Lab 3 - Spark

Big Data Management for Data Science (2024)

Natalia Beltrán, Mikel Gallo

Under the supervision of Besim Bilalli

June 16, 2024

June 16, 2024

A Tasks for Data Management Backbone

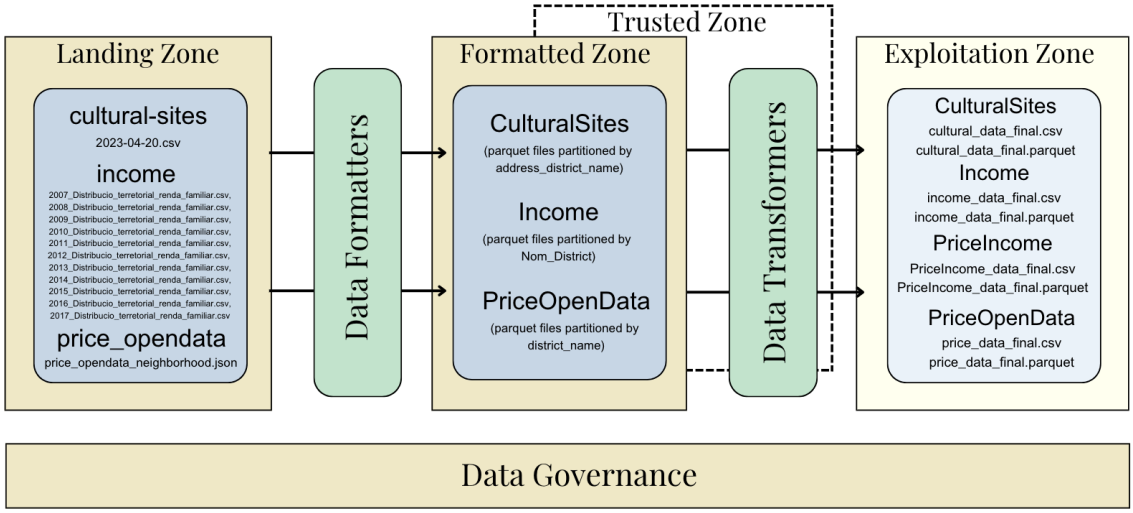

Figure 1: Data Lake Implementation

After a thorough review of the available raw datasets, we selected three sources for our analysis: cultural sites data, income data, and prices data. To set up the data management backbone, we established three local file system directories corresponding to the Landing Zone, Formatted Zone, and Exploitation Zone. We manually placed our raw data into the Landing Zone to initiate the data processing workflow.

Data Formatting Process

Cultural Sites Data:

We formatted the cultural sites dataset by selecting essential columns: addresses road name, values category,values attribute name, addresses type, addresses zip code, addresses town, addresses district name, addresses neighborhood name, and values id. The data was cleaned using a custom comprehensive clean function to resolve text formatting issues. We created a DataFrame from the cleaned data and saved it to the Formatted Zone.

Prices Data:

The prices dataset, provided in JSON format, required restructuring. We extracted details from the 'info' column, flattening it into distinct columns. We created a DataFrame from this processed data. The DataFrame was then saved to the Formatted Zone.

Income Data:

We began by concatenating multiple CSV files into a single file and renamed two of the columns for clarity. A DataFrame was created from this consolidated data. This DataFrame was also saved to the Formatted Zone.

For each DataFrame in the Formatted Zone, we conducted three separate Spark queries to

verify data quality and ensure proper formatting.

Data Transformation Process

During the data transformation process for the cultural sites data, we renamed all columns to ensure consistency with other datasets, facilitating easier data joins. Once renamed, the data was saved to the Exploitation Zone, partitioned by district.A similar process was applied to the prices and income data. The columns were renamed to match those in other datasets for ease of joining. The data was then saved to the Exploitation Zone, also partitioned by district name.

Additionally, we created a new dataset by combining the income and prices datasets. This was done via an inner join on the year and district fields. We saved this combined dataset in the Exploitation Zone as it contains relevant information that could be useful for future analysis. Including it now simplifies potential future analytical processes. For all four datasets, we performed queries to validate the data and ensure proper formatting.

B Tasks for the Data Analysis Backbone

B.1 Descriptive Analysis and Dashboarding

Descriptive Analysis

Based on our exploratory data analysis of the PriceIncome dataset, we identified several noteworthy patterns. Both the amount of meters and the price per meter are skewed to the right, indicating the presence of properties with significantly larger dimensions. This skew likely contributes to the higher prices per square meter. To gain a deeper understanding, we will examine other factors that may influence these differences.

When analyzing the size of the houses and the price per meter across different districts, a strong correlation between these variables and the district location becomes evident. Specifically, Sarria and Les Corts stand out as the districts with the largest houses on average and the highest prices per square meter.

Examining the evolution of price per meter and the size of houses over the years reveals a consistent increase in both variables, with a significant spike from 2016 to 2017. This trend can be attributed to the current housing policies in Barcelona and the increasing demand from tourism, both of which have had a considerable impact on prices.

Additionally, it is important to address missing data in the dataset. The columns diffAmount and diffPermeter are missing almost half of their observations, leading us to decide to drop these columns. In contrast, Used Amount and UserPerMeter have less than 1 percent of missing values, which we will impute using the KNN method.

In terms of correlation, there is a high correlation between the price per meter and the amount (size of the house), with a correlation score of 0.9. This indicates that larger houses tend to have a higher price per square meter, further emphasizing the influence of house size on pricing.

June 16, 2024

Dashboarding

1. Evolution of Population Distribution per District
    - Description: Monitor the size of the population for each district over time.
2. Evolution of Average Price per Meter per District
    - Description: Track the average price per meter for each district over time.
3. Average Price per Meter by Neighborhood
    - Variables: Year, District
    - Description: Analyze each district to determine the best prices per neighborhood.
4. Distribution of the Amount of Meters vs. Amount of Used Meters
    - Variables: Year, District
    - Description: Examine house sizes and their used space in detail.

These metrics will help provide a comprehensive view of population trends, housing prices, and space utilization across different districts and neighborhoods.