CS-839 NoSQL Systems

Dr. Vinu Venugopal, Divyam Agrawal, Aniket Mitra & Nandakishore S Menon

End semester project guidelines

The purpose of this use-case study is to learn how to formulate a large-scale data processing and analytics problem using existing NoSQL systems and to gain experience in solving it by using the algorithms, system designs and techniques taught in the class lectures and paper presentations. This task replaces the final exam and will contribute to ***25% of your final grade*** *for the course.*

Instructions

- The project's overall goal is to create **reusable and user-friendly data processing pipelines** using the NoSQL technologies covered in the course. From the perspective of an end user, your system should offer a simple and fast method for loading, processing, and analysing data.
- Your project should adhere to the overall objective and fit under one of the following five sub-categories: multi-modal data processing & analytics, consistency model for distributed systems, structured data processing & analytics, graph data processing & analytics, and optimisations of data processing pipelines.
- You may use any publicly accessible dataset to test your system, while using synthetic datasets is also encouraged. You can also use the datasets that were introduced in the course.
- It is necessary to use at least one of the tools (Apache Hadoop, PIG, HIVE, and MongoDB) covered in the course. You must seek permission from the designated TAs/course instructor if you prefer to utilise additional tools.
- A **group or an individual** can work on a project. There will be more expectations for the group projects. There can be no more than three people in a group. By **April 14th**, the group details must be decided upon and sent to your TAs.
- A project proposal document (using the ACM double column format– template available on overleaf[1]) should be submitted on **April 20th, 8:00 pm**. The student or the group will be informed about the acceptance of their idea within a few days.
- The final project submission is due on **May 15 at 8:00 p.m**. It must include all of the following: (i) source code (excluding the dataset) or the appropriate github link, (ii) a written report, and (iii) a video presentation of your use-case study. and all of which must be submitted in one zip file via Moodle. Below are some specifics on the submission requirements.
- If there is any plagiarism or similarity to other projects or projects submitted to other courses, the submission will not be considered. The instructor or teaching assistants will decide the final points in these situations.
- If the project is simply an expanded version of any of the assignment questions or papers discussed in class, the students' contributions should be clearly stated in both the report and the presentation. The student or group should also explain how the project relates to the overall theme and the sub-categories mentioned above.
- The final report should clearly describe the project's goals and outcomes.
- The submission must also include a short video presentation (approximately 10 minutes) summarising the main steps (and pipeline details) involved in developing your project.
- Your presentation should clearly show the effort that you put in, the challenges you encountered, your hypothesis, and the results you obtained.
- The marks assigned will be based on the quality and depth of (i) your submitted implementation, (ii) the final report, and (iii) the accompanying video presentation.

[1] http://surl.li/gatvk

Report Format

The written report of at least 4 pages (and no more than 6 pages) in the ACM double column format – mentioned above.

1. *Title*: Provide a concise title of your use-case study.
2. *Abstract*: One paragraph which summarizes your main contributions and findings.
3. *Problem Definition:*
   - This part should include one or two carefully crafted paragraphs that state and highlight the problem setting. It should be defined in a way that answers the following questions:
     - What is the problem you are trying to solve? This should include the background description of the problem, and how is it related to the overall goal/theme and sub-categories.
     - Why is this problem/task/application interesting from a NoSQL systems perspective?
     - What are the main results/insights you intend to obtain?
4. *Approach:*
   - Next, your proposed approach to solve the problem should be described in more detail. It should include:
     - The algorithms/techniques/models that you have used in this study.
     - Justifications for using a specific tool or a specific combination of tools.
     - List one or two specific examples.
     - How does your approach make it simple and easy to build reusable pipelines that are user-friendly?
     - The dataset(s) that you have used in this project.
5. *Evaluation & Results:*
   - The experimental results (in terms of both result quality and efficiency/runtimes) should be briefly evaluated on different parameter settings. The students are expected to use large datasets.
6. *References:*
   - You must cite every source you used, including any datasets, libraries, and books, if any, in your report.

Video Presentation

A short video presentation of at least 10 minutes (but, no more than 20 minutes) should once more summarize the main features of your project.

- Your presentation should cover the content discussed in your report and implementation. Please create at least 7 slides, including:
  1. Slide 1: Short self-introduction[2]
  2. Slide 2: Problem Statement and Motivation
  3. Slide 3: Approach
  4. Slide 4: Implementation
  5. Slide 5: System demo
  6. Slide 5: Evaluation
  7. Slide 6: Conclusions & Lessons Learned

[2]Please briefly identify yourself with your webcam on. After this, the webcam is no longer mandatory, but please provide at least a voice recording of your further presentation and slides. In case you have any privacy concerns regarding this point, we will try to arrange for a remote presentation of your project via a live Teams meeting instead.