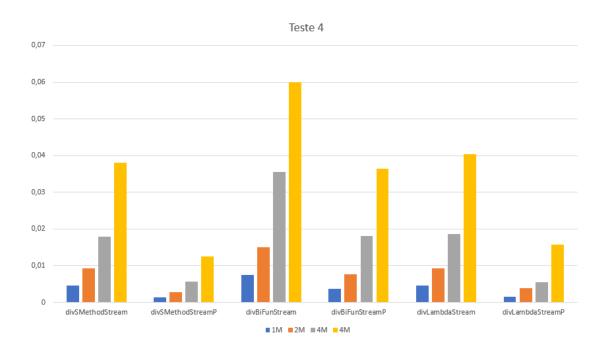


Figura 4.1.: Representação gráfica destes resultados

Embora neste dataset todos os métodos tenham uma performance excelente, na casa das centésimas de segundo, embora a implementação do divSMethodStreamP e divLambdaStreamP sejam as melhores, sendo a sua performance idêntica (diferenças na casa das milésimas de segundo). O pior método de implementação trata-se da divBiFunStream que utiliza uma BiFunction para o cálculo da metade do número em causa.

# 5. Comparação da ordenação através de um TreeSet e do método sorted

#### Observações

Listing 5.1: Ordenação através de um TreeSet

```
public List<TransCaixa> sortList() {
    Comparator<TransCaixa> byDate =
        Comparator.comparing(TransCaixa::getData);

return this.transactions.stream()
        .sorted(byDate).collect(Collectors.toList());
}
```

Listing 5.2: Ordenação através do método sorted

Input	(1) sortTreeSet	(2) sortList	
1000000 transactions	1,510102	0,849679	_
2000000 transactions	3,449318	1,840600	
4000000 transactions	8,242589	4,047586	
8000000 transactions	19,584347	8,851000	

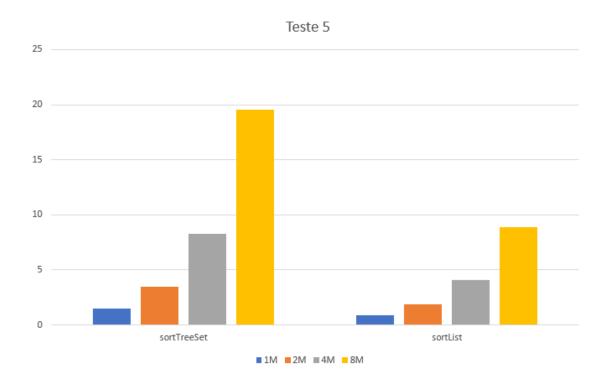


Figura 5.1.: Representação gráfica destes resultados

#### Análise e conclusões

Como podemos verificar pela breve análise do gráfico e tabela, a função sortList aplica o método sorted a uma Stream, tornado a sua performance vastamente superior ao TreeSet usado pela função sortTreeSet. Pelo gráfico gerado a partir da tabela, verificamos que a função sortList é, em média, 2x mais rápida que a sortTreeSet.

# 6. Catálogo de transações por Mês, Dia e Hora

#### Observações

```
public Map<LocalDateTime, List<TransCaixa>> catalog() {
    Map<LocalDateTime, List<TransCaixa>> catalog = new TreeMap<>();

    for (TransCaixa transaction : this.transactions) {
        if (!catalog.containsKey(transaction.getData())) {
            catalog.put(transaction.getData(), new ArrayList<>());
        }

        catalog.get(transaction.getData()).add(transaction);
    }

    return catalog;
}
```

Listing 6.1: Obtenção do catálogo recorrendo a técnicas do JAVA7

Listing 6.2: Obtenção do catálogo recorrendo a Streams

Input	(1) catalog	(2) catalogStream	
1000000 transactions	1,110954	0,283002	
2000000 transactions	$2,\!578776$	0,634677	
4000000 transactions	4,683762	1,300415	
8000000 transactions	11,734155	3,324781	

