NoSQL Project Description

* **Project Objective and Requirements**
* **Objective**: The objective of this project is to learn NoSQL concepts and develop coding skills by doing a hands-on project of your own interest.
* **NoSQL database**: You can choose any NoSQL database and programming language that you like. However, you are fully responsible for setting up and managing the NoSQL database and implementing your NoSQL queries yourself. Most open-source NoSQL databases, such as MongoDB or Apache Spark, are very well documented with a lot of examples and tutorials. If not, you can find something else. We don’t have time and resources to assist you in this regard.
* **Dataset**: The dataset that you choose to analyze must include not only tabular but also non-tabular data, such as text data or images. For example, product reviews or job posting data that include English descriptions, in addition to tabular attributes, are acceptable. You can find many such datasets online (e.g., Kaggle, Hugging Face, or Google dataset search engine).
* **How to find an unstructured/non-tabular dataset:** For example, look at this dataset: [https://huggingface.co/datasets/jingwora/unstructured-data-multilingual/viewer/default/en?row=0](https://huggingface.co/datasets/jingwora/unstructured-data-multilingual/viewer/default/en?row=0" \t "_new). This dataset is unstructured due to the "product\_detail" and "review" attributes. The data in both of these attributes has no fixed size or data type, i.e., no fixed structure. You need to choose a dataset that has these types of unstructured attributes and use them for analysis. Each proposed analysis task must involve at least one unstructured attribute, along with others. Remember that data stored in a .csv file is not unstructured/ non-tabular. CSV is a text format that uses commas to separate values, storing tabular data in plain text. However, this doesn't magically make tabular data non-tabular.
* **Dataset size**: Your dataset needs to have at least 20 attributes and 500 samples.
* **N+1 requirements**: *To complete the project, a team of N (max 3) students needs to implement N+1 nontrivial NoSQL queries and present the results by plotting graphs or tables.*
  + Implement them using the SQL-like commands provided by most NoSQL databases. However, you are not allowed to use Python data science or data analysis libraries/tools to do data analysis. Neither can you use any machine learning functions or libraries unless you implement them yourself from scratch.
  + Your NoSQL queries need to extract new information from the raw data in the dataset. To this end, each NoSQL query must analyze at least 3 attributes together. Furthermore, one of the attributes analyzed need to be nontabular data.
  + Trivial tasks, such as word count, data format conversion, data load/store, don’t count toward the N+1 requirement.
  + Computing simple statistics, such as the min, max, average, median, rankings, and correlation, doesn’t count either.
* **Plagiarism:** Making minor changes to existing software is plagiarism that is a violation of the Academic Honesty Code.

1. **Example Datasets**

* Kaggle Datasets:<https://www.kaggle.com>
* Google Dataset Search Engine: <https://datasetsearch.research.google.com/>
* [Tweet datasets](https://github.com/shaypal5/awesome-twitter-data#id5)
* Wikipedia Data:<https://towardsdatascience.com/wikipedia-data-science-working-with-the-worlds-largest-encyclopedia-c08efbac5f5c>

1. **Popular NoSQL Databases**

* Batch systems: Apache Spark (If you are interested in Hadoop, use Apache Spark instead of Hadoop, as it is much more efficient.)
* Stream processing engines: Spark Streaming, Apache Storm
* Document database: MongoDB, CouchDB