# ESOF-3675 Database Systems

## Group Project: Data Mining, Database Design, and Implementation

**Presentation Date:** March 27 during class time
**Report Due Date :** March 27, at 11:59 PM

**Objective:** This project aims to develop a complete database system based on a non-RDBMS (NoSQL) and utilize its advanced features of fast massive data querying.

**Task:**

1- **Data Collection:** The first part of this project is to find a proper large-scale dataset to build your application on.

a. You may choose a real application from the local industry/businesses/ administrations or a generic (well-known) application. A few possible applications are as follows:

- Smart house energy consumption predictive analytics. (#1)
- A sports statistics application. (#2)
- Doctor appointment and disease predictions. (#3)
- Movie reviews analytics. (#4)
- Weather forecasting. (#5)
- Student feedback review system. (#6)
- Music genre classification. (#7)

For instance, in application #1, the dataset will be household energy consumption patterns (context-rich data of smart houses), which provide great insight into consumers' energy consumption behavioral traits. The goal is to extract information that can be used for predictive analysis of power consumption patterns for a high-level task. In the case of application #1, you can use the Almanac of Minutely Power dataset Version 2 (AMPds2), which is a dataset collected from a smart house in Vancouver over a long period of time (more than 2 years) and provides 11 measurements characteristics for electricity.

Publication reference of Ampds2: Makonin, S., Ellert, B., Bajić, I. V., & Popowich, F. (2016). Electricity, water, and natural gas consumption of a residential house in Canada from 2012 to 2014. Scientific data, 3(1), 1-12.

Data source: http://ampds.org/. This link is the original place, where the data is located.

There is some useful information on this site. There you will also see a link to an excel spreadsheet of millions of records of the data in a well-curated format. However, the flat files of the dataset (.dat files) are moved to the following link: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/FIE0S4.

From this link you need to download the files to work on this project. As you can see there are many files of data collected from the smart house.

Please note you need to read the earlier mentioned publication to familiarize yourself with the dataset.

## Progress Stages and Deadlines:

2- **Feb/11/2025 - Data Pre-processing (20%):** You will use data mining techniques to explore and report the characteristics of the big data aka. exploratory data analysis (EDA) in order to define your domain-specific problem and proper solution using classification, association rule mining, and clustering techniques.

a. List all data types and standard stats (mean, min, max, std., etc.).

b. Make sure all attributes (per variable) have the same scale and data format.

c. Report the % of missing values and suggest and discuss a method to impute the missing values.

Note: Implementation of the missing value imputation is optional for this project. A successful implementation will be awarded a bonus mark of 2.

d. Perform outlier detection and discuss how you handle them in your implementation.

e. If you have ordinal variables, discuss how you handle them in your implementation.

f. Visualize all attributes after prepressing using histograms, boxplots, qq-plots. Make sure to add appropriate axis labels and a title to the visualizations.

g. Perform a correlation analysis among the attributes and visualize them using an attribute heatmap (A Python code e.g. can be found here). You can also use the Chi-squared test.

h. Class distribution analysis: Check if the data is imbalanced. Discuss the data distribution across classes. For instance, the percentage of samples found for the positive class and negative class. Elaborate on how you handle the class imbalanced data distribution.

**3- Feb/25/2024 - Database Design (15%):** Based on the collected data, design a database. Show the corresponding E-R diagram to illustrate entities, their data sharing, and possible constraints.

4- **Mar/11/2024 - Database Implementation (25%):** The next step is to implement a NoSQL database management system and store the cleaned data in this database. In the case of a high volume of data, you can use the *sharding* technique in a non-relational DBMS to distribute your large-scale data.

5- **Mar/25/2024 - Data Mining (15%):** After the database building, implement statistical analysis, estimation, or prediction, and visualize the discovered results (e.g., the projections or forecasts are shown in a plot).

**6- Mar/27/2024 - User Interface Design (25%):** A web-based/desktop UI is required to facilitate data-mining querying and result visualization. Using a proper API, connect the database to a UI. Try to implement at least ten comprehensive queries.

## Group dynamics and progress:

The workload should be equally divided among the group members (two members per group). However, each student is expected to be fully conversant with all aspects of the project, including those for which he/she is not responsible for. The following are indicative steps for developing your system:

1. Requirements elicitation
2. System specification, data collection, and modeling
3. Database design and E-R diagram
4. Database implementation
5. Query creation
6. Query implementation
7. User Interface
8. Project management, system testing
9. Report writing
10. A presentation (15 minutes): Presentations need to show all the development steps of the project, the problems you faced, how you solved them, and a critical evaluation of your work.

For assessment rubric, please go to the next page.

## Assessment Rubric:

**Database Systems Final Project Rubric**

| Criteria | Level 4 | Level 3 | Level 2 | Level 1 | Score |
|---|---|---|---|---|---|
| Presentation of Ideas<br>(10%) | 2 points<br>Student is able to express their idea clearly and answers the questions | 1.5 points<br>Student is able to express their ideas and answer most of the questions | 1 point<br>Student is somehow able to express their ideas and answer few questions | 0.5 point<br>Student is not presenting their ideas well and not able to answer most questions | /2 |
| Design<br>(20%) | 4 points<br>The approach of solving the problem clearly described with relevant details | 3 points<br>The approach of solving the problem clearly described with most relevant details | 2 points<br>The approach of solving the problem clearly described but lack some relevant details | 1 point<br>The approach of solving the problem is not clearly described with minimum relevant details | /4 |
| Demonstration of the product<br>(40%) | 8 points<br>Implementation is complete well written with additional features beyond requirements. Demonstration works perfectly as intended | 6 points<br>Implementation is complete with intended solution the demo works | 4 points<br>Implementation complete, some of the project objectives are missing. The demo works with few flaws | 2 points<br>Implementation is questionable, demo is not working properly | /8 |
| Analysis of results<br>(30%) | 6 points<br>Analysis is well presented and shows excellent understanding of the project's results | 4.5 points<br>Analysis is well presented and shows good understanding of the project's results | 3 points<br>Analysis is presented and shows some understanding of the project's results | 1.5 points<br>Analysis is weak, the student does not understand the project's results | /6 |

**Total = /20**