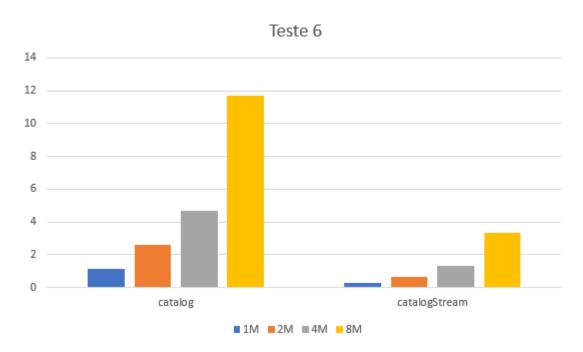


Figura 6.1.: Representação gráfica destes resultados

# 7. Comparação do cálculo da soma com ou sem partições

#### Observações

```
public double sum() {
                        double res = 0.0;
                         \begin{tabular}{ll} \bf for & \tt (TransCaixa transCaixa : \bf this.transactions) & \tt (TransCaixa transCaixa : \bf this.transactions) & \tt (TransCaixa transCaixa transCaix
                                                  res += transCaixa.getValor();
                                                                              return 0.0
                        return res;
public double sumPartition() {
                        int[] pos = {
                                                 this.transactions.size() - 3 * transactions.size() / 4,
                                                this.transactions.size() -2* transactions.size() /4, this.transactions.size() -1* transactions.size() /4,
                                                  this.transactions.size() -0 * transactions.size() / 4,
                         \texttt{partitions.add}(\textbf{this}.\texttt{transactions.subList}(\texttt{pos}\hspace{.05cm}[\hspace{.05cm}0\hspace{.05cm}]\hspace{.1cm},\hspace{.1cm}\texttt{pos}\hspace{.05cm}[\hspace{.05cm}1\hspace{.05cm}]));
                         partitions.add(\mathbf{this}.transactions.subList(pos[1], pos[2]));
                         partitions.add(this.transactions.subList(pos[2], pos[3]));
                         partitions.add(this.transactions.subList(pos[3], pos[4]));
                         // Compute the result
                        double res = 0.0;
                        \begin{tabular}{ll} \textbf{for} & (List < TransCaixa > list : partitions) &
                                                  for (TransCaixa transCaixa : list) {
                                                                           res += transCaixa.getValor();
                        return res;
```

Listing 7.1: Cálculo iterativo sem e com partições

```
public double sumStream() {
    return this.transactions.stream()
    .mapToDouble(TransCaixa::getValor)
```

Input	(1) sum	(2) sum Partition	(3) sum P	(4) sum PartitionStream	(5) sum StreamP	(6) sum PartitionStreamP
1M	0,009782	20,010309	0,020080	0,019041	0,010039	0,010972
2M	0,016814	0.020771	0,038469	0,038853	0,016470	0,016523
4M	0,036431	0,033048	0,079529	0,073569	0,043543	0,030857
8M	0,075378	30,064377	$0,\!152512$	0,144809	0,062982	0,058897

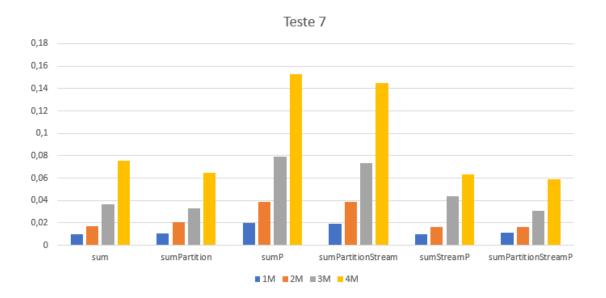


Figura 7.1.: Representação gráfica destes resultados

# 8. Determinação da maior transação entre uma determinada hora

#### Observações

```
String biggestTransaction7() {
    List<TransCaixa> transactions = new ArrayList<>(this.transactions);

transactions.sort(new Comparator<TransCaixa>() {
    @Override
    public int compare(TransCaixa t1, TransCaixa t2) {
        return Double.compare(t1.getValor(), t2.getValor());
    }
});

for (TransCaixa transaction : transactions) {
    int hour = transaction.getData().getHour();

    if (hour >= 16 && hour <= 20) {
        return transaction.getTrans();
    }
}

return null;
}</pre>
```

