Data engineer case study

Introduction

The analytical team is trying to analyze the housing market and decided to enhance their models with publicly available data from advertising portals.

Data has been already scrapped from portals and are stored as text files in html format within the data lake.

Assignment

Analytical team requires the following data to be extracted[1]:

- price value of `class="norm-price ng-binding"`
- location value of `class="location-text ng-binding"`
- all parameters labels and values in `class="params1"` and `class="params2"`

You are in the role of a data engineer responsible to extract needed information from the unstructured data, clean and transform it into structured dataset(s) that will be stored in a relational database.

The target platform for both analysts and data engineers is Databricks (Apache Spark).

Output specification

Create a data pipeline solution that would extract the required information from text files.

The solution should contain a diagram of the target relational data model, a working code that loads, parses, cleans and extracts the information from html files into dataset(s).

The target dataset(s) should be written into a structured file format of your choice.

It should also provide a short description of how the code is intended to be run in the target environment with its dependencies.

You can select the language of your choice that can be executed on Apache Spark (Python, Scala, Java, SQL, .NET), use any publicly available libraries and solve the problem the way you think is the best in your opinion and you are comfortable with.

[1] Attribute of CSS class via which the raw information can be located within the file specified using courier font family

The only limitation is that the code should be runnable on Apache Spark[2] and try to reflect that the solution should be processing large amounts of data.

Environment specification

If you are familiar with Apache Spark, you can either install a standalone Spark on your local PC[3], use Databricks Community edition or utilize your own cloud trial or paid instance.

There is a manual on how to get started with Databricks Community edition as part of case study materials that you can use if you want to.

Beside that you are free to choose any IDE, operating system or tools you like.

Data

Data for processing are included within the compressed data.zip file.

Evaluation criteria

Your solution will be evaluated based on following criteria:

1. Overall solution design
2. Code conciseness
3. Effectivity
4. Reusability of the solution
5. Easy of use
6. Performance
7. Parallel execution
8. Ability to justify the solution
9. Completeness

[2] In case you are not familiar with Apache Spark, write the code in plain Scala, Java or Python.
[3] https://spark.apache.org/docs/latest/api/python/getting_started/install.html