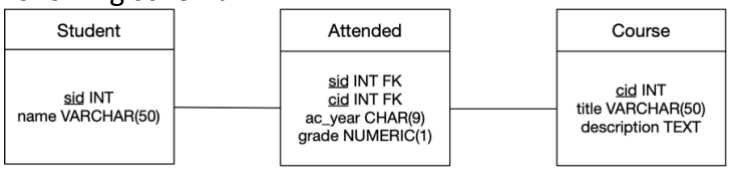
Assignment

MSc in Business Analytics
Advanced Topics in Data Engineering
(Individual assignment)
Deadline: Thu July 7th 2022, 23:59

Task 1 [25 points]
Among Hive, Impala, and Drill, which is the one the implements more precisely the concept of data virtualization? Elaborate.

Task 2 [25 points]
You started working for a large bookstore company. Your client has a large data center containing data in various formats. More specifically, all client data (e.g., personal information, orders) are stored in a Mongo DB database, e-books are stored on HDFS, and social media metadata (likes, ratings, reviews) are stored in a Hive database. They would like to simplify the queries used by various User Interface elements. What would you suggest for their case? Elaborate.

Task 3 [40 points]
Another client of yours has an Impala database for their needs. They want to have a new database with the following schema:



Please provide detailed answers to the following:

- 3a) Create the Impala database & the required tables.
- 3b) Give an example command that inserts an entry to the Student table (use your own details for that entry).
- 3c) Write a statement that retrieves all the names of the students that have attended the course having title “Artificial Intelligence” during the academic year “2021-2022”.
- 3d) Write a statement that retrieves the titles and the average grades of all the courses for which the average grade of the students that attended them is lower than 6.

Task 4 [10 points]
A particular query in the previous Impala database is too slow. Describe what you are going to do to investigate what is going wrong and what can be done to improve efficiency. Provide any commands that you are going to run.

GOOD LUCK!