Hand-in exam

Big Data Management 2022

This is the exam for Big Data Management, Fall 2022. You will get your grade based on your written answers to the questions below. The hand-in is a report submitted on LearnIt as a *pdf file* by:

22 December 2022, 14:00.

The exam is composed of four questions, with multiple subquestions, each counting 10% of the final grade. Including figures, plots, or drawings to explain your thoughts is useful and will be marked higher; you can use any source to answer the questions, besides the lecture notes and reading material; if you use quotations or figures other than your own, please make that clear and *cite the source*. Each subquestion has a limit of half a page (250 words, excluding figures). Similarly to the assignments, you will not lose marks for going over the page limit if it is not extreme; it should give you an intuition on how much detail you need to include.

Submission requirements:

- *one pdf file*
- *title: your_name.pdf*

A) Pipelines (10%)

In the first assignment you had to implement a ML pipeline for a wind energy prediction system.

- Which steps does your *sklearn* pipeline include? Show a pipeline in code and briefly explain each step. Show your results for 2 experiments (different models/the same model with different hyperparameters/different preprocessing steps).

B) Scalable processing (20%)

For the second assignment you ran Spark queries on the Yelp Academic Dataset using a cluster.

- Explain how you found the businesses that have been reviewed by more than 5 influencers. Show both your code and the results, including the number of businesses.
- Did you find a difference in the amount of authenticity language used in different geographical areas?

C) ML lifecycle (20%)

In the third assignment you designed a wind power prediction system that did model selection by experimental evaluation and packaged the final model in a format that ensures reproducibility.

- Explain what parameters you logged and why. Show your results.

- Imagine that you were developing a wind prediction model for each grid connected turbine in Denmark (6-7000 in total). The turbines have independent datasets and diverse scales and conditions. How would you set up this experiment in a distributed manner using MLflow?

D) Lecture material (50%)

- Define the concepts *data lake* and *data warehouse* and describe the differences between the two.
- Explain the differences between ETL and ELT pipelines; give practical examples for when you would use one over the other.
- Give an overview of one computational method for processing high-dimensional data that we discussed in class (BFR, PCA, or LSH). Give an example when it would be applicable.
- Suppose you would like to build a predictive maintenance system. What type of data architecture discussed in class would you use and why?
- Describe 3 challenges of working with big data. These can refer to any aspect of distributed systems, running algorithms in parallel, privacy, bias, deep learning, edge computing etc.

E) Bonus question (extra 5%) - describe your favourite concept, tool, technique, or algorithm from the class (that you haven't already discussed in the previous questions). Mention why you think it is important in the context of big data.