PONTIFÍCIA UNIVERSIDADE CATÓLICA DO PARANÁ

BIG DATA

PROF. JEAN PAUL BARDDAL

Big Data Processing with Apache SparkSQL

You and your team were hired to process data using Apache SparkSQL. Your company has access to a dataset with commercial transactions between countries during the past 30 years. For each transaction, the dataset contains the following variables:

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 has over 8 million instances (rows, or commercial transactions). The dataset is made available in CSV format. Columns are separated by semi-colons (";"). The image below exhibits the first 5 rows of the dataset:

```
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
```

Given the aforementioned context, you are in charge of developing a set of solutions that allow the company to answer the following demands using SparkSQL statements:

1. (Easy) The number of transactions involving Brazil;
2. (Easy) The number of transactions per year;
3. (Easy) The number of transactions per flow type and year;
4. (Easy) The average of commodity values per year;
5. (Easy) The average price of commodities per unit type, year, and category in the export flow in Brazil;
6. (Medium) The maximum, minimum, and mean transaction price per unit type and year;
7. (Hard) The most commercialized commodity (summing the quantities) in 2016, per flow type.

Given your knowledge and skills in Python and Apache SparkSQL, for each item above, provide:

1. **The source code for solving the problem using Apache Spark programming and Python**
2. **The result of your code run in a separate text file (.txt). If more than 5 rows of results are available, you must report only the 5 first rows of such result.**

Important:

- **The use of views and standard SQL statements inside SparkSQL is strictly forbidden!**
- **The grading of this activity is conditioned to the audit test.**