4th Phase LocalStore: Reports

(Project not assessable to develop in class)

So far, the sales management system has used a single database that temporarily stored all the information regarding sales made and shipments. Every year, to avoid a deterioration in the performance of the system, an archiving process was carried out by which the oldest information was exported to external files so that the load of information in the database was reduced. Each export file has a first header line that indicates the type of data that corresponds to each column and then comes the data. For each line, a sale and the sales data are stored separated by tabs.

It is not planned to include the old data in the new system due to the cost of integration. However, it has been decided to create a consultation system that allows access to the most relevant information. For this purpose, the export files have been stored in a Hadoop cluster that allows efficient queries when it is necessary to retrieve information from that period. You want to implement a series of mapreduce queries that allow you to obtain the desired reports.

Consultations

Two files will be developed for each query, one for the mapper and one for the reducer. The file names should be mX.py for the mapper and rX.py for the reducer where X corresponds to the query number.

The file must be executable from the terminal, so it will need to start with the line #!/usr/bin/python (or the path corresponding to the python executable).

The localstore.data file will contain the repository of information on which the queries will be made. Querying on the cluster will be simulated through the localstore.data file. Under no circumstances will it be necessary to make inquiries about Hadoop. To simulate the mapreduce, the following command will be used on a Linux kernel system.

```
cat testfile.txt | ./mapper.py | sort | ./reducer.py
```

In Windows there are similar to cat and sort that can be used.

The Python sys library is required to access the information passed through stdin to both the mapper and the reducer: sys.stdin

Please note that there may be errors in the data provided. If an error is found, you must:

- More or less columns: The row is skipped.
- Text strings where numeric values are expected: The row is skipped.
- Blank spaces at the beginning and end of a string: Strings of text with the same content with or without spaces before and/or after the string must be equivalent.

Consultations

1. Total billing from a supplier.
2. Average billing per month for a supplier.
3. Increase per year (in percentage) of a supplier's turnover.
4. Top three suppliers that spend the most on shipments in a given year and total shipping cost for each supplier (Shipping cost: Type 1: €10, Type 2: €5, Type 3: €3.)
5. Sales of higher and lower value made in a given year and suppliers who have made them.
6. Average sales density with type 1 shipping.
7. Different types of shipments made by a vendor in a given month of a given year.
8. Shipments (origin – destination) with the most traffic in a given year.
9. City with the most shipments. Shipment inbound plus shipment outbound.
10. Billing balance for a given city. Sales that are shipped from a city minus sales shipped to that city.