PONTIFÍCIA UNIVERSIDADE CATÓLICA DO PARANÁ
BIG DATA – BSI
PROF. JOELTON DEONEI GOTZ

MapReduce – Versão Português

Você foi contratado para integrar a equipe de análise de dados de uma grande empresa multinacional. Nessa equipe, é utilizada a tecnologia Hadoop para processar grandes bases de dados utilizando a linguagem de programação Java. No projeto atual, você deverá utilizar o modelo de programação MapReduce para extrair uma série de informações sobre transações comerciais internacionais realizadas pela empresa nos últimos 30 anos. Essas transações estão armazenadas em um dataset estruturado com 10 colunas, conforme a descrição apresentada na tabela abaixo:

Variable (column)	Description
Country	Country involved in the commercial transaction
Year	Year in which the transaction took place
Commodity code	Commodity identifier
Commodity	Commodity description
Flow	Flow, e.g. Export or Import
Price	Price, in USD
Weight	Commodity weight
Unit	Unit in which the commodity is measured, e.g. Number of items
Amount	Commodity amount given in the aforementioned unit
Category	Commodity category, e.g. *Live animals*

O dataset contêm mais 8 milhões de exemplos (ou seja, 8 milhões de linhas que representam as transações comerciais da empresa). Esse dataset é disponibilizado no formato CSV em que as colunas são separadas por ponto e vírgula ";". Na imagem abaixo, são apresentadas as primeiras 5 linhas do arquivo, cada uma com um total de 10 colunas.

```
Afghanistan;2016;010410;Sheep, live;Export;6088;2339;Number of items;51;01_live_animals
Afghanistan;2016;010420;Goats, live;Export;3958;984;Number of items;53;01_live_animals
Afghanistan;2008;010210;Bovine animals, live pure-bred breeding;Import;1026804;272;Number of items;3769;01_live_animals
Albania;2016;010290;Bovine animals, live, except pure-bred breeding;Import;2414533;1114023;Number of items;6853;01_live_animals
Albania;2016;010392;Swine, live except pure-bred breeding > 50 kg;Import;14265937;9484953;Number of items;96040;01_live_animals
```

De acordo com o contexto apresentado acima, você e sua equipe são responsáveis por desenvolver soluções em MapReduce capazes de responder as seguintes perguntas:

1. **(1,0 ponto)** Número de transações envolvendo o Brasil.
2. **(1,0 ponto)** Número de transações por ano.
3. **(1,0 ponto)** Número de transações por categoria.
4. **(1,0 ponto)** Número de transações por tipo de fluxo (flow).
5. **(1,5 ponto)** Valor médio das transações por ano somente no Brasil.
6. **(1,5 ponto)** Transação mais cara e mais barata no Brasil em 2016.
7. **(1,5 ponto)** Valor médio das transações por ano, considerando somente as transações do tipo exportação (Export) realizadas no Brasil.
8. **(2,0 ponto)** Transação com o maior e menor preço (com base na coluna amount), por ano e país.

Para cada um dos itens acima, forneça:

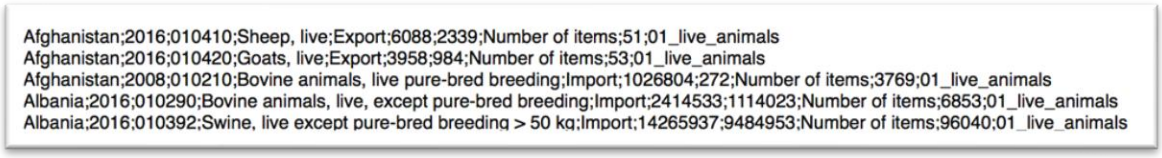
1. Será necessário retirar o cabeçalho.
2. Será necessário tratar dados faltantes.
3. Código fonte para a resolução do problema utilizando MapReduce em Java. **ATENÇÃO**: não serão consideradas como corretas, soluções que realizam a concatenação de strings para a formação de chaves ou valores compostos.
4. O resultado da execução em um arquivo separado e no formato txt.

MapReduce – English Version

You have been hired to join the data analysis team of a large multinational company. In this team, Hadoop technology is used to process large databases using the Java programming language. In the current project, you will need to use the MapReduce programming model to extract a series of information about international business transactions conducted by the company over the last 30 years. These transactions are stored in a structured dataset with 10 columns, as described in the table below:

Variable (column)	Description
Country	Country involved in the commercial transaction
Year	Year in which the transaction took place
Commodity code	Commodity identifier
Commodity	Commodity description
Flow	Flow, e.g. Export or Import
Price	Price, in USD
Weight	Commodity weight
Unit	Unit in which the commodity is measured, e.g. Number of items
Amount	Commodity amount given in the aforementioned unit
Category	Commodity category, e.g. *Live animals*

The dataset contains over 8 million examples (i.e., 8 million rows representing the company's business transactions). This dataset is provided in CSV format, where the columns are separated by a semicolon ";". In the image below, the first 5 rows of the file are shown, each with a total of 10 columns.



According to the context presented above, you and your team are responsible for developing MapReduce solutions capable of answering the following questions:

1. (1.0 point) Number of transactions involving Brazil.
2. (1.0 point) Number of transactions per year.
3. (1.0 point) Number of transactions per category.
4. (1.0 point) Number of transactions per flow type (flow).
5. (1.5 points) Average value of transactions per year only in Brazil.
6. (1.5 points) Most expensive and cheapest transaction in Brazil in 2016.
7. (1.5 points) Average value of transactions per year, considering only export-type (Export) transactions conducted in Brazil.
8. (2.0 points) Transaction with the highest and lowest price (based on the amount column), per year and country.

For each of the above items, provide:

1. The header must be removed.
2. Missing data must be handled.
3. Source code to solve the problem using MapReduce in Java. ATTENTION: solutions that concatenate strings to form composite keys or values will not be considered correct.
4. The result of the execution in a separate file and in txt format.