TP3: Comparison between “Big Data – NoSQL” data models.

This assignment focuses on the subject of Software Architectures for Big Data. As discussed in class, NoSQL databases follow various data models, each with its own advantages and disadvantages. In this assignment, you are tasked with writing an academic paper to present a comprehensive comparison between 3 of these models and their respective implementations. More specifically, you have the following tasks:

1. Select **three** different data models and present their details and their specifications. Compare the models with respect to their documented properties and the general attributes of Big Data architectures.
2. Select one concrete implementation for each data model from the previous step. Compare the implementations with respect to their documented properties and the general goals of Big Data architectures.
3. Use the Yahoo! Cloud Service Benchmark (https://github.com/brianfrankcooper/YCSB). Deploy an instance of each of your selected implementations and test their performance using the workloads provided by the benchmark
4. Write a script in any programming language to automate the benchmarking process. The script is just responsible to perform different kind of workloads with one command. You can put the bash command inside of the script. The simplest way is to write a small bash script.
5. Repeat the experiment at least for 3 times to get the average throughput.
6. Push your script and results in a Bitbucket/GitHub repository and make it public.
7. Provide a comprehensive comparative evaluation for all three implementations and discuss the differences in the performance.

To further help you with the assignment, here are some questions that you may want to answer in your paper:

- What are the basic challenges and properties of Big Data? Why are traditional database systems insufficient in handling Big Data? What makes NoSQL systems better?
- For each of the three data models: What are the nominal and documented properties of the data model? How do they address the challenges of Big Data? What are their benefits and what are their weaknesses?
- How do the three data models compare based on their documented properties? What use cases are benefited the most by each data model?
- For each of the three implementations: What are the nominal and documented properties of the implementation? How do they implement the basic properties of the respective data model? What are their benefits and weaknesses?
- How does the performance of the three implementations compare against similar workloads? Which one is best in read latency, write latency, throughput? How do the three implementations compare against different workloads? What kind of workloads are more suitable for each implementation?

Submission Guidelines

- Submit a paper in IEEE format (https://www.ieee.org/conferences/publishing/templates.html) of at least 8 pages (maximum 10 pages).
- Your paper should definitely have an abstract, keywords, an Introduction, related works (minimum 5 papers), methodology, results and conclusions. You can organize the rest of the content as you see fit.
- Provide a meaningful title (not just “Assignment 3”) and give the author names and affiliations as specified by the template.
- This is an academic paper and you will need to consult numerous sources (other papers, documentation, possibly online posts). Make sure you properly cite your sources and give credit. Anything that does not have a citation will be considered your contribution and will be subjected to judgement. So, try to support as many of your arguments as possible with proper sources. DO NOT CITE WIKIPEDIA! Instead you can consult the references that a Wikipedia page already cites. They usually contain the information you are looking for. If you provide online documentation or blog posts (as a last resort), also provide the date you last accessed the source.
- Allocate enough time to set up and run the benchmark experiments with YCSB.
- Along with the paper, you will need to prepare and submit a presentation on your paper. The presentation will be given on the 1th of December during regular class time by all team members. Each team will be allocated 12 minutes for the presentation and 2-3 minutes for questions.
- After the presentation you will have another week to submit the final version of your paper on the 8th of December. You may have to take into account feedback from the presentations to finish your paper.
- You will receive 70% for the paper and 30% for the presentation.
- **Presentation submission deadline: December 1th 23:59**
- **Paper submission deadline: December 8th 23:59**