Listing 8.1: Determinação da maior transação entre uma determinada hora apenas com funcionalidades do JAVA7

```
Optional < String > biggestTransaction8() {
    t.getData().getHour() >= 16 && t.getData().getHour() <= 20;

return this.transactions.stream()
    .filter(timeInRange)
    .max(Comparator.comparing(TransCaixa::getValor))
    .map(TransCaixa::getTrans);
}</pre>
```

Listing 8.2: Determinação da maior transação entre uma determinada hora com auxílio de Streams

Input	(1) biggestTransaction7	(2) biggestTransaction8
1000000 transactions	0,280259	0,017792
2000000 transactions	0,499207	0,035590
4000000 transactions	1,011563	0,071187
8000000 transactions	2,011931	$0,\!142256$

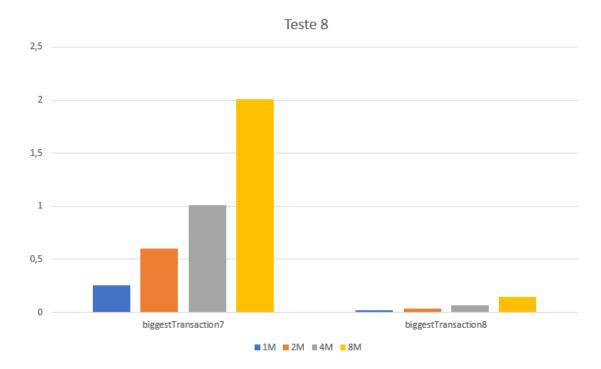


Figura 8.1.: Representação gráfica destes resultados

# 9. Cálculo do total facturado numa determinada semana do ano

#### Observações

```
public double totalInWeekList() {
    final int week = 12;
    List < List < TransCaixa>> by Week = new ArrayList <>();
    // Inicializar cada uma das listas
    for (int i = 0; i < 54; i++) {
        byWeek.add(i, new ArrayList <>());
    for (TransCaixa transaction : this.transactions) {
        byWeek.get(transaction.getData()
                .\ get \ (\ ChronoField\ .ALIGNED\_WEEK\_OF\_YEAR)\ )
                 .add(transaction);
    }
    // Calcular total faturado
    double total = 0.0;
    for (TransCaixa transaction : byWeek.get(week)) {
        total += transaction.getValor();
    return total;
```

Listing 9.1: Cálculo do total facturado na semana 12 do ano

Listing 9.2: Cálculo do total facturado na semana 12 do ano com recurso a streams

Input	(1) totalInWeekList	(2) totalInWeekStream
1000000 transactions	0,033511	0,038533
2000000 transactions	0,070856	0,083489
4000000 transactions	$0,\!144393$	0,163294
8000000 transactions	$0,\!284878$	0,322149

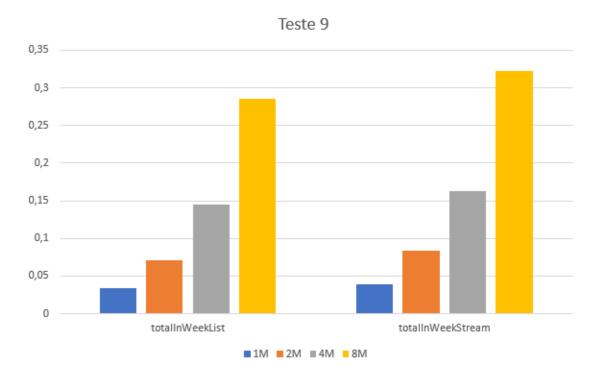


Figura 9.1.: Representação gráfica destes resultados

# 10. Cálculo do total de IVA associado a cada mês

#### Observações

Listing 10.1: Cálculo do total de IVA para cada mês

Listing 10.2: Cálculo do total de IVA para cada mês, com recurso a streams

Input	(1) iva	(2) ivaStream
1000000 transactions	0,027518	0,070057
2000000 transactions	$0,\!058240$	0,134494
4000000 transactions	$0,\!107619$	0,283225
8000000 transactions	$0,\!210951$	0,475746



