

Please do these task no more than 3 days:

1. Create a pipeline flowchart diagram that consist of at least:
  - a. Data Source / Data Lake with DB (stack/tool) name that you already have experiences at (e.g. API, FTP, Elasticsearch, etc.)
  - b. ETL tools that you already have experiences at (e.g. Airflow, Talend, etc.)
  - c. Data Warehouse that you already have experiences at (e.g. MySQL, Bigquery, etc.)
  - d. Visualization/Dashboard that you already have experiences at (e.g. Tableau, Kibana, etc.)
  - e. NOTES: Just write the tagging symbol for part that you do not have any experience with yet (e.g. Data Source, Viz Tools, etc.)
2. Provided a table "Warehouse" as shown as below:

itemid	category	date	price	sold	stock	priority
23		2021-08-01	5000	5	95	High
23	medior	2021-08-02	5000	6	94	
23	medior	2021-08-03				High
55	lower	2021-09-10		12	88	High
55	lower	2021-09-11	1000	22	78	High
55		2021-09-12	1000			High
10	upper	2021-08-15		32	68	Low
10	upper	2021-08-16	3000	32	68	

Create SQL (preferably query for MySQL or PostgreSQL) to fill NULL values that meet these conditions:

- a. Every items must have single category and exact one priority
  - b. Price of item cannot be changed
  - c. Assume the number of item sold is exactly same as the previous day for missed one
3. Build a script (preferably using python / R and could be in IPython Notebook .ipynb format) that could convert/transform the file "**Logic Test - Raw Data.csv**" to be like "**Logic Test - Result Data.csv**".  
Please, write down your logic or full steps to produce the Serial Number from the given pair of Number and Letter ! (Those files are attached)

4. As general stated in machine learning, focused on terms of building a model:
  - a. What is the main difference between *underfitting* and *overfitting* ?
  - b. Provide simple *training dataset*, *validation dataset*, and *testing dataset* that could possibly lead to the **overfitting** model ones !  
(No more than 9 distinct data for each of dataset, means only 27 distinct data in total at most is allowed)
  - c. Explain in simple words (for being informed to non data-geeks and non IT-guys) what is regularization? And the differences between L1-regularization & L2-regularization?
5. Tell me at least **one thing** (could be a *dark secret* or *thrilling secret*) in Data Science that most people do not know yet these days !



insignia