SQL & NOSQL

Tasks:

1. Setup a Pod or two isolated containers comprising (HINT: Docker Hub):
    a. a relational SQL database
    b. a NOSQL database
2. Import testset_B.tsv into the relational DB and calculate the following KPIs with SQL commands:
    a. Ranks based on column Price, grouped by column brand
    b. min and max of column HDD_GB
    c. median of column GHz, grouped by column RAM_GB
3. Represent the results of 2) in the NOSQL database
4. Commit all code & results to a Git repo.

Spark

Tasks:

1. Use the latest version of Apache Spark (HINT: Docker Hub)
2. Create two artificial datasets "A" & "B" in pySpark. Hint: set a seed.
    a. Dataset 1:
        i. initialize a data frame with 10k rows and 2 columns "id" & "price"
        ii. create variables: "id" (int) & "price" (double) from random draws from appropiate distributions
    b. Dataset 2:
        i. initialize a data frame with 10k rows and 2 columns "id" & "sales"
        ii. create variables: "id" (int) & "sales" (int) from random draws from appropiate distributions
3. Save the two datasets as flat file "result"
4. Load "result" into pySpark, merge by "id" and save as flat file "merged-result"
5. Commit all code & results to a Git repo.

Analysis 1

Dataset C: testset_C.csv
Dataset Description:
Dataset C is supposed to contain records with article texts and the belonging product group. The following information is known about the columns:

Column	Info
id	A unique record identifier
product group	Product category
main_text	a describing text about the article
add_text	an additional describing text about the article
manufacturer	the manufacturer belonging to the article

Tasks:

1. Create a machine learning modell in order to predict the product category based on appropriate features. Use a machine learning algorithm of your choice.
2. Present the result in a vivid way (e.g. in a Jupyter notebook) and explain your model from a statistical PoV.
3. Create a web service on top of your model which obtains an article text and predicts the product category.
4. Commit all code & results to a Git repo.

Analysis 2

Data:

House Price in $1000s (Y)	Square Feet (X)
245	1400
312	1600
279	1700
308	1875
199	1100
219	1550
405	2350
324	2450
319	1425
255	1700

Tasks:

1. Describe/Explain an efficient procedure that predicts house prices (Y) by square feets as input parameter (X).