# Data Management III – ETL and Smart Data Exam Guidelines and Requirements

* Each student will be given a dataset and it is expected that you use it for each step described below and not any other dataset\*.
* The exam will be presentation + demo, each student will be allocated 15 min + 5 min question answer session.
* The exam will be on 13th, 14th and 15th November from 14:00 to 18:00. 15th Nov is reserved for the Part-time students. The exam slot assigned to the student will be made available in Moodle.

**Step 1 Deducing the question**

* The student has to profile the dataset and find a relevant and appropriate question that can be answered using any of the data analytics techniques.
* At least one question is necessary for the exam. You can find and present any additional questions that could be answered using this dataset.
* Until 16th October Wednesday 23:55, each student needs to upload the document in Moodle with the question that will be addressed via their dataset. I will verify the question and if there is any discrepancy in the question then a Skype call can be scheduled with me until 20th October Sunday.
  + Till 20th October Sunday each one of you should be finalized with the question.

**Step 2 ETL Workflow**

* The student has to individually design and implement an ETL workflow using any tool or programming language
  + Any data quality steps necessary for curating/transforming the data must be part of the workflow, i.e., they should be automated and not performed manually external to your primary ETL workflow.
  + The result of your ETL workflow must be at least a Start schema or a Snowflake schema containing a Fact table and arbitrary number of Dimension tables.

**Step 3 Data Analytics**

* Applying data analytics techniques using any programming language
  + You will require the ETL result generated in Step 2 and apply the appropriate data analytics technique to answer the question that you have proposed in Step 1.

**Step 4 Submission**

* Submitting your results in Moodle: Your results have to be reproducible, thus make sure that you submit,

1. the complete ETL workflow source code,
2. the source code for data analysis,
3. the raw dataset (only if you have made some modifications on the raw dataset) or have selected a dataset other than the allocated one.