

Análise de Dados com MapReduce

Pontifícia Universidade Católica do Paraná

Big Data

Prof. Jean Paul Barddal

Implemente as soluções em rotinas MapReduce para extrair o conjunto de informações solicitadas.

Descrição:

Você foi contratado por uma empresa para efetuar uma análise de dados. Esta empresa possui acesso a uma base de dados com dados sobre as transações comerciais entre países nos últimos 30 anos. Sendo que, para cada transação comercial presente nesta base de dados os seguintes campos são fornecidos:

Campo	Descrição
País	País envolvido na transação comercial
Ano	Ano em que a transação foi efetuada
Código	Código da mercadoria
Mercadoria	Descrição da mercadoria
Fluxo	Fluxo, e.g. <i>Exportação</i> ou <i>Importação</i>
Valor	Valor em dólares
Peso	Peso da mercadoria
Unidade	Unidade de medida da mercadoria, e.g. <i>Quantidade de itens</i>
Quantidade	Quantidade conforme a unidade especificada da mercadoria
Categoria	Categoria da mercadoria, e.g. <i>Produto Animal</i>

No total, a base de dados possui mais de 8 milhões de transações comerciais. A base de dados foi fornecida no formato CSV, sendo que cada entrada (transação comercial), é representada por uma linha no arquivo. Enquanto que cada linha possui os campos listados previamente, estes separados pelo caractere “;”. A imagem a seguir exibe as 5 primeiras transações comerciais presentes na base.

```
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
```

Diante desse contexto, você foi encarregado pelo desenvolvimento de um conjunto de soluções que permitam a extração das seguintes informações sobre a base:

1. O número de transações, por mercadoria, envolvendo o Brasil (como a base de dados está em inglês utilize Brasil, com Z);
2. Quantidade de transações financeiras realizadas por ano;
3. Mercadoria mais transacionada em 2016, no fluxo de importação e no Brasil (como a base de dados está em inglês utilize Brasil, com Z); **Dicas:** assumo que a mesma commodity não aparecerá em duas linhas e também use a coluna de quantidade.
4. Média de peso por mercadoria, separadas de acordo com o ano;
5. Média de peso por mercadoria comercializadas no Brasil (como a base de dados está em inglês utilize Brasil, com Z), separadas por ano;
6. Mercadoria com o maior preço por unidade de peso;
7. Quantidade de transações comerciais de acordo com o fluxo, de acordo com o ano;

Com base no seu conhecimento em Java e MapReduce, para cada uma das soluções requisitadas, forneça:

1. **O código fonte do script MapReduce**
2. **O resultado da execução (fornecer apenas os 5 primeiros resultados)**