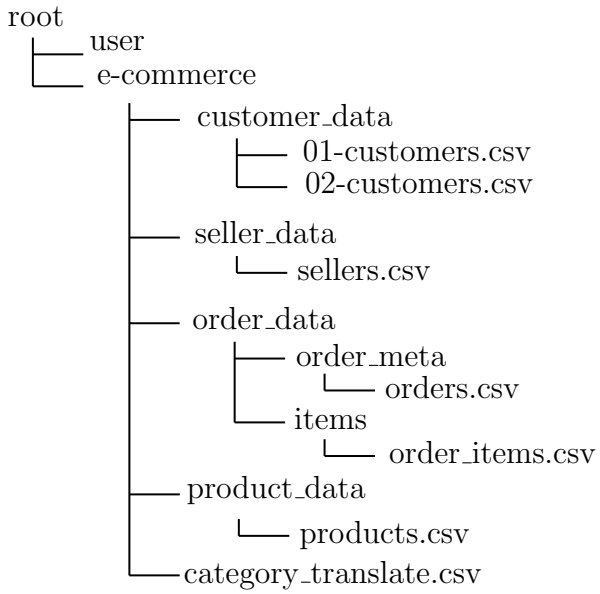
Practise - Big Data Analytics and Reasoning

The *Big Data Analytics Company* wants to analyze its e-commerce sails. Data are stored into their cluster and different formats. Here the file-system tree of their hdfs.



Problem 1

Replicate e-commerce file system tree on your hdfs and define hive tables on top of this data.

Problem 2

Compute the following queries:

- For each customer, find the number of order with at least 2 items
- Find active customers for each year. A customer is active if it has at least three order in a given year.
- For each year and for each customer city, compute the total income for the company (i.e. the sum of the total price of each order)

- Find the three most frequent categories (possibly english) among e-commerce product
- Find for each product the number of sold items and the total income
- Find product category (possibly english) compute of sold items and the total income

Note: implements these queries both using intermediate tables (managed or view or temporary tables) and also directly on external tables.

For this lecture you can clean data using python or whatever tools