Hello,

As mentioned earlier during classes and via the announcements on Canvas, the exam will be a 'take home exam with verbal explanation'

The assignment:

We've received the following business question from a customer that should be the base of all the further steps:
"How much revenue did we generate this year, and how is it distributed across products, customers, order statuses, and geography?"

To solve this business question you'll need to create a medaillon architecture with their required pipelines. Make sure you'll pay attention to all topics that are mentioned during classes like SCD, Keys, Importing functions, code efficiency, ...

Sourcefiles:

All the sourcefiles that we've received from the customer are on Canvas in the files section. You can find the files in the 'Take Home exam source files' map.

Allowed technology:

Google Colab
Code has to be written in PySpark
For the plots use Spark (or pandas/matplotlib)

Folderstructure:

Please use the following folderstructure to put all the results per layer.

/Raw
/Bronze
/Silver
/Gold

Plots:

- Revenue by Category (all)
- Top-10 Subcategories
- Top-10 Customers
- Revenue by Order Status (all)
- Top-10 Countries revenue
- Top-10 States revenue
- Top-10 Cities revenue

Deliverables:

- 1 notebook or script with all the steps from Raw to Gold in .ipynb
- Per step you'll add a comment with explenation and show a dataframe
- 7 plots
- A powerpoint presentation with the following content:
  - Your overall approach to solve the problem.
  - The data models you've designed for each layer.
  - Seven (7) relevant visualizations (plots).
- Code and presentation should be delivered on canvas before Sunday, June 8th at 12:00PM.
- Presentation length will be 10 minutes
- Questions 5 minutes

Points of attention:

- Don't forget to implement
  - SCD
  - Keys
  - Date dimension
  - Writing to Gold should be done using Upsert