Professor Eduardo Kugler Viegas
Frameworks de Big Data

Prática MapReduce

Considerando o caso a seguir, implemente as soluções em Python 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. *Exportação* ou *Importação*
Valor	Valor em dólares
Peso	Peso da mercadoria
Unidade	Unidade de medida da mercadoria, e.g. *Quantidade de itens*
Quantidade	Quantidade conforme a unidade especificada da mercadoria
Categoria	Categoria da mercadoria, e.g. *Produto Animal*

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 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. País com a maior quantidade de transações comerciais efetuadas;
2. Mercadoria com a maior quantidade de transações comerciais no Brasil (como a base de dados está em inglês utilize Brazil, com Z);
3. Quantidade de transações financeiras realizadas por ano;
4. Mercadoria com maior quantidade de transações financeiras;
5. Mercadoria com maior quantidade de transações financeiras em 2016;
6. Mercadoria com maior quantidade de transações financeiras em 2016, no Brasil (como a base de dados está em inglês utilize Brazil, com Z);
7. Mercadoria com maior total de peso, de acordo com todas transações comerciais;
8. Mercadoria com maior total de peso, de acordo com todas transações comerciais, separadas de acordo com o ano;
9. Média de peso por mercadoria, separadas de acordo com o ano;
10. Média de peso por mercadoria comercializadas no Brasil (como a base de dados está em inglês utilize Brazil, com Z), separadas de acordo com o ano;
11. Média de peso por mercadoria comercializadas no Brasil (como a base de dados está em inglês utilize Brazil, com Z), em relação ao fluxo, separadas de acordo com o ano;
12. Média de valor por peso, de acordo com a mercadoria comercializadas no Brasil (como a base de dados está em inglês utilize Brazil, com Z), separadas de acordo com o ano;
13. Valor máximo de código;
14. Mercadoria com o maior preço por unidade de peso;

15. Quantidade de transações comerciais de acordo com o fluxo, de acordo com o ano;

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

1. **O código fonte da função de Map**
2. **O código fonte da função de Reduce**
3. **O resultado da execução (forneça apenas os 5 primeiros resultados)**

Dicas

- *Localização dos arquivos*

Máquina	***Base parcial***	***Base completa***
JobTracker	*/home/docker/Desktop/data/operacoes_comerciais/base_100_mil.csv*	*/home/docker/Desktop/data/operacoes_comerciais/base_inteira.csv*
NameNode	*/data/operacoes_comerciais/base_100_mil.csv*	*/data/operacoes_comerciais/base_inteira.csv*
Host	*/home/docker/Desktop/data/operacoes_comerciais/base_100_mil.csv*	*/home/docker/Desktop/data/operacoes_comerciais/base_inteira.csv*

- *Crie uma pasta em /home/docker/Desktop/data/operações_comerciais, com o código das suas funções de map e reduce para submissão da tarefa via TaskTracker, a pasta mapeada é a mesma pasta nos containers*
- *Utilize o arquivo base_100_mil.csv para testar a sua solução*
- *O resultado deve ser obtido sobre o arquivo base_inteira.csv através* ***da execução no cluster hadoop!***
- *Para exibir os resultados ordenados, utilize o comando "sort -k 2 -g -r ARQUIVO"*