Barcelona School of Economics

23D020: Big Data Management for Data Science
Lab 3: Spark

In this lab you will develop two critical backbones of a data-driven architecture: a Data Management Backbone and a Data Analysis Backbone. This will involve creating structured zones within a data lake and performing either descriptive or predictive analysis.

In the training document, we provide instructions for setting up the environment and here we list the exercises to be solved. One group member, in the name of the group, must upload the solution. Remember to include the name of all group members in your solutions. Please check the assignment deadline and be sure to meet it. It is a strict deadline!

Data Management Backbone

Landing Zone

Description: This zone stores raw data that has been ingested into the data lake in a structured or semi-structured format. This could be data directly extracted from source systems without significant transformation. See the Data Sources section below for more details on the datasets that should be assumed to be present in this zone.
Implementation: In a real project this zone would be implemented in a Distributed File System, however for the sake of this project we will assume it is your local file system.

Formatted Zone

Description: This zone stores data in a standardized format, according to a canonical data model (syntactic homogenization). The data could be potentially enriched in a consumption-ready form.
Implementation: An option to implement this zone is by storing data in Parquet files or using Delta files for added benefits such as ACID transactions and schema enforcement.
Location: For this lab, the Formatted Zone can also be on the local file system.

Exploitation Zone

Description: This zone contains processed and refined data (e.g., features, KPI's) that is typically cleaned, enriched, and optimized for analysis.
Implementation: This zone can also be implemented using Delta files, or Parquet files for efficient storage and fast reads. The data can also be stored in CSV files.
Location: For this lab, the Exploitation Zone can also be stored on the local file system.

Data Sources

In the folder datasets, you can find a set of datasets as potential candidates. Most of them are extracted from the Open Data BCN portal[1], and they include:

- Income data
- Idealista data[2]
- Incidence data
- Unemployment data
- Cultural sites data
- Prices data
- Lookup data for Idealista and Income[3]

You are allowed to choose any additional dataset from the BCN portal in case these are not sufficient for your analysis. It is required that you work with **at least three different datasets** in the Landing Zone, and at least one of them should be in JSON format. You may also consider any additional dataset that can later be used as lookup or master data for data reconciliation (i.e., integrating different datasets).

A Tasks for the Data Management Backbone

A.1 Explore the data and choose the KPI's

Study and explore the provided datasets. Choose three datasets to be uses as data for the Landing Zone. Decide which analysis you will perform (i.e., fix the KPIs).

[1] `https://opendata-ajuntament.barcelona.cat/data/en/dataset`

[2] This dataset is downloaded from the Idealista website and contains information about apartments.

[3] This dataset contains reconciled information about neighborhoods and districts to serve as lookup data for joining Idealista and Income.

A.2 Data Formatting Process

Create a directory on your local file system to serve as the Formatted Zone. Write the required Spark jobs to read the raw data from the Landing Zone (i.e. 3 datasets from the local file system), perform necessary transformations, and write data in Parquet or Delta format in the Formatted Zone. Ensure the data is partitioned appropriately for efficient querying.

A.3 Move Data to the Exploitation Zone

Create another directory on your local file system to serve as the Exploitation Zone. Develop a data pipeline using Spark to perform the necessary transformations and cleaning depending on the type of analysis you want to perform, and write data back in Parquet, Delta or CSV format in the Explotation Zone. Ensure the data in the Exploitation Zone is prepared for the analysis that will follow.

A.4 Validate the Data

Write Spark queries to validate the integrity and quality of the data in both the Formatted and Exploitation Zones. Perform basic analytics and print them to ensure that the data has been transformed correctly and is ready for consumption.

B Tasks for the Data Analysis Backbone

You can choose one of the following analysis paths.

B.1 Descriptive Analysis and Dashboarding

Descriptive Analysis

Perform exploratory data analysis (EDA) to summarize and understand the data in the Exploitation Zone.

Dashboarding

Create interactive dashboards using tools like Tableau, Power BI, or Jupyter Notebooks with matplotlib/seaborn to visually tell a ”data story”. Ensure the dashboards provide meaningful insights and support decision-making.

B.2 Predictive Analysis via Model Training and Management

Model Training

Use `spark.mllib` or the `spark.ml` to train predictive models with the data (i.e., matrix) from the Exploitation Zone.

- Create two datasets (training and validation) and format them according to what is expected by the classifier algorithms.
- Create and store at least two classification models together with their corresponding hyperparameters and evaluation metrics over the validation sets. The evaluation metrics can be the traditional ones like predictive accuracy and recall.
  - Use a framework for model management. See the description below.
- Rank the models by their performance in the validation set (e.g., using predictive accuracy).
- Automatically deploy the model that performs best (i.e., set the model with the highest predictive accuracy as the default model to be used when performing predictions with new data).

Model Management

- It is recommended to use the MLflow[4] or any other Model Management framework to manage the data analysis pipeline (e.g., to store/-track the models, hyperparameters and evaluation metrics), as well for tagging the best performing model which will be deployed.

C Bonus Point

Use an orchestration mechanism/framework to orchestrate the Spark Jobs.[5]

[4] `https://mlflow.org`

[5] Apache Airflow: `https://airflow.apache.org`

Note 1: Even if the analytical aspects of the problem fall out of scope (e.g., we are not searching the absolute best predictive model, instead we want to try a couple of models and store the data associated with them), we do consider if you follow good practices when preparing the data to feed the analysis.

Note 2: Even if this lab is a simplified version of what would be required in a realistic case, it is still representative. In the following we list a set of characteristics that can be expected in practice:

- Typically, the features used in the predictive are not decided beforehand. Instead, first a model is created and interpreted. Then, according to this interpretation, these pipelines are iteratively modified to consider new features and fine-tune the model. However, this part of the problem falls out of scope of this lab and we set the features beforehand.
- For the descriptive analysis, typically a data warehouse (i.e., a cube) is created in the Exploitation Zone.
- The files are not stored in the local file system. Instead, a distributed file system or an appropriate DBMS is typically deployed.

Deliverables

1. Three Python scripts or notebooks corresponding to each pipeline described above. The scripts must be included in a single zip file.
    - The Python code must include comments to facilitate the rationale you followed. At the header of each file, include an overall comment explaining WHAT are the steps implemented in the pipeline, and refer to these steps when explaining the Spark code in the subsequent comments.
    - The execution of the three pipelines should be facilitated. For instance, the code should not include absolute paths or fixed user credentials (e.g., they should be requested by the user or stored in configuration files).
2. A PDF file (max three A4 pages, 2.5cm margins, font size 12, inline space 1.15) with all assumptions made and justifying the decisions you made (if any).
    - Use one page to sketch the three pipelines at a higher abstraction level (i.e., you can group a couple of related Spark operations into one single box). Use the notation that you find more appropriate.

- In the rest of the document you need to explain the selected datasets and the analysis you chose to perform together with the necessary justification. You should also elaborate on any assumption not stated in the lab statement but that you followed. This can be done for each one of the pipelines and should refer to any specificity of your solution that should help to understand the decisions you made in your code.

Assessment criteria

i) Conciseness of explanations

ii) Understandability

iii) Coherence

iv) Soundness