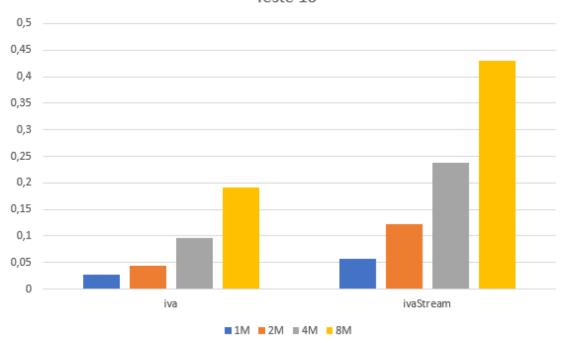


Figura 10.1.: Representação gráfica destes resultados

# 11. Comparação de performance entre JDK8 e JDK9

#### Observações

#### Métodos a testar

Dos testes anteriores, foram selecionados os seguintes, que manipulam conjuntos enormes de registos TransCaixa:

- byDateStream do teste 4;
- sortList do teste 4;
- biggestTransaction8 do teste 8;
- ivaStream do teste 10;

	(1)	(2)	(3)	(4)
Input	sumStreamP	byDateStreamP	sortList	${\it biggestTransaction8}$
1000000	0,013631	1,522605	0,792593	0,019558
transactions 2000000	0,028075	3,481223	1,772310	0,040336
transactions	,	,	,	,
4000000 transactions	0,049482	7,990621	4,034171	0,083384
8000000 transactions	0,099308	17,342737	8,943635	0,162232

	(1)	(2)	(3)	(4)
Input	sumStreamP	by Date Stream P	sortList	${\it biggestTransaction8}$
1000000	0,014094	1,664687	0,844504	0,020016
transactions				
2000000	$0,\!028293$	3,817031	1,939572	0,035174
transactions				
4000000	0,054420	8,699993	$4,\!425399$	0,068976
transactions				
8000000	$0,\!108938$	$19,\!293742$	9,910055	0,138288
transactions				

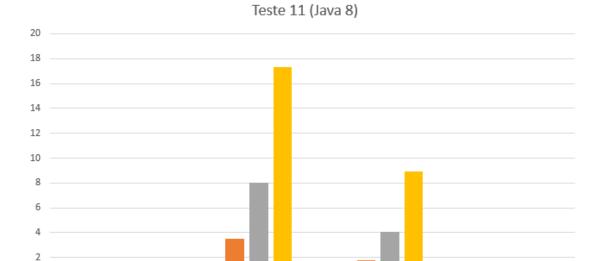


Figura 11.1.: Representação gráfica destes resultados (Java8)

■1M ■2M ■4M ■8M

sortList

biggestTransaction8

byDateStreamP

sumStreamP

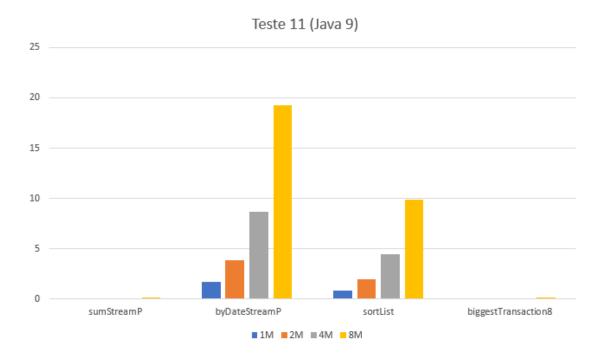


Figura 11.2.: Representação gráfica destes resultados (Java